CS 573 - Data Visualization

Final Project

Analysis and Visualization of Accident Hotspots in Boston

Xiao Du, Lei Li, Umesh Nair

I. Introduction

Over the years, traffic accidents have become a significant issue to public safety. Approximately 1.35 million people die each year as a result of road traffic crashes. The ability to predict future accidents or highlight areas of high accident risk is thus very useful not only to public safety stakeholders (e.g., police) but also transportation administrators and individual travelers. A potential application of such technique would be real-time safe route recommendation for drivers. With the rapid development of data collection techniques and the availability of big urban datasets in recent years, predicting traffic accidents has become more plausible. Detailed weather data, district information, and motor vehicle crash reports could provide more valuable information for traffic accident analysis.

However, this problem is very challenging due to a few issues. (1) Class imbalance. Traffic accidents are rare incidents. If we construct class labels based on accident vs. no-accident for each road, the classes will be severely imbalanced. (2) Spatial heterogeneity, i.e., the prediction model parameters may vary from place to place. For example, factors causing traffic accidents in large cities with dense population and lower speed limits might be very different from those in rural areas with low population density but high speed limit. A global model might not be very accurate everywhere. (3) The relationship between environmental factors and accidents might be complex and nonlinear. Simple linear models might not achieve good performance.

This paper presents our explorations on effective ways to improve traffic accident prediction results, which is an essential step towards building robust and reliable traffic accident predictive models. Specifically, we consider an unsupervised problem. We collect datasets, including all the motor vehicle crashes in Boston, between 2005 and 2019, detailed accidents location with district information. We first, conduct exploratory data analysis, including accidents patterns in different time period or crash modes, then, we conduct experiments on four classical classification models, i.e., k-dimensional tree, k-means and k means plus plus. At last, we end up with a web application that, user can input their location and different clustering method, and top-k dangerous locations will be marked as circles showing on the map.

II. Related Work

In the previous work, classification models have been widely used. They aim to classify each given road segment at given time into binary classes {Accident, No

Accident}. Zhuoning Yuan [1] incorporates spatial structure of the road network into the classifier through eigen-analysis and significantly improves the performance of all the models examined, including Support Vector Machine, Decision Tree, Random Forest, and Deep Neural Network.

The second group of works aim at fitting regression or other models to predict the number of traffic accident on specific roads or in certain regions. Many of them try to identify correlations between attributes (e.g., weather, road conditions) and the accident risk. Caliendo et al. [2] developed Poisson, Negative Binomial, and Negative Multinomial regression models to predict the number of accidents on given roads. Li YenChang [3] developed a Classification and Regression Tree (CART) model and a negative binomial regression model to establish the empirical relationship between traffic accidents and highway geometric variables, traffic characteristics, and environmental factors. There are lots of work, which indicates the correlation of crash data, for example, unobserved factors, crash frequencies and types observed. Many papers mainly focus on solving this problem by applying multilevel binomial logistic models for predicting the probability of certain types of crashes.

The third group of works view this as unsupervised problem. Isabelle[4] develops an unsupervised categorical model-based accident clustering, like k-means clustering and hierarchical clustering. Then, they uses a more appropriate density-based similarity to assign the accidents to the different clusters. Our work is based on this group and implemented three clustering methods to list the top-k dangerous location near the input location.

III. Dataset and Features

We acquired the dataset for this project from multiple sources. We procured the accidents records from the Vision Zero Boston program, which contains records of the date, time, location, and type of crash for incidents requiring public safety response which may involve injuries or fatalities.

Date range: 01/01/2015 to 02/28/2019
Number of records: 11187

Once we collected the data, we proceeded to perform some preprocessing steps. There were entries in the dataset having 'Other' in the location_type, which had very little information that could help in the analysis, and hence we decided to filter out those records. Next, the records with location type as Intersection did not have any street information, but had the street names(xstreet1 and xstreet2) forming the intersection in other columns. Since, we are using street as an indicator for

identifying the streets and the intersections, we concatenated the values in xstreet1 and xstreet2 to fill in the street information for intersections.

We were also interested in performing analysis based on the time of day. For that, we extracted the time component from the accident date, and categorized the times into 4 different times of day.

