LOVELY PROFESSIONAL UNIVERSITY

Transforming Education Transforming India

Name : Eshwar Anugandula

Roll No: RK20KTB41

Reg. no:12019502

Course: INT 353

E- Commerce Shipping Data

Introduction:

This dataset is about the products sent by shipping mode. It consists about the information of different means of transport of which product has been delivered, Cost of product, Rating of customer, discount offered, weight of the product, It completely gives information about the details of products have been shipped.

Reason for choosing this data set:

An international e-commerce company based wants to discover key insights from their customer database. They want to use some of the most advanced machine learning techniques to study their customers. The company sells electronic products. Now a days everybody is experiencing the door delivery of the products. We can analyse whether the product received with safety, time taken for delivery, cost of shipping, It is easier to know which mode is best to order a particular product. There are 12 columns and 11000 rows which helps me analyse, visualise and interpret the data set in effective manner.

Data set column description:

1	ID	ID Number of Customers.
2	Warehouse block	The Company have big Warehouse which is divided in to block such as A,B,C,D,E.
3	Mode of Shipment	The Company Ships the products in multiple way such as Ship, Flight and Road.
4	Customer care calls	The number of calls made from enquiry for enquiry of the shipment.
5	Customer rating	The company has rated from every customer. 1 is the lowest (Worst), 5 is the highest (Best).
6	Cost of the Product	Cost of the Product in US Dollars.
7	Prior purchases	The Number of Prior Purchase.
8	Product importance	The company has categorized the product in the various parameter such as low, medium, high.
9	Gender	Male and Female
10	Discount offered	Discount offered on that specific product.
11	Weight in grams	It is the weight in grams.
12	Reached on time	It is the target variable, where 1 Indicates that the product has NOT reached on time and 0 indicates it has reached on time.

Insights from the data set:

Importing necessary imports from library

Loading data

Finding no of rows and columns

Description of data

Information of columns

Checking null values

Data cleaning by dropping unwanted columns

heatmap of the data for checking the correlation between the features and target column

Checking value counts of categorical columns

Exploring relation of categorical columns with reached on time or not
looking at the warehouse column and what are the categories present in it

making a count plot of warehouse column and see the effect of Reached on time or not on the warehouse column.

looking at the gender column and what are the categories present in it

making a count plot of gender column and see the effect of Reached on time or not on the warehouse column.

making a count plot of mode of shipment column and see the effect of Reached on time or not on the warehouse column.

looking at the product importance column and what are the categories present in it

making a count plot of product importance column and see the effect of Reached on time or not on the warehouse column.

looking at the customer ratings column and what are the categories present in it

making a count plot of prior purchases column and see the effect of Reached on time or not on the warehouse column.

making a dist. plot of cost of the product column

making a dist. plot of discount offered column

Which mode of shipment carries most weights?

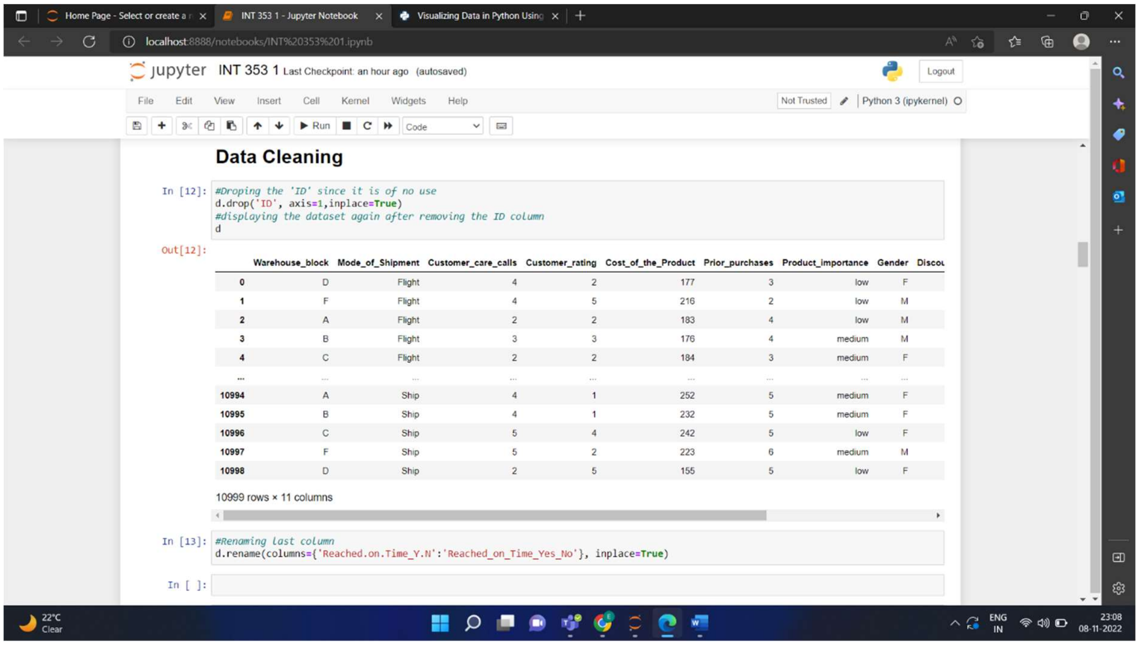
Effect of Warehouse on Cost of Product

Does Mode of Shipment effect Cost of Product?

the relation between cost of the product and the discount offered and the relation with whether or not the product will reach on time.

Conclusion

Data Cleaning:


Home Page - Select or create a n
INT 353 1 - Jupyter Notebook
Visualizing Data in Python Using
localhost:8888/notebooks/INT%20353%201.ipynb
jupyter INT 353 1 Last Checkpoint: an hour ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
22°C
Clear
ENG
IN
23:08
08-11-2022

Data Cleaning

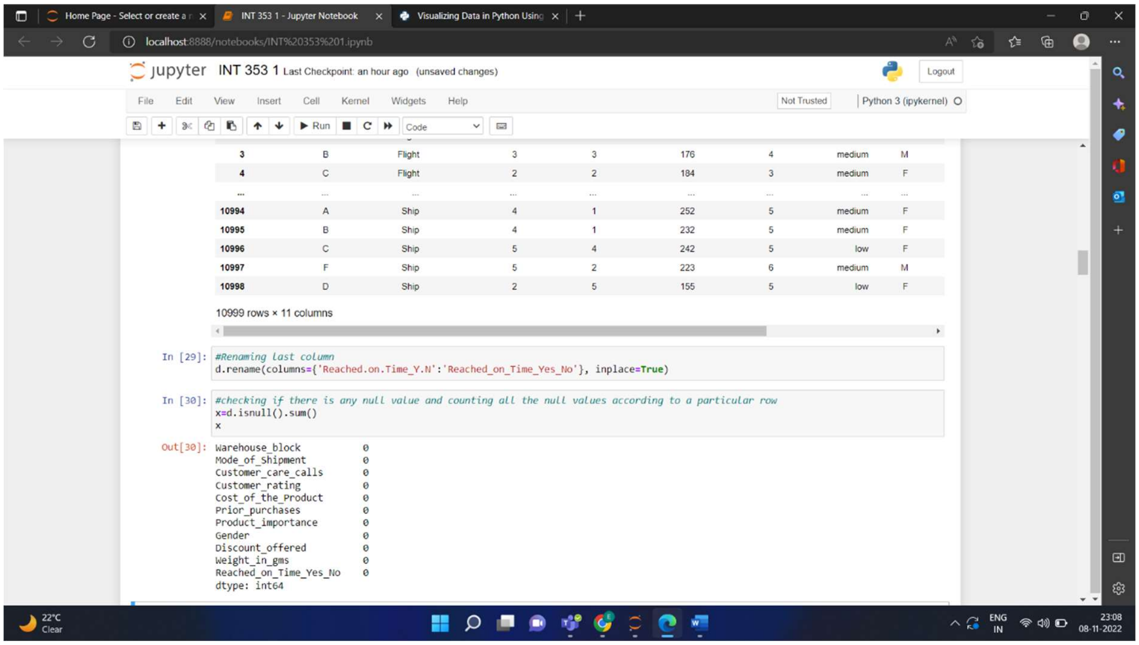
```
In [12]: #Droping the 'ID' since it is of no use
         d.drop('ID', axis=1,inplace=True)
         #displaying the dataset again after removing the ID column
         d
```

Out[12]:

	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender	Discou
0	D	Flight	4	2	177	3	low	F	
1	F	Flight	4	5	216	2	low	M	
2	A	Flight	2	2	183	4	low	M	
3	B	Flight	3	3	176	4	medium	M	
4	C	Flight	2	2	184	3	medium	F	
...	...	...	...	...	...	...	...	...	
10994	A	Ship	4	1	252	5	medium	F	
10995	B	Ship	4	1	232	5	medium	F	
10996	C	Ship	5	4	242	5	low	F	
10997	F	Ship	5	2	223	6	medium	M	
10998	D	Ship	2	5	155	5	low	F	

10999 rows × 11 columns

```
In [13]: #Renaming last column
         d.rename(columns={'Reached.on.Time_Y.N':'Reached_on_Time_Yes_No'}, inplace=True)

In [ ]:
```


Home Page - Select or create a n
INT 353 1 - Jupyter Notebook
Visualizing Data in Python Using
localhost:8888/notebooks/INT%20353%201.ipynb
jupyter INT 353 1 Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
22°C
Clear
ENG
IN
23:08
08-11-2022

