EDA AND SALES PREDICTION ON BLACK FRIDAY SALES

TEAM NO: 5

Team Members:

Mehbooba C

Sooraj S

R Kannan Pillai

Naheeda Kallan

Mohamed Ishaq Kuttiyadan

ABSTRACT

Black Friday is an informal name for the day following Thanksgiving Day in United States. It is traditionally the busiest shopping day of the year and it is the ignition of one of the busiest shopping season in a year. Many stores offer highly promoted sales on this day because consumers are eager to spend so much money during this period, retailers seriously look forward to good preparation for the shopping holiday. From data science point of view, the ability to recognize and understand pattern from data using exploratory data analysis and sales prediction will help the retailers for their preparation on the same. The data set here is on Black Friday Sales for selected high volume of products of last season which contains customer demographics (age, gender, marital status, city type, stay in current city), product details (product id and product category) and total purchase amount. The purpose of this project is to understand customer behavior in Tableau and effectively predict how much a customer is probably to spend at a store based on historical purchasing patterns using python. If retailers comprehensively understand their customers in terms of characteristics, behaviors and motivations in the previous shopping seasons, they can implement and develop more effective marketing strategies for specific customer categories.

INDEX

CHAPTER 1 : INTRODUCTION

The largest shopping day of the year in America is the Friday following the Thanksgiving holiday. It is recognized as the ignition of one of the busiest shopping seasons in a year. It started back in the 1940s in the U.S. The day after Thanksgiving, retailers would offer huge discounts, and families would flock to these stores in order to pick up some bargains. This discount day has become to be known as “Black Friday”, named after the point in the year in which retailers moved out of debt, from the “red” to the “black”.

The issue with discounts are that price promotions largely stimulate temporary elasticity of demand in terms of purchases, but do little to improve elasticity of consumption. In simple terms, people will buy more and more people will buy in a given moment, but, most of the increase is represented by people who have simply time-shifted a purchase they would have made anyway. These are discounted sales, which mean low profit margins. Volume is the key if profit margins are low. But high margin, high volume is much, much better. So, consumers enjoy the discounts. The only thing that matters at the end of the day is the profit the retailer has made during the whole of the critical Christmas trading period. The future of Black Friday will rely on customer experience and the offerings driving demand. Black Friday sales are still increasing, which is promising for retail. Effectively utilizing this trend is key for retailers to ensure Black Friday is still a profitable event.

Because consumers are eager to spend so much money during this period, retailers seriously look forward to good preparation for the shopping holiday. In preparation for this day, retailers will typically hire more employees, stock their commodities, prepare new promotions, and decorate store layouts. Retailers rely on designing advertising campaigns to attract more customers into their stores and/or their online shops. In order to maximize their efforts and revenues, retailers enthusiastically understand how the consumers make shopping decisions that will assist them to achieve the most profits during the shopping season.

From the data science point of view, one of the most interesting applications of machine learning in the retail industry is to effectively predict how much a customer is probably to spend at a store based on historical purchasing patterns. If retailers comprehensively understand their customers in terms of characteristics, behaviors and motivations in the previous shopping seasons, they can implement and develop more effective marketing strategies for specific customer categories.

CHAPTER 2 : PROBLEM STATEMENT

This is a multiple or multivariate linear regression model which is used to predict the purchase amount a customer is expected to spend on Black Friday. This is a supervised machine learning problem since we are using available target values to train the model. Here we already know the target, how much a customer spend on a specific product. Hence the response is a continuous value. The ideal outcome is to provide retailers with information on how much is the expected purchase amount.

CHAPTER 3: DATASET DISCRIPTION

Data set link: https://www.kaggle.com/debasish05/black-friday-anallytics-vidhya (Train Data)

Dataset consist of 550068 observations about black Friday sales in a retail store. The set possesses 12 different features. There are 2 features of float type, 5 features of integer type and 5 features of object type (String). Only product categories 2 and 3 have missing values. The explanation of columns is as follows.

```
Data columns (total 12 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   User_ID                     550068 non-null  int64
 1   Product_ID                  550068 non-null  object
 2   Gender                      550068 non-null  object
 3   Age                         550068 non-null  object
 4   Occupation                  550068 non-null  int64
 5   City_Category               550068 non-null  object
 6   Stay_In_Current_City_Years  550068 non-null  object
 7   Marital_Status              550068 non-null  int64
 8   Product_Category_1          550068 non-null  int64
 9   Product_Category_2          376430 non-null  float64
 10  Product_Category_3          166821 non-null  float64
 11  Purchase                    550068 non-null  int64
dtypes: float64(2), int64(5), object(5)
```

3.1 User_ID

This column indicates ID of user and it is of data type integer. There are 5891 unique values in this column. This means for 550068 purchases, the distinct customers are 5891. Even though, it is an integer type, it's value should not be considered during analysis, as one value simply indicates a user. Count of this column gives how many unique customers are there for the retail store.

3.2 Product_ID

This column indicates ID for a particular product. Here the data type is object. There are only 3631 unique values which means no. of products considered in the data set is 3631. This column simply indicates which product a customer is purchasing and it is not considered for further analysis of the problem

3.3 Gender

This column describes whether customer is male (M) or female (F). Here data type is object and it is a categorical feature.

3.4 Age

Here age is object data type. This column indicates age group in which a customer belongs to. The age groups are 0-17, 18-25, 26-35, 36-45, 46-50, 51-55 and 55+.

3.5 Occupation

The occupation column contains integer data, where the occupation of the customer is encoded as integers from 0 to 20. Here occupation is a nominal data (categorical feature).