Features:

In order to apply the machine learning techniques, we primarily made use of the following features from the data: -

1. Location Type: Street or Intersection
2. Accident mode Type: Pedestrian, Motor Vehicle, Bike
3. Time of day: Early Morning(10pm to 6am), Morning(6am to 12pm), Afternoon(12pm to 5pm), Evening(5pm to 10pm).
4. Location: Latitude and Longitude
5. District: Name of the district in which the street is located.

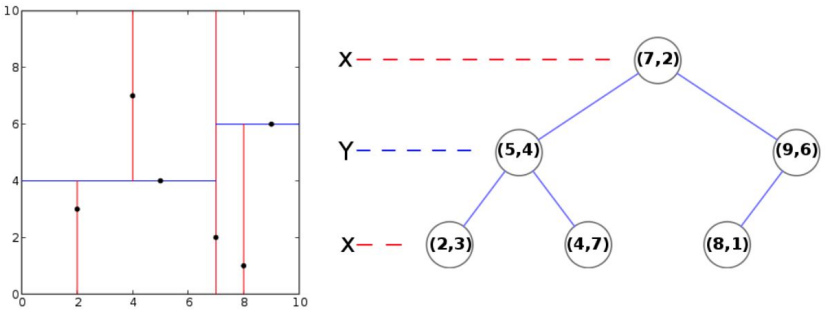
IV. Methodology

1. Machine Learning Methods

Clustering is one of the most widely used data mining techniques in unsupervised learning. The result of clustering is a group of clusters contain data objects that are similar within the same cluster and are dissimilar to the objects in other clusters. There are many clustering algorithms such as k-means, and k-modes. We implement three clustering methods in this project, namely k-dimensional tree, k means and k-means plus plus.

1.1. k-d tree: -

A k-d tree, or k-dimensional tree, is a data structure used for organizing points in a space with k dimensions. It is a binary search tree with other constraints imposed on it. K-d trees are very useful for range and nearest neighbor searches. Each level of a k-d tree splits all children along a specific dimension, using a hyperplane that is perpendicular to the corresponding axis. At the root of the tree all children will be split based on the first dimension (i.e. if the first dimension coordinate is less than the root it will be in the left subtree and if it is greater than the root it will be in the right subtree). Each level down in the tree divides on the next dimension, returning to the first dimension once all others have been exhausted.



Ten areas with higest accident frequency rate are as follows: -

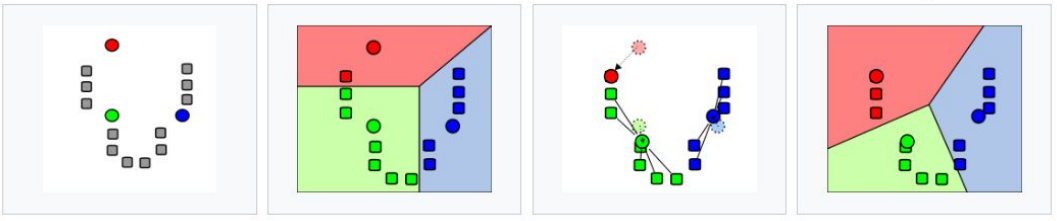
Blue Hill Missionary Baptist Church, Blue Hill Avenue, Mattapan, Dorchester, Boston, Suffolk County, Massachusetts, 02126, USA
Bowdoin Street, Downtown Crossing, West End, Boston, Suffolk County, Massachusetts, 02133, USA 42.360986322532895,-71.062884425
40, North Beacon Street, North Brighton, Allston, Boston, Suffolk County, Massachusetts, 02135, USA 42.35359430593892,-71.13985
99, Devon Street, Mount Bowdoin, Dorchester, Boston, Suffolk County, Massachusetts, 02121, USA 42.30887892989368,-71.078491437
99, Lamartine Terrace, Forest Hills, Jamaica Plain, Boston, Suffolk County, Massachusetts, 02130, USA 42.31396746796662,-71.107
32, Hautevale Street, Clarendon Hills, West Roxbury, Boston, Suffolk County, Massachusetts, 02131, USA 42.27388504376939,-71.13
86, Glenrose Road, Fields Corner, Dorchester, Boston, Suffolk County, Massachusetts, 02124, USA 42.29130280621054,-71.059393991
Southampton Street, Uphams Corner, South Boston, Boston, Suffolk County, Massachusetts, 01125, USA 42.3294923351393,-71.0592435
Economy Parking, Cottage Street, Eagle Hill, East Boston, Boston, Suffolk County, Massachusetts, 02150, USA 42.376665416756516,
Northeastern University, ISEC Pedestrian Bridge, Roxbury Crossing, Fenway, Boston, Suffolk County, Massachusetts, MA 02118, USA

