FLIP ROBO

A Project Report on Black Friday Prediction

Submitted By:
Satu V. Pole

ACKNOWLEDGMENT

I would like to express my special gratitude to my SME Mr. Shwetank Mishra as well as “Flip Robo” team for letting me work for a project named “Black Friday Prediction” . Also thanks to my institute ‘Data Trained’ .

Also I would like to thank websites such as StackOverflow, geeksforgeeks and Youtube who has helped me in solving issues and errors.

[1] INTRODUCTION

{1.1} **Business Problem Framing.**

Large or most number of people buy any kind of products for themselves or others when different retaoil companies or brands provides heavy discounts on their products. In India many online retail websites such as Flipkart and Amazon have their shopping festivals such as Big billion sale or Amazon prime sale wherein these retail companies provide highest discounts during this time. Outside India, most retail companies have their shopping festival named as Black Friday Sale where they provide highest discounts possible on all their products. So nowdays using data science as a medium companies wants to know which products were sold last year in highest quantity and products that provided them with highest profit. This inturn is finding out customer behaviour I.e. how much customer is ready to spend on these flash sales.

A retail company “ABC Private Limited” wants to understand the customer purchase behaviour specifically, purchase amount against various products of different categories. They have shared purchase summary of various customers for selected high volume products from last month. The data set also contains customer demographics (age, gender, marital status, city_type, stay_in_current_city), product details (product_id and product category) and Total purchase_amount from last month.

{1.2} **Conceptual Background of the Domain Problem**

In this project using machine learning models The retail company wants maximize sales and revenue during Black Friday Sale as this shopping event presents a great opportunity for them to boost their sales, it also poses challenges such as managing inventory, pricing products effectively and predicting customer behaviour to monitor their purchases. This can predict purchase patterns of customers and their preferences and using then companies can make decisions on inventory management and pricing strategies that can lead to increase sales and revenue. This model will also help in choosing effective marketing strategies and channels and target the right customers with personalized promotions to drive engagement and loyalty. This will ultimately increase companies sales and help in deliver a better shopping experience to their customers during Black Friday and beyond.

{1.3} **Review of Literature**

The Black Friday sales event is one of the most significant shopping days in the retail industry. As a result, it has been studied extensively by researchers to gain insights into consumer behaviour and sales trends. The key findings from literature in Black Friday sale are:

→ Customer behaviour: Consumers tend to engage in more impulsive buying behaviour during Black Friday sales. They are also more likely to make purchases online then in store dues to convenience of shopping from home. Retailers use various pricing strategies including price bundling and discounts, to lure customers in buying more increasing sales during this event.

→ Inventory management: Black Friday sales require retailers to manage their inventory effectively to avoid stock-outs or overstocking. Accurate demand forecasting can help retailers manage their inventory better and optimize their sales.

→ Marketing and Advertising: Black Friday sales require effective marketing and advertising campaigns to attract customers. Social media is a powerful tool to reach to various customers and based on their past purchases effective marketing strategy can be applied so that customers lure in to buy products.

→ Customer Loyalty: Black Friday sales can be opportunity for retailers to improve customer loyalty by offering personalized promo codes and discounts as these customers are more likely to repeat purchases.

Overall, the literature on Black Friday sale highlights the importance of effective pricing, inventory management, marketing and customer loyalty strategies in driving sales during this event. Using this retailers can optimize their sales performance and improve the shopping experience for their customers.

{1.4} **Motivation for the Problem Undertaken**

The motivation behind our Black Friday sale project is to optimize retail organizations sales performances during the sale time. This event is excellent opportunity for businesses to increase revenue, attract new customers and improve customer loyalty. However, it also presents challenges such as managing inventory, pricing products effectively and predicting customer behaviour.

[2] Analytical Problem Framing

{2.1} **Mathematical/ Analytical Modeling of the Problem**

For this project we had to provide with Exploratory Data Analysis for dataset that was provided to us. This analysis will determine the trends shown bu various customers and what do they have purchased in past black Friday sales. Two dataset were provided to us namely test and train data. The exploratory analysis has been done on train dataset. The 2 columns from dataset contained null values which were treated and graphical representation/analysis of dataset was done. Further the dataset was treated if there were any skewness and outliers present in it and eventually treated.

