FLIP ROBO

NAME OF THE PROJECT:

A Project Report on

Black Friday Price Prediction Sales

Submitted by:

A KISHORE KUMAR

Batch No.:

Internship_33

Under the guidance of:

Mr. Shwetank Mishra

(FlipRobo SME)

ACKNOWLEDGMENT

I wish to express my sincere thanks and deep sense of gratitude to “Flip Robo” team, who has given me this opportunity to deal with an informative dataset and it has helped me to improve my problem analyzation skills and Machine Learning modelling.

Also, I want to express my huge gratitude to Mr. Shwetank Mishra (SME Flip Robo) for his tremendous support, to get me out of all the difficulties I faced while going through this project and for the successful completion of this work. He has been a great source of inspiration to work with and I shall always cherish my association with him with immense pleasure.

Finally, A huge thanks to “Data trained” who gave me the opportunity to get the Internship at Flip-Robo.

References use in this project:

1) SCIKIT Learn Library Documentation.
2) Blogs from towardsdatascience, Analytics Vidya, Medium.
3) Andrew Ng Notes on Machine Learning (GitHub).
4) Data Science Projects with Python Second Edition by Packt.
5) Hands on Machine learning with scikit learn and tensor flow by Aurelien Geron.
6) C. M. Wu, P. Patil and S. Gunaseelan, "Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data," 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), 2018, pp. 16-20, doi: 10.1109/ICSESS.2018.8663760
7) Odegua, Rising. (2020). Applied Machine Learning for Supermarket Sales Prediction
8) S. Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," 2016 IEEE 6th International Conference on Advanced Computing (IACC), 2016, pp. 78-83, doi: 10.1109/IACC.2016.25.

INTRODUCTION

Black Friday marks the beginning of the Christmas shopping festival across the US. On Black Friday big shopping giants like Amazon, Flipkart, etc. lure customers by offering discounts and deals on different product categories. The product categories range from electronic items, Clothing, kitchen appliances, Décor. Research has been carried out to predict sales by various researchers. The analysis of this data serves as a basis to provide discounts on various product items. With the purpose of analyzing and predicting the sales, we have used three models. The dataset Black Friday Sales Dataset available on Kaggle has been used for analysis and prediction purposes. The models used for prediction are linear regression, lasso regression, ridge regression, Decision Tree Regressor, and Random Forest Regressor. Mean Squared Error (MSE) is used as a performance evaluation measure. Random Forest Regressor outperforms the other models with the least MSE score.

For a long history of several decades, Black Friday has been recognized as the largest shopping day of the year in the US. It is the Friday after Thanksgiving and for American consumers, it ignites the Christmas holiday shopping. For most retailers, it is the busiest day of the year. Black Friday is traditionally known for long lines of customers waiting outdoors in cold weather before the open hours. Sales are so high for Black Friday that it has become a crucial day for stores and the economy in general with approximate 30% of all the annual retail sales occurring in the time from Black Friday through Christmas making it the kick-off day for the busiest and most profitable season for many businesses. It is unofficially a public holiday in more than 20 states and is considered the start of the US Christmas shopping season. In 2018, US shoppers expected to drop $483.18 on the shopping holiday of holidays, which equates to $90.14 billion [1]. Although Black Friday is originally from America, it has become a universal recognition worldwide. Because consumers are eager to spend so much money during this period, retailers seriously look forward to good preparation for the shopping holiday [2]. In preparation for this day, retailers will typically hire more employees, stock their commodities, prepare new promotions, and decorate store layouts. Retailers rely on designing advertising campaigns to attract more customers into their stores and/or their online shops. In order to maximize their efforts and revenues, retailers enthusiastically understand how the consumers make shopping decisions that will assist them to achieve the most profits during the shopping season [3]. Many possible parameters that have been considered are presented in Figure (1). If retailers comprehensively understand their customers in terms of characteristics, behaviors and motivations in the previous shopping seasons, they can implement and develop more effective marketing strategies for specific customers categories [4,5,6]. The authors place the Black Friday challenge is an interesting opportunity to investigate the performance of several machine learning models. We decide to apply boosting-based models to the problem and see how would they perform. The objective is to predict the amount of purchase a consumer is willing to pay given several categorical and numerical features. The rest of this paper is as follows. First of all, in Section II, we briefly discuss the overview of technical background including ensemble learning, bagging and boosting that summarizes critical materials existing in the literature that is essential to solving the problems. In Section III, we evaluate and perform the approach to our experiment dataset. In Section IV, the authors discuss some important remarks of the proposed approach. Finally, Section V recapitulates the approaches, discuss achievements and further investigation.

The shopping sector has greatly evolved due to the Internet revolution. Most of the population takes into consideration online shopping more than the traditional method of shopping. The biggest perks of online shopping are convenience, better prices, more variety, easy price comparisons, no crowds, etc. The pandemic has boosted online shopping. Though online shopping keeps growing every year, the total sales for the year 2021 are expected to be much higher [16]. Black Friday originated in the USA and is also referred to as Thanksgiving Day. This sale is celebrated on the fourth Thursday of November once every year. This day is marked as the busiest day in terms of shopping. The purpose of organizing this sale is to promote customers to buy more products online to boost the online shopping sector. The prediction model built will help to analyze the relationship among various attributes. Black Friday Sales Dataset is used for training and prediction. Black Friday Sales Dataset is the online biggest dataset and the dataset is also accepted by various e-commerce websites [1]. The prediction model built will provide a prediction based on the age of the customer, city category, occupation, etc. The prediction model is implemented based on models like linear regression, ridge regression, lasso regression, Decision Tree Regressor, Random Forest Regressor. The paper further walks through various sections. Section I gives an introduction to the problem, section II illustrates the prior research done in this field, section III provides the data set description, section IV presents the proposed model, with the conclusion in the last section.

LITERATURE REVIEW

Ample research is carried out on the analysis and prediction of sales using various techniques. There are many methods proposed to do so by various researchers. In this section, we will summarize a few of the machine learning approaches.

C. M. Wu et al. have proposed a prediction model to analyze the customer's past spending and predict the future spending of the customer. The dataset referred is Black Friday Sales Dataset from analyticsvidhya. They have machine learning models such as Linear Regression, MLK classifier, Deep learning model using Keras, Decision Tree, and Decision Tree with bagging, and XGBoost. The performance evaluation measure Root Mean Squared Error (RMSE) is used to evaluate the models used. Simple problems like regression can be solved by the use of simple models like linear regression instead of complex neural network models.

Odegua, Rising have proposed a sales forecasting model. The machine learning models used for implementation are K-Nearest Neighbor, Random Forest, and Gradient Boosting. The dataset used for the experimentation is provided by Data Science Nigeria, as a part of competitions based on Machine Learning. The performance evaluation measures used are Mean Absolute Error (MAE). Random Forest outperformed the other algorithms with a MAE rate of 0.409178.

Singh, K et al have analyzed and visually represented the sales data provided in the complex dataset from which we ample clarity about how it works, which helps the investors and owners of an organization to analyze and visualize the sales data, which will outcome in the form of a proper decision and generate revenue. The data visualization is based on different parameters and dimensions. The result of which will enable the end-user to make better decisions, ability to predict future sales, increase the production dependencies on the demand, and also regional sales can be calculated.