1.2. k-means: -

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. A cluster refers to a collection of data points aggregated together because of certain similarities. K-means algorithm needs a parameter K to determine the number of clusters. At first, the clusters are initiated with random values of data objects as cluster centers. These cluster centers are the centers around which the data objects centered, data objects are assigned to the clusters by calculating the distance between each object and all other centers based on Euclidean distance and is given by the equation stated below, then the nearest distance is chosen. Cluster center is updated by the mean value of objects in the cluster. The process of updating the centers and reassigning the cluster objects are an iterative process until the assignment is stable.

$$d(i,j) = \sqrt{(xi1 - xj1)^2 + (xi2 - xj2)^2 + ...(xip - xjp)^2}$$

where i and j two objects decribed by p numeric attributes.



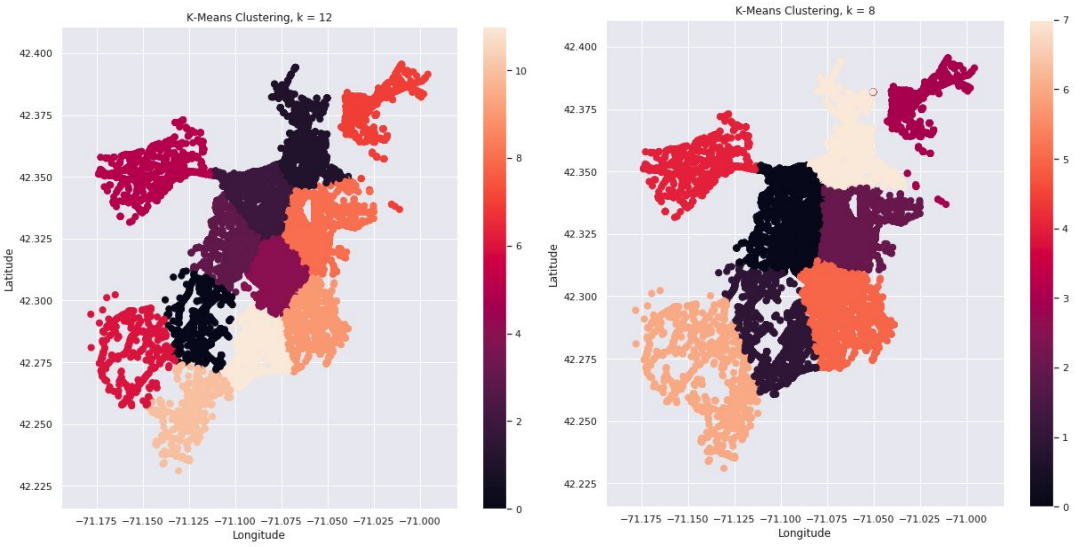
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

The following plots are the clusters based on different k using our dataset.



1.3. k means plus plus: -

In data mining, k-means++ is an algorithm for choosing the initial values (or seeds) for the k-means clustering algorithm. This is because the traditional k-means has at least two major shortcomings:

- First, it has been shown that the worst case running time of the algorithm is super-polynomial in the input size.
- Second, the approximation found can be arbitrarily bad with respect to the objective function compared to the optimal clustering.

The k-means++ algorithm addresses the second of these obstacles by specifying a procedure to initialize the cluster centers before proceeding with the standard k-means optimization iterations. With the k-means++ initialization, the algorithm is guaranteed to find a solution that is O(log k) competitive to the optimal k-means solution.

