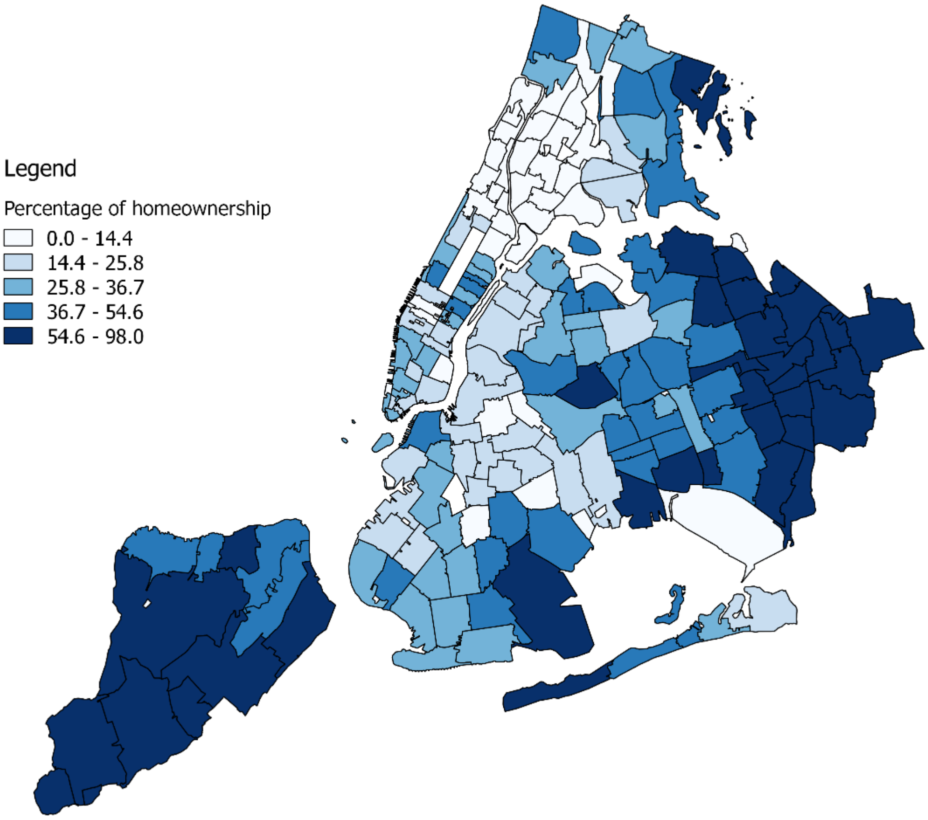
1. Is there a spatial relationship between home ownership and selected demographic characteristics at the zip code level in NYC?

Location: New York City

Level of detail: Zip codes

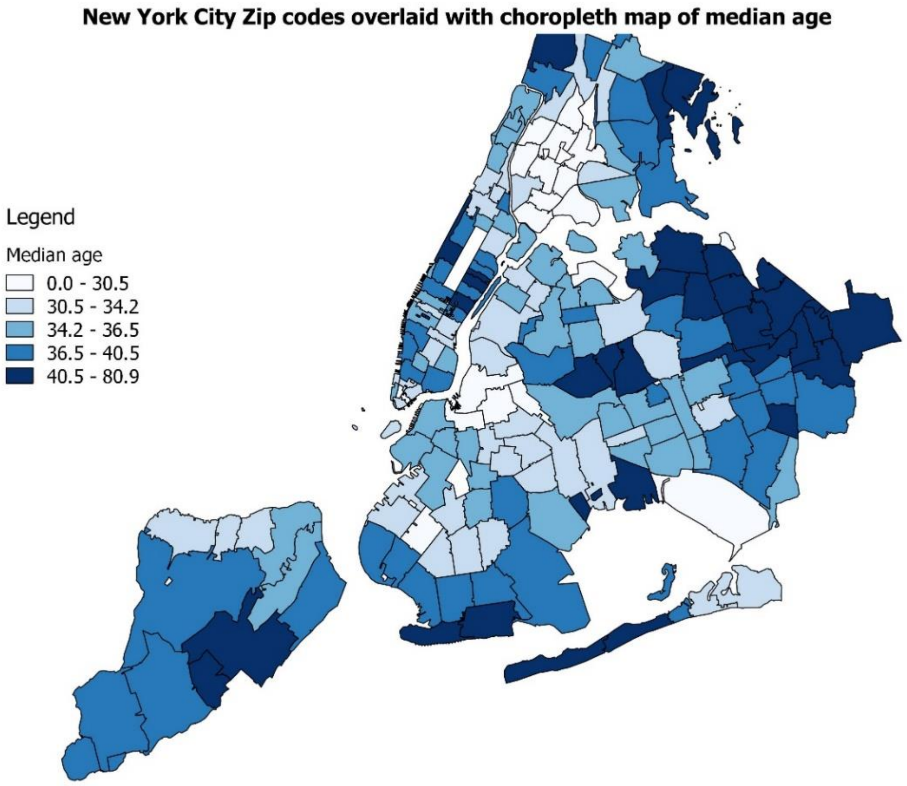
Coordinate Reference System (CRS): NAD83 / New York Long Island (EPSG: 2263)

