Data Analysis

Code

ARC Capstone Team

4/19/2022

1. Basic look at our dataset

To take an overview to our data, we have two census-tract level datasets for 2019 and 2020, both with 1 dependent variables: eviction rate, calculate by dividing total eviction cases to total renting households, and 7 independent variables, poverty rate, education rate, uninsurance rate, minority rate, renter rate, unemployment rate, rent burden rate. Our goal is using regression to find out which of those are the top factors contributed to high eviction rate among 5 counties in Atlanta area. We have 622 observations for our 2019 data, and 318 observations for our 2020 data. Since our observations are based on 11 digits GEOID assigned by the Census Bureau and other state and federal agencies, the reason why we have more observations for 2019 than 2020 is that we are missing dependent variablels for some of our tract area. For example, we have GEOID 13063040408 (Clayton county) for 2019, but not for 2020. Since our goal is to find top factors for eviction rate, but not in a specific time line. We think it would be better to combine our 2019 and 2020 dataset for more obervations.

To take a deeper look into our data, we first calculate the average of all our variables. The first chart shows the combination of 2019 and 2020, the second shows 2019, and the third shows 2020.

Code

	Meanvalue
EvictionRate	0.098
PovertyRate	0.152
EducationRate	0.398
UninsurRate	0.141
MinorityRate	0.640
RenterRate	0.446
UnempRate	0.062
RentBurdenRate	0.488

Code

	Meanvalue
EvictionRate	0.117
PovertyRate	0.148
EducationRate	0.397
UninsurRate	0.141
MinorityRate	0.633
RenterRate	0.443
UnempRate	0.059
RentBurdenRate	0.494

Code

	Meanvalue
EvictionRate	0.063
PovertyRate	0.158
EducationRate	0.399

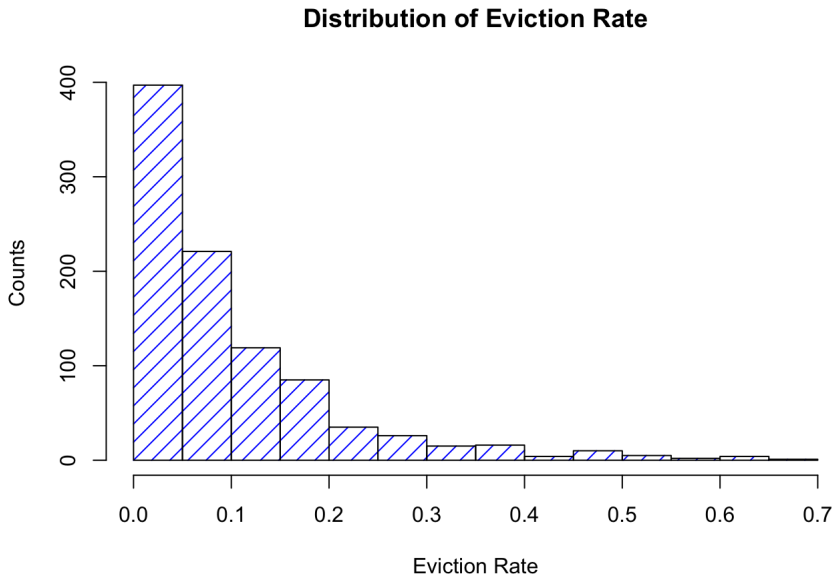
	Meanvalue
UninsurRate	0.141
MinorityRate	0.653
RenterRate	0.453
UnempRate	0.068
RentBurdenRate	0.478

2. Analyzing the 2019 and 2020 data together

1. Distribution Histogram

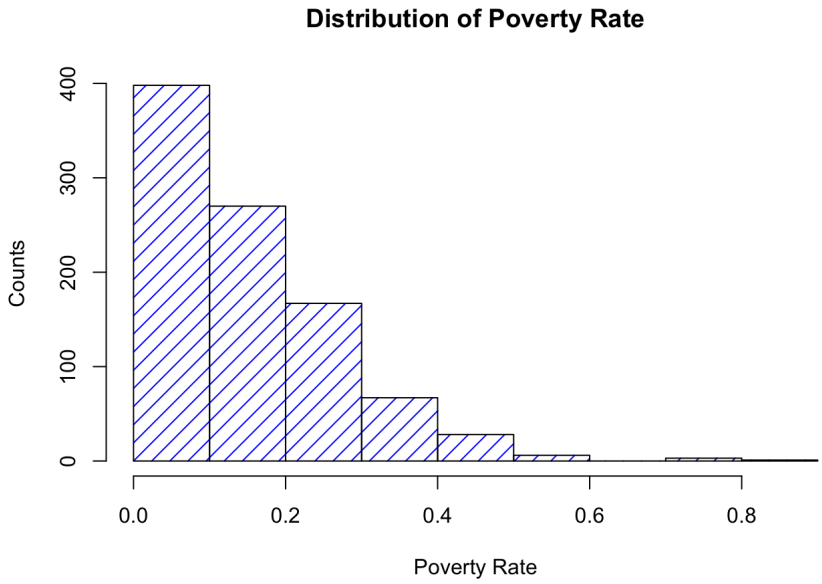
We also want to see the skewness of our data, and the best way to do so is to generate density plot for our dependent and independent variables.

