Rakamin Academy

Fantastic Two

Dokumen Laporan Final Project

(dipresentasikan setiap sesi mentoring)

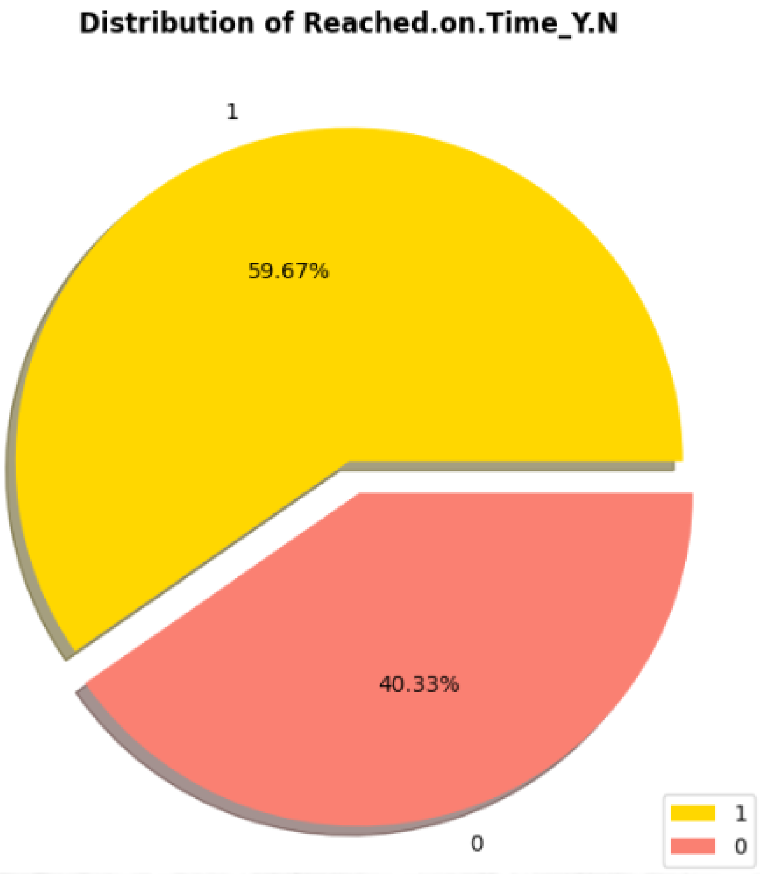
Latar Belakang Masalah

Problem :

Sebuah perusahaan e-commerce berbasis internasional ingin menemukan insight dari data pelanggan. Berdasarkan data dari perusahaan tersebut, terdapat 59,67% yang mengalami keterlambatan dalam penerimaan barang. Pihak e-commerce ingin meningkatkan performa mereka dikarenakan banyaknya pelanggan yang melakukan complain mengenai ketepatan waktu pengiriman.

Role :

Sebagai konsultan Data Scientist, kami diminta untuk memprediksi apakah pengiriman tepat waktu atau tidak berdasarkan data yang tersedia serta kami diminta untuk menganalisis faktor-faktor apa saja yang mempengaruhi ketepatan waktu pengiriman dan juga memberikan insight dan rekomendasi untuk meningkatkan performa perusahaan.



Latar Belakang Masalah

Goal :

Meningkatkan persentase ketepatan pengiriman.

Objective :

Membuat model machine learning untuk memprediksi ketepatan waktu pengiriman barang agar persentase keterlambatan menurun. Dengan demikian perusahaan dapat menggunakan model tersebut untuk menentukan keputusan bisnis sehingga meningkatkan tingkat kepuasan pelanggan.

