Capstone Project 1: Milestone Report

1. Problem Statement

A building permit is an official approval issued by the local governmental agency that allows you or your contractor to proceed with a construction or remodeling project on your property. Those who would like to construct, enlarge, alter, repair, move, remove, improve, convert, or demolish a building or other structure typically apply for a building permit. Unfortunately, building permits are often times the longest item in a project schedule; e.g. 2-3 years for a newly constructed low-rise building. Cities review applications on a first-come, first-serve basis and must check relevant plans to ensure they meet local building codes. Could I predict the building permit issue time for New York City? Some important use cases for this problem may include:

- **Real estate developers** may use the prediction to better optimize their project portfolio to consider reduced project timelines, which should improve their bottom line.
- **Homeowners**, especially those inexperienced in the building permitting process, will obtain more accurate time estimates for permitting and reduce reliance on anecdotal advice from building contractors/building department.
- **Building departments** could better triage requests to potentially reduce response turnaround time.

2. Data Wrangling

a. Cleaning Steps

The data was downloaded as of May 9th, 2018 from NYC Open Data. The .csv data contains a list of permits issued for a particular day and associated data. Prior weekly and monthly reports are archived at DOB and are not available on NYC Open Data. The raw building permit data contains 60 unique columns and nearly 3.37 million rows of data. A separate .csv file provides additional detailed descriptions about the data.

The data was then imported into a Pandas DataFrame for ease of data manipulations. Feature names were adjusted to be short yet meaningful, free of spaces via replacement using underscores and converted to lowercase. Twenty features out of the original 60 were kept for use in the EDA and machine learning stages to keep only relevant features in the scope of the study.

Feature relevance was based on visual inspection of the records and partially due to the proportion of null/empty records. The raw data contained building permits issued dating back to the 1980s, however I decided to only use data from the past five years because it is likely to be more relevant from a prediction standpoint; processes have likely changed to reduce building permit issue times. However, I may use permits issued up to 10 years ago if additional data is required to correct class imbalance issues at the machine learning stage.

https://data.cityofnewyork.us/Housing-Development/DOB-Permit-Issuance/ipu4-2q9a

b. Missing Values

Missing values were primarily due to whether the feature is required or optional for a building permit. Even if the data is required, sometimes null values were still present as a small proportion of all of the records. In Python, specifically Pandas, NumPy and Scikit-Learn, the standard practice is to mark missing values as NaN. Values with a NaN value are ignored from operations like sum, count, etc. We can mark values as NaN easily with the Pandas DataFrame by using the replace() function on a subset of the columns we are interested in. After we have marked the missing values, we can use the isnull() function to mark all of the NaN values in the dataset as True and get a count of the missing values for each column.

After marking these records as NaN, I decided that it was acceptable to remove records where the proportion of nulls to total records is extremely small, e.g. the nulls in the latitude feature were <1% of the total number of records. For other features, I've decided to use a decision tree classifier to automatically categorize the empty records into their own category and calculate their log likelihoods, which avoids data imputation, data removal, and throws no errors using machine learning algorithms. This decision tree approach will be done at a later stage.

c. Outliers

Outliers may be due to measurement/input error, data corruption, or perhaps actually represent legitimate behavior. I used Tukey fences as a simple way to determine outliers, i.e. a record is considered an outlier if it is 1.5 times the interquartile range. The dependent variable, issue time (days), is heavily right skewed based on its histogram. From the histogram, I observed negative values for the issue time, which I subsequently removed from the analysis because of suspected input error. On the other hand, outliers are present for permits requiring multiple years to be issued, and I suspect these represent legitimate records. For example, new buildings, especially high-rise and skyscrapers, typically require a much longer time to obtain a building permit compared to an average project. From the data, I notice that in the last five years some of these projects needed up to four years to obtain a building permit. One approach is to simply trim the bottom 1-5% of data, but there may be value in understanding the factors that lead to multi-year building permit issue times. However, it would be a good idea to trim permits with less than a month to issue. Permits with issue times less than a month are primarily from "quick-fix" projects, e.g. minor electrical work, which are generally more predictable and likely don't provide much value to the client if included in this study.