{2.2} **Data Sources and their formats**

As mentioned above two datasets were provided to us namely train and test data. The analysis was performed on train dataset. The train dataset was a csv file which was uploaded using pandas library in Jupyter notebook. The dataset contained 12 columns in it whose details is provided below:

Column name	Column description
UserID	User ID of customer
ProductID	Product ID of every product bought by customer
Gender	Male or Female
Age	Age of customer in bins/groups
Occupation	Occupation (Masked)
City_Category	Category of the City (A,B,C)
Stay_In_Current_City_Years	Number of years stay in current city
Marital_Status	Marital Status of customer
Product_Category_1	Product Category (Masked)
Product_Category_2	Product may belongs to other category also (Masked)
Product_Category_3	Product may belongs to other category also (Masked)
Purchase	Amount (Target Variable)

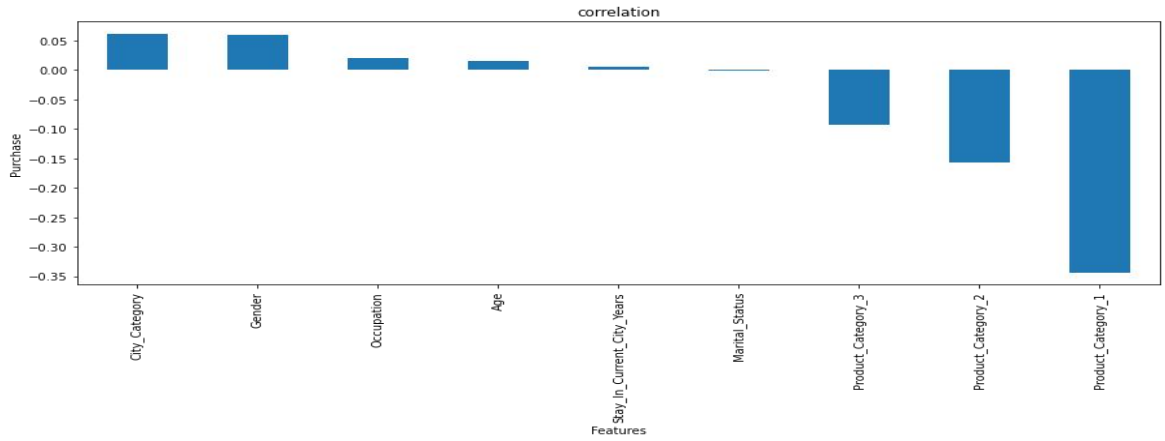
{2.3} **Data Preprocessing Done**

→ The train dataset was uploaded using pandas library and dataset was printed.
→ The shape of the dataset was 550068 rows and 12 columns.
→ Further the dataset was checked if there were any null values present and it occurred that 2 columns had null values present. Upon further analysis these null values were treated by filling them with median values of those particular columns. The columns that had null values present were Product_Category_2 and 3.
→ Further the dataset was checked if there are any duplicate values and it was known that there were no duplicates in dataset.

→ Further 2 columns that are UserID and ProductID were dropped from dataset as these columns contains unique values which are related to unique customer and unique product.
→ Further the dataset was described using graphical representation which showed relationships within columns. This graphs/plots included countplots, bar plots, line plots, heatmaps, distplots and boxplots.
→ Further the columns were encoded using Label Encoder method.
→ Further the dataset was checked for outliers by plotting box plots. It showed that 2 columns had outliers present. These were treated using z score method. The data loss after removing outliers was 6.03%.
→ Further the dataset was checked for skewness and 2 columns were skewed. This was treated by performing power transformer method on those 2 columns.

{2.4} **Data Inputs- Logic- Output Relationships**

The Purchase column or target variable is of continuous kind, hence it is a type of regression machine learning problem. Bar plots and line plots were used to devise a relationship between features and target columns.