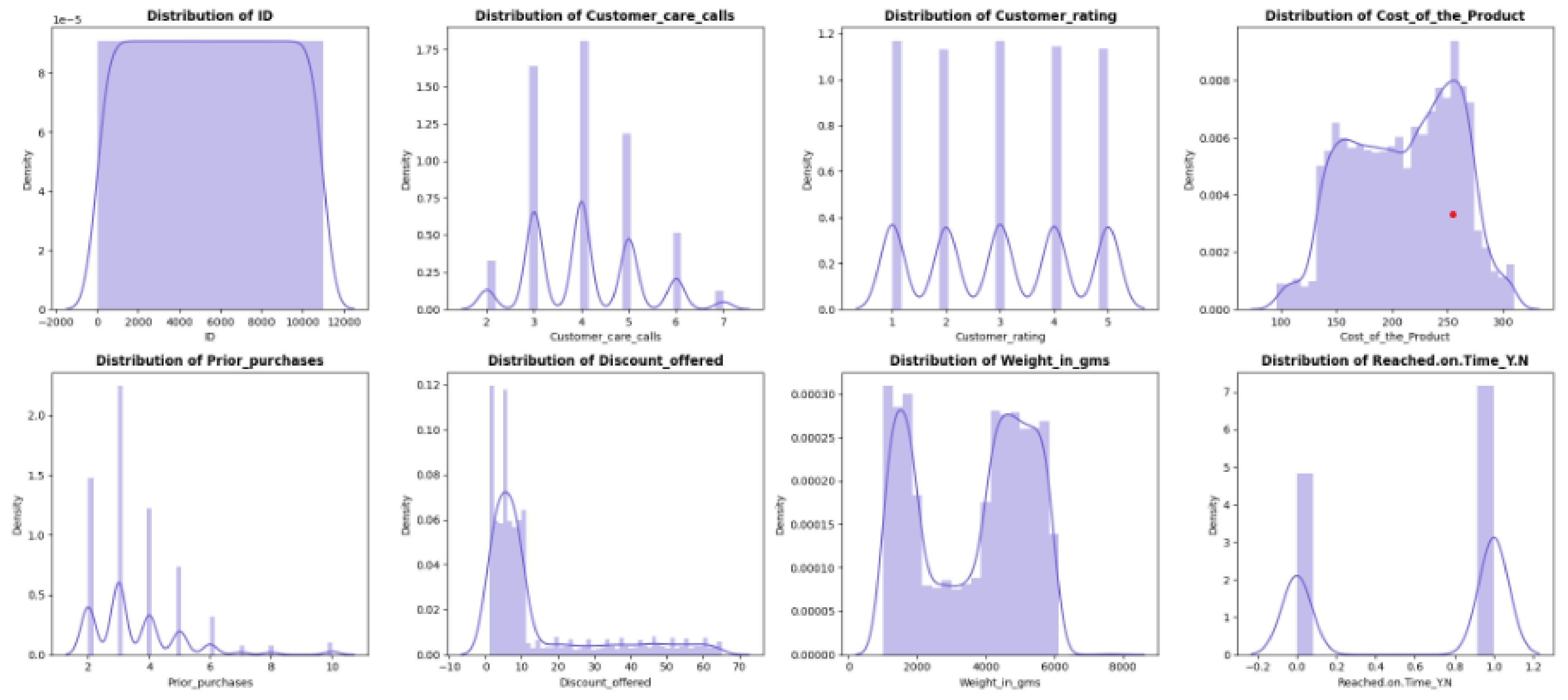
Bussiness Metrics :

Persentase ketepatan waktu pengiriman.

Rakamin Academy
School for Career Acceleration

Exploratory Data Analysis

Numeric Data Visualization

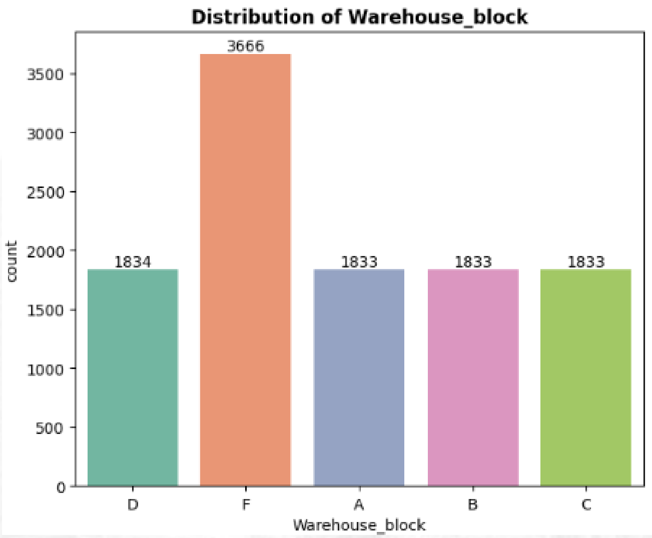


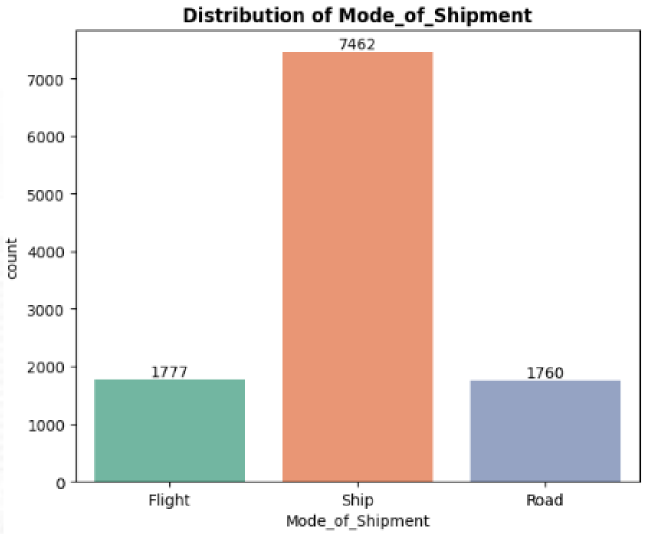
Observation:

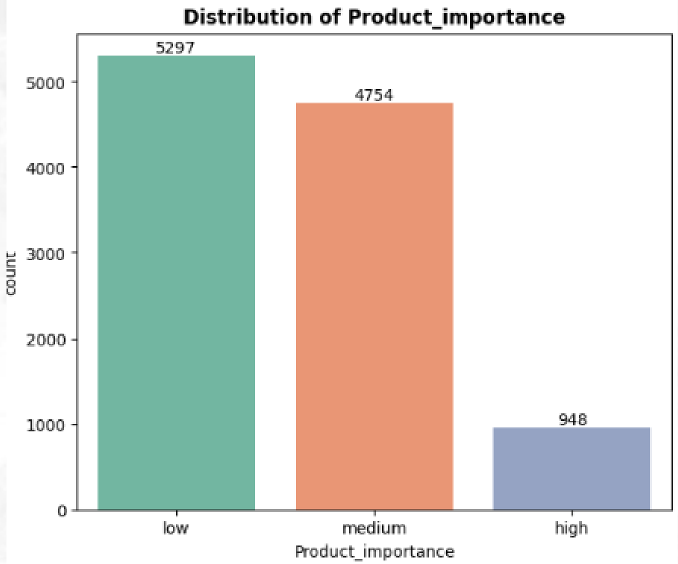
- ✓ Discount_offered and prior_purchases have distribution skew.
- ✓ Weight_in_gms and cost_of_the_product have a bimodal distribution.
- ✓ Customer_care_calls has a normal distribution.

