University of Bridgeport

Spring 2019



Project Phase #3

Black Friday Sales

Prof. Jeongkyu Lee

Data Mining – CPSC552

Submitted by:

Name: Hindhuja Dhanasekharan

Name: Jagrithi Palachanda Chittiappa

Name: Shreya Satish Nikkam

Problem Definition:

Finding the targeted customer

* A specific group of consumers at which a company aims its products and services. Your target customers are those who are most likely to buy from you
* This will influence the marketing strategies and profit by concentrating on advertising their products to their targeted customer.

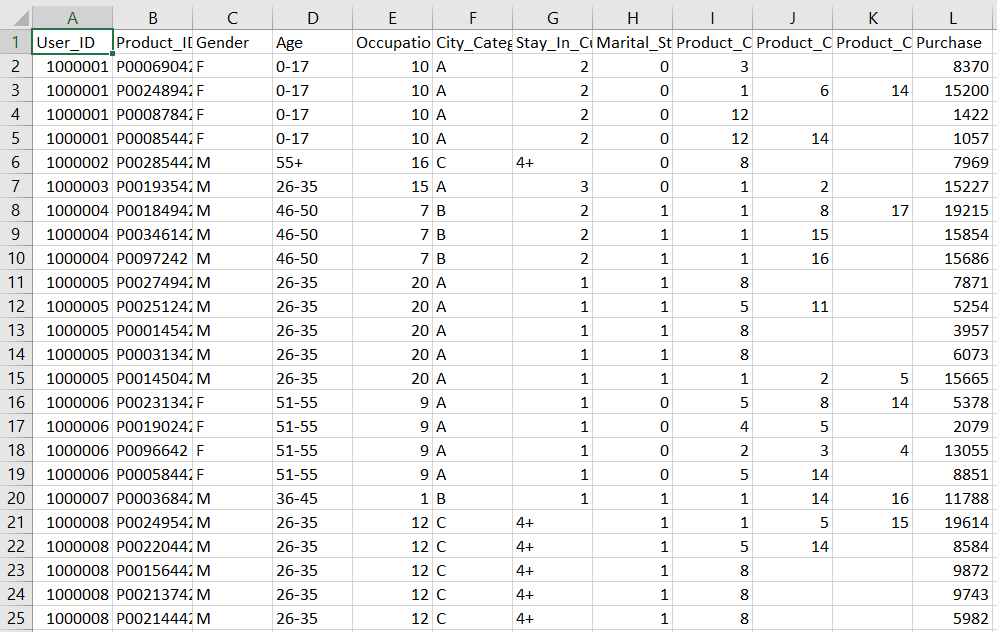
Characteristics:

The characteristics of the dataset are as follows:

* Dataset of 550 000 observations about the black Friday in a retail store.
* It has 550k rows x 12 columns which is available in the .csv format
* It contains different kinds of variables either numerical or categorical.
* It contains missing values

Dataset:

The snapshot below shows the sample of the Black Friday dataset.



LINK TO DATASET:

<https://drive.google.com/open?id=1IPcOuabwyZh_Tv0mtz-C6EVioEF1J3B8>

Algorithm and Evaluation:

EDA

The primary purpose of EDA is to examine a dataset without making any assumptions about what it might contain which maximizes the insight of the data patterns to spot anomalies and test hypothesis from graphical representations

Random Forest Algorithm

This is a regression problem and we'll use random forest. We'll choose number of trees (n\_estimators) in our forest and max\_depth for each tree by calculating scores for each combination and choosing the best one. Scoring metric, we'll use is RMSE.

RMSE

The RMSD represents the sample standard deviation of the [differences between predicted values and observed values](https://gerardnico.com/data_mining/residual).

