Black Friday Data Analysis

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 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

Acquiring and Cleaning Data

Acquiring and Cleaning Data

Our dataset is divided into csv files:

1. **Train.csv**

   550068 rows 12 columns

2. **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

```
'data.frame':   550068 obs. of  12 variables:
 $ User_ID                   : int  1000001 1000001 1000001 1000001 1000002 1000003 1000004 1000004 1000004 1000005 ...
 $ Product_ID                : Factor w/ 3631 levels "P00000142","P00000242",..: 673 2377 853 829 2735 1832 1746 3321 3605 2632 ...
 $ Gender                    : Factor w/ 2 levels "F","M": 1 1 1 1 2 2 2 2 2 2 ...
 $ Age                       : Factor w/ 7 levels "0-17","18-25",..: 1 1 1 1 7 3 5 5 5 3 ...
 $ Occupation                : int  10 10 10 10 16 15 7 7 7 20 ...
 $ City_Category             : Factor w/ 3 levels "A","B","C": 1 1 1 1 3 1 2 2 2 1 ...
 $ Stay_In_Current_City_Years: Factor w/ 5 levels "0","1","2","3",..: 3 3 3 3 5 4 3 3 3 2 ...
 $ Marital_Status            : int  0 0 0 0 0 0 1 1 1 1 ...
 $ Product_Category_1        : int  3 1 12 12 8 1 1 1 1 8 ...
 $ Product_Category_2        : int  NA 6 NA 14 NA 2 8 15 16 NA ...
 $ Product_Category_3        : int  NA 14 NA NA NA NA 17 NA NA NA ...
 $ Purchase                  : int  8370 15200 1422 1057 7969 15227 19215 15854 15686 7871 ...
```

Let’s see some records of our data

```
  User_ID Product_ID Gender   Age Occupation City_Category Stay_In_Current_City_Years Marital_Status Product_Category_1 Product_Category_2
1 1000001  P00069042      F  0-17         10             A                          2              0                  3                 NA
2 1000001  P00248942      F  0-17         10             A                          2              0                  1                  6
3 1000001  P00087842      F  0-17         10             A                          2              0                 12                 NA
4 1000001  P00085442      F  0-17         10             A                          2              0                 12                 14
5 1000002  P00285442      M   55+         16             C                         4+              0                  8                 NA
6 1000003  P00193542      M 26-35         15             A                          3              0                  1                  2
  Product_Category_3 Purchase
1                 NA     8370
2                 14    15200
3                 NA     1422
4                 NA     1057
5                 NA     7969
6                 NA    15227
```

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

```
  User_ID Product_ID Gender   Age Occupation City_Category Stay_In_Current_City_Years Marital_Status Product_Category_1 Product_Category_2
1 1000001  P00069042      F  0-17         10             A                          2              0                  3                  0
2 1000001  P00248942      F  0-17         10             A                          2              0                  1                  6
3 1000001  P00087842      F  0-17         10             A                          2              0                 12                  0
4 1000001  P00085442      F  0-17         10             A                          2              0                 12                 14
5 1000002  P00285442      M   55+         16             C                         4+              0                  8                  0
6 1000003  P00193542      M 26-35         15             A                          3              0                  1                  2
  Product_Category_3 Purchase
1                  0     8370
2                 14    15200
3                  0     1422
4                  0     1057
5                  0     7969
6                  0    15227
```

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

```
'data.frame':	233599 obs. of  11 variables:
 $ User_ID                   : int  1000004 1000009 1000010 1000010 1000011 1000013 1000013 1000013 1000015 1000022 ...
 $ Product_ID                : Factor w/ 3491 levels "P00000142","P00000242",..: 1145 995 2673 1300 520 3241 1400 3438 1459 639 ...
 $ Gender                    : Factor w/ 2 levels "F","M": 2 2 1 1 1 2 2 2 2 2 ...
 $ Age                       : Factor w/ 7 levels "0-17","18-25",..: 5 3 4 4 3 5 5 5 3 2 ...
 $ Occupation                : int  7 17 1 1 1 1 1 1 7 15 ...
 $ City_Category             : Factor w/ 3 levels "A","B","C": 2 3 2 2 3 3 3 3 1 1 ...
 $ Stay_In_Current_City_Years: Factor w/ 5 levels "0","1","2","3",..: 3 1 5 5 2 4 4 4 2 5 ...
 $ Marital_Status            : int  1 0 1 1 0 1 1 1 0 0 ...
 $ Product_Category_1        : int  1 3 5 4 4 2 1 2 10 5 ...
 $ Product_Category_2        : num  11 5 14 9 5 3 11 4 13 14 ...
 $ Product_Category_3        : num  0 0 0 0 12 15 15 9 16 0 ...
```

Some records of the test data

```
  User_ID Product_ID Gender   Age Occupation City_Category Stay_In_Current_City_Years Marital_Status Product_Category_1 Product_Category_2
1 1000004  P00128942      M 46-50          7             B                          2              1                  1                 11
2 1000009  P00113442      M 26-35         17             C                          0              0                  3                  5
3 1000010  P00288442      F 36-45          1             B                         4+              1                  5                 14
4 1000010  P00145342      F 36-45          1             B                         4+              1                  4                  9
5 1000011  P00053842      F 26-35          1             C                          1              0                  4                  5
6 1000013  P00350442      M 46-50          1             C                          3              1                  2                  3
  Product_Category_3
1                 NA
2                 NA
3                 NA
4                 NA
5                 12
6                 15
```

So, we have also **NA** values for product_Category_2 and product_Category_3 columns that need to be **zeros**

```
  User_ID Product_ID Gender   Age Occupation City_Category Stay_In_Current_City_Years Marital_Status Product_Category_1 Product_Category_2
1 1000001  P00069042      F  0-17         10             A                          2              0                  3                  0
2 1000001  P00248942      F  0-17         10             A                          2              0                  1                  6
3 1000001  P00087842      F  0-17         10             A                          2              0                 12                  0
4 1000001  P00085442      F  0-17         10             A                          2              0                 12                 14
5 1000002  P00285442      M   55+         16             C                         4+              0                  8                  0
6 1000003  P00193542      M 26-35         15             A                          3              0                  1                  2
  Product_Category_3 Purchase
1                  0     8370
2                 14    15200
3                  0     1422
4                  0     1057
5                  0     7969
6                  0    15227
```

Explanatory Data Analysis

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

"train_cleaned.csv"

"test_cleaned.csv"

Explanatory Data Analysis

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

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

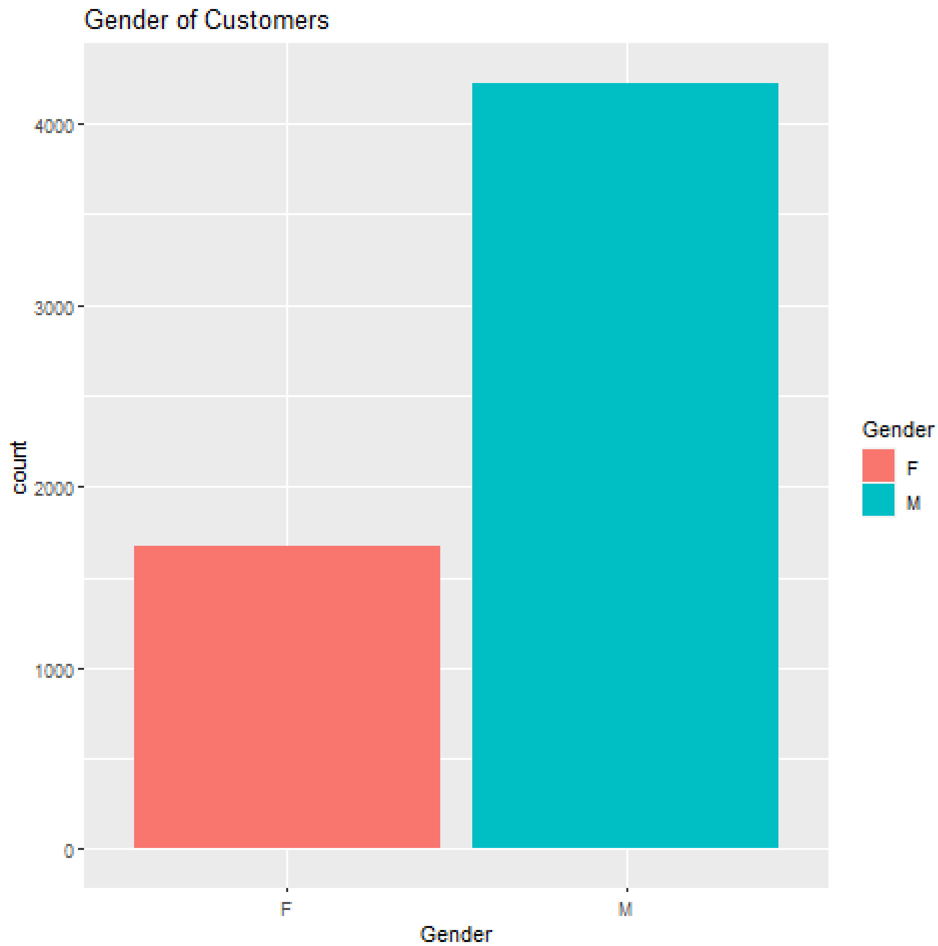
Let's take a look on the data summary

```
    User_ID           Product_ID     Gender       Age          Occupation     City_Category Stay_In_Current_City_Years Marital_Status
 Min.   :1000001   P00265242:  1880   F:135809   0-17 : 15102   Min.   : 0.000   A:147720      0 : 74398                  Min.   :0.0000
 1st Qu.:1001516   P00025442:  1615   M:414259   18-25: 99660   1st Qu.: 2.000   B:231173      1 :193821                  1st Qu.:0.0000
 Median :1003077   P00110742:  1612              26-35:219587   Median : 7.000   C:171175      2 :101838                  Median :0.0000
 Mean   :1003029   P00112142:  1562              36-45:110013   Mean   : 8.077                 3 : 95285                  Mean   :0.4097
 3rd Qu.:1004478   P00057642:  1470              46-50: 45701   3rd Qu.:14.000                 4+: 84726                  3rd Qu.:1.0000
 Max.   :1006040   P00184942:  1440              51-55: 38501   Max.   :20.000                                            Max.   :1.0000
                   (Other)  :540489              55+  : 21504
 Product_Category_1 Product_Category_2 Product_Category_3    Purchase
 Min.   : 1.000     Min.   : 0.000     Min.   : 0.000     Min.   :   12
 1st Qu.: 1.000     1st Qu.: 0.000     1st Qu.: 0.000     1st Qu.: 5823
 Median : 5.000     Median : 5.000     Median : 0.000     Median : 8047
 Mean   : 5.404     Mean   : 6.735     Mean   : 3.842     Mean   : 9264
 3rd Qu.: 8.000     3rd Qu.:14.000     3rd Qu.: 8.000     3rd Qu.:12054
 Max.   :20.000     Max.   :18.000     Max.   :18.000     Max.   :23961
```

