Black Friday
A Study of Consumer Behavior and Sales

Submitted by

Pranjal Bhalla

Registration No: 12016264

Programme Name: B. Tech. CSE With Specialization in AI/ML (3rd Year)

School of Computer Science & Engineering

Lovely Professional University, Phagwara

Acknowledgement

I want to acknowledge and thank all the people who played a significant role in my academic accomplishments to date.
Special thanks to Mr. Abhijeet Dutta for guidance throughout the
research, his support and guidance in the research methodologies and how to articulate a research study, and guidance in the initial shaping of research. I want to thank all the faculty and staff at Lovely Professional University for being incredibly supportive and making my experience memorable. I would also like to thank my parents for always having trust
in me and supporting me throughout my journey. I would also like to thank all my friends for their support. Thank you all for your unwavering support.

Abstract

The purpose of this study was to observe and analyze consumer behaviors and purchasing patterns. Black Friday is the biggest holiday shopping season in the United States as an excellent opportunity for retailers to increase their sales. Through the years, this practice is extending to other countries in the world. Although globally widespread, there has been a lack of research attention to consumers' attitudes toward and behaviors during Black Friday. Therefore, the study's purpose was to study Greek consumers' behaviors, attitudes, and intentions towards Black Friday. Moreover, this research's objective was to understand differences between shoppers on Black Friday and non-shoppers and suggest how retailers can harness these marketing differences. The online survey research method used, and data gathered from consumers. The study results show that there is a vast switch in the shopping methods during Black Friday practice. Moreover, the findings show that only 1/3 of consumers participate in in-store Black Friday sales, and the rest are shifting to online sales. Finally, purchasing patterns, as well as future consumer intentions, are presented.

Table of Contents

Introduction

The purpose of this study is to observe and analyze the consumer behaviors of the Black Friday customer. The day after Thanksgiving, Black Friday is a term used by the retail industry in the United States that signifies the Christmas holiday shopping season. (Swilley, 2013) Thanksgiving Day is on the last Thursday of November; therefore, the holiday shopping season runs from the Friday after Thanksgiving Day and continues until Christmas eve, the day before Christmas. Black Friday is not a national holiday. However, many employees have Thanksgiving Day holidays and the following day, increasing the number of potential shoppers on that Friday. Black Friday is famously known for long lines with customers waiting outdoors in cold weather waiting for the store to open, confusion, and customers' chaos. Once the retail doors open for business, the challenges faced are Heavily crowded stores, limited products available at a reduced price, long lines, and the lack of advertised sale products. 6 Black Friday was originated in Philadelphia around the year 1961. The police described it as the day where there would be heavy pedestrians and vehicular traffic after Thanksgiving Day. Since the 21st century, retailers have been attempts in the US to introduce Black Friday in other countries. Retailers outside the US have promoted black Friday sales to compete with the US retailers in online sales. Black Friday is considered to be the most significant sale that happens in the United States. Black Friday and Cyber Monday combined have a revolutionary history in the shopping industry. However, there have been many changes in the trends and the shoppers. There are many advancements and changes in shopping approaches, and there have also been many factors that influenced the traditional shopping methods. Since this is a digital era, there have been many online shopping activities compared to in-store shopping.

Problem Statement

A Retail Company wants to understand the customer purchase behaviour specifically purchase amount against various products ofdifferent categories.
They have shared purchase summary of various customers for selected high-volume products from last month. The dataset also contains customer demographics (age, gender, marital status, city_type, stay_in_current_city), product details (product_id, product category) and total purchase amount from last month.
Now they want to analyse purchase amount of customer against various products which will help them create personalized offer forcustomers against different products.

Dataset

Source: https://www.kaggle.com/datasets/sdolezel/black-friday

Reason for choosing this industry and dataset

With the holiday season fast approaching, I found it intriguing to examine a dataset revolving around a hypothetical store and data of itsshoppers. As described by the author, "The dataset is comprised of 550,000 observations

about Black Friday shoppers in a retail store, it contains different kinds of variables either numerical or categorical. It contains missingvalues."

Asking right questions is the key part of the analysis process. This will define what you are going to present to the audience.
Answers to the right questions will provide key inputs to thecompany or the audience to improve their business.

If we consider our Black Friday data set the appropriate questions and their results will lead to better functioning of theretail store on their sales.

1. Which age group of customers is more likely to purchase with more amount per

person?

2. Which age group and gender have high visiting rate to the retail store?
3. Which occupation type have high purchase rate?
4. Who has high purchase rate newly settled or people staying from long time?
5. Based on marital status and gender who has high purchase rate?
6. Which product is popular for each age group?
7. What is the purchase percent for each age group and for gender Group in total purchase amount?
8. Which cities spend the most?
9. Which user spends the most in black Friday sales list top 20.
10. Which products are most popular during Black Friday, list the top 20

Available data

This is the current data they have available:

Data made available “ABC Private Limited”

A closer look at the features

1. **User_ID:** A distinct ID is given to the customers to identify them uniquely.
2. **Product_ID:** A distinct ID is given to products to identify them uniquely.
3. **Gender:** M or F can be used as a binary variable.
4. **Age:** Age is given in bins with 6 categories.
5. **Occupation:** The type of occupation a user does, it is already masked.
6. **City_Category:** The category of the city out of A, B, C. Should be used as Categorical Variable.
7. **Stay_In_Current_City_Years:** It has 5 values: 0, 1, 2, 3, 4+ and may be used as categorical variables.
8. **Marital_Status:** 0: Unmarried and 1: Married. It is expected that marital status does affect the Purchase value.

9. **Product_Category_1:** The primary category that a product belongs to. It can be a useful feature as some certain category of products are sold more often than others.

10. **Product_Category_2:** The Secondary category of a product. If there is no secondary category this will be *Null.*

11. **Product_Category_3:** The Tertiary Category of a product. This will be only occupied when Category 1 and 2 are occupied. Also, if a product does not have a tertiary category, it will be *Null.*

12. **Purchase:** This is the target variable.

If we analyse it individually, we see that we do not have any information regarding the stores. Moreover, there is some information related to the customer such as age group, sex, occupation and marital status. On the other hand, we have data

on the city's size and how many years the customer has lived init whereas on the product's side there is only information regarding the categories and the amount spent. It is my belief that Gender , Age , City_Category , Product_Category_1 are the predictors that will influence more the amount spent by a customer on this day.

The target variable is Purchase.

Now that we have understood our data, we can start visualizing and gain some more insights.

Wrangle

Wrangling is the part where we make sure that the data, we collectedfor analysis is of good quality. Here we assess the data quality and clean the data. Wrangling is the part where we take care of missing data, duplicate data, incorrect datatypes etc.,

let us check the basic summary which our dataset tells us.

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
0	1000001	P00069042	F	0-17	10	A	2	0	3	NaN	NaN	8370
1	1000001	P00248942	F	0-17	10	A	2	0	1	6.0	14.0	15200
2	1000001	P00087842	F	0-17	10	A	2	0	12	NaN	NaN	1422
3	1000001	P00085442	F	0-17	10	A	2	0	12	14.0	NaN	1057
4	1000002	P00285442	M	55+	16	C	4+	0	8	NaN	NaN	7969

train.**head()**

```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 550068 entries, 0 to 550067
Data columns (total 12 columns):
User_ID                     550068 non-null int64
Product_ID                  550068 non-null object
Gender                      550068 non-null object
Age                         550068 non-null object
Occupation                  550068 non-null int64
City_Category               550068 non-null object
Stay_In_Current_City_Years  550068 non-null object
Marital_Status              550068 non-null int64
Product_Category_1          550068 non-null int64
Product_Category_2          376430 non-null float64
Product_Category_3          166821 non-null float64
Purchase                    550068 non-null int64
dtypes: float64(2), int64(5), object(5)
memory usage: 50.4+ MB
```

What we can understand from information?