K-Means

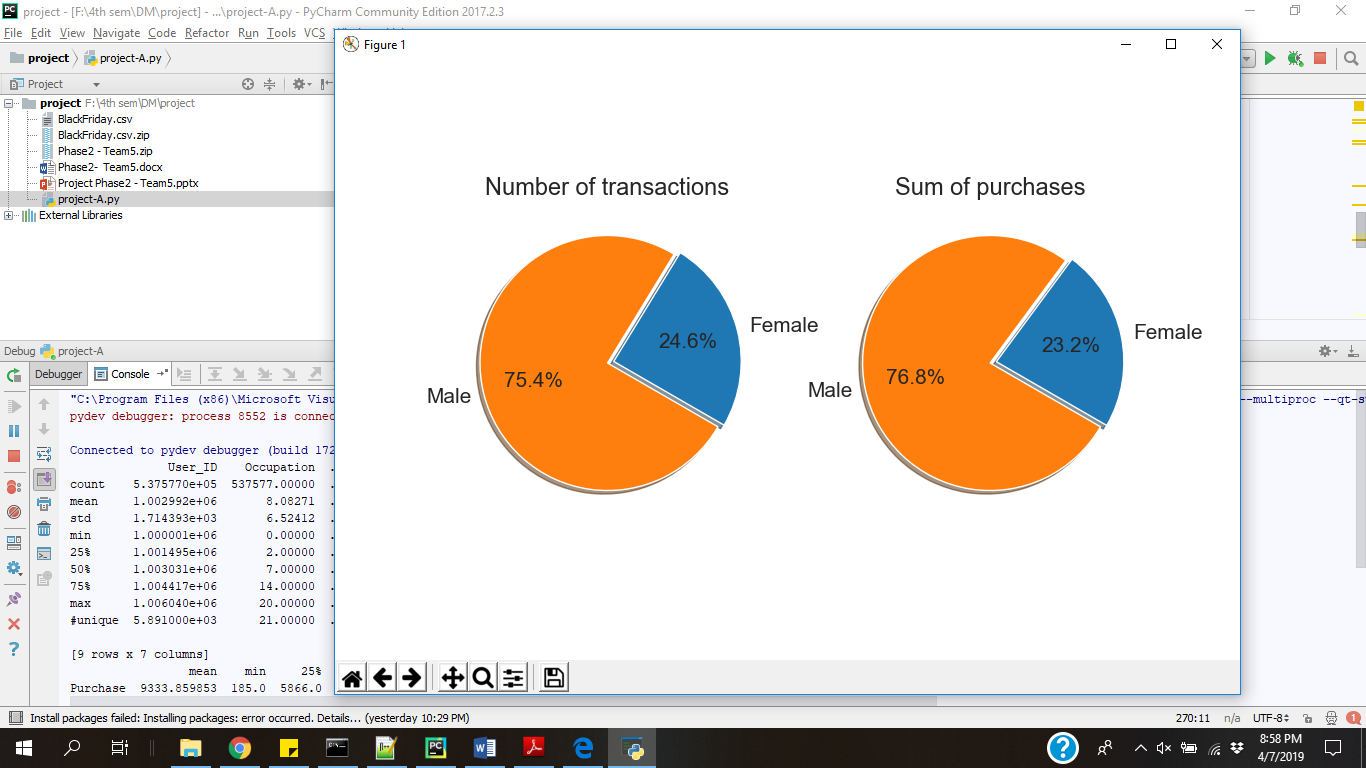
We group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids. After the number of clusters for predictive analytics we set K-means the algorithm to populate the clusters.

Implementation

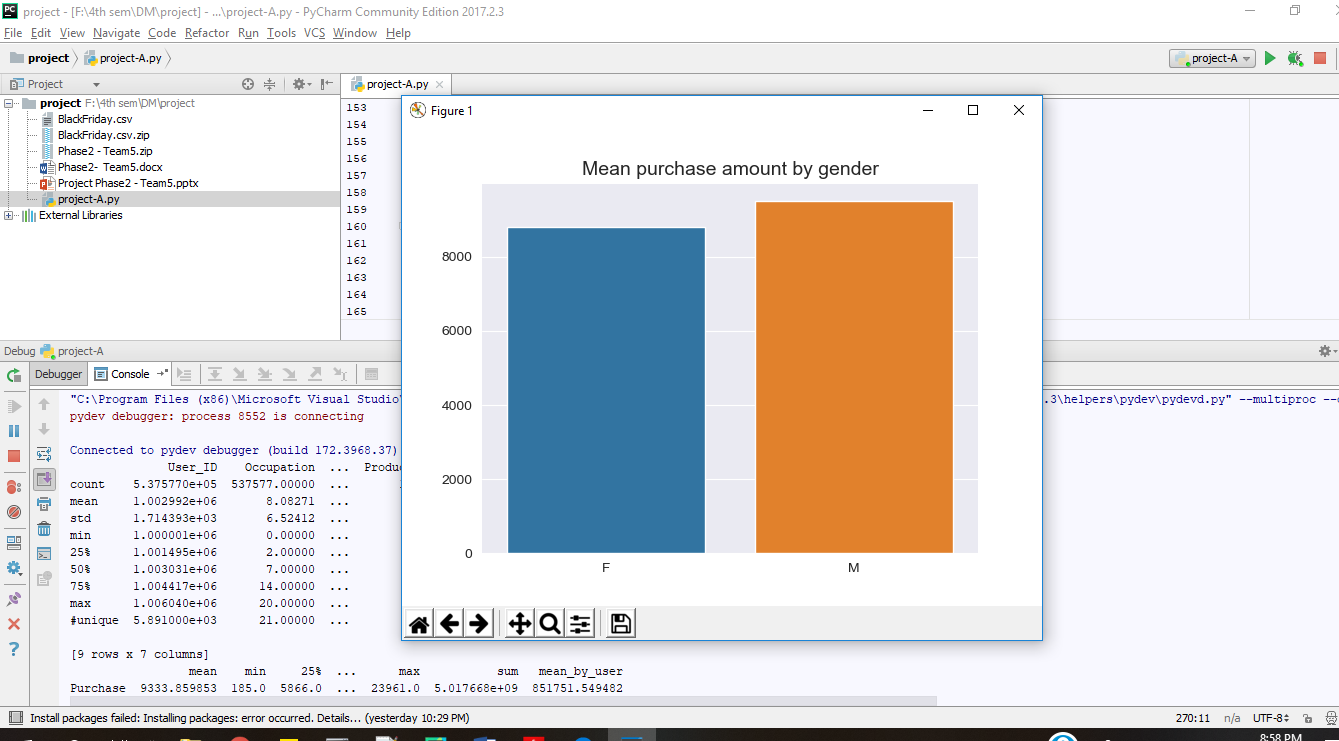
We imported the seaborn, IPython and jinja2 package to run the code. The snapshots below shows the implementation of the data mining techniques used in our project:

1. EDA

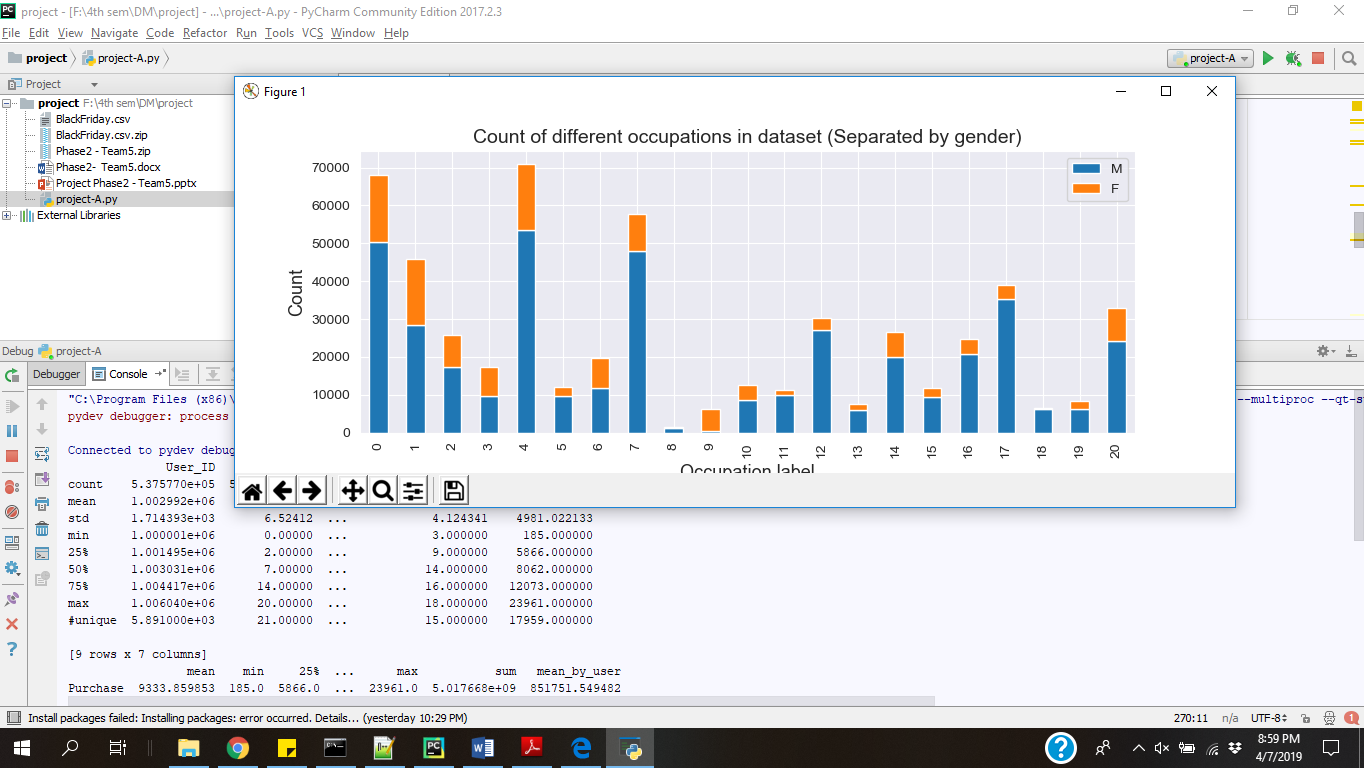
The analysis in the below graphical representation shows the amount of purchases made by men and women, category of occupations and age groups during the Black Friday Sales.



