PRE-TEST DATA SCIENTIST INTERN

PT Bina Pertiwi

Nama : Nathania Christy Nugraha
Email : nathania.nugraha03@gmail.com
No. HP : 0878-2050-1325

LATIHAN 1

Dataset = penjualan tiket auditorium sepanjang tahun
Masalah = terdapat nilai hilang, nilai tidak sesuai, dan duplikasi
Tujuan = membersihkan dataset agar siap digunakan untuk analisis lebih lanjut

Penyelesaian:

1. **Exploratory Data Analysis (EDA)**

   Langkah awal pembersihan data yaitu dengan melakukan eksplorasi singkat untuk memahami struktur dan pola umum data. EDA berguna untuk mengetahui tipe data, distribusi nilai, konsistensi, serta mendeteksi anomali yang dapat mengganggu proses pembersihan dan analisis selanjutnya.

2. **Pembersihan Data**

   Urutan pembersihan data dilakukan secara bertahap yaitu:

   - Pertama, data duplikat dihapus terlebih dahulu untuk menghindari perhitungan berulang yang bisa menyebabkan bias dalam analisis.
   - Kedua, data yang tidak sesuai diperbaiki agar proses selanjutnya hanya menggunakan data yang valid secara logika.
   - Terakhir, menangani nilai hilang, karena proses imputasi akan lebih akurat jika didasarkan pada data yang bersih dan valid.

   Berikut adalah detail proses dari ketiga tahapan pembersihan:

   a. **Menangani duplikasi data**

      Data duplikat dapat menyebabkan overrepresentasi dalam analisis, perhitungan statistik yang bias, serta menurunkan kualitas hasil analisis. Oleh karena itu, data duplikat harus dihapus agar hasil analisis lebih akurat dan tidak terdistorsi.

b. Menangani ketidaksesuaian data.

Data yang tidak sesuai merupakan nilai-nilai yang tidak logis (berada di luar batas yang wajar). Ketidaksesuaian data dapat menyebabkan kesalahan dalam pengambilan keputusan dan kesimpulan analisis. Perbaikan dapat dilakukan jika sumber kesalahannya dapat diidentifikasi, misalnya:

- durasi_film negatif yang seharusnya bernilai positif.
  contoh: "- 90 menit" dikoreksi menjadi "90 menit".
- jumlah_tiket terjual melebihi kapasitas
  contoh: kapasitas maksimal auditorium adalah 300, namun tiket terjual sebanyak 500, maka dapat dikoreksi dengan membatasi nilainya sesuai kapasitas maksimal (yaitu 300) atau menyesuaikan dengan pola data serupa (misalnya menggunakan rata-rata atau distribusi nilai dari film dengan kapasitas sejenis).

c. Menangani data yang hilang = yaitu perlu diperbaiki

Nilai hilang *(missing values)* dapat mengganggu proses analisis dan menyebabkan error saat melakukan perhitungan statistik atau pemodelan. Penanganannya dapat dilakukan berdasarkan jenis data, yaitu:

- Apabila datanya adalah <u>data numerik</u> *(durasi_film, kapasitas_auditorium, harga_tiket)*, maka dapat ditangani dengan mengisi menggunakan:
  - nilai mean, jika distribusi data nya normal (karena nilai mean cocok untuk mewakili pusat data), atau
  - median, jika distribusi datanya skewed/miring (karena median tidak terpengaruh dengan angka yang ekstrem dan lebih representatif)
- Apabila datanya adalah <u>data kategorikal</u> *(judul_film)*, maka dapat ditangani dengan mengisi menggunakan:
  - Kategori judul baru, seperti "others" atau "lainnya" – untuk merepresentasikan data tidak diketahui, atau
  - Kategori judul yang paling sering muncul (modus) – agar nilai hilang dapat mengikuti pola mayoritas.

Namun, jika nilai hilang atau nilai yang tidak sesuai tidak memungkinkan untuk diperbaiki secara logis, maka baris data terkait sebaiknya dihapus untuk menjaga kualitas dan akurasi analisis.

LATIHAN 2

Dataset = informasi perjalanan pelanggan selama satu bulan

Tujuan = mengeksplorasi pola dan hubungan antar fitur data menggunakan visualisasi

Penyelesaian:

1. **Visualisasi Distribusi Data**

   Langkah awal visualisasi data umumnya dilakukan dengan melihat distribusi dari data numerik (seperti: jarak, durasi, dan harga). Visualisasi distribusi data dapat menggunakan **histogram** untuk melihat bentuk sebaran data, seperti apakah nilainya normal, skewed, atau memiliki outlier. Tujuan dari visualisasi ini adalah untuk memahami karakteristik awal data numerik, yang dapat memengaruhi langkah pembersihan, pemilihan metode analisis, maupun interpretasi hasil selanjutnya.

2. **Visualisasi Hubungan Antar Variabel**

   Langkah selanjutnya adalah mengeksplorasi hubungan antar variabel numerik menggunakan **scatterplot** tergantung dengan tujuan analisisnya.

Pasangan Variabel	Tujuan Analisis
jarak vs harga	Mengetahui apakah semakin jauh jarak perjalanan, tarif yang dikenakan juga semakin tinggi
jarak vs durasi	Mengevaluasi apakah jarak tempuh sebanding dengan durasi perjalanan
durasi vs harga	Melihat apakah durasi perjalanan menjadi salah satu faktor yang memengaruhi tarif
harga vs driver_rating	Mengidentifikasi apakah tarif perjalanan berkaitan dengan kepuasan pelanggan terhadap pengemudi
harga vs customer_rating	Mengetahui apakah tarif perjalanan memengaruhi penilaian pengemudi terhadap pelanggan

   Selain itu, **boxplot** dan **bar chart** juga dapat digunakan saat fitur numerik dikelompokkan terlebih dahulu, misalnya untuk melihat variasi harga atau rata-rata rating berdasarkan kelompok jarak.

