ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

Classification and Clustering Methods

Athens University of Economics and Business

Master of Science in Business Analytics

Statistics for BA II

Christos Vlassis

f2822204

With ‘**1**’ we classify the observations that have more than 50% of the votes for Trump.
With ‘**0**’ we classify the observations that have less than 50% of the votes for Trump.

Contents

With ‘**1**’ we classify the observations that have more than 50% of the votes for Trump.
With ‘**0**’ we classify the observations that have less than 50% of the votes for Trump.

1. Abstract

In part 1 of the paper the best models that were found was the lasso logistic regression and random forest, both of them having equal good results in terms of accuracy. In part 2 of the paper we found that from the 4 clusters that were created three of them did not have significant economic differences. But the fourth group was found to be a lot wealthier and more populated from the rest 3 clusters that were created.

2. Introduction

The following paper is spited into 2 parts. In part 1 we will create 3 models to predict whether Trump got more than 50% of the votes. In part 2 of this paper we will create 2 different methods to cluster our demographic data. Finally, we will use the economic related data to try and describe the economic situation of these clusters.

3. Data Preparation

To prepare our data for the modelling part we made the same manipulations as the previous paper. So we will not get in details of what was those data manipulation were, because they have already been explained in the previous paper. But, we also had to test whether the proportions of event of interest and not interest are almost the same. If this is not the case, we had to take measures such under Pampling or oversampling. This would be done for our model to be able to predict both cases at its full potential both cases. We found that 37% of the data concern the event ('1') of interest and 63%('0') of the data concern the non-event of interest. With these proportions, and also by taking into account the result of the models(we will see it in the later chapters), we can conclude that there is no need to make any changes in our data concerning the event of interest.

Finally, we randomly split our data to test and training with the use of a seed in order, if needed, to reproduce the results. For the training data, for the training data we used 70% of the observations, and for the test data the rest 30%.

With '**1**' we classify the observations that have more than 50% of the votes for Trump.
With '**0**' we classify the observations that have less than 50% of the votes for Trump.

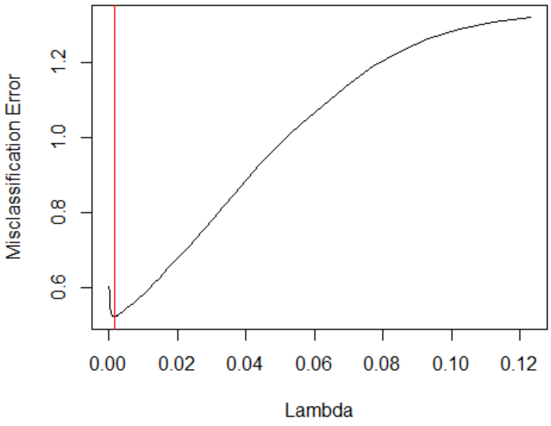
4.Part 1

4.1. Logistic Regression

The main goal of this paper is to predict using the same set of variables as those that were used in the previous paper. These variables contain information about demographic and economic characteristics for each county of the U.S.A for different time periods.

To find the appropriate variables we used lasso regression for binomial. We used the coefficients for the min lambda since it produces the smallest classification error, and we did find very good results in our test data. It is worth mentioning, because there is no cost function provided, that we will use the following hypothesis. If the probability of Trump to get more than 50% of the votes is larger than 50% then we classify these observations as the event of interest. This means that these observations are classified as 'Trump got more the 50% of the votes – has dominated'.

Graph 1: Misclassification error Vs Lambda Values



As can be seen from Graph 1 the smallest misclassification error has a lambda of 0.0015. This means that we will use the coefficients and their values for lambda equal to 0.0015.

Now that we have created our lasso model we will go straight to test the predictive capability. But before we start with the predictions, we will mention the metrics that we will use to judge the predictive capability of the models.

Classification error = (False Positive + False Negative) / Total Observations Used

Where, False Positives are the observations that the model decided that are positives but, they were Negatives and False Negative are the observation that the model decided that were Negative but, they were Positives.

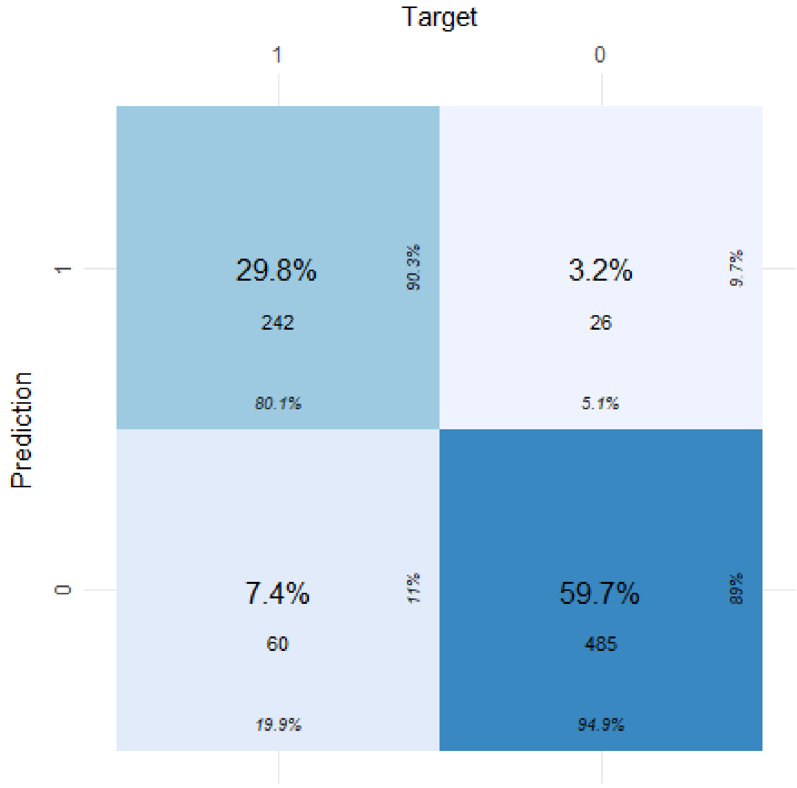
Model Accuracy = (Correct Predictions) / Total Observations Used

Now that we have mentioned the metric which will be used to judge the predictive capability of the models we can continue with our analysis. The following Graph is the confusion matrix for the test data set. It shows the combination of actual vs predicted values. Actual values(target) are the value that we have at our disposal from the test data set and

With '**1**' we classify the observations that have more than 50% of the votes for Trump.
With '**0**' we classify the observations that have less than 50% of the votes for Trump.

predicted are those that were produced by the model. It’s the most important Graph for classification models because it shows the predictive power of the model.

Graph 2: Confusion Matrix, logistic regression



The test data set contains 813 observations. From those 813 the 242 were correctly predicted as ‘1’ from the model, and the 485 were correctly predicted as ‘0’. On the other hand, 60 observations were falsely predicted as ‘0’(false negative) and 26 as ‘1’(false positive). The **accuracy** of the model is very good, 90% and the **misclassification error** is very low, 10%. This shows, that the predictive capability of our model is good and it should be used for predictions. Also, we can see that our model does a better job of predicting the ‘0’ not such a good job of predicting the ’1’. For the target = ‘1’ we can see that 19.9% observations have been falsely predicted as ‘0’. But, on the other hand, for the target = ‘0’ only 5.1% have been falsely predicted as 0.