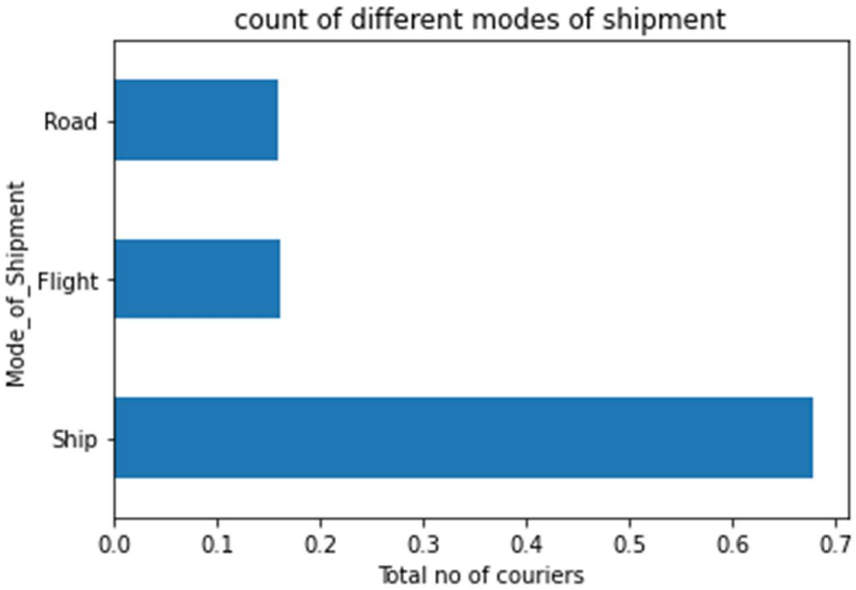
3	B	Flight	3	3	176	4	medium	M
4	C	Flight	2	2	184	3	medium	F
...	...	...	...	...	...	...	...	...
10994	A	Ship	4	1	252	5	medium	F
10995	B	Ship	4	1	232	5	medium	F
10996	C	Ship	5	4	242	5	low	F
10997	F	Ship	5	2	223	6	medium	M
10998	D	Ship	2	5	155	5	low	F

10999 rows × 11 columns

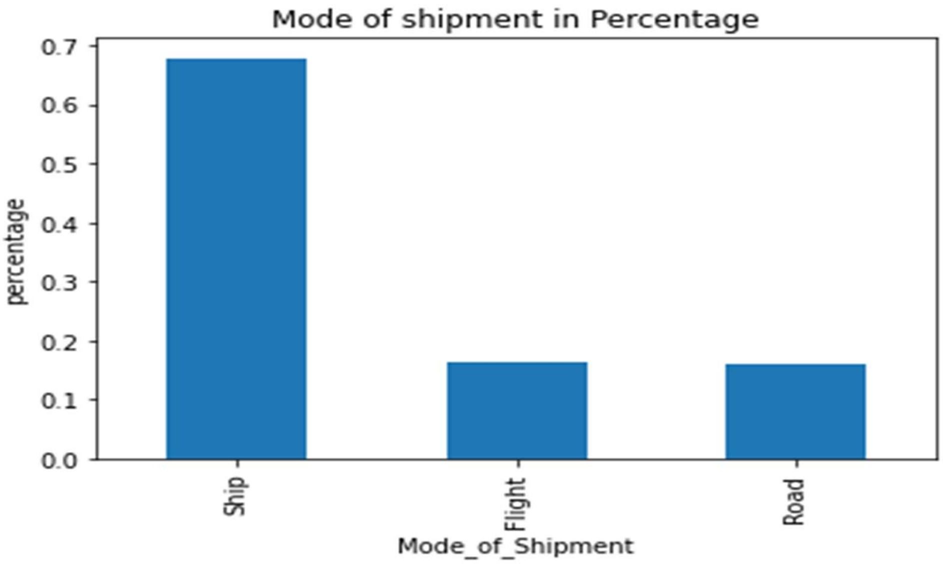
```
In [29]: #Renaming last column
         d.rename(columns={'Reached.on.Time_Y.N':'Reached_on_Time_Yes_No'}, inplace=True)

In [30]: #checking if there is any null value and counting all the null values according to a particular row
         x=d.isnull().sum()
         x

Out[30]: 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_Yes_No    0
         dtype: int64
```

Univariate Analysis:

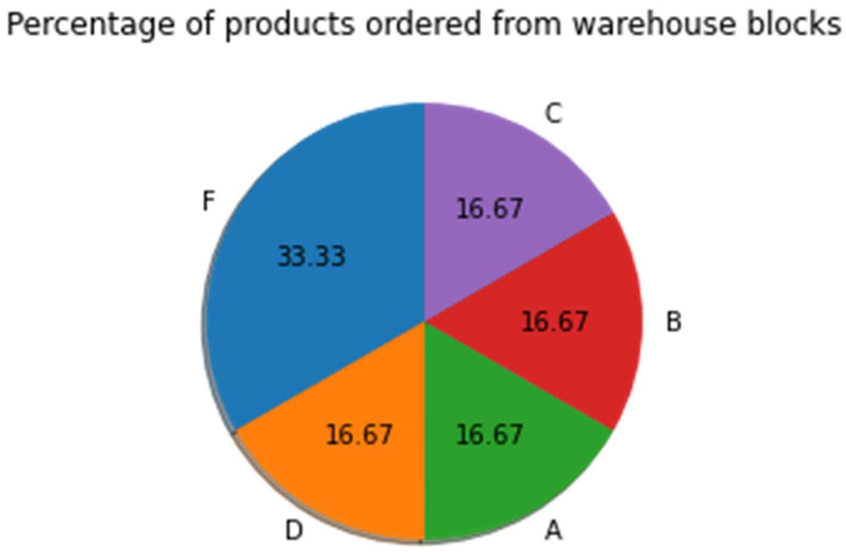
Categorical features:



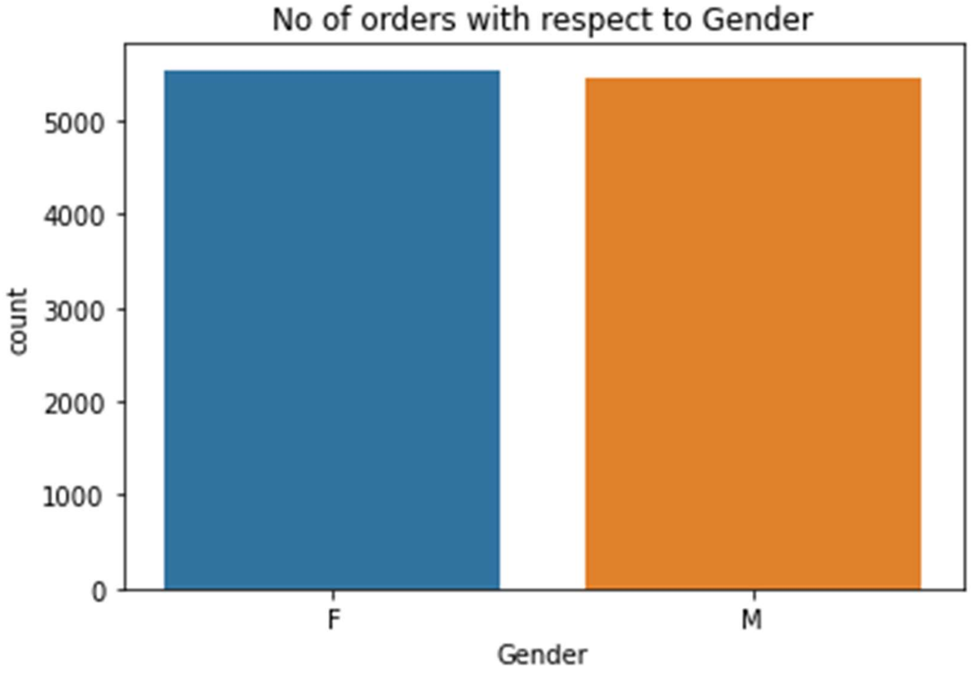
From the above bar graph we can see that most of the shippings are done by “Ship” that is in water ways



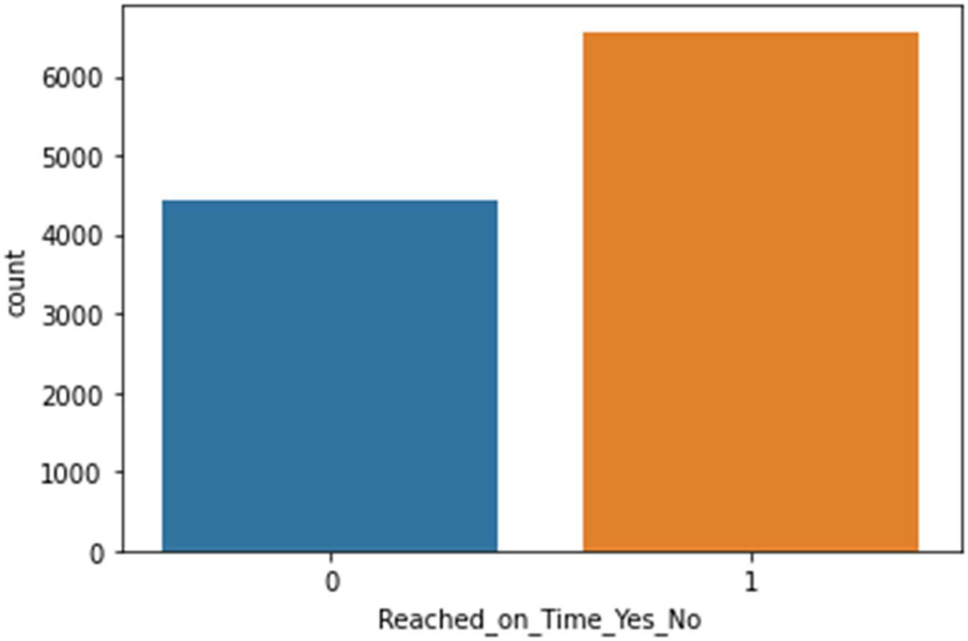
The above bar graph shows in percentage. About 68% of the shipments are done by ship and 16% of shipments are done by both roads and flights



