



Topik

1. Ekplorasi Data

Tujuan

- Mahasiswa mampu memahami konsep eksplorasi data
- Mahasiswa mampu memahami jenis data dan statistic data
- Mahasiswa mampu mengidentifikasi tren, pola, outlier
- Mahasiswa mampu menguasai teknik dasar eksplorasi manual

Eksplorasi Data

Eksplorasi data (data exploration) adalah tahap awal dalam analisis data untuk memahami struktur, kualitas, dan pola dari data sebelum dilakukan pemodelan atau analisis lanjutan. Eksplorasi data adalah tahap awal dalam analisis data yang bertujuan untuk:

- Memahami struktur dan karakteristik data.
- Menemukan pola, tren, dan outlier.
- Menyiapkan data untuk analisis lebih lanjut atau pemodelan.

Dalam konteks Business Intelligence, eksplorasi data menjadi fondasi pengambilan keputusan berbasis data (data-driven decision making).

Pengertian Pengambilan Keputusan Bisnis

Pengambilan keputusan Bisnis merupakan serangkaian cara atau proses pengambilan keputusan yang logis, sistematis, dan terstruktur dalam upaya mengidentifikasi faktor-faktor penting dalam bisnis untuk memperoleh keputusan terbaik pencapaian visi dan misi suatu entitas bisnis. Pengambilan keputusan bisnis yang baik harus didasarkan pada data-data yang dapat diukur serta memiliki parameter yang jelas dan dapat dipertanggungjawabkan/ akuntabel.

Kerangka Kerja Pengambilan Keputusan Bisnis

Untuk mendapatkan keputusan bisnis yang tepat dan berguna, diperlukan suatu tahapan yang menjadi pedoman dalam pengambilan keputusan bisnis perusahaan. Terdapat 6 tahap pengambilan keputusan bisnis dengan skema sebagai berikut:



1. Perumusan Masalah

Pada tahap ini, seorang *business intelligence analyst* dituntut mampu memetakan permasalahan yang sedang dihadapi dan menemukan permasalahan utama yang harus diselesaikan oleh pengambil keputusan bisnis. Kondisi yang seringkali terjadi adalah adanya permasalahan minor yang dianggap sebagai permasalahan utama dan dapat menimbulkan bias pada saat pengambilan keputusan bisnis tersebut. Seorang *business intelligence analyst* dituntut memiliki *curious* yang tinggi terhadap

sesuatu untuk terus menggali kebenaran informasi. Selain itu seorang *business intelligence analyst* dituntut memiliki pemikiran yang terbuka (open minded), dimana mampu menanggapi informasi dari berbagai pihak yang dapat berkaitan satu sama lain. Seorang bisnis intelligence analyst diharapkan mampu memilah informasi penting dan tidak, menemukan tingkat korelasi atau keterhubungan data satu dengan lainnya dan menentukan sebatas mana informasi dapat digunakan. Selain itu, seorang *business intelligence analyst* harus berfokus pada detail yang sangat dimungkinkan sebagai permasalahan pokok (root cause) yang harus diselesaikan. Satu hal terakhir yang harus dipikirkan adalah kemungkinan solusi yang dapat diberikan oleh seorang *business intelligence analyst* dalam memecahkan permasalahan yang terjadi. Sebagai seorang internal consultant seorang business intelligence analyst harus dapat berdiskusi dengan baik bersama stakeholder guna memutuskan dan merumuskan saran pengambilan keputusan bisnis berdasarkan data data yang dimiliki sebelumnya.

2. Pemutakhiran Data (Data Cleaning)

Dalam proses pengambilan keputusan bisnis, bagian paling penting setelah penentuan permasalahan adalah proses pemutakhiran dan pembersihan data (Data Cleaning). Proses pemutakhiran data meliputi pemilahan jenis data yang diperlukan, banyaknya data yang dibutuhkan, rentang waktu, bentuk data, dan proses filterisasi atau pembersihan data yang dapat menimbulkan bias pada saat dilakukan visualisasi data. Sebagai contoh adalah jika diperlukan data penjualan pada hari kerja dalam satu tahun, diperlukan proses pembersihan dengan melihat hari libur lain selain ketentuan libur pada hari minggu. Proses pembersihan data inilah yang menjadi kunci dan jaminan bahwa analisis yang digunakan dalam pengambilan keputusan bisnis benar dan dapat dipertanggungjawabkan. Kualitas data yang harus dipenuhi oleh seorang *business intelligence analyst* harus memenuhi minimal 4 syarat kualitas data yakni Objective, Representatif, Uptodate, dan dapat dipercaya. Kualitas data Objective adalah data tersebut harus menggambarkan keadaan nyata dan disajikan secara wajar atau tidak dilakukan manipulasi sebelumnya. Kualitas data Representative adalah data tersebut memiliki kecukupan baik dalam jumlah, jenis, dan macam data sehingga dapat memberikan gambaran yang komprehensif dan utuh serta mewakili populasi jika menggunakan data sampling. Kualitas data Uptodate maksudnya adalah data tersebut harus menggambarkan informasi terkini dan terbaru, sehingga keputusan digunakan sesuai dengan masa dan waktunya. Kualitas data dapat dipercaya maksudnya ialah data tersebut harus bersumber dari pihak atau orang yang credible dan dapat dipercaya sehingga dapat memberikan jaminan keandalan data.

3. Visualisasi Data

Proses visualisasi data adalah tahapan penyajian data yang telah diolah agar dapat dengan mudah dipahami dan dimengerti oleh pengambil keputusan bisnis. Data tersebut dapat berbentuk diagram, trend, excel, sebaran, maupun data data lain yang pada intinya memberikan gambaran isi dari data mentah (raw data) yang telah dilakukan cleaning maupun pengolahan sesuai dengan kebutuhan. Visualisasi data dalam hal pengambilan keputusan bisnis harus memiliki relevansi dengan permasalahan yang akan dipecahkan dan diputuskan oleh pengambil keputusan. Sebagai contoh analisis berbentuk trend dapat digunakan untuk melihat tingkat penjualan suatu produk tertentu.

4. Hipotesis Umum

Hipotesis Umum adalah simpulan yang dihasilkan oleh seorang *business intelligence analyst* yang berupa kesimpulan umum, korelasi data, hubungan kondisi aktual dengan strategi perusahaan yang sudah dijalankan sebelumnya. Pada tahap ini disimpulkan jawaban atas permasalahan pada perusahaan atau entitas bisnis untuk dapat diputuskan secara baik dan benar dalam suatu keputusan bisnis. Hipotesis umum harus memuat permasalahan, jawaban, alasan kejadian, serta penjelasan kausalitas secara akuntabel dan terkorelasi.

5. Analisis

Analisis teknis adalah analisis lanjutan yang dilakukan oleh seorang *business intelligence analyst* yang berisi konfirmasi lanjutan kepada stakeholder yang berkaitan dan tahapan perumusan saran yang akan diberikan sebagai tindak lanjut atas hasil analisis data yang dilakukan oleh seorang *business intelligence analyst*. Dalam melakukan analisis teknis dan detil data dapat digunakan beberapa metode analisis berikut:

- a. Analisis deskriptif: Adalah salah satu metode analisis yang paling sering digunakan sebagai analisis keputusan bisnis yang berisi rangkuman penggambaran kondisi atau keadaan tertentu.
- b. Analisis Eksplanatory: adalah analisis yang digunakan sebagai upaya untuk menyelidiki fenomena tertentu, sebagai contoh fenomena pasar yang sedang berkembang.
- c. Analisis Infrensial merupakan analisis yang digunakan untuk mengeneralisasikan data sample pada keadaan populasi secara umum
- d. Analisis Prediktif analisis yang menggunakan data saat ini dan data historis untuk memprediksi kondisi masa depan (digunakan oleh data scientist)
- e. Analisis Preskriptif adalah metode analisis yang menggabungkan beberapa data dan analisis lain untuk pengambilan keputusan masa datang (digunakan oleh data scientist)
- f. Analisis Causal adalah analisis yang digunakan untuk menilai ada tidaknya korelasi suatu kejadian tertentu.

