APPLIED DATA SCIENCE

"*Battle of Neighborhoods*" – Coursera Capstone Project

Project Title

Recommending Hospital need in the near vicinity of high prone Accident Areas

Author: Deepen Virani

Published date: Oct 28, 2018

Table of Contents

1. **Introduction/Business Problem**

1.1 *Abstract Idea:* Many of the times we observed that due to lack of availability of hospitals in the close vicinity of high prone accident areas it takes lot of time for the patient to get immediate first aid in case of road accidents. This leads to loss of many lives due to the time taken to reach to the hospital.

This project aims to aid the user/government with choosing an ideal location for opening a new medical center or hospital in any city. It uses the data of all the accidents occurred in a target city or location on daily basis and check if there are adequate hospitals / medical facilities available nearby to the high prone accident zones. In case of no availability of hospitals / medical facilities within in specified distance it recommends opening a new hospital near that area.

This study uses Austin city's daily accident data statistics from the U.S. govt. website and identifies the top most accident-prone location. Now, using Foursquare API it checks the number of hospitals present near to that location. This model can be extended to any City / State.

1.2 *Target Audience:* The specific target users for this kind of application would include individuals / organizations / firms who are planning to open a new hospital of medical center. Also, it can be used to alert government authorities to recommend opening a new hospital in any location.

Based on this project, I have demonstrated 2 Use cases:

a) Among all the accident spots identified, taking the highest number of accident spots and identifying hospitals surrounding it.

b) K-means on all the coordinates (Latitudes and Longitudes) of accident areas.

2. Data Collection

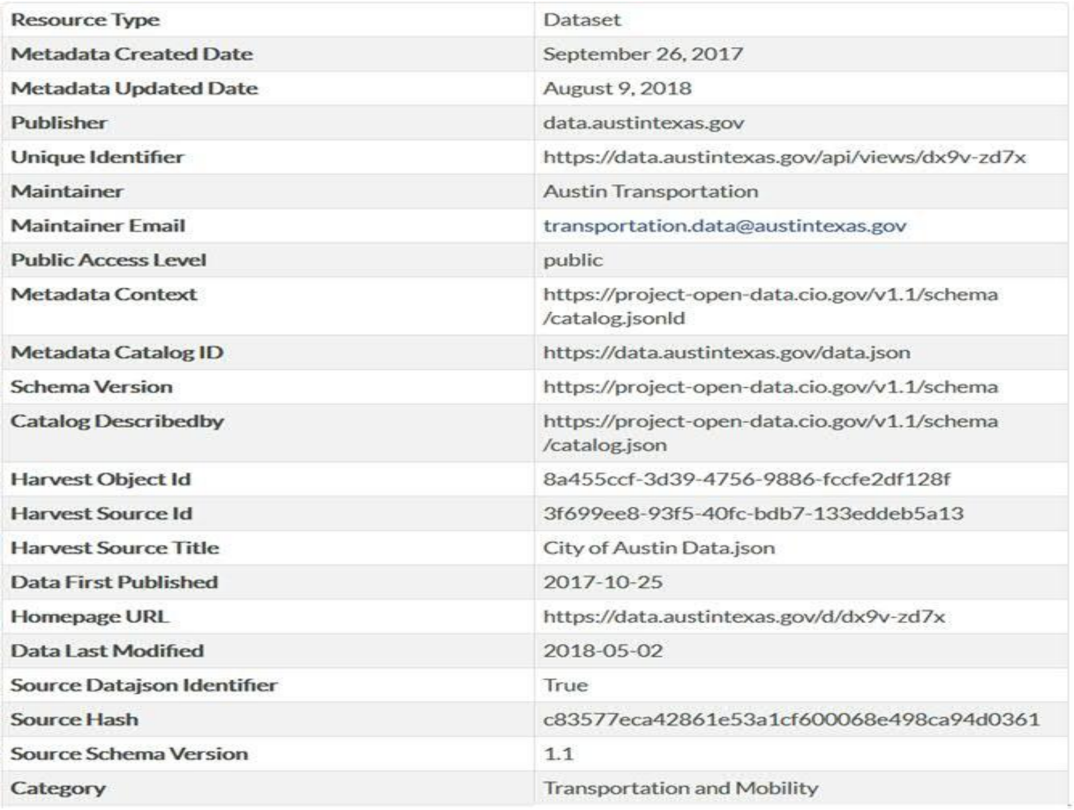
The two main sources of data used for this project are:

a) Real time Traffic Incident reports
b) Foursquare API data.

a) **Real time Traffic Incident reports:** This data sour This data set contains traffic incident information from the Austin-Travis County traffic reports RSS feed.
This is available at https://catalog.data.gov/dataset/traffic-reports.
Publisher : data.austintexas.gov