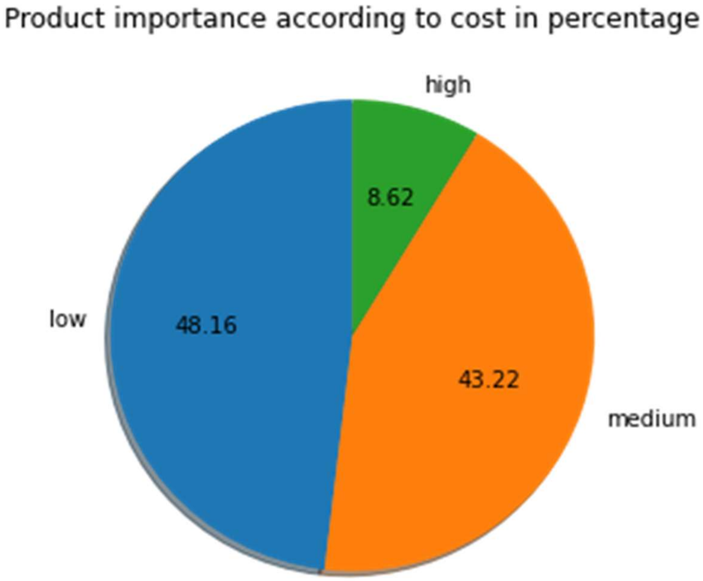
In this pie chart we can clearly depict that most of the shipments are delivered to block “F” (33.33%) and rest of the blocks has equal weightage.



In this bar graph it is clear that count of female’s shipments are slightly higher than male’s.

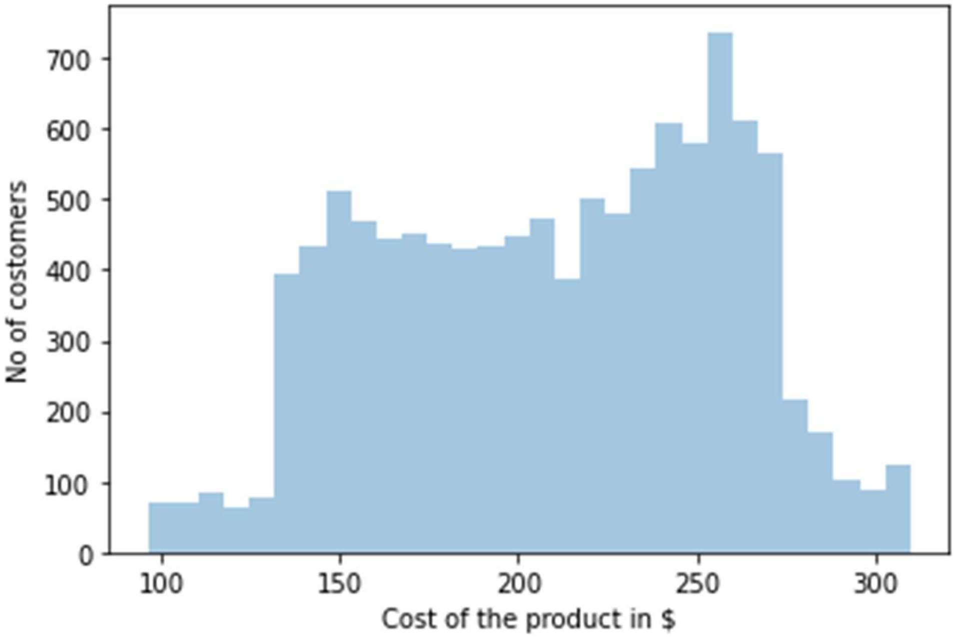


In this bar graph we can conclude that most of the products are reached on time but the count of the products which are not reached on time is also high.

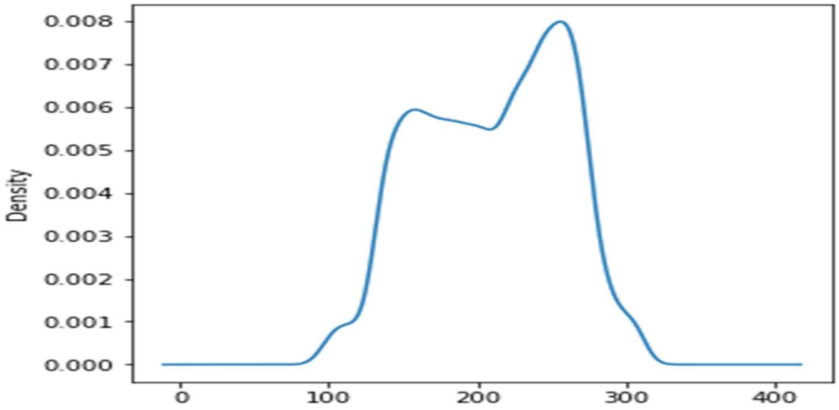


In this pie chart it is clear that most of the products ordered where low cost, and medium cost products. Only 8.62% of the high cost products were shipped.

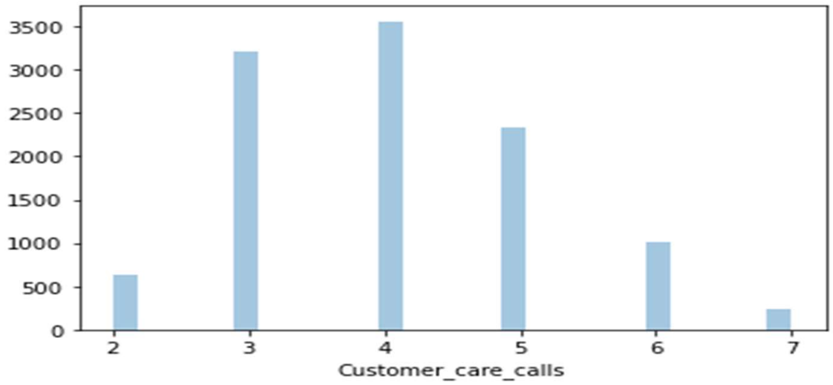
Numerical features



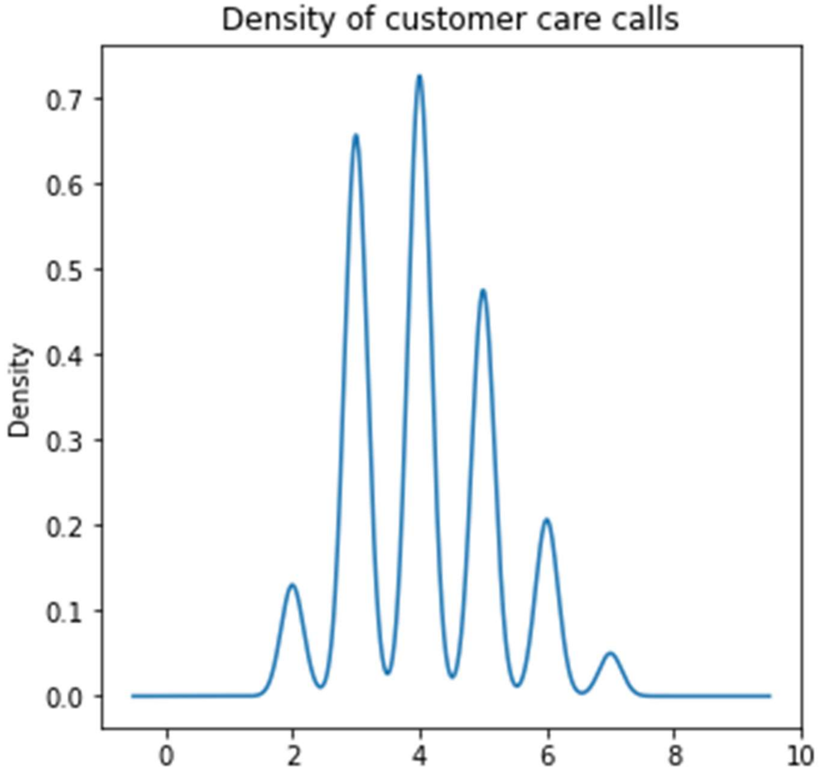
The histogram clearly shows that most of the products ordered were in between the range of $250 to $255 where as least no products ordered were around $100 to $110



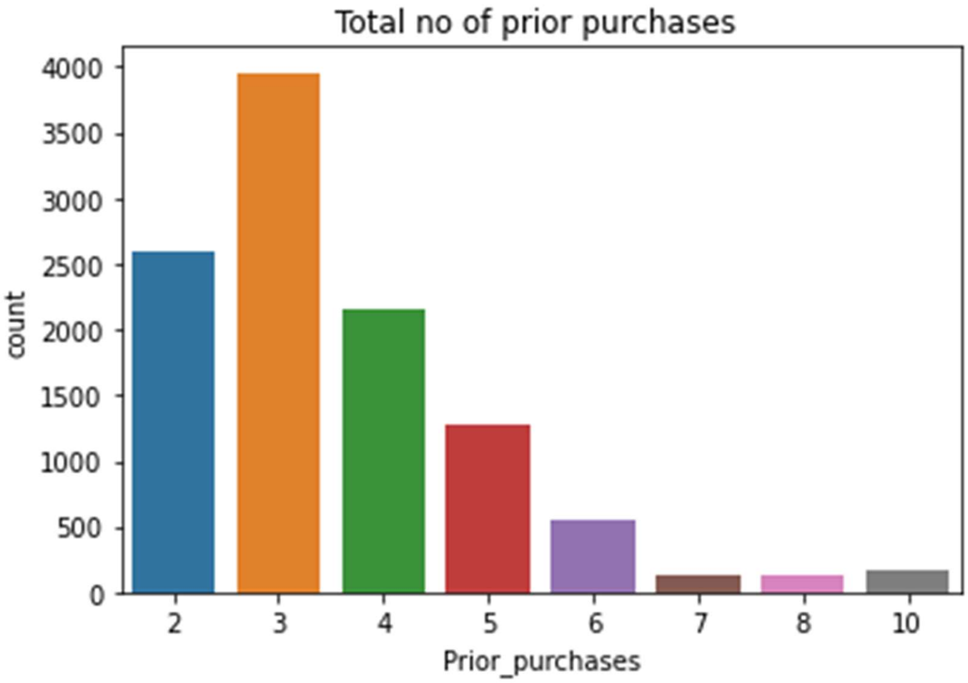
This is the density plot for cost of the products where can see that density is high at $250 to $300.



