

ANALISIS DAN VISUALISASI DATA



Penulis:

Wulan Sri Lestari, S.Kom., M.Kom.
Mustika Ulina, S.Kom., M.Kom.
Manto Lumban Gaol, S.Pd., M.Pd.

ANALISIS DAN VISUALISASI DATA

Penulis:

Wulan Sri Lestari, S.Kom., M.Kom.

Mustika Ulina, S.Kom., M.Kom.

Manto Lumban Gaol, S.Pd., M.Pd.



GET PRESS INDONESIA

ANALISIS DAN VISUALISASI DATA

Penulis :

Wulan Sri Lestari, S.Kom., M.Kom.

Mustika Ulina, S.Kom., M.Kom.

Manto Lumban Gaol, S.Pd., M.Pd.

ISBN : 978-623-125-492-4

Editor : Henda Kusuma, M.Kom.

Desain Sampul dan Tata Letak : Tri Putri Wahyuni, S.Pd.

PENERBIT GET PRESS INDONESIA

Anggota IKAPI No. 033/SBA/2022

Jl. Palarik RT 01 RW 06 Kelurahan Air Pacah
Kecamatan Koto Tangah Padang Sumatera Barat

website: www.getpress.co.id

email: adm.getpress@gmail.com

Cetakan pertama, November 2024

Hak cipta dilindungi undang-undang

Dilarang memperbanyak karya tulis ini dalam bentuk
dan dengan cara apapun tanpa izin tertulis dari penerbit.

KATA PENGANTAR

Dengan mengucapkan puji syukur kehadiran Allah SWT, atas limpahan rahmat dan hidayahNya, maka Penulisan Buku dengan judul Analisis dan Visualisasi Data dapat diselesaikan. Buku ini berisikan bahasan tentang Literasi Data, Big Data, Statistika, Analitika Data, Data Storytelling, Visualisasi Data, Alat Visualisasi Data, dan Dashboard dan Story.

Buku ini masih banyak kekurangan dalam penyusunannya. Oleh karena itu, kami sangat mengharapkan kritik dan saran demi perbaikan dan kesempurnaan buku ini selanjutnya. Kami mengucapkan terima kasih kepada berbagai pihak yang telah membantu dalam proses penyelesaian Buku ini. Semoga Buku ini dapat menjadi sumber referensi dan literatur yang mudah dipahami.

Medan, Oktober 2024

Penulis

DAFTAR ISI

KATA PENGANTAR i

DAFTAR ISI..... ii

DAFTAR GAMBAR v

BAB 1 LITERASI DATA..... 1

 1.1 Pendahuluan 1

 1.2 Karakteristik Literasi Data 3

 1.3 3 C’s Literasi Data..... 6

 1.4 Payung Literasi Data 9

 1.5 Tantangan Literasi Data 13

 1.6 Rangkuman..... 14

DAFTAR PUSTAKA 16

BAB 2 BIG DATA 17

 2.1 Pendahuluan 17

 2.2 Karakteristik Big Data 18

 2.3 Sumber Big Data 22

 2.4 Arsitektur Big Data 25

 2.5 Teknologi dan Alat Big Data..... 27

 2.6 Implementasi Big Data di Berbagai Industri 30

 2.7 Rangkuman..... 33

DAFTAR PUSTAKA 35

BAB 3 STATISTIKA 37

 3.1 Pendahuluan 37

 3.2 Populasi dan Sampel 38

 3.3 Teknik Sampling 40

3.4 Jenis-Jenis Data.....	43
3.5 Statistika Deskriptif.....	45
3.6 Rangkuman.....	50
DAFTAR PUSTAKA.....	52
BAB 4 ANALITIKA DATA.....	53
4.1 Pendahuluan.....	53
4.2 Jenis-Jenis Analitika Data.....	56
4.3 Tahapan Analitika Data.....	61
4.4 Rangkuman.....	82
DAFTAR PUSTAKA.....	84
BAB 5 DATA STORYTELLING.....	85
5.1 Pendahuluan.....	85
5.2 Komponen Data Storytelling.....	86
5.3 Contoh Data Storytelling.....	88
5.4 Rangkuman.....	90
DAFTAR PUSTAKA.....	92
BAB 6 VISUALISASI DATA.....	93
6.1 Pendahuluan.....	93
6.2 Prinsip Visualisasi Data.....	95
6.3 Jenis-Jenis Visualisasi Data.....	101
6.4 Tahapan Visualisasi Data.....	103
6.5 Rangkuman.....	105
DAFTAR PUSTAKA.....	106
BAB 7 ALAT VISUALISASI DATA.....	107
7.1 Pendahuluan.....	107

7.2 Komponen Fungsional Tableau.....111

7.3 Jenis-Jenis Grafik Tableau.....112

7.4 Rangkuman.....132

DAFTAR PUSTAKA..... 133

BAB 8 DASHBOARD dan STORY 135

8.1 Pendahuluan135

8.2 Dashboard.....136

8.3 Story140

8.4 Rangkuman.....142

DAFTAR PUSTAKA..... 144

DAFTAR GAMBAR

Gambar 1. 1 Karakteristik Literasi Data.....	3
Gambar 1. 2 3 C's Literasi Data.....	7
Gambar 1. 3 Payung Literasi Data	9
Gambar 2. 1 Karakteristik Big Data	18
Gambar 2. 2 Arsitektur Big Data	25
Gambar 3. 1 Ilustrasi Populasi dan Sampel	39
Gambar 3. 2 Populasi dan Sampel	39
Gambar 4. 1 Jenis-Jenis Analitika Data	56
Gambar 4. 2 Analisis Deskriptif.....	57
Gambar 4. 3 Analisis Diagnostik.....	58
Gambar 4. 4 Analisis Prediktif.....	59
Gambar 4. 5 Analisis Preskriptif.....	60
Gambar 4. 6 Tahapan Data Analitik	61
Gambar 4. 7 Ilustrasi Data Requirement Gathering.....	62
Gambar 4. 8 Metode Pengumpulan Data	65
Gambar 4. 9 Proses Umum Data Cleaning.....	67
Gambar 5. 1 Komponen Data Storytelling.....	86
Gambar 6. 1 Kategori Grafik.....	97
Gambar 7. 1 Tampilan Workspace Tableau Public.....	109
Gambar 7. 2 Ilustrasi Komponen Fungsional Tabelau	111
Gambar 7. 3 Menghubungkan Data ke Tableau	113
Gambar 7. 4 Grafik Bar Chart Horizontal.....	114
Gambar 7. 5 Grafik Bar Chart Vertical	115
Gambar 7. 6 Grafik Stacked Bar	116
Gambar 7. 7 Garfik Side by Side Bars	117
Gambar 7. 8 Grafik Line Chart.....	118
Gambar 7. 9 Grafik Dual Line Chart	119
Gambar 7. 10 Grafik Pie Chart.....	119
Gambar 7. 11 Grafik Text Tables.....	120
Gambar 7. 12 Grafik Heat Maps.....	121
Gambar 7. 13 Grafik Highligt Tables.....	121
Gambar 7. 14 Filter 10 Produk	122

Gambar 7. 15 Grafik Tree Maps.....	123
Gambar 7. 16 Grafik Tree Maps.....	123
Gambar 7. 17 Grafik Area Chart	124
Gambar 7. 18 Grafik Dual Combination	125
Gambar 7. 19 Grafik Scatter Plots	126
Gambar 7. 20 Cara Membaca Box Plot.....	126
Gambar 7. 21Grafik Box Plot.....	127
Gambar 7. 22 Grafik Bullet Graphs	129
Gambar 7. 23 Grafik Gantt Chart.....	130
Gambar 7. 24 Grafik Histogram.....	130
Gambar 7. 25 Grafik Packed Bubbles	131
Gambar 8. 1 Dashboard Penjualan Superstore	140
Gambar 8. 2 Story Overview Penjualan	141
Gambar 8. 3 Story Perbandingan Discount-Sales-Profit	142

DAFTAR TABEL

Tabel 1. 1 Pentingnya Literasi Data	5
---	---

BAB 1

LITERASI DATA

1.1 Pendahuluan

Data telah menjadi bagian integral dari kehidupan kita sehari-hari, dari penggunaan teknologi hingga pengambilan keputusan dalam bisnis. Secara sederhana, data dapat didefinisikan sebagai informasi mentah yang belum diolah. Data dapat berupa angka, teks, gambar, atau sinyal-sinyal lain yang dapat diolah untuk menghasilkan informasi yang lebih berguna (Morrow, 2021). Dalam skala yang lebih besar, data dapat membantu organisasi, pemerintah, dan individu untuk memahami fenomena dan membuat keputusan yang lebih tepat. Misalnya, dalam bidang kesehatan, data dapat digunakan untuk melacak penyebaran penyakit dan menentukan metode penanganan yang paling efektif.

Pentingnya data semakin terlihat jelas saat kita memasuki era Revolusi Industri 4.0, di mana teknologi-teknologi canggih seperti kecerdasan buatan (AI), *Internet of Things* (IoT), dan analisis *Big Data* telah memainkan peran utama. Saat ini, pengambilan keputusan berbasis data menjadi hal yang umum, bukan hanya di dunia bisnis, tetapi juga di bidang pendidikan, kesehatan, pemerintahan, dan lain-lain (Marr, 2022). Data memberikan wawasan yang tidak dapat diperoleh hanya dengan perkiraan atau insting, sehingga pengelolaan data yang baik sangat penting untuk mencapai keunggulan kompetitif. Namun, pentingnya data juga menuntut keahlian khusus. Tidak hanya kemampuan teknis dalam mengumpulkan, mengelola, dan menganalisis data,

tetapi juga keterampilan dalam memahami dan menginterpretasi data tersebut (Morrow, 2021).

Banyak perusahaan mengalokasikan dana besar untuk investasi dalam data dan analitik, namun hasil yang mereka peroleh sering kali tidak sesuai harapan. Morrow (2021) mengidentifikasi beberapa faktor penyebab utama, termasuk kurangnya pelatihan yang memadai, kesulitan dalam membaca dan menganalisis data, serta minimnya pemahaman tentang bagaimana data dapat mendukung pengambilan keputusan yang tepat. Dalam konteks ini, literasi data muncul sebagai keterampilan kunci. Literasi data bukan hanya tentang mampu membaca data, tetapi juga memahami, mengelola, dan menggunakannya untuk pengambilan keputusan yang tepat. Menurut penelitian yang dilakukan oleh Harvard Business Review (2018), literasi data adalah kemampuan yang sangat penting di era digital ini karena individu dan organisasi yang tidak memiliki literasi data berisiko tertinggal dalam persaingan global. Literasi data memungkinkan individu untuk mengajukan pertanyaan yang tepat, mengenali pola dalam data, serta memahami keterbatasan dan potensi data. Hal ini menjadikan literasi data sebagai salah satu keterampilan yang paling dicari dalam dunia kerja saat ini.

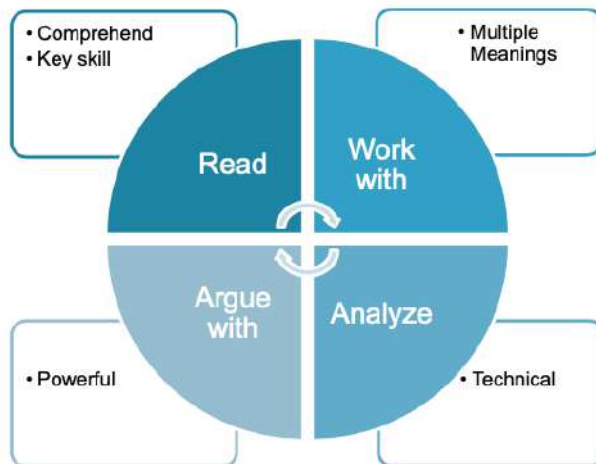
Namun, meski data memiliki potensi besar, penggunaannya juga memiliki tantangan. Salah satu tantangan terbesar adalah validitas dan integritas data. Data yang tidak valid atau diolah secara salah dapat menghasilkan kesimpulan yang keliru dan berdampak buruk pada pengambilan keputusan. Oleh karena itu, penting untuk tidak hanya mengumpulkan data dalam jumlah besar, tetapi juga memastikan kualitas data yang digunakan.

Dengan demikian, literasi data bukan hanya menjadi kompetensi teknis, tetapi juga mencakup pemahaman etika dan tanggung jawab dalam menggunakan data. Dalam banyak kasus, data yang salah kelola dapat berdampak buruk, seperti penyalahgunaan privasi atau pengambilan keputusan yang

merugikan banyak pihak (Cathy O'Neil, 2016). Oleh karena itu, para profesional yang bekerja dengan data harus memiliki keterampilan yang baik dalam menganalisis dan memitigasi risiko ini. Melalui pengembangan literasi data, individu dapat mengambil peran yang lebih aktif dalam dunia digital, menghindari jebakan kesalahan interpretasi, dan memberdayakan diri mereka dengan alat yang diperlukan untuk mendorong inovasi. Literasi data menjadi fondasi dalam era data-driven, di mana setiap orang memiliki kesempatan untuk memanfaatkan data secara efektif dalam kehidupan sehari-hari maupun dalam karier mereka.

1.2 Karakteristik Literasi Data

Memiliki literasi data yang baik memungkinkan kita untuk mengambil keputusan yang lebih tepat berdasarkan fakta dan bukti. Untuk lebih memahami pentingnya literasi data, mari kita pahami empat karakteristik utama literasi data seperti terlihat pada Gambar 1.1.



Source: QlikTech International (2019), *A Culture of Data Literacy*

Gambar 1. 1 Karakteristik Literasi Data

a. Membaca Data

Membaca data adalah langkah pertama dan paling mendasar dalam literasi data. Kemampuan ini berarti dapat memahami dan menginterpretasikan informasi yang disajikan dalam bentuk angka, tabel, grafik, atau visualisasi lainnya. Membaca data tidak hanya sekadar melihat angka, tetapi juga memahami konteks dan pesan di balik data tersebut. Misalnya, ketika melihat grafik tren penjualan, kita harus mampu mengidentifikasi pola, perubahan, atau anomali yang mungkin terjadi. Dengan pemahaman yang baik tentang cara membaca data, kita dapat lebih mudah menemukan informasi penting yang berguna dalam pengambilan keputusan.

b. Bekerja dengan Data

Setelah membaca data, langkah selanjutnya adalah bekerja dengan data. Ini mencakup kemampuan untuk mengumpulkan, mengelola, dan memanipulasi data. Bekerja dengan data memerlukan pemahaman tentang berbagai alat dan teknik yang dapat membantu dalam pengolahan data, seperti menggunakan Excel, database, atau perangkat lunak analitik lainnya. Kemampuan ini penting untuk memastikan bahwa data yang kita kelola adalah data yang akurat, relevan, dan siap untuk dianalisis lebih lanjut. Dengan bekerja secara efektif dengan data, kita dapat mempersiapkan informasi yang tepat untuk mendukung analisis yang lebih mendalam.

c. Menganalisis Data

Menganalisis data adalah kemampuan untuk menggali wawasan dari data yang telah dikumpulkan dan diolah. Proses ini melibatkan penggunaan teknik-teknik statistik atau analisis lainnya untuk menemukan pola, tren, atau hubungan dalam data. Menganalisis data memungkinkan kita untuk membuat kesimpulan berdasarkan bukti yang

ada, bukan hanya dugaan atau asumsi. Misalnya, analisis data penjualan dapat membantu kita memahami faktor-faktor yang mempengaruhi kenaikan atau penurunan penjualan, sehingga kita dapat membuat strategi yang lebih tepat untuk masa depan. Kita akan membahas lebih detail tentang menganalisis data pada bab berikutnya.

d. Berkomunikasi dengan Data

Kemampuan untuk berkomunikasi dengan data adalah keterampilan penting dalam literasi data. Hal ini berarti dapat menyampaikan hasil analisis data kepada orang lain dengan cara yang jelas dan mudah dipahami. Komunikasi data sering kali melibatkan penggunaan grafik, diagram, atau infografis untuk menyederhanakan informasi yang kompleks. Kemampuan ini sangat penting karena tidak semua orang memiliki keahlian yang sama dalam memahami data mentah. Dengan menyampaikan data secara efektif, kita dapat memastikan bahwa pesan yang ingin disampaikan diterima dengan baik dan dapat digunakan untuk pengambilan keputusan. Kita akan membahas lebih detail tentang menganalisis data pada bab berikutnya.

Tabel 1, menjelaskan mengapa literasi data sangat penting.

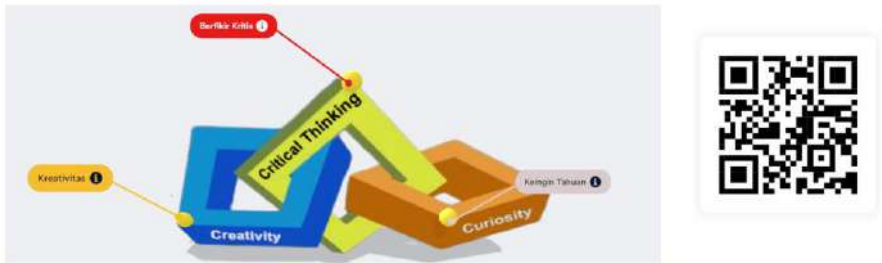
Tabel 1. 1 Pentingnya Literasi Data

Literasi Data adalah kemampuan untuk	Tetapi banyak perusahaan yang kesulitan untuk
Membaca data	Memahami cara menggunakan sebagian besar data yang mereka kumpulkan dari pelanggan dan sistem internal.

Literasi Data adalah kemampuan untuk	Tetapi banyak perusahaan yang kesulitan untuk
	• Mengajukan pertanyaan yang tepat tentang data mereka dan membangun eksperimen untuk menjawab pertanyaan-pertanyaan tersebut.
Bekerja dengan data	Menyimpan, membersihkan, dan mengelola data dengan efisien untuk menggunakan data yang tepat pada waktu yang tepat.
Analisis data	• Memfilter, menyortir, membandingkan, dan menguji validitas data. • Mengembangkan analisis yang bermakna dan dapat diandalkan dari temuan data.
Berkomunikasi dengan data	• Membuat visualisasi data agar semua pemangku kepentingan dapat memahaminya. • Menceritakan temuan data sehingga perusahaan dapat mengambil tindakan yang menghasilkan hasil bisnis yang positif.

1.3 3 C's Literasi Data

3 C's literasi data adalah kerangka kerja yang dapat membantu individu memahami dan memanfaatkan data dengan lebih efektif. 3 C's ini mencakup tiga aspek penting dari literasi data yaitu: Curiosity (keingintahuan), Creativity (kreativitas), dan Critical Thinking (berpikir kritis) seperti terlihat pada Gambar 1.2. Pada Gambar 1.2, Anda dapat mengscan barcode yang diberikan untuk melihat 3 C's literasi data dalam bentuk *Augmented Reality* (AR).



Gambar 1. 2. 3 C's Literasi Data

Mari kita bahas masing-masing komponen tersebut.

a. **Curiosity (keingintahuan)**

Keingintahuan adalah dasar dari literasi data. Anak-anak secara alami memiliki rasa ingin tahu yang tinggi, selalu bertanya tentang hal-hal di sekitar mereka. Namun, seiring bertambahnya usia, banyak orang kehilangan rasa ingin tahu ini. Dalam konteks literasi data, keingintahuan mendorong kita untuk tidak hanya menerima data begitu saja, tetapi juga untuk mengajukan pertanyaan yang mendalam dan mencari makna di balik data tersebut.

Misalnya, seorang analis data yang melihat laporan penjualan mungkin merasa penasaran mengapa penjualan meningkat tajam pada bulan tertentu. Dengan rasa ingin tahu, analis tersebut akan menyelidiki lebih dalam untuk menganalisis data, mengevaluasi tren, dan mencari faktor-faktor yang memengaruhi lonjakan penjualan. Keingintahuan ini dapat membantu menemukan pola atau masalah yang mungkin tidak terlihat sebelumnya, yang dapat mengarah pada keputusan yang lebih baik.

Contoh: Misalkan sebuah perusahaan e-commerce mencatat lonjakan penjualan produk tertentu selama kampanye promosi. Dengan rasa ingin tahu, analis dapat mengeksplorasi data kampanye secara mendalam, mencari tahu faktor-faktor apa yang berkontribusi pada

keberhasilan kampanye tersebut, seperti penempatan iklan, waktu promosi, atau diskon yang ditawarkan.

b. *Creativity* (kreativitas)

Kreativitas dalam literasi data melibatkan kemampuan untuk berpikir di luar kebiasaan dan mengembangkan cara baru dalam menyajikan dan menganalisis data. Hal ini termasuk menciptakan visualisasi data yang inovatif, menyusun laporan yang menarik, dan merancang solusi analitis yang unik. Kreativitas membantu kita menyajikan data dengan cara yang lebih menarik dan mudah dipahami, serta menemukan solusi baru untuk masalah yang kompleks.

Contoh: Dalam konteks data visualisasi, alat seperti Tableau atau Power BI memungkinkan seorang data scientist dapat menciptakan dashboard interaktif yang memungkinkan pengguna untuk mengeksplorasi data secara real-time. Misalnya, dalam sektor kesehatan, kreativitas dapat digunakan untuk membangun *heat map* yang menunjukkan penyebaran penyakit di berbagai daerah, yang membantu tim medis dalam merencanakan intervensi dengan lebih efektif.

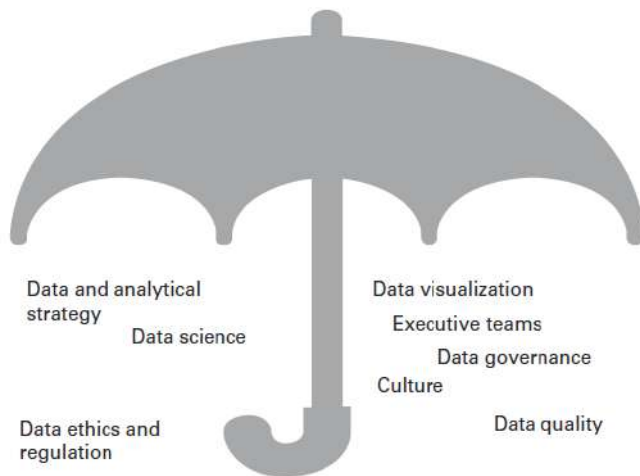
c. *Critical Thinking* (berpikir kritis)

Berpikir kritis adalah kemampuan untuk mengevaluasi data secara objektif, mempertanyakan keakuratan, relevansi, dan sumber informasi. Hal ini melibatkan analisis mendalam untuk mengidentifikasi bias, kesalahan, dan keterbatasan dalam data. Berpikir kritis memungkinkan kita untuk membuat keputusan yang lebih baik dan memastikan bahwa kesimpulan yang diambil didasarkan pada analisis yang solid dan terpercaya.

Contoh: Seorang peneliti yang menganalisis data survei kepuasan pelanggan perlu menggunakan pemikiran kritis untuk mengevaluasi kualitas data. Misalnya, mereka harus memeriksa apakah ada bias dalam sampel survei, memastikan bahwa pertanyaan yang diajukan tidak mempengaruhi jawaban, dan menilai apakah hasil survei benar-benar mewakili opini pelanggan secara keseluruhan.

1.4 Payung Literasi Data

Setiap komponen dalam payung literasi data memainkan peran penting dalam membangun kapabilitas data yang kuat dalam organisasi. Gambar 1.3 merupakan payung literasi data (Marrow, 2021).



Gambar 1. 3 Payung Literasi Data

a. **Data and Analytical Strategy**

Data and analytical strategy merupakan fondasi utama bagi organisasi untuk memulai perjalanan data mereka. Strategi yang kuat membantu memastikan bahwa

semua inisiatif data sejalan dengan tujuan bisnis dan dapat dieksekusi dengan efektif. Data and analytical strategy melibatkan perencanaan dan pelaksanaan pendekatan untuk mengumpulkan, mengelola, dan menganalisis data. Hal ini termasuk penentuan tujuan analitik, pemilihan alat dan metode yang sesuai, serta pengembangan rencana untuk menyelaraskan kegiatan analitik dengan kebutuhan bisnis atau penelitian. Strategi ini penting untuk memastikan bahwa data yang dikumpulkan dan dianalisis memberikan wawasan yang berharga dan relevan.

b. Data Science

Literasi data sangat terkait dengan ilmu data (data science). Data Science adalah bidang yang menggabungkan teknik statistika, pemrograman, dan analisis data untuk mengekstrak wawasan dan pengetahuan dari data besar. Data Science melibatkan pembuatan model prediktif, algoritma pembelajaran mesin, dan analisis statistik untuk membantu organisasi membuat keputusan berbasis data. Hal ini juga mencakup pembersihan data, eksplorasi data, dan interpretasi hasil analisis. Meskipun ilmu data fokus pada teknik dan alat untuk menganalisis data, literasi data memastikan bahwa individu dalam organisasi dapat memahami dan menggunakan hasil analisis tersebut untuk pengambilan keputusan yang lebih baik.

c. Data Ethics and Regulation

Aspek etika dan regulasi sangat penting dalam literasi data. Etika dan Regulasi Data mengacu pada prinsip dan peraturan yang mengatur penggunaan dan perlindungan data pribadi dan sensitif. Hal ini mencakup kepatuhan terhadap hukum privasi seperti GDPR (General Data Protection Regulation) dan HIPAA (Health Insurance Portability and Accountability Act), serta pertimbangan

etis seperti privasi, transparansi, dan keadilan dalam pengolahan data. Memahami etika dan regulasi ini penting untuk memastikan bahwa data digunakan secara bertanggung jawab dan sesuai dengan standar hukum. Organisasi harus memastikan bahwa mereka mematuhi semua regulasi terkait data dan mengadopsi praktik etis dalam pengelolaan dan penggunaan data, termasuk privasi data, keamanan data, dan penggunaan data untuk tujuan yang baik.

d. Data Visualization

Visualisasi data adalah komponen kunci dari literasi data. Visualisasi yang efektif membantu dalam menyampaikan informasi kompleks dengan cara yang mudah dipahami. Visualisasi Data adalah teknik untuk menyajikan data dalam bentuk grafis atau visual untuk memudahkan pemahaman dan interpretasi. Hal ini termasuk penggunaan grafik, diagram, peta, dan dashboard untuk menampilkan pola, tren, dan hubungan dalam data. Visualisasi yang efektif membantu dalam menyampaikan informasi kompleks dengan cara yang mudah dimengerti dan dapat mendukung pengambilan keputusan yang lebih baik.

e. Executive Teams

Tim eksekutif harus memiliki pemahaman yang baik tentang literasi data untuk membuat keputusan strategis. Literasi data di tingkat eksekutif memastikan bahwa strategi data diintegrasikan dengan baik ke dalam keseluruhan strategi bisnis. Tim Eksekutif memainkan peran penting dalam mendorong budaya dan strategi data di organisasi. Mereka bertanggung jawab untuk membuat keputusan strategis terkait dengan investasi dalam teknologi data, pengembangan kebijakan data, dan pengelolaan sumber daya data. Dukungan dan komitmen dari tim eksekutif sangat penting untuk kesuksesan

inisiatif data dan untuk memastikan bahwa data digunakan untuk mencapai tujuan bisnis yang lebih besar.

f. Data Governance

Data governance mencakup struktur dan kebijakan yang mengatur pengelolaan data dalam organisasi. Hal ini melibatkan penetapan tanggung jawab, standar kualitas data, prosedur akses, dan kontrol keamanan untuk memastikan data dikelola dengan baik dan digunakan secara efektif. Data governance bertujuan untuk memastikan integritas, akurasi, dan kepatuhan data terhadap kebijakan internal dan regulasi eksternal. Data governance yang baik membantu dalam mengurangi risiko dan memastikan bahwa data digunakan secara etis dan sesuai dengan regulasi. Data governance yang kuat melibatkan peran serta dari seluruh tingkat organisasi.

g. Culture

Budaya organisasi yang mendukung literasi data sangat penting untuk keberhasilan inisiatif data. Membentuk budaya yang menghargai dan memprioritaskan penggunaan data dalam pengambilan keputusan membuat organisasi lebih cepat beradaptasi dengan perubahan. Untuk mencapai ini, pelatihan dan pengembangan berkelanjutan dalam literasi data sangat diperlukan. Budaya ini mencerminkan sikap dan nilai yang mendasari cara data digunakan dan dihargai dalam organisasi. Budaya data yang kuat menjadikan data sebagai aset strategis, mendorong kolaborasi dalam analisis, dan mendukung pembelajaran teknologi serta metode analitik yang terus berkembang. Dengan membangun budaya ini, organisasi dapat berhasil dalam transformasi digital mereka.

h. Data Quality

Kualitas data adalah aspek penting dalam literasi data karena data yang berkualitas tinggi menjamin analisis yang andal dan valid. Untuk menjaga kualitas data, dibutuhkan beberapa langkah yang dapat dilakukan antara lain pembersihan, validasi, dan pemantauan data secara berkelanjutan. Kualitas data mengukur seberapa akurat, konsisten, relevan, dan dapat dipercaya data tersebut. Data yang berkualitas tinggi memastikan analisis dan keputusan yang dibuat berdasarkan data tersebut tepat dan dapat diandalkan. Proses ini mencakup penghapusan duplikasi, perbaikan kesalahan, dan menjaga integritas data sepanjang siklus hidupnya.

