DATA ANALYSIS

TEAM YANKEES

NAKUL PACHERIWALA – NP2455

SAYANTAN MUKHOPADHYAY – SM9752

ABOUT

We will analyze various patterns, form clusters, and visualize the data to make conclusions by performing exploratory data analysis. We will be visualizing the results using charts and maps. By analyzing the data using big data technologies, we can gain some insights into some of the important issues and co-related problems in the city and draw up conclusions that might help the authorities in various administrations.

THE DATASET

The dataset we have used is New York City's 311 data. It contains a list of service requests and 311 calls made by the residents of New York City to various departments in the city. It is available for free from NYC OpenData at It is available at https://opendata.cityofnewyork.us/. It has around 28.4 million rows and 41 columns.

TECHNOLOGIES USED

- PySpark for processing data
- Pandas to store the results
- Plotly express make visualizations
- SodaPy to get the dataset from an API
- GeoJSON to mark co-ordinates of each neighborhood based on its zip code

LOADING THE DATA

We have loaded the data into a Spark data frame. We have used SodaPy to scrape the data from the Socrata API. To expedite the processing of data for this experiment, we have limited the number of rows to 200,000, however, SodaPy allows any number of rows as required to be loaded, including the whole dataset.

PRE-PROCESSING THE DATA

We converted the date and time into Pandas date-time format and calculated the day of the week.

While most ZIP codes were in 5-digit format, some were in 5 + 4-digit format. We removed incorrect zip codes with the incorrect format and converted them all to a 5-digit integer format.

We calculated the time taken to resolve a complaint by taking the difference between the start date and resolution date.

We observed that there are multiple types of noise complaints like residential, commercial, street horn, etc. We combined all of these into a common noise complaint section as they are a subpart of the same problem.

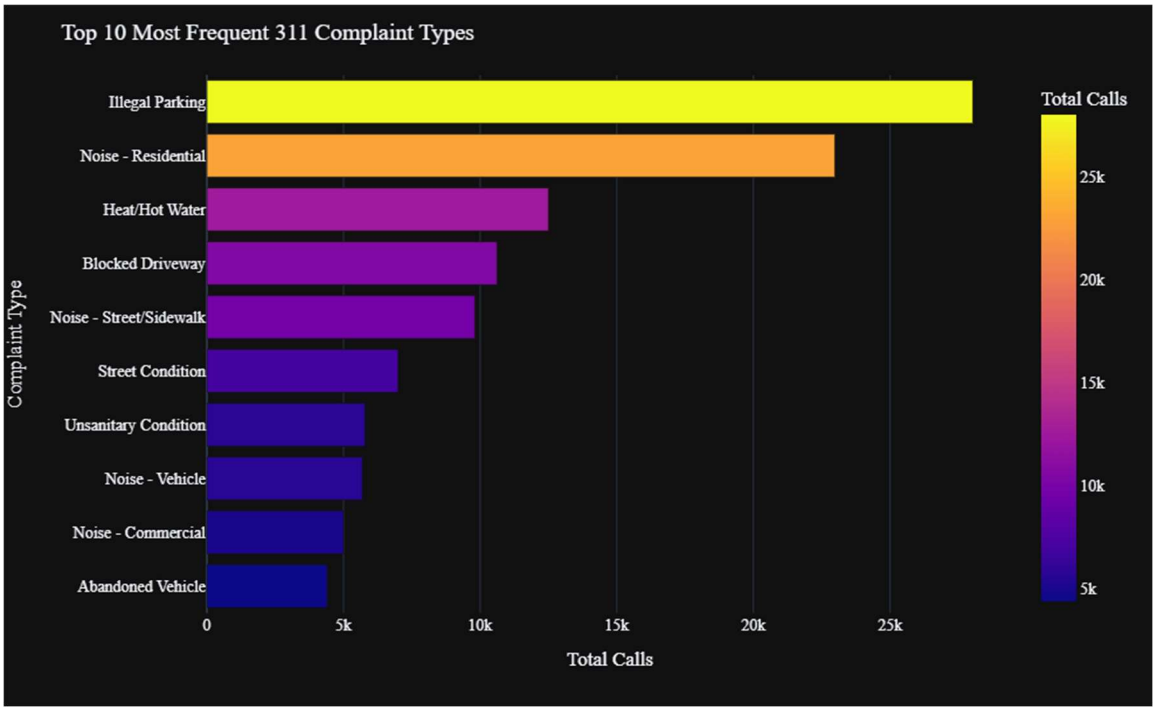
After performing all aggregations, we transformed the output to a Pandas data frame which allows us to make visualizations using plotly. After experimenting we concluded that toPandas() is the most efficient method.

