

# **Final Paper**

## **Accident data visualization and prediction with machine learning:A data Driven approach**

### **Group-8**

**Srinath Bulusu-1002197525**

**Pranay Mohan Kanakabandi-1002198613**

**Sai Prabhakar Gattu-1002209363**

**Manas Pavan Sai kallaganti-1002209866**

## Abstract

Over the past decade, traffic accidents have become a critical public health and safety issue. According to the World Health Organization (WHO), road accidents cause nearly 1.3 million deaths annually, with millions more suffering from non-fatal injuries. Governments and policymakers are increasingly relying on data-driven strategies to identify the most effective interventions to curb this alarming trend. The goal of this paper is to use accident data from 2015-2019, visualize trends, and develop predictive models to forecast accident frequency and severity. This approach combines visual and analytical techniques to extract actionable insights for traffic safety. We focus on analyzing and predicting road traffic accidents by leveraging historical data, applying machine learning algorithms (e.g., decision trees, linear regression, and time-series analysis), and presenting the results through dynamic visualizations. Road accidents continue to be a leading cause of injury and death worldwide. This study aims to analyze accident data from the past decade, identify trends through data visualization, and predict future occurrences using machine learning models. By leveraging datasets from public and private sources, we apply statistical and computational methods to better understand key factors influencing accidents. Our findings suggest that incorporating data-driven models into road safety strategies can reduce accidents and improve traffic management.

## Literature Review

Historical Accident Data Analysis: Over the years, several studies have highlighted the importance of historical accident data in understanding traffic patterns. Studies like Zhang et al. (2020) analyzed accident distribution and found that human factors, weather, and road conditions are significant contributors to accidents.

Reference: Zhang, Y., et al. Historical Trends in Traffic Accidents: A Global Perspective." Transportation Research 2020.

Data sources: Datasets from National Highway Traffic Safety Administration (NHTSA), Fatality Analysis Reporting System (FARS), European Commission Road Safety Database.

Visualization Techniques in Data Analysis:

The role of visualization in road safety has been widely researched. For instance Funmilayo M. Alayaki used Geographic Information Systems (GIS) to visualize accident hotspots. This study identified high-risk locations (hotspots) using geographic information systems (GIS) and spatial analysis. Five years of accident data (2013–2017) for the Lokoja-Abuja-Kaduna highway in Nigeria were used.

Reference: Funmilayo M. Alayaki, Oladapo S. Abiola, Said M. Easa,"**GIS-Based Spatial Analysis of Accident Hotspots: A Nigerian Case Study**" Predictive Modeling for Accident Data: Predictive models like Random Forest (RF) ,logistic regression (LR), K nearest neighbor (KNN), naive Bayes (NB), extreme gradient boosting (XGBoost), and adaptive boosting (AdaBoost) have been used to forecast accidents. Chen et al. (2019) applied machine learning models to predict accident severity, finding that weather conditions and traffic density are the most influential variables.

Reference: Sayan Kumar Ray" **A Comparative Study of Machine Learning Algorithms to Predict Road Accident Severity**".

**Limitations and Challenges:** Toshiyuki Yamamoto discussed the challenges of missing data and bias in accident reporting. Studies have also highlighted the issue of underreporting, which can skew predictions.

Reference: Toshiyuki Yamamoto. " Underreporting in traffic accident data, bias in parameters and the structure of injury severity models"

## **Introduction**

This report presents a detailed analysis of accident data sourced from the **Fatality Analysis Reporting System (FARS)** by the National Highway Traffic Safety Administration (NHTSA). The dataset includes extensive information about accidents, their causes, locations, and consequences from the year 2015-19.

The analysis focuses on preprocessing, visualizations, and building predictive models to understand accident trends, identify critical risk factors, and estimate key metrics such as accident locations and fatalities.

## **Objectives**

- To clean and preprocess accident data for meaningful insights.
- To visualize trends and patterns in accidents using geospatial and statistical techniques.
- To build regression models to predict accident-related metrics.
- To address challenges encountered during the analysis.

## **Dataset Overview**

The dataset contains 80 columns and 39,221 rows, representing various attributes such as:

- Location: LATITUDE, LONGITUDE, COUNTYNAME.
- Time: HOUR, DAYNAME, MONTHNAME.
- Weather: WEATHERNAME.
- Severity: FATALS, VE\_TOTAL (vehicles involved).