1.5 Tantangan Literasi Data

Berikut merupakan beberapa tantangan literasi data:

- a. **Keterampilan dan Pengetahuan Terbatas:** Banyak individu dan organisasi mengalami kesulitan karena keterbatasan keterampilan dan pengetahuan dalam analisis data. Ini termasuk kurangnya pemahaman tentang teknik statistik, alat analitik, dan metode pemrograman. Tanpa keterampilan yang memadai, sulit untuk mengolah dan menginterpretasi data secara efektif.
- b. **Kualitas Data yang Buruk:** Data yang tidak akurat, tidak konsisten, atau tidak lengkap dapat menyebabkan kesalahan dalam analisis dan keputusan. Memastikan kualitas data memerlukan upaya untuk membersihkan, memvalidasi, dan memantau data secara terus-menerus.
- c. **Integrasi Data dari Sumber yang Berbeda:** Menggabungkan data dari berbagai sumber yang berbeda dapat menjadi tantangan, terutama ketika format data, struktur, atau kualitasnya bervariasi. Integrasi data yang efektif memerlukan teknik dan alat khusus untuk menyelaraskan data yang heterogen.

- d. Kepatuhan terhadap Regulasi dan Etika Data:** Mematuhi regulasi perlindungan data dan etika penggunaan data adalah tantangan besar, terutama dengan peraturan yang terus berkembang seperti GDPR dan CCPA. Organisasi harus memastikan bahwa data digunakan secara sah dan etis.
- e. Kurangnya Dukungan Organisasi dan Budaya Data:** Tanpa dukungan dari tingkat eksekutif dan tanpa adanya budaya yang mendukung penggunaan data, inisiatif literasi data cenderung kurang efektif. Budaya organisasi yang mengabaikan pentingnya data dapat menghambat adopsi dan penggunaan analitik.
- f. Teknologi dan Alat yang Cepat Berkembang:** Teknologi dan alat analitik data terus berkembang dengan cepat, sehingga memerlukan pembelajaran berkelanjutan dan penyesuaian dari para profesional. Mengikuti perkembangan ini dapat menjadi tantangan besar.
- g. Komunikasi dan Interpretasi Data:** Menerjemahkan data kompleks ke dalam informasi yang mudah dipahami dan relevan untuk berbagai pemangku kepentingan bisa menjadi sulit. Komunikasi yang efektif memerlukan keterampilan untuk menyajikan data dengan cara yang jelas dan informatif.

1.6 Rangkuman

Data telah menjadi elemen penting dalam kehidupan sehari-hari dan berbagai bidang seperti bisnis, kesehatan, dan pemerintahan, dengan kemampuan untuk meningkatkan pengambilan keputusan dan efisiensi. Di era Revolusi Industri 4.0, data, bersama dengan teknologi canggih seperti AI dan Big Data, semakin memegang peranan krusial. Namun, keberhasilan pemanfaatan data tidak hanya bergantung pada pengumpulan data yang besar, tetapi juga pada kualitas dan keahlian dalam mengelolanya. Literasi data, yang mencakup

kemampuan membaca, bekerja, menganalisis, dan berkomunikasi dengan data, menjadi keterampilan kunci yang dicari di dunia kerja saat ini.

Meskipun penting, literasi data menghadapi beberapa tantangan, termasuk keterampilan dan pengetahuan yang terbatas, kualitas data yang buruk, dan integrasi data dari sumber yang berbeda. Kepatuhan terhadap regulasi dan etika data juga menjadi perhatian utama, bersama dengan perlunya dukungan organisasi dan budaya data yang mendukung. Selain itu, perkembangan teknologi yang cepat dan kebutuhan untuk komunikasi data yang efektif menambah kompleksitas.

Untuk mengatasi tantangan ini dan memanfaatkan potensi data secara optimal, penting bagi individu dan organisasi untuk mengembangkan literasi data yang komprehensif. Ini termasuk memahami strategi data, etika, dan kualitas data, serta memiliki kemampuan untuk beradaptasi dengan teknologi terbaru dan berkomunikasi secara efektif. Dengan membangun kompetensi literasi data yang kuat, organisasi dapat memastikan bahwa data digunakan secara strategis untuk mencapai tujuan bisnis dan inovasi yang lebih baik.

DAFTAR PUSTAKA

- Morrow, J. (2021). Be Data Literate: The Data Literacy Skills Everyone Needs to Succeed. Kogan Page.
- Marr, B. (2022). Data Strategy: How to Profit from a World of Big Data, Analytics and the Internet of Things. Kogan Page.
- Harvard Business Review. (2018). Data Literacy: What It Is and Why It Matters. Harvard Business Publishing.
- O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown.
- Apa itu literasi data? – Shout Hub . (2019). Shout-Hub.eu. <https://shout-hub.eu/sfwd-lessons/what-is-data-literacy/>

BAB 2

BIG DATA

2.1 Pendahuluan

Di era digital saat ini, data menjadi salah satu aset terpenting bagi individu, perusahaan, dan organisasi di seluruh dunia. Setiap aktivitas yang kita lakukan, baik secara online maupun offline, menghasilkan data. Mulai dari penggunaan media sosial, transaksi e-commerce, hingga sensor pada perangkat Internet of Things (IoT), semua itu berkontribusi pada ledakan data yang kita kenal sebagai Big Data. Big Data merujuk pada kumpulan data yang sangat besar, beragam, dan terus bertambah dengan cepat, sehingga sulit untuk diolah menggunakan metode tradisional. Big Data pada dasarnya mengacu pada:

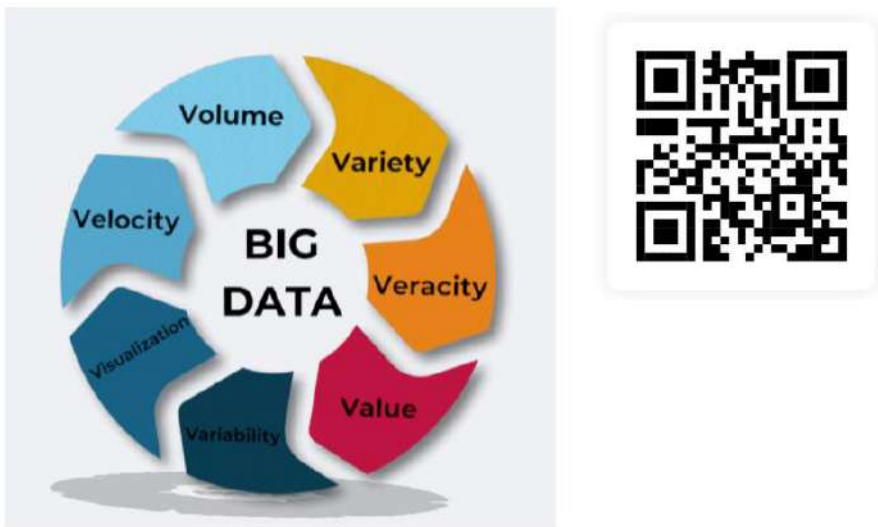
- a. Fakta bahwa kita memiliki lebih banyak data tentang apa pun.
- b. Kemampuan kita yang telah meningkat untuk menyimpan dan menganalisis data apa pun.

Dalam beberapa tahun terakhir, telah terjadi pertumbuhan eksponensial dalam data terstruktur dan tidak terstruktur yang dihasilkan oleh teknologi informasi, industri, perawatan kesehatan, Internet of Things, dan sistem lainnya. Big Data memiliki potensi untuk mendukung aplikasi pintar generasi berikutnya yang akan memanfaatkan kekuatan data untuk membuat aplikasi menjadi cerdas. Data telah menjadi elemen penting dalam kehidupan sehari-hari dan berbagai bidang seperti bisnis, kesehatan, dan pemerintahan, dengan kemampuan untuk meningkatkan pengambilan keputusan dan efisiensi. Di era Revolusi Industri 4.0, data, bersama

dengan teknologi canggih seperti AI dan Big Data, semakin memegang peranan krusial.

2.2 Karakteristik Big Data

Big Data dikenal dengan berbagai karakteristik yang membedakannya dari data tradisional. Untuk memahami sifat dan tantangan yang dihadirkan oleh Big Data, para ahli sering merangkum karakteristik ini ke dalam konsep yang dikenal sebagai 7 V's seperti terlihat pada Gambar 2.1. Pada Gambar 2.1, Anda dapat menscan barcode untuk melihat karakteristik big data dalam bentuk 3D *Augmented Reality* (AR).s



Gambar 2. 1 Karakteristik Big Data

a. *Variety* (Keberagaman)

Variety merujuk pada beragam jenis data yang ada. Big Data datang dari berbagai sumber dan format yang berbeda. Big Data tidak hanya berupa teks, tetapi juga

dapat berupa gambar, video, audio, dan data sensor. Selain itu, data dapat datang dari sistem terstruktur seperti database, maupun yang tidak terstruktur seperti media sosial dan log website. Misalnya:

- Data dari **sistem transaksi** (terstruktur) yang mencatat pembelian pelanggan.
- **Tweet di media sosial** (tidak terstruktur) berisi opini pengguna tentang suatu produk.
- **Gambar dan video** yang diunggah ke platform seperti YouTube atau Instagram (semi-terstruktur).

b. Volume (Volume Data)

Volume mengacu pada jumlah data yang sangat besar yang dihasilkan setiap hari. Di era digital ini, volume data yang dihasilkan berkembang secara eksponensial, sehingga menuntut sistem penyimpanan dan pemrosesan yang efisien. Misalnya:

- Pada tahun 2023, diperkirakan ada 5,4 miliar pengguna internet di seluruh dunia, meningkat dari 5,3 miliar pada tahun sebelumnya, yang mewakili 67% dari populasi global.
- Facebook memiliki lebih dari 3,07 miliar pengguna aktif bulanan, mencakup lebih dari setengah pengguna internet global.
- Google memproses lebih dari 3,5 miliar pencarian per hari.
- Sensor IoT (Internet of Things) menghasilkan data dalam jumlah besar dari perangkat seperti jam tangan pintar atau kendaraan yang terhubung.

c. Velocity (Kecepatan)

Kecepatan merujuk pada seberapa cepat data dihasilkan dan perlu diproses untuk memberikan nilai. Dalam Big Data, data seringkali dihasilkan secara real-time

dan membutuhkan analisis langsung untuk memberikan hasil yang bermanfaat.

Misalnya:

- Transaksi perbankan online harus diproses secara real-time untuk mencegah penipuan.
- Sensor kendaraan otonom yang menghasilkan data dalam milidetik untuk memastikan kendaraan bisa merespon dengan cepat terhadap perubahan lingkungan.
- Feed berita di media sosial seperti Twitter yang terus diperbarui dalam waktu singkat berdasarkan aktivitas pengguna.

d. *Variability* (Variabilitas)

Variabilitas mengacu pada inkonsistensi dalam data. Data tidak selalu datang dengan pola yang sama, dan jenis data yang berbeda mungkin memerlukan metode analisis yang berbeda. Data yang berubah-ubah atau tidak konsisten menambah tantangan dalam pemrosesan Big Data.

Misalnya:

- Sistem cuaca yang berubah setiap jam, menyebabkan data yang dihasilkan selalu berbeda.
- Data opini pengguna di media sosial yang bisa berbeda-beda tergantung situasi atau trend yang sedang berkembang.
- Transaksi musiman seperti penjualan yang melonjak selama liburan atau diskon besar-besaran.

e. *Value* (Nilai)

Value dalam Big Data merujuk pada potensi nilai yang bisa dihasilkan dari data tersebut, misalnya untuk membuat keputusan yang lebih baik atau mengoptimalkan proses bisnis.

Misalnya:

- Data pelanggan yang dianalisis untuk menghasilkan rekomendasi produk yang dipersonalisasi (seperti di Amazon).
- Data penggunaan mesin di pabrik yang dianalisis untuk mengoptimalkan produksi dan meminimalkan downtime.
- Data kesehatan yang diolah untuk memprediksi penyakit dan memberikan diagnosis lebih cepat.

f. Veracity (Ketepatan)

Veracity merujuk pada kualitas atau akurasi data. Tidak semua data yang dihasilkan dapat diandalkan, karena dapat saja terdapat kesalahan, bias, atau ketidakakuratan dalam data tersebut. Penting untuk memilah dan membersihkan data sebelum digunakan dalam analisis.

Misalnya:

- Review produk di platform e-commerce yang dapat dipalsukan atau diberi ulasan negatif yang bias.
- Data sensor yang mungkin tidak akurat jika alat pengukurnya rusak.
- Data media sosial yang penuh dengan rumor/hoaks atau informasi yang tidak diverifikasi.

g. Visualization (Visualisasi)

Visualization adalah cara untuk mempresentasikan data yang kompleks dalam bentuk grafik, diagram, atau peta, sehingga dapat lebih mudah dipahami oleh manusia. Visualisasi yang efektif membantu dalam menganalisis data secara cepat dan membuat keputusan yang lebih baik.

Misalnya:

- Dashboard bisnis yang menampilkan data penjualan dalam bentuk grafik dan indikator KPI.
- Peta interaktif yang menunjukkan persebaran penyakit di wilayah tertentu berdasarkan data kesehatan.

- Heatmap website yang menunjukkan bagian mana dari halaman web yang paling banyak diinteraksi oleh pengguna.

2.3 Sumber Big Data

Big Data berasal dari berbagai sumber yang beragam yang mencakup hampir setiap aspek kehidupan modern kita. Sumber-sumber ini memberikan data yang kaya dan bervariasi yang memungkinkan analisis mendalam dan wawasan yang berharga.

Berikut beberapa sumber Big Data yang paling umum:

a. Data Transaksional

Data ini berasal dari setiap transaksi yang terjadi di berbagai bidang, seperti keuangan, ritel, dan e-commerce. Setiap kali seseorang melakukan pembelian, pembayaran, atau transfer uang, data dihasilkan dan disimpan dalam bentuk yang terstruktur.

Contoh:

- E-commerce: Setiap pembelian yang dilakukan melalui platform seperti Amazon menghasilkan data tentang produk, jumlah, metode pembayaran, dan lokasi pengiriman.
- Perbankan: Transaksi melalui ATM atau online banking menghasilkan data tentang saldo, waktu transaksi, dan jumlah uang yang ditransfer.

b. Media Sosial

Data dari media sosial merupakan salah satu sumber Big Data yang tumbuh paling cepat. Setiap kali pengguna memposting status, membagikan foto, memberi komentar, atau menekan tombol "like," data baru dihasilkan. Data ini tidak terstruktur dan seringkali berupa teks, gambar, atau video.

Contoh:

- Facebook, Twitter, Instagram: Informasi yang dihasilkan berupa teks (posting atau tweet), gambar, video, dan interaksi pengguna (likes, comments, shares).
- Sentimen Analisis: Data dari media sosial digunakan untuk menganalisis pendapat publik tentang suatu produk atau layanan.

c. Sensor dan IoT (*Internet of Things*)

Internet of Things merujuk pada jaringan perangkat fisik yang terhubung ke internet, seperti sensor cuaca, kendaraan pintar, atau peralatan rumah tangga yang terkoneksi. Perangkat ini terus-menerus mengirimkan data dalam jumlah besar.

Contoh:

- Sensor Cuaca: Mengirimkan data tentang suhu, kelembaban, dan kecepatan angin setiap detik dari berbagai lokasi di dunia.
- Mobil Otonom: Menghasilkan data dari sensor seperti GPS, radar, dan kamera yang membantu navigasi dan pengambilan keputusan.

d. Data dari Mesin dan Log Aktivitas

Setiap perangkat atau aplikasi digital, seperti server, komputer, atau mesin industri, menghasilkan log aktivitas. Data ini sangat penting untuk pemantauan performa, keamanan, dan optimasi sistem.

Contoh:

- Situs Web: Server website seperti Google Analytics menyimpan data mengenai pengunjung, termasuk halaman yang diakses, waktu akses, dan lokasi pengunjung.

- Manufaktur: Mesin industri modern menghasilkan data mengenai kecepatan produksi, jumlah output, serta kerusakan atau kegagalan operasi.

e. Data Kesehatan

Sektor kesehatan juga merupakan sumber Big Data yang signifikan. Data ini berasal dari rekam medis elektronik (Electronic Health Records), perangkat medis, dan riset klinis. Data kesehatan seringkali terstruktur namun bisa juga tidak terstruktur, seperti hasil MRI atau X-ray.

Contoh:

- Rumah Sakit: Data pasien yang mencakup riwayat medis, hasil lab, diagnosis, dan perawatan.
- Wearables: Perangkat seperti Fitbit atau Apple Watch yang memantau detak jantung, jumlah langkah, dan pola tidur.

f. Data Publik dan Pemerintah

Data dari lembaga pemerintah sering digunakan untuk berbagai analisis, termasuk kebijakan publik, penelitian sosial, dan perencanaan kota. Data ini biasanya tersedia secara publik melalui portal data terbuka.

Contoh:

- Data Sensus Penduduk: Memberikan informasi tentang demografi, status pekerjaan, tingkat pendidikan, dan distribusi penduduk.
- Transportasi: Data lalu lintas yang dikumpulkan dari sensor jalan raya atau sistem transportasi publik.

g. Data dari Komunikasi Digital

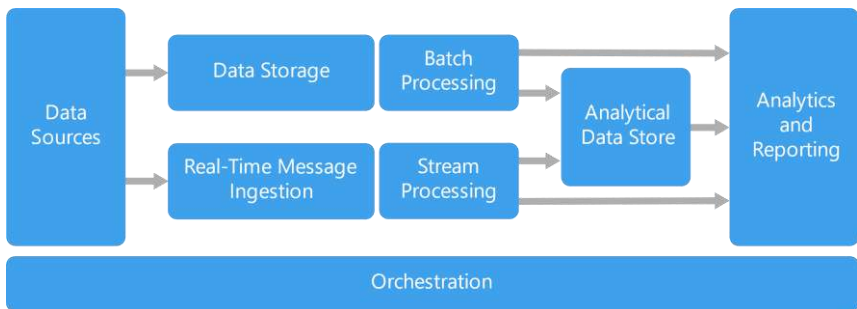
Data yang dihasilkan dari komunikasi seperti email, pesan instan, dan telepon juga merupakan sumber Big Data yang signifikan. Data ini sering digunakan untuk analisis tren, deteksi penipuan, dan keamanan siber.

Contoh:

- Email: Analisis terhadap spam atau pola komunikasi bisnis.
- Telekomunikasi: Data dari panggilan telepon, SMS, atau penggunaan data mobile untuk memahami pola penggunaan dan peningkatan layanan.

2.4 Arsitektur Big Data

Sebagian besar arsitektur Big Data mencakup beberapa atau semua komponen berikut:



Gambar 2. 2 Arsitektur Big Data (Microsoft, 2024)

- Data Sources:** Semua solusi Big Data dimulai dengan satu atau lebih sumber data. Misalnya database, File statis seperti log server, sumber data real-time seperti IoT.
- Data Storage:** Data untuk proses batch processing biasanya disimpan dalam penyimpanan berkas terdistribusi yang dapat menyimpan data dalam jumlah

yang besar dengan berbagai format. Jenis penyimpanan ini disebut dengan data lake.

- c. **Batch Processing:** Dataset yang sangat besar sering kali membuat solusi big data harus memproses berkas data menggunakan *batch processing* dimana data dibagi per batch dan berjalan berulang-ulang secara otomatis sampai seluruh data selesai diproses
- d. **Real-time Message Ingestion:** Jika solusinya mencakup sumber data real-time, arsitektur harus mencakup cara untuk menangkap dan menyimpan pesan real-time untuk pemrosesan aliran.
- e. **Stream Processing:** Setelah menangkap pesan real-time, solusi harus memprosesnya dengan menyaring, menggabungkan, dan mempersiapkan data untuk analisis. Data aliran yang telah diproses kemudian ditulis ke tempat penyimpanan keluaran.
- f. **Analytical Data Store:** Banyak solusi Big Data mempersiapkan data untuk dianalisis dan kemudian menyajikan data yang telah diproses dalam format terstruktur yang dapat ditanya dengan alat analisis.
- g. **Analysis and Reporting:** Tujuan dari sebagian besar solusi Big Data adalah memberikan wawasan ke dalam data melalui analisis dan pelaporan.
- h. **Orchestration:** Sebagian besar solusi Big Data terdiri dari operasi pemrosesan data yang diulang-ulang, yang terkapsulasi dalam alur kerja, yang mengubah data sumber, memindahkan data antara

2.5 Teknologi dan Alat Big Data

Berikut merupakan beberapa teknologi yang sering digunakan dalam Big Data.

a. Hadoop

Hadoop adalah salah satu teknologi Big Data yang paling terkenal dan digunakan secara luas. Hadoop adalah framework open-source yang memungkinkan pemrosesan data dalam skala besar dengan mendistribusikannya ke beberapa komputer atau server dalam sebuah jaringan. Hadoop memiliki beberapa fitur utama seperti:

- Hadoop Distributed File System (HDFS): Sistem penyimpanan yang memungkinkan penyimpanan data dalam jumlah besar di berbagai node (komputer). Data disimpan dalam potongan-potongan kecil yang disebar di seluruh jaringan.
- MapReduce: Model pemrograman yang membagi tugas pemrosesan data menjadi dua tahap: Map, yang membagi tugas menjadi beberapa bagian, dan Reduce, yang menggabungkan hasil dari semua bagian.

Hadoop digunakan oleh perusahaan seperti Facebook dan Yahoo! untuk memproses data pengguna dalam jumlah besar dan menganalisisnya untuk kebutuhan internal, seperti analisis tren pengguna atau preferensi iklan.

b. Spark

Apache Spark adalah alat Big Data yang lebih cepat daripada Hadoop dalam pemrosesan data karena menggunakan pemrosesan in-memory, yang berarti data diproses di memori (RAM) alih-alih ditulis dan dibaca dari disk seperti di Hadoop. Spark memiliki beberapa fitur utama seperti:

- Pemrosesan In-Memory: Spark dapat menyimpan data dalam memori sehingga mempercepat pemrosesan data yang besar.
- Batch dan Streaming Processing: Spark mendukung baik pemrosesan batch (data diproses dalam blok) maupun streaming (data diproses secara real-time).
- Kompatibilitas dengan Hadoop: Spark dapat bekerja di atas Hadoop, menggunakan HDFS untuk penyimpanan.

Uber menggunakan Spark untuk menganalisis data perjalanan pengguna secara real-time, seperti memperkirakan waktu kedatangan dan mengoptimalkan rute.

c. NoSQL Databases

Basis data NoSQL (Not Only SQL) dirancang untuk menangani data yang tidak terstruktur atau semi-terstruktur, seperti data media sosial, dokumen, atau grafik. NoSQL berbeda dari basis data relasional tradisional (seperti MySQL) karena lebih fleksibel dalam menangani berbagai jenis data. Contoh NoSQL Databases:

- MongoDB: Digunakan untuk menyimpan data berbentuk dokumen yang fleksibel dan bisa dengan mudah berkembang seiring bertambahnya jumlah data.
- Cassandra: Basis data terdistribusi yang dirancang untuk menangani volume data yang sangat besar dan tetap responsif meskipun server mengalami kegagalan.

Netflix menggunakan Cassandra untuk menyimpan dan mengelola data dalam jumlah besar, termasuk preferensi penonton dan rekomendasi film.

d. Data Warehousing: Amazon Redshift

Data warehousing adalah proses penyimpanan data dalam jumlah besar yang berasal dari berbagai sumber ke dalam satu tempat terpusat untuk analisis lebih lanjut.

Salah satu alat populer untuk ini adalah Amazon Redshift. Redshift adalah layanan cloud yang memungkinkan pengguna menyimpan dan menganalisis data dalam skala petabyte dengan cepat dan efisien.

Amazon Redshift memiliki beberapa fitur utama seperti:

- Kecepatan dan Skalabilitas: Redshift dapat menangani query (permintaan data) dalam jumlah besar dan sangat cepat berkat arsitektur kolomnya.
- Integrasi dengan Layanan AWS: Redshift terintegrasi dengan layanan Amazon lainnya, seperti S3, untuk penyimpanan data yang mudah.

Airbnb menggunakan Amazon Redshift untuk menyimpan dan menganalisis data pengguna mereka, seperti pencarian properti dan reservasi, untuk meningkatkan pengalaman pelanggan.

e. Alat Pemvisualisasian Data: Tableau

Setelah data dianalisis, langkah penting berikutnya adalah menyajikan hasil analisis dalam bentuk visual yang mudah dipahami. Tableau adalah salah satu alat pemvisualisasian data paling populer yang memungkinkan pengguna membuat grafik, peta, dan dashboard interaktif tanpa memerlukan keahlian pemrograman. Tableau memiliki beberapa fitur utama seperti:

- Drag-and-Drop Interface: Memudahkan pengguna untuk membuat visualisasi dengan cepat hanya dengan menyeret dan melepas data.
- Real-Time Data Analysis: Tableau dapat menghubungkan data secara real-time dan memperbarui visualisasi secara otomatis.
- Kompatibilitas dengan Berbagai Sumber Data: Tableau dapat terhubung ke berbagai sumber data, seperti SQL, cloud, atau spreadsheet.

Tesla menggunakan Tableau untuk memantau performa pabrik mereka, termasuk produktivitas mesin dan pemeliharaan, sehingga mereka dapat membuat keputusan berdasarkan data secara cepat.

2.6 Implementasi Big Data di Berbagai Industri

Big Data telah membawa revolusi dalam berbagai industri yang memungkinkan pengambilan keputusan yang lebih baik, efisiensi operasional, serta inovasi produk dan layanan. Di bawah ini adalah beberapa implementasi big data di industri.

a. Industri Ritel

Dalam industri ritel, Big Data digunakan untuk memahami perilaku konsumen, mengoptimalkan inventaris, dan meningkatkan pengalaman pelanggan. Data dari transaksi penjualan, interaksi online, dan feedback konsumen membantu peritel dalam membuat keputusan yang lebih baik dan lebih cepat. Misalnya Amazon menggunakan Big Data untuk merekomendasikan produk kepada pelanggannya. Algoritma Big Data menganalisis riwayat pencarian, pembelian, dan kebiasaan belanja pelanggan untuk memberikan rekomendasi yang dipersonalisasi. Walmart menggunakan analitik Big Data untuk memprediksi pola permintaan berdasarkan cuaca, tren musiman, dan data penjualan historis, yang membantu mereka menjaga stok produk secara efisien di seluruh toko.

b. Industri Kesehatan

Big Data memainkan peran penting dalam meningkatkan perawatan kesehatan, terutama dengan berkembangnya rekam medis elektronik (Electronic Health Records - EHR), wearable devices, dan data genomik. Data yang dihasilkan dari sumber-sumber ini membantu dalam diagnosis lebih cepat, pencegahan penyakit, dan pengembangan obat. Misalnya IBM Watson Health

menggunakan Big Data dan kecerdasan buatan untuk membantu dokter dalam menganalisis data medis dan memberikan diagnosis atau rekomendasi pengobatan yang lebih tepat berdasarkan bukti klinis dari jutaan rekam medis. Fitbit dan perangkat kesehatan wearable lainnya menghasilkan data real-time seperti detak jantung, aktivitas fisik, dan pola tidur yang dianalisis untuk memberikan wawasan kesehatan bagi pengguna.

c. Industri Keuangan

Industri keuangan menggunakan Big Data untuk mendeteksi penipuan, mengelola risiko, dan mempersonalisasi layanan pelanggan. Dengan menggunakan data transaksi, perilaku pasar, dan informasi kredit, lembaga keuangan dapat membuat keputusan yang lebih cerdas dan lebih aman. Misalnya PayPal menggunakan Big Data untuk mendeteksi aktivitas penipuan dalam real-time dengan menganalisis pola transaksi yang mencurigakan dan segera mengambil tindakan pencegahan.

