DATA 606 Final Project

Zach Alexander

I. INTRODUCTION

Labeled by many news outlets as the "biggest upset in U.S. history", a large narrative about a "divided" America continued to develop in the days following the 2016 presidential election. A week after the election, The Upshot by the New York Times published an article that examined the stark differences in votes by candidate relative to where voters lived in the U.S. – it became obvious very quickly that many living in rural areas cast their vote for Republican candidate, Donald Trump, while many living in urban centers cast their vote for the Democratic candidate, Hilary Clinton. Many would argue that this "divided America" narrative continues to dominate current news headlines and will be an influential factor in the way candidates run their campaigns over the next 12 months. With 2020 presidential candidates currently pooring hundreds of thousands of dollars into rural states, vying for their votes, we can see just how pivotal voters in these regions could be for the outcome of the election next year.

For my project, I thought it would be interesting to take a look back at the election data from 2016 in order to think about the urban/rural divide and it's implications for the 2020 election.

Research question(s)

> Was there a significant difference in the mean proportion of GOP votes in the 2016 presidential election based on county type (Rural, Suburban, or Urban)?
>
> Is educational attainment and/or ruralness predictive of the proportion of GOP votes by county in the 2016 presidential election?

The "urban/rural divide" and educational differences will be big talking points over the next 12 months. As a way to explore these talking points further, I'll see if there are large differences based on these factors, and if so, determine whether or not it possible to build some type of predictive model for the 2020 election.

II. DATA

Data collection

In order to prepare the data for analysis, I had to do a bit of tidying by selecting the relevant columns from the raw datasets and recoding some of the education and ruralness data from numeric to categorical variables. By using `ifelse` statements and the `tidyverse` package I was able to accomplish this:

```
# The dataset has many columns, I'm subsetting down to only columns of interest
election_results_df <- election_results_df %>%
  dplyr::select(combined_fips, county_name, state_abbr, per_dem, per_gop)
```

```
# I'm taking the rural/urban continuum codes and creating a separate variable
# with the qualitative values
education_data_df <- education_data_df %>%
  dplyr::select(fips, lesscollege_pct, ruralurban_cc) %>%
```

```
  mutate(ruralurban_grp = ifelse(ruralurban_cc == 1,
  "Counties in metro areas of 1 million population or more",
  ifelse(ruralurban_cc == 2,
         "Counties in metro areas of 250,000 to 1 million population",
  ifelse(ruralurban_cc == 3,
         "Counties in metro areas of fewer than 250,000 population",
  ifelse(ruralurban_cc == 4,
         "Urban population of 20,000 or more, adjacent to a metro area",
  ifelse(ruralurban_cc == 5,
         "Urban population of 20,000 or more, not adjacent to a metro area",
  ifelse(ruralurban_cc == 6,
         "Urban population of 2,500 to 19,999, adjacent to a metro area",
  ifelse(ruralurban_cc == 7,
         "Urban population of 2,500 to 19,999, not adjacent to a metro area",
  ifelse(ruralurban_cc == 8,
         "Completely rural or less than 2,500 urban population, adjacent to a metro area",
  ifelse(ruralurban_cc == 9,
         "Completely rural or less than 2,500 urban population, adjacent to a metro area",
         NA))))))))))
```

Since we have the common identifier of FIPS county codes across both datasets, I was able to merge the 2016 election data with the education and ruralness data:

```
# since the FIP codes are found in both data frames, I can use merge to
# join the information from both data frames into one data frame while
# creating a few columns to calculate the party winner by county as well
# as the proportion of individuals in each county with a bachelor's degree or more

election_education_df <- merge(election_results_df, education_data_df,
                               by.x = "combined_fips", by.y = "fips" ) %>%
  mutate(college_or_more_pct = ((100 - lesscollege_pct)/ 100),
         party_winner = ifelse(per_gop > per_dem, 'Republican', 'Democrat')) %>%
  dplyr::select(combined_fips:per_gop, college_or_more_pct, party_winner, ruralurban_cc,
         ruralurban_grp)
```

Here's a snapshot of some of the combined dataset, ready for analysis:

combined_fips	county_name	state_abbr	per_gop	college_or_more_pct	ruralurban_grp
1001	Autauga County	AL	0.7343579	0.2459277	Counties in metro areas of 250,000 to 1 million population
1003	Baldwin County	AL	0.7735147	0.2954711	Counties in metro areas of fewer than 250,000 population
1005	Barbour County	AL	0.5227141	0.1286779	Urban population of 2,500 to 19,999, adjacent to a metro area
1007	Bibb County	AL	0.7696616	0.1200000	Counties in metro areas of 1 million population or more
1009	Blount County	AL	0.8985188	0.1304976	Counties in metro areas of 1 million population or more
1011	Bullock County	AL	0.2422889	0.1025501	Urban population of 2,500 to 19,999, adjacent to a metro area
1013	Butler County	AL	0.5631549	0.1608001	Urban population of 2,500 to 19,999, adjacent to a metro area
1015	Calhoun County	AL	0.6923970	0.1765297	Counties in metro areas of fewer than 250,000 population
1017	Chambers County	AL	0.5663376	0.1248428	Urban population of 2,500 to 19,999, adjacent to a metro area
1019	Cherokee County	AL	0.8387127	0.1396170	Urban population of 2,500 to 19,999, adjacent to a metro area
1021	Chilton County	AL	0.8254177	0.1485455	Counties in metro areas of 1 million population or more
1023	Choctaw County	AL	0.5643919	0.1196020	Completely rural or less than 2,500 urban population, adjacent to a metro area
1025	Clarke County	AL	0.5495516	0.1213723	Urban population of 2,500 to 19,999, not adjacent to a metro area
1027	Clay County	AL	0.7958004	0.1106288	Completely rural or less than 2,500 urban population, adjacent to a metro area
1029	Cleburne County	AL	0.8784446	0.1152807	Completely rural or less than 2,500 urban population, adjacent to a metro area
1031	Coffee County	AL	0.7714620	0.2366672	Urban population of 20,000 or more, adjacent to a metro area
1033	Colbert County	AL	0.6788760	0.1847090	Counties in metro areas of fewer than 250,000 population
1035	Conecuh County	AL	0.5216262	0.0866258	Urban population of 2,500 to 19,999, not adjacent to a metro area
1037	Coosa County	AL	0.6463718	0.0994056	Completely rural or less than 2,500 urban population, adjacent to a metro area
1039	Covington County	AL	0.8358832	0.1492149	Urban population of 2,500 to 19,999, adjacent to a metro area
1041	Crenshaw County	AL	0.7215291	0.1464768	Completely rural or less than 2,500 urban population, adjacent to a metro area
1043	Cullman County	AL	0.8781050	0.1496902	Urban population of 20,000 or more, adjacent to a metro area
1045	Dale County	AL	0.7411506	0.1606870	Urban population of 20,000 or more, adjacent to a metro area
1047	Dallas County	AL	0.3088094	0.1378430	Urban population of 20,000 or more, adjacent to a metro area
1049	DeKalb County	AL	0.8348923	0.1136594	Urban population of 2,500 to 19,999, adjacent to a metro area
1051	Elmore County	AL	0.7483810	0.2223840	Counties in metro areas of 250,000 to 1 million population
1053	Escambia County	AL	0.6758693	0.1162565	Urban population of 2,500 to 19,999, adjacent to a metro area
1055	Etowah County	AL	0.7391084	0.1649185	Counties in metro areas of fewer than 250,000 population
1057	Fayette County	AL	0.8180820	0.1414535	Urban population of 2,500 to 19,999, adjacent to a metro area
1059	Franklin County	AL	0.7918026	0.1239010	Urban population of 2,500 to 19,999, adjacent to a metro area
1061	Geneva County	AL	0.8548761	0.1166179	Counties in metro areas of fewer than 250,000 population
1063	Greene County	AL	0.1723571	0.0997485	Completely rural or less than 2,500 urban population, adjacent to a metro area
1065	Hale County	AL	0.3960050	0.1402874	Counties in metro areas of fewer than 250,000 population
1067	Henry County	AL	0.7013846	0.1614434	Counties in metro areas of fewer than 250,000 population
1069	Houston County	AL	0.7272662	0.2101790	Counties in metro areas of fewer than 250,000 population
1071	Jackson County	AL	0.8007217	0.1257846	Urban population of 2,500 to 19,999, adjacent to a metro area
1073	Jefferson County	AL	0.4502208	0.3144025	Counties in metro areas of 1 million population or more
1075	Lamar County	AL	0.8395614	0.1306261	Completely rural or less than 2,500 urban population, adjacent to a metro area
1077	Lauderdale County	AL	0.7145802	0.2167923	Counties in metro areas of fewer than 250,000 population
1079	Lawrence County	AL	0.7339123	0.1030862	Counties in metro areas of fewer than 250,000 population

The education percentages were calculated by Stephen Pettigrew at Harvard University's Dataverse. It appears that the percent with a college degree or more by county population are calculations based on a dataset created by IPUMS NHGIS, University of Minnesota. The rural/urban continuum codes (ruralurban_cc) variable, was adopted from the United States Department of Agriculture Economic Research Service in 2013. These variables were combined by Stephen Pettigrew into one dataset, which also stored other types of election data - outside the scope of this investigation.

Cases

The cases are the number of counties in the United States (the id is FIPS code) that have educational attainment percentages, ruralness continuum codes, and voting percentages available from the 2016 election. There are 3112 counties in this dataset that contain this information out of a total of 3141 counties in the United States. Although we are missing some county-level data, this dataset still comprises about 99% of the counties that make up the United States.

Type of study

This is an observational study.

Data Source

I found the dataset with 2016 election results on GitHub:

- 2016 Election Results

I found the dataset with the education and ruralness data on GitHub as well:

- Education & Ruralness Data

Although this second dataset houses 2018 election data, I am only using the American Community Survey education data (5-year estimates).

