VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY
HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

# DATA ENGINEERING (055240)

# Getting insights from Service Request

| | | |
|---|---|---|
| Lecturer: | Phan Trọng Nhân | |
| Group: | 2 | |
| Student: | Lý Gia Huy | 2010289 |
| | Lục Gia Huy | 2011268 |

Ho Chi Minh City, 04/2023

# Contents

# 1 Introduction

The NYC311 service is a vital part of the public service system in New York City, providing citizens with a platform to report non-emergency issues and request services from various government agencies.

NYC 311's mission is to provide the public with quick and easy access to all New York City government services and information while offering the best customer service. Each day, NYC311 receives thousands of requests related to several hundred types of non-emergency services, including noise complaints, plumbing issues, and illegally parked cars. These requests are received by NYC311 and forwarded to the relevant agencies such as the police, buildings, or transportation. The agency responds to the request, addresses it, and then closes it.

The city government of New York stores all service requests and creates a dataset for each year. From this dataset, we can extract valuable insights by analyzing different aspects.

# 2 Problem Statement

The objective of this assignment is to gain insights into the different types of SRs, their frequency, and any patterns or trends in the data. Specifically, we aim to answer questions such as:

- **Which are the most/least common types of SRs received by NYC311?**

  We will examine the data to determine which types of SRs are the most and least frequent. This information can be useful for the relevant agencies to allocate their resources and respond effectively to the most common types of requests. Additionally, identifying the least frequent types of requests can help the agencies to investigate any potential reasons for the low volume of reports and take action to address any issues

- **From which borough most SRs come from?**

  We will analyze the data to determine which boroughs in New York City generate the most SRs. This information can be useful for the relevant agencies to allocate resources and prioritize their response efforts.

- **Which SRs peaks at what time of year or time of day?**

  We will examine the data to determine if there are any patterns or trends in the timing of SRs. For example, do certain types of SRs peak during certain times of the year or day? This information can be useful for the relevant agencies to plan their staffing and response efforts accordingly.

- **What are the most efficient agencies responsible for addressing these requests?**

  We will analyze the data to identify the agencies that are more efficient in resolving SRs. This information can be useful for the relevant agencies to improve their internal processes and allocate their resources more effectively

- **From which type of location we get most number of complaints?**

  We will analyze the data to determine which types of locations generate the most SRs. For example, do most SRs come from residential areas, commercial areas, or public spaces? This information can be useful for the relevant agencies to identify areas that require more attention and resources.

Next thing we want to find out or predict is:

- **Predict time required in terms of range of days to resolve a specific complaint in a specific borough.**

  We will analyze the data to predict the time required to resolve specific complaints in different boroughs. By examining the data of past requests, we can identify patterns and trends that can help us predict the time required to resolve future complaints. This information can be useful for the relevant agencies to allocate their resources and plan their response efforts accordingly.

- **A time series analysis to forecast the volume of calls to be expected on given future date.**

We will conduct a time series analysis of the data to forecast the volume of calls that can be expected on future dates. By analyzing past trends in the volume of calls, seasonal variations, and other factors, we can make accurate predictions of the expected volume of calls on future dates. This information can be useful for the relevant agencies to plan for future demand and allocate their resources and staffing accordingly.

The insights gained from this analysis can help improve the efficiency and effectiveness of the service and provide better customer service to the public.

# 3 Motivation

It is very important to understand and analyze data from the NYC311 service. Specifically, this may stem from the following reasons

- Improving service quality: government agencies better understand citizen issues and requests. When issues are resolved quickly and effectively, citizens are more satisfied with public services.
- Optimizing resources: government agencies better understand resource allocation and unit operations. When resources are distributed more effectively, government agencies can save costs and optimize their operations.
- Forecasting demand: government agencies forecast future demand and prepare for potential issues. When government agencies are able to forecast and prepare for future demand, they can minimize negative impacts on the community and ensure that public services are provided efficiently and meet the needs of citizens.
- Researching user behavior: government agencies better understand citizen behavior and needs when using public services. When government agencies understand citizen behavior and needs, they can improve public services to better meet those needs.

For example, we can identify patterns and trends in the types of service requests received, the frequency of requests in different neighborhoods, and the response times for different types of requests. We can also use this dataset to forecast future demand for certain types of services and identify areas where additional resources may be needed. Additionally, we can analyze the data to identify areas where service quality can be improved, such as reducing response times or increasing the resolution rates of certain types of requests.

In summary, analyzing NYC311 data can help government agencies improve service quality, optimize resources, forecast future demand, and research user behavior. This benefits the community and enhances citizen trust in public services.

# 4 Challenges in terms of Data Engineering

There are several challenges in terms of data engineering that may arise when working with the NYC311 service request dataset. Some of these challenges include:

- **Data Cleaning**: The NYC311 dataset may contain errors, missing values, or inconsistencies that need to be addressed before conducting analysis. This requires thorough data cleaning and preprocessing to ensure accurate insights.
- **Data Integration**: The NYC311 dataset may need to be integrated with other datasets, such as weather data or demographic data, to obtain a comprehensive understanding of the factors affecting service request trends. This requires expertise in data integration and management.
- **Data Volume**: The NYC311 dataset is large and may require significant computing power to analyze. This requires expertise in data processing and management to efficiently handle large datasets.
- **Data Visualization**: To effectively communicate insights, it is important to create clear and visually appealing data visualizations. This requires expertise in data visualization and communication.

# 5 Data Preparation

The NYC311 service request dataset was obtained from the official NYC Open Data website. The dataset contains information on service requests made to the city, including the type of request, the date and time of submission, the location of the request, and the outcome of the request.

Before conducting any analysis, the dataset underwent several data preparation steps to ensure that it was clean and ready for analysis.

## 5.1 Data Cleaning

The dataset contained missing values, inconsistencies, and errors such as NaN and "Unspecified" that needed to be addressed .

Based on the problem statement, we will extract data by selecting only the columns that we consider valuable for evaluating the NYC311 service request dataset. Typically, we follow the principles:

- Columns with a high percentage of missing values will not have much insight value, so we will remove those columns. Specifically, we will remove the following columns: 'School Name', 'School Number', 'School Region', 'School Code', 'School Phone Number', 'School Address', 'School City', 'School State', 'School Zip', 'School Not Found', and 'School or Citywide Complaint'.

