Relationship between Temperature, Homeless Encampments, and Criminal Summons in New York City from 2018 to 2019

Abstract

In recent years in New York City, there has been an increased sense of insecurity among citizens as well as an increase in homelessness. Feeling an urge to investigate the underlying causes and correlations related to crimes and homelessness, the authors of this paper combined the temperature, the number of criminal summons, and the number of homeless encampment records datasets from NYPD and sought to find the true relationship between them. To eliminate the influence of the pandemic, the study selected the data from the years 2018 and 2019. Results of the analysis showed low correlations both between summons and temperature (0.25) and between the summons and homeless encampments (no greater than 0.11). The correlation between homelessness and temperature was relatively high (0.618). In terms of different geographic regions, no prominent difference in correlation results between different boroughs was observed.

Introduction

Public safety surrounds the different needs of different community groups. Since the entering of the pandemic period, however, there had been a steady increase in crime in New York City. In 2021, the situation “deteriorated to the point where the total for all such offenses”, according to Newsday,“ passed 100,000 for the first time since 2016”. More recently, for the month of February 2022, New York City saw a 58.7% increase in overall index crime compared to February 2021 (9,138 vs. 5,759), according to NYPD. With the recent increased amount of criminal activities in New York City, while concerns were arousing, the need to understand what types of factors are potentially related to the criminal trend became more urgent than ever.

Many factors could affect the number of criminal incidents. Intuitively, the pandemic-related factors may have played a major role in the increase in criminal activities, but other factors cannot be ignored as well. In our study, we intended to investigate how an environmental factor (temperature) and a social factor (homeless encampments) correlate with criminal activities in New York City. Through observation and analysis of the correlation between the number of criminal activities, temperature, and the number of homeless encampments, we were able to examine how these factors may or may not correlate with each other.

For the big data tools used in this project, we mainly utilized MapReduce for initial profiling to clarify the structure and content, verify that the information in these datasets matches the descriptions, and get a general sense of what the datasets are like. We then used MapReduce again for cleaning, dropping the information that was not useful for our analysis and filling

missing values. Then, we used Spark for the analysis of each individual dataset, examining the potential patterns that existed. After merging the three datasets using Hive, we used Spark for the merged analysis, calculating the correlation index between different factors, and, lastly, we used Matplotlib for visualization.

Motivation

The size of the homeless population and the count of criminal records (arrest and summons) in New York City have always been huge. Recent news reported that “systematic killings of the unhoused” happened more often in the past few months, raising safety concerns for the homeless people (Moses, D. 2022). One purpose of our project was to explore the relationship between the homeless encampment and the amount of criminal summons in the city, with an additional climatological variable, temperature, to remove potential confounds. To eliminate the special condition under COVID, we only selected the data from 2018 to 2019. With the two historic datasets from NYPC, we wanted to synthesize the available pieces of evidence on how the homeless encampment rate relates to the rates of criminal behaviors and contacts with the justice system. Joining them with the temperature dataset, we would like to investigate and, potentially, conclude a relationship between homelessness, criminal summons, and temperature in NYC.

Related Work

Burton, B., Pollio, D. E., & North, C. S. (2018). A longitudinal study of housing status and crime in a homeless population. Annals of Clinical Psychiatry.

Potential relations between homelessness and crime have been widely measured in previous studies. A longitudinal study (Burton, B. et.al, 2018) followed 225 homeless (at start) individuals for 3 years and evaluated the pattern of their criminal records. Their results showed that the individuals who were unhoused had more criminal records compared to those that were housed. Further, homelessness was shown to be a strong predictor of homeless status offenses, such as vagrancy and trespassing.

Berk, R., & MacDonald, J. (2010). Policing the homeless: An evaluation of efforts to reduce homeless-related crime. Criminology & Public Policy, 9(4), 813-840.

This study examined the implementation of the place-based policing intervention in Los Angeles, which focused on crime and disorder associated with homeless encampments, to

achieve the goal of reducing crimes. The research provided evidence that geographically targeted police interventions could meaningfully reduce crimes associated with homeless encampments, but there was no evidence that crimes are simply displaced to other areas.