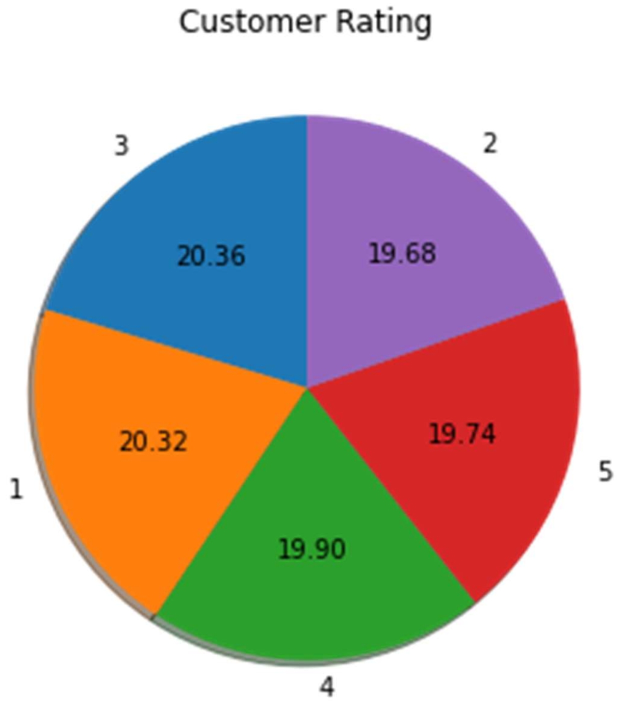
The no of customer care call received are 4 with highest frequency.



This is the density plot for customer care calls

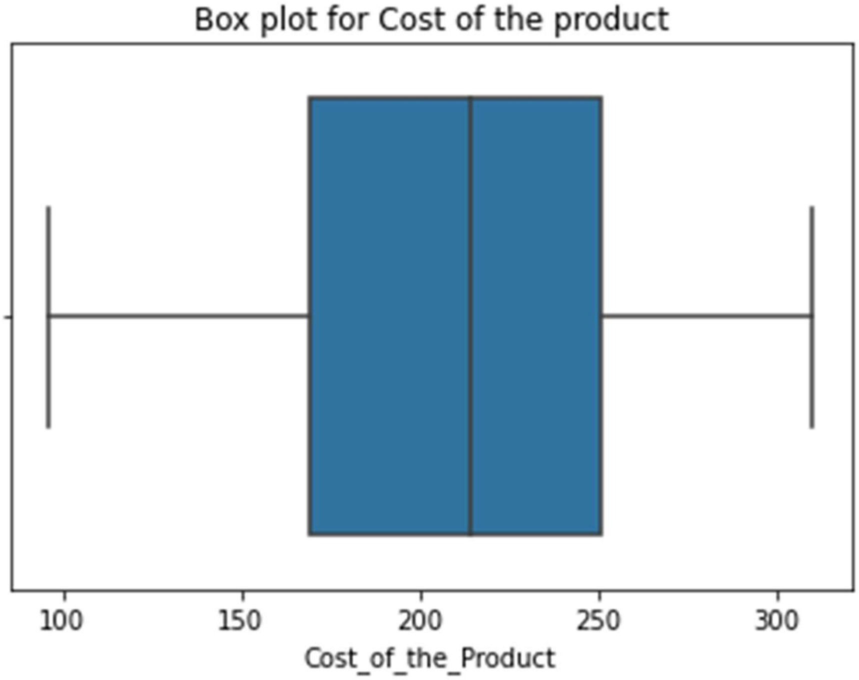


Highest prior purchases are 7 and lowest prior purchases are 3.



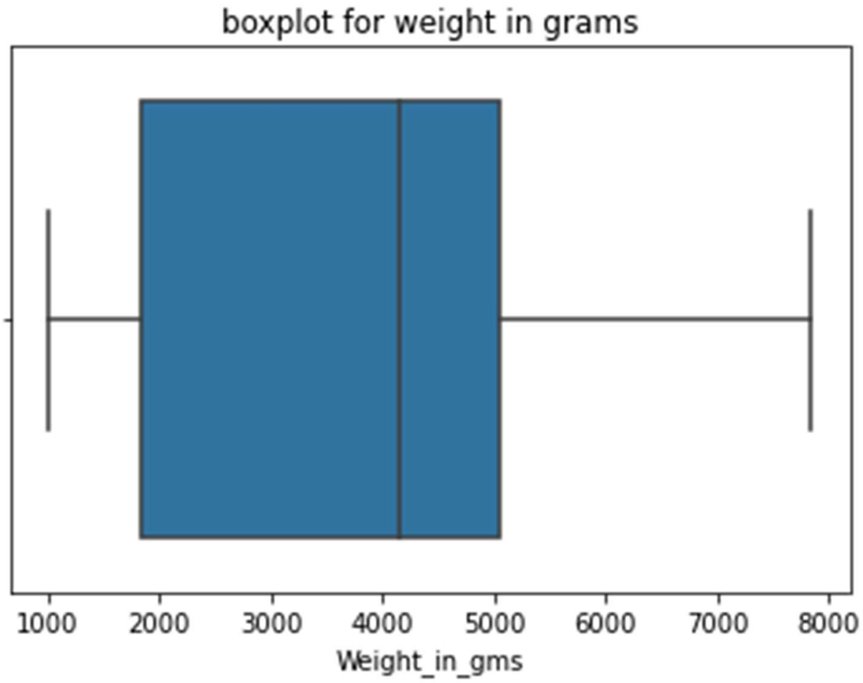
From the above pie chart we can conclude that rating of “3” is given by most of the customers.

Box plot for the “Cost of the product”



```
       count    10999.000000
mean       210.196836
std         48.063272
min         96.000000
25%        169.000000
50%        214.000000
75%        251.000000
max        310.000000
Name: Cost_of_the_Product, dtype: float64
```

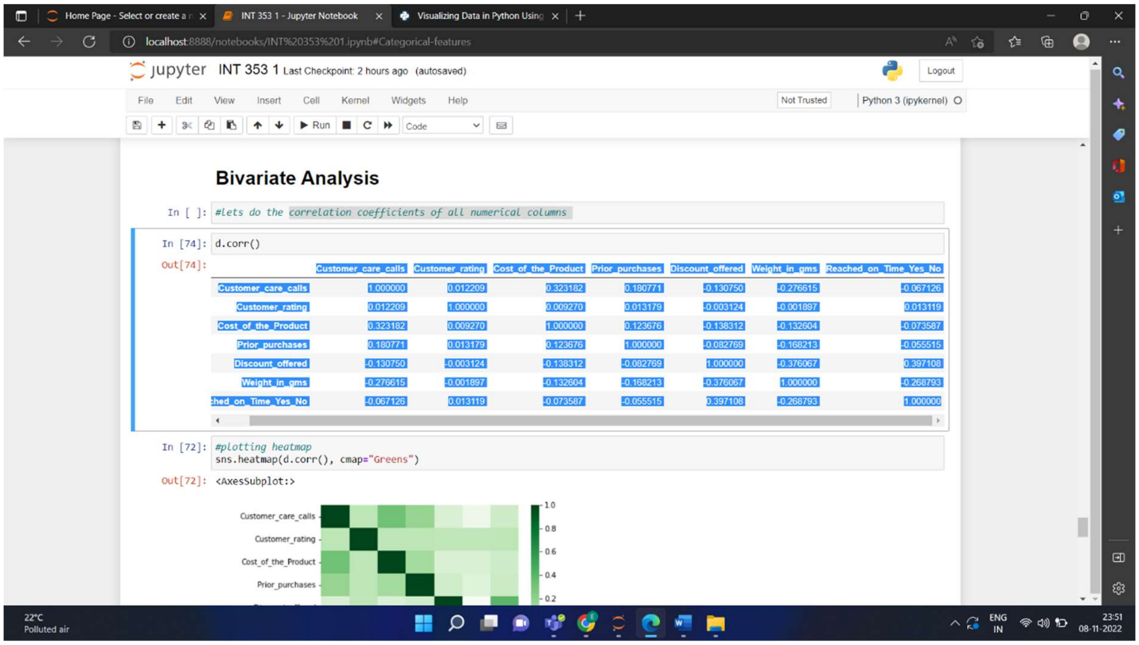
Boxplot for weight in grams



```
count    10999.000000
mean      3634.016729
std       1635.377251
min       1001.000000
25%       1839.500000
50%       4149.000000
75%       5050.000000
max       7846.000000
Name: Weight_in_gms, dtype: float64
```

Bivariate Analysis

correlation coefficients of all numerical columns:


Home Page - Select or create a
INT 353 1 - Jupyter Notebook
Visualizing Data in Python Using
localhost:8888/notebooks/INT%20353%201.ipynb#Categorical-features
jupyter INT 353 1 Last Checkpoint: 2 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
22°C
Polluted air
ENG
IN
23:51
08-11-2022