Code

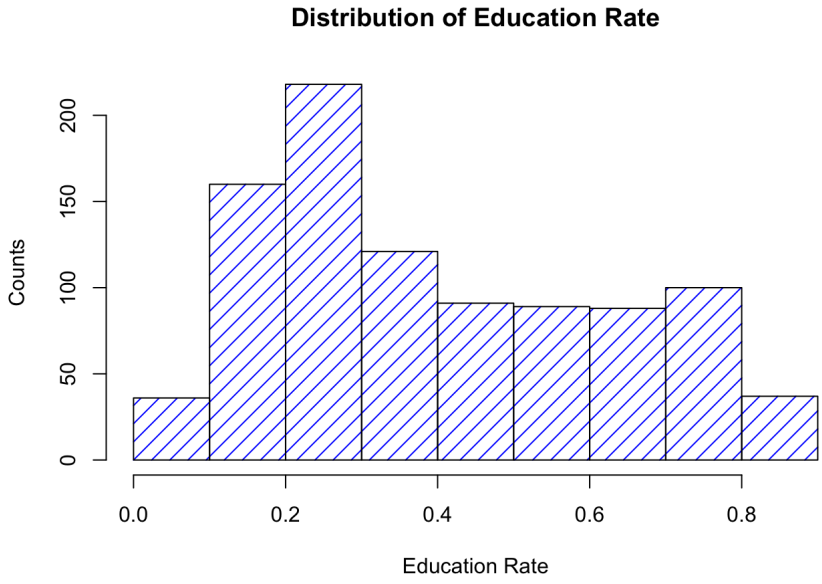


The histogram plot shows the overall distribution of our dependent variables. In this plot, x axis is a number line that has been split into number ranges.the y-axis represents the number count of occurrences in the data for each column and are used to visualize data distributions. From the graph we could see the majority of our observations are between 0 to 0.2, meaning the distribution of eviction rate is skewed to right, This skewness means when we're trying to run the mlr later in our regression part, we need to take log to our dependent variable to bring the skewed dependent variable to be more normal.

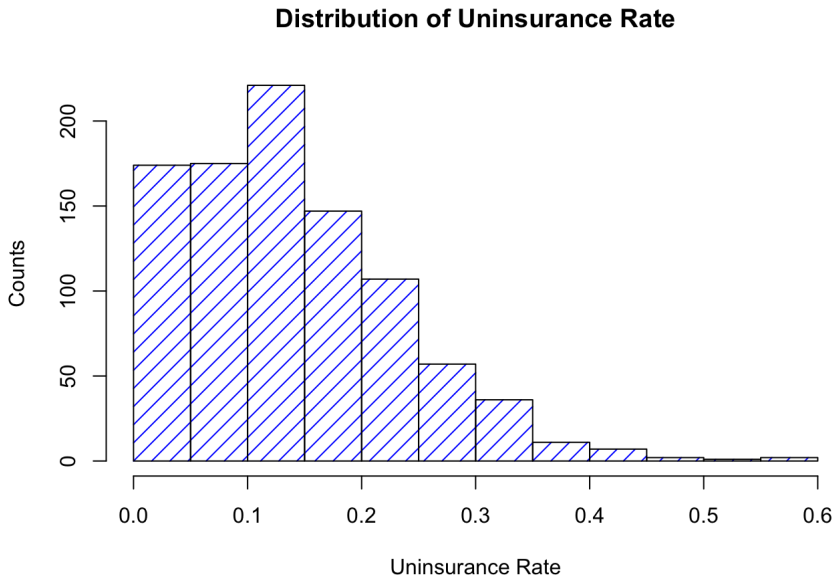
Code


Distribution of Poverty Rate
Counts
0
100
200
300
400
0.0
0.2
0.4
0.6
0.8
Poverty Rate

