Final Write-Up

Apache Juntion Armchairs: Ellie, Ryan, Sude and Darren

4/20/2020

Section 1: Introduction

In the past 10 years, development in the Research Triangle area has skyrocketed, and the face of downtown has changed drastically. Alongside the rapid growth, gentrification has become a major issue in the region. This 2018 New York Times article sums up the changes in Durham well: "What has been largely overlooked is the cultural displacement that can accompany rapid urban change: the sense that home is not home anymore, at least for a portion of the population" (https://nyti.ms/2VyxTfs). Where black and brown people are being pushed out, they are being replaced by white homeowners in different tax brackets. A 2019 New York Times article describes the type of demographic replacement that is occuring: "In South Park, a neighborhood with picturesque views of the Raleigh skyline, the white home buyers who have recently moved in have average incomes more than three times that of the typical household already here" (https://nyti.ms/2zuVPb). Higher incomes, higher home prices, and more white-collar workers tend to be associated with this kind of demographic change. Our research captures what factors lead to the displacement of the black population in the Research Triangle.

Our analysis focuses on the percentage point change between 2010 and 2017 in several demographic characteristics for census tracts in Chatham, Durham, Orange, and Wake counties. We look at the college educated white population, the population of several different education levels, the percentage of white collar workers, the median income, median home price, the percentage of children in private school, among other variables. The response variable is a binary variable that categorizes census tracts as either "gentrified" or "not gentrified," based on the percentage point change in the black population. The distribution of the percentage point changes in the black population across all census tracts (omitting any NAs in the dataset) is 6.76. The mean is very close to zero, so we have decided to classify a census tract as "gentrified" if its decline in black population is greater than 6.76 percentage points.

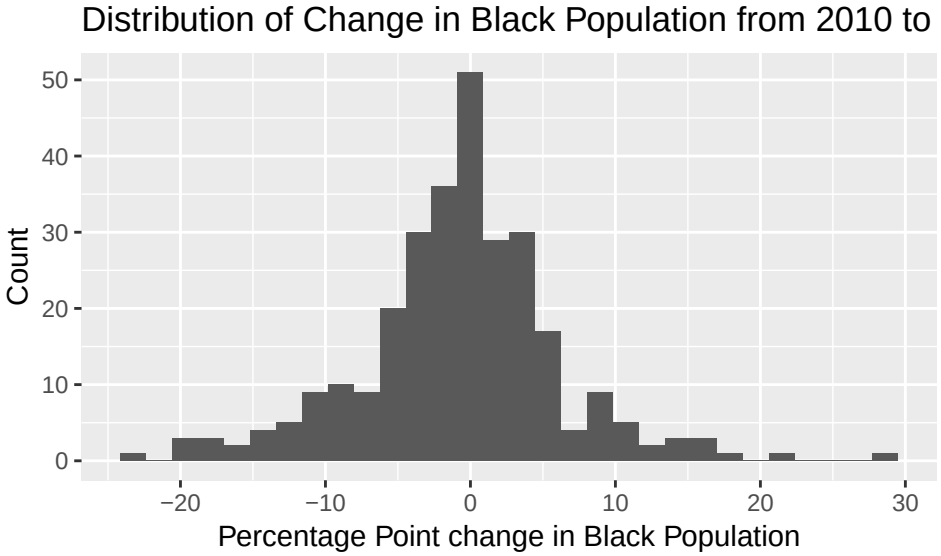
We hypothesize that census tracts in which there is an increase in college educated whites, an increase in median household price and an increase in the amount of people who obtain a bachelor degree or higher will correspond to a decrease in Black population. In addition to exploring the mechanisms behind this displacement, we are interested in investigating where this displacement occurs. We want to be able to pinpoint the neighborhoods with the highest rate of incoming college-educated white people and the highest rate of Black exodus.

We put together a dataset of relevant variables from The Census Bureau's annual American Community Survey, which conducts a wide range of demographic surveys that covers almost everything a decennial census covers. The data can be found here: https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.

All of our demographic data is framed in terms of percentage point change from 2010-2017. By measuring our variables in percentage point changes, we are making our variable more robust to population changes in the years between 2010-2017. We are more focused on the proportion of demographic changes, as opposed to changes of a particular demographic group in a vacuum. Our "metadata" file explains what each variable is actually measuring. Median home price and median household income are the only two variables measured in monetary terms, as opposed to PP change.

