CS-GY 6513 BIG DATA
FINAL REPORT

Topic : Analysis of NYC 311 Service Requests

Team Members

Hardi Ramani	**hpr2017@nyu.edu**
Hrituja Khatavkar	**hk3219@nyu.edu**
Sakshi Garg	**sdg8172@nyu.edu**
Upasana Mehta	**um2024@nyu.edu**

Table of Content

Introduction

NYC 311 is a non-emergency phone number and online service that allows residents of New York City to report issues and request city services. 311 service requests cover a wide range of issues, including noise complaints, graffiti, parking violations, and abandoned vehicles.

To analyze 311 service requests, it is important to first obtain data on the requests. The city of New York publishes data on 311 service requests on its Open Data portal, which includes information on the type of request, the location of the request, and the status of the request.

Analyzing 311 service requests can provide valuable insights into the needs and concerns of New York City residents and help the city improve its services. For example, analyzing the data can reveal trends in the types of issues being reported, as well as the efficiency of the city's response to these issues. It can also help identify areas of the city where certain types of issues are more prevalent, and allow city officials to prioritize resources and address these issues more effectively.

The analysis of 311 service requests can be a powerful tool for improving city services and addressing the needs of the community. By examining the data and using the insights gained from the analysis, city officials and policymakers can make informed decisions that benefit both the city and its residents.

Background

NYC 311 is a service provided by the city of New York that allows residents to report non-emergency issues and request city services. The service was launched in 2003 as a way to improve communication between the city and its residents, and to provide a single point of contact for non-emergency issues.

Since its launch, NYC 311 has become a valuable resource for residents and a key source of data for the city. The service receives over 20 million calls and online requests each year, covering a wide range of issues including noise complaints, graffiti, parking violations, and abandoned vehicles.

In addition to providing a single point of contact for non-emergency issues, NYC 311 also serves as a centralized database for service requests. The data collected through 311 service requests is made available to the public through the city's Open Data portal, and can be used to gain insights into the needs and concerns of residents, as well as the efficiency of city services.

Overall, NYC 311 is a valuable resource for both residents and city officials, and the data collected through the service has the potential to inform decision-making and improve city services. Analyzing 311 service requests can provide valuable insights into the needs of the community and help the city address these issues more effectively.

Objective

The objective of analyzing NYC 311 service requests was to to gain insights into the needs and concerns of New York City residents, and to help the city improve its services. Some specific goals of analyzing 311 service requests -

- Identifying trends in the types of issues being reported: Analyzing the data can reveal which types of issues are most common, and whether there are any patterns or trends over time. This information can help city officials prioritize resources and address the most pressing issues in the community.

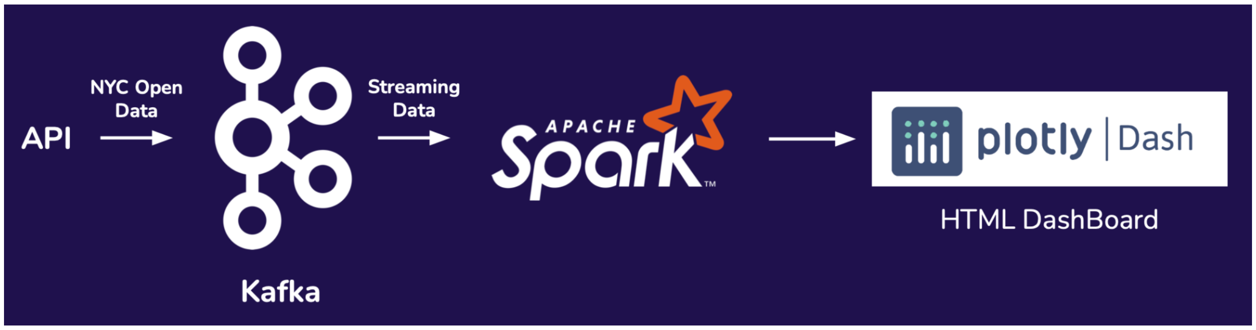
- Evaluating the efficiency of the city's response to 311 service requests: By analyzing the data on the status of requests, it is possible to determine how quickly requests are being resolved, and whether certain types of requests are more likely to be resolved quickly than others. This information can help identify any bottlenecks in the process and suggest ways to improve efficiency.

- Identifying areas of the city where certain types of issues are more prevalent: By mapping the location of 311 service requests, it is possible to identify hotspots where particular types of issues are more common. This information can help city officials target resources and address specific issues in these areas.

- Improving communication with residents: Analyzing 311 service requests also provides insights into how residents are interacting with the city and using city services. This information can help city officials improve communication with residents and ensure that they are aware of the services available to them.

The objective of analyzing NYC 311 service requests is to use the insights gained from the analysis to inform decision-making and improve city services for the benefit of the community.

High Level Design



High-level design of the application:

- **Connect to the API:** An API (Application Programming Interface) is used to retrieve data on 311 service requests from the city's Open Data portal.
- **Set up a Kafka cluster:** The first step is to set up a Kafka cluster, which will be used to stream the data on 311 service requests in real-time.
- **Connect to the Kafka cluster:** Once the Kafka cluster has been set up, PySpark is used to connect to the cluster and consume the data streams.
- **Preprocess and clean the data:** The data is cleaned and preprocessed before it can be analyzed. This includes removing missing or invalid values, formatting dates and times, etc.
- **Analyze the data:** Once the data has been cleaned and preprocessed, it is analyzed using PySpark.
- **Visualize the data:** The analyzed data is visualized using iplot, a library within the Plotly library that allows for the creation of interactive graphs and charts.
- **Display findings:** We create a Dash application to display all the visualizations and findings on a single page.

