Spatial Data Science Final Project: Assessing how the Relationship Between Economic Well-Being and Commute Time Varies in Near and Far from Manhattan

By Will Fisher

Introduction

For this project, I wanted to explore the relationship between economic well-being and commute time in New York City. In general, there is not always a clear relationship between the two, as the layout of metropolitan areas can affect where the most desirable areas to work and to live are. However, in NYC, the lower half of Manhattan has both a high concentration of wealthy businesses and wealthy residents. This led me to hypothesize that as areas get further from the lower half of Manhattan, economic well-being and commute time will have a positive relationship, but closer to Manhattan, the two will have a negative relationship. To test this hypothesis, I used 2008-2012 ACS data from NYC, downloaded from the Geoda Data and Lab Web Page(https://geodacenter.github.io/data-and-lab/NYC-Nhood-ACS-2008-12/). To measure economic well-being in an area, I used three variables: unemployment rate(uemprate), percentage of households on public assistance income(wpa_pct), and percentage of households with income under $10,000(under10). These are my independent variables. To measure commute time, I used the percentage of the working-age population with a commute to work of over 45 minutes(over45_pct).

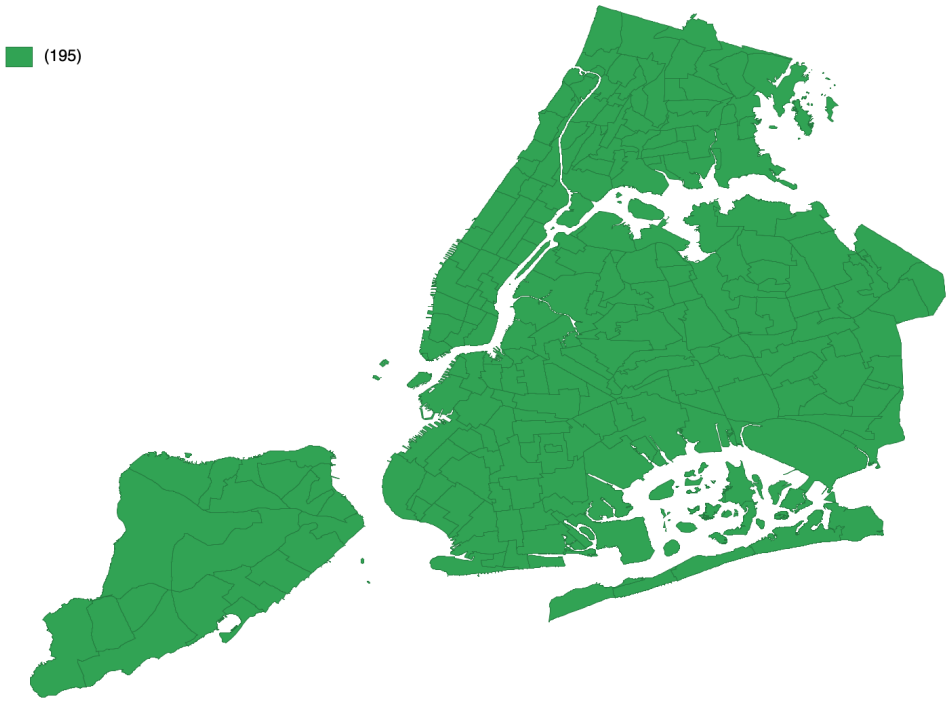
Projection

The projection used for this data set is WGS84 (+proj=longlat +datum=WGS84 +no_defs)

Geography (Figure 1)

The geography used for this project is the 195 districts the ACS divides NYC into, which is shown below in Figure 1. The first reason for this is that the number of districts provides a sufficiently large sample of spatial observations, which helps to control for outliers and provide an accurate representation of statistical patterns across the city, which should help for the data to avoid issues involving the modifiable areal unit problem. Additionally, NYC is one of the most densely populated areas in the country, and small districts are appropriate to represent it as there are large changes between relatively small spatial areas.

Figure 1: Geography of NYC ACS Districts



Data and Variables (Table 1)

- over45_pct: This is a continuous variable, with a possible range from 0 to 100. It was created by adding up the variables of the totals of people commuting for more than 45 minutes to work, and then dividing by the total working population. This is the dependent variable.
- uemprate: This is a continuous variable, with a possible range from 0 to 1, measuring the rate that people in a district are unemployed. This is an independent variable.
- wpa_pct: This is a continuous variable, with a possible range from 0 to 100, which was created by dividing the total number of households on public assistance income by the total number of households. This is an independent variable.
- under10: This is a continuous variable, with a possible range from 0 to 100, which was created by dividing the total number of households with income under $10,000 by the total number of households. This is an independent variable.

The three independent variables were chosen because they measure economic well-being in three different ways: by employment, by government support, and by income. Since these variables are inversely related to economic well-being, my hypothesis would support these variables and commute time having a positive relationship in lower Manhattan, and a negative relationship in

areas far from lower Manhattan. The dependent variable was chosen because it represented the upper half of commuting intervals, as surveyed by the ACS.
Table 1: Descriptive Statistics

Variable	Min	Max	Mean	Std Dev
over45_pct	0	35.692	22.251	6.743
uemprate	0	.5698	.1043	.0514
wpa_pct	0	13.862	4.334	3.254
under10	0	14.960	4.956	3.171

Note: There were four, out of 195 total districts, that did not have complete data, which is why the minimum of all these variables is 0. However, despite this, the large number of observations should make the effect of these outliers minimal.

Exploratory Data Analysis and Linking and Brushing(Figures 2-7)

For the individual variable maps, natural breaks were used for the bins, as it helps to determine the break points that yield groups with the largest internal similarity. 5 quantiles were also used for the individual maps as a way to divide the large number of observations without creating outlier bins. Figure 2 displays what might be expected from the introduction, which is that generally, as one gets further from the lower half of Manhattan, commute times increase, as that is where the highest concentration of jobs is. However, this is not a perfect description of what Figure 2 shows, as there is a lot of variation in commute time within the outer boroughs. Figures 3, 4, and 5, mapping out unemployment rates, percentage on public assistance income, and percentage with incomes under $10,000, show very different patterns from Figure 2, suggesting the variables do not have the same relationship with commute time across the city. Those variables are generally highest in the south Bronx, central Brooklyn, and northern Staten Island, area where the commute percentage is higher than in Manhattan, but not as high as in east Queens, where those variables are generally lower.

Figures 6 and 7 explore this analysis further by linking and brushing a parallel coordinate plot and a scatter plot of unemployment versus commute time, respectively, with a base map of the city. Figure 6 shows that the areas with the highest commute times tend to be on the lower ends of the economic variables (meaning that they may have better economic well-being), and that they are spread out over the city. Other than them all being outside Manhattan, there is not a particularly clear spatial pattern to them, suggesting that it is not just about distance from Manhattan. Figure 7 selects observations with high commute times and low unemployment rates from a scatter plot of the two variables. Some of the areas selected are different from the areas selected in Figure 6, but there does seem to be some similar patterns in east Brooklyn, east Queens, and the north Bronx.

