Learning From U.S. Fatal Road Traffic Crash Data

Chun Wang
Department of Computer Science
Stanford University
chunw@stanford.edu

Kexin Yu
ICME
Stanford University
kexinyu@stanford.edu

ABSTRACT

In this project, we deep dive into the fatal car crash records in the U.S. in year 2016, a Google BigQuery public dataset [1] collected by National Highway Traffic Safety Administration (NHSTA) and encoded using the Fatality Analysis Reporting System [2]. By wielding this dataset and research, we develop an interactive document attempting to thoroughly explore the top risk factors that are highly correlated to fatal motor vehicle crashes. We also present interactive tools for readers to explore the full dataset and further understand impact of the identified risk factors on road traffic fatality.

1 INTRODUCTION

Each year fatal motor vehicle crashes in the U.S. lead to an estimated societal burden of more than $230 billion from medical and other costs [3]. Motor vehicle crashes are also the leading cause of death for persons every age from 5 to 32 years old [4]. Identifying the top risk factors that contribute to fatal motor vehicle crashes is important in developing interventions that can reduce the risks associated with those factors and promoting safe alternatives.

Analyzing and visualizing the NHSTA traffic record dataset is a challenging task: the dataset consists of 20 data tables which collectively describe over 400 known attributes of 34,619 fatal motor vehicle crashes. Furthermore, as a motor vehicle crash usually results from a combination of factors including vehicles, road users, environment, and the way they interact, it poses challenges to underpin their individual influences in road traffic fatalities. It is our goal to create visual explainers and exploratory tools that help people digest and learn from this dataset more easily.

2 RELATED WORK

There are a few attempts in the past to create interactive visualizations on road accidents data. Yang [5] developed an interactive map visualizing New York City motor vehicle collisions, complemented by a collection of bar charts and pie charts summarizing key statistics. His work pivots some interesting questions to ask from a traffic accident dataset and is typical of the approach used to visualize traffic dataset. Our work is different in that we attempt to create a visual explainer clearly summarizing the takeaways from the visualizations. Similarly, we also propose an interactive map as a tool for users to explore the full NHSTA dataset.

NHTSA's own research offices have also summarized the 2016 fatal accidents data and studied vehicle safety and driving behaviors to reduce vehicle crashes. Their findings are mainly published as fact sheets and research notes. We attempt to create a more interactive and accessible visualization work based on the same dataset. We cite their statistics and safety tips in some of our visualizations.

3 METHODS

3.1 Data Analysis

3.1.1 Dataset Analysis

On a high level, the NHSTA traffic fatalities dataset schema consists of 20 tables (e.g.`accident_2016`, `person_2016`, `vehicle_2016`, `damage_2016`, etc), where each table can be seen as an entity involved in the car crash. All tables contain `consecutive_number`, a data element representing the unique case number assigned to each crash, which can be used to merge information across data tables.

Records in the `accident_2016` table contain high-level description for each fatal car crash that occurred in 2016, such as time and place of the crash, number of fatalities involved, and atmospheric conditions of the crash. The rest of the tables can be divided into two categories:

1. Tables identifying specific entities involved in the crash and describing their behaviors in details. For example, `person_2016` identifies whether a person is a driver, pedestrian or passenger, the persons alcohol usage, seating position and so on; `vision_2016` describes the entities that obscured driver's vision at the time of crash (if any existed). These are the tables that most directly relate describe contributing factors to car crashes and relate to our task.
2. Tables describing the consequences of the crash, such as the violation charged and damaged areas of the vehicle.

After reviewing the content of all 20 tables, we choose to focus our analysis on tables from category 1, mainly: `person_2016` (85.5K rows x 91 columns) and `vehicle_2016` (52.2K rows x 115 columns). Together with `accident_2016` (34.4K rows x 70 columns), these tables collectively give us a good-sized pool of data attributes to be considered as contributing factors in fatal road traffic crashes.

3.1.2 Data Exploration

3.1.2.1 Data Transformation

To reduce noise in data and ensure accuracy, we construct our SQL queries to filter out invalid data points (Examples: latitude not in range [-90, +90], data logged as "Not Reported" or "Not Applicable"). Where applicable we also apply transformation, aggregation and normalization to raw data in the SQL queries. In various analyses such as the exploration of road conditions, rescue delays, we combine data from external datasets (U.S. Census Bureau Data, Homeland Infrastructure Foundation-Level Data, etc.) to deduce the metrics present in the visualizations as needed.

In order to measure fatality risk, we define four metrics based on data schema present in the dataset, and use one of these metrics to represent fatality risk in analysis of correlation with different data attributes:

1. Raw count of fatal car crashes. This metric can be queried from table `accident_2016` and is used to analyze the impact of atmospheric conditions (e.g. time of crash, weather) on fatality risk.
2. Fatal injury percentage among all persons involved in a car crash, i.e. {[number of fatally injured persons involved (injury

```
select state_name,
avg(crash2notification) as crash2notification,
avg(notification2arrival) as notification2arrival,
avg(arrival2hospital) as arrival2hospital,
avg(crash2notification + notification2arrival + arrival2hospital) as total
from
(select * from
(select state_name,

(case when (hour_of_notification - hour_of_crash) < 0 then (24 + hour_of_notification - hour_of_crash) else (hour_of_notification - hour_of_crash) end) * 60
(case when (hour_of_arrival_at_scene - hour_of_notification) < 0 then (24 + hour_of_arrival_at_scene - hour_of_notification) else (hour_of_arrival_at_scene
(case when (hour_of_ems_arrival_at_hospital - hour_of_arrival_at_scene) < 0 then (24 + hour_of_ems_arrival_at_hospital - hour_of_arrival_at_scene) else (hou

hour_of_crash, minute_of_crash,
hour_of_notification, minute_of_notification,
hour_of_arrival_at_scene, minute_of_arrival_at_scene,
hour_of_ems_arrival_at_hospital, minute_of_ems_arrival_at_hospital

from `bigquery-public-data.nhtsa_traffic_fatalities.accident_2016`
where hour_of_crash <> 99 and hour_of_notification <> 99 and hour_of_arrival_at_scene <> 99 and hour_of_ems_arrival_at_hospital <> 99
and hour_of_crash <> 88 and hour_of_notification <> 88 and hour_of_arrival_at_scene <> 88 and hour_of_ems_arrival_at_hospital <> 88

and minute_of_crash <> 99 and minute_of_notification <> 99 and minute_of_arrival_at_scene <> 99 and minute_of_ems_arrival_at_hospital <> 99
and minute_of_crash <> 88 and minute_of_notification <> 88 and minute_of_arrival_at_scene <> 88 and minute_of_ems_arrival_at_hospital <> 88

and land_use_name = "Urban"
)

where crash2notification > 0 and crash2notification < 1000)
group by state_name
```

Figure 1: SQL query for rescue response time by state.

severity = 4)] / [number of fatally injured persons involved in the crash + number of survived persons (injury severity = 0, 1, 2, 3, 5)]} x 100%. This metric can be computed from table `person_2016` and takes into account all types of persons (pedestrians, drivers, other vehicle occupants) involved in a car crash. It is fitted to analyze the impact of road user behaviors on their fatality risk.

3. Fatality percentage in vehicle among all vehicle occupants, i.e.{[fatality in vehicle] / [number of occupants in vehicle]} x 100%) of a vehicle. This metric can be computed from table `vehicle_2016`, and is suitable for measuring the effect of vehicle conditions, such as vehicle age and model year on fatality risk of vehicle occupants.

4. Fatality rate with respect to population. This metric is calculated by combining data from `accident_2016` and external population census data. It uncovers influences of geographical/administrative regions on road traffic fatality.

When presenting the visualizations in our interactive document, we report the exact methodology we use for calculated metrics in the footnote of the visualizations.

3.1.2.2 Exploration Process

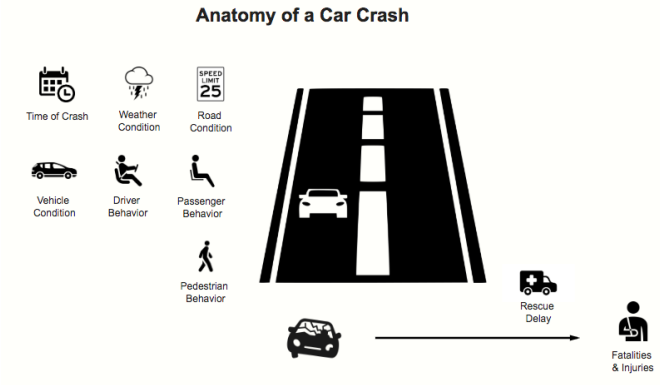

Figure 2: Main risk factor categories.

We then apply exploratory data analysis process on the dataset via Google Colab[1] interface using the BigQuery syntax. For each risk factor category present in **Figure 2**, i.e. Time of Crash, Weather Condition, etc., we query in the dataset for related data attributes, along with data needed for computing fatality risk. These data queries produce results that are readily consumable by Altair[2] to generate interactive plots demonstrating the existence of correlation between the data attributes and fatality risk. We observe the data patterns emerged from these visualizations and look for indications of correlation in the data. When the visualization indicates existence of data outliers (e.g. County of Loving, Texas shows exceedingly high fatality rate, Virginia has exceptionally long crash-to-notification response time, etc.) or trends not immediately explainable (e.g. why does fatal road crashes occur more often during 17-19 p.m. in November and December compared to other times?), we proceed to search for related research, other data attributes or data sources that might help explaining the aberrations, and make our data-backed hypothesis about their causes.

From the exploratory data analysis process, we are able to identify risk factors that are correlated with fatality in motor vehicle accidents. We collect the most interesting insights uncovered from the data for presentation in our interactive document. Particularly, we highlight insights and visualizations that reveal outliers or unusual patterns, or have educational values of promoting safe road user behaviors.

3.2 Visualization Design

We formulate an interactive document structured as a visual essay followed by a section of open exploration. The visual essay loosely follows a scrollytelling structure and consists mainly of interactive plots. The text corpus in the essay consists of cited news stories and research data, our observations and explanations for patterns and outliers shown in the visualizations, as well as actionable takeaways such as safety tips for drivers, passengers and pedestrians.

3.2.1 Navigation Features

We employ various design elements to facilitate reading and navigating through the visual essay, and reduce visual clutters.

As shown in **Figure 3**, after user scrolls to the main analysis section of the document, a *navigation menu* will appear and be locked to the left of the page. The menu will automatically highlight the related category based on user's scroll position. With this menu, user is able to see an overview of the topics discussed at a glance and quickly jump between topics. *Tooltips and hyperlinks* embedded inline are designed to offer additional information or links to external resources (**Figure 4**). *Navigation tabs* are used to group together similar plots and allow users to easily compare the data patterns under different conditions (**Figure 11**).

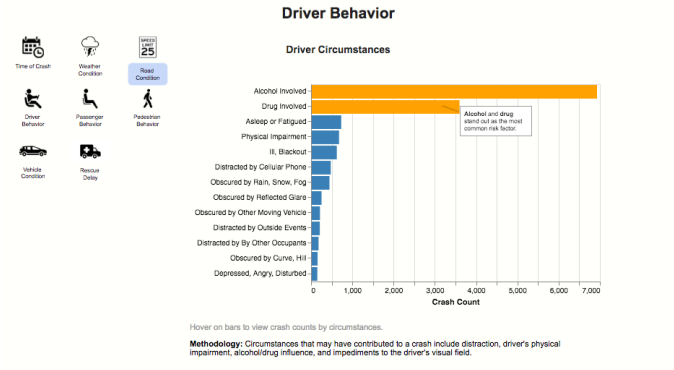

Figure 3: Sticky Navigation Menu

[1]Colab is a collaborative Jupyter Notebook developed by Google. A Google BigQuery client can be easily initiated from within a Colab notebook.

[2]Altair is a Python visualization library built on top of the Vega-Lite visualization grammar. Altair charts can be easily embedded into Web pages using JavaScript.

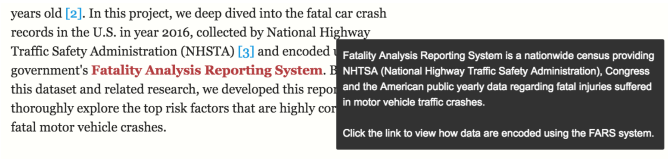

Figure 4: Example Inline Tooltip

3.2.2 Interactive Visualizations

We primarily include interactive plots of the following types to demonstrate the correlation between certain risk factors to fatality risk in road traffic:

- **Normalized stacked bar charts** make it easy to compare percentage difference across multiple data categories. These are primarily used in analyses that use fatality risk definition of type 2 and 3, as discussed in **section 3.1.2.1**.
- **Sorted bar charts** help identify the most significant data elements in a data categories. We apply this type of charts to highlight the prominence of some risk factors.
- **Scatter plots** reveal statistical relationship between two variables. We use these charts to demonstrate strong correlation between some risk factors and fatality in motor vehicle crashes.
- **2-D bubble plots** allow us to incorporate a third data dimension into the visualization. For example, we use them to highlight how fatal car crash counts change over both hour of the day and month of the year.
- **Colored maps** demonstrate data variations across the U.S..

Wherever applicable, we annotate the charts to help readers to quickly identify the key takeaways from the charts. Readers can hover on any data points present in the charts to view tooltips about the data points, which contain exact data values and sometimes useful information from additional data dimensions.

3.2.3 Interactive 3D Map

Due to the sheer volume of the data present in the NHSTA dataset, our analysis is likely to have overlooked some risk factors contributing to fatal car crashes. To overcome this shortcoming, we have designed an interactive map using Mapbox API[3] that allows readers to learn about case details of every fatal car crashes occurred in 2016 and filter records by selected attributes (**Figure 5, 6 & 8**). We include this map in the *Further Exploration* section of our interactive document.

We merge all of the 20 tables in the dataset to generate a comprehensive case report for each fatal car crash occurred in 2016, and map it to a data point on the map at its exact geographic location of crash. Geographic patterns emerge immediately in this map so that users can easily spot regions or roads where fatal car crashes tend to cluster. The map is zoomable and rotatable - when zoomed in to a certain level, the map features a 3D street and buildings view (**Figure 7**) of the U.S. which facilitates reliving the road traffic scenario at which the fatal crash happened. The ultra-zoomed-in view is helpful to visualize additional information about the specific road conditions at the location of crash which may have also contributed to the crash (e.g. a sharp turn that obscures driver's view of incoming traffic). These geographic details are not specified to this level in the dataset and not easily describable with language. Further, we provide user with the ability to search for related records by location of user interest or data attributes of their choice. With the filtered view we display the statistics of how many fatal crashes occurred in 2016 that meet the filtering condition. We hope this map will pique user curiosity in the dataset and make discoveries that tailor to their own needs and interests.

[3]Mapbox is an open source mapping platform for custom designed maps. https://www.mapbox.com/

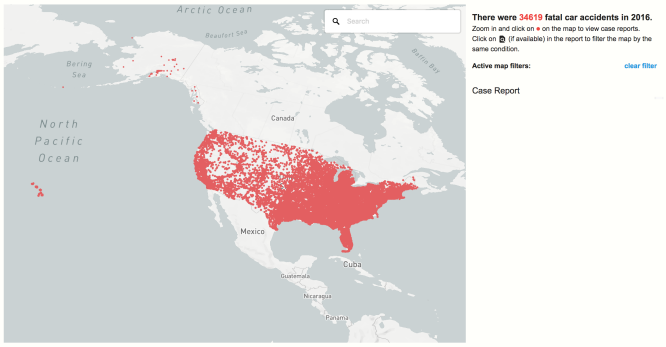

Figure 5: Interactive 3D Map: zoomed-out view with all records plotted

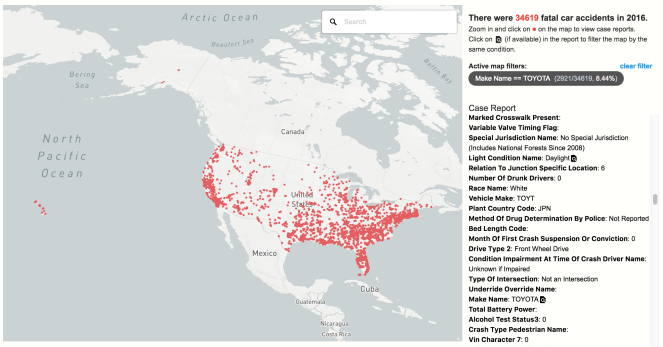

Figure 6: Interactive 3D Map: zoomed-out filtered view on filtering condition 'Vehicle Make == TOYOTA'

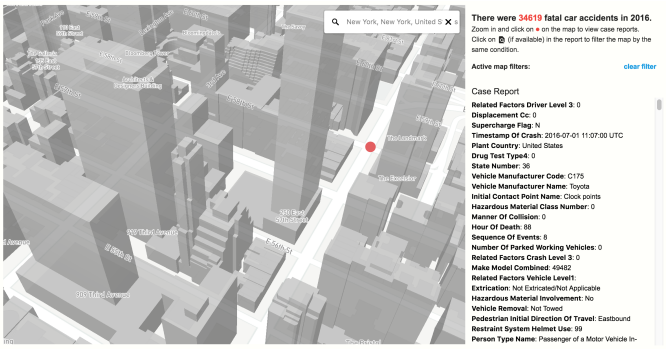

Figure 7: Interactive 3D Map: zoomed-in buildings view

4 Results

The interactive document we developed is accessible online[4] and we encourage reader to interact with it over there. For completeness, we include below the visualizations and analysis we produced for the visual essay.

[4]http://chunw.github.io/carcrash.html

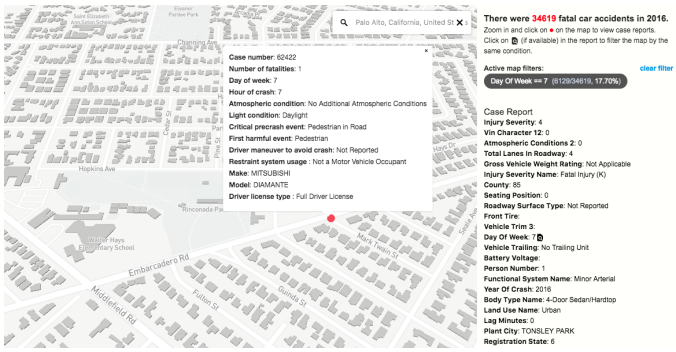

Figure 8: Interactive 3D Map: zoomed-in case report and tooltip view

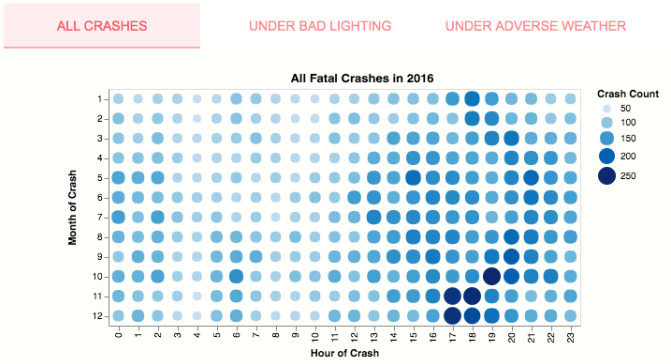

Figure 9: Time Difference. The size and color indicates crash frequency during the given hour in the given month.

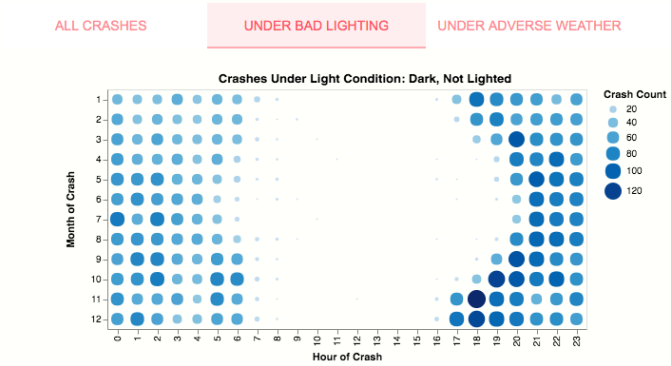

Figure 10: Influence of Light. This plot highlights fatal crashes that happened in a dark and unlighted environment.

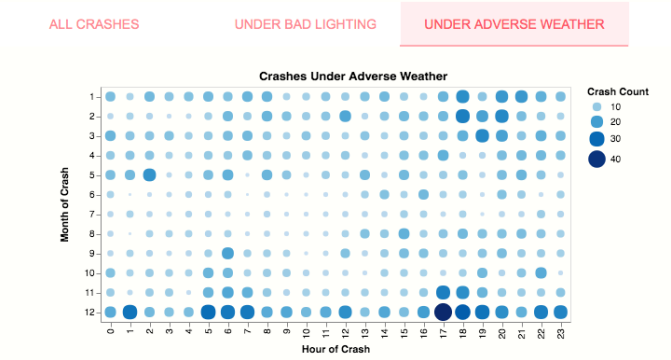

Figure 11: Influence of Weather. Weathers other than clear or cloudy (i.e. rain, sleet, snow, fog, severe crosswinds, or blowing snow/sand/debris) are considered as adverse weathers.

4.1 Time of Crash

Figure 9 shows that fatal crashes mostly happen in rush hours with busy traffic caused by commuting, e.g. from **17:00** to **19:00**. They also tend to occur more frequently in winter months, e.g. **November** and **December**, where adverse weather conditions (e.g. snow on the East Coast and rain on the West Coast) increase risks.

However, we do not observe the same high frequency during rush hours in the morning. As we further explore the level of light that existed at the time of the crash in **Figure 10**, we find that **bad light condition** appears to be a risk factor during evening rush hours highly correlated with fatal car crashes. Due to the changing length of daylight, we see a bell shaped curve: the "dangerous" period starts early and lasts longer in winter, but starts late and lasts shorter in summer. **Figure 11** shows weather-related accidents prevail particularly in **December**.

4.2 Weather Condition

Approximately 21% of all vehicle crashes are weather-related and on average, nearly 5,000 people are killed in weather-related crashes each year (source: Ten-year averages from 2007 to 2016 analyzed by Booz Allen Hamilton, based on NHTSA data). Weather acts through visibility impairments, precipitation, high winds, and temperature extremes to affect driver capabilities, vehicle performance (i.e. traction, stability and maneuverability), pavement friction, roadway infrastructure, crash risk, traffic flow, and agency productivity. **Figure 12** shows that the vast majority of most weather-related crashes happen on wet pavement and during rainfall: 70% on wet pavement and 46% during **rainfall**.

4.3 Road Condition

Figure 13 reveals one notable outlier: **Loving**, Texas, a sparsely populated county with a soaring fatality rate of 3.158%. The culprit is one West Texas highway that is notorious for being dangerous - Highway 285, which runs from New Mexico right on through the small town of Fort Stockton. According to the New Mexico Department of Transportation, there were another three fatal crashes from the Texas state line to Loving in 2017. "*Because there are so many trucks coming off the side roads and you know they don't see everybody and they will just pull off*," said Sherrif Cliff Harris, "*or pull right in front of people sometimes*." It was said the construction zones in Fort Stockton"*are not going anywhere anytime soon*."

4.4 Driver Behavior

4.4.1 Alcohol and Drug Influence

Figure 14 shows that **alcohol** and **drugs usage** are the most common risk factors. **Figure 15** and **Figure 16** confirm that they do highly correlate to fatality risk.

4.4.2 Speeding

Speeding endangers everyone on the road: According to NHSTA, for more than two decades, speeding has been involved in approximately one-third of all motor vehicle fatalities. In 2016,

- Speeding killed 10,111 people, accounting for more than a quarter (27%) of all traffic fatalities.
- 37% percent of all speeding drivers were alcohol-impaired in fatal crashes.

Thus, we are interested in the impact of on fatality risk of the vehicle occupants. From **Figure 17**, there is a clear trend that as vehicle's

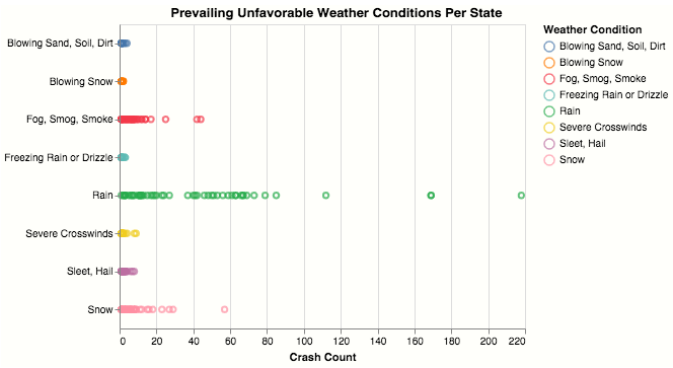

Figure 12: Crashes Under Adverse Weather Conditions. Occurrence of various adverse weather conditions that existed at the time of the crash is counted for each state. Weathers other than clear or cloudy (i.e., rain, sleet, snow, fog, severe crosswinds, or blowing snow/sand/debris) are considered as adverse weathers.

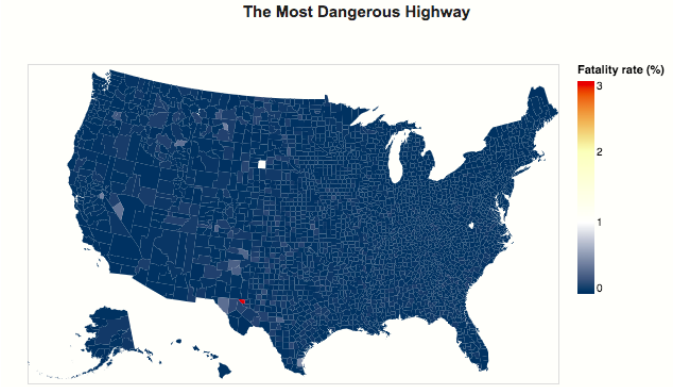

Figure 13: The fatality rate is computed as {[total fatalities in car accidents in year 2016] / [county population]} * 100%.

overspeeding amount goes up, there is also a higher fatality risk associated for vehicle occupants.

Since speeding tickets are the most commonly used tool to deter speeders, we further explore their effectiveness in reducing car crashes related to speeding. Overall, it can be observed from **Figure 18** that speeding citations are effective at reducing the likelihood of receiving subsequent speeding tickets, as significant drops in number of fatal crashes for drivers who have previously received speeding citations. However, while the total speeding-related crash count sees a downward trend, the absolute crash count remains significant, especially for reckless speeders, suggesting that speeding citations have limited effects on deterrence in the context of the current traffic enforcement system.

4.5 Passenger Behavior

4.5.1 Restraint System Use

Of the 23,714 passenger vehicle occupants killed in 2016, there were 11,282 (48%) who were restrained and 10,428 (44%) who were unrestrained at the time of the crashes. Restraint use was not known for the remaining 2,004 (8%) of the occupants. For passenger vehicle occupants involved in fatal crashes in 2016, nearly half of those who were killed were **unrestrained** in the crash, compared to only 14 percent of those that survived. **Figure 19** highlights the reduction in injury severity with restraint system use.

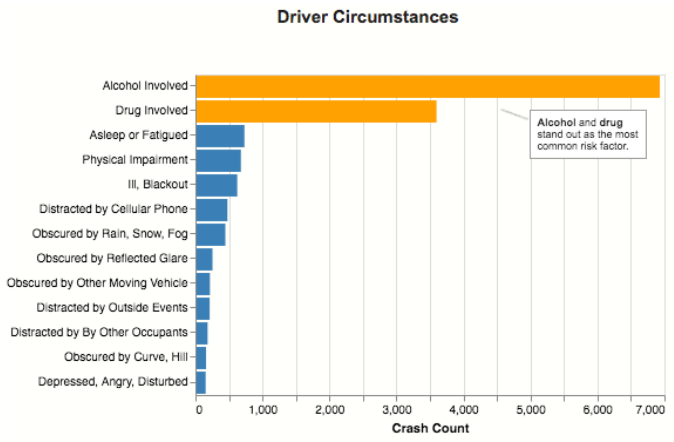

Figure 14: Driver Circumstances. Circumstances that may have contributed to a crash include distraction, driver's physical impairment, alcohol/drug influence, and impediments to the driver's visual field.

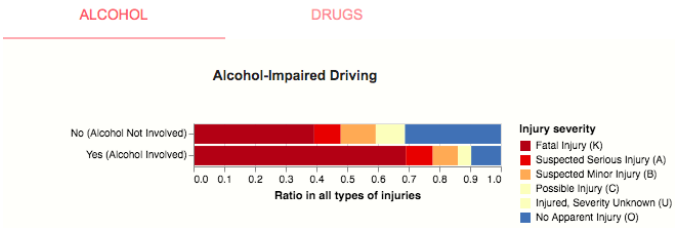

Figure 15: Alcohol-Impaired Driving.

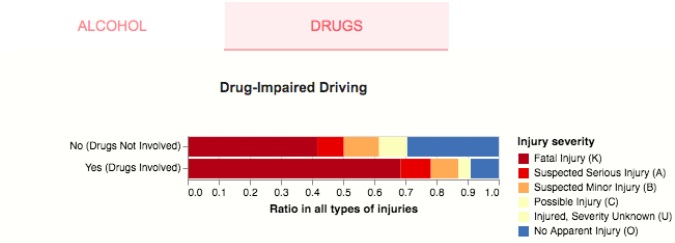

Figure 16: Drug-Impaired Driving

Ejection refers to occupants being totally or partially thrown from the vehicle as a result of an impact or rollover. **Figure 20** shows that ejection from the vehicle is one of the most injurious events that can happen to a person in a crash. Seat belts are very effective in preventing total ejections; in 2016 only 1 percent of all passenger vehicle occupants (those killed as well as survivors) in fatal crashes reported to have been using restraints were totally ejected, compared to 29% of those unrestrained.

4.5.2 Seating Position

Figure 21 shows that among all vehicle occupants, people riding on **vehicle exterior** have the highest fatality rate, followed by passengers sitting in cargo areas or trailing unit, a result from lacking safety protection measures. Inside the vehicle, seats in the **front row (Front Seat, Second Seat positions)** are shown to be more dangerous than seats in the back row (Third Seat, Fourth Seat positions).

4.6 Pedestrian Behavior

In 2016,

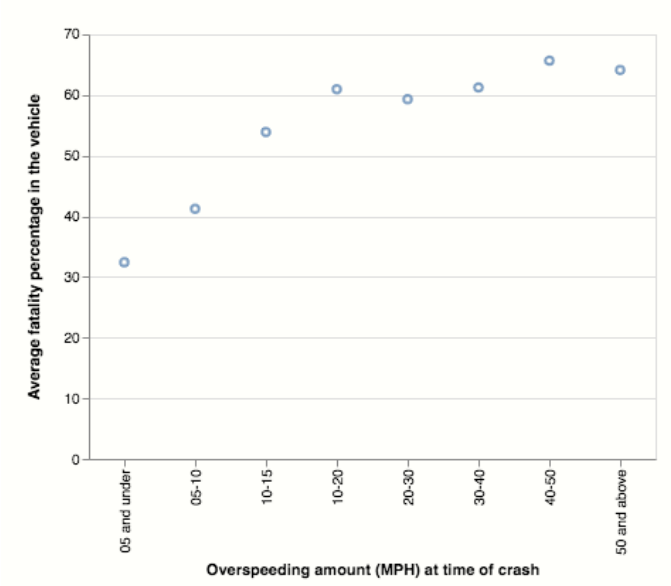

Figure 17: The overspeeding amount of a vehicle is computed as {[vehicle travel speed] - [road speed limit]} at the time of crash. The average fatality percentage for a overspeeding range is computed as {[fatality in vehicle] / [all occupants in vehicle]} * 100% averaged over all vehicles involved in fatal crashes in that overspeeding range at the time of crash.

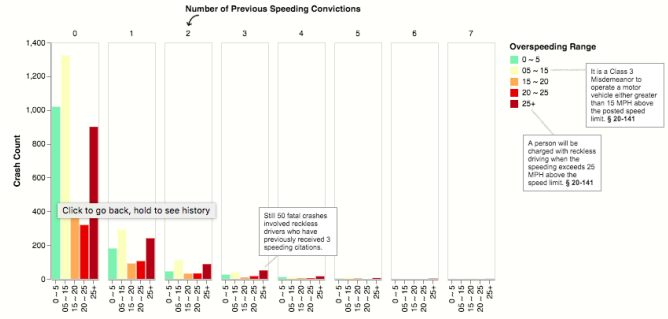

Figure 18: This visualization relates the vehicle's overspeeding amount at the time of crash to the drivers past speeding convictions. We trace the change in ratio of drivers involved in fatal crashes by the number of previous (within five years from the crash date) speeding convictions of the driver.

- 5,987 pedestrians were killed in traffic crashes, accounting for 16% of all traffic fatalities.
- On average, a pedestrian was killed in traffic crashes nearly every 1.5 hours.

In **Figure 22**, we see that **pedestrians**, along with **cyclists** and **persons on personal conveyances**, are one of the road user groups with the highest fatality risks in road traffic crashes.

In **Figure 23**, we show how the number of fatal crashes correlate to pedestrian actions immediately prior to the crash and how the actions relate to the orientation of their collision with respect to the striking vehicle. We find that far more people involved in fatal crashes when dealing with **disabled vehicles** and it's important to know how to remain safe during a roadside emergency.

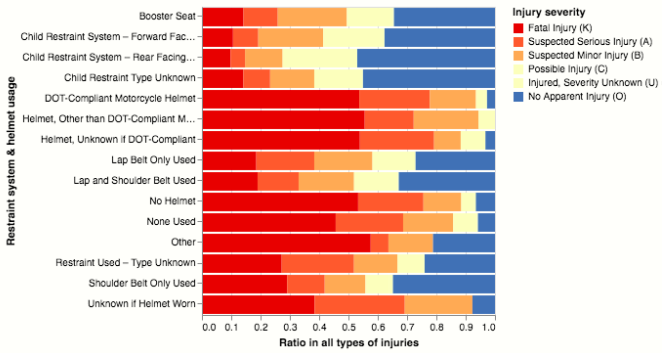

Figure 19: Restraint System Use. The data describes restraint usage of passengers of a motor vehicle in-transport.

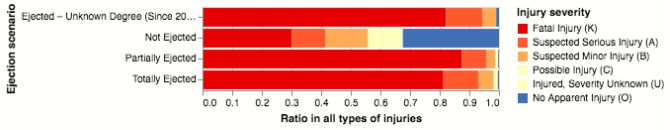

Figure 20: Ejection. The ejection scenario describes the ejection status and degree of ejection for this person, excluding motorcycle occupants.

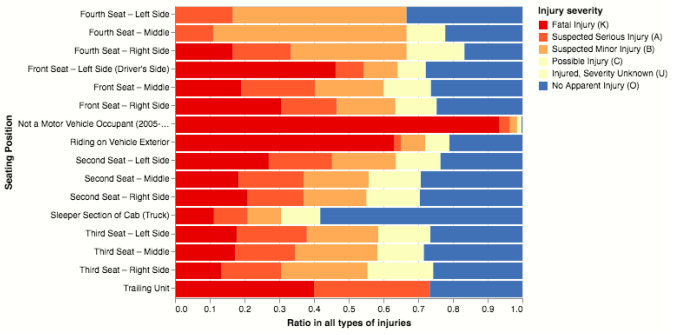

Figure 21: Seating Position.

4.7 Vehicle Condition

4.7.1 Vehicle Age and Model Year

We investigated how passenger vehicle occupant fatality correlates to various vehicle conditions in fatal crashes. **Figure 24** shows that the proportion of occupants who were fatally injured was higher among occupants of **older model year** vehicles as compared to the occupants of newer model year vehicles. **Figure 25** shows that among all passenger vehicle (passenger cars, SUVs, pickup trucks or vans) occupants involved in a fatal crash, the proportion who were fatally injured increases with **vehicle age**.

This suggests that the improved quality of cars is a bit of a double-edged sword when it comes to the age issue. As cars get better on multiple fronts, such as safety technology and crashworthiness, they also become more reliable and longer lasting, leading to Americans driving cars until they're much older and pushing the average age of a car in the U.S. up to 11.6 years at last count. Despite these gains in car quality, the older a car gets, the less reliable and less safe it becomes.

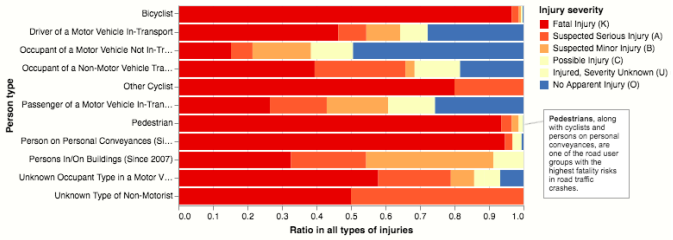

Figure 22: Person Type and Injury Severity.

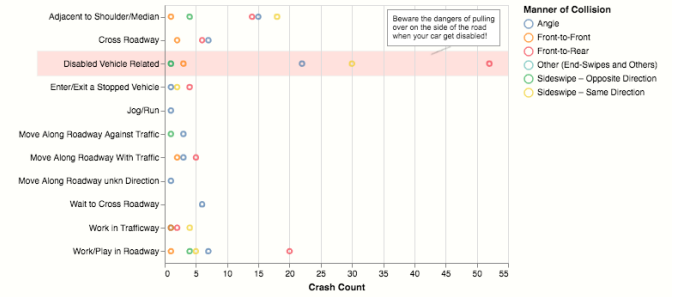

Figure 23: Manner of Collision.

4.7.2 Airbag Deployment

On the surface, the data in **Figure 26** deviates from our expectation that airbags deployed are associated with lower fatality rate. Our hypothesis for this pattern is that airbags are generally designed to deploy only in **moderate to severe crashes**, therefore when the airbags are observed to be deployed, we can assume that the crash is much more severe than when airbags are not deployed. If factoring in crash severity, airbags may have been effective in lowering the fatality risk of vehicle occupants in severe crashes. In fact, NHTSA estimated that 2,756 lives were saved by frontal air bags in 2016.

It is important, however, to follow guidance on how to safely position ourselves to prevent injury from air bags in a crash. Generally, when there is a moderate to severe crash, a signal is sent from the air bag system's electronic control unit to inflate the air bag within the blink of an eye (less than 1/20th of a second). Because air bags deploy very rapidly, serious or sometimes fatal injuries can occur if the driver or passenger is too close to or comes in direct contact with the air bag when it first begins to deploy. Side-impact air bags inflate even more quickly since there is less space between the driver or passengers and the striking object, whether the interior of the vehicle, another vehicle, a tree, or a pole.

4.7.3 Vehicle Body Type

We have also studied which vehicle body types render their occupants vulnerable to fatality risk in road traffic crashes. It was observed that fatality risk varies widely based on vehicle body types, and notably, **motorcycles (of all types) and snowmobiles** occupants suffer higher fatality risk by a large margin. **Trucks, vans, and buses** have some of the lowest fatalities.

4.8 Rescue Delay

In **Figure 27** and **Figure 28**, "crash2notification" tracks the time from the crash occurred to the emergency medical service was notified. "notification2arrival" tracks how long it took for the emergence medical service to arrive on the crash scene."arrival2hospital" tracks how long it took for the emergence medical service to transport victims of the crash to the treatment facility. The time is averaged over all crashes that happened in the urban area of the given state.

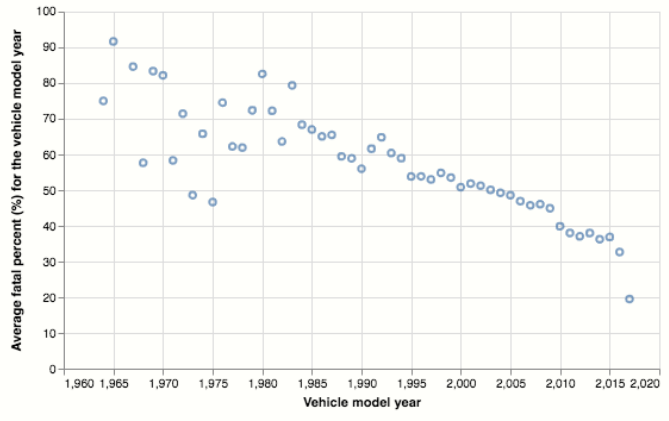

Figure 24: The average fatality percentage for a vehicle model year is computed as {[fatality in vehicle] / [all occupants in vehicle]} 100% averaged over all vehicles of that model year involved in fatal car crashes.

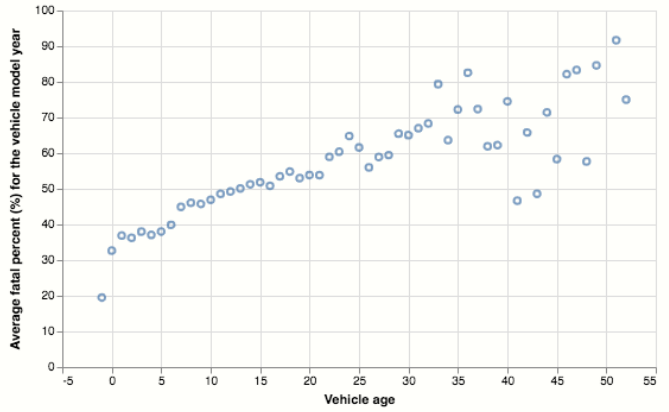

Figure 25: The age of a vehicle is measured by subtracting the vehicle model year from the calendar year at the time of the crash (vehicle whose age was calculated to be -1 was recoded to be age 0). The average fatality percentage for a vehicle age is computed as {[fatality in vehicle] / [all occupants in vehicle]} * 100% averaged over all vehicles of that vehicle age involved in fatal car crashes.

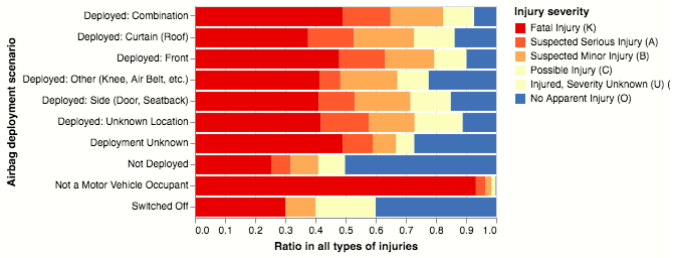

Figure 26: Airbag Deployment.

In **Virginia, Kansas and Vermont**, there existed noticeable delays between crashes and notification. In **Oklahoma**, the emergency service reacted relatively slowly. In **Washington**, transferring victims from the scene to the treatment facility was a long trip; the state also has the longest total response time.

Combining Homeland Infrastructure Foundation-Level Data

(HIFLD) and U.S. Census Bureau data, we further relate response time to the per-capita number of emergency medical service (EMS) stations which provide ambulances. In general, the response time **decreases** as the number of EMS stations increases and we immediately see that Washington has low number of EMS stations per capita. The HIFLD data also includes details on the facility conditions, such as the number of ambulances, the number of EMS members, level of professionalism etc., which are all factors that potentially contributed to the delay. We also found out that **Othello**, the city that reacted the slowest in Washington has no EMS station records according to the HIFLD data.

5 Discussion

Readers of the interactive document we create will learn about risk factors correlated with fatal road traffic crashes and safer alternatives. Some factors, such as speeding and alcohol or drug impaired driving, contribute to the occurrence of a collision and are therefore part of crash causation. Other factors, such as old vehicles and unused restraints systems, aggravate the effects of the collision and thus contribute to trauma severity. Some causes are immediate, but they may be underpinned by medium-term and long-term structural causes as with the case of Highway 285 and states with emergency response delays. Helping readers to identify these risk factors is an important step towards promoting the interventions via interactive visualizations that reduce the risks associated with those factors.

Readers will also benefit from exploring the fatal motor vehicle crash records in a 3D geographic map view. The geographic details are better contextualized and visualized with the map, while detailed reports and the ability to search records by location and attributes allow readers to sympathize with the cases and direct their own researches.