DeFronzo, James. “Climate and Crime: Tests of an FBI Assumption.” Environment and Behavior, vol. 16, no. 2, Mar. 1984, pp. 185–210, doi:10.1177/0013916584162003.

The study aimed to test an assumption by the FBI that climatic factors had an independent impact on crime rates. To evaluate this hypothesis, the researchers conducted multivariate analyses of the possible impacts of three climatic and twelve non-climatic factors on variation in seven serious forms of violent crime and property crime among the 142 largest American SMSAs in order. The findings revealed that climatic conditions had uniformly weak relationships with different types of crimes when compared to the impacts of other variables. In the study, the temperature was evaluated as a major climatic factor, and the results showed that the number of days with high temperatures (above 32.20°C) and the number of days with low temperatures (below 0°C) had positive associations with the variation in certain crime rates. The results also indicated that freezing temperatures and significant precipitation had weak indirect negative associations with crime rates.

Datasets

Data Source 1: NYPD Criminal Court Summons (Historic)

The Criminal Court Summons dataset contains all NYC historical criminal summons data from 2006 to May 3, 2021. Our project, however, only used the crime data from 2018 to 2019, which removes the potential influence of COVID. Each row recorded one incidence of a summon and included the description of the person, location, and summon type data. Examples of criminal summons included alcohol, marijuana, disorderly conduct, knife, trespassing, and vending. The NYC borough that which the location belonged was also recorded.

Column Name	Type	Description
SUMMONS_KEY	Plain Text	Randomly generated persistent ID for each violation
SUMMONS_DATE	Date & Time	Exact date of violation for the reported event
OFFENSE_DESCRIPTION	Plain Text	Description of the violation committed
LAW_SECTION_NUMBER	Plain Text	NYS Penal Law and local law section number

LAW_DESCRIPTION	Plain Text	Description of the law dictionary where the cited violation came from. Blank fields represent errors in the database
SUMMONS_CATEGORY_TYPE	Plain Text	General description of the violation category
AGE_GROUP	Plain Text	Perpetrator's age within a category
SEX	Plain Text	Perpetrator's sex code. M(Male), F(Female), U (Unknown or violation issued to a business)
RACE	Plain Text	Perpetrator's race description. U(Unknown or violation issued to a business)
JURISDICTION_CODE	Number	Jurisdiction responsible for the issued violation. 0(Patrol), 1(Transit) and 2(Housing) represent NYPD whilst codes 3 or more represent non-NYPD jurisdictions
BORO	Plain Text	New York City Boroughs
PRECINCT_OF_OCCUR	Number	Precinct where the violation was issued
Latitude	Number	Latitude coordinates for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude	Number	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Lon_Lat	Point	Longitude and Latitude coordinates for mapping

Data Source 2: 311 Service Requests from 2010 to Present: Homeless Encampments

NYC 311 serves the public and handles all requests for government and non-emergency services, connecting residents, business owners, and visitors with the information and people who can help them best. The Homeless Encampments dataset recorded 311 service requests that related to homeless incidents from 2010 to the present. Our project, again, only used the 311 data from 2018 to 2019 to remove the potential influence of COVID.

Column Name	**Type**	**Description**
Unique Key	Plain Text	Unique identifier of a Service Request (SR) in the open data set

Created Date	Date & Time	Date SR was created
Closed Date	Date & Time	Date SR was closed by responding agency
Agency	Plain Text	Acronym of responding City Government Agency
Descriptor	Plain Text	This is associated to the Complaint Type, and provides further detail on the incident or condition. Descriptor values are dependent on the Complaint Type, and are not always required in SR.
Location Type	Plain Text	Describes the type of location used in the address information
Incident Zip	Plain Text	Incident location zip code, provided by geo validation.
Incident Address	Plain Text	House number of incident address provided by submitter.
Street Name	Plain Text	Street name of incident address provided by the submitter
Address Type	Plain Text	Type of incident location information available.
City	Plain Text	City of the incident location provided by geovalidation.
Borough	Plain Text	Provided by the submitter and confirmed by geovalidation.