6. Pengkomunikasian hasil analisis

Pada tahapan ini seorang *business intelligence analyst* dituntut mampu menyajikan sebuah laporan yang dapat memenuhi keinginan stakeholder dalam hal ini pengambil keputusan bisnis secara ringkas, jelas, akuntable dan applicable. Laporan harus memiliki isi berupa simpulan, penjabaran temuan, saran perbaikan maupun keputusan terbaik yang harus diambil oleh pimpinan perusahaan atau entitas bisnis dalam membuat suatu keputusan bisnis.

Jenis Data dalam Kecerdasan Bisnis

Berdasarkan sumber: internal (transaksi, laporan keuangan) dan eksternal (pasar, demografi).

Berdasarkan struktur: terstruktur (SQL, Excel), semi-terstruktur (JSON, XML), tidak terstruktur (teks, video, gambar).

Berdasarkan skala pengukuran:

- Nominal (kategori tanpa urutan: jenis kelamin, kota)
- Ordinal (kategori dengan urutan: kepuasan rendah–tinggi)
- Interval (angka dengan jarak sama, nol tidak mutlak: suhu)
- Rasio (angka dengan nol mutlak: pendapatan, jumlah barang)

Data Descriptive Statistics

Teknik deskriptif digunakan untuk meringkas data sebelum analisis lebih lanjut:

- Ukuran pemusatan: mean, median, mode.
- Ukuran penyebaran: range, variance, standard deviation.
- Distribusi: normal, skewness, kurtosis.
- Frekuensi & proporsi: tabel frekuensi, persentase.

Ukuran Pemusatan

Beberapa ukuran pemusatan data dapat digunakan untuk melakukan analisis statistik antara lain:

1. Mean
Perhitungan untuk menghitung nilai rata-rata pada data.
2. Median
Perhitungan untuk menghitung nilai tengah pada data.
3. Modus

Perhitungan untuk mencari nilai yang paling sering muncul pada distribusi data.

Ukuran Penyebaran

Penyajian data statistik dalam berbagai bentuk tabel dan grafik, masih belum bisa membuat data menjadi “berbicara”. Untuk dapat membuat data berbicara dan bermakna, dapat menggunakan ukuran pemusatan data, yaitu mean, median, dan modus. Akan tetapi cara tersebut dirasa masih belum cukup tajam dan teliti. ukuran pemusatan data (mean, median, modus) tidaklah cukup untuk menggambarkan sekumpulan data dan bahkan bisa menyesatkan. Untuk menghindari hal ini diperlukan informasi lain berupa variasi atau penyebaran data.

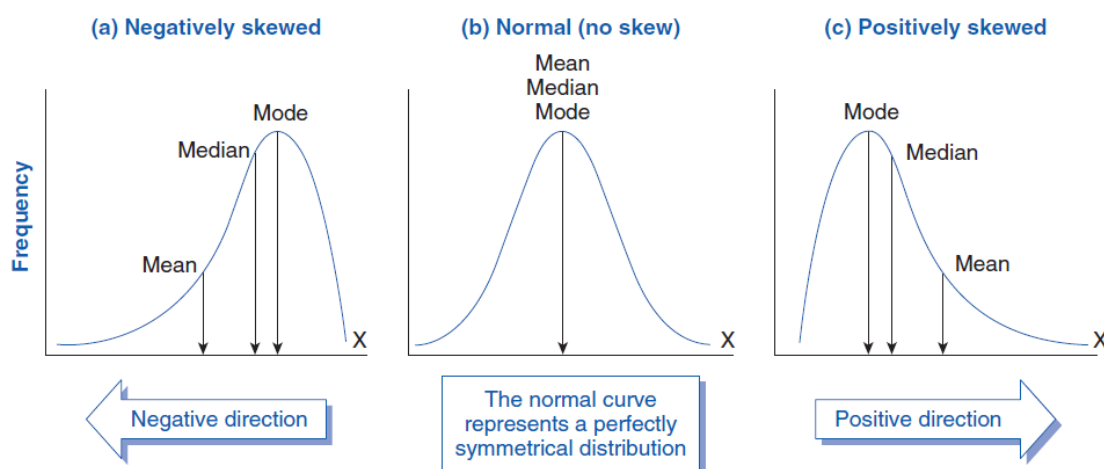
Contohnya pada kasus pengambilan keputusan sekelompok tentara yang menyeberang sungai. Terdapat pilihan untuk menyeberang langsung saat itu juga tanpa kapal atau menunggu bantuan kapal untuk menyeberang. Pemimpin tentara mendapatkan informasi rata-rata kedalaman sungai adalah 1.4m. Diketahui bahwa tentara-tentara tersebut tidak ada yang tingginya dibawah 1.4m. Berdasar informasi tersebut, keputusannya adalah menyeberang sungai tanpa kapal. Apa yang terjadi? Ada beberapa tentara yang tidak bisa berenang akan kesulitan menyeberang karena rata-rata kedalaman sungainya yang 1,4m bukan keseluruhan sungai. Padahal bisa jadi sebagian kedalamannya 2m yang mana dapat membuat sebagian tentara yang tidak bisa berenang akan tenggelam. Sehingga diperlukan pengetahuan mengenai distribusi atau sebaran data untuk menyajikan informasi yang lebih akurat. Distribusi atau penyebaran data adalah suatu fungsi yang menunjukkan semua nilai (atau interval nilai) dari sebuah data dan seberapa sering nilai tersebut terjadi.

Beberapa ukuran penyebaran data yang dapat digunakan dalam analisis kecerdasan bisnis adalah:

- Range (selisih nilai max-min)
contoh: penjualan tertinggi Rp 10 juta, terendah Rp 2 juta → range = 8 juta.
- Standar Deviasi (seberapa besar data menyebar)
SD kecil → data transaksi mirip-mirip, SD besar → bervariasi
- Variance (ragam sebaran data dalam satuan kuadrat)

Distribusi

Kemencengan atau kecondongan (skewness) adalah tingkat ketidaksimetrisan atau kejauhan simetri dari sebuah distribusi. Skewness diartikan sebagai kemiringan distribusi data. Sebuah distribusi yang tidak simetris akan memiliki rata-rata, median, dan modus yang tidak sama besarnya sehingga distribusi akan terkonsentrasi pada salah satu sisi dan kurvanya akan menceng.



Gambar 2 Kecondongan kurva distribusi

No Skew artinya model distribusi yang simetrik atau distribusi data normal. Pada jenis ini nilai mean, median, dan modus adalah sama. Contoh data dengan distribusi normal adalah data rata-rata tinggi badan orang Indonesia. Positif skewness artinya model distribusi yang lebih condong ke arah kanan (arah positif). Pada jenis ini nilai mean lebih besar dari nilai median, dan nilai median lebih besar dari nilai modus. Contoh data pada positif skewness adalah penghasilan orang Indonesia, orang yang berpenghasilan tinggi semakin sedikit. Negatif skewness artinya model distribusi yang lebih condong ke arah kiri (arah negative). Pada jenis ini nilai mean lebih kecil dibanding nilai median, nilai median lebih kecil dari nilai modus. Contoh data dengan sebaran negative skewness adalah kebutuhan kalori usia bayi sampai usia anak-anak.

