NYC 311 complaints data analysis

VISHAL BHARTI
The Graduate Center, CUNY

***Abstract*—In this project, we used the NYC's 311 complaints dataset to do Big data analytics. The dataset used was from 2010 to 2016. Using Pyspark and visual analytic tool Tableau, we tried to study the peculiar trends in the complaints based on time, location, complaint types, etc. The Tableau worksheet can be accessed at link. All the visual analytics were done using this workbook. In the last section, we tried to build a prediction model that uses the Random Forest algorithm to predict the response time of complaints by using some important features in each record.**

***Index Terms*—311 data, NYC Complaints, Big Data Analysis**

I. Introduction

NYC's 311 complaints data service is New York city's main source for non-emergency complaints. NYC311 complaints data is open source and is currently available from January 2010 and is updated at regularly. The dataset used had around 14.2 million records and 53 columns.

For this project the data used was from January 2010 to 7^{th}December 2016. The total size of the dataset was around 9 Gigabytes. Given the large size, it was not possible to perform data analysis using the conventional techniques. The analysis was done using the clusters at 'NYU CUSP', which is a parallel computing large scale cluster center.

Using visual analytics and Pyspark, some hypothesis were tested. Following were some of the proposed hypothesis:

- Complaints will be vary for each location.
- Complaints will vary based on time of day, the day of week and time of the year.
- Noise complaints will be higher in residential areas.
- Parking violations complaints will increase during weekends.
- Street light conditions complaints will be higher during night time.
- Correlation between the number of complaints and response time.

II. Proposed Methodology

To study the trends in the dataset, the best approach is visual analytics, but given the size of the dataset conducting visual analysis on such large dataset is not feasible. The approach used here was to use data filtering to extract information about some essential features and then use the filtered data to conduct visual analytics.

A. Data filtering

The important feature for each record were "Agency", "Agency Name", "Created Date", "Closed Date", "Descriptor", "Incident Zip", "Borough", "X Coordinate", "Y Coordinate". However, for different analysis different subsets of these features were extracted.

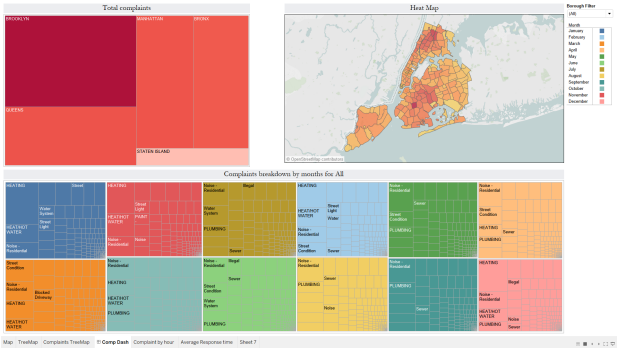

Fig. 1. A snapshot of the dashboard

The data had many missing and discrepant records. The first step was to drop out all those records or make appropriate corrections. For example, many records had incorrect Borough names, we assumed that the Incident Zip was correct in each record and the Borough was assigned was using the Incident Zip. Additionally all records that had missing entries were filtered.

B. Visual Analytics

To perform visual analytics on the filtered data, Tableau was used. The first important analysis was to study how the complaints were distributed based on the location, i.e., by Borough or Zip Code, or how frequent was a particular complaint in a given location. The next analysis was to study the trends in complaints with time, i.e. with month, day or hour of the day.

For both these analysis a single filtered dataset was obtained by using Pyspark. The key here was a tuple of "Borough", "Agency Name", "Complaint Type", "Created Month", "Created Year", "Created Day" and "Incident Zip". "Created Month", "Created Year" and "Created Day" was extracted from the "Created Date" feature. To study the hourly trend of complaints, another filtered dataset was obtained by filtering based on a key, which was a tuple of "Borough", "Agency Name" and "Created Hour". The "Created Hour" was extracted from the "Created Date" field.

Finally to compute the average response times for each "Agency", the difference in seconds was computed between the "Created Date" and "Closed Date". Then using "Agency" as the key the average response time was computed. All these tasks were incorporated into a single Tableau workbook.

III. NYC311 Data Visual Analytics

Using the Tableau workbook built from the filtered dataset, that came from the Pyspark jobs runs on the CUSP clusters, we looked at various trends in the 311 data.

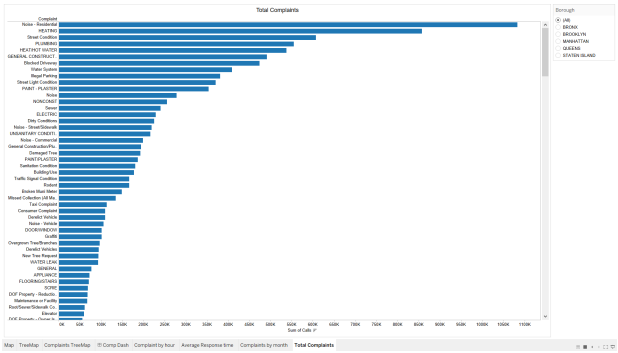

Fig. 2. Top Complaints-All Boroughs

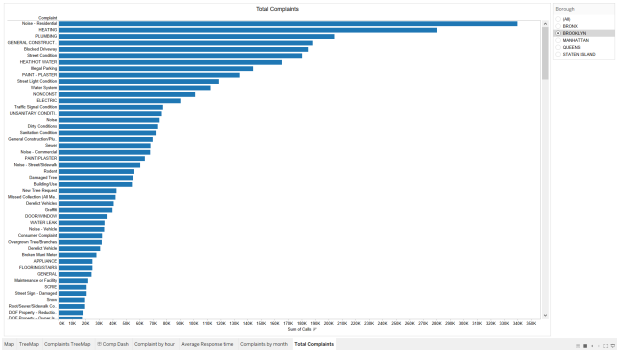

Fig. 3. Top Complaints-Brooklyn

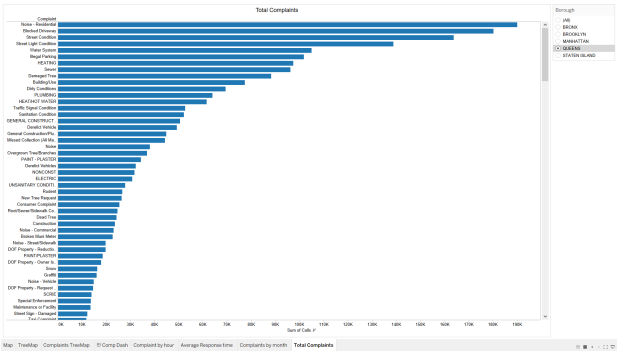

Fig. 4. Top Complaints-Queens

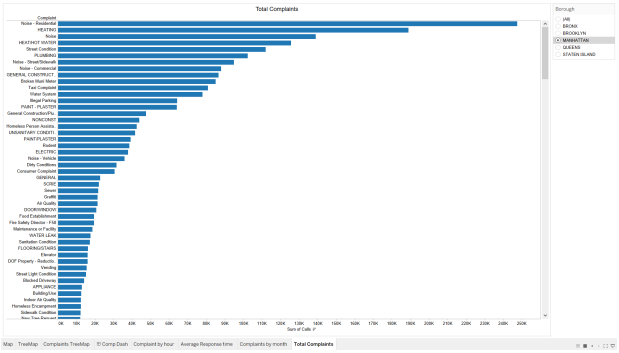

Fig. 5. Top Complaints-Manhattan

A. Complaint trends based on location

Figure 1 shows a snapshot of the dashboard that is built into the workbook.The tree map on the top left shows that Brooklyn received the most number of complaint calls, followed by Queens, Manhattan, Bronx and Staten Island. After filtering the data had around 12.7 million complaint records. Out of these 4.16 million complaints were from Brooklyn, 2.67 million were from Queens, 2.65 million were from Manhattan, 2.55 million were from Bronx and 0.67 million were from Staten Island. The heat map on top right shows that the two areas(having the darkest red color) with zip '11226'(0.25 million) in Brooklyn and '10467'(0.22 million) in Bronx received the most complaint calls. The most frequent complaints in these areas were "Heating" and "Noise-Residential".

Using the sheet "Total complaints" in the workbook, the most frequent complaints can be listed. In the entire dataset, the top three complaints were "Noise-Residential", "Heating" and "Street Condition". For Brooklyn the top three complaints were "Noise-Residential", "Heating" and "Plumbing". For Queens, "Noise-Residential", "Blocked Driveway" and "Street condition" were the top three complaints. For Manhattan, "Noise-Residential", "Heating" and "Noise" were top three complaints. For Bronx, "Heating", "Noise-Residential" and "Heat/Hot water" were top three complaints and for Staten Island, "Street Condition", "Street lighting conditions" and "Water system" were top three complaints. Figure 2-7 show the snapshot of these.

B. Complaint trends based on time

Figure 8 shows the complaints(top 20) trends across months. As we can see, heating complaints are highest during cold months, while for all other months the 'Noise-Residential' complaint is the most frequent one. Some complaints, such as "Noise-Residential", "Blocked Driveway", "Plumbing", etc. follow a regular trend throughout the year.

The sheet "complaint by hour" shows trend of complaints by hour of the day. Figure 9 shows the complaint trends for all complaint by hour of the day. Most complaints are made at midnight, which seems counter-intuitive. Figure 10,11,12 and 13 show the trends for "Illegal parking", "Noise-residential", "Illegal parking" and "Blocked Driveway" complaints. "Blocked Driveway" and "Illegal parking complaints" show very similar trends, with maximum complaints in early morning and late night. "Noise-residential" complaints show a peak around midnight. "Violation of parking rules" complaints are maximum during the day time.

The sheet "Complaints by Day" shows number of complaints for days of the week. Figure 14 shows number of complaints on each day of the week for all complaints. More calls were made in weekdays as compared to weekends. Figure 15 and 16 show that the complaint calls for "Noise" and "Violation of parking rules" show a different trend with more complaint calls in the weekends. For "Construction" complaints figure 17 shows that very few complaints were made in the weekends.

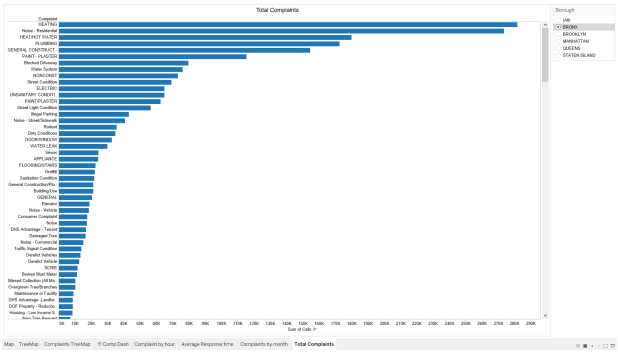

Fig. 6. Top Complaints-Bronx

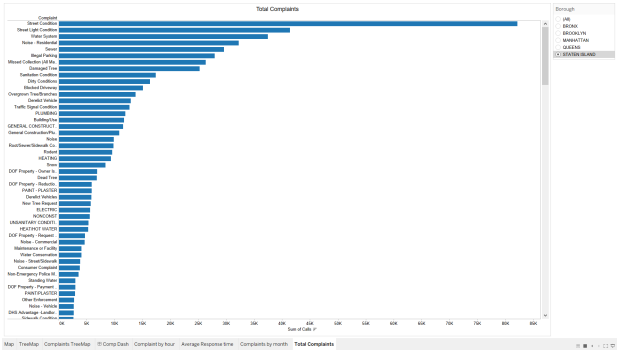

Fig. 7. Top Complaints-Staten Island

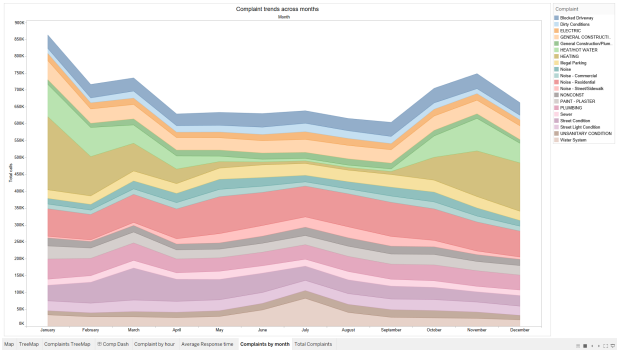

Fig. 8. Complaint trends across months

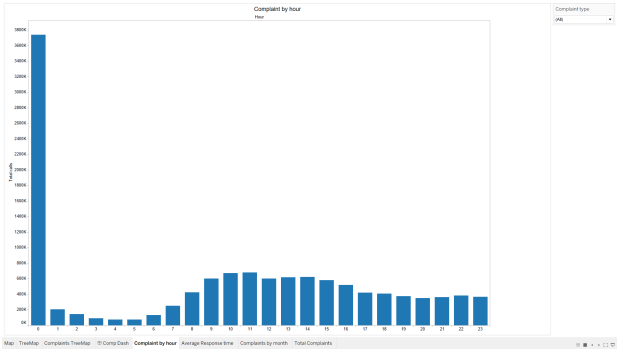

Fig. 9. Complaint trend with hour of day

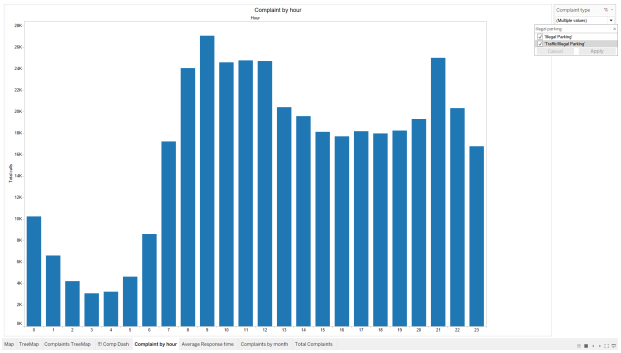

Fig. 10. “Illegal parking” complaint trend with hour of day

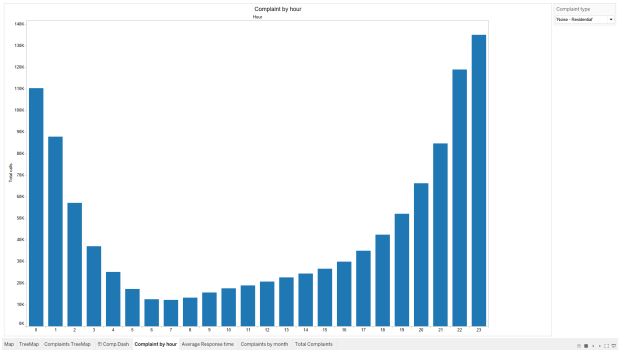

Fig. 11. “Noise-Residential” complaint trend with hour of day

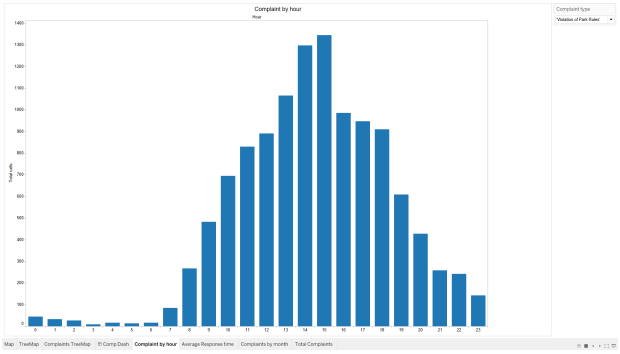

Fig. 12. “Violation of parking” complaint trend with hour of day

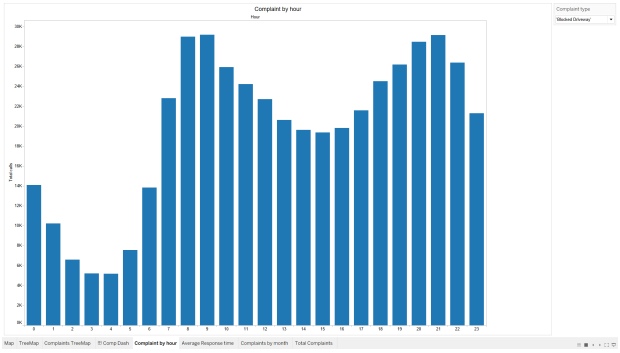

Fig. 13. “Blocked Driveway” complaint trend with hour of day

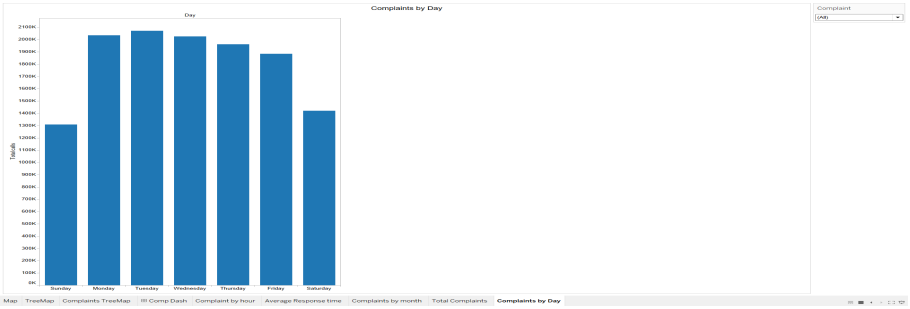

Fig. 14. All complaints by day of week

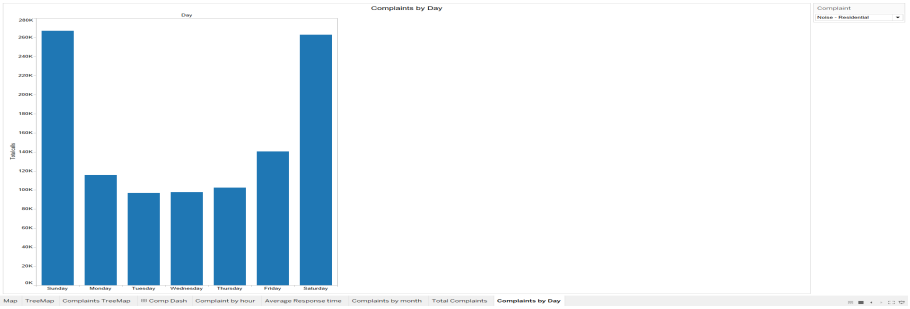

Fig. 15. "Noise-Residential" complaints by day of week

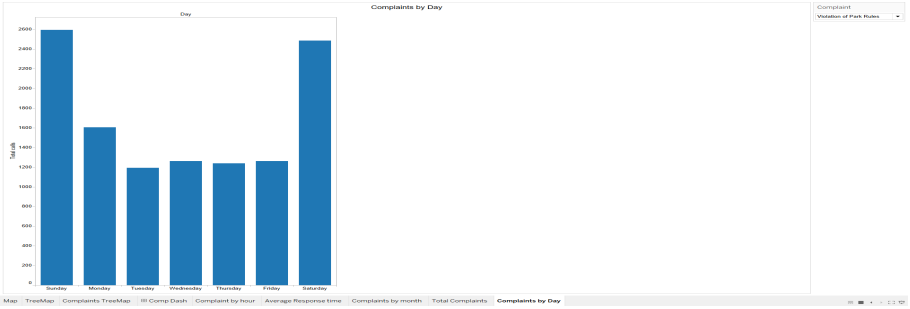

Fig. 16. "Violation of parking" complaints by day of week

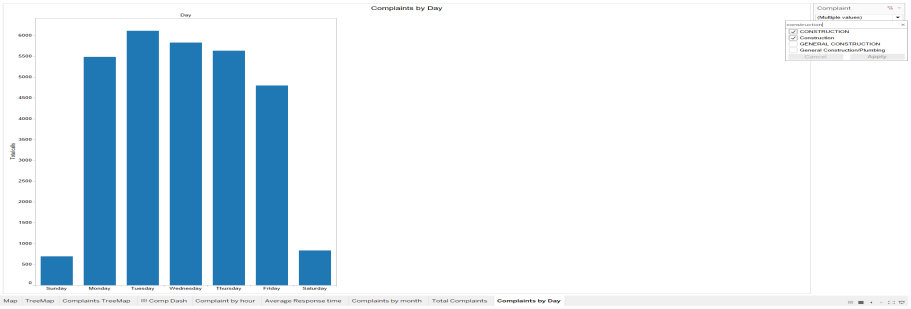

Fig. 17. "Noise-Residential" complaints by day of week

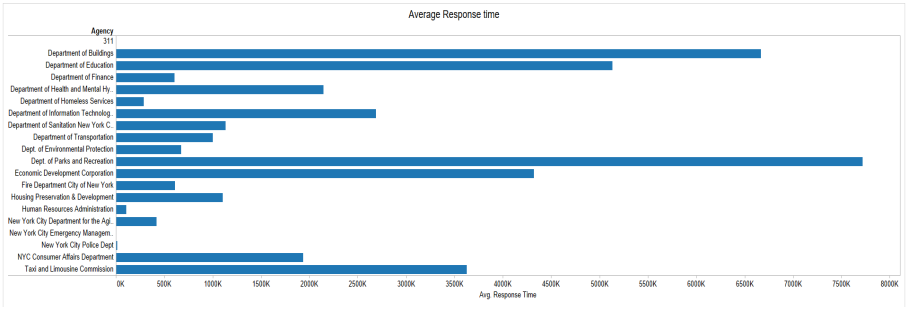

Fig. 18. Average response time for agencies

IV. Predicting response time for a complaint

Figure 18 shows response times for different agencies, DEP(Dept. of Environmental Protection) has the worst average response time, followed by DOB(Dept. of Building) and DOE(Dept. of Education).

The response time for different agencies were very different. Using some important features in the 311 data and the Random Forest regression algorithm, we will build a model that will predict the response time for a complaint. For training and testing we use 70:30 split, i.e. 70% of the data to train the Random Forest algorithm and 30% to test the built model.

The features that were used include : “Agency”, “Agency Name”, “Complaint Type”, “Descriptor”, “Incident Zip”, “Borough”, “X-coordinate” and “Y-coordinate”. The feature “Agency Name” has 1664 different values and hence the max bins parameter was set as 1664, the depth was 10 and number of trees were taken as 3,7, and 16.

The results was computed in terms of RMSE(root mean square error). For 3 trees the RMSE was around 1560 hours (65 days), for 4 trees 1791 hours (74.5 days) and for 16 tress 1782 hours (74.3 days).

V. Conclusion

In this project we tried to study complaint trends in NYC311 complaints dataset. The trends varied with location and time. The most common complaints were “Noise-Residential” and “Heating”. Brooklyn was the borough that received most complaint calls. It was also noted that the complaint calls varied for months, day and time of the day. For each agency the average response time was and computed. The response time varied significantly for each agency. Most of the preliminary hypothesis were tested positive in the analysis. The “street light” complaints hypothesis and the correlation hypothesis, however, were tested false. The correlation between the “Total complaints” and “Response time” was negligible (-0.015).

Finally, Random Forest algorithm was used to predict the response time for complaints, using 8 features in the dataset. The error associated was significantly high, because of the highly variable nature of response times. For future work, it would be interesting to consider different factors that might influence the response time.

References

[1] Link to NYC311 dataset