Rakamin Academy

Artha Data - Kelompok 1

Anggota Kelompok:

1. Yoga Aprila
2. Zaima Syarifa Asshafa
3. Muhammad Fauzi Fayyad
4. Moch Siswan Afandi
5. Faris Rahmatullah
6. Deni Yuniawan
7. Rafa Kamila
8. Nijma Fuaiyida Hanum

EDA & DATA PRE-PROCESSING

1. Descriptive Statistics
2. Univariate Analysis
3. Multivariate Analysis
4. Data Cleansing
5. Feature Engineering

Rakamin Academy

Dataset (

Rakamin Academy

Dataset : **Product Exclusive Classification**

Deskripsi: Memprediksi apakah suatu produk eksklusif atau tidak berdasarkan fitur yang tersedia

id	brand	category	rating	number_of_reviews	love	price	value_price	exclusive
50	SEPHORA COLLECTION	no category	5	46	0	50	50	0
304	SEPHORA COLLECTION	no category	0	0	0	50	50	0
404	SEPHORA COLLECTION	no category	0	0	0	50	50	0
443	SEPHORA COLLECTION	no category	0	0	0	50	50	0
463	SEPHORA COLLECTION	no category	0	0	0	50	50	0

1. Descriptive Statistics

1. Kesesuaian tipe data

   Berdasarkan informasi diatas bahwa 9 fitur pada dataset ini telah sesuai tipe datanya. Terdapat 5 fitur bertipe float (rating, number_of_reviews, love, price, value_price), 2 fitur bertipe int (id, exclusive), 2 fitur bertipe object (brand, category)

2. Missing value

   Missing value terdapat pada kolom category dengan jumlah 13, rating dengan jumlah 95, number_of_review dengan jumlah 9, love dengan jumlah 34, price dengan jumlah 8, value_price dengan jumlah 17.

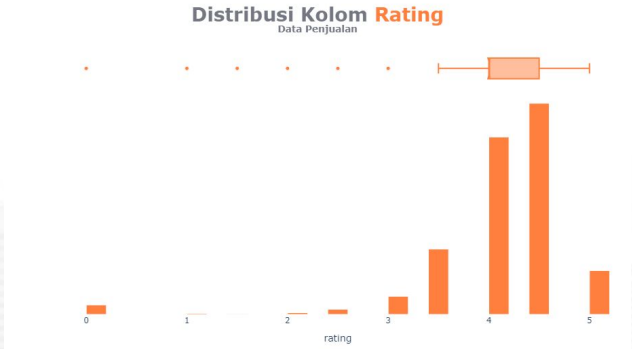
3. Kolom statistics

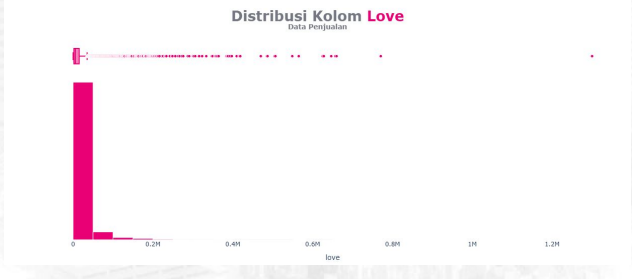
   Terdapat keanehan pada kolom number of review dan love di mean dan std, dimana **std yang sangat besar dibandingkan mean**. Hal ini dikarenakan, data yang digunakan sangat bervariasi sehingga std besar.

