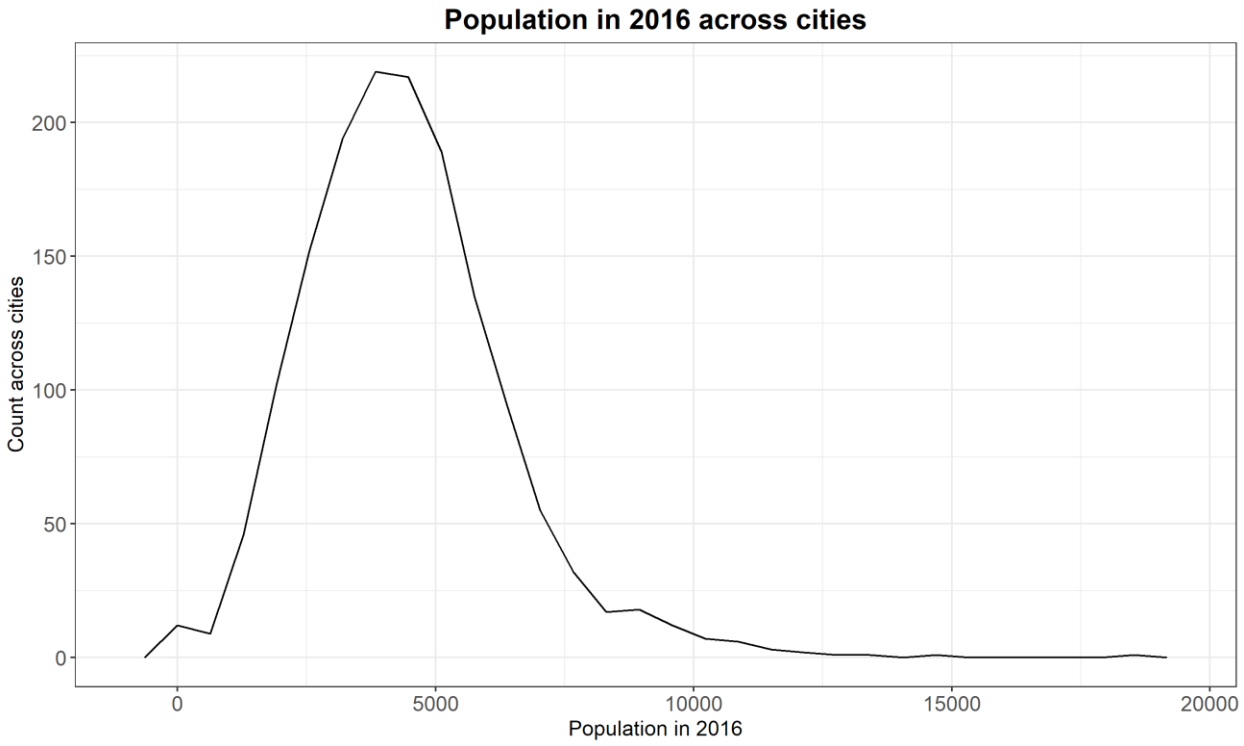
Adam Roen

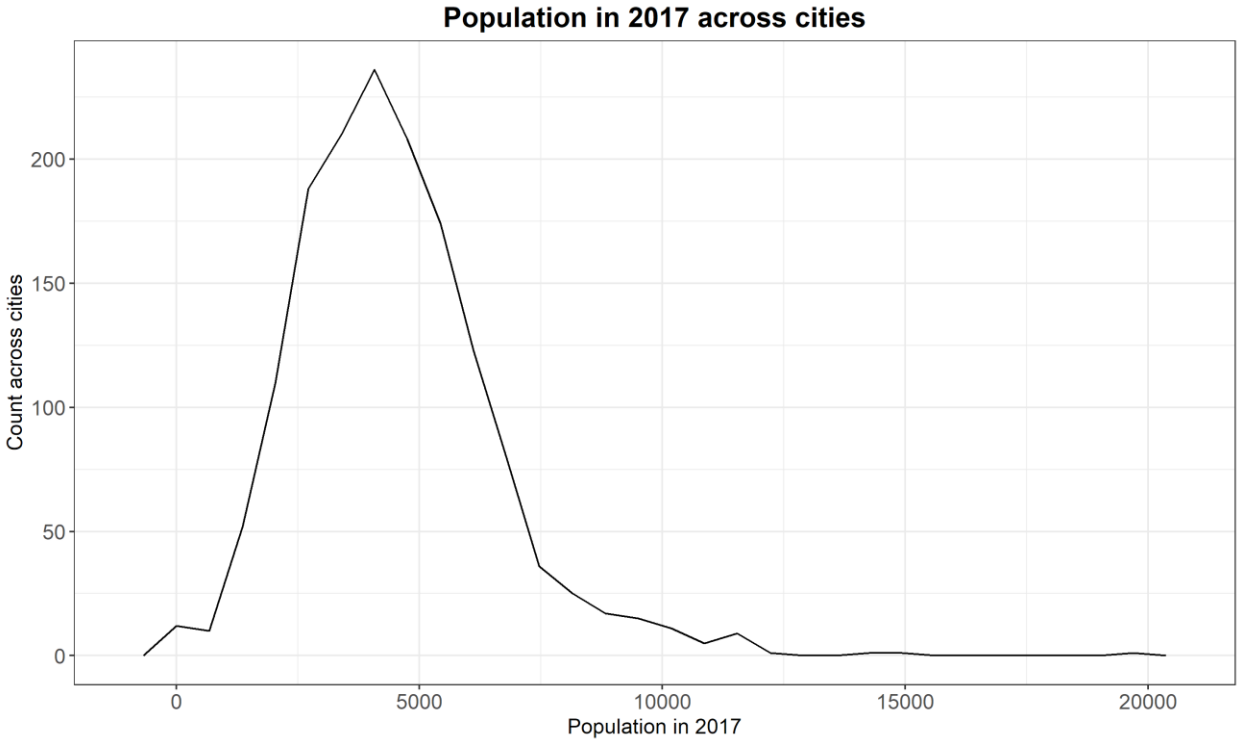
GIS 470

Final Project

To start off with on the project, I wanted to get some quick graphs done to show what the population throughout the state looked like in 2016 and 2017, the change in the population from 2016 to 2017 as well, and I wanted to graph the independent variables that I chose to study. For the amenities category I chose to look at air quality, and greenness. I considered these amenities because they both, in a sense, do not affect the wallets of those living there. For the economic category, I chose to look at income inequality and college degrees. Income inequality I chose because that will determine where some communities have more money and where some have less, etc. And college degrees because throughout the semester we looked at how areas with higher college graduates have a higher income.

To start, here are the graphs for population of 2016 and 2017:





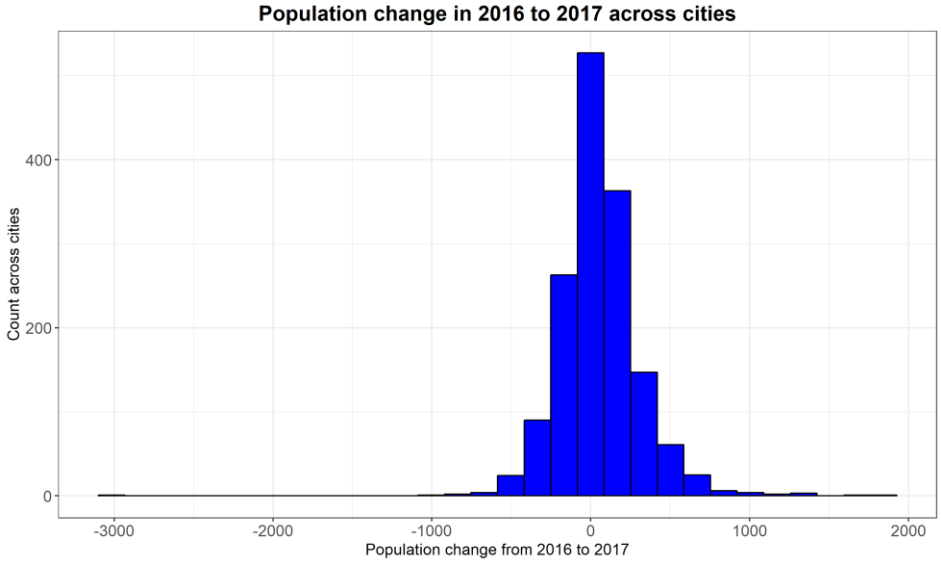
As we can see, in 2017 there are several cities that have gotten closer to the 20,000 mark and from these summary outputs in RStudio, the mean, median and max have all increased from 2016 to 2017

```
> summary(health_data$Pop_2016)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    3060    4227    4410    5446   18531
```

```
> summary(health_data$Pop_2017)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    3078    4267    4463    5507   19689
```

I have also created histograms of the 4 variables I chose for my independent variables just to see if there was any sort of trend with the data, and to visualize the data I am working with.

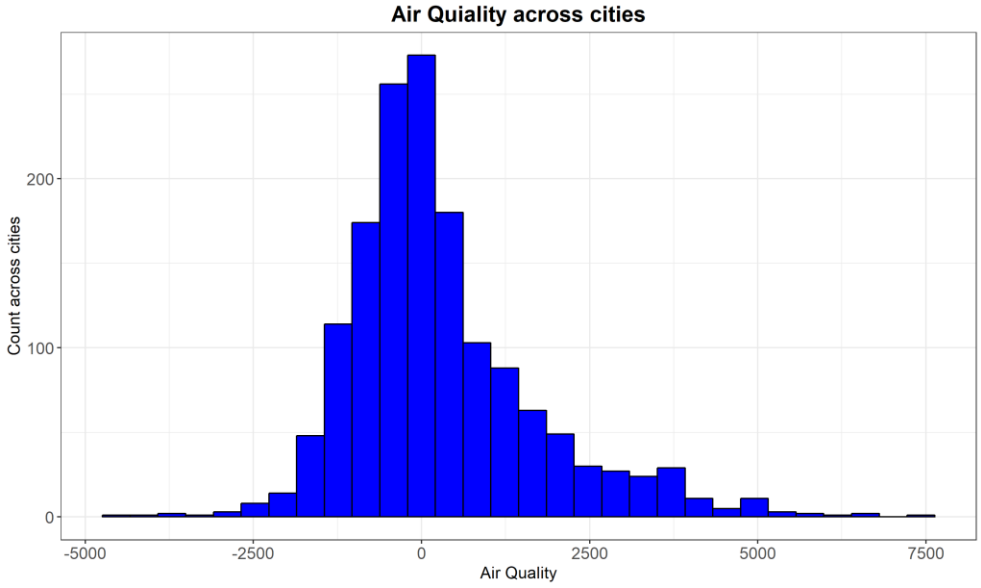
I also created a new variable in the Healthy Tracts data as the difference between the 2016 population and the 2017 population. I then went ahead and plotted that as well in the form of histogram.

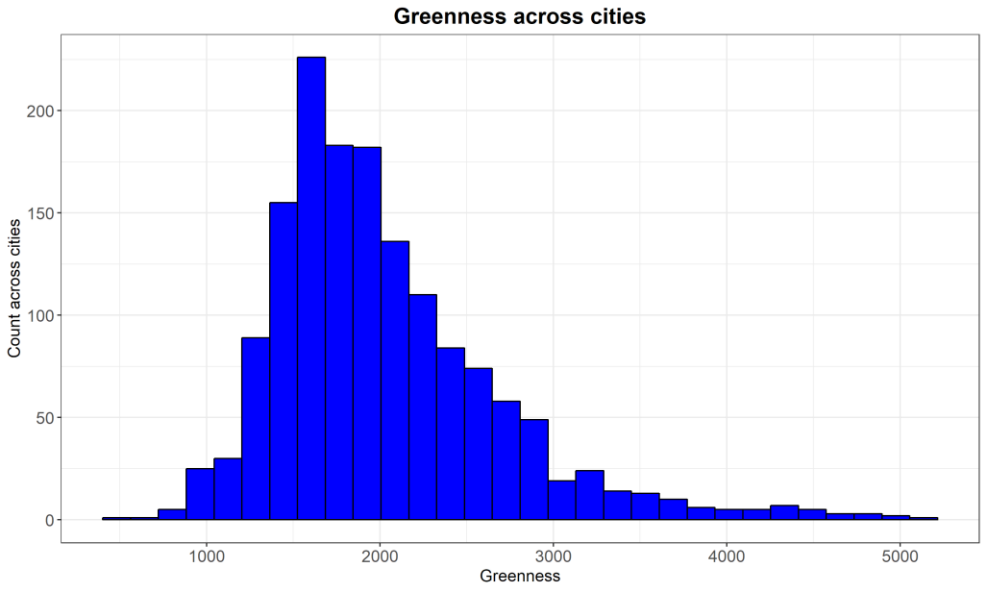


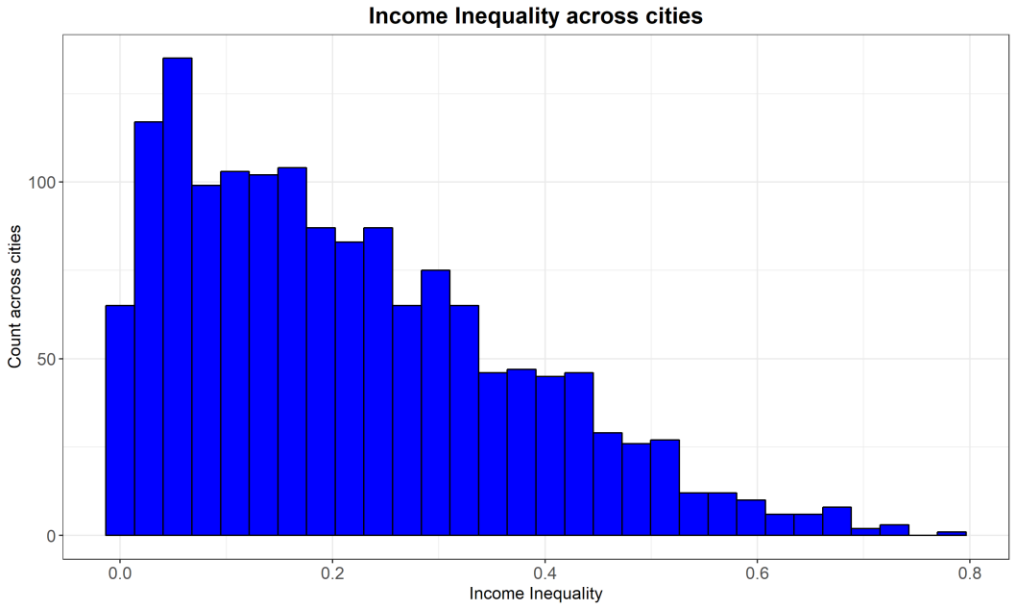
Here is the summary output for that as well. The mean change between all the cities is 53.29 or 54 people. The largest change was 1779 people.

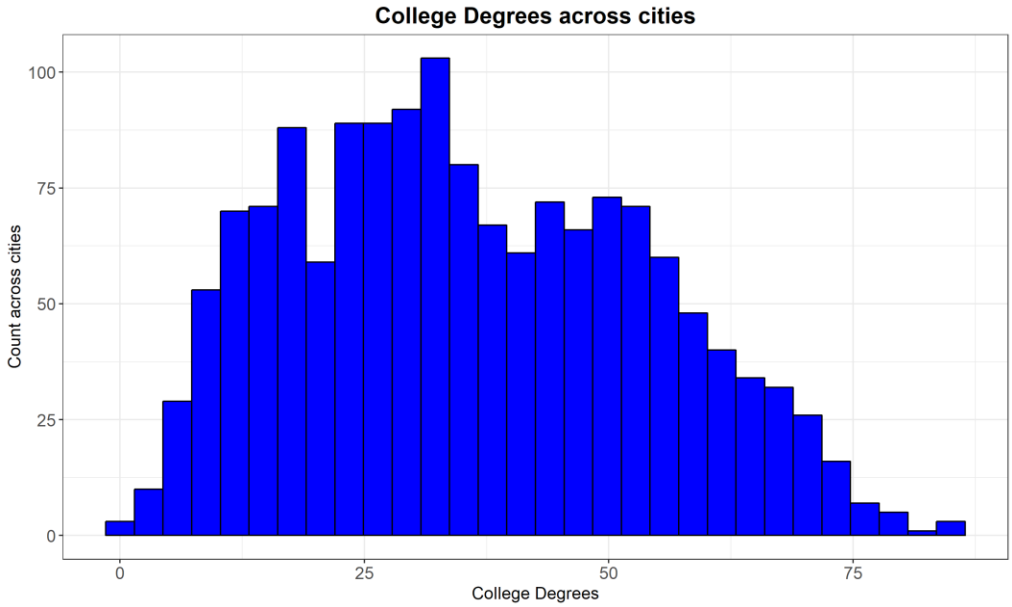
```
> summary(health_data$popDiff)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3081.00   -85.00    34.00    53.29   168.00  1779.00
```

I also created histograms of the four independent variables that I am going to be working with, that I mentioned at the beginning of my report. I just wanted to visualize my data.




Greenness across cities
Count across cities
200
150
100
50
0
1000
2000
3000
4000
5000
Greenness


Income Inequality across cities
Count across cities
100
50
0
0.0
0.2
0.4
0.6
0.8
Income Inequality



Next, I created my models for the multiple regression that I was going to perform. Per the instructions for the assignment, I used my popDiff variable (difference in population from 2016 to 2017) as the dependent variable. I then created 6 different models, using the four categories I mentioned at the beginning of my report as the independent variables. I made sure that each of them interacted with each other, even if it was an amenity interacting with another amenity and vise versa. I also made sure to scale each variable within the model, because there is obviously a difference in scale between air quality and income inequality, for example.

The first model I made was for air quality, and greenness which are the two amenities.

```
> a1 <- lm(popDiff ~ scale(AirQuality_v) + scale(Greenness_v), data = health_data)
> summary(a1)

Call:
lm(formula = popDiff ~ scale(AirQuality_v) + scale(Greenness_v),
    data = health_data)

Residuals:
     Min       1Q   Median       3Q      Max
-3131.78  -140.66   -16.45   112.93  1671.53

Coefficients:
                   Estimate Std. Error t value            Pr(>|t|)
(Intercept)         53.3251     6.7107   7.946 0.00000000000000371 ***
scale(AirQuality_v) -0.3957     6.8306  -0.058               0.9538
scale(Greenness_v)  13.9406     6.8288   2.041               0.0414 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 262 on 1521 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.002861,  Adjusted R-squared:  0.001549
F-statistic: 2.182 on 2 and 1521 DF,  p-value: 0.1132
```

From the output in R, I can see that the multiple R^2 result is 0.002861, meaning that 0.2861% of the population change can be explained by air quality and greenness, which is not meaningful at all.

Next up is air quality, and income inequality. So an amenity, and an economic factor.

```
> a2 <- lm(popDiff ~ scale(AirQuality_v) + scale(IncomeInequality_v), data = health_data)
> summary(a2)

Call:
lm(formula = popDiff ~ scale(AirQuality_v) + scale(IncomeInequality_v),
    data = health_data)

Residuals:
     Min       1Q   Median       3Q      Max
-3126.04  -142.06   -17.02   113.77  1726.18

Coefficients:
                          Estimate Std. Error t value            Pr(>|t|)
(Intercept)                 53.563      6.750   7.935 0.00000000000000407 ***
scale(AirQuality_v)         -1.510      6.765  -0.223               0.8234
scale(IncomeInequality_v)   15.717      6.787   2.316               0.0207 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 262.6 on 1510 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.003677,  Adjusted R-squared:  0.002357
F-statistic: 2.786 on 2 and 1510 DF,  p-value: 0.06197
```

From the output in R, I can see that the multiple R^2 result is 0.003677, meaning that 0.3677% of the population change can be explained by air quality and income inequality, which even though this is higher than air quality and greenness, it is not meaningful at all.

The next model is air quality and college degrees. This is an amenity with an economic factor again.

```
> a3 <- lm(popDiff ~ scale(AirQuality_v) + scale(CollegeDegree_v), data = health_data)
> summary(a3)

Call:
lm(formula = popDiff ~ scale(AirQuality_v) + scale(CollegeDegree_v),
    data = health_data)

Residuals:
     Min       1Q   Median       3Q      Max
-3112.49  -138.69   -17.11   114.46  1721.84

Coefficients:
                          Estimate Std. Error t value            Pr(>|t|)
(Intercept)                 53.524      6.739   7.942 0.00000000000000383 ***
scale(AirQuality_v)          9.211      9.656   0.954               0.3403
scale(CollegeDegree_v)      17.078      9.674   1.765               0.0777 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 262.6 on 1515 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.002185,  Adjusted R-squared:  0.0008676
F-statistic: 1.659 on 2 and 1515 DF,  p-value: 0.1907
```

From the output in R, I can see that the multiple R^2 result is 0.002185, meaning that 0.2185% of the population change can be explained by air quality and college degrees. This is even less than air quality and greenness, and so clearly is not meaningful at all.

The next model I looked at is greenness with income inequality. Amenity and economic factor, again.

```
> a4 <- lm(popDiff ~ scale(Greenness_v) + scale(IncomeInequality_v), data = health_data)
> summary(a4)

Call:
lm(formula = popDiff ~ scale(Greenness_v) + scale(IncomeInequality_v),
    data = health_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3123.1  -142.0   -15.6   112.0  1671.7

Coefficients:
                          Estimate Std. Error t value            Pr(>|t|)
(Intercept)                 53.456      6.741   7.931 0.00000000000000421 ***
scale(Greenness_v)          14.425      6.766   2.132              0.0332 *
scale(IncomeInequality_v)   16.038      6.743   2.379              0.0175 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 262.2 on 1510 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.006634,  Adjusted R-squared:  0.005319
F-statistic: 5.042 on 2 and 1510 DF,  p-value: 0.006567
```

From the output in R, I can see that the multiple R^2 result is 0.006634, meaning that 0.6634% of the population change can be explained by greenness and income inequality. This is the highest R^2 value so far, but it is still less than 1% and is still not meaningful.

The fifth model I looked at is greenness with college degrees, so amenity with economic factor.

```
> a5 <- lm(popDiff ~ scale(Greenness_v) + scale(CollegeDegree_v), data = health_data)
> summary(a5)

Call:
lm(formula = popDiff ~ scale(Greenness_v) + scale(CollegeDegree_v),
    data = health_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3123.11  -140.87   -15.82   114.94  1675.31

Coefficients:
                         Estimate Std. Error t value            Pr(>|t|)
(Intercept)                53.473      6.735   7.940 0.00000000000000391 ***
scale(Greenness_v)         11.852      7.019   1.688               0.0915 .
scale(CollegeDegree_v)      7.199      7.008   1.027               0.3045
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 262.4 on 1515 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.003461,  Adjusted R-squared:  0.002145
F-statistic:  2.63 on 2 and 1515 DF,  p-value: 0.07237
```

From the output in R, I can see that the multiple R^2 result is 0.003461, meaning that 0.3461% of the population change can be explained by greenness and college degrees. This is still low, and is still not meaningful.

The last model I looked at compares income inequality with college degrees, so these are both economic factors.

```
> a6 <- lm(popDiff ~ scale(IncomeInequality_v) + scale(CollegeDegree_v), data = health_data)
> summary(a6)

Call:
lm(formula = popDiff ~ scale(IncomeInequality_v) + scale(CollegeDegree_v),
    data = health_data)

Residuals:
     Min       1Q   Median       3Q      Max
-3117.63  -142.08   -16.14   113.03  1722.70

Coefficients:
                          Estimate Std. Error t value           Pr(>|t|)
(Intercept)                 53.539      6.748   7.934 0.0000000000000041 ***
scale(IncomeInequality_v)   14.134      6.956   2.032             0.0423 *
scale(CollegeDegree_v)       7.216      6.977   1.034             0.3012
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 262.5 on 1510 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.004349,  Adjusted R-squared:  0.00303
F-statistic: 3.298 on 2 and 1510 DF,  p-value: 0.03722
```

From the output in R, I can see that the multiple R^2 result is 0.004349, meaning that 0.4349% of the population change can be explained by income inequality and college degrees. This is the second highest R^2 value that I have found, but it is still less than 1% and is meaningless.

The final step of the project was to look at an interaction model. Per the instructions, if the variables we chose were continuous (as most were) then we needed to turn one of them into a dummy variable. So I decided to turn one of the amenities into a dummy variable, and one of the economic selections into a variable. So I turned air quality and college degree into dummy variables, by saying that areas with air quality above the mean are categorized using 0, and less than the mean is a 1, and the same for college

degrees. I then used the package stargazer to interpret the results.

	Dependent variable:			
	popDiff			
	(1)	(2)	(3)	(4)
AirQualityDummy	-0.150	24.388		
	(13.893)	(20.901)		
scale(IncomeInequality_v)	25.954***			
	(7.980)			
AirQualityDummy:scale(IncomeInequality_v)	-35.271**			
	(14.930)			
scale(CollegeDegree_v)		17.292*		
		(10.048)		
AirQualityDummy:scale(CollegeDegree_v)		1.906		
		(20.662)		
CollegeDegreeDummy			10.354	-1.032
			(13.864)	(18.667)
scale(Greenness_v)			11.435	
			(9.426)	
CollegeDegreeDummy:scale(Greenness_v)			2.677	
			(13.899)	
scale(AirQuality_v)				13.154
				(9.347)
CollegeDegreeDummy:scale(AirQuality_v)				-51.944**
				(20.858)
Constant	54.041***	44.870***	48.313***	38.597***
	(8.562)	(9.926)	(9.447)	(10.619)
Observations	1,513	1,518	1,518	1,518
R2	0.007	0.003	0.003	0.005
Adjusted R2	0.005	0.001	0.001	0.003
Residual Std. Error	262.177 (df = 1509)	262.582 (df = 1514)	262.520 (df = 1514)	262.267 (df = 1514)
F Statistic	3.708** (df = 3; 1509)	1.364 (df = 3; 1514)	1.603 (df = 3; 1514)	2.579* (df = 3; 1514)
Note:				*p<0.1; **p<0.05; ***p<0.01

Shown here, the only significant result from the test is how places with higher income inequality affects the difference in population.

I think the biggest challenges that I faced while doing this were data management, and interpretation of the results. By data management, I mean because of how many different formulas, processes, etc that I needed to use to do the project I sometimes felt overwhelmed trying to keep it all organized on paper and in my brain. This, I don't think, was detrimental to me at all though as I think this will help me with future classes and in my career. I think the interpretation was difficult as well for me because heading into the project, I didn't fully understand interaction models, and how they related to the multiple regression modeling and analysis. I think, however as the project went on I started to understand how each piece is a key for the rest of the models.