Dataset

The data used in this project was obtained from NYC Open Data's 311 Service Requests from 2010 to Present. 311 Service Requests encompass all non-emergency requests from the city, including but not limited to noise complaints, air quality issues and reports of unsanitary conditions etc. The 311 calls in New York City (NYC) are publicly available. 311 service request dataset is available at https://data.cityofnewyork.us/Social-Services/311- Service-Requests-from-2010-to-Present/erm2-nwe9 .

This project uses a subset of the data from 2022 that was accessed with the API - https://data.cityofnewyork.us/resource/erm2-nwe9.json which has fetched the data for the months of November and December 2022.

The data is streamed in real time and stored. There are about 1M records stored and the analysis is conducted on this data. The data that is fetched from the API is JSON data. Some of the important fields in the dataset are unique request ID, date, time of request , different location based fields like latitude and longitude, city, borough and the specific address from where the Service Request was made , status of the request whether the request in in progress, closed or open, and a few more details.

The JSON file also includes information about the agency responsible for addressing the issue and the actions taken to resolve it. This data can be used to gain insights into the most common types of issues being reported and the areas of the city where they are most prevalent.

Kafka Streaming

Apache Kafka is a distributed streaming platform that is used to build real-time data pipelines and streaming applications. It is designed to handle high volumes of data with low latency and provide a seamless experience for processing data streams.

Kafka works by using a publish-subscribe model, where producers write data to Kafka topics and consumers read from these topics. A topic is a category or feed to which messages are published. Producers write data to topics and consumers read from these topics.

Kafka stores all published messages for a configurable amount of time, allowing consumers to read messages that were published while they were offline. This makes Kafka a highly durable and scalable platform for handling data streams.

Kafka is composed of a number of components, including:

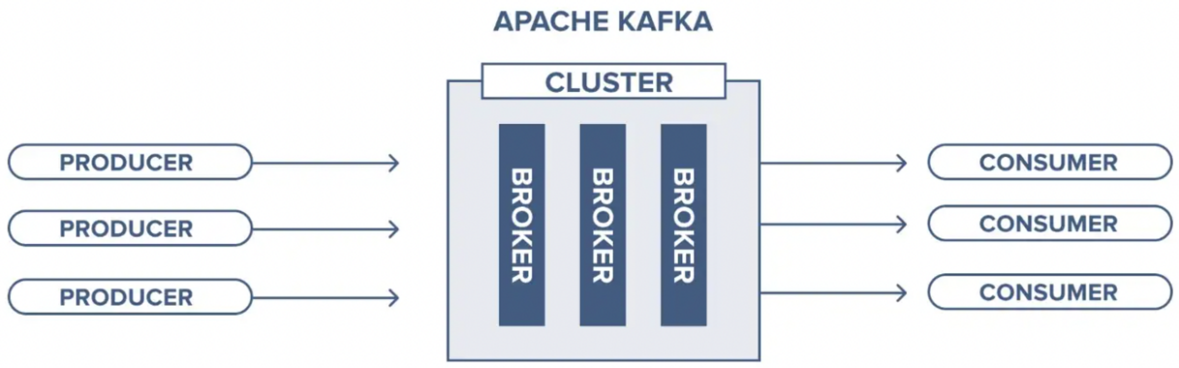
- **Producers:** Producers are applications that generate data and write it to Kafka topics. Producers can write data to multiple topics and have a range of options for specifying the partition to which a record is sent.

- **Topics:** Topics are the categories or feeds to which producers write data. Topics are divided into a number of partitions, which allows for parallel processing of the data by consumers.

- **Brokers:** Brokers are the servers that make up the Kafka cluster. They are responsible for storing published messages and serving them to consumers.

- **Consumers:** Consumers are applications that read data from Kafka topics. Consumers can read data from multiple topics and have a range of options for specifying the partitions to read from and the starting offset (position) in the partition.

- **Zookeeper:** Zookeeper is a distributed coordination service that is used by Kafka to store metadata about the Kafka cluster and consumer offsets (positions).



Source:
https://www.cloudkarafka.com/blog/part1-kafka-for-beginners-what-is-apache-kafka.html

Kafka is used in a variety of applications, including real-time data processing, real-time analytics, and event-driven architectures. It is highly reliable and scalable, making it an attractive choice for organizations looking to build data pipelines and streaming applications.

Why have we used Kafka?

Since the data is being streamed in real time and is being stored , over time an enormous amount of data will be collected. We face two main challenges when dealing with this data:
Since the data is being streamed in real time the data needs to be brought to the datastore quickly.

The data is in huge volumes, to tackle this problem we can use Kafka which as Kafka has a simple API and supports a wide range of programming languages, making it easy to integrate into existing applications.This helps to tackle the problem of mobility of the data. Kafka allows for real-time processing of data streams, making it an ideal choice for applications that require immediate action on incoming data. Kafka is designed to handle high volumes of data with low latency, making it a high-performance platform for handling data streams.

All in all, Kafka acts as a good messaging system to handle huge volumes moving of data.

Analysis

Once the setup is done, then comes the analysis part. Since we have continuous streaming of data from an API, it becomes necessary to find some useful output or insights of this enormous data.

