Kelompok Data 200

PRODUCT ANALYSIS

WEEK 10 - ML PREPARATION

rakamin academy

Start Slide

Our Great Team

Theofilus Arifin

Christofer Bryan N. K.

Ramlan Apriyansyah

Muhammad Iqbal

Hanifah Arrasyidah

Christopher Stephen

Muhammad Rizq N. A.

Ujang Pian

Descriptive Statistics

Data 200

Descriptive Analysis

Data Info

Berdasarkan pengematan yang telah dilakukan,
Semua tipe data sudah sesuai

```
<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
```

Data 200

Descriptive Analysis

Null Values

Terdapat beberapa kolom kosong yaitu:

- category : 13 data kosong
- rating : 95 data kosong
- number of reviews : 9 data kosong
- love : 34 data kosong
- price : 8 data kosong
- value_price : 17 data kosong

Rating memiliki data kosong yang paling banyak.

```
data.isna().sum()
```

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

Data 200

Descriptive Analysis

Keanehan Pada Data

- Nilai max dan min pada rating dan number_of_reviews memiliki jarak yang sangat jauh dari rata-rata, hal ini kemungkinan merupakan nilai outlier

- Descriptive price dan value_price sangat mirip, kedua fitur tersebut kemungkinan memiliki relasi yang sangat kuat

- Nilai max pada price dan value_price sangat jauh dari rata-rata, kemungkinan terpadat suatu product outlier yang harganya sangat mahal dibandingkan produk lainnya.

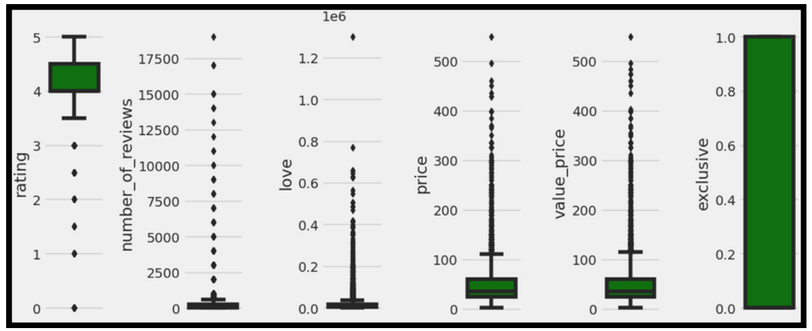
Univariate Analysis

Data 200

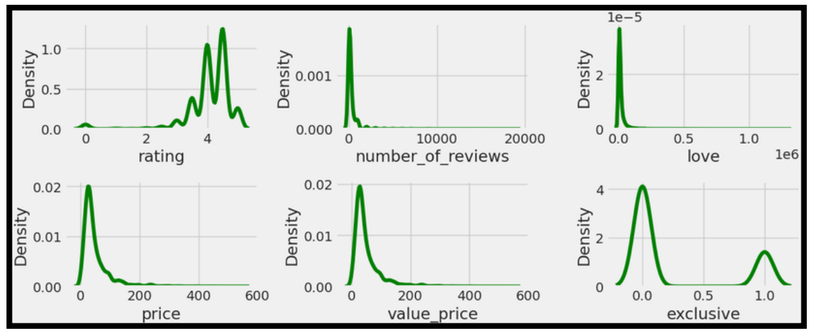
Numeric Data Visualization

Visualization Result

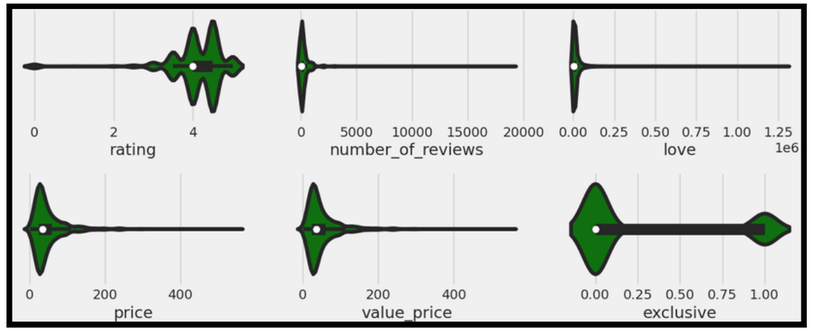
Berikut merupakan hasil visualisation dari fitur numerik yang ada pada data. Visualisasi yang digunakan antara lain adalah boxplot, kde plot, dan violin plot



Boxplot



KDE Plot



Violin Plot

Data 200

Numeric Data Insight

Insight

Result

Rating

Banyak produk dengan rating 0 yang tampaknya seperti outlier. Ini mungkin mengindikasikan produk yang belum mendapatkan rating atau ada alasan lain.

