**IS 733 – Data Mining**

**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**

**Need to format**

|  |  |
| --- | --- |
| Introduction | 3 |
| Motivation | 3 |
| Data Collection | 3 |
| Data Preprocessing | 4 |
| Implementation | 5 |
| Future Scope | 14 |
| Conclusion | 15 |
| References | 15 |

**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. Gerrymandering as a concept was affirmed in 1812 Boston Congressional primaries where such an act as filed. 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.

**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.

**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).

1. U.S. Census Bureau :

<https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk>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).

1. 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).
2. 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).
3. 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).
4. 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).

# 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:

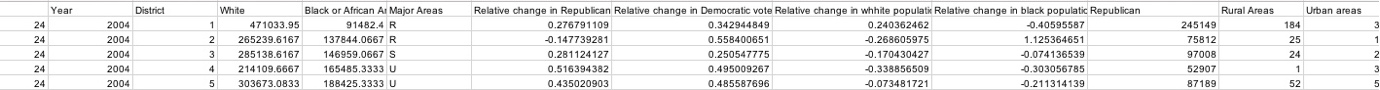
1. We had to extract the population of each race for every county which was available in different columns and calculate the race population for every district in each state. While converting from count level to district, we had to consider the situations where a county can fall under multiple districts. 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.

**Implementation :**

From the data preprocessing step we obtained a Congressional elections datasets 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.



In order to predict the factors involved in Gerrymandering ,Initially we analyzed 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

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.

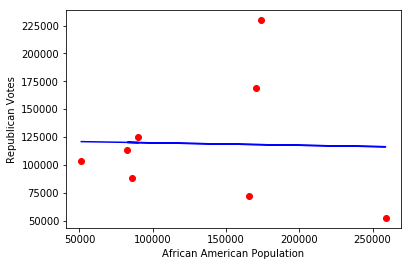
Step 4:Predicted the values Y\_pred for test values X\_test.

**Analysis :**

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

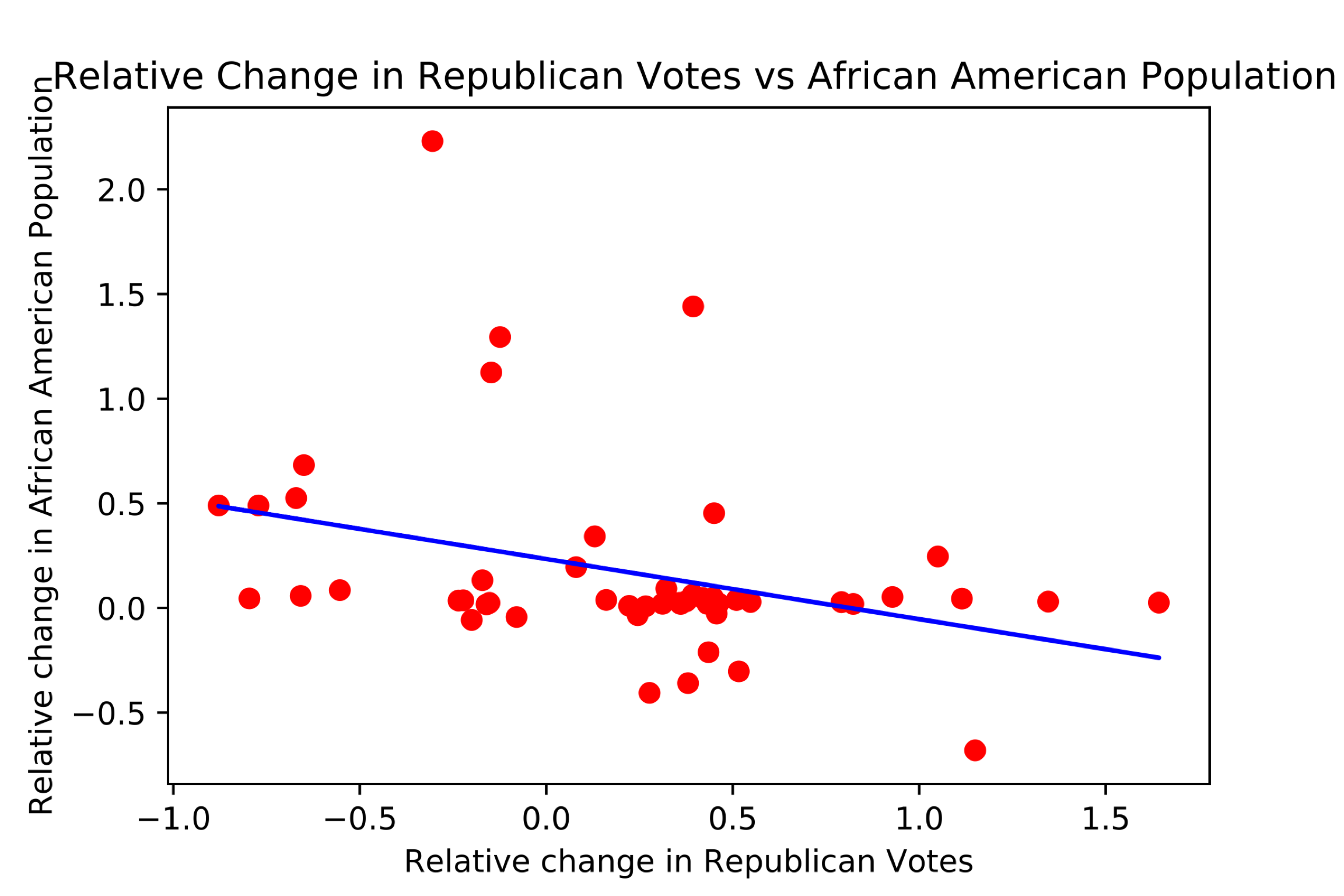
Maryland:

