FinalWriteUp

Apache Juntion Armchairs: Ellie, Ryan, Sude and Darren

4/20/2020

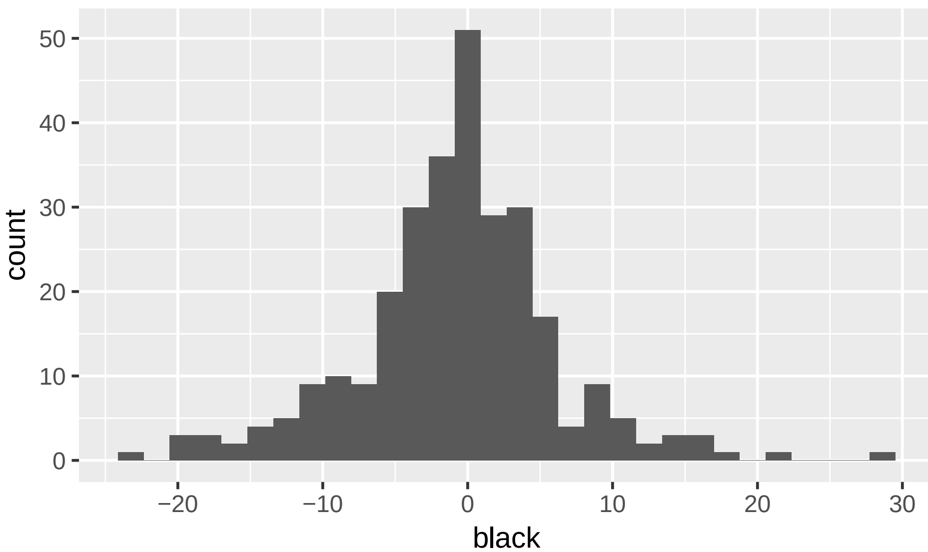
Section 1: Introduction

Section 2: Regression Analysis:

manually imputing the mean value for homeprice and income, and setting Na values of PP change variables to zero.

The distribution of change in Black population:

In order to assess whether gentrification is taking place in an area we looked at the change in the Black population. First we looked at the distribution of change in Black population to determine an appropriate threshold.




```
## [1] 6.88333
```

std deviation is = 6.88333. We will use this value (-6.88333) as the threshold to determine if gentrification has occurred in a census tract. If a tract has experienced more than a -6.88333 PP change in the Black population, we will consider that census tract "gentrified." Even though the mean is not exactly at 0, it is close enough that we feel one standard deviation away from 0 is a sufficient threshold for gentrification.

Next, we create a new variable "gent" to represent whether a census tract is gentrified or not. As described above, ff a census tract have less than or equal to -6.88333 percent change in Black population then we classify the region as gentrified and the variable gent will be "1" otherwise the region will not be classified as gentrified and gent will be "0".

```
## # A tibble: 2 x 2
##    gent     n
##   <dbl> <int>
## 1     0   246
## 2     1    42
```

In our data set we have 246 observations that are not considered gentrified and 42 that are.

After examining the Univariate and Bi-variate EDA (located in section 6) we proceeded with our analysis without any additional transformations because each predictor variable is normally distributed around 0 and the relationship between the response variable "gent" and the predictor variables are all each roughly normal.

###Part I: Location of Gentrification

In part I, the following research question will be examined:

