NYC Complaints DataSet Analysis

Authors
Aparajita Choudhury - ac5901
Akshay Kalia - ak5641
Storm Avery Ross - sar516

Abstract: *Launched on March 9, 2003, NYC311 provides services to New York residents - 24 hours a day, 7 days a week in nearly 180 languages. On an average, it receives more than 50,000 calls, texts and emails combined daily. This study explores the relationship amongst different columns, such as complaint types, time to close a complaint, etc and how they are related with borough, time and agency. This study also aims to analyze the 311 complaints dataset and find causations of different anomalies found in data. A copy of the report is available at:*
https://docs.google.com/a/nyu.edu/document/d/1CZEp-McMnW7ZK9JqylFL_TwXBL46ZXatwkkljy86KBY/edit?usp=sharing
The code, results and plots are available on github at:
https://github.com/aparajita2930/NYC_Complaints_Analysis

1. Introduction

Since its launch, NYC311 has received more than 158 million calls and has been a clearinghouse for all things New York City government, providing information on more than 4,000 topics, routing details to the appropriate City agencies and providing customers with service request numbers for use in tracking the progress of their inquiry.

In doing so, vast amount of data is collected, which can be used to drive valuable insights. In this project we summarize the data across all columns. With over 15M rows, the distributed framework that Big Data technologies provide prove to be beneficial to analyze the data. We have use the NYU Hadoop cluster to perform all the processing, analysis and aggregating our dataset.

2. Data Summary and Quality

The first step that has been performed to set up the data environment for the project was to join the two files - one for the year 2009 and the other for the year 2010-2017 into a single file using the script https://github.com/aparajita2930/NYC_Complaints_Analysis/blob/master/src/join_csvs.py (Spark). This script saves the single merged file into HDFS.

After going through the 311 dataset for 2009 to 2017 and assigning a base, a semantic, and a validity type we derived a summary of the data. A few of the things that we noted are:

- No Invalid or Null values in two columns: “Created Date” and “Agency”
- Two invalid values in “Unique Key” column - two of the keys could be tagged as invalid as they are duplicates and hence violate the functional dependency requirement

- The column “Resolution Action Updated Date” has the highest number of invalid records with dates lying outside of the range or even dates that are before the created date or after the closed date for the complaint
- Data referring to the same place or thing represented in slightly different ways from each other - eg: lowercase and uppercase, punctuations, etc. - To deal with this issue, the the data in each column has been standardized and dictionaries and lists of the domain of the particular columns have been maintained wherever possible to perform lookups
- A few of the zips are invalid meaning that they are outside of the areas of NYC

The details of each of the 52 columns in the dataset is as below. It shows the number of Valid, Invalid and Null elements in the particular column as well as the number of elements having a particular base data type. The semantic type column below shows the count of each semantic type of elements in a column, separated by ‘|’. Obtaining the below information has been a two step process:

- First, the scripts whose names begin with 1 to 52 in the github repository folder: https://github.com/aparajita2930/NYC_Complaints_Analysis/tree/master/src/column_summary were executed to generate output.
- These outputs were fed to the script https://github.com/aparajita2930/NYC_Complaints_Analysis/blob/master/src/column_summary/0_column_summary.py to generate the summary.

