KELOMPOK 2

Stage 1

Mentor : Mas Mirza

Elvis Muh. Rizqy

Fuji Resti M

Ni Kadek Yulia Cyntia Dewi

Haolia

Luthfi Adnan Rahmantyo

Descriptive Statistics

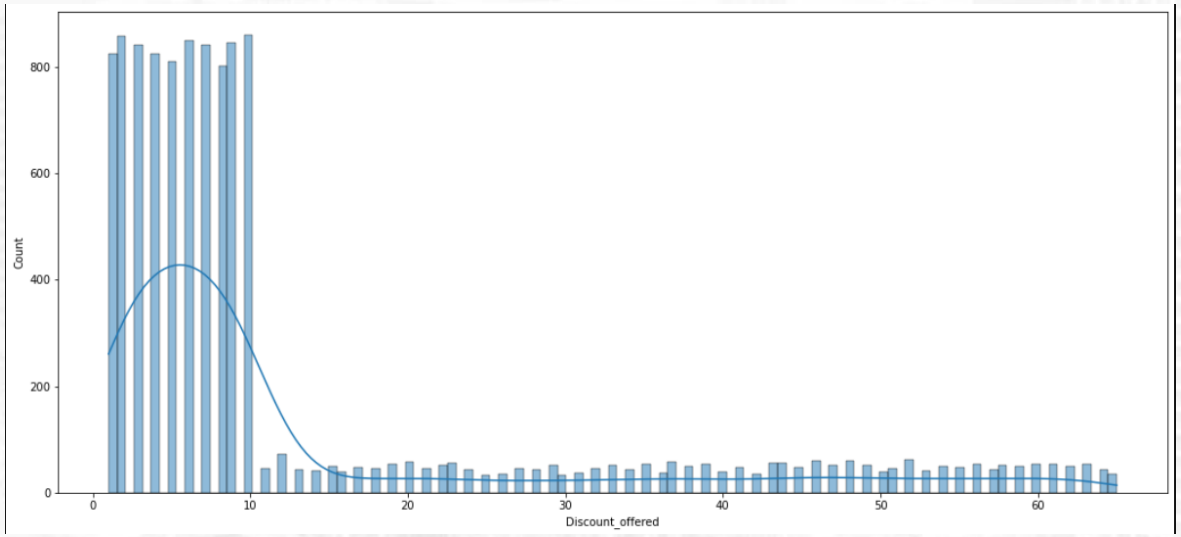
Melakukan pengecekan data dan menunjukkan rangkuman statistic dari dataset dengan function info() dan describe()

```
 #   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)
```

	ID	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
count	10999.00000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000
mean	5500.00000	4.054459	2.990545	210.196836	3.567597	13.373216	3634.016729	0.596691
std	3175.28214	1.141490	1.413603	48.063272	1.522860	16.205527	1635.377251	0.490584
min	1.00000	2.000000	1.000000	96.000000	2.000000	1.000000	1001.000000	0.000000
25%	2750.50000	3.000000	2.000000	169.000000	3.000000	4.000000	1839.500000	0.000000
50%	5500.00000	4.000000	3.000000	214.000000	3.000000	7.000000	4149.000000	1.000000
75%	8249.50000	5.000000	4.000000	251.000000	4.000000	10.000000	5050.000000	1.000000
max	10999.00000	7.000000	5.000000	310.000000	10.000000	65.000000	7846.000000	1.000000

Visualisasi

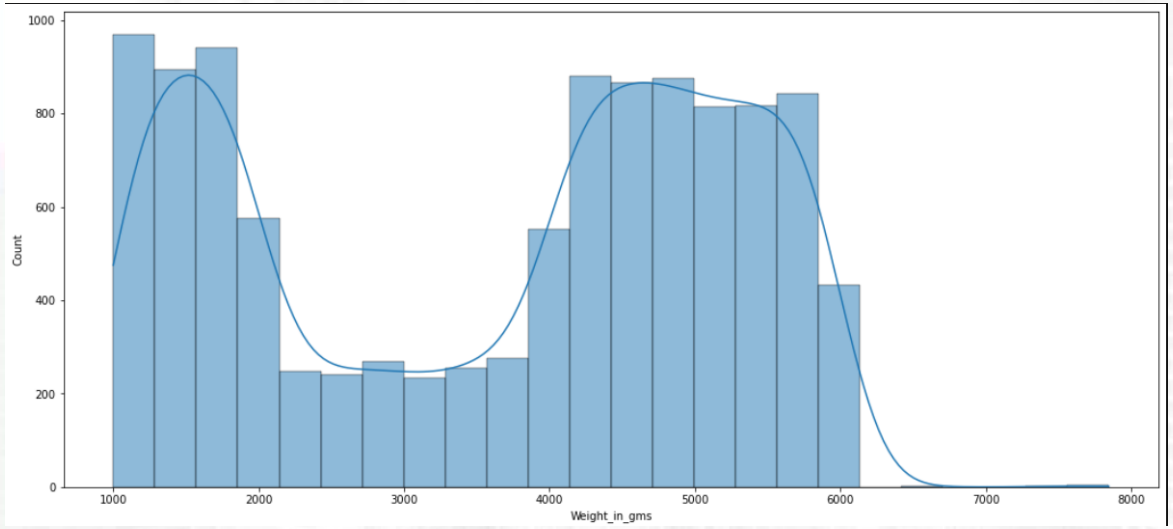
```
plt.figure(figsize=(18,8))
ax = sns.histplot(df, x='Discount_offered', linewidth=0.4, kde = True)
```