Number of Review

Ada banyak outlier di sebelah kanan, yang menunjukkan beberapa produk dengan jumlah ulasan yang sangat tinggi dibandingkan dengan produk lainnya.

Love

Ada banyak outlier di sebelah kanan, yang menunjukkan beberapa produk dengan jumlah ulasan yang sangat tinggi dibandingkan dengan produk lainnya.

Price & Value Price

Terdapat beberapa produk dengan price dan value price yang sangat tinggi dibandingkan dengan produk lainnya, yang ditunjukkan sebagai outlier di sebelah kanan dari boxplot.

Exclusive

Mayoritas produk tampaknya bukan eksklusif (dengan label 0). Hanya sebagian kecil produk yang eksklusif (dengan label 1).

Data 200

Numeric Data Follow Up

Follow Up Result

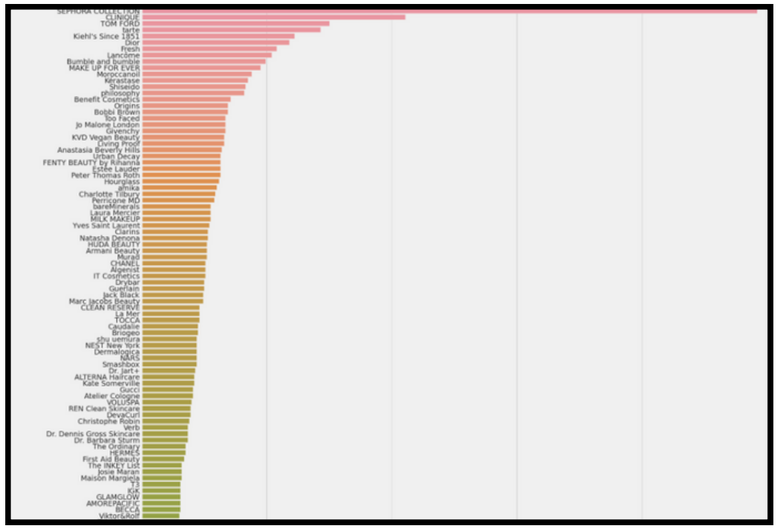
Rating	Number of Review	Love	Price & Value Price	Exclusive
Perlu ditangani produk dengan rating 0. Apakah ini berarti produk tersebut belum mendapatkan rating atau ada alasan lain?	Mengingat distribusi yang sangat skewed, transformasi mungkin diperlukan untuk mengatasi skewness (misalnya, transformasi log).	Mengingat distribusi yang sangat skewed, transformasi mungkin diperlukan untuk mengatasi skewness (misalnya, transformasi log).	Outlier perlu ditangani, dan mungkin diperlukan normalisasi atau standarisasi.	Mengingat ketidakseimbangan label, perlu mempertimbangkan teknik seperti oversampling, undersampling.

