Rakamin Academy

Data Pre-Processing

Data Set E-Commerce Shipping

Final Project - Laporan Stage 2

Kelompok 8 - Decentraland

Anggota Kelompok:

- Dharma Setiawan
- Ilham Ibnu A.
- M. Farhan Atmawinanda
- Fikri Diva S.
- Ahmad Ilham H.

STAGE 2 DISCUSSION

- Data Cleaning
- Feature Engineering

Data Cleaning

- Menghapus column “ID”
- Pembersihan Data Outlier

Informasi isi dataset ecommerce shipping

```
1 # info semua variabel pada dataset
2 df_shipping.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10999 entries, 0 to 10998
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   10999 non-null  int64
 1   Warehouse_block      10999 non-null  object
 2   Mode_of_Shipment     10999 non-null  object
 3   Customer_care_calls  10999 non-null  int64
 4   Customer_rating      10999 non-null  int64
 5   Cost_of_the_Product  10999 non-null  int64
 6   Prior_purchases      10999 non-null  int64
 7   Product_importance   10999 non-null  object
 8   Gender               10999 non-null  object
 9   Discount_offered     10999 non-null  int64
 10  Weight_in_gms        10999 non-null  int64
 11  Reached.on.Time_Y.N  10999 non-null  int64
dtypes: int64(8), object(4)
memory usage: 1.0+ MB
```

```
1 # cek missing value
2 df_shipping.isna().sum()
```

```
ID                     0
Warehouse_block        0
Mode_of_Shipment       0
Customer_care_calls    0
Customer_rating        0
Cost_of_the_Product    0
Prior_purchases        0
Product_importance     0
Gender                 0
Discount_offered       0
Weight_in_gms          0
Reached.on.Time_Y.N    0
dtype: int64
```

```
1 # cek duplicated data
2 df_shipping.duplicated().sum()
```

```
0
```

Informasi isi dataset ecommerce shipping

Rakamin Academy

- Data terdiri dari **10.999 sampel** (baris).
- **Tidak ada null value** di semua kolom.
- **Tidak ada baris yang terduplikasi**.
- Terdapat **10 fitur** (variabel independen) dan **1 target variabel** (variabel dependen).
- **Berdasarkan jenisnya**, fitur terdiri dari **4 fitur kategorik (‘Warehouse_block’, ‘Mode_of_Shipment’, ‘Product_importance’, dan ‘Gender’)** dan **6 fitur numerik (‘Customer_care_calls’, ‘Customer_rating’, ‘Cost_of_the_Product’, ‘Prior_purchases’, ‘Discount_offered’, dan ‘Weight_in_gms’)**.
- Satu **target variabel** yaitu **‘Reached.on.Time_Y.N’** yang bertipe **numerik**.

Delete column ID

```
1 # Drop columns ID
2 df_shipping.drop(columns="ID", inplace=True)
```

- Menghapus column **ID** dikarenakan column tersebut tidak memiliki arti penting untuk kegunaan proses modelling nantinya.

Descriptive statistik numeric data

```
# Descriptive statistik untuk numeric data
df_shipping[numerical_cols].describe()
```

	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
count	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000
mean	4.054459	2.990545	210.196836	3.567597	13.373216	3634.016729	0.596691
std	1.141490	1.413603	48.063272	1.522860	16.205527	1635.377251	0.490584
min	2.000000	1.000000	96.000000	2.000000	1.000000	1001.000000	0.000000
25%	3.000000	2.000000	169.000000	3.000000	4.000000	1839.500000	0.000000
50%	4.000000	3.000000	214.000000	3.000000	7.000000	4149.000000	1.000000
75%	5.000000	4.000000	251.000000	4.000000	10.000000	5050.000000	1.000000
max	7.000000	5.000000	310.000000	10.000000	65.000000	7846.000000	1.000000

Descriptive statistik categorical data

```
1 # Descriptive statistik untuk kategorikal data
2 df_shipping[category_cols].describe()
```