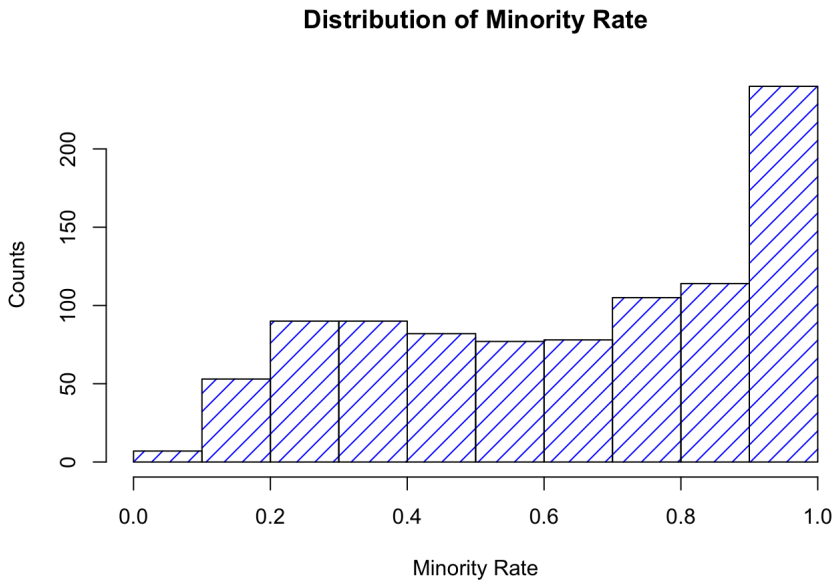
Code


Distribution of Education Rate
Counts
0
50
100
150
200
0.0
0.2
0.4
0.6
0.8
Education Rate

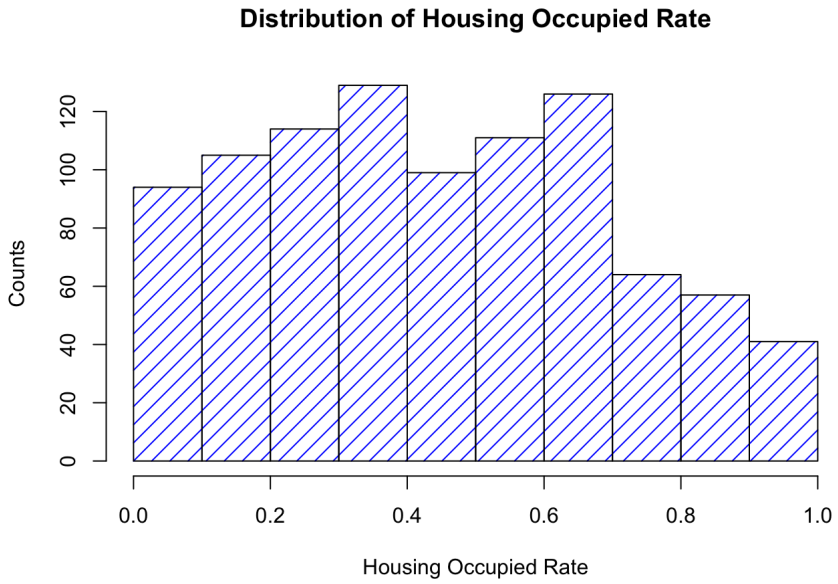
Code


Distribution of Uninsurance Rate
Counts
0
50
100
150
200
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Uninsurance Rate

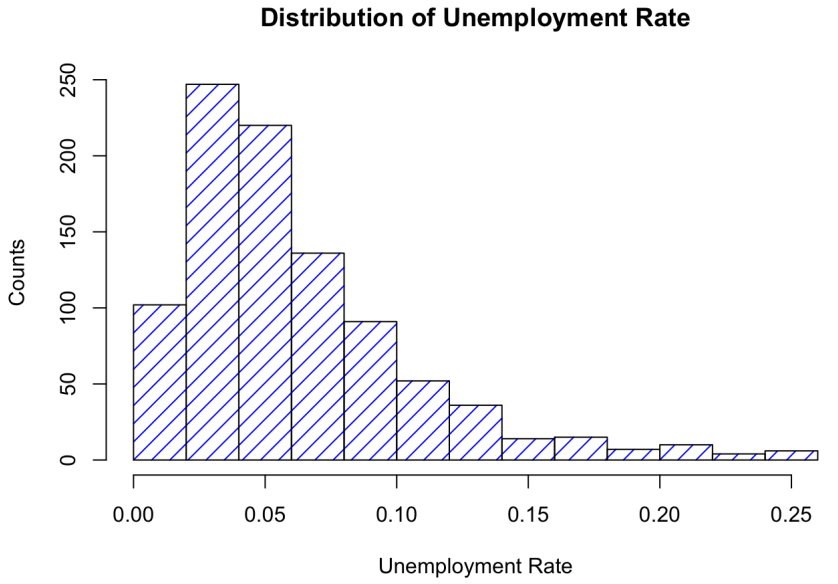
Code


Distribution of Minority Rate
Counts
0
50
100
150
200
0.0
0.2
0.4
0.6
0.8
1.0
Minority Rate