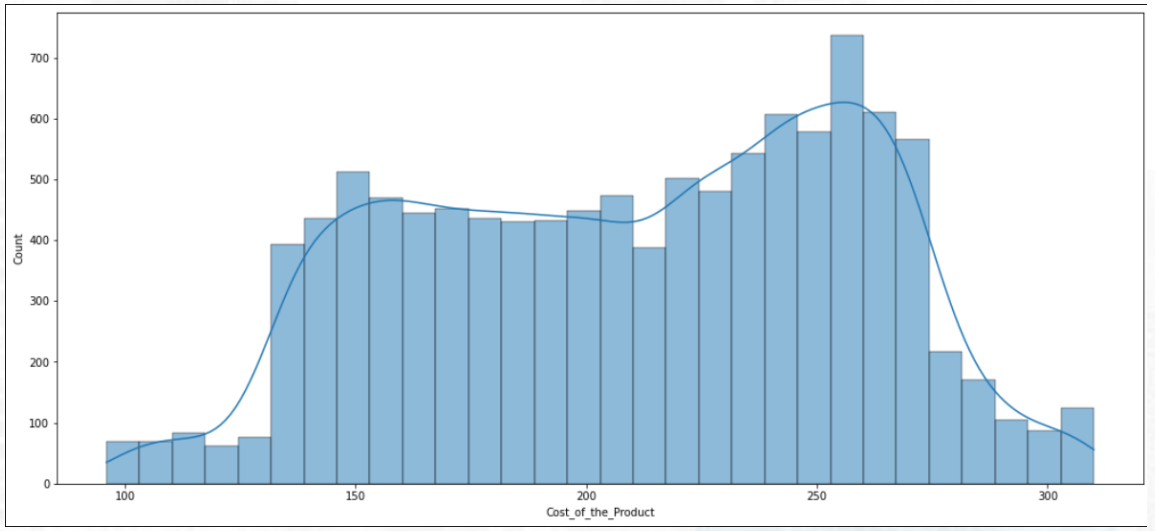
- Tidak ada kolom dengan tipe data kurang sesuai dan memiliki nilai kosong
- Kolom Discount offered membentuk Positive Skewed karena nilai Mean>Median

Visualisasi

```
plt.figure(figsize=(18,8))
ax = sns.histplot(df, x='Weight_in_gms', linewidth=0.4, kde = True)
```




```
plt.figure(figsize=(18,8))
ax = sns.histplot(df, x='Cost_of_the_Product', linewidth=0.4, kde = True)
```



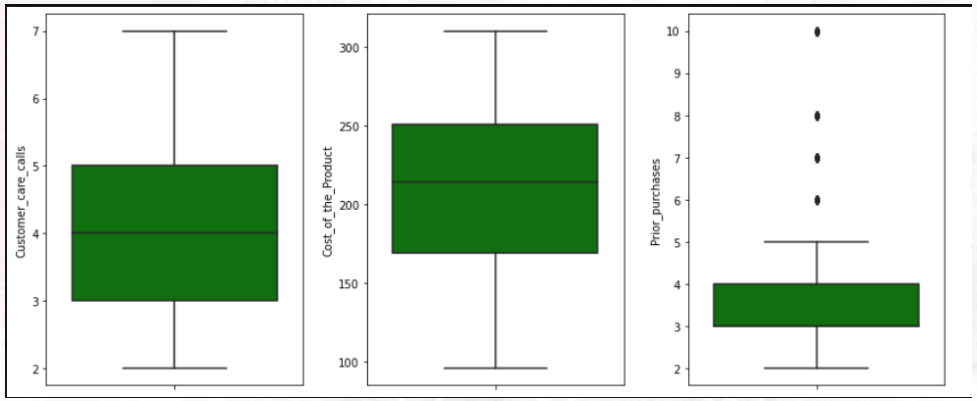
- Pada kolom Cost of_Product dan weight_in_gms membentuk Negatif Skewed karena nilai Median > mean.
- Pada kolom Cost_of_Product dan weight_in_gms membentuk Negatif Skewed sehingga harus dilakukan Feature Transformation agar nantinya menjadi Normal Skewed
- Pada kolom Discount_offered membentuk Positif skewed sehingga harus dilakukan Feature Transformation agar nantinya menjadi Normal Skewed.
- Kolom ID memiliki banyak nilai Unique sehingga harus di drop.

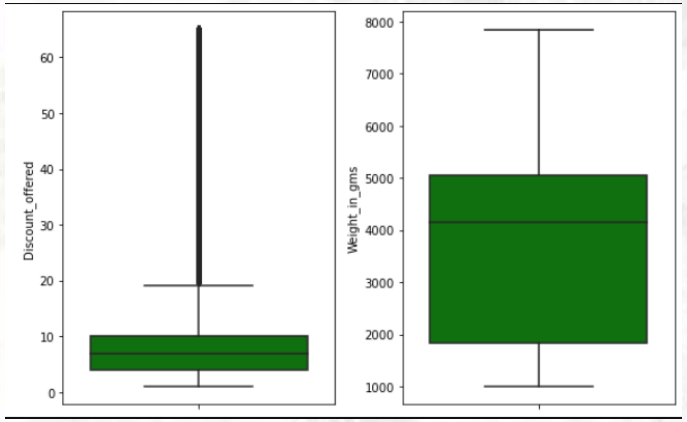
Univariate Analysis

Terdapat 2 jenis data pada dataset E-Commerce Shipping Data

Numerical

```
plt.figure(figsize=(20,5))
for i in range(0, len(numerical)):
    plt.subplot(1, len(numerical), i+1)
    sns.boxplot(y=df[numerical[i]], color='green', orient='v')
    plt.tight_layout()
```