1. Total samples in our dataset are 5,50,068 (no of rows)
2. This dataset has 12 features (no of columns)
3. Only Product Category 2 and 3 have missing values => total noof non-nulls are less than total samples
4. There are 2 features of float type, 5 features of int type, 5features of object type (String)
5. Stay_In_Current_City_Years is a number but the type it isshowing is object (String) let us find why?
6. Age in dataset is range so it is Object (String)
7. Product category 2,3 are shown float let us find why?

```
There are 544177 duplicate IDs for 550068 total entries
```

	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

data.**describe()**

i) Missing data

```
df.isnull().sum() ### gives the sum of missing values with respect to columns
```

```
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        166986
Product_Category_3        373299
Purchase                       0
dtype: int64
```

Only product category 2,3 have missing values. As the details of thesecolumns were not specified *I **assume*** that the particular product doesn't come into any category in that field so let us replace them with
0. (why with 0, in the below code you can check different uniquevalues in that feature)

```
df['Product_Category_2'].unique()  ## gives different values in that column
```

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

```
df['Product_Category_3'].unique()
```

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

```
df.fillna(0,inplace=True) ### replace nan with 0 in that same dataframe
```

ii) **Incorrect datatypes***

All the unique values after handling the missing values in product category 2,3 are integers. But the data type shown in info is float sowe can change it by converting the numbers in float to integers.

```
df.Product_Category_3 = df.Product_Category_3.astype('int64')
df.Product_Category_2 = df.Product_Category_2.astype('int64')
```

```
df["Stay_In_Current_City_Years"].unique()
```

```
array(['2', '4+', '3', '1', '0'], dtype=object)
```

Stay_In_Current_City_Years feature contains few values asstrings("4+") so the data type remains Object.

iii) **Duplicate Values***

```
sum(df.duplicated())
```

```
0
```

So the dataset doesn't contains any duplicated values. It is agood news

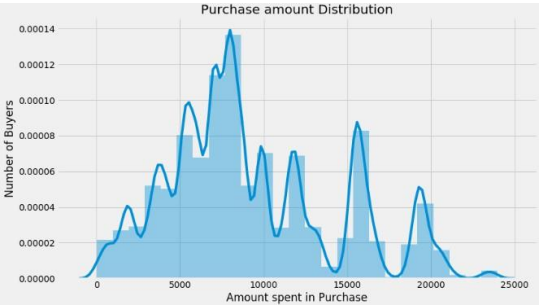
1. Exploratory Data Analysis (EDA)

Univariate Analysis

To get an idea of the distribution of numerical variables, histogramsare an excellent starting point. Let's begin by generating one

for Purchase, our target variable.

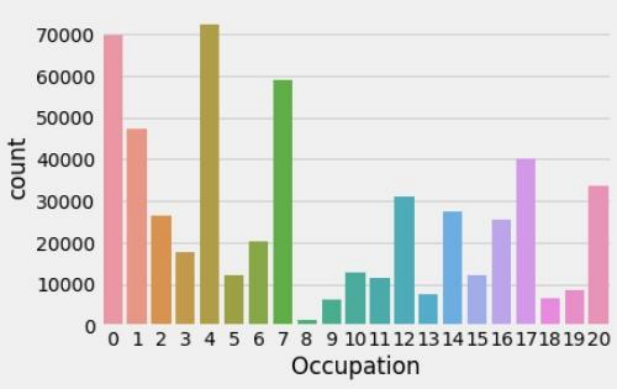
Distribution of the target variable: Purchase



It seems like our target variable has an almost Gaussian distribution.

Distribution of the variable Occupation

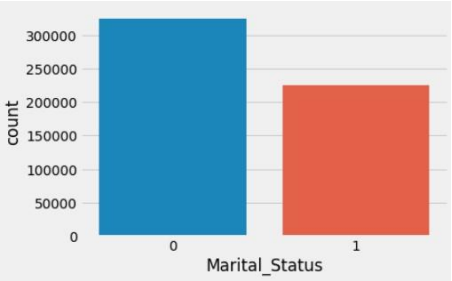
As seen in the beginning, Occupation has at least 20 different values. Since we are not known to each occupation each number corresponds,is difficult to make any analysis.



Distribution of the variable Marital_Status

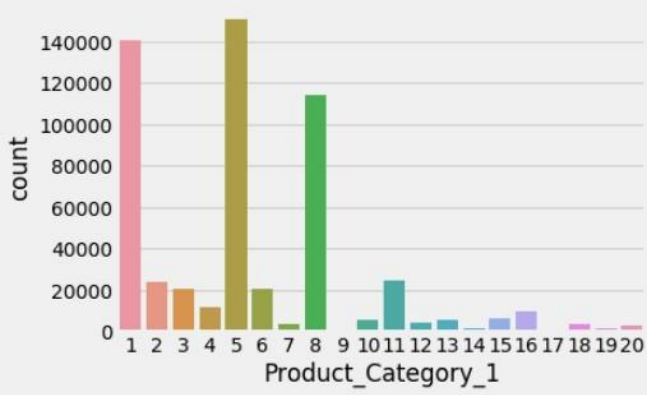
As expected, there are more single people buying products on BlackFriday than married people, but do they spend more?

It is possible that the commodities that married people prefer to buy did not have attractive offers and perhaps the company can work on that in the next sale. It is also possible that couples choose not to fritter away their income in the sale and focus more on themselves and their family.

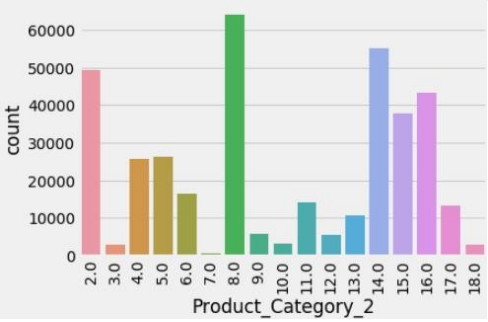


Distribution of the variable Product_Category_1

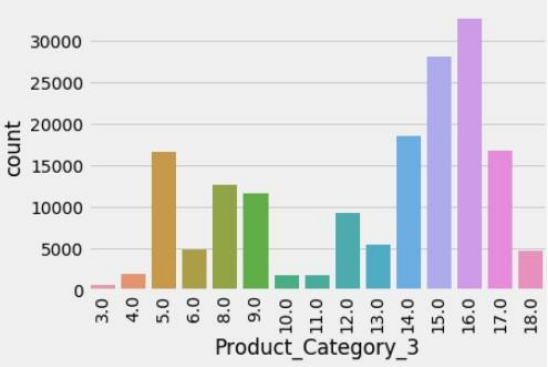
From the distribution for products from category one, it is clear that three products stand out, number 1, 5 and 8. Unfortunately, we do notknow which product each number represents.



Distribution of the variable Product_Category_2



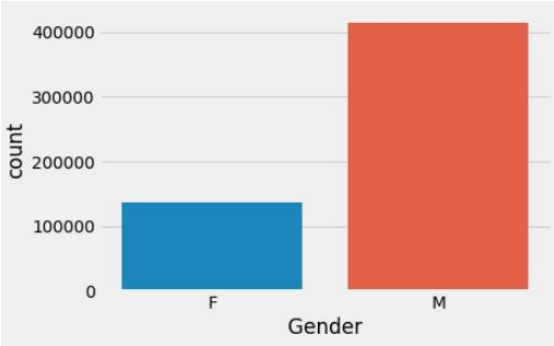
Distribution of the variable Product_Category_3



Distribution of the variable Gender

Most of the buyers are males, but who spends more on each purchase:man or woman?

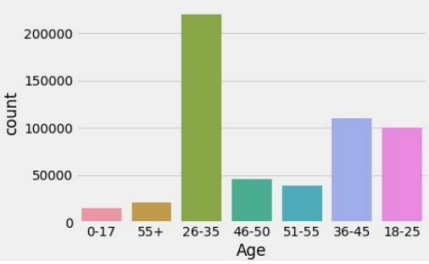
This is a very peculiar observation as one will not expect such great disparity between the genders and the company must get behind the reason why there is this disparity and what can



be done to entice female shoppers.

