Table of Contents

Abstract

The purpose of Data Analysis is to discover meaningful information from a dataset supporting decision-making in businesses. Conducting Data Analysis includes a huge range of activities such as cleaning, transforming and modelling data.
Data visualisation is about using tools and technologies to interpret data visually. It aims to recognise trends, patterns, and relationships between elements in the dataset and deliver meanings from data.

The planned approach to investigate the problem start with understanding the dataset and cleaning data. Going through the description of elements helps me to have a sense of the dataset in relation to the problems the company is facing. The next plan will be conducting data analysis comprising of cleaning and transforming data and then making data visualisation to interpret the information the company is looking for.

From the Data analysis task, there is an expectation that the business question which is reaching on time of deliveries is determined by which factors will be solved.

1. Introduction and Background

1.1 The problem you tried to solve

Nowadays, E-Commerce is getting more and more popular because of its advantages such as timesaving and convenience. Shipping including delivering items to customers on time is considered one of the most important things making contributions to the success of companies who operate in the E-Commerce field. In the report, how the warehouse location, mode of shipping, the weight of items, the importance of products, product prices, and offered discounts affect the possibility of shipping coming on time and how delivery time affects customers' satisfaction will be explored. From the analysis, it is expected that the company will have observations of which factors/attributes affect the possibility of delivering on time.

1.2 Business Question

As a result of the data exploratory analysis, it is expected to find out answers to the main question which is the factor/attribute that has the most significant impact on the possibility of orders to reach on time. In addition to the main question, the list of below questions is also expected to be answered from the analysis.

1. The number of orders reached on time versus not on time in the dataset
2. How to allocate orders in warehouse block and its effects on the delivery time
3. The distribution of mode of shipment and its effects on the delivery time
4. The relationship between the importance of products and their costs, and the possibility of the products being delivered on time.
5. Are customer care calls and customer ratings affected by the delivery time or other factors, for example, discount offers?
6. Whether the weight of products have an impact on the mode of delivery and delivery time?

1.3 Dataset

The dataset is about shipping data in an e-commerce company from Kaggle dataset (**https://www.kaggle.com/datasets/prachi13/customer-analytics?resource=download**). It contains 12 attributes as follows:

1. ID: The number of customers who are recorded in the dataset
2. Warehouse block: The company's warehouse block (including blocks A,B,C,D,E)
3. Mode of Shipment: How products are shipped (by Ship, Flight or Road)
4. Customer care calls: The number of calls to make inquiries about the shipment
5. Customer rating: Rating from customer (Rank from the lowest: 1 to the highest: 5)
6. Cost of the product: Cost of the product (in US Dollars)
7. Prior Purchase: The number of prior purchase
8. Product importance: How important products (ranked as low, medium and high)
9. Gender: customers' gender (Male or Female)
10. Discount offered: Discount on products
11. Weight in gms: The weight of products (in grams)
12. Reached on time: Products are shipped on time or not (0: it reached on time, 1: it does not reach on time)

In order to answer the above business questions, the analysis will be focused on the attributes such as Warehouse block, Mode of Shipment, Customer care calls, Cost of the project, Prior Purchase, Discount offered, Weigh in gms and especially Reached on time.

2. Overview of the Data Analysis Pipeline

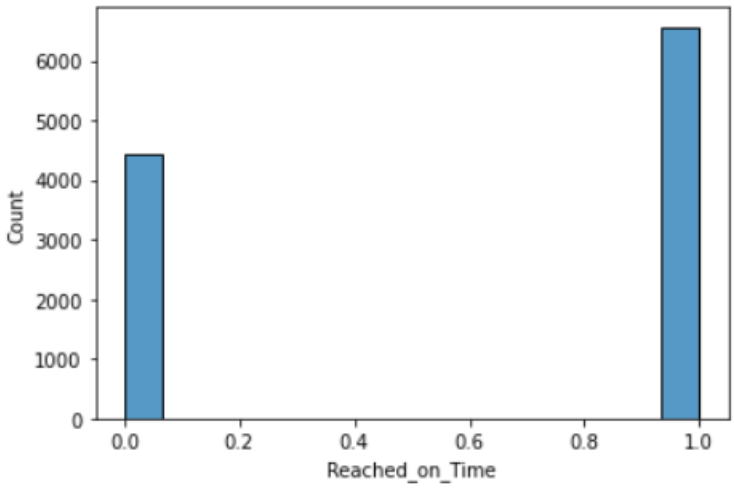
2.1 Data Preparation

For data preparation, the techniques have been used are head, tail, rename and replace. Statistical methods are also applied and shown as below:

	ID	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Discount_offered	Weight_in_gms	Reached_on_Time	F
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	

With a focus on the possibility of delivering on time, the statistics indicate that mean of reaching on time is 0.59. The mean of customer care calls who called for asking the delivery is more than 4, which is high. The mean cost of the product is also high at around USD $210 while the discount offered is quite low with the mean of about USD $13.

