Predicting Voters' Behavior in US Presidential Elections

Chidi Agbaeruneke, Badr Albrikan, Martin Ferreiro, Zachary Lessner

2020-12-10

Contents

1 Executive summary

Political analysts tend to agree that recent changes in the demographic composition of the US have reshaped the electorate and that this has had an impact in voting patterns. This hypothesis implies that economic and demographic characteristics are important determinants of voting behavior. If this is true, then we should be able to predict election results by analyzing variables such as voters' age, ethnicity, level of education, or income.

Based on this assumption, we attempt to build a powerful model to anticipate presidential elections outcome in the US. In order to achieve this, we conduct a regression analysis using an MIT database which contains demographic and past election data at the county level. By selecting different combinations of potential explanatory variables we fit different models, and then select among them based on their predicting power and simplicity.

In the second part of this project, we use our model to compare between different explanatory variables to identify which has the greatest impact on voting behavior. Contrary to what we expected based on our literature review, we find that that the percentage of college educated voters in a county has greater influence on the election results than voters' race.

2 Introduction

The demographic composition of the electorate in an area is often considered a good predictor of the election outcome. This does not imply that there are no other variables that influence voters' decisions. Electors are also susceptible to circumstantial factors such as the charisma and track record of a candidate, or how strong the economy is performing on election year. Still, analyst tend to agree that voting patterns are associated with demographic variables such as voters' education level, income or ethnicity.

The Pew Research Center, a non-partisan fact tank, has published numerous articles analyzing how the composition of the US electorate has changed over time and how this has impacted the elections. In a recent publication , John Gramlich (2020), senior writer of the Center, analyzed the profile of registered voters in terms of race, age, education, religion and how each of these variables was associated with party id. Gramlich showed that white voters have consistently accounted for a much larger share of Republican registered voters than of Democratic voters, and that voters who identify with the Democratic Party or lean toward it are much more likely than their Republican counterparts to have a college degree.

Prominent scholars have also taken a data driven approach to analyze the relation between voters' behavior and demographics. Barilla and Levernier (2006), from Georgia Southern University, analyzed the 2000 US election and concluded that economic and demographic characteristics were important determinants of the observed voting patterns. Hill, Hopkins

and Huber (2019) examined whether demographic changes at low levels of aggregation were associated with vote shifts between 2012 and 2016. They showed that influxes of Hispanics or non citizen immigrants benefited democrats over republicans. Diggs, Farooq, Kidd, and Murray (2006), on the other hand, analyzed black voter's preferences and concluded the influence of Democratic Party allegiance is a very powerful cue for them. Our literature review reveals most researchers focus on the influence of race on voting behavior, highlighting the tendency of Hispanic and particularly black voters to support Democratic candidates.

If, as previous research on the field suggests, economic and demographic variables influence voting behavior, then we should be able to build a powerful model to predict election results in the US. In order to achieve this, we conduct a regression analysis using a data set from the MIT Election Data and Science Lab (MEDSL). This file contains the results for the 2012 and 2016 Presidential, Senate and Congressional elections at the county level. It also includes 18 different demographic variables such as percentage of black population, median household income or percentage of rural population. By selecting different combinations of potential explanatory variables we fit different models, and then select among them based on their predicting power and simplicity. Due to time constrains, the analysis presented here focused only on the 2016 presidential election, and specifically on Trump's performance. Our model could and should be adjusted by replicating the analysis on the 2020 election results (updated demographics would be required).

The approach we have chosen enables us to build a model that can predict results at different levels: county, state and nation. Our model also allows to compare between different explanatory variables to see which has the greatest impact on voting behavior. In the second part of this project, we test whether or not race, the variable most commonly cited as a predictor of voting patterns in the US, is in fact more influential than other variables included in our model, such as education, gender and age.

3 Question of Interest

What demographic variables help us to predict Presidential election results in the United States?

3.1 Hypothesis

According to our research on which predictors are most important for a candidate's ability to capture vote share, we expect the variable associated with the category of race to have the most influence on our regression model.

4 Data Preparation

As explained, our original data set contained results for the 2012 and 2016 presidential, senate and congressional election, as well as 18 different demographic variables with information at a county level (all quantitative continuous variables, except for one that was ordinal). It is worth noting the Alaska counties are missing. To facilitate the manipulation and tidying of the data, we created two different dataframes: one for the election results and the other for demographic variables.

- Below we show the first three rows for our original Election Results Data frame

```
## # A tibble: 3 x 9
##   state  county  fips trump16 clinton16 otherpres16 romney12 obama12 otherpres12
##   <chr>  <chr>  <dbl>   <dbl>     <dbl>       <dbl>    <dbl>   <dbl>       <dbl>
## 1 Alaba~ Autau~  1001   18172      5936         865    17379    6363         190
## 2 Alaba~ Baldw~  1003   72883     18458        3874    66016   18424         898
## 3 Alaba~ Barbo~  1005    5454      4871         144     5550    5912          47
```

As the output shows, our election results dataframe was originally formatted in such a way that the name of each column contained a candidate's name and the year of the election in which he competed, while the values corresponded to the number of votes obtained in each county. Since these are actually three different variables, we did pivot longer on the data frame and created three different columns: one referencing the candidate's names, one referencing the year of the election and one referencing votes obtained in each county.

Another challenge that we faced was that the "votes" variable contained the **number** of votes won by each candidate, not the percentage. With this format, we would not have got any meaningful results from a regression analysis: regardless of how we fitted our model, we would have seen all candidates doing better on bigger counties, were there is a larger number of votes, and worse on smaller counties. To avoid this, we applied some basic arithmetic to transform the votes variables into a vote share variable, containing percentages. Output below shows the Election results data frame after concluding the tidying process.

```
## # A tibble: 3 x 7
##   state   county   fips candidate year  votes Vote_share
##   <chr>   <chr>   <dbl> <fct>     <fct> <dbl>      <dbl>
## 1 Alabama Autauga  1001 trump     16    18172      72.8
## 2 Alabama Autauga  1001 clinton   16     5936      23.8
## 3 Alabama Autauga  1001 otherpres 16      865       3.46
```

As explained, due to time constrains, the analysis presented here focused only on the 2016 presidential election, and specifically on Trump's performance. Therefore, when fitting

the regression model, we built an auxiliary data frame containing only information on the president's vote share and the demographic variables, both at a county level.

5 Data Exploration

We first looked at what variables we wanted to considered for our model. Our response variable was (Trump's) Vote Share. Our potential explanatory variables were

```
##  [1] "total_population"       "cvap"                   "white_pct"
##  [4] "black_pct"              "hispanic_pct"           "nonwhite_pct"
##  [7] "foreignborn_pct"        "female_pct"             "age29andunder_pct"
## [10] "age65andolder_pct"      "median_hh_inc"          "clf_unemploy_pct"
## [13] "lesshs_pct"             "lesscollege_pct"        "lesshs_whites_pct"
## [16] "lesscollege_whites_pct" "rural_pct"              "ruralurban_cc"
```

Cvap is citizen voting age population. The inclusion of this variable made total population redundant, so we removed the latter. We also removed nonwhite_pct, which was basically the sum of black (population) percentage and hispanic (population) percentage.

5.1 Descriptive Statistics

To analyze the distribution of the variables and the correlation between them, we built a correlation matrix (available in the Appendix A). This showed some multicollinearity present between some of the predictor variables. This was to be expected because certain predictors were actually subsets from the same category of demographics. According to the histogram charts in the matrix, our response variable Vote_share appeared closely normally distributed. All of the predictor variables except age29andunder_pct exhibited skewness of varying degree.

6 Model Fitting

Our first approach to try to build the most powerful model was to implement a Stepwise Automatic Selection method.

6.1 Stepwise Automatic Selection Method of Model 0

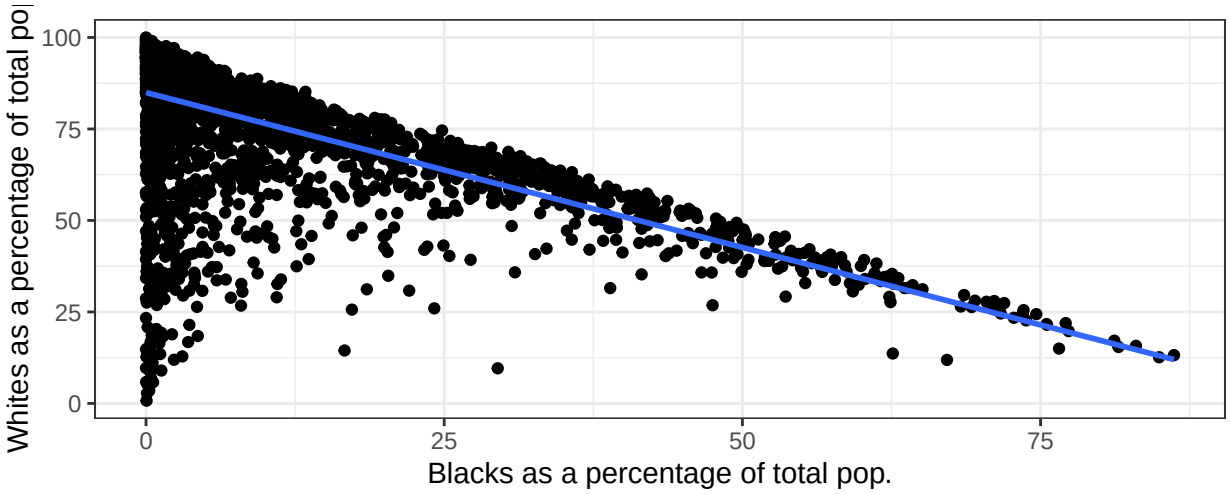
The formula for the model suggested by the Stepwise Automatic Selection Method (shown below) included 13 explanatory variables.

```
## lm(formula = Vote_share ~ white_pct + black_pct + hispanic_pct +
##     foreignborn_pct + age29andunder_pct + age65andolder_pct +
##     median_hh_inc + clf_unemploy_pct + lesshs_pct + lesscollege_pct +
##     lesscollege_whites_pct + rural_pct + ruralurban_cc, data = county_elec_data)
```

```
## [1] "Adjusted R-squared = 0.62471357337579"
```

After reviewing the model, we encountered concerns of potential multicollinearity: as our exploratory analysis indicated, some of the variables included in the model are highly correlated due to the fact that they are actually subsets from the same category of demographics. For instance, percentage of white population and percentage of black population are both racial variables and, as the plot bellows shows, they are strongly negatively correlated:

```
## `geom_smooth()` using formula 'y ~ x'
```



The correlation matrix included in Appendix A shows that other variables were also strongly correlated. To reduce multicollinearity, we decided to divide the variables into different categories- race, gender, age, income, education, and locality- and fit two simpler models.

6.1.1 Categories and Respective Variables of Predictors

- Race: Whites, Blacks, Hispanics, and Non-whites as a percentage of total county population.
- Gender: Female as a percentage of total county population.
- Age: (age 29 and under) and (age 65 and older) as a percentage of total county population.
- Income
  - median household income in the past 12 months (in 2016 inflation-adjusted dollars).

  - Unemployed as a percentage of total labor force by county.

- Education
  - Less than (regular high school diploma) and (bachelor's degree) as a percentage of total county population.
  - White population with less than (regular high school diploma) and (bachelor's degree) as a percentage of total county population.

- Locality: rural population as a percentage of total county population.

We built the two alternative models by selecting one variable from each group. Our aim was to find a simpler model with similar predicting power to model 0 (the one suggested by the Stepwise Automatic Selection Method).

6.2 Model 1

```
## lm(formula = Vote_share ~ white_pct + female_pct + age29andunder_pct +
##     clf_unemploy_pct + lesscollege_pct + rural_pct, data = county_elec_data)
```

```
## [1] "Adjusted R-squared = 0.585432940614355"
```

6.3 Model 2

```
## lm(formula = Vote_share ~ black_pct + female_pct + age65andolder_pct +
##     median_hh_inc + lesshs_pct + rural_pct, data = county_elec_data)
```

```
## [1] "Adjusted R-squared = 0.402611538404421"
```

After comparing the predicting power of Model 1 and Model 2, we elected to proceed with Model 1, which resulted in a higher Adjusted R-squared value of 0.5854. Model 1 also proved more efficient in explaining the data as compared to Model 0, the one generated by the Stepwise Automatic Selection Method. With half the number of variables, we lost very little predicting power (Adjusted R-squared decreased from 0.624 to 0.585).

Based on the importance existing literature assigns to the racial composition of the electorate and the presence of minorities, we tried a small adjustment in model 1, which consisted in replacing only white_pct with black_pct. We also experimented with Hispanic_pct instead of white_pct. None of these alternative models (included in the Appendix B) showed greater predicting power than model1.