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

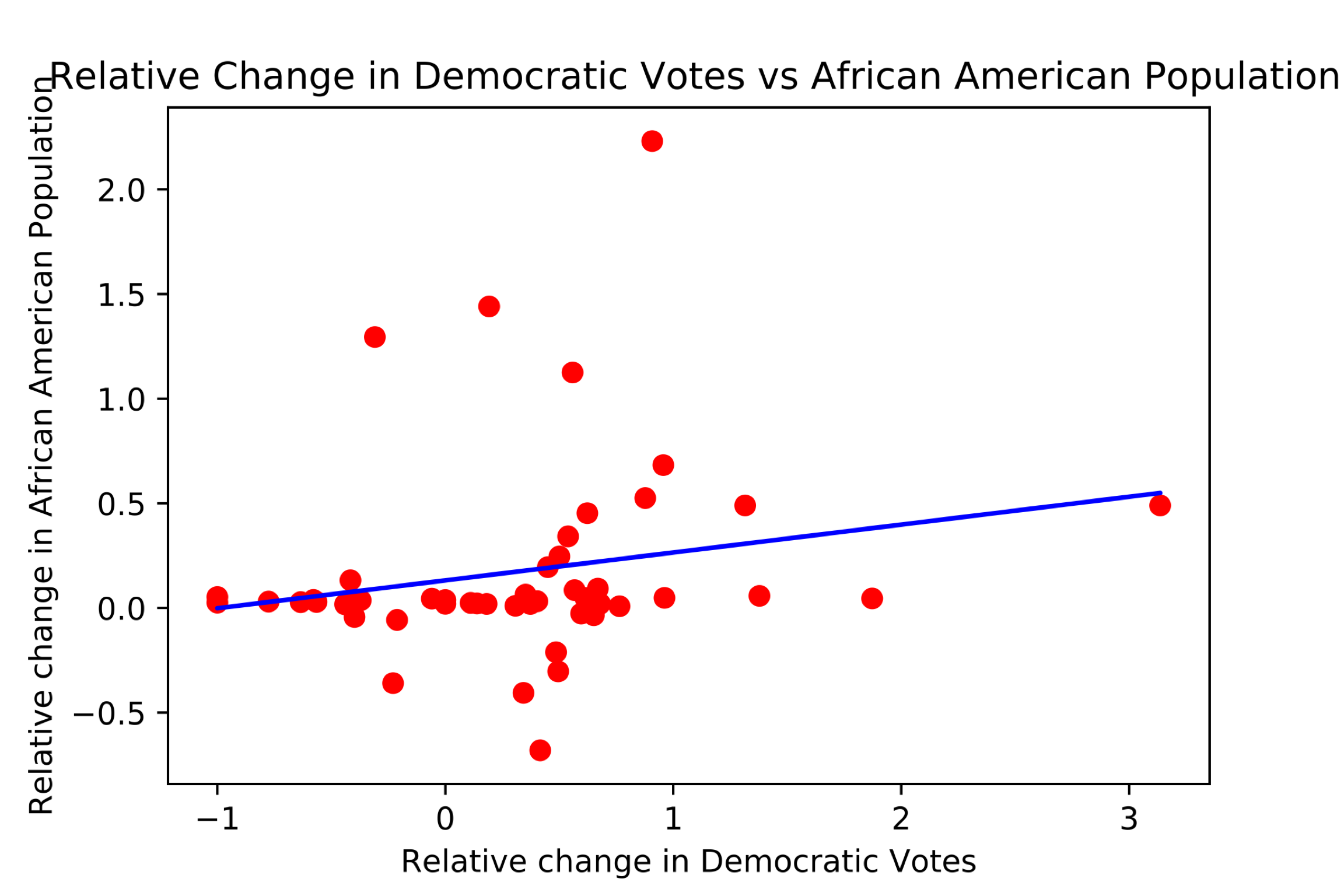


As shown above we couldn’t find the direct relation between these factors. So we found the relative changes in population and votes and then we found the following interesting factors.

