



IS 733 DATA MINING

Final Project Report

**Predicting Relations Between
the Factors Involved In
Gerrymandering**

Members

**Abishek Varadarajan
Aditi Sharma
Asmita Ranasinghe
Rashmi Gangadariah
Savio Kay
Tejaswini Yella**

Mentor

Dr. James Foulds

Topic	Page No
Introduction	3
Motivation	3
Data Collection	3
Data pre-processing	4
Implementation	5
Analysis	6
Prediction	12
Future Scope	12
Conclusion	13
References	13

1. Introduction

You would believe that one of the largest democracy in the world would conduct their elections and elections proceedings with integrity and fair opportunity agenda in mind. But that's not the case with the American congressional elections which is criticised to be heavily profiting a single party, usually the one currently in power. Here, the party currently in power abuse their power and privileges bestowed to them by the people who got them elected, to make the election favour to their side by redrawing district boundary lines. This unjust process of redistricting so as to secure an unjust advantage to a certain party challenging another for the election is commonly known as Gerrymandering. This term was first termed in the Boston Gazette on 26 March 1812 to the reaction of redrawing district lines. Gerrymandering as a concept was affirmed in 1812 Boston Congressional primaries where such an act as filed. In that election, the Governor Elbridge Gerry signed a bill which allowed redrawing district lines for that particular senate election. The word Gerrymandering is a portmanteau of the Governor's last name and the word salamander. It was originally written as 'Gerry-mander'. There are many factors and aspect to which this process builds up to, some of which may have greater influence over the other for redrawing district lines. In this study, we try to examine the various factors involved with gerrymandering. By having a better perception of these factors we can predict and make moves to decrease the effect of it.

2. Motivation

With factors like population, votes, party affiliation, type of location, the income of living, we wanted to see the association between these factors and many other factors with each other. With certain running assumptions about population and party-race affiliations, we wanted to analyse information throughout eight congressional elections to determine how they link and affect the decision of gerrymandering. We believe the above-mentioned factors can have major correlations between each other and this analysis tries to delve into this in a much more deeper sense.

3. Data Collection

Data Collection refers to the process of collecting data from one or various sources for the purpose of analysis. For this project, we have collected data from multiple sources as follows

1. U.S. Census Bureau :
<https://www.census.gov/programs-surveys/popest/data/tables.All.html>
From here, we collected the population of each race from every county for the two states from analysis(Maryland, North Carolina) for the years 2002 to 2009.
Data Format: HTML Object(Bureau, 2018).
2. U.S. Census Bureau :
https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bk_mnk
From here, we collected the population of each race from every county for the two states from analysis(Maryland, North Carolina) for the years 2010 to 2016. Data Format: HTML Object(DADS), 2018).

3. History, Art & Archives, United States House of Representatives : <http://history.house.gov/Institution/Election-Statistics/Election-Statistics/>
This is the website from where we had retrieved the congressional election votes for each party in every district for the two states, Maryland and North Carolina, which we have considered in this project for analysis and prediction. Data Format: Unorganized PDF("Election Statistics, 1920 to Present | US House of Representatives: History, Art & Archives", 2018).
4. District Shapefiles : https://www.census.gov/geo/maps-data/data/cbf/cbf_cds.html
Shapefiles for all the congressional elections. From which we were able to retrieve all the zip codes falling under each election. Data Format: Shapefiles(Branch, 2018).
5. Urban Area Shapefiles : https://www.census.gov/geo/maps-data/data/cbf/cbf_ua.html
Shapefiles for all the election years. From which we were able to retrieve all the urban area zip codes falling under each district. Data Format: Shapefiles("Cartographic Boundary Shapefiles - Urban Areas - Geography - U.S. Census Bureau", 2018).
6. Kaggle : <https://www.kaggle.com/laa283/evidence-of-gerrymandering/data>
This is from where we collectively got an idea of the types of factors we can consider for the analysis("Evidence of Gerrymandering | Kaggle", 2018).