3. **Analisis Korelasi antar Fitur**

   Analisis korelasi dilakukan untuk mengetahui kekuatan dan arah hubungan antar fitur numerik secara keseluruhan. Korelasi dihitung menggunakan nilai koefisien (seperti

Pearson) dan divisualisasikan menggunakan **heatmap**, sehingga hubungan positif atau negatif antar fitur dapat dengan mudah diidentifikasi. Korelasi positif antara fitur dapat dijadikam bahan pertimbangan awal jika akan dilakukan pemodelan prediktif di tahap berikutnya.

INSIGHT TAMBAHAN

Selain mengeksplorasi hubungan antar fitur-fitur yang tersedia, analisis tambahan dapat dilakukan dengan menurunkan fitur baru dari kombinasi data yang ada. Salah satunya adalah kecepatan perjalanan (dihitung dari jarak dibagi durasi). Fitur ini dapat digunakan untuk melihat apakah terdapat hubungan antara kecepatan dan driver_rating, seperti apakah pengemudi dengan rating tinggi cenderung berkendara lebih cepat atau sebaliknya.

Selain itu, bisa juga dengan menggabungkan tanggal_waktu dan durasi, dapat dilakukan eksplorasi awal terhadap indikasi waktu sibuk atau kemacetan. Misalnya, durasi perjalanan dapat dibandingkan antar jam atau hari untuk mendeteksi adanya pola waktu tertentu yang memiliki waktu tempuh lebih lama. Analisis ini memungkinkan identifikasi waktu-waktu potensial yang padat dan dapat berdampak pada performa operasional.

LATIHAN 3

Dataset = informasi karyawan perusahaan

Tujuan = Memahami pola dan hubungan antar fitur menggunakan statistik deskriptif dan inferensial

Penyelesaian:

1. **Statistik Deskriptif**

   Langkah awal dilakukan dengan menganalisis ringkasan statistik untuk masing-masing fitur. Tujuannya adalah untuk mendapatkan pemahaman umum mengenai karakteristik data.

   a. Fitur numerik seperti umur, lama_bekerja, dan gaji dianalisis menggunakan nilai minimum dan maksimum, mean atau rata-rata, median atau nilai tengah, kuartil (q3, q1), dan standar deviasi. Hal ini membantu memahami sebaran dan pola umum

data numerik. Misalnya analisis pada fitur *lama_bekerja*, dapat memberikan informasi mengenai tingkat retensi dan pengalaman karyawan di perusahaan. Atau analisis pada fitur *umur*, untuk mengetahui persebaran usia karyawan di perusahaan. Bisa juga analisis pada fitur *gaji*, untuk mengetahui range atau skala gaji karyawan di perusahaan.

b. Fitur kategorikal seperti jenis_kelamin dan pendidikan dianalisis menggunakan frekuensi atau jumlah kemunculan tiap kategori, dan persentase distribusi per kategori. Hal ini digunakan untuk melihat dominasi kategori tertentu dalam populasi, seperti proporsi karyawan berdasarkan jenis_kelamin atau jenjang pendidikan.

2. Analisis Hubungan dan Korelasi Antar Fitur

Setelah mendapatkan informasi dari hasil analisis statistik deskriptif, langkah selanjutnya yaitu mengidentifikasi apakah terdapat hubungan antar fitur numerik. Misalnya, apakah karyawan yang lebih tua cenderung memiliki masa kerja lebih lama, atau apakah lama bekerja berkaitan dengan tingkat gaji.

Untuk mengevaluasi hubungan dua variabel secara visual, dapat menggunakan scatterplot agar dapar melihat pola kecenderungan linear atau keberadaan outlier. Selanjutnya, untuk mengetahui kekuatan dan arah hubungan secara kuantitatif antar seluruh fitur numerik, dapat menggunakan analisis korelasi Pearson yang divisualisasikan dalam bentuk heatmap.

3. Statistik Inferensial

Setelah mengetahui adanya hubungan antar fitur dalam dataset, dilakukan pendekatan inferensial untuk menguji hipotesis dan menilai kekuatan pengaruh antar fitur secara kuantitatif. Dua metode yang digunakan dalam analisis ini adalah independent t-test dan regresi linear.

a. Independent t-test, yaitu uji statistik untuk membandingkan rata rata dua kelompok berbeda. Misalnya, untuk mengetahui apakah terdapat perbedaan rata-rata gaji antara karyawan pria dan wanita, dilakukan uji t dua sampel independen. Hasil uji ini menunjukkan apakah perbedaan yang terlihat secara deskriptif memang signifikan secara statistik atau hanya terjadi karena variasi acak.

b. Regresi Linear, untuk mengukur seberapa besar pengaruh satu atau lebih variabel independent terhadap variabel dependen. Misalnya untuk menganalisis pengaruh

usia, lama bekerja, dan pendidikan terhadap gaji. Regresi menghasilkan informasi seperti koefisien (arah dan besar pengaruh), nilai p (signifikansi statistik), dan R^2 (kemampuan model menjelaskan variabilitas data).

Dengan menggabungkan kedua metode ini, analisis inferensial tidak hanya dapat menjawab *apakah ada perbedaan*, tetapi juga *apa yang memengaruhi apa* dan *seberapa kuat pengaruhnya*.