We used regex to eliminate special characters from the complaint description so use it for further analysis.

OUR ANALYSIS

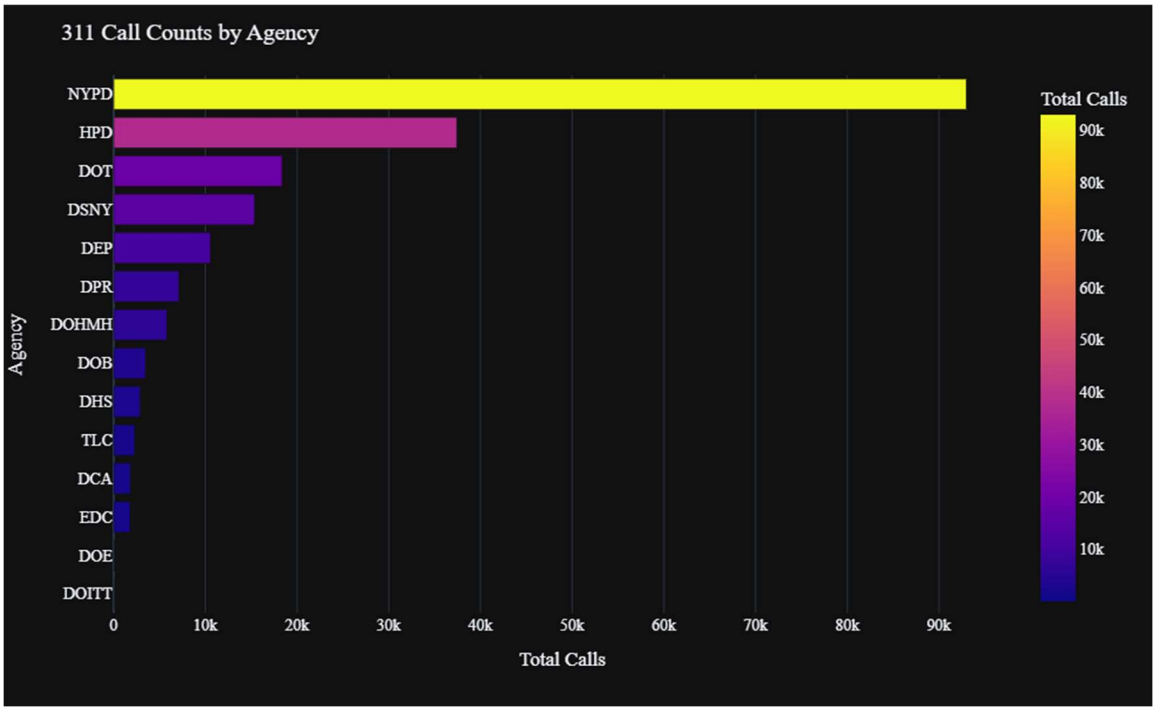
We performed various aggregations to try and gain meaningful insight from the dataset.

We found the most common complaint type for the city. Then to analyze it further we tried to find the most recurring issue in each borough by using a window function to make partitions of the database based on the borough and calculated the most common issue in each partition (borough).

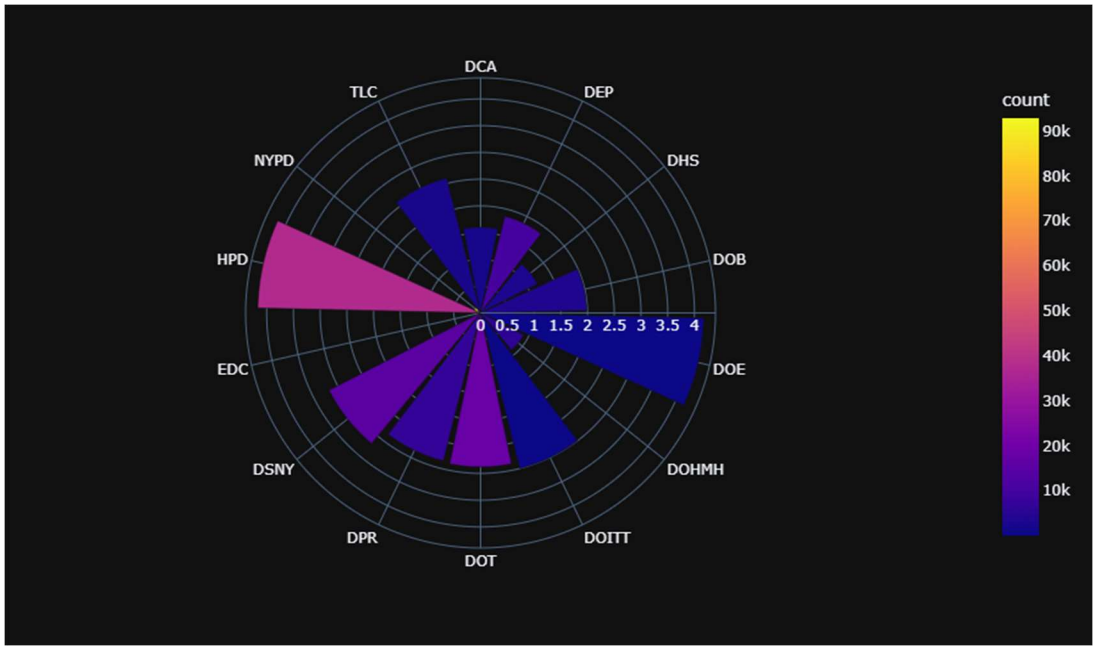


Top 10 most frequent complaint types

We considered which department handles most issues and tried to understand their efficiency by visualizing the average time taken to resolve a complaint. Here, we observed that some departments, like New York Police Department (NYPD), are very quick to resolve complaints, even though they address the highest number of complaints. On the other hand, some departments like the New York City Department of Education (DOE) take a long time to fix issues even though they receive a much lower number of complaints.



Number of 311 calls / service requests per agency

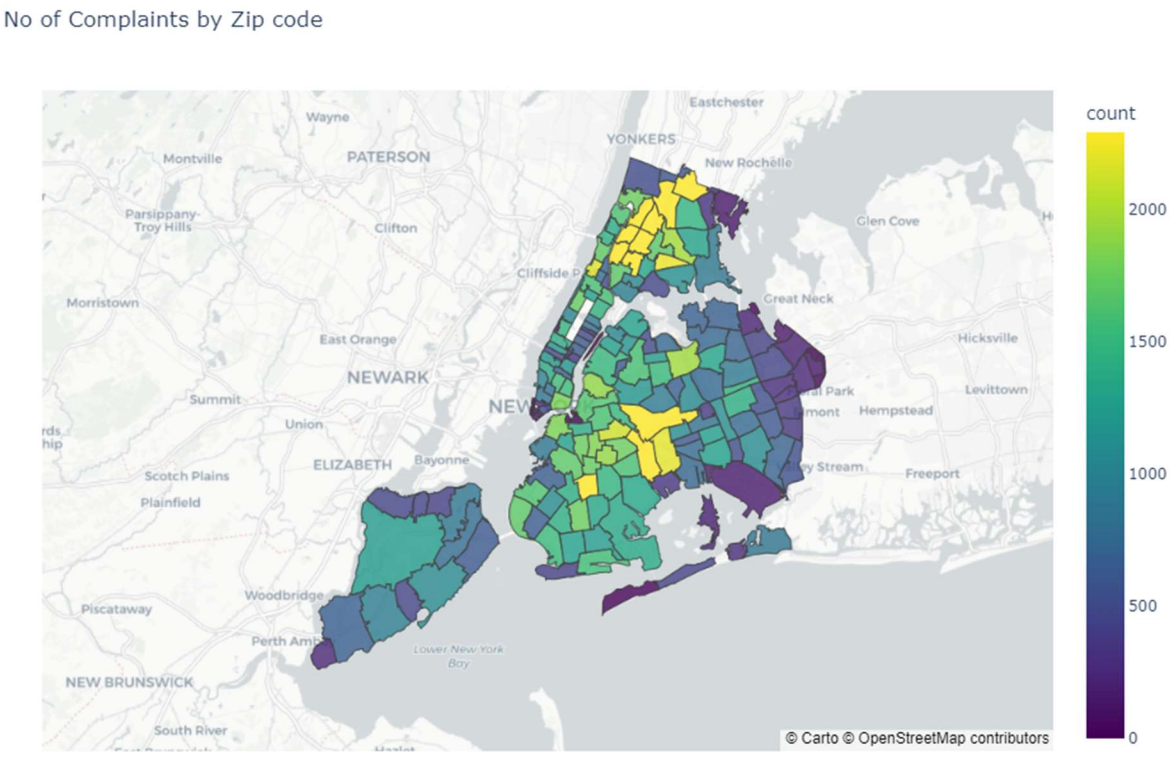


Average time taken by each agency to close complaints

MAPPING THE NOISE

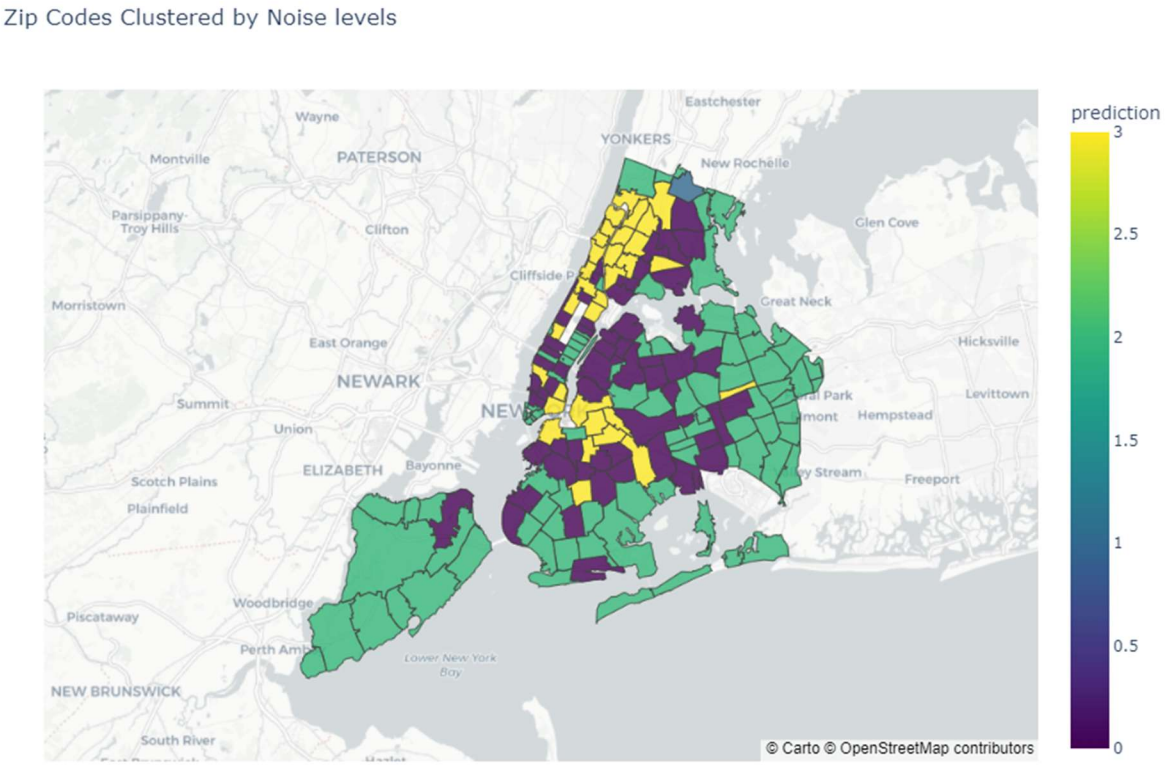
As a demonstration, we wanted to use NYC 311 dataset to help potential renters/buyers decide which area is suitable for them.

So, we made a map showing complaint levels by zip code, to know which neighborhoods have a higher or lower number of complaints.



