Rakamin Academy

E-Commerce Shipping Data

"Asklepios"

Awalsyah Rinanto Putra

Fathah Oscar

M Rizky Septiansyah

Hermawan Febrianto

Devi Puji Ayuningsih

Anggita Citanegara Lubis

Stage 2 (Pre-Processing)

1. Data Cleansing

A. Handle Missing Values

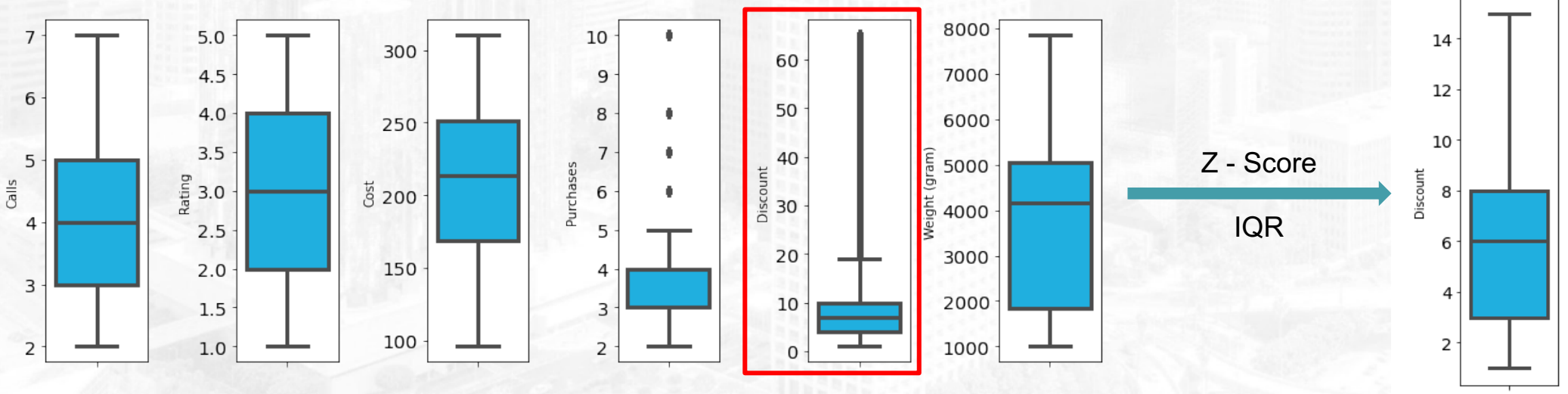
Tidak perlu dilakukan karena tidak ada missing values pada dataset

B. Handle Duplicated Data

Tidak perlu dilakukan karena tidak ada data yang terduplikasi pada dataset

C. Handle Outliers

Dilakukan pada feature Discount

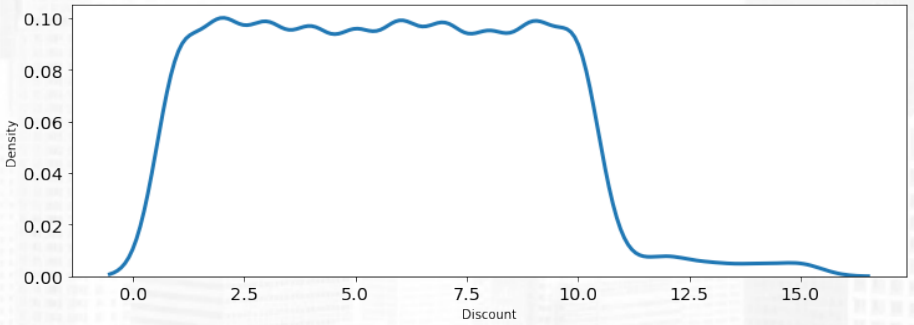


Jumlah data sebelum outlier dihilangkan : 10999

Jumlah data setelah outlier dihilangkan : 8604

D. Feature Transformation

- ***Log Transformation***




```
skewness feature Discount:  0.22926712487076678
```

Tidak perlu dilakukan karena nilai skew pada feature Discount sudah mendekati distribusi normal

- ***Standardization***

```
# Standardization pada kolom numerik
df['Std_Cost'] = StandardScaler().fit_transform(df['Cost'].values.reshape(len(df), 1))
df['Std_Disc'] = StandardScaler().fit_transform(df['Log_Discount'].values.reshape(len(df), 1))
df['Std_Weight'] = StandardScaler().fit_transform(df['Weight (gram)'].values.reshape(len(df), 1))
```

