Project Report

Marketing Strategy Development for a Retail Store

Abhijit Menon

Jing Wang

Phaniraj Bhatkal Goverdhan

Abstract

Data analysis of sold goods and customers' information is an effective marketing strategy for retail stores to increase sales amount. In this project, we select a database of retail stores as the object, including the amount of purchase of different types of goods, as well as the information of customers' gender, age, city, occupation, marital status and so on. We use Python and Excel as data processing tools. The main data analysis steps are data acquisition, data cleaning, data visualization and data modelling. This report shows the process and results of data analysis, including the code of python and data distribution image. From the above analysis, we get some conclusions on how to develop this retail store's sales amount.

Data Acquisition

Data Acquisition is the process of collecting and storing the data from various sources. We have taken the data from Kaggle, a website with multiple free datasets. Our dataset consists of a retail stores data on the sales from Black Friday 2017.

Weblink:https://www.kaggle.com/mehdidag/black-friday

Columns of the data set include:

1) User ID
2) Product ID
3) Gender
4) Age
5)Occupation -> 20 different occupations have been categorised numerically between 1-20.
6) Category of City -> Cities are classified as tier A,B and C.
7) Number of years in a city -> Classified as 1,2,3, and 4+ years in a city.
8) Marital Status -> 0 is unmarried and 1 is married.
9) Product Category -> There are 3 different type of product categories in 3 different columns and the specific product in the product type picked up in each category is given.
10) Purchase -> The amount spent by the customer is given under purchase.

Data Cleaning

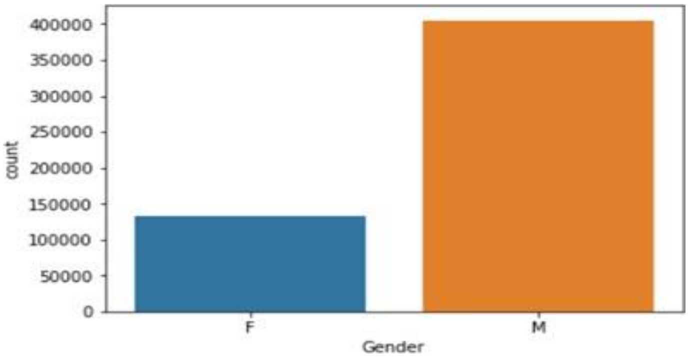
It is the process of analyzing, identifying and correcting messy, raw data from the database. Here we are using excel and python to clean the records so as to keep only the required ones. First we replace na/nan values in the data frame with 0. Then replace 4+ years in the Stay_In_Current_City_Years column with a 4 because it is a string and an ML model requires integers to draw conclusion. Similar techniques are used in the other columns to replace string instances with numeric data. Categories are replaced with 0,1,2 hence creating numeric bins of data.

Data Visualization

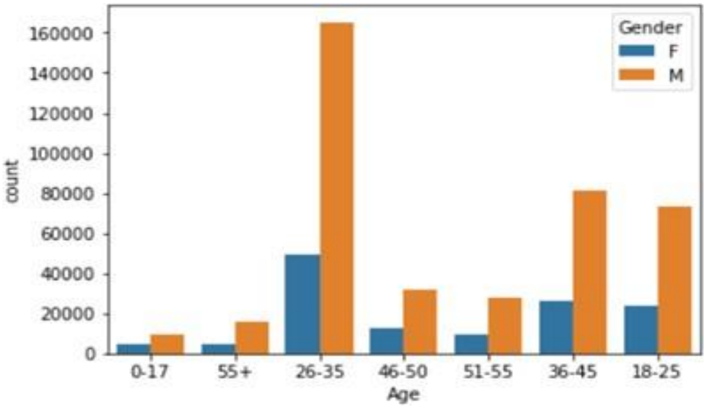
Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. With interactive visualization, you can take the concept a step further by using technology to drill down into charts and graphs for more detail, interactively changing what data you see and how it's processed.

The following are the the visualisations of the data from our data set and the insights we can draw from them.