The intuition behind this approach is that spreading out the k initial cluster centers is a good thing: the first cluster center is chosen uniformly at random from the data points that are being clustered, after which each subsequent cluster center is chosen from the remaining data points with probability proportional to its squared distance from the point's closest existing cluster center.

2. Web Framework

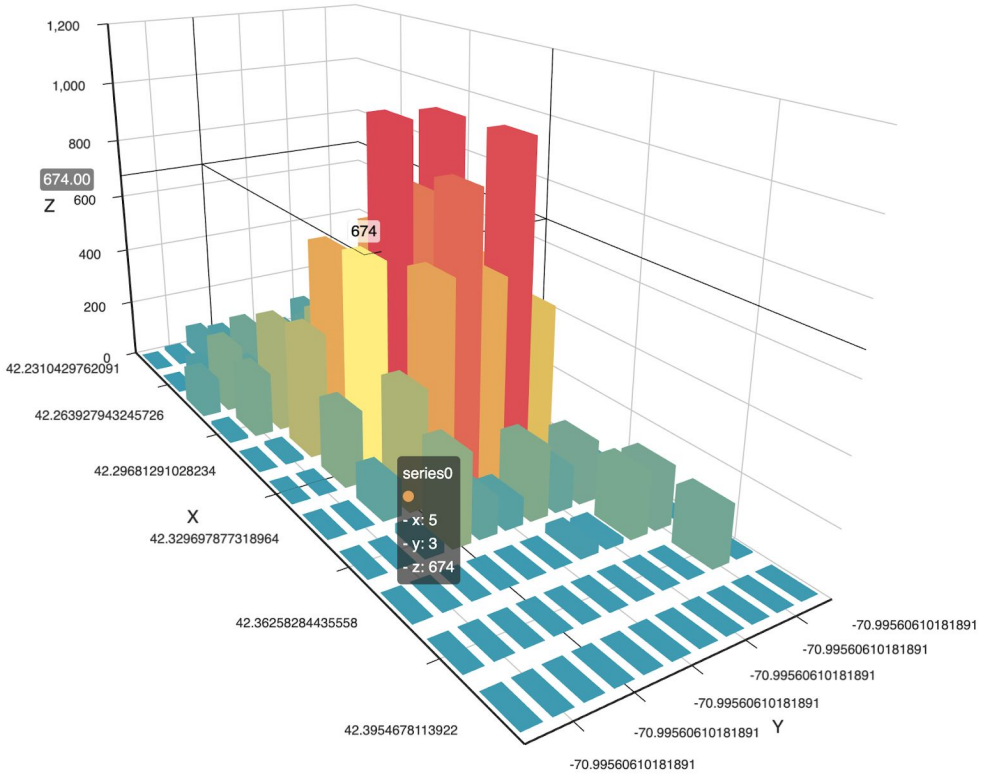
Flask framework is implemented to build a web apps in this project. In the backend we used python to do data acquisition, data cleaning, feature generation, machine learning, some parts of data visualization, and processing user's input location via Request class. In the frontend, we used d3.js, and google map API to visualize the results from data analysis.

V. Visualization

1. Exploratory Data Analysis

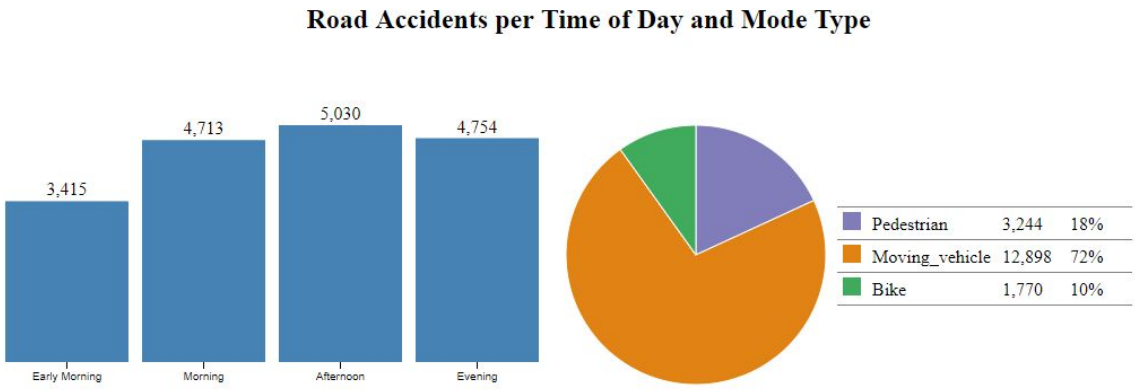
1.1. Road accidents in 3-D :