S. Yadav et al have analyzed and compared the performance of K-Fold cross-validation and hold-out validation method. The result of the experimentations where k-fold cross-validation gives more accurate results. The accuracy results of K - Fold cross-validation were around 0.1 - 3% more accurate as compared to hold-out validation for the same set of algorithms.

Purvika Bajaj et al. have performed sales prediction based on a dataset collected from a grocery store. The algorithms used for experimentations are Linear Regression, K-Nearest Neighbors algorithm, XGBoost, and Random Forest. The result precision is based on Root Mean Squared Error (RMSE), Variance Score, Training, and Testing Accuracies. The Random Forest algorithm outperforms the other three algorithms with an accuracy of 93.53%.

Ramasubbareddy S. et al. have applied machine learning algorithms to predict sales. The dataset for the experimentation purpose is taken from Kaggle, named as Black Friday Sales Dataset. The algorithms used for the implementation of the system are linear regression, Ridge Regression, XGBoost, Decision Tree, Random Forest, and Rule-Based Decision Tree. Root Mean Squared Error is used as the performance evaluation measure. As per RMSE lower the RMSE value better the prediction. As a result, based on the RMSE rate Rule-Based DT outperforms other machine learning techniques with a RMSE rate of 2291.

Aaditi Narkhede et al. has applied machine learning algorithm in tracking sales at places like shopping center big mart to anticipate the demand of customers and handle the management of inventory accordingly the methods presented here are an effective method for data shaping and decision making. New ways that can better identify consumer needs and calculate marketing plans which will improve sales.

M.Sahaya Vennila et al. have analyzed, preprocessed, and applied machine learning techniques to predict sales. The dataset used for the analysis and experimentation purpose is Black Friday Sales Dataset from Kaggle. The dataset is preprocessed. K - Fold method is used for the purpose of splitting the dataset into training and testing datasets. The prediction model is implemented using Linear Regression, Decision Tree, Random Forest, Gradient Boost, and XGBoost. Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are used as the accuracy evaluation measures. As a result of experimentation, the Random Forest performed significantly with an accuracy of 77%, with an RMSE value of 2730 and MAE value of 2349.

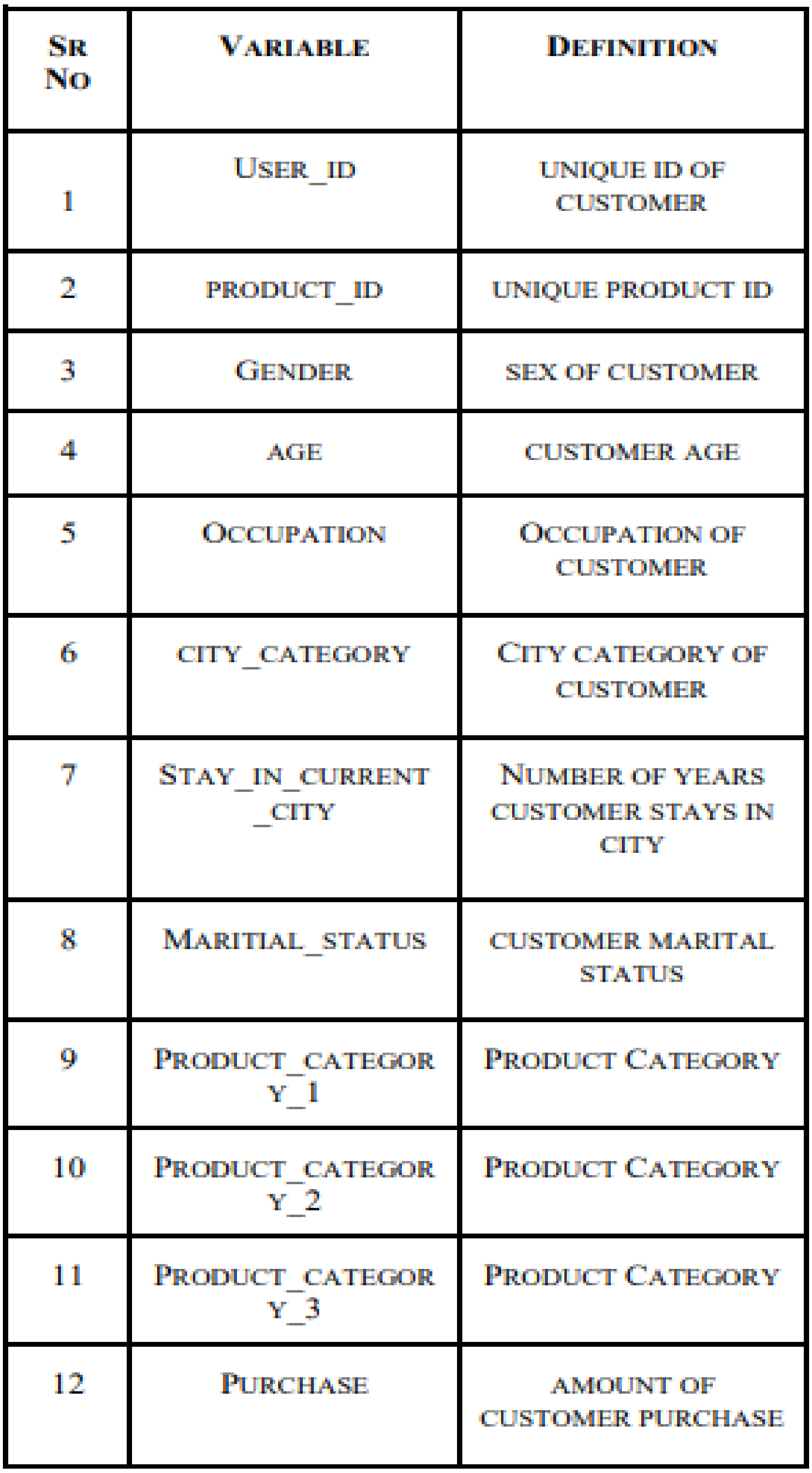
IMPLEMENTATION

A retail company “ABC Private Limited” wants to understand the customer purchase behavior (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.

The study uses Black Friday Sales Dataset [9] publicly available on Kaggle. The dataset consists of sales transaction data. The dataset consists of 5,50,069 rows. The dataset consists of attributes such as user_id, product_id, martial_status, city_category, occupation, etc. The dataset definition is mentioned in Table 1. The Black Friday Sales dataset is used for training various machine learning models and also for predicting the purchase amount of customers on black Friday sales [1]. The purchase prediction made will provide an insight to retailers to analyze and personalize offers for more customer's preferred products.

