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. Jounralist Amanda Abrams 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, white homeowners in different tax brackets are coming in. 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.

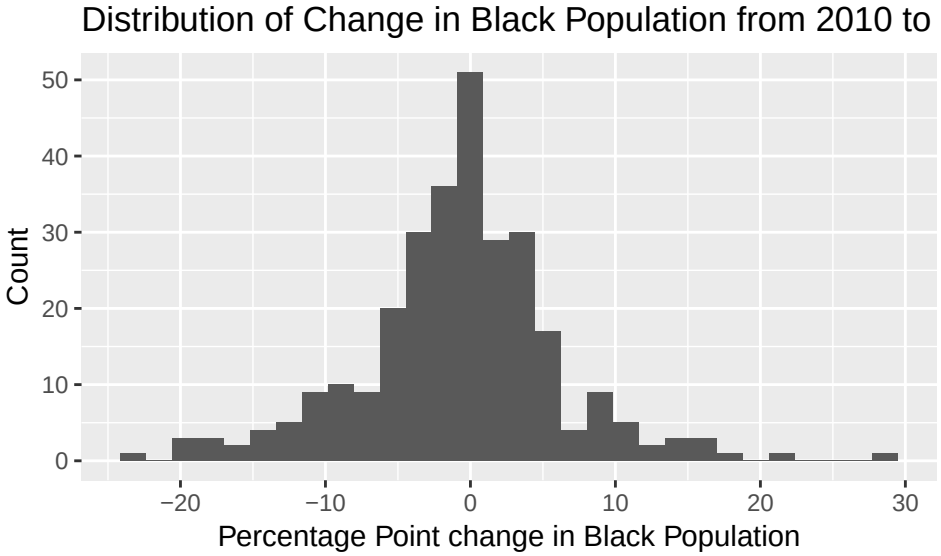
Our analysis focuses on the percentage point change between 2010 and 2017 in several demographic characteristics for census tracts in Chatham, Durham, Orange, and Wake counties. We look at the college educated white population, the population of several different education levels, the percentage of white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

We hypothesize that census tracts in which there are increases in college educated whites, median household price and 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 cover almost everything a decennial census covers. The data can be found here: https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.

By measuring our demographic variables in percentage point changes, we are making our variables 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 of 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. We use this distribution to determine an appropriate threshold for gentrification.



The standard deviation, omitting NAs, is 6.765474. We will use this value as the gentrification threshold. If a tract has experienced more than a -6.765 percentage point change in the Black population, we will consider that census tract "gentrified." Even though the mean is not exactly 0, it is close enough that we feel one standard deviation away from 0 is a sufficient threshold for gentrification.

gent	n
0	244
1	44

In our data set we have 244 non-gentrified census tracts and 44 gentrified ones.

After examining the rest of the univariate and bivariate 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 roughly normal.

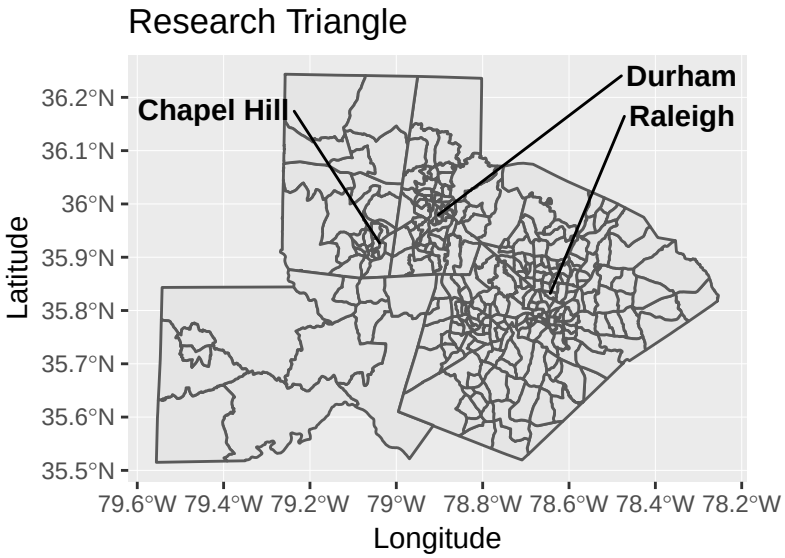
Section 2: Regression Analysis

Part I: Location of Gentrification

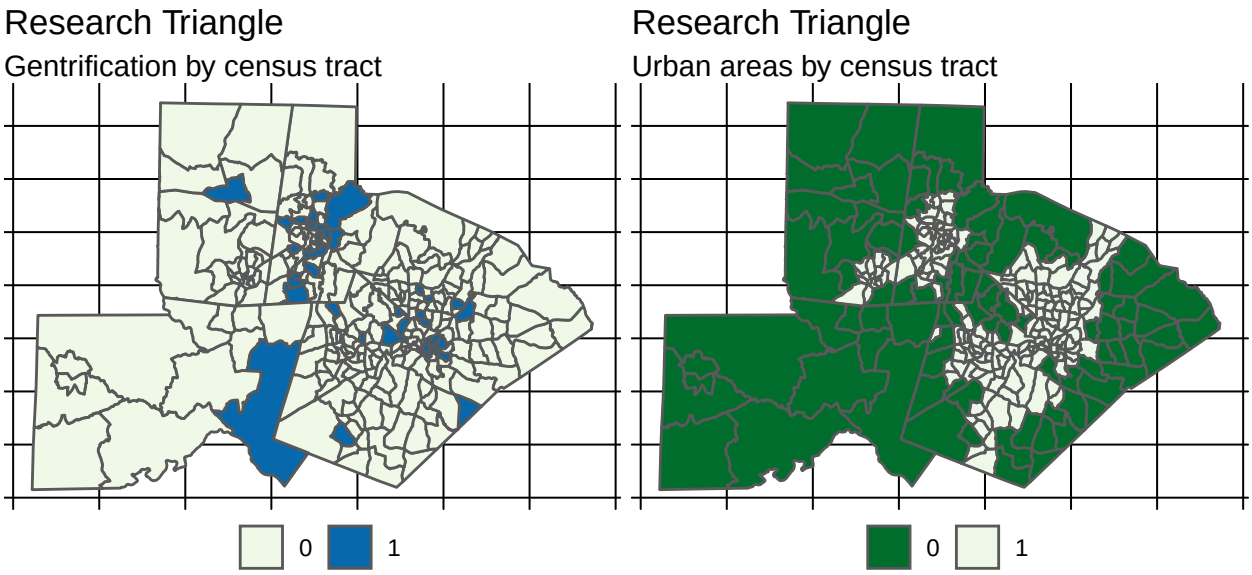
In part I, we ask: where in the Research Triangle is gentrification occurring the most?

In order to determine the location of gentrification we analyze spatial data:

Shown below is a plot of the research triangle area (Chatham, Durham, Orange, and Wake counties) with the three major cities identified.



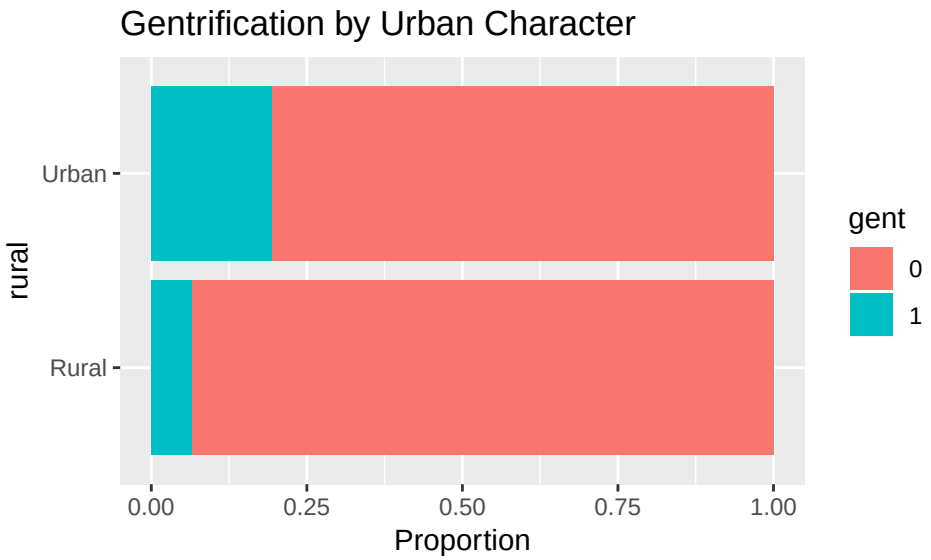
We hypothesize that urban character of an area will impact the odds that it has experienced gentrification. Below, we visualize the urban/rural breakdown of the research triangle, and which tracts have experienced gentrification. In the lefthand map, the blue census tracts are gentrified and the white tracts are not. On the righthand map, the dark green tracts are rural, and the white tracts are urban.



By comparing the locations of gentrified tracts to urban areas, we can see that almost all gentrified tracts are in urban areas.

rural	gent	n
Rural	1	6
Urban	1	38

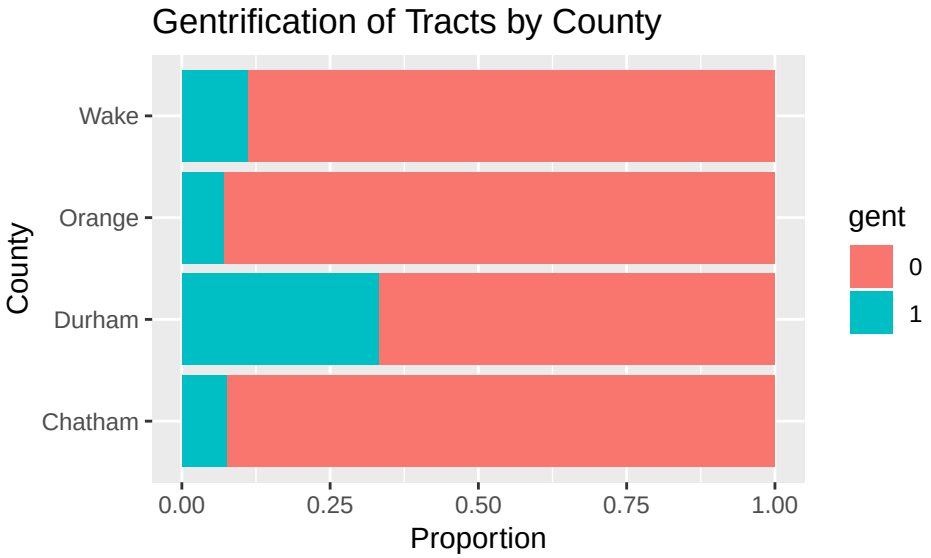
Of the 44 gentrified tracts, only 6 are rural. Moreover, many of the gentrified tracts appear to be in and around city centers. This makes sense–we tend to think of gentrification as affecting highly urbanized downtown areas.



We can see that gentrified tracts tend to be around city centers, but we do not know which cities. The Triangle is a diverse region with three major cities. Raleigh is the heart of Wake county, Durham is the major city in Durham county, Chapel Hill is the population center of Orange county, and Chatham county is mostly rural. In order to get an idea of which areas in the triangle are most affected by gentrification, we look at which counties have the highest number and proportion of gentrified tracts.

County	gent_tracts	census_tracts	proportion_gent
Chatham	1	13	0.077
Durham	20	60	0.333
Orange	2	28	0.071
Wake	21	187	0.112

Unsurprisingly, Chatham county only has one gentrified tract. Every census tract in the county is rural, and its one gentrified tract is one of the relatively rare rural gentrified tracts. Orange county, interestingly, has just two gentrified tracts, and only 7.1% of its tracts are gentrified. Wake County has the most gentrified tracts in the region, but as the biggest county in the region, its percentage of gentrified tracts is also relatively low, at just 11.2%. Durham county clearly experiences the most gentrification. Its number of gentrified tracts is about equal to that of Wake county, but its proportion is by far the highest at 33.3%.



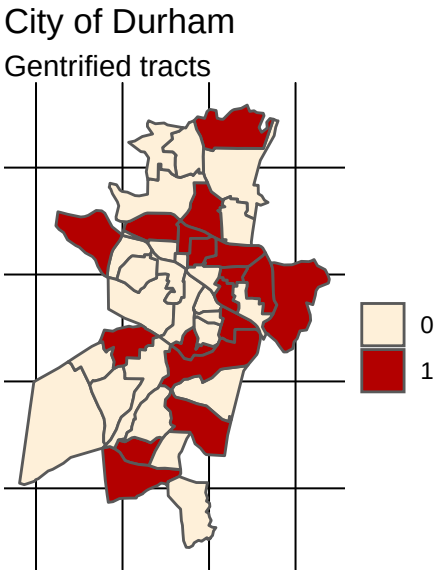
Looking at the breakdown of gentrified tracts by county is useful, but it still only hints at broad strokes. Based on our spatial analysis, we suspect that gentrification is most prevalent in cities. We know which cities are in which counties, and we know which counties experience the most gentrification, but we want to be more specific in our analysis. In order to pinpoint the scale of gentrification within each city, we filter out the rural counties to look at the proportion of gentrified urban tracts within each county. The only urban census tracts in Durham county are in the city of Durham, and the same goes for Orange county and Chapel Hill. The only urban tracts in Wake county are in Raleigh's sprawling metropolis, which technically incorporates other cities but is commonly understood as the greater Raleigh area.

City	gent_tracts	census_tracts	proportion_gent	nn
Chapel Hill	1	15	0.067	1
Durham	18	43	0.419	1
Raleigh	19	138	0.138	1

In Chapel Hill, where gentrification is effectively a non-factor, the percentage of gentrified urban tracts is roughly the same as for its whole county. In both Durham and Raleigh, where there are around 20 gentrified tracts, the proportion of gentrified tracts increases when we take out the rural tracts. Raleigh shows just a slight increase to from 11.2% to 13.8%, but Durham's percentage jumps from 33.3% to 41.9%.

These three cities show us three different gentrification stories. Durham, a city with a storied African-American heritage and a historically large Black population, has experienced significant gentrification. In seven years, over 40% of its neighborhoods have seen at least a 6.7% decline in the black population. That is an incredibly significant demographic shift. In Raleigh, the cultural displacement is not nearly as severe. About 14% of its neighborhoods have been gentrified, which is not so significant of a shift. While Raleigh has certainly developed rapidly over this time, it has not undergone the same sort of demographic shellacking as Durham. This could be due to either its demographic history or responsible housing and growth policies. Chapel Hill, a city that has historically been predominantly white, has not experienced gentrification like Raleigh or Durham.

Durham is clearly the city most affected by gentrification. Since over 40% of its tracts have been gentrified between 2010 and 2017, it will be useful to zoom in on the city to see which parts of town are being gentrified.



East Durham, a hisotrically Black part of town, appears to be experiencing the most gentrification. Below, we look at the city's racial divide and how it compares to gentrified areas.

side	2010 Black Population	proportion_black_pop	n
East	49690	0.71	1
West	20257	0.29	1

side	gent_tracts	census_tracts	proportion_gent	n
East	13	24	0.542	1
West	5	19	0.263	1

In 2010, East Durham contained 71% of the city's Black population. Between 2010 and 2017, 54% of those census tracts became gentrified. In West Durham, where 29% of the Black population lives, only 26% of the census tracts became gentrified. That is a significant disparity. Within Durham, we can see that gentrification is primarily happening in the predominantly Black part of town. This is evidence of what Amanda Abrams calls "the sense that home is not home anymore, at least for a portion of the population." All across this part of town, the Black population is decreasing at an unusually high rate. Comparing East to West Durham can give us some insight into why Raleigh, which has experienced a similar rate of growth to Durham over the past decade, has much lower gentrification percentages. East Durham and West Durham are governed by the same municipal government. While some zoning and permit differences might apply, both parts of town are mostly subject to the same development regulations. We speculated that Raleigh's low rate of gentrification might be due to thoughtful governance and sound policy, but the East/West Durham comparison suggests that demographic history is a important factor in determining this specific measure of demogrgaphic change. In other words, where there are more Black people, there is more potential for a decrease in the Black population.

Now that we have a sense of how cities are impacted by gentrification and further explored the case study of Durham, we will take a closer look at what factors are associated with this demographic change.

Part 2: Factors Associated with Gentrification

In Part 2, We aim to examine the factors associated with and the strongest predictors of gentrification.

In order to predict where gentrification is taking place, we fit a model with gentrified as the response variable. Since gentrified is a categorical 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 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 hypothesized that whether a census tract is urban or rural would impact the gentrification of an area. In order to determine urban vs rural impact we compared a map of urban tracts to gentrified tracts. Many of the gentrified tracts appeared to be in and around city centers. Now, we want to determine if urban character 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

term	estimate	std.error	statistic	p.value
(Intercept)	-2.877	0.450	-6.392	0.000
collegewhite	0.094	0.040	2.349	0.019
whitecollar	-0.055	0.032	-1.716	0.086
nodiploma	0.118	0.063	1.869	0.062
highschoolgrad	0.061	0.041	1.494	0.135
homeprice_med	0.000	0.000	2.309	0.021
ruralUrban	0.891	0.476	1.872	0.061

```
## [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.

We saw earlier that not all counties are gentrified equally. To test whether or not to incorporate the variable County into the model, we run another drop in deviance test, this time adding County to the model.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.925	1.079	-2.710	0.007	-5.869	-1.208
collegewhite	0.093	0.043	2.167	0.030	0.010	0.179
whitecollar	-0.059	0.032	-1.843	0.065	-0.124	0.002
nodiploma	0.145	0.065	2.226	0.026	0.021	0.277
highschoolgrad	0.060	0.042	1.431	0.152	-0.020	0.144
homeprice_med	0.000	0.000	2.722	0.006	0.000	0.000
ruralUrban	1.002	0.544	1.844	0.065	0.011	2.178
CountyDurham	1.073	1.191	0.901	0.368	-0.985	4.129
CountyOrange	-1.223	1.438	-0.850	0.395	-4.080	2.082
CountyWake	-0.623	1.198	-0.520	0.603	-2.712	2.437

```
## [1] 20.12013
```

```
## [1] 0.0001602844
```

Something interesting is going on. While the test statistic is very high (20.12) and the p-value is very low (.000), which encourages us to add County to our model, the county p-values are extremely high. We decide to check for multicollinearity to see if something is amiss in the model that might account for this anomaly.

In order to avoid collinearity of predictor 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).

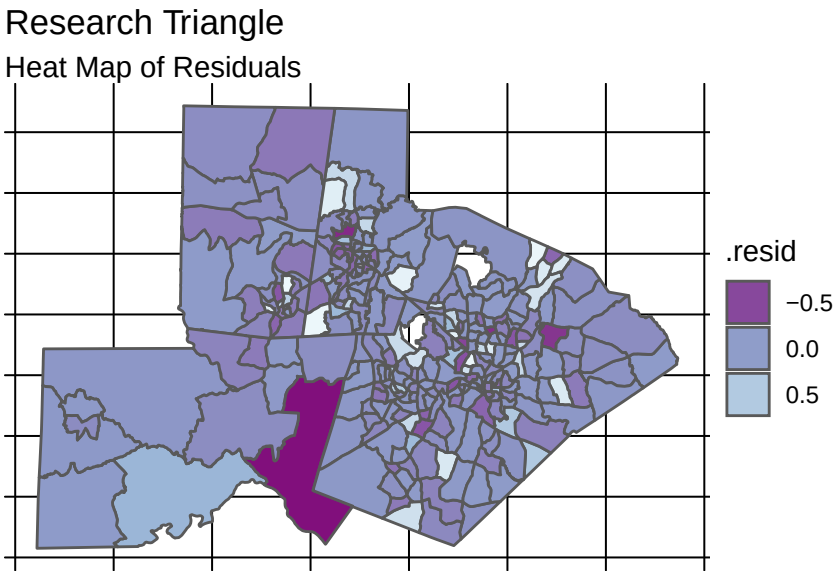
names	x
collegewhite	1.521
whitecollar	1.178
nodiploma	1.204
highschoolgrad	1.408
homeprice_med	1.310
ruralUrban	1.243
CountyDurham	9.935
CountyOrange	2.588
CountyWake	10.614

Multicollinearity appears to be a problem because both Wake and Durham counties have a vif of around 10. However, since these are categorical variables, multicollinearity does not quite make sense. Since the vast majority of tracts are in Durham and Wake counties, their standard errors are artifically inflated. This accounts for the unusally high p-values and vifs. We have seen that there are major differences in gentrified tracts between the counties, so we are going to continue with County included in our model.

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.

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:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-2.925	1.079	-2.710	0.007	-5.869	-1.208
collegewhite	0.093	0.043	2.167	0.030	0.010	0.179
whitecollar	-0.059	0.032	-1.843	0.065	-0.124	0.002
nodiploma	0.145	0.065	2.226	0.026	0.021	0.277
highschoolgrad	0.060	0.042	1.431	0.152	-0.020	0.144
homeprice_med	0.000	0.000	2.722	0.006	0.000	0.000
ruralUrban	1.002	0.544	1.844	0.065	0.011	2.178
CountyDurham	1.073	1.191	0.901	0.368	-0.985	4.129
CountyOrange	-1.223	1.438	-0.850	0.395	-4.080	2.082
CountyWake	-0.623	1.198	-0.520	0.603	-2.712	2.437

```
log(pi_hat/(1-pi_hat)) = -2.925 + (0.093)*(collegewhite) - (0.059)*(whitecollar) + (0.145)*(nodiploma)
+ (0.060)*(highschoolgrad) + (0.00001)*(homeprice_med) + (1.002)*(ruralUrban) + (1.073)*(CountyDurham)
- (1.223)*(CountyOrange) - (0.623)*(CountyWake)
```

The intercept, -2.925, 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 workers, those with no

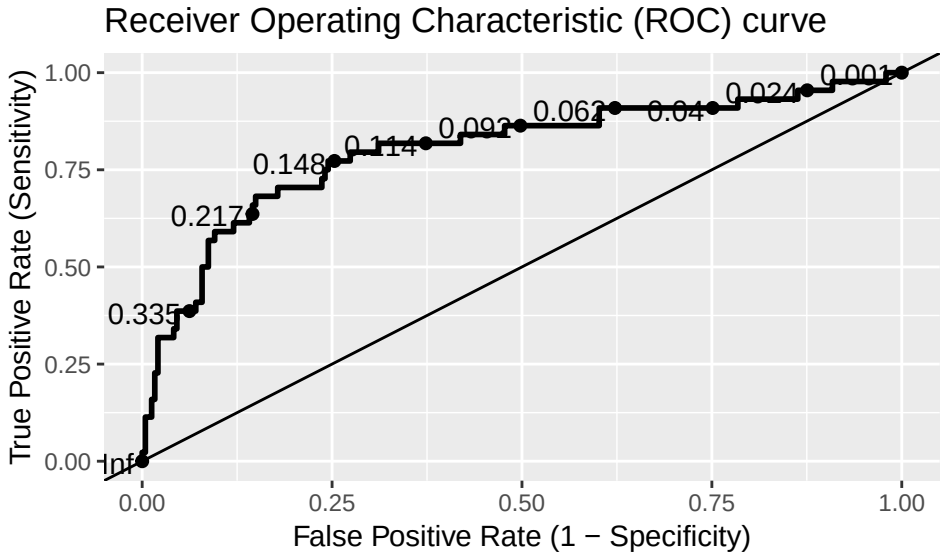
diploma, highschool graduates, and median home price is equal to zero and the census tract is rural in Chatham county.

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 change in college educated whites seems to have a reliably strong impact on gentrification: holding all other variables constant, with a percentage point change in college educated whites, the odds of gentrification are expected to multiply by a factor of exp(0.093).

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(1.002) 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.

Although the p-value for Durham county might be artifically high, it is noteworthy that the odds of gentrification for a Durham census tract are exp(1.073) that of a census tract in Chatham county.

In order to use our model to predict whether an area is gentrified or not, we create 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] 0.8016786
```

Since our AUC is 0.8016 we can see that the logistic model fits the data fairly well.

In order to pick an appropriate threshold for predicting gentrification, 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. The risk of false-positive classification is not very high compared to the risk of a false-negative outcome.

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. Given these concerns, we are picking .148 as the threshold to predict gentrification.

Section 4: Limitations

The biggest limitation of our study is the definition of gentrification. Many prior 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 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. 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 also pinpointed the difference in gentrification across the region's cities. Durham experiences the most gentrification by far, Raleigh experiences mild gentrification, and Chapel Hill is mostly immune to it. We zoomed in even further on Durham and discovered that gentrifiaction is most prevalent in East Durham, a historically Black part of town. Our analysis of Durham suggest that gentrification is more than just a change in demographics – it affects the social fabric of a city as well.

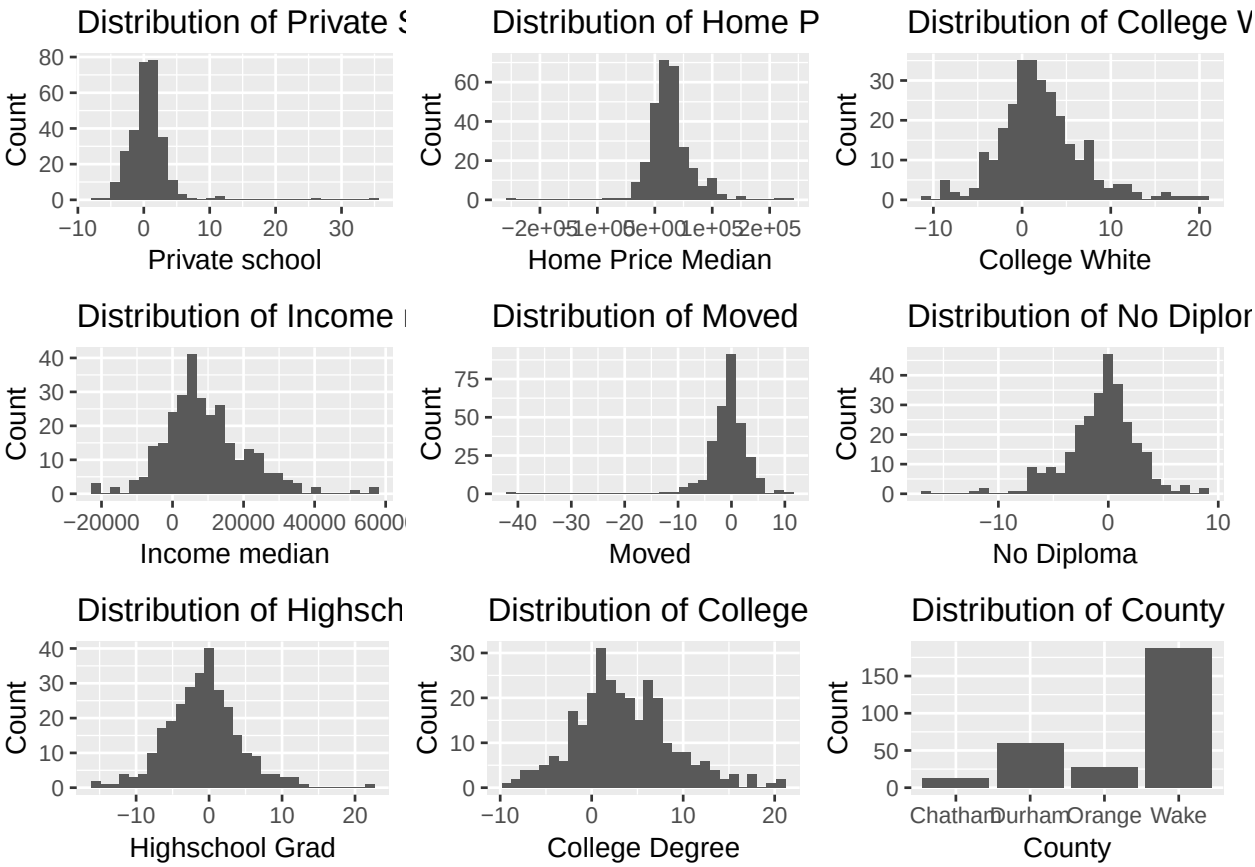
We created a logistic model that used urban character, county, median home price, and percentage changes in college-educated whites, white-collar jobs, those with no diploma, and those with just a high school education to predict gentrification. From our hypothesis—-that the change in college educated 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 urban character 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 (broken down 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. In addition, more comparative analysis could be done between high-growth cities to determine the root causes of gentrification. Government policy, demographic history, and type of growth could be analyzed. This would allow policymakers, non-governmental organizations, and developers to better understand and mitigate the impacts of gentrification on local communities.