LATIHAN 4

Dataset = Informasi pelanggan perusahaan kartu kredit

Tujuan = Membangun model machine learning untuk memprediksi apakah pelanggan akan tertarik pada penawaran kartu kredit baru

Penyelesaian:

1. **Memahami Masalah**

   Pada kasus ini, diminta untuk membangun model machine learning untuk memprediksi apakah pelanggan akan tertarik pada penawaran kartu kredit baru. Yang berarti masalah pada kasus ini merupakan **klasifikasi biner** – karena target prediksi model nya bersifat kategorikal, yaitu "tertarik" atau "tidak tertarik". Model yang akan dibangun harus mampu mempelajari pola dari data historis pelanggan untuk memprediksi kecenderungan dalam menerima penawaran kartu kredit baru.

2. **Pra Proses Data**

   Sebelum dilakukan model, beberapa persiapan perlu dilakukan untuk memastikan bahwa data berada dalam format yang dapat diproses secara efektif oleh algoritma machine learning. Proses ini bertujuan untuk meningkatkan akurasi, stabilitas, dan efisiensi pelatihan model, serta mencegah kesalahan akibat perbedaan skala atau format data yang tidak seragam. Beberapa langkah yang dilakukan meliputi:

   a. Cek dan penanganan nilai hilang, untuk memastikan tidak ada data kosong yang menyebabkan error saat pelatihan model.
   b. Normalisasi atau standarisasi fitur numerik, seperti pendapatan dan pengeluaran_bulanan, agar berada dalam skala yang seimbang dan tidak mendominasi proses pembelajaran model.

c. Encoding fitur kategorikal, seperti jenis_kelamin, agar dapat dikonversi dari format teks ke format numerik yang dapat dibaca oleh model (Wanita dikonversi menjadi 0, dan Pria dikonversi menjadi 1)
d. Split data menjadi data latih dan data uji (misalnya dengan rasio 80:20) agar performa model dapat diuji pada data yang tidak dilihat saat pelatihan.

3. Pemilihan Model

Untuk menyelesaikan permasalahan klasifikasi biner ini, dipertimbangkan dua model yang umum digunakan dan cukup efektif, yaitu **Decision Tree** dan **Random Forest**.

a. Decision Tree

   Model ini sederhana dan mudah diinterpretasi, dengan cara membagi data berdasarkan aturan logika "jika – maka", dan dapat memberikan pemahaman yang jelas tentang bagaimana keputusan dibuat. Model ini cocok untuk dataset yang tidak terlalu besar atau kompleks.

b. Random Forest

   Model ini merupakan pengembangan dari decision tree karena menggunakan banyak pohon atau tree dan melakukan voting untuk menentukan prediksi akhir, sehingga Random Forest lebih stabil dan akurat. Model ini cocok digunakan jika dataset memiliki banyak fitur atau data noise. Keunggulan dari model ini adalah untuk mengetahui fitur mana yang paling berpengaruh melalui feature importance.

4. Pelatihan dan Validasi Model

Setelah menentukan model yang akan digunakan, langkah selanjutnya yaitu pelatihan (training) menggunakan data latih yang sudah di siapkan pada pra proses sebelumnya. Selama proses ini, model mengidentifikasi pola dari data historis untuk menentukan kombinasi nilai fitur yang berpengaruh terhadap keputusan pelanggan dalam menerima atau menolak penawaran kartu kredit baru.

Setelah dilatih, dilakukan proses validasi untuk menghindari hasil yang bias serta memastikan model tidak hanya menghafal dari data latih. Proses ini dilakukan dengan menguji performa model pada data yang tidak digunakan saat pada pelatihan. Hasil validasi memberikan gambaran sejauh mana model mampu dalam melakukan generalisasi terhadap data baru.

5. **Evaluasi Hasil**

   Untuk mengevaluasi hasil kinerja model klasifikasi, digunakan metrik evaluasi seperti Accuracy (untuk mengetahui seberapa banyak prediksi benar dari total data), Precision & Recall (untuk melihat ketepatan dan kemampuan model dalam menangkap kelas positif)), F1-Score (untuk menyeimbangkan precision dan recall), ROC AUC (untuk menilai kemampuan model dalam membedakan antar kelas). Selain itu juga evaluasi dapat dilihat melalui visualiasi dari confusion matrix untuk melihat detail jumlah prediksi benar dan salah pada masing-masing kelas.

   Apabila hasil evaluasi belum memuaskan, dapat dilakukan hyperparameter tuning untuk meningkatkan performa model dan mencegah overfitting.

6. **Kesimpulan**

   Model klasifikasi biner dibangun untuk memprediksi apakah pelanggan akan tertarik pada penawaran kartu kredit baru berdasarkan fitur fitur yang tersedia dalam dataset. Model ini dapat digunakan sebagai dasar pengambilan keputusan dalam strategi pemasaran perusahaan.

