**Black Friday Data Analysis**

**Big Data project**

**CMP-4TH YEAR**

**Team 10**

**Team members:**

**Ahmed Salama Hamed**

**Mohamed Talaat Mohamed**

**Moamen Hassan Attia**

**Ahmed Maher**

**Problem Description**

Black Friday is the busiest shopping day of the year, many stores offer highly promoted sales on that day and open very early, such as at midnight, It’s a very great chance for customers to buy many products such as clothes, electronic-devices and even food for a very little amount of money because all shops make huge discounts, All shop owners start to prepare for that day buy analyzing their customers’ purchasing behavior so that they can find strategies to promote their sales and try to find answers to many questions like Who is more likely to spend more in a black Friday sale? And Which type of products are common among men and which among women?

In this project we try to analyze customers behavior on black Fridays and try to find solutions to many questions each shop owner might have by giving insights and visualizations and building predictive models.

**Project Pipeline**

****

**Getting and Cleaning Data**

Our dataset is divided into csv files:

1. Train.csv

550068 rows 12 columns

1. Test.csv

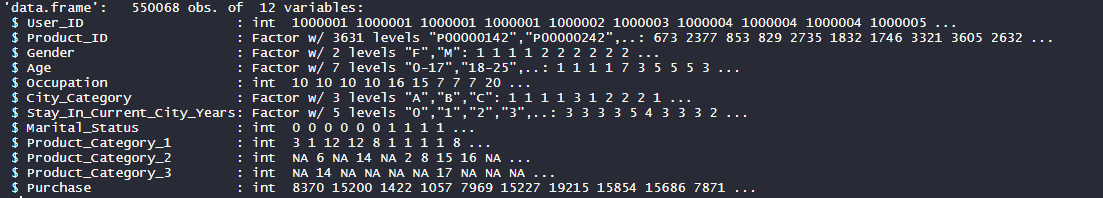
233599 rows 11 columns

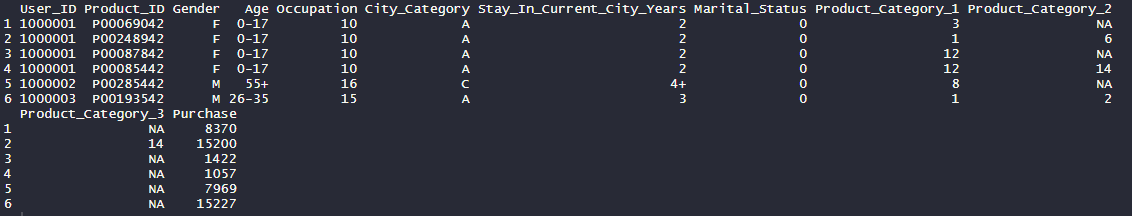
Let’s have a look on the train data

Variables:

|  |  |
| --- | --- |
| Variable | Description |
| "User\_ID" | Unique identifier of shopper. |
| "Product\_ID" | Unique identifier of product (No key given) |
| "Gender" | Sex of shopper. |
| "Age" | Age of shopper split into bins. |
| "Occupation" | Occupation of shopper (No key given) |
| "City\_Category" | Residence location of shopper (No key given) |
| "Stay\_In\_Current\_City\_Years" | Number of years stay in current city |
| "Marital\_Status" | Marital status of shopper |
| "Product\_Category\_1" | Product category of purchase |
| "Product\_Category\_2" | Another Product category of purchase |
| "Product\_Category\_3" | Another Product category of purchase |
| "Purchase" | Purchase amount in dollars |

Let’s see the structure of our data

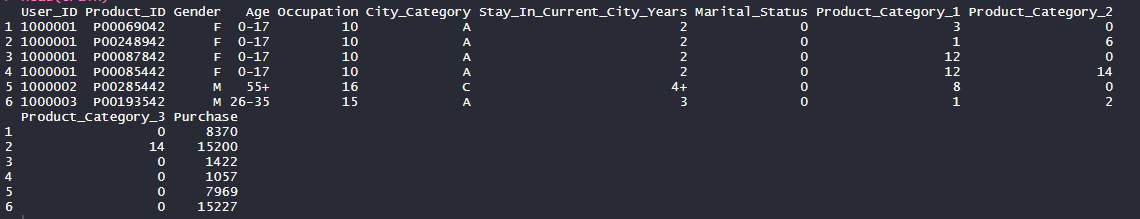


Let’s see some records of our data 

From this observation and after checking each column if it contains NA values, we found out that only columns with product\_Category\_2 and product\_Category\_3 have NA values

we can assume that having NA value means that the user did not purchase this product from those categories hence, we can replace the NA values by zeros

