SAN DIEGO STATE UNIVERSITY

BDA 594: Big Data Science and Analytics Platforms

USA ACCIDENT ANALYSIS

Final Project

Big Data Analytics, San Diego State University

Project Website:
https://sites.google.com/sdsu.edu/titans

Group 12(TITANS)

Naveen Reddy Sama

Vijay Bhaskar Kothapally

Chandra Teja Peddi

Akshita Nimmala

December 15, 2022

Table of Contents

1. PROBLEM STATEMENT

In 2022, there will be close to 300 million vehicles in operation in the United States, according to a research article published by Mathilde Carlier in Statista. According to the Annual United States Road Crash Statistics (ASIRT), more than 46,000 individuals die in car accidents yearly. With the increase in the number of vehicles operating on the road, there is a rise in the number of accidents and fatalities across the country every day. So, the need for Accident Analysis has increased tremendously as reducing road accidents is a vital public safety concern on a worldwide scale. Road accidents have become a concern for both the government and individuals as these accidents are typically fatal and pose a risk to society.

There are multiple challenges with the methods by which the government is trying to prevent accidents. Hence, a robust analysis of road accidents enables us to identify and categorize the causes of road car accidents. The analysis and identification of the accident causes would help prevent similar accidents from happening again and paves a road map for future projects. This report aims to address this problem and examine the primary causes of the rise in car accidents.

2. PURPOSE

Automobiles have become part and parcel of everyone's lives these days. The number of automobiles is expected to grow with the population. So, analyzing the previous accident and building prediction models to identify the critical factors that can lead to an accident and determine the severity of it based on those factors. Eliminating the possibility of accidents is entirely out of hand for now. Still, we can take specific steps to minimize the number and the severity with the results from our analysis.

The main objectives of this project are:

- To analyze the road accident data in The United States and find the road accident patterns, potential causes, and any other trends.
- Visualize the accident patterns by state, highway, city, and county.
- Identify the top highways, cities, counties, states, and time zones having the highest accidents.
- Determine which time of the day, month, and year records more accidents.
- Recognize which weather conditions like visibility, temperature, and precipitation, also nearby transport hubs,
- Analyze the Covid-19 impact on the number of accidents and their severity.
- Build an accident severity prediction model.

3. LITERATURE REVIEW

Literature review plays an important role in any research work. Several researchers have conducted their research work related to road accident analysis. Each of them analyzed the data in different ways. Some of them identified the hotspot regions where most accidents took place, and a few of them created machine-learning models to predict future accidents. Furthermore, they presented road safety solutions.

I. "The paper titled "A review of the effect of traffic and weather characteristics on road safety", Accident Analysis & Prevention" by Theofilatos A., Yannis G(2014) tries to give an overview of how traffic and weather conditions affect road safety. There are many new patterns observed in this research. For example, this paper says that road traffic appears to share a non-linear relationship with accident rates. In terms of meteorological effects,

precipitation often increases accident rates but does not appear to have a consistent impact on severity.

II. According to the paper “Effects of weather conditions, light conditions and road lightening on vehicle speed” by Annika K Jagerbrand and Jonas Sjobergh (2016) claimed that driving speed is higher during daytime hours and in clear weather. Additionally, rain also had a major impact on deaths and serious injuries. However additional studies like rush hours and extremely dark conditions weren't included in the study.

III. According to the paper “A Review of Traffic Accidents and Related Practices Worldwide” by Ali Ahmed Mohammed provides information that numerous factors, including a deficient planning system, weak safety standards, and a lack of public awareness, are connected to traffic accidents. To assess the crash frequency on-road segments related to traffic flow, road length, and risk factors including cross-sectional design components, crash prediction models (CPMs) were utilized. These models can be used to determine which road segments in a network are considered to be particularly risky as well as to estimate how safety changes will affect a particular road segment.

4. TOOLS USED

Tableau: As the saying goes, “a picture is worth a thousand words”, Tableau does exactly the same with the data. Tableau is a visualization tool for data analytics that is used to extract meaningful insights from the beautiful charts and graphs made from it. Tableau was used in this project using the accident data for exploratory data analysis and to create visualizations on the US map with accident statistics by states, cities, counties, highways, pin codes, and several relational bar charts and dashboards.