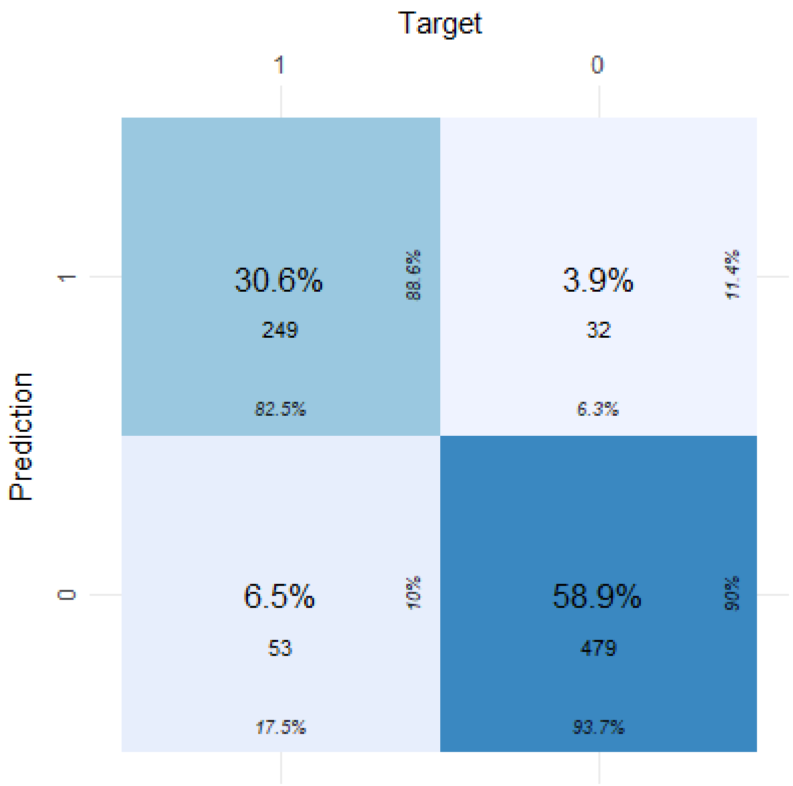
With ‘**1**’ we classify the observations that have more than 50% of the votes for Trump.
With ‘**0**’ we classify the observations that have less than 50% of the votes for Trump.

4.2. Random Forest

The second model that was created was a Random Forest. The random forest is a combination of many decision trees but it cannot have an interpretation such as a logistic regression or a decision tree. The random forest creates smaller decision trees and we use 'average decision' to calculate if the observation is of value '1' or '0'. For example, if we have 3 decision trees in our random forest and the two decision trees have set a particular observation to '1' and the third to '0' then this observation will take the value of '1'.

In our case we used 1000 decision trees in our random forest and we found an **accuracy,** approximately, of 90% and **misclassification error,** approximately**,** of 10%. This model has also great results for prediction purposes and should be used for prediction.

Graph 3: Confusion Matrix, Random Forest



As can be seen from Graph 3, from the accuracy and from the misclassification error the results are identical to the logistic regression. We can see that from the total observations only 53 observations have been falsely classified as '0'(false negative) and only 32 have been falsely classified as '1'(false positive). Also, we can see the same issue as the one that we encountered in the logistic regression. The model does a better job of predicting the '0' from the '1'.

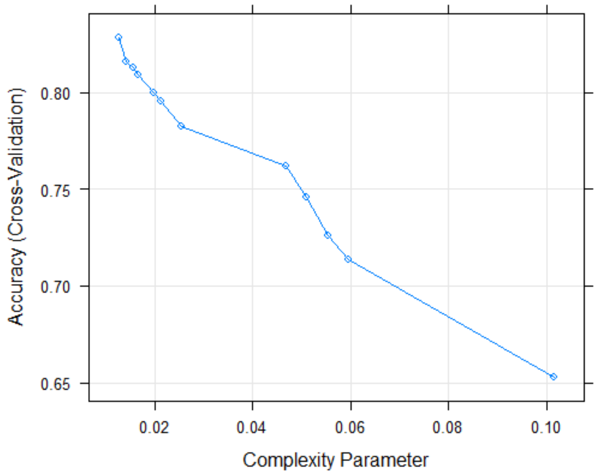
With '**1**' we classify the observations that have more than 50% of the votes for Trump.
With '**0**' we classify the observations that have less than 50% of the votes for Trump.

4.3. Decision Tree

The main advantage of the Decision Tree method is the interpretability and the lac of assumptions that comes with them. This makes them extremely easy to interpret to business users. The main idea of the decision tree method is that we split our data to nodes depending on information that is gained for each level of the tree.

In our case, we used a decision tree to create a model with the use of cross-validation. Also, we used 12 different complexity parameters to test the accuracy of the tree. Finally, a smaller complexity parameter value results in a larger tree with more nodes, which may lead to overfitting the training data.

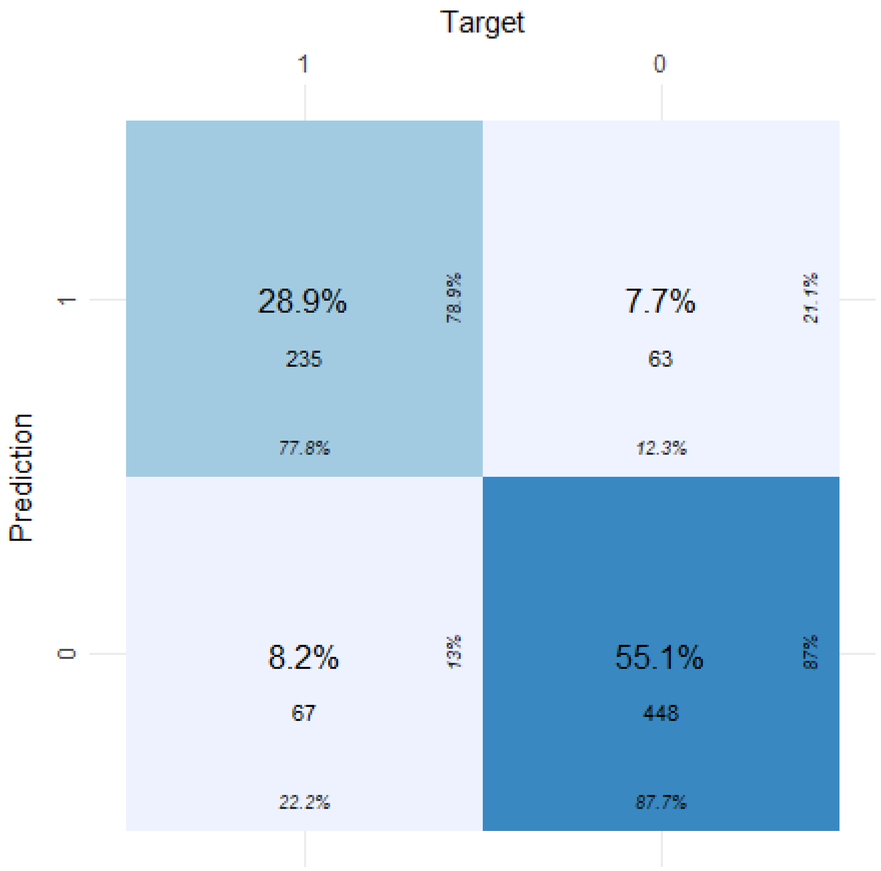
Graph 4: Accuracy Vs Complexity Parameter