now check again we will have no NA values

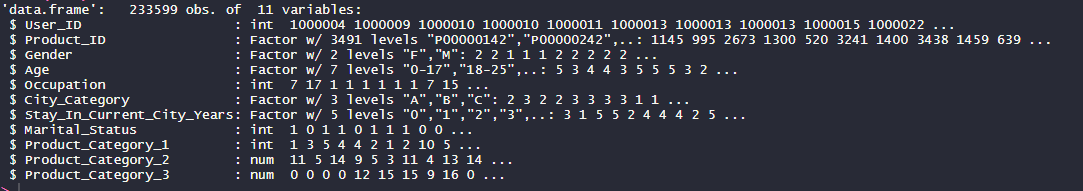


Now let’s check and explore our test set

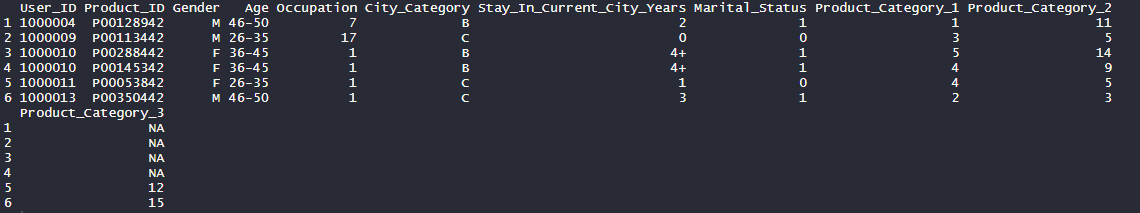
Variables:

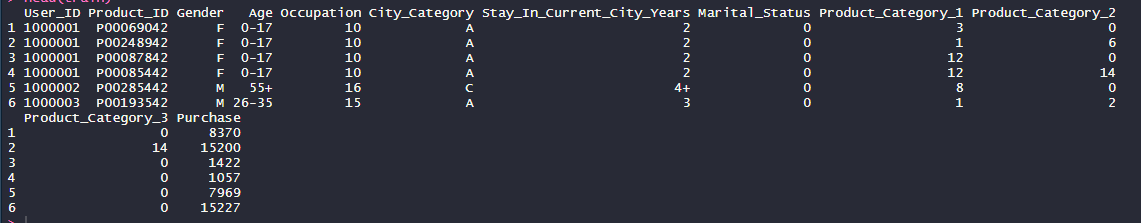
|  |  |
| --- | --- |
| Variable | Description |
| "User\_ID" | Unique identifier of shopper. |
| "Product\_ID" | Unique identifier of product (No key given) |
| "Gender" | Sex of shopper. |
| "Age" | Age of shopper split into bins. |
| "Occupation" | Occupation of shopper (No key given) |
| "City\_Category" | Residence location of shopper (No key given) |
| "Stay\_In\_Current\_City\_Years" | Number of years stay in current city |
| "Marital\_Status" | Marital status of shopper |
| "Product\_Category\_1" | Product category of purchase |
| "Product\_Category\_2" | Another Product category of purchase |
| "Product\_Category\_3" | Another Product category of purchase |

Structure of the test data



Some records of the test data



So, we have also NA values for product\_Category\_2 and product\_Category\_3 columns that need to be zeros 

Now having our train and test data cleaned let’s output them to new files

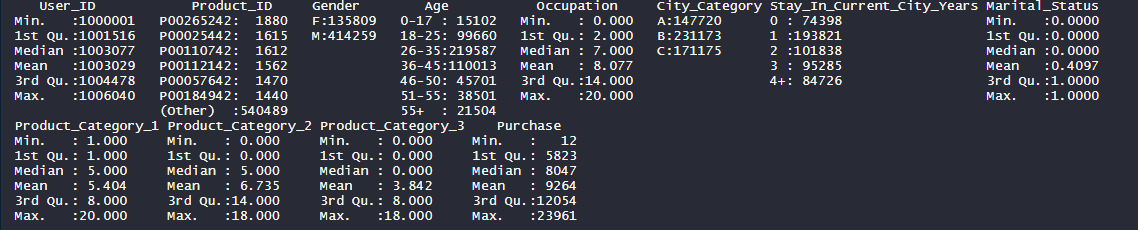
“train\_cleaned.csv”

“test\_cleaned.csv”

**Explanatory Data Analysis**

Now we have our data cleaned let’s start our data analysis phase to get some insights from our data

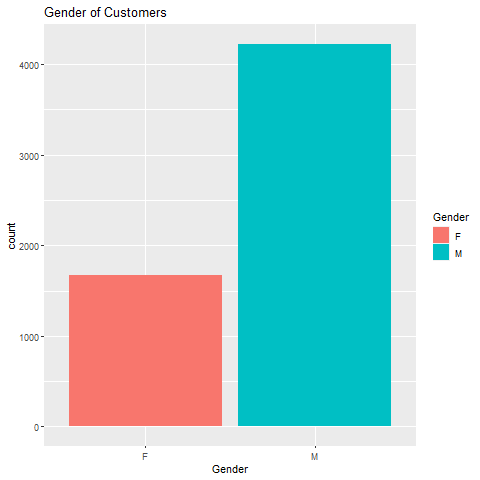
Here we focus only on explanatory analysis so we will only use train data

Let’s take a look on the data summary

Having an overview of our data, we have many variables that might affect the purchasing behavior of the customers let’s examine each of them in details