The above figure shows that column 'Product_category_1' is highest negatively related to label while column 'City_Category' is highest positively related to label. There were no presence of multicollinearity seen within the dataset.

{2.5} **Hardware and Software Requirements and Tools Used**

Hardware used to complete the analysis and model building:
Model: ASUS TUF A15
Processor: AMD RYZEN 5 4600H OCTA CORE
RAM: 8GB
ROM:500 GB SSD

Software used to complete the analysis and model building is:

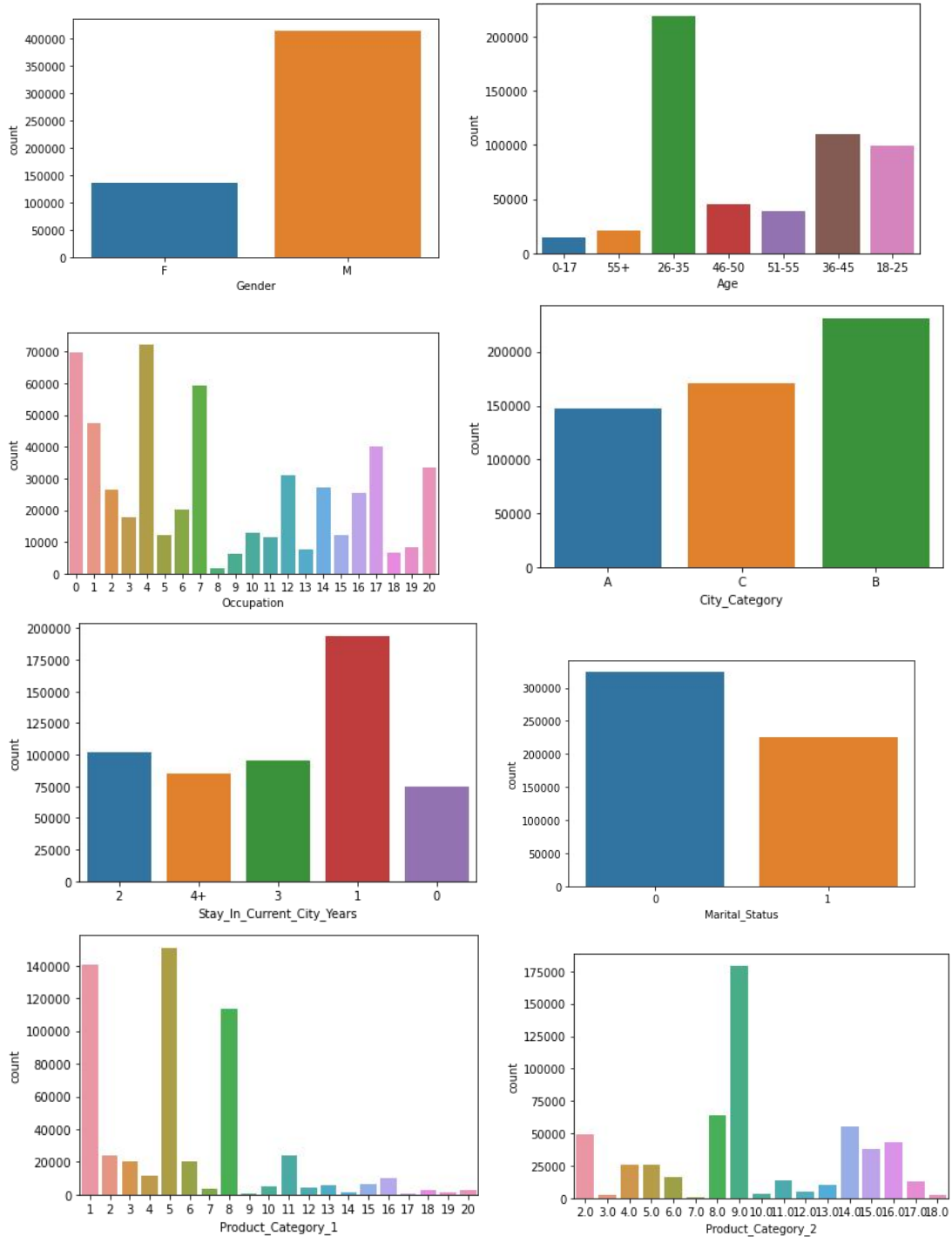
Jupyter Notebook (via Anaconda Navigator) - used for all the analysis

