PROJECT REPORT ON:

"BLACK FRIDAY PREDICTION"

SUBMITTED BY

RAHUL M

ACKNOWLEDGMENT

I would like to express my special gratitude to "Flip Robo" team, who has given me this opportunity to deal with a beautiful dataset and it has helped me to improve my analyzation skills. And I want to express my huge gratitude to Shwetank Mishra (SME Flip Robo) who encouraged me a lot with his valuable words and with his unconditional support I have ended up with a beautiful Project.

A huge thanks to my academic team "Data trained" who are the reason behind what I am today. Last but not least my parents who have been my backbone in every step of my life. And also thank you for many other persons who has helped me directly or indirectly to complete the project.

1.INTRODUCTION

1.1 Business Problem Framing:

A retail company “ABC Private Limited” wants to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high volume products from last month. The data set also contains customer demographics (age, gender, marital status, city_type, stay_in_current_city), product details (product_id and product category) and Total purchase_amount from last month.

Now, they want to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

1.2 Motivation for the Problem Undertaken

There are purchase data on multiple products and data of multiple category of cities. And data on people purchasing data as well, So that we can do the EDA on the data.

2.Analytical Problem Framing

2.1 Mathematical/ Analytical Modeling of the Problem

As a first step I have loaded the datasets in the jupyter Notebook.

In this particular problem we have Purchase as the target column and it was a continuous column. So clearly it is a regression problem .Anc checked for

duplicate values in the dataset. There was null values in the dataset. To get better insight on the features I have used ploting like distribution plot, bar plot, reg plot, line plot and count plot. With these ploting I was able to understand the relation between the features in better manner. Also, I found outliers and skewness in the dataset so I removed outliers using z-score method and I removed skewness using yeo-johnson method and checked the data for collinearity in the dataset.

2.2 Data Sources and their formats

The data was available in csv format and loaded in the data.Also my dataset was having 550068 rows and 12 columns in train dataset including target and 233599 rows and 11 columns in test data. In this particular datasets I have object type of data which has been changed as per our analysis about the dataset. The information about features is as follows.

Features Information:

- Data
- Variable Definition
- User_ID User ID
- Product_ID Product ID
- Gender Sex of User
- Age Age in bins
- Occupation Occupation (Masked)
- City_Category Category of the City (A,B,C)
- Stay_In_Current_City_Years Number of years stay in current city
- Marital_Status Marital Status
- Product_Category_1 Product Category (Masked)
- Product_Category_2 Product may belongs to other category also(Masked)
- Product_Category_3 Product may belongs to other category also
- Purchase Purchase Amount (Target Variable)

2.3 Data Preprocessing Done

- ✓ I have imported required libraries and I have imported the dataset which was in csv format.
- ✓ Then I did all the statistical analysis like checking shape, nunique, value counts, info etc…..
- ✓ While checking for null values I found null values in the dataset and I replaced them using imputation technique.
- ✓ I have also dropped the column with more than 50% null values in the dataset:0 column as I found they are useless.
- ✓ Next as a part of feature extraction I converted the data types of all the columns and I have extracted usefull information from the raw dataset. Thinking that this data will help us more than raw data.

2.4 Data Inputs- Logic- Output Relationships

- ✓ Since I had numerical columns I have plotted dist plot to see the distribution of skewness in each column data.
- ✓ I have used bar plot for each pair of categorical features that shows the relation between label and independent features.
- ✓ I have used reg plot and line plot to see the relation between numerical columns.
- ✓ I can notice there is a linear relationship between maximum columns and target.

2.5 Hardware and Software Requirements and Tools Used

While taking up the project we should be familiar with the Hardware and software required for the successful completion of the project. Here we need the following hardware and software.

Hardware required: -

1. Processor — core i5 and above
2. RAM — 8 GB or above

3. SSD — 250GB or above

Software/s required: -

1.Anaconda

Libraries required :-

```
:  1 import pandas as pd
   2 import numpy as np
   3 import seaborn as sns
   4 import matplotlib.pyplot as plt
   5 %matplotlib inline
   6 import seaborn as sns
   7
```

To run the program and to build the model we need some basic libraries as follows:

- ✓ **import pandas as pd**: **pandas** is a popular Python-based data analysis toolkit which can be imported using import pandas as pd. It presents a diverse range of utilities, ranging from parsing multiple file formats to converting an entire data table into a numpy matrix array. This makes pandas a trusted ally in data science and machine learning.
- ✓ **import numpy as np**: NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.
- ✓ **import seaborn as sns:** Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.
- ✓ **Import matplotlib.pyplot as plt:** matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.
- ✓ from sklearn.preprocessing import LabelEncoder

- ✓ from sklearn.preprocessing import StandardScaler
- ✓ from statsmodels.stats.outliers_influence import variance_inflation_factor

With this sufficient libraries we can go ahead with our model building.