Data 200

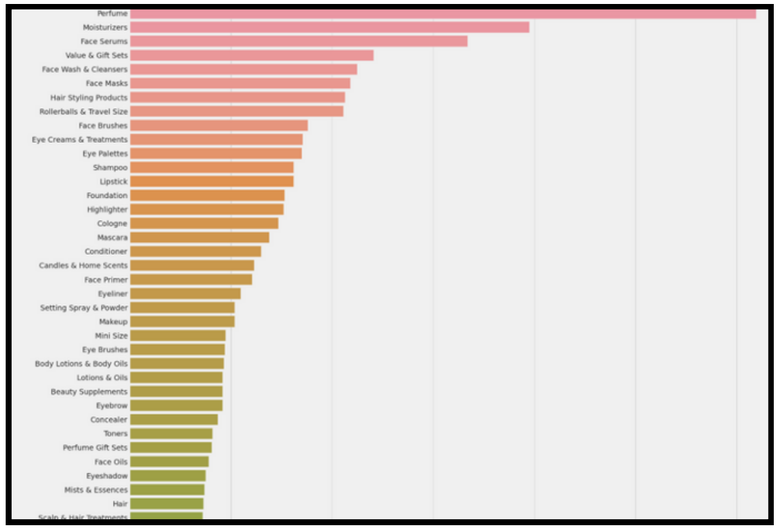
Categorical Data Visualization

Visualization Result

Berikut merupakan hasil visualisation dari fitur categorical yang ada pada data. Visualisasi yang digunakan adalah count plot.



Count Plot Brand



Count Plot Category

Data 200

Numeric Data Insight

Insight

Result

Brand

- Terdapat 310 merek unik.
- Mayoritas produk berasal dari merek "SEPHORA COLLECTION".
- Terdapat beberapa merek lainnya tetapi jumlah produk mereka jauh lebih sedikit dibandingkan dengan "SEPHORA COLLECTION".

Category

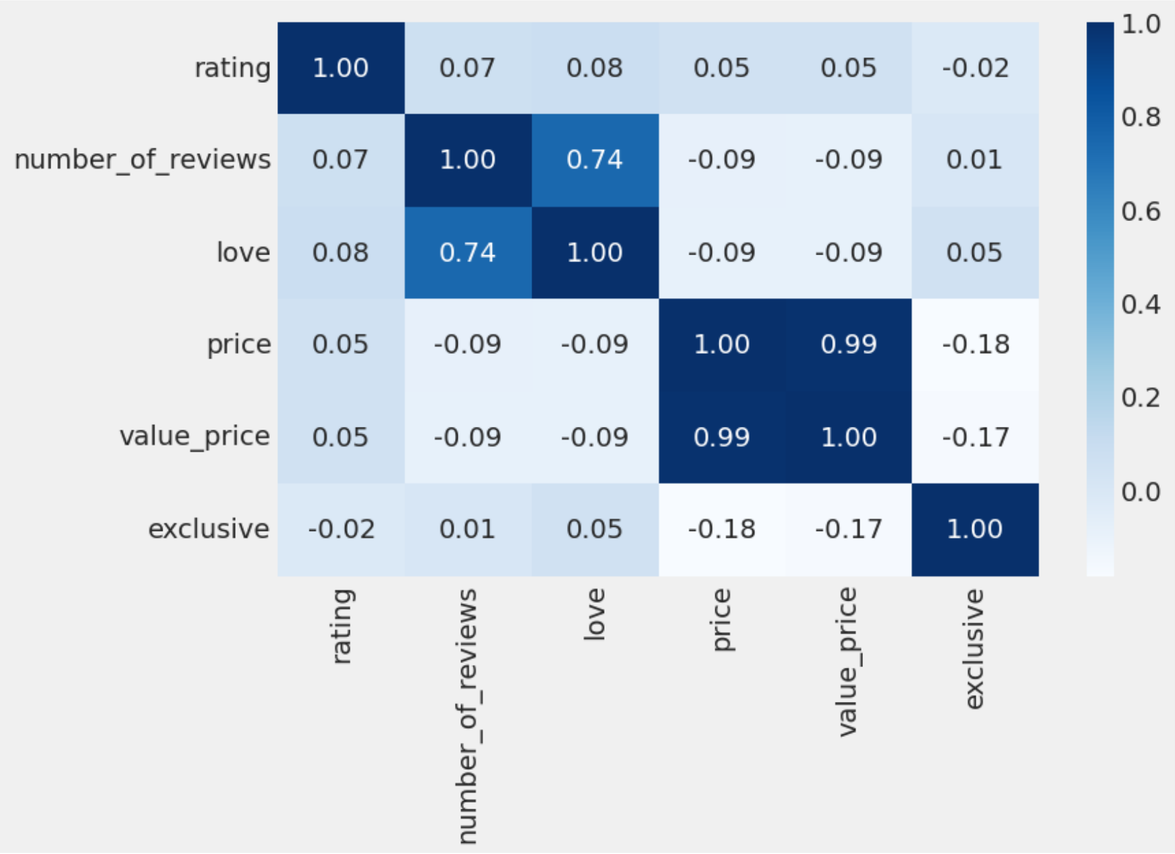
- Terdapat 142 kategori unik
- Mayoritas produk berasal dari merek "Perfumme".
- Terdapat beberapa kategori lainnya tetapi jumlah produk mereka jauh lebih sedikit dibandingkan dengan "Perfumme".