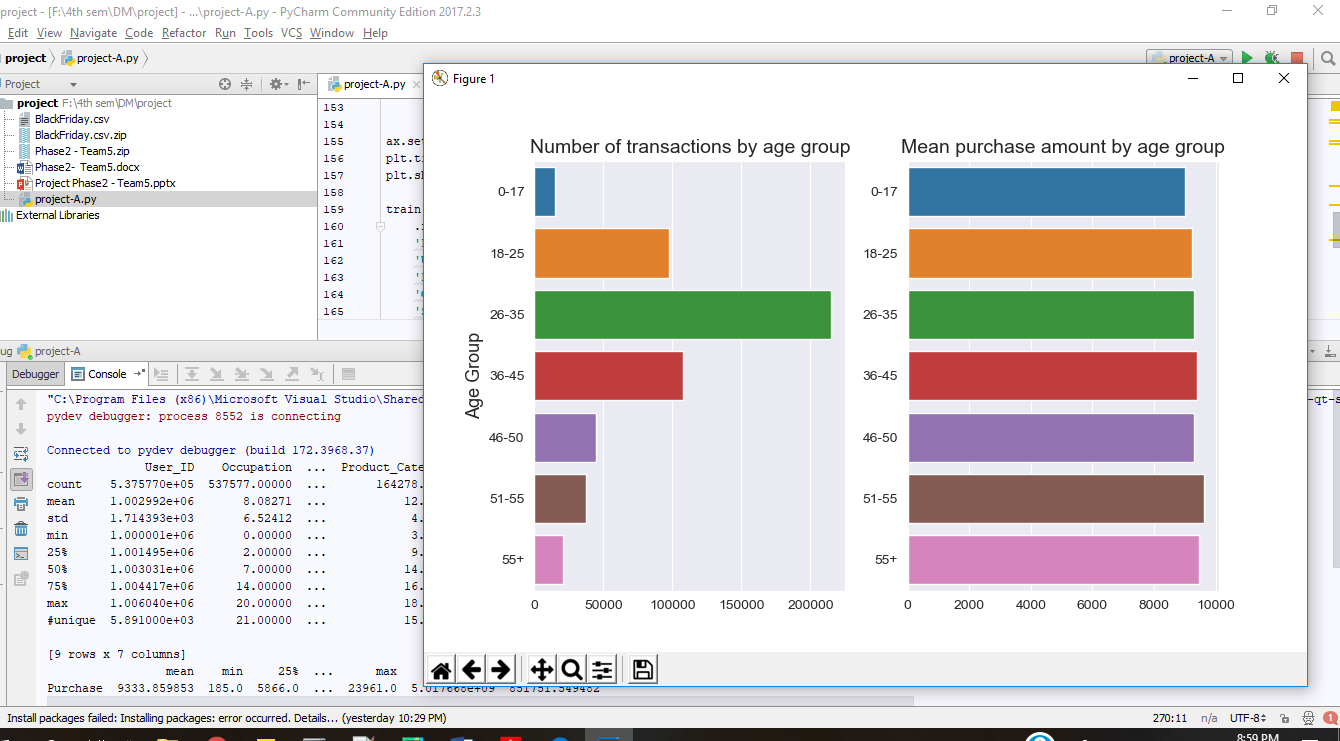
We can derive, that the amount of purchases made by males is much greater than that of the females.