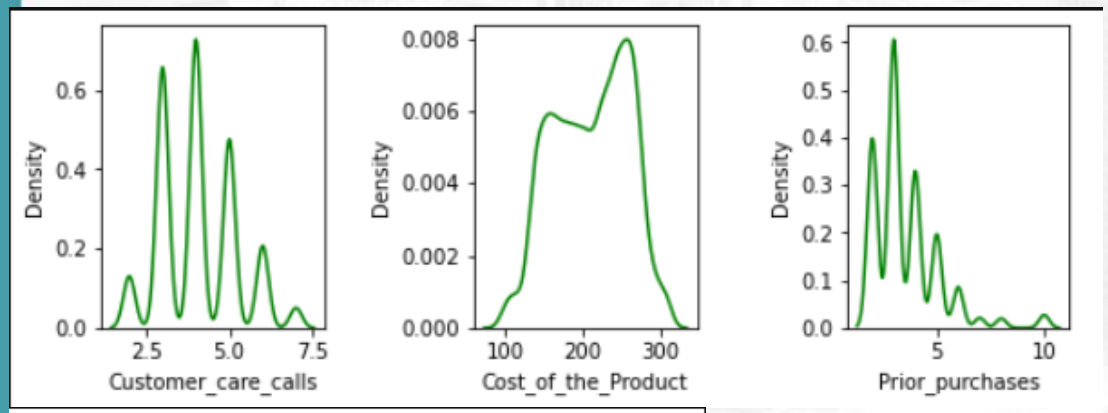
Untuk boxplot, hal paling penting yang harus kita perhatikan adalah keberadaan outlier.

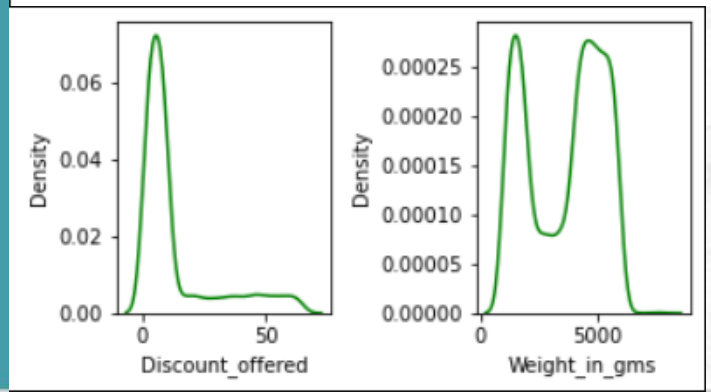
- Outlier pada kolom prior purchases tidak perlu dibuang dikarenakan nilainya masih dalam batas wajar (kecuali ada nilai yang < 0 sehingga harus dilakukan drop pada kolom tersebut).
- Outlier pada kolom Discount offered tidak perlu dibuang dikarenakan discount yang diberikan masih dalam batas wajar

Univariate Analysis

Numerical

```
plt.figure(figsize=(20,5))
for i in range(0, len(numerical)):
    plt.subplot(2, 8, i+1)
    sns.kdeplot(x=df[numerical[i]], color='green')
    plt.xlabel(numerical[i])
    plt.tight_layout()
```





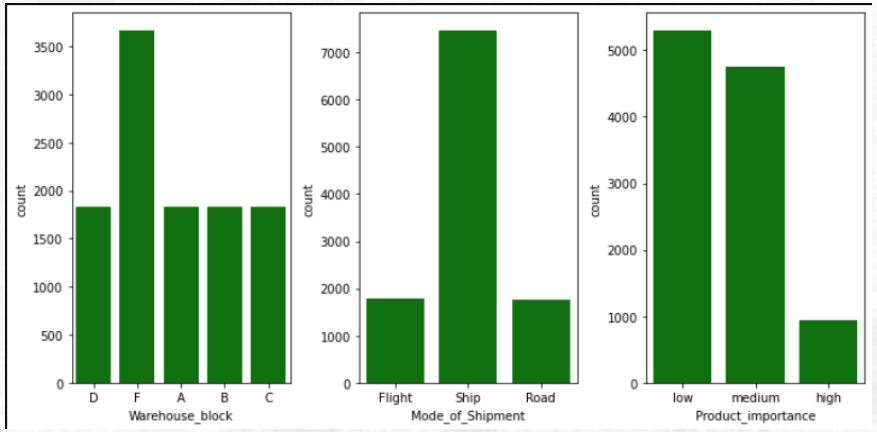
Untuk distribution plot, hal utama yang perlu diperhatikan adalah bentuk distribusi:

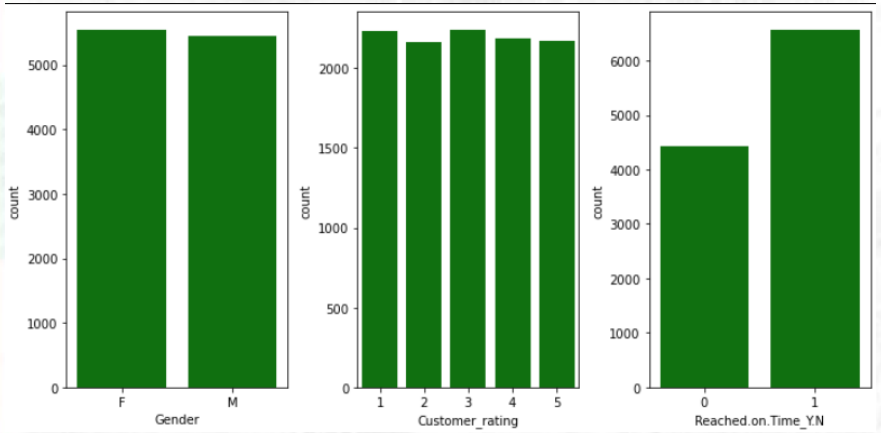
- Pada kolom cost_of_product dan weight_in_gms membentuk grafik berbentuk bimodal sehingga harus di tranformasi agar nantinya berbentuk Normal skewed .
- Pada kolom customer_care_calls dan prior purchases membentuk grafik berbentuk multimodal sehingga harus dilakukan transofrmasi agar nantinya menjadi Normal Skewed.
- Pada kolom Discount_offered membentuk Positif skewed sehingga harus dilakukan transformation agar nantinya menjadi Normal Skewed.

Univariate Analysis

Categorical

```
plt.figure(figsize=(20,5))
for i in range(0, len(categorical_1)):
    plt.subplot(1, len(categorical_1), i+1)
    sns.countplot(x=df[categorical_1[i]], color='green', orient='v')
    plt.tight_layout()
```





- Data yang terdapat dalam kolom categoricals masih dalam batas wajar karena tidak ada nilai yang mendominasi dan kategori dari tiap kolom tidak terlalu banyak sehingga feature masih bisa dipertahankan.

Hal yang harus dilakukan ketika preprocessing :

- Pada kolom cost_of_product dan weight_in_gms membentuk grafik berbentuk bimodal sehingga harus di tranformasi agar nantinya berbentuk Normal skewed
- Pada kolom customer_care_calls dan prior purchases membentuk grafik berbentuk multimodal sehingga harus dilakukan transofrmasi agar nantinya menjadi Normal Skewed
- Pada kolom Discount_offered membentuk Positif skewed sehingga harus dilakukan transformati agar nantinya menjadi Normal Skewed

Multivariate Analysis

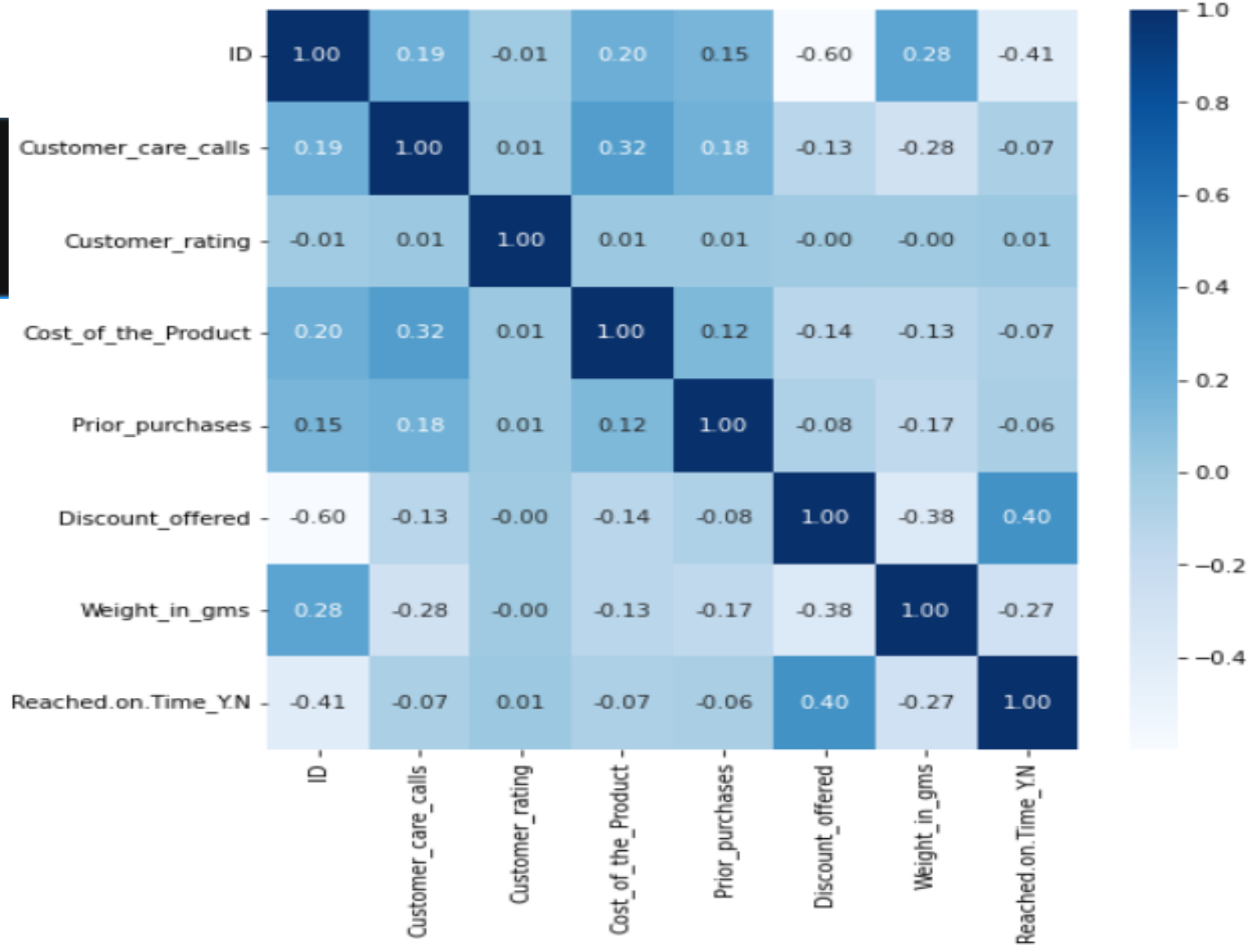
Analisis yang menggunakan lebih dari atau sama dengan tiga variabel

```
df.corr()
```