We started by providing answers to the more general questions before moving on to the more specific ones. We examined the many 311 complaints received to determine which was the biggest issue that the public was dealing with. The majority of New Yorkers struggle with the heat/hot water issue. Next, we examined the distribution of complaints among New York's various boroughs. We saw from the pie chart that Brooklyn Borough received the majority of the complaints. Next, we considered which Brooklyn Borough issue was of utmost importance. That is unlawful parking. The following step was to determine which month or season of the year receives the most complaints. The month is December. After that, we responded to the query regarding the agency that received the most 311 requests. The NYPD was there. The New York Police Department has also successfully completed the most service requests.

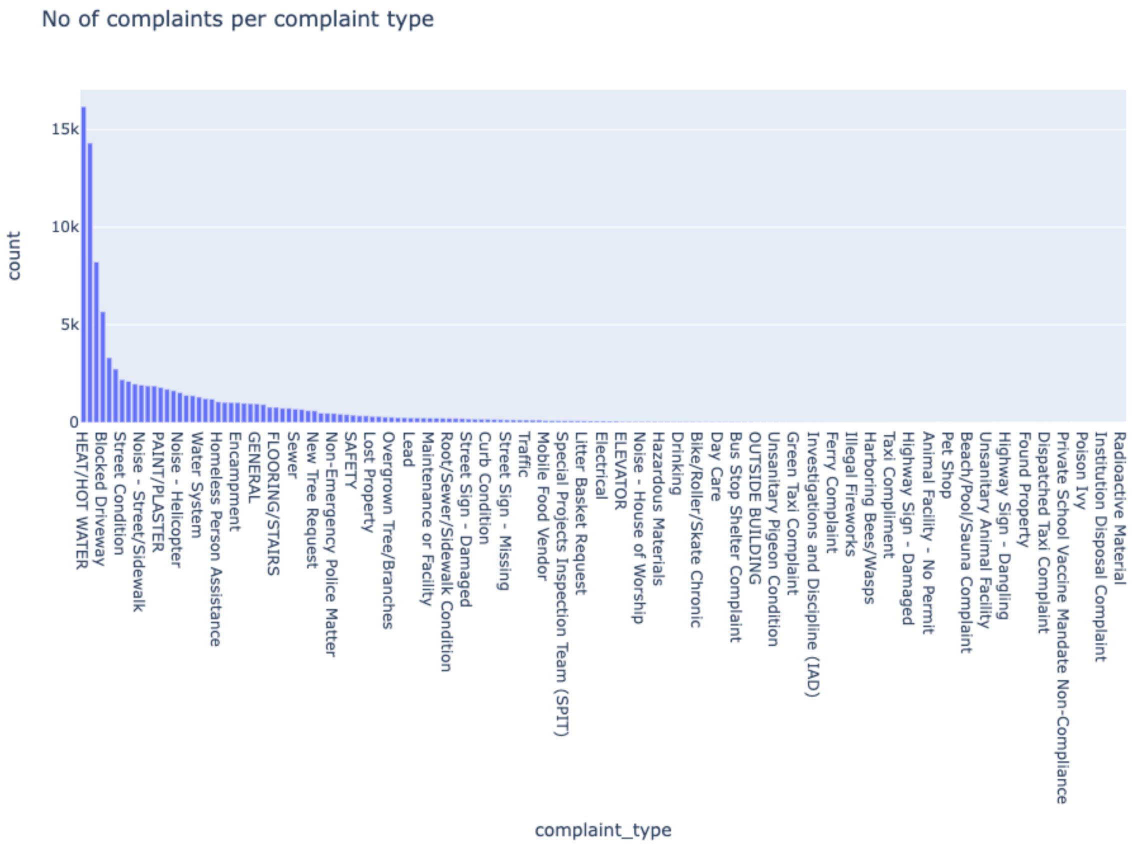
We did analysis on how many complaints are solved, in progress and are yet to be solved and saw that the maximum number of complaints are closed. We can say that all NYC agencies are working efficiently round the clock. Next is how many various types of complaints are filed at each location. Here location means street/sidewalk, residential building, park, restaurant, etc. Majorly, issues with rodents are the most filed complaint at each location. Previously, we did an analysis on which borough has the most number of complaints. But now, we will perform an analysis on which city has the most number of complaints and found that again Brooklyn city has the most number of complaints files. Moving further, we found that Illegal parking is the most common issue in Brooklyn city. Next is based on the intersection of longitudinal and latitudinal, which combination has the most filed complaints and the outcome is. The visual for this analysis is more interesting than its analysis. At last, which channel type is the most common source to file complaints. People prefer online sources to report the complaints.

Visualization

We created the visualizations in Python and viewed them as HTML Dashboards using Plotly Dash. We started with answering broader questions and further went on narrowing our approach. We created the following visualizations to draw insights from the NYC 311 service requests data.

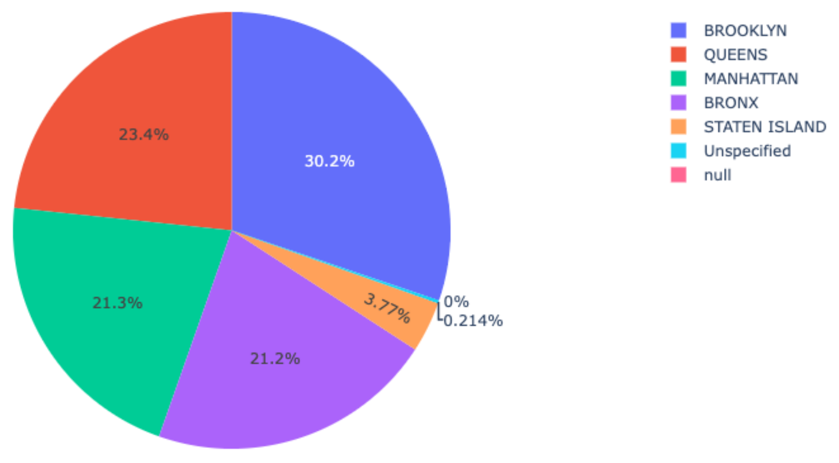
1. What are different types of Service Requests(SRs)/Complaint Type? Which is the most frequent?

