

UNIVOLATATORS

Machine Learning
Preparation

OUR TEAM

M Arvin Fadriansyah

galih refa

Zulfikar fauzi

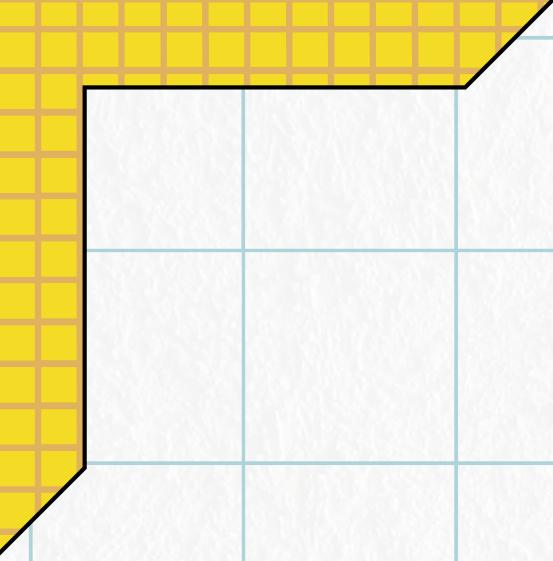
Annisa Sulistyaningsih

M Rizqi Fadhilah

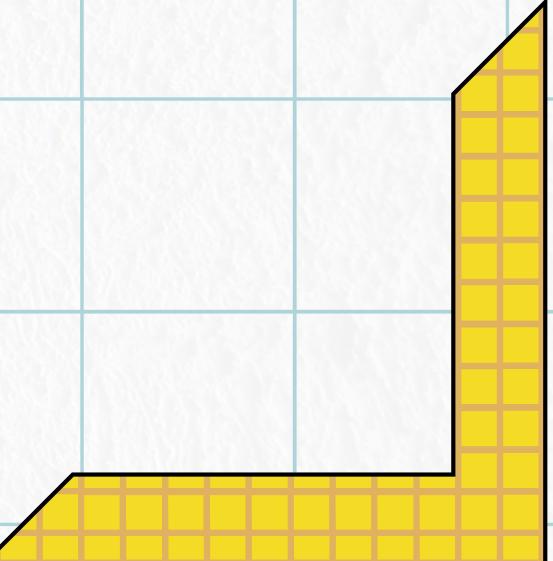
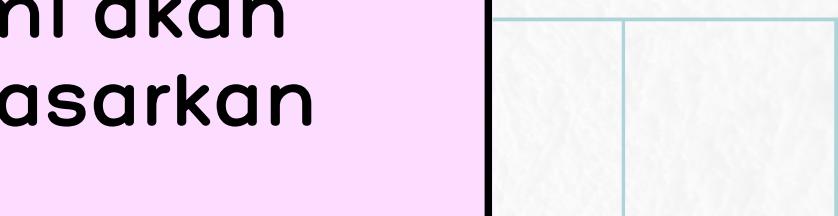
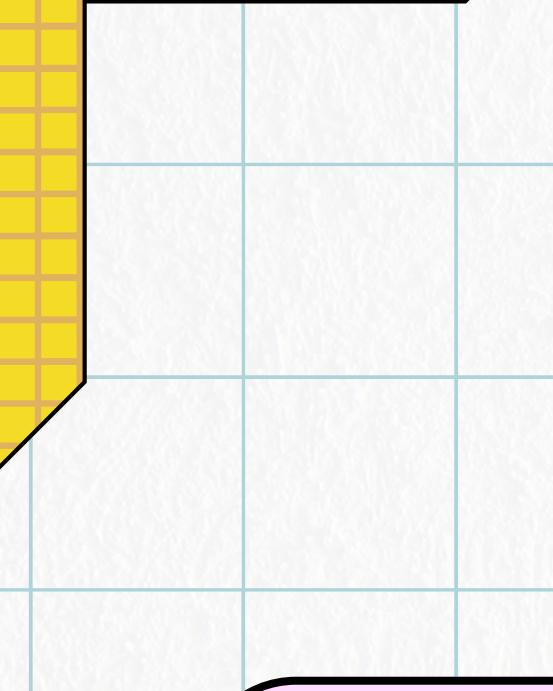
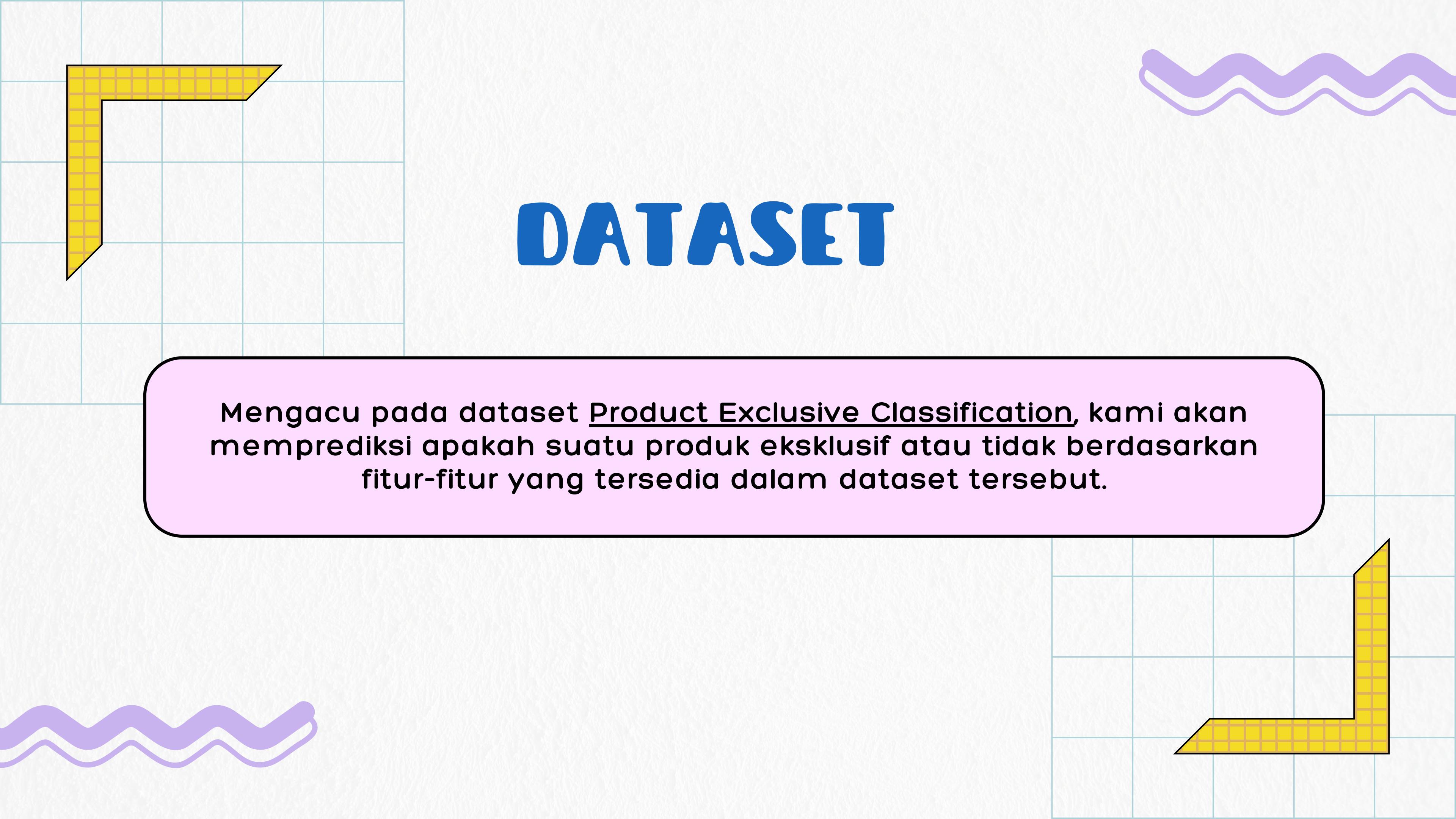
Thufael Bintang Alfattah

Melliza Nastasia Izazi

Niken Mustikaweni



DATASET



Mengacu pada dataset Product Exclusive Classification, kami akan memprediksi apakah suatu produk eksklusif atau tidak berdasarkan fitur-fitur yang tersedia dalam dataset tersebut.