LATIHAN 5

Dataset = Data produk di Sephora

Tujuan = Mendapatkan wawasan terkait dengan model pembelajaran mesin dan visualisasi data

Penyelesaian: *eksplorasi data menggunakan Python – Google Colab*

1. **Memahami Dataset**

   Dataset ini terdiri dari 8000 baris data produk dari Sephora, dengan total 17 fitur. Setiap baris mewakili satu produk yang dijual, dengan informasi tambahan seperti brand, nama produk, kategori, harga, rating, dan jumlah ulasan. Tipe data dalam dataset ini mencakup **numerik** (*id, rating, number_of_reviews, love, price, value_price, exclusive*), **kategorikal** (*brand, category, name, size, URL, options, details, how_to_use, ingredients*), dan **boolean** (*MarketingFlags*). Dataset tidak memiliki nilai yang hilang maupun baris duplikat. Hal ini menunjukkan bahwa kualitas data cukup baik untuk langsung digunakan dalam proses eksplorasi dan analisis lebih lanjut tanpa perlu pembersihan data tambahan.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8000 entries, 0 to 7999
Data columns (total 17 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 8000 non-null   int64  
 1   brand              8000 non-null   object 
 2   category           8000 non-null   object 
 3   name               8000 non-null   object 
 4   size               8000 non-null   object 
 5   rating             8000 non-null   float64
 6   number_of_reviews  8000 non-null   int64  
 7   love               8000 non-null   int64  
 8   price              8000 non-null   float64
 9   value_price        8000 non-null   float64
 10  URL                8000 non-null   object 
 11  MarketingFlags     8000 non-null   bool   
 12  options            8000 non-null   object 
 13  details            8000 non-null   object 
 14  how_to_use         8000 non-null   object 
 15  ingredients        8000 non-null   object 
 16  exclusive          8000 non-null   int64  
dtypes: bool(1), float64(3), int64(4), object(9)
memory usage: 1007.9+ KB
```

```
# Cek missing value
missing_values = df.isnull().sum()
print("Missing Value per Kolom:\n", missing_values)

# Cek duplikat
duplicate_count = df.duplicated().sum()
print("\nJumlah Baris Duplikat:", duplicate_count)
```

```
Missing Value per Kolom:
 id                   0
brand                 0
category              0
name                  0
size                  0
rating                0
number_of_reviews     0
love                  0
price                 0
value_price           0
URL                   0
MarketingFlags        0
options               0
details               0
how_to_use            0
ingredients           0
exclusive             0
dtype: int64

Jumlah Baris Duplikat: 0
```

Untuk menambah wawasan dalam memahami dataset, perlu diketahui informasi dari masing-masing fitur, yaitu:

- id = ID unik untuk setiap produk
- brand = Nama merek/brand dari produk
- category = Kategori produk
- name = penanda tipe/varian produk dalam satu brand atau category
- size = Ukuran atau kemasan produk (dalam oz/mL)
- rating = Rating rata-rata dari pengguna terhadap produk (skala 0–5)
- number_of_reviews = Jumlah total ulasan yang diterima produk
- love = Jumlah pengguna yang menyukai produk
- price = Harga aktual dari produk
- value_price = Nilai harga yang tertera (harga awal sebelum diskon)
- URL = Link ke halaman produk di web Sephora
- MarketingFlags = tanda pemasaran khusus (True/False)
- options = Variasi atau opsi produk, jika ada
- details = Deskripsi umum tentang produk
- how_to_use = Instruksi penggunaan produk
- ingredients = Daftar bahan atau kandungan dalam produk
- exclusive = Menunjukkan apakah produk eksklusif di Sephora (1 = eksklusif, 0 = tidak)

2. Pemilihan Fitur

Berdasarkan pemahaman awal terhadap masing-masing fitur, dilakukan seleksi terhadap fitur yang akan digunakan dalam proses eksplorasi data. Fitur seperti id, name, size, URL, options, details, how_to_use, dan ingredients dikecualikan karena tidak mengandung informasi kuantitatif yang dapat langsung dianalisis tanpa proses pra-pemrosesan lanjutan. Sehingga fitur yang hanya digunakan pada tahapan ini yaitu: brand, category, rating, number_of_reviews, love, price, value_price, MarketingFlags, dan exclusive.

3. Statistik Deskriptif

Namun, sebelum mengidentifikasi hubungan antar fitur dalam data, perlu dilakukan analisis statistik deskriptif terlebih dahulu. Untuk fitur numerik, dihitung ukuran pusat dan sebaran seperti rata-rata, median, minimum, maksimum, dan standar deviasi. Sedangkan untuk fitur kategorikal, dihitung jumlah kategori unik dan kategori yang paling sering muncul (modus). Langkah ini membantu memahami karakteristik awal data dan menentukan pendekatan eksplorasi yang tepat pada tahap selanjutnya.

a. Statistik Numerik

- Rating rata-rata adalah 4.087 dari skala 5, menunjukkan mayoritas produk memiliki rating tinggi.
- Jumlah ulasan sangat bervariasi, dari 0 hingga 19.000, dengan median hanya 56. Yang berarti distribusinya sangat miring (banyak produk dengan ulasan sedikit).
- Jumlah love juga sangat bervariasi, dari 0 hingga 1.300.000, dengan median 5.500. Ini menunjukkan ada produk sangat populer, tapi sebagian besar tidak
- Harga produk (price) berkisar antara 2 hingga 549, dengan rata-rata sekitar 50
- value_price sangat mirip dengan price, menandakan bahwa sebagian besar produk tidak sedang diskon

