W200 Summer 2021 Project 2 Final Report - DSNY Graffiti Tracking

Alejandro Pelcastre, Henry Wang, Shanie Hsieh

Introduction

For our W200 project 2, we decided to research data that resonated to each of us personally. We pulled data from the City of New York Department of Sanitation, also referred to as DSNY. The data we decided to focus on was the DSNY graffiti tracking in the City of New York. The NYC Department of Sanitation is the world's largest sanitation department. DSNY collects more than 10,500 tons of residential and institutional garbage and 1,760 tons of the recyclables – each day. While efficiently managing solid waste and clearing litter or snow from 6,300 miles of streets, the Department is also a leader in environmentalism — committing to sending zero waste to landfills.

The Graffiti-Free NYC Program removes graffiti and other blight across the five boroughs. Graffiti-Free NYC is a cooperative effort among the NYC Economic Development Corporation, the NYC Department of Sanitation, and the Office of the Mayor.

Questions

The motivation behind our analysis is centered around our overall question, " How is the graffiti in New York City impacted by the city and its people?" After looking through the data we began narrowing down our questions and decided to try and answer the following questions: "What areas of New York City have the most graffiti?", "What unique geographical qualities contribute to the difference in graffiti counts and time it takes to remove graffiti in different boroughs?", and "Why is there a surge in graffiti activity at the end of the year?". These are the questions our analysis tries to explore and answer.

Steps to analyze questions / Data Cleaning

The DSNY data contains key variables which we used in order to form our analysis and make such predictions from the data. Some of these variables include the BOROUGH (where in the five boroughs graffiti is located), CREATED_DATE (date graffiti was reported), CLOSED_DATE (date issue was closed, whether it was cleaned up or covered), X_COORDINATE AND Y_COORDINATE (latitude and longitude of the complaint address), and ZIP_CODE (zip code address where graffiti is located). The first step we took before beginning any analysis was to determine the significance of NaN values or unspecified values contained in the data so it could be removed from our analysis if insignificant. We found negligible amounts of NaN and unspecified values and therefore concluded we could eliminate it from our analysis without much influence.

Next, we want to take a look at specific graffiti locations and use the x- and y-coordinates to map out a graph. Before beginning, we found that there are 1345 missing x-coordinates and 1345 missing y-coordinates where removing them still left us with 99.68% of the original data size data. With the remaining data, we created a scatter plot graph with the x- and y-coordinates of the graffiti incidents to reveal a density map of where the graffiti was occuring. We can use this graph to compare to a population density heat map from 2015 under the assumption that the 2015 population did not change much compared to 2020. As for the zip code, removing those NaN values left us with 93.78% of the original data. With this, we identified the zip codes with the highest number of graffiti incidences and were also able to create a distribution of graffiti incidences.

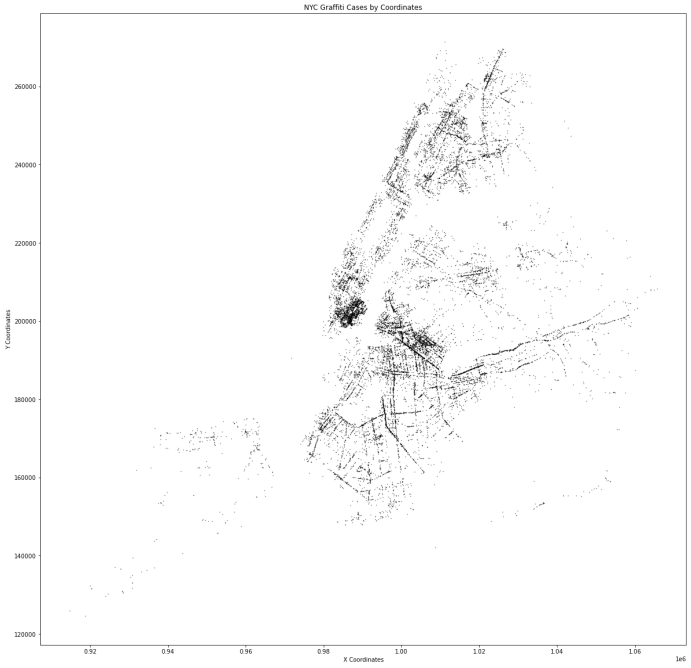
We also decided to conduct a couple different analyses with the borough locations. For some data cleaning, we found only 12 unspecified borough locations and thus, the 12 compared to the 20,000+ data points we have seemed conclusive to remove. The first step we took was

