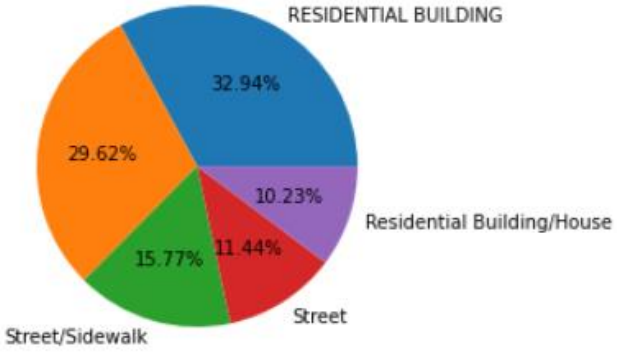
New York 311 Data Challenge

Data Visualizations

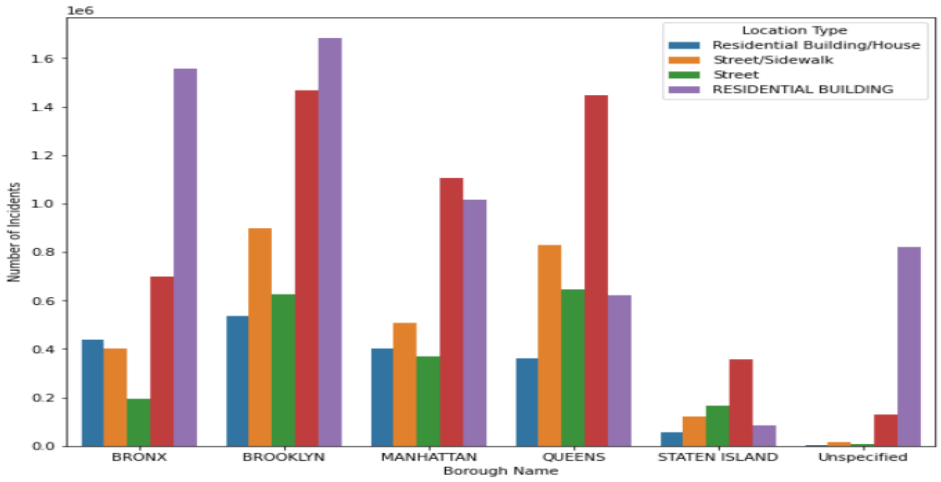
- Since we know there are 27Million Records creating a visualization of all the records will lead to Unimpressive Visualizations. Hence for this purpose I have shown TOP 5 OR TOP 10 records for each attribute.
- This gives a good visualization diagram and conveys sound information.

- First Visualization I have shown is a PIE CHART of Top 5 Location_Type, this includes “Residential Building” , “Street/Sidewalk” , “Residential Building/House” , “Street” and an empty name.



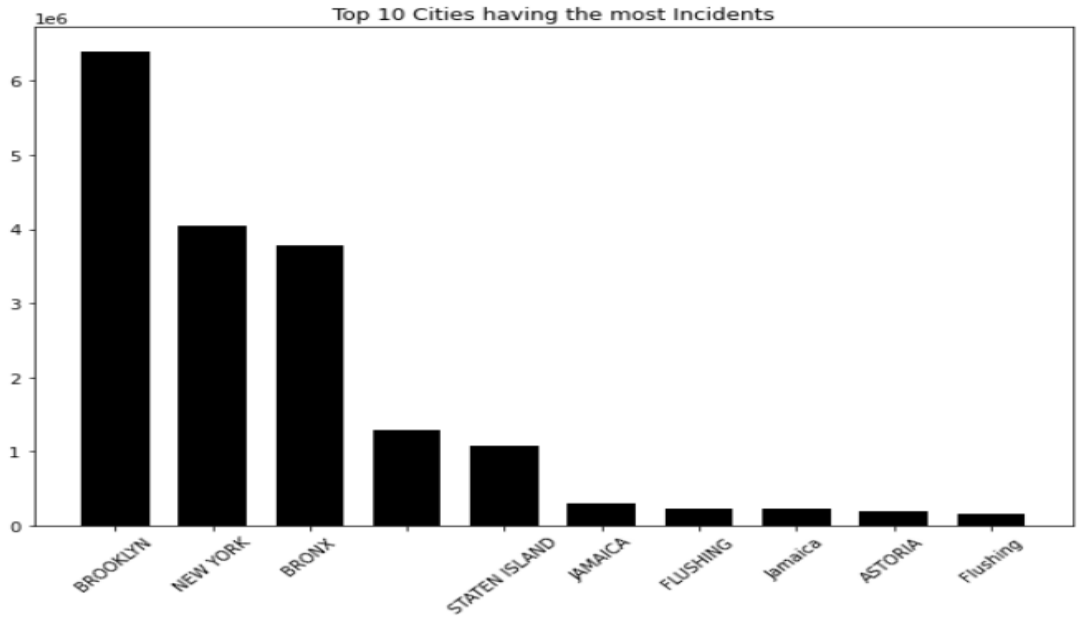
- It is clear what the problem will be if not taken care before model creation. The RESIDENTIAL BUILDING and Residential Building/House convey the same message but while creating incidents they were reported differently.
- We can also see that there is 29.62% with no name indicating that this field is not mandatory while creating incident. This kind of problem needs to be handle by either substituting with a value or dropping such rows. In the later part of the project, we will see this.