Sebelum Standardization

```
variance :

Weight (gram)    2613757.66
Cost                2306.67
Discount               9.68
dtype: float64
```

```
standard deviation :

Weight (gram)    1616.71
Cost               48.03
Discount            3.11
dtype: float64
```

Setelah Standardization

```
variance :

Std_Weight    1.0
Std_Disc      1.0
Std_Cost      1.0
dtype: float64
```

```
standard deviation :

Std_Weight    1.0
Std_Disc      1.0
Std_Cost      1.0
dtype: float64
```

D. Feature Transformation

- *Normalization*

	Norm_Weight	Norm_Cost	Norm_Disc
count	8604.000000	8604.000000	8604.000000
mean	0.585027	0.551658	0.569361
std	0.316940	0.224429	0.261301
min	0.000000	0.000000	0.000000
25%	0.188198	0.364486	0.405684
50%	0.693982	0.584112	0.661642
75%	0.834983	0.738318	0.767874
max	1.000000	1.000000	1.000000

```
df['Norm_Cost'] = MinMaxScaler().fit_transform(df['Cost'].values.reshape(len(df), 1))
df['Norm_Disc'] = MinMaxScaler().fit_transform(df['Log_Discount'].values.reshape(len(df), 1))
df['Norm_Weight'] = MinMaxScaler().fit_transform(df['Weight (gram)'].values.reshape(len(df), 1))
```

Semua nilai min dan max pada feature yang dinormalisasi sudah bernilai 0 dan 1

E. Feature Encoding

- *Label Encoding*

```
{'low' : 0,
 'medium' : 1,
 'high' : 2,}
```

```
Value Counts feature Product Importance :
0    4185
1    3720
2     699
Name: Importance, dtype: int64
```

```
Value Counts feature Gender :
0    4313
1    4291
Name: Gender, dtype: int64
```

```
{'F' : 0,
 'M' : 1}
```

Dilakukan pada kolom Importance yang mempunyai tipe data ordinal dan kolom Gender

E. Feature Encoding

- *One Hot Encoding*

Warehouse_A	Warehouse_B	Warehouse_C	Warehouse_D	Warehouse_F	Shipment_Flight	Shipment_Road	Shipment_Ship
0	0	0	1	0	1	0	0
0	0	0	0	1	1	0	0
1	0	0	0	0	1	0	0
0	1	0	0	0	1	0	0
0	0	1	0	0	1	0	0

Dilakukan pada feature Warehouse dan Shipment method

F. Handle Class Imbalance

	Late	Jumlah	Ratio
0	0	4436	51.557415
1	1	4168	48.442585

Tidak perlu dilakukan karena proportion of minority class >40%

2. Feature Engineering

A. Feature Selection

- Menghapus feature **ID** dikarenakan feature tersebut tidak memiliki arti penting untuk kegunaan proses modelling.
- Menghapus feature **Warehouse** dan **Shipment** karena sudah dilakukan feature encoding
- Dari heatmap plot, tidak ada feature lain yang perlu dihapus karena tidak ada feature yang redundant dengan nilai korelasi antar feature > 0.7

B. Feature Exraction

Tidak ada fitur yang bisa diekstraksi dari dataset

C. Feature Tambahan

1. Waktu pengiriman

Bisa dilakukan analisis regresi untuk memprediksi waktu pengiriman customer di waktu yang akan datang.

2. Alamat customer (Kota-Provinsi/Luar negeri)

Jika jauh, potensi terlambat makin besar karena makin banyak peluang mengalami kendala pengiriman

Jika di luar negeri, potensi terlambat makin besar karena penyesuaian regulasi import dan eskport pengiriman barang

3. Alamat Warehouse

Bisa digunakan untuk merekomendasikan warehouse mana yang paling dekat dengan alamat customer agar potensi keterlambatan dapat direduksi

2. Feature Engineering

C. Feature Tambahan

4. Musim

Pada musim hujan atau musim dingin, moda pengiriman kapal bisa terkendala karena cuaca buruk bisa mengakibatkan dilarangnya kapal berlayar.

5. Kapasitas Pengiriman Per Hari

Makin sedikit kapasitas, potensi terlambat makin besar karena makin sedikit pengiriman dilakukan.

6. Traffic Route

Makin padat rute yang dipilih, potensi terlambat makin besar karena durasi pengiriman makin lama.

3. Git

https://github.com/anggita0712/Asklepios_Preprocessing