|  |
| --- |
| Houston Pothole Analysis |
| Large rock formation in rocky landscape under blue sky |

|  |  |
| --- | --- |
| 7/4/2021 | A look into the predictability of pothole creation based on weather conditions |

Houston Pothole Analysis

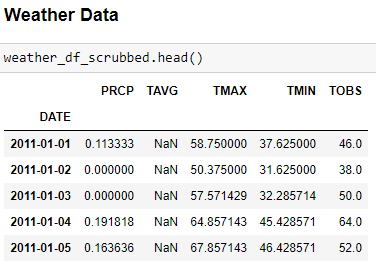
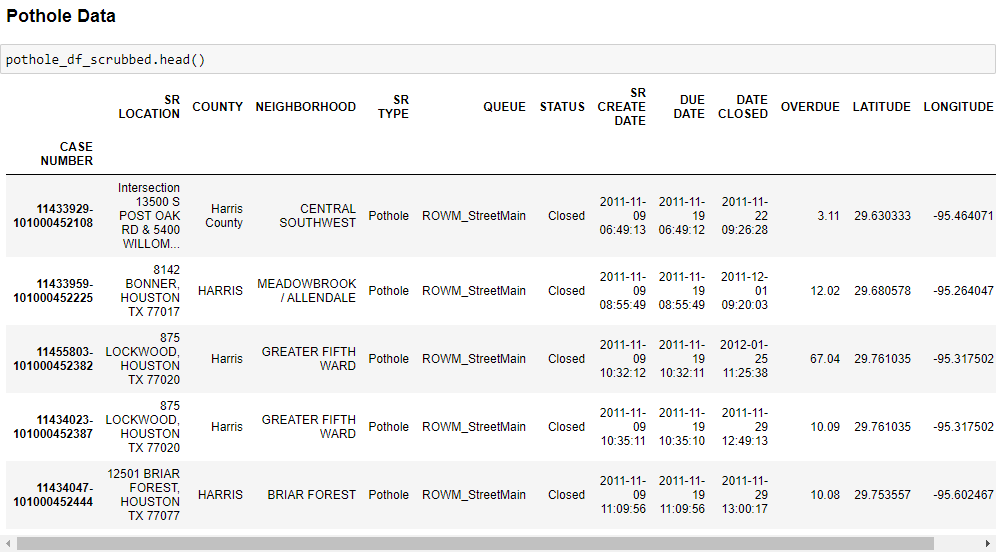
A look into the predictability of pothole creation based on weather conditions

# Purpose

The purpose of this deep dive is to come up with a predictive model for determining pothole creation location in Houston, Texas while utilizing the existing pothole data and weather forecast data. If potholes could be predicted it could allow for better planning for commuters and better planning for the city to distribute resources more effectively. The City of Houston has a next day policy for pothole repairs, so the more notice, the better.

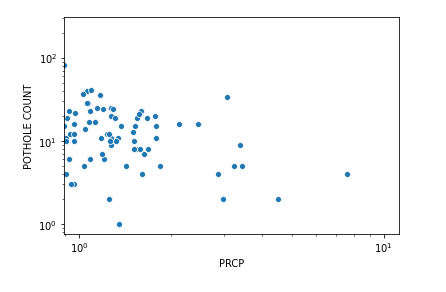
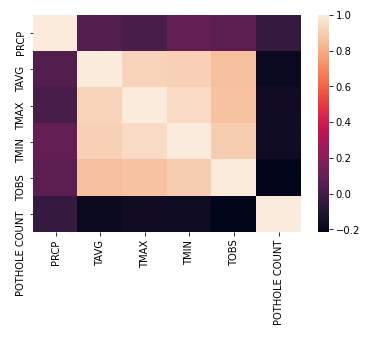
## Data Wrangling

The pothole data is supplied by the City of Houston Pothole record site. Each year is split into a separate JSON file. Iterating through each page and appending to a pandas dataframe produced this data set. The weather data for Houston was taken from the National Oceanic and Atmospheric Administration. Pulling directly from an FTP into a pandas dataframe populated this data set. Both data sets were sampled to a daily timestamp. No days were skipped as there were reported potholes and temperatures/precipitation for each day. The data was cleaned of null values and miscellaneous columns that were categorical but not of interest. Null values for mean temperatures being substituted with the mean of the High Temp/Low Temp of the day as these values did not contain missing data. The datasets were then combined with the common index of the date. A total of 10 years of data was collected.



## EDA

A deeper dive into the data was performed to see if any statistical correlation could be seen between the precipitation/temperature and potholes reported. A statistical correlation would need to be determined before further predictive analysis could be trusted. Pothole count and precipitation showed an exceptionally long right tailed distribution and unfortunately also showed no correlation, even when adjusted on a logarithmic scale. This was a bit surprising given how the Houston website states how pothole creation relies heavily on rainfall and temperature. Even with an analysis on a more granular level by analyzing the correlation on 96 Houston regions showed the same non-correlation. Because of this, future predictions could not be confidentially created but the exercise was done anyways to provide practice.



## Training, Modelling, Results

Preprocessing of the data was done by first utilizing the StandardScaler() utility class. The data was split into a training and test group with the test group being 30%. Three different models were run, and the mean absolute error calculated to determine the most optimal model. Results can be seen below. With a lower mean absolute error and standard deviation, the linear regression model was selected for use in predicting future pothole counts based on precipitation.

|  |  |  |
| --- | --- | --- |
| Model | Mean Absolute Error | Standard Deviation |
| Linear Regression | 9.43 | 0.27 |
| Random Forest | 9.96 | 0.33 |
| Logistic Regression | 10.15 | 0.43 |