3.Data Analysis and Visualization

3.1 Identification of possible problem-solving approaches (methods)

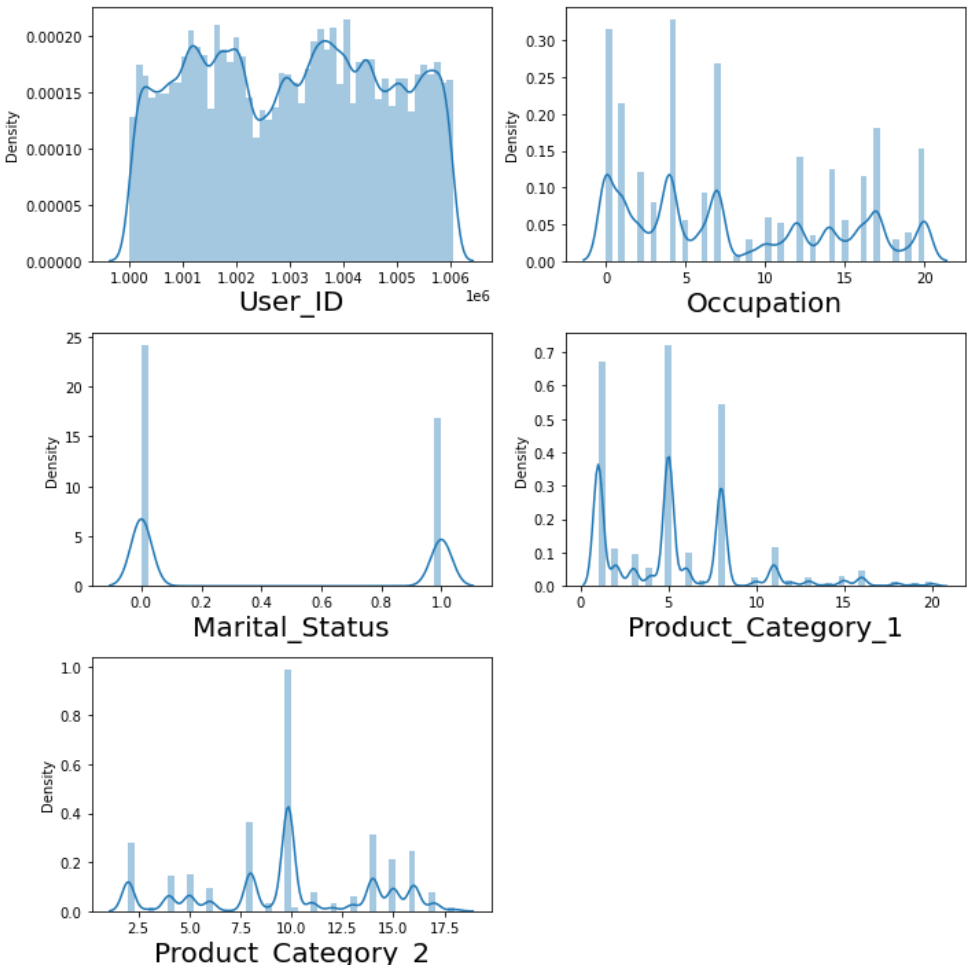
- ✓ cleaned the data and bring it to the proper format for our analysis. To remove outliers I have used z-score method. And to remove skewness I have used yeo-johnson method. We have dropped all the unnecessary columns in the dataset according to our understanding. Use of Pearson's correlation coefficient to check the correlation between dependent and independent features. Also I have used Standardisation to scale the data. After scaling we have to remove multicolinearity using VIF.

3.4 Visualizations

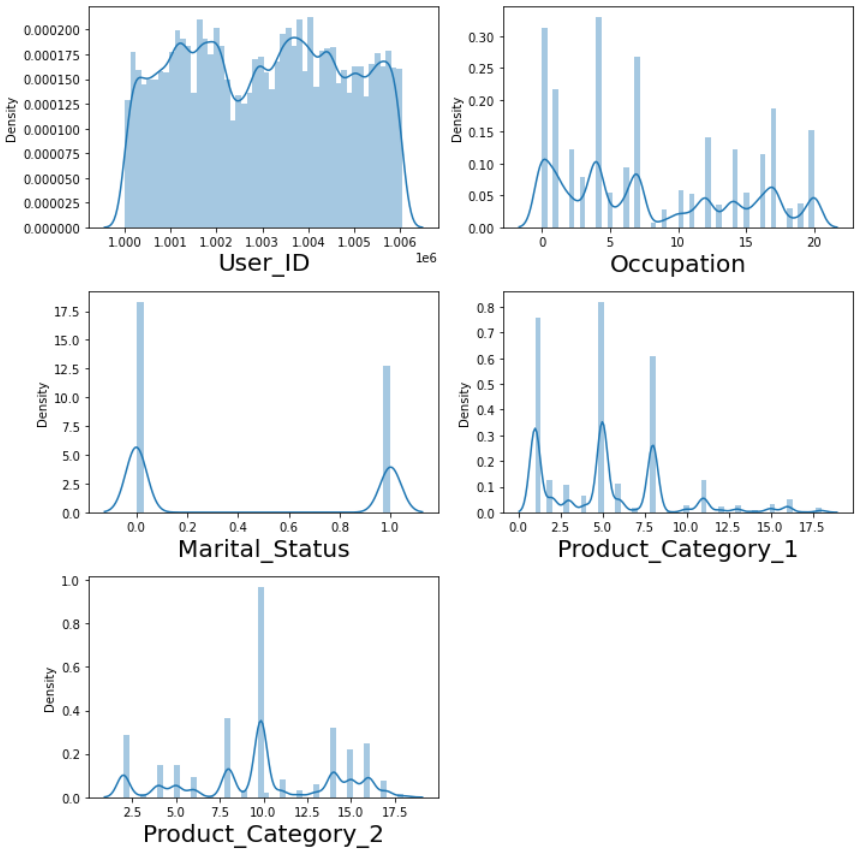
I have used bar plots to see the relation of categorical feature with target and I have used 2 types of plots for numerical columns one is disp plot for univariate and reg plot, line plots for bivariate analysis.

