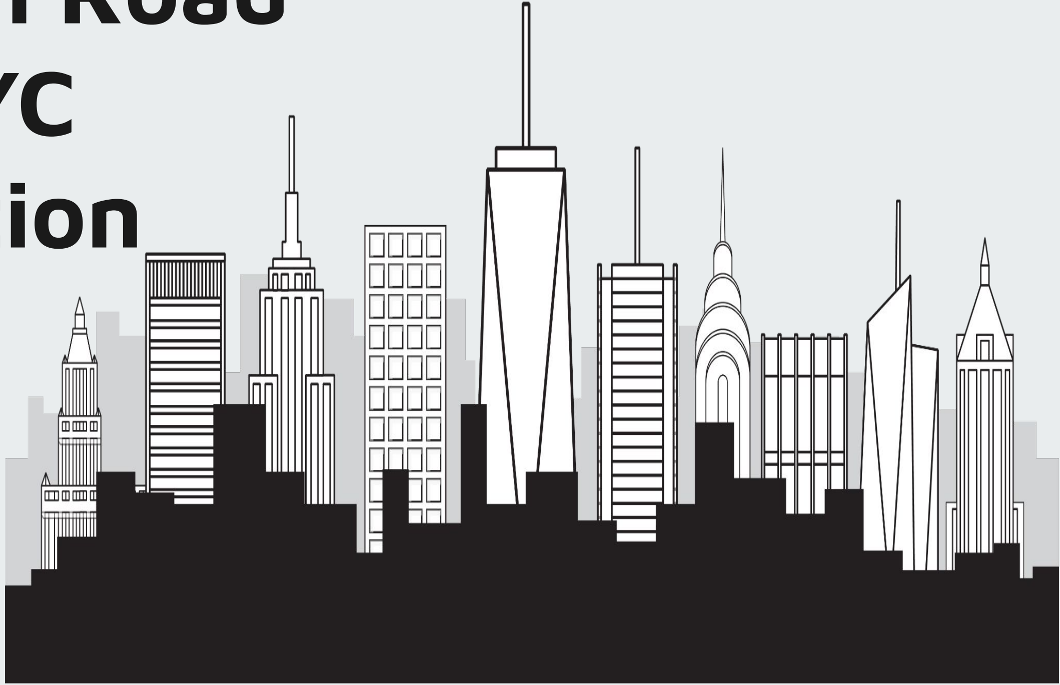
# Unveiling Urban Road Challenges - NYC Pothole Prediction

A Data-Driven Exploration of Potholes in NYC with Predictive Modeling, Spatial Analysis, and Strategic Recommendations

By Sarmad Maqbool



# Outline

The Problem

Solution Proposal

Introduction to Data

Exploratory Data Analysis

# The Problem

In the bustling streets of New York City, where individuals spend an average of 2 hours daily navigating the vibrant urban landscape, the quality of their commuting experience is a critical factor.

Imagine a scenario where this daily journey is marred by the presence of potholes, and unpredictable road hazards that not only pose a threat to vehicle safety but also significantly impact commute times.

This project has a serious potential as there is no work done in this space earlier and I really want to help the authorities to address this issue effectively.

# Solution Proposal

This data science project aims to tackle the issue head-on by leveraging advanced predictive modeling techniques to identify and forecast pothole locations across NYC. By analyzing historical data, weather patterns, road maintenance schedules, and other relevant factors, we seek to develop a robust predictive model that can anticipate the likelihood of pothole formation in different areas of the city. "The ultimate objective is to equip municipal authorities with real-time insights, enabling them to proactively address the issue by identifying and prioritizing potential pothole-prone areas in advance.

Through this project, we aim to contribute to the development of a smarter and more resilient urban transportation system in NYC, ultimately improving the quality of life for its residents and visitors.

# The Impact

The occurrence of potholes is a common challenge faced by both drivers and pedestrians alike, leading to not only potential vehicle damage but also contributing to traffic congestion and increased travel times.

By efficiently managing potholes, the city can potentially save millions of dollars in repair costs and minimize the environmental impact associated with frequent road maintenance. The project's societal value lies in improving the overall quality of transportation infrastructure in NYC, positively impacting businesses, residents, and the environment.

