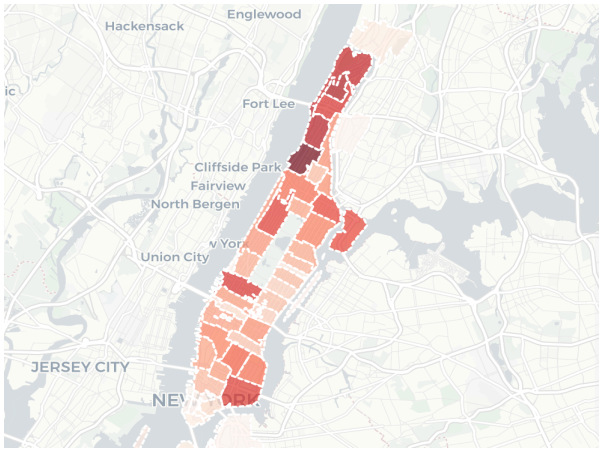
Zachery Macintyre
Jiaxin Li

Intro:

Being Columbia students, we have access to one of the best cities in the entire world. With 8 million people during non-Covid times NYC is the most densely packed metropolitan in the US. With all those people in such a tight area, there are bound to be complaints and things that go wrong. In our project we investigate how NYC 311 data is related to NYC payroll data. Our investigation turned out an interesting finding on how the total number of important 311 complaints are related negatively to the budget of the department. In other words, as complaints go up, total budget goes down.

Data Description:

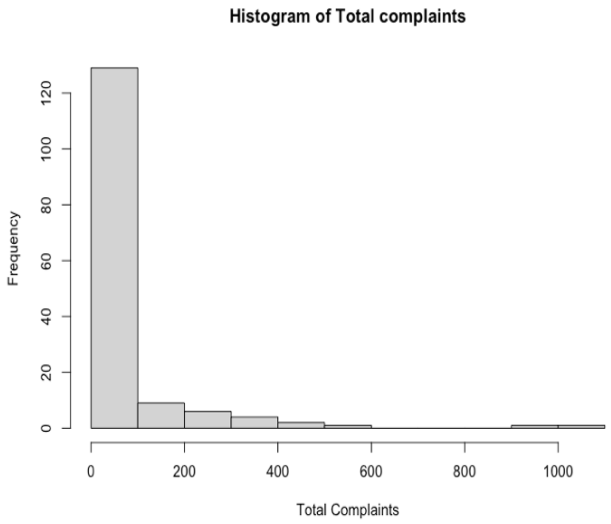
For the data we used a Socrata API to pull publicly available data off NYC OpenData. We focused on Manhattan, as it is where we both live, and where Columbia is located. Additionally, we kept our focus to a 3-year period between 2015-2017, this limiter was used to control the total amount of data. The 3 years gave us over 1.4 million rows of 311 complaints and 590,000 rows of NYC payroll data and we felt this sufficient to do analyses on.



Now, on to the actual data sets. Jiaxin Li was 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 Manhattan from 2015-2017. We felt this was a good way to visualize complaints. This is more of exploratory analysis, and our results did not focus on neighborhoods within the borough. We also used the 311 data to get very common complaints. Common complaints being ones that occur more than 15 times a day. There is more on this later in the section.

As for the NYC payroll data, Zac was in charge of cleaning it. He made a new column for total pay and changed all the numeric values into doubles. For total pay, it was base salary + overtime pay, if there were any missing values the total pay would be missing too. With so many rows, it did not seem too important to remove missing values, just skip over them.

In order to do any sort of analysis we needed to merge the data set somehow. The difficulty in this lies with the fact the department names do not match up between the two sets. So, we



created a new data frame. We sampled 10,000 random rows of 311 complaints, this is approximately 10 days of complaints. From there we focused on complaints that happened more than 150 times within those 10,000 samples. This would mean that the complaint was received on average 15 times a day, and we could eliminate complaints that do not occur very often. We then got the department that was linked to these complaints, and we considered them important departments if they deal with such a high volume of complaints. We ran this loop 10 times to get roughly 100 days of samples.

Finally, we combined the NYC payroll dataset and NYC 311 dataset by department name. Since the department names do not match in each dataset, we got the names for all the departments manually, and fixed them. Additionally, 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 were removed and all their records were transferred to the New York City Police Department and the Department of Housing Preservation and Development.

Feature Engineering:

In this new data frame, the columns are [**per complaint, total complaints, average salary (of dept), factor, max salary (of dept), employee count (of dept), percent of budget**] and the rows corresponded to the departments. We feature engineered 2 of these, the first being per complaint. This feature was made by doing average salary/total complaints in the sample. We created this because we felt like it gives dollar value of what a 311 complaint costs an average employee to deal with. These values fell anywhere between 20 to 350 dollars depending on the department that responded. We additionally engineered another feature, factor. Factor is how many of the major complaints correspond to one department. For example, the department of housing responds to heat/hot water, plumbing, and other things, but the department of homeless services only deals with homeless person assistance. Seeing, as how it hard to rate which complaint might cost more to deal with. We created a factor measurement where every major complaint you deal with increases your departments factor by 1. The police and housing department usually had the highest factors. Again, we did this to signal importance of a department outside of trying to create a dollar amount for what the departments respond to. Percent of budget is the average total budget over the years divided by NYC average budget over the years. It was engineered too, but we did not consider it a unique feature. It was just a way to make a good dependent variable. The other values are

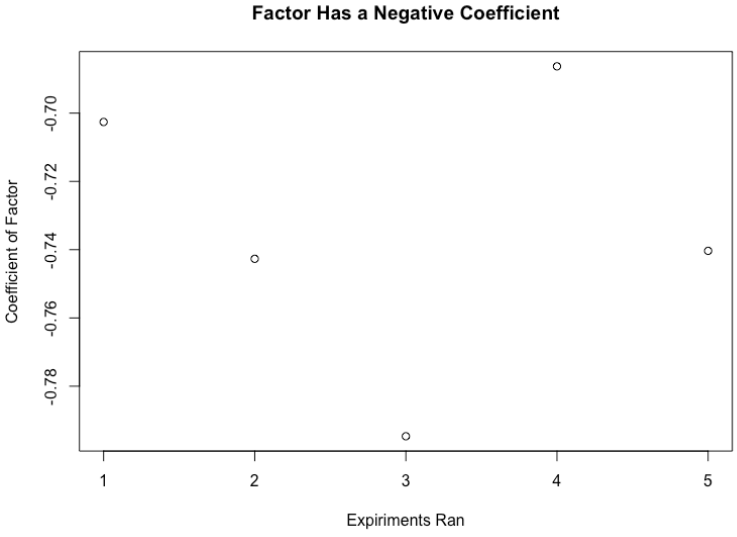
exactly what their title states. We broke everything down by department that was represented in the sample.

Model Building:

With the data frame now in place, we scaled the data down. The argument for this is that things like average salary and max salary are in the 10s of thousands to 100s of thousands and things like factor were rarely above 15. We did not want to have extreme coefficients in a linear regression model because of the extreme differences in features. Thus, we normally scaled all of the data. We planned on running a regression model using percent of budget, as our independent variable. However, we were also interested in all of the columns as the independent variable and seeing if anything interesting happened to the coefficients or significance. Turns out there is a very interesting pattern that emerged. When running the regression model lm(percent of budget ~ the rest of the columns) we saw that factor had negative coefficient. We then, tried controlling for things that are highly correlated to factor such as total complaints, and average salary. However, the results still showed factor to be a negative coefficient.

Take-Aways (Conclusion):

New York City has over 300 departments that have taken and responded to 311 calls. Of those 300+ departments only about 8 get more than 15+ calls every day about some issue. Of those departments 8 were in every sample [**DEPARTMENT OF BUILDINGS, DEPT OF ENVIRONMENT PROTECTION, DEPT OF HEALTH/MENTAL HYGIENE, HOUSING PRESERVATION & DVLPMNT, DEPARTMENT OF TRANSPORTATION, POLICE DEPARTMENT, DEPT. OF HOMELESS SERVICES, and TAXI & LIMOUSINE COMMISSION**]. When controlling just for these high volume 311 departments and running a regression against the percentage of total budget, an interesting phenomenon occurs. As factor increases, we expect total budget percentage to go down. This is counter intuitive because factor measures how many major complaints fall into a department's jurisdiction. One would expect as factor increases the budget to follow. The argument is as follows, if you deal with more types of complaints every day, you need the people to solve different issues. With more people or different types of complaints you would expect costs to go up. However, this is not what we found. Even in situations when you control for other features, factor stays a negative. While the values look small, one must remember all of these have been normally scaled, so these negative values do indeed have pull.



Another interesting takeaway is that our other new feature gives interesting values in general. For example, if the police or housing preservation responds to a complaint, the cost is in the range of 35 dollars. Whereas, if the Taxi and Limousine Commission are the main responders it costs about 335 dollars. There is a factor of 10 difference between the 2. It must be noted that the police respond to a factor of 10 times more complaints than the taxi commission. Ultimately, it is interesting to note the difference in job title and responsibilities for what a 311 complaint can cost the city.