Rakamin Academy

Team Cobra (Kelompok 4)

Dokumen Laporan Final Project

Rakamin Academy

STAGE 0
PREPARATION

Latar Belakang Masalah

Sebuah perusahaan e-commerce skala internasional seringkali mengalami masalah **keterlambatan** dalam proses pengiriman barang. Hal ini dapat berdampak kepada tingkat kepuasan pelanggan (*customer satisfaction*) dan memungkinkan terjadinya penurunan total penjualan. Perusahaan membutuhkan sebuah **model *machine learning*** yang dapat memprediksi apakah suatu pengiriman akan terlambat atau tidak sebagai upaya untuk meningkatkan rasio pengiriman tepat waktu.

Peran Tim Kami

Tim kami berperan sebagai ***Data Scientists*** dalam perusahaan e-Commerce Internasional yang menjual produk elektronik, yang ditugaskan untuk melakukan analisis data dan membangun model machine learning yang dapat membantu perusahaan e-commerce dalam memprediksi keterlambatan pengiriman dan meningkatkan customer satisfaction untuk meningkatkan penjualan

Tujuan

- Mengurangi rasio keterlambatan pengiriman barang dari sekitar 40% menjadi 20% pada akhir tahun

- Meningkatkan total sales sebesar 10-20% dibanding tahun sebelumnya dengan meningkatnya jumlah order customer, seiring dengan meningkatnya kepuasan customer terhadap pengiriman barang

Sasaran

- Mencari fitur-fitur yang memiliki korelasi/hubungan dengan keterlambatan pengiriman barang
- Membuat model *machine learning* untuk memprediksi apakah suatu pengiriman akan terlambat atau tidak
- Memberikan rekomendasi tindakan terhadap pengiriman yang diduga memiliki potensi keterlambatan yang tinggi

Business Metrics

Main metric: Delivery on time ratio

Metric ini menghitung jumlah pengiriman barang yang tiba tepat waktu dibagi dengan total pengiriman barang dalam periode waktu tertentu, sehingga dapat memberikan gambaran tentang seberapa efektif perusahaan dalam memenuhi kewajiban pengiriman barang tepat waktu.

Supporting metric: Customer satisfaction (rating/call/complaints)

Metric ini dapat dihitung dengan memperhatikan beberapa faktor, seperti rating yang diberikan oleh pelanggan, jumlah panggilan layanan pelanggan, dan jumlah keluhan yang diterima oleh perusahaan. Dengan metric ini perusahaan dapat mengetahui tingkat kepuasan pelanggan dan mengambil tindakan yang sesuai untuk meningkatkan pengalaman pelanggan.

STAGE 1
Exploratory Data Analysis

Descriptive Statistics

Column	Description	Non-Null Count	Data Type	Categories of Data
ID	ID Number of Customers.	10999 non-null	int64	Discrete Data
Warehouse_block	The Company have big Warehouse which is divided in to block such as A,B,C,D,E.	10999 non-null	object	Nominal Data
Mode_of_Shipment	The Company Ships the products in multiple way such as Ship, Flight and Road.	10999 non-null	object	Nominal Data
Customer_care_calls	The number of calls made from enquiry for enquiry of the shipment.	10999 non-null	int64	Discrete Data
Customer_rating	The company has rated from every customer. 1 is the lowest (Worst), 5 is the highest (Best)	10999 non-null	int64	Ordinal Data
Cost_of_the_Product	Cost of the Product in US Dollars.	10999 non-null	int64	Continuous Data
Prior_purchases	The Number of Prior Purchase.	10999 non-null	int64	Discrete Data
Product_importance	The company has categorized the product in the various parameter such as low, medium, high.	10999 non-null	object	Nominal Data
Gender	Male and Female.	10999 non-null	object	Nominal Data
Discount_offered	Discount offered on that specific product.	10999 non-null	int64	Discrete Data
Weight_in_gms	Weight of product in grams	10999 non-null	int64	Continuous Data
Reached.on.Time_Y.N	Boolean of whether shipment reached in time or not (0 = No, 1 = Yes)	10999 non-null	int64	Nominal Data

Tidak ada null values dan **tidak ada duplicate ID** (setiap baris adalah unique customer)