# Steps Involved in the Solution

- Acquiring Data and Intro
- Data Cleaning
- Data Analysis
- Presenting Insights
- Modelling Data
- Testing the Model
- Predicting Future Potholes
- Deployment

# Acquiring Data

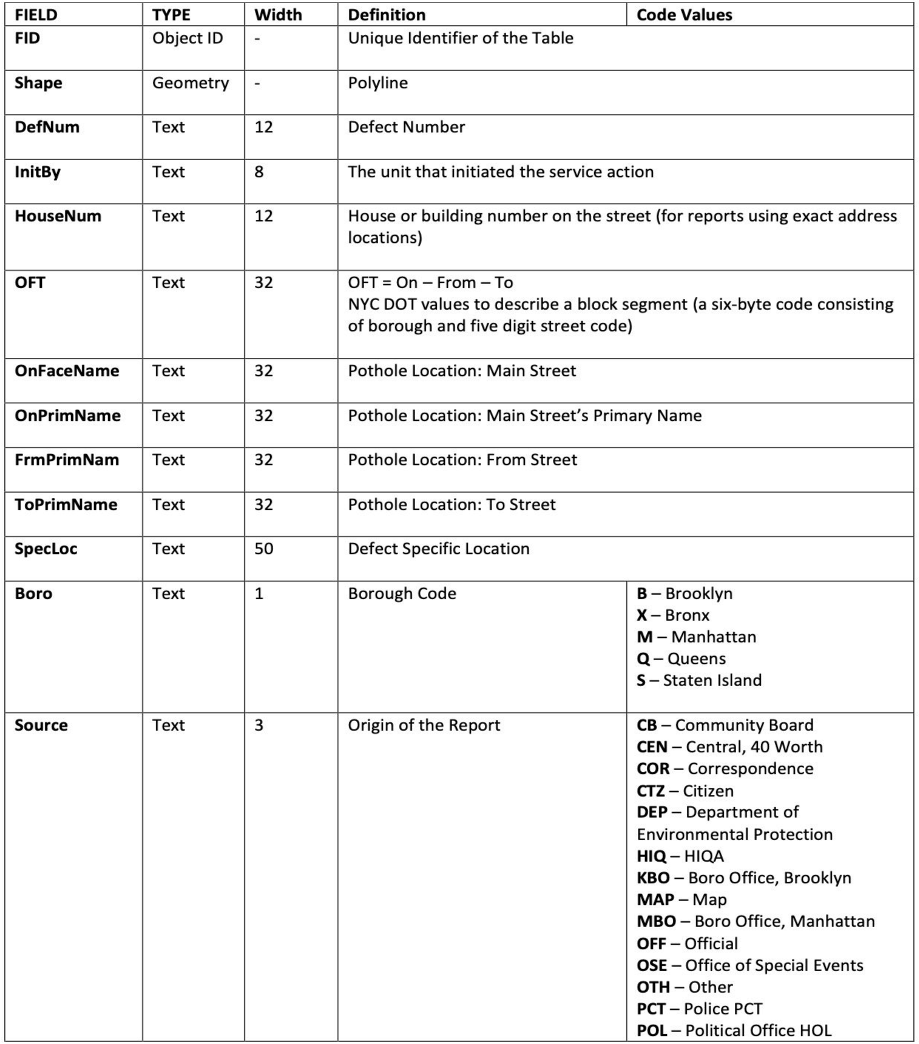
NYC Open Data 311 Service Requests Dataset: This dataset, available on NYC Open Data, contains information about service requests, including those related to potholes. It includes details such as the date of the request, location, and the status of the reported issue. ● Street Pothole Work Orders - Closed (Dataset) | NYC Open Data,

https://data.cityofnewyork.us/Transportation/Street-Pothole-Work-Orders-Closed-Dataset-/x9wy-ing4/data

