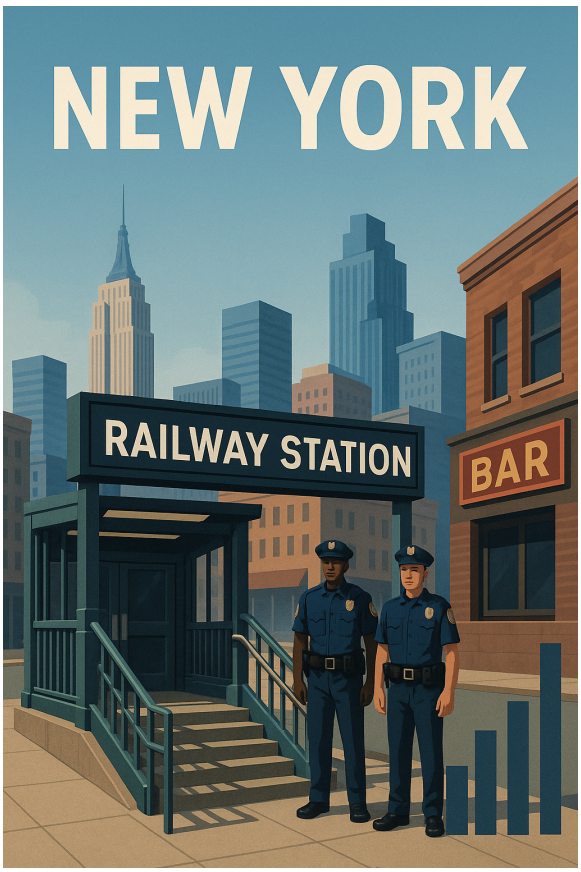
Geospatial Data Science: Analysis of New York Crimes



Jacob Grum (jacg)

Joakim Goyle Dissing (jodi)

Thor Liam Møller Clausen (tcla)

Deadline for submission: May 23, 2025

28504 characters including spaces ≈ 12 pages
(excluding front page, table of contents, references and appendices)

Abstract

Crime can be a complex subject with many different factors. Finding those factors can be challenging and even often misleading. This report examines the spatial relationship between crimes committed in New York City neighbourhoods and the presence of specific amenities, namely bars, restaurants, and railway stations. Utilizing geospatial data analysis methods such as Ripley's K-function, Hopkin's statistic, and LISA spatial autocorrelation, significant clustering of crimes around bars and railway stations was identified, while restaurants did not consistently show similar patterns. These findings align closely with existing research conducted in other urban areas.

Contents

1 Introduction

1.1 Problem description

The goal of this project is to research the spatial properties of crimes committed in New York City (NYC). We aim to explore how crimes in NYC are related to certain amenities. The investigation will be utilising amenity location data from Open Street Map (OSM), specifically bars, restaurants and railway stations, to study the spatial clustering of crimes related to these amenities. In this report the overall term "crimes" refers to the American crime classifications being felonies, misdemeanours, and violations. We will explore the spatial relations between these data, to get a better understanding of the factors that could influence or prevent crime in large urban areas.

2 Background

During our exploratory research on crime, we found some areas of interest, we wanted to dig deeper into:

- Railway stations
- Bars
- Restaurants

Crime is a multi-faceted complex area, with an enormous range of factors contributing, where just finding a single motivation for each is a highly paid job in the FBI. What our areas of interest have in common, is their frequent occurrence in crime related discussions, acting both for and against crime [5].

By selecting New York we had plenty of crime and amenities data available. In 2023 there were 2190 crimes per 100.000 population in New York[22].

2.1 Do not leave bags unattended

Stations are concentration points of large volumes of people at particular places and times, thus increasing the number of potential theft targets. Studies show that this type of environment, increases amount of thefts[23]. Theft under 1000 dollars are considered a "petty larceny" which is under the category misdemeanour[8]. In general this is the dominant crime category in the transit system[10].

2.2 Drunken knuckles

A lot of assaults in general occur near bars[17]. Some of the factors contributing might be low ratio of staff to patrons, alcohol and higher acceptance of disorderly conduct. Additionally, our source also tells us that crimes near bars are under-reported, due to bars often handling their own patrons with their own security, instead of reporting them to the officials[17]. In most states these results in a misdemeanour[21]. Even those problems escalating to assaults are often classified as misdemeanours[1].

2.3 The eyes are watching

Jane Jacobs proposes a "eyes on the street" theory, which in short says that certain environments and their amenities, acts like a "natural surveillance". For example the staff of the restaurant and their patrons would organically observe their surroundings, paired up with the perspective on the criminal being "A criminal who thinks he will be seen should be less inclined to offend"[7][5].

2.4 Easy targets in crowded streets

Initially, looking at population density itself as a crime factor, wasn't a result we were looking for. But during our research the keyword "density" together with "crime" occurs frequently. Such as above, regarding crime in stations, the source does specify, that people being densely packed itself leads to more crime. This is also mentioned by Jane Jacobs, that people in densely packed situations have a harder time, being natural observant for any criminal activity. In other words, one way for the criminal to be seen less, is to have other people being seen more, and as such they use the crowd as a hiding mechanism.

This might also be why the top spots for most pickpockets in Europe, are very densely packed areas with high foot traffic, such as Rome and Paris being number one and two[6].

3 Data

This section concerns the data we used throughout this project, its purposes, sources, and how we processed it. All data sets cover only New York City administrative boundaries. We work with three main data sets:

- **Crime** data for determining where crimes occur and the type of them. Violations are the least severe, misdemeanours are a step above, and felonies are the most severe crimes.

- **Amenities** for mapping out the amenity's location and type. We ended up using bars, restaurants, railway stations, and parks.
- **Neighbourhood tabulation areas (NTA's)** are the smallest administrative divisions of New York City that we could find. We use these to get estimates of the population distribution or population density. We also believe that NTA's give a more nuanced result, than looking at the broader New York boroughs, namely Manhattan, Brooklyn, Queens, The Bronx, and Staten Island.

The meta data tables explaining the data columns can be found in appendix A.

3.1 Data acquisition

3.1.1 NYC Crime data

We obtained crime data from the New York Police Department (NYPD)[12]. The data set contains all valid crimes reported to the NYPD from 2006 to the end of 2019. The data set contains 9.49 million rows and 35 columns, where each row is a complaint. The columns we use in this project contain data concerning the type of crime committed, and the latitude and longitude in EPSG 4326 format.

3.1.2 Amenity data

We use the OSMnx library to fetch data from Open Street Map (OSM)[11]. With the library, we can query all data from New York City with custom tags like key "amenity" and values of our own selection (like bar, restaurant, etc.). The dataset contains 1321 bars, 7521 restaurants, 551 railway stations, 2054 parks and all with a geometry in EPSG 4326.

3.1.3 NTA data

The NTA data consists of two data sets merged into one. We get both datasets from NYC's department of city planning. The first of the two contains NTA names and multipolygons[3]. The second dataset contains NTA names and population[2]. How these data sets were combined is described in the next section, and was a major part of the data preprocessing.

3.2 Data preprocessing

The crime and amenity data did not require much preprocessing. There were some amenities that were registered as polygons and other geometric objects. We decided to convert them to points using their centroid for simplicity and consistency in the data. We also decided to only keep crime

data from the latest year 2019[1], because nine million data points were too computationally slow to work with. There were around 460 thousand data points from 2019.

3.2.1 Struggles with NTA data

All NTA's supposedly have an NTA code. The codes were included in both data sets. However, they did not use the same NTA code scheme. This made it impossible to join the data sets on NTA codes. We ended op joining them based on their NTA names instead. These did not match up one to one, but in short we paired them based on how many words they had in common. More details can be found in the source code (Appendix C).

The NTA population data set contains fewer rows than the NTA polygons data set (195 to 262), which means some polygons are without a population. These are mostly parks, and other uninhabited areas like airports and certain islands.

In some cases the two data sets disagreed on how the NTA's were divided. For instance the polygons data set contained "Bushwick (East)" and "Bushwick (West)" while the population data set contained "Bushwick North" and "Bushwick South". There were also instances where one row in the first data set would correspond to two rows in the other. For example the row "Allerton-Pelham Gardens" in population data corresponded to "Allerton" and "Pelham Gardens" on separate rows in the polygons data. In case our algorithm couldn't handle the rows, they were often manually handled, by merging the two conflicting NTA's into a single NTA. This was a necessary trade off where some granularity was lost, since the result was fewer rows. It required a lot of manual work, but with only $\sim$450 rows in total it was manageable.

3.3 Assumptions

To argue for the validity of our analysis we had to make some assumptions about the data:

- The population data is from 2010. We assume there has been no drastic changes up until 2019 where the crime data is from.
- The amenities from Open Street Map is continuously updated, meaning the data is from 2025. We assume they were also there in 2019 when the crimes happened. (E.g. if there is a bar in a certain location today we assume there was a bar at the same location in 2019.)

[1]While working on this project the data set was updated (April 15th). This means 2019 is no longer the latest year, however we kept using 2019 data to avoid the extra work of changing.

- Our method of merging the data may have imperfections and introduced some false data, like some NTA's having an incorrect population reading. We assume the error is minimal.

4 Results

4.1 Interactive map

To get an initial overview and a better understanding of the data, we started by creating an interactive map with a bunch of different views. It was made primarily with Folium[19].

4.1.1 Point data maps

We started with a simple scatter plot as seen in appendix E. There were too many crime points to get a proper overview. Therefore we made kernel density estimates (KDE)/heat-maps of crimes and amenities. This made it easier to comprehend where the different point data were concentrated. Figure 1 shows the heat map of of the crime data.

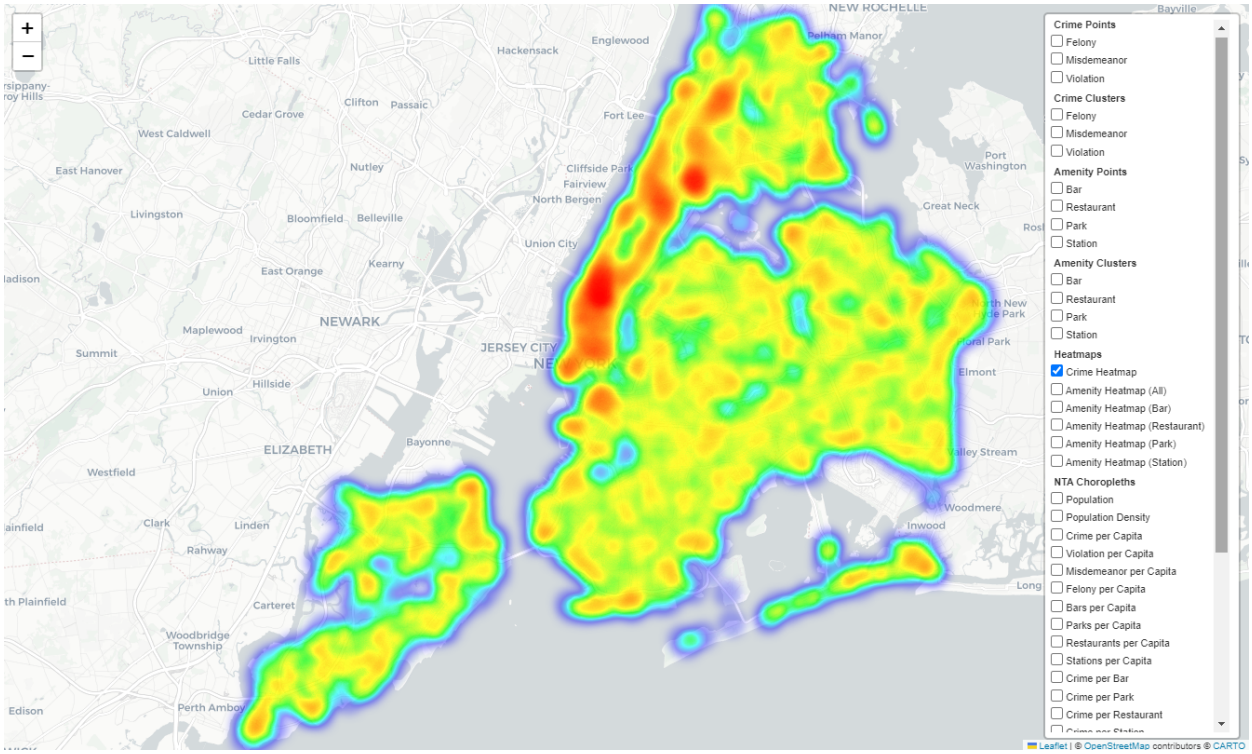


Figure 1: Our interactive map with crime heat map toggled.