Descriptive Statistics - Numerical Data

	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

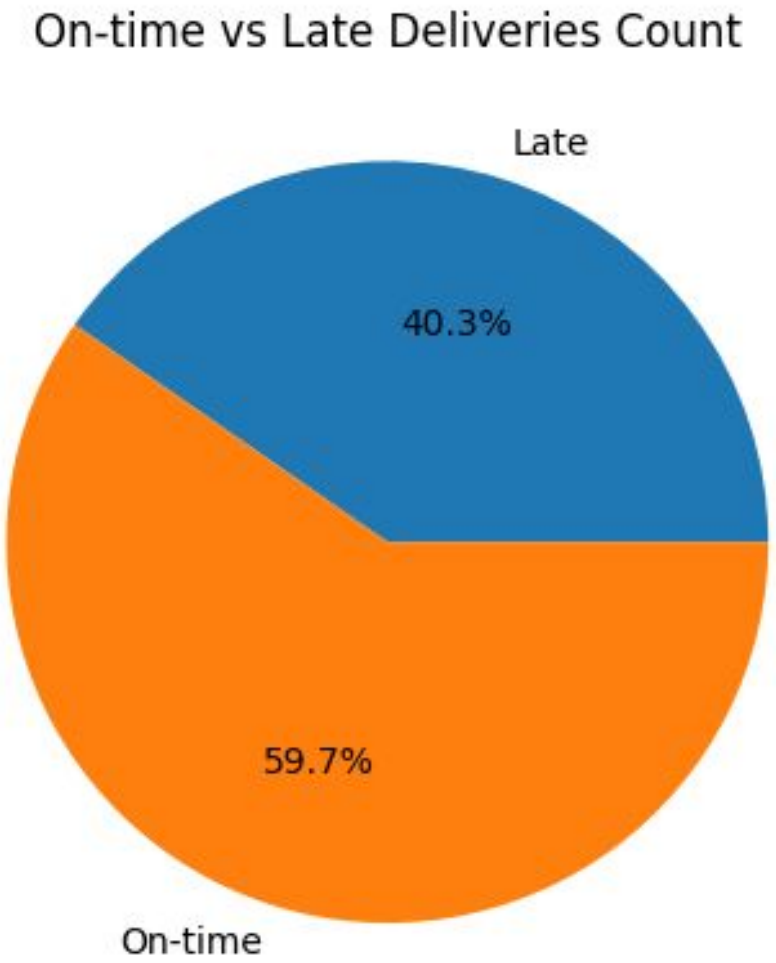
Terdapat sebuah kejanggalan pada feature ***Discount_offered***, dimana nilai maksimalnya sebesar **65** sedangkan selisih jaraknya dengan **Q3 (10)** atau **mean (13)** sangat tinggi, sehingga diduga terdapat beberapa ***outliers***

Descriptive Statistics - Categorical Data

	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

Tidak terdapat kejanggalan pada *Categorical Data* karena setiap kolom memiliki jumlah data yang sama dengan jumlah total baris data dan tidak terdapat nilai yang tidak diharapkan seperti huruf atau karakter yang dianggap sebagai kejanggalan.