3.6 City_Category

The occupation column contains integer data, where the occupation of the customer is encoded as integers from 0 to 20. Here occupation is a nominal data.

3.7 Stay_In_Current_City_Years

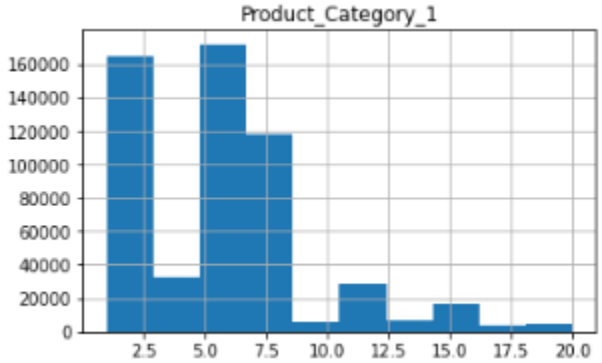
This column indicates the no. of years of stay in the current city of the customer. It contains object data type and there are 5 distinct values – 0, 1, 2, 3 and 4+.

3.8 Marital_Status

This column indicates whether the customer is married (1) or non-married (0). The values are integer data type with 2 unique values 0 and 1.

3.9 Product_Category_1

This column represents the category in which a product belongs to. There are 20 distinct values from 1 to 20. Each integer indicates a particular category. The data is already encoded here. From the frequency distribution, it is clear that the data is skewed. There is a chance of outliers in this case. The outliers in this case simply indicate that the particular product less sold.

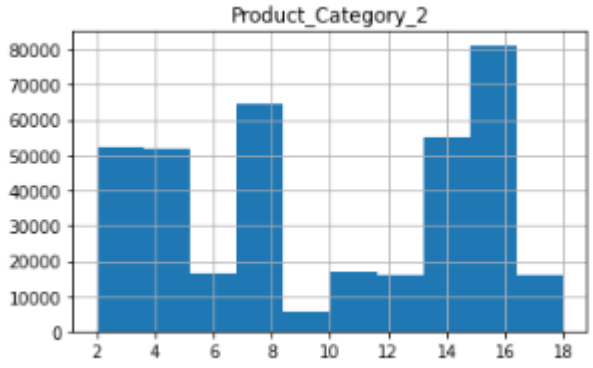


3.10 Product_Category_2

This column indicates a second product category in which a product belongs to. There are 17 distinct values (2-18), 31.6 % missing values in this column. Here the data type is float since the integer values are given with decimal points. This column can be removed from further analysis, since missing data is above 20% of total data.

```
# Checking why Product_Category_2 is showing float data type
data['Product_Category_2'].unique()
```

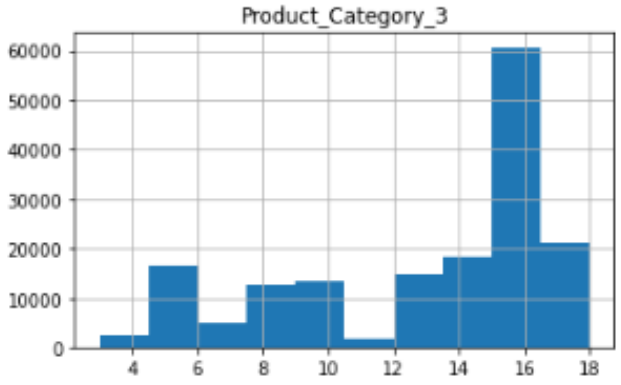
```
array([nan,  6., 14.,  2.,  8., 15., 16., 11.,  5.,  3.,  4., 12.,  9.,
       10., 17., 13.,  7., 18.])
```



Product_Category_1	0.000000
Product_Category_2	31.566643
Product_Category_3	69.672659

3.11 Product_Category_3

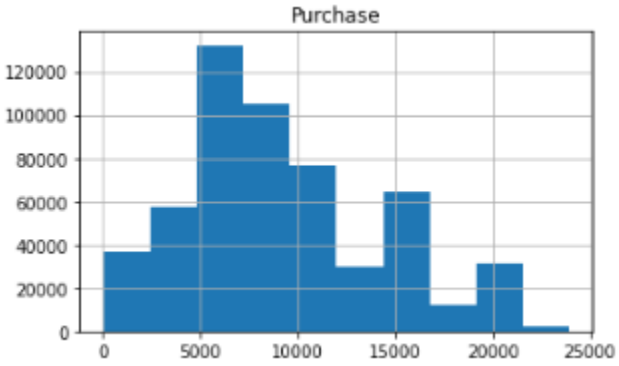
This column indicates a third product category in which a product belongs to. There are 15 distinct values (3-18), 69.7 % missing values in this column. Here the data type is float since the encoded integer values are given with decimal points. This column also can be removed from further analysis, since missing data is above 20% of total data.




```
array([nan, 14., 17.,  5.,  4., 16., 15.,  8.,  9., 13.,  6., 12.,  3.,
       18., 11., 10.])
```

3.12 Purchase

This column gives the purchase amount in dollar as an integer data. Most frequent sales are in the range 5000 – 7500 dollar. The total purchase amount in the last season was 5,095,812,742.00 dollar. Median of purchase amount is 8000 dollar approximately




```
# Checking total amount of purchase
data['Purchase'].sum()
```

```
5095812742
```