Distribution of the variable Age

Most purchases are made by people between 18 to 45 years old.



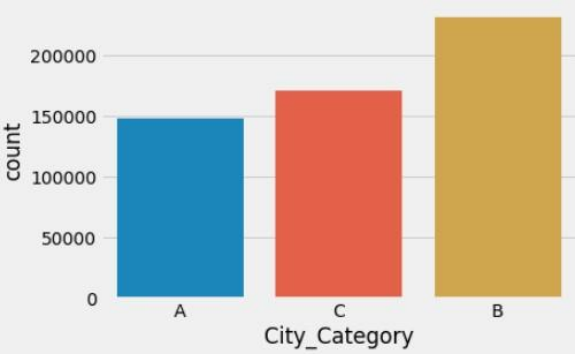
We can therefore infer that people of age group 26–35 shopped the most followed by 36–25, 18–25, 51–55, 55+ and then 0–17.

It is easy to speculate on this data. Since people of age 0–17 are usually dependent on elders, their numbers as customers are the lowest. Also, people of the age group 26–35 are generally independent and have income sources, they make the largest population in our data.

Distribution of the variable City_Category

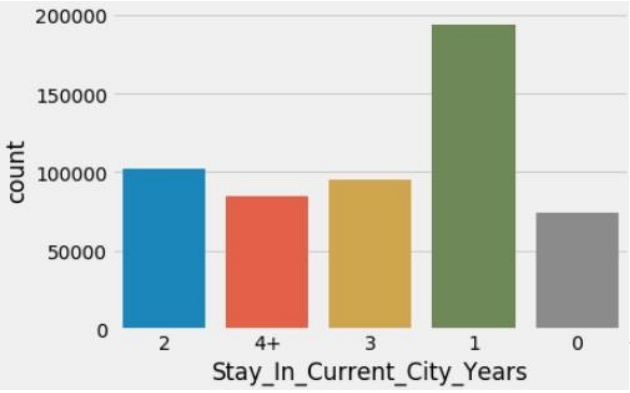
Now we are not sure what these categories mean and on what basis are these categories made. However, we can get some idea after a more detailed analysis. For example, if we assume that cities are divided based on income range of people, so it is possible that high-income people are less interested in the sale and so belong to City A, also, people with low-income will be interested but their hands are tied because of their low remuneration and so can be categorized in City C. People with wages not too high and not too low can freely participate in this sale and so are the major shoppers, hence belong to the City Category B.

But do they also spent more?



Distribution of the variable Stay_In_Current_City_Years

The tendency looks like the longest someone is living in that city theless prone they are to buy new things. Hence, if someone is new in town and needs a great number of new things for their house that they'll take advantage of the low prices in Black Friday to purchase all the things needed.



Bivariate Analysis

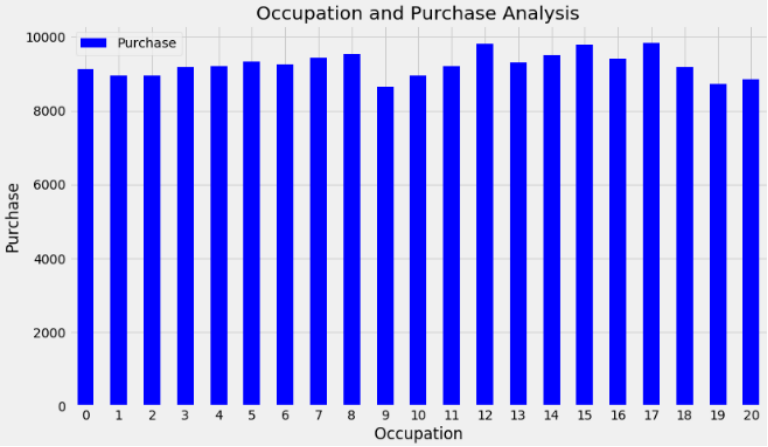
Firstly, we individually analyzed some of the existent features, now itis time to understand the relationship between our target variable andpredictors as well as the relationship among predictors.

Numerical Variables

Occupation and Purchase analysis

It is evident that there is some trend in the occupation-purchase graph. Customers with occupation 17 have spent the most and that with occupation 9 have spent the least.

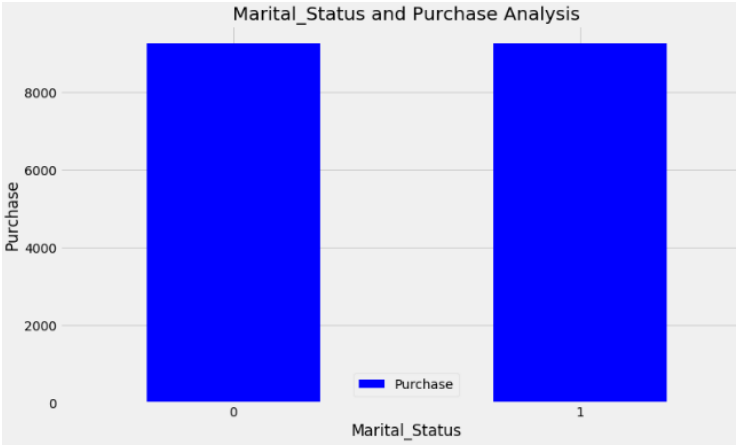
Although there are some occupations which have higher representations, it seems that the amount each user spends on averageis more or less the same for all occupations. Of course, in the end, occupations with the highest representations will have the highestamounts of purchases.



Marital_Status and Purchase analysis

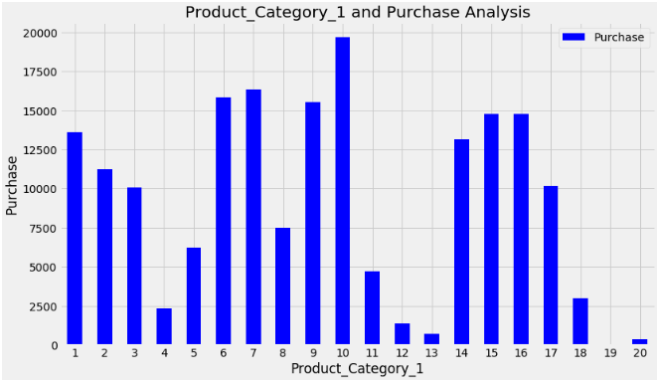
Now, this is an interesting aspect. We had more single customers thanmarried. However, on average an individual customer tends to spend the same amount independently if his/her is married or not. Again, if you had all the purchases the single group, since has a higher representation, will have the highest purchase values.

Although unmarried people did more of the shopping, the average amount spent by both unmarried and married people is nearly the same.

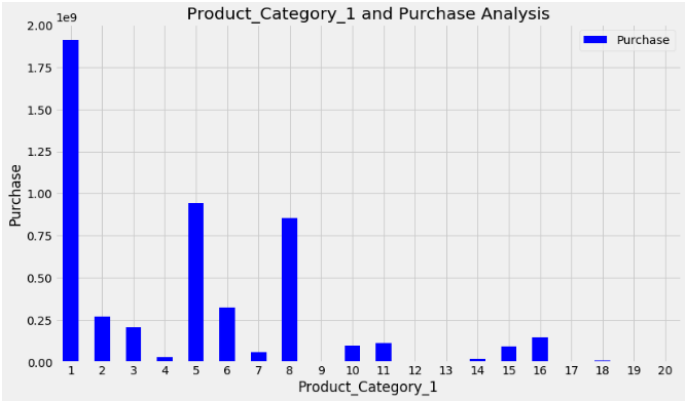


Product_category_1and Purchase analysis

If you see the value spent on average for Product_Category_1 you seethat although there were more products bought for categories 1,5,8 theaverage amount spent for those three is not the highest. It is interesting to see other categories appearing with high purchase values despite having low impact on sales number.