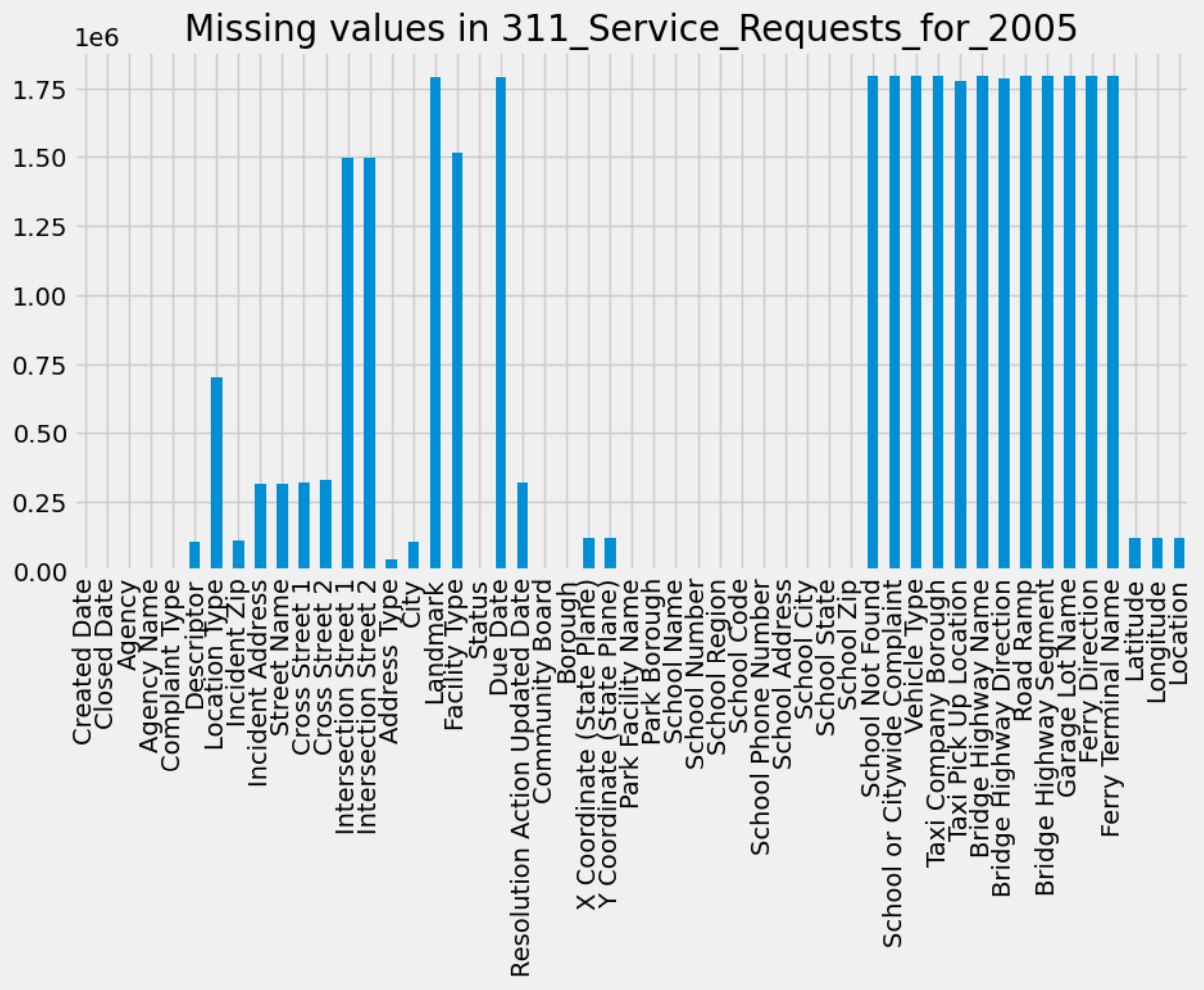


**Figure 1:** *Missing values of dataset in 2005*

- We will select only the service requests that have a "Closed" status and non-missing "Created Time" and "Closed Time" values.

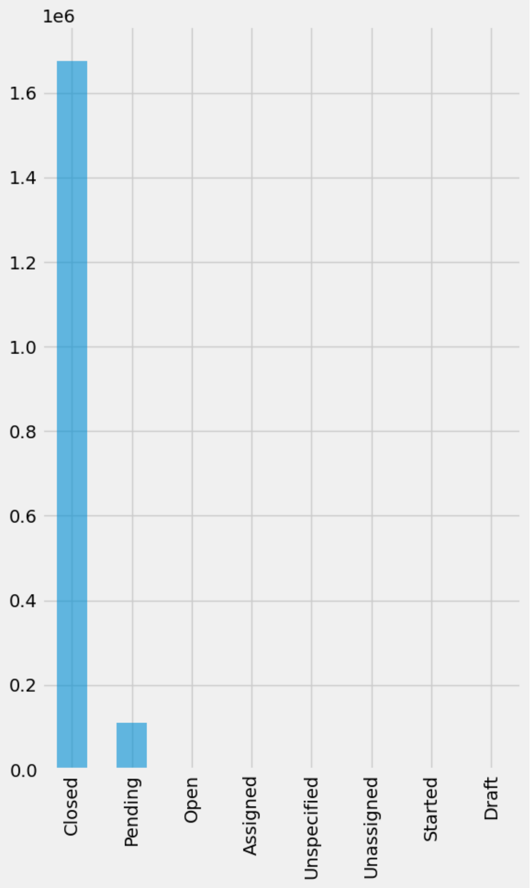


**Figure 2:** *Status values of dataset in 2006*

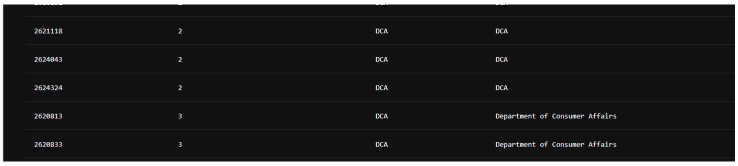
## 5.2 Data Transformation

The dataset was transformed into a format suitable for analysis. This included converting the data into a structured format, such as a table, and aggregating the data to the appropriate level of granularity.

First, we need to create additional columns to support the analysis and prediction process later on. These columns include:

- Resolution Time: The time taken to resolve a service request, calculated as the difference between the Closed Date and Created Date.
- Day of Week: A column that extracts the day of the week from the Created Date column.
- Date: A column that extracts the date from the Created Date column.
- Month: A column that extracts the month from the Created Date column.
- Year: A column that extracts the year from the Created Date column.

Also, an issue that may arise is the inconsistency between the values of the "agency" and "agency name" columns. For example, there may be different agency names that have the same agency value, which can make future analysis difficult. Therefore, we will synchronize the data in these two columns by transforming the value of the "agency name" column as follows: selecting an agency name that has a broad enough value to sufficiently describe the agency.