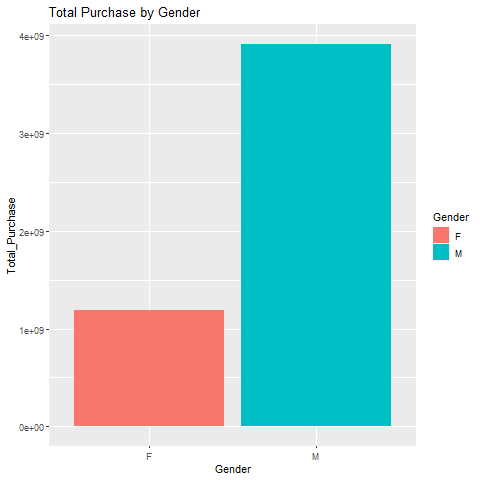
1. **Gender**

Let’s see the gender distribution of our customers



Here we can find that more males than females are shopping on black Friday, this can be useful for shop owners as they can modify their store layout, product selection, discounts and other variables differently depending on the gender proportion of their shoppers.

Let’s go deeper and compute the total spending amount corresponding to gender to see if we really should focus on males in promotions and discounts and do not focus on females or what.



Here we can see that the most purchasing amount comes from males so for sure we have to focus on male customers in discounts and offers.

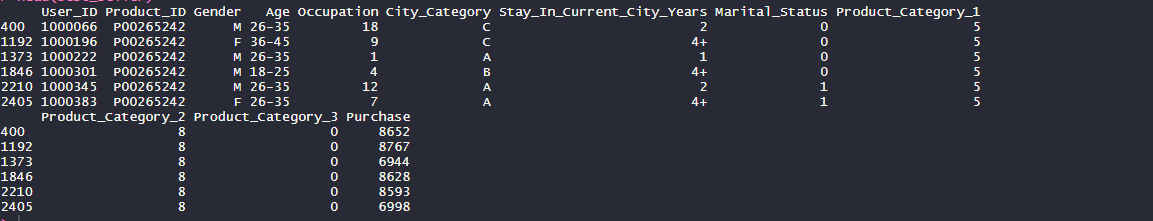
1. **Best Sellers**

Now we are going to find out the best seller products and investigate them

Our 3 best seller products

|  |  |
| --- | --- |
| Product\_ID | count |
| P00265242 | 1880 |
| P00025442 | 1615 |
| P00110742 | 1612 |

Let’s have a deeper look at our best seller product

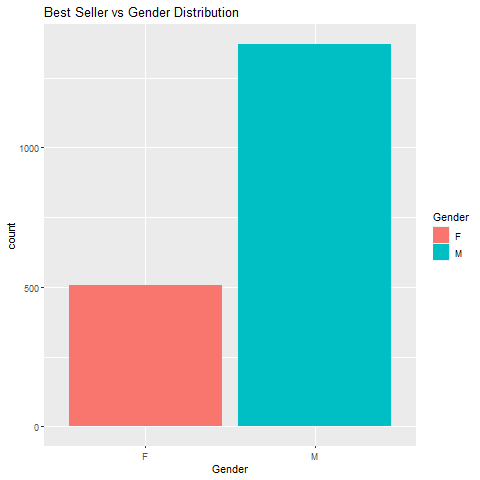


Now we notice that our best seller product falls into

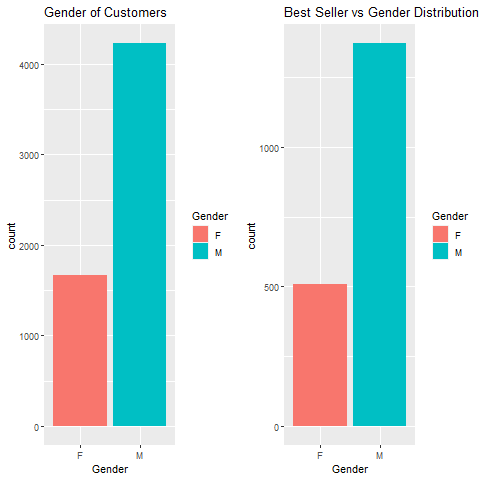
product\_category\_1 = 5 and product\_category\_2 = 8

A very interesting point to catch here is that our best seller product does not have the same price, this could be due to various Black Friday promotions, discounts, or coupon codes. In other cases, investigation would need to be done regarding the reason for different purchase prices of the same product between customers.

Let’s now see the relation between our best seller and the customer gender



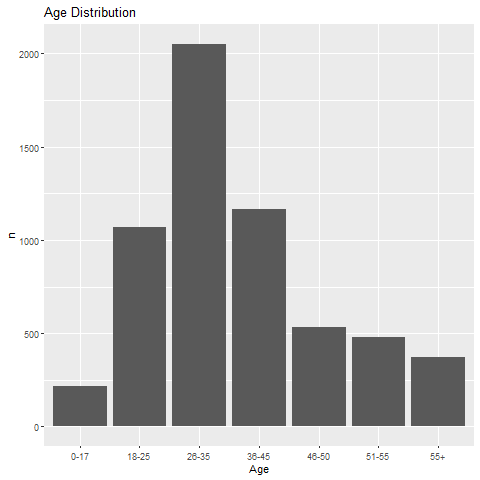
From here we can see that our best seller is more bought by males but let’s have a look at our best seller distribution by gender and the total purchasing distribution by gender



Although we concluded that the best seller is more bought by males but after comparing the above graphs we can see they are very similar which means all male buyers buy the best seller product but also all female buyers buy the best seller product so we can conclude that our best seller product doesn’t favor a specific gender.