For examples, if instead of the average spent we look at the amountspent on purchase, as illustrated in the chart below, that distributionthat we saw for this predictor previously appears here. For example,those three products have the highest sum of sales since their were three most sold products.

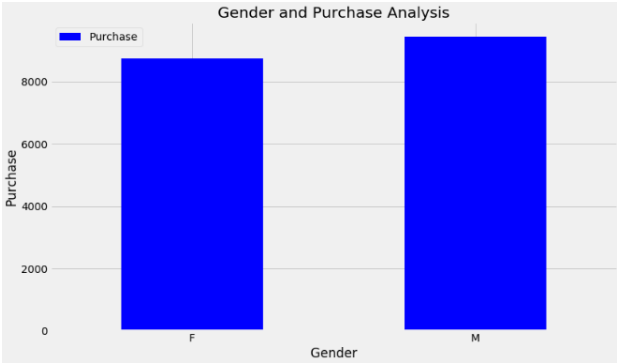


We can see the same behaviour for the other two categories.

Gender and Purchase analysis

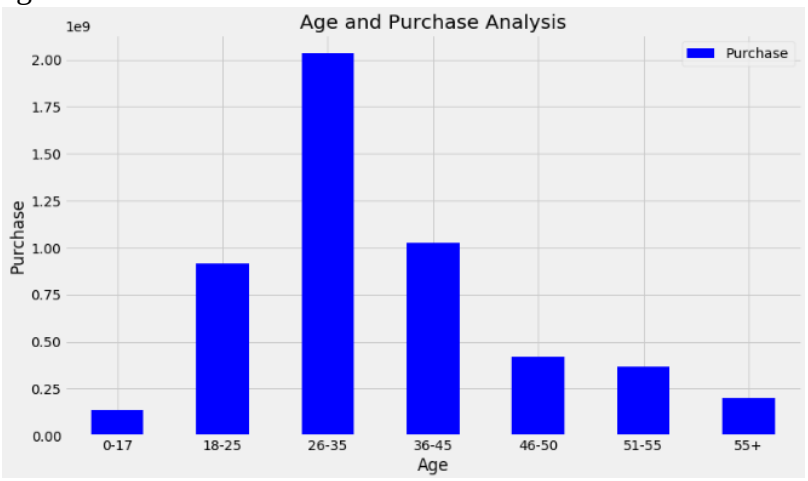
On average the male gender spends more money on purchase contraryto female, and it is possible to also observe this trend by adding the total value of purchase. This last conclusion is more reasonable since the percentage of male buyers is higher than female buyers.

Not only did Males shop more, but also spent more on an average than the Females. Also, the amount spent by an average male is more than an average female, though not by much amount.

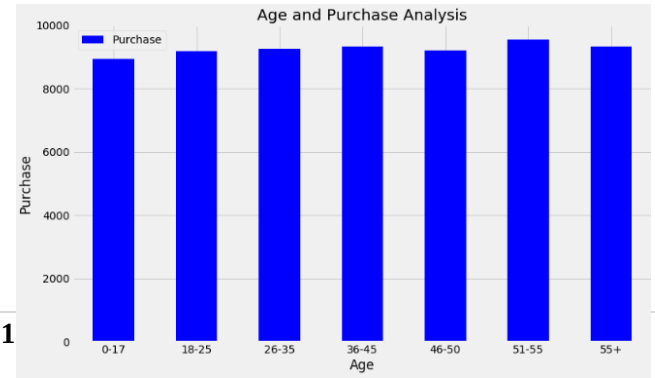


Age and Purchase analysis

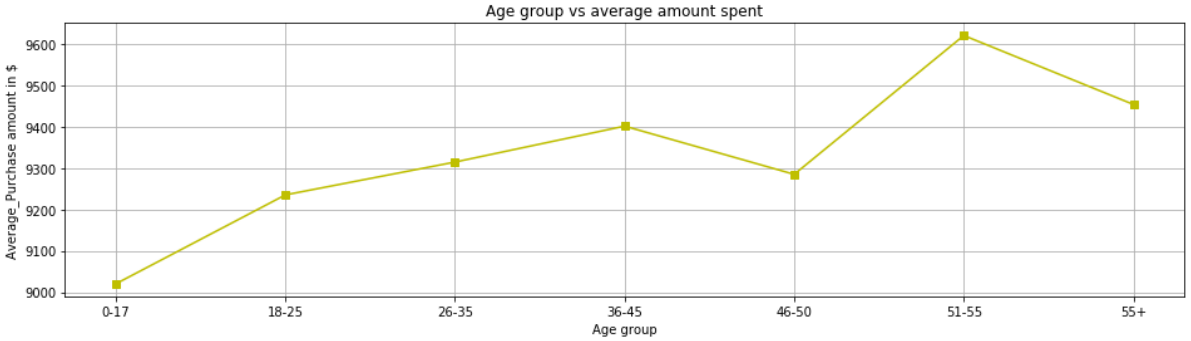
Total amount spent in purchase is in accordance with the number ofpurchases made, distributed by age.



Nevertheless, regarding the average spend by group age we can see the amount spent is almost the same for everyone. Curiously, on average customer with more than 50 years old are the ones who spentthe most. A general reason can be that they don’t need to save up anymore and can freely spend whatever amount they wish.



Having a closer look

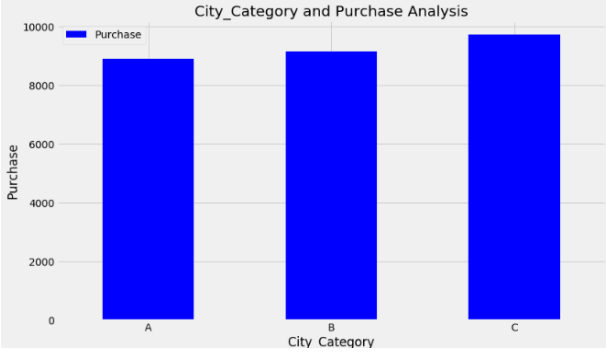


Approximately 9600$ on average spent by People between age 51-55.

The graph values tend to increase so higher the age group higher the interest in the sale. But there is a slight purchase variation in 46-50 and 50-55 age people.

City_Category and Purchase analysis

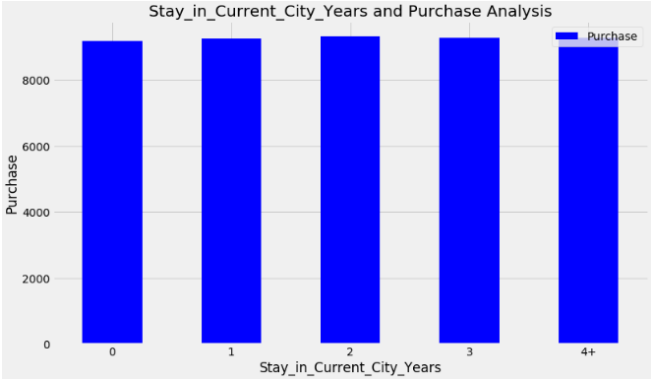
We saw previously that city type ‘B’ had the highest number of purchases registered. However, the city



whose buyers spend the mostis city type ‘C’.

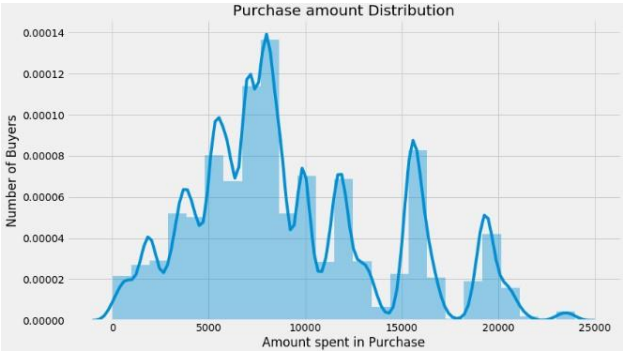
Stay_In_Current_City_Years and Purchase analysis

Again, we see the same pattern seen before which show that on average people tend to spend the same amount on purchases regardless of their group. People who are new in city are responsible for the higher number of purchases, however looking at it individuallythey tend to spend the same amount independently of how many yearsthey have lived in their current city.