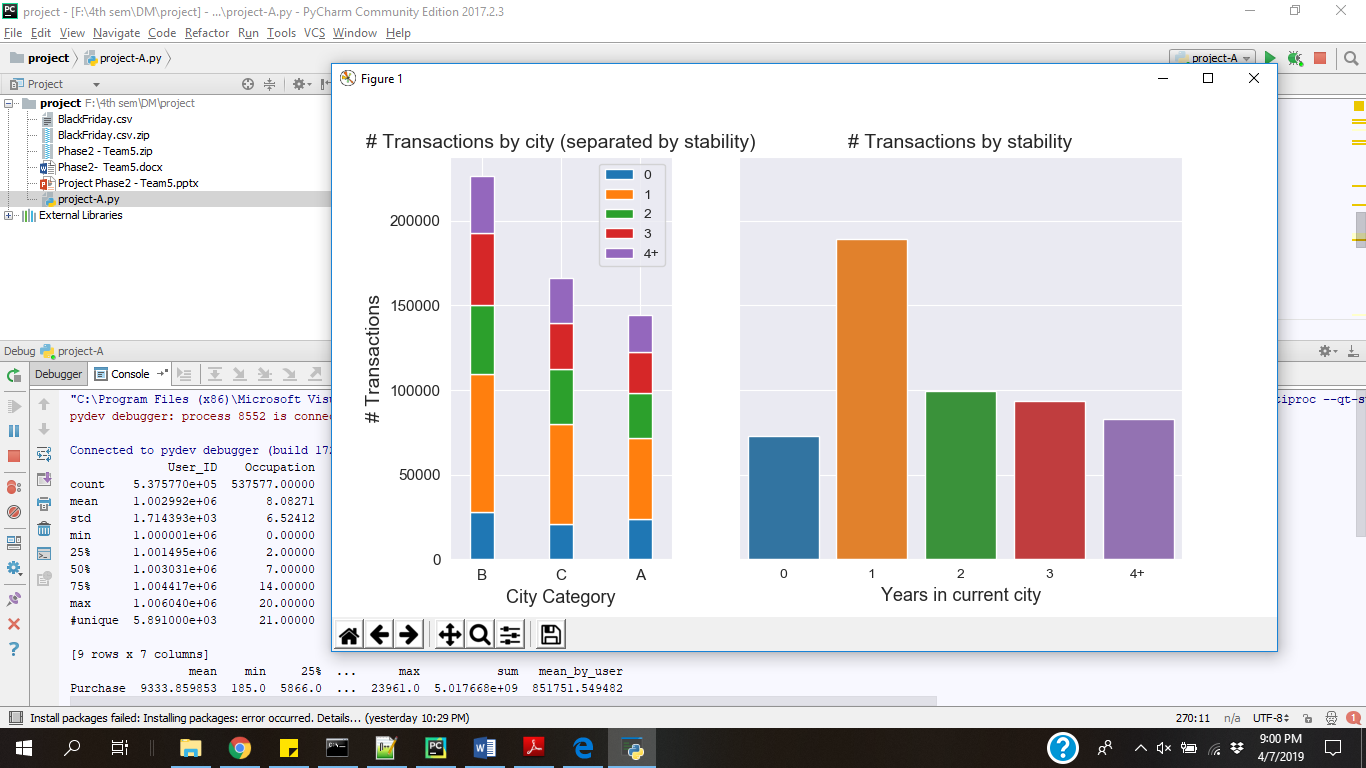
We can observe in the below graphical representation that the Occupation category 0,4,7 makes the highest number of purchases which can be focused more at the time of sales for better sales.