		Validity			Datatype				
No.	Column Name	VALID	INVALID	NULL	INT	DECIMAL	DATETIME	TEXT	Semantic Type
1	Unique Key	15405235	2	0	15405237	0	0	0	Key:15405237
2	Created Date	15405239	0	0	0	0	15405239	0	Created_Date:15405239
3	Closed Date	9659702	5201708	543829	0	0	14867225	538014	Text:5745537 | Closed_Date:9659702
4	Agency	15405239	0	0	0	0	0	15405239	Agency:15405239
5	Agency Name	15388894	16345	0	0	0	0	15405239	None:16345 | Agency_Name:15388894
6	Complaint Type	15404536	702	1	0	0	0	15405239	None:702 | Complaint_Type:15404537
7	Descriptor	15043169	219044	143026	0	0	0	15405239	Descriptor:15186195 | None:219044
8	Location Type	11310491	29468	4065280	0	0	0	15405239	None:29468 | Loc_Type:15375771
9	Incident Zip	14222212	90824	1092203	14311176	0	0	1094063	Incident_Zip:15403800 |None:1439
10	Incident Address	11744922	235008	3425309	164	0	0	15405075	Address:15170231 | None:235008
11	Street Name	11147745	831105	3426389	989	1	0	15404249	Street:14574134 | None:831105
12	Cross Street 1	9522359	1361938	4520942	50	0	0	15405189	Street:14043301 | None:1361938
13	Cross Street 2	9507944	1312925	4584370	40	0	0	15405199	Street:14092314 | None:1312925
14	Intersection Street 1	1871099	573071	12961069	322	0	0	15404917	Street:14832168 | None:573071
15	Intersection Street 2	1963690	477528	12964021	53	0	0	15405186	Street:14927711 | None:477528
16	Address Type	14687020	0	718219	0	0	0	15405239	Address_Type:15405239
17	City	3134816	11183846	1086577	25	0	0	15405214	City:4221393 | None:11183846
18	Landmark	8490	0	15396749	0	0	0	15405239	None:15396749 | Landmark:8490
19	Facility Type	3606883	0	11798356	0	0	0	15405239	Facility_Type:15405239
20	Status	15405131	0	108	0	0	0	15405239	Status:15405239
21	Due Date	6065893	10381	9328965	0	0	6076274	9328965	Text:9339346 | Due_Date:6065893
22	Resolution Action Updated Date	379163	14709076	317000	0	0	15088239	317000	Text:15026076 | Res_Date:379163
23	Community Board	12754172	16458	2634609	0	0	0	15405239	None:16458 | Community_Board:15388781
24	Borough	13886183	0	1519056	0	0	0	15405239	Borough:15405239
25	X Coordinate (State Plane)	13766562	0	1638677	13766562	0	0	1638677	State_Plane_Cord:13766559 | Text:1638680
26	Y Coordinate (State Plane)	13766562	0	1638677	13766562	0	0	1638677	State_Plane_Cord:13766559 | Text:1638680
27	Park Facility Name	24101	69387	15311751	0	0	0	15405239	None:69387 | Park:15335852
28	Park Borough	13886183	0	1519056	0	0	0	15405239	Park_Borough:15405239
29	School Name	15083	78405	15311751	0	0	0	15405239	School:15326834 | None:78405
30	School Number	15221	75080	15314938	15229	0	0	15390010	None:75080 | School_Number:15330159
31	School Region	15215	6	15390018	0	0	0	15405239	School_Region:15405233 | None:6
32	School Code	15083	146	15390010	0	0	0	15405239	None:146 | School_Code:15405093
33	School Phone Number	76200	0	15329039	76200	0	0	15329039	School_Phone_Number:15405239
34	School Address	90944	2536	15311759	0	0	0	15405239	Address:15402703 | None:2536
35	School City	19311	74177	15311751	0	0	0	15405239	None:74177 | School_City:15331062
36	School State	93488	0	15311751	0	0	0	15405239	School_State:15405239
37	School Zip	93387	101	15311751	93487	0	0	15311752	School_Zip:15405238 | None:1
38	School Not Found	5921449	0	9483790	0	0	0	15405239	School_Not_Found_Indicator:15405239
39	School or Citywide Complaint	3992	0	15401247	0	0	0	15405239	School_Or_Citywide_Complaint Indicator:15405239
40	Vehicle Type	7974	0	15397265	0	0	0	15405239	Vehicle_Type:15405239
41	Taxi Company Borough	13350	0	15391889	0	0	0	15405239	Taxi_Company_Borough:15405239
42	Taxi Pick Up Location	123079	0	15282160	0	0	0	15405239	Taxi_Pickup_Loc:15405239
43	Bridge Highway Name	42651	0	15362588	0	0	0	15405239	Bridge_Highway_Name:15405239
44	Bridge Highway Direction	42589	0	15362650	0	0	0	15405239	Bridge_Highway_Direction:15405239
45	Road Ramp	42245	0	15362994	0	0	0	15405239	Road_Map:15405239
46	Bridge Highway Segment	18716	29880	15356643	0	0	0	15405239	None:29880 | Bridge_Highway_Segment:15375359
47	Garage Lot Name	5086	0	15400153	0	0	0	15405239	Garage_Name:15405239
48	Ferry Direction	3562	0	15401677	0	0	0	15405239	Ferry_Direction:15405239
49	Ferry Terminal Name	37	10227	15394975	7	0	0	15405232	None:10227 | Ferry_Terminal_Name:15395012
50	Latitude	13766559	3	1638677	0	13766562	0	1638677	None:15405239
51	Longitude	13766499	63	1638677	60	13766502	0	1638677	None:15405239
52	Location	13766559	3	1638677	0	0	0	15405239	Geo_Code:13766562 | Text:1638677