Bank of America menggunakan analitik Big Data untuk memahami perilaku nasabah, menawarkan produk perbankan yang sesuai, serta meningkatkan layanan melalui analisis sentimen dari interaksi media sosial.

d. Industri Transportasi

Dalam industri transportasi, Big Data digunakan untuk mengoptimalkan rute, memprediksi lalu lintas, dan meningkatkan efisiensi operasional. Dengan data yang dihasilkan dari sensor kendaraan, GPS, dan sistem manajemen transportasi, perusahaan transportasi dapat merencanakan rute yang lebih efisien dan mengurangi biaya operasional. Misalnya Uber, Grab, Gojek menggunakan Big Data untuk memperkirakan permintaan perjalanan di berbagai lokasi, menentukan harga dinamis,

serta mengoptimalkan rute bagi pengemudi sehingga mereka dapat tiba lebih cepat di lokasi penjemputan.

e. Industri Energi

Big Data digunakan di industri energi untuk memantau dan mengelola sumber daya secara efisien, mengurangi biaya, dan meningkatkan keandalan jaringan listrik. Data dari sensor, meter pintar, dan kondisi cuaca dianalisis untuk meningkatkan efisiensi energi dan mengurangi pemborosan. Misalnya General Electric (GE) menggunakan Big Data untuk menganalisis data dari sensor yang dipasang pada turbin angin, mesin jet, dan peralatan energi lainnya. Analisis ini membantu mereka dalam pemeliharaan prediktif dan meningkatkan efisiensi operasional.

Tesla menggunakan Big Data untuk memonitor performa kendaraan listrik mereka. Data dari setiap kendaraan Tesla dikumpulkan dan dianalisis untuk meningkatkan efisiensi baterai, keamanan, dan pengalaman pengguna.

f. Industri Pertanian

Big Data di bidang pertanian membantu petani mengoptimalkan hasil panen dengan cara menganalisis data tentang kondisi cuaca, tanah, penggunaan air, dan faktor lainnya. Ini dikenal dengan istilah precision farming atau pertanian presisi. Misalnya John Deere, perusahaan manufaktur peralatan pertanian, menggunakan sensor pada mesin-mesin mereka untuk mengumpulkan data tentang kelembaban tanah, kualitas biji, dan efisiensi mesin. Data ini kemudian digunakan untuk membantu petani memaksimalkan hasil panen mereka. Climate Corporation menggunakan Big Data untuk membuat prediksi cuaca dan memberikan rekomendasi kepada petani tentang waktu tanam terbaik, penggunaan pupuk, dan manajemen air untuk meningkatkan produktivitas.

g. Industri Hiburan

Big Data digunakan oleh perusahaan di industri hiburan untuk menganalisis preferensi pengguna dan mempersonalisasi konten yang ditawarkan. Data streaming, preferensi tontonan, dan interaksi pengguna dianalisis untuk meningkatkan pengalaman pelanggan. Misalnya Netflix menganalisis data tontonan dari jutaan pengguna untuk memahami preferensi mereka. Dengan Big Data, Netflix dapat merekomendasikan film dan serial yang disesuaikan dengan kebiasaan dan minat pengguna, serta memutuskan konten apa yang perlu diproduksi.

Spotify menggunakan Big Data untuk menganalisis kebiasaan mendengarkan musik pengguna dan memberikan rekomendasi lagu yang sesuai dengan selera musik individu.

2.7 Rangkuman

Big Data telah menjadi aset yang sangat penting bagi individu, perusahaan, dan organisasi di seluruh dunia. Dengan perkembangan pesat dalam teknologi informasi, Internet of Things (IoT), dan berbagai sistem lainnya, data yang dihasilkan setiap hari semakin besar, beragam, dan cepat. Big Data, yang dicirikan oleh karakteristik 7 V's *Variety*, *Volume*, *Velocity*, *Variability*, *Value*, *Veracity*, dan *Visualization* memiliki potensi luar biasa dalam meningkatkan pengambilan keputusan, efisiensi, dan inovasi di berbagai sektor industri.

Sumber Big Data meliputi data transaksi, media sosial, sensor IoT, data mesin, data kesehatan, data publik, dan komunikasi digital, masing-masing memberikan wawasan yang berharga untuk analisis dan pengambilan keputusan. Arsitektur Big Data melibatkan beberapa komponen kunci seperti sumber data, penyimpanan, pemrosesan batch dan real-time, serta analisis dan pelaporan.

Teknologi dan alat seperti Hadoop, Spark, NoSQL databases, Amazon Redshift, dan Tableau memainkan peran penting dalam mengelola dan menganalisis data dalam skala besar. Implementasi Big Data di berbagai industri seperti ritel, kesehatan, keuangan, transportasi, energi, pertanian, dan hiburan telah menunjukkan bagaimana analisis data dapat memengaruhi keputusan strategis, meningkatkan efisiensi operasional, dan mengembangkan produk serta layanan yang lebih baik. Dengan memanfaatkan Big Data secara efektif, berbagai sektor industri dapat meraih keunggulan kompetitif dan beradaptasi dengan perubahan yang cepat dalam lingkungan bisnis mereka.

DAFTAR PUSTAKA

- Santoso J.T. (2020). Analisis Big Data. Yayasan Prima Agus Teknik. ISBN: 9786239 504243.
- Matellio. How companies are using big data and analytics. Available from: <https://www.matellio.com/blog/how-companies-are-using-big-data-and-analytics/>
- SlideTeam. Big data sources & technologies PPT PowerPoint presentation clipart. Available from: <https://www.slideteam.net/big-data-sources-technologies-ppt-powerpoint-presentation-clipart.html>
- Feldaprodata. Available from: www.feldaprodata.com.my
- Microsoft. Big data architecture guide. Available from: <https://learn.microsoft.com/id-id/azure/architecture/guide/architecture-styles/big-data>

BAB 3

STATISTIKA

3.1 Pendahuluan

Bayangkan kamu sedang mengembangkan aplikasi baru yang sudah mencapai ribuan pengguna (#). Setiap klik, setiap interaksi, dan durasi penggunaan aplikasi menjadi data. Pengguna aplikasi meninggalkan banyak jejak data yang dapat dianalisis. Mulai dari waktu yang mereka habiskan di halaman atau fitur tertentu, hingga frekuensi dan pola navigasi penggunaan. Data yang ditinggalkan juga dapat mencakup kapan dan darimana mereka menggunakan aplikasi, perangkat apa yang mereka gunakan, serta bagaimana mereka berinteraksi terhadap konten di aplikasi seperti artikel, gambar atau video serta pengguna lainnya. Lebih lanjut, terdapat informasi penting lainnya yang mungkin dapat kamu peroleh seperti rasio konversi, feedback langsung, istilah atau kata kunci pencarian dan lainnya. Semua data tersebut dapat bermanfaat untuk memahami kebiasaan dan preferensi pengguna yang selanjutnya menjadi pertimbangan dalam pengoptimalan aplikasi yang sedang kamu kembangkan. Tentu data-data tersebut dapat bermanfaat bila makna didalamnya dapat dipahami, bila tidak semua itu hanya sekedar data (tumpukan data). Memahami makna yang ada pada data-data yang banyak akan sulit jika hanya dilihat begitu saja. Dalam membermaknakan data-data tersebutlah Statistika berperan.

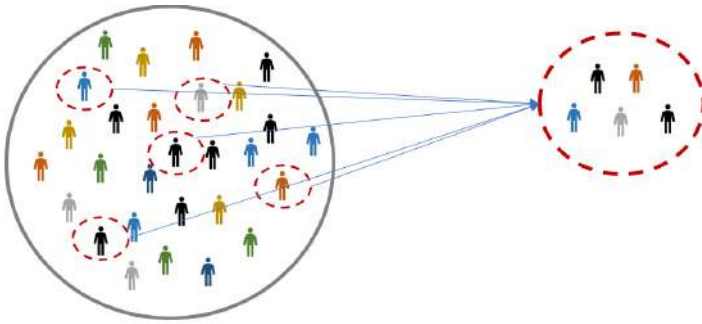
Statistika merupakan ilmu yang mempelajari tentang cara pengumpulan, analisis, penyajian, dan menarik kesimpulan dari data. Statistika bukan hanya tentang angka

dan perhitungannya, tetapi juga tentang bagaimana memahami makna (cerita) yang ada dibalik data. Dalam konteks kamu sebagai pengembang aplikasi tadi, kemampuan memahami data yang ditinggalkan pengguna aplikasi dengan ilmu Statistika yang mumpuni, sangat mungkin menentukan jalan cerita aplikasi yang kamu kembangkan.

3.2 Populasi dan Sampel

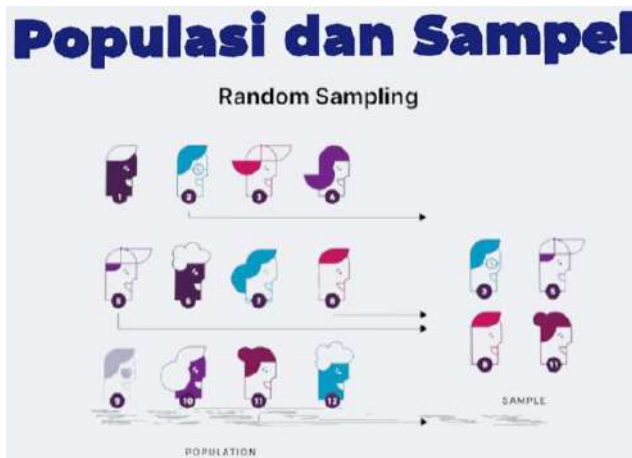
Kembali ke #. Aplikasi yang kamu kembangkan, mengalami pertumbuhan pengguna yang pesat. Misalkan sudah mencapai ratusan ribu pengguna. Pengguna aplikasi tersebut berasal dari berbagai penjuru dunia. Semua orang yang menggunakan aplikasi ini adalah apa yang kita sebut dengan Populasi (Populasi Pengguna). Jadi dalam konteks ini, populasi adalah semua orang yang menggunakan aplikasi, baik yang aktif setiap hari maupun yang hanya sekali-sekali.

Misalkan dalam rangka pengembangan berkelanjutan, kamu ingin mengetahui apa kelebihan dan kekurangan aplikasi yang kamu kembangkan menurut para pengguna. Oleh karena itu, kamu berencana melakukan wawancara singkat dengan para pengguna aplikasi tersebut. Namun, masalahnya adalah kamu tidak memiliki waktu atau sumber daya untuk melakukan wawancara singkat dengan semua orang pengguna aplikasi (populasi). Maka solusinya, kamu memutuskan untuk melakukan wawancara singkat dengan sebagian dari populasi, yang kemudian kita sebut dengan **sampel**. Jadi sampel adalah bagian dari populasi yang dipilih dengan metode tertentu sehingga tetap dapat mewakili (menggambarkan) populasi secara keseluruhan. Ilustrasi populasi dan sampel dapat diperhatikan pada Gambar 3.1 berikut.



Gambar 3. 1 Ilustrasi Populasi dan Sampel

Gambar 3.2, merupakan contoh random sampling populasi dan sample dan Anda dapat scan barrcode yang diberikan untuk melihat dalam bentuk *Augmented Reality* (AR).



Gambar 3. 2 Populasi dan Sampel

Jadi, mengapa kita membutuhkan sampel dari suatu populasi? Sampel kita butuhkan bila melakukan kajian/penelitian terhadap suatu populasi dan populasi tersebut relatif besar. Ada beberapa hal yang membuat kita

perlu menarik sampel dari suatu populasi yang relatif besar, yaitu:

- a. **Untuk efisiensi waktu:** bilamana penelitian terhadap populasi membutuhkan waktu yang lama. Bayangkan kamu ingin mengetahui pengalaman pengguna aplikasi dari seluruh dunia. Mengumpulkan data dari setiap pengguna akan memakan waktu lama. Jadi, kamu dapat menarik sampel pengguna dari berbagai negara.
- b. **Untuk efisiensi tenaga:** bilamana penelitian terhadap populasi membutuhkan tenaga yang banyak. Misalkan kamu ingin menguji performa fitur baru pada aplikasi. Menguji performa fitur pada setiap pengguna akan membutuhkan banyak tenaga. Jadi kamu dapat memilih sampel pengguna yang aktif menggunakan fitur tersebut.
- c. **Untuk efisiensi biaya:** bila mana penelitian terhadap populasi membutuhkan biaya yang besar. Misalkan kamu ingin menguji performa aplikasi diberbagai jenis/type perangkat smartphone. Membeli semua jenis/type perangkat tersebut akan membutuhkan biaya besar. Jadi kamu dapat memilih sampel dari berbagai perangkat tersebut.

Melakukan penelitian terhadap suatu populasi tidak harus menarik sampel atau sebagian dari populasi tersebut. Penelitian terhadap populasi dapat dilakukan terhadap keseluruhan anggota dari populasi tersebut. Dalam hal ini populasi menjadi (sama dengan) sampel, yang kemudian disebut dengan sampel jenuh. Jadi, pada intinya penarikan sampel dari suatu populasi dilakukan untuk menjamin fisibilitas (keterlaksanaan) dari penelitian tersebut.

3.3 Teknik Sampling

Selain fisibilitas dan efisiensi penelitian, hal yang sangat penting dari penarikan sampel adalah **teknik sampling** yang digunakan. Dalam penarikan sampel

diperlukan teknik sampling yang tepat untuk memastikan bahwa sampel yang diperoleh dapat mewakili populasi (representatif). Secara garis besar, terdapat dua teknik sampling, yaitu sampling probabilitas dan sampling non-probabilitas. **Sampling probabilitas** merupakan teknik sampling di mana setiap anggota populasi memiliki peluang yang sama untuk dipilih menjadi sampel. Sedangkan **sampling non-probabilitas** merupakan teknik sampling di mana tidak semua anggota populasi memiliki peluang yang sama untuk dipilih menjadi sampel. Kedua teknik sampling tersebut memiliki beberapa jenis seperti berikut.

a. Jenis-jenis Sampling Probabilitas

Berikut merupakan jenis-jenis sampling probabilitas.

- **Acak Sederhana (Simple Random Sampling):** memilih anggota populasi secara acak dan sederhana menjadi sampel. Misalkan aplikasi yang kamu kembangkan memiliki 70.923 pengguna yang memiliki No. ID sesuai dengan nomor pendaftaran. Dengan random sampling, kamu menggunakan generator angka acak untuk memilih 100 nomor ID dari 70.923 nomor ID pengguna tersebut.
- **Sistemik (Systemic Sampling):** memilih setiap anggota ke- n dari populasi menjadi sampel. Misalkan kamu punya daftar pengguna aplikasi yang diurutkan berdasarkan waktu pendaftaran. Kamu memilih setiap pengguna ke-10 atau ke-100 untuk menjadi sampel.
- **Berstrata (Stratified Sampling):** membagi populasi menjadi beberapa kelompok berdasarkan strata, kemudian sampel diambil dari setiap strata. Misalkan kamu punya daftar pengguna aplikasi yang dibagi ke dalam beberapa strata berdasarkan kelompok usia. Kemudian, kamu memilih sampel secara acak dari kelompok usia tersebut.

- **Berkelompok (Cluster Sampling):** membagi populasi menjadi beberapa kelompok, kemudian memilih beberapa kelompok secara acak. Misalkan kamu mengelompokkan pengguna berdasarkan kota tempat tinggal. Kemudian, kamu memilih beberapa kota secara acak dan menganalisa data pengguna di kota-kota tersebut.
- **Multistage Sampling:** menggabungkan beberapa teknik probabilitas sampling. Misalkan kamu memilih beberapa kota tempat pengguna secara acak (cluster sampling), kemudian kamu memilih pengguna secara acak dari kota-kota tersebut (simple random sampling).

b. Jenis-Jenis Sampling Non-probabilitas

Berikut merupakan jenis-jenis sampling non-probabilitas.

- **Snowball Sampling:** memilih sampel berdasarkan rekomendasi sampel sebelumnya. Misalkan kamu memulai dengan satu pengguna aplikasi yang cocok dengan kriteria kemudian melanjutkan dengan meminta pengguna tersebut merekomendasikan pengguna lainnya yang cocok dengan kriteria.
- **Quota Sampling:** memilih sampel yang memenuhi kriteria tertentu sampai kuota terpenuhi. Misalkan kamu menetapkan kuota untuk setiap kelompok, misalnya 20 pengguna dari setiap kelompok rentang usia.
- **Sequential sampling:** memilih sampel secara berurutan hingga memperoleh informasi yang cukup. Misalnya kamu mengumpulkan testimoni pengguna terhadap suatu fitur baru aplikasi dengan satu per satu pengguna hingga mendapatkan feedback yang cukup untuk membuat keputusan.

- **Purposive sampling:** memilih sampel berdasarkan tujuan tertentu. Misalnya kamu memilih pengguna yang sangat aktif untuk mendapatkan feedback atau testimoni terhadap fitur baru dari aplikasi yang kamu kembangkan.

Bagaimana kita memilih teknik sampling dalam melakukan kajian atau penelitian terhadap suatu populasi? Pemilihan antara probabilitas sampling dan non-probabilitas sampling tergantung pada tujuan penelitian, sumber daya yang tersedia, dan kebutuhan spesifik dari suatu kajian atau penelitian. Probabilitas sampling sebaiknya dipilih ketika: penelitian yang dilakukan memerlukan representasi yang sangat akurat, hasil penelitian perlu di generalisasi ke seluruh populasi, diperlukan hasil yang akurat dan dapat diandalkan, atau terdapat sumber daya (waktu, dana, dan tenaga) yang memadai. Probabilitas non-sampling sebaiknya dipilih ketika: penelitian yang dilakukan memiliki keterbatasan sumber daya, memiliki tujuan khusus yang tidak memerlukan representasi populasi secara keseluruhan, membutuhkan hasil yang cepat, atau keseluruhan anggota populasi target sulit dijangkau.

3.4 Jenis-Jenis Data

Data merupakan informasi yang dikumpulkan dan dikelola untuk dianalisis dan digunakan dalam pengambilan keputusan. Data terdiri dari berbagai jenis, bisa berupa angka, teks, gambar atau bentuk lainnya. Jenis-jenis data dapat dibedakan berdasarkan beberapa kategori, diantaranya: berdasarkan sifatnya, cara memperolehnya, waktu pengumpulannya, dan skala pengukurannya.

a. Jenis data berdasarkan sifatnya.

- **Data kualitatif:** Data yang dinyatakan dalam bentuk non-angka atau non-numerik atau biasa juga disebut atribut. Misalnya feedback pengguna tentang

pengalaman mereka menggunakan aplikasi, seperti “antarmuka yang intuitif” atau “fitur yang sangat bermanfaat”.

- **Data kuantitatif:** data yang dinyatakan dalam bentuk angka (data numerik). Data kuantitatif ini terdiri atas dua jenis yaitu data diskrit dan data kontinu. Data Diskrit merupakan data hasil pencacahan dan berupa bilangan bulat (dalam komputer dikenal sebagai integer). Misalnya, jumlah pengguna aplikasi dari setiap kota. Sedangkan data kontinu merupakan data hasil proses pengukuran dan dapat berupa bilangan pecahan (bilangan real). Misalnya, data durasi penggunaan dalam satu hari dari para pengguna aplikasi.

b. Jenis data berdasarkan cara memperolehnya.

- **Data primer:** data yang dikumpulkan langsung dari sumber aslinya. Contohnya, data yang diperoleh melalui survei atau wawancara langsung dari pengguna aplikasi untuk mendapat feedback.
- **Data sekunder:** data yang diperoleh dari sumber yang sudah ada. Contohnya, laporan analitik dari pihak ketiga tentang tren penggunaan aplikasi yang dikembangkan.

c. Jenis data berdasarkan waktu pengumpulannya.

- **Data Cross-Sectional:** data yang dikumpulkan pada satu titik waktu tertentu. Contohnya, data survei kepuasan pengguna aplikasi yang dilakukan setelah peluncuran fitur baru.
- **Data time-series/longitudinal:** data yang dikumpulkan dari waktu ke waktu. Contohnya, data penggunaan aplikasi selama beberapa bulan untuk melihat tren penggunaan fitur tertentu.

- **Data panel:** data yang dikumpulkan dengan kombinasi data cross-sectional dan data longitudinal.

d. Jenis data berdasarkan skala pengukurannya.

- **Data nominal:** data berupa atribut yang dikategorikan tanpa urutan atau peringkat. Contohnya, jenis kelamin (laki-laki atau perempuan) dan kota asal pengguna aplikasi.
- **Data ordinal:** data berupa atribut/kategori yang memiliki urutan atau peringkat tetapi jarak antar kategori tidak sama. Contohnya, tingkat pendidikan (SD, SMP, SMA, Sarjana) dan tingkat kepuasan (puas, netral, tidak puas) pengguna aplikasi.
- **Data interval:** data yang memiliki urutan dan jarak yang sama antar nilai, tetapi tidak memiliki nol mutlak. Contohnya, skor (skala 1-10) yang diberikan pengguna aplikasi untuk menilai fitur baru dari suatu aplikasi.
- **Data rasio:** data yang memiliki urutan, jarak yang sama antar nilai dan nol mutlak. Contohnya, durasi waktu penggunaan fitur tertentu dari aplikasi atau jumlah transaksi yang dilakukan pengguna aplikasi.

3.5 Statistika Deskriptif

Secara umum, statistika terbagi atas statistika deskriptif dan statistika inferensial. Perbedaan utama antara keduanya terletak pada fokus tujuan analisis data masing-masing. Analisis data pada statistika deskriptif bertujuan untuk meringkas dan menyajikan data sehingga lebih mudah dipahami. Sedangkan analisis data pada statistika inferensial bertujuan untuk pengambilan keputusan atau kesimpulan dari sampel yang dianalisis, yang kemudian diterapkan secara umum pada populasi.

Statistika deskriptif merupakan cabang statistika yang berfokus pada bagaimana meringkas dan menyajikan data sehingga lebih mudah dipahami. Sekelompok data perlu diringkas dan disajikan dalam bentuk yang lebih mudah

dimengerti, sehingga memudahkan pembaca untuk memaknai data tersebut. Sekelompok data (kuantitatif) dapat diringkas dalam bentuk ringkasan numerik. Ringkasan numerik ini terdiri atas ukuran pemusatan data dan ukuran penyebaran data. Bentuk atau metode penyajian data cukup beragam. Bentuk penyajian data yang paling sederhana adalah dengan menyajikan data dalam keadaan terurut dari yang terkecil hingga terbesar atau sebaliknya. Misalnya dengan mengurutkan data usia pengguna aplikasi, kita dapat memahami usia termuda atau usia tertua dari pengguna aplikasi tersebut. Adapun bentuk lain dari penyajian data adalah dalam bentuk tabel, grafik, dan/atau infografis. Penyajian atau visualisasi data tidak kita bahas secara detail pada bab ini karena akan kita bahas secara khusus pada bab berikutnya.

a. Ukuran Pemusatan Data

Ukuran Pemusatan Data merupakan ringkasan numerik yang menggambarkan letak (nilai) dimana data berpusat. Data sering menunjukkan kecenderungan terpusat di sekitar suatu nilai (nilai pusat). Terdapat tiga ukuran pemusatan data yang paling umum, yaitu: mean, median, dan modus. Ketiga ukuran ini akan kita bahas untuk data tunggal (unrouted data) dan data berkelompok (grouped data).

Mean

Mean merupakan nilai rata-rata dari semua nilai pada kumpulan data. Dengan kata lain, mean (rata-rata) merupakan jumlah semua data dibagi banyaknya data. Mean biasa dilambangkan dengan $\bar{x}$ (baca x bar). Secara matematis, mean dirumuskan sebagai berikut.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (\text{ungrouped data})$$

$$\bar{x} = \frac{\sum_{i=1}^k f_i x_{m,i}}{\sum_{i=1}^k f_i} = \frac{\sum_{i=1}^k f_i x_{m,i}}{n} \quad (\text{grouped data})$$

di mana:

$\bar{x}$ = mean atau nilai rata-rata

x_i = nilai data ke- i

n = banyaknya data

$x_{m,i}$ = nilai tengah dari interval kelas ke- i

k = jumlah interval kelas

f_i = frekuensi

Median

Median merupakan nilai tengah dari kumpulan data yang telah diurutkan, dari data terkecil hingga terbesar. Jika kumpulan data berjumlah genap sehingga terdapat 2 nilai tengah (ungrouped data), maka median adalah rata-rata dari kedua nilai tersebut. Untuk data berkelompok, median dapat dihitung dengan rumus berikut.

$$M_e = T_b + \frac{\frac{n}{2} - F}{f} \cdot p$$

di mana:

M_e = median atau nilai tengah

T_b = tepi bawah kelas median

n = banyak data

F = frekuensi kumulatif sebelum kelas median

f = frekuensi kelas median

p = panjang interval

Modus

Modus merupakan nilai yang paling sering muncul dalam kumpulan data. Suatu kumpulan data mungkin memiliki satu modus (unimodal), dua modus (bimodal), atau bahkan lebih (multimodal). Untuk data tak terkelompok (ungrouped data), modus merupakan data yang paling sering muncul atau

memiliki frekuensi paling besar. Untuk data berkelompok, modus dapat dihitung dengan rumus berikut.

$$M_o = T_b + \frac{d_1}{d_1 + d_2} \cdot p$$

di mana:

M_o = modus

T_b = tepi bawah kelas modus

d_1 = selisih frekuensi kelas modus dengan kelas sebelumnya

d_2 = selisih frekuensi kelas modus dengan kelas sesudahnya

p = panjang kelas/interval

b. Ukuran Penyebaran Data

Ukuran dispersi atau penyebaran data merupakan ringkasan numerik yang menggambarkan seberapa “beragam” atau “bervariasi” data-data dalam suatu kumpulan. Ukuran ini juga menggambarkan seberapa tersebar data dari ukuran pemusatan (mean). Terdapat tiga ukuran penyebaran data yang paling umum, yaitu: jangkauan, simpangan baku, dan varian. Ketiga ukuran ini akan kita bahas untuk data tidak terkelompok (ungrouped data) dan data berkelompok (grouped data).

Jangkauan

Jangkauan (range) menggambarkan perbedaan (jarak) antara nilai terbesar dengan nilai terkecil dalam suatu kumpulan data. Secara matematis, range dirumuskan dengan:

$$R = x_{max} - x_{min}$$

di mana:

R = jangkauan (range)

x_{max} = nilai data terbesar

x_{min} = nilai data terkecil

Simpangan Baku

Simpangan baku atau deviasi standar (standard deviation) menggambarkan jarak antara data-data dengan nilai rata-ratanya. Semakin besar nilai deviasi standar, semakin jauh data-data tersebut menyebar dari nilai rata-rata. Mayoritas nilai data cenderung dalam satu deviasi standar dari mean, dan hanya sebagian kecil saja yang terletak di luar dari tiga deviasi standar meannya. Secara matematis, deviasi standar dirumuskan dengan:

$$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}} \quad \text{(ungrouped data)}$$

$$S = \sqrt{\frac{\sum_{i=1}^k f_i (x_{m,i} - \bar{x})^2}{n - 1}} \quad \text{(grouped data)}$$

di mana:

s = simpangan baku

$\bar{x}$ = mean atau nilai rata-rata

x_i = nilai data ke- i

n = banyaknya data

$x_{m,i}$ = nilai tengah dari interval kelas ke- i

k = jumlah interval kelas

f_i = frekuensi

Varian

Varian merupakan kuadrat dari simpangan baku. Oleh karena itu, varian ini juga menggambarkan seberapa jauh data-data menyebar dari nilai rata-ratanya. Secara matematis, varians dirumuskan dengan:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (\text{ungrouped data})$$

$$s^2 = \frac{\sum_{i=1}^k f_i (x_{m,i} - \bar{x})^2}{n - 1} \quad (\text{grouped data})$$

di mana:

S^2 = simpangan baku

$\bar{x}$ = mean atau nilai rata-rata

x_i = nilai data ke-i

n = banyaknya data

$x_{m,i}$ = nilai tengah dari interval kelas ke-i

k = jumlah interval kelas

f_i = frekuensi

3.6 Rangkuman

Aplikasi yang memiliki ribuan hingga jutaan pengguna menghasilkan sejumlah besar data dari setiap interaksi pengguna. Data ini mencakup pola navigasi, perangkat yang digunakan, hingga konversi dan feedback yang dapat diolah untuk optimalisasi aplikasi. Namun, tanpa analisis yang tepat, data hanya akan menjadi tumpukan informasi yang tidak berarti. Di sinilah peran ilmu statistika menjadi penting. Statistika memfasilitasi pengumpulan, analisis, dan penyajian data untuk mengungkap cerita dan tren yang tersembunyi di balik angka-angka, membantu pengembang aplikasi dalam membuat keputusan berbasis data.

Dalam penelitian yang melibatkan populasi besar, seperti pengguna aplikasi global, sering kali diperlukan sampel sebagai representasi dari populasi tersebut. Sampel dipilih untuk alasan efisiensi waktu, tenaga, dan biaya. Ada

beberapa teknik penarikan sampel yang dapat digunakan, baik secara probabilitas maupun non-probabilitas, tergantung pada tujuan penelitian dan sumber daya yang tersedia. Teknik probabilitas menjamin representasi yang lebih akurat, sementara teknik non-probabilitas lebih cocok untuk penelitian dengan sumber daya terbatas.

