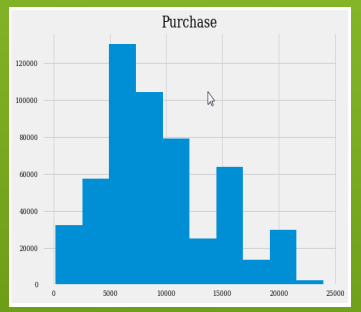
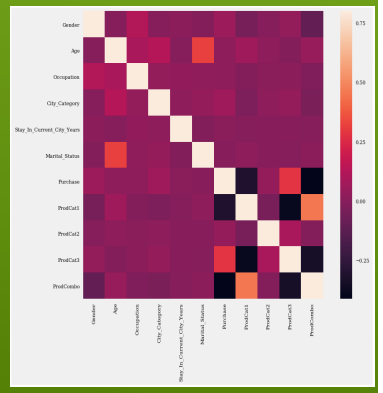
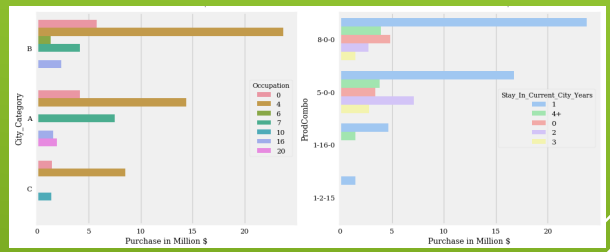
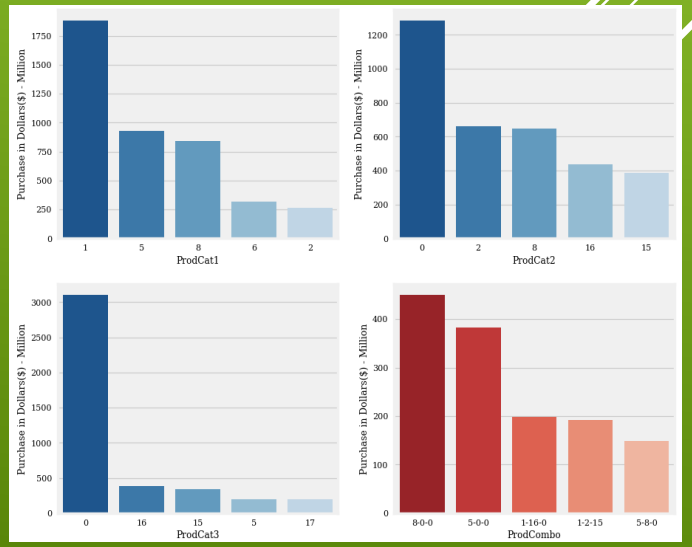
BLACK FRIDAY SHOPPING PREDICTION









Rishi Phule

Problem Statement

Kaggle dataset - a sample of the transactions made in a retail store

Store wants to know better the customer purchase behavior against different products

Try to answer:

- What are the characteristics of my top revenue generators?
- Who buys the top products?
- To what accuracy can we predict purchase amount

Regression Problem – Trying to predict dependent variable – Purchase Amount

Step 1: Data Wrangling

- Fixing Data type and Null values – Selecting 0 for Product_Category_2 & Product_Category_3.
- Rename columns
- User_ID represents a unique person – Group by User_ID and Gender, compare counts
- Product_ID – User_ID is a unique combination – No observation repetition
- Product_ID has no relationship with Product_Categories.

