Identifying and Understanding High-Risk Accident Zones in the United States: A Data-Driven Approach

Akanksha Patil (avp5576) Gauravjit Singh Gill (gsg7817) Koyel Das (kd3075) Sanyogita Deshmukh (ssd9792) Satyajit Sriram (ss17492)

Project Github Link: https://github.com/gauravjit112/iADS-Project/tree/main

Abstract

Our study employs a data-driven approach to comprehensively analyze the persistent issue of high road accident numbers and their severity across the USA. While mishaps can manifest anywhere, there are specific regions with a persistent trend of elevated accident rates and those of higher severity (Johnson Phillips, 2021). We are utilizing a dataset of over 1.5 million accident records from 49 states from 2016. Our project research focuses on identifying factors influencing accidents. We undertake a chain of processes involving preparing the data for analysis involving data cleaning and exploratory data analysis. For further hypothesis testing and machine learning algorithms, we compute the correlations among multiple attributes and present the results in correlation matrices. Now we undertake a hypothesis to conduct testing on whether to reject or accept it. A Random Forest classifier demonstrates its efficiency in severity prediction with a 96% accuracy rate. Policymakers can benefit from the project's integrated approach, which offers a roadmap towards a future with fewer and avoidable traffic accidents, ultimately improving public safety.

Introduction

Road accidents in the United States claimed over 42000 lives last year and several over 100 thousand people have been hospitalized due to the same reason. This kind of loss of human life in the day and age of data where such numbers can be reduced with the study of data publicly available in these fields. Road safety means more than just maintaining roads. It involves identifying high-risk zones and causes of accidents and studying the causes. A few simple steps taken in the right direction can help mitigate the risks of accidents and thus reduce the number of accidents significantly.

Problem Statement

The primary concern that is being studied here is the long-standing problem of consistently high numbers and severity of accidents that occur in particular areas of the United States. Although the accident data has been well documented, the task at hand is to perform a thorough analysis to identify the exact factors contributing to these concerning statistics. The goal of this study is to not point fingers at presumptions or anecdotal evidence but towards a data-centric perspective. The objective is to get a better knowledge of the elements that directly cause accidents in these areas. This study looks for important factors and complex correlations crucial to the frequency and severity of accidents through a thorough analysis of accident data. Through the application of data, we seek to develop evidence-based tactics that can successfully lower the frequency and severity of accidents in not just the designated high-risk areas, but also reduce accident frequency all across the United States, ultimately raising road safety standards and saving millions of lives.

Research Question

Our primary research question, through this study, seeks to be:

"Which primary factors and patterns underlie the high incidence of accidents in identified risk-prone zones in the US, and how can this intel be channeled for road safety enhancement?"

Literature Review

There is a great deal of research out there that addresses accidents occurring the world over. The fact that there are still more accidents occurring despite all of this advancement in research is a major concern for everyone. But most of them focus on accident analysis, and the prediction has made use of limited resources that don't fully understand the problem and don't influence the outcomes we are looking for. In one of the research papers "A Countrywide Traffic Accident Dataset" (Moosavi, Samavatian, Nandi, Parthasarathy, &Rajiv Ramnath,2019), they tried to address this issue by collecting data from API resources available from various sources and records of 2.25 million instances of traffic accidents that took place within the contiguous The United States, and over the last three years. Every accident record includes a range of contextual and intrinsic elements, including time, place, weather, natural language description, day of the week, and points of interest (Moosavi, Samavatian, Nandi, Parthasarathy, & Rajiv Ramnath, 2019). Chang et al. (Chang, 2005) used weather data, road geometry, and annual average daily traffic to build a neural network model that predicted the frequency of accidents on a highway.

Despite all these studies, results were not available for further research. The main thing about the dataset is though it is available publicly the attributes are not enough for analysis. To address this issue, we propose in-depth data-driven research with predictive analysis and accuracy testing which shows the findings within the top 25 states and cities that are accident-prone, also analyzing day and time, street conditions, severity, and weather conditions.

The Data

This dataset spans 49 US states, including information on traffic accidents nationwide. Since February 2016, an ongoing collection of data has been conducted through a number of data providers, including several APIs that offer streaming traffic event data. The US and state departments of transportation, law enforcement organizations, traffic cameras, and traffic sensors inside the road networks are just a few of the organizations that record and broadcast traffic events. This dataset contains approximately 1.5 million accident records at the moment. For more specific details, read the descriptions below. (Moosavi, 2023). This dataset contains over 40 columns describing the nature of the accident and the circumstances under which it occurred. It includes the weather and location along with several binary indicators. We study many of these in detail further in the project.

Methodology and Steps

I. Data Cleaning

In the process of cleaning data, we start off by evaluating the relevance of the attributes. The presence of over 45 columns meant we were bound to leave out a few. We noticed quite a few columns contained data that was irrelevant to our cause of studying accident data- such as Twilight, Airport

code, etc. A few columns were also removed due to inconsistencies in the data. We went on to remove entries with NaN values in them so as to not come to mistaken conclusions.

II. Exploratory Data Analysis

To conduct a basic exploratory analysis of the data, we decided to visualize the top few of each case- such as by state, hour of the day, etc.

Exploratory Analysis was conducted on the Top City, State, Street, Hour, and Weather conditions (categorized into four main elements). In our project, which focused on analyzing road accidents in the USA, we utilized Python's Pandas library for this Exploratory Data Analysis (EDA). Our specific aim was to investigate the top 25 cities, states, and streets with the highest incidence of road accidents. Figures 1 & 2 show the study of top accident-prone cities and states. Figure 3 shows hour hour-wise analysis of accidents throughout the day. Figure 4 shows accident analysis for the top highways in the USA. Figure 5 shows the % of the severity of accidents with 2 & 3 severity- mid severity leading the charts. Figures 6 through 8 show temperature, humidity, and visibility factors contribute on accidents. To visually represent our findings, we employed Matplotlib, for creating insightful bar graphs and other visualizations. These graphical representations, which are linked at the end of this report, effectively illustrate the critical areas and factors contributing to road accidents in the USA. They not only demonstrate our analytical methodology but also vividly highlight the significant insights derived from our data exploration.