Feature Description of Dataset:



SR NO	VARIABLE	DEFINITION
1	USER_ID	UNIQUE ID OF CUSTOMER
2	PRODUCT_ID	UNIQUE PRODUCT ID
3	GENDER	SEX OF CUSTOMER
4	AGE	CUSTOMER AGE
5	OCCUPATION	OCCUPATION OF CUSTOMER
6	CITY_CATEGORY	CITY CATEGORY OF CUSTOMER
7	STAY_IN_CURRENT_CITY	NUMBER OF YEARS CUSTOMER STAYS IN CITY
8	MARITIAL_STATUS	CUSTOMER MARITAL STATUS
9	PRODUCT_CATEGORY_1	PRODUCT CATEGORY
10	PRODUCT_CATEGORY_2	PRODUCT CATEGORY
11	PRODUCT_CATEGORY_3	PRODUCT CATEGORY
12	PURCHASE	AMOUNT OF CUSTOMER PURCHASE

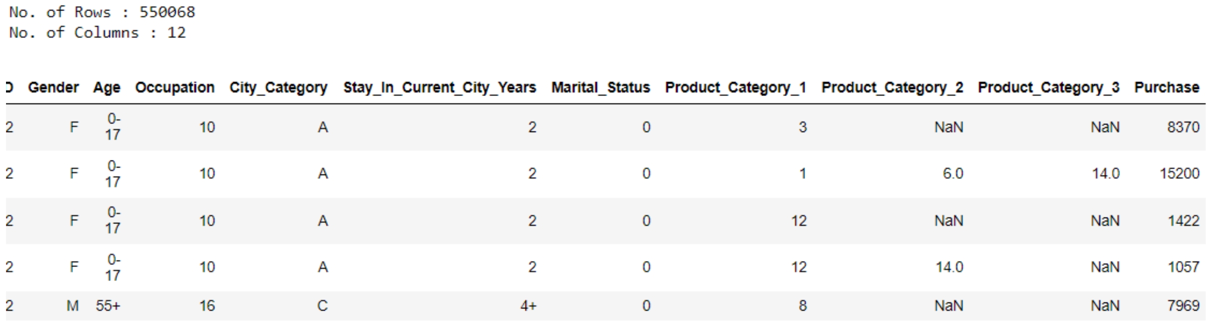
The Purchase Variable will be the predictor variable. The Purchase Variable will predict the amount of purchase made by a customer on the occasion of black Friday sales.

Data Sources and their formats

Data set provided by Flip Robo was in the format of CSV (Comma Separated Values). There are 2 data sets that are given. One is training data and one is testing data.

1) Train file will be used for training the model, i.e., the model will learn from this file. It contains all the independent variables and the target variable. The dimension of data is 5,50,068 rows and 12 columns.

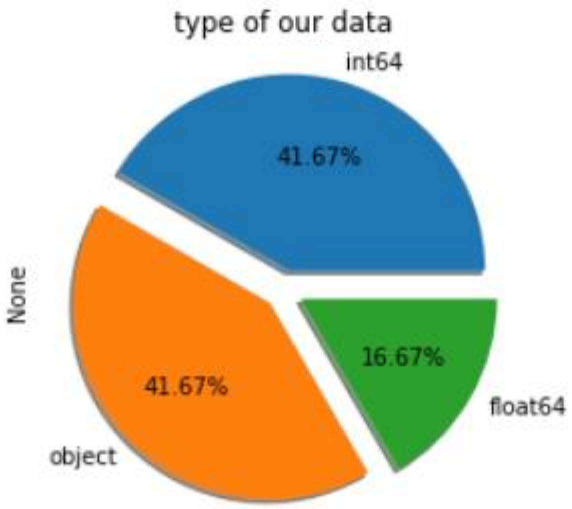
2) Test file contains all the independent variables, but not the target variable. We will apply the model to predict the target variable for the test data. The dimension of data is 2,33,600 rows and 12 columns.

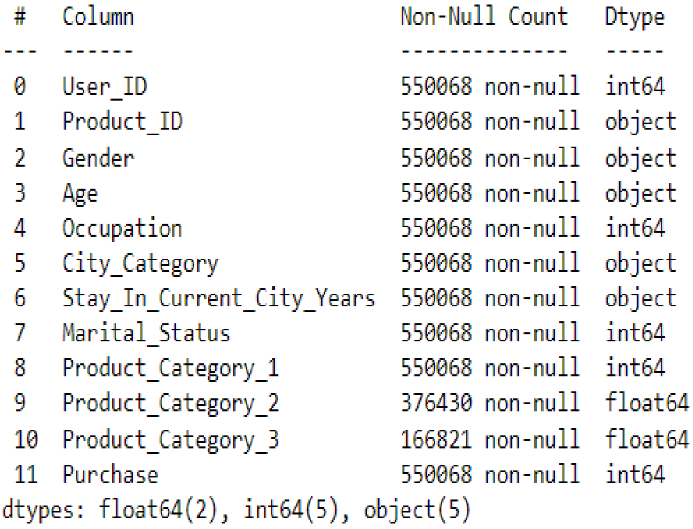


```
No. of Rows : 550068
No. of Columns : 12
```

D	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
2	F	0-17	10	A	2	0	3	NaN	NaN	8370
2	F	0-17	10	A	2	0	1	6.0	14.0	15200
2	F	0-17	10	A	2	0	12	NaN	NaN	1422
2	F	0-17	10	A	2	0	12	14.0	NaN	1057
2	M	55+	16	C	4+	0	8	NaN	NaN	7969

- The data types of different features are as shown below:



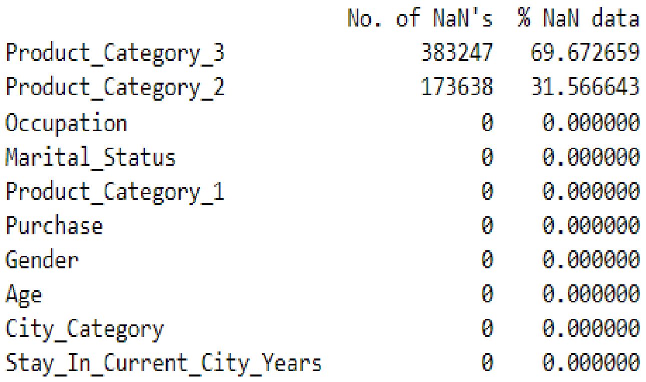


```
#   Column                      Non-Null Count   Dtype
---  ------                     --------------   -----
0   User_ID                     550068 non-null  int64
1   Product_ID                  550068 non-null  object
2   Gender                      550068 non-null  object
3   Age                         550068 non-null  object
4   Occupation                  550068 non-null  int64
5   City_Category               550068 non-null  object
6   Stay_In_Current_City_Years  550068 non-null  object
7   Marital_Status              550068 non-null  int64
8   Product_Category_1          550068 non-null  int64
9   Product_Category_2          376430 non-null  float64
10  Product_Category_3          166821 non-null  float64
11  Purchase                    550068 non-null  int64
dtypes: float64(2), int64(5), object(5)
```

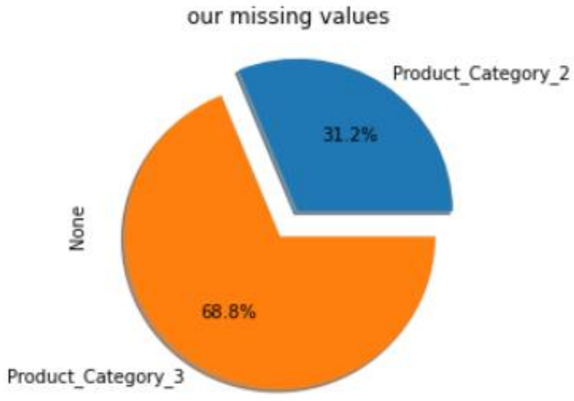
Dataset Information:

- Dataset consist of two CSV files, one for training model & other for testing dataset.
- Training dataset contain **550068 Rows** & **12 Columns**.
- There can be some **missing values** that are present in the dataset. With which we'll have to deal.
- We can **drop 'User_ID','Product_ID'**, because we want to be able to predict the buying behavior of any customer, not just those who are already recorded in the dataset.