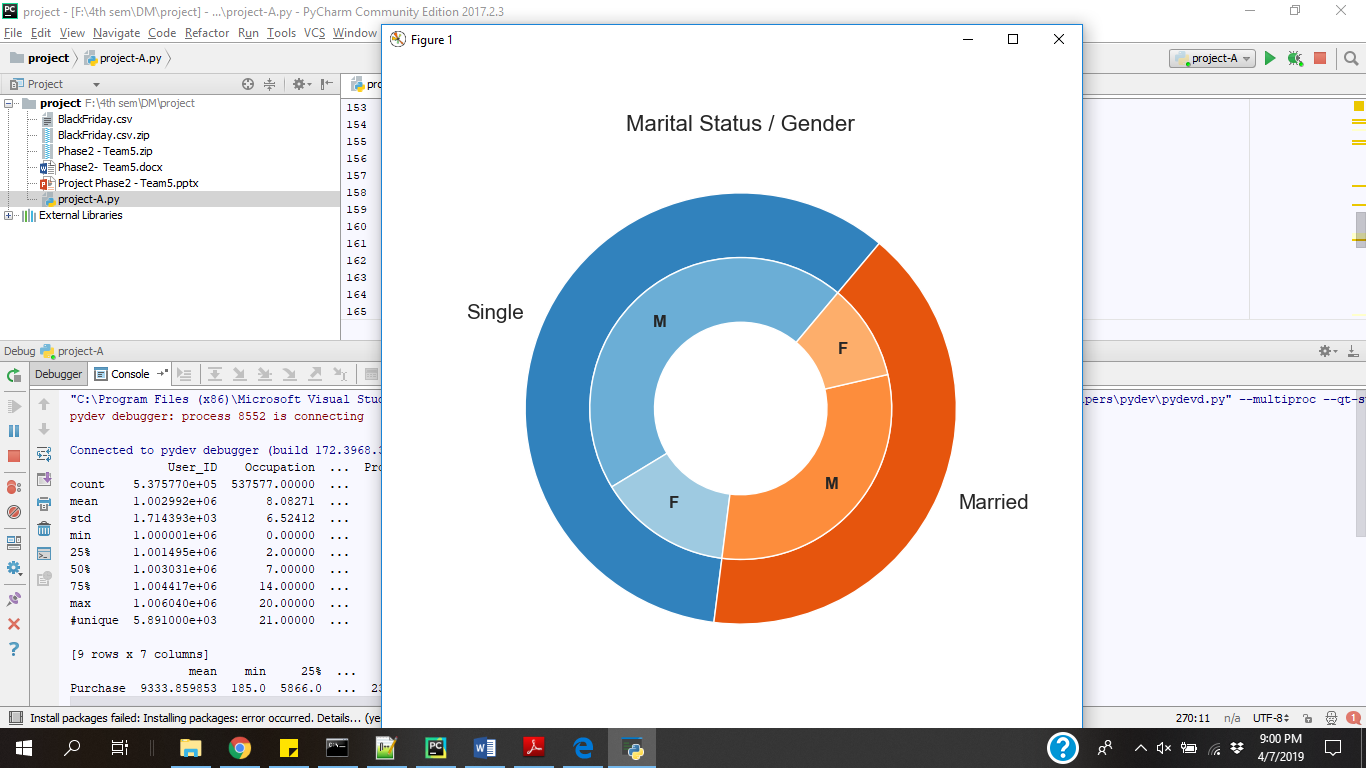
We see that the customers Age group ranging 26-35 makes the greatest number of purchases.



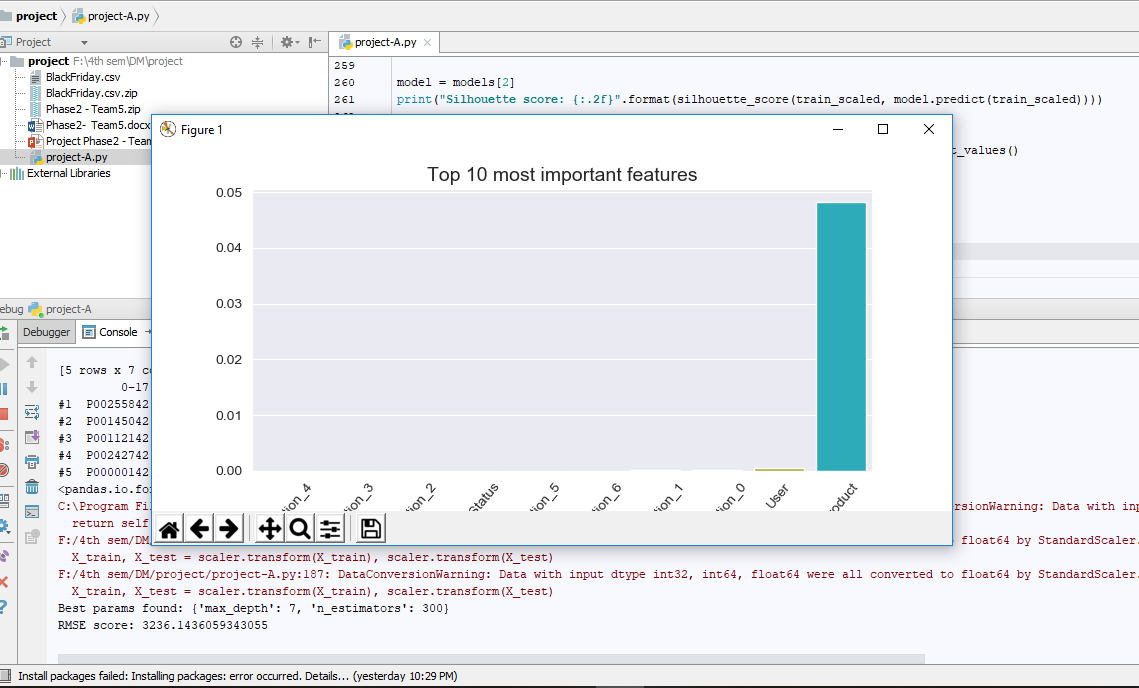
We can observe that City B makes more purchases when compared to the others and also customers staying in the city for 1 year also increases the sales made during Black Friday.



