

Milestone Report

Step 1: Data Wrangling

Data acquisition:

Data has been provided and is taken from Kaggle competition at the below location:

<https://www.kaggle.com/mehdidag/black-friday>. There is a file 'BlackFriday' of size ~25 Mb containing shopping patterns of a store on black Friday with 537,577 records which is made available for the task at hand.

Data type and null values:

From the total of 12 columns, 2 columns have null values 'Product_Category_2' and 'Product_Category_3'. These product categories have float64 as type compared to int64 of 'Product_Category_1' column.

```
# Lets see the data types and null values
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 537577 entries, 0 to 537576
Data columns (total 12 columns):
User_ID                537577 non-null int64
Product_ID             537577 non-null object
Gender                 537577 non-null object
Age                    537577 non-null object
Occupation              537577 non-null int64
City_Category          537577 non-null object
Stay_In_Current_City_Years  537577 non-null object
Marital_Status          537577 non-null int64
Product_Category_1      537577 non-null int64
Product_Category_2      370591 non-null float64
Product_Category_3      164278 non-null float64
Purchase                537577 non-null int64
dtypes: float64(2), int64(5), object(5)
memory usage: 49.2+ MB
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 537577 entries, 0 to 537576
Data columns (total 13 columns):
User_ID                537577 non-null int64
Product_ID             537577 non-null object
Gender                 537577 non-null object
Age                    537577 non-null object
Occupation              537577 non-null int64
City_Category          537577 non-null object
Stay_In_Current_City_Years  537577 non-null object
Marital_Status          537577 non-null int64
Purchase                537577 non-null int64
ProdCat1                537577 non-null int64
ProdCat2                537577 non-null int32
ProdCat3                537577 non-null int32
ProdCombo               537577 non-null object
dtypes: int32(2), int64(5), object(6)
memory usage: 49.2+ MB
```

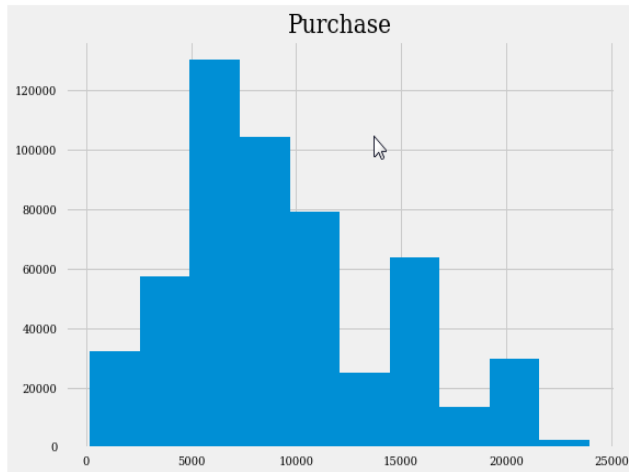
Before

After

1. First, let's handle the null values. Upon checking the null columns, it is found that none of them have the value 0 as product category. So, a decision to replace the null values to 0 is taken.
2. Next, the columns of Product_Category_* are renamed for ease of understanding to ProdCat* and during the process, Product_Category_2 and Product_Category_3 are converted from float64 to int64 type. The original Product_Category_* columns are dropped.
3. We checked if column User_ID represents a unique person by grouping by User_ID and Gender. The count of the unique User_ID was compared to the group by count and it matched. It can therefore be concluded that User_ID is associated with a unique person.
4. We can now check if the Product_ID – User_ID is a unique combination by same means. It can be seen that group by count also matches with the total number of records. It can therefore be concluded that Product_ID – User_ID is a unique key for the dataset.
5. We can now investigate if the product combination is related to a unique product ID. We do a group by of the ProdCat* columns and compare the counts to that of Product_ID. They do not match. Therefore, Product_ID does not seem to be directly related to the combination of categories.
6. The above point can be expanded to the ordering of ProdCat* columns can be permanently captured in a separate called 'ProdCombo'. This order is compared to that of the Product_ID and again there doesn't seem to be any specific relation. It can also be said that the same combo of records could exist with a different Product_ID.
7. Let's check out some rows and confirm this. It can be seen that for User_ID 1001015, the same ProdCombo has 3 different Product_IDs.

Step 2: Exploratory Data Analysis - Inferential Statistics

Data Visualization:



❖ Looking at the spread of the data by doing a Histogram, it can be inferred that most of the purchases are between \$3,000 and \$12,000 and around \$16,000 mark.

