CS504 Report : Revere Data Analysis

Abstract :

We are working with The city of Revere to make the city safer by better understanding traffic collisions. We try to use statistical tools learnt from Data Mechanics and Web development skills to analyze 16 years of traffic collision data.

Introduction :

For this project, our goal is to we find the factors that contribute the most to causing traffic collisions by using State traffic collision data from 2002-2018 for the city of Revere.

We focus on using difference in differences technique. Try to figure out the relation between traffic collisions and different factors such as weather, speed limits, natural ambient light, flooding and extreme weather and so on.

And we have made a website to show the result. After using those technique, We draw a conclusion which includes the places where the traffic collisions are most likely to happen, and the patterns behind the occurrences behind the scene.

Datasets :

Weather Condition	Ambient Light	Road Surface	X	Y
Rain	Dark - lighted roadway	Wet	239952.5656	910285.2124
Clear	Dark - lighted roadway	Dry	239672.0104	907624.6692
Clear	Daylight	Wet		
Clear	Dark - lighted roadway	Dry	239532.4905	909013.2871
Clear	Daylight	Dry		
Cloudy/Fog, smog, smoke	Dark - lighted roadway	Wet	240751.406	908019.6877

Revere Crash Data 01-19

This is the main dataset we used. It includes traffic collisions data from 2002 to 2018 in Revere. Typically, the data has weather condition, road surface, and where it happened.

# of Units Added		
Year	# of Units	Difference
2008	24343	
2009	24360	17
2010	24375	15
2011	24379	4

Number of units Added

The additions of new units by year in the data.

Service and Software :
Programming Language : python 3.7
Database : MongoDB
Frameworks : Chart.js, Angular, Flask, Google Map

Technique:

We use python to process the raw dataset and use MongoDB as our database. Based on intuition by visualizing the data. We want to find out the places where car accidents are most likely to happen and find out what factors contributes to car accidents.

We use google map to visualize the car accidents geographical distribution. More specifically, we draw a heat map of the car accidents in Revere city. We let the heat map changes dynamically based on different year, different hour, different month. We use sliding window algorithm to find out 5 place where car accidents are most likely to happen.

Then We use chart.js to draw charts about car accidents. We use a factor as a baseline and then add other factors to draw a new line on the same chart. In order to use difference in differences techniques to find out the influence of each factor.

Finally, we use Flask as a RESTful Server and use Angular and chart.js as the frontend, building up a website to show our result.

Issue encountered :
Issue1 :

After using google map to visualize crash spots, we found several spots in

the car accidents heat map where car accidents are most likely to happen, such as traffic circles. We believe that there are still other spots which we can not discover simply by observing the heat map of car accidents.

Solution :

We abstract this real world problem to a mathematical model. To simplify the problem, we divide the map into grid map and calculate the car accidents happened in each grid.

Then we use a slide window to go throw the grids and find the grid where car accidents are most likely to happen. For example n=5, then we have C_1, C_2, C_3, C_4, C_5 for 5 positions of the slide window.

Issue2 :

The number of injuries depends on the severity of the crash. Here in the dataset, we have columns of data about injuries in every accident, non-fatal and fatal ones. We also have columns about number of cars involved in the accident. We want to find out if there is any relation between them.

Solution :

First, we calculate average and standard deviation of non-fatal injuries, fatal injuries, injuries and number of cars. Then we calculate correlation coefficient between these dimensions.
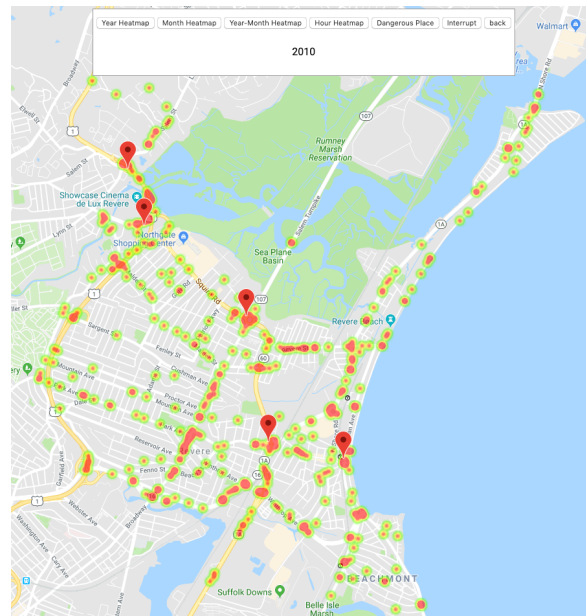
Results :