6.4 Assessment of variable importance to Model 1

Once we decided on a model, we proceeded to test our initial hypothesis concerning the influence of our explanatory variables on voting patters. Specifically, we tested whether or not our racial variable was in fact more influential than other variables included in our model, such as those associated with education, age, gender, or economic status.

In order to achieve this, we fitted different auxiliary models by removing in every case only the variable whose impact we wanted to measure. By comparing the decrease of the adjusted R squared for every model, we determined which variable contributed the most to the predictive power of model 1, that is, which explained the largest part of the variability in our response variable.

Comparison of Model Strength

```
##                        Model       rsq   adj.rsq      aic      bic    press
## 1                 Full Model 0.5862327 0.5854329 14396.43 14438.73 318423.2
## 2         Without female_pct 0.5854633 0.5847957 14400.21 14436.46 318807.4
## 3 Without age29andunder_pct 0.5780954 0.5774160 14455.02 14491.27 324463.6
## 4          Without rural_pct 0.5755387 0.5748552 14473.81 14510.07 326350.9
## 5  Without clf_unemploy_pct 0.5501244 0.5494000 14654.72 14690.97 345992.8
## 6          Without white_pct 0.4600314 0.4591619 15222.60 15258.86 415397.0
## 7   Without lesscollege_pct 0.4239368 0.4230091 15423.90 15460.16 443011.9
```

The output above shows that the largest decrease in the predicting power of the model occurred when we removed the variable associated with education, that is, "population with an education of less than a bachelor's degree". Based on these results, we rejected our original hypothesis that our racial variable is the one that has the largest influence in voting behavior.

The comparison of each of the variables' contribution to the model's predictive power not only revealed that education is the most influential variable, but also that some of the variables we had selected could easily be removed without weakening the model. The output above showed that the gender, age and locality predictors explained a very small portion of the variability on vote share once the other variables were already considered.

In this particular model, gender as a percentage of county population predictably does not tell us much, given most counties are pretty close to 50-50 between male and female. A different approach would be required to evaluate the relationship between gender and voters' behavior.

In the interest of parsimony, we did a new adjustment to our model and removed the variables female_pct, age29andunder_pct, rural_pct.

6.5 Model 3

```
## lm(formula = Vote_share ~ white_pct + lesscollege_pct + clf_unemploy_pct,
##     data = county_elec_data)
```

```
## [1] "Adjusted R-squared = 0.570371931585727"
```

When comparing the fit of Model 1 (our original model with six explanatory variables) against Model 3 (our new adjusted model) on the basis of Adjusted R-squared, we noticed a small difference of 0.0158327. This proved that the addition of the three variables- female_pct, rural_pct, and age29andunder_pct- contributed little to predicting the outcome of the election once the other variables were already considered. Based on this finding, we decided that model 3 would be our final model.

- $\hat{Votes}hare = -31.8722751 + 0.3440012(whitepct) + 0.9700893(lesscollegepct) - 1.1795349(clfunemploypct)$

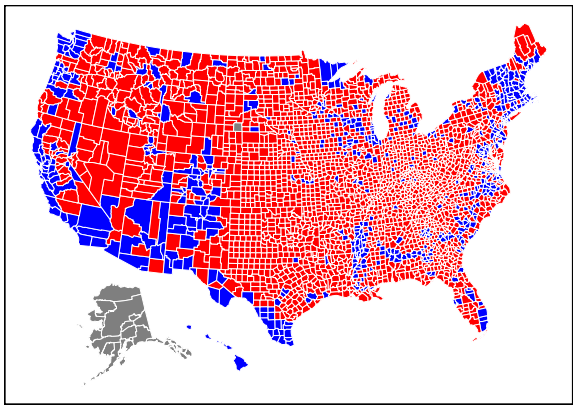
According to our prediction equation, there is a positive relationship between percentage of white population and vote share for the republican candidate: as one increases, so does the other. There is also a positive relationship between the percentage of the population in a county without a college degree and the electoral support the republican candidate receives there. These findings are in line with the reviewed literature.

Finally, our model reveals a negative correlation between the unemployment percentage and the republican candidate's vote share. Since our analysis is limited only to the 2016 election, it is possible that unemployment has a negative relationship not with republican candidate's vote share but with the incumbent's vote share, that is, that people in counties with high unemployment tend to blame the sitting president for their economic struggles, regardless of what party he belongs to.

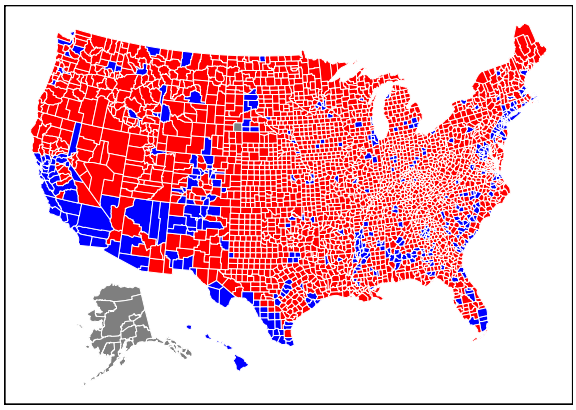
7 Testing the model

Besides analyzing the Adjusted R-squared of model 3, we decided to test its predictive power in a more practical manner: comparing the predicted winner of the 2016 election in every county with the actual winner of the 2016 election.

2016 Election results



2016 predicted Election results



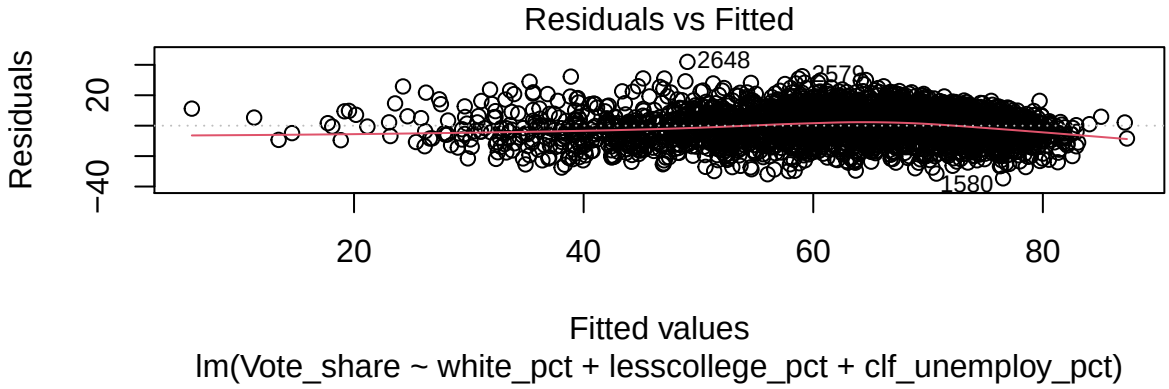
Both maps show similar patterns. It looks, however, as if our model over estimated Trump's vote share in the north east as wells as in the north west. One could hypothesize that this phenomenon results form the large percentage of white population in those areas that lean democrat, contrary to what we observe in the rest of the country.

Software calculations showed us that the predicted winner of the election was not the same as the actual winner in only 347 out of the 3111 counties. That is a percentual error barely over 11%.

8 Model Diagnostics

Linearity

- Graphical test: Residuals vs Fitted Plot (e_i vs $\hat{Y}$)



- Interpretation of Residual vs Fitted Values Plot

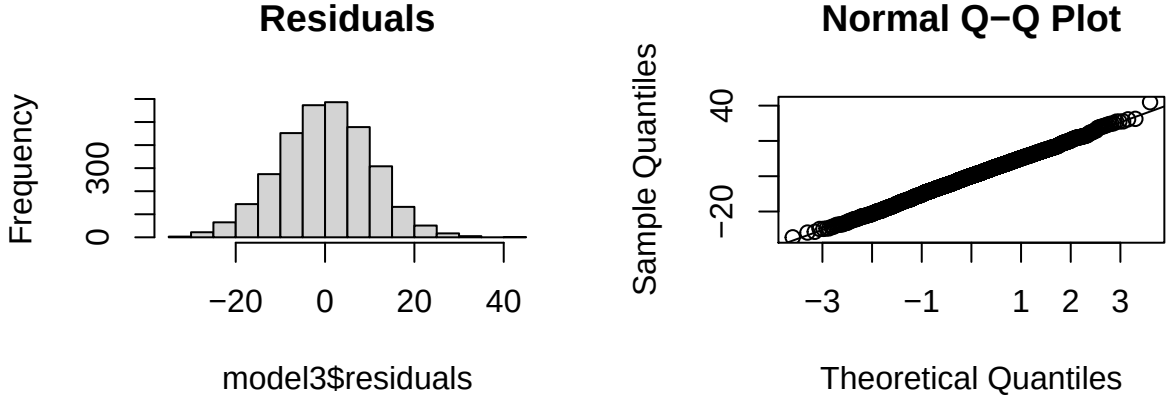
- This plot helped us evaluate the assumption of linearity. The red line looks relatively flat and does not deviate much from the dotted horizontal line of 0, meaning there does not appear to be a systematic pattern present. There is no clear violation of the linearity assumption. The plot was also useful in evaluating whether or not the assumption of homoscedasticity was violated. If there was constant variance about the line at 0 (homoscedasticity), the spread of residuals would be approximately the same across the x-axis. The plot shown above suggested there may be a violation of constant variance. We evaluated this by conducting a Brensch-Pagan test.

Constant Variance test

- Statistical test: Brensch-Pagan test
- Hypothesis
  - H_0 : The error variance is constant
  - H_a : The error variance is not constant
- Conclusion (calculations available in the appendix)
  - The test resulted in a p-value = 2.2e-16, which led us to reject H_0. This backed our interpretation of the plot: that the error variance is not constant.

Normality

- Graphical tests: Histogram and Q-Q plot



- The histogram plot of residuals displays a normal distribution and the dots on the Normal Q-Q plot are roughly scattered around the reference line randomly. There is minor deviation near the bottom left tail; however, it is not severe. This suggested the normality assumption was not grossly violated.

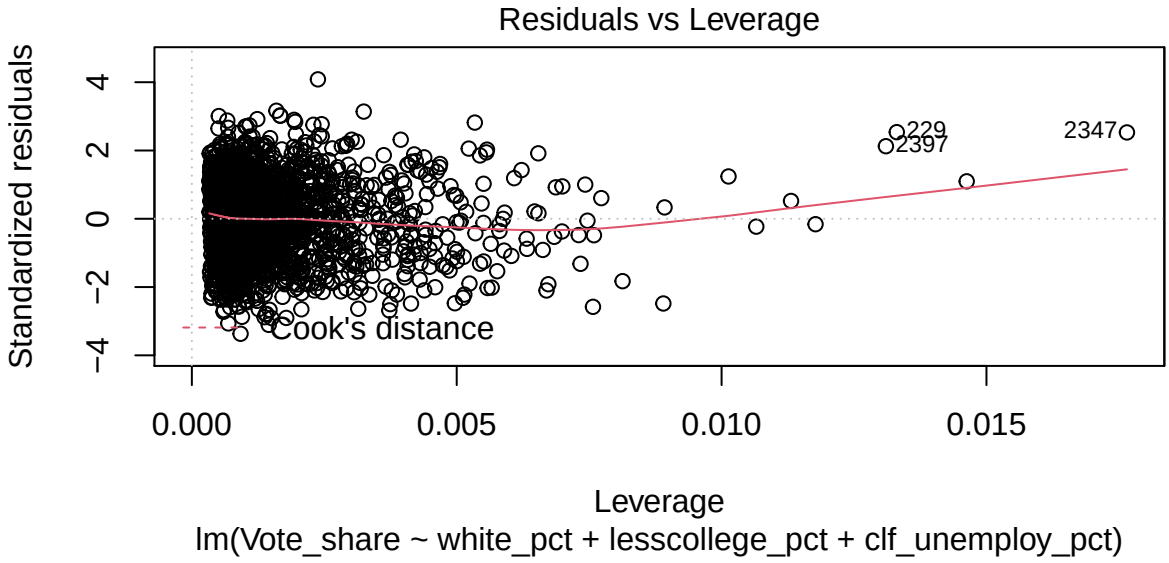
We ran a Shapiro-Wilks test to confirm our interpretation of the plot.

- Hypothesis
  - H_0 : Residuals follow normal distribution
  - H_a : Residuals do not follow normal distribution
- Conclusion (calculations available in the appendix)
  - The Shapiro test resulted in a p-value = 0.2981 $>$ $\alpha = 0.05$. Therefore, we did NOT reject H_0. We concluded the residuals followed a normal distribution and that the normality assumption was not violated.

Outliers in Data

Finally, we checked for the presence of outliers that could have a large influence on the fit of our model.