4. Data Pre-processing

Data Pre-processing helps in resolving issues that relate to data that is available from various places that are often incomplete or inconsistent. It is a data mining technique that helps in transforming the raw real-world data into an understandable format that also helps in reducing errors further during analysis or prediction.("Data pre-processing", 2018)

During the data pre-processing stage we faced a lot of challenges, as the data that we required for analysis and prediction for this project was not readily available and added to that the information/data we retrieve was not available to us in an easily usable format. We performed various steps to form the structured data that we now have used for this project as follows:

1. We had to extract the population of each race for every county which was available in different columns and calculate the race population for every district in each state. While converting from count level to district, we had to consider the situations where a county can fall under multiple districts. For that case we decided to split the county against each district and calculate the mean of the population and substitute it for that county under each district.
2. Certain county and district data contained multiple districts for several counties (E.g. Baltimore County, 1-3,7). While preparing the dataset, a few of them were imported as a Date, for e.g. '01-03' county became date as '03-January'. To solve that we used level function to replace the county/district.
3. The election votes data we collected was available in an unorganized PDF file, from where we had to extract the required data into the final CSV(Comma Separated Value) file. We had extracted the votes received by each party for every district in each state for all the years required for analysis and prediction.
4. We used a software called ArcGIS to open the district shapefiles. With the help of ArcGIS we were able to extract the counties from certain districts. we also were able to extract the area type of every zip code to know whether it is a rural area or an urban area. Once that was formed we had to convert the data from zip code level to county level, then from county level to a district level.

5. We extracted the Urban Area zip codes with the help of the shapefiles and also manually had to find the rural area zip codes by collecting all the zip codes and removing the zip codes that fall under urban area for every district for each election year.
6. From the above extracted data, we calculated the relative change of each and every factor that were under consideration(White Population, African American Population, Rural Areas, Urban Areas) for the purpose of analysis and prediction.
7. Once we retrieved all the relevant information, we clubbed them together into one CSV file that we could refer to during the analysis and prediction.
8. From the final CSV file we then handled certain data like “DIV/0 ERROR” and other missing values, which was performed with the help of python.

The final dataset created for this project involved State Code, Year, District, Major Areas, White Population, African American Population, Rural Areas, Urban Areas, Republican Votes, Democratic Votes, Relative Change in White Population, Relative Change in African American Population, Relative Change in Rural Areas, Relative Change in Urban Areas, Relative Change in Republican Votes and Relative Change in Democratic Votes in a comma separated value file.

5. Implementation

From the data preprocessing step we obtained a Congressional elections dataset for highly gerrymandered states in united states of America (i.e. for Maryland and North Carolina) from 107th to 115th congressional elections which covers the years from 2002 to 2016.

The data set obtained is as follows.

Year	District	White	Black or African A/ Major Areas	Relative change in Republican	Relative change in Democratic vote	Relative change in white populati	Relative change in black populati	Republican	Rural Areas	Urban areas		
24	2004	1	471033.95	91452.4	R	0.276791159	0.342294489	0.240362462	-0.40595587	245149	184	3
24	2004	2	285239.6167	137844.0667	R	-0.147739281	0.558400651	-0.268605975	1.125364651	75812	25	1
24	2004	3	285139.6167	146959.0667	S	0.281124127	0.250547775	-0.170430427	-0.074136539	97008	24	2
24	2004	4	214109.6667	165485.3333	U	0.516394382	0.495009267	-0.338856509	-0.303056785	52907	1	3
24	2004	5	303673.0833	188425.3333	U	0.435020903	0.485587696	-0.073481721	-0.211314139	87189	52	5

In order to predict the factors involved in Gerrymandering, Initially we analysed the relations between various factors like White American population vs republican votes ,African American Population vs republican votes etc. and then developed a prediction system for number of Republic/Democratic Votes.

we wrote a Python code on spyder IDE to analyze, predict and to show graphical representation of these relations.

Steps for Analyzing the factors involved in Gerrymandering using python:

- Step 1: selected the particular attributes to which we need to find out the relation
- Step 2: Assigned the independent attribute values to X and dependent variable values to Y
- Step 3: Used Linear Regression from Scikit-learn and trained the model with X and Y
- Step 4: using matplotlib.pyplot , scatter plot with red color is plotted for all the X,Y points

Step 5: Now Y_{pred} is predicted from the trained regression model for X values and plotted the line in blue for X, Y_{pred} values

Step 6: Correlation coefficient between the population (race wise) v/s votes and areas (rural/urban) v/s votes were calculated to know the factors which have positive linear relation and support our assumptions.

Step 7: Correlation coefficient between the factors were calculated in excel using data analysis add-ins.

Steps for Predicting the Factors involved in Gerrymandering using python:

Step 1: we chose Republican Votes and Democratic vote as Dependent variables and assigned them to Y and then Population by race, Rural Areas, Urban Areas, Year, District, relative changes in population by race, relative changes in rural/urban areas to X .

Step 2: As the data is already preprocessed from data preprocessing step, dataset is splitted into training and testing sets using scikit-learn libraries.

Step 3: Used the Multiple Linear Regression model and fitted the model with X_{train}, Y_{train} values.