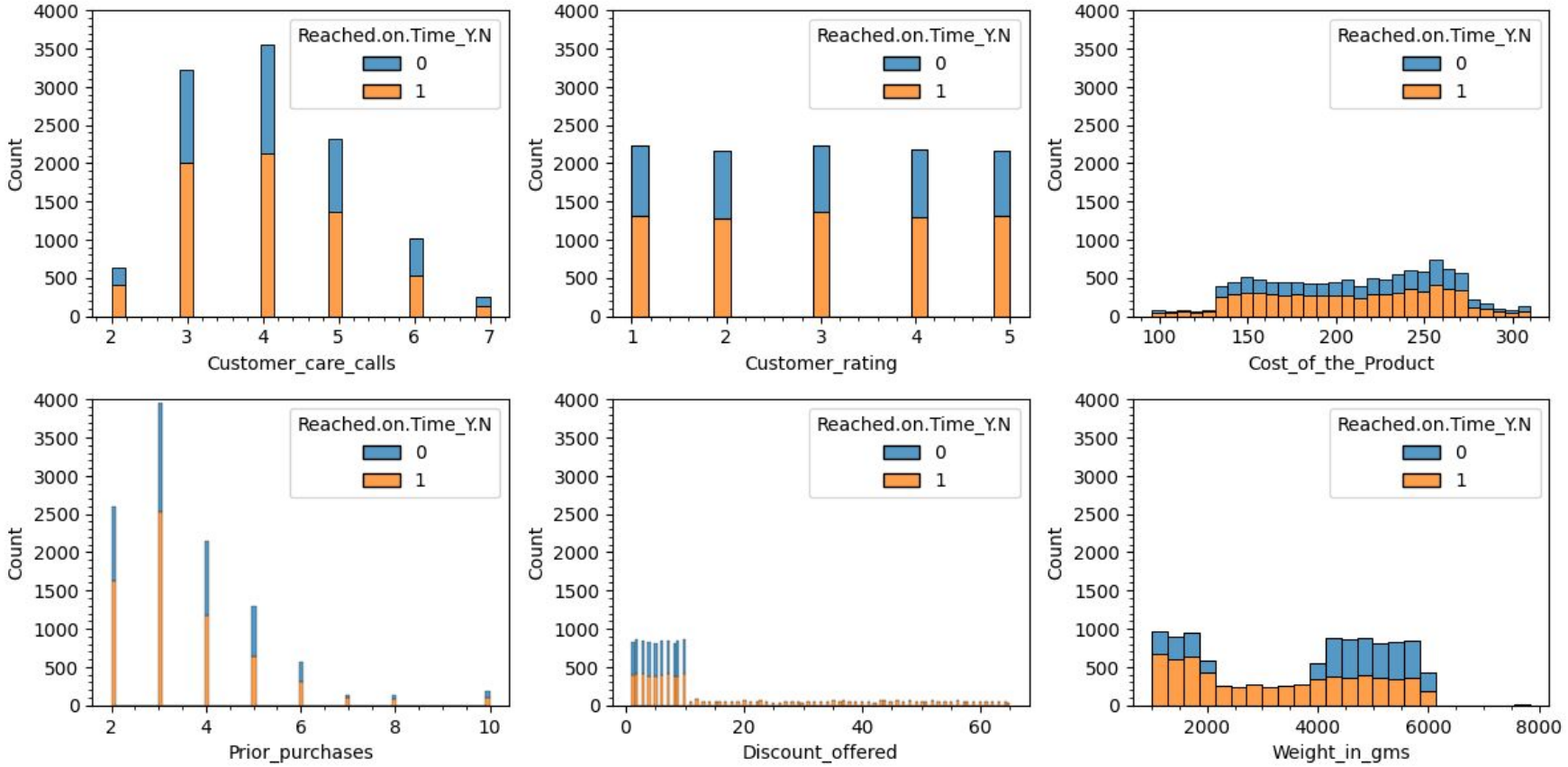
Univariate Analysis



Diketahui bahwa proporsi pengiriman yang tepat waktu (*On-time*) sebesar 59,7% dan proporsi pengiriman yang terlambat (*Late*) sebesar 40,3%. Dapat disimpulkan bahwa sebagian besar pengiriman dilakukan tepat waktu, namun masih ada sebagian kecil pengiriman yang terlambat. Hal ini dapat menjadi fokus perbaikan untuk meningkatkan kualitas layanan dan kepuasan pelanggan.

