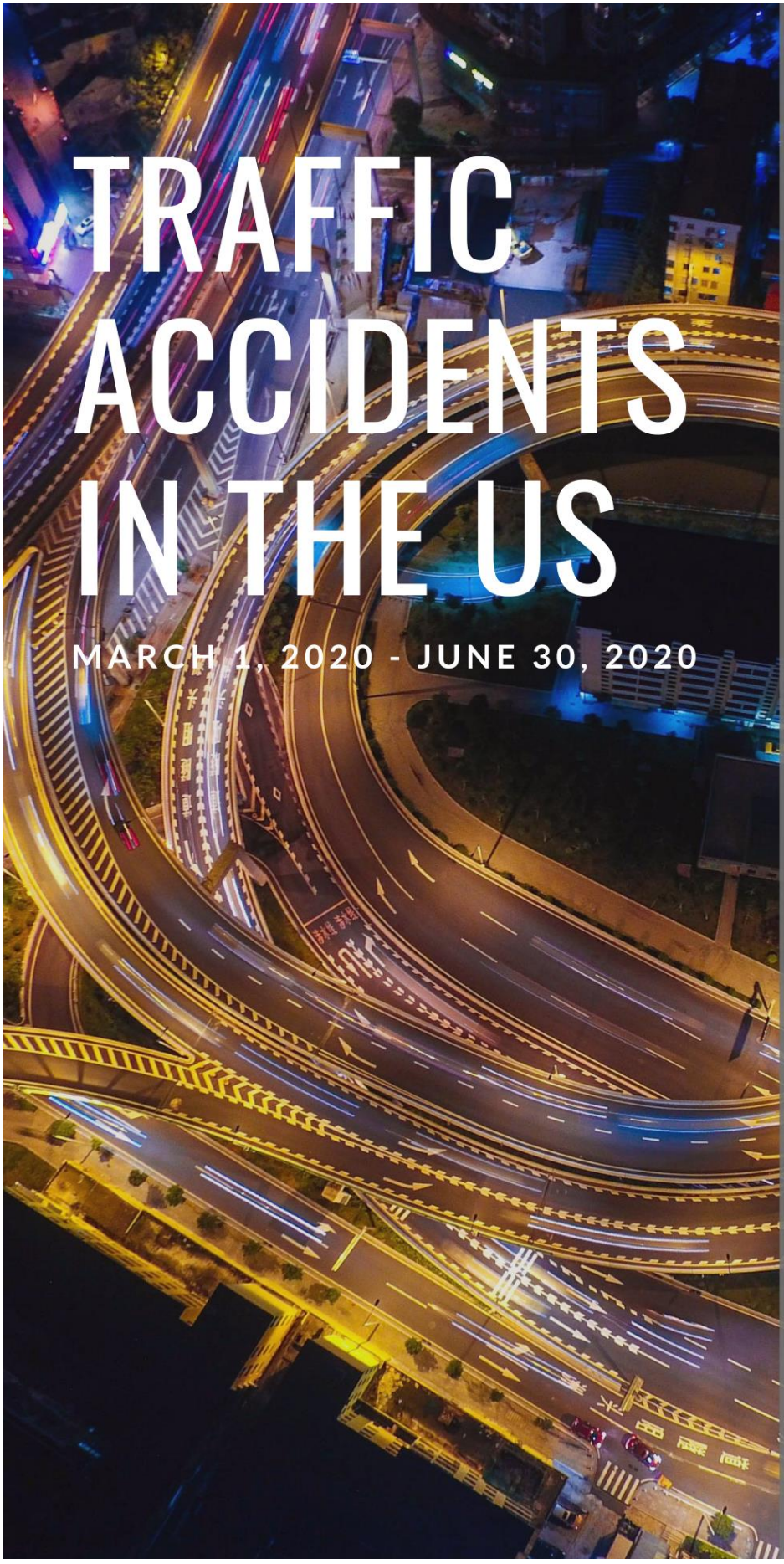


TRAFFIC ACCIDENTS IN THE US

MARCH 1, 2020 - JUNE 30, 2020

David Obembe
Ryan Thomas
Brian Yu
Will Pumphrey
John Dilligard

Contents

Summary

This project was on accidents in the United States of America and required the collection and analysis of road accidents from February 2016 to June 2020. Our initial dataset contained 3.5 million rows of data in csv format and provided detailed information on the location of accidents, time of accidents, weather conditions, severity etc. Our efforts as a team were spent on cleaning and filtering this data as well as using it to plot visualisations that reveal accident trends.

The tools used for the project were Pandas (a Python Library), JavaScript's Leaflet (for creating maps), Plotly JS, D3, HTML and CSS.

1. Introduction

1.1 Background

Motor vehicle accidents continue to be one of the leading causes of accidental deaths and injuries in the United States. They are responsible for billions in property damage and other economic losses each year [1].

Accidents on the road happen for many reasons. The national Highway traffic Safety Administration (NHTSA) released reports that fatality rates for various types of accidents have increased in recent years. They have attempted to identify the major causes of accidents and these were a few of their findings:

- As of 2016, alcohol related accidents accounted for roughly 30% to 40% of fatal car accidents in the USA.
- Speeding contributes to one in three of fatal accidents.
- Aggressiveness and reckless driving also contribute to 1 in 3 accidents.
- About 90 people die every day in the U.S from vehicle accidents.
- Almost one-fourth of all fatal accidents involved a driver under the age of 25.
- Drivers reported distracted driving in about one of every five vehicle accidents in the U.S. in 2017.

Regardless of all the information and driving tests that qualify people to drive a car, accidents will continue to occur due to human negligence and errors. Our project goal was to identify the States with the most negligent drivers, the streets that are most accident prone (so that the reader may avoid), the weather conditions that cause accidents and the time of day with the most accidents.