As we began our analysis we explored each of the variables in our data set in order to get a better sense of univariate and bivariate distributions in the data. Most importantly, in order to assess whether gentrification is taking place in an area we looked at the change in the Black population because we would use this change as our response to determine if the census tract was considered gentrified or not.

Here the distribution of change in Black population is shown, we use this distribution to determine an appropriate threshold for gentrification.



The standard deviation is = 6.765474. We will use this value (-6.765474) as the threshold to determine if gentrification has occurred in a census tract. If a tract has experienced more than a -6.765 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.

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

In our data set we have 244 observations that are not considered gentrified and 44 that are.

After examining the rest of 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.

Section 2: Regression Analysis:

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?

First, we create the response variable "gent" to represent whether a census tract is gentrified or not. As described above, if a census tract has less than or equal to -6.765 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". The response variable "gent" will be used in order visualize and eventually create a logistic model.

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

Show below is a plot of the 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

I coulnt knit–> make sure to make code chunck again!!!

“{r} shapeurban <- read_sf(dsn = “data”, layer = “MunicipalBoundaries”)

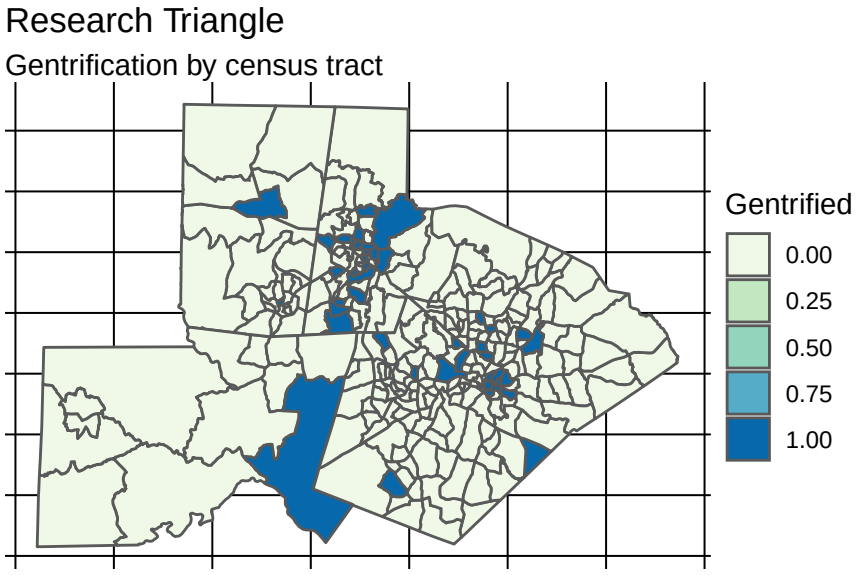
shapeurban_aea <- st_transform(shapeurban, st_crs(shape))

cities <- cbind(shapeurban_aea, st_coordinates(st_centroid(shapeurban_aea)))

cities<- as.data.frame(cities)

```{r}
ggplot(data = merged) +
  geom_sf() +
  labs(title = "Research Triangle",
       x = "Longitude", y = "Latitude") +
  geom_text_repel(data = cities, aes(x = X, y = Y, label = MunicipalB),
        fontface = "bold", nudge_x = c(1, -1.5, 2, 2, -1), nudge_y = c(0.25,
            0.25, 0.5, 0.5, 0.5))
```

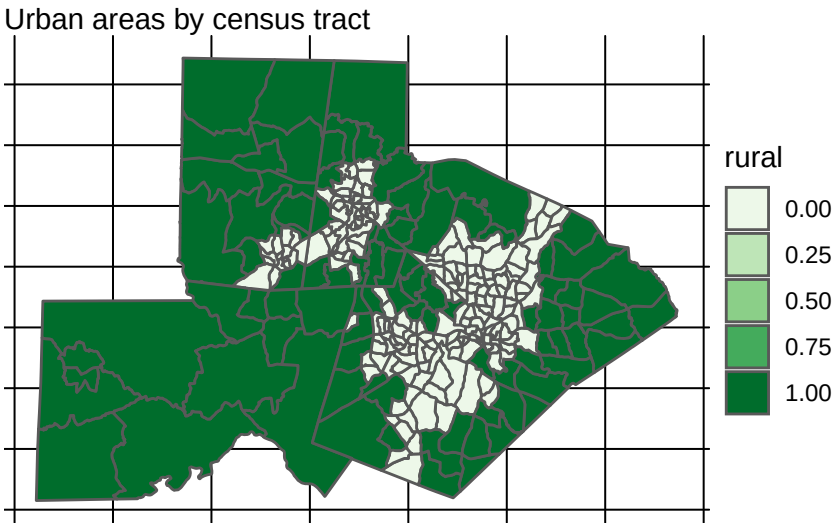
Next we want to visualize which regions in the research triangle area have experienced gentrification by plotting “gent” across the region:



The census tracts shown 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” if the census tract is considered to be rural and “0” if the census tract is in an urban area.

Next, we want to visualize which regions in the research triangle area are considered urban/rural by plotting our new variable, “rural”:



The census tracts shown above in dark green are classified as rural and those that are light green are urban.

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.

Now that we have a better sense of where gentrification is happening in thr research triangle, we would like to take a closer look at what factors are associated with this change.

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?

In order to predict where gentrification is taking place we again looked at the change in the Black population. We use the catagorical variable created in part one above, "gent" where a "1" is coded for census tracts that we classify as gentrified and "0" for census tracts which have not been gentrified. Since gent is a catagorical variable with 2 outcomes we fit a logistic model.

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

It is clear that not every variable is a significant predictor of gentrification, so we next use backward selection, with AIC criterion, to find a reduced, optimal model.

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

Logistic model predicting gentrification:

`log(pi_hat/(1-pi_hat)) = -2.2308 + (0.1084)*(collegewhite) - (0.0569)*(whitecollar) + (0.1347)*(nodiploma) + (0.0694)*(highschoolgrad) + (0.00001)*(homeprice_med)`

We originally 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 we visualized where the gentrified areas were and compared to the map of urban vs. rural regions. In the visualizations, many of the gentrified tracts appeared to be in and around city centers. Now, we want to determine if the variable "rural" is significant and should be added to our model.

We create a full model that includes "rural" and conduct a drop in deviance test to determine if we should add "rural" to the model:

The null hypothesis is that the coefficient for rural does not add significant information to our model. The alternative hypothesis is that the coefficient for rural does add significant information to our model.

H0: B_rural = 0 HA: B_rural != 0

```
## # 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
```

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

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

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

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

The drop in deviance test has a test statistic of 4.014 and a p-value of 0.0451.

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 remainder of our analysis.

Now that we have determined Rural is a significant predictor variable, we consider adding interaction terms with rural to our model. We use k-fold cross validation to determine if we should add the interactions homeprice_med*rural and collegewhite*rural. We start by looking at the 5-fold cross validation results for the model above using the predictors: collegewhite, whitecollar, nodiploma, highschoolgrad, homeprice_med and rural. We fit a model on each training set and calcualte the testing error. Next, we repeat the same process using the predictors: collegewhite, whitecollar, nodiploma, highschoolgrad, homeprice_med, rural, and our new interaction terms, homeprice_med*rural and collegewhite*rural.

```
## # A tibble: 1 x 2
##   mean_train_mse mean_test_mse
##            <dbl>         <dbl>
## 1           5.37          5.37
```

```
## # A tibble: 1 x 2
##   mean_train_mse mean_test_mse
##            <dbl>         <dbl>
## 1           5.67          5.67
```

Finally we compared the estimated testing error for both models:

Model 1 (excluding interaction terms): 5.369539 Model 2 (including interaction terms): homeprice_med*rural and collegewhite*rural: 5.671572

Although the testing errors are very close, Model 1 performs better than Model 2 when predicting if a census tract is gentrified. Therefore, we will continue with the model that does not include the interaction terms.

Before we begin using our model we want to make sure that multicollinearity is not a problem in our model. In order to avoid collinearity of response variables we want to confirm that no two variables are highly correlated with one another. We do so by checking multicollinearity and looking for variables with high VIF (Variance Inflation Factor).

```
## # 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
```

Since the VIF for all of our variables is relatively low and none even come close to 10+, we can be confident that multicollinearity is not a problem in our model and none of the variables are highly correlated with one another.

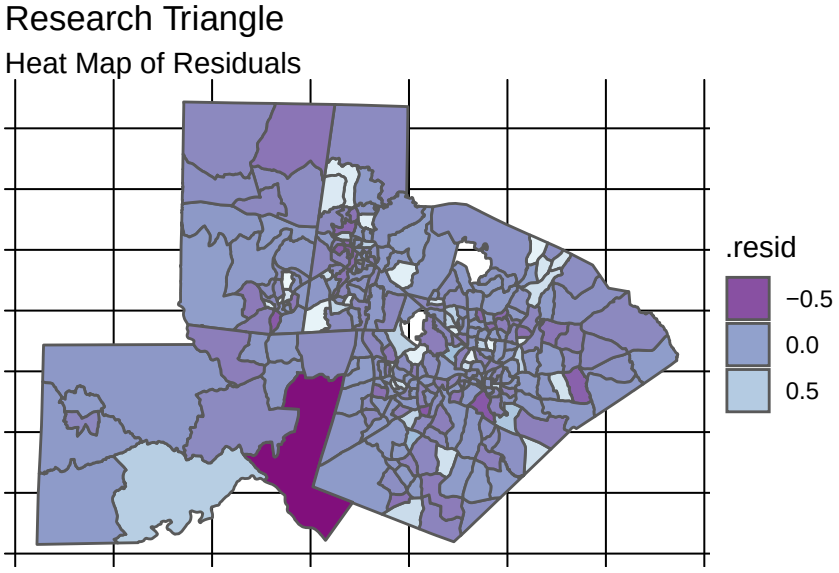
Assumptions

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. A complete look at assumptions is present in section 6 (Additional work), for now we will just discuss the independence assumption.

```
## # 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>
```

We created a heat map of residuals in order to examine the independence assumption:



We are able to see from the map that the distribution of residuals is fairly random. From this, we are able to conclude the independence is satisfied.

Section 3: Discussion

Now that we've confirmed that our model satisfies assumptions, let's take another 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

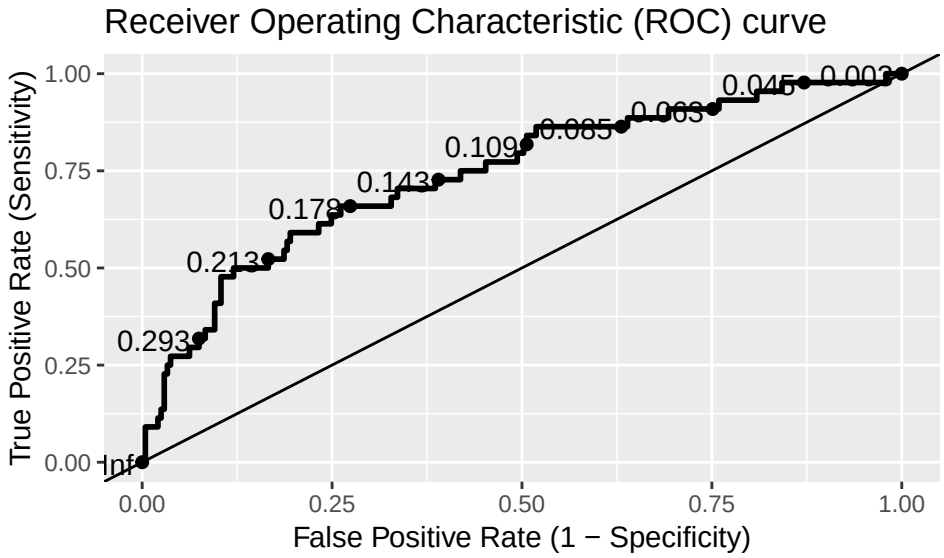
```
log(pi_hat/(1-pi_hat)) = -2.877 + (0.094)*(collegewhite) - (0.055)*(whitecollar) + (0.118)*(nodiploma)
+ (0.061)*(highschoolgrad) + (0.00001)*(homeprice_med) + (0.891)*(ruralUrban)
```

This entire process produced the final logistic model, shown here. The intercept, -2.877, describes the log-odds that a census tract has experienced gentrification when the percentage point change (from 2010-2017) in college educated whites, white collar, no diploma, highschool grad, and home price median is equal to zero and the census tract is in a rural area.

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(0.094).

In addition, the rural variable also appears to have a strong impact. 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 exp(0.891) 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).

In order to use our model to predict whether an area is gentrified or not, we creatie a Receiver Operating Characteristic (ROC) curve. The ROC curve will both help us assess how well the model fits the data as well as pick a threshold for our logistic model.




```
## [1] 0.7424557
```

Since our AUC 0.742 we can see that the logistic model fits the data fairly well.

In order to pick an appropriate threshold, we have to consider the potential consequences of misclassifying an area as gentrified or not gentrified. If we erroneously classify an area as not gentrified, we could miss an opportunity to control the effects of gentrification and we risk letting growth have far-reaching cultural consequences. It we call an area gentrified, somebody will likely conduct further research on the area before making any policy decisions, so the risk of false-positive classification is not very high.

The Apache Junction Armchairs, being socially responsible policymakers, are more worried about falsely classifying an area as not gentrified than we are about falsely classifying an area as gentrified. The social costs are higher in the former scenario. With the desire to reflect there priorities in mind, we used the ROC curve above to determine this threshold of 0.143 which will prioritize sensitivity when determining if an area is gentrified or not gentrified.

Section 4: Limitations

The biggest limitation of our study is the definition of gentrification. Much prior research has been done on the subject, and many of those studies define gentrification as some combination of demographic characteristics and poverty rate. Our characterization of gentrification is hamstrung by the fact that it does not incorporate any economic factors. In addition, it would be helpful to analyze more variables. Gentrification is a process that incorporates a variety of different socioeconomic groups, so more racial and economic data would have been helpful for analyzing gentrification from a number of different angles.

In addition, we turned 'Black percentage point change' into a logistic model by picking an arbitrary threshold for determining if a tract was gentrified or not based on a rough measure of the Standard Deviation. Since the average census tract (omitting NA's) fell 6.76 PPs away from the mean of the distribution of "black" ,

which was roughly 0, we decided to convert any tract with a value "black" of less than or equal to 6.76 into a gentrified tract, and all others were not gentrified. This is not grounded in any social science research, but rather it is relative to the distribution. A more complete analysis would determine an appropriate threshold of gentrification based on some mixture of quantitative and qualitative research.

Section 5: Conclusion

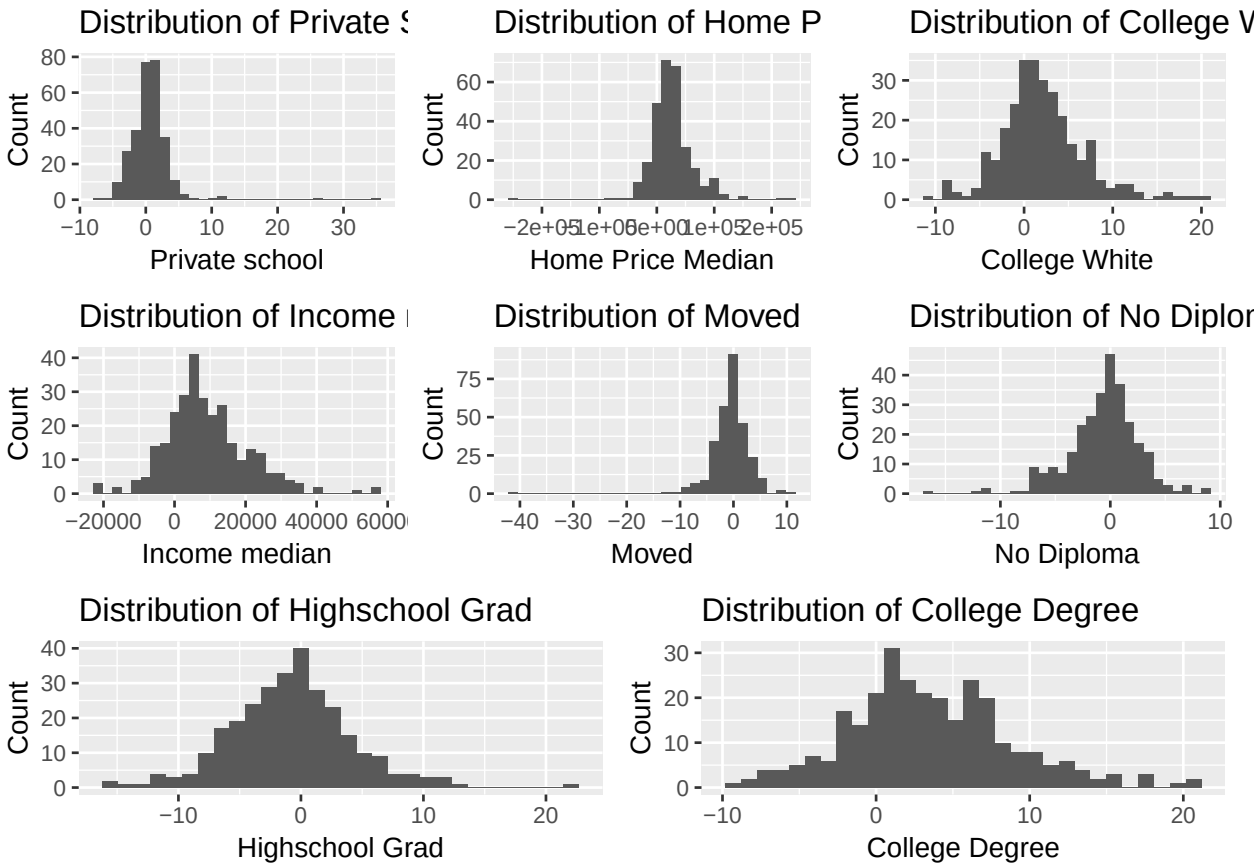
Our study was motivated by two questions: where in the Research Triangle does displacement of black population occur, and which factors contribute to the displacement of the black population in the Research Triangle? The spatial data suggests a strong association between gentrification and urban character–most gentrified tracts were in urban locales, and many were also in city centers. We created a logistic model that used rural/urban classification and percentage change in college-educated whites, white-collar jobs, those with no diploma, those with just a high school education, and median home price to predict gentrification. From our hypothesis—-that the change in college educated whites, median home price, and those with a bachelor degree would be the strongest predictors of gentrification—-median home price change and proportional college-educated white change made it into our model. Classification as urban had the strongest impact on gentrification, which aligns with our spatial analysis. Overall, we conclude that the proportional change in college educated whites and an area's character as urban or rural are the two strongest predictors of gentrification in the Research Triangle area. In other words, the displacement of the urban black population is associated with an influx of college-educated whites in those urban areas.

Of all the variables we started with, three education-based variables—-proportional change in college-educated whites, those with no high school diploma, and those with only a high school diploma—- made it into our final model. This is an interesting link—-further research could be done on the relationship between gentrification and education at a variety of socioeconomic levels.

Gentrification is a complex phenomenon. While we focused specifically on the change in Black population to determine gentrification within our study, gentrification can impact a variety of different socioeconomic groups. For a more rigorous and complete understanding of demographic shifts and gentrification in the Research Triangle region or North Carolina at large, the different socioeconomic groups should be taken into consideration. It would be very productive to study the demographic shifts of the Hispanic population and low-income population at large (and by specific racial categories). Studying shifts in public housing projects and populations in the Research Triangle region would also be productive in gaining a more complete picture of gentrification. This would allow policymakers, non-governmental organizations, and other interested parties to understand and mitigate the impacts of gentrification on the local communities.

