**Final Project Analysis Report**

**Author’s Name:** Group A (Shivani Sharma and Rupali Somani)

**Course Number:** ALY6015

**Course Title:** Intermediate Analytics

**Academic Term:** Fall 2019 CPS Analytics

**Instructor’s Name:** Joe Manseau

**Assignment Completion Date:** 12-12-2019



**Introduction**

In our final project we want to implement clustering techniques on a data set containing census data in various countries and building linear regression model on top of it. I would like to give a brief introduction about “Clustering”. It’s an approach of grouping similar things together in a group named cluster. It is most popular method used by data analysts and its mainly a data mining concept. Clustering is mainly used for the data mining application areas like pattern recognition, image processing, computer graphics, information retrieval. Basically, clustering is of two type i.e soft and hard clustering. In hard clustering each data object can either belong to a cluster completely or do not belong to cluster at all. On the other hand, in soft clustering a data point can belong to more than one cluster with little probability or similarity. There are different categories of clustering algorithms.

1. Connectivity based clustering algorithm: Clusters are created by connecting data points according to the distance between them. “Hierarchical Clustering” is based on connectivity-based clustering algorithm.
2. Centroid based clustering algorithm: Clusters are represented by centroids or a central vector. K-mean clustering is based on this algorithm.
3. Distribution based clustering algorithm: It is based on the concept that what is the degree of randomness for a point to belong to a distribution like gaussian distribution. Gaussian mixture model is based on this algorithm.
4. Density Based algorithm: Clusters are defined by the areas of high density within a region as compared to other regions. DBSCAN is based on this algorithm.

**Analysis**

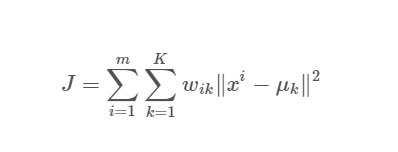
In our final project we are implementing k-mean clustering on our census data set and linear regression model. We have picked up K-mean clustering technique. In order to make uniform data set I have mapped the country names to country codes. And now the data set is univariant. Now coming to the analysis of code and results of our project. So let’s begin with the K-mean. I would like to give a brief about few parameters used in both the techniques:

In order to calculate k-mean we can use either Euclidean distance or manhattan distance. K-mean is based on iterative approach to partition the data set into k pre-defined unique non overlapping clusters where each point belongs to only one group. This algorithm makes inter cluster data points similar and also tries to make all clusters as far as possible. It assigns data points to a cluster so that the sum of squared distance between the points and centroid of the cluster is minimum. Following are the steps of this algorithm:

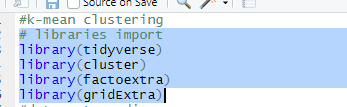
1. Randomly select K rows i.e k centroids.
2. Now assign each data point to the nearest centroid.
3. The algorithm again calculates centroids as the mean of all the data points present in a cluster.
4. Assign the data points to the nearest centroids.
5. Now, repeat steps 3 and 4 till the time observations are not assigned again.

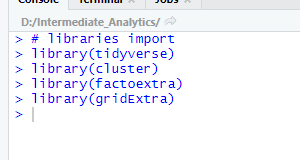
The problem solving algorithm is based on Expectation Maximization which can be mathematically expressed as follows:

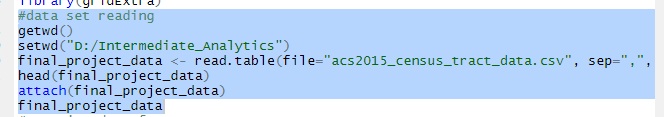
Objective function:

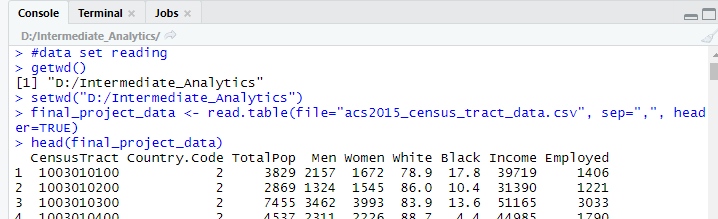


So our step one is loading all the required libraries and reading the test data file i.e csv and storing it in a variable.

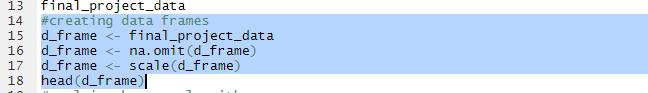


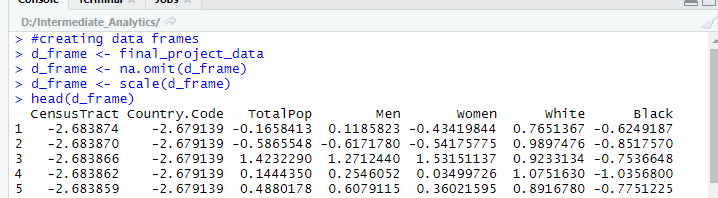




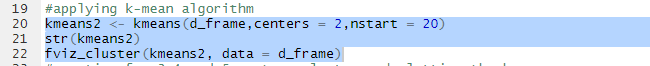


Now we have to create a data frame so that we can feed this to k-mean algorithm.