1.2 Objectives

- Create a dynamic map of the U.S showing locations where accidents occur and the number of accidents that occur in that region—filtered by month (from March to June 2020) and by severity.
- Display stacked bar charts that compare number of accidents by State as well as visualises the effects of Covid-19 on number of accidents.
- Visualise cumulative accidents by day over the past four months (from March 1st, 2020 to June 30, 2020) on an interactive scatter plot.
- Generate number of accidents over time with a line chart.
- Create a project dashboard that shows visualisations that reveal the time of day with the highest number of accidents and top 10 roads with the highest number of accidents.

2. Hypothesis

We expect that most accidents will occur during the day due to the number of cars on the road, total number of accidents should drop in 2020 due to Covid-19, accidents should be more common in rainy and snowy weather and accidents should be more likely to occur in populated cities and states.

3. Sources of Data

Our CSV was a countrywide car accident dataset, which covered **49 states of the USA**. Accident Data was collected from **February 2016 to June 2020**, using two APIs that provided streaming traffic incident (or event) data.[2] The APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. There were about **3.5 million** accident records in this dataset.

4. Cleaning & Exploration Methodology

Due to the extremely large size of the file that our team dealt with as well as the format of the file (CSV), each individual in the group had to perform their separate data cleaning process specific to the final product of the Individual.

Generally the file was imported into Jupyter notebook and converted into a Pandas data frame, all null fields were removed, date columns were converted into a date type format, time range was selected to further reduce size of file and finally each individual had to create their own separate csv file with Subsets of columns that they needed for their visualisations.

5. Results

5.1 Geographical Map

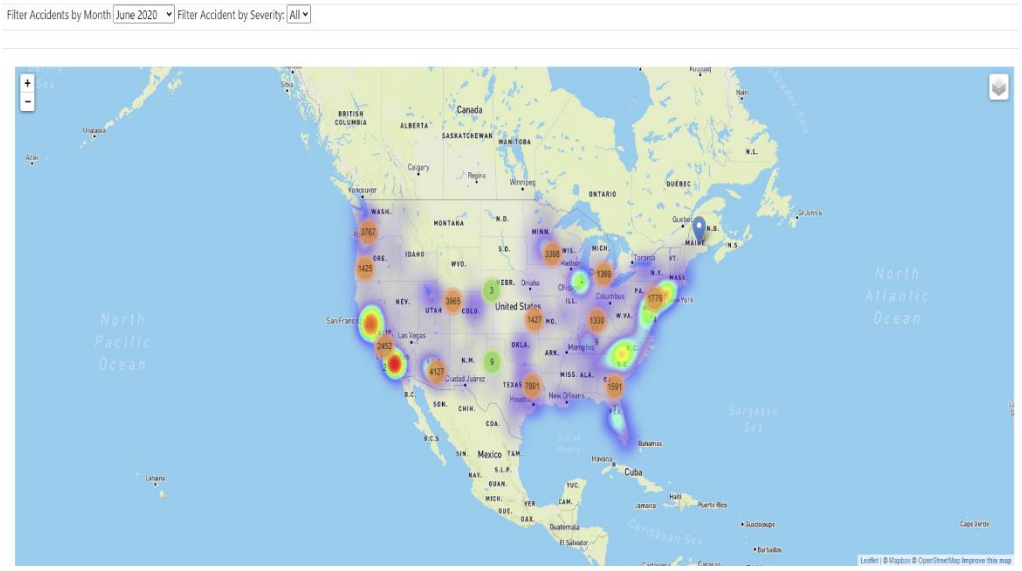


Fig 1: Leaflet heatmap of accidents

Taking a look at the 4-month period of March 2020 to June 2020, the largest volume of car accidents took place on the East Coast. States such as Virginia, Massachusetts, New Jersey, and Pennsylvania reported the highest incidence. On the West Coast, Los Angeles was a standout city with 7,886 accidents and Oregon had a significant volume as well.

It is logical that states and cities with the highest population densities would have a corresponding number of incidents. Pivoting to states with lower populations such as Montana, Wyoming, North and South Dakota, we see little to no accidents. There appears to be correlation behind the distribution of accidents and population. However, as this dataset did not include population specific data, this cannot be said definitively.

5.2 Bar Chart : Number of Accidents by State

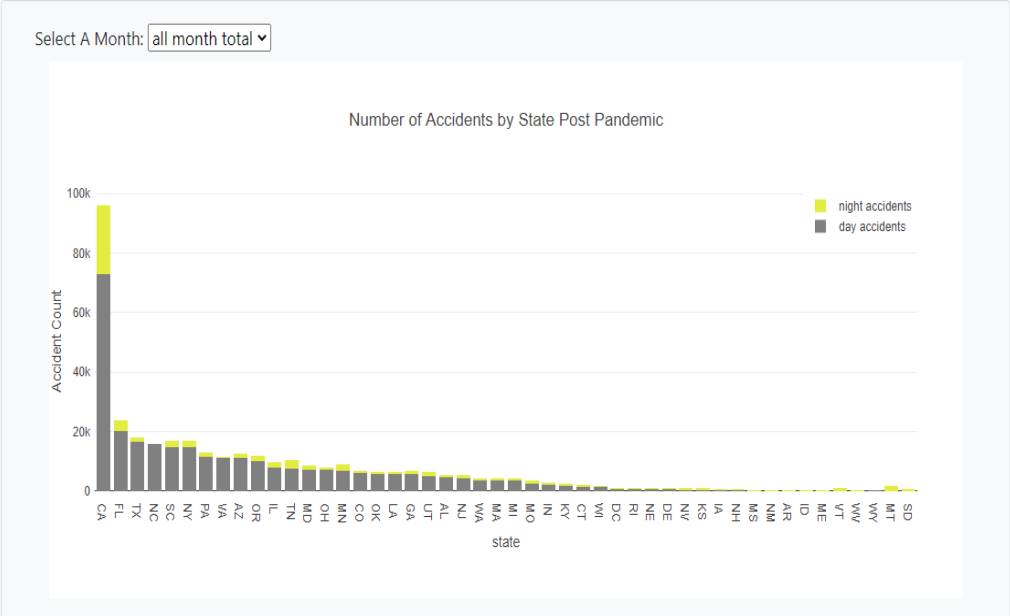


Fig 2: Plotly stacked bar chart of accidents by state

The bar chart plotted from march to June for the pre-pandemic and post-pandemic eras identified that for majority of the states, there were more accidents in the day than there were in the night. The total number of accidents per state reduced during the pandemic period except for California (CA).