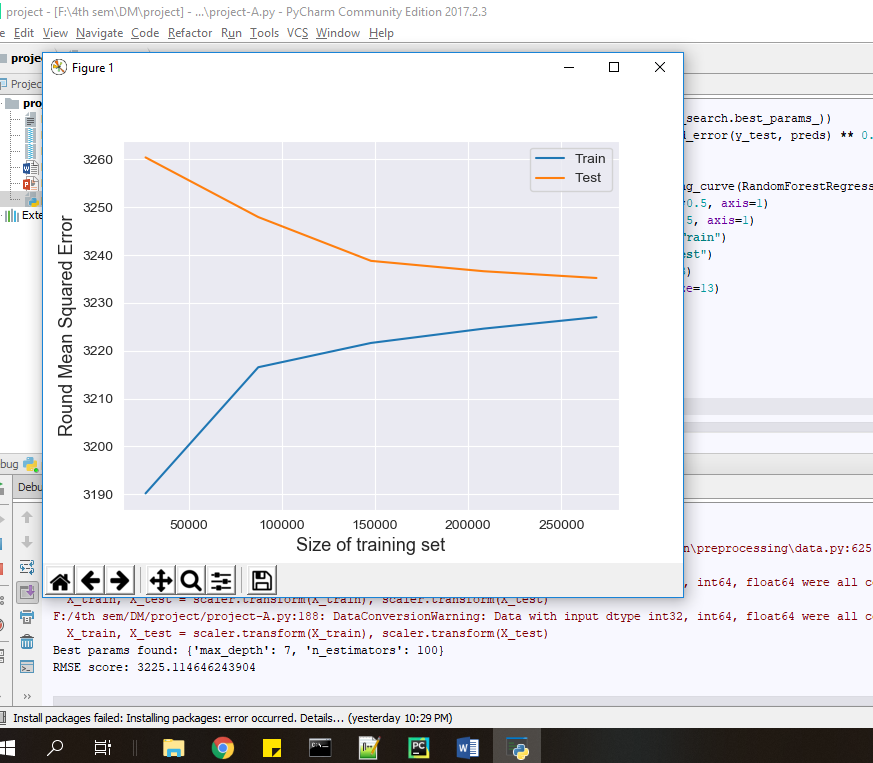
Here, we analyze that the customers who are single make more purchase than the married ones.