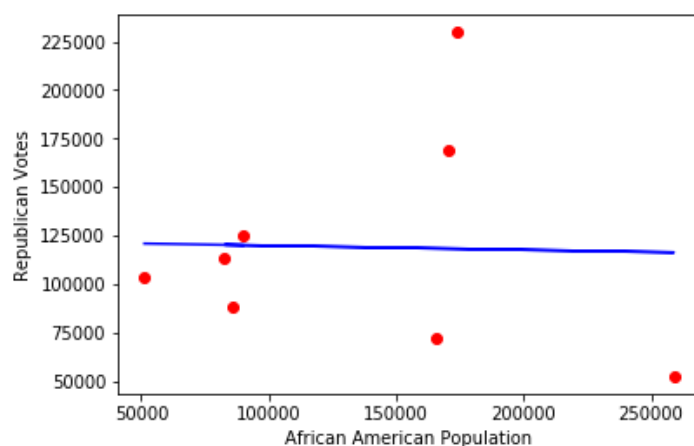
Step 4: Predicted the Y_{pred} for test values X_{test} .

7. Analysis

We analyzed the Maryland and North Carolina Datasets and found some interesting relations between various factors we considered for gerrymandering

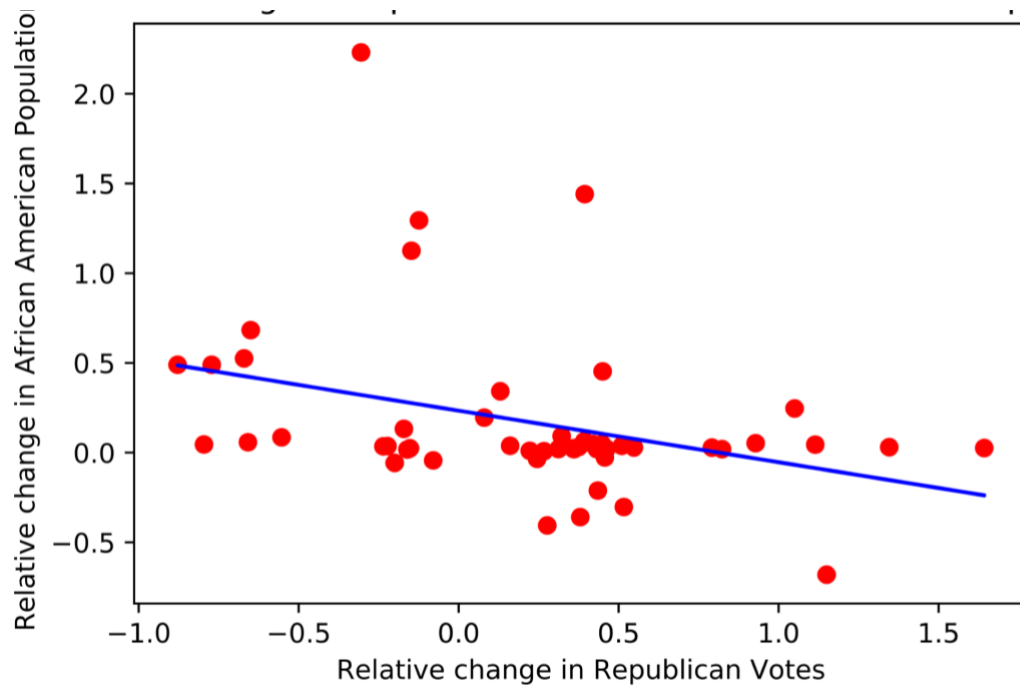
7.1 Maryland:

We tried to Analyze direct relation between White American or African American Population and Republican or Democratic votes.