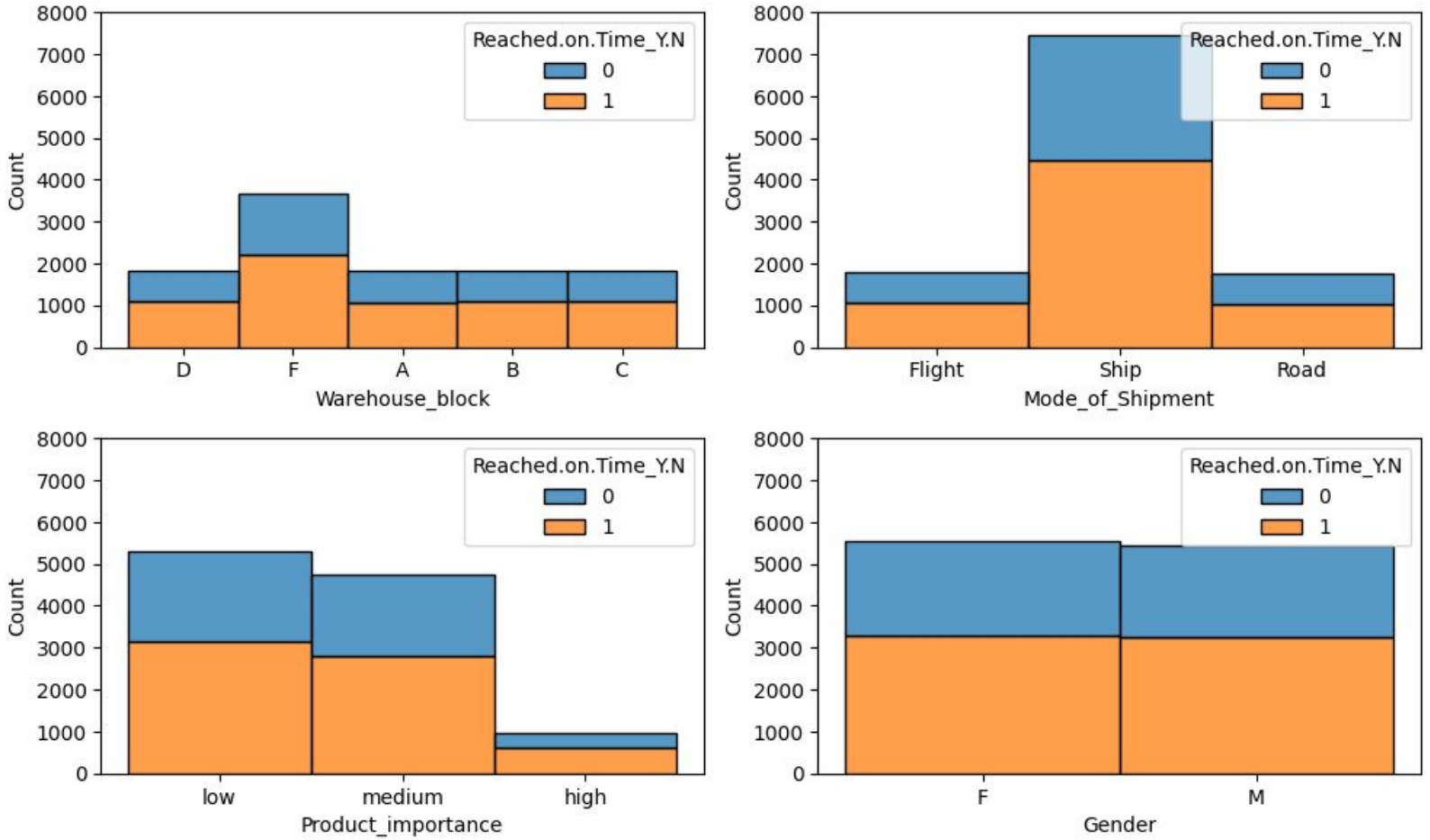
Karena data sudah memiliki jumlah sampel yang cukup banyak dan tidak terlalu berbeda antara kelas satu dengan yang lain (tidak terlalu signifikan *class imbalance*-nya), maka tidak perlu dilakukan penyeimbangan lebih lanjut terhadap pada tahap *pre-processing* data.

Univariate Analysis



- Pada feature *Prior_purchase* membentuk positive skew
- Pada feature *Weight_in_gms* juga terdapat beberapa outlier pada nilai diatas 7500 (tidak terlalu terlihat pada grafik)
- Pada feature *Discount_offered*, terdapat nilai yang mendominasi yaitu pada nilai 0 - 10
- Saat data pre-processing, perlu dilakukan scaling pada data numerik agar mempunyai range yang seragam.

