METHOD AND RESULT

How the data was collected:
Three datasets were chosen to include in the project:

First of all, the crime data from Chicago Data Portal is selected (link: https://data.cityofchicago.org/P-ublic-Safety/Crimes-2001-to-present/ijzp-q8t2). It reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system.

Secondly, the Public health and socio-economic indicators dataset is selected, which is imported from the Geoda Center, which is the Public health and socio-economic indicators for the 77 community areas of Chicago, IL, 2014. (link: https://geodacenter.github.io/data-and-lab//comarea_vars/).

Finally, we selected the Chicago community areas data set from Chicago Data Portal. The dataset was imported from Chicago Data Portal. It is a boundary file for the community areas. (link: https://data.cityofchicago.org/Facilities-GeographicBoundaries/Boundaries-Community-Areas-current-/cauq-8yn6).

Research Questions
We are targeting the following research question(s):

- **Q1:** Is there a statistical relationship between crime rate versus education level, unemployment rate, and income level?
- **Q2:** Is there a spatial correlation with respect to the variables of interest?

Data and Variables of Interest
After the datasets have been imported, they have been modified by aggregating and rearranging the variables; community ID, community name, number of incidents of crime per thousand capita in 2014 (calculated by the count of crime divided by population of community, then times one thousand), number of people living in poverty per thousand capita in 2014, number of people unemployed per thousand capita in 2014, number of adults without a high school diploma in 2014, and per capita income in 2014, which we named ComAreaID, community, CRIMP1000, POVP1000, UNEMPP1000, NOHSP1000, PerCInc14, respectively (For more detailed description of variables, refer to the table below). According to our interest, the explanatory variables used in our analysis were no high school diploma rate, unemployment rate, income per capita, and poverty rate, and crime rate were used as the response variables. Table 1 provides the name and a brief description of each variable that were created, including the associated levels and the type and usage of the variables in the future analysis.

Variable Attributes

Variable	Description	Levels	Comments
ComAreaID	The ID number of the community in Chicago	No.1-77	ID variable
community	The community name in Chicago	text name	ID variable
CRIMP1000	Number of incidents of crime per thousand capita in 2014	[22, 366]	Response variable
POVP1000	Number of people living in poverty per thousand capita in 2014	[24, 628]	Explanatory variable
UNEMPP1000	Number of people unemployed per thousand capita in 2014	[28, 153]	Explanatory variable
NOHSP1000	Number of adults without a high school diploma in 2014	[13, 273]	Explanatory variable
PerCInc14	Per capita income in 2014	[14905, 2075859]	Explanatory variable

TABLE 1. The table includes the name, description, levels, comments for each variable. Explanatory variable, response variable, and variable type are noted in the comments.

Exploratory Data Analysis (EDA)

The data was reviewed and modified prior to the spatial and statistical analysis followed by the procedure stated in the former section. There were no notable errors or missing values. We looked at the descriptive statistics of some variables including their histograms, shown in Table 2 (refer Appendix A for the histograms).

Descriptive Statistics for Variables with respect to the Whole Sample

Variable	N	Mean	Standard Deviation	Minimum	Maximum
CRIMP1000	77	109.811	75.4334	22.014	365.317
POVP1000	77	233.46	122.773	24.470	627.677
UMEMPP1000	77	74.5495	31.1279	28.6728	152.69
NOHSP1000	77	120.104	63.3869	13.2498	272.405
PerCInc14	77	289557	360033	14905	2075860

TABLE 2. The table displays some variables with respect to the whole sample that were meaningful to include. There were a total of 77 observations across the study. The table also displays additional information pertaining to the mean, standard deviation, maximum and minimum for each variable.

As we can see from the table 2 and the histograms in Appendix A, the distribution of all the five variables are right skewed, since there is a lower bound for all the variables, which is zero.

Then, some quantile maps were created in order to see if there exists some potential spatial patterns in the dataset (refer the maps in Appendix B). As we can see from the maps, people in South part of Chicago tend to have higher crime rates, higher unemployment rate, lower income, and lower education level, while people in North part of Chicago including Loop area tend to have lower crime rates, lower unemployment rate, higher income, and higher education level. In addition, as we can see from the crime histogram, most of the communities have a value less than 90.7. Then, we selected the bins that are higher than 90.7. As we can see from the equal intervals map, the distribution of the communities are in the South part of Chicago, downtown and the communities on its West (refer to Appendix B for the histogram and map). Thus, there is potentially some spatial correlations to these variables that we can investigate later on.

Then, we drew the scatterplot matrix among unemployment rate, high school education rate, income per capita, poverty rate, and crime rate (refer the scatterplot matrix in Appendix C). As we can see from the scatterplot matrix, crime rate has a positive relationship with unemployment rate and poverty rate, and a negative relationship with income level. And the LOWESS smoother does not imply any structural break with respect to crime versus other variables except crime versus income level. In addition, we can see that after we selected the higher income per capita bins in the income histogram, the communities are all located in the North part of Chicago. They are all with relatively low crime rate (lower than 154.6). In addition, the points shown in the scatterplot can be potential outliers.

To check the high income per capita communities, we draw the box plot of Income per capita and the box map of it with hinge = 1.5 (as shown in Appendix D). As we can see from the box plot and box map, there is no lower outlier, but there are some upper outliers, which are exactly the communities to the North of downtown Chicago.

Then, we plot the parallel coordinate plot (as shown in Appendix D). We select the observations with highest income per capita. We find that it is the four communities located to the North of downtown Chicago, and it makes sense in the plot that high income per capita implies a low poverty rate and a low crime rate.

In addition, we draw the conditional map (as shown in Appendix D), since income level and poverty rate are variables with same idea of representation of the population. The horizontal axis is the crime per thousand capita; the vertical axis is the income per capita; and the base box map is based on the people living in poverty per thousand capita. We can see from the map on the first row and first column that communities with higher income per capita and lower poverty rate tend to have a low crime rate. Similarly, as we can see from the maps on the second row and third column and on the third row and third column, communities

with lower income per capita and higher poverty rate tend to have a higher crime rate.

Finally, we created the rook contiguity weight for the community polygon map (as shown in Appendix D). As we can see from the weight table and the histogram, the distribution of the number of neighbors of community is approximately normal.

Statistical and Spatial Analysis
First, we specifically look at the scatterplot of crime per thousand capita versus the other four explanatory variables (as shown in Appendix E). As we can see from the scatterplots, the p-value for the linear fit of the 77 data points is less than 0.05 for scatterplots of crime rate versus unemployment rate and crime rate versus poverty rate, which are significant. On the other hand, the scatterplots of crime rate versus no high school diploma rate and crime rate versus income level are not significant, which means crime rate is not significantly correlated with education level and income level. In addition, we found that if we select all the north communities, we can see that the Chow test is significant for all the scatterplots, which means the slopes we select and the slopes of the rest of the map are significantly different to each other, suggesting spatial heterogeneity between North part of Chicago and the rest. For the two significant scatterplots, we can see that the slopes are both positive, which implies that crime rate and unemployment rate are positively correlated, as well as crime rate and poverty rate. Thus, the higher poverty rate and unemployment rate will have a higher crime rate.

After fitted the scatterplots and did the summary, Table 3-6 shown below had been generated:

Statistical Summary for Poverty Rate

Parameter	**Coefficient**	**Standard Error**
Intercept	2.770	12.440
POVP1000	0.458	0.047

TABLE 3. The table displays the coefficient and the standard error for each variable.

Statistical Summary for Unemployment Rate

Parameter	**Coefficient**	**Standard Error**
Intercept	-7.794	17.144
UNEMPP1000	1.578	0.212

TABLE 4. The table displays the coefficient and the standard error for each variable.

Statistical Summary for No High School Diploma Rate

Parameter	**Coefficient**	**Standard Error**
Intercept	101.230	18.602

POVP1000	0.071	0.137

TABLE 5. The table displays the coefficient and the standard error for each variable.

Statistical Summary for Income Level

Parameter	**Coefficient**	**Standard Error**
Intercept	121.213	10.935
POVP1000	-0.000	0.000

TABLE 6. The table displays the coefficient and the standard error for each variable.

Then, we draw the Moran's I scatterplots of the five variables including crime rate, no high school diploma rate, income level, unemployment rate, and poverty rate with respect to Rook weight (as shown in Appendix F). Among the five Moran scatterplot, the slopes of all the scatterplots are significant, since the p-value is less than 0.05, which implies that for all the five variables, there exists significant spatial correlation for the variables, since from the simulation, we can see that the p-values are all 0.001. In addition, we can see from the Moran Scatterplots that there is a potential structural break in the map between north and the rest, since the p-value for Chow test is less than 0.05 for all Moran scatterplots after we select the north part of Chicago.

Next, we drew the cluster map and significance map of the five variables with respect to Rook weight. We chose the p-value cutoff to be 0.05, and the number of permutation to be 999 (as shown in Appendix G). As we can see from the maps, for crime rate, the communities on the North seem to have a lower crime rate cluster as shown in blue color to be the core of the cluster. In addition, the middle South and West communities seem to have a high crime rate cluster with red colors as the core of the cluster. Moreover, we can see that there is no low-high or high-low communities in the map, which implies that the communities' crime rates are positively correlated. For poverty rate, the communities on the North seem to have a lower poverty rate cluster as shown in blue color to be the core of the cluster. In addition, the middle South communities seem to have a high poverty rate cluster with red colors as the core of the cluster. Moreover, we can see that there is just two low-high or high-low communities in the map, which implies that the communities' poverty rates are mostly positively correlated. For no high school diploma rate, some north communities and two communities on south seem to have a lower no high school diploma rate cluster as shown in blue color to be the core of the cluster. In addition, the West communities seem to have a high no high school diploma rate cluster with red colors as the core of the cluster. Moreover, we can see that there is no low-high or high-low communities in the map, which implies that the communities' no high school diploma rates are positively correlated. For unemployment rate, the communities on the North seem to have a lower unemployment rate cluster as shown in blue color to be the core of the cluster. In addition, the

South communities seem to have a high unemployment rate cluster with red colors as the core of the cluster. Moreover, we can see that there are only two low-high or high-low communities in the map, which implies that the communities' unemployment rates are positively correlated. Finally, for income level, the communities on the North seem to have a higher income level cluster as shown in red color to be the core of the cluster. In addition, the South and West Loop communities seem to have a lower income level cluster with blue colors as the core of the cluster. Moreover, we can see that there are only one low-high or high-low communities in the map, which implies that the communities' income level are positively correlated.

In order to find spatial clusters, we also tried Principle Component Analysis on the five variables. I set the transformation option to be Standardize. As we can see from the output, the 5 variables are reduced to 3 dimensions based on 95% threshold criterion, i.e. the first three principle explain 90.56% of the total variance of the original variables. According to the squared correlations table (as shown in Appendix H), the first principle component (PC1) is mainly made up of the 3 variables including crime rate, poverty rate, and unemployment rate. The second principle component (PC2) is mainly explained by three variables including no high school diploma rate, crime rate, and income level. The third component (PC3) is explained mainly by the no high school diploma rate and income level.

Then, we use the max-p regionalization with respect rook contiguity weight. In order to have not too many cluster, we set the minimum bound to be 17% of the 2014 population so that the number of clusters is appeared to be 5 (as shown in Appendix H). We set the iteration number to 1000. As we can see from the cluster map, south, west, loop, and north have been clustered roughly, which is identical to our prior analysis.

Conclusion

Our first research question was: Is there a statistical relationship between crime rate versus education level, unemployment rate, and income level?

The answer is poverty rate and unemployment rate are statistically related to crime rate, since from the analysis in section 3, we found that the p-values for the slope of the scatterplot with unemployment rate versus crime rate and that of the scatterplot with poverty rate versus crime rate are both significant. In addition, they are all positively correlated, which means a higher unemployment rate and higher poverty rate will make the crime rate higher.

Our first research question was: Is there a spatial correlation with respect to the variables of interest?

The answer is yes, since as we can see from the Moran's I scatterplots, the p-values are all less than 0.05 for all the five variables. In addition, there is a structural break between north part of Chicago and the rest. Moreover, the spatial clusters are clustered roughly by north, south, loop and west. Also, we can see that the neighbors of Chicago are influencing each other positively, which also implies similarity among neighbor communities.

Appendix A. Histograms of the Variables

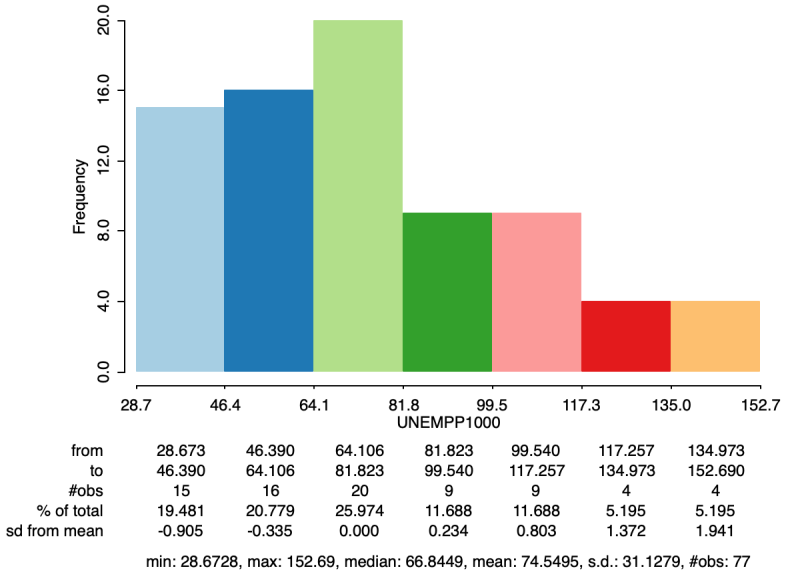
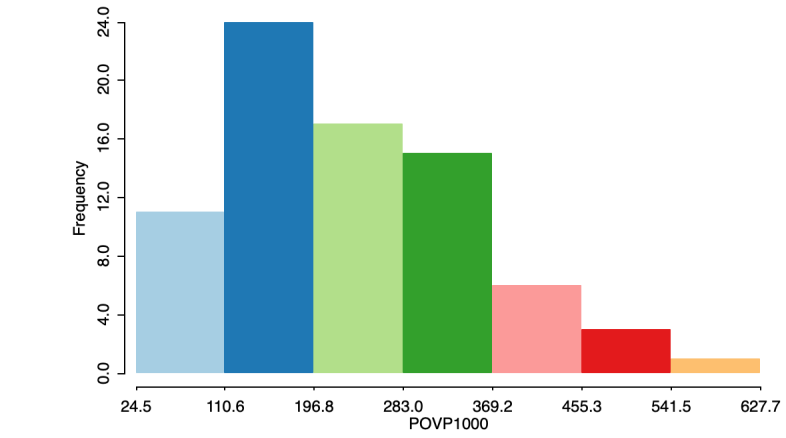


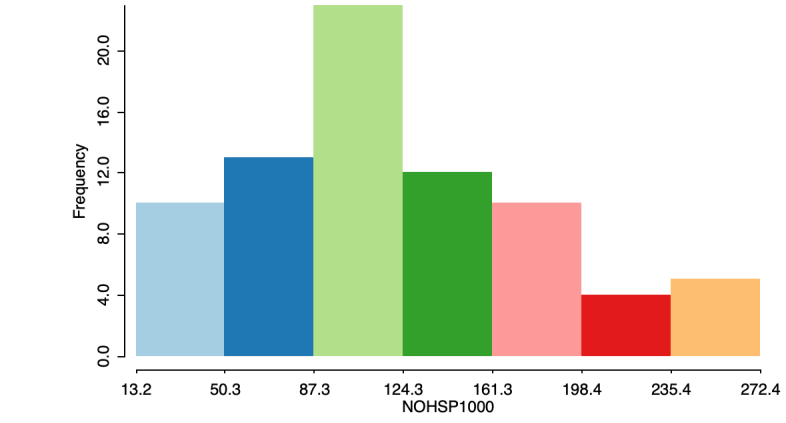
Figure 1. Histogram of Number of Unemployment Population per thousand capita in 2014.



from	24.470	110.643	196.815	282.987	369.160	455.332	541.505
to	110.643	196.815	282.987	369.160	455.332	541.505	627.677
#obs	11	24	17	15	6	3	1
% of total	14.286	31.169	22.078	19.481	7.792	3.896	1.299
sd from mean	-1.000	-0.298	0.000	0.403	1.105	1.807	2.509

min: 24.4701, max: 627.677, median: 209.526, mean: 233.46, s.d.: 122.773, #obs: 77

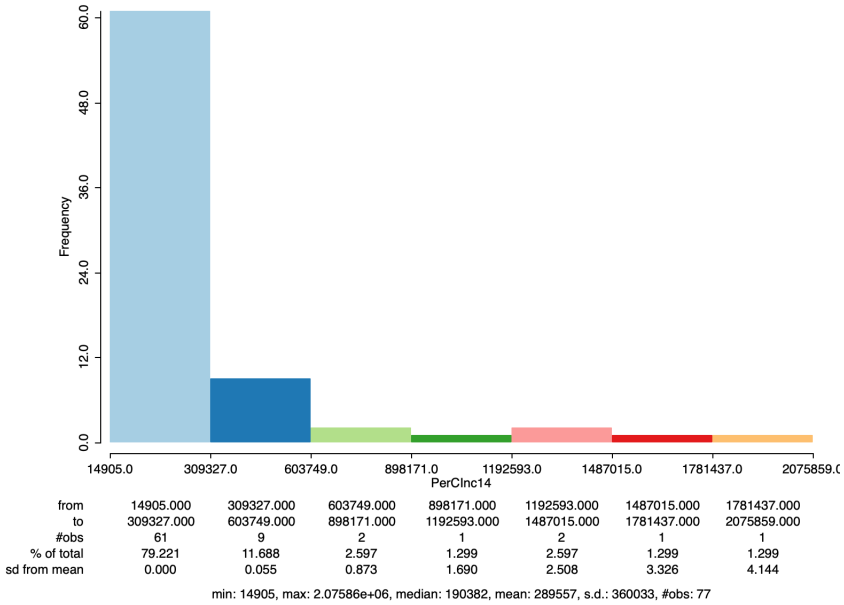
FIGURE 2. Histogram of Number of People Living in Poverty per thousand capita in 2014.



from	13.250	50.272	87.294	124.316	161.339	198.361	235.383
to	50.272	87.294	124.316	161.339	198.361	235.383	272.405
#obs	10	13	23	12	10	4	5
% of total	12.987	16.883	29.870	15.584	12.987	5.195	6.494
sd from mean	-1.102	-0.518	0.000	0.066	0.651	1.235	1.819

min: 13.2498, max: 272.405, median: 110.76, mean: 120.104, s.d.: 63.3869, #obs: 77

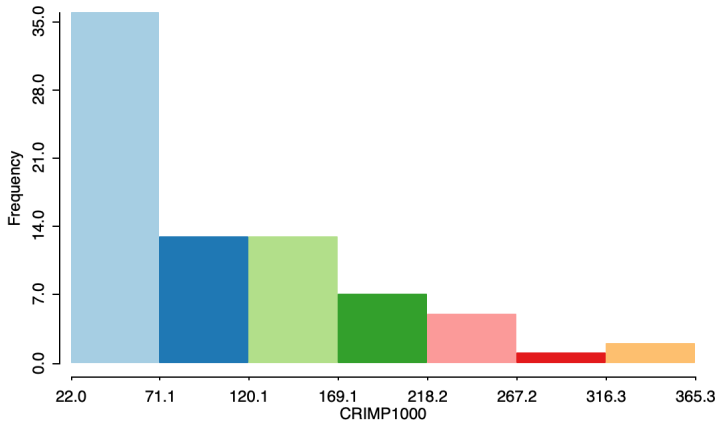
FIGURE 3. Histogram of Number of People without High School Diploma per thousand capita in 2014.



from	14905.000	309327.000	603749.000	898171.000	1192593.000	1487015.000	1781437.000
to	309327.000	603749.000	898171.000	1192593.000	1487015.000	1781437.000	2075859.000
#obs	61	9	2	1	2	1	1
% of total	79.221	11.688	2.597	1.299	2.597	1.299	1.299
sd from mean	0.000	0.055	0.873	1.690	2.508	3.326	4.144

min: 14905, max: 2.07586e+06, median: 190382, mean: 289557, s.d.: 360033, #obs: 77

FIGURE 4. Histogram of Income per capita in 2014.



from	22.014	71.057	120.100	169.144	218.187	267.230	316.273
to	71.057	120.100	169.144	218.187	267.230	316.273	365.317
#obs	36	13	13	7	5	1	2
% of total	46.753	16.883	16.883	9.091	6.494	1.299	2.597
sd from mean	-0.514	0.000	0.136	0.787	1.437	2.087	2.737

min: 22.014, max: 365.317, median: 74.9393, mean: 109.811, s.d.: 75.4334, #obs: 77

FIGURE 5. Histogram of Number of Crime happed per thousand capita in 2014.

Appendix B. Quantile Maps of the Variables

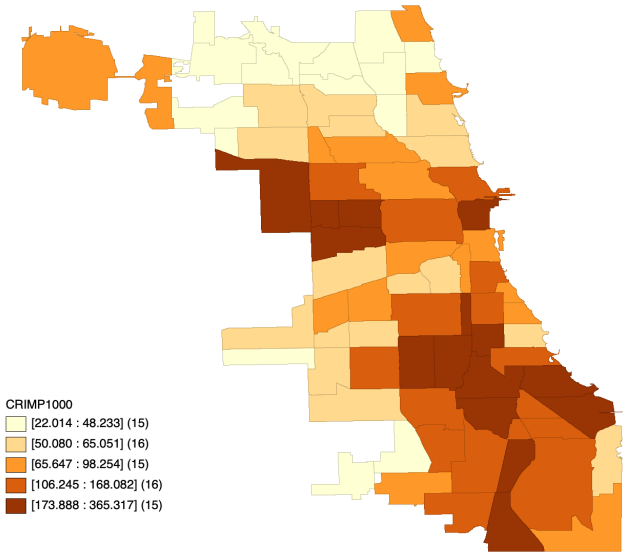


Figure 6. Quantile Map of Number of Crime happed per thousand capita in 2014.

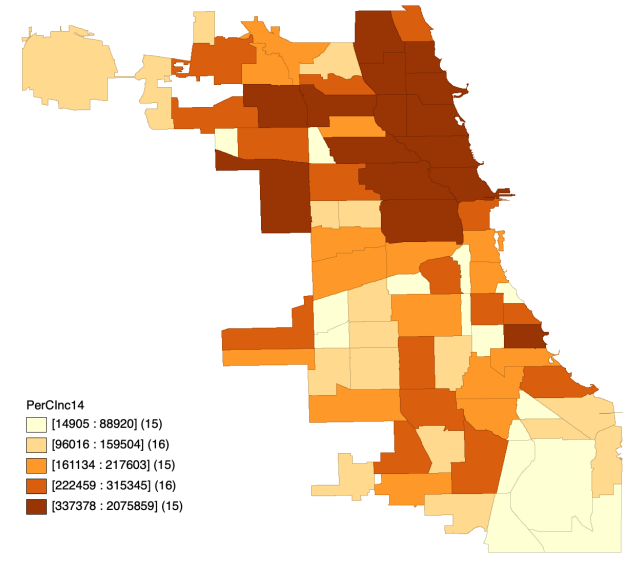


Figure 7. Quantile Map of Income per capita in 2014.

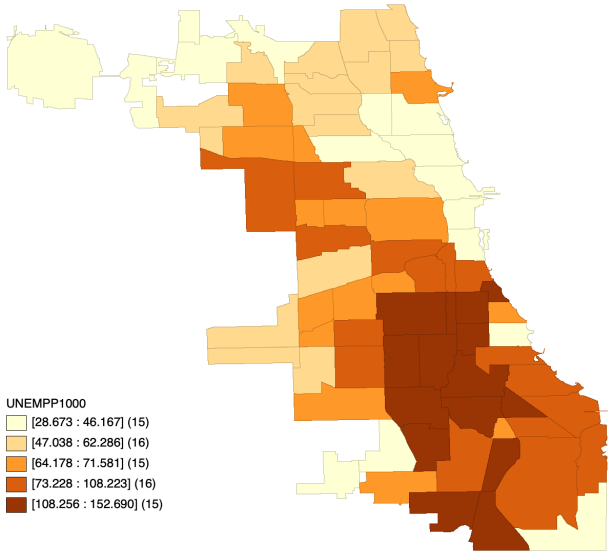


FIGURE 8. Quantile Map of Number of Unemployment Population per thousand capita in 2014.

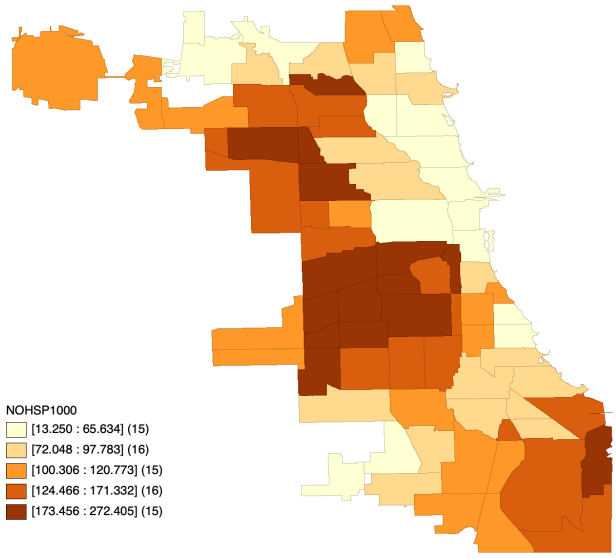


FIGURE 9. Quantile Map of Number of People without High School Diploma per thousand capita in 2014.

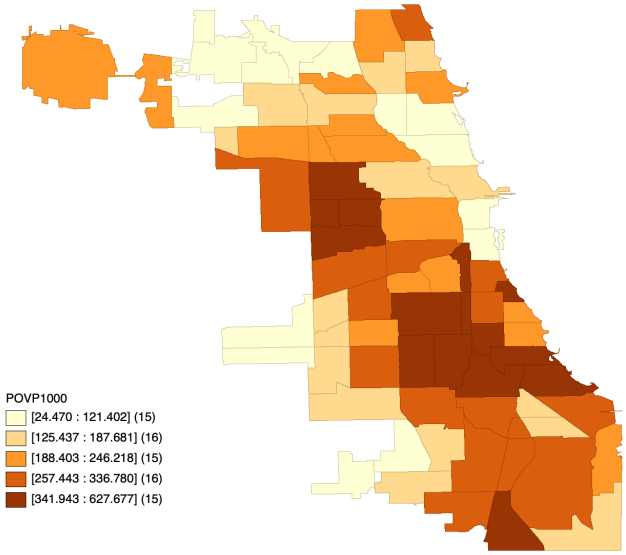


FIGURE 10. Quantile Map of Number of People Living in Poverty per thousand capita in 2014.

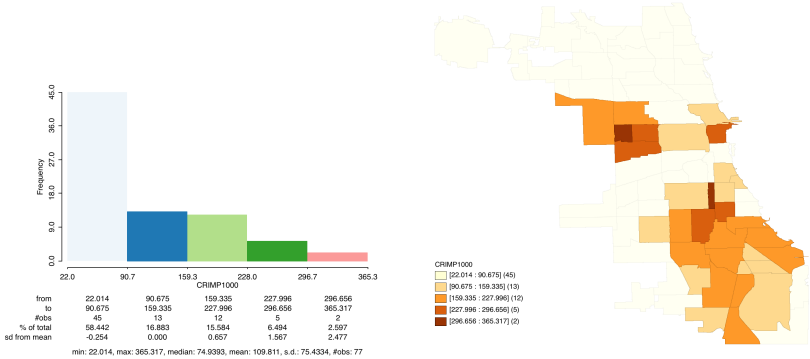


FIGURE 11. Histogram of Incidents of crime per thousand capita and the Corresponding Equal Intervals Map.

Appendix C. Scatterplot Matrix

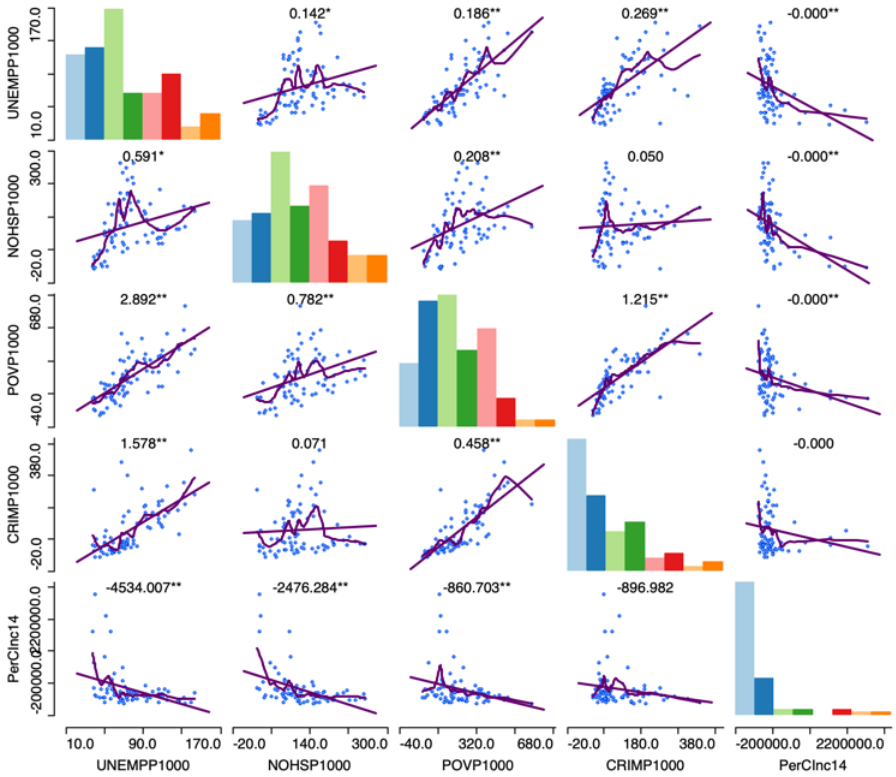


Figure 12. Scatterplot Matrix of Variables of Interest.

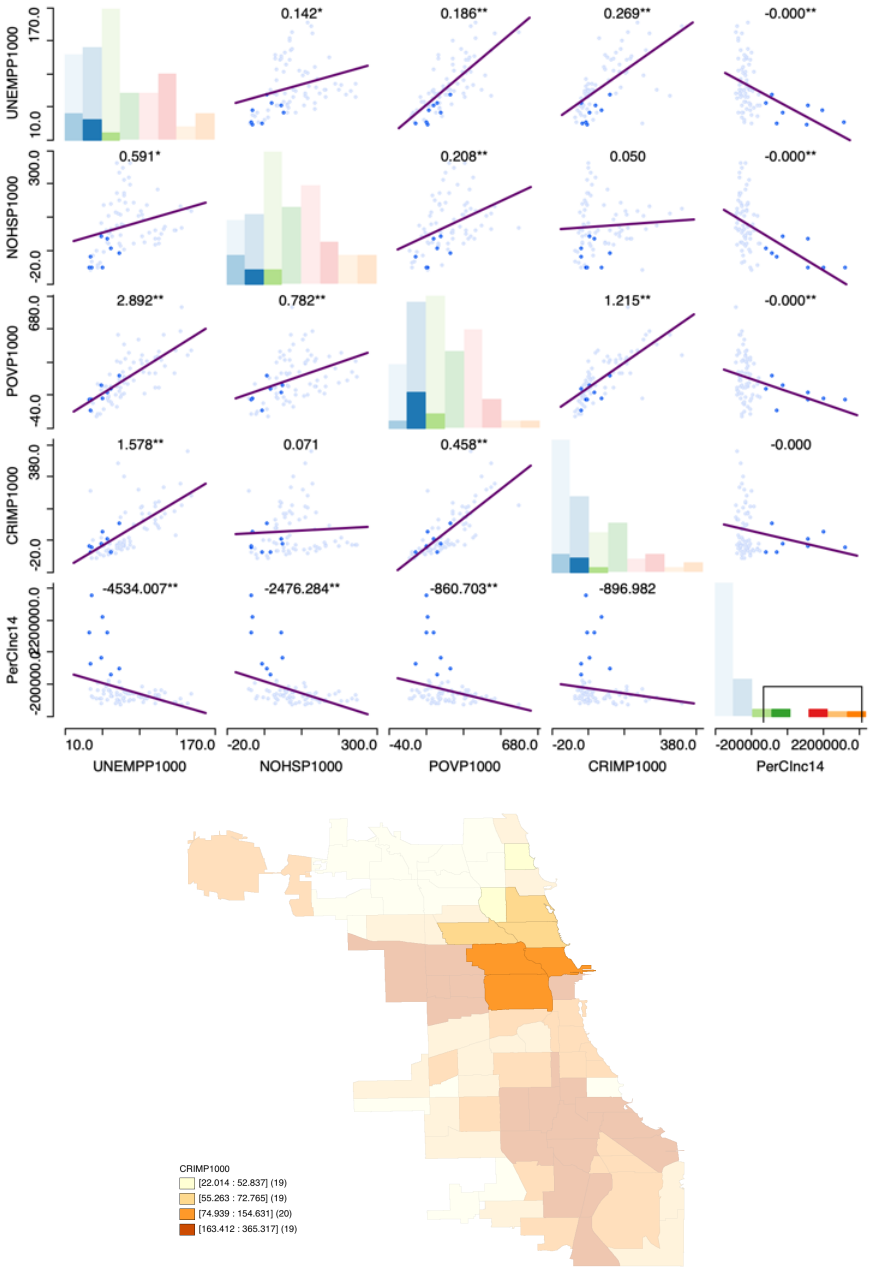


FIGURE 13. Scatterplot Matrix of the Five Variables and Crime Choropleth Map.

Appendix D. Boxplots, PCP, Conditional Map, and Rook Weight

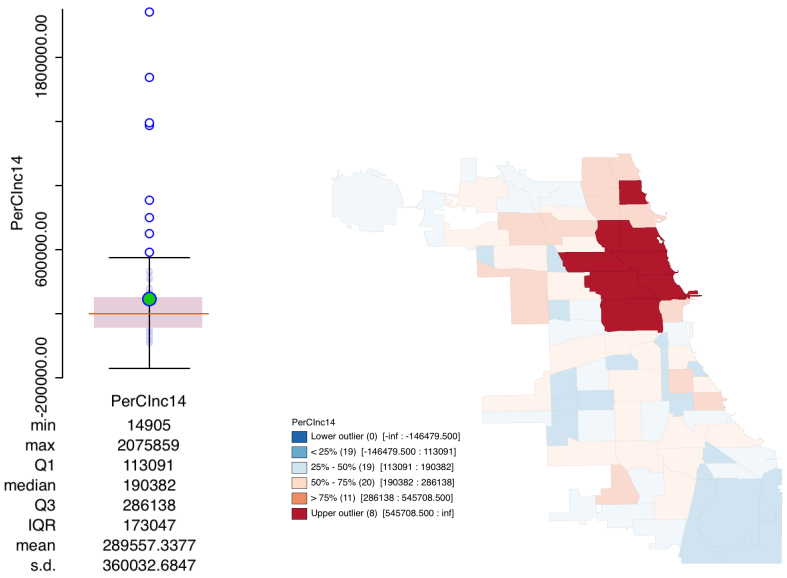


Figure 14. Boxplot of Income per capita and Income per capita Box Map.

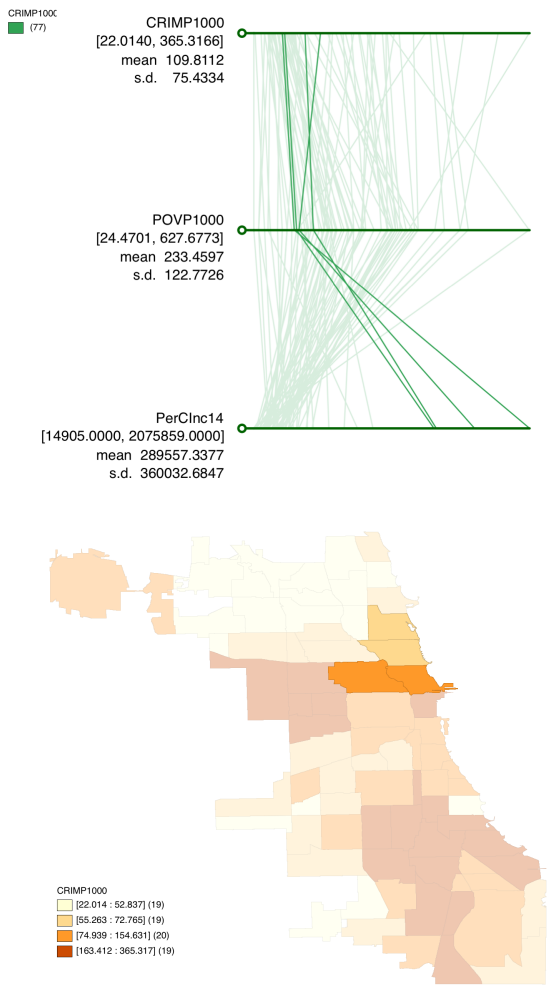


FIGURE 15. Parallel Coordinate Plot and Crime Choropleth Map.

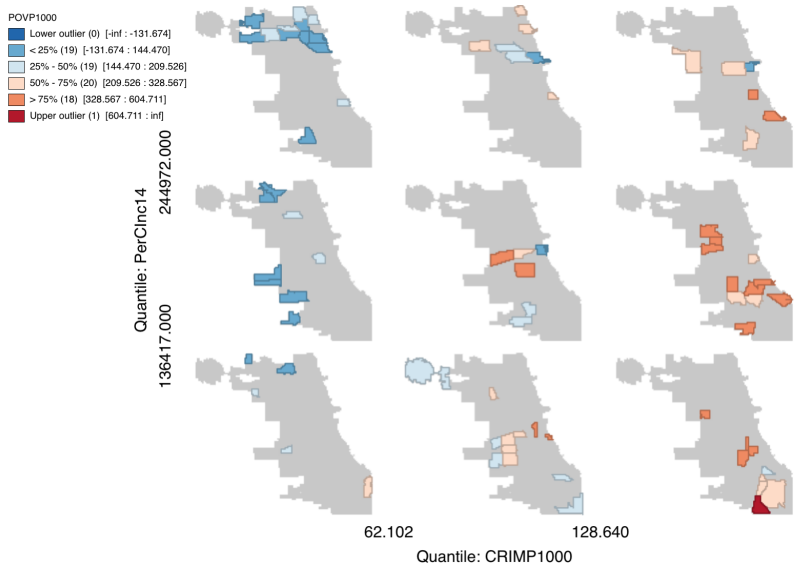
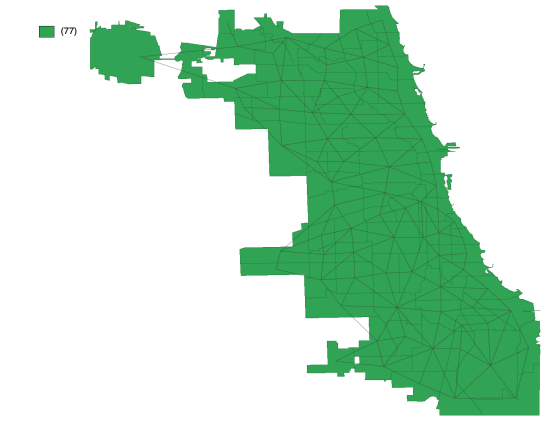
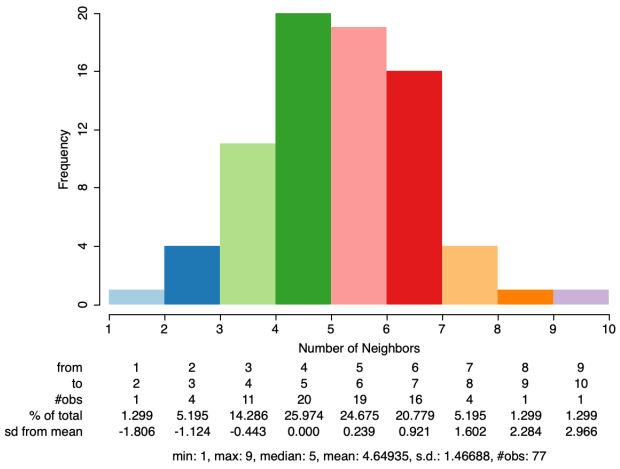


FIGURE 16. Conditional Map with Crime Rate x-Axis, Income per capita y-Axis, and Poverty Rate Base.

Property	Value
type	rook
symmetry	symmetric
file	OriginPolygonRook.gal
id variable	ComAreaID
order	1
# observations	77
min neighbors	1
max neighbors	9
mean neighbors	4.65
median neighbors	5.00
% non-zero	6.04%





from	1	2	3	4	5	6	7	8	9
to	2	3	4	5	6	7	8	9	10
#obs	1	4	11	20	19	16	4	1	1
% of total	1.299	5.195	14.286	25.974	24.675	20.779	5.195	1.299	1.299
sd from mean	-1.806	-1.124	-0.443	0.000	0.239	0.921	1.602	2.284	2.966

min: 1, max: 9, median: 5, mean: 4.64935, s.d.: 1.46688, #obs: 77

FIGURE 17. Rook Weight with respect to Original Geography.

Appendix E. Scatterplot

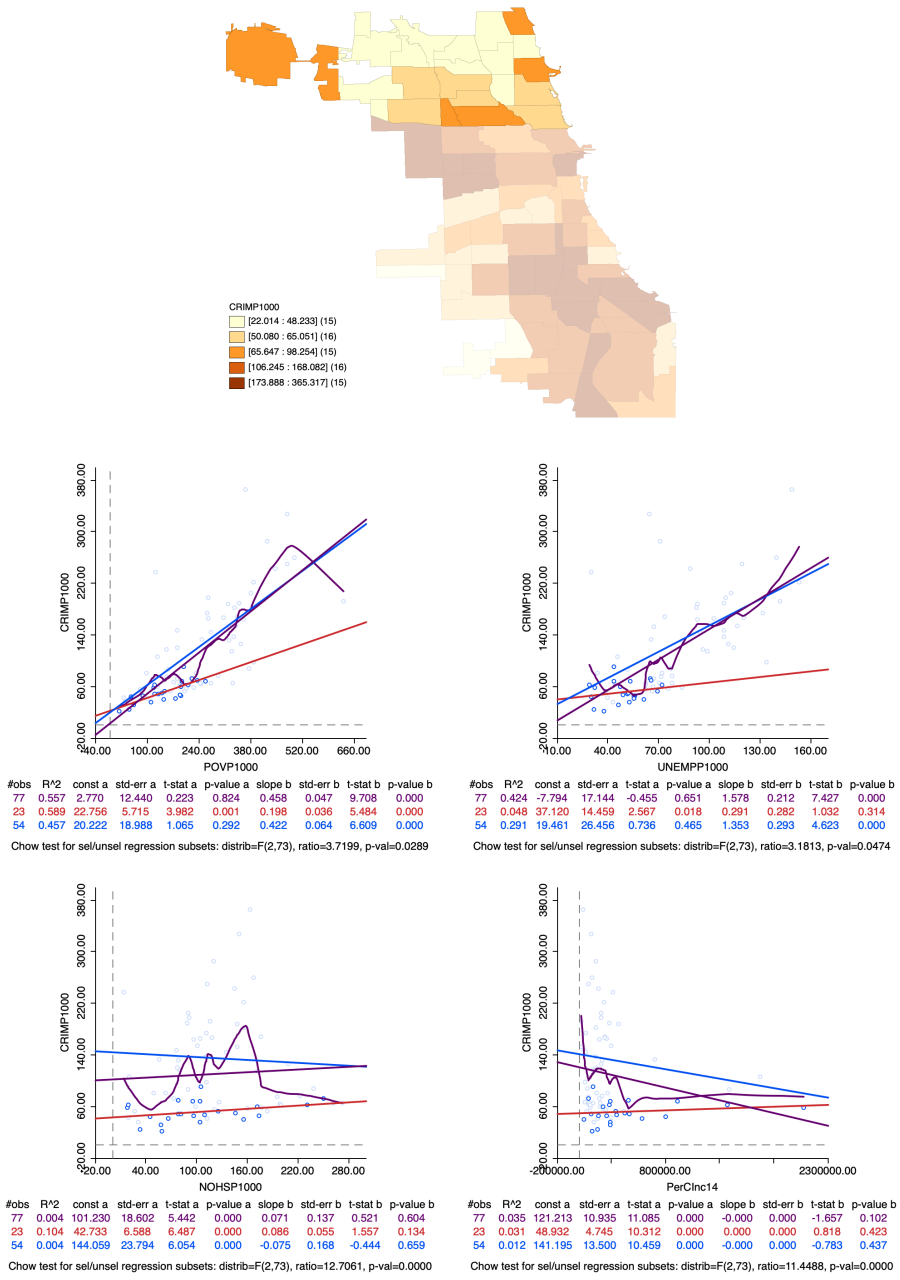


Figure 18. Scatterplot of Crime versus Variables of Interest and Crime Choropleth Map.

Appendix F. Moran's I Scatterplot

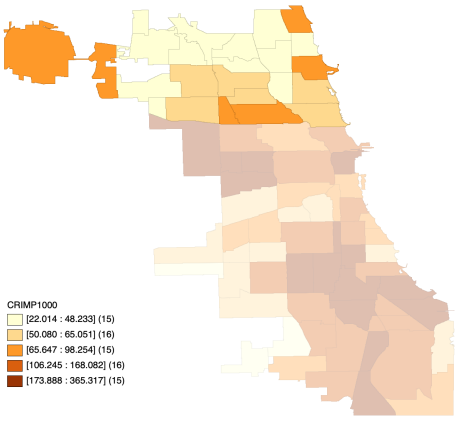


Figure 19. Crime Choropleth Map with North Part Selected.

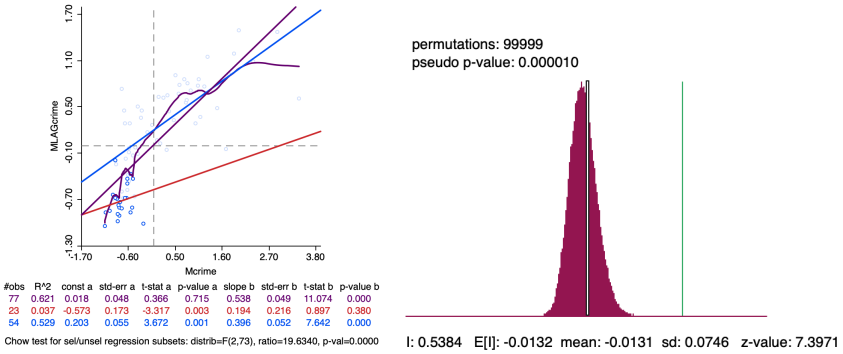


Figure 20. The Moran Scatterplot of Crime Rate with respect to Rook Weight.

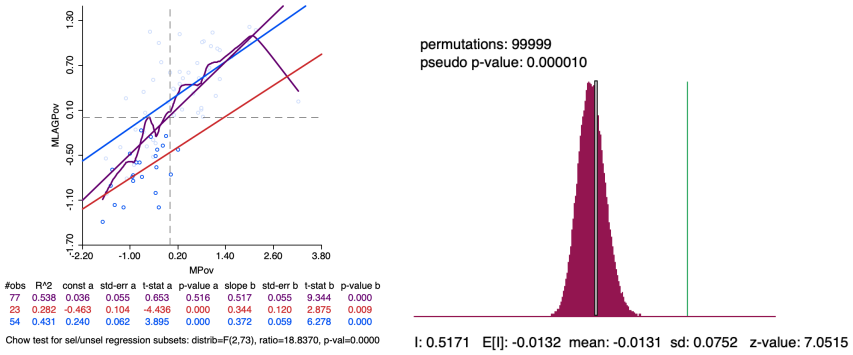


FIGURE 21. The Moran Scatterplot of Poverty Rate with respect to Rook Weight.

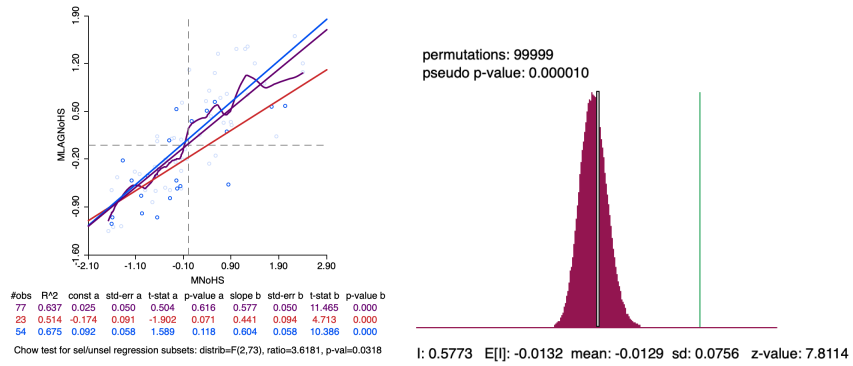


FIGURE 22. The Moran Scatterplot of No High School Diploma Rate with respect to Rook Weight.

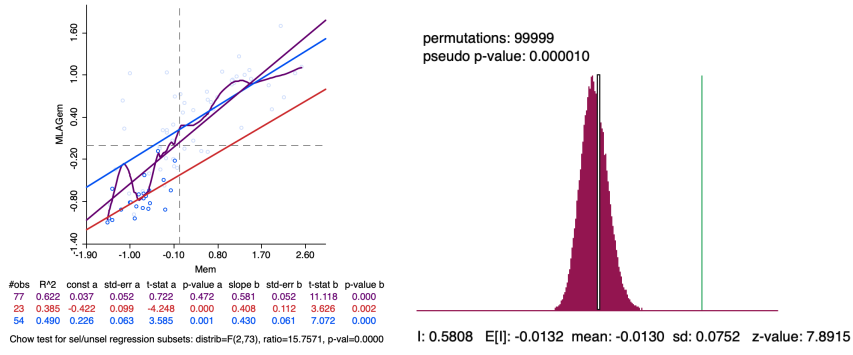


FIGURE 23. The Moran Scatterplot of Unemployment Rate with respect to Rook Weight.

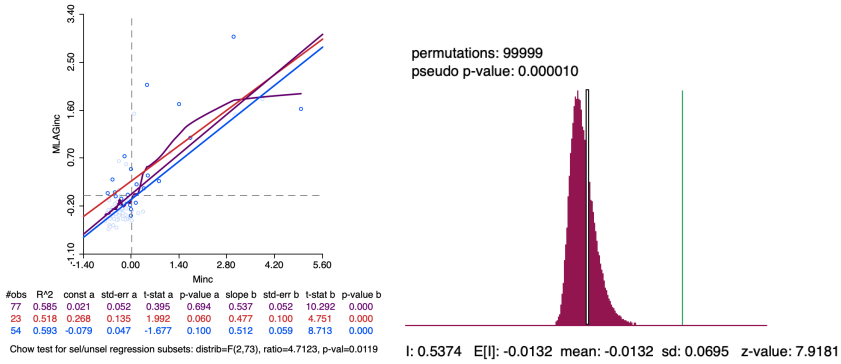


FIGURE 24. The Moran Scatterplot of Income Level with respect to Rook Weight.

Appendix G. Cluster and Significant Map

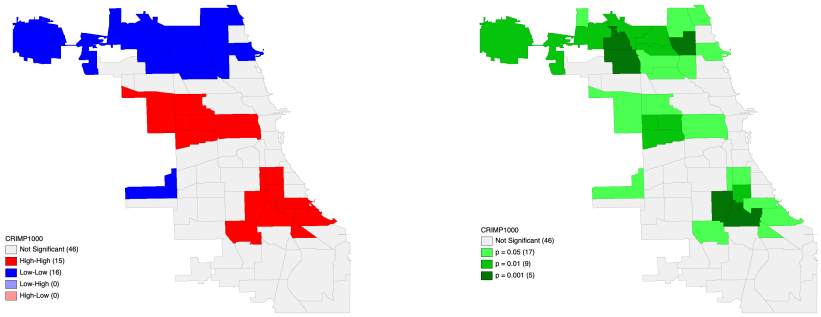


Figure 25. The Cluster Map and Significance Map of Crime Rate with respect to Rook Weight.