Frekuensi dan Proporsi

Teknik menyajikan analisis data dalam kecerdasan bisnis lainnya adalah berdasarkan frekuensi atau proporsinya. Frekuensi adalah jumlah kemunculan suatu kategori/nilai dalam sebuah dataset. Sedangkan perbandingan antara frekuensi suatu kategori dengan total keseluruhan data. Contoh:

Data asal kota pelanggan (10 orang):

[Jakarta, Bandung, Jakarta, Surabaya, Jakarta, Bandung, Medan, Jakarta, Surabaya, Bandung]

Berdasarkan data tersebut, maka berikut ini adalah contoh tabel frekuensi:

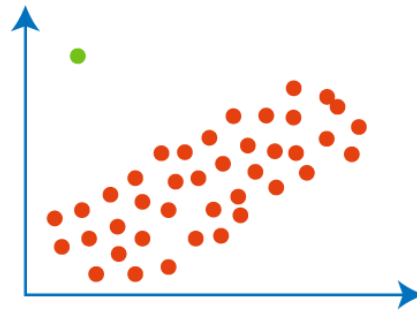
Kota	Frekuensi
Jakarta	4
Bandung	3
Surabaya	2
Medan	1

Berdasarkan data dan tabel frekuensi di atas, berikut ini merupakan tabel proporsi:

Kota	Frekuensi	Proporsi (%)
Jakarta	4	40%
Bandung	3	30%
Surabaya	2	20%
Medan	1	10%

Identifikasi Pola, Trend, dan Outlier

Trend menunjukkan perkembangan data dari waktu ke waktu (time series). Tujuan analisis terhadap trend adalah mengetahui arah perubahan (naik, turun, musiman, stabil). Dari trend tersebut dapat diambil pola-pola data. Pola dalam analisis data menunjukkan keterkaitan antar variabel yang berulang. Tujuannya untuk menemukan hubungan antar faktor. Outlier atau pencilan adalah data yang menyimpang secara ekstrim dari sekumpulan data. Contohnya: Rata-rata nilai dikelas antara 70-100, ada 2 mahasiswa nilainya 5. Gambar 3 menunjukkan ilustrasi outlier.



Gambar 3 Outlier

Adanya outlier, membuat analisis berdasarkan satu ukuran (misal ukuran pemusatan saja) tidak dianjurkan. Sebagai contoh, apabila menarik Kesimpulan berdasarkan hanya nilai mean (rata-rata) saja, maka kesimpulan yang diambil berkemungkinan salah. Karena nilai Mean sangat dipengaruhi oleh validitas data. Nilai Mean dapat berubah jika ada Outlier pada data. Perhatikan tabel berikut ini:

Tabel Angka pendapatan perbulan penduduk Jakarta dan Bandung

	Jakarta	Bandung
	Rp1.000.000,00	Rp1.000.000,00
	Rp2.000.000,00	Rp2.000.000,00
	Rp3.000.000,00	Rp3.000.000,00
	Rp4.000.000,00	Rp4.000.000,00
	Rp5.000.000,00	Rp5.000.000,00
	Rp16.000.000,00	
Mean	Rp5.166.666,67	Rp3.000.000,00
Median	Rp3.500.000,00	Rp3.000.000,00

Angka pendapatan perbulan setiap penduduk di kota Jakarta dan Bandung pada data sample tersebut bernilai sama. Seharusnya nilai Mean pada kedua data juga sama. Namun karena terdapat Outlier, yang ditunjukkan oleh highlight warna abu-abu, nilai Mean kedua data menjadi sangat berbeda.

Mayoritas kejadian outlier disebabkan karena kesalahan input data (human error), kesalahan pengukuran, eksperimental error, kesengajaan (memasukkan data dummy agar tidak terdeteksi), kesalahan pemrosesan data (data manipulation), sampling error, atau data memang menyimpang jauh secara natural.

Korelasi dan Kausalitas

Korelasi adalah pola hubungan antara dua hal atau variabel yang muncul bersamaan dalam satu waktu. Korelasi adalah hubungan statistik antar variabel, ketika satu variabel berubah, variabel yang lain juga berubah. Tipe correlation antara lain:

1. positive correlation

Korelasi positif adalah perubahan nilai yang diikuti pada nilai variabel secara teratur pada arah yang sama, sehingga mengalami kenaikan. Pada korelasi positif variabel X mengalami penambahan nilai maka akan diikuti penambahan nilai variabel Y. Contoh: Apabila dilakukan penambahan jumlah pupuk (X), maka produksi jagung akan semakin meningkat (Y).

2. negative correlation

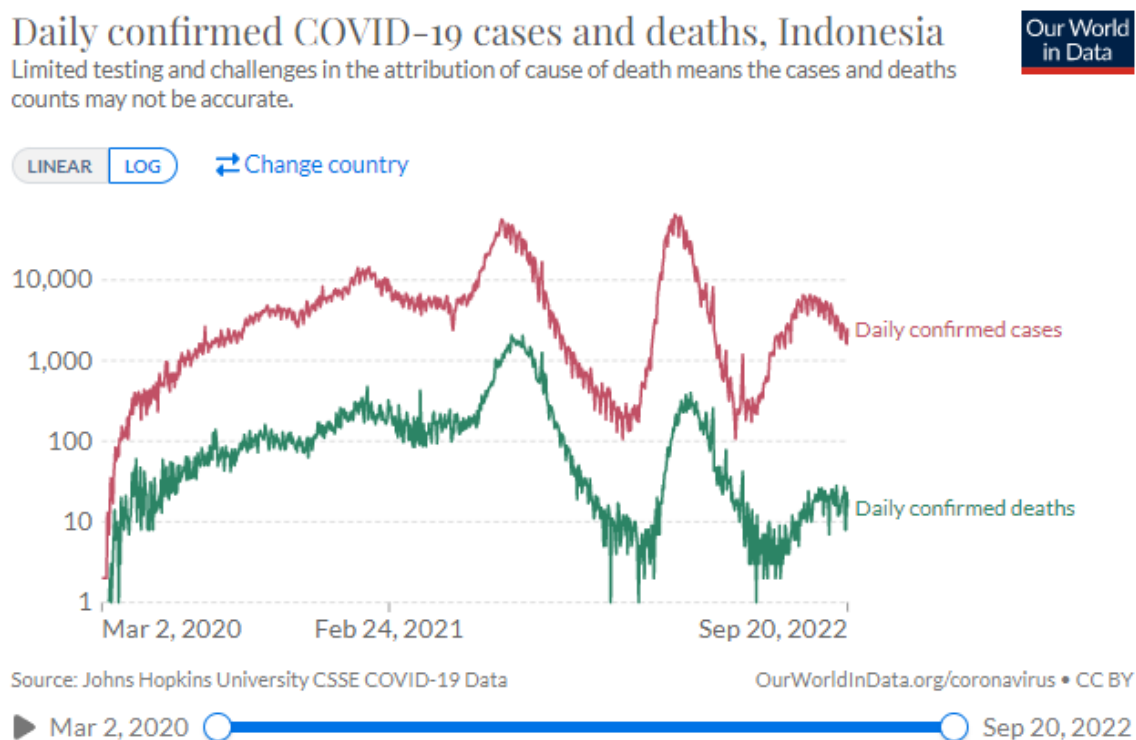
Korelasi negatif adalah perubahan nilai yang diikuti pada nilai variabel secara teratur, tetapi memiliki arah yang berlawanan dengan kenaikan variabel yang tidak teratur. korelasi negatif ini berbanding

terbalik dengan korelasi positif. Korelasi negatif adalah hubungan antara variabel X dan variabel Y. Jika nilai variabel X meningkat, maka nilai variabel Y akan menurun. Contoh: Jika harga barang elektronik naik (X) maka permintaan terhadap barang tersebut akan menurun (Y).

3. no correlation

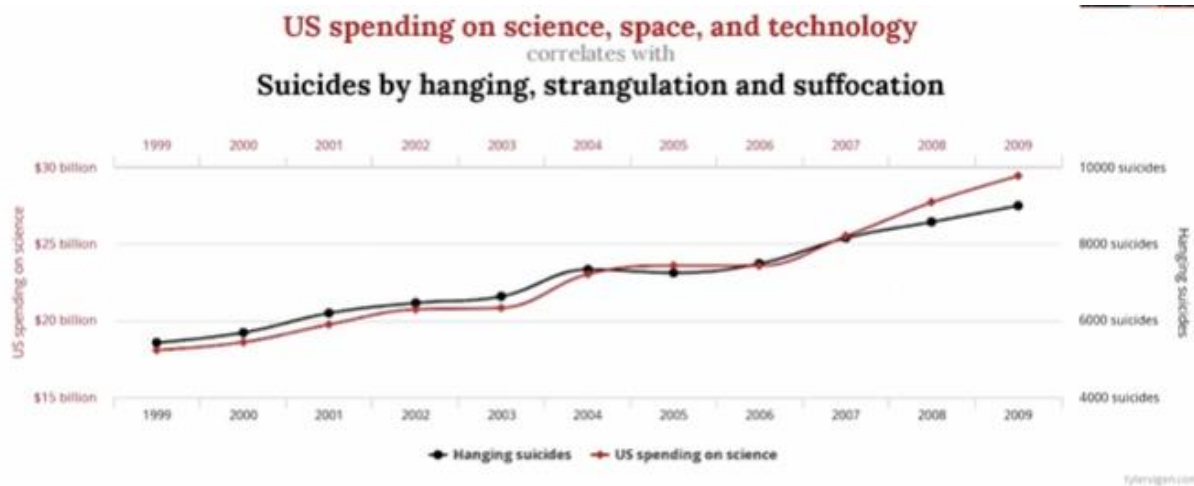
variabel yang tidak berkorelasi terjadi bila kenaikan nilai diikuti dengan penurunan data yang berlawanan atau tidak saling berhubungan. Dalam bentuk korelasi ini, nilai koefisien memiliki pasangan data dengan korelasi yang lemah. Hasil analisis korelasi dari dua variabel (X dan Y) tidak menunjukkan adanya hubungan linear. Contoh: Kenaikan harga BBM (X) dan kenaikan nilai siswa (Y)