III. Hypothesis Analysis

After an extensive Exploratory Data Analysis (EDA) in our project, we seamlessly transitioned into the realm of hypothesis building and testing. EDA lays the groundwork by revealing patterns in the data, prompting a structured investigation into the observed relationships. In this pivotal phase, our focus shifts towards crafting testable hypotheses, specifically addressing the impact of weather conditions on traffic accident severity.

i. Data Loading and Cleaning

The analysis begins with the loading of the population data and relevant mappings for U.S. states. The population data is loaded into a pandas DataFrame named population_df, containing information about states and their populations in the year 2023. This data is then cleaned to include only the necessary columns, specifically the state name (state) and the population (pop2023) (Refer Figure 15). Next, a mapping DataFrame named state_mapping_df is created, associating each state's name with a corresponding StateID. This mapping is crucial for standardizing state names across datasets. Subsequently, a DataFrame named pop2023_df is created, including state names and corresponding populations for the year 2023. This DataFrame is merged with state_mapping_df to replace state names with StateIDs, enhancing consistency and facilitating further analysis (Refer Figure 16).

ii. Incorporating Weather Data

To analyze the relationship between weather conditions and accident severity, the analysis utilizes a dataset named data containing information about traffic accidents. The dataset is filtered to include only records from the year 2023, enhancing the focus on the relevant timeframe. Weather conditions are introduced by merging the filtered data DataFrame with the pop2023_df DataFrame on the 'State' column, ensuring a unified dataset. Additionally, datetime features such as the year, month, and day are extracted from the 'Start_Time' column for further temporal analysis.

iii. Correlation Analysis and Visualization

The severity column is the target variable, and 'Temperature(F),' 'Humidity (%),' 'Visibility(mi),' 'Precipitation(in),' 'Month,' and 'Day' columns are selected for analysis. For correlation analysis, we carried out Pearson, Spearman and Normal correlation to capture key factors that may influence each other within the dataset.

Pearson, Spearman and normal correlation coefficients are calculated to understand the linear and rank relationships between selected features. This information aids in identifying potential patterns and dependencies within the dataset.

For Pearson Correlation when used for linear relationships under normal distribution, assessing how weather factors are linearly related to severity we find (Refer Figure 10):

- Temperature and Humidity have a moderate negative correlation (-0.351).
- High Humidity correlates with lower Visibility (-0.510).
- Precipitation and Visibility show a moderate negative correlation (-0.308).

In Spearman Correlation when applied to non-normal data, capturing monotonic (increasing/decreasing) relationships between weather factors and severity (Refer Figure 11):

- Temperature and Humidity exhibit a moderate negative correlation (-0.351).
- A strong negative correlation between Humidity and Visibility (-0.598).
- Precipitation and Visibility show a notable negative correlation (-0.484).

Whereas Normal Correlation provides insights into linear relationships, enhancing understanding when combined with Spearman analysis (Refer Figure 12):

- Temperature and Humidity have a moderate negative correlation (-0.351).
- Humidity is strongly negatively correlated with Visibility (-0.510).
- Visibility and Precipitation show a moderate negative correlation (-0.308).

A pair plot is generated using Seaborn to visualize the relationships between selected features. This provides a graphical representation of the interactions between variables, offering insights into potential trends (Refer Figure 13):

- Severity Distribution: The histogram on the diagonal shows the frequency of different severity levels.
- Temperature vs. Severity: Scatterplot indicates a very weak negative correlation, as evidenced by the slight dispersion trend.
- Humidity vs. Visibility: Noticeable pattern suggesting a strong negative correlation; higher humidity is associated with lower visibility.

iv. Correlation P-values

Prior to hypothesis testing on the data_subset DataFrame, missing values are handled by filling them with the mean of their respective columns, ensuring a complete dataset for subsequent correlation analysis. P-values for Pearson correlation coefficients between variables in the data_subset are then computed using the stats. pearsonr function from the SciPy library. These p-values convey the likelihood of observing the correlation by chance alone, with lower values indicating stronger evidence against the null hypothesis. The resulting p_values DataFrame displays these calculated p-values (Refer Figure 17), offering additional context to identify statistically significant relationships between variables in the correlation analysis that are -

Significant Correlations: Statistically significant relationships were found where p-values were less than 0.05; for instance, between 'Severity' and 'Humidity(%)', 'Severity' and 'Precipitation(in)', and 'Visibility(mi)' and 'Day'.

Insignificant Temperature Factor: Temperature showed no statistically significant correlation with 'Severity', as indicated by a p-value greater than 0.05.

Strong Month Correlations: The correlation between 'Month' and other factors like 'Temperature(F)' and 'Humidity(%)' was statistically significant, with p-values of 0.0.

v. Hypothesis Formulation

The primary hypothesis under consideration is whether weather conditions significantly impact the severity of traffic accidents. This is formalized into a null hypothesis (H0) stating no linear relationship and an alternative hypothesis (H1) proposing a linear relationship between weather conditions and accident severity.

vi. Linear Regression Modeling

Linear regression modeling is used to quantify the impact of weather conditions on accident severity. It is trained using relevant features, and key performance metrics such as mean squared error (MSE) are calculated to assess model accuracy. The performance, as measured by the Mean Squared Error (MSE) = 0.10438187929740494, appears to be relatively low suggesting that the features did not have a linear relationship and no p-values, so this did not work out for our model.

vii. Machine Learning - Random Forest Classifier

In the context of our project, we used a Random Forest classifier to predict accident severity based on carefully selected features, including temperature, humidity, visibility, precipitation, month, day, and start hour. The model, trained on a subset of the dataset, harnesses the strength of an ensemble of decision trees, collectively contributing to a robust and accurate predictive tool. Post-training, the model's accuracy is evaluated to gauge its performance in predicting accident severity compared to actual outcomes. A high accuracy score of 96% indicates the model's success in capturing patterns and relationships within the data, establishing it as a reliable tool for predicting accident severity

viii. Feature Importance Analysis and Visualization

Beyond accuracy, understanding the contribution of each variable to the prediction is important. A feature importance analysis offers insights into which variables significantly impact accident severity predictions which showed that humidity (0.350906), temp(0.299810), and day(0.224425) were the top 3 contributing factors (Refer Figure 14). This knowledge is valuable for prioritizing interventions or focusing on specific aspects to mitigate the severity of traffic accidents. To enhance interpretability, a bar plot is employed to visually represent the importance of each feature.

ix. Ordinal Logistic Regression

Now we developed an ordinal logistic regression model to predict traffic accident severity. The dataset is split into subsets—with and without severity levels. The 'Severity' column is transformed into an ordered categorical variable, and features like temperature and humidity are defined as predictors. The model, successfully trained with a low objective function value of 0.125411, accurately predicts severity levels, achieving a high accuracy of 97.23%. This suggests that the chosen features, including temperature, humidity, visibility, and others, effectively contribute to the model's ability to discern different levels of severity in traffic incidents (Refer Figure 18).