- Graphical tests: Residuals vs Leverage Plot



- Interpretation of Residuals vs Leverage Plot
  - In this plot there is no evidence of outliers. Those Cook's Distance dashed curves do not appear on the plot. This means that none of the points exhibited both high residual and leverage nor were they influential to the regression model.

Conclusion of Model Diagnostics

- In conclusion, Model 3 upheld the assumptions of Linearity and Normality, and did not contain any Influential cases. We did detect a small violation of the Constant Variance assumption. However, it did not appear egregious with respect to the size of our sample.

9 Summary

Based on the assumption that economic and demographic characteristics are important determinants of the voting patterns, we built a model to predict voting behavior in the presidential elections of the US. By experimenting with different variable selection processes and criteria, we designed a powerful and simple model, with only 3 explanatory variables: whites as a percentage of total population, percentage of unemployed population and population with an education of less than a bachelor's degree. We tested the model for the 2016 election and found that in almost 90% of all counties the predicted winner matched the actual winner of the election. This high predicting accuracy suggests that there is in fact a relation between demographics and voting behavior.

Our model also allowed us to compare between different explanatory variables to see which had the greatest impact on voting behavior. By comparing how much of the variability in the election results by county could be explained by each variable, we determined that "population with an education of less than a bachelor's degree" actually contributed more to the predicting power of the model than "percentage of white population". This suggests that researchers on the field, who have mainly been focused on the influence of race on voting patterns, should devote more time to studying the impact of education.

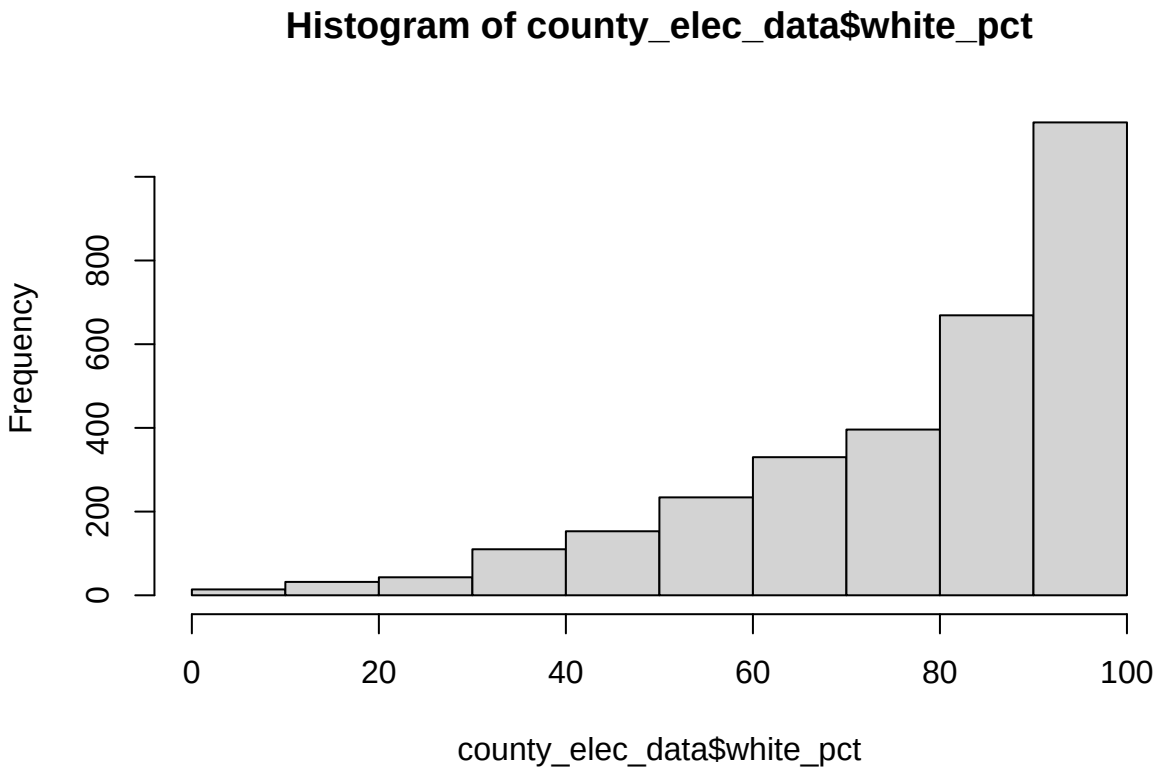
Since, as explained previously, the analysis presented here focused only on the 2016 presidential election, our model could and should be adjusted by replicating the analysis on the 2020 election results. Future researcher could also attempt to make longer run predictions: by analyzing census projections, they could anticipate how underlying demographic changes will affect the winning chances of the two major parties.

10 Appendices

10.1 Appendix A: Supplementary Descriptive Statistics

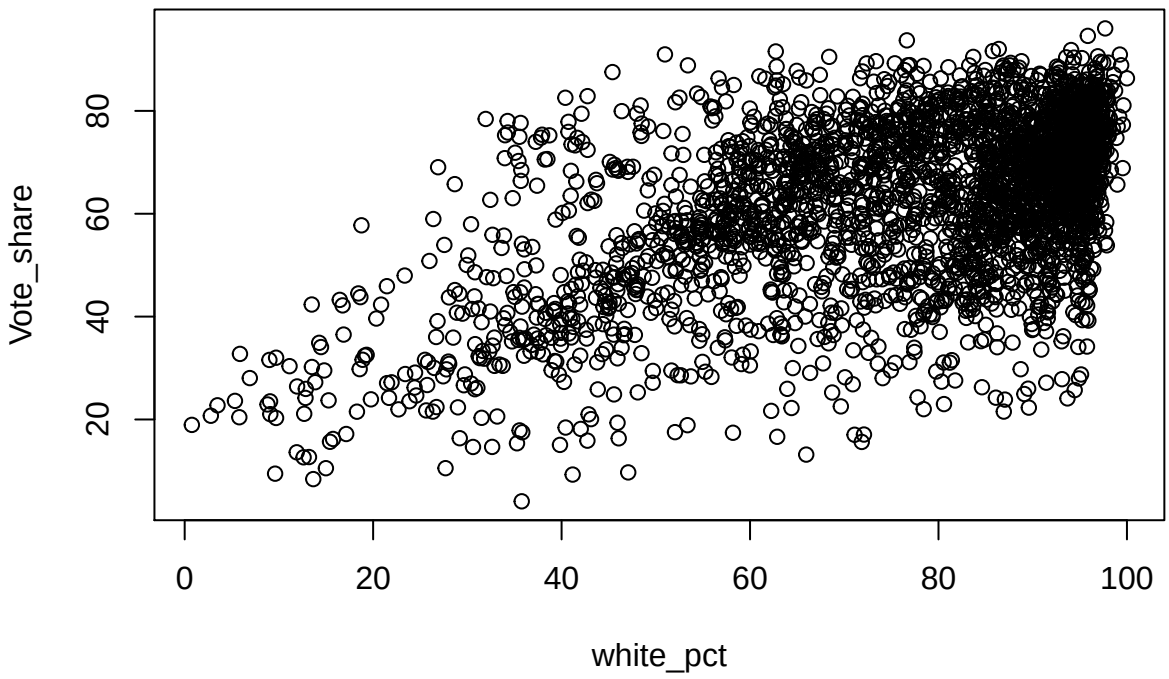
Evaluation of demographic predictor and relationship between response variable and demographic predictor

10.1.1 Plot 1



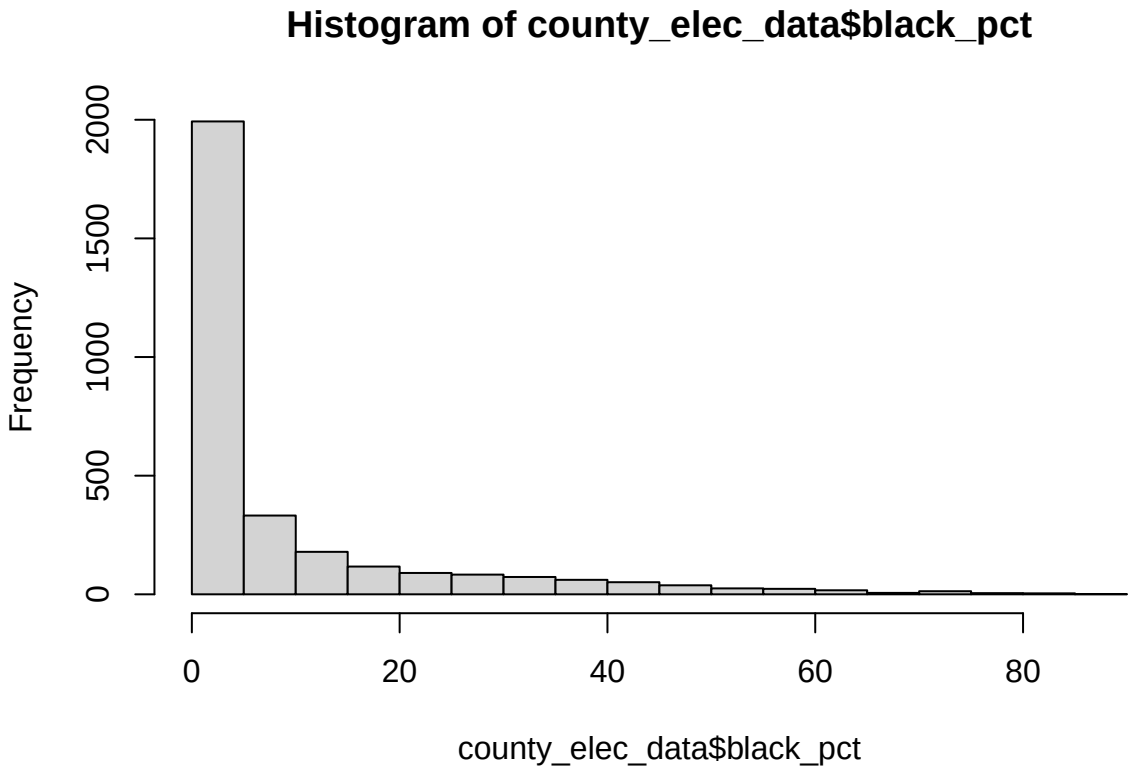
- The histogram is right-skewed. There are no gaps in the data.

10.1.2 Plot 2



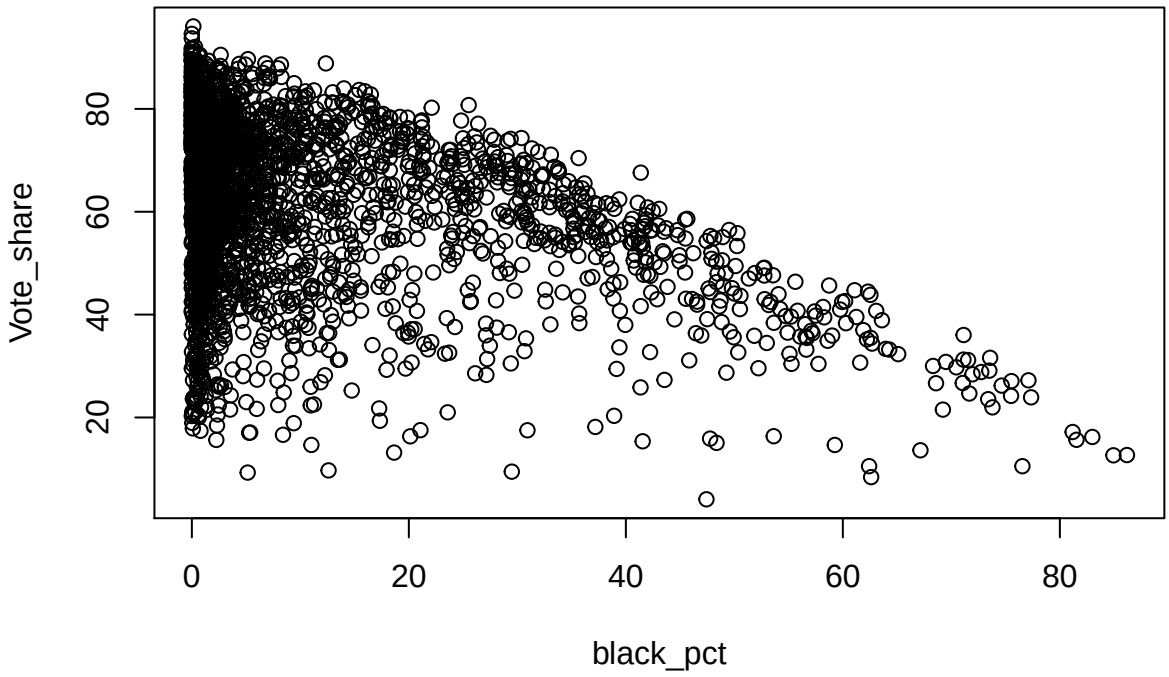
- According to this plot, there appeared to be a linear relationship between white_pct and Vote_share. Vote_share increased as white_pct increased. We also noticed that the variation in Vote_share increased as white_pct increased.
- We can infer from this association that counties that were majority white led to greater Vote_share for Donald Trump until the white_pct reached 90 and we saw a larger variation in the concentration of Vote_share.