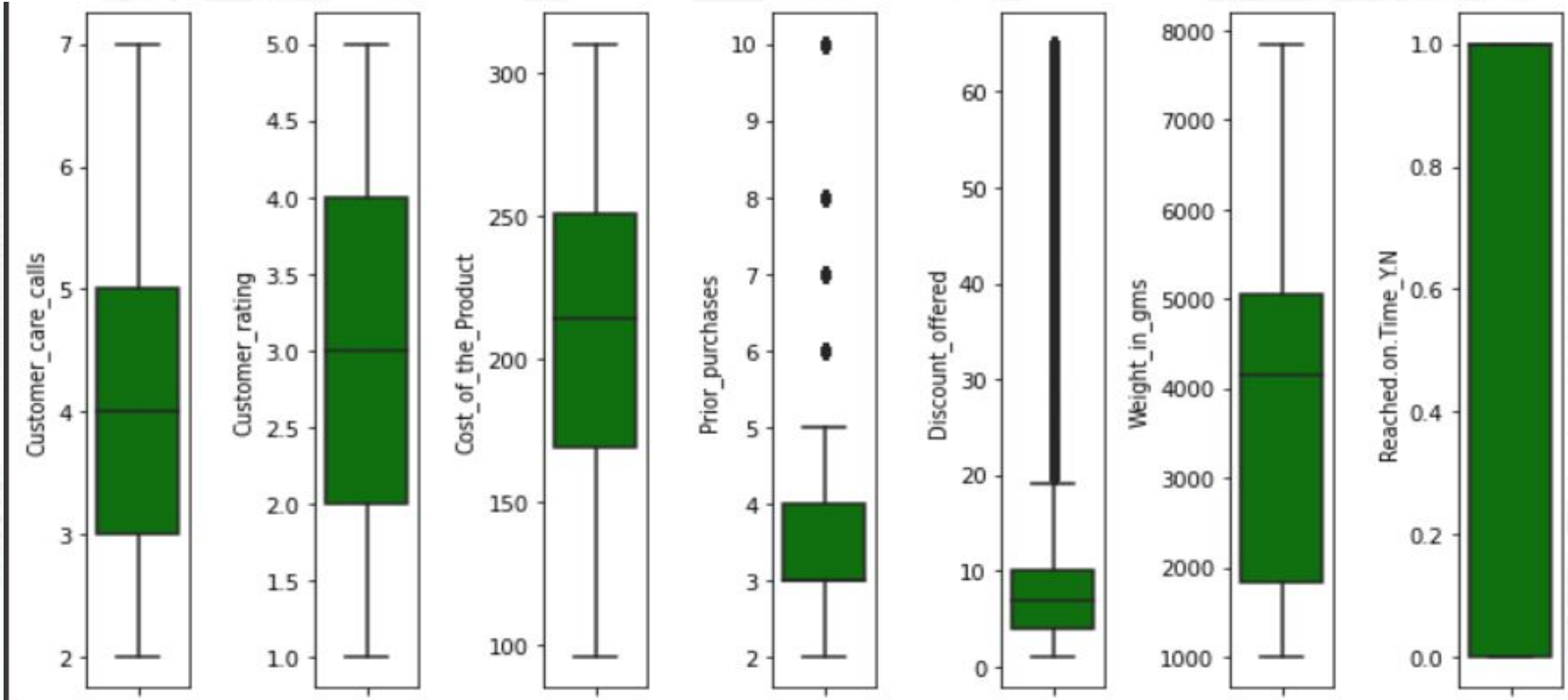
	Warehouse_block	Mode_of_Shipment	Product_importance	Gender
count	10999	10999	10999	10999
unique	5	3	3	2
top	F	Ship	low	F
freq	3666	7462	5297	5545

Informasi hasil descriptive statistik

- Data terdiri dari **10.999 sampel** (baris).
- **Tidak ada null value** di semua kolom.
- **Tidak ada invalid values** di semua column.

