PRABAV MURALI

Halifax, NS – B3L 4P2

prabavmurali1@gmail.com | www.linkedin.com/in/prabavmurali | https://github.com/prabav639 | +1 (782)-882-2715

Problem Statement:

The 311 NYC Service requests data has inbound call patterns available for all days of the year 2016. The goal is to predict the next day's inbound call patterns and find whether the weather parameters have an impact or causality on the inbound call patterns

About the data:

311 NYC Service requests dataset consists of all 311 calls made in the year 2016 in New York City starting from 00:00 to 23.59 every day. The columns present are unique_key, created_date, agency, complaint_type, location_type, incident_zip, borough. The information about the columns is as follows.

1. unique_key – Unique ID of every call received
2. created_date – Complaint created date and time
3. agency – Agency (or) department to which the call was forwarded or intended to
4. complaint_type – Type of complaint reported, for e.g., Heat/Hot water, Rodent etc.
5. location_type – Type of location for e.g., Residential Building
6. incident_zip – Zip code or Postal code from which the complaint was reported
7. borough – Borough is the area/town of a city. For e.g., Manhattan

Additionally, data on the weather of all cities were also provided. A weather dataset is a group of multiple datasets having information about city attributes, temperature, weather description, humidity, pressure, wind speed and wind directions. The information is available from 2012 to 2017 for multiple cities like New York City, Vancouver etc.,

Design Approach:

- Data Preprocessing & Data Cleaning:

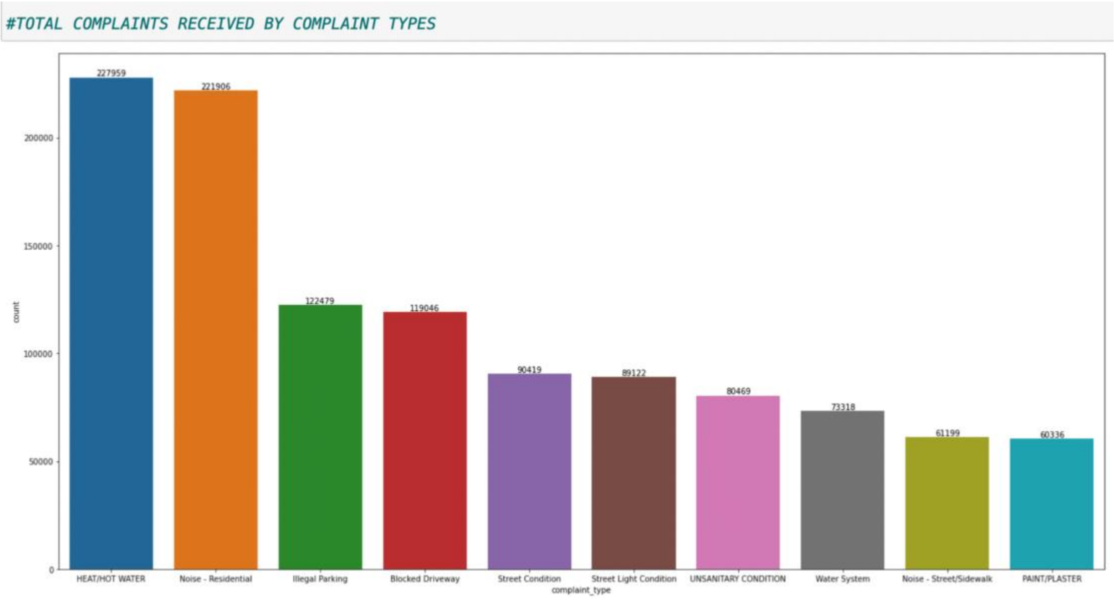
It is the most important step to make the data ready for analysis and prediction. The data preprocessing was performed to filter out the date, month and year from the timestamp which would be sufficient to forecast the next day’s incoming calls. Hence, the month, date and year are filtered.