	ID	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
ID	1.000000	0.188998	-0.005722	0.196791	0.145369	-0.598278	0.278312	-0.411822
Customer_care_calls	0.188998	1.000000	0.012209	0.323182	0.180771	-0.130750	-0.276615	-0.067126
Customer_rating	-0.005722	0.012209	1.000000	0.009270	0.013179	-0.003124	-0.001897	0.013119
Cost_of_the_Product	0.196791	0.323182	0.009270	1.000000	0.123676	-0.138312	-0.132604	-0.073587
Prior_purchases	0.145369	0.180771	0.013179	0.123676	1.000000	-0.082769	-0.168213	-0.055515
Discount_offered	-0.598278	-0.130750	-0.003124	-0.138312	-0.082769	1.000000	-0.376067	0.397108
Weight_in_gms	0.278312	-0.276615	-0.001897	-0.132604	-0.168213	-0.376067	1.000000	-0.268793
Reached.on.Time_Y.N	-0.411822	-0.067126	0.013119	-0.073587	-0.055515	0.397108	-0.268793	1.000000

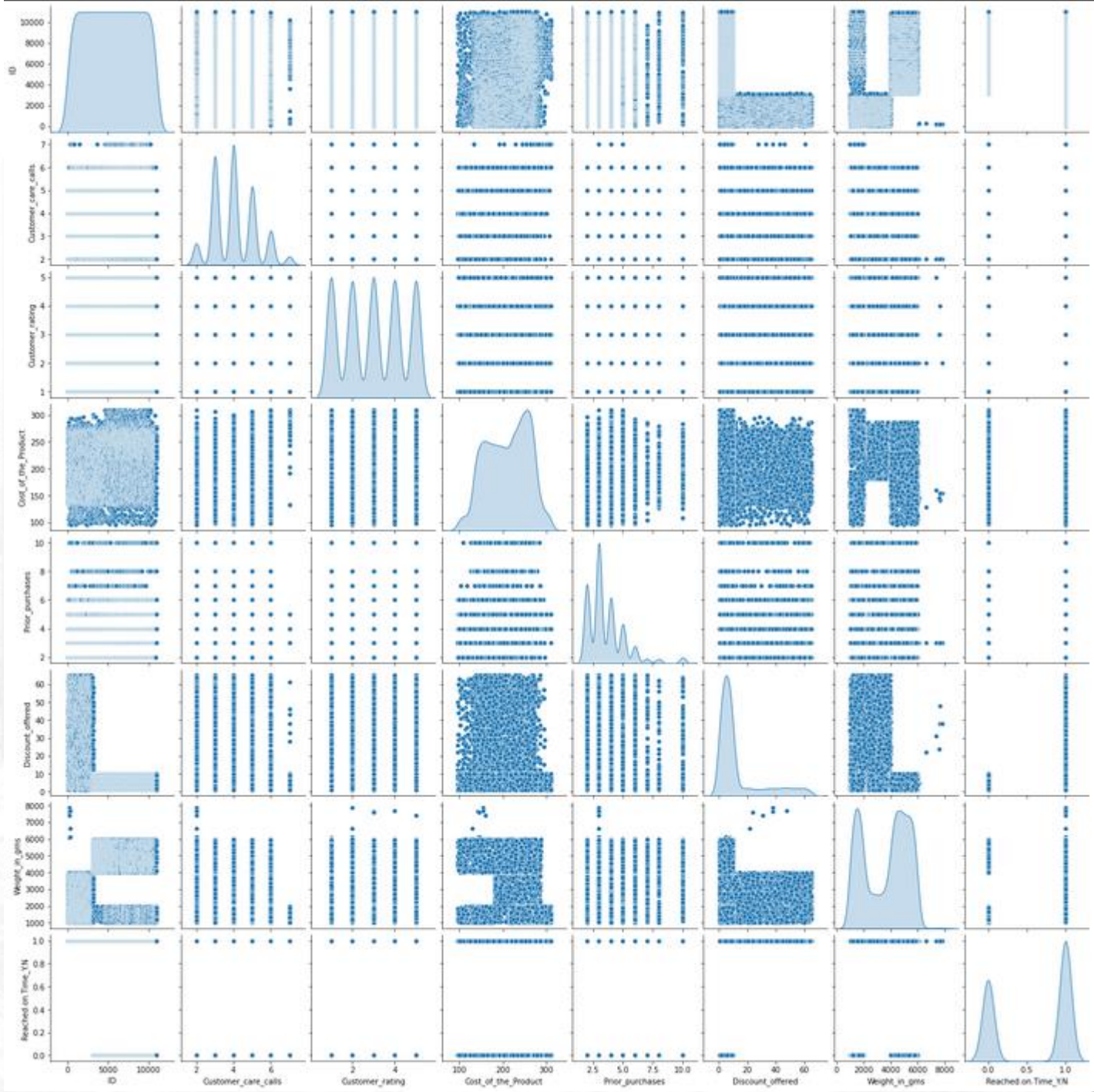
Multivariate Analysis

```
# correlation heatmap
plt.figure(figsize=(8, 8))
sns.heatmap(df.corr(), cmap='Blues',
            annot=True, fmt='.2f');
```