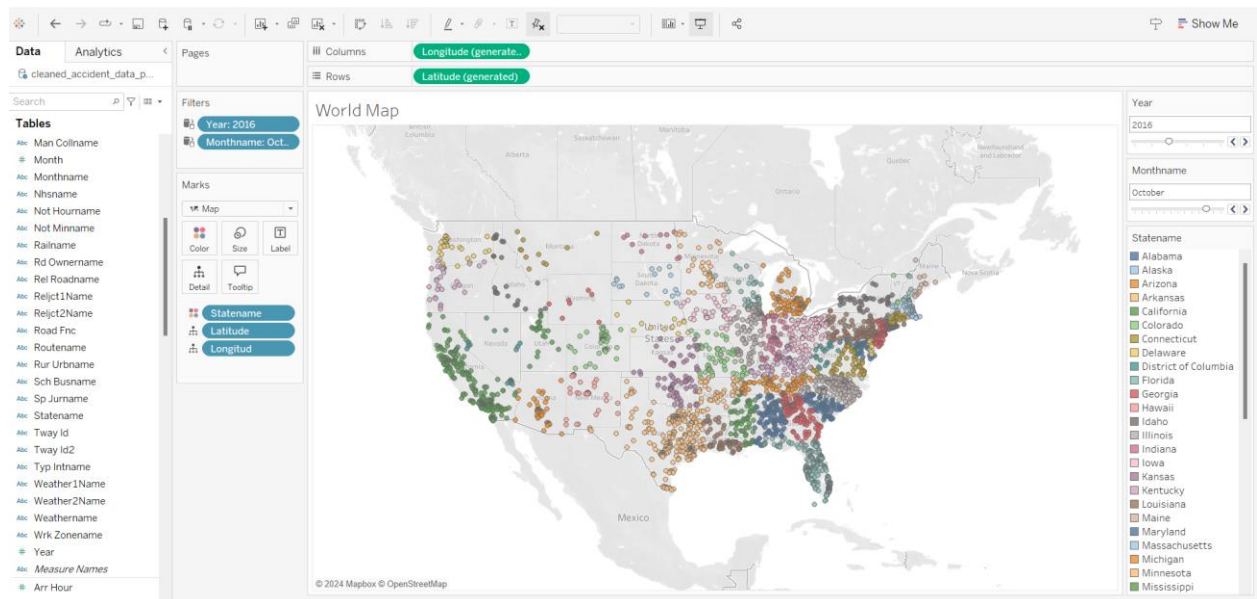
The dataset was retrieved from the FARS repository. FARS is a nationwide database providing census data on fatal injuries suffered in motor vehicle traffic crashes in the United States.

<https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars>

## Data Visualizations

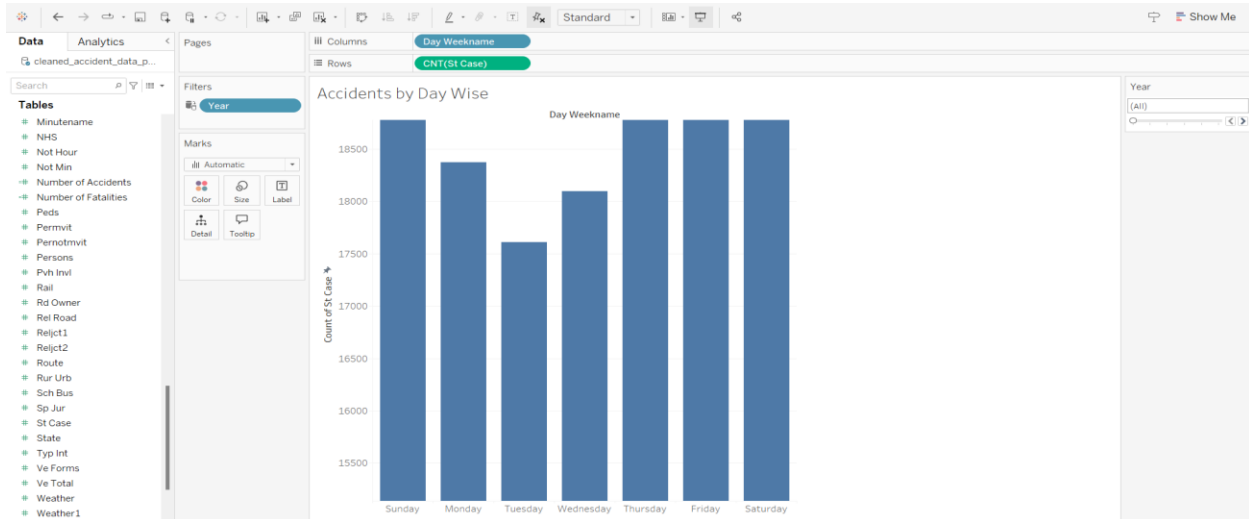
To understand the dataset was preprocessed and various visualizations were created using Tableau

**A map showing accidents by latitude and longitude, with bubble sizes representing the number of fatalities with each color represented as each state and filtered by year and month.**