Conclusion

Traffic accidents, a major public safety concern, have been extensively studied for analysis and prediction. Through this research, we've identified key contributing factors to accidents. The analysis establishes a meaningful connection between weather conditions and accident severity. Initial data loading and cleaning ensure data consistency, while exploratory data analysis reveals potential patterns. Regression modeling and statistical analysis validate hypothesized relationships, and the Random Forest classifier adds a predictive dimension. The feature importance analysis not only highlights significant factors but also provides actionable insights for policymakers and traffic authorities. This integrated approach, combining statistical techniques and machine learning, offers a comprehensive understanding of factors influencing accidents. The results contribute valuable information for enhancing road safety measures and optimizing resources for road accident prevention with respect to weather conditions.

Member's Contribution

Member Name	Task Assigned
Sanyogita Deshmukh (ssd9792)	Led the coding efforts by writing essential code for data analysis and model development using Python.
Koyel Das (kd3075)	Contributed to detailed report writing, assisted in refining the presentation, and brainstorming with code development.
Satyajit Sriram (ss17492)	Worked on the EDA part of the code and drafted the final report

Gauravjit Singh Gill (gsg7817)	Worked on the code for data analysis and model development using Python -Random Forest
Akanksha Patil (avp5576)	Worked on Geospatial Analysis, Prepared a Literature Review, drafted a presentation, and brainstorming for hypothesis testing,