Libraries used:

→ Seaborn and matplotlib for visualization
→ Numpy
→ Pandas
→ Scipy.stats to import z score
→ Sklearn to import Power Transformer

[2] Visualization

Countplots



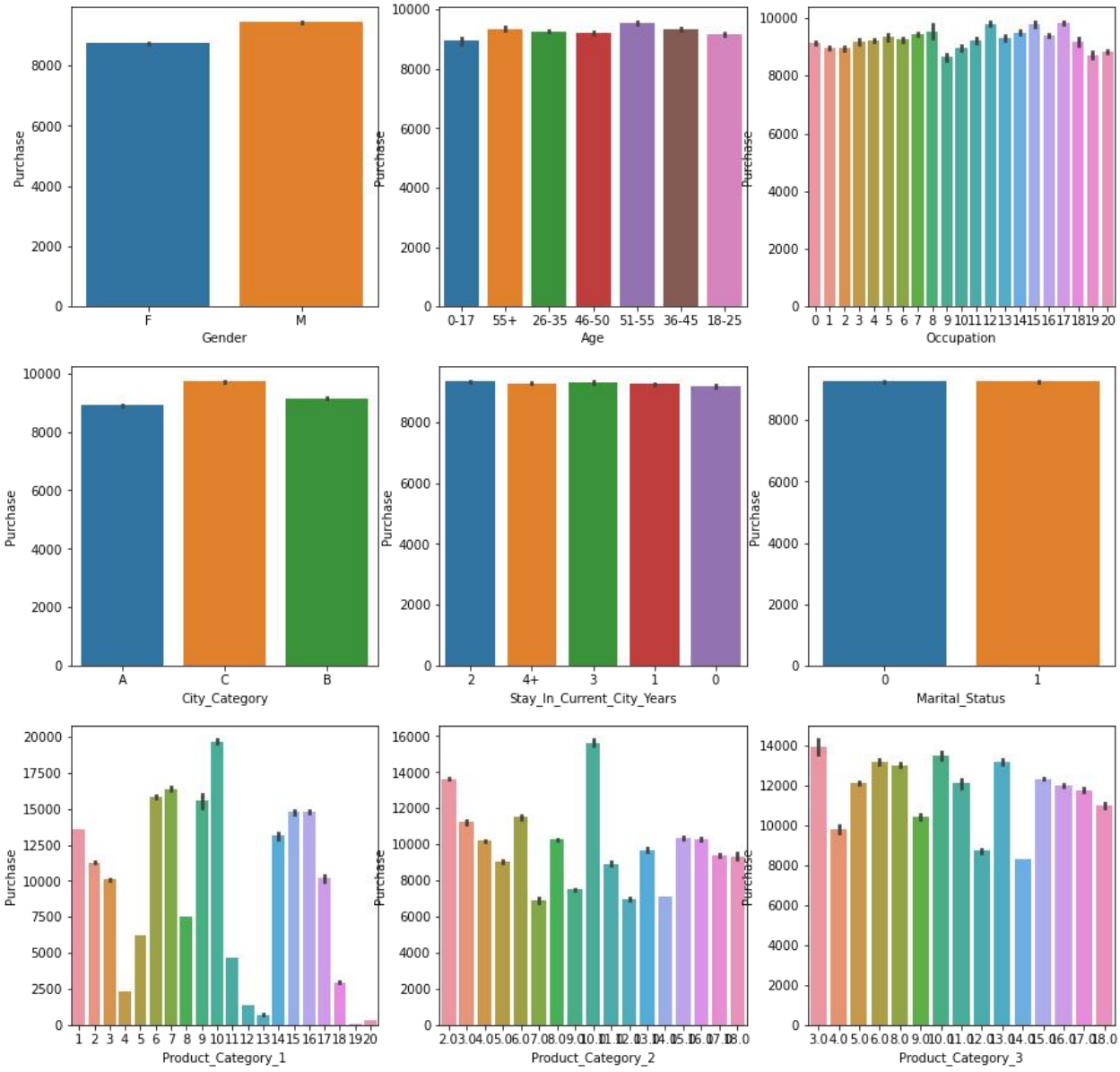
Observations for above plots:

→ Most number of customers are male while less than half of customers are female

→ The persons in age group of 26-35 are more compared to other age groups that participated in black friday sale.

→ The persons having occupation of value 0,4 and 7 are more compared to other occupations that participated in black friday sale
→ The customers staying in B city category are more compared to other city groups that participated in black friday sale.
→ Customers who are staying in current city for past 1 year are more compared to other persons that participated in black friday sale.
→ Most number of unmarried customers attended black friday sale.
→ In Product category 1, 1,5 and 8 categorized products were in large numbers compared to other products during black friday sale.
→ In Product category 2, 2,8 and 014 categorized products were in large numbers compared to other products during black friday sale.

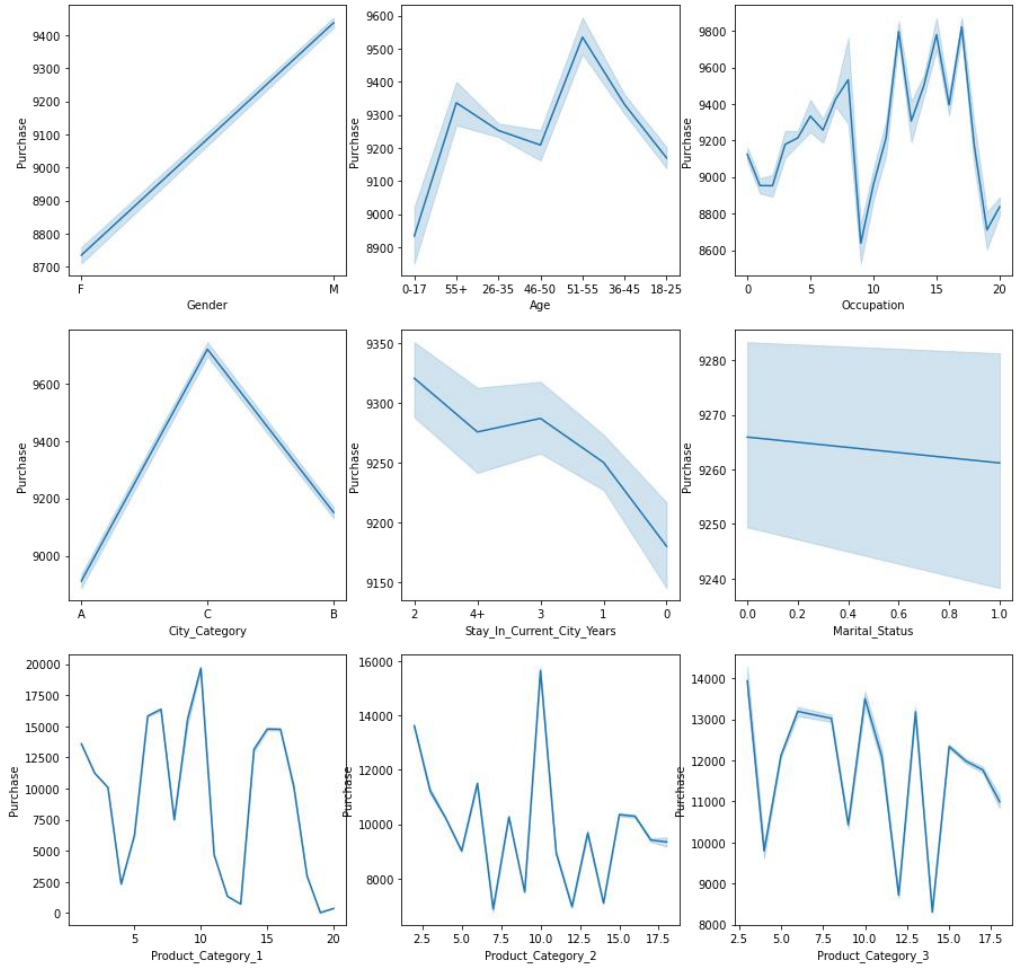
Barplots



Observations for above plots:

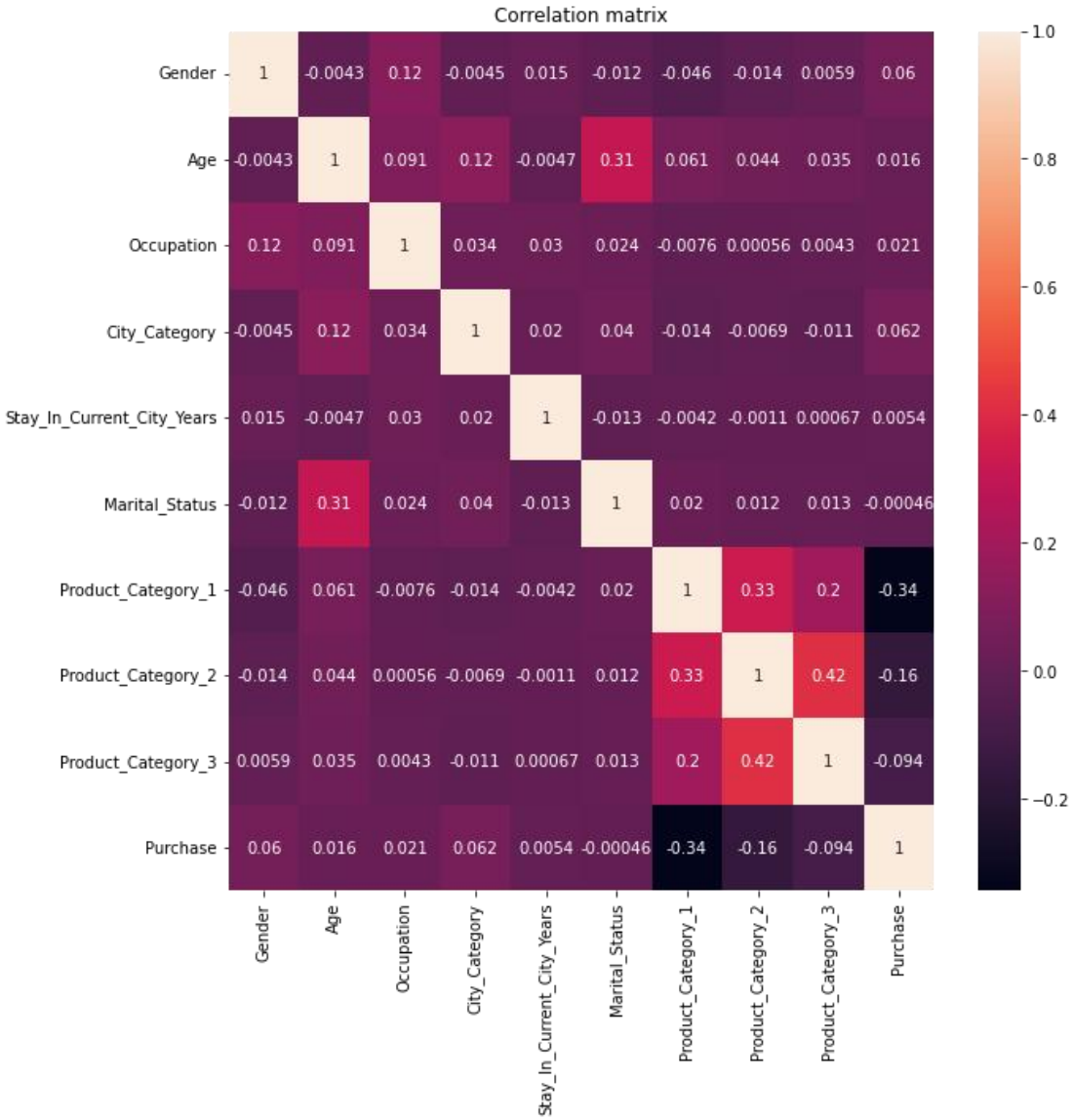
→ The males purchased more products then females during black friday sales.
→ Purchases for 0-17 age group were slightly lower compared to other age groups.Purchases are highest for 51-55 age group people.
→ Customers having occupation of 12, 15 and 17 purchased more compared to people from other occupations.
→ The persons from city category c have purchased more items for black froday then A and B category.
→ Almost equal number of purchases has been seen for people staying in current city for black friday sale.
→ Equal number of purchases has been seen in married as well as unmarried people for black friday sale.
→ In product category 1, products with masked code of 10 have purchased highest by people.
→ In product category 2, products with masked code of 10 have purchased highest by people compared to other products.
→ In product category 3, products with masked code of 3 and 10 have purchased highest by people.

