Lions and Tigers and…Rats?! Oh My!

A Visual Exploration of New York City Rats

Jessica Guan

Abstract

In the spring of 2023, the New York City mayor appointed the city's first ever "rat czar", a role specifically created to spearhead rat reduction. In this paper, the rat population in New York City will be explored, including factors that might affect the rat density in different areas such as subway locations and restaurant density. This is a visual analysis that uses datasets on New York City rat sightings, subway infrastructure, and restaurants. With focuses on visualisations, spatial analysis, temporal analysis, and modelling will be used to explore the nature of New York City rats.

I. Problem Statement

New York City is home to around 8.4 million New Yorkers. When people think of New York City, they often think of its impressive urban sprawl of people, public transportation, shopping centres, restaurants–and rats. An urban myth says that for every person in New York City there is a rat. In reality, the rat to population ratio is closer to one third. Although the urban myth has been debunked, the New York City rat population is still large and thriving, but what factors of the city contribute to this number?

With a focus on visual analysis, this paper will explore the NYC rat population by answering the following questions:

1. What are the time patterns of rat sightings in NYC?
2. What areas have the most rats?
3. Do factors such as subway infrastructure and restaurants play influence the rat population?

This analysis will use geospatial and temporal data on rat sightings to address the time patterns and geographic spread of rats. It will also use geospatial data on restaurants and the subway infrastructure and the data on population density and socioeconomics to investigate contributing factors to the rat population.

II. State of the Art

Rats have long been a hot topic in New York City. In the first paper analysed, Jonathan Auerbach aims to answer the urban myth of the rat population having a one to one ratio with New Yorkers. In his paper, he points out the difficulty of counting the real number of rats in the city. Unlike humans, rats do not participate in an annual census. Alternatively, he looks to the recapture estimation method, a method used by wildlife researchers in which animal populations are measured in two samples. In the first sample, the animals are captured, marked, and released. Then, a second round of capturing is performed. It is assumed that because they are released, it is equally as likely that the marked animals are recaptured. Using these two samples, the population is estimated. However, the New York City government is unlikely to approve an experiment like this. Instead, Auerbach uses the rat sightings data (2010 - 2014) to perform a pseudo recapture estimate by using the data from the first half of 2010 as the first sample and the data from the first half of 2011 as the second sample. With this method, he visualises the data with bar graphs and choropleth maps, and estimates the rat population to be around two million [1].

In the second paper, Michael G. Walsh explores the correlation between New York City rats and sociodemographics and housing. He uses kernel density estimation to visualise rat sightings in proximity to factors such as subway lines, housing units, open spaces, and human population with density and choropleth maps. He also implements an autoregressive model to predict geospatial rat counts across the city [2].

In the third paper, Helen Bailey et al studies dolphin sightings with generalised estimating equations and temporal modelling. They model the temporal data of dolphin sightings by time of day, day of year, and tidal wave measure [3].

These studies will help in this analysis. From the first paper, bar graphs and choropleth maps will be used to visualise the geospatial data of rat sightings by zip code in New York City. Then, from the second paper, density maps will be used to visualise the correlation of rat sightings, subway infrastructure, and restaurants. To build on top of Walsh's autoregressive model, we will implement an autoregressive moving average (ARMA) model and a visualisation of the model to predict rat count. An ARMA model combines an autoregressive model and a moving average model to robustly capture time series data. Lastly, from the third paper, temporal analysis with time series graphs and temporal heat maps will be used to investigate the time patterns of rat sightings.

III. Properties of the Data

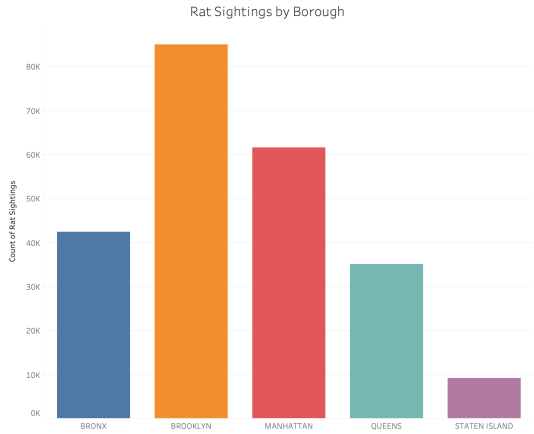
This analysis uses NYC Open Data datasets on rat sightings from 2010 to 2023 [4], subway line and entrances [6], and restaurant inspections from 2015 to 2023 [5]. It will also use a dataset on New York City's population numbers [7]. NYC Open Data is run by New York City agencies to provide public data about the city. The rat sightings dataset has 233,527 total sightings recorded over the thirteen year span. The recorded date and time is provided which will be used in the temporal analysis, and the latitude, longitude, and zip code is provided which will be used in the spatial analysis. The restaurants and human population datasets provide the zip code which will be used in a choropleth map. The subway dataset provides the latitude and longitude of each subway entrance and exit. Figure 1 shows the summary of numbers for each factor by borough. Figure 2 shows the aggregate number of rat sightings in the entire dataset by borough.