1. Univariate Analysis for numerical columns:

Testdata



Traindata

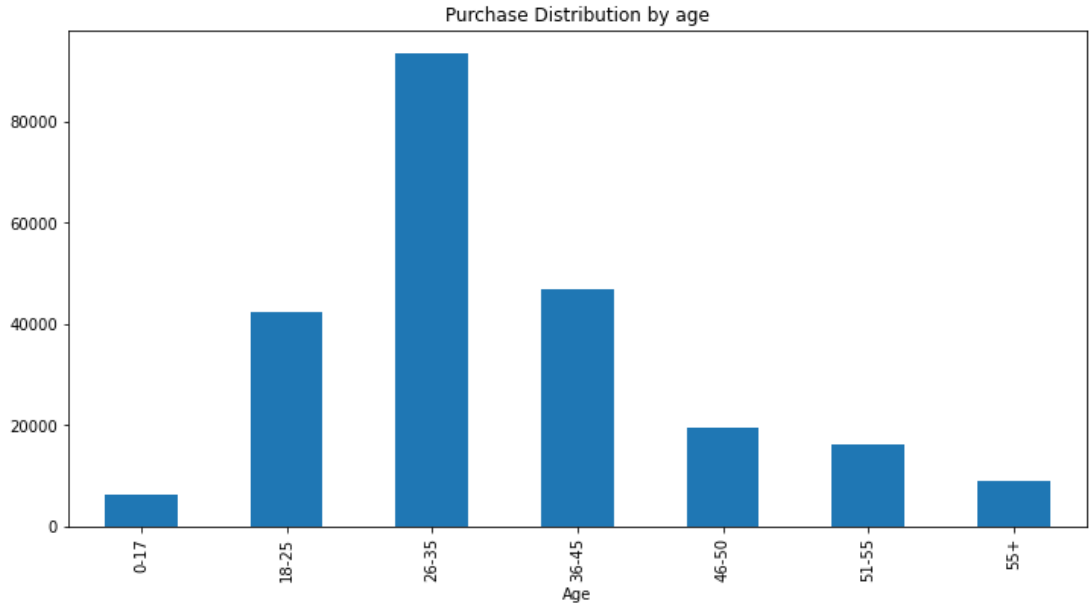


Observations:

- ✓ We can clearly see that there is skewness in most of the columns so we have to treat them using suitable methods.

2. Univariate analysis for categorical column:

Train data



#We can see that in the 0-17 age group there are 6232 count

#We can see that in the 18-25 age group there are 42293 count

#We can see that in the 26-35 age group there are 93428 count

#We can see that in the 36-45 age group there are 46711 count

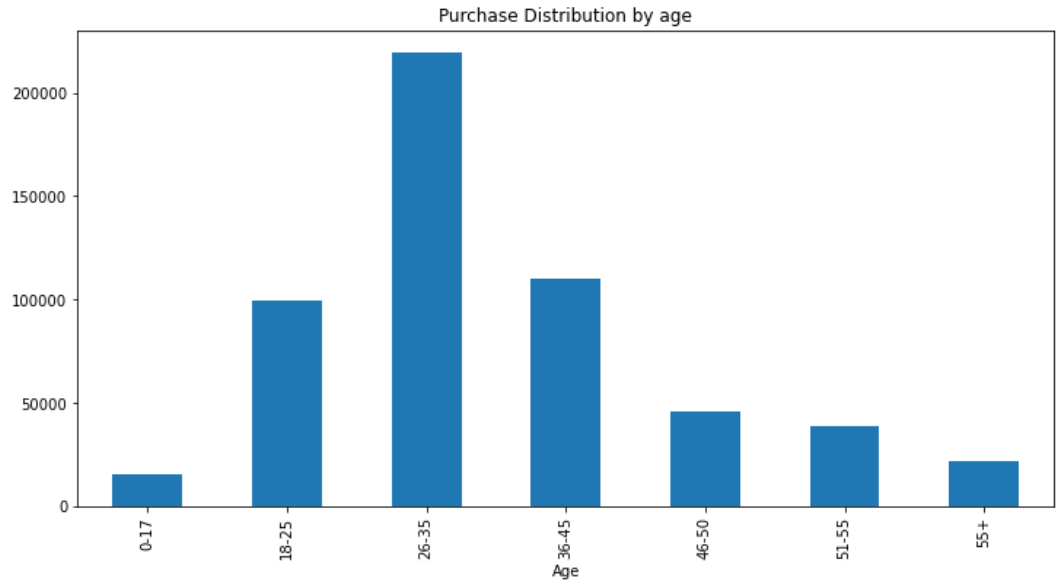
#We can see that in the 46-50 age group there are 19577 count

#We can see that in the 51-55 age group there are 16283 count

#We can see that in the 55+ age group there are 9075 count

#So max data is in 18-25 age group and least in 0-17 age group

Testdata



#We can see that in the 0-17 age group there are 15102 count

#We can see that in the 18-25 age group there are 99660 count

#We can see that in the 26-35 age group there are 219587 count

#We can see that in the 36-45 age group there are 110013 count

#We can see that in the 46-50 age group there are 45701 count

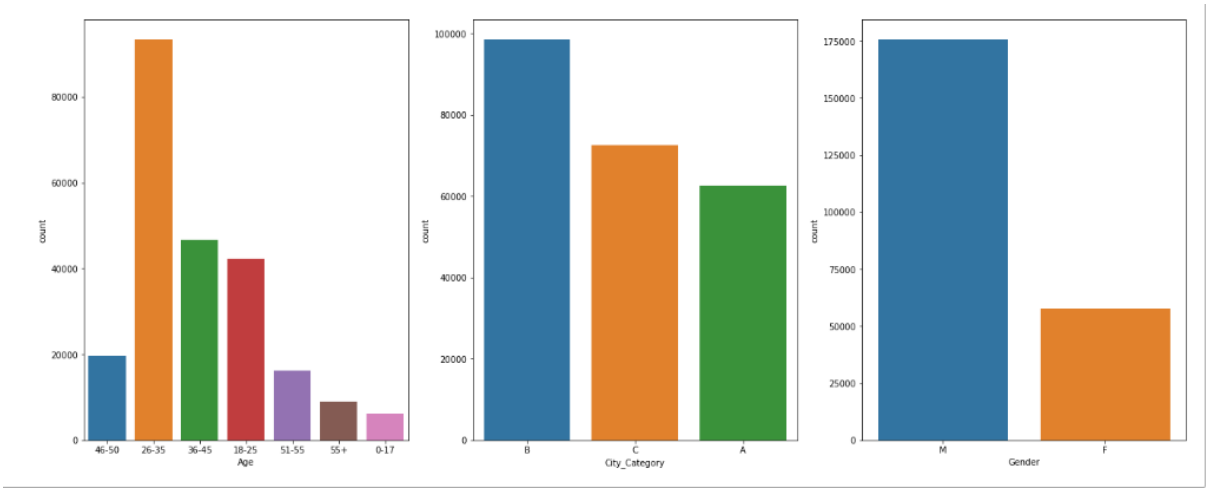
#We can see that in the 51-55 age group there are 38501 count