California has always had a very high accident rate. As a matter of fact, in 2019 its accident rate was almost double that of the state with the second highest number of accidents. During the pandemic months(march-June), accident rates in California doubled and reached its all-time high of roughly 96k accidents throughout the period (this was almost four times the accident count of the next highest state—Florida). Based on news reports, accidents have risen in California due to over-speeding reckless drivers on the now empty roads in big cities like Los Angeles

5.3 Scatter Plot: Comparison of Day & Night Accidents

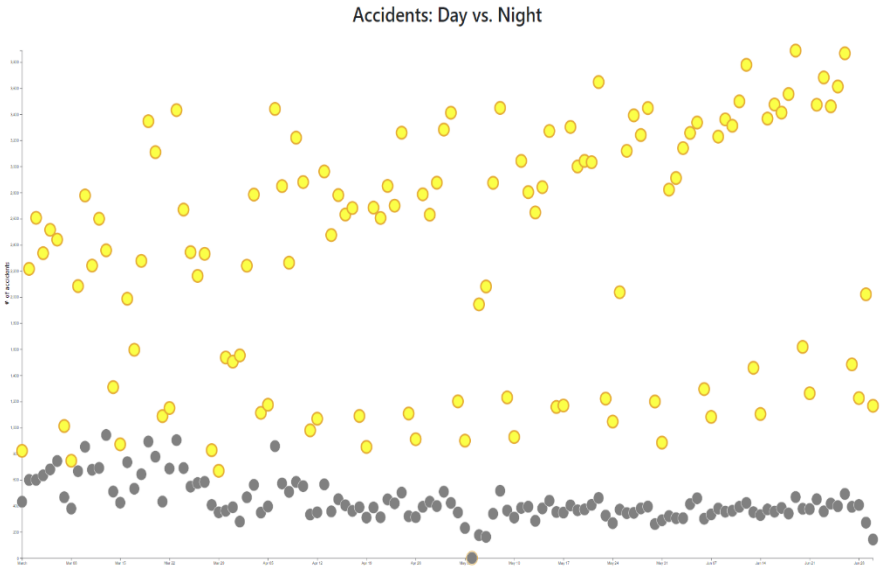


Fig 3: D3 scatter Plot of accidents over time

The scatter plot above visualizes cumulative accidents by day for 4 months starting March 1st, 2020.

Accidents that occur during the day are represented by yellow dots. These averaged 2,315 accidents and peaked on June 19, 2020 with 3,888. The nadir of accidents occurring on May 04, 2020 (1).

Nocturnal accidents are represented by the grey dots and averaged 943 with accidents with a peak on March 13, 2020 (943). The nadir of accidents was on May 04, 2020 (1).

Overall, the plot indicates that not only do most accidents occur during the day but there is an upward trend. In contrast, night accidents are trending down.

5.4 Line chart: United states Accident Severity over time

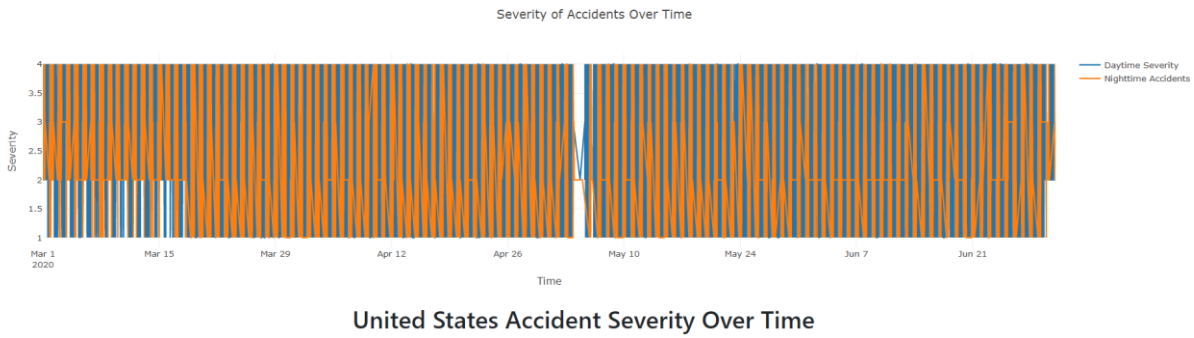


Fig 4: Plotly line chart of accident severity over time

The line chart was inconclusive. It was recommended that an average be taken of all of the data, split into day and night. However, considering California accounted for almost 50% of the accidents, there would be a large skew with the data, and would not portray an accurate story. Due to the inconclusive nature of the line chart, a data table was created breaking down the numbers based on severity.

