CS 418 Project 02 Report

1. We partitioned the given merged dataset into a training set and a validation test using the holdout method. We kept the training set size to be 75% and the size of the test set is 25%.

2. We used StandardScaler and scaler.transform to standardize the training set and the validation set.

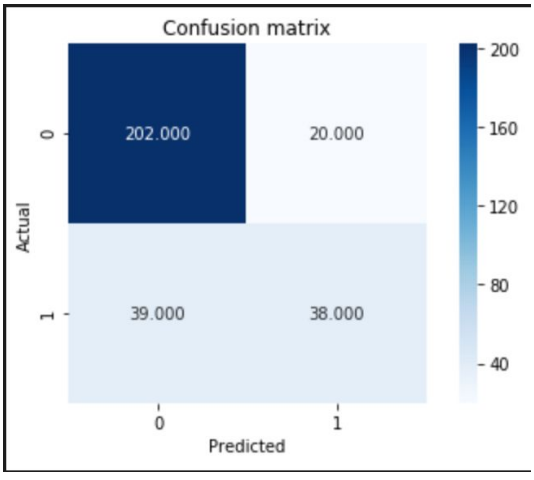
3. The best performing linear regression model is where we consider the columns 'Percent Age 29 and Under', 'Percent Age 65 and Older', 'Total Population', 'Percent Foreign Born', 'Percent Hispanic or Latino',' Percent Black, not Hispanic or Latino', 'Percent White, not Hispanic or Latino', 'Percent Female' and 'Percent Unemployed' for building the model.

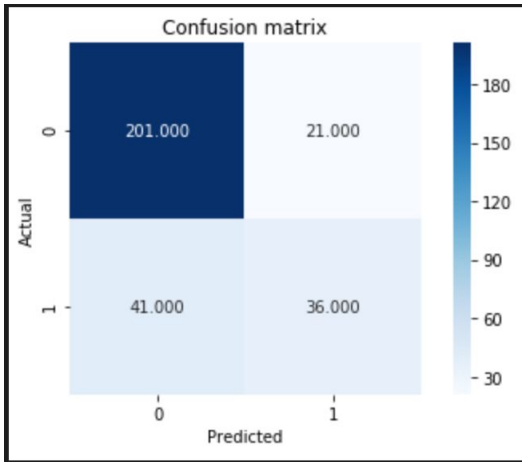
   The score of this model is 87.27% for Democratic.

   All the other columns which are 'Median Household Income', 'Percent Rural', 'Percent Less than Bachelor\'s Degree' and 'Percent Less than High School Degree' are decreasing the score, hence we do not take them as predictor variables. By trial and error, we were able to select the variables for building this model.

   The score of the model is 84.71% for Republicans.

4. The best performing classification model is Support Vector Machines using all variables. The F1_score for the model where we use all variables in SVM is [0.9094, 0.6504]. We selected the parameters following the general steps by trying different numbers of combinations of the number of variables and arguments like k for k-nearest neighbors, kernels for SVM.





Some of the confusion matrixes, the others can be seen on the pdf for the code.

5. We tried Single Linkage Hierarchical Clustering with variables being ‘Percent White, not Hispanic or Latino’, ‘Percent Black, not Hispanic or Latino’ and ‘Percent Hispanic or Latino’.

   Adjusted_rand_index = 0.0071

   Silhouette Coefficient = 0.8360

   We tried Single Linkage Hierarchical Clustering with variables being ‘Percent White, not Hispanic or Latino’, ‘Percent Black, not Hispanic or Latino’ and ‘Percent Hispanic or Latino’, ‘Percent Age 29 and Under’, ‘Percent Age 65 and Older’, ‘Percent Less than High School Degree’, ‘Percent Less than Bachelor\s’ Degree’.

   Adjusted_rand_index = 0.0071

   Silhouette Coefficient = 0.7392

   We tried Complete Linkage Hierarchical Clustering with variables being ‘Percent White, not Hispanic or Latino’, ‘Percent Black, not Hispanic or Latino’ and ‘Percent Hispanic or Latino’.

   Adjusted_rand_index = 0.0396

   Silhouette Coefficient = 0.7463

   We tried Complete Linkage Hierarchical Clustering with variables being ‘Percent White, not Hispanic or Latino’, ‘Percent Black, not Hispanic or Latino’ and ‘Percent Hispanic or Latino’, ‘Percent Age 29 and Under’, ‘Percent Age 65 and Older’, ‘Percent Less than High School Degree’, ‘Percent Less than Bachelor\s’ Degree’.