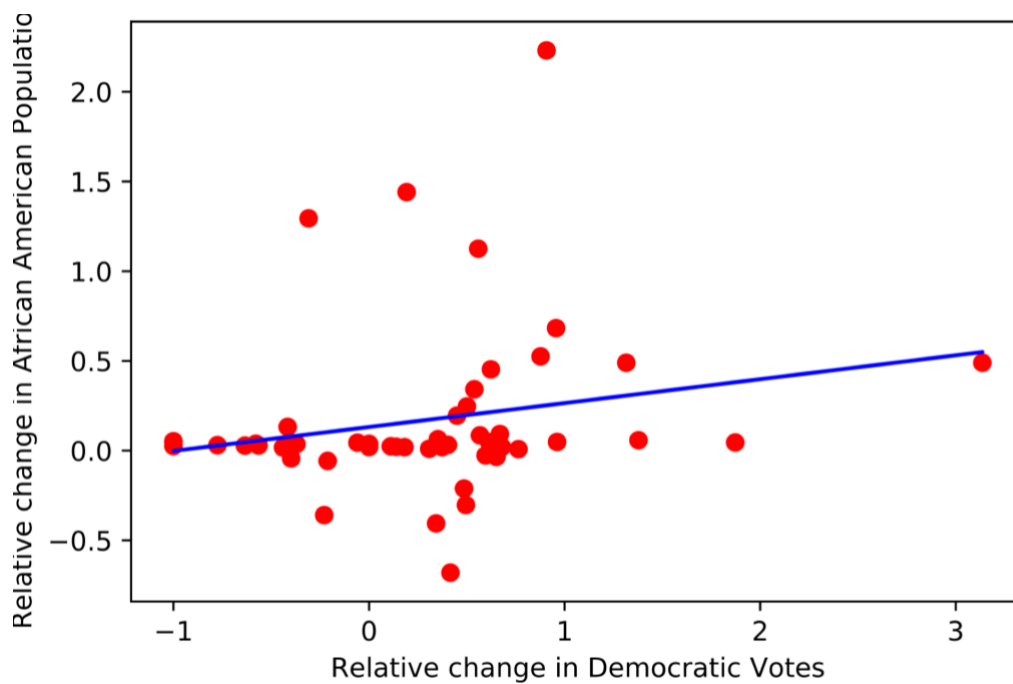


As shown above we couldn't find the direct relation between these factors. So we found the relative changes in population and votes and then we found the following interesting factors.

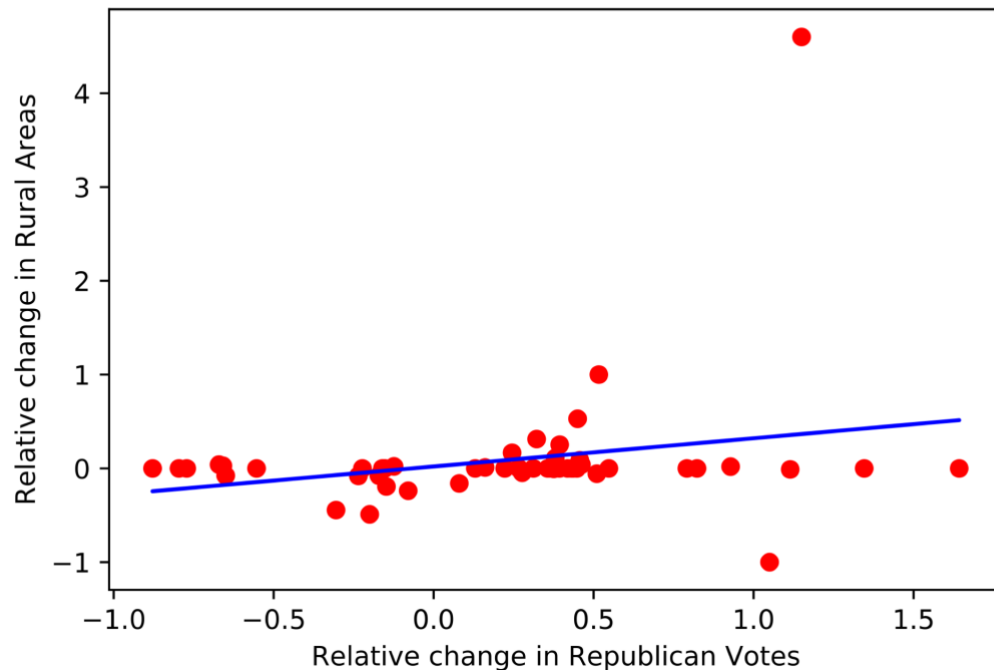
- Relation between Relative change in African American population and Relative change in Republican votes are decreasing linearly with more points concentrated around linearity from year 2002 to 2016 (i.e. 107th to 115th congressional elections)



- Relation between Relative change in African American population and Relative change in Democratic votes are increasing linearly with more points concentrated around linearity from year 2002 to 2016 (i.e. 107th to 115th congressional elections)



- Relation between Relative change in Rural Areas and Relative change in Republican Votes are slightly increasing linearly with little noise around linearity from year 2002 to 2016 (i.e. 107th to 115th congressional elections)