10.1.3 Plot 3



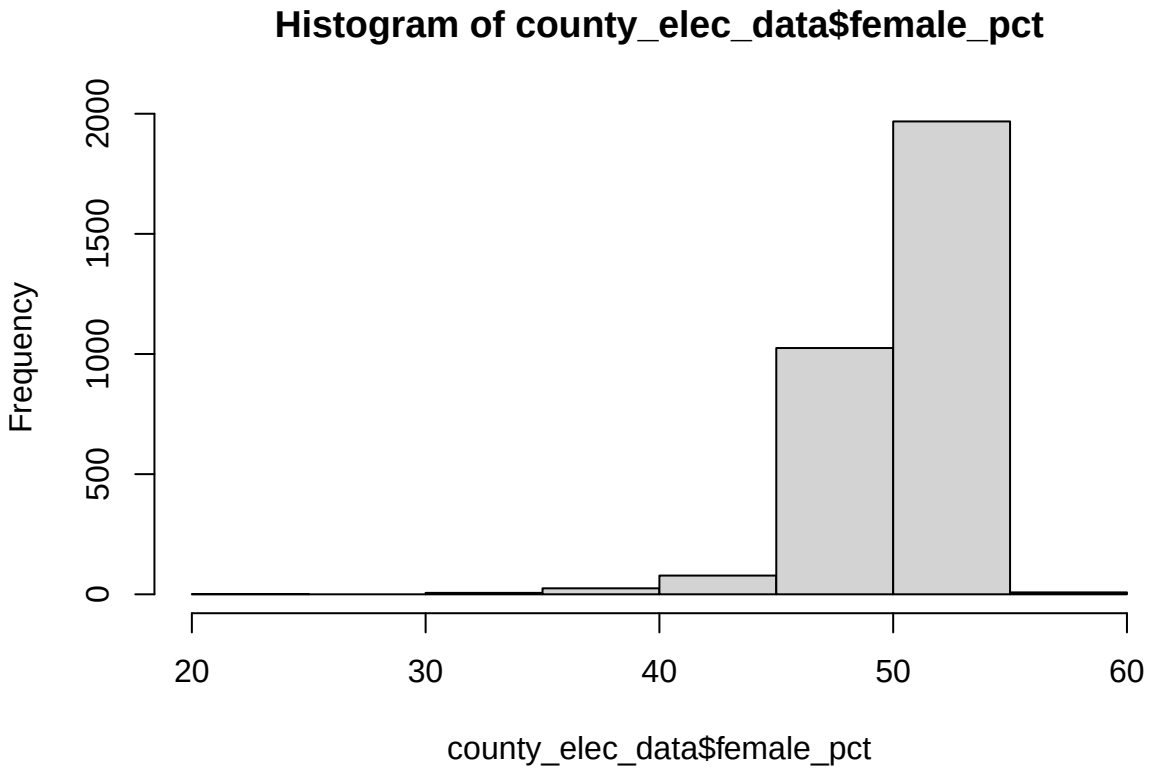
- The histogram is left-skewed. There are no gaps in the data.

10.1.4 Plot 4



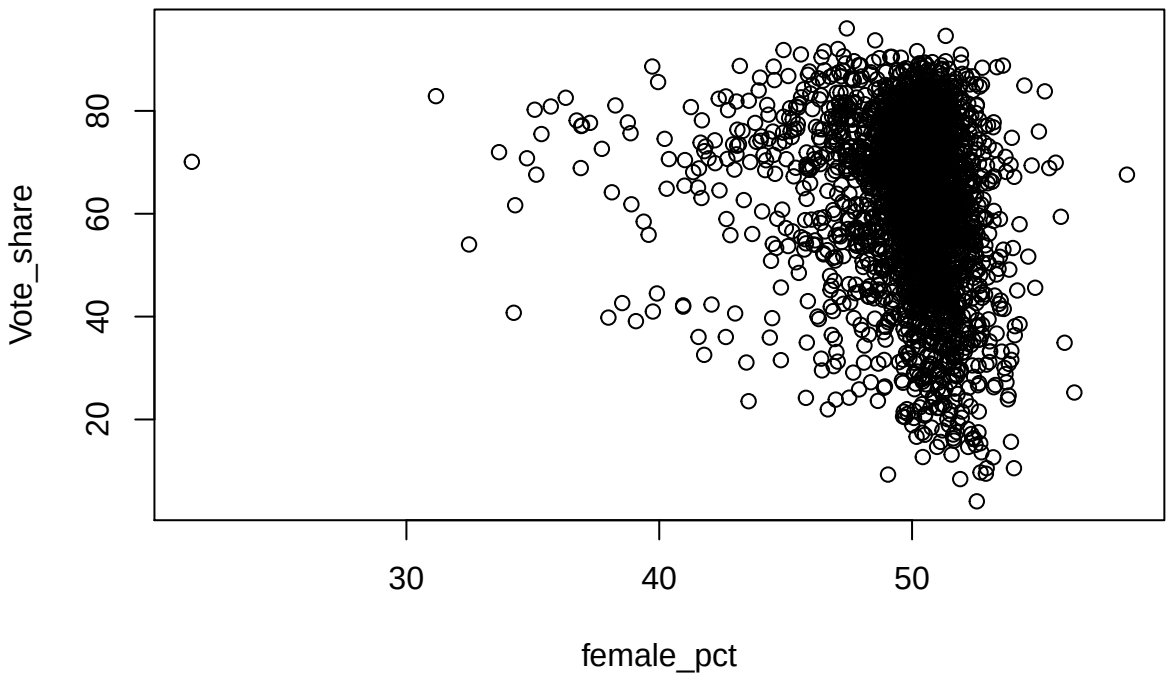
- The greatest variability in Vote_share occurred within counties that had near 0 black_pct. There is a negative relationship present between Vote_share and black_pct. As black_pct increased, Vote_share trended downwards steadily.

10.1.5 Plot 5



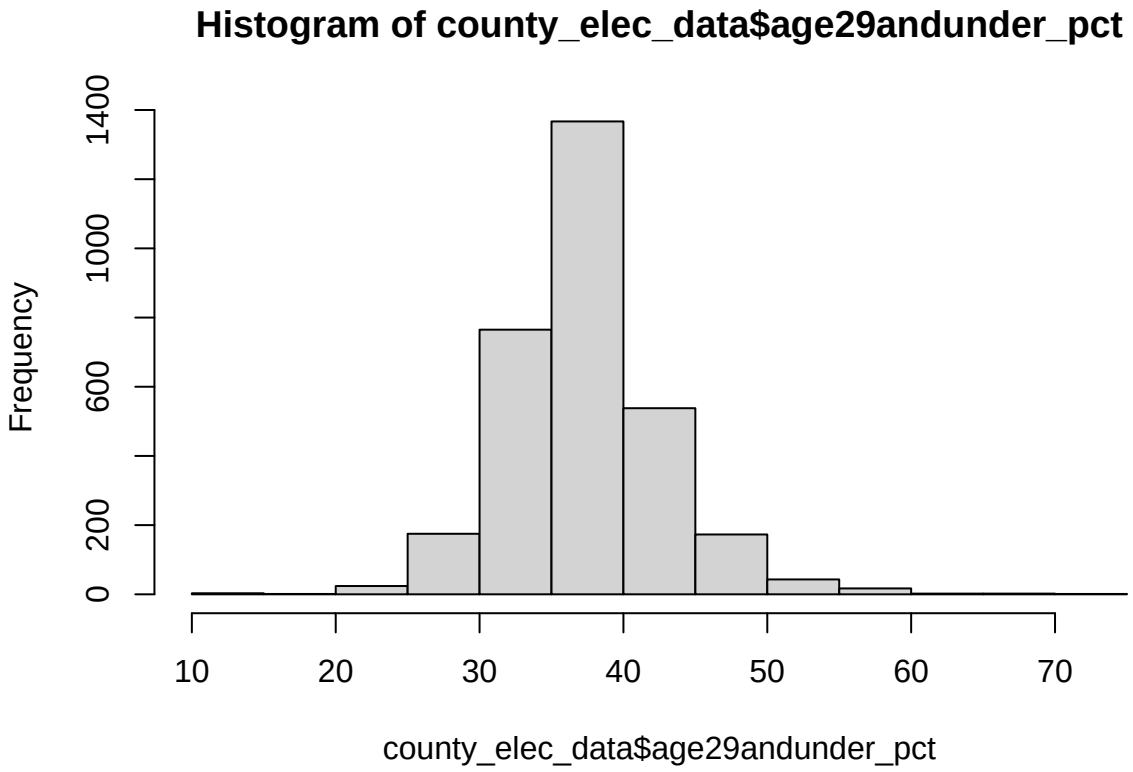
- The histogram is left-skewed. There are no gaps in the data.

10.1.6 Plot 6



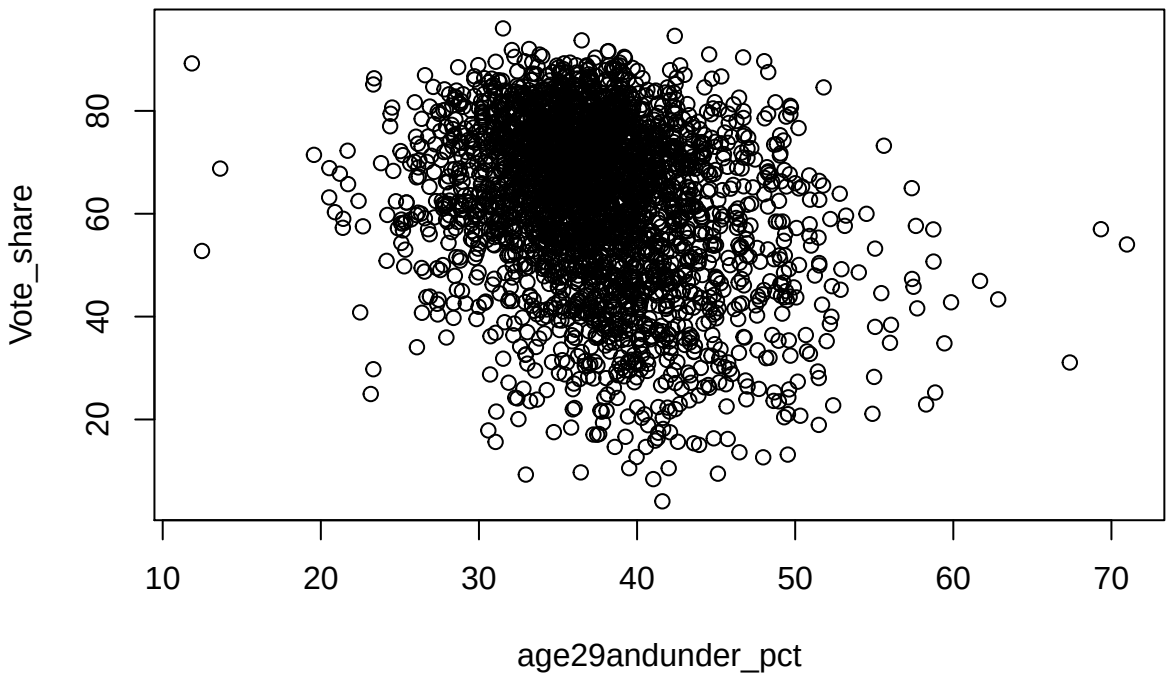
- According to this plot, counties with at least 50 percent of female voters led to large variation in Vote_share, evident by the dense concentration between 20 and 90 according to the y-axis. There was no linear association present between these two variables.
- There was little to no appreciable association between female_pct and Vote_share in counties where female_pct was under 30.

