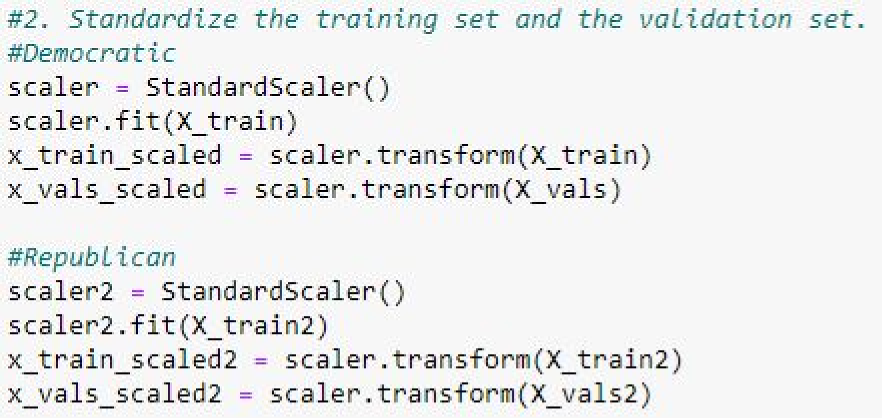
1. **Partition the merged dataset into a training set and a validation set using the holdout method or the cross-validation method. How did you partition the dataset?** We partitioned the dataset using the holdout method. We did that by creating a training set, validation set, and the holdout set a.k.a the ‘test’ set. The common ratio we used to split the data was 75:25.

2. **(5 pts.) Standardize the training set and the validation set** We standardized each partition, one for democrats and the other for republicans.



```
#2. Standardize the training set and the validation set.
#Democratic
scaler = StandardScaler()
scaler.fit(X_train)
x_train_scaled = scaler.transform(X_train)
x_vals_scaled = scaler.transform(X_vals)

#Republican
scaler2 = StandardScaler()
scaler2.fit(X_train2)
x_train_scaled2 = scaler.transform(X_train2)
x_vals_scaled2 = scaler.transform(X_vals2)
```

3. **(25 pts.) Build a linear regression model to predict the number of votes cast for the Democratic party in each county. Consider multiple combinations of predictor variables. Compute evaluation metrics for the validation set and report your results. What is the best performing linear regression model? What is the performance of the model? How did you select the variables of the model? Do the same for Republicans.**

For democrats, the best performing linear regression model had an R^2 value of 0.9514, and it contained just 3 variables. The variables were ‘Total Population’, ‘Percent Age 29 and Under’, and ‘Percent Unemployed’. These variables produced the highest value R^2 score out of all the combinations of variables we tried. The performance of the model is indicated by the R^2 value, which again was 0.9514. Given this score, we can argue the model fits the data well. This also entails about 95% of our variance is being explained by our model. The 3 variables mentioned above are solid predictors for predicting the number of votes cast for the Democratic party in each county. We selected these variables through trial and error and good use of intuition.

For republicans, the best performing linear regression model had an R^2 value of 0.51562, and it contained just 8 variables. The variables were ‘Total Population', 'Percent White, not Hispanic or Latino', 'Percent Female', 'Percent Age 29 and Under', 'Median Household Income', 'Percent Unemployed', "Percent Less than Bachelor's Degree", 'Percent Rural’. These variables produced the highest value R^2 score out of all the combinations of variables we tried. The performance of the model is indicated by the R^2 value, which again was 0.51562. Given this score, we can argue the model fits the data okay. The R^2 value isn't as high as we would like it to be. This entails about 52% of our variance is being explained by our model. The 8 variables mentioned above are ok predictors for predicting the number of votes cast for the Republican party in each county. We selected these variables through trial and error and good use of intuition. The results for Republican are not as strong as the democratic ones, which is most likely due to the fact that we have less data for democratic counties than republican counties.

4. **(25 pts.) Build a classification model to classify each county as Democratic or Republican. Consider at least two different classification techniques with multiple combinations of parameters and multiple combinations of variables. Compute evaluation metrics for the validation set and report your results. What is the best performing classification model? What is the performance of the model? How did you select the parameters of the model? How did you select the variables of the model?**

   The results Decision tree we got for the F1 score was 0.875 for Republicans and 0.625 for Democrats. We see that the precision and the recall for the Republicans and Democrats provided the most accurate information out of all the other models created. The metrics for Decision Tree error were the lowest value compared to other models, which was 0.1875 compared to the other models that were around 0.200 or more in errors. The one with the most errors was Naives Bayes with a measurement of 0.2545. Although K-Nearest Neighbors and SVM were the best models, SVM had a problem with its F1 Score result for Democrats because it was extremely low compared to others because of the inaccurate results. Thus, making the model Decision Tree a better choice.

