EDA & Data Pre Processing of Product Classification

Fahrizan Rasyad

EDA

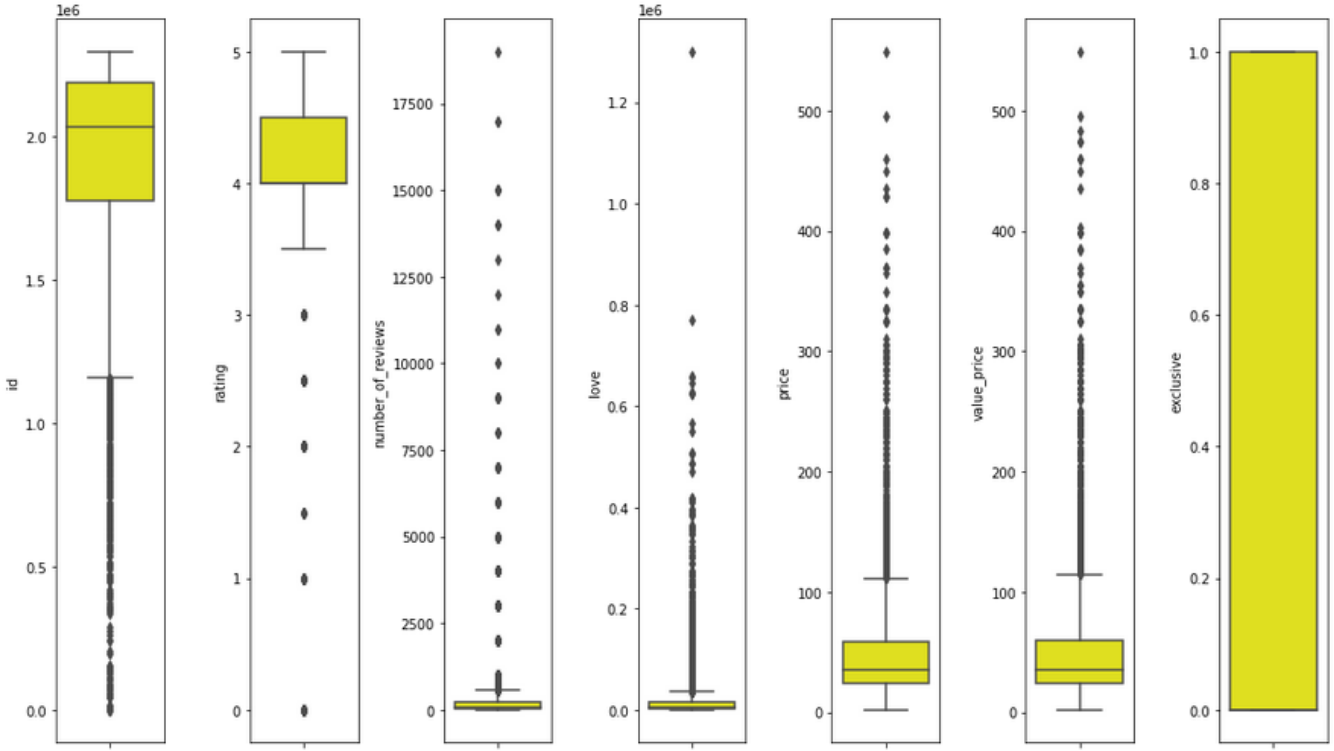
Exploratory Data Analysis

```
for col in cats:
    print('Value count columns {col}:')
    print(df[col].value_counts())
    print()
```