4.1.2 Choropleth maps

We used the NTA polygons as spatial units for the choropleth maps. We wanted it to be as granular as possible with the MAUP in mind. The choropleth maps helped us show statistics like crime per capita.

We tried different classification schemes. We did not use standard deviations because the data did not generally have a normal distribution. In appendix D there is an example of the NTA population density being right skewed. When using even intervals a few outliers were highlighted, while the rest were hard to distinguish from one another. This was useful to spot great outliers in crime per capita for instance. But, we found quantiles to be the best overall classification scheme. It does hide the outliers more, but it makes it easier to see the more general distribution. Figure 2 shows an example of a choropleth map with crime per capita as value.

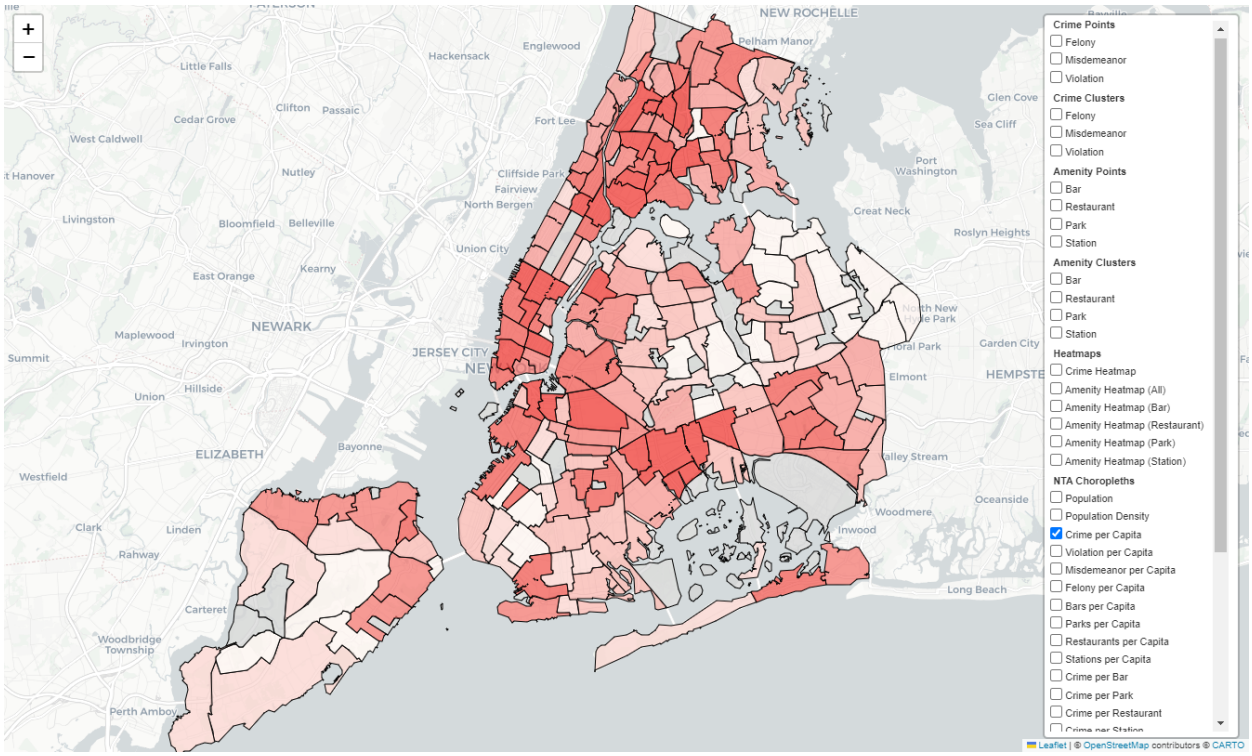


Figure 2: Our interactive map with population density per NTA toggled, shown with a choropleth map.

We could not get the legends to appear and hide with the choropleth maps, so we hid them altogether. The alternative would have been to re-render the map every time with a new legend, which would have been too time costly. In general, darker colours means it belongs to a higher quantile.

4.2 Clustering

If there was no clustering in our data, this would already be a strong hint that crime is not centered around amenities. It wouldn't be centered around anything if it was regularly spaced instead of clustered.

When computing clustering scores like Hopkin's H and Ripley's functions, the results are sensitive to the window that the points are in. Therefore, we set the window shape to be the New York borough boundaries, which we got using the geodatasets library[4]. This prevents comparing the observed points to simulated points potentially placed in the ocean.

4.2.1 Hopkin's H

We checked for clustering tendency with Hopink's H. We could not find a library that computes Hopkin's H, and had to rely on an implementation we found on an online forum[9]. The results are shown in Table 1.

Data points	Hopkin's H score
Crimes	0.999
Bars	0.978
Restaurants	0.973
Stations	0.907
Parks	0.728

Table 1: Hopkin's H score for each of the point data sets.

Most of the data has a high tendency to cluster, except for parks which is closer to a random distribution with only a small cluster tendency.

4.2.2 Ripley's K

To get a better insight into how the different data were clustered we wanted to use Ripley's functions. We used the pointpats library[15], which has both the F, G, K and L functions. We decided to focus on the K and L functions, and avoided F and G because they are more focused on empty spaces. Reason being, our scatter plot from appendix E revealed data points all over the map leaving only tiny empty spaces.

We tried the L function, but the results were a bit strange. As seen in appendix F the L function did not give the results described in the documentation of the tool. For this reason we ended up using Ripley's K function, which measures the same as the L function but without normalization.

It took a long time to compute even with small samples of the data. It was infeasible to run it on all $\sim$460 thousand crime data points. By sampling 10 thousand points we could compute the results in just around 10 minutes.

An unfortunate limitation, but hopefully it did not affect the results too drastically.

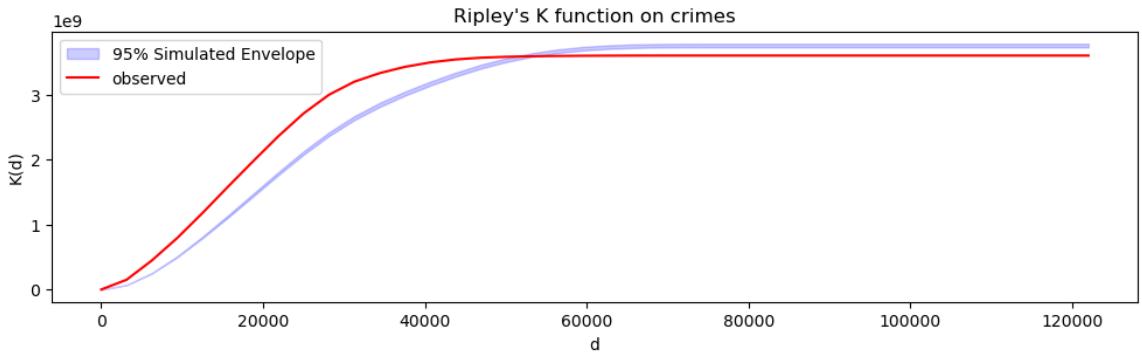


Figure 3: Graph showing Ripley's K score on crime data vs randomly simulated data, as distance d increases.

Figure 3 shows the result of Ripley's K on the crime data, with the distance metric d being in meters. Initially it was in degrees, but that was harder to interpret, so we converted the data to ESPG 3857 because it uses meters as metric[13].

For some unknown reason the observed value and simulated values don't align as distance approaches max[2]. This would have been expected since max distance guarantees that all points have all other points as neighbours within this distance. It could have something to do with the way the function estimates expected number of pairs.

All the results are shown in appendix G. In general there is a high cluster tendency in the data. Already from small distances the data points show a higher number of neighbouring pairs than the simulated data. Bars especially show a high deviation, where most bars are within a 20-30 kilometre radius while the simulated data requires around 40-50 kilometres to get the same number of neighbour pairs.

4.2.3 Plotting the clusters

To get a better understanding of where the clusters were located we computed the clusters using the sklearn.cluster library[16]. There are a lot of different algorithms to choose from. We excluded DBSCAN because it may struggle with varying density in the data, which is exactly what we have, as observed on the point plot from appendix E. Instead we chose agglomerative clustering with ward linkage, because it is good at separating a dense mass

[2]As max distance, we used the diagonal of our point data's bounding box. This is the highest possible distance between two points and should catch all neighbouring pairs.

of data from a less dense. We used mean shift to get a second view that doesn't require a pre-specified number of clusters.

We could not run the algorithms on all the crime data points, because it couldn't allocate enough data according to the error message we got. Instead we sampled 50 thousand. The result is shown in Figure 4.

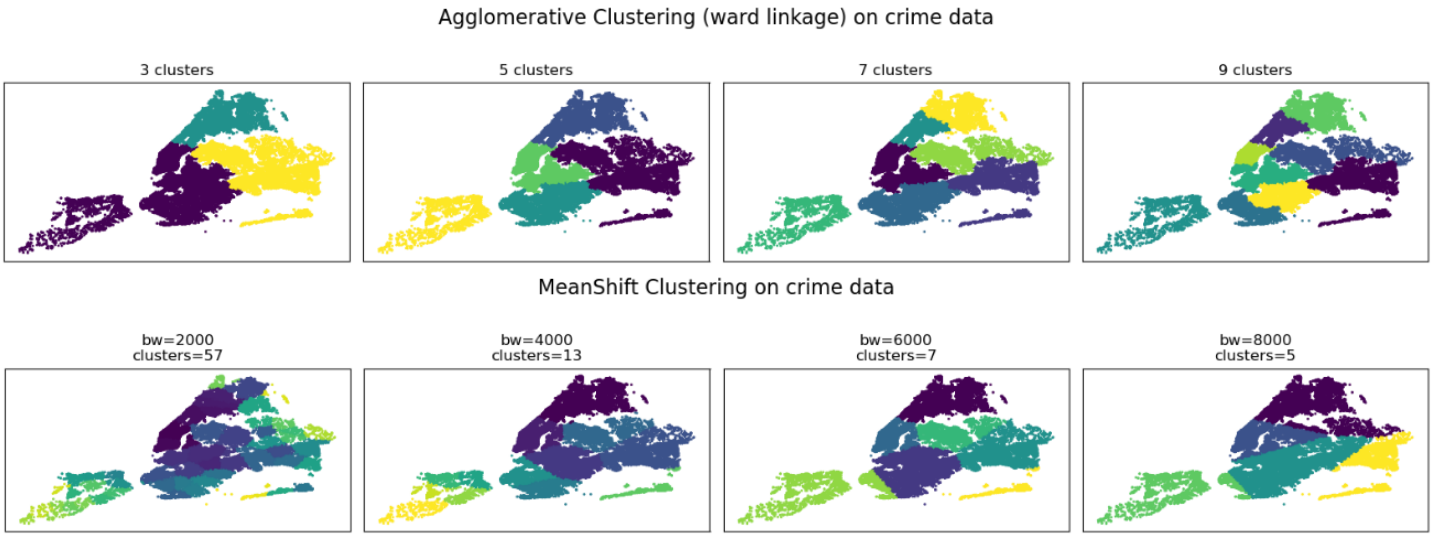


Figure 4: The clusters computed for the crime data.