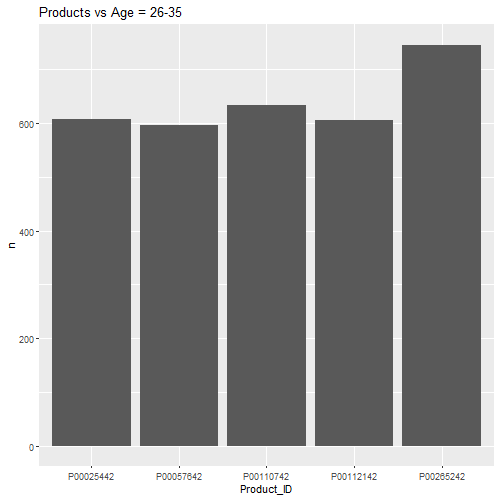
1. **Age**

Let’s now see the age distribution in our dataset



From here we can conclude that our buyers are between 26-35 years old

Let’s now see the products that buyers between 26-35 most buy.

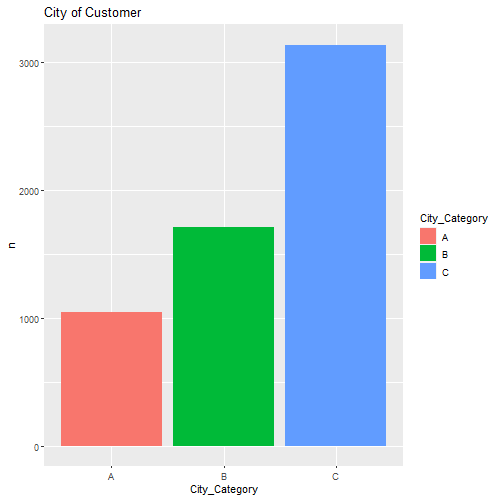


From here we can see the most bought 5 products by our most buyers with Age = 26-35 so we can promote these products, and for sure the most bought product by our most buyers is our best seller that we obtained before with

Product\_ID = P00265242

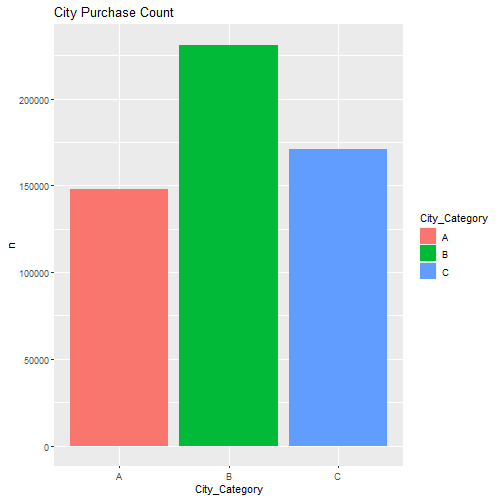
1. **City**

Let’s now see the location of our customers

****

Form this graph we can see that our most customers come from city category “C”, let’s now see the total purchase amount and count of each city to make sure to focus on the right city when making discounts and offers to increase sales.

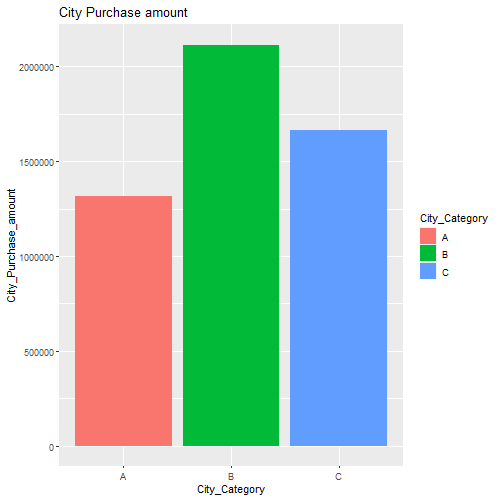
Total purchase count by city



Here we notice that, although city “C” has the most customers but the most purchase count come from city “B”

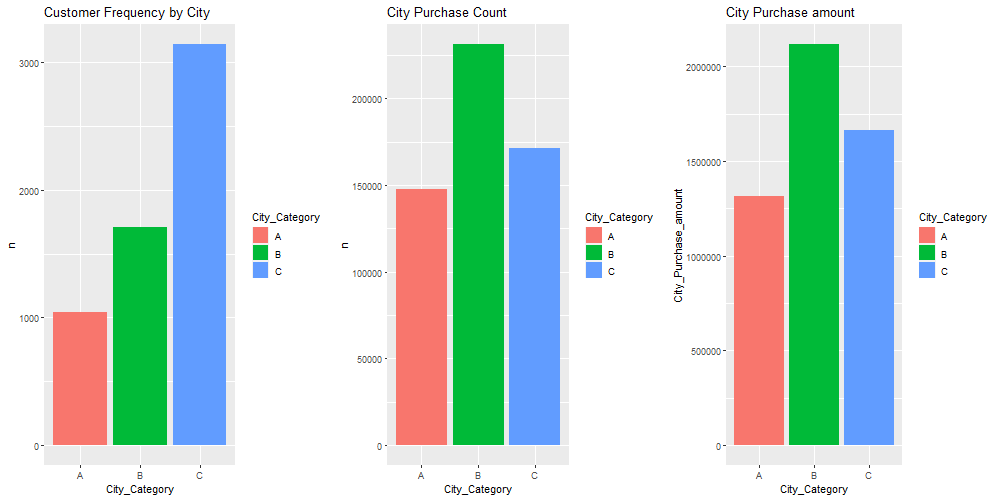
Let’s have a final check by visualizing the total purchase amount by city