10.1.7 Plot 7



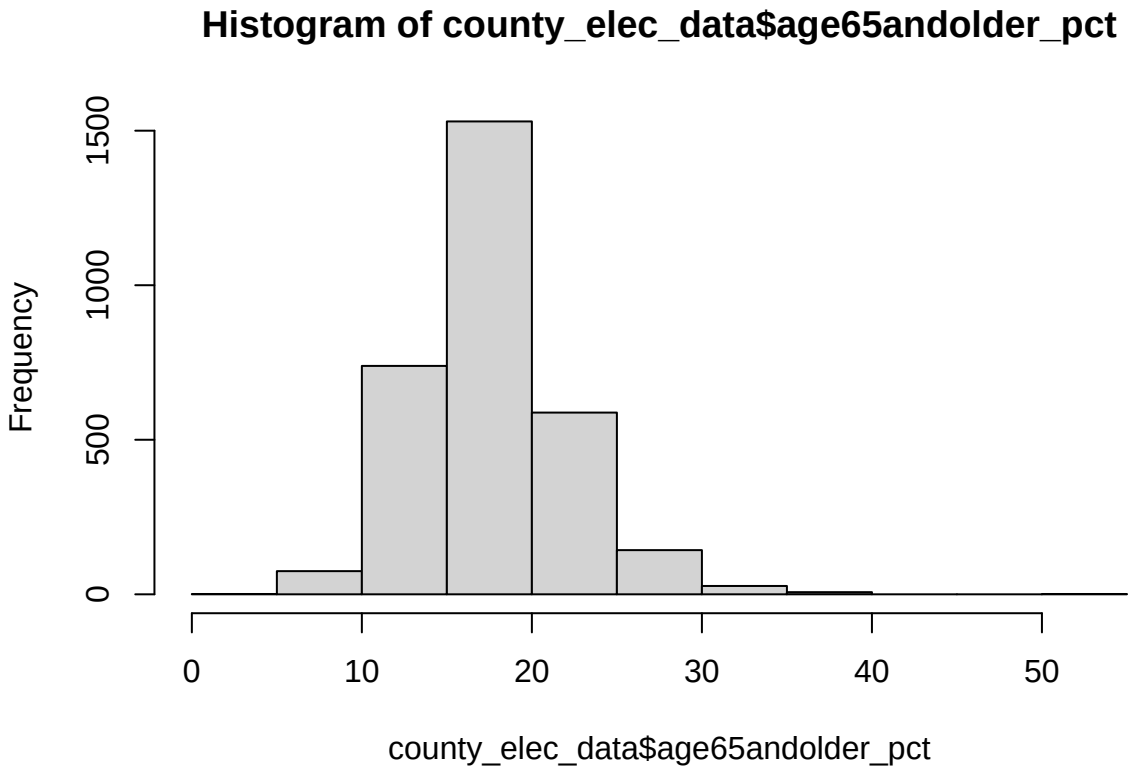
- The histogram is normally distributed. There is a gap on the left tail suggesting potential outliers.

10.1.8 Plot 8



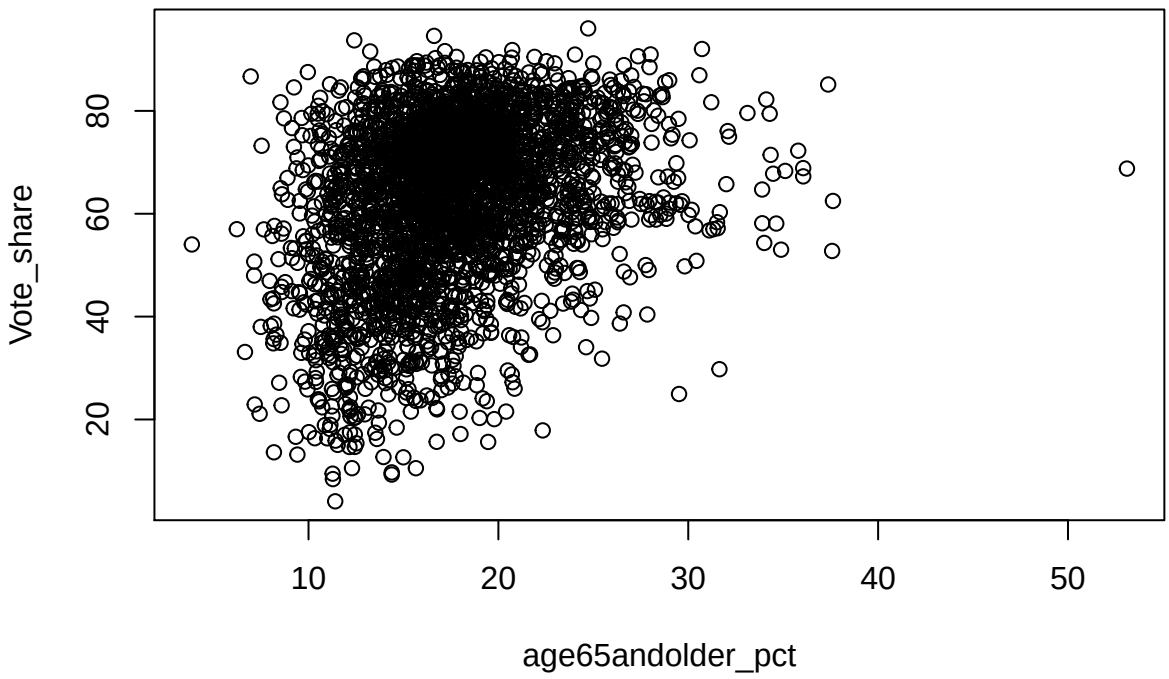
- There was no linear association between these two variables
- According to this plot counties with age29andunder_pct between 30 and 45 were associated with high Vote_share.

10.1.9 Plot 9



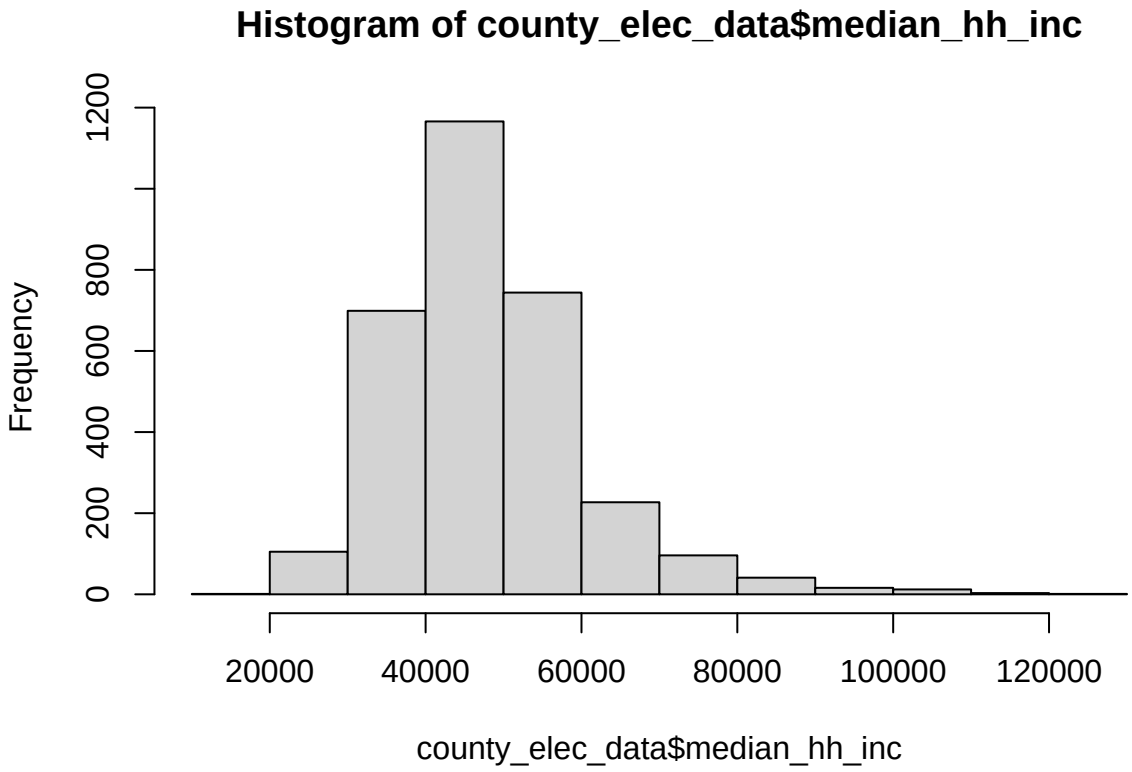
- The histogram is normally distributed. There are no gaps in the data.

10.1.10 Plot 10



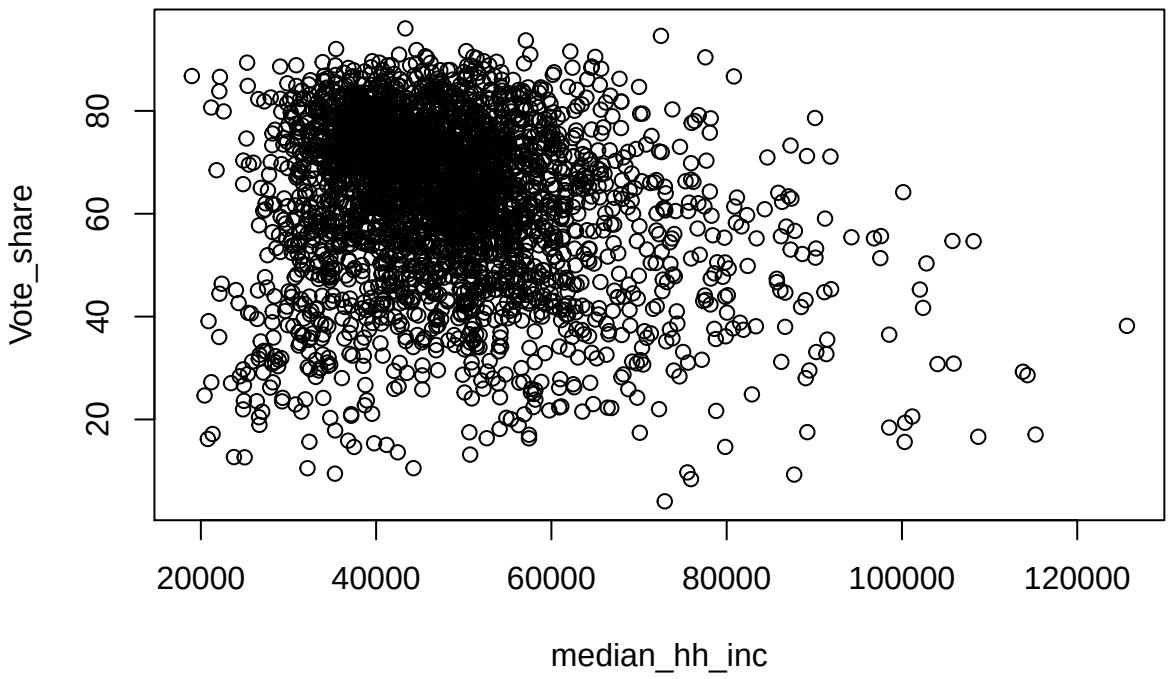
- There appeared to be a possible linear association between these two variables.
- According to this plot counties as age65andolder_pct increased, Vote_share increased up until about when age65andolder_pct reached 40; excluding the outlier past 50.

10.1.11 Plot 11



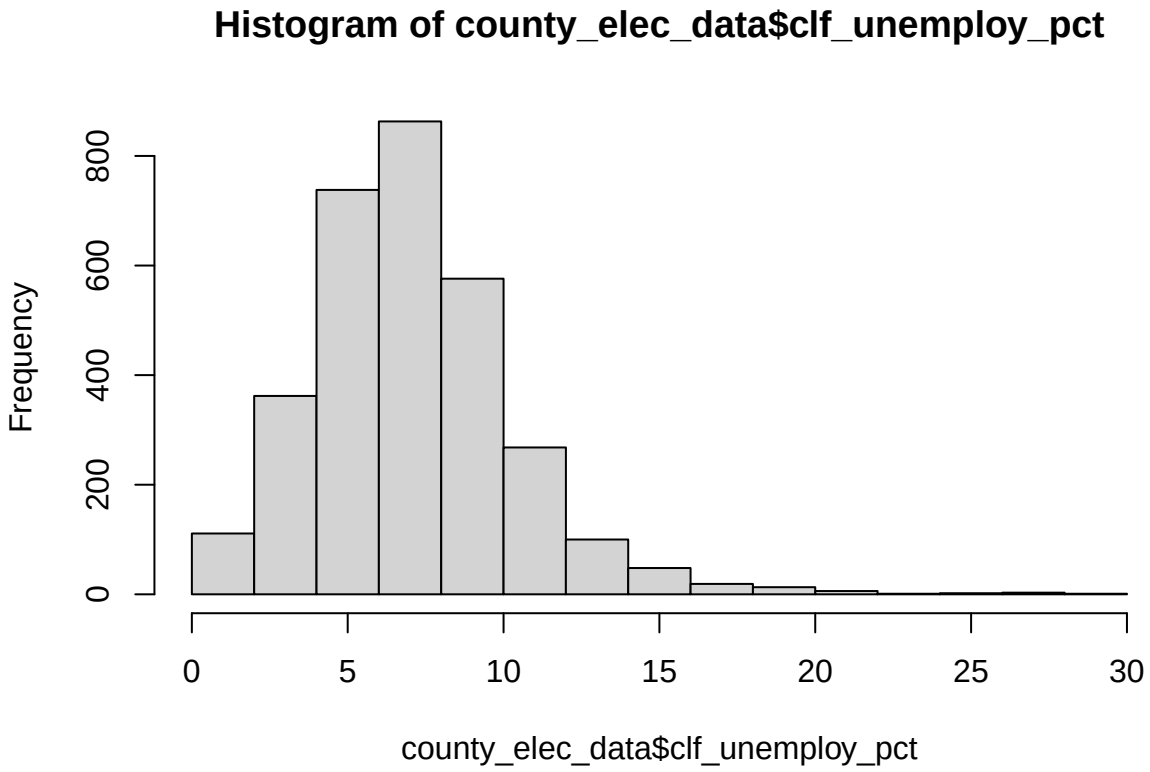
- The histogram is right-skewed. There are no gaps in the data.