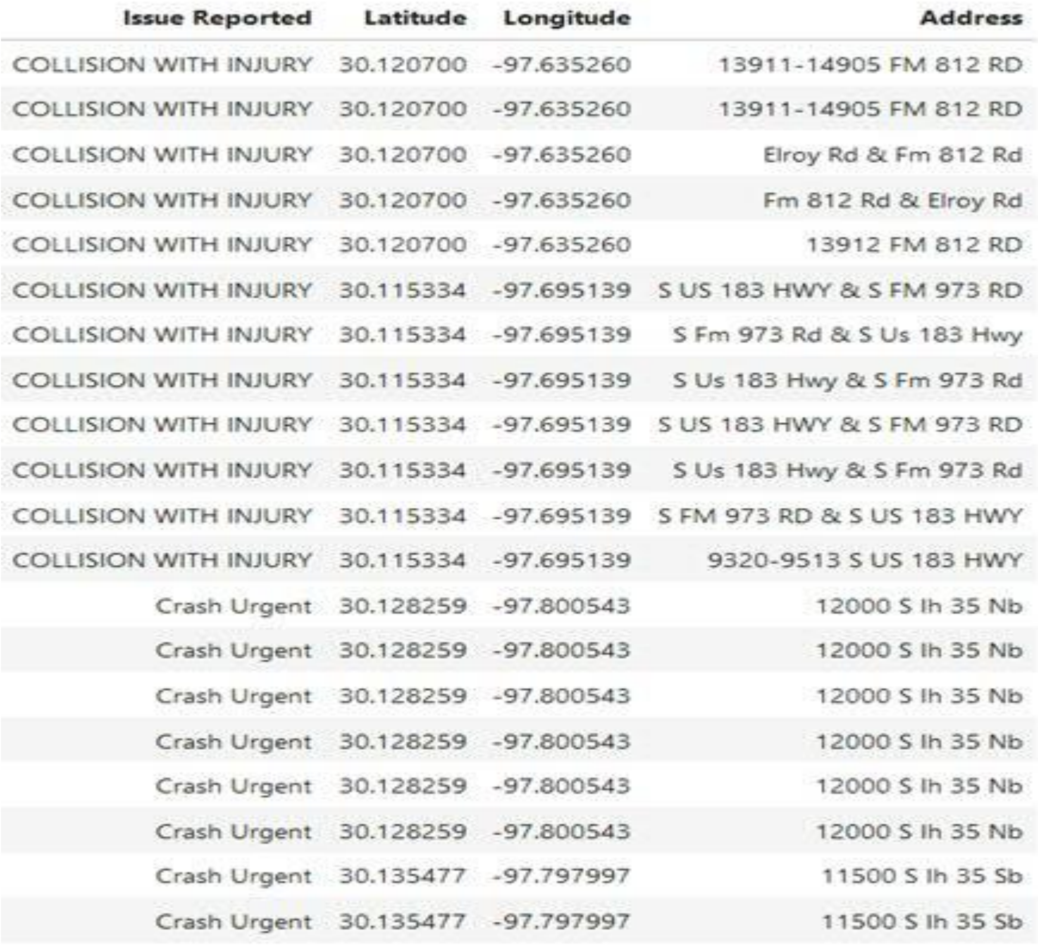
Check for the data set I have added in the repository of Github which is used for this project.

Additional Metadata:



Resource Type	Dataset
Metadata Created Date	September 26, 2017
Metadata Updated Date	August 9, 2018
Publisher	data.austintexas.gov
Unique Identifier	https://data.austintexas.gov/api/views/dx9v-zd7x
Maintainer	Austin Transportation
Maintainer Email	transportation.data@austintexas.gov
Public Access Level	public
Metadata Context	https://project-open-data.cio.gov/v1.1/schema /catalog.jsonld
Metadata Catalog ID	https://data.austintexas.gov/data.json
Schema Version	https://project-open-data.cio.gov/v1.1/schema
Catalog Describedby	https://project-open-data.cio.gov/v1.1/schema /catalog.json
Harvest Object Id	8a455ccf-3d39-4756-9886-fccfe2df128f
Harvest Source Id	3f699ee8-93f5-40fc-bdb7-133eddeb5a13
Harvest Source Title	City of Austin Data.json
Data First Published	2017-10-25
Homepage URL	https://data.austintexas.gov/d/dx9v-zd7x
Data Last Modified	2018-05-02
Source Datajson Identifier	True
Source Hash	c83577eca42861e53a1cf600068e498ca94d0361
Source Schema Version	1.1
Category	Transportation and Mobility

Sample records from the Accident information Dataframe:



Issue Reported	Latitude	Longitude	Address
COLLISION WITH INJURY	30.120700	-97.635260	13911-14905 FM 812 RD
COLLISION WITH INJURY	30.120700	-97.635260	13911-14905 FM 812 RD
COLLISION WITH INJURY	30.120700	-97.635260	Elroy Rd & Fm 812 Rd
COLLISION WITH INJURY	30.120700	-97.635260	Fm 812 Rd & Elroy Rd
COLLISION WITH INJURY	30.120700	-97.635260	13912 FM 812 RD
COLLISION WITH INJURY	30.115334	-97.695139	S US 183 HWY & S FM 973 RD
COLLISION WITH INJURY	30.115334	-97.695139	S Fm 973 Rd & S Us 183 Hwy
COLLISION WITH INJURY	30.115334	-97.695139	S Us 183 Hwy & S Fm 973 Rd
COLLISION WITH INJURY	30.115334	-97.695139	S US 183 HWY & S FM 973 RD
COLLISION WITH INJURY	30.115334	-97.695139	S Us 183 Hwy & S Fm 973 Rd
COLLISION WITH INJURY	30.115334	-97.695139	S FM 973 RD & S US 183 HWY
COLLISION WITH INJURY	30.115334	-97.695139	9320-9513 S US 183 HWY
Crash Urgent	30.128259	-97.800543	12000 S Ih 35 Nb
Crash Urgent	30.128259	-97.800543	12000 S Ih 35 Nb
Crash Urgent	30.128259	-97.800543	12000 S Ih 35 Nb
Crash Urgent	30.128259	-97.800543	12000 S Ih 35 Nb
Crash Urgent	30.128259	-97.800543	12000 S Ih 35 Nb
Crash Urgent	30.128259	-97.800543	12000 S Ih 35 Nb
Crash Urgent	30.135477	-97.797997	11500 S Ih 35 Sb
Crash Urgent	30.135477	-97.797997	11500 S Ih 35 Sb

b) **<u>Foursquare API data:</u>** The Foursquare API allows application developers to interact with the Foursquare platform. The API itself is a RESTful set of addresses to which one can send requests and get responses. The APIs allows searching for places and users (feedback and ratings), exploring popular places and checking out reviews and images for the places.
In this project we will be using Foursquare places API to identify the available hospitals nearby to a latitude and longitude. This is possible using “explore” call that returns a list of available hospitals.

Foursquare webpage is available at: https://foursquare.com/

3. Methodology

In this project I have covered 2 Use cases / approach to derive results using the same set of source data information.

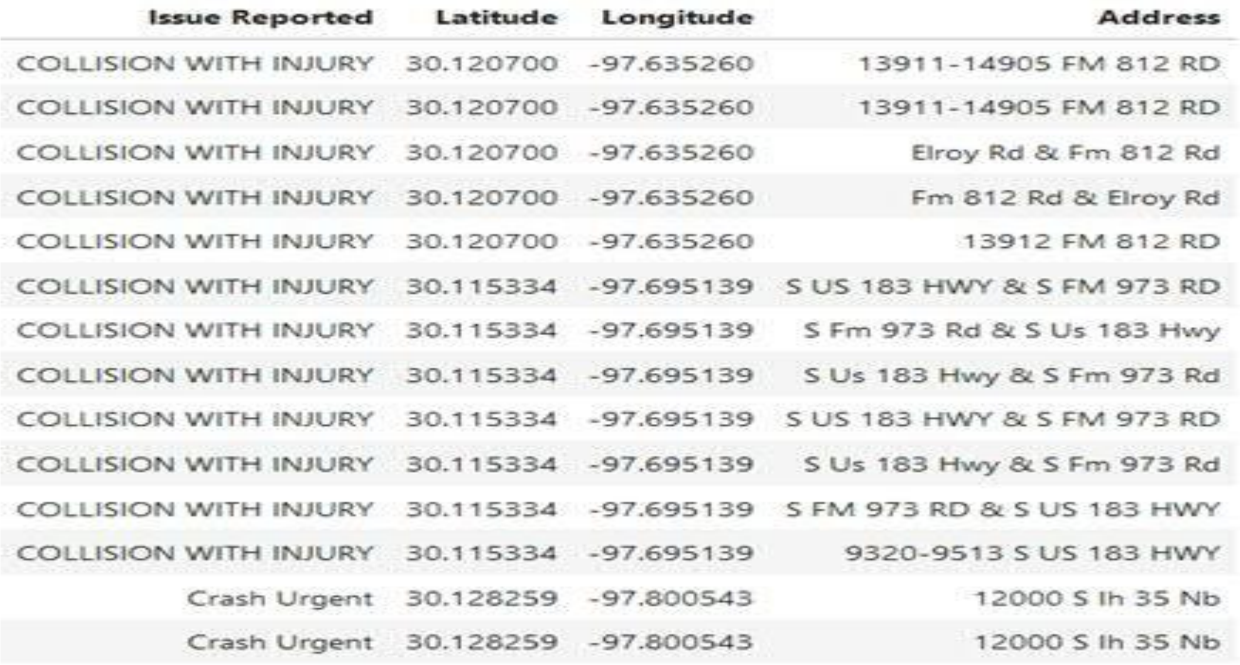
They are:

a) Among all the accident spots identified, taking the highest number of accident spots and identifying hospitals surrounding it.