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

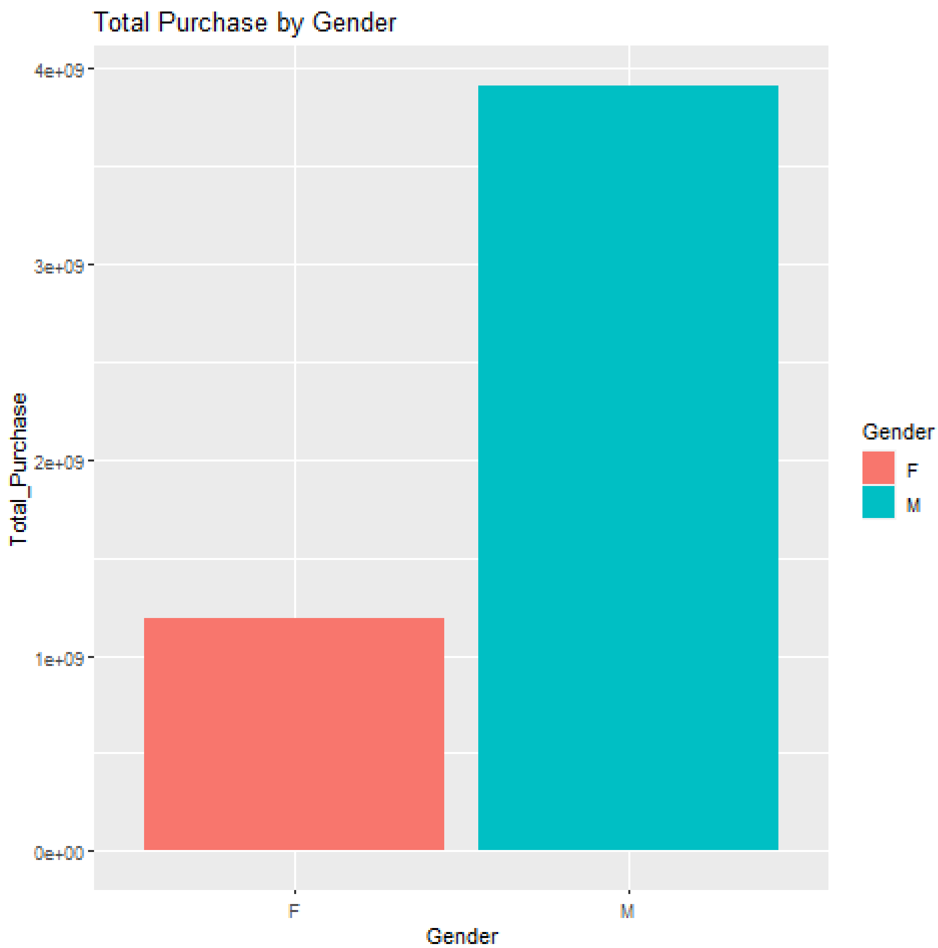
1- Gender

Let’s see the gender distribution of our customers



Here we can find that **<u>more males than females are shopping on black Friday,</u>** 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.



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

2- 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

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1
400	1000066	P00265242	M	26-35	18	C	2	0	5
1192	1000196	P00265242	F	36-45	9	C	4+	0	5
1373	1000222	P00265242	M	26-35	1	A	1	0	5
1846	1000301	P00265242	M	18-25	4	B	4+	0	5
2210	1000345	P00265242	M	26-35	12	A	2	1	5
2405	1000383	P00265242	F	26-35	7	A	4+	1	5

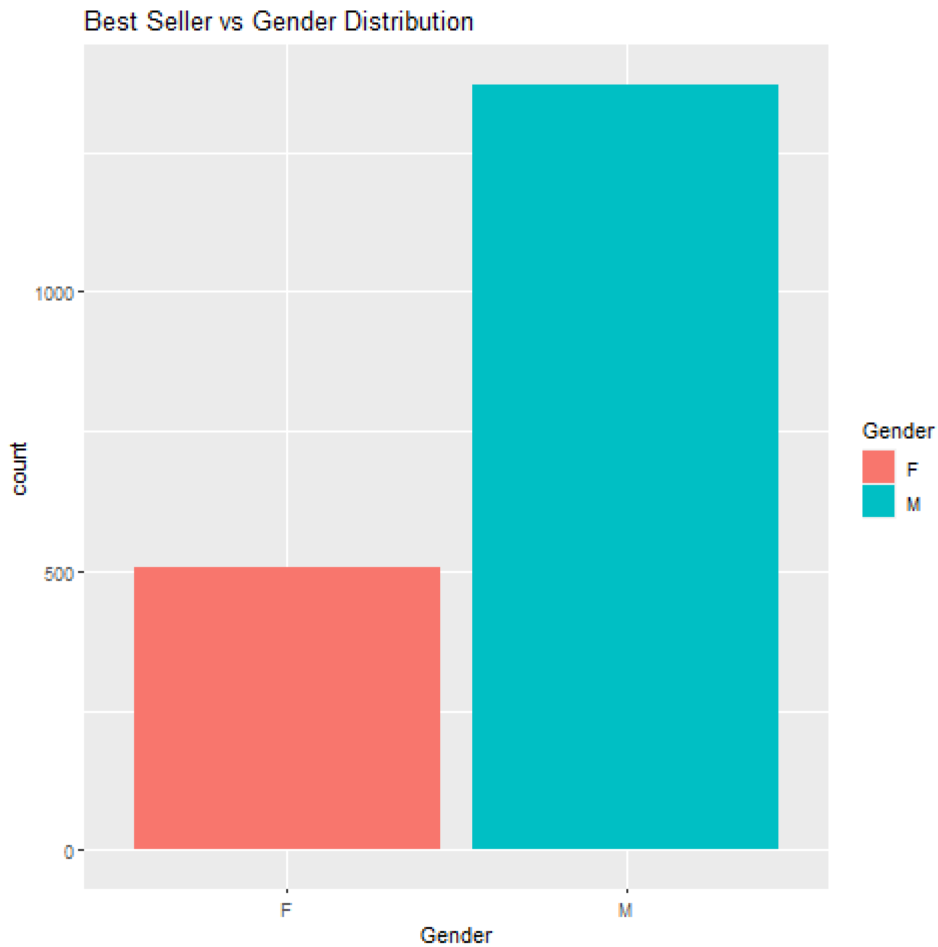
	Product_Category_2	Product_Category_3	Purchase
400	8	0	8652
1192	8	0	8767
1373	8	0	6944
1846	8	0	8628
2210	8	0	8593
2405	8	0	6998

