3. Identifying Best Model -

By running the code with all features it was determined that the best parameter is 20, which give us a decision to built a Random Forest of 20.

For that porpoise we changed “n\_estimate” variable to equal 20, and run RandomForest with that parameter.

That give us significant features. TO build a confusion matrix, we need to pick the most significant feature, the most top features.

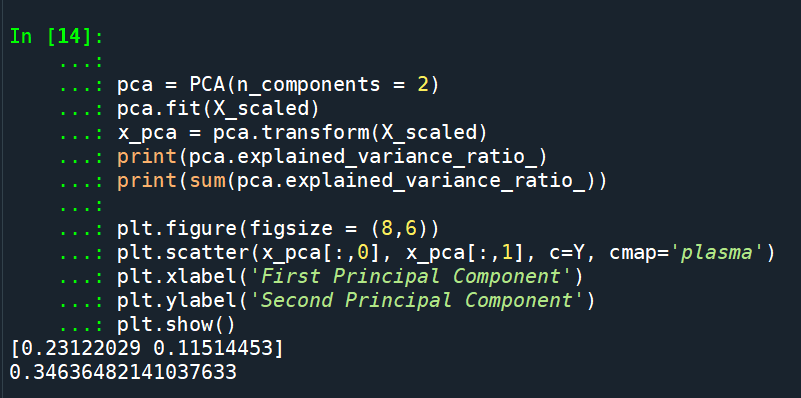
As we decided for approach of “minimising false positive”, when we run the analysis with all of the features, number of false positives was 268.

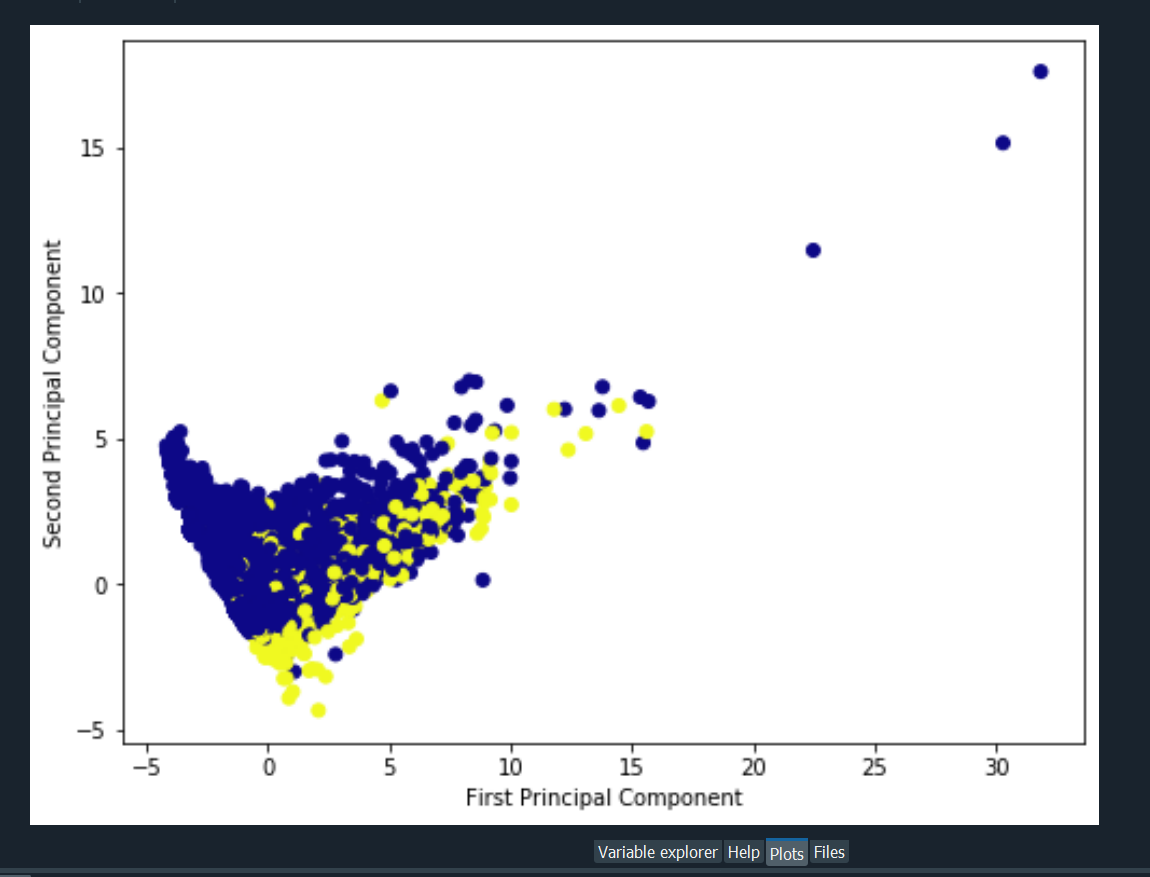
We have chosen top 5 variables out of 23. But that didn’t give good results. Number of false positives didn’t decrease, but it went up to 300. In the end we needed to add more significant features. When we added top 15 features and got false positives down to 278.