❖ A count-plot is used to see how the data is spread amongst each of the individual variables of Gender, Marital_Status, City_Category, Stay_In_Current_City_Years, Age and Occupation. Following inferences can be made:

1. Men shopped 3 times as much as women. This could be due to the fact that men paid for the purchases too.
2. Singles shopped more times than married ones. May be they have more disposable income.
3. Maximum number of purchases were made by people from City B, and people from cities A and C shop

approximately the same number of times.

4. People who have been in the city for a year shop most frequently.
5. People between the age group of 26-35 shop the most followed by 36-45 and 18-25 groups
6. People in occupations 0, 4, 7 shop the most and 8 shop the least.

❖ Looking at the count plots of Product categories, these observations can be made:

1. It can be seen that products 1, 5 and 8 from Category 1 dominate the sales numbers by far and seem to be the favorites.
2. There are a large number of products in Categories 2 and 3 that are either missing or not shopped for. Products in category 3 have very low numbers.
3. Although Category 2 is not all that popular, there are some products that seems to be popular. Such as 2,8,14 and 16.

❖ Next we look at the influence of the same variables on total purchase value to see how it compares.

Comparing this to the counts plots, it is now clear that the distribution in amounts is sort of similar to the distribution in counts. This means that we cannot see any variable disproportionately affecting the purchase. For example, it could have been possible that women bought all the expensive products and even though they shopped less number of times, they could have generated the most revenue. The two plots seems somewhat similar and we can conclude that no variable/distribution needs special attention.

❖ By looking at the total purchase value breakdown by product category, it is clear that there could be lot of hidden information from various combination of the categories. So we ask 2 questions and create a master plot of top 50 product combinations to uncover more info. But this needs to be looked in conjunction with the how other variables influence these 50 combos.

1. What are the characteristics of my top revenue generators or who is my target audience?

Driven by data of top 50 combos, it can be inferred that:

- a) Single males buy the most
- b) These single males are mainly from city B, in age group of 18-35. They are followed by city A in age groups of 18-25 and 25-36.
- c) A disproportionate number of these people are in Occupation 4, mainly buying from city B and A.
- d) People who mainly buy ProdCombos 8-0-0 and 5-0-0 have been living in the city around 1 year.

Answer: The target/valuable audience for Black Friday sale are single males from cities A and B in the age group of 18-35 in occupation 4 who have been living in their cities for around 1 year.

2. Who buys the top product combinations?

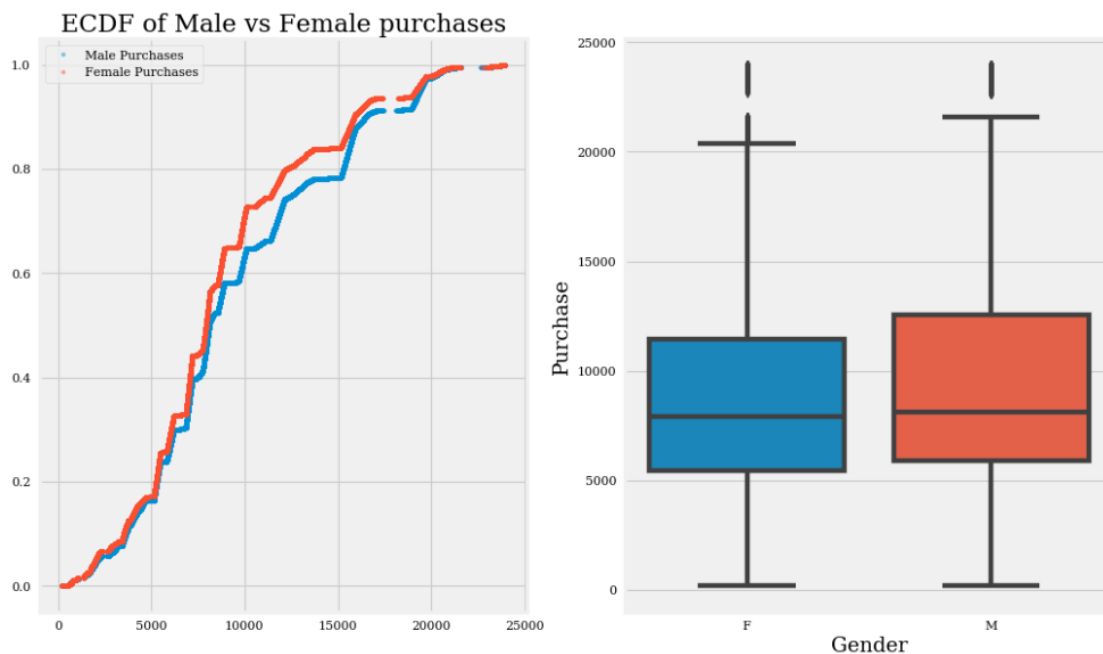
For this analysis, we restrict the top 50 to further of top 20 product combinations to get some meaningful conclusion. In the top 20 combos, you can now confirm that product 1 in Category 1 still has the most number of combos and generates the highest revenue, followed by 8 and 5. In terms of combos across all three categories, 1-2-15, 1-15-16 and 1-2-5 are frequently bought together. Overall, assuming that these 3 categories have independent products, the top products are:

- a) Category 1: 1,8,5
- b) Category 2: 8,2,16
- c) Category 3: 15,5,16

Inferential Statistics: To find out hidden relationship, we ask some interesting questions.