# Introduction to Data

The most important features which are going to help us perform our analysis are

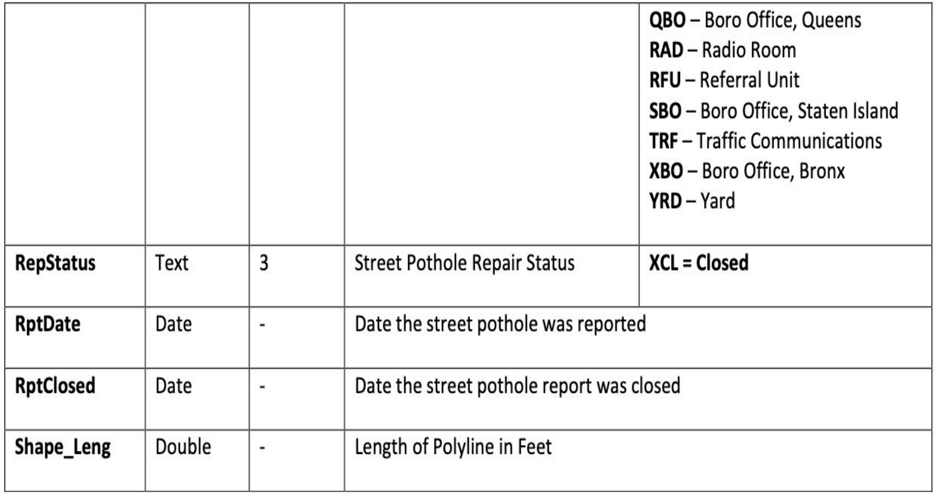
- Location
- Source
- Date Reported
- Date Resolved
- Status of Report
- Size of Pothole



| FIELD | TYPE | Width | Definition | Code Values |
|---|---|---|---|---|
| **FID** | Object ID | - | Unique Identifier of the Table | |
| **Shape** | Geometry | - | Polyline | |
| **DefNum** | Text | 12 | Defect Number | |
| **InitBy** | Text | 8 | The unit that initiated the service action | |
| **HouseNum** | Text | 12 | House or building number on the street (for reports using exact address locations) | |
| **OFT** | Text | 32 | OFT = On – From – To<br>NYC DOT values to describe a block segment (a six-byte code consisting of borough and five digit street code) | |
| **OnFaceName** | Text | 32 | Pothole Location: Main Street | |
| **OnPrimName** | Text | 32 | Pothole Location: Main Street's Primary Name | |
| **FrmPrimNam** | Text | 32 | Pothole Location: From Street | |
| **ToPrimName** | Text | 32 | Pothole Location: To Street | |
| **SpecLoc** | Text | 50 | Defect Specific Location | |
| **Boro** | Text | 1 | Borough Code | **B** – Brooklyn<br>**X** – Bronx<br>**M** – Manhattan<br>**Q** – Queens<br>**S** – Staten Island |
| **Source** | Text | 3 | Origin of the Report | **CB** – Community Board<br>**CEN** – Central, 40 Worth<br>**COR** – Correspondence<br>**CTZ** – Citizen<br>**DEP** – Department of Environmental Protection<br>**HIQ** – HIQA<br>**KBO** – Boro Office, Brooklyn<br>**MAP** – Map<br>**MBO** – Boro Office, Manhattan<br>**OFF** – Official<br>**OSE** – Office of Special Events<br>**OTH** – Other<br>**PCT** – Police PCT<br>**POL** – Political Office HOL |

# Introduction to Data

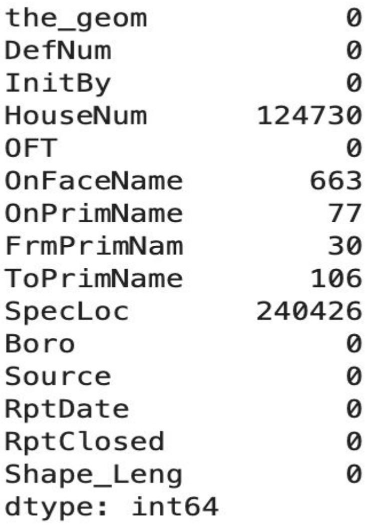
Shape and length of a pothole is given in "Polyline". A polyline in GIS (geographic information system) consists of interconnected line segments representing linear features like rivers, roads, trails, and administrative boundaries.