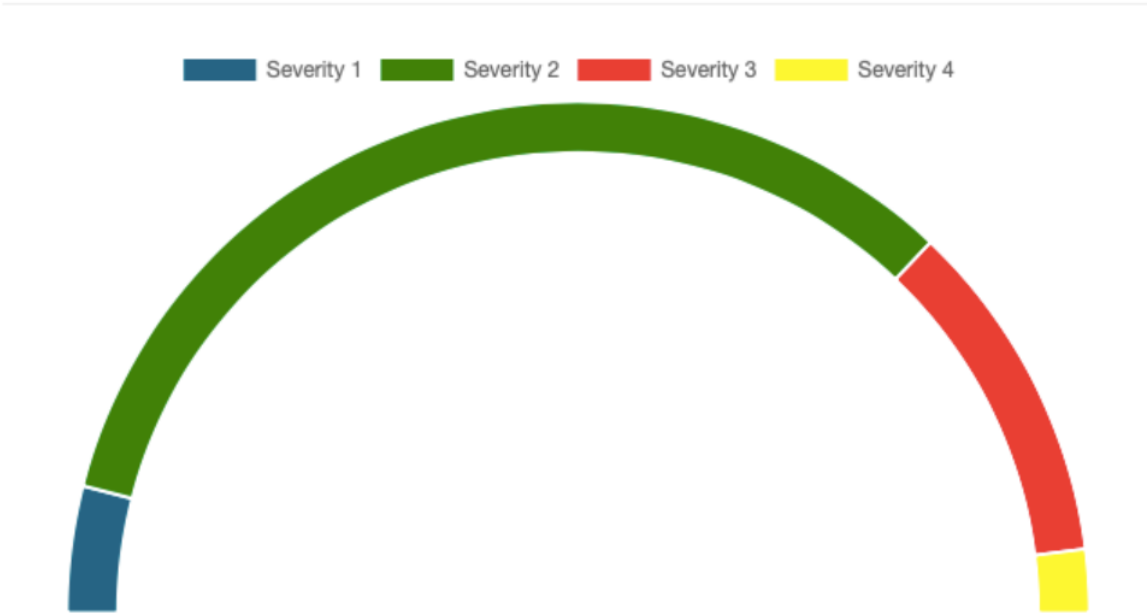


Fig 4.5: JS gauge chart

These charts allowed for a much clearer story to be told. Most accidents occur during the day, with severity 2 accounting for 76% of all accidents across the two charts. Based on the data, it was reasonable to assume, before analysis was performed, that severity two would account for most of the accidents, but 76% was a much larger piece than expected.

5.5 Accident Data Dashboard

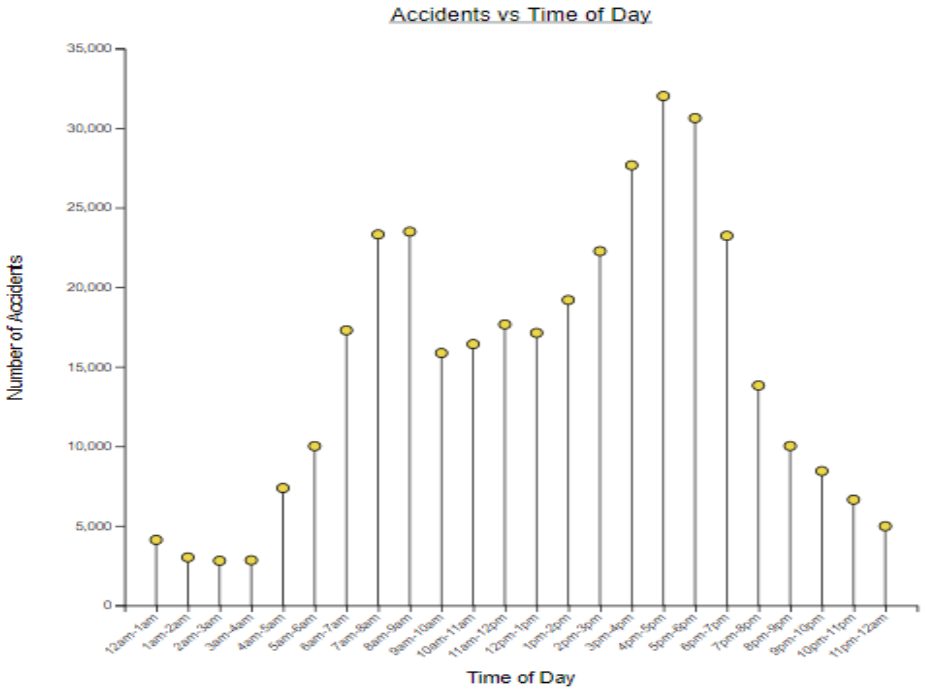


Fig 5: D3 lollypop chart of number of accidents per hour in the day

The data dashboard was built to reveal information concerning the most accident-prone time in the day (Fig. 5), what weather conditions yielded the highest number of accidents (Appendix I) and what roads had the highest number of accidents (Appendix G).

To address the question of what time of day accidents occurred, a lollipop graph was used. According to the graph, 4pm-5pm was the most problematic time of day for accidents. This was created using D3.js.

The most common weather conditions during accidents were visualised using a bar graph. According to the graph, weather is not a factor in most accidents. Most accidents happen on regular days with fair weather. The graph was generated using the Plotly library in JavaScript.

The roads with the highest number of accidents were visualised using two charts (a bar chart and a Leaflet map). The first graph shows the 15 most accident-prone roads on a map whereas the later (created with Plotly) ranked the roads based on number of accidents.

Lastly, the Choropleth map (Appendix J) on the dashboard page was used to further prove what was already identified on Fig. 1 and Fig. 2 — that the highest number of accidents were in California and on the East Coast.

6. Limitations

6.1 File size and Website Responsiveness

The web site leveraged a CSV file as the data sources rather than an API. This resulted in significant client-side processing. When rendering the Leaflet map, users observed high memory usage which impacted the performance of their device (e.g. laptop, desktop). In some cases, the browser halted.

To minimize a negative user experience, the default behavior for loading the map was modified to filter the time frame to one month and severity. In this way, the volume of data retrieved and processed was reduced.

6.2 Correlations

The original factor in determining how to construct the scatter plot was based off elements of weather. However, after the initial correlation analysis done in python, it was determined that there was not a strong enough correlation to plot the results and a decision was made to pivot.

	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)	Precipitation(in)
Temperature(F)	1.000000	0.994648	-0.455798	0.078754	0.209071	0.099080	-0.056208
Wind_Chill(F)	0.994648	1.000000	-0.438279	0.092058	0.209302	0.065498	-0.052248
Humidity(%)	-0.455798	-0.438279	1.000000	0.218951	-0.385988	-0.214082	0.210233
Pressure(in)	0.078754	0.092058	0.218951	1.000000	-0.040603	-0.061041	0.021947
Visibility(mi)	0.209071	0.209302	-0.385988	-0.040603	1.000000	0.034712	-0.299217
Wind_Speed(mph)	0.099080	0.065498	-0.214082	-0.061041	0.034712	1.000000	0.016575
Precipitation(in)	-0.056208	-0.052248	0.210233	0.021947	-0.299217	0.016575	1.000000

6.4 Radar Chart

When constructing the radar chart, the data console logged, but the data did not populate into the graph. Much of the syntax for creating the chart was simple, but there was a disconnect between the data and the chart itself. As a result, the idea of creating a radar chart depicting changes from day to night was scrapped in favor of the line and gauge charts.

