Final Project: SOCI 20253

Robert Crump, 12231546 December 11, 2020

Research Question:

- Are commute times spatially clustered and can they be explained by other variables related to socio-economic status and/or vehicle traffic?

This project explores the causal relationship and potential spatial clustering of commute times, reported vehicle crashes, median income, and median rent within the Census tracts of Chicago. By choosing these variables, I hope to pose questions related to quality of life and the experience of using cars within the city.

I want to learn about different ways to advocate for expanding transit within cities. I was inspired by data I found on the Chicago Metropolitan Area Planning (CMAP) website related to transit needs. Originally, I was hoping to reproduce their maps using their datasets, which are based on current housing density and land use surveys. Their data is not especially well described, so I could not be confident in my interpretation. I decided to switch focus to a more human level analysis, not to replace CMAP's data but to enhance it with added dimensionality.

Hypothesis:

- *Null hypothesis:* commute times are spatially random and are not associated with other variables related to socio-economic status and/or vehicle collisions.

- *Alternative hypothesis:* commute times are not spatially random and are associated with other variables related to socio-economic status and/or vehicle collisions.

Data Description:

My dependent variable is Average Commute Times in Minutes as recorded by the American Community Survey in 2018 (ACS) collected using the R package `tidycensus`. Two of my Independent variables are Median Rent and Median Income from the same source. My third independent variable is Vehicle Crashes Per 1000 People collected from the Chicago Open Data Portal. My areal unit is Census tracts within the City of Chicago, excluding tracts with zero or missing values for one or more variables. This exclusion appears most notably with the absence of O'hare International Airport which did not have a measure for commute times.

The ACS data is designed to be evaluated at the Census tract level, so compiling those variables requires just a few lines of code in R. Associating the crash data to the areal unit took a few more steps and wrangling to include as a continuous variable amongst the others. Each incident (between 2017-2019) has a latitude/longitude coordinate pair, so with a few steps they can be aggregated within a Census tract polygon. The motivation behind using crash incidence is to try to identify roadways that may be poorly designed or otherwise prone to collisions.

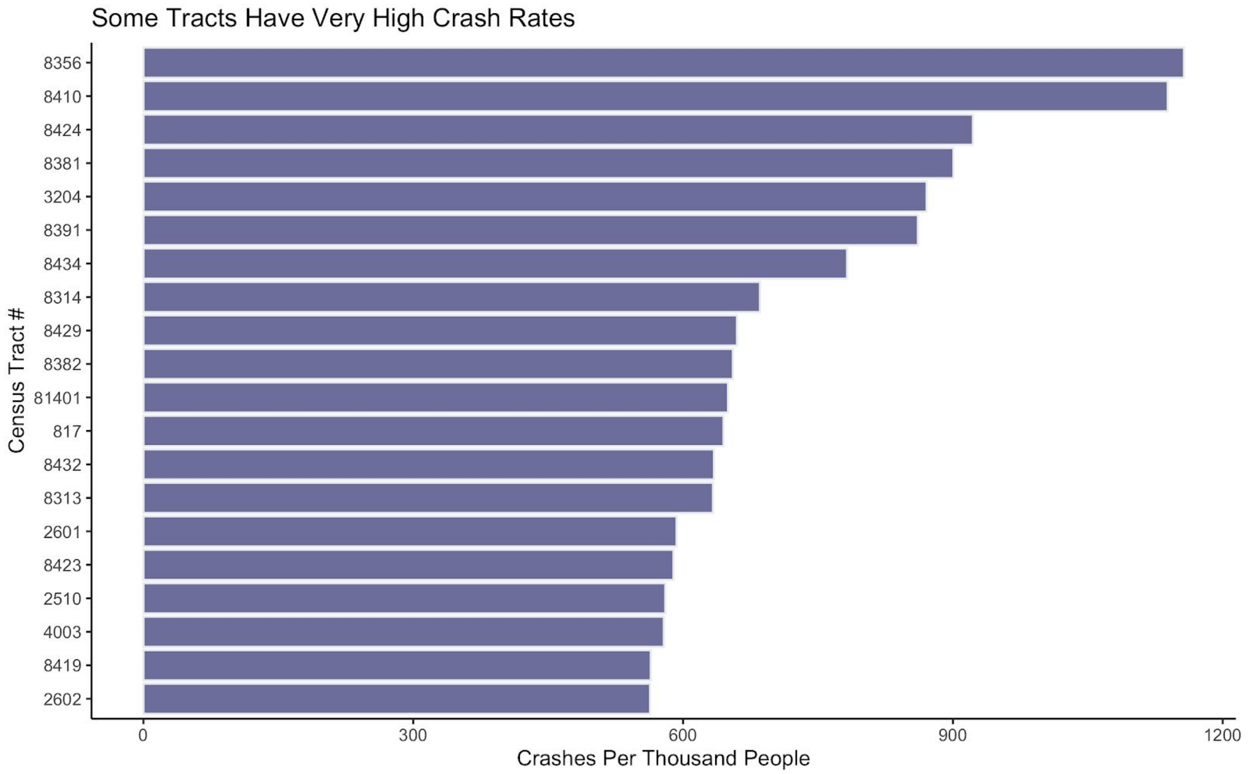
First, using the `st_as_sf ` function, I transformed the lat/lon columns into a point geometry layer in its own column. Then I used `st_join` in conjunction with `st_within` to link the crash data with the Chicago Census tract polygon layer, giving me a raw count of crashes within each tract. Then, I joined the `crash per tract` data frame with the other Census variables plus total population per tract to generate an intensive variable of crashes per one thousand people per tract. After omitting any NA or zero values, we're ready to proceed to exploratory data analysis.