Tables & Figures

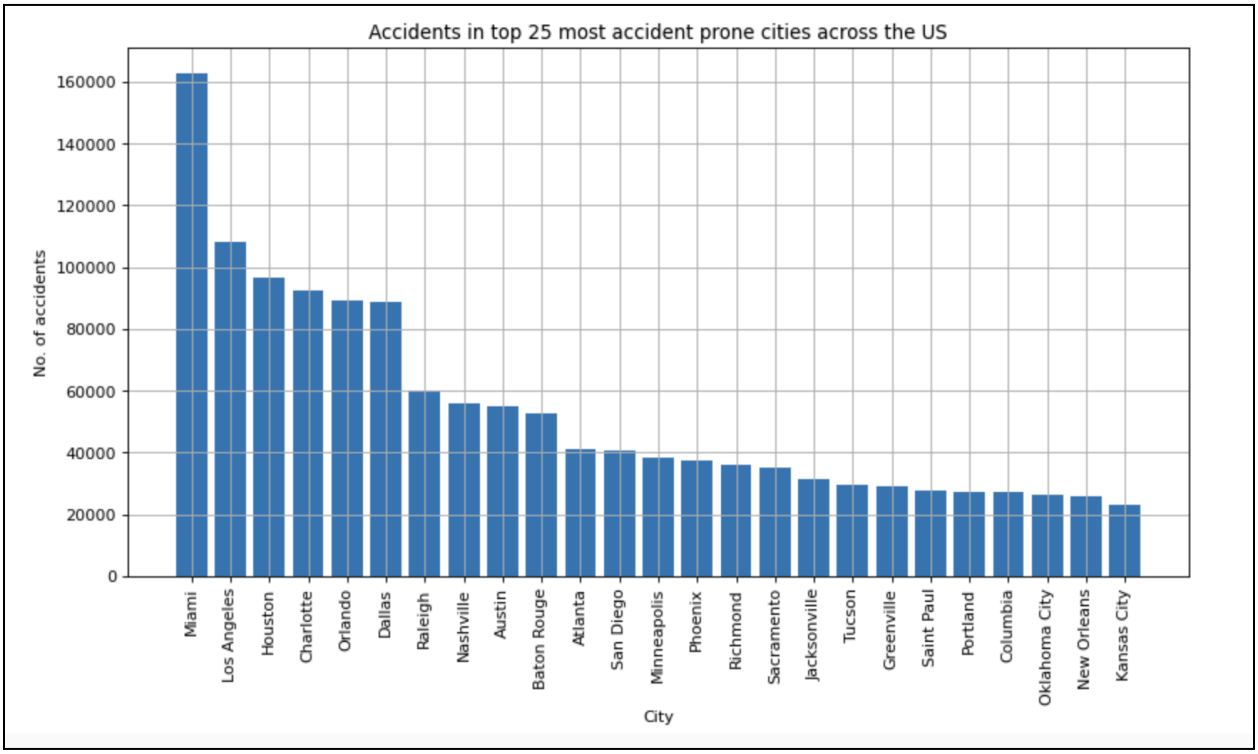

Figure 1- Top 25 Accident prone cities in the USA

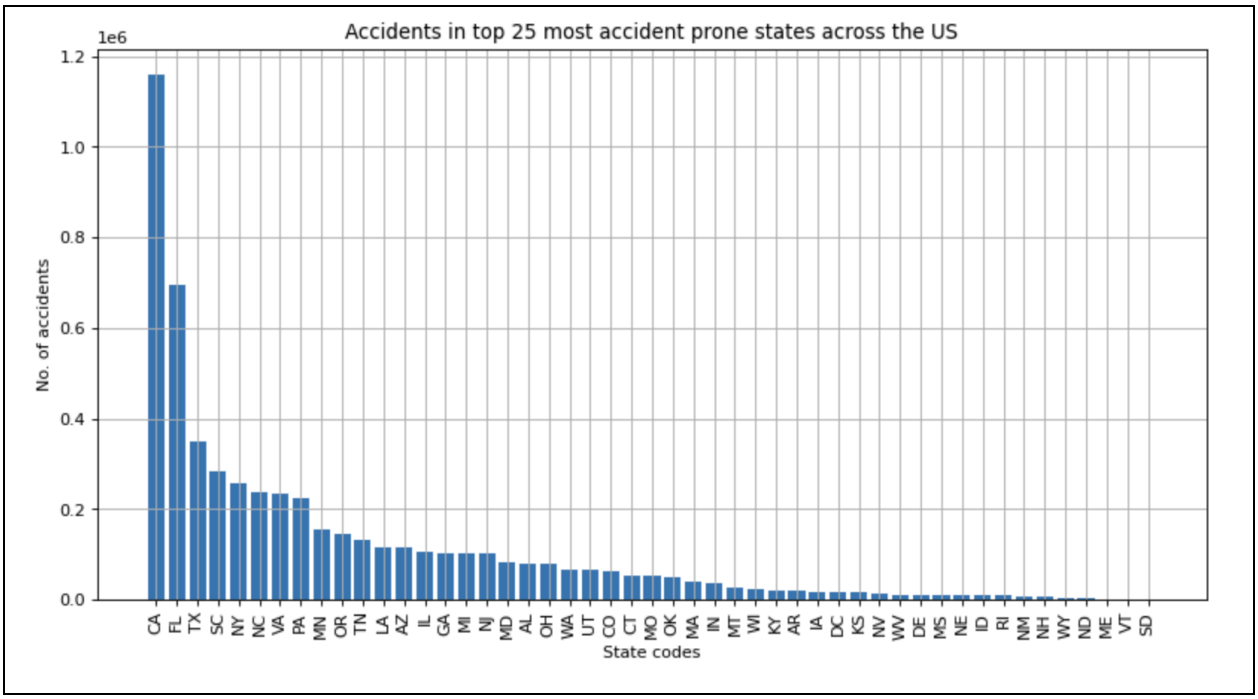

Figure 2- Top 25 Accident prone states in the USA

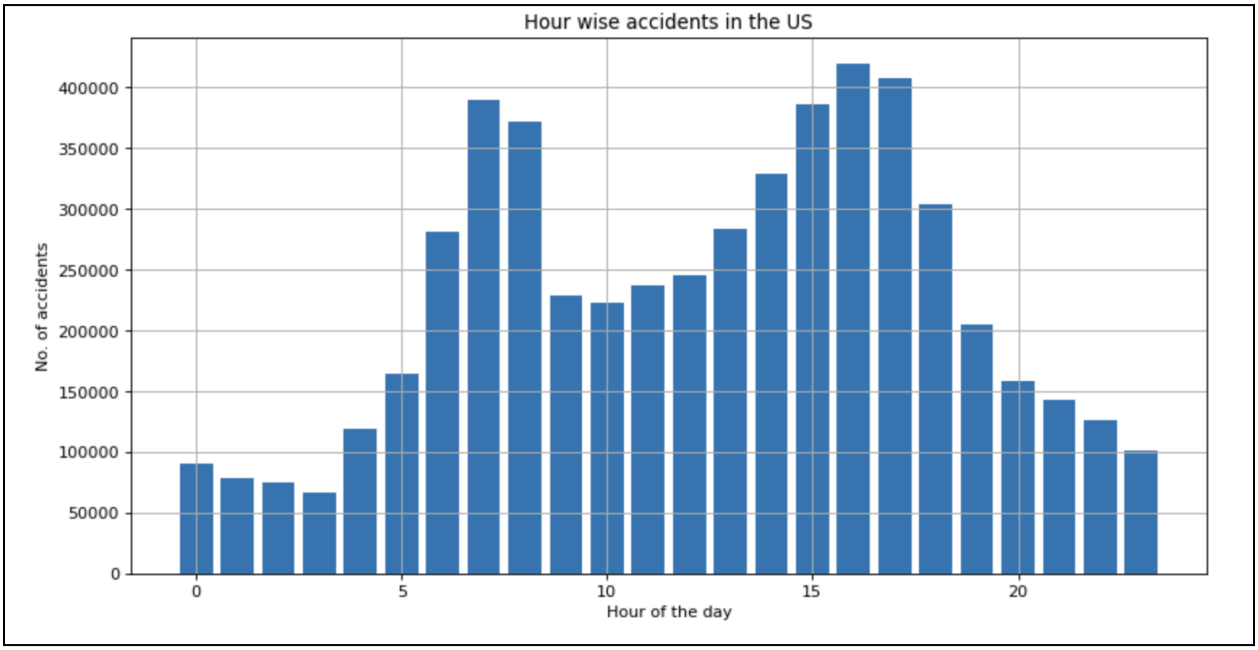