   Adjusted_rand_index = 0.0180

   Silhouette Coefficient = 0.6977

   We tried K-Means with 5 clusters, 10 iterations using variables being ‘Percent White, not Hispanic or Latino’, ‘Percent Black, not Hispanic or Latino’ and ‘Percent Hispanic or Latino’.

   Adjusted_rand_index = 0.0481

   Silhouette Coefficient = 0.4262

   We tried K-Means with 5 clusters, 10 iterations using variables being ‘Percent White, not Hispanic or Latino’, ‘Percent Black, not Hispanic or Latino’ and ‘Percent Hispanic or Latino’, ‘Percent Age 29 and Under’, ‘Percent Age 65 and Older’, ‘Percent Less than High School Degree’, ‘Percent Less than Bachelor\s’ Degree’.

   Adjusted_rand_index = 0.0408

Silhouette Coefficient = 0.1966

We tried K-Means with 10 clusters, 25 iterations using variables being ‘Percent White, not Hispanic or Latino’, ‘Percent Black, not Hispanic or Latino’ and ‘Percent Hispanic or Latino’.

Adjusted_rand_index = 0.0341

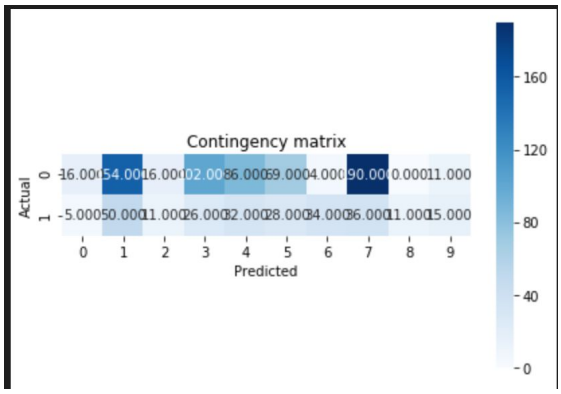
Silhouette Coefficient = 0.4097

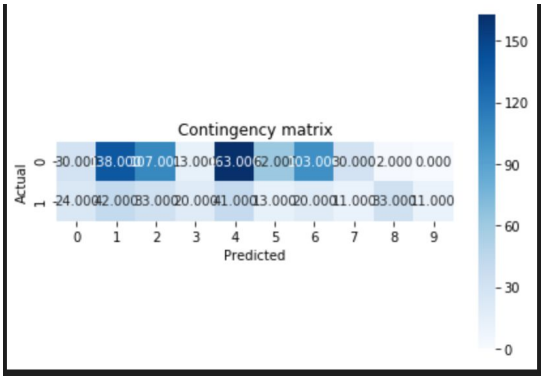
We tried K-Means with 10 clusters, 25 iterations using variables being ‘Percent White, not Hispanic or Latino’, ‘Percent Black, not Hispanic or Latino’ and ‘Percent Hispanic or Latino’, ‘Percent Age 29 and Under’, ‘Percent Age 65 and Older’, ‘Percent Less than High School Degree’, ‘Percent Less than Bachelor\s’ Degree’.

Adjusted_rand_index = 0.0406

Silhouette Coefficient = 0.2236

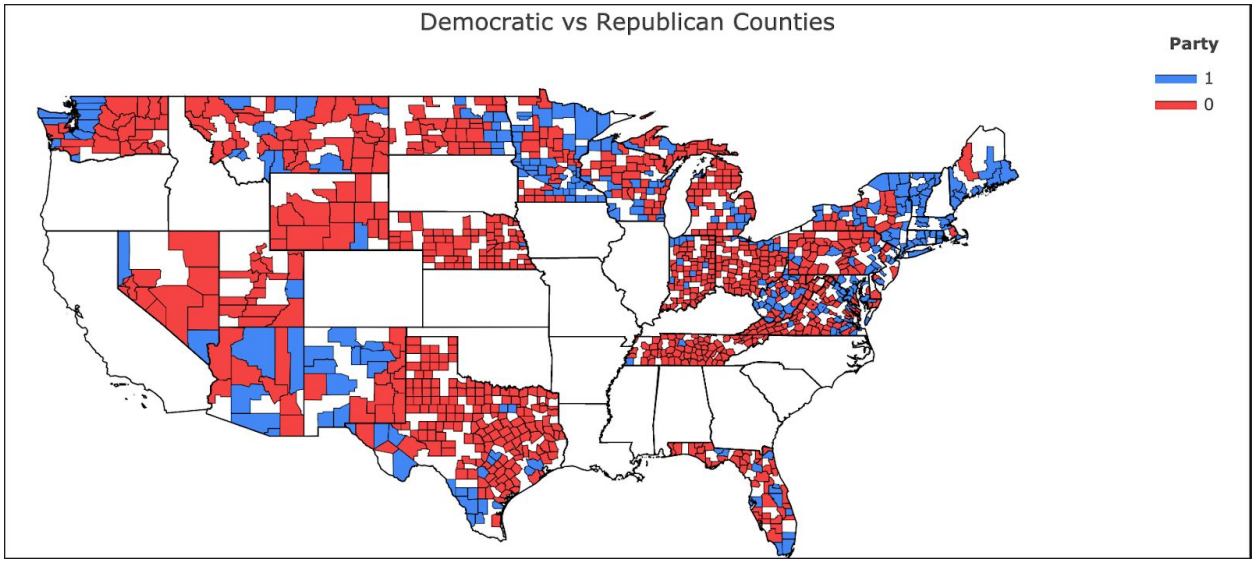
As none of the models give a considerable performance, none of them can be taken as a good performance clustering model.