Selain itu, pemilihan jenis data juga krusial dalam analisis statistika. Data dapat dikategorikan berdasarkan sifatnya, cara pengumpulannya, waktu pengumpulan, serta skala pengukuran. Pemahaman terhadap berbagai jenis data seperti kualitatif, kuantitatif, data primer, sekunder, dan skala pengukuran, membantu dalam menentukan metode analisis yang tepat. Statistika deskriptif, sebagai cabang dari statistika, memberikan kerangka untuk merangkum dan menyajikan data melalui ukuran pemusatan seperti mean, median, dan modus, serta ukuran penyebaran seperti jangkauan, simpangan baku, dan varian, sehingga tren data menjadi lebih jelas dan dapat diinterpretasikan dengan mudah. Melalui penggunaan statistika, pengembang aplikasi dapat memahami pola perilaku pengguna dengan lebih baik, memungkinkan keputusan yang lebih informatif terkait pengembangan dan peningkatan fitur.

DAFTAR PUSTAKA

- Harinaldi, D. I., & Eng, M. (2005). Prinsip-prinsip statistik untuk teknik dan sains. Jakarta: Erlangga.
- Mendenhall, W., Beaver, R., & Beaver, B. (2019). Introduction to Probability and Statistics. Boston: Cengage Learning.
- Bluman, A. G. (2023). Elementary statistics: A step by step approach: Eleventh Edition. New York, NY, USA: McGraw-Hill.

BAB 4

ANALITIKA DATA

4.1 Pendahuluan

Analitika data adalah proses komprehensif yang melibatkan pengumpulan, pengolahan, dan interpretasi data mentah untuk menemukan pola, hubungan, dan wawasan yang berguna dalam mendukung pengambilan keputusan yang lebih baik. Saat ini, dengan perkembangan teknologi dan semakin besarnya volume data yang dihasilkan setiap hari, analitika data menjadi bagian penting dari berbagai industri, seperti bisnis, kesehatan, pemerintahan, dan pendidikan. Analitika data membantu mengubah data yang kompleks dan beragam dapat diubah menjadi informasi yang lebih jelas dan berguna untuk meningkatkan efisiensi, mengoptimalkan strategi, dan membuat keputusan yang lebih baik.

Mengapa analitika data menjadi begitu penting di era modern ini? Salah satu alasan utamanya adalah ledakan volume data yang belum pernah terjadi sebelumnya, yang dikenal dengan istilah **big data**. Saat ini, perusahaan tidak hanya mengumpulkan data dari sumber tradisional seperti penjualan dan survei pelanggan, tetapi juga dari berbagai sumber digital seperti media sosial, sensor Internet of Things (IoT), dan aplikasi berbasis web. Dengan data yang begitu beragam, organisasi memiliki peluang besar untuk mendapatkan wawasan yang lebih mendalam tentang perilaku konsumen, tren pasar, dan bahkan operasional internal.

Sebagai contoh, di dunia bisnis, perusahaan e-commerce menggunakan analitika data untuk memahami pola belanja pelanggan mereka. Melalui analisis data transaksi, data perilaku dari situs web, dan umpan balik pelanggan, mereka dapat menciptakan pengalaman yang lebih personal dan relevan bagi konsumen. Amazon, misalnya, memanfaatkan analisis prediktif untuk merekomendasikan produk berdasarkan riwayat pencarian dan pembelian pengguna, yang secara signifikan meningkatkan penjualan melalui personalisasi tersebut.

Di sektor kesehatan, analitika data sangat membantu dalam mendiagnosis penyakit lebih awal dan meningkatkan kualitas perawatan. Dengan memanfaatkan data medis pasien dan menerapkan teknik machine learning, rumah sakit dapat memprediksi kemungkinan kambuhnya suatu penyakit atau mengidentifikasi risiko penyakit kronis sebelum gejalanya muncul. Misalnya, analitika data digunakan untuk mengembangkan algoritma yang mampu memprediksi kemungkinan serangan jantung atau diabetes berdasarkan data biometrik pasien, sehingga intervensi medis dapat dilakukan lebih awal.

Tidak hanya di bisnis dan kesehatan, di pemerintahan pun analitika data digunakan untuk pengambilan keputusan strategis. Pemerintah kota dapat menggunakan data lalu lintas dan sensor IoT untuk merancang kota pintar yang lebih efisien. Dengan memantau data real-time dari kamera lalu lintas, aplikasi navigasi, dan sensor udara, kota dapat meningkatkan manajemen lalu lintas, mengurangi kemacetan, dan meningkatkan kualitas udara. Misalnya, Singapura telah memanfaatkan analitika data untuk mengoptimalkan transportasi umum dan mengurangi waktu perjalanan warganya.

Analitika data juga memainkan peran penting dalam industri keuangan, di mana risiko dan peluang harus dikelola dengan cermat. Perusahaan asuransi, misalnya, menggunakan

analitika data untuk menilai risiko klaim pelanggan dan menetapkan premi asuransi yang sesuai. Bank juga memanfaatkan analitik untuk mendeteksi fraud atau penipuan, dengan menganalisis pola transaksi yang mencurigakan dan mengidentifikasi anomali yang mungkin menunjukkan aktivitas ilegal.

Pentingnya analitika data juga tercermin dalam pengambilan keputusan berbasis data (*data-driven decision making*). Dalam pendekatan ini, keputusan diambil berdasarkan bukti yang didapat dari analisis data, bukan intuisi atau pengalaman semata. Pendekatan ini memungkinkan organisasi membuat keputusan yang lebih akurat, mengurangi ketidakpastian, dan meminimalkan risiko. Google, misalnya, menggunakan analitika data untuk mengoptimalkan strategi pemasaran mereka, dengan memanfaatkan data pengguna untuk menyesuaikan kampanye iklan mereka sesuai dengan preferensi dan perilaku target audiens.

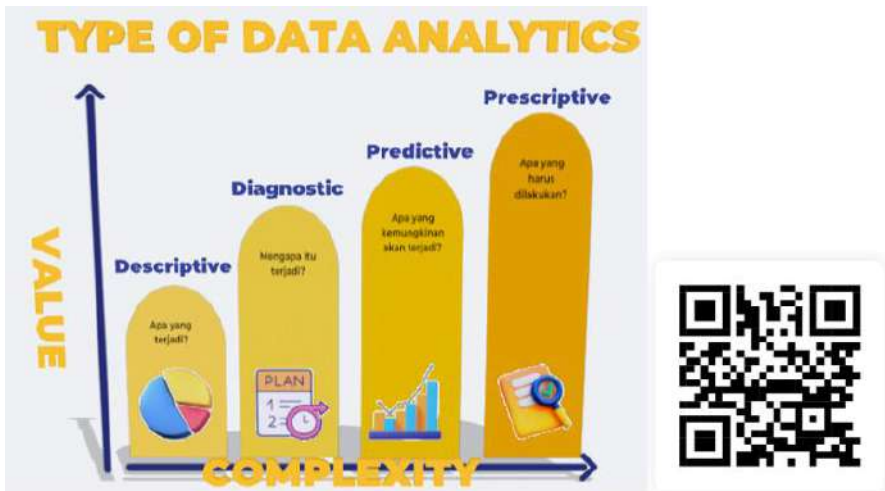
Seiring dengan berkembangnya kecerdasan buatan (AI) dan machine learning, peran analitika data semakin kuat dalam mendukung inovasi teknologi. AI yang dilatih dengan data yang relevan mampu mengotomatiskan proses, mempercepat analisis, dan memberikan rekomendasi yang lebih cerdas. Dalam industri manufaktur, analitika data digunakan untuk memprediksi kegagalan mesin sebelum terjadi, sehingga biaya perbaikan yang besar dapat dihindari dan waktu produksi dapat lebih dioptimalkan.

Secara keseluruhan, analitika data memberikan keunggulan kompetitif bagi organisasi yang mampu memanfaatkannya dengan baik. Di tengah perkembangan teknologi yang pesat, kemampuan untuk mengumpulkan, menganalisis, dan menggunakan data secara efektif menjadi keterampilan yang sangat berharga. Dari e-commerce hingga kesehatan, dari pemerintahan hingga manufaktur, analitika data telah terbukti memberikan dampak signifikan terhadap

berbagai sektor industri, dan pentingnya hanya akan terus meningkat di masa depan.

4.2 Jenis-Jenis Analitika Data

Ada empat jenis utama analitika data yang digunakan untuk memahami dan mengolah data seperti terlihat pada gambar 4.1. Masing-masing jenis analitika data memiliki fungsi dan tujuan yang berbeda dalam membantu kita memahami data dan membuat keputusan yang lebih baik. Anda dapat scan barcode yang diberikan untuk melihat dalam tampilan *Augmented Reality* (AR).



Gambar 4. 1 Jenis-Jenis Analitika Data

a. Analisis Deskriptif

Analisis deskriptif menjadi pondasi atau dasar untuk melakukan analisis diagnostik, analisis prediktif dan analisis preskriptif. Analisis deskriptif berfokus pada **apa yang telah terjadi** seperti terlihat pada Gambar 4.2.



Gambar 4. 2 Analisis Deskriptif

Tujuan dari analisis ini adalah memberikan gambaran umum dari data historis, seperti kinerja perusahaan atau tren dalam periode waktu tertentu. Misalnya, ketika sebuah perusahaan e-commerce melihat laporan penjualan bulanan, mereka menggunakan analisis deskriptif untuk mengetahui berapa banyak produk yang terjual, kategori produk mana yang paling laku, dan berapa total pendapatan bulan itu. Analisis deskriptif sering digunakan sebagai langkah awal untuk memahami pola-pola dasar dalam data sebelum melangkah ke analisis yang lebih mendalam.

Analisis deskriptif digunakan untuk memahami kinerja keseluruhan pada tingkat **agregat** dan jauh lebih mudah bagi perusahaan untuk memulai karena data cenderung mudah didapatkan untuk membangun laporan dan aplikasi. Sebagian besar analisis yang dilakukan saat ini adalah analisis berbasis deskriptif melalui penggunaan fungsi statistik seperti Sum, Mean, Median, Variance, Counts, Top-N dan Distinct. Visualisasi yang digunakan

untuk Analisis Deskriptif berupa diagram lingkaran, diagram batang, tabel, atau grafik garis.

b. Analisis Diagnostik

Jika analisis deskriptif menjelaskan apa yang terjadi, analisis diagnostik menjawab pertanyaan **mengapa hal tersebut terjadi** seperti terlihat pada Gambar 4.3.



Gambar 4. 3 Analisis Diagnostik

Analisis ini bertujuan untuk mencari penyebab dari suatu fenomena atau perubahan yang terlihat dalam data. Misalnya, jika penjualan suatu produk menurun tajam pada bulan tertentu, analisis diagnostik akan mencari faktor-faktor yang mungkin menyebabkan penurunan tersebut, seperti adanya pesaing baru, perubahan harga, atau ulasan negatif dari pelanggan. Dengan analisis diagnostik, kita dapat menggali lebih dalam untuk menemukan akar masalah dan mengambil tindakan yang tepat untuk memperbaikinya. Teknik umum yang digunakan untuk Analisis Diagnostik adalah Penambangan data, Korelasi, dan lain-lain.

c. Analisis Prediktif

Analisis prediktif digunakan untuk **memperkirakan apa yang akan terjadi di masa depan** seperti terlihat pada Gambar 4.4.



Gambar 4. 4 Analisis Prediktif

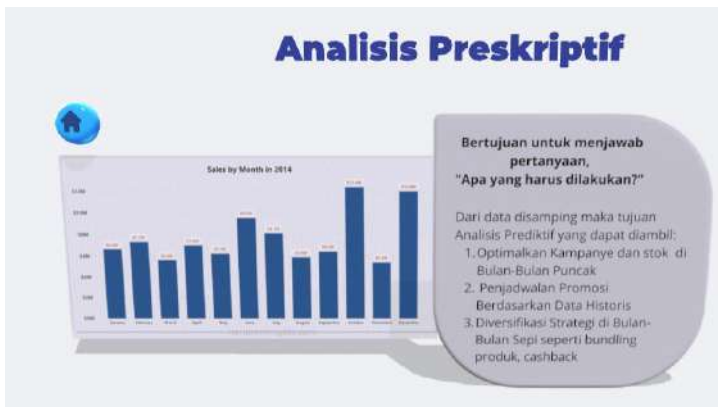
Hal ini melibatkan penggunaan data historis dan model statistik untuk memprediksi tren atau kejadian mendatang. Misalnya, perusahaan ritel dapat menggunakan analisis prediktif untuk memperkirakan penjualan di masa liburan mendatang berdasarkan data penjualan di tahun-tahun sebelumnya. Analisis prediktif sangat berguna dalam perencanaan dan pengambilan keputusan strategis, karena membantu organisasi mempersiapkan diri untuk peluang atau tantangan yang mungkin terjadi di masa depan.

Topik mendasar dan penting dalam analisis prediktif adalah pemodelan prediktif dan analisis dan optimasi keputusan. Analisis Prediktif melibatkan teknik-teknik tertentu seperti Analisis regresi, Statistik multivariat,

Pencocokan pola, Pemodelan prediktif, dan Peramalan. Teknik-teknik ini sulit dicapai oleh organisasi, karena mereka memerlukan jumlah data yang besar dan berkualitas tinggi. Teknik-teknik ini memerlukan pemahaman mendalam tentang statistika dan bahasa pemrograman seperti R dan Python. Meskipun sulit dicapai, nilai yang dihasilkan oleh Analisis Prediktif sangat besar.

d. Analisis Preskriptif

Tingkat terakhir dan paling tinggi dalam analitika adalah Preskriptif. Analisis Preskriptif adalah metode analitika yang menganalisis data untuk menjawab pertanyaan, **apa yang harus dilakukan** seperti terlihat pada Gambar 4.5.



Gambar 4. 5 Analisis Preskriptif

Analisis ini tidak hanya memprediksi masa depan, tetapi juga memberikan saran mengenai tindakan terbaik yang dapat diambil untuk mencapai hasil yang optimal. Misalnya, setelah memprediksi lonjakan permintaan produk selama musim liburan, analisis preskriptif dapat

merekomendasikan strategi inventaris terbaik, seperti berapa banyak stok yang harus disediakan dan kapan waktu terbaik untuk memesan lebih banyak produk. Analisis preskriptif sering kali menggunakan teknologi canggih seperti kecerdasan buatan (AI) untuk memberikan solusi yang lebih efektif dan berbasis data dalam pengambilan keputusan.

4.3 Tahapan Analitika Data

Proses analitika data melibatkan beberapa tahapan yang harus dilakukan secara berurutan untuk menghasilkan wawasan yang bermakna dan mendukung pengambilan keputusan. Setiap tahapan penting untuk memastikan data yang diolah akurat, relevan, dan dapat menghasilkan informasi yang bermanfaat. Gambar 4.6 merupakan langkah-langkah utama dalam proses analitika data. Anda dapat scan barcode yang diberikan untuk melihat dalam visualisasi AR.



Gambar 4. 6 Tahapan Data Analitik

a. *Data Requirement Gathering*

Data Requirement Gathering adalah langkah awal yang sangat penting dalam analisis data. Pada tahap ini, tim analisis bekerja sama dengan para pemangku kepentingan (stakeholder) untuk memahami tujuan proyek dan mengidentifikasi data yang dibutuhkan. Gambar 4.7 merupakan ilustrasi proses *data requirement gathering* dan Anda dapat scan barcode untuk melihat dalam visualisasi AR.



Gambar 4. 7 Ilustrasi Data Requirement Gathering

Tujuan utama dari proses ini adalah memastikan bahwa data yang dikumpulkan sesuai dengan kebutuhan agar bisa menghasilkan analisis yang akurat dan relevan. Jika proses pengumpulan kebutuhan data ini tidak dilakukan dengan benar, keputusan yang dihasilkan mungkin tidak sesuai dengan tujuan proyek. Hal ini dikenal dengan prinsip "Garbage In, Garbage Out" (GIGO), yang berarti bahwa data yang buruk akan menghasilkan hasil analisis yang juga buruk. Oleh karena itu, penting untuk memastikan bahwa data yang digunakan benar dan berkualitas untuk mendukung hasil analisis yang tepat. Berikut merupakan tahapan *data requirment gathering*.

- **Identifikasi Pemangku Kepentingan**

Pemangku kepentingan adalah pihak-pihak yang terdampak oleh hasil proyek. Pemangku kepentingan bisa saja Manajer, Pemilik Produk, Suplier, Pengguna, Bagian Pemasaran, dan lain-lain. Pastikan Anda memberikan pertanyaan yang efektif kepada setiap pemangku kepentingan.

- **Menentukan Tujuan**

Menentukan masalah bisnis atau pertanyaan analitis yang ingin dijawab. Memahami kebutuhan pemangku kepentingan

Menentukan hasil atau output yang diharapkan dari analisis data. Tentukan tujuan yang mudah dipahami, sederhana, singkat dan dapat diukur. Misalnya tujuannya adalah menentukan perbedaan produk pesaing dengan produk kita, maka kita dapat mulai dengan pertanyaan: **Bagaimana produk pesaing berbeda dengan produk kita?**

Untuk menjawab pertanyaan di atas kita butuh beberapa data seperti: data dari pelanggan mengenai preferensi produk dan spesifikasi produk perlu dikumpulkan untuk menjawab pertanyaan tersebut. Jangan membuang waktu dan uang dengan mengumpulkan data yang tidak relevan dengan tujuan karena analisis data merupakan prosedur yang Panjang dan terkadang mahal.

- **Klarifikasi Tujuan Analitika Data**

Pastikan semua pemangku kepentingan memahami tujuan analitik data, tetapkan ruang lingkup proyek, identifikasi kriteria keberhasilan, estimasi biaya, serta analisis risiko dan manfaat yang akan diperoleh.

- **Identifikasi Sumber Data**

Sumber data internal: Data yang sudah dimiliki oleh organisasi, seperti data penjualan, data pelanggan, dan data operasional.

Sumber data eksternal: Data dari luar organisasi yang mungkin diperlukan, seperti data pasar, data media sosial, dan data dari pihak ketiga.

- **Identifikasi Format dan Kualitas Data yang dibutuhkan.**

Pada tahap ini, detail spesifik tentang data yang dibutuhkan harus diidentifikasi:

- Jenis data: Apakah data yang diperlukan adalah data terstruktur, semi-terstruktur, atau tidak terstruktur.
- Atribut data: Detail spesifik tentang atribut atau kolom data yang diperlukan, seperti nama pelanggan, tanggal transaksi, jumlah penjualan, dll.
- Format data: Format data yang diinginkan, seperti CSV, JSON, XML, dll.
- Standar kualitas data (misalnya, tanpa nilai yang hilang, akurat).
- Seberapa banyak jumlah yang dibutuhkan (sampling).
- Durasi waktu data.

b. Data Collection

Data collection adalah proses pengumpulan data dari berbagai sumber yang besar (big) dan beragam. Tujuan utama data collection adalah untuk mengumpulkan informasi dan data yang relevan dan berkualitas tinggi, yang akan dianalisis untuk membuat sebuah keputusan bisnis yang krusial. Terdapat dua metode pengumpulan data seperti terlihat pada gambar 4.8 berikut dan anda dapat scan barcode yang diberikan untuk melihat visualisasi AR.



Gambar 4. 8 Metode Pengumpulan Data

- **Data Primer (Primary)**

Metode ini menggunakan data yang langsung diterima dari sumber utama. Data yang dihasilkan biasanya sangat akurat dan masih berhubungan erat dengan arah penelitian. Sayangnya, metode primer umumnya sangat makan waktu dan cukup mahal untuk dilakukan. Metode pengumpulan data primer biasanya: Survey, Kuesioner, Wawancara, Observasi, Experiment, Studi Kasus, Focus Group Discussions, Sampling, dan lain-lain

- **Data Sekunder (Secondary)**

Metode sekunder tidak membutuhkan pengumpulan data langsung dari responden. Metode ini menggunakan data dan informasi yang sebelumnya dihimpun dari sejumlah sumber: Literature Review, Database Pemerintah, Commercial Database (Kaggle, Uci, dll), Web Scraping (Sosial Medial).

Terdapat beberapa tantangan dalam pengumpulan data seperti:

- Mengetahui data apa yang harus dikumpulkan, ini seringkali memerlukan pemahaman atau pengetahuan tentang domain masalah.
- Pencarian Data
- Sumber data yang teridentifikasi
- Penilaian kualitas data
- Relevansi Data
- Tidak ada Pengenal Umum
- Data yang hilang, kesalahan penginputan data
- Perubahan jenis data seiring waktu atau antar sampel
- Validasi dan Keandalan data

c. Data Cleaning

Data cleaning adalah proses mengidentifikasi dan memperbaiki (atau menghapus) kesalahan atau inkonsistensi dalam dataset seperti duplikasi, format yang tidak tepat, outlier dll. Proses data cleaning untuk memastikan bahwa data yang digunakan dalam analisis atau pelaporan adalah akurat, konsisten, dan relevan. Kesalahan dalam data dapat muncul karena berbagai alasan, seperti kesalahan input manusia, kerusakan perangkat keras atau perangkat lunak, atau kesalahan selama proses pengumpulan data. Ketika menggabungkan beberapa sumber data, ada banyak kemungkinan duplikasi data dan pelabelan yang tidak tepat.

Jika data tidak benar, maka hasil dan algoritma yang digunakan juga tidak dapat efektif.

Tujuan data cleaning adalah:

- **Akurasi Data:** Memastikan data akurat dan dapat diandalkan.
- **Konsistensi Data:** Menyamakan format dan unit pengukuran data.
- **Relevansi Data:** Menghapus atau memperbaiki data yang tidak relevan atau usang.
- **Kelengkapan Data:** Mengisi nilai yang hilang atau memperbaiki data yang tidak lengkap.
- **Integritas Data:** Memeriksa integritas referensial dan keterkaitan antar data.

Dalam data cleaning terdapat beberapa proses yang umumnya dilakukan seperti terlihat pada Gambar 4.9 dan Anda dapat scan barcode yang diberikan untuk melihat visualisasi AR.



Gambar 4. 9 Proses Umum Data Cleaning

- Remove Duplicate Data

Ketika menggunakan dataset besar atau menggunakan dataset dari berbagai sumber akan terdapat banyak kemungkinan data duplikat. Data duplikat dapat menambah redundansi pada data sehingga dapat mengakibatkan hasil perhitungan menjadi tidak efektif. Misalnya Kode Produk yang duplikat akan memberikan Anda jumlah produk yang lebih banyak dibandingkan data yang sebenarnya atau alamat email/telepon yang duplikat membuat komunikasi Anda terlihat seperti spam.

Misalnya: Dataset Startup Success Prediction merupakan dataset dari kumpulan perusahaan startup yang berada di negara Amerika Serikat. <https://www.kaggle.com/datasets/manishkc06/startup-success-prediction>

Tahap awal sebelum melakukan proses remove duplicate adalah membaca dataset.

```
#Membaca Dataset
data = pd.read_csv('/content/sample_data/startup_data.csv')
```

Melakukan pengecekan duplicate data.

```
#Cek Duplicate Data
data.duplicated(subset=['name']).sum()
```

```
#Cek Duplicate Data
data.duplicated(subset=['name']).sum()
```

	A	B	C	D	E	F	G	H	I	J	
Unnamed: 0	state	city	latitude	longitude	zip	code	id	city	Unnamed: 0	name	labels
506	CA		3.754.827	-121.908.572	94538	c:28482	Fremont	Fremont	C	Redwood Systems	1
505	CA		374.815.098	-1.219.453.283	94538	c:28482	Fremont			Redwood Systems	1

- Remove Irrelevant Data

Misalkan Anda sedang menganalisis Prediksi Kesuksesan Startup, maka Anda akan mendapatkan data-data yang berisi berbagai atribut penting yang berhubungan dengan startup seperti nama startup, kategori, pendanaan, milestones, funding rounds, jumlah mitra/relasi, tahun pertama kali mendapatkan pendanaan, investor, dan lama berdirinya startup. Selain atribut yang disebutkan di atas, Anda juga tentu mendapatkan berbagai data lainnya seperti kode negara berdirinya startup, latitude, longitude, id startup, tanggal berdirinya, tanggal tutup, dan kode category startup.

- Nah, pertanyaannya selanjutnya adalah apakah semua informasi tersebut kita butuhkan untuk proses analisis kesuksesan startup? Jawabannya adalah tidak. Berikut ini merupakan atribut yang tidak relevan dengan tujuan analisis yang akan dilakukan

No	Variabel	Keterangan
1	Unnamed: 0	Kode unik startup
2	state_code	Kode negara berdirinya startup
3	latitude	Latitude negara berdirinya startup
4	longitude	Longitude negara berdirinya startup
5	zip_code	Kode pos berdirinya startup
6	id	Id startup
7	city	Kota tempat berdirinya startup
8	Unnamed: 6	Gabungan variabel state code, city dan zip_code
9	name	Nama startup
10	Labels	Berisi nilai 1 dan 0 yang menyatakan status startup acquired atau closed [sama dengan variabel status]
11	founded_at	Tanggal berdirinya startup
12	closed_at	Tanggal tutupnya startup
13	first_funding_at	Tanggal pendanaan pertama
14	last_funding_at	Tanggal pendanaan terakhir
15	state_code_1	Kode negara [sama dengan variabel state_code]
16	category_code	Jenis kode kategori startup
17	object_id	Id startup [sama dengan variabel id]

- Standardize Capitalization

Anda harus memastikan bahwa teks dalam data Anda konsisten. Jika kapitalisasi Anda tidak konsisten,

hal ini dapat menghasilkan pembuatan kategori palsu. Misalnya memiliki nama kolom sebagai "Total_Sales" dan "total_sales" berbeda (kebanyakan bahasa pemrograman bersifat case-sensitive). Untuk menghindari kebingungan dan menjaga keseragaman antara nama kolom, kita seharusnya mengikuti cara standar dalam memberikan nama kolom.

Misalnya saat mengumpulkan nama dan alamat email melalui formulir online, survei, atau metode lainnya, kita mendapatkan input dalam gaya yang beragam.

Kita dapat memperbaikinya untuk menghindari kesalahan entri duplikat. Alamat email 'wulan@mikroskil.ac.id' dan 'WULAN@MIKROSKIL.AC.ID' dapat dianggap sebagai alamat email yang berbeda, sehingga lebih baik membuat semua huruf alamat email di kolom tersebut menjadi huruf kecil. Demikian pula, untuk nama, kita dapat mengikuti gaya title case di mana semua kata ditulis dengan huruf kapital.

Pada case dataset startup prediction success seluruh data sudah sesuai dengan startand sehingga tidak perlu untuk diperbaiki

- **Convert Data Type**

Ketika bekerja dengan data CSV di Python, Pandas akan mencoba menebak jenis data yang kita gunakan pada dataset. Jenis data paling umum yang kita temui dalam data adalah jenis data teks, numerik, dan tanggal. Jenis data teks dapat menerima segala jenis nilai campuran termasuk huruf, angka, atau

karakter khusus. Contoh data teks Nama orang, jenis produk, lokasi toko, alamat email, kata sandi, dll.

Jenis data numerik berisi nilai bilangan bulat atau angka dengan titik desimal, juga disebut float. Memiliki kolom tipe data numerik berarti Anda dapat melakukan komputasi matematika seperti menemukan nilai minimum, maksimum, rata-rata, dan median, atau menganalisis distribusi menggunakan histogram, box plot, q-q plot, dll. Namun untuk kolom id, telepon sebaiknya tetap diset sebagai data teks. Jika jenis data tidak diidentifikasi dengan benar, data akan diidentifikasi sebagai kolom string atau teks. Dalam hal ini, kita perlu secara eksplisit menentukan jenis data kolom dan format tanggal yang disebutkan dalam data. Kolom tanggal dapat direpresentasikan dalam berbagai format: 2 Oktober 2023, 02-10-2023, 2023/10/02, 2-Okt-2023.

Berikut adalah kolom tanggal pada dataset startup yang sudah sesuai dengan format.

founded_at	closed_at	first_funding_at	last_funding_at
06/01/2012		6/30/2012	6/30/2012
07/01/2010		01/01/2010	2/16/2011
01/01/2004		05/08/2006	04/11/2012
01/01/2007		04/01/2007	7/23/2009
07/01/2007	09/01/2011	01/01/2007	10/01/2008
01/01/2007	11/01/2012	07/06/2010	12/15/2010
02/05/2013		02/05/2013	02/05/2013
08/01/2008	05/01/2013	01/01/2008	12/04/2009
01/01/2002		9/16/2005	12/14/2010
03/01/2007		06/03/2008	05/04/2010
02/01/2000		05/01/2000	07/01/2007
11/01/2006		02/01/2007	3/28/2011
01/01/2003	02/01/2001	09/01/2004	05/01/2012
01/01/2003		02/01/2006	02/01/2006
06/01/2011		10/01/2012	11/01/2012

- **Handling Outlier**

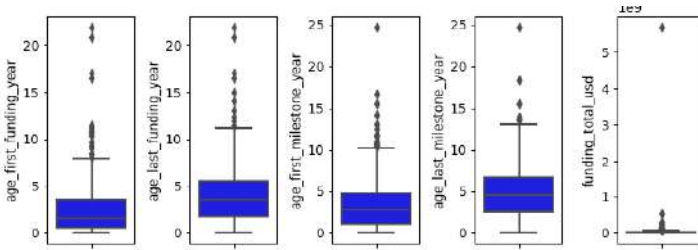
Outlier adalah titik data dalam statistik yang sangat berbeda dari pengamatan lainnya. Outlier mungkin mencerminkan variabilitas pengukuran, atau dapat menunjukkan kesalahan eksperimental yang terakhir kadang-kadang dihapus dari set data. Misalnya kita sedang melakukan analisis harga pizza. Harga pizza bervariasi antara INR 100 hingga INR 7500 di wilayah tersebut setelah melakukan survei di sekitar 500 restoran. Tetapi setelah analisis, kita menemukan bahwa ada hanya satu catatan dalam dataset dengan harga pizza sebesar INR 7500, sedangkan sisa harga pizza lainnya berkisar antara INR 100 hingga INR 1500. Oleh karena itu, pengamatan dengan harga pizza sebesar INR 7500 adalah outlier karena secara signifikan berbeda dari populasi.