3. Exploratory Data Analysis

a. Initial Trends and Questions Explored

I initially explored the categorical data for any potential outliers and trends by creating and answering questions, including:

- **What are the most frequent permit types submitted and are there outliers?** From the subplots below, electrical work is the most popular permit type. Permit types for DM (Demolition and Removal) and FO (Foundations) may be considered outliers. They are considered outliers because their share of the data is much smaller compared to the other categories in their feature.
- **Are most permits issued?** Yes, it seems like the vast majority of permits are issued. Permit statuses that are in-process and re-issued may be considered outliers.
- **Are most submittals first-time or renewals?** It seems like initial submittals are more than two times than higher than renewals.
- **How do the boroughs rank in building permit frequency?** 1) Manhattan, 2) Brooklyn, 3) Queens, 4) Bronx, 5) Staten Island. Based on this ranking, it'd be interesting to further explore geographic hotspots for building permits and issue times.

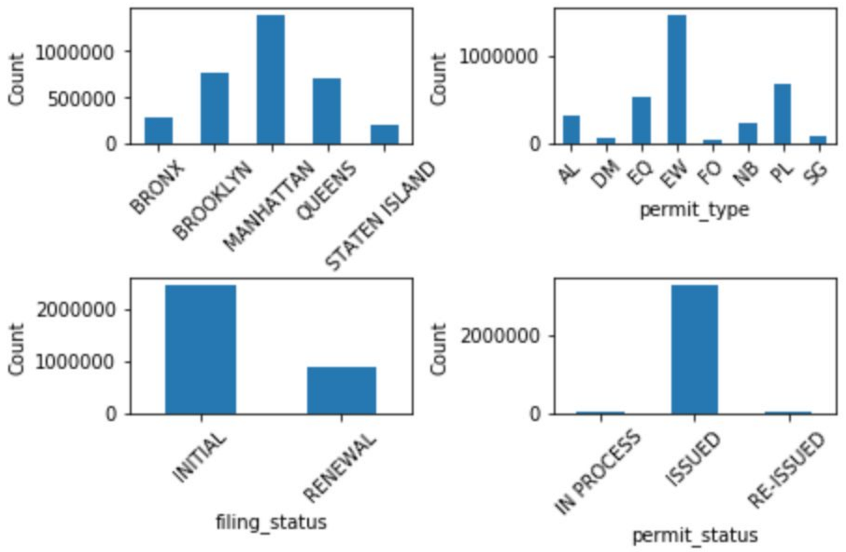

Figure 1. Categorical Features Bar Graphs

Now we turn our attention to some important numerical features to initially explore the questions below. I've plotted an overlapping time series to inspect any anomalies in the frequency of building permits over the years and identify trends.

- **Are most filed permits issued as well?** The issuance and filing are practically following the same trend showing that almost all of the building permits that are filed are also issued.
- **How has building permits issuance frequency changed over time?** From visual inspection I notice that since the 2008 U.S. recession, building permits have been relatively chaotic in their issuance and filing, but generally on a slow upwards trend. This trend makes sense because the building industry is a notoriously cyclical industry that has a strong correlation with the U.S. economy. I may explore how the recession impacted permit issuance time; I hypothesize that building permit issue times are lengthier due to the economic slowdown.