	User_ID	Occupation	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
count	5.500680e+05	550068.000000	550068.000000	550068.000000	376430.000000	166821.000000	550068.000000
mean	1.003029e+06	8.076707	0.409653	5.404270	9.842329	12.668243	9263.968713
std	1.727592e+03	6.522660	0.491770	3.936211	5.086590	4.125338	5023.065394
min	1.000001e+06	0.000000	0.000000	1.000000	2.000000	3.000000	12.000000
25%	1.001516e+06	2.000000	0.000000	1.000000	5.000000	9.000000	5823.000000
50%	1.003077e+06	7.000000	0.000000	5.000000	9.000000	14.000000	8047.000000
75%	1.004478e+06	14.000000	1.000000	8.000000	15.000000	16.000000	12054.000000
max	1.006040e+06	20.000000	1.000000	20.000000	18.000000	18.000000	23961.000000

CHAPTER 4: PROPOSED METHOD

The proposed system has a data which has customer demographics, product information and purchase amount. This proposed system can use this data to understand the customer behavior and predict purchase amount as per the features.

4.1 UNDERSTAND DATA

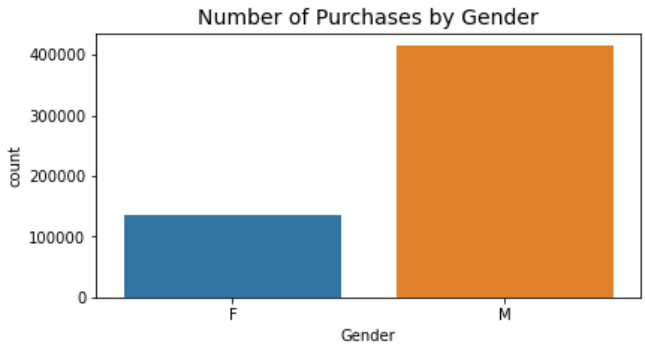
From the data we are having, it should be converted in to desired format for creating regression model to predict purchase amount. For this, first we need to understand the data fully. That is we need to address the questions: what each column represents, what is data type, is there any problem in data type,

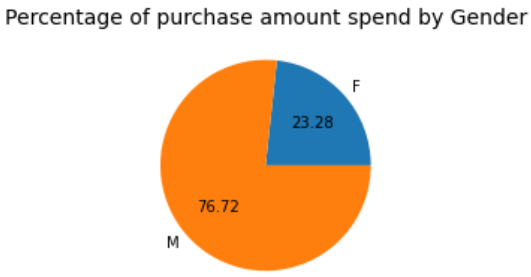
4.2 EXPOLORATORY DATA ANALYSIS

In statistics, exploratory data analysis is the process of summarizing the main characteristics of data set to understand what the data can tell us beyond the formal modelling task. With EDA, we can do univariate analysis for understanding the distribution of each features and bivariate analysis is used for understanding the features relation with target variable.

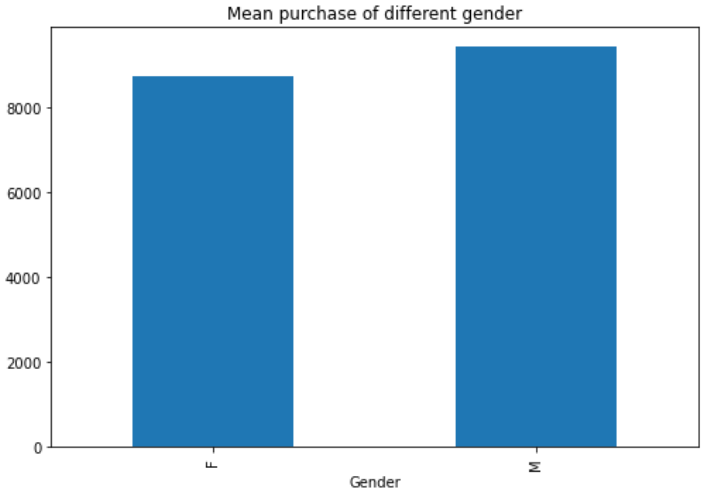
A count plot can be used to understand the density of underlying distribution of a single numerical or categorical data. Also relation with target(purchase amount) can be found using a bar plot or pie plot by using group by method.

4.2.1 Age column



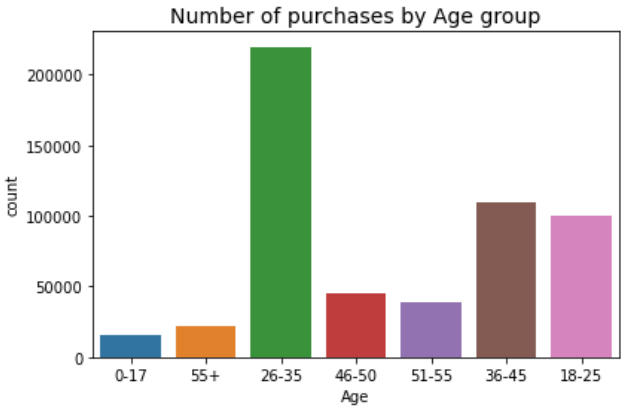


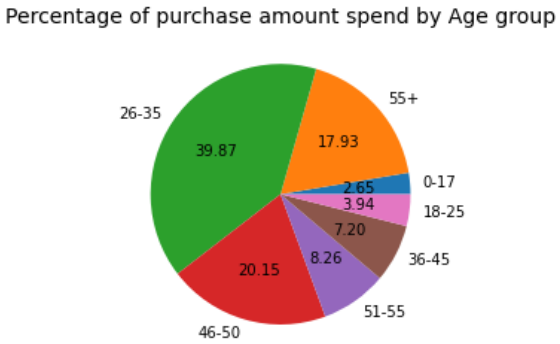
The graph shows that there are almost 3 times more male customers than female customers. Maybe male visitors are more likely to go out and buy something for their ladies when more deals are present. The purchase amount contributed by male customers is 76%.