Using the histogram technique, the plot of Reached_on_Time attribute is as below:



2.2 Missing value exploration

Since the dataset should not have missing or null values, checking both null and missing values have been applied. Tables show the values after checking in the below:

```
#checking null values
df.isnull().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         0
Price_after_discount    0
Percentage_of_discount  0
dtype: int64
```

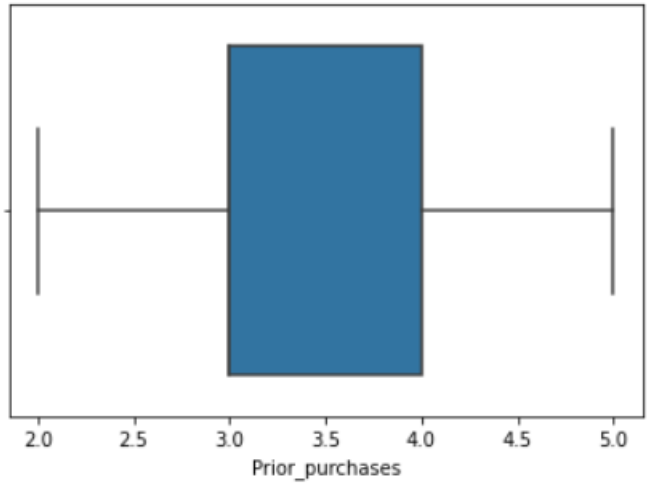
```
#checking missing values
df=df.dropna()
print(df.isnull().sum())
```

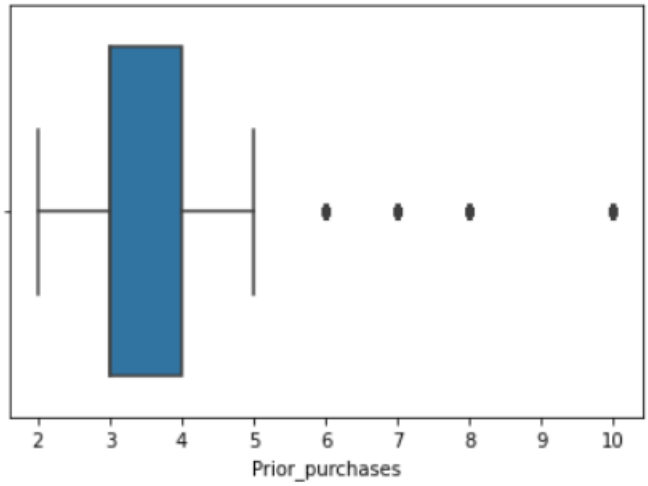
```
ID                      0
Warehouse_block         0
Mode_of_Shipment        0
Customer_care_calls     0
Customer_rating         0
Cost_of_the_Product     0
Prior_purchases         0
Product_importance      0
Gender                  0
Discount_offered        0
Weight_in_gms           0
Reached_on_Time         0
Price_after_discount    0
Percentage_of_discount  0
dtype: int64
```

From the table, it can be said that the dataset has no missing or null values. Therefore, it is unnecessary to use techniques to handle them.

