Rakamin Academy

W3.Solutions()

Dokumen Laporan Final Project

Latar Belakang Masalah

Problem :

Sebuah perusahaan *e-commerce* berbasis internasional ingin menemukan *insight* dari data pelanggan.. Berdasarkan data dari perusahaan tersebut, terdapat kurang lebih 60% yang mengalami keterlambatan dalam penerimaan barang. Berdasarkan studi dari voxware yang melakukan survey terhadap 600 orang, sebanyak 62% pelanggan akan cenderung berkurang atau berhenti berbelanja dari retailer online jika barang yang mereka beli terlambat 2-3 hari dari tanggal yang dijanjikan. Maka dari itu pihak *e-commerce* ingin menjaga dan meningkatkan customer retention dan meningkatkan performa logistik dikarenakan banyak dari pelanggan yang melakukan komplain mengenai ketepatan waktu pengiriman. (https://www.supplychainbrain.com/articles/14912-impact-of-late-or-inaccurate-deliveries-can-be-disastrous-study-shows)

Role :

Sebagai konsultan data scientist untuk perusahaan *e-commerce*, kami diminta memprediksi apakah penerimaan tersebut tepat waktu atau tidak berdasarkan data yang tersedia dan kami diminta untuk menganalisis faktor-faktor yang mempengaruhi ketepatan waktu penerimaan serta memberikan insight dan rekomendasi berdasarkan hasil analisis.

Goal :
Menurunkan persentase keterlambatan

Objective :
Membuat model *machine learning* untuk memprediksi ketepatan waktu pengiriman barang agar mencegah keterlambatan agar persentase keterlambatan menurun. Perusahaan diharapkan dapat menggunakan model tersebut untuk menentukan keputusan bisnis sehingga customer retention dan tingkat kepuasan pelanggan tetap atau meningkat.

Business Metrics :
- Persentase keterlambatan

Data Exploration

```
df.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
```

Pengamatan:

1. Data terdiri dari 10999 baris
2. Tidak terdapat data yang null atau missing value
3. Sepertinya tidak ada issue yang mencolok pada tipe data untuk setiap kolom (sudah sesuai)

Dan dari beberapa kali sample

Sepertinya tidak ada anomali pada setiap entri kolom sudah sesuai

```
[4] df.sample(5)
```

	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
1440	1441	D	Ship	3	2	273	3	low	M	41	3464	1
6032	6033	A	Flight	6	3	283	5	medium	M	10	1483	0
3838	3839	C	Flight	4	1	190	4	medium	F	5	5570	1
9812	9813	A	Ship	4	3	253	3	low	F	5	5703	0
8148	8149	D	Ship	3	2	176	2	medium	F	1	5292	0

Dan dari beberapa kali sample :

Sepertinya tidak ada anomali pada setiap entri kolom sudah sesuai

Exploratory Data Analysis

```
[6]  df[nums].describe()
```

	Customer_care_calls	Customer_rating	Prior_purchases	Discount_offered	Cost_of_the_Product	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	3.567597	13.373216	210.196836	3634.016729	0.596691
std	1.141490	1.413603	1.522860	16.205527	48.063272	1635.377251	0.490584
min	2.000000	1.000000	2.000000	1.000000	96.000000	1001.000000	0.000000
25%	3.000000	2.000000	3.000000	4.000000	169.000000	1839.500000	0.000000
50%	4.000000	3.000000	3.000000	7.000000	214.000000	4149.000000	1.000000
75%	5.000000	4.000000	4.000000	10.000000	251.000000	5050.000000	1.000000
max	7.000000	5.000000	10.000000	65.000000	310.000000	7846.000000	1.000000

Beberapa pengamatan:

1. Kolom Customer_care_calls, customer_rating, dan Cost_of_the_Product tampak sudah cukup simetrik distribusinya (mean dan median tak berbeda jauh)
2. Kolom Discount_offered dan Prior_purchases tampaknya skew ke kanan (long-right tail)
3. Kolom Reached.on.Time_Y.Nbernilai boolean/binary