- Second Visualization is a Bar Plot for each Borough reporting incident under Top 5 Location Type.



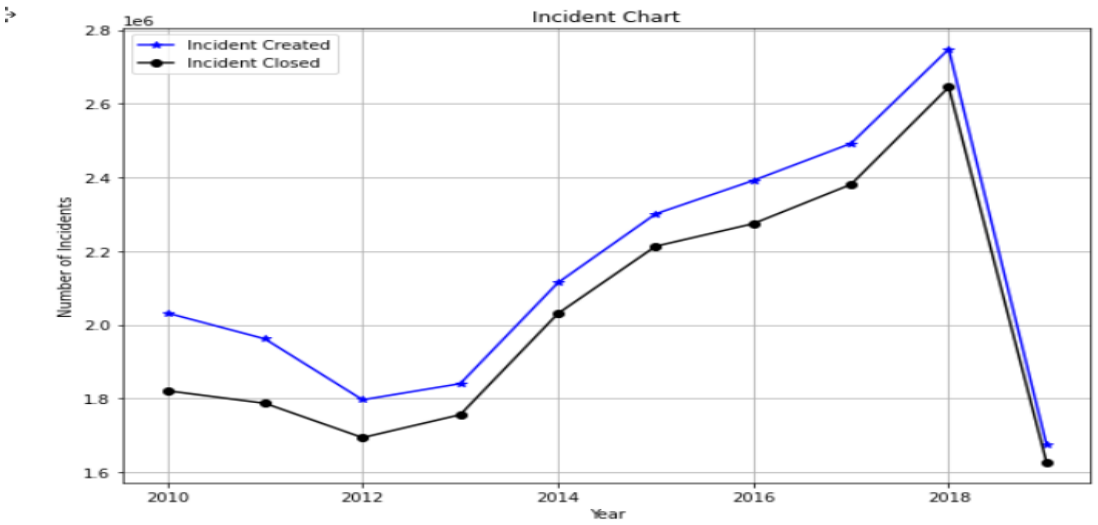
- We can see that RESIDENTIAL BUILDING location type incident is raised highest in each borough except for QUEENS, STATEN ISLAND and MANHATTAN.

- This visualization is a Bar Chart for TOP 10 cities reporting the incident.



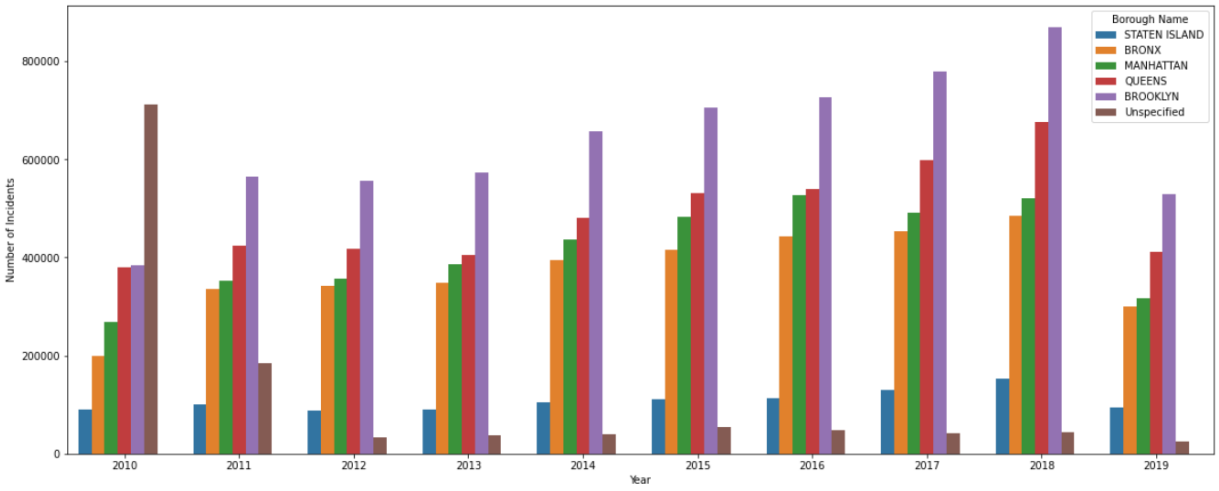
- We can see clearly BRROKLYN reports the highest number of Incidents.

- This represents a line chart showing number of incidents created and closed for each year.



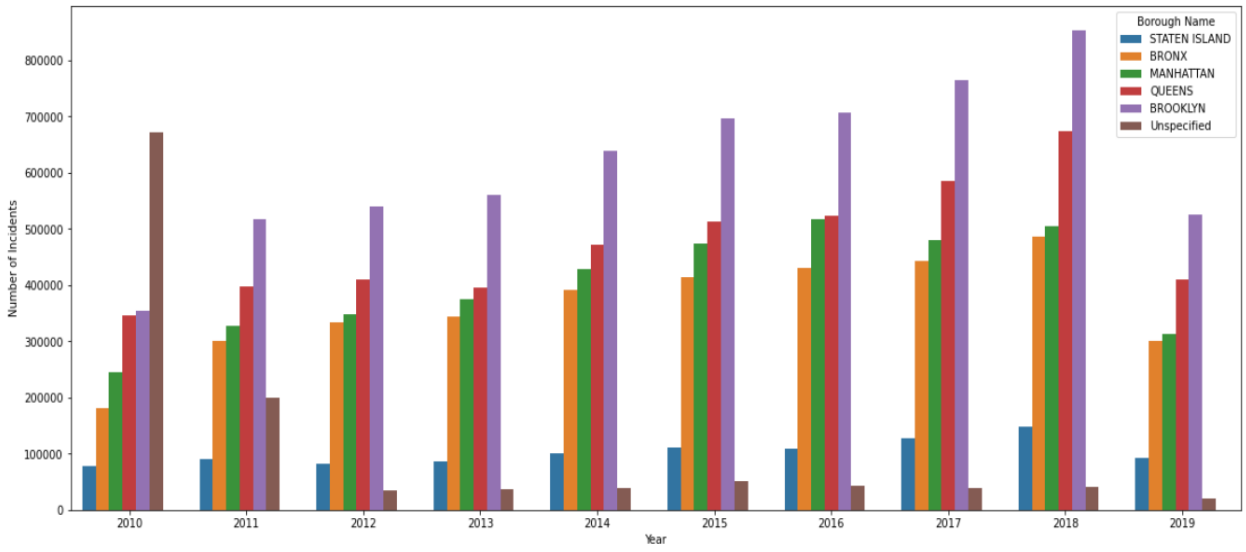
- Before creating the chart I have also seen inconsistent data. There were records for the year 1800, 3200 in created_date and closed_date column. Hence I have filtered out such rows and then plotted the line chart as attached above.

- Next I was trying to visualize what are the number of Incidents created Borough wise for each year and the plot came to as follows.



- We can also see that Number of Unspecified Borough Name Incidents kept on decreasing as people started mentioning Borough Name in their SR from 2011.
- Also BROOKYLN had highest number of incidents for every year.

- I have also tried to visualize, which city solved the most Incidents for each year and it turns out BROOKYLN solved the most incidents.

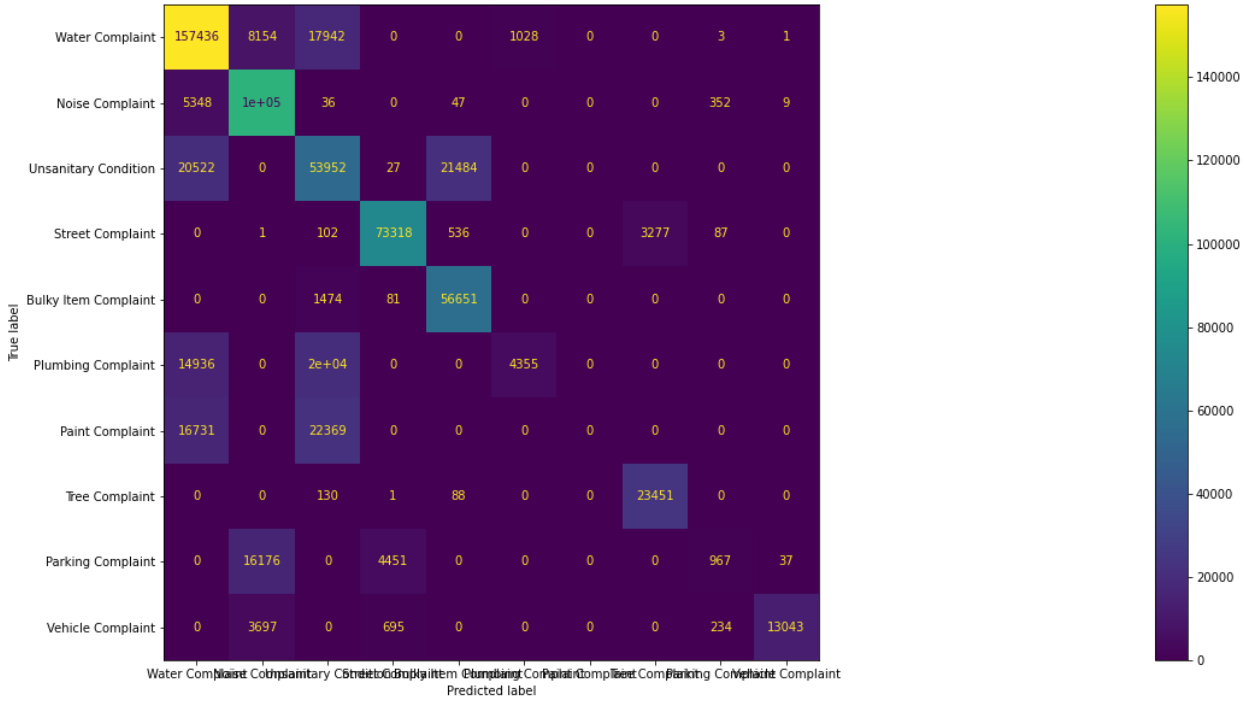


- These are the few of selected visualizations I have made and more visualizations are included in the Code File.

Model Creation

- Even though Google Collab runs faster than my personal laptop it is still not capable to process 27 Million records.
- To tackle this problem, I have selected a subset of records for creating a Model.
- The records are selected from the year 2016 to 2020.
- This subset represents our population data and hence I have created model on this subset of data.
- I have selected created date, closed date, agency, city, borough as independent variables and complaint type as dependent variable which is what we are trying to predict here.
- The reason I haven't taken Descriptor variable is because first it is dependent on Complaint type [dependent variable] also it may have null values. Now in order to handle null values we need to drop such rows which is as good as not selecting the feature as there are lot of null values.

- Even after subset of data being used for model creation which is around 6.9 Million, it is still not possible to process huge amount of data.
- Therefore, I tried to downsize the complaint type records. Since there were too many complaints type to predict, I used regex to narrow them down. For e.g RESIDENTIAL BUILDING is same as Residential Building/House such kind of changes were made.
- After that I am only trying to predict TOP 10 Complaint Type and for TOP 5 City.
- After data preparation the final dataset is of 2.5 Million Record.
- I split the dataset in 70% training and 30% testing.
- I have used Gradient Boosting Classifier to create the ML model.
- The Accuracy is 72.9%.
- I have plotted Confusion Matrix for Model Validation purpose.



- I also tried GridSearchCV to find the best parameters but even after 5 hours of running the GridSearchCV showed no output as the processing power is too low.
- I have also used Random Forest Classification Algorithm and it had accuracy of 73%