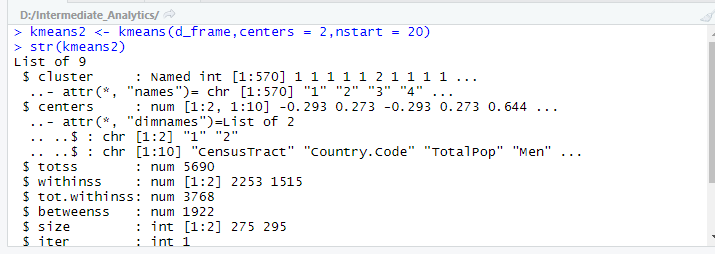


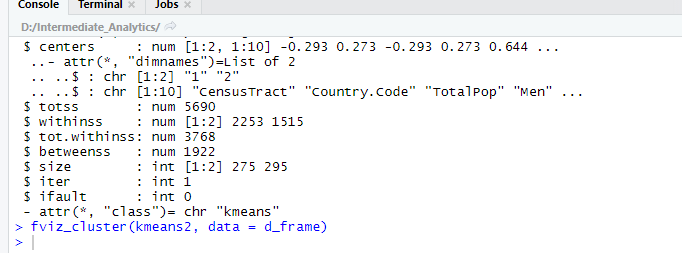


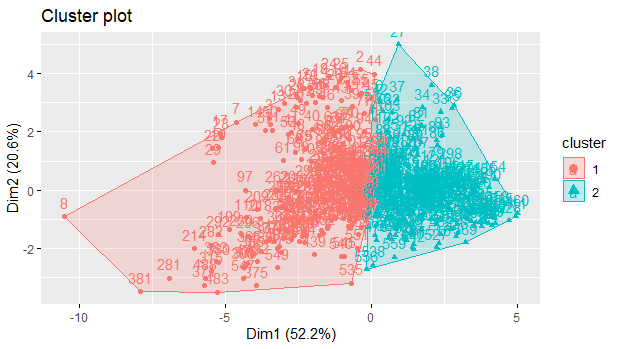
Now applying k mean algorithm on the data frame we created



The results are showing the cluster details and array details. It also displays the column header names.

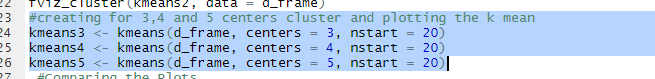


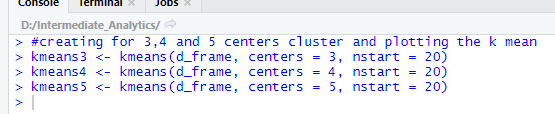


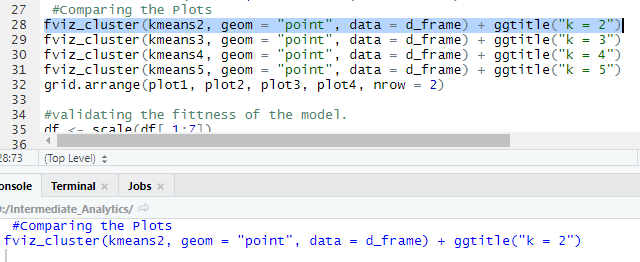


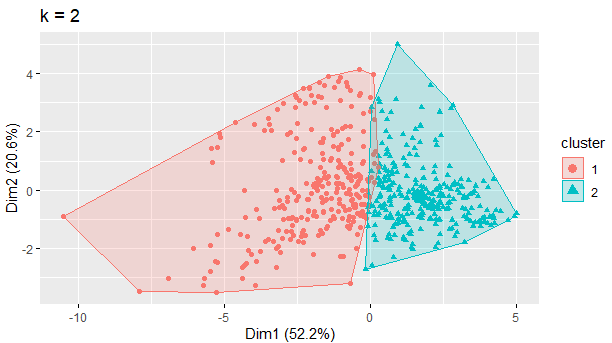
These components are illustrating 52.2% of the point variability. Now I am creating kmeans for 3,4 and 5 centers and plotting the respective kmeans for them.

NOTE: we are observing the behavior and pattern of our data set by plotting it for k=3,4,5 and in order to confirm the optimal number of clusters will do a second test i.e WSS test and based on the shape of the graph we will confirm the number of optimal clusters and then further build the model on it and find out the fitness of the algorithm.