All the plots of the clusters can be found in appendix H. There were some patterns that kept repeating. First of all there is a natural clustering caused by the land that makes up New York City. The ocean makes natural separations causing for instance Staten Island to have its own clusters. Another observation is that even within the islands, the clusters seem to form in similar ways across the different data sets. An example is lower Manhattan. Even as the number of clusters increase, the area in lower Manhattan remains in one cluster and is never separated.

4.3 Spatial autocorrelation

From the point data analysis, it was clear that both crimes committed and amenities were highly concentrated in areas with high population. To countermeasure this we aggregated the crimes into NTA polygons to make variables like, crimes per capita, crimes per bar within an NTA etc.

With this we could measure if there are groups of NTA's where crimes etc. are located when normalized by population, or if there are significant outliers. We did this by measuring the correlation between neighbouring NTA's with Moran's I and LISA from the ESDA library[14]. We chose Queen contiguity-based spatial weights for the computation. Our reasoning

is that we want to analyse connected polygons. The Manhattan polygons tend to form a grid, making polygons that touch in one point, which Rook based wouldn't have caught. Table 2 shows the results of Moran's I.

Variable	Moran's I score	p-value
Population density	0.362	0.001
Crime per capita	0.297	0.001
Bar per capita	0.587	0.001
Restaurant per capita	0.637	0.001
Station per capita	0.175	0.004
Park per capita	0.17	0.003
Crime per bar	0.303	0.003
Crime per restaurant	0.422	0.001
Crime per station	0.156	0.018
Crime per park	0.141	0.01

Table 2: Moran's I score for different variables in the NTA polygons.

Table 2 shows the Moran's I scores for a number of different variables. We included metrics like crime per bar to identify NTA's where there might be many bars but few crimes or vice versa. In general the scores are positive, meaning no strong dispersion tendency. Stations per capita and parks per capita scores are low indicating a more random distribution, whereas bars and restaurants per capita show a higher clustering tendency.

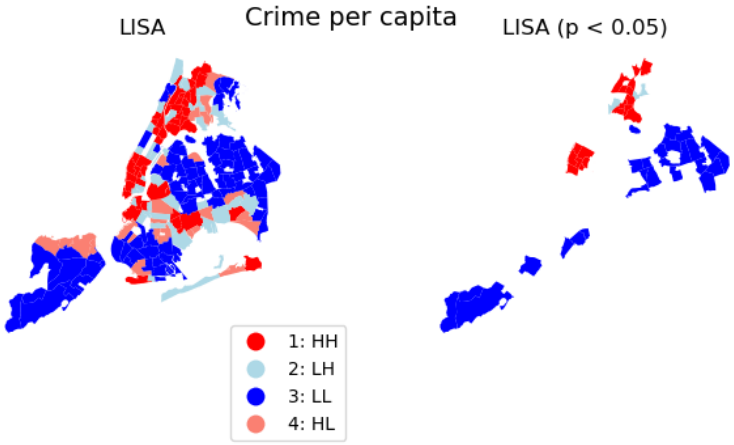


Figure 5: The result of computing LISA on NTA polygons with crime per capita as variable.

Figure 5 shows the result of computing LISA on the polygons with crime per capita as variable. All LISA results can be found in appendix I. A pattern we recognized in crimes and amenities per capita is similar to what

we saw in the point clusters. For instance, there are many NTA's grouped in lower Manhattan with high crime and amenity counts resulting in High-High LISA scores. Something different we noticed here is that Bronx tends to have Low-Low scores for bars and restaurants but High-High scores for crime per capita.

When comparing the population density result to the crime per capita result, they tend to agree in having High-High values in lower Manhattan and Bronx, but there are a few counter examples of statistically significant ($p < 0.05$) Low-High values in Bronx in the crime per capita result indicating outliers with high population density but low crime rate.

Restaurant per capita and bar per capita show Low-Low values in Bronx where crime per capita has High-High values, indicating that other factors must have had an influence. The crime per restaurant results show High-High values in Bronx confirming the high crime to restaurant ratio, whereas Manhattan has Low-Low values meaning more restaurants per crime.

4.4 Exploratory findings

We wanted to include more analysis using bi-variate methods, but found them to be out of scope for this course. So rather than skipping this desire, we ended up including this very small section, that doesn't use ESDA tools, but just our own simple bi-variate exploration results combined with Pearson R correlation analysis. As such, these results are not included in the discussion later. Therefore it can be seen as a secondary investigation into the hypotheses mentioned in the background.

For each of these comparisons we define "high" as being in the top 20 percent of the variables own dataset, and "low" as bottom 20 percentage. For the scatter plots, the variables have been normalized.

Each of these comparisons is a direct matching of a subsection in the background section.

4.4.1 Misdemeanours at stations?

Most crimes at stations are petty larceny, being a misdemeanour. Therefore we compare misdemeanours per capita to stations per capita.

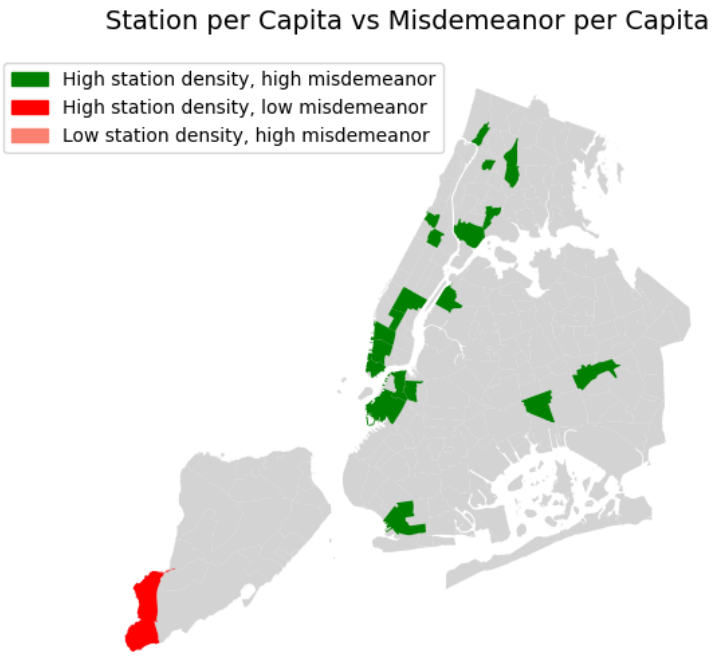


Figure 6: NTA's where the number of misdemeanours per capita and number of stations per capita are high or low.

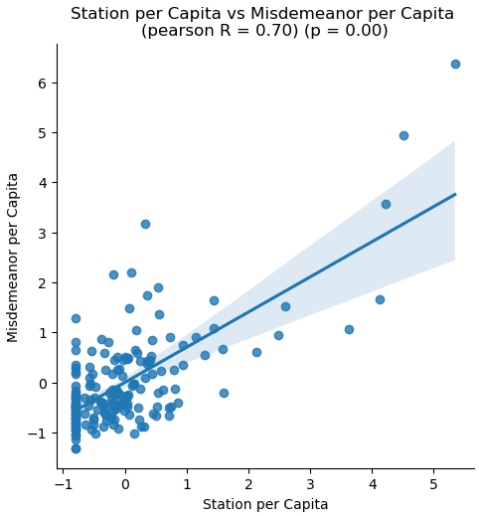


Figure 7: Stations per capita correlated with misdemeanours per capita across NTA's.

Bare in mind, that even if both variables completely matched for every single NTA, then a maximum of 20 percentage of the NTA's could be marked green. Then it is clear that Figure 6 agrees with our sources, and shows that places

with many stations also have a lot of misdemeanour crimes, and only a single counter-example where a neighbourhood has many stations but low crime. This is also confirmed in Figure 7, showing the two variables to be significantly linearly correlated.

4.4.2 Misdemeanours in dense populations?

Many of our sources mentioned the density of people being a crime relevant factor. Therefore we compare the population density of the NTA's against crime rate.

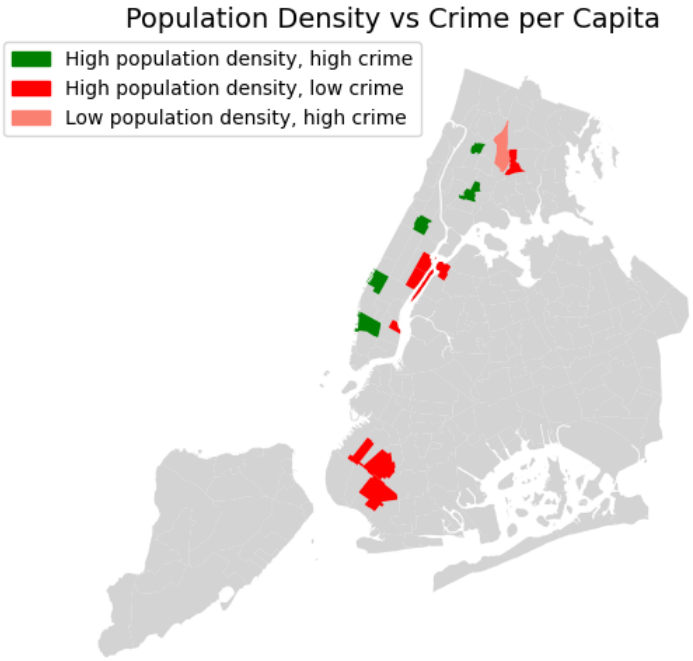


Figure 8: NTA's where the number of crimes per capita and population density are high, as well as counter examples of the opposite.

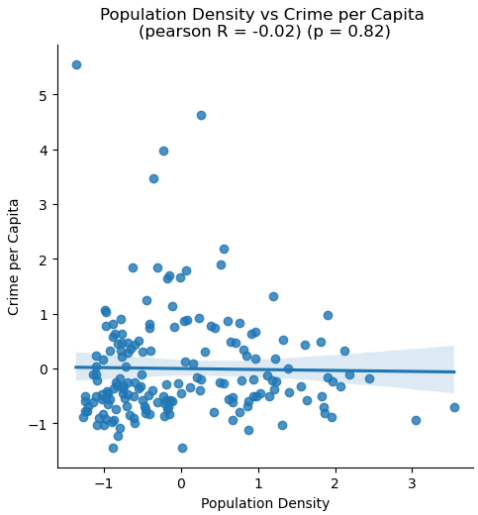


Figure 9: Population density correlated with crime per capita across NTA's.

Our result seen in Figure 8 does not concur with this conclusion. There are more counter-examples, such as high crime rate for the least dense neighbourhoods, than there are examples supporting this hypothesis. We also see from Figure 9, that there is no significant correlation between population density and crime per capita.