Data Source 3: U.S. Local Climatological Data: NY CITY CENTRAL PARK

Local Climatological Data (LCD) are summaries of climatological conditions from airports and other prominent weather stations managed by NWS, FAA, and DOD. The datasets include hourly observations along with daily and monthly summaries of maximum, minimum, and average temperature, temperature departure from normal, dew point temperature, average station pressure, ceiling, visibility, weather type, wet bulb temperature, relative humidity, degree days (heating and cooling), daily precipitation, average wind speed, fastest wind speed/direction, sky cover, and occurrences of sunshine, snowfall and snow depth. The datasets we used were recorded and summarized by the weather station in Central Park, New York City from 2018 to 2019. As we only laid our focus on daily temperature-related data types, the table below documented parts of the schema of the dataset that is related to temperature indication.

Column Name	Type	Description
DATE	Date & Time	Exact date and time of the record
DailyAverageDewPointTemperature	Number	The daily average dew point temperature
DailyAverageDryBulbTemperature	Number	The daily average dry bulb temperature
DailyAverageWetBulbTemperature	Number	The daily average wet bulb temperature
DailyCoolingDegreeDays	Number	The daily cooling degree days
DailyDepartureFromNormalAverageTemperature	Number	The daily departure from normal average temperature
DailyHeatingDegreeDays	Number	The daily heating degrees
DailyMaximumDryBulbTemperature	Number	The daily maximum dry bulb temperature
DailyMinimumDryBulbTemperature	Number	The daily minimum dry bulb temperature

Analytic process

As multiple tools were used in the profiling, cleaning, and analysis process, the data was transported from place to place at different stages. As mentioned in the introduction, we used MapReduce for initial profiling for purposes of clarifying the structure and content, verifying that the information in these datasets matches the descriptions. Doing so allowed us to interpret the general structure of our datasets. We then used MapReduce again for cleaning, dropping the information that was not useful for later analysis, and filling missing values. After cleaning, we used Spark for the analysis of each individual dataset, examining the potential patterns that existed. After merging the three datasets using Hive, we used Spark for the merged analysis, calculating the correlation index between different factors, and, lastly, we used Matplotlib for visualization.

Individual dataset analysis

Temperature

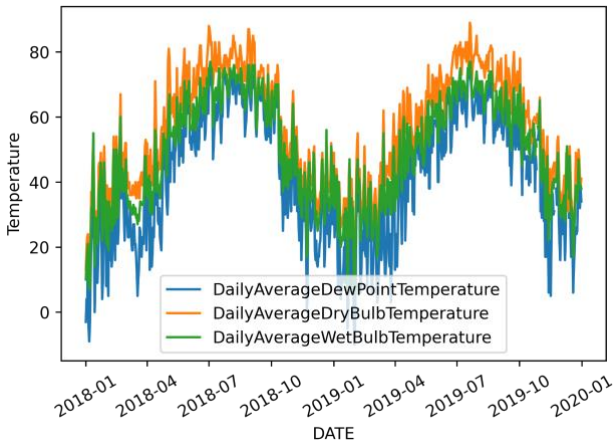

Figure 1: visualization of three indications of temperature in 2018 and 2019

For the Local Climatological Dataset, we used MapReduce for the initial profiling to get a general sense of how large the dataset was, counting how many records there were. We then again used MapReduce to drop the unwanted columns and rows, only keeping the daily temperature-related information. There were three types of indications of temperature in the LCD dataset: the daily average dew point temperature, the daily average dry-bulb temperature, and the daily average wet-bulb temperature, which used different methodologies to record the temperature. However, the correlation between these three columns was really high, with a pair-wise correlation of around 0.96. Therefore, we decided to use only one of the three types for our analysis. We chose the daily average dry-bulb temperature because the other two columns both had missing values.