- Some features contain missing values as shown below:

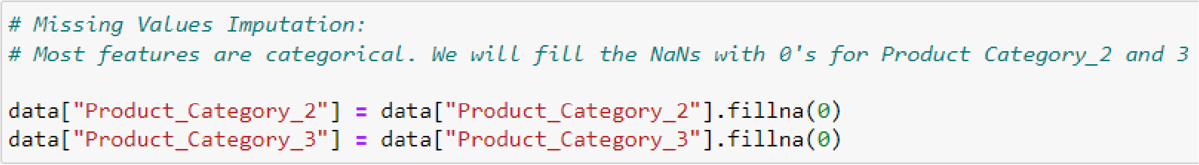


```
                           No. of NaN's  % NaN data
Product_Category_3               383247   69.672659
Product_Category_2               173638   31.566643
Occupation                            0    0.000000
Marital_Status                        0    0.000000
Product_Category_1                    0    0.000000
Purchase                              0    0.000000
Gender                                0    0.000000
Age                                   0    0.000000
City_Category                         0    0.000000
Stay_In_Current_City_Years            0    0.000000
```



Observations:

- Dataset contains both **Object & Numerical data types**.
- Dataset contains **numerical as well as categorical variables**.
- It would make sense to look at the summary statistics of the purchase variable. We see that the average purchase amount is 9263.97. The remainder of the variables don't make much sense to check summary statistics, as they are categories. Nevertheless, we have included them but won't examine them for insight.
- The dataset contains **many null values** in Product_Category_3 & Product_Category_2 coumns.
- *It makes sense for there to be missing values because some customers might only buy one category of product and not buy from the other 2 product categories. We should fill the missing value with a 0 to show this.*

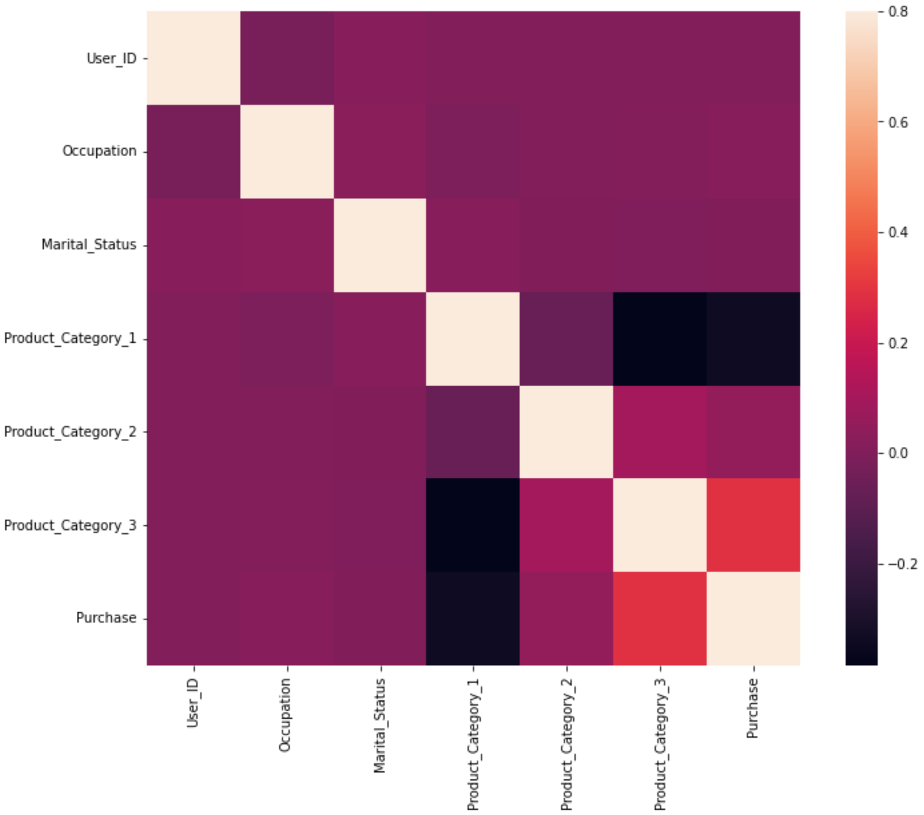


```
# Missing Values Imputation:
# Most features are categorical. We will fill the NaNs with 0's for Product Category_2 and 3

data["Product_Category_2"] = data["Product_Category_2"].fillna(0)
data["Product_Category_3"] = data["Product_Category_3"].fillna(0)
```

Exploratory Data Analysis (EDA)

Heatmap is used for determining the correlation between dataset attributes. The data of a given dataset can be easily represented graphically by using a Heatmap. It uses a color system to represent the correlation among different attributes. It is a data visualization library (Seaborn) element. Heatmap color encoded matrix can be described as lower the intensity of the color of an attribute related to the target variable, higher is the dependency of target and attribute variables. Based on the Black Friday Sales Dataset [9] the heatmap obtained gives output as Figure 2. The observation based on the heatmap is the attributes age and marital_status, product_category_3 and purchase have a correlation.



Variables correlated with Purchase:

- Product_Category_3 (positive correlation)
- Product_Category_1 (negative correlation)

Perhaps Product_category 3 is positively correlated because it offered cheaper products. So, let's see if this is true by checking the average price of the product categories.

The count plots for different attributes are visualized as different figures given below. The count plot for gender attributes is as Figure 3. Based on the count plot for gender attribute it is observed that feature M (Male) has the maximum count. The count for F features is less.

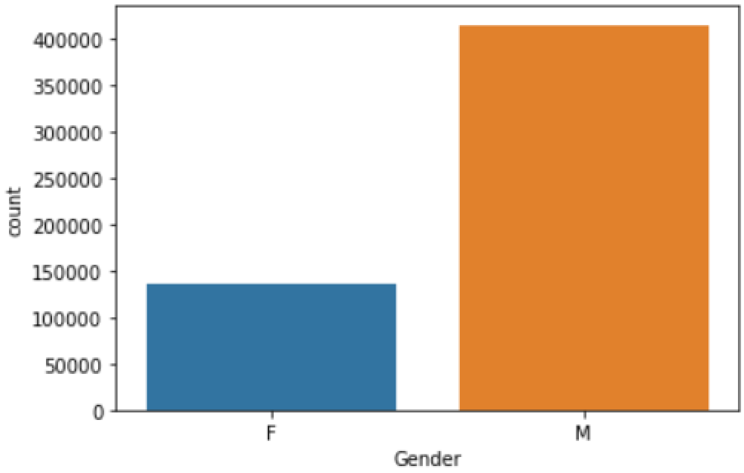
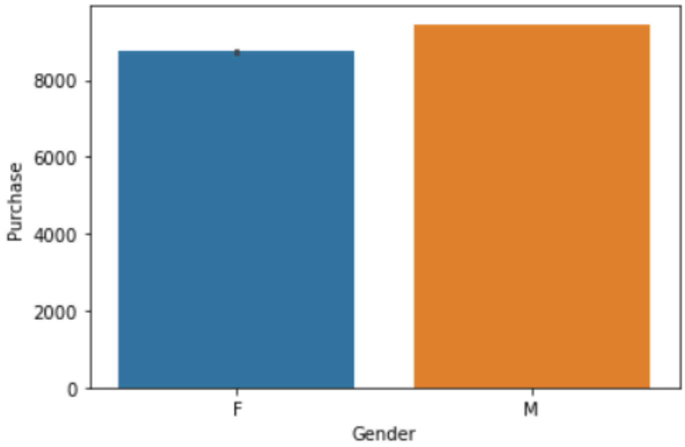


Fig.: Count Plot for Gender.