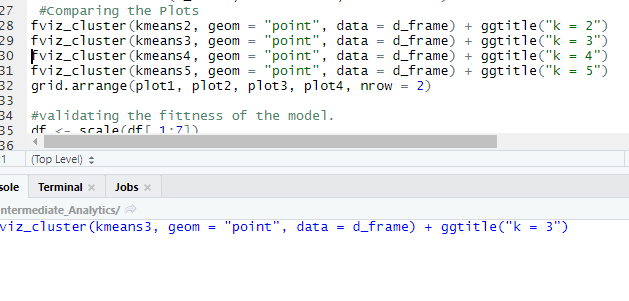


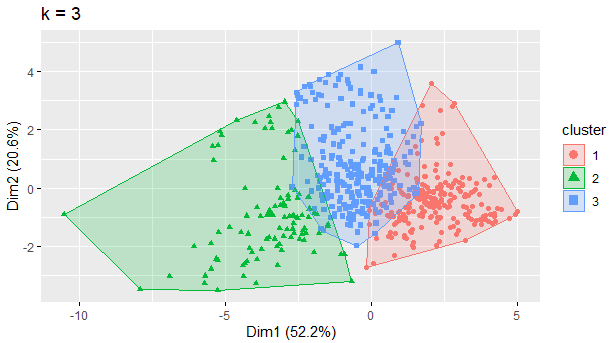




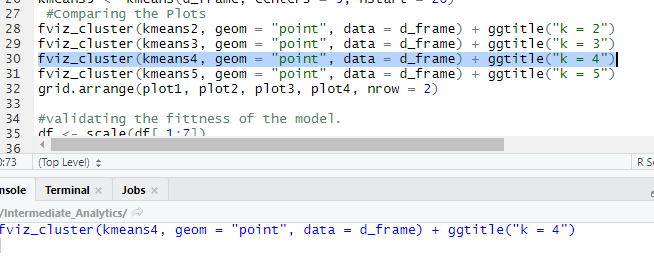


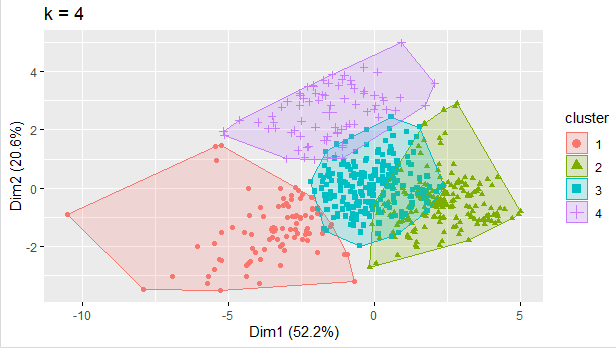
From the above plot we can observe the point of variablity is 52.2 % between the components and also the clusters are overlapping a bit so we will plot if for k=3 again and see the difference among the clusters.



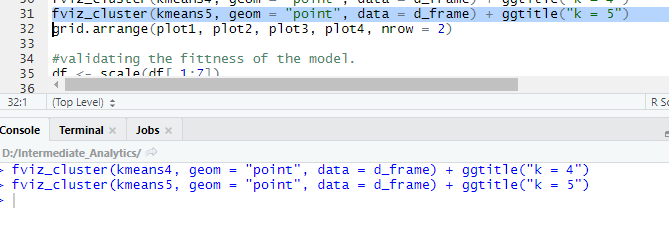


From the above graph the clusters are clear and the centroids are having clear distance from each other. However we will still plot it for k=4 and 5 in order to validate the behavior of the clusters and data set.

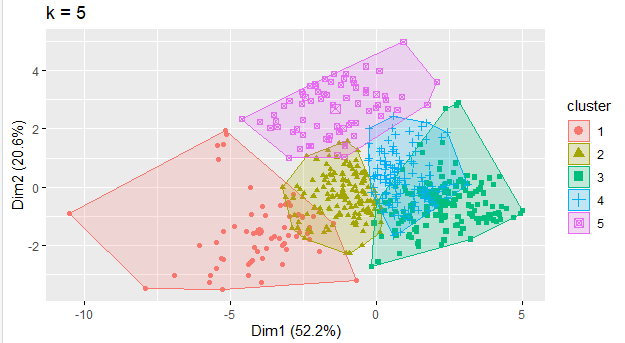




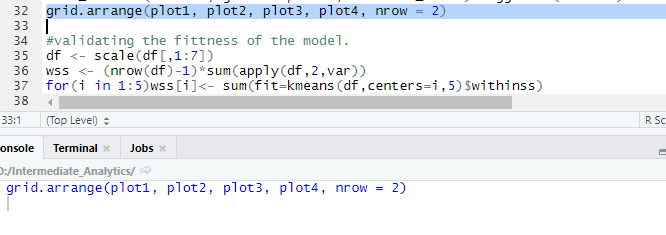
For k=4 the results are not that optimal as the centroids are not having clear and uniform distances between them. It is clearly observed and even if we find out the values using mathematical formulae then it will show up the overlapping region.