Exploratory Data Analysis:

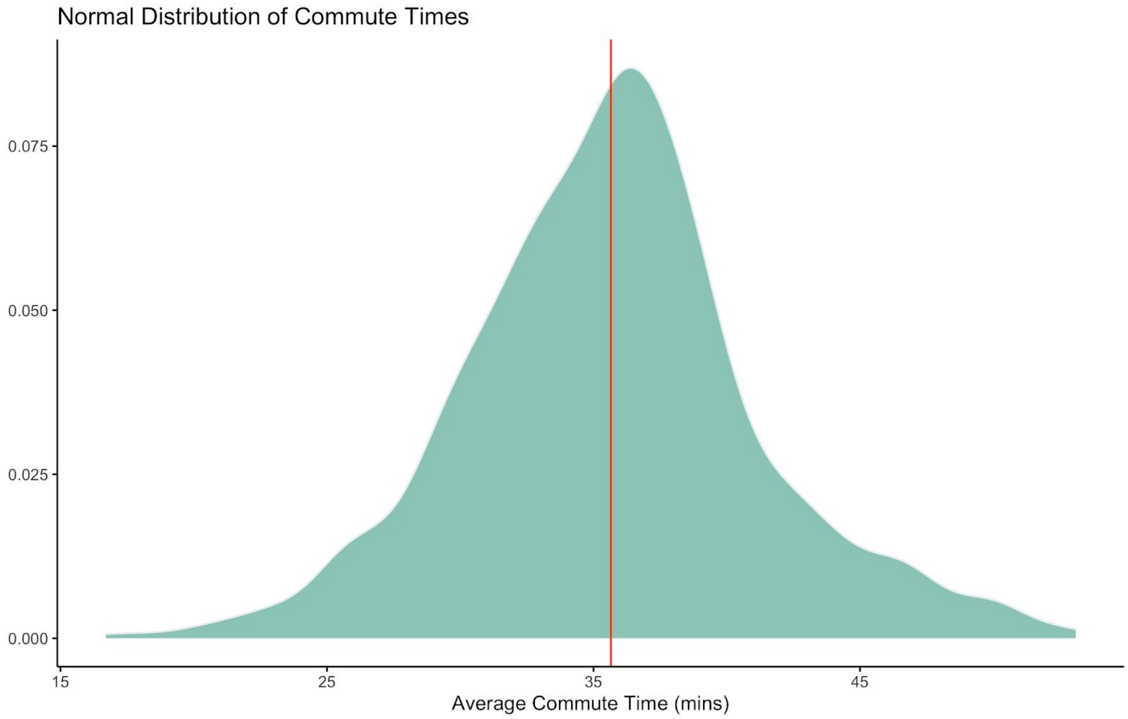
I initially intended to use crash incidents from the Chicago data portal as my dependent variable. My original hypothesis was that socio-economic factors (median rent and median income) would be associated with higher incidence of crashes. I thought it would be interesting to include commute times as a third independent variable as a means of adding more substance to the 'quality of life' analysis. The longer you are on the road, the more likely you are to be part of a collision. However, none of the standard regression analysis methods produced meaningful or significant results. When I ran the data through GeoDa and different spatial autocorrelation metrics, I could clearly see significant regions, so I figured something had to be wrong in my empirical specification.

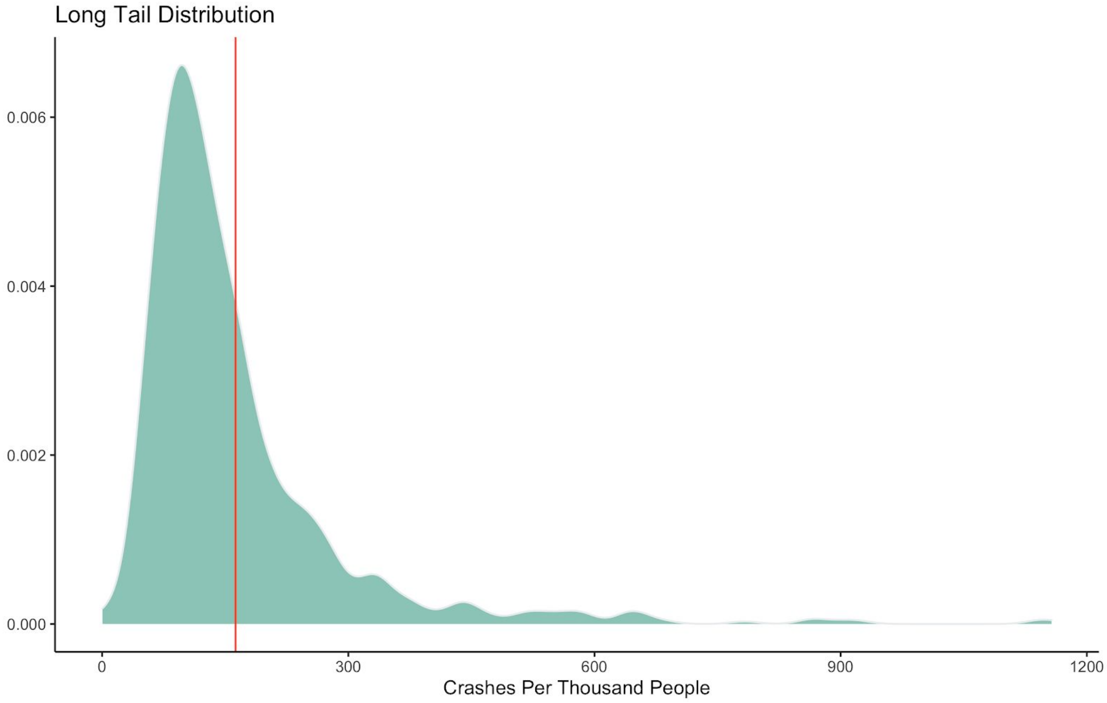
After many hours of testing different types of regression specifications, I realize my basic error in selecting my variables. The reason crashes and commute times are not correlated is because they are identifying very different sorts of information. Crashes within tracts are events not necessarily associated with people living in the tract because people living outside the tract drive through it, and areas with a higher concentration of traffic and crashes are more likely associated with roadways, not population statistics. Average commute time of residents is likely to be associated with spatial concentration of crashes, but not the other way around. One variable is measuring aggregate behavior of a population within a geographic region, while the other is aggregating event data in the same that is not dependent on that behavior. But, we should anticipate that a region with higher crash propensity would lead to longer commute times. Building from this hypothesis, we can ask a deeper question about socio-economic status of the region as it may or may not correlate to commute time.