4.4.3 Misdemeanours at Bars?

Some of the most frequent crime related problems with bars are public intoxications and assaults, which are both typically classified as misdemeanours. Therefore, we compare misdemeanours per capita to bars per capita.

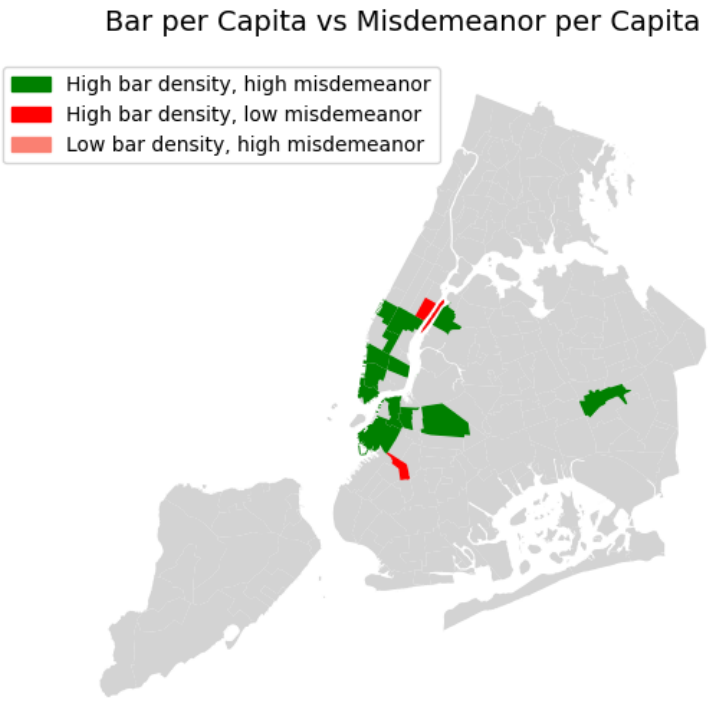


Figure 10: NTA's where the number of misdemeanours per capita and number of bars per capita are high, as well as counter examples of the opposite.

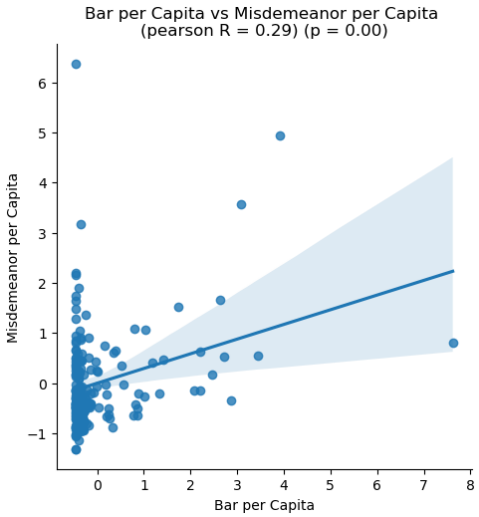


Figure 11: Bars per capita correlated with misdemeanours per capita across NTA's.

Our result seen in Figure 10 agrees with this conclusion. There are a

lot of neighbourhoods where there is a lot of crime and bars. There are no counter examples of neighbourhoods where there was low bar density and high amount misdemeanours. Figure 11 also shows a significant mild linear correlation, supporting this hypothesis.

4.4.4 Low crime rates around restaurants?

The "eyes on the street" theory specifically mentions that many restaurants and their related busy streets, contributes to low crime. Therefore we compared restaurants per capita to crimes per capita.

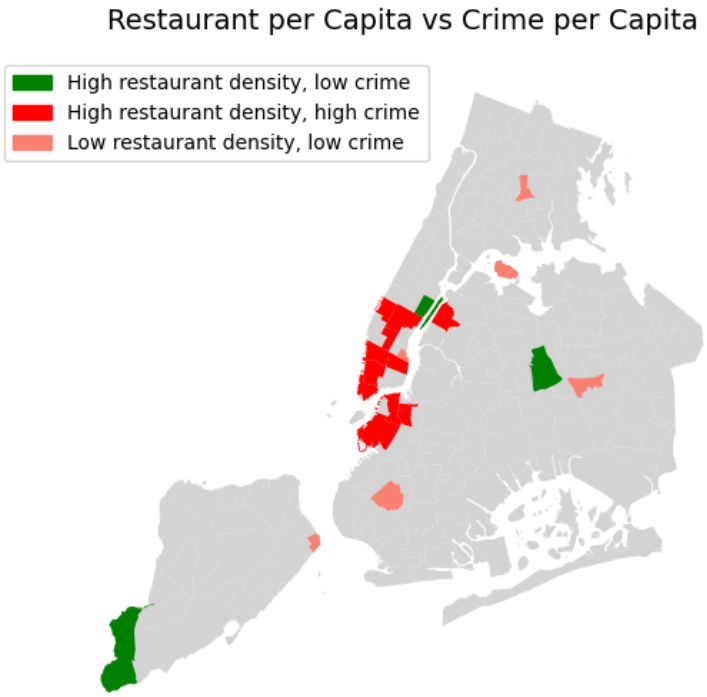


Figure 12: NTA's where the number of restaurants per capita is high and number of crimes per capita are low, as well as counter examples of the opposite.

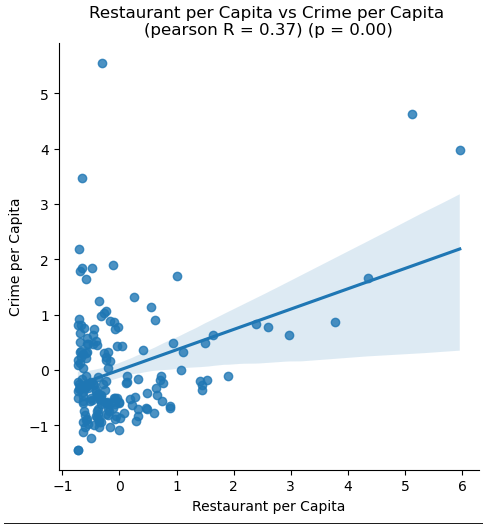


Figure 13: Restaurants per capita correlated with crime per capita across NTA's.

Our results seen in Figure 12 does not agree with this part of the theory. We have many neighbourhoods that have both a high amount of restaurants and crime. But our Pearson R seen in Figure 13, does find a significant positive linear correlation, where the theory expected it to be negative, and thus our Pearson R result does not support the hypothesis.

5 Discussion

5.1 The point pattern analyses

The point pattern analyses helped get an overview of our data points. It confirmed that crimes tends to occur in some places more than others. In other words they are clustered rather than dispersed.

However, it had some limitations. All the point data had a tendency to cluster in the densely populated places. This could be seen both from scatter plots, KDE heat maps and the point clustering analysis. A common mistake when making maps with this kind of data, is that they end up being "basically just population maps"[20]. This makes it hard to conclude anything other than where there are a lot of people. The NTA aggregation was therefore necessary and allowed us to divide by population to get results that avoided this issue.

However, this does not mean the cluster analysis was mostly useless. For instance, it could have been the case that crimes mostly happen outside the city where there are fewer people to witness and less surveillance. Our point analysis helped deny that theory.

5.2 Population vs people

A very significant limitation of our analysis is that it uses population as a metric. It does not include people who visit the places but are not residents there. This means it doesn't include for instance tourists and people who go to places for work during the day while living elsewhere. Having this data could yield different results for crowded areas around new york city like busy stations and tourist attractions etc.

5.3 Parks

Initially we also had parks included in our focus area. However, there were a number of problems regarding the parks that made the analysis difficult.

- Many parks don't have a population. And thus we cannot use our "per capita" metrics.
- According to the NYPD data foot notes "Offenses occurring in open areas such as parks or beaches may be geo-coded as occurring on streets or intersections bordering the area"[12], which makes it difficult to tell how many crimes actually occurred in the parks.
- The ∼2000 parks that we fetched from OSM include small polygons that together make up larger parks. This is explained in detail in appendix J. Getting a better representation of where parks are located, would require more data pre-processing.

We judged these problems to be too challenging, and decided to focus more on the other amenities instead.

5.4 The spatial autocorrelation

It is important to keep in mind that our Moran and LISA results are affected by how the polygons are divided, like the modifiable areal unit problem (MAUP). Densely populated areas tend to be more divided into smaller polygons resulting in many neighbouring polygons with high amenity and crime scores. Less dense areas are less divided resulting in large polygons, fewer neighbours and lower significance. Staten island for instance, generally has larger polygons than Manhattan.

The LISA analysis helped identify larger areas with a general high population density like Manhattan vs low density areas like the east coast and Staten Island. This was often also the areas with high crime rates. The LISA analysis did however reveal counter examples of some NTA's being highly populated but having relatively low crime rates in Bronx. Furthermore, it helped reveal areas with few bars and restaurants per capita but high crime rates.

LISA cannot be used to analyse how two different values correlate with each other. While the tool was useful it required some manual work of comparing the graphs. We focused on the LISA values that were statistically significant with a p-value below 0.05. Our findings may though be biased by the fact that humans tend to find patterns where there are none. Therefore we added the extra exploratory correlation analysis.

5.5 Threats to validity

We know a lot about WHAT, but almost no WHY. There might be no causation at all. Or even if there is a causation, we may have the order wrong as well. For all we know, it could be: High inequality can lead to dense zones of different economical ranks, leading to more crime, and more crime will lead to lower house prices. The problem is, with no WHY, it could even be the opposite way around, for example: higher house prices deter a certain segment of people who commit crimes often. The point is that crime is a really complex area with a wide range of factors. There may even be very different motivations for committing the same category of crime. Such as a misdemeanour being stealing food from a supermarket for oneself, versus drunken disorderly conduct such as jaywalking across the street, to get home faster.

As an example, of how other variables might be the causation, in Figure 14, we can see how it looks like house prices also correlates with crime negatively. But in reality, we don't know the causation of it all.

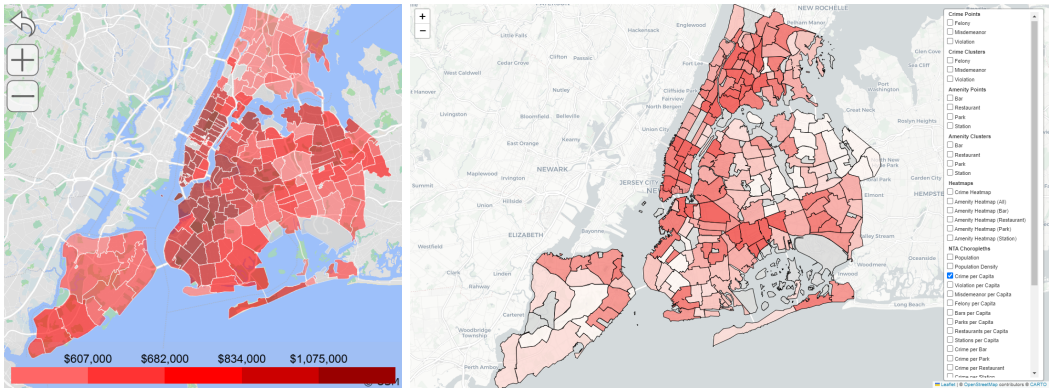


Figure 14: Average housing prices in NYC zip code areas. Source: [18]. Vs Crime rates per NTA.