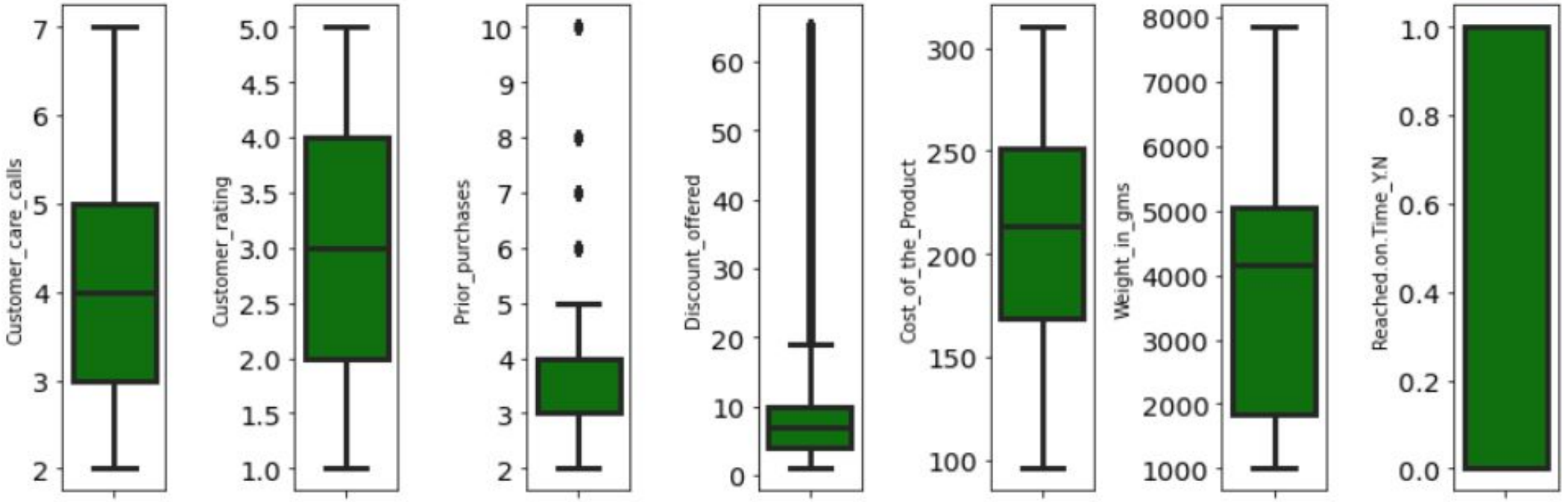
```
df[cats].describe()
```

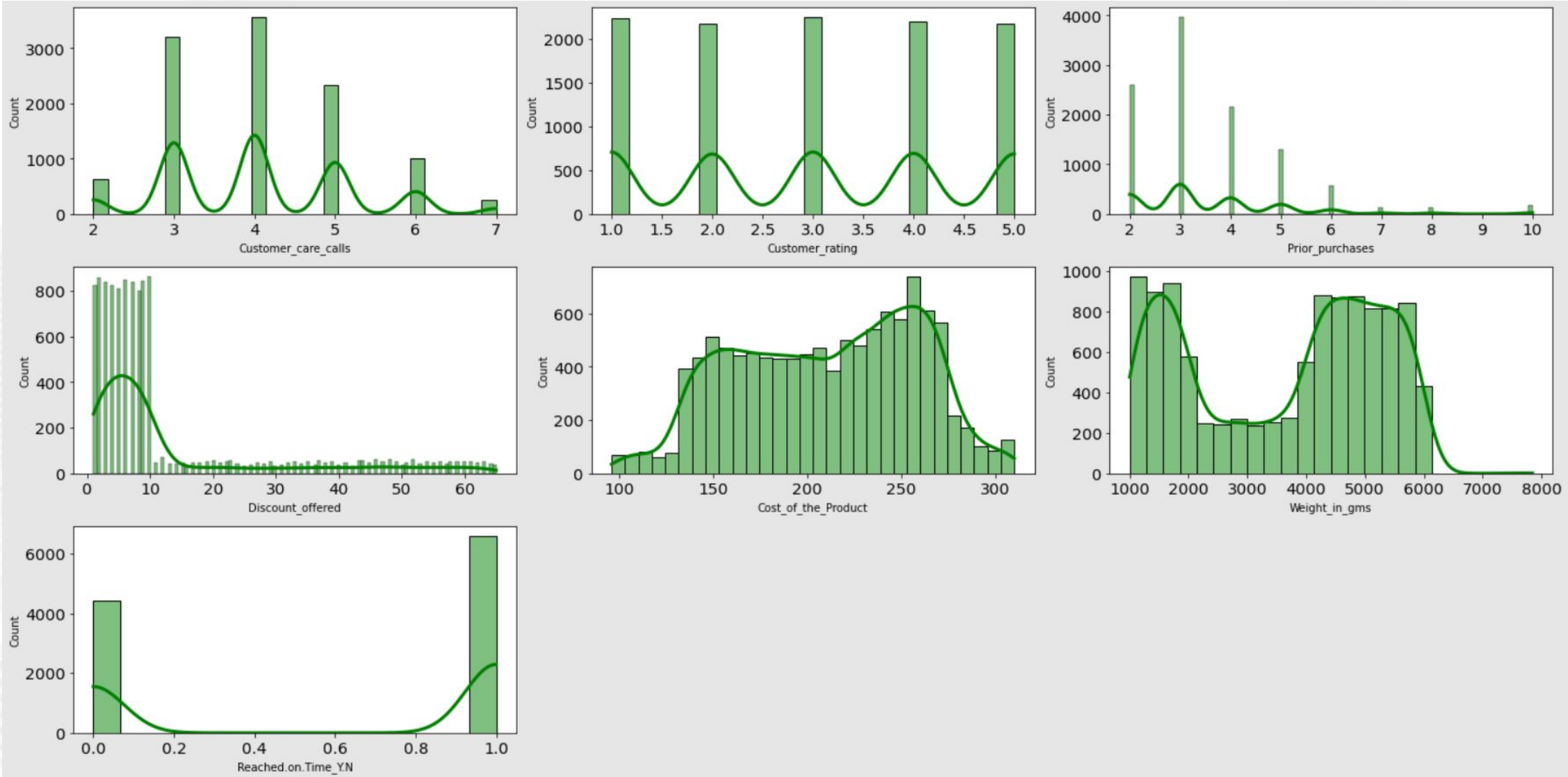
	Mode_of_Shipment	Product_importance	Gender	Warehouse_block
count	10999	10999	10999	10999
unique	3	3	2	5
top	Ship	low	F	F
freq	7462	5297	5545	3666

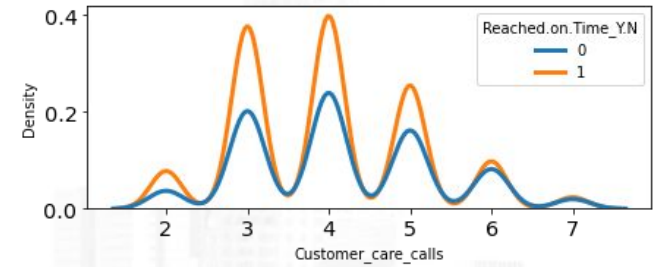
Beberapa pengamatan:

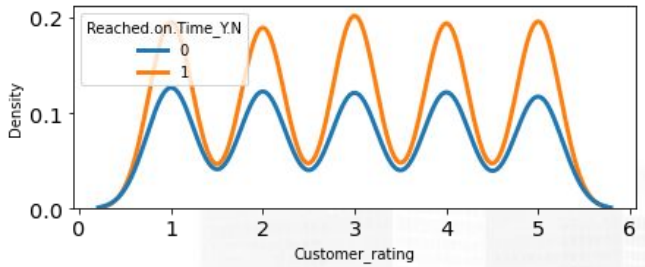
- Untuk kategori gender **perempuan** lebih dominan,
- untuk kategori product importance di dominasi oleh **kategori low**
- untuk kategori mode pengiriman di dominasi oleh **pengiriman menggunakan kapal (ship)**
- untuk warehouse_block didominasi oleh **block F**
- Semua unique value tiap kategori masih dalam kategori normal sekitar **2-5 unique values**

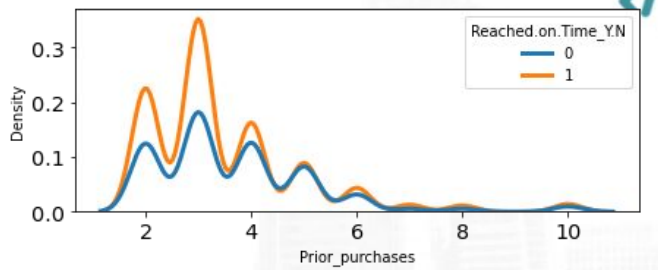
Exploratory Data Analysis - Univariate Analysis