Figure 3- Hour wise accident analysis in the USA

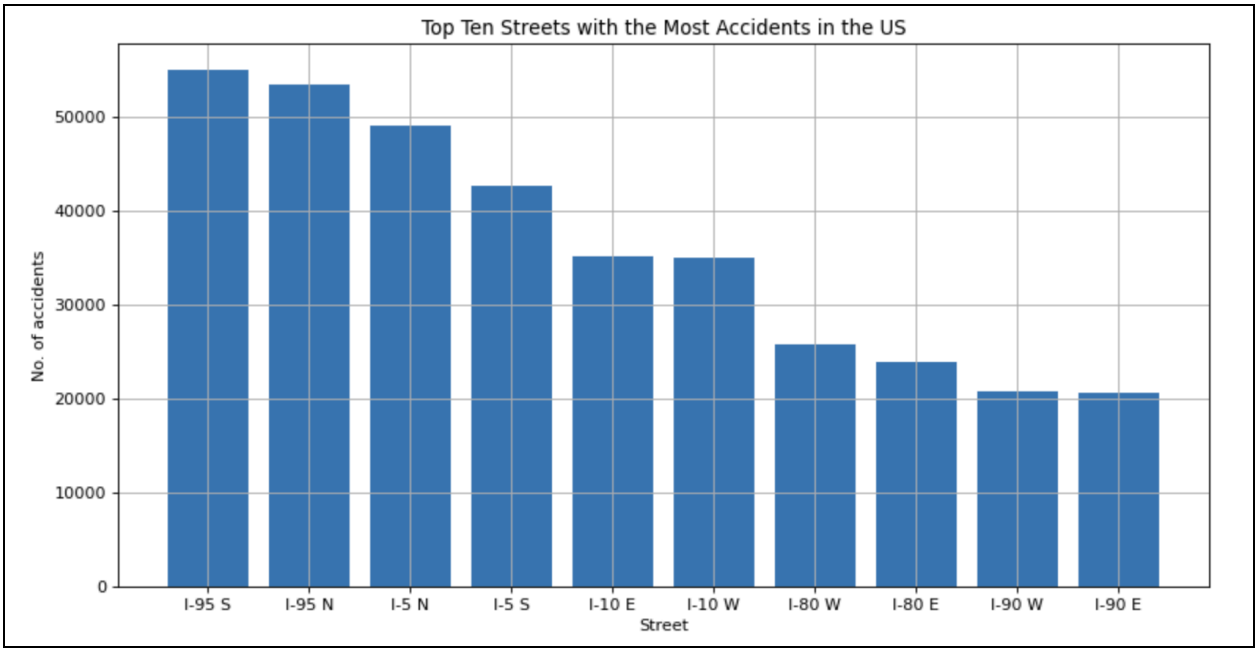

Figure 4- Top 10 Highways with most Accident cases in the USA

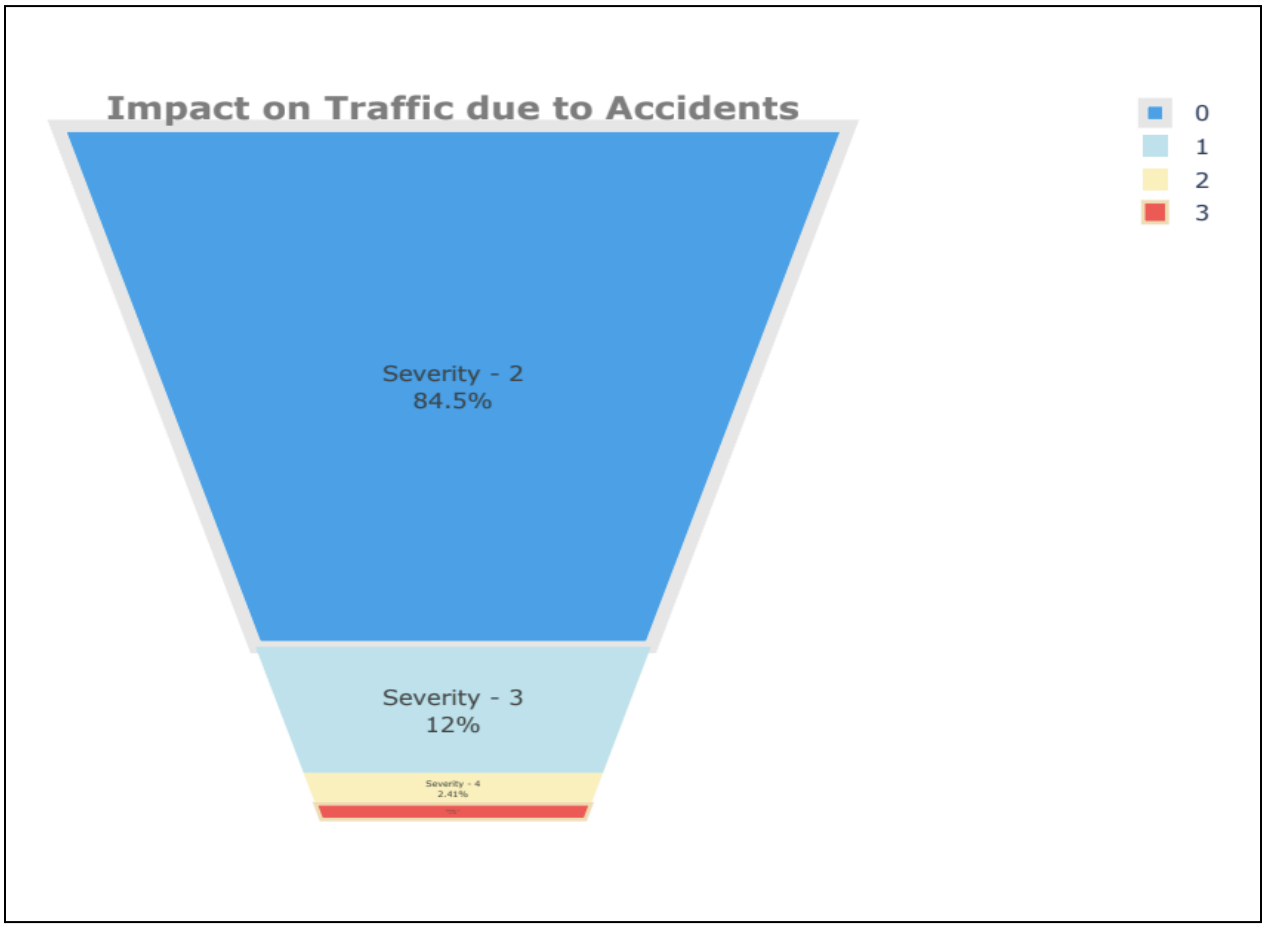

Figure 5- % of cases based on severity of accident