Dependent Variable

The response variable is `per_gop` (proportion of GOP vote out of total votes by county) and it is quantitative.

Independent Variable

My two independent variables are `college_or_more_pct` (quantitative) and `ruralurban_grp` (qualitative).

Scope of Inference – generalizability

Population of interest – the individuals that voted for a presidential candidate in the 2016 election.

Because we are specifically looking at election data in this project, and whether or not someone votes is not completely random, we cannot generalize these findings to all voting-age individuals that were in the U.S. during the time of the 2016 presidential election. In this study, we'll merely be looking at differences within our population of voters and not attempting to generalize outside of this subset to all voting-age individuals that were in the U.S. during the time of the 2016 presidential election.

Some areas of bias that would prevent generalizability to the broader voting-age population are:

- Whether or not someone has access to a polling station
- Whether or not someone has accurate information about where to go to vote
- Whether or not someone is a justice-involved individual
- Whether or not someone was intimidated or harrassed prior to voting
- Whether or not someone was subjected to strict voter ID and ballot requirements

And many other items. As you can see, many areas of bias would prohibit us from generalizing our findings to all voting-age individuals.

Scope of Inference – causality

Given that this is an observational study, and we are merely observing the data that arise from the 2016 election, we **cannot** imply causation. Causation can only be inferred from a randomized experiment.

III. EXPLORATORY DATA ANALYSIS

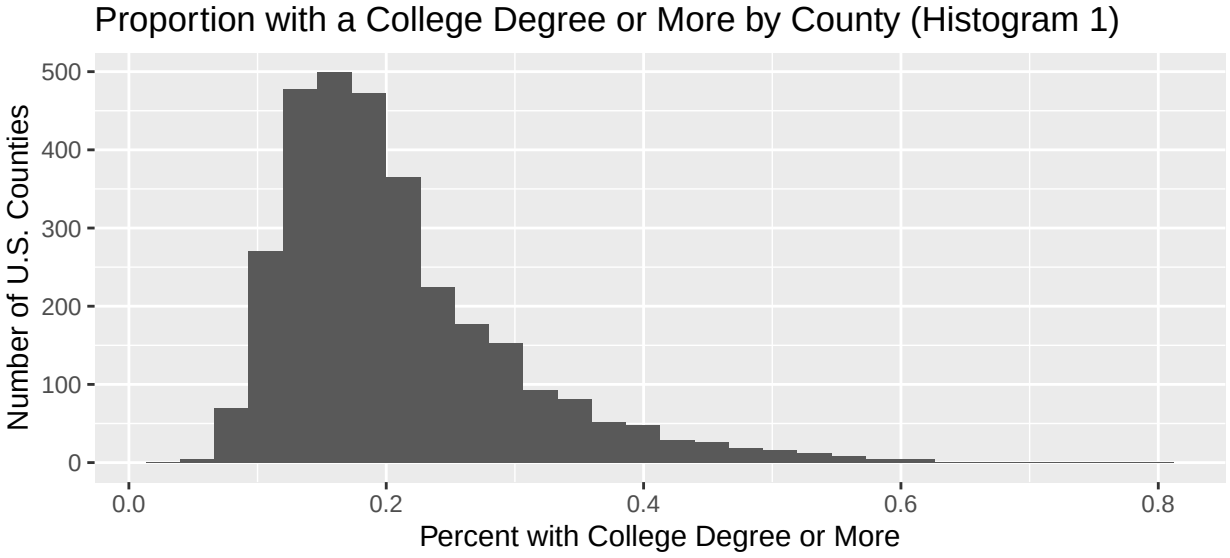
Relevant summary statistics

First, I thought it would be interesting to take a quick look at how many counties Donald Trump won in 2016:

Var1	Freq
Democrat	487
Republican	2625

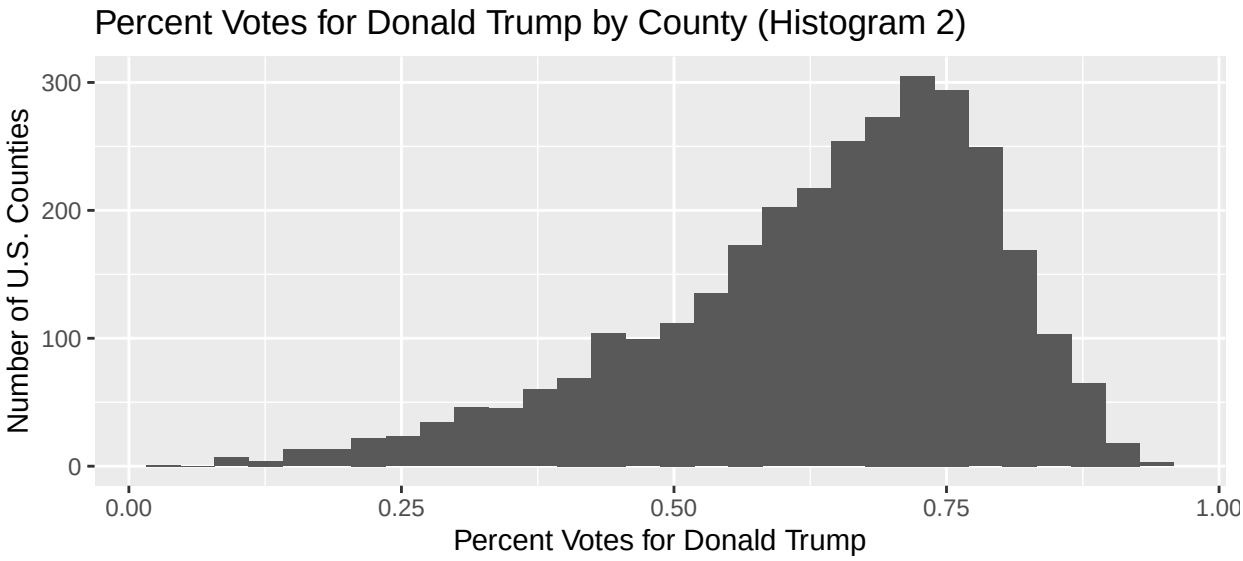
We can see that when solely looking at the vote proportions by county, Donald Trump won a much larger percentage of U.S. counties in 2016 (albeit, this doesn't take into account population size, just the number of counties). It was great to see that this split was confirmed by the Associated Press (helps to check that my data merges and wrangling weren't prone to errors - the AP reports one extra "county" in Louisiana going to Donald Trump to make their count at 2626, but this extra "county" is a parish in Lousiana, so it's up for interpretation).

Then, I wanted to get a better sense of the data, so I decided to plot a histogram of the percent of each county's population that has a college degree or higher.



	vars	n	mean	sd	median	min	max	skew	kurtosis	se
X1	1	3111	0.2078352	0.0913877	0.1852531	0.0298507	0.8021012	1.520067	3.096609	0.0016385

Next, I thought it would be helpful to plot a histogram of the percent of each county that voted for Donald Trump in 2016.



	vars	n	mean	sd	median	min	max	skew	kurtosis	se
X1	1	3112	0.6361341	0.1565173	0.667431	0.0412207	0.9527273	-0.8364531	0.3906748	0.0028057

To take a look at my rural variable, I first thought it would be helpful to see some summary statistics related to my qualitative variable `ruralurban_grp`, broken down by county:

	group1	n	mean	sd	se
X11	Completely rural or less than 2,500 urban population, adjacent to a metro area	627	0.7155089	0.1409475	0.0056289
X12	Counties in metro areas of 1 million population or more	432	0.5355758	0.1775423	0.0085420
X13	Counties in metro areas of 250,000 to 1 million population	376	0.5806173	0.1447696	0.0074659
X14	Counties in metro areas of fewer than 250,000 population	354	0.6137479	0.1395124	0.0074150
X15	Urban population of 2,500 to 19,999, adjacent to a metro area	593	0.6630747	0.1318980	0.0054164
X16	Urban population of 2,500 to 19,999, not adjacent to a metro area	425	0.6729490	0.1431233	0.0069425
X17	Urban population of 20,000 or more, adjacent to a metro area	214	0.6060651	0.1260898	0.0086193
X18	Urban population of 20,000 or more, not adjacent to a metro area	91	0.6062969	0.1546884	0.0162157

I may be able to compare means across these 8 different groups, but I also thought it would be helpful to subset this data a bit more based on broader categories of ruralness. I decided to recode the ruralurban codes so that those that are characteristic of urban areas (`ruralurban_cc` = 1 - 3) would be grouped together, those characteristic of rural areas (`ruralurban_cc` = 7 - 9) would be grouped together, and those characteristic of suburban areas (`ruralurban_cc` = 4 - 6) would be grouped together. This'll hopefully make mean comparisons easier later on. I saved these recodes in a new variable called `ruralurban_grp_3_way`.

```
# recoding the ruralurban_grp and ruralurban_cc variable into a 3_way grouping variable
election_education_df$ruralurban_grp_3_way <- ifelse(
  election_education_df$ruralurban_cc <= 3, 'Urban Counties',
  ifelse(election_education_df$ruralurban_cc >= 4 &
        election_education_df$ruralurban_cc < 7, 'Suburban Counties',
  ifelse(election_education_df$ruralurban_cc >= 7, 'Rural Counties', NA)))
```