#We can see that in the 55+ age group there are 21504 count

#So max data is in 18-25 age group and least in 0-17 age group in train dataset

Train data

- ✓ the Highest data is in 26-35 in the testdataset
- ✓ the least data is in 26-35 in the testdataset
- ✓ Among the city category B has the highest count and
- ✓ Least count in city category A
- ✓ The male Count is the highest in dataset



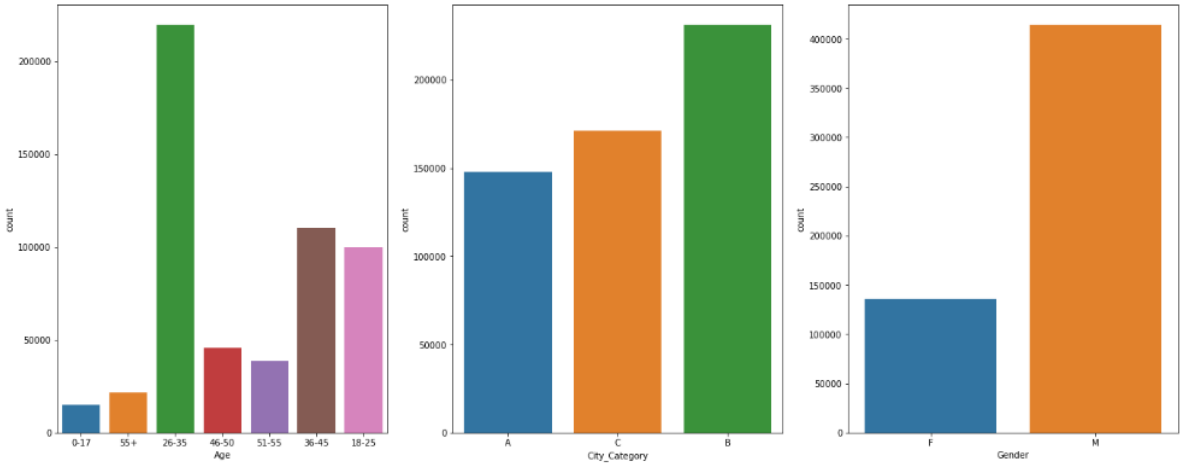
Testdata

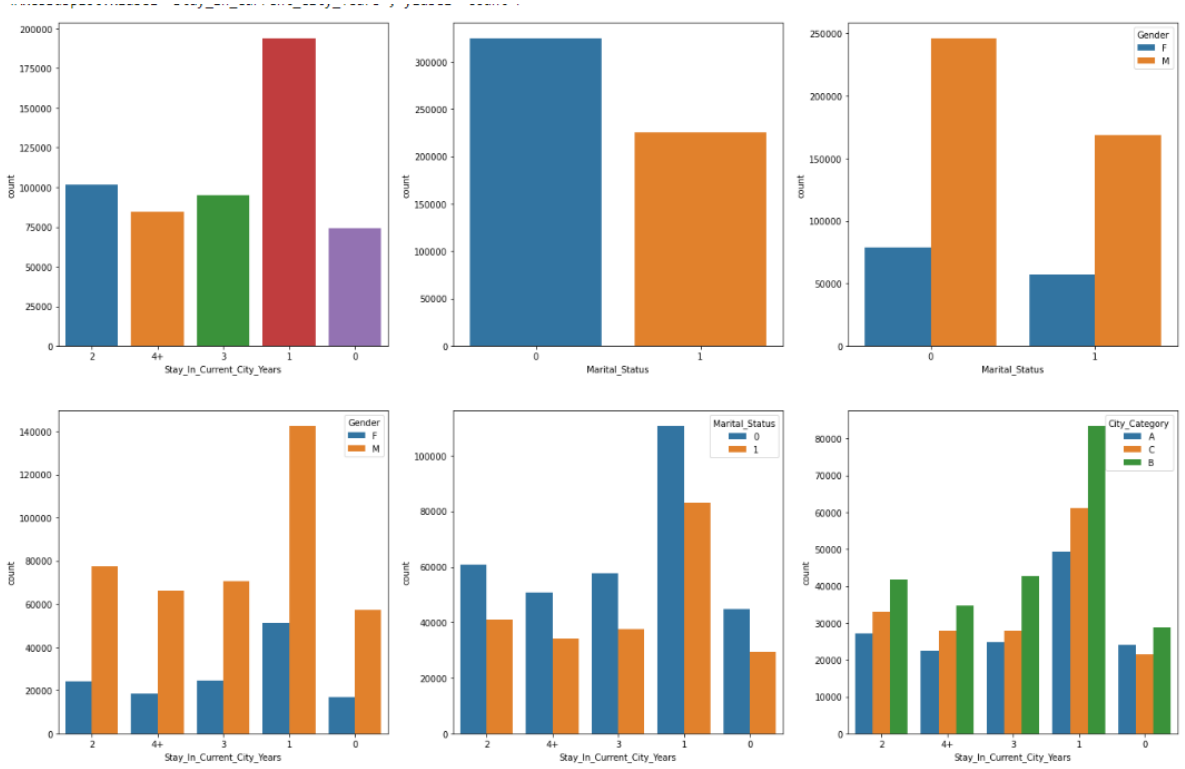
the Highest data is in 26-35 in the traindataset

the least data is in 26-35 in the traindataset

-Among the city category B has the highest count and
-Least count in city category A

The male Count is the highest in dataset





Observations:

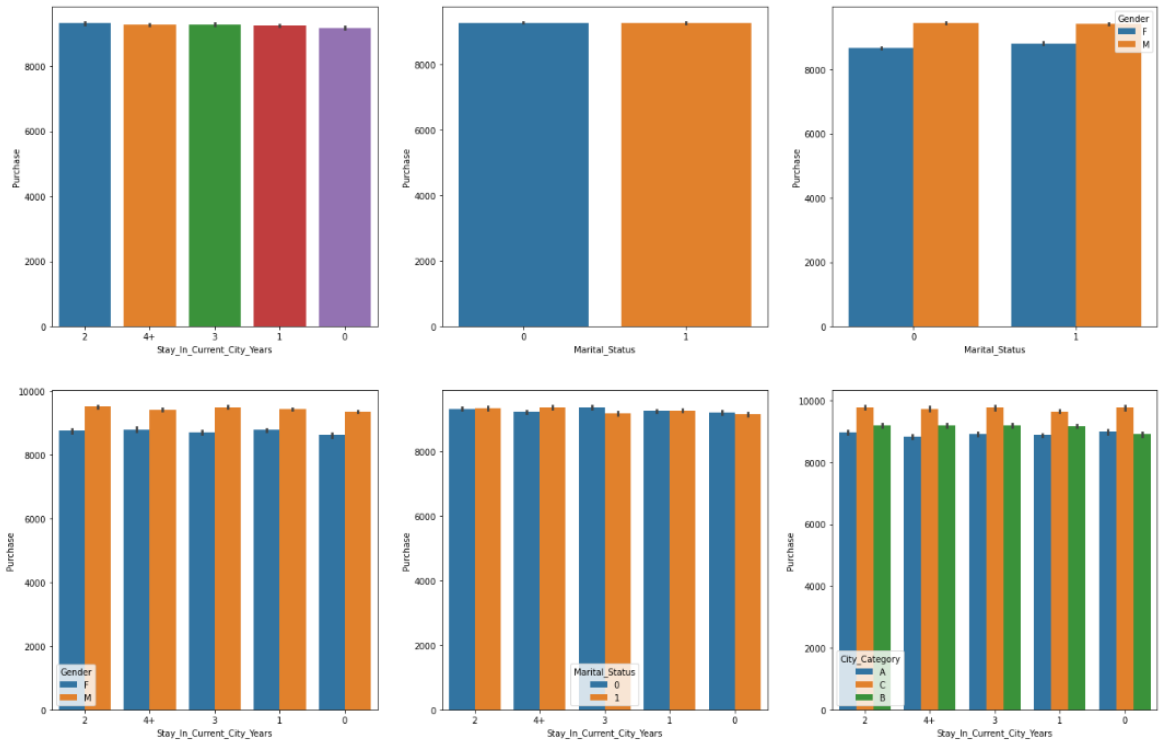
- We can see Stay_In_Current_City_Years has the highest count in 1 year the train dataset

- and Stay_In_Current_City_Years is least count in less than 1 year the train dataset

- We can see the data with married people less in the dataset

- Among the datastet males are dominating in the married and nonmarried people in the dataset

- We can see least dataAmnong the females with in married people



- We can see Stay_In_Current_City_Years has almost equal no of purchase count in the train dataset

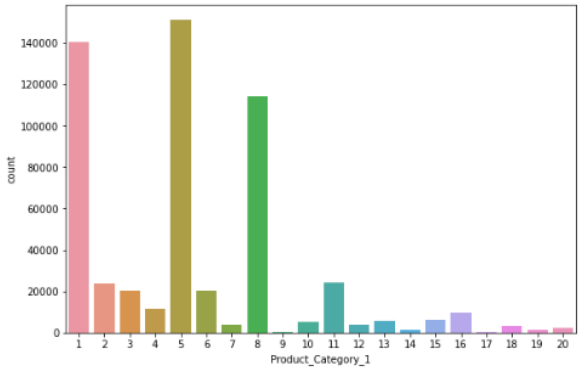
- and Stay_In_Current_City_Years has the highest count for males in the train dataset

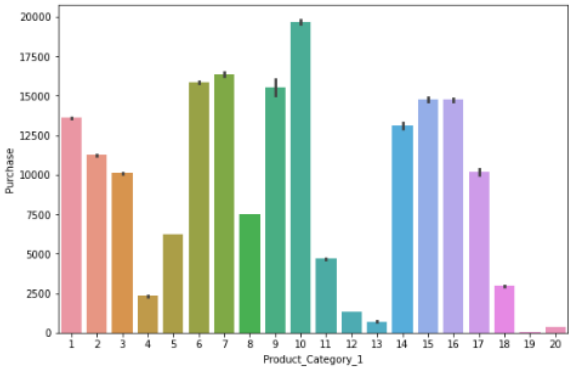
- Comparing the data difference of percentage of males and female is small among teh dataset

- We can see the data with married people having almost equal no of purchase count in the train dataset

- Among the datastet males are dominating in the married and nonmarried people in the dataset in Purchase counts

- We can see least dataAmnong the females with in married people with Purchase counts





Observations:

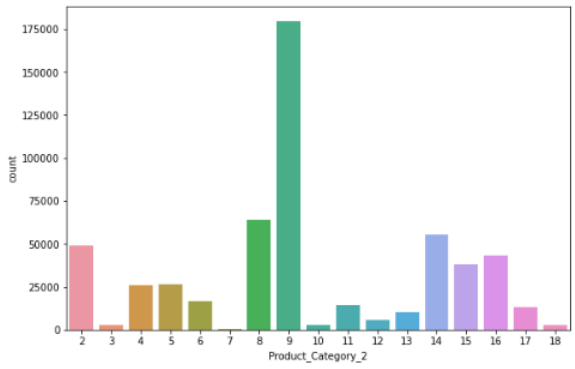
- In Product_Category_1 Category 5 has the highest count and least in Category 9 and 17.

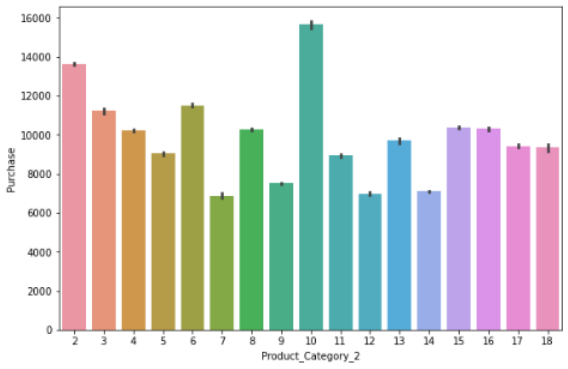
- The second highest is in Category 1 in the train dataset

- the categories 1,5,8 has predominence in the train dataset

- The purchase count is highest for the category for 10 and

- least in categories 4 and 12,13,19,20 in the dataset





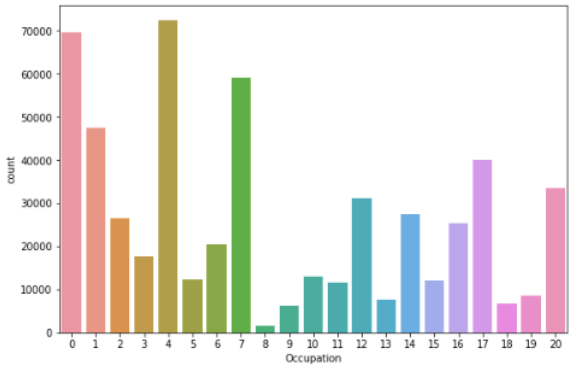
- In Product_Category_2 Category 9 has the highest count and least in Category 7.