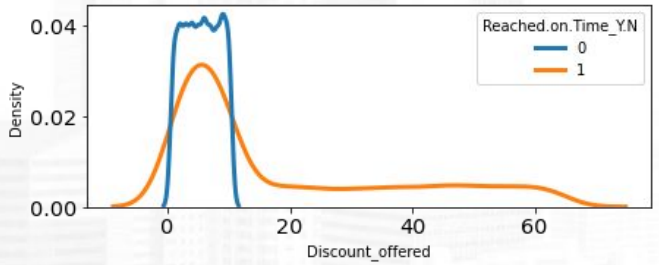


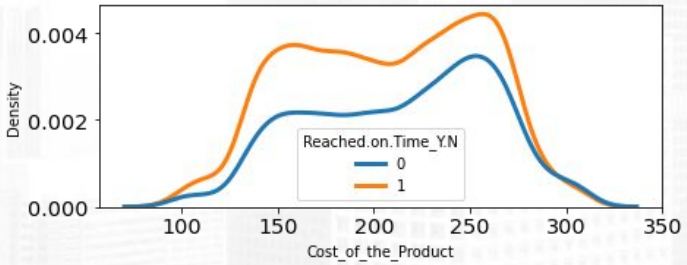

Count
Customer_care_calls
Customer_rating
Prior_purchases
Discount_offered
Cost_of_the_Product
Weight_in_gms
Reached.on.Time_Y.N

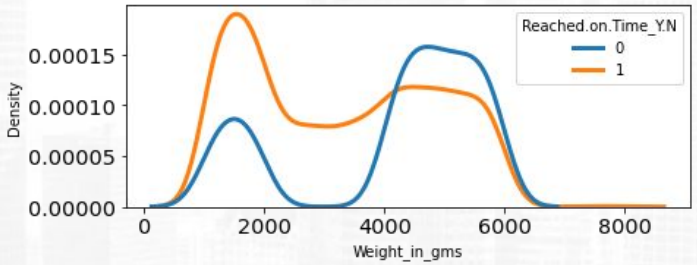


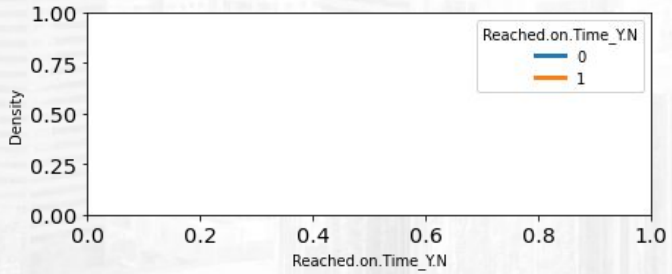






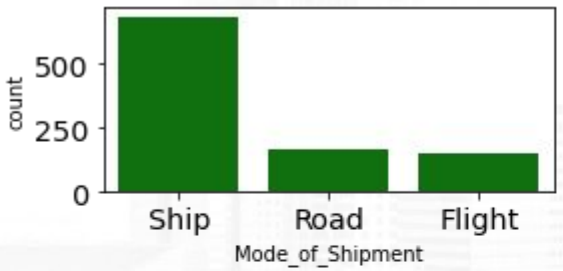


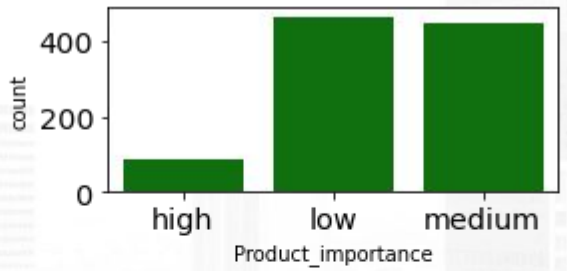


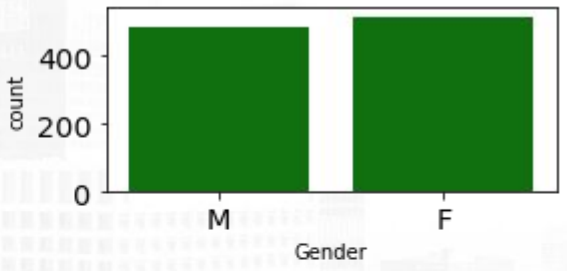


Dari distribution plot terlihat bahwa:

- Kolom cost_of_the_product tampak sudah mendekati distribusi normal
- Seperti dugaan kita ketika melihat boxplot di atas, kolom Prior_purchases, dan Discount_offered sedikit skewed Berarti ada kemungkinan kita perlu melakukan sesuatu pada kolom2 tersebut nantinya
- Kolom-kolom Reached.on.Time sejatinya adalah biner, sehingga tidak perlu terlalu diperhatikan bentuk distribusinya
- Untuk kolom weigh_in_gms terdapat ketidakpastian distribusi karena berbentuk u-shape.
- untuk kolom customer_care_calls dan customer_rating distibusi merata







