

**ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ**



**ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS**

Course: Statistics II for Business Analytics

Professor: Karlis Dimitrios

Student: Sergouloupoulou Maria

Student ID: f2822016

Assignment: 2nd Assignment

Table of Contents

Introduction	3
Data.....	3
Aim.....	3
Background Information.....	3
About the data	5
Predictive Models	5
Decision Tree.....	5
Random Forest	7
Logistic Regression	9
Comparison of the Models.....	9
Clustering	10
The Data.....	10
K-means	11
Silhouette	11
Description of the Clusters with the Economic Related Variables	14
<i>Figure 1-Decision Tree</i>	6
<i>Figure 2-Decision Tree Confusion Matrix</i>	6
<i>Figure 3-Mean Decrease Gini</i>	7
<i>Figure 4-OOB Error</i>	8
<i>Figure 5-Random Forest Confusion Matrix</i>	8
<i>Figure 6-Confusion Matrix Logistic Regression</i>	9
<i>Figure 7-Within Clusters Sum of Squares</i>	11
<i>Figure 8-Silhouette Plot with 5 clusters</i>	12
<i>Figure 9-Silhouette Plot with 6 Clusters</i>	13
<i>Figure 10-Number of Observations in Clusters</i>	13
<i>Figure 11-Parallel Plot</i>	14

Introduction

In this assignment we are called to work with prediction models and clustering. I am using a dataset that contains demographic and economic characteristics of the population during the American election of 2016. Following I am trying to find the best prediction model for the people that voted for Trump and to do so I am using and analyzing three different prediction models. Continuing the analysis, I move forward to clustering, where I determine the optimal number of clusters and then complete the assignment by explaining the clusters with the economic characteristics of the dataset.

Data

The data we are using for this assignment are a mixture of the socioeconomic characteristics such as age, education, living conditions, size of income of the voters and the other sheet the votes casted for Donald Trump per county and state.

Aim

This study is divided in two separate parts. In the first part, the main aim is to predict if Donald Trump got more than 50% of the votes. This must be achieved by creating a predictive model using three different methods in order to collect enough proof about the quality of the models. For the second part, we are called to separate the dataset into demographic and economic characteristics and create clusters that contain the counties and can be explained and described by the economic characteristics.

Background Information

In this assignment I am going to rely a lot on the analysis that I did in the previous one. For the first assignment I had to create a model that was explaining the best way possible the profile of a voter in the election of 2016. In order to conclude to the best model, I started with a model that contained all the covariates in the dataset and I had a response variable called success which explained whether the voter voted for Donald Trump or not.

In the process that I followed I did stepwise procedure directed bothways to find the model with the best AIC that explains the quality of the model. Moving forward, I had to check for multicollinearity and I used the variance inflation factor as metric. Each time I would remove the covariate with the highest vif, create a new model and repeat the command again to find the next one with the highest vif until there was no covariate with vif over than 5.

Once I solved the multicollinearity problem, I checked the residual deviance, the AIC value and the pseudo R2 metrics to make sure that the quality and the fit of my model was improving. Continuously, I used the command summary to see if all my covariates included in my model were statistically significant. At this point, I noticed that 4 covariates were not statistically significant and I removed them. I checked again the vif, aic and pseudo r2 and everything was improving. I also compared the two models and found out that the complex model was not more accurate than the nested one and the fit of the nested was better.

I concluded to the final model after removing two more covariates that were not statistically significant. Again, I repeated the process by checking the metrics to make sure the quality and the fit was improving, AIC and mcfadden scores were better and the test showed that the nested model was better than the complex.