- ❖ We created a bar graph and have put different types of Service requests on x-axis and the count on y-axis.



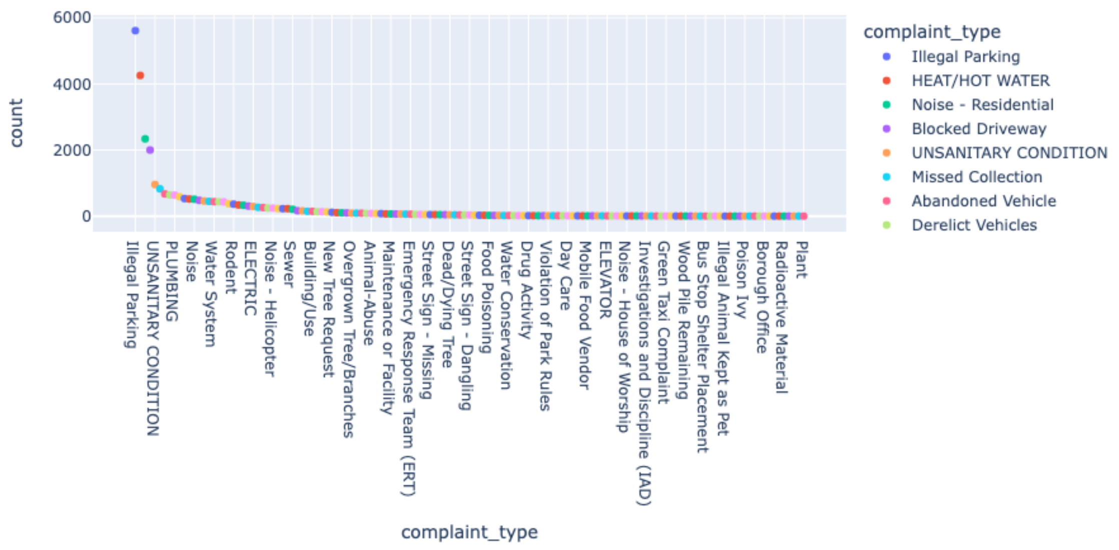
2. What is the percentage of complaints in each borough? From which borough most SRs come from?
    - ❖ We created a pie chart with each cut representing one borough. We count the number of complaints from each borough.

Complaints distribution across Boroughs



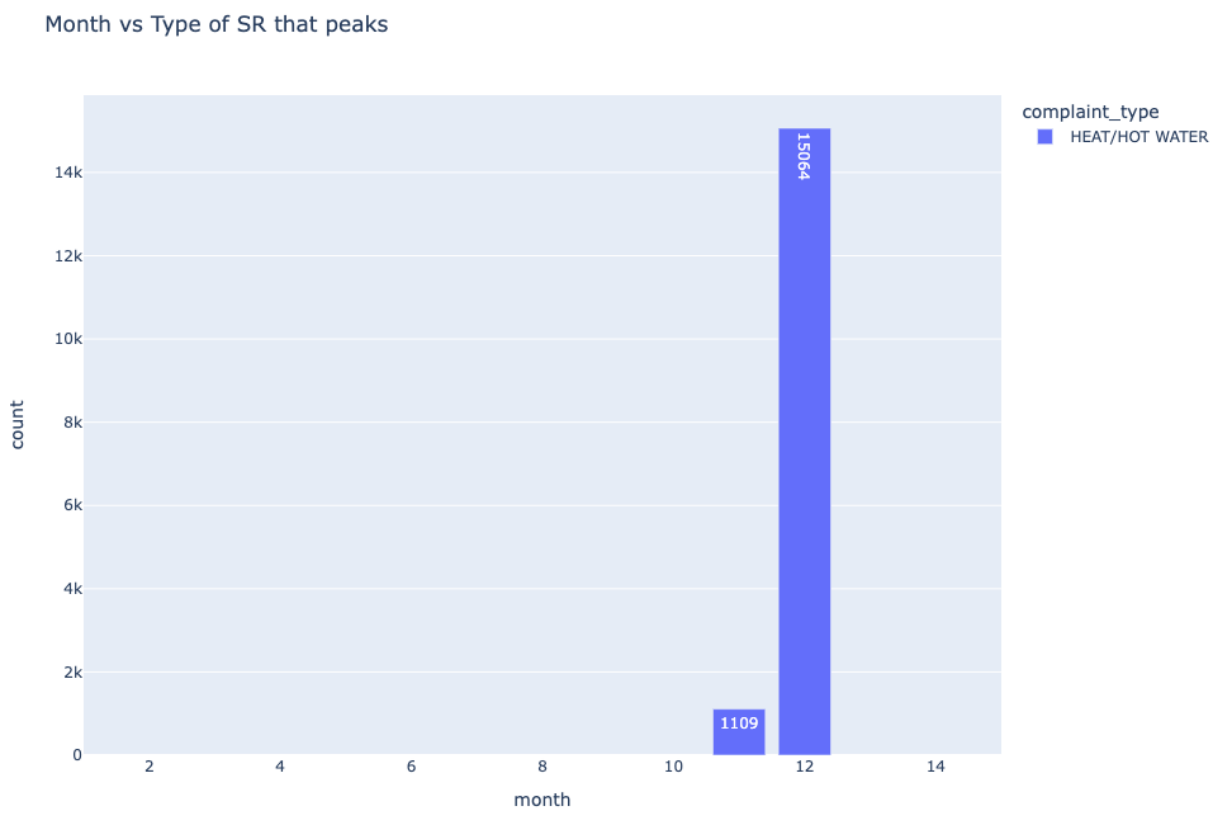
3. Which SR is most common in the borough with max SR?
    - ❖ After getting the result is Brooklyn from the previous question. We further deep dived to understand that which is the major issue faced by people in Brooklyn borough. We created a scatter plot to understand this trend.