	User_ID	Product_ID	Gender	Age	ProdCat1	ProdCat2	ProdCat3	Occupation	City_Category	Stay_In_Current_City_Years
6243	1001015	P00226142	M	36-45	8	9	0	3	A	4+
396201	1001015	P00190642	M	36-45	8	9	0	3	A	4+
435036	1001015	P00367442	M	36-45	8	9	0	3	A	4+

Before

```
# 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
```

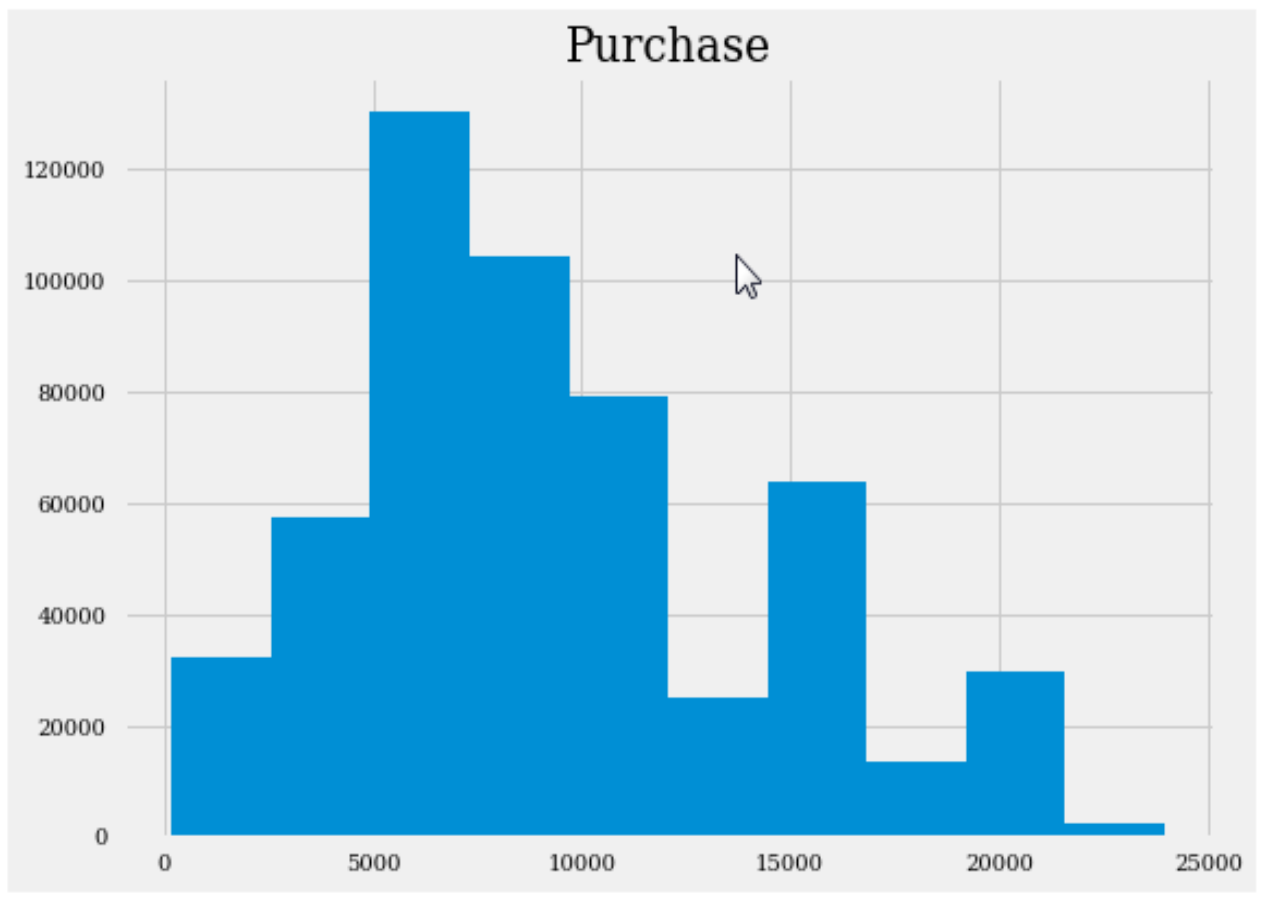
After

```
data.info()
```

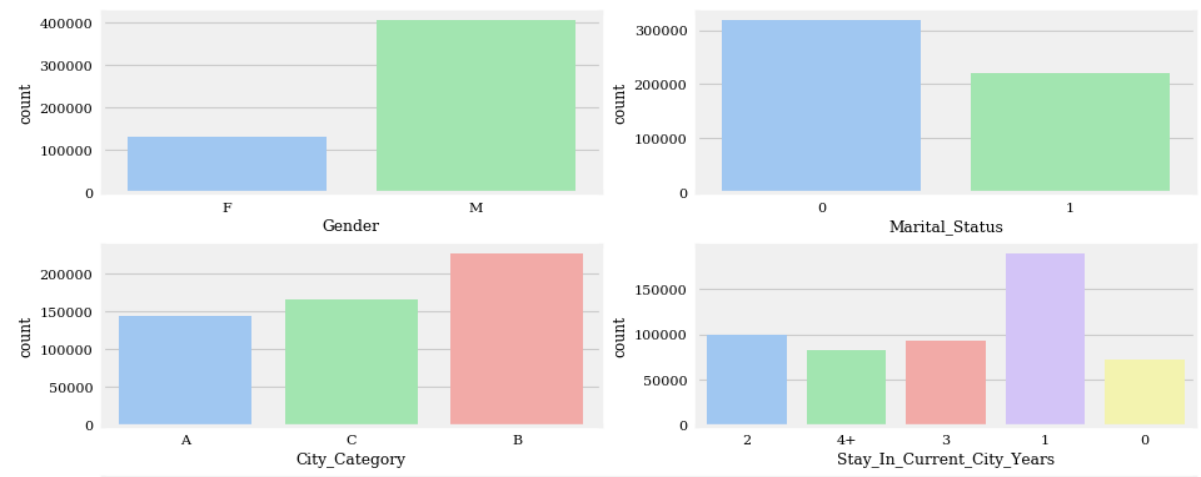
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 537577 entries, 0 to 537576
Data columns (total 13 columns):
User_ID                     537577 non-null int64
Product_ID                  537577 non-null object
Gender                      537577 non-null object
Age                         537577 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
```

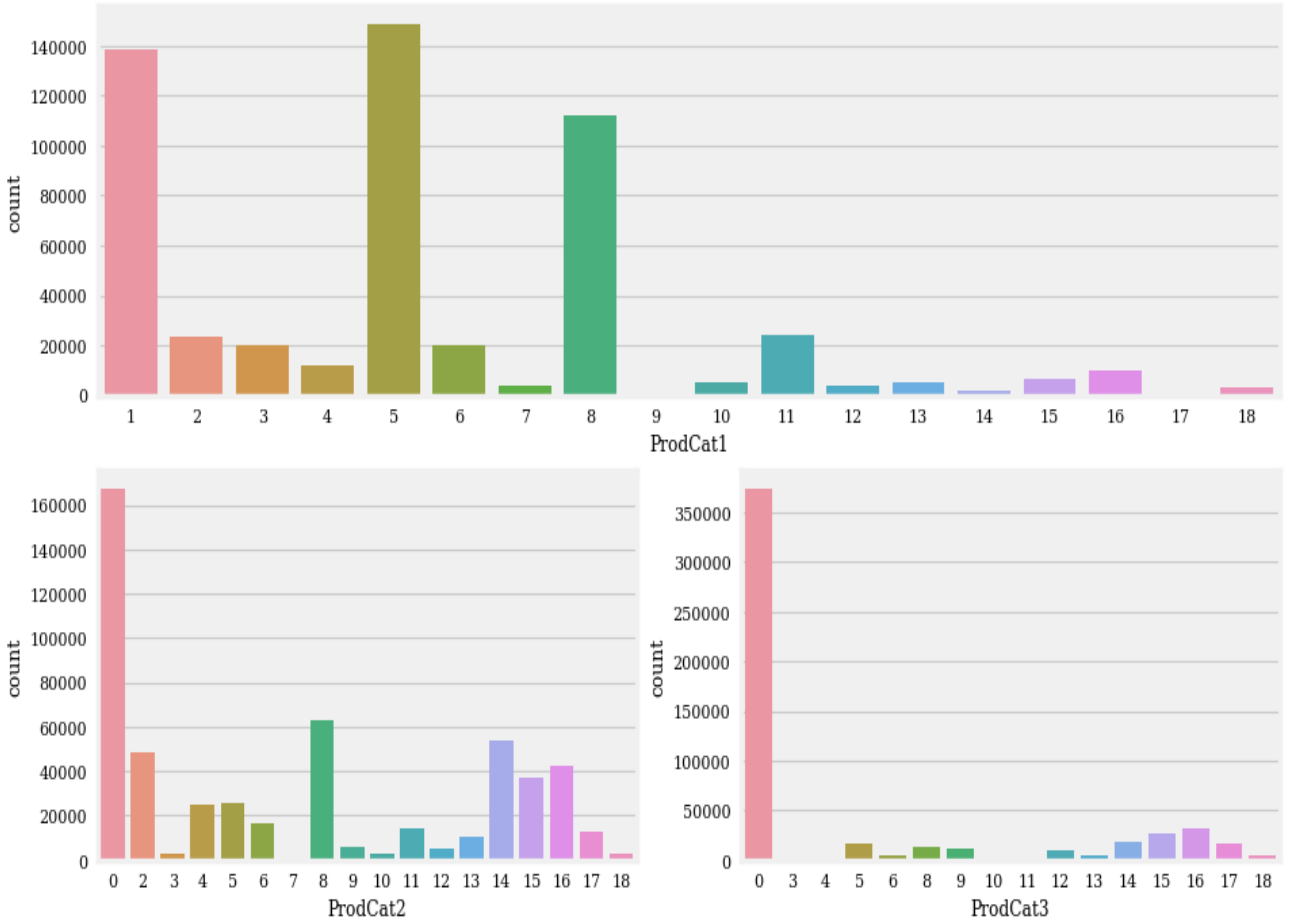
Step 2: Exploratory Data Analysis

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.