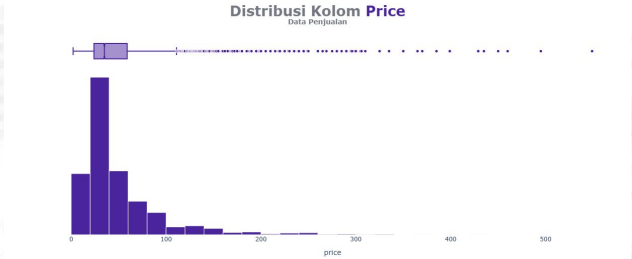
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8000 entries, 0 to 7999
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 8000 non-null   int64
 1   brand              8000 non-null   object
 2   category           7987 non-null   object
 3   rating             7905 non-null   float64
 4   number_of_reviews  7991 non-null   float64
 5   love               7966 non-null   float64
 6   price              7992 non-null   float64
 7   value_price        7983 non-null   float64
 8   exclusive          8000 non-null   int64
dtypes: float64(5), int64(2), object(2)
memory usage: 562.6+ KB
```

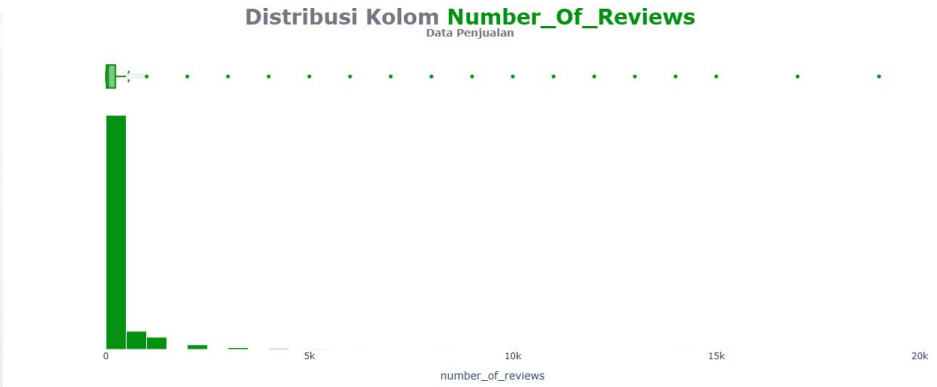
```
Banyak missing value tiap kolom adalah