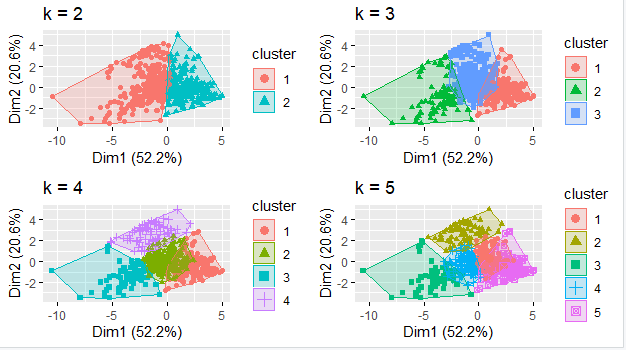


Though we are clear with the results of k=4 and need not a validation for k=5 but still we are checking it for making a fair selection of optimal number of clusters and comparison.

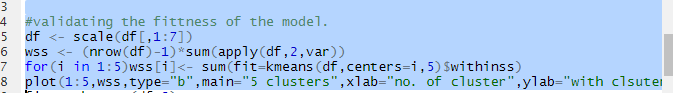


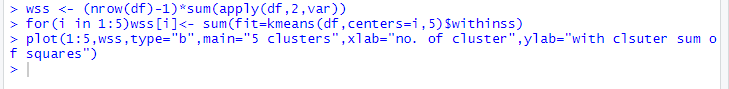
In order to compare the behavior of clusters I am adding them on a grid using below mentioned function.

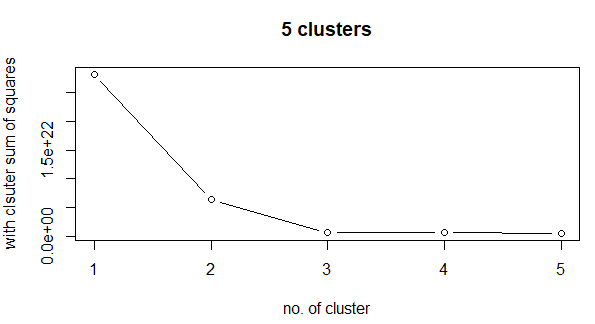




Now we will find out the optimal number of clusters using the WSS test and plotting the number of clusters against the sum of squares.

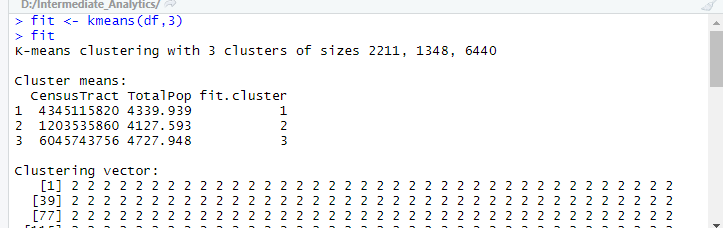


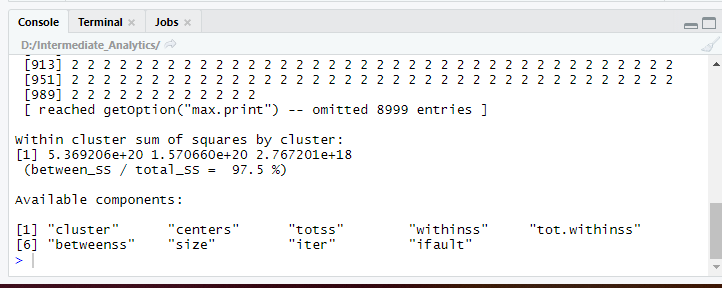




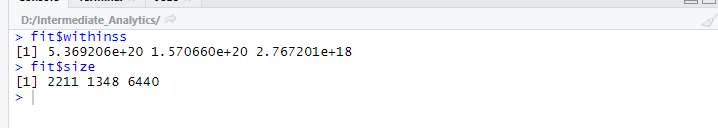
From the above graph we can conclude that the optimal number of clusters is 3 as the curve is showing a elbow point on this one so we will build the model on k=3 and find out the fitness of the algorithm on k=3