We divide the Boston city map into 100 parts according to the range of longitude and latitude, and the height of bar means the number of historical traffic accident count. In the following case, the yellow bar means there were 674 traffic accidents happened in that area (center coordinates: -70.995, 42.362)



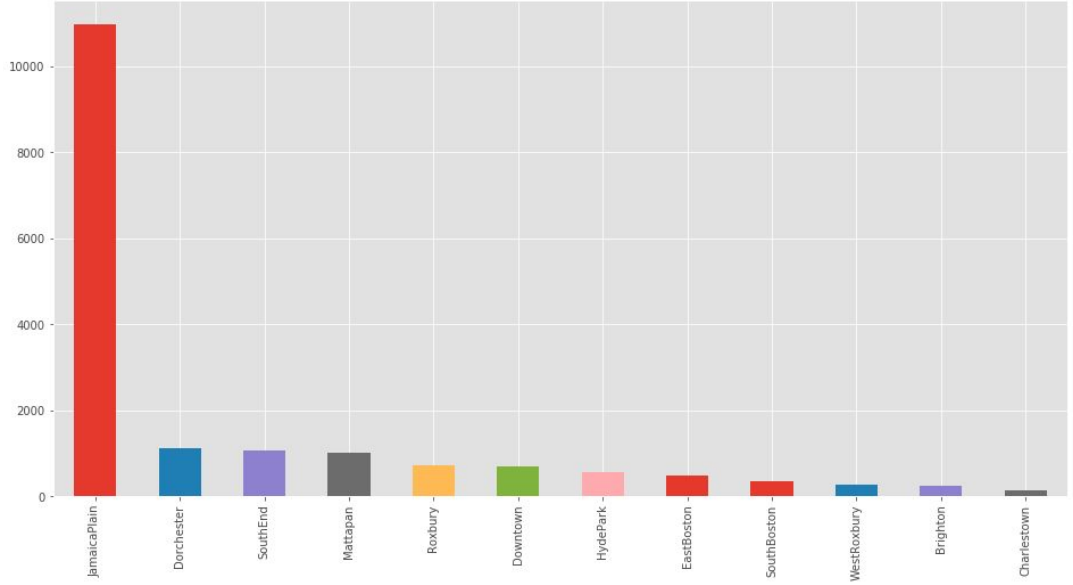
1.2. Road accidents with time and mode types:

The following interactive charts plot the road accidents per time of day regarding to different mode types. The bar chart in blue shows the distribution of accidents in one day using our whole dataset. Clicking the each section in pie chart, the bar chart will show the number of accidents by corresponding types in one day. The time period shows as Early Morning(10pm to 6am), Morning(6am to 12pm), Afternoon(12pm to 5pm), Evening(5pm to 10pm).