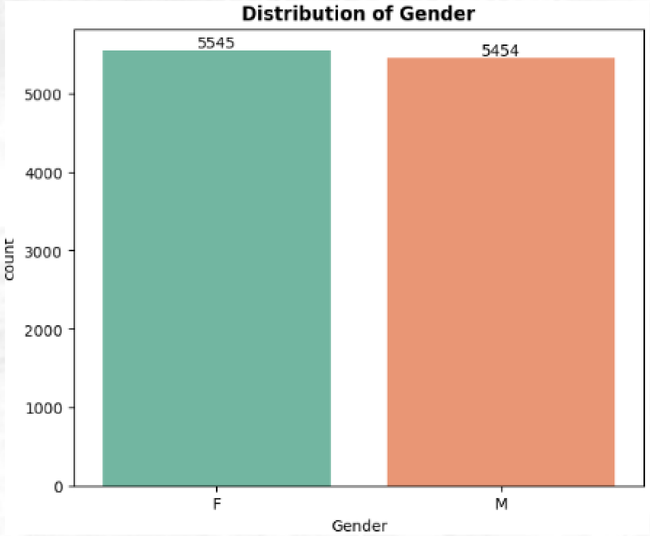
Exploratory Data Analysis

Categorical Data Visualization







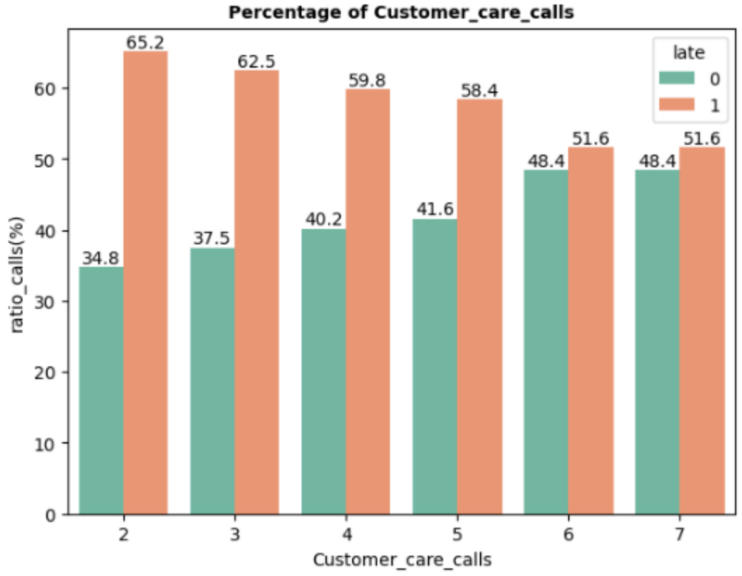


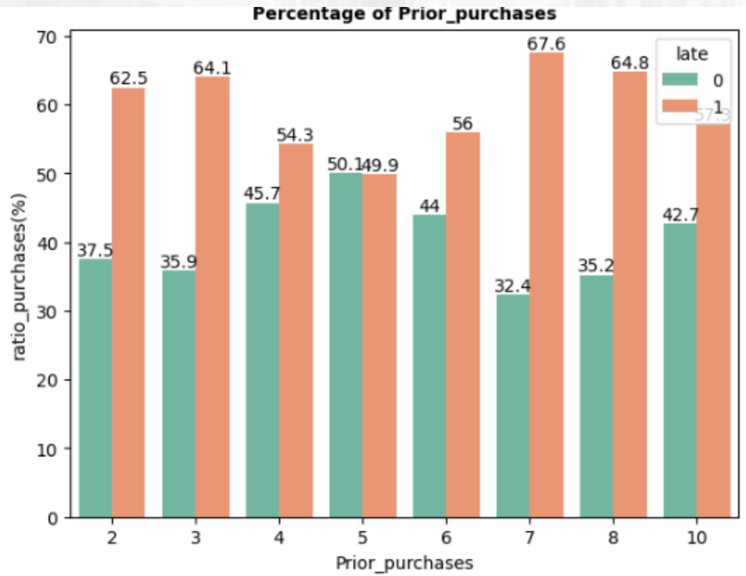
Observation:

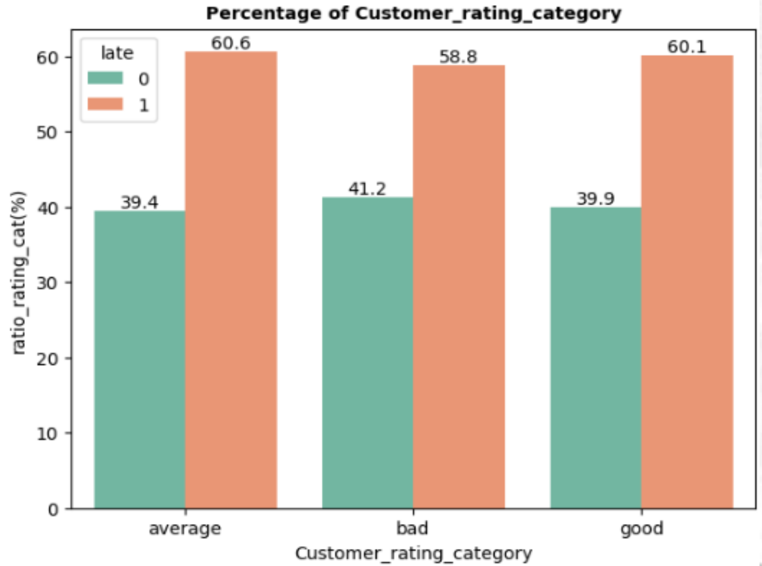
- ✓ In addition to the top value in the warehouse_block and mode _of_Shipment, they have the same value
- ✓ In product_importance, it can be seen that the low category has a large number, inversely proportional to the high category which has a small amount
- ✓ The difference in the number of women and men is not too much, almost equal

Exploratory Data Analysis

Percentage of Numerical Data Based on Unique Value







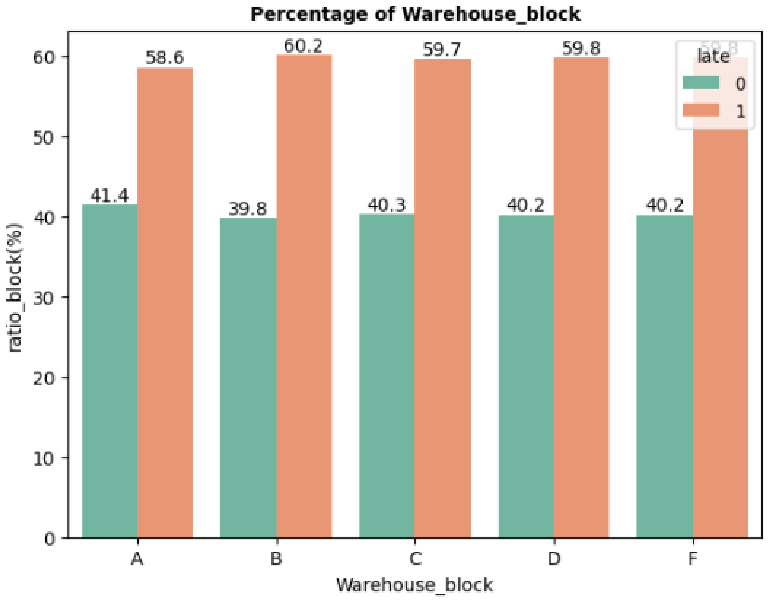
Observation:

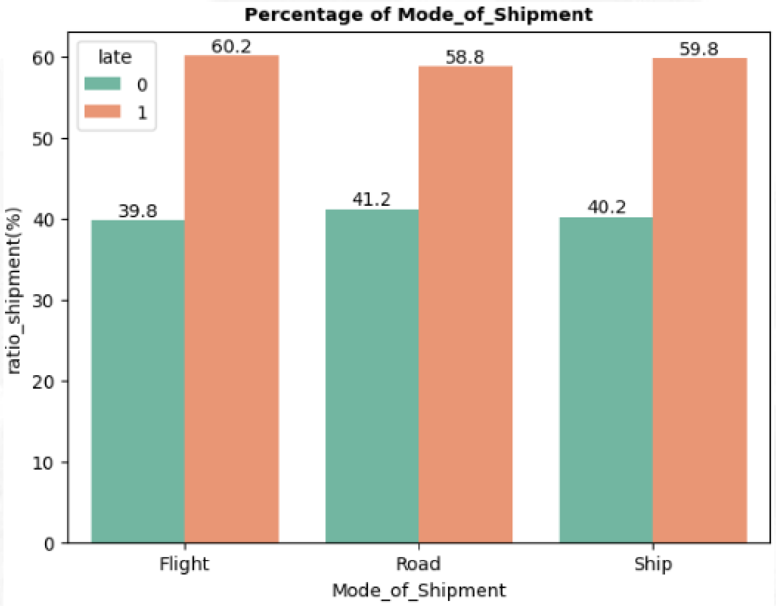
- ✓ More than 80% of customers make 3-5 calls during the shipment process.
- ✓ 60% of shipment delays are on products that have a rating of 3-5.
- ✓ The highest shipment delay occurs in customers who previously made 2-3 purchases.
- ✓ This is also influenced by the high volume of shipments, which is 60%.
- ✓ Prior_purchases above 4 times tends to experience on time shipment.
- ✓ Bad ratings are given to shipments that tend to be on time compared to good ratings.

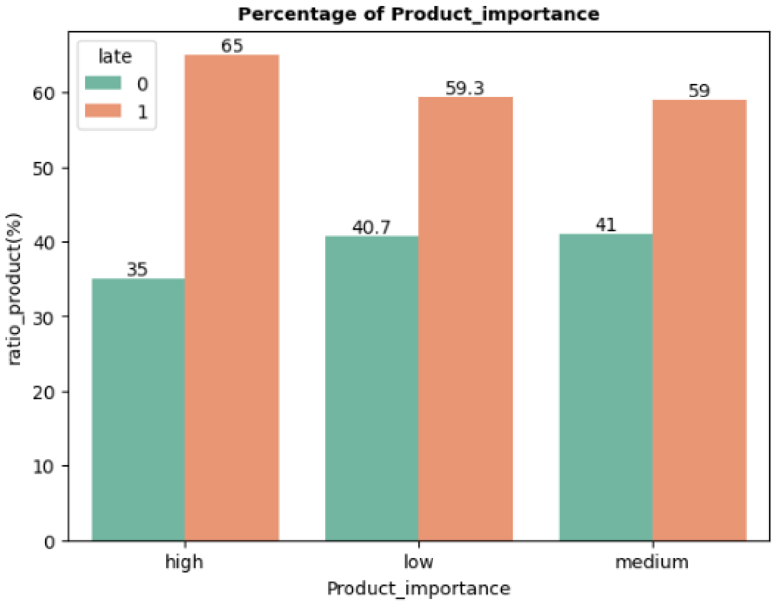
Exploratory Data Analysis

Rakamin Academy

Percentage of Categorical Data Based on Unique Value







Observation:

- ✓ Shipments from warehouse_block F have a higher volume of shipments compared to other blocks even though they have almost the same difference in the percentage of lates (< 1%). But, warehouse_block 'F' can accounts for 33% of all shipment volume.
- ✓ 68% of all deliveries are made by ship. Shipment delays by Ship tend to be higher due to higher shipping volumes.
- ✓ In contrast to product_importance high which tends to be late shipment with small shipping volume, it's only 9% of the total shipping volume.

Data Cleansing

```
# Returns the sum of the unique values for each column
df.nunique()
```

```
ID                      10999
Warehouse_block             5
Mode_of_Shipment            3
Customer_care_calls         6
Customer_rating             5
Cost_of_the_Product       215
Prior_purchases             8
Product_importance          3
Gender                      2
Discount_offered           65
Weight_in_gms            4034
Reached.on.Time_Y.N         2
dtype: int64
```

```
# Missing Value Check
df.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
```

```
# Check Duplicates Data
df.duplicated().sum()
```

```
0
```

Observation :

- ✓ Data contains 12 column with 10999 rows
- ✓ The data type in each column is appropriate
- ✓ No missing values found
- ✓ No Duplicate Data

Pre-processing (Feature Encoding)

Rakamin Academy
School for Career Acceleration

```python
# Label Encoding
label_encoder = preprocessing.LabelEncoder()

# Fit and transform
df_encod['Product_importance'] = label_encoder.fit_transform(df_encod['Product_importance'])
df_encod['Gender'] = label_encoder.fit_transform(df_encod['Gender'])
```

```python
df_encod.head()
```

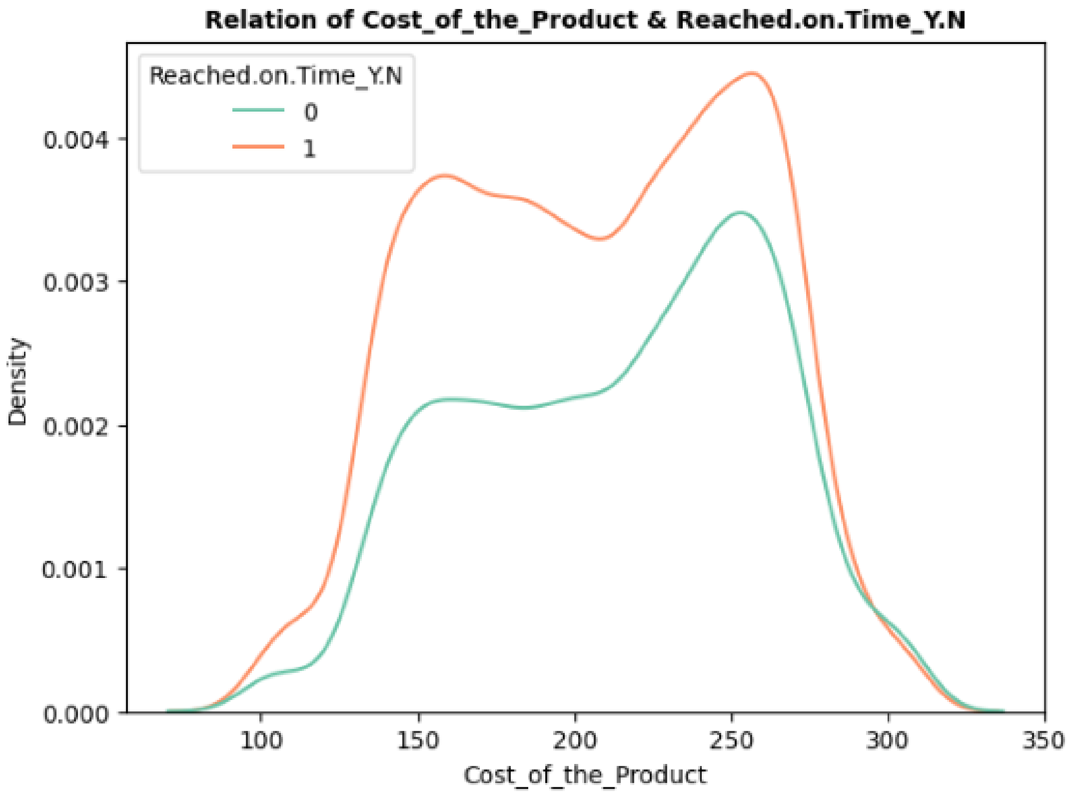
	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Cost_of_the_Product	Prior_purchases	Product_importance	Gender
0	D	Flight	4	177	3	1	0
1	F	Flight	4	216	2	1	1
2	A	Flight	2	183	4	1	1
3	B	Flight	3	176	4	2	1

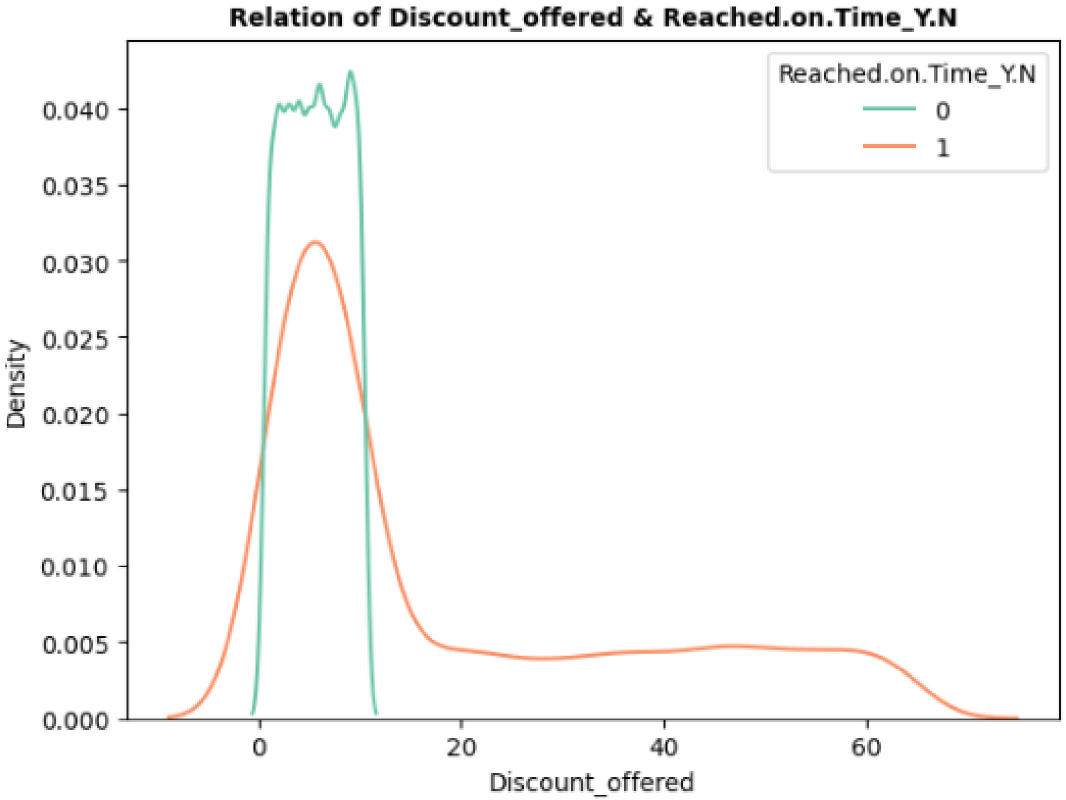
```python
# One-hot Encoding
for column in ['Mode_of_Shipment', 'Warehouse_block']:
  df_encod = onehot_encode(df_encod, column=column)
```

Customer_rating_category	Mode_of_Shipment_Flight	Mode_of_Shipment_Road	Mode_of_Shipment_Ship	Warehouse_block_A	Warehouse_block_B	Warehouse_block_
1	1	0	0	0	0	
3	1	0	0	0	0	
1	1	0	0	1	0	
2	1	0	0	0	1	

- ✓ Product Importance and Gender use encoding labels.
- ✓ The customer rating category uses ordinal encoding.
- ✓ The Mode of Shipment and Warehouse Block use One Hots Encoding.

STAGE2 - Insights

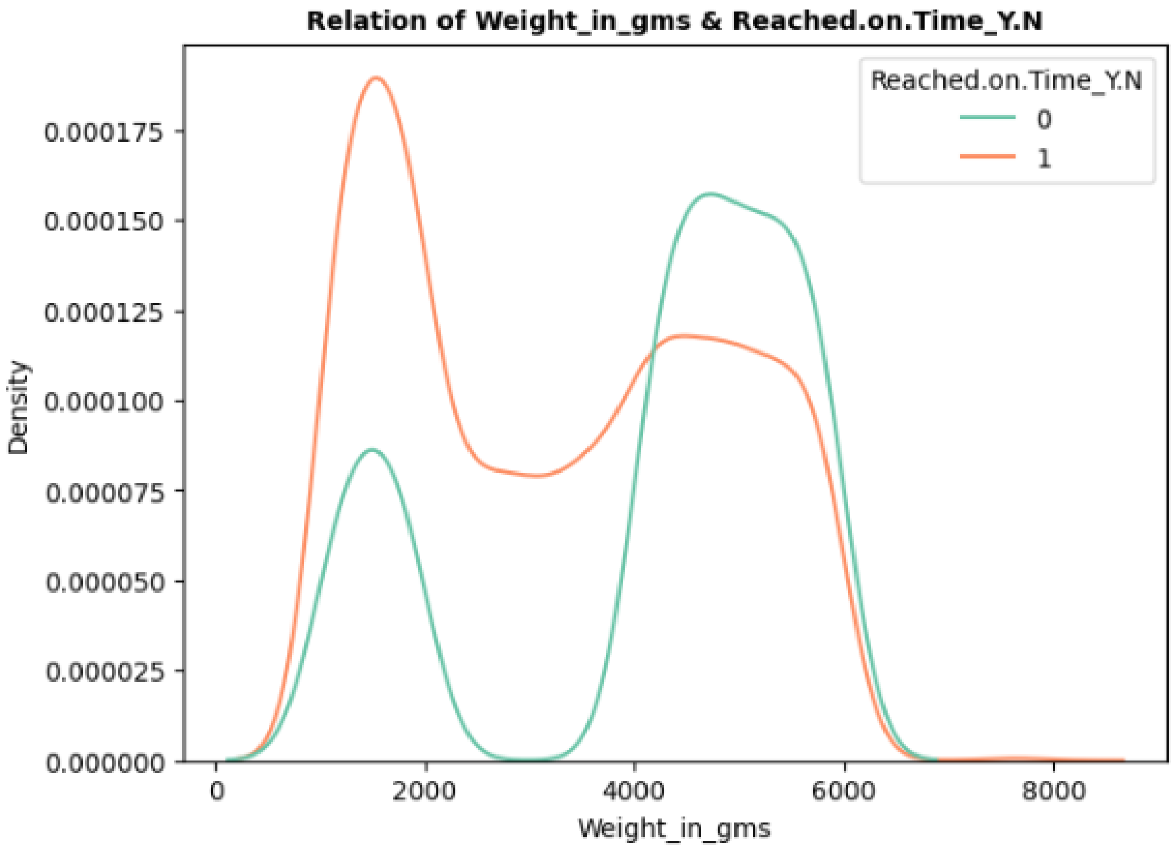




- ✓ The higher the cost_of_the_product, the greater the possibility of on time delivery

- ✓ Shipments that have a discount of less than 13.8 tend to experience on time shipments, these shipments account for 77% of the total volume.
- ✓ All shipments that have a discount offer greater than 13.8 experience delays.

STAGE2 - Insights



Observation:

- ✓ Shipment of products weighing less than 4000 grams (4kg) tends to be late, while those more than 4kg tend to be on time. Shipments more than 4kg account for 56% of the total volume
- ✓ All shipment of products weighing 2370-3739 grams and more than 6477 grams are delayed

STAGE2 - Modelling Experiments

Model / Original Data	Before Recall		After Recall	
	Test	Train	Test	Train
Logrec	0.661	0.673	0.671	0.674
Knn	0.686	0.796	0.657	1
Decision Tree	0.697	1	0.69	1
XGBoost	0.643	0.887	0.632	1
AdaBoost	0.55	0.579	0.614	0.623
Random Forest	0.633	1	0.632	1

Model / Log-Transformation	Before Recall		After Recall	
	Test	Train	Test	Train
Logrec	0.747	0.754	0.767	0.774
Knn	0.683	0.802	0.662	1
Decision Tree	0.698	1	0.663	0.827
XGBoost	0.643	0.887	0.669	0.721
AdaBoost	0.55	0.579	0.614	0.623
Random Forest	0.64	1	0.634	1

Model / IQR	Before Recall		After Recall	
	Test	Train	Test	Train
Logrec	0.499	0.501	0.498	0.501
Knn	0.510	0.699	0.466	1
Decision Tree	0.574	1	0.571	1
XGBoost	0.471	0.835	0.546	0.946
AdaBoost	0.4	0.423	0.428	0.428
Random Forest	0.466	1	0.462	0.999

Model / Z-Score	Before Recall		After Recall	
	Test	Train	Test	Train
Logrec	0.646	0.664	0.66	0.668
Knn	0.674	0.781	0.646	1
Decision Tree	0.712	1	0.694	1
XGBoost	0.637	0.901	0.69	0.736
AdaBoost	0.594	0.606	0.623	0.616
Random Forest	0.614	1	0.616	1

Recall values before & after hyperparameters with multiple dataset conditions:

- ✓ Without handle outlier & skew
- ✓ With log transformation
- ✓ Handle outliers with IQR
- ✓ Handle outliers with Z-Score

STAGE2 - Modelling Experiments

Rakamin Academy
School for Career Acceleration

Observation:

- ✓ Based on the recall values obtained by each model, we chose logistic regression as the best model for predicting on time delivery. The recall value increases significantly by using features that have been log transformed.
- ✓ After modeling, it was found that the percentage of on time delivery increased by 50.7%.

```
----- Existing -----
                          count percentage
Delivery :                10999
Late :                    6563 , 59.7 %
On Time :                 4436 , 40.3 %

----- After Modeling -----
                          count percentage
Delivery :                10999
Late :                    6563 , 59.7 %
  Predicted Late :        5034 , 76.7 %
  Predicted On Time :     1529 , 23.3 %
Late After Pred :         1529 , 13.9 %
On Time :                 4436 , 40.3 %
On Time After Pred :      9470 , 81.79499999999999 %
On Time Growth rate :     50.7 %
```

STAGE3 - Executive Summary & Recommendation

Business recommendations

- ✓ Add estimated delivery time feature delivery status
- ✓ Notifications in real time
- ✓ Notification to customers regarding delivery delays due to Harbolnas
- ✓ Expand the choice of expeditions, especially during big events

Pembagian Tugas

Hafidz Alawy
- ✓ Ketua kelompok
- ✓ Peran utama dalam pengerjaan kodingan tiap stage
- ✓ Presenter saat final project

Robby Dipomiharjo
- ✓ Moderator saat mentoring
- ✓ Bantu dalam kodingan
- ✓ presenter saat final project

Annisa Yovinda
- ✓ Moderator yang membuat jadwal mentoring
- ✓ Moderator saat final project
- ✓ Pembuat ppt final project

Jedi Manullang
- ✓ Pembuat laporan tiap stage
- ✓ Bantu dalam kodingan
- ✓ Penjawab pertanyaan saat final project

F. Artha
- ✓ Notulensi saat mentoring
- ✓ Pembuat ppt final project
- ✓ Penjawab pertanyaan saat final project

Marius Iddo
- ✓ Notulensi saat mentoring
- ✓ Pembuat ppt final project
- ✓ Penjawab pertanyaan saat final project

Romaito Silalahi
- ✓ Pembuat laporan tiap stage
- ✓ Bantu dalam kodingan
- ✓ Penjawab pertanyaan saat final project