University of California, Los Angeles

Analysis of LA Traffic Collision and Route Finding Application

Bowen Liu

Statistics 199:Directed Research in Statistics

Mentor Professor: Vivian Lew

December 8th, 2019

Summary:

This research consists of two parts: the first section focuses on the analysis of factors that are related to the traffic collision in the great Los Angeles Area, by finding the chronological, regional and climatical patterns of the collisions reported from 2010 to 2019, and investigating the possible reasons underneath; As the second part, this research is aimed to design a route-finding application that provides the users with a faster and safer route, which has the least potential places that are more likely to have the heavy traffic caused by traffic collision.

Introduction:

Los Angeles, as the second-largest city in the United States, is not only famous for its developed entertainment industry but also its annoying traffic condition. Such heavy traffic makes millions of people stuck on their way to work and home. According to Curbed LA, "more than 150,000 people now spend 90 minutes or longer commuting to work in each direction, Between 2009 and 2017, the number of LA County residents with commutes longer than 90 minutes increased by nearly 30,000—a 22 percent spike". It is no denying that traffic collision is not the only reason leading to horrible traffic, it is the worst case to handle especially during the peaks of the common commuting time on the freeway, where most vehicles have to stop and wait for cleaning up. Therefore, the very first purposes of this research is exploring the reasons that might result in traffic collisions using the data from the City of Los Angeles. Some of them might be common sense, but this research will prove them by the data. The second part of this research is helping the user choose the route with the least amount of potential traffic collision spots by building a proper model so that they could reach their destination faster and safer, under

an important assumption that traffic collision is a compound result of multiple external effects such as weather, road condition, and traffic flow.

Part one:

Data Cleaning:

In general, there are 30 variables of this dataset with around half of a million observations and this research only takes care of variables include Date Occured, Time Happened, Location, and Premise Description for the analysis of time and location with high frequency, which will be furthered processed as below:

1. "Date Occur": The first thing to do is transforming the system date format to the real date-time format as we usually see, which consist of Year-Month-Day only; then separating such a combination into variable "year" and "month" using "tidyr" in order to do chronological analysis later; in addition, one variable, named "weekday" was created using "Date Occur" to specify which weekday it was when the collision happened.
2. "Time Happened": The cleaning here is transforming its format such as "7:30" into numeric form with hours kept only(no minutes here), by inserting "0" in front of all single unit-digit hour inputs, like "07:30", and then extracting the first 2 characters, which represent hour, of all observations.
3. "Location": This column is in the form of reviews, which have latitude as well longitude as characters inside. However, some of the latitude and longitude are of 3 decimal (110 meters accurate) while some are of 4 decimals(11 meters accurate), thus in the purpose of

making the model later as accurate as possible, the cleaning process will discard those observations with 3 decimals latitude or longitude.

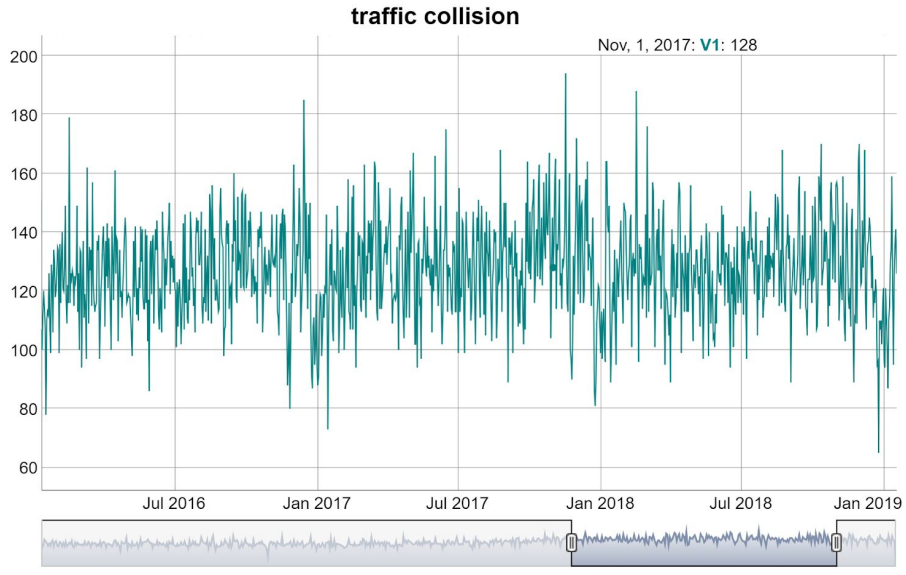
4. “Premise Description”: Traffic collision happened everywhere including driveway, gas station and so on. Since this research has its preference for the route finding as the second part, it only deals with the traffic happened in the street and highway.