complaints of each type in the Borough with maximum Service Requests

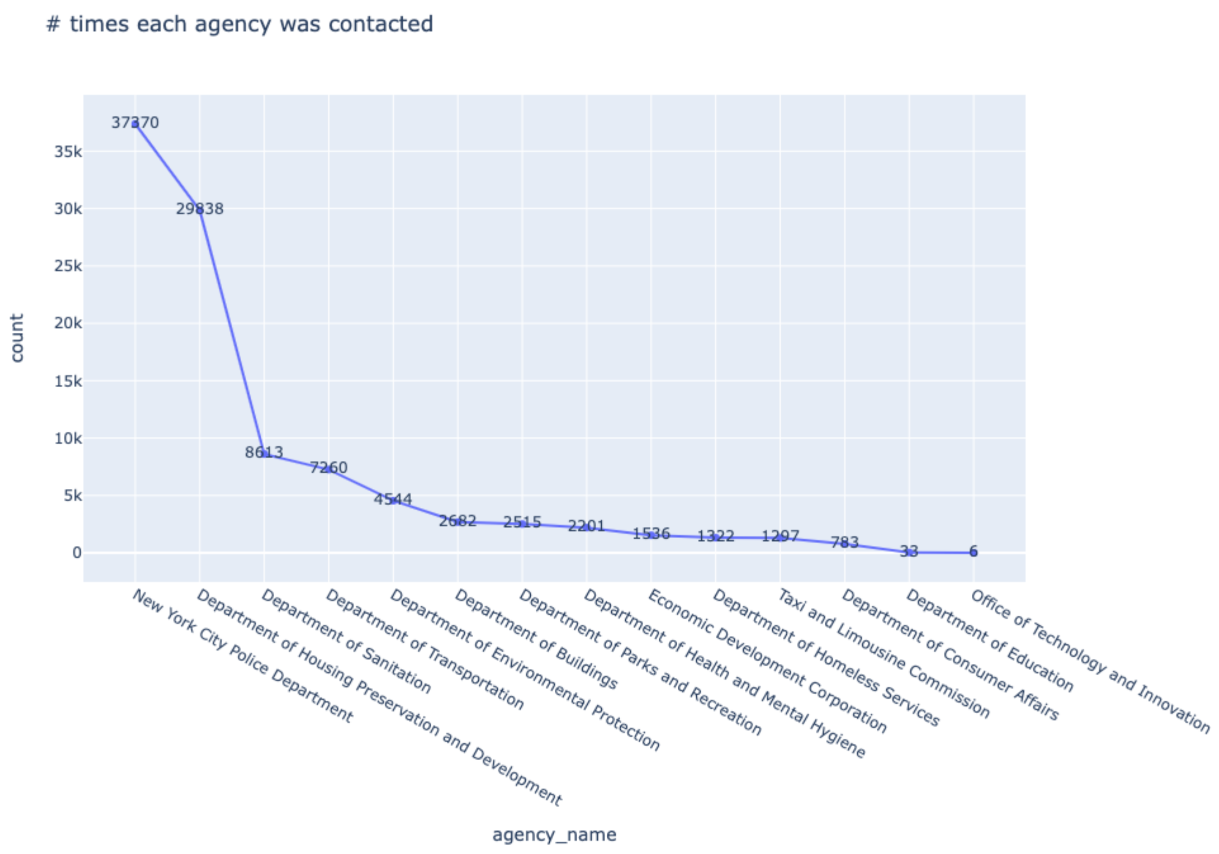


4. How many complaints were made in each month and what was the count of each complaint? Which SRs peaks at what time of year?