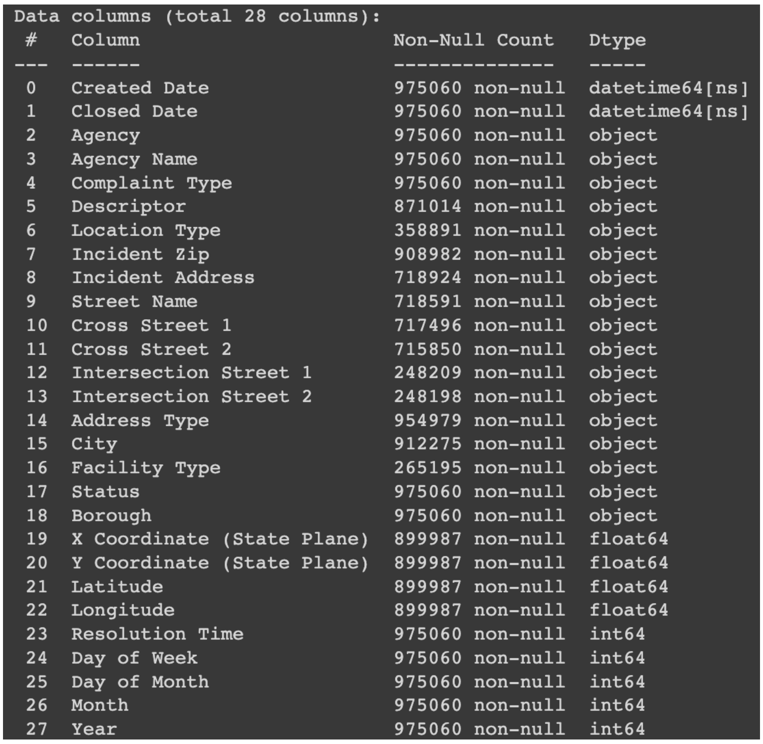


| | | | |
|---|---|---|---|
| 2621118 | 2 | DCA | DCA |
| 2624043 | 2 | DCA | DCA |
| 2624324 | 2 | DCA | DCA |
| 2620813 | 3 | DCA | Department of Consumer Affairs |
| 2620833 | 3 | DCA | Department of Consumer Affairs |

**Figure 3:** *Example for inconsitence data*

In this case, we will transform value "DCA" to "Department of Consumer Affairs".

Then, we will load the entire cleaned dataset into a single table in the staging environment that contains all the fields selected after cleaning.

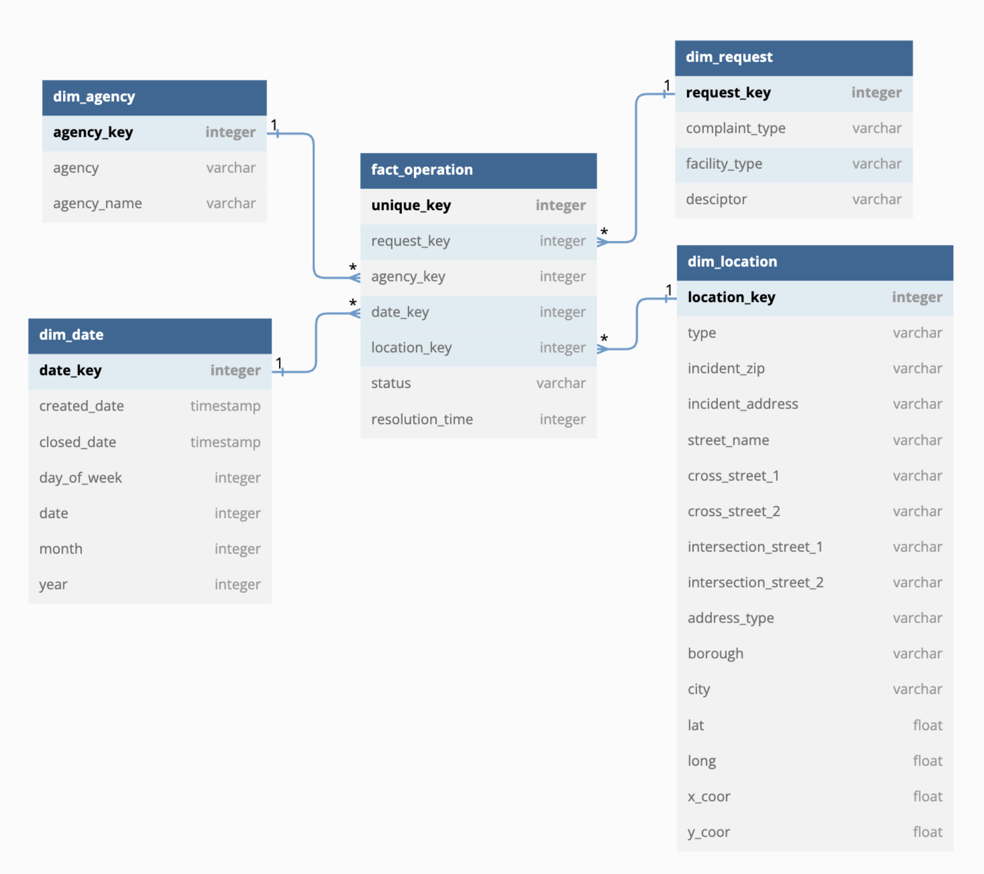


```
Data columns (total 28 columns):
 #   Column                       Non-Null Count   Dtype
---  ------                       --------------   -----
 0   Created Date                 975060 non-null  datetime64[ns]
 1   Closed Date                  975060 non-null  datetime64[ns]
 2   Agency                       975060 non-null  object
 3   Agency Name                  975060 non-null  object
 4   Complaint Type               975060 non-null  object
 5   Descriptor                   871014 non-null  object
 6   Location Type                358891 non-null  object
 7   Incident Zip                 908982 non-null  object
 8   Incident Address             718924 non-null  object
 9   Street Name                  718591 non-null  object
 10  Cross Street 1               717496 non-null  object
 11  Cross Street 2               715850 non-null  object
 12  Intersection Street 1        248209 non-null  object
 13  Intersection Street 2        248198 non-null  object
 14  Address Type                 954979 non-null  object
 15  City                         912275 non-null  object
 16  Facility Type                265195 non-null  object
 17  Status                       975060 non-null  object
 18  Borough                      975060 non-null  object
 19  X Coordinate (State Plane)   899987 non-null  float64
 20  Y Coordinate (State Plane)   899987 non-null  float64
 21  Latitude                     899987 non-null  float64
 22  Longitude                    899987 non-null  float64
 23  Resolution Time              975060 non-null  int64
 24  Day of Week                  975060 non-null  int64
 25  Day of Month                 975060 non-null  int64
 26  Month                        975060 non-null  int64
 27  Year                         975060 non-null  int64
```

**Figure 4:** *All column selected after cleaning*

By loading the data into a single table, we can easily access all the relevant fields and perform any necessary additional cleaning or transformations before loading the data into the fact and dimension tables of the **star schema**. This approach also allows us to maintain the integrity of the original data and ensures that we do not lose any important information during the transformation process

Once the data is loaded into the staging environment, we can then begin the process of transforming the data into a star schema format for use in the database. This will involve identifying the appropriate fact and dimension tables and mapping the relevant fields from the staging table to these tables. By using a **star schema**, we can ensure that the data is organized in a way that facilitates efficient querying and analysis.



| dim_agency | |
|---|---|
| **agency_key** | **integer** |
| agency | varchar |
| agency_name | varchar |

| dim_date | |
|---|---|
| **date_key** | **integer** |
| created_date | timestamp |
| closed_date | timestamp |
| day_of_week | integer |
| date | integer |
| month | integer |
| year | integer |

| fact_operation | |
|---|---|
| **unique_key** | **integer** |
| request_key | integer |
| agency_key | integer |
| date_key | integer |
| location_key | integer |
| status | varchar |
| resolution_time | integer |

| dim_request | |
|---|---|
| **request_key** | **integer** |
| complaint_type | varchar |
| facility_type | varchar |
| desciptor | varchar |

| dim_location | |
|---|---|
| **location_key** | **integer** |
| type | varchar |
| incident_zip | varchar |
| incident_address | varchar |
| street_name | varchar |
| cross_street_1 | varchar |
| cross_street_2 | varchar |
| intersection_street_1 | varchar |
| intersection_street_2 | varchar |
| address_type | varchar |
| borough | varchar |
| city | varchar |
| lat | float |
| long | float |
| x_coor | float |
| y_coor | float |