Graph 4 shows us the decision tree accuracy and complex parameters for the 12 different combinations of hyperparameters. The complexity parameter is a very important metric. The higher the complexity parameter the more probable it is to have overfit the model. Also, a more complex tree (with more nodes) will be created with a high complexity parameter. In our case we can see that the best value of cp to be used is 0.0127 because in this value we have the largest accuracy for the cross validation. Now that we have found the best cp for our tree its time to test the accuracy of the model in the test data.

With ‘**1**’ we classify the observations that have more than 50% of the votes for Trump.
With ‘**0**’ we classify the observations that have less than 50% of the votes for Trump.

Graph 5: Confusion Matrix for decision tree



From graph 5 we can see that from the 813 total observations the 130 have been misclassified. More specifically, we found an **accuracy** of 84 % and a **misclassification error** is 16%. For the values that were '1' in the test data set the model has misclassified the 22.2% of them and 12.3% for the '0'.

5. Conclusions Part 1

For the 1st part of the paper, we created 3 different models and tested the predictive capability using the test data set. We found that the best models, according to the accuracy and misclassification error are the random forest and the lasso logistic regression for the min lambda. But all models had better predictive capability for predicting the '0' than the '1'. Perhaps, this issue can be resolved using oversampling because, as said before, 37% of the total observation are classified as '1' and the rest as '0'. So, by adding more 'fake' observations with them classified as '1' we may resolve this issue.

With '**1**' we classify the observations that have more than 50% of the votes for Trump.
With '**0**' we classify the observations that have less than 50% of the votes for Trump.

6. Part 2

6.1. Data manipulation

For this part we used only the county facts dataset and we split it according to the needs of the exercise. The first data frame contains the demographic data that will be used for the clustering and the second data frame contains the economic data that will be used for explaining, as best as can be done, the groups. Also, the groups are created from the clustering part.

6.2. Problems that were encountered.

As we will see later, 2 clustering models were created. The first one used the Euclidean distance to create the matrix and we used as linkage the ward. The second clustering model used the Manhattan distance to create the matrix and we used as linkage the ward.

For the first and second model the best linkage is the 'ward.D'. All other linkages that were tried-used (complete and average) gave results that were not logical. Such as creating a big cluster with all the observations and some smaller clusters with very few observations. This happened to various combinations of clusters. See the appendix for examples of this problem. Also, we have to mention that for those clustering models the average silhouette width is high, which is a good sign, but if the results of the clustering is not realistic then they cannot be used for our purposes.

6.3 Explaining the logic and metrics to be used for clustering.

To create a clustering model with the use of distances, we have to create a distance matrix. It is also, very important to scale the data especially when the variables that we have in our disposal have very different scales. This is something that we encountered in our dataset, so we scaled the data. This must be done because the algorithms use distances to calculate the results.

In the context of clustering with distance, there are many ways to calculate the distance matrix, such as Mahalanobis distance, Euclidean distance, Manhattan, and many others.

Another, important method that is used in cluster analysis is the linkage. The linkage determines how the distance between clusters is calculated when performing hierarchical clustering. We have different types of linkages such as simple, complete, average, centroid. E.g simple linkage, which calculates the distance between clusters from the closest observation of the one cluster with the closest observation of the other cluster.

Also, the silhouette width of an observation show if this observation has been assigned to correctly to a cluster. Overall, we want positive silhouette width for the observations and a high average silhouette width. The average silhouette width takes values from -1 to 1 and it can be calculated in total for all clusters and for each cluster independently. A good

With '**1**' we classify the observations that have more than 50% of the votes for Trump.
With '**0**' we classify the observations that have less than 50% of the votes for Trump.

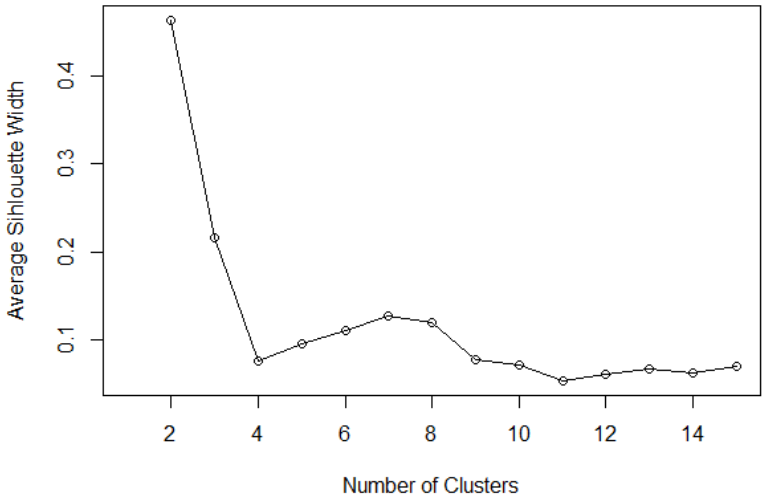
clustering should have high average silhouette width, but we have to take into account if the clusters that are created are realistic.

Finally, the 'height' that we see mainly in dendrograms shows the similarity or dissimilarity of the clusters. Overall, big 'jumps' on heights shows that the clusters that being merged are dissimilar and it is advisable not to merge them in this case.

6.4 Clustering with Euclidean Distance and ward linkage

As mentioned before, using different linkage such as average does not give us good results. The best linkage that was found is the ward linkage (ward.D in R). To decide the best clusters according to the Average Silhouette Width we created the following Graph:

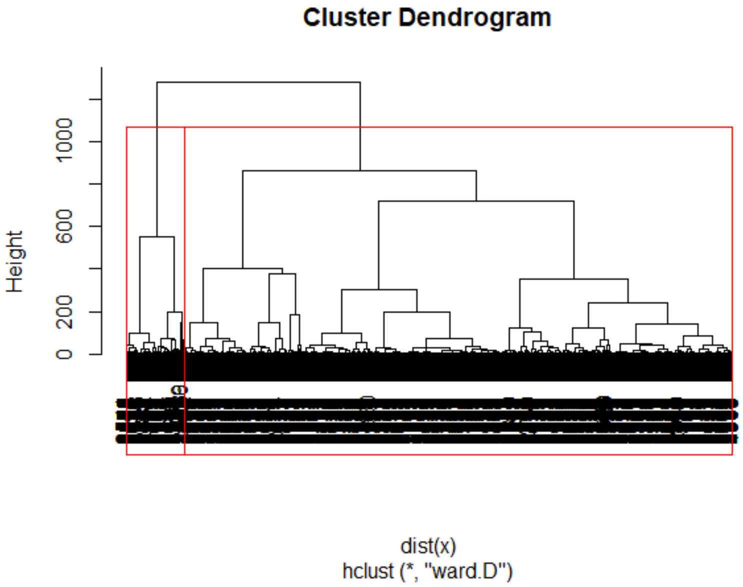
Graph 6: Average Sihlouette Width Vs Clusters



According to Graph 6, the optimal number of clusters when using the Euclidean distance and ward linkage is 2. Let's continue the analysis by showing the Cluster Dendrogram.