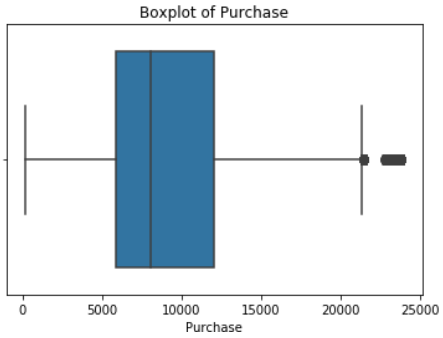


Statistical Analysis

We already observed that purchase amount is repeating for many customers. This may be because on Black Friday many are buying discounted products in large numbers and kind of follows a Gaussian Distribution.




```
sns.boxplot(data["Purchase"])
plt.title("Boxplot of Purchase")
plt.show()
```




```
data["Purchase"].skew()
```

```
0.6242797316083074
```

```
data["Purchase"].kurtosis()
```

```
-0.34312137256836284
```

```
data["Purchase"].describe()
```

```
count    537577.000000
mean       9333.859853
std        4981.022133
min         185.000000
25%        5866.000000
50%        8062.000000
75%       12073.000000
max       23961.000000
Name: Purchase, dtype: float64
```

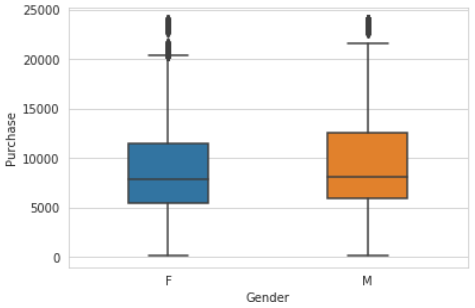
We can observe that there are some outliers above 21000

The purchase is right skewed and we can observe multiple peaks inthe distribution

Plotting relationships between 'Purchase' and various otherattributes in the dataset

1. Purchase vs. Gender

```
sns.boxplot(x='Gender',y='Purchase', data = df, width=0.4)
plt.show()
```



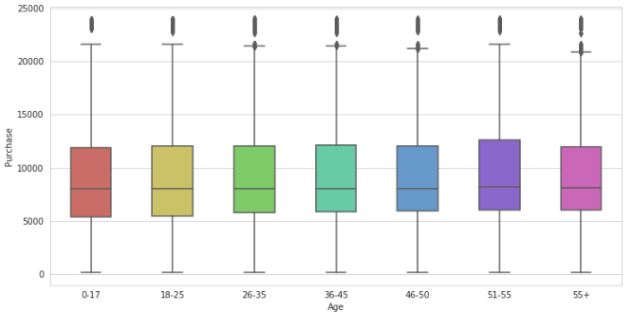

```
df.groupby('Gender').agg({'Purchase':['max','min','mean','median']}).round(3)
```

	Purchase			
	max	min	mean	median
Gender				
F	23959	185	8809.761	7929
M	23961	185	9504.772	8112

It is seen that on an average males spent more money thanfemales shopping on a Black Friday.

2. Purchase vs. Age

```
plt.figure(figsize=(12,6))
sns.boxplot(x = 'Age',y='Purchase', data = df,palette='hls',
            order=['0-17','18-25','26-35','36-45','46-50','51-55','55+'],width=0.5)
plt.show()
```

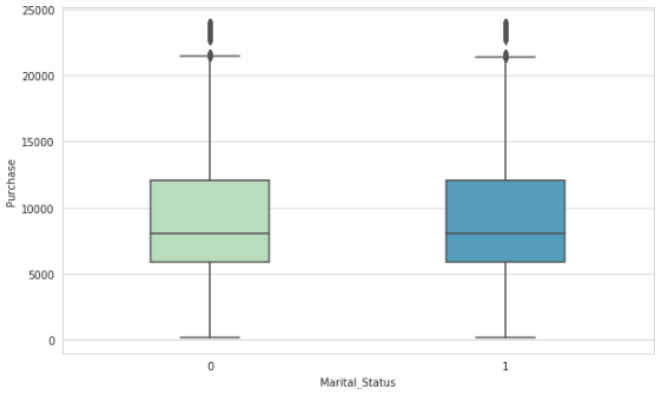


	Purchase		
	min	max	mean
Age			
0-17	187	23955	9020.127
18-25	185	23958	9235.198
26-35	185	23961	9314.589
36-45	185	23960	9401.479
46-50	186	23960	9284.872
51-55	187	23960	9620.617
55+	187	23960	9453.899

From the given data, people from all age groups except people in age range '51-55' spent almost the same average amount on shopping during a black friday. People in range '51-55' had a slightly higher average purchase amount than other age groups.

Category C city made relatively higher amount purchases whichmay mean they like to stalk up on things during sale

3. Purchase vs. Marital_Status

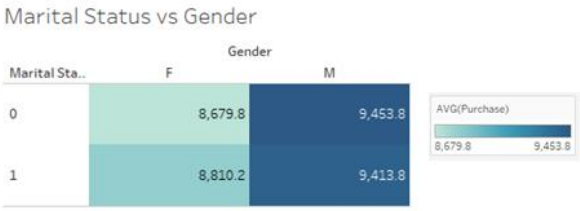


	Purchase		
	min	max	mean
Marital_Status			
0	185	23961	9333.325
1	186	23961	9334.633

We cannot observe any significant difference here.
Both married as well as unmarried people spend about the sameamount of money on shopping.

Multi-Variate Analysis

But there was a variable ‘Marital Status’ for which we could not figure out how it affects our target. So, we will do multivariate analysis and gain a deeper insight.



In *Figure*, we see the comparison of Marital Status and Gender w.r.t Average Purchase. We see that for an unmarried person if they are Female, they spend a lot less than if they are male. The same trend is shown for a married person with just a slight difference.

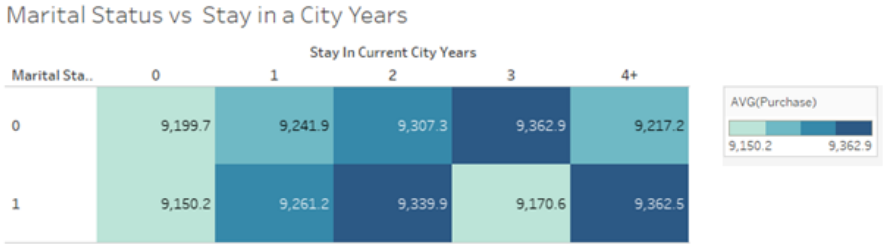
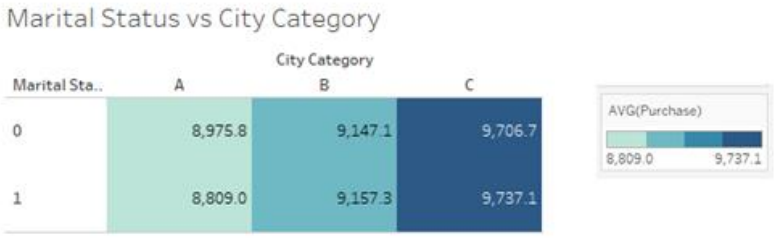


Figure 3.8.2

In *Figure* we see the comparison of Marital Status and Stay in Current City Years (SCCY) w.r.t Average Purchase. For SCCY 0, the value of avg. purchase is nearly the same. For SCCY 1, the value of avg. purchase there is just a slight difference between married and unmarried. For SCCY 2 also there is just a slight difference between married and unmarried. For SCCY 3 however, there is a little more difference between married (9,170.6) and unmarried (9,362.9). For SCCY 4+, there is just a slight difference between married and unmarried.