b) K-means on all the coordinates (Latitudes and Longitudes) of accident areas.

Till half way of methodology for both the Use cases the approach remains the same, so, covering it commonly then I will split them separately.

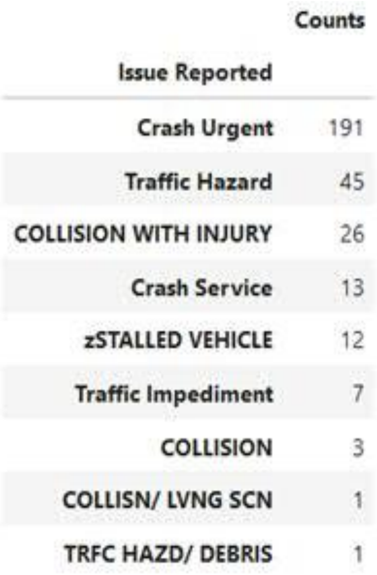
I have collected the data for a span of last 3 months from the data.austintexas.gov and loaded in to a dataframe with the required attributes / columns. Source data file used is Austin_Accident_Data_Final.csv. Sample data as below:



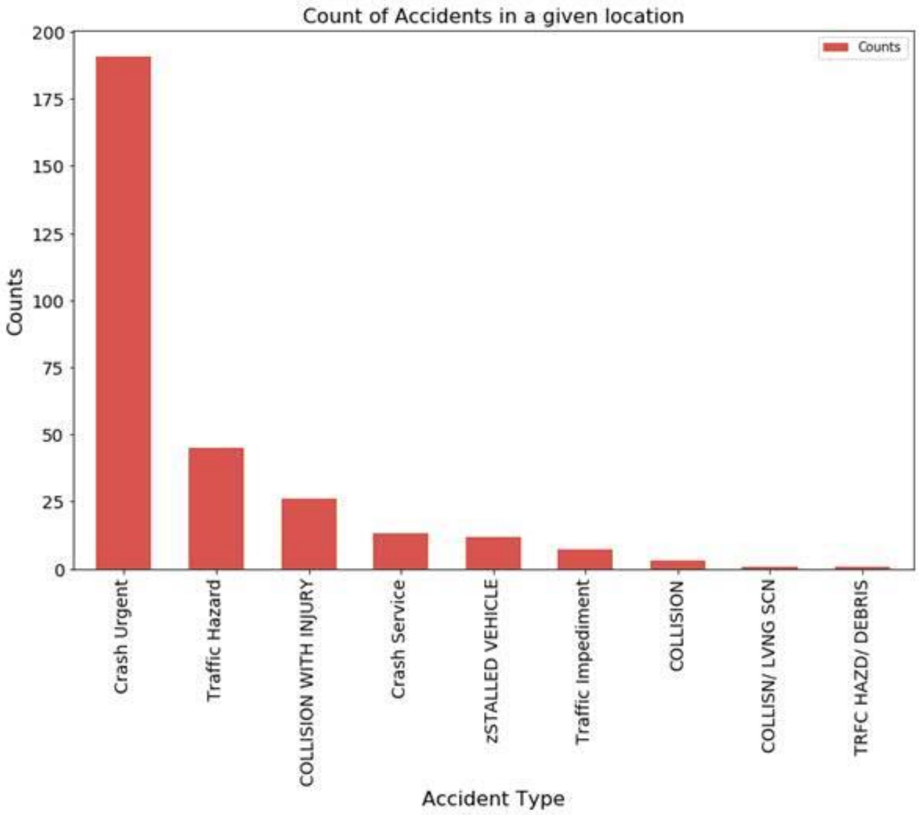
Issue Reported	Latitude	Longitude	Address
COLLISION WITH INJURY	30.120700	-97.635260	13911-14905 FM 812 RD
COLLISION WITH INJURY	30.120700	-97.635260	13911-14905 FM 812 RD
COLLISION WITH INJURY	30.120700	-97.635260	Elroy Rd & Fm 812 Rd
COLLISION WITH INJURY	30.120700	-97.635260	Fm 812 Rd & Elroy Rd
COLLISION WITH INJURY	30.120700	-97.635260	13912 FM 812 RD
COLLISION WITH INJURY	30.115334	-97.695139	S US 183 HWY & S FM 973 RD
COLLISION WITH INJURY	30.115334	-97.695139	S Fm 973 Rd & S Us 183 Hwy
COLLISION WITH INJURY	30.115334	-97.695139	S Us 183 Hwy & S Fm 973 Rd
COLLISION WITH INJURY	30.115334	-97.695139	S US 183 HWY & S FM 973 RD
COLLISION WITH INJURY	30.115334	-97.695139	S Us 183 Hwy & S Fm 973 Rd
COLLISION WITH INJURY	30.115334	-97.695139	S FM 973 RD & S US 183 HWY
COLLISION WITH INJURY	30.115334	-97.695139	9320-9513 S US 183 HWY
Crash Urgent	30.128259	-97.800543	12000 S Ih 35 Nb
Crash Urgent	30.128259	-97.800543	12000 S Ih 35 Nb