City purchase amount



Having those 3 graphs let’s combine them and have our final conclusion.

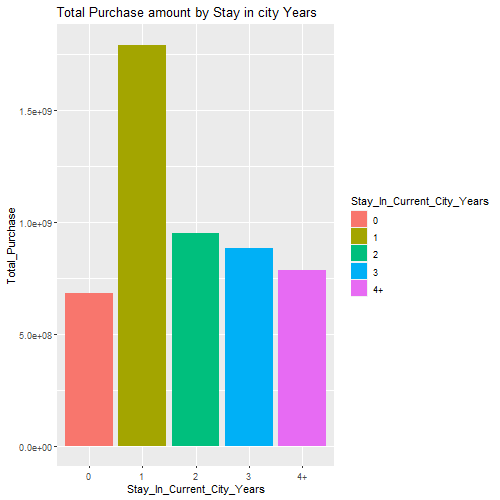
Comparison check



Now we see that, although most of our customers come from city “C” but the total purchase amount and count come from city “B”, so we have to focus on city “B” in our discounts and promotions.

1. **Stay in Current City**

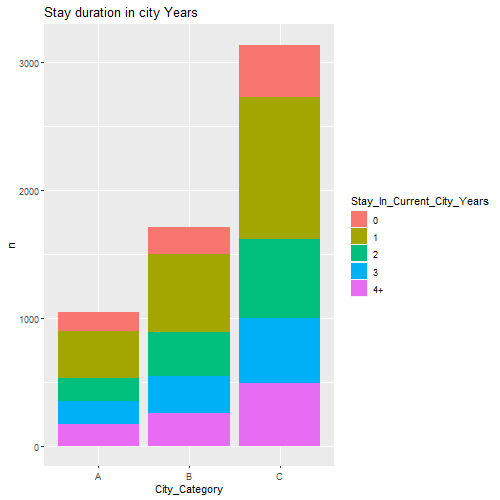
Let’s see the relationship between staying no. of years in the current city and the purchase amount



It looks like that the most purchase amount come from customers who stay only 1 year in their cities.

This is a very interesting observation as it seems that the more a person stays in his city the less the purchase amount, so it seems like we are losing customers every year but let’s be sure of this interesting observation by considering the amount of years a person stays in each city.

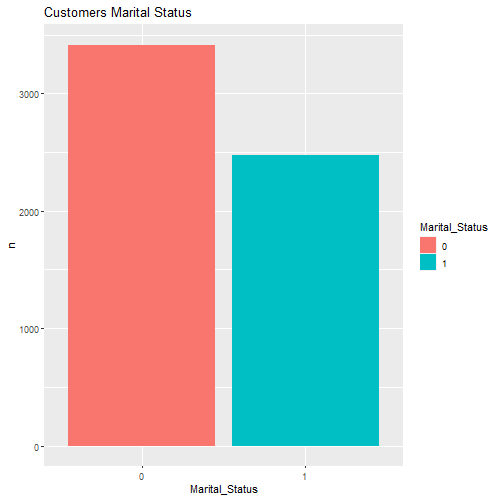
Stay in City Duration in Years



From this graph we can see that the most common stay length in each city is 1 year, so regarding the fact that we are losing customers is not true because as years go each city loses citizens so it’s normal to lose customers as years go.

1. **Marital Status**

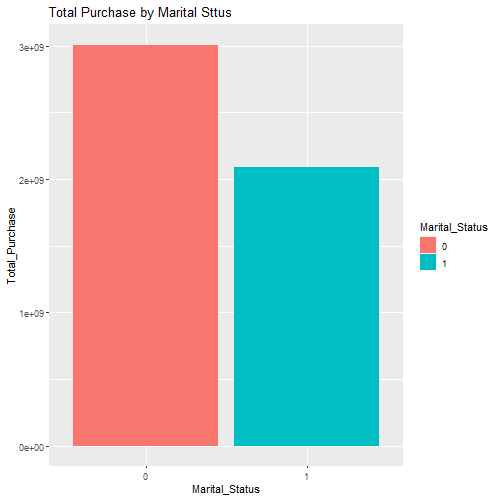
Now let’s whether most of our customers are single or married



As it’s not specified in the dataset what 0 and 1 correspond to, we will assume that 0 = single and 1 = married.

So, it looks like that our most customers are singles, but let’s check the total purchase amount by each marital status.