10.1.12 Plot 12



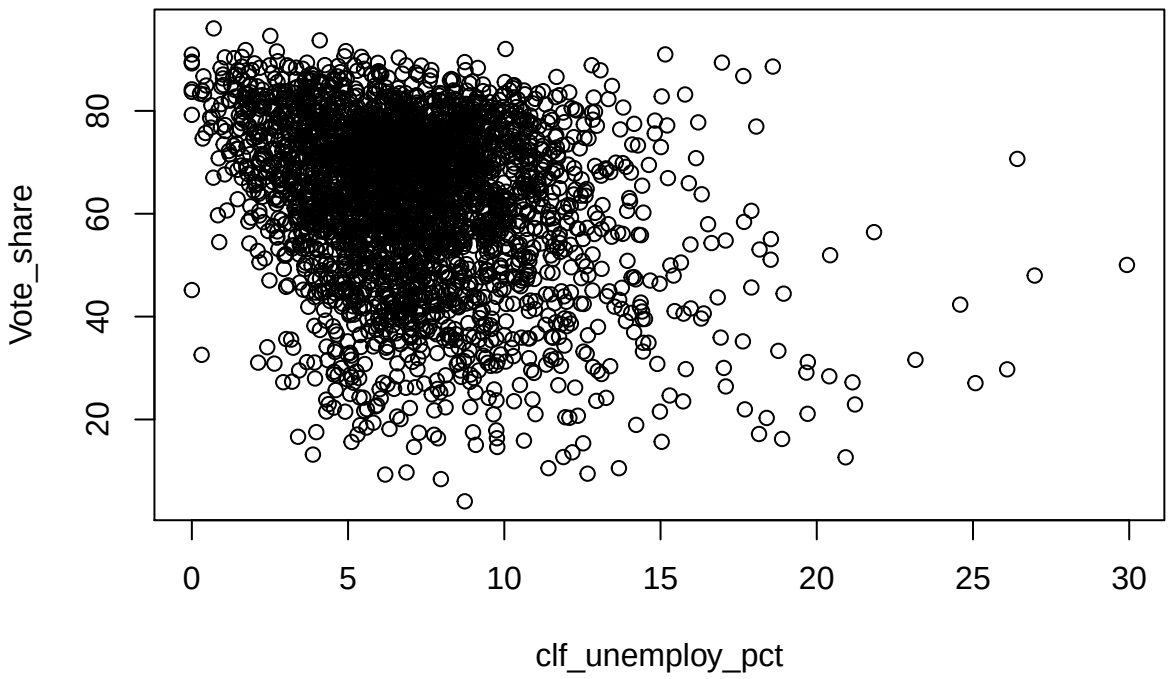
- There appeared to be a possible linear association between the two variables.
- Counties with median household income between 20,000 and 70,000 were associated with higher Vote_share.

10.1.13 Plot 13



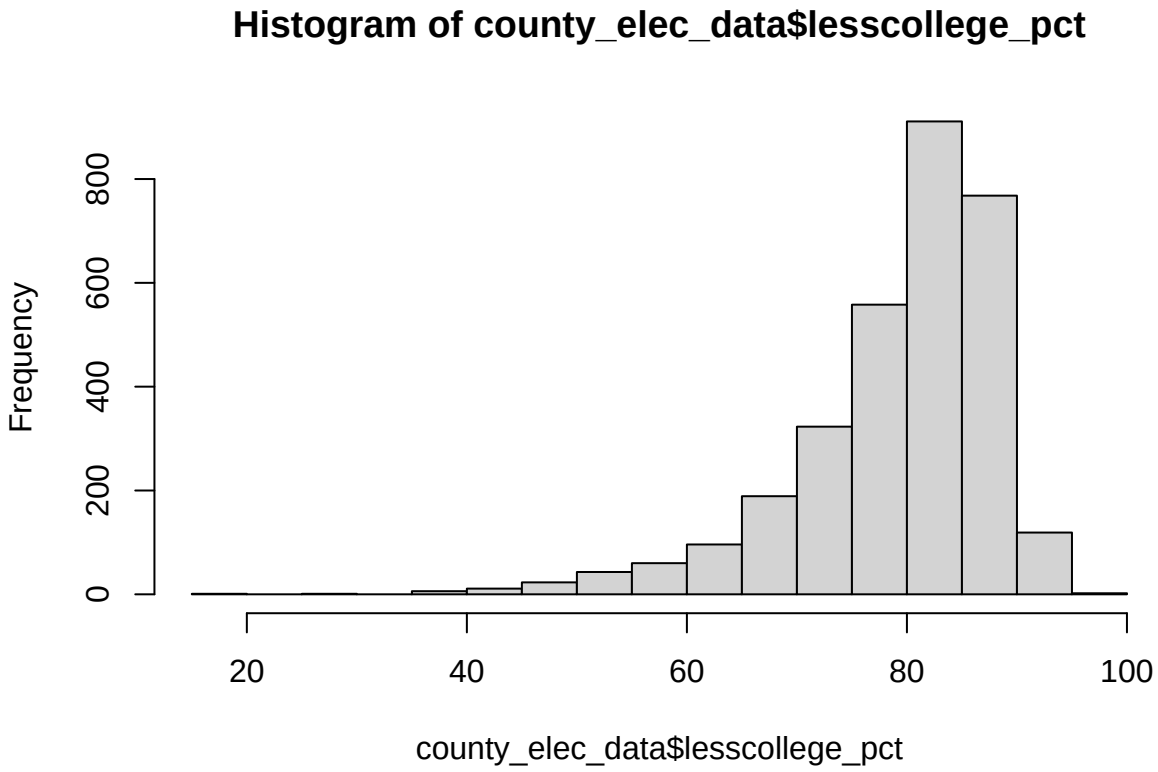
- The histogram is right-skewed. There are no gaps in the data.

10.1.14 Plot 14



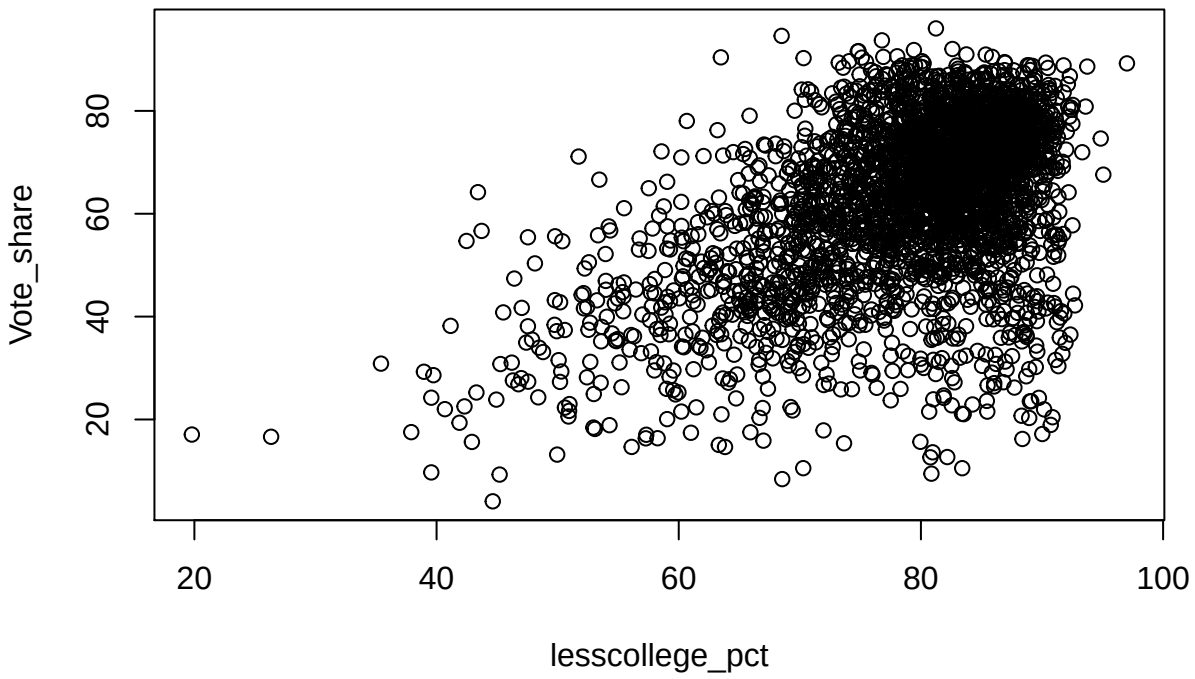
- There did not appear to be a linear association between the two variables.
- Counties with approximatly 3 to approximately 10 percent of the voters unemployed were associated with high Vote_share.

10.1.15 Plot 15



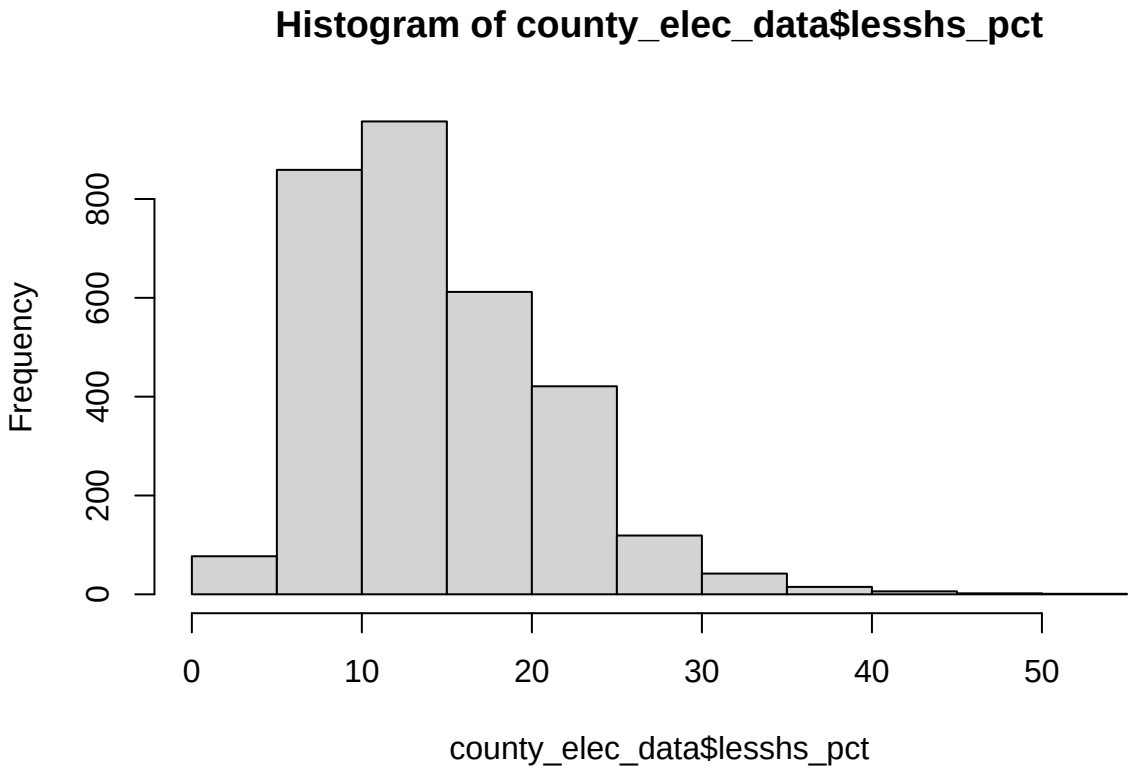
- The histogram is left-skewed. There is a gap on the left tail suggesting potential outliers.