Now we notice that our best seller product falls into

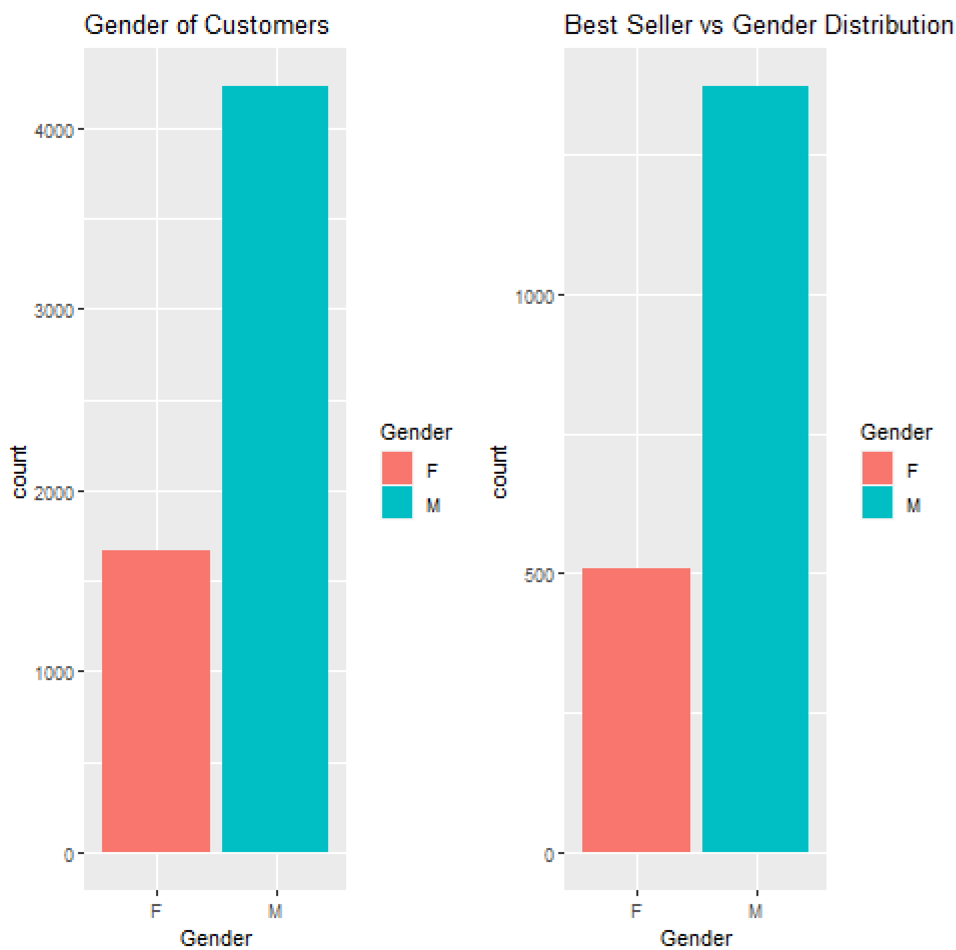
product_category_1 = 5 and product_category_2 = 8

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



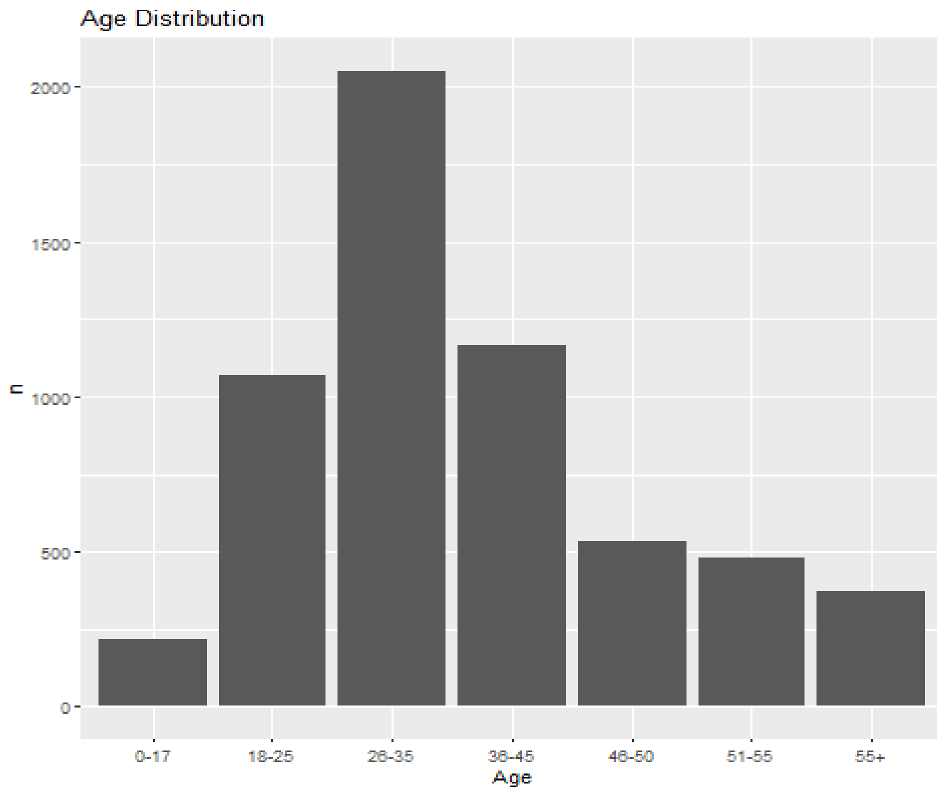
From here we can see that our best seller is more bought by males but let’s have a look at our best seller distribution by gender and the total purchasing distribution by gender