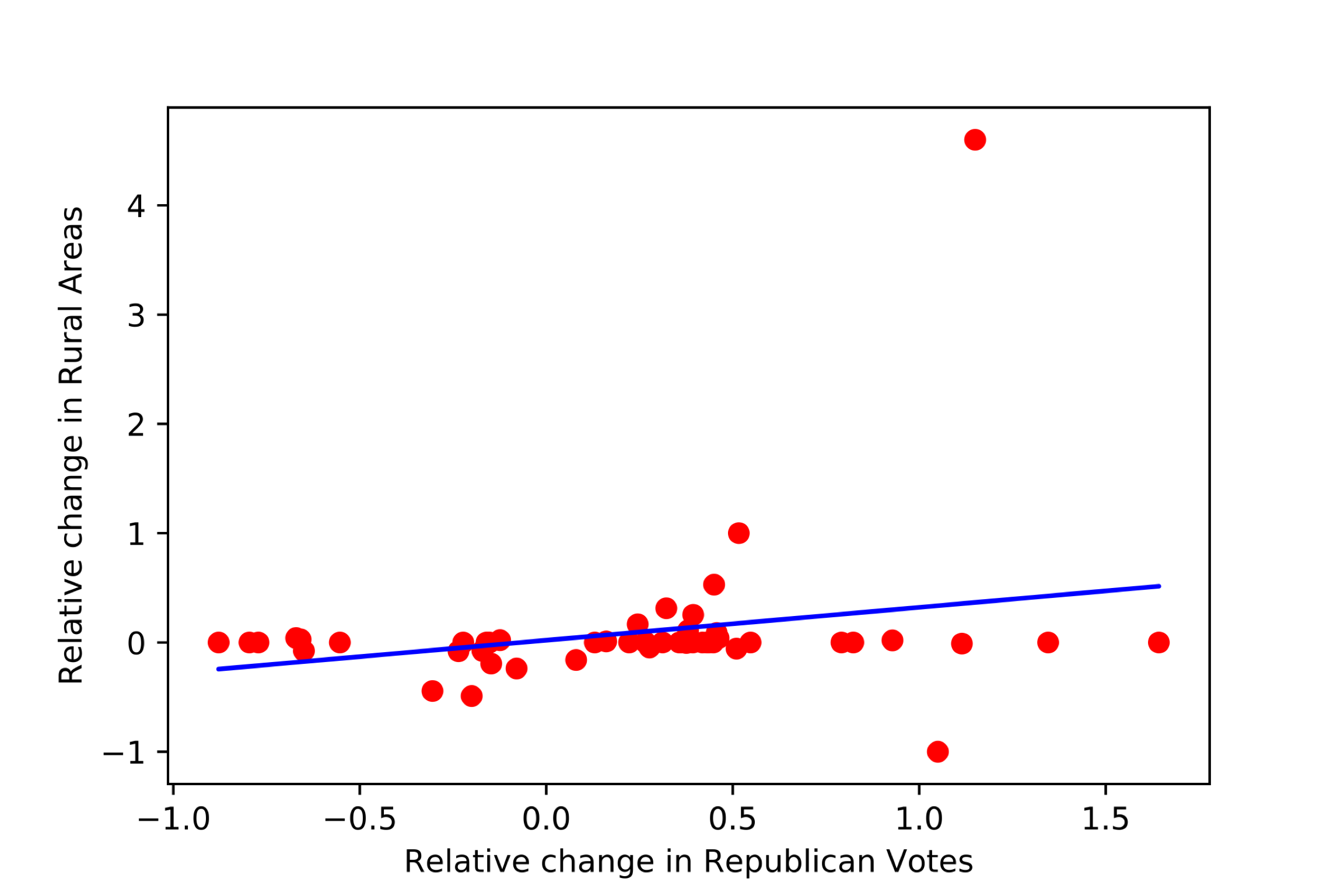
1. Relation between Relative change in African American population and Relative change in Republican votes are decreasing linearly with more points concentrated around linearity from year 2002 to 2016 (i.e. 107th to 115th congressional elections)



1. 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)



1. 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)