checking for any trends over time and creating a time series analysis by borough and month from the dataset. We also decided to do an analysis of the city response to the boroughs by taking the created_date and closed_date and identifying the proportion of not closed incidents by borough. We also took the two dates to calculate response time per incident, then aggregated by borough and found the average. Lastly, we wanted to take a look at other potential factors to the boroughs so we introduced a second data set from the U.S. Census Bureau containing information such as population, income, and poverty by borough. The first thing we found out with this dataset, however, was that every element was a string and not types we could use to create analyses to. So, our first step was adjusting the data itself to allow for manipulation and analysis, and through that, we were able to create bar charts for comparison to the amounts of graffiti in each borough.

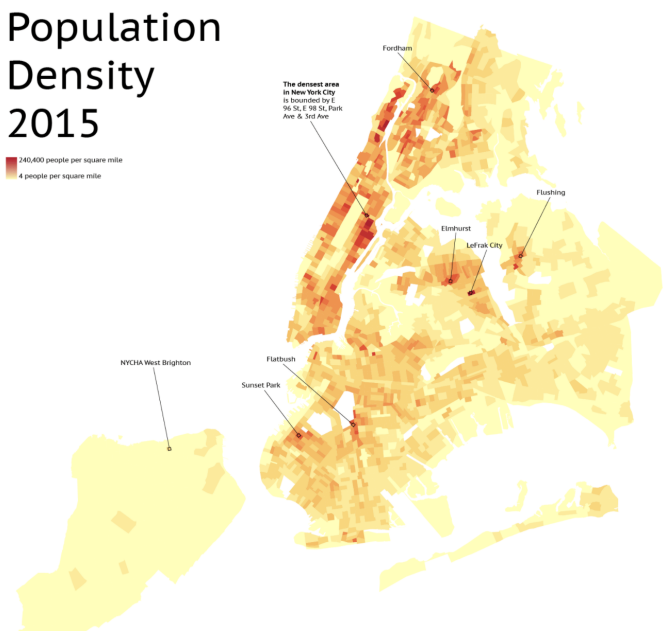
Compelling Text and Data Stories with Charts

Populations and Maps

We begin our analysis by plotting the x and y coordinates of where the graffiti incidents occured in NYC. Below in Figure 1 each black dot represents a single graffiti case in NYC from the dataset. Underneath we have Figure 2 representing the population of NYC using a heatmap to emphasize the difference in numbers. Our intuition made us believe that graffiti cases in boroughs would be proportional to their population. Comparing figure 1 and 2 we see that is not necessarily true. There is a ton of activity near the center of the map (In Manhattan and near the border of the Queens and Brooklyn boroughs) even though there are more people living in the regions above.



(Figure 1: Scatter-plot of XY coordinates of reported NYC Graffiti Cases)

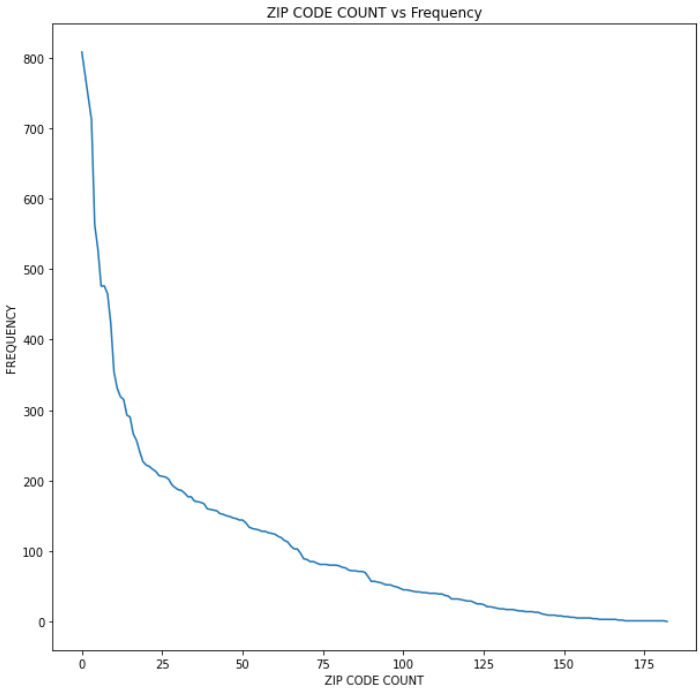


(Figure 2: Heatmap of population density in NYC)

This distinction between our assumptions and data led us to investigate the zip codes of NYC. By exploring the zip codes we get a better understanding of what's going on in specific regions of each borough.