Figure 1: Summary of data by borough

Borough	Population	Population Density	Rat Sightings	Restaurants	Subway Stations
Bronx	1,382,480	1,077,441	42,431	18,399	238
Brooklyn	2,504,700	1,543,302	83,877	54,990	542
Manhattan	1,575,590	3,321,669	61,395	75,609	763
Queens	2,233,270	1,586,882	34,916	48,807	325
Staten Island	468,730	117,167	9,254	7,271	

Population Density, Population, Restaurants, Rat Sightings and Subway Stations broken down by Borough.

Figure 2: Bar graph - rat sightings by borough



Using python, the number of missing values was investigated in the rat sightings dataset. Of the relevant columns, there were 433 missing zip code values and 2266 missing latitude and longitude values. These null values will be excluded from the spatial analysis. There are no missing values for the recorded dates, so all values will be used for the temporal analysis. The restaurant dataset also has 2688 missing zip code values and 280 missing latitude and longitude values. Like for the rat sightings dataset, these null values will be excluded from the spatial analysis. The subway and human population dataset has no relevant missing values.

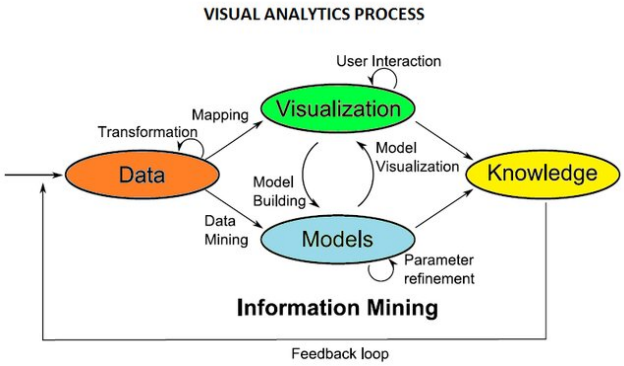
IV. Analysis

A. Analysis Approach

For this analysis, I will be using Python for data preprocessing and exploration. Python will also be used for the autoregressive moving average model. Then, I will be using Tableau to visualise the spatial and temporal graphs. After creating the visualisations, human interpretation and analysis

will be used to gather knowledge around the graphs. I will be following the process depicted in Figure 3 by Chate, Parinita & Patkar, Uday (2015) [8]. After data preprocessing, the visualisation and modelling will be implemented, and then human interpretation will be applied to gather knowledge. The process will be a feedback loop of these steps.

Figure 3: Visual analytics process diagram, diagram from Chate, Parinita & Patkar, Uday (2015) [8]



Spatial Analysis Approach

In the spatial analysis, the latitude, longitude, and zip code data will be used to visualise the rat sightings, subway infrastructure, and restaurants data in density maps and choropleth maps. These visualisations will be used to determine the influence of the subways and restaurants on the rat population throughout the city.

Temporal Analysis Approach

In the temporal analysis, the recorded dates in the rat sightings dataset will be used in time series graphs. The time patterns such as what time of year, what time of month, and what time of day has the most rat sightings will be explored. Temporal line graphs and temporal heatmaps will be created to investigate.

Modelling Approach

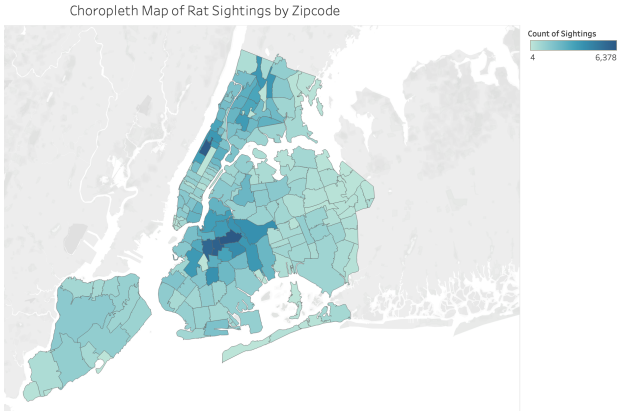
The occurrence of rat sightings over time are modelled using an autoregressive moving average (ARMA) model. This forecasting method will generate a predicted time series. Then, the performance will be evaluated by calculating the error between the predicted and actual data.

B. Analysis Process

Spatial Analysis

In the spatial analysis, let's first look at the distribution of rat sightings by borough as shown in Figure 2. Based on the aggregate data from 2010 to 2023, Brooklyn has the most number of rat sightings at 85,022, followed by Manhattan with 61,629 rat sightings. To understand the rat population more comprehensively, we'll cluster the rat sightings data by zip code in a choropleth map. The choropleth map in Figure 4 visualises the count of rat sightings geographically across New York City clustered by zip codes. The visualisation shows that most rat sightings occurred in zip codes 11221 and 10025 which are in Brooklyn and Manhattan respectively. The most dense zip codes are in Brooklyn. This aligns with the bar graph in Figure 2 in the initial data exploration. The third borough with the most rat sightings is the Bronx, which can also be visualised in the choropleth map where there are more darker sections in the Bronx than in Queens or Staten Island.

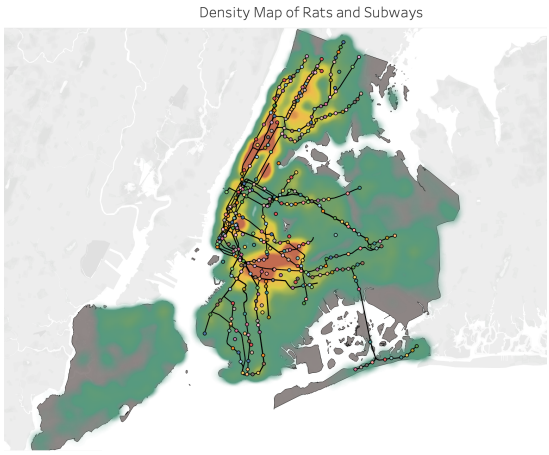
Figure 4: Choropleth map - rat data by zip code



After determining the zip codes with the most dense rat populations, let's look at the influence of the subway infrastructure on rats. Subways are often associated with rats, as it is difficult to take the subway regularly without spotting a rodent. Figure 5 shows the subway stations and paths data

overlaid on top of a density map of the rat sightings data. We can see that the subway infrastructure has an effect on the rat sightings, as the subway entrances and the lines are mostly in areas of the geographic map where it shows the most density. To get to Staten Island, people must take the ferry across the Hudson River, so there are no subway lines that go to the borough. Figure 5 shows that in addition to not having subway access on Staten Island, there is also the smallest density of rats, aligning with the idea that the subway infrastructure contributes to rat sightings. Manhattan and Brooklyn have the most number of subway stations and lines throughout, and the two boroughs also show the highest densities of rats.

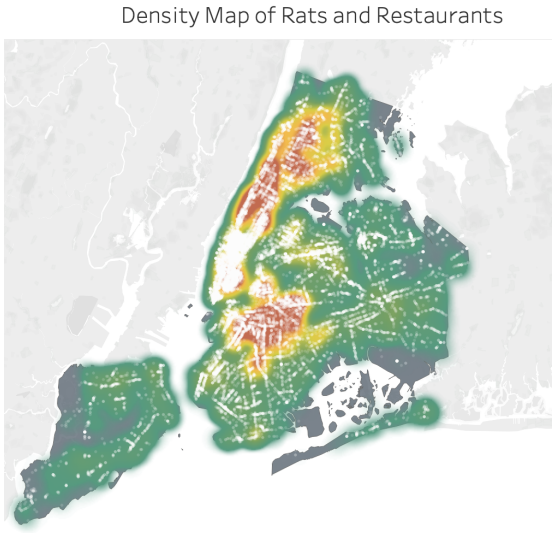
Figure 5: Density map - rat density and subway infrastructure



Next, we will look at the spatial analysis of rats and restaurants in New York City. A choropleth map was also created to visualise the density of restaurants by zip code in a choropleth map. Although this choropleth map is not shown in this report, the visualisation shows that lower Manhattan has the highest number of restaurants, followed by a few areas in Brooklyn and Queens.

Comparing the choropleth maps, it is difficult to draw a correlation between restaurants and rat sightings. To visualise these data points better, a density map of the rat countings with an overlay of the restaurants data was created and shown in Figure 6. In the density map of rat countings and restaurants in New York City, it is still difficult to draw a correlation between the two. Figure 6 shows a high density of restaurants in lower and middle Manhattan, but the choropleth map in Figure 4 show that the highest rat countings were in zip codes 11221 and 10025, which are in Brooklyn and upper Manhattan respectively.

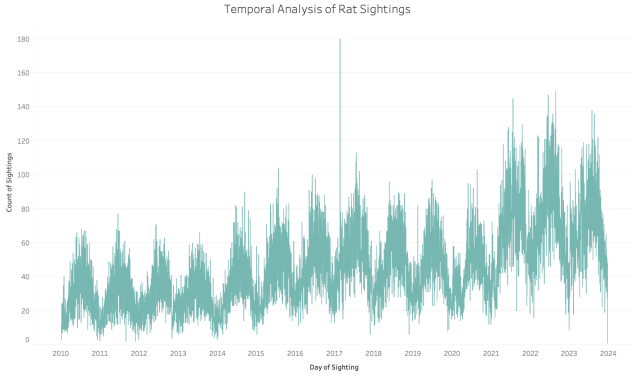
Figure 6: Density map - rat density and restaurants



Temporal Analysis

In the temporal analysis, we will explore the time patterns of the rat sightings. In the initial exploration of the temporal data, a time series graph was implemented to show the general time pattern from 2010 to 2023, as shown in Figure 7.

Figure 7: Time series line graph



The time series graph shows a general increase in rat sightings over the years and a clear batten in each year. From the beginning of each year, the rat sightings go up and look like they reach a peak in the middle of the year before going back down towards the end. To explore this more in depth, Figure 8 shows a heatmap of the temporal data monthly by each year. As suspected, the more dense areas are mostly concentrated in months that are in the middle of the year. We can observe that the rat sightings go up as the year goes on, and goes back down after the summer months.

Figure 8: Temporal heat map (monthly by each year)

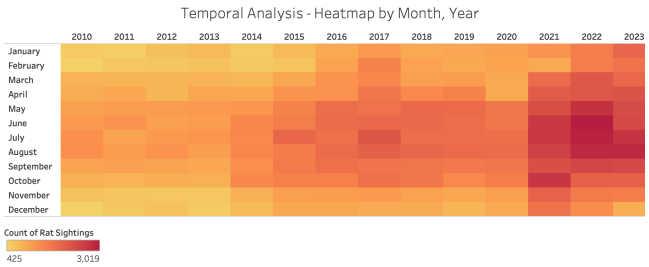
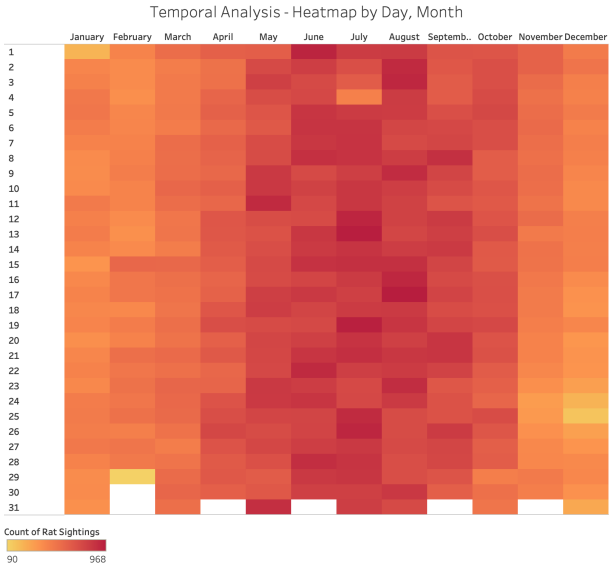


Figure 9 shows another heatmap of the temporal data daily by each month. This visualisation uses the aggregate data points of rat sightings each month across the years of 2010 to 2023. We can again observe that the months with the most rat sightings are in the middle of the year, with peaks in June, July and August. Based on these visualisations and analysis, we can draw a correlation between rat sightings and warmer months of the year. This makes sense as rats are most likely more active during warmer weather to scavenge for food and mate with each other, while in the winter months they are conserving energy.

Figure 9: Temporal heat map (daily by each month)



In another heatmap that was created to explore the time patterns of rat sightings, it showed that the hour with the most number of rat sightings was at midnight. At the aggregate level, 3000 - 4000 rat sightings were recorded at midnight of each day, whereas the number of rat sightings recorded at other hours ranged from around 40 to 300. This time pattern also makes sense as rats are nocturnal animals.

Modelling

Plotting the occurrence of rat sightings over time gives us a stationary time series which we can then model using forecasting methods. The number of rats seen on a given day is largely influenced by temporal factors such as the season and the sociological climate (e.g. the pandemic lockdown or increased resourcing to rat extermination). Given this, the forecasting method used should take into account previous data in determining future predicted values. One such method is the autoregressive moving average (ARMA) model.

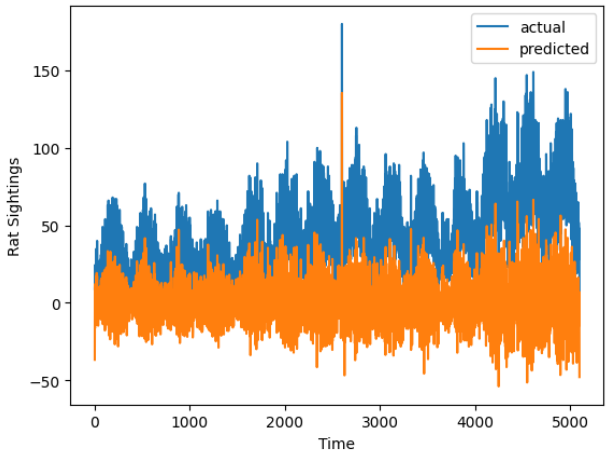
The ARMA model is a linear combination of an autoregressive and a moving average model. The autoregressive analysis calculates the regression

of past time series, generating a linear relationship between past and present values [9]. A moving average model uses the residuals of the past forecast errors in a weighted moving average that captures the impact of random dips or jumps in rat sightings at various time points. Taken together, the combined model predicts the future values based on both the previous data values and past forecasting errors.

The time series data was confirmed to be stationary using an Augmented Dicky-Fuller test, and the resulting parameters for the ARMA model were optimised by analysing the autocorrelation and partial autocorrelation plots, in accordance with the Box-Jenkins method.

The performance of the model was evaluated by finding the difference between the actual rat sighting values and the predicted rat sighting values. The errors were plotted alongside the model prediction in Figure 10.

Figure 10: ARMA model visualisation



Despite the robustness of the ARMA model, it largely fails to accurately predict the rat sighting data. While it captures the general shape and trends of the time series, the error between the prediction and the actual values remains high. There are various potential reasons for this - for example, the time series data likely exhibits underlying complex patterns that cannot be sufficiently modelled by autoregressive and moving average terms. If nonlinear relationships between underlying variables exist, these would also not be captured by an ARMA model, which assumes linear relationships between past and current data. This would make sense given that rat sightings are influenced by multiple factors and therefore this data likely requires a more sophisticated model such as an LSTM (Long Short-Term Memory) model. Such deep learning methods, however, would require a much larger dataset than the one provided. Additionally, because ARMA models smooth data in relation to previous errors, they can sometimes fail at predicting turning points. However, given the limited data available, the model still adequately forecasted the general trends of the time series.

C. Analysis Results

This analysis answers our initial research questions. The zip codes with the most rat sightings are 11221 (Bushwick, Brooklyn) and 10025 (Upper West Side, Manhattan). The geospatial visualisations of rat sightings and subway infrastructure show that there is a correlation. The areas with the most subway stations and where the subway lines go through have the highest density of rats. However, when we look at the geospatial map for rat density and restaurant data, there is no visualisation that shows that restaurants have an influence on the rat population.

In the temporal analysis, there is a time pattern of rat sightings. The time series graph shows a repeated pattern in each year. This is confirmed by the temporal heat maps. From January, rat sightings are at a low, and steadily increases and reaches its peak in the summer months. After August, the rat sightings steadily decrease and reach a low again in December.

In the ARMA model, the errors were high but the model was able to capture the general trend of the

time patterns of rat sightings. This could be because of a more complex pattern of the data that could not be captured in the model.

V. Critical Reflection

The spatial analysis using density maps and choropleth maps successfully conveys the areas and zip codes of New York City that are most densely populated with rats according to the supplementary dataset. The choropleth map provides a geographic representation of New York City and its individual zip codes based on the rat data variables. It is easy to see which zip codes were most darkly coloured, representing the most densely rat populated zip codes. While the choropleth map represents the geographic areas of New York City, the density maps are used to compare two data variables. In this case, the density maps examine rat population and subway infrastructure correlation and rat population and restaurant density correlation. Although the density map is not as granular as the choropleth map as it does now clearly show which areas have the highest rat population, it shows the continuous spatial distribution of rat density while using an overlay technique in Tableau to visualise two variables at once. The subway stations and lines layer over the density map successfully shows the influence of the subway infrastructure of the rat population in different areas. However, the restaurant data layer over the density map is slightly harder to interpret. Because the number of restaurants is dense in certain areas such as lower Manhattan it is difficult to see the density map of the rat data beneath the restaurant data layer.

Using a time series line graph and temporal heat maps, the temporal analysis shows the time patterns of rat sightings in New York City. The time series line graph provides a clear look into the overall trend of rat sightings from 2010 to 2023. Then, the temporal heat maps explore the time patterns more deeply by visualising the data on a more granular level, showing the time patterns on a monthly and daily basis. With the temporal heat maps, the time pattern initially observed in the time series line graph can be confirmed. However, it should be taken into consideration the time pattern of behaviour of humans in New York City. Although the temporal analysis shows a pattern in rat sightings, rat sightings are also dependent on when humans are outside to witness and report the rats. For example, the graphs show that the most rat sightings occur in the summer months, but this could be because humans are also more likely to be outside during warmer weather.

Lastly, an autoregressive moving average (ARMA) model is used to predict future rat sightings. An ARMA model was implemented because it is built to capture time patterns and dependencies [9]. However, the results yielded a high error rate but managed to capture the general pattern of the time series data. As discussed in the analysis, this could be due to underlying complex nonlinear patterns in the data. This would make sense as the rat population is influenced by many factors outside of time such as spatial reasons, food resources, human movement, and socioeconomic factors.

References

[1] J. Auerbach, “Does New York City Really Have as Many Rats as people?,” *Significance*, vol. 11, no. 4, pp. 22–27, Oct. 2014, doi: https://doi.org/10.1111/j.1740-9713.2014.00764.x.
[2] M. G. Walsh, “Rat Sightings in New York City Are Associated with Neighborhood sociodemographics, Housing characteristics, and Proximity to Open Public Space,” *PeerJ*, vol. 2, p. e533, Aug. 2014, doi: https://doi.org/10.7717/peerj.533.
[3] H. Bailey, R. Corkrey, B. Cheney, and P. M. Thompson, “Analyzing Temporally Correlated Dolphin Sightings Data Using Generalized Estimating Equations,” *Marine Mammal Science*, vol. 29, no. 1, pp. 123–141, Mar. 2012, doi: https://doi.org/10.1111/j.1748-7692.2011.00552.x.

[4] “Rat Sightings | NYC Open Data,” *data.cityofnewyork.us*. https://data.cityofnewyork.us/Social-Services/Rat-Sightings/3q43-55fe (accessed Jan. 04, 2024).
[5] “DOHMH New York City Restaurant Inspection Results | NYC Open Data,” *data.cityofnewyork.us*. https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j
[6] “MTA NYCT Subway Entrances and Exits: 2015 | State of New York,” *data.ny.gov*. https://data.ny.gov/Transportation/MTA-NYCT-Subway-Entrances-and-Exits-2015/i9wp-a4ja/about_data (accessed Jan. 04, 2024).
[7] “NYC Neighborhoods - nyc_zip_borough_neighborhoods_pop.csv - BetaNYC’s Community Data Portal,” *data.beta.nyc*. https://data.beta.nyc/en/dataset/pediacities-nyc-neighborhoods/resource/7caac650-d082-4aea-9f9b-3681d568e8a5
[8] Chate, Parinita & Patkar, Uday. (2015). Big Data Visualization: Challenges and SAS Visual Analytics.
[9] R. Li, “Research on Applications of ARMA in Forecasting of Time Series,” *Atlantis Press*, 2015. https://www.atlantis-press.com/article/18690.pdf
[10] A. Cheng, “Manrattan - A Look into NYC’s Rats,” *Cmu.edu*, Dec. 12, 2022. https://www.stat.cmu.edu/capstoneresearch/fall2022/315files_f22/team6.html

Table of Word Counts

Problem Statement	**188/250**
State of the Art	**441/500**
Properties of the Data	**269/500**
Analysis Approach	**257/500**
Analysis Process	**1289/1500**
Analysis Results	**198/200**
Critical Reflection	**495/500**