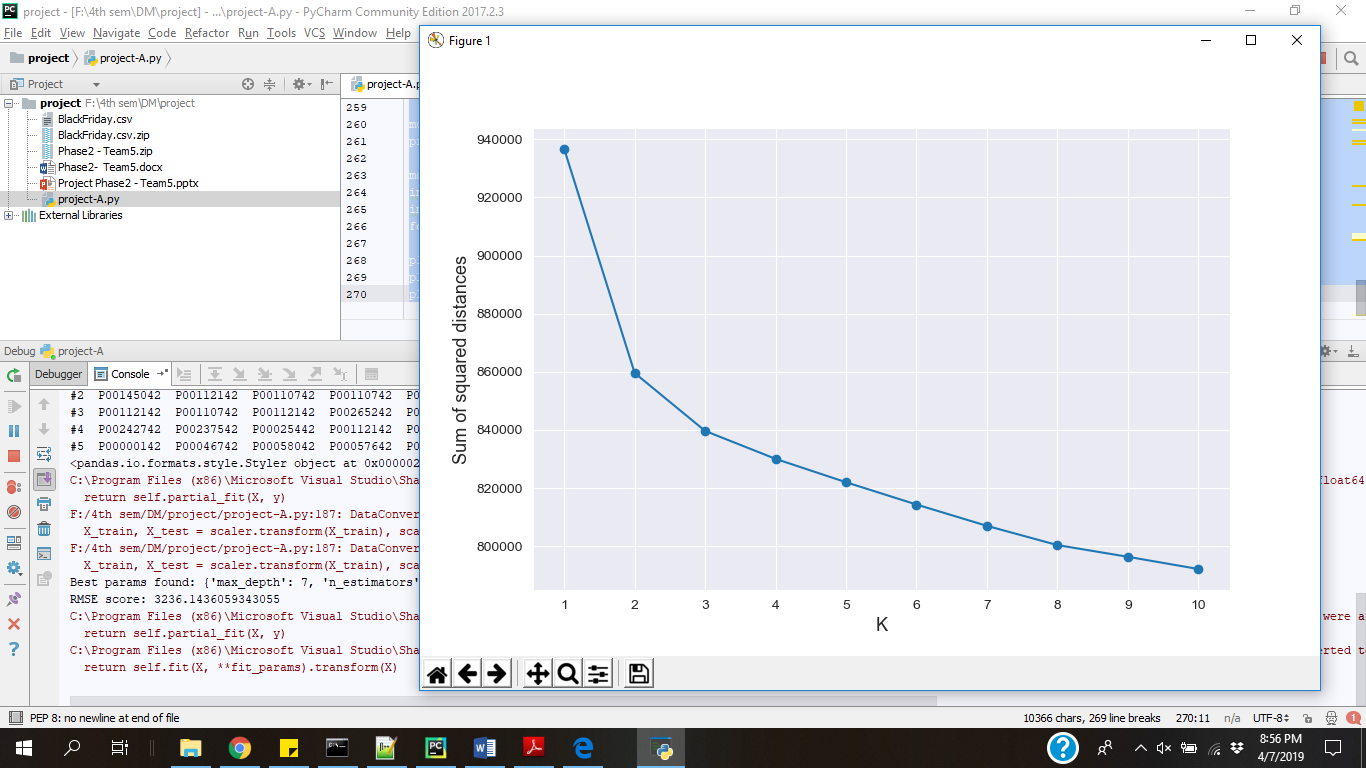
1. Random Forest Algorithm & RMSE



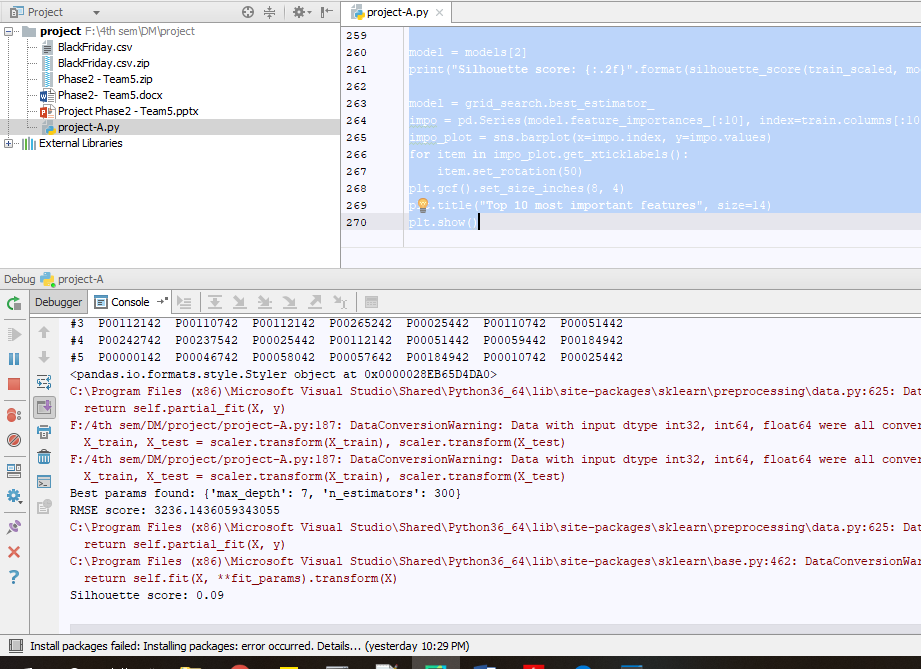
In the above snapshot, after running the Random forest regression we can observe that Product plays an important role in influencing the Black Friday sales. We see the evaluation score of the RMSE(Root Mean Square Error) to be 3236.14.



1. K-Means



After performing, K-means cluster customers based on products they have purchased. We cluster the users according to the transactions and their ID’s.



Conclusion:

In this project we performed the exploratory data analysis and Random Forest Regressor to predict purchase amount based on user id, product id and other features available in dataset. Our model had RMSE around 3236.14 mean of the target value based on Product\_ids.  
The customers are clustered using K-Means algorithm with 3 clusters and a silhouette score of 0.09.