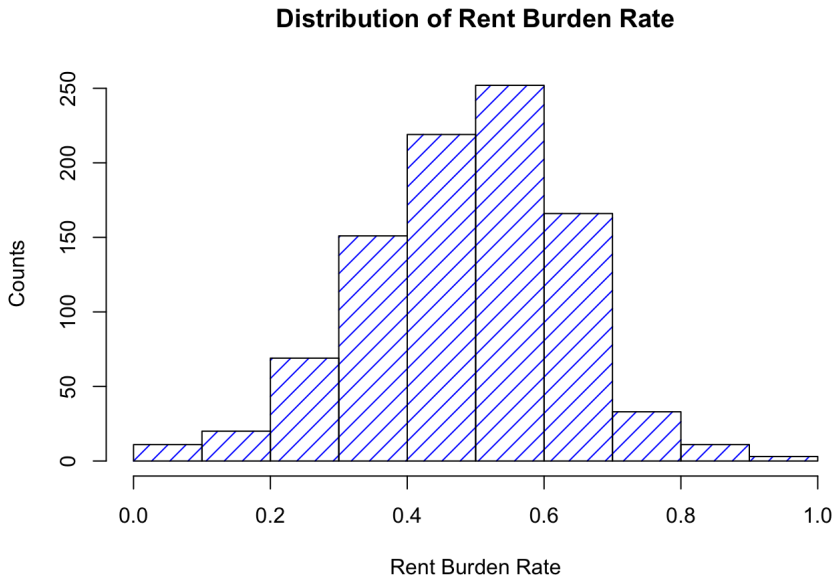
Code


Distribution of Housing Occupied Rate
Counts
0
20
40
60
80
100
120
0.0
0.2
0.4
0.6
0.8
1.0
Housing Occupied Rate

Code


Distribution of Unemployment Rate
Counts
0
50
100
150
200
250
0.00
0.05
0.10
0.15
0.20
0.25
Unemployment Rate

Code



From the graph, we could see the distribution plots for our seven x variables, we could look at the plot for poverty, the majority of our observations are between 0 to 0.3, meaning the distribution of eviction rate is skewed to right, and take plot for rent burden for example, the graph is approximately bell-shaped and symmetric about the mean, it's a perfect normal distribution.

2. Multiple Linear Regression

Since we've discussed that the distribution of eviction rate is skewed to right and we need to take log to our dependent variable to bring the skewed dependent variable to be more normal. We could generate basic multiple linear regression equation:

Model 1

$$log(EvictionRate) = \beta_0 + \beta_1 * Poverty + \beta_2 * Edutacion + \beta_3 * RentBurden + \beta_4 * Uninsurance + \beta_5 * Minority + \beta_6 * Renter + \beta_7 * Unemplo$$

Code

	Dependent variable:
	log(Eviction rate)
Poverty	-2.506***
	(0.390)
Education	-1.077***
	(0.265)
Rent Burden	0.291
	(0.206)
Uninsurance	-0.169
	(0.418)
Minority	2.019***
	(0.199)
Housing Occupied	3.261***
	(0.153)
Unemployment	-1.126
	(0.787)
Constant	-5.024***
	(0.249)
Observations	931
R^2	0.688
Adjusted R^2	0.686
Residual Std. Error	0.794 (df = 923)
F Statistic	291.062*** (df = 7; 923)
Note:	*p<0.1;* **p<0.05;** p<0.01

Code

	x
PovertyRate	3.027796
EducationRate	5.033308
RentBurdenRate	1.378443
UninsurRate	2.200675

	x
MinorityRate	4.414054
RenterRate	2.220299
UnempRate	1.781776

From the chart we could see the results of our linear regression model, we first look at the R sqr at the bottom of the chart, in multiple linear regression, the R2 represents the correlation coefficient between the observed values of the outcome variable (y) and the fitted (i.e., predicted) values of y. Since the R^2 will increase as more x is included in the model, we introduce "Adjusted R Square", which is the adjustment value in the summary output is a correction for the number of x variables included in the prediction model. In our model, the adjusted R^2 is around 69, which means the model explains 69% of the variance in the outcome variable, eviction rate. We then look back at the numbers on the top, those numbers outside the parentheses indicate the estimate of regression beta coefficients in our model, the star sign on the right of the number indicates the significance level, and the 3 stars indicates a factor to be the most significant. The number below inside the parentheses indicates the standard error of regression, representing the average distance that the observed values fall from the regression line. Looking back at regression beta coefficients, a positive coefficient means that this predictor variable will positively affect the eviction rate, while a negative coefficient means that this predictor variable will negatively affect the eviction rate, in other words, when it increases, the eviction rate will drop. From the chart, we could see poverty rate, education rate, uninsurance rate, and unemployment rate are negatively correlated, while housing occupied, minority rate and rent burden rate are positively correlated, we could also generate our models to be:

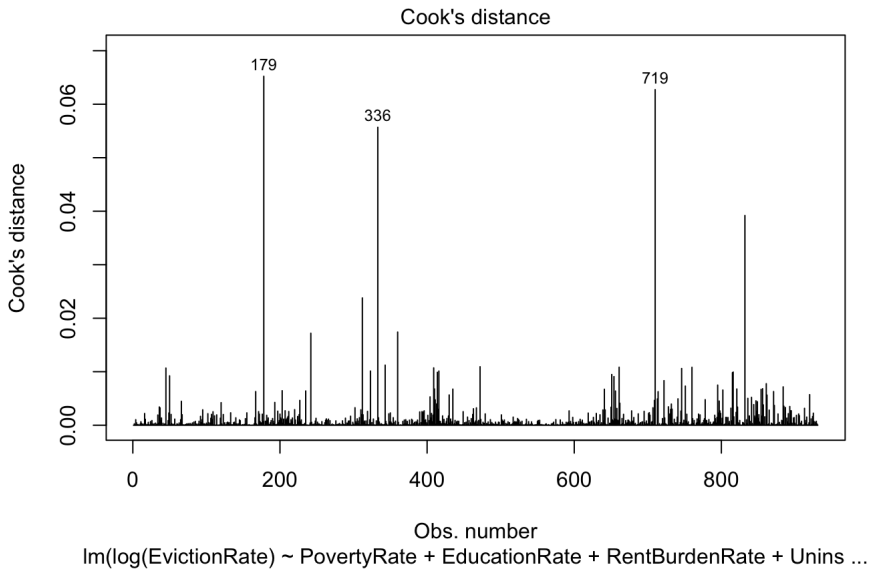
Model

$$log(Evictionrate) = -2.5060 * Poverty - 1.0766 * Edutacion + 0.2907 * RentBurden + -0.1695 * Uninsurance + 2.0186 * Minority + 3.2613$$

From the equation and the significance level in the chart, we could see poverty, education, minority and housing occupied are the most significant factors.

3. Finding Outliers

Code



This graph was plotted to indicate the outliers, In this graphs, x axis is the observation numbers of our datasets, and the y axis is Cook's Distance, which is an estimate of the influence of a data point. We defined our observations with top 3 cook's distance as outliers. From these two graphs, we could see For year 2019, outliers are oberservation number 179, 336, and 719.

Code

	Dependent variable:	
	log(Eviction rate)	
	(1)	(2)
Poverty	-2.506***	-2.462***
	(0.390)	(0.386)
Education	-1.077***	-1.055***
	(0.265)	(0.261)
Rent Burden	0.291	0.344*
	(0.206)	(0.203)
Uninsurance	-0.169	0.142
	(0.418)	(0.424)

Minority	2.019***	2.020***
	(0.199)	(0.196)
Housing Occupied	3.261***	3.258***
	(0.153)	(0.151)
Unemployment	-1.126	-1.684**
	(0.787)	(0.780)
Constant	-5.024***	-5.070***
	(0.249)	(0.245)
Observations	931	928
R^2	0.688	0.700
Adjusted R^2	0.686	0.697
Residual Std. Error	0.794 (df = 923)	0.780 (df = 920)
F Statistic	291.062*** (df = 7; 923)	306.018*** (df = 7; 920)
Note:		*p<0.1; **p<0.05; ***p<0.01

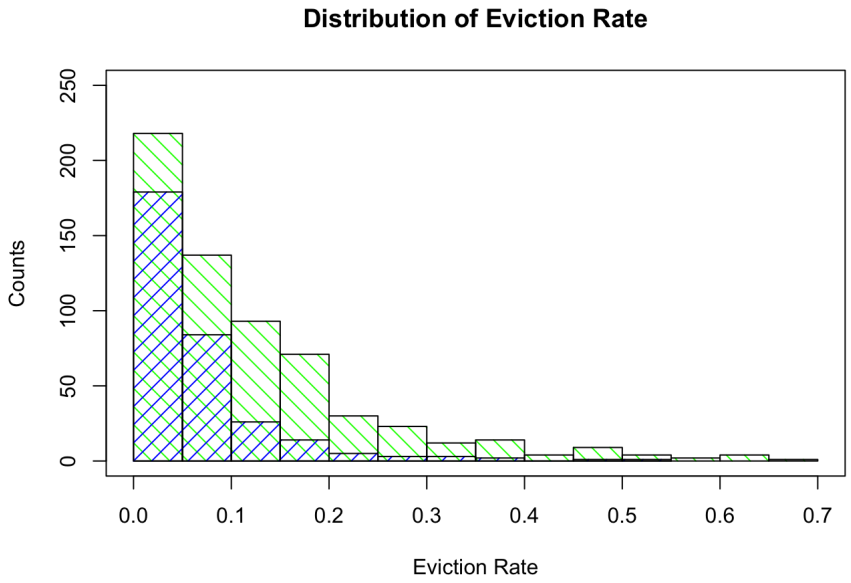
We want to test the model after removing the outliers, so we reran the model without outliesr, we could see after removing the outliers, the adjusted R^2 from 0.6859 to 0.6973; and adjusted R^2 for 2020 incresed from 0.734 to 0.753. The estimate of regression beta coefficients and standard error also changed a little bit, the significance level of unemployment rate increases.