Kausalitas adalah hubungan sebab akibat antara dua variabel. Dalam hubungan kausal ada variabel independen (penyebab) dan variabel dependen (akibat). Jika terdapat dua variabel yang berkorelasi, tidak harus terdapat hubungan kausalitas. contoh korelasi. Setiap hujan deras turun, kecelakaan lalu lintas juga terjadi. Artinya, dalam satu waktu, ada dua variabel yang muncul. Tetapi musim hujan bukan penyebab kecelakaan lalu lintas, atau keduanya tidak memiliki hubungan kausalitas. Kecelakaan lalu lintas sering terjadi di musim hujan karena jalanan licin, padamnya lampu jalan, orang yang tetap ngebut saat hujan, atau lubang di jalan yang tak tampak karena tergenang air. Sedangkan contoh yang menunjukkan adanya korelasi yang memiliki hubungan kausalitas adalah ditunjukkan pada gambar 5.



Gambar 5 Hubungan korelasi dan kausalitas

Grafik pada Gambar 5 menunjukkan adanya hubungan korelasi dan kausalitas. Grafik tersebut memaparkan fluktuasi konfirmasi kasus covid-19 dan konfirmasi kematian yang disebabkan oleh covid-19 di Indonesia dari kurun waktu 2020-2022. Gambar 6 menunjukkan gambaran hubungan korelasi, tetapi tidak memiliki hubungan kausalitas.



Gambar 6 Hubungan Korelasi tetapi tidak kausalitas

Terlihat pada grafik yang ditunjukkan pada Gambar 6 mengenai hubungan kemajuan IPTEK di US dan kasus bunuh diri. Pada tahun 1999 – 2004 terlihat bahwa trend kemajuan IPTEK dan angka bunuh diri memiliki korelasi positif, pada kurun waktu 2004 – 2006 data menunjukkan tidak adanya korelasi, Selanjutnya pada tahun 2006 – 2009 data menunjukkan adanya korelasi positif. Meskipun demikian kedua kejadian tersebut tidak serta merta bisa dikaitkan sebagai hubungan kausalitas. Karena memang kasus kematian akibat bunuh di US tidak ada kaitan sebab akibat dengan kemajuan IPTEK.

Bagian 1: Memahami konsep eksplorasi data

Jawablah beberapa pertanyaan berikut ini:

1. Jelaskan konsep eksplorasi data!
2. Menurut pendapat Anda apa itu pengambilan keputusan bisnis?
3. Jelaskan bagaimana tahapan pengambilan keputusan bisnis?

Jawaban:

- a. Eksplorasi data adalah tahap awal analisis untuk memahami struktur, kualitas, dan pola pada data: cek tipe kolom, nilai hilang/aneh, ringkas statistik, lihat sebaran, tren, dan outlier. Tujuannya menyiapkan data agar siap dianalisis/ dipakai modeling dan menghasilkan wawasan yang valid.
- b. Serangkaian proses logis, sistematis, dan terukur untuk memilih tindakan terbaik demi mencapai tujuan bisnis. Keputusan yang baik berbasis data (terukur, akuntabel), bukan asumsi.
- c. Tahapan pengambilan keputusan bisnis (ringkas & runtut)
 1. Rumusan masalah (root cause, scope)
 2. Pemutakhiran/Pembersihan data (data cleaning)
 3. Visualisasi & eksplorasi (mencari pola/tren/outlier)
 4. Hipotesis umum (jawaban awal + alasan/kausalitas)
 5. Analisis (deskriptif, eksploratori, inferensial/prediktif/preskriptif/kausal)
 6. Komunikasi hasil (temuan, rekomendasi, keputusan yang actionable)

Bagian 2: Melakukan teknik deskripsi data

Diberikan dataset: dataset_penjualan_ecommerce.csv

Deksripsi kolom:

1. Transaction_id
2. Transaction_value (Rp)
3. Product_category (Fashion, Elektronik, Makanan, Kesehatan, Olahraga)
4. Customer_age (16-49 tahun)
5. Month (1-12)

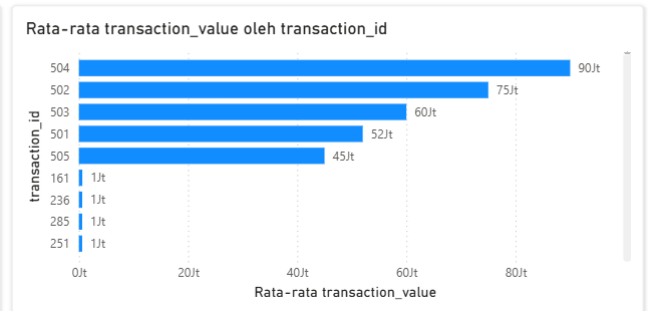
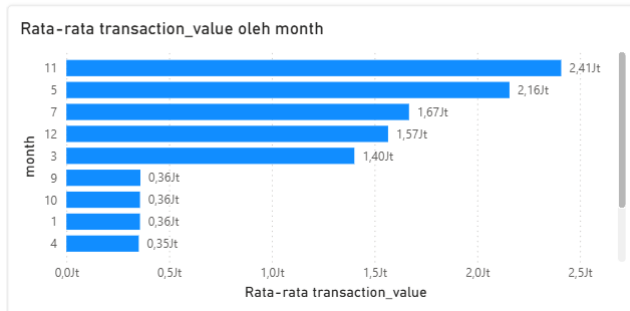
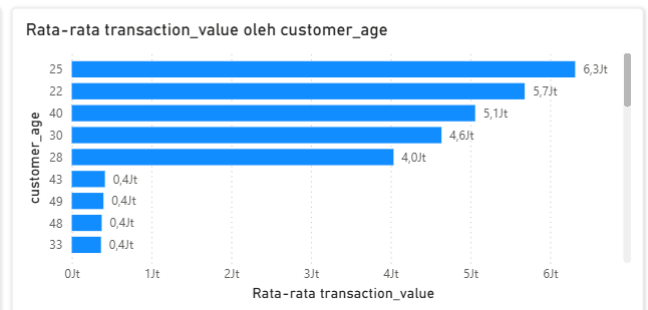
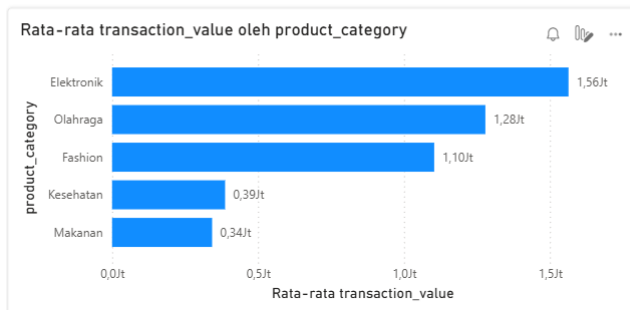
Statistik Deskriptif

1. Hitung mean, median, modus, nilai maksimum, nilai minimum, standar deviasi dari transaction_value!

Rata-rata (Mean)