Univariate Analysis



- Berdasarkan warehouse, barang paling banyak disimpan/dilayani oleh warehouse F, sedangkan warehouse lainnya kurang lebih menampung jumlah barang yang sama
- Mayoritas pengiriman dilakukan melalui jalur laut (Ship)
- Jumlah barang yang tingkat kepentingannya tinggi (high) relatif sedikit
- Jumlah customer pria hampir setara dengan jumlah customer wanita, dengan rasio keterlambatan yang juga serupa

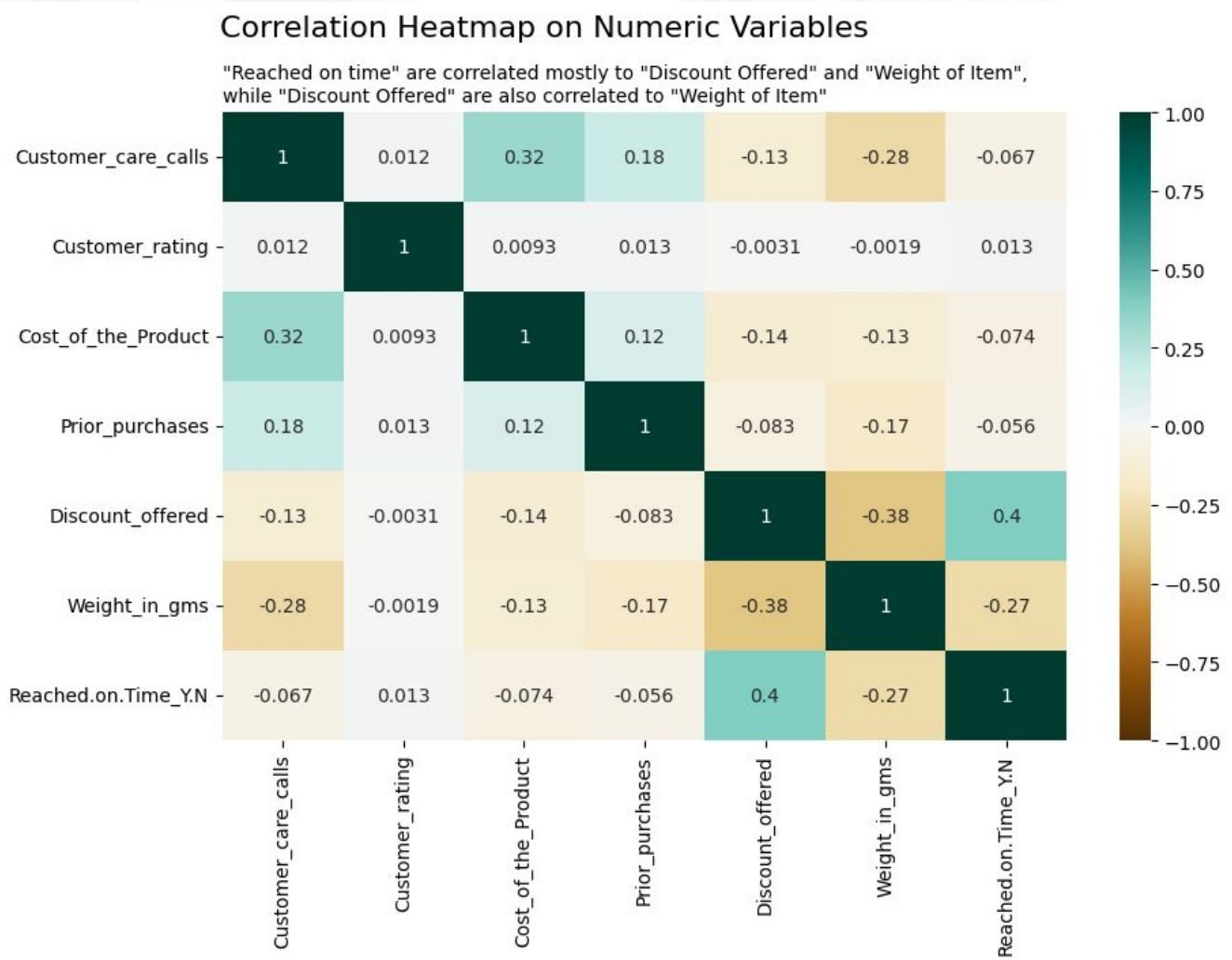
Univariate Analysis: Next Steps

- Beberapa hal yang harus ditindaklanjuti pada saat data *pre-processing*:
  - Untuk kolom Gender, dapat dilakukan mapping agar nilai "M" dan "F" menjadi 0 dan 1.
  - Untuk kolom Product_importance, dapat dilakukan label encoding karena kedua kolom tersebut memiliki nilai-nilai yang berurutan dan memiliki hubungan ordinal antara nilai-nilainya.
  - Untuk kolom Warehouse_block dan Mode_of_Shipment, dapat dilakukan one-hot encoding karena kolom tersebut tidak memiliki hubungan ordinal antara nilai-nilainya dan setiap nilai kategorikal dianggap sama pentingnya.

