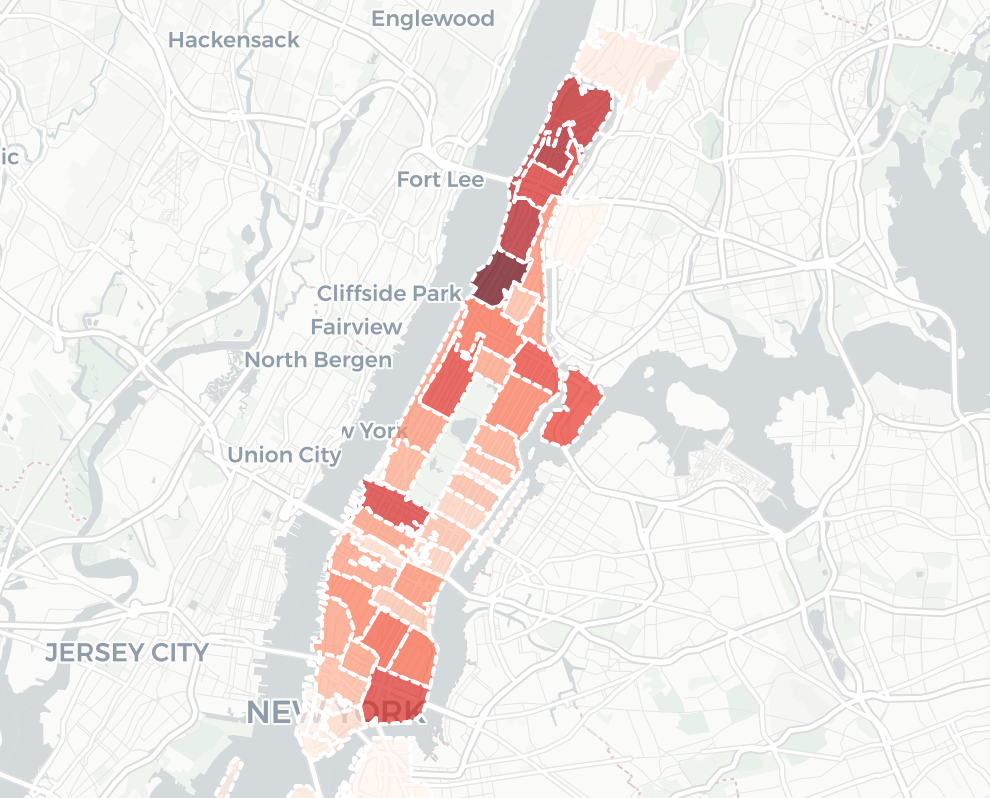
Part One - Data description

Both of the two datasets have features about borough, and we found four boroughs are in both datasets, which are Bronx, Brooklyn, Manhattan, Queens. We decided to only focus on Manhattan first and study the relationship between average salary and cases of complaints across different department to study if resources are allocated properly, then the same logic can apply to other boroughs.

Jiaxin Li is mainly responsible for the NYC 311 dataset. She generated zip\_code from latitude and longitude in the dataset and built a choropleth map to visualize the cases of complaints across different area in Manhattan.

Zachery is mainly responsible for the NYC payroll dataset. He looked up NYC fiscal year, which starts on July 1 and he wrote a function that takes the date and gives back the year. And he also wrote a function that converts pay data from character into doubles.

Since there are more than 200 types complaints in NYC 311 data, we only focus on complaints whose frequency is over 150. There are a little over 1000 complaints a day on average in Manhattan. At 10,000 samples that is close to 10 days of random samples. There we have complaints that come in at over 15 times a day in those samples with the complaints being above 150. Then we get departments are responsible to solve those complaints. We repeat this 10 times to get 100 random days throughout the year. This out put will give us department names and the frequency with which they happened in each of the 10 times the for loop ran.

We combine the NYC payroll dataset and NYC 311 dataset by department name. Since the department names do not match in each dataset, we get the names for all the departments manually. What’s more, we also removed the departments that fall under larger departments. For example, the Traffic Management Center is under the New York City Police Department, and the Division of Alternative Management is under the Department of Housing Preservation and Development. In these cases, the Traffic Management Center and the Division of Alternative Management are removed and all their records are transferred to the New York City Police Department and the Department of Housing Preservation and Development.