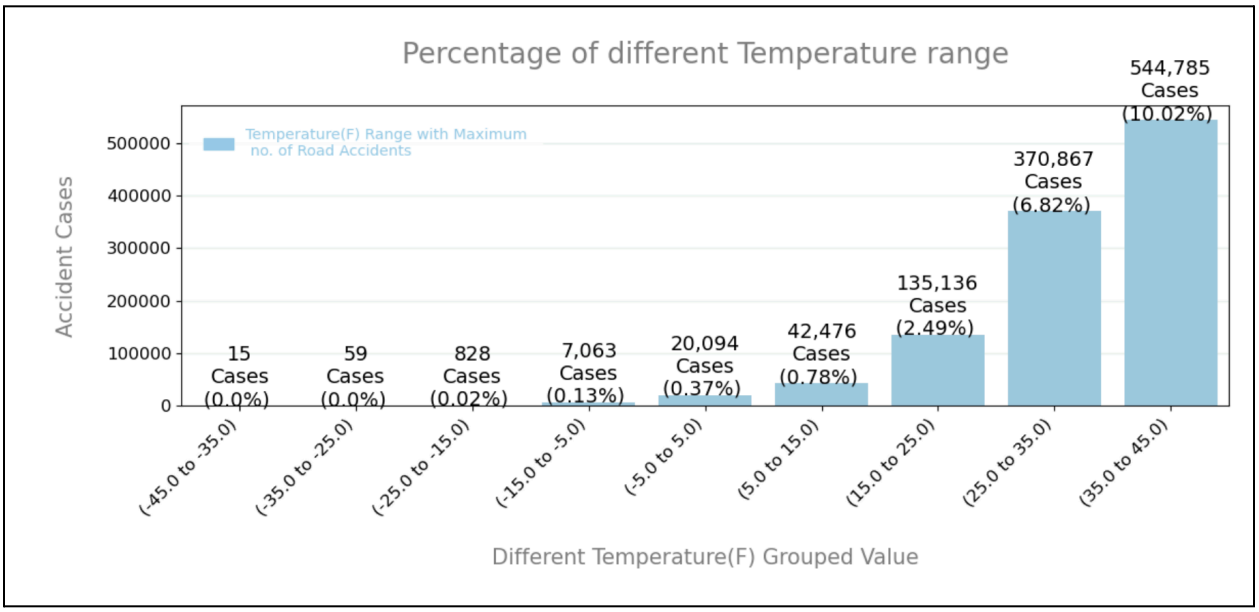

Figure 6- Temperature range relation with accident cases

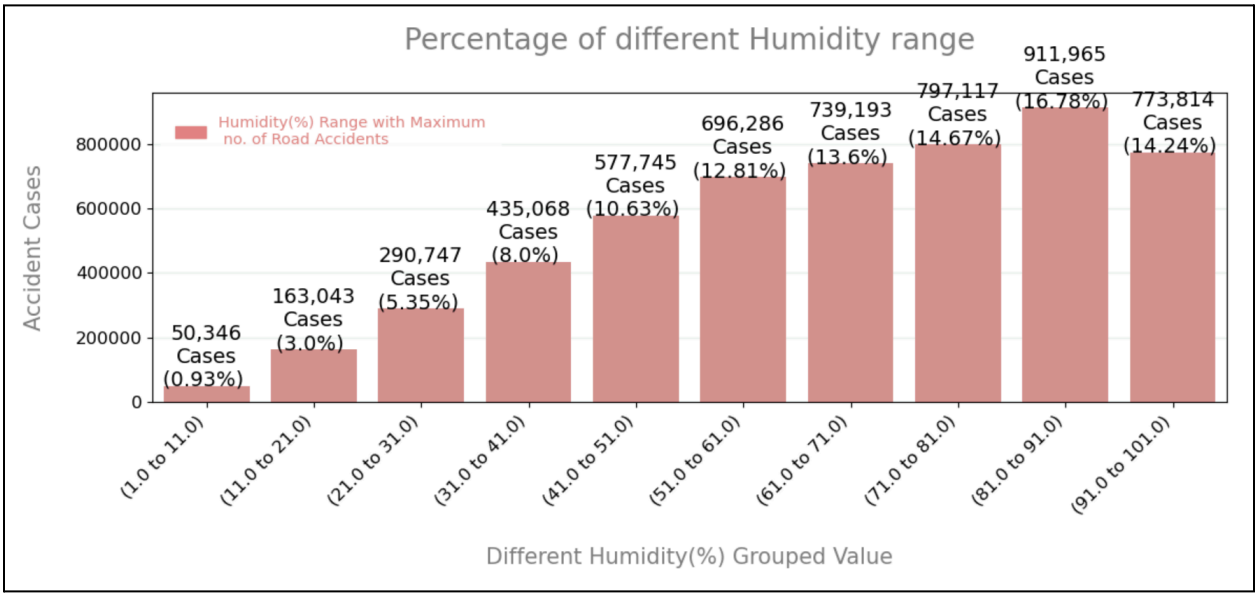

Figure 7- % humidity relation with accident cases

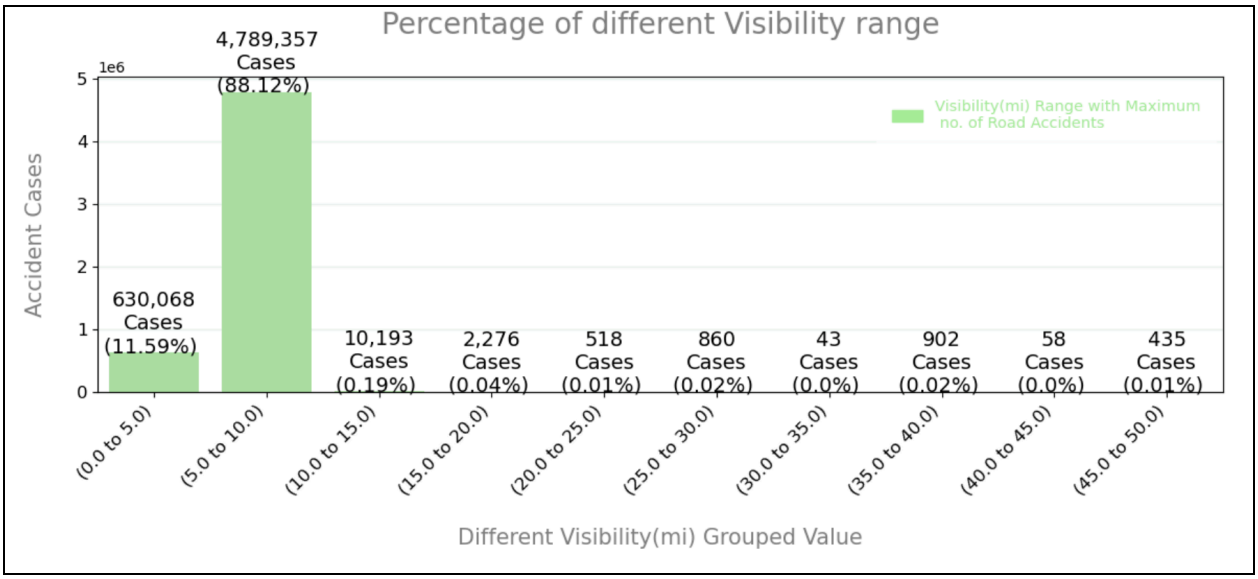