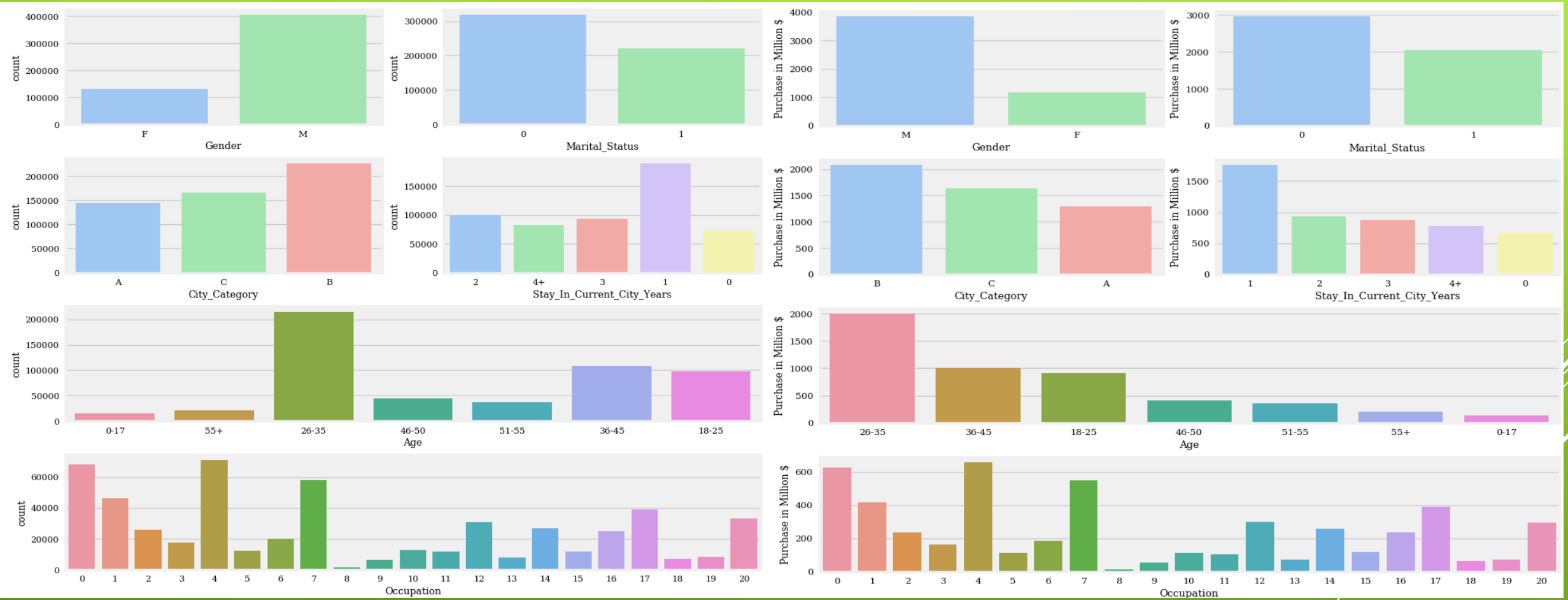
Count Plot of all variables