Figure 2: Natural Breaks Map of over45_pct, 5 quantiles

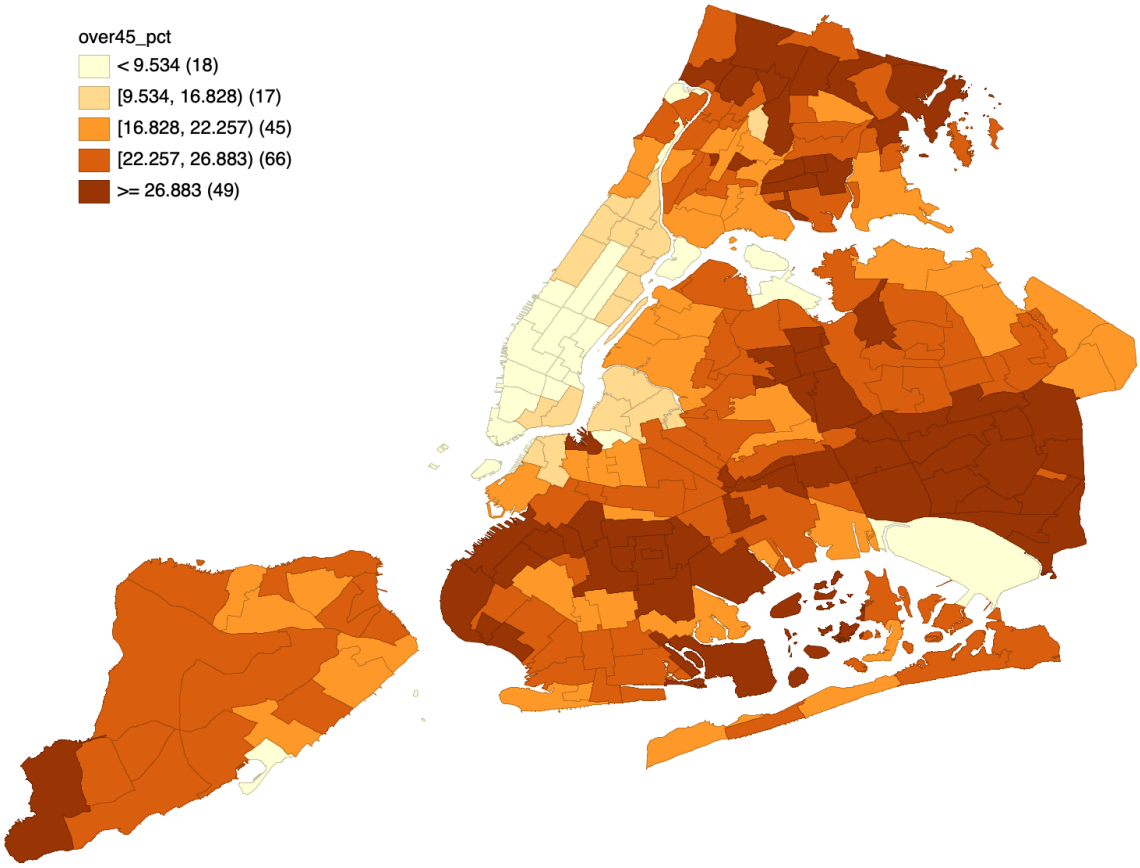


Figure 3: Natural Breaks Map of uemprate, 5 quantiles

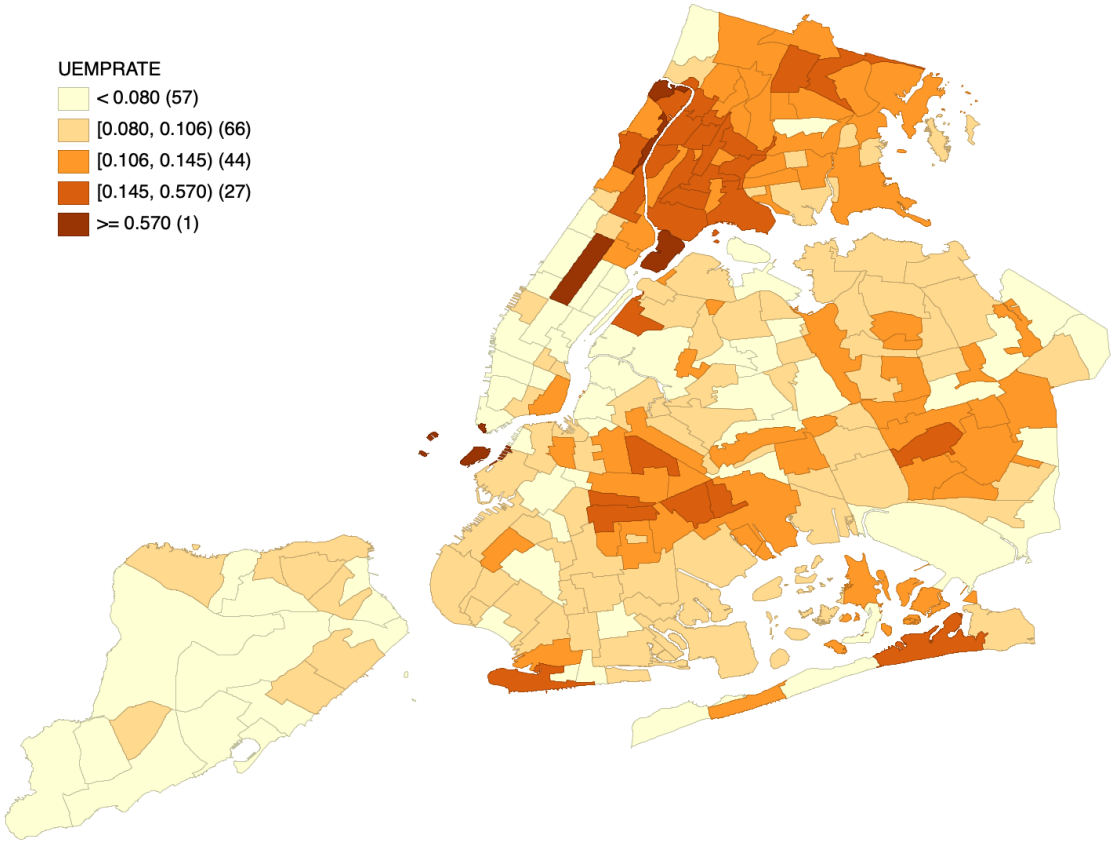


Figure 4: Natural Breaks Map of wpa_pct, 5 quantiles

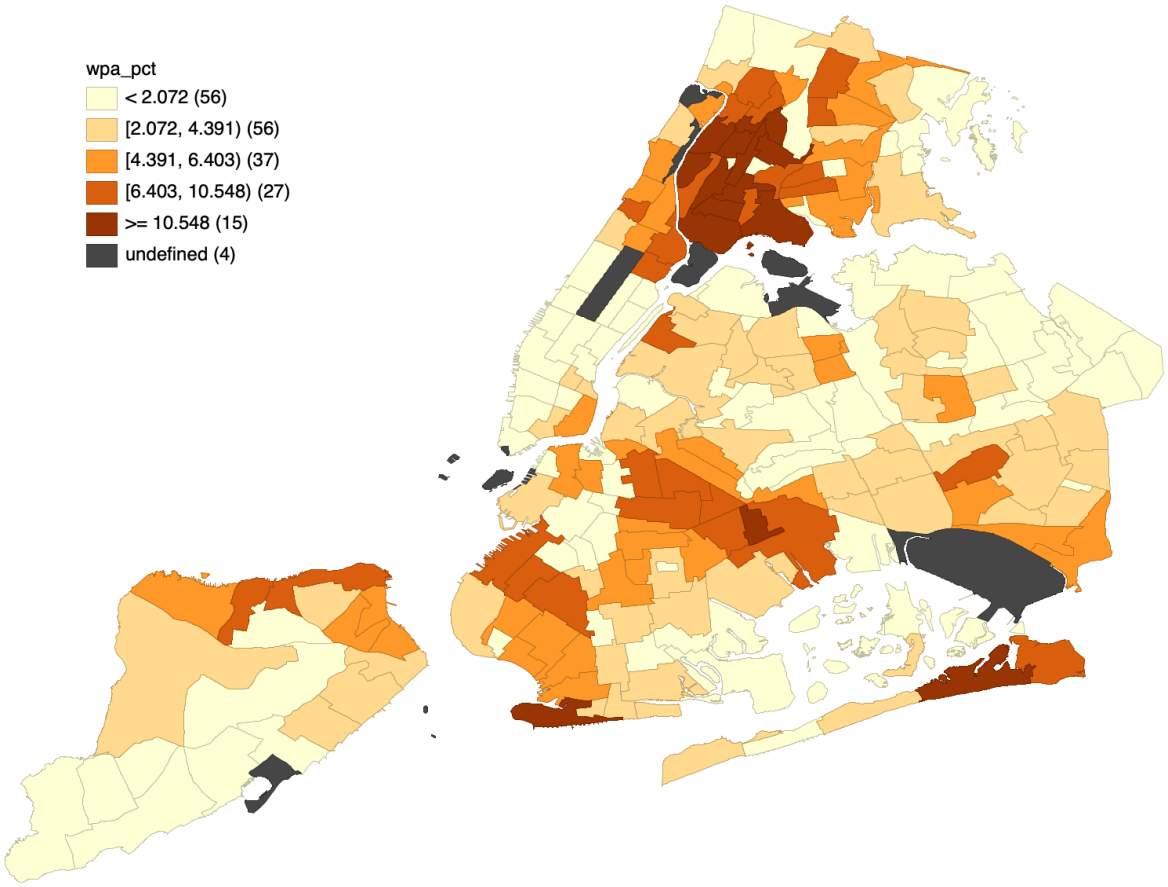


Figure 5: Natural Breaks Map of under10, 5 quantiles

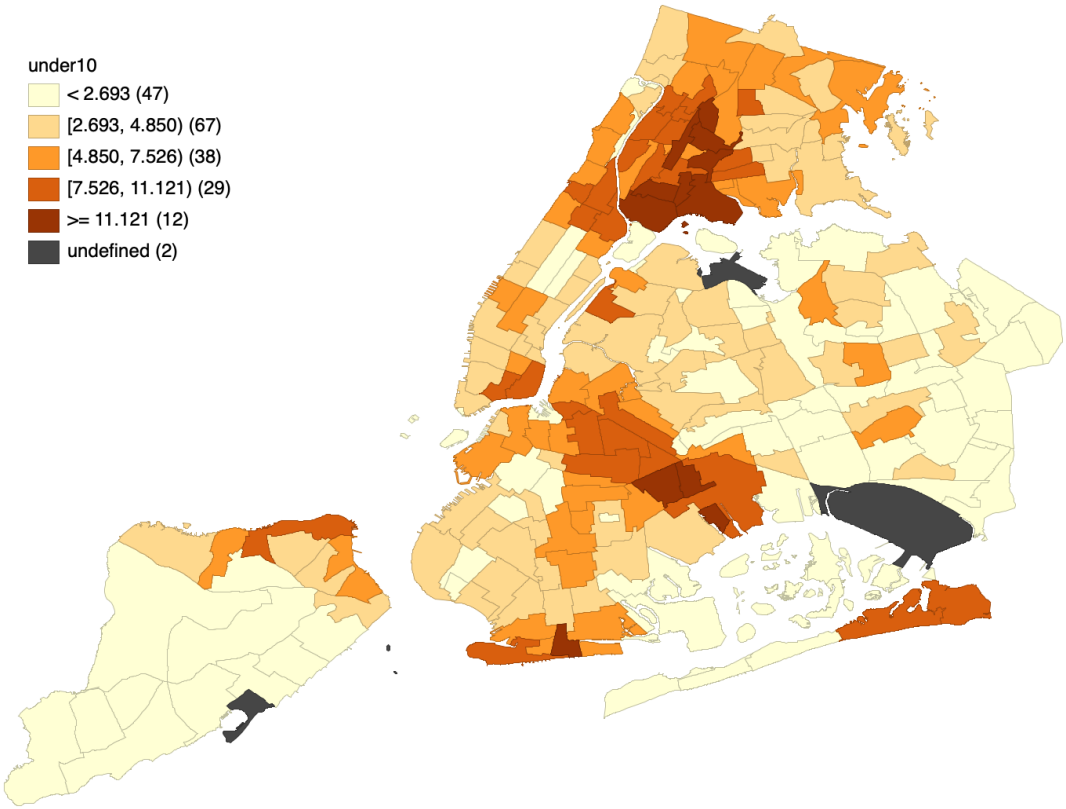


Figure 6: Parallel coordinate plot, brushed with basemap

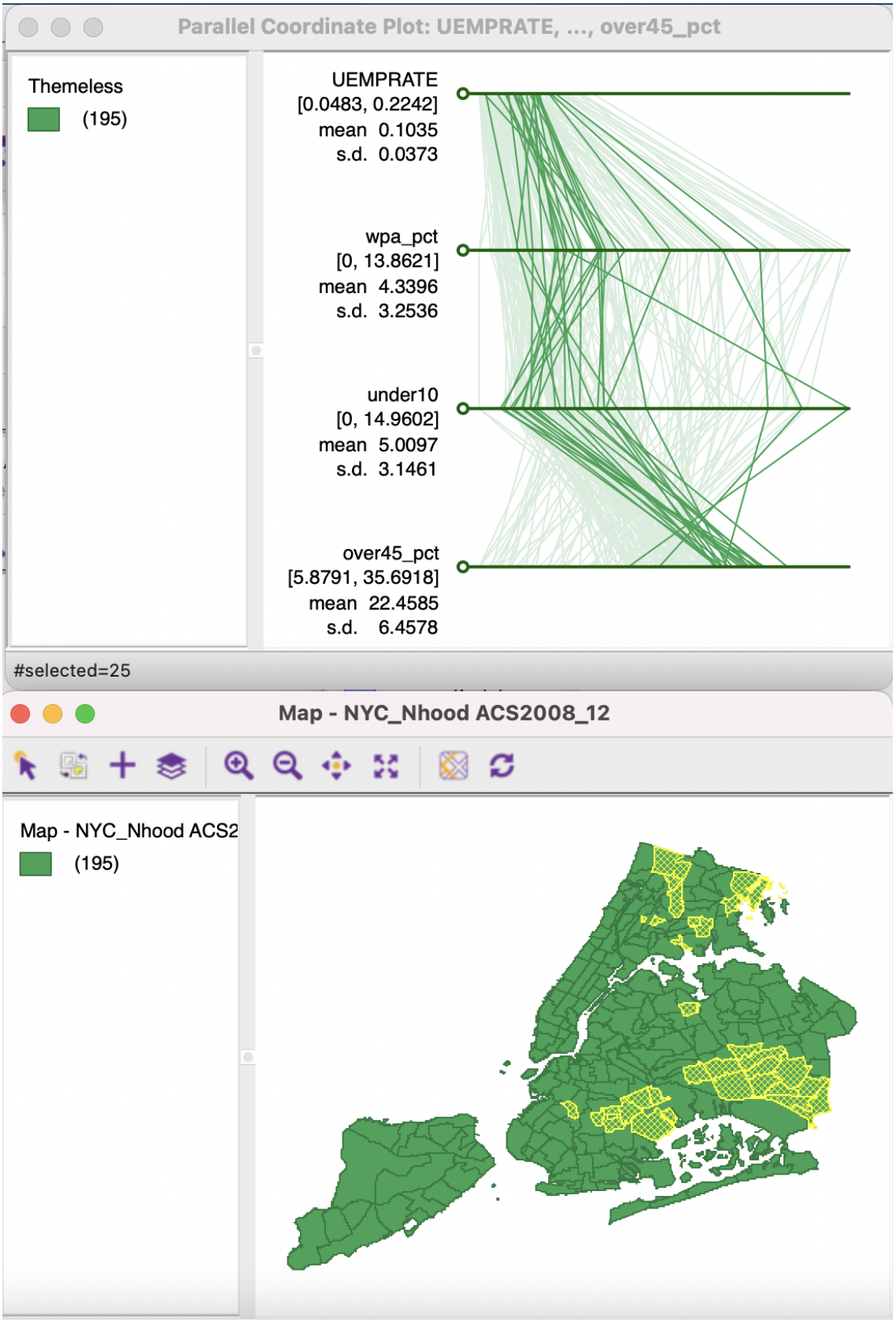
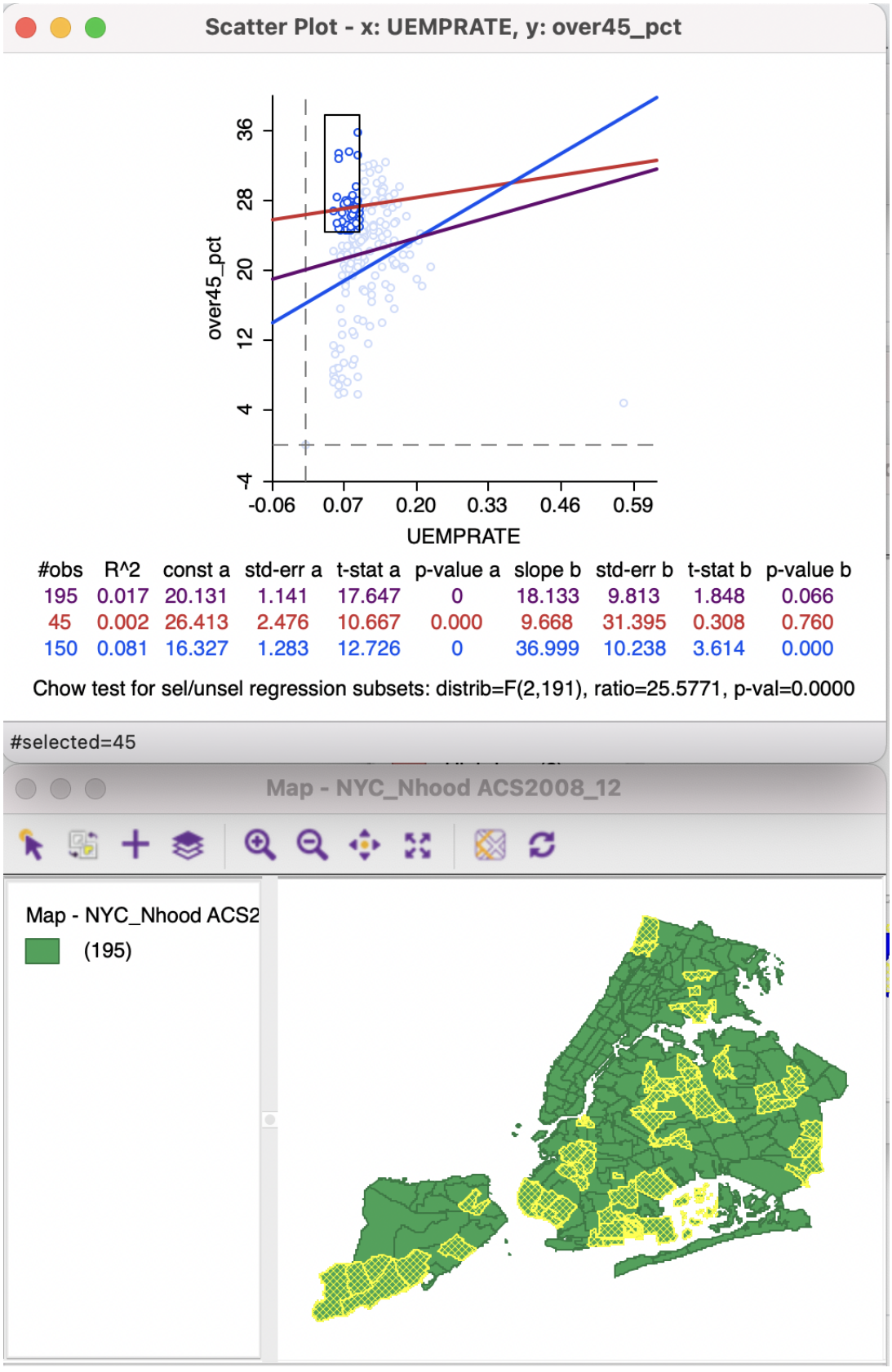


Figure 7: Unemployment vs commute scatter plot, brushed with basemap



Weights, Sensitivity Analysis, and Univariate Local SAC (Table 2, Figures 8-17)

Weights

I selected three different weights for sensitivity analysis, shown in Table 2, to measure clusters of the different variables in a variety of ways and test to make sure the clusters were consistent. I used a queen first order contiguity weight to measure immediate neighbors of ACS districts, which resulted in an average of 5.59 neighbors per district. I used a distance band=6km weight to capture clusters in dense areas, using 6km since it was just above the minimum necessary to make sure every district had at least one neighbor, but it still resulted in an average of 24.01 neighbors per district. I used a K nearest neighbors = 10 weight to capture the middle of these two weights, which can be difficult given the dense nature of New York City.

Sensitivity Analysis

The variable measuring commute time, over45_pct, consistently had a high level of spatial correlation across all three weights, with a z-score of at least 15 for all of them. Additionally, across all three weights, the ranking of the four variables in terms of Moran's I and z-score was very consistent across the different weights, suggesting consistent spatial patterns. The Moran's I and z-score for wpa_pct and under10 were both significant across the different variables, although lower than the values for over45_pct, and very similar to one another. For example, for the Knn=10 weight, wpa_pct had Moran's I=0.529(z=17.525), and under10 had Moran's I=0.527(z=17.652). The unemployment rate variable generally had the lowest Moran's I and z-score values(for example, for the Knn=10 weight, it had Moran's I=0.218(z=8.266)), but it still demonstrated significant spatial autocorrelation across all the weights. From this table, the Knn=10 weight seems like the most appropriate one to use, since, with the exception of the unemployment rate, that is the one where the variables showed the highest spatial autocorrelation values, and even then, unemployment still had significant SAC. Additionally, it will likely serve as a helpful median in between the Queen 1 weight with relatively few neighbors, and the distance band that includes so many.

Table 2: Sensitivity Analysis Results

#	Weight	Variable	Min/max/avg No. Of neighbors	% non-zero	Global Moran's I	z-stat
1	Queen first order	over45_pct	1/38/5.59	2.87	0.676	15.233
2	Queen first order	uemprate	1/38/5.59	2.87	0.285	7.096
3	Queen first order	wpa_pct	1/38/5.59	2.87	0.598	13.020
4	Queen first order	under10	1/38/5.59	2.87	0.633	14.056
5	Distance band = 6km	over45_pct	1/39/24.01	12.31	0.453	19.018
6	Distance band = 6km	uemprate	1/39/24.01	12.31	0.187	8.793
7	Distance band = 6km	wpa_pct	1/39/24.01	12.31	0.360	14.719
8	Distance band = 6km	under10	1/39/24.01	12.31	0.374	15.451
9	Knn=10	over45_pct	10/10/10	5.13	0.569	19.435
10	Knn=10	uemprate	10/10/10	5.13	0.218	8.266
11	Knn=10	wpa_pct	10/10/10	5.13	0.529	17.575
12	Knn=10	under10	10/10/10	5.13	0.527	17.652

Univariate Local SAC

For all the maps below, the significance filter of .01 was used with 99999 permutations. This was done to ensure more accurate results than the .05 filter with fewer permutations, while still capturing clusters that could inform the overall spatial analysis.

Same Methods, Different Weights

For maps of the over45_pct variable, there are three LISA maps under the three different variables(Figures 8, 10, 11) to see how cluster locations vary based on weight. As expected, the weights with more neighbors capture more observations. Additionally, the Knn=10 and Distance Band=6km weights, which average more neighbors than the Queen 1 weight, include districts with negative spatial autocorrelation in certain areas. This might have been predicted from the opening hypothesis, since it was predicted that areas far from Manhattan might vary in their commute time depending on their economic well-being. Additionally, the wpa_pct and under10 variables were tested to see how they vary under different weights under the same Local Geary method, since their z-scores varied a lot from weight to weight. For both of these variables, the areas of positive SAC are largely the same under different weights (Figures 14 and 15 for wpa_pct, Figures 16 and 17 for under10), but the areas of negative SAC are different. On the one hand, the clusters of negative spatial autocorrelation might be expected given the hypothesis, but changing location of those areas could be a cause for concern of consistency.

Different Methods, Same Weights

For the maps of the over45_pct variable under different methods, but the same Knn=10 weight(Figures 8 and 9), there was significant overlap between the identified clusters of positive and negative SAC, suggesting a consistent spatial pattern, although the LISA map included more observations of significant clusters than the Local Geary. For the LISA and Local Geary maps of the unemployment rate variable(Figures 12 and 13), there is again overlap, but this time the Local Geary captures many more significant observations than the LISA map. Although both of these variables see differences under different methodologies, there are plenty of clusters that are significant regardless of method.

Figure 8: LISA map of over45_pct, Knn=10

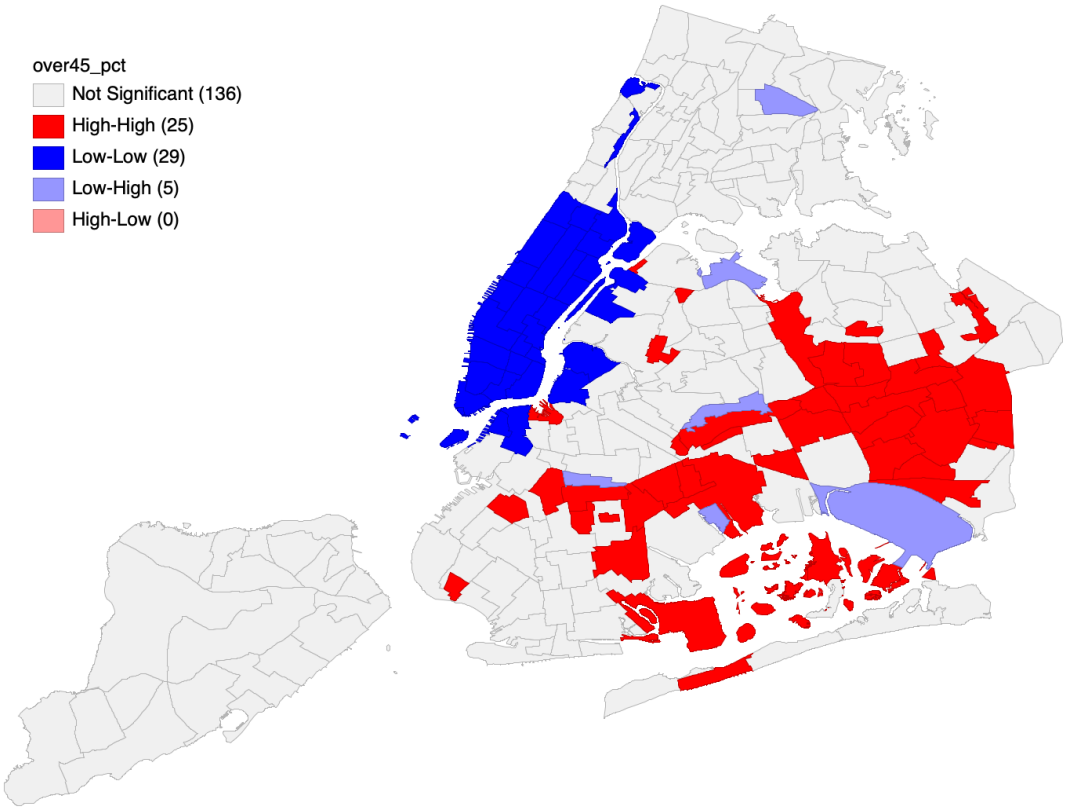


Figure 9: Local Geary map of over45_pct, Knn=10

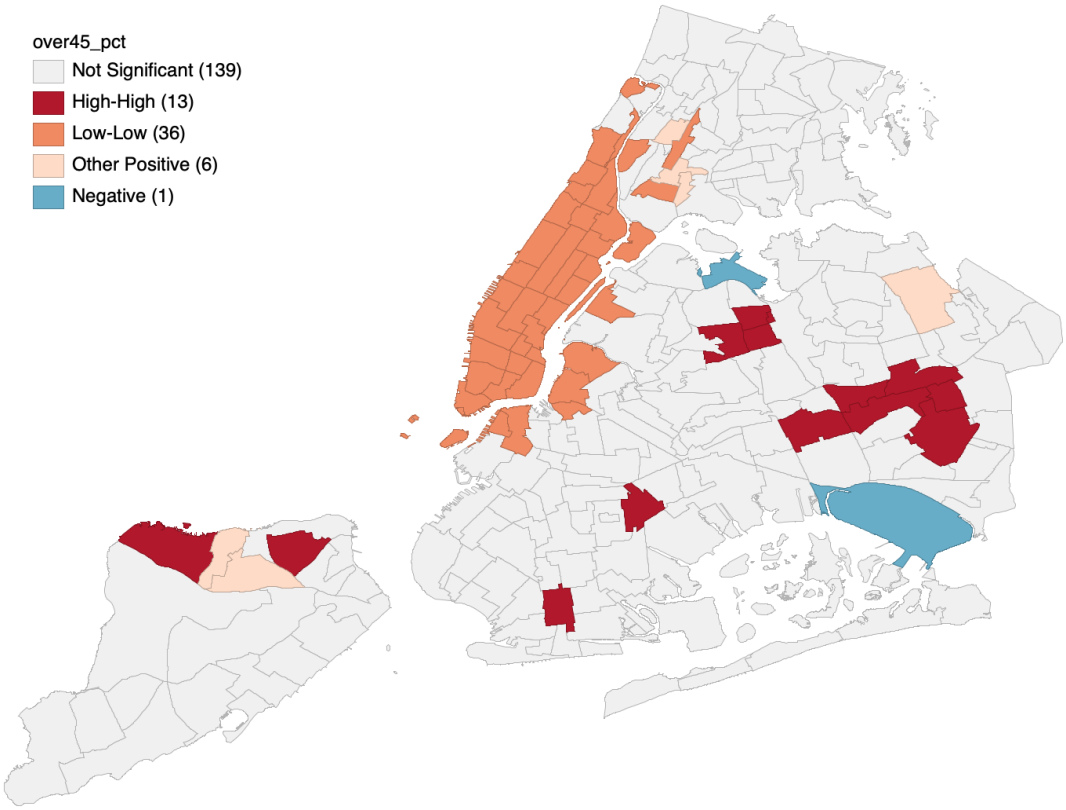


Figure 10: LISA map of over45_pct, Queen 1

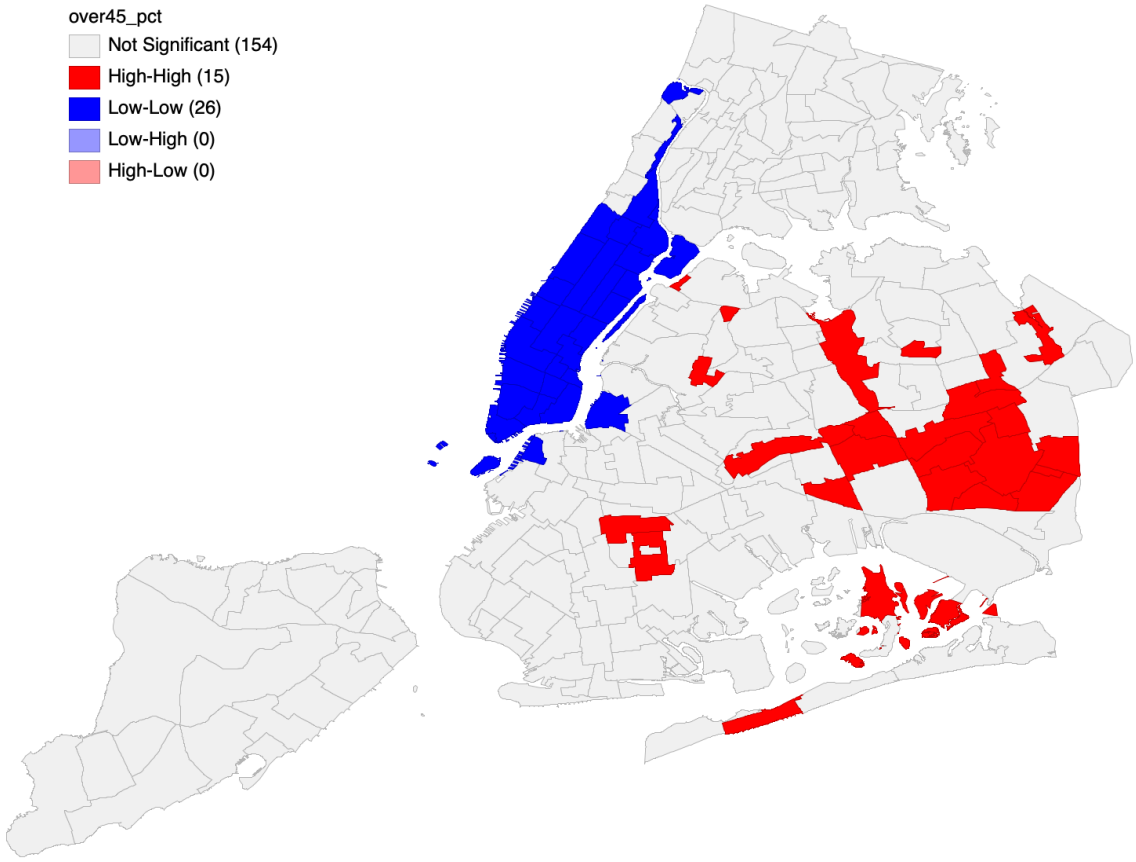


Figure 11: LISA map of over45_pct, Distance Band=6km

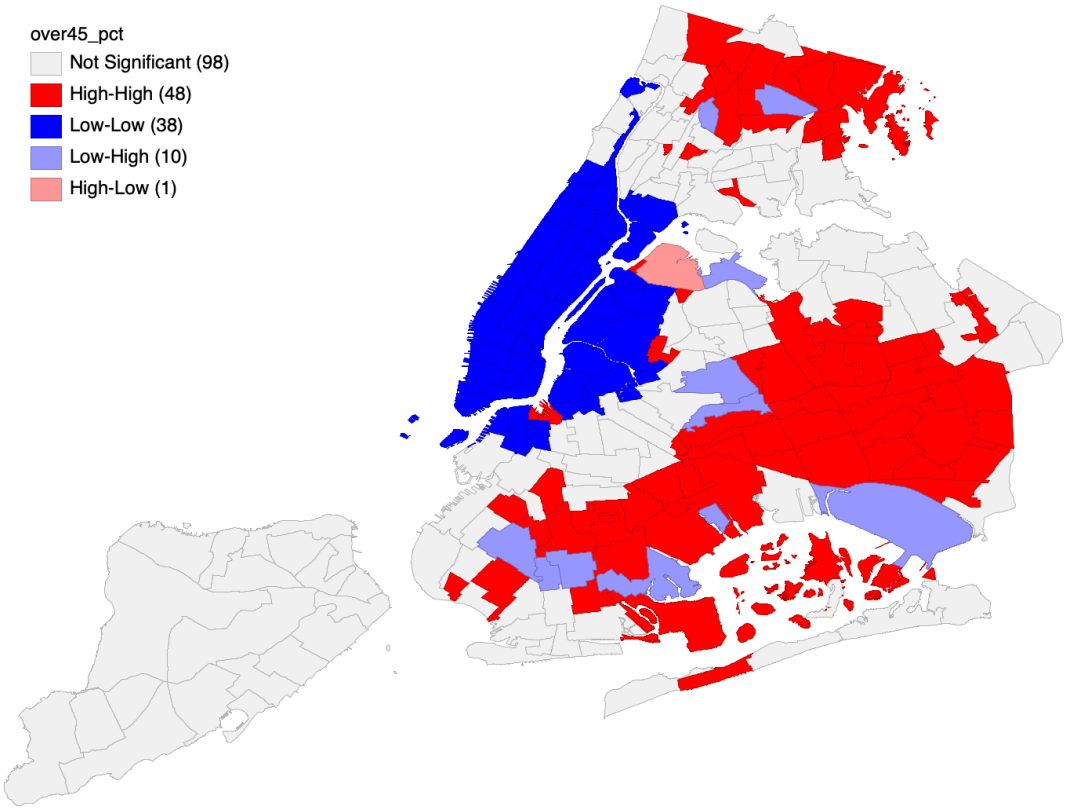


Figure 12: LISA map of uemprate, Knn=10

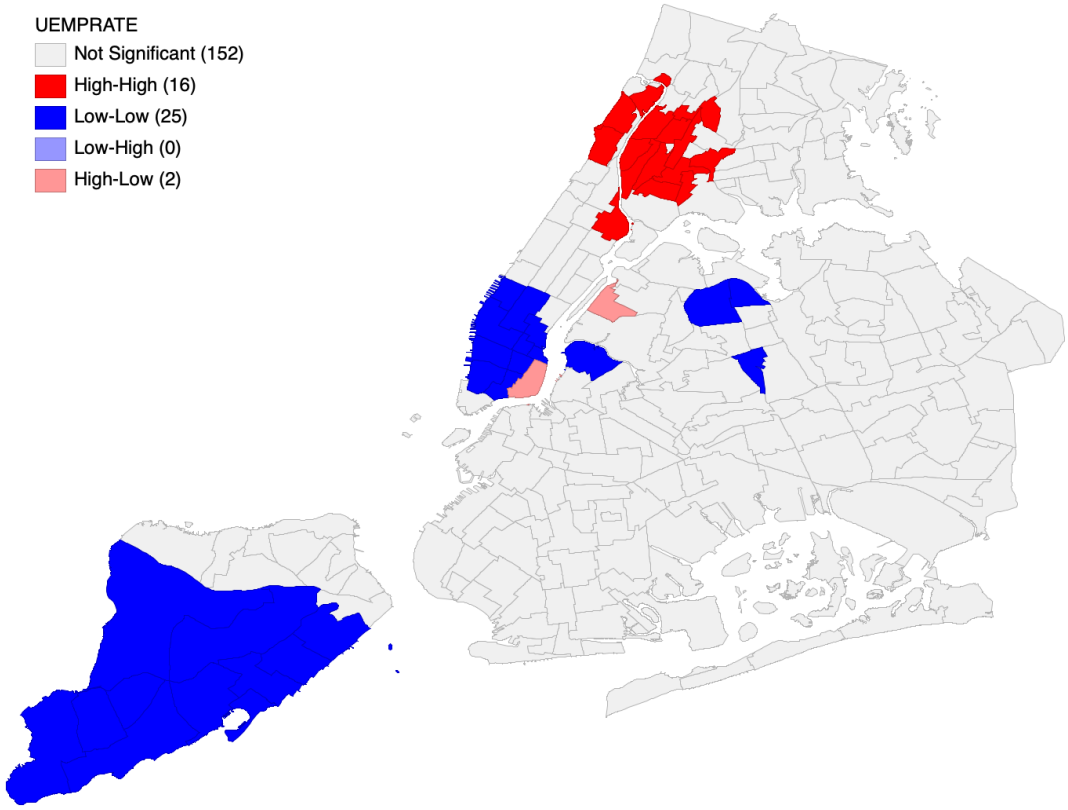


Figure 13: Local Geary map of uemprate, Knn=10

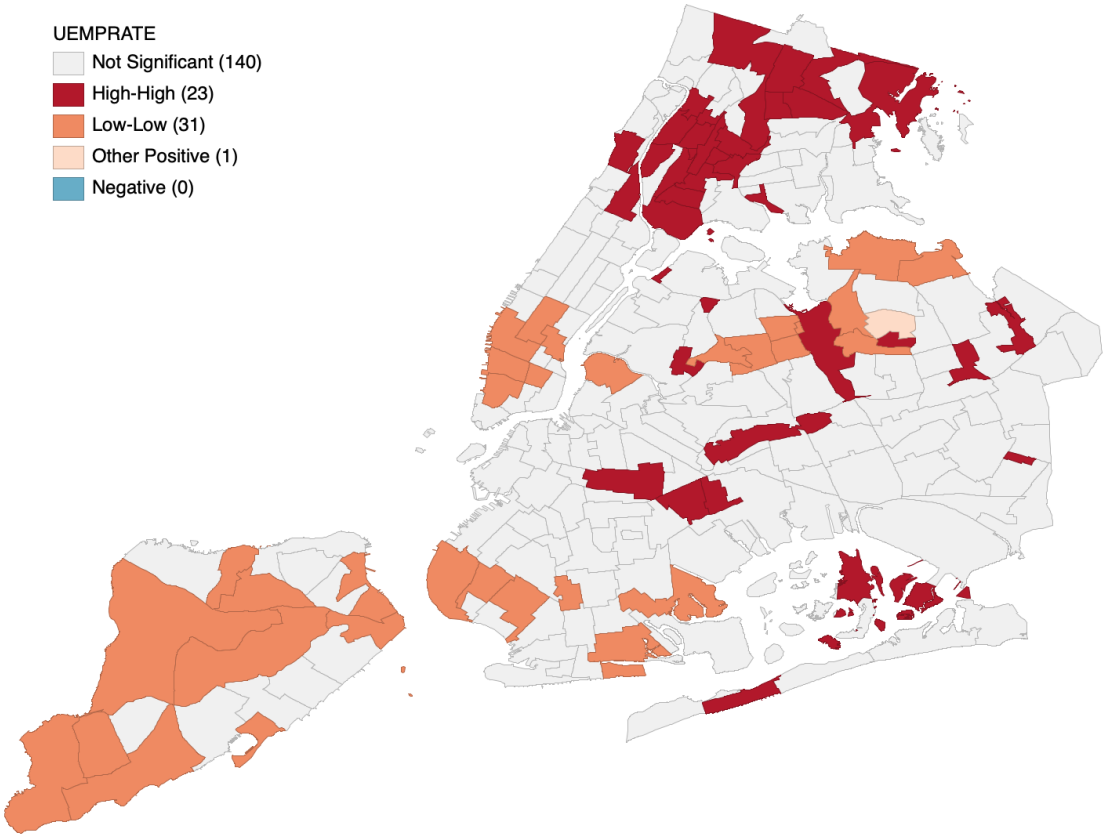


Figure 14: Local Geary map of wpa_pct, Knn=10

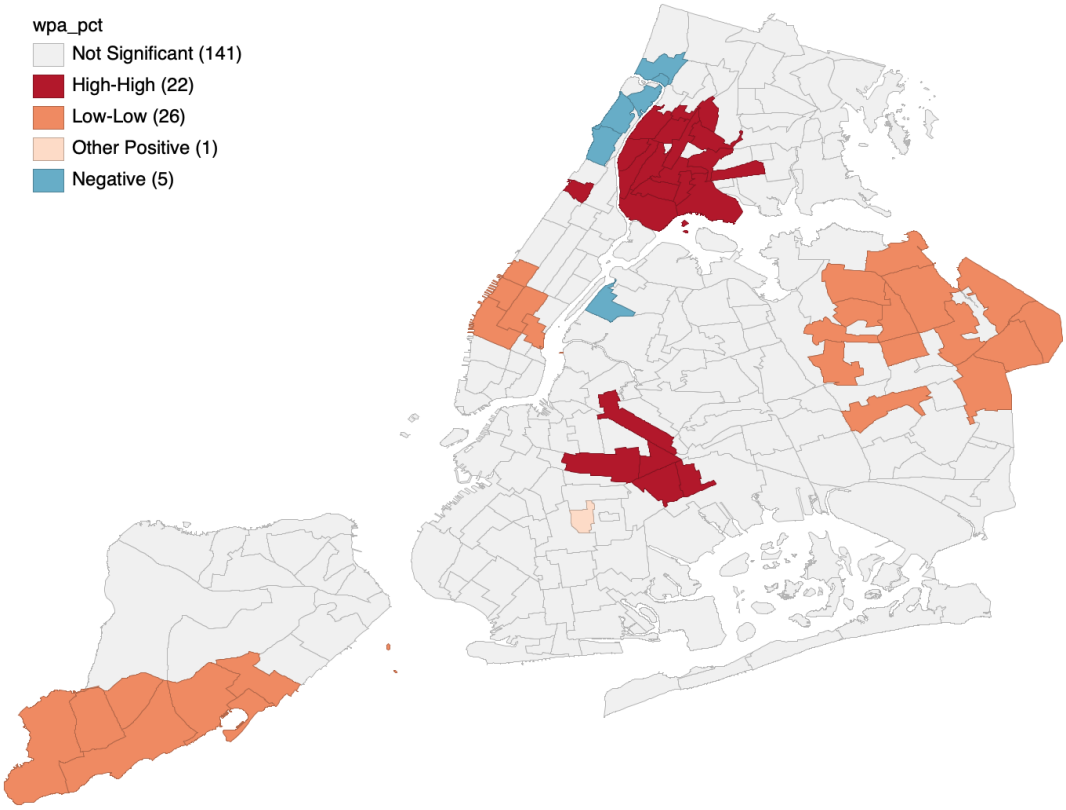


Figure 15: Local Geary map of wpa_pct, Queen 1

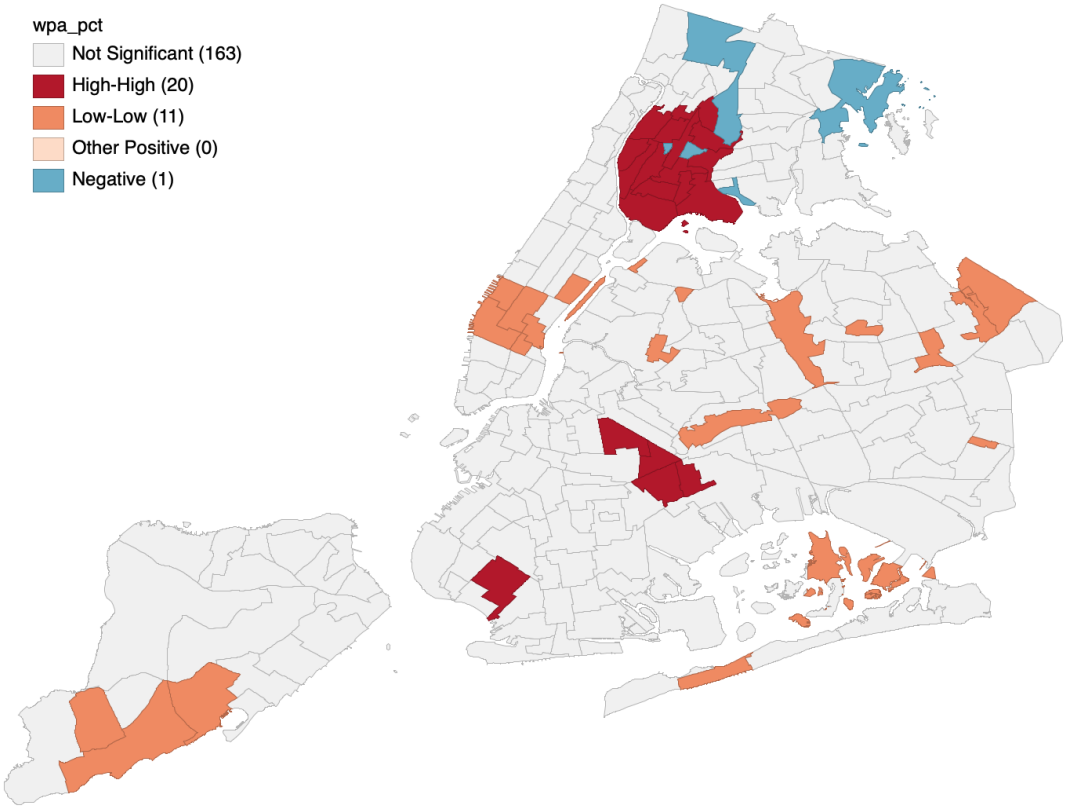


Figure 16: Local Geary map of under10, Knn=10

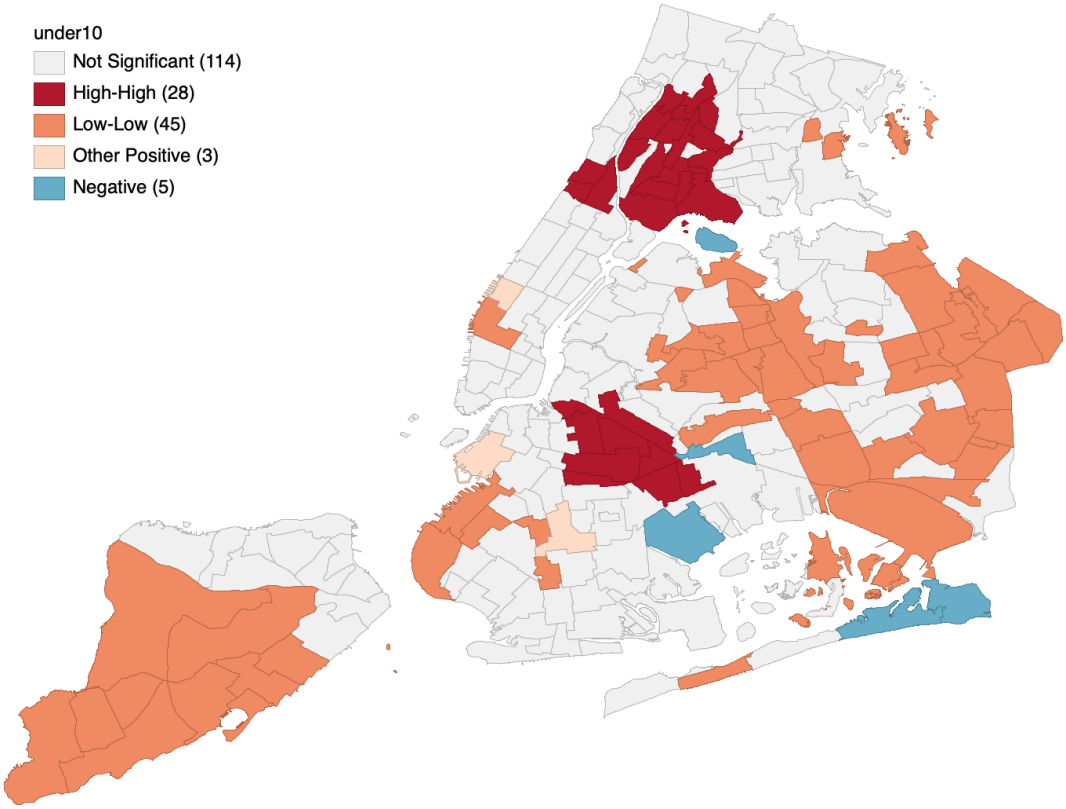
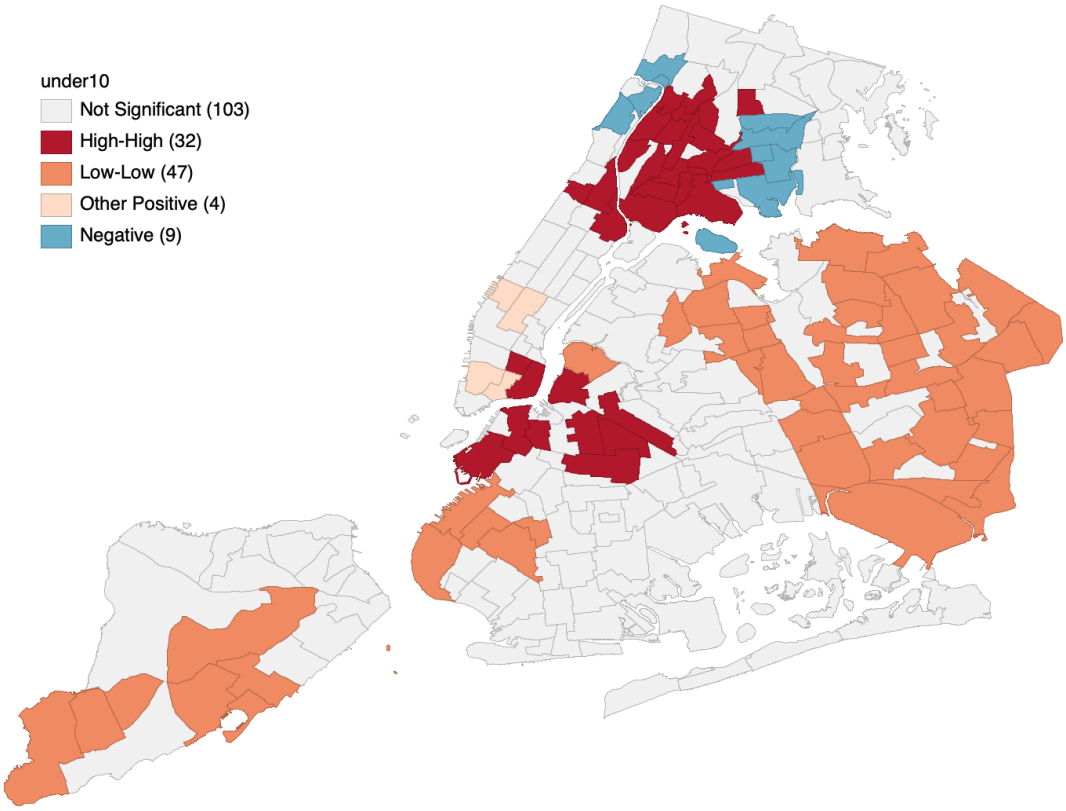


Figure 17: Local Geary map of under10, Distance Band=6km



Multivariate Local SAC(Figures 18-22)

Methods

To test multivariate local SAC among the economic variables and commute time, I thought it would be most appropriate to use a multivariate Local Geary analysis since it measures the distance between variables relative to spatial randomness, as opposed to a method like Quantile LISA, which might be used when focusing on extremes or variables without co-location. Additionally, I tested multivariate local SAC on each individual economic variable against commute time, as well as all four together, since the three economic variables were likely to be closely linked. Based on the above univariate analysis, I thought Knn=10 was the most

appropriate weight to use, since it generally had observed a number of clusters in between the Queen 1 and Distance Band weights. Finally, a significance filter of .01 with 99999 permutations was once again used in order to capture as many multivariate clusters as possible while avoiding the uncertainty of the .05 significance filter and fewer permutations.

Results

The Local Geary Map of all four variables, Figure 18, showed only instances of positive SAC between the four variables, which was expected since all the three economic variables were closely linked to one another. However, both the maps of over45_pct and wpa_pct and over45_pct and under10, Figures 19 and 20, respectively, detected clusters negatively related to each other. In both cases, there were clusters in an area where the commute time was low, but the percent of households either on public assistance income or having under $10,000 in income were high, which supports the hypothesis, although there were only single locations identified. The Local Geary of over45_pct and unemployment rate did not identify any negative clusters, but the positive clusters don't necessarily suggest more positive correlation between the two variables then in general. For example, Figure 21 shows the map brushed with a scatter plot of the two variables of a cluster of areas identified as having positive SAC between the two variables. However, in that cluster, which is in an area in Queens that had low unemployment, commute time and unemployment were more negatively correlated than on average. Overall, the bivariate maps provide some support for the hypothesis, but more analysis is needed to determine if the economic variables are generally positively correlated with commute time, or if there is more variation among wealthier neighborhoods outside of Manhattan.

Figure 18: Local Geary Map of over45_pct, wpa_pct, uemprate, under10, Knn=10

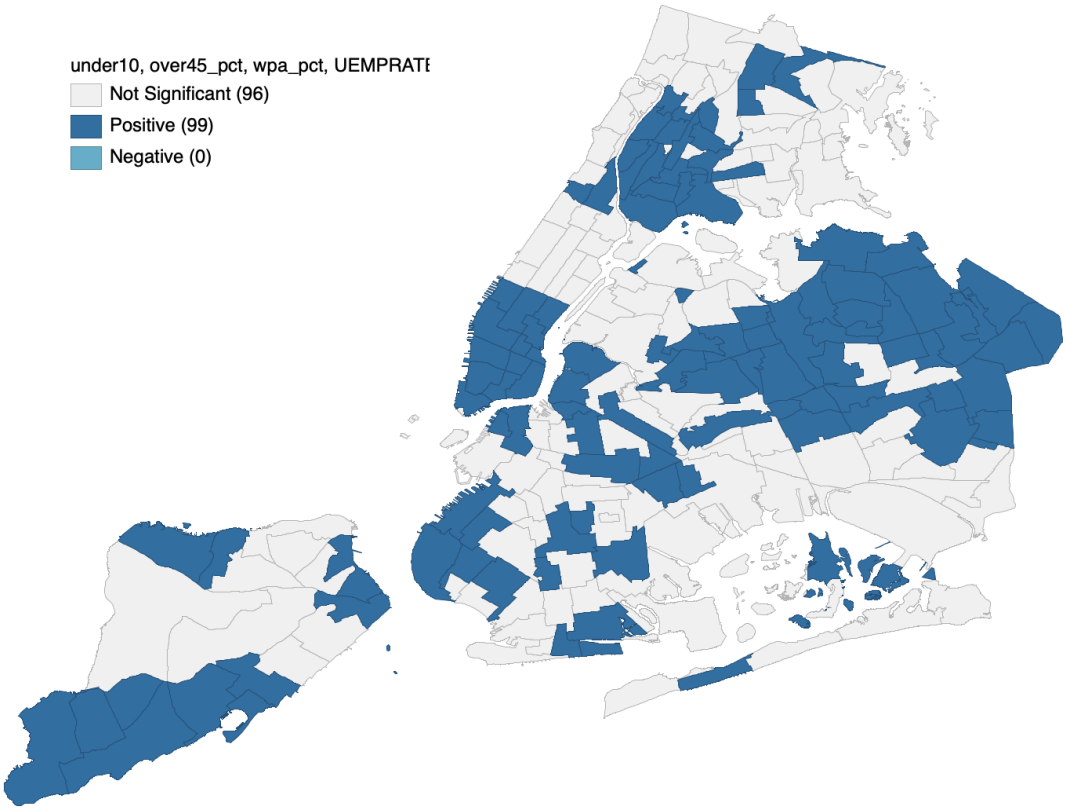


Figure 19: Local Geary Map of over45_pct, wpa_pct, Knn=10

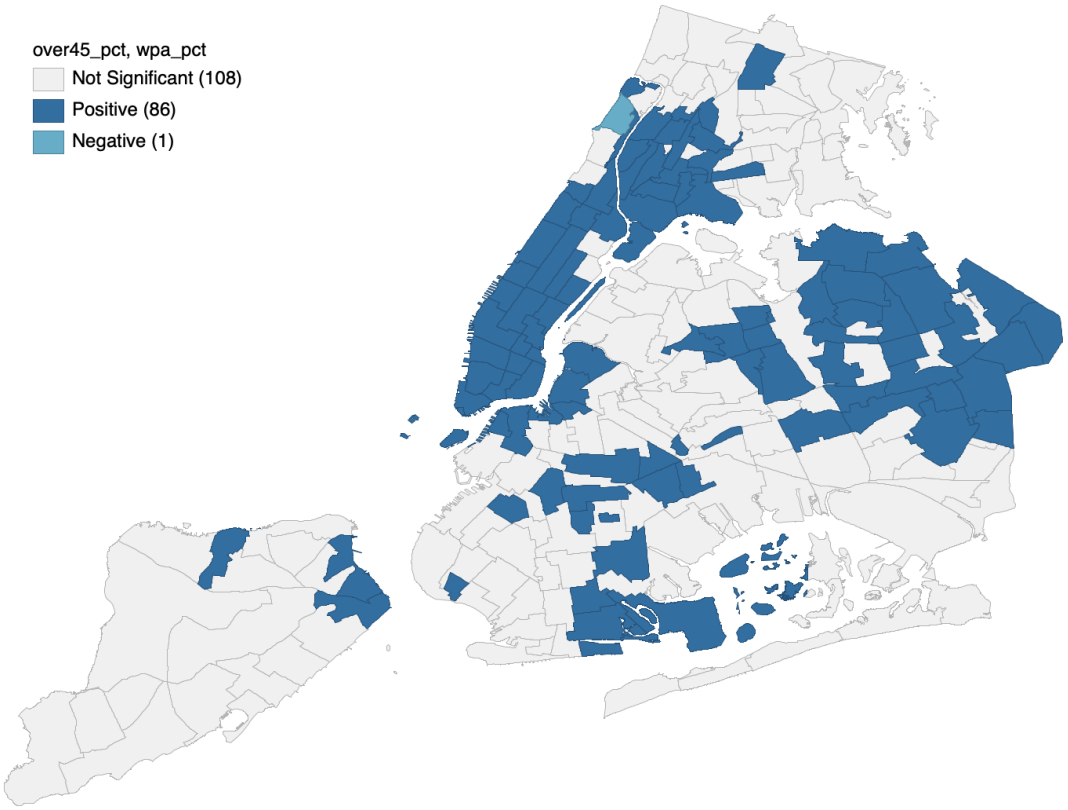


Figure 20: Local Geary Map of over45_pct, under10, Knn=10

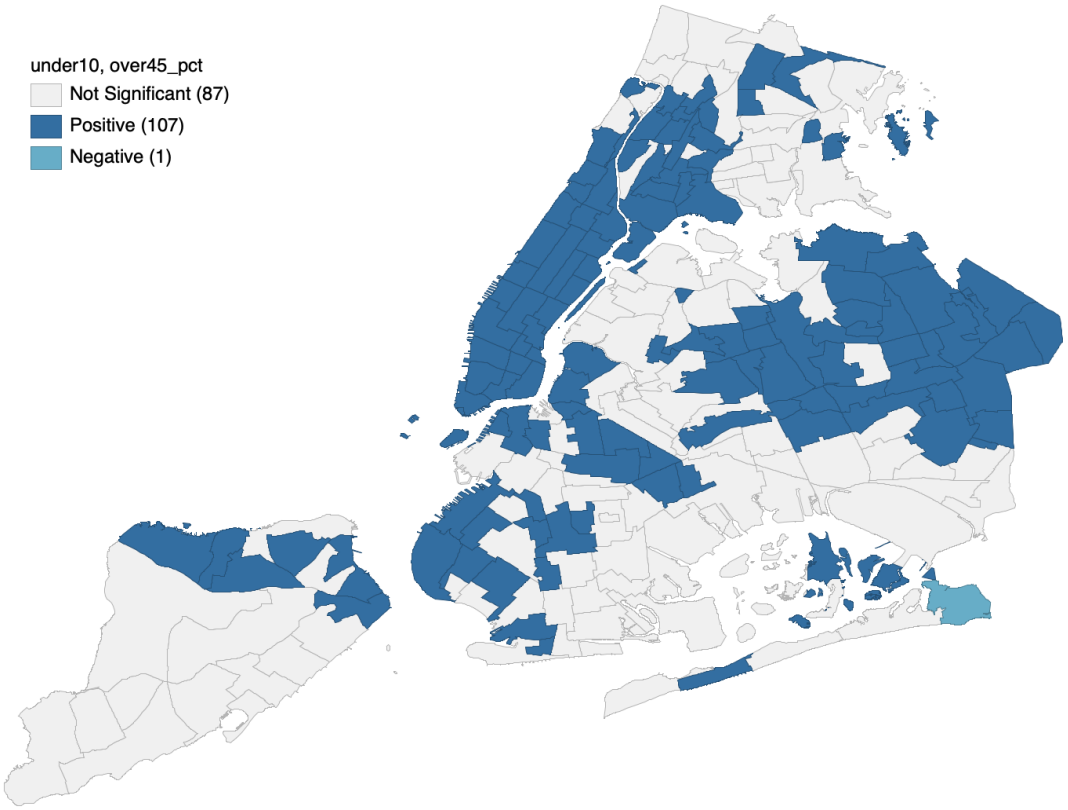
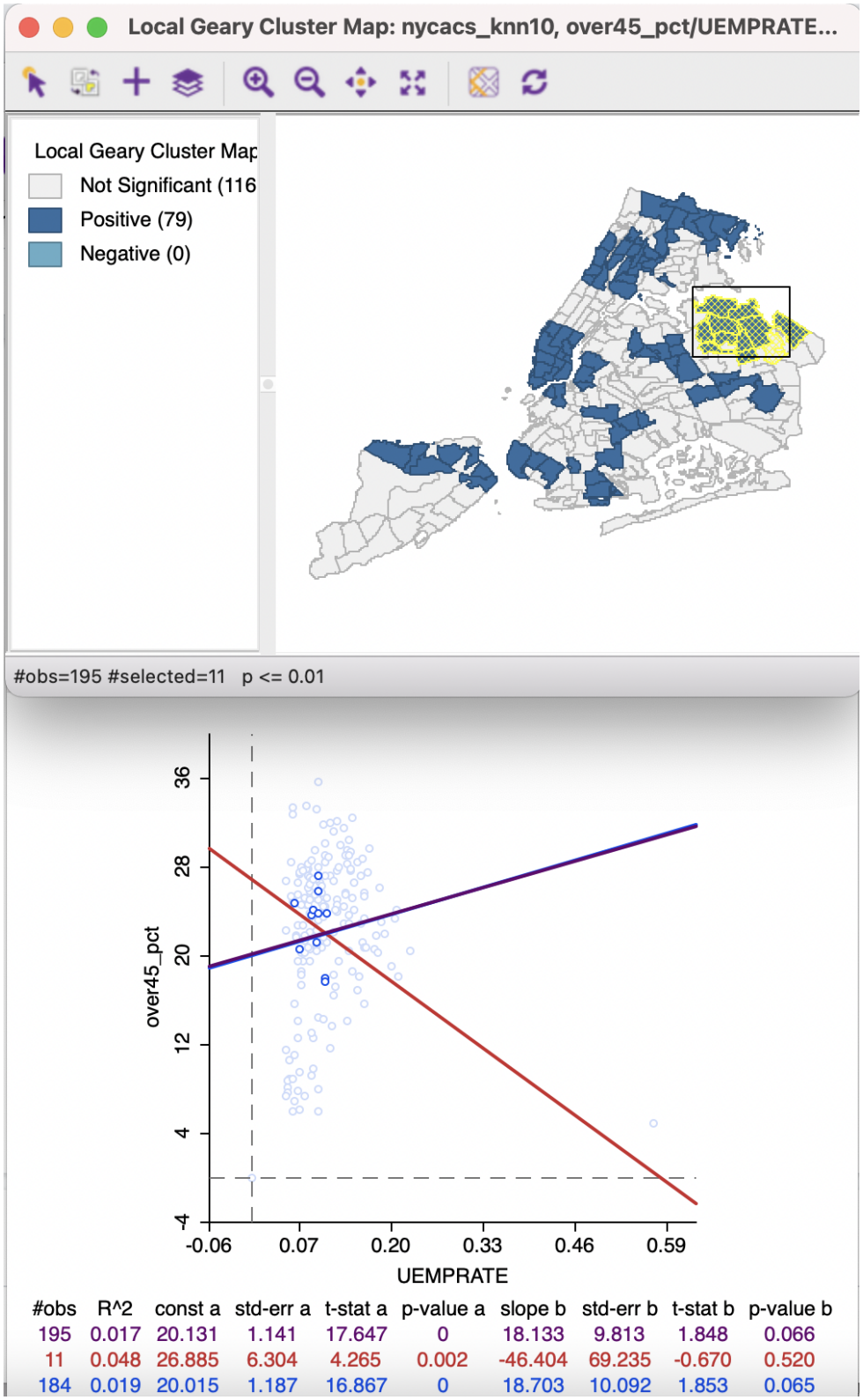


Figure 21: Local Geary Map of over45_pct, uemprate, Knn=10, brushed with scatter plot



#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
195	0.017	20.131	1.141	17.647	0	18.133	9.813	1.848	0.066
11	0.048	26.885	6.304	4.265	0.002	-46.404	69.235	-0.670	0.520
184	0.019	20.015	1.187	16.867	0	18.703	10.092	1.853	0.065

Cluster Analysis (Figures 23-24)

Methods

In order to determine group similarity among the ACS districts in NYC and see how clusters of districts are similar to one another, a K-means clustering analysis was performed on the data for all four of the variables in question. Six clusters were used as a way to capture the diverse array of areas in the city and to create distinct clusters. Unfortunately, there was one isolated area, which mapped to large park areas in NYC, such as Central Park and Brooklyn Bridge Park, but it was unlikely to be grouped even if there were fewer clusters.

Results

The K Means Cluster Map, Figure 23, shows five distinct types of areas, as the sixth one is really an outlier observation not representing many actual people. Manhattan and nearby areas are one cluster, with a couple exceptions included in it in east Queens and Staten Island. The south Bronx is another largely contiguous cluster, with a few exceptions in Brooklyn. The upper half of Manhattan and the middle parts of Brooklyn are part of another cluster. The other clusters have pockets of contiguity, but overall are not quite as contiguous.

Figure 24 displays the statistics of the K Means Cluster Map. The ratio of the sum of squares between clusters is 0.78177, meaning the different clusters are significantly distinct from one another. Additionally, it shows that while cluster 4, which is largely the lower half Manhattan, has low commute times and high economic well-being, cluster 2, which includes many areas towards the outer edges of the city, has the inverse relationship, providing support for the hypothesis that areas outside of lower Manhattan have the reverse relationship between commute time and economic well-being. Not only that, clusters 3 and 5, which are on the lower end of commutes outside of Manhattan, have higher values on the economic variables, which are inversely related to economic well-being.

Figure 23: K Means Cluster Map

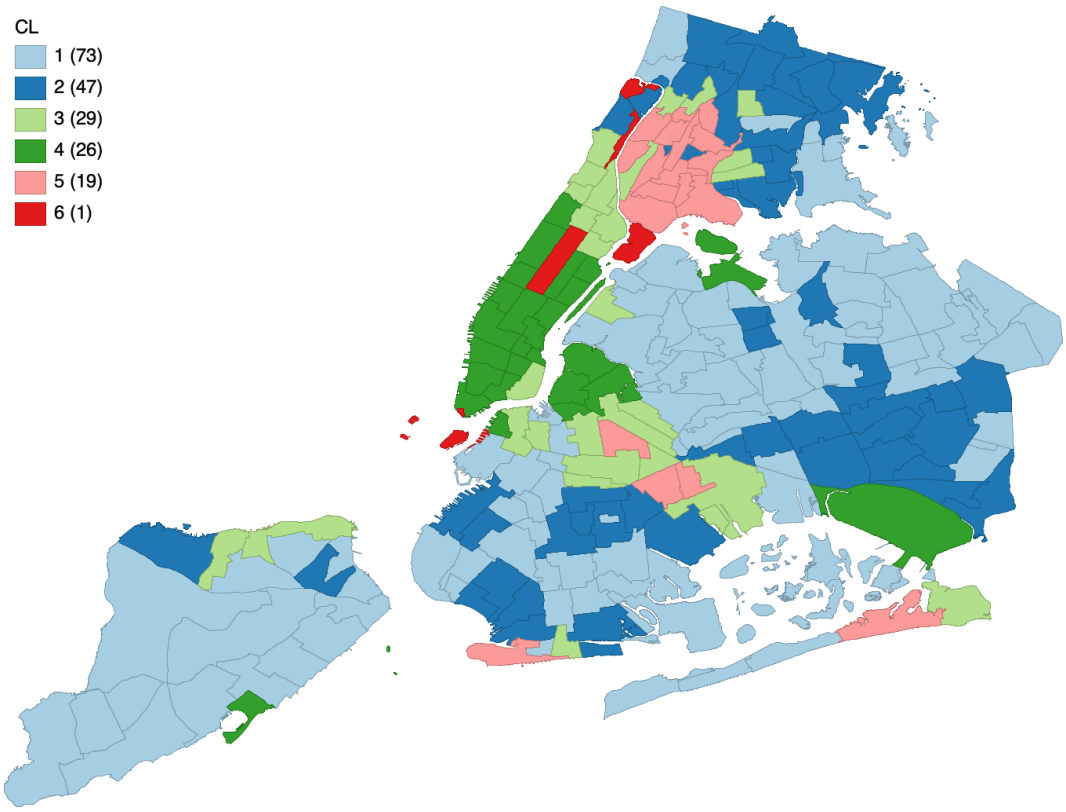


Figure 24: K Means Cluster Statistics

```
------
Method: KMeans
Number of clusters: 6
Initialization method:  KMeans++
Initialization re-runs: 150
Maximum iterations: 1000
Transformation: Standardize (Z)
Distance function:  Euclidean

Cluster centers:
|   |under10|over45_pct|wpa_pct|UEMPRATE  |
|---|-------|----------|-------|----------|
|C1 |2.83656|23.785    |2.17886|0.0812702 |
|C2 |4.06709|28.2004   |4.49103|0.115788  |
|C3 |8.39873|20.6189   |6.78162|0.123641  |
|C4 |4.04931|8.21199   |1.64622|0.0580641 |
|C5 |11.0415|21.9159   |11.5398|0.17369   |
|C6 |0      |4.82337   |0      |0.569811  |

The total sum of squares:  776
Within-cluster sum of squares:
|   |Within cluster S.S.|
|---|-------------------|
|C1 |42.3453            |
|C2 |40.4937            |
|C3 |39.0008            |
|C4 |29.818             |
|C5 |17.6888            |
|C6 |0                  |

The total within-cluster sum of squares:   169.346
The between-cluster sum of squares:    606.654
The ratio of between to total sum of squares:  0.78177
```

Discussion

The univariate analysis provided some support for the hypothesis that the variables indicating lack of economic well-being would have a positive relationship with commute in and near lower Manhattan, but a negative relationship far away from it. While commute times increased outside of Manhattan, it was not uniform and there were lots of variations across the boroughs. Additionally, many of the areas with the highest commute times had lower than average values of the economic variables. Not only that, the local SAC measurements indicated areas far from Manhattan with negative SAC, indicating areas where commutes could be high with economic variables low, since they were near places where those values were higher. The bivariate analysis was mixed, in that it shows clusters where commute time and the economic

variables had a positive relationship even outside of Manhattan, but it identified negative relationships as well. The cluster analysis gave strong support for the hypothesis, as it identified many similar areas outside of Manhattan that had an inverse relationship between commute time and economic well-being.

Interpretation

There is evidence to support the hypothesis that, outside of lower Manhattan, commute time and economic well-being are inversely related, but there is some doubt to fully reject the null hypothesis, particularly in the analysis of bivariate local SAC. Nevertheless, the univariate analysis, through the identification of areas with high commutes and high economic well-being on the edges of the city, and the cluster analysis, with its identification of clusters of areas with the relationship, supported it. Additionally, the bivariate analysis did not necessarily go against it, since some of the positive clusters actually showed that commute time and the negative economic variables were negatively correlated in those clusters. There are multiple ways New York City could respond to this analysis. It could decide to invest more in public transit from the outer reaches of New York City into Manhattan, since it seems like in areas where people are making that longer commute, they have lower unemployment, higher incomes, and are less likely to be on public assistance income. Helping more people to commute could help those poorer neighborhoods improve economically. After all, residents of the wealthier outer borough areas may be able to make the commute into Manhattan because they are more likely to be able to afford cars. This investment in public transit could include increased bus and train routes into Manhattan, as well as public transport across Brooklyn and Queens that could make the commute more accessible. Another way New York City officials could respond to these findings would be to increase investment in businesses outside of Manhattan, so that people in the outer boroughs don't have to commute in order to be economically well-off. Then maybe lower commutes could be a positive thing throughout the city, instead of just within Manhattan. Although the data for this analysis is a little dated, the relationship between commute and economic well-being for areas far from lower Manhattan may very well still be in place and will still likely require increased investment in areas where making a long trip to work is a significant factor in bettering oneself financially.