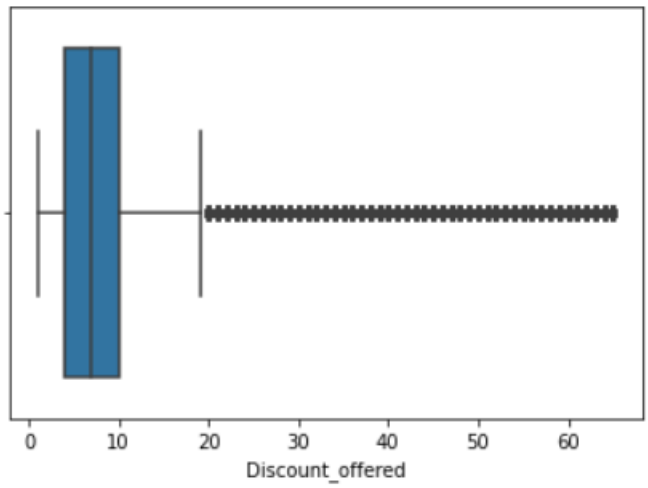
2.3 Outlier identification

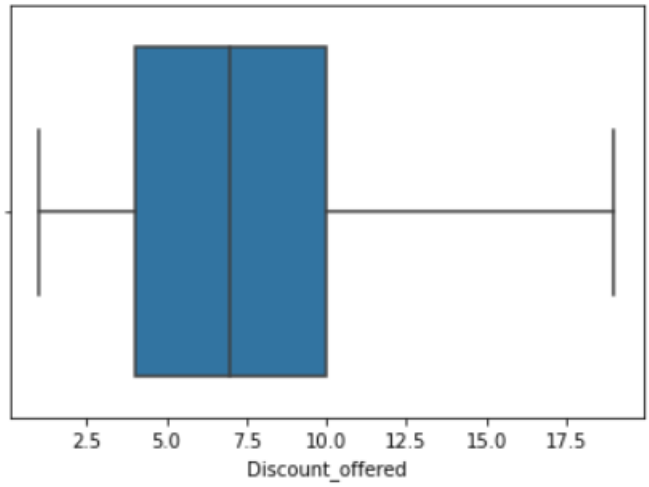
For outliner identification, first, using box plots for all numeric values in the dataset to check outliners. Observations are only prior purchases and discounts offered have outliners. Second, using Interquartile Range, it is indicated that there are 1003 outliers in Prior purchases. Because of the high value, if these outliers are removed, visualisation of the dataset will not fully reflect the current situation. Due to these reasons, replacing the outliners with mean is applied for both Prior purchases and Discount offered attributes. Box plots indicating the outliers in Prior purchases before and after using the handling outliner technique are as follows:





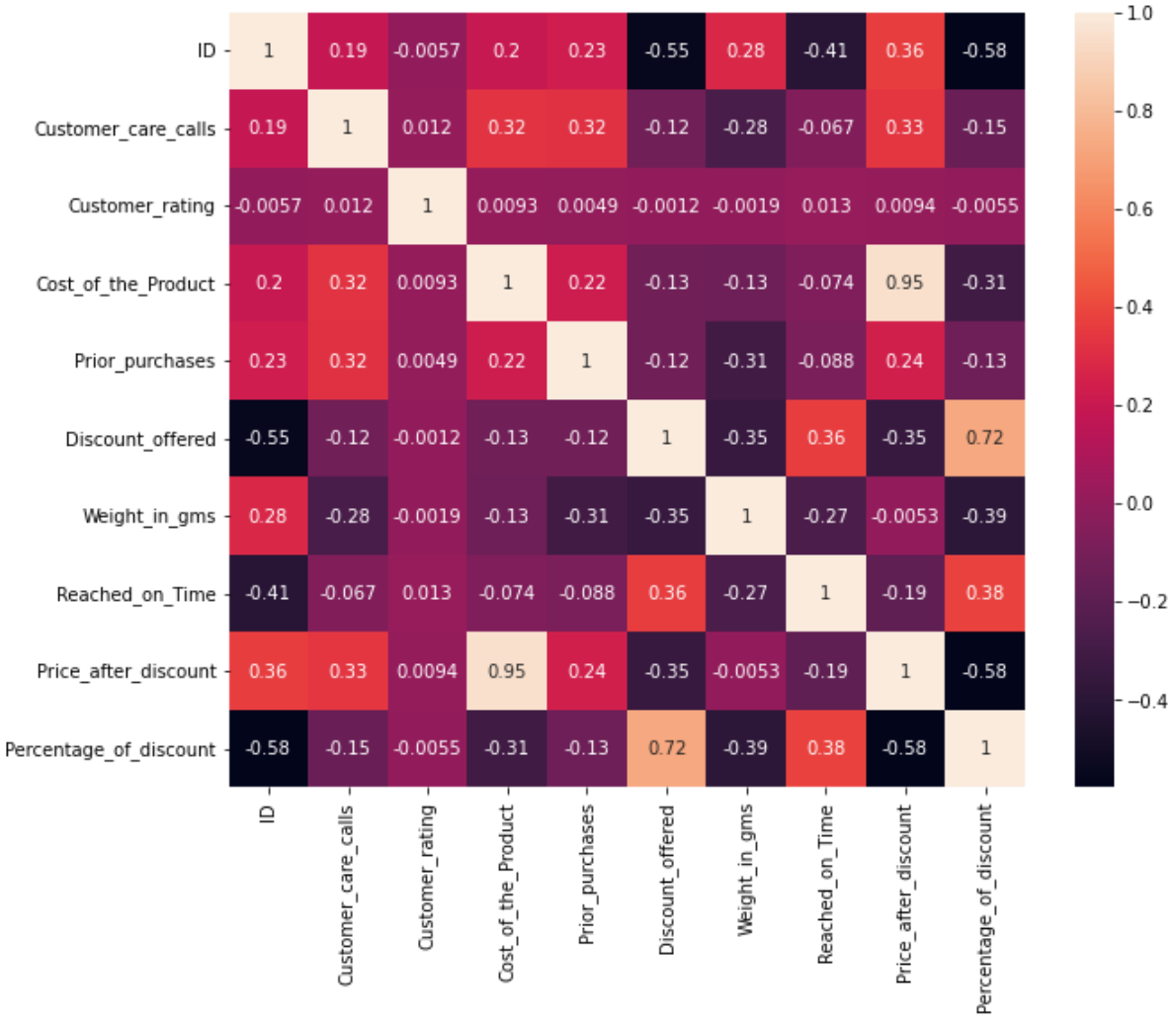
Box plots indicating the outliers in Discount offered before and after using handling outliner technique are as follows:



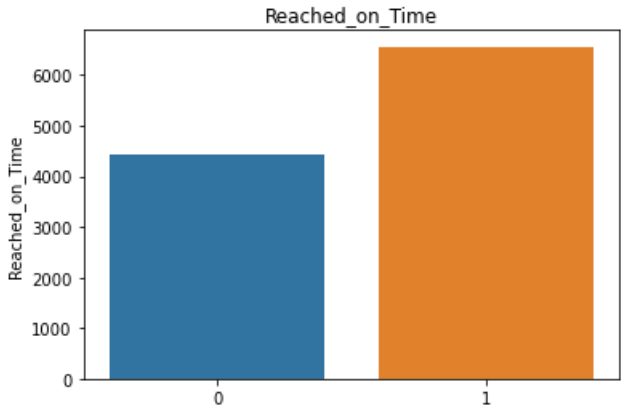


2.4 Data Visualization

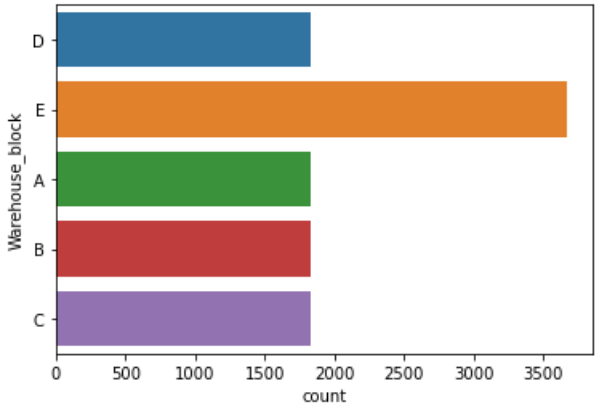
Using the heatmap to show the correlation between different attributes in the dataset. If the number in the below heatmap is more nearly to 1, the more correlation between the attributes. From the visualisation, it can be seen that the discount offered has the most significant effect on the possibility of delivering on time although the effect is not high.



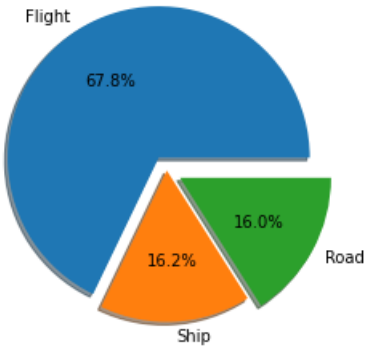
The below bar plot demonstrates the number of deliveries that reached on time versus not on time. The deliveries reached on time is about 1500 deliveries higher than not on time.