New York City zip codes overlaid with choropleth map of homeownership



The top map is a choropleth map of the percentage of owner-occupied housing units for each zip code in New York City. This variable was obtained by dividing the number of owner-occupied units by the total number of units in each zip code. The density was then plotted for 5 classes using the Quantile mode. In this case, using quantiles was found to provide clearer visual separation between clusters. We notice three main clusters of homeownership in New York: First, in the Bronx and Washington Heights there is a cluster of zip codes where very few units are occupied by their owners. Second, in East Queens there is a cluster of zip codes where units are mostly occupied by their owners, which contrasts strongly with the first cluster. Third and last, on Staten Island, almost all the zip codes have a high percentage of owner-occupied units. In Manhattan and Brooklyn, the density across neighboring zip codes is a lot more heterogeneous so that no clear cluster emerges from the map.

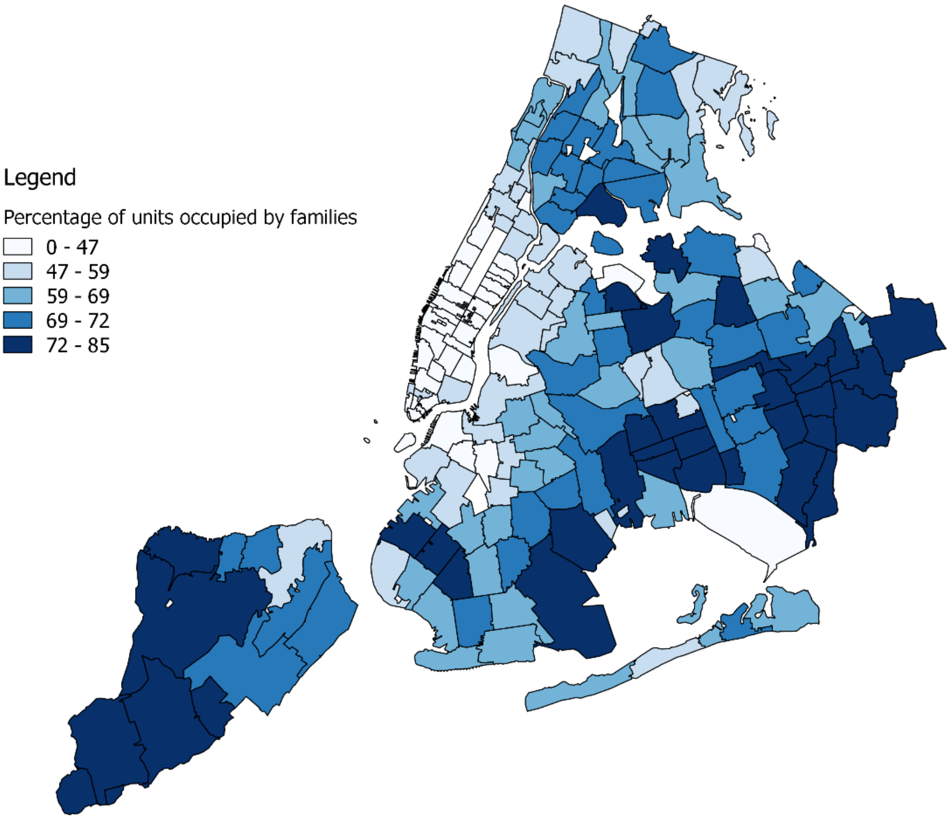
Next, we want to find 2 demographic variables that are spatially correlated with the fraction of owner-occupied units per zip code. To do so, we need to pick two candidates among all variables available and then plot them to assess their similarity with the distribution of homeownership. We hypothesize that older people are more likely to own housing units due to the longer time they've had to earn and save money. Thus, we pick the variable Median Age for Both Sexes among the total population (item DP0020001 in the table). The map is plotted below using quantiles distributed among 5 classes, i.e the same method used for the first map.