**Figure 5:** *Schema of database*

# 6 Getting Insights

## 6.1 The most/least common types of SRs

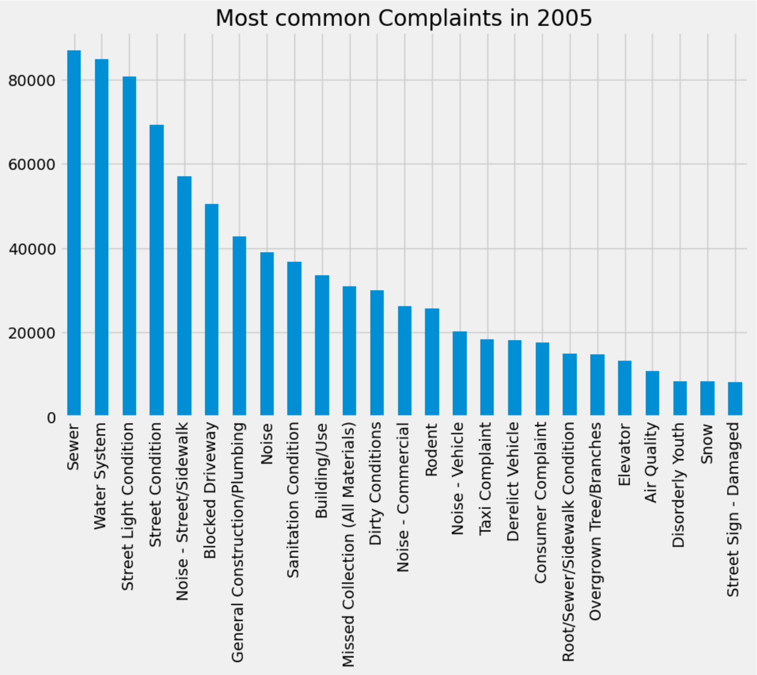


**Figure 6:** *The most common complaints in 2005*

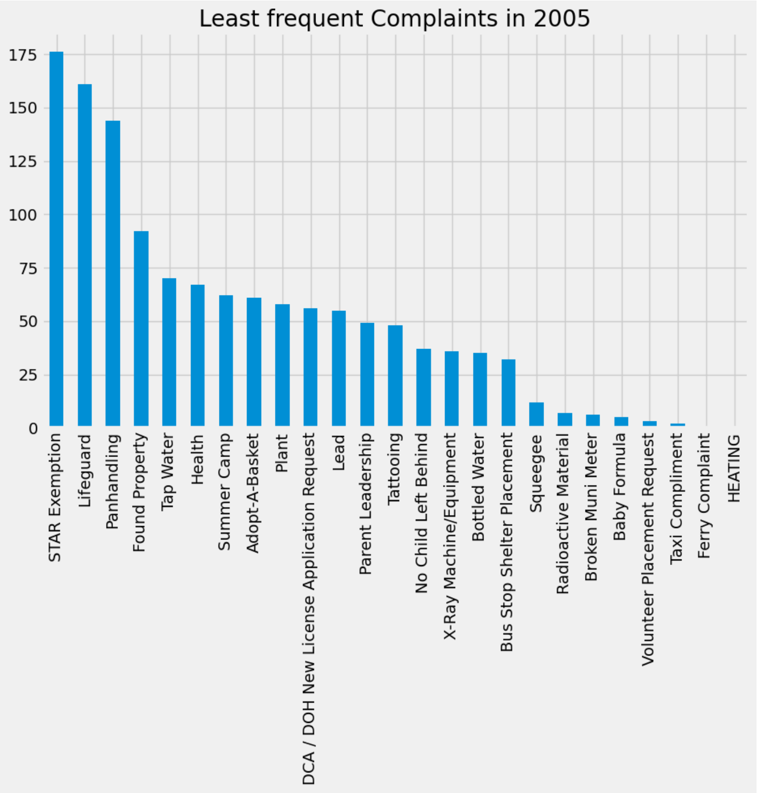


**Figure 7:** *The least frequent complaints in 2005*

## 6.2 Borough that most SRs come from

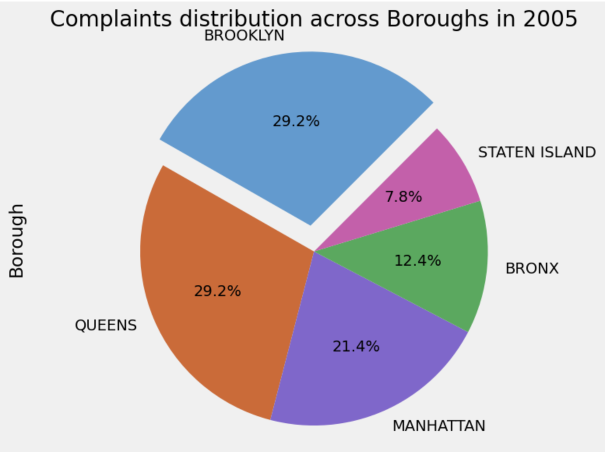


**Figure 8:** *Borough getting most number of complaints in 2005*

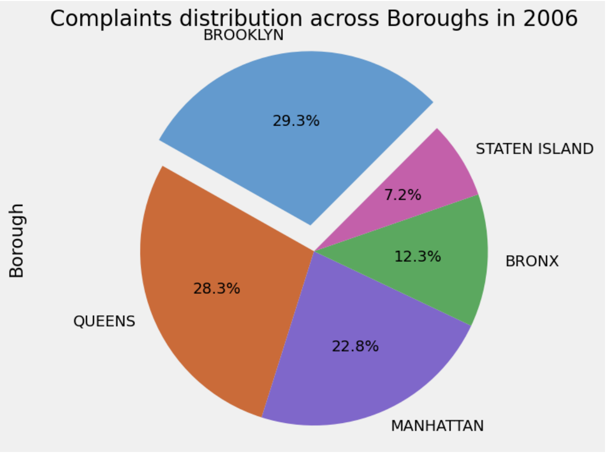


**Figure 9:** *Borough getting most number of complaints in 2006*

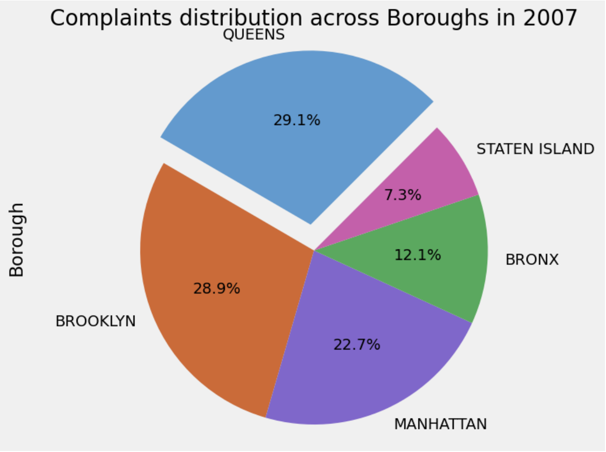


**Figure 10:** *Borough getting most number of complaints in 2007*

## 6.3 The most common types of SRs in Brooklyn

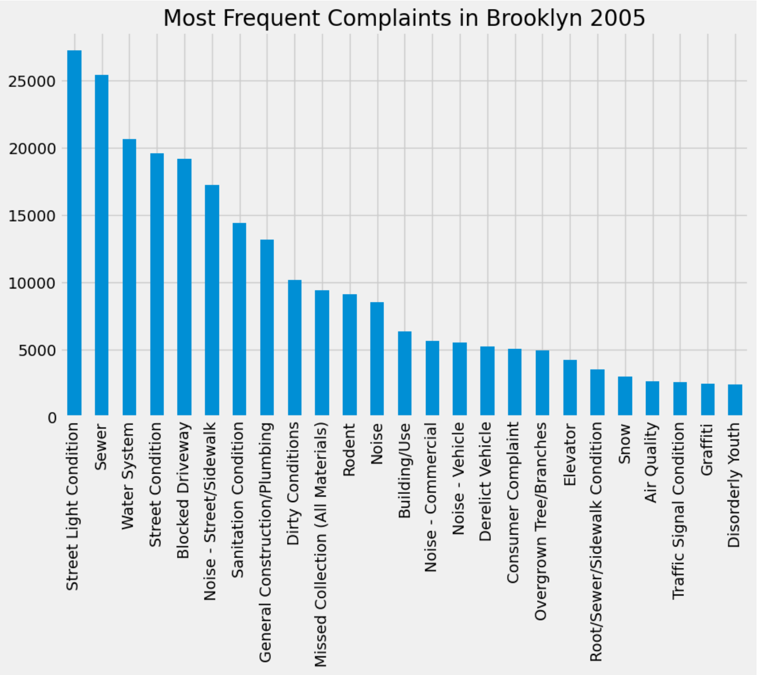