| | | | | |
|---|---|---|---|---|
| | | | | **QBO** – Boro Office, Queens<br>**RAD** – Radio Room<br>**RFU** – Referral Unit<br>**SBO** – Boro Office, Staten Island<br>**TRF** – Traffic Communications<br>**XBO** – Boro Office, Bronx<br>**YRD** – Yard |
| **RepStatus** | Text | 3 | Street Pothole Repair Status | **XCL = Closed** |
| **RptDate** | Date | - | Date the street pothole was reported | |
| **RptClosed** | Date | - | Date the street pothole report was closed | |
| **Shape_Leng** | Double | - | Length of Polyline in Feet | |

# Size of Data and Data Cleaning

We have 360730 rows and 15 columns in this data. Out of 15 Features, we have 13 objects, 1 Float and 1 Int.



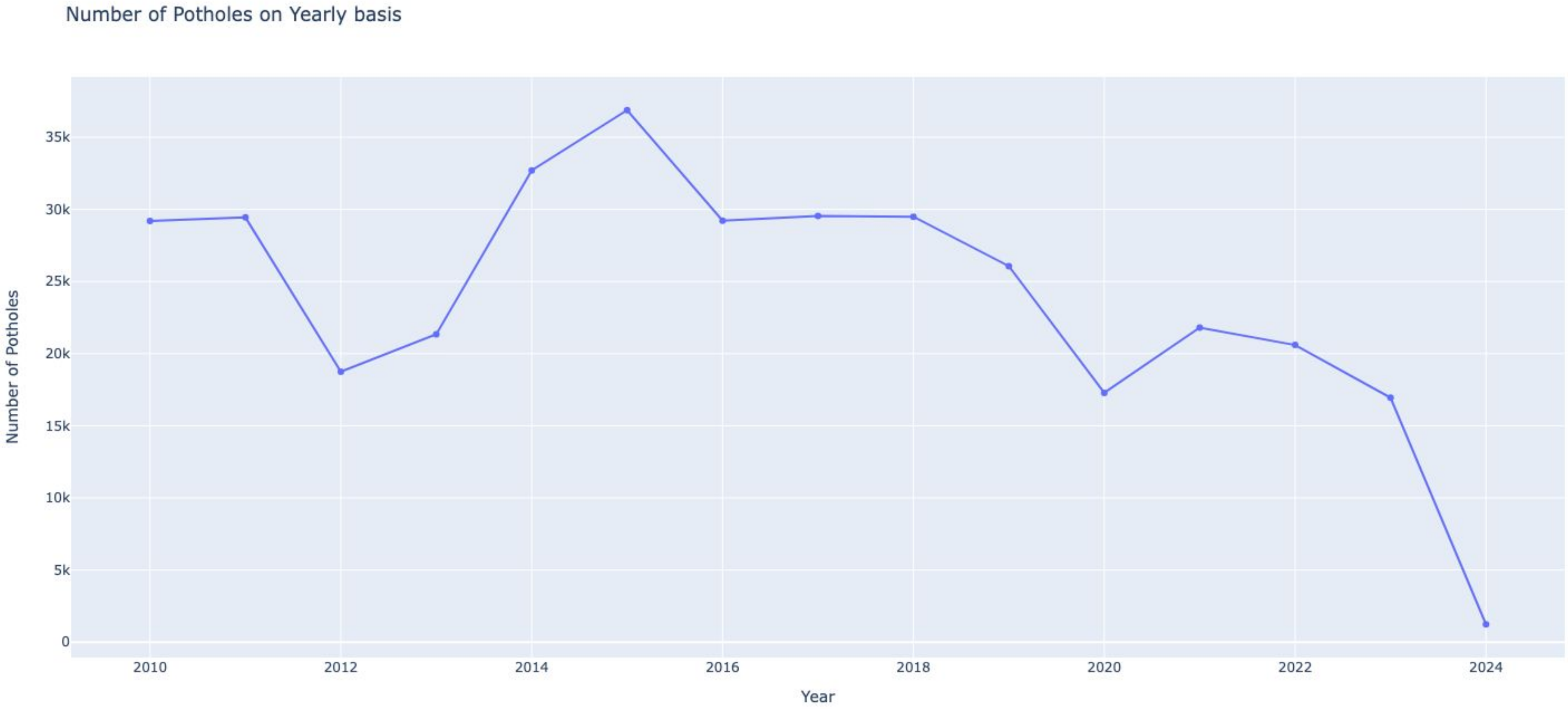
```
the_geom             0
DefNum               0
InitBy               0
HouseNum        124730
OFT                  0
OnFaceName         663
OnPrimName          77
FrmPrimNam          30
ToPrimName         106
SpecLoc         240426
Boro                 0
Source               0
RptDate              0
RptClosed            0
Shape_Leng           0
dtype: int64
```