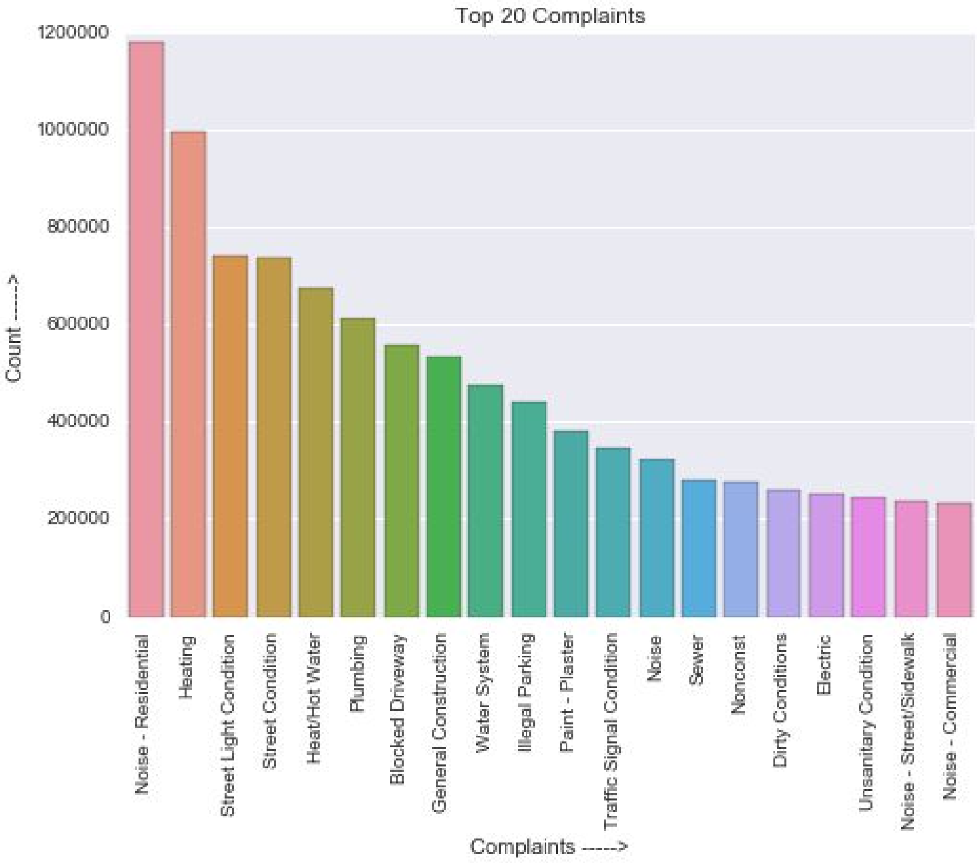
3. Complaint Trends

(All the graphs in this section have been created using the script: https://github.com/aparajita2930/NYC_Complaints_Analysis/blob/master/results/plots/visualizations.ipynb)
In this section, the data has been summarized along various dimensions - like hour of the day, year or day of week as well as by location, city or borough.

The data to plot the graphs in section 3.1 to 3.6 have been generated using the script: https://github.com/aparajita2930/NYC_Complaints_Analysis/blob/master/use_cases/complaint_type_distribution.py.

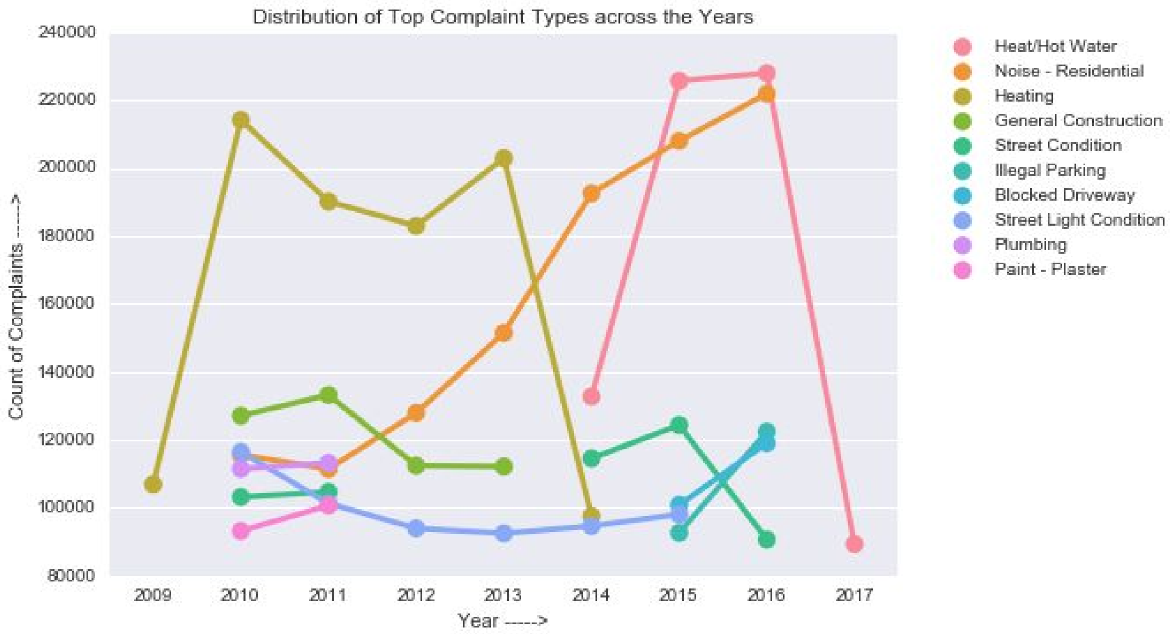
3.1. Distribution of Top 20 Complaints

Though 311 records complaints for more than 4000 categories, Noise –Residential (7.66%), Heating (6.46%), Street Light Conditions (4.8%) and Street Conditions (4.8%) are the most complained about.



3.2. Distribution of Top Complaint Types across the Years

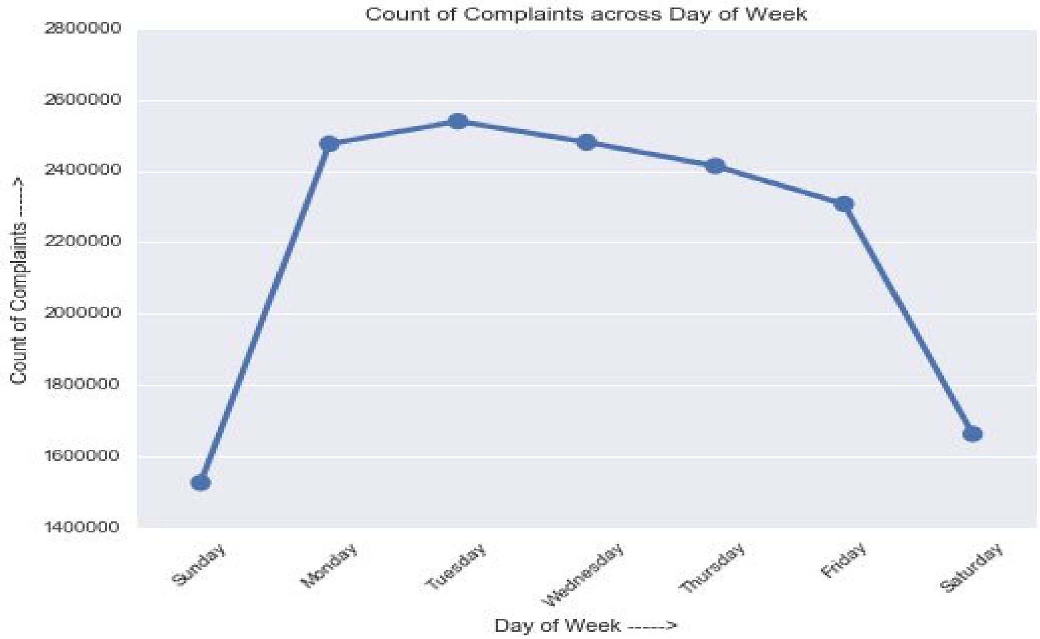
Over the years, some trends can be observed. Each year, Noise and Heating were most complained about.



It is also observed from data, that label 'Heating' was changed to 'Heat/Hot Water' in 2014.

3.3. Distribution of Complaints across Day of Week

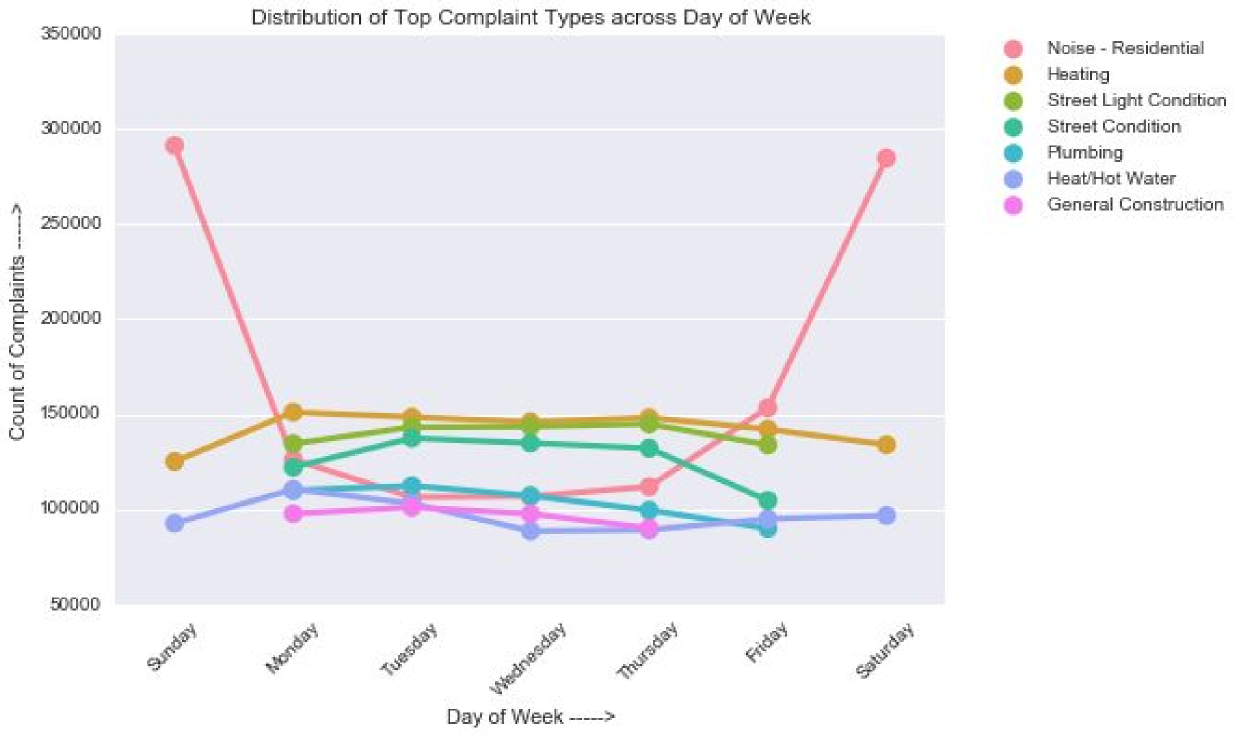
Maximum number of complaints, 16.48% of the total complaints, were registered on a Tuesday. Least number of complaints were registered on Saturday (10.79%) and Sunday (9.90 %).



3.4. Distribution of Top Complaint Types across Day of Week

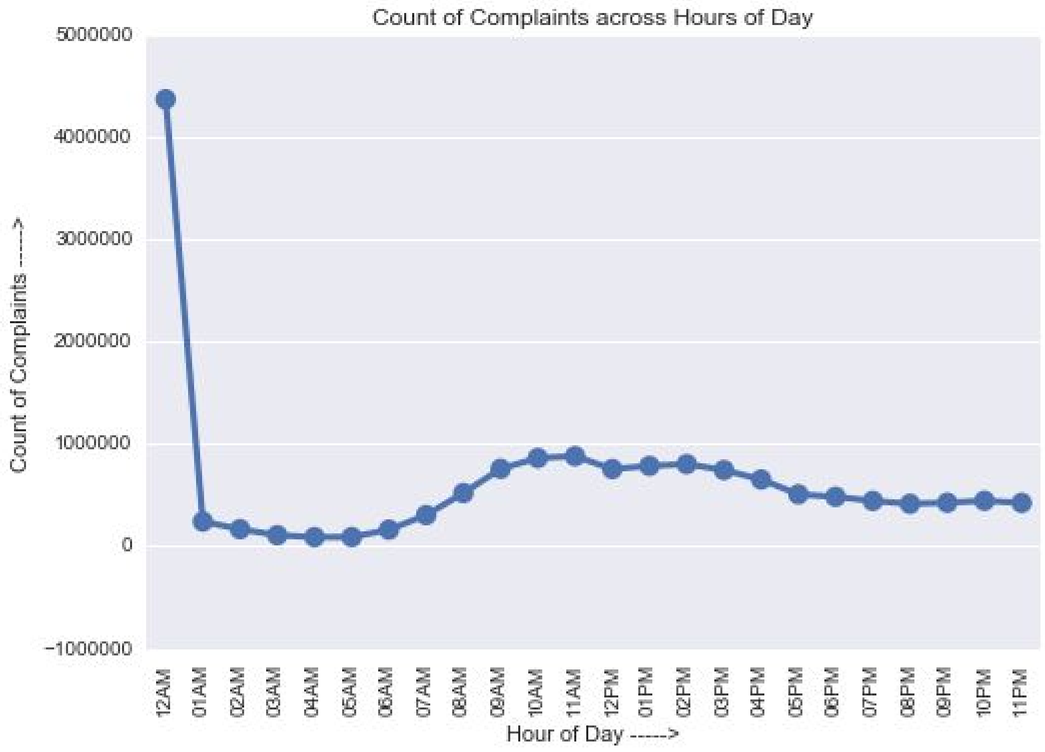
As expected, Noise from Residential buildings consists of maximum number of 311 complaints on Friday and the weekend as people often party over the weekend. Noise complaints are followed by Heating and Hot Water complaints.

On the weekdays, Heating issues form the maximum number of 311 complaints. Street Light Conditions and Street Conditions are 2nd and 3rd most complained about on the weekdays. This might be due to the fact that residents drive to their work over the weekdays and might observe bad street conditions.



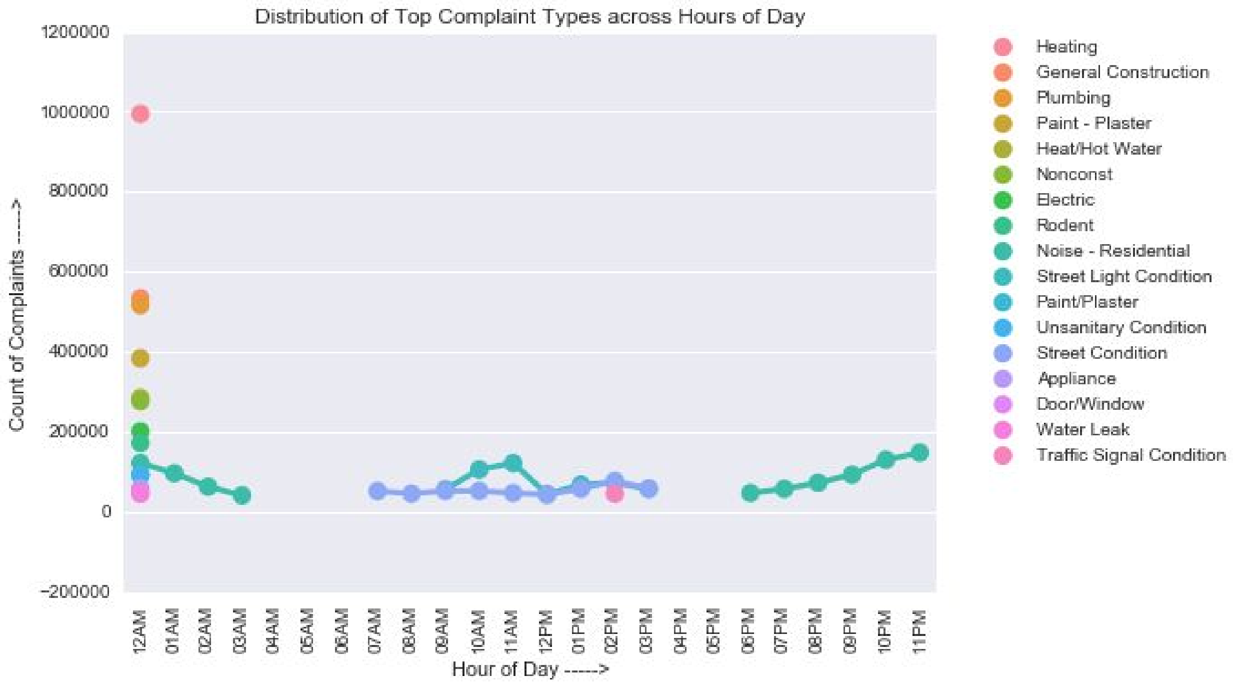
3.5. Distribution of Complaints across Hour of Day

Maximum number of complaints, about 28.34% of total, were registered at midnight. A marginal increase in 311 Complaints is also observed from 7AM to 6PM.



3.6. Distribution of Top Complaint Types across Hours of Day

Though we expected Noise complaints to be the dominant complaints at midnight, we observed that heating and plumbing complaints were the most reported.

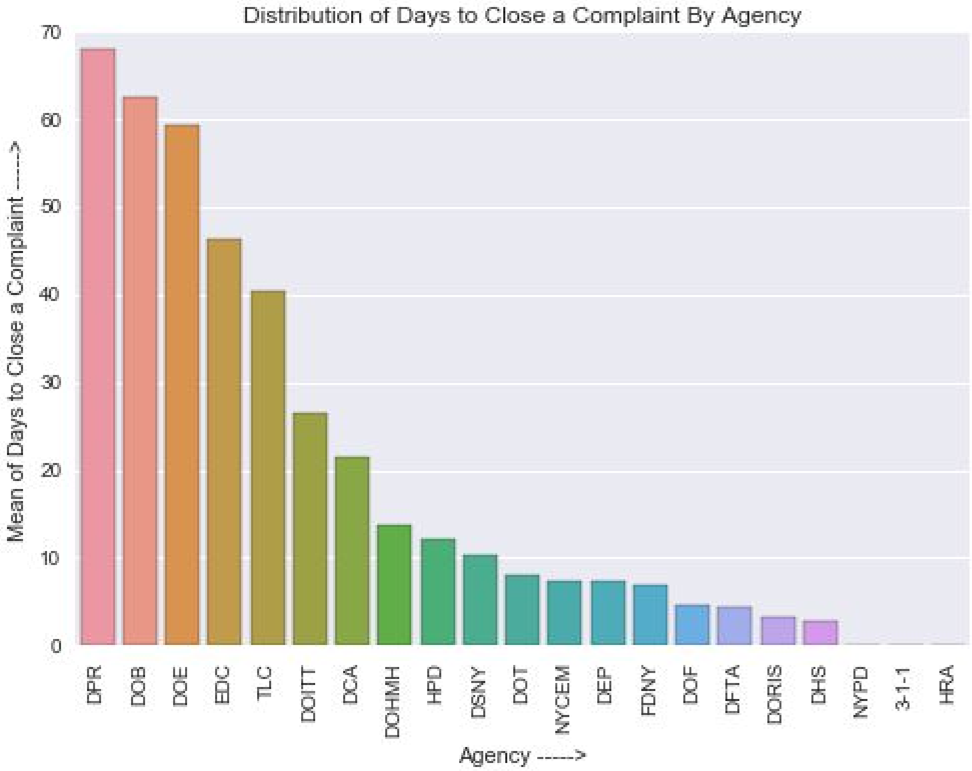


We can also see an increase in 311 complaints for street conditions during the day as observed previously.

3.7. Distribution of Mean Days to Close a Complaint by Agency

The data to plot the graph has been generated using the script: https://github.com/aparajita2930/NYC_Complaints_Analysis/blob/master/use_cases/closing_time_distribution.py.

We observe that the average number of days to close a complaint varies highly across the different agencies with agency DPR taking the longest amount of time to close a complaint and with HRA taking the shortest time.

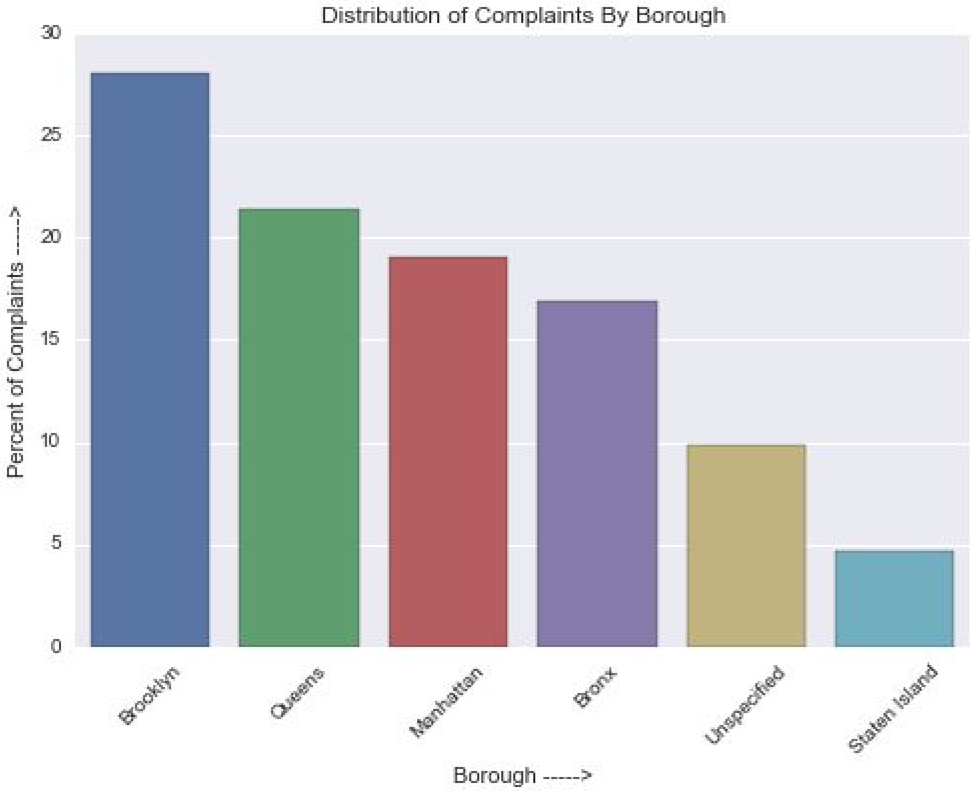


However, the short time to close complaints by NYPD, 3-1-1 or HRA as observed above can also be due to the effect of lack of valid closing times on many records.

3.8. Distribution of Complaints by Borough

The data to plot the graph has been generated using the script: https://github.com/aparajita2930/NYC_Complaints_Analysis/blob/master/use_cases/comp_count.py.

We observe that Brooklyn has the highest number of complaints and Staten Island the lowest. This can be due to the fact that Brooklyn is the highest populated borough in NYC and Staten Island, the least populated one [6].



We also see that about 10% of the complaints do not have the borough information associated with them.

4. Individual Contributions

Each of the us in the team have worked on brainstorming the use cases that would give us valuable insights from the data. We divided the task amongst ourselves with each of us responsible for a number of scripts.

5. Conclusion

We observed various data quality issues - like missing data, same data represented in various forms, invalid data, etc. With over 15M rows, Big Data technologies proved to be beneficial in analyzing and aggregating the data over various dimensions. Also, we could identify various trends in the complaints, the way people complaint and how complaints are dealt with.

6. References

1. https://data.cityofnewyork.us/City-Government/Street-name-Dictionary/w4v2-rv6b/data?pane=manage
2. https://www.mapdevelopers.com/geocode_bounding_box.php
3. http://schools.nyc.gov/schoolsearch/
4. https://en.wikipedia.org/wiki/List_of_New_York_City_parks
5. https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City
6. https://en.wikipedia.org/wiki/Borough_(New_York_City)