Outlier-outlier ini biasanya diidentifikasi menggunakan box plot atau scatter plot. Outlier-outlier ini mengakibatkan data yang miring dan mempengaruhi hasil algoritma sehingga penting untuk ditangani sebelum proses lebih lanjut. Terdapat dua cara untuk menangani outlier yaitu (1) menghapus pengamatan yang terdiri dari nilai-nilai outlier dan (2) menggunakan transformasi seperti logaritma, akar kuadrat, box-cox, dll., untuk membuat nilai-nilai data mengikuti distribusi normal atau mendekati normal.

Pada dataset prediksi startup terdapat beberapa atribut yang memiliki outlier seperti berikut.

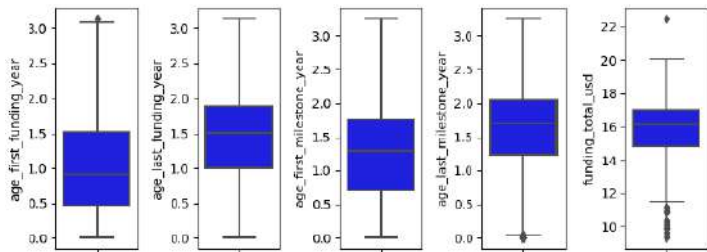
```
#outlier
features_humfinal = ['age_first_funding_year', 'age_last_funding_year', 'age_first_milestone_year', 'age_last_milestone_year', 'funding_total_usd']

plt.figure(figsize=(15, 7))
for i in range(0, len(features_humfinal)):
    plt.subplot(1, len(features_humfinal), i+1)
    sns.boxplot(y=df[features_humfinal[i]], color='blue', orient='v')
plt.tight_layout()
```



Untuk mengatasi outlier pada data maka digunakan fungsi log-transformation untuk setiap variable, sehingga data dapat terdistribusi normal seperti berikut.

```
df["age_first_funding_year"] = np.log1p(df["age_first_funding_year"])
df["age_last_funding_year"] = np.log1p(df["age_last_funding_year"])
df["age_first_milestone_year"] = np.log1p(df["age_first_milestone_year"])
df["age_last_milestone_year"] = np.log1p(df["age_last_milestone_year"])
df["funding_total_usd"] = np.log1p(df["funding_total_usd"])
```



- Fix Errors

Kesalahan dalam data dapat membuat Anda kehilangan temuan-temuan kunci. Hal ini harus dihindari dengan memperbaiki kesalahan yang mungkin ada dalam data Anda. Sistem-sistem yang memasukkan data secara manual tanpa adanya

pemeriksaan data hampir selalu akan mengandung kesalahan. Untuk memperbaikinya, pertama-tama kita perlu memahami data tersebut. Setelah itu, kita dapat mendefinisikan logika atau memeriksa data dan kemudian memperbaiki kesalahan data. Pada case dataset startup prediction success terdapat 4 atribut/variable yang memiliki nilai negative.

No	Variabel	Negative Value
1	age first milestone year	46
2	age last milestone year	12
3	age first funding year	46
4	age last funding year	12

```
age=["age_first_funding_year","age_last_funding_year","age_first_milestone_year","age_last_milestone_year"]
for a in range(len(age)):
    print("Is there any negative value in '{}' column : {}".format(age[a],(df[age[a]]<0).any()))
```

Pada case ini seluruh nilai negative akan dihapus sehingga tidak ada lagi yang bernilai negatif.

```
df=df.drop(df[df.age_first_funding_year<0].index)
df=df.drop(df[df.age_last_funding_year<0].index)
df=df.drop(df[df.age_first_milestone_year<0].index)
df=df.drop(df[df.age_last_milestone_year<0].index)

for a in range(len(age)):
    print("Is there any negative value in '{}' column : {}".format(age[a],(df[age[a]]<0).any()))

Is there any negative value in 'age_first_funding_year' column : False
Is there any negative value in 'age_last_funding_year' column : False
Is there any negative value in 'age_first_milestone_year' column : False
Is there any negative value in 'age_last_milestone_year' column : False
```

- Handling Missing Value

Selama pembersihan dan transformasi data dalam ilmu data, menangani nilai yang hilang adalah salah satu tugas paling umum. Data dunia nyata mungkin mengandung nilai-nilai yang hilang yang perlu diperbaiki sebelum data dapat digunakan untuk analisis.

Kita dapat menangani nilai-nilai yang hilang dengan:

- Menghapus catatan yang memiliki nilai-nilai yang hilang atau
- Mengisi nilai-nilai yang hilang menggunakan beberapa teknik statistik atau dengan memahami pemahaman data.

Aturan umum dalam **missing value** adalah Anda dapat menghapus nilai-nilai yang hilang jika mereka mencapai **kurang dari lima persen dari total jumlah catatan**, tetapi hal ini tergantung pada analisis, pentingnya nilai-nilai yang hilang, ukuran data, dan kasus penggunaan yang sedang Anda kerjakan. Pertimbangkan sebuah dataset yang berisi parameter kesehatan tertentu seperti glukosa, tekanan darah, insulin, IMT, usia, diabetes, dll. Tujuannya adalah membuat model klasifikasi terawasi yang memprediksi apakah seseorang mungkin memiliki diabetes atau tidak berdasarkan parameter kesehatan. Jika data memiliki nilai-nilai yang hilang untuk kolom glukosa dan tekanan darah untuk beberapa individu, maka data ini tidak dapat dihapus wajib untuk mengisi nilai-nilai yang hilang dengan Teknik yang ada.

Pada dataset prediksi startup terdapat beberapa atribut yang memiliki missing value seperti berikut.

```
#Pengecekan Missing Value
null=pd.DataFrame(data.isnull().sum(),columns=["Null Values"])
null["% Missing Values"]=(data.isna().sum()/len(data)*100)
null = null[null["% Missing Values"] > 0]
null.style.background_gradient(cmap='viridis',low =0.2,high=0.1)
```

Variabel	Jumlah Missing Value	Persentasi
closed_at	588	63.71%
Unnamed: 6	493	53.41%
age_last_milestone_year	152	16.47%
age_first_milestone_year	152	16.47%
state_code.1	1	0.11%

Berdasarkan table di atas **closed_at** dan **Unnamed: 6** memiliki persentase terbanyak. **Closed_at** adalah variabel yang berisi tanggal tutupnya sebuah startup dan **Unnamed:6** berisi gabungan beberapa variabel yaitu **state_code**, **city** dan **zip_code**. Kedua variabel tersebut tidak mendukung dalam pembuatan model prediksi sehingga dilakukan penghapusan variabel.

Selain kedua variabel tersebut, variabel **state_code.1** juga dihapus meskipun persentase missing value-nya kecil karena memiliki nilai yang sama dengan variabel **state_code**.

```
df.drop(["Unnamed: 6"], axis=1, inplace=True)
df.drop(["closed_at"], axis=1, inplace=True)
df.drop(["state_code.1"], axis=1, inplace=True)
```

Untuk mengatasi missing value pada variabel **age_last_milestone_year** dan **age_first_milestone_year** maka dilakukan analisis terhadap kedua variabel tersebut.

Variabel **age_first_milestone_year** memiliki nilai rata-rata sebesar 3.055 dengan median 2.520 dan **age_last_milestone_year** memiliki nilai rata-rata sebesar 4.754 dengan median 4,477 yang menunjukkan bahwa data dari kedua variabel tersebut terdistribusi tidak normal. Sehingga missing value pada

kedua variabel tidak dapat diisi dengan nilai rata-rata. Kemudian dilakukan analisis kedua variabel dengan variabel milestones sehingga diperoleh hasil bahwa startup yang belum memiliki milestones atau nilai milestones = 0 maka seluruh variabel age_last_milestone_year dan age_first_milestone_year memiliki nilai null (NaN).

Untuk mengatasi seluruh missing value pada variabel **age_first_milestone_year** dan **age_last_milestone_year** dilakukan dengan **teknik imputation**. Salah satu cara untuk melakukan imputation adalah dengan memberikan nilai konstanta yang merupakan bagian dari kumpulan nilai yang mungkin dari variabel yang mengalami missing value. Pada Dataset Startup ini nilai konstanta yang digunakan adalah angka 0 yang artinya startup tersebut belum memiliki milestones sama sekali selama beroperasi.

```
#Isi dengan nilai 0
df['age_first_milestone_year'] = df['age_first_milestone_year'].fillna(value="0")
df['age_last_milestone_year'] = df['age_last_milestone_year'].fillna(value="0")
```

d. Data Analysis

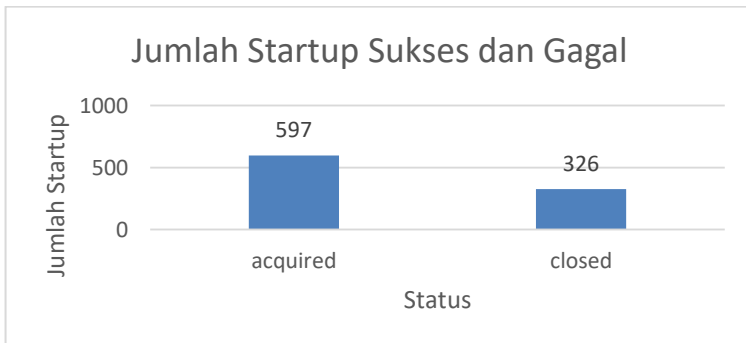
Analisis Data merupakan proses untuk mengorganisir, menginterpretasikan, memahami, dan menggali informasi dan wawasan dari data. Proses analisis data merupakan langkah penting untuk mendapatkan pemahaman yang mendalam tentang masalah atau fenomena yang akan diteliti. Tujuan analisis data adalah untuk mengidentifikasi pola, tren, hubungan, dan informasi penting lainnya yang dapat membantu dalam pengambilan keputusan yang informasional dan rasional.

Misalnya kita akan melakukan analisis dataset startup success prediction :

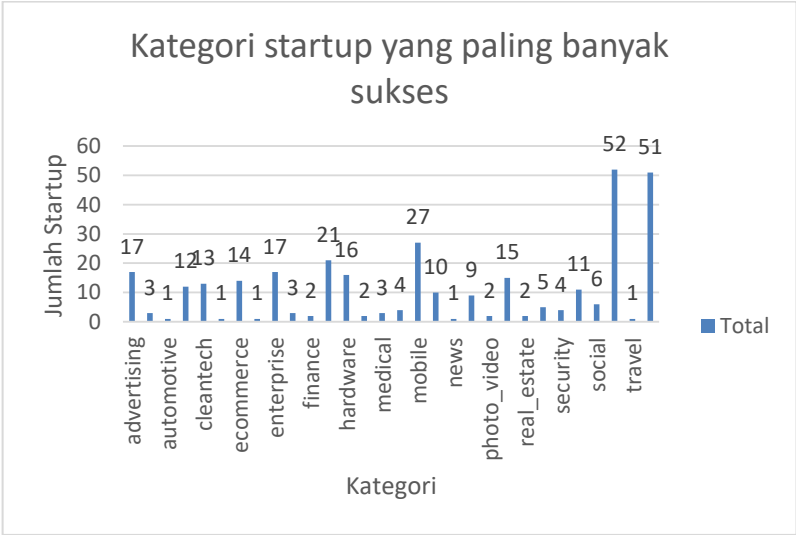
<https://www.kaggle.com/datasets/manishkc06/startup-success-prediction>

Berdasarkan data tersebut kita ingin mengetahui:

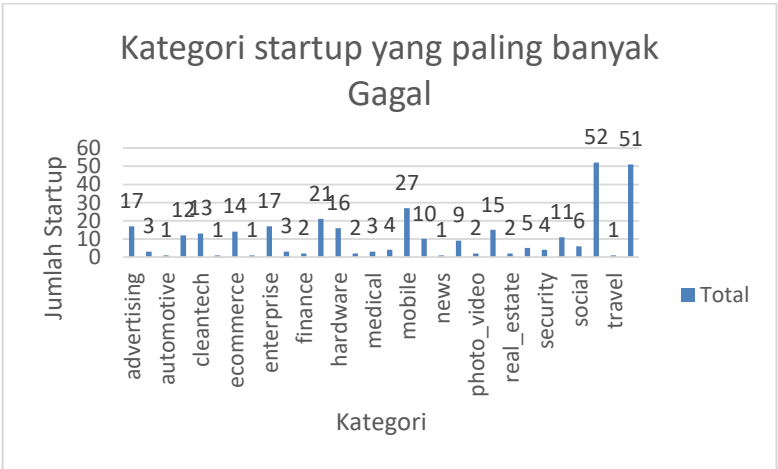
- a. Berapa banyak jumlah startup yang sukses dan gagal?
Untuk menjawab pertanyaan di atas kita dapat menggunakan fungsi sum untuk kolom status sehingga diperoleh 597 jumlah yang sukses dan 326 jumlah yang gagal.



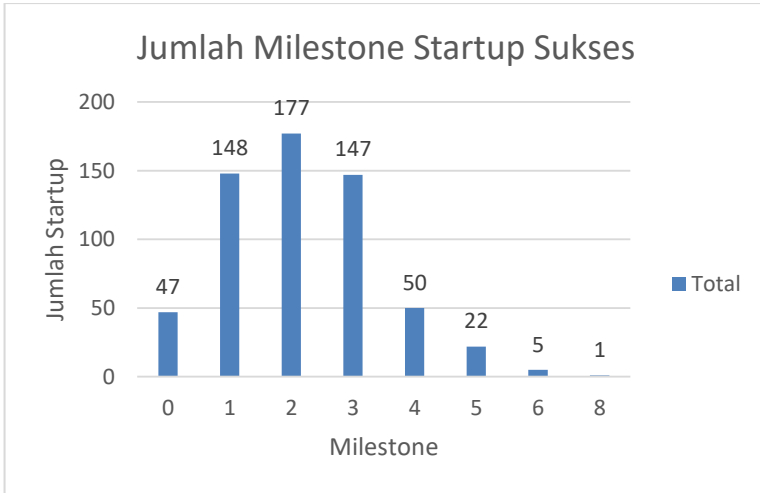
- b. Kategori startup seperti apa yang paling banyak sukses?
Untuk menjawab pertanyaan di atas kita dapat menggunakan kolom **status** dan **kategori** sehingga kita mengetahui bahwa kategori **Software** dan **Web** yang paling banyak sukses.



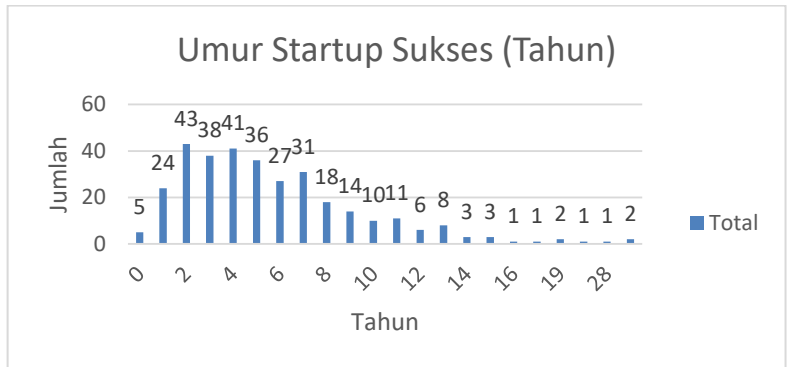
- c. Kategori startup seperti apa yang paling banyak gagal? Untuk menjawab pertanyaan di atas kita dapat menggunakan dari kolom **status** dan **kategori** sehingga kita mengetahui bahwa kategori **Software** dan **Web** yang paling banyak sukses.



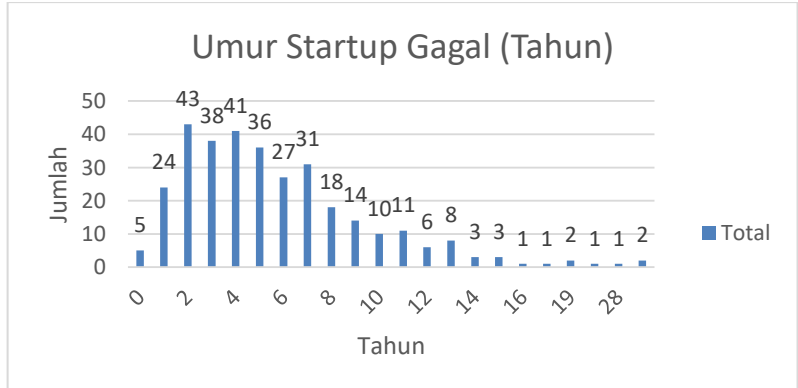
- d. Berapa jumlah milestone startup yang sukses?
Kita dapat menggunakan kolom **milestones** dan **status**.



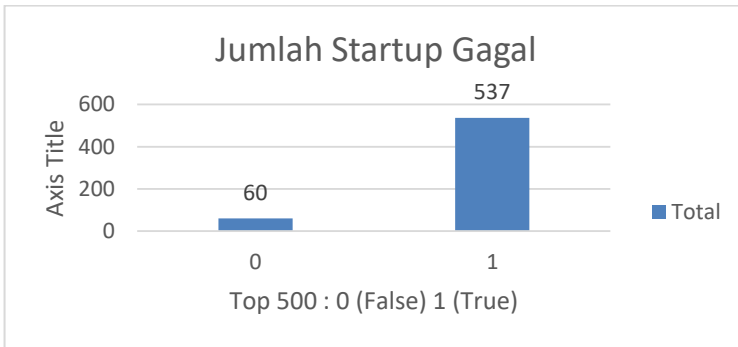
- e. Berapa minimal dan maximal umur startup yang sukses?
Kita perlu lakukan pengurangan untuk kolom **founded_at** dan **closed_at**.



- f. Berapa minimal dan maximal umur startup yang gagal? Kita perlu lakukan pengurangan untuk kolom **founded_at** dan **closed_at**.

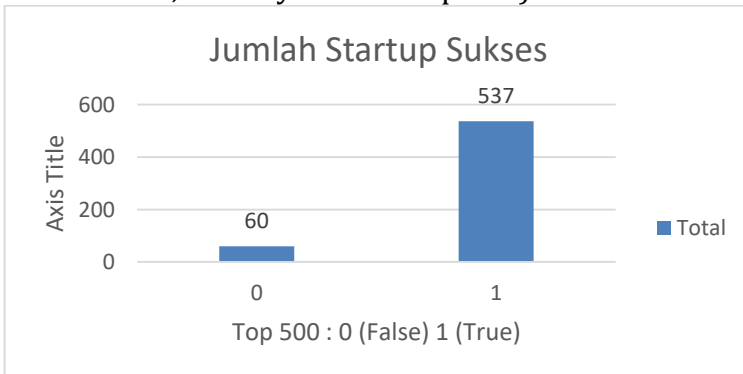


- g. Berapa banyak startup yang masuk top 500 dunia namun gagal? Kita dapat menggunakan kolom **is_top500** (0 artinya tidak masuk, 1 artinya masuk top 500) dan **status**.



- h. Berapa banyak startup yang tidak masuk top 500 dunia namun sukses?

Kita dapat menggunakan kolom **is_top500** (0 artinya tidak masuk, 1 artinya masuk top 500) dan **status**.



4.4 Rangkuman

Analitika data merupakan proses penting dalam mengolah data mentah untuk mendapatkan wawasan yang dapat mendukung pengambilan keputusan. Di era modern ini, dengan ledakan volume data yang dikenal sebagai big data, analitika data menjadi kunci dalam berbagai sektor seperti bisnis, kesehatan, pemerintahan, dan pendidikan. Melalui analisis data yang terstruktur, organisasi dapat

mengoptimalkan strategi, meningkatkan efisiensi operasional, dan membuat keputusan yang lebih tepat.

Proses analitika data terdiri dari berbagai tahapan penting. Tahap awal adalah Data Requirement Gathering, yang bertujuan untuk memahami kebutuhan analitik dan memastikan data yang dikumpulkan sesuai dengan tujuan analisis. Setelah itu, data dikumpulkan melalui dua metode utama, yaitu data primer dan data sekunder. Selanjutnya, data yang dikumpulkan harus dibersihkan (Data Cleaning) untuk menghapus duplikasi, memperbaiki kesalahan, dan memastikan konsistensi data. Terakhir analisis data untuk menghasilkan wawasan yang akurat dan relevan, yang sangat berguna dalam pengambilan keputusan yang berbasis data.

DAFTAR PUSTAKA

- Lestari, W.S. and Halim, A. (2022) 'Prediksi Kesuksesan Startup menggunakan deep neural network', Jurnal SIFO Mikroskil, 23(2), pp. 99–110. doi:10.55601/jsm.v23i2.885.
- KC, Manish, "Startup Success Prediction", kaggle, September 2020, [Online], Tersedia: <https://www.kaggle.com/datasets/manishkc06/startup-success-prediction> [Diakses: 1 Juli 2024].
- Tmarch. (2020, February 4). 4 types of data analytics for educators. Tom March. <https://tommarch.com/2020/01/4-types-data-analytics-for-educators/>.

BAB 5

DATA STORYTELLING

5.1 Pendahuluan

Data storytelling adalah seni mengubah data menjadi cerita yang dapat dipahami oleh semua orang, tidak hanya oleh ahli data. Proses ini menggabungkan analisis data dengan narasi dan visualisasi untuk menyampaikan pesan yang lebih jelas, menarik, dan relevan. Pada dasarnya, data storytelling membantu kita menjelaskan apa yang terjadi, mengapa hal tersebut penting, dan tindakan apa yang sebaiknya diambil berdasarkan data yang telah dianalisis.

Konsep data storytelling bertujuan membangun narasi menarik dari data dan analisis kompleks yang membantu mengkomunikasikan temuan penting kepada audiens yang spesifik, seperti pemilik bisnis. Praktik ini menggabungkan keterampilan dari berbagai disiplin ilmu, termasuk komunikasi, analisis, dan desain. Data storytelling digunakan di berbagai bidang untuk mengatasi berbagai tantangan dengan menyajikan informasi yang relevan secara lebih efektif dan berdampak.

Data storytelling sangat penting karena dapat menjembatani kesenjangan antara analisis data yang kompleks dan pemahaman audiens yang lebih luas. Dalam dunia bisnis, pemerintahan, atau bahkan pendidikan, tidak semua orang memiliki kemampuan untuk memahami data secara langsung. Data storytelling menyederhanakan hal ini, sehingga semua pihak—termasuk yang non-teknis—dapat memahami dan menggunakan informasi yang dihasilkan dari data untuk membuat keputusan yang lebih baik.

Lebih jauh lagi, data storytelling dapat memengaruhi keputusan dan mendorong aksi. Narasi yang didukung oleh data yang kuat dan visualisasi yang tepat membantu membangun kepercayaan, memperkuat pesan, dan memudahkan audiens untuk terhubung dengan informasi yang disampaikan. Oleh karena itu, data storytelling bukan hanya alat untuk mempresentasikan data, tetapi juga strategi penting dalam menyampaikan wawasan yang dapat berdampak besar pada pengambilan keputusan.

Sebelum kita mulai menceritakan data, ada beberapa pertanyaan dasar yang perlu dijawab:

1. **Untuk siapa Anda menulis ini?**

Apakah ini disajikan kepada CEO, Manajer Proyek, Kepala Tim, atau mungkin kepada pelanggan/pengguna?

2. **Apa yang ingin Anda sampaikan?**

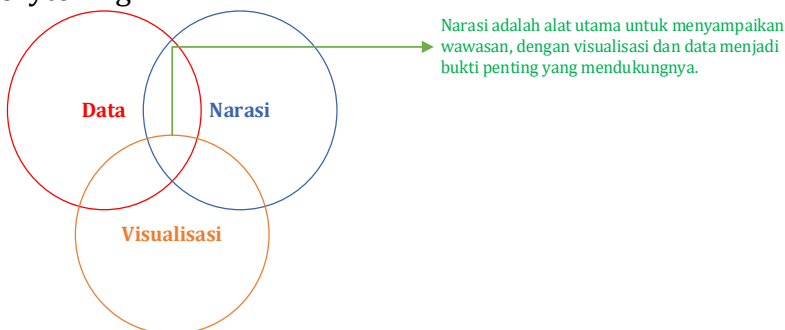
Tentukan bagian dan fitur data yang ingin dipresentasikan.

3. **Mengapa ini penting?**

Motivasi untuk membentuk cerita harus menjelaskan mengapa informasi ini sangat penting.

5.2 Komponen Data Storytelling

Berikut ini merupakan komponen utama data storytelling.



Gambar 5. 1 Komponen Data Storytelling

a. Data

Elemen pertama yang penting dalam data storytelling adalah **data yang kuat**. Data ini harus akurat, relevan, dan mampu menjawab pertanyaan yang diajukan. Data yang baik merupakan fondasi dari cerita yang kuat.

b. Narasi

Setelah memiliki data yang valid, elemen kedua yang perlu diperhatikan adalah **narasi yang jelas**. Narasi membantu menghubungkan data dengan audiens. Cerita yang baik harus logis, mudah dipahami, dan menjawab pertanyaan-pertanyaan penting seperti apa yang terjadi, mengapa itu penting, serta bagaimana kita dapat bertindak berdasarkan informasi tersebut.

c. Visualisasi

Visualisasi menjadi elemen penting ketiga dalam data storytelling. **Visualisasi yang tepat** membantu menyederhanakan data kompleks dan membuatnya lebih mudah dimengerti. Grafik, diagram, atau peta dapat membantu memperjelas pesan utama yang ingin disampaikan, sehingga audiens dapat dengan cepat memahami inti dari cerita yang disampaikan.

Misalnya, jika kita ingin menceritakan tren penjualan bulanan dari sebuah perusahaan, kita bisa memulai dengan data penjualan dalam bentuk tabel. Namun, narasi yang efektif akan menghubungkan angka-angka tersebut dengan konteks, seperti mengapa terjadi peningkatan penjualan pada bulan tertentu. Mungkin karena ada promosi khusus atau perubahan tren konsumen. Untuk memperkuat cerita ini, kita bisa menambahkan grafik garis yang menunjukkan perubahan penjualan dari bulan ke bulan, sehingga audiens bisa melihat pola yang terjadi secara visual.

Kesuksesan data storytelling tidak hanya bergantung pada angka-angka, tetapi juga pada kemampuan kita untuk menyampaikan pesan utama dengan cara yang mudah

dipahami dan menarik. Dengan memadukan **data yang kuat, narasi yang jelas, dan visualisasi yang mendukung**, data storytelling mampu mengubah data menjadi wawasan yang dapat dipahami oleh siapa saja, sehingga membantu dalam pengambilan keputusan yang lebih baik. Oleh karena itu, data storytelling tidak hanya sekadar alat untuk mempresentasikan data, tetapi juga strategi penting dalam menyampaikan wawasan yang dapat berdampak besar pada pengambilan keputusan.

5.3 Contoh Data Storytelling

Berikut adalah beberapa contoh storytelling data yang efektif, menggambarkan cara menyajikan data dalam format naratif yang menarik:

a. Dampak Pendidikan terhadap Tingkat Pendapatan

- **Narasi:** Sebuah cerita yang menjelaskan hubungan antara pencapaian pendidikan dan tingkat pendapatan di berbagai demografis.
- **Data Point:** Tampilkan statistik tentang gaji rata-rata berdasarkan tingkat pendidikan (SMA, S1, S2, dll.) dan soroti perbedaan di antara kelompok ras dan etnis.
- **Visual:**
 - Grafik batang membandingkan pendapatan rata-rata untuk berbagai kualifikasi pendidikan.
 - Grafik garis menunjukkan pertumbuhan pendapatan dari waktu ke waktu berdasarkan pendidikan.
- **Insight:** Narasi ini menekankan pentingnya akses terhadap pendidikan berkualitas sebagai jalan menuju pendapatan yang lebih tinggi, mendorong diskusi tentang reformasi kebijakan pendidikan.

b. Analisis Preferensi Konsumen Selama COVID-19

- **Narasi:** Analisis mendetail tentang bagaimana perilaku pembelian konsumen berubah selama pandemi, dengan

fokus pada pertumbuhan e-commerce dan pergeseran permintaan produk.