Multivariate Analysis - Correlation

Berdasarkan hasil heatmap yang dibuat korelasi antar feature beragam dengan range 1 sampai -1. Semakin mendekati 1 atau -1 maka korelasi semakin kuat, sedangkan semakin mendekati 0 maka korelasi semakin lemah. Beberapa nilai korelasi yang paling relevan adalah sebagai berikut:

- *Discount_offered* dengan *Reached.on.Time_Y.N* berkorelasi sedang positif, sedangkan *Weight_in_gms* dengan *Reached.on.Time_Y.N* berkorelasi lemah negatif
- *Discount_offered* dengan *Weight_in_gms* berkorelasi sedang negatif (diduga dapat mengakibatkan **multikolinearitas**)
- *Cost_of_the_Product* dengan *Customer_care_calls* berkorelasi sedang positif
- *Customer_rating* memiliki korelasi yang sangat kecil terhadap seluruh fitur lainnya, termasuk keterlambatan pengiriman



Multivariate Analysis - Correlation

- Berdasarkan hasil korelasi antara feature, terdapat beberapa hal yang perlu dilakukan, yaitu:
    - *Discount_offered* dan *Weight_in_gms* memiliki korelasi yang cukup signifikan, sehingga perlu dilakukan pengecekan terhadap adanya multikolinearitas antara kedua feature tersebut. Jika ditemukan adanya multikolinearitas, salah satu feature dapat dihapus atau digabungkan dengan feature lain.
    - Korelasi antara *Cost_of_the_Product* dengan *Customer_care_calls* perlu diperhatikan pada saat modelling. Jika terdapat multikolinearitas, feature yang memiliki korelasi lebih rendah dengan target (*Reached.on.Time_Y.N*) dapat dihapus atau digabungkan dengan feature lain.
    - Feature yang memiliki korelasi rendah dengan target (*Reached.on.Time_Y.N*) namun memiliki korelasi yang tinggi dengan feature lain juga perlu diperhatikan pada saat modelling. Pada beberapa kasus, feature tersebut mungkin dapat dihapus atau digabungkan dengan feature lain untuk menghindari multikolinearitas dan meningkatkan akurasi model.

Multivariate Analysis - Pair Plot

Terdapat segmentasi data yang secara visual cukup jelas terlihat pada beberapa pair plot. Secara visual, dapat diambil beberapa insight sebagai berikut:

- Pada *discount offered* terhadap *weight*, pada umumnya barang-barang yang beratnya di atas 4000 gram tidak diberikan diskon lebih besar dari 10% (kecuali untuk beberapa outlier).
- **Tidak ditemukan barang terlambat** pada barang yang diberikan **diskon lebih dari 10%**
- Barang dengan **berat di antara 2000-4000 gram** harganya ada di kisaran ~200 sampai ~300 dollar, dan **tidak ada yang terlambat pengirimannya**
- Terdapat beberapa data outlier jika dilihat berdasarkan berat barang (Weight_in_gms), yaitu barang-barang yang beratnya melebihi 6000 gram, namun untuk barang-barang tersebut **tidak ada satupun yang mengalami keterlambatan**