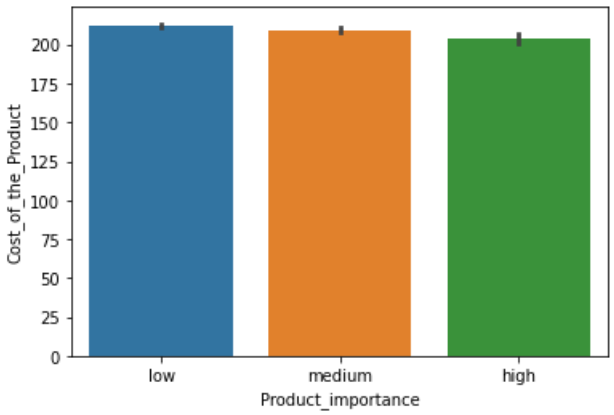
In order to demonstrate the number of orders for each warehouse block, the below bar plot is applied. It can be seen that orders are distributed with the same number of orders except block E.



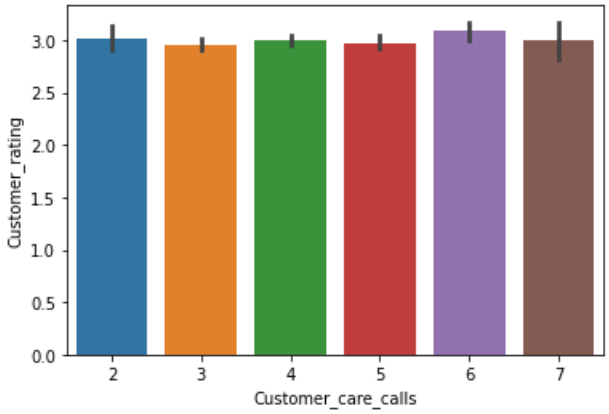
The pie chart illustrates the mode of shipment in the dataset. Most of the deliveries are shipped by flight, taking account 67.8% whereas deliveries using ship and road are 16.2% and 16% respectively.

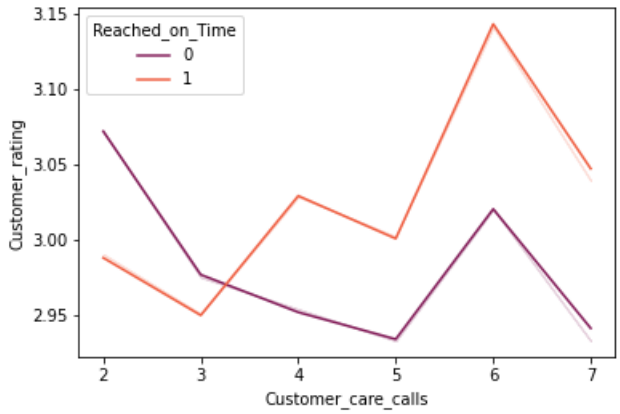


The relationship between the cost of the Project and Product importance is illustrated as in the below bar plot. It can be said that the importance of the products does not depend on the cost of the product since the importance of the products varies from high to low cost.



To visualise customer rating and customer care calls and their impact on the possibility of delivering on time, the bar chart and line chart are as below used. From the visualisation, it is seen that customer care calls do not affect customer rating and it is considered that the possibility of reaching on time does not depend on customer rating and customer care calls.





3.Discussion and Conclusions

The data exploration experiment answers the main question which is discount offered is the discount offered has the most significant impact on the delivery to reach on time. The additional questions that are being asked in the above section are also answered:

1. The number of orders reached on time is more than not on time in the dataset. It is a positive indication for e-commerce shipping.
2. Order is allocated in different warehouses nearly equal and it does not affect the delivery time.
3. Mode of shipment is used most is by plane but despite the speed of the mode of shipment, it does not affect the delivery time.
4. Although the higher cost of the product is considered to be a high level of importance, the reality reflected in the dataset is the assumption is incorrect. Furthermore, both the cost of the product and the product's importance does not have much impact on the possibility of reaching on time.
5. Assuming that there are more customer care calls results in lower customer ratings but from the data analysis, the relationship between these attributes is unclear. These factors also have a low impact on reaching on time.
6. The heavy products are highly likely to be delayed. However, the data analysis indicates that the weight of products does not impact the mode of delivery and delivery time.

In conclusion, data analysis and data visualization are used to discover and interpret meaningful insights from a dataset with the aim of supporting business decision-making. Analysing E-Commerce Shipping Data with data analysis and data visualisation techniques is exploring the performance of shipping in an international e-commerce company and how it affects customers' satisfaction; as such, the organisation could make adjustments in shipping plans to improve the overall business performance.