Capstone Project 1: Data Story

Github Jupyter Notebook Link:
https://github.com/dtse91/Springboard/blob/master/Capstone%201%20Project/Capstone%20Project%201%20Data%20Wrangling.ipynb

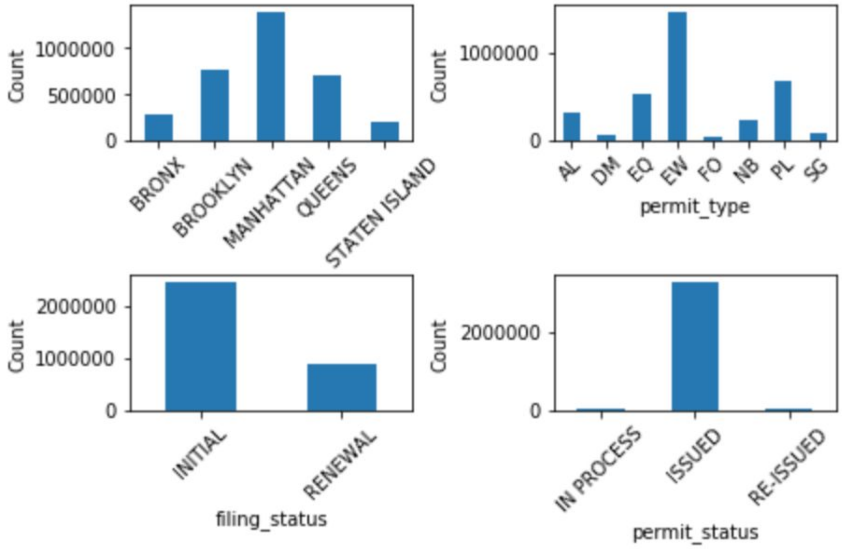
I initially explored the categorical data for any potential outliers and trends to better prepare for the EDA phase.

- **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.

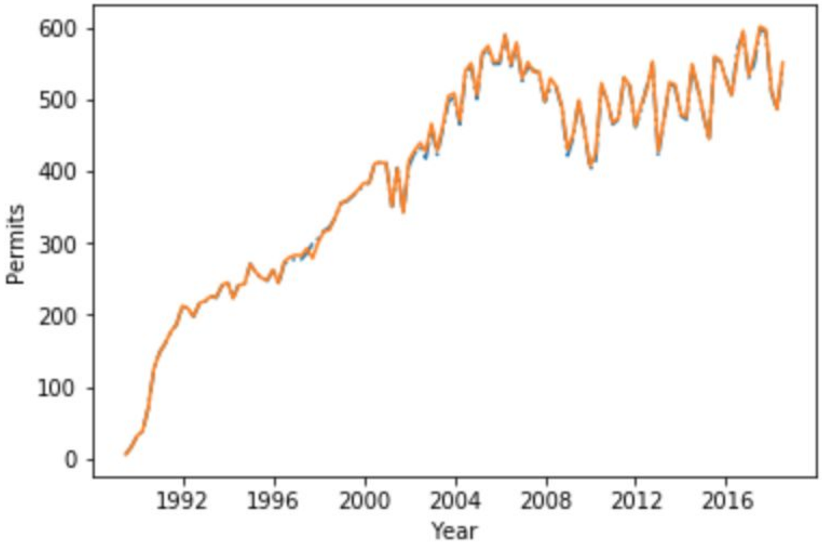
- **How do the boroughs rank in building permit frequency?** 1) Manhattan, 2) Brooklyn, 3) Queens, 4) Bronx, 5) Staten Island. Based on this ranking, it'd be interesting to further explore geographic hotspots for building permits and issue times.



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.

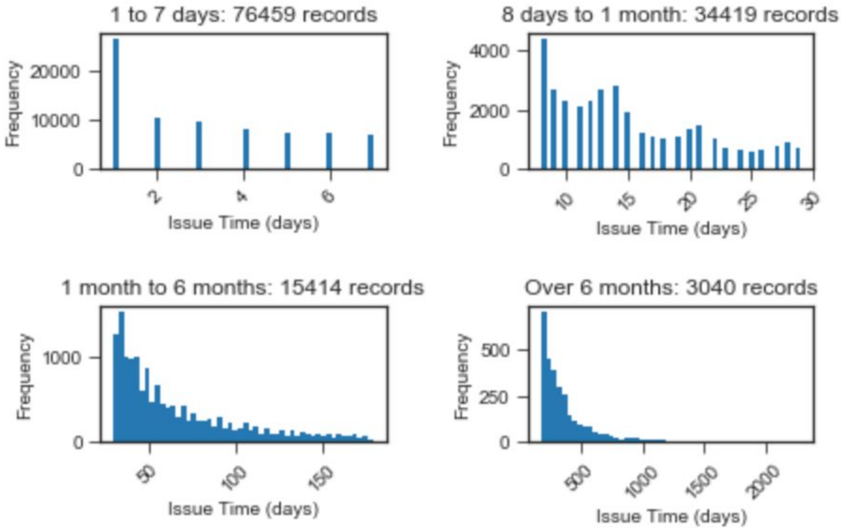
- **How has building permits issuance frequency changed over time?** From visual inspection I notice that since the 2008 U.S. recession, building permits have been relatively chaotic in their issuance and filing, but generally on a slow upwards trend. This trend makes sense because the building industry 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.



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 question:

- **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).


1 to 7 days: 76459 records
Frequency
Issue Time (days)
8 days to 1 month: 34419 records
Frequency
Issue Time (days)
1 month to 6 months: 15414 records
Frequency
Issue Time (days)
Over 6 months: 3040 records
Frequency
Issue Time (days)