Using this dataframe I analyzed which are the main types of accidents happened to identify the level of fatal injuries and plotted a Bar plot for the same. Along with this I also plotted a Bar plot to visualize which areas of Austin are facing high rate of accidents so that I can consider them for my further analysis. Dataframes and their corresponding plots are as follows:

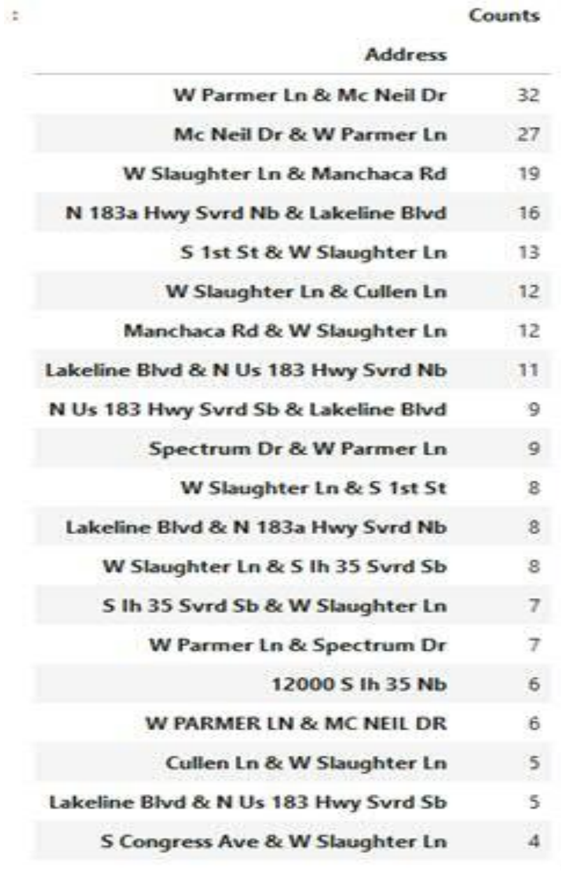
Accident types wise counts and its Bar plot:



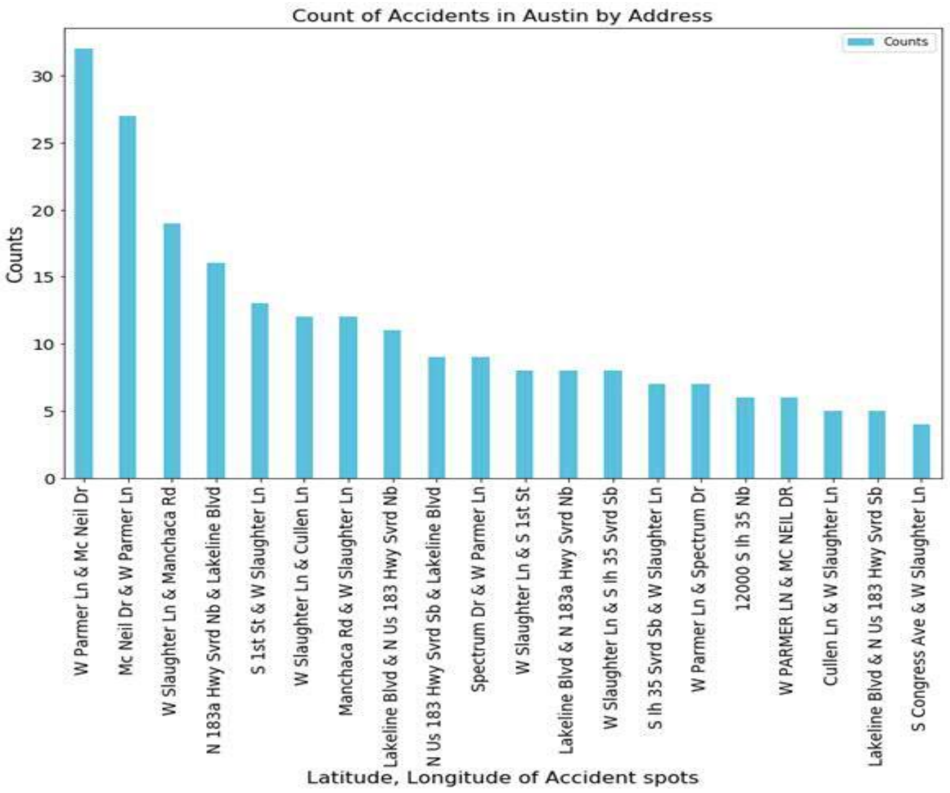
	Counts
Issue Reported	
Crash Urgent	191
Traffic Hazard	45
COLLISION WITH INJURY	26
Crash Service	13
zSTALLED VEHICLE	12
Traffic Impediment	7
COLLISION	3
COLLISN/ LVNG SCN	1
TRFC HAZD/ DEBRIS	1