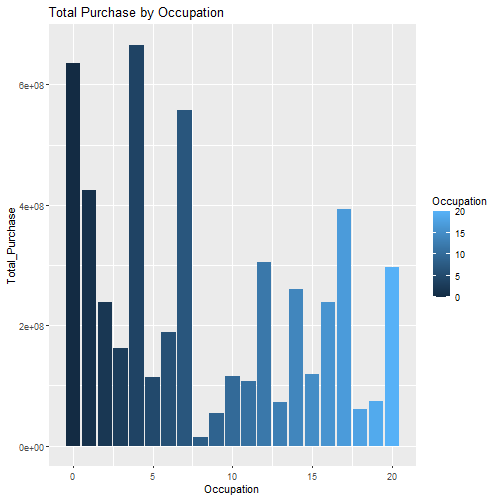
Marital Status vs Purchase amount



So from here also we can see that that most purchase amount come from single people :’(

1. **Occupation**

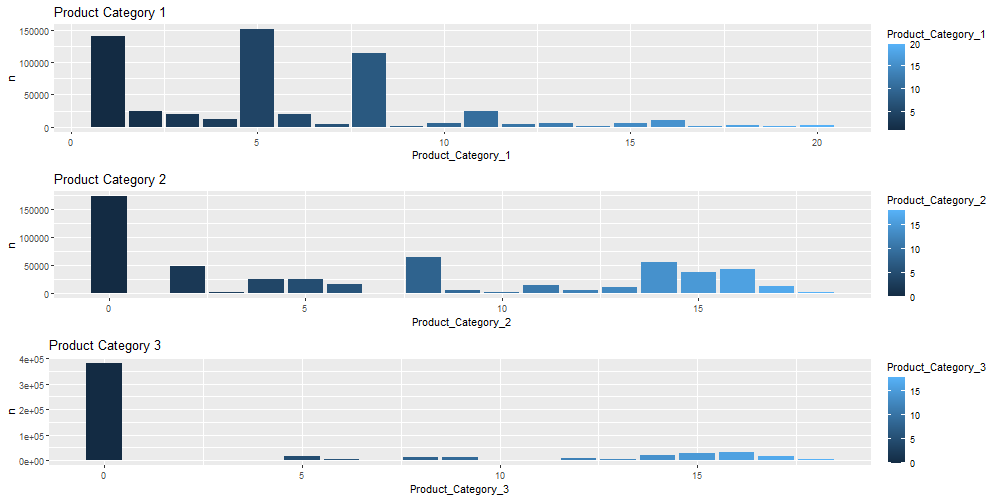
Let’s now see the effect of the occupation on the total purchase amount



From here we can see that the most purchasing customers are from occupation = 4 then occupation = 0 and occupation = 7

The data set does not show what exactly each occupation code means.

1. Product Categories



From this graph we can see that the most purchased product falls in categories:

Product Category 1 = 5,1,8

Product Category 2 = 8,14

Product Category 3 = 16,15

Now as we Notice that our best seller product falls into product category 1 = 5 and product category 2 = 8 which are the highest two product categories.

Note: Product Category = 0 is not actually a category but it was replacement for NA values meaning that the product wasn’t bought from this category.

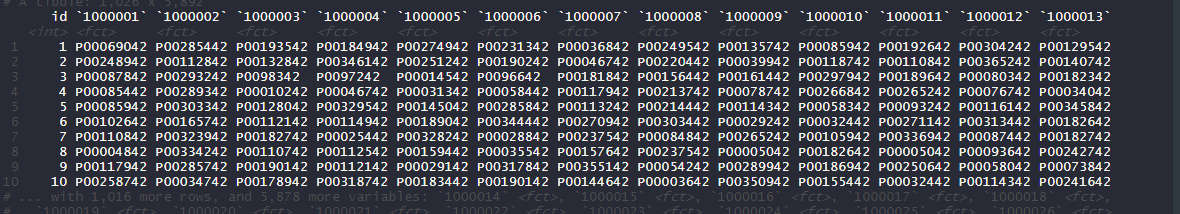
**Conclusions**

1. Purchasing behavior of males are larger than behavior of females in black Friday
2. Our best seller product “P00265242” that falls into product category 1 = 5 and product category 2 = 8
3. Our best seller product is bought by males as females so it does not favor specific gender
4. Most of customers are between 26-35 years old and our best seller product is most bought by this category
5. Although most of the customers come from city “C” but the most purchase count and amount come from “B”
6. Purchases from customers who live in their current city for only one years and the more the years the less the purchases but we actually the common stay length in each city is only 1 year
7. Single people have larger purchasing behavior than married people
8. Most purchased products fall in product category 1 = 5 and product category 2 = 8 which contains our best seller product as well

**Association Rule Mining**

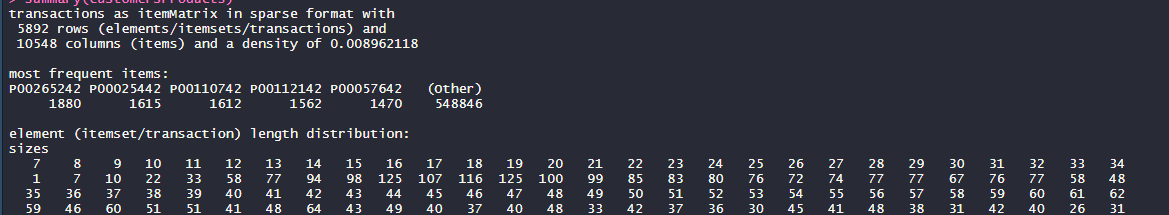
**Data Pre-Processing**

In order to be able to apply the Apriori algorithm we need to change the structure of the data.