Question 1: Do men buy more expensive things than women?

To find out this, we plot ECDF of Male vs Female Purchases and compliment that with a box plot.



It seems like men tend to make more expensive purchases by looking at ECDF. Expensive purchase = \$10,000.

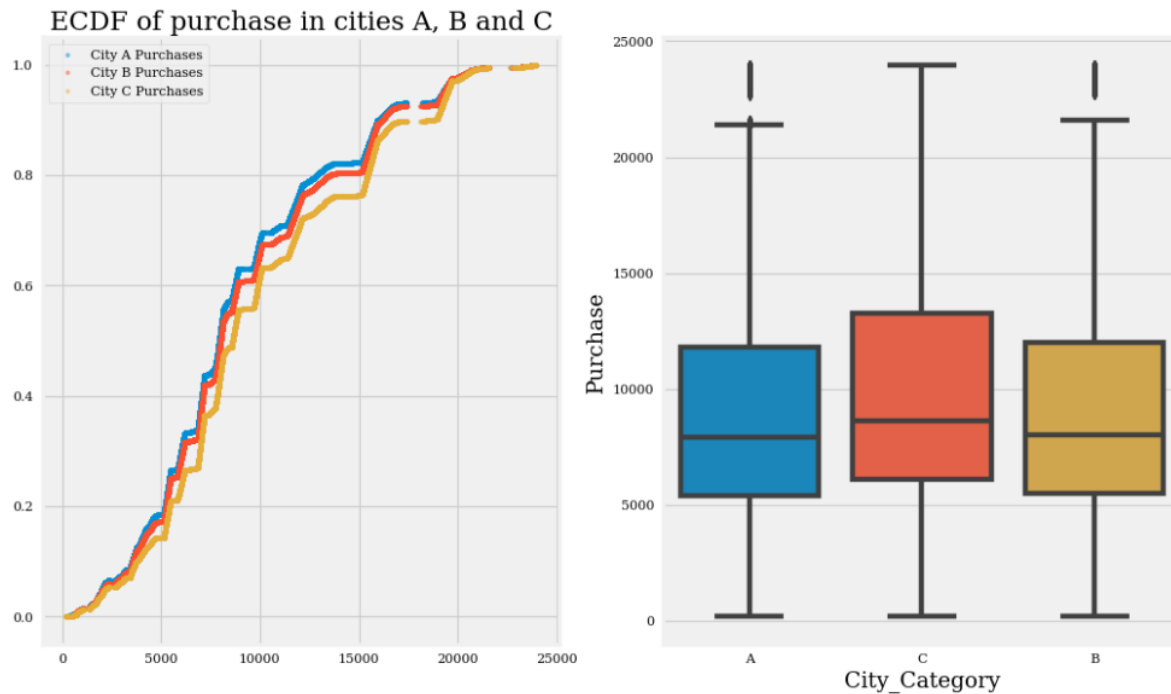
- ❖ **H0:** The average purchase price by men is the same as women for expensive items.
- ❖ **H1:** The average purchase price by men is more than women for expensive items.

Findings:

- Mean of men's purchase is: 15068.7163
- Mean of women's purchase is: 14931.9739
- Mean of total purchase is: 15082.73146804466
- The probability that average purchases by men is equal to that of women is 0.002
- This negates the null hypothesis. Meaning the average purchase price by men is more than women for expensive items.

Question 2: Do people shop equally in all three cities?

To find out this, we plot ECDF of Purchases in cities A, B and C along with a box plot.



From the ECDF and box plots, we can see that in general people from city C generate more revenue. Whereas people from city A and B seems to have similar buying patterns. But does it mean people from city C also spend more? We will take the "Average Purchase value" as the measure for this. Let's see the statistical significance of this happening.

- ❖ **H0:** The average purchase price by people in cities A, B and C is the same.
- ❖ **H1:** The average purchase price by people in cities A, B and C is NOT the same.