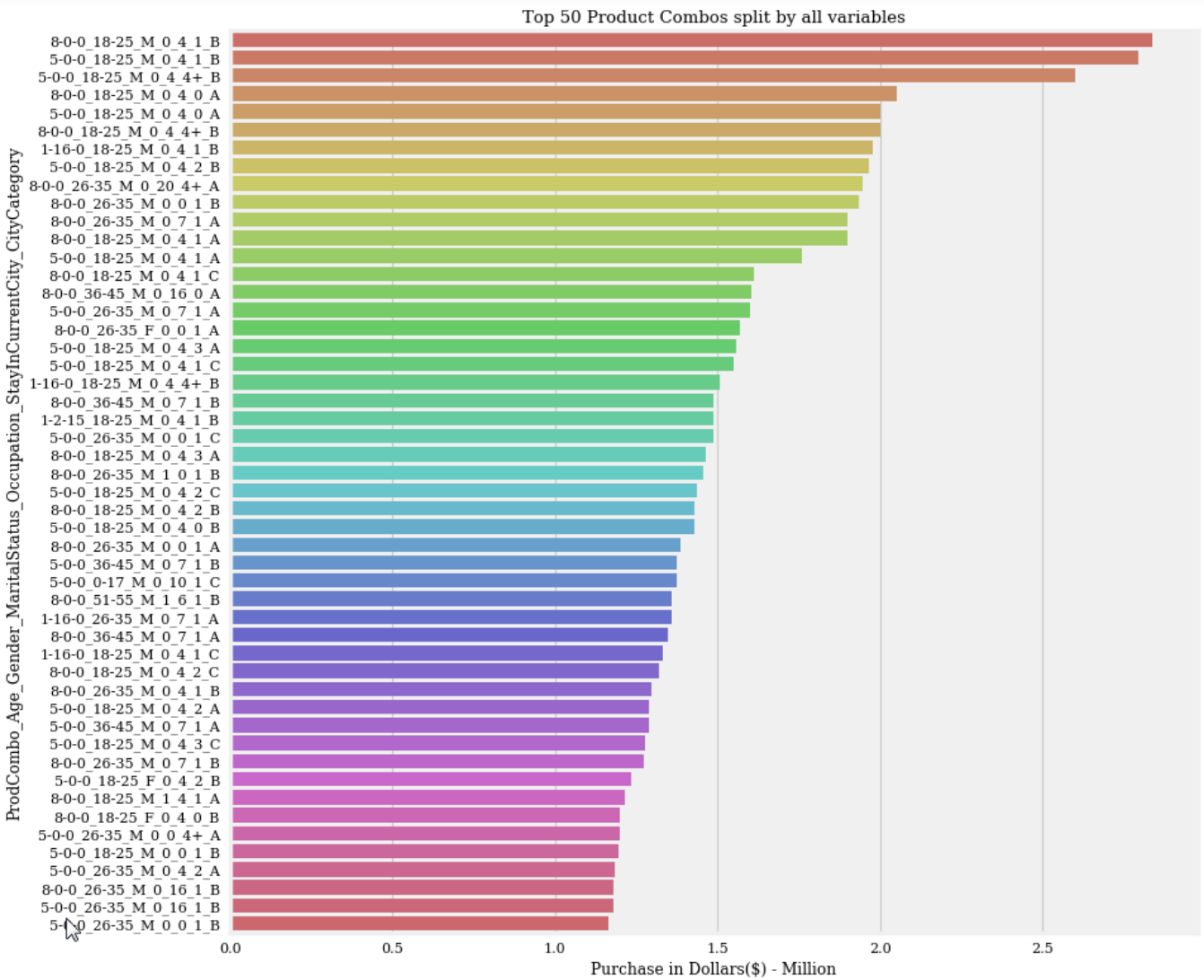
1. Men shopped 3 times as much as women. This could be due to the fact that men paid for the purchases too.
2. Singles shopped more times than married ones. May be they have more disposable income.
3. Maximum number of purchases were made by people from City B, and people from cities A and C shop approximately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

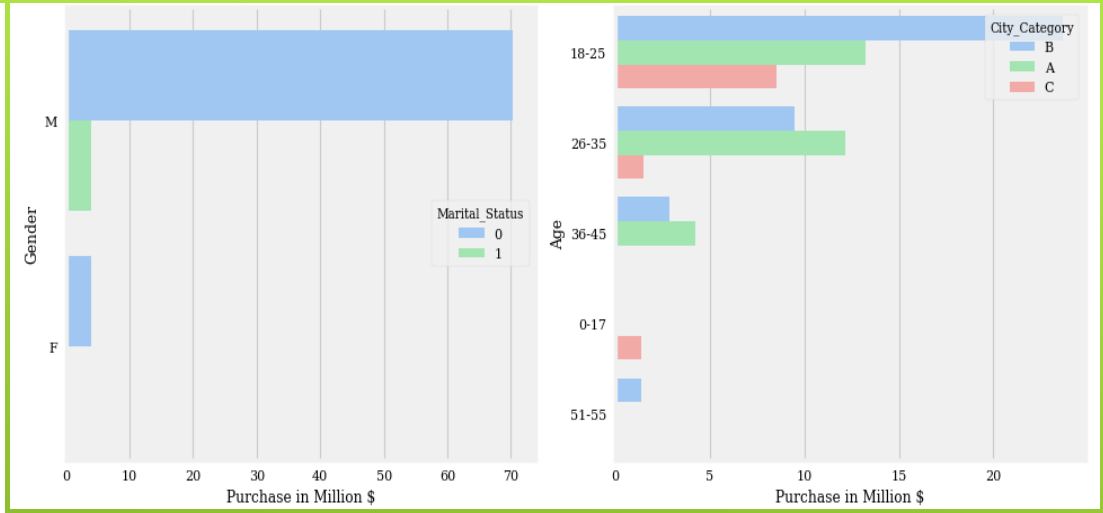
Count Plot vs Purchase value

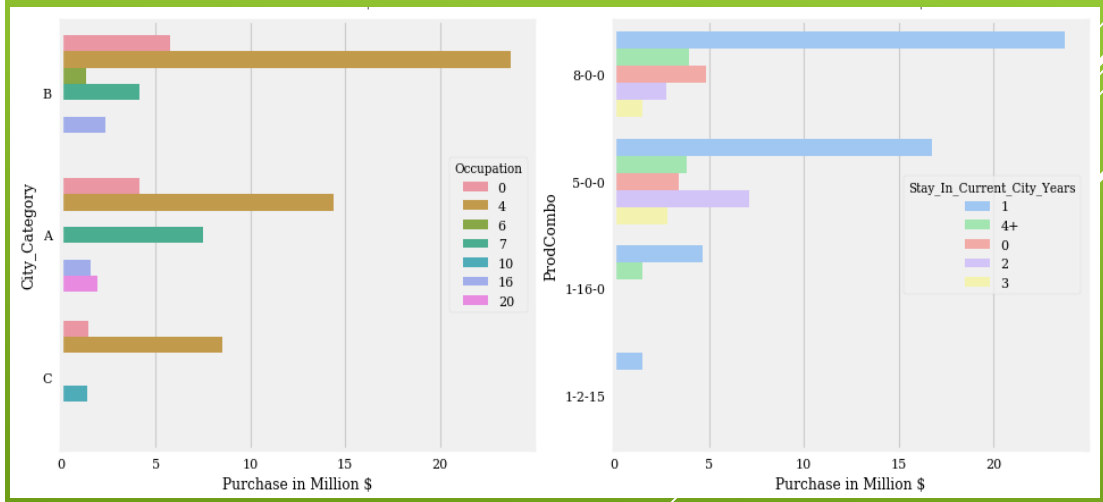


Distribution in amounts is similar to the distribution in counts.
No variable disproportionately affecting the purchase.
The two plots seems somewhat similar and we can conclude that no further analysis needed

Master Plot Insights – Who is my target audience?