Bivariate Analysis

```
In [ ]: #lets do the correlation coefficients of all numerical columns
```

```
In [74]: d.corr()
```

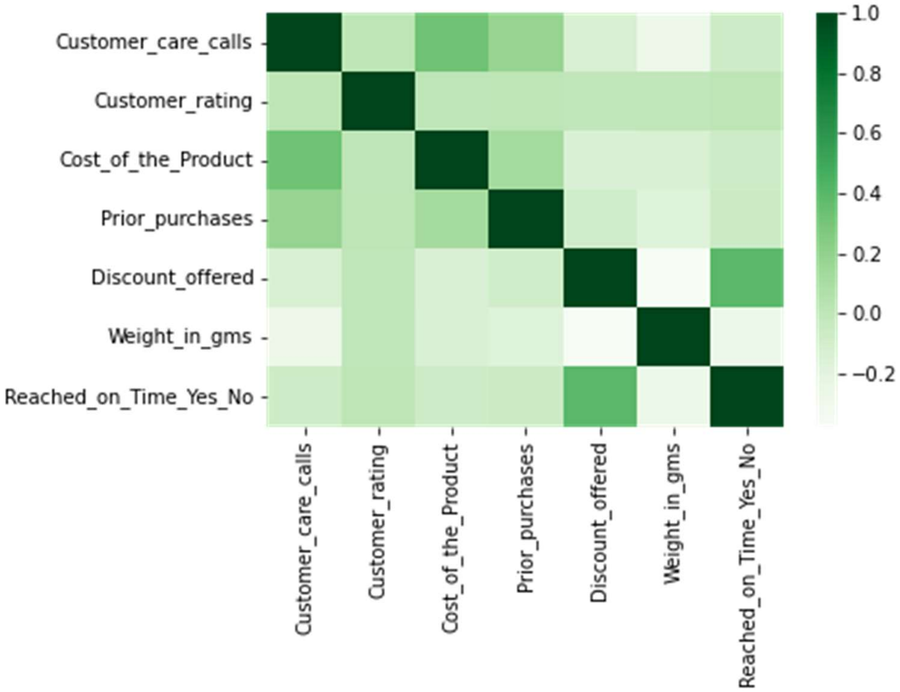
Out[74]:

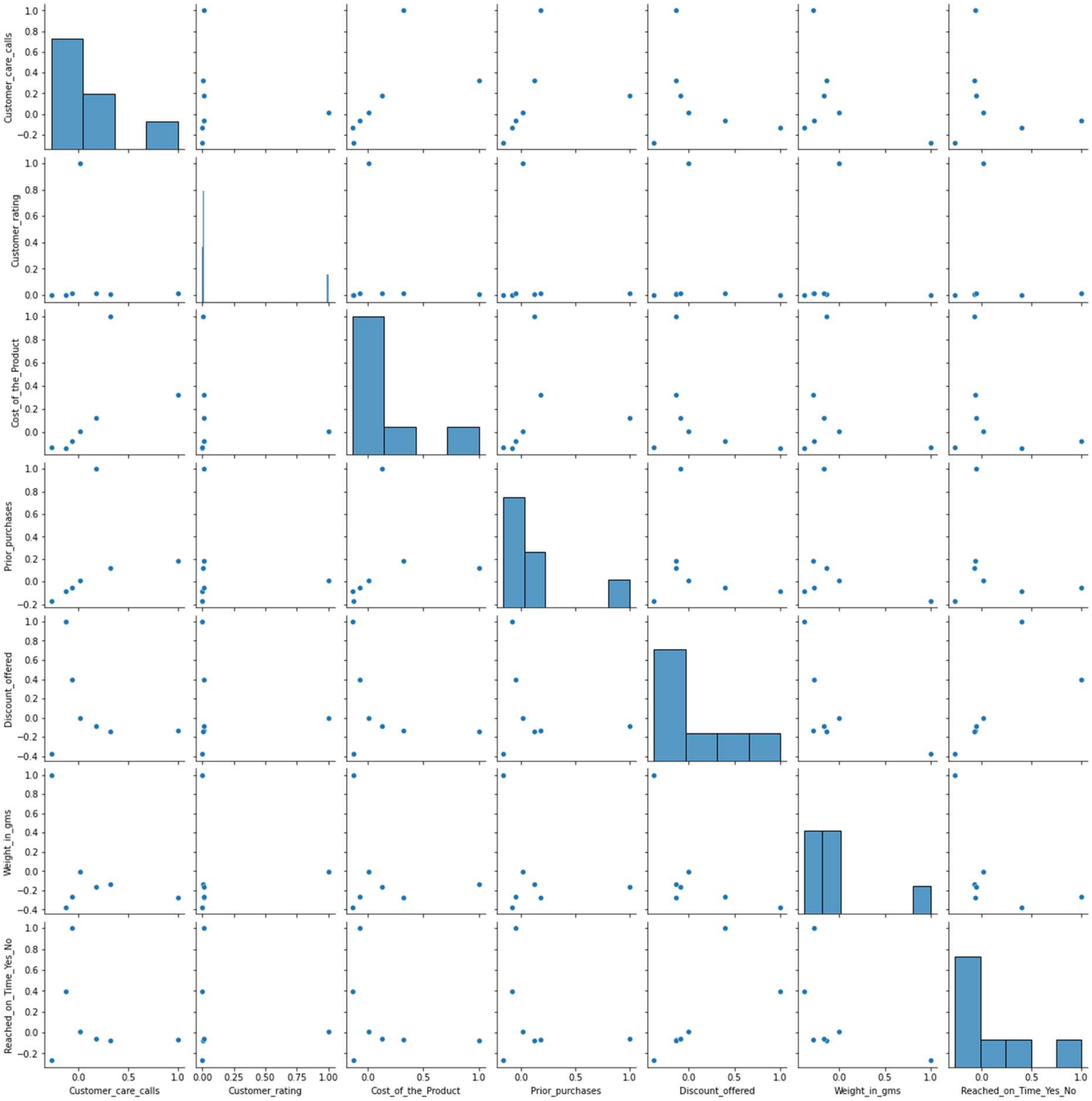
	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Discount_offered	Weight_in_gms	Reached_on_Time_Yes_No
Customer_care_calls	1.000000	0.012209	0.323182	0.180771	-0.130750	-0.276615	-0.067126
Customer_rating	0.012209	1.000000	0.009270	0.013179	-0.003124	-0.001897	0.013119
Cost_of_the_Product	0.323182	0.009270	1.000000	0.123676	-0.138312	-0.132604	-0.073587
Prior_purchases	0.180771	0.013179	0.123676	1.000000	-0.082769	-0.168213	-0.055515
Discount_offered	-0.130750	-0.003124	-0.138312	-0.082769	1.000000	-0.376067	0.397108
Weight_in_gms	-0.276615	-0.001897	-0.132604	-0.168213	-0.376067	1.000000	-0.268793
hed_on_Time_Yes_No	-0.067126	0.013119	-0.073587	-0.055515	0.397108	-0.268793	1.000000

```
In [72]: #plotting heatmap
         sns.heatmap(d.corr(), cmap="Greens")
```

```
Out[72]: <AxesSubplot:>
```

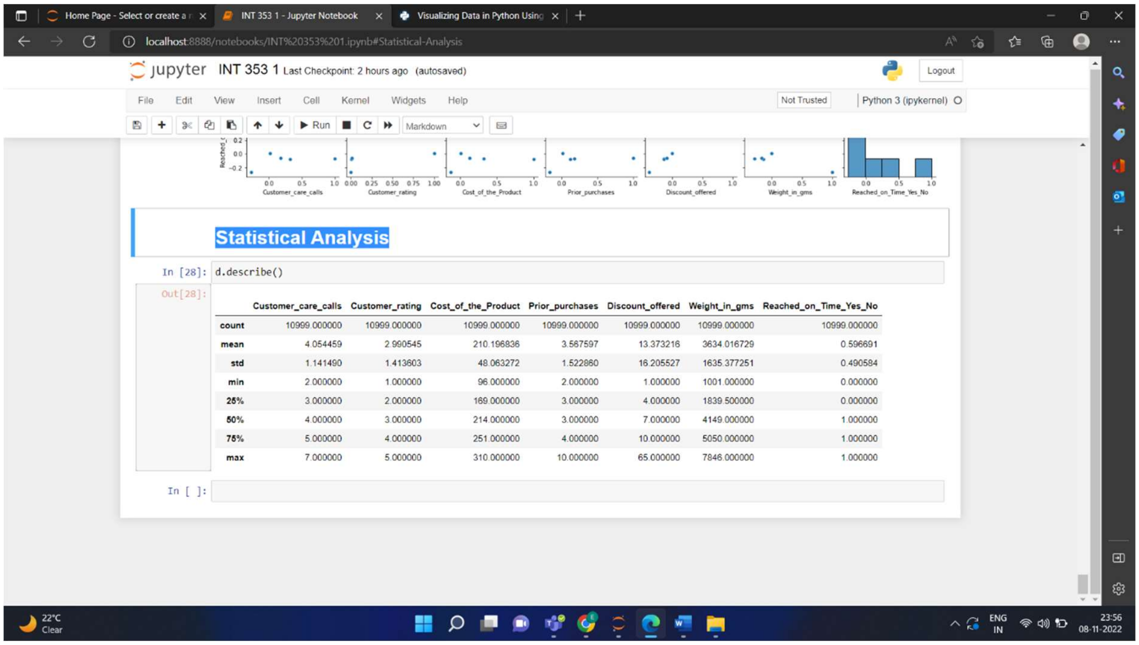
Heatmap for corelation coefficients:





This is pair plot for ear and every column of the data set. A very few graphs shows linear regression.