   We focused on all four classification techniques, so we used Decision Trees, K-Nearest Neighbors, Naive Bayes, and Support Vector Machine. We first created and observed the decision tree and got the results that the predictions were majorly accurate for republicans, but lacked prediction for democrats.

   The best model that we had is the one where we used decision trees. The decision tree model contained much more accurate data than the rest of the models. The accuracy was at 0.8125 while the others only had at least 0.79 accuracy. The precision was higher for both parties with Decision Trees results being at 0.8352. We first partitioned the party and standardized the data and scale X_train3 and X_vals3. We went through each column and saw what gave us the best score to work, so that it would be the most accurate. Thus, we came up with 7 variables.

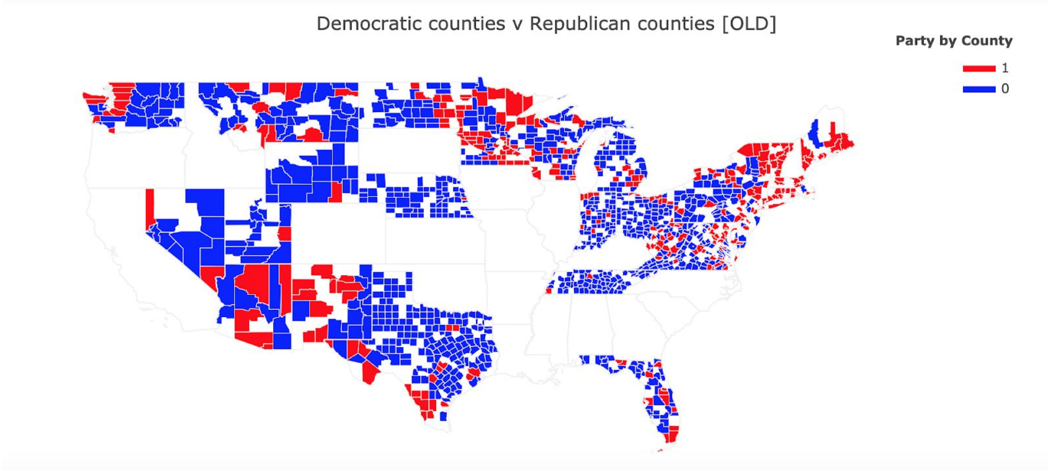
5. **(25 pts.) Build a clustering model to cluster the counties. Consider at least two different clustering techniques with multiple combinations of parameters and**

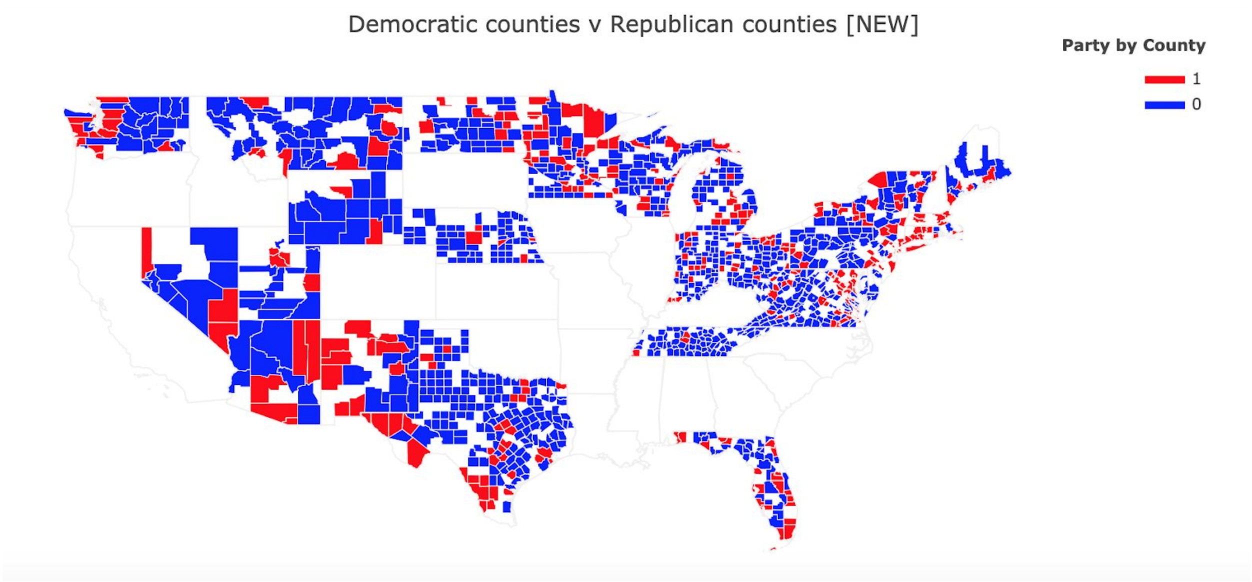
multiple combinations of variables. Compute unsupervised and supervised evaluation metrics for the validation set with the party of the counties (Democratic or Republican) as the true cluster and report your results. What is the best performing clustering model? What is the performance of the model? How did you select the parameters of the model? How did you select the variables of the model?