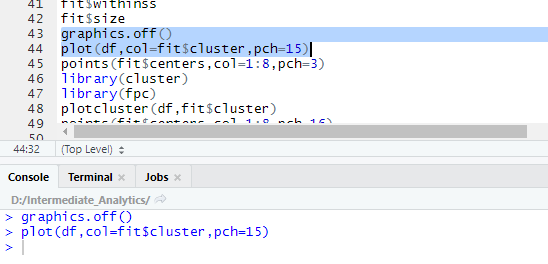


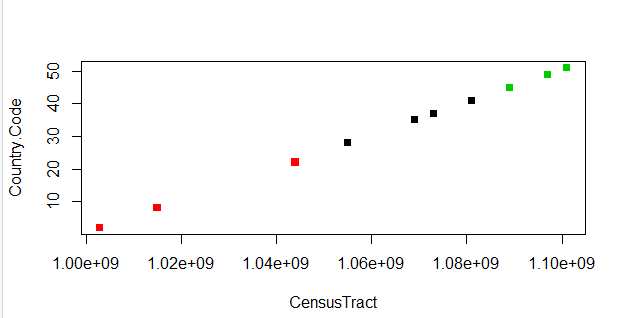






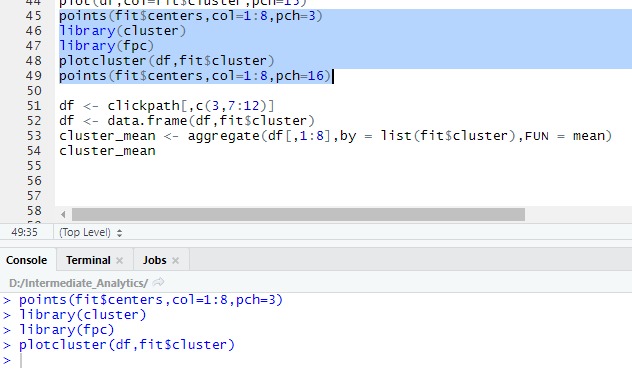
From the above results we can observe the values of sum of squares and size of the clusters. Also the most important point here is the percentage of the individual sum of square’s of a cluster and total sum of squares is 97.5 % which is a strong value to support the fitness of the model.

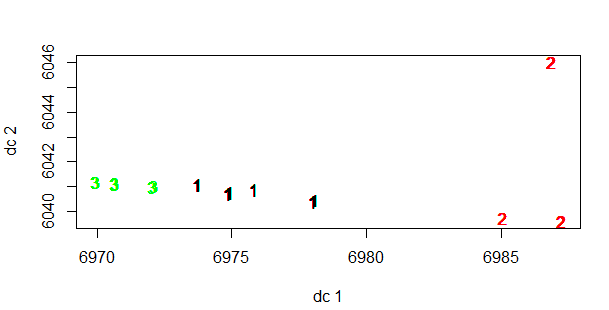




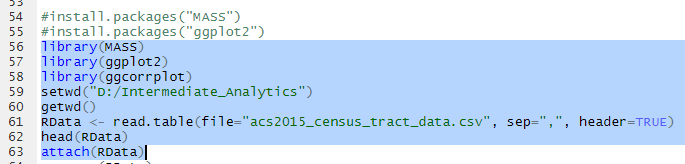
From the above graph we can conclude that there is a linear relationship between country codes and census Tract.

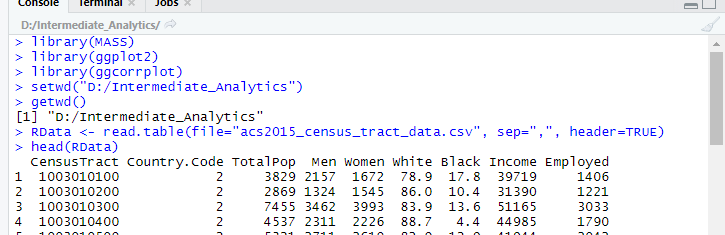
Now in order to give a more clear insight and deep dive into the algorithm and showing the behavior for k=3. I am plotting the fitness against cluster for showing the behavior

