NYC Pothole Prediction

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

By Sarmad Maqbool

Problem

In NYC's busy streets, where people spend 40 mins daily commuting, road quality is crucial. Picture this: Daily journeys disrupted by potholes and unpredictable hazards, endangering safety and delaying commutes. **Can we predict the number of days required to fix a pothole?**

Motivation

My recent encounter with a pothole on my way home in NYC sparked a realization: the need for proactive measures to prevent such incidents. I tried to learn and research more about potholes, most affected routes, what are the authorities doing about it, but there is not enough work done in this area.

Solution

This project aims to predict the **NUMBER OF DAYS** required to fix a pothole using zip codes, street names, traffic data, weather patterns leveraging machine learning.

Impact

Potholes pose a common challenge for drivers and pedestrians, causing vehicle damage and traffic congestion. Efficient pothole management can save millions in repair costs and reduce environmental impact. This project aims to enhance NYC's transportation infrastructure, benefiting businesses, residents, and the environment.

Overview of Dataset

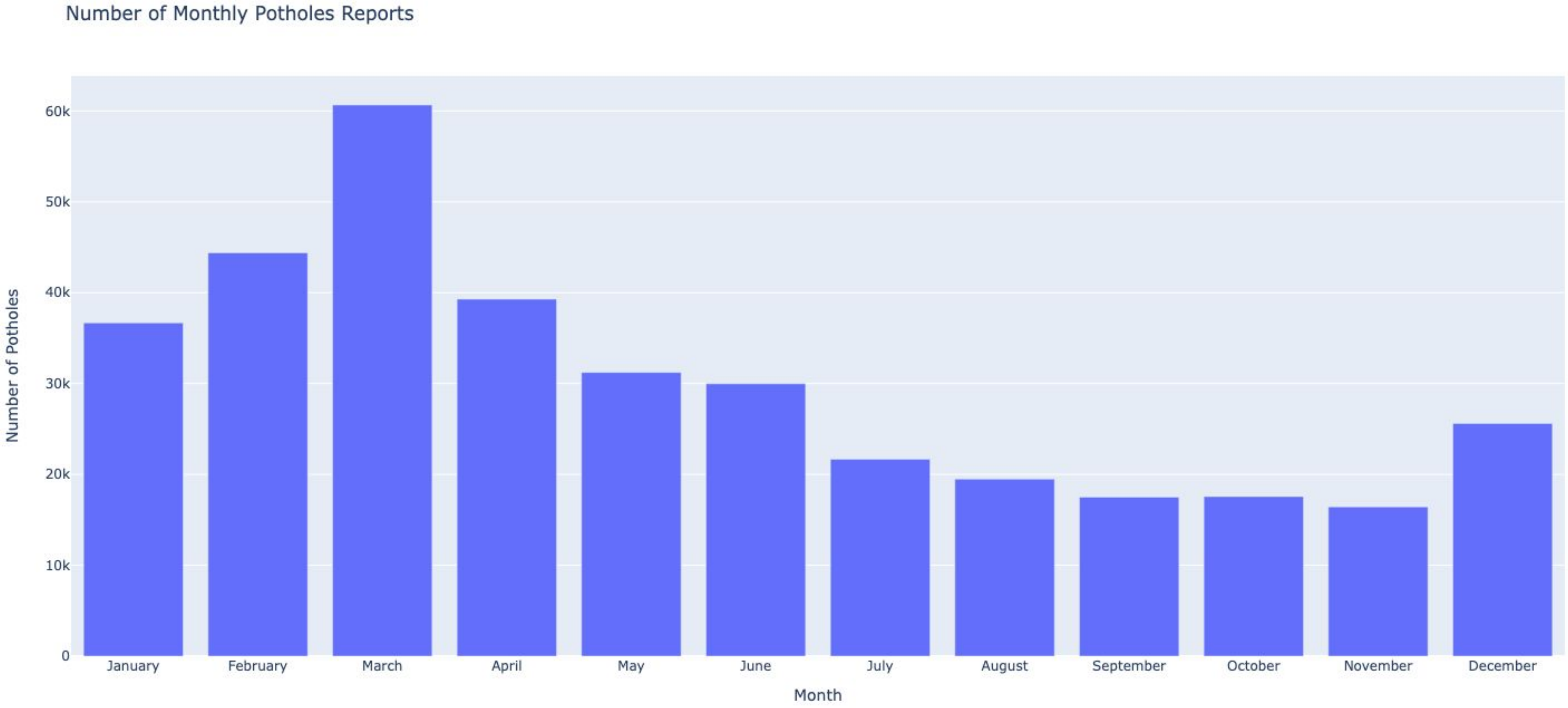
- **NYC Open Data 311 Service Requests Dataset:** This dataset, available on NYC Open Data, contains information about service requests, including those related to potholes. It includes details such as the date of the request, location, and the status of the reported issue ranging from 2004 to 2023. There are 370000 rows and 15 columns,