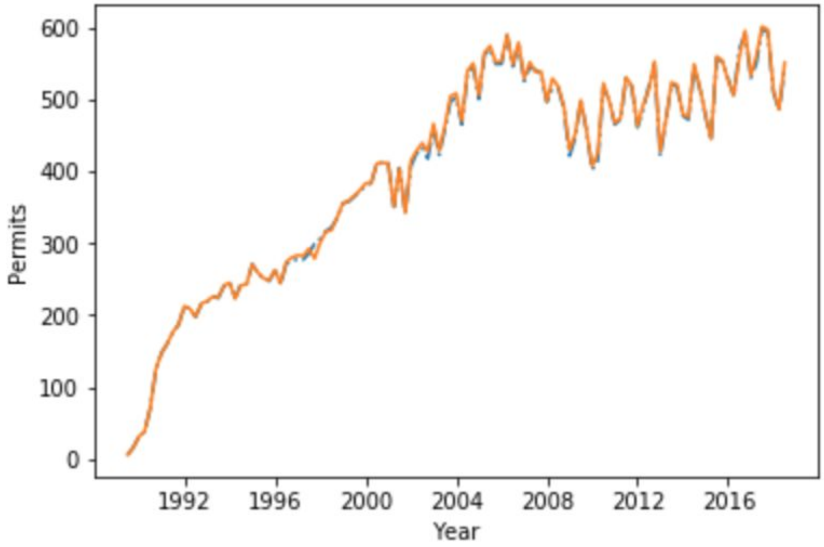

Figure 2. Issued and Filed Permits Time Series

When I initially explored the issue time feature, I found that a majority of the permits were issued on the same-day, and that the data was highly skewed to the right. I decided to omit same-day issue times because I do not believe there is much value in knowing if they could get a permit the same-day compared to permits that take months if not years. I then broke up the issue time data into four discrete time ranges to plot as histograms: 1) one week, 2) one week to one month, 3) one to six months, and 4) over six months. Based on the visualization, I asked and answered the following questions:

- **Is there enough data to continue EDA after filtering the data?** Yes, there seems to be thousands of records still available even after filtering. I will likely have to use oversampling/undersampling techniques given the large skewness.
- **Is the data still skewed to the right and why?** Yes, and this is likely because buildings that require longer issuance times generally require a more careful inspection of plan sets. I hypothesize that the data follows an exponential distribution which is the time taken between two events occuring (filing time to issuing time).

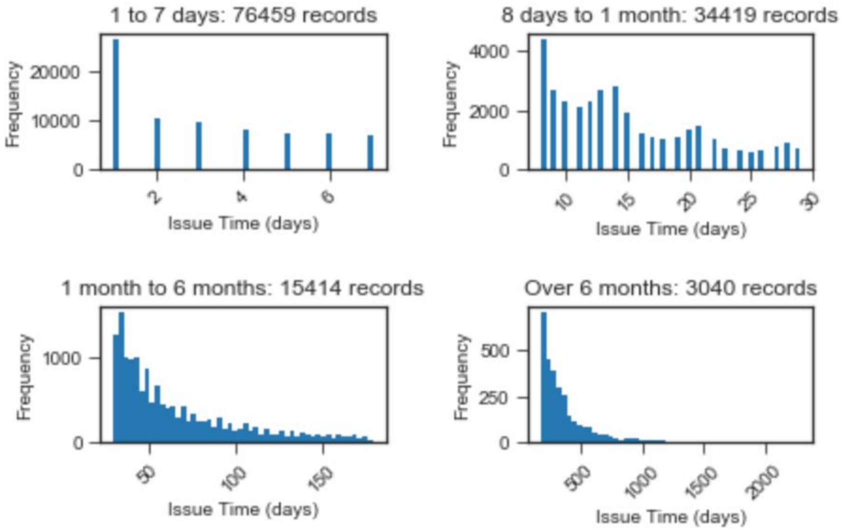

Figure 3. Issue Time Divided into Four Time Ranges

b. Statistical Analyses

I started with exploring the summary statistics and plotting boxplots, bar graphs, and empirical CDFs to better understand the data. The key finding was that the wait time is on average four months and that the distribution of issue times is highly variable with a standard deviation to mean, or coefficient of variation, greater than 1.0. Because the data has such a large spread, I anticipate issues satisfying regression assumptions and will likely need to consider classification-based ML algorithms. I have preemptively created a two and three class feature based on the continuous dependent variable, issue time.