Follow Up

Mengingat sebagian besar produk berasal dari brand dan category tertentu saja mungkin perlu dipertimbangkan untuk mengelompokkan brand dan category jumlah produk yang sangat sedikit ke dalam satu kategori "Lainnya".

Multivariate Analysis

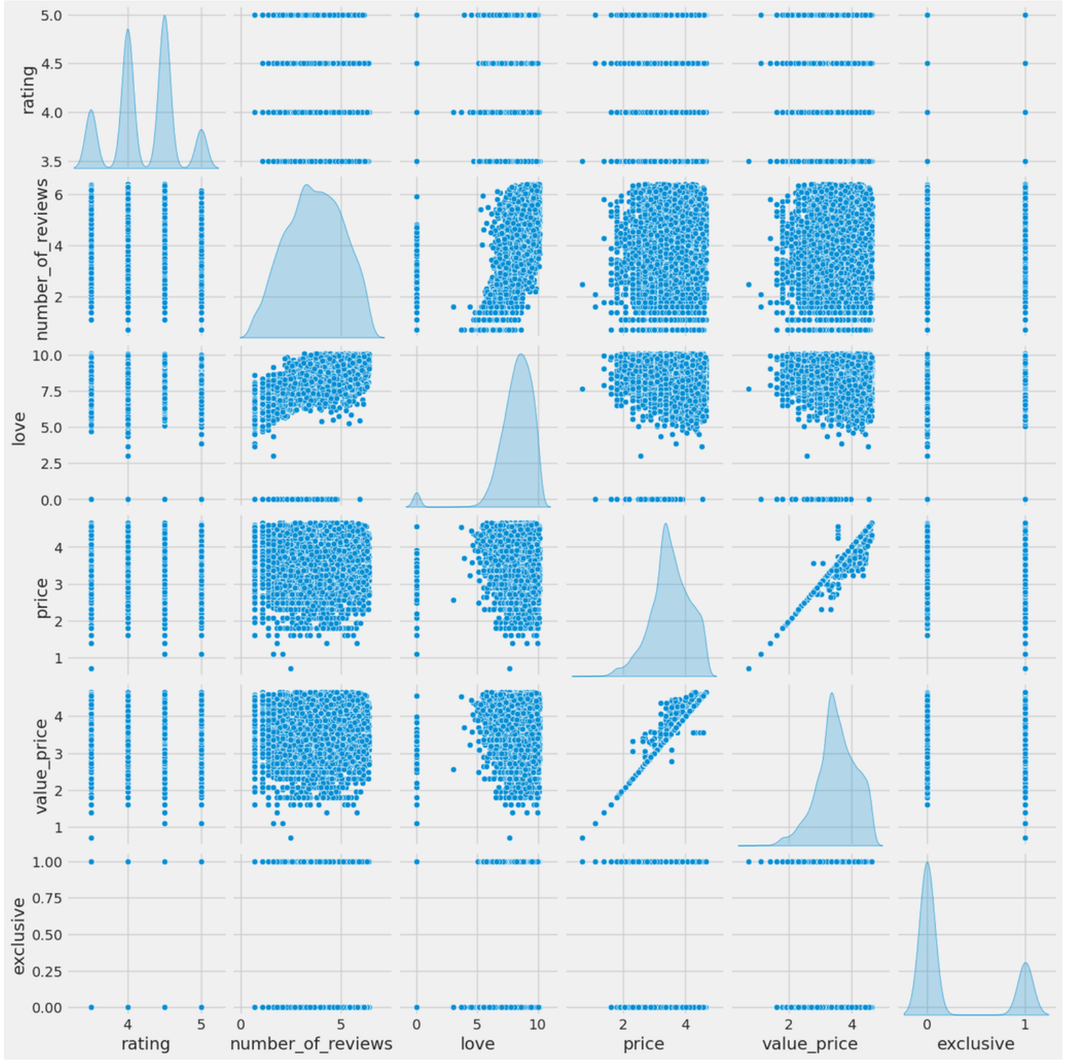
Correlation map
(Heat map)