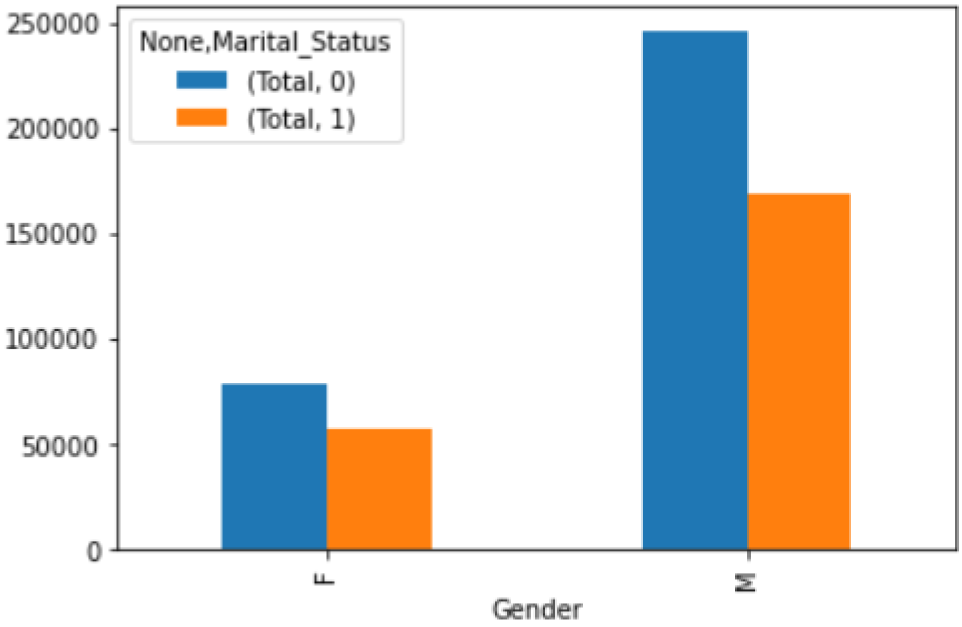
The count plot for the age attribute is as Figure. Based on the count plot the observations noted are the age group 26-35 has a maximum count. The second maximum count observed is for the age group 36-45. The third maximum count observed is for the age group 18-25.

We notice there are more male customers than there are female customers, including recurring customers.

Let's now check to see the female to male ratio against Purchase.



Male customers tend to have purchased more than female customers. There may be different reasons for this. First, we saw that the ratio of males to females is more which would mean more purchases. It could also be that married couples shop together and it is the husband that tends to purchase the products for his wife. Therefore, the reflection of more male purchasers doesn't necessarily reflect for whom the products are being purchased for. Also, there may be more male customers because the products target males more. It is interesting to see that although there are fewer female customers by count, the purchase amount is much more similar to the male's. This could mean that the products the females purchase are more expensive.



We notice that there are more single shoppers than there are married shoppers across gender. This does not support our previous belief that a reason for more male purchasers is attributable to the fact that more couples go shopping togther and the husband purchases the products. We clearly see that single shoppers are more prevalent across both groups.

Gender has potential to be a decent predictor for purchase. Next, we would like to explore the age category. While exploring, we should consider:

- if age has potential to be a good predictor for purchase

- check differences in gender across age groups

- check to see if there is much variation in the different age groups for purchases

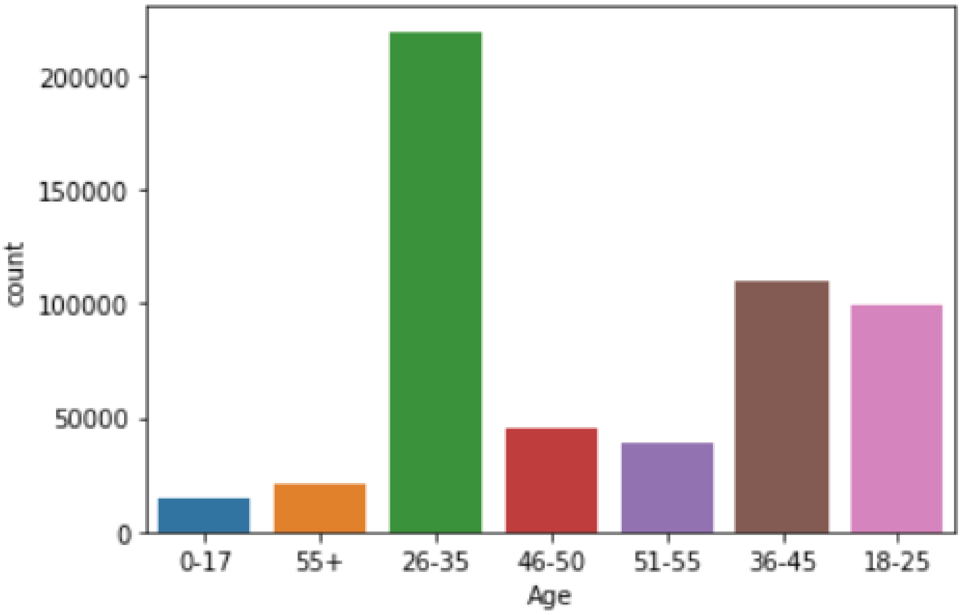
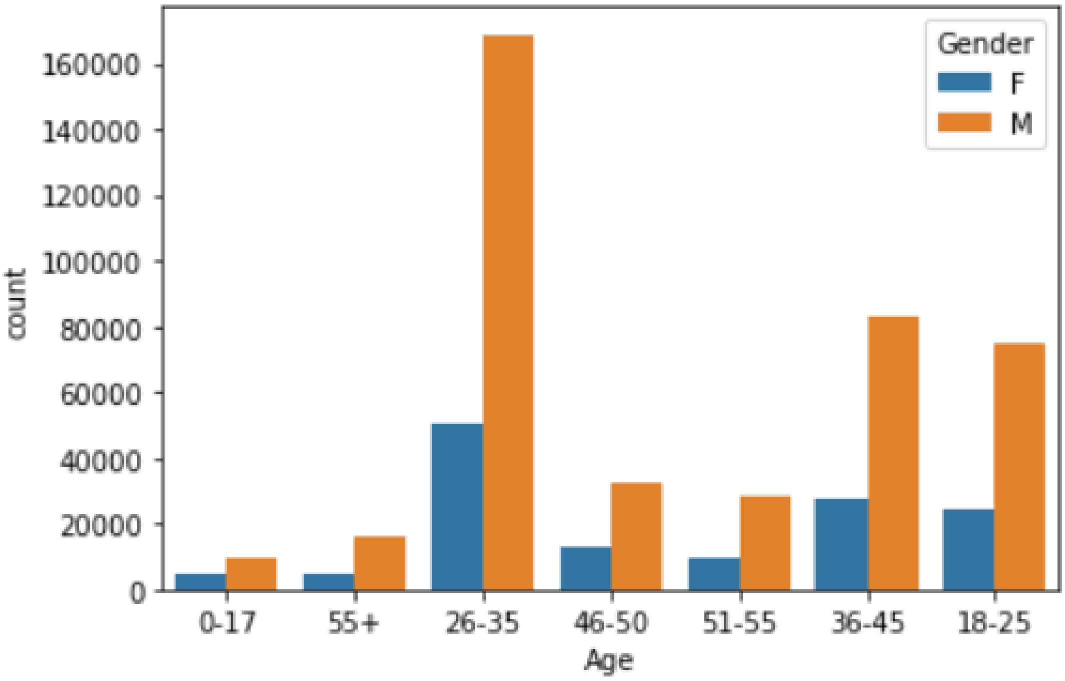