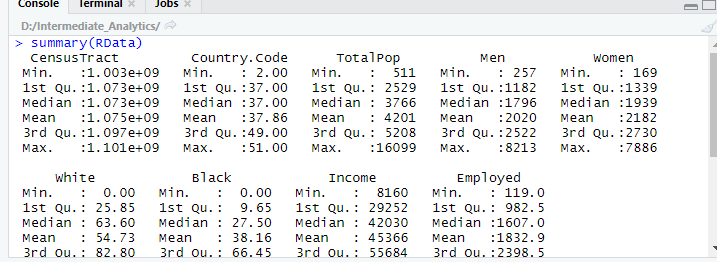


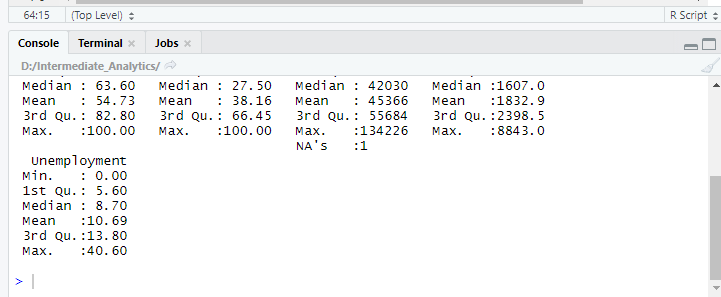


Secondly, I wanted to build a linear regression model on top of this data set in order to confirm the presence of extent of linear relationship between the parameters. Also, we will plot the correlation between the parameters. First of all we need to load the libraries i.e MASS, ggplot2 and ggcorrplot. Summary of the data set is also calculated.

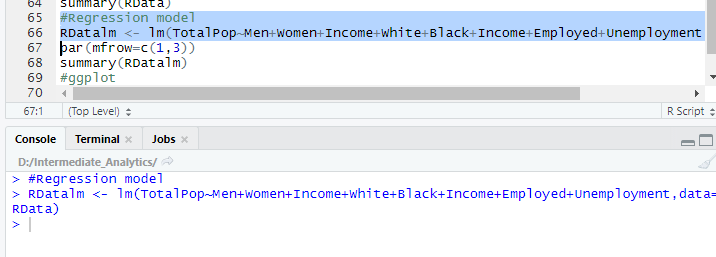


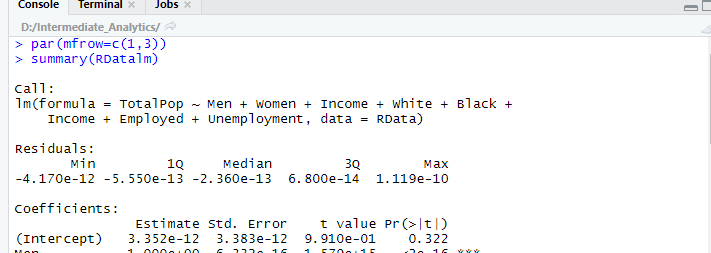


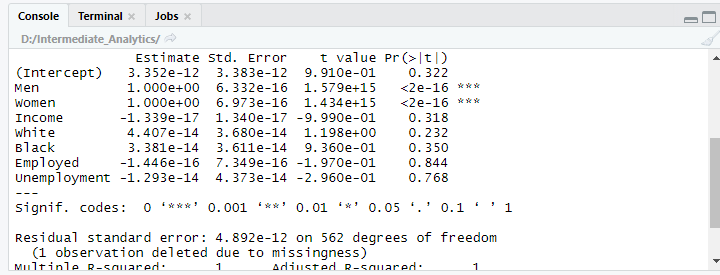


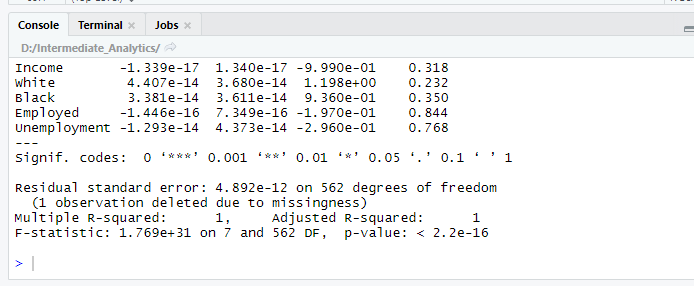


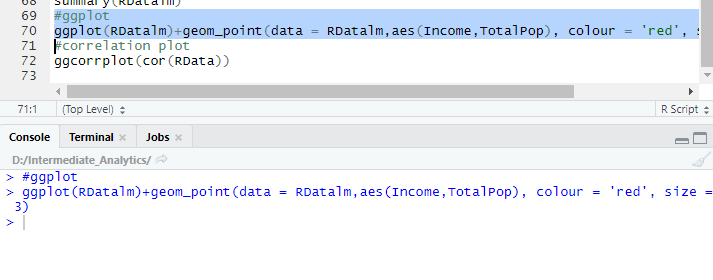
Now in order to check the relationship between the parameters we need to build a linear model on top of data sets