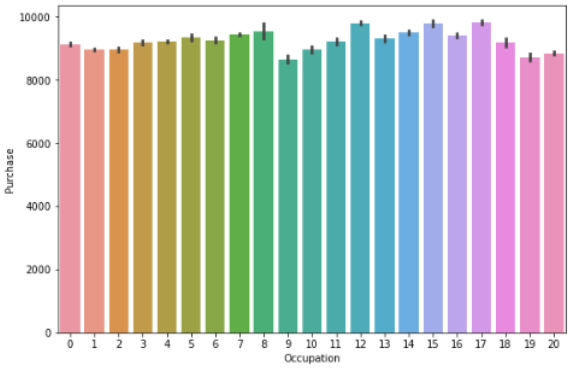
- The second highest is in Category 8 and 14 in the train dataset

- the categories 9,8,2,14,15,16 has predominence in the train dataset

- The purchase count is highest for the category for 10 and

- least in categories 7,12,14 in the dataset





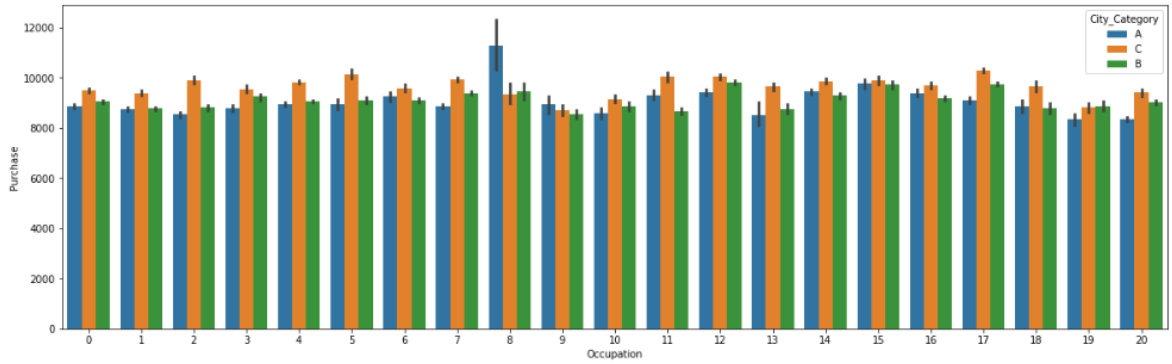
- In Occupation Category 4,0,7 have the highest counts and least in Category 8.

- The second highest is in Category 8 and 14 in the train dataset

- the categories 9,8,2,14,15,16 has predominence in the train dataset

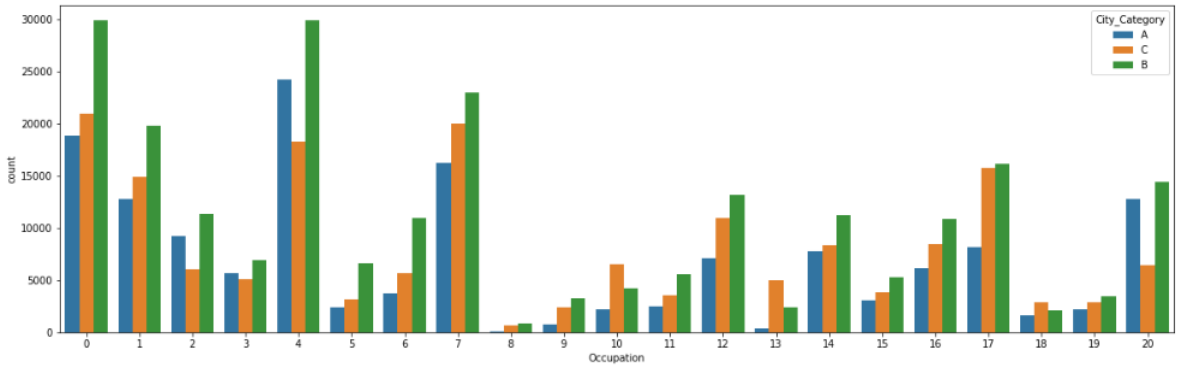
- The purchase count have the highest counts for 12,15,17 Category and

- least in categories 9,19 in the dataset



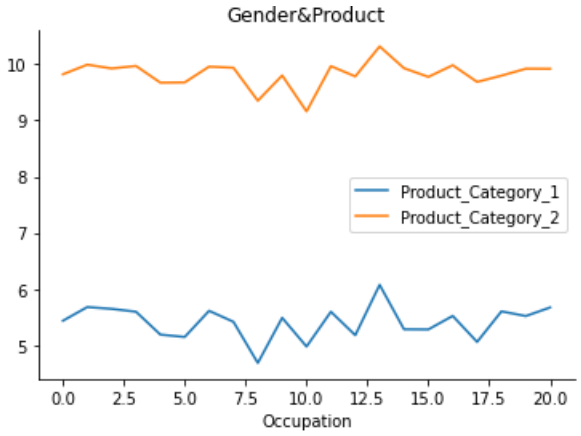
- Among the Occupation categories we can see that except in category 8 all the occupation categories are having highest in City category of C in terms of purchase counts

- In almost all the occupation categories the City category A is least purchase counts