- **Data Poin:** Statistik tentang pertumbuhan penjualan online, produk terlaris, dan pergeseran demografis dalam pola pembelian.
- **Visuals:**
Grafik garis menunjukkan lonjakan penjualan e-commerce selama periode penguncian.
Grafik lingkaran yang membagi penjualan berdasarkan kategori (misalnya, bahan makanan, peralatan kebugaran rumah).
- **Insight:** Narasi ini membantu bisnis memahami tren yang muncul dan menyesuaikan strategi pemasaran mereka sesuai, menekankan pentingnya transformasi digital.

c. Evolusi Penggunaan Media Sosial

- **Narasi:** Eksplorasi bagaimana penggunaan media sosial telah berkembang selama dekade terakhir, dengan fokus pada perubahan demografis dan preferensi platform.
- **Data Poin:** Statistik tentang pertumbuhan pengguna untuk platform utama (Facebook, Instagram, TikTok) dan pembagian demografis pengguna berdasarkan usia, jenis kelamin, dan lokasi.
- **Visuals:**
Grafik garis menunjukkan tren pertumbuhan pengguna untuk setiap platform dari waktu ke waktu.
Grafik batang membandingkan tingkat keterlibatan di antara berbagai demografis.
- **Insight:** Narasi ini memberikan wawasan tentang bagaimana merek dapat berinteraksi secara efektif dengan audiens target mereka di berbagai platform, menekankan perlunya strategi konten yang disesuaikan.

Contoh-contoh ini menunjukkan bagaimana storytelling data dapat menyampaikan informasi kompleks dengan cara yang menarik, melibatkan audiens, dan menghasilkan wawasan yang berarti.

Tips untuk menceritakan data yang efektif

- a. **Identifikasi Pesan yang Jelas:** Mulailah dengan tema atau wawasan sentral yang ingin Anda komunikasikan.
- b. **Gunakan Visual yang Menarik:** Pilih visual yang melengkapi dan meningkatkan narasi, membuat data kompleks lebih mudah dipahami.
- c. **Kontekstualisasikan Data:** Berikan konteks untuk membantu audiens memahami signifikansi data dan implikasinya.
- d. **Sertakan Contoh Kehidupan Nyata:** Hubungkan data dengan skenario dunia nyata atau studi kasus untuk membuat cerita lebih mudah dipahami.
- e. **Dorong Interaksi:** Jika memungkinkan, masukkan elemen interaktif yang memungkinkan audiens menjelajahi data lebih lanjut.

5.4 Rangkuman

Data storytelling adalah proses yang mengubah data menjadi cerita yang mudah dipahami oleh semua orang, tidak hanya oleh ahli data. Proses ini menggabungkan analisis, narasi, dan visualisasi untuk menyampaikan pesan yang jelas dan menarik, membantu menjelaskan situasi, mengapa informasi itu penting, dan tindakan apa yang perlu diambil berdasarkan analisis data. Dengan menyederhanakan data kompleks, teknik ini memungkinkan audiens non-teknis untuk memahami informasi dan membuat keputusan yang lebih baik. Selain itu, narasi yang kuat dan visualisasi yang tepat dapat mempengaruhi keputusan dan mendorong aksi, menjadikan data storytelling strategi penting dalam pengambilan keputusan.

Tiga komponen utama dalam data storytelling adalah data yang kuat dan akurat sebagai fondasi cerita, narasi yang jelas untuk menghubungkan data dengan audiens, serta visualisasi yang membantu menyederhanakan data dan menyampaikan pesan dengan jelas. Agar data storytelling efektif, penting untuk mengidentifikasi pesan yang jelas, menggunakan visual yang menarik, memberikan konteks, menyertakan contoh kehidupan nyata, dan mendorong interaksi untuk meningkatkan pemahaman audiens.

DAFTAR PUSTAKA

- Knaflig, C.N. (2015) *Storytelling with data: The Effective Visual Communication of Information*. Hoboken, NJ: John Wiley & Sons.
- Joshi, P. and Mahalle, P.N. (2023) *Data storytelling and visualization with tableau: A hands-on approach*. Boca Raton, FL: CRC Press.
- Hwang, J. and Yoon, Y. (no date) *Data Analytics and visualization in quality analysis using tableau*. CRC Press.

BAB 6

VISUALISASI DATA

6.1 Pendahuluan

Dalam era informasi yang semakin melimpah, data telah menjadi aset berharga bagi berbagai sektor kehidupan. Namun, data mentah yang begitu banyak seringkali sulit dipahami dan diinterpretasikan oleh manusia. Di sinilah visualisasi data memainkan peran krusial. Visualisasi data adalah proses penggambaran data dalam bentuk visual, seperti grafik, diagram, atau peta, dengan tujuan memudahkan pemahaman dan analisis informasi. Dengan mengubah data numerik menjadi representasi visual, kita dapat mengidentifikasi pola, tren, dan insight yang tersembunyi di dalam data tersebut. Secara sederhana visualisasi data adalah representasi dari **data-data bisnis** yang sangat **kompleks** dan ribet, yang kemudian diubah ke dalam format **grafik** yang lebih **sederhana**.

Konsep visualisasi data sebenarnya telah ada sejak zaman dahulu. Grafik sederhana seperti diagram batang dan lingkaran telah digunakan sejak abad ke-17. Namun, dengan perkembangan teknologi komputer dan perangkat lunak, visualisasi data mengalami evolusi yang signifikan. Kini, kita dapat membuat visualisasi yang jauh lebih kompleks dan interaktif, serta memanfaatkan berbagai jenis data, mulai dari data numerik hingga data spasial.

Memasuki era Big Data telah mengubah cara kita memandang visualisasi data. Saat ini, volume, kecepatan, dan variasi data yang dihasilkan oleh sistem digital sangat besar dan terus bertambah dengan cepat. Visualisasi data menjadi lebih penting karena manusia memiliki keterbatasan dalam memproses informasi yang disajikan dalam bentuk numerik yang sangat besar. Visualisasi memungkinkan data dalam jumlah besar dapat dipersepsikan secara intuitif, sehingga informasi yang terkandung di dalamnya dapat dipahami dengan lebih cepat dan efisien. Di era Big Data, visualisasi tidak hanya memudahkan eksplorasi data, tetapi juga berfungsi sebagai alat untuk mengkomunikasikan hasil analisis data kepada audiens yang lebih luas.

Pentingnya visualisasi data dalam dunia modern juga tercermin dalam berbagai bidang. Dalam dunia bisnis, misalnya, dashboard visualisasi data memungkinkan para pemimpin bisnis untuk mengakses laporan kinerja secara real-time dan mengambil keputusan strategis berdasarkan data tersebut. Dalam bidang kesehatan, visualisasi data mempermudah para peneliti dan tenaga medis untuk memantau epidemi, memodelkan penyebaran penyakit, dan menganalisis hasil penelitian klinis. Di bidang pendidikan, visualisasi data digunakan untuk mengevaluasi hasil belajar siswa, mengidentifikasi kesenjangan dalam pendidikan, dan membantu perencanaan kebijakan pendidikan yang lebih tepat sasaran.

Selain fungsinya dalam pengambilan keputusan, visualisasi data juga memiliki manfaat dalam meningkatkan pemahaman dan keterlibatan. Data yang divisualisasikan dengan baik dapat mengkomunikasikan informasi dengan cara yang lebih menarik dan mudah dipahami oleh berbagai lapisan audiens. Hal ini sangat penting dalam lingkungan yang membutuhkan komunikasi yang cepat dan tepat, seperti di dunia bisnis dan pemerintahan. Sebuah grafik yang sederhana

namun dirancang dengan baik dapat memberikan gambaran yang lebih jelas dibandingkan dengan halaman penuh angka.

Tidak hanya itu, visualisasi data juga memfasilitasi penemuan pola dan anomali dalam data yang mungkin tidak terlihat jika data hanya dipandang dalam bentuk tabel atau teks. Misalnya, dalam penelitian ilmiah, pola hubungan antara variabel dapat lebih mudah dikenali melalui scatter plot atau heatmap, dibandingkan dengan melihat deretan angka dalam tabel statistik. Pola-pola seperti tren, korelasi, atau outliers dapat langsung terlihat, memungkinkan peneliti atau analis untuk fokus pada bagian yang paling relevan dari data tersebut.

6.2 Prinsip Visualisasi Data

Agar visualisasi data efektif dan mudah dipahami, ada beberapa prinsip dasar yang perlu diikuti. Prinsip-prinsip ini membantu menghindari kesalahan dalam penyajian data dan memastikan bahwa visualisasi dapat mengkomunikasikan pesan dengan jelas.

a. Identifikasi Pembaca

Masalah yang paling sering terjadi dalam visualisasi data adalah ketidakterbacaan grafis data yang disajikan. Untuk itu, kita perlu memastikan beberapa hal sebelum melakukan visualisasi data.

- Pertama, tampilan data disesuaikan dengan job role atau target pembacanya.
- Kedua, pertimbangkan latar belakang pendidikan dan keahlian para pembaca dalam memilih topik visualisasi data.
- Ketiga, perhatikan faktor-faktor seperti budaya dan etnis.
- Terakhir, sesuaikan visualisasi data dengan tingkat literasi pembacanya mengenai technical terms, statistika dan bahasa.

b. Diagram Pertama

Prinsip ini mungkin yang paling tidak teknis tetapi sangat penting. Sebelum membuat visual, prioritaskan informasi apa yang ingin Anda bagikan, bayangkan, dan desain. Fokuslah pada informasi dan pesan terlebih dahulu sebelum menggunakan perangkat lunak yang mungkin membatasi atau mempengaruhi alat visual yang Anda gunakan. Dalam kata lain, jangan terlalu memikirkan geometri (titik, garis) yang akan digunakan nanti, tetapi fokus pada inti informasi yang perlu disampaikan dan apa yang akan mendukung poin Anda. Langkah ini dapat dilakukan secara mental atau dengan menggunakan pena dan kertas untuk kebebasan berpikir yang maksimal.

c. Gunakan Software yang Tepat

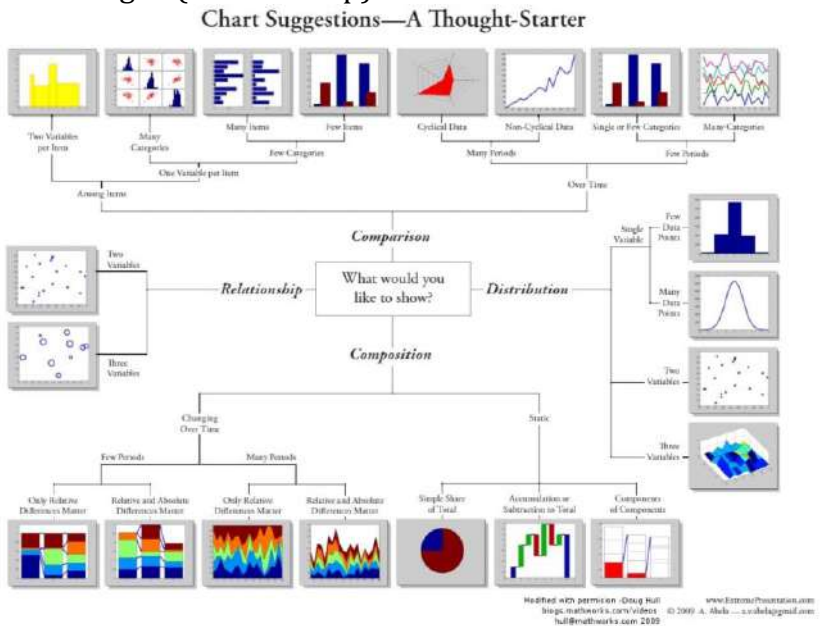
Visual yang efektif umumnya membutuhkan penguasaan software. Mungkin tidak realistis mengharapkan visual yang kompleks, teknis, dan efektif jika Anda hanya menggunakan software sederhana seperti spreadsheet. Anda mungkin perlu mempelajari software baru atau memperluas pengetahuan Anda tentang software yang sudah dikenal seperti Tabelau, Power BI, Looker Studi, dan lain-lain.

d. Gunakan Geometri yang Efektif dan Tampilkan Data

Geometri adalah bentuk dan fitur yang sering dikaitkan dengan jenis visual tertentu. Penting untuk memilih geometri yang sesuai dengan data Anda. Grafik adalah representasi data dalam berbagai bentuk, dan sering kali terdapat lebih dari satu grafik yang perlu dipertimbangkan. Utamakan fungsi dibandingkan bentuk dan hindari penggunaan elemen visual yang tidak diperlukan.

Grafik biasanya dibagi menjadi beberapa kategori berikut:

- Perbandingan (Comparison)
- Komposisi (Composi
- Distribusi (Distribution)
- Hubungan (Relationship)



Gambar 6. 1 Kategori Grafik

e. Warna Selalu Memiliki Makna

Penggunaan warna dalam visualisasi bisa sangat bermanfaat, dan jarang ada alasan untuk tidak menggunakan warna. Visualisasi data yang penuh warna memiliki skor daya ingat yang lebih tinggi dan tujuh warna atau lebih adalah yang terbaik. Warna mewakili informasi, baik secara langsung dan jelas, atau secara tidak langsung dan halus. Contoh langsung penggunaan warna dapat

dilihat pada peta yang perairannya berwarna biru dan daratannya hijau atau coklat. Skema warna bisa berupa skala sekuensial (dari terang ke gelap), divergen (dengan dua skala yang bertemu di tengah), atau kualitatif (untuk membedakan kelompok).

f. Sertakan Ketidakpastian

Ketidakpastian adalah bagian penting dari pemahaman sebagian besar sistem. Visual tanpa ketidakpastian bisa menyesatkan. Ketidakpastian bisa disajikan dengan garis error bar atau area yang diarsir. Cara lain untuk menampilkan ketidakpastian adalah dengan memasukkannya secara tidak langsung ke dalam bentuk visual yang sudah ada, seperti box plot atau distribusi. Atau, kita bisa menambahkannya secara langsung sebagai elemen visual tambahan, misalnya garis error bar atau area yang diarsir.

g. Gunakan Panel Kecil (Small Multiples)

Salah satu cara yang sangat baik untuk melakukan small multiples atau peneling adalah dengan mengulang grafik yang sama beberapa kali untuk menunjukkan perbedaan-perbedaannya. Ide di balik "small multiples" adalah bahwa karena banyak elemen desainnya sama (seperti sumbu, skala sumbu, dan bentuk grafik), perbedaan dalam data akan lebih mudah terlihat. Dengan kata lain, setiap panel mewakili perubahan dalam satu variabel, yang biasanya adalah waktu, kelompok, atau faktor lainnya. Tujuan dari "small multiples" adalah untuk membuat data menjadi mudah dibandingkan, dan "small multiples" yang efektif selalu berhasil mencapai tujuan ini.

h. Data dan Model adalah Hal yang Berbeda

Ada tiga bentuk utama dalam visualisasi data yaitu: data mentah (seperti scatterplot), data yang telah diringkas (seperti box plot), atau statistik inferensial (seperti garis regresi yang telah disesuaikan). Data mentah dan data yang diringkas biasanya cukup mudah dipahami. Namun, model yang diplot mungkin memerlukan penjelasan lebih lanjut agar pembaca dapat sepenuhnya memahami dan mengulang penelitian tersebut. Setiap model dalam suatu studi harus dilaporkan secara lengkap untuk memastikan dapat diulang. Selain itu, setiap visualisasi model harus dijelaskan dalam keterangan gambar atau di bagian lain dokumen, sehingga pembaca dapat menemukan detail lengkap tentang apa yang diwakili oleh visualisasi model tersebut.

i. Visual yang Sederhana dan Informasi yang Detail

Pastikan visual Anda sederhana dan mudah dipahami. Menambahkan informasi yang tidak diinginkan dapat membuatnya membingungkan, sehingga menggagalkan tujuan visualisasi data. Ingatlah bahwa tujuan akhir visualisasi data adalah kesederhanaan. Anda dapat membuat visual tambahan jika ingin menyampaikan cerita yang memiliki banyak sisi

j. Pertimbangkan Infografis

Meskipun kadang sulit untuk membedakan antara grafik dan infografik, biasanya grafik lebih fokus pada penyajian data dan model, sementara infografik menggabungkan teks, gambar, dan berbagai elemen diagram. Meskipun tidak disarankan untuk mengubah semua grafik menjadi infografik, penelitian menunjukkan bahwa infografik lebih mudah diingat. Selain itu, diagram cenderung lebih diingat dibandingkan bentuk grafik seperti

titik, batang, garis, dan tabel. Para ilmuwan bisa meningkatkan cara mereka menyampaikan informasi dengan mempertimbangkan penggunaan infografik yang mengombinasikan berbagai informasi secara efektif. Infografik juga dapat lebih efektif ketika digunakan di luar publikasi yang ditinjau oleh rekan sejawat, seperti dalam presentasi lisan atau poster, di mana visual perlu mencakup lebih banyak elemen penelitian tetapi dengan informasi yang lebih sederhana. Walaupun infografik mungkin tidak digunakan di banyak situasi, visualisasi teknis sering kali bisa lebih baik dengan tambahan teks atau anotasi.

k. Mintalah Pendapat Orang Lain

Visualisasi data yang efektif sangat penting agar pembaca dapat memahami dan terhubung dengan informasi yang disajikan. Meskipun ada pedoman dalam menciptakan visualisasi yang baik, penting untuk mendapatkan umpan balik dari orang lain untuk memastikan efektivitasnya. Penulis disarankan untuk meminta rekan kerja atau pihak luar memeriksa visualisasi data secara terpisah, bukan keseluruhan dokumen, agar mereka dapat memberikan penilaian objektif tanpa dipengaruhi oleh teks. Umpan balik ini akan membantu meningkatkan kualitas visualisasi dan memastikan bahwa pesan yang ingin disampaikan dapat diterima dengan baik oleh pembaca.

6.3 Jenis-Jenis Visualisasi Data

Berikut merupakan jenis-jenis dalam visualisasi data:

a. Visualisasi Data Temporal

Visualisasi data temporal digunakan untuk menggambarkan informasi yang berkaitan dengan waktu. Teknik ini sering diterapkan dalam bentuk **grafik garis** atau **diagram garis**, yang menunjukkan bagaimana suatu variabel berubah seiring berjalannya waktu. Contohnya, diagram garis dapat menampilkan fluktuasi suhu harian selama sebulan atau pertumbuhan populasi tahunan. Dengan menggunakan beberapa garis dalam satu diagram, kita juga dapat membandingkan berbagai faktor atau kategori yang terjadi pada periode yang sama, seperti penjualan produk di berbagai wilayah.

b. Visualisasi Data Hirarkis

Visualisasi data hirarkis digunakan untuk menggambarkan hubungan antara elemen-elemen dalam struktur yang memiliki tingkat atau lapisan berbeda. Salah satu contoh yang umum digunakan adalah **diagram pohon** (tree diagram) atau **sunburst chart**, yang menunjukkan bagaimana kategori utama dibagi menjadi subkategori. Teknik ini sangat berguna untuk menampilkan data yang memiliki struktur berjenjang, seperti organisasi perusahaan, klasifikasi biologis, atau sistem klasifikasi produk. Visualisasi ini membantu pengguna memahami bagaimana bagian-bagian yang lebih kecil berhubungan dengan keseluruhan.

c. Visualisasi Data Jaringan

Visualisasi data jaringan digunakan untuk menggambarkan hubungan antar entitas dalam bentuk **graf**. Setiap entitas diwakili sebagai simpul (*node*), sementara hubungan di antara mereka diwakili sebagai

garis penghubung (*edges*). Teknik ini sering digunakan dalam analisis sosial, seperti menunjukkan koneksi antara individu dalam jejaring sosial, atau dalam analisis jaringan transportasi dan komunikasi. Dengan visualisasi ini, pengguna dapat melihat pola, keterhubungan, dan pengaruh antar elemen dalam jaringan tersebut.

d. Visualisasi Data Multidimensi

Visualisasi data multidimensi digunakan untuk menyajikan informasi dengan lebih dari dua variabel dalam satu visualisasi. Teknik ini sering mencakup penggunaan **scatter plot 3D**, **diagram paralel**, atau **heatmap**. Misalnya, diagram paralel dapat menunjukkan hubungan antara beberapa variabel dalam satu set data, memungkinkan pengguna untuk melihat pola dan tren yang mungkin tidak terlihat dalam visualisasi dua dimensi. Ini berguna dalam analisis data yang kompleks, di mana banyak faktor saling berinteraksi.

e. Visualisasi Data Geospasial

Visualisasi data geospasial berfokus pada penyajian data yang terkait dengan lokasi geografis. Teknik ini sering menggunakan peta untuk menunjukkan distribusi dan pola data berdasarkan lokasi. Contohnya, **peta panas (*heatmap*)** dapat digunakan untuk menunjukkan kepadatan populasi atau insiden kejahatan di suatu area. Visualisasi ini sangat efektif untuk analisis data yang berhubungan dengan geografi, seperti perencanaan kota, analisis lingkungan, atau pengelolaan sumber daya alam. Dengan mengaitkan data dengan lokasi, pengguna dapat memahami konteks dan implikasi yang lebih baik dari informasi yang disajikan.

6.4 Tahapan Visualisasi Data

Berikut tahapan dalam membuat visualisasi data.

a. Tentukan Tujuan Visualisasi

Tahap pertama adalah menentukan tujuan dari visualisasi. Di sini, penting untuk memahami apa yang ingin disampaikan melalui visualisasi dan siapa audiensnya. Tujuan visualisasi dapat bervariasi, seperti memberikan wawasan tentang tren data, membuat perbandingan, atau menyoroti pola tertentu. Dengan memahami tujuan yang jelas, kita dapat memilih pendekatan yang tepat untuk menyajikan data secara efektif.

b. Kumpulkan Data

Setelah tujuan ditentukan, tahap selanjutnya adalah mengumpulkan data yang relevan. Data ini bisa berasal dari berbagai sumber seperti database internal, survei, sensor, atau dataset eksternal. Pastikan data yang dikumpulkan sesuai dengan kebutuhan visualisasi dan cukup untuk mendukung analisis yang akan dilakukan. Data yang lengkap dan akurat sangat penting untuk mendapatkan visualisasi yang bermakna.

c. Bersihkan Data (*Data Cleaning*)

Tahap ini melibatkan proses pembersihan data, di mana data yang kotor, tidak lengkap, atau tidak relevan diidentifikasi dan diperbaiki. Data cleaning termasuk menghilangkan duplikasi, menangani nilai yang hilang, mengoreksi kesalahan dalam data, dan memastikan konsistensi format. Data yang bersih adalah fondasi untuk membuat visualisasi yang akurat dan bermakna.

d. Pilih Visualisasi

Setelah data siap, pilihlah jenis visualisasi yang paling sesuai dengan tujuan dan jenis data yang dimiliki. Ada dua jenis utama visualisasi data.

Visualisasi statis: hanya menyediakan satu tampilan cerita data tertentu. Infografis adalah contoh visualisasi statis.

Visualisasi interaktif: memungkinkan pengguna berinteraksi dengan grafik dan diagram. Pengguna dapat mengubah variabel dalam parameter visualisasi untuk menemukan wawasan baru atau mengakses informasi yang mendalam. Software visualisasi data biasanya menyertakan dasbord untuk interaksi pengguna dengan sistem.

Jenis visualisasi yang dipilih harus mampu menyampaikan informasi dengan jelas dan mudah dipahami. Misalnya, jika ingin menunjukkan perubahan dari waktu ke waktu, grafik garis mungkin paling cocok. Jika ingin memperlihatkan distribusi, histogram mungkin lebih sesuai. Pemilihan yang tepat akan membantu audiens dengan cepat menangkap insight yang ingin disampaikan.

e. Buat (*Create*) Visualisasi

Tahap terakhir adalah membuat visualisasi menggunakan alat atau software visualisasi data seperti Tableau, Power BI, Python (matplotlib, seaborn), atau Excel. Di sini, data yang sudah siap dan jenis visualisasi yang dipilih akan diimplementasikan. Selama pembuatan visualisasi, perhatikan prinsip desain yang baik, seperti pemilihan warna yang tepat, label yang jelas, serta penyederhanaan visual untuk

meminimalkan kebingungan. Hasil akhir harus menyampaikan informasi dengan cara yang menarik dan mudah dipahami oleh audiens.

6.5 Rangkuman

Visualisasi data memegang peran penting dalam menyederhanakan dan mengomunikasikan informasi yang kompleks di era big data. Dengan mengubah data mentah menjadi representasi visual, seperti grafik atau diagram, visualisasi data membantu pengguna dalam mengidentifikasi pola, tren, dan wawasan yang tersembunyi. Evolusi teknologi telah memungkinkan pembuatan visualisasi yang lebih interaktif dan kompleks, mempermudah analisis data besar yang sulit dipahami jika hanya dalam bentuk numerik. Visualisasi data kini digunakan di berbagai bidang, seperti bisnis, kesehatan, dan pendidikan, untuk mendukung pengambilan keputusan yang lebih cepat dan akurat.

Tahapan visualisasi data meliputi penentuan tujuan visualisasi, pengumpulan data yang relevan, dan pembersihan data agar siap untuk dianalisis. Setelah data bersih, langkah berikutnya adalah memilih jenis visualisasi yang sesuai dengan tujuan, misalnya grafik garis untuk perubahan waktu atau histogram untuk distribusi. Akhirnya, visualisasi tersebut dibuat menggunakan alat atau perangkat lunak khusus, dengan mempertimbangkan desain yang efektif, penggunaan warna, serta elemen visual yang mudah dipahami oleh audiens.

DAFTAR PUSTAKA

- Midway, S.R. (2020) 'Principles of Effective Data Visualization', *Patterns*, 1(9), p. 100141. doi:10.1016/j.patter.2020.100141.
- Joshi, P. and Mahalle, P.N. (2023a) Data storytelling and visualization with tableau: A hands-on approach. Boca Raton, FL: CRC Press.
- Hwang, J. and Yoon, Y. (no date) Data Analytics and visualization in quality analysis using tableau. CRC Press.
- Anggie Irfansyah. (2023) Jenis-Jenis visualisasi data Yang Sering Digunakan, Eduparx Blog. Available at: <https://eduparx.id/blog/insight/jenis-jenis-visualisasi-data-yang-sering-digunakan/> (Accessed: 08 October 2024).
- Chart Chooser in Matlab (2020) Extreme Presentation. Available at: <https://extremepresentation.com/2009/04/chart-chooser-in-matlab/> (Accessed: 08 October 2024).

BAB 7

ALAT VISUALISASI DATA

7.1 Pendahuluan

Alat visualisasi data adalah perangkat lunak yang digunakan untuk mengubah data mentah menjadi representasi visual yang lebih mudah dipahami, seperti grafik, diagram, peta, dan dashboard interaktif. Alat ini sangat penting dalam analisis data karena mampu menyajikan data yang kompleks dalam format yang lebih sederhana dan menarik, sehingga mempermudah pengambilan keputusan berdasarkan informasi yang dihasilkan. Salah satu alat visualisasi data yang populer dan banyak digunakan adalah Tabelau.

Tableau didirikan pada tahun 2003 oleh tiga lulusan Universitas Stanford: Christian, Chris, dan Pat. Awalnya, mereka mengalami kesulitan mengoperasikan perangkat lunak Business Intelligence (BI) saat bekerja pada proyek analisis data pada tahun 1999. Meskipun memiliki latar belakang dalam bidang komputer, mereka menyadari perlunya perbaikan dalam perangkat lunak BI. Hal inilah yang mendorong pendirian Tableau dengan filosofi manajemen utama "Membantu orang melihat dan memahami data." Tableau menjadi salah satu perusahaan terkemuka dalam platform analitika, memungkinkan pengguna dari berbagai tingkat keterampilan untuk menggunakan dan bekerja dengan data.

Tableau adalah alat visualisasi data yang kuat, yang mengubah data mentah menjadi format interaktif, intuitif, dan menarik secara visual. Kekuatan Tableau terletak pada kemampuannya untuk terhubung ke berbagai sumber data dan berbagai format data seperti Microsoft Excel, Text File, JSON File, CSV File, Microsoft Access, PDF File, Spatial File, dan Statistical File dan banyak lainnya, melakukan pembaruan secara real-time, serta menawarkan antarmuka seret dan lepas (drag-and-drop) untuk kemudahan penggunaan. Dengan Tableau, pengguna dapat membuat dasbor, diagram, grafik, dan peta yang dinamis, yang memungkinkan mereka untuk dengan cepat menemukan wawasan dan tren. Tableau juga cerdas dalam merekomendasikan jenis visualisasi atau grafik yang paling sesuai dengan data yang kita gunakan. Tableau memiliki beberapa versi seperti Tableau Desktop, Tableau Public, dan lain-lain tergantung pada fungsionalitas yang dibutuhkan. Pada buku ini, kita akan menggunakan Tableau Public yang dapat diakses pada laman <https://public.tableau.com/app/discover> secara gratis dan Anda cukup membuat akun dan login untuk memulai membuat visualisasi. Gambar 7.1 merupakan tampilan workspace Tableau Public.