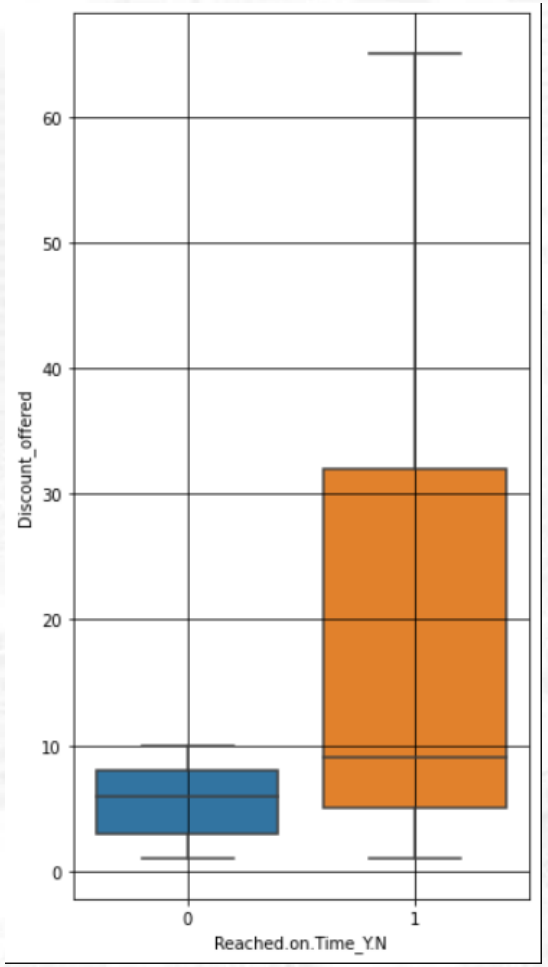
Multivariate Analysis

```
plt.figure(figsize=(12,8))
sns.pairplot(df, diag_kind='kde');
```



Multivariate Analysis

```
plt.figure(figsize=(5,10))
sns.boxplot(x='Reached.on.Time_Y.N', y='Discount_offered', data=df);
pyplot.grid(True,color='black')
plt.show()
```