10.1.16 Plot 16



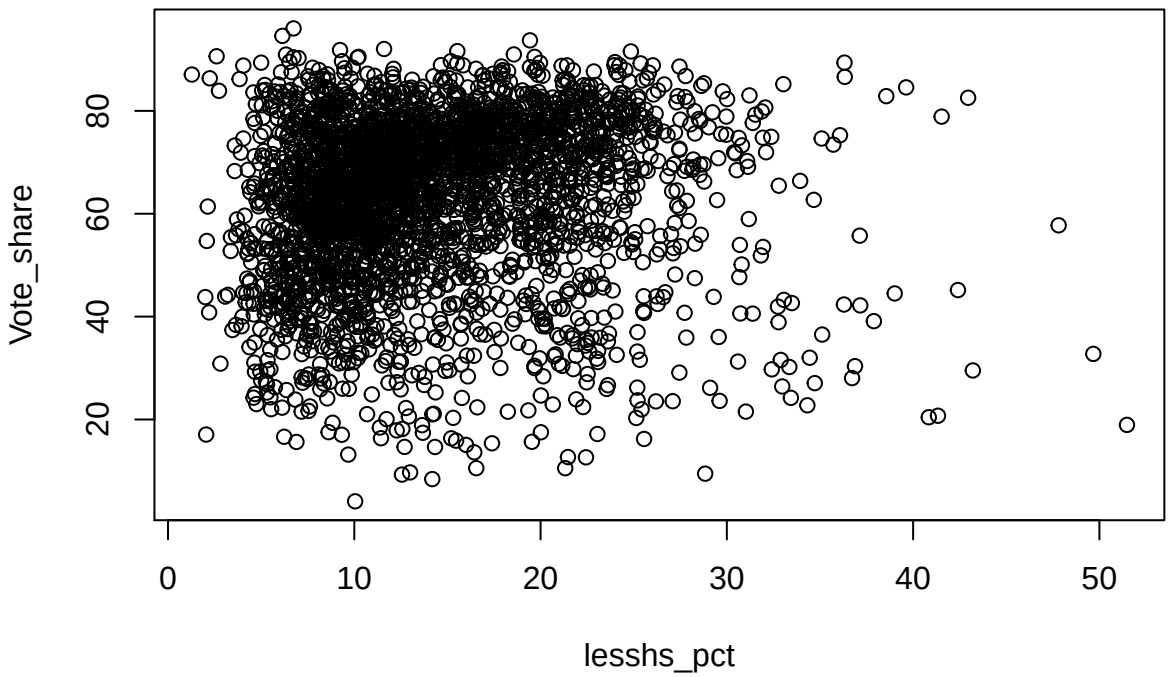
- According to this plot, there appeared to be a linear relationship between lesscollege_pct and Vote_share. Vote_share increased as lesscollege_pct within respective counties increased. We also noticed that the variation in Vote_share increased as lesscollege_pct increased.
- A point to take note of is it appeared not much Vote_share was captured until the percentage reached at least 40 at which point there was a steady increase in Vote_share. This led us to infer that counties where at least half the percentage of voters had less than a college degree were associated with a higher concentration of Vote_share for Donald Trump.

10.1.17 Plot 17



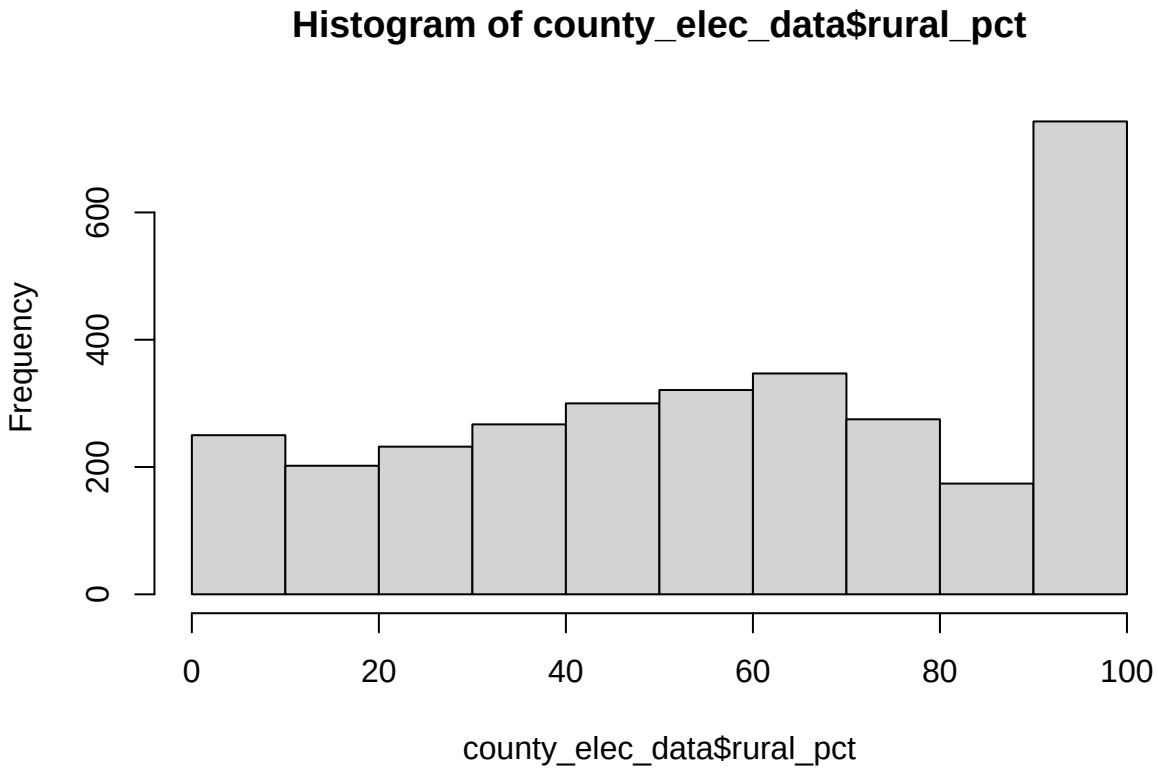
- The histogram is right-skewed. There are no gaps in the data.

10.1.18 Plot 18



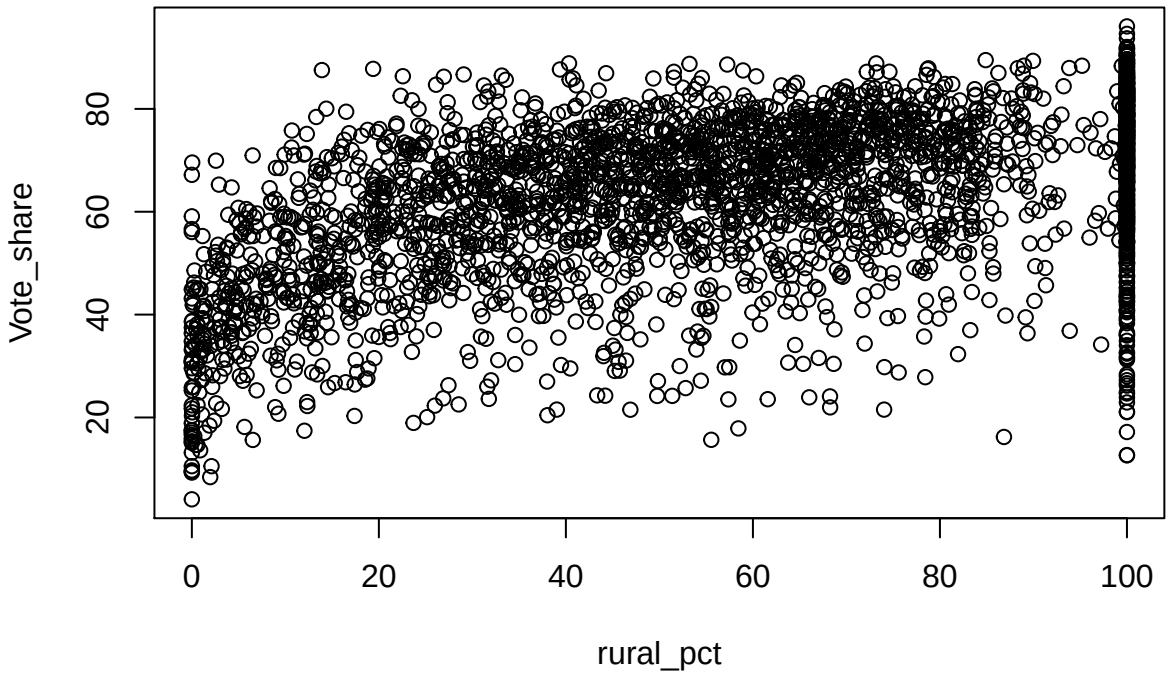
- There did not appear to be a linear association between the two variables.

10.1.19 Plot 19



- The histogram is left-skewed. There are no gaps in the data.

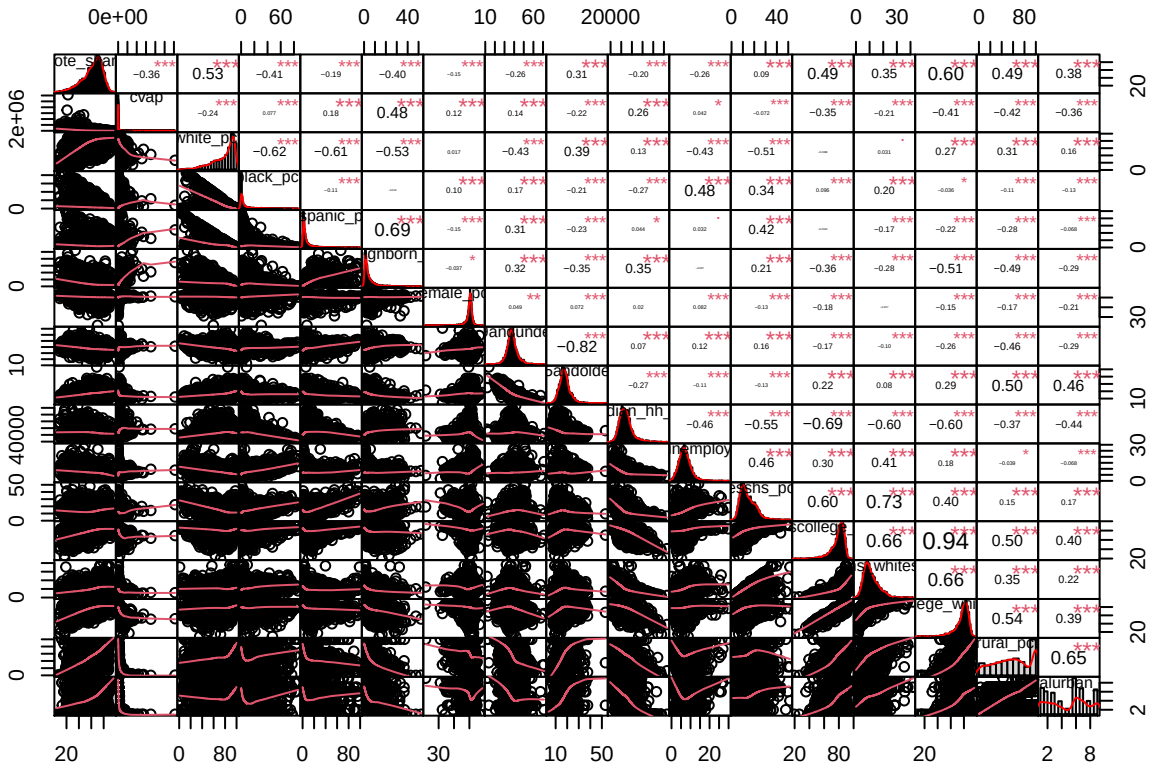
10.1.20 Plot 20



- There appeared to be a linear association between the two variables.
- The was high variability present at 0 and 100 on the x-axis.