TOPICS



Descriptive
Statistics



Univariate
Analysis



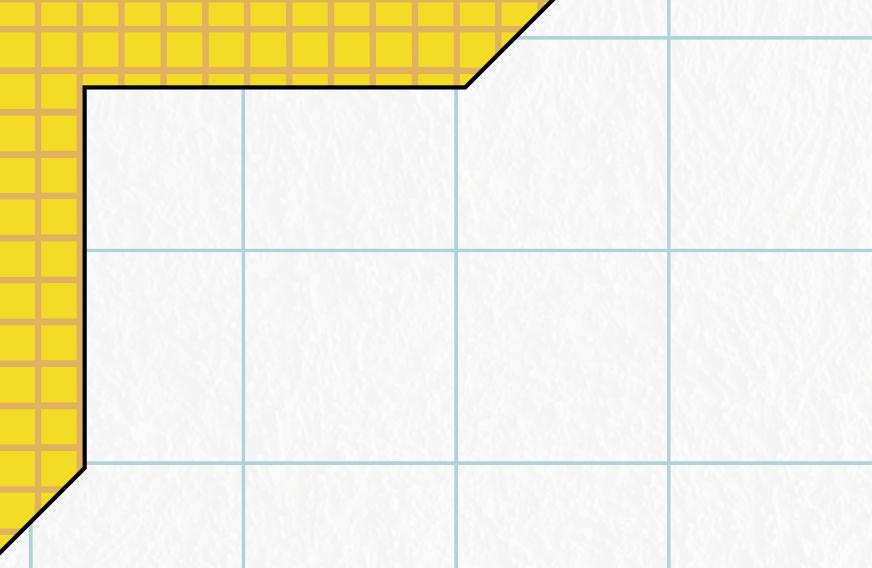
Multivariate
Analysis



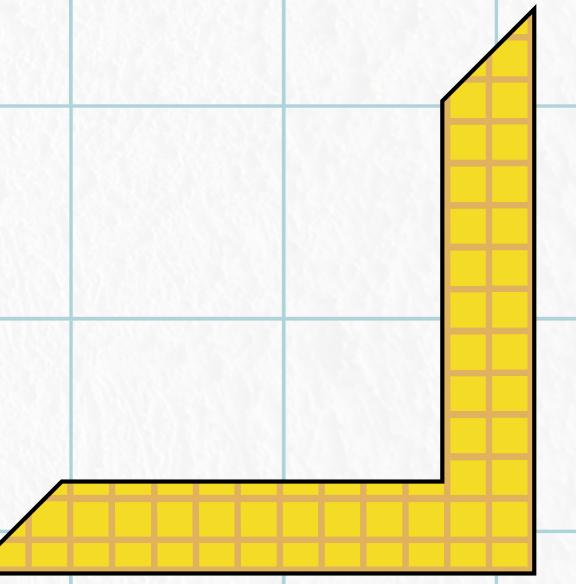
Data
Cleansing



Feature
Engineering



1. DESCRIPTIVE STATISTICS



FINDINGS

Dari data yang disajikan tidak ada yang kurang sesuai dengan tipe datanya.

```
df.info()  
  
-> <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
```



beberapa kolom seperti category, rating, number_of_reviews, love, price, dan value price memiliki nilai kosong.

```
df.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
```

ANALISA KOLOM NUMERICAL

1. Kolom rating, number_of_reviews, love, price, value_price, dan exclusive tidak ada issue pada nilai min/max
2. Semua kolom nilainya masih masuk akal
3. Kolom number_of_reviews ada perbedaan yang cukup janggal pada bagian Mean dan std
4. Kolom love juga memiliki perbedaan yang jauh dari mean dan std.

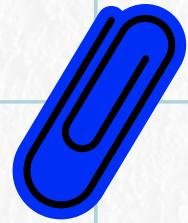
```
# Melakukan pengkategorian
cats = ['id', 'brand', 'category']
num = ['rating', 'number_of_reviews', 'love', 'price', 'value_price', 'exclusive']

Analisa Kolom Numerical

[ ] df[num].describe().T
```

	count	mean	std	min	25%	50%	75%	max
rating	7905.0	4.085136	0.761069	0.0	4.0	4.0	4.5	5.0
number_of_reviews	7991.0	303.574396	931.724460	0.0	14.0	56.0	231.5	19000.0
love	7966.0	17563.958951	44253.391743	0.0	2000.0	5500.0	15300.0	1300000.0
price	7992.0	49.900935	46.864764	2.0	24.0	35.0	59.0	549.0
value_price	7983.0	50.983300	48.473049	2.0	24.0	35.0	60.0	549.0
exclusive	8000.0	0.255875	0.436379	0.0	0.0	0.0	1.0	1.0

ANALISA KOLOM CATEGORY



Kolom brand didominasi oleh Sephora Collection,



Kolom category didominasi oleh kategori Perfume

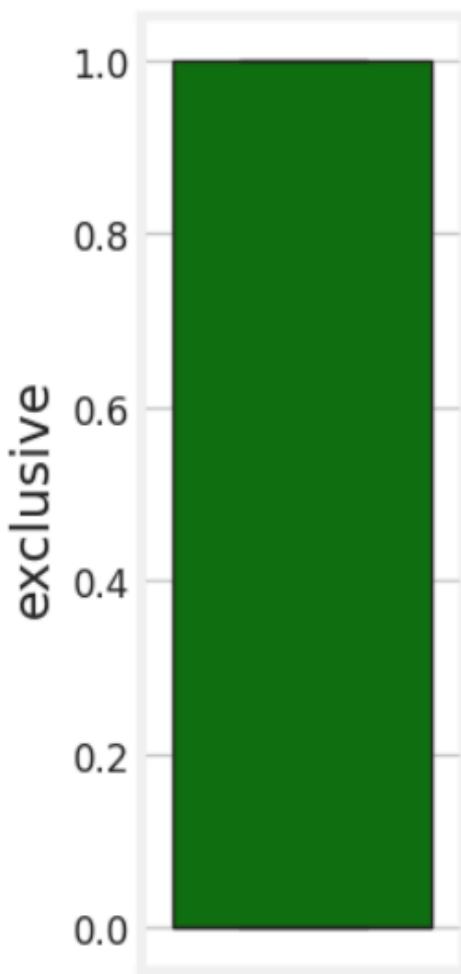
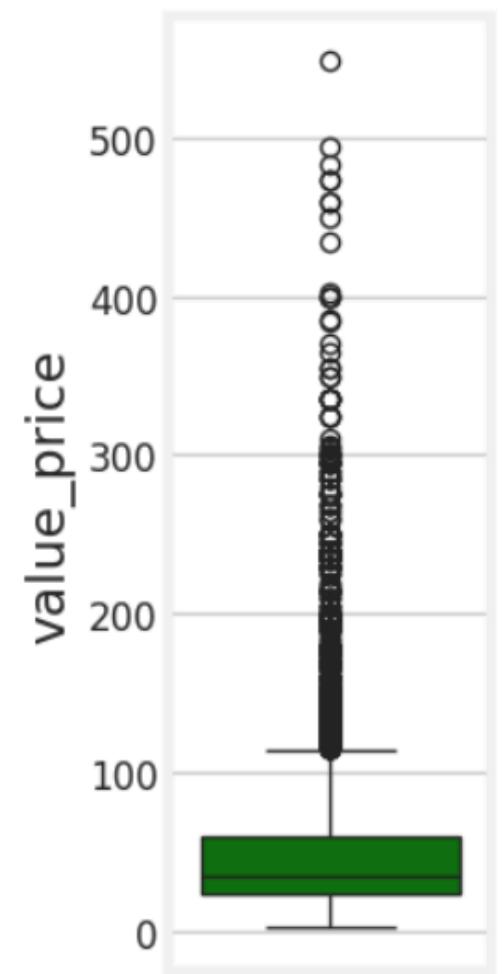
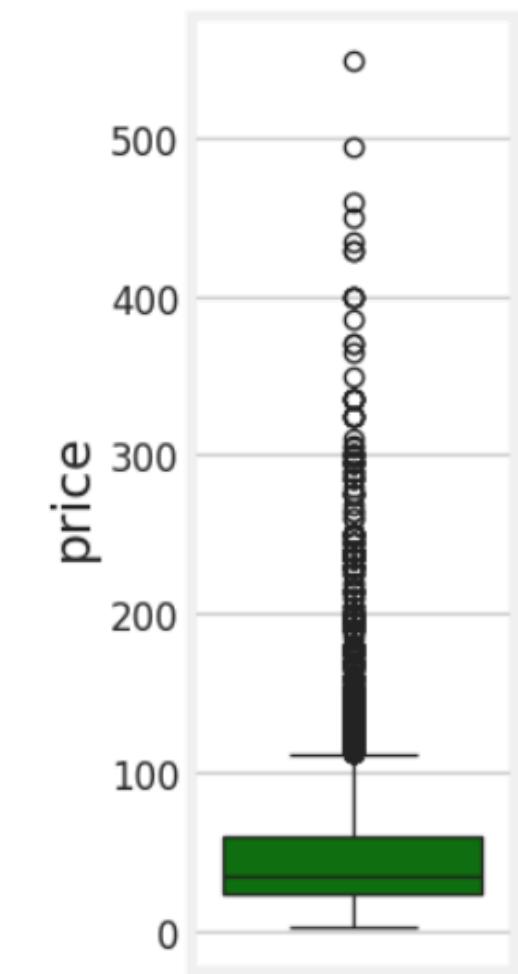
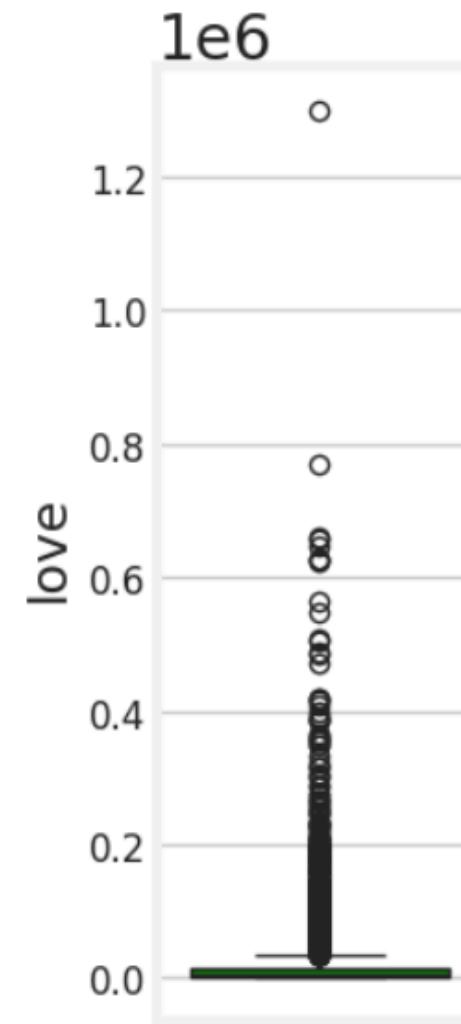
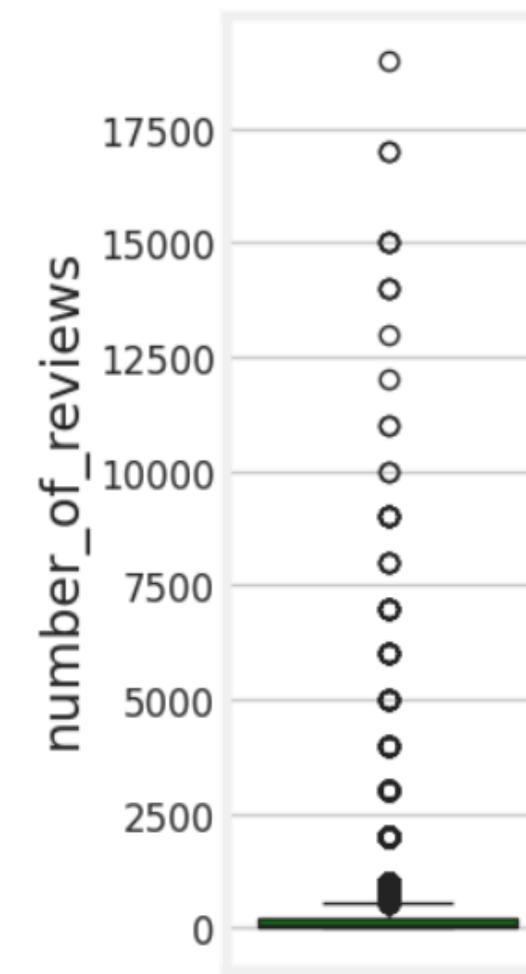
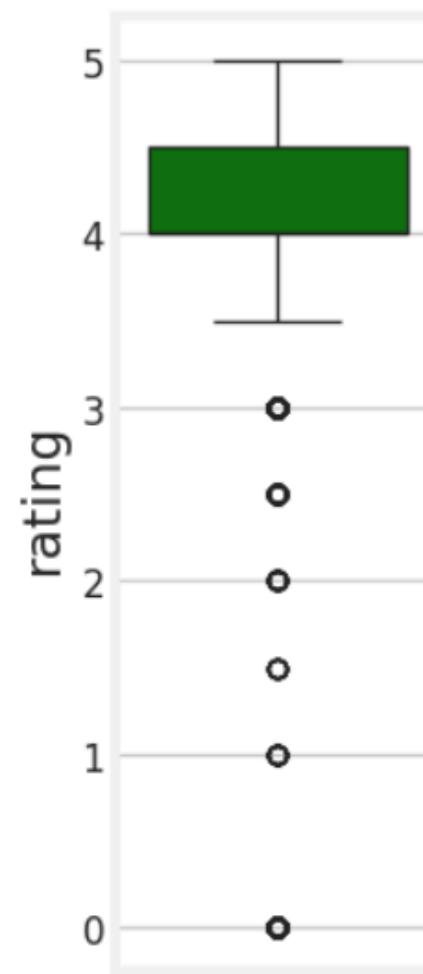
```
[ ] df[cats] = df[cats].astype(str)

[ ] df[cats].describe().T
```

