# A picture containing clipart Description automatically generated



**Black Friday Sales Analysis and Prediction**

by

House Stark

Submitted by

A. Priyanka

K. Pruthvi raj

P. Anish

**INDEX**

1.Introduction--------------------------------------------------------3

* Python
* Machine Learning

2.Objectives of Research-----------------------------------------3-4

3.Problem Statement------------------------------------------------4

4.Review of Literature----------------------------------------------5

5.Data Collection----------------------------------------------------6

6.Methodology---------------------------------------------------8-23

* Exploratory Data Analysis
* Figures and Tables
* Data Modelling

7.Findings and Suggestions-----------------------------------24

8.Conclusion-------------------------------------------------------25

**1.Introduction**

**Python**

python v 3.2 is used in this project because of many reasons like it is open source, free to use. machine independent , and it contains many modules with many functionalities.

Python is the mostly used language to do machine learning because of its numerous modules and methods.

**Machine Learning**

Machine learning is the science of getting computers to act without being explicitly programmed. It is a branch of [artificial intelligence](https://www.sas.com/en_in/insights/analytics/what-is-artificial-intelligence.html) based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

Machine Learning basically consists of two types of learning they are Supervised learning (parametric/non-parametric algorithms, support vector machines, kernels, neural networks) and Unsupervised Learning (clustering, dimensionality reduction, recommender systems, deep learning), based on the availability of the outputs

(In the data set that is available since we are having both the inputs and the output that we want to predict so we are using Supervised Learning.)

**2.Objectives of Research**

Some of the main objectives for this research are mentioned below.

>Implementing several Machine learning algorithms using python and jupyter notebook.

>In this research we are analysing some data so that we can bring forth the relationship between many independent variables with the respective dependent variable.

>To generate models that are trained on some data sets so that they can predict some accurate value based on the inputs.

>creating a user interface/website for the same model by using watson studio and node red, so that everyone can use is use this model to predict values.

**3.Problem Statement**

Observation and analysis of sales on normal days is boring and linear in nature, so let us take Black friday sales.

Black friday sales is generally celebrated on the fourth Friday of November after thanksgiving , where one can find huge discounts and offers on many products.

On this day the sales of different products are huge because of the discounts and offers.so the seller need to predict the sales before the day based on general data of a person living nearby to the place where the sales are going on. various independent values that effect these sales are the persons age, gender, city they are living in, since how many years they are living there, occupation, etc.

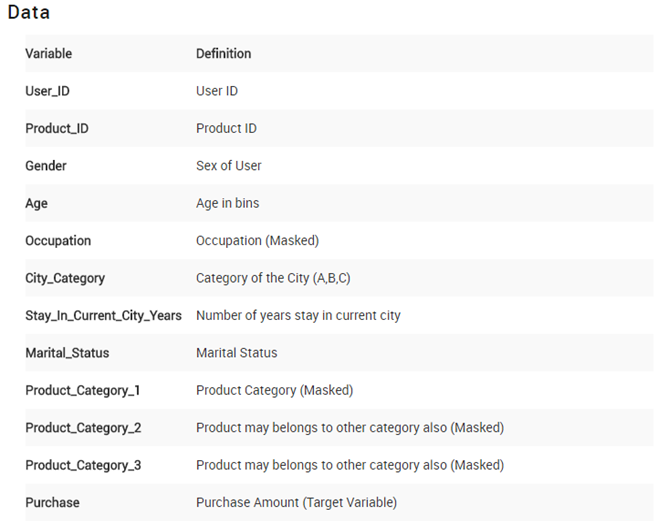
We are also going to analyse the impact of each variable on the sales on this day.

We are going to generate a model using different algorithms that can predict accurately the sales values based on several independent variables like Age, Gender, City, Occupation, etc.

**4.Review of Literature**

Available data

This is the current data they have available:



If we analyse it individually we see that we do not have any information regarding the stores. Moreover, there is some information related to the customer such as age group, sex, occupation and marital status. On the other hand, we have data on the city’s size and how many years the customer has lived in it whereas on the product’s side there is only information regarding the categories and the amount spent. It is my belief that Gender , Age , City\_Category , Product\_Category\_1 are the predictors that will influence more the amount spent by a customer on this day.

The target variable is purchase .

**Age :** should be treated as numerical. It presents age groups.

**City\_Category:**We can convert this to numerical as well, with dummy variables. Should take a look at the frequency of the values.

**Occupation**: It seems like it has at least 16 different values, should see frequency and try to decrease this value.

**Gender:** There are possibly two gender, we can make this binary.

**Product\_ID:** Should see if the string “P” means something and if there are other values.

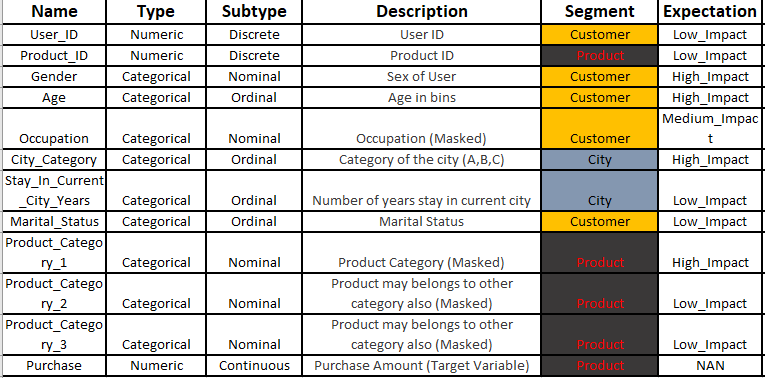
**Stay\_In\_Current\_City\_Years:**We should deal with the ‘+’ symbol.