```
# Statistik numerik
numerical_features = ['rating', 'number_of_reviews', 'love', 'price', 'value_price']
numerical_stats = df[numerical_features].describe()

print("Statistik Deskriptif - Fitur Numerik:")
print(numerical_stats)
```

```
Statistik Deskriptif - Fitur Numerik:
            rating  number_of_reviews          love        price  value_price
count  8000.000000        8000.000000  8.000000e+03  8000.000000  8000.000000
mean      4.087313         303.694250  1.757584e+04    49.903846    51.082835
std       0.758233         931.236885  4.421160e+04    46.854991    48.513402
min       0.000000           0.000000  0.000000e+00     2.000000     2.000000
25%       4.000000          14.000000  2.000000e+03    24.000000    24.000000
50%       4.000000          56.000000  5.500000e+03    35.000000    35.000000
75%       4.500000         232.000000  1.540000e+04    59.000000    60.000000
max       5.000000       19000.000000  1.300000e+06   549.000000   549.000000
```

b. Statistik Kategorikal

- Terdapat 310 brand unik dalam dataset, namun brand SEPHORA COLLECTION mendominasi dengan 492 produk, jauh di atas brand lainnya seperti CLINIQUE (211) dan TOM FORD (150). Hal ini menunjukkan adanya ketimpangan distribusi produk antar brand.
- Kategori produk berjumlah 142 jenis, dengan kategori Perfume paling banyak (620 produk), diikuti Moisturizers (398) dan Face Serums (334). Artinya, fokus produk yang terjual di Sephora cenderung berada pada kategori fragrance dan skincare.

```
# Statistik kategorikal
categorical_features = ['brand', 'category']

print("Statistik Deskriptif - Fitur Kategorikal:")
for col in categorical_features:
    print(f"\nFitur: {col}")
    print("Jumlah unik:", df[col].nunique())
    print("Top 10 terbanyak:\n", df[col].value_counts().head(10))
```

```
Statistik Deskriptif - Fitur Kategorikal:

Fitur: brand
Jumlah unik: 310
Top 10 terbanyak:
 brand
SEPHORA COLLECTION    492
CLINIQUE              211
TOM FORD              150
tarte                 143
Kiehl's Since 1851    122
Dior                  118
Fresh                 108
Lancôme               104
Bumble and bumble      99
MAKE UP FOR EVER       95
Name: count, dtype: int64

Fitur: category
Jumlah unik: 142
Top 10 terbanyak:
 category
Perfume                     620
Moisturizers                398
Face Serums                 334
Value & Gift Sets           241
Face Wash & Cleansers       225
Face Masks                  218
Hair Styling Products       213
Rollerballs & Travel Size   212
Face Brushes                176
Eye Creams & Treatments     171
Name: count, dtype: int64
```

4. Exploratory Data Analysis (EDA)

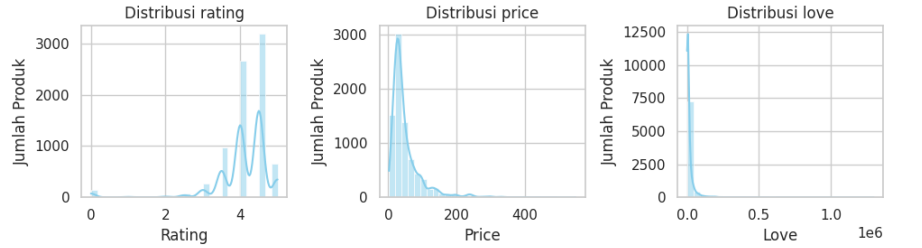
Karena dataset ini terdiri dari berbagai jenis fitur (numerik, kategorikal, dan boolean), maka analisis hubungan antar fitur dilakukan secara bertahap berdasarkan tipe datanya.

a. Distribusi

Jika memahami data dengan baik, dapat diketahui bahwa ada 3 fitur numerik utama yang harus dilakukan eksplorasi datanya, yaitu fitur rating, price, dan love. Ketiga fitur tersebut memiliki nilai numerik yang bervariasi, dan relevan untuk menjawab

pertanyaan umum terkait dataset ini, seperti: “Produk seperti apakah yang paling disukai pelanggan?”, “Apakah harga produk berpengaruh terhadap rating pelanggan?”, dan pertanyaan lainnya. Oleh karena itu, penting untuk melihat visualisasi distribusi diantara ketiga fitur tersebut.

```
# Set style
sns.set(style="whitegrid")

# Buat histogram untuk 3 fitur numerik utama
features = ['rating', 'price', 'love']
plt.figure(figsize=(10, 3))

for i, col in enumerate(features, 1):
    plt.subplot(1, 3, i)
    sns.histplot(df[col], kde=True, bins=30, color='skyblue')
    plt.title(f'Distribusi {col}')
    plt.xlabel(col.capitalize())
    plt.ylabel('Jumlah Produk')