From this graph mean purchase amount of males is larger than females.

4.2.2 Gender Column

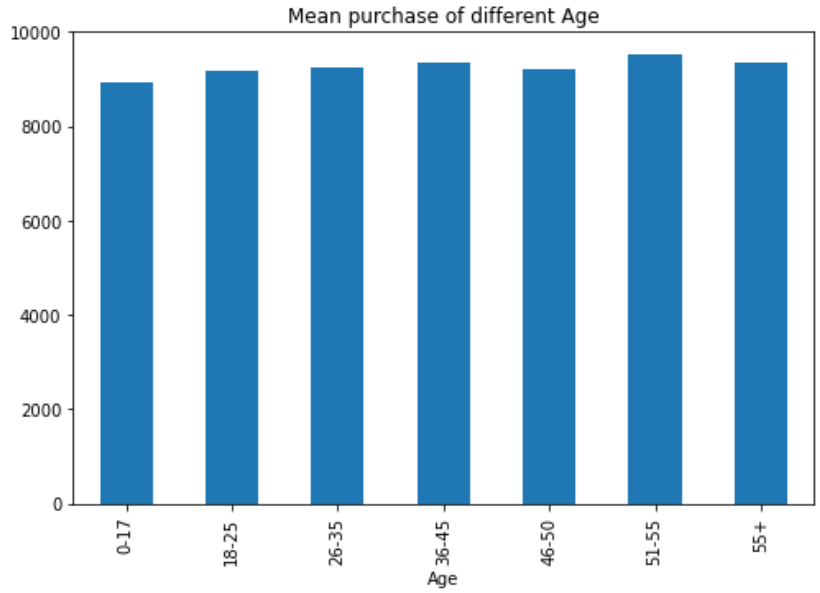




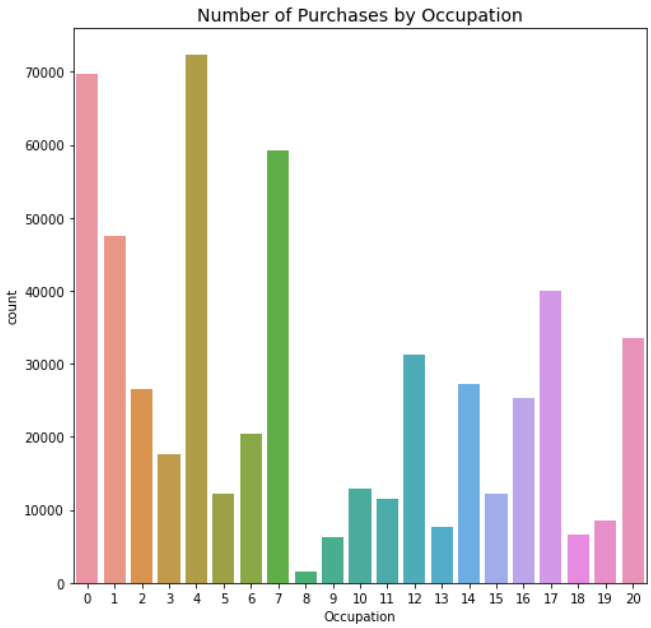
From the figure above, we can easily conclude that the highest number of customers belong to the age group between 26 and 35. Also more purchase amount is contributed by the same age group. Even though, the frequency of visit of 46-50 and 55+ age group is less, they contribute 20% and 18% of the total purchase amount

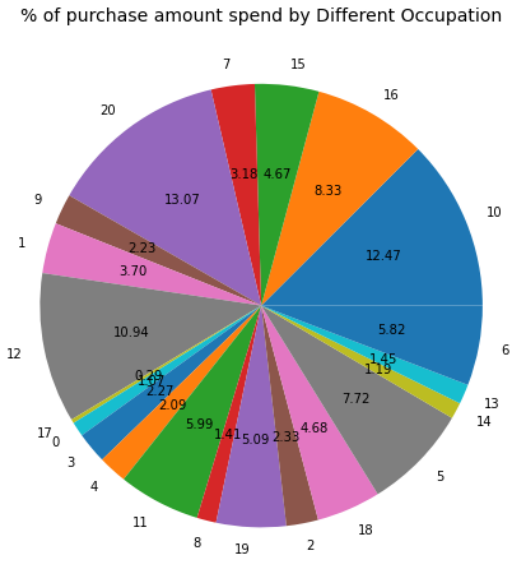
Based on these results, the retail store should sell most of the products that target people in their late twenties to early thirties. To increase profits, the number of products targeting people around their thirties can be increased.

From the mean of purchase amount by different age group graph, it is clear that all age group have near mean values. From this it is clear that, the age group 46-50 and 55+ spend more in a single purchase.