Zip Code Count Versus Frequencies

Geographic analysis is another method we used to analyze the data. First, we inspected the Zip Code data and only ran analysis on the non-empty data, as there was not that much data that was null. We grouped by zip codes and counted the number of graffiti occurrences by zip code. From this, we are able to see what proportion of the graffiti incidents are accounted for by a proportion of zip codes. This is more clearly visualized in the graph below.



(Figure 3: Zip Code Count vs Frequency)

This graph above (Figure 3) demonstrates that the majority of graffiti occurrences happened within a smaller subset of the zip codes. We saw that the top 25% of zip codes held 67% of the graffiti occurrences. Thus, it appears that graffiti occurrences are highly concentrated in a small number of zip codes.

We wanted to investigate further what could cause the top five zip codes to have such a high number of graffiti counts. We thus create the following table using data from an additional source.[1]

	Zip Code	Median Income Zipcode	Median Income in Borough	Median Zipcode income percent of Median Borough Income
0	10002	$33,218	$86,553	0.383788
1	11237	$40,372	$60,231	0.670286
2	11206	$28,559	$60,231	0.474158
3	11221	$39,178	$60,231	0.650462
4	11211	$46,848	$60,231	0.777805

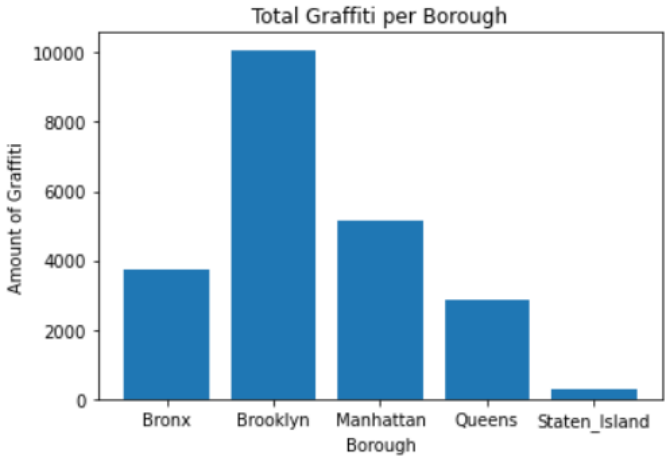
(Figure 4: Top 5 Zip codes with graffiti and relevant stats)

[1] (https://www.unitedstateszipcodes.org/10002/ note: the last number of the link can be changed to the zipcode of interest)

From here we can see that all five of the zip codes have lower median income levels than the average of the boroughs that they are in. The rightmost column is meant to portray just how much lower the zipcodes median income is compared to the borough's. Our sample here is small so it is hard to conclusively say income is a significant factor in explaining the high rates of graffiti in these areas, but the data suggest it is a possibility and is worth noting.

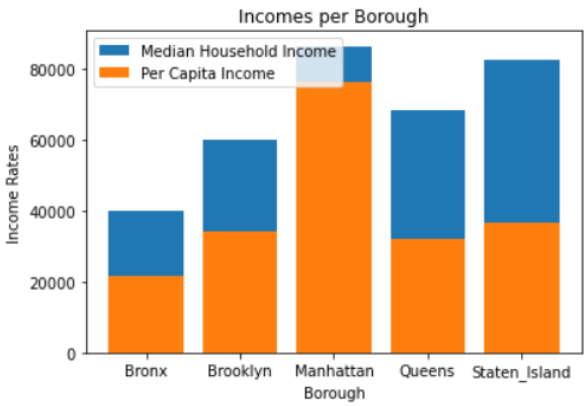
Census Analysis

The first step to gaining an idea of the census data is first gaining an idea of the graffiti reports from the original dataset for each borough. In this graph (Figure 5), we see a much higher level of total graffiti count in Brooklyn compared to all the other boroughs. Manhattan follows Brooklyn, followed by the Bronx, then Queens, and finally, at the lowest level, is Staten Island. Using the census data hopefully gives us insight on potential reasons why there is more graffiti in certain boroughs.