7. Conclusion

By the end of the project, each team member gained experience cleaning and exporting data using Pandas, using D3 to create dynamic websites and finally Plotly.

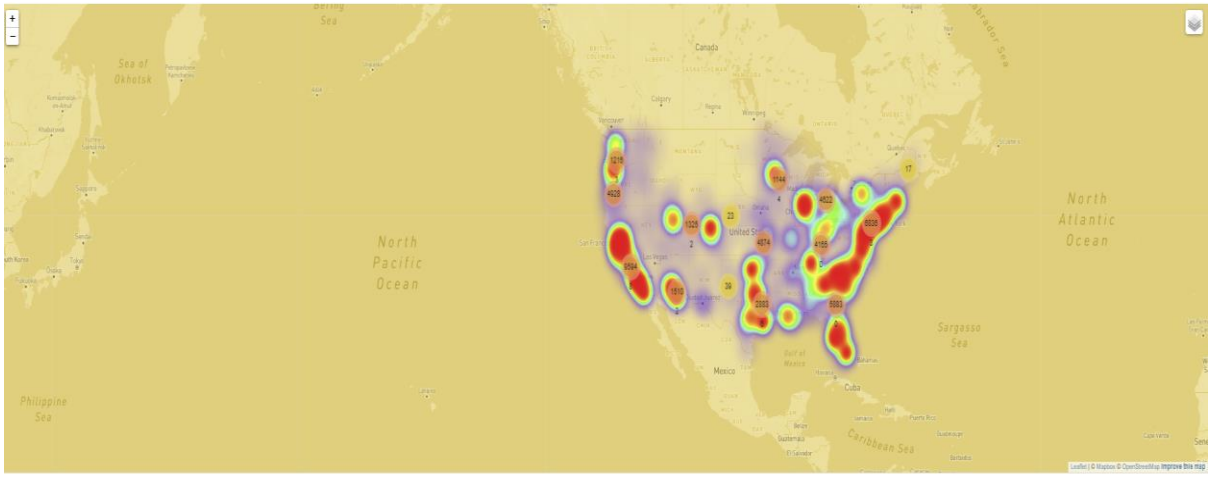
We realised that accidents were more prevalent in the East Coast and California, every state (except for California) experienced a reduction in accident rates during the pandemic months, accidents occur more often during the day and between 4p to 5pm.

References

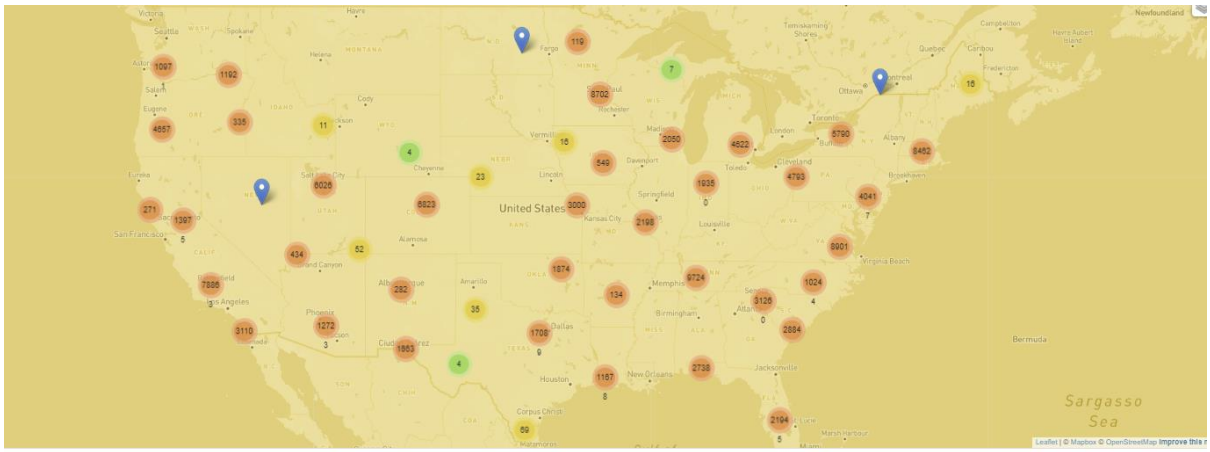
[1] https://www.sweeneymerrigan.com/car-accident-statistics-in-the-united-states/#:~:text=More%20than%20six%20million%20car,accidents%20result%20in%20nonfatal%20injuries.

[2] https://www.kaggle.com/sobhanmoosavi/us-accidents

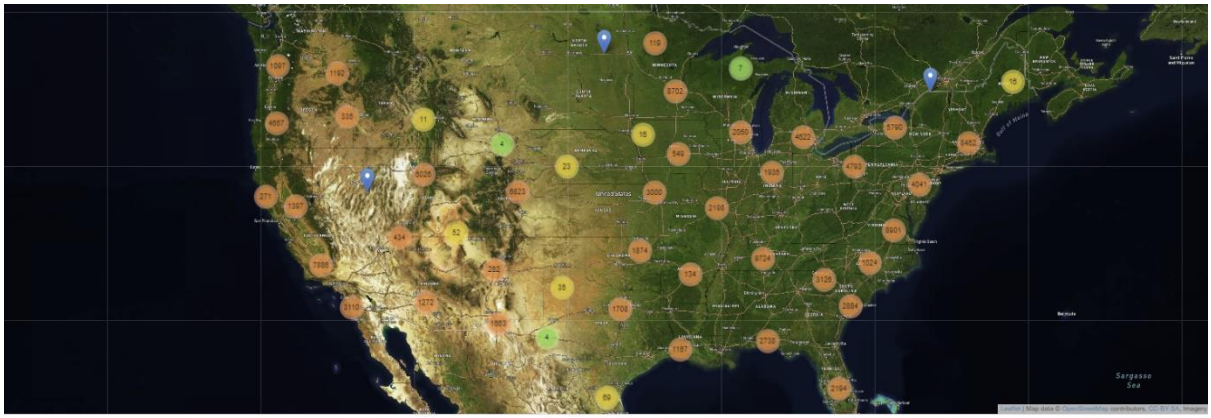
Appendices



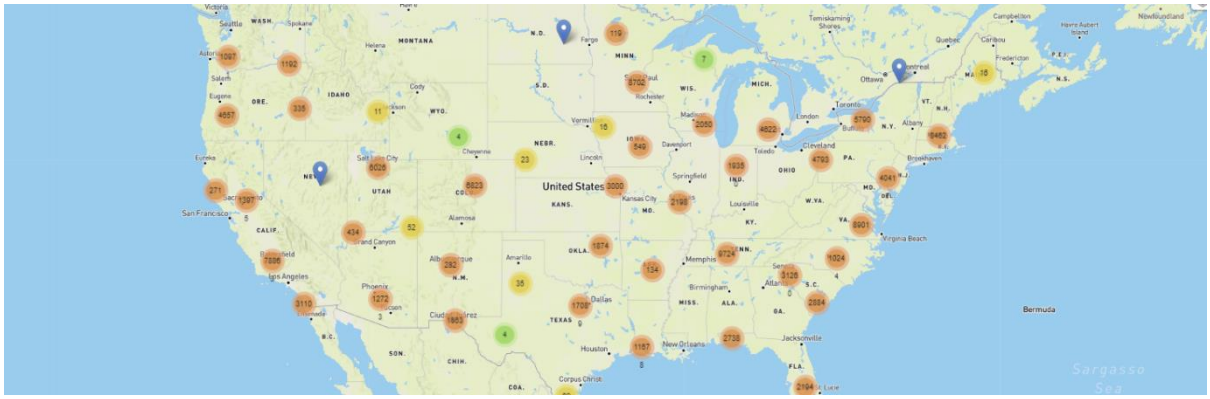
Appendix A: Leaflet map showing heatmap layer of all accidents during the 4 months of Covid-19 around the U.S with a golden overlay design



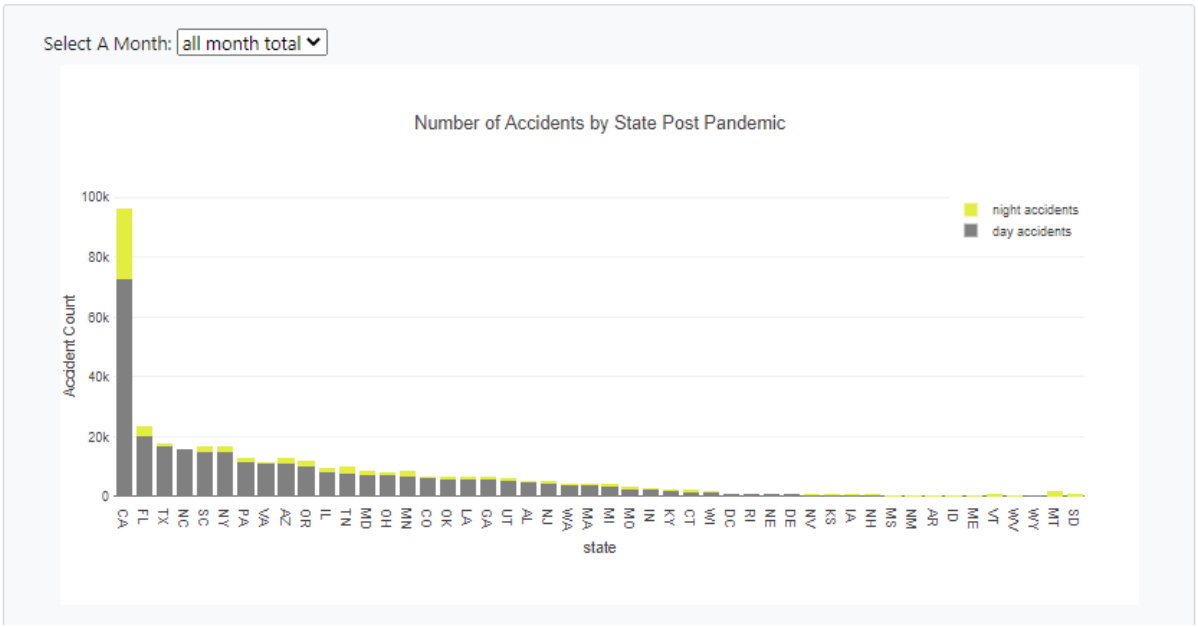
Appendix B: Leaflet map showing marker layer of all accidents during the 4 months of Covid-19 around the U.S with a golden overlay design.



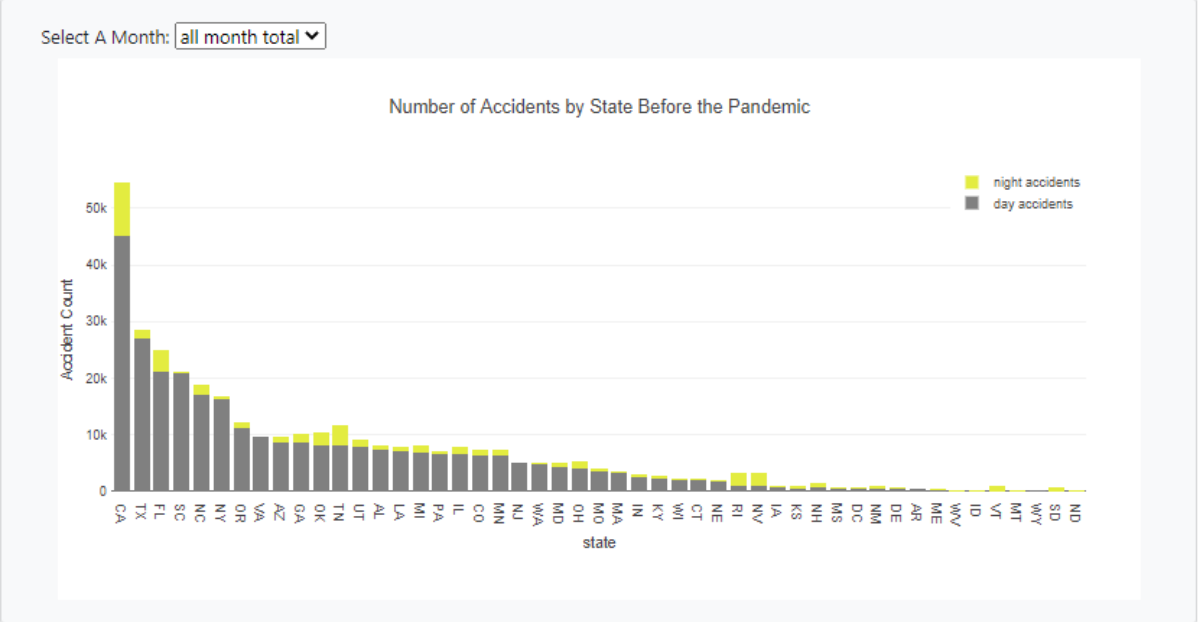
Appendix C: Leaflet map showing marker layer of all accidents during the 4 months of Covid-19 around the U.S in satellite view.



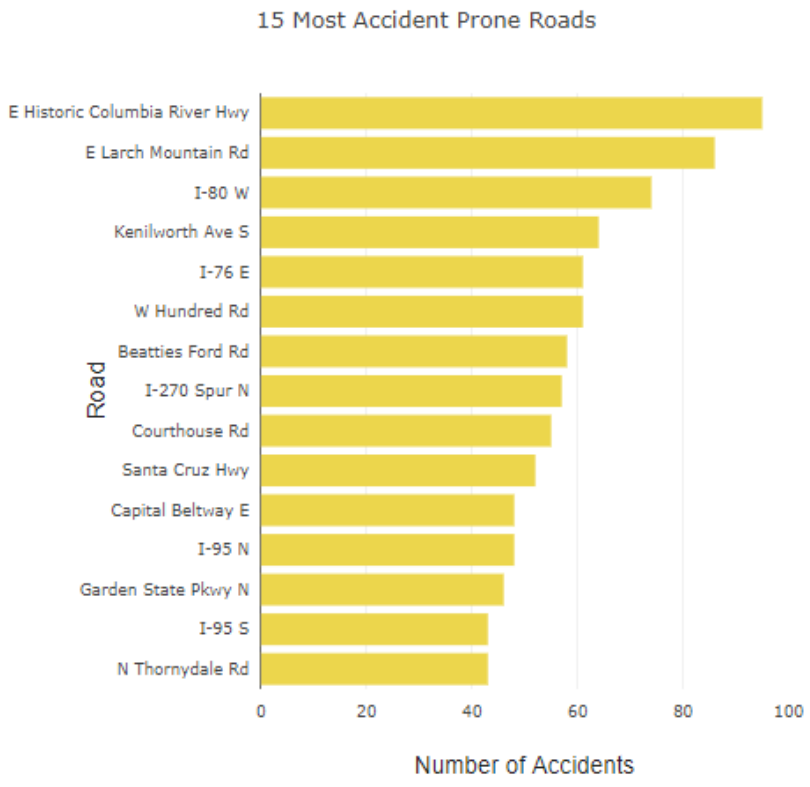
Appendix D: Leaflet map showing marker layer of all accidents during the 4 months of Covid-19 around the U.S in street view.



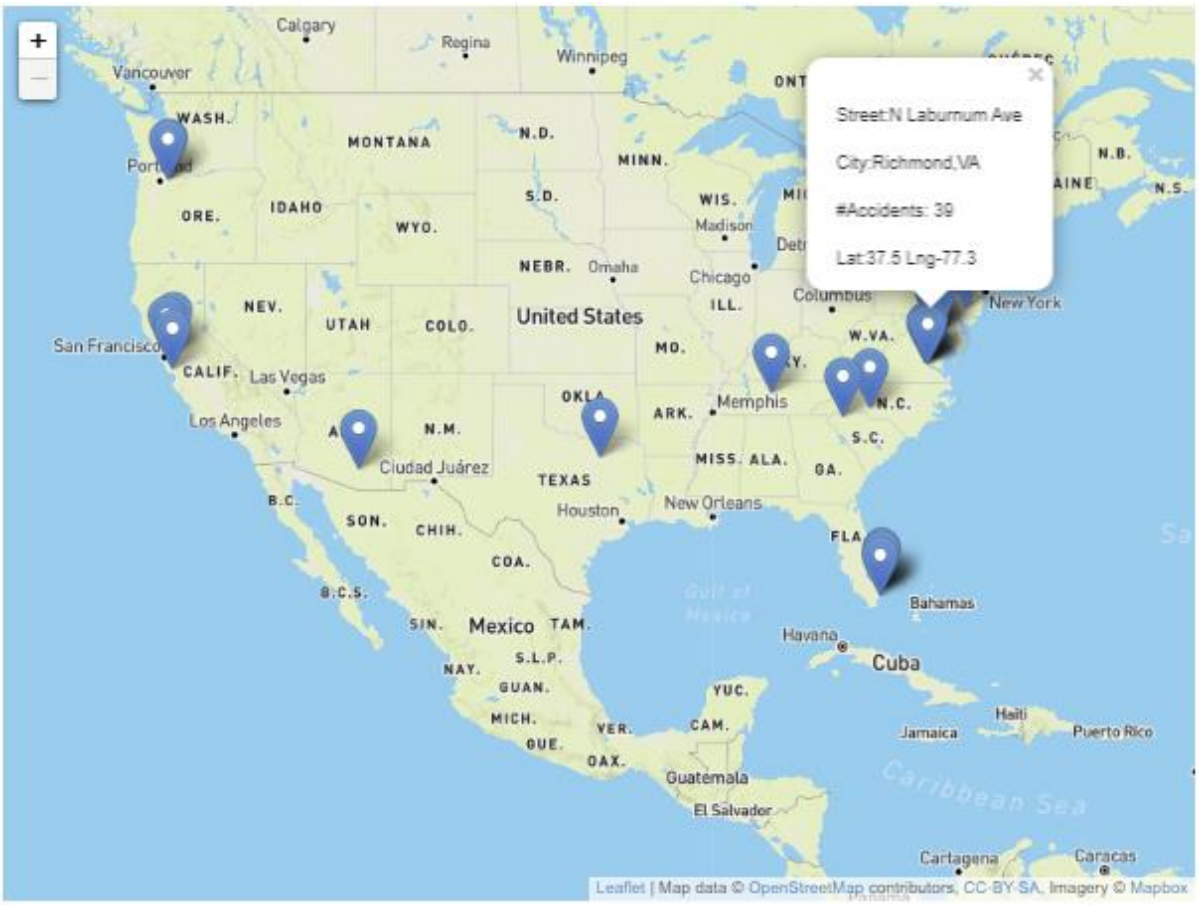
Appendix E: Bar chart showing distribution of accidents across all the states for the Covid-19 period (i.e. from March 2020 to June 2020).