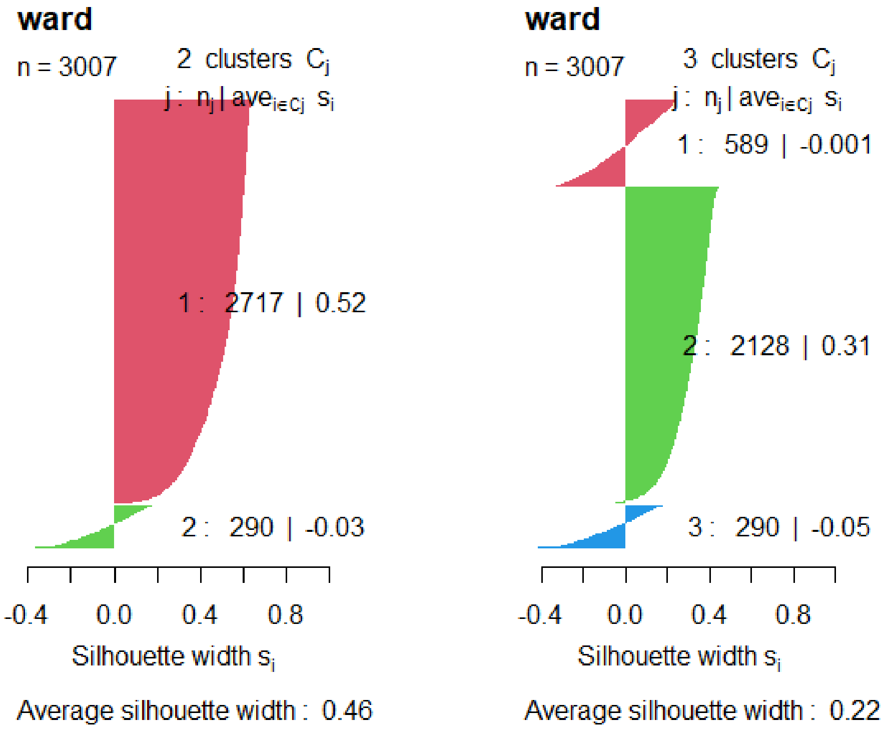
With '**1**' we classify the observations that have more than 50% of the votes for Trump.
With '**0**' we classify the observations that have less than 50% of the votes for Trump.

Graph 7: Cluster Dendrogram, where red lines are the 2 clusters



As we can see from Graph 7, the clusters have been created well and they are balanced. Also, they do not have any sudden 'spikes' on the heights. Let's continue our analysis with the Average silhouette width plot.

Graph 8: Silouete Width Graph



In graph 8, we can see the silhouette width for each observation for 2 clusters and for 3 clusters. In the right side we find the number of observations for each cluster and the average width for this cluster. We find some negative values for the observation, meaning

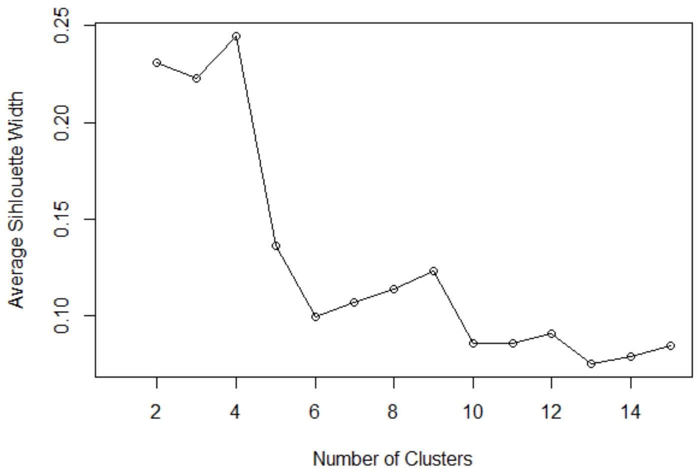
With '**1**' we classify the observations that have more than 50% of the votes for Trump.
With '**0**' we classify the observations that have less than 50% of the votes for Trump.

that they are not well classified. For both 2 and 3 clusters we can find for some clusters a negative average silhouette width.

6.5 Clustering with Manhattan distance and ward linkage

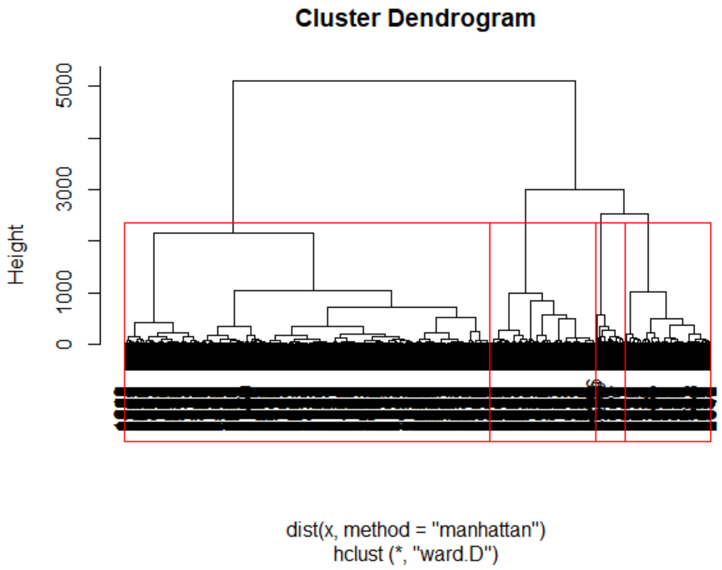
In this chapter we will cluster our data with the use of Manhattan distance and for the linkage we will the ward method. As mentioned before, by using some other linkage such as average the results are not logical so we will continue with the ward linkage.

Graph 9: Average Sihlouette Width Vs Clusters



According to the Average Silhouette Width, the optimal number of clusters is 4. But we will analyse the use of 2 and 3 clusters since they have high average width, in comparison to 5 or more clusters.

Graph 10: Cluster Dendrogram, where red lines are the 4 clusters



With '**1**' we classify the observations that have more than 50% of the votes for Trump.
With '**0**' we classify the observations that have less than 50% of the votes for Trump.