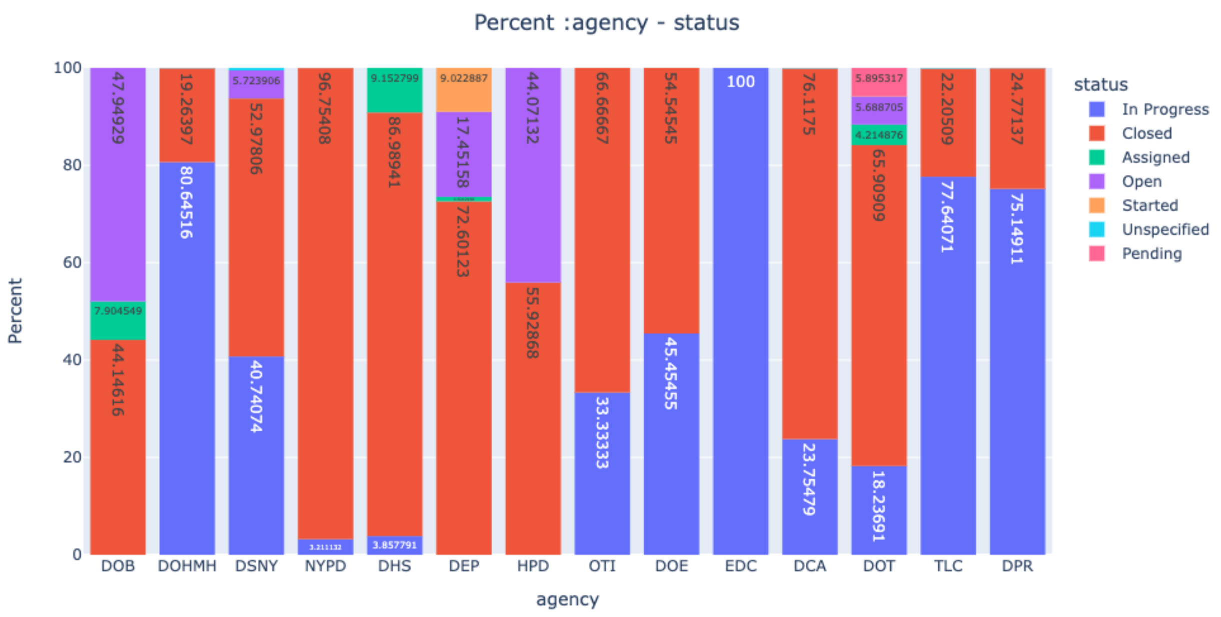
- We further went on to analyze during which time of the year most complaints were made. We created a bar graph with months on x-axis and counts on y-axis. We also put a limit on x-axis to hold only 12 values.



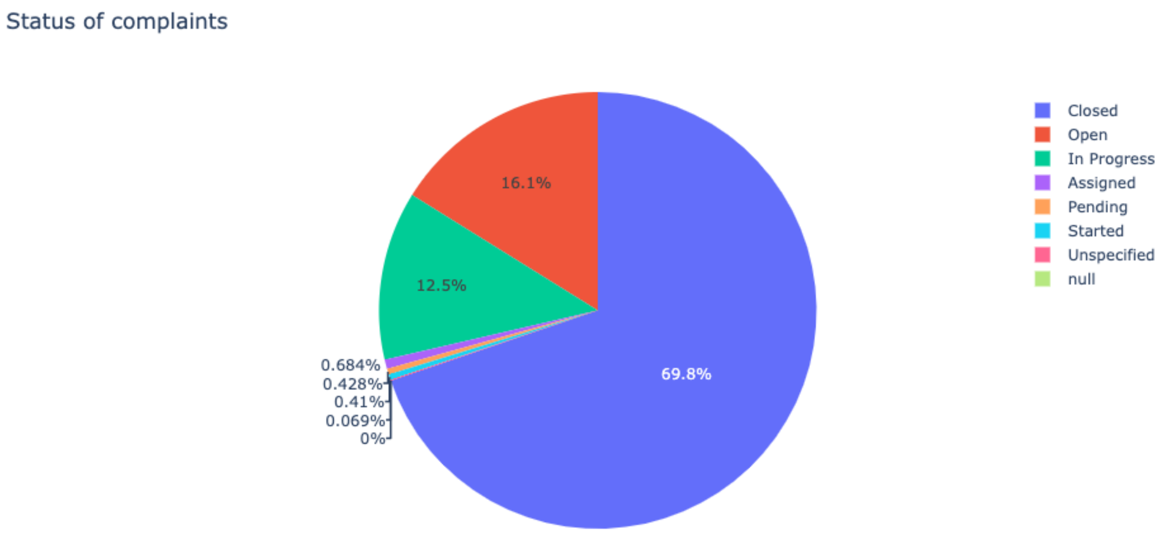
5. Which agency is contacted the most number of times?
    - There are a lot of different agencies who handle these 311 requests. To understand if they are able to serve the people on time. And if there was an agency in particular which needed more allocation of resources or were lacking in their service. We created a line chart to understand this trend.



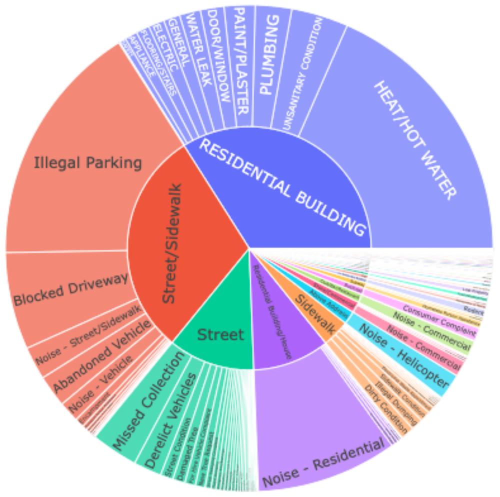
6. How many of the SR have been closed by all the agencies and their percentage?
    - ❖ Every Service Request made has a status like Closed/Open/In Process. We wanted to understand which agency has the most number of Closed and Open requests to look at the data more closely. We created a stacked bar chart for this purpose.