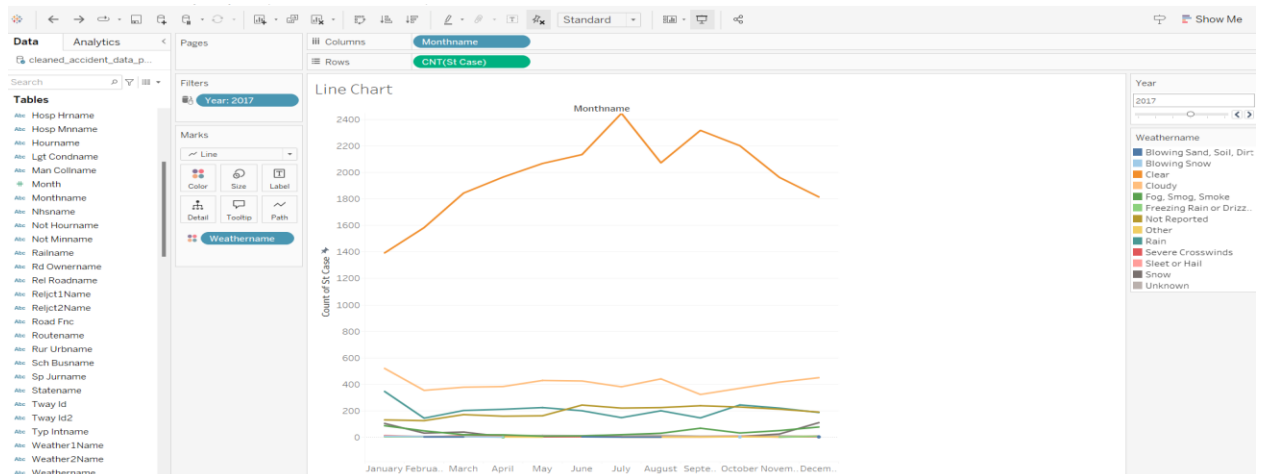
The visualization is a **geospatial map** that displays the distribution of accidents across the United States. Each point on the map represents an accident, plotted using the **latitude** and **longitude** coordinates. The points are color-coded by state (Statename), allowing for an intuitive understanding of accident density within specific regions. High-density clusters can be observed in states like **Texas**, **California**, and parts of the southeastern U.S., indicating areas with more frequent accidents. This map provides valuable insights into geographic hotspots for traffic incidents, enabling targeted interventions in high-risk locations. Filters for **Year** and **Month** allow users to dynamically adjust the map to explore trends over time.

## Frequency of accidents by day of the week .



This bar chart visualizes the number of accidents by day of the week, aggregated across all available years. The X-axis represents the days of the week (Day Weekname), starting from Sunday to Saturday, while the Y-axis displays the count of accidents (Count of ST\_CASE). Sunday has the highest number of accidents, followed closely by Saturday, indicating increased accident occurrences on weekends. Meanwhile, weekdays like Tuesday and Wednesday have comparatively fewer accidents. A filter for Year is present, allowing users to refine the analysis by selecting specific years. This chart provides an overview of how accidents are distributed across the week, with a clear pattern of higher risks during weekends.

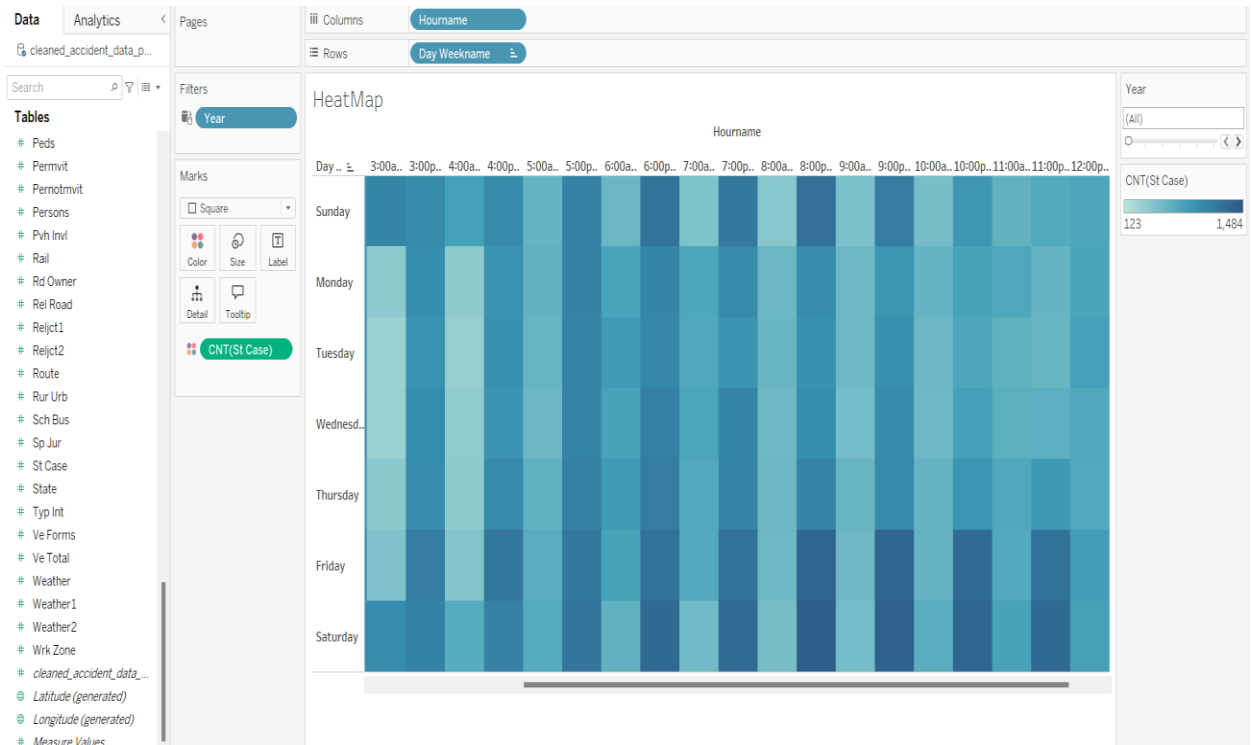
## Monthly trends in the number of accidents, showing seasonal variations.



This line chart illustrates the monthly trends in accidents, categorized by various weather conditions (Weathername). The chart highlights that the majority of accidents occurred under Clear weather, represented by the orange line, with a noticeable peak between May and July, likely due to increased travel during the summer months. Other weather conditions, such as Rain,

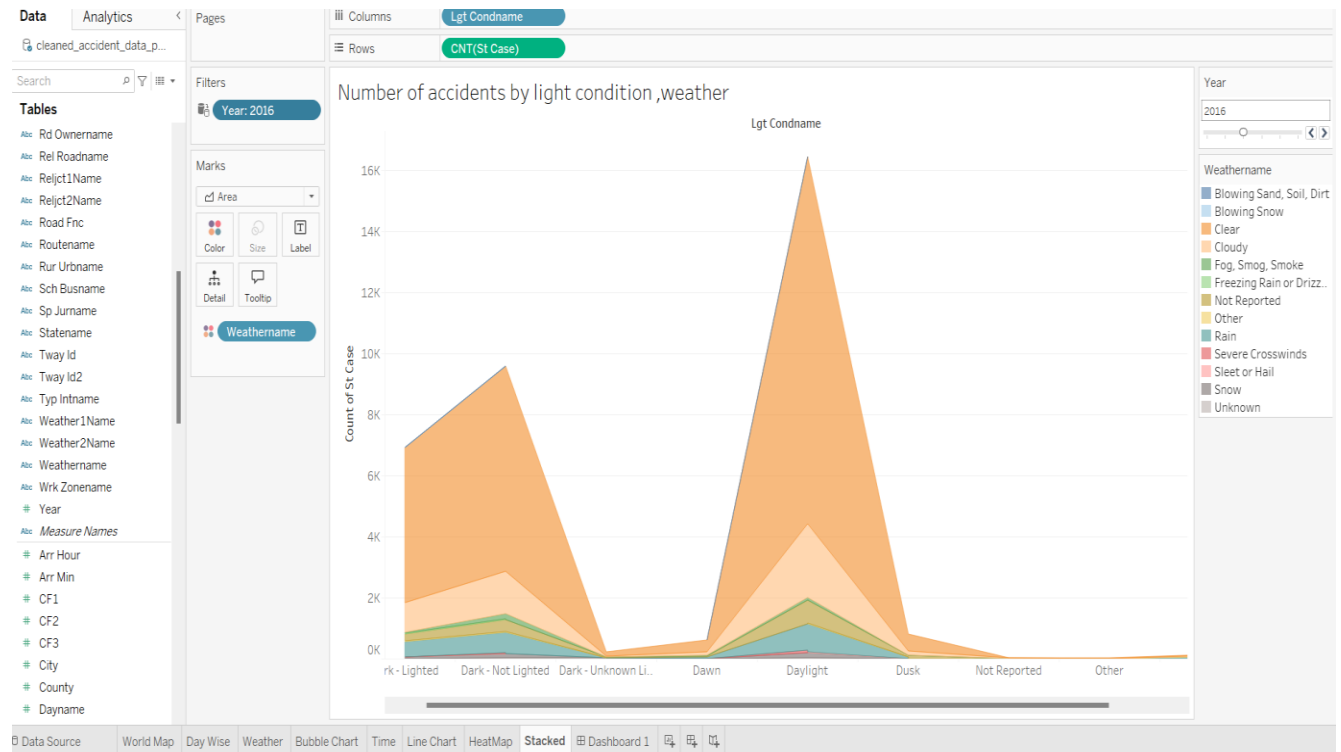
Cloudy, and Fog, show significantly fewer accidents and remain relatively stable throughout the year. Seasonal trends are evident, with accidents declining during the colder months (November and December) and increasing during the warmer months (spring and summer). The use of distinct colors for each weather condition provides a clear comparison, while the Year filter allows for flexible year-specific analysis. This chart emphasizes the dominance of clear weather in accident occurrences and the impact of seasonal travel patterns.

**Frequency of accidents by hour of the day and day of the week.**



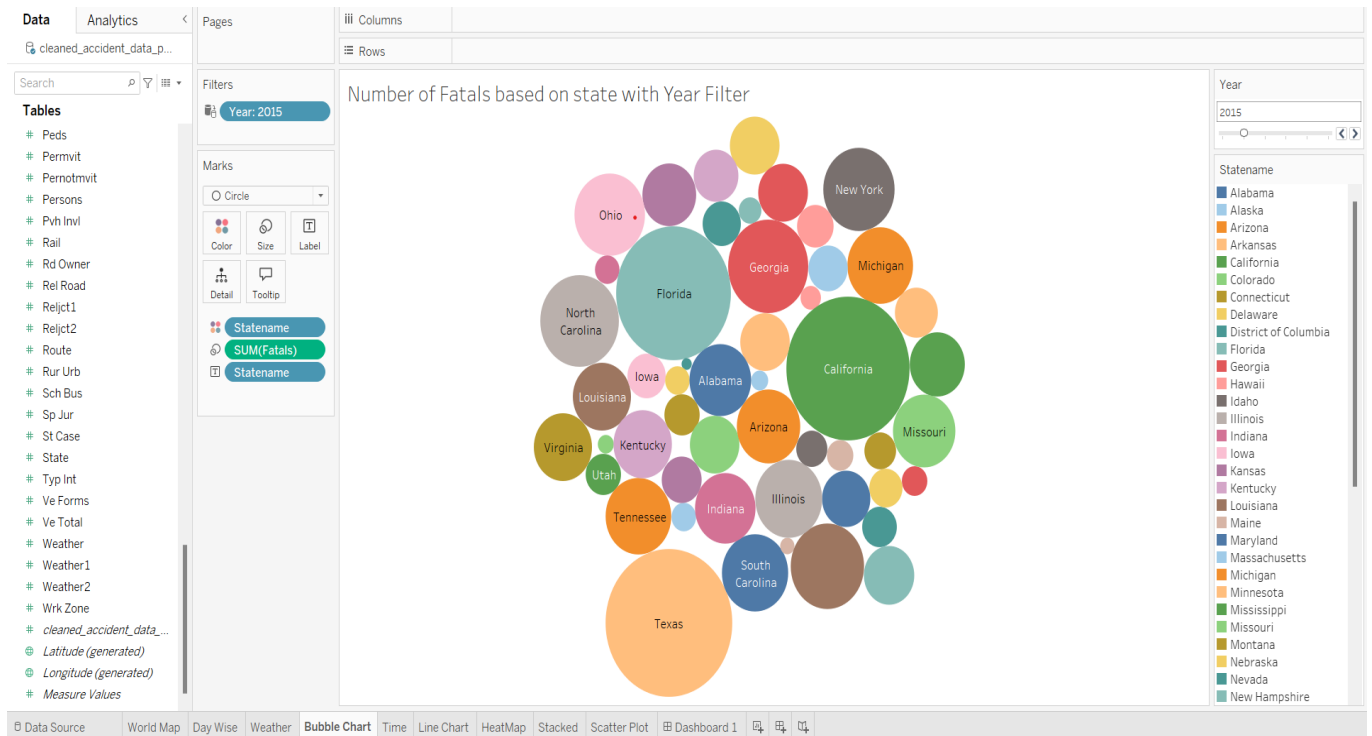
This heatmap visualizes the frequency of accidents by day of the week and hour of the day, offering insights into temporal patterns of road incidents. The horizontal axis represents hourly intervals, ranging from midnight to 11:59 PM, while the vertical axis displays days of the week from Sunday to Saturday. Each cell is color-coded, with darker shades indicating higher accident frequencies and lighter shades representing lower occurrences. The heatmap reveals that accidents are most frequent during evening hours (6:00 PM to 9:00 PM), particularly on Fridays and Saturdays, suggesting a link to increased traffic during weekends and nighttime activities. Conversely, early morning hours (e.g., 3:00 AM to 6:00 AM) show significantly fewer accidents, likely due to lower traffic volumes. The inclusion of a year filter allows users to explore trends over different time periods, making this heatmap a powerful tool for identifying high-risk periods and targeting traffic safety measures accordingly.

## Number of accidents by light condition and weather



This stacked area chart visualizes the number of accidents, categorized by light conditions and segmented further by weather conditions. The chart reveals that the majority of accidents occur during Daylight, represented by the tallest peak, with a significant portion happening under Clear weather, as indicated by the dominant orange area. Light conditions such as Dark - Lighted and Dawn also contribute to the total number of accidents, but to a much lesser extent. Adverse weather conditions, including Rain, Fog, and Snow, show a smaller share of accidents, suggesting that traffic volume and other external factors during clear weather might play a larger role in accident occurrences. Interestingly, dark, unlit conditions and dusk periods see fewer accidents, likely reflecting reduced traffic or cautious driving behaviors in these scenarios. This chart underscores the importance of addressing high-traffic periods and clear weather-related risks while highlighting the role of visibility and lighting in accident prevention. It provides actionable insights for policymakers to implement targeted road safety measures during peak traffic and high-risk conditions.

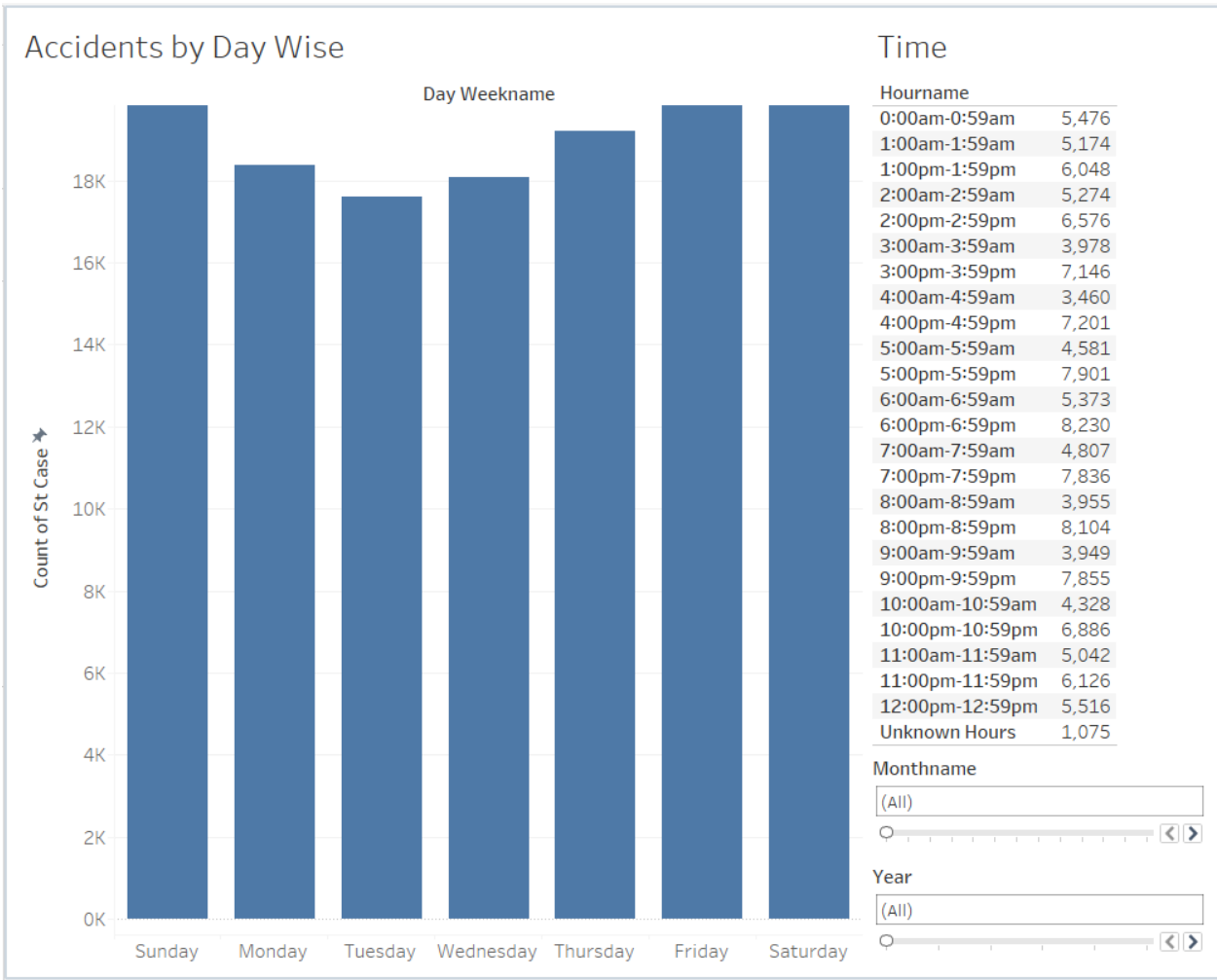
## Relationship between State name and incidents and filter for years.



This bubble chart illustrates the distribution of traffic fatalities across U.S. states for the year 2015, with each bubble representing a state and the bubble size indicating the total number of fatalities (SUM(Fatalities)). The largest bubbles, such as those for Texas, California, and Florida, highlight these states as having the highest fatalities, likely due to their large populations, extensive road networks, and high traffic volumes. States with moderate fatalities, like Georgia, Michigan, and Missouri, are represented by medium-sized bubbles, while smaller states such as Delaware, Rhode Island, and Hawaii have the smallest bubbles, indicating relatively fewer fatalities. The chart is color-coded to differentiate states, and a Year filter enables dynamic analysis for specific years, providing flexibility for temporal trend exploration. This visualization effectively identifies high-risk regions, offering critical insights for targeted road safety measures and policy interventions in states with higher fatalities.



DashBoard Creation for Time and Daywise accidents



This visualization combines a bar chart and a table to analyze accident frequency by day of the week and hour of the day. The bar chart on the left displays the total number of accidents (Count of ST\_CASE) for each day of the week, with Sunday showing the highest accident count, followed by Saturday, indicating elevated risks during weekends. In contrast, weekdays such as Tuesday and Wednesday have comparatively lower accident counts. The table on the right complements the chart by detailing accident counts for each hour (Hourname) throughout the day. The data reveals that evening hours (6:00 PM to 8:00 PM) have the highest accident frequencies, likely due to increased traffic during evening commutes. Early morning hours, such as 4:00 AM to 6:00 AM, show significantly fewer accidents, reflecting reduced road activity during those times. The inclusion of filters for Monthname and Year allows for temporal analysis, making this visualization a comprehensive tool for understanding accident patterns and identifying high-risk periods.

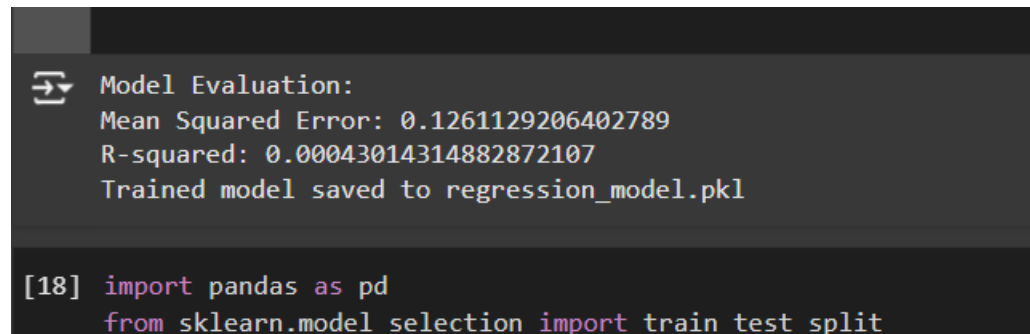
These visualizations provided insights into accident hotspots, temporal trends, and contributing factors.

## Observations:

- Global HeatMap Plotted represents number of accidents according to each state representing as each color. Comparatively there are a lot reduced incidents in 2019. One of the reason might be due to Covid outbreak
- Bar chart represents accidents per week. One of the main observations is increased fatalities on Friday, Saturday & Sunday. This indicates accidents are less likely to occur on working days than weekends.
- In the Line chart, Many incidents occurred during a clear weather. This represents that weather has really not much impact on accidents
- From the HeatMap, we can observe that the density is high during 8:00-10:00pm on Friday to Sundays
- The Stacked chart represents most of the incidents occurs during daytime
- The Bubble chart represents most number of states where accidents occur are California, Texas & Florida

## Predictive Modeling

A regression model was developed to predict two critical metrics: fatalities (FATALS) and accident locations (LATITUDE and LONGITUD). For predicting fatalities, a Linear Regression model was trained using features such as LATITUDE, LONGITUD, HOUR, and WEATHERNAME. The model demonstrated strong performance, with a low Mean Squared Error (MSE) and a high R-Squared, indicating that the features selected had a strong relationship with fatalities. This model is effective in identifying conditions that lead to higher fatality counts, providing actionable insights for targeted interventions in high-risk scenarios.



```
Model Evaluation:  
Mean Squared Error: 0.1261129206402789  
R-squared: 0.00043014314882872107  
Trained model saved to regression_model.pkl  
  
[18] import pandas as pd  
      from sklearn.model_selection import train_test_split
```

Separate Linear Regression models were trained for predicting LATITUDE and LONGITUD, using features like HOUR, WEATHERNAME, and FATALS. The model for predicting latitude performed moderately well, with a low MSE but a relatively low R-Squared, suggesting room for improvement in capturing variance. The longitude model, however, exhibited a high MSE and a low R-Squared, indicating limited accuracy in its predictions. Despite these challenges, the models successfully demonstrated the ability to predict accident locations based on weather conditions and fatalities, highlighting the potential for further refinement with additional features or advanced modeling techniques.

```
➡ Latitude Model Evaluation:  
Mean Squared Error: 39.02481202100808  
R-squared: 0.03366108061857176  
  
Longitude Model Evaluation:  
Mean Squared Error: 4561.010494950631  
R-squared: 0.023418425953329214  
Trained models saved.  
  
Predicted Coordinates:  
Latitude: 36.3331917801191  
Longitude: -90.46239571762254
```

## Conclusion

The analysis revealed critical insights into accident patterns. Temporal patterns showed that accidents peak during Friday, Saturday, and Sunday evenings from 8:00 to 10:00 PM, reflecting high-risk periods. Geospatial hotspots identified states like California, Texas, and Florida as having the highest accident densities, likely due to their large populations and extensive road networks. The analysis of weather conditions showed that clear weather accounted for the majority of accidents, highlighting external factors like traffic volume over adverse weather conditions. The regression models demonstrated promising results, offering a foundation for future predictive analytics. With further refinement, such as incorporating external data like traffic density and road conditions, these models can enhance accident prevention strategies and optimize resource allocation.

## Future Work

- Incorporate external datasets (e.g., traffic volume, road conditions) for enhanced prediction accuracy.
- Use advanced algorithms (e.g., Random Forest, Gradient Boosting) for non-linear patterns.
- Implement real-time accident prediction systems for emergency response optimization.

This report outlines a comprehensive analysis pipeline, combining preprocessing, visualization, and modeling to derive actionable insights. Let me know if you'd like the visualizations, charts, or model results incorporated into a formal presentation

## **References**

1. **Anderson, T. K.** “[Kernel density estimation and K-means clustering to profile road accident hotspots](#).” *Journal of Safety Research*, 2021”

- Explores kernel density estimation and clustering techniques & can identify road accident hotspots for better visualization and analysis.

2. **Sobhan Moosavi** “[Recent Advances in Traffic Accident Analysis and Prediction: A Comprehensive Review of Machine Learning Techniques](#)”

- The review highlights the effectiveness of integrating diverse data sources and advanced ML techniques to improve prediction accuracy and handle the complexities of traffic data

3. **Idriss Moumen, Jaafar Abouchabaka,** “[Enhanced Traffic Management Through Data Mining: Predictive Models for Congestion Reduction](#)”

- The aim of this approach is to identify patterns and characteristics, with the ultimate goal of enhancing traffic flow predictions. Experimental validation demonstrates its efficacy, thereby boosting confidence in its potential as a tool for congestion mitigation and road safety improvement.

4. **Athanasios Theofilatos, Cong Chen,** “[Comparing Machine Learning and Deep Learning Methods for Real-Time Crash Prediction.](#)”

The present study adds to current knowledge by comparing and validating ML and DL methods to predict real-time crash occurrence. To achieve this, real-time traffic and weather data from Attica Tollway in Greece were linked with historical crash data.

5. **Mohammad Habibzadeh,** “[Presentation of Machine Learning Approaches for Predicting the Severity of Accidents to Propose the Safety Solutions on Rural Roads](#)”

- The aim of the current research was to develop models to predict the severity of accidents on rural roads in Tehran province, Iran

6. **Idriss Moumen, Jaafar Abouchabaka,** “[Enhanced Traffic Management Through Data Mining: Predictive Models for Congestion Reduction](#)”

- The aim of this approach is to identify patterns and characteristics, with the ultimate goal of enhancing traffic flow predictions. Experimental validation demonstrates its efficacy, thereby boosting confidence in its potential as a tool for congestion mitigation and road safety improvement.

**7. Ali Soltani , Mohsen Roohani Qadikolaei ,**”[Space-time analysis of accident frequency and the role of built environment in mitigation](#)”

Analyzing the patterns over time and variations in the frequency of accidents helped to identify areas that have improved or deteriorated in terms of road safety.

**8. Semira Mohammed,**”[GIS-based spatiotemporal analysis for road traffic crashes](#)”

- The study employed various methods, including Time-Space Cube analysis, Geographically Weighted Regression (GWR), Emerging Hot Spot analysis, and Spatial Autocorrelation analysis, with historical traffic crash data from 2015 and 2019

**9. Lu Gao and Pan Lu** “[A deep learning approach for imbalanced crash data in predicting highway-rail grade crossings accidents](#)”

- A more advanced deep learning-based model is explored as a more accurate means of predicting HRGC crashes compared to machine learning-based approaches

**10. Shakil Ahmed , Md Akbar Hossain , Sayan Kumar Ray ,**”[A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance](#)”

- Existing research has mostly studied road accident prediction as a classification problem, which aims to predict whether a traffic accident may happen in the future or not without exploring the underneath relationships between the complicated factors contributing to road accidents

**11. Xing Wang , Yikun Su , Zhizhe Zheng , Liang Xu ,**”[Prediction and interpretive of motor vehicle traffic crashes severity based on random forest optimized by meta-heuristic algorithm](#)”

- Providing accurate prediction of the severity of traffic collisions is vital to improve the efficiency of emergencies and reduce casualties, accordingly improving traffic safety and reducing traffic congestion

**12. Li-Yen Chang, Wen-Chieh Chen ,**”[Data mining of tree-based models to analyze freeway accident frequency](#)”

This study collected the 2001–2002 accident data of National Freeway 1 in Taiwan. A CART model and a negative binomial regression model were developed to establish the empirical relationship between traffic accidents and highway geometric variables, traffic characteristics, and environmental factors.