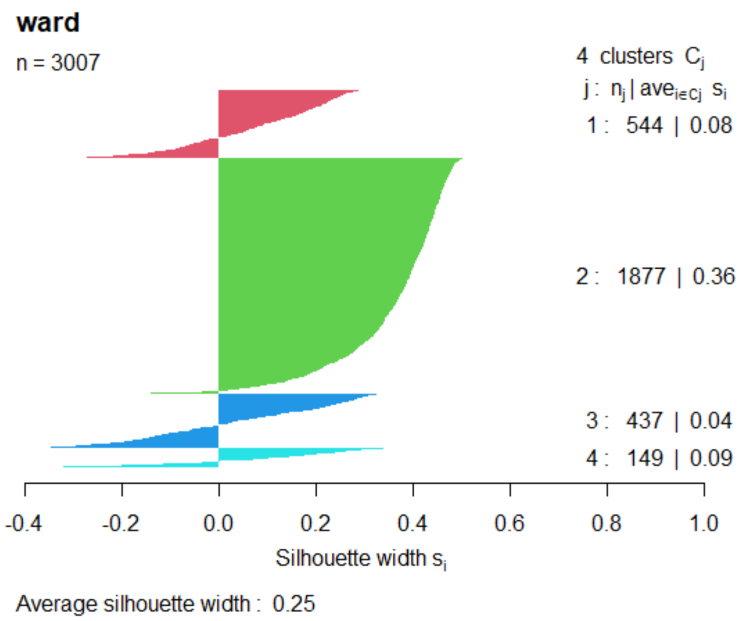
From Graph 10 we can see how the observations are clustered for each value of the height. We find the optimal number of clusters to be 4 but we could use 2 or 3 clusters since they all have close results (see silhouette plot in the appendix).

From Graph 11, in the Appendix. We find that the average silhouette width is the almost same for 2, 3 and 4 clusters. We can also see, that if we use 2 clusters we have many observations with negative silhouette width, so we will not use for sure 2 clusters.

6.7 Deciding the best clustering method.

Clustering in an unsupervised technique and it I not easy to find the best solution. As we saw in previous chapters many times, we have misclassified observations something that we would not want to happen in clustering. But we have to decide according to the silhouette width criterion which is the best clustering method and the best number of clusters. In all methods that we saw, we found some clusters with a negative average width, at least one. In more detail, at least one cluster had many observations misclassified. The method that will be used for clustering and explaining the economic data is the Manhattan distance with Ward linkage for 4 clusters. I prefer this method and this number of clusters because there are no negative values for any of the cluster that has been created. It might not have the higher overall Average Silhouette Width, but all other methods have misclassified many observations of at least one cluster.

Graph 12: Final Silhouette Width Graph



From Graph 12 we find 4 clusters. The first cluster contains 544 observations with some of them being misclassified, but the average silhouette width is positive for this group. For the second group, we find 1877 observations with a strong positive average silhouette width. For the third group, we find 437 observations, but we have some misclassification there. The same can be said for the fourth group which contains 149 observations.

With '**1**' we classify the observations that have more than 50% of the votes for Trump.
With '**0**' we classify the observations that have less than 50% of the votes for Trump.

7. Explaining the economic data with the use of the clusters

Our data set contains 18 economic related variables. Because of the time limit that we have in our disposal, we will analyse only the 5 of these variables for the 4 clusters that were created. The following variables are going to be analysed. Retail sales of 2007 in thousand units , total number of firms for 2007, the Accommodation and food services sales for 2007 in thousand units, total number of housing units for 2014, population per square mile 2010. Also, the first group contains 544 counties, the second 1877 counties, the third 437 counties and the fourth 149 counties. Let's begin by taking a look for the mean values of each group

Table 1: Average Values for each group

	Number of Counties	Retail sales	Accommodation and food service sales	Total number of firms	Total number of housing units	Population per square mile
Cluster 1	544	673,346	1,192	4,513	25,834	116
Cluster 2	1877	609,864	1,195	4,427	23,622	94
Cluster 3	437	677,637	1,243	4,701	24,544	64
Cluster 4	149	13,357,840	25,483	93,994	407,012	2,839

We can find some interesting facts in Table 1. The fourth cluster contains only 149 counties and, on average, it has the most intense economic activity. It shows huge Retail Sales and huge Accommodation and food service sales. Also, it has a lot of firms, and houses and it has very dense population. The average value is used mostly for the business users and it is very sensitive to extreme values. Also, we must take into consideration that the range of these variables is extremely big, so it is difficult for us to make a visual representation such as box plot or even a simple histogram. We will continue with the analysis of the median.

Table 2: Median Values for each group

	Number of Counties	Retail sales	Accommodation and food service sales	Total number of firms	Total number of housing units	Population per square mile
Cluster 1	544	204,407	420	1,778	11,166	46
Cluster 2	1877	210,610	495	1,997	11,574	43
Cluster 3	437	272,673	557	2,053	11,474	25
Cluster 4	149	10,007,613	18,930	66,632	306,954	1295

Interesting enough, we can find the same pattern while using the Median values for the 4 clusters. Always, the fourth cluster has a lot larger median value for all variables that are being analysed. As we can see the mean values for all the variables for all the clusters are always larger than the median values for the same clusters and the same variables. This shows as two possible scenarios. The first and most common, is that we have longer tails in the end but we might have some outliers that effect the distributions. Until this point, we have seen some small differences between the first 3 clusters and a big difference between the fourth

With '**1**' we classify the observations that have more than 50% of the votes for Trump.
With '**0**' we classify the observations that have less than 50% of the votes for Trump.

cluster with all the others. We can see that the most wealthy and more populated group is the fourth one. The third cluster is second in terms of wealth with the second group following. The least wealth group is the first group. But, there are no significant differences between the first three groups.

Now it's time to do the pairwise comparison with the use of pairwise Wilcoxon test. We will not include all the result here, but all the relative tables can be found in the Appendix section. In Wilcoxon test the null hypothesis assumes that the medians of the groups to be compared are equal. We reject the null hypothesis for a=5% when the p-value is less 5%(we choose the rejection level). According to the Wilcoxon test, we find that we have statistically significant evidence to support that the fourth class is different for all of the variables from all the other groups. In the Appendix, one can find those 149 counties which could be considered as more wealthy and more populated to analyse them more extensively.

8. Conclusions Part 2

In part 2 of the paper, we clustered with the use the demographic data and we found that appropriate method was the Manhattan distance with the Ward linkage for 4 clusters. We found that the first 3 clusters have similar economic related data but the fourth one was a lot different from the rest of the clusters. We could describe the fourth cluster as more wealthy and more populated, overall.