plt.tight_layout()
plt.show()
```



Berdasarkan distribusi numerik fitur utama dataset, terlihat bahwa:

- ***rating*** sangat terkonsentrasi di angka 4–5, dengan puncak di 4.0 dan 4.5. Artinya, mayoritas produk dinilai positif oleh pengguna, dan menunjukkan bahwa produk Sephora memiliki persepsi kualitas yang baik.
- ***price*** memiliki distribusi sangat miring ke kanan, yaitu berada di rentang harga rendah hingga menengah. Artinya produk Sephora umumnya memiliki harga yang terjangkau.
- ***love*** juga memiliki distribusi sangat miring ke kanan, yang berarti popularitas produk tidak merata, atau hanya beberapa produk saja yang sangat dominan disukai pelanggan. Namun hal ini bisa dicari penyebabnya saat dilakukan analisis bersama fitur lainnya.

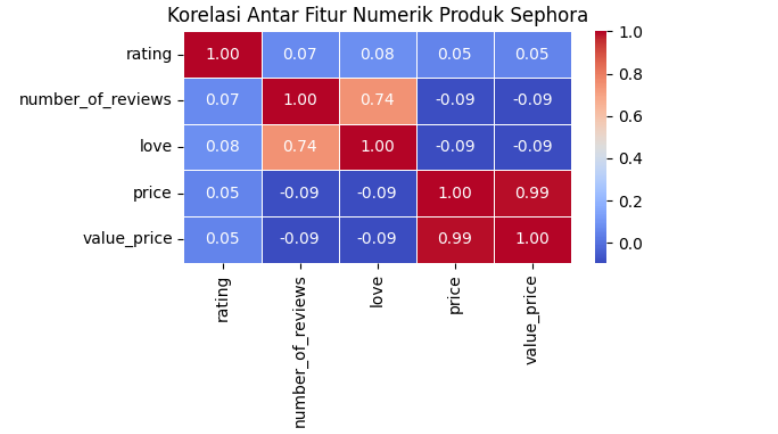
b. Korelasi

Untuk mengetahui hubungan antar fitur-fitur numerik secara keseluruhan, dapat dilihat melalui visualisasi heatmap.

Berdasarkan analisis korelasi antar fitur numerik, hanya ditemukan korelasi yang cukup kuat antara *number_of_reviews* dan *love*. Hal ini menunjukkan bahwa semakin banyak review yang dimiliki sebuah produk, semakin besar pula kemungkinan produk tersebut mendapatkan banyak “love” dari pengguna.

Sementara itu, fitur-fitur lain seperti *price*, dan *rating* tidak menunjukkan hubungan yang signifikan dengan fitur numerik lainnya.

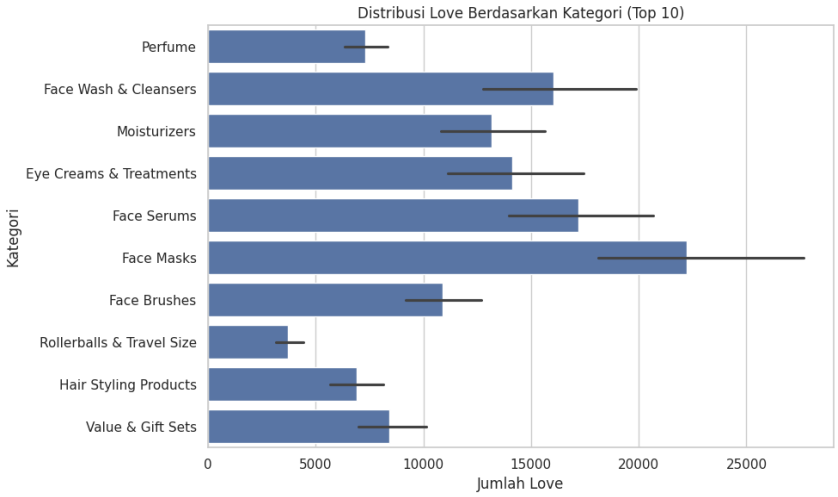
```
import seaborn as sns
import matplotlib.pyplot as plt