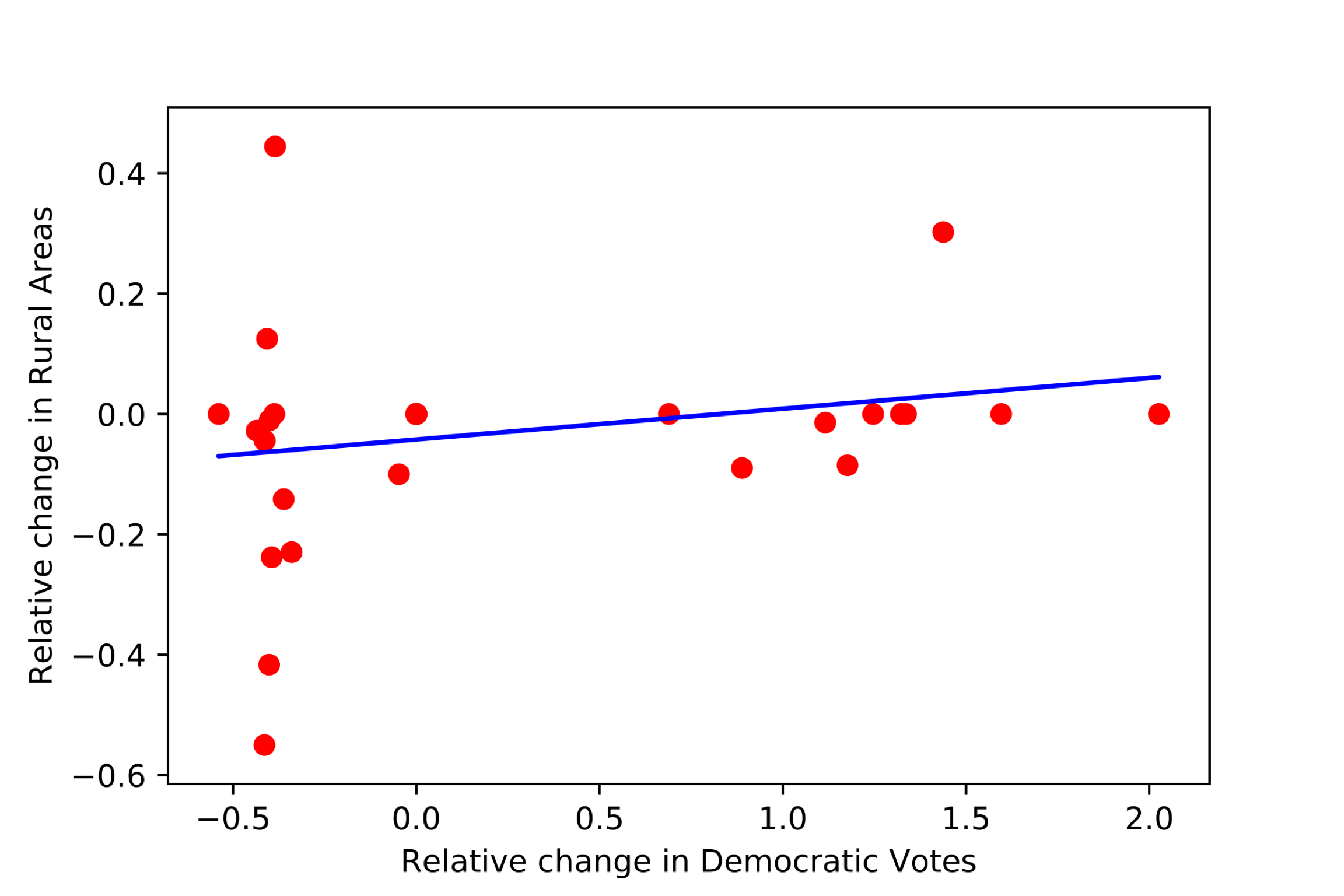


From the above observations it is evident the population and change in area resulted on redrawing the district boundaries for each congressional elections has a relation with the relative change in votes. In addition to this it is also evident that Relative change in African American population is positively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