**Figure 11:** *The most common complaints across Brooklyn in 2005*

## 6.4 Response time of the most complaint types in Brooklyn

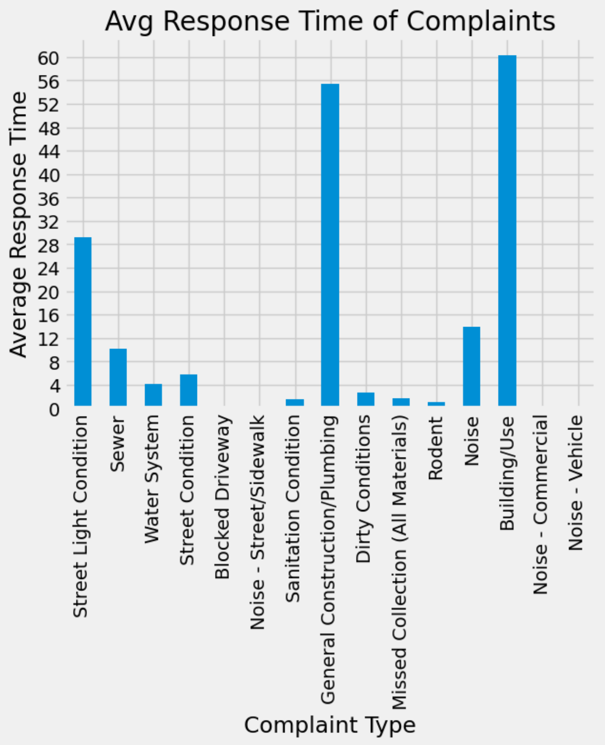


**Figure 12:** *Response time of the most complaint types in 2005*

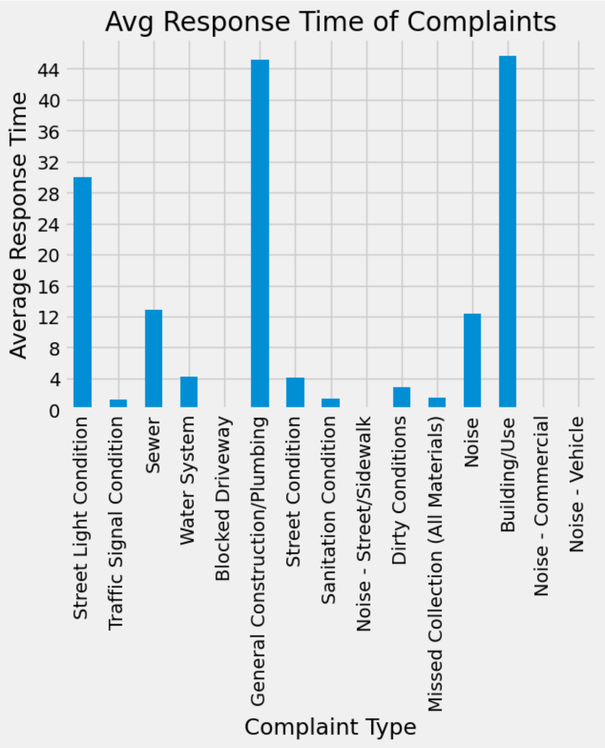


**Figure 13:** *Response time of the most complaint types in 2006*

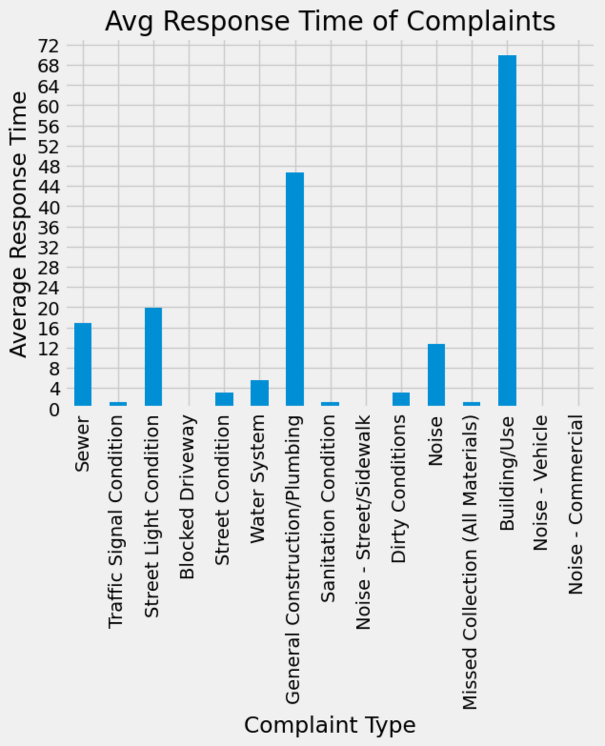


**Figure 14:** *Response time of the most complaint types in 2007*

We can see here that, generally Noise and illegal parking issues are resolved on the same day. While issues like Street light condition take a rather long time to resolve.