The results of the clustering model using single linkage found a low adjusted rand index of 0.0056, and the silhouette coefficient for this model was 0.95. The K-means clustering model found a higher adjusted rand index of 0.174, and the silhouette coefficient was 0.87. The results of the DBSCAN clustering are identical to the clusters found using hierarchical clustering with single linkage, with an adjusted rand index of 0.0056 and the silhouette coefficient 0f 0.95.

The best performing clustering model was hierarchical clustering with single linkage and DBSCAN clustering, because the silhouette coefficient was higher. This means that the clusters are better separated and more cohesive than the clusters found using K-means. To determine the parameters of the model we partitioned the Party and standardized the data, then we tested all variables and examined each variable individually to determine which clustering model would give us the best result. We found that the total population gave the best results for the clustering models.

6. **(10 pts.) Create a map of Democratic counties and Republican counties using the counties' FIPS codes and Python's Plotly library (plot.ly/python/county-choropleth/). Compare with the map of Democratic counties and Republican counties created in Project 01. What conclusions do you make from the plots?**

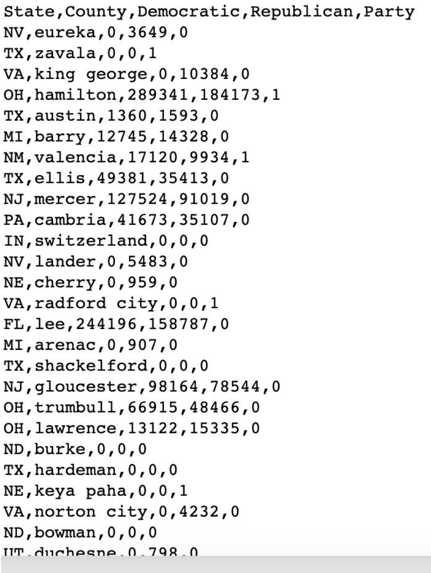




Our best classification model used Decision Trees as the classification technique to produce it, along with 7 variables 'Total Population', 'Percent White, not Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Unemployed', 'Percent Less than High School Degree', and 'Percent Rural'. Compared to the map of Democratic counties and Republican counties created in Project 01, our model did a pretty good job at predicting Democratic counties. Our model pinpointed which areas were Democratic just right. Compared to the old map, the areas almost align. However, the dataset itself does lack a lot of data on Democratic counties. We saw this from the map we produced in Project 01. The amount of Republican counties overshadowed the amount of Democratic counties. This helps explain why our F1-scores for our best model were great for Republican counties (0.875), but ok for Democratic counties (0.625). This is also the case regardless of the technique used to make the model. As such, it makes sense why we missed the mark on some counties. Overall, however, our model did a really good job at matching the old map.

7. **(5 pts.) Use your best performing regression and classification models to predict the number of votes cast for the Democratic party in each county, the number of votes cast for the Republican party in each county, and the party (Democratic or Republican) of each county for the test dataset (demographics_test.csv). Save the output in a single CSV file. For the expected format of the output, see sample_output.csv.**

   We used the best models to predict the number of votes as well as the rightful classification. Below is a snippet of our results. For the full results, consult the output.csv file in our project folder.



```
State,County,Democratic,Republican,Party
NV,eureka,0,3649,0
TX,zavala,0,0,1
VA,king george,0,10384,0
OH,hamilton,289341,184173,1
TX,austin,1360,1593,0
MI,barry,12745,14328,0
NM,valencia,17120,9934,1
TX,ellis,49381,35413,0
NJ,mercer,127524,91019,0
PA,cambria,41673,35107,0
IN,switzerland,0,0,0
NV,lander,0,5483,0
NE,cherry,0,959,0
VA,radford city,0,0,1
FL,lee,244196,158787,0
MI,arenac,0,907,0
TX,shackelford,0,0,0
NJ,gloucester,98164,78544,0
OH,trumbull,66915,48466,0
OH,lawrence,13122,15335,0
ND,burke,0,0,0
TX,hardeman,0,0,0
NE,keya paha,0,0,1
VA,norton city,0,4232,0
ND,bowman,0,0,0
UT,duchesne,0,798,0
```