TEAM 1:
AYUSH SHIRSAT
SARTHAK
ARORA
JAINAM SHAH

DRAFT ANALYSIS REPORT

Haverhill 311 Service Request Analysis

Summary

City of Haverhill is interested in improving and refining their decision-making process from insights born out of data analysis and visualization. The 311-system set up receives non-emergency requests such as Highway street improvements, street light repairs, etc. All relevant details regarding the issue are stored in QAlert. QAlert is a citizen request management solution that has been deployed by the City of Haverhill. The data stored in QAlert serves as a good repository to analyze and visualize. Having access to such data will help us find the underlying patterns and points of interest. The data was made available to us in a .csv format.

Problem Statement

To assist City of Haverhill in assembling and using 311 data, complaint and information for better insights into providing more effective and efficient customer and municipal services in Haverhill. Specifically, we target the following points:

- To recognize the trends regarding 311 requests and closures of the request by type, time of the year, day of the year, day of the week, time of day.
- Mapping Geographic Information System (GIS) data.
- To observe how Haverhill is performing in managing and responding to 311 requests, by issue, by department, by the time of year, day of week, etc.
- Based on past records compute request response time, which would be communicated to citizens.

Collection Methods

Two datasets were used for this project and they were obtained as follows:

- QScend Technologies using QAlert API provided us .csv containing necessary 311 complaint and activity data.
- GIS data shapefile - provided by the client.

Relevant fields of data

Table 1 indicates the relevant fields of data in the QAlert data that we worked on. The GIS data is explained in the Methodology section.

Field Name	Field Description
Request ID	The unique identifier for the request
Create Date	The date the request was entered into the system
Status	The status of the request
Close Date	The date on which the request was executed
Request Type	The category of the Requests
Department	The department under which the request is assigned
Longitude	The longitude coordinate of the Request origin
Latitude	The latitude coordinate of the Request origin
Origin	The source from which the Request was made

Table 1: Fields of data in the QAlert data.

Methodology

The project was divided as follows:

1) Data pre-processing: To remove NaN values and extract relevant features.
2) Data visualization: To study our data and get insights.
3) Machine Learning: To regress and find estimates for future use.

Data Pre-Processing

- After glancing over the data, the useful columns were first determined as features. Some of them which were not required were dropped.
- Some columns have ID and name, only the ID's were retained as we can always look up their names when required.
- Dates were present in string and were converted to datetime objects with the appropriate format which enabled us to calculate completion time.
- To plot latitude and longitude all entries with these values as 0 were dropped.
- To find completion time only "closed" entries were taken into consideration as others did not have a completion date.
- Request completion time was computed by subtracting "date created" and "close date" for request type. This completion time was in format of day, hr:min:sec. It was converted to seconds while training model for regression.

Geodatabase

- We were provided with shapefiles of the ward and precinct polygons of Haverhill. Shapefiles generally contain a column called 'geometry'. This column can contain Points, Lines or Polygons. These are Shapely objects describing the different geometrical sections within Haverhill.
- The files provided to us mainly contained of rows depicting the ward, precinct and its corresponding polygon. Ward and geometry were also the only columns used from the shapefiles.
- There are 7 different wards in Haverhill.
- The geometry column comprised of Polygons defining the wards. The Latitude and Longitude from the requests were zipped together to form corresponding Point.
- Mapping GIS data requires projecting the geometries (Points & Polygons) to the same crs (Coordinate Reference Systems).
    - The shapefiles were in EPSG:2249 format, which is the Spatial Reference for Massachusetts Mainland. The Longitude and Latitude are in EPSG:4326 (or WGS84 World Geodetic System).
    - The latitude and longitude from the requests were checked in which polygon (ward) they are contained. This resulted in the corresponding data frame for 7 wards.

Machine Learning

Linear SVM was used as regression to find estimates of time. The parameters used to train were C (regularization parameter) and tolerance. The data here was completion time and labels were request type. Models were trained separately for departments with high number of requests.

Motivation: The variance in completion time was very high for all request types. Just averaging or finding median did not give good estimates. Hence, the team thought it would be helpful in creating an ML model that would regress to give estimates. In future as more data would come in there would be regularities in data and a model would be able to differentiate between an outlier and actual data point. This would give robust time estimates.

Difficulty faced: The variance in data was very high. For example, completion time for a request type would range from 1 min to a few months. Training this was extremely challenging as the error would be quite high.

Models were trained for each department with completion time as data and their respective request type as label. The train-test split for data for generally between 10% to 20% based on the size of data. We trained linear SVM models for top 8 departments (based on frequency of requests).

Completion times with 0 values were not used for training. Error used to evaluate was root mean squared error (RMSE).

Model testing:

- Train the model and observe train-test error. High difference in these two errors would imply overfitting.
- If model is overfitting change regularization parameter till the two errors converge and we can say model is generalizing well.
- Change parameters (C and tolerance) till least possible test RMSE is achieved.
- Compare the estimated/predicted times with average and median times to see if estimate is appropriate.

Data Manipulation and Feature Extraction

There were two csv files Request and Activity. Our primary interaction was with the Request data, with its relevant fields of data provided in Table 1. Certain columns were generated to analyze the data.

- Dates were mapped to datetime format
- For the requests that have been closed, we calculated the difference between the ‘Create Date’ and ‘Close Date’ and mapped this data to a new column as ‘Completion Time’
- Extracted the day from the datetime formatted ‘Create Date’ and appended to a ‘DayofWeek’ column
- Extracted the month from the datetime formatted ‘Create Date’ and appended to a ‘Month’ column
- A dictionary for seasons was generated to map the seasons to their months as seen in Figure 6.

Data visualizations

The dataset consists of 70,060 requests which are categorized as “Closed”, “In Progress” or “Open”. The distribution of such requests is shown in Figure 1. Closed requests help determine the time taken for a request to be completed. The trends of “closed” requests will determine the approximate time required for “In Progress” requests to be completed.

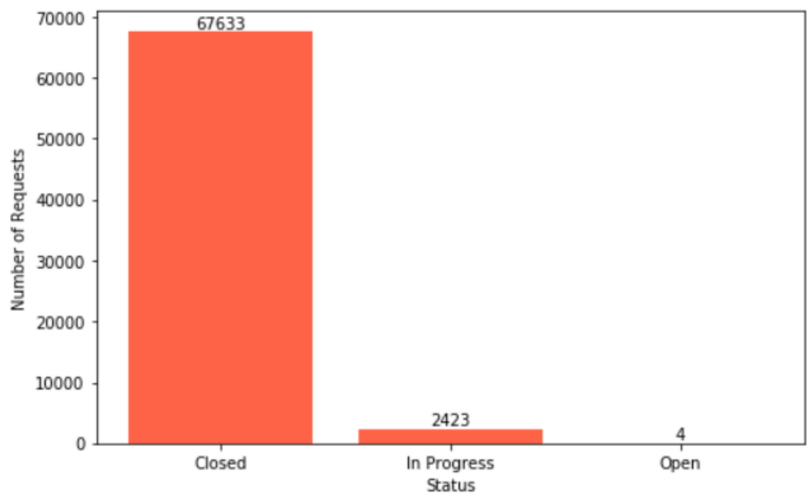


Fig. 1: Status of Requests

Figure 2 demonstrates the origin of requests and the platform people used to report a 311 request.