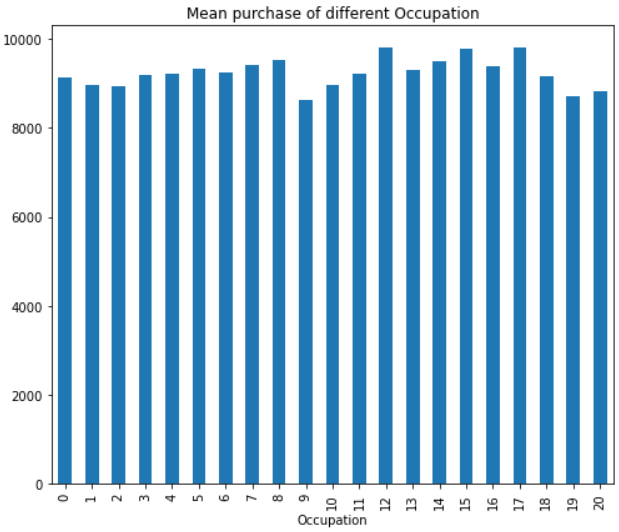


4.2.3 Occupation Column

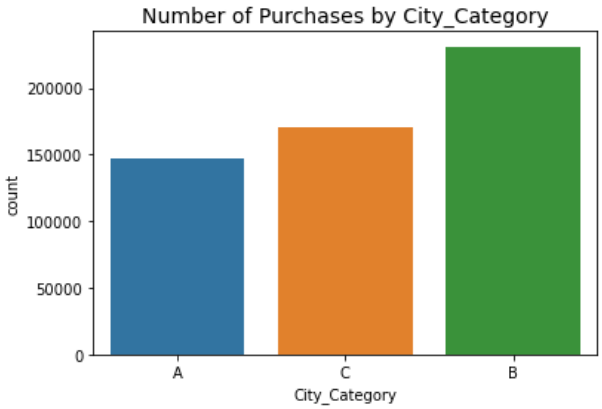


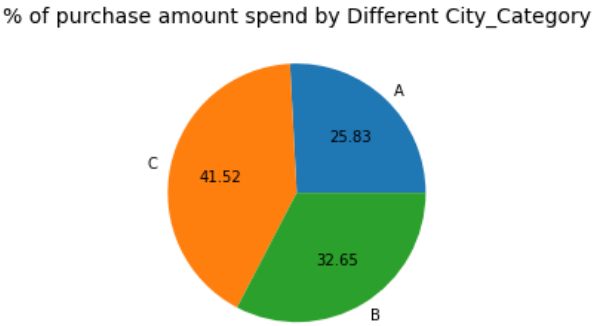


The customers belonging to occupation category 4 are more frequent visitors and customers belonging to occupation category 20,12 and 10 spends more. The mean purchase of different occupation are in the range 8000 to 10000.



4.2.4 City Category Column



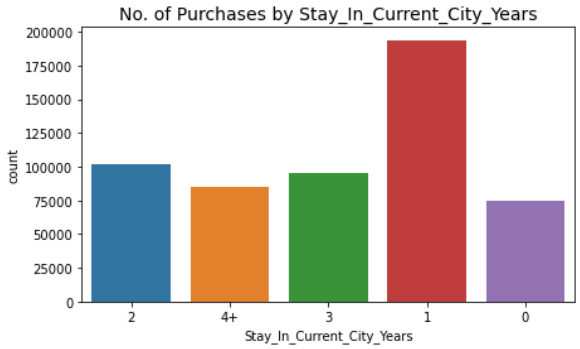


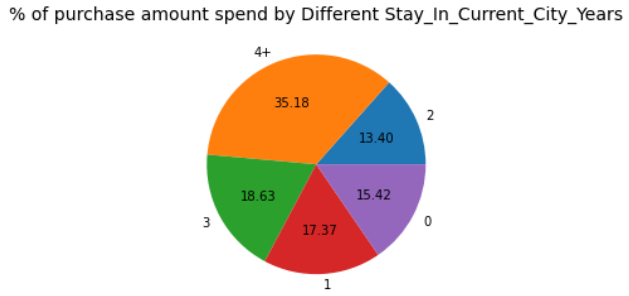
It is evident from the pie chart that all the three cities are almost equally represented in the retail store during Black Fridays. Maybe the store is somewhere between these three cities, is easily accessible and has good road connections from these cities.

Customers from C_category cities make up more than half of our black friday sales even though, most frequent customers are from B_type city. On the contrary, we didn't get very many customers from A_type city and they spent the least in our store. This can be noted when making future marketing plans.

4.2.5 Stay in Current city

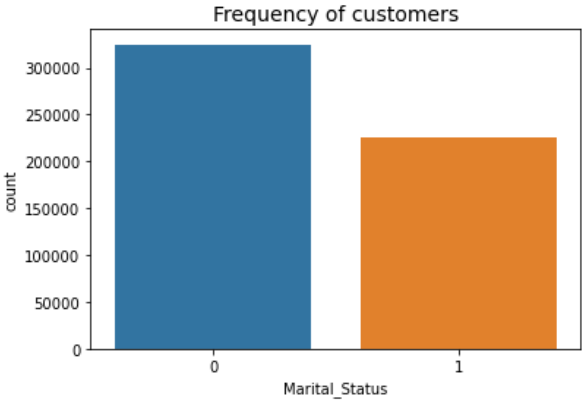
From the graph, the frequent customers in the store new residents (one year or less). That is store apperas to be popular among new residents. People who have been living in current city for longer spend a bit more than new comers. Since they chose to stay with the store, we do need to find out what kept them loyal so that better plans can be made to keep more customers instead of losing them over time.