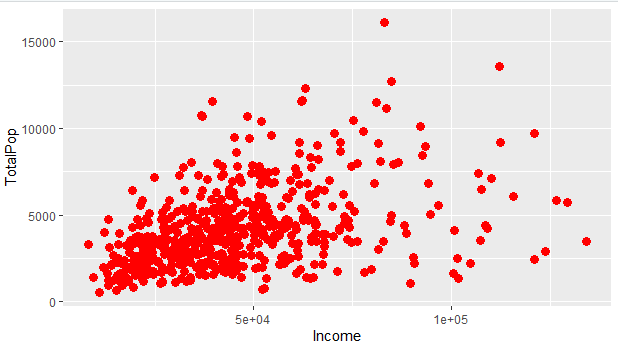


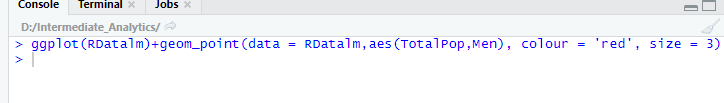


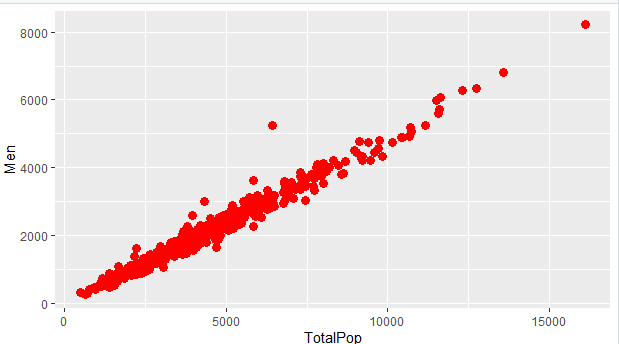


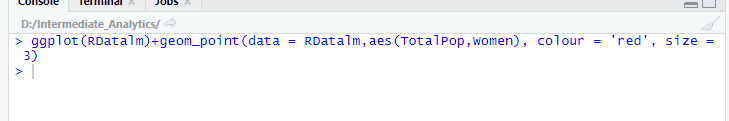


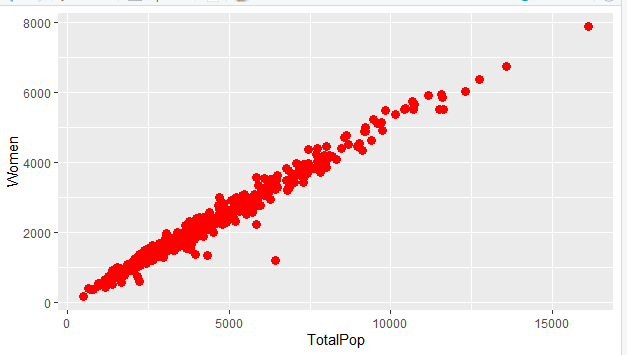
From the results we can observe that the p value is very high and ggplot is showing the relationship between total population and income. Similarly, we can plot the relationship between other parameters too.