In *Figure*, we see the comparison of Marital Status and City Category w.r.t Average Purchase. There is not a significant difference between married and unmarried people living in different categories of the city.

From the above analysis, we can conclude that Marital Status does not affect the target much and can be dropped from further use.

Purchase Mean by Age Group and City

If we plot a straightforward visualization to answer this question, we will come up with the graph below.

Even though the plot above gives us the information we want, it's a bit confusing. So, in order to make a clear comparison between those groups we will plot each of them separately.

Now, we can clearly see that, for example, from age 0 to 55 the mean purchase of city C is greater than others. We can also see that city B takes the first spot only in the 55+ age group.

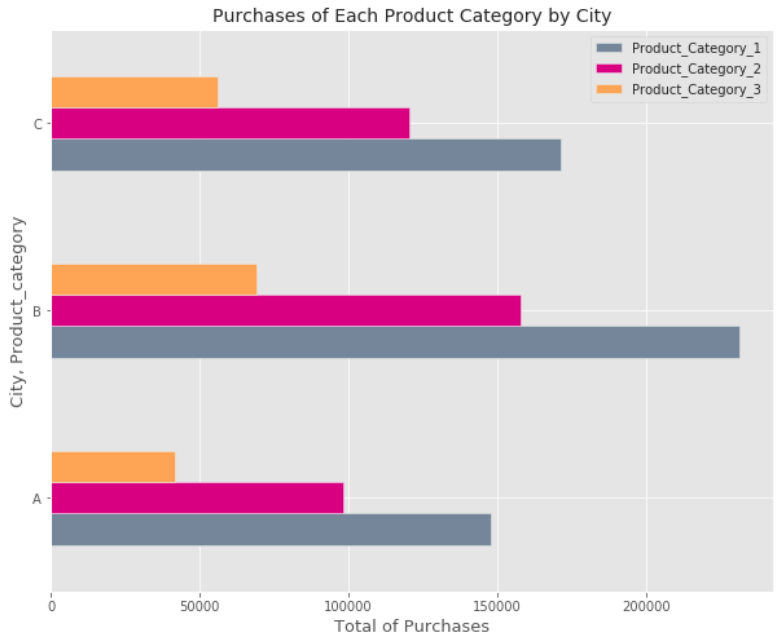
Product Category Analysis

It's important to mention that one product may belongs to different category and groups. For example, the product **P00248942** belongs to all categories but is named as group 1 in the first category, group 6 in the second and 14 in the third one. With that said, let's start our analysis.

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

Number of Purchase of Each Product Category by City:

The chart below shows that we have sort of a pattern in terms of quantity of purchases. As we can see, the product category 1 takes the top spot, followed by product category 2 and product category 3 in all of the cities. What differs each city regarding the number of purchases is the fact that consumers of city B buy more product of all category than city A and C.



Purchase Mean of Product_Category 1 by City

Well, we won't be scrutinizing all those plots. Instead, let's just focus on the main information we can get from them. That being said, the first thing that catches my eye is the product 19 with its way lower purchase mean compared to the other products (less than 40 USD on average). In addition, we can notice that product 4 has the greatest purchase mean with city C taking the top spot followed by city B which has a slight advantage over city C.

Purchase Mean of Product_Category 2 by City

As mentioned in the description of the data, the product may belong to other category and we can actually notice this by looking at the plots below. We have a group of 17 products that also belong to the first category and, as can be seen, product 10 has the greatest purchase mean. Once again, city C takes the first position, followed by cities B and A (similar to product category 1).

Purchase Mean of Product_Category 3 by City

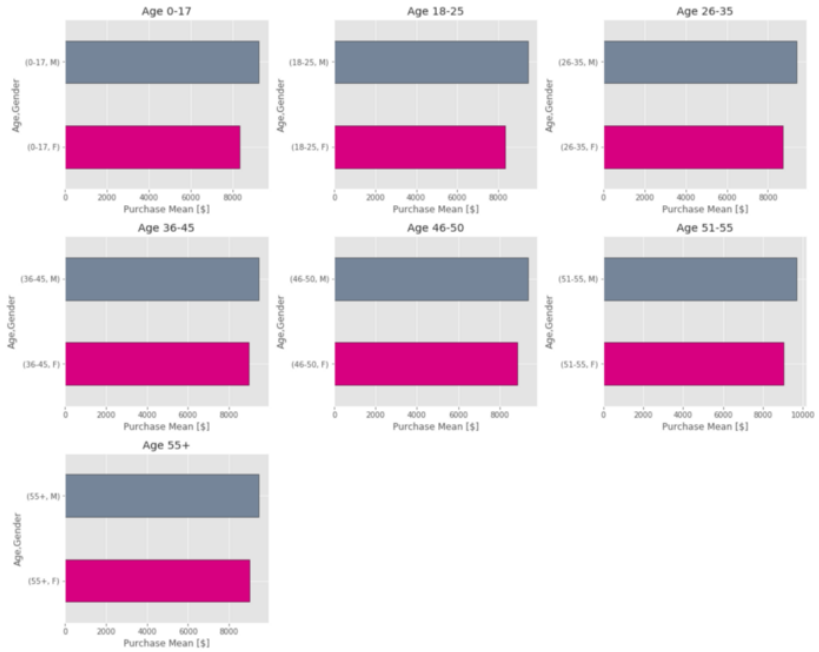
We can see that there are 15 products in this category and product 3 has the greatest purchase mean in all of the cities.

Gender Analysis

Firstly, we will plot the purchase Mean by age group, and then, we will see how men and women differ from each other in terms of product acquisition in cities A, B, and C.

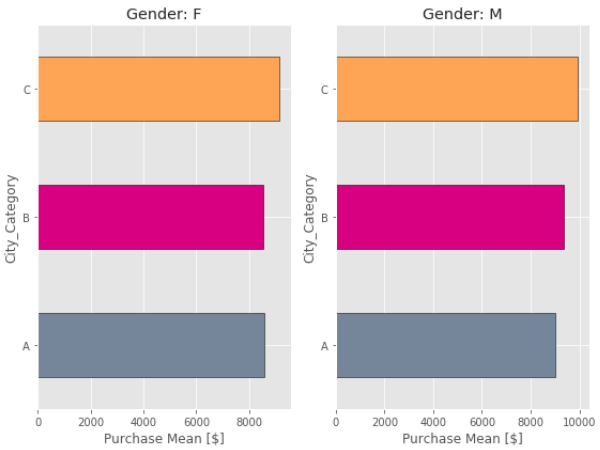
Purchases Mean by Gender and Age Group:

With the purpose of comparing the purchase pattern of different age groups and genders, we can look at the bar charts below. They show that the purchase mean of the men's group is more than 1000 USD greatest than the purchase mean of the women's group for customers aged from 18 to 25. We can also see that male consumer take the first spot in all age groups, though the difference is not so noticeable as in the 18–25 age group.