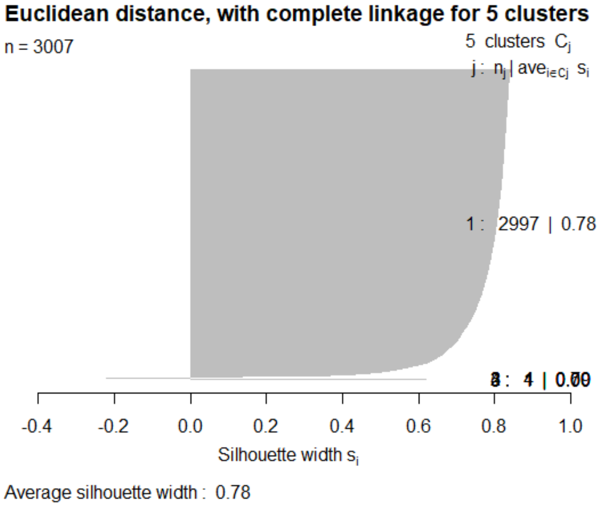
With '**1**' we classify the observations that have more than 50% of the votes for Trump.
With '**0**' we classify the observations that have less than 50% of the votes for Trump.

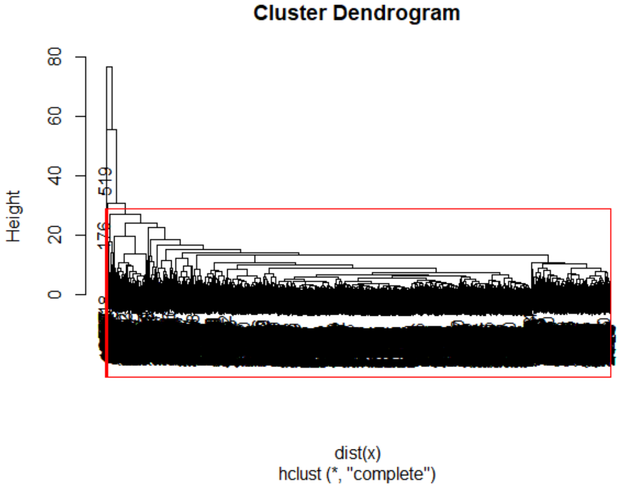
9. Appendix

1) Euclidean Distance with complete linkage:

Here we can see that there are 5 clusters. From the total 3193 observations the 3180 are contained in the first group and the rest 13 to the other 4 clusters. This result has no realistic meaning for us.

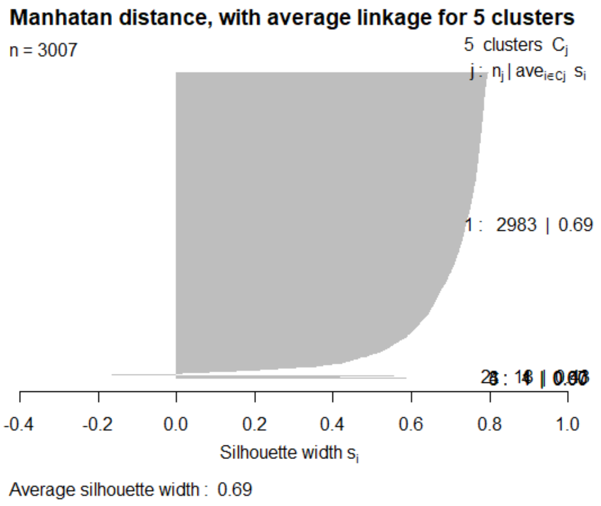
# Note: the same has been done for larger number of clusters with all of them giving the same 'weird' results.

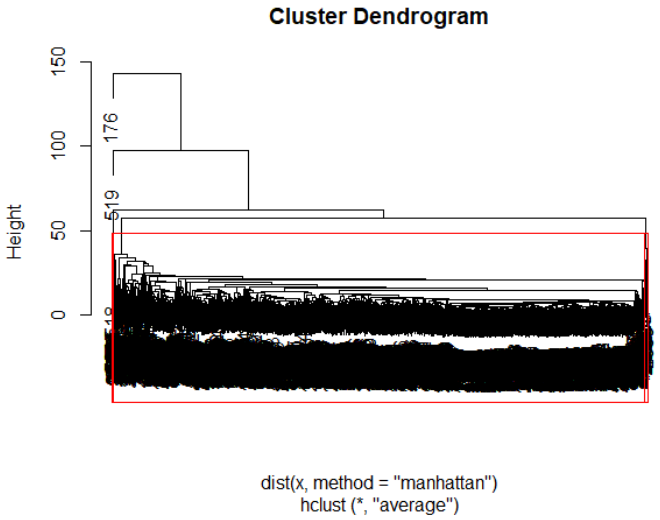




2) Manhattan Distance with average linkage:

With '**1**' we classify the observations that have more than 50% of the votes for Trump.
With '**0**' we classify the observations that have less than 50% of the votes for Trump.