4. Generating Recommendations -

Part 2. PCA and K-means visualisation

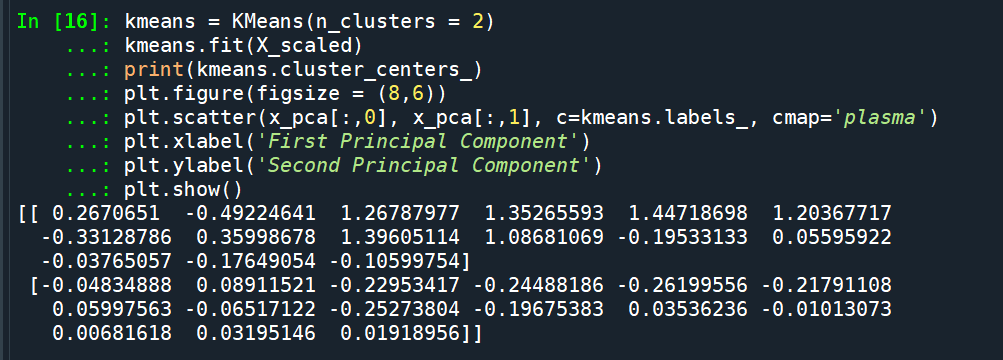
After running PCA – We notice that first principle component is showing that amount of variance explained by first princile component is 23% and amount of variance explained by second principle componen is 11.5%, whcih together is 35% explained variance in the data. . It is not enough, as we expect to get 70% . That means we cannot trust this visualisation. PCA has failed.

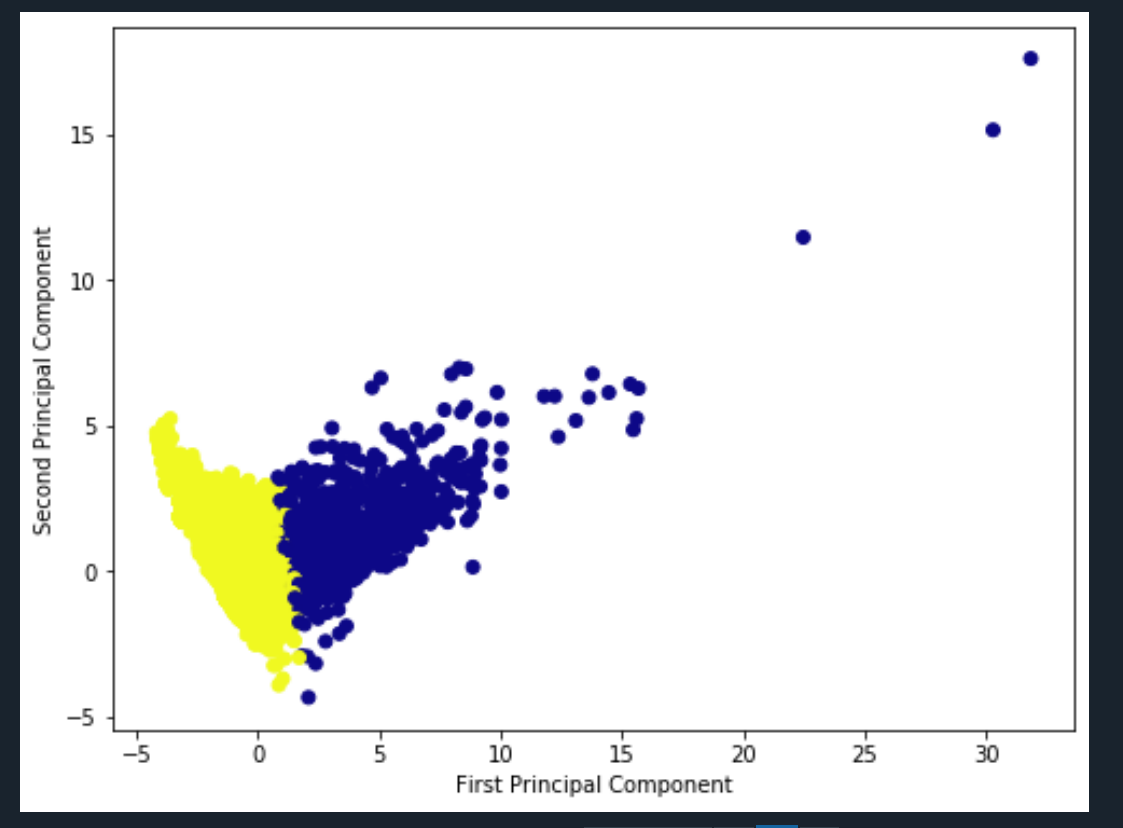




K-mean -

After running K-mean algorithm, we can see there are 2 clusters whcih we expected. We have 2 clusters, transactions made, and transactions not made.





Elbo Plot

Elbo plot does not show clearly number of clusters. There is slight elbo showing there coudl be 3 clusters, but it is not reliable presentation.