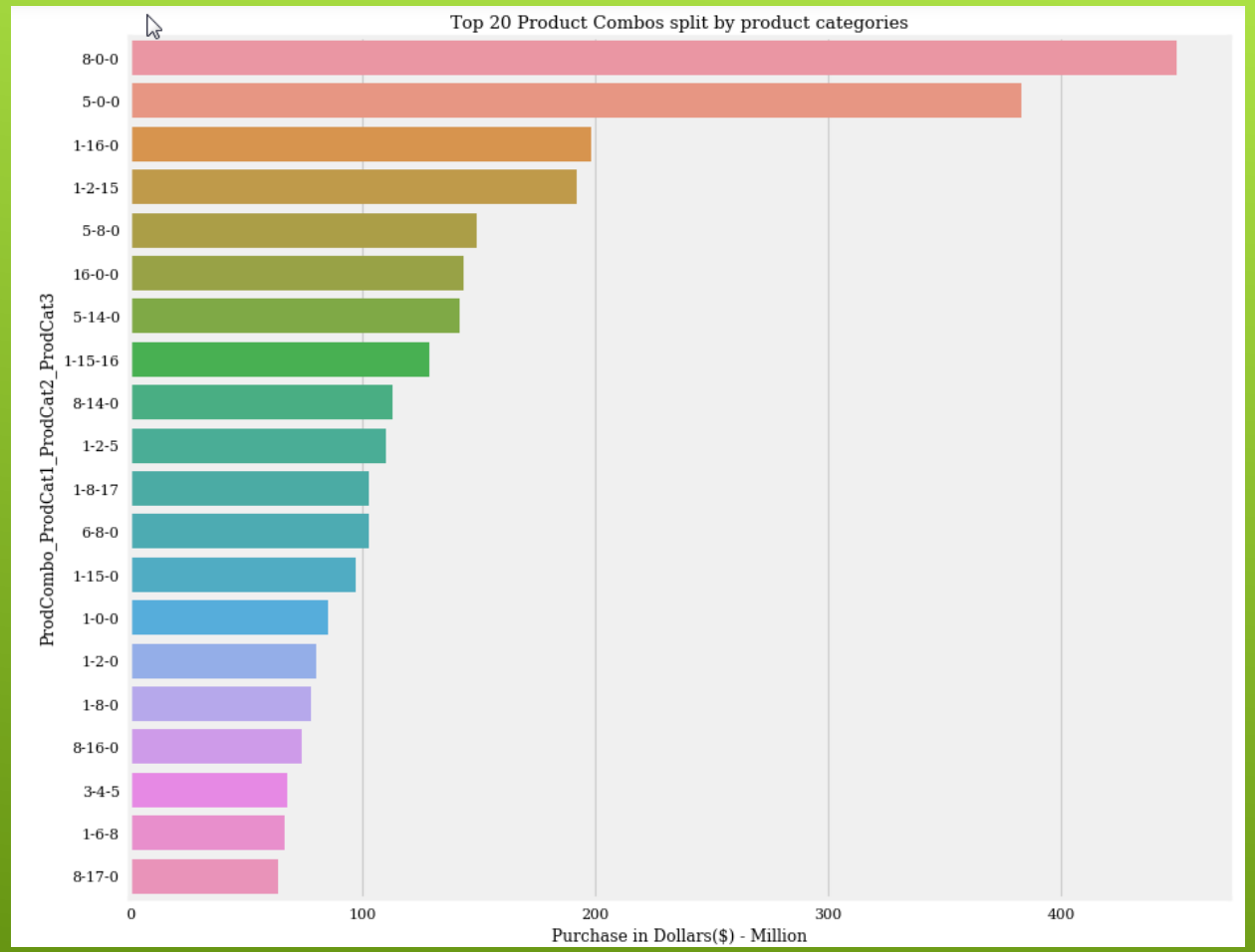


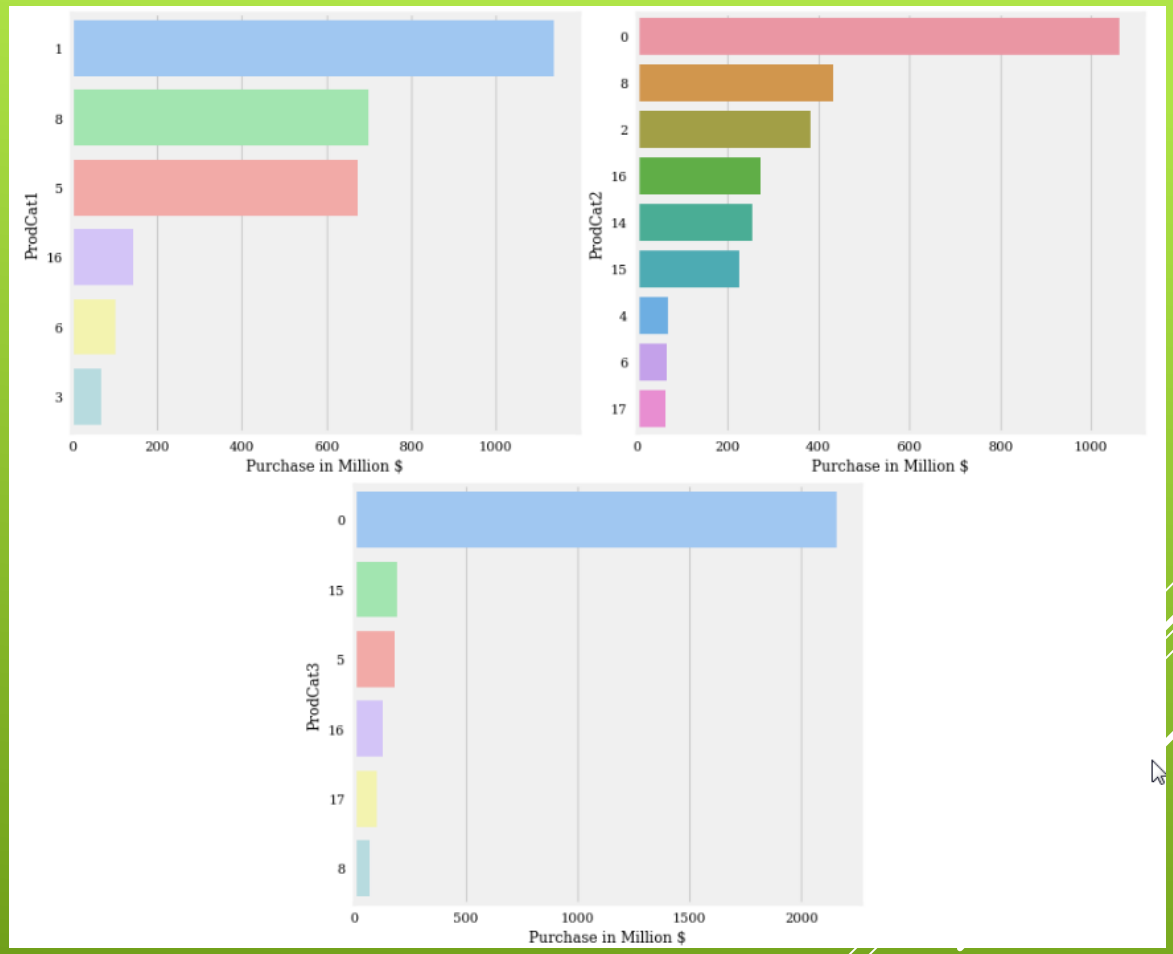


- Single males buy the most from city B, in age group of 18-35. They are followed by city A in age groups of 18-25 and 25-36.