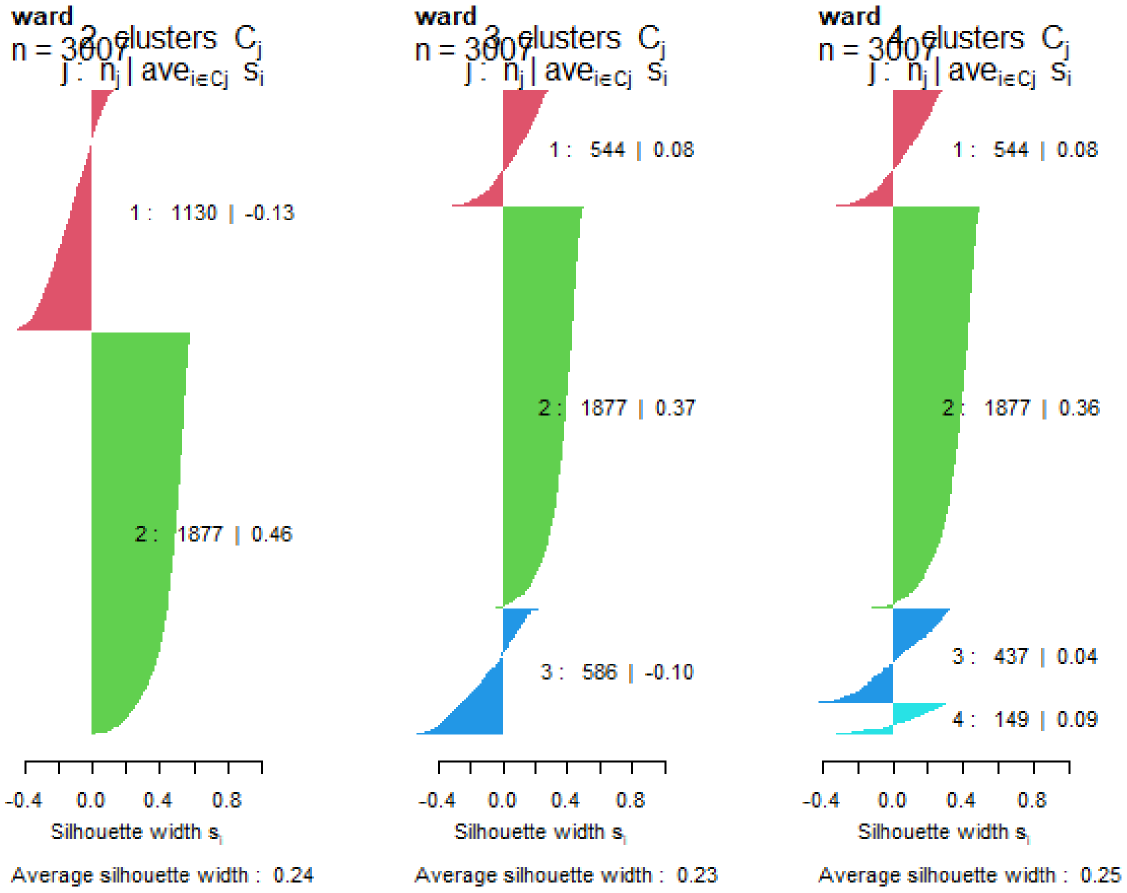




3) Clustering with Manhattan distance and ward linkage

Graph 11: Silouete Width Graph

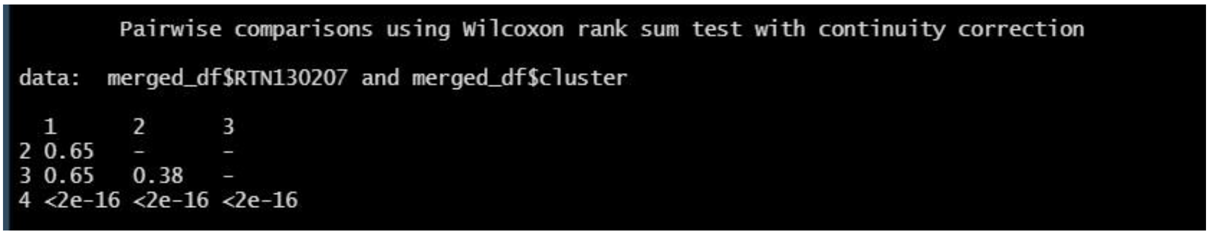
With ‘**1**’ we classify the observations that have more than 50% of the votes for Trump.
With ‘**0**’ we classify the observations that have less than 50% of the votes for Trump.



With ‘**1**’ we classify the observations that have more than 50% of the votes for Trump.
With ‘**0**’ we classify the observations that have less than 50% of the votes for Trump.

Pairwise Comparisons for the 4 Clusters:

retail sales

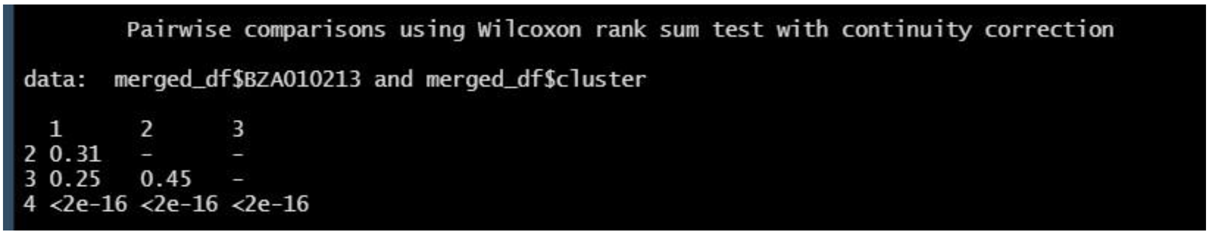


```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction 

data:  merged_df$RTN130207 and merged_df$cluster 

  1      2      3     
2 0.65   -      -     
3 0.65   0.38   -     
4 <2e-16 <2e-16 <2e-16
```

Accommodation and food services sales

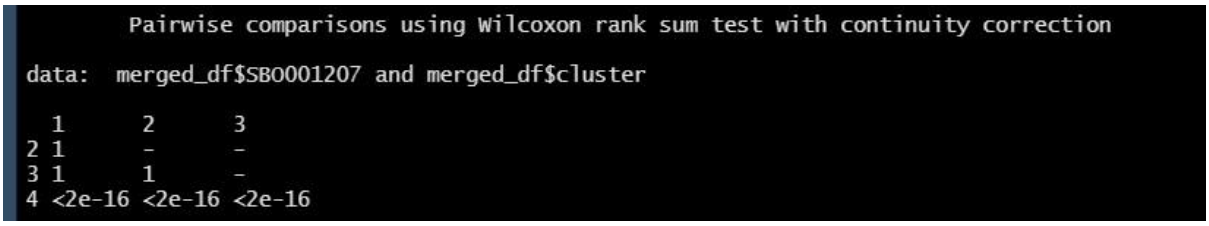


```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction 

data:  merged_df$BZA010213 and merged_df$cluster 

  1      2      3     
2 0.31   -      -     
3 0.25   0.45   -     
4 <2e-16 <2e-16 <2e-16
```

total number of firms

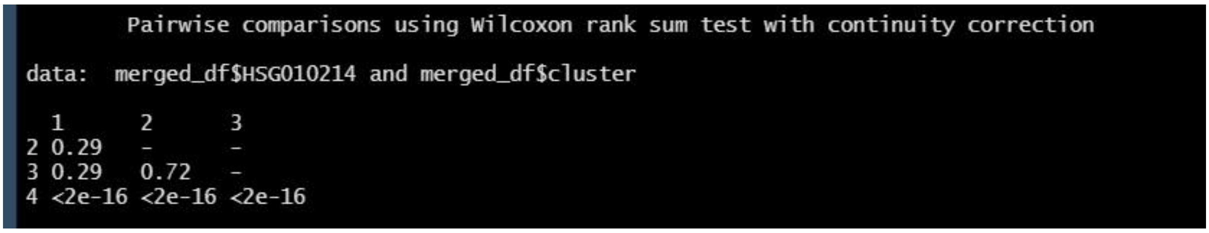


```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction 

data:  merged_df$SBO001207 and merged_df$cluster 

  1      2      3     
2 1      -      -     
3 1      1      -     
4 <2e-16 <2e-16 <2e-16
```

total number of Housing units

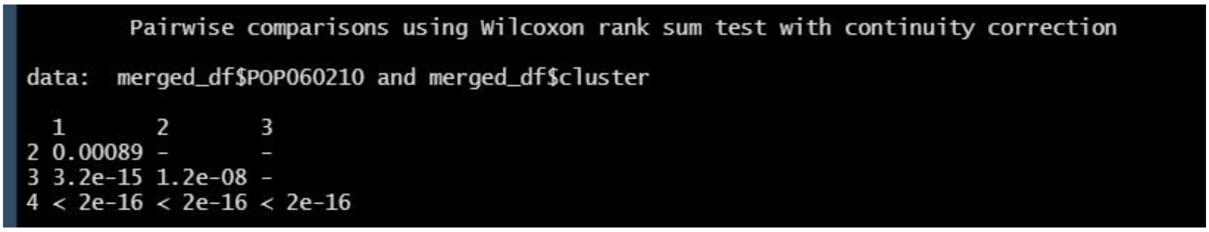


```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction 

data:  merged_df$HSG010214 and merged_df$cluster 

  1      2      3     
2 0.29   -      -     
3 0.29   0.72   -     
4 <2e-16 <2e-16 <2e-16
```