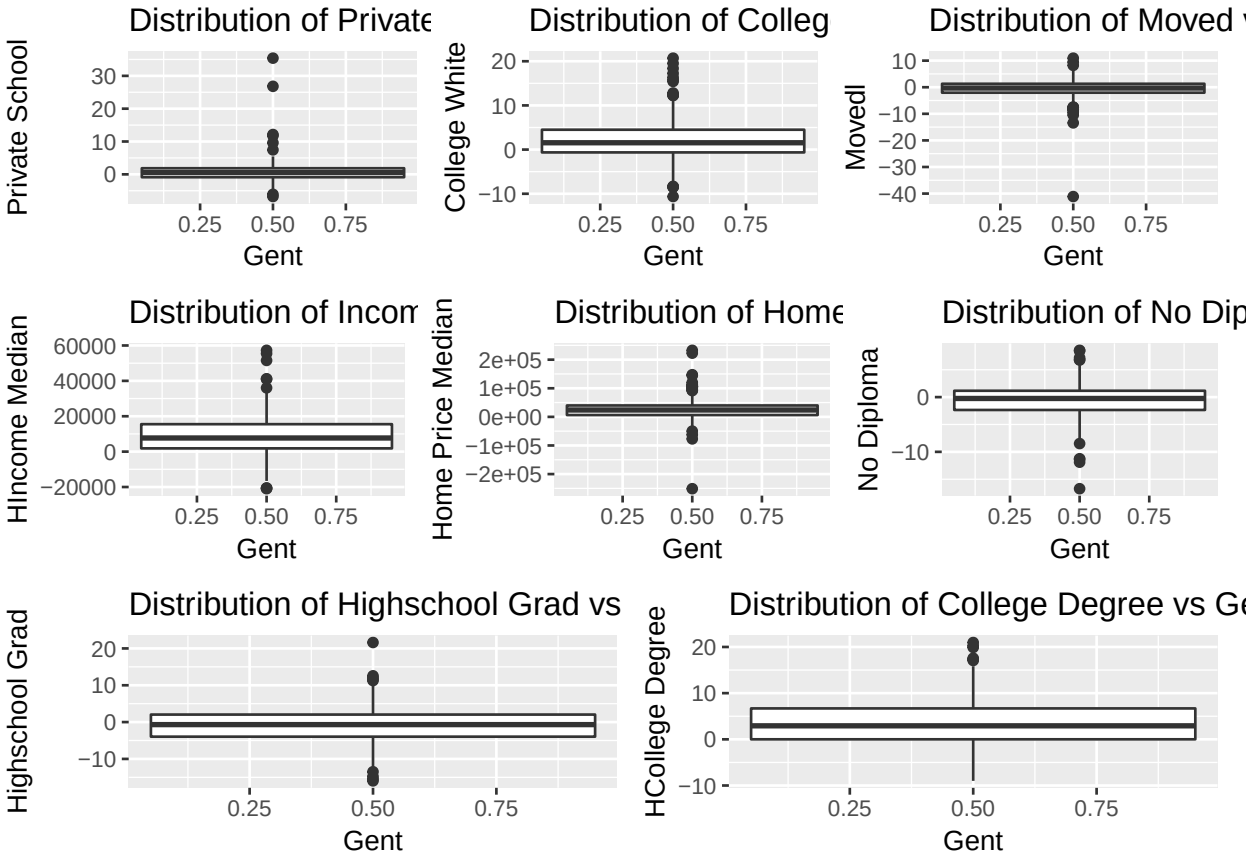
Section 6: Additional Work

Univariate EDA



Each predictor variable is normally distributed around 0.

Bivariate EDA:



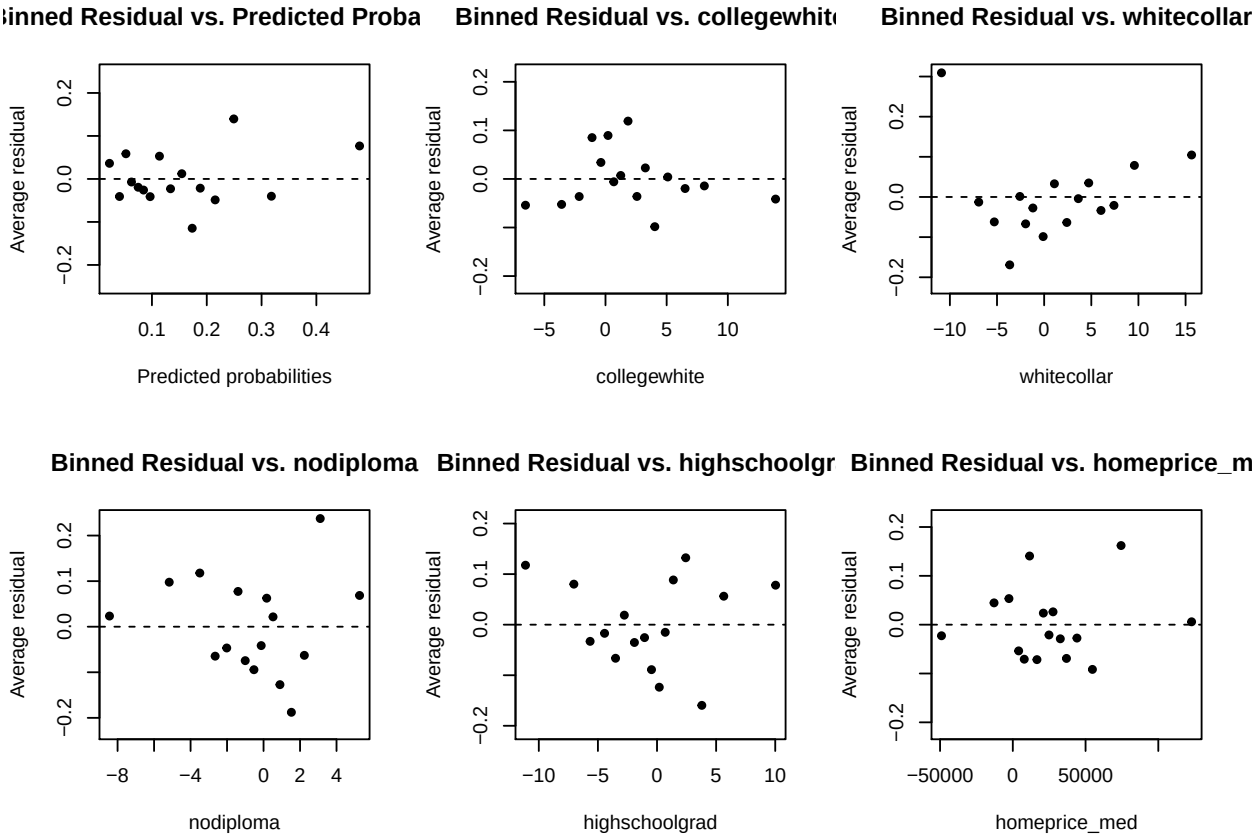
The relationship between the response variable "gent" and the predictor variables are all each roughly normal.

###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.

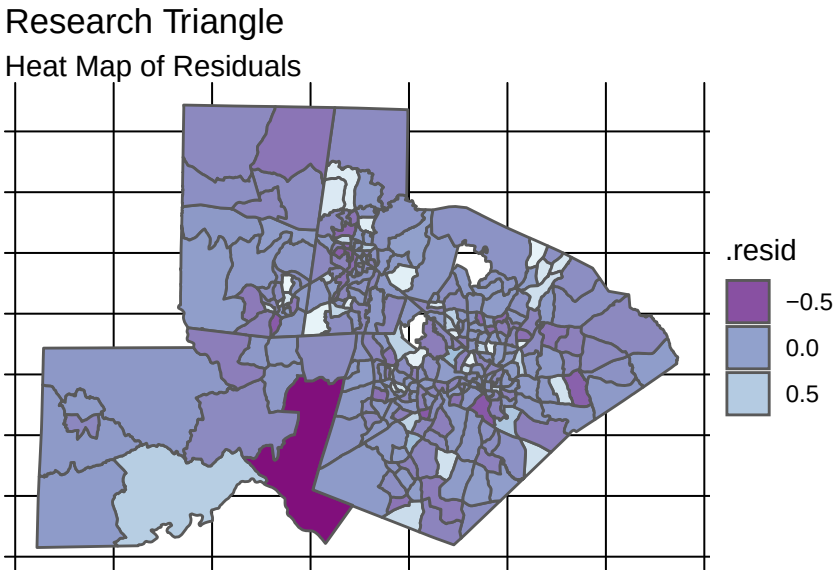
```
## # A tibble: 285 x 15
##    .rownames  gent collegewhite whitecollar nodiploma highschoolgrad
##  * <chr>     <dbl>        <dbl>       <dbl>     <dbl>          <dbl>
##  1 1             0       -8.46        -16.4    7.20           -4.89
##  2 2             0        2.96          3.2   -0.640          -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.

We created a heat map of residuals to examine the independence assumption:



We are able to see from the map that the distribution of residuals is fairly random. From this, we are able to conclude the independence is satisfied.

To discuss randomness, 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 is satisfied. Read more here: https://www.census.gov/programs-surveys/sipp/methodology.html