Figure 8- Visibility relation with accident cases

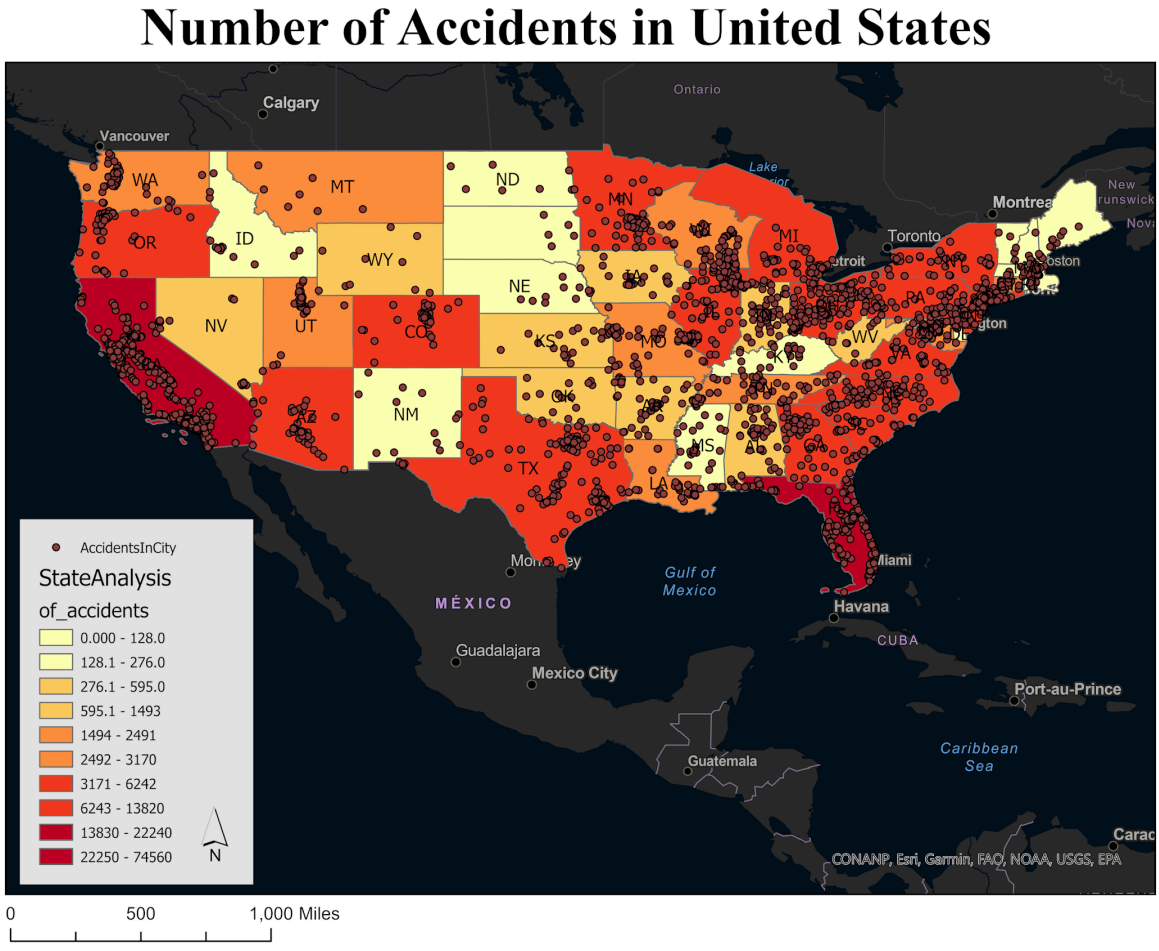

Figure 9- Geospatial plotting of accidents in the USA

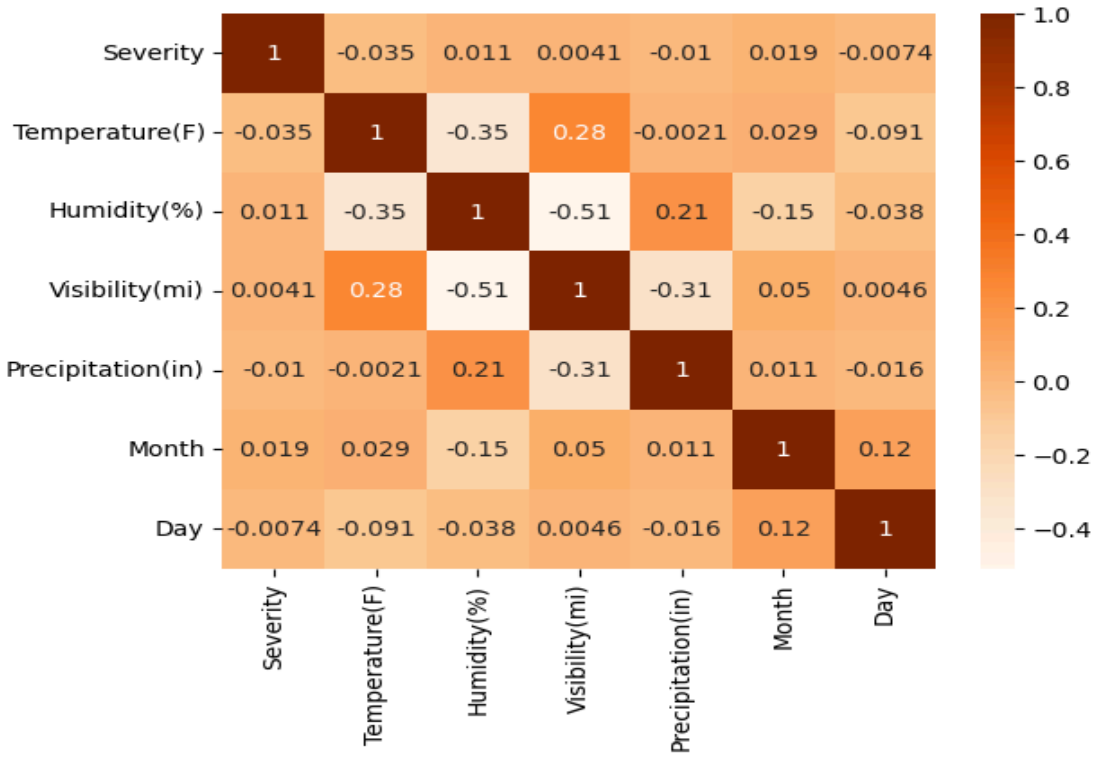

Figure 10- Correlation chart 1 (Pearson Correlation)