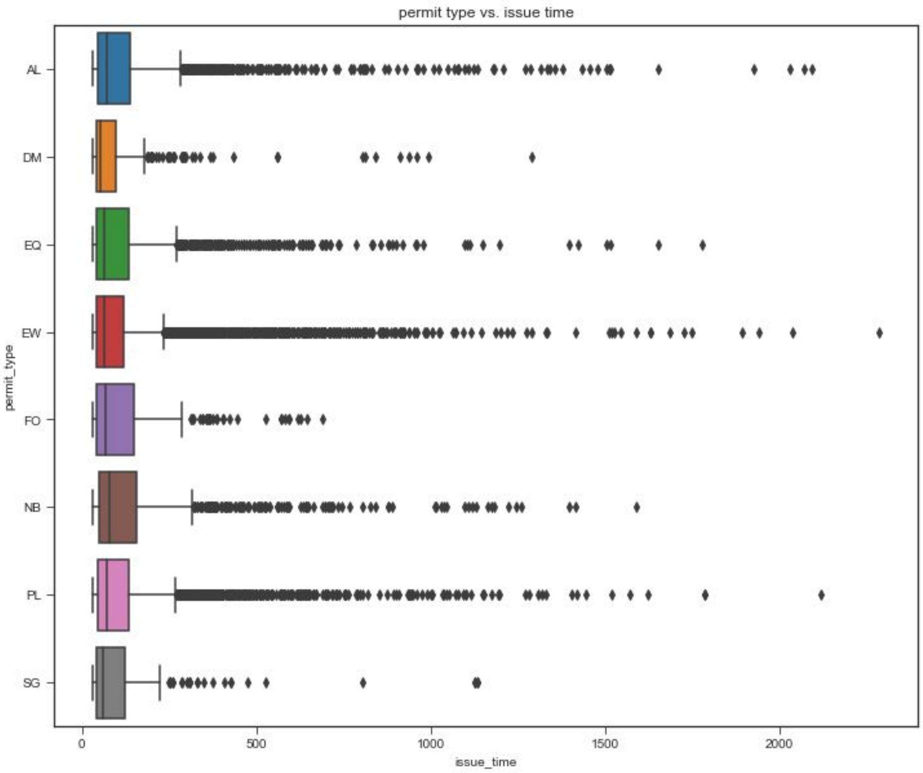

Figure 4. Permit Type Box and Whisker Plots

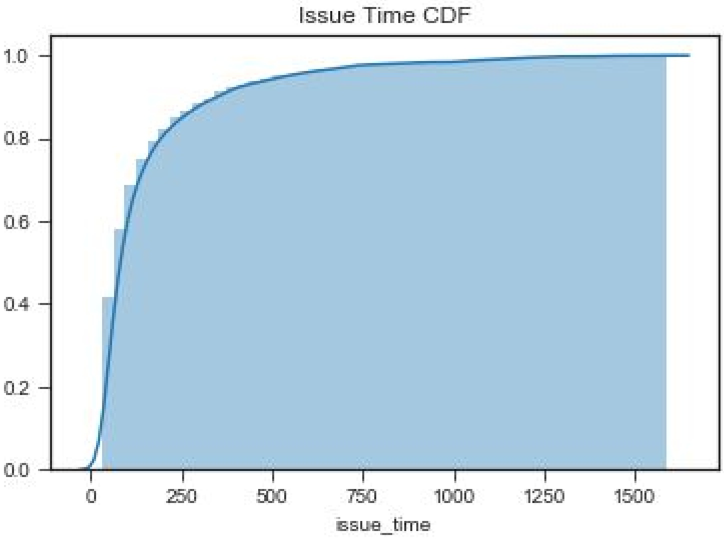

Figure 5. "New Building" Permit Type Empirical CDF

Next, I looked for strong correlations between pairs of independent variables or between an independent and a dependent variable for the numerical features. Plotting a correlation coefficient heat map, we can see that the most negative correlation between council district and latitude. The most positive correlation is between the census tract and the longitude. I found that location-based features, e.g. Council district and longitude, seem to have a statistically significant correlation with the issue time. I used a spearman rank correlation coefficient instead of the pearson rank correlation coefficient. A pearson's correlation works well if the variables are roughly normal and outliers are not present. It is a measure of the linear relationship between variables. Spearman's rank correlation, however, is a better alternative that mitigates the effect of outliers and skewed distributions.