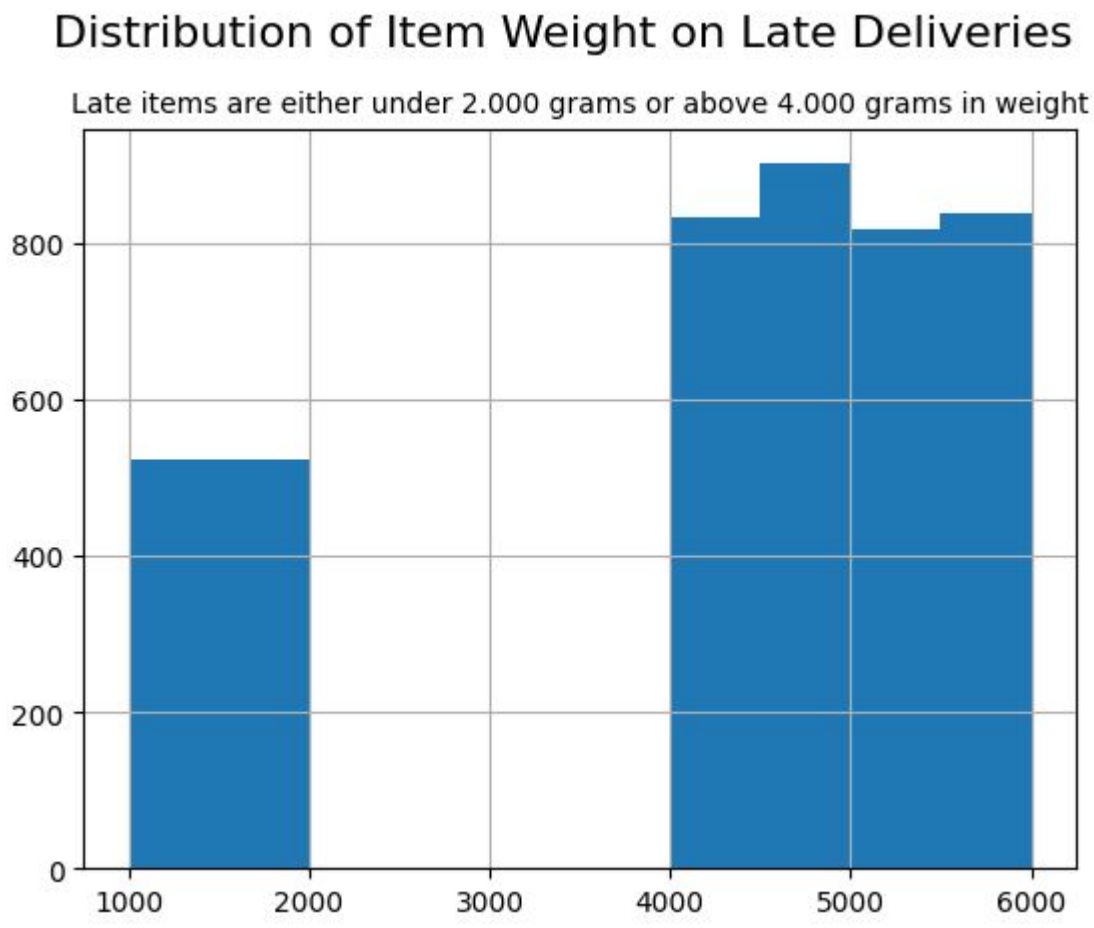
Multivariate Analysis - Pair Plot

- Berdasarkan hasil korelasi antara feature, terdapat beberapa hal yang perlu dilakukan, yaitu:
    - Beberapa data outlier pada berat barang (*Weight_in_gms*) yang perlu diobservasi lebih lanjut untuk memastikan apakah data tersebut valid atau tidak. Jika data tersebut valid, maka dapat dipertimbangkan untuk menggunakan teknik *pre-processing* seperti pengurangan dimensi (PCA) atau penanganan outlier untuk memperbaiki performa model.