We have developed a website provide statistical charts and heatmap.
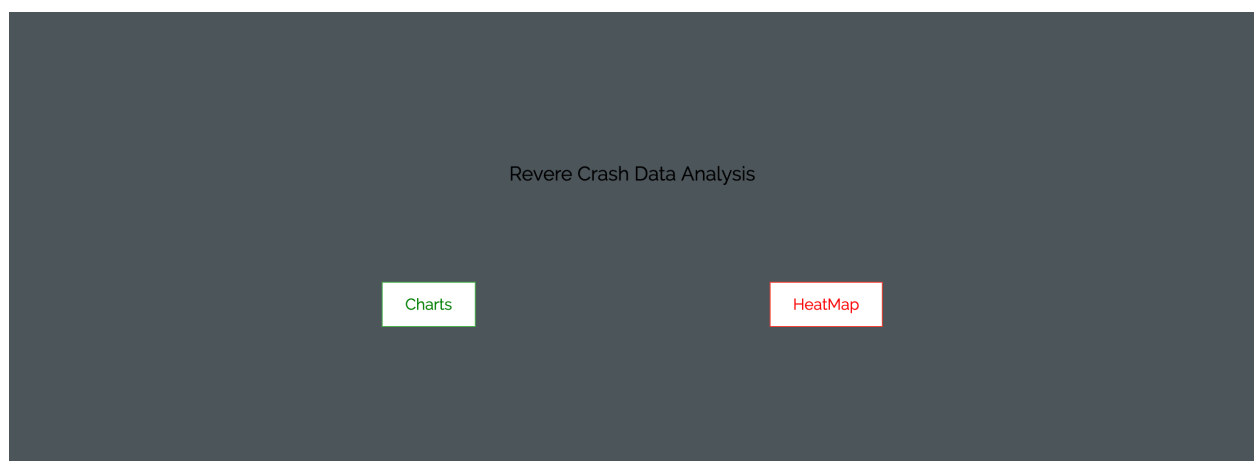
The dataset of all car-crashing spots in revere city and we use google map to visualize these spots. We provide several different view modes for the crash

spots. You can view it dynamically by years, months and hours. You can also find out the places where car accidents are most likely to happen.

A website where you can choose the data you are interested in and then it will show you the result charts. You can select a category as your baseline, and then select different factors in order to figure out how the baseline would be changed by these factors.
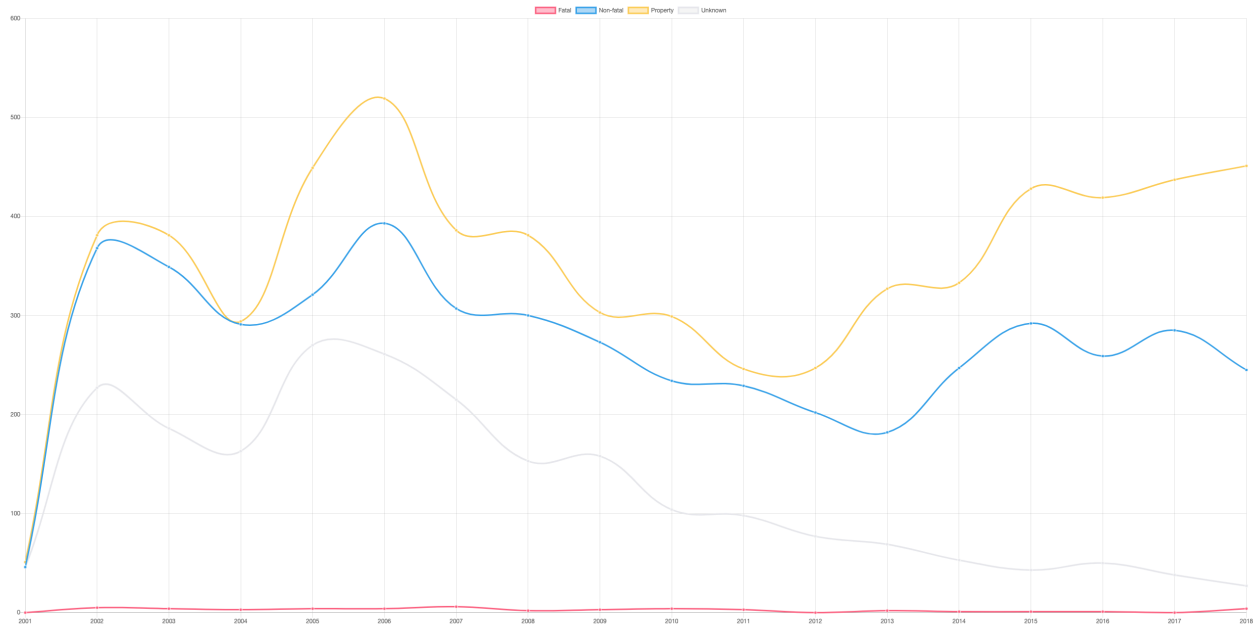
Number of occurrences of traffic collisions in Revere in different years.

Conclusion :

We have made a website which contains all the data from our dataset and visualized those data by charts. You can easily find all the data you are interested in and compare the occurrences of traffic collisions with different factors.

We also draw the heat map of all the occurrences in Revere and found that there was an obviously increase in collisions during commute time in the Broadway. And there Were three places where the traffic collisions are most likely happen which are 1. intersection of Broadway and Squire Road 2. intersection of Beach St and American Legion Hwy 3. intersection of Northeast Expy and Squire Road.

Future Work :

we can keep adding more data to our project to see if there are more interesting patterns we could find.

And since we want to know the patterns behind the data, It should be useful to try some machine learning technique to the datasets and draw some more patterns from it.

For the web-site, we could make it more useful by add more features to it like to predict the possibility of crashes each day in a specific area.