After filtering, the data is aggregated based on the count of unique_key in each day from which we prepare the data for the year 2016 as 366 rows. The weather data is available from 2012-2017 in which the temperature is present as Kelvin. So, it is advisable to convert the temperature column from Kelvin to Celsius for better understanding. Also, the weather data is collected for the whole at different time periods, which must be aggregated for a single day.

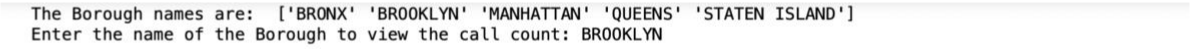
Finally, the data frames are merged with the 311 data which gives us aggregated data on the Count of calls each day, temperature, wind speed and humidity for 366 days of 2016.

After visualizing the results, it is important to prepare the data for modelling. If the data is continuous and has values of wide ranges, it is difficult for the model to learn from the data for future predictions. These are called outlier values. These outlier values are brought closer to the median or within the quartiles for feeding the model with consistent data to make better predictions.

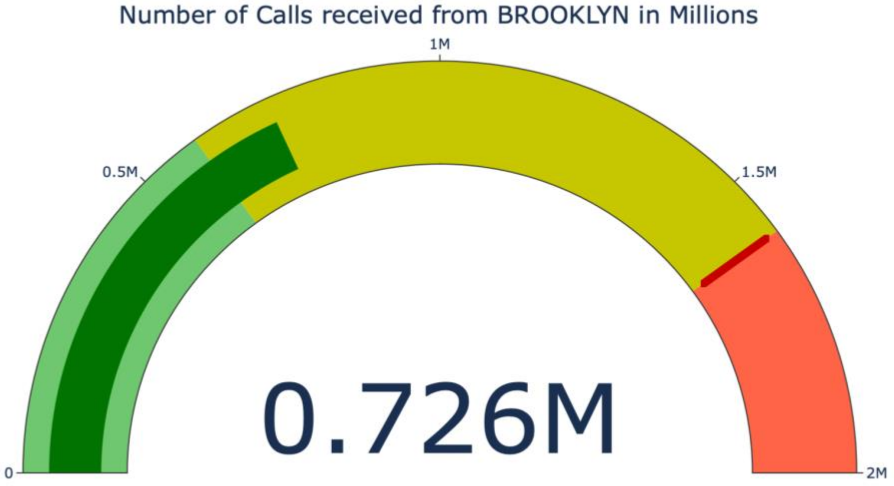
Important Visualizations:

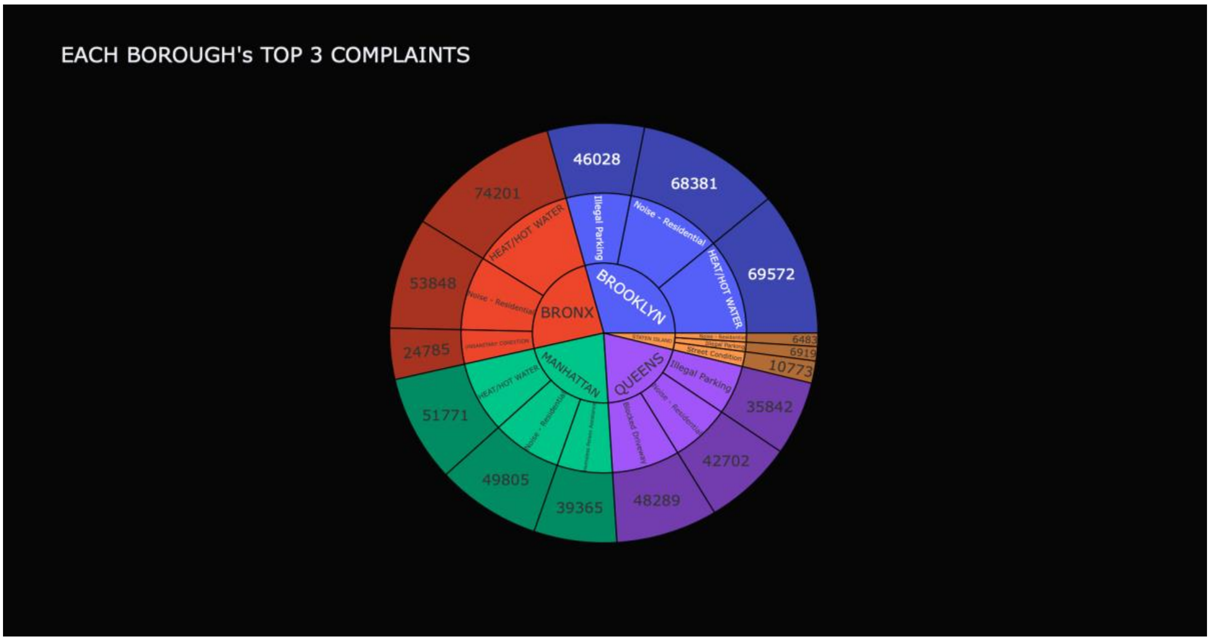


```
#TOTAL COMPLAINTS RECEIVED BY COMPLAINT TYPES
```

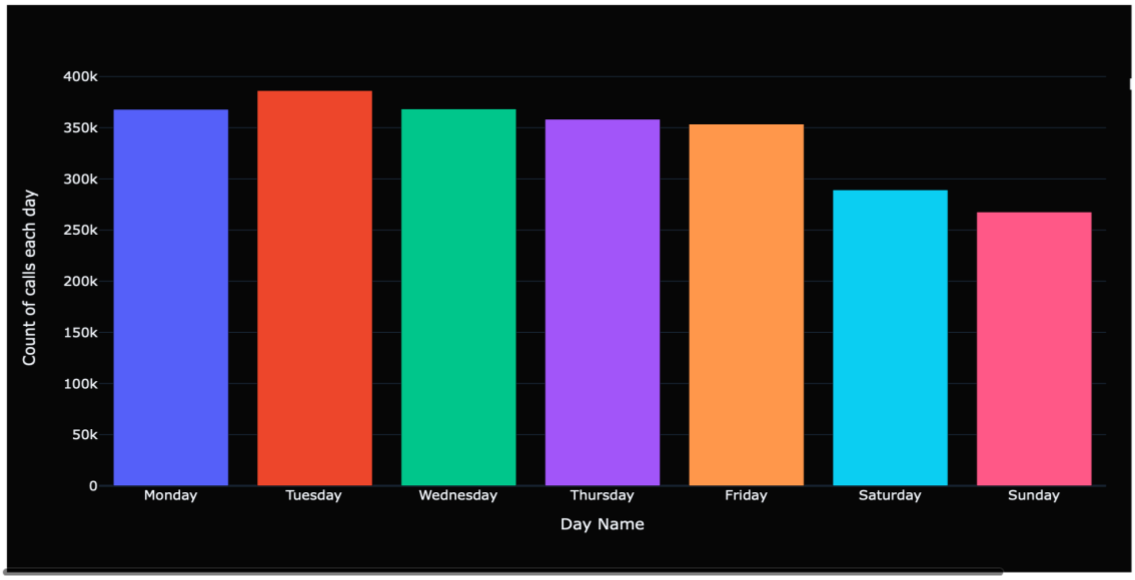


```
The Borough names are:  ['BRONX' 'BROOKLYN' 'MANHATTAN' 'QUEENS' 'STATEN ISLAND']
Enter the name of the Borough to view the call count: BROOKLYN
```

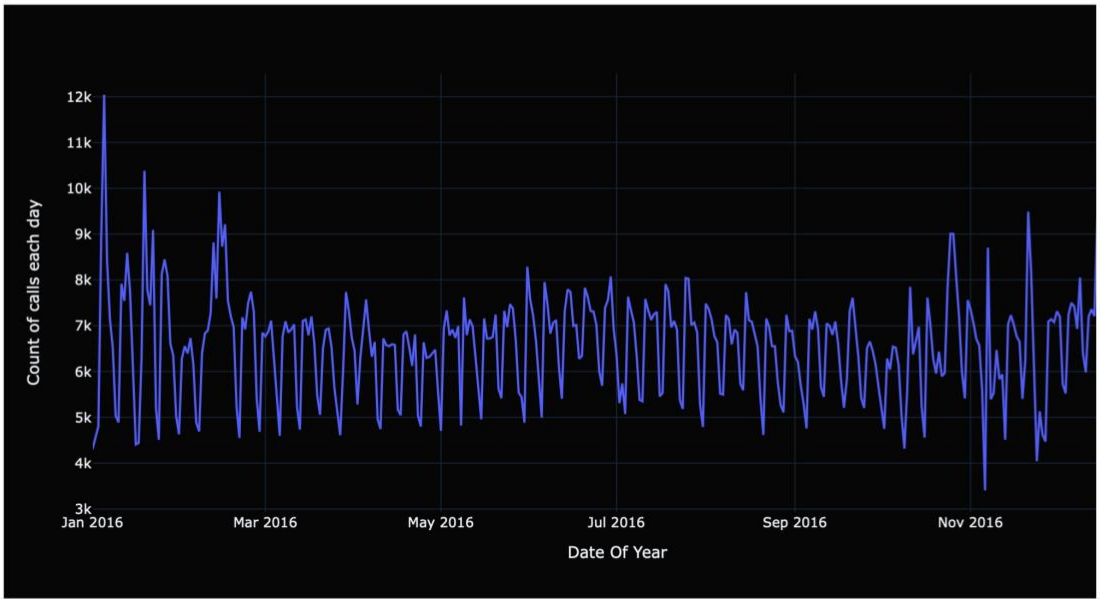


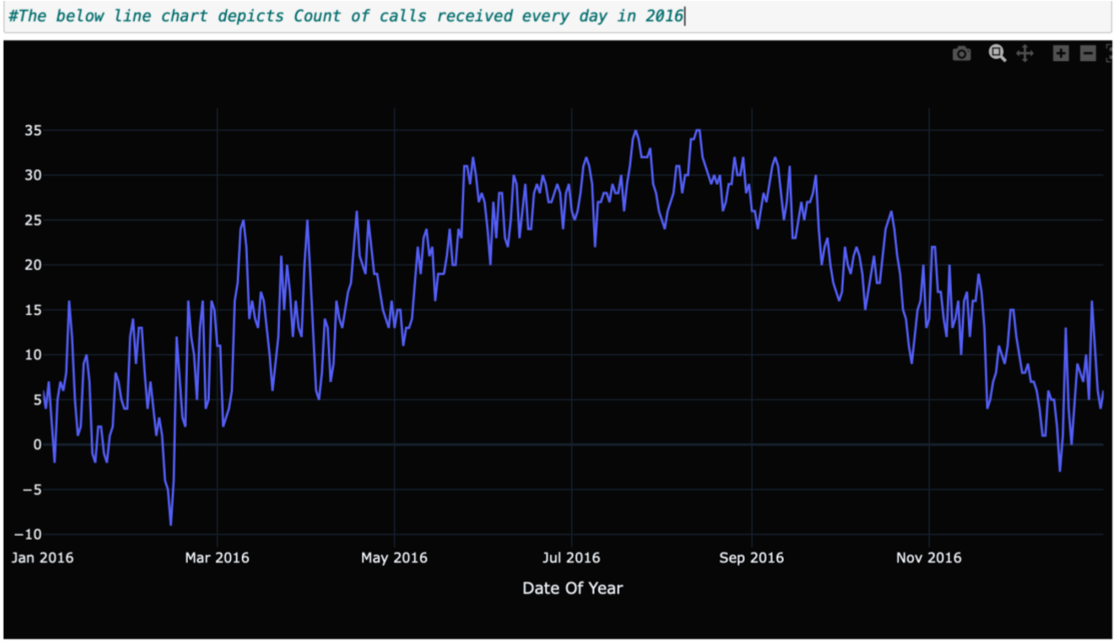


```
#From the below bar chart, we can infer that Tuesdays had the most number of calls in 2016
#----------------------------------------------------------------------------------------------------------
```



```
#The below line chart depicts Count of calls received every day in 2016
```





```
#The below line chart depicts Count of calls received every day in 2016
```

Statistical Tests:

To conduct an experiment on the causality of weather data on the count of calls received by 311 NYC each day, multiple statistical tests are performed. A test for correlation is performed to check whether there is any relationship between the two variables. A positive value of 0.8 and above shows a good positive correlation and a negative value of -0.8 and above shows a good negative correlation. In our case, we have obtained weak positive and negative correlations from which we can understand that the relationship between weather and number of calls is very weak.

To conduct a time series forecasting on the data, it is necessary to conduct a test called ADF (Augmented Dickey-Fuller) Test to check whether the data is stationary or seasonal. If the ADF's should_diff function returns True, the data is not stationary and needs differencing, whereas if it returns False, the data does not need differencing and is stationary. The test for checking the autocorrelation is performed to investigate the behaviour of present values based on past values. This is done using ACF and PACF plots.

Observed Patterns:

The observed patterns from the visualizations are given below.

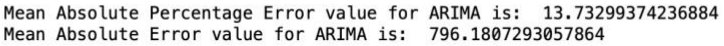
1. Heat/Hot water has been the most reported complaint.
2. Brooklyn borough reported the most complaints
3. Heat/ Hot water complaint was the most reported from Brooklyn
4. New York Police Department received the most complaints
5. January month received the maximum complaints

Model Building:

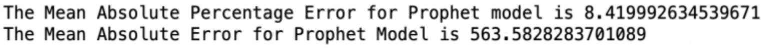
To predict/forecast a future value, a forecasting model becomes necessary. The forecasting model is selected on observed statistical results and values. The data is split into training and test model to help the model learn the data and forecast a future value. Models used for forecasting were: ARIMA (Auto regressor, Integrated, Moving Average) and Prophet. The ARIMA model requires three parameters for it to forecast the values which are (p,d,q) each for AR-I-MA. The parameters are selected using a function called Auto-Arima which selects the best parameters. The Prophet model is a model developed by Facebook for time series forecasting purposes. The model performs well on time series data and has performed well for the 311 NYC data giving good results.

Model Evaluation and Model Performance:

The ARIMA and Prophet models have performed better with the data provided. The model is evaluated using the metrics called MAPE (Mean Absolute Percentage Error) and MAE (Mean Absolute Error). The model scores for ARIMA and Prophet are shown below.

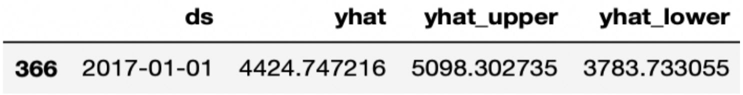


```
Mean Absolute Percentage Error value for ARIMA is:  13.73299374236884
Mean Absolute Error value for ARIMA is:  796.1807293057864
```



```
The Mean Absolute Percentage Error for Prophet model is 8.419992634539671
The Mean Absolute Error for Prophet Model is 563.5828283701089
```

The results a show good MAPE score with which the forecasted count for the next day (01-01-2017) would be.



	ds	yhat	yhat_upper	yhat_lower
366	2017-01-01	4424.747216	5098.302735	3783.733055

Where first column is 'ds' which is the period and second is 'yhat' which is the forecasted value for the period.
yhat_upper and yhat_lower are third and fourth columns which is the uncertainty values falling between confidence intervals

Future goals:

Building and fitting ARIMA and Prophet models produced good results and with more data, the models can produce better results giving accurate predictions. An enhancement to Prophet model would be to implement the Neural Prophet model which enhances the time series forecasting using deep learning techniques.

Conclusion:

To sum up, the 311 NYC service request data of the year 2016 had multiple complaints received from different boroughs of New York City and weather had no major impact on the number of calls received. The forecasted inbound calls to 311 NYC Service for the next day i.e., January 1, 2017, would be 4425 with an uncertainty ranging between 3784 to 5098.