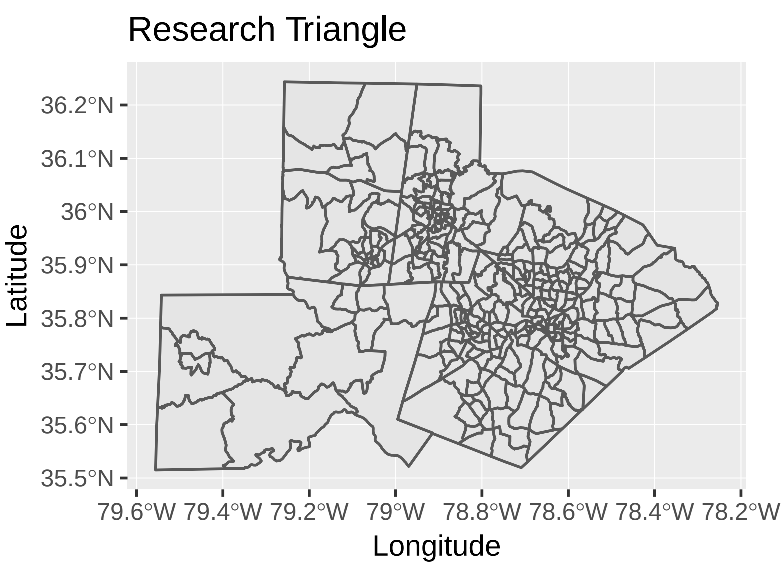
Where in the Research Triangle (counties including Durham, Wake, Orange and Chatham) is gentrification occurring the most?

Recoding our response variable to "1" if change in black population is <= "-6.765" or one standard deviation below 0 (roughly the mean) and equal to "0" if > "6.765" in order visualize and eventually create a logistic model:

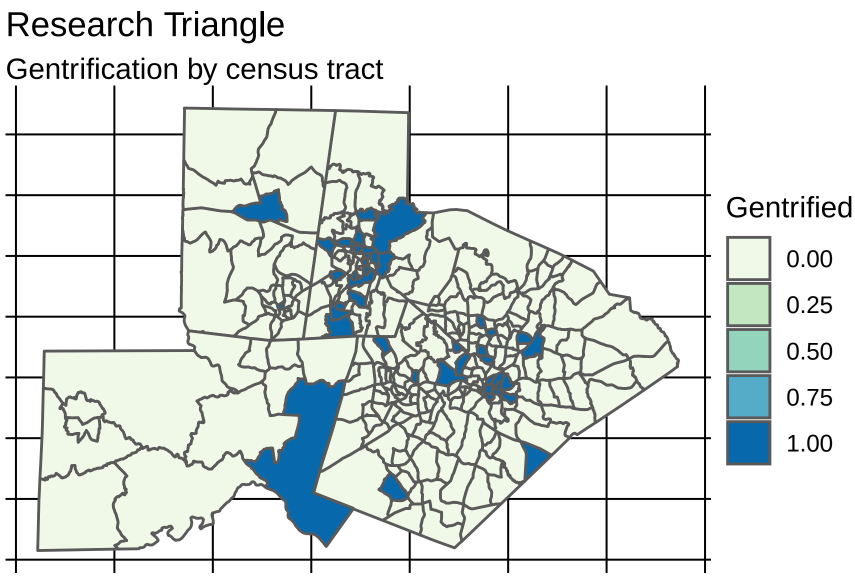
```
## # A tibble: 2 x 2
##    gent     n
##   <dbl> <int>
## 1     0   244
## 2     1    44
```

In order to determine the locations of gentrification we use spatial data to conduct our analysis:

Plotting research triangle area (counties: Chatham, Durham, Orange and Wake): We will be looking only at this region and conducting our analysis of the census tracts shown below