# Hitung korelasi
correlation_matrix = df[numerical_features].corr()

# Tampilkan heatmap
plt.figure(figsize=(6, 4))
sns.heatmap(correlation_matrix, annot=True, fmt=".2f", cmap="coolwarm", linewidths=0.5)
plt.title("Korelasi Antar Fitur Numerik Produk Sephora")
plt.tight_layout()
plt.show()
```



c. Hubungan Antara 2 Variabel

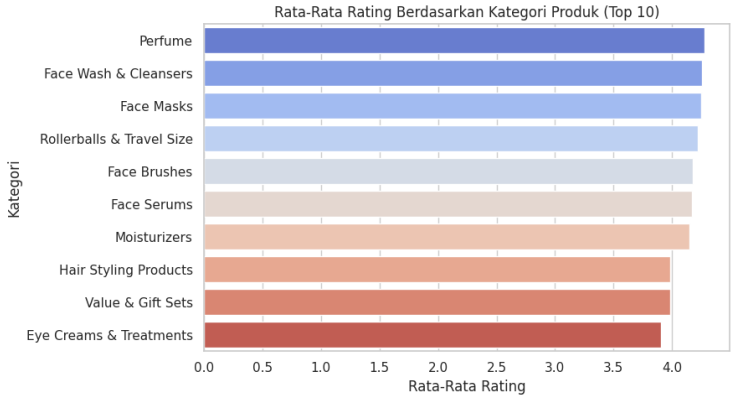
- `category` vs `love`

  Dari visualisasi dapat disimpulkan bahwa kategori-kategori skincare seperti Face Masks, Face Serums, dan Face Wash & Cleansers memiliki rata-rata love yang tinggi, menunjukkan bahwa produk perawatan wajah merupakan kategori paling disukai oleh pengguna Sephora. Sementara itu, kategori seperti Rollerballs & Travel Size dan Perfume memiliki rata-rata love yang lebih rendah meskipun secara jumlah produk cukup tinggi.



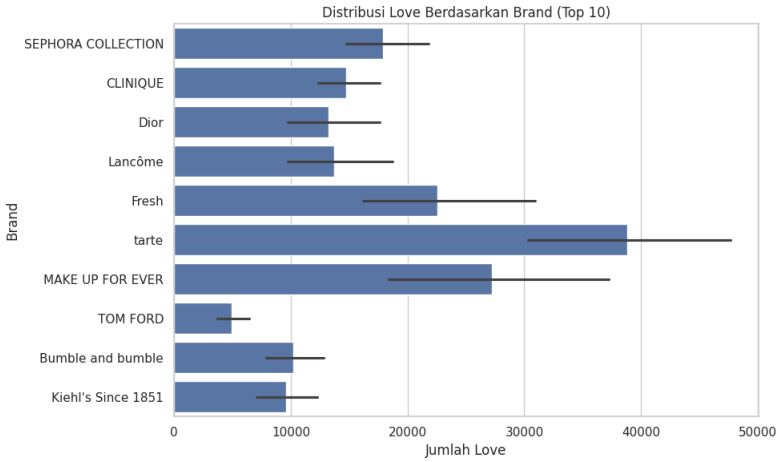
- `category` vs `ratings`

  Analisis rata-rata rating menunjukkan bahwa produk dengan penilaian terbaik oleh pelanggan berasal dari kategori Perfume, Face Wash & Cleansers, serta Face Masks. Sementara itu, kategori seperti Eye Creams & Treatments dan Value & Gift Sets memiliki rata-rata rating yang sedikit lebih rendah, meskipun masih tergolong tinggi (sekitar 4.0). Secara keseluruhan, produk di Top 10 kategori ini dinilai sangat baik oleh pengguna.



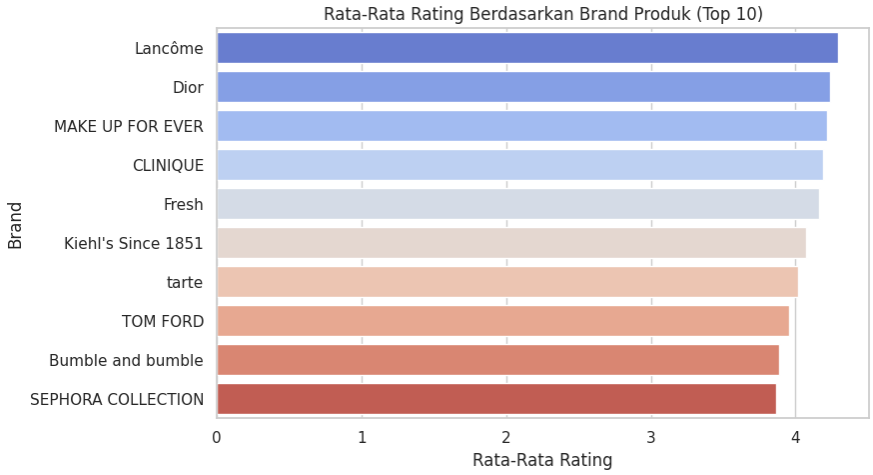
- `brand` vs `love`

  Berdasarkan analisis rata-rata love pada 10 brand dengan jumlah produk terbanyak, ditemukan bahwa brand Tarte mendominasi dalam hal popularitas di mata pengguna. Di sisi lain, meskipun SEPHORA COLLECTION memiliki jumlah produk terbanyak, tingkat love-nya masih kalah dibandingkan brand lain seperti MAKE UP FOR EVER dan Fresh. Hal ini menunjukkan bahwa brand dengan banyak produk tidak selalu menjadi yang paling disukai.



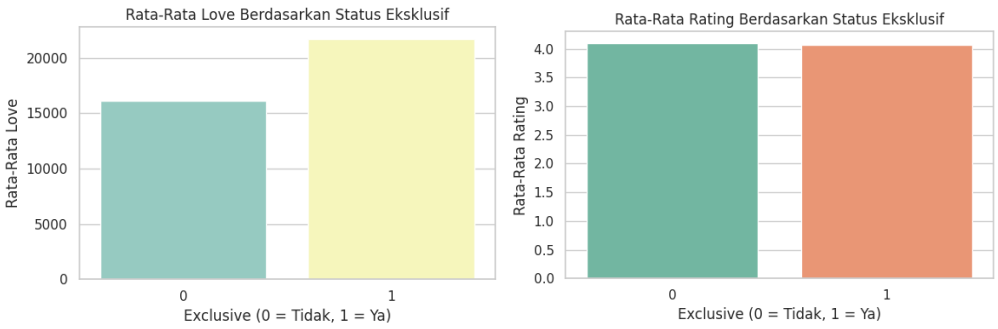
- `brand` vs `ratings`

  Analisis terhadap rata-rata rating menunjukkan bahwa brand-brand premium seperti Lancôme, Dior, dan MAKE UP FOR EVER mendapat penilaian terbaik dari pengguna. Sebaliknya, SEPHORA COLLECTION — meskipun memiliki jumlah produk terbanyak — justru menempati posisi terendah dalam hal rata-rata rating. Ini menunjukkan bahwa persepsi kualitas pelanggan terhadap suatu brand tidak selalu sejalan dengan banyaknya jumlah produk atau tingkat popularitasnya.



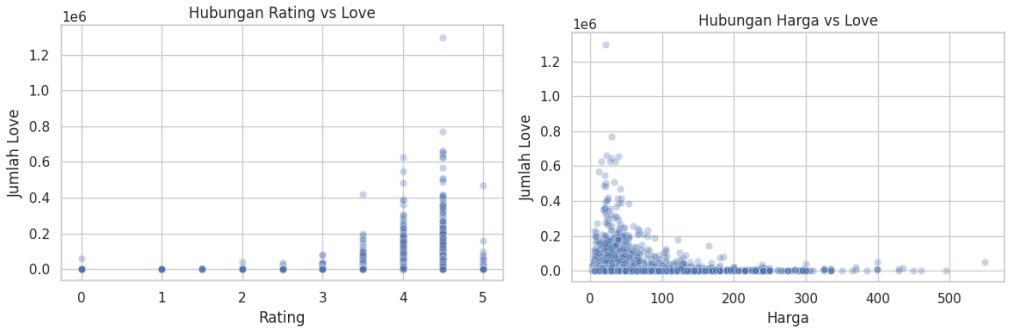
- `exclusive` vs `rating` dan `exclusive` vs `love`

  Berdasarkan analisis status eksklusif produk, terlihat bahwa produk eksklusif cenderung lebih populer dibanding produk non-eksklusif, ditunjukkan oleh rata-rata jumlah love yang lebih tinggi. Namun, hal ini tidak berlaku pada persepsi kualitas, karena rating rata-rata kedua kelompok produk ini relatif serupa. Yang berarti, status eksklusif lebih memengaruhi preferensi atau minat pelanggan, bukan penilaian terhadap kualitas produk itu sendiri.



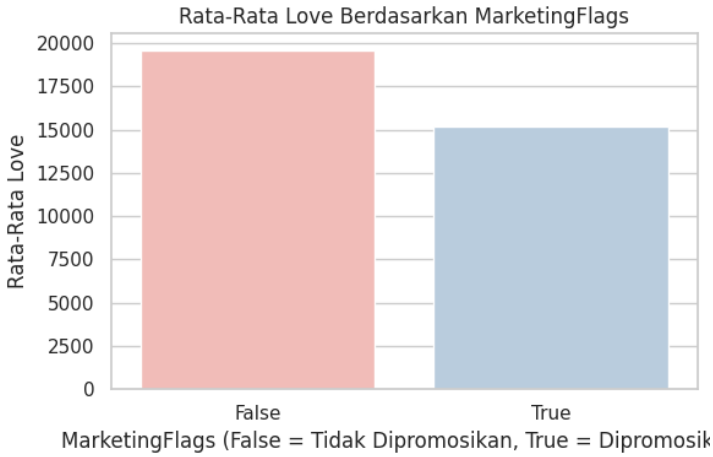
- `rating` vs `love` dan `price` vs `love`

  Produk dengan rating tinggi cenderung lebih disukai oleh pengguna, walaupun tidak semua produk populer punya rating bagus. Selain itu, produk dengan harga murah hingga sedang justru lebih populer dibandingkan produk yang mahal. Jadi, harga dan rating memang berpengaruh, tapi bukan satu-satunya faktor yang menentukan popularitas.