Fig.: Count Plot for Age.

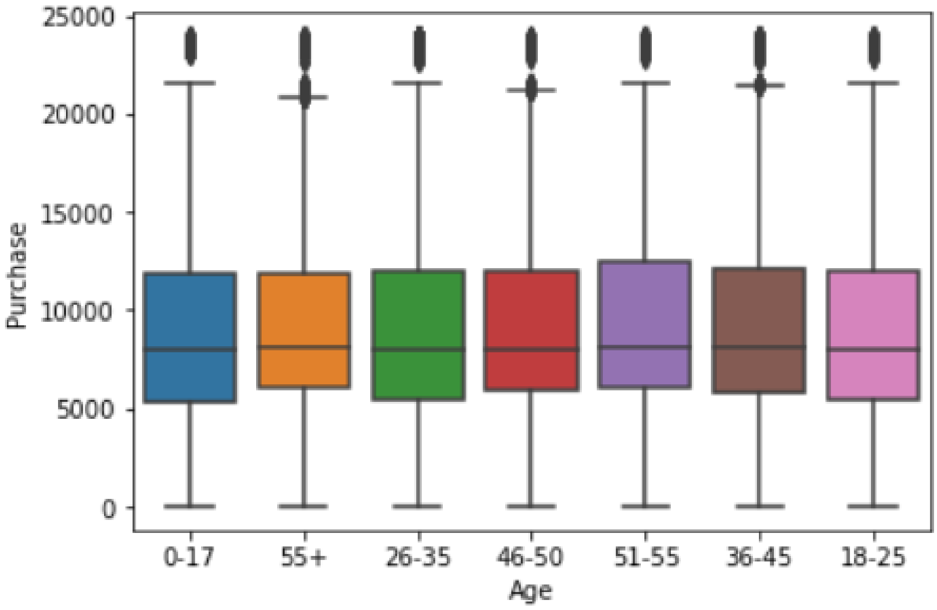
The count plot for the occupation attribute is as Figure 5. The observation based on the count plot is that the masked occupation 4 has maximum count. The second maximum based on the count plot is occupation 0.

From the bar plot above, we see that there are much more shoppers between the ages of 26-35 years than any of the other age groups. We understand from this that the product offerings are catered more towards this age group and one age group above (36-45) and one age group below (18-25) mainly. Since most shoppers are from these age groups, it would make sense for the business to continue stocking their inventory with more of these popular products to cater to these age groups. Knowing this, the business can focus on marketing for these specific age groups and spend the most on them since it is their largest customer composition. Furthermore, it can consider adding more variety in products if it wants to cater to all age groups and potentially increase sales.

Next, let's check the gender decomposition across the popular age groups to see if males are consistently more across the 3 popular age groups.

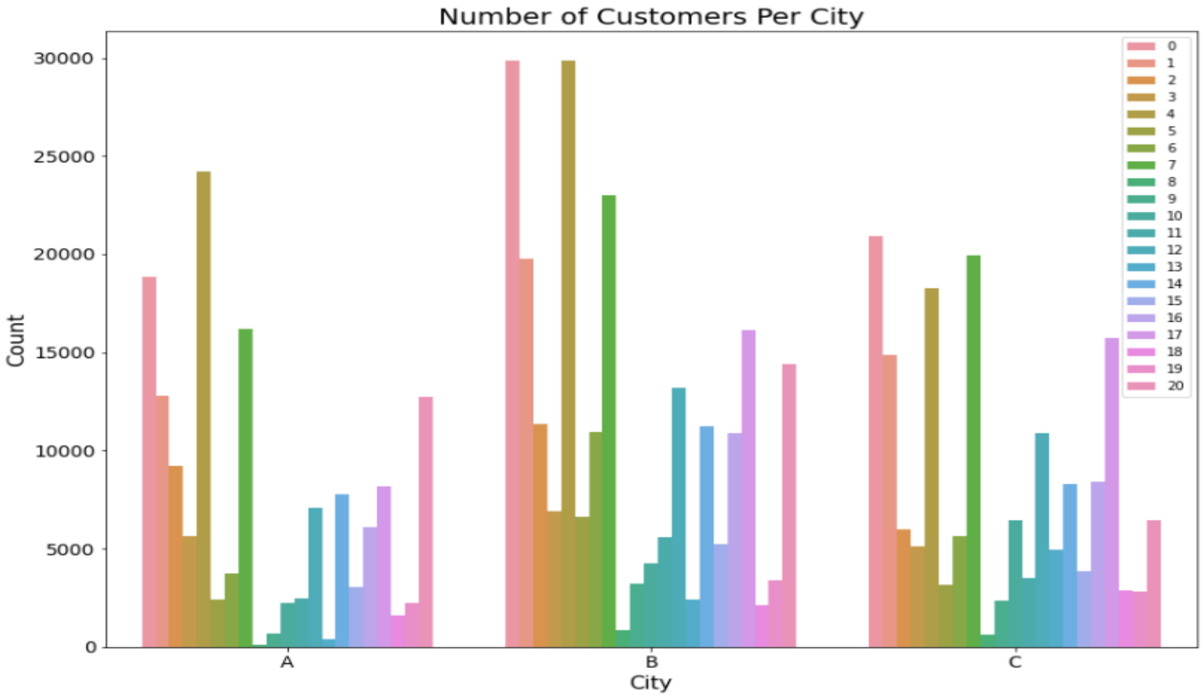


We verify that there are more male shoppers. The difference is especially noticeable in the age group of 26-35. Males between ages of 26-35 constitute the largest portion of clients. This can help the business better allocate its energy and money towards targeting males between these age groups and increasing inventory count of these products or increase product offering similar to the ones already selling. Knowing this age group uses social media, the business can also use social media as another way of reaching and attracting customers.