Python: It is one of the finest programming languages for data analysis and model building. Several packages in Python such as Pandas, NumPy, and Sci-Kit learn could be used to play with the data and build models. Python is used in this project for data cleaning and severity prediction model building.

ArcGIS: It is a mapping solution based on cloud technology. It uses interactive maps to analyze the data and share them with others through web maps, apps, etc. ArcGIS was used in this project for spatial analysis through heatmaps with accident severity in the US to study the covid-19 impact on accident patterns.

iMovie: iMovie is a video editing application from Apple Inc. It is free software available on any Apple device. It was used to create the project overview video, the link to which was provided on the title page of this report, it can also be found on the project website.

5. DATABASE MANAGEMENT AND DATA PREPROCESSING PROCEDURES

The dataset used in this project is sourced from Kaggle, an online community of data scientists. The data was collected by a Kaggle user and published for free use. It was collected through several APIs from different entities like the Federal and State DoT which monitors traffic data and accidents through a variety of sources. The data set contains nearly 2.8M records of previous accidents in various states across the United States from February 2016 to December 2021.

Here is the link to the data source.

The dataset includes:

- Start and end time of the accident, with start and end latitude and longitudes.
- Street address with state, city, county, time zone, and Pincode where the accident took place.

- Whether any nearby airports, stations, bumps, amenities, crossings, stops, junctions, exits, loops, traffic signals, and roundabouts.
- Weather conditions at the time of the accident like temperature, wind speed, wind direction, visibility, pressure, relative humidity, etc.

The dataset was downloaded from Kaggle directly in CSV format. Python was used for data preprocessing, and then Tableau was used for exploratory data analysis.

Several Data Pre Processing steps for modeling were performed on this dataset:

- First and foremost, is Data Sampling. As the dataset is huge, working on it is a pretty difficult task. So, a sample size of the entire dataset, say 20 percent is very convenient to work with initially.
- Next is Data Cleaning. Like every other dataset, our dataset also contains missing values. Of these, three variables had high missing values, which were removed from consideration.
- Followed by Data Selection, where usually ineffective columns are dropped. Here, the weather timestamp is a duplicate of the accident start time. Also, the ID column is just a unique row number which is not required in the model as there is no significance for it. Similarly, pin code, latitude, and longitude were omitted as there are already columns for the address.
- Next is Data Transformation, many of the columns in the data are in character format. Since we can run our models on the data with that format, the character variables in the data are label encoded. Also, the start time in the data is in raw format, and from that date is extracted.

- Finally, feature engineering is a crucial step in data preprocessing. As the model performs based on the input features given to the model, if we generate more features with the existing features then we can improve the model performance. In this project Year, month, day, and hour are extracted from the time column. Also, road type is extracted from the street column.