	count	unique	top	freq
id	8000	7951	1723881	2
brand	8000	310	SEPHORA COLLECTION	492
category	8000	143	Perfume	619

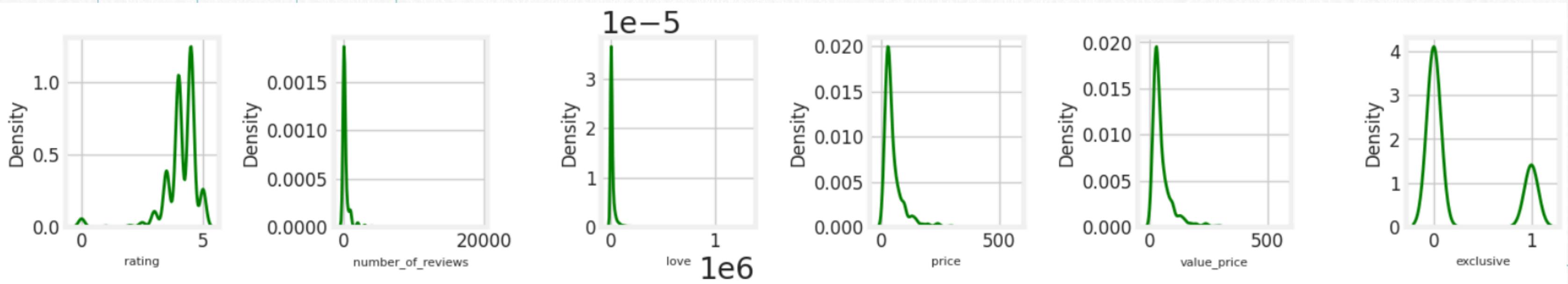
2. UNIVARIATE ANALYSIS

PENGAMATAN BOXPLOT



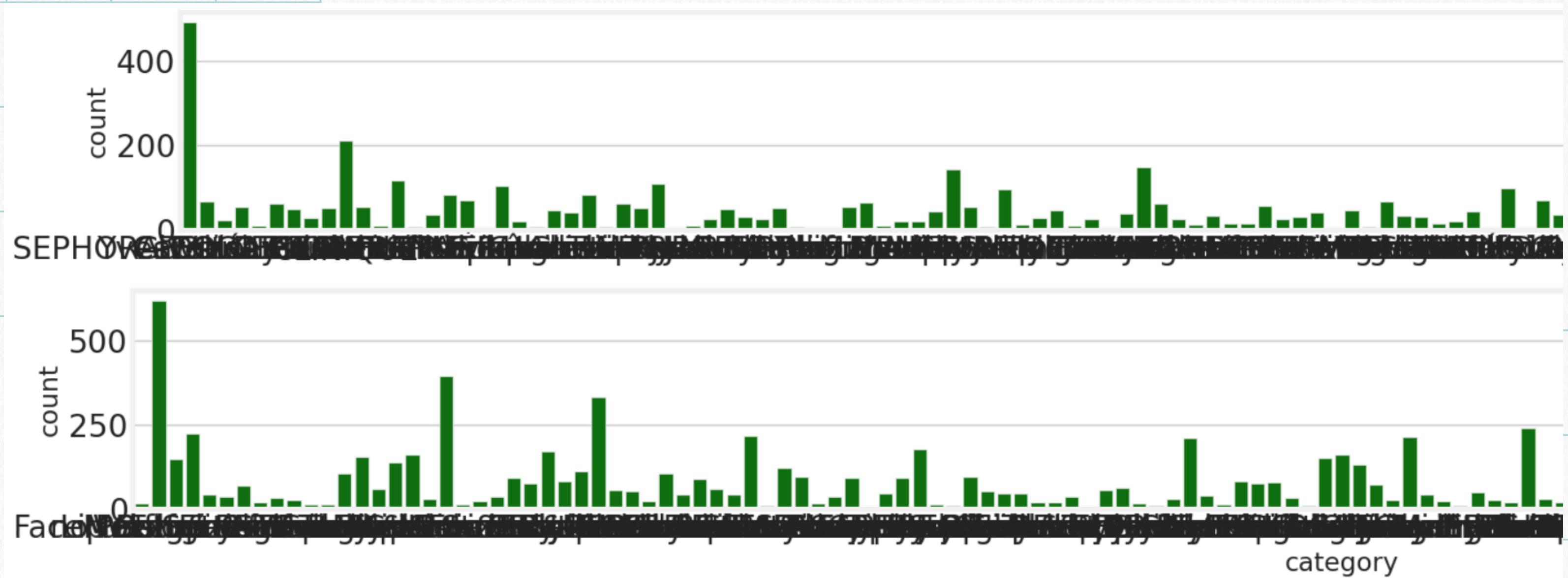
Kolom rating dan love memiliki outlier signifikan, menunjukkan beberapa produk dengan nilai yang sangat berbeda dari mayoritas. Distribusi data pada rating, number_of_reviews, love, price, dan value_price tidak simetris; price dan number_of_reviews cenderung skew ke kanan (lebih banyak produk dengan nilai rendah), sementara rating cenderung skew ke kiri (lebih banyak produk dengan nilai tinggi).

PENGAMATAN BOXPLOT

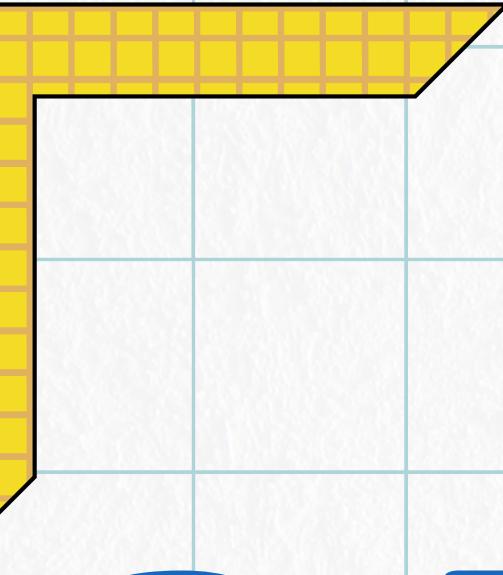


- Seperti observasi pada box plot sebelumnya, kolom `number_of_reviews`, `love`, `price`, `value_price` memiliki distribusi positive skew,
- Kolom `rating` memiliki arah distribusi hampir mendekati normal
- Kolom `exclusive` memiliki arah distribusi bimodal.

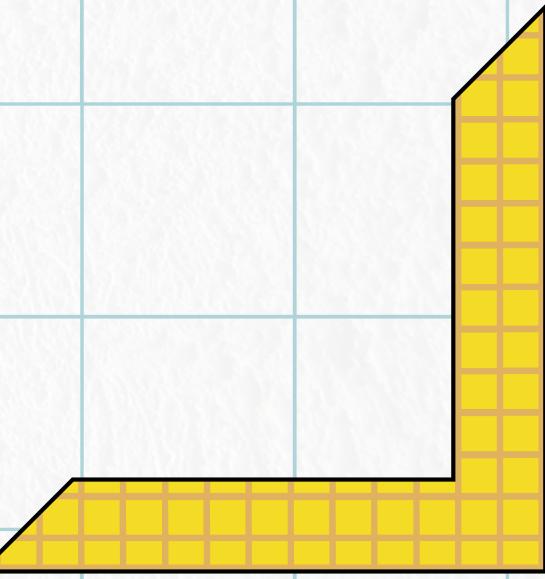
PENGAMATAN COUNTPLOT



Kolom brand dan category memiliki unique yang sangat banyak seperti pada gambar grafik dan perlu melakukan peninjauan kembali jika ingin membangun pemodelan



3. MULTIVARIATE ANALYSIS



Pengamatan HeatMap

Dari Correlation heatmap di sampaing dapat dilihat bahwa :

- 1.Target dari analisis ini adalah price dan memiliki korelasi positif
- 2.sangat kuat dengan value_price (strong potential feature)
- 3.Tetapi target value_pirce juga memiliki korelasi negatif dengan
- 4.number_of_reviews, love, dan exclusive (decent potential feature),
- 5.Price memiliki korelasi negatif cukup kuat dengan exclusive

	rating	number_of_reviews	love	price	value_price	exclusive
rating	1.00	0.07	0.08	0.05	0.05	-0.02
number_of_reviews	0.07	1.00	0.74	-0.09	-0.09	0.01
love	0.08	0.74	1.00	-0.09	-0.09	0.05
price	0.05	-0.09	-0.09	1.00	0.99	-0.18
value_price	0.05	-0.09	-0.09	0.99	1.00	-0.17
exclusive	-0.02	0.01	0.05	-0.18	-0.17	1.00

0.8

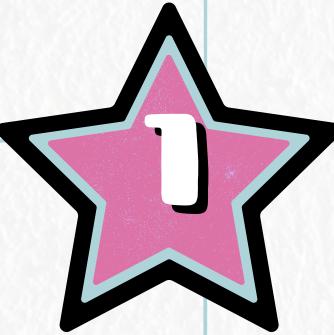
0.6

0.4

0.2

0.0

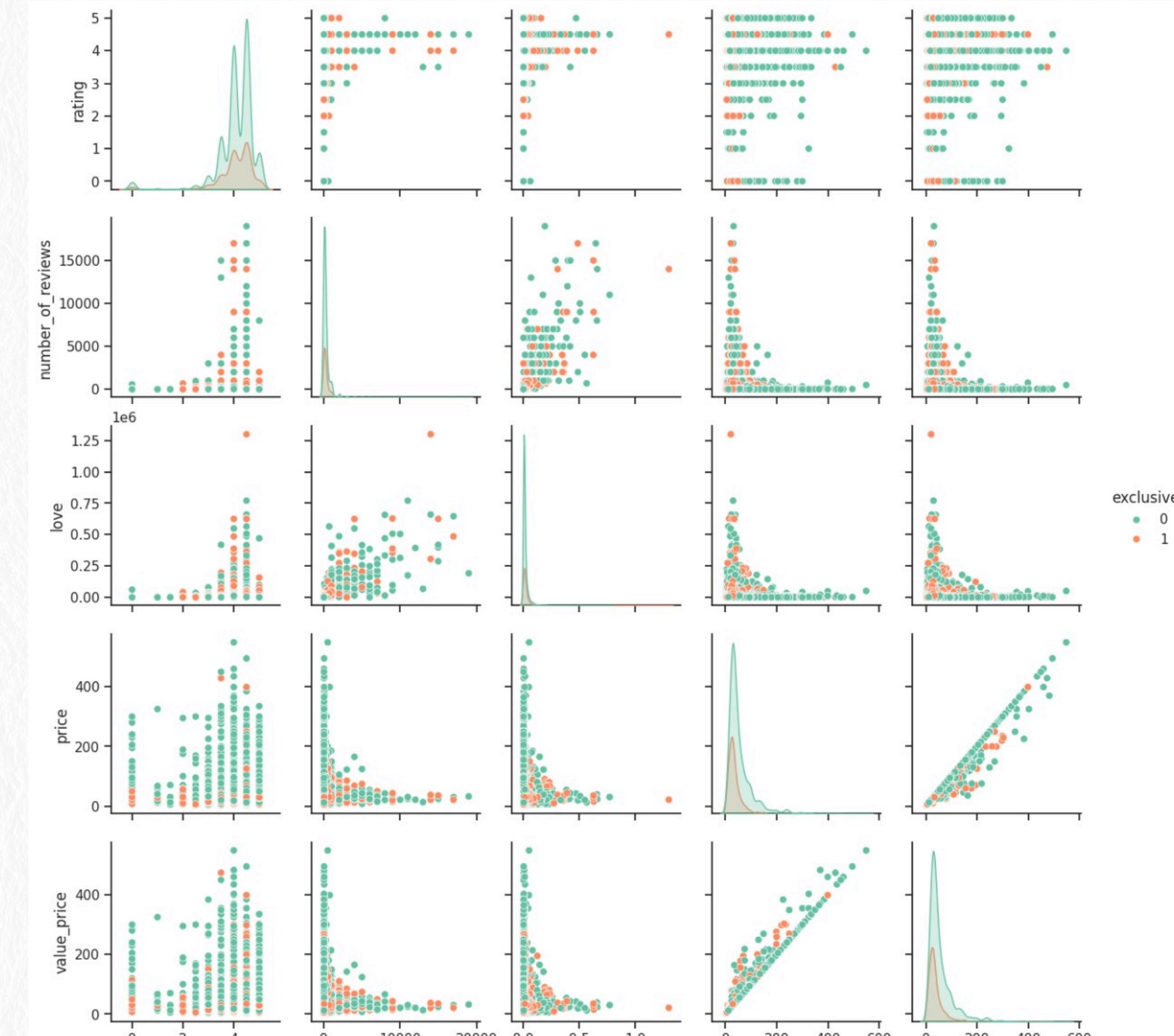
PROJECT GOALS



Dari *correlation heatmap* di atas dapat dilihat bahwa:

- Jika melihat target (price) dari analisis ini adalah price dan memiliki korelasi positif sangat kuat dengan value_price (strong potential feature),
- Tetapi target value_price juga memiliki korelasi negatif dengan number_of_reviews, love, dan exclusive (decent potential feature),
- price memiliki korelasi negatif cukup kuat dengan exclusive. (Paling negatif)

PAIRPLOT



HASIL GRAFIK PAIRPLOT

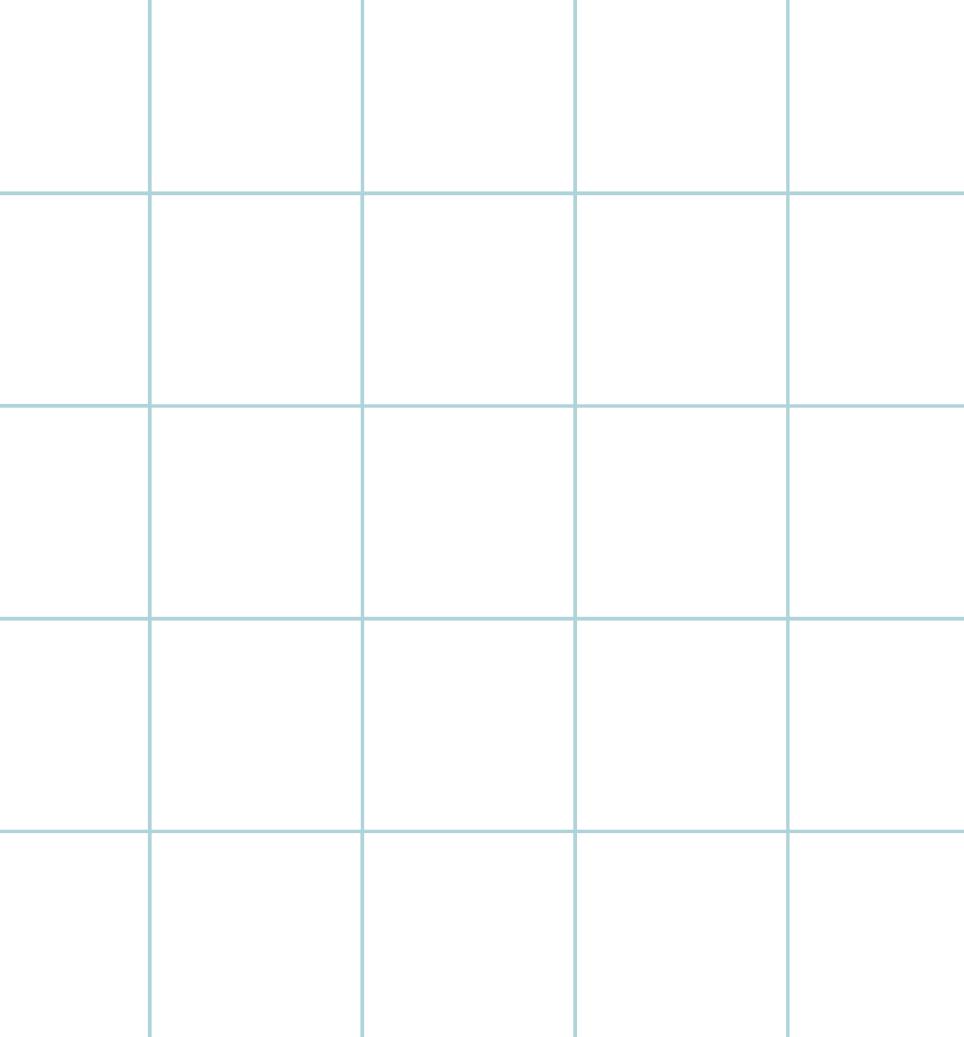
Hasil Grafik pairplot di atas, Menunjukan bahwa :

- `price` jelas memiliki korelasi linear dengan `value_price`
- Barang yang `exclusive = 0` memiliki nilai `rating`, `price`, `value_price`, dan `number_of_reviews` lebih besar dibanding barang `exclusive = 1`

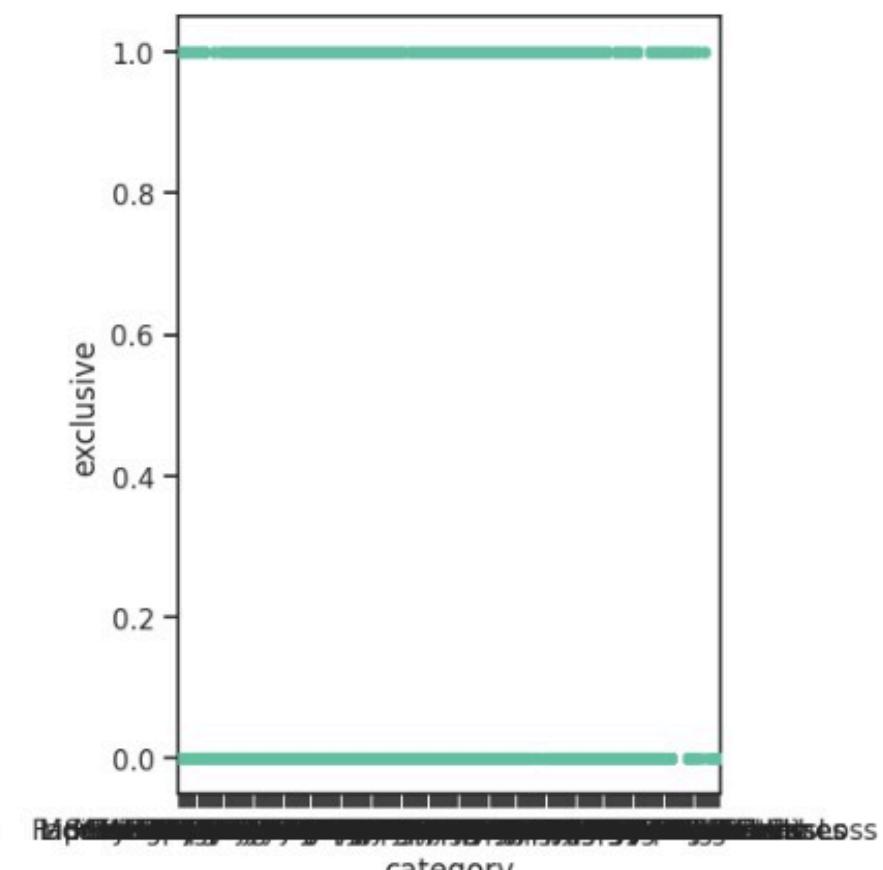
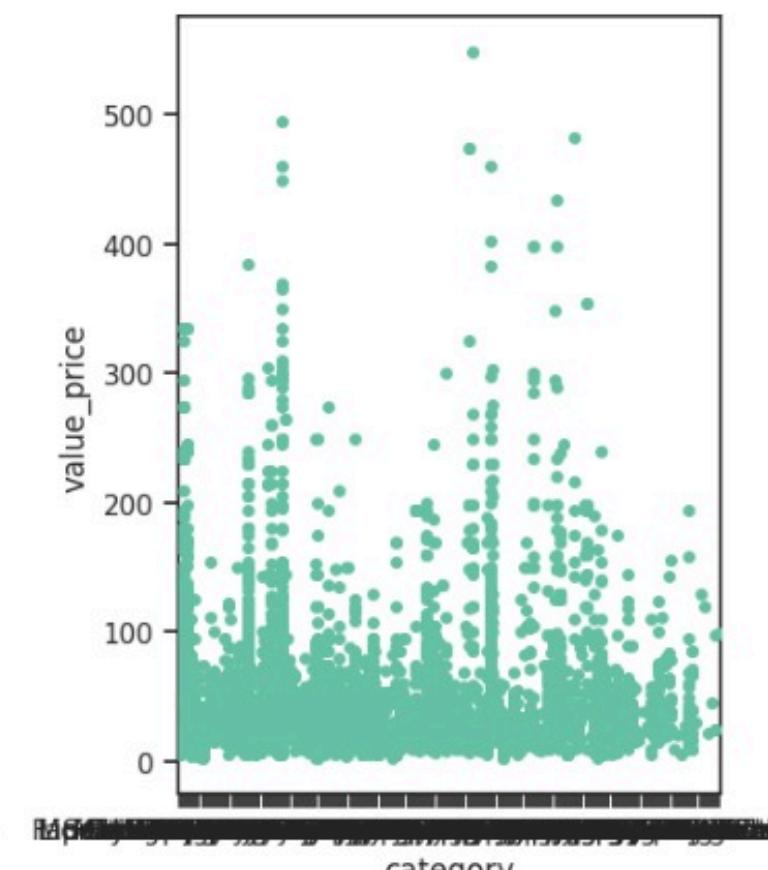
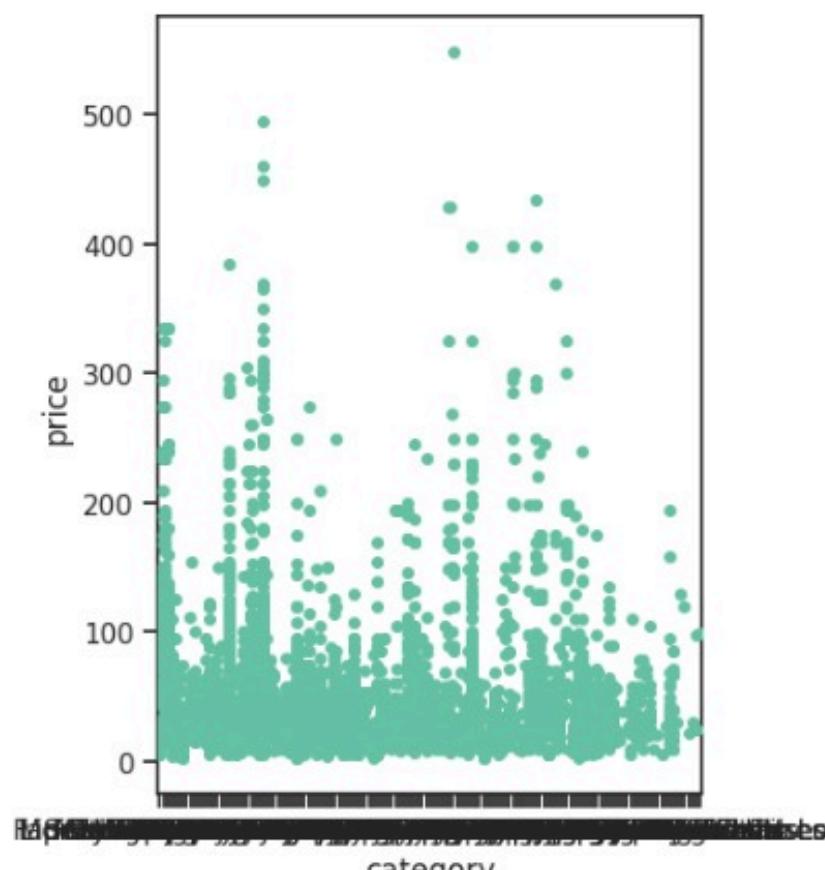
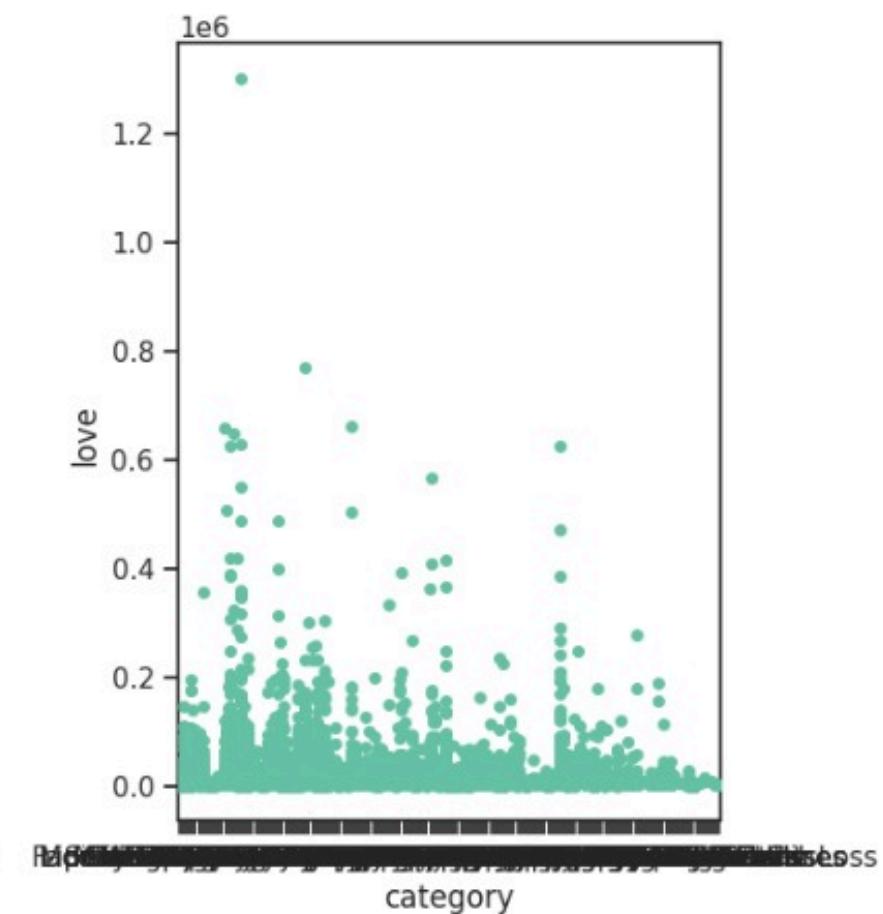
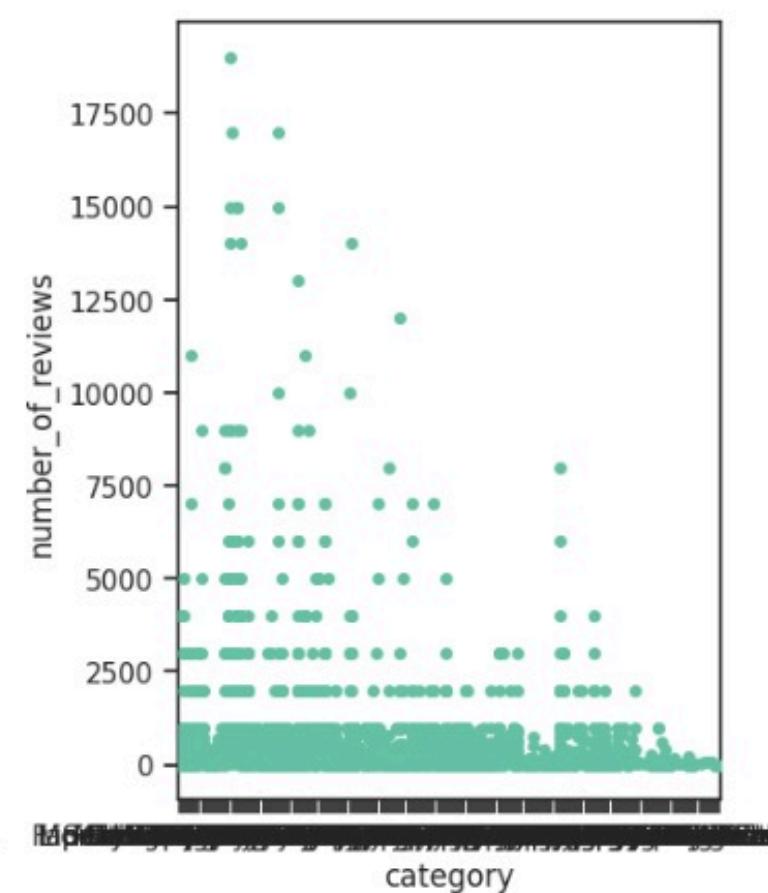
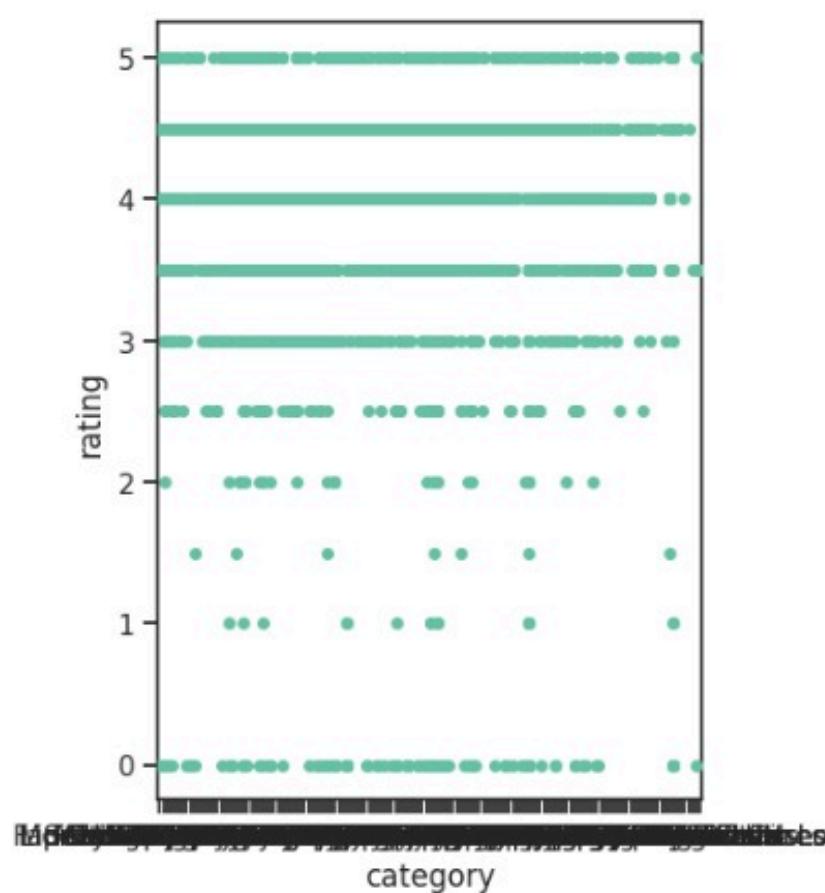
HASIL GRAFIK PAIRPLOT

Hasil Grafik pairplot di atas, Menunjukan bahwa :

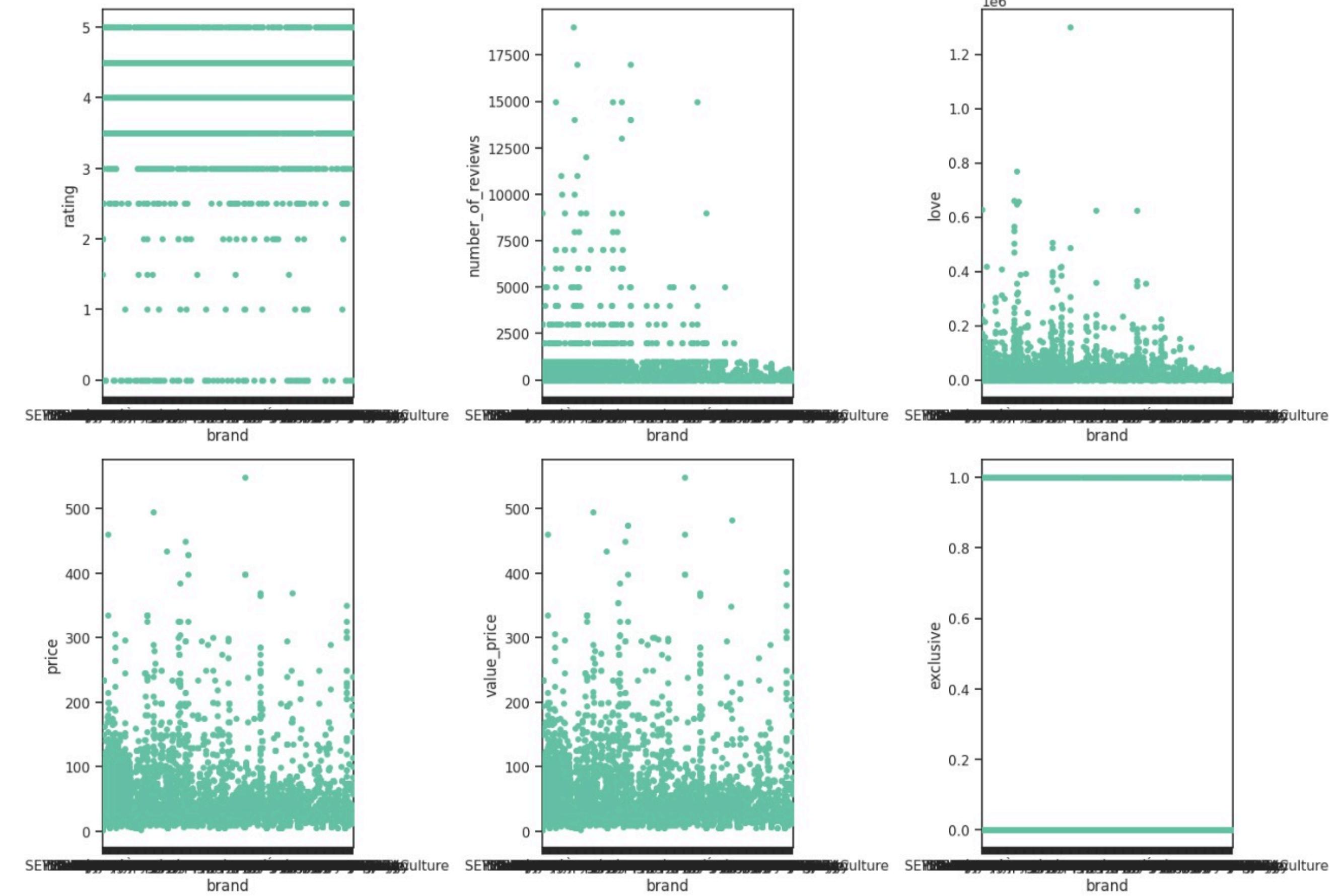
- `price` jelas memiliki korelasi linear dengan `value_price`
- Barang yang `exclusive = 0` memiliki nilai `rating`, `price`, `value_price`, dan `number_of_reviews` lebih besar dibanding barang `exclusive = 1`