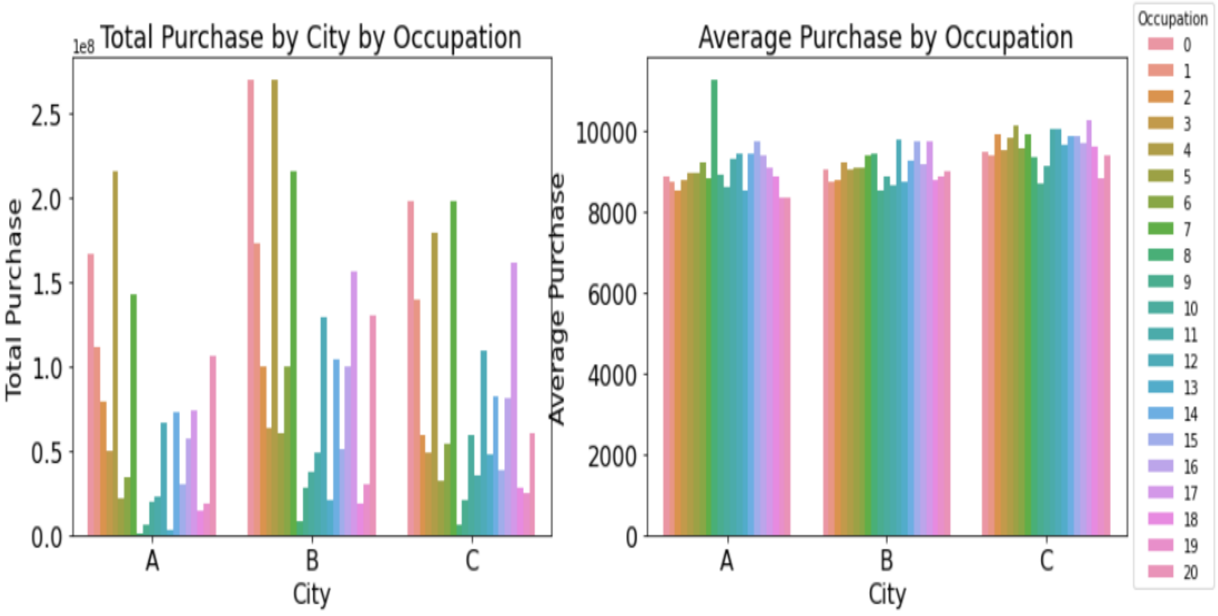


It is interesting to see that from the boxplot above there doesn't seem to be much variation in purchase amount across the different age groups. Maybe the less popular clients that fall outside the popular age groups stated earlier purchase more expensive products while the popular clients purchase the less expensive items which would balance and create small variation across age groups.

Check to see if there is variation in occupation across cities.

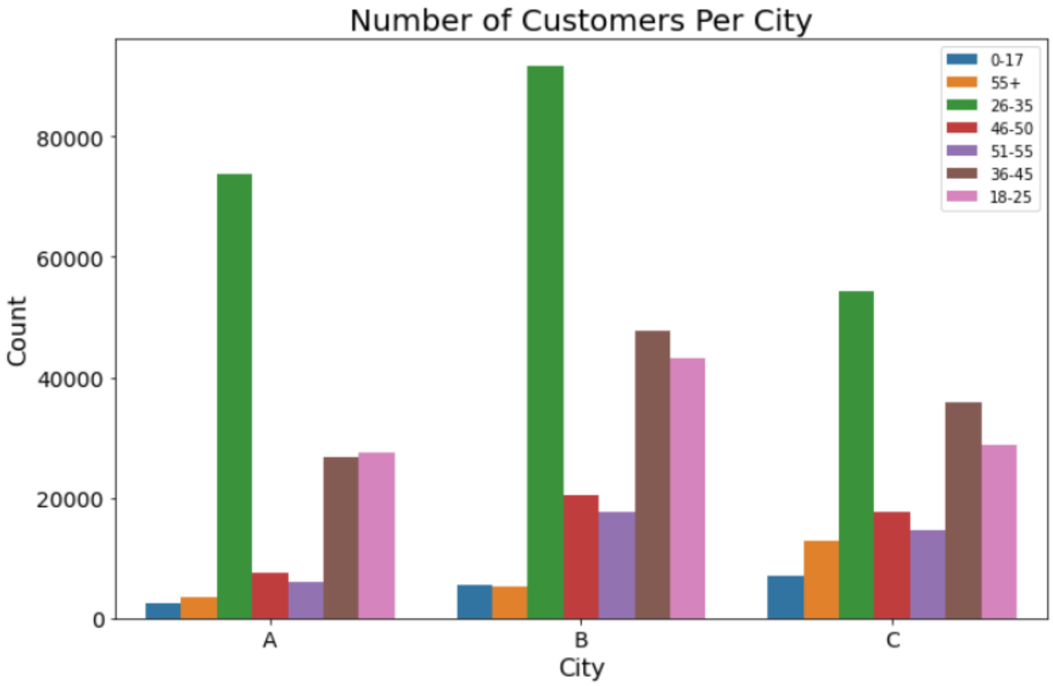


We see that in each of the cities, there seems to be a similar occupation distribution. Occupation 0, 4, 7 seem to be popular occupations among the customers across the three different cities. There doesn't seem to be much fluctuation across cities. There is more variation in occupation rather than across cities. Therefore, occupation seems to be a better indicator of likelihood of being a customer rather than the city defining it.



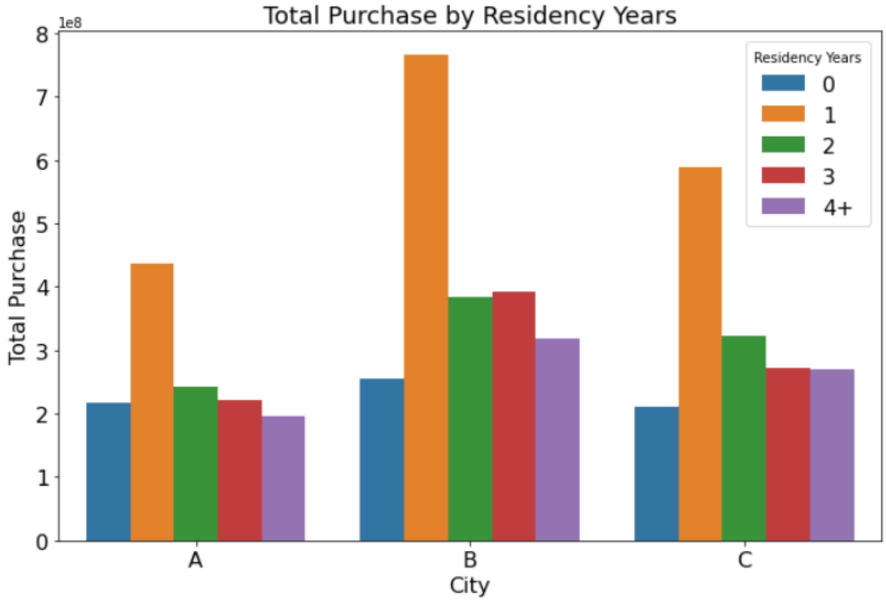
From the previous plot, we notice the same distribution and see how total number of customers reflects total purchases so they are positively correlated. From the plot on the right of the average purchase by occupation, we can see how occupation 8 from city A stands out from the rest. It is interesting that the average purchase was higher for Occupation 8 only in City A. This seems to be an outlier in the data.