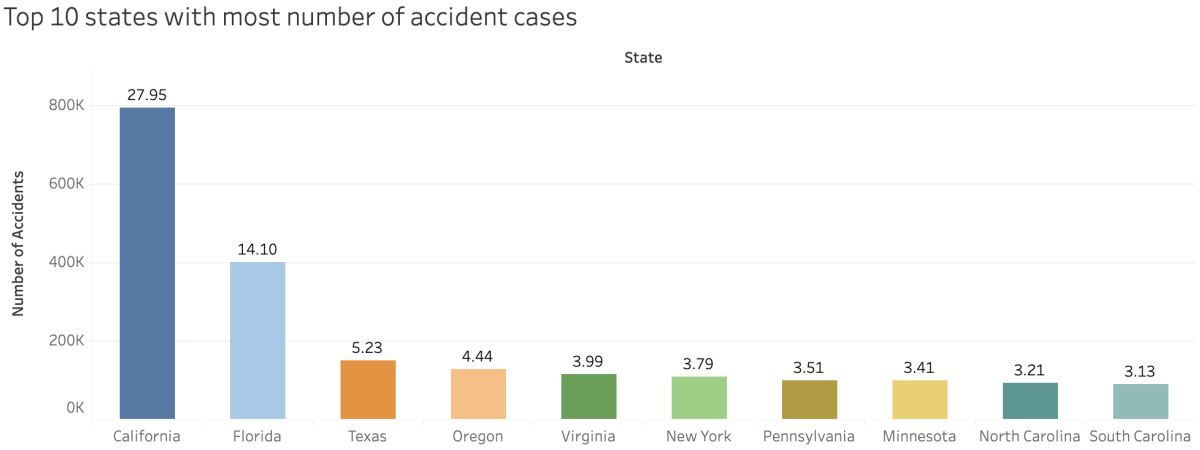
6. DATA ANALYSIS RESULTS

In this project, Data Analysis was done using Tableau and ArcGIS. The main goal of this analysis was to extract meaningful insights from data by identifying which factors contribute most to causing an accident and also visualize which regions (State, City, County, Pincode, and Timezone) have more accidents. Analysis was done in four parts, namely, Spatial Analysis, Temporal Analysis, Other Analysis, and Covid-19 Impact Analysis.

6.1. SPATIAL ANALYSIS

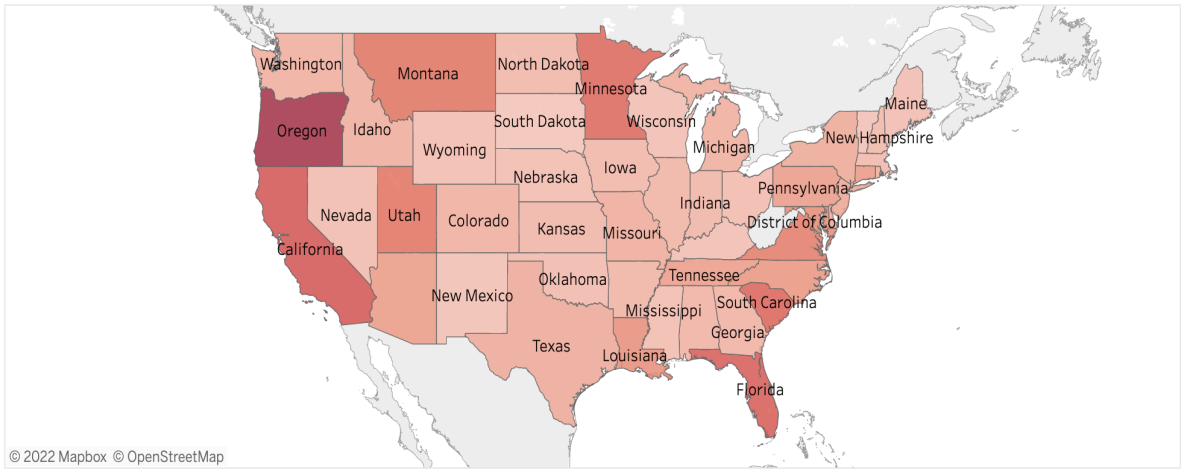
Considering the individual states in the US, visualizations were created by aggregating the number of accidents by each state. Below is the choropleth map (created using the open street map in Tableau) showing the darker-colored states with a higher number of accidents.