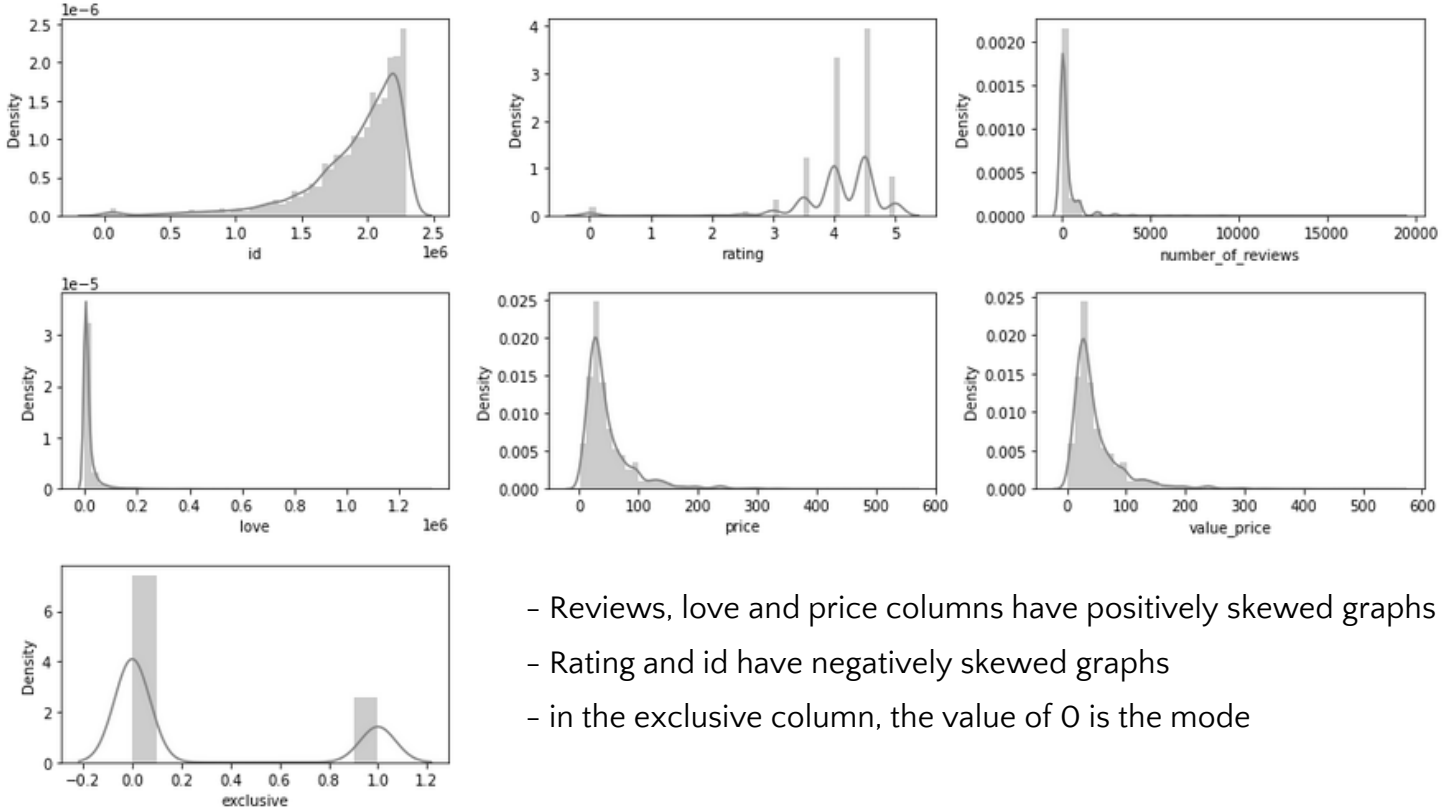
```
Value count columns {col}:
SEPHORA COLLECTION     492
CLINIQUE               211
TOM FORD               150
tarte                  143
Kiehl's Since 1851     122
                      ...
bkr                      1
DL.MD                    1
High Beauty              1
Too Cool For School      1
Cocofloss                1
Name: brand, Length: 310, dtype: int64

Value count columns {col}:
Perfume                  632
Moisturizers             395
Face Serums              334
Value & Gift Sets        241
Face Wash & Cleansers    225
                        ...
Powder Brush               1
Cleansing Brushes          1
Curls & Coils              1
Lid Shadow Brush           1
Body Moisturizers          1
Name: category, Length: 142, dtype: int64
```