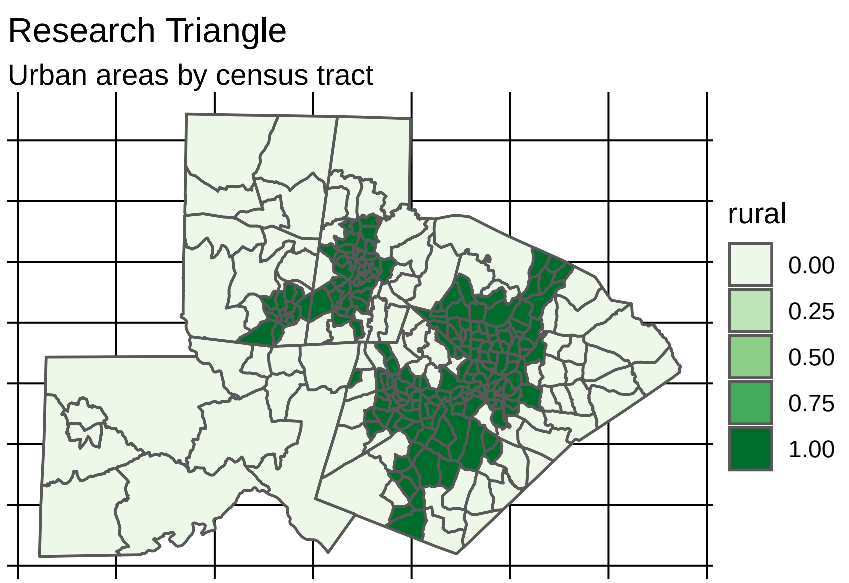


Next we want to visualize which regions in the research triangle area have experienced gentrification:



We can see that the census tracts show above in blue are classified as gentrified and those that are yellow are not.

We hypothesized that whether a census tract is in an urban or rural area would impact whether that region had also experienced gentrification. In order to determine urban vs rural impact on gentrification, we recoded the variable "rural" to be equal to 1



By comparing the locations of gentrified tracts to urban areas, we can see that almost all gentrified tracts are in urban areas. Moreover, many of the gentrified tracts appear to be in and around city centers. This makes sense–we tend to think of gentrification as affecting highly urbanized downtown areas.

ryans mapping stuff

###Part 2: Factors Associated with Gentrification

In part 2, the following research question will be examined:

What factors are associated with and what are the strongest predictors of the gentrification of these areas?

We already determined a model using aic and drop in deviance tests . . . .

##Using Logistic Regression

Creating the logistic model using mutated variable "gent" as our response variable:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-3.00216	1.76126	-1.70455	0.08828	-6.50546	0.42759
collegewhite	0.14487	0.06063	2.38927	0.01688	0.02889	0.26801
whitecollar	-0.05600	0.03098	-1.80772	0.07065	-0.11837	0.00337
privateschool	0.01855	0.05029	0.36882	0.71226	-0.09732	0.10785
nodiploma	0.12286	0.06408	1.91736	0.05519	0.00077	0.25299
highschoolgrad	0.05705	0.04424	1.28955	0.19721	-0.02836	0.14590
collegedegree	-0.03647	0.06220	-0.58638	0.55762	-0.16062	0.08438
income_med	-0.00002	0.00002	-0.86680	0.38605	-0.00005	0.00002
homeprice_med	0.00001	0.00000	2.44891	0.01433	0.00000	0.00002
early_late	-0.00940	0.01852	-0.50727	0.61197	-0.04578	0.02715
moved	0.00344	0.04475	0.07697	0.93865	-0.07898	0.10019

Using backward selection to find the optimal model:

```
## Start:  AIC=241.3
## gent ~ collegewhite + whitecollar + privateschool + nodiploma +
##     highschoolgrad + collegedegree + income_med + homeprice_med +
##     early_late + moved
##
##                  Df Deviance    AIC
## - moved           1   219.31 239.31
## - privateschool   1   219.43 239.43
## - early_late      1   219.56 239.56
## - collegedegree   1   219.65 239.65
## - income_med      1   220.07 240.07
## - highschoolgrad  1   221.00 241.00
## <none>                219.30 241.30
## - whitecollar     1   222.71 242.71
## - nodiploma       1   223.19 243.19
## - collegewhite    1   225.34 245.34
## - homeprice_med   1   225.64 245.64
##
## Step:  AIC=239.31
## gent ~ collegewhite + whitecollar + privateschool + nodiploma +
##     highschoolgrad + collegedegree + income_med + homeprice_med +
##     early_late
##
##                  Df Deviance    AIC
## - privateschool   1   219.44 237.44
## - early_late      1   219.56 237.56
## - collegedegree   1   219.66 237.66
## - income_med      1   220.08 238.08
## - highschoolgrad  1   221.00 239.00
## <none>                219.31 239.31
## - whitecollar     1   222.73 240.73
## - nodiploma       1   223.19 241.19
## - collegewhite    1   225.38 243.38
## - homeprice_med   1   225.65 243.65
##
```

```
## Step:  AIC=237.44
## gent ~ collegewhite + whitecollar + nodiploma + highschoolgrad +
##     collegedegree + income_med + homeprice_med + early_late
##
##                  Df Deviance    AIC
## - early_late      1   219.66 235.66
## - collegedegree   1   219.76 235.76
## - income_med      1   220.21 236.21
## - highschoolgrad  1   221.11 237.11
## <none>                219.44 237.44
## - whitecollar     1   222.86 238.86
## - nodiploma       1   223.25 239.25
## - collegewhite    1   225.44 241.44
## - homeprice_med   1   225.82 241.82
##
## Step:  AIC=235.66
## gent ~ collegewhite + whitecollar + nodiploma + highschoolgrad +
##     collegedegree + income_med + homeprice_med
##
##                  Df Deviance    AIC
## - collegedegree   1   219.98 233.98
## - income_med      1   220.52 234.52
## - highschoolgrad  1   221.25 235.25
## <none>                219.66 235.66
## - whitecollar     1   223.04 237.04
## - nodiploma       1   223.43 237.43
## - collegewhite    1   225.57 239.57
## - homeprice_med   1   226.71 240.71
##
## Step:  AIC=233.98
## gent ~ collegewhite + whitecollar + nodiploma + highschoolgrad +
##     income_med + homeprice_med
##
##                  Df Deviance    AIC
## - income_med      1   220.93 232.93
## <none>                219.98 233.98
## - highschoolgrad  1   222.65 234.65
## - whitecollar     1   223.39 235.39
## - nodiploma       1   224.51 236.51
## - homeprice_med   1   226.87 238.87
## - collegewhite    1   228.82 240.82
##
## Step:  AIC=232.93
## gent ~ collegewhite + whitecollar + nodiploma + highschoolgrad +
##     homeprice_med
##
##                  Df Deviance    AIC
## <none>                220.93 232.93
## - highschoolgrad  1   224.01 234.01
## - whitecollar     1   224.36 234.36
## - nodiploma       1   225.70 235.70
## - homeprice_med   1   226.98 236.98
## - collegewhite    1   228.92 238.92
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.23080	0.24515	-9.09985	0.00000	-2.74176	-1.77646
collegewhite	0.10841	0.03943	2.74954	0.00597	0.03290	0.18858
whitecollar	-0.05692	0.03137	-1.81461	0.06958	-0.11996	0.00321
nodiploma	0.13470	0.06339	2.12476	0.03361	0.01355	0.26289
highschoolgrad	0.06938	0.04021	1.72554	0.08443	-0.00792	0.15022
homeprice_med	0.00001	0.00000	2.37718	0.01745	0.00000	0.00002

Creating a full model to determine if we should add "rural" to the model:

```
## # A tibble: 7 x 5
##   term             estimate  std.error statistic  p.value
##   <chr>               <dbl>      <dbl>     <dbl>    <dbl>
## 1 (Intercept)    -2.88      0.450          -6.39 1.64e-10
## 2 collegewhite    0.0939    0.0400          2.35 1.88e- 2
## 3 whitecollar    -0.0546    0.0318         -1.72 8.61e- 2
## 4 nodiploma       0.118     0.0631          1.87 6.16e- 2
## 5 highschoolgrad  0.0606    0.0406          1.49 1.35e- 1
## 6 homeprice_med   0.0000108 0.00000468      2.31 2.10e- 2
## 7 ruralUrban      0.891     0.476           1.87 6.12e- 2
```

Drop in deviance test:

```
## [1] 220.9284
```

```
## [1] 216.9145
```

```
## [1] 4.013908
```

p-value:

```
## [1] 0.04512643
```

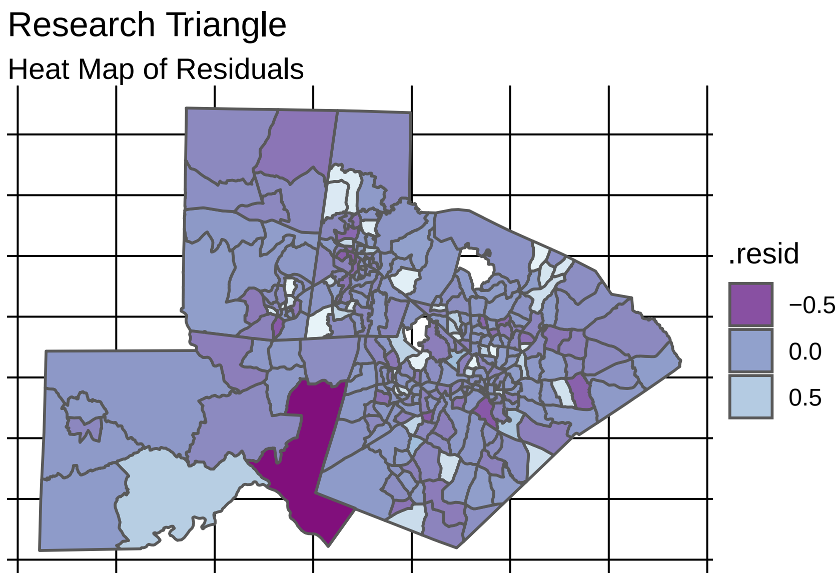
Since the chisq p-value for adding "Rural" to the model is less than .05, we reject the null hypothesis that "Rural" is not a significant predictor of whether or not a region has experienced gentrification.

Therefore we will continue with this full model for the remained of our analysis.

Checking multicollinearity

```
## # A tibble: 6 x 2
##   names              x
##   <chr>          <dbl>
## 1 collegewhite    1.53
## 2 whitecollar     1.17
## 3 nodiploma       1.16
## 4 highschoolgrad  1.41
## 5 homeprice_med   1.18
## 6 ruralUrban      1.03
```

heat map of residuals for independence

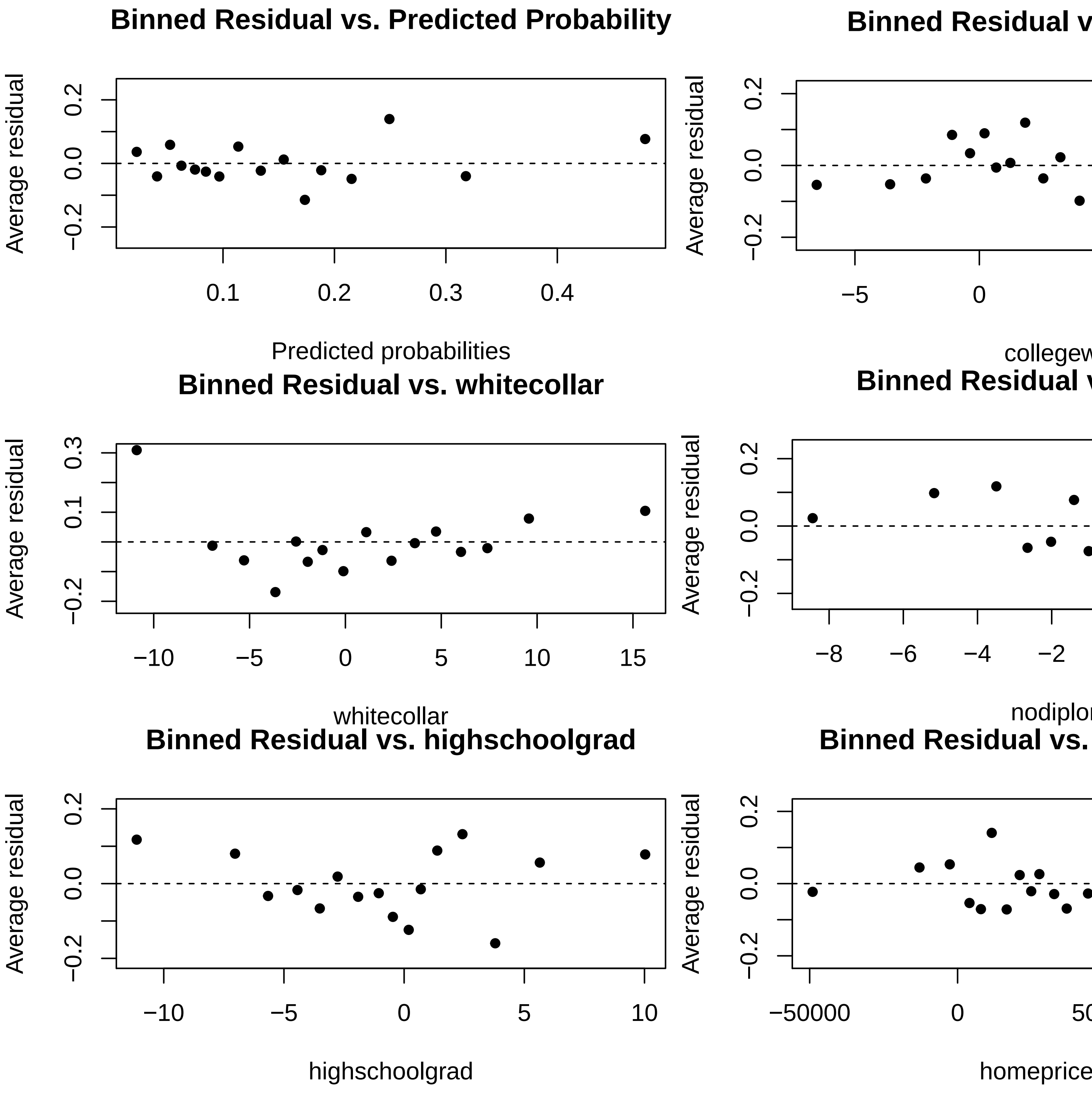


###Assumptions

In order to use the full model with the predictor variables collegewhite, whitecollar, nodiploma, highschoolgrad, homeprice_med, and rural, we must first test how well this model satisfies assumptions.

For testing linearity, we will augment the model with predicted probabilities and residuals in order to examine binned residual plots for predicted probability and numeric variables.

```
## # A tibble: 285 x 15
##    .rownames  gent collegewhite whitecollar nodiploma highschoolgrad
##   * <chr>     <dbl>        <dbl>       <dbl>     <dbl>          <dbl>
##   1 1             0       -8.46        -16.4    7.20           -4.89
##   2 2             0        2.96          3.2   -0.640          -0.658
##   3 3             0       -0.735         4.3    0.807          -1.44
##   4 4             0       -4.08          6.6    0.0800         -1.86
##   5 5             0        2.31         -2.8    1.16           -4.08
##   6 6             0       -1.25          0.5   -0.388          -4.55
##   7 7             0       -0.224         4.3   -7.19            5.87
##   8 8             0        2.17         18.4   -5.19           11.3
##   9 9             0       -1.92         -5.6    2.05            0.321
## 10 10             0        5.44          6.9   -0.837           0.687
## # ... with 275 more rows, and 9 more variables: homeprice_med <dbl>,
## #   rural <chr>, .fitted <dbl>, .se.fit <dbl>, .resid <dbl>, .hat <dbl>,
## #   .sigma <dbl>, .cooksd <dbl>, .std.resid <dbl>
```




```
## # A tibble: 2 x 2
##   rural mean_resid
##   <chr>      <dbl>
## 1 Rural  -5.38e-11
## 2 Urban  -4.71e-12
```

The linearity assumption is satisfied. The binned residuals vs. predicted probability plot shows irregularity with a very slight clustering of residual values below 0.0. The binned residuals vs. collegewhite plot shows irregularity. The binned residuals vs. whitecollar plot shows irregularity, with a slight clustering of residual values below 0.0 and a slight increase in residual values as you move right. The binned residuals vs. nodiploma, binned residuals vs. highschoolgrad, and binned residuals vs. homeprice_med show complete irregularity. For the predictor variable rural, which has two categories rural and urban, both mean residuals are very close to zero. There is no strong indication of nonlinearity; therefore, we can assume that there is a linear relationship between log(gent) and the predictor variables.

To discuss randomness and independence, we must go back to the source of our data. All of the data we are using is sourced from the Census Bureau's annual American Community Survey and official North Carolina demographic data. According to the census sampling techniques and methodology, we can reasonably assume that randomness and independence are satisfied. Read more here: https://www.census.gov/programs-surveys/sipp/methodology.html

Interpreting Model Coefficients

Now that we've confirmed that it satisfies assumptions, let's take a look at our chosen logistic model again:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.877	0.450	-6.392	0.000	-3.867	-2.076