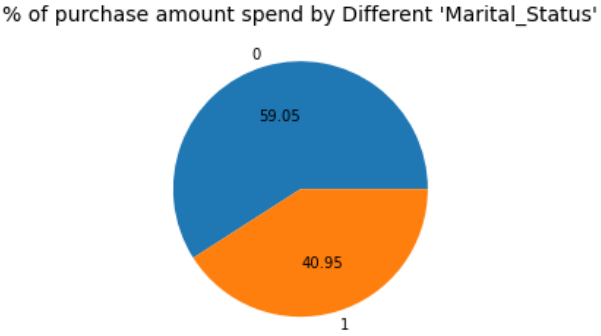




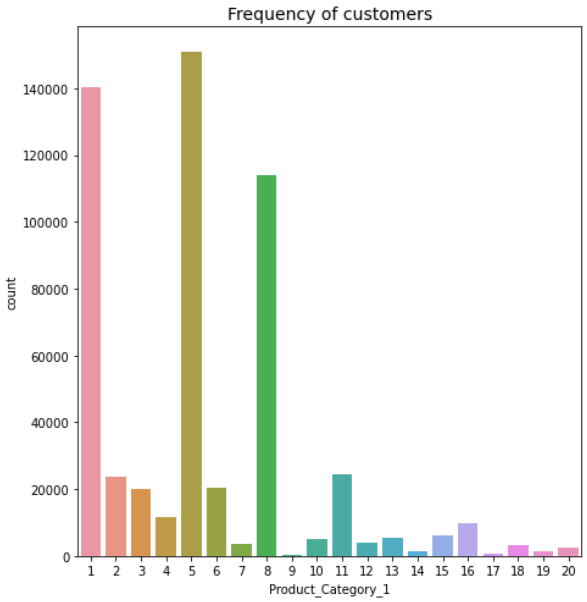
4.2.7 Marital Status

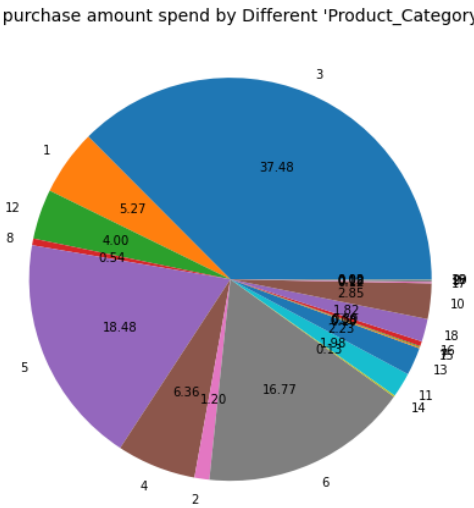
As expected there are more single people buying products on Black Friday than married people, also they spend more than married people.