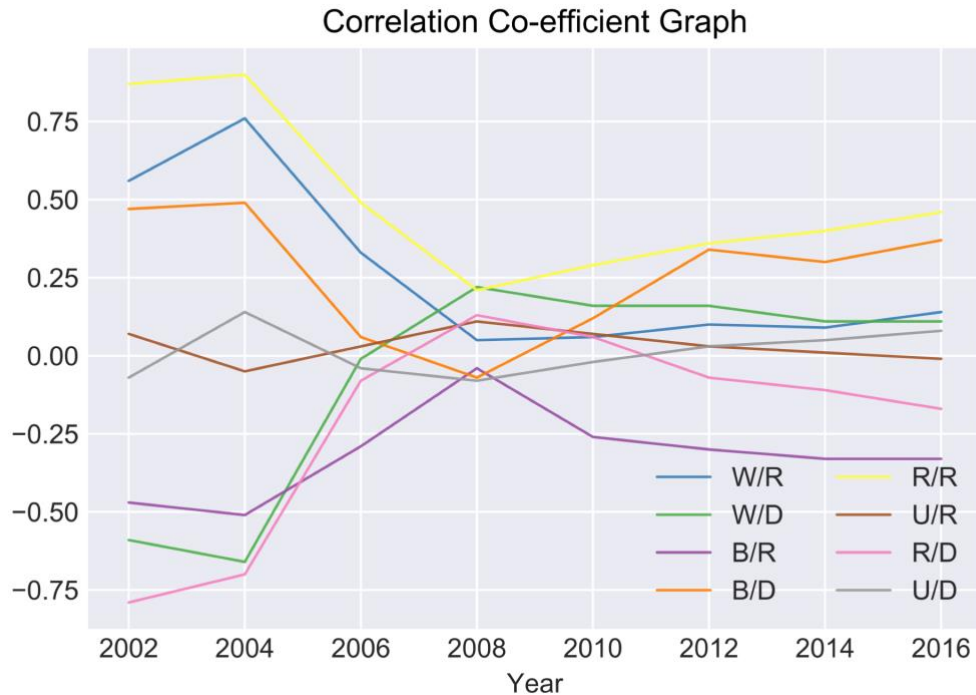
From the above observations it is evident the population and change in area resulted on redrawing the district boundaries for each congressional elections has a relation with the relative change in votes. In addition to this it is also evident that Relative change in African American population is positively correlated with relative change in Democratic Votes and relative change in African American Population is negatively correlated with Republican Votes.

7.1.1 Calculation of Correlation coefficient for Maryland :

As we found the linear relation between the factors from the scatter plot, we further calculated the correlation coefficients to know numeric relationship. Below is the correlation coefficient table.

Year	W/R	W/D	B/R	B/D	R/R	U/R	R/D	U/D
2002	0.56	-0.59	-0.47	0.47	0.87	0.07	-0.79	-0.07
2004	0.76	-0.66	-0.51	0.49	0.9	-0.05	-0.7	0.14
2006	0.33	-0.01	-0.29	0.06	0.49	0.03	-0.08	-0.04
2008	0.05	0.22	-0.04	-0.07	0.21	0.11	0.13	-0.08
2010	0.06	0.16	-0.26	0.12	0.29	0.07	0.06	-0.02
2012	0.1	0.16	-0.3	0.34	0.36	0.03	-0.07	0.03
2014	0.09	0.11	-0.33	0.3	0.4	0.01	-0.11	0.05
2016	0.14	0.11	-0.33	0.37	0.46	-0.01	-0.17	0.08

The correlation graph was plotted using matplotlib to analyse the factors visually.



From the year 2002 to 2006 and 2008 to 2016 there is a change in the correlation of the below aspects.