  229 unique zip codes, 5800 street names, average time to fix a pothole is 4.5 days

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

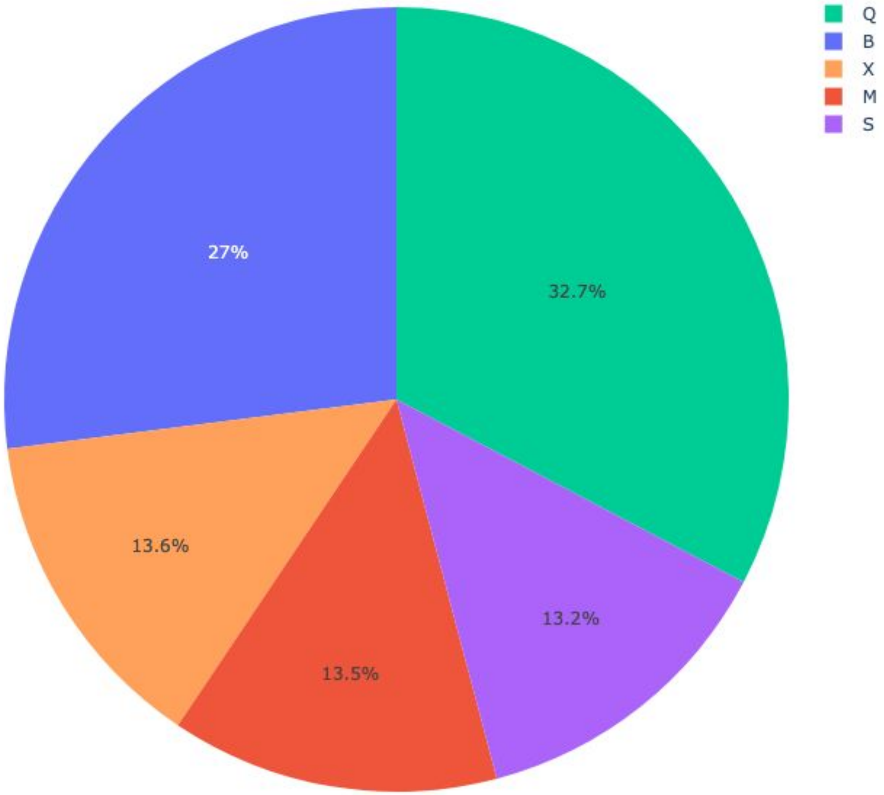
- **New York City Weather: A 154-Year Retrospective:** This Dataset is obtained from Kaggle, contains information about Central Park NYC daily weather from year 1869 to 2022. It includes columns like daily TMIN, TMAX, PRECP, SNOW.