New York City Zip codes overlaid with choropleth map of median age

We notice that the area of the Bronx and Washington Heights (i.e, the first cluster described previously) showcases a cluster of low median age which reinforces the hypothesis made above that homeownership is related to age. Then we observe that East Queens has a concentration of zip code with high median age. This echoes the high levels of homeownership previously observed in this area and further supports the case that age is influential on homeownership. Last, we look at Staten Island and do not observe any cluster of high median age. This does not contradict our assumption because median age is somewhat high and relatively homogeneous on the island, but not as high as expected given the significant homeownership levels observed there. Outside of these three main locations of study, median age displays the same variability observed with homeownership in areas like Brooklyn and Manhattan.

Now let’s turn our attention to the second variable we want to select. We further hypothesize that living with a family is a strong motivation to buy and inhabit a unit, whereas individuals who live alone most likely rent their unit. To check this assumption, we create a new variable which expresses the percentage of family households per zip code (we create it as a fraction of number of family households by total households). Let’s plot the results using 5 classes of quantiles, as we did before.

New York City zip codes overlaid with choropleth map of density of units occupied by families



This time, the pattern observed in the Bronx and Washington Heights is the opposite of what we expected: this area shows a relatively high level of family households although few of them own the place. We would have expected to see a cluster of zip codes with low proportion of family households, such as what can be seen below Central Park on this map. Then we look at East Queens and Staten Island and we correctly observe the expected trend: these two locations have a cluster of zip codes displaying high percentage of family households. We can conclude that this second variable, just like the first one, is not a perfect predictor of homeownership but it is still able to explain some clusters observed.

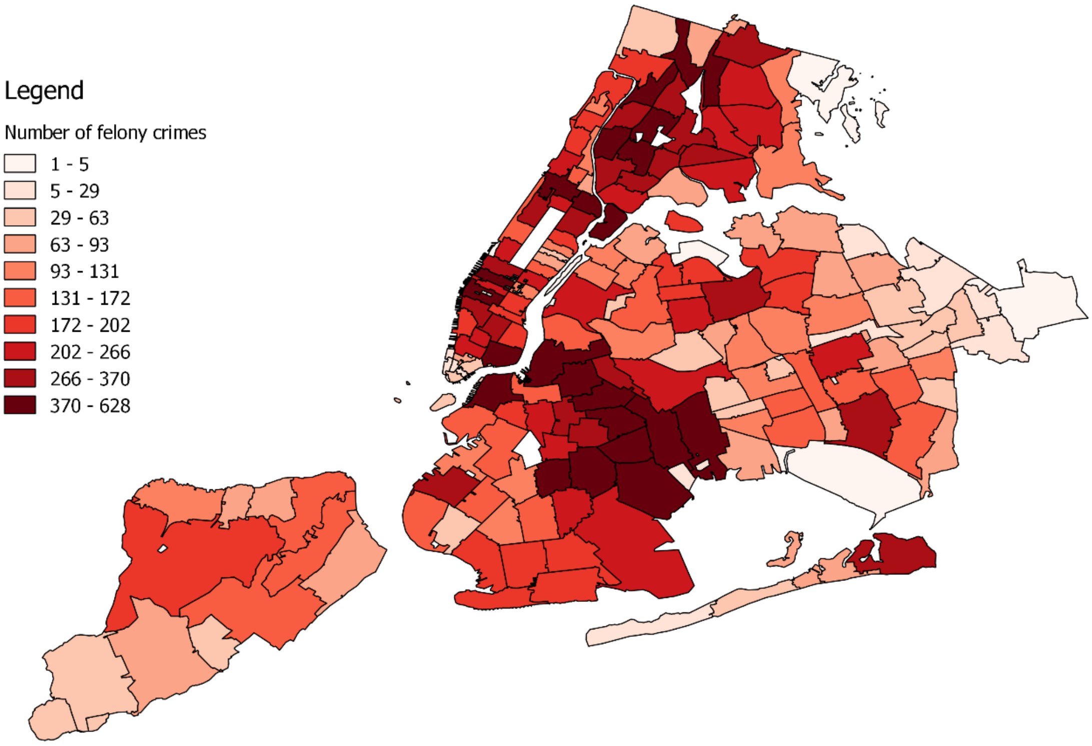
2. Is there a spatial relationship when we compare serious crimes in NYC?