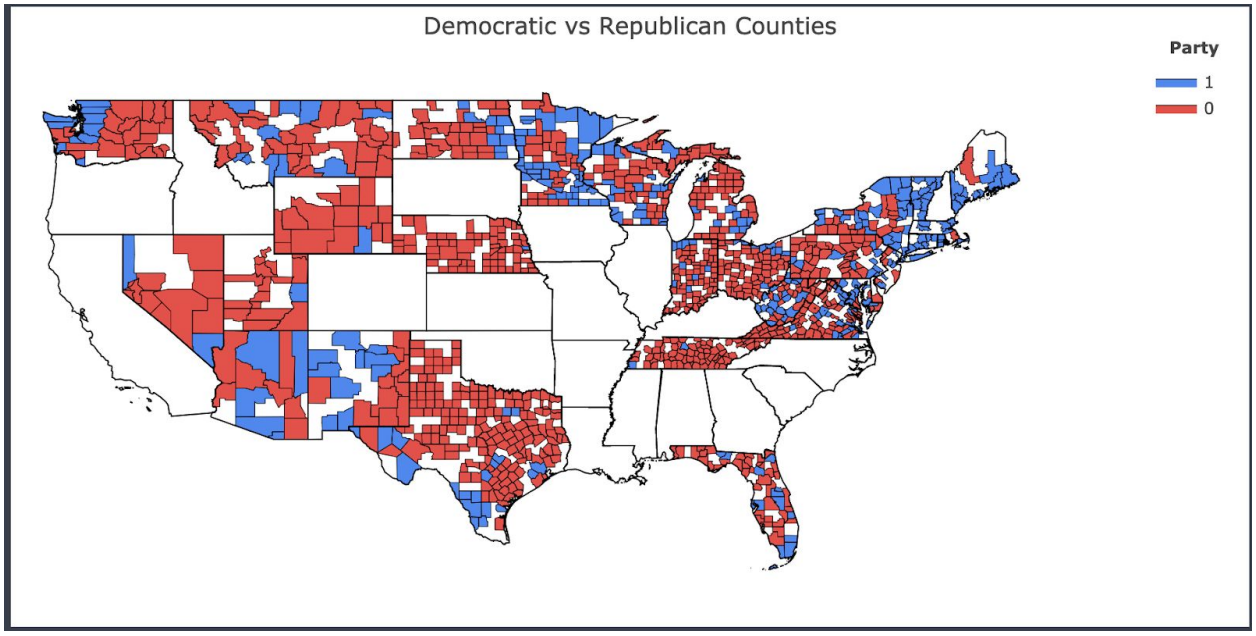


Some of the matrixes from task 5, the others can be seen on the report for the code.

6. Project 2 Graph



Project 1 Graph



The Conclusion that can be made from the two maps is that it pretty much stayed the same and there were only minimal changes in the number of votes count which is a very little thing to see on the map, as even after scaling it.
There was one republican county added in Wisconsin on the border of Illinois and Wisconsin

7. Look at the output file called output.csv

output

State	County	Democratic	Republican	Party
NV	eureka	6554	7835	0
TX	zavala	0	4878	1
VA	king george	7931	17140	0
OH	hamilton	171680	112207	1
TX	austin	8270	4562	0
MI	barry	11117	14116	0
NM	valencia	0	18142	1
TX	ellis	32419	28087	0
NJ	mercer	104044	54625	1
PA	cambria	27288	29167	0
IN	switzerland	0	0	0
NV	lander	0	15806	0
NE	cherry	4690	7242	0
VA	radford city	0	2116	1
FL	lee	151373	98789	0
MI	arenac	1314	3554	0
TX	shackelford	0	1996	0

Sample from Output.csv