The model I concluded:

Success = -1.468e+00 -6.379e-02PST120214 +4.461e-02AGE135214 +9.423e-02AGE775214 -1.339e-01SEX255214 -2.139e-02RHI225214 -7.824e-02RHI625214 -7.824e-02RHI625214+1.016e-01POP715213-4.679e-02POP645213 -3.685e-02EDU685213 -6.067e-02HSG445213 +2.104e-02HSG0962 +1.097e-05HSG495213 +4.311e-02PVY020213 +1.908e-02SB0315207 -4.952e-01SB0515207 +4.937e-03 SBO415207 +1.512e-04 LND110210

The characteristics that are involved in this model are the change of the population between 2010 and 2014, the population with age under 5 years who did not vote, the population with age 65+ most of whom supported Trump, the percentage of female population in 2014 who did not support Trump, black African and American Indian percentages of population, foreign born persons, percentage of people with stable home for over a year who did not support Trump, percentage of people with bachelors degree and higher who were not Trump

supporters as well, the rate of home ownership, the percentage of housing units in multi-unit structures, the median value of owner-occupied housing units, the persons below the poverty level who voted for Trump, the percentage of black-owned, Hispanic and native Hawaiian firms, the land area in square miles and the population per square miles.

About the data

The data I am going to use for creating the predictive model are the covariates that were included in the best model that was produced in the previous assignment. This way I am going to avoid adding noise in my dataset and focus on the variables that contain useful information. For the modeling, I am going to split my dataset into 80% training set and 20% testing set.

Predictive Models

Decision Tree

A Decision Tree is a Supervised Machine Learning algorithm which looks like an inverted tree, wherein each node represents a predictor variable (feature), the link between the nodes represents a Decision and each leaf node represents an outcome (response variable). The decision tree is built through a process known as binary recursive partitioning, this is an iterative process of splitting the data into partitions and then splitting it further on each of the branches. A decision tree is inexpensive to construct and easy to interpret if it has a small size. On the other hand, it is easy to overfit and small changes in the training data can result in big changes to decision logic.

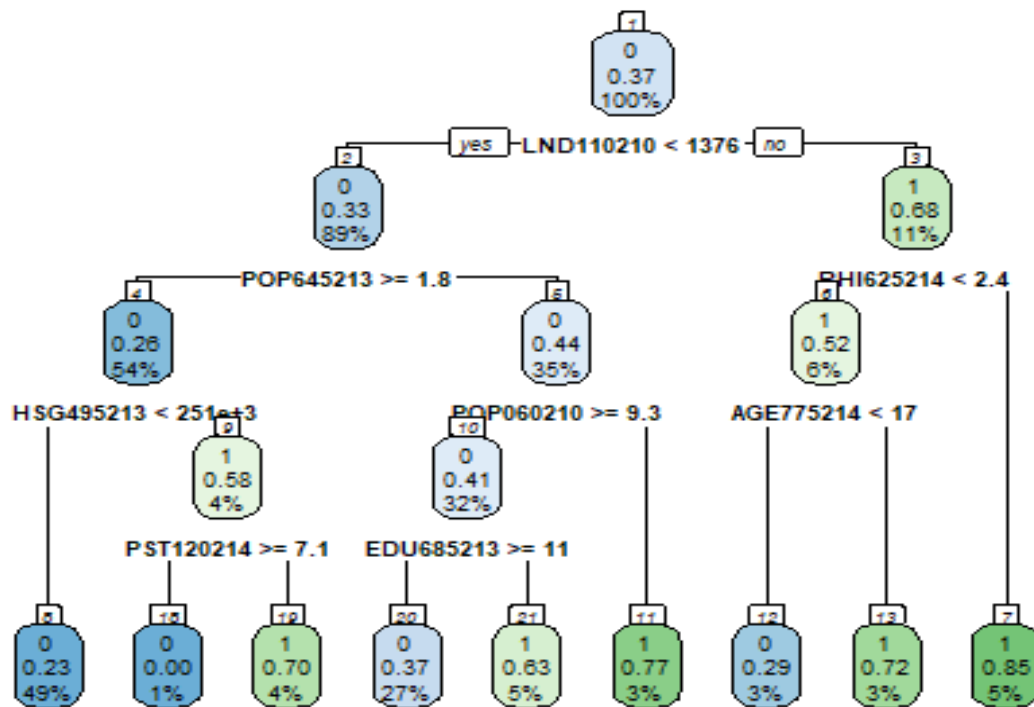


Figure 1-Decision Tree

The packages used for the tree model and the plot are rpart and rpart.plot.