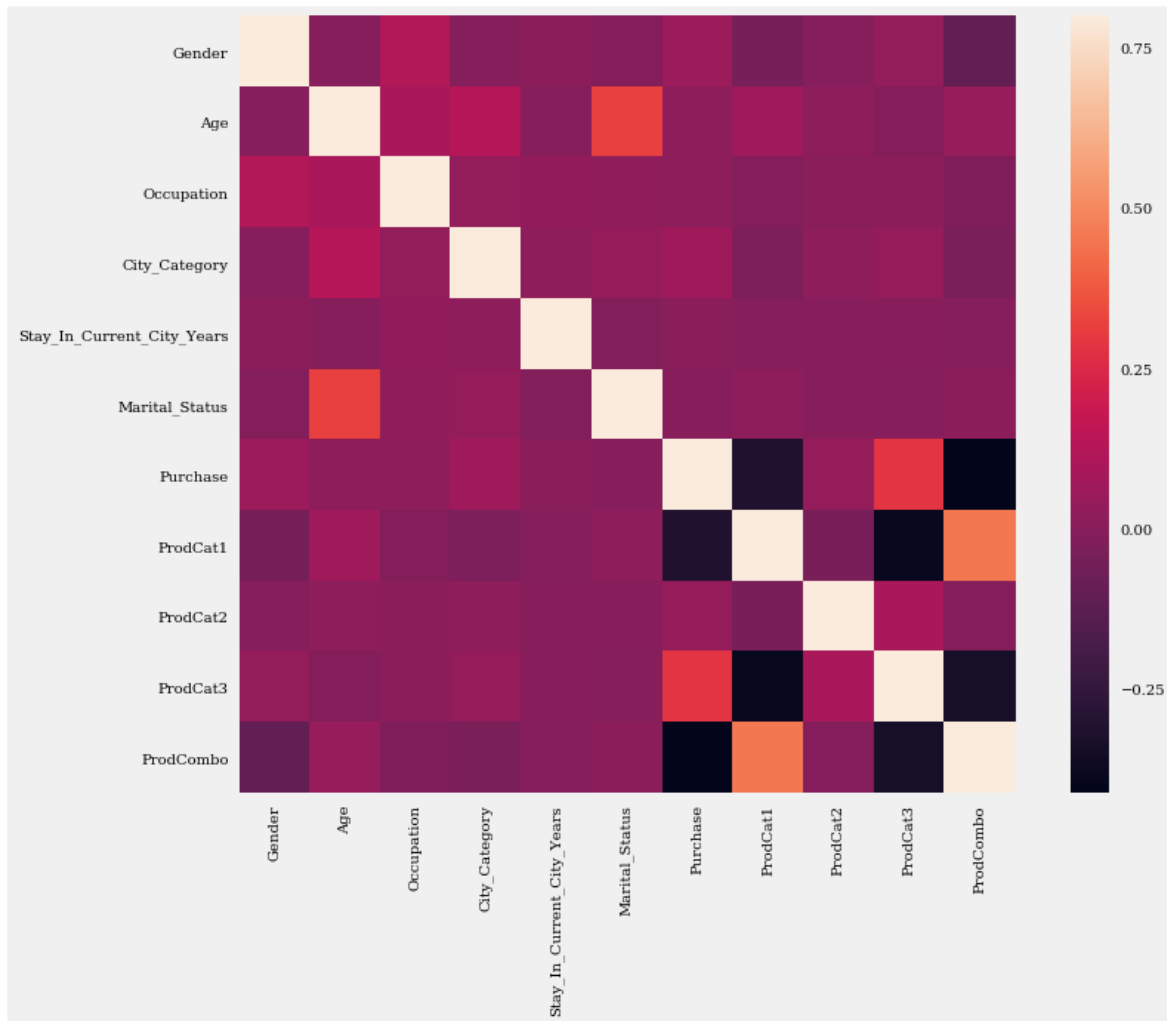
Findings:

The mean purchase value by people from city A is: 9131.392

- The mean purchase value by people from city B is: 9366.509
- The mean purchase value by people from city C is: 9792.204
- The probability that average purchases from City A are equal to City C are: 0.001
- The probability that average purchases from City A are equal to City B are: 0.141
- The probability that average purchases from City B are equal to City C are: 0.028
- This negates the null hypothesis. Meaning, the average purchase price by people in cities A, B and C is NOT the same.

Correlations

Let's see if there are any correlations in the variables. In order to do that, let's plot them in a heat map.



Findings: Marital status seems to have strong positive correlation with Age. And ProdCat1 seems to be strongly correlated with ProdCombo, but this does not mean much. It could be possible that there are a lot of products from Category1 in ProdCombo column. There is slight negative correlation between Purchase and ProdCat3 and ProdCombo, which means there are a lot fewer items of category 3. Other than that, there doesn't seem to be any strong correlation between the variables.