3. Analyzing the 2019 and 2020 data seperately to see COVID impact

Since we considered year 2019 as before COVID, and 2020 after COVID. We want to see how COVID-19 impacts the eviction rate, how the weights of our top factors vary. In order to do so, we separate our dataset into year 2019, and 2020, run the mlr again to see the difference.

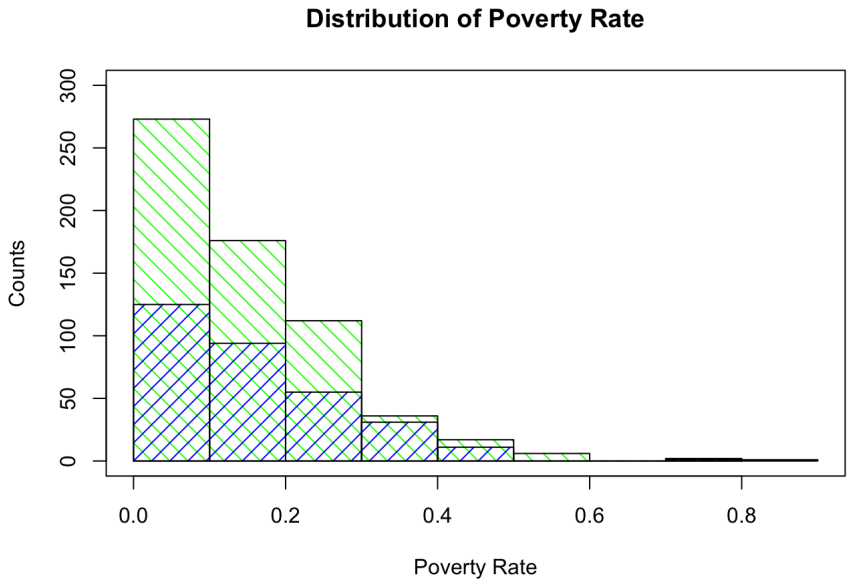
1. Distribution Histogram

Code

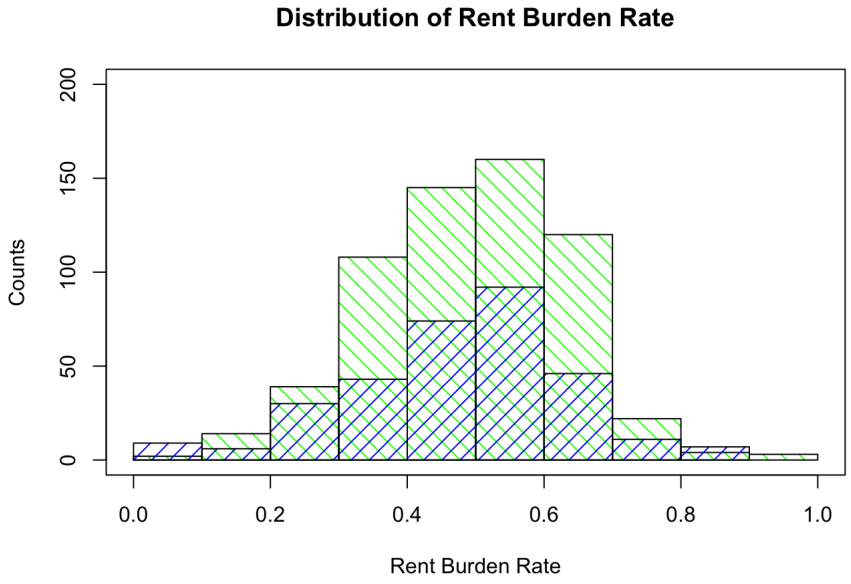


We made a distribution plot to see the skewness of eviction rate for year 2019 and 2020 differently, in this graph, the green region indicates 2019 eviction rate, and blue indicates 2020. Since we have more observations in 2019 than 2020, we could see the columns for 2019 are higher, however, the distribution are similar, we could see for both 2019 and 2020, majority of our observations are between 0 to 0.2, meaning the distribution of eviction rate is skewed to right, This skewness means when we're trying to run the mlr later in our regression part, we need to take log to our dependent variable to bring the skewed dependent variable to be more normal.