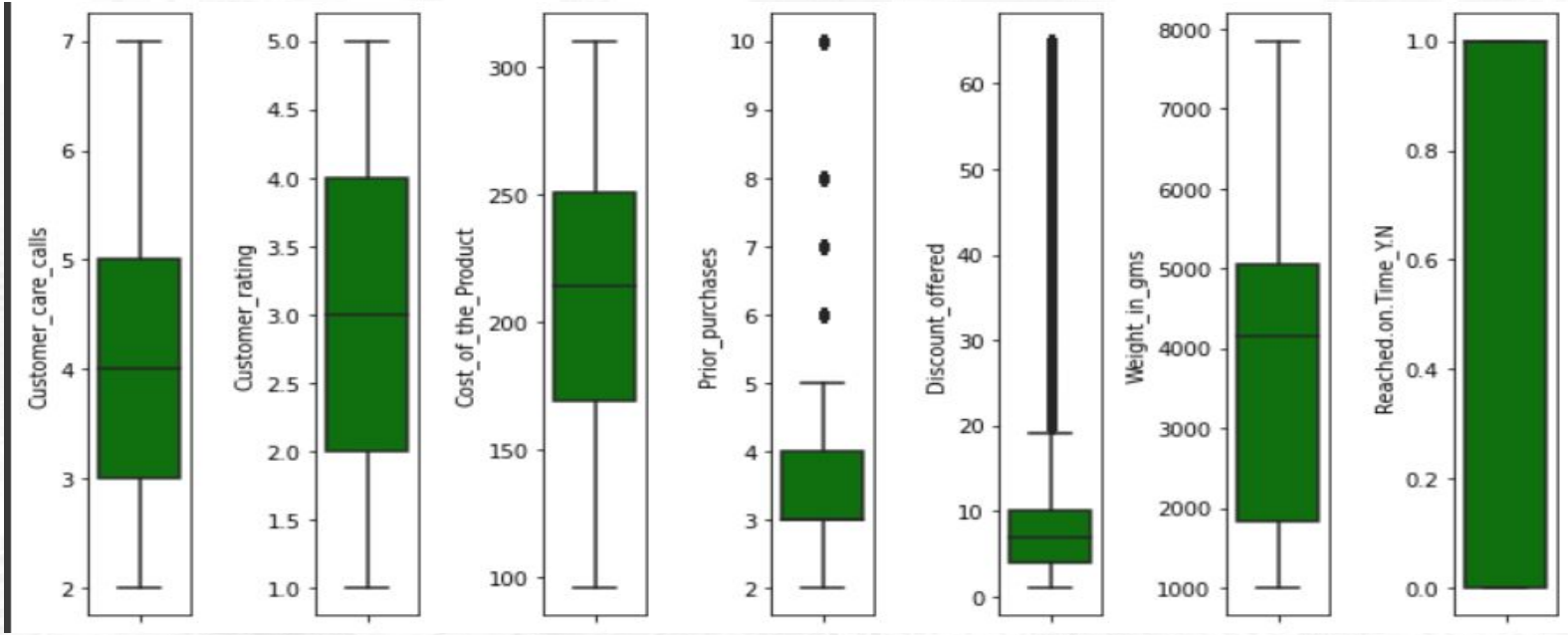
Remove Outlier using Z-score

Sebelum menghapus data outlier dengan z-score



Remove Outlier using Z-score

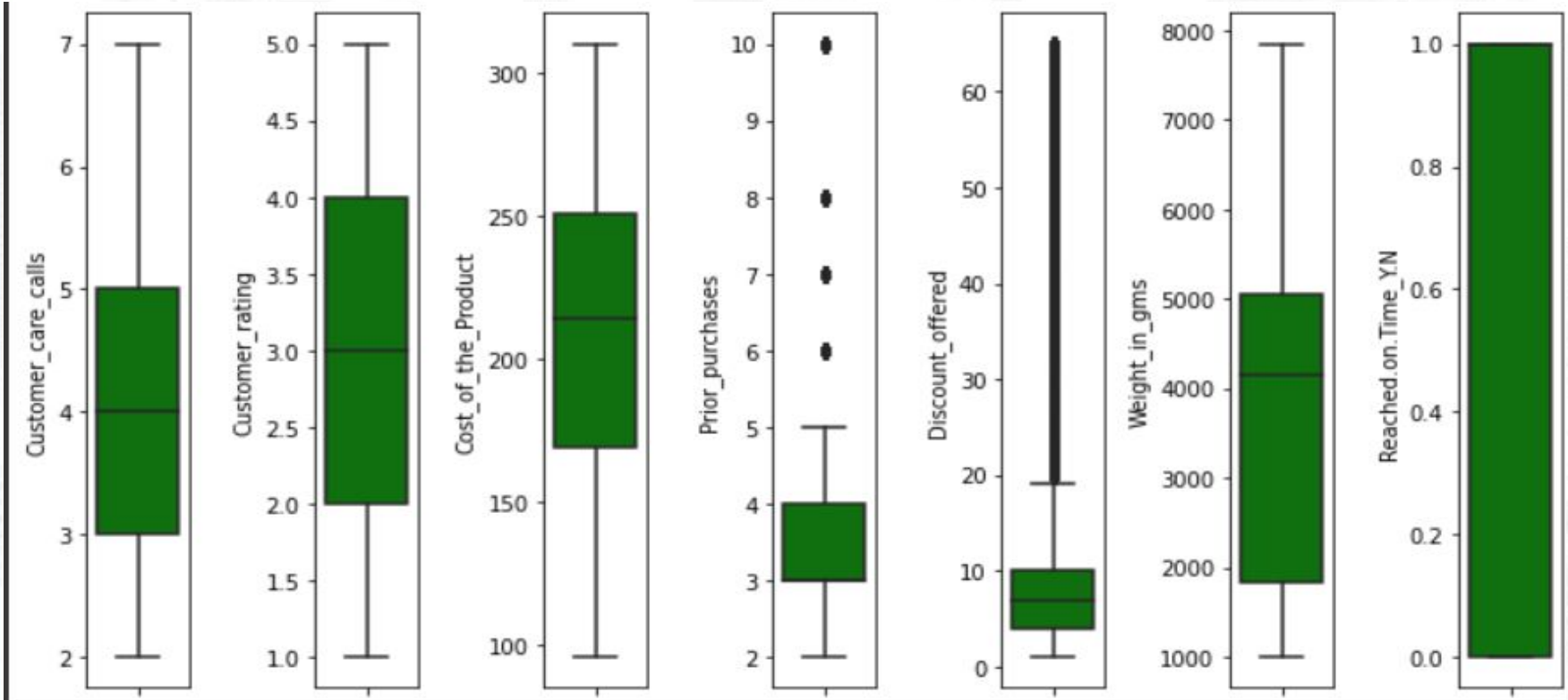
Setelah menghapus data outlier dengan z-score




```
Jumlah baris sebelum memfilter outlier: 10999
Jumlah baris setelah memfilter outlier: 10642
```