- longitude: test-statistic = 0.023   p-value = 0.0026
- council_district: test-statistic = 0.031   p-value = 0.0000
- census_tract: test-statistic = 0.020   p-value = 0.0083

In the machine learning phase, I will use feature engineering based on the building permit location to potentially include:

- Median income
- Crime data
- Community board socio-economic and demographic data

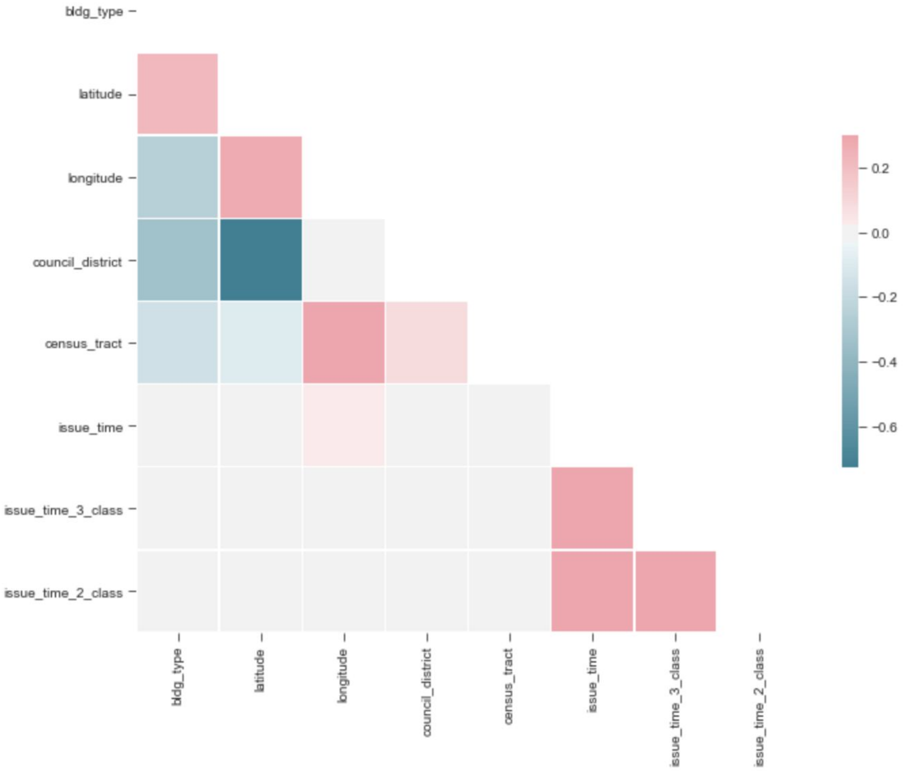

Figure 6. Correlation Matrix

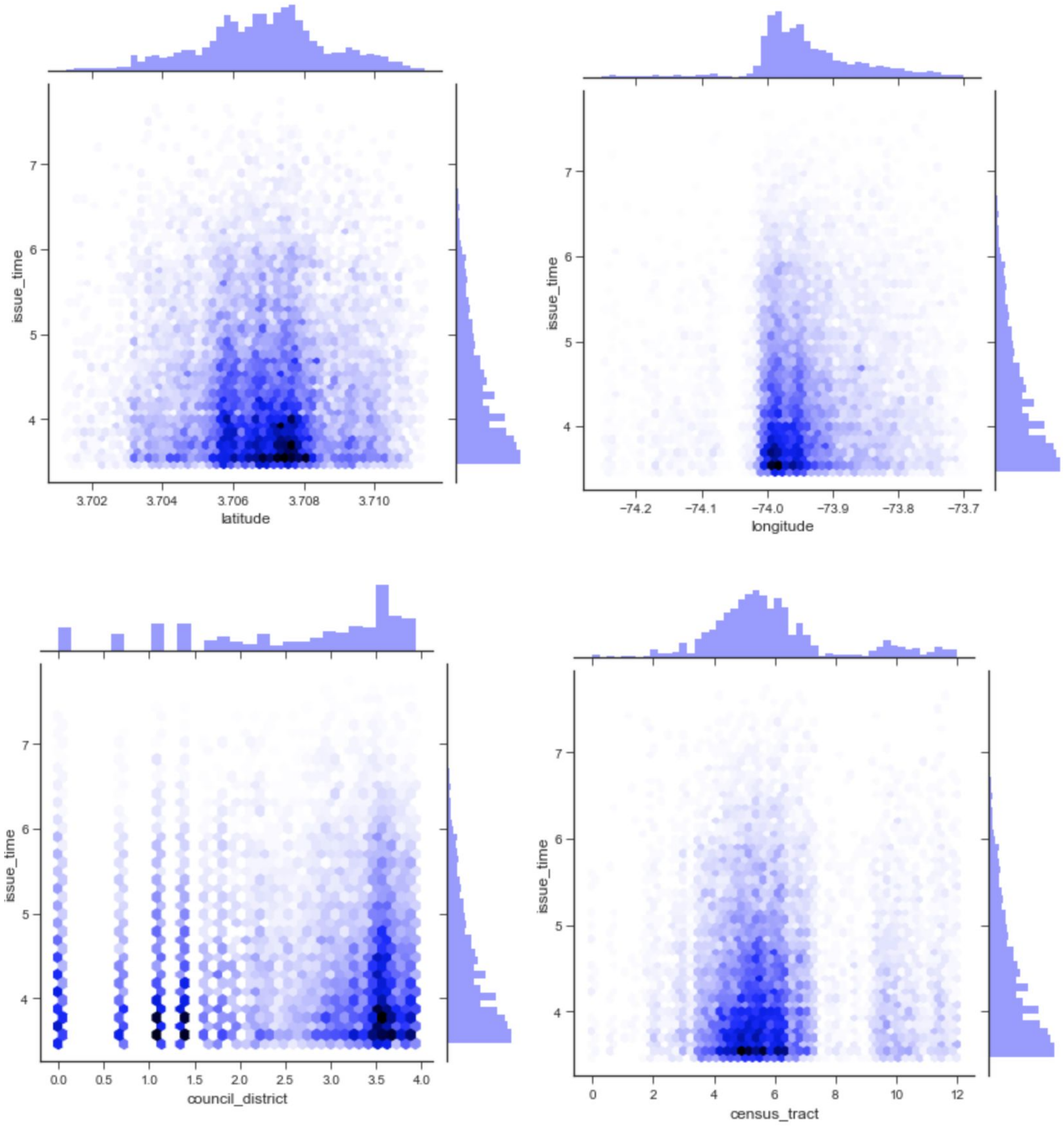

Figure 7. Hex-binned Scatter Plots

From hypothesis testing, I tested a claim regarding the average wait time for building permits. Quoting a building permit expediting services company, they claimed that "complex projects," i.e. new building construction, could take 6 months or more to obtain an issued permit. Using a left-tailed Z-Test and bootstrapping methods, I found that building permits for "complex projects" actually require less than 6 months on average based on a sample size of 1109 and p-value of 0.000.

I used a heat map visualization based on the Google Maps API to obtain some insight into the neighborhood hotspots for building permits. These neighborhoods seem to coincide with regions of high economic activity based on my domain knowledge growing up in NYC. Economic data would be a great addition to this project and may help with the prediction of building permit issue times.