collegewhite	0.094	0.040	2.349	0.019	0.017	0.175
whitecollar	-0.055	0.032	-1.716	0.086	-0.119	0.006
nodiploma	0.118	0.063	1.869	0.062	-0.002	0.246
highschoolgrad	0.061	0.041	1.494	0.135	-0.018	0.142
homeprice_med	0.000	0.000	2.309	0.021	0.000	0.000
ruralUrban	0.891	0.476	1.872	0.061	0.018	1.915

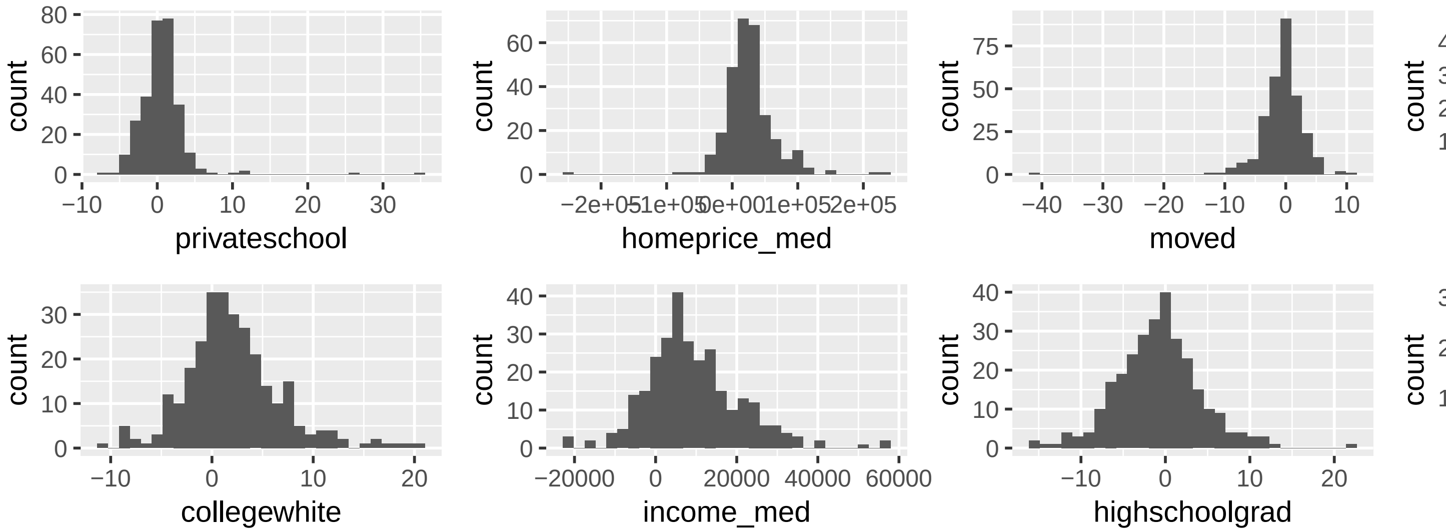
We would like to discuss the variables that have the most impact on the response variable gent. Therefore, we will discuss variables with p-values of <0.05. The variable collegewhite seems to have a reliably strong impact on gent: holding all other variables constant, with a unit change in collegewhite, the odds of gentrification are expected to multiply by a factor of $\exp(0.089) = 1.093$. However, this impact is not as strong as that of the rural variable. According to the model coefficient for the term ruralUrban, holding all other variables constant, the odds of gentrification for an urban area is expected to be 2.55 that of a rural locale. We would like to suggest that the change in college-educated whites in a county and urban character likely greatly impact "gentrification" as we have classified it (a significant decrease in black population).

Section 4: Limitations

Section 5: Conclusion

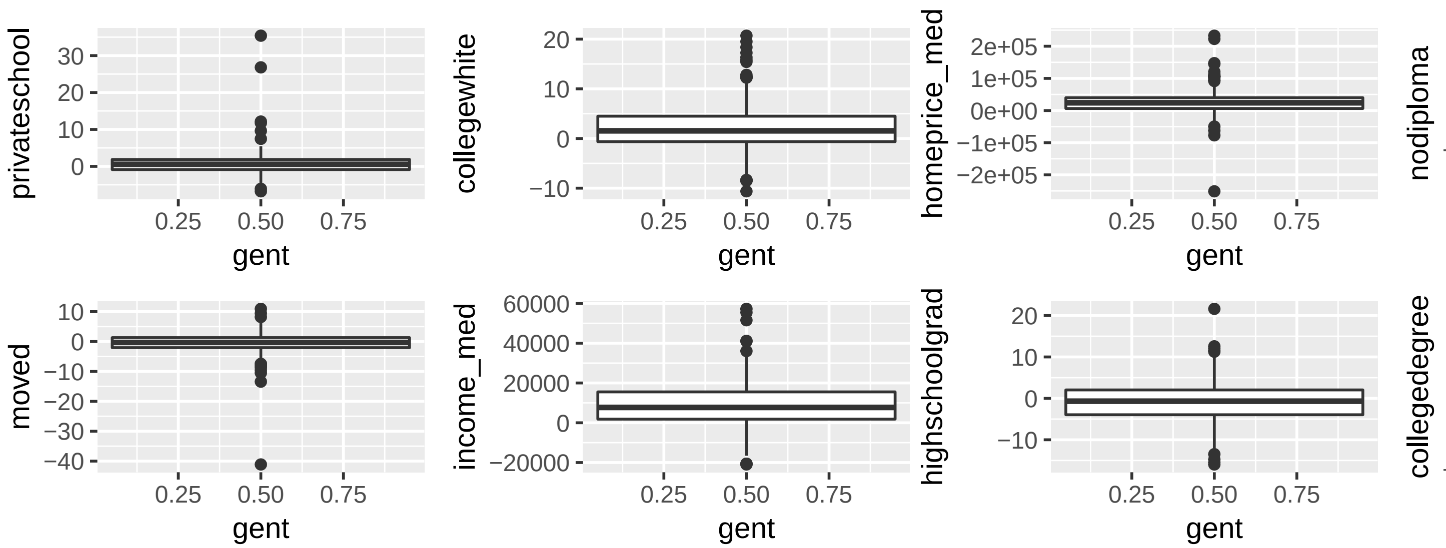
Section 6: Additional Work

More Univariate EDA:



Each predictor variable is normally distributed around 0.

Bivariate EDA:



The relationship between the response variable “gent” and the predictor variables are all each roughly normal.