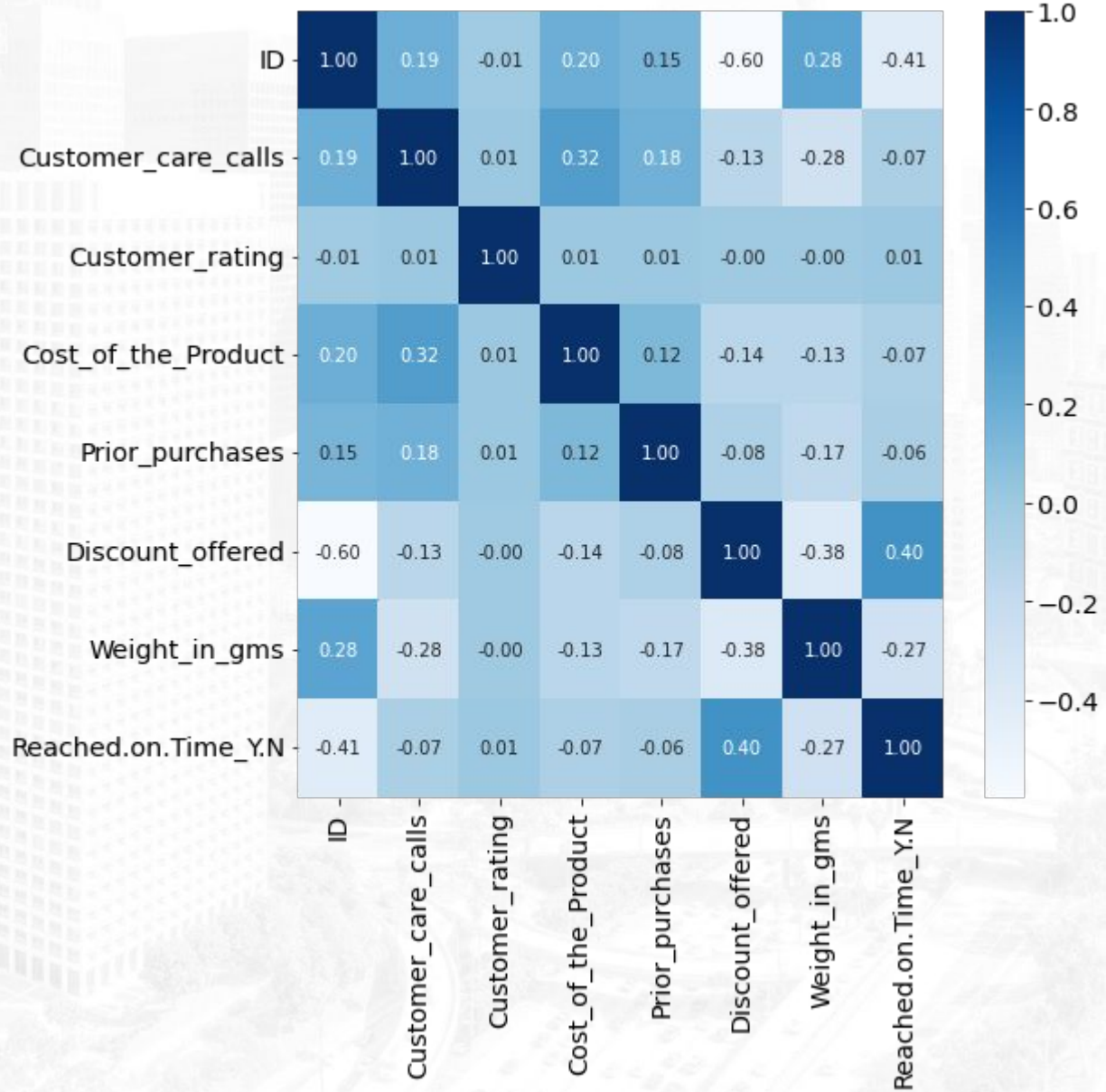
Seperti pengamatan kita sebelumnya, distribusi kategori low (Product_importance), gudang penyimpanan(warehouse_block) dan kategori Ship (Mode_of_Shipment) didominasi 1-2 value.

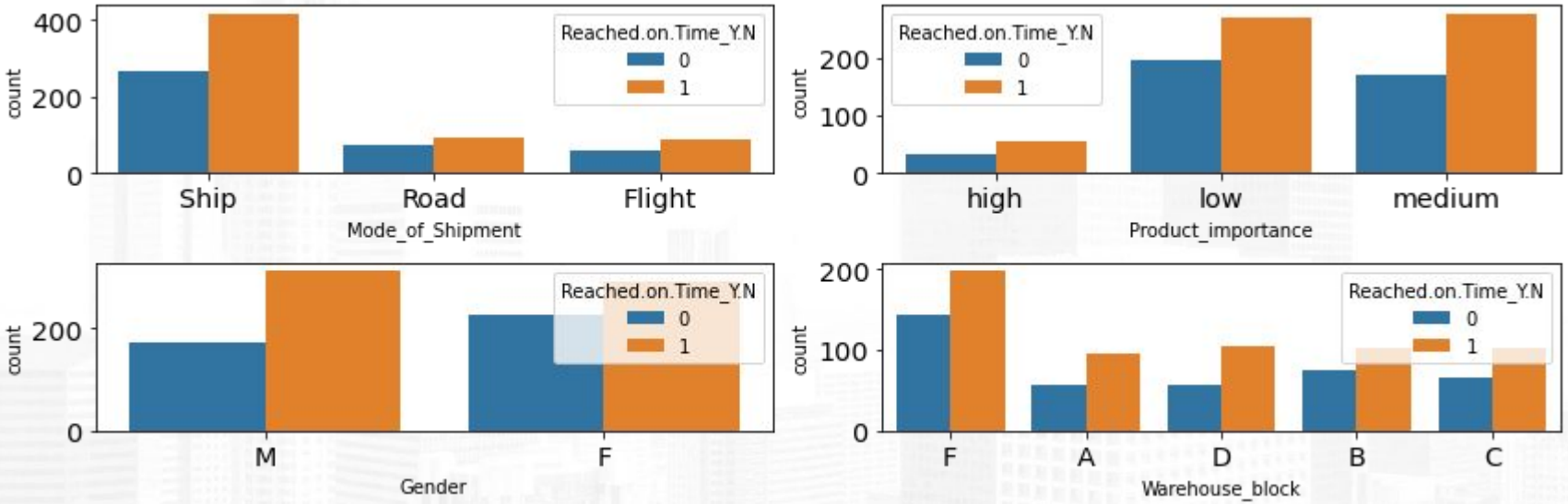
Exploratory Data Analysis - Bivariate Analysis

Rakamin Academy

Dari correlation heatmap di atas dapat dilihat bahwa:

- Target kita Reached.on.Time_Y.N memiliki korelasi positif lemah dengan customer_rating, cost_of_the_product, customer_care_calls dan prior_purchases
- Ia juga memiliki korelasi positif cukup kuat dengan Discount_offered
- Ia juga memiliki korelasi negatif cukup kuat dengan weight_in_gms