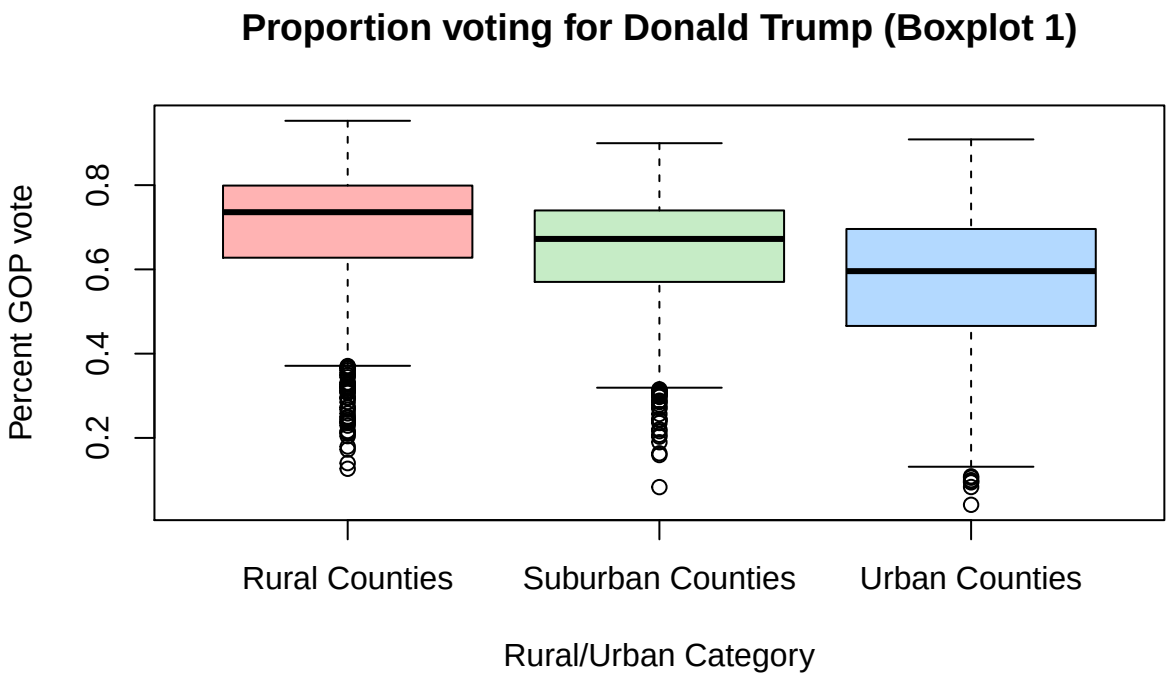
After this was set up, I could then look at summary statistics of the `per_gop` variable, based on these new ruralness types:

	group1	n	mean	sd	median	min	max	skew	kurtosis	se
X11	Rural Counties	1052	0.6983151	0.1432941	0.7357744	0.1267168	0.9527273	-1.2389639	1.5434319	0.0044179
X12	Suburban Counties	898	0.6437352	0.1355943	0.6723085	0.0832182	0.8995612	-0.9857139	0.9619270	0.0045248
X13	Urban Counties	1162	0.5739652	0.1594815	0.5958650	0.0412207	0.9085546	-0.5956667	-0.0557926	0.0046785

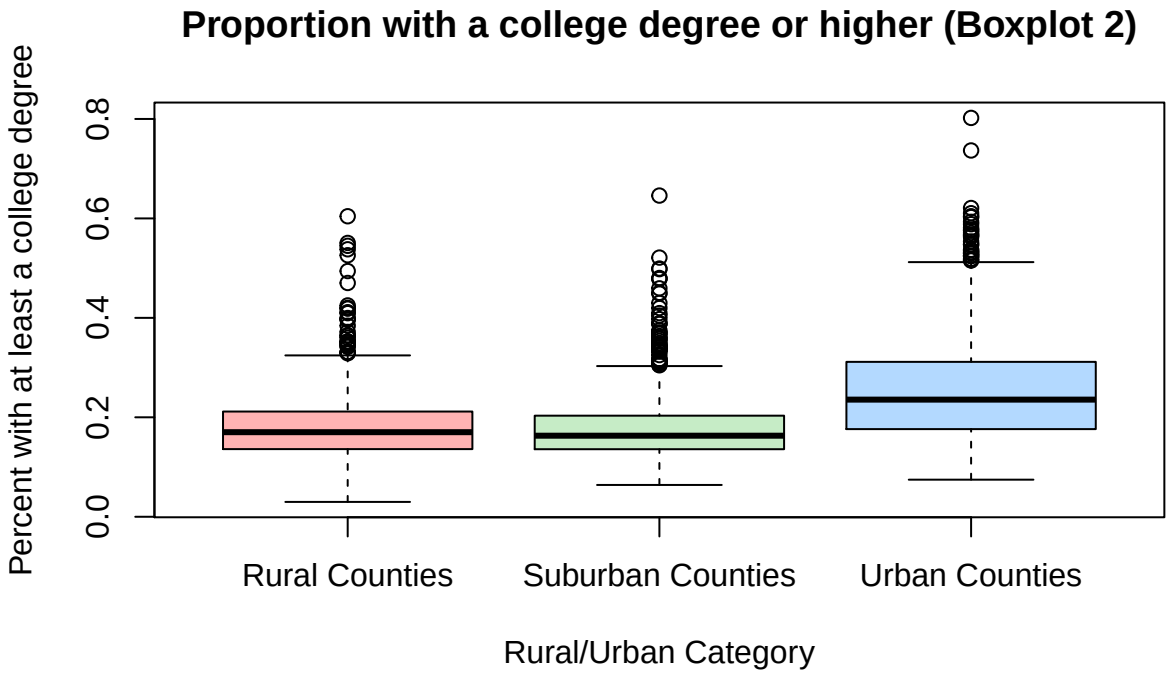
And the `college_or_more_pct` variable split by this new variable.

	group1	n	mean	sd	median	min	max	skew	kurtosis	se
X11	Rural Counties	1052	0.1799575	0.0659897	0.1701078	0.0298507	0.6043459	1.6415236	5.665799	0.0020346
X12	Suburban Counties	897	0.1792921	0.0681997	0.1630310	0.0639714	0.6459119	1.9812014	6.103309	0.0022771
X13	Urban Counties	1162	0.2551077	0.1061605	0.2355707	0.0745510	0.8021012	0.9741119	1.152331	0.0031143

Here are these two tables as box plots. See below for the box plot of the of per_gop split by ruralness (3-way):

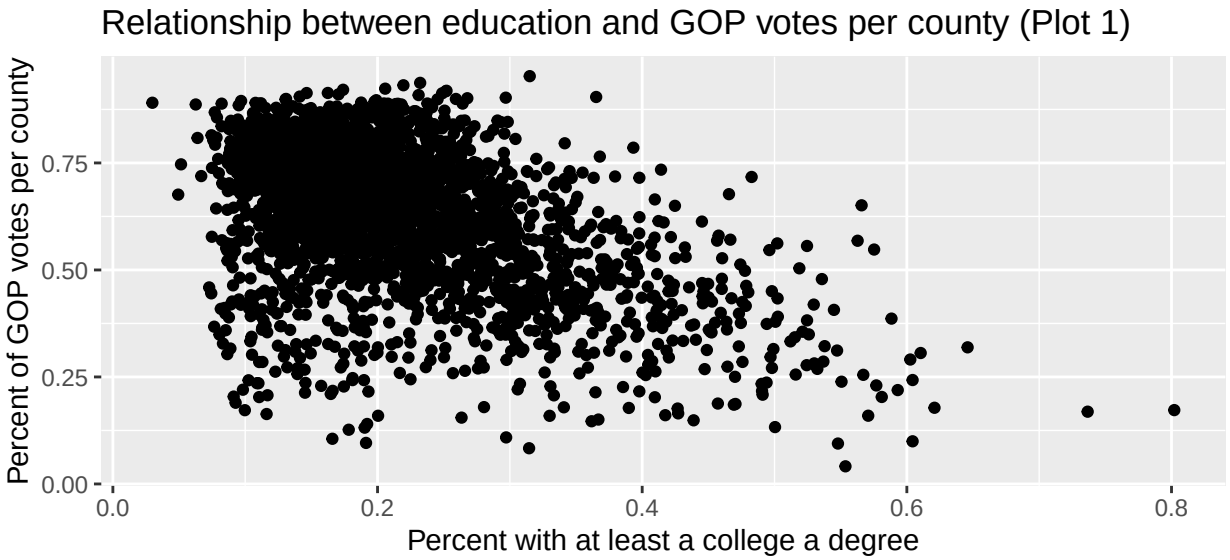


And, here is a boxplot for percent with a bachelor's degree or higher by ruralness (3-way).



Although this second boxplot is a bit outside the scope of the research question, I thought it would be interesting to see if there are any noticeable differences in educational attainment across the ruralness continuum as well to build a bit more context.

Finally, we'll look at the relationship between the percent votes for Donald Trump per county in 2016, and the percent with a college degree or more:



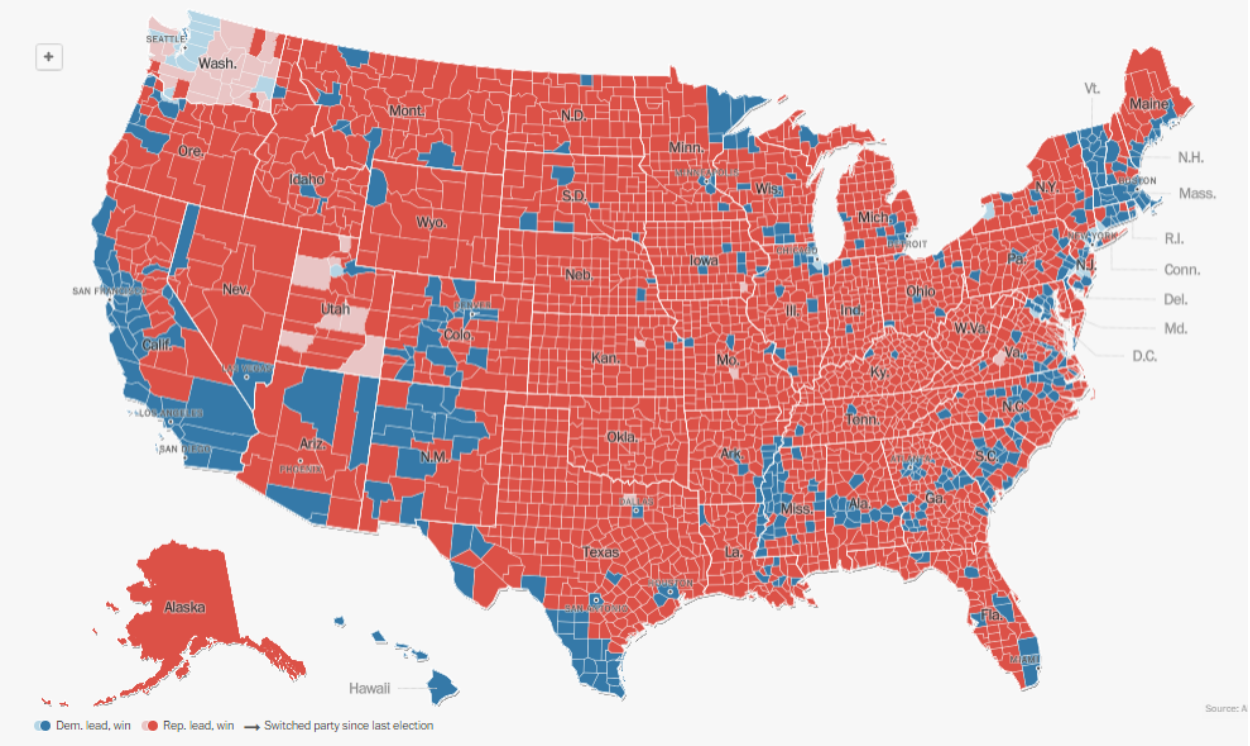


Figure 1:

Summary of Exploratory Analysis

From the exploratory data analysis we can see the following:

- It appears that the percentage of individuals that voted with a college degree or more has a right-skewed, unimodal distribution. We can see that a majority of counties have around 20% of their voting population identify as having a college degree or more – in other words, for a majority of counties, 1 in 5 voters tended to be college-educated in 2016 (Histogram 1).
- It also appears that a majority of counties voted for Donald Trump in 2016 (Figure 1). From our distribution, it appears to be left-skewed, with a majority of counties with an electorate that had a vote share for Donald Trump around 75% – meaning he won big on a *county* level. As a reminder, this does not take into consideration population size or the number of votes, but just the proportion of Trump voters per county (Histogram 2).
- After recoding the rural variable to a 3-way grouping variable and plotting the boxplot above, we can see noticeable differences in the percent of GOP votes per county based on ruralness.
- Finally, after plotting the relationship between my education variable (`college_or_more_pct`) and the percent of GOP votes per county (`per_gop`), we can see a large clustering of instances of counties that have about 20% of their population with college degree or more and their GOP vote proportions from 60-75%. (Plot 1).

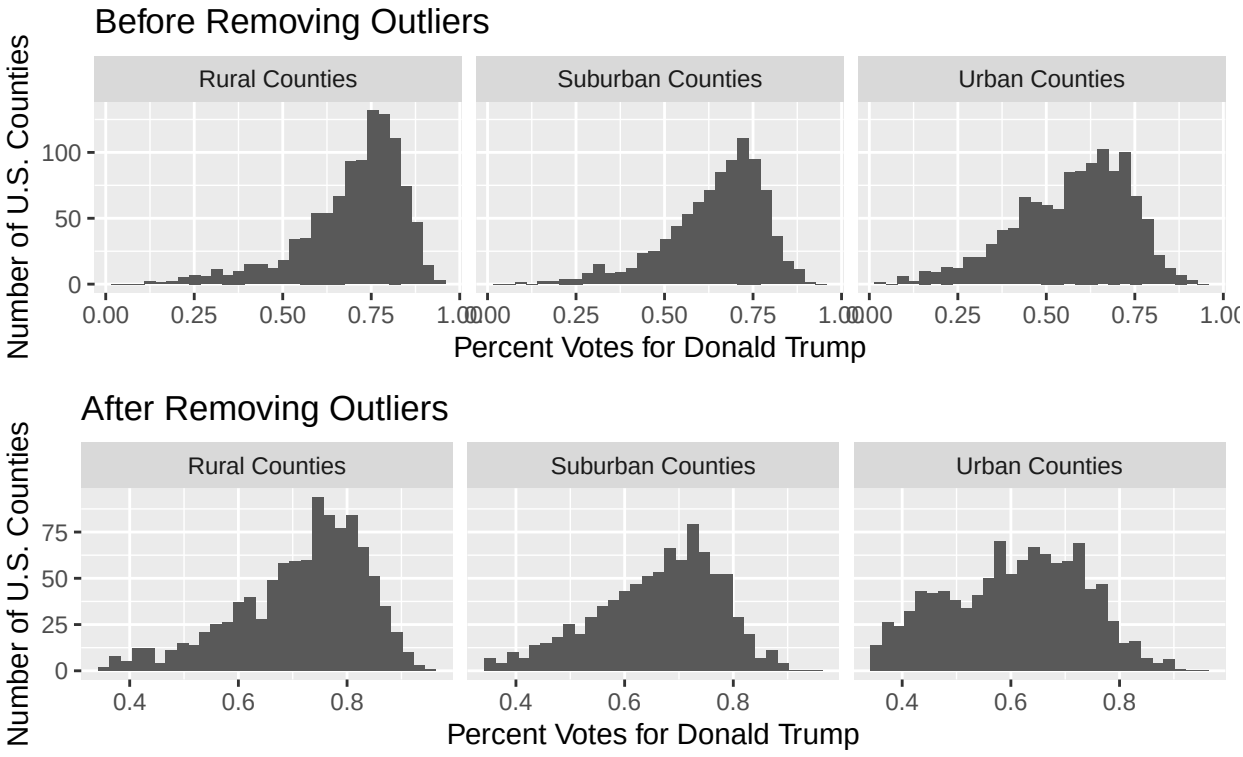
IV. MORE DATA ANALYSIS AND INFERENCE

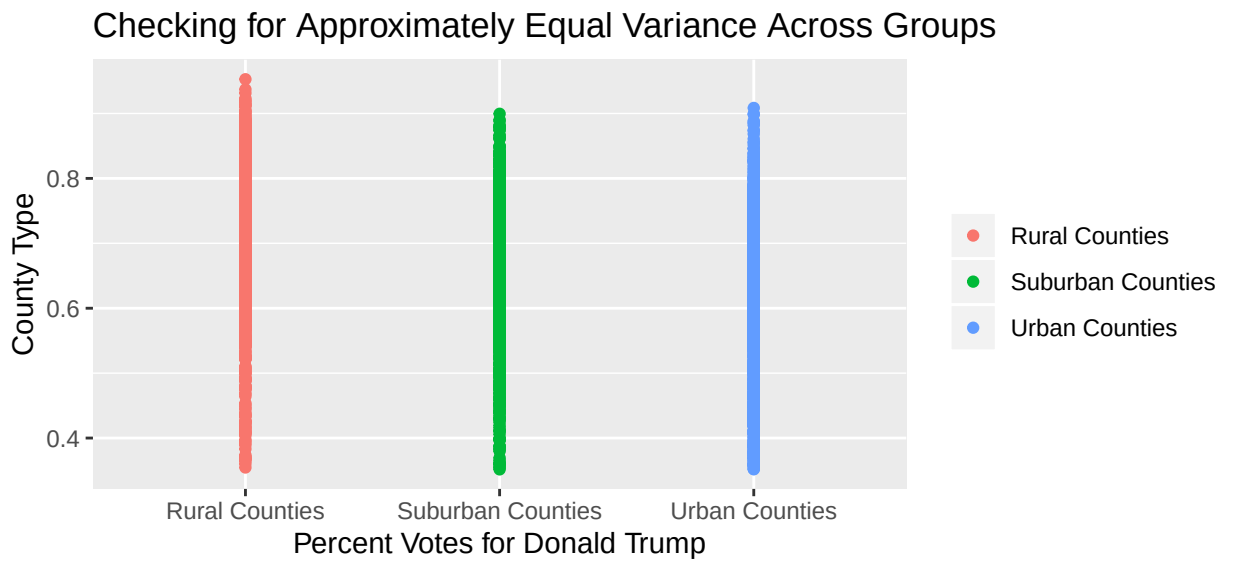
Now that we've taken a look at some of the summary statistics, and are starting to see some trends in the data, I thought it would be beneficial to run some statistical analyses to see if some of these visual differences hold weight.

Analysis of County Type and Percent of Trump Voters

To run some tests on our 3-way ruralness variable (`ruralurban_grp_3_way`), we'll be using Analysis of Variance (ANOVA) to compare means. There are three conditions that need to be checked before running an ANOVA.

- All observations must be independent.
  - The election and ruralness data were not sampled, however, there is not an obvious reason why independence of observations would not hold for most or all of the values in the dataset. **This condition is met.**
- The data in each group must be nearly normal.
  - Given the large sample size for each group, we should be fine, but in order to be safe I removed a few outliers from the full dataset in order to ensure that each group was approaching a normal distribution. In total, I removed 189 counties from the dataset. **This is satisfied.**
- The variance within each group must be approximately equal.
  - We can see from the side-by-side plot below that there is approximately equal variance across groups. **This is satisfied.**





After checking our conditions, we are ready to run the ANOVA to see if there is a significant difference between the mean percentage of Trump voters in the 2016 election broken down by Rural, Suburban, and Urban counties.

Our hypothesis test will look like this:

- H0: There is no difference between the average percent of Trump voters by county type (ruralness).
- HA: There is a difference between the average percent of Trump voters by county type (ruralness).

ANOVA - Ruralness Variable

```
##                          Df Sum Sq Mean Sq F value Pr(>F)
## ruralurban_grp_3_way      2   5.94  2.9690   209.7 <2e-16 ***
## Residuals              2922  41.38  0.0142
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this ANOVA, we can see that the p-value is less than 0.001, which is less than 0.05 given our 95% confidence level. Therefore, we can **reject the null hypothesis**. There is a difference between the average percent of Trump voters by county type (ruralness) in the 2016 election.

Now, although this shows that there is a difference in mean values across our ruralness groups, we do not know where the differences are occurring between groups. In order to determine this, we'll run a post-hoc analysis test:

```
TukeyHSD(aov1)
```

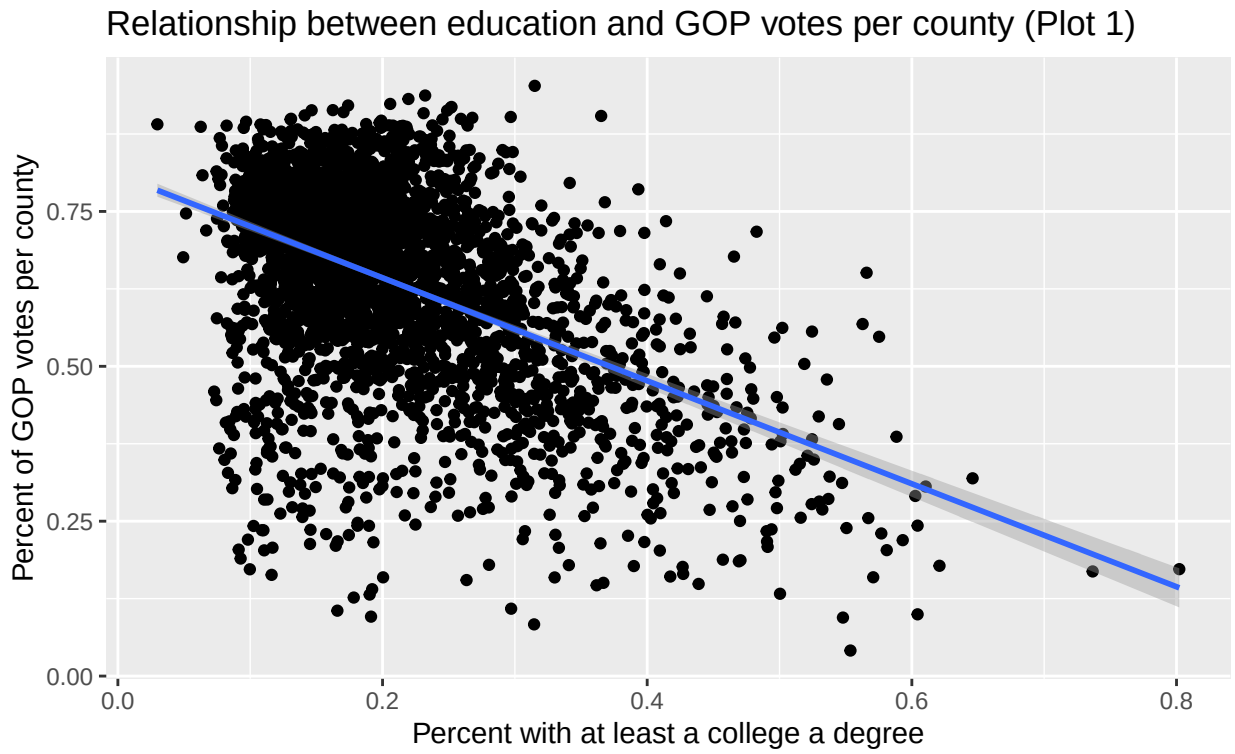
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = per_gop ~ ruralurban_grp_3_way, data = election_education_df_rmoutliers)
##
## $ruralurban_grp_3_way
##                                       diff         lwr         upr p adj
```

```
## Suburban Counties-Rural Counties -0.05455412 -0.06749214 -0.04161610     0
## Urban Counties-Rural Counties    -0.10726745 -0.11955034 -0.09498457     0
## Urban Counties-Suburban Counties -0.05271333 -0.06554077 -0.03988590     0
```

After running the post-hoc analysis, we can see that all three groups have signficantly different means from one another. The p-adjusted value for each group is less than 0.001 (which is less than 0.05) given our 95% confidence level.

Analysis of Educational Attainment Level and Percent of Trump Voters

When we fit a least squares line to our plot from earlier that examines the relationship between educational attainment and the percent of GOP votes per county, there is a negative correlation between these two variables. In other words, as the percent of individuals in a given county with a college degree or more increases, the percent of GOP votes in a given county decreases.




```
##
## Call:
## lm(formula = per_gop ~ college_or_more_pct, data = election_education_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56562 -0.07144  0.01850  0.09072  0.40550
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.809042   0.006083  133.01   <2e-16 ***
## college_or_more_pct -0.831090   0.026792  -31.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.1365 on 3109 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2363, Adjusted R-squared:  0.2361
## F-statistic: 962.2 on 1 and 3109 DF,  p-value: < 2.2e-16
```

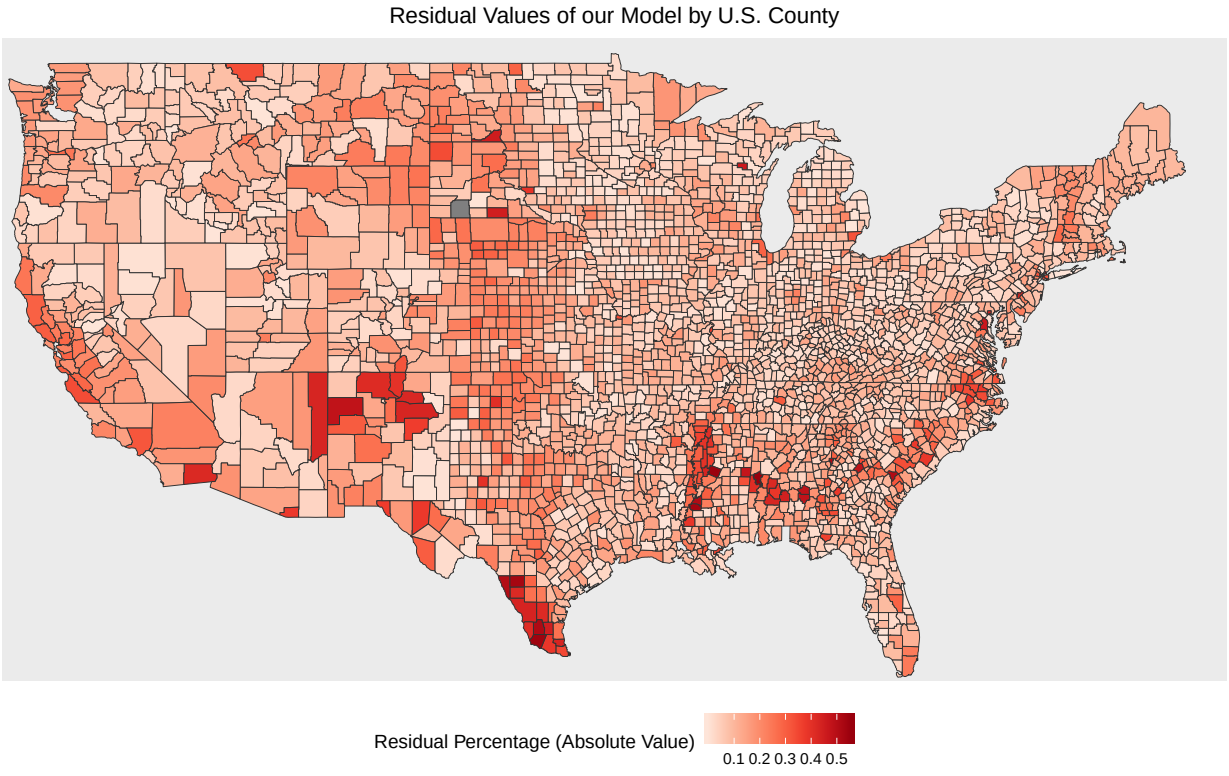
From the summary of the regression, we can see that the p-value is less than 0.001 (which is less than 0.05), meaning that the slope of the least squares line is not equal to zero. Additionally, we can see that there is an adjusted R^2 value equal to 0.2361, indicating that the amount of variation in the percent of votes for Trump in a given county in the 2016 election that can be explained by the percent of individuals in a given county with a college degree or more is around 24%.

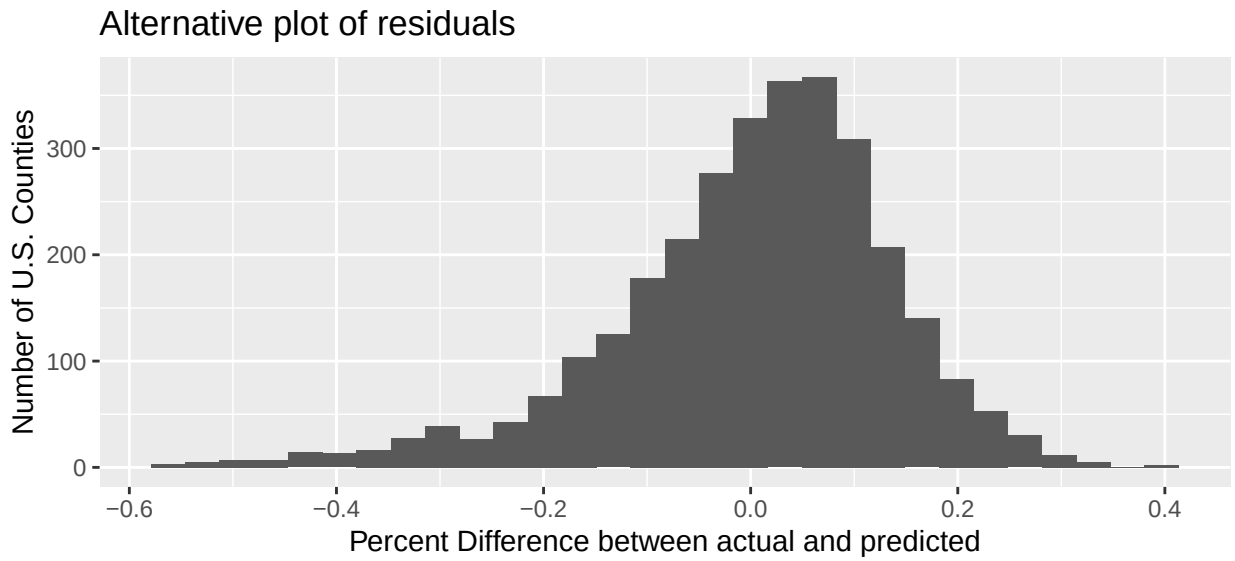
Predictability of Educational Attainment and Ruralness in the Proportion of GOP Votes per county in the 2016 election

The summary from our regression model in the previous section also contains values to create a linear regression equation that we can use to predict the percent of votes for Trump in the 2016 election given the percentage of individuals with a college degree or more for a given county:

$\hat{PercVotesGOP} = 0.809042 + (-0.831090 \times$ `college_or_more_pct`$)$

Although this may seem valuable, if we were to run a prediction for each county using this linear regression equation, we'd find that our residuals would be quite high. To test this, I ran this calculation on the dataset and have plotted the residual values by county on the map below:





Yikes! Although the distribution of residuals seems to be centered around zero, we can see that for some counties our model predicted the proportion of GOP votes 50% higher or more than the actual! As we can see, the educational attainment variable has some predictability, but if we were to use it as the only variable to predict the percent of GOP votes per county in the 2016 election, we'd have large residuals for some of the counties.

In an attempt to improve our model, and to examine the predictability of ruralness on the proportion of GOP votes by county in the 2016 presidential election, we'll need to create a few dummy variables from our 3-way ruralness variable.

county_name	state_abbr	ruralurban_grp_3_way	county_rural	county_suburban	county_urban
Autauga County	AL	Urban Counties	0	0	1
Baldwin County	AL	Urban Counties	0	0	1
Barbour County	AL	Suburban Counties	0	1	0
Bibb County	AL	Urban Counties	0	0	1
Blount County	AL	Urban Counties	0	0	1
Bullock County	AL	Suburban Counties	0	1	0
Butler County	AL	Suburban Counties	0	1	0
Calhoun County	AL	Urban Counties	0	0	1
Chambers County	AL	Suburban Counties	0	1	0
Cherokee County	AL	Suburban Counties	0	1	0
Chilton County	AL	Urban Counties	0	0	1
Choctaw County	AL	Rural Counties	1	0	0
Clarke County	AL	Rural Counties	1	0	0
Clay County	AL	Rural Counties	1	0	0
Cleburne County	AL	Rural Counties	1	0	0

Although I found out later that R creates dummy variables automatically if adding in a categorical value to a regression model, I wanted to make sure I tested this assumption. You can see below two identical outputs of my regression model, one using my created dummy variables, the other generated from R directly using my `ruralurban_grp_3_way` categorical variable.

```
regression2 <- lm(formula = per_gop ~ college_or_more_pct + county_suburban + county_urban,
                  data = county_fnl_shp)
summary(regression2)
```

```
##
## Call:
## lm(formula = per_gop ~ college_or_more_pct + county_suburban +
##     county_urban, data = county_fnl_shp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58453 -0.06892  0.02184  0.08926  0.35193
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.829211   0.006625 125.164   <2e-16 ***
## college_or_more_pct -0.725063   0.028987 -25.013   <2e-16 ***
## county_suburban     -0.052569   0.006034  -8.713   <2e-16 ***
## county_urban        -0.067638   0.006054 -11.172   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1323 on 3064 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.261,  Adjusted R-squared:  0.2603
## F-statistic: 360.7 on 3 and 3064 DF,  p-value: < 2.2e-16
```

```
regression3 <- lm(formula = per_gop ~ college_or_more_pct + ruralurban_grp_3_way,
                  data = county_fnl_shp)

summary(regression3)
```

```
##
## Call:
## lm(formula = per_gop ~ college_or_more_pct + ruralurban_grp_3_way,
##     data = county_fnl_shp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58453 -0.06892  0.02184  0.08926  0.35193
##
## Coefficients:
##                                      Estimate Std. Error t value
## (Intercept)                           0.829211   0.006625 125.164
## college_or_more_pct                  -0.725063   0.028987 -25.013
## ruralurban_grp_3_waySuburban Counties -0.052569   0.006034  -8.713
## ruralurban_grp_3_wayUrban Counties    -0.067638   0.006054 -11.172
##                                      Pr(>|t|)
## (Intercept)                            <2e-16 ***
## college_or_more_pct                    <2e-16 ***
## ruralurban_grp_3_waySuburban Counties  <2e-16 ***
## ruralurban_grp_3_wayUrban Counties     <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1323 on 3064 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.261,  Adjusted R-squared:  0.2603
```

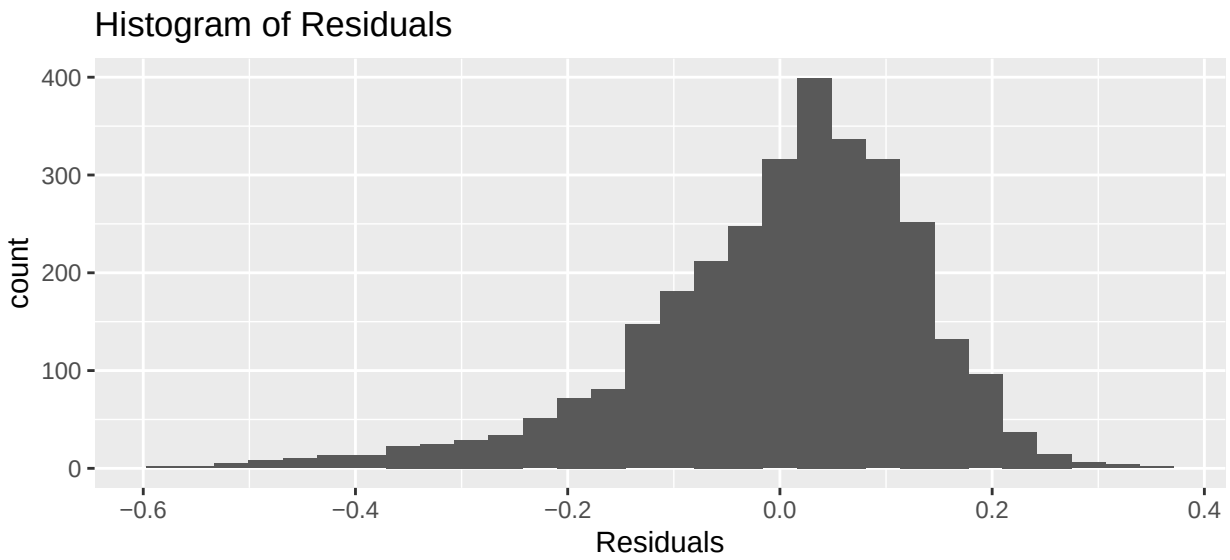
```
## F-statistic: 360.7 on 3 and 3064 DF,  p-value: < 2.2e-16
```

As we can see here, adding in the ruralness variable improved our R^2 value slightly, from 0.2361 to 0.2603. Now, before we can interpret the model, we need to check some of the model diagnostics and conditions.

Model Diagnostics - Checking Conditions

1. Nearly normal residuals

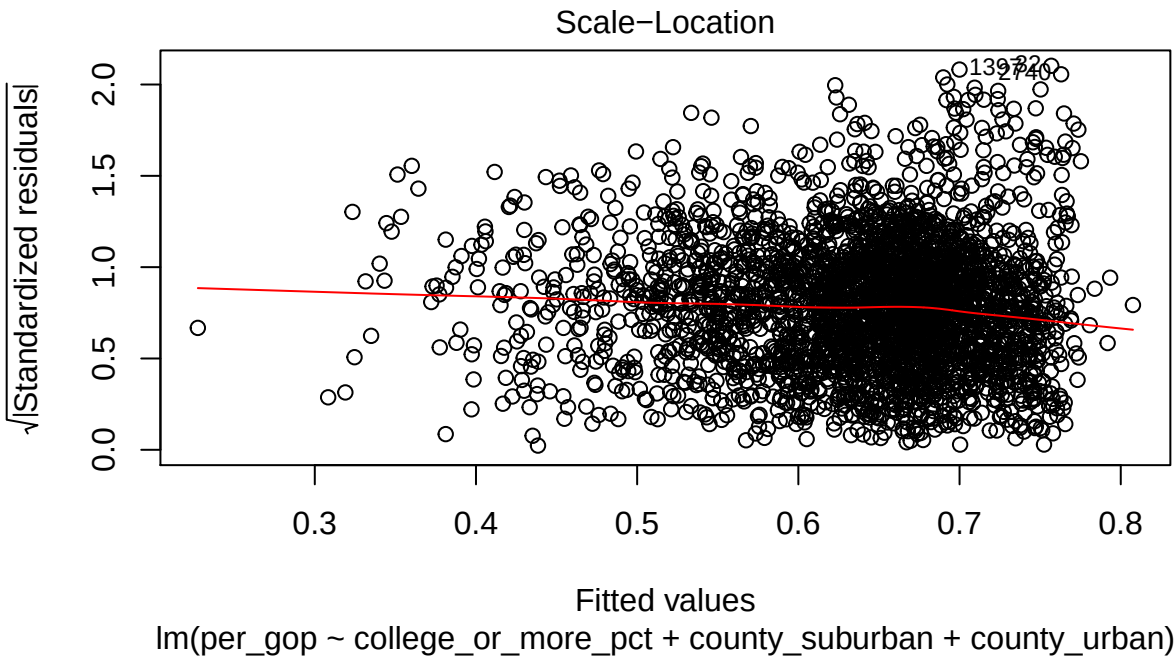
- The first condition we'll need to check is whether or not the residuals of the model are nearly normal:



As we can see from the above plot, the residuals do appear to be nearly normal. Although it looks slightly left skewed, the data set is quite large which makes this condition a little less substantial. **This condition is met.**

2. Nearly constant variability of residuals

- The next condition we'll need to check is whether or not the variability of the residuals is nearly constant. To check this, we can use the `plot` function in R and utilize the Scale-Location plot:



From the plot above, it appears that the red line through the center of the plot is not very horizontal and flat – which would indicate that the variability is not nearly constant. We'll run another test from the `car` package, the `ncvTest`, to test whether or not there is "non-constant variance" in the residuals:

```
library(car)
ncvTest(regression2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 7.335863, Df = 1, p = 0.0067592
```

Unfortunately, the null hypothesis for this test is that the variance of the residuals *is* constant. We can see from our test that the p-value is less than 0.05, indicating that there *is not constant variability of residuals.* **This condition is NOT met**.

3. Residuals are independent

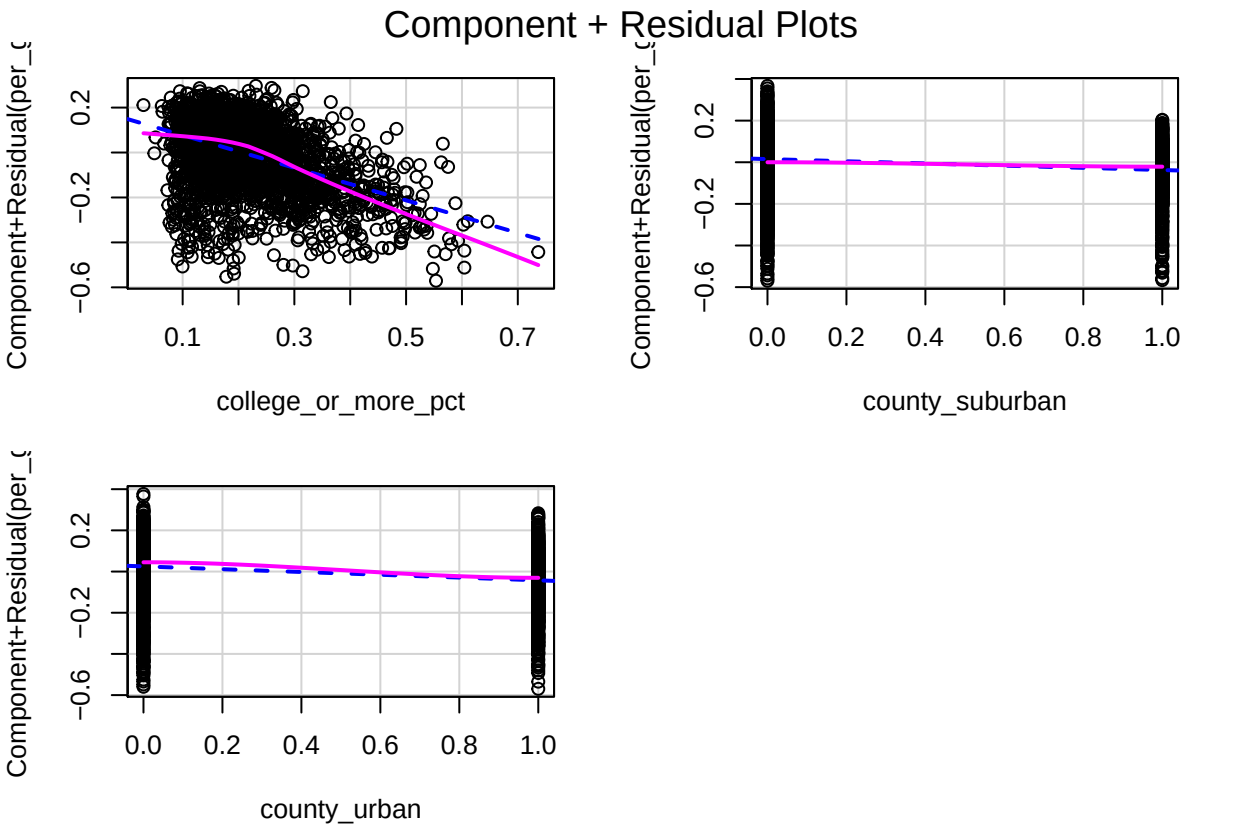
- The next condition to test is whether or not the residuals are independent. Given that these election results were happening all at one time (individuals were going to their polling stations and casting their votes on the same day and roughly around the same time), it is difficult to think of a logical reason as to why these residuals would not be independent from one another. **This condition is met**.

4. Each variable is linearly related to the outcome

- Finally, we'll need to test whether or not each variable is linearly related to the outcome. To test this, we can create individual plots of each variable against the `per_gop` variable to ensure that there is a linear relationship.

A great tool to investigate this, is to use the `crPlots()` function in the `car` package to show component residual plots of our predictors. These types of plots attempt to model the residuals of one predictor against the dependent variable and adds a line indicating where the line of best fit lies. If we can visibly see a significant divergence between the residual line (pink) and the component line (blue), then we can see that the predictor does not have a linear relationship with the dependent variable – in our case `per_gop`. We can see these plots below with our latest regression model:

```
library(car)
crPlots(regression2)
```

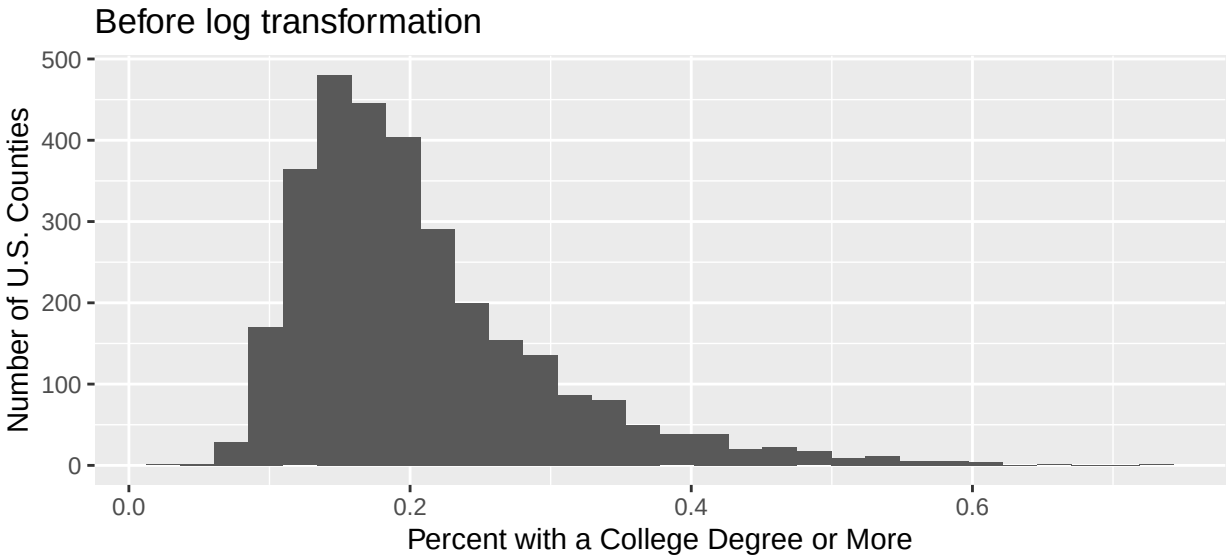


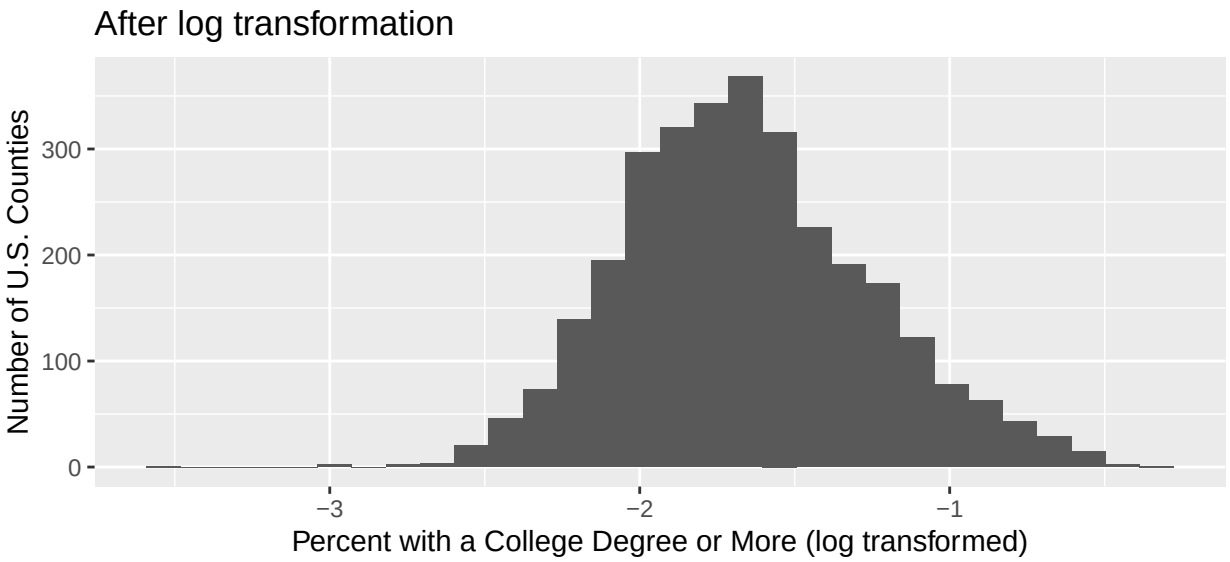
As we can see, it looks like our ruralness variable residuals seem to map onto the best fit line pretty well. However, the educational attainment variable (`college_or_more_pct`) residuals seem to drastically diverge from the best fit line. **This condition is NOT met.**

Attempts at adjusting the model

In summary, it looks like two of our conditions out of the four for multiple regression are not met. Although the R^2 value is quite low, and we wouldn't be able to utilize this latest model as a valid predictive tool, I did want to try to adjust the model a bit to see if I could resolve some of the conditions that were not met above.

One way to do this would be to perform a transformation on the predictors. Based on my work above, I already know that my `college_or_more_pct` distribution is significantly right skewed. As an attempt to make this more of a normal distribution, I'll transform the values from x to $log(x)$.





Now that the education variable is transformed, and shows a much more normal distribution, let's insert it back into our model to see if it has an impact:

```
##
## Call:
## lm(formula = per_gop ~ transform + county_suburban + county_urban,
##     data = county_fnl_shp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60000 -0.07217  0.02227  0.09151  0.34001
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.452362   0.012640  35.789   <2e-16 ***
## transform        -0.138820   0.006717 -20.667   <2e-16 ***
## county_suburban -0.052380   0.006203  -8.445   <2e-16 ***
```

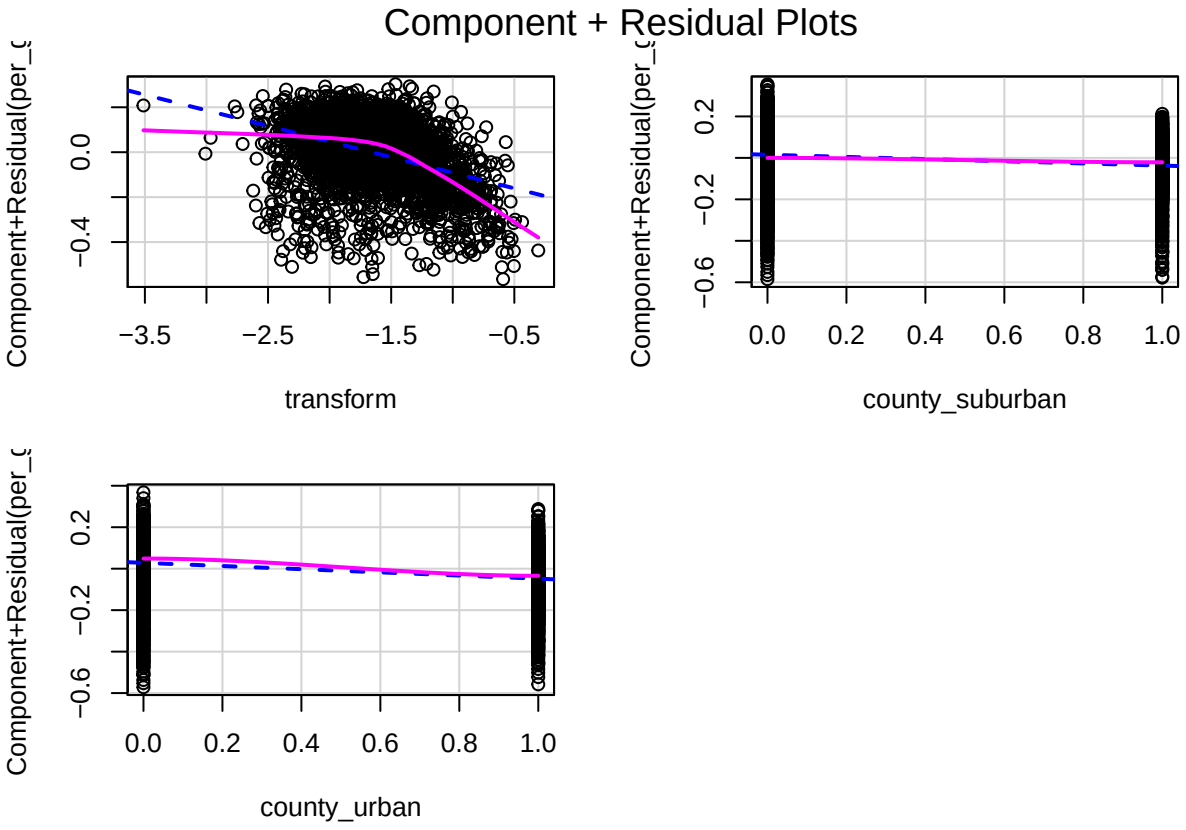
```
## county_urban    -0.076474   0.006209 -12.316   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.136 on 3064 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.219,  Adjusted R-squared:  0.2182
## F-statistic: 286.3 on 3 and 3064 DF,  p-value: < 2.2e-16
```

It appears to have affected our R^2 value slightly. Let's see if it helped resolve our heteroscedasticity issue (condition 2):

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 7.068916, Df = 1, p = 0.0078433
```

It did improve our p-value slightly, from 0.006 to 0.008. However, we still cannot reject the null hypothesis that variability of the residuals is nearly constant since it is less than 0.05.

Additionally, when we look at the component residuals plot, we actually see it is worse (`transform`) – the new `college_or_more_pct` variable. I'm assuming that the log transformation made our residual line deviate even more from the best fit line.



V. CONCLUSION & FUTURE RESEARCH

Conclusion

After working through a pretty extensive analysis of the relationship between the `per_gop` variable and the `college_or_more_pct` and `ruralurban_grp_3_way` variable, looking at educational attainment and ruralness respectively and their impact on the proportion of Trump votes in the 2016 election, we can circle back to our research questions:

> Was there a significant difference in the mean proportion of GOP votes by county in the 2016 presidential election based on county type (Rural, Suburban, or Urban)?

- When we ran an Analysis of Variance (ANOVA), utilizing a 95% confidence interval, we found that there were significant differences in the mean proportion of voters that voted for GOP candidate Donald Trump when testing across the ruralness continuum (Rural, Suburban, Urban) in the 2016 election.
- After running post-hoc analysis, it is confirmed that there was a significantly higher mean proportion of voters that voted for Donald Trump in rural counties compared to the mean proportion that voted for Donald Trump in suburban counties and urban counties in the 2016 election.
- Additionally, a significantly higher mean proportion of voters voted for Donald Trump in suburban counties than the mean proportion of voters that voted for him in urban counties.

As we are already seeing in the lead up to the 2020 election, votes in more rural counties will be highly contested and the Democratic party will seek to regain votes that they lost in the 2016 election. As campaign managers spend money on advertisements and events in places like Iowa and the Midwest – typically more rural areas – they are likely understanding these factors that rural/suburban votes are crucial and need to be won back in this upcoming election.

> Is educational attainment and/or ruralness predictive of the proportion of GOP votes by county in the 2016 presidential election?

- After initially running a linear regression to see if educational attainment (`college_or_more_pct`) had a linear relationship with the proportion of GOP votes by county in the 2016 election, we could see that this was the case when we added a least squares line. The R^2 value was quite low, at roughly 0.24. There is an argument to be made that educational attainment alone was predictive of the proportion of GOP votes by county in the 2016 election, given that roughly 24% of variation in the percent of votes for Donald Trump in a given county in the 2016 election could be explained by the percent of individuals with a college degree or higher for that county.
- This is also the case when we added ruralness to the regression model. The R^2 value increased slightly, indicating that roughly an added 2% of variation in the percent of votes for Donald Trump in a given county in the 2016 election could be explained by the combination of the education variable and the ruralness variable.
- However, when checking the model diagnostics, a few conditions were not able to be met. It appeared that there wasn't nearly constant variability of residuals and it is debateable whether or not each variable, specifically the education variable, was linearly related to the outcome.

Overall, I'd say that both of these factors did show some level of predictive capacity during the analysis. However, I'd caution against using the education variable until it is better known what relationship it has with the proportion of GOP votes by county in the 2016 election (i.e. is it a linear relationship?, non-linear relationship?, etc.).

Predictive models are often built to forecast elections, and utilizing types of variables such as the ones in this project are quite common. However, in the case of the regression model that was created, we'd need to incorporate more variables and data to increase the predictive capacity before comfortably using it as a predictive tool. With an R^2 value around 0.26, we'd have to boost this value by at least 0.50 in order to make respectable predictions of the proportion of votes for Donald Trump in the 2016 election by county.

Future Research

There are many ways that this project can be extended. Namely, continuing to add variables to the model in order to increase its predictive capacity. Some variables that could be added are those from the Census around the time of the election. Adding in variables such as median household income, age, race/ethnicity breakdowns, employment information, etc. by county could really help the predictive capacity since it'll likely subset the data more and allow the model to utilize more nuanced information for its predictions.

Additionally, it would be helpful to narrow in on why there was not constant variability of the residuals of the model.

REFERENCES

- OpenIntro Statistics
  - Particularly, chapter 7 (ANOVA), chapter 8, and chapter 9 (Regression)
- Regression Diagnostics
- Regression Diagnostics
- Creating Dummy Variables