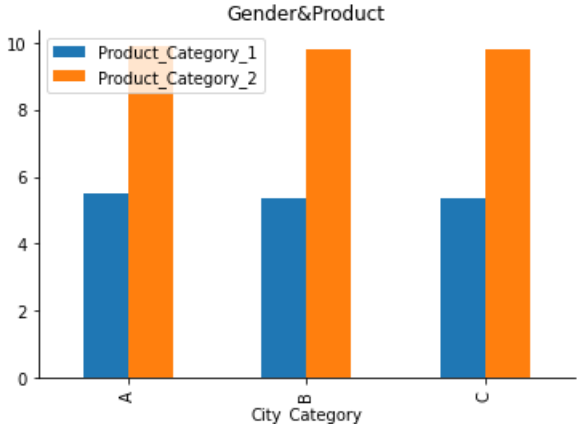
- Among the Occupation categories 0 ,4 7 has the highest percentage of data in the three city categories

- In almost all the occupation categories 8 8,18,19 are tyhe least counts with least data in the City category A



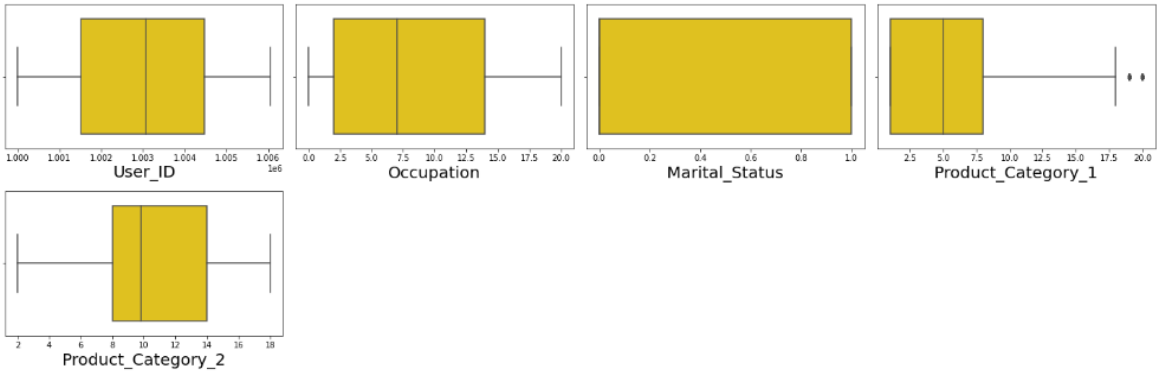

```
1
```

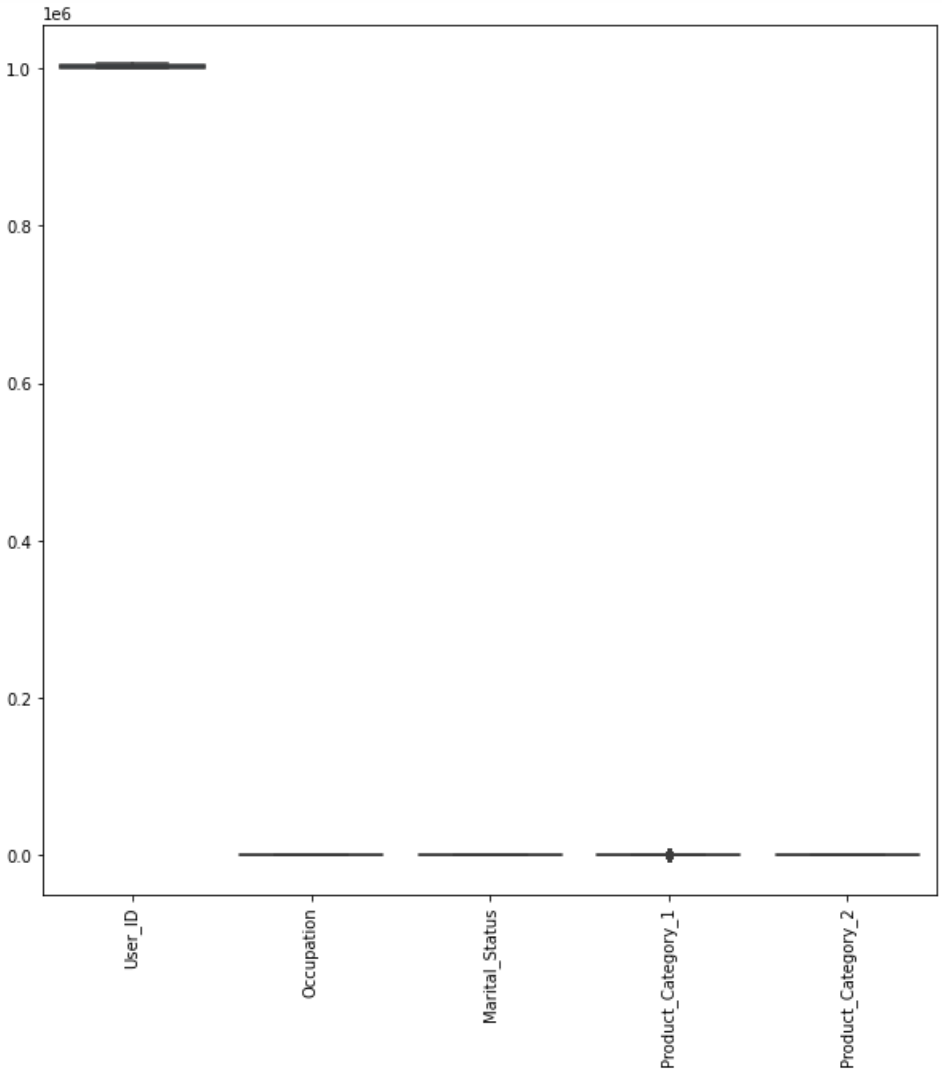
```
1 data.groupby(["City_Category"]).mean()[["Product_Category_1", "Product_Category_2"]].plot.bar(title="Gender
2 sns.despine()
3 #Products under category 2 are our most popular items, and this is true for all kinds of customers coming f
```