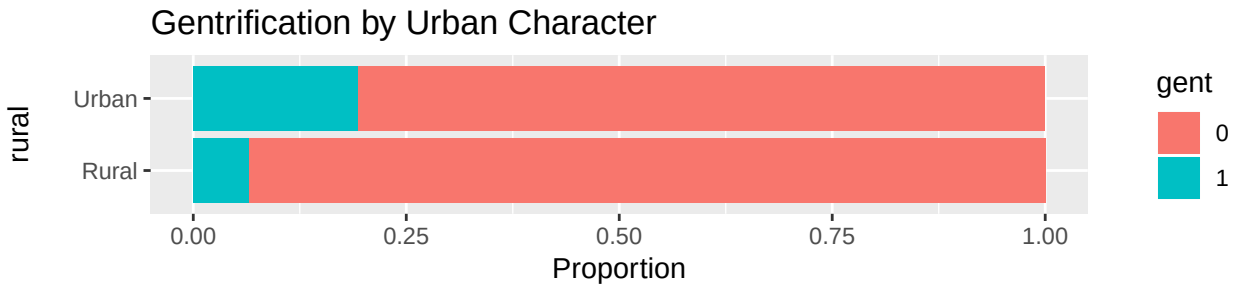
Section 6: Additional Work

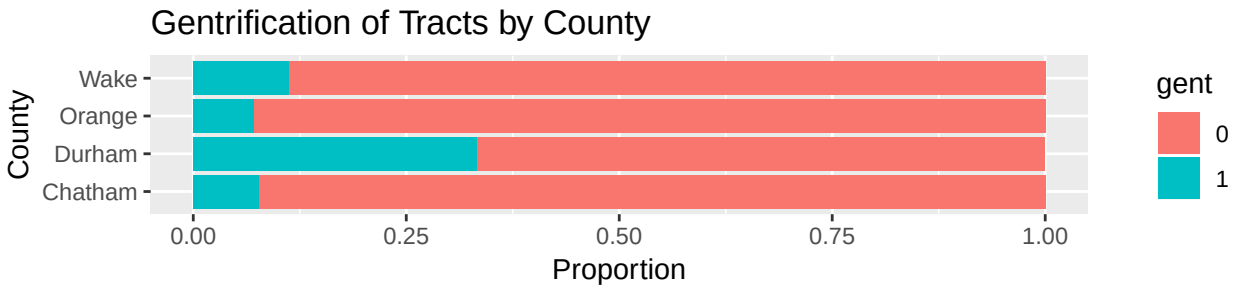
Univariate EDA

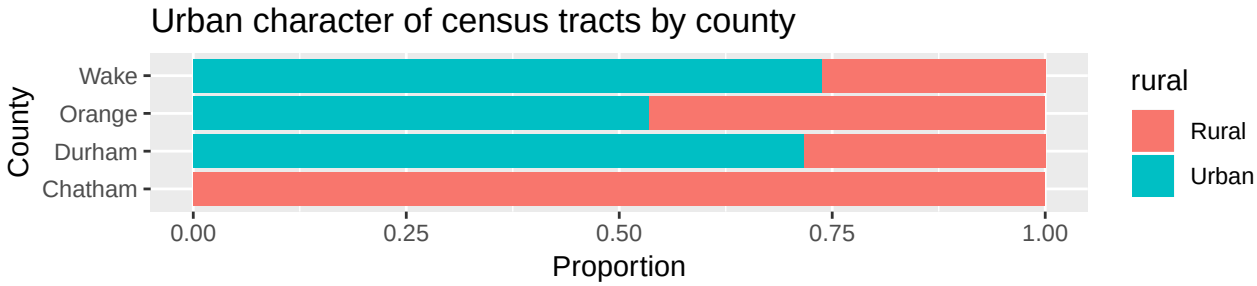


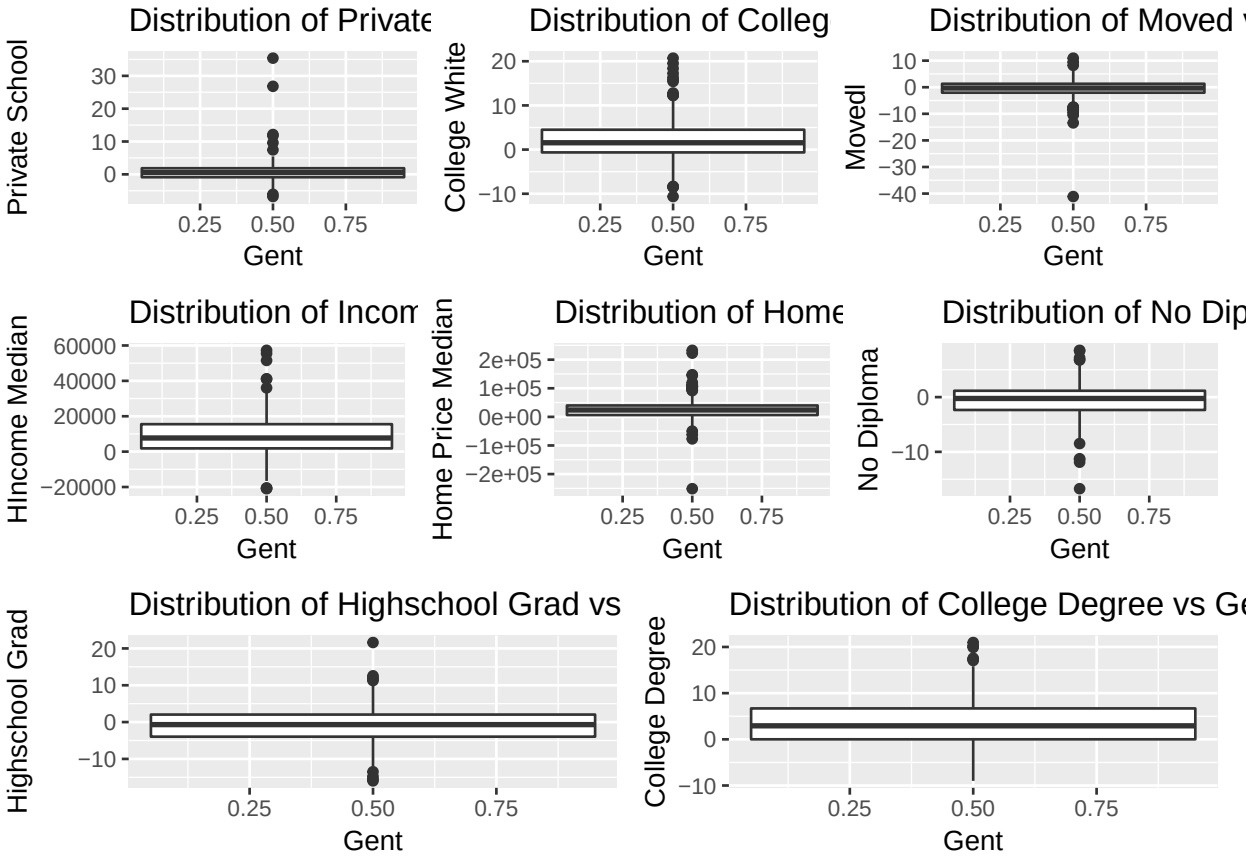
Each predictor variable is normally distributed around 0.

Bivariate EDA:


Gentrification by Urban Character
rural
Urban
Rural
0.00
0.25
0.50
0.75
1.00
Proportion
gent
0
1


Gentrification of Tracts by County
County
Wake
Orange
Durham
Chatham
0.00
0.25
0.50
0.75
1.00
Proportion
gent
0
1


Urban character of census tracts by county
County
Wake
Orange
Durham
Chatham
0.00
0.25
0.50
0.75
1.00
Proportion
rural
Rural
Urban



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

Considering interaction terms:

After determining Rural and County are significant predictor variables, we considered adding interaction terms with rural to our model. We used k-fold cross validation to determine if we should add the interactions between median home price and rural and that between college educated whites and rural. We start by looking at the 5-fold cross validation results for the model above, and then we run the same test with our new interaction terms.

Considering interaction between rural and county:

mean_train_mse3	mean_test_mse3
7.028	7.028

mean_train_mse4	mean_test_mse4
7.038	7.037

Model 1 (excluding interaction terms) testing error: 7.028 Model 2 (including interaction terms) testing error: 7.038

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 between county and rural character.

Then we did another K-fold cross validation considering interactions between median homeprice and rural,

and between college educated whites and rural:

mean_train_mse	mean_test_mse
5.368	5.37

mean_train_mse	mean_test_mse
5.671	5.672

Model 1 (excluding interaction terms) testing error: 5.369539 Model 2 (including interaction terms)testing error: 5.671572

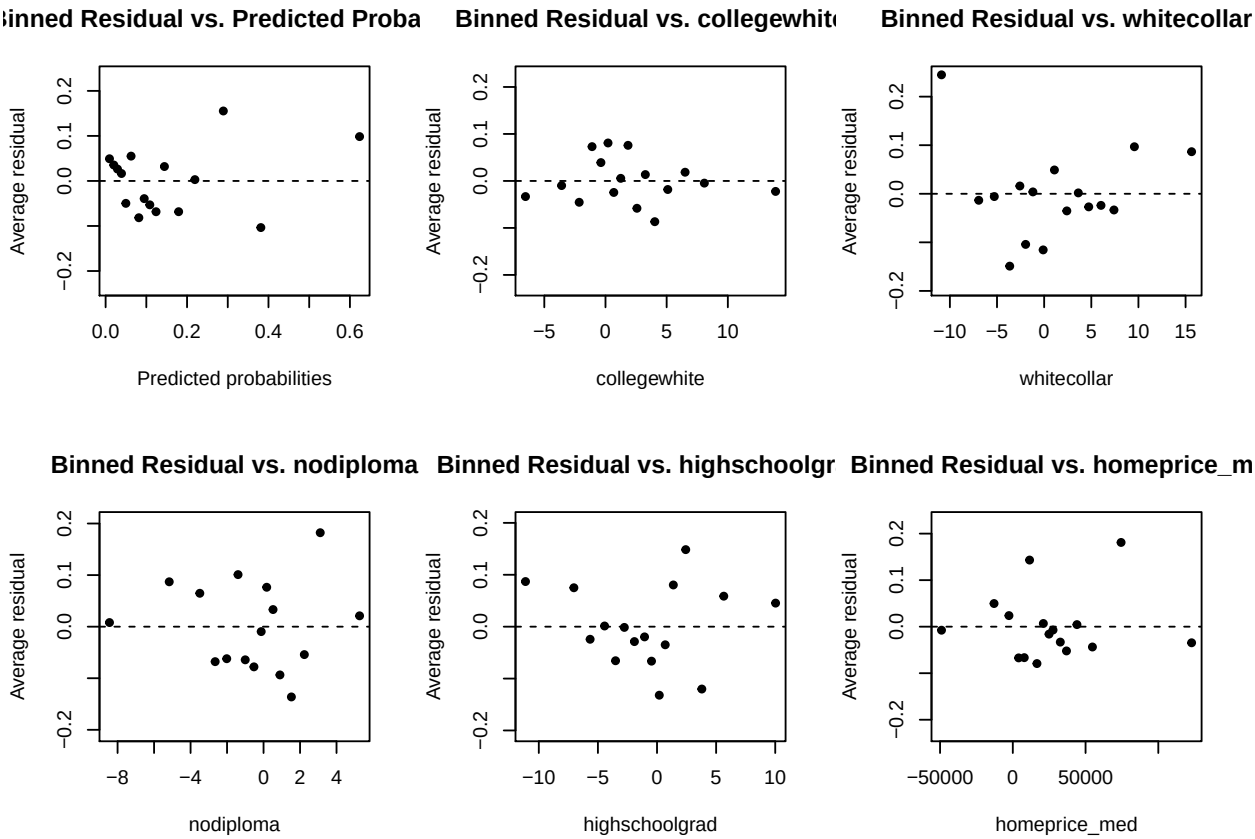
Although the testing errors are very close, Model 1 performs better than Model 2 when predicting if a census tract is gentrified. Therefore, we will continue with the model that does not include the interaction between median homeprice and rural character and between college educated whites and rural character.

###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 16
##    .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 10 more variables: homeprice_med <dbl>,
## #   rural <chr>, County <chr>, .fitted <dbl>, .se.fit <dbl>, .resid <dbl>,
## #   .hat <dbl>, .sigma <dbl>, .cooksd <dbl>, .std.resid <dbl>
```

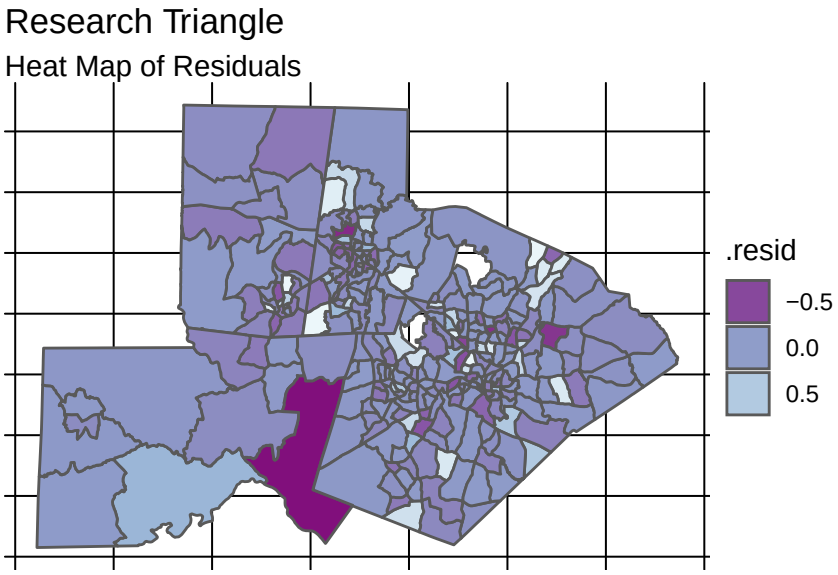



```
## # A tibble: 2 x 2
##   rural mean_resid
##   <chr>      <dbl>
## 1 Rural  -2.24e-15
## 2 Urban  -1.24e-15
```

```
## # A tibble: 4 x 2
##   County  mean_resid
##   <chr>        <dbl>
## 1 Chatham  -1.97e-16
## 2 Durham   -3.65e-16
## 3 Orange   -1.31e-14
## 4 Wake     -2.85e-16
```

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