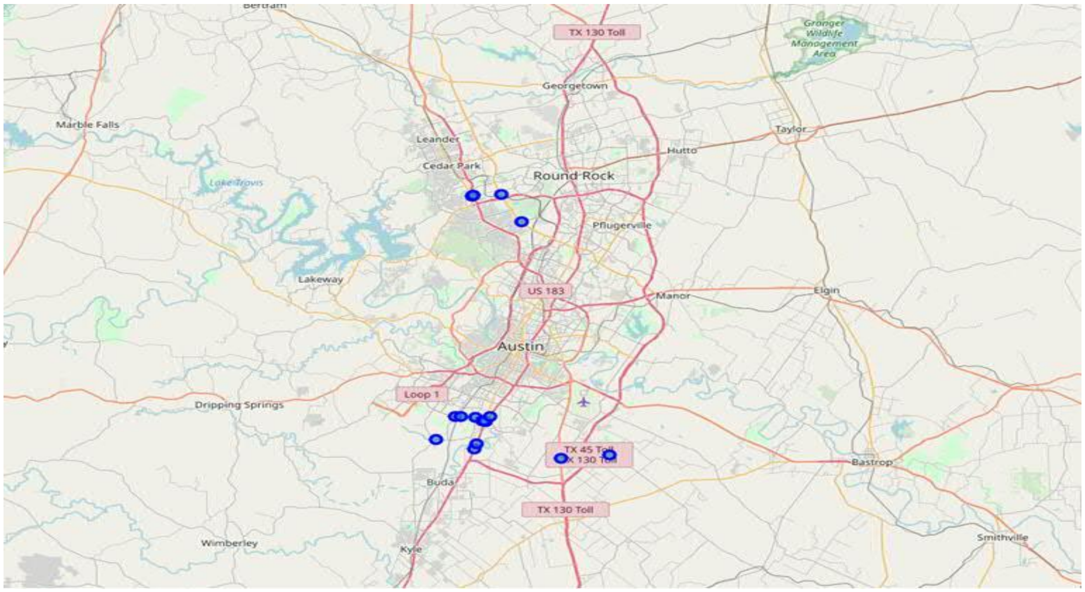
Address wise counts and its Bar plot:



	Counts
Address	
W Parmer Ln & Mc Neil Dr	32
Mc Neil Dr & W Parmer Ln	27
W Slaughter Ln & Manchaca Rd	19
N 183a Hwy Svrd Nb & Lakeline Blvd	16
S 1st St & W Slaughter Ln	13
W Slaughter Ln & Cullen Ln	12
Manchaca Rd & W Slaughter Ln	12
Lakeline Blvd & N Us 183 Hwy Svrd Nb	11
N Us 183 Hwy Svrd Sb & Lakeline Blvd	9
Spectrum Dr & W Parmer Ln	9
W Slaughter Ln & S 1st St	8
Lakeline Blvd & N 183a Hwy Svrd Nb	8
W Slaughter Ln & S Ih 35 Svrd Sb	8
S Ih 35 Svrd Sb & W Slaughter Ln	7
W Parmer Ln & Spectrum Dr	7
12000 S Ih 35 Nb	6
W PARMER LN & MC NEIL DR	6
Cullen Ln & W Slaughter Ln	5
Lakeline Blvd & N Us 183 Hwy Svrd Sb	5
S Congress Ave & W Slaughter Ln	4