Dari hasil correlation Heat map dapat dilihat bahwa:

- **number_of_reviews dan love memiliki korelasi yang signifikan hingga 74%**
- **price dan value_price memiliki korelasi yang hampir sama yaitu 99%**

Correlation map
(Pair Plot)



Dari hasil correlation pair plot dapat dilihat bahwa:

- **tidak ada fitur yang memiliki korelasi yang tinggi dengan target(exclusive)**

<u>Follow Up</u>

- **pasangan-pasangan fitur yang memiliki korelasi tinggi ini kemungkinan redundan dan bisa dijadikan satu fitur saja**
- **model linear tidak bisa digunakan karena tidak ada fitur yang memiliki korelasi yang tinggi dengan target (exclusive)**

Data Cleansing

Handling Missing Values

Fitur yang memiliki missing values yaitu **category (13), rating (95), number of reviews (9), love (34), price (8), value price (17).**
Missing values pada fitur category diisi dengan value (unknown), sedangkan pada fitur lain dilakukan penghapusan (drop)

Duplicate Values

Tidak terdapat duplicate pada dataset.

Duplicate Values

Outliers dihilangkan menggunakan rumus IQR.

Feature Transformation

Dilakukan dengan log transformation, dengan tujuan agar distribusi data mendekati normal.

Feature Encoding

Tidak perlu feature encoding karena semua feature sudah numerical (Brand dan Category tidak digunakan)

Class Imbalance

Handling class imbalance dilakukan dengan manual oversampling sehingga class pada target menjadi seimbang.

Feature Engineering

Feature Selection

Value price dan price memiliki korelasi yang sangat tinggi, fitur yang akan dibuang adalah value_price dan yang akan dipertahankan adalah price karena price memiliki korelasi yang lebih tinggi dengan exclusive

Love dan number_of_reviews memiliki korelasi yang cukup tinggi, fitur yang akan dibuang adallah numbber_of_reviews dan yang akan dipertahankan adalah love karena love memiliki korelasi yang lebih tinggi dengan exclusive.

Feature extraction

Di proses ini, kita menambah feature baru dari feature price yang sudah ada. Feature baru ini adalah price_type yang adalah kategori harga produk yang terbagi dalam tiga kelompok, “Cheap”, “Medium”, dan “Expensive.”

Feature tambahan

- mass_produced. Feature ini adalah indikator apakah suatu produk diproduksi secara massal atu tidak. Feature ini bertipe categorical dan mempunyai dua value 0 (False) untuk menandakan tidak diproduksi secara massal dan 1 (True) menandakan produk diproduksi secara massal.
- kota_asal. Feature ini adalah kota tempat produk diproduksi. Feature ini bisa menjelaskan apakah suatu produk eksklusif atau tidak. Contohnya, produk yang diproduksi di kota asal usulnya dapat menandakan bahwa produk tersebut lebih eksklusif.

Feature tambahan

- **bahan_baku**

Jenis bahan baku yang digunakan dalam pembuatan produk juga dapat mempengaruhi tingkat eksklusivitas. Penggunaan bahan baku premium atau langka dapat menandakan produk yang lebih eksklusif dibandingkan dengan produk yang menggunakan bahan baku umum atau murah.

- **limited_edition**

Fitur yang menunjukkan apakah produk merupakan edisi terbatas atau tidak. Produk dengan label "Limited Edition" cenderung dianggap lebih eksklusif karena ketersediaannya terbatas.

THANK YOU

HOMEWORK - ML INTRODUCTION