**Product\_Category\_2**and **Product\_Category\_3**: Have NaN values.

New variables to have in consideration:

**User\_Count**: There are duplicate User\_ID , so it would be a good idea to create a feature with number of observations of the user

**Product\_Count**: Number of observations of the product



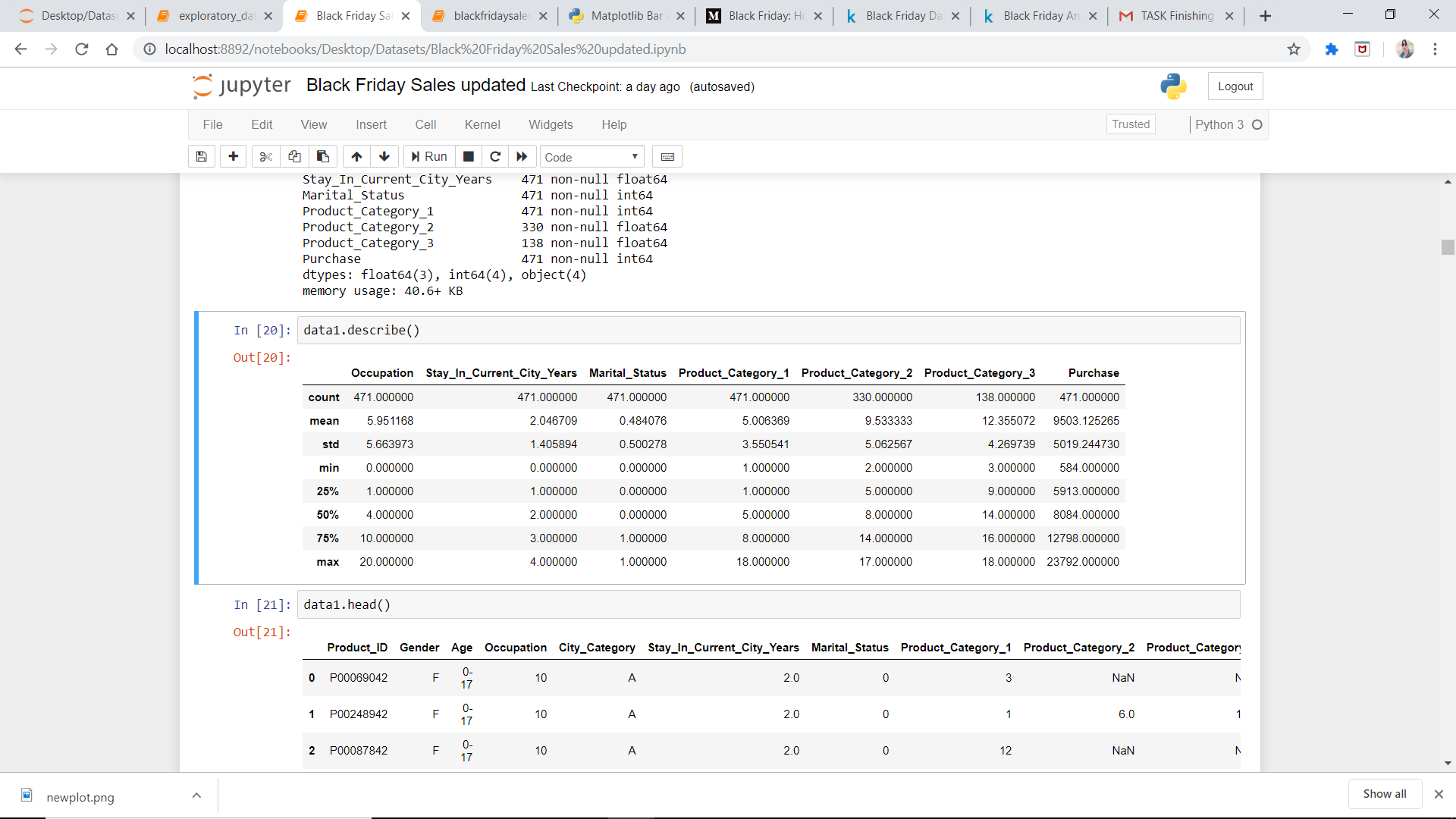
**5.Data Collection**

The data set that we got is from Kaggle.com

**6.Methodology**

**Exploratory Data Analysis:** It is an approach to analysing data sets to summarize their main characteristics, often with visual methods.

**Describe**



Min-It represents min value of each column.

Max-It represents max value of each column.

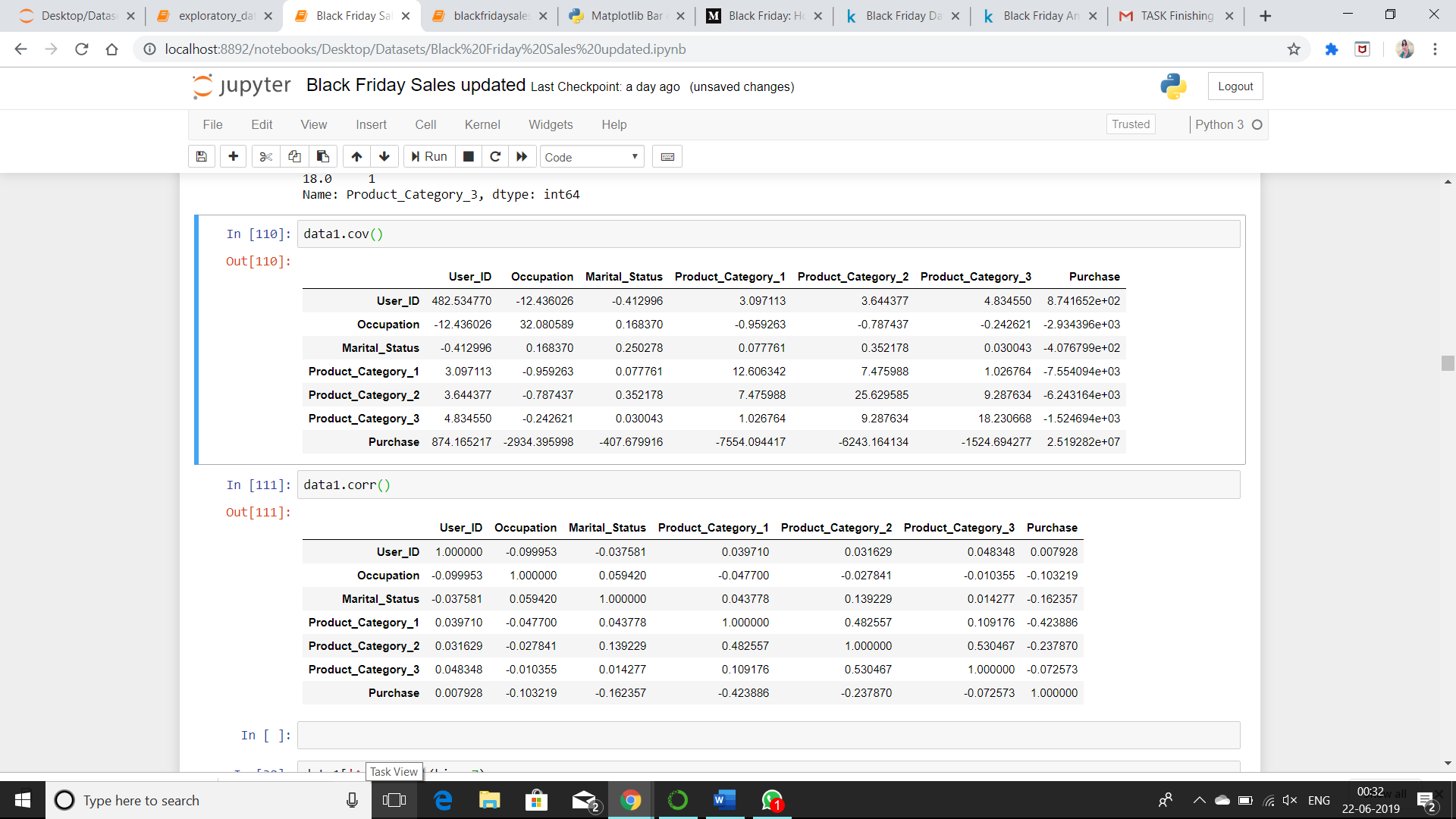
25%-It represents the 1st quartile.