Location: New York City

Level of detail: Zip codes

Coordinate Reference System (CRS): WGS84 / New York Long Island (EPSG: 4326)

New York City zip codes overlaid with a choropleth map of felony crimes counts

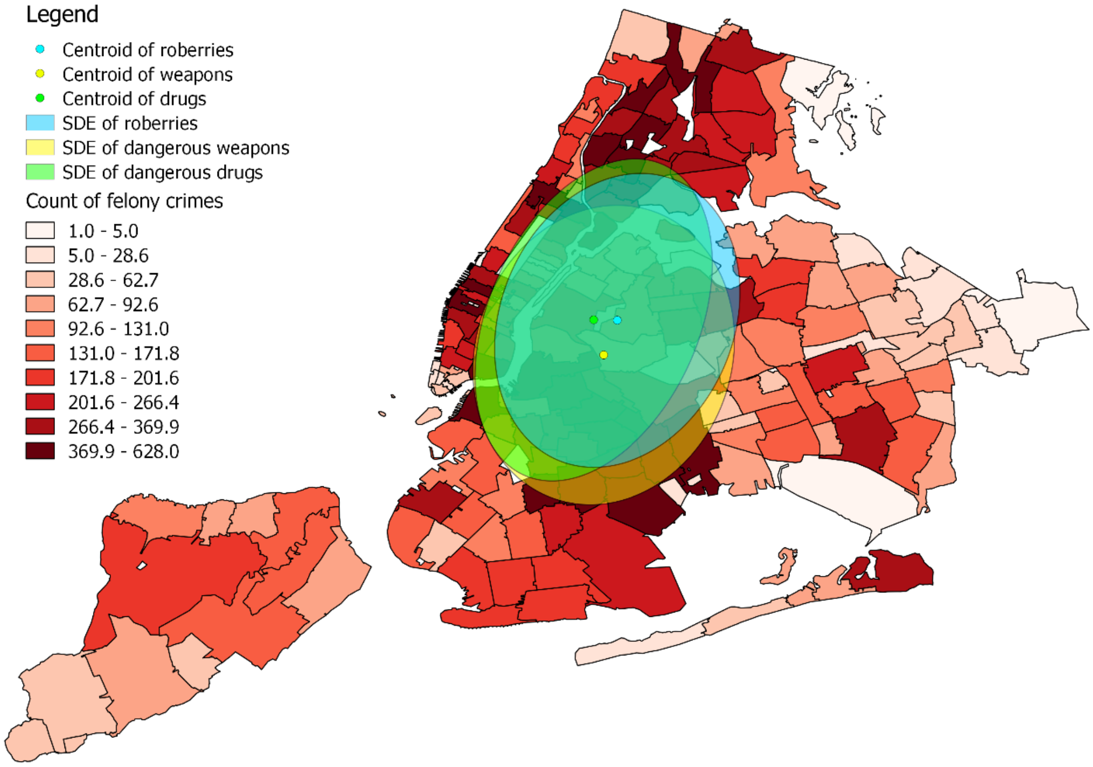


The top figure is a choropleth map of the number of felony crimes for each zip code in New York City. This variable was obtained by projecting a point layer of crimes onto the New York zip code layer. The density was then plotted for 10 classes using the Quantile mode. In this case, using quantiles was found to provide clearer visual separation between clusters. We notice two main clusters of felony crimes in New York:

First, in the Bronx there is a cluster of zip codes with very large number of felony crimes. Second, in Brooklyn there is a cluster of zip codes spanning from Williamsburg to East New York where felonies are also highly concentrated. These neighborhoods are notoriously underserved in terms of access to education and economic opportunities which could be an argument to explain their high level of criminality.

To better understand the distribution of felony crimes in New York. We focus our attention on 3 types of felonies: robberies, felonies involving the use of dangerous drugs, and felonies involving the use of dangerous weapons. To compare the central spatial locations and distributions of each of these 3 types, we create a point layer for each and then plot its centroid and standard deviational ellipse. The results are plotted below.