- A disproportionate number of these people are in Occupation 4, mainly buying from city B and A.
- People who mainly buy ProdCombos 8-0-0 and 5-0-0 have been living in the city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

Verdict: The target 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.

Who buys top products?



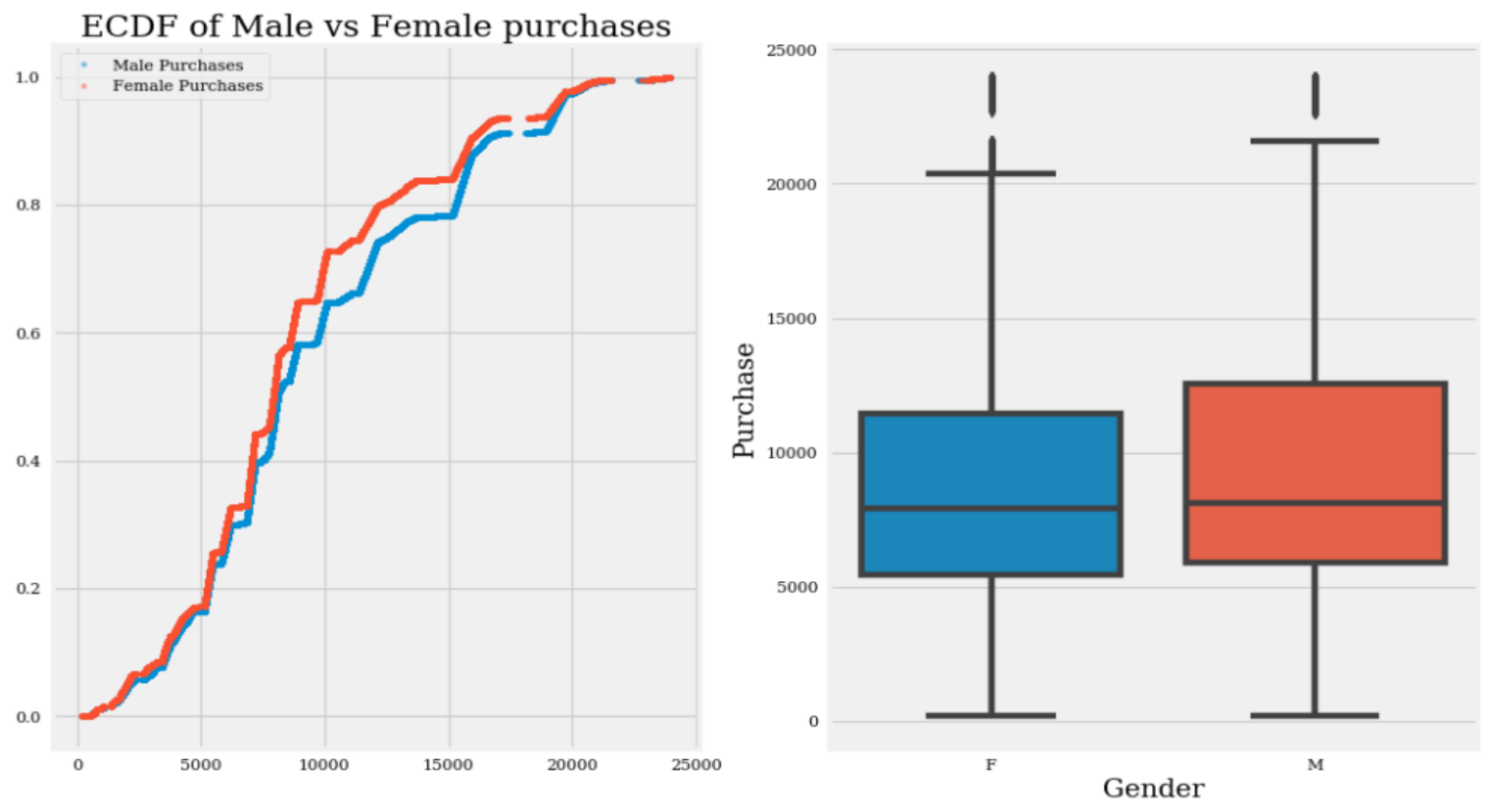


- In the top 20 combos, you can now confirm that product 1 in Category 1 still has the most number of combos and generates the highest revenue, followed by 8 and 5.
- In terms of combos across all three categories, 1-2-15, 1-15-16 and 1-2-5 are frequently bought together.

Verdict: Top Products are:

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

Inferential Statistics 1. Do men buy more expensive things than women?