Part Two - Feature engineering

The first feature we generated is called total complaints, which can give us a sense of how many complaints each department is responsible for.

The second feature we generated is called average salary, which describes how much money people in each department can earn.

The tertiary feature is called per complaint(which means the salary per case of complaints), which is to divide average salary of a certain department by cases of complaints. This feature is built to scale salary across different departments.

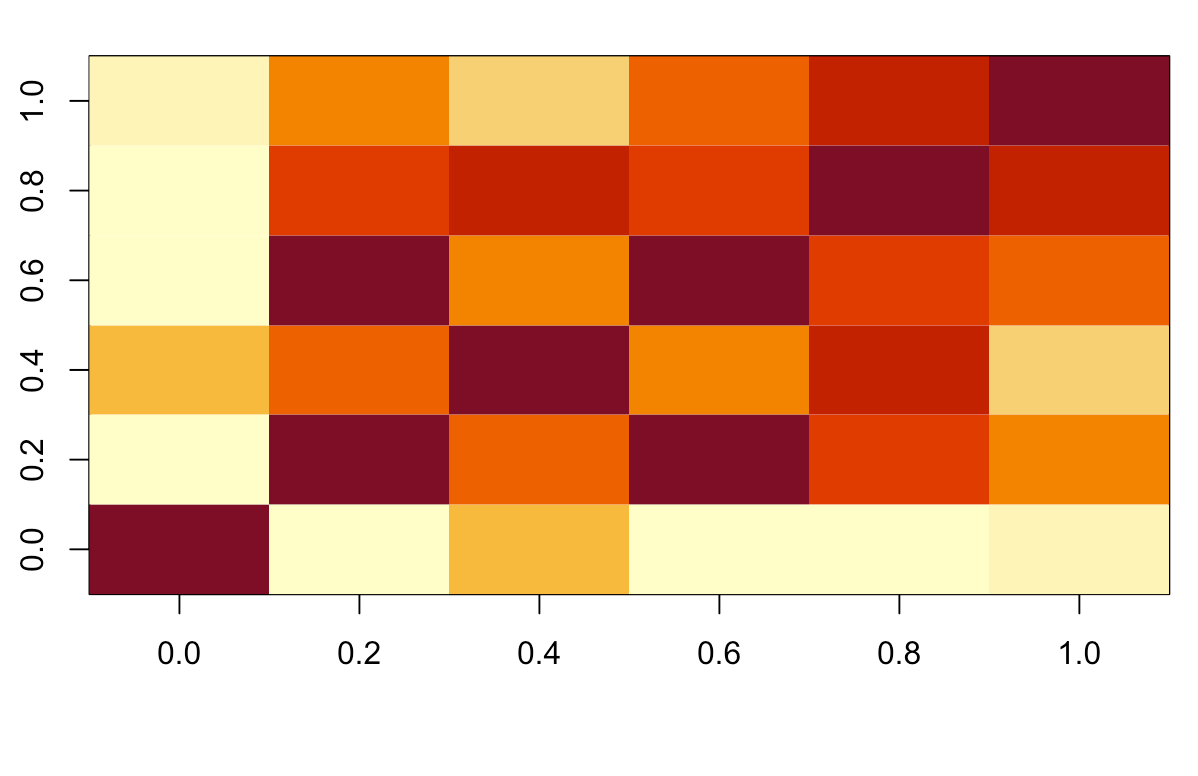
The fourth feature is called factor, and it determines how many types complaints that each department is responsible for.

The fifth feature is called employee count, which represents the number of employees in each department.

PART THREE MODEL BUILDING

After all the feature engineering process, we got a table containing description of each departments. Each row represents a department,

and each column is a feature of the department.

 Descriptions of each department

We treated each feature as dependent variable in turn and do the linear regression again the remaining features. When we took employee count as dependent variables, we found the coefficient of total complaints is negative. So as complaints go up, the amount of people in the department is supposed to go down. It is very interesting. We went further to understand why there is a negative relationship between employee count and total complaints, and we found it might because there is collinearity between variables. For example, there is strong collinearity between per complaint and average salary.