1. Let’s extract User\_ID and Product\_ID from the dataset and arrange them
2. Now let’s spread our User\_ID and Product\_ID into a matrix where each columns have all products bought by a certain User
3. Finally, we transpose this matrix and write the new data into customers\_products.csv

We now have each row has all products bought by a certain user

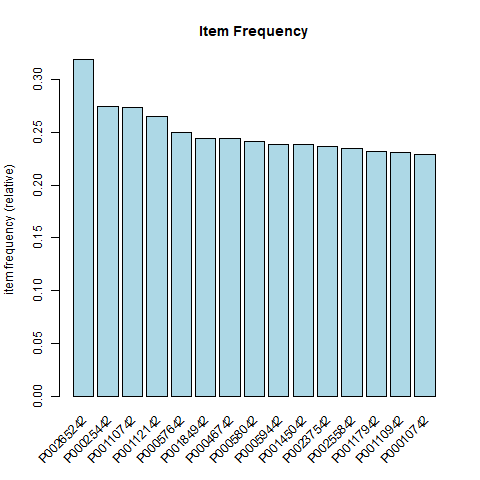
Having our data ready, let’s now see a summary



Here we notice that we have total of 5892 transactions and the most frequent items are the same Items we got in our explanatory data analysis while discovering our best seller product

|  |  |
| --- | --- |
| Product\_ID | count |
| P00265242 | 1880 |
| P00025442 | 1615 |
| P00110742 | 1612 |

Now let’s have a look at the item frequency plot

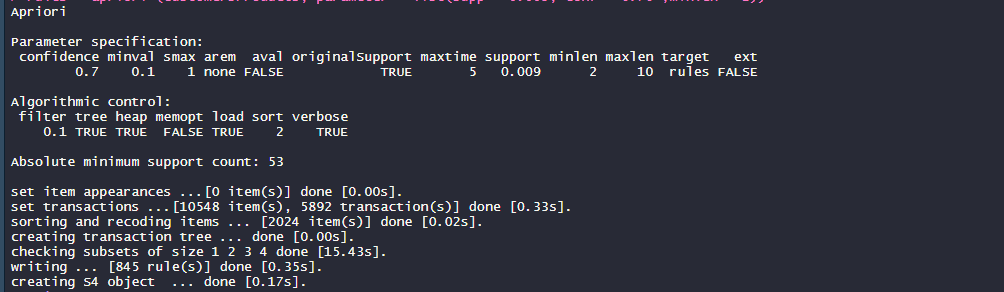


Now let’s determine our support and confidence values for the Apriori algorithm:

Support = no of item transactions / total no. of transactions

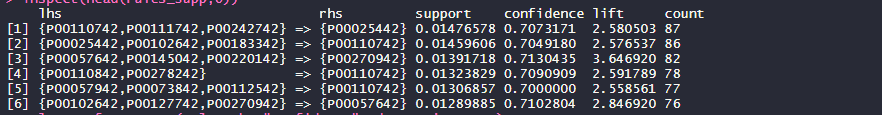
Let’s assume that we want to choose a product which was purchased by at least 55 (median no of customers buying a certain product) different customers.

Support = 55/5892(total transactions) = 0.009, and set confidence = 0.7



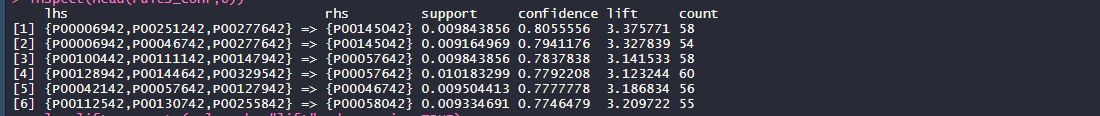
Looks like we have got 845 rules, let’s now discover our obtained rules.

Let’s sort our rules based on support



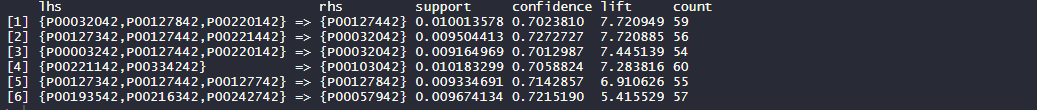
From this set if rules, I will choose the rule no. 3 with the highest lift and confidence since all the rules have almost the same support = 0.01

Let’s sort our rules based on confidence



From this set of rules, I will choose the first 3 rules with the highest confidence and almost the same lift.