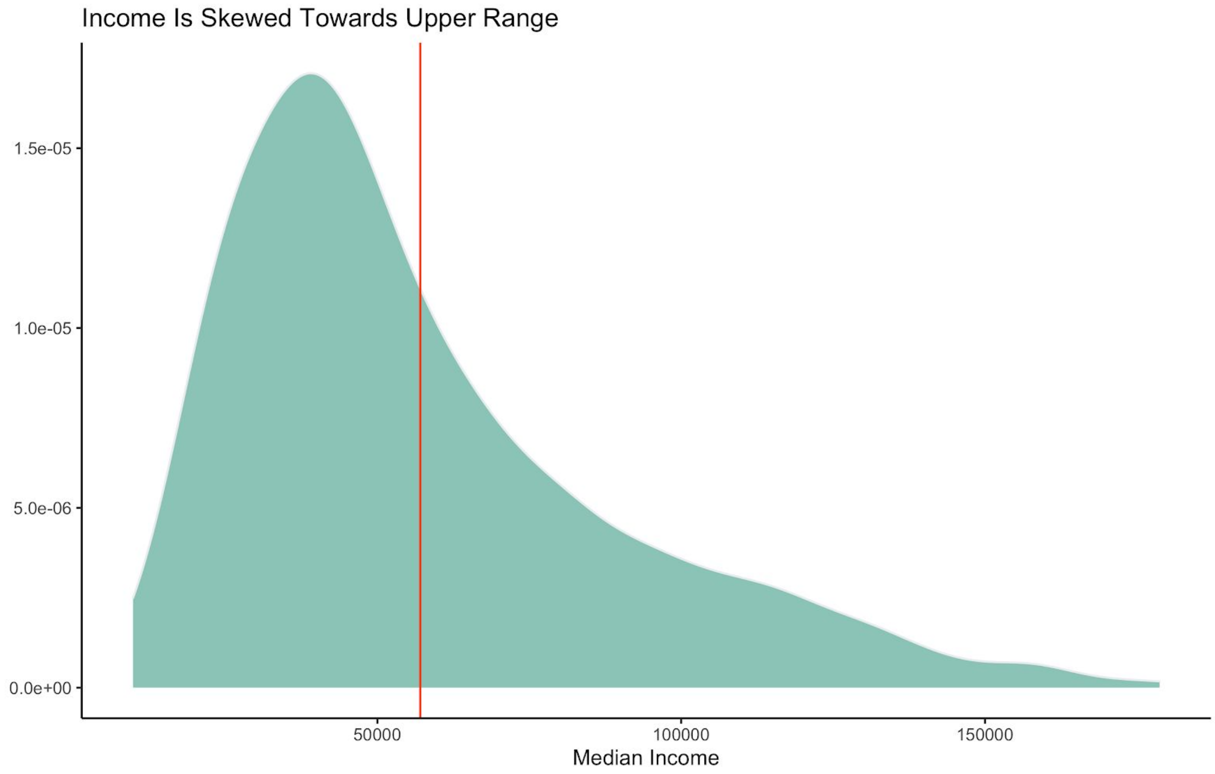
To begin exploring the data, I sorted and filtered the top twenty rankings of all four variables and plotted them on histograms. Of these four variables, only crash density had any meaningful dispersion among the top twenty tracts. The other three were basically flat.

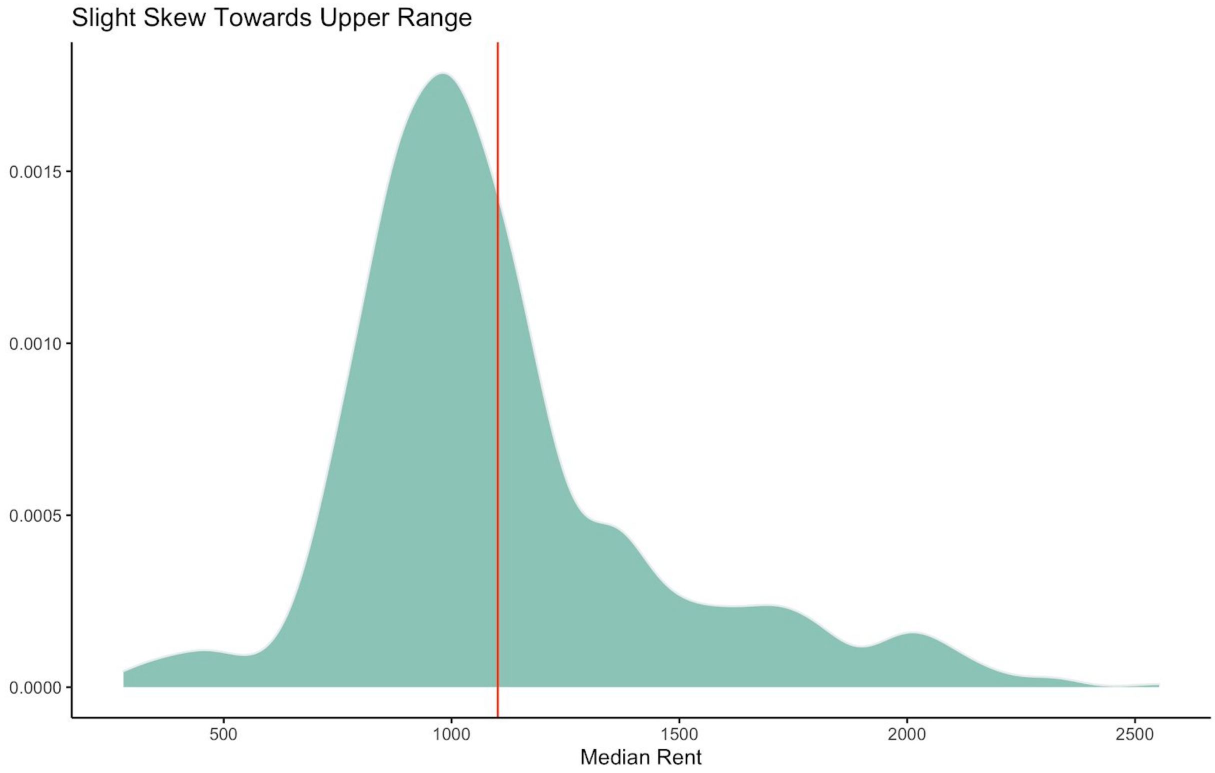


Then I plotted distribution graphs of each variable with the mean plotted as a red line, starting with the dependent variable, average commute times in minutes.