Statistical Analysis:



Statistical Analysis

```
In [28]: d.describe()
```

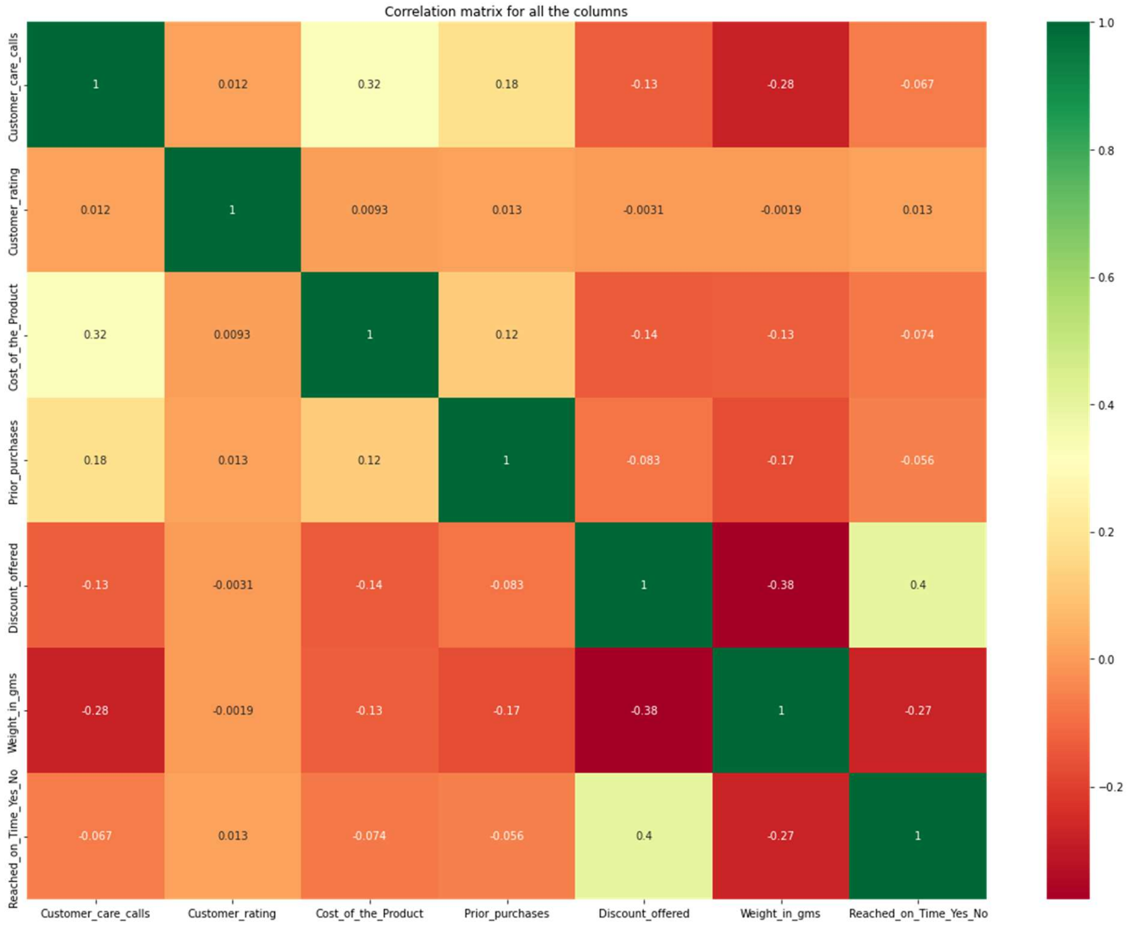
Out[28]:

	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Discount_offered	Weight_in_gms	Reached_on_Time_Yes_No
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

```
In [ ]:
```

Multivariate Analysis:

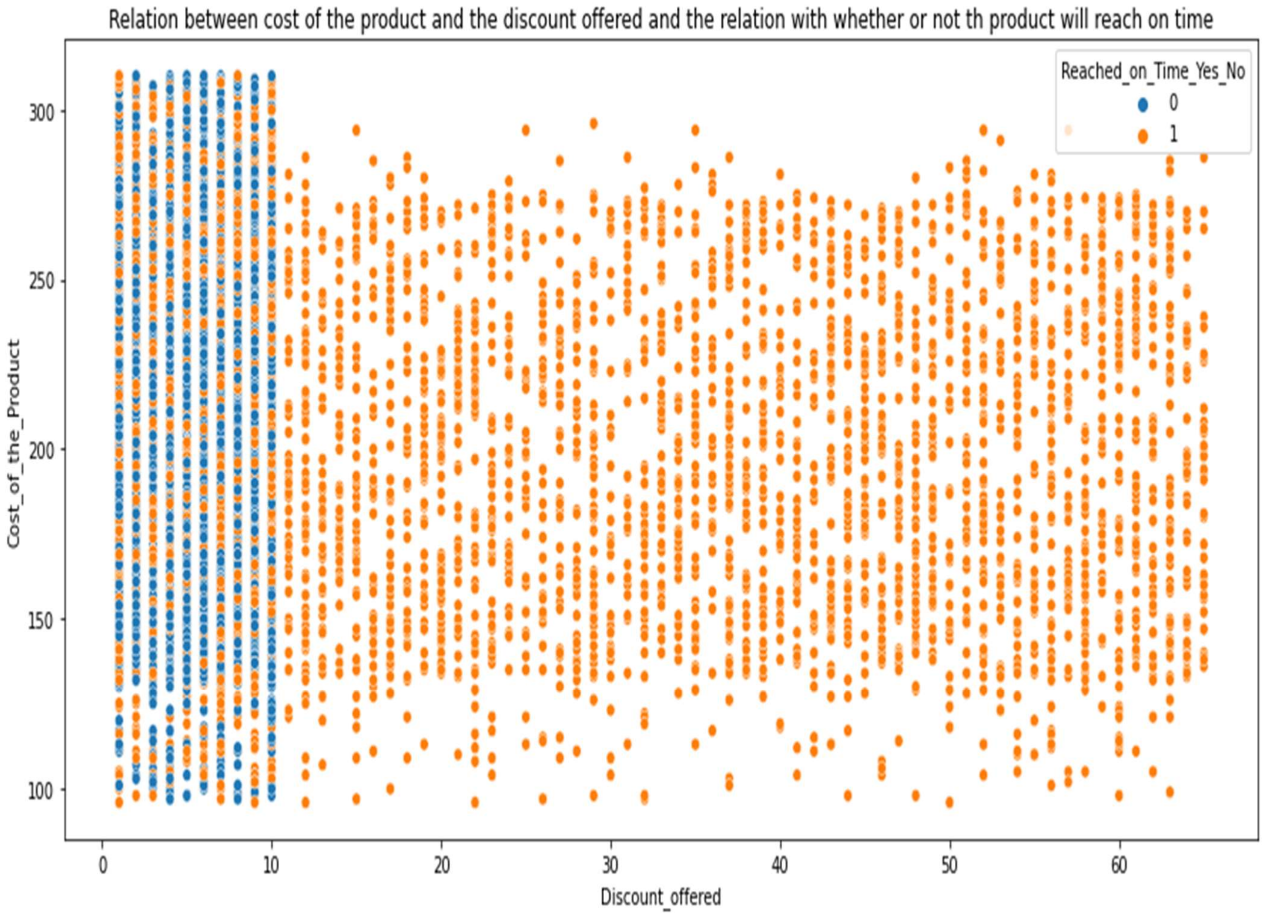
Correlation matrix for the columns in the data set



	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Discount_offered	Weight_in_gms	Reached_on_Time_Yes_No
Customer_care_calls	1	0.012	0.32	0.18	-0.13	-0.28	-0.067
Customer_rating	0.012	1	0.0093	0.013	-0.0031	-0.0019	0.013
Cost_of_the_Product	0.32	0.0093	1	0.12	-0.14	-0.13	-0.074
Prior_purchases	0.18	0.013	0.12	1	-0.083	-0.17	-0.056
Discount_offered	-0.13	-0.0031	-0.14	-0.083	1	-0.38	0.4
Weight_in_gms	-0.28	-0.0019	-0.13	-0.17	-0.38	1	-0.27
Reached_on_Time_Yes_No	-0.067	0.013	-0.074	-0.056	0.4	-0.27	1

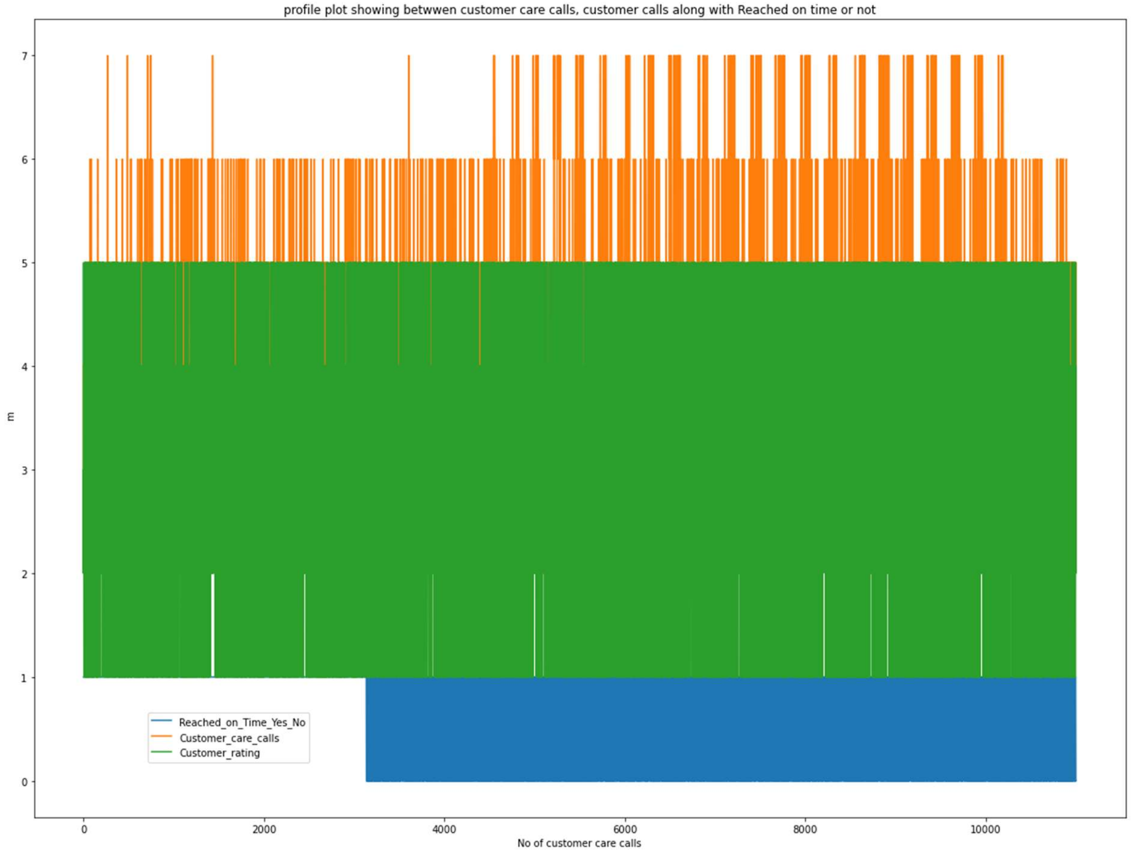
We can clearly depict that find that in most of the cases the correlation is almost equal to zero or negative. There is no cases where it showed positive correlation.

Scatter plot to see the relation between cost of the product and the discount offered and the relation whether the product reached on time

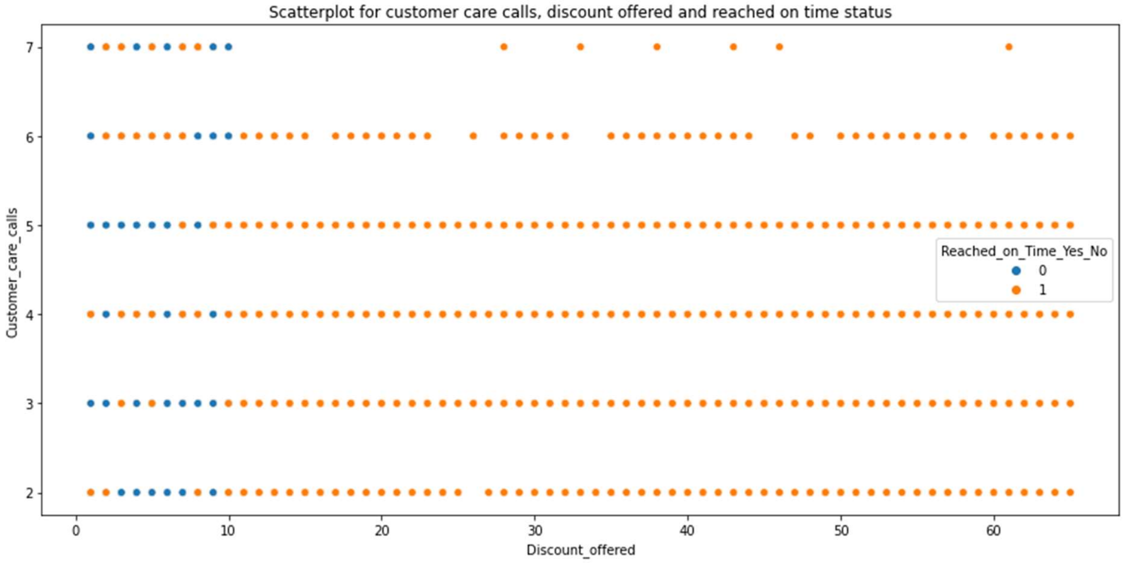


From the above graph we can conclude that discount offered between 0%-10% were only reached on time and rest of them were not reached time.

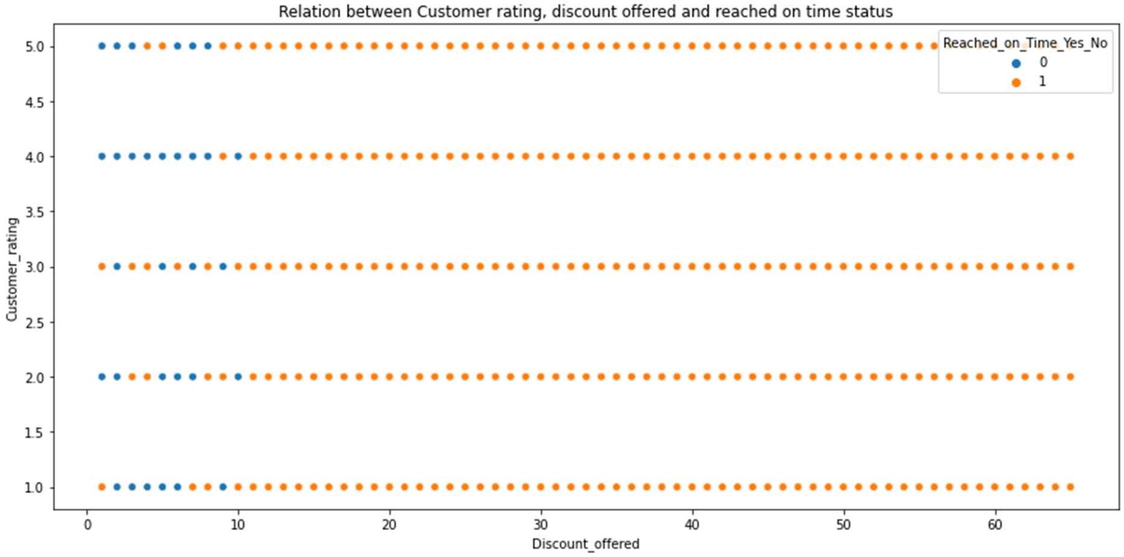
Profile plot:



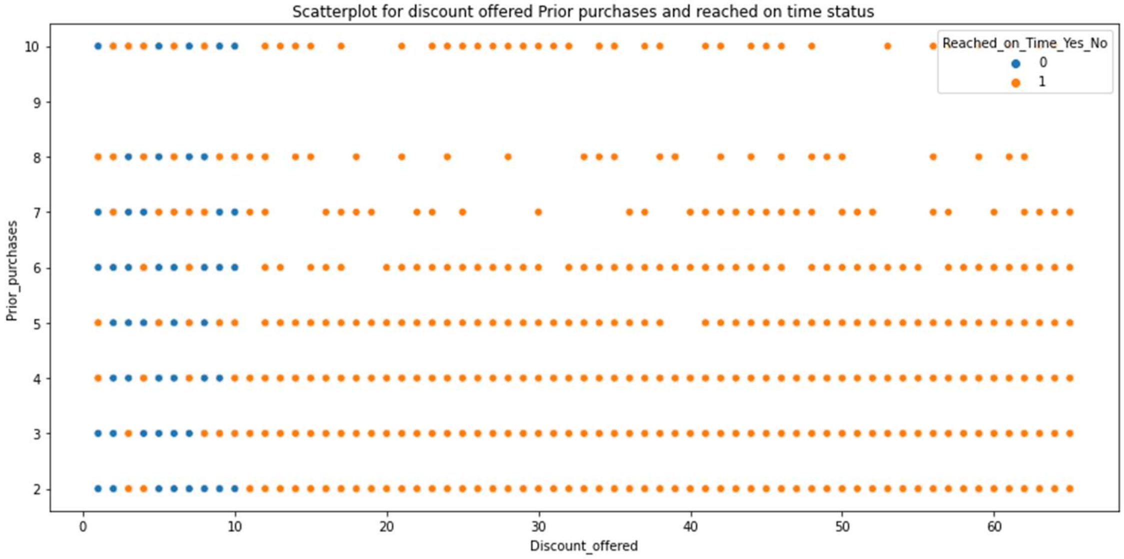
profile plot showing between customer care calls, customer calls along with Reached on time or not.



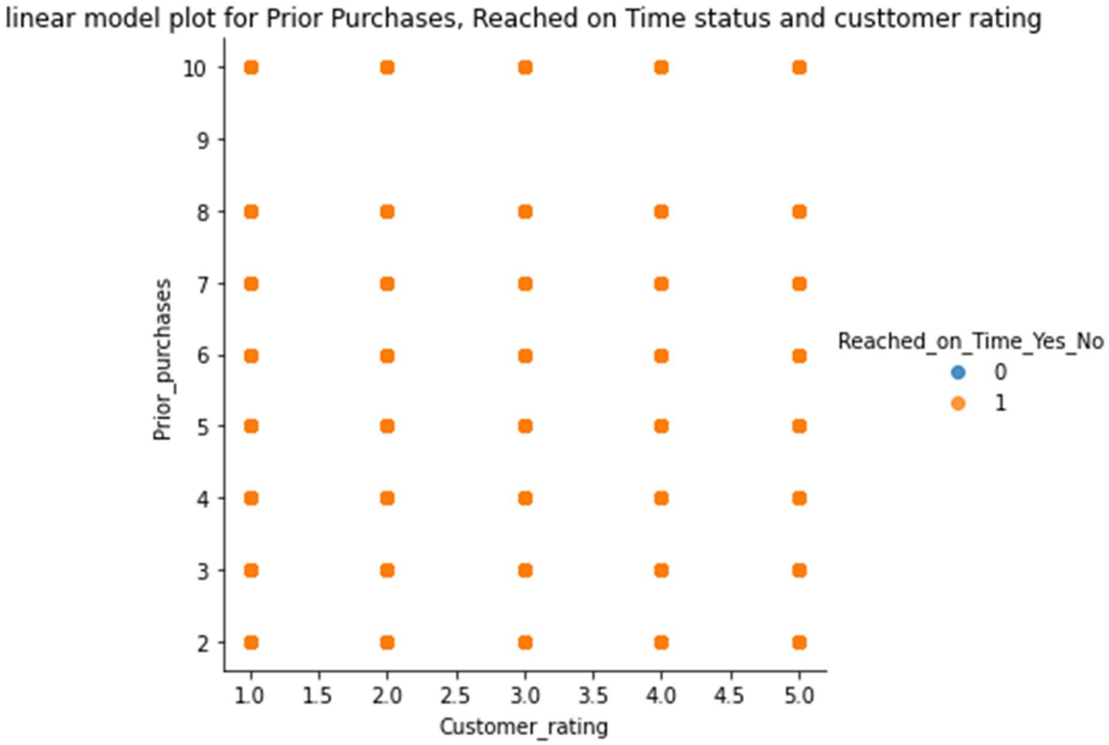
From the above scatterplot it is clear that discount offered between 0%-10% were only reached on time and rest of them were not reached time.



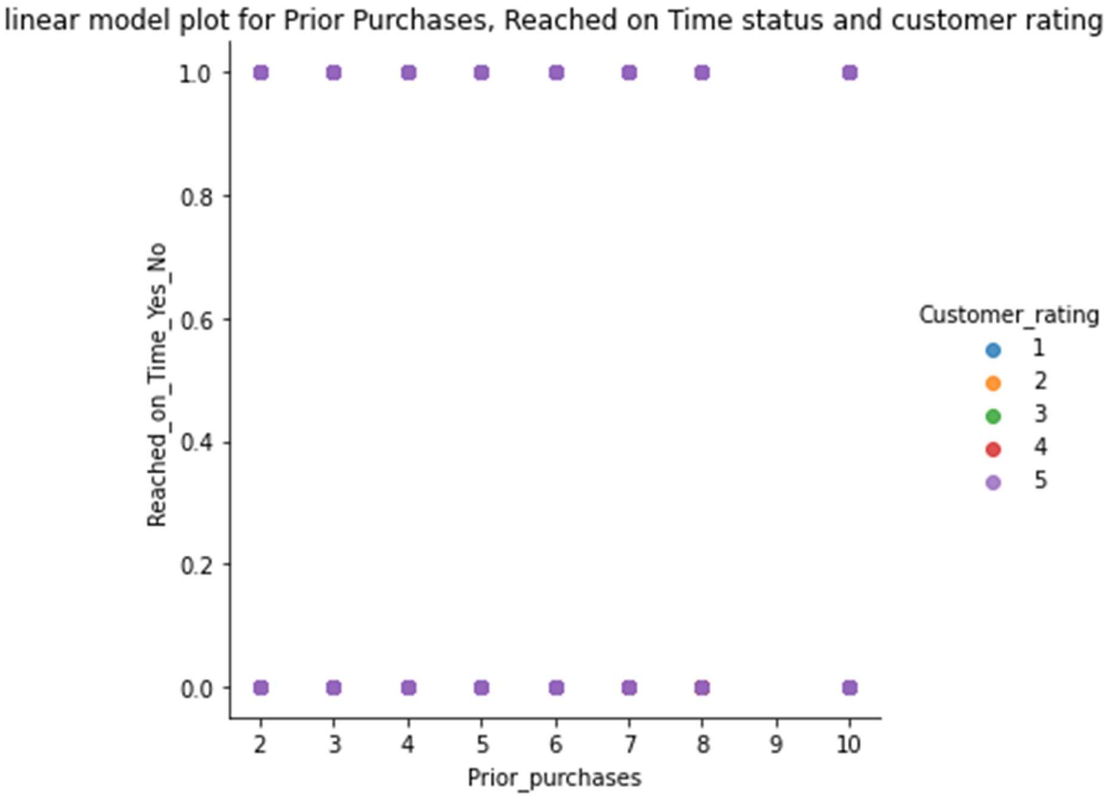
It is clear that the discount offer between 0 &10 where most of the products reached on time has customer rating as 4.0



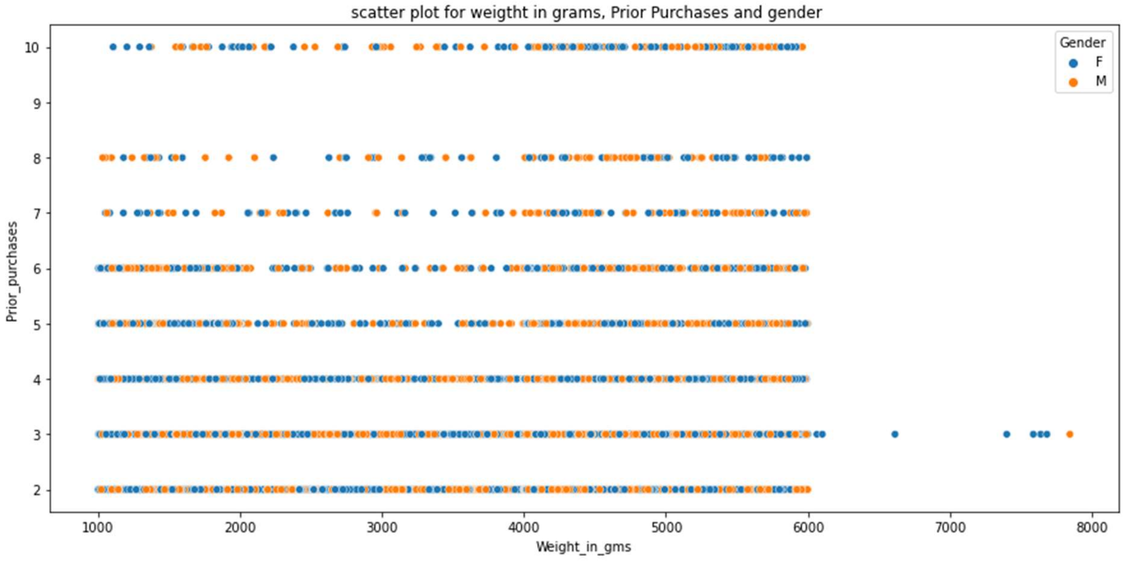
It is clear that there were zero number of prior purchases where prior purchases count is equla to 9.



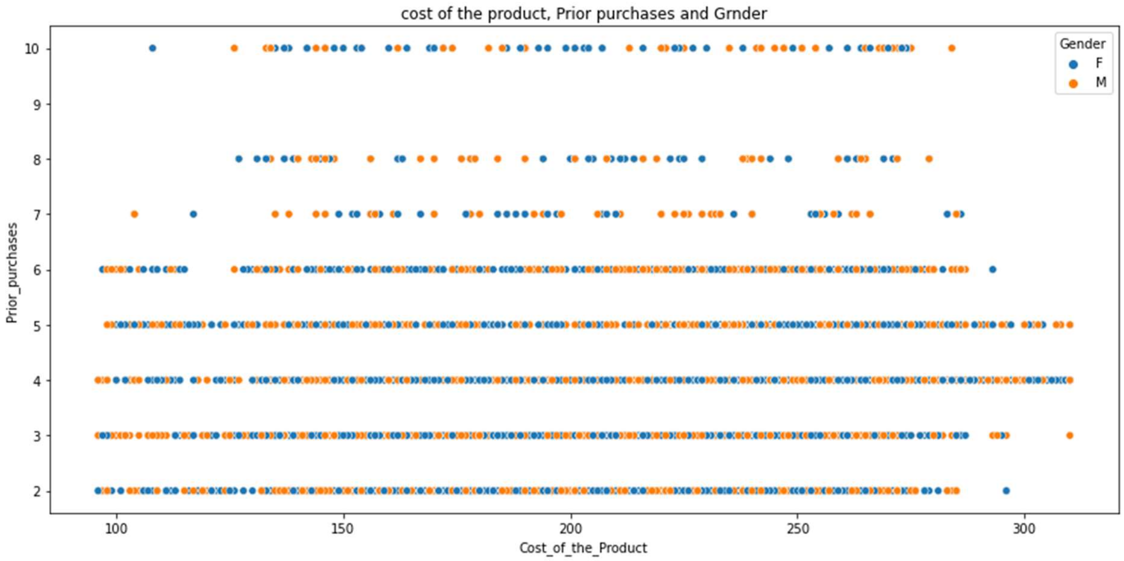
All the products which are were reached late irrespective of Prior purchases and customer rating.



Customer rating is 5 either it is reached on time or not reached.



scatter plot for weight in grams, Prior Purchases and gender



Scatterplot for cost of the product, Prior purchases and Gender.

Prediction Analysis:

The rating is low due to most of the products did not reached on time. So in order to increasing rating products must be reached on time. Prior purchases playing a crucial role most the people whose chosed shipping were actually the onces who ordered the products before. There are very few people who ordered for first time.

Conclusion:

Most of the shipping is done by “Ship” that is in water ways. About 68% of the shipments are done by ship and 16% of shipments are done by both roads and flights. The Count of female’s shipments are slightly higher than male’s shipments.

The count of the products which are not reached on time is also high. Most of the products ordered where low cost, and medium cost products. Only 8.62% of the high-cost products were shipped. Most of the products ordered were in between the range of $250 to $255 where as least no products ordered were around $100 to $110.The no of customer care call received are 4 with highest frequency. Highest prior purchases are 7 and lowest prior purchases are 3. Rating of “3” is given by most of the customers. A very few graphs showed linear regression.