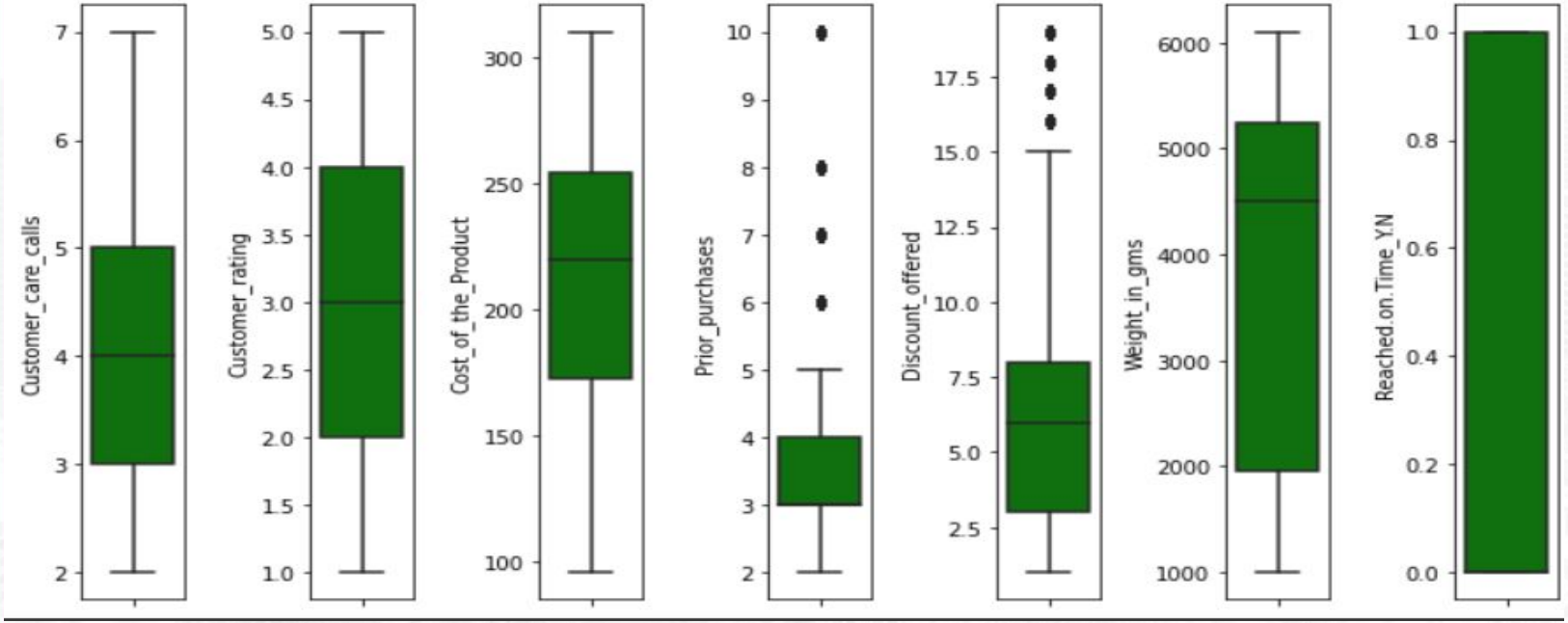
Remove Outlier using IQR untuk column discount

Sebelum menghapus data outlier dengan IQR



Remove Outlier using IQR untuk column discount

Setelah menghapus data outlier dengan IQR




```
Jumlah baris sebelum memfilter outlier: 10999
Jumlah baris setelah memfilter outlier: 8790
```

Informasi hasil remove outlier

- Data outlier yang dihapus menggunakan teknik z-score hanya terbuang sedikit yang bermula dari **10999** menjadi **10642** data.
- Data outlier yang dihapus menggunakan teknik IQR terbuang lebih banyak dibandingkan menggunakan z-score yang bermula dari **10999** menjadi **8790** data.

Rakamin Academy

Feature Engineering

- Label Encoding
- One hot encoding

Label encoding

	Gender(num)	Product_Importance(num)
3	1	1
5	0	1
6	0	0
8	0	0
10	1	1
12	0	1
16	0	1
18	1	2
22	1	0
23	1	2

- Label encoding digunakan pada colum **“Gender”** dan **“Product importace”** dikarenakan column tersebut bersifat ordinal dan juga nilainya tidak lebih dari 2

One hot encoding

	Warehouse_block_B	Warehouse_block_C	Warehouse_block_D	Warehouse_block_F	Mode_of_Shipment_Road	Mode_of_Shipment_Ship
3	1	0	0	0	0	0
5	0	0	0	1	0	0
6	0	0	1	0	0	0
8	0	0	0	0	0	0
10	0	1	0	0	0	0
12	0	0	1	0	0	0
16	0	1	0	0	0	0
18	0	0	1	0	0	1
22	0	1	0	0	0	1
23	0	0	0	1	0	1

One hot encoding

- One hot encoding digunakan pada column **“warehouse block”** dan **“Mode of shipment”** dikarenakan kedua column tersebut tidak bersifat ordinal dan memiliki value lebih dari 2. Pada teknik one hot encoding ini ditambahan parameter **drop_first=True** yang mana berfungsi untuk mengurangi 1 kolom dari hasil one hot encoding agar tidak redundant dan tidak memiliki korelasi yang tinggi antar column.