Ringkasan cepat
dataset_penjualan_ecommerce

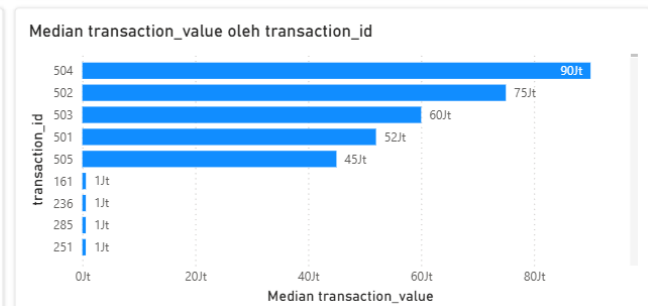
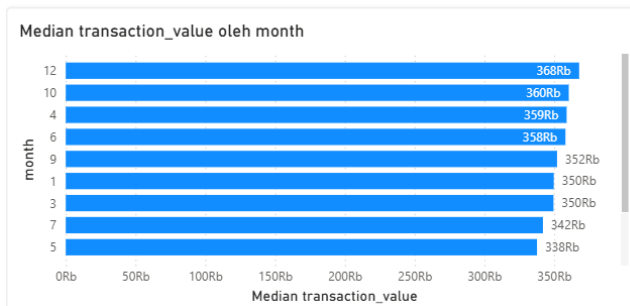
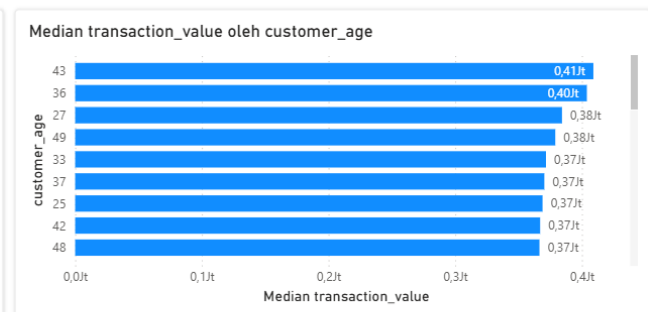
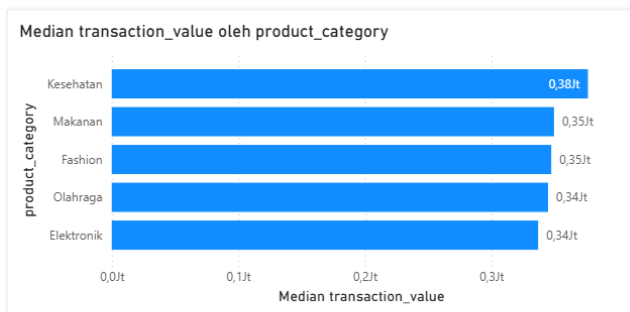
979293,33
Rata-rata transaction_value



Nilai Tengah (Median)

Ringkasan cepat
dataset_penjualan_ecommerce

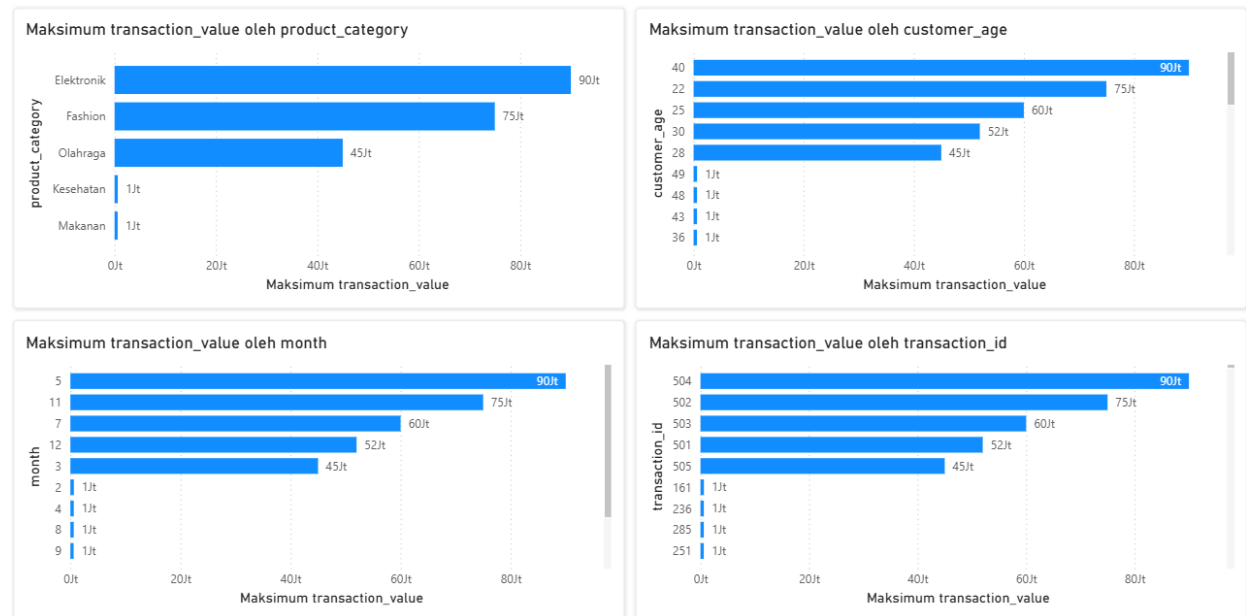
349334
Median transaction_value



Nilai Maksimum

Ringkasan cepat
dataset_penjualan_ecommerce

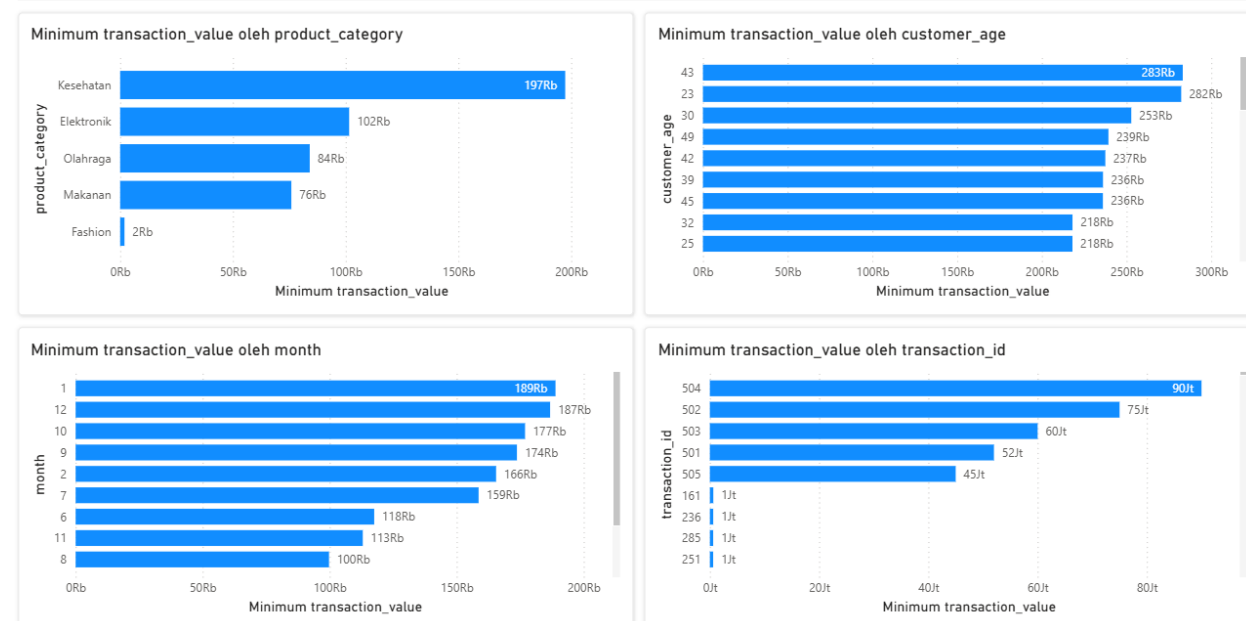
90000000
Maksimum transaction_v...