Let’s sort our rules based on lift

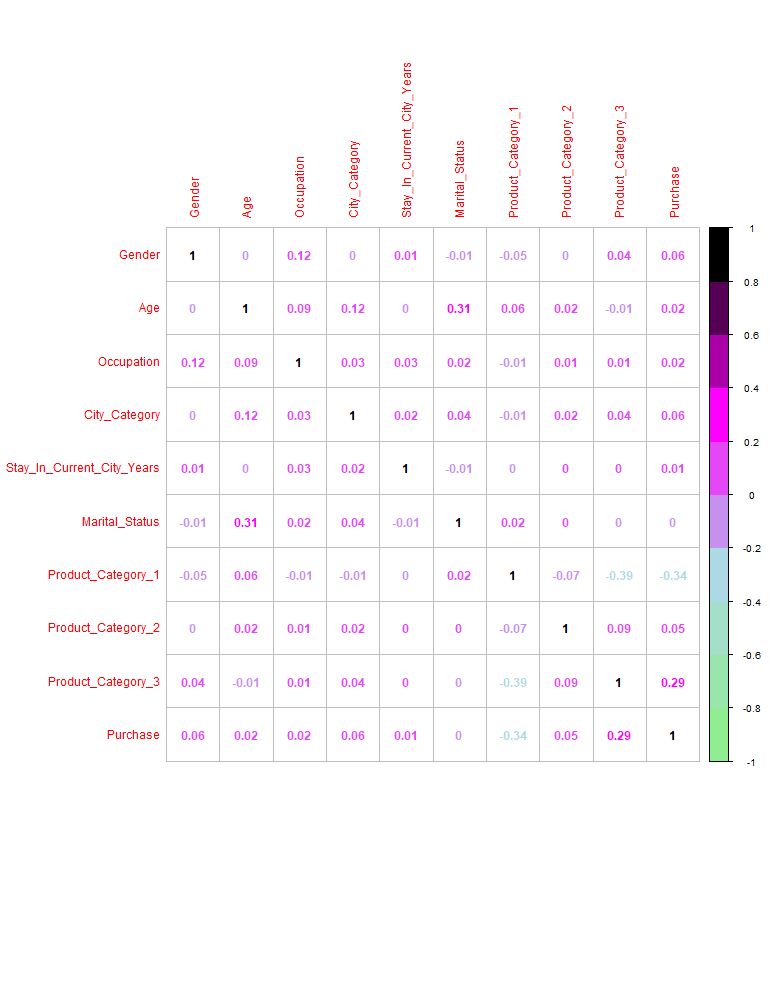


From this set of rules, I will choose rule no. 2 with the highest confidence and lift.

**Model Building**

After exploring data and having initial insights of how each variable can affect the purchase behavior of customers in black Fridays let’s now build our model to predict the purchase value.

First let’s have a look on the correlation matrix

For function “cor” to work correctly we will convert factors into integers

There does not seem to be any variable that would have a high impact on Purchase, since the highest correlation is given by Product Category 3 with 0.29. On the other hand, Product Category 1 has a negative correlation with our target with the value -0.34 which is somehow odd

**Some pre-processing steps:**

1. Converting Product categories, Occupation from integers to factors as they are categorical data.
2. Drop User\_ID and Product\_ID, as unique identifiers are not of our interest. Instead, customer’s general attributes are more of interest in estimating the influences on the target variable, purchase.

Now, let’s divide our data set into train/test sets with ratio 80/20

Although we have two csv files train.csv and test.csv, the dataset does not provide the labels for the test set, so we need to divide train.csv into train and test in order to evaluate our model.

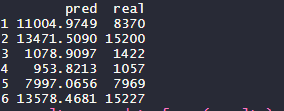
**Model 1 Linear Regression**

Let’s see the model R squared value



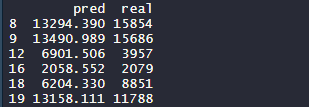
Looks like we have R squared almost 65% which is not very satisfying, but let’s have a look on some predicted values and compute the normalized root mean squared error on the train and test data.

Train data



Normalized RMSE = 0.5895921

Test data

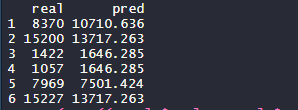


Normalized RMSE = 0.591615 which is good.

**Model 2 Decision Tree**

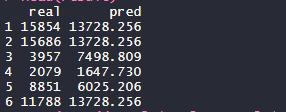
Let’s have a look on some predicted values and compute the normalized root mean squared error.

Train data



Normalized RMSE = 0.6115422 which is also good.

Test data



Normalized RMSE = 0.6133471 which is also good.

Looks like we got almost the same NRMSE from the two models, although the linear regressor was slightly better.

**Conclusion**

Overall, we have made some insightful discoveries from our EDA of this Black Friday dataset. We saw how customers at our store were distributed across multiple categorical classifications such as Gender, Age, Occupation, Stay in Current City, etc. We have also determined who our top selling Products and our “best seller” product, after our EDA, we applied Association Rule Learning and identified some association rules for our store on Black Friday, finally we build linear regression model and decision tree model to predict the purchase amount and end up with normalized root mean squared error almost 0.6.

**Future work**

We will build more complex models such as random forest and neural networks, we also can try to build classification models to predict categorical data such as gender of customer, city of customer and the Product that the customer is more likely to purchase, depending upon his gender, age, and occupation.