Pengamatan

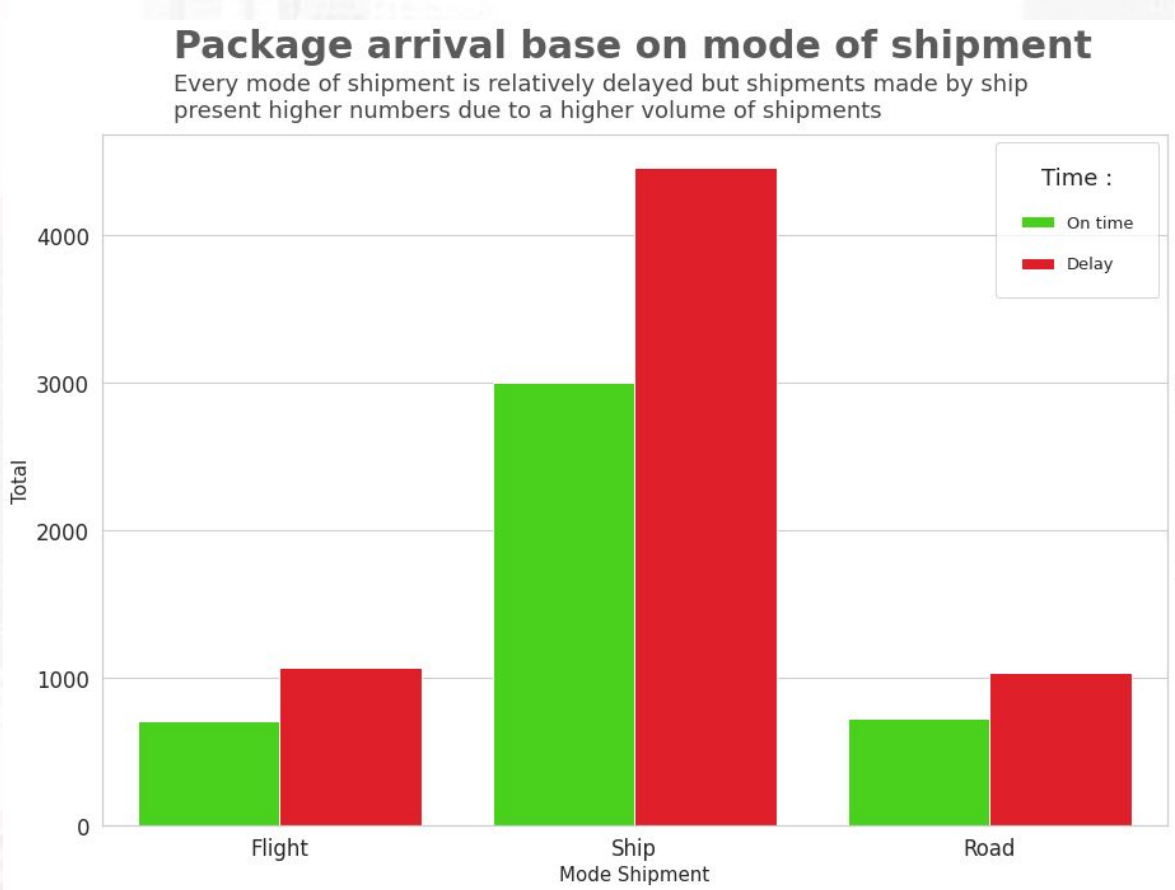
- shipment dengan ship cenderung akan mengalami telat pengiriman
- untuk produk_importance dengan kategori low dan medium cenderung akan mengalami telat pengiriman
- untuk warehouse_block dengan kategori F cenderung mengalami telat pengiriman

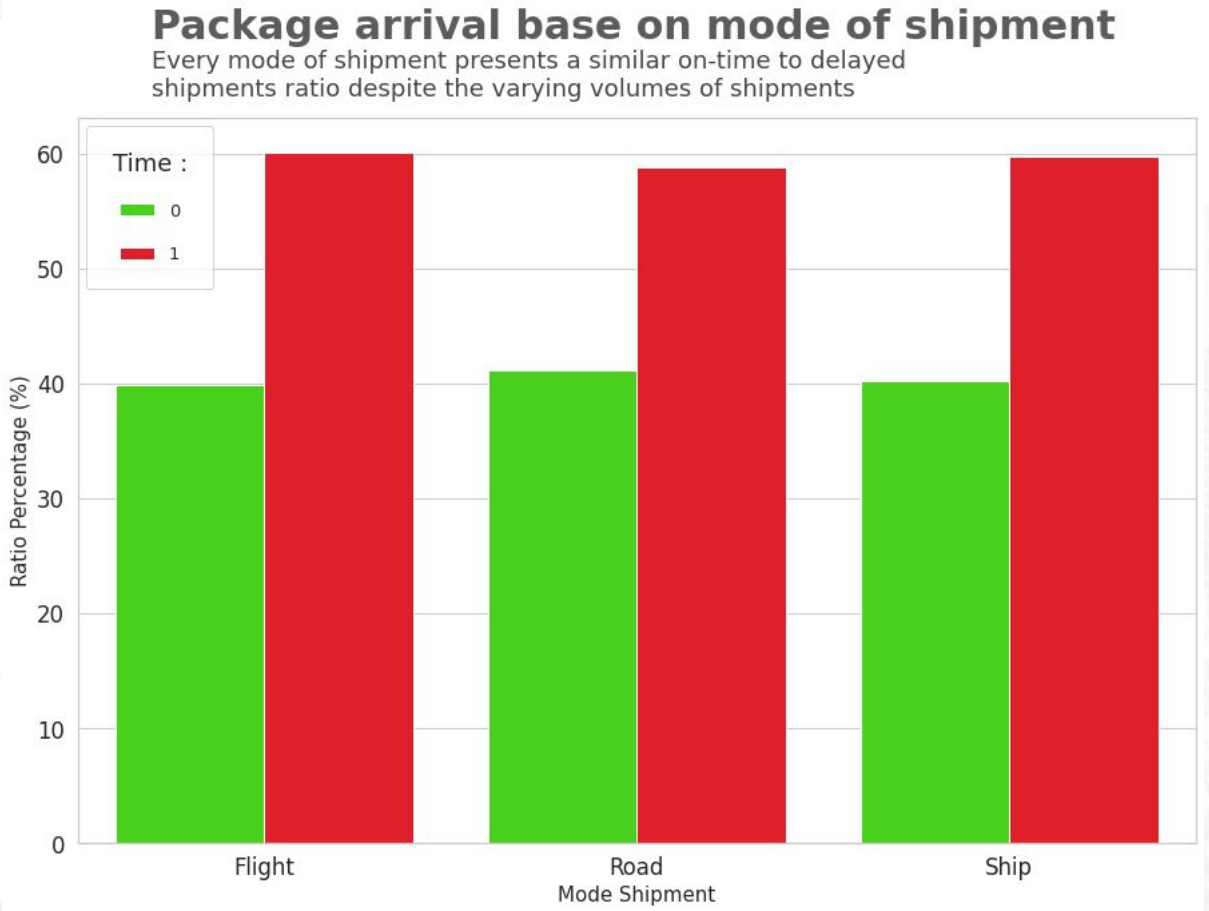
EDA Conclusion

Beberapa hal yang kita temukan dari EDA dataset ini adalah:

- Data terlihat valid dan tidak ada kecacatan yang major/signifikan
- Ada beberapa distribusi yang sedikit skewed, hal ini harus diingat apabila kita ingin melakukan sesuatu atau menggunakan model yang memerlukan asumsi distribusi normal
- Beberapa feature memiliki korelasi yang jelas dengan target, mereka akan dipakai
- Beberapa feature terlihat sama sekali tidak berkorelasi, mereka sebaiknya diabaikan
- Dari fitur kategorikal, “mode_of_shipment” ,” warehouse_block”, dan “product_importance” sepertinya berguna untuk menjadi prediktor model

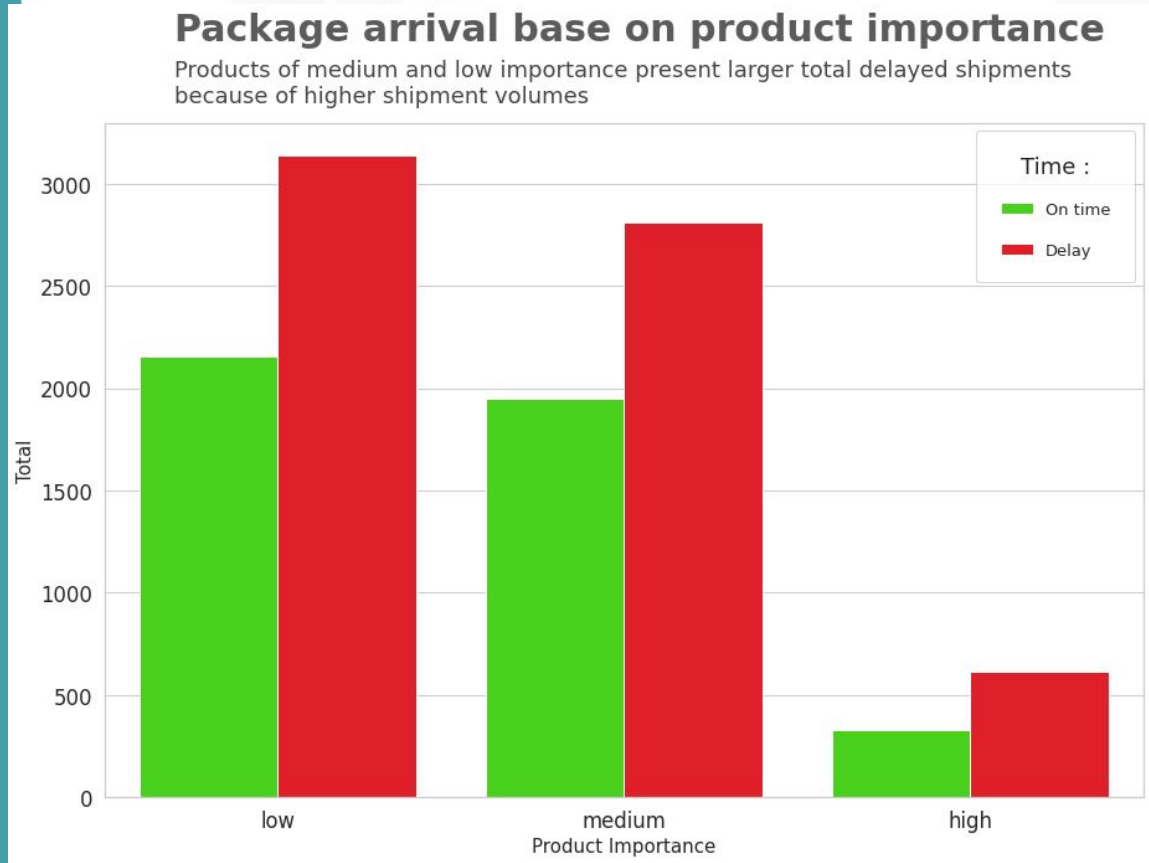
Insight and Visualization

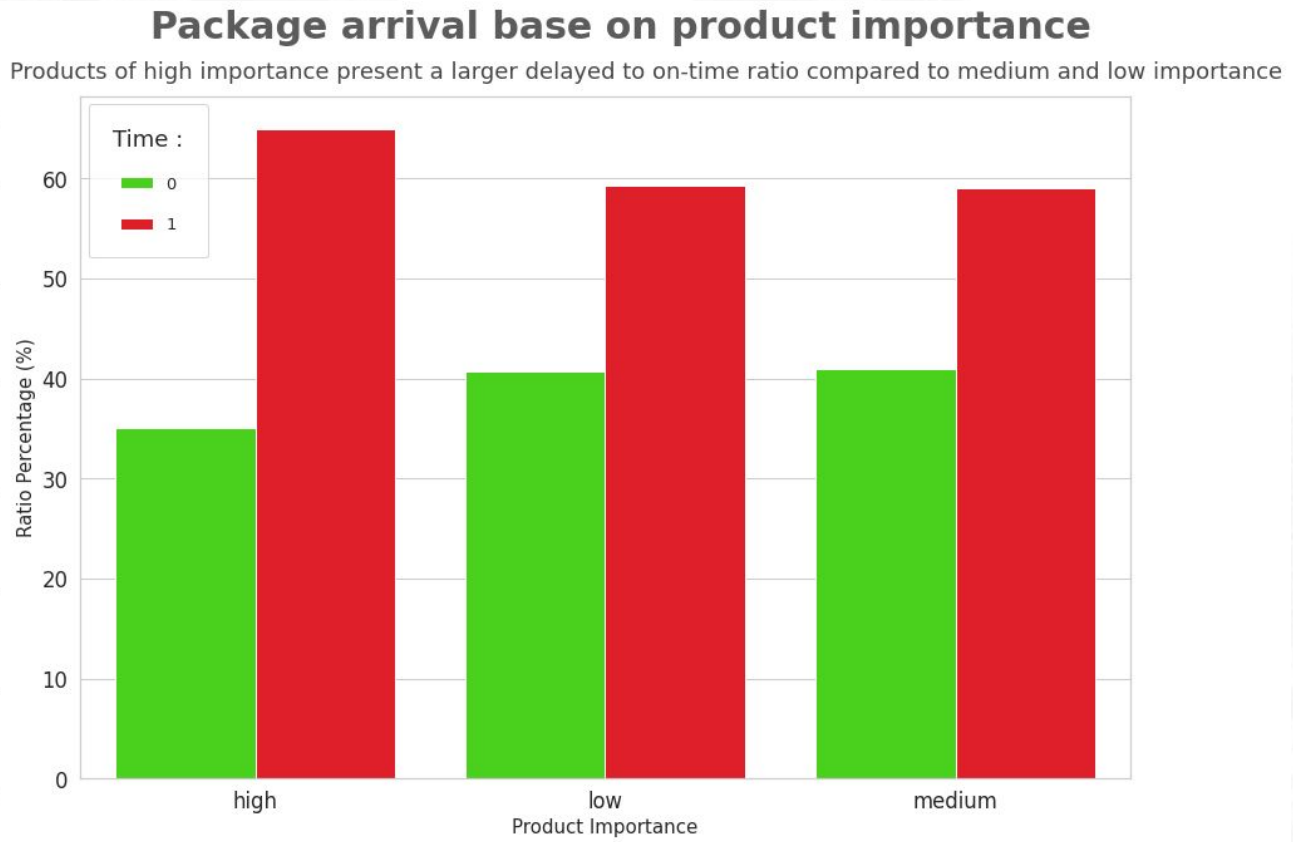




Rekomendasi bisnis:

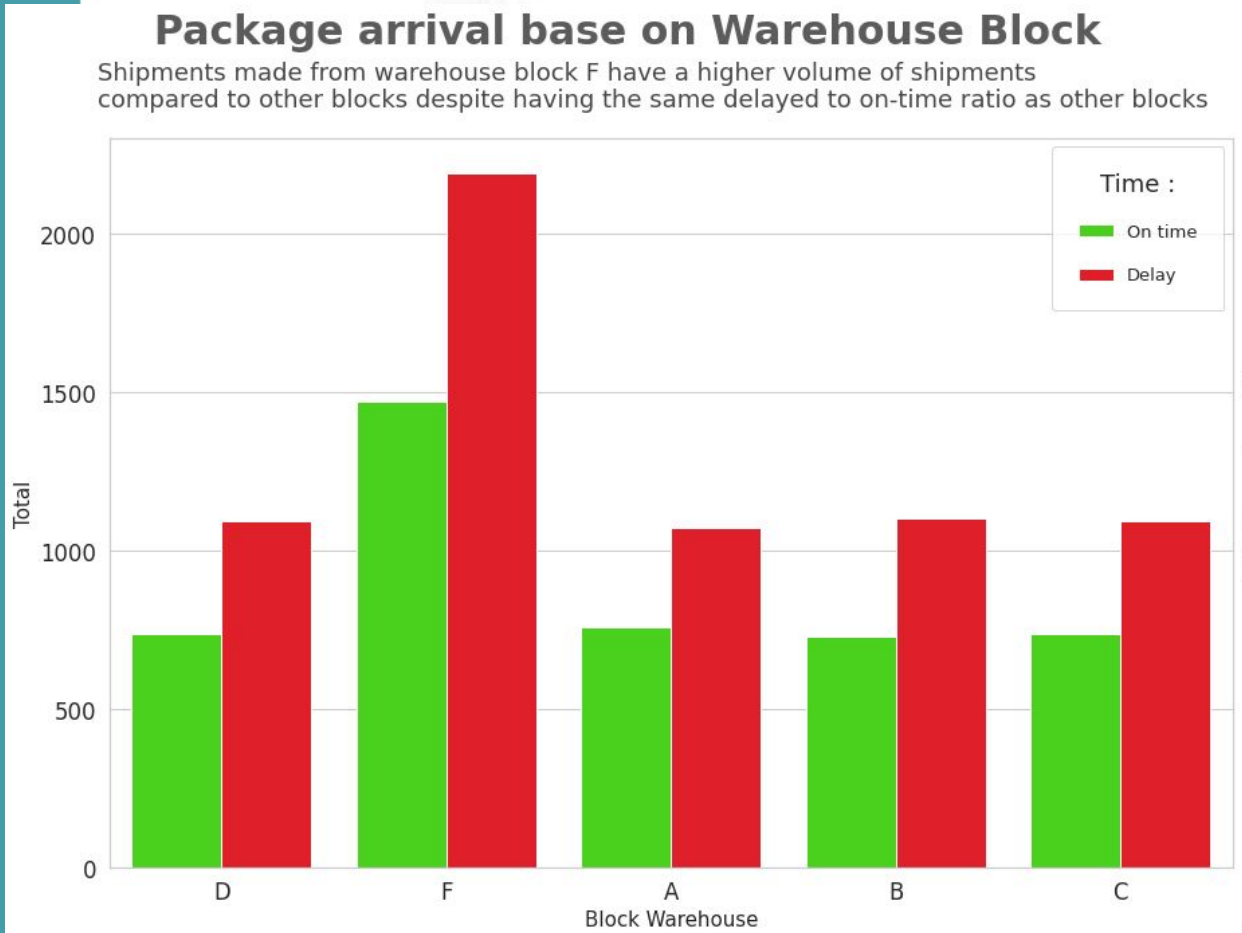
Client dapat meningkatkan performa logistik dan evaluasi internal untuk semua metode pengiriman karena dapat dilihat underperformance dari ketiga-tiganya