Nilai Minimum

Ringkasan cepat
dataset_penjualan_ecommerce

1919
Minimum transaction_val...



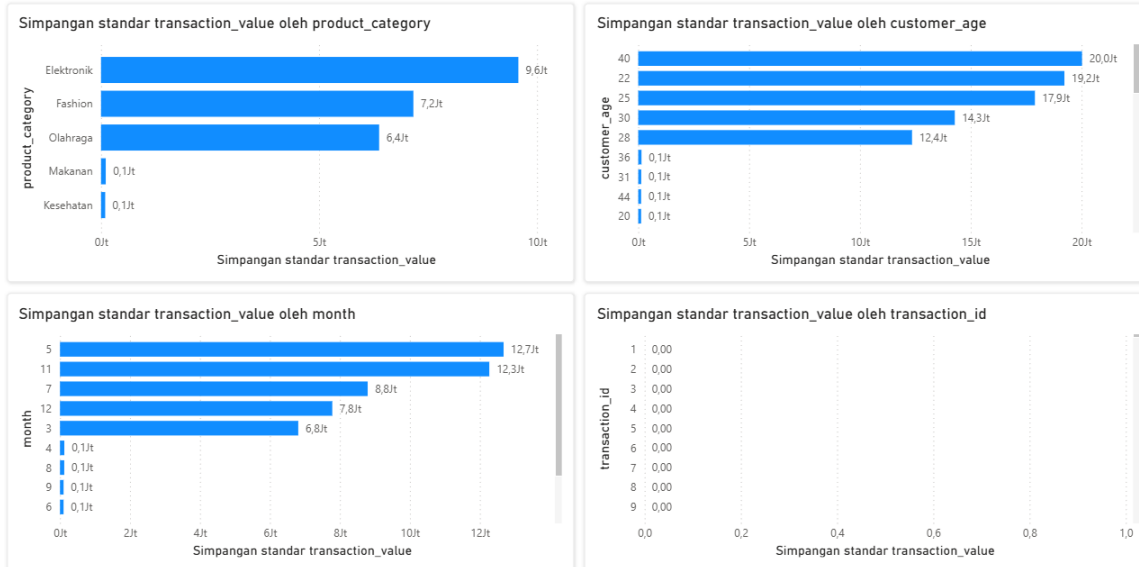
Standar Deviasi

Ringkasan cepat

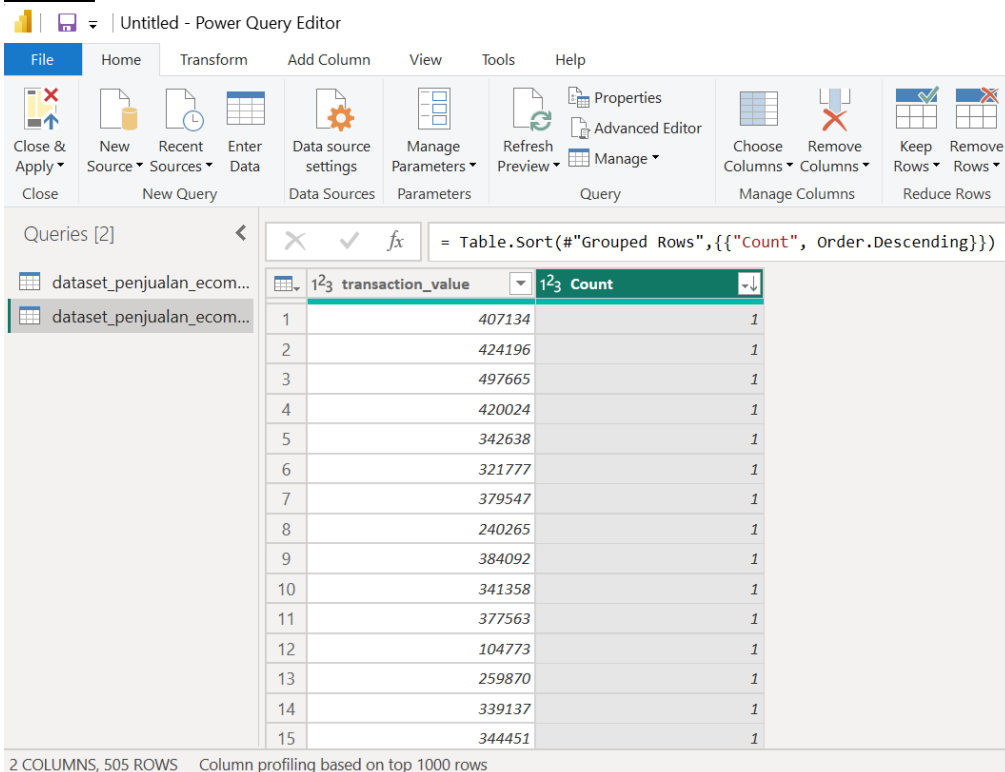
dataset_penjualan_ecommerce

6545282,78

Simpangan standar trans...



Modus



Dari hasil pengelompokan (Group By) pada kolom transaction_value, seluruh nilai transaksi hanya muncul 1 kali. Tidak ada nilai yang berulang. Oleh karena itu, tidak ada modus tunggal pada data transaksi ini.

Analisis:

Nilai rata-rata transaksi cukup tinggi (hampir 1 juta rupiah), namun median hanya sekitar Rp 349 ribu. Perbedaan besar antara mean dan median ini menunjukkan adanya distribusi data miring ke kanan (right-skewed) akibat keberadaan beberapa transaksi dengan nilai sangat besar.

Hal ini juga didukung dengan nilai maksimum yang mencapai Rp 90 juta, jauh di atas nilai tipikal transaksi, sehingga transaksi tersebut dapat dikategorikan sebagai outlier.

Standar deviasi yang besar (sekitar Rp 6,5 juta) menunjukkan bahwa penyebaran nilai transaksi sangat bervariasi, dari transaksi kecil di bawah Rp 10 ribu hingga transaksi besar puluhan juta.

Karena semua nilai transaksi unik, tidak ditemukan modus. Dengan demikian, tidak ada satu nilai pun yang paling sering muncul di dataset.

2. Buatlah tabel frekuensi transaksi berdasarkan product_category!

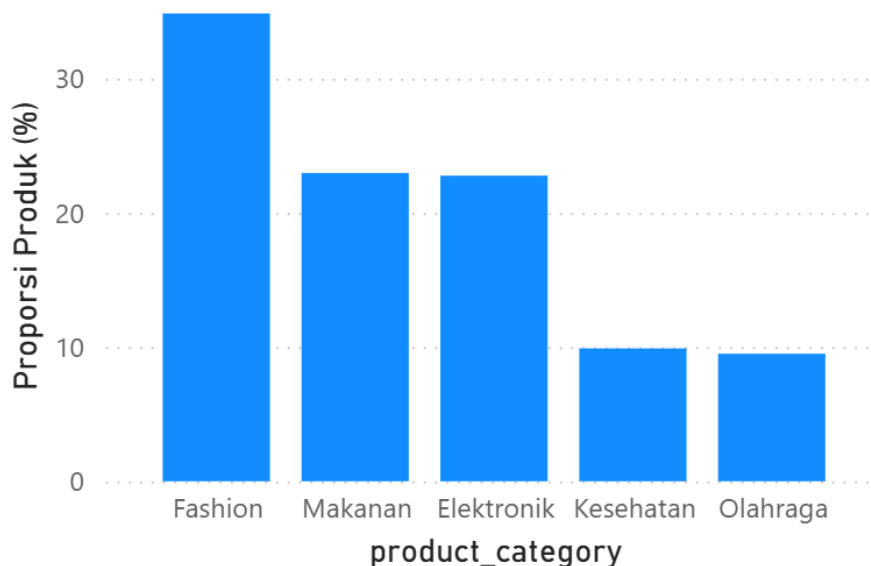
product_category	Count of transaction_id
Fashion	176
Makanan	116
Elektronik	115
Kesehatan	50
Olahraga	48
Total	505

Analisis

Kategori Fashion memiliki frekuensi transaksi paling tinggi dibanding kategori lain (176 kali). Kategori dengan transaksi paling sedikit adalah Olahraga (48 kali). Dari hasil ini dapat disimpulkan bahwa pelanggan e-commerce lebih banyak melakukan transaksi pada kategori Fashion, sementara minat pada produk Kesehatan dan Olahraga lebih rendah.