Variable Selection:

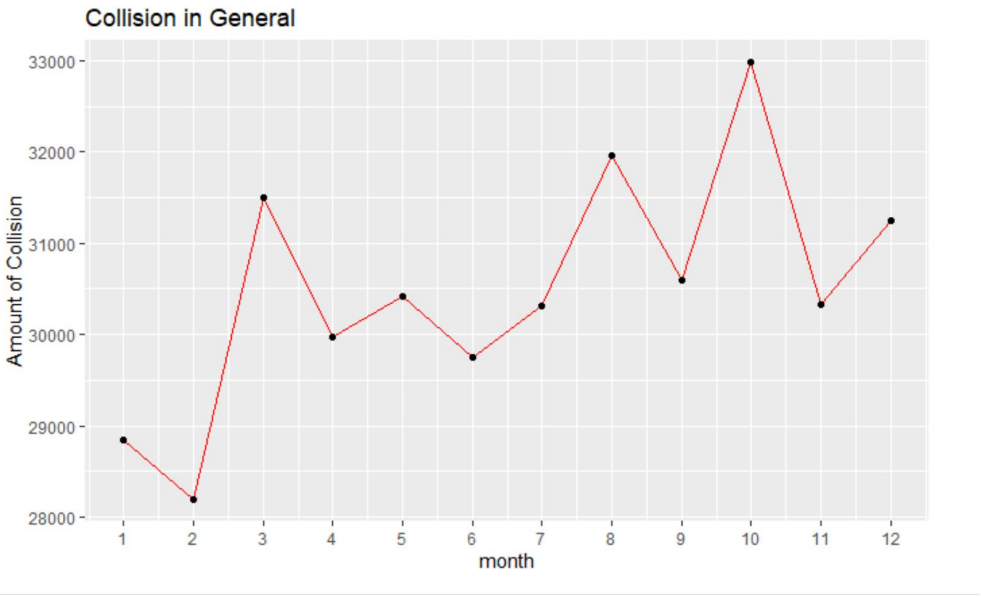
The list of variables that will be used next consists of year, month, weekday, latitude, longitude, Data Occur, and Area Name, which are able to be summarized to observe some pattern since they are scaled and nature-ordered, while other variables such as Victim profiles and premise description are out of the consideration of research due to their obvious uncertainty.

Data Visualization:

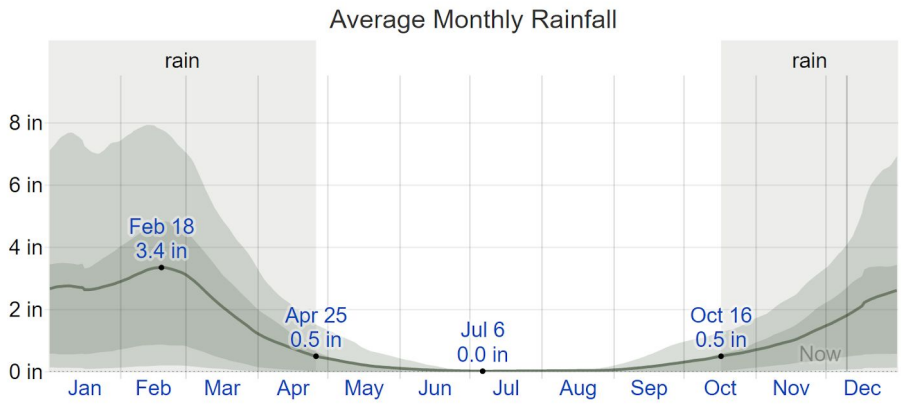
Seasonality:



The graph above is made by “dygraphs” using the variable “Date Occur”, which we can drag the timeline at the bottom to check the trend of the collisions within or among the specific time period. As a result, it is clear the traffic collision has a seasonality in each year, and it will be shown in the unit of month as below:

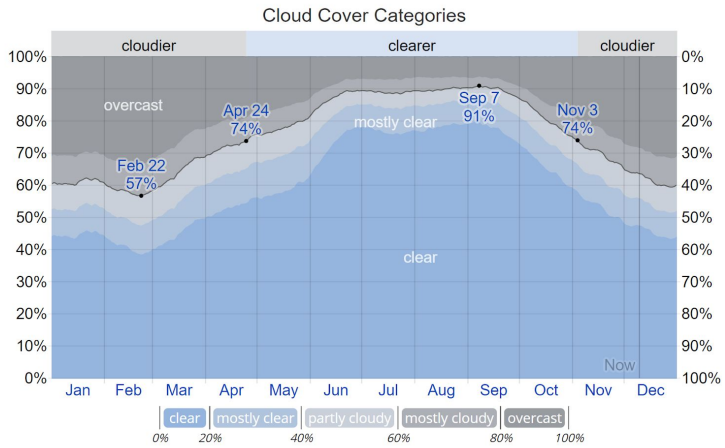


According to this graph, March, August, and October are the local maximum month for the traffic collision. There are some resources from the internet that could explain this phenomenon:



The average rainfall (solid line) accumulated over the course of a sliding 31-day period centered on the day in question, with 25th to 75th and 10th to 90th percentile bands. The thin dotted line is the corresponding average liquid-equivalent snowfall.

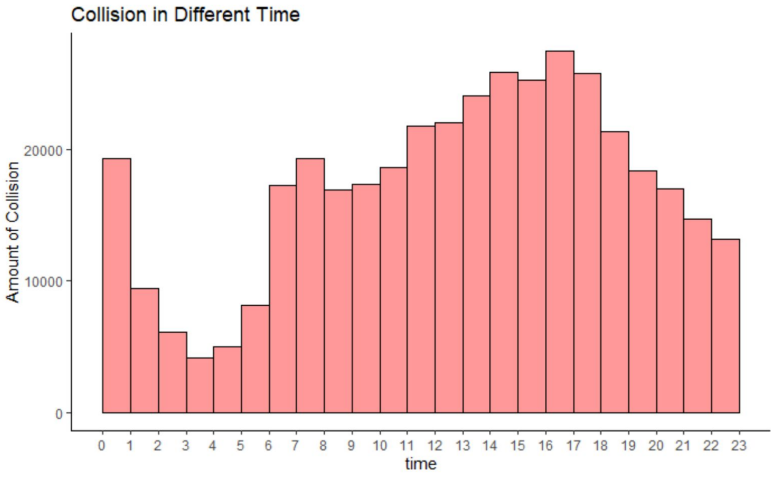
1. The picture above is the record of rainfall from 1980-2016. From the graph, it is clear that around February and March, there were most raining in Los Angeles Area; in addition, Starting from October, sunny weather had over and raining reason started



The percentage of time spent in each cloud cover band, categorized by the percentage of the sky covered by clouds.