Similarly, we next would like to see the variation in age groups across the different cities.



We see a similar distribution of the different age groups across the 3 cities. The most popular age group of 26-35 is mostly from City B. It seems that the most purchases come from City B so the business can target this city more by increasing its advertising to City B.

We next would like to see how the relation between residency years and total purchase.



We once again see a similar residency duration distribution across the 3 cities. It seems to be that customers who have been a resident for 1 year have contributed to total purchases. This is apparent across all 3 cities in orange. New residents are probably not as acquainted with the city and business and have the lowest contribution to total purchases.

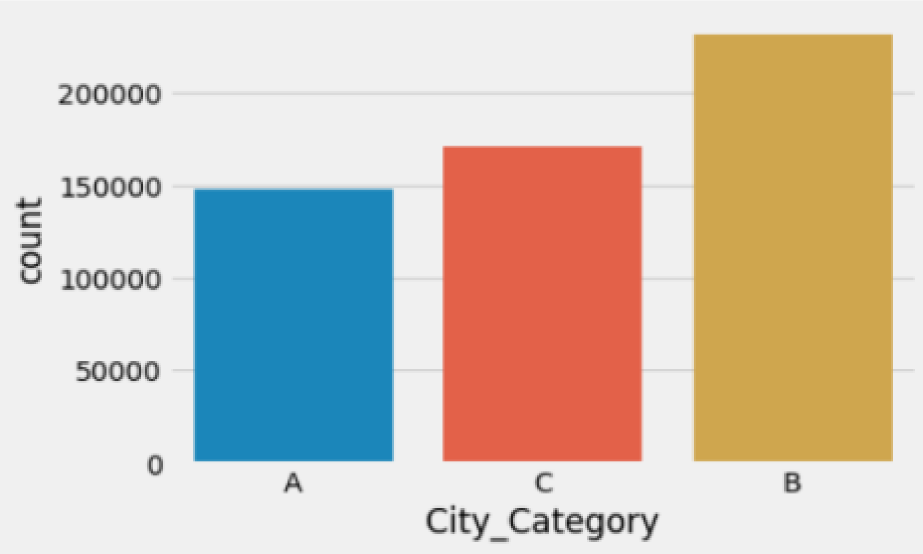


Fig.: Count Plot for City_Category

The count plot for Stay_In_Current_City is as given in Figure 7. The observations based on the count plot can be stated as the maximum count is for 1 year. The minimum count is for 0 years.

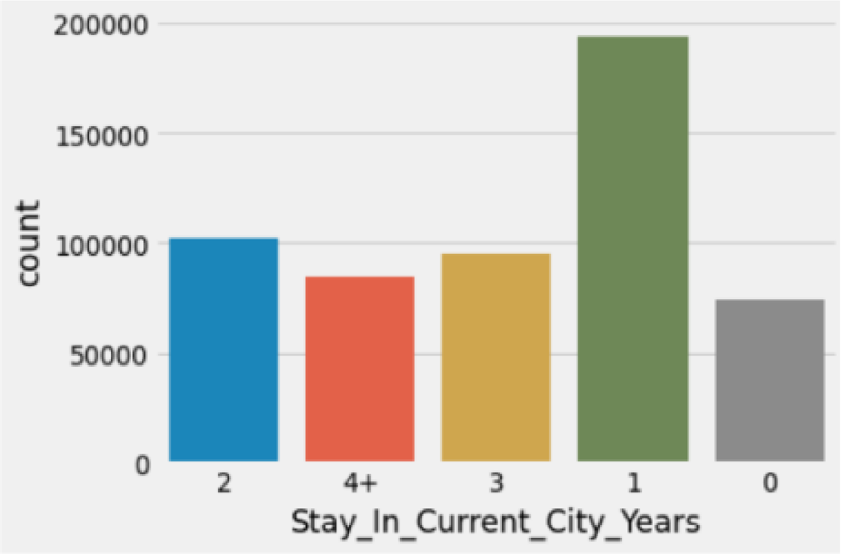
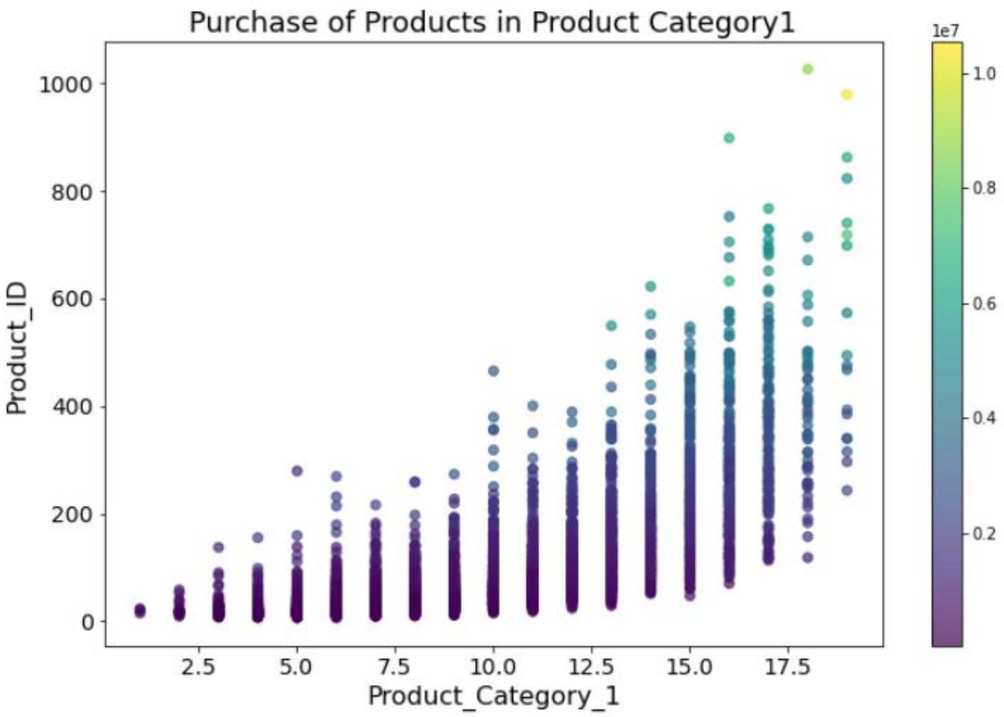


Fig.: Count Plot for Stay_In_Current_City_Years



Product ID 1000 from Product Category 1 seems to be the highest purchase amount.

CONCLUSION

The key purpose of this study is to investigate the Exploratory Data Analysis (EDA) techniques in extremely noise data. More specifically, the current results have confirmed the effective strategy to predict the amount of purpose. The Black Friday challenge is still operating, so much further consideration can be made to improve the result. This project strongly indicates the need of more advanced model's tuning and feature engineering.

---***--