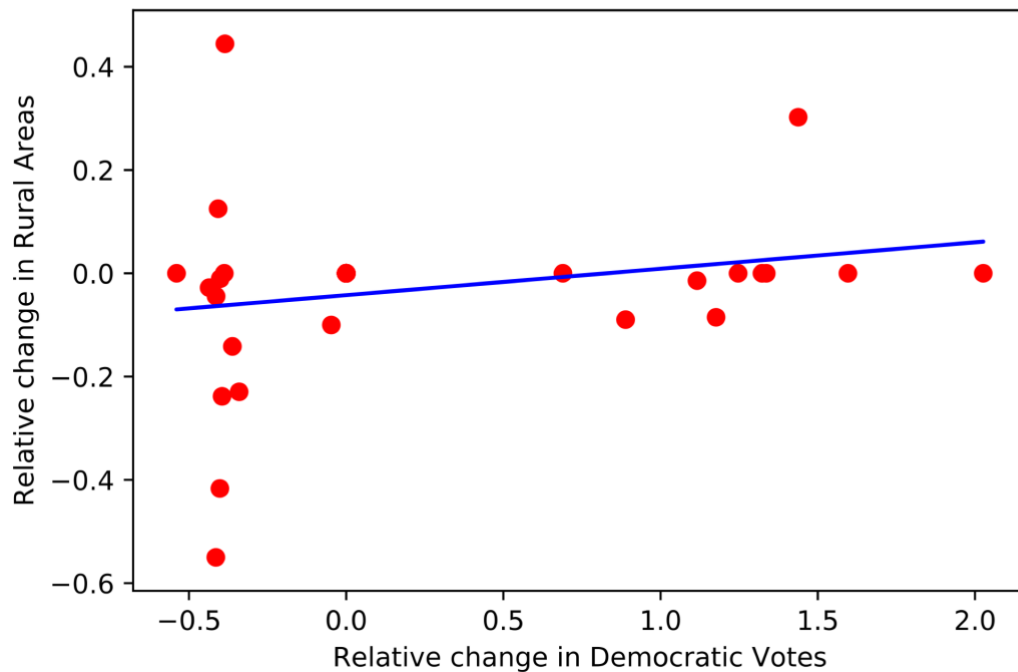
- As we can see, in the year 2002 the coefficient between white population and democratic votes is negative -0.59. And, a substantial increase in the correlation between white population and democratic votes i.e., 0.11
- Correlation between African American and Republican votes is always negative, it is seen that there is a negatively large variation over the years i.e. from 2002 to 2008 there is small negative increase from -0.47 to -0.04 but again it decreases to -0.33 in 2016.
- Correlation between rural area and republican votes is positive and it is highly correlated in 2002 with correlation coefficient of 0.87 but gradually it has decreased to 0.46 in 2016.
- Correlation between rural area and republican votes is positive and it is highly correlated in 2002 with correlation coefficient of 0.87 but gradually it has decreased to 0.46 in 2016.
- Correlation between White American Population and republican votes is always positive from 2002 to 2016, even though the amount of correlation gradually decreasing.

7.2 North Carolina:

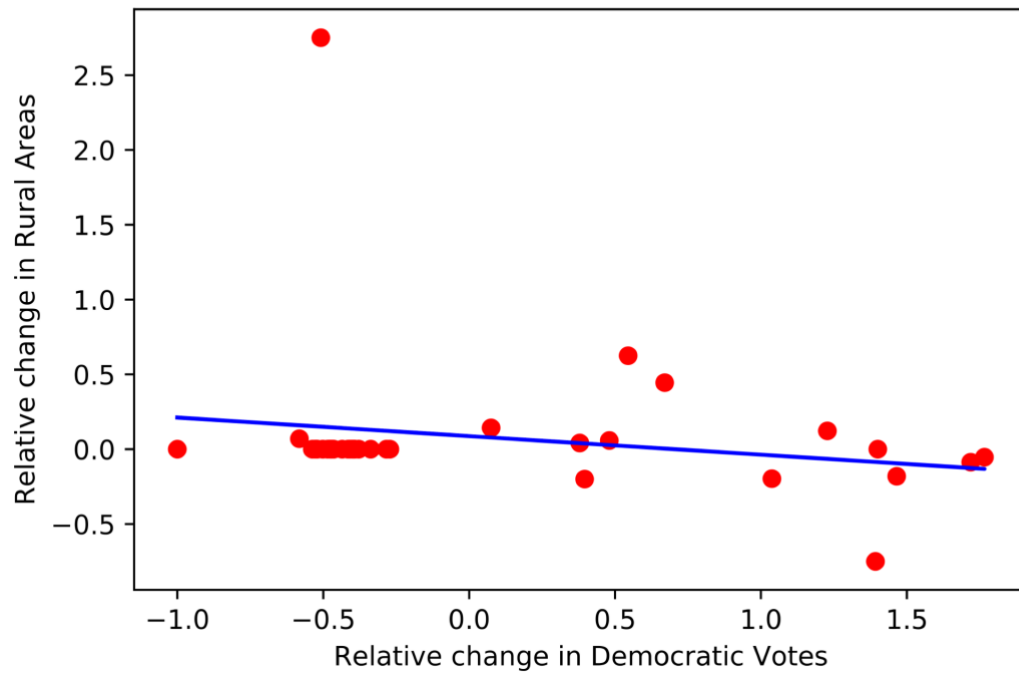
In North Carolina, the thing which interests more is unlike Democratic Party always winning race in Maryland from 107th to 115th congressional elections, Republican party has won the

race from 111th to 115th congressional elections. so the relations which we observed varies from 2002 to 2008 and 2008 to 2016. Following are some interesting relations which we observed.