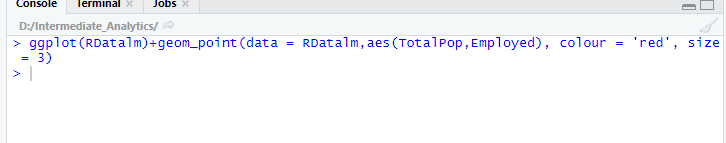


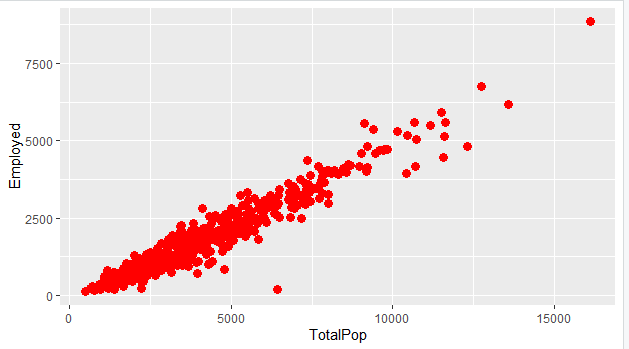


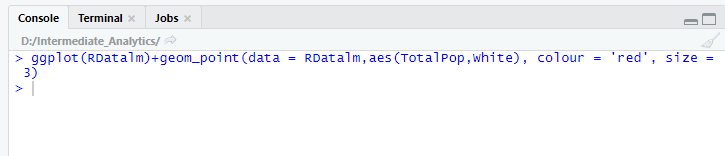


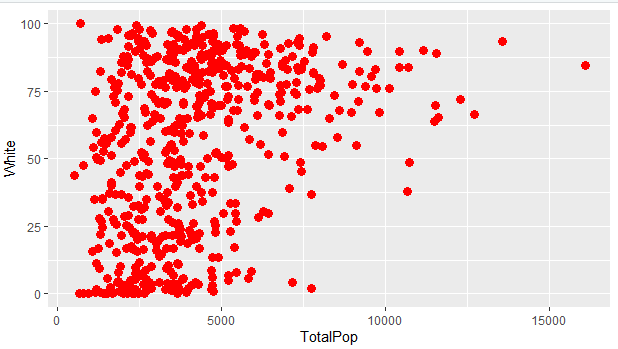


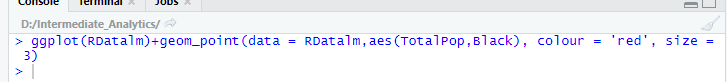


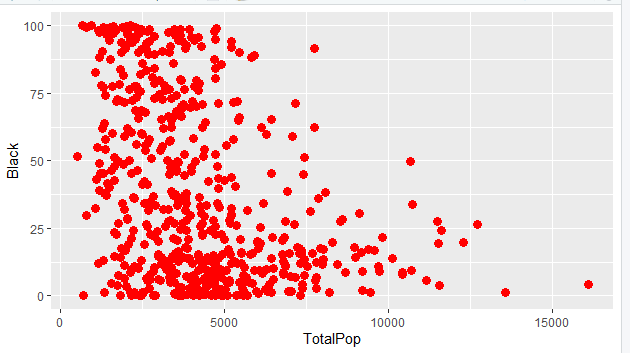


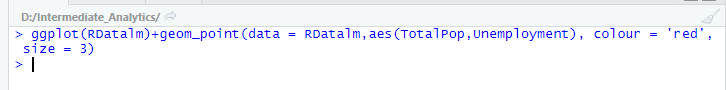


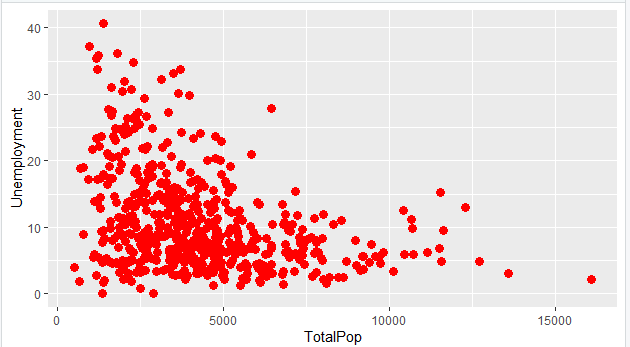






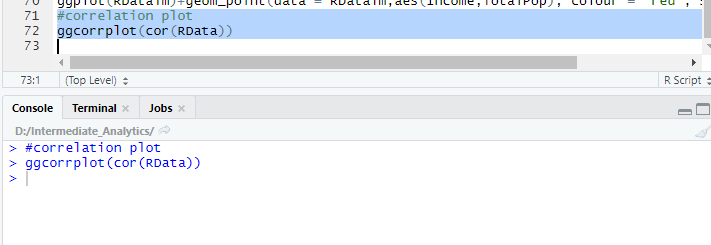


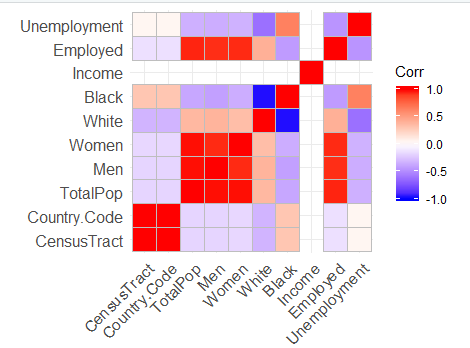




From the above graphs we can conclude that the most linear relationship exist between total population and Men, total population and Women, total population and employed and a straight line relationship exist between these parameters.

Now I want to show the correlation between the parameters using below mentioned logic.





From the graph we can the red spots which states that these are maximum correlated to each other.

**Conclusion**

From the results of above two algorithms i.e k-mean and linear regression we can conclude following:

1. For k mean clustering we have observed that the cluster size 3 is showing most optimal results as the elbow point of WSS graph results shows the point at k=3 and clustering is best illustrated using k=3 and model is showing up the clusters clearing. And the relationship between data sets i.e total population vs categories i.e Men ,women and employed can be clearly figured out.
2. From the results of linear regression and ggplot we can conclude that a linear relationship only exist between total population, Men ,women and employed values.

# References

1. Garbade, Michael J. “Understanding K-Means Clustering in Machine Learning.” *Medium*, Towards Data Science, 12 Sept. 2018, <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>.
2. *CS221*, <https://stanford.edu/~cpiech/cs221/handouts/kmeans.html>.
3. Dabbura, Imad. “K-Means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks.” *Medium*, Towards Data Science, 3 Sept. 2019, <https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>.
4. “Sign In.” *RPubs*, <https://rpubs.com/AnanyaDu/361293>.
5. “Sign In.” *RPubs*, <https://rpubs.com/gabrielmartos/ClusterAnalysis>.
6. MuonNeutrino. “US Census Demographic Data.” *Kaggle*, 3 Mar. 2019, <https://www.kaggle.com/muonneutrino/us-census-demographic-data>.