DATA 606 Data Project Proposal

Zach Alexander

Data Preparation

```
library(dplyr)
library(ggplot2)
library(psych)
library(knitr)
library(kableExtra)
```

```
election_link <- paste0('https://raw.githubusercontent.com/zachalexander/',
                        'data_606_cunysps/master/ProjectProposal/',
                        'county_level_votes.csv')

education_link <- paste0('https://raw.githubusercontent.com/zachalexander/',
                         'data_606_cunysps/master/ProjectProposal/',
                         'education_data_fip.csv')
# load data
election_results_df <- read.csv(election_link)

education_data_df <- read.csv(education_link)
```

```
# the dataset has many columnes, I'm subsetting doum to only columns of interest
election_results_df <- election_results_df %>%
  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 %>%
  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))))))))))
```

```r
# 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),
         party_winner = ifelse(per_gop > per_dem, 'Republican', 'Democrat')) %>%
  select(combined_fips:per_gop, college_or_more_pct, party_winner,
         ruralurban_cc, ruralurban_grp)
```

```r
# a quick look at the data frame
head(election_education_df)
```

```
##   combined_fips    county_name state_abbr    per_dem   per_gop
## 1          1001 Autauga County         AL 0.23956855 0.7343579
## 2          1003 Baldwin County         AL 0.19565310 0.7735147
## 3          1005 Barbour County         AL 0.46660250 0.5227141
## 4          1007    Bibb County         AL 0.21422039 0.7696616
## 5          1009  Blount County         AL 0.08469902 0.8985188
## 6          1011 Bullock County         AL 0.75090406 0.2422889
##   college_or_more_pct party_winner ruralurban_cc
## 1            24.59277   Republican             2
## 2            29.54711   Republican             3
## 3            12.86779   Republican             6
## 4            12.00000   Republican             1
## 5            13.04976   Republican             1
## 6            10.25501     Democrat             6
##                                             ruralurban_grp
## 1    Counties in metro areas of 250,000 to 1 million population
## 2      Counties in metro areas of fewer than 250,000 population
## 3 Urban population of 2,500 to 19,999, adjacent to a metro area
## 4       Counties in metro areas of 1 million population or more
## 5       Counties in metro areas of 1 million population or more
## 6 Urban population of 2,500 to 19,999, adjacent to a metro area
```

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

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

I thought these two items would be interesting to look into in lieu of the upcoming 2020 election. I had read a lot of articles after 2016 about the "urban/rural" divide, and educational differences between Republicans and Democrats, and thought it would be interesting to explore the election data a bit more to see if these factors really do seem to have an affect on voter choice. If I do see large differences based on these factors, it would be neat to build some type of predictive model in the lead up to the 2020 election.

Cases

What are the cases, and how many are there?

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.

Data collection

Describe the method of data collection.

The education percentages were calculated by Stephen Pettigrew at Harvard University 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 proposal.

Type of study

What type of study is this (observational/experiment)?

This is an observational study.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

I found the dataset with 2016 election results on GitHub: https://github.com/tonmcg/US_County_Level_Election_Results_08-16/blob/master/2016_US_County_Level_Presidential_Results.csv

I found the dataset with the education and ruralness data on GitHub as well: https://github.com/MEDSL/2018-elections-unoffical/blob/master/election-context-2018.md

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

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

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

Independent Variable

You should have two independent variables, one quantitative and one qualitative.

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

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

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

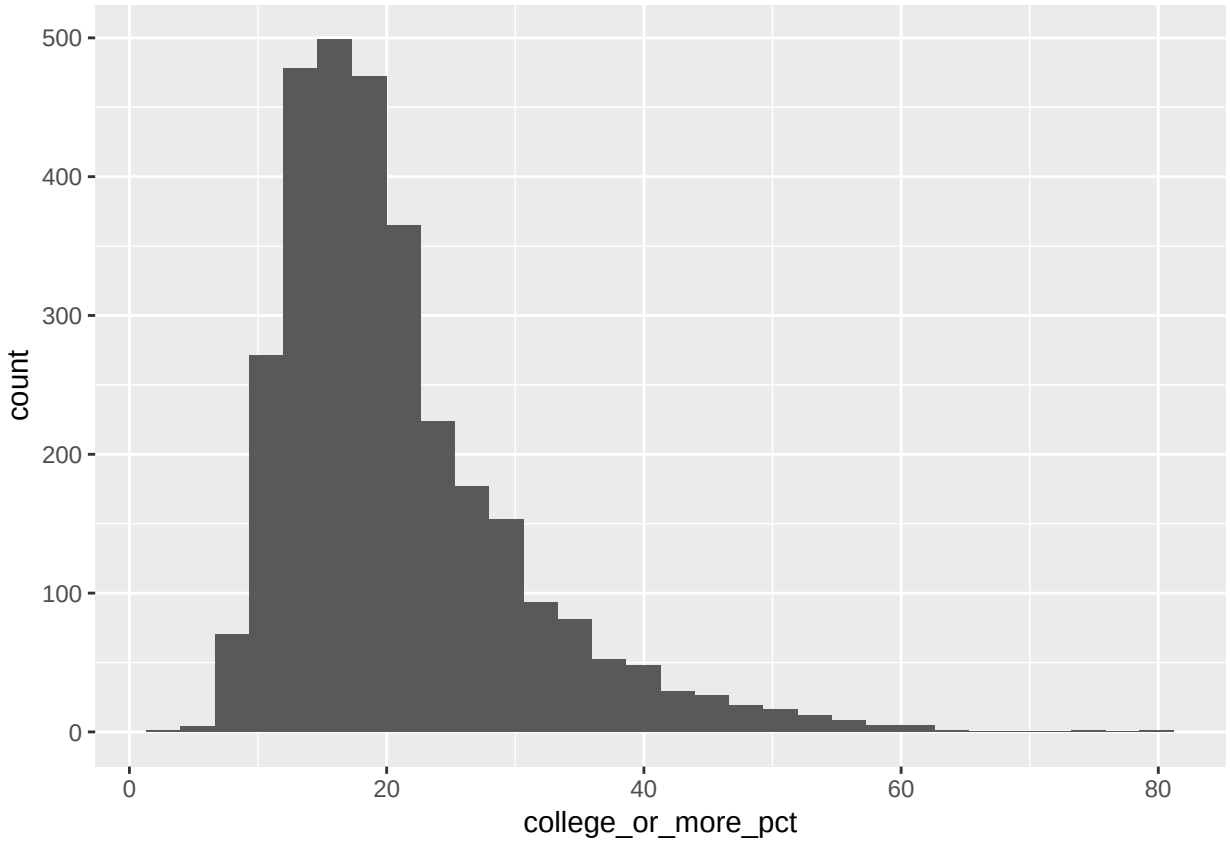
```
kable(table(election_education_df$party_winner), align = rep('c', 2)) %>%
  kable_styling(bootstrap_options = c("striped"), full_width = F)
```

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 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. We can see that the histogram is right skewed, with a center around 20%. The histogram is unimodal.

```
ggplot(election_education_df, aes(x=college_or_more_pct)) + geom_histogram()
```

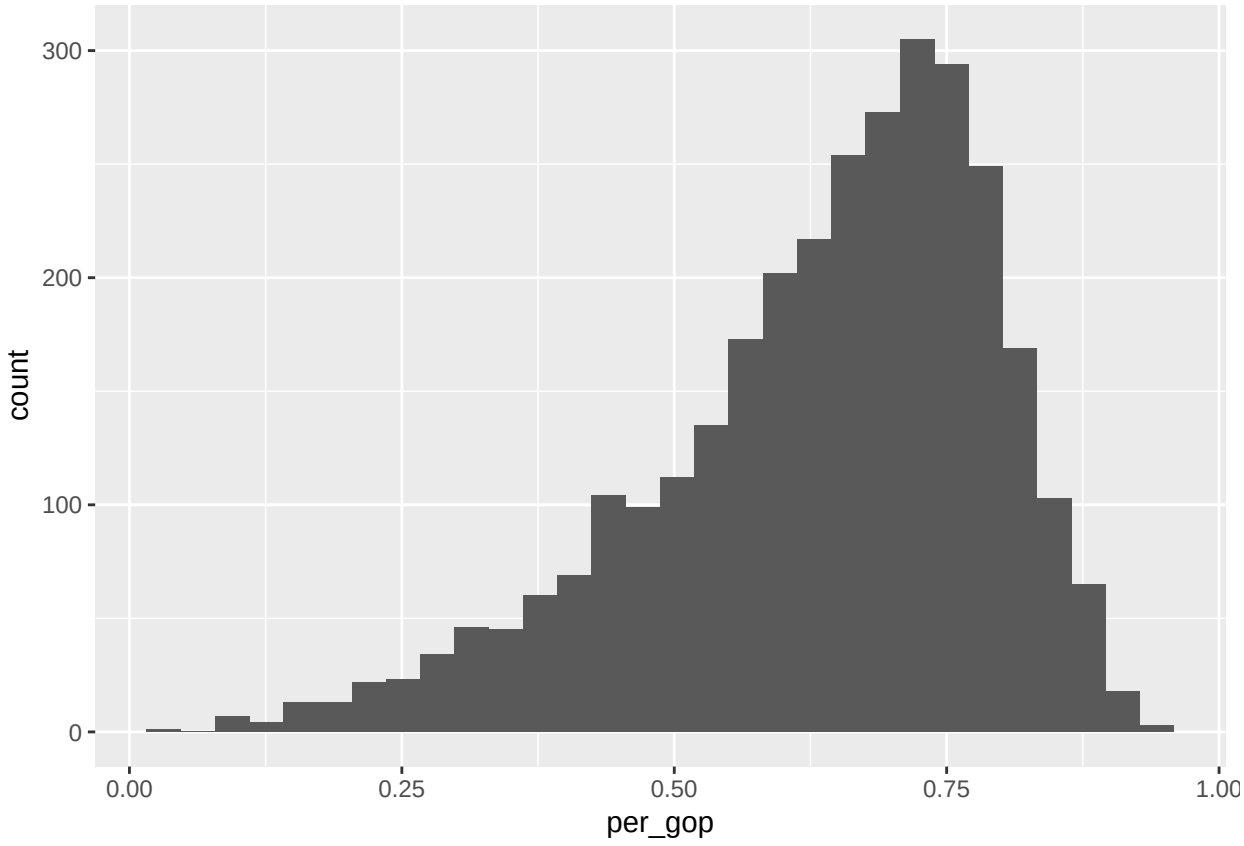



```
describe(election_education_df$college_or_more_pct)
```

```
##    vars    n  mean   sd median trimmed  mad  min   max range skew kurtosis
## X1    1 3111 20.78 9.14  18.53   19.53 6.99 2.99 80.21 77.23 1.52      3.1
##      se
## X1 0.16
```

Next, I thought it would be helpful to plot a histogram of the percent of each county that voted for Donald Trump in 2016. Again, this distribution is unimodal, however it is left skewed with a center around 65%.

```
ggplot(election_education_df, aes(x=per_gop), bins = 30) + geom_histogram()
```




```
describe(election_education_df$per_gop)
```

```
##    vars    n mean   sd median trimmed  mad  min  max range  skew kurtosis
## X1    1 3112 0.64 0.16   0.67    0.65 0.14 0.04 0.95  0.91 -0.84     0.39
##    se
## X1  0
```

To take a look at my rural variable, I first thoguht it would be helpful to see some summary statistics related to my qualitative variable 'ruralurban_grp':

```
describeBy(election_education_df$per_gop,
           group = election_education_df$ruralurban_grp,
           mat = TRUE)
```

```
##     item
## X11    1
## X12    2
## X13    3
## X14    4
## X15    5
## X16    6
## X17    7
```

```
## X18    8
##                                                                        group1
## X11 Completely rural or less than 2,500 urban population, adjacent to a metro area
## X12                         Counties in metro areas of 1 million population or more
## X13                      Counties in metro areas of 250,000 to 1 million population
## X14                        Counties in metro areas of fewer than 250,000 population
## X15                   Urban population of 2,500 to 19,999, adjacent to a metro area
## X16               Urban population of 2,500 to 19,999, not adjacent to a metro area
## X17                    Urban population of 20,000 or more, adjacent to a metro area
## X18                Urban population of 20,000 or more, not adjacent to a metro area
##     vars   n      mean        sd    median   trimmed       mad        min
## X11    1 627 0.7155089 0.1409475 0.7492596 0.7340160 0.1094827 0.12671677
## X12    1 432 0.5355758 0.1775423 0.5677114 0.5468153 0.1944945 0.04122067
## X13    1 376 0.5806173 0.1447696 0.5946537 0.5872617 0.1605669 0.17758944
## X14    1 354 0.6137479 0.1395124 0.6314410 0.6234976 0.1378521 0.13286210
## X15    1 593 0.6630747 0.1318980 0.6929303 0.6782067 0.1077754 0.08321823
## X16    1 425 0.6729490 0.1431233 0.7066179 0.6911833 0.1209096 0.17953822
## X17    1 214 0.6060651 0.1260898 0.6264930 0.6149296 0.1160750 0.18971859
## X18    1  91 0.6062969 0.1546884 0.6518507 0.6192606 0.1285414 0.20724168
##           max     range       skew   kurtosis          se
## X11 0.9527273 0.8260105 -1.3462052  2.0365845 0.005628899
## X12 0.8985188 0.8572981 -0.4841845 -0.4595109 0.008542009
## X13 0.9085546 0.7309651 -0.3810779 -0.3713940 0.007465923
## X14 0.8866245 0.7537624 -0.6371559  0.1957401 0.007415002
## X15 0.8995612 0.8163430 -1.2020007  1.7648060 0.005416403
## X16 0.8885465 0.7090083 -1.1623278  1.0857824 0.006942501
## X17 0.8781050 0.6883865 -0.7028702  0.3655221 0.008619320
## X18 0.8797324 0.6724907 -0.6847745 -0.2889630 0.016215748
```

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 setup, I could then look at summary statistics of the 'per_gop' variable and 'college_or_more_pct' variable split by this new variable.

```
# summary statistics of the percent voting for Donald Trump split by ruralness
describeBy(election_education_df$per_gop,
           group = election_education_df$ruralurban_grp_3_way,
           mat = TRUE)
```

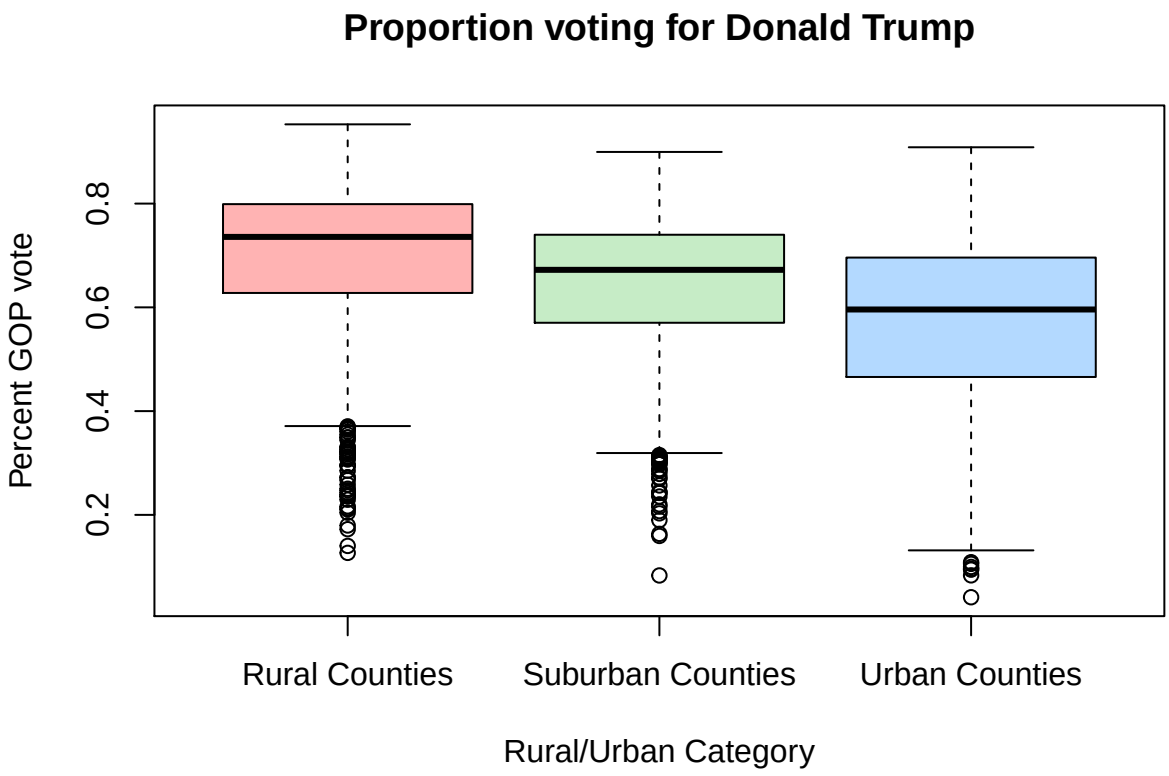
```
##     item            group1 vars    n      mean        sd    median
## X11    1    Rural Counties    1 1052 0.6983151 0.1432941 0.7357744
## X12    2 Suburban Counties    1  898 0.6437352 0.1355943 0.6723085
## X13    3    Urban Counties    1 1162 0.5739652 0.1594815 0.5958650
##       trimmed       mad        min       max     range       skew
## X11 0.7167167 0.1134308 0.12671677 0.9527273 0.8260105 -1.2389639
## X12 0.6575258 0.1185858 0.08321823 0.8995612 0.8163430 -0.9857139
## X13 0.5848342 0.1671507 0.04122067 0.9085546 0.8673339 -0.5956667
##        kurtosis          se
## X11  1.54343191 0.004417946
## X12  0.96192696 0.004524839
## X13 -0.05579262 0.004678506
```

```
# summary statistics of the percent with a bachelor's degree or higher split by ruralness
describeBy(election_education_df$college_or_more_pct,
           group = election_education_df$ruralurban_grp_3_way,
           mat = TRUE)
```

```
##     item            group1 vars    n     mean        sd   median  trimmed
## X11    1    Rural Counties    1 1052 17.99575  6.598972 17.01078 17.34035
## X12    2 Suburban Counties    1  897 17.92921  6.819967 16.30310 16.94540
## X13    3    Urban Counties    1 1162 25.51077 10.616054 23.55707 24.47652
##          mad      min      max    range      skew kurtosis        se
## X11 5.545929 2.985075 60.43459 57.44951 1.6415236 5.665799 0.2034550
## X12 4.694267 6.397138 64.59119 58.19406 1.9812014 6.103309 0.2277121
## X13 9.827727 7.455096 80.21012 72.75502 0.9741119 1.152331 0.3114297
```

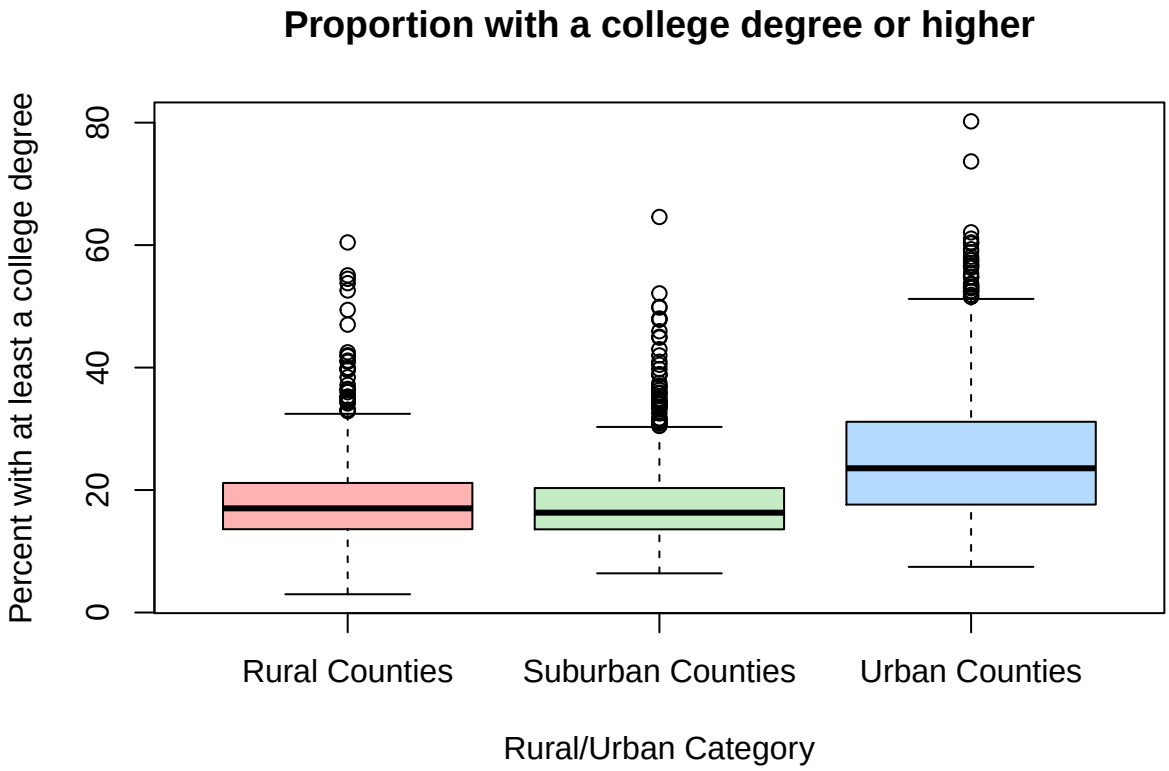
Finally, I thought it would be interesting to plot these two tables as box plots. See below for the box plot of the of per_gop split by ruralness (3-way):

```
boxplot(per_gop~ruralurban_grp_3_way,data=election_education_df,
   col = c("#ffb3b3", "#c6ecc6", "#b3d9ff"),
   main="Proportion voting for Donald Trump",
   xlab="Rural/Urban Category",
   ylab="Percent GOP vote")
```



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

```
boxplot(college_or_more_pct~ruralurban_grp_3_way,data=election_education_df,
   col = c("#ffb3b3", "#c6ecc6", "#b3d9ff"),
   main="Proportion with a college degree or higher",
   xlab="Rural/Urban Category",
   ylab="Percent with at least a college degree")
```



Although this second boxplot is a bit outside the scope of the reseearch 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.