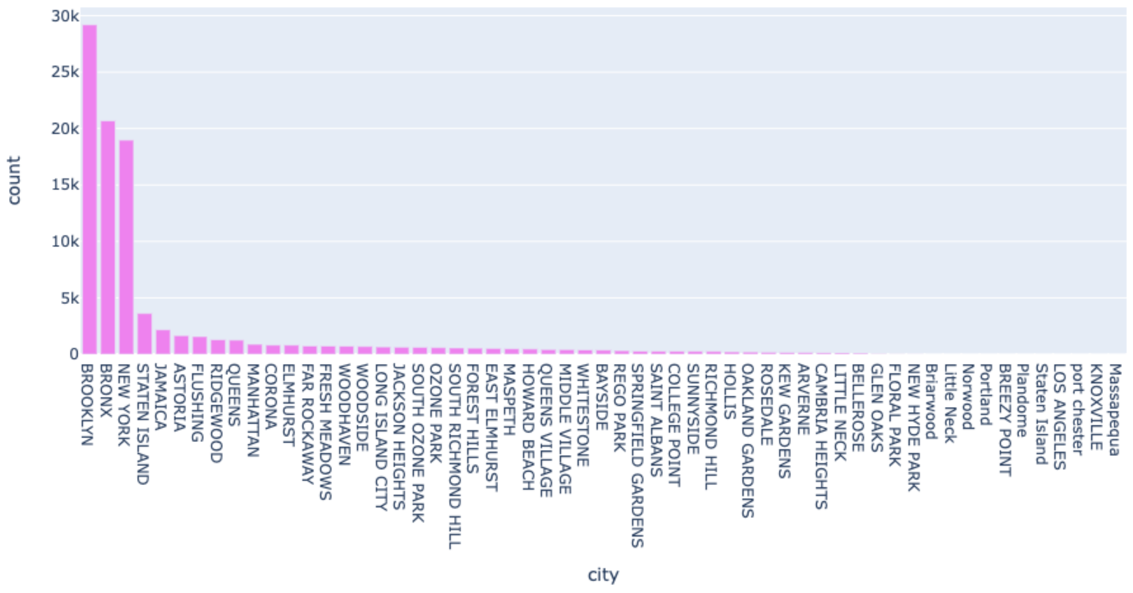
7. How many complaints are resolved, in progress and yet to be resolved?
   - ❖ We created a pie chart to show the status of complaints.



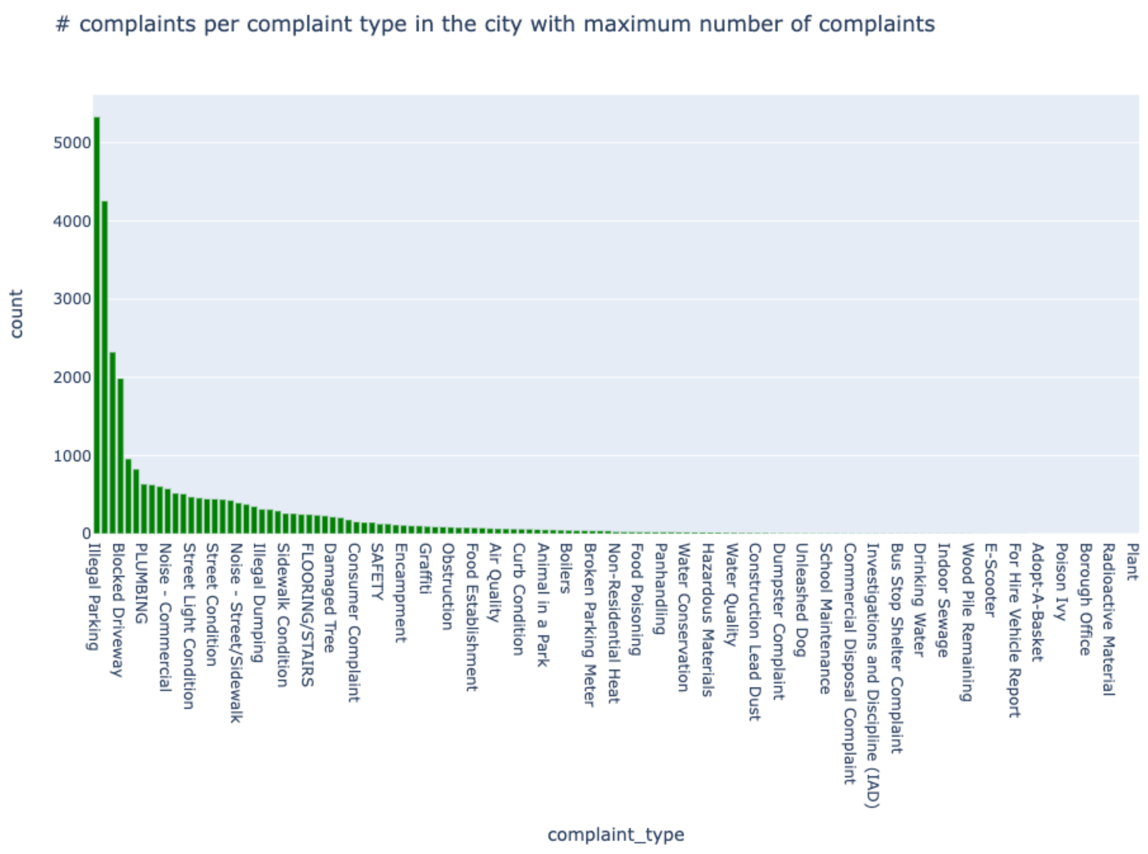
8. What are the different types of complaints with its count at each location in NY?
   - ❖ We created a pie chart within a pie chart. The inner pie chart has different locations while the outer pie chart is the type of complaints.