- `MarketingFlags` vs `love`

  Berdasarkan hasil analisis, rata-rata jumlah love pada produk yang tidak dipromosikan ternyata lebih tinggi dibandingkan produk yang dipromosikan. Hal ini menunjukkan bahwa promosi tidak selalu menjamin popularitas produk, dan ada kemungkinan bahwa produk yang populer memang sudah disukai tanpa perlu didorong oleh strategi pemasaran khusus.



5. Pelatihan Model dan Evaluasi

Setelah melakukan eksplorasi data dan visualisasi, dilakukan pemodelan machine learning untuk menjawab pertanyaan:

"Apakah kita bisa memprediksi apakah suatu produk akan populer atau tidak berdasarkan informasi dasarnya?"

Langkah ini bertujuan untuk menunjukkan bahwa data historis bisa digunakan untuk memprediksi potensi suatu produk, serta memberikan wawasan tambahan yang actionable, terutama dalam pengambilan keputusan pemasaran, penempatan produk, atau rekomendasi.

Tahapan dalam Pembangunan Model ini adalah sebagai berikut:

a. Menentukan Target Klasifikasi

   Dibuat target klasifikasi is_popular berdasarkan nilai love > 10.000, sebagai indikator bahwa produk dapat dianggap populer oleh pengguna.

b. Memilih Fitur

   Fitur yang digunakan adalah rating, price, number of reviews, dan exclusive. Fitur ini dipilih karena tersedia di kedua file (train.csv, dan test.csv), tidak memerlukan preprocessing tambahan, dan juga relevan terhadap konteks analisa ini

c. Memilih dan Melatih Model

   Model yang digunakan adalah Random Forest Classifier, karena mudah digunakan dan proses pelatihannya cepat, serta bisa menangani data numerik dan biner tanpa banyak preprocessing.

d. Mengevaluasi Model

   Model dievaluasi pada data training karena test.csv tidak memiliki label. Hasil evaluasi model menunjukkan bahwa model mampu mempelajari pola dengan baik dari fitur fitur yang tersedia serta dapat memprediksi produk populer dengan sangat akurat.

```
Evaluasi Model di Data Train:
              precision    recall  f1-score   support

       False       0.99      0.99      0.99      5214
        True       0.99      0.99      0.99      2786

    accuracy                           0.99      8000
   macro avg       0.99      0.99      0.99      8000
weighted avg       0.99      0.99      0.99      8000

Confusion Matrix:
[[5187   27]
 [  30 2756]]
```

6. Kesimpulan

Seluruh hasil analisa diatas memebrikan berbagai wawasan berbasis data (bukan asumsi). Semua prosesnya menunjukkan bahwa analisis deskriptif bisa dilanjutkan hingga ke insight prediktif. Model yang dibangun menjadi bagian penting dalam menunjukkan bahwa data historis memiliki nilai strategis jika diolah menggunakan pendekatan machine learning yang tepat. Meski hanya model sederhana, hasilnya dapat dimanfaatkan dalam berbagai aspek pengambilan keputusan bisnis seperti, meningkatkan efisiensi promosi, membantu pengambilan stock, serta riset pengembangan produk baru dari ciri ciri produk populer yang dijadikan sebagai acuan.

Link Google Colab:
https://colab.research.google.com/drive/1lSvJQSPnnLApsGkhSHCLicz8QPFWkGHI?usp=sharing