**Week 5: Final Project**

**Author’s Name:** Team 6

**Course Number:** ALY6110

**Course Title:** Data Management and Big Data

**Academic Term:** Fall 2019 CPS Analytics

**Instructor’s Name:** Mr. Daya Rudhramoorthi

**Assignment Completion Date:** 05-10-2020



**Analysis**

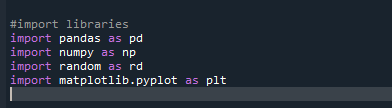
As a part of analysis I performed clustering of data using K-mean algorithm. The problem statement that we have considered here is that which age group is performing the maximum trip duration. The results obtained with K-mean are as follows:

Here I would like to mention why we use K-mean algorithm. Since our data set is unlabeled multidimensional dataset. So, k-mean would be one of the best algorithm. Also, I have implemented K-mean in two ways i.e first by building k-mean from scratch using numpy and pandas and another way by using inbuilt function for k-mean in sklearn library . Reason behind using these 2 different approaches are as follows:

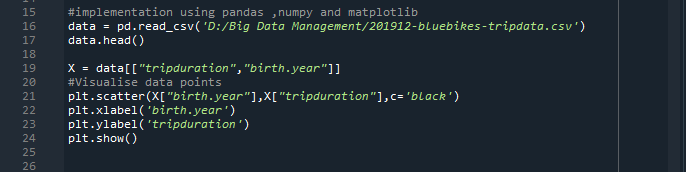
1. By building model from scratch wanted to represent the centroids and distances for each cluster for all the iterations.
2. Also wanted to represent the results with different visuals.

Will explain the results for both.

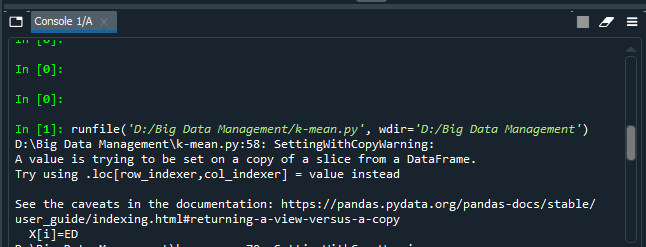
For building k-mean from scratch we require following libraries.



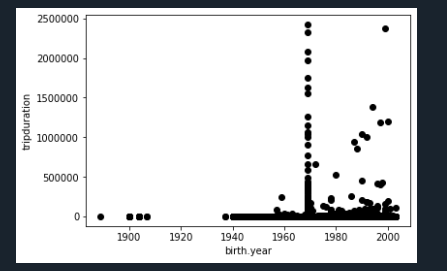
Now we will do data reading and processing as follows. We are using 2 variables i.e trip duration and birth year and plotting the result obtained on X and Y axis.



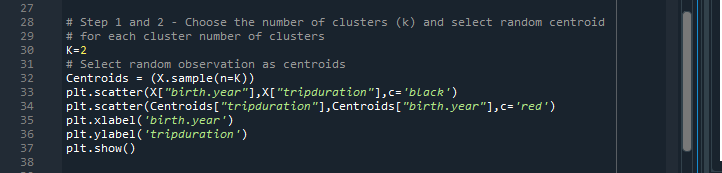
Here are the results:



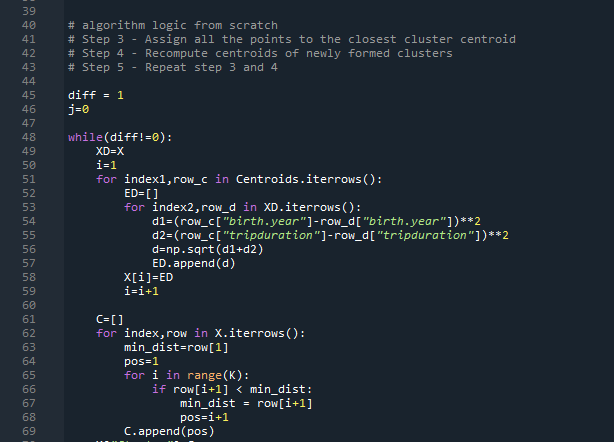
From the obtained results the visuals which are plotted clearly represents that those born in year 1969 have travelled maximum trips for long durations and rest others are showing different behaviors.

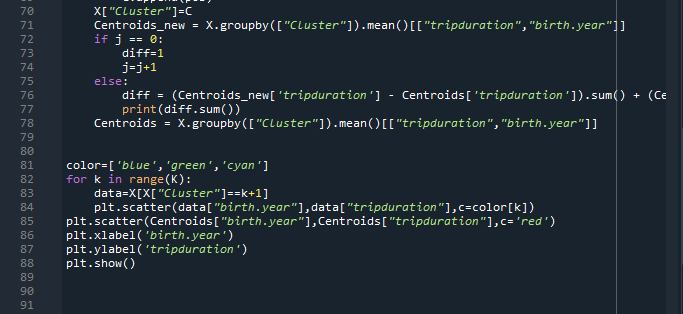


Now for k-mean algorithm we have selected cluster size 2 as the optimized results are obtained with this size. We have tried with cluster size 3 and 4 as well. Below is the logic for same.

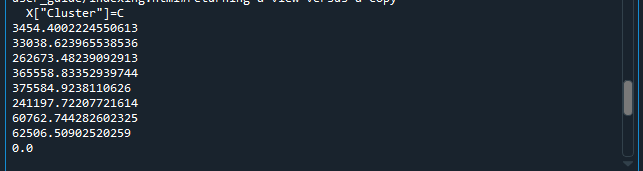


Now for implementing k-mean algorithm from scratch we have to assign all the points to the respective closest cluster. Also we have to recompute the centroids of the newly created clusters.

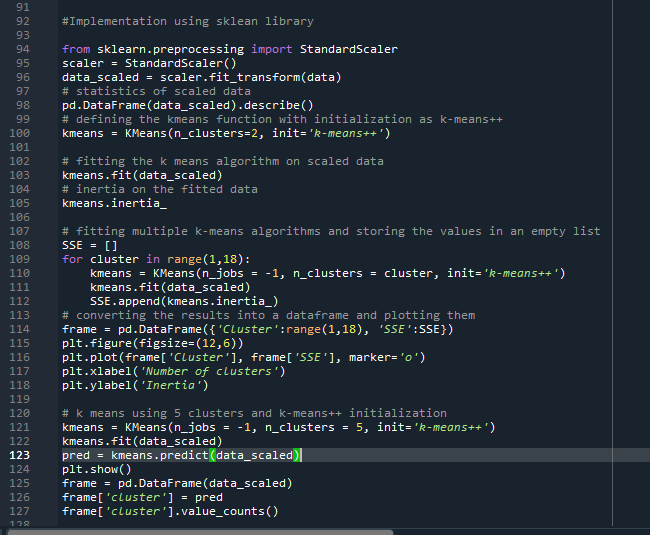




Here are the results of sizes of the clusters which are created for each iteration.



Now I would like to explain another method for k-mean implementation using sklearn library. Here is the logic for the same:



From the results obtained after applying the algorithm we can generalize that those born in 1969 are taking maximum long trips and the elbow curve is also showing up the ideal shape for k-mean.