id                   0
brand                0
category            13
rating              95
number_of_reviews    9
love                34
price                8
value_price         17
exclusive            0
dtype: int64
```

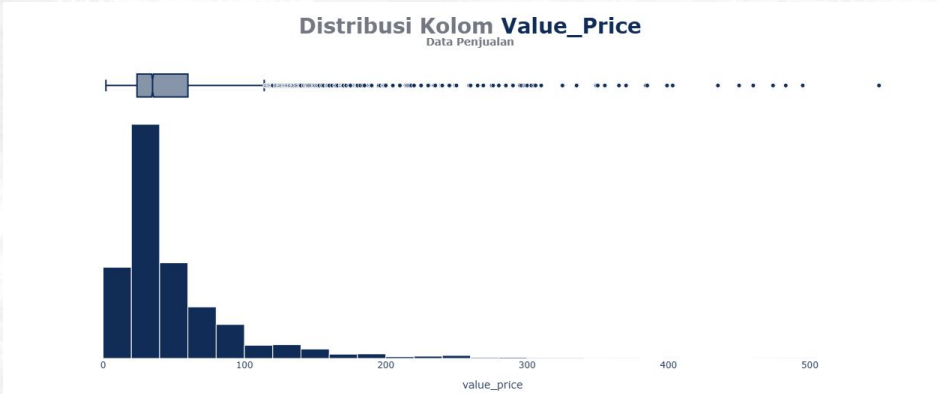
2. Univariate Analysis









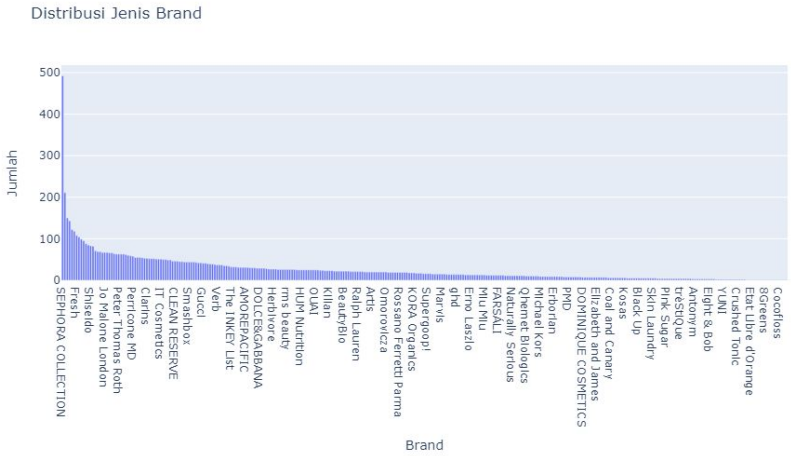


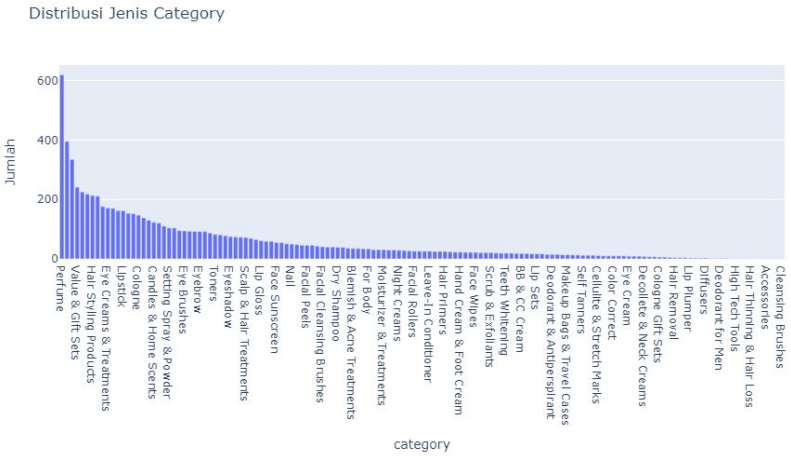
2. Univariate Analysis

- Kolom Rating :
  - Kolom Rating menunjukkan data yang tersebar dari nilai 0 hingga 5. Sebagian besar nilai rating berada pada nilai 4 dan 5. Terdapat sedikit sekali data pada rating 0, 1, 2, dan 3.
  - Distribusi right-skewed : Mayoritas pelanggan memberikan rating yang tinggi (4 dan 5) terhadap produk yang dijual.
  - Rating yang rendah sangat jarang terjadi, menunjukkan kepuasan pelanggan yang tinggi secara umum.
  - Ada beberapa outliers pada nilai rating yang rendah, tetapi tidak banyak.
- Kolom number_of_reviews :
  - Distribusi right-skewed : Mayoritas produk menerima sedikit ulasan.
  - Beberapa produk populer menerima banyak ulasan, menunjukkan tingkat perhatian atau popularitas yang tinggi di antara pelanggan.
  - Ada banyak outliers pada jumlah ulasan yang tinggi, menunjukkan bahwa beberapa produk sangat populer dan menarik banyak ulasan.
- Kolom love :
  - Pada kolom love didapatkan hasil visualisasi right-skewed dimana banyaknya data ada di rentang kecil.
  - Outliers mencakup nilai "Love" yang mencapai lebih dari 1,2 juta, menunjukkan adanya produk yang sangat populer atau mendapat perhatian besar dari pengguna.
- Kolom price :
  - Pada kolom price didapatkan hasil visualisasi right-skewed dimana banyaknya data ada di rentang kecil (harga rendah), tetapi ada juga data yang harganya mahal.
  - Banyak outliers dengan nilai yang sangat tinggi, menunjukkan ada beberapa produk dengan harga yang jauh lebih tinggi dari rata-rata.
- Kolom value_price :
  - Distribusi right-skewed : Kebanyakan produk memiliki nilai Value_Price yang rendah.
  - Banyak outliers dengan nilai yang sangat tinggi, menunjukkan ada beberapa produk dengan Value_Price yang jauh lebih tinggi dari rata-rata.

2. Univariate Analysis

Distribusi kolom kategorikal



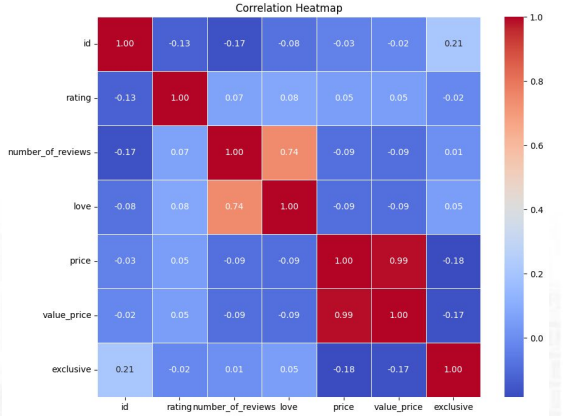


