HOUSTON POTHOLE ANALYSIS

7/4/2021

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

Houston Pothole Analysis

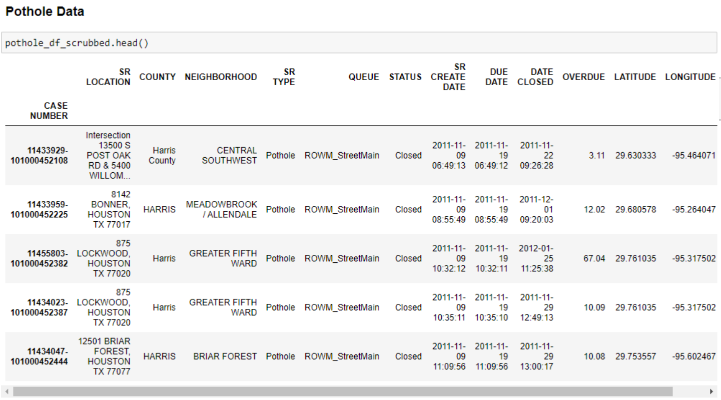
A LOOK INTO THE PREDICTABILITY OF POTHOLE CREATION BASED ON WEATHER CONDITIONS

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

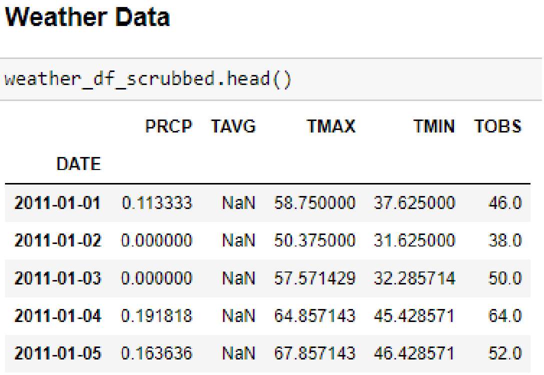
The pothole data is supplied by the City of Houston Pothole record site. Each year is split into a separate JSON file. Iterating through each page and appending to a pandas dataframe produced this data set. The weather data for Houston was taken from the National Oceanic and Atmospheric Administration. 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.



Pothole Data

```
pothole_df_scrubbed.head()
```

CASE NUMBER	SR LOCATION	COUNTY	NEIGHBORHOOD	SR TYPE	QUEUE	STATUS	SR CREATE DATE	DUE DATE	DATE CLOSED	OVERDUE	LATITUDE	LONGITUDE
11433929-101000452108	Intersection 13500 S POST OAK RD & 5400 WILLOM...	Harris County	CENTRAL SOUTHWEST	Pothole	ROWM_StreetMain	Closed	2011-11-09 06:49:13	2011-11-19 06:49:12	2011-11-22 09:26:28	3.11	29.630333	-95.464071
11433959-101000452225	8142 BONNER, HOUSTON TX 77017	HARRIS	MEADOWBROOK / ALLENDALE	Pothole	ROWM_StreetMain	Closed	2011-11-09 08:55:49	2011-11-19 08:55:49	2011-12-01 09:20:03	12.02	29.680578	-95.264047
11455803-101000452382	875 LOCKWOOD, HOUSTON TX 77020	Harris	GREATER FIFTH WARD	Pothole	ROWM_StreetMain	Closed	2011-11-09 10:32:12	2011-11-19 10:32:11	2012-01-25 11:25:38	67.04	29.761035	-95.317502
11434023-101000452387	875 LOCKWOOD, HOUSTON TX 77020	Harris	GREATER FIFTH WARD	Pothole	ROWM_StreetMain	Closed	2011-11-09 10:35:11	2011-11-19 10:35:10	2011-11-29 12:49:13	10.09	29.761035	-95.317502
11434047-101000452444	12501 BRIAR FOREST, HOUSTON TX 77077	HARRIS	BRIAR FOREST	Pothole	ROWM_StreetMain	Closed	2011-11-09 11:09:56	2011-11-19 11:09:56	2011-11-29 13:00:17	10.08	29.753557	-95.602467



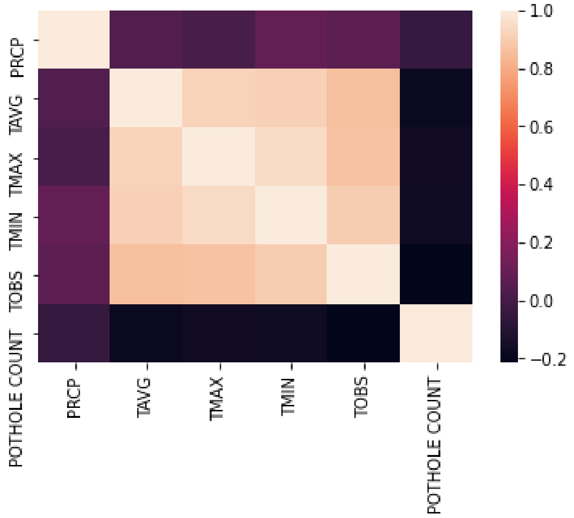
Weather Data

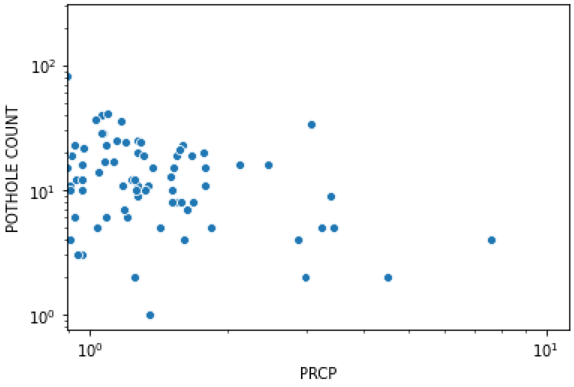
```
weather_df_scrubbed.head()
```

DATE	PRCP	TAVG	TMAX	TMIN	TOBS
2011-01-01	0.113333	NaN	58.750000	37.625000	46.0
2011-01-02	0.000000	NaN	50.375000	31.625000	38.0
2011-01-03	0.000000	NaN	57.571429	32.285714	50.0
2011-01-04	0.191818	NaN	64.857143	45.428571	64.0
2011-01-05	0.163636	NaN	67.857143	46.428571	52.0

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