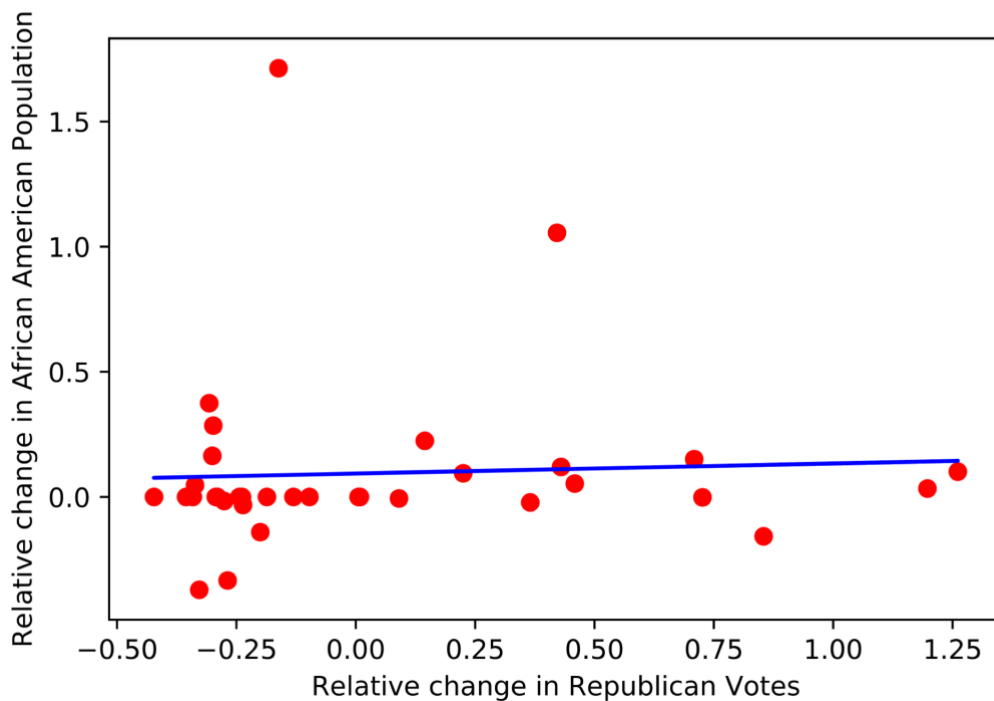
- Relative change in Democratic votes is linearly increasing with relative change in Rural areas with more points concentrated around linearity from year 2002 to 2008 (i.e. 107th to 110th congressional elections)



- Relative change in Democratic votes is linearly decreasing with relative change in Rural areas with more points concentrated around linearity from year 2010 to 2016 (i.e. 111th to 115th congressional elections)



- Relative change in Republican votes is linearly increasing with relative change in African American population with more points concentrated around linearity from year 2010 to 2016 (i.e. 111th to 115th congressional elections)



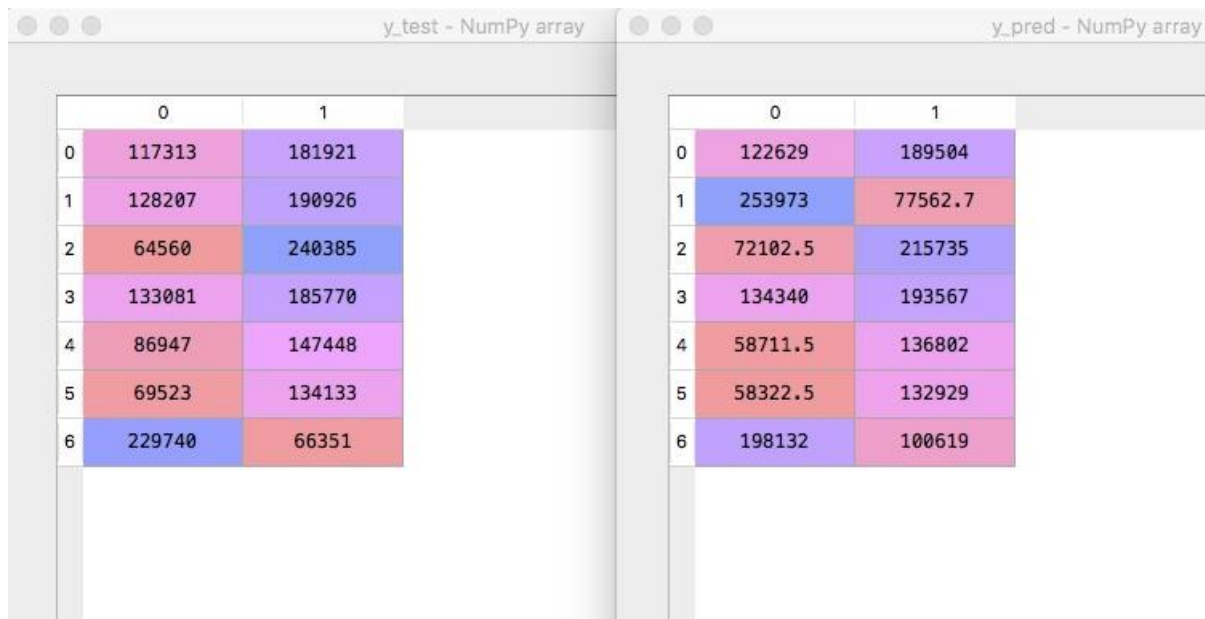
From the above observations it is evident that from 2002 to 2008 population in rural areas also supported Democratic Party and played a role for Democratic party in winning the race. From 2010 to 2016 the change in rural area resulted on redrawing the district boundaries for each congressional elections has a negative relation with the relative change in Democratic