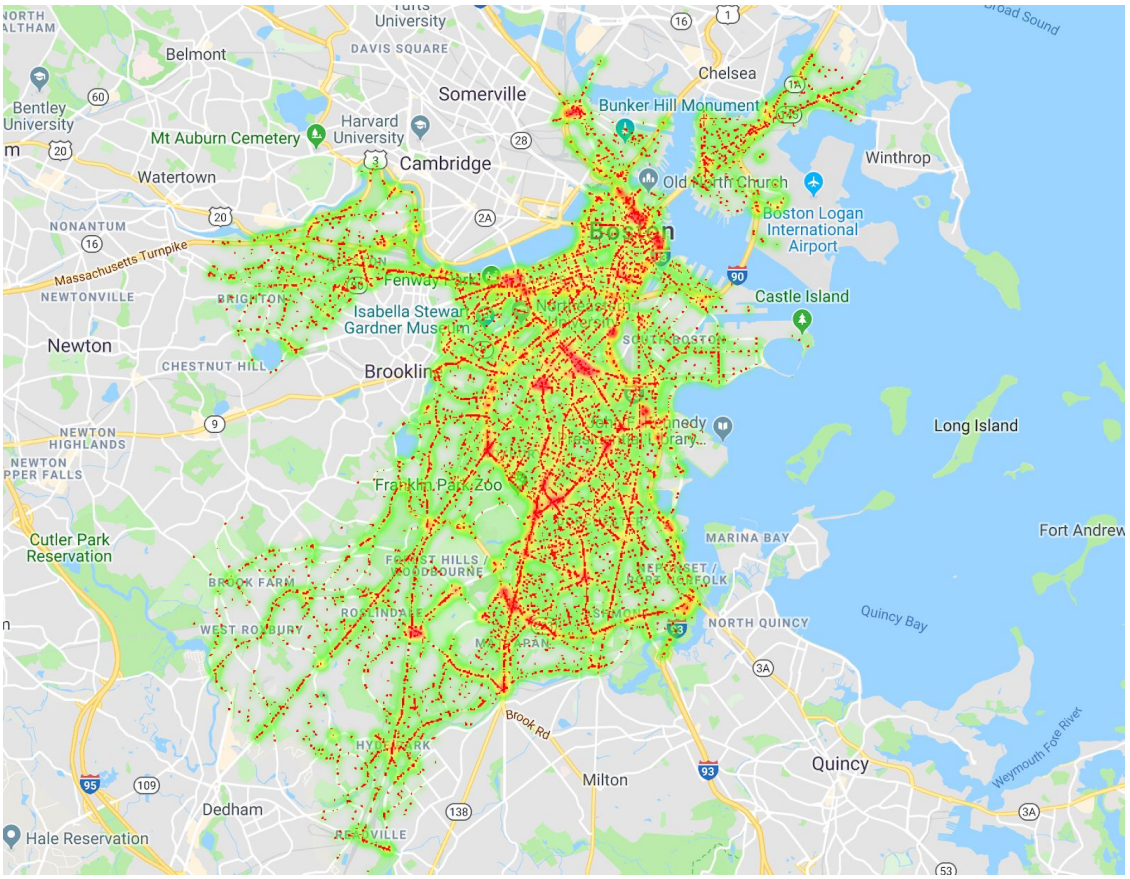
1.3. Road accidents with districts:

We divided the boston in to 12 districts.The following bar chart shows the number of accidents in different districts.