Code



Code



We also take poverty rate and rent burden rate as an example to see the distribution of our independent variables, we could see the distribution of 2019 and 2020 for these two variables are also similar, while poverty rate is skewed to right, and rent burden rate is normally distributted

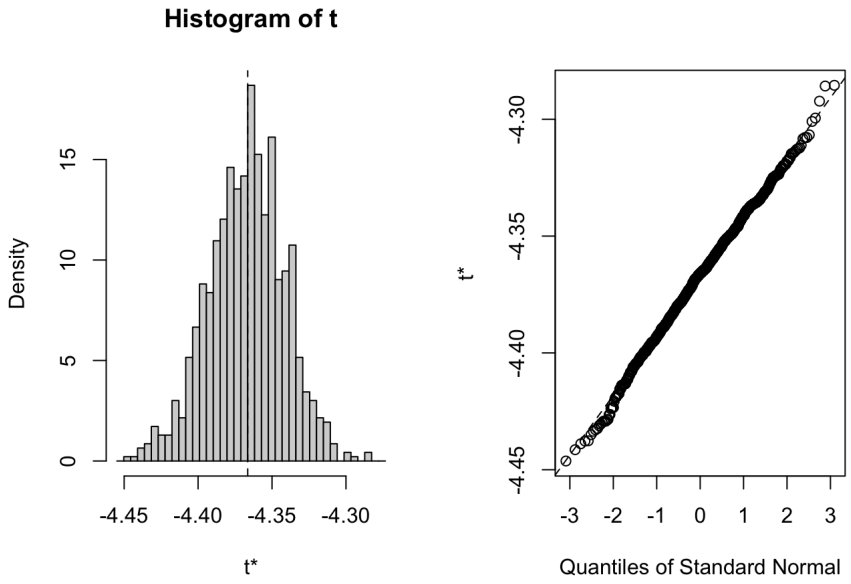
2. Bootstrapping to examine change in weekly evictions during pandemic

We used bootstrapping method, which replicates sampling with replacements, to examine if trend in eviction rate during the pandemic is truly different from 2019. Sample mean of difference in weekly evictions between 2019 and 2020 was computed with 10000 bootstrap replicates. Observation after March was only used because 1) The pandemic begun on April and 2) to control for seasonal trend in evictions.

Code

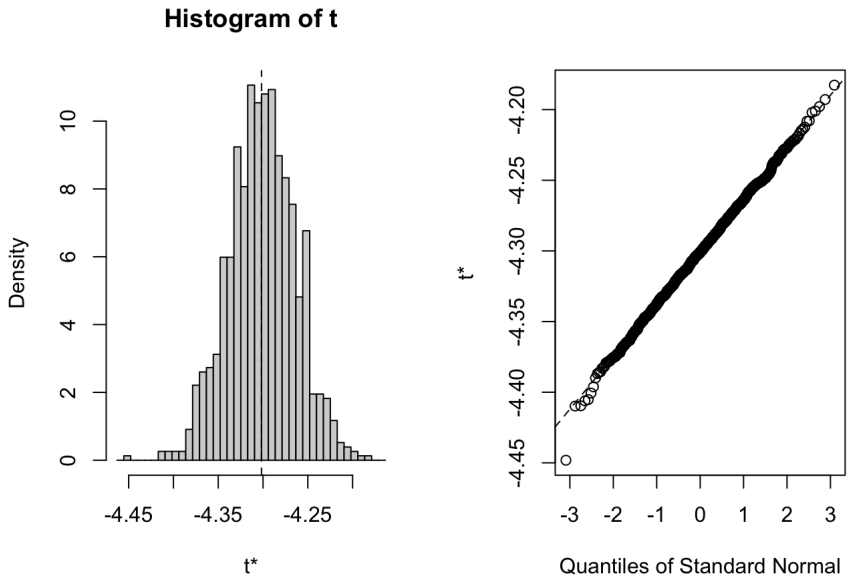
```
## `summarise()` has grouped output by 'TractID', 'Year', 'Month'. You can
## override using the `.groups` argument.
```