50%-It represents 2nd quartile.

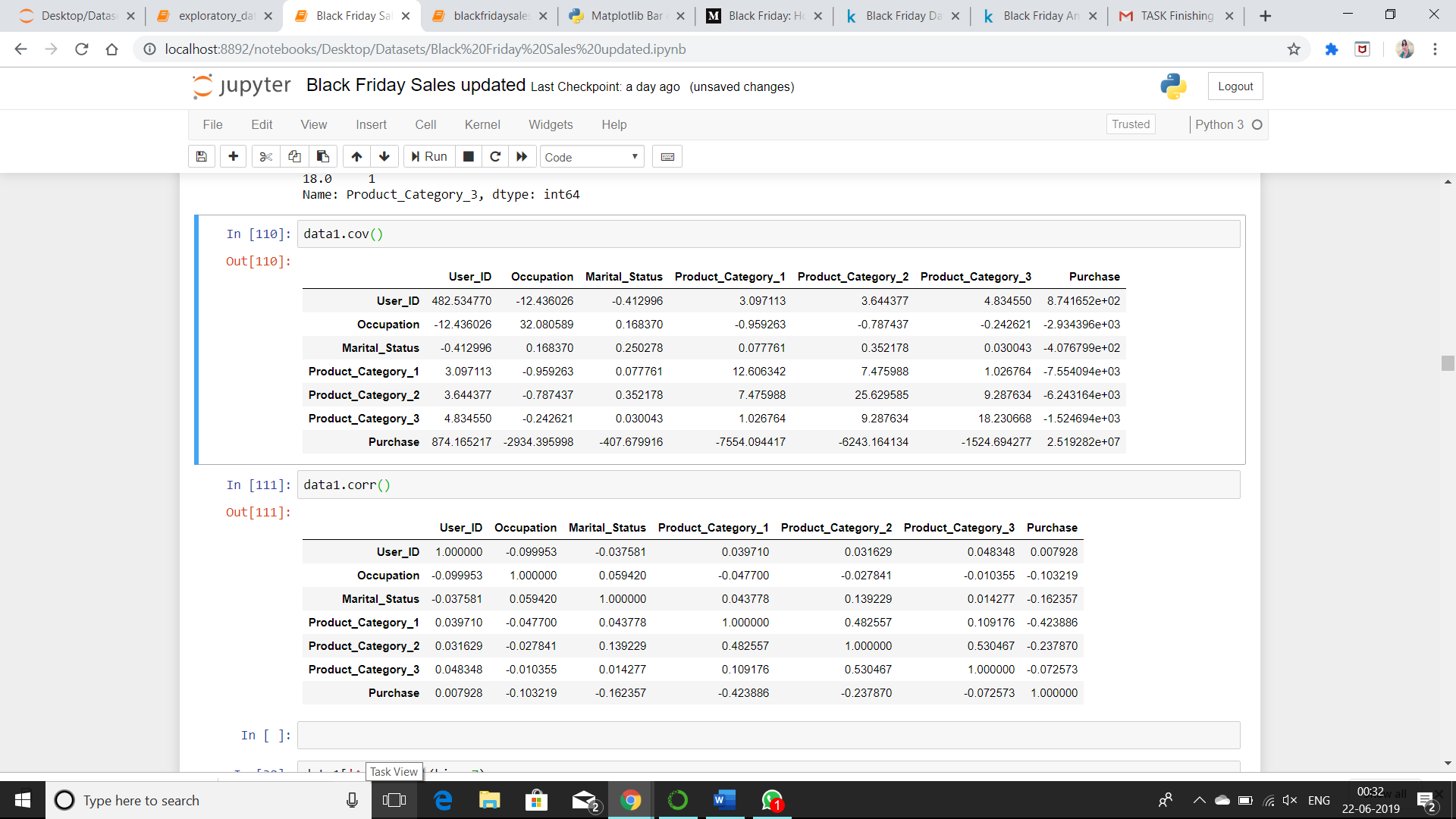
75%-It represents 3rd quartile.

Count-It represents number of rows in each column.

**Covariance** :-Covariance indicates how two variables are related. A positive covariance means the variables are positively related, while a negative covariance means the variables are inversely related.



**Correlation:-** Correlation is another way to determine how two variables are related. In addition to telling you whether variables are positively or inversely related, correlation also tells you the degree to which the variables tend to move together



0-No correlation

+1-strong positive correlation

-1-strong negative correlation

>0.5-good positive correlation

<0.5-weak positive correlation

>-0.5-week negative correlation

<0.5-good negative correlation

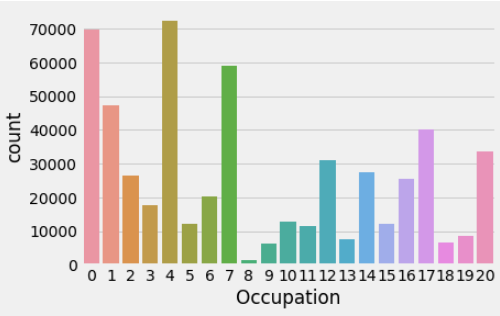
**Figures and Tables**

#### **1.Univariate Analysis:** Studyofonevariable

**Numerical Predictors:**

In this we are performing the distribution of single numeric variable and analyse through representation.