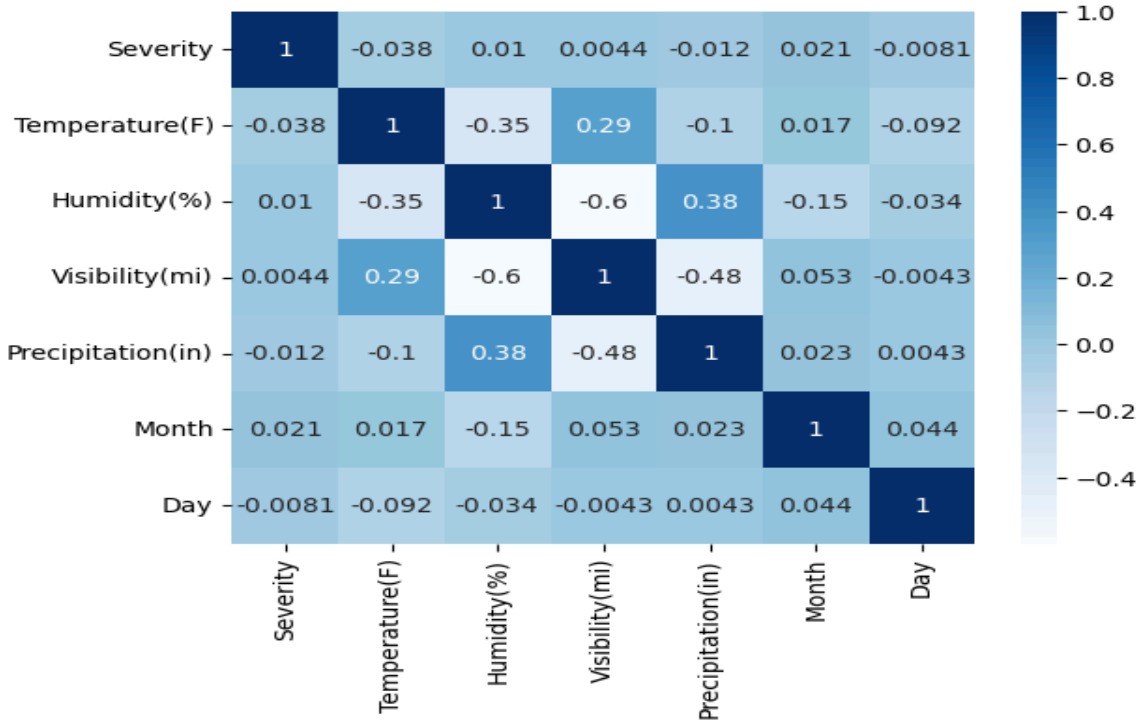

Figure 11- Correlation chart 2 (Spearman Correlation)

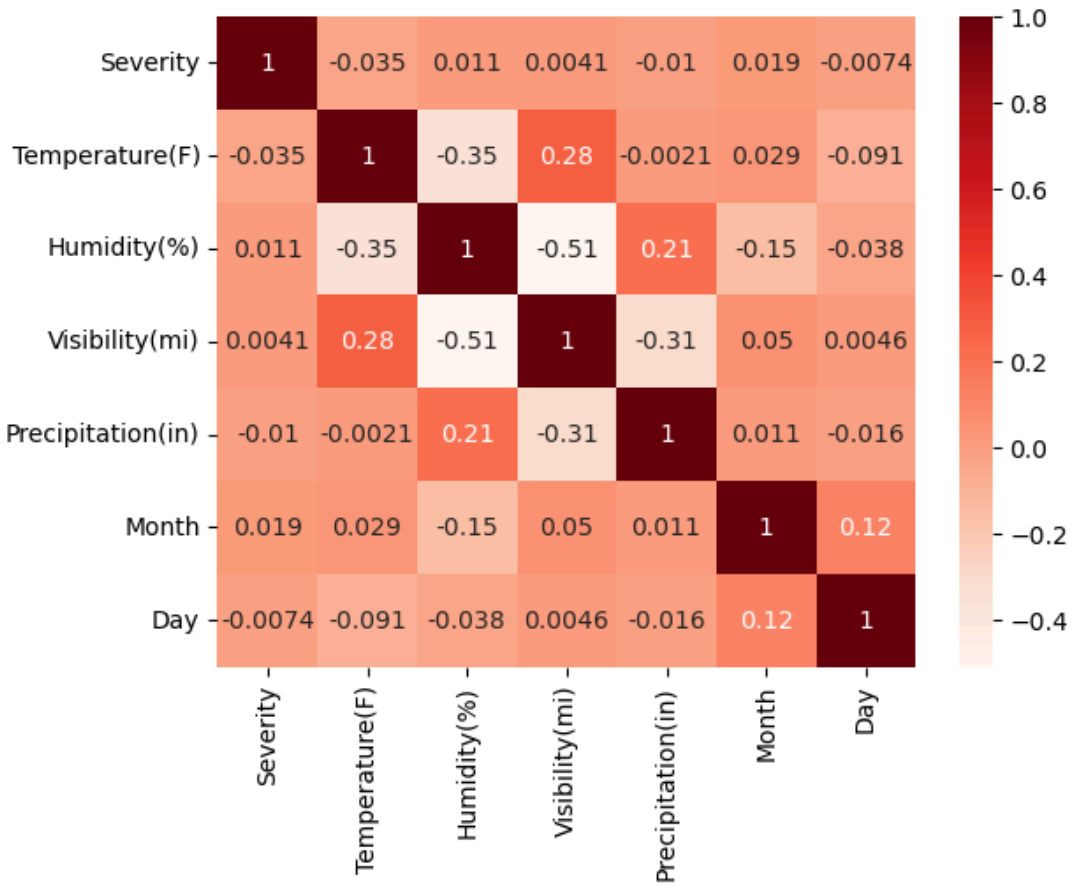

Figure 12- Correