Sydney Lieske

Isabella Morales

Final Project

Intelligent Mobility

Fall 2022

Car Accident Dataset

*Submitted to*

Mahta Zakaria

Florida Polytechnic University

*Submitted by*

Sydney Lieske

Isabella Morales

Department of Data Science

Florida Polytechnic University

**12/11/2022**

Table of Contents

[Executive Summary: 3](#_Toc121674250)

[Introduction 4](#_Toc121674251)

[Literature Review 4](#_Toc121674252)

[Methodology 6](#_Toc121674253)

[Results 7](#_Toc121674254)

[Conclusions 12](#_Toc121674255)

[References 13](#_Toc121674256)

**Table of Figures**

[Figure 1 - Accident distribution in the US 7](#_Toc121677210)

[Figure 2 - Top 10 States with most no. of Accidents 7](#_Toc121677211)

[Figure 3 - Top 10 States with least no. of Accidents 8](#_Toc121677212)

[Figure 4 - Top 10 Accident Prone Streets 8](#_Toc121677213)

[Figure 5 - Road Accident Percentage for Different Months 9](#_Toc121677214)

[Figure 6 - Road Accident Percentage for Different Days 9](#_Toc121677215)

[Figure 7 - Road Accident Percentage for Different Hours 10](#_Toc121677216)

[Figure 8 - Presence of Bump, Crossing, Give\_Way, or Junction 10](#_Toc121677217)

[Figure 9 - Presence of Stop, No\_Exit, Traffic\_Signal, or Turning\_Loop 11](#_Toc121677218)

[Figure 10 - Road Accident Percentage for Weather Conditions 11](#_Toc121677219)

Executive Summary:

This project is an analysis of the US Accidents dataset which covers 49 states of the USA. The objective of our data analysis on U.S. accidents was done in order to find preventative measures. Preventative measures can afterwards be used in order to reduce the number of accidents that occur. Understanding this dataset would allow for the ability to limit the number of crashes that occur during a year. The US Accidents dataset can be used for real-time car accident prediction, casualty analysis, studying the impact of environmental stimuli on the occurrence of accidents, and even car accident hotspot locations. It also can be used to study the Covid-19 pandemic on traffic accidents and behavior

## Introduction

Research Questions:

* What is the distribution of accidents across the country?
* Which states have the most cases of accidents? Least?
* Which streets are the most accident prone?
* Which months have the most accidents?
* What time and day are safer to travel?
* What factors are most responsible for these accidents
* During what weather conditions do the most accidents occur?
* What can be implemented to help reduce these accidents?
* How could these accidents be minimized?
* What kind of IoT devices were used to report the incidents?
* Was an ATMS used?

We chose this dataset because of how prominent vehicle accidents are within our everyday lives. Crashes that occur impact not only everyone involved, but also those around them. We wanted to look at these statistics to understand what causes and what things can be done to prevent the number of crashes that occur in a year.

Literature Review

Dataset:

[US-Accidents: A Countrywide Traffic Accident Dataset](https://smoosavi.org/datasets/us_accidents)

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, arXiv preprint arXiv:1906.05409 (2019).

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. “Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.” In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

The data is from Feb. 2016 to Dec. 2021 which is collected from multiple APIs that stream traffic incident data. With these APIs, they broadcast the traffic data captured by various entities such as the US and state DoT, traffic cameras and sensors within road-networks, and law enforcement agencies. Currently, there are about 2.8 million accident records in this dataset.

This dataset includes 47 columns of variables: ID, Severity, Start Time, End Time, Start Latitude, Start Longitude, End Latitude, End Longitude, Distance (mi), Description, Number, Street, Side, City, County, State, Zipcode, Country, Timezone, Airport Code, Weather Timestamp, Temperature (F), Wind Chill (F), Humidity (%), Pressure (in), Visibility (mi), Wind Direction, Wind Speed (mph), Precipitation (in), Weather Condition, Amenity, Bump, Crossing, Give Way, Junction, No Exit, Railway, Roundabout, Station, Stop, Traffic Calming, Traffic Signal, Turning Loop, Sunrise Sunset, Civil Twilight, Nautical Twilight, Astronomical Twilight.

Methodology

For our project we are utilizing R and Python to do our data cleaning and analysis. We first started by researching more about our dataset and what we wanted to do with it.

In data cleaning, many columns were removed from the original dataset that we felt were not as useful as others. From 47 columns it was cut down to 36 for one portion of data analysis. The unused dataset was used for a smaller amount of analysis, for example, the Longitude and Latitude variables were not used originally, but in order to see the distribution of accidents across the United States (Figure 1) it was necessary to observe.

The removed columns include ID, Start Longitude, End Latitude, End Longitude, Number, Description, Country, Timezone, Airport Code, Wind Direction, Precipitation (in), Amenity, Bump, Crossing, Give Way, Junction, No Exit, Railway, Roundabout, Station, Stop, Traffic Calming, Traffic Signal, Turning Loop, Sunrise Sunset, Civil Twilight, Nautical Twilight, Astronomical Twilight.

One thing that we wanted to learn was time of day that the most accidents occurred. In order to properly read the Start and End time, we had to transform it into a data type that could be read. We created categorical variables for all variables that need to be, like Traffic Signal or Junction. Then, Severity was grouped together by number. Severity Levels were from 1-4.

Null or NA values were read for some columns, so those needed to be filtered out from columns like Weather Condition and Temperature.

With the use of Python, we utilized some data visualization packages such as pandas and matplotlib. First, we imported the csv file containing all 2.8 million records and converted the start and end time variables into a datetime feature. Then we created a dataframe of the city and their corresponding accident cases. We also created a data dictionary using the US state codes and their corresponding names. Then used a function to convert the state code with the actual corresponding name. This helped us create the top 10 cities and states with the most no. of accidents and least no. of accidents figures.

We also did some data visualizations just to gain some more insight. Some more things we wanted to discover were which streets had the most accidents, the severity level of accidents, which months had the most and least accidents, as well as the days of the week and hours.

Results

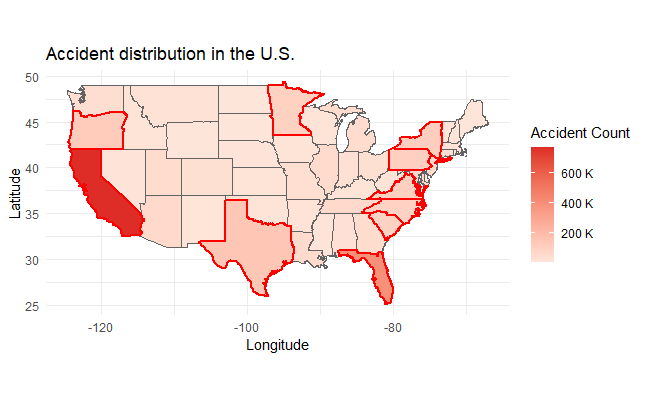


Figure 1 - Accident distribution in the US

This figure, made with R, shows the distribution of the data throughout the United States. As we can see California has the most accidents of all 49 states, followed by Florida and Texas.

Chart, histogram

Description automatically generated

Figure 2 - Top 10 States with most no. of Accidents

This figure shows the top 10 states with the most number of accidents in the United States, as figure 1 shows the distribution on a map, here we can see the numbers. California has the greatest at 795,868 (28%) of the accidents occurring, followed by Florida with 401,388 (14.1%) of accidents, and Texas with 149,037 (5.2%).

Chart, bar chart, histogram

Description automatically generated

Figure 3 - Top 10 States with least no. of Accidents

In this figure it shows the top 10 states with the least number of accident cases in the US. The state with the least number of accidents is South Dakota with 201 (0.01%) of accidents occurring, Vermont with 365 (0.01%), and Wyoming with 990 (0.03%).

Chart, bar chart

Description automatically generated

Figure 4 - Top 10 Accident Prone Streets

This figure shows the top 10 streets that are the most accident prone in the US. 39,853 accidents occurred on I-95 N, followed by 39,402 on I-5 N, and 36,425 on I-95 S.

Chart, bar chart

Description automatically generated

Figure 5 - Road Accident Percentage for Different Months

The months with the highest percentage of road accidents are November at 12.68% and December at 16.66%. The least percentage is in July at 5.59% and August at 6.28%

Chart, bar chart

Description automatically generated

Figure 6 - Road Accident Percentage for Different Days

The days with the highest road accident percentages are Friday with 17.29% and Thursday with 16.29%. The lowest percentages are the weekends, Saturday with 10.95% and Sunday with 9.11%.

Chart, histogram

Description automatically generated

Figure 7 - Road Accident Percentage for Different Hours

During morning hours, the highest percentage of accidents occurring, happen during 7am with 4.75% followed by 4.6% at 8am. During the night hours, the highest occurring accidents happen during 3pm – 5pm with 3pm being 7.53%, 4pm with 7.68%, and the highest being 5pm with 7.74%.

Chart, bubble chart

Description automatically generated

Figure 8 - Presence of Bump, Crossing, Give\_Way, or Junction

These figures show if there was a presence of a bump, crossing, give way, or junction when an accident occurred. As we can see only 0.04% of cases had a bump, 7.04% were at a crossing, 0.24% were at a give way, and 10.21% were at a junction.

Chart, bubble chart

Description automatically generated

Figure 9 - Presence of Stop, No\_Exit, Traffic\_Signal, or Turning\_Loop

These figure show if there was a presence of a stop sign, no exit, traffic signal, or turning loop. With 1.77% of cases there was a stop, 0.15% of it being a no exit, 9.32% having a traffic signal, and 0.00% having a turning loop.

Chart, waterfall chart

Description automatically generated

Figure 10 - Road Accident Percentage for Weather Conditions

Here we wanted to see how many accidents occurred during different weather conditions. As we cans see most accidents occurred when the weather condition was fair at 38.91%, followed by mostly cloudy (12.79%), and cloudy at 12.26%. The least number of accidents occurred with light snow at 1.54% and fog at 1.45%. We would want to see how many days of the year were fair, cloudy, snowy, or foggy as it is hard to come to a solid conclusion as some weather conditions do not occur as frequently as others.

Conclusions

Through the analysis of our chosen dataset, we were able to answer many of our research questions such as:​

* Fridays are the most accident prone with cases at 17.29%, vs Sundays 9.11%, the lowest of the days of the week. Hours 3-5pm are the most in the evening time with 5pm being the highest, and 6-8am with 7am at the most in the morning.​
* The presence of traffic signals are just one of the preventive measures that we were able to analyze within this data set.​
* Accidents occur more on highways or state road than on local streets, preventive measures like speed limits ​
* Roadside messages and if the vehicles are equipped with ADAS would help to minimize accidents.​

Some limitations and future work with the data are, seeing how frequently each weather condition occurred. As most days seem to be fair or cloudy, with less conditions such as fog or snow occurring, we want to come to a solid conclusion about the number of accidents occurring during different weather conditions with how frequent they are.

​

​

​

​

​

​

References

Dataset:

[US-Accidents: A Countrywide Traffic Accident Dataset](https://smoosavi.org/datasets/us_accidents)

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, arXiv preprint arXiv:1906.05409 (2019).

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. “Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.” In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.