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.**

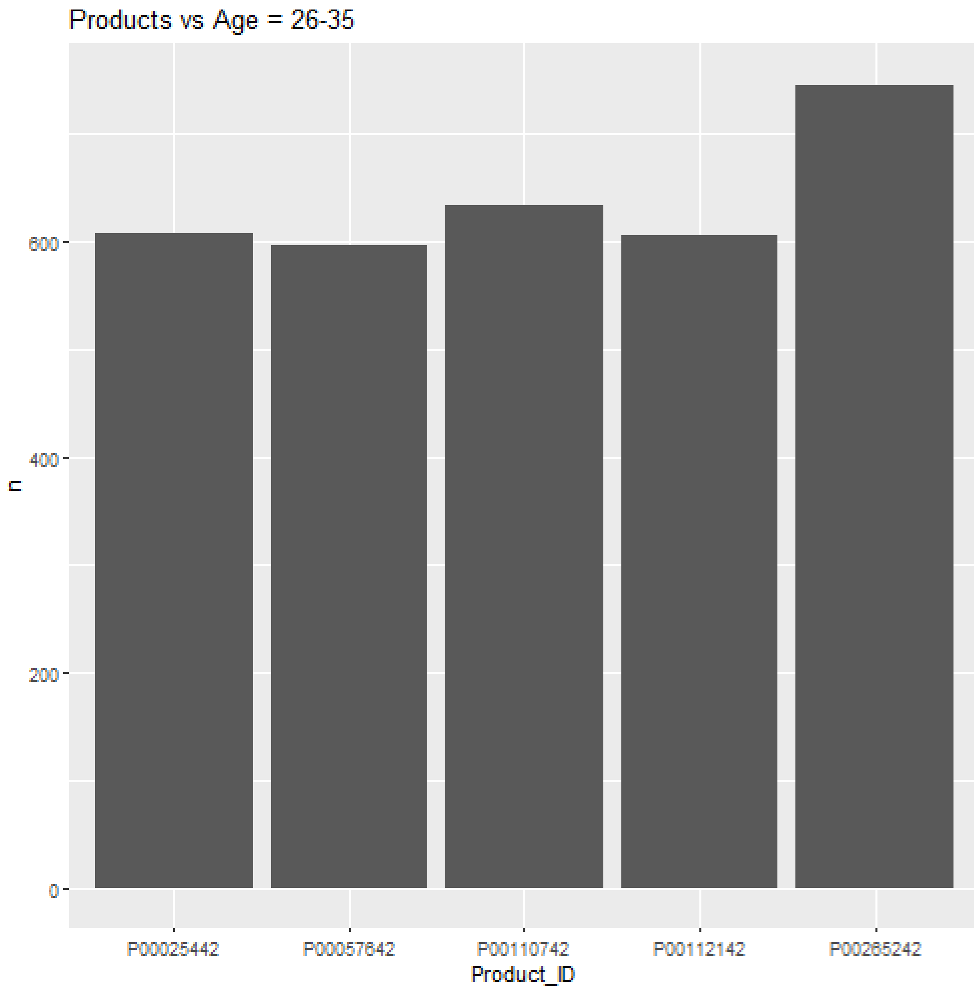
3- Age

Let’s now see the age distribution in our dataset



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

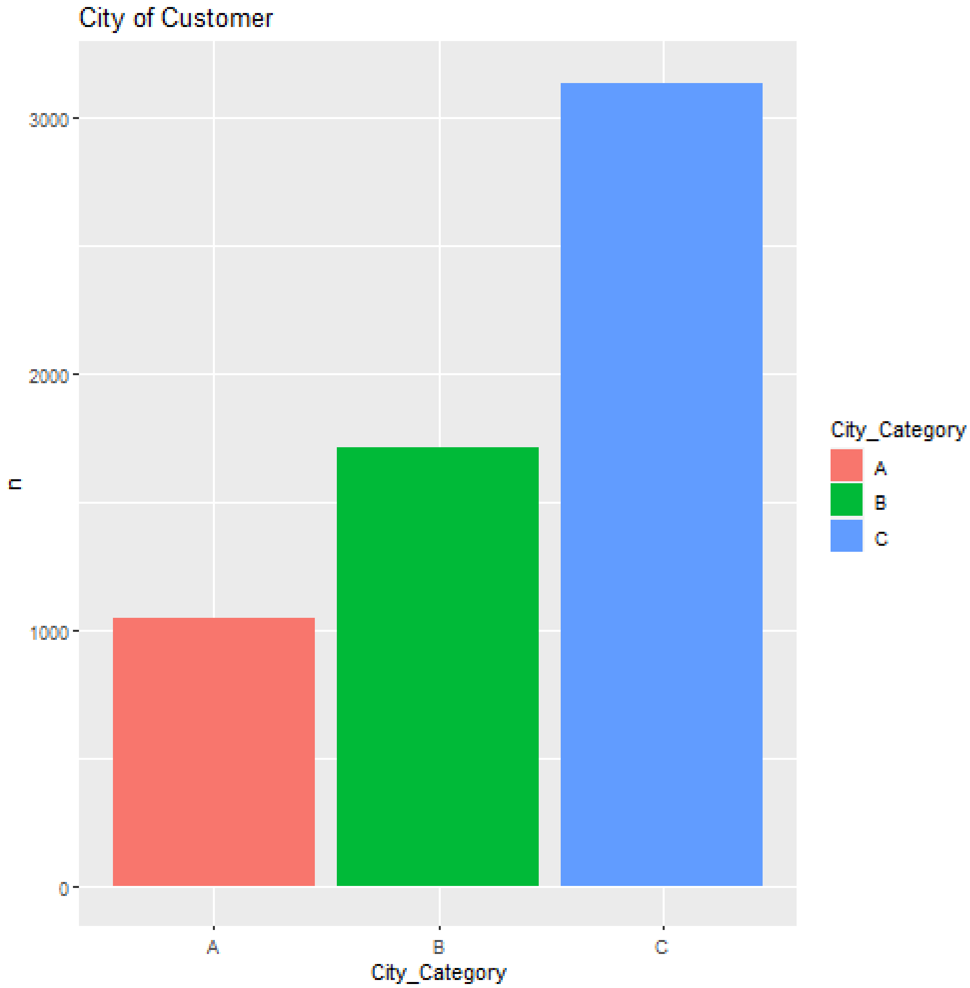
Let’s now see the products that buyers between 26-35 most buy.



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

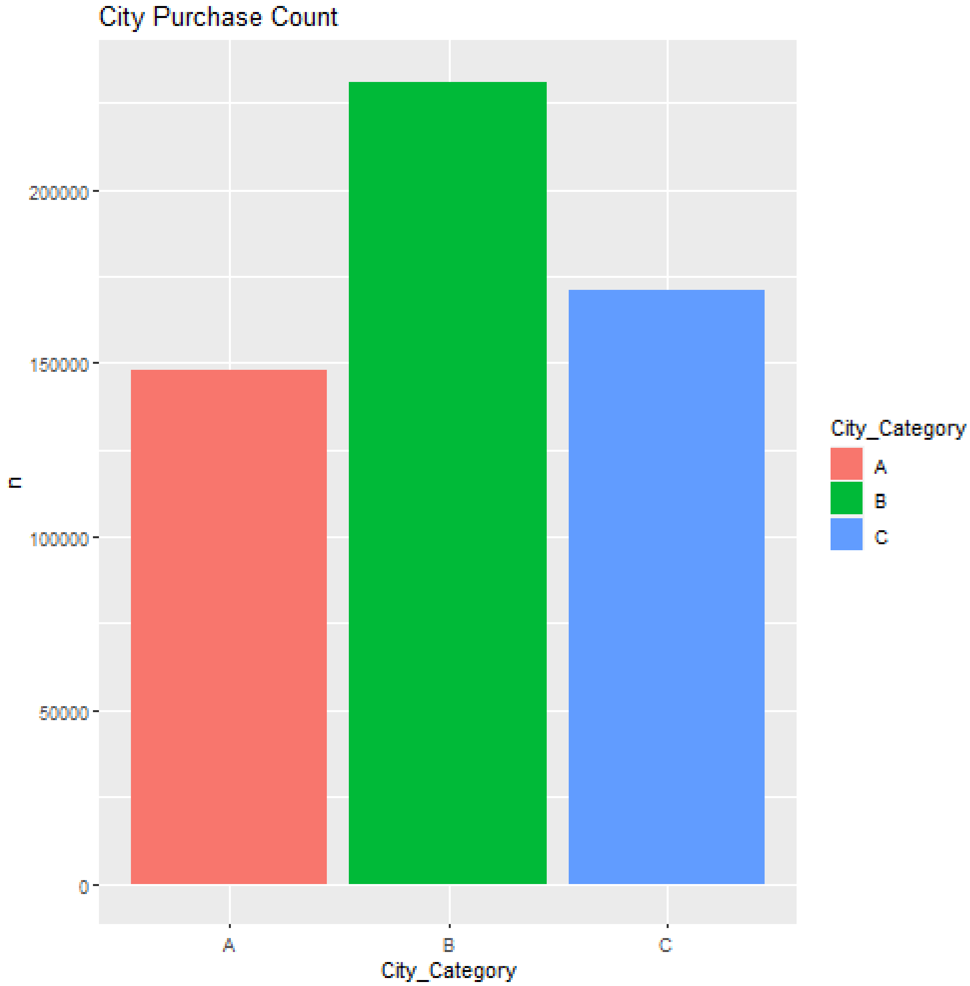
4- City

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.