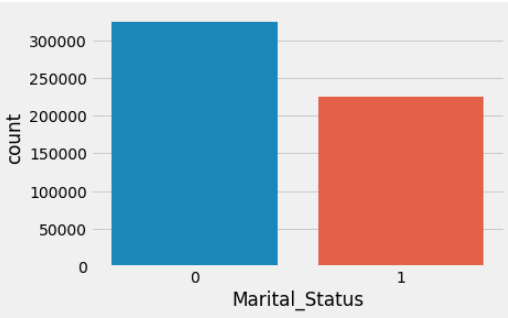
Occupation



From the above figure it is clear that people at occupation 4 are buying more products on black Friday than the people at other occupations.

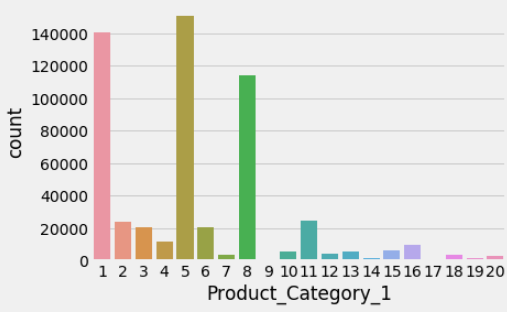
Marital Status

0-Single People 1-Married People



From the above figure it is clear that single people are buying more products on black Friday than the married people.

Product Category

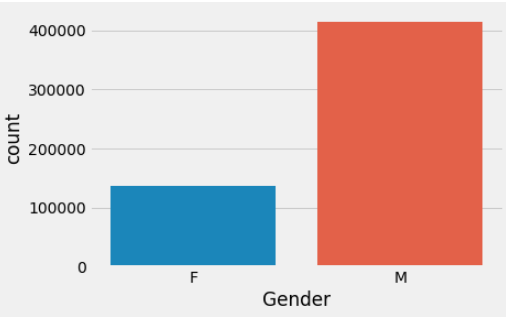


 From the above figure it is clear that three products stand out, number 1, 5 and 8 represents more number of buying products.

Categorical Predictors:

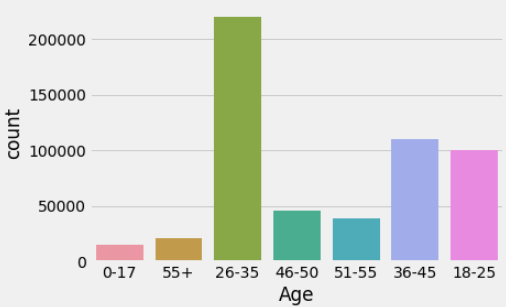
Non numeric variables. In this we can predict the values using bar charts, frequency

Gender



From the above figure it is clear that Male are buying more products on black Friday than the Female.

Age



From the above figure it is clear that the peoples age between 26-35 are buying more products and the people age between 0-17 are buying less products.

City Category

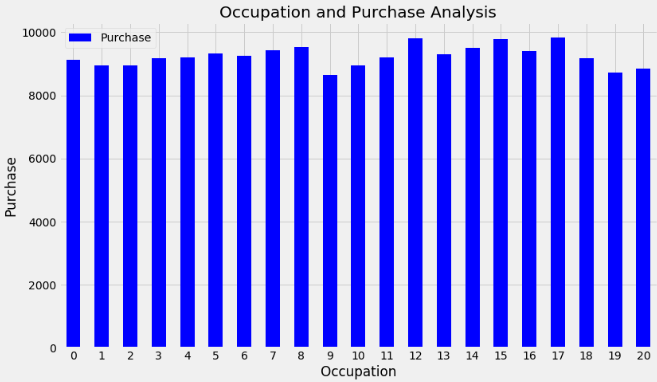


From the above figure it is clear that the peoples in city B are buying more products and than other two cities.

2. Bivariate Analysis

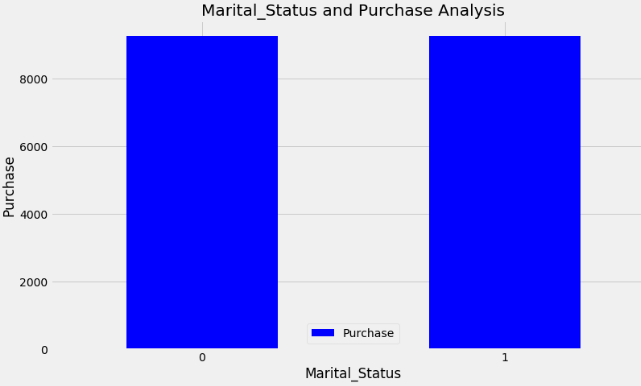
Using this analysis we can understand the relationship between our target variable and predictors as well as the relationship among predictors.

Occupation and Purchase Analysis



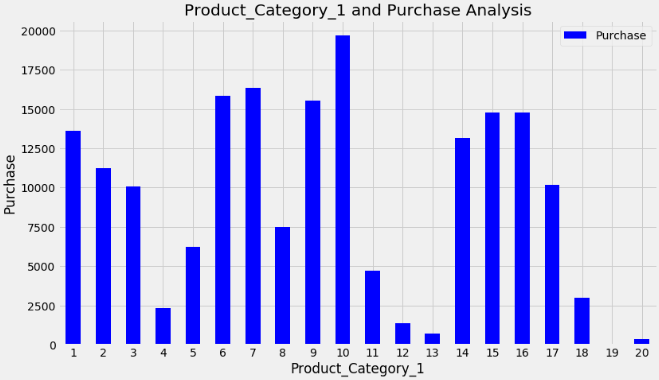
In the above figure the amount each user spends on average is more or less the same for all occupations.

Marital Status and Purchase Analysis



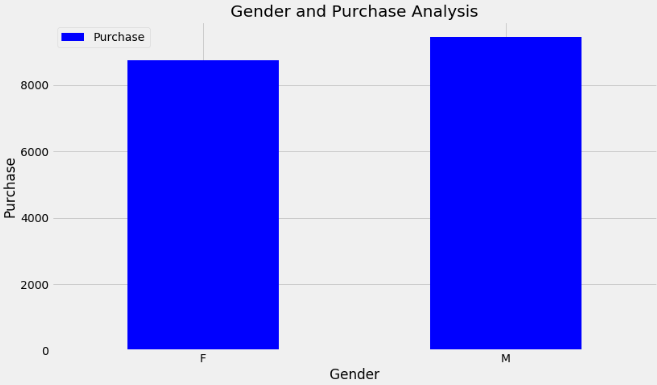
Previously we had more single customers buying more products than married. However, on average an individual customer tends to spend the same amount independently if his/her is married or not.