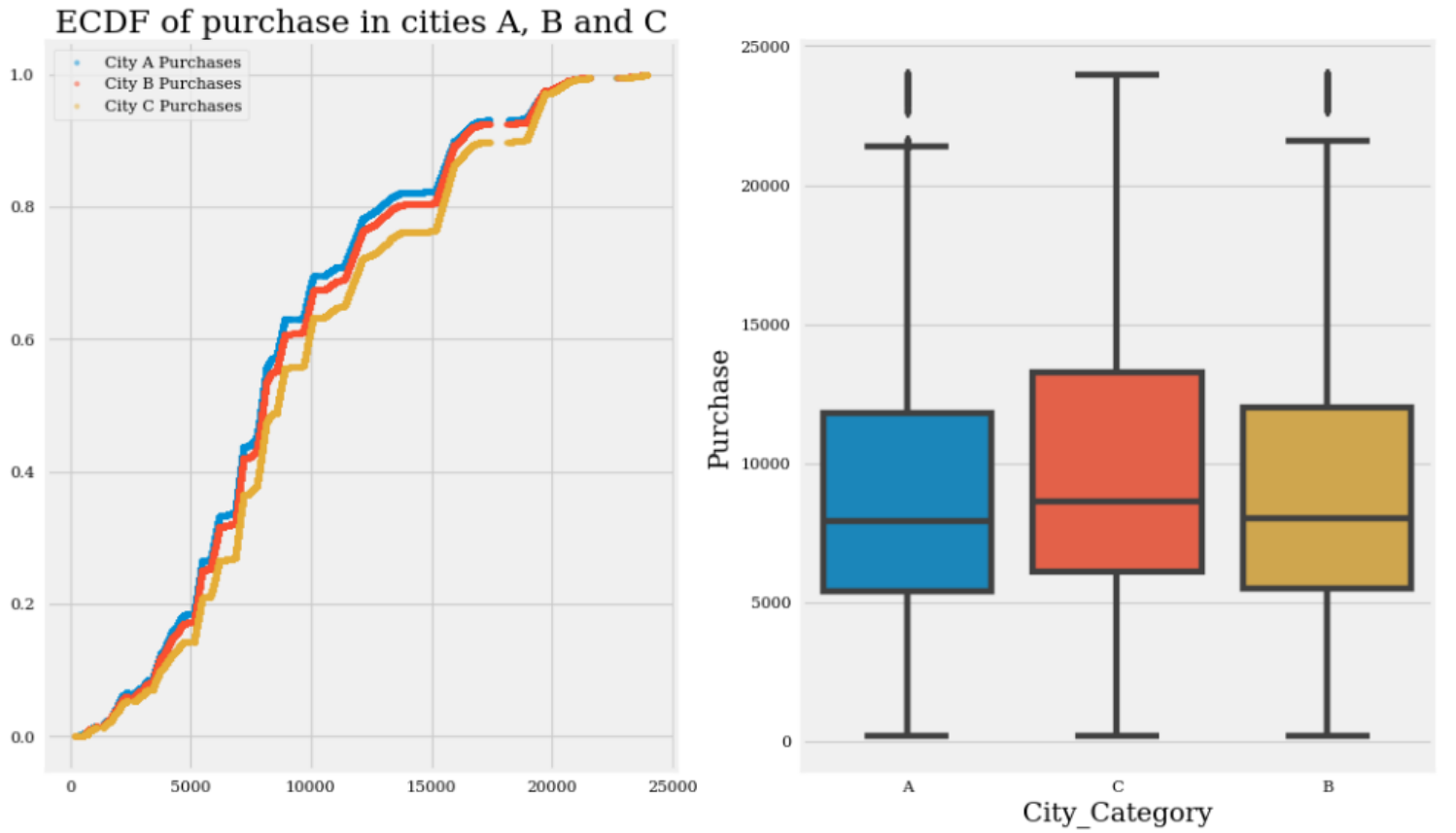
It seems like men tend to make more expensive purchases by looking at ECDF.
Expensive purchase = $10,000.

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.

Inferential Statistics 2. Do people shop equally in all three cities?



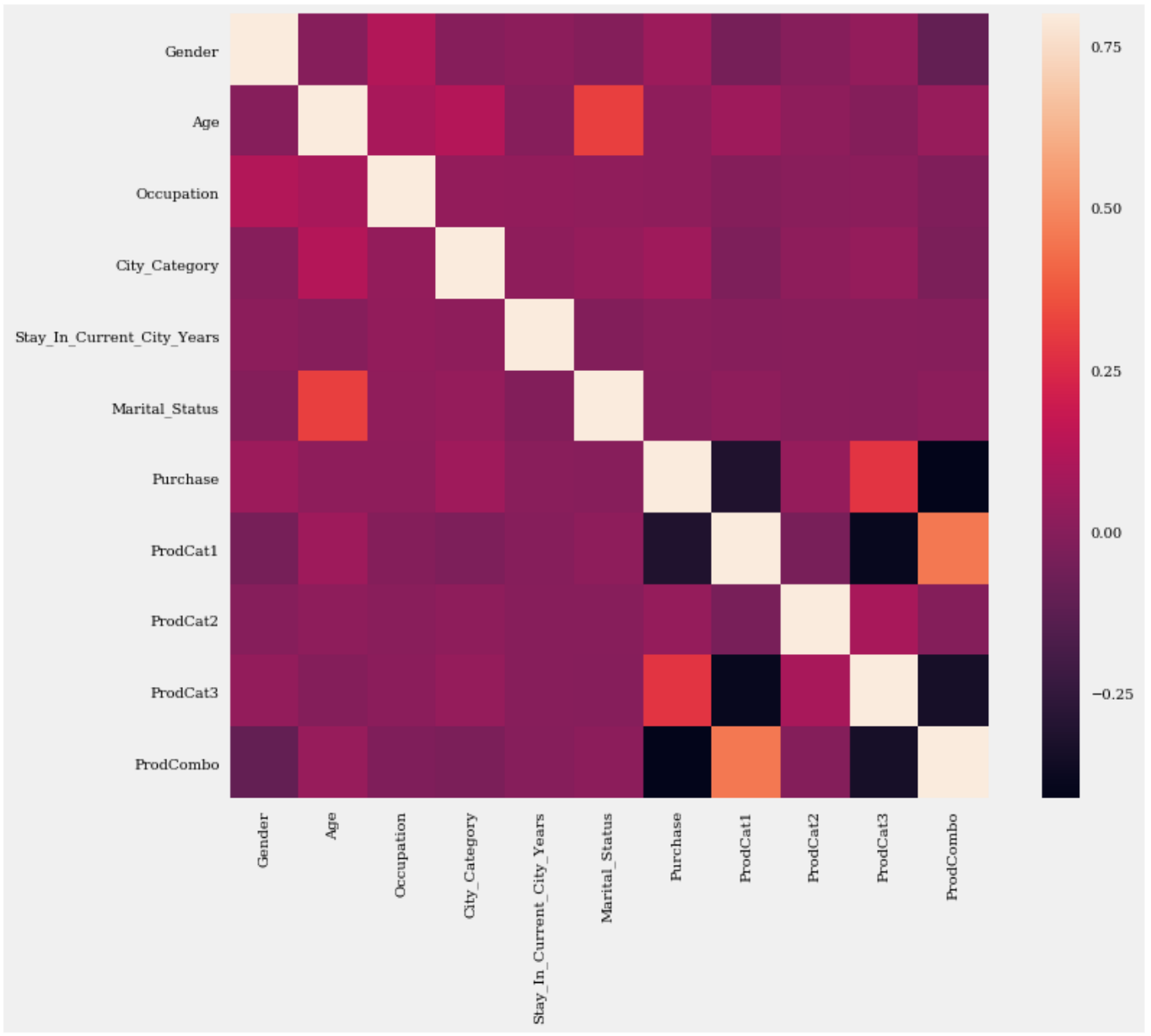
H0: The average purchase price by people in cities A, B and C is the same.