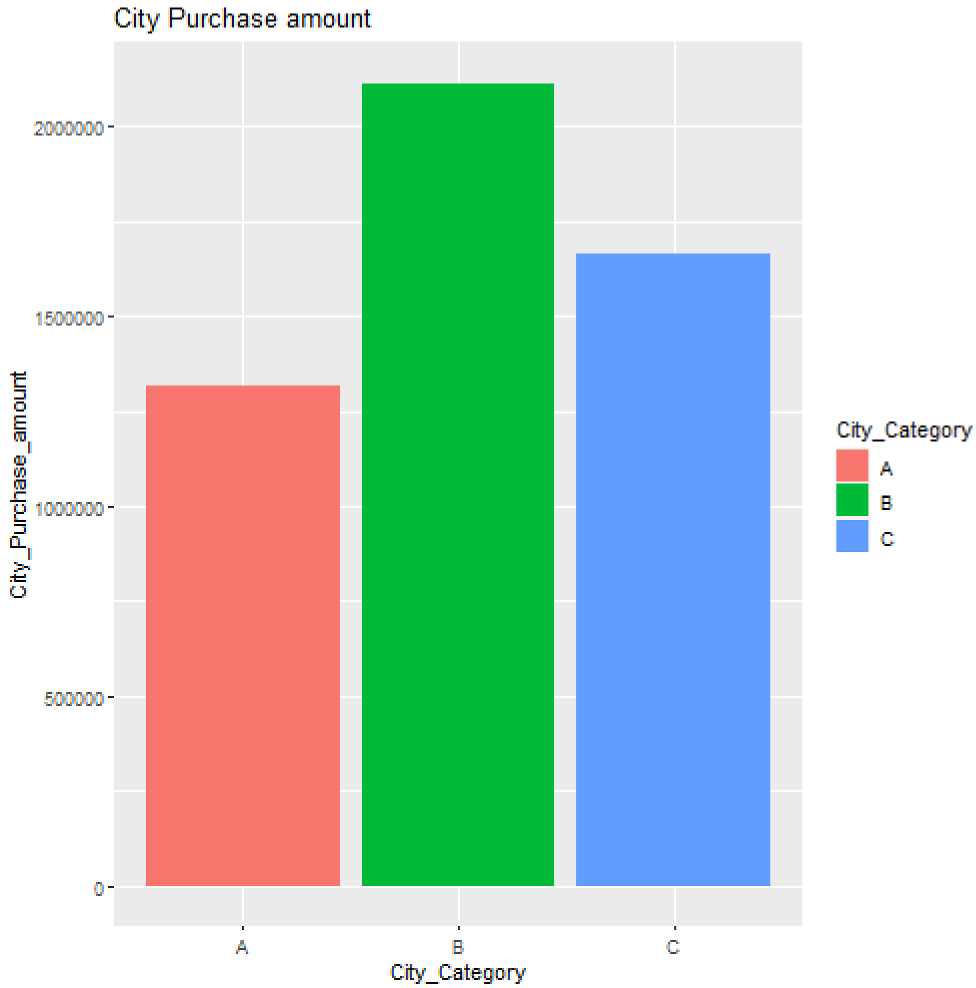
Total purchase count by city



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

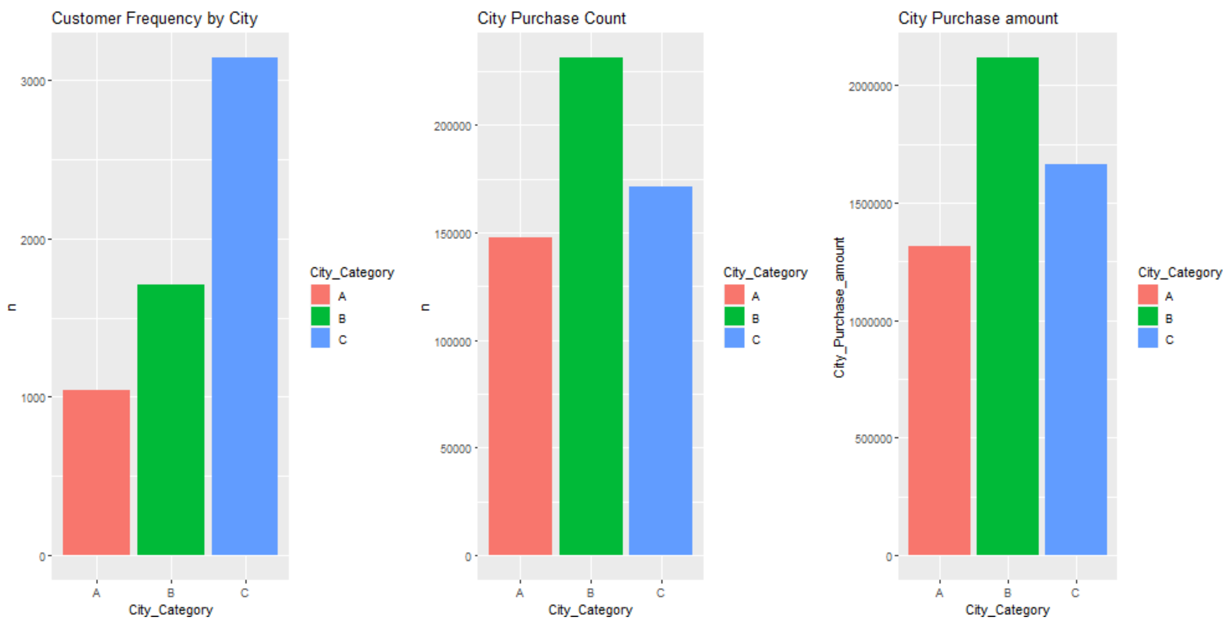
Let’s have a final check by visualizing the total purchase amount by city

City purchase amount



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

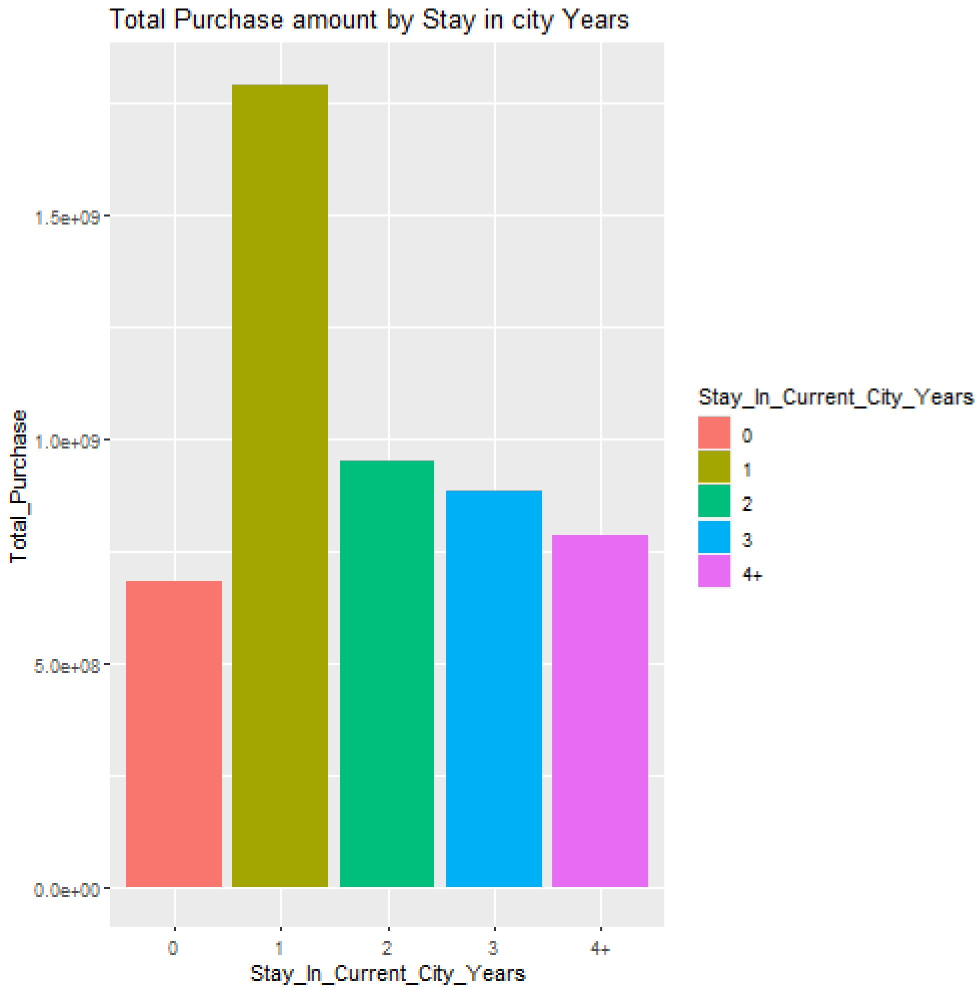
Comparison check



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**.

5- Stay in Current City

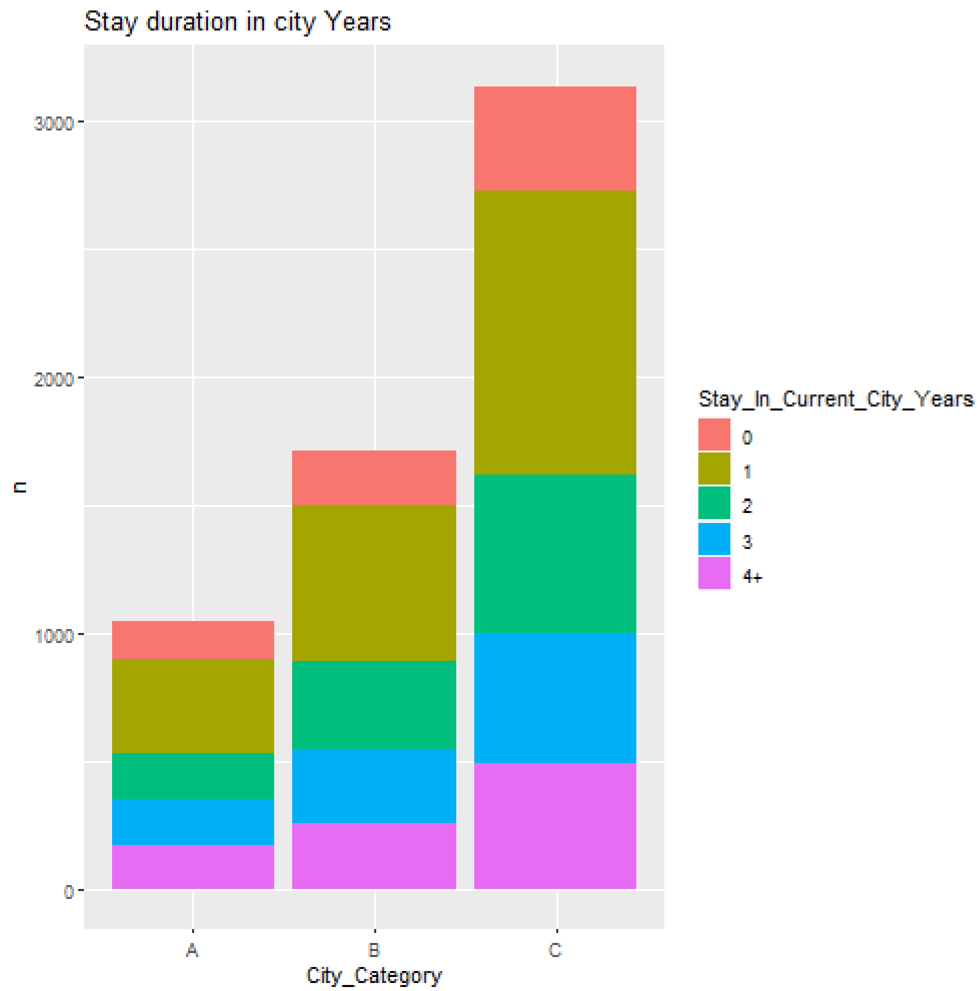
Let's see the relationship between staying no. of years in the current city and the purchase amount



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.