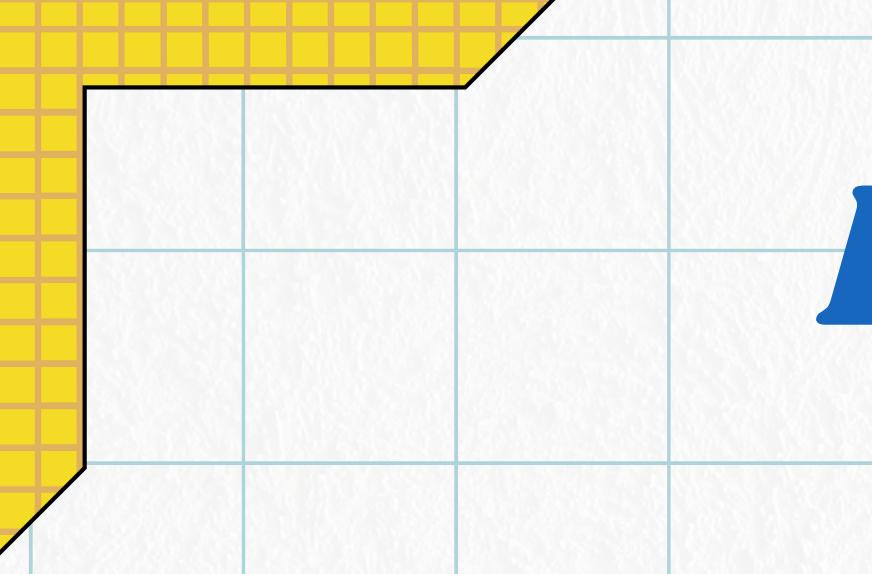


Dari Gambar plot disamping,
plot category tidak memiliki
korelasi dengan kolom
numeric



Dari Gambar plot disamping,
plot category tidak memiliki
korelasi dengan kolom
numeric





KESIMPULAN



1. Masih ada data yang kosong sehingga di handle saat pre-processing
 2. Ada beberapa distribusi yang skew
 3. Beberapa feature seperti brand dan kategori tidak memiliki korelasi dengan kolom numerical
- 
- 



4. DATA **CLEANSING**

- A. HANDLE MISSING VALUES
- B. HANDLE DUPLICATED DATA
- C. HANDLE OUTLIERS
- D. FEATURE TRANSFORMATION
- E. FEATURE ENCODING
- F. HANDLE CLASS IMBALANCE

A. HANDLE MISSING VALUES



```
# Disini kita melakukan handle Missing Value, Dengan mengisi kolom kategorikal agar tidak ada nilai NaN  
df['category'].fillna(df['category'].mode()[0], inplace=True)  
  
# Melakukan drop value NaN pada kolom yang memiliki Missing Value  
df = df.dropna(subset=['rating', 'number_of_reviews','love','price','value_price'])
```



Disini kita melakukan handle Missing Value, Dengan mengisi kolom kategorikal agar tidak ada nilai NaN, Melakukan drop value NaN pada kolom yang memiliki Missing Value.

B. HANDLE DUPLICATED DATA

```
[24] df.duplicated().sum()
```

```
→ 0
```

Tidak ada data terduplicate



Tidak ada data duplicated

C. HANDLE OUTLIERS



```
from scipy import stats

#menggunakan IQR
print(f'Jumlah baris sebelum memfilter outlier: {len(df)}')

filtered_entries = np.array([True] * len(df))
for col in ['rating','number_of_reviews','love','price','value_price']:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    low_limit = Q1 - (IQR * 1.5)
    high_limit = Q3 + (IQR * 1.5)

    filtered_entries = ((df[col] >= low_limit) & (df[col] <= high_limit)) & filtered_entries

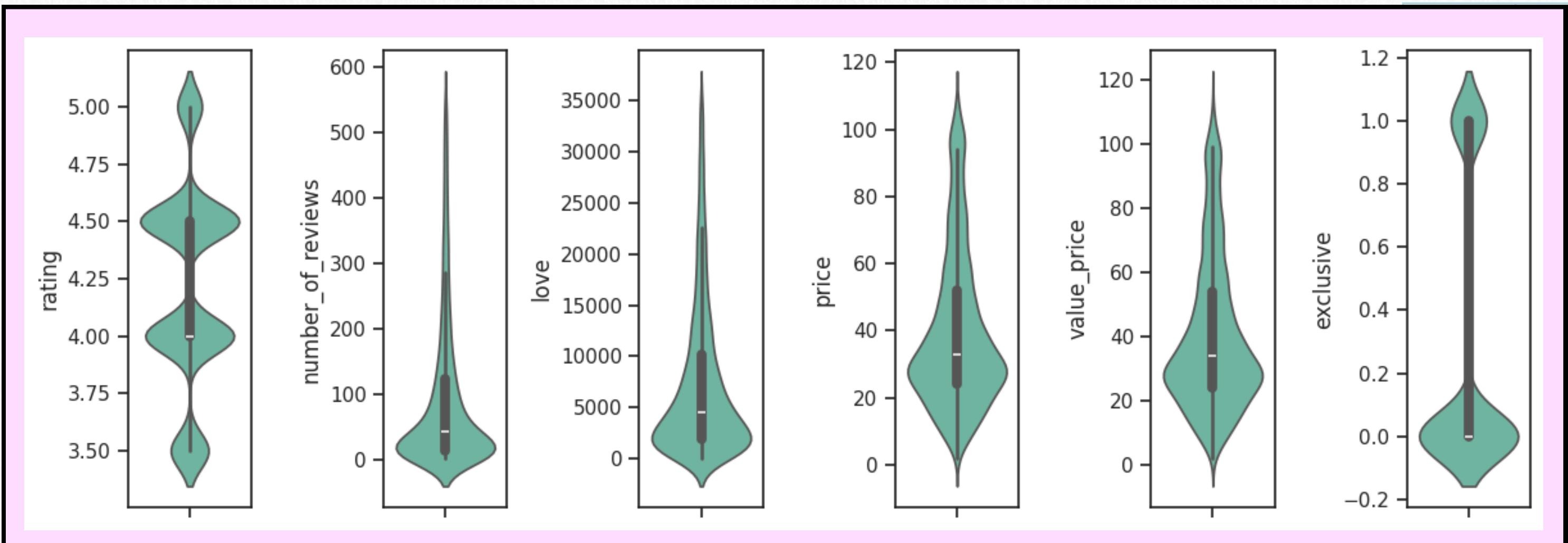
df = df[filtered_entries]

print(f'Jumlah baris setelah memfilter outlier: {len(df)}')
```