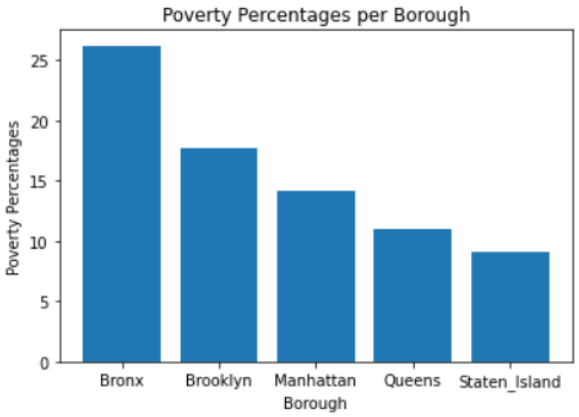


(Figure 5: Total Graffiti per Borough)

Taking the data from the U.S. Census Bureau, we created corresponding bar graphs for each variable and borough to use for comparison. We found that the most relevant variables include average income, poverty levels, and population.

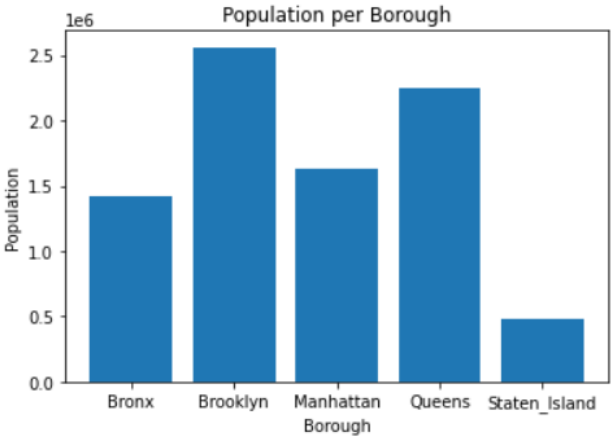


(Figure 6.1: Incomes per Borough)



(Figure 6.2: Poverty Percentages per Borough)

Taking a look at Figure 6.1 about the incomes per borough, we have a double bar chart with the blue showing median household income and the orange showing per capita income. As for Figure 6.2, we see the poverty percentages per borough. We found that there is a sort of correlation between income and poverty. For example, the Bronx has the lower income averages and highest levels of poverty. Manhattan has the highest levels of average income however still has medium percentages of poverty, and some background information about the area and economy in manhattan shows reasons why there is a higher level of poverty than expected. We see that there are most graffiti reports from Brooklyn and though Figure 6.1 and 6.2 give us an idea for all boroughs, it is not enough to make conclusions about the effects on graffiti in boroughs.

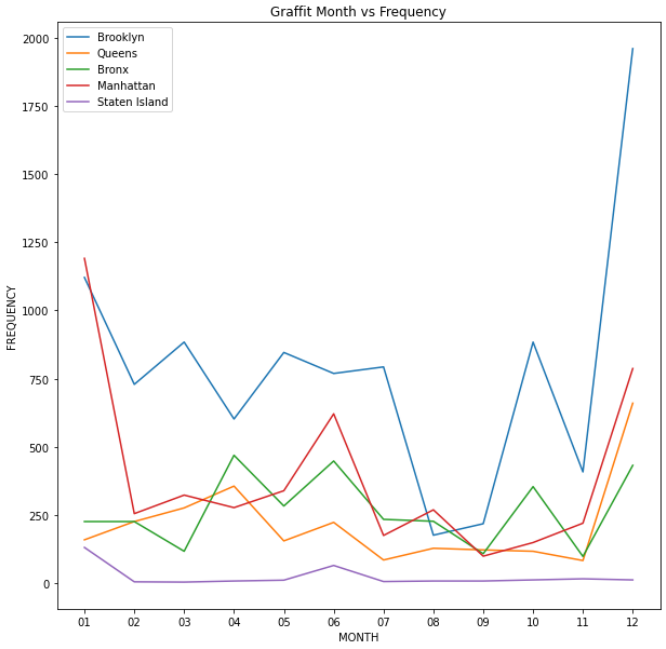


(Figure 7: Populations per Borough)