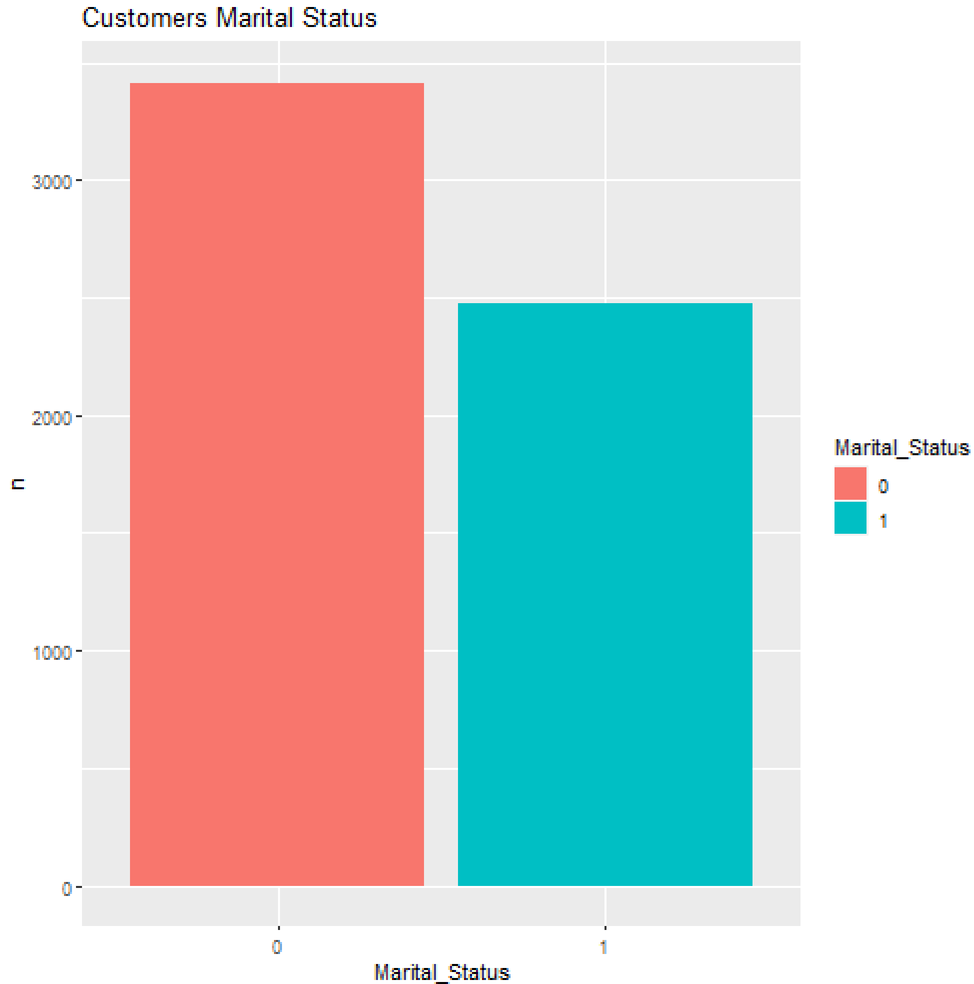
Stay in City Duration in Years



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.

6- Marital Status

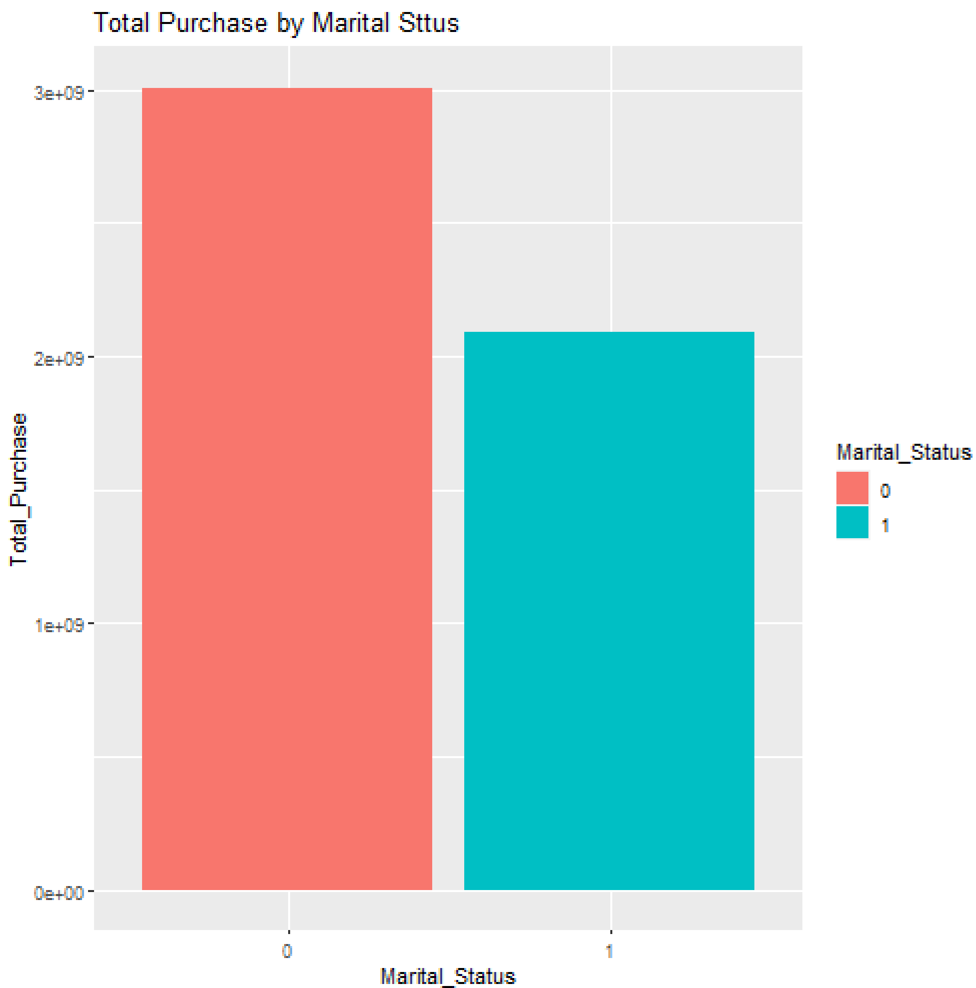
Now let’s whether most of our customers are single or married



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.

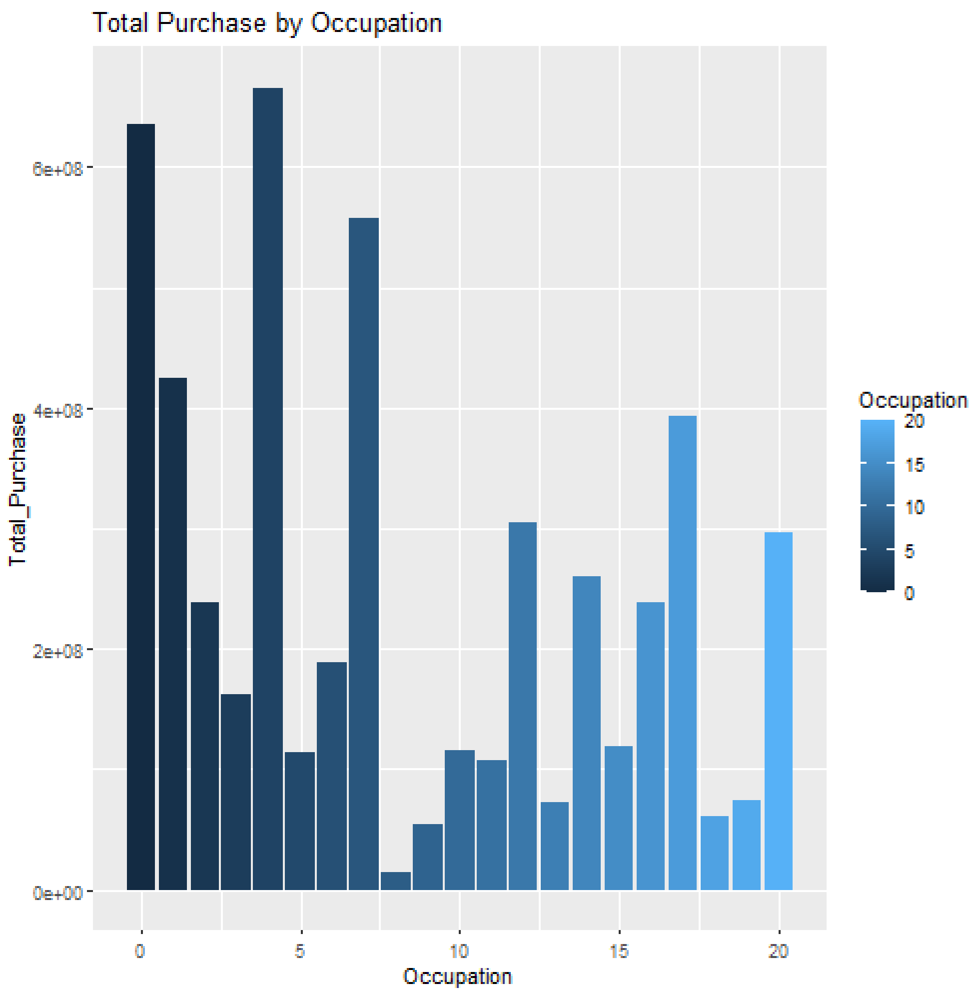
Marital Status vs Purchase amount



So from here also we can see that that **<u>most purchase amount come from single people :'(</u>**