In Figure 1, the graph reaches its peak around July and August, and it reaches its lowest around December and January. To retrieve the average of the daily average temperature in 2018 and 2019, we used the aggregate function and used the avg function imported. The average temperature in 2018 is 56.205, and in 2019 it is 55.962. We then further calculated the average temperature in each month by filtering by both year and month, and we found that, both in 2018 and 2019, the average monthly temperature reached its highest in July and August, and it reached its lowest in January. To retrieve the dates with the lowest and the highest daily average temperature in 2018 and 2019, respectively, we sorted the dates by the temperature by filtering by years. We found that, in 2018, the dates with the lowest temperature were mostly in January, while in 2019, the dates with the lowest temperature were distributed in January, February, and March more evenly. For the dates with the highest temperature in 2018, they were distributed mostly in June, July, August, and September, while the dates with the highest temperature in 2019 mostly came from July.

Homeless Encampments

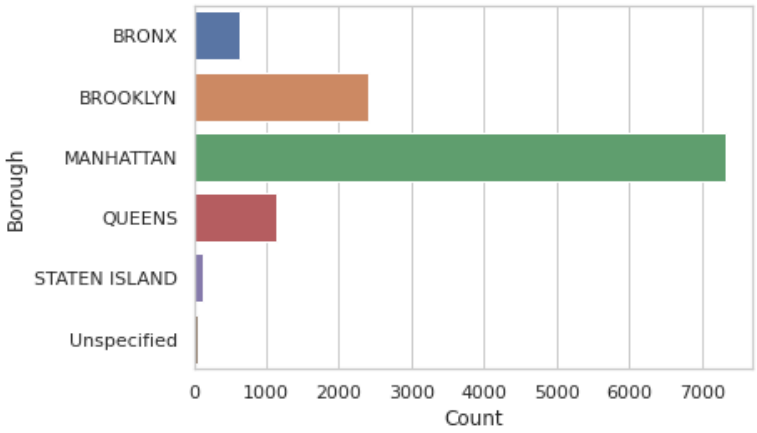

Figure 2: Number of encampments in different boroughs in 2018 and 2019

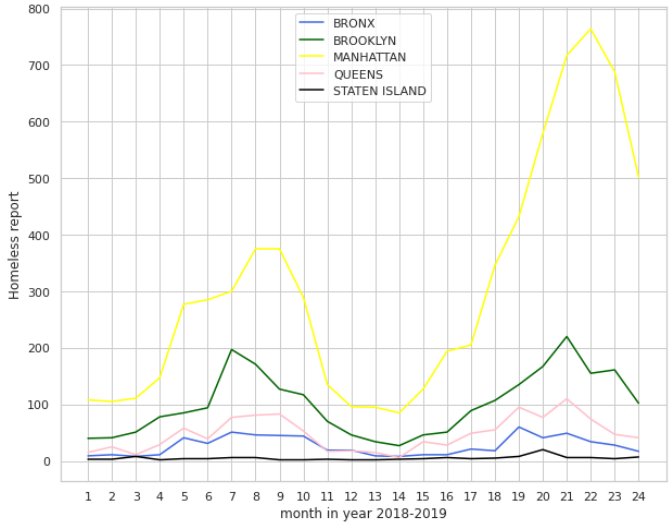

Figure 3: Number of encampments in different months of 2018 and 2019

The original Homeless Encampments dataset recorded 311 service requests from 2010 to the present. We first applied MapReduce to profile the original dataset to get the general information. Then we did data cleaning to drop the unwanted columns while keeping the non-null columns that cover the basic information about homeless encampments. Specifically, the preprocessed dataset retained the information of the location and date of service request(SR) being reported. We further preprocessed the data by utilizing PySpark. We extracted the year, month, and date from the date column in the cleaned dataset so that we could constrain the dataset in the years 2018-2019. Moreover, we grouped the dataset by day to generate a dataset that was suitable for merging, including the information of date, the daily count of homeless encampments reported, and the borough.