As a motor vehicle results from a combination of various factors including vehicles, road user behaviors and atmospheric conditions, it is challenging to isolate the effect of a single factor on fatality risk in road traffic crashes. In the exploratory data analysis process, we have not strictly controlled all other variables when studying the correlation between one variable and fatality risk.

To allow readers perceive the effects of individual factors more easily, we propose a predictive model using BigQuery Machine Learning[5]. The logistic regression model predicts a person's fatality risk (represented as a probability) in a road traffic crash scenario given input features about this person, such as alcohol involvement, seating position, person type and so on. Trained on a set of features selected based on our previous analysis, our best performance model currently achieves an accuracy of 79.38% (**Figure 31**).

We publish our model as a Colab notebook[6] (**Figure 30**). Readers may study, modify and interact with this model after copying it to their Google Drive[7]. Users of this model may change the value of a single input feature to the model and see how much it will affect the predicted fatality risk quantitatively.

6 Future Work

We wish to extend this project in these directions when time permits:

1. Evolve the visual essay by delving into existing stories and discovering new directions, e.g. the impact of holidays/special events, driver age/experience on risks or any insights that disprove common misconceptions.
2. Design quizzes on road traffic risk factors and motivate users to seek answers in our visualization.
3. Incorporate the predictive model into the interactive document for easy access. Also, how might we better capture the way various factors interact with each other, e.g. identify factors that work together to contribute to fatal crashes?
4. Improve the user experience of the open-ended exploration

[5] See https://cloud.google.com/bigquery/docs/bigqueryml-intro for more details

[6] Accessible at https://colab.research.google.com/drive/1ukF799LA-K9qPdf2D89N8mM8UDu7RJK4scrollTo=ITTmbrIIdFOY

[7] Due to permission restriction imposed by Google, users will need to copy the notebook into their own Drive and run the model from there.

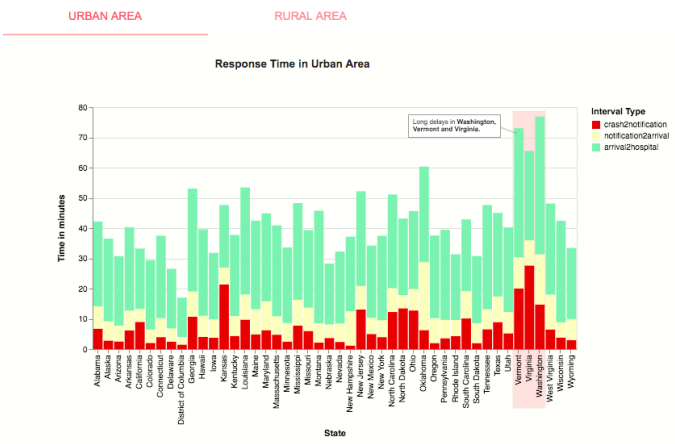

Figure 27: Rescue Delay in Urban Area.

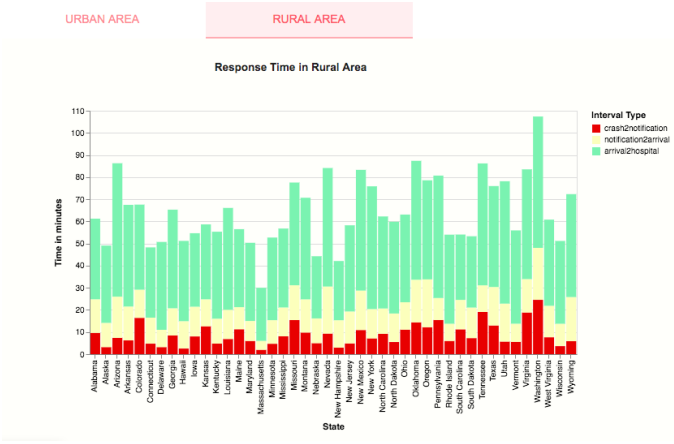

Figure 28: Rescue Delay in Rural Area.

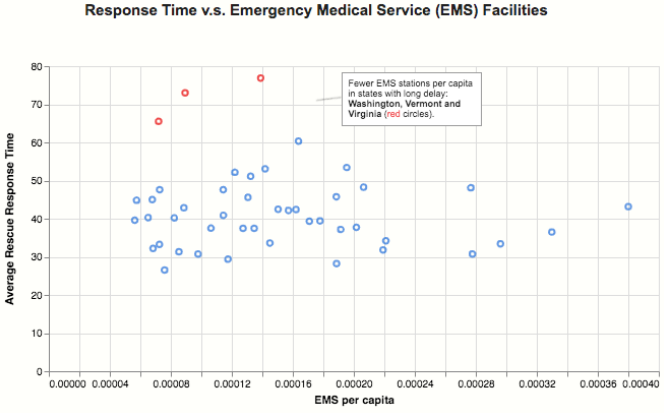

Figure 29: Importance of Emergency Medical Service Facilities.

Figure 30: The Colab notebook interface to view and interact with our model.

precision	recall	accuracy	f1_score	log_loss	roc_auc
0.838045	0.437928	0.79381	0.575252	0.494914	0.765963

Figure 31: Predictive Model Performance Statistics

map. Allow user to filter on all data attributes appearing in case reports[8]. Conduct usability tests to find out how readers would like to interact with the map. What can we do better to facilitate their exploration?

ACKNOWLEDGMENTS

The authors wish to thank the instructor Maneesh Agrawala and the course assistant Gracie Young, Vera Lin for their valuable guidance throughout the quarter.

REFERENCES

[1] *NHSTA Traffic Fatalities*. https://bigquery.cloud.google.com/dataset/bigquery-public-data.

[2] *Fatality Analysis Reporting System (*FARS*) Analytical Users Manual*, volume DOT HS 812 092. US Department of Transportation, National Highway Traffic Safety Administration, December 2014.

[3] David A. Lombardi, William J. Horrey, and Theodore K. Courtney. Age-related differences in fatal intersection crashes in the united states. *Accident Analysis Prevention*, 99:20 – 29, 2017.

[4] *Motor Vehicle Traffic Crashes as a Leading Cause of Death in the United States, 2015*, volume DOT HS 812 499. US Department of Transportation, National Highway Traffic Safety Administration, February 2018.

[5] *New York City Motor Vehicle Collision Data Visualization*. https://nycdatascience.com/blog/student-works/new-york-city-motor-vehicle-collision-data-visualization/.

[8] Currently, restriction is imposed by Mapbox on number of data dimensions to upload for the map.