Code



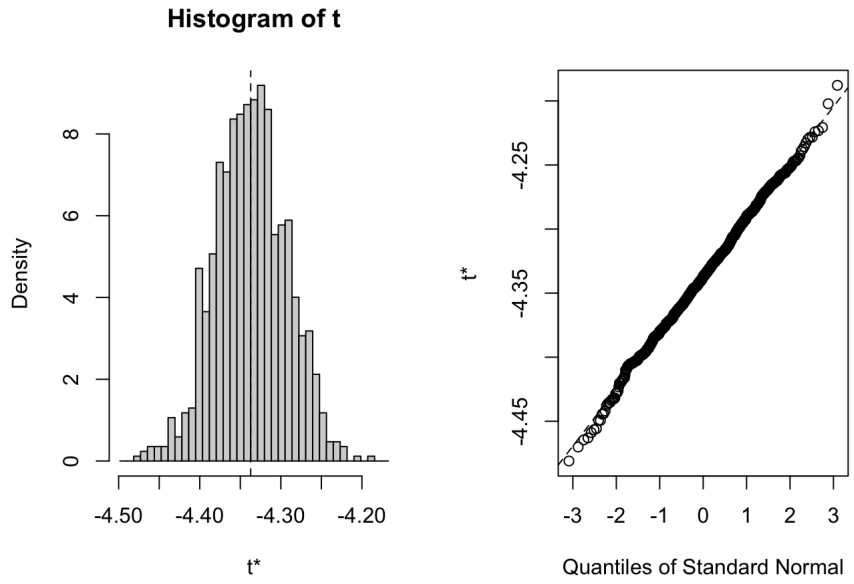
This time, we bootstrapped using data only during CARES ACT for 2020, and data between March and July for 2019

Code



Then, we bootstrapped using data after Augusts to rule out the effect of CARES Act

Code



Distribution of sample means suggest that the weekly eviction filings reported was lower in 2020 than 2019 (CI: 95%).

3. Multiple linear regression

We took log to our dependent variable to bring the skewed dependent variable to be more normal. We could generate two basic multiple linear regression equations:

Model 1

$$log(Eviction2019) = \beta_0 + \beta_1 * Poverty + \beta_2 * Edutacion + \beta_3 * RentBurden + \beta_4 * Uninsurance + \beta_5 * Minority + \beta_6 * Renter + \beta_7 * Unemple$$

Model 2

$$log(Eviction2020) = \beta_0 + \beta_1 * Poverty + \beta_2 * Edutacion + \beta_3 * RentBurden + \beta_4 * Uninsurance + \beta_5 * Minority + \beta_6 * Renter + \beta_7 * Unemple$$

Code

	Dependent variable:	
	log(Eviction rate)	
	(1)	(2)
Poverty	-2.510***	-2.356***
	(0.456)	(0.554)
Education	-1.224***	-0.219
	(0.290)	(0.418)
Rent Burden	-0.201	0.429
	(0.230)	(0.311)
Uninsurance	-0.587	0.847
	(0.448)	(0.676)
Minority	1.936***	2.545***
	(0.207)	(0.345)
Housing Occupied	3.378***	3.065***
	(0.170)	(0.237)
Unemployment	0.555	-0.822
	(0.994)	(1.028)
Constant	-4.506***	-6.368***
	(0.269)	(0.400)
Observations	613	318
R^2	0.740	0.740
Adjusted R^2	0.737	0.734
Residual Std. Error	0.686 (df = 605)	0.747 (df = 310)
F Statistic	246.250*** (df = 7; 605)	126.072*** (df = 7; 310)
Note:		*p<0.1; **p<0.05; ***p<0.01

From the chart we could see the results of two multiple linear regression model, the column on the left shows the result of 2019 and the column on the right shows the result of 2020. We first look at the R sqr at the bottom of the chart, the adjusted R^2 for both 2019 and 2020 are around 73, which means the model explains 73% of the variance in the outcome variable, eviction rate. We want to find some difference between 2019 and 2020, we could see in year 2019, housing occupied, unemployment are positively correlated, the rest of predictor variables are negatively correlated, while in year 2020, minority and housing occupied are positively correlated, the rest of predictor variables are negatively correlated. We could also generate our models to be:

Model 1

$$log(Eviction2019) = -2.510 * Poverty + -1.224 * Edutacion + -0.201 * RentBurden + -0.587 * Uninsurance + -1.936 * Minority + 3.378 *$$

Model 2

$$log(Eviction2020) = -2.356 * Poverty + -0.219 * Edutacion + 0.429 * RentBurden + 0.847 * Uninsurance + 2.525 * Minority + 3.065 * Rent$$

From the equation and the significance level in the chart, we could see education is significant in year 2019 but not in 2020.

4. Conclusion

Based on our regression analysis and connecting our data with what happened in real life, our team believe that poverty rate, education rate, race (minority rate), renter occupied unit rate, and rent burden are the top five factors most highly associated with evictions on a census tract level in the Atlanta region. Our team chose these five explanatory variables because poverty rate, education rate, minority race, and renter occupied unit rate are the most significant variables based on the regression analysis we conducted on both the combined (2019 & 2020) data sets and also the separate ones. We chose rent burden as the fifth one because it is most related to eviction among the three variables left.