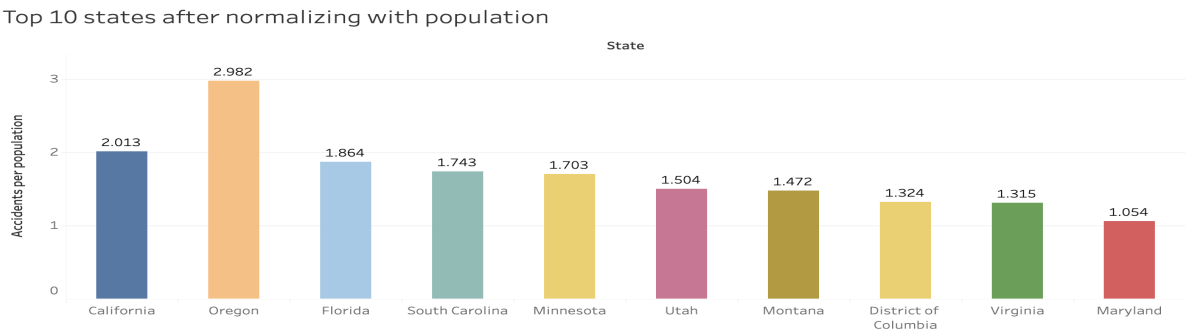
Total Number of Accidents Across USA



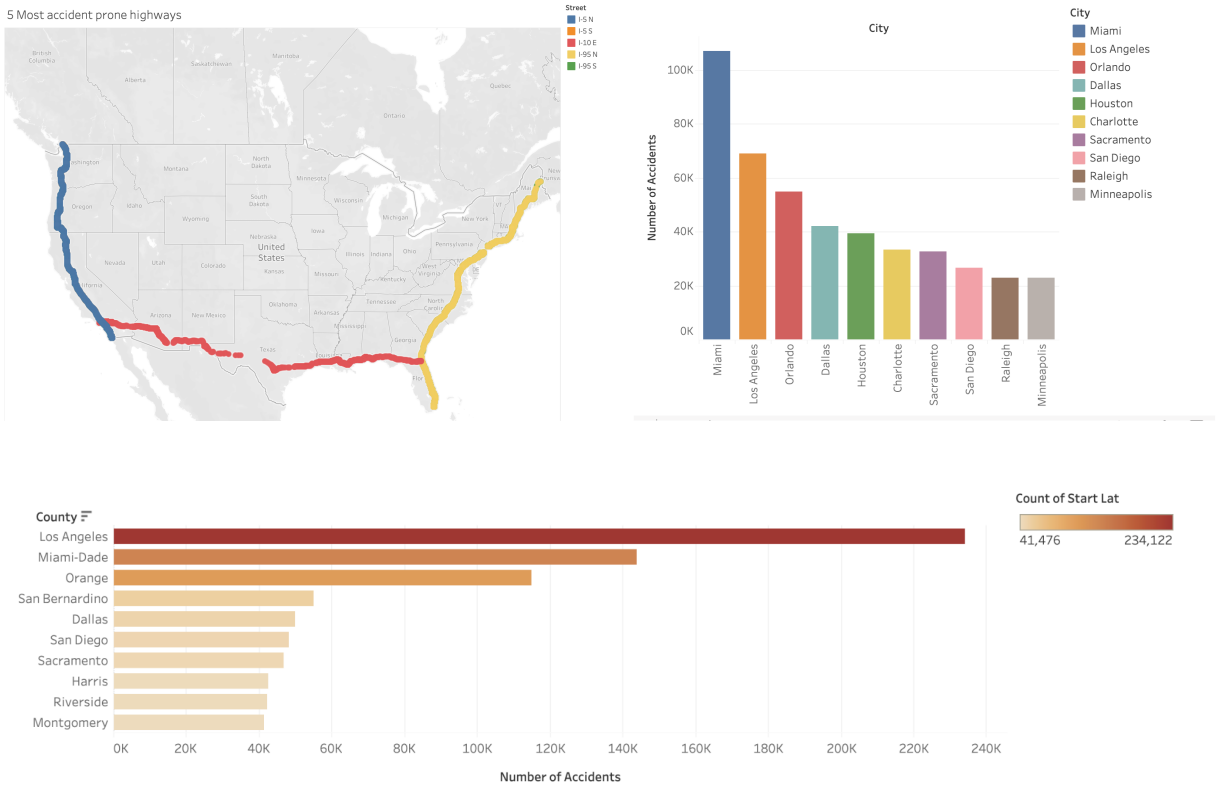
The above bar chart shows the ten states with the highest number of accidents recorded. From the above two visualizations, it can be inferred that the highest number of accidents is recorded in California; About 30% of the total accident records of the past five years in the US are only from California. Florida is next to it.

As California is a highly populated state in the USA, it is evident that there will be more accidents happening. So, we tried to divide the accidents by the population of each state and created further visualizations using the normalized data to observe the accident rate per state with respect to the population.





After normalizing with population, Oregon came on the top, followed by California. So, the rate of accidents is more in Oregon. But California is second after Oregon, meaning the rate of accidents is also more in California along with the number.



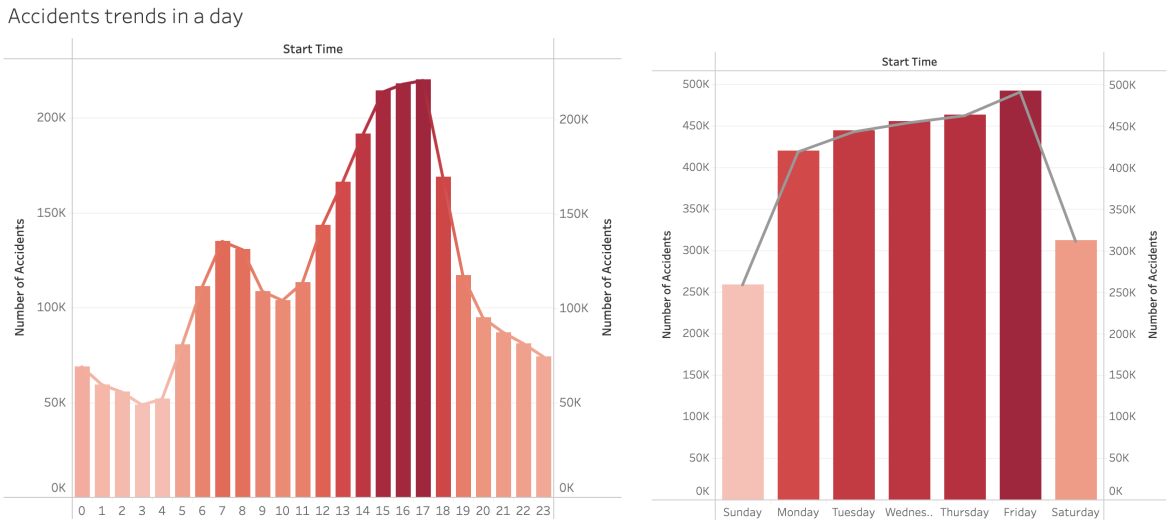
From the above visualizations, it can be gathered that out of the total accidents recorded in the USA, most are from Los Angeles county. Also out of the top 10 counties, three counties are from

California state. San Diego is the sixth one among them in the top ten. When it comes to the cities, Miami tops the list. And out of the top ten, three are from California. There is also another map, showing the top five accident-prone highways identified by the numbers, out of which I-5 (including North and South) got a deadly first place. Also, when zip codes in San Diego were analyzed, more accidents happened in the 92108 area.

6.2. TEMPORAL ANALYSIS

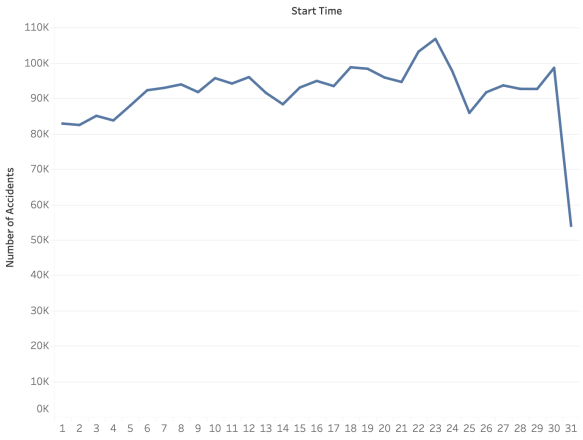
When analyzing accidents, time is one of the most important factors to be considered. For analyzing the accidents with respect to time, Year, month, day, and hour are extracted from the start time provided in the data. By using these, patterns of accidents in a day, week, and month are visualized.

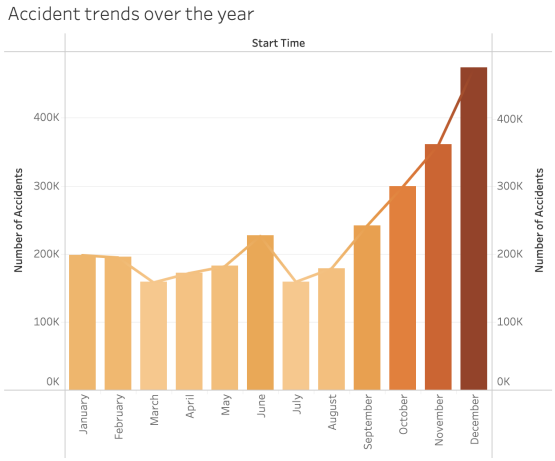
The accident patterns in a day and a week are shown below respectively,



From the visualizations, it can be inferred that most of the accidents are occurring between 3-5 PM. It is obvious that many people will return home from their work at this time. Also when observing patterns over the week it can be seen that the accident rate is high on Friday compared to other weeks and is very low on Sunday as people will be staying in their homes.

The accident patterns in a month and a year are shown below respectively,

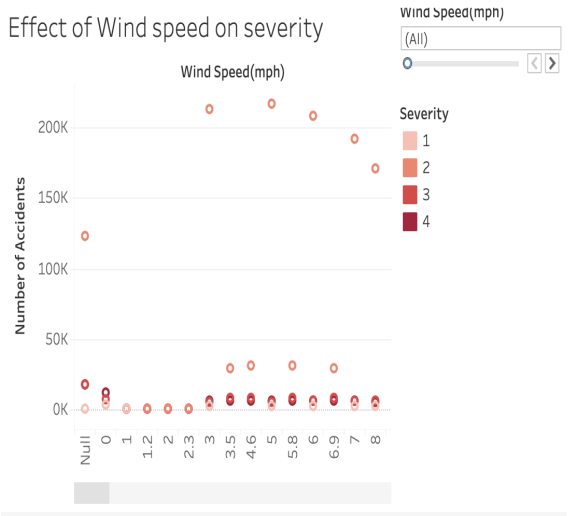


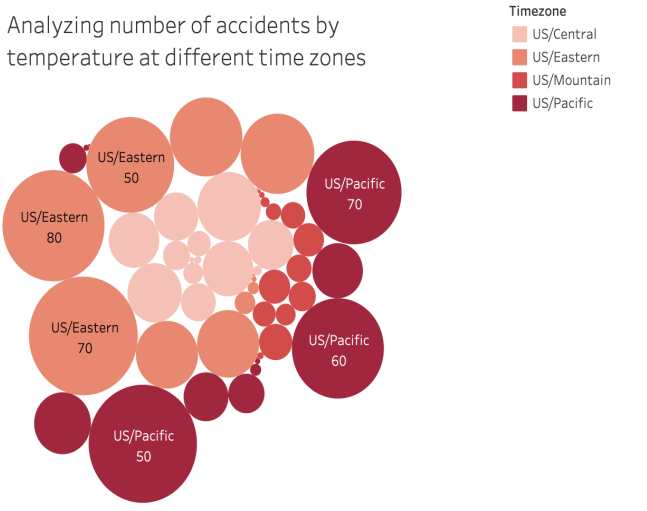


It can be inferred that there is not any regular pattern when it comes to a month but there is a very less number of accidents that are getting recorded at month end. Coming to a year there is a clear pattern that more accidents are occurring in the month of December as it is a festival month. Also, a spike can be observed in the month of June as it is the holiday season.

6.3 OTHER ANALYSIS

There are various other factors that are responsible for road accidents. There are many meteorological factors that affect road accidents such as wind speed, temperature, wind direction, visibility, and precipitation.

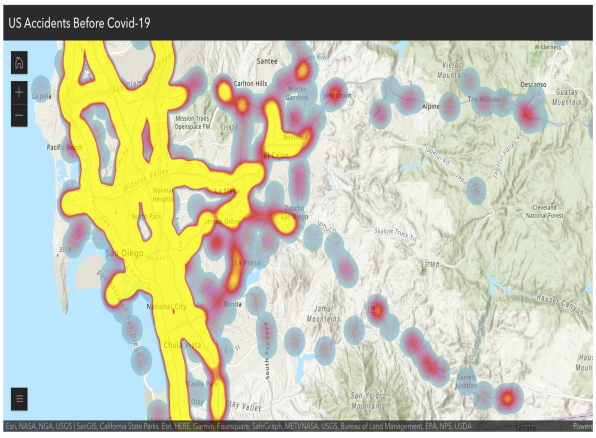


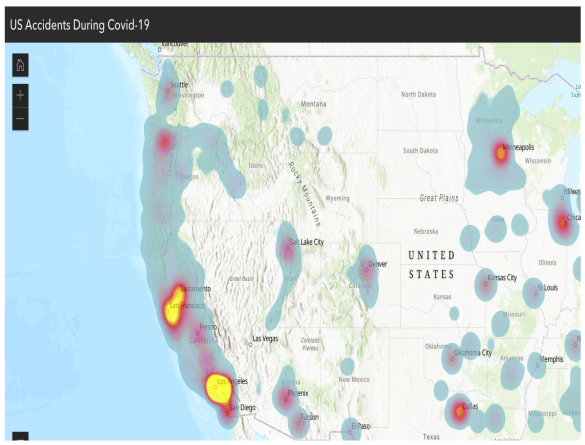


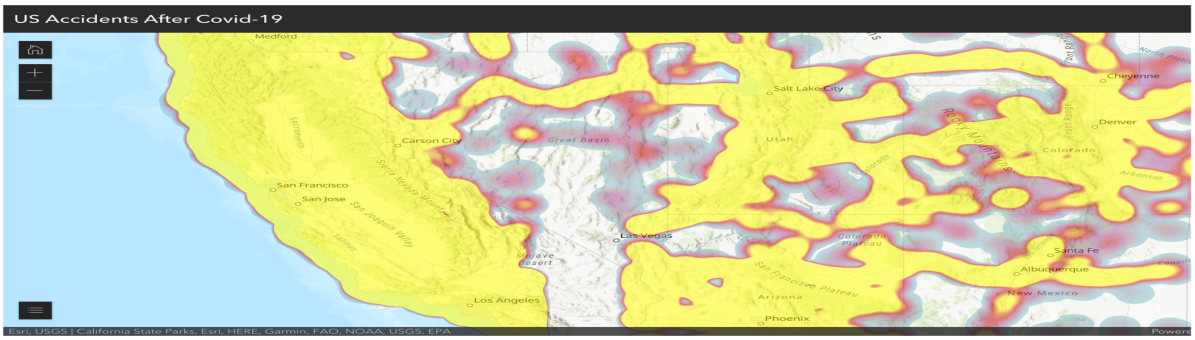
It can be inferred from the above plot that the major number of accidents took place when the wind speed is negligible. Also from the above bubble plot, it can be seen that a good number of accidents took place when the temperature range is 50 to 70 Fahrenheit which is normal temperature. Hence it is concluded that the temperature and weather conditions don't make a huge impact on road accidents. As the weather cannot be controlled, human behavior is essential in these situations. With self-discipline, accidents can be reduced.

6.4. COVID-19 IMPACT

Covid-19 has dramatically changed everyone's lives more than we can imagine. To study how covid-19 pandemic has changed accident trends and patterns, we have created heatmaps for three time periods, viz., before the pandemic(Feb'16 - Feb'20, during the pandemic(Feb'20 - Dec'20), and after the pandemic(Dec'20-Dec'21).





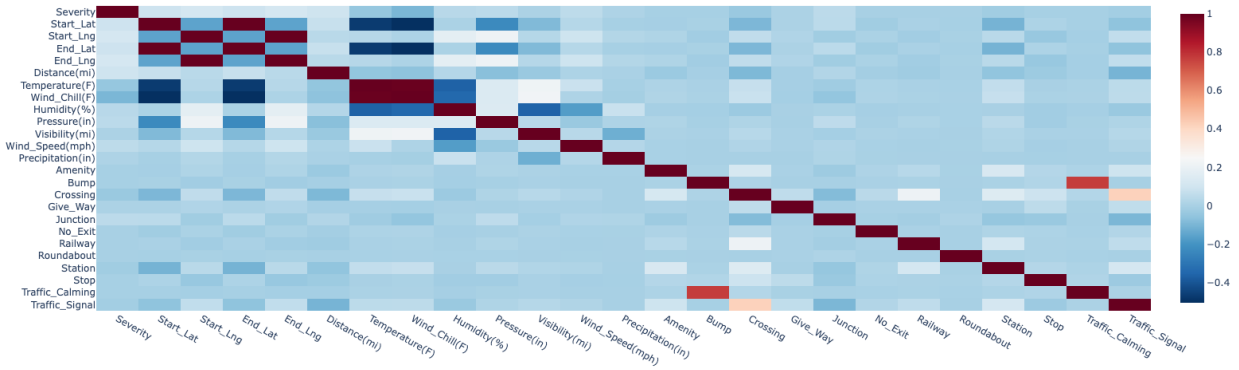


By observing the above heatmaps, we can conclude that accidents have declined during the pandemic due to the imposed lockdown restrictions but have risen more than ever following the lockdown in 2021. The US Government Accountability Office reported that "The Department of Transportation's National Highway Traffic Safety Administration (NHTSA) reported that traffic fatalities during the first half of 2021 increased by 18.4% since the first half of 2020. The estimated 20,160 deaths during the first half of last year are the highest since 2006".

7. SEVERITY PREDICTION MODEL

The last part of our project consists of predicting the traffic severity of the accident. As there are a total of 2.8 million rows and 47 columns, For doing a predictive model we do not need all the features available in the data. For example Id, state, country, and zip code may not affect the severity of the accident so those variables were removed.

Then we observed the correlation between all the variables and plotted them and the results are shown in the below correlation matrix.

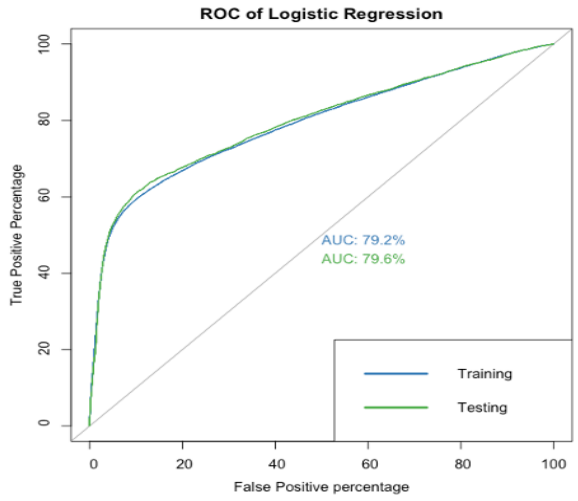


From the image, it can be seen that the start and end latitudes and longitudes are correlated to each other, so they were not considered in the end parameters for the modeling. Then the null

values were dropped from the data as they are very low compared to the length of the data we dropped them. The next problem is there are many object data types and boolean data types in the data available, as the models can't handle this data we converted them to 0 and 1 using the label encoding technique. Also after the data cleaning is completed the data looks very odd as some of the parameters have high median and some have very low so MinMaxScaler() was used to normalize the data.

The data is trained and tested on different machine learning algorithms and the results are given below

Model	Accuracy	AUC (Training)	AUC(Testing)
Logistic Regression	78	79.2	79.6
Naive Bayes	58	65	73.2
K-Nearest Neighbor	67	72.8	52.2
Decision Tree	73	83	76.6



Ultimately, logistic regression performed well on this data and the ROC of that model is as given above.

8. LIMITATIONS AND CHALLENGES:

- The dataset is based on accident data that are only recorded and hence sometimes there may be a chance that some of them were not recorded or could not be collected which could not be accounted for in the analysis and prediction model.
- Since the number of records is very large, the models took a very long time to train.
- As the data doesn't contain the population, analyzing accidents per population density became difficult.
- As the data is more leaned toward severity 2, improving the model performance is a very tough task.

9. CONCLUSION

This study aims to understand the significance of the factors that contribute to vehicle accidents and offer remedies to reduce them. Results showed that the major number of accidents are happening in the evening 3 to 5 pm when adults or students are either traveling to or from work, school, or returning home during rush hour, accidents seem to happen more frequently.

Additionally, more populated cities tend to have accidents. The effects of different weather, wind speed, visibility, temperature, and lightning on accidents were minimal. Other attributes present in the dataset like state, city, and street will help the government to predict severity ahead to lessen the traffic. Refer to https://sites.google.com/sdsu.edu/titans/ for more information and interactive dashboards.

10. FUTURE DEVELOPMENT PLAN

As far as future work goes, The prediction model needs to be deployed, and there is a need to feed live data to improve the performance of the model. The prediction model may include several neural network-based elements that make use of a range of data attributes, including traffic events, weather data, location-based data, and timing-related info.

Future studies could look at several countries to see if they will have equivalent results and add more details about the driver and the vehicle, age, gender, occupation, type of vehicle, and ownership information are all important data for accident analysis and modeling.

11. REFERENCES

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 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.
- https://www.gao.gov/blog/during-covid-19-road-fatalities-increased-and-transit-ridership-dipped.
- https://www.researchgate.net/publication/354238856_Road_Traffic_Accident_Data_Analysis_and_Its_Visualization.