In this plot we can see that 8 variables were used for the creation of the tree and we can identify them. In my case the first split has to do with the land area in square miles, then the population of foreign born persons and the percentage of two or more races, the median value of owner-occupied house units, the population per square miles and the persons with age 65 and over, the percentage of the population change and the percentage of persons over 25 years old with bachelors degree.

Another thing that we can learn from the plot is that the number of loops of the tree are nine.

		Actual	
Predicted		0	1
	0	289	131
	1	47	75

Figure 2-Decision Tree Confusion Matrix

The Decision Tree method predicts correctly 364(diagonal) out of 542 instances.

Random Forest

The forest is built from an ensemble of decision trees based on the idea that a combination of learning models increases the overall result. Since the results of each tree are merged the disadvantage of overfitting of the decision tree can be corrected. Another great quality of the random forest algorithm is that it is very easy to measure the relative importance of each feature on the prediction. Finally, its ability in prediction can be more accurate.

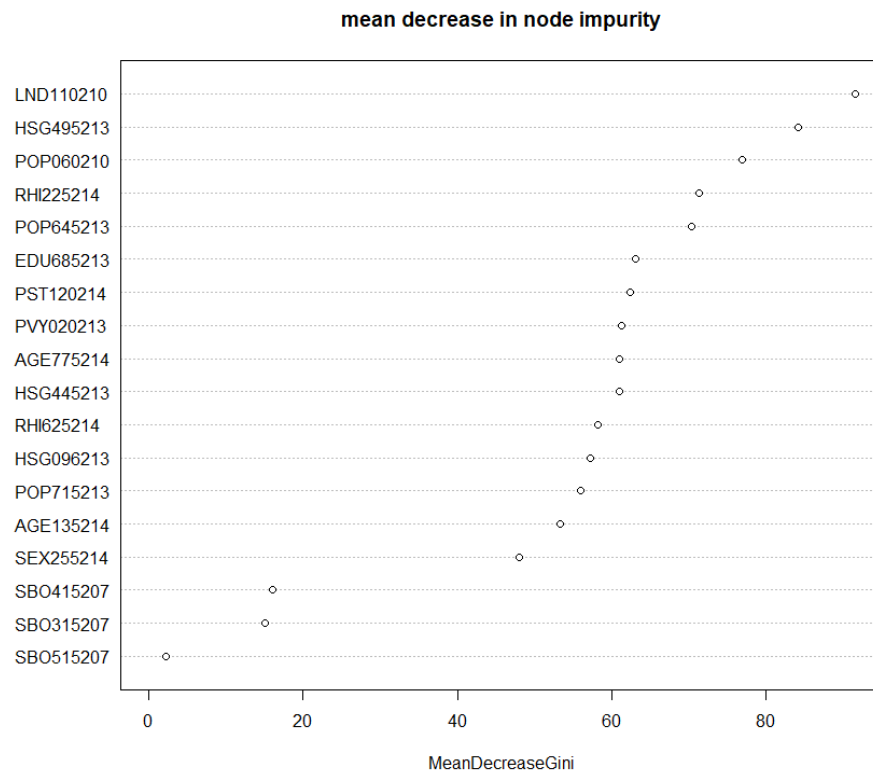


Figure 3-Mean Decrease Gini

Gini index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen. The more the Gini Index decreases for a feature, the more important it is.