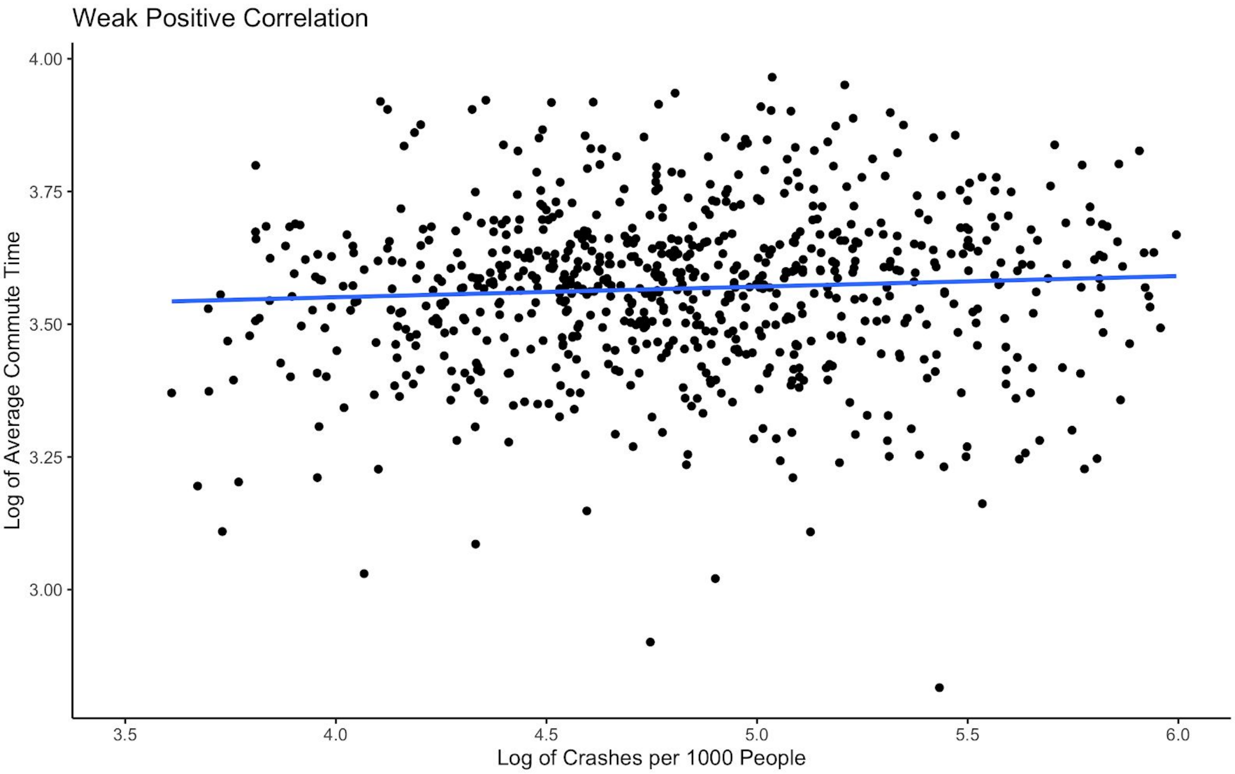




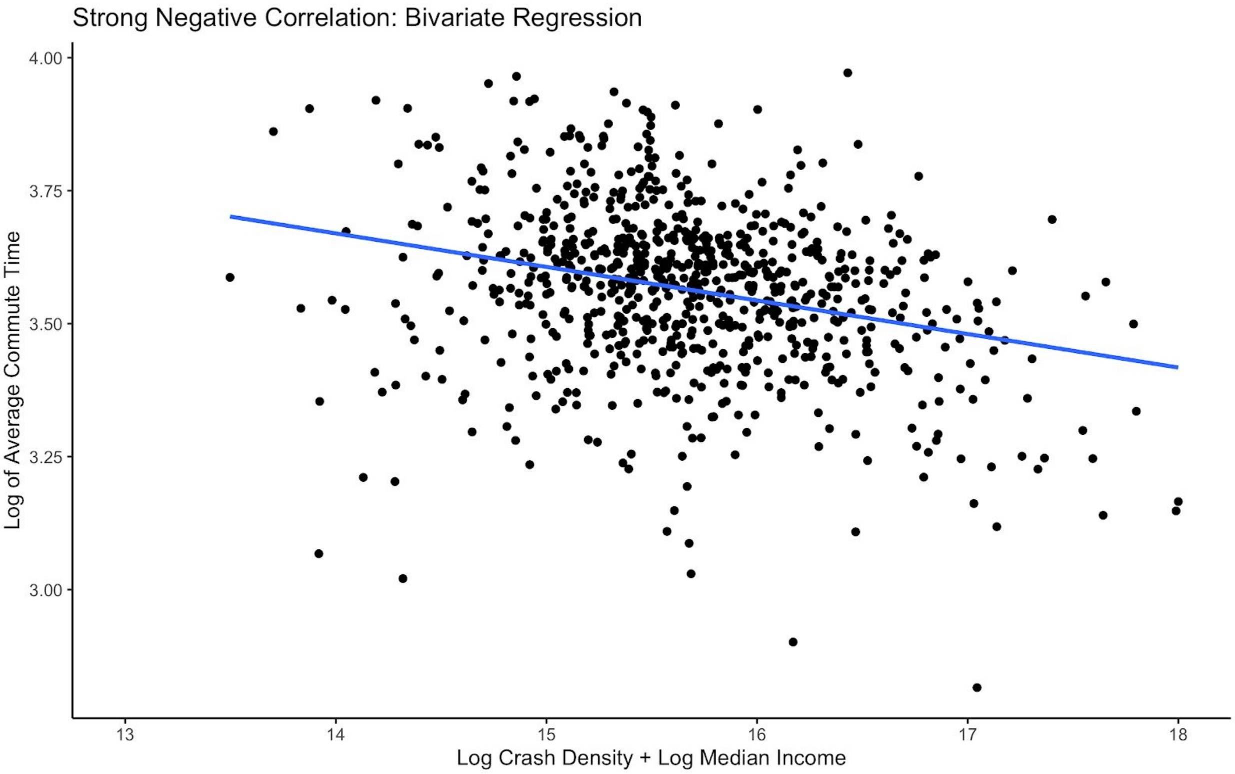


When I started plotting regression lines, it was difficult to discern any meaningful interaction between the dependent and independent variables. After trying a number of different options like eliminating outliers and demeaning the data, I decided finally to evaluate the logarithms of each variable. The density graphs above show major dispersion differences and very different numeric scales for each variable of interest. Averages are highly affected by outliers, so the dependent commute time variable is highly skewed. The crash density variable is also highly skewed, but we should expect this to occur in areas of high traffic and/or poor road conditions. Median rent and income are in dollar amounts, whereas commute time is measured in minutes and crashes are rates of incidence. Logging these different variables measured in different units and scales allows us to examine the effects of marginal changes to various combinations of variables.

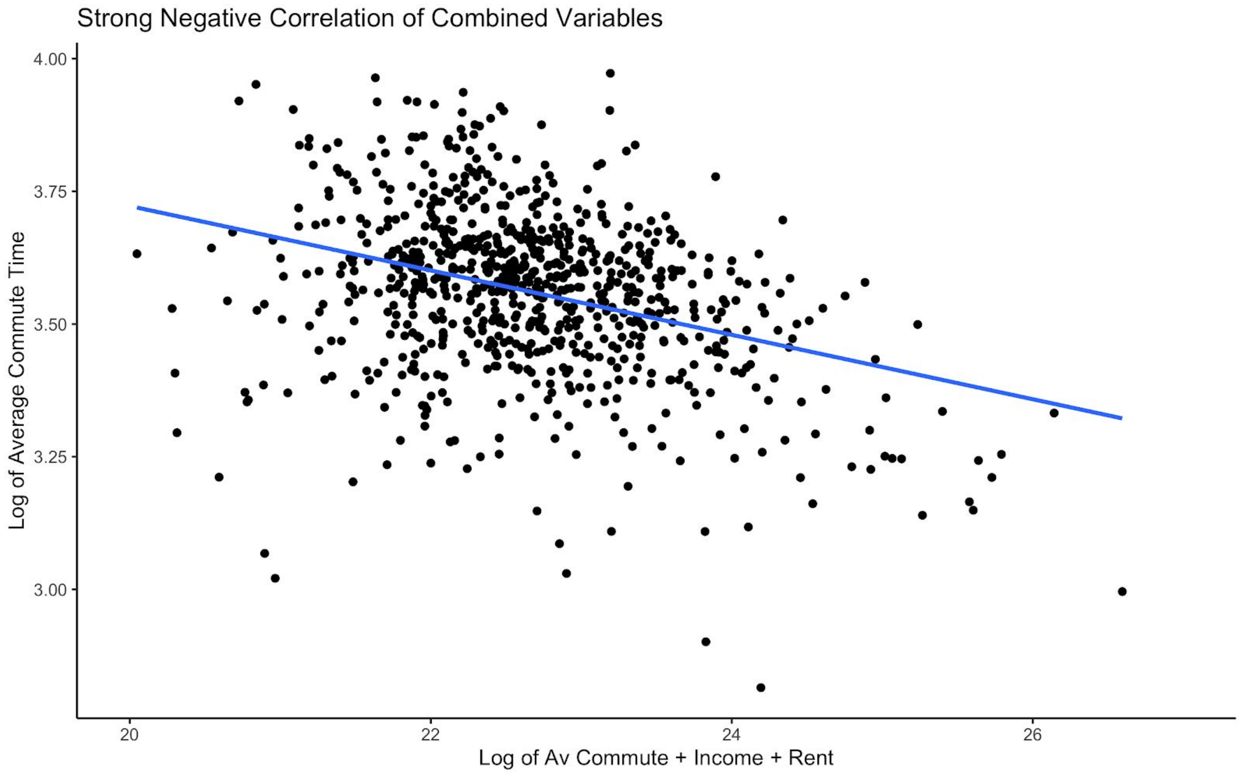
All of the graphs below use a standard significance threshold of $p < 0.05$.



Weak Positive Correlation
Log of Average Commute Time
4.00
3.75
3.50
3.25
3.00
3.5
4.0
4.5
5.0
5.5
6.0
Log of Crashes per 1000 People



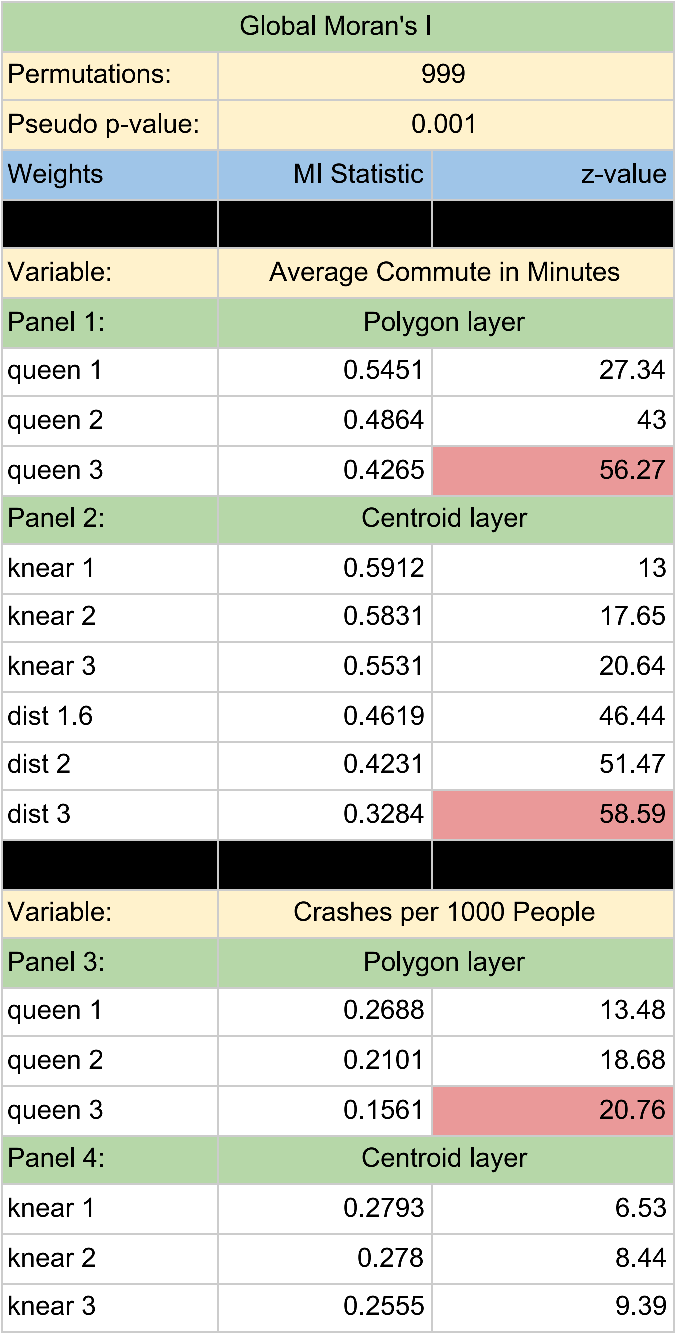
Strong Negative Correlation: Bivariate Regression
Log of Average Commute Time
4.00
3.75
3.50
3.25
3.00
13
14
15
16
17
18
Log Crash Density + Log Median Income



When evaluated as a single regression crash density is only weakly correlated. When combined with the other area population statistics in multivariate regressions, the correlation is strongly negative. Higher rents and incomes reduce average commute time, while crash density increases commute times slightly.

Global Spatial Autocorrelation, Global Moran's I:

I tested nine different weights for Global Moran's I to inform my decision of which weights to use for local spatial analysis. Consistent across all 8 panels, the queen 3 weight and 3 miles distance-band weights were the most significant. These were unexpectedly high z-scores, which are reflected in the cluster maps of Chicago's nearly 793 Census tracts used for this project.



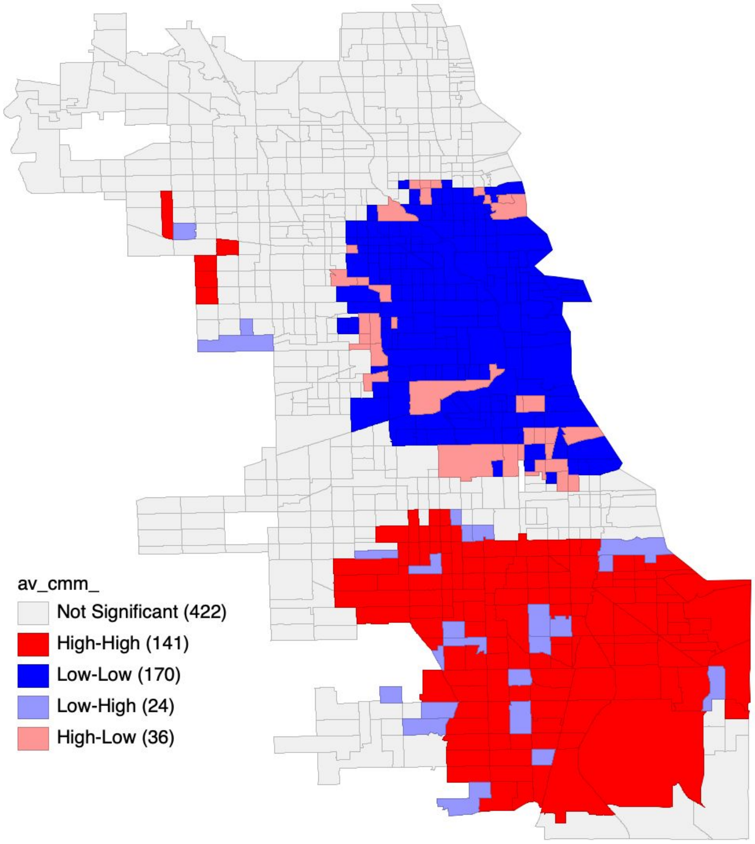
Global Moran's I		
Permutations:	999	
Pseudo p-value:	0.001	
Weights	MI Statistic	z-value
Variable:	Average Commute in Minutes	
Panel 1:	Polygon layer	
queen 1	0.5451	27.34
queen 2	0.4864	43
queen 3	0.4265	56.27
Panel 2:	Centroid layer	
knear 1	0.5912	13
knear 2	0.5831	17.65
knear 3	0.5531	20.64
dist 1.6	0.4619	46.44
dist 2	0.4231	51.47
dist 3	0.3284	58.59
Variable:	Crashes per 1000 People	
Panel 3:	Polygon layer	
queen 1	0.2688	13.48
queen 2	0.2101	18.68
queen 3	0.1561	20.76
Panel 4:	Centroid layer	
knear 1	0.2793	6.53
knear 2	0.278	8.44
knear 3	0.2555	9.39

dist 1.6	0.168	16.45
dist 2	0.153	18.29
dist 3	0.1026	19.35
Variable:	Median Income	
Panel 5:	Polygon layer	
queen 1	0.762	38.68
queen 2	0.6985	64.49
queen 3	0.6301	81.17
Panel 6:	Centroid layer	
knear 1	0.8348	19.16
knear 2	0.8101	23.77
knear 3	0.7843	28.11
dist 1.6	0.6446	65.57
dist 2	0.6159	74.6
dist 3	0.5007	92.89
Variable:	Median Rent	
Panel 7:	Polygon layer	
queen 1	0.6255	30.29
queen 2	0.5759	51.34
queen 3	0.509	66.67
Panel 8:	Centroid layer	
knear 1	0.6819	15.73
knear 2	0.6483	19.29
knear 3	0.6455	23.42
dist 1.6	0.5325	54.23
dist 2	0.5034	62.14
dist 3	0.4189	76.76

Univariate Local Spatial Autocorrelation

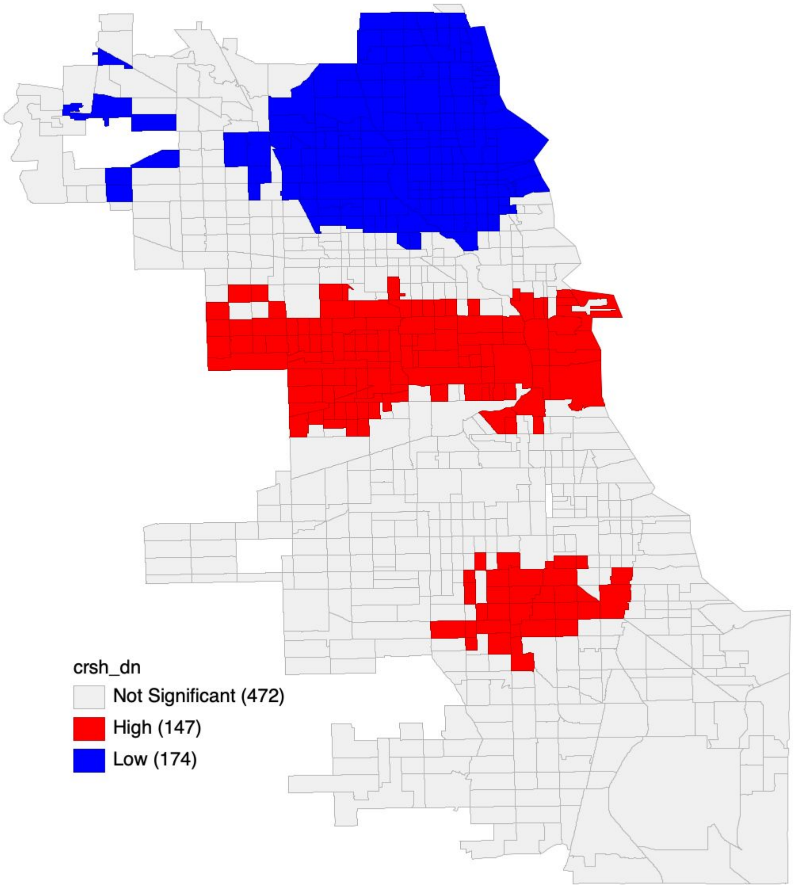
Map 1, Average Commute Times (Minutes):

Local Moran's I, 3 Mile Distance Band Weight, p-value = 0.001



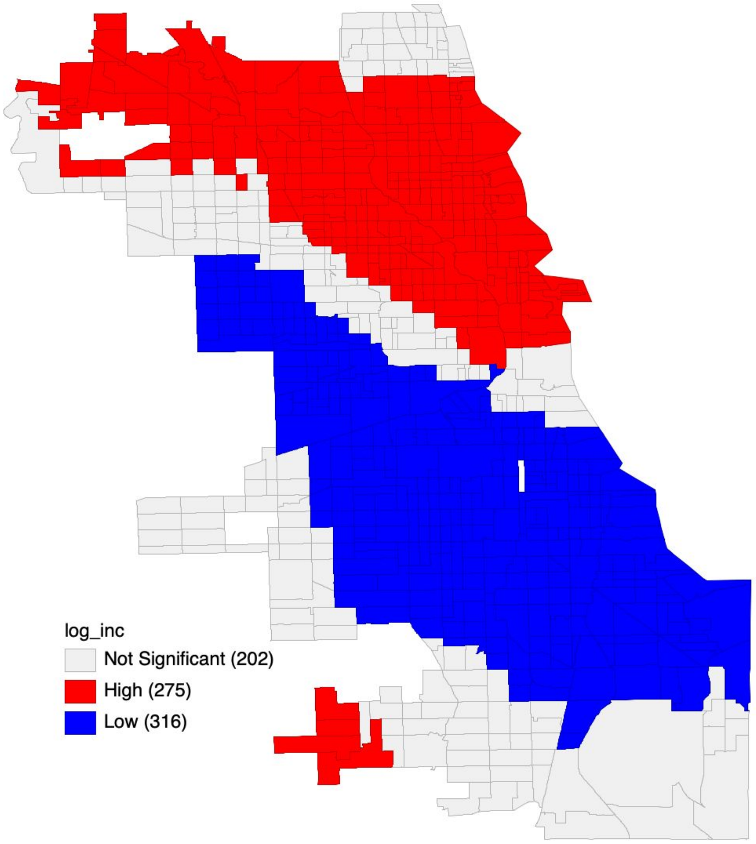
Map 2, Crash Density (per 1000 People):

Local Geary's C, 3 Mile Distance Band Weight, p-value = 0.001



Map 3, Median Household Annual Income (Dollars):

Local Geary's C, 3 Mile Distance Band Weight, p-value = 0.001



Interpretation:

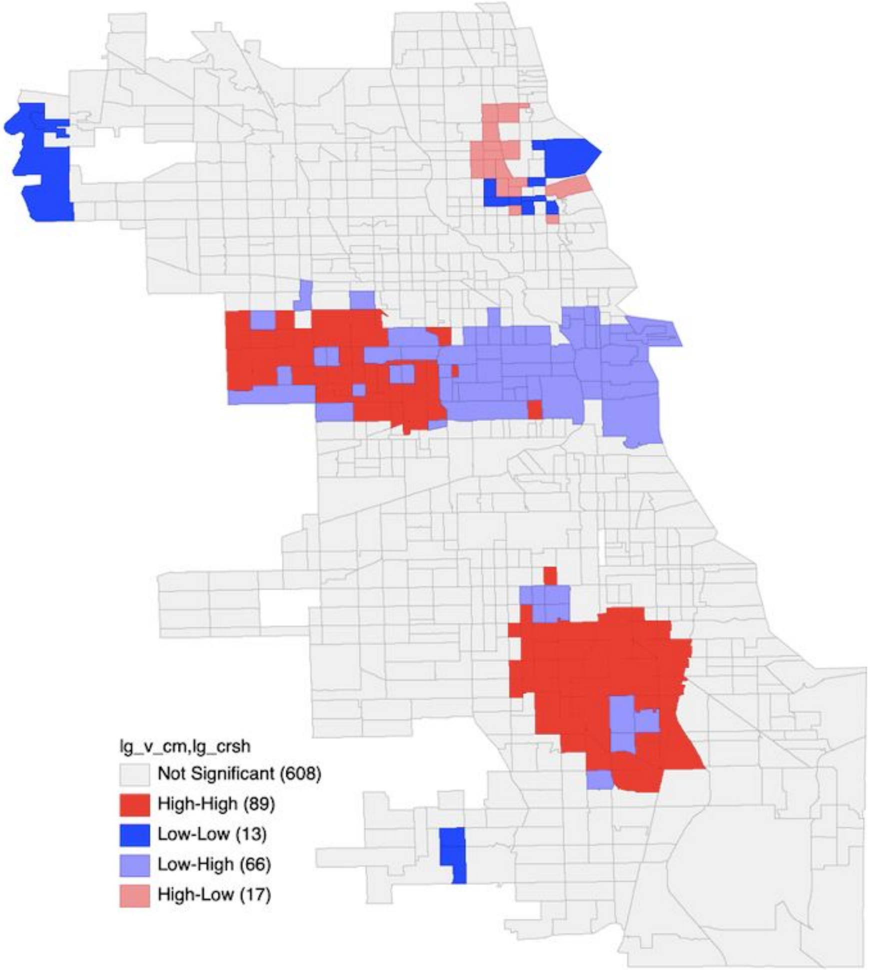
Maps 1-3 represent measures taken at a 3 mile distance band weight with a significance level of 0.001. Lower or higher significance levels did not produce interesting clusters, and in some cases, blank maps. Queen 3 weights produce nearly identical clusters across each variable.

The most interesting map to me is the concentration of crash density cutting through the center of the city. These tracts are located near Interstate 290 headed into The Loop. The income cluster map is nearly identical to the rent map, so I only included one. Ultimately, these two variables are interchangeable with regard to spatial trends. Average commute times are low in and around The Loop, but high commute times are concentrated on the Southside. I would interpret this as illustrating the daily commutes of high and low income workers entering the job center downtown, keeping in mind that these data represent a pre-COVID ear.

Multivariate Local Spatial Autocorrelation

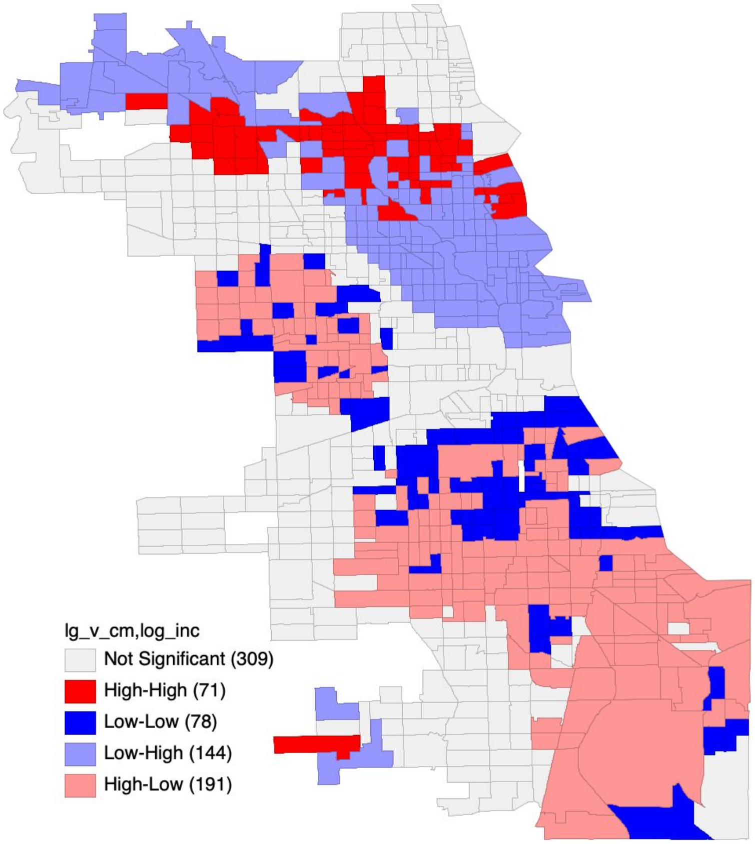
Map 4, Log of Commute Time/ Log of Crash Density:

Local Bivariate Moran's I, Queen 3 Weight, p-value = 0.001



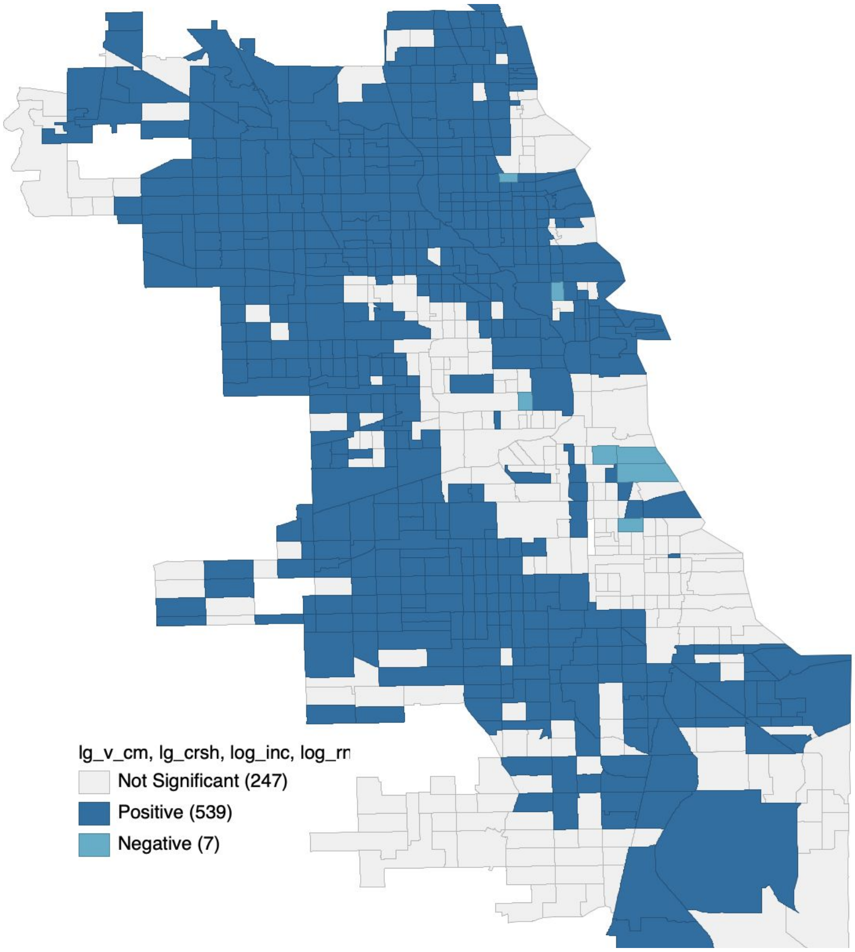
<u>Map 5, Log of Commute Time/ Log of Median Income</u>:

Local Bivariate Moran's I, Queen 3 Weight, p-value = 0.001



<u>Map 6, Combination of All Logged Variables</u>:

Local Multivariate Geary's C, 3 Mile Distance Band Weight, p-value = 0.001



Interpretation:

Finally, we can interpret multivariate spatial autocorrelation of the four variables. Since these variables are very different in unit, scale, and type, they are expressed here as in the above regression graphs as logged variables. Without this step, interactions between the gross quantities are difficult to discern or altogether not significant.

Displayed above are the most interesting patterns achieved using the Bivariate Local Moran's I and Multivariate Geary's C statistics. In Map 4, we can observe the same cross-cutting cluster along Interstate 290, as well a Southern cluster compared to Map 2. Interestingly, the high-high cluster is only on the West side of town because commute times become significantly less pronounced as one nears The Loop.

Map 5 illustrates an interesting pattern following Interstate 90. The bivariate cross-products of logged commute times and logged median income cluster around the highway, but in divergent ways. All the tracts on the Northside are of relatively higher income compared to the mean, but experience both high and low commute times. This leads me to believe that some sections or possibly different on/off ramps to Interstate 90 impact traffic differently.

I tried many different combinations using the Multivariate Local Geary's C statistic, and each produced a similar pattern that doesn't show much in the way of clusters except that the central part of the city is not significant compared to the outer edges. I decided to include the map of all logged variables to illustrate this consistent pattern. In part, this demonstrates to me that mixing of different variables can have a diluting effect on analysis. None of the other combinations I tried, even with k-nearest neighbor and other weights with smaller z-scores, illuminated a different pattern if any pattern at all.

At the same time, teasing out each of these variables does isolate one particular cluster on the Southside that is negatively affected by each of these measures of quality of life. Maps 2 and 4 identify a cluster of high crash rates and high commute times outside the Interstate 290 band that is also represented in low income/ rent clusters. After some quick searching on Google Maps, I have identified this neighborhood as Englewood at the terminus of the Green Line L Train. Perhaps if this line could be extended further West or South or if service from the end of the line were increased commute times and crash rates may decrease and quality of life may increase.