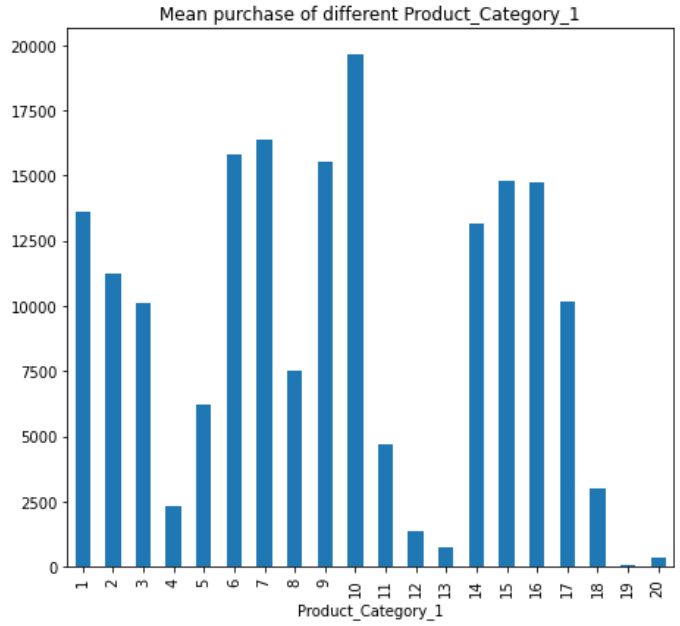




4.2.8 Product Category 1 column

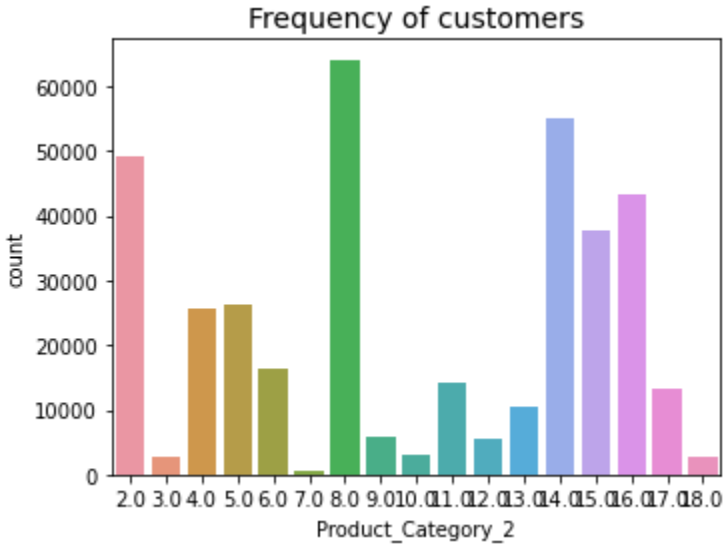






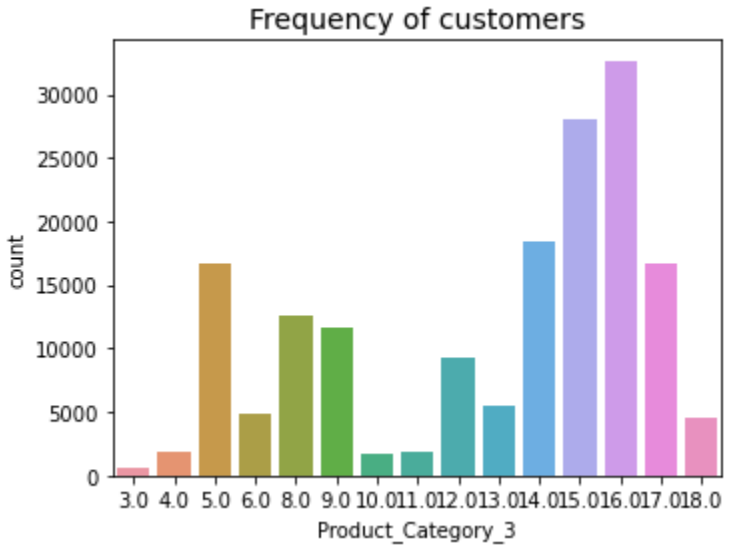
From the above graphs most selling product category 1 is 5th and 3rd category is contributing 40% on total purchase amount. The product category with highest purchasing amount is 10th. Even though 5th & 8th category is selling more, its mean value is less.

4.2.9 Product Category 2 column



From the above graph the most selling product category 2 is 8th category.

4.2.10 Product Category 3 column



In the product category 3 most selling is 16th category

4.3 DATA PRE-PROCESSING

Data preprocessing is the method of converting raw data to make it suitable for machine learning model. Within preprocessing, the main steps are handling missing values, feature engineering and dimensionality reduction, handling outliers, encoding categorical features and checking correlation.

4.3.1 Handling missing values

Presence of missing values in the dataset, will affect our ML model. Therefore, it is necessary to check and handle those values.

Here the columns Product_Category_2 and Product_Category_3 have missing values. The missing value percentages of these columns are above 20% of data. If we fill these values with test static, more than 70% data will be artificial. Also, for every single product it is non-realistic to have a second and third product category. Therefore, we can drop these columns.

```
# Checking for missing values
data.isna().sum()
```

```
User_ID                          0
Product_ID                       0
Gender                           0
Age                              0
Occupation                       0
City_Category                    0
Stay_In_Current_City_Years       0
Marital_Status                   0
Product_Category_1               0
Product_Category_2          173638
Product_Category_3          383247
Purchase                         0
dtype: int64
```

	percent_missing
User_ID	0.000000
Product_ID	0.000000
Gender	0.000000
Age	0.000000
Occupation	0.000000
City_Category	0.000000
Stay_In_Current_City_Years	0.000000
Marital_Status	0.000000
Product_Category_1	0.000000
Product_Category_2	31.566643
Product_Category_3	69.672659
Purchase	0.000000

4.3.2 Encoding Categorical Features

For building ML model all features must be numerical. Therefore, categorical columns must be encoded to numerical values. There are 4 object data type columns now.

```
User_ID                      int64
Product_ID                  object
Gender                      object
Age                         object
Occupation                   int64
City_Category               object
Stay_In_Current_City_Years  object
Marital_Status               int64
Product_Category_1           int64
Purchase                     int64
```

Here we can use either one hot encoding or label encoding. One hot encoding in this case will result in more no. of columns. Therefore, label encoding is used in this case. The output of label encoding is as follows.

```
Gender : [0 1]
Age : [0 6 2 4 5 3 1]
City_Category : [0 2 1]
Stay_In_Current_City_Years : [2 4 3 1 0]
```

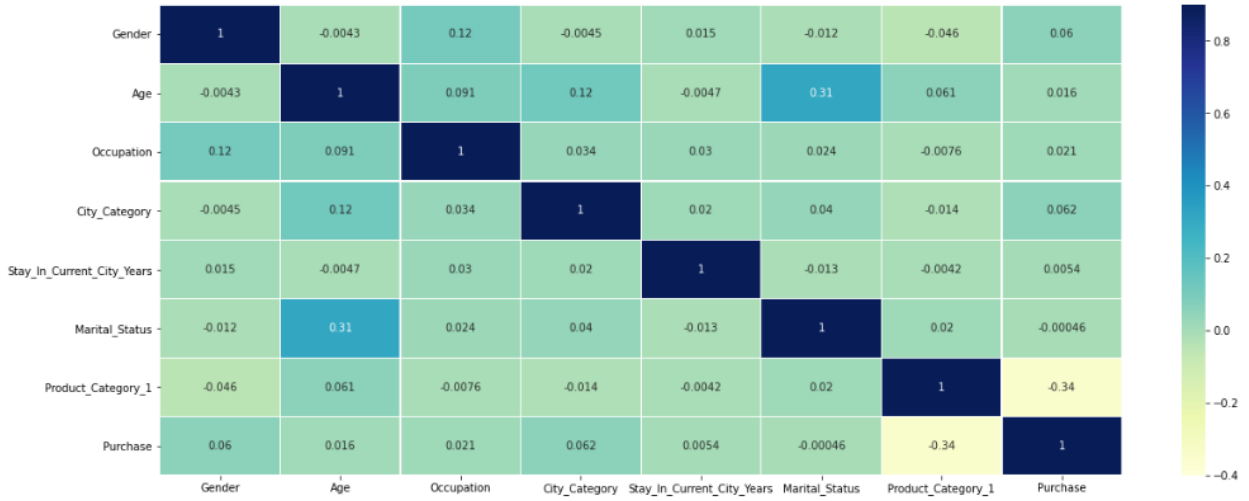
4.3.3 Dimensionality Reduction

The lesser the no.of columns, the better will be ML model. Dimensionality reduction is done by removing unnecessary columns. Also checking correlation will help in the reduction of

unnecessary columns. Here encoding is done prior to correlation checking so that those columns are also there in the correlation heat map.