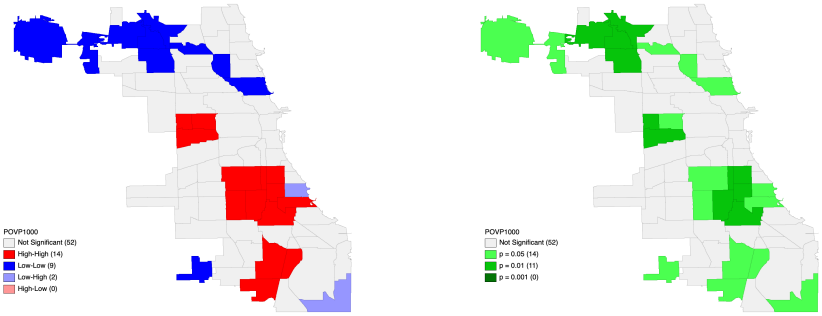


Figure 26. The Cluster Map and Significance Map of Poverty Rate with respect to Rook Weight.

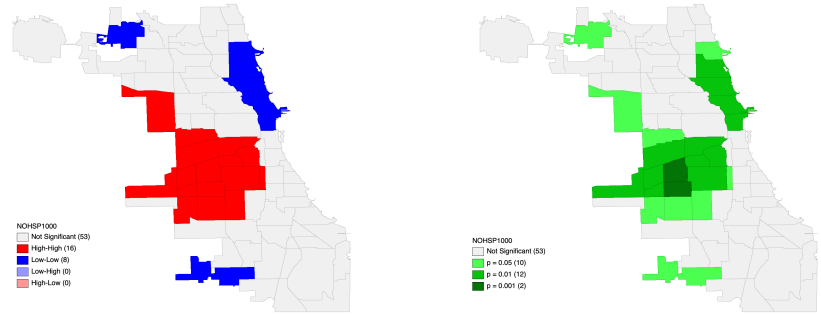


Figure 27. The Cluster Map and Significance Map of No High School Diploma Rate with respect to Rook Weight.

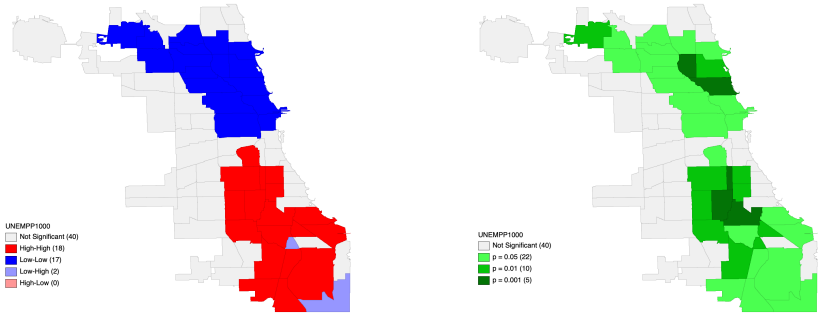


FIGURE 28. The Cluster Map and Significance Map of Unemployment Rate with respect to Rook Weight.

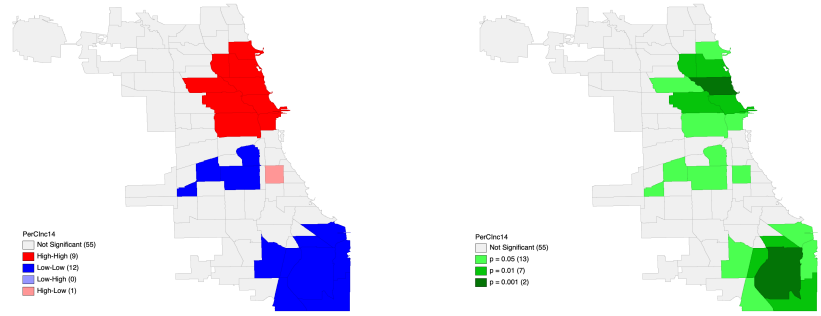


FIGURE 29. The Cluster Map and Significance Map of Income Level with respect to Rook Weight.

Appendix H. Principle Component Analysis

```
---

PCA method: svd

Standard deviation:
1.662251 1.077649 0.776793 0.554215 0.406241

Proportion of variance:
0.552615 0.232266 0.120682 0.061431 0.033006

Cumulative proportion:
0.552615 0.784881 0.905563 0.966994 1.000000

Kaiser criterion: 2.000000

95% threshold criterion: 3.000000

Eigenvalues:
 2.76308
 1.16133
0.603408
0.307155
0.165032

Variable Loadings:
                    PC1        PC2         PC3        PC4        PC5
UNEMPP1000     0.525054   0.141586  -0.119044  -0.806963  -0.197257
NOHSP1000      0.309563  -0.649624   0.613058  0.0745908  -0.317418
POVP1000       0.543921   0.170144   0.278339   0.157682   0.756879
CRIMP1000      0.470223   0.468109 -0.0772764   0.527739  -0.524676
PerCInc14      -0.33396   0.556657   0.725634  -0.199687  -0.110386


Squared correlations:
                     PC1          PC2          PC3          PC4          PC5
UNEMPP1000      0.761731    0.0232805   0.00855118     0.200016   0.00642154
NOHSP1000       0.264784     0.490095     0.226785   0.00170894    0.0166278
POVP1000        0.817455    0.0336194    0.0467474   0.00763702    0.0945407
CRIMP1000       0.610944     0.254477   0.00360334    0.0855454    0.0454309
PerCInc14       0.308164     0.359857     0.317721    0.0122479   0.00201093
```

Figure 30. PCA Result Table.

```
------
Weights:    OriginPolygonRook
Minimum bound:  464752(Pop2014)
# iterations:   99
Local search:   Greedy
Distance function:  Euclidean
Transformation: Standardize (Z)
Cluster centers:
|   |PC1       |PC2       |PC3       |
|---|----------|----------|----------|
|C1|0.0561576|-0.215519|-0.590258|
|C2|0.728434  |0.7676    |-0.368025|
|C3|-0.791796|-0.427121|-0.407591|
|C4|0.464558  |-0.95763  |0.975966  |
|C5|-1.44443  |1.53039   |1.55175   |

The total sum of squares:   228
Within-cluster sum of squares:
|   |Within cluster S.S.|
|---|-------------------|
|C1|32.0975             |
|C2|26.5897             |
|C3|10.5987             |
|C4|16.0555             |
|C5|16.6824             |

The total within-cluster sum of squares:    102.024
The between-cluster sum of squares: 125.976
The ratio of between to total sum of squares:   0.552527
```

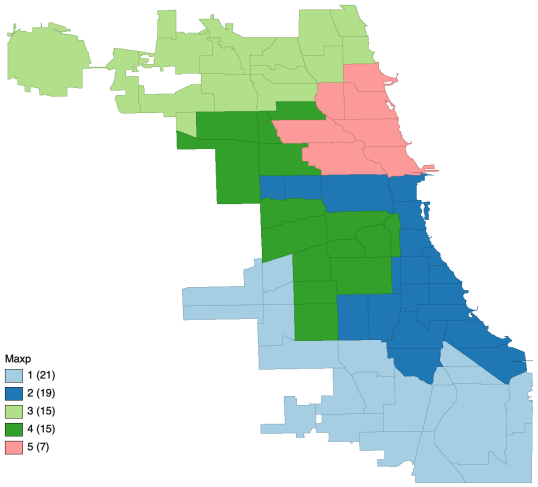


FIGURE 31. Max-p cluster map and characteristics (Minimum population bound = 17%)