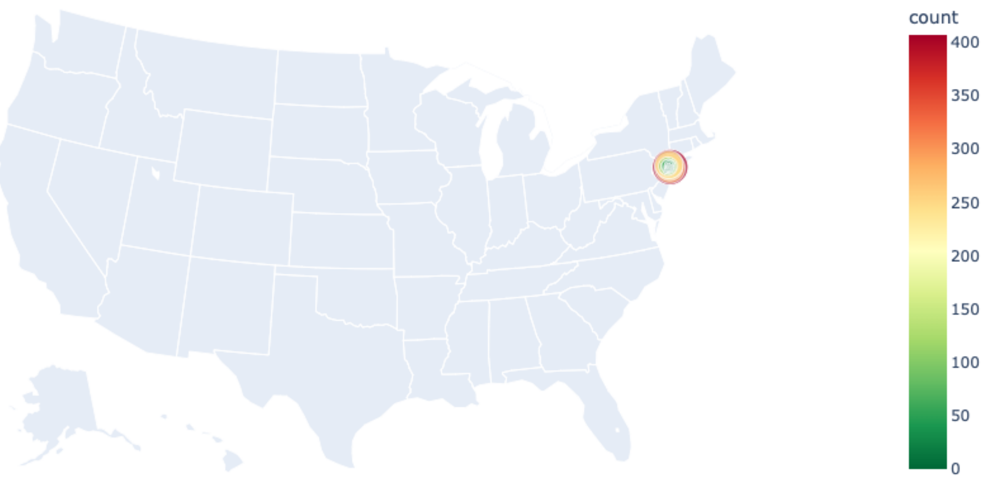
9. Which city has the most number of service requests?
    - ❖ We created a bar graph and put the count on the x-axis and the city on the y-axis.



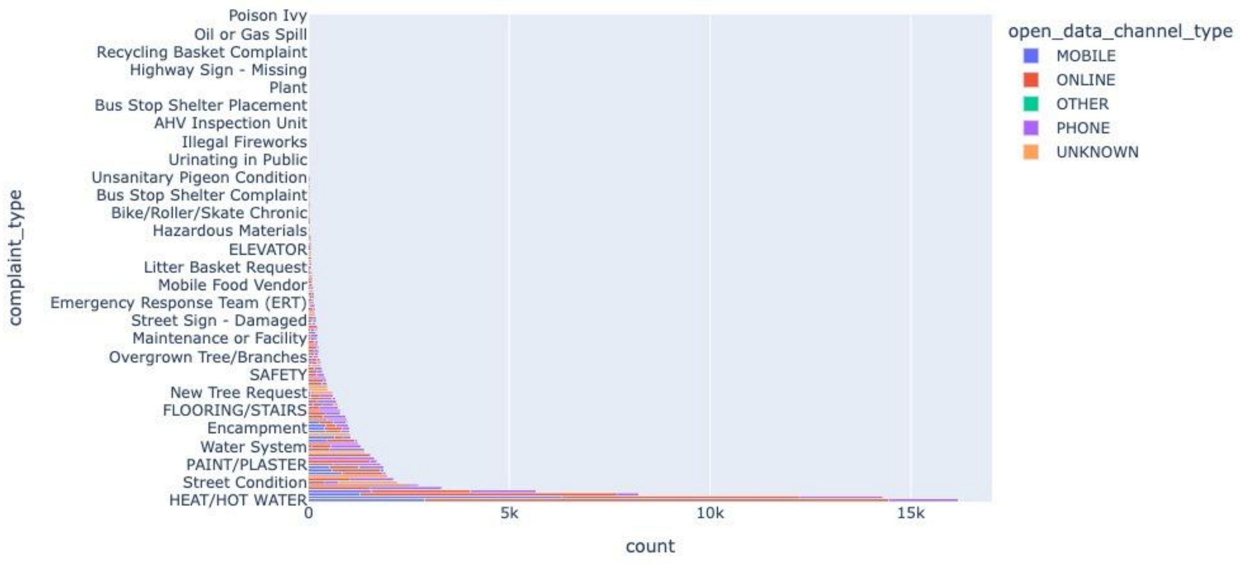
10. Which complaint is most common in the address type with the maximum number of complaints?
    - ❖ We here created a bar graph with count of complaints on the x-axis and types of complaints on the y-axis.



11. Based on longitudinal and latitudinal intersections, where are the most complaints filed from?
    - ❖ We created a bubble chart for the intersection of latitude and longitude points.



12. From which channel type most of the complaints were made and its distribution?
    - We created a bar graph with count on the x-axis and complaint type on the y-axis with different colors denoting various channel types.



Insights

There are a number of questions that have been answered. There were many interesting facts that came into picture. This can be used by the officials of NYC to better understand the issues faced by NYC residents. Now, considering the dataset and the technologies used, we've obtained the following insights:

Since we have used an API, Kafka helps in streaming the data which later is used in Spark for analysis. The analysis conducted infers that most complaints are coming from the area of Brooklyn and Heat/Hot Water Related complaints are the most received. This analysis might help in understanding the problems faced by the people of New York, their urgency to be addressed and their current status right now.

Future Work

Right now, we have to use shell to initialize the deployment of our application. But moving forward, we are planning to use docker to automate the flow and deploy the application without any human interference. Moreover, we are storing our data on a local machine but since the data is enormous, it needs to be stored on some cloud platform so that it can be accessed anytime anywhere and for that Cassandra comes into the frame. Finally, when everything is done, we need some visualization tool that takes real time analytics of our data to display it on the dashboard and for that we will be using databricks.

References

1. https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9
2. https://medium.com/geekculture/integrate-kafka-with-pyspark-f77a49491087
3. https://medium.com/inspiredbrilliance/kafka-basics-and-core-concepts-5fd7a68c3193
4. https://kafka.apache.org/documentation/
5. https://spark.apache.org/docs/latest/api/python/index.html
6. https://docs.docker.com/
7. https://cassandra.apache.org/doc/latest/
8. https://www.databricks.com/learn
9. https://www.cloudkarafka.com/blog/part1-kafka-for-beginners-what-is-apache-kafka.html