Population per square mile



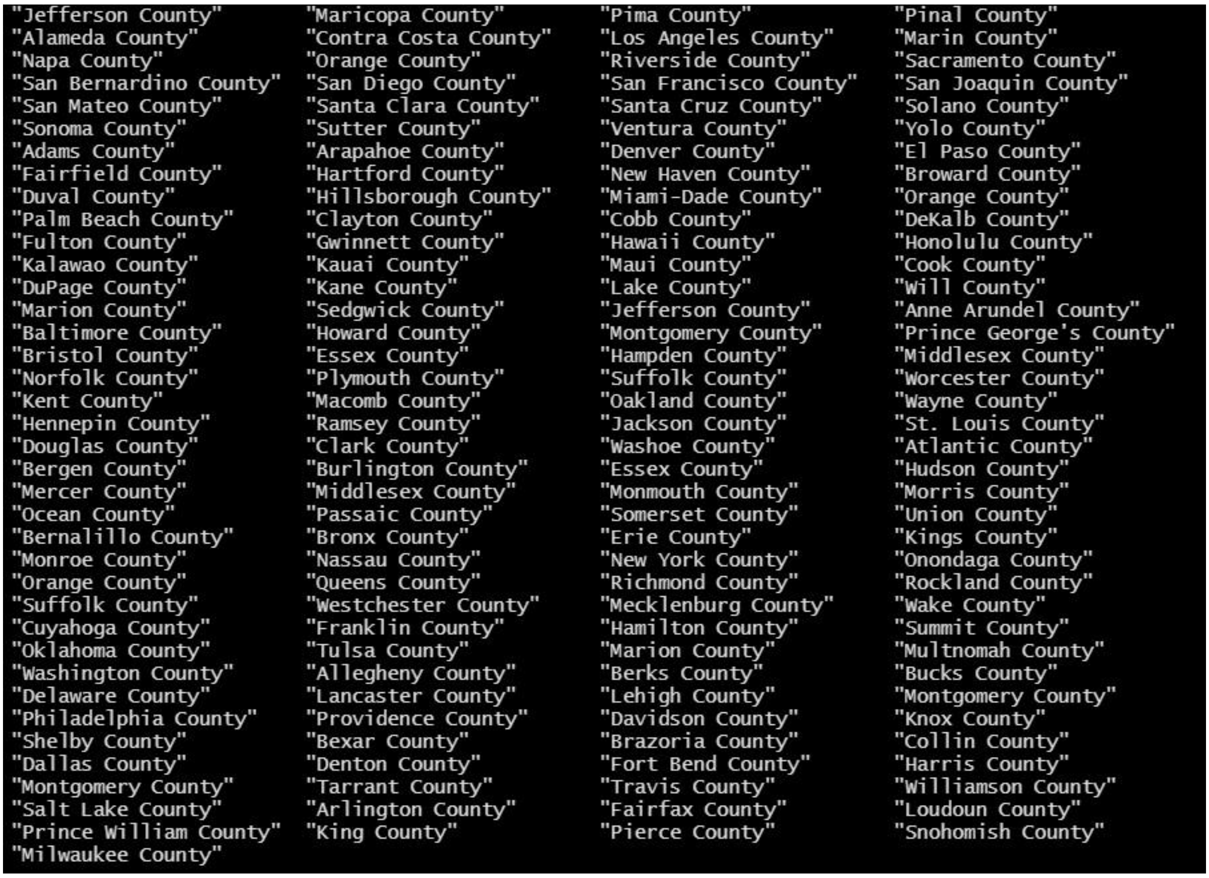
```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction 

data:  merged_df$POP060210 and merged_df$cluster 

  1       2       3      
2 0.00089 -       -      
3 3.2e-15 1.2e-08 -      
4 < 2e-16 < 2e-16 < 2e-16
```

With ‘**1**’ we classify the observations that have more than 50% of the votes for Trump.
With ‘**0**’ we classify the observations that have less than 50% of the votes for Trump.

Cluster 4 counties:



```
"Jefferson County"       "Maricopa County"        "Pima County"            "Pinal County"
"Alameda County"         "Contra Costa County"    "Los Angeles County"     "Marin County"
"Napa County"            "Orange County"          "Riverside County"       "Sacramento County"
"San Bernardino County"  "San Diego County"       "San Francisco County"   "San Joaquin County"
"San Mateo County"       "Santa Clara County"     "Santa Cruz County"      "Solano County"
"Sonoma County"          "Sutter County"          "Ventura County"         "Yolo County"
"Adams County"           "Arapahoe County"        "Denver County"          "El Paso County"
"Fairfield County"       "Hartford County"        "New Haven County"       "Broward County"
"Duval County"           "Hillsborough County"    "Miami-Dade County"      "Orange County"
"Palm Beach County"      "Clayton County"         "Cobb County"            "DeKalb County"
"Fulton County"          "Gwinnett County"        "Hawaii County"          "Honolulu County"
"Kalawao County"         "Kauai County"           "Maui County"            "Cook County"
"DuPage County"          "Kane County"            "Lake County"            "Will County"
"Marion County"          "Sedgwick County"        "Jefferson County"       "Anne Arundel County"
"Baltimore County"       "Howard County"          "Montgomery County"      "Prince George's County"
"Bristol County"         "Essex County"           "Hampden County"         "Middlesex County"
"Norfolk County"         "Plymouth County"        "Suffolk County"         "Worcester County"
"Kent County"            "Macomb County"          "Oakland County"         "Wayne County"
"Hennepin County"        "Ramsey County"          "Jackson County"         "St. Louis County"
"Douglas County"         "Clark County"           "Washoe County"          "Atlantic County"
"Bergen County"          "Burlington County"      "Essex County"           "Hudson County"
"Mercer County"          "Middlesex County"       "Monmouth County"        "Morris County"
"Ocean County"           "Passaic County"         "Somerset County"        "Union County"
"Bernalillo County"      "Bronx County"           "Erie County"            "Kings County"
"Monroe County"          "Nassau County"          "New York County"        "Onondaga County"
"Orange County"          "Queens County"          "Richmond County"        "Rockland County"
"Suffolk County"         "Westchester County"     "Mecklenburg County"     "Wake County"
"Cuyahoga County"        "Franklin County"        "Hamilton County"        "Summit County"
"Oklahoma County"        "Tulsa County"           "Marion County"          "Multnomah County"
"Washington County"      "Allegheny County"       "Berks County"           "Bucks County"
"Delaware County"        "Lancaster County"       "Lehigh County"          "Montgomery County"
"Philadelphia County"    "Providence County"      "Davidson County"        "Knox County"
"Shelby County"          "Bexar County"           "Brazoria County"        "Collin County"
"Dallas County"          "Denton County"          "Fort Bend County"       "Harris County"
"Montgomery County"      "Tarrant County"         "Travis County"          "Williamson County"
"Salt Lake County"       "Arlington County"       "Fairfax County"         "Loudoun County"
"Prince William County"  "King County"            "Pierce County"          "Snohomish County"
"Milwaukee County"
```

With ‘**1**’ we classify the observations that have more than 50% of the votes for Trump.
With ‘**0**’ we classify the observations that have less than 50% of the votes for Trump.