5.6 For future analysis

This project was limited to a few variables in a single city. Further analysis could include accounting for many more variables, specifically, economic variables such as housing prices, income, and purchasing power. Luckily,

this data is available as seen in Figure 14. Other variables one could extend the analysis with are demographics such as age, gender, etc. as well as time of the crime. The NYC data includes the exact date and time of the crimes, so we could include these to see if there is a temporal spatial correlation of certain crimes near certain amenities.

For more general conclusions on this research topic, one ought to look at more than just one city. The analysis could be replicated in other US. cities, and the results compared. Furthermore, it would be interesting to compare US. cities to other worldwide cities, as this would give an indication of how much culture, or general city design plays a role in the correlation of crimes and amenities.

Finally, to establish the actual "why", one would need to look at causations. This might include the use interdisciplinary techniques such as Probabilistic Programming etc., but also using more qualitative data rather than quantitative data.

6 Conclusion

Our analysis indicates spatial correlations between crimes and particular amenities, notably bars and railway stations.

Although we established meaningful crime clustering around the aforementioned bars and railway stations, we found no consistent evidence supporting Jane Jacobs' "eyes on the street" theory, as NTA's with a higher density of restaurants did not have a strong correlation with reduced crime rates.

Our research and relative weakness of observed correlations suggest complex interactions involving additional unexplored variables. Furthermore, it is important to remember that this is only based on data across New York City. Future studies should extend the range to other cities and countries, and include other factors such as time of crime, socioeconomics, and cultural patterns etc.

With this paper, we have shown how complex crime is and that there is a lot more research to be done on this very interesting topic in the future.

References

[1] aidalaw. 3 things you should know if you get into a bar fight. https://aidalalaw.com/3-things-you-should-know-if-you-get-into-a-bar-fight/, 2023. [Online; accessed 22-May-2025].

[2] Department of City Planning (DCP). New york city population by neighborhood tabulation areas. https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Neighborhood-Tabulatio/swpk-hqdp/about_data, 2020. [Online; accessed 24-April-2025].

[3] Department of City Planning (DCP). 2020 neighborhood tabulation areas (ntas). https://data.cityofnewyork.us/City-Government/2020-Neighborhood-Tabulation-Areas-NTAs-/9nt8-h7nd/about_data, 2025. [Online; accessed 24-April-2025].

[4] M. Fleischmann. Introduction of geodatasets. https://geodatasets.readthedocs.io/en/latest/introduction.html, 2024. [Online; accessed 21-May-2025].

[5] Z. Golden. Eyes on the street: Testing jane jacobs. https://equilibriumecon.wisc.edu/2024/07/17/eq-vol-14-eyes-on-the-street-testing-jane-jacobs/, 2024. [Online; accessed 22-May-2025].

[6] R. A. Hughes. Italy, france, spain: Which european country is worst for pickpockets? https://www.euronews.com/travel/2023/09/05/italy-france-spain-which-european-country-is-worst-for-pickpockets, 2023. [Online; accessed 22-May-2025].

[7] J. Jacobs. *The Death and Life of Great American Cities*. Vintage Books, A Division of Random House, inc., New York, 1961.

[8] Leagal Information Institute. petty larceny. https://www.law.cornell.edu/wex/petty_larceny, 2020. [Online; accessed 22-May-2025].

[9] Maverick Meerkat. Cluster tendency using hopkins statistic implementation in python. https://datascience.stackexchange.com/questions/14142/cluster-tendency-using-hopkins-statistic-implementation-in-python, 2023. [Online; accessed 15-May-2025].

[10] T. McNicholas. Nypd data shows decline in major crimes in transit system, rise in misdemeanor assaults and petit larceny. https:

//www.cbsnews.com/newyork/news/nypd-mta-transit-crime/?utm_source=chatgpt.com, 2023. [Online; accessed 22-May-2025].

[11] OpenStreetMap. Openstreetmap amenities. https://wiki.openstreetmap.org/wiki/Key:amenity, 2025. [Online; accessed 08-May-2025].

[12] Police Department (NYPD). Nypd complaint data historic. https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i/about_data, 2024. [Online; accessed 10-April-2025].

[13] P. Pridal, T. Pohanka, A. Ashraf, and R. Kacer. Epsg:3857. https://epsg.io/3857, 2020. [Online; accessed 21-May-2025].

[14] pysal developers. Esda: Exploratory spatial data analysis. https://pysal.org/esda/, 2018. [Online; accessed 20-May-2025].

[15] pysal developers. pointpats.l - ripley's l function. https://pysal.org/pointpats/generated/pointpats.l.html, 2018. [Online; accessed 15-May-2025].

[16] scikit-learn developers. 2.3 clustering. https://scikit-learn.org/stable/modules/clustering.html, 2025. [Online; accessed 20-May-2025].

[17] M. S. Scott and K. Dedel. Crime rate by country 2025. https://popcenter.asu.edu/content/assaults-and-around-bars-2nd-ed, 2006. [Online; accessed 13-May-2025].

[18] simplemaps. Home value. https://simplemaps.com/city/new-york/zips/home-value, 2019. [Online; accessed 22-May-2025].

[19] R. Story. Folium documentation. https://python-visualization.github.io/folium/latest/index.html, 2025. [Online; accessed 20-May-2025].

[20] The Map Room. The end of maps in seven charts. https://www.maproomblog.com/2016/02/the-end-of-maps-in-seven-charts/, 2016. [Online; accessed 20-May-2025].

[21] Wikipedia. Public intoxication. https://en.wikipedia.org/wiki/Public_intoxication, 2025. [Online; accessed 22-May-2025].

[22] World Population Review. Crime rate by state 2025. https://worldpopulationreview.com/state-rankings/crime-rate-by-state, 2025. [Online; accessed 13-May-2025].

[23] H. Zhang, R. Zahnow, Y. Liu, and J. Corcoran. Crime at train stations: The role of passenger presence. *Applied Geography*, 140:102666, 2022.

Appendix

A Meta data tables

The tables shows only the columns from the different datasets that were used for this project. For information about all the columns, we refer to the respective sources from where the data sets were obtained.

NYC Crime data

Source: [12]

Variable Name	Explanation
CMPLNT_FR_D	Exact date of occurrence for the reported event (or starting date of occurrence, if it spans multiple days)
LAW_CAT_CD	Level of offense: felony, misdemeanor, violation
Latitude	Midblock Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude	Midblock Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

Table 3: Meta data table of the crime data set.

Amenity data

Source: [11]

Variable Name	Explanation
geometry	The geometric representation of the object
leisure	Denotes places where people go in their spare time. E.g. park
amenity	Describes useful and important facilities for visitors and resident. E.g. bar and restaurant
railway	Used to tag rails and infrastructure for many kinds of railways. railway=station is a railway facility where trains stop to load and unload passengers and/or freight

Table 4: Meta data table of the amenity data set.

NTA polygon data

Source: [3]

Variable Name	Explanation
the_geom	The geometric representation of the NTA as a GeoJSON multipolygon
BoroName	Name of the borough that the NTA is located in
NTA2020	2020 Neighborhood Tabulation Area Code.
NTAName	2020 Neighborhood Tabulation Area Name

Table 5: Meta data table of the NTA data set with multi polygons.

NTA population data

Source: [2]

Variable Name	Explanation
Borough	Name of the borough that the NTA is located in
Year	The year of the population data acquisition. (Includes 2000 and 2010 data)
NTA Name	Neighborhood Tabulation Area Name
Population	Number of residents living in the NTA

Table 6: Meta data table of the NTA data set with population.

B Contribution statement

The entire project was done in close collaboration with all three group members involved, making it hard to determine primary/secondary contributors on individual sections. However, since it is a requirement, we have made an estimate of who spend a majority of their time on what.

Data collection and pre-processing

Primary person: Jacob
- Secondary person: Joakim
- Tertiary person: Thor

Interactive map

Primary person: Jacob
- Secondary person: Thor
- Tertiary person: Joakim

Clustering analysis

Primary person: Joakim
- Secondary person: Thor
- Tertiary person: Jacob

Spatial autocorrelation

Primary person: Thor
- Secondary person: Jacob
- Tertiary person: Joakim

Background & Exploratory analysis

Primary person: Jacob
- Secondary person: Joakim
- Tertiary person: Thor

C Code repository url

https://github.com/Grumlebob/ExamGDS

D NTA population density distribution

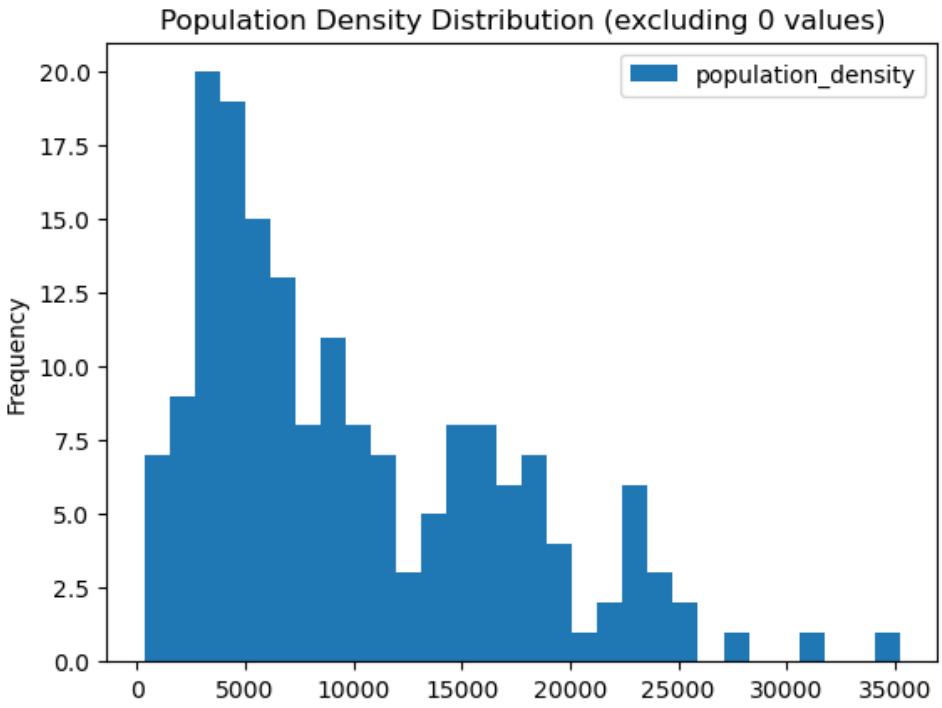


Figure 15: The distribution of populated NTA's by their population density.

E Scatter plot of the crime data

Figure 16: Scatter plot of the crime data points that was used throughout the project.

F Ripley's L tool giving strange results

According to the pointpats's documentation[15], random point patterns should be close to 0 for all distances. However, as seen in Figure 17 our result showed a growing value for increasing distances for both crime data points and randomly simulated samples. We tried setting the `linearized` parameter to true, but this resulted in a curve that went towards large negative numbers as distance increased.
Code snippet used to obtain data for figure 17:

```
result = pointpats.l_test(
    coordinates=points,
    support=support,
    n_simulations=n_sims,
    keep_simulations=True,
    hull=hull)
```

If for some reason it were showing the results of Ripley's K, we would expect the median simulation value to go towards the same value as the crime points value. Since they have an equal number of points, the number of neighbours at max distance would be equal to the number of points. This is however, not the case either.