Checking for outliers

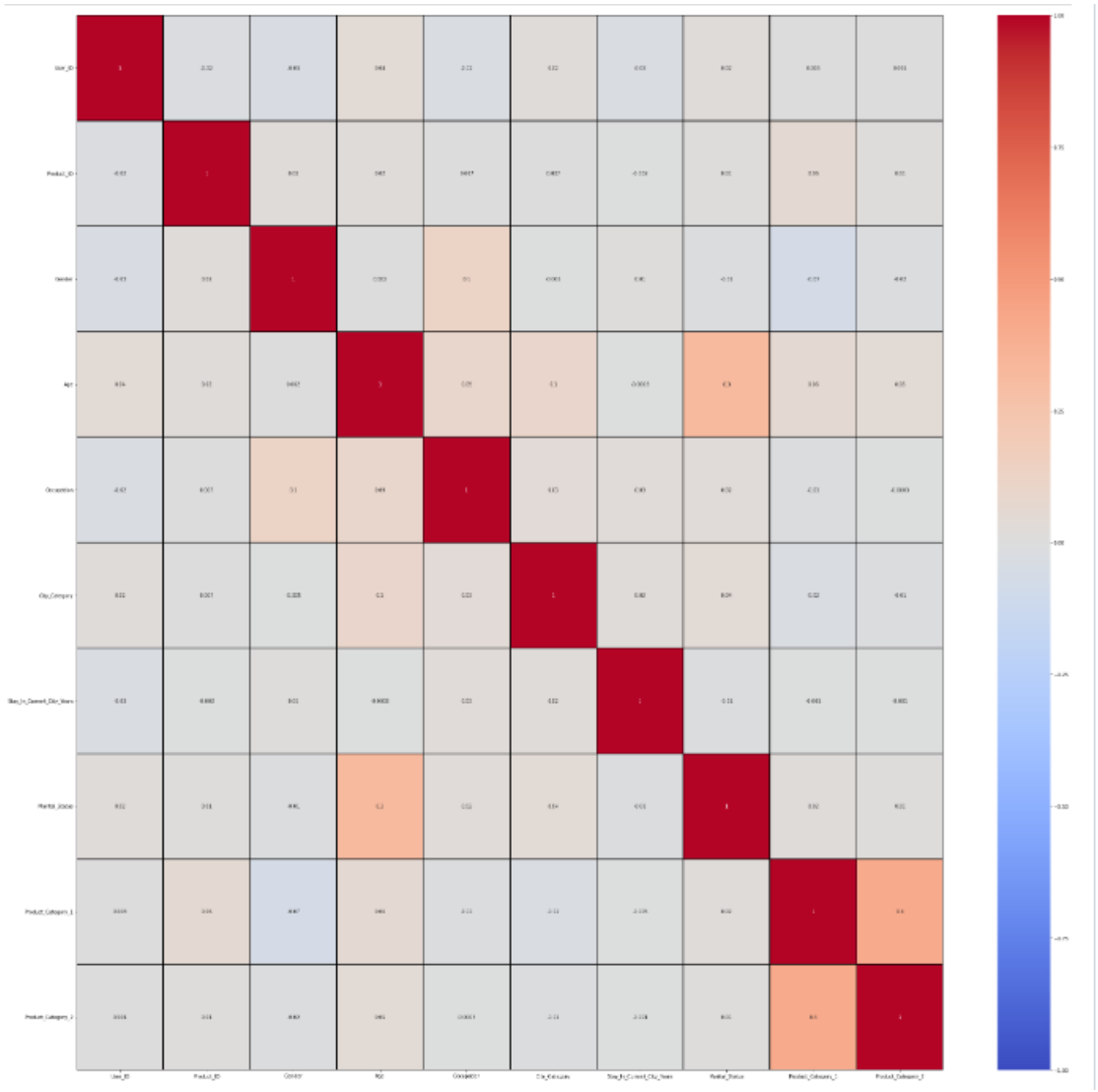
Train data





We can see some outier in Product category1 in the dataset in the dataset

Heatmap



Scaling the train data using standard scaler:

```
from sklearn.preprocessing import StandardScaler
scaler=StandardScaler()
X = pd.DataFrame(scaler.fit_transform(x), columns=x.columns)
```

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
vif=pd.DataFrame()



vif["vif_Features"]=[variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif["Features"]=X.columns
vif.sort_values('vif_Features')
```

	vif_Features	Features
6	1.002505	Stay_In_Current_City_Years
1	1.004725	Product_ID
0	1.005194	User_ID
5	1.012193	City_Category
2	1.020348	Gender
4	1.024402	Occupation
7	1.113365	Marital_Status
3	1.136018	Age
9	1.200805	Product_Category_2
8	1.212307	Product_Category_1

```
#No multicolinearity issue  in  dataset.
```

3.6 Interpretation of the Results

- ✓ The dataset was very challenging to handle it had 11 features with 550068 samples.
- ✓ Firstly, the datasets were having any null values, so I have used imputation method to replace the nan values.
- ✓ And there was number of unnecessary entries in all the features so I have used feature extraction to get the required format of variables.
- ✓ And proper ploting for proper type of features will help us to get better insight on the data. I found both numerical columns and categorical columns in the dataset so I have choosen line and bar plot to see the relation between target and features.

- ✓ I notice a few outliers and skewness in the data so we have choose proper methods to deal with the outliers and skewness. If we ignore this outliers and skewness we may end up with a bad model which has less accuracy.
- ✓ Then scaling dataset has a good impact like it will help the model not to get biased. Since we have removed outliers and skewness from the dataset so we have to choose Standardisation.

4.CONCLUSION

4.1 Key Findings and Conclusions of the Study

In this project report, we have analysed data to predict the trend in the dataset. We have mentioned the step by step procedure to analyze the dataset and finding the correlation between the features. Thus we can select the features which are correlated to each other and are independent in nature. The Data is ready for Model Building

4.2 Learning Outcomes of the Study in respect of Data Science

I found that the dataset was quite interesting to handle as it contains all types of data in it. Improvement in computing technology has made it possible to examine social information that cannot previously be captured, processed and analysed. New analytical techniques of machine learning can be used in used for Black friday Prediction research. The power of visualization has helped us in understanding the data by graphical representation it has made me to understand what data is trying to say. Data cleaning is one of the most