votes. In addition to this it is also evident that Relative change in African American population is positively correlated with relative change in Republican Votes in the years 2010 to 2016.

8. Prediction

After Analyzing congressional election data of Maryland and North Carolina, we applied Multiple Regression algorithm to train the model and predict the Democratic and Republican votes considering limited features like Population by race, Year of elections, District , Rural and Urban areas to the Maryland state.

Following are the test values and predicted values obtained for Maryland given X_test values



	0	1
0	117313	181921
1	128207	190926
2	64560	240385
3	133081	185770
4	86947	147448
5	69523	134133
6	229740	66351

	0	1
0	122629	189504
1	253973	77562.7
2	72102.5	215735
3	134340	193567
4	58711.5	136802
5	58322.5	132929
6	198132	100619

Above is the screenshot of test and predicted values of y. Here first column and second column indicates predicted values of Republican votes and Democratic Votes respectively. From the limited features we have considered we cannot say that Multiple Linear regression model has accurately predicted the values but can definitely say that this model has made the sensible predictions with the limited features it has for the Maryland state.

9. Future Scope

In our work, we analyzed the relations between major things like population by race, change in count of rural and urban areas by considering zip codes and votes. We calculated change in area using zip codes after redrawing district boundaries but this study could be more accurate if we consider the unit change in area after redrawing district boundaries and type of area that is added after gerrymandering like rural/urban part of zip code. This work can also be extended in the scenario where the same zip code falls under 2 districts, then which part of the zip code (i.e. urban/rural) is included in that district and influence of that on votes. In addition to this, study will be more accurate if we consider the population by race that is added to that district because of redrawing district boundaries separately and influence of that population in winning the race for the party.

10. Conclusion

After the detailed analysis of the data and the graphs obtained from the study of gerrymandered states, the following conclusions are drawn

- It makes us believe that the assumption made about the Maryland democratic party influencing the congressional elections to their favour through redistricting boundaries lines appears reasonable.
- In case of North Carolina state, the hypothesis made about the republican party of controlling the congressional election in their favour also looks plausible by analysing the graphs and dataset of the state.
- These assumptions can be solidifying if granular viewpoints like the unit change in area is used and not just change via zip code which will improve the accuracy of results to the great extent

11. References

1. Data Preprocessing. (2018). Retrieved from <https://www.techopedia.com/definition/14650/data-preprocessing>
2. Bureau, U. (2018). Population and Housing Unit Estimates Tables. Retrieved from <https://www.census.gov/programs-surveys/popest/data/tables.All.html>
3. (DADS), D. (2018). American Factfinder - Results. Retrieved from <https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkml>
4. Election Statistics, 1920 to Present | US House of Representatives: History, Art & Archives. (2018). Retrieved from <http://history.house.gov/Institution/Election-Statistics/Election-Statistics/>
5. Branch, G. (2018). Cartographic Boundary Shapefiles - Congressional Districts - Geography - U.S. Census Bureau. Retrieved from https://www.census.gov/geo/maps-data/data/cbf/cbf_cds.html
6. Cartographic Boundary Shapefiles - Urban Areas - Geography - U.S. Census Bureau. (2018). Retrieved from https://www.census.gov/geo/maps-data/data/cbf/cbf_ua.html
7. Evidence of Gerrymandering | Kaggle. (2018). Retrieved from <https://www.kaggle.com/laa283/evidence-of-gerrymandering/data>