Business Insights

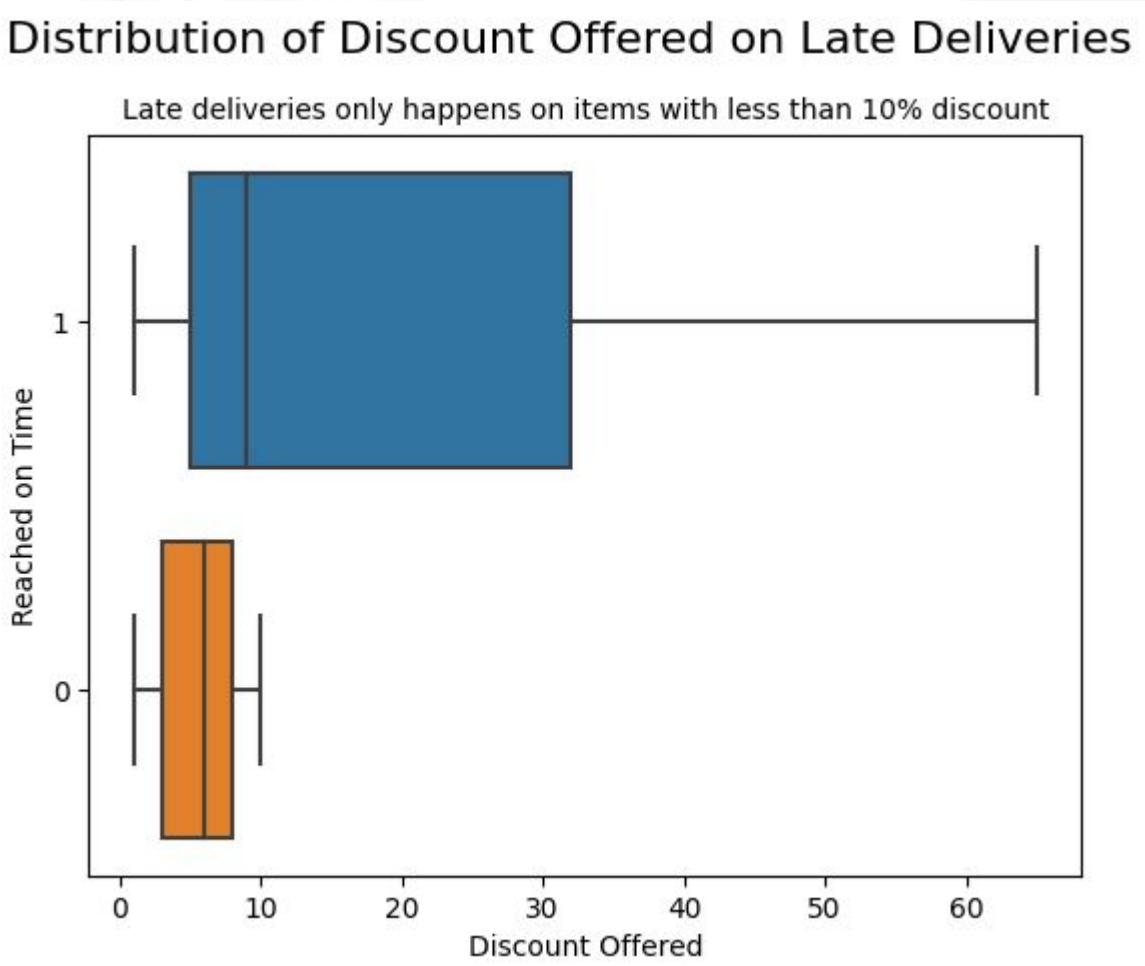
- Berdasarkan data yang tersedia, rasio keterlambatan pengiriman berada di 40.3% dari total pengiriman
- Beberapa fitur yang diduga memiliki korelasi dengan keterlambatan pengiriman adalah:
  - Berat barang (*Weight_in_gms)*
  - Diskon (*Discount_offered*)
- Terdapat beberapa data outlier jika dilihat berdasarkan berat barang (*Weight_in_gms*), yaitu barang-barang yang beratnya melebihi 6000 gram, namun untuk barang-barang tersebut tidak ada satupun yang mengalami keterlambatan (walaupun jumlah sampel outlier tersebut tidak banyak untuk dapat mengambil kesimpulan yang bermakna)
- Nilai *Customer rating* tidak memiliki korelasi dengan keterlambatan (*Reached_on_time*) bahkan juga dengan *feature-feature* lain.

Business Insights - Recommendations



- Keterlambatan hanya terjadi pada barang yang beratnya ada **di bawah 2000 gram** atau di antara **4000-6000 gram**. Diperlukan analisis lanjutan untuk mengetahui penyebab keterlambatan. Bisa jadi, perusahaan tidak memiliki partner shipment yang secara khusus mengantarkan paket produk untuk perusahaannya. Sehingga pengiriman barang bercampur dengan perusahan lain berdasarkan berat barang, *mode of shipment* ataupun faktor lain dan akhirnya terjadi keterlambatan.

Business Insights - Recommendations



- Rekomendasi : memberikan **pengawasan** / **perhatian** khusus pada **barang-barang** yang diberikan **discount <10%**. Perlu diperiksa apakah ada ketentuan perusahaan atau situasi yang berbeda pada barang-barang yang diberikan diskon dibawah 10% dibandingkan dengan jumlah lainnya, yang dapat mempengaruhi keterlambatan

Git

Link Repository Github: https://github.com/mezkymy/ecommerce-ds