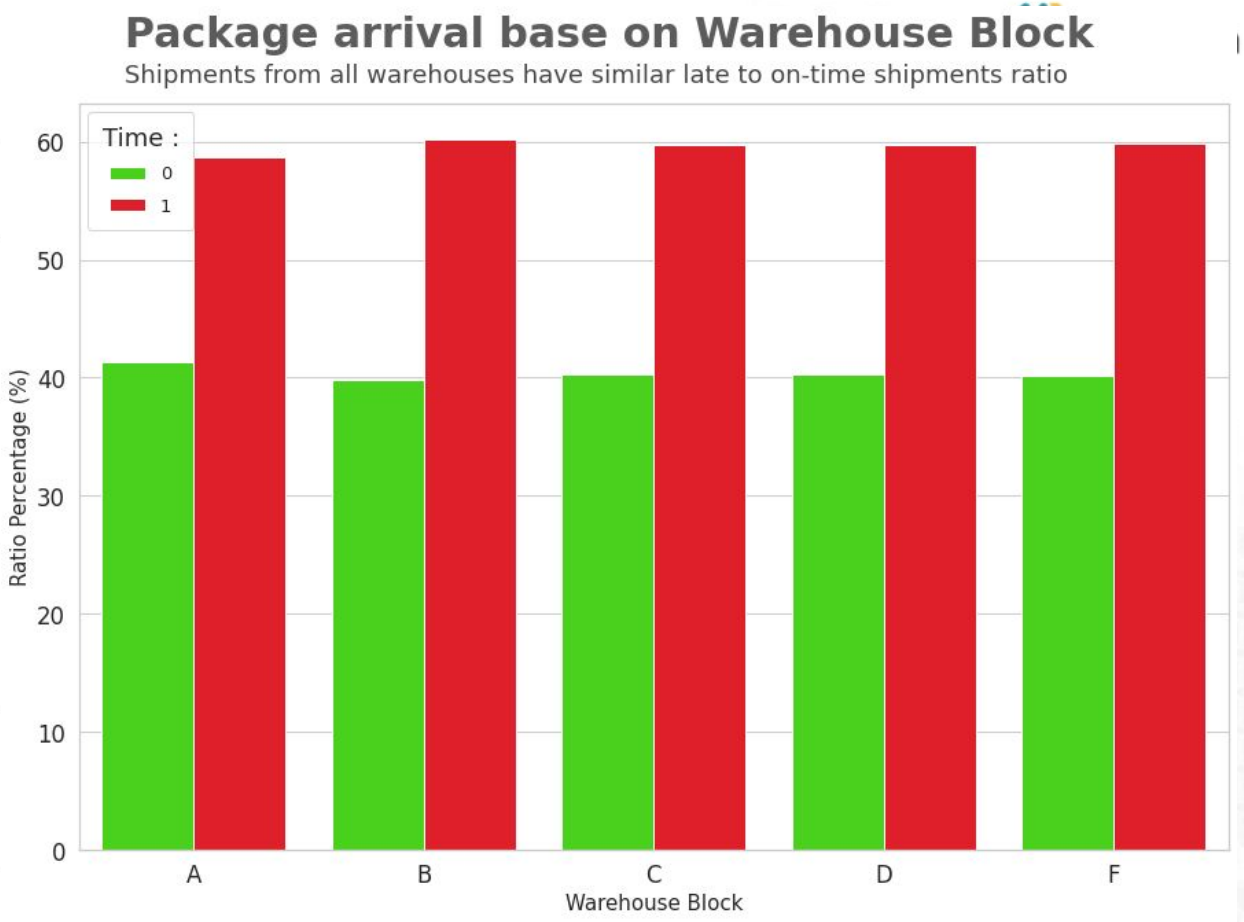




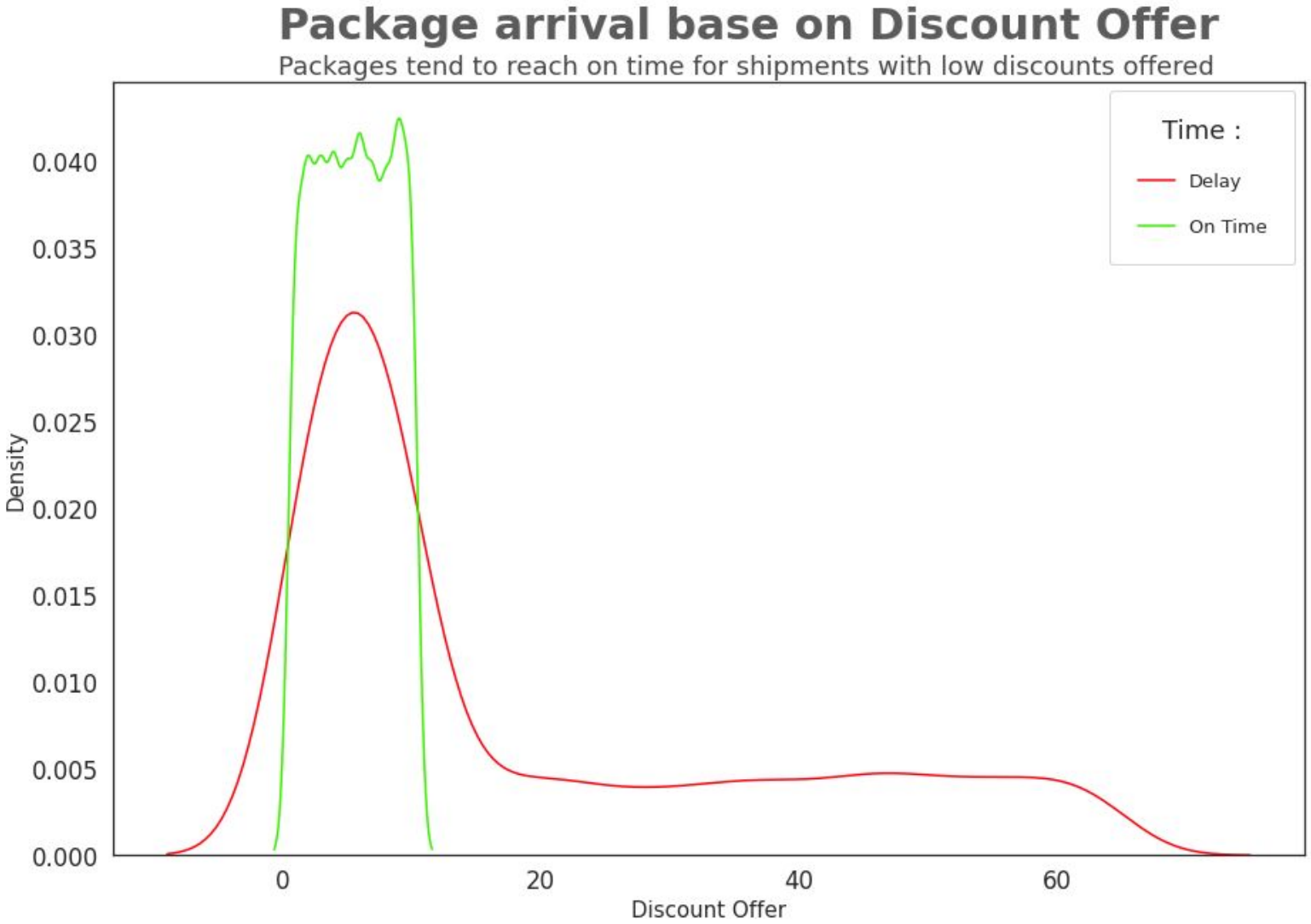
Rekomendasi bisnis:

Client dapat meningkatkan performa logistik untuk semua *product importance* terutama ‘high’ karena dapat dilihat underperformance dari ketiga-tiganya





Rekomendasi bisnis:
Client dapat meningkatkan performa logistik dan evaluasi internal untuk semua warehouse block karena dapat dilihat underperformance dari ketiga-tiganya



Rekomendasi bisnis:

Client dapat meningkatkan SDM dalam pengiriman saat customer diberikan banyak discount seperti ‘Black Friday’, Harbolnas, dan event-event khusus lainnya