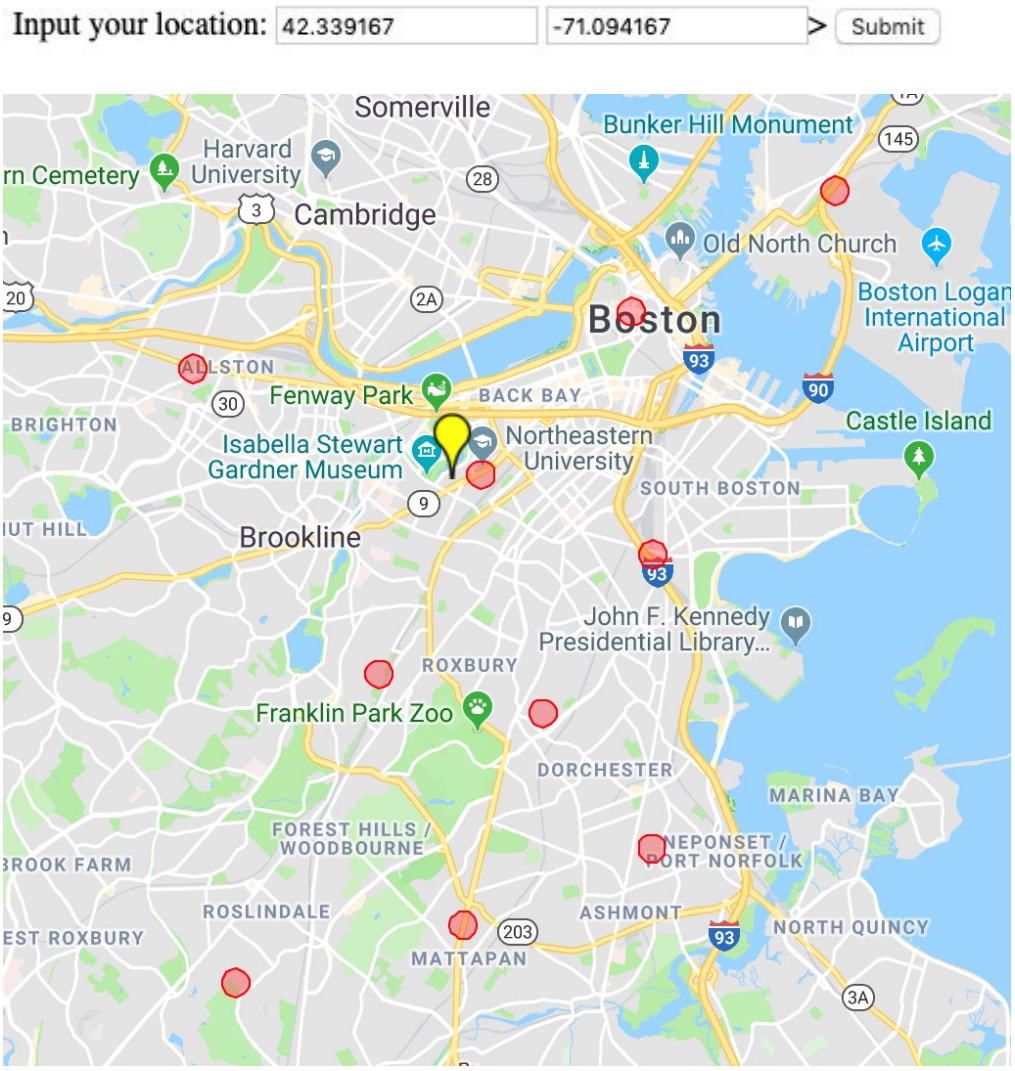


2. Dashboard

For the dashboard of our web application, we map our accidents data into Google map. The following plot shows the heatmap of accidents. We can easily view the locations where happens accidents a lot as the color indicates.



Our web application also allows the user to input his current location using longitude and latitude format. It will generate the map with circling top-k hazardous streets or intersections. The following plot shows the input bar and generated results. The current location is marked as Yellow.



VI. Conclusion

In this project, we were able to combine machine learning and data visualization techniques to analyse accidents data in Boston, to provide some descriptive analysis of past 5 years with the help of various charts, and then apply clustering techniques on the accidents statistics for different streets to identify and mark potentially accident-prone areas in Boston.

This could be used as an application for helping drivers to understand dangerous areas while driving around Boston, so that they could drive safely in those areas, or completely avoid those streets and possibly take a safer route to get to their

destination. This work can also be extended and used by road-safety authorities to identify these accident-prones streets and areas, and focus their safety measures in those areas, in order to reduce the number of accidents in the future.

This project is also testament to the power of the combination of machine learning and data visualization, and how these subject areas can complement each other to produce outstanding and actionable results, where one can derive insights from data, otherwise impossible to the human eye, and the other can present those insights to the common man in a simple, convenient and understandable manner.

VII. Further Work

In future, this work can be taken forward by applying other machine learning techniques like time-series analysis and regression to try and forecast or predict the number of accidents on these streets for future dates. The results can then be visualized on a map along with a timeline, which would show the historical accident statistics and then the user can slide the bar onto future dates, and the map visualizes a rough estimate of the number of accidents on different streets that the regression model predicts to see for that time period.

VIII. References

[1] Zhuoning Yuan, Xun Zhou and Tianbao Yang. 2017. Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study. In Proceedings of 6th International Workshop on Urban Computing, Halifax.

[2] Ciro Caliendo, Maurizio Guida, and Alessandra Parisi. 2007. A crash-prediction model for multilane roads. Accident Analysis & Prevention 39, 4 (2007), 657–670.

[3] Li YenChang and Wen ChiehChen. Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research, Volume 36, Issue 4, 2005, Pages 365-375.

[4]. Isabelle Thomas.Road Traffic Accident Clustering With Categorical Attributes. Proceedings of the 83th Annual Meeting of the Transportation Research Board.