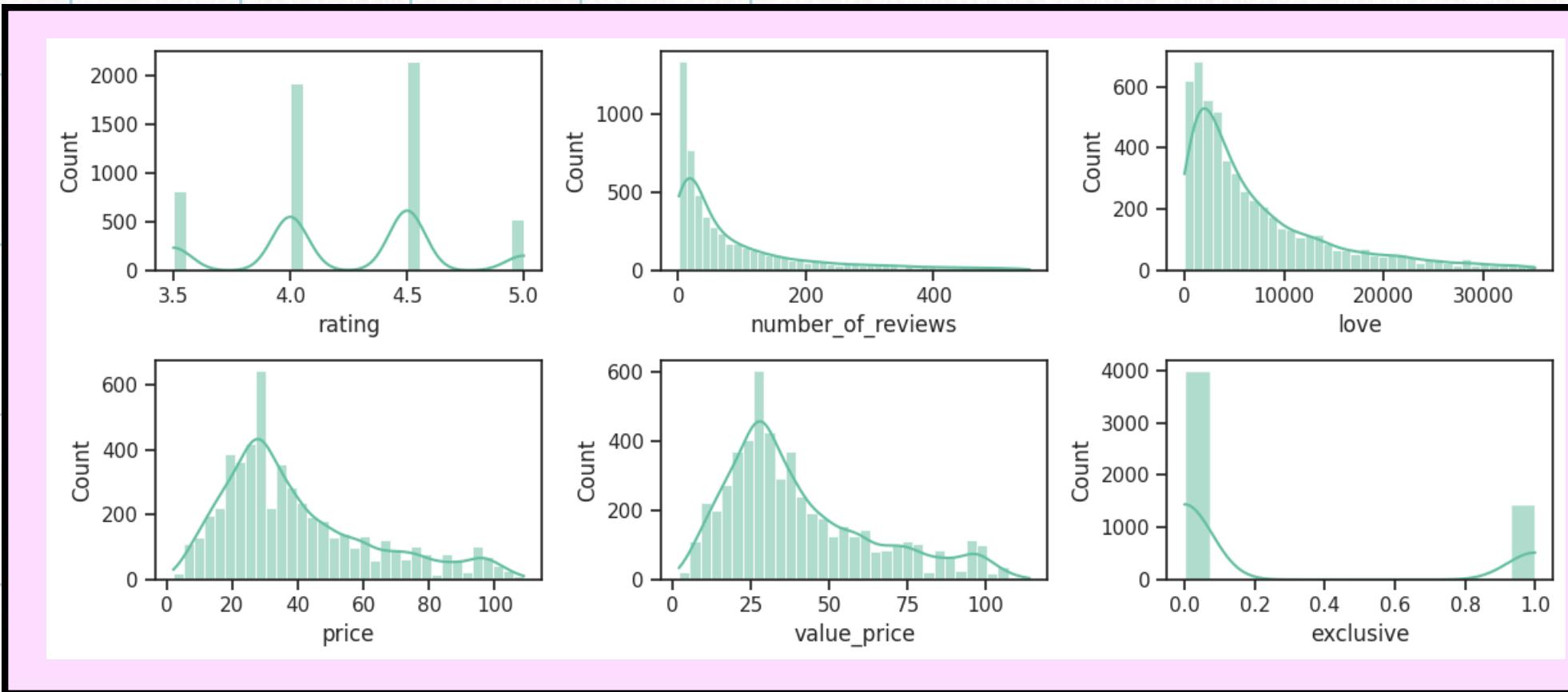
Jumlah baris sebelum memfilter outlier: 7841
Jumlah baris setelah memfilter outlier: 5425

DISTRIBUSI DATA SETELAH DIFILTER OUTLIER

```
#Data setelah oulier difilter menggunakan violin plot  
plt.figure(figsize = (12,4))  
for i in range(0, len(num)):  
    plt.subplot(1, 6, i+1)  
    sns.violinplot(y = df[num[i]], orient='v')  
    plt.tight_layout()
```

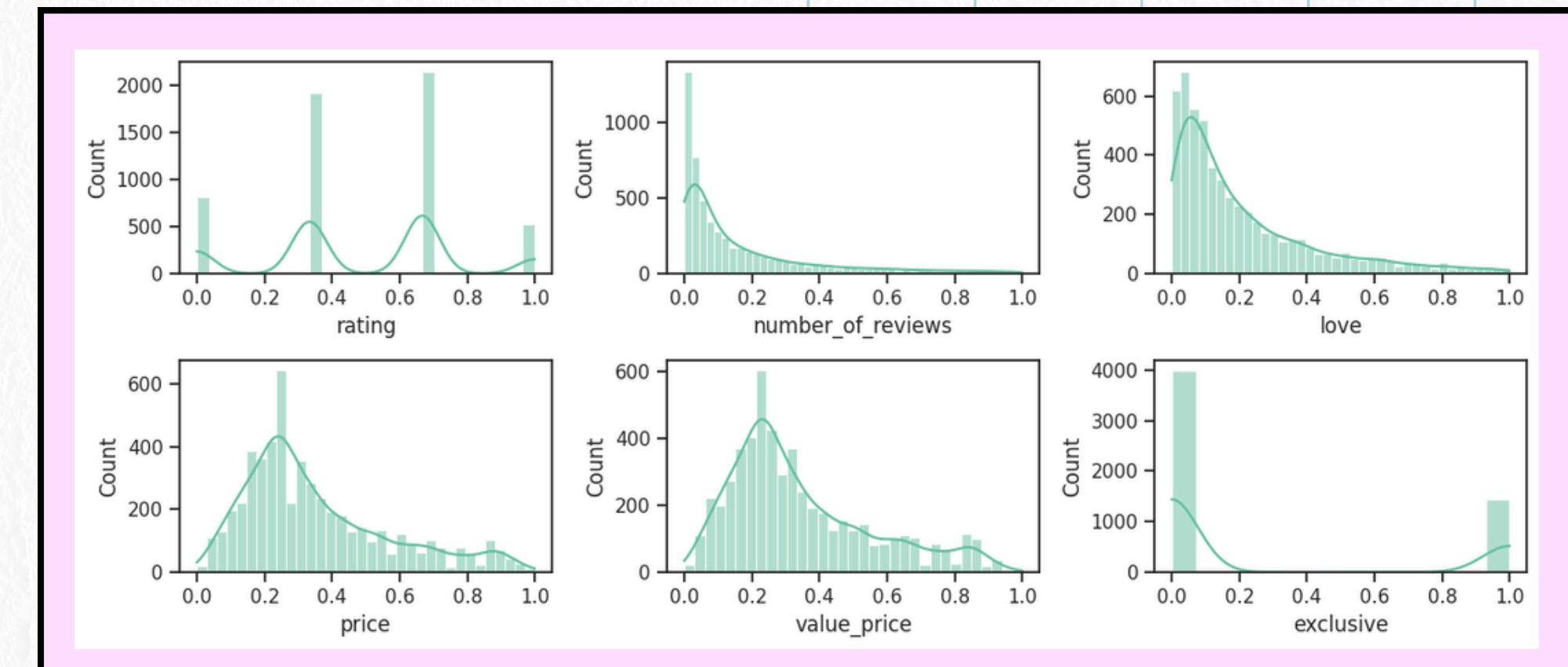


D. FEATURE TRANSFORMATION



Data Sebelum Normalisasi

Data Sesudah Normalisasi



E. FEATURE ENCODING



Tidak dilakukan feature encoding dikarenakan value counts sangat besar dalam artian count pada kolom kategorikal seperti `brand` dan `category` itu memiliki jumlah nama yang sangat banyak dan tidak sesuai dengan ketentuan yang harus dipenuhi jika ingin melakukan encoding.



F. HANDLE CLASS IMBALANCE

```
[93] # Memberikan nama baru pada data exclusive yang memenuhi standar diatas 0.8
df['Tru_Exclusive'] = df['exclusive'] > 0.8

[94] # menghitung hasil data yang
df['Tru_Exclusive'].value_counts()

Tru_Exclusive
False    3989
True     1436
Name: count, dtype: int64
```

Cek keseimbangan data
(Exclusive Vs
Non-Exclusive)

HANDLING IMBALANCE DATA MENGGUNAKAN TEKNIK SMOTE

```
from imblearn.under_sampling import RandomUnderSampler
from imblearn.over_sampling import RandomOverSampler, SMOTE
import pandas as pd

# Random Undersampling
X_under, y_under = RandomUnderSampler(sampling_strategy=0.5).fit_resample(X, y)

# Random Oversampling
X_over, y_over = RandomOverSampler(sampling_strategy=0.5).fit_resample(X, y)

# SMOTE
X_over_SMOTE, y_over_SMOTE = SMOTE().fit_resample(X, y)

print(pd.Series(y_under).value_counts())
print(pd.Series(y_over).value_counts())
print(pd.Series(y_over_SMOTE).value_counts())
```

False	2872
True	1436
Name: count, dtype: int64	
False	3989
True	1994
Name: count, dtype: int64	
False	3989
True	3989
Name: count, dtype: int64	

Dari proses data pre-processing tersebut yang paling memberikan impact adalah data imbalanced dan juga data outliers.

5. FEATURE ENGINEERING

Feature Selection

Feature Extraction

Feature Tambahan

01.

Feature yang kurang relevan :

1. id (Karena sangat tidak relevan dan tidak berkorelasi dengan data manapun)
 2. love
 3. number_of_reviews
- Sehingga perlu dilakukan selection pada features tersebut.

02.

1. Mengategorikan Kolom rating dan price

Kita akan membuat kategori untuk kolom rating dan price menjadi low, medium, dan high. Ini akan memberikan kita wawasan lebih mengenai distribusi data dalam setiap kategori.

*2. Membuat Kolom worst_review

Kita akan membuat kolom baru bernama worst_review. Jika nilai rating adalah 0, number_of_reviews adalah 0, dan love adalah 0, maka kita akan melabelinya sebagai "worst review". Menghitung Korelasi Kita akan melihat korelasi antara kolom worst_review dengan kategori lainnya untuk mendapatkan wawasan lebih lanjut.

03.

Features tambahan yang bisa kita lakukan untuk meng optimalkan model machine learning adalah dengan menambahkan feature sales, gender, umur, rekomendasi, lokasi.



THANK YOU

Presented by : Involfators