7- Occupation

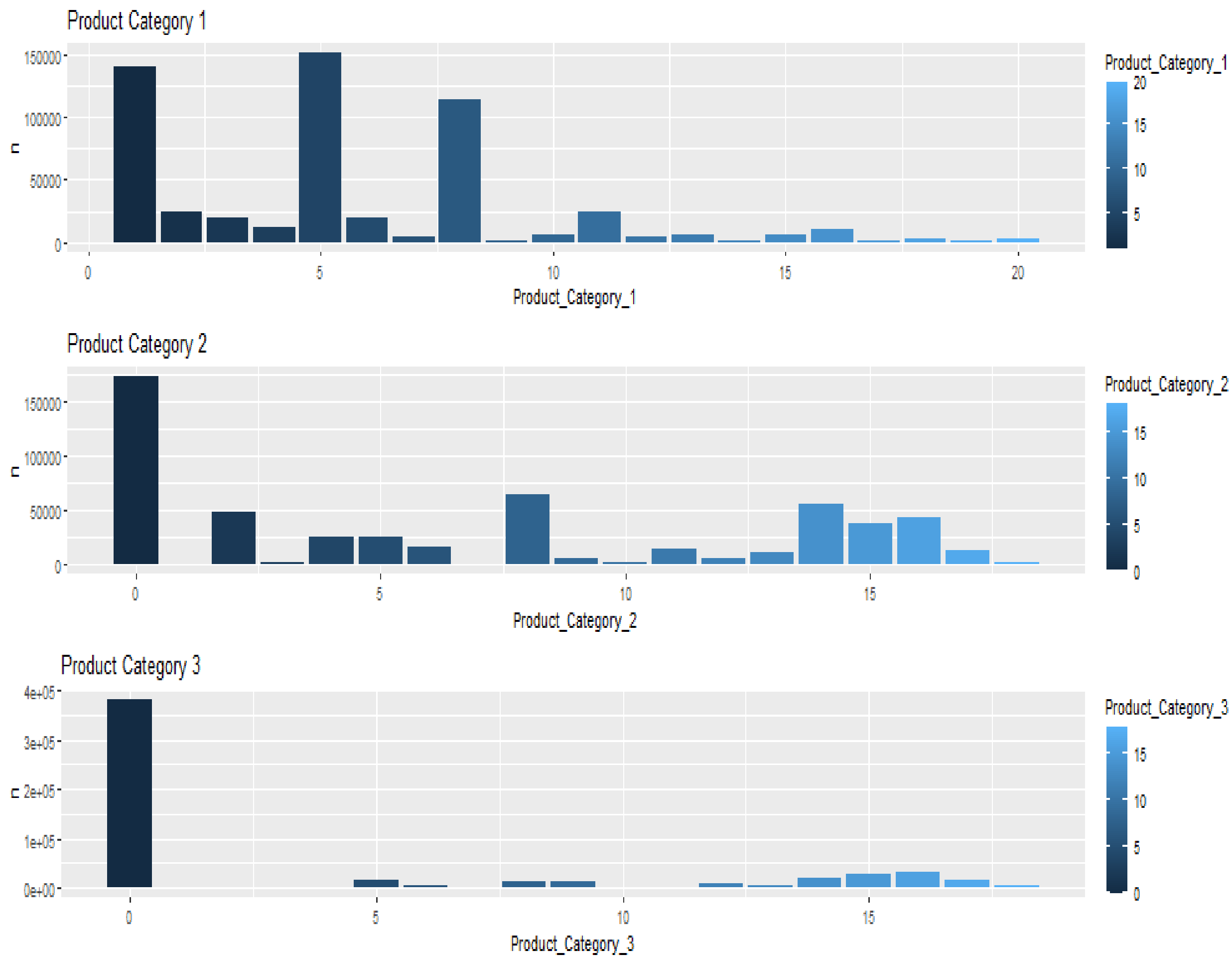
Let's now see the effect of the occupation on the total purchase amount



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.

8- 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

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

```
   User_ID Product_ID
     <int> <fct>
 1 1000001 P00069042
 2 1000001 P00248942
 3 1000001 P00087842
 4 1000001 P00085442
 5 1000001 P00085942
 6 1000001 P00102642
 7 1000001 P00110842
 8 1000001 P00004842
 9 1000001 P00117942
10 1000001 P00258742
```

2- Now let's spread our User_ID and Product_ID into a matrix where each of the columns has all products bought by a certain User

```
# A tibble: 1,026 x 5,892
      id `1000001` `1000002` `1000003` `1000004` `1000005` `1000006` `1000007` `1000008` `1000009` `1000010` `1000011` `1000012` `1000013`
   <int> <fct>     <fct>     <fct>     <fct>     <fct>     <fct>     <fct>     <fct>     <fct>     <fct>     <fct>     <fct>     <fct>
 1     1 P00069042 P00285442 P00193542 P00184942 P00274942 P00231342 P00036842 P00249542 P00135742 P00085942 P00192642 P00304242 P00129542
 2     2 P00248942 P00112842 P00132842 P00346142 P00251242 P00190242 P00046742 P00220442 P00039942 P00118742 P00110842 P00365242 P00140742
 3     3 P00087842 P00293242 P0098342  P0097242  P00014542 P0096642  P00181842 P00156442 P00161442 P00297942 P00189642 P00080342 P00182342
 4     4 P00085442 P00289342 P00010242 P00046742 P00031342 P00058442 P00117942 P00213742 P00078742 P00266842 P00265242 P00076742 P00034042
 5     5 P00085942 P00303342 P00128042 P00329542 P00145042 P00285842 P00113242 P00214442 P00114342 P00058342 P00093242 P00116142 P00345842
 6     6 P00102642 P00165742 P00112142 P00114942 P00189042 P00344442 P00270942 P00303442 P00029242 P00032442 P00271142 P00313442 P00182642
 7     7 P00110842 P00323942 P00182742 P00025442 P00328242 P00028842 P00237542 P00084842 P00265242 P00105942 P00336942 P00087442 P00182742
 8     8 P00004842 P00334242 P00110742 P00112542 P00159442 P00035542 P00157642 P00237542 P00005042 P00182642 P00005042 P00093642 P00242742
 9     9 P00117942 P00285742 P00190142 P00112142 P00029142 P00317842 P00355142 P00054242 P00289942 P00186942 P00250642 P00058042 P00073842
10    10 P00258742 P00034742 P00178942 P00318742 P00183442 P00190142 P00144642 P00003642 P00350942 P00155442 P00032442 P00114342 P00241642
# ... with 1,016 more rows, and 5,878 more variables: `1000014` <fct>, `1000015` <fct>, `1000016` <fct>, `1000017` <fct>, `1000018` <fct>,
#   `1000019` <fct>, `1000020` <fct>, `1000021` <fct>, `1000022` <fct>, `1000023` <fct>, `1000024` <fct>, `1000025` <fct>, `1000026` <fct>
```