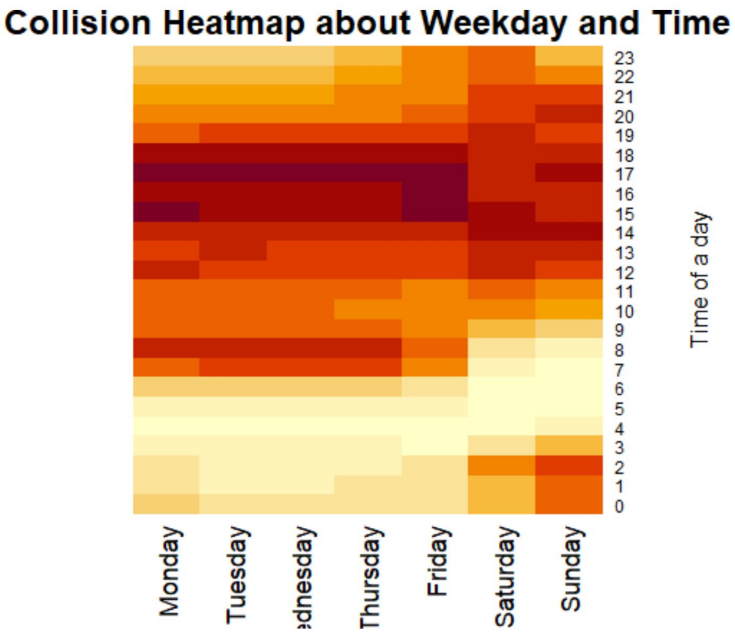
2. The picture above is about cloudiness from 1980 to 2016, it is also obvious that around February and November, the visibility of the weather is the worst.

Difference in Time Happened:



According to this graph, there does exist a difference in collision frequency, in which 7-9 AM and 2-7 PM are the local maximum of traffic collisions, while 3 - 6 AM and 9-11 AM are local minimum. Furthermore, such a pattern is very close to the commuting schedule of most people who need to go to work 8 or 9 o'clock in the morning and clock off around 4 or 5 PM.

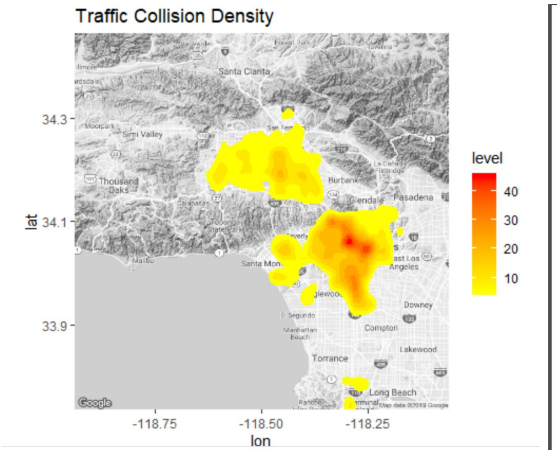
Weekday with Time:

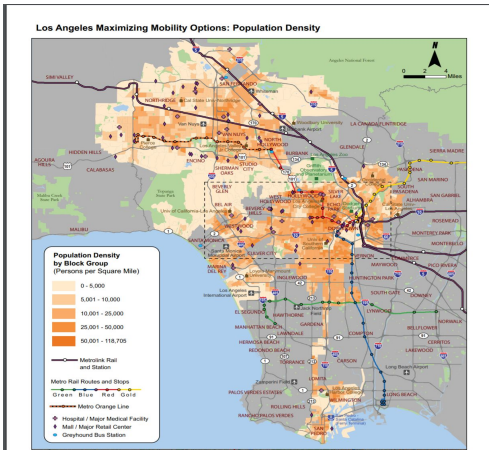


According to this graph, it confirms that the frequency of the traffic collision is related to the commuting time. During the working weekday, it is obvious that between 7 AM and 9 AM as well as 2 PM and 7 PM, collision happened more frequently than those hours before 7 AM or after 8 PM.

On the other hand, one interesting phenomenon is that Saturday and Sunday morning have more collisions compared with other weekdays when most people do not need to work so that they could stay up late at night.

Regional Distribution:





By looking at the population density map of Los Angeles, which is the picture on the left, the places where relatively more collision happened, the density of population is also obviously high.