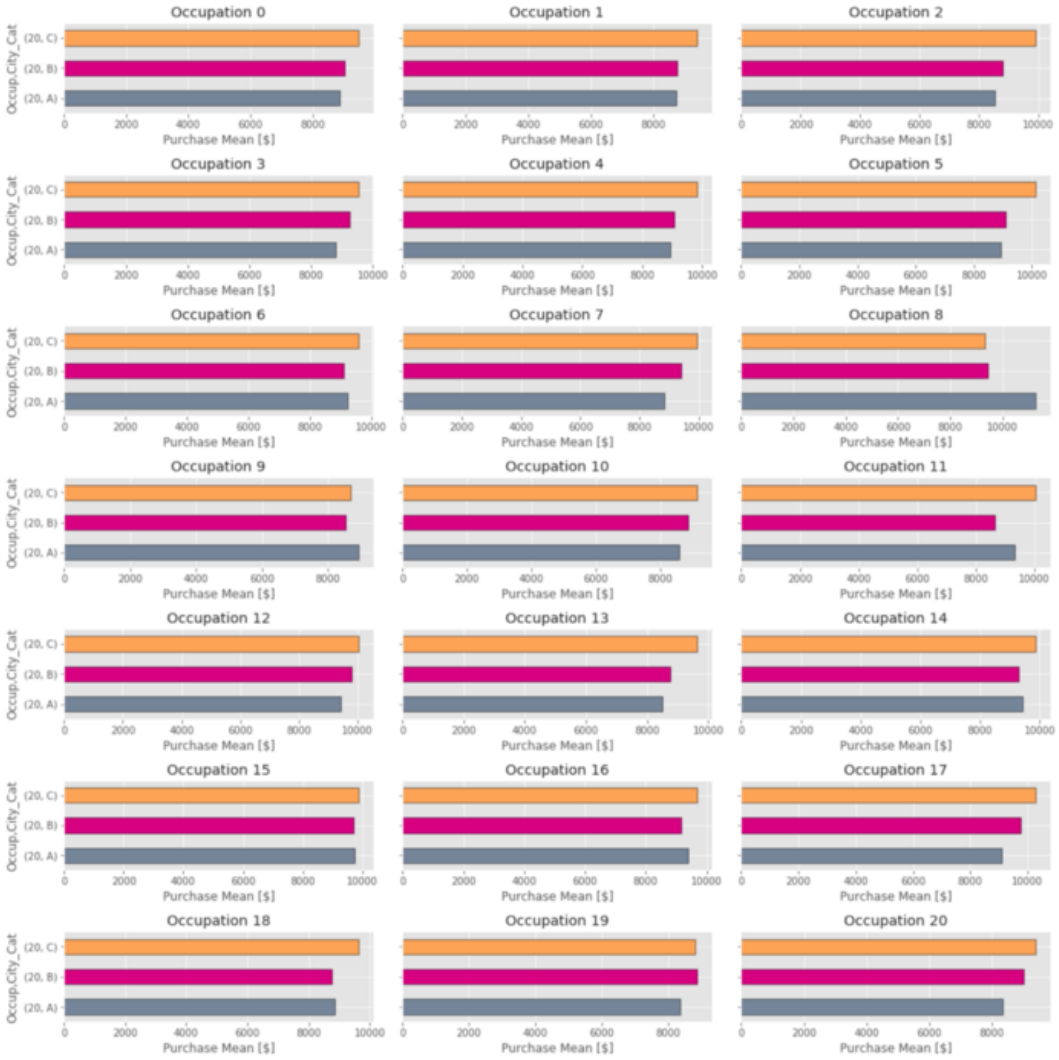
Purchase Mean by Gender and City:

The last gender analysis consists in understanding how purchases were made for both genders in each city. In order to achieve that, let's take a moment to analyze the bar charts below that depict exactly what we are looking for. They show that males spend more money than women in all of the cities (the difference is significant only in city C) and that city C takes the first spot in regards to purchase means for both genders. It is also worth noting that men based in city B bought more than those based in city A, whereas women from city A acquired more than those who live in city B.



Occupation by City:

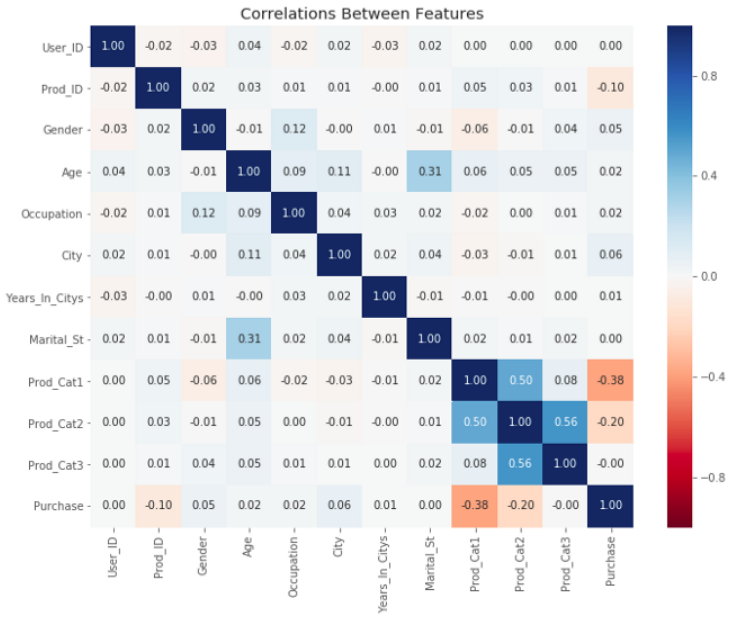
The bar charts below depict the purchase mean of each occupation in each city. It's interesting that city C has the greatest purchase means in almost all of the occupations. We see a change in this pattern only with occupations 8, 9, and 19. Another interesting fact is that we have the greatest amount of money (overall) spent in city A by people which the number 8 is their occupation.



HeatMap:

Basically, Heatmap is a graphical data representation that uses a scheme of color-coding to represent different values. In our case, we are generating a heatmap of correlations between features of our dataset.

From the correlation heatmap , we can conclude that Gender & City_Category are most positive related to Purchase comparing to other features, while all Product_Category features are negative related to Purchase. Marital_Status, Stay_In_Current_City_Years are not so important features that relate to Purchase. All three Product_Category are highly correlated to each other. Besides that, we can also find that Martital_Status are highly related to Age, which is quite reasonable.



Result

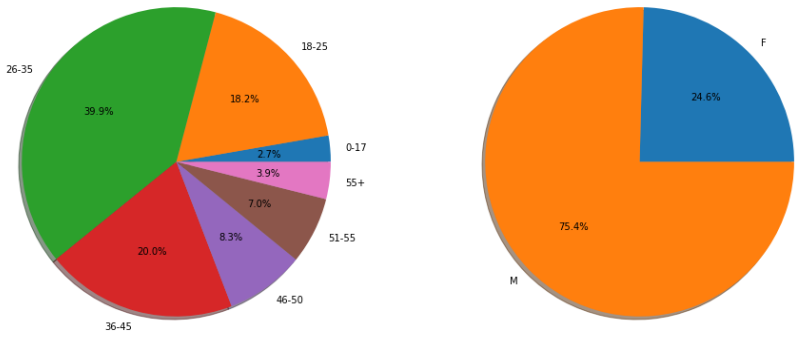
With this we come to the end of the analysis of Black Friday Sales

Let us see if we are able to answer the questions asked in the starting

1. Which age group of customers is more likely to purchase with more amount per person?

   Answer: People of Age group 51-55 have spent more on purchase. Approximately 9600$ on average spent by People between age 51-55.

2. Which age group and gender have high visiting rate to the retail store?



If we check the second pie, By this stat we can tell that the store gets most of the male customers (75.4% male 24.6% Female).

This shows 40% of customers are 26-35 age group and 20% are from 36-45 => 60% of customers from 26-45 age group.

only 7% of customers are of 51-55 Age group.

From 1st and 2nd questions we can tell 60% of customers from 26-45 who have a

medium purchase rate. 7% of customers are from 51-55 who have high purchase rate. This gives an interesting insight on sales to store owners.