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

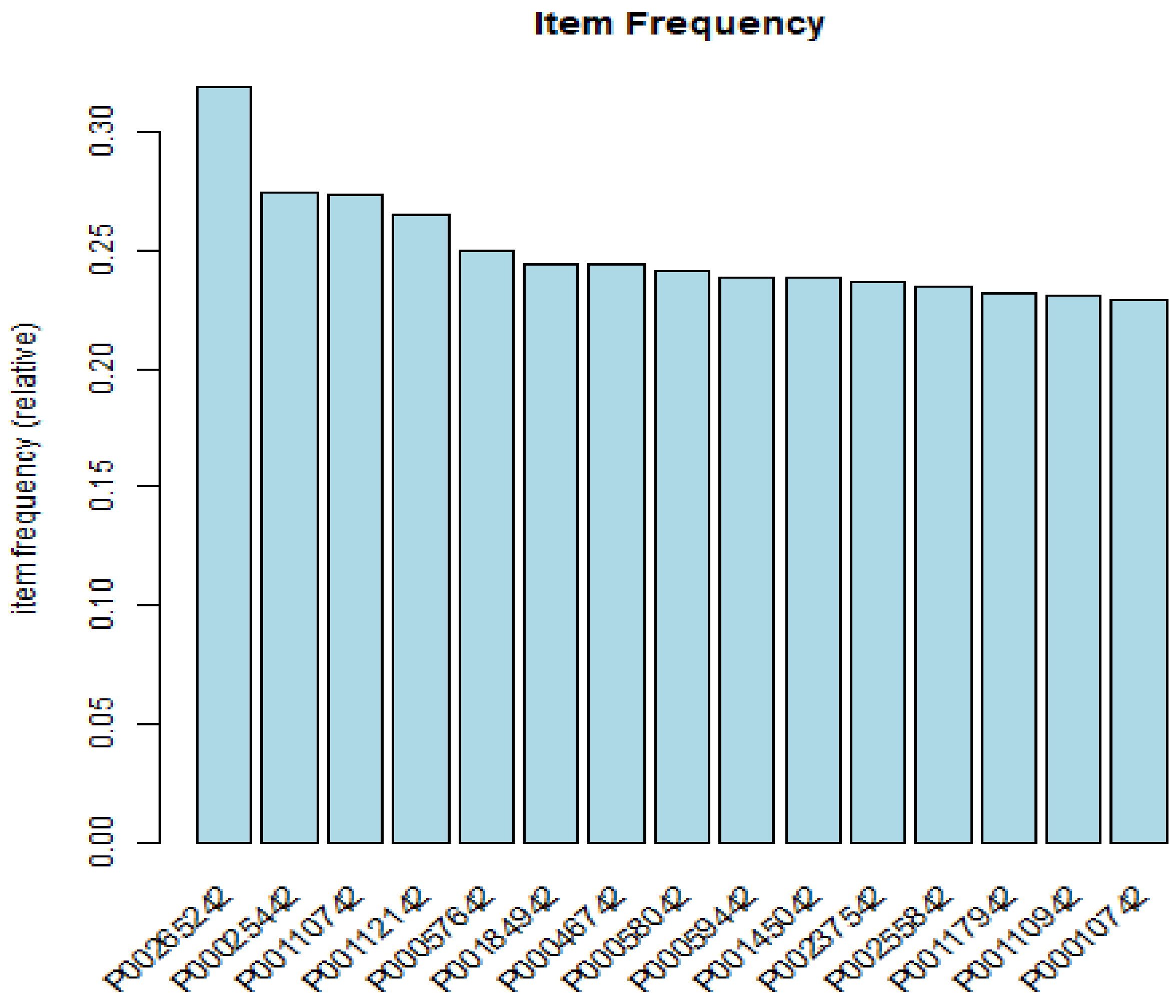
Having our data ready, let’s now see a summary

```
> summary(customersProducts)
transactions as itemMatrix in sparse format with
 5892 rows (elements/itemsets/transactions) and
 10548 columns (items) and a density of 0.008962118

most frequent items:
P00265242 P00025442 P00110742 P00112142 P00057642   (other)
     1880      1615      1612      1562      1470    548846

element (itemset/transaction) length distribution:
sizes
  7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34
  1   7  10  22  33  58  77  94  98 125 107 116 125 100  99  85  83  80  76  72  74  77  77  67  76  77  58  48
 35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62
 59  46  60  51  51  41  48  64  43  49  40  37  40  48  33  42  37  36  30  45  41  48  38  31  42  40  26  31
```

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

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.7    0.1    1 none FALSE            TRUE       5   0.009      2     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 53

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[10548 item(s), 5892 transaction(s)] done [0.33s].
sorting and recoding items ... [2024 item(s)] done [0.02s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 done [15.43s].
writing ... [845 rule(s)] done [0.35s].
creating S4 object  ... done [0.17s].
```

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

Let’s sort our rules based on support

```
    lhs                             rhs         support    confidence lift     count
[1] {P00110742,P00111742,P00242742} => {P00025442} 0.01476578 0.7073171  2.580503 87
[2] {P00025442,P00102642,P00183342} => {P00110742} 0.01459606 0.7049180  2.576537 86
[3] {P00057642,P00145042,P00220142} => {P00270942} 0.01391718 0.7130435  3.646920 82
[4] {P00110842,P00278242}           => {P00110742} 0.01323829 0.7090909  2.591789 78
[5] {P00057942,P00073842,P00112542} => {P00110742} 0.01306857 0.7000000  2.558561 77
[6] {P00102642,P00127742,P00270942} => {P00057642} 0.01289885 0.7102804  2.846920 76
```

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

```
    lhs                                   rhs          support     confidence lift     count
[1] {P00006942,P00251242,P00277642} => {P00145042} 0.009843856 0.8055556  3.375771 58
[2] {P00006942,P00046742,P00277642} => {P00145042} 0.009164969 0.7941176  3.327839 54
[3] {P00100442,P00111142,P00147942} => {P00057642} 0.009843856 0.7837838  3.141533 58
[4] {P00128942,P00144642,P00329542} => {P00057642} 0.010183299 0.7792208  3.123244 60
[5] {P00042142,P00057642,P00127942} => {P00046742} 0.009504413 0.7777778  3.186834 56
[6] {P00112542,P00130742,P00255842} => {P00058042} 0.009334691 0.7746479  3.209722 55
```

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

```
    lhs                                   rhs          support     confidence lift     count
[1] {P00032042,P00127842,P00220142} => {P00127442} 0.010013578 0.7023810  7.720949 59
[2] {P00127342,P00127442,P00221442} => {P00032042} 0.009504413 0.7272727  7.720885 56
[3] {P00003242,P00127442,P00220142} => {P00032042} 0.009164969 0.7012987  7.445139 54
[4] {P00221142,P00334242}           => {P00103042} 0.010183299 0.7058824  7.283816 60
[5] {P00127342,P00127442,P00127742} => {P00127842} 0.009334691 0.7142857  6.910626 55
[6] {P00193542,P00216342,P00242742} => {P00057942} 0.009674134 0.7215190  5.415529 57
```

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

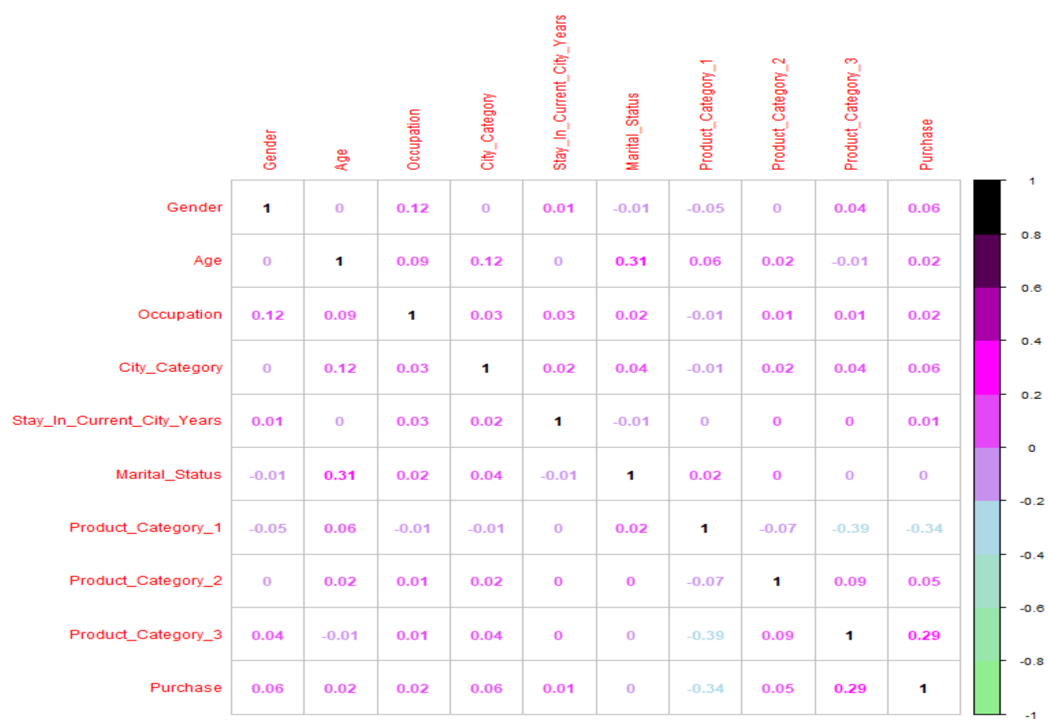
Model Building

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



It 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

```
Residual standard error: 2981 on 385001 degrees of freedom
Multiple R-squared:  0.648,	Adjusted R-squared:  0.6479
F-statistic:  8337 on 85 and 385001 DF,  p-value: < 2.2e-16
```

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

```
        pred  real
1 11004.9749  8370
2 13471.5090 15200
3  1078.9097  1422
4   953.8213  1057
5  7997.0656  7969
6 13578.4681 15227
```

Normalized RMSE = 0.5895921

Test data

```
        pred  real
8  13294.390 15854
9  13490.989 15686
12  6901.506  3957
16  2058.552  2079
18  6204.330  8851
19 13158.111 11788
```

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

	real	pred
1	8370	10710.636
2	15200	13717.263
3	1422	1646.285
4	1057	1646.285
5	7969	7501.424
6	15227	13717.263

Normalized RMSE = 0.6115422 which is also good.

Test data

	real	pred
1	15854	13728.256
2	15686	13728.256
3	3957	7498.809
4	2079	1647.730
5	8851	6025.206
6	11788	13728.256

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

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.