We have few nulls in the data but mostly in House_num and specific location and it is not going to help much with analysis because they have around 30% and 60% nulls respectively. As we already have the other columns displaying the same information "OFT, OnFaceName , OnPrimName, FrmPrimNamToPrimName" so I am going to drop columns "House_num" and 'SpecLoc'. Also Imputing null values of location from other columns.

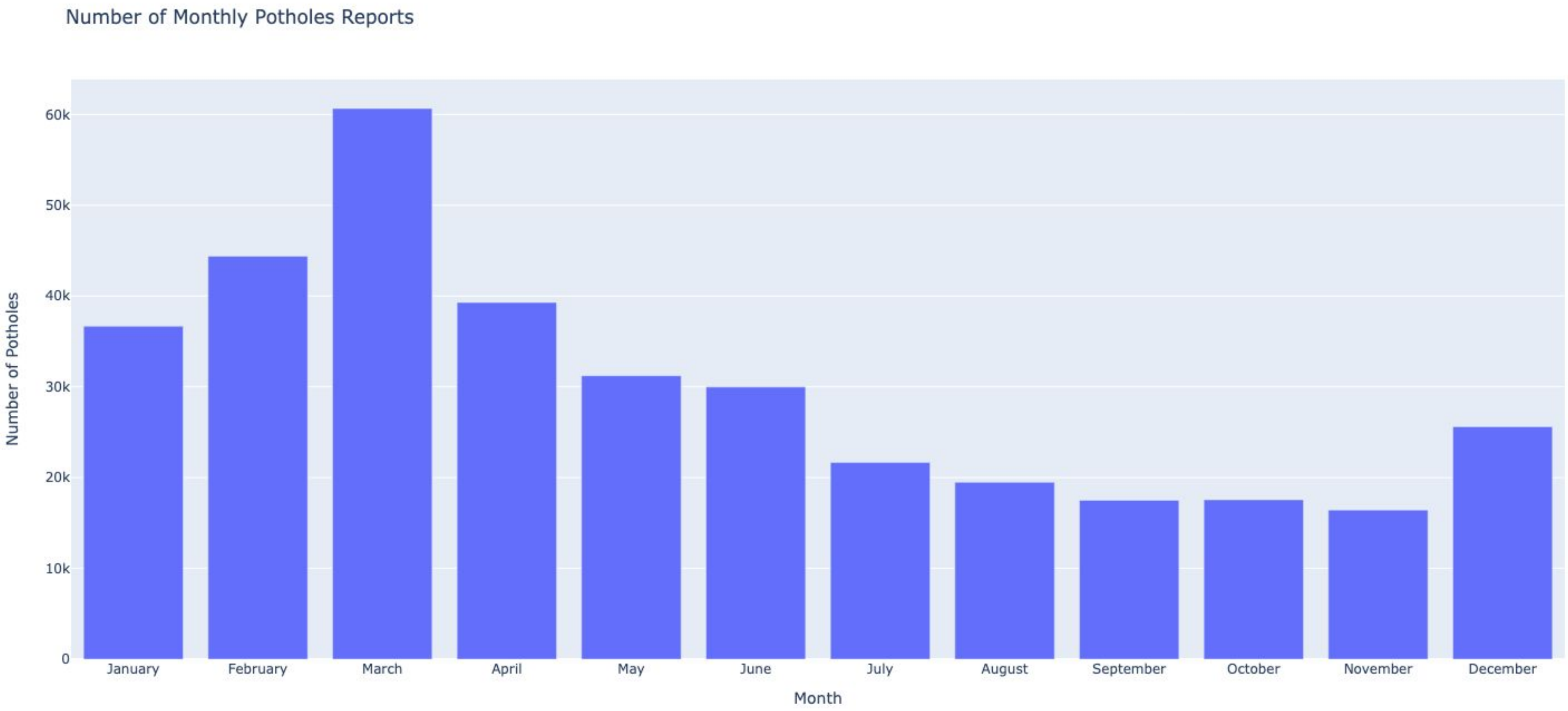
# Data Visualization

The tool I have used for the visualization is the **PLOTLY** Library. I have covered the yearly, monthly and departmental distribution of the potholes.

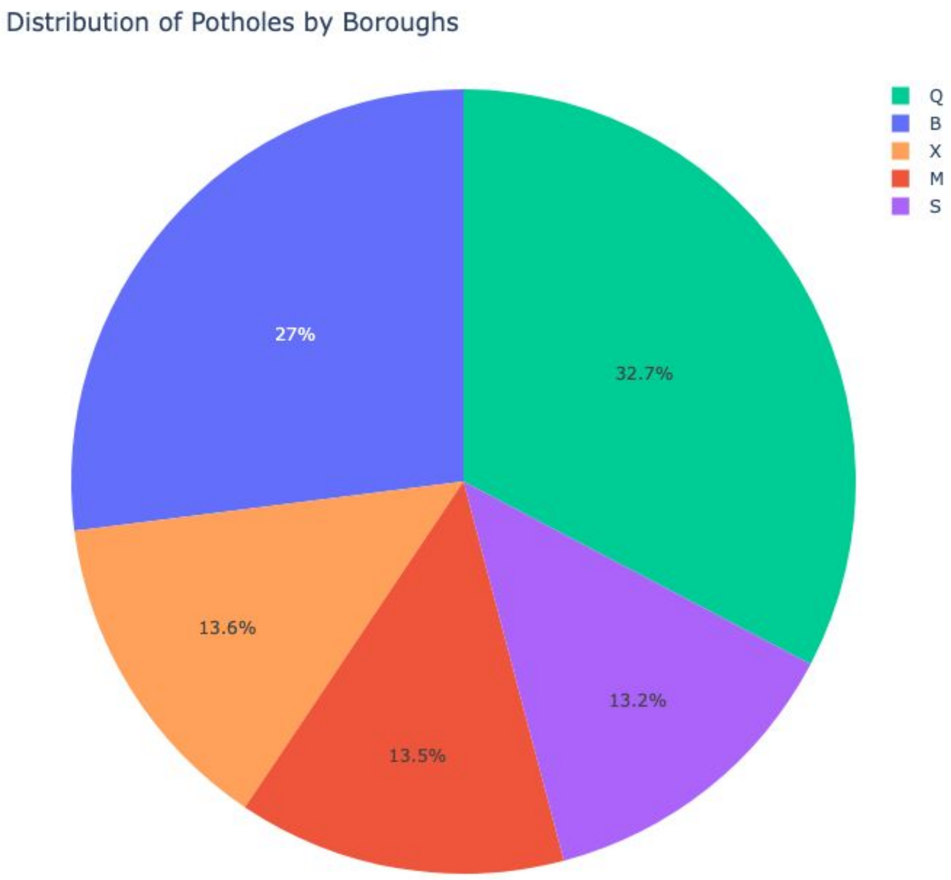
Also covered which Boroughs are most prone to potholes.