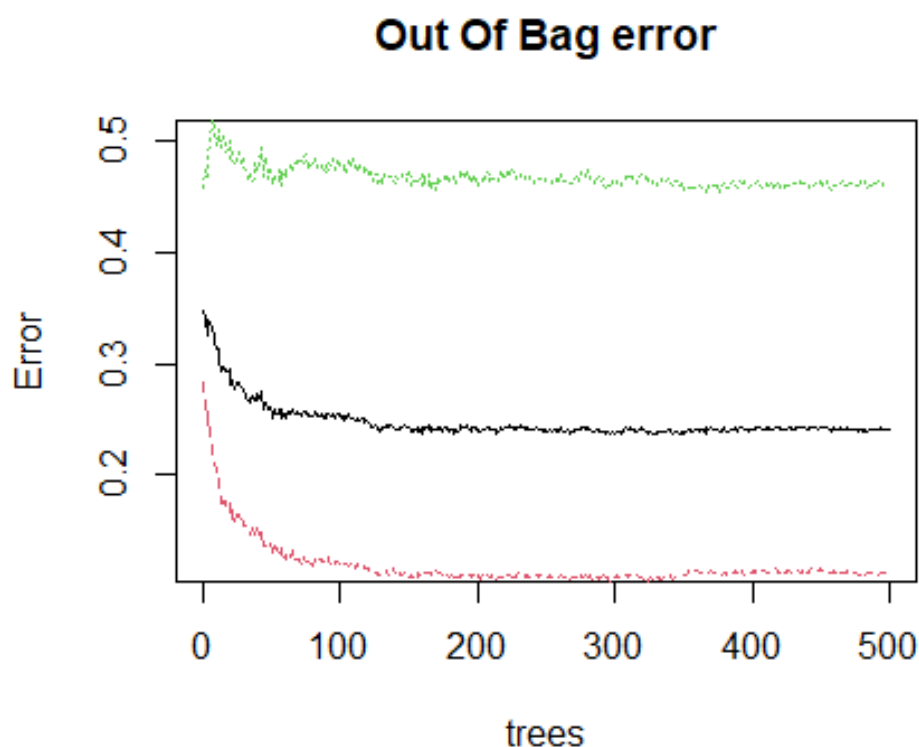


Figure 4-OOB Error

Out of bag (OOB) score is a way for the Random forest model to be fit and validated whilst being trained. The random forest predictor is constructed by bootstrapping a fraction of the samples, the OOB error is calculated on the samples that were not selected (out of the bag) on each iteration of the algorithm. The plot above shows how the OOB can be measured each time a tree is added during training and we can see where the error stabilizes and observe its performance. Our score stabilizes approximately after 150 trees.

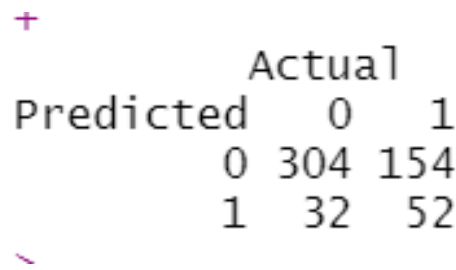
		Actual	
Predicted		0	1
	0	297	102
	1	39	104

Figure 5-Random Forest Confusion Matrix

The Random Forest method predicts correctly 401(diagonal) out of 542 instances.

Logistic Regression

Linear Regression is a way of modelling the relationship between one or more variables in a linear way. The purpose that needs to be achieved in our case is to predict which people are going to vote for Trump or not since our response answers to yes or no.



	Actual	
Predicted	0	1
0	304	154
1	32	52

Figure 6-Confusion Matrix Logistic Regression

Logistic Regression method predicts correctly 356(diagonal) out of 542 instances.

Comparison of the Models

	DECISION TREE	RANDOM FOREST	LOGISTIC REGRESSION
ACCURACY	67%	73%	65%
SENSITIVITY	86%	88%	90%
SPECIFICITY	36%	58%	25%

The table above shows the quality of the model based on three terms

- Accuracy: How many observations were predicted correctly.
- Sensitivity: The proportion of observed positives that were predicted to be positive.
- Specificity: The proportion of observed negatives that were predicted to be negatives.

In more words, we observe that the best model in terms of accuracy is the random forest with the score of 73%. The model with the best score of sensitivity which means the percentage of persons that the model predicted correctly that will vote for Trump is Logistic Regression with 90% and the model with the best specificity score which is the correct prediction percentage of people that will not vote for Trump is Random Forest with 58%.