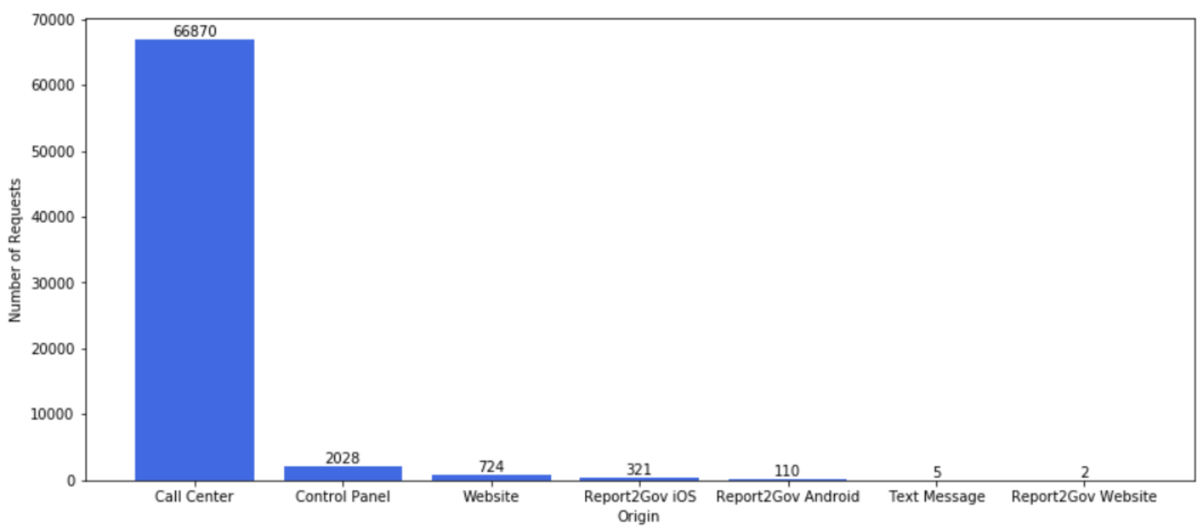


Fig. 2: Origin of Requests

Figure 3 is a plot of all requests based on their geographical location. All points are mapped as per their respective latitude-longitude coordinates. There are 18,080 non-zero coordinates that are mapped.

Figure 4 demonstrates these points on actual map of Haverhill.

Note that Folium package is used to plot these points on actual map. Only 1000 samples of data are plotted because of performance issues. However, one can notice some trends on either side of the Merrimack river where most people reside. The density of requests is higher in this region.

For further visualizations and findings regarding GIS data, refer to the section 'Haverhill Wards and their requests'.

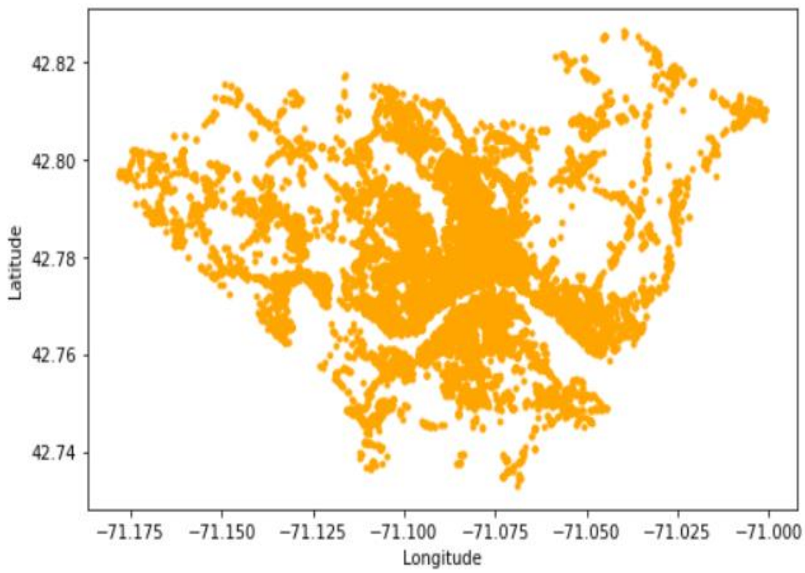


Fig. 3: Distribution of all requests

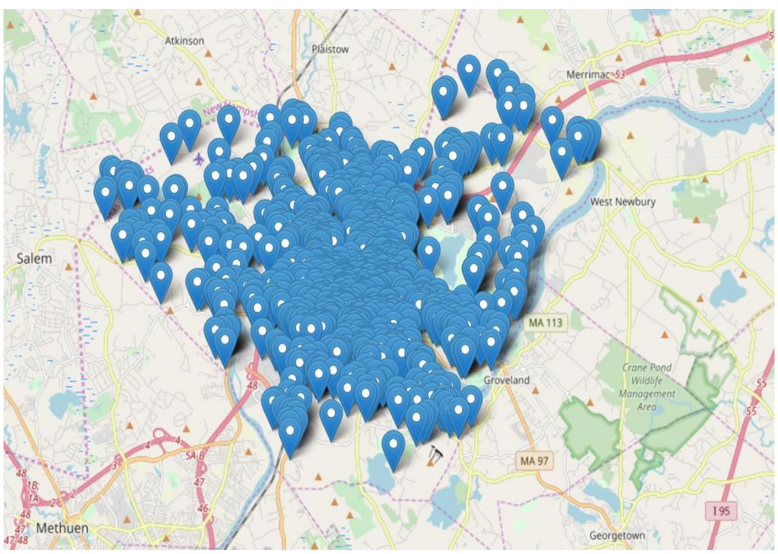


Fig. 4: Distribution of requests (1000 samples) on map

Figure 5 shows the count of requests department wise. Most requests are handled by 311 call center and Highway. In future one might consider grouping departments with less counts as one category.

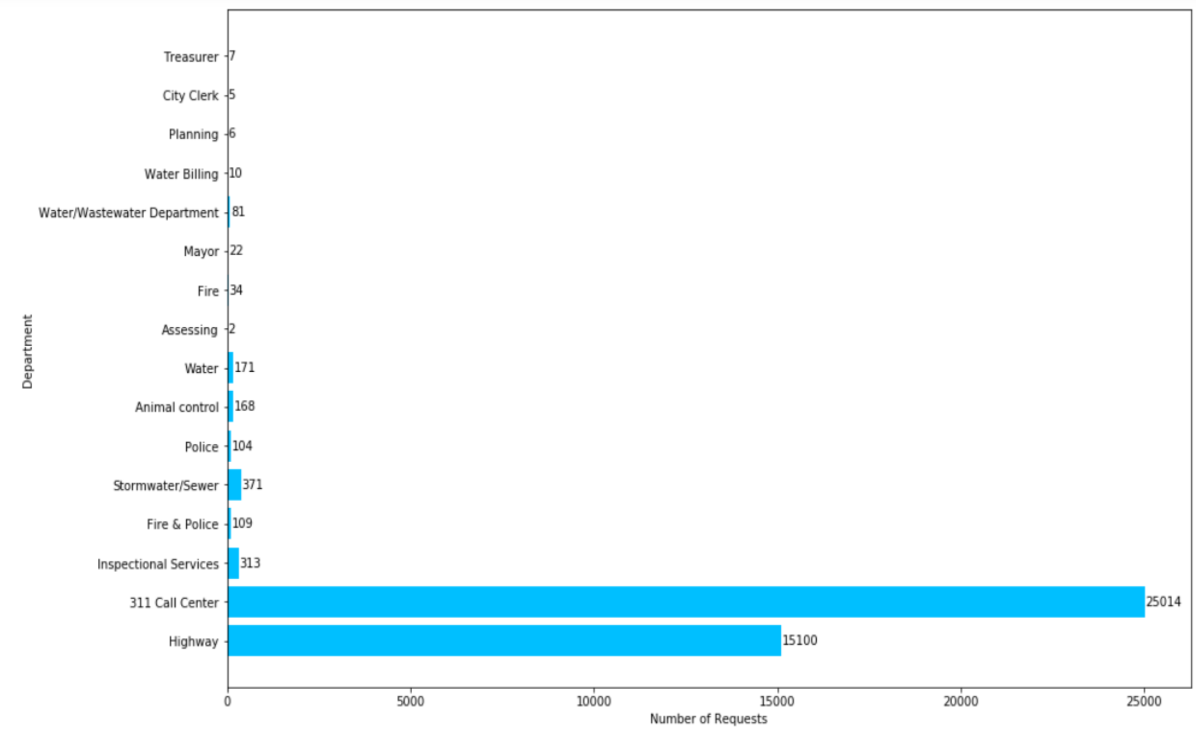


Fig. 5: Department requests

Figure 6 shows the number of requests as per season. This will be useful in predicting if a certain type of request is seasonal. Note the percentage of requests in Winter is high, however, it balances out due to the mapping of 4 months to winter.

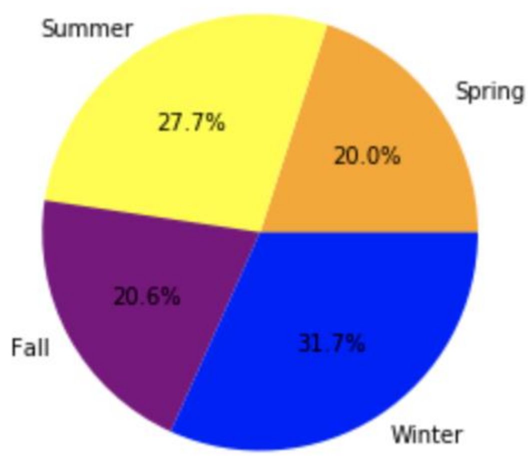


Fig. 6: Percentage of requests based on season

Spring(Apr,May), Summer(Jun, Jul, Aug),
Fall(Sep, Oct, Nov), Winter(Dec, Jan, Feb, Mar)

After having looked at top 30 requests (including Information inquiries), it was mutually decided to separate the Information requests (leading with “A - Information”) as from the non-information requests because the Information requests comprised around 80% of the 311 requests. Figure 7 demonstrates the top 30 Request Types excluding Information Requests. Its observed that Highway - Pothole, Snow Removal Issues and Tree issues dominate the requests made by people.

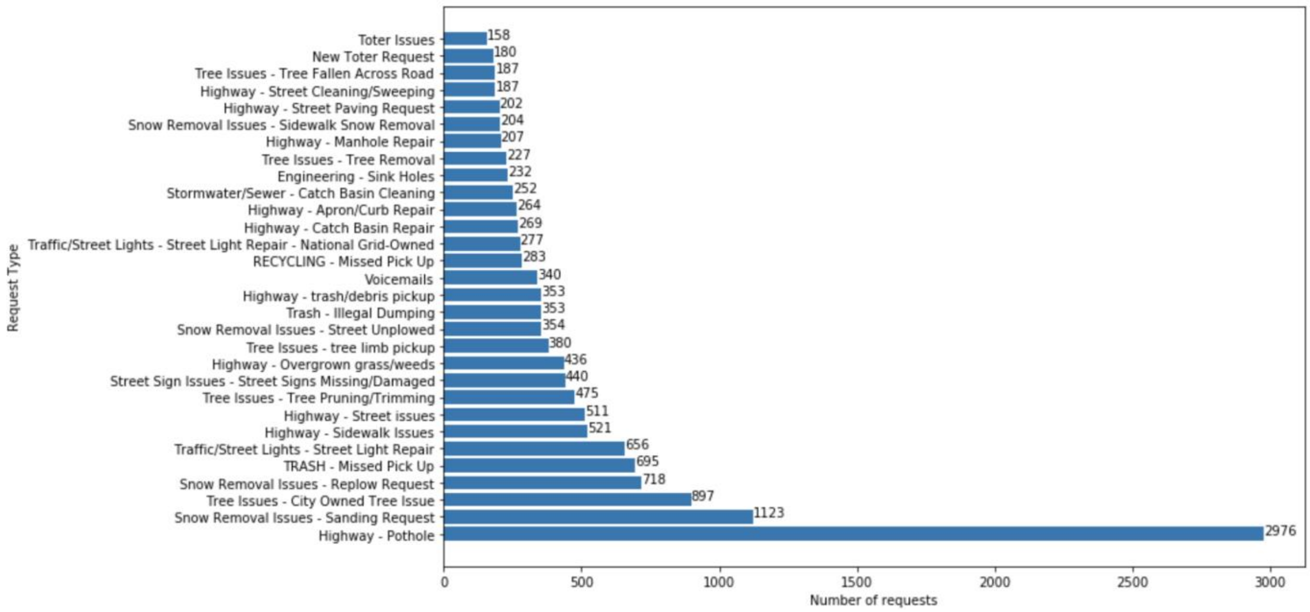


Fig. 7: Request Types (top 30) excluding Information Requests

Table 2 shows the average and median time taken by Haverhill to complete the Request types in Figure 7. This gives us a rough idea as to how different types of request vary in their times and can also act as a good indicator to take action on requests.

Request Type	**Avg time**	**Median Time**
Highway - Pothole	18 days	6 days
Snow Removal Issues - Sanding Request	3 days	2 hours
Tree Issues - City Owned Tree Issue	49 days	15 days
Snow Removal Issues - Replow Request	8 days	5 hours
TRASH - Missed Pick Up	3 days	3 hours
Traffic/Street Lights - Street Light Repair	45 days	14 days
Highway - Sidewalk Issues	98 days	29 days
Highway - Street issues	59 days	10 days
Tree Issues - Tree Pruning/Trimming	54 days	17 days
Street Sign Issues - Street Signs Missing/Damaged	34 days	10 days
Highway - Overgrown grass/weeds	72 days	40 days
Tree Issues - tree limb pickup	31 days	10 days
Snow Removal Issues - Street Unplowed	3 days	2 hours
Trash - Illegal Dumping	16 days	4 days
Highway - trash/debris pickup	53 days	8 days
Voicemails	1 day	3 hours
RECYCLING - Missed Pick Up	3 days	2 hours
Traffic/Street Lights - Street Light Repair - National Grid-Owned	91 days	1 day
Highway - Catch Basin Repair	53 days	14 days
Highway - Apron/Curb Repair	101 days	41 days

Stormwater/Sewer - Catch Basin Cleaning	70 days	15 days
Engineering - Sink Holes	294 days	232 days
Tree Issues - Tree Removal	51 days	14 days
Highway - Manhole Repair	39 days	10 days
Snow Removal Issues - Sidewalk Snow Removal	18 days	5 days
Highway - Street Paving Request	134 days	7 days
Highway - Street Cleaning/Sweeping	80 days	10 days
Tree Issues - Tree Fallen Across Road	65 days	10 days
New Toter Request	1 days	0 hours
Toter Issues	3 days	0 hours

Table 2: Average time per request type (top 30 types)

Figure 8 demonstrates the Information inquiry types and the number of requests made for each type. Haverhill 311's majority (around 80% of 70060) of requests are based on these types.

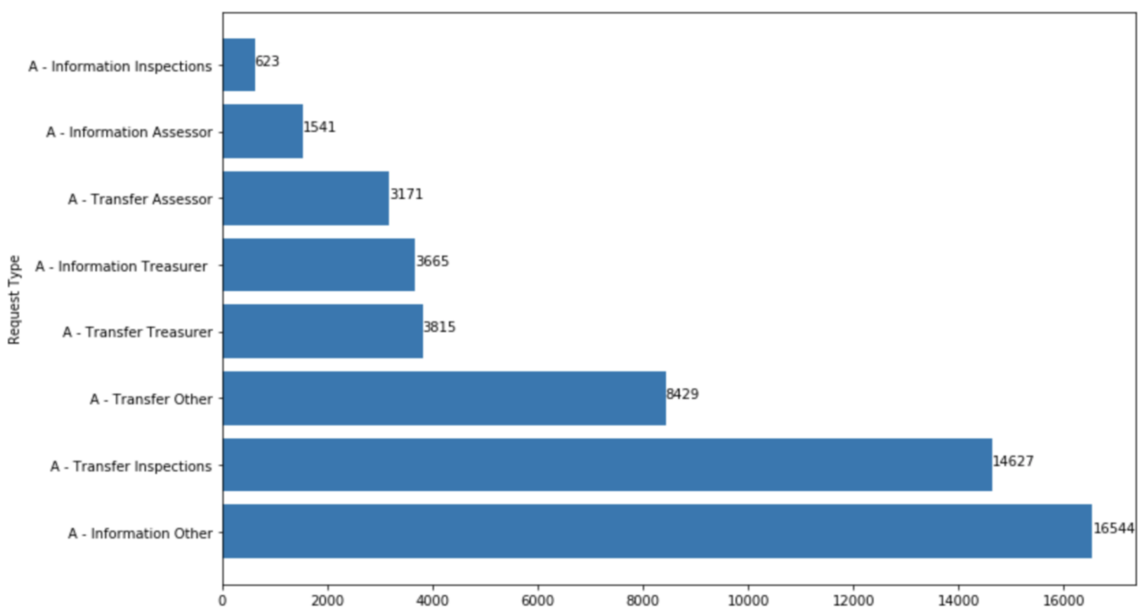


Fig. 8: Information Inquiries/Requests

Table 3 displays the average time taken by the city of Haverhill to respond to Information requests made by people. The median of these requests were 0, which indicates the presence of outliers.

Request Type	Average Time Taken
A - Information Assessor	1 hour
A - Information Inspections	1 day
A - Information Other	13 hours
A - Information Treasurer	1 hours
A - Transfer Assessor	11 hours
A - Transfer Inspections	1 hour
A - Transfer Other	2 hours
A - Transfer Treasurer	1 hour

Table 3: Time taken for Information Requests
Median time taken for all these requests was 0 hours.

Table 4 depicts the average time, the median time and the estimated time taken by each department to close a request. There is a lot of difference between average and median time and none of them act as a good measure to give an estimated time. We tried to perform regression and gave the best possible estimates.

The fourth column gives us an estimated completion time for a department to handle requests. This time is only computed for departments that had large enough number of requests for a machine learning model to generalize.

Department	Average time taken	Median time taken	Predicted time taken
311 Call Center	10 day	0 days	2 hours
Animal control	34 days	7 days	14 days
Assessing	3 hours	3 hours	-
City Clerk	3 hours	0 days	-
Fire	44 days	10 days	-
Fire & Police	10 days	3 days	-
Highway	35 days	4 days	34 days
Inspectional Services	92 days	33 days	89 days
Mayor	159 days	35 days	-
Planning	100 days	42 days	-
Police	41 days	14 days	14 days
Stormwater/Sewer	54 days	7 days	33 days
Treasurer	114 days	52 days	-
Water	154 days	49 days	28 days
Water Billing	6 days	1 day	-
Water/Wastewater Department	67 days	1 day	-

Table 4: Average, Median, Estimated completion time for departments

Department and their request type time

Regression is performed using Support Vector Machine to predict the time taken for a request to be completed. All predicted times are in format of days or hours. For a department if there were too many requests, we have shown tables only for the top 10 requests types based on their frequency. Note that the code gives an estimate time for all requests handled by the department.

Department: Highway

There is a total of 65 request types in Highway Department. Table 5 shows the time taken for top 10 requests.

Number of training samples: 11953

C = 3000

Tolerance = 3000

Request type	Predicted Time taken for completion
Traffic/Street Lights - Street Light Repair - National Grid-Owned	18 days
Snow Removal Issues - Sanding Request	16 days
Highway - Pothole	13 days
Snow Removal Issues - Replow Request	16 days
TRASH - Missed Pick Up	17 days
Tree Issues - Tree Pruning/Trimming	19 days
Trash - Illegal Dumping	18 days
Recycling Drop-off Center	15 days
Highway - Street Paving Request	13 days
Tree Issues - Tree Removal	6 days

Table 5. Time taken by Highway department

Department: Inspectional Services

There are 19 request types in this department. Table 6 shows the time taken by top 10 requests.

Number of training samples: 205

C = 12500

Tolerance = 3000

Request type	**Predicted Time taken for completion**
Trash - Enforcement	53 days
Poor Property Maintenance	38 days
Traffic/Street Lights - Street light/pole request	50 days
Graffiti - Private Property	13 days
Traffic/Street Lights - Street Light Request - National Grid-Owned	47 days
Accumulation of Litter/Garbage	4 days
Vacant Building	59 days
Health & Inspections	16 days
Traffic/Street Lights - Street Light Request	44 days
Excessive Noise/Disturbances	10 days

Table 6: Time taken by Inspectional Services department

Department: 311 Call center

311 call center has 7 request types. While, this department handles the most number of requests, there are only a few training examples. This is because all 0-time values which were mostly pertaining to information-based request were dropped.

Number of training samples: 396

C = 1000

Tolerance = default

Request type	Predicted Time taken for completion
A - Information Other	2 hours
A - Transfer Other	7 hours
LAZ - Downtown Parking Issue	16 hours
Parking Meter Malfunction	1 day
Handicapped Parking Sign Request	12 hours
Train Idling Complaint	1 day
Resident Feedback	1 day

Table 7: Time taken by 311 call center department

Department: Fire & Police

There are only 2 types of requests in this department. Table 8 shows the time taken by all these requests.

Number of training samples: 94

C = 100000

Tolerance = 100000

Request type	Predicted Time taken for completion
Traffic/Street Lights - Traffic Light Not Working	6 days
Downed Electric Power Lines	3 days

Table 8: Time taken by Fire & Police department

Department: Stormwater/Sewer
There are 3 requests types in this department.
Number of training samples: 212
C = 2500000
Tolerance = 3000000

Request type	Predicted Time taken for completion
Stormwater/Sewer - Manhole issue	5 days
Stormwater/Sewer - Catch Basin Cleaning	16 days
Stormwater/Sewer - Sewer Problem	7 days

Table 9: Time taken by Stormwater/Sewer department

Department: Police
This department handles 6 request types.
Number of training samples: 95
C = 15000,
Tolerance = 15000

Request type	Predicted Time taken for completion0
Abandoned Vehicles	3 days
Police Department	30 days
Illegal Parking	14 days
Illegal Out-of-State License	9 days
Motorcycle Noise	19 days
Noise/Disturbance	25 days

Table 10: Time taken by Police department

Department: Animal Control

There are 5 types of request handled by Animal Control department.

Number of training samples: 141

C = 20000

Tolerance = 10000

Request type	Predicted Time taken for completion
Dead Animal on Street	14 days
Wildlife	36 days
Barking Dog	7 days
Pet Cruelty	21 days
Stray Animal	29 days

Table 11: Time taken by Animal Control department

Department: Water

Water department handles 8 types of request.

Number of training samples: 100

C = 200000

Tolerance = 100000

Request type	Predicted Time taken for completion
Water - Leak	19 days
Water - Valve Problem	37 days
Water - Watergate Valve	41 days
Water - Discolored Water	10 days
Water - Pressure Problem	28 days
Water - Hydrant Issue	14 days
Water - No Water	23 days
Water - Quality Problem	32 days

Table 12: Time taken by Water department

Haverhill Wards and their requests

Note: The dataset from QScend had a total of 70060 requests, however only 18080 requests were provided with their longitude and latitude. Therefore, the following visualizations and analysis have been performed only on those 18080 requests.

According to the geodatabase provided to us, there are 7 Wards in Haverhill. Below we've visualized the requests as per their geographic coordinates overlaid on the Haverhill map. Also, each ward is mapped with all its requests and we visualized the top 10 request type of each ward. Highway - Pothole requests is the topmost issue in each ward.

Figure 9 provides us with a representation of 18080 requests mapped to their corresponding wards. Corresponding base map of Haverhill was obtained from OpenStreetMap by inserting the bounds (min longitude, max longitude, min latitude, max latitude) which were taken from the shapefiles given to us.

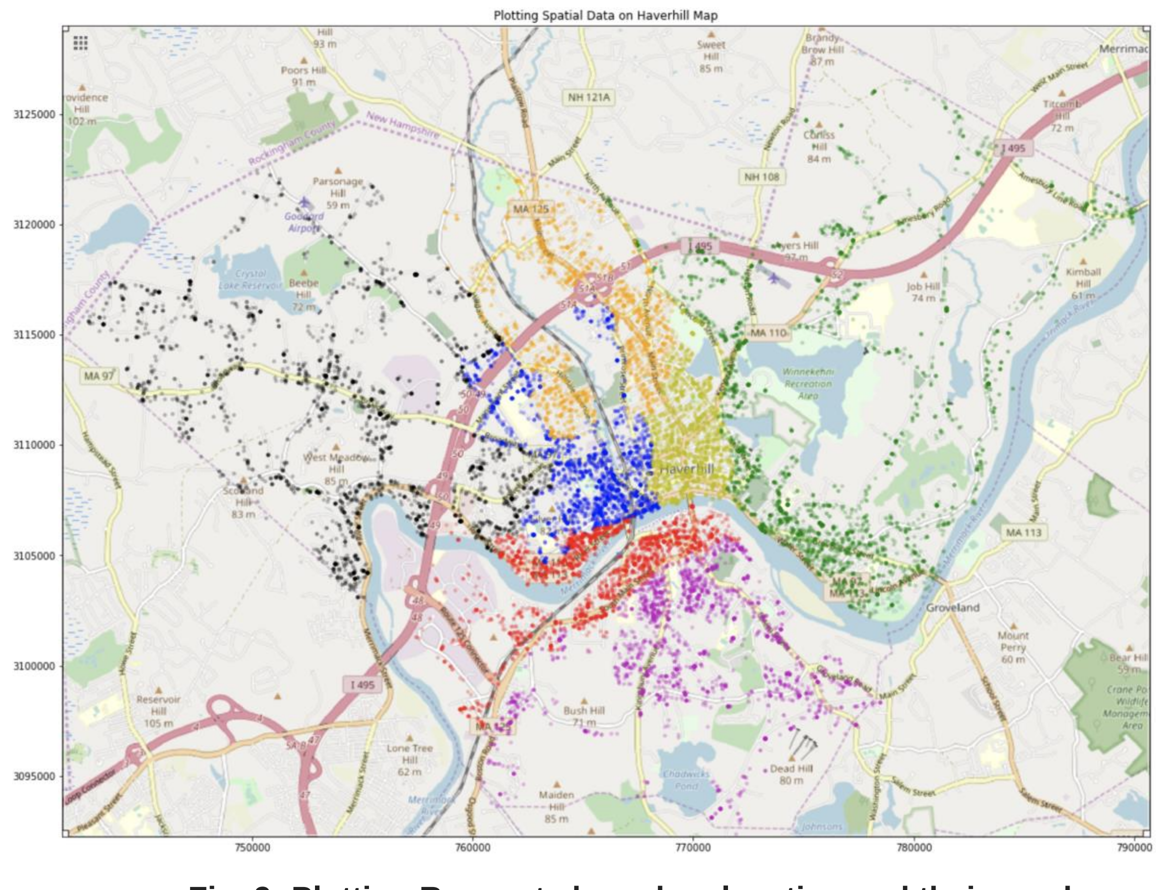
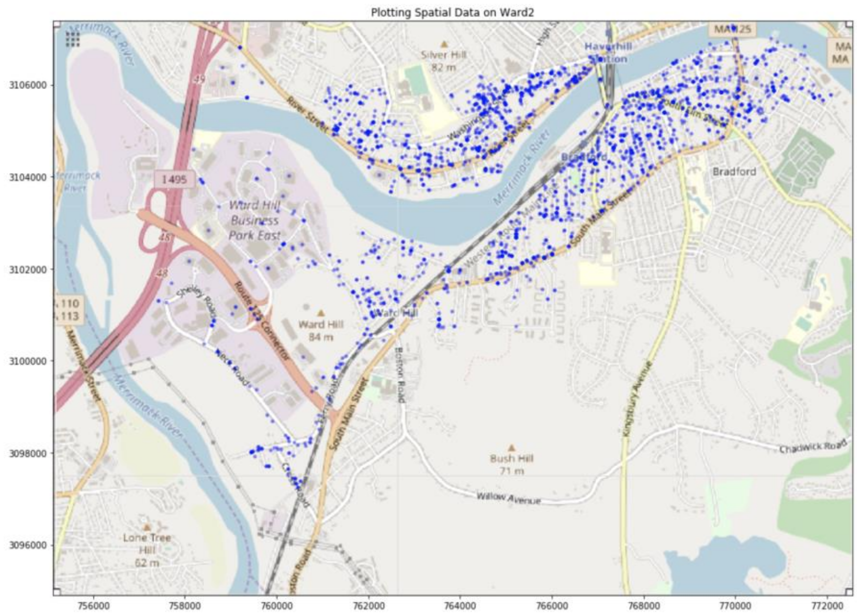


Fig. 9: Plotting Requests based on location and their ward
Ward 1-Blue, Ward 2-Red, Ward 3-Yellow, Ward 4-Green
Ward 5-Black, Ward 6-Orange, Ward 7-Magenta
X and Y axes are the bounds of the Shapely objects as provided in the geodatabase epsg:2249

Figure 10 provides us with the requests in Ward 1 and top 10 request types in Ward 1

Fig. 10: Ward 1 requests and top 10 requests in Ward 1
2649 requests

Figure 11 provides us with the requests in Ward 2 and top 10 request types in Ward 2



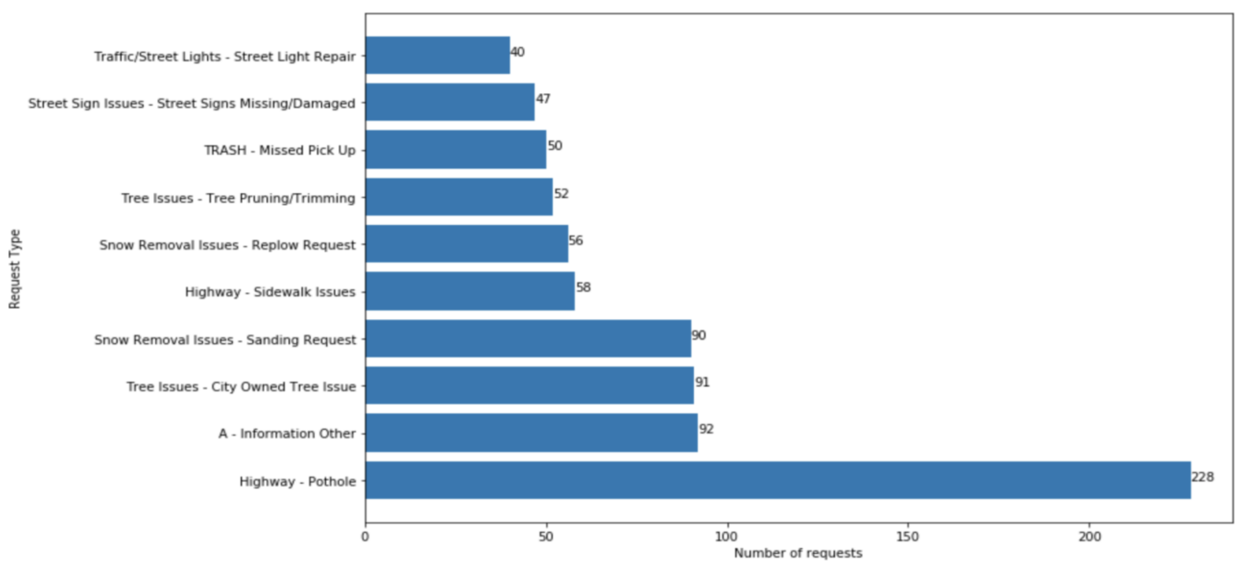
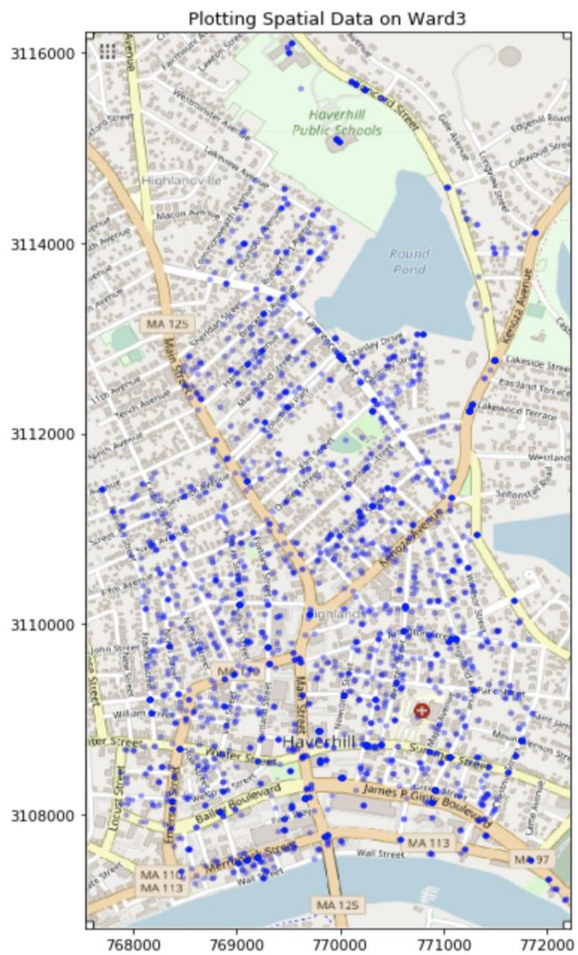


Fig. 11: Ward 2 requests and top 10 requests in Ward 2
2509 requests

Figure 12 provides us with the requests in Ward 3 and top 10 request types in Ward 3



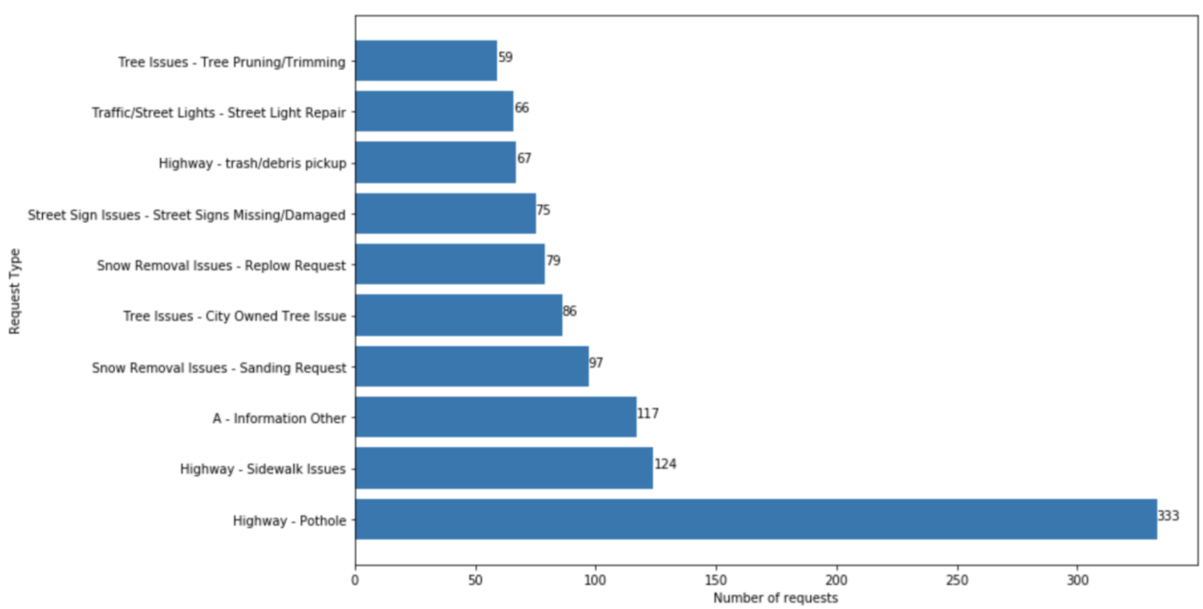
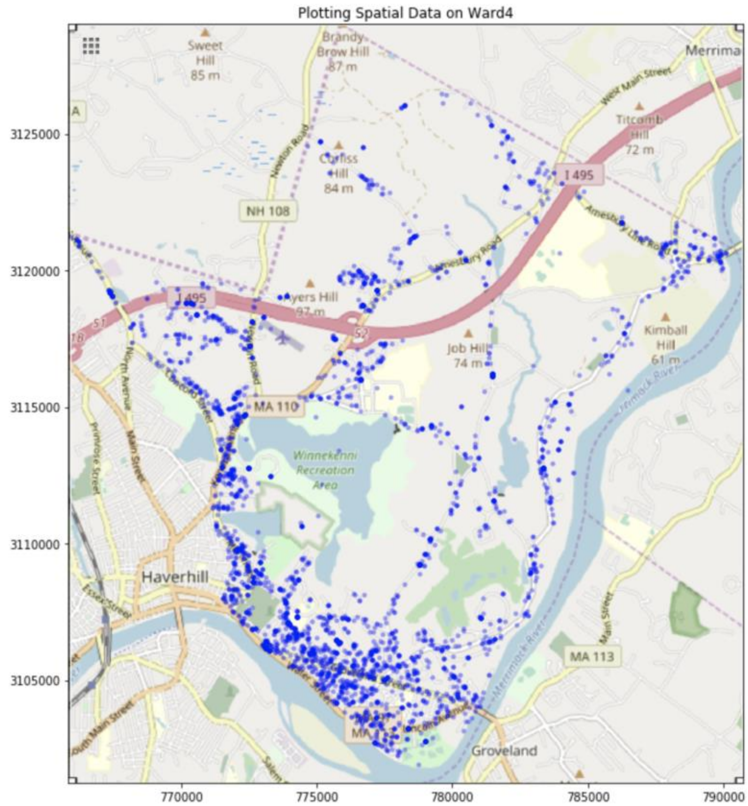


Fig. 12: Ward 3 requests and top 10 requests in Ward 3
2291 requests

Figure 13 provides us with the requests in Ward 4 and top 10 request types in Ward 4



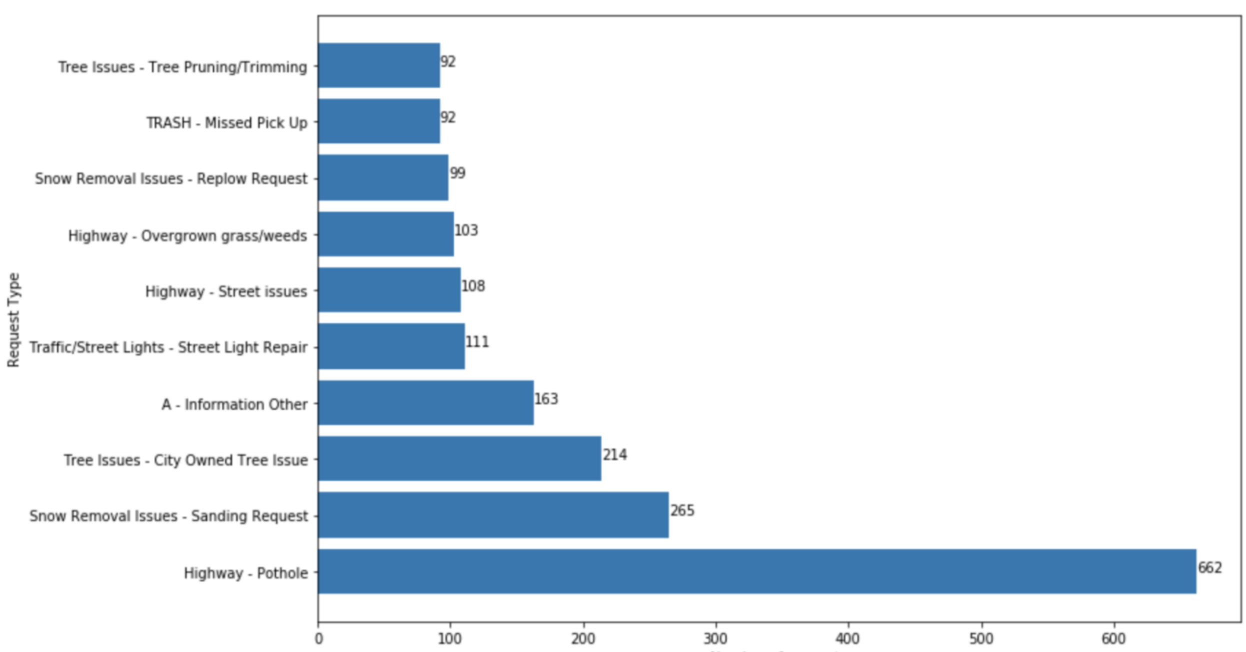
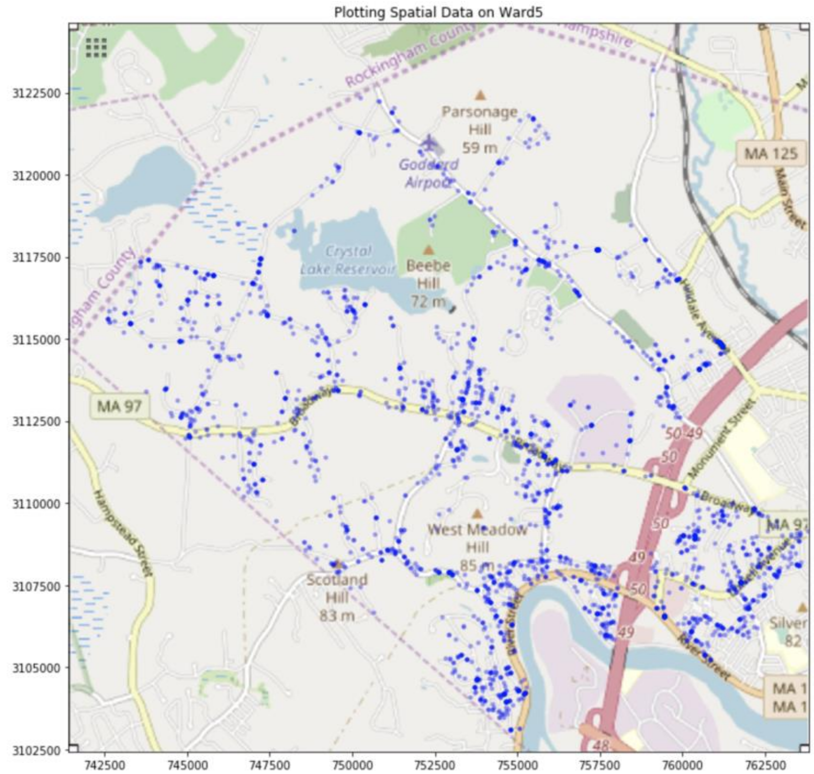


Fig. 13: Ward 4 requests and top 10 requests in Ward 4
3427 requests

Figure 14 provides us with the requests in Ward 5 and top 10 request types in Ward 5



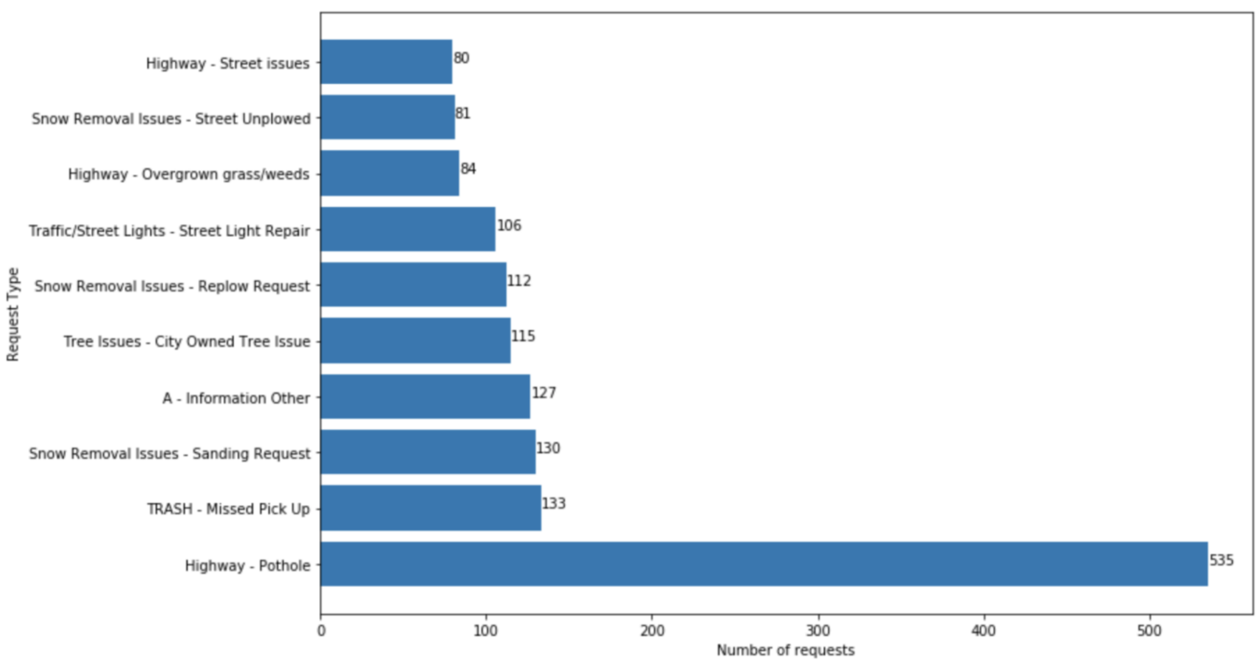
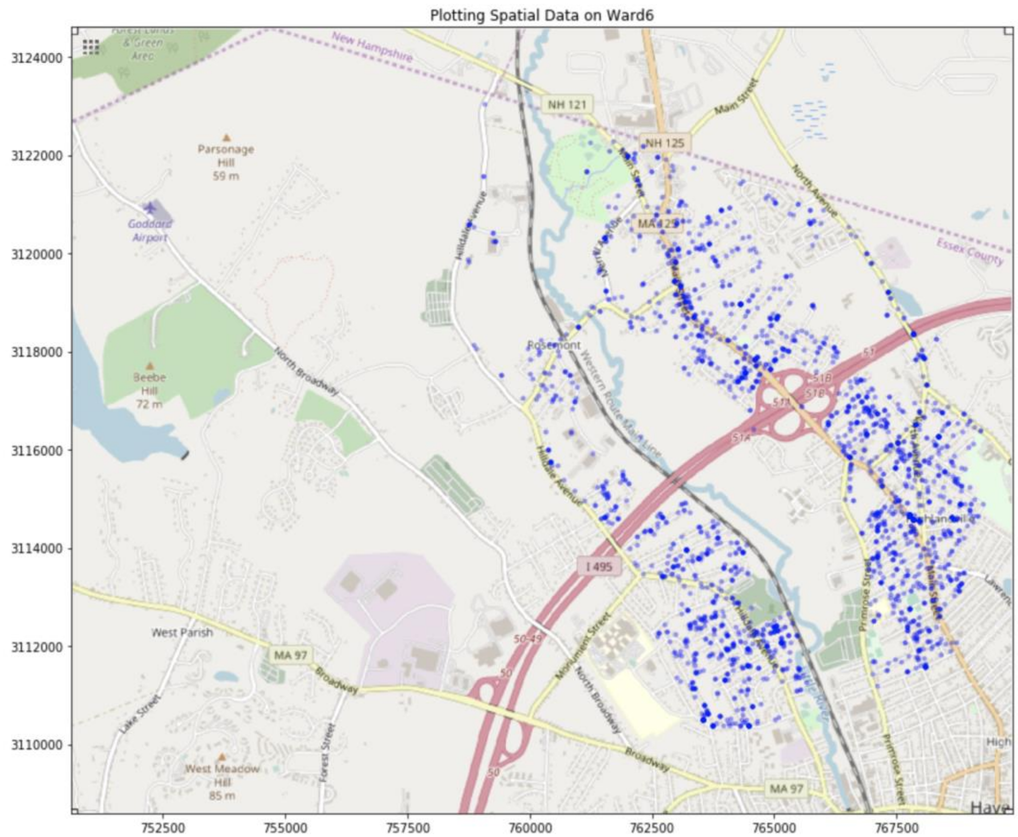


Fig. 14: Ward 5 requests and top 10 requests in Ward 5
2737 requests

Figure 15 provides us with the requests in Ward 6 and top 10 request types in Ward 6



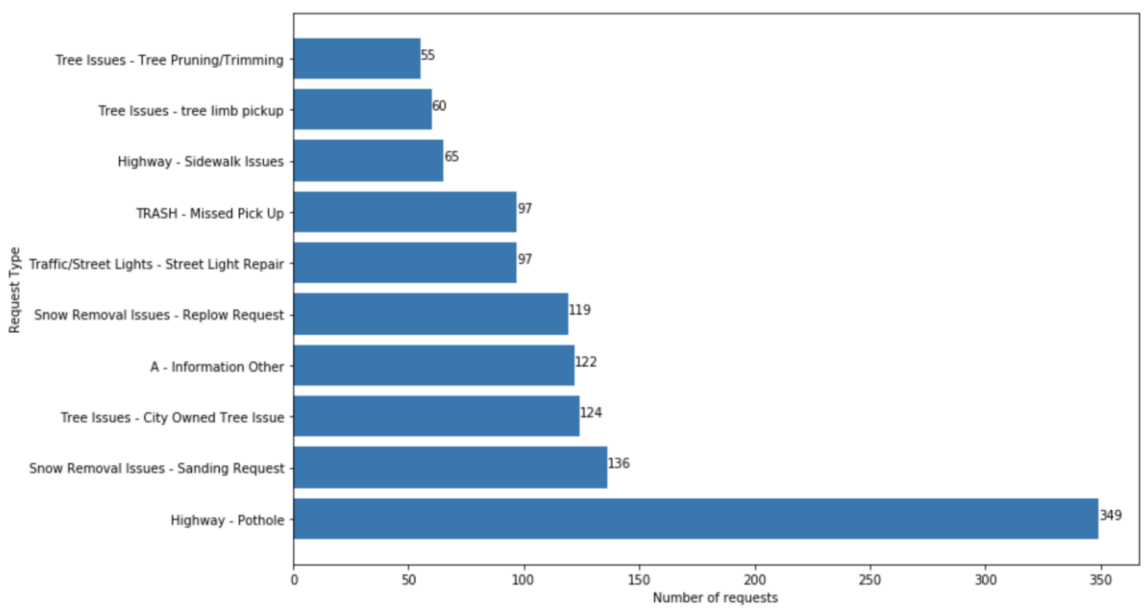


Fig. 15: Ward 6 requests and top 10 requests in Ward 6
2309 requests

Figure 16 provides us with the requests in Ward 7 and top 10 request types in Ward 7

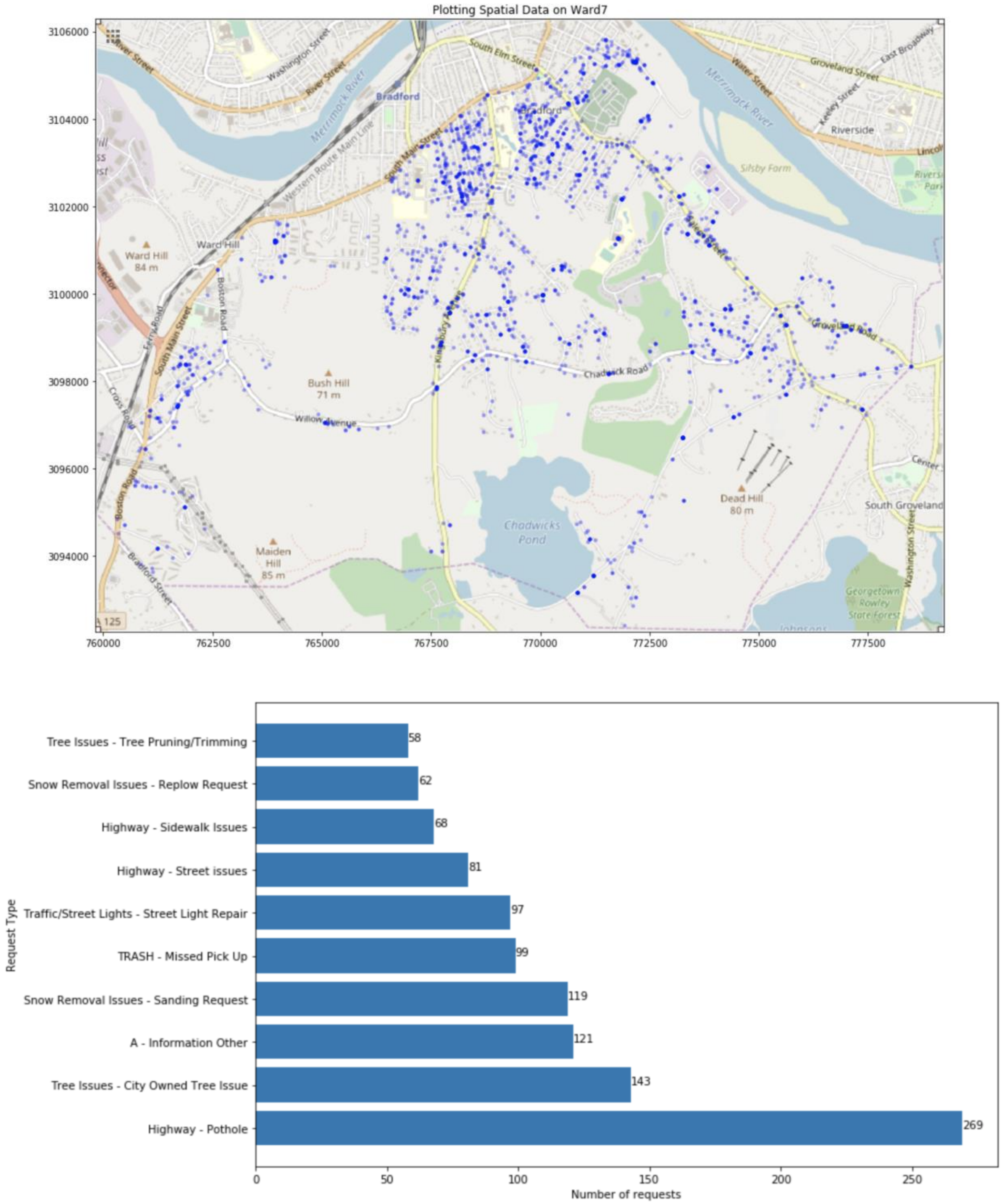


Fig. 16: Ward 7 requests and top 10 requests in Ward 7
2158 requests

Conclusion & Future work

After assessing and analyzing the data set we were successfully able to answer the questions put forth by the client. We were able to get an insight into the trends of 311 requests by their type, time of the year and their location.

Most 311 requests as expected, come from places that are densely populated. When number of requests were plotted with season it was observed that most requests came during winter and summer. Apart from information requests the most popular requests are related to highway potholes, snow removal or tree issues. Snow issues are consistent with the high number of requests during winter.

Plotting the top 10 requests coming from the 7 wards of Haverhill gave us a better idea how these requests varied from region to region. Calculating the average, median completion time for all departments requests gives us a good indicator as to the time certain requests take and how it can be improved going forward.

Regression was performed using Support Vector Machines which predicted the time taken for a request to be completed by a department. This time is only computed for departments that had large enough number of requests for a machine learning model to generalize. In future as more requests come in there would be some regular patterns in the data and the results can be made more robust and accurate.

We propose the following suggestions that can be taken up at some point in the future which will benefit the City of Haverhill and its citizens:

- Use a WeatherAPI & MapsAPI to correlate it with the various requests which will help us predict certain types of requests that originate in certain areas during certain weather conditions. This will help teams be more prepared.
- Build a dashboard that will let a citizen track the status of a complaint and also look through the past complaints and have the ability to quickly reopen it.