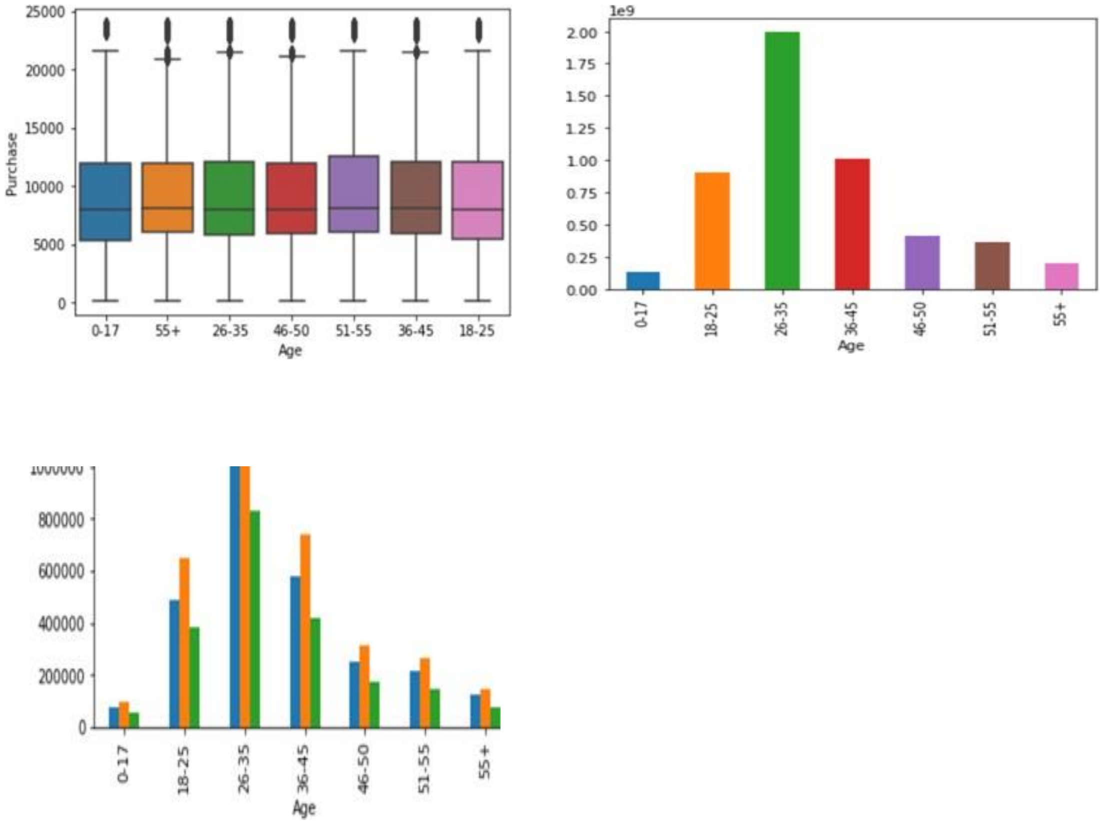
Age,Gender and Population visualisation



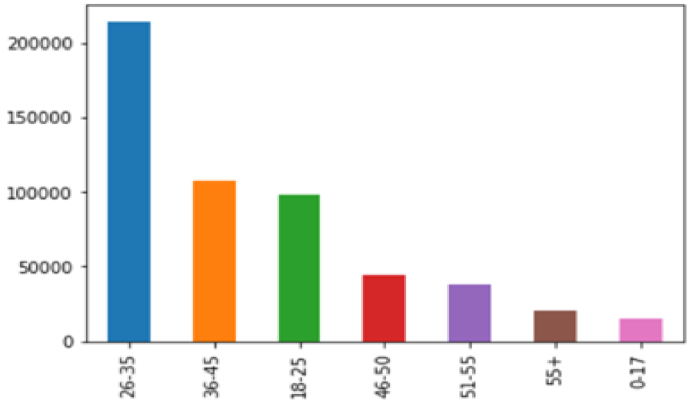
The above graph depicts the number of females and males who purchased on black Friday.



The above graph depicts the count ratio of female and male age groups.



Box plot and Bar chart plotting amount spent in each product category by various age groups.



The above graph depicts the purchase done by certain age group during Black Friday.

Insights Drawn:

1. For female customers, the age group of 51-55 bought most, the age group of 0-17 bought least.

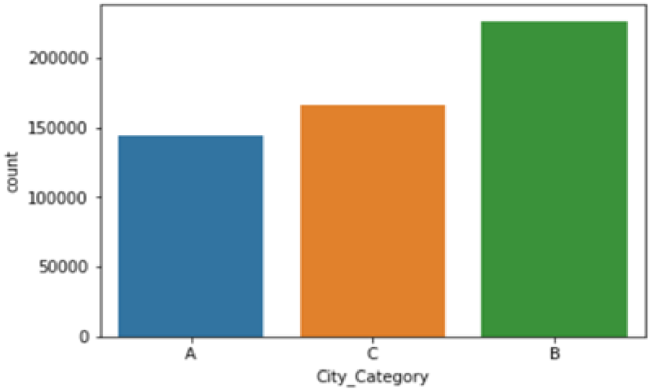
2. For male customers, the age group of 51-55 bought most, the age group of 0-17 bought least.

3. For total customers, the age group of 51-55 bought most, the age group of 0-17 bought least.

4. Male bought more than female.

<u>City wise purchase</u>

City	A	B	C
Purchase	8958	9199	9844

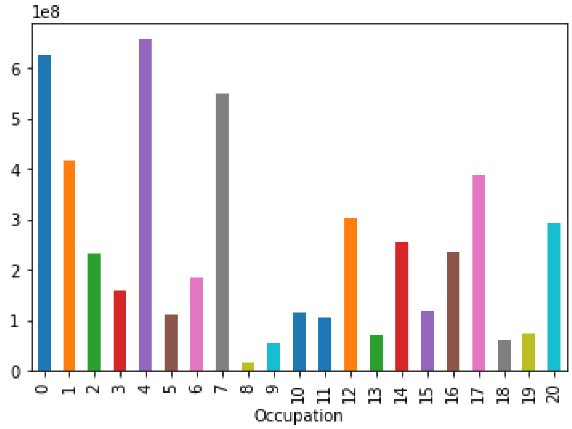


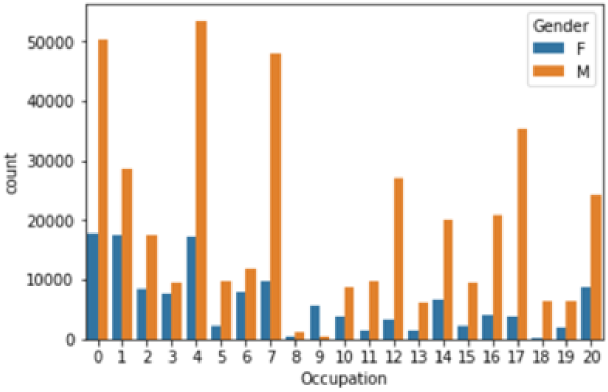
The above graph depicts the count ratio of three city tiers.

Insight Drawn:

The city C bought most, and city A bought least.

Occupation :

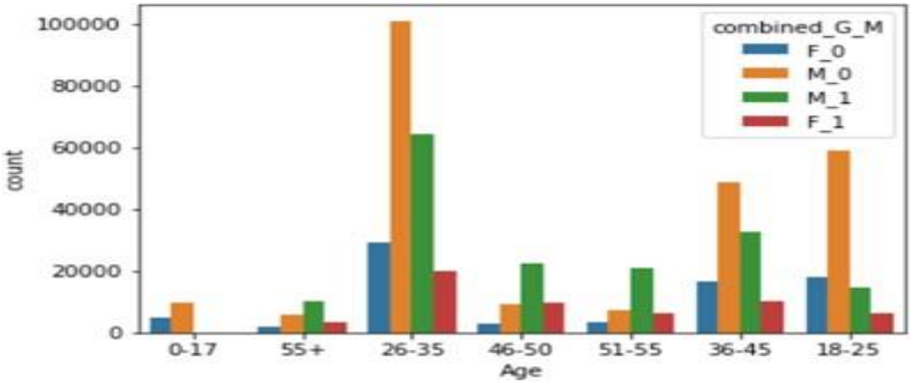




Insight Drawn:

1. Occupation 4 bought most and occupation 8 bought least.
2. Marketing can be skewed towards occupation 4.
3. Females from Occupation 0 and 1 buy the most.

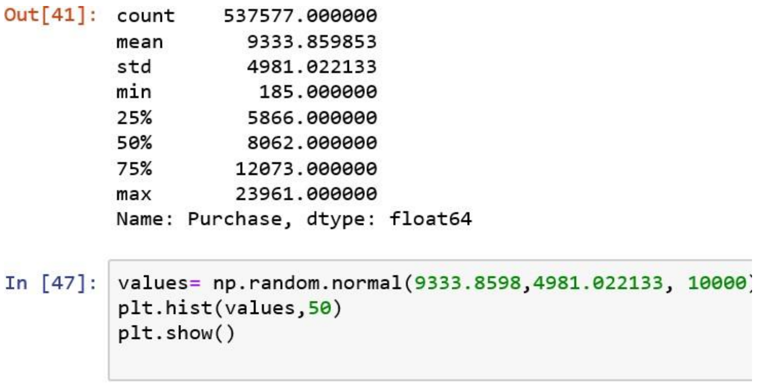
Marital Status :



The above graph depicts the count of female and males who are either married or unmarried.

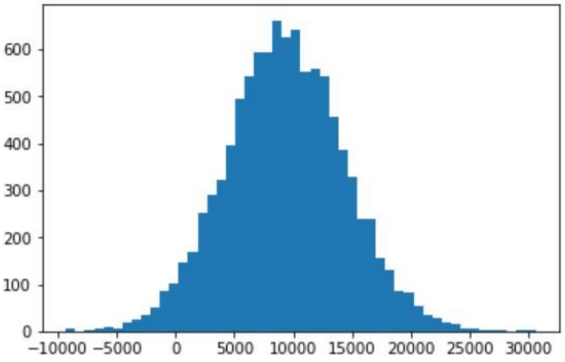
Insight Drawn : Unmarried males of 18-25 surprisingly purchase a lot.

Converting our data into a normally distributed curve on the basis of purchase from each customer.



```
Out[41]: 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
```

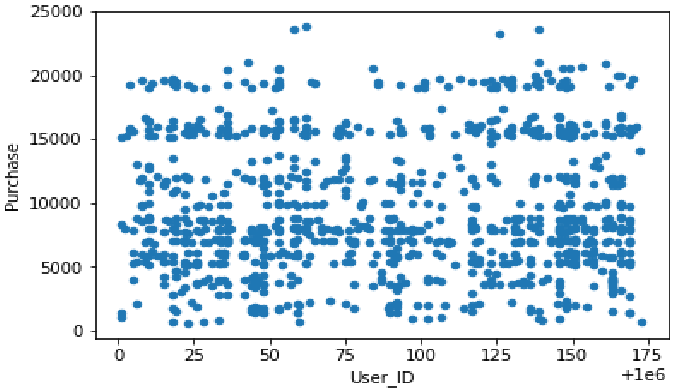
```
In [47]: values= np.random.normal(9333.8598,4981.022133, 10000)
         plt.hist(values,50)
         plt.show()
```

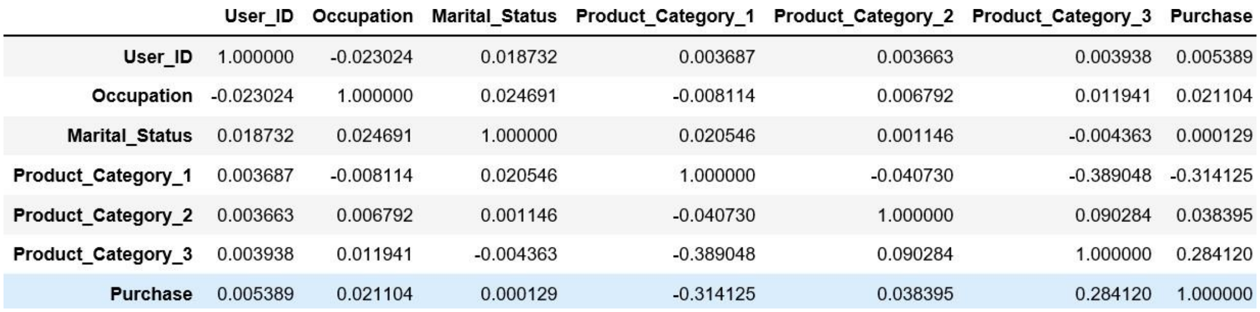


Data Modelling

Data modelling is a process to be able to develop a function using an algorithm that can be used to predict the the nature of the outcome.

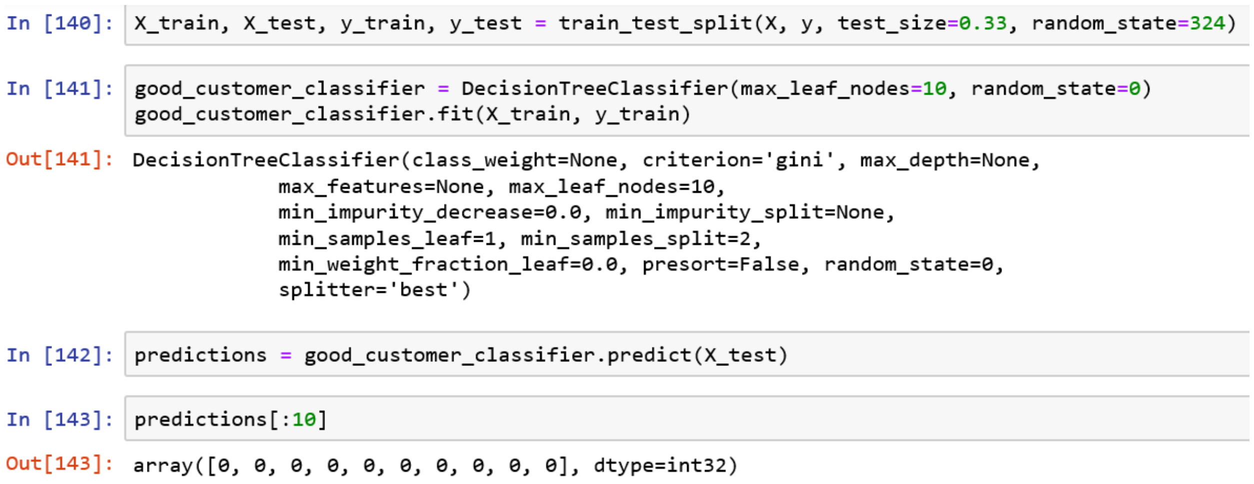
In the case of our data, we had a lot of string inputs for various column classifications, so as a part of data cleaning we converted them to numerical input. Further, we needed to determine the type of algorithm that will be best suited for this kind of data set. Below is a scatter plot and the correlation statistics of buying customers' characteristics .





	User_ID	Occupation	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
User_ID	1.000000	-0.023024	0.018732	0.003687	0.003663	0.003938	0.005389
Occupation	-0.023024	1.000000	0.024691	-0.008114	0.006792	0.011941	0.021104
Marital_Status	0.018732	0.024691	1.000000	0.020546	0.001146	-0.004363	0.000129
Product_Category_1	0.003687	-0.008114	0.020546	1.000000	-0.040730	-0.389048	-0.314125
Product_Category_2	0.003663	0.006792	0.001146	-0.040730	1.000000	0.090284	0.038395
Product_Category_3	0.003938	0.011941	-0.004363	-0.389048	0.090284	1.000000	0.284120
Purchase	0.005389	0.021104	0.000129	-0.314125	0.038395	0.284120	1.000000

From what we can see, we notice the buying pattern is far too scattered and the correlation is too weak to get an accurate regression line to pass through it.Hence we move onto the next algorithm. We picked decision trees. A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.



```
In [140]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=324)

In [141]: good_customer_classifier = DecisionTreeClassifier(max_leaf_nodes=10, random_state=0)
          good_customer_classifier.fit(X_train, y_train)

Out[141]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                      max_features=None, max_leaf_nodes=10,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, presort=False, random_state=0,
                      splitter='best')

In [142]: predictions = good_customer_classifier.predict(X_test)

In [143]: predictions[:10]

Out[143]: array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0], dtype=int32)
```

For this model we build a testing data and a training data. The testing data consists of 33% of the whole data set in this case and is used to train the model as to how customer behaviour changes according to their background.

Our purchase data summary was as follows :

count	mean	std	min	25%	50%	75%	max
537577.000000	9333.859853	4981.022133	185.000000	5866.000000	8062.000000	12073.000000	23961.000000

Based on these statistics we decided that customers who bought for above the 75% of the people on the normal distribution to be good customers (Denoted by 1) and the rest to be not so good customers(Denoted by 0).

After running the decision tree model we were able to predict the nature of the customer(Good or not so good) with an 84.411% accuracy.

Conclusion

Based on our visualization and modelling the marketing strategy becomes clearer for the retail store. Based on our inferences we can have the store target certain customers that are buying more of a certain type of product while also being able to determine if a person is more likely to be a good customer or not through the decision tree model. This information can be used for targeted advertising through social media such as facebook and google adwords to be able to generate further profit.

Github link to code:

https://github.com/Phaniraj201095/Marketing-Strategy-Development-for-a-Retail-Store/blob/main/Black_Friday_Analysis.ipynb