3. Hitung proporsi (%) untuk masing-masing kategori produk!

Proporsi Penjualan per Kategori Produk (%)



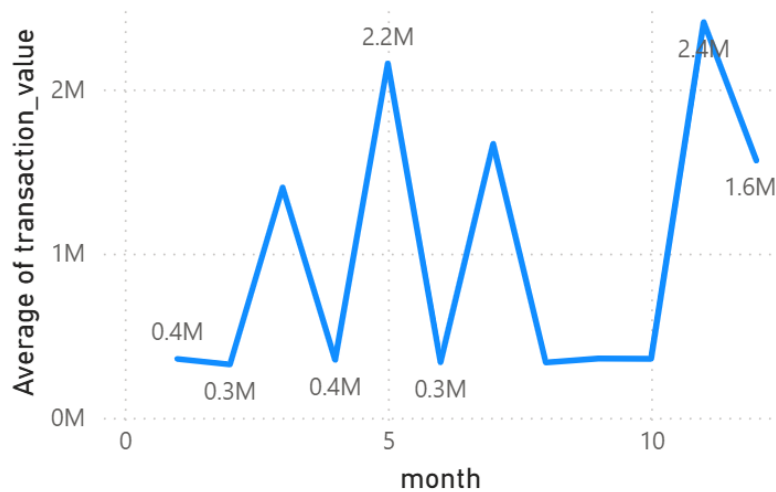
Analisis

Kategori produk dengan transaksi terbanyak adalah Fashion (34,85%), diikuti Makanan (22,97%) dan Elektronik (22,77%). Kategori dengan transaksi paling sedikit adalah Olahraga (9,50%). Hal ini menunjukkan bahwa konsumen e-commerce dalam dataset lebih sering bertransaksi untuk produk Fashion dibanding kategori lainnya.

Trend dan Outlier

1. Buatlah grafik line chart rata-rata transaksi per bulan!

Rata-rata Transaksi per Bulan



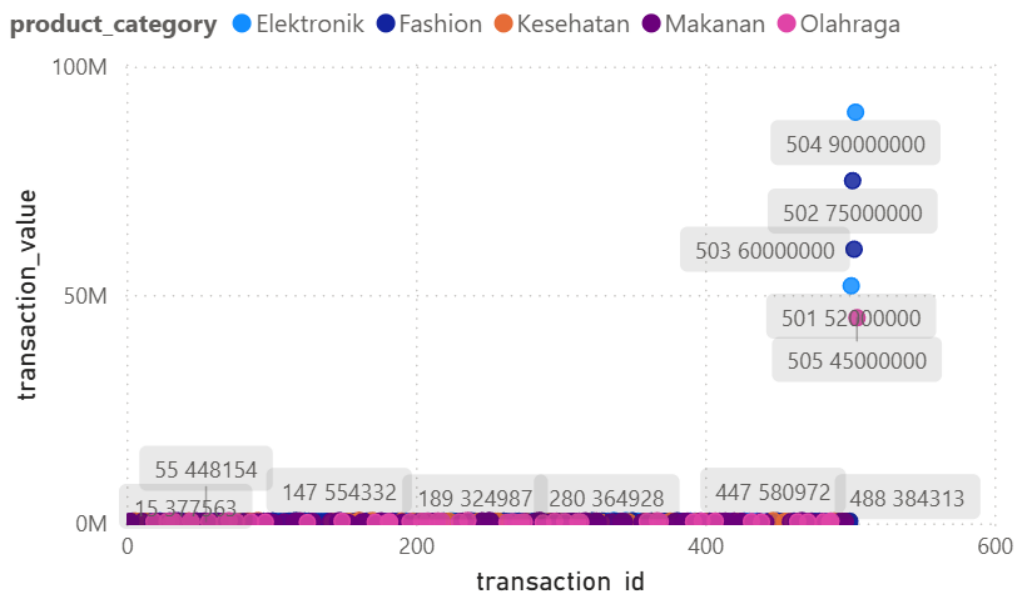
Analisis

Grafik line chart menunjukkan rata-rata nilai transaksi per bulan. Dari hasil pengolahan data, rata-rata transaksi terlihat naik signifikan di bulan 5 dan bulan 11. Hal ini bukan karena transaksi normal meningkat, melainkan karena terdapat transaksi bernilai sangat besar (outlier) pada bulan tersebut sehingga mempengaruhi rata-rata.

Jika menggunakan median sebagai pembanding, pola transaksi tipikal justru lebih stabil, dengan puncak terjadi di bulan 12 (akhir tahun) dan bulan 10, yang kemungkinan berkaitan dengan periode promo akhir tahun atau musim liburan. Jadi kesimpulannya, tren bulanan berdasarkan rata-rata sangat dipengaruhi oleh outlier, sehingga median lebih tepat digunakan untuk menggambarkan kondisi transaksi tipikal.

2. Buat boxplot/scatterplot untuk mendeteksi outlier nilai transaksi. Identifikasi minimal 3 transaksi yang tergolong outlier!

Distribusi Nilai Transaksi per ID Transaksi



Analisis

Scatter chart menampilkan titik-titik transaksi. Sebagian besar transaksi ada di bawah Rp 1 juta, sehingga titiknya bergerombol di bagian bawah grafik.

Tapi ada beberapa titik yang melonjak jauh ke atas. Tiga transaksi yang jelas-jelas outlier adalah:

- ID 504 – Rp 90.000.000 (Elektronik)
- ID 502 – Rp 75.000.000 (Fashion)
- ID 503 – Rp 60.000.000 (Fashion)

Nilai transaksi ini jauh lebih besar dibanding transaksi lainnya. Akibatnya, rata-rata transaksi jadi ikut naik tinggi. Kemungkinan transaksi ini terjadi karena pembelian dalam jumlah banyak sekaligus, atau mungkin ada input data khusus.

Pola dan Segmentasi

1. Buat tabel pivot untuk menghitung:

a. Rata-rata nilai transaksi per kelompok usia (16-25, 26-25, 36-49)

```
1 kelompok Usia =  
2 SWITCH(  
3 TRUE(),  
4 'dataset_penjualan_ecommerce'[customer_age] >= 16 && 'dataset_penjualan_ecommerce'[customer_age] <= 25, "16-25",  
5 'dataset_penjualan_ecommerce'[customer_age] >= 26 && 'dataset_penjualan_ecommerce'[customer_age] <= 35, "26-35",  
6 "36-49"  
7 )  
8
```

transaction_id	transaction_value	product_category	customer_age	month	Kelompok Usia
4	447052	Fashion	36	9	36-49
7	420024	Fashion	26	11	26-35
9	342638	Fashion	39	7	36-49
10	321777	Fashion	39	6	36-49
12	240265	Fashion	37	11	36-49
16	377872	Fashion	17	11	16-25
17	104773	Fashion	36	4	36-49
21	375938	Fashion	40	2	36-49
25	318697	Fashion	30	10	26-35
29	456044	Fashion	24	5	16-25
32	319613	Fashion	40	1	36-49
35	361486	Fashion	41	6	36-49
36	505126	Fashion	17	3	16-25
38	465324	Fashion	43	5	36-49
40	366608	Fashion	23	10	16-25

Table: dataset_penjualan_ecommerce (505 rows) Column: Kelompok Usia (3 distinct values)