Product\_ Category1 and Purchase



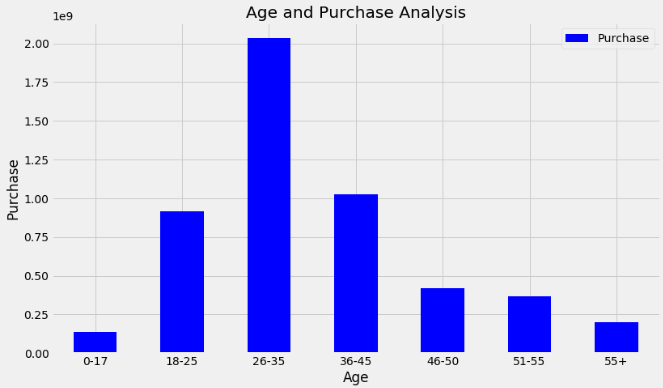
We can observe that the value spent on average for Product\_Category\_1 we can also see that though there were more products bought for categories 1,5,8 the average amount spent for those three is not the highest. We can observe that the other categories appearing with high purchase values despite having low impact on sales number

Gender and Purchase



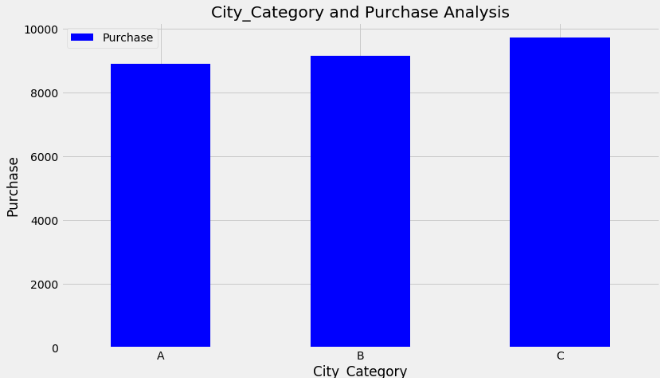
From the above figure the male gender spends more money on purchase than the female gender.

Age and Purchase



From the above figure we can observe that total amount spent in purchase is in accordance with the number of purchases made, distributed by age.

City Category and Purchase Analysis



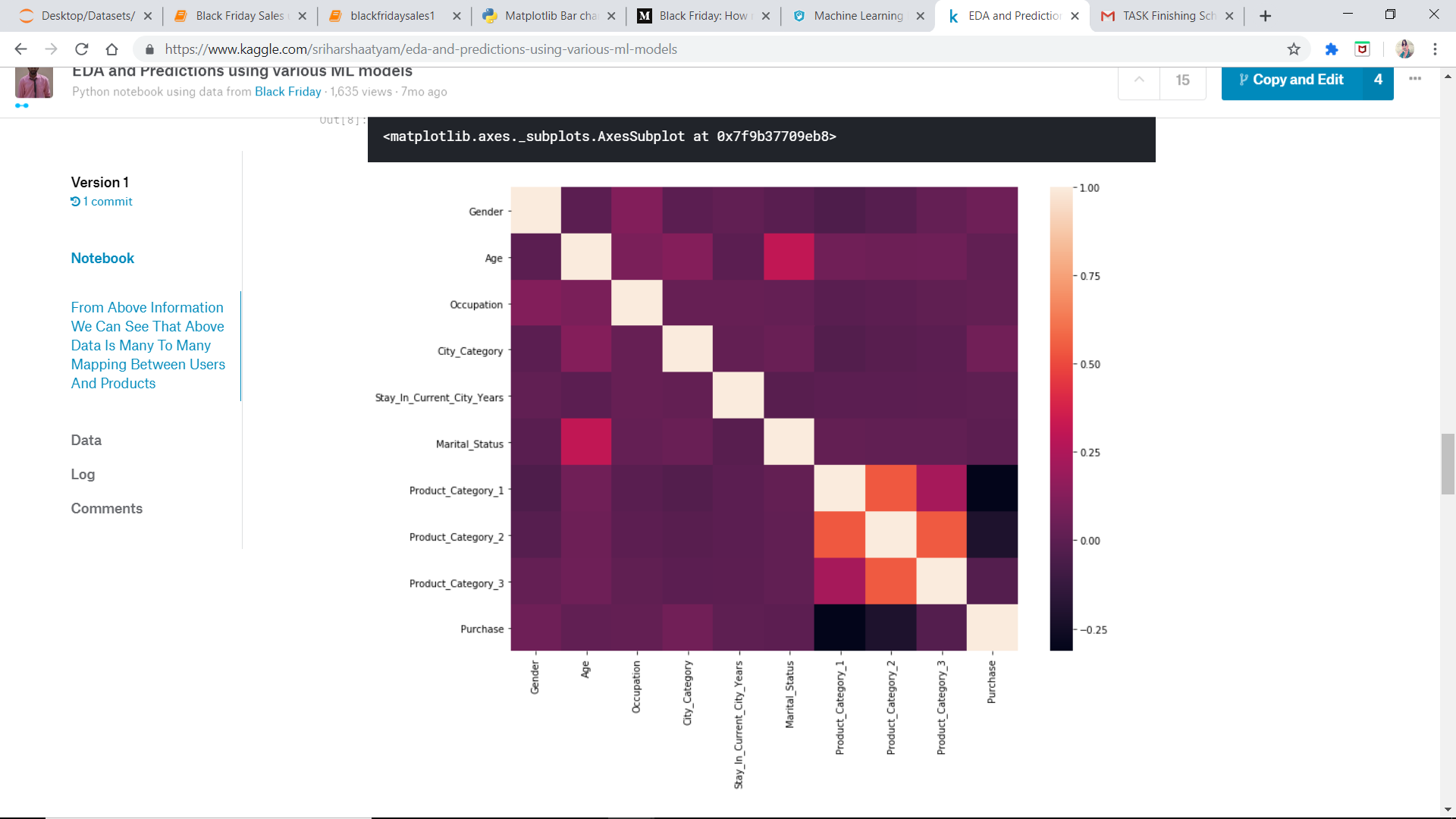
We can observe that city type ‘B’ had the highest number of purchases registered. However, the city whose buyers spend the most is city type ‘C’.