Number of complaints by NYC ZIP codes

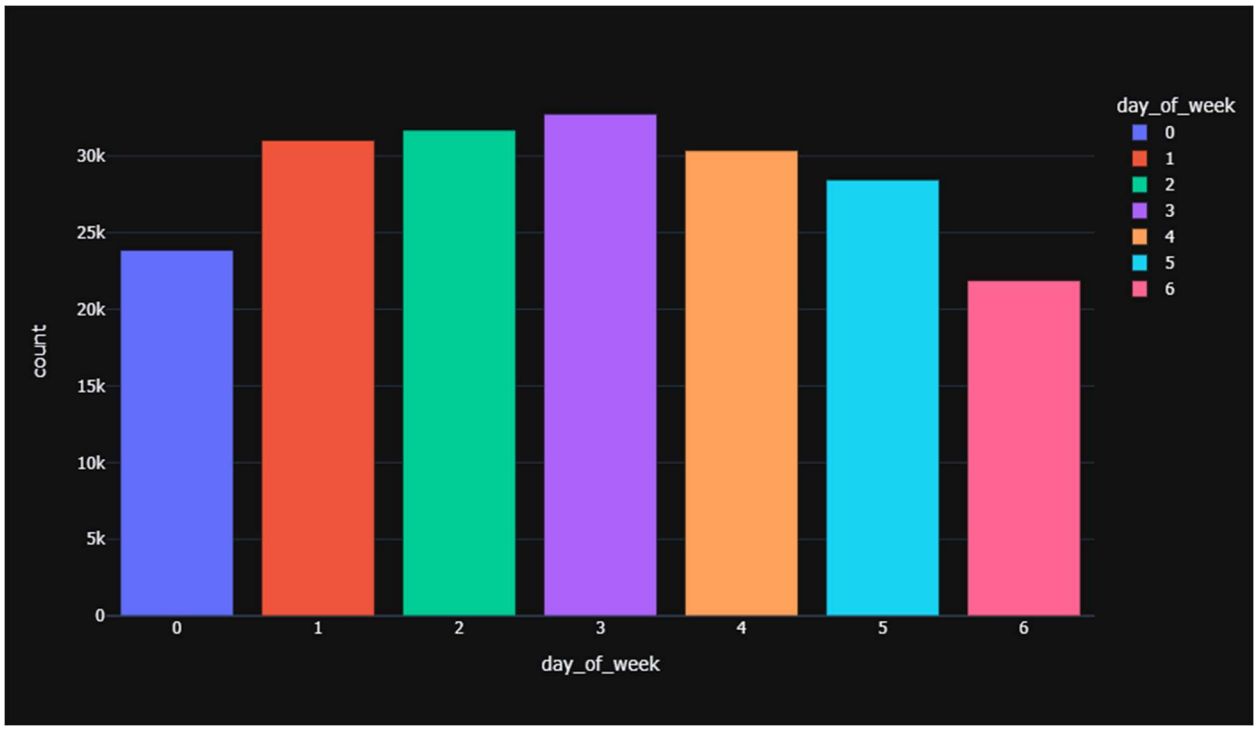
Noise complaints are the most prominent type of complaints in NYC. So, we decided to use the unsupervised Machine Learning model K-Means clustering to cluster neighborhoods with similar noise complaint levels. We formed 4 clusters which classifies each ZIP code into one of the 4 groups – ranging from a one with lower number of noise complaints to a one with a lot.



ZIP Codes clustered by Noise Complaints using K-Means

NOISE VS DAY OF THE WEEK

We have further tried to analyze which days have the most complaints and which have the least. We observed that it forms a bell curve with the middle of the week having the greatest number of complaints while weekends having much less.



Distribution of Noise Complaints by Day of the Week

We made a word cloud to find the most common words using complaint description. We observed that ‘loud’, ‘blocked’, ‘music’ and ‘party’ were the most common words.

borough	complaint_type	count
BRONX	Noise - Residential	7853
BROOKLYN	Illegal Parking	9836
MANHATTAN	Noise - Residential	4434
QUEENS	Illegal Parking	8159
STATEN ISLAND	Street Condition	847
Unspecified	Broken Parking Meter	47

Noise and Parking Complaints were the most common type of complaints

CONCLUSION

By observing these graphs, we can see that New York City departments such as the Department of Housing Preservation and Development, and the Department of Education may need more manpower and budget to reduce their turnaround time for closing complaints.

In places where illegal parking is the most common complaint, it means there is more personal vehicular traffic in the area. The administrations can provide solutions like increasing public transport options to that area, allocating more designated parking spots in the vicinities, and increasing fine for illegal parking.

Just like we have demonstrated for noise complaints, other types of complaints can also be analyzed in the same way. The competent authorities can draw inferences from such analysis and make improvements to the administration and thereby to the lives of the residents of New York City.