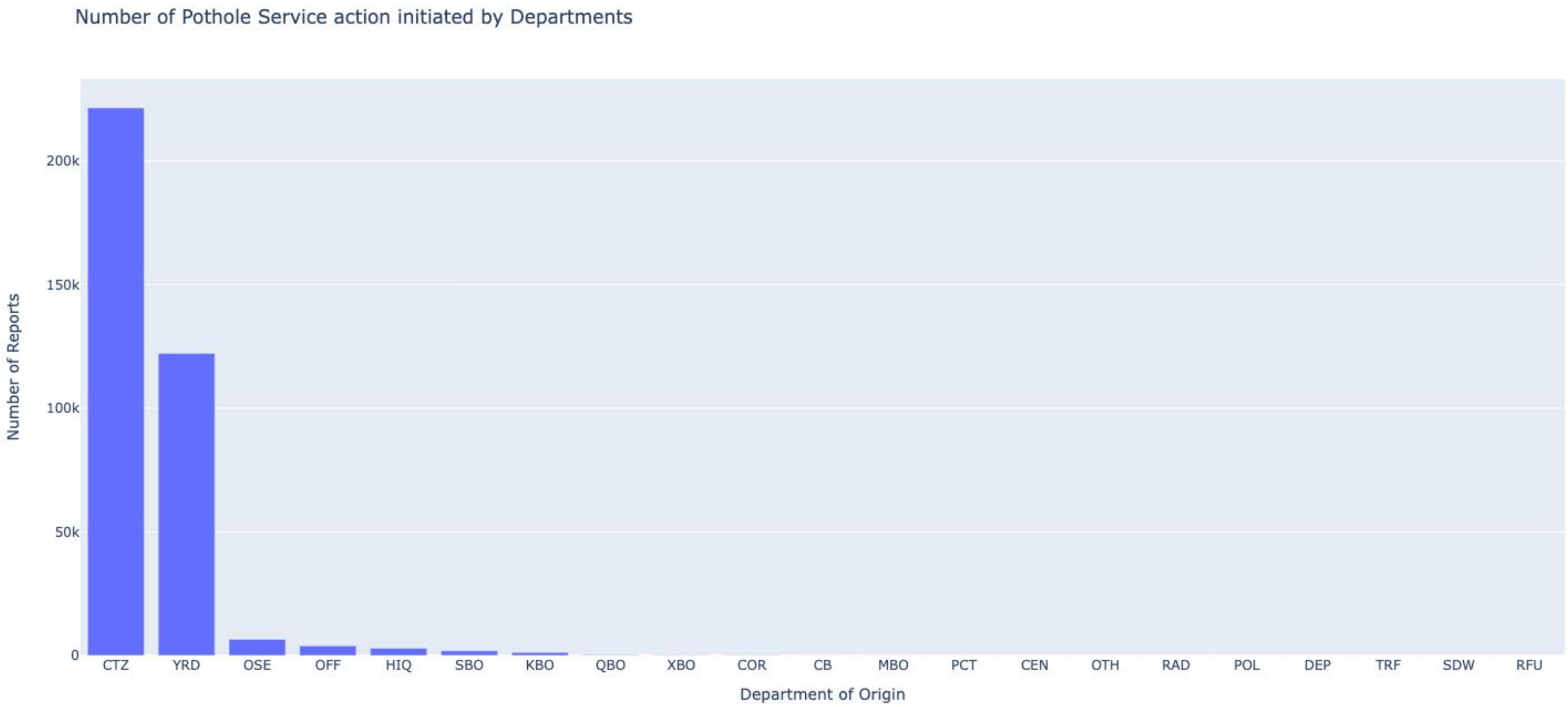
The highest number of potholes recorded in NYC in a calendar year was in 2015 with slightly over 36K potholes. Overall the yearly average is 24032 potholes per year.



The trend to note here is that the pothole numbers are maximum in the spring season which helped me to unveil an interesting fact that If the water freezes and thaws over and over, the pavement will weaken and continue cracking.



We can see that the most number of potholes occurrences are in Queens and then Brooklyn. We can assume that number of potholes are dependent on the area of Boroughs and population.



The citizens department has handled the most number of complaints regarding to potholes and then Yard department contributed a lot with the pothole repair. Both departments contributed almost 90% to the reports generated for potholes in this dataset.

# Next Steps

- Feature Engineering (check relationship between features)
- Modeling (Time Series)
- Testing (Scikit Learn Train Test)
- Predicting
- Deployment

# Questions?