We then analyzed the further preprocessed dataset. Firstly, we grouped the dataset by boroughs and applied the aggregation function of sum to see the distribution of homeless encampments among boroughs. We could see from Figure 2 that Manhattan had the highest number of homeless encampments and followed by Brooklyn and Bronx. We also visualized the trend of

changes in the number of homeless encampments through months from 2018 to 2019. From Figure 3, we could observe that there are two peaks of encampments count from 2018 to 2019. And encampments count in Manhattan was way higher than in the other boroughs. We could also get a general sense that the peak in 2019 was much higher than in 2018, suggesting that the homeless phenomenon was exacerbated in 2019. Moreover, if we specified the month where the peak appeared, we could then conclude that the boost of encampments occurred in the hot season (from July to September). In contrast, the local minima of encampments count occurred in the cold season (from December to January). In addition, we could see that the gap between encampments counts in Manhattan and that in other boroughs started boosting in month 17 (May in 2019) and reached the highest in month 22 (October in 2019), indicating that there were some factors that drove the homelessness to be further researched on.

Criminal Summons

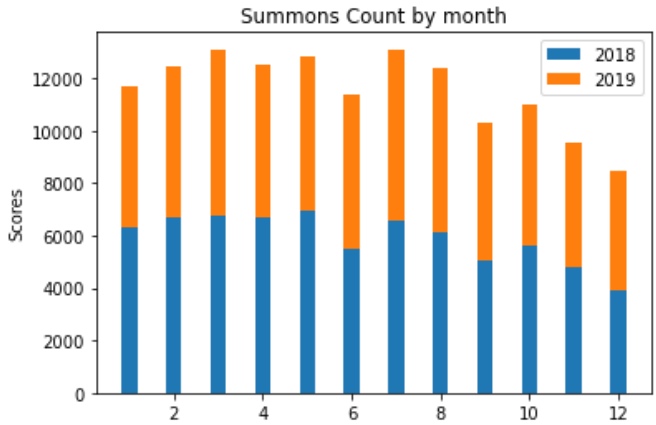

Figure 4: Summon counts by months in 2018 and 2019

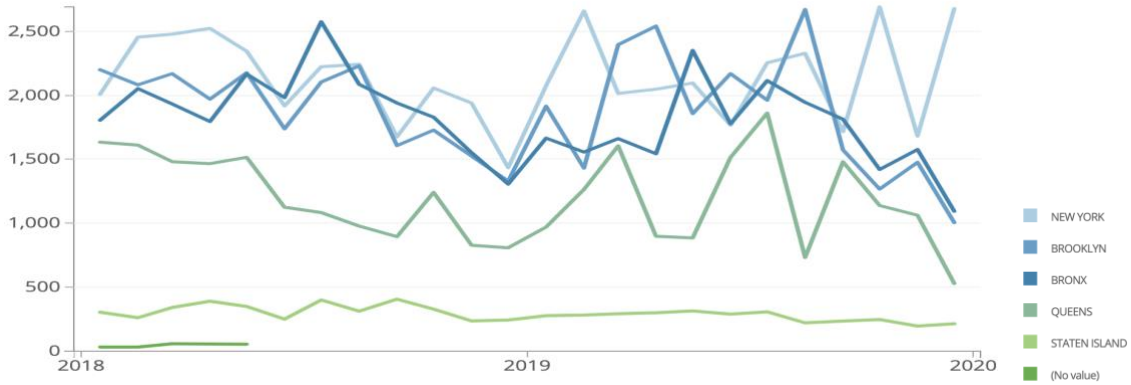

Figure 5: Summon counts in different boroughs in 2018 and 2019

Originally, the criminal summons dataset had 17 columns, including some record numbers as primary keys for identification. During the data profiling stage, MapReduce removed a small number of records with incorrect numbers of elements in a row, which left us with 5116597 records. Further, the data was imported into Spark, cleaned, and aggregated via PySpark.

Eliminating the rows with no borough value and selecting only the records in 2018 and 2019, we were left with 138993 rows of records. Though there were only five boroughs in New York City, we observed that there was an extra type named “New York.” After plotting all the dots labeled “New York” on the map using the longitude-latitude data, we found that most of the records were in Manhattan, with only a small amount of exceptions. We thus decided to include “New York” in the “Manhattan” category.