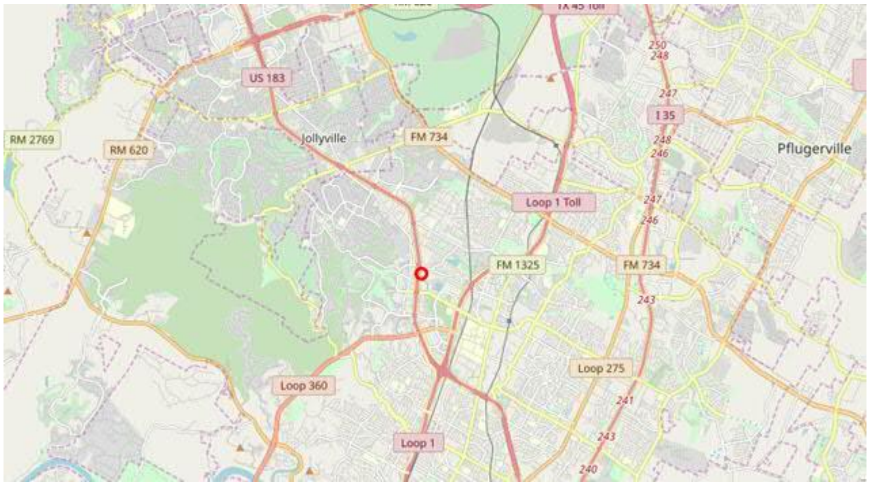


I have also shown the accident area in the map of Austin, Texas for better visualization which is as follows:



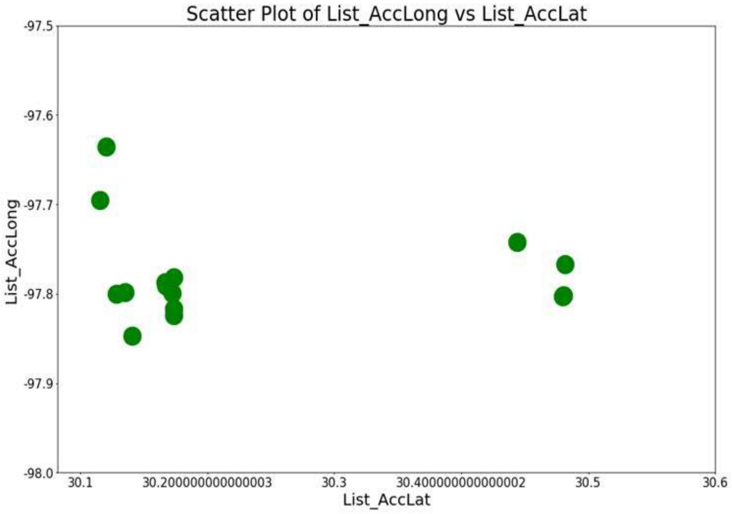
3.1 Use case – 1: Among all the accident spots identified, taking the highest number of accident spots and identifying hospitals surrounding it.

Here, I have found the highest accident coordinate (Latitude and Longitude) among all the accident spots by taking the counts of accidents occurred at each Latitude and Longitude. Now, for these coordinates, I have checked for the number of medical centers available in the radius of 1 km. If there are no medical centers available, then recommending opening a new to the desired user of this application. Based on my analysis of one coordinate, I identified one hospital encircled in red in the map below:



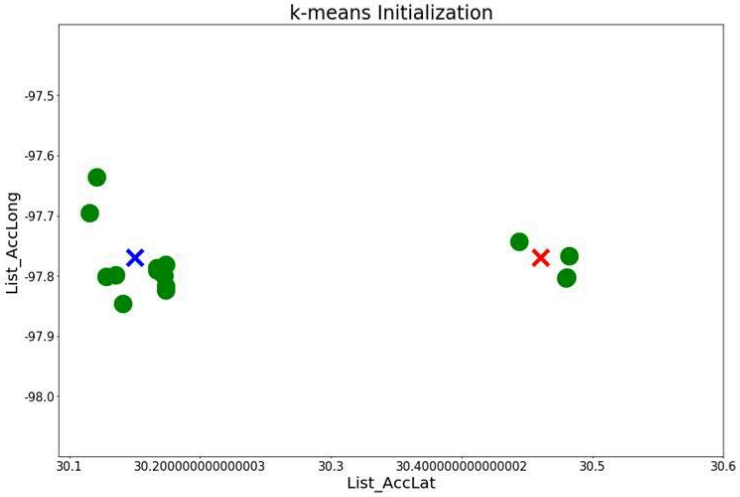
3.2 Use case – 2: K-means on all the coordinates (Latitudes and Longitudes) of accident areas.