Pada kolom kategorikal brand dan category memiliki jumlah ketegori yang sangat banyak, yaitu terdapat 310 brand dan 142 jenis category produk.

3. Multivariate Analysis

I. Bagaimana korelasi antara masing-masing feature dan label. Kira-kira feature mana saja yang paling relevan dan harus dipertahankan?

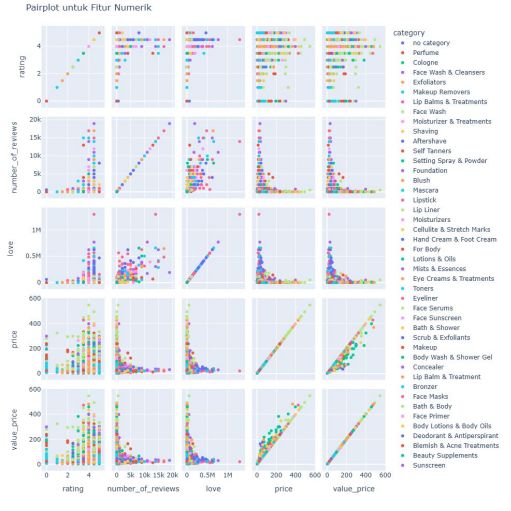
- Kolom exclusive memiliki **korelasi negatif yang rendah** dengan rating (-0.02) dan number_of_reviews (0.01).
- Korelasi dengan fitur lain seperti love, price, dan value_price hampir nol, menunjukkan **tidak ada hubungan linier** yang signifikan.
- Tidak ada fitur yang menunjukkan korelasi kuat dengan label exclusive, jadi tidak ada fitur yang secara khusus perlu dipertahankan berdasarkan korelasi ini saja.



II. Bagaimana korelasi antara feature, apakah ada pola yang menarik? Apa yang perlu dilakukan terhadap feature itu?

A. Correlation Heatmap
   1. price dan value_price memiliki korelasi yang sangat tinggi (0.99) menunjukkan bahwa satu nilai hampir dapat diprediksi dari yang lain.
   2. Korelasi lainnya antara fitur-fitur lain sangat rendah, menunjukkan tidak ada hubungan linier yang kuat di antara mereka.

B. PairPlot
   1. Menunjukkan beberapa pola distribusi dan hubungan antara fitur-fitur numerik.
   2. Tidak ada pola menarik lainnya yang terlihat selain korelasi tinggi antara price dan value_price.



3. Multivariate Analysis

	Variable	VIF
0	id	13.690906
1	rating	13.845562
2	number_of_reviews	2.487240
3	love	2.591358
4	price	70.572847
5	value_price	69.457074
6	exclusive	1.462115

C. The Variance Inflation Factor (VIF)

1. Fitur price dan value_price menunjukkan multikolinearitas yang sangat tinggi dan harus diprioritaskan untuk dihapus atau ditransformasi untuk meningkatkan kinerja model.
2. id dan rating juga menunjukkan multikolinearitas tinggi dan mungkin perlu disesuaikan atau dihapus.
3. Fitur lainnya memiliki multikolinearitas yang dapat diterima atau rendah dan kemungkinan tidak memerlukan tindakan segera.

4. Data Cleansing

Rakamin Academy

1. Handle missing values

- Pada kolom category menggunakan metode ‘Imputasi’, diganti dengan nilai **modus**
- Pada kolom number_of_reviews, love, price, rating, dan value_price menggunakan metode ‘imputasi’, diganti dengan nilai **median** karena distribusi data yang sangat miring.

```
from sklearn.impute import SimpleImputer

# Pastikan kolom 'category' ada dalam DataFrame
category_imputer = SimpleImputer(strategy='most_frequent')

# Imputasi nilai yang hilang dalam kolom 'category'
product_exc['category'] = category_imputer.fit_transform(product_exc[['category']]).ravel()

# Imputasi 'number_of_reviews', 'love', 'price', 'value_price' dengan median
median_imputer = SimpleImputer(strategy='median')
product_exc[['number_of_reviews', 'love', 'price','rating', 'value_price']] =
median_imputer.fit_transform(product_exc[['number_of_reviews', 'love', 'price','rating', 'value_price']])
```

2. Handle duplicate data

Tidak terdapat data duplikat pada dataset

```
# pengecekan data duplikat
duplicated_data = product_exc.duplicated().any()

if duplicated_data == True:
    print('Terdapat data duplikat')
else:
    print('Tidak terdapat data duplikat')
```

```
Tidak terdapat data duplikat
```

4. Data Cleansing

Rakamin Academy

3. Handle outliers

Dilakukan penghapusan nilai outlier dengan menggunakan **‘metode IQR’** dari beberapa kolom yaitu number_of_reviews, love, price, dan value_price.

```
columns = ['rating', 'number_of_reviews', 'love', 'price', 'value_price']

for col in columns:
    # Hitung kuartil 1 (Q1), kuartil 3 (Q3), dan IQR
    Q1 = product_exc[col].quantile(0.25)
    Q3 = product_exc[col].quantile(0.75)
    IQR = Q3 - Q1

    # Identifikasi batas bawah dan atas untuk outlier
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    # Identifikasi outlier
    outliers = product_exc[(product_exc[col] < lower_bound) | (product_exc[col] > upper_bound)].index

    # Hapus outlier dari dataset
    product_exc = product_exc.drop(outliers)
```

4. Feature transformation

Dengan menggunakan transformasi log dapat mengurangi kemiringan dari distribusi kolom number_of_reviews, love, price, dan value price.

```
# menggunakan transformasi log untuk mengurangi kemiringan distribusi
product_exc['number_of_reviews'] = product_exc['number_of_reviews'].apply(lambda x: np.log1p(x))
product_exc['love'] = product_exc['love'].apply(lambda x: np.log1p(x))
product_exc['price'] = product_exc['price'].apply(lambda x: np.log1p(x))
product_exc['value_price'] = product_exc['value_price'].apply(lambda x: np.log1p(x))
```

4. Data Cleansing

5. **Feature encoding**

   Dengan menggunakan metode one hot encoding dapat merubah kolom kategorik menjadi numerik. Pemilihan one hot encoding karena fitur memiliki banyak kategori dan bukan ordinal, sehingga lebih cocok menggunakan one hot encoding.

6. **Handle class imbalance**

   Kolom exclusive menunjukkan ketidakseimbangan kelas dengan lebih banyak produk yang tidak eksklusif, maka untuk mengatasi masalah Imbalance Class dapat dengan menggunakan Oversampling pada kelas minoritas menggunakan SMOTE (Synthetic Minority Over-sampling Technique) jika diperlukan dalam analisis lebih lanjut.

5. Feature Engineering

1. **Feature selection**

   Menghapus fitur **value_price** dan **id**, karena fitur value_price dan price memiliki korelasi yang sangat tinggi (0.99) dan fitur id dirasa kurang relevan sebagai data modelling

2. **Feature extraction** (membuat feature baru dari feature yang sudah ada)

   - **log_reviews_per_rating** -> Fitur baru yang menghitung log dari ulasan per rating dengan Rumus : log_reviews_per_rating = log(number_of_reviews) / rating
   - **price_to_rating** -> Fitur baru yang mengukur rasio harga terhadap rating dengan Rumus :

     price_to_rating = price / rating.
   - **is_high_end** -> Fitur biner baru yang menunjukkan apakah produk memiliki harga di atas rata-rata dengan Rumus : (is_high_end = 1 jika price > mean(price)).

3. **Additional 4 feature**

   - Popularity Index : Indeks popularitas berdasarkan kombinasi love, number_of_reviews, dan rating.
   - Time on Market : Fitur yang menunjukkan berapa lama produk telah berada di pasar.
   - Seasonal Demand : Fitur yang menunjukkan apakah produk cenderung lebih populer di musim-musim tertentu.
   - Customer Loyalty : Fitur yang menunjukkan tingkat loyalitas pelanggan berdasarkan frekuensi pembelian ulang.

Kesimpulan

1. Kualitas Data :

   Data memiliki beberapa missing values pada kolom penting seperti kategori, rating, jumlah ulasan, love, harga, dan value_price. Hal ini perlu ditangani sebelum melakukan pemodelan prediktif. Tipe data dan penamaan kolom sudah sesuai, menunjukkan kualitas data yang baik secara umum.

2. Distribusi Data :

   Sebagian besar produk menerima rating tinggi (4 dan 5), menunjukkan kepuasan pelanggan yang tinggi. Mayoritas produk memiliki jumlah ulasan yang rendah, tetapi ada beberapa produk yang sangat populer dan menerima banyak ulasan. Distribusi harga dan value_price menunjukkan bahwa kebanyakan produk memiliki harga dan nilai yang rendah, tetapi ada beberapa produk dengan harga dan nilai yang sangat tinggi.

3. Korelasi Antar Fitur :

   Terdapat korelasi positif yang kuat antara jumlah ulasan dan love, menunjukkan bahwa produk yang lebih banyak diulas cenderung lebih disukai. Korelasi positif antara harga dan value_price menunjukkan bahwa produk yang lebih mahal cenderung memiliki nilai yang lebih tinggi. Tidak ada korelasi yang kuat antara rating dan fitur lainnya, menunjukkan bahwa rating mungkin tidak dipengaruhi oleh faktor-faktor lain seperti harga atau jumlah ulasan.

Kesimpulan

4. Analisis Multivariat :

   Heatmap korelasi dan pairplot menunjukkan hubungan antar fitur numerik. Analisis kategori menunjukkan distribusi brand dan kategori produk, dengan beberapa brand dan kategori yang lebih dominan daripada yang lain. Rekomendasi untuk Data Pre-processing:

   Penanganan Missing Values: Missing values pada kolom penting perlu ditangani dengan metode yang sesuai, seperti imputasi atau penghapusan baris data. Penanganan Outliers: Outliers pada fitur numerik seperti jumlah ulasan, love, harga, dan value_price perlu ditangani untuk menghindari bias dalam pemodelan. Encoding Fitur Kategorikal: Fitur kategorikal seperti brand dan kategori perlu diubah menjadi bentuk numerik agar dapat digunakan dalam pemodelan.

   Analisis data ini memberikan wawasan berharga tentang karakteristik produk eksklusif. Hasil analisis dapat digunakan untuk mengembangkan strategi pemasaran yang lebih efektif, mengoptimalkan harga produk, dan meningkatkan kepuasan pelanggan. Pemodelan prediktif dapat dilakukan untuk memprediksi popularitas atau nilai produk berdasarkan fitur-fitur yang relevan.

Rakamin Academy

Selamat Mengerjakan!