  https://www.kaggle.com/datasets/danbraswell/new-york-city-weather-18692022



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.**

Distribution of Potholes by Boroughs



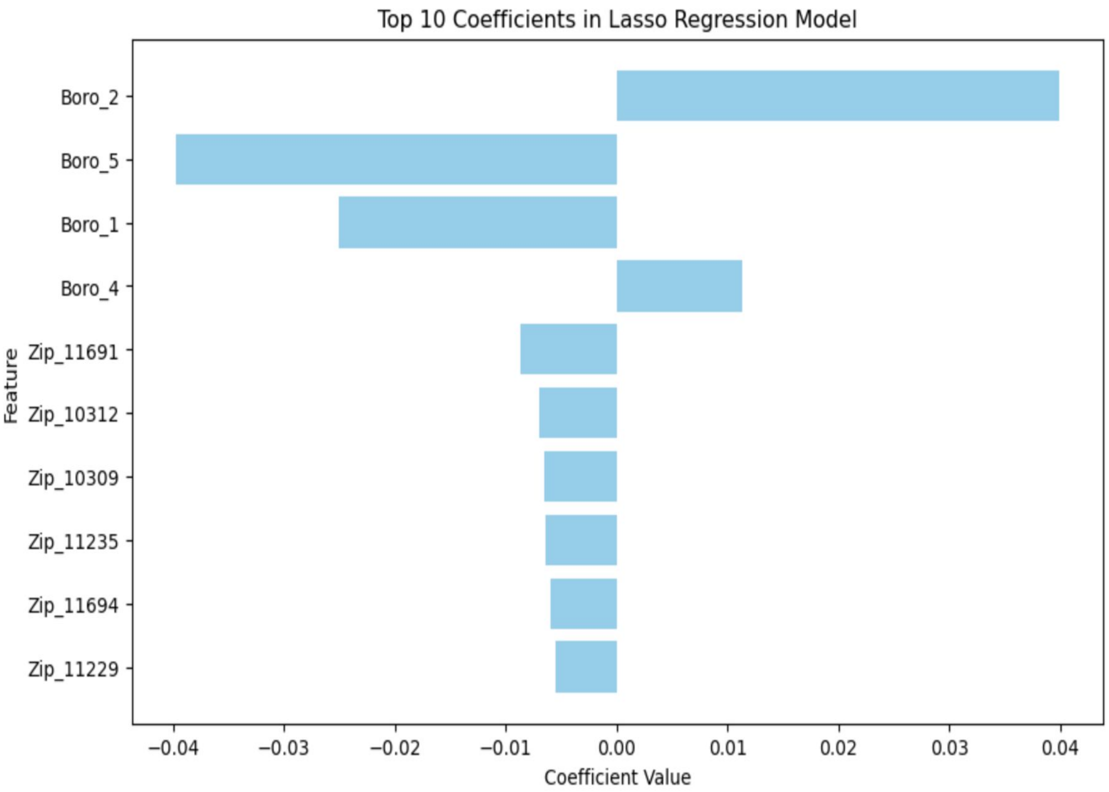
We can see that the most number of potholes occurrences are in Queens and then Brooklyn. We can assume that number of potholes are dependent on the area of Boroughs, Population and Traffic.

Model Performance

ML Algorithm	**R^2**	**MSE**
Linear Regression	0.2696	34.38
Ridge Regression	0.7705	0.0017
Lasso Regression	**0.9712**	**0.0002**
Random Forest Regressor	0.3185	32.08

The best model I got after multiple iterations is Lasso Regression. After several iteration and tuning our data, we have a decent R^2 of 97%.

Model Discussion



```
R-squared (R2) Score: 0.9712053015299478
Mean Squared Error (MSE): 0.00022185674732133983
```

Lasso Regression Model has a **R^2 score of 97%** which is pretty decent. The Features with the most predictive power are **Boroughs and Zip Codes**. It should be noticed that some zip codes and Boroughs are bigger in size and population hence contributing more to the feature importance.

We can also interpret that some locations have more quicker fixes of potholes despite of the large number of potholes so are less contributing.

Boro_2: Bronx

Boro_5: Staten Island

Boro_1: Brooklyn

Boro_4: Manhattan

Next Steps:

- Introduce Traffic Data
- Introduce Population and Area distribution of Zip Code
- Optimizing hyperparameters
- Engineering more features
- More data

Thank you

Connect with me on LinkedIn

Sarmad Maqbool

BrainStation Alumni | Data Scientist - Python - Machine Learning - Neural Networks - Supervise...