Among all boroughs, Manhattan had the highest number of summoning cases from 2018 to 2019. The second and third highest were Brooklyn and Bronx. The most frequent categories of summoning included alcohol consumption, marijuana, trespassing, MTA-related summons, knife, and misbehaviors like fighting. Age-group-wise, the group that committed the most cases is 25-44 (noted that there were 33680 cases with unknown age, which may make the data inaccurate). The racial groups that committed the most cases of summons were African American, then White Hispanic, White, and Black Hispanic. Regarding gender, the male cases were significantly higher than females (male=43056, female=6080).

Only looking at the years 2018 and 2019, the distribution of summons count by date seemed to correlate with the time of the year, especially in December. One natural generalization was that the end-of-the-year decrease in summons count was due to the drop in temperature. However, there were still a lot of confounding variables. For instance, the decrease in summons records may result from the decrease in NYPD officer activity instead of the *actual* reduction of summons cases. If we further group the summons count by borough, there are still no prominent trends that accounted for the variation of summons counts throughout the year. The total number of summons was relatively consistent. Had the summons dataset included the level of severity of each summon case and more detailed categorizations, we could further analyze the data distribution by summons categories and separating the influence of various unaccounted variables.

Analysis of the merged datasets

After exporting the cleaned, join-ready datasets as CSV files and uploading them to HDFS, we created external tables in Hive out of the three CSV files. Then, the tables were joined within Hive using SQL commands, creating a new table, which was then exported as a CSV file. Further analysis of the merged dataset was based on the file. The datasets of criminal summon count and encampment count were joined on the date and borough columns, and the temperature dataset was joined on the date only with the previous two.

Criminal summons vs. homeless encampments

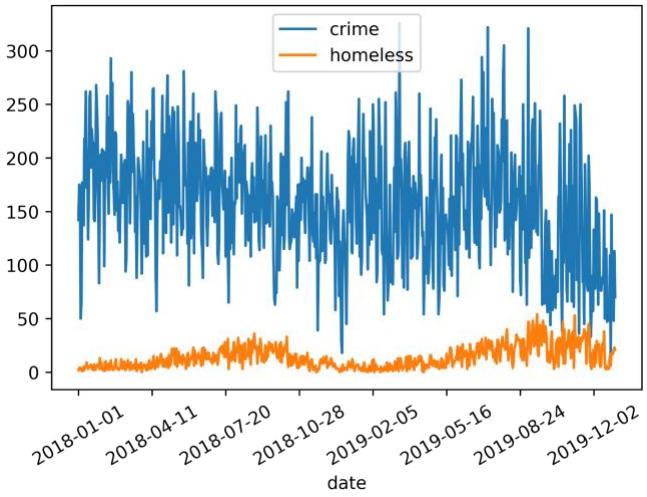

Figure 6: Summon and encampment counts in 2018 and 2019

The correlation between the summon count and the encampment count was overall weak. The overall correlation coefficient between daily summon count and daily encampment count in 2018 was 0.1136, and the correlation coefficient of that in 2019 is 0.0391. The correlation was also analyzed with respect to each borough. The regional correlation analysis also indicates really low correlation scores. The correlation coefficients between daily summon count and daily encampment count in Bronx, Brooklyn, Manhattan, Queens, and Staten Island were 0.1172, -0.0286, 0.0866, 0.0715, -0.0098 in two years. Then, the correlation coefficients between monthly summon count and monthly encampment count in 2 years were calculated, with the highest correlation coefficient being 0.3764 in July, and the lowest being 0.1471 in March.