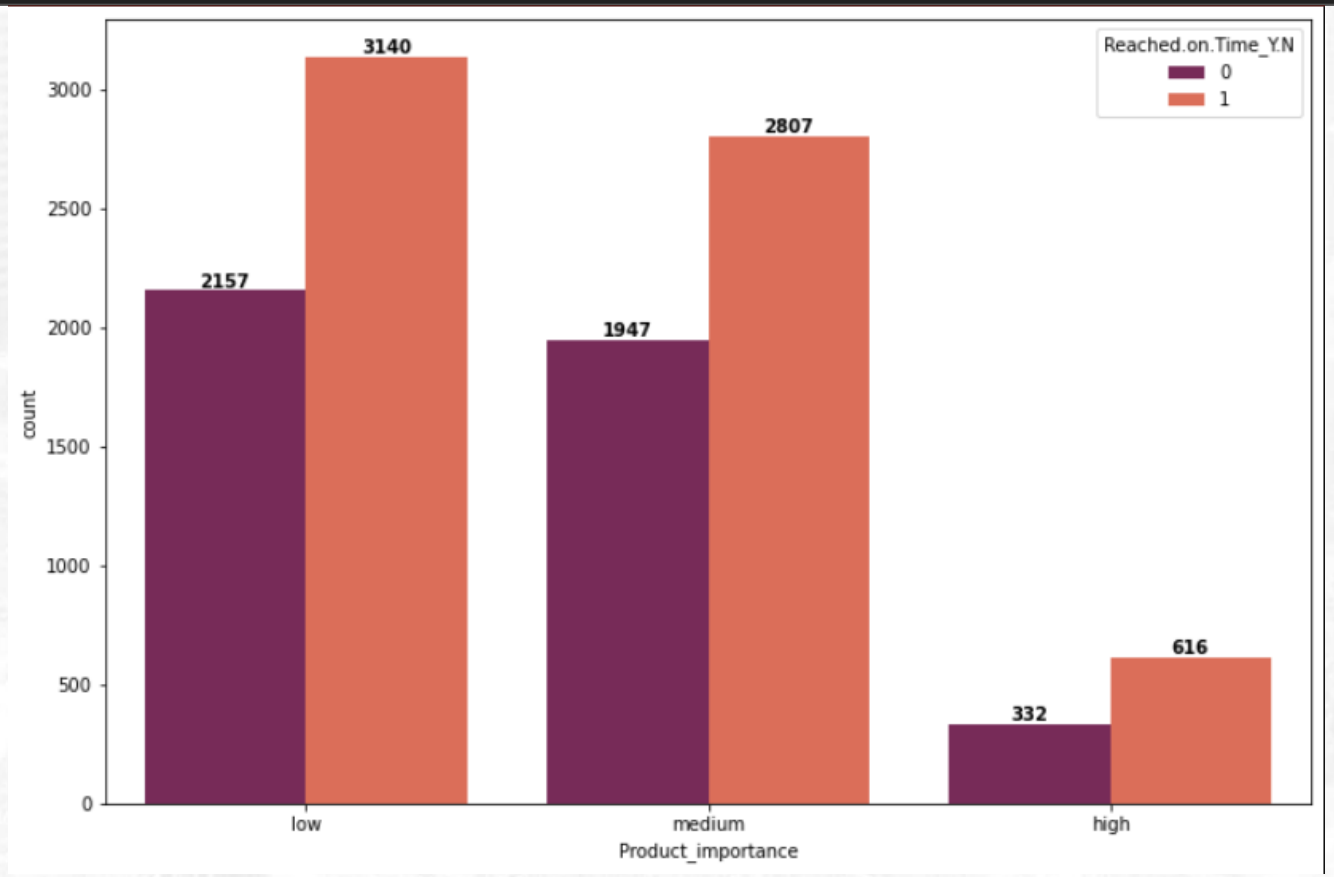


- Korelasi dari kolom Reached.on.Time_Y.N dan Discount_offered menujukkan hubungan korelasi positif cukup kuat sehingga merupakan strong potential feature dan harus dipertahankan. (Semakin banyak discount yang diberikan dapat menyebabkan keterlambatan dalam pengiriman paket)
- Korelasi dari kolom Weight_in_gms dan Reached.on.Time_Y.N menunjukkan hubungan korelasi negatif cukup kuat sehingga berpontensi menjadi potential feature.
- Tidak ada fitur yang redundan dikarenakan nilai korelasi antar fitur tidak ada yang lebih besar dari 0.4.
- Korelasi dari kolom Customer_rating, Weight_in_gms, dan Discount_offered sangat lemah, menandakan fitur tersebut tidak dapat dijadikan feature.
- Berdasakan grafik yang ditujukkan diatas kolom Reached on Time Y.N dan Discount_offered menujukkan hubungan korelasi positif, kami telah melakukan visualisasi data yang menunjukkan bahwa semakin banyak diskon yang diberikan makan semakin banyak juga keterlambatan dalam pengiriman.

Business Insight

Setelah melakukan EDA, kami menemukan beberapa business insight sebagai berikut.

```
fig, ax = plt.subplots(figsize=(12,8))
sns.countplot('Product_importance', hue = 'Reached.on.Time_Y.N', data = df, palette='rocket')
for label in ax.containers: #Untuk bikin angka diatas bar
    ax.bar_label(label, fontweight='bold')
plt.show()
```

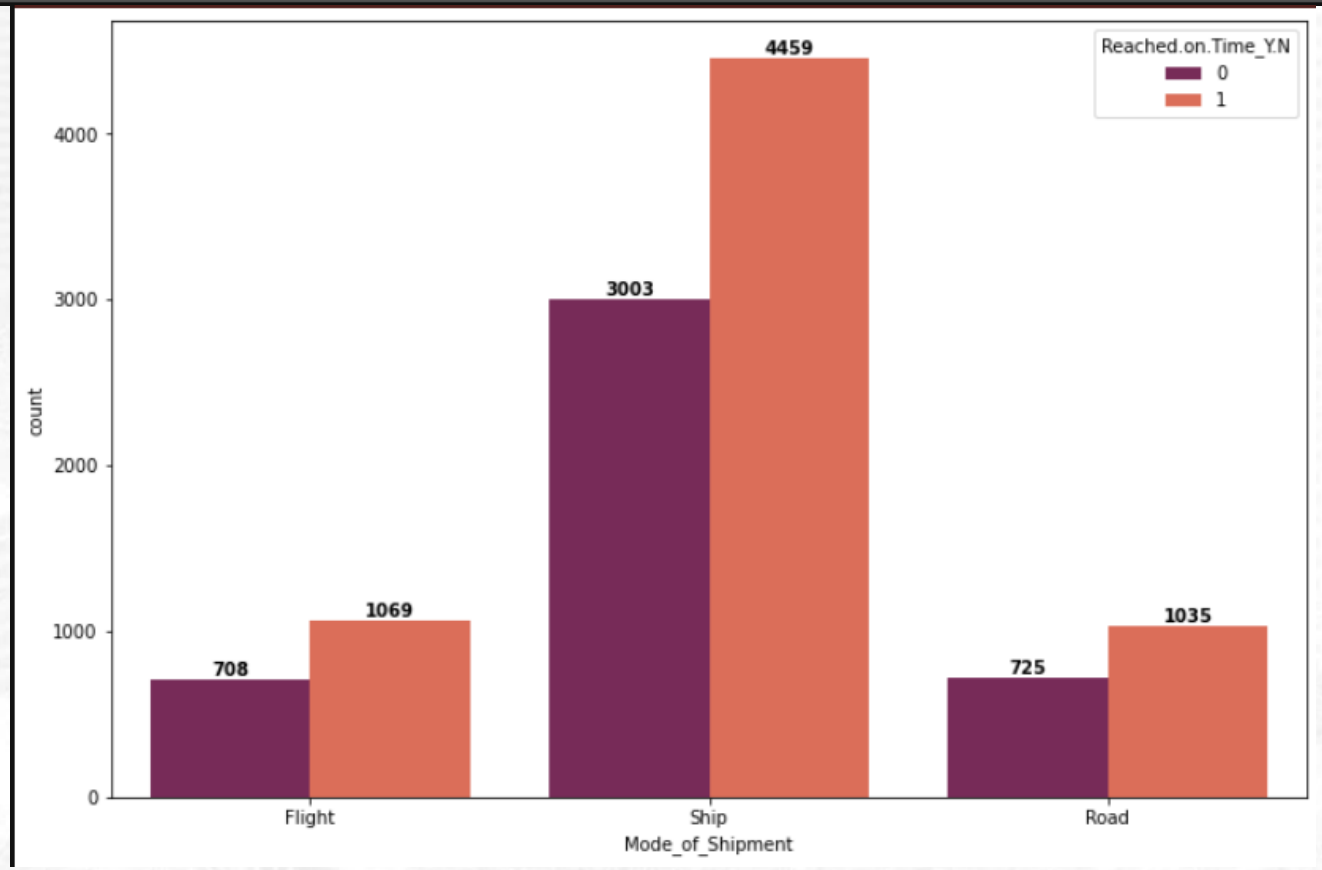


Untuk paket yang memiliki importance high banyak yang mengalami keterlambatan dalam pengiriman