### 6.5 The most efficient agencies responsible for addressing SRs across Brooklyn

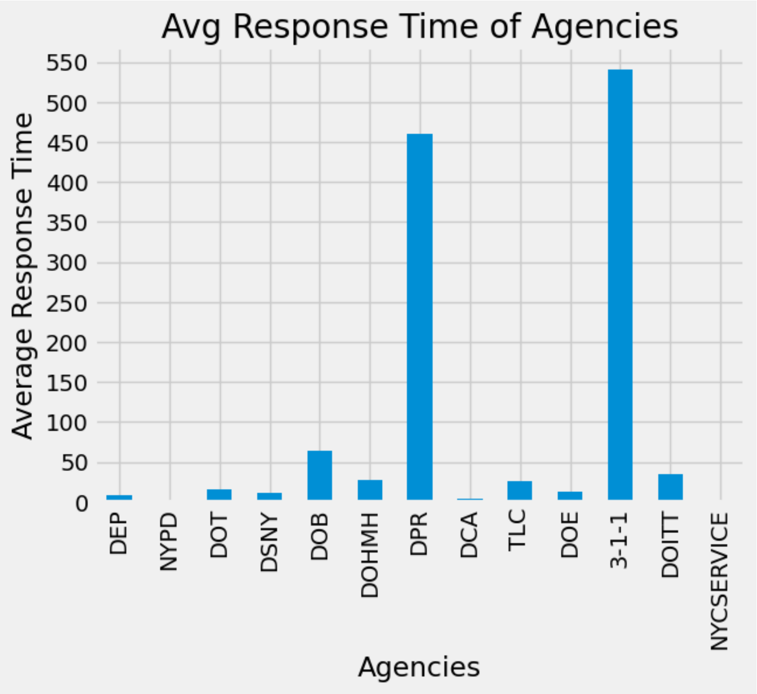


**Figure 15:** *The most efficient agencies across Brooklyn in 2005*

From the above graph we can conclude that NYPD and NYCSERVICE are the most efficient agencies among the rest. These are also the agencies which generally solve the most number of complaints.

### 6.6 Location types that we get most number of complaints across Brooklyn

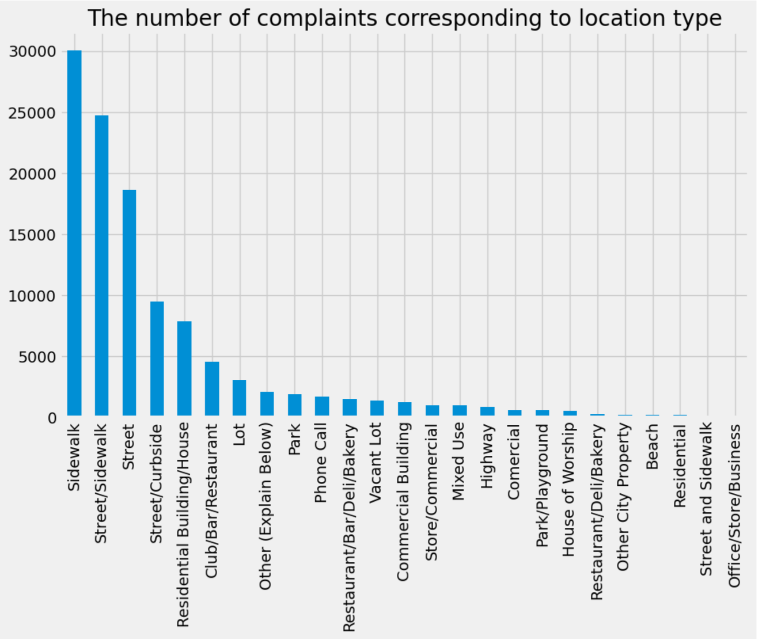


**Figure 16:** *Location types getting most number of complaints in 2005*

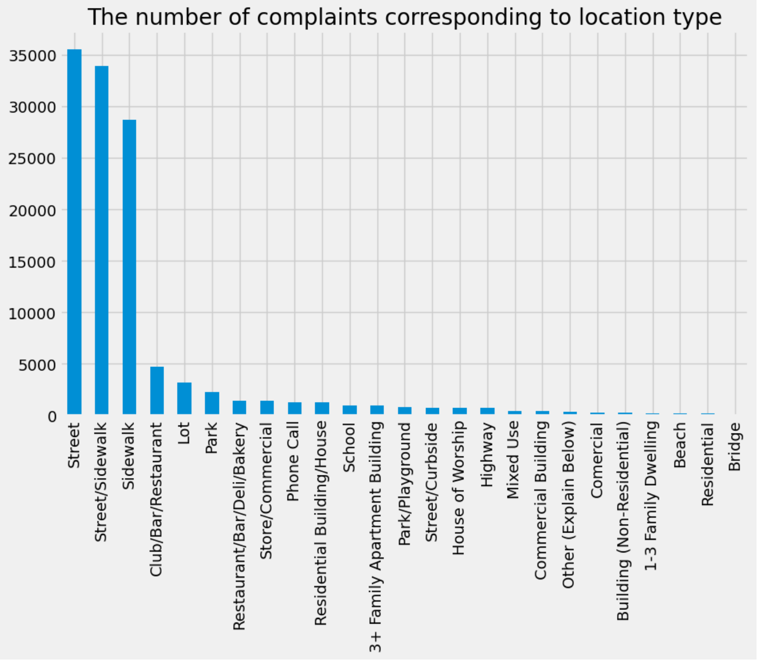


**Figure 17:** *Location types getting most number of complaints in 2006*

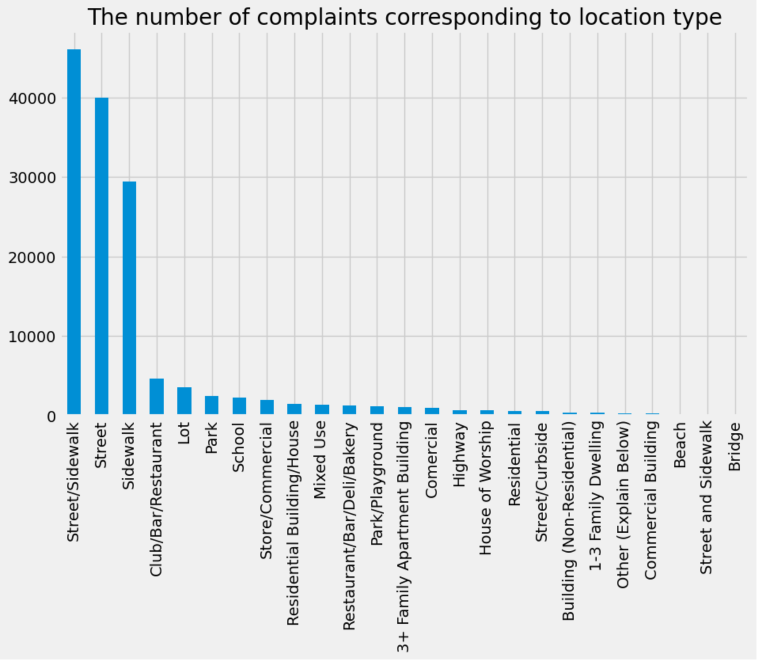


**Figure 18:** *Location types getting most number of complaints in 2007*

This graph is telling us that most number of complaints through 2005-2007 came from sidewalk and street. We can relate that most common complaint type is Street Light Condition and Sewer.

## 6.7 SRs concentration across Brooklyn

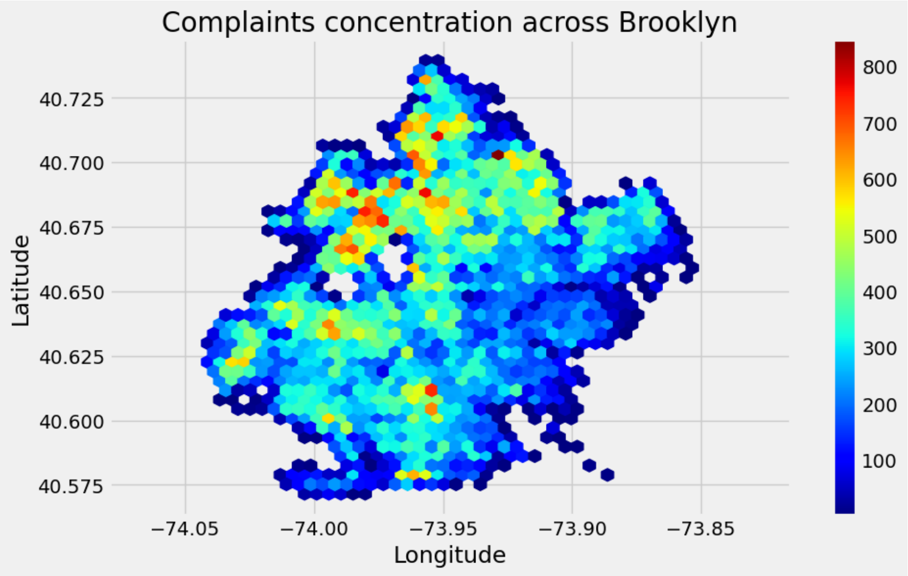


**Figure 19:** *Complaints concentration in 2006*

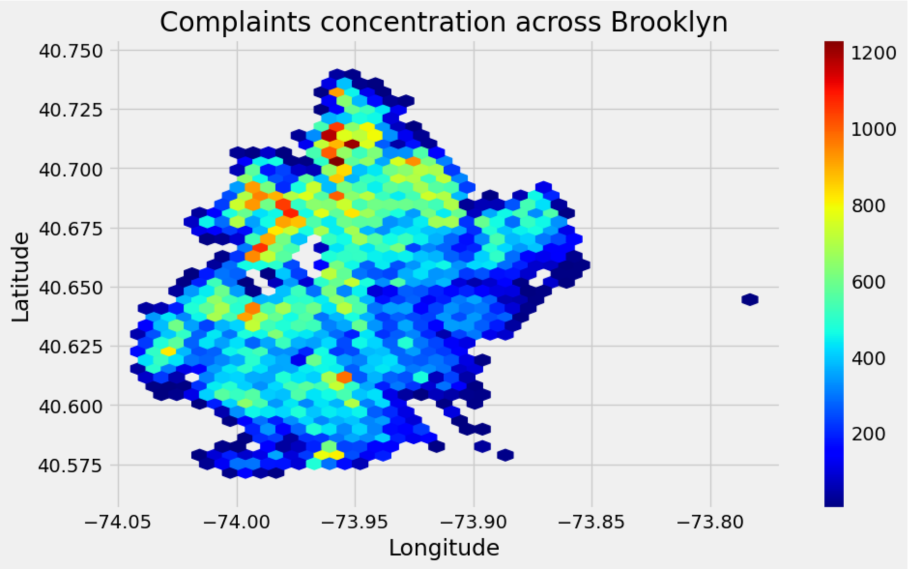


**Figure 20:** *Complaints concentration in 2007*

### 6.8 Analyse the most frequent complaint across Brooklyn

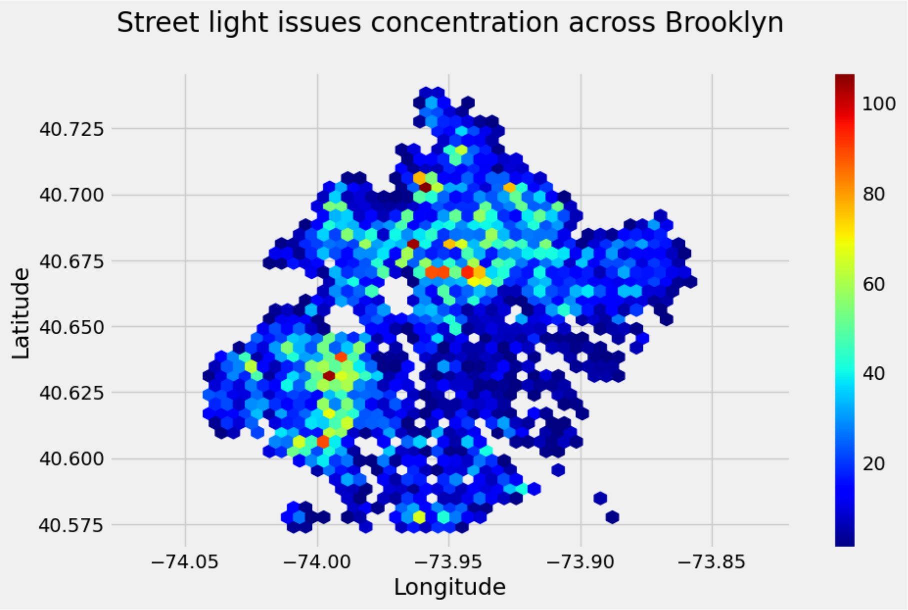


**Figure 21:** *Street light concentration in 2005*

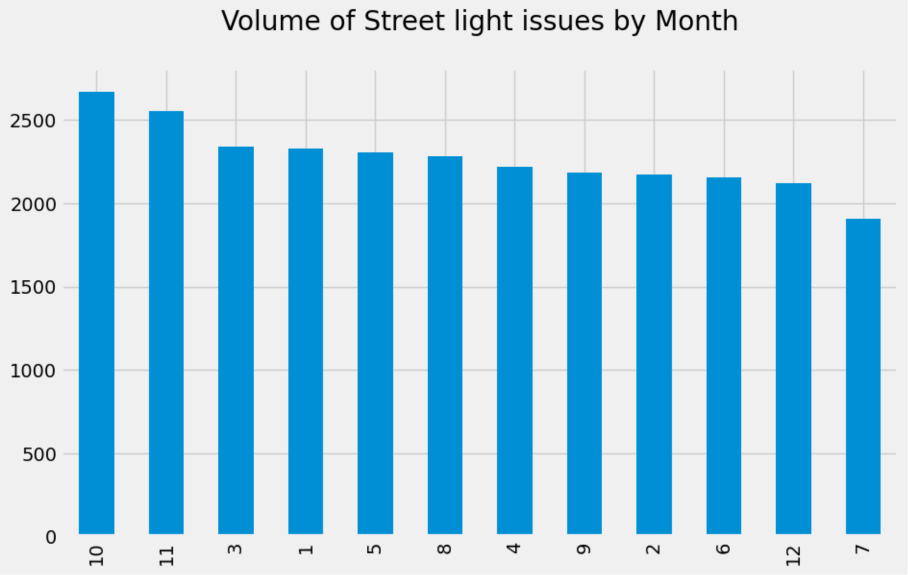


**Figure 22:** *Volume of street light issues in 2005*

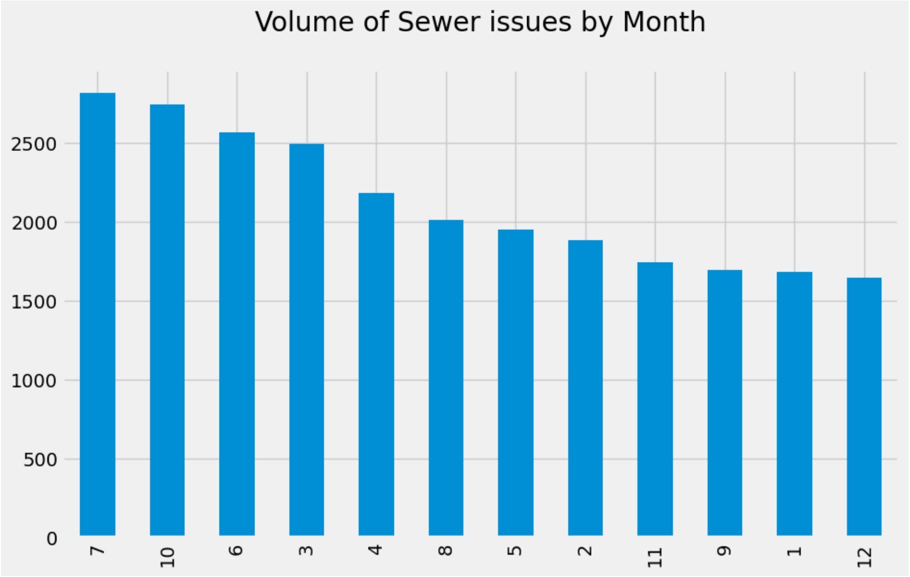


**Figure 23:** *Volume of sewer issues in 2005*

## 6.9 Predict the volume of service requests in the future for New York

We will train on the volume request data of the years 2005, 2006, and 2007, and then predict the volume request data of the year 2008.

To do this, we perform the following steps:

- We transform the volume request values by taking the natural logarithm of them. This transformation is used to reduce the difference between large and small values in the data and make it easier to analyze and model.
- Next, we train the model on the number of requests generated for each day in the years 2005, 2006, and 2007, and use this model to predict the requests for each day in the year 2008.
- Finally, we will compare the predicted results with the known volume request data of the year 2008. In addition, we will use the mean squared error metric to evaluate the performance of this model.

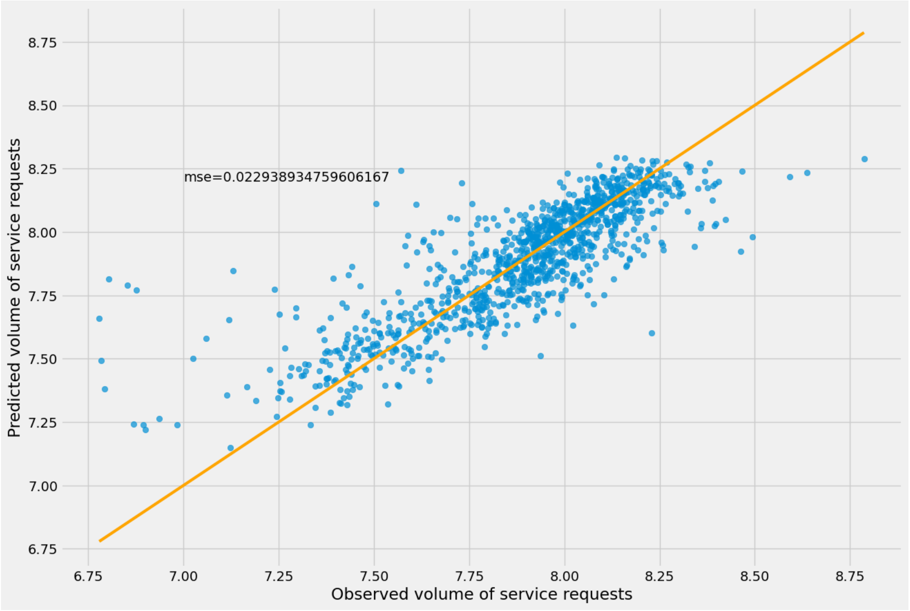


**Figure 24:** *Compare observed volume to predicted volume*

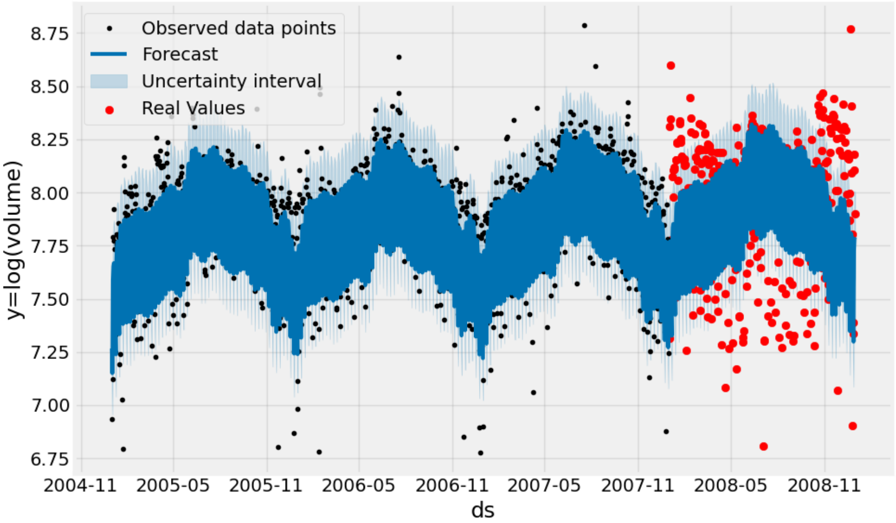


**Figure 25:** *Predicted result of model*

This plot is drawing the observed data through 2005-2007 (black dots), the model (blue line), the error of the forecast (uncertainly interval) and the observed data in 2008. The plot is showing the forecasted values for entire data as well as the new data for 2008. The shaded area in graph is the uncertaintity estimation of forecast.

## 7 Conclusion

The analysis of service request data has provided valuable insights into the volume, frequency, timing, location, and types of requests received by the goverment. The analysis revealed that there are notable trends and patterns in the data, such as an increase in request volume on some days, months in year or a correlation between the location and type of request.

Based on these findings, we recommend that the organization consider increasing staffing in areas with high request volumes on predicted days or implementing different measures to address different types of requests.

It is important to note that there are limitations to the analysis, such as the assumption that the observed patterns in the data will continue into the future.

Overall, this analysis provides a solid foundation for further research into service request data and highlights the potential benefits of using data-driven insights to optimize organizational performance.

# References

[1] Wes McKinney (2017). Python for Data Analysis, 2e: Data Wrangling with Pandas, Numpy, and Ipython.

[2] Jake Vander Plas (2016). Python Data Science Handbook: Tools and Techniques for Developers.

[3] Ralph Kimball, Margy Ross (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling