Car Accident Severity Analysis

Sahil Rajendrakumar Raut
Rutgers University
Piscataway, NJ, USA
sahil.raut@rutgers.edu

Kaushik Vakadkar
Rutgers University
Piscataway, NJ, USA
kaushik.v@rutgers.edu

Sohailabbas Saiyed
Rutgers University
Piscataway, NJ, USA
ss3723@scarletmail.rutgers.edu

***Abstract*— Reducing traffic accidents is an essential public safety challenge all over the world; therefore, accident analysis has been a subject of much research in recent decades. The objective of the project is to analyze the US road accidents and provide an interactive platform which could be used by US government agencies and the public to quickly visualize trends and possible causes of traffic accidents and take relevant actions to reduce such events. We establish a multilevel interactive system to visualize trends over a million accident records in the United States across 49 states. Some of the visuals include the number of accidents by year, number of accidents by state, the best time to travel by month, day and hour, an accident-prone area in each state, factors responsible for the accidents like weather, wind flow, temperature, location, etc.**

***Keywords*— Traffic accidents, data visualization, data interaction, exploratory analysis, statistical analysis, dashboard, Plotly, Flask, HTML, CSS**

I. Project Description

The economic and social impact of road accidents is affecting U.S. citizens. hundreds of billions of dollars each year. And much of the loss is due to several serious accidents. Road crashes cost the U.S. $230.6 billion per year or an average of $820 per person [1]. Reducing road accidents, especially the worst accidents, however, remains an important challenge. The acceleration method, one of the two main ways to deal with road safety problems, focuses on preventing unsafe road conditions from happening in the first place. For effective use of this method, risk forecasting and difficulty forecasting are essential. If we can identify patterns of how these bad accidents occur and the key factors, we can take informed actions and better share financial and human resources. The primary objective of this project is to recognize key factors contributing to road accidents, without any detailed information about itself, like driver attributes or vehicle type. This visualization is supposed to be able to find the pattern of the accident in the United States. The aim of this project is to establish an interactive dashboard to demonstrate information about accidents and discover patterns with respect to weather conditions, location, and time of the year.

A. Data

Our dataset [2] is based on data collected from "A Countrywide Traffic Accident Dataset" by Moosavi et. al. [3]. This dataset has been collected in real-time, using multiple Traffic APIs. Currently, it contains accident data that has been collected from February 2016 to Dec 2021 covering 49 states of the USA. For this project, we are making use of a subset of the entire dataset by only considering data from the year 2021.

- Size: 619 MB
- Total number of records: 1,511,745
- Number of features: 20
- Format: CSV
- Streaming Data: No
- Time Variant: Yes

The goal of this project is to provide interactive statistical graphs which give detailed information of the data based on three major entities: Accident spatial position (State, City, Street, etc.), weather conditions (Temperature, Humidity, Wind Speed, etc.), and the time of occurrence (Month, Day, Hour, etc.). The ER diagram is shown in Fig. 1

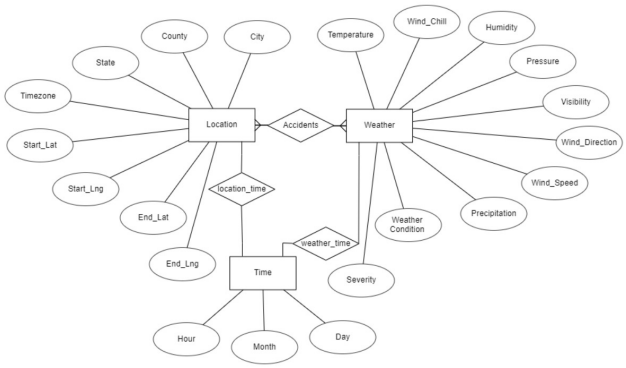

Fig. 1. The ER diagram.

B. Questions

1) Motivation:
Nearly 1.35 million people die in road crashes each year, resulting in an average of 3,700 deaths a day. In the USA itself, over 38,000 people die in road crashes each year, and 4.4 million are injured or disabled. These statistics urge for a way to analyze this issue of grave concern by providing a way to effortlessly visualize the data and trends.

2) What are the fundamental questions you want to answer about the data?
   - What key factors affect accident severity?
   - What's the impact of different weather conditions or environmental stimuli on severity of accidents in USA?
   - Are there any spatial patterns in terms of area size?
   - What are the accident hot spot locations?
   - Trend analysis based on various attributes

3) What is the fundamental data representation that will allow the system and users to query extract information about the data?
The dashboard will allow the users to interact with the data and the visualizations via drop down menus, zooming, panning and more interactive features. Based on the option selected from the drop down menu, the corresponding plots and visualizations update in a coordinated and synchronized fashion.

4) Create a name that faithfully describes the purpose of your interactive system

Car Accident Severity Analysis. The purpose of the system is to enable its users to explore, analyze and draw insights about car accidents in the US by using interactive visualization. Hence, the system has been named 'Car Accident Severity Analysis'.

5) Expected Target Users:
The main recommendations from the visualization focus on Infrastructure, Policy, Administrative, and Human behaviour-related changes that can be implemented by the state and the federal government. Government agencies and the public can leverage these insights and take preventive measures which can reduce road accidents in the US. Apart from this, the visualization can help state transportation agencies and logistics companies to select the safest and optimal route, weather, and time of the year to reduce losses.

C. Mode of Processing

1) Data Representation:
The Data which is downloaded from the Kaggle system is first processed using the Dataset.ipynb file. This preprocessed data is stored in CSV format and used as a static input to pandas data frame using python.
Data is processed in pandas data frame for plotting plots suing plotly.

2) Data Input:
The data is inputted as a static CSV file to the system.

II. Dashboard

A. Sections

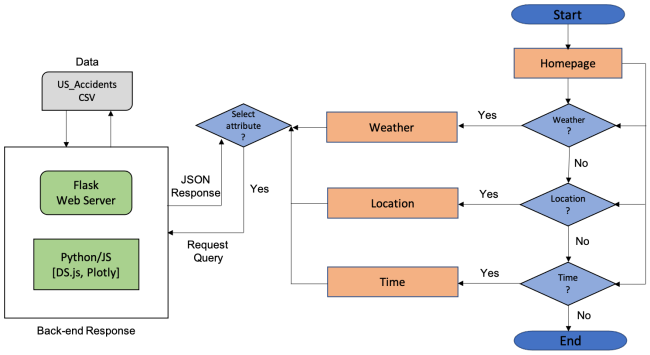

Fig. 2. System flow diagram

The application focuses on analyzing data across three primary entities - Weather, Location and Time. The system flow diagram is shown in Fig. 2. The navigation between these sections and the details of each, have been been explained below.

Our Application website is distributed into 5 sections. The user can select from these options through a menu bar on the top right side as shown in Fig. 3.

- **Home**: The homepage gives us a generalized summary of the project with links to the dataset and source code for the dashboard (Fig. 3).
- **Weather**: The 'Weather' page visualizes trends across different weather conditions like temperature, humidity, wind speed, visibility, etc. that impact the number of accident cases across the country (Fig. 4).
- **Location**: The 'Location' page provides granular analysis starting from Timezone to States to Cities, right up till Streets. In this section we analyze these four features based on the number of accident cases for each distinct location (Fig. 5).
The Page also gives the user the ability to filter on Month and Weather for additional interactivity. The user is able to analyze all the plots for filtered values. These filters are global filters. Along with the filters the plots have tool tips, pan, zooming features for detailed analysis and insights.
- **Time**: The 'Time' page provides an in-depth analysis on the number of accident cases based on the month of year, day, and hour. In this dataset we have `Start_Time` and `End_Time` for the timings of each accident. `Start_Time` shows start time of the accident, while `End_Time` provides the end time of the accident in local time zone (Fig. 6).
- **More (Prediction Analysis)**: The goal here is to identify and precisely predict what factors are majorly responsible for causing road accidents in the US.

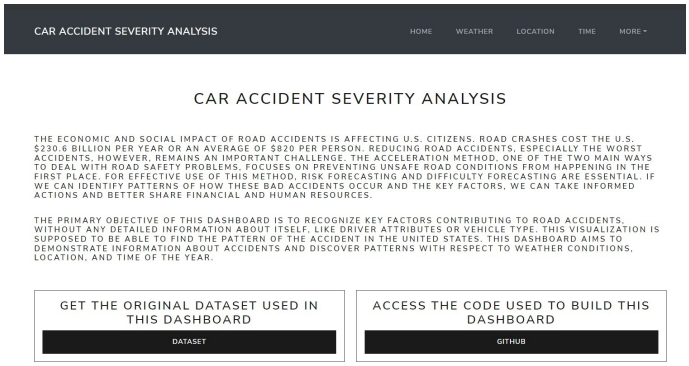

Fig. 3. Home page

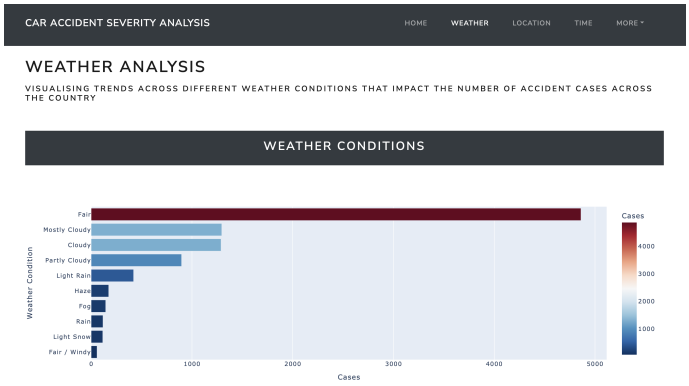

Fig. 4. Weather page

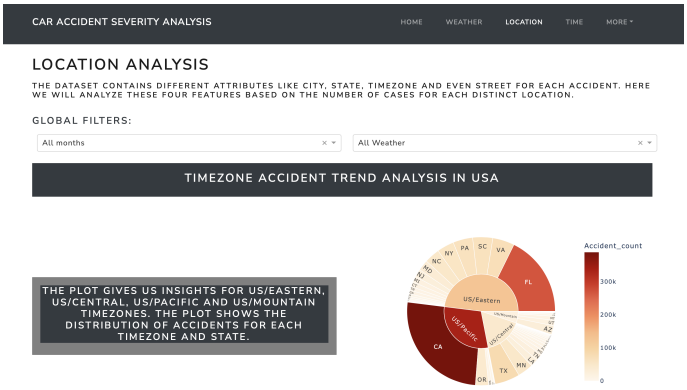

Fig. 5. Location page

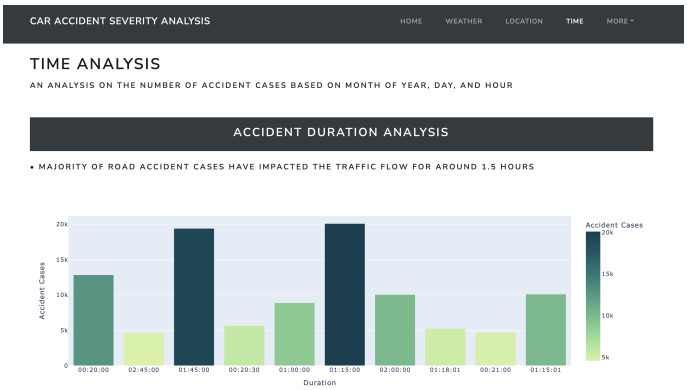

Fig. 6. Time page

B. Visual Representation

Different kinds of plots have been plotted to visually represent the data and uncover interesting trends. These include bar plots, donut plots, pie charts, maps, tree maps, and funnel plots.

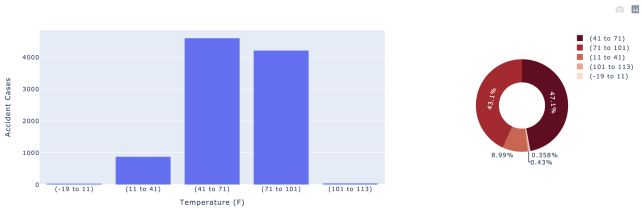

Fig. 7. Temperature (F) vs. Accident Cases

The barplot and the donut plot shown in Fig. 7 gives us insights on the effect of temperature on the number of accidents caused in USA. It is seen that 47% of cases occurred in the temperature range of 41°F to 71°F.

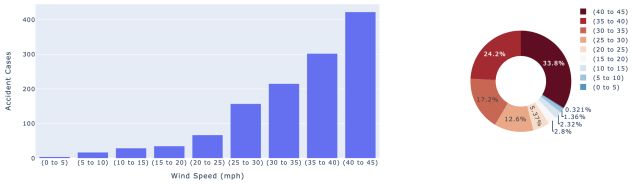

Fig. 8. Wind Speed (mph) vs. Accident Cases

As for the effect of wind speed on road accidents (Fig. 8), a positive correlation can be seen with a majority of cases occurring when the speed is between 40 and 45 mph.

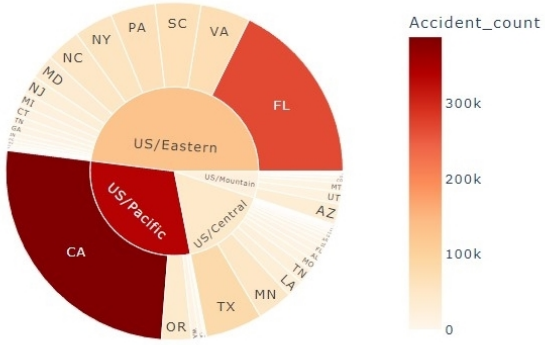

Fig. 9. Timezone Analysis for accidents in USA.

The Sunburst plot (Fig 9) gives us analysis about the distribution of accidents across different time zones US/Eastern, US/Central, US/Pacific and US/Mountain in USA and its states.

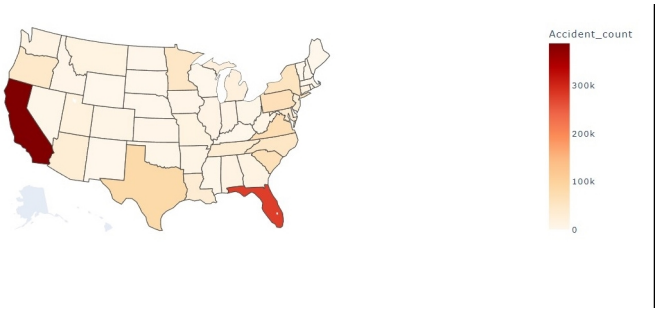

Fig. 10. State wise accident trend in USA.

The map plot (Fig. 10) gives us insights about the geographical distribution of accidents across each state in the US.

Fig. 11. Top 10 States contributing to accidents.

The bar plot shown in Fig. 11 shows the trend for the top states prone to road accidents. California (CA) has the most number of accidents followed by Florida (FL) and Texas (TX) when we consider the whole year.

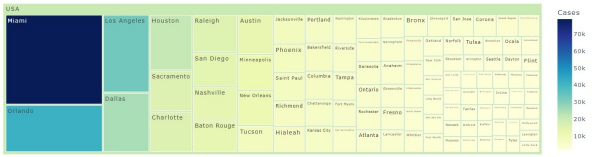

Fig. 12. City wise accident trend analysis.

The treemap (Fig. 12) gives the distribution of accidents across cities in USA.

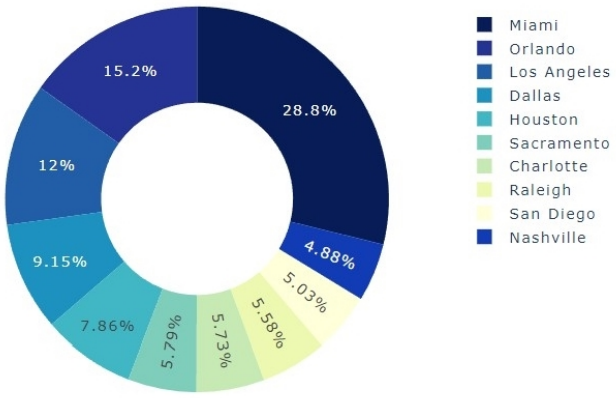

Fig. 13. Top 10 Cities with the highest number of Accidents

The Plot (Fig. 13) gives insights for the Top Cities prone to accidents. Government can focus more on these states to build stricter laws and take actions to prevent accidents.

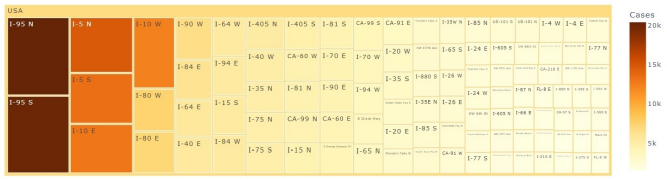

Fig. 14. Street wise Accident trend analysis in USA

The Treemap (Fig. 14) gives the distribution of accidents across streets in USA.

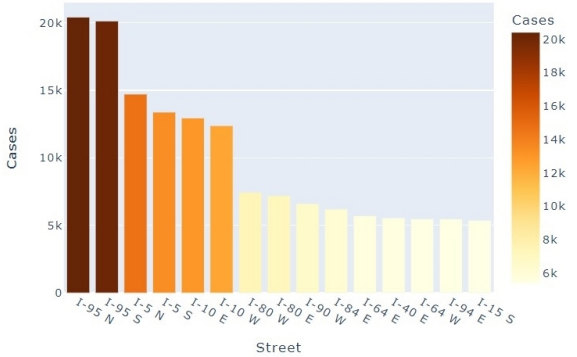

Fig. 15. Top 15 Streets with the highest number of Accidents

The Plot (Fig. 15) gives insights for the Top streets prone to accidents. Interstate 95 S and N are the most accident prone followed by Interstate 5 N and S for the year. The Top 15 streets constitute a large amount of accident cases.

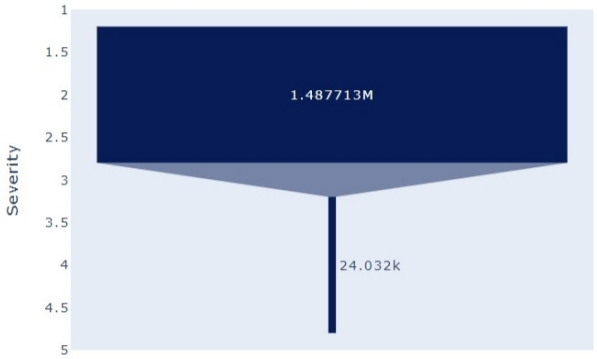

Fig. 16. Severity analysis of Accidents in USA

The Plot (Fig. 16) gives insights of the severity of accidents that takes place in USA. A large amount of accident populations is of Moderate Severity for the year 2021.

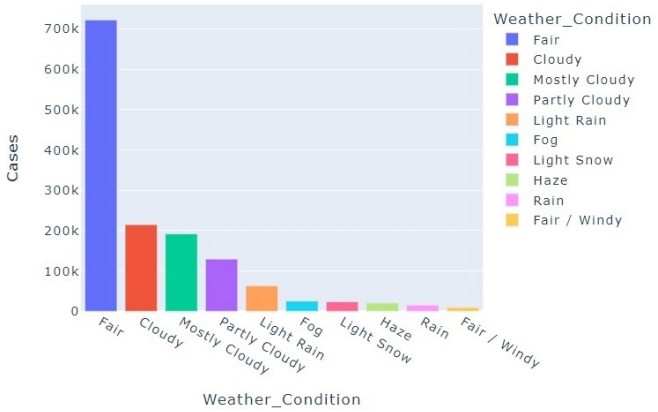

Fig. 17. Weather analysis of Accidents in USA

The Plot (Fig. 17) gives insights of the top 10 weather contributing to accidents that takes place in USA.

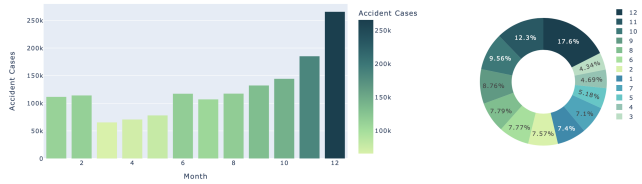

Fig. 18. Month vs. Accident Cases

The Barplot and Donut plot (Fig. 18) gives the distribution and percentage of accidents over each month.

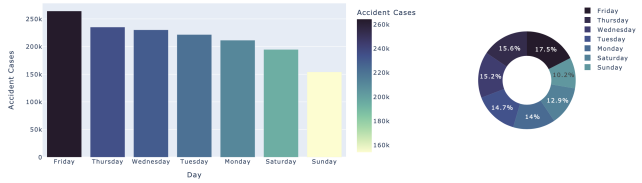

Fig. 19. Day vs. Accident Cases

The Barplot and Donut plot (Fig. 19) gives the distribution and percentage of accidents for each Day in a week.

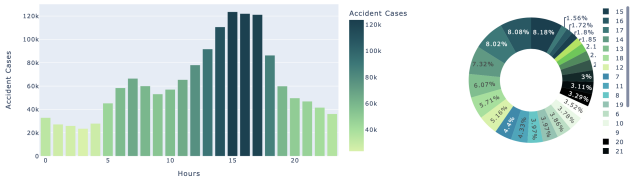

Fig. 20. Hour vs. Accident Cases

The Barplot and Donut plot (Fig. 20) gives the distribution and percentage of accidents over the course of 24 hours in a day.

C. Interactivity

Each action on average takes anywhere between 1 to 2 seconds to render the plots. Depending on the selected filter and the plot, the rendering time can either be low or high.

- **Interface Layout**: The layout of the system has been described in Section II-A and is shown in Fig. 3. It is a dashboard having a navigation bar on the top right side.
- **Interface Interaction mechanisms and answers representation for further user interaction**:
  - Mouse Clicks, Mouse Hovering, Mouse Selection and their combinations: These features are present in every plot on our dashboard. Upon hovering over the bar plots, map, pie chart, or even tree maps, the value of the corresponding item is displayed.
  - Menu Driven Bottoms and Tabs: In addition to the primary navigation menu, the 'Location' page has been provided with two global filters as drop-down menus to further drill down based on the month and weather condition.
  - Zooming and panning: These two features have been employed in all the plots. Zooming is especially useful in the case of the lower levels of the tree maps.
  - Linking of different views: All the plots in the 'Location' page are linked with each other and coordinated with the help of the two global month and weather condition filters.

D. Data Analytics

The Data Analytic is done in python. The implementation done over here is of logistic Regression.we have applied logistic regression to identify which accidents are fatal and which are not.We have Use advanced re sampling techniques like SMOTE and AD SYN to handle the imbalanced data.The main objective of Data Analytic was to identify the weather conditions,time and locations which causes these accidents.It helps Road authorities, citizens and insurance companies to take informed decisions.Steps followed in implementing the Machine Learning models are as below:

- Data Collection
- Data Preprocessing
- Exploratory Data Analysis (univariate bivariate)
- Modeling
- Evaluation
- Deployment

Our data set is divided into two major parts, accident spatial position and the corresponding weather conditions. Therefore, we use attributes from both these parts to plot charts for the following statistics.

- Standardize the data type
- Remove null values and incorrect records
- One Hot encoding for Categorical Variables
- Standard Scaling

A severity map highlights the severity of the accident in terms of geographical location

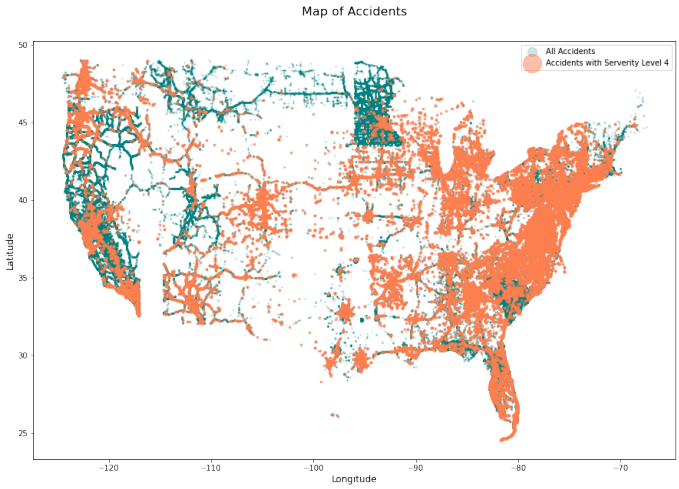

Fig. 21. Severity Map.

A correlation plot is used to analyse the multi-collinearity between the features.

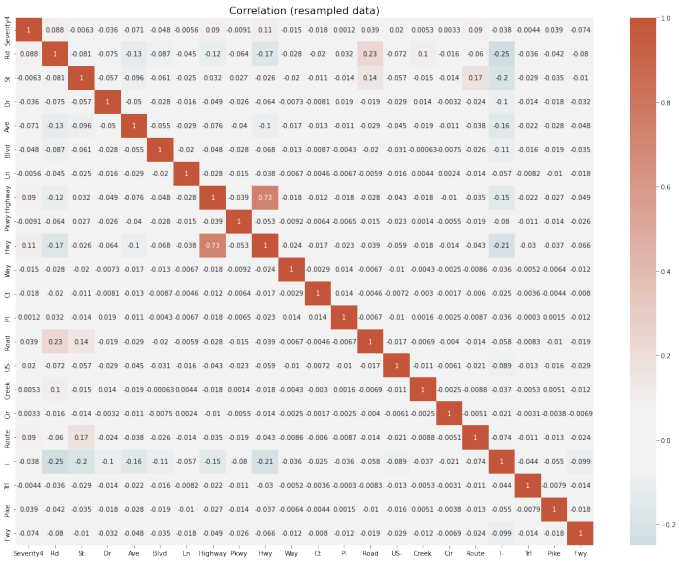

Fig. 22. Correlation Plot.

A confusion matrix describe the prediction power of our model and shows how many False Positive and False negative exist in the data.

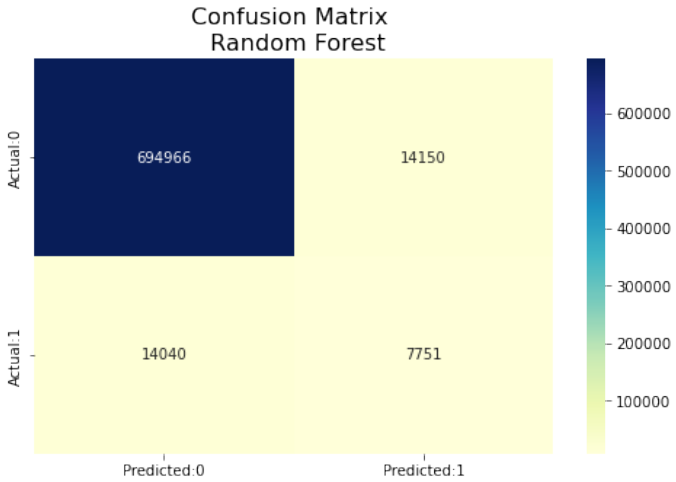

Fig. 23. Top 10 States by the number of accidents.

AUC score shows the balance between False positive vs False negative rate.

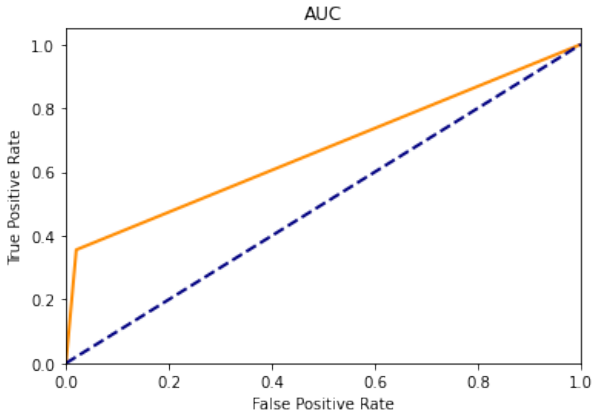

Fig. 24. AUC Curve

Precision Recall curve shows the balance between Precision and recall metrics.

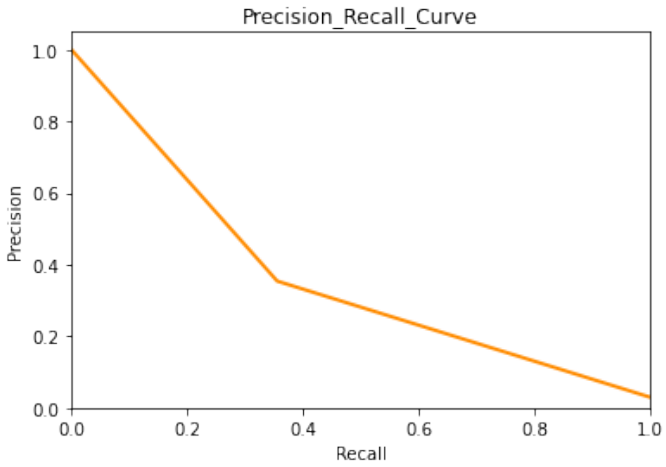

Fig. 25. PRC curve.

E. Development

The components forming the front-end and back-end along with the flow of data and query between them can be seen in Fig 26. The data is read from the CSV flat file. As for visualization, Plotly is primarily used to generate all the plots. In order to create a menu-driven interactive dashboard, HTML/CSS, Bootstrap, and JavaScript has been used.

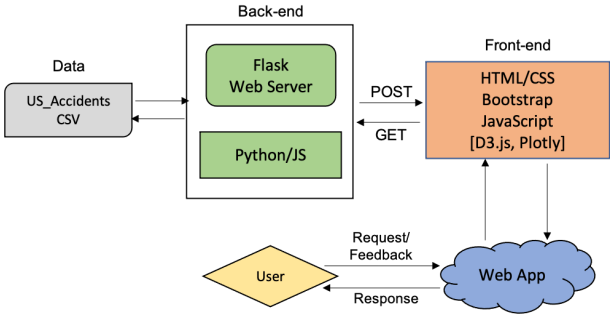

Fig. 26. Component integration diagram.

- Language(s) used:
  - Backend: Python
  - Frontend: HTML 5, CSS, JavaScript
- Software platform (Back-end + front-end)
  - Data: CSV flat file
  - Web server: Flask [4]
  - Front-end framework: Dash Bootstrap Components
- Software libraries
  - Plotting and visualization libraries: Plotly [5], Matplotlib, Seaborn
  - Analysis: NumPy, Pandas, Scikit-learn, SciPy

Division of Labor:

- Sahil Raut: Dataset manipulation, prototype, location analysis
- Kaushik Vakadkar: Dashboard front-end, weather analysis, time analysis
- Sohailabbas Saiyed: Predictive analysis

Project Timeline:

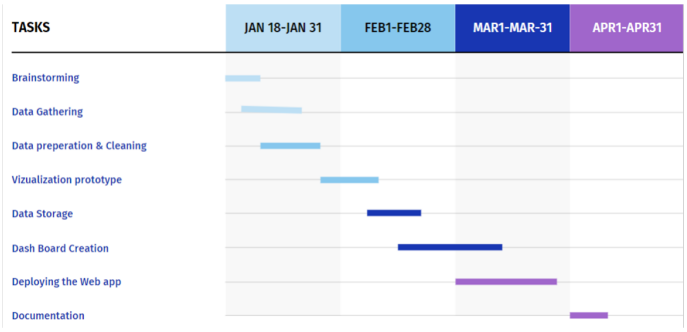

Fig. 27. Gantt chart depicting tasks and estimated timeline

III. Project Highlights

Our system aims to uncover some essential insights from the data. Some of the most important questions to be answered include:

1) What key factors affect accident severity?
2) What's the impact of different weather conditions or environmental stimuli on severity of accidents in USA?
3) Are there any spatial patterns in terms of area size?
4) What are the accident hotspot locations?

IV. Future work

The focus of our current system is on visualizing patterns found in the data and facilitating exploratory analysis. If time permits, we would like to incorporate the following features to further augment the system's functionality as well as usability.

- Prediction UI: Although we have performed exhaustive predictive analysis by employing logistic regression to accurately predict the most probable weather conditions that cause road accidents in the US, we plan to build a user-interface on the dashboard for the same. This would enable user driven inputs to generate the predicted output.
- User Query Interface: With a user query interface, researchers from different disciplines can apply their own analysis methods to explore the data. The data can be accessed by querying some attributes like the weather condition, location, day of week, etc.
- Feedback: Through a feedback interface we can gather valuable information from different types of users, which can help us to constantly improve user experience, interface layout, and the findings they are interested in.

V. Acknowledgements

We would like to thank Professor James Abello who has given excellent guidelines and has constantly motivated us throughout the project. We would also like to thank Haoyang Zhang for answering and clearing out all queries regarding the project.

References

[1] (2022) Road safety facts. [Online]. Available: https://www.asirt.org/safe-travel/road-safety-facts/
[2] Us accidents (updated) - a countrywide traffic accident dataset. [Online]. Available: https://www.kaggle.com/sobhanmoosavi/us-accidents
[3] S. P. Sobhan Moosavi, Mohammad Hossein Samavatian and R. Ramnath. (2019) A countrywide traffic accident dataset. [Online]. Available: https://arxiv.org/pdf/1906.05409.pdf
[4] Flask. [Online]. Available: https://flask.palletsprojects.com/en/2.0.x/
[5] Plotly python open source graphing library. [Online]. Available: https://plotly.com/python/