Pie plot

Total purchase by Age

A picture containing vector graphics

Description automatically generated

**Heat Map**

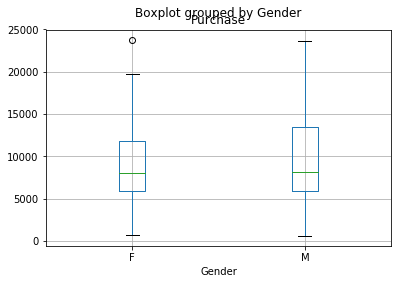
From the above heatmap we can see that the correlation between product category 1 and product category 2 is high compared to product category 1 and product category 3.

we can also observe that there is a negative correlation between purchase and product category 1 and product category 2 which might indicate the prices of product category 1 and product category 2 are high and few people purchase them.

we can also check that age and marital status have high correlation.we can also observe that no variable is highly correlated with the purchase variable which is our target .It shows that purchase variable depends on ensemble of all the variables.

Box Plot:

Gender and Purchase



Males are more active in purchasing than females and we can observe some outliers in female

Male Female

Lower whisker-500 Lower whisker-1000

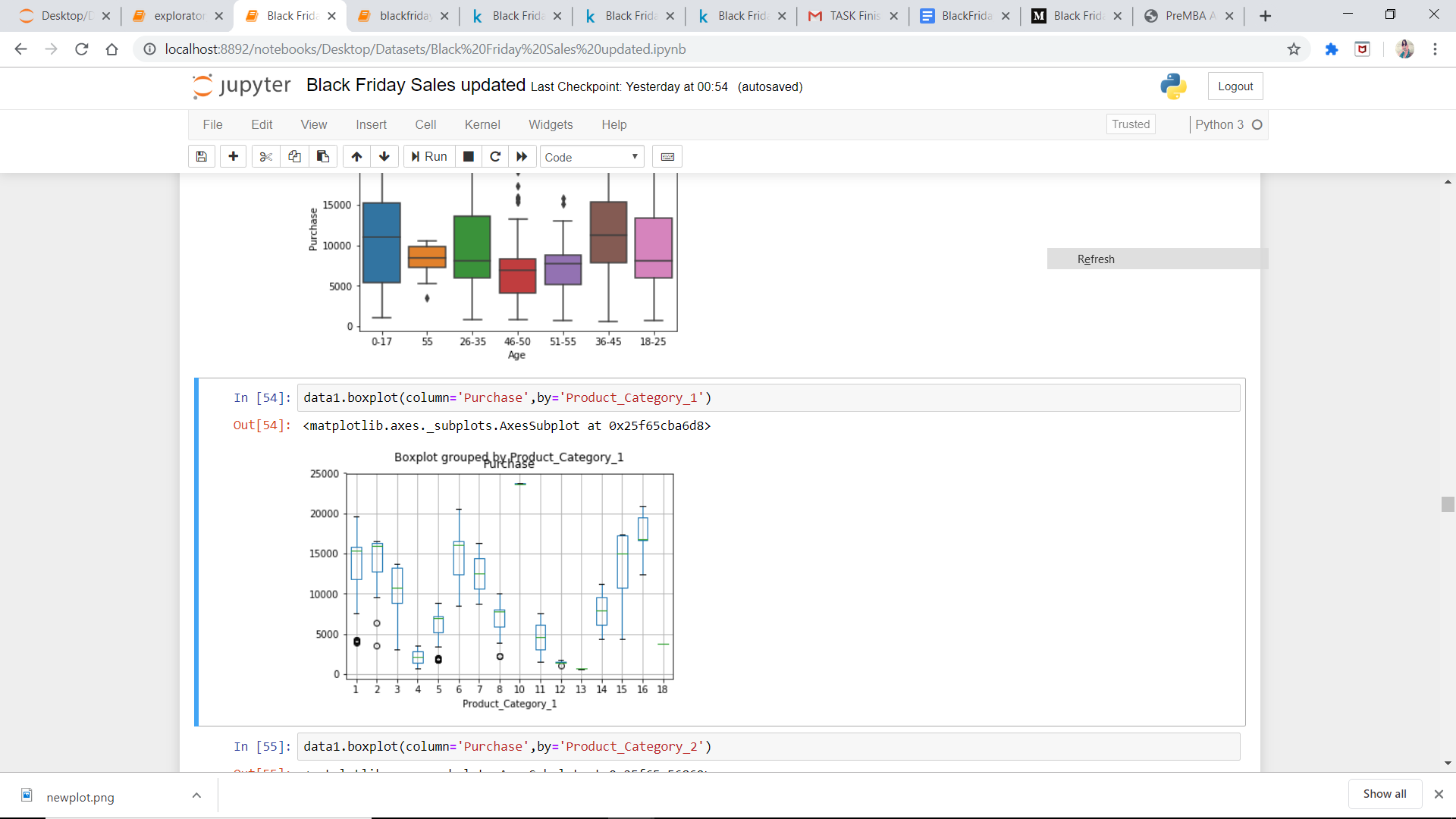
Lower Hinge-6000 Lower Hinge-6000

Median-8000 Median-8000

Upper Hinge-14000 Upper Hinge-12000

Upper whisker-24000 Upper whisker-20000

Product purchase1and purchase



There are many outliers in above box plot

**Data Modelling**

We can train a model by using

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. SVM
5. Naive Bayes
6. kNN

Since we have many algorithms to choose from lets calculate the accuracy scores first.

Accuracy Score of Linear regression on test set 7.861742005464478

Accuracy Score of Decision Tree on test set 50.93775785935173

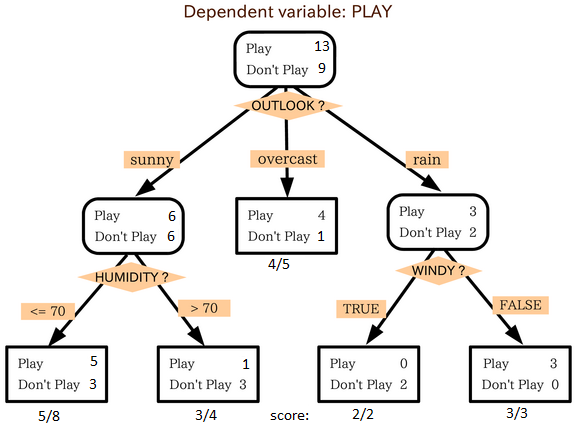
Accuracy Score of Random Forests on test set 63.81975515943545

These are the three highest accuracy scores that we are getting for the data taken so let us first create 2 models one for Decision Tree and other for the Random Forest

**Decision Tree**

It is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.

Example:



In the image above, you can see that population is classified into four different groups based on multiple attributes to identify ‘if they will play or not’. To split the population into different heterogeneous groups, it uses various techniques like Gini, Information Gain, Chi-square, entropy.

After training the model with 70% of the data

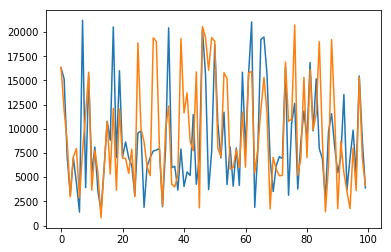
These are test set data values

[6931, 9564, 8027, 15172, 12014, 7824, 9846, 10007, 4028, 19525, 8159, 11788, 12015, 5183, 15859, 5897, 16662, 1747, 3702, 16500, 15361, 1747, 15859, 7190, 15705, 3993, 7089, 12850, 12409, 3664, 5329, 7796, 1848, 3624, 6931, 11463, 9564, 7801, 3667, 5900, 6855, 5839, 3993, 16866, 7796, 16306, 11562, 6122, 8027, 12098, 3897, 15524, 9564, 18974, 16306, 3594, 8770, 15957, 6927, 7959, 9894, 19525, 1828, 8804, 7766, 15859, 10766, 1784, 15549, 2010, 16306, 7796, 10058, 12909, 3449, 15361, 13551, 11404, 1828, 6407, 3055, 13253, 5329, 7802, 3897, 5915, 8804, 1447, 7909, 1629, 15246, 2010, 8599, 15581, 5349, 6073, 5897, 1539, 7802, 7801, 7772]

These are predicted data values

[ 8637 4290 6171 15872 7159 6078 5996 5875 1414 11023 15244 15333 16137 8702 19251 6109 15365 3580 8696 15903 15715 3696 11937 5252 11755 8007 5152 23603 5280 8621 6996 10021 7176 7061 8685 15770 3581 7969 8714 9695 16420 9788 5899 20442 7820 20474 15426 7786 7973 15517 7854 11518 15227 15212 15943 5407 8858 5234 8851 16579 9938 7746 7159 6852 7965 19416 11605 12131 11853 3482 20529 9856 758 16352 5168 11896 13400 11430 4325 9734 3978 15371 8701 15900 17391 6025 8598 1446 5973 1752 4121 6975 12422 19215 13337 3957 3990 4664 15319 4570 4124]

Since we cannot conclude the accuracy between them let us draw a graph



This graph shows the similarities between the test and predicted values.

**Random Forest**

Random Forest is a trademark term for an ensemble of decision trees. In Random Forest, we’ve collection of decision trees (so known as “Forest”). To classify a new object based on attributes, each tree gives a classification and we say the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Each tree is planted & grown as follows:

1. If the number of cases in the training set is N, then sample of N cases is taken at random but *with replacement*. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

After training the model with 70% of the data

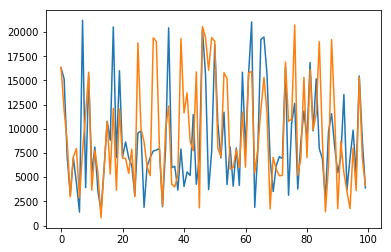
These are test set data values

[ 7824, 9846, 10007, 4028, 19525, 8159, 11788, 12015, 5183, 15859, 5897, 16662, 1747, 3702, 16500, 15361, 1747, 15859, 7190, 15705, 3993, 7089, 12850, 12409, 3664, 5329, 7796, 1848, 3624, 6931, 11463, 9564, 7801, 3667, 5900, 6855, 5839, 3993, 16866, 7796, 16306, 11562, 6122, 8027, 12098, 3897, 15524, 9564, 18974, 16306, 3594, 8770, 15957, 6927, 7959, 9894, 19525, 1828, 8804, 7766, 15859, 10766, 1784, 15549, 2010, 16306, 7796, 10058, 12909, 3449, 15361, 13551, 11404, 1828, 6407, 3055, 13253, 5329, 7802, 3897, 5915, 8804, 1447, 7909, 1629, 15246, 2010, 8599, 15581, 5349, 6073, 5897, 1539, 7802, 7801, 7772]

These are predicted data values

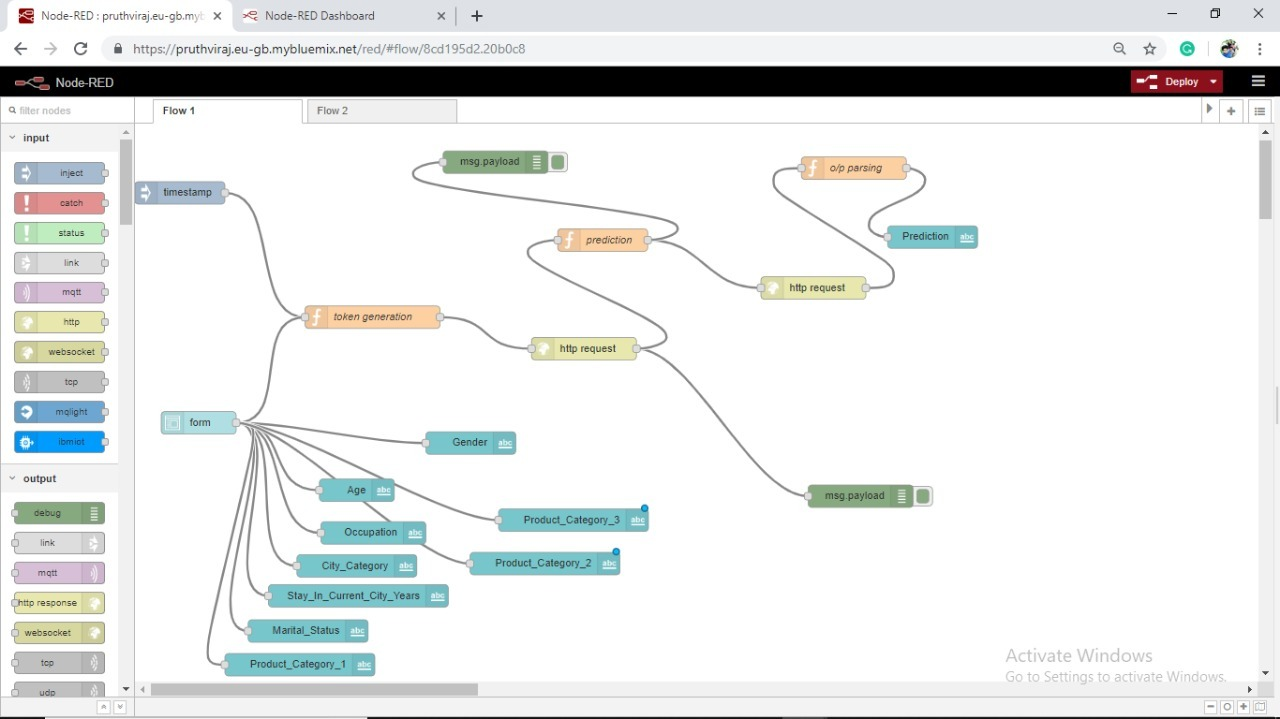
[ 6078 5996 5875 1414 11023 15244 15333 16137 8702 19251 6109 15365 3580 8696 15903 15715 3696 11937 5252 11755 8007 5152 23603 5280 8621 6996 10021 7176 7061 8685 15770 3581 7969 8714 9695 16420 9788 5899 20442 7820 20474 15426 7786 7973 15517 7854 11518 15227 15212 15943 5407 8858 5234 8851 16579 9938 7746 7159 6852 7965 19416 11605 12131 11853 3482 20529 9856 758 16352 5168 11896 13400 11430 4325 9734 3978 15371 8701 15900 17391 6025 8598 1446 5973 1752 4121 6975 12422 19215 13337 3957 3990 4664 15319 4570 4124]