Business Insight

Setelah melakukan EDA, kami menemukan beberapa business insight sebagai berikut.

```
fig, ax = plt.subplots(figsize=(12,8))
sns.countplot('Mode_of_Shipment', hue = 'Reached.on.Time_Y.N', data = df, palette='rocket')
for label in ax.containers: #Untuk bikin angka diatas bar
    ax.bar_label(label, fontweight='bold')
plt.show()
```

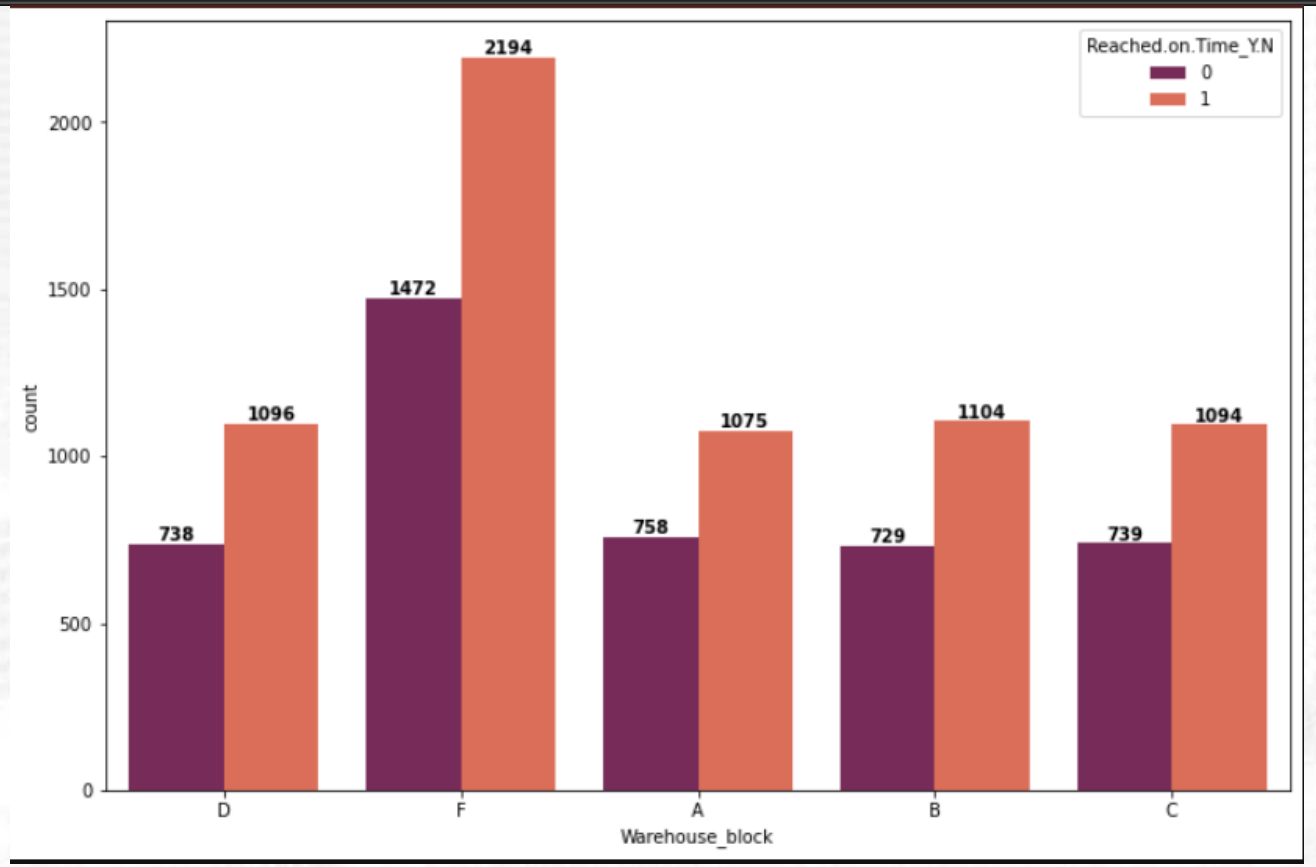


Pengiriman dengan kapal yang paling banyak mengalami keterlambatan dibandingkan dengan metode pengiriman yang lain

Business Insight

Setelah melakukan EDA, kami menemukan beberapa business insight sebagai berikut.

```
fig, ax = plt.subplots(figsize=(12,8))
sns.countplot('Warehouse_block', hue = 'Reached.on.Time_Y.N', data = df, palette='rocket')
for label in ax.containers: #Untuk bikin angka diatas bar
    ax.bar_label(label, fontweight='bold')
plt.show()
```



Setiap warehouse sering mengalami keterlambatan dalam pengiriman, pada warehouse block f sering mengalami keterlambatan hal ini mungkin terjadi karena banyak metode pengiriman yang menggunakan kapal