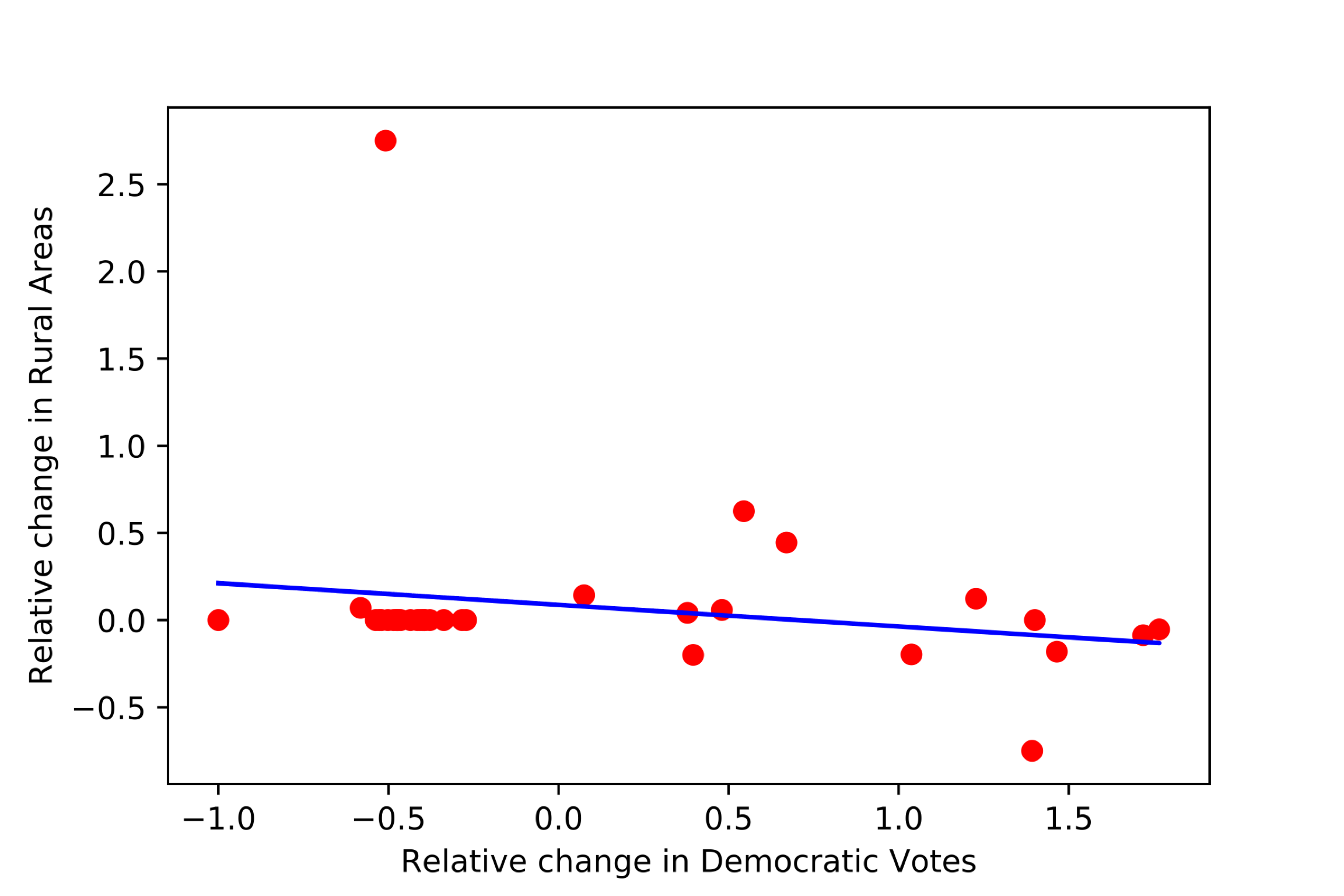
North Carolina:

In North Carolina, the thing which interests more is unlike Democratic Party always wining 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.