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



- Marital status seems to have strong positive correlation with Age.
- 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 seems to be any strong correlation between the variables.

In-Depth Analysis

Linear Regression:
R^2 with Prod Categories: 0.6376223400425602
Root Mean Squared Error with Prod Categories: 2989.4402029238413
Cross validation score is with Prod Categories: 0.6389821282926864

Decision Tree: A configuration of (max_depth=15, min_samples_leaf=100) gave the below results.
R^2: 0.635121867047957
Root Mean Squared Error: 2999.736317822526

Elastic CV: A configuration of (cv=5, alphas=np.linspace(0.001,1,50)) gave the below results
R^2: 0.6400716117954796
Root Mean Squared Error: 2993.6060388057804

Random Forest: A configuration of (max_depth=16, n_estimators=90) gave the below result
R^2: 0.6424529101374172
Root Mean Squared Error: 2983.686705716457

- In general we are not getting past the 0.65 accuracy mark.
- Need for feature engineering to improve the scores.

Feature Engineering

Four functions were created that return

1. Average Purchase Value per column.
2. Average Count of column
3. Total count of column
4. Median purchase value of column

Linear Regression:
R^2: 0.7176789707438745
Root Mean Squared Error: 2638.6438593275593 (vs. 2989.44)
Cross validation score with new features is: 0.7191447714961551 (vs. 0.6389)

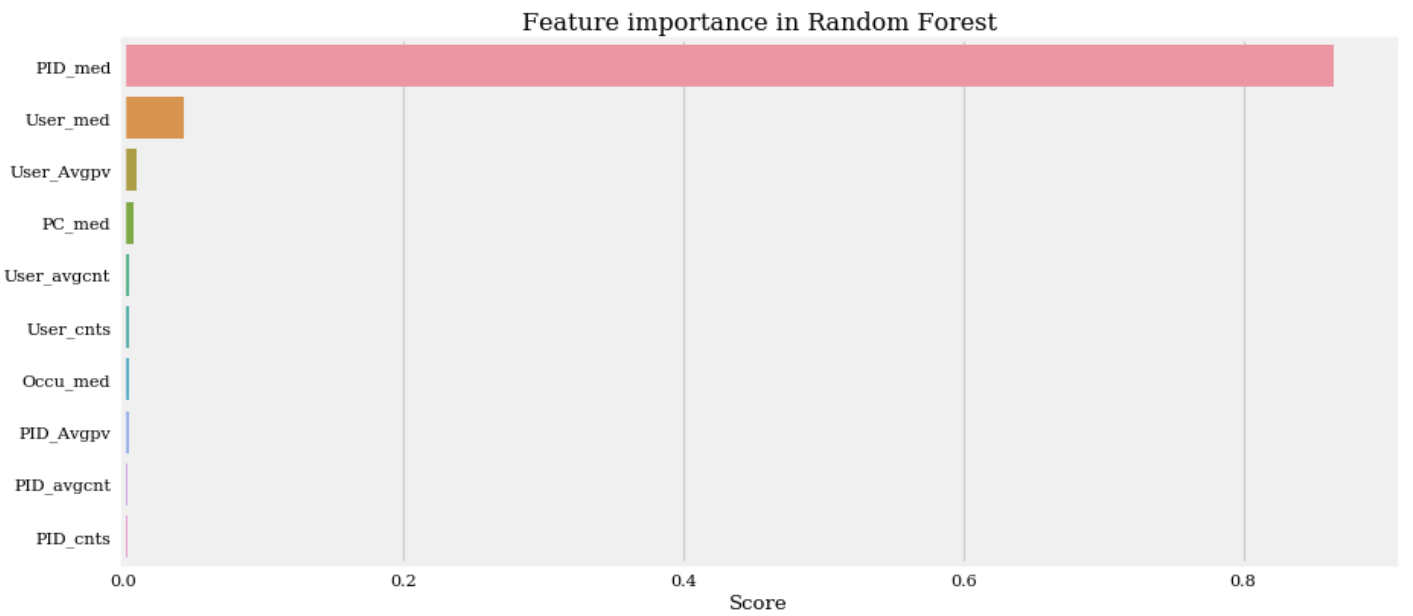
Decision Tree: Configuration of **(**max_depth=400, min_samples_leaf=112, min_samples_split=40) gave
R^2 Test: 0.7373382707106608 (vs. 0.6351)
Root Mean Squared Error Test: 2552.277662635797 (vs. 2999.73)

Elastic CV: A configuration of (cv=5, alphas=np.linspace(0.001,1,50)) gave the below results
R^2: 0.7197624634394806 (vs. 0.64007)
Root Mean Squared Error: 2641.494731114654 (vs. 2993.60)

Random Forest: A configuration of (max_depth=16, n_estimators=90) gave the below result
R^2: 0.7467830310391308 (vs. 0.64)
Root Mean Squared Error: 2510.7525519630262 (vs. 2983.68)

Conclusion

Feature engineering helped boost the overall accuracy score past 70% to 74.6%.



Further investigation:

1. Instead of setting the missing values to 0, we can try putting in mean/median values in the product category and checking if the prediction would be any different.
2. Try more parameters to see if Random Forest can be more effective in increasing accuracy.
3. Try Gradient Boosting and XGB to see if it is any better than Random Forest.
4. Try Neural Network Regression – feature engineering is not usually needed with neural networks as they are really good at detecting hidden features. Try to see if gives better accuracy.