Gambar 7. 1 Tampilan Workspace Tableau Public

(<https://help.tableau.com/>)

Keterangan setiap istilah:

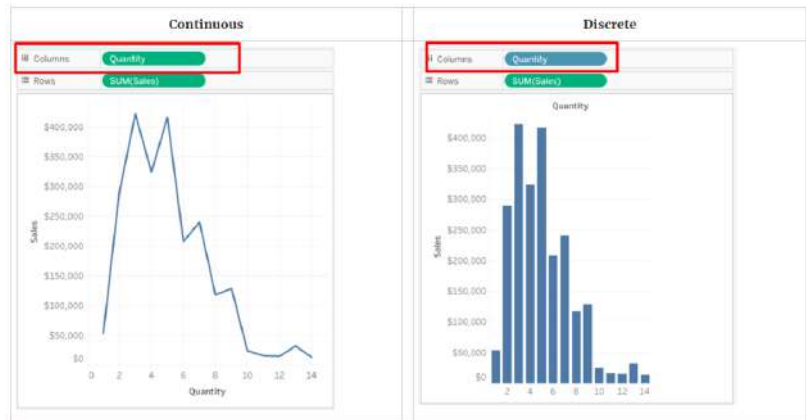
- A. **Nama Workbook** - Sebuah workbook berisi lembar kerja (sheets) seperti Microsoft Excel. Sheet dapat berupa worksheet, dashboard, atau story. Sebuah worksheet berisi satu tampilan (view) yang dilengkapi dengan shelves, cards, legend, serta Data dan Analytics Pane di sidebar-nya. Sebuah dashboard adalah kumpulan tampilan dari beberapa worksheet. Di sidebar-nya, tersedia Dashboard Pane dan Layout Pane. Sebuah story berisi rangkaian worksheet atau dashboard yang bekerja bersama untuk

menyampaikan informasi. Sidebar-nya menyediakan Story Pane dan Layout Pane.

- B. **Cards dan Shelves** - Seret bidang (fields) ke cards dan shelves di ruang kerja untuk menambahkan data ke tampilan Anda.
- C. **Toolbar** - Gunakan toolbar untuk mengakses perintah serta alat analisis dan navigasi.
- D. **View** - adalah kanvas di ruang kerja tempat Anda membuat visualisasi (juga disebut sebagai "viz").
- E. **Icon** - Klik ikon ini untuk menuju halaman Start dan menghubungkan ke data. Untuk informasi lebih lanjut, lihat Start Page.
- F. **Side Bar** - Di worksheet, area sidebar berisi Data Pane dan Analytics Pane.
- G. **Data Source** - Klik tab ini untuk menuju halaman Data Source dan melihat data Anda. Untuk informasi lebih lanjut, lihat Data Source Page.
- H. **Status Bar** - Menampilkan informasi tentang tampilan saat ini.
- I. **Sheet Tabs** - Tab ini mewakili setiap sheet dalam workbook Anda. Ini bisa mencakup worksheet, dashboard, dan story. Untuk informasi lebih lanjut, lihat Workbooks dan Sheets.

7.2 Komponen Fungsional Tableau

Tableau memiliki beberapa komponen fungsional yang sering digunakan seperti terlihat pada Gambar 7.2.



Gambar 7. 2 Ilustrasi Komponen Fungsional Tabelau

a. Measures dan Dimensions

Measure: Nilai kuantitatif yang dapat diagregasi. Misalnya, nilai agregasi measure dapat berupa jumlah (sum), rata-rata, median, jumlah, minimum, maksimum, persentil, deviasi standar, dan varians. Saat menggeser Measure ke lembar kerja Tableau, data secara otomatis diagregasi sebagai jumlah (sum) dalam pengaturan default. Measures biasanya bersifat **continous**.

Dimensions: Nilai kualitatif yang digunakan untuk mendeskripsikan data dengan lebih rinci. Misalnya, nama produk dan tanggal produksi dapat dianggap sebagai Dimensi. Umumnya digunakan untuk mengkategorikan data. Dimensions umumnya bersifat **discrete**.

b. Rows dan Columns

Untuk membuat visualisasi, Anda cukup drag and drop field dari Data Pane. Jika Anda drag sebuah dimensi ke bagian row, akan muncul header untuk anggotanya. Sementara itu, jika Anda menggunakan sebuah measure, itu akan dianggap sebagai parameter kuantitas. Dalam Gambar 7.2, dimension Quantity didrag and drop ke Columns, dan Sales didrag and drop ke Rows (Agregasi diambil secara otomatis.) Tableau kemudian akan secara otomatis menghasilkan bar chart berdasarkan data tersebut.

c. Discrete dan Continous

Discrete: Nilai-nilai bersifat terpisah dan berbeda satu sama lain. Bisa berupa Ukuran maupun Dimensi. Ditampilkan sebagai **Blue pill** di Tableau. Misalnya, peristiwa cacat produk (Ya atau Tidak) atau warna produk (biru dan merah) dapat bersifat Diskrit.

Continuous: Nilai-nilai tidak bersifat terpisah dan dapat dibagi menjadi nilai pecahan dan desimal yang lebih kecil seperti angka 7 diikuti oleh 8 dengan jarak yang sama menuju 9, dan 7,5 berada di tengah antara 7 dan 8. Bisa berupa Measure maupun Dimension. Ditampilkan sebagai **Green pill** di Tableau. Misalnya, berat, tinggi, lebar, dan panjang dapat bersifat Kontinu.

7.3 Jenis-Jenis Grafik Tableau

Tableau menawarkan berbagai jenis chart atau grafik, tetapi bagaimana cara memilih yang paling cocok untuk data Anda? Di Tableau, bentuk visualisasi mengikuti fungsi yang ingin dicapai. Visualisasi (atau viz) yang Anda buat bergantung pada:

- a. Pertanyaan yang ingin Anda ajukan
- b. Karakteristik data Anda
- c. Cara Anda ingin menyajikan dan mengkomunikasikan wawasan kepada orang lain

Sebagai contoh, untuk menunjukkan pertumbuhan penjualan setiap tahun, Anda memerlukan visualisasi yang berbeda dibandingkan dengan menunjukkan hubungan antara item yang didiskon dan profitabilitasnya. Memahami apa yang ingin Anda tampilkan akan membantu Anda menentukan cara penyajiannya. Sebelum menggunakan grafik pada tableau, langkah awal yang perlu dilakukan akan menghubungkan data yang digunakan ke dalam tableau dengan cara Klik Data -> New Data Source, Drag and drop a file atau Upload file from computer. Pada buku ini kita menggunakan sample dataset superstore-sales yang dapat diakses pada <https://community.tableau.com/s/question/0D54T00000CW eX8SAL/samplestore-sales-excelxls>.

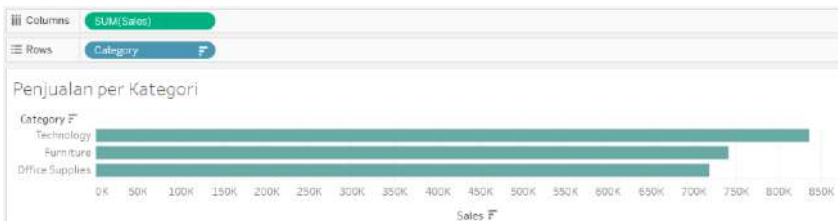
Setelah data berhasil diupload atau dihubungkan ke Tableau, maka kita perlu untuk memasukan tabel yang akan kita gunakan dalam pengelolaan data selanjutnya. Pada tahapan ini, kita akan menggunakan semua tabel yang ada di dataset, maka kita perlu untuk mendrag seluruh tabel seperti terlihat pada Gambar 7.3.



Gambar 7. 3 Menghubungkan Data ke Tableau

a. Bar Chart

Bar Chart adalah grafik yang menggunakan batang vertikal atau horizontal untuk menunjukkan perbandingan antara berbagai kategori. Meskipun bar chart / grafik batang merupakan visualisasi data yang paling sederhana dan umum, namun desainer visualisasi data cenderung menggunakan bar chart secara luas dengan alasan “**Simplicity is key**”. Bar chart dapat digunakan untuk menampilkan total penjualan masing-masing kategori. Untuk menghasilkan grafik tersebut, maka kita menggunakan variabel **Category** dan **Sales**. Variabel Category sebagai Rows dan Sales sebagai Coloums. Gambar 7.4 merupakan grafik yang dihasilkan.



Gambar 7. 4 Grafik Bar Chart Horizontal

Kita dapat menampilkan hasilnya dalam bentuk vertical dengan mengklik menu **swaps rows and coloums** dan hasilnya dapat dilihat pada gambar 7.5.

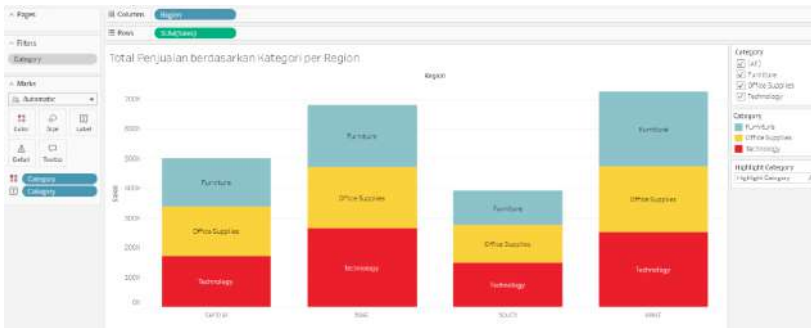


Gambar 7.5 Grafik Bar Chart Vertical

Pada kedua grafik di atas, kita dapat mengetahui bahwa kategori Technology memiliki jumlah penjualan paling banyak dibandingkan dua kategori lainnya.

b. Stacked Bar Chart

Stacked Bar Chart menampilkan nilai-nilai dalam satu batang yang dibagi menjadi beberapa bagian yang memungkinkan analisis komposisi kategori dalam total. Misalnya, kita ingin menunjukkan total penjualan berdasarkan kategori yang dibagi berdasarkan region. Hal ini dapat membantu melihat kontribusi setiap region terhadap total penjualan per kategori. Untuk menghasilkan grafik tersebut, maka kita menggunakan variabel **Category**, **Sales**, dan **Region**. Gambar 7.6 merupakan tampilan stacked bar chart yang dihasilkan.

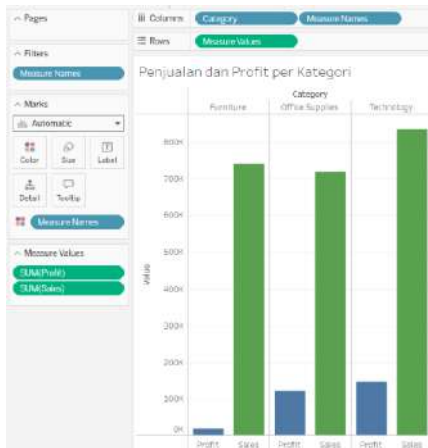


Gambar 7. 6 Grafik Stacked Bar

Pada grafik di atas, kita dapat mengetahui bahwa region west dan east memberikan kontribusi terbesar untuk penjualan masing-masing kategori.

c. Side by Side Bars

Grafik side by side bars menampilkan dua atau lebih batang berdampingan untuk membandingkan nilai kategori yang berbeda dalam satu tampilan. Misalnya kita dapat melihat perbandingan penjualan dan profit untuk kategori produk yang sama. Untuk menampilkan grafik side by side bars, kita dapat menggunakan variable **profit**, **sales**, dan **category** seperti terlihat pada Gambar 7.7 berikut.

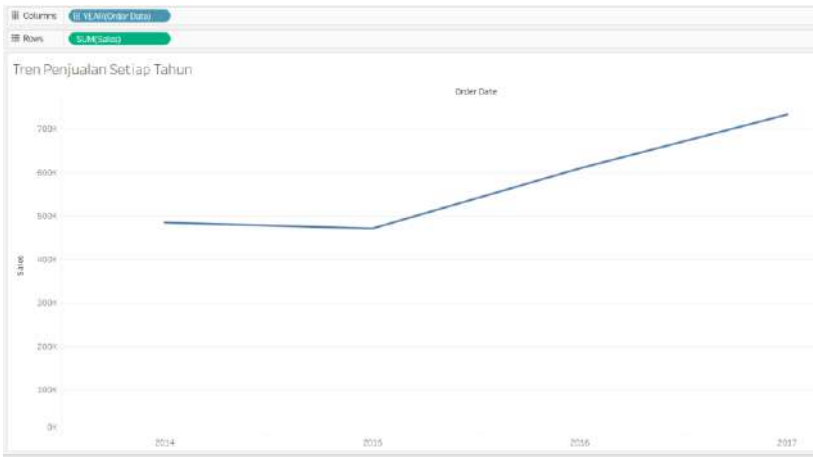


Gambar 7. 7 Garfik Side by Side Bars

Pada grafik di atas, kita dapat mengetahui bahwa kategori Technology memiliki profit dan sales paling banyak dibandingkan dua kategori lainnya.

d. Line Chart

Line Chart menunjukkan tren data sepanjang waktu dengan menghubungkan titik data dengan garis. Grafik garis digunakan dan disarankan ketika kita ingin menggunakan variabel temporal. Misalnya, kita ingin mengetahui tren penjualan selama periode tertentu. Pada latihan kali ini, kita akan menggunakan line chart untuk mengetahui tren penjualan sesuai dengan order date. Variabel yang digunakan adalah **order date** dan **sales** seperti terlihat pada Gambar 7.8.



Gambar 7.8 Grafik Line Chart

Pada grafik di atas, kita mengetahui bahwa penjualan tahun 2014 dan 2015 terjadi penurunan sedikit dan kemudian naik pada tahun 2016 hingga 2017.

e. Dual Line Chart

Grafik ini memungkinkan perbandingan dua tren data dalam satu tampilan dengan dua garis berbeda. Misalkan kita ingin menampilkan perbandingan penjualan dan profit masing-masing kategori per quarter, maka kita dapat menggunakan variabel **sales**, **order data**, **profit** dan **quarter** seperti terlihat pada gambar 7.9.

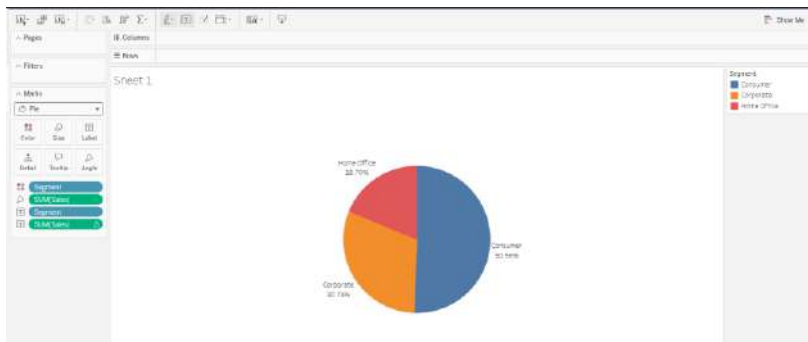


Gambar 7. 9 Grafik Dual Line Chart

Pada grafik di atas, kita mengetahui penjualan dan profit setiap kategori per quarter.

f. Pie Chart

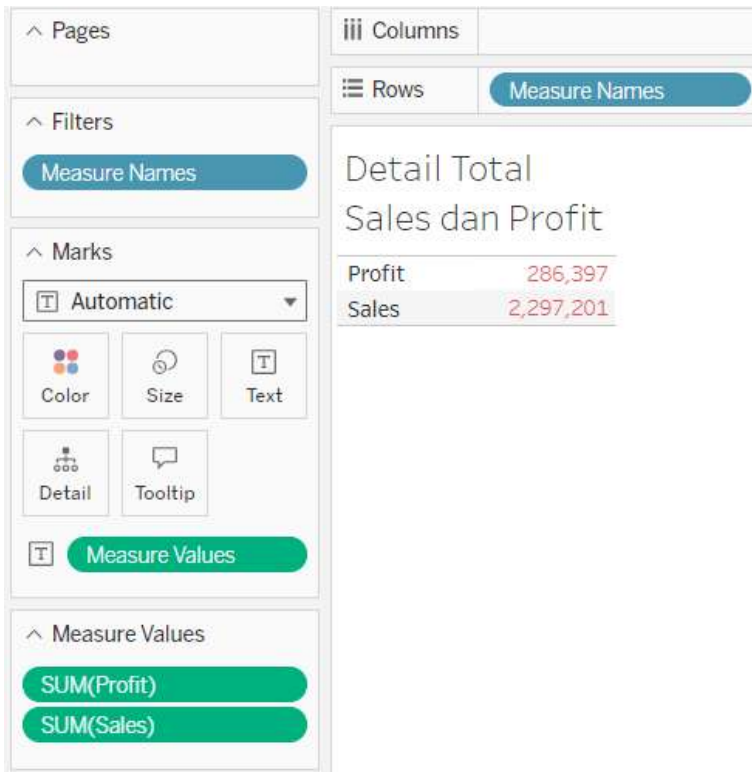
Pie Chart adalah jenis diagram lingkaran yang menunjukkan proporsi data kategorikal. Pie Chart menggambarkan proporsi dari keseluruhan dengan membagi lingkaran menjadi beberapa irisan. Misalnya, kita ingin menampilkan persentasi penjualan masing-masing segmen, maka kita dapat menggunakan variabel **Segment** dan **Sum** seperti pada Gambar 7.10.



Gambar 7. 10 Grafik Pie Chart

g. Text Tables

Text Tables menyajikan data dalam format tabel yang lebih mudah dibaca, sering digunakan untuk menampilkan nilai yang tepat. Misalnya kita ingin menampilkan detail total penjualan dan keuntungan, maka kita dapat menggunakan variabel **Profit** dan **Sales** seperti pada Gambar 7.11.

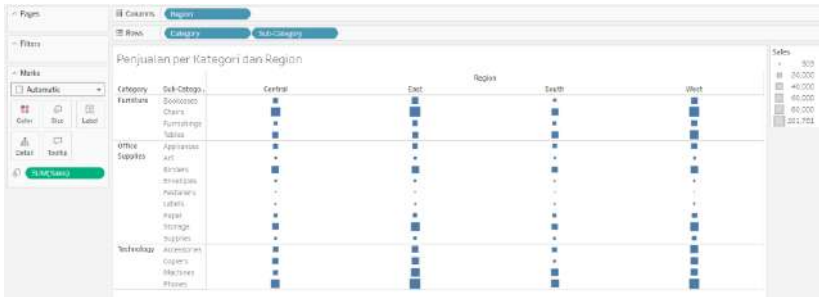


Gambar 7. 11 Grafik Text Tables

h. Heat Maps

Heat Maps menggunakan warna untuk menunjukkan nilai data dalam matriks. Pada latihan kali ini, kita ingin menampilkan penjualan berdasarkan kategori dan region. Misalnya, Anda dapat melihat mana yang paling sukses di

setiap region, maka kita dapat menggunakan variabel **category**, **region**, dan **sales** seperti terlihat pada Gambar 7.12.



Gambar 7. 12 Grafik Heat Maps

Pada grafik heat maps di atas, semakin besar kotaknya, semakin tinggi nilainya.

i. Highlight Tables

Highlight Tables adalah tabel dengan warna untuk menonjolkan nilai tertentu, membuat analisis lebih cepat dan mudah. Misalnya kita ingin menyajikan profit per kategori dan bulan dengan warna untuk menunjukkan mana yang paling menguntungkan di waktu tertentu. Kita dapat menggunakan variabel **profit** dan **order date** seperti pada Gambar 7.13 berikut.



Gambar 7. 13 Grafik Highlight Tables

Pada grafik di atas, semakin gelap warna biru menunjukkan bahwa profit semakin tinggi dan warna orange menunjukkan penurunan profit.

j. Tree Maps

Tree Maps menunjukkan hierarki data menggunakan persegi panjang yang memiliki ukuran dan warna berbeda. Tree Maps cocok digunakan untuk menunjukkan perbandingan proporsi data di dalam hierarki. Namun, Tree Maps tidak cocok digunakan untuk melihat tingkat hierarki antara kategori terbesar dan setiap titik data. Pada contoh kali ini kita akan melihat perbandingan penjualan 10 produk terbanyak. Variabel yang digunakan adalah **Product Name** dan **Sales**. Drag & drop variabel Product Name ke Bagian rows, kemudian drag & drop variabel sales ke bagian marks -> Text. Untuk menampilkan 10 produk dengan penjualan terbanyak, drag & drop variabel **Product Name** ke bagian filter, kemudian pada pop up filter setting seperti tampilan berikut.



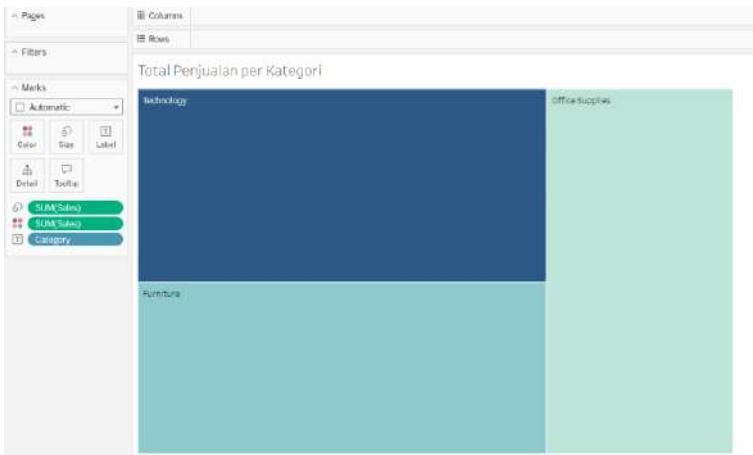
Gambar 7. 14 Filter 10 Produk

Pada Bagian Show Me, pilih Tree Maps. Gambar 7.15 adalah hasil tree maps untuk 10 produk dengan penjualan terbanyak.



Gambar 7. 15 Grafik Tree Maps

Contoh lainnya, misal kita ingin menampilkan total penjualan masing-masing kategori. Untuk menghasilkan grafik tersebut, maka kita menggunakan variabel **Category** dan **Sales** seperti terlihat pada Gambar 7.16.

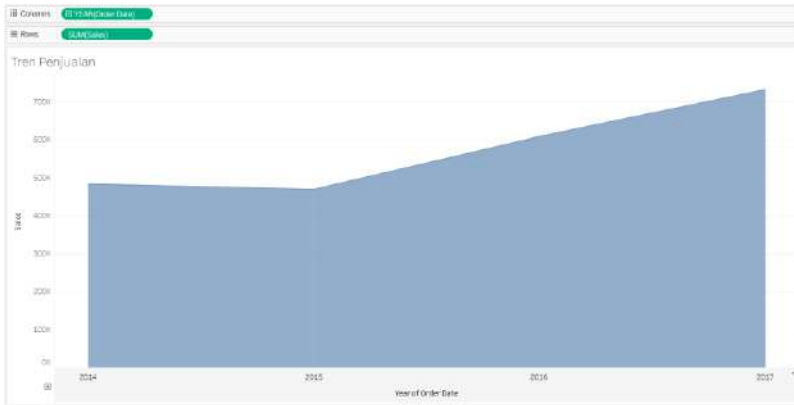


Gambar 7. 16 Grafik Tree Maps

k. Area Chart

Area Chart mirip dengan Line Chart tetapi mengisi area di bawah garis untuk menunjukkan volume data. Misalnya, kita ingin mengetahui tren penjualan selama

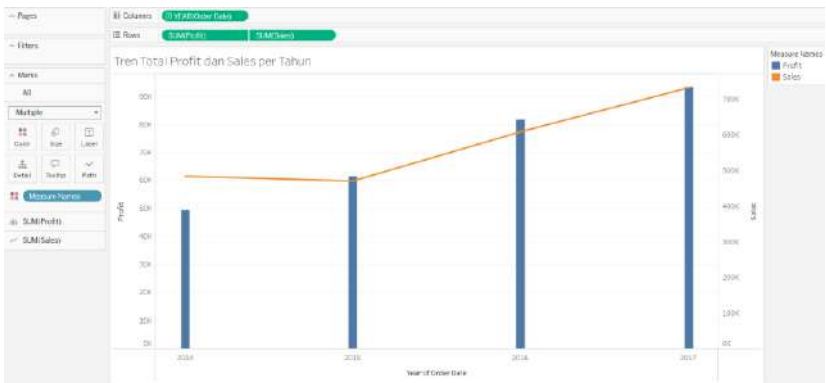
periode tertentu. Pada latihan kali ini, kita akan menggunakan area chart untuk mengetahui tren penjualan sesuai dengan order date. Variabel yang digunakan adalah **order date** dan **sales** seperti terlihat pada Gambar 7.17.



Gambar 7. 17 Grafik Area Chart

I. Dual Combination

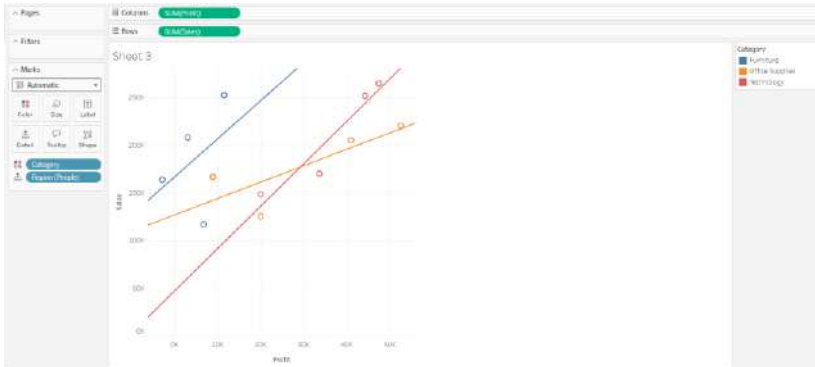
Grafik ini menggabungkan dua jenis grafik, seperti batang dan garis, dalam satu visualisasi untuk perbandingan yang lebih efektif. Misalnya kita ingin menampilkan penjualan (garis) dan tren margin keuntungan (batang) pada waktu yang sama untuk analisis yang lebih komprehensif. Kita dapat menggunakan variabel **Profit**, **Sales**, dan **Order Date** seperti terlihat pada Gambar 7.18.



Gambar 7. 18 Grafik Dual Combination

m.Scatter Plots

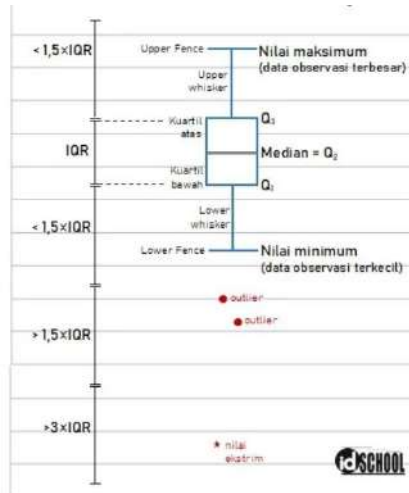
Scatter Plots menampilkan nilai dua variabel dengan titik untuk melihat pola atau hubungan antara keduanya. Scatter plot adalah jenis grafik yang digunakan untuk menampilkan hubungan antara dua variabel numerik. Fungsi scatter plot mencakup beberapa aspek utama yang dapat membantu menganalisis dan memahami korelasi atau pola yang mungkin ada di antara pasangan variabel tersebut. Pada contoh kali ini, kita akan melihat hubungan antara variabel Profit dan Sales untuk setiap category dan region. Pada bagian analytics, pilih Trend Line -> Linear. Hasil Scatter Plot untuk variabel Profit dan Sales dapat dilihat pada Gambar 7.19.



Gambar 7. 19 Grafik Scatter Plots

n. Box and Wishker Plots

Box and Whisker plot adalah jenis grafik yang membantu kita memahami distribusi data. Box and Whisker Plots menunjukkan distribusi data berdasarkan lima angka ringkasan (minimum, kuartil pertama, median, kuartil ketiga, maksimum). Gambar 7.20 merupakan cara membaca box plot.



Gambar 7. 20 Cara Membaca Box Plot

Pada contoh kali ini, kita akan melihat penyebaran jumlah diskon untuk setiap segmen dan region. Drag and drop variabel diskon ke Rows, Segmen dan Region ke Columns. Pada menu Analysis -> Aggregate Measures. Gambar 7.21 merupakan tampilan hasil Box Plots.