New York City zip codes overlaid with a choropleth map of felony crimes counts, centroids and standard deviational ellipses of 3 main types of crime

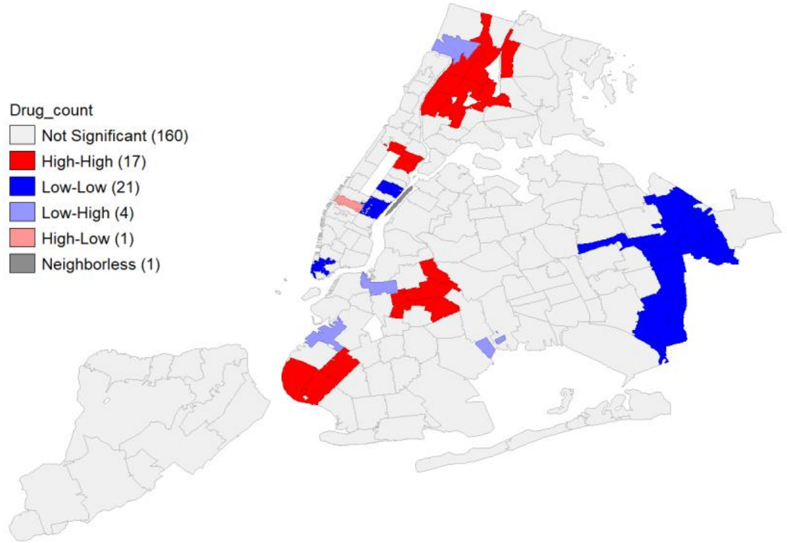


First, we notice that the centroid for dangerous drugs (green) is shifted towards the North West compared to the other 2, which suggests higher use of drugs towards Harlem. Second, the centroid for robberies (blue) is shifted towards the North East which denotes higher number of robberies in the Bronx cluster. Third and last, the centroid for felonies involving the use of weapons is shifted towards the South which may indicate higher occurrence of such crime in the Brooklyn cluster. The standard deviational ellipses are not significantly differing in their orientation, although we do notice that robberies and dangerous weapons have larger minor axis than dangerous drugs. This suggests that robberies and dangerous weapons are more prevalent in Brooklyn and Queens than dangerous drugs, which seem to be concentrated in Manhattan.

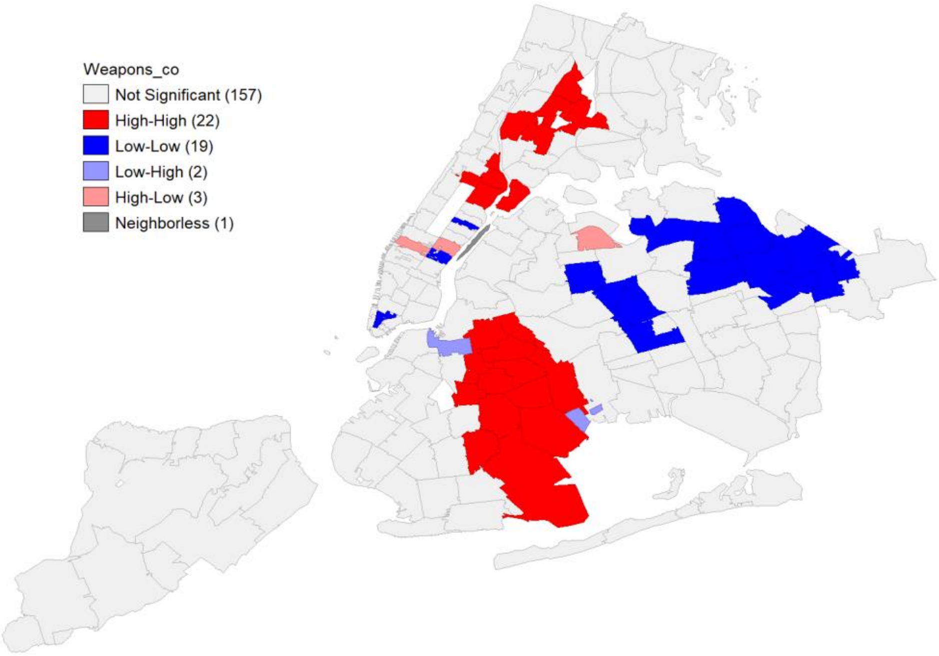
3. **Use the layer you created at the end of section two to calculate and interpret statistically significant clusters of robberies, dangerous weapons, and dangerous drugs in GEODA using LISA clusters.**

Now let’s study the statistical significance of the clusters in GeoDA. Below are three LISA cluster maps generated using univariate local Moran’s I, one for each type of felony. I first intended to display these results in QGIS but instead I had to visualize them in GeoDA because it would crash when I tried to save data as a new shapefile.