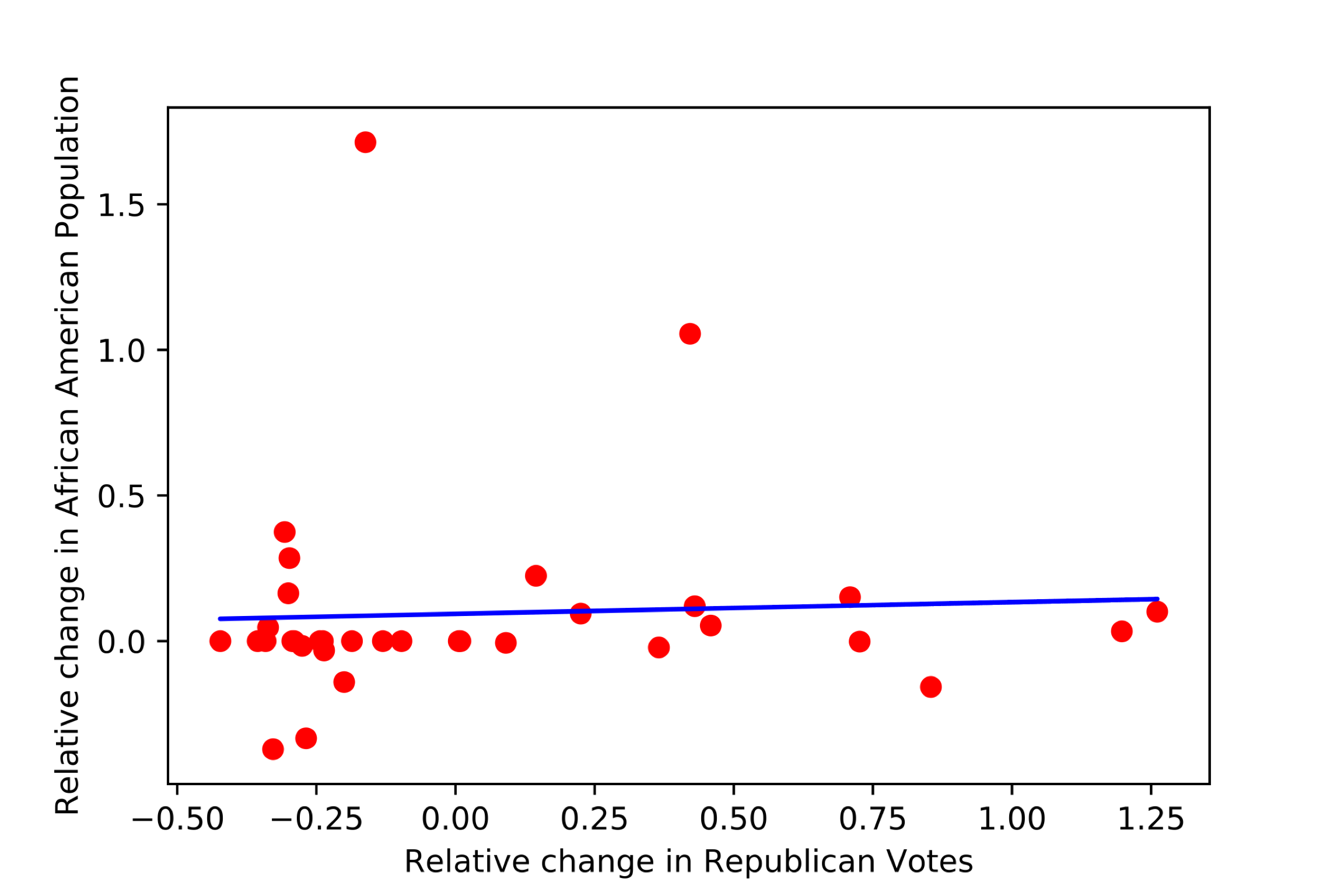
1. Relative change in Democratic votes is linearly increasing with relative change in Rural areas with more points concentrated around linearity from year 2002 to 2008 (i.e. 107th to 110th congressional elections)



1. 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)



1. 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)

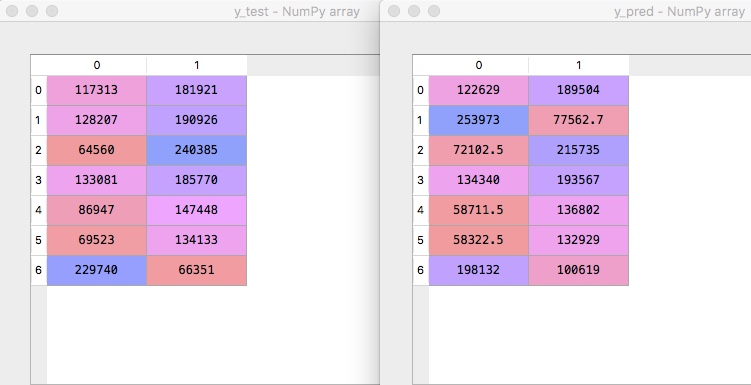


From the above observations it is evident that from 2002 to 2008 population in rural areas also supported Democratic Party and played a role for Democratic party in winning the race. From 2010 to 2016 the change in rural area resulted on redrawing the district boundaries for each congressional elections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

**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 given X\_test



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.

**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.

**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