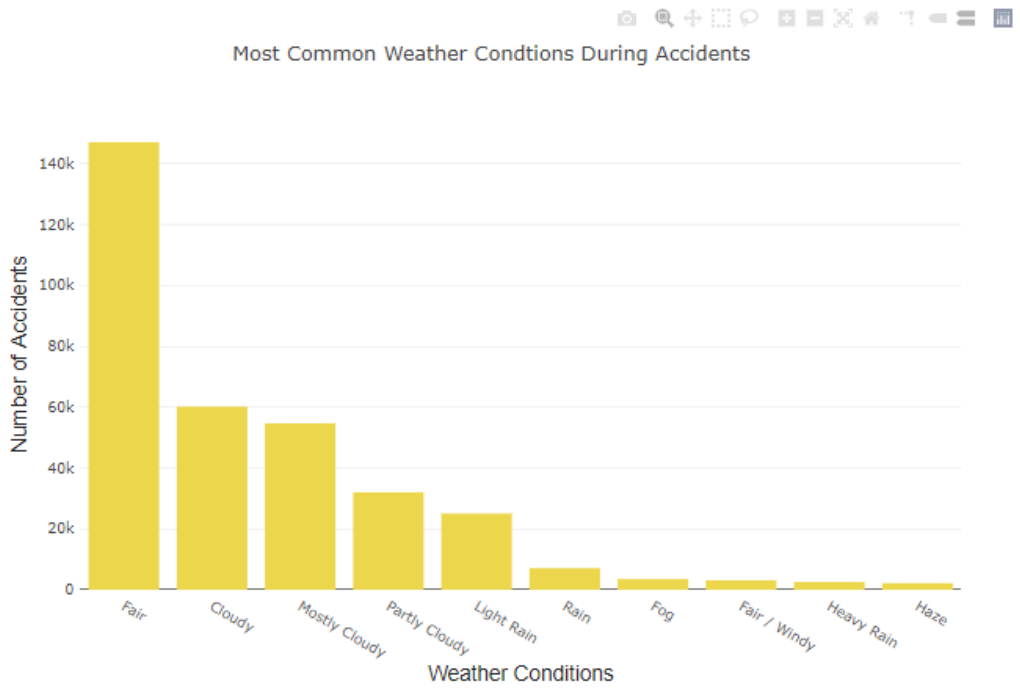
Appendix F: Bar chart showing distribution of accidents across all the states for the prior to the Covid-19 period (i.e. from March 2019 to June 2019).



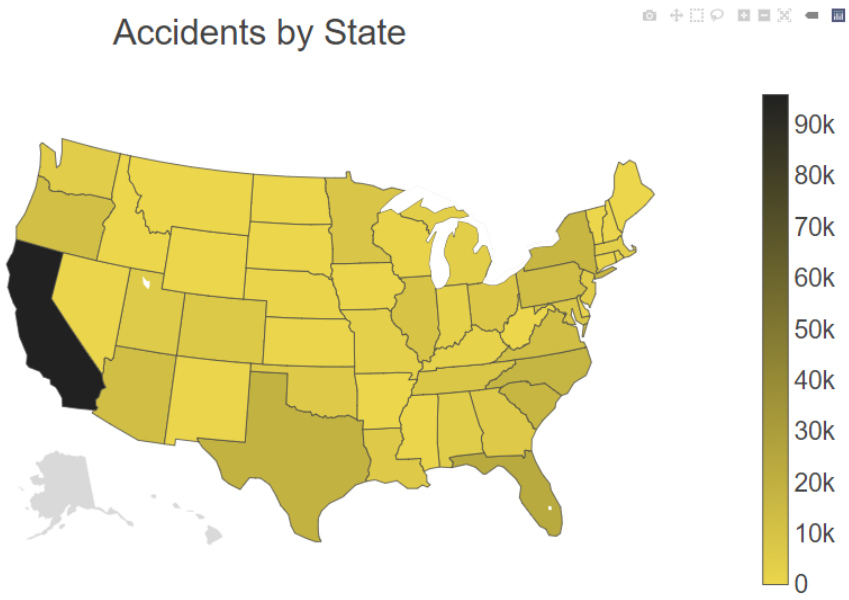
Appendix G: Dashboard page horizontal bar chart showing top 10 roads with the highest number of accidents in the USA.



Appendix H: Dashboard page Leaflet chart showing top 10 roads with the highest number of accidents in the USA.



Appendix I: Dashboard page vertical bar chart showing accident rate vs weather condition.



Appendix J: Dashboard Plotly Choropleth map showing states with accidents envisioned with a colour scale.