While checking unnecessary columns, the User_ID and Product_ID columns have more unique values. These columns simply represent a particular user and product respectively. There are 5891 distinct customers and 3631 distinct products. Therefore, these 2 columns can be removed from further analysis.

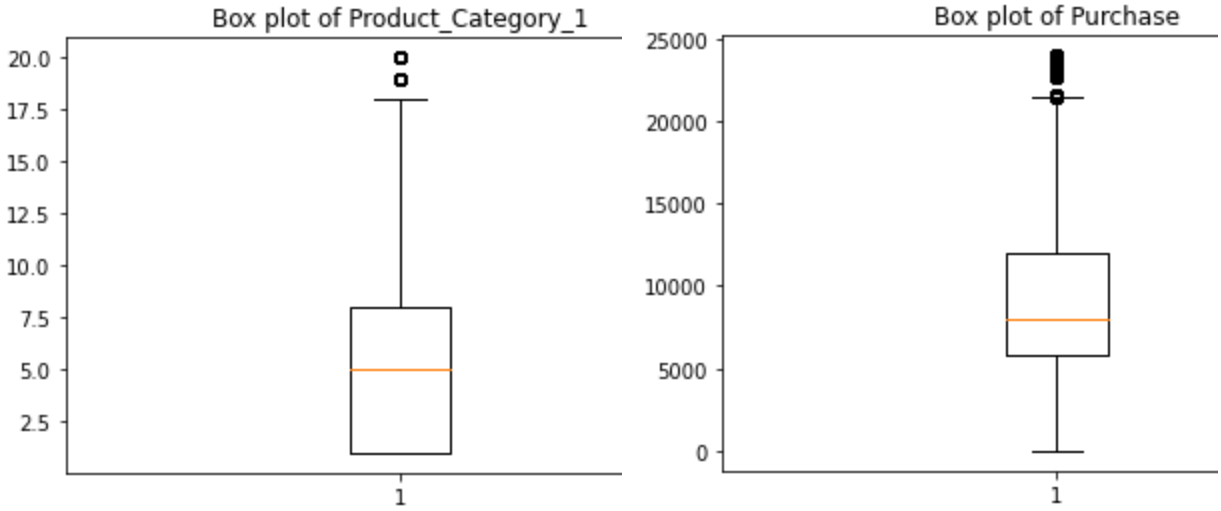
After this process correlation check is done to know whether we can remove any columns based on correlation.



Here marital status and age have a low positive correlation of 0.32. Also age and city category have low positive correlation. All other columns have negligible correlation. With the target column, the more correlated feature is Product category 1. Unfortunately, there is no single feature that shows strong correlation with purchase directly, so we can understand this as that purchase depends on the ensemble of all features.

4.3.3 Handling Ouliers

Outliers in dataset will affect our ML model. To find outliers in the dataset, boxplot is used. Here only numerical features are checked since categorical columns already have only specified no. of values.



Only Product_category 1 and Purchase column have outliers. After removing outliers using IQR method, distribution plot is checked agin for each feature. There is no significant change detected. Also correlation matrix heat map is plotted. Even in the heat map, no signifant change is there.

4.6 MODELING

Here the problem is a multivariate linear regression. The target feature is purchase amount and it is a continuous value and it's value is available in the data set. Therefore, here a supervised machine learning model is used to predict the purchase amount.The regression models used here are linear regression model, decision tree regression model, random forest regression model and XGB regression model.

The evaluation parameters used to evaluate the models are Mean squared error (MSE) and R squared value. The mean squared error tells how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the "errors") and squaring them. The squaring is necessary to remove any negative signs. It also gives more weight to larger differences. There is no correct value for MSE. Simply put, the lower the value the better and 0 means the model is perfect. The lower the MSE value, the better the model. R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model. R-squared is always between 0 and 1: 0 indicates that the model explains none of the variability of the response data around its mean and 1 indicates that the model explains all the variability of the response data around its mean. In general, the higher the R-squared, the better the model fits your data.

Before creating the model, first data set is divided in to features (X) and target (y). After defining features and target, the data is divided to train set and test set to train and test the model respectively. Then each model is created and checked the MSE and R-squared value. The obtained values of MSE and R-squared value for various machine learning regression models is as follows:

Model	**MSE**	**R-squared value**
Linear Regressor	4579.4394	0.1240603
Decision Tree Regressor	2953.5882	0.6356242
Random Forest Regressor	2926.03191	0.642391
XGB Regressor	2941.18112	0.6386791

From the table, it is clear that the best model is Random Forest Regression model since this model have highest R-squared value and low MSE.

CHAPTER 5 : CONCLUSION

In this project, we used a machine learning algorithm to predict the amount that a customer is likely to spend on Black Friday. We also performed exploratory data analysis to find interesting trends from the dataset. Here we have developed a model with Random Forest with high accuracy to predict the sales for coming year. From that analysis we can observe that It didn't matter if the group was male, female, young, old, married or unmarried, the median purchase by the customers hovered around $8000. However, some groups were more present than others. Males shopped more than females. The marital status 0 shopped more than the marital status 1. Also customers between the ages of 18 and 45 shopped the most. Finally, the models were tested to find the model that makes the best predictions. When analyzing the data from the testing model, it is revealed that Random Forest with high accuracy to predict the sales for coming year. This model produced an RMSE of 2926.03191and accuracy of 64.2%. Therefore, it was the model used to make the final predictions for the testing dataset. Retailers can use this model to predict their sales.