Lineplots:



The observations for line plots are same as bar plots.These plots are plotted to show another way representing feature columns with respect to target columns.

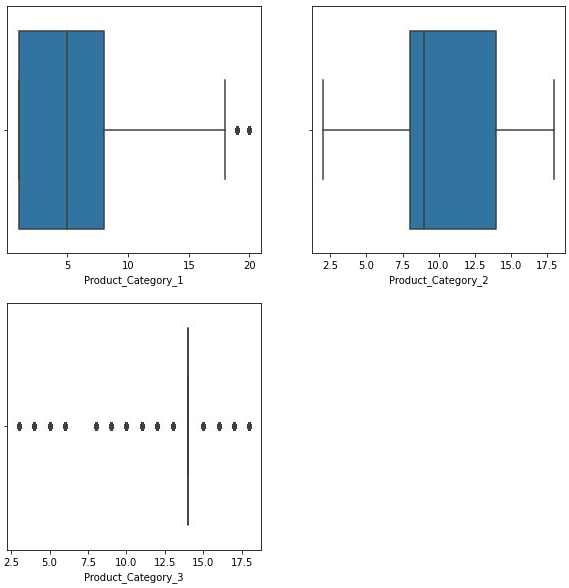
Heatmap:



Observations:

→ The above figure shows that no columns shows signs of multicollinearity.

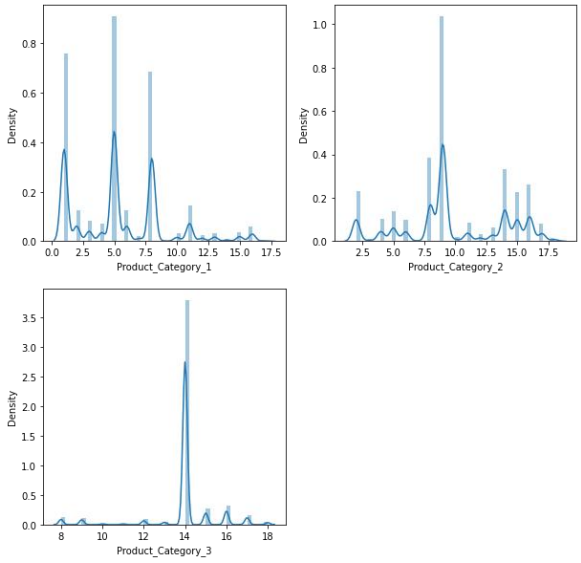
Boxplots:



Observations:

→ It can be seen from above plots that product category 1 and 3 columns have outliers present, which were further treated using z score method.

Distplots:



Observations:

→ It can be seen from above plots that product category 1 and 3 columns have skewness present, which were further treated using power transformer method.

Remarks/Conclusion:

This report does not have any conclusion as such as its machine learning model was not made. Hence how all the features mentioned above will affect the purchases during black Friday sale is not known to us. The observations as per graphical analysis/ visualizations are stated above.