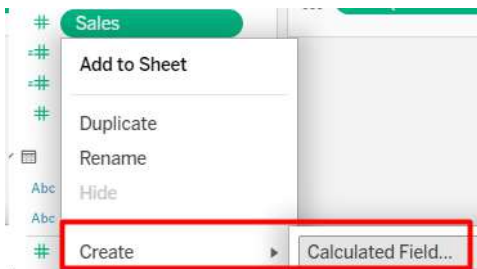
Gambar 7. 21Grafik Box Plot

o. Bullet Graphs

Bullet Graphs adalah jenis visualisasi data yang dirancang khusus untuk menyajikan dan membandingkan kinerja suatu ukuran dengan tujuan atau target tertentu. Grafik ini umumnya digunakan untuk menggantikan indikator atau meter dasbord yang lebih tradisional. Fungsi utama dari bullet graphs adalah memberikan gambaran yang jelas dan efektif tentang sejauh mana suatu kinerja

mencapai tujuan atau target yang ditetapkan. Contoh ini digunakan untuk menampilkan target penjualan masing-masing category pada tahun 2014 dan 2015.

Pertama: Klik kanan sales, kemudian lakukan langkah berikut ini untuk membuat variabel baru penjualan tahun 2015.



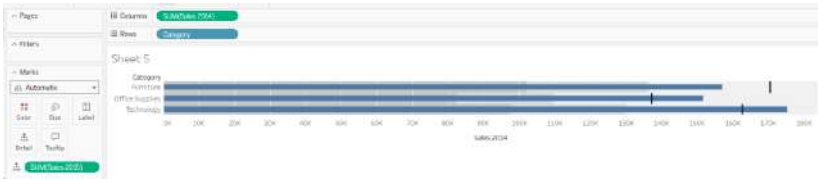
Kemudian lakukan tahapan berikut.



Lakukan tahapan yang sama untuk membuat variabel Sales 2014. Setelah selesai dibagian kiri akan muncul dua variabel yang telah dicreate sebelumnya.



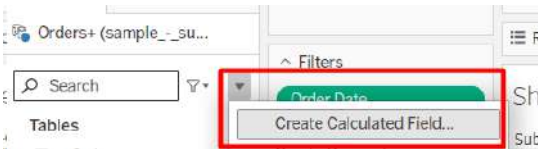
Kemudian pilih variabel **Sales 2014** dan **Sales 2015** lalu klik **Show Me** pilih **Bullet Graps**. Drag & Drop variabel **Category** ke Bagian rows. Gambar 7.22 merupakan, bullet graphs yang telah dibuat.



Gambar 7. 22 Grafik Bullet Graphs

p. Gantt Chart

Gantt chart, adalah jenis grafik yang membantu kita melihat jadwal kegiatan atau tugas proyek secara visual. Gantt chart memungkinkan kita menampilkan durasi kegiatan atau tugas dengan mudah. Pada contoh kali ini kita akan membuat sebuah gantt chart untuk menampilkan durasi pengiriman 3 bulan sejak 1/1/2014 sampai 31/3/2013. Pertama create Calculated Field



Kemudian buat variabel baru **OrderUntilShip** seperti berikut lalu apply dan Ok.



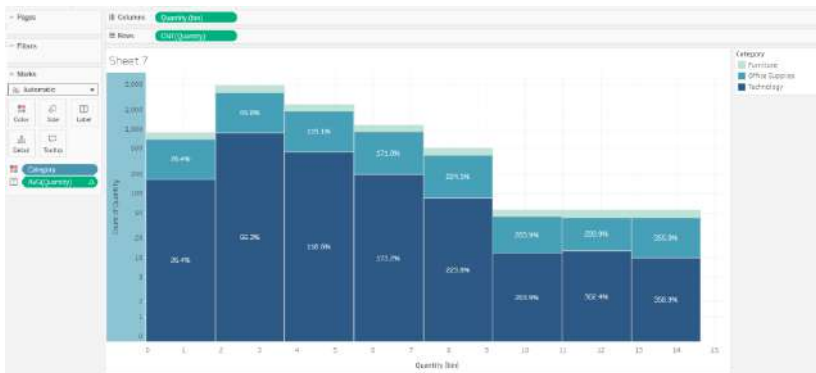
Pada Bagian filter, ubah order date dari tanggal 1/1/2014 sampai 31/3/2014. Gambar 7.23 merupakan tampilan dari Gantt Chart.



Gambar 7. 23 Grafik Gantt Chart

q. Histogram

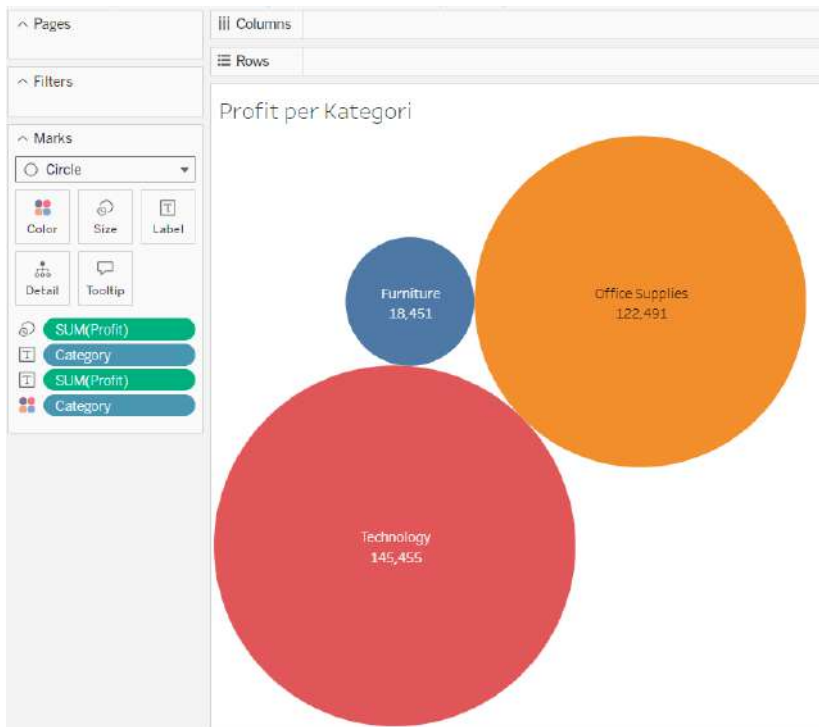
Histogram berfungsi untuk menunjukkan sebaran frekuensi nilai dalam data numerik. Dengan membagi data ke dalam interval atau "bin," grafik ini memberikan gambaran tentang seberapa sering nilai-nilai tertentu muncul. Kita akan membuat histogram dari jumlah quantity masing-masing **category**. Gambar 7.24 merupakan histogram yang telah dihasilkan.



Gambar 7. 24 Grafik Histogram

r. Packed Bubble

Grafik Packed Bubble adalah jenis visualisasi yang menampilkan data dalam bentuk lingkaran yang terisi, di mana ukuran setiap lingkaran mencerminkan nilai kuantitatif dari data yang ditampilkan. Lingkaran tersebut saling berdekatan dan tidak tumpang tindih, menciptakan tampilan yang menarik dan mudah dipahami. Grafik ini efektif untuk menunjukkan hubungan antara variabel, terutama saat membandingkan kategori yang berbeda. Misalnya kita ingin menampilkan total profit per kategori, maka kita dapat menggunakan packed bubble seperti Gambar 7.25.



Gambar 7.25 Grafik Packed Bubbles

7.4 Rangkuman

Alat visualisasi data adalah perangkat lunak yang mengubah data mentah menjadi representasi visual seperti grafik dan dashboard interaktif, yang memudahkan pemahaman dan pengambilan keputusan. Tableau, salah satu alat visualisasi yang populer, didirikan pada tahun 2003 oleh tiga lulusan Universitas Stanford yang ingin memperbaiki perangkat lunak Business Intelligence (BI) yang ada. Dengan filosofi "Membantu orang melihat dan memahami data," Tableau memungkinkan pengguna dari berbagai latar belakang untuk mengakses dan menganalisis data secara efektif. Kemampuan Tableau untuk terhubung ke berbagai sumber data dan format, serta antarmuka seret dan lepas, menjadikannya alat yang sangat berguna dalam analisis data.

Tableau menawarkan beragam jenis grafik untuk visualisasi, dan pemilihan bentuk grafik sangat bergantung pada tujuan analisis dan karakteristik data. Beberapa jenis grafik yang disediakan antara lain Bar Chart, Line Chart, dan Pie Chart, yang masing-masing memiliki fungsi dan cara penyajian data yang berbeda. Dalam menggunakan Tableau, penting untuk memahami komponen fungsional seperti Measures, Dimensions, Rows, dan Columns, yang membantu dalam membuat visualisasi yang efektif. Dengan memanfaatkan berbagai jenis grafik dan fungsionalitas Tableau, pengguna dapat menyampaikan informasi secara jelas dan menarik, serta mendukung proses pengambilan keputusan berbasis data.

DAFTAR PUSTAKA

- Joshi, P. and Mahalle, P.N. 2023. Data storytelling and visualization with tableau: A hands-on approach. Boca Raton, FL: CRC Press.
- Hwang, J. and Yoon, Y. 2020. Data Analytics and visualization in quality analysis using tableau. CRC Press.
- Tableau Community Forums. Available at: <https://community.tableau.com/s/question/0D54T00000CWeX8SAL/sample-superstore-sales-excelxls> (Accessed: 09 October 2024).
- Start page Tableau. Available at: https://help.tableau.com/current/pro/desktop/en-us/environment_startpage.htm (Accessed: 09 October 2024).

BAB 8

DASHBOARD dan STORY

8.1 Pendahuluan

Dashboard merupakan tampilan yang menggabungkan berbagai visualisasi (charts, graphs, maps) ke dalam satu layar atau antarmuka. Hal ini memungkinkan pengguna untuk melihat, menganalisis, dan membandingkan berbagai aspek dari data secara bersamaan. Setiap elemen dalam dashboard dapat saling berinteraksi, memungkinkan pengguna untuk mengeksplorasi data lebih dalam dengan mudah. Misalnya, ketika salah satu visualisasi diperbarui, elemen lain dalam dashboard juga dapat diperbarui secara otomatis. Dashboard berguna untuk menyajikan laporan yang komprehensif dan memberikan gambaran yang lebih jelas dan menyeluruh mengenai data yang sedang dianalisis.

Selain dashboard, tableau juga memiliki **story** yang digunakan sebagai kumpulan dashboard atau visualisasi (chart, graphs, maps, dan lain-lain) yang diorganisasikan untuk menyampaikan narasi atau cerita berbasis data. Dengan Story, pengguna dapat membawa audiens mereka melalui alur cerita yang logis dan terstruktur, membantu menjelaskan hubungan antar data, mengungkapkan tren atau anomali, serta mendukung insight yang didapat dari analisis data. Setiap langkah dalam Story bisa difokuskan pada poin tertentu, yang memungkinkan audiens mengikuti perjalanan data dari satu kesimpulan ke kesimpulan lainnya.

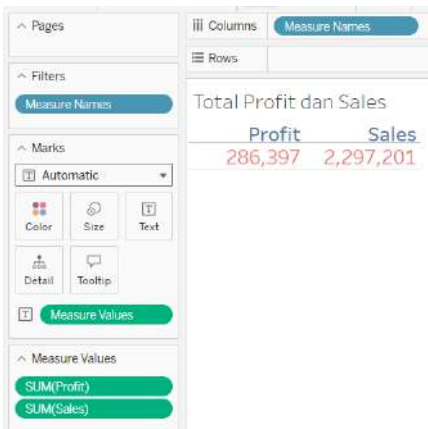
Dashboard dan story dapat membantu memudahkan pengambilan keputusan berbasis data. Dashboard menyajikan data dalam bentuk yang lengkap dan interaktif, sementara story menyusun data menjadi sebuah alur cerita yang jelas, sehingga memudahkan penyampaian informasi yang ingin disampaikan dari data tersebut.

8.2 Dashboard

Dashboard adalah tampilan visual yang digunakan untuk menyajikan data secara komprehensif dan ringkas dalam satu layar. Pada buku kali ini kita akan membuat dashboard penjualan menggunakan dataset **sample superstore sales**. Pada saat pembuatan dashboard, kita perlu membuat grafik-grafik pada sheet yang terpisah kemudian seluruh grafik digabungkan dalam dashboard sebagai berikut.

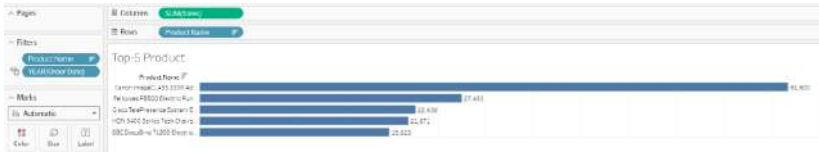
a. Membuat Tabel total Sales dan Profit

Untuk membuat tabel ini, maka kita dapat menggunakan variabel **sales** dan **profit** dan grafik text tables seperti berikut.



b. Membuat Grafik Top-5 Penjualan Product

Untuk membuat grafik ini, maka kita dapat menggunakan variabel **Sales**, **Product Name**, **Order Date** dan grafik Horizontal bar seperti berikut.



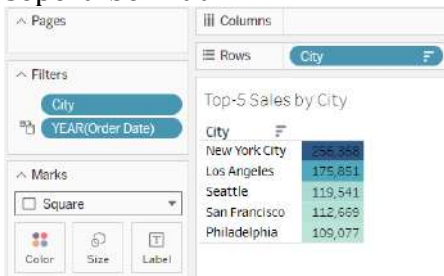
c. Membuat Grafik Top-5 Customer

Untuk membuat grafik ini, maka kita dapat menggunakan variabel **Sales**, **Customer Name**, **Order Date** dan grafik Horizontal bar seperti berikut.



d. Membuat Grafik Top-5 Sales by City

Untuk membuat grafik ini, maka kita dapat menggunakan variabel **Sales**, **City**, **Order Date** dan grafik text tables seperti berikut.



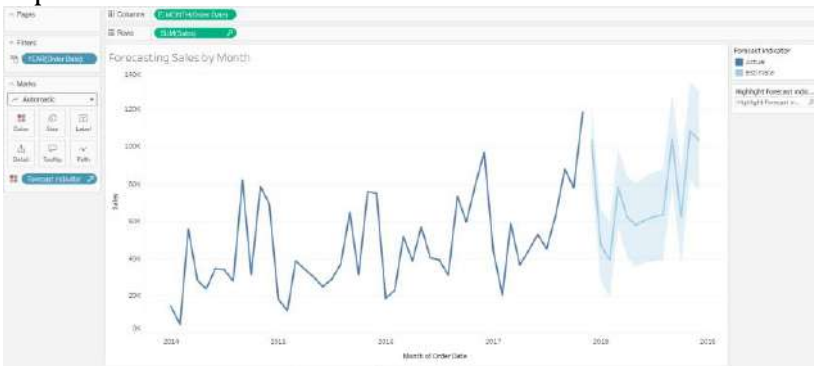
e. Membuat Grafik Perbandingan Profit dan Sales

Untuk membuat grafik ini, maka kita dapat menggunakan variabel **Sales**, **Profit**, **Order Date** dan grafik dual combinations seperti berikut.



f. Membuat Grafik Forecasting Sales

Untuk membuat grafik ini, maka kita dapat menggunakan variabel **Sales**, **Order Date** dan grafik lines continous seperti berikut.

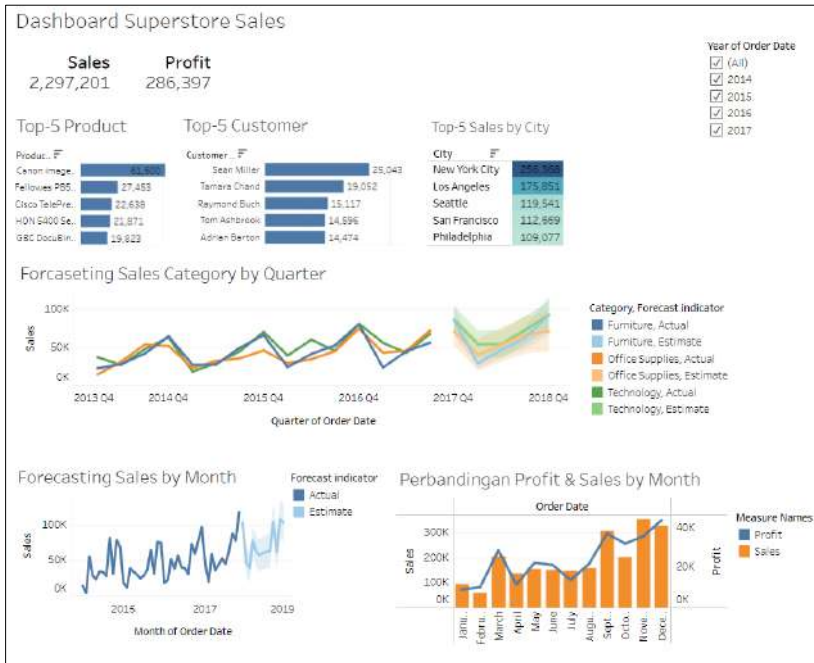


g. Membuat Grafik Forecasting Sales per Kategori

Untuk membuat grafik ini, maka kita dapat menggunakan variabel **Sales**, **Order Date** dan grafik lines continous seperti berikut.



Setelah membuat 7 grafik tersebut untuk menghubungkan semua grafik agar menampilkan informasi yang lebih interaktif dan memudahkan proses analisis data, maka tahap selanjutnya adalah menggabungkan seluruh grafik masing-masing sheet ke dalam tab **dashboard** pada tableau sehingga diperoleh hasil seperti berikut.



Gambar 8. 1 Dashboard Penjualan Superstore

Dashboard di atas dapat difilter berdasarkan tahun untuk melihat informasi-informasi yang lebih detail.

8.3 Story

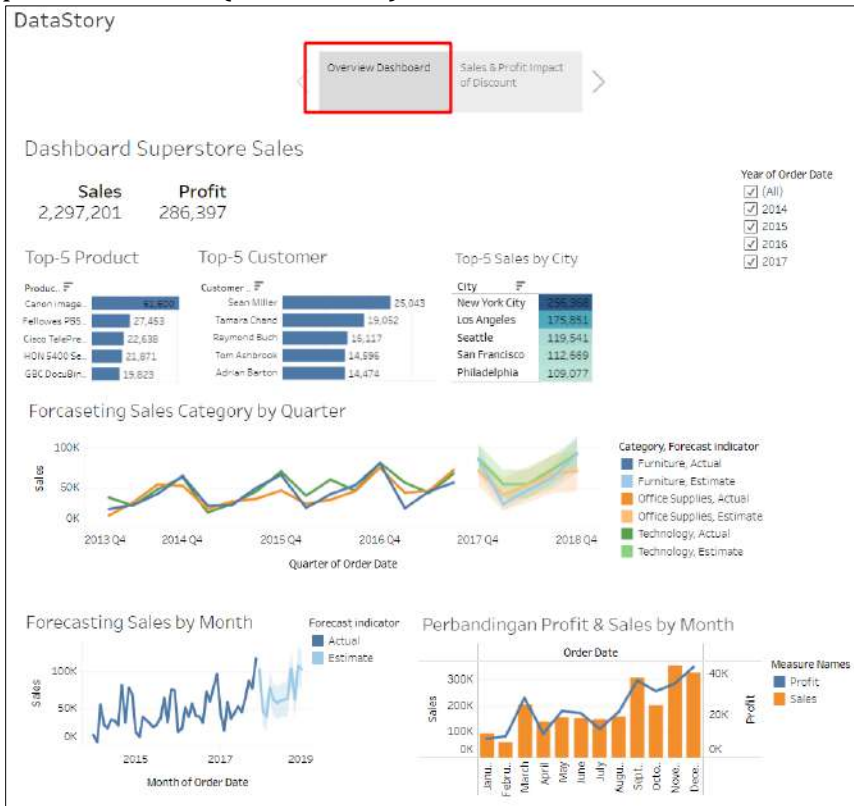
Story adalah cara untuk menyusun dan menyajikan narasi atau alur cerita berdasarkan visualisasi data. Hal ini memungkinkan pengguna untuk membuat narasi yang selaras dan menjelaskan temuan atau wawasan yang ditemukan dari data. Story memiliki beberapa fungsi sebagai berikut:

- Komunikasi Naratif:** Story membantu dalam menyusun narasi logis dan bersifat naratif yang memandu pemirsa melalui serangkaian visualisasi untuk menyampaikan pesan atau temuan tertentu.
- Alur Analisis:** Story memberikan struktur yang jelas untuk memandu pengguna melalui tahap-tahap analisis data,

mulai dari pengenalan masalah hingga penyelesaian atau temuan akhir.

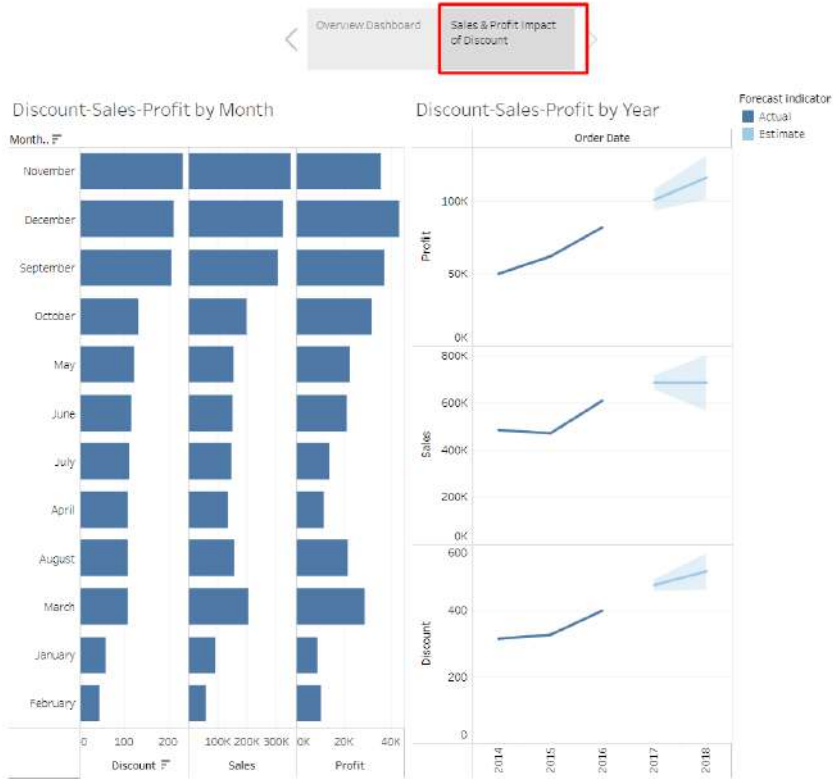
- c. **Presentasi:** Story dapat digunakan untuk membuat presentasi interaktif yang memungkinkan pemirsa untuk mengikuti pemikiran analisis data dengan baik.

Pada latihan kali ini, kita akan membuat story berdasarkan seluruh grafik dan dashboard yang telah dibuat sebelumnya pada sub-bab dashboard (Gambar 8.2) dan insight yang berisi informasi tentang perbandingan dampak diskon terhadap profit dan sales (Gambar 8.3).



Gambar 8. 2 Story Overview Penjualan

DataStory



Gambar 8. 3 Story Perbandingan Discount-Sales-Profit

Dalam sebuah story, kita dapat membuat lebih banyak cerita sesuai dengan kebutuhan.

8.4 Rangkuman

Dashboard dan story adalah dua fitur penting dalam visualisasi data yang memungkinkan pengguna menganalisis serta menyampaikan informasi secara efektif. Dashboard menggabungkan berbagai visualisasi seperti grafik dan tabel dalam satu tampilan interaktif, memungkinkan pengguna untuk mengeksplorasi dan membandingkan data secara komprehensif. Di sisi lain, story membantu menyusun narasi

berdasarkan visualisasi tersebut, sehingga audiens dapat mengikuti alur analisis data secara logis dan memahami wawasan yang disampaikan.

Pada contoh praktisnya, dashboard penjualan dibangun menggunakan dataset Superstore Sales, melibatkan berbagai grafik seperti total sales dan profit, top-5 penjualan produk, dan forecasting sales. Story kemudian menyusun grafik-grafik ini menjadi narasi yang fokus pada temuan, seperti perbandingan dampak diskon terhadap profit dan sales. Kombinasi antara dashboard dan story ini memudahkan penyajian data yang interaktif serta mendukung proses pengambilan keputusan yang lebih informatif.

DAFTAR PUSTAKA

- Joshi, P. and Mahalle, P.N. 2023. Data storytelling and visualization with tableau: A hands-on approach. Boca Raton, FL: CRC Press.
- Hwang, J. and Yoon, Y. 2020. Data Analytics and visualization in quality analysis using tableau. CRC Press.
- Tableau Community Forums. Available at: <https://community.tableau.com/s/question/0D54T00000CWeX8SAL/sample-superstore-sales-excelxls> (Accessed: 09 October 2024).
- Start page Tableau. Available at: https://help.tableau.com/current/pro/desktop/en-us/environment_startpage.htm (Accessed: 09 October 2024).
- Dewi, Kartika Sari. 2012. Kesehatan Mental, Semarang: UPT UNDIP Press.

BIODATA PENULIS



Wulan Sri Lestari, S.Kom., M.Kom.

Dosen Program Studi Teknologi Informasi
Fakultas Informatika Universitas Mikroskil

Penulis lahir di Tanjung Merawa tanggal 31 Maret 1993. Penulis adalah dosen tetap pada Program Studi Pendidikan Teknologi Informasi Fakultas Informatika, Universitas Mikroskil. Menyelesaikan pendidikan S1 pada Jurusan Teknik Informatika dan melanjutkan S2 pada Jurusan Teknologi Informasi. Penulis telah menekuni bidang penelitian terkait Keamanan Data, Deep Learning, dan Kecerdasan Artifisial serta telah menulis beberapa buku dengan judul Dasa-Dasar Pemrograman, Pengantar Basis Data, dan Artificial Intelligence: Teori dan Penerapan AI di Berbagai Bidang. Penulis dapat dihubungi melalui e-mail: wulan.lestari@mikroskil.ac.id

BIODATA PENULIS



Mustika Ulina, S.Kom., M.Kom.

Dosen Program Studi Teknologi Informasi
Fakultas Informatika Universitas Mikroskil

Seorang penulis dan dosen tetap Prodi Teknologi Informasi Fakultas Informatika pada Universitas Mikroskil Medan. Lahir di Medan, 28 September 1994. Pendidikan program Sarjana (S1) STMIK Mikroskil Jurusan Teknik Informatika dan menyelesaikan program Pasca Sarjana (S2) di STMIK Mikroskil dengan jurusan Teknologi Informasi. Penulis telah menjadi dosen Fakultas informatika selama 5 tahun. Beberapa judul penelitian yang telah dilakukan oleh penulis dalam bidang machine learning yaitu Foreign exchange prediction using CEEMDAN and improved FALSTM, Deteksi Serangan DDoS Menggunakan Deep Q-Network, dan Optimizing Deep Neural Networks Using ANOVA for Web Phishing Detection. Penulis juga telah menulis buku lainnya seperti Buku Ajar Pengantar Basis Data dan Artificial Intelligence: Teori dan Penerapan AI di Berbagai Bidang. Penulis dapat dihubungi melalui e-mail: mustika.ulina@mikroskil.ac.id

BIODATA PENULIS



Manto Lumban Gaol, M.Pd

Kepala Bagian Pengembangan Keunggulan Pembelajaran
Universitas Mikroskil

Penulis lahir di Dolok Sanggul, Sumatera Utara. Penulis menyelesaikan pendidikan S1 dan S2 pada Program Studi Pendidikan Matematika. Penulis telah memiliki pengalaman mengampu mata kuliah yang berkaitan dengan Matematika dan Statistika seperti: Kalkulus, Matematika Diskrit, Metode Numerik, Statistika dan Probabilitas, dan Analisis Multivariat. Penulis dapat dihubungi melalui e-mail: manto.lumbangaol@mikroskil.ac.id

SINOPSIS BUKU

Buku "Analisis dan Visualisasi Data" merupakan buku yang dirancang untuk memberikan pemahaman mendalam tentang cara mengelola, menganalisis, dan memvisualisasikan data dalam konteks yang relevan. Buku ini menggabungkan teori dan praktik dengan pendekatan yang terstruktur untuk membantu pembaca dari berbagai latar belakang memahami dan menerapkan konsep-konsep kunci dalam analisis data. Dimulai dengan literasi data untuk membangun fondasi yang kuat, buku ini melanjutkan dengan eksplorasi Big Data, statistika, dan teknik analitika data yang canggih. Selanjutnya, buku ini mengajarkan seni Data Storytelling untuk menyampaikan informasi secara efektif, diikuti oleh prinsip-prinsip visualisasi data dan penggunaan alat-alat modern. Pada bagian akhir, buku ini menjelaskan pembuatan dashboard untuk memantau data secara real-time, memberikan pembaca keterampilan praktis dan teori yang diperlukan untuk membuat keputusan berbasis data yang informasional dan strategis.