Both categorical columns (object datatypes) have a lot of **unique values**

Outliers Detected



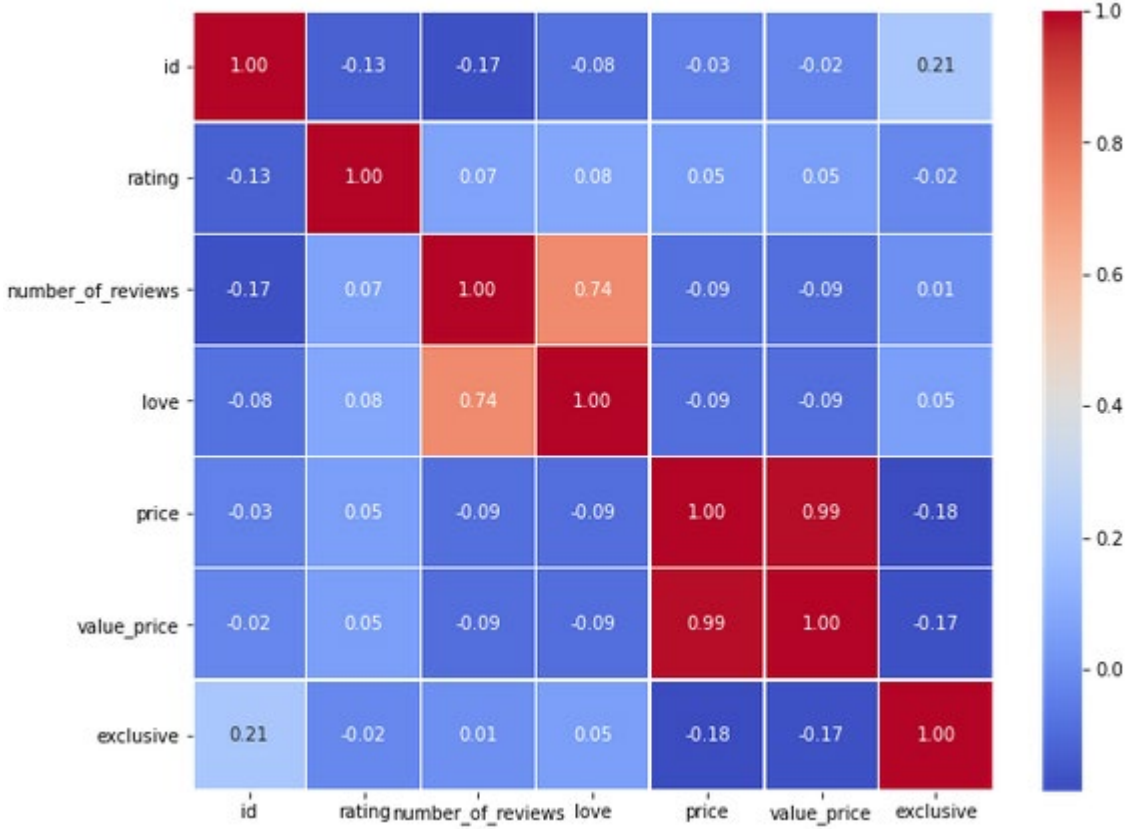
Distribution Plot



- Reviews, love and price columns have positively skewed graphs
- Rating and id have negatively skewed graphs
- in the exclusive column, the value of 0 is the mode

Correlation Heatmap

- *'Exclusive'* as a target has a weak positive correlation with *'review'*, *'love'*, and has a negative correlation with *'price'* and *'value price'*
- *'Value Price'* and *'price'* features might categorized to be redundant due to strong correlation.
- *'number_of_review'* and *'love'* also have a very strong correlation above 0.7, it is also likely to be redundant features.

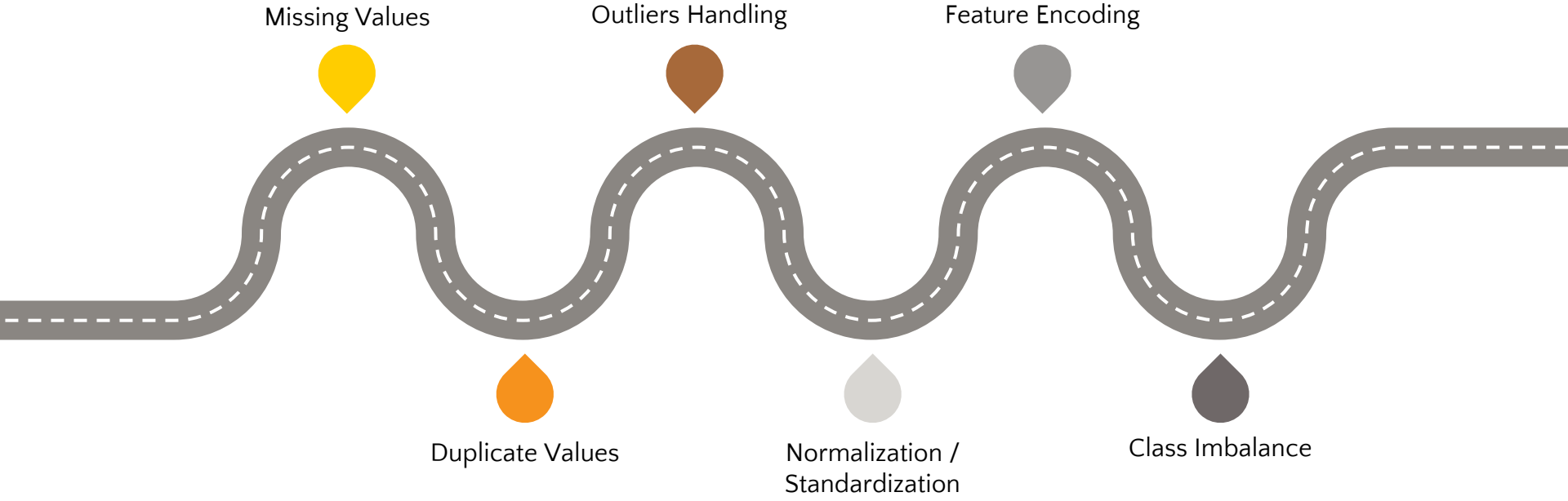


Insights

- In the correlation between features, it can be seen that there are some features that are redundant or have a very high correlation to the other. This causes one of the redundant features to be removed because it has the same effect on the target.
- Based on all feature correlations to the target, it can be concluded that there are no features that are strongly correlated to the target (exclusive), all of them have a relatively weak correlation (<0.5)
- Additional treatment is needed to be able to find features that have a strong enough correlation in influencing the target (exclusive)

Data Preparation

Data Preparation



Data Preparation

All pre-processing aims to make the data as clean and good as possible before it is entered into the machine learning model. However, the pre-processing that gives the most impact are 'Missing Value', 'Outlier', and 'Class Imbalance'. In addition, by grouping categorical data types, it can be seen that the unique values number in the hundreds so that they cannot be used as a feature for further analysis.

Thanks!

Looking forward to positive feedback!

- indianajanss on twitter
- Fahrizan rasyad on linkdin