Conclusion:

The frequency of the traffic collision is related to both climate and human activity: driving under the weather with less visibility or more humidity is more likely to cause traffic collisions. Furthermore, human activity plays an important role in traffic collisions. Places with higher population density are more likely to have traffic collision than those with less; in addition, time during the day when most people commute to work and school has more chance of traffic collisions than those time with few people around; last but not least, weekends have more collision in the morning and night compared with weekdays, but less at daytime.

<u>Part Two:</u>

Basic Idea:

1. Build a model based on the traffic collision dataset and tune the model
2. Find the routes from the start point to the destination, in term of a list of coordinates

3. Use the model to predict which points(coordinates) in each route could have a traffic collision
4. Choose the route with the least number of potential traffic collision points and plot the map

Build Model:

In this section, there will be three steps: Dataset management, model selection, and tuning, which are detailed as below:

Step1:

This is a classification regression which should have positive result 1(collision will happen), and negative result 0(collision will not happen), whereas there are only positive result in this dataset, which became a new variable names “crash”, thus we need to generate some negative dummy variables. On the other hand, since the traffic collision is the type of accident with relatively high uncertainty, generating random samples for every point that has traffic collisions before makes a little sense here.

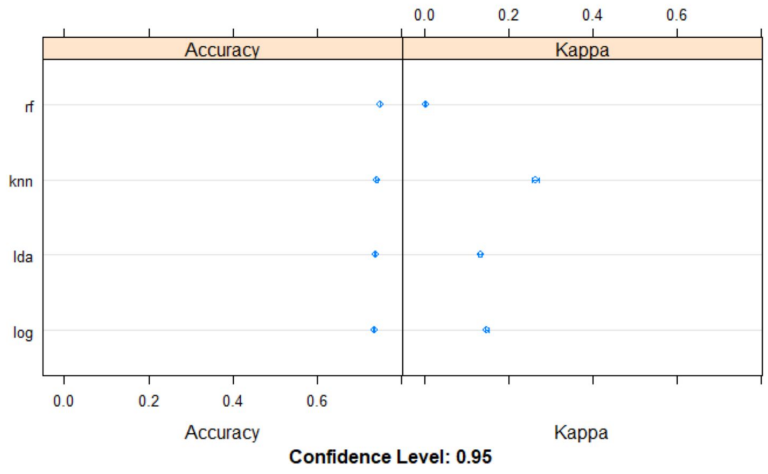
Therefore, the first thing to do here is finding places with the high frequency of traffic collisions as a potential hotspot: according to “Manual on Uniform Traffic Control Devices”, if an intersection of streets has more than 5 times of collisions in a period of 12 months, a signal placed will be placed at that specific place. Therefore, places that have 5 or more times of collision will be filtered out as the potential hotspots; in addition, with respect to the data

effectiveness such as road reconstruction as well as integrity with all months included, the collision happened only in 2018 will be considered.

Secondly, since these potential hotspots are not "active" all the time, we need to generate dummy variables when they are not "active" for testing purposes. The method using here is ***Yuan et al,*** which generates 3 dummy variables at the same location of each hotspot, but at different times, month, and weekday when there is no traffic collision happened, thus these new generated observations all have 0, as a negative result in column "crash".

Step2:

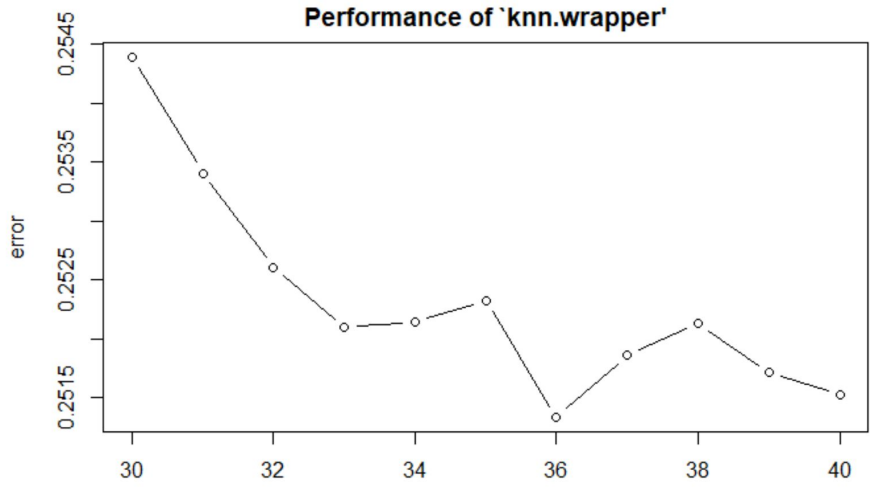
The model considered in this case are LDA, Logistic, KNN, and Random Forest, by running the cross-validation method and comparing their final accuracy, which is shown as below:



In consideration of both accuracy and Kappa value, KNN will be the model to use.

Step 3:

After choosing KNN as the model, the next step is tunning the parameter, which is the number of k. By using the "tune.knn" method from package "e1071", the model will test the standard error of k values vary from 30 - 40 using the bootstrap method. The result is shown below:



According to the graph, k = 36 is the best k value for the least stand error.

Find Route:

The R package used here is **"googleway"**, which can find the routes, based on orientation and destination input, **in terms of coordinates** and plot the route. Two things need to be clarified here:

1. Not all inputs have multiple routes. If there is only one available route, the system will only print out the route map with only potential traffic collision spots.
2. Not all routes have potential traffic collision spots. If there are no such spots, the system will simply choose the one with the least distance.

Run Model:

According to the previous step, there should be a/several list(s) of coordinates available representing a unique route, thus the next thing to do is plugging these coordinates, which consists of latitude and longitude with the month, weekday and time, into the model from step 3, then recording the coordinates with positive result 1 for the using of next step as active hotspots.

Choose Route:

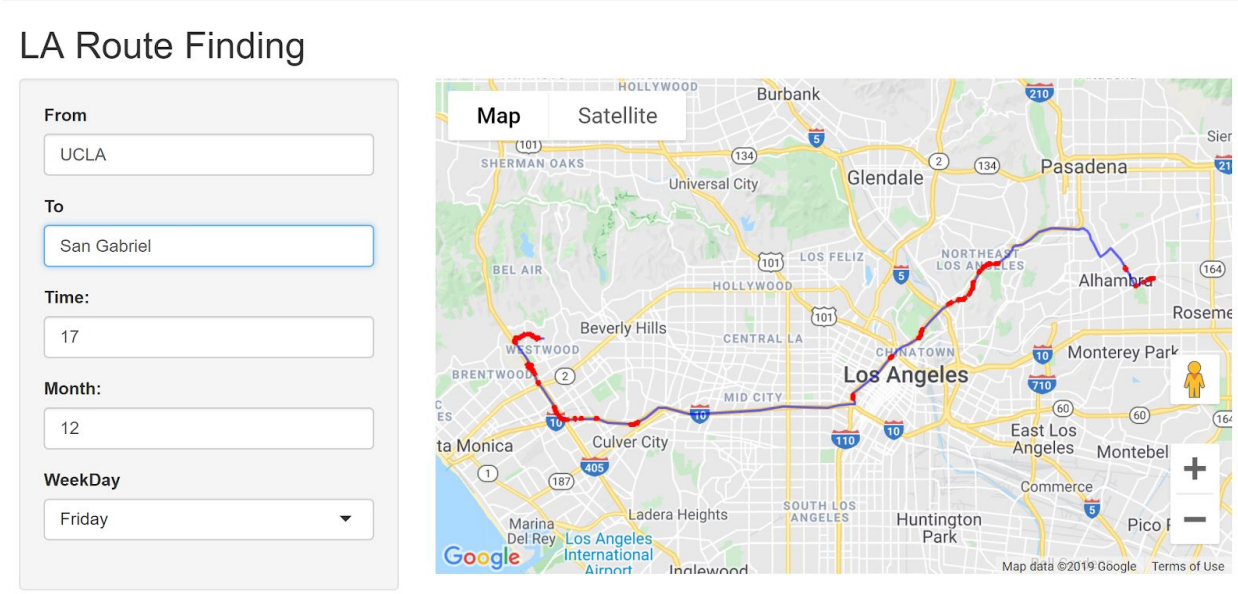
Having the list(s) of the potential traffic collision points of each route from the previous step, there is one more consideration before the comparison of the number of points. According to the previous step, the accurate rate of the model is around 80%, and again, traffic collision is an event with a lot of uncertainty. Therefore, for the purpose of accuracy, the system will perform the method **DBSCAN** to filter those points again. **DBSCAN** is a clustering method which requires two input: least number of observations to consist a group(minPts), and the distance among each observation(eps) within each group. The system set eps to 0.001(Second most accurate coordinate) and eps to 2, which means within 110 meters, if there are at least 2 collision happened, the points within are valid, otherwise, the collision happened at that isolated points will be regarded as coincident and deleted.

In the end, the final comparison could have 3 scenarios:

1. Only one route, no comparison with map output only.
2. Several routes without any potential traffic collision spots, so only compare the distance and pick up the route with the least distance.
3. Several routes with potential traffic collision spots, compare the amount of potential traffic collision spots and pick up the route with the least amount.

Shiny: Route Visualization

Since this is an application designed for the user, it will be more visible for the user to input their demand and view the route easily. Therefore, a Shiny is made as below with the input of Starting Point, Destination, Departing time, Month, and weekday:



The red points are potential traffic collision places in this route, which is very close to what google map provides in a live routing. Therefore, the assumption that traffic collision is a result of multiple conditions is held in this case.

Limitation:

Even though the model and algorithms work pretty well in this case, there are two important improvements that can be made for a better model building:

1. A significant assumption I made in this research: traffic collision is a compound result of multiple situations such as weather and road conditions. However, it is not always true,

thus the related datasets should also be applied in this case in order to make the model more accurate.

2. The range that a traffic collision can effect is unclear. As we can see from the Shiny output above, the red points are potential traffic collision spots with equal effect on the traffic, which is not always true. For example, a traffic collision will definitely cause a longer waiting on a road with 1000 vehicles than one with only 10 vehicles. Therefore, a historical traffic flow dataset is necessary for further study.

Citation

1.Antonio, Meraldo. "Live Prediction of Traffic Accident Risks Using Machine Learning and Google Maps." *Medium*, Towards Data Science, 24 Oct. 2019, https://towardsdatascience.com/live-prediction-of-traffic-accident-risks-using-machine-learning-and-google-maps-d2eeffb9389e.

2.Sarah Zhang. "How precise of one degree of longitude and latitude." GIZMODO, 05 Sep. 2014, https://gizmodo.com/how-precise-is-one-degree-of-longitude-or-latitude-1631241162

3. Weather Spark. "Average Weather in Los Angeles", Cedar Lake Ventures Inc, 31 Dec. 2016, https://weatherspark.com/y/1705/Average-Weather-in-Los-Angeles-California-United-States-Year-Round

4."The Manual on Uniform Traffic Control Devices". Federal Highway Administration, 2003, "https://mutcd.fhwa.dot.gov/pdfs/2003r1r2/mutcd2003r1r2complet.pdf"

5.Elijah Chiland, “Number of LA residents dealing with commute times over 90 minutes surges”, CURBED Los Angeles, 15 Aug.2019, https://la.curbed.com/2019/8/15/20807275/los-angeles-commute-times-traffic

6. “Los Angeles Maximizing Mobility Options: Population Density”, SCAG ™, http://www.scag.ca.gov/Documents/map10-PopulationDensity.pdf