Correlation matrix

10.1.21 Plot 21

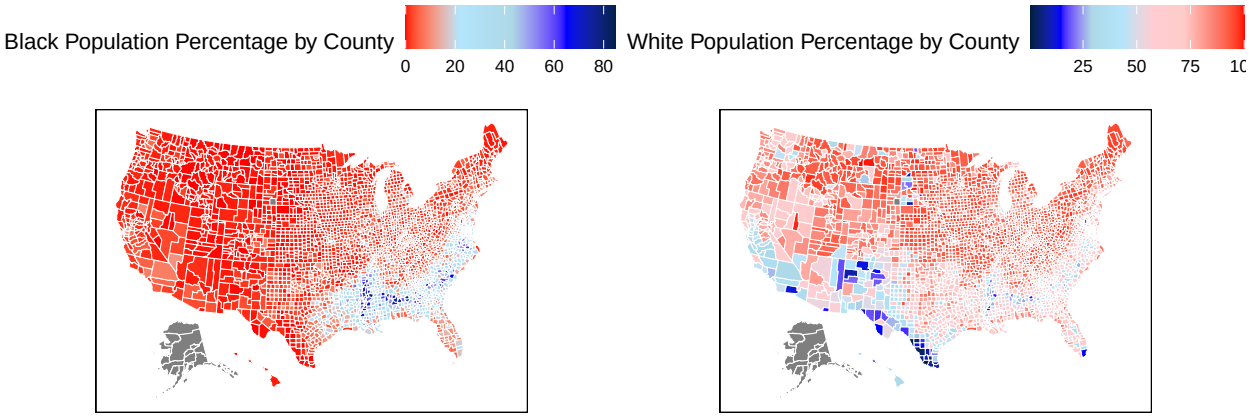


10.2 Appendix B: R Code for Model Fitting and Supplemental Visual Variable Analysis

10.2.1 Geographical Representation of racial variables.

- As you can also see from the maps below, counties with large white populations for the most part do not have large black populations, and vice-versa. This makes sense given that these two variables are negatively correlated to each other, with both subsets coming from the same demographic category.

10.2.2 Maps 1 & 2



10.2.3 Model with black population

```
##
## Call:
## lm(formula = Vote_share ~ black_pct + female_pct + age65andolder_pct +
##     median_hh_inc + lesshs_pct + rural_pct, data = county_elec_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.006  -6.341   2.022   8.036  31.666
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.575e+01  5.404e+00  12.167  < 2e-16 ***
## black_pct         -4.665e-01  1.688e-02 -27.644  < 2e-16 ***
## female_pct        -1.880e-01  9.758e-02  -1.927   0.0541 .
## age65andolder_pct  1.032e-01  6.229e-02   1.657   0.0977 .
## median_hh_inc     -1.236e-04  2.398e-05  -5.154 2.70e-07 ***
## lesshs_pct         3.039e-01  4.370e-02   6.954 4.32e-12 ***
## rural_pct          1.855e-01  8.595e-03  21.583  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.13 on 3104 degrees of freedom
## Multiple R-squared:  0.4038, Adjusted R-squared:  0.4026
## F-statistic: 350.3 on 6 and 3104 DF,  p-value: < 2.2e-16
```

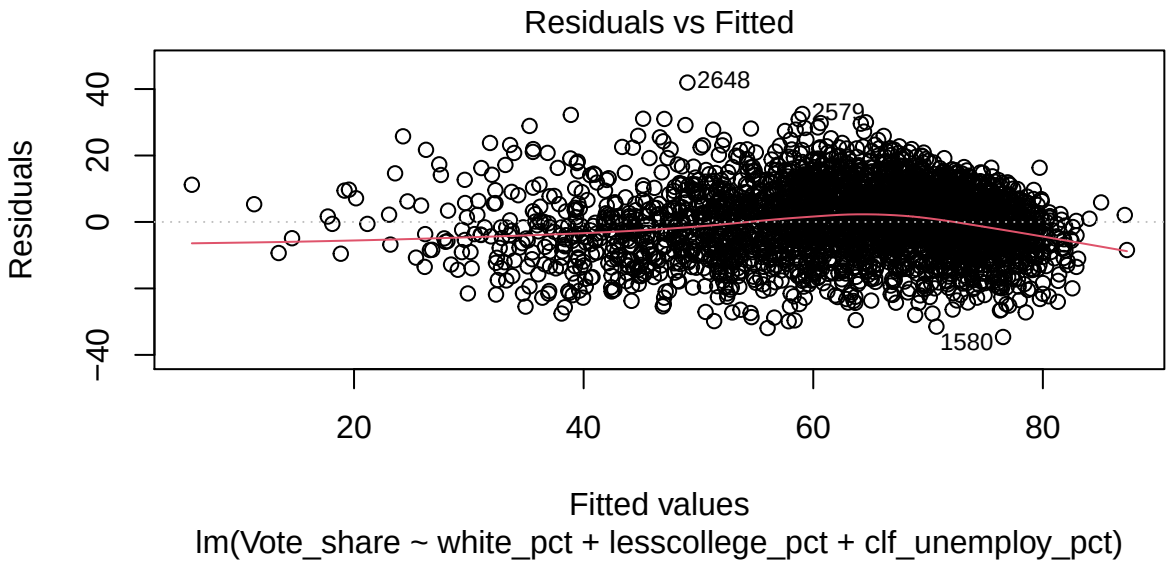
10.2.4 Model with Hispanic population

```
##
## Call:
## lm(formula = Vote_share ~ hispanic_pct + female_pct + age65andolder_pct +
##     median_hh_inc + lesshs_pct + rural_pct, data = county_elec_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.024  -7.664   2.350   9.477  38.774
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.198e+01  6.003e+00  11.991  < 2e-16 ***
## hispanic_pct      -1.392e-01  2.217e-02  -6.281 3.83e-10 ***
## female_pct        -6.765e-01  1.086e-01  -6.231 5.26e-10 ***
## age65andolder_pct  4.761e-01  6.832e-02   6.968 3.90e-12 ***
## median_hh_inc      6.302e-05  2.701e-05   2.333   0.0197 *
## lesshs_pct         2.768e-01  5.601e-02   4.942 8.13e-07 ***
## rural_pct          1.878e-01  1.004e-02  18.709  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.45 on 3104 degrees of freedom
## Multiple R-squared:  0.2663, Adjusted R-squared:  0.2649
## F-statistic: 187.8 on 6 and 3104 DF,  p-value: < 2.2e-16
```

10.3 Appendix D: Model 3 Diagnostics

10.3.1 Linearity

- Graphical test: Residuals vs Fitted Plot (e_i vs $\hat{Y}$)



- Interpretation of Residual vs Fitted Values Plot
  - This plot helped us evaluate the assumption of linearity. The red line looks relatively flat and does not deviate much from the dotted horizontal line of 0, meaning there does not appear to be a systematic pattern present. There is no clear violation of the linearity assumption. The plot was also useful in evaluating whether or not the assumption of homoscedasticity was violated. If there was constant variance about the line at 0 (homoscedasticity), the spread of residuals would be approximately the same across the x-axis. The plot shown above suggested there may be a violation of constant variance. We evaluated this by conducting a Brensch-Pagan test.

10.3.2 Constant Variance test

- Statistical test: Brensch-Pagan test
- Hypothesis
  - H_0 : The error variance is constant
  - H_a : The error variance is not constant
- Calculation

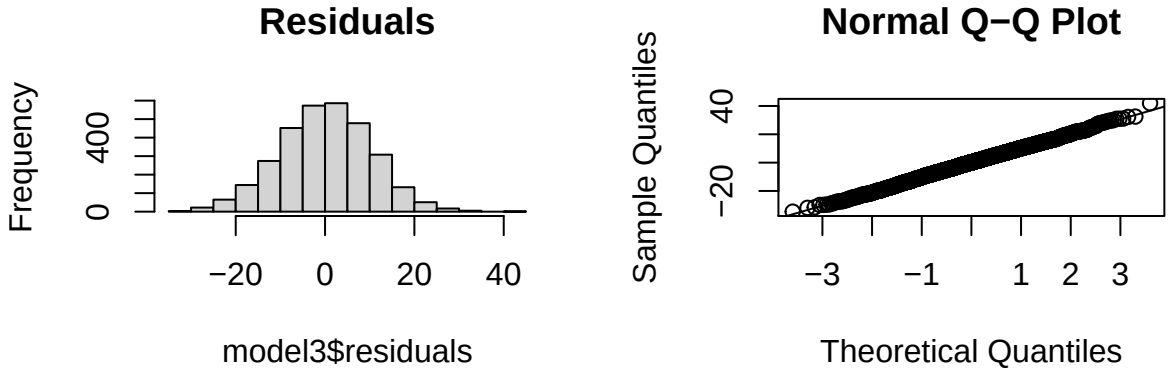
```
##
##  Breusch-Pagan test
##
## data:  model3
## BP = 118.68, df = 3, p-value < 2.2e-16
```

```
## [1] 7.814728
```

- Decision Rule
  - Chi-square distribution $X^2_{crit} = X^2(1-\alpha; 1)$.
  - If $X^2_{BP} \leq X^2_{crit}$, do not reject H_0.
  - If $X^2_{BP} > X^2_{crit}$, conclude H_a.
  - Alternatively, use p-value = $P(X^2_{df=1} > X^2_{BP})$.
- Conclusion
  - According to $X^2_{BP} = 118.68$ is larger than $X^2_{crit} = 7.814728$ and p-value = 2.2e-16 $< \alpha = 0.05$, we reject H_0. Therefore, as the plot suggested, the error variance is not constant.

10.3.3 Normality

- Graphical tests: Histogram and Q-Q plot



- The histogram plot of residuals displays a normal distribution and the dots on the Normal Q-Q plot are roughly scattered around the reference line randomly. There is minor deviation near the bottom left tail; however, it is not severe. This suggested the normality assumption was not grossly violated.

We ran a Shapiro-Wilks test to confirm our interpretation of the plot.

- Hypothesis
  - H_0 : Residuals follow normal distribution
  - H_a : Residuals do not follow normal distribution

- Calculation

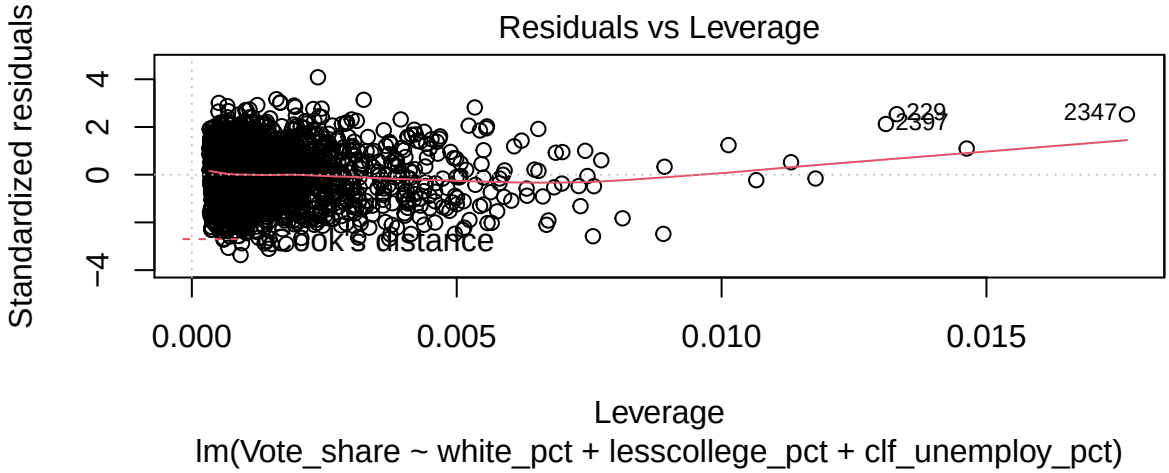
```
##
##  Shapiro-Wilk normality test
##
## data:  model3$residuals
## W = 0.99931, p-value = 0.2981
```

- Decision Rule
  - Reject H_0 if p-value $< \alpha$
- Conclusion
  - The Shapiro test resulted in a p-value = 0.2981 $> \alpha = 0.05$. Therefore, we did NOT reject H_0. We concluded the residuals followed a normal distribution and that the normality assumption was not violated.

10.3.4 Outliers in Data

Finally, we checked for the presence of outliers that could have a large influence on the fit of our model.

- Graphical tests: Residuals vs Leverage Plot



References

Barilla, A and Levernier, W. 2006. "The Effect of Region, Demographics, and Economic Characteristics on County-Level Voting Patterns in the 2000 Presidential Election. The Review of Regional Studies." https://www.researchgate.net/publication/254449508_The_Effect_of_Region_Demographics_and_Economic_Characteristics_on_County-Level_Voting_Patterns_in_the_2000_Presidential_Election.

Diggs, H and Farooq, M and Kidd, Q and Murrary, M. 2007. "Black Voters, Black Candidates, and Social Issues: Does Party Identification Matter?" https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6237.2007.00452.x?casa_token=qlk7bsncz14AAAAA%3ADXU_NK1M8Nhe5jay9P_ITel94ARlWoHidkYNoNpEEGlMc-zBlCd2PNB3rZhf1vx4E1KUCHtccLWJ

Gramlich, J. 2020. "What the 2020 Electorate Looks Like by Party, Race and Ethnicity, Age, Education and Religion." https://www.pewresearch.org/fact-tank/2020/10/26/what-the-2020-electorate-looks-like-by-party-race-and-ethnicity-age-education-and-religion/.

Hill, S and Hopkins, D and Huber, G. 2019. "Local Demographic Changes and Us Presidential Voting, 2012 to 2016." https://www.pnas.org/content/116/50/25023.