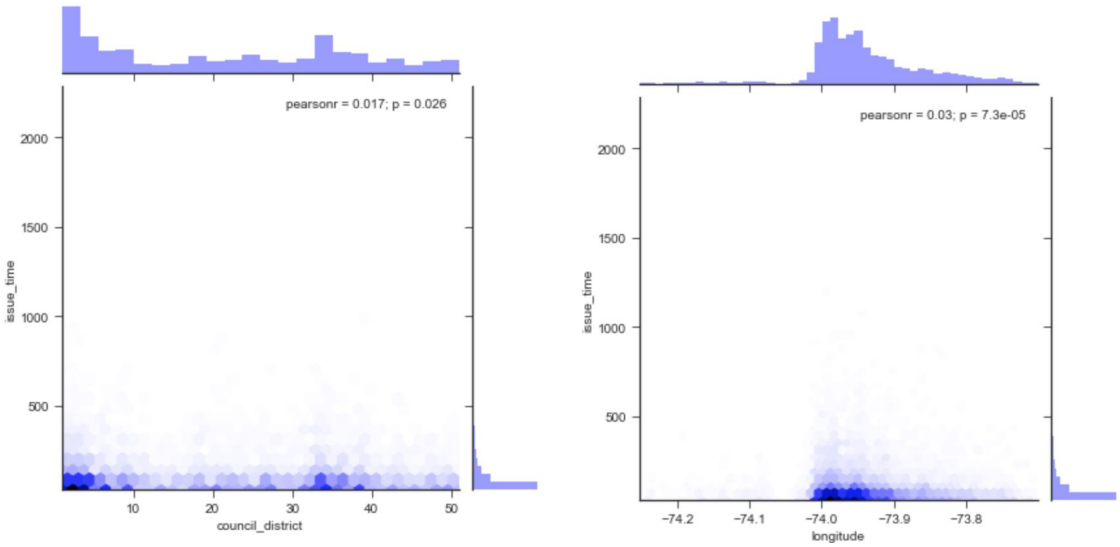
Capstone Project 1: Exploratory Data Analysis (EDA)

Github Jupyter Notebook Link:
https://github.com/dtse91/Springboard/blob/master/Capstone%201%20Project/Capstone%20Project%201%20Exploratory%20Data%20Analysis%20(EDA).ipynb

- Are there variables that are particularly significant in terms of explaining the answer to your project question?

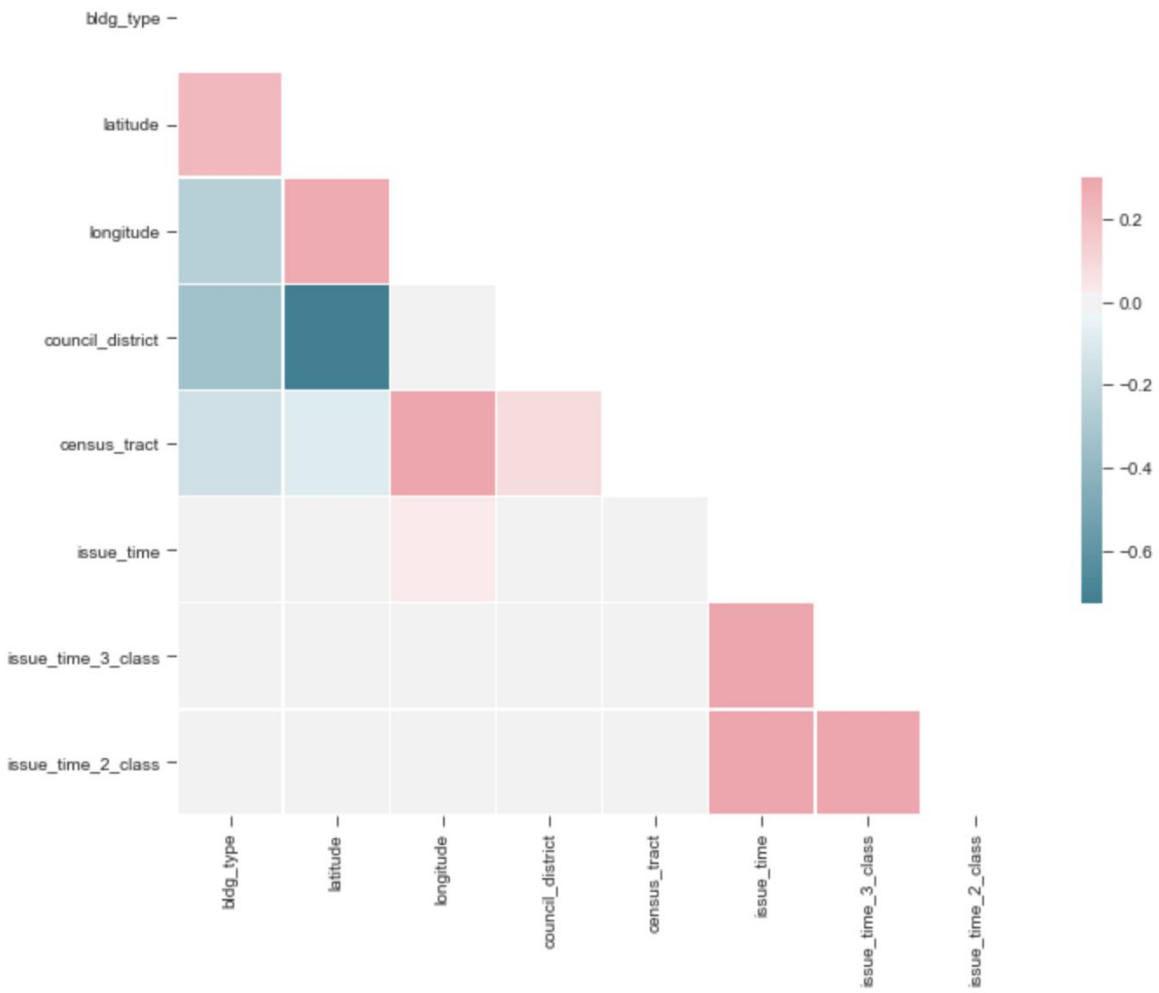
Location-based variables seem to be important in predicting building permits. I found that location-based features seem to have a statistically significant correlation with the issue time. In the machine learning phase, I will use feature engineering based on the longitude and latitude to potentially include:

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



- Are there strong correlations between pairs of independent variables or between an independent and a dependent variable

Plotting a pearson 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.



- What are the most appropriate tests to use to analyse these relationships?

I primarily used Pearson correlation hypothesis testing to determine statistical significance. I also used a left-tailed z-test to determine whether or not a claim made by a building permit expeditor is statistically significant. I determined that the average building permit time is less than six months.