3. Which occupation type have high purchase rate?

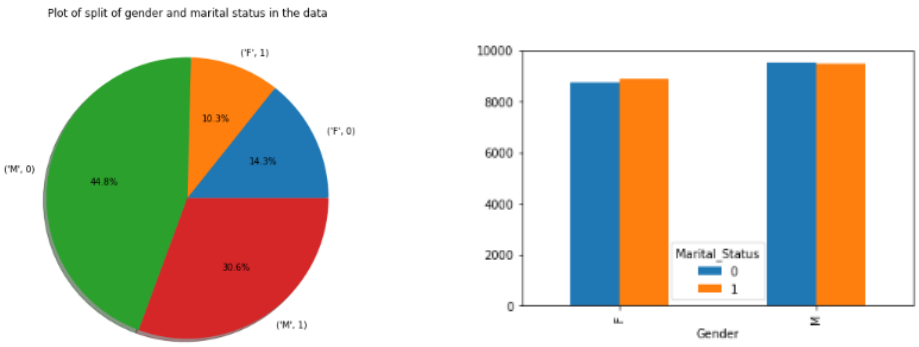
   A few points are observed while analysing Occupation variable alone and with average purchase amount:

   1. Occupation type 12,15,17 have high purchase rates but no of people with those occupations are not in large amount.
   2. Occupations 0,4 and 7 have highest amount of people but their average amount of purchases are less.
   3. If we observe occupation 8 and 9 no of people in 8 and lesser than in 9 but the average spending of occupation 8 is roughly 800 dollars more than occupation 9.
   4. If we observe 11 and 12 total people and average spending both are more for 12. So no of people is not correlated with purchase.
   5. Occupation 8 which does not even have 10,000 no of people have average spending just 300 dollars less than occupation 17 with 40,000 people. So occupation 8 looks like heavy spenders.
   6. Occupation 4 which has highest amount of people has spending 600 dollars less than the highest Spender.
   7. Occupation 1 has 20,000 more people than occupation 2 but they seem to have roughly same amount of spending.

4. Who has high purchase rate newly settled or people staying from long time?

   People who are 2 years residents spent more on average in the sale. Even though the 1 year residents visited more in sale they have not spent much but 2 years residents who are around 100,000 people visited, each have spent 9400$ on average.

5. Based on marital status and gender who has high purchase rate?



   Males tend to purchase more and unmarried Males are around 45% in the data and they show to purchase 9000$ on average.

6. Which product is popular for each age group?

	Age	Product_ID
0	0-17	P00255842
1	18-25	P00265242
2	26-35	P00265242
3	36-45	P00025442
4	46-50	P00265242
5	51-55	P00265242
6	55+	P00265242

The product P00265242 has attracted most age groups. From 18-35,46-55+ all of them have this product has highest purchased in their age group. For Age group 0-17 P00255842 is more frequently purchased and for 36-45 P00025442 is the most frequently purchased.

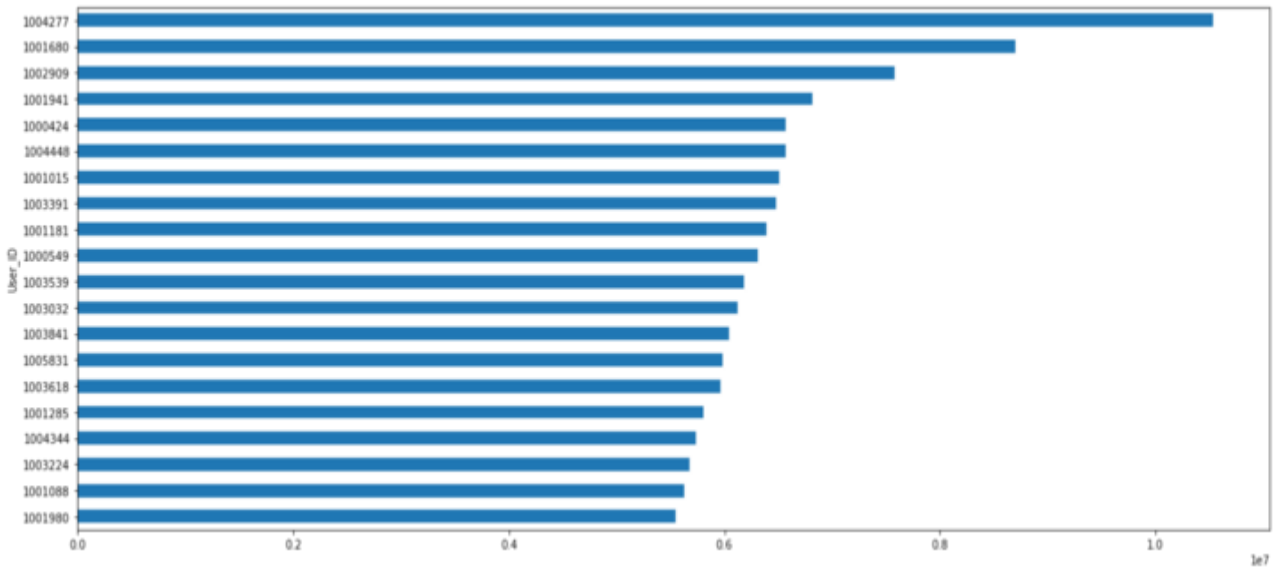
7. What is the purchase percent for each age group and for gender Group in total purchase amount?

   It looks like count of people in different age groups in data is in correlation with total percent of amount spent. Similarly with gender males were 75% their spendings in total is 76.8%, females were 25% their spendings in total is 23.2%.

8. Which cities spend the most?

   The people living in city category B made the most number of purchases but the amount spent is the most in case of city category C.

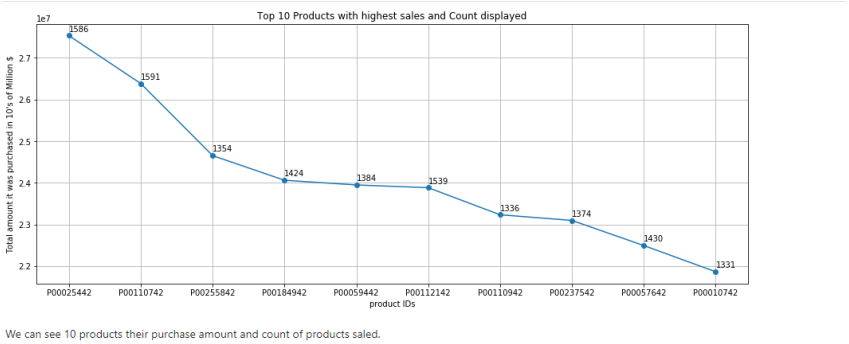
1. Which user spends the most in black Friday sales list top 20.



It's important for the seller to identify high quality customers. These customers with higher purchase amount should be valued. Understanding the needs of these customers will help the merchant to make more suitable operational decisions, such as product type, pricing,

after-sales, etc. Loyalty program, advertisements should be made to keep these customers continuing shopping with the merchant.

9. Which products are most popular during Black Friday, list the top 20



We can see 10 products their purchase amount and count of products saled.

We can see 10 products their purchase amount and count of products saled.

1. 1st product has 1586 pieces sold with total sale of 27.5 million and 2nd product with 1591 pieces sold but with 26.5 million. which means 1st product might have higher product cost.
2. 3rd product has 1354 units sold and 4th product has 1424 units sold but 4th product has low price than 3rd product so even it has higher products sold it has lesser sale amount than 3rd.

Similarly, we can observe for all the products.

Conclusion

From the questions and Solutions lets write a summary of our findings.

Findings

- People of Age group 51-55 have purchased with high amount per person (9600 dollars per person).
- 75% of total people visited were Male and 60% of total people were between Age 26-45.
- People from Age group 26-35 collectively have spent more amount (40% of sale purchase is from this group).
- P00265242 was the product which attracted most of the adults and P00255842 attracted 0-17 Age group.
- Unmarried Male who are 45% in the dataset have spent 9000 dollars per person.
- Even though less no of customers are of Occupation 12,15,17 the spend more roughly 9800 dollars per person.
- Highest No of customers are from Occupation 0,4 and 7.
- High no of customers is of newly settled people but customers who are 2 years residents have spent 9400 dollars per person.
- Product P00025442 has got highest total sale amount of about 27.5 million but it is not the highest repeated product in sale P00265242 was highest repeated with 1858 times (1858 customers have bought this).

References

1 http://google.com/

2 https://www.kaggle.com/

3 https://github.com/

4 https://www.youtube.com/