Kelompok Usia Average of transaction_value

16-25	1289069.63
26-35	1043036.82
36-49	747322.62
Total	979293.33

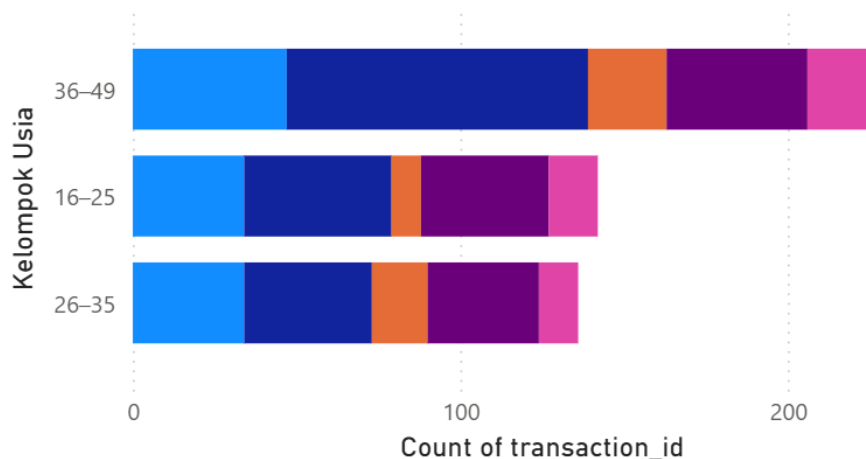
b. Distribusi kategori produk pada masing-masing kelompok usia

Kelompok Usia	Elektronik	Fashion	Kesehatan	Makanan	Olahraga	Total
16-25	34	45	9	39	15	142
26-35	34	39	17	34	12	136
36-49	47	92	24	43	21	227
Total	115	176	50	116	48	505

2. Buat visualisasi bar chart hubungan customer_age berdasarkan kelompok umur pada point (a) dengan product_category

Distribusi Produk Berdasarkan Kelompok Usia

product_category ● Elektronik ● Fashion ● Kesehatan ● Makanan ● Olahraga



Analisis untuk Pola dan Segmentasi

- Rata-rata transaksi per kelompok usia:
Kelompok usia 16–25 tahun memiliki rata-rata transaksi paling tinggi (Rp 1,28 juta), tetapi nilai ini dipengaruhi oleh adanya outlier (transaksi puluhan juta dari usia muda). Jika dilihat dari median, nilai transaksi cenderung stabil dan bahkan sedikit meningkat seiring bertambahnya usia (Rp 334 ribu → Rp 360 ribu).
- Distribusi kategori produk per kelompok usia:
 1. Usia 16–25 tahun: paling banyak membeli Fashion dan Makanan.
 2. Usia 26–35 tahun: paling banyak membeli Fashion, lalu diikuti Makanan dan Elektronik.
 3. Usia 36–49 tahun: pola belanja lebih merata, dengan Fashion dan Elektronik hampir seimbang.
- Visualisasi bar chart:
Bar chart menunjukkan bahwa setiap kelompok usia punya kecenderungan berbeda dalam memilih produk. Usia muda dominan Fashion, usia menengah mulai banyak ke Elektronik, dan usia lebih tua belanja lebih seimbang di semua kategori.

Analisis Korelasi & Kausalitas

1. Berdasarkan data apakah pelanggan yang lebih tua cenderung bertransaksi lebih besar?

Jawaban:

Jika kita bandingkan nilai transaksi berdasarkan kelompok usia:

- Rata-rata (mean) justru lebih tinggi pada usia muda (16–25 tahun ≈ Rp 1,28 juta) dibanding usia 26–35 (Rp 1,04 juta) dan 36–49 (Rp 747 ribu).
- Namun, rata-rata ini dipengaruhi oleh outlier (transaksi puluhan juta rupiah) yang banyak terjadi di kelompok usia muda.
- Jika menggunakan median (nilai tengah), terlihat hasil yang lebih stabil:
 - Usia 16–25 tahun ≈ Rp 334 ribu
 - Usia 26–35 tahun ≈ Rp 340 ribu
 - Usia 36–49 tahun ≈ Rp 360 ribu

Dengan median, transaksi cenderung sedikit meningkat seiring bertambahnya usia.

Kesimpulan:

- Jika melihat rata-rata, tampak seolah usia muda bertransaksi lebih besar (karena outlier).
- Tapi jika melihat pola umum (median), pelanggan yang lebih tua justru cenderung bertransaksi lebih besar secara konsisten.

2. Apakah ada bulan tertentu dengan nilai transaksi lebih tinggi?

Jawaban:

Dari line chart rata-rata transaksi per bulan:

- Bulan 5 (Mei) dan 11 (November) menunjukkan rata-rata transaksi yang sangat tinggi.
- Namun, lonjakan ini dipengaruhi adanya transaksi outlier (puluhan juta rupiah) di bulan tersebut.
- Jika melihat median (transaksi normal), nilai transaksi tipikal paling tinggi ada di bulan 12 (Desember) dan bulan 10 (Oktober).

Kesimpulan:

- Ada bulan tertentu dengan transaksi besar (5 dan 11), tapi itu karena ada transaksi ekstrem.
- Pola umum menunjukkan transaksi cenderung meningkat di akhir tahun (bulan 10–12), kemungkinan dipengaruhi oleh promo akhir tahun dan musim belanja liburan.

Analisis dan Diskusi

Berikan kesimpulan:

- Berapa rata-rata transaksi tipikal?

Jawaban:

Rata-rata transaksi adalah sekitar Rp 979 ribu. Namun angka ini terlalu tinggi karena ada transaksi yang nilainya sangat besar. Jika dilihat dari median (nilai tengah), transaksi normal pelanggan ada di sekitar Rp 349 ribu. Jadi transaksi tipikal pelanggan sebenarnya berada pada kisaran Rp 300–400 ribu.

- Pada bulan apa tren penjualan tertinggi?

Jawaban:

Berdasarkan rata-rata, bulan 5 (Mei) dan bulan 11 (November) terlihat paling tinggi, tetapi ini karena ada transaksi yang nilainya sangat besar. Jika dilihat dari median, tren penjualan normal justru paling tinggi pada bulan 12 (Desember) dan bulan 10 (Oktober). Artinya, penjualan cenderung meningkat di akhir tahun.

- Apa penyebab outlier transaksi di atas Rp 50 juta?

Jawaban:

Transaksi di atas Rp 50 juta muncul karena ada beberapa transaksi dengan nilai yang sangat jauh di atas transaksi biasa. Hal ini kemungkinan disebabkan oleh pembelian dalam jumlah banyak sekaligus atau pencatatan transaksi khusus.

- Produk apa yang dominan di usia 18–25 tahun?

Jawaban:

Pada kelompok usia 18–25 tahun, produk yang paling banyak dibeli adalah Fashion. Produk berikutnya yang juga cukup sering dibeli adalah Makanan. Jadi produk yang dominan untuk kelompok usia ini adalah Fashion.

- Bagaimana korelasi usia dan nilai transaksi?

Jawaban:

Jika dilihat dari rata-rata, kelompok usia muda terlihat punya transaksi lebih besar, tetapi ini dipengaruhi oleh transaksi yang sangat tinggi (outlier). Jika melihat nilai tengah (median), transaksi

justru sedikit lebih besar pada kelompok usia yang lebih tua. Jadi secara umum, pelanggan yang lebih tua cenderung punya transaksi yang lebih besar dan lebih stabil dibandingkan pelanggan yang lebih muda.

Kesimpulan Akhir:

Berdasarkan hasil eksplorasi data penjualan e-commerce:

1. Nilai transaksi tipikal pelanggan berada pada kisaran Rp 300–400 ribu. Rata-rata transaksi (Rp 979 ribu) lebih tinggi karena dipengaruhi oleh beberapa transaksi dengan nilai yang sangat besar.
2. Tren penjualan bulanan menunjukkan peningkatan pada akhir tahun, khususnya di bulan Oktober dan Desember, yang kemungkinan dipengaruhi oleh promo dan musim belanja akhir tahun.
3. Outlier transaksi dengan nilai di atas Rp 50 juta terjadi karena adanya pembelian dalam jumlah besar atau transaksi khusus, sehingga nilainya jauh lebih tinggi dibandingkan transaksi normal.
4. Segmentasi usia memperlihatkan bahwa kelompok usia muda (16–25 tahun) lebih banyak membeli produk Fashion, sedangkan usia yang lebih tua menunjukkan pola belanja yang lebih merata dan konsisten.
5. Korelasi usia dan transaksi menunjukkan bahwa meskipun rata-rata pada usia muda tampak lebih besar (karena outlier), secara umum pelanggan yang lebih tua cenderung memiliki nilai transaksi tipikal yang sedikit lebih tinggi dan stabil.

~sekian dan selamat mengerjakan~