Since we cannot conclude the accuracy between them let us draw a graph

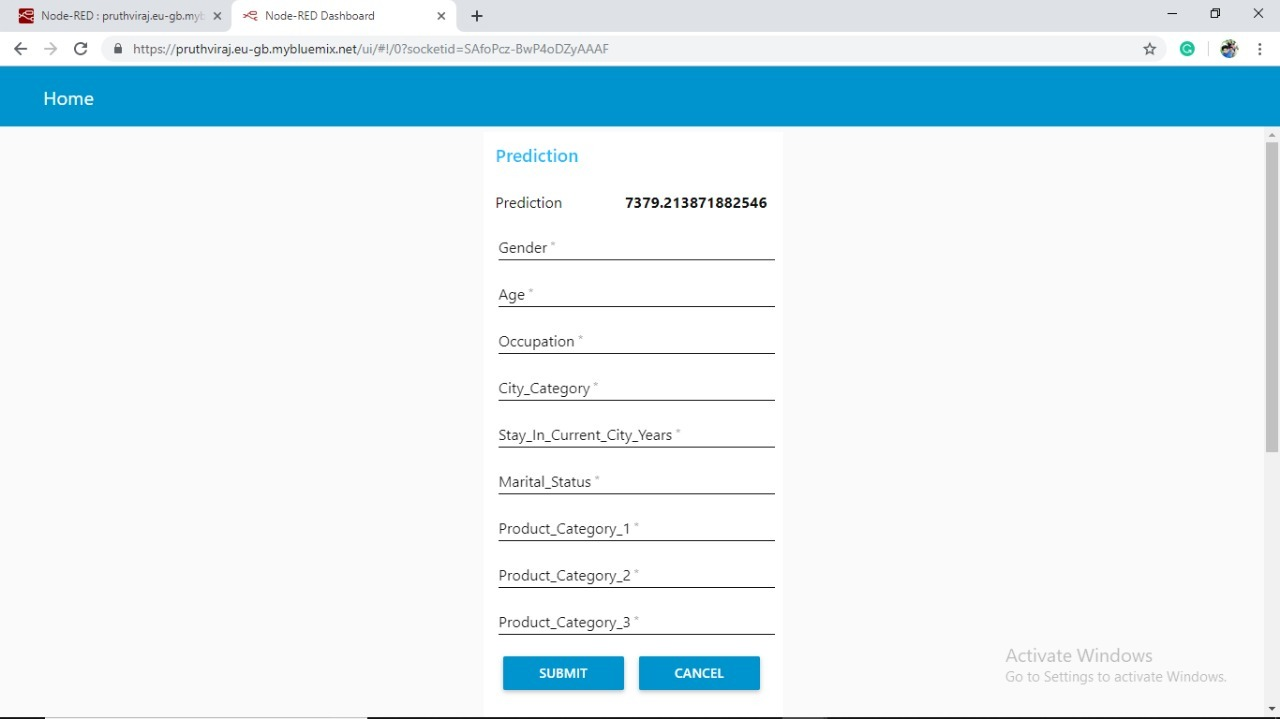
This graph shows the similarities between the test and predicted values

Using the Node red in IBM Watson we can get the below flow this flow takes the input values and give prediction values .by this json file we get a UI in which we can give inputs and get prediction.

Node Red



**UI**



**7. Findings and Suggestions**

From our exploratory data analysis we have observed the following things.

* more single people buying products on Black Friday than married people but we had more single customers than married. However, on average an individual customer tends to spend the same amount independently if his/her is married or not. Again, if you had all the purchases the single group, since has a higher representation, will have the highest purchase values.
* >From the distribution for products from category one, it is clear that three products stand out, number 1, 5 and 8.
* >Male bought up 3 times more purchases than Females,but they both ended up producing similar (nearly equal) sales.
* >most purchases are made by people between 18 to 45 years old.
* >cities ‘B’ had higher sales than the others instead of having lesser number of people than city ‘A’ but we saw previously that city type ‘B’ had the highest number of purchases registered. However, the city whose buyers spend the most is city type ‘C’.
* >The tendency looks like the longest someone is living in that city the less prone they are to buy new things. Hence, if someone is new in town and needs a great number of new things for their house that they’ll take advantage of the low prices in Black Friday to purchase all the things needed.
* >Although there are some occupations which have higher representations, it seems that the amount each user spends on average is more or less the same for all occupations. Of course, in the end, occupations with the highest representations will have the highest amounts of purchases.

Suggestions

* >It is suggested to decrease the discount at places where there are more buyers and increase it at cities having less people because people are going to buy the product whatever may be the discount.this ends up giving more profits to the buyers.
* >It is also suggested to change the discounts depending on the nature of the place such as average people age,occupation,how many members are living in the city since only one year,

**8.Conclusion**

The ML algorithm that perform the best was Random Forest Model with RMSE = 4293 which got me in the first 63.5%. The next step will be looking at Ensembling.