- **Brooklyn**: Park Slope, Bushwick, Brighton Beach, Williamsburg
- **Manhattan**: Majority of neighborhoods
- **Queens**: Flushing, Jackson Heights, Elmhurst, Astoria, College Point
- **Bronx**: N/A
- **Staten Island**: N/A

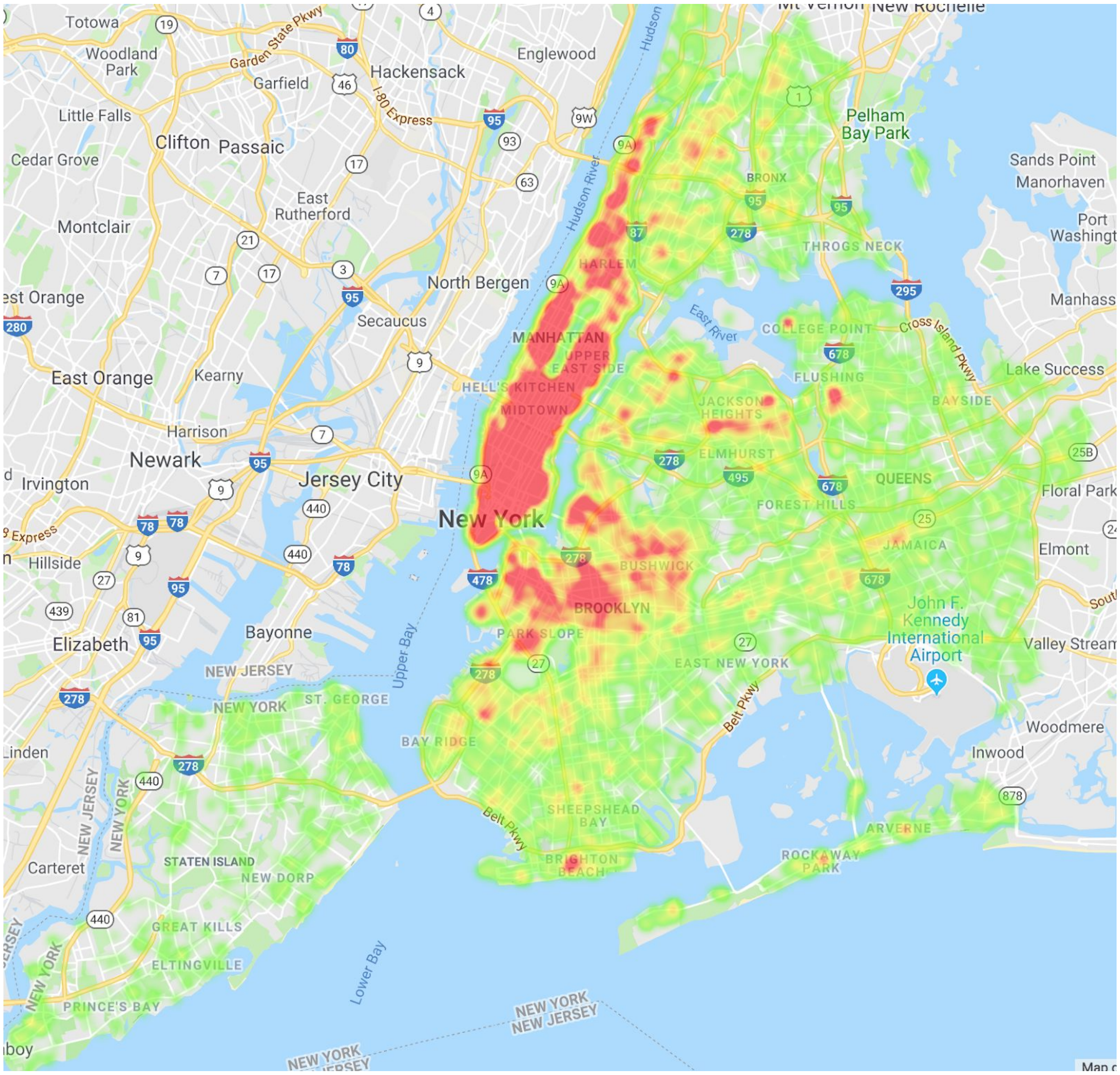

Figure 8. Building Permit Heatmap