In Figure 7, we see the population sizes per borough and that Brooklyn actually has the highest population of all the boroughs. With this, we come to the conclusion that Brooklyn having the second highest poverty rate and the highest population size contributes the most to it having the highest number of graffiti. The Bronx had the highest poverty rate yet a lower population size, thus leading to lower graffiti rates. We can also assume that having a higher poverty rate makes reporting graffiti in an area not significant, thus we can also assume there is a large percentage of graffiti that goes unreported. The Manhattan borough has the highest income average and is the second highest borough with graffiti reports. This suggests support for the conclusion that more people report graffiti in the borough since it is a higher-class neighborhood. Queens has the second highest population but has moderately high income averages as well as low rates of poverty which can attribute to low levels of graffiti reports. Finally, Staten Island has extremely low rates of graffiti reports which can be due to their much lower population compared to other boroughs or their high income average and lowest poverty rates.

Graffiti Month vs Frequency

Next, we can inspect the graffiti frequency over time and see the number of occurrences for each month. We can also group by the different boroughs in our dataset, which gives us a graph that displays how much graffiti was reported in each borough for each month of 2020.



(Figure 8: Graffiti Month vs Frequency)

It looks like graffiti incidents are high around the end/beginning of the year. According to the NYCEDC (https://edc.nyc/program/graffiti-free-nyc) the reason there is a spike in the winter months is because "graffiti cannot be removed during the winter months because painting and power washing is not successful when the temperature is close to freezing or below".

We can also see that Brooklyn (the blue line in the graph) has the highest occurrence of graffiti incidences overall even when we compare proportional differences between the boroughs.

Now, we can examine the response times of each borough to see which boroughs respond the fastest and slowest. Looking at the figure below, we can see that Queens has the longest response time. Most of the boroughs have similar response times with the exception of Staten Island, which has a relatively shorter response time compared to the others. Looking at the proportion of closed graffiti cases, we can also see that Staten Island has the lowest number of open cases at 33%.

The other boroughs have a higher open case rate and longer response time. Staten Island responds very well to graffiti reports; however, they only have 305 graffiti reports total in the year which is a small sample size and probably makes it easier for the authorities to respond.

BOROUGH	RESPONSE TIME
QUEENS	137.811636
MANHATTAN	128.699660
BROOKLYN	124.104716
BRONX	123.931112
STATEN ISLAND	116.182266

(Figure 9: Response Time of each borough by days)

BOROUGH	count	proportion of not closed	not closed
Unspecified	12	1.000000	12
BRONX	3742	0.526724	1971
QUEENS	2855	0.518389	1480
MANHATTAN	5149	0.485920	2502
BROOKLYN	10076	0.484418	4881
STATEN ISLAND	305	0.334426	102

(Figure 10: Open Case Proportion)

Conclusion

In essence, our investigation of the NYC graffiti and census datasets revealed a few important findings. First, population was not entirely proportional to the number of graffiti incidents, and census and zip code analysis hinted that after taking into account population, low-median-income levels and poverty were better predictors of graffiti counts. We discovered that the increase in graffiti incidents in the winter was due to the difficulty for cleaning services to resolve the graffiti incidents in cold temperatures. Finally, we saw that all the boroughs in New York City have a similar response time and case closing rate for graffiti with the exception of Staten Island (which has a small sample size). We uncovered interesting characteristics of New York City and its graffiti. Data is a powerful tool we can continue to leverage to gain more insights about ourselves and the places we live in.

Works Cited

Department of Sanitation (DSNY) (2013). *DSNY Graffiti Tracking,* NYC OpenData.

https://data.cityofnewyork.us/City-Government/DSNY-Graffiti-Tracking/gpwd-npar

United States Census Bureau (2020). *QuickFacts New York city, New York*.

https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork,bronxcountybronxboroughnewyork,kingscountybrooklynboroughnewyork,newyorkcountymanhattanboroughnewyork,queenscountyqueensboroughnewyork,richmondcountystatenislandboroughnewyork/PST045219

Population Density Map:

https://viewing.nyc/this-density-map-shows-how-we-crowd-85-million-people-in-new-york-city/

Graffiti Cleaning Information:

https://edc.nyc/program/graffiti-free-nyc

Zip Code Supplemental Data:

https://www.unitedstateszipcodes.org/