Generally, our models have really good scores except in terms of specificity that they are lower than what we would want them to be. The best model turns out to be Random Forest since it had better forecast scores than the others.

Clustering

For the second part of this assignment, we are called to use the demographic part of our data to cluster the counties. What we are looking for, is to determine if there is an economic heterogeneity in the population, if this heterogeneity can be divided into distinct groups, how many of those groups exist and the total observations that exist in each group.

For the purpose of clustering the characteristics of Trump voters and any other part of our previous analysis do not have any information needed so they are removed from the process.

The Data

The data in this part of the assignment were divided into two groups the demographics and the economics. Before the division, I deleted from the dataset any column that did not belong to neither of those two categories. Continuously, I scaled my data which centers and scales the columns of a numeric matrix. Scaling normalizes the data in order to have zero mean and variance equal to 1. This is important to make sure that our calculations will not be biased either to the very high or to the very low values.

K-means

For my clustering process I used K-means method. k-means is an iterative method for minimizing the within-class sum of squares for a given number of clusters. The algorithm starts with an initial guess for cluster centers, and each observation is placed in the cluster to which it is closest. The cluster centers are then updated, and the entire process is repeated until the cluster centers no longer move.

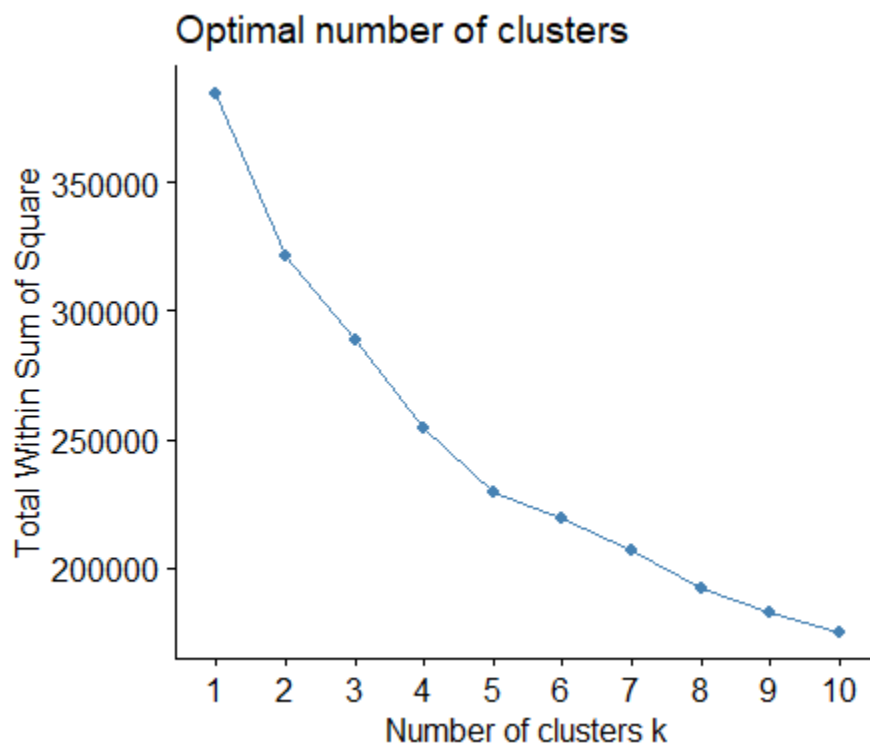


Figure 7-Within Clusters Sum of Squares

I used the plot of within clusters sum of squares to help me determine the number of clusters I should choose to have. WSS means the sum of distances between the points and the corresponding centroids. In this case it is difficult to be sure if the optimal number of clusters is 5 or 6 because the correct number of clusters should be before the lines' falling abruptly. For this reason I decided to do more searching to determine what fits the best.

Silhouette

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a graphical representation of how well each object has been classified.