Criminal summons vs. temperature

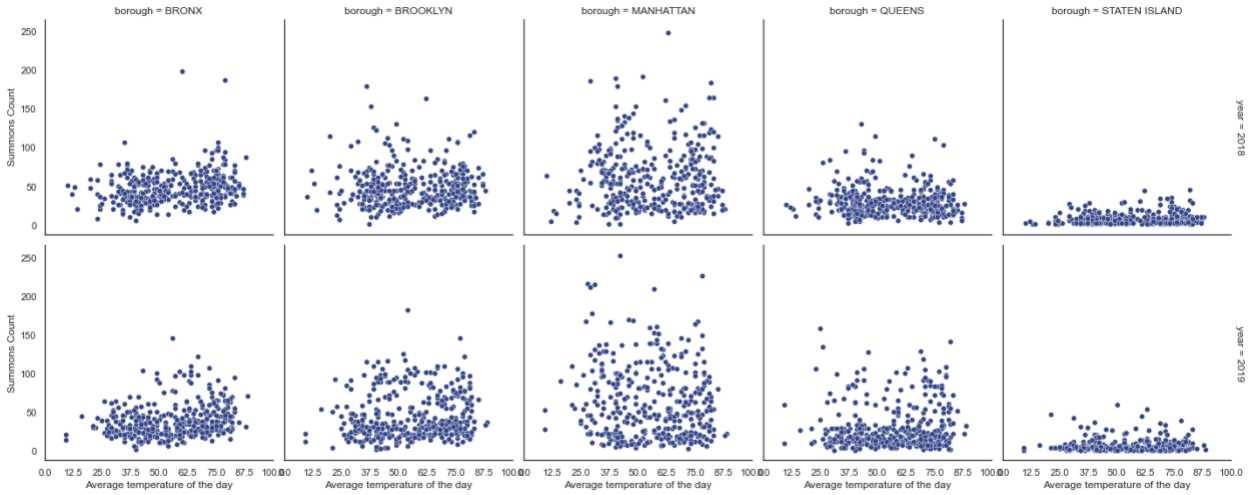

Figure 7: Temperature and summon counts in 2018 and 2019

Similar to that between the summons and homeless encampment, the correlation coefficient between these two datasets was relatively low. The correlation coefficient between the monthly summons count and the monthly average temperature was 0.25. The correlation between daily temperature and the total number of summons was only 0.033. From the grouped scatter plot, we could also find that the shape of the scatters depended more on the total number of summons and the population within a borough, not on the variation of temperature throughout the year.

This low level of correlation was partly expected before processing the data due to the fact that summons fell into many different categories, and not all of them happened outdoor. In addition, the police activity patterns, which were not recorded in the dataset, could also be a great determiner of the number of summons on a given day.

Temperature vs. Homeless Encampments

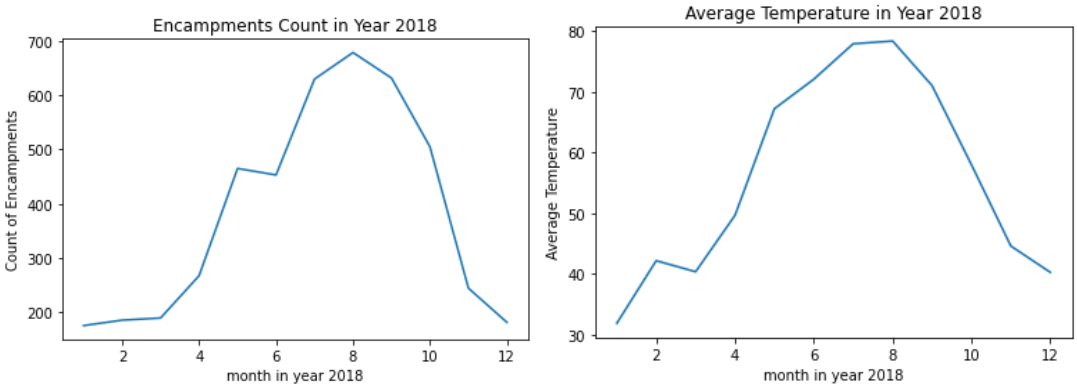

Figure 8: Temperature and encampment count in 2018

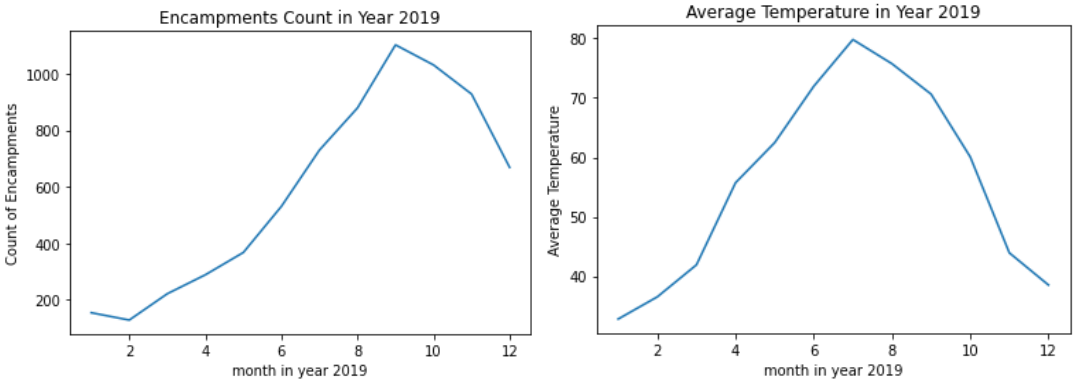

Figure 9: Temperature and encampment count in 2019

The correlation between temperature and homeless encampments was relatively strong compared to the two analyses above. In 2018, the correlation between monthly average temperature and homeless encampments count was up to 0.951, while in 2019, the correlation between them was 0.540. In general, the correlation coefficient between average temperature and homeless encampments count was 0.618.

By visualizing the trend of changes in average temperature and homeless encampments count, we could easily form a general sense of the relationship between temperature and homeless encampments. From the Figure 8, we could see that the trend silhouettes were similar: they were both a bell curve centered around month equals eight, indicating that the number of encampments increased as the temperature rises. From the Figure 9, we could see that although the trend in plots are similar in general, the peak of the curve resided in different months. It could be depicted that the peak of homeless encampments came after the peak of average temperature. That was the major reason why the correlation dropped in 2019 compared to that in 2018.

Conclusion

Results of the analysis showed weak correlations both between summons and temperature and between the summons and homeless encampments. The correlation between homelessness and temperature was relatively high, agreeing with common sense. In terms of different geographic regions, no prominent difference in correlation results between different boroughs was observed.

While the study provided an initial insight into how these factors are correlated, the low correlation between summons, homelessness, and temperature may be due to the variety and different levels of severity of summons cases. Many factors residing in the original dataset were discarded in the analysis process due to limitations of time, such as different summon categories and police activity patterns. Therefore, we wish to further research how other potential factors could influence the number of criminal activities in New York City. We hope this study, in combination with future research, will further elucidate what factors play a determining role in criminal activities and help to protect our community.

References

Burton, B., Pollio, D. E., & North, C. S. (2018). A longitudinal study of housing status and crime in a homeless population. Annals of Clinical Psychiatry.

Berk, R., & MacDonald, J. (2010). Policing the homeless: An evaluation of efforts to reduce homeless-related crime. Criminology & Public Policy, 9(4), 813-840.

Calsyn, R. J., Yonker, R. D., Lemming, M. R., Morse, G. A., & Klinkenberg, W. D. (2005). Impact of assertive community treatment and client characteristics on criminal justice outcomes in dual

disorder homeless individuals. Criminal behaviour and mental health : CBMH, 15(4), 236–248. https://doi.org/10.1002/cbm.24

DeFronzo, Jamesw. “Climate and Crime: Tests of an FBI Assumption.” Environment and Behavior, vol. 16, no. 2, Mar. 1984, pp. 185–210, doi:10.1177/0013916584162003.

Roy, L., Crocker, A. G., Nicholls, T. L., Latimer, E. A., & Ayllon, A. R. (2014). Criminal behavior and victimization among homeless individuals with severe mental illness: a systematic review. Psychiatric services (Washington, D.C.), 65(6), 739–750. https://doi.org/10.1176/appi.ps.201200515

Moses, D. (2022, March 23). Not safe anywhere: NYC's homeless residents, advocates decry dangerous shelter and street conditions. amNewYork. Retrieved May 1, 2022, from https://www.amny.com/news/not-safe-anywhere-homeless-advocates-decry-dangerous-shelter-and-street-conditions/

Yoo, Y., & Wheeler, A. P. (2019). Using risk terrain modeling to predict homeless related crime in Los Angeles, California. Applied geography, 109, 102039.