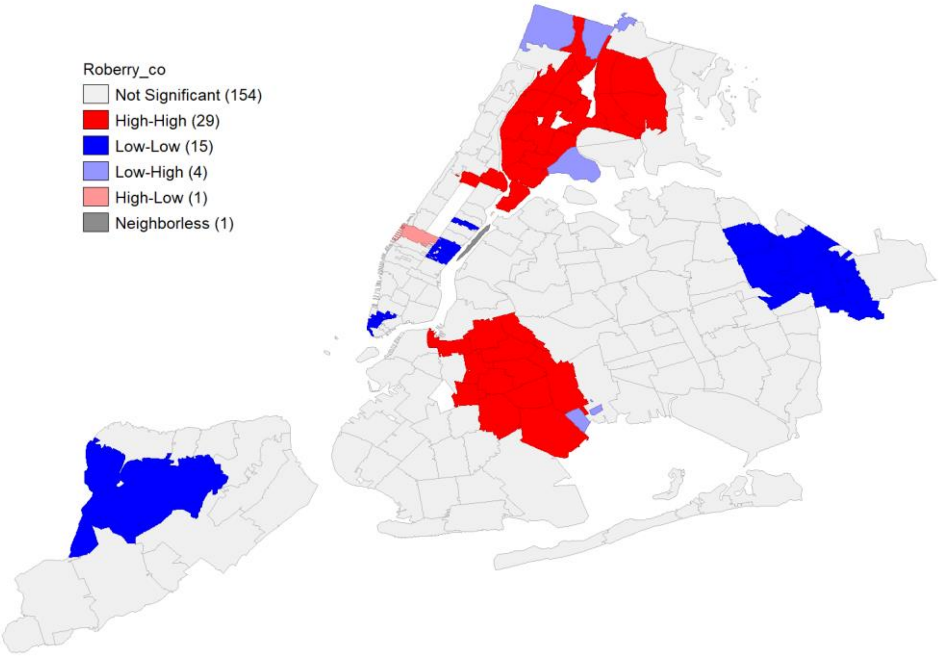
Fist, we look at the LISA cluster map for dangerous drugs. We see that there is a statistically significant cluster of high criminality in the Bronx and two others in Brooklyn (Bushwick and Bay Ridge). Additionally, there is a statistically significant cluster low criminality in East Queens.



Then we look at dangerous weapons. We still observe a statistically significant cluster of high criminality in the Bronx, but we see a new one in Harlem and an even larger cluster in East Brooklyn. The cluster of low criminality in East Queens is still present.



Third, we look at robberies. Here again, we find two clusters of high criminality in the Bronx and Brooklyn and a cluster of low criminality in East Queens. However, we notice a fourth cluster of low criminality in Staten Island that was not visible before.



Overall, for the 3 crime types studied, there is a statistically significant trend for high criminality clusters in the Bronx/Harlem and East Brooklyn, and a cluster of low criminality in East Queens.

4. **Why do social scientists need to pay attention to the Modifiable Areal Unit Problem? (Answer in no more than two to three paragraphs.)**

The Modifiable Areal Unit Problem (MAUP) is a zoning effect in spatial analysis which occurs when data measured at a certain scale are aggregated at a higher scale, resulting in data being blended. Social scientists pay attention to this serious issue because changing the scale of statistical values can lead to vastly different results and conclusions. In other words, spatial results are scale-dependent. Therefore, social scientists need to identify which scale is most appropriate for the data, and how aggregated data are interpreted.

5. **What is the Ecological fallacy and why do social scientists need to pay attention to it? How does the Modifiable Areal Unit Problem relate to Ecological fallacy? (Answer in no more than two to three paragraphs)**

The ecological fallacy is a form of fallacy that describes incorrect statistical reasoning. It occurs when individual data is aggregated into groups and when the conclusions drawn from these groups are incorrectly attributed the individual units. Social Scientists must be careful about the ecological fallacy as this is a wrong interpretation of data which causes incorrect statistical inferences on individual units. The MAUP is a form of ecological fallacy where observations are grouped into areal units, and the results obtained from these spatial units are used to make inferences about the individual observations.

6. **Why is it important to pay attention to the datum and projection that was used to generate location data attached to a shapefile?**

A datum is the coordinate reference system used to project data into a (x,y) coordinate system. It is important to know which datum was used to generate new data because the GIS must be set up in this very same datum prior to performing any analysis. Failure to do so will result in an incorrect data projection, i.e the location of data as seen in the GIS will not reflect their true location. This can lead to significant problems during the data analysis because conclusions drawn from the GIS for a certain area will be wrong in reality.