The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters. Its values are from -1 to 1 where positive numbers means that the observations are classified in the right cluster, zero or close to zero observations are in the edge of right classification and the negative numbers means that those observation probably belong to another cluster.

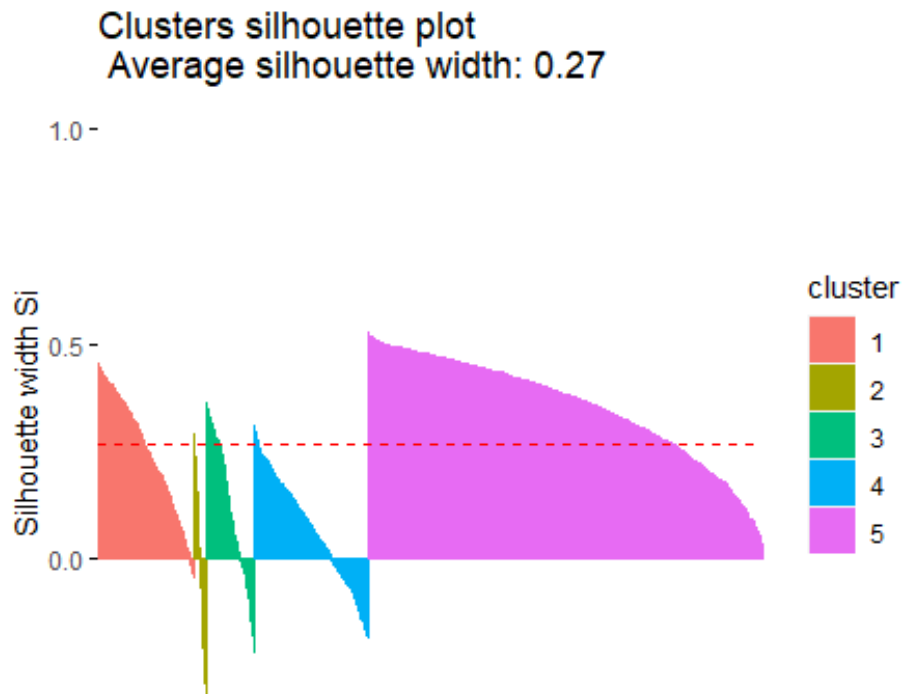


Figure 8-Silhouette Plot with 5 clusters

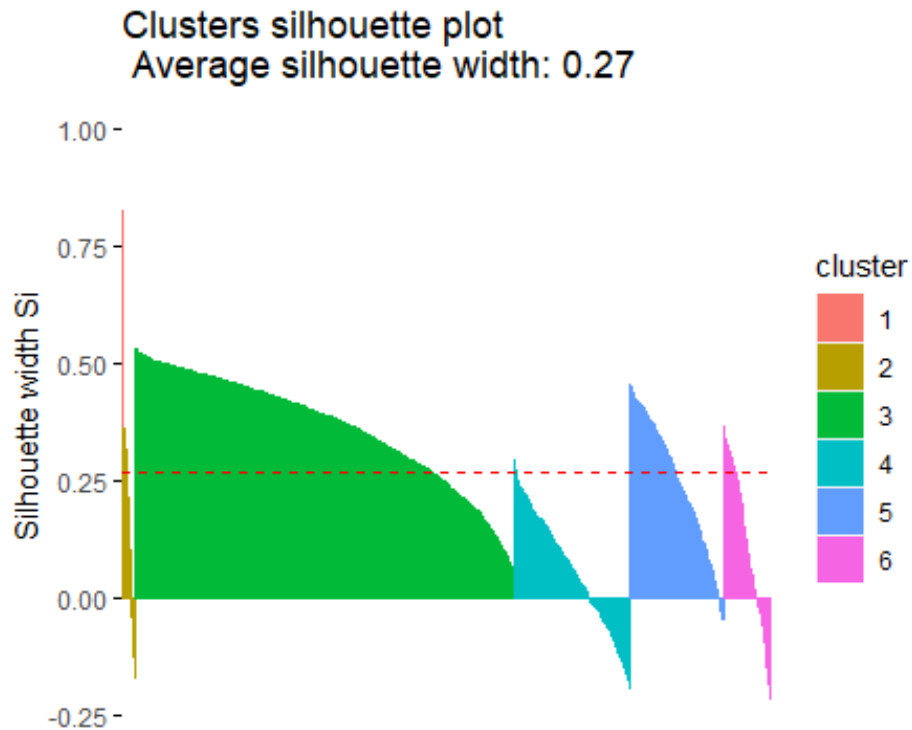


Figure 9-Silhouette Plot with 6 Clusters

Those silhouette plots represent the classifications of observations to 5 and to 6 clusters respectively.

I will choose to have 5 clusters since in the second plot the cluster number 1 cannot be found and in general comparison the presence of clusters with below average silhouette scores are the same. Another thing to consider here is the fluctuation in the size, it is similar between the two plots here.

	cluster	size	ave.sil.width
1	1	2515	0.25
2	2	319	-0.02
3	3	1278	0.11
4	4	2981	0.07
5	5	10386	0.36

Figure 10-Number of Observations in Clusters

Here is the number of clusters chosen and the division of the observations for each cluster along with the average silhouette width.

Description of the Clusters with the Economic Related Variables

For describing the clusters, I merged the cluster with the economics data frame. Then I plotted the relationships to understand better the characteristics of each cluster.

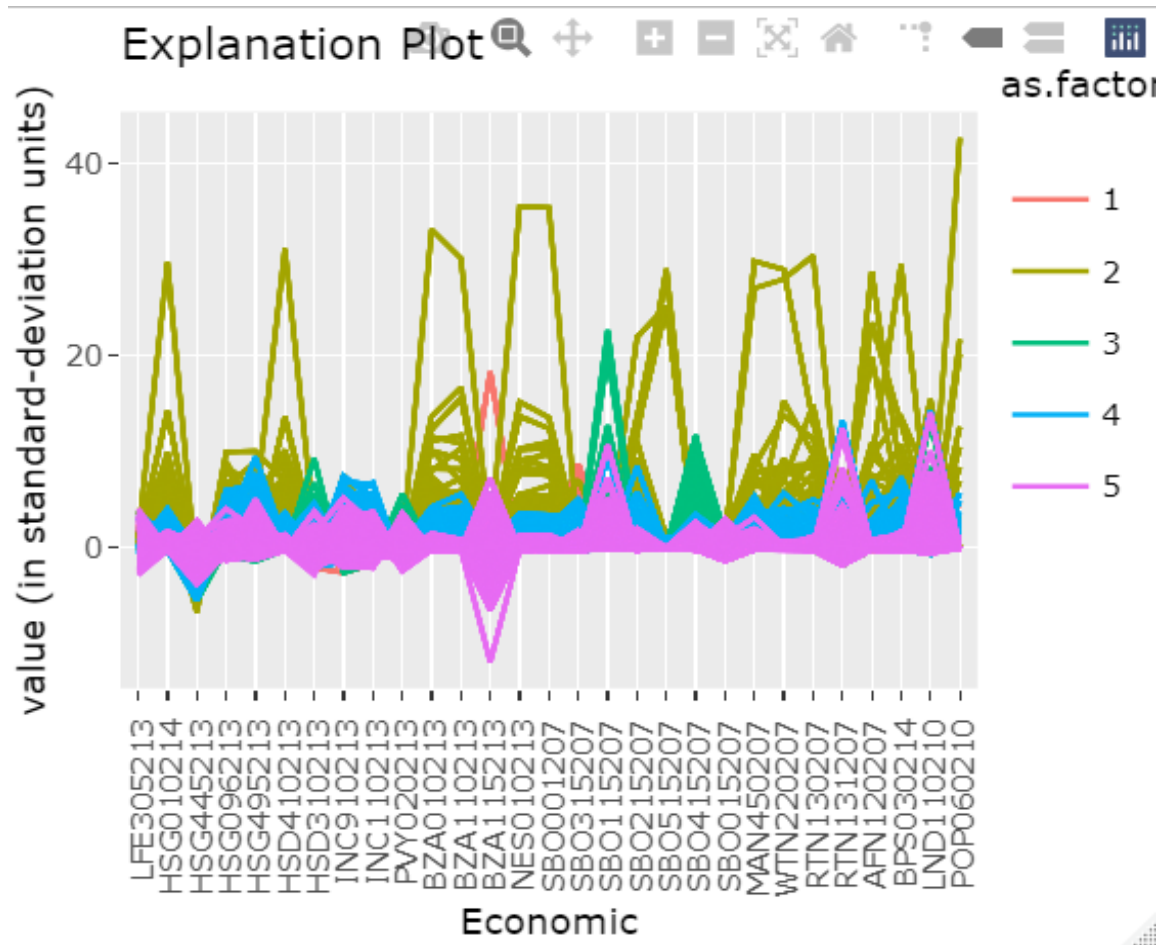


Figure 11-Parallel Plot

Cluster 1: Clusters' one observations are not really visible in the plot and so we understand that in terms of mean travel to work and housing circumstances as house unit value, home ownership percentage, persons per house unit, this county holds average values in relation to the other counties. It includes a large percentage of persons who have non-farm private employment and an above average percentage of black owned firms. Also, the values of persons per household are slightly below average. In terms of other non-Caucasian owned firms, retail sales of 1000\$ and per capita, sales in different business areas, land area in square miles, building permits and population per square mile, its values range to the general average.

So, by those characteristics we understand that this county is probably an industrial territory, sparsely habituated, average in size, where black people have flourished business wise.

Cluster 2: Clusters' two lines differ a lot that the ones of other clusters and show a lot of spikes in the plot. This cluster holds the highest values in mean travel time to work, in the number of housing units and households, median value of housing units. Also, it holds high values in private non-farm establishments and employment, in non-employer establishments and in the number of firms. Also, there are high numbers in women owned firms, high number of shipments and sales, accommodations and food services and building permits.

This county is really big and overpopulated, so we are probably referring to large city, with a lot of establishments and non-farm work opportunities. There are plethora of firms and it holds the highest number for women owned firms, so we can say that it is a progressive city. Also, it is safe to say that it has expensive living with great opportunities and experiences in accommodations and services. Last, the homeownership rate is way below average, so we assume most people live on rent.

Cluster 3: Cluster three spikes in characteristics such as the number of households, and in multi racial owned firms. In all other characteristics its values are average in relation to the other clusters.

This county is medium sized, it has average expense of living and it has a lot of different race owned firms. This could mean that in this county there are a lot of immigrants. Its population is average but there is a high number of households so we could assume that there are a lot of families.

Cluster 4: Cluster four differentiates the least from the average. It shows very little spikes, so we assume the most of its characteristics have values close to the average. The little spikes that it shows refer to the percentage of households in multi-unit structures, the per capita income, private non-farm employment, non-employer establishments and slightly above average firms and multiracial owned firms.

This county is not very large neither in land nor in population. It has slightly more than average multi-unit structures and non-farm employment so we can assume that there are a lot of working-class people who do not own their houses, which can be confirmed by having more than average number of firms. Last, its population is mixed race.

Cluster 5: The characteristics of this cluster is that it shows the highest spikes below the average in relation with the other clusters. The values below average are in homeownership rate, in the mean time travel to work and the lowest spikes are in non-farm establishments and non-farm employment. Above average spikes are noticed in native American and Alaska owned firms, in retail sales per capita and land area in square miles.

This county is above average in size but sparsely populated, people live close to their work. This makes sense because the people are mostly farmers so they could live in their farms where they work. Last, there is a larger percentage of native Americans and Alaska natives than in other counties. In terms of other characteristics this county is close to the average.