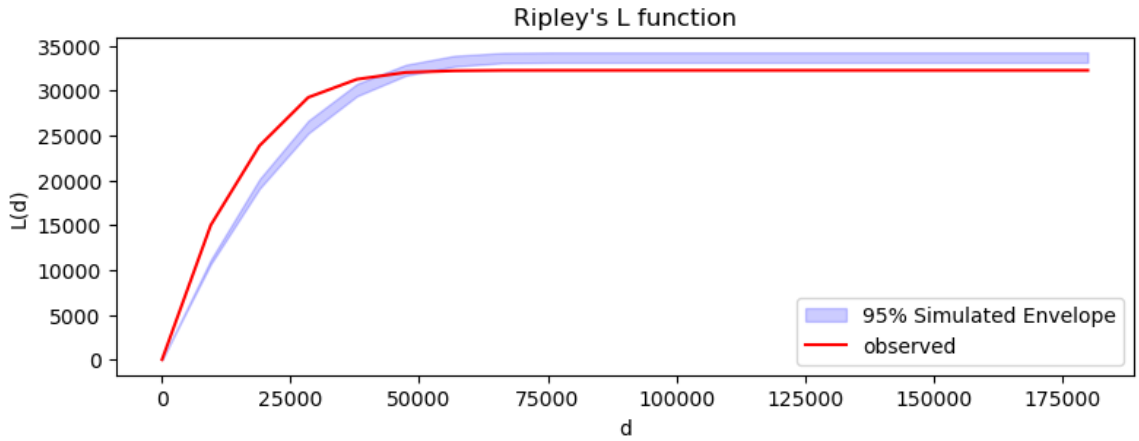


Figure 17: The result of using Ripley's L on points from the crime data.

G Ripley's K results

Figure 18-22 show the results we got from computing the Ripley's K function on our data. Distance metric d is in meters. 99 simulations were used to create the expected values area.

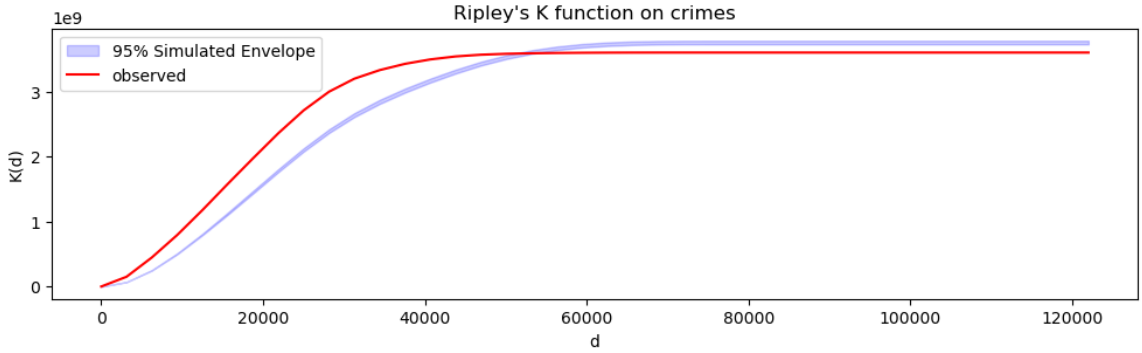


Figure 18: Graph showing Ripley's K score on crime data vs randomly simulated data, as distance d increases.

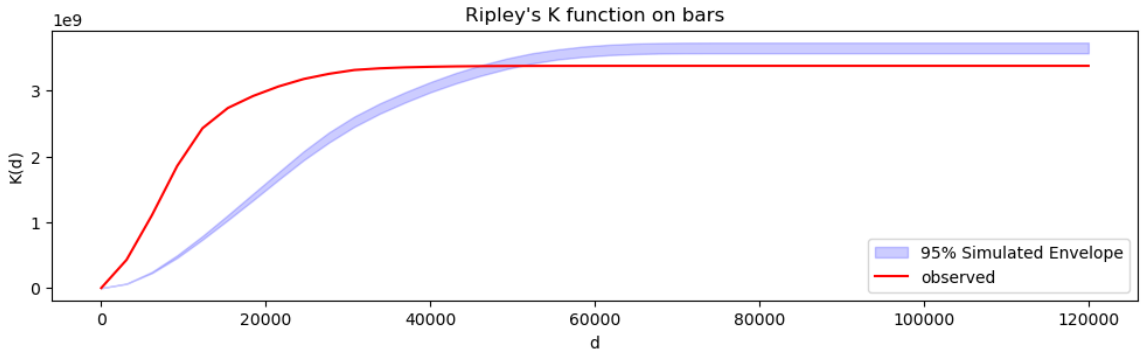


Figure 19: Graph showing Ripley's K score on bar data vs randomly simulated data, as distance d increases.

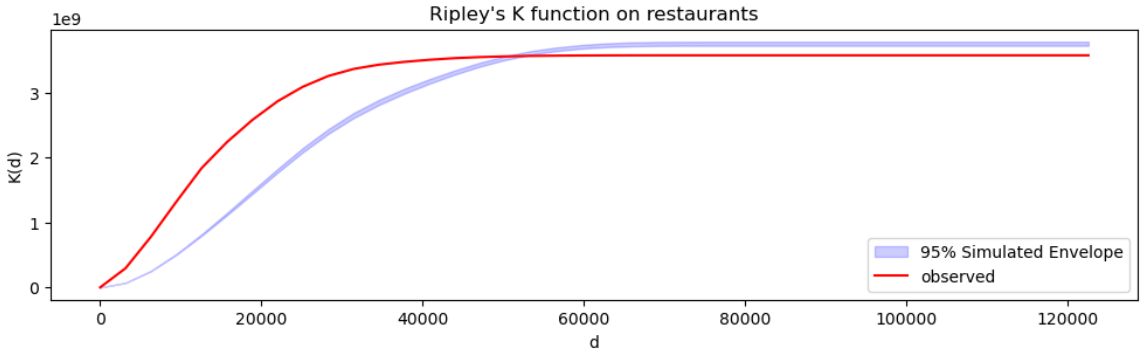


Figure 20: Graph showing Ripley's K score on restaurant data vs randomly simulated data, as distance d increases.

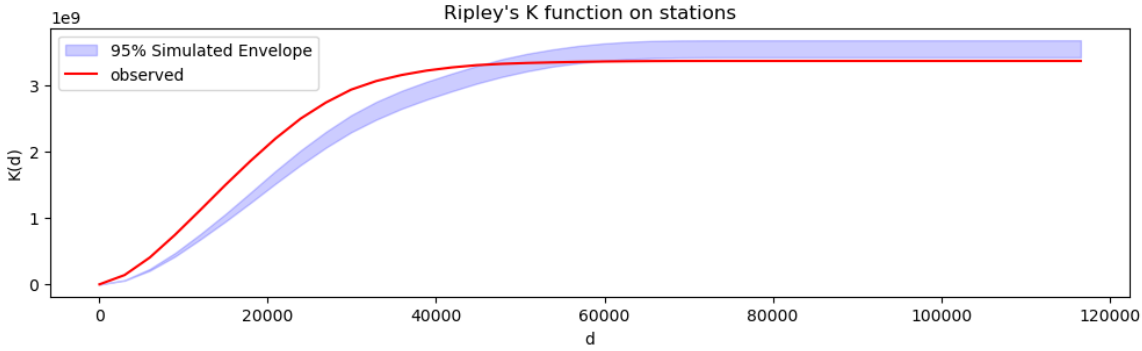


Figure 21: Graph showing Ripley's K score on station data vs randomly simulated data, as distance d increases.

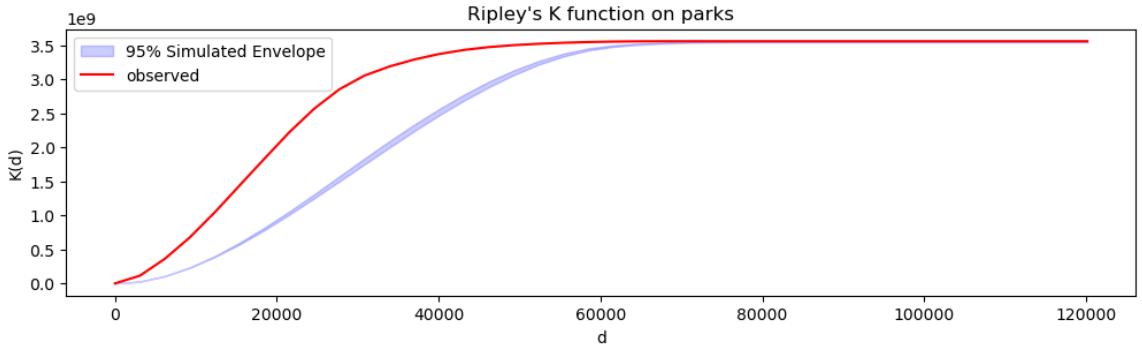


Figure 22: Graph showing Ripley's K score on park data vs randomly simulated data, as distance d increases.

H Clusters plotted

Figure 23 to 27 shows the clusters obtained with agglomerative and means shift clustering. The number of clusters for agglomerative and the bandwidth values for mean shift were found experimentally.

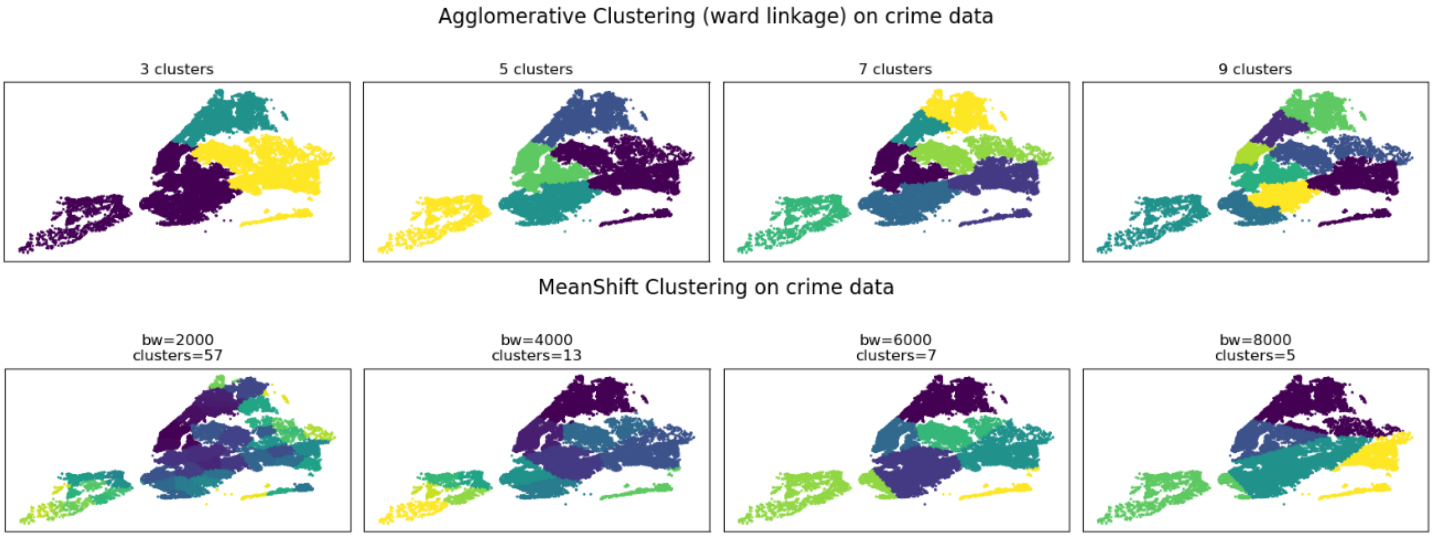


Figure 23: The clusters computed for the crime data.

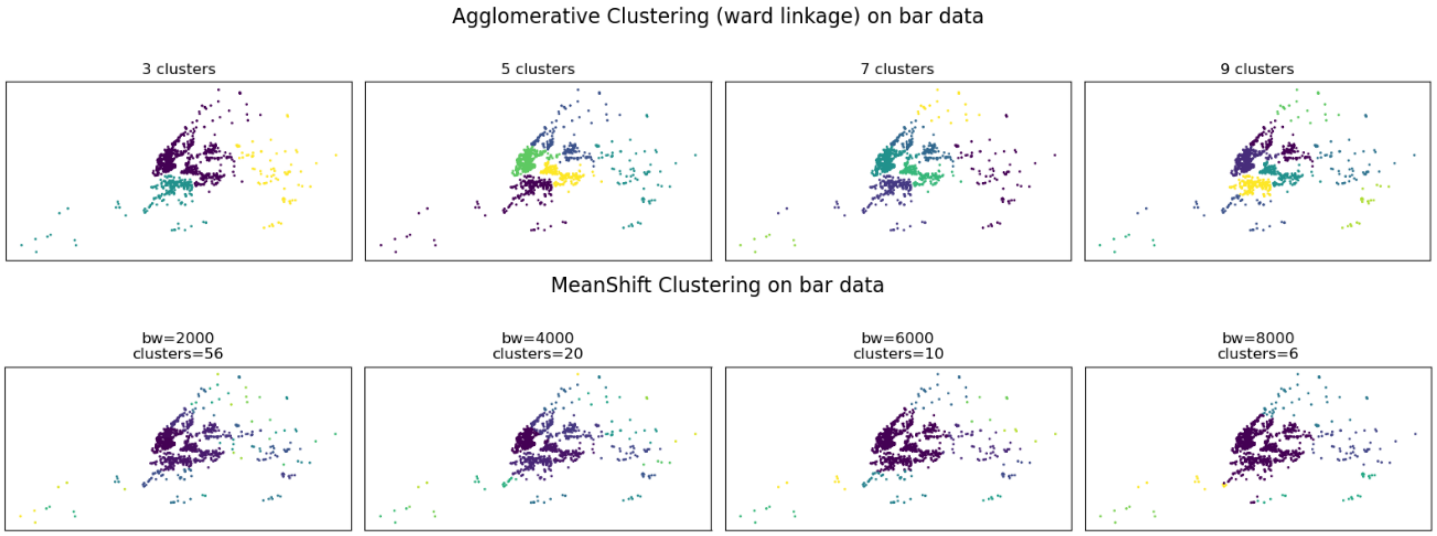


Figure 24: The clusters computed for the bar data.

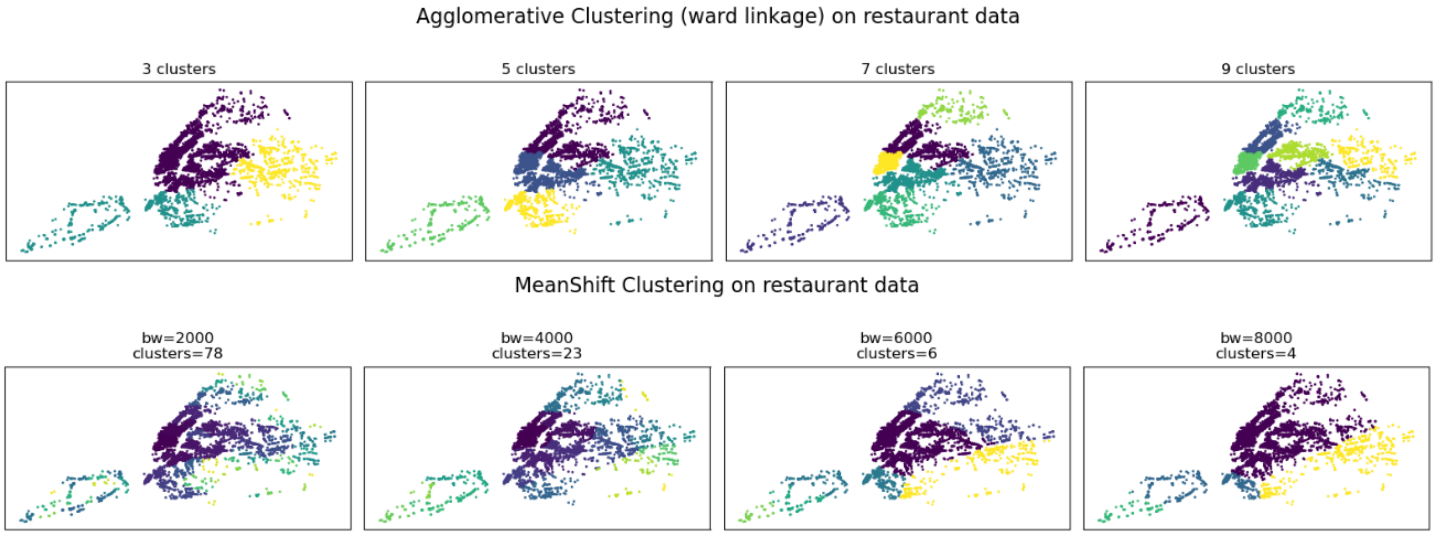


Figure 25: The clusters computed for the restaurant data.

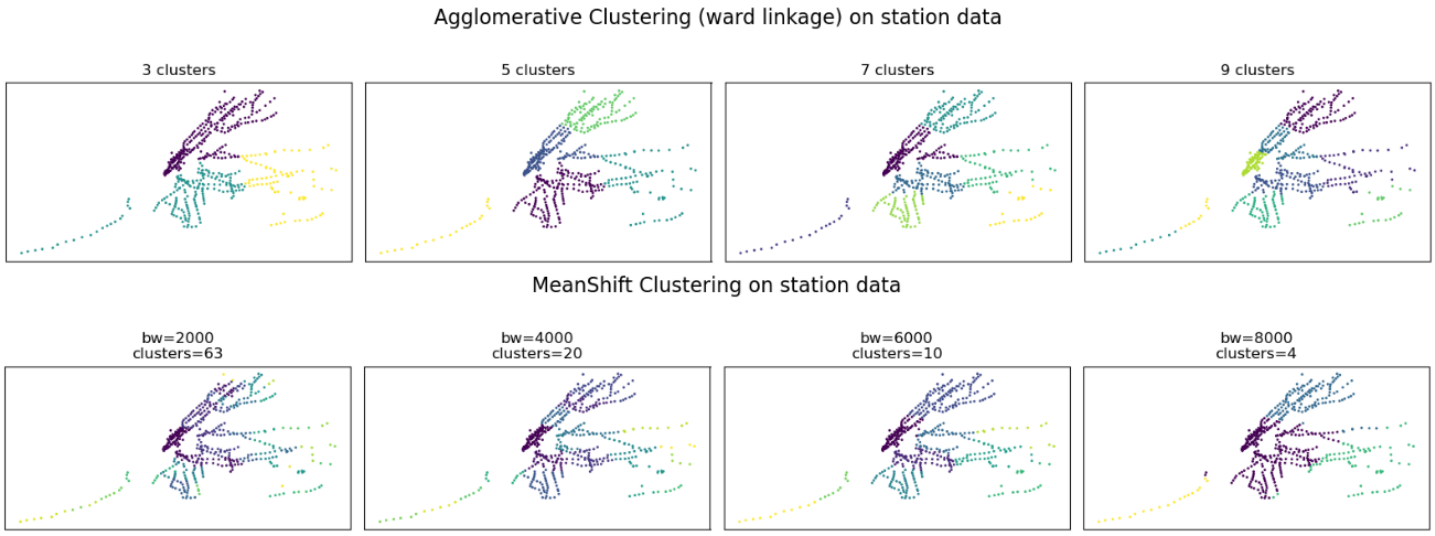


Figure 26: The clusters computed for the station data.

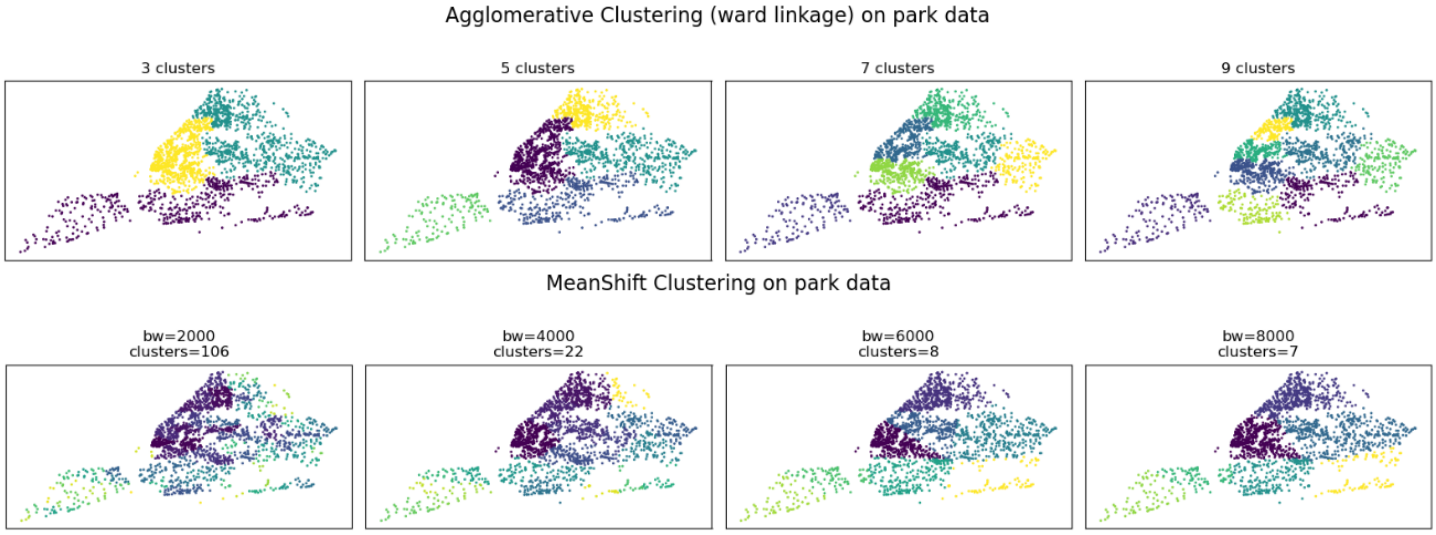


Figure 27: The clusters computed for the park data.

I LISA results

Figure 28-37 show the results of computing LISA on the NTA polygons with different variables. Missing polygons on the left is caused by some polygons having 0-values, like no population or no stations etc.

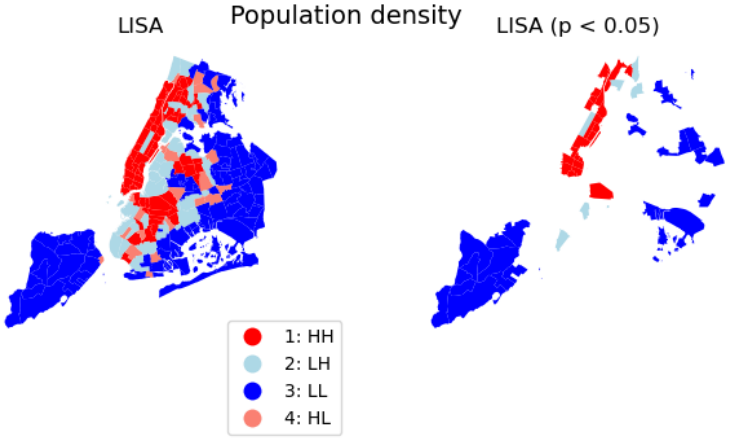


Figure 28: Result of running LISA with population density as variable.

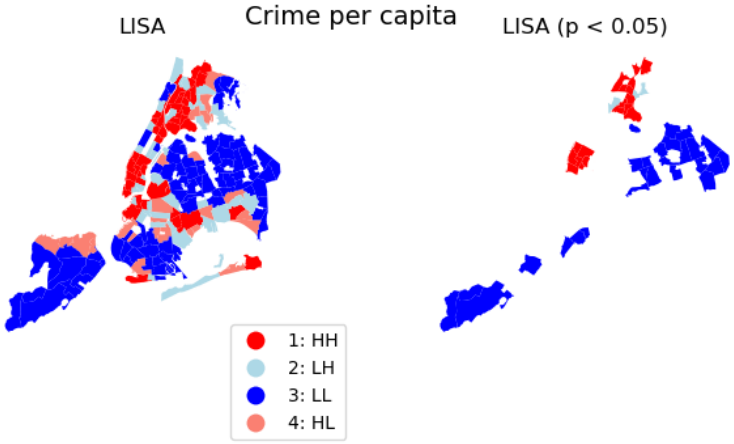


Figure 29: Result of running LISA with crime per capita as variable.

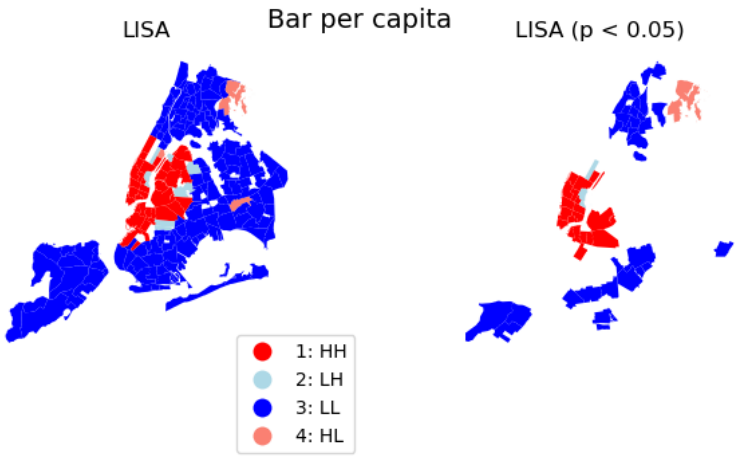


Figure 30: Result of running LISA with bar per capita as variable.

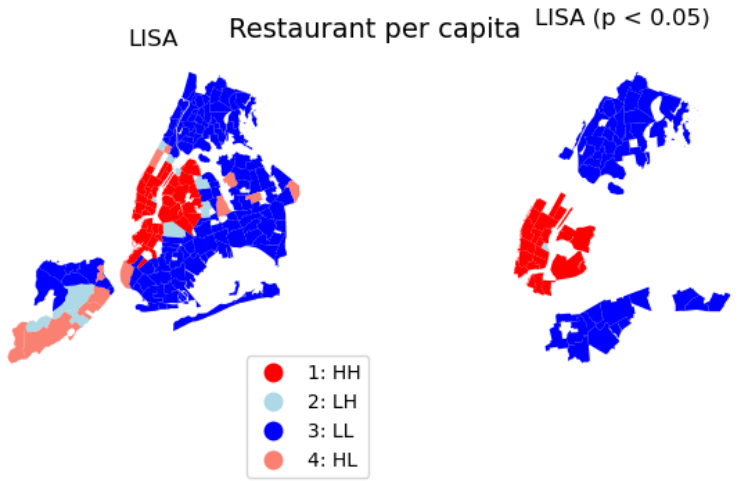


Figure 31: Result of running LISA with restaurant per capita as variable.

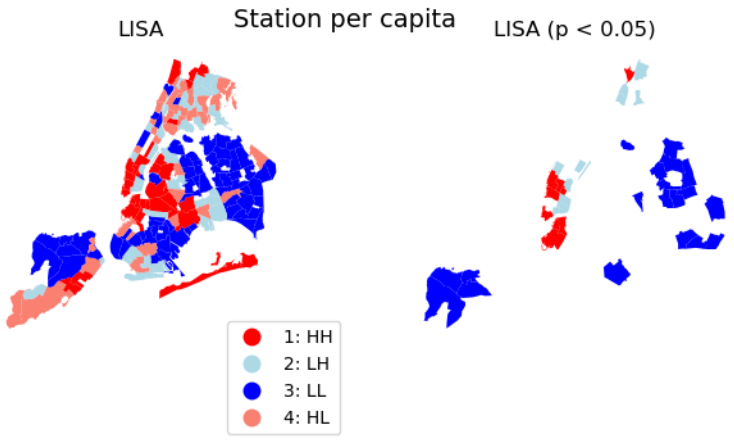


Figure 32: Result of running LISA with station per capita as variable.

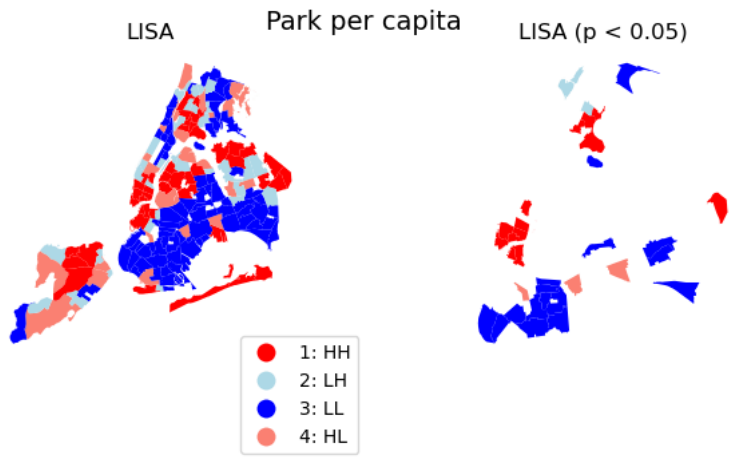


Figure 33: Result of running LISA with park per capita as variable.

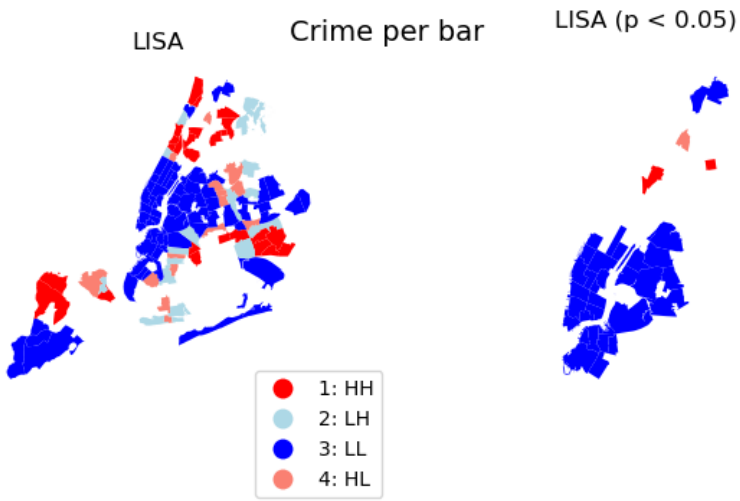


Figure 34: Result of running LISA with crime per bar as variable.

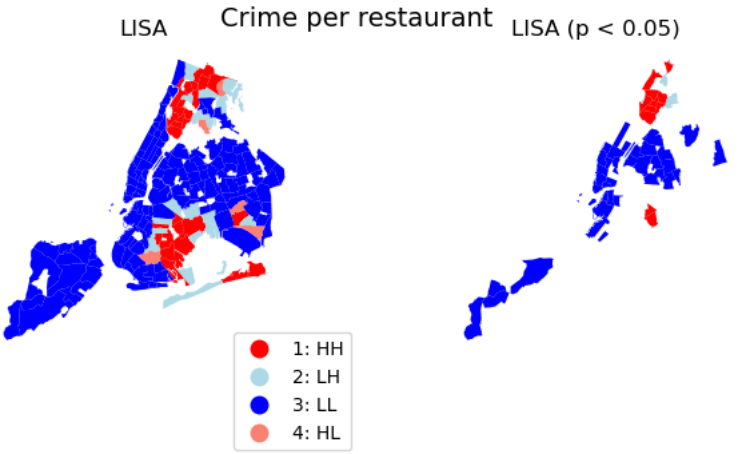


Figure 35: Result of running LISA with crime per restaurant as variable.

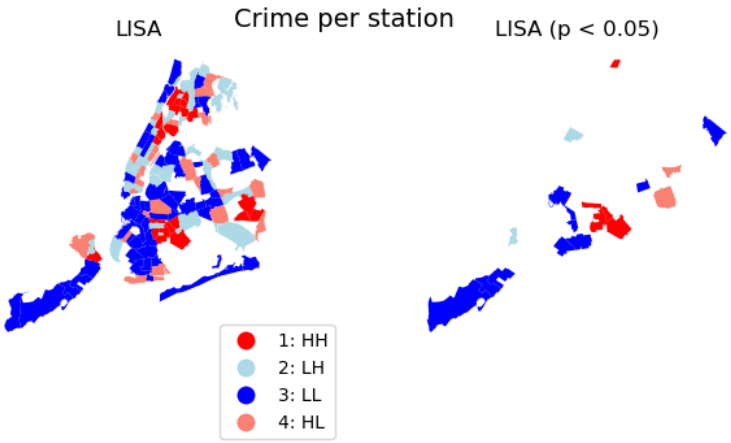


Figure 36: Result of running LISA with crime per station as variable.

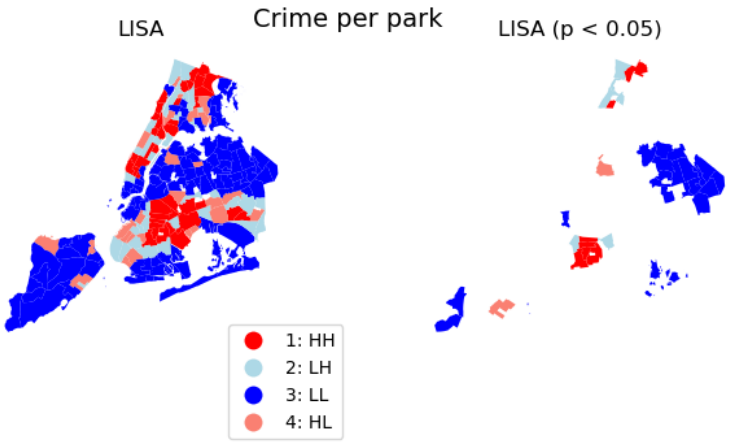


Figure 37: Result of running LISA with crime per park as variable.

J Open Street Map being too granular

We fetched over 2000 parks from Open Street Map, which seemed overwhelming. We investigated our interactive map and found that our "parks" also included polygons that make up parks. Figure 38 shows an example of a tiny park that makes up nine rows in our dataset. This observation limits the usefulness of the OSM park data.

Figure 38: Klapper Circle is around 50 by 30 meters, and counts as nine parks in our dataset.