Here, in this approach I have taken all the accident coordinates and created clusters for the same and using the K-means method I will be find the précised data points (coordinates) for opening a new Medical center. For this, using the accident information dataframe and transforming the Latitudes and Longitudes to different lists. Then, I defined a function that assigns each data point to a cluster and a function that updates the centroid of each cluster. Using this data set I have initialized the K-mean and plotted a scatter plot between the Latitudes and Longitudes for visualization purpose as below:

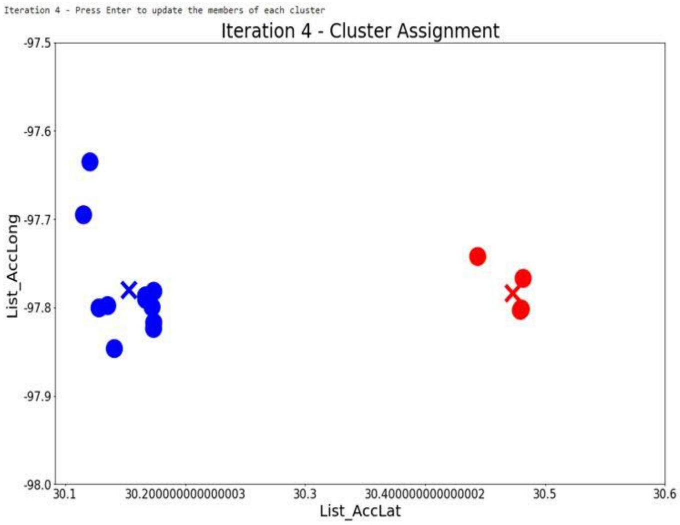


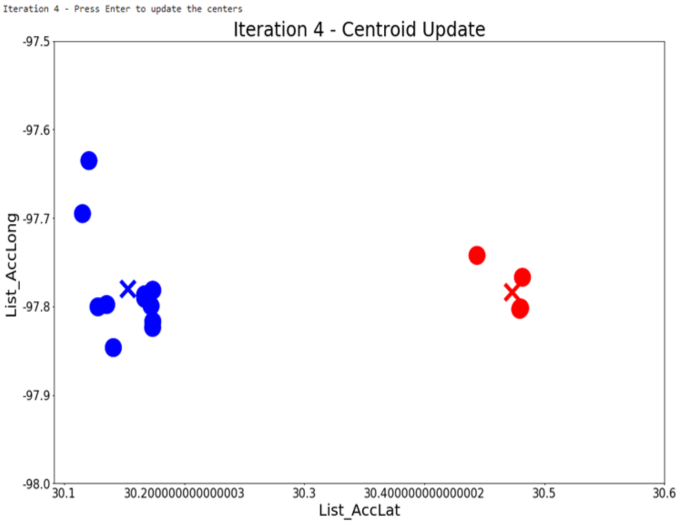
To the above plot, I have added 2 random data points as cluster centers and ran the K-means methodology for 4 iterations.

So, before 4 iterations the clusters and centroid data points were as below:



After 4 iterations, the Cluster assignment and Centroid Update scatter plots are as below:





4. Results

Based on the above 2 use cases results, I can recommend the location where a Medical center can be opened for emergency treatments.

Based on the model of use case 1, I was able to get the location corresponding to the high prone accident zones vicinity where there is a need of medical centers. However, I have considered only the top most accident zone for analysis, but it can be tweaked to consider top 10 or more such locations where there is a need of a new medical centers.

Based on the model of use case 2, I was able to get the Centroid locations for a cluster of high prone accident zones where there is a need of a new medical center. In my date set, I got 2 clusters of data points basis which I got 2 different locations to recommend. If the input dataset has more number of data points in scattered manner, then the model will result as multiple locations where there is a need to new medical center.

5. Discussion Section

The raw dataset which I got from the above-mentioned website of govt. of U.S., there were many columns which I thought was of no use for this analysis, so I decided to drop them to consider in analysis and also there were any such records where partial or incomplete data was present like address, latitude, longitude, so I decided to exclude them as well to consider in analysis.

Another way to make use of this model and its results could be restructuring or improving the zone / location where large number of accidents happens, however, both the audiences / Users and their purposes are different so I opted to go ahead with suggesting opening a new medical center.

6. Conclusion

With the results of the 2 approaches shown above, I would like to conclude that this project can be used as a preemptive measure by the government to identify the need of opening new medical centers so that many deaths happening due to lack of immediate relief to the needful person can be avoided.

7. References

Accident data sources: https://catalog.data.gov/dataset/traffic-reports

Maintainer Email: transportation.data@austintexas.gov

Foursquare APIs : https://foursquare.com/ , https://api.foursquare.com/v2/

Jupyter notebook (actual code): Capstone_Project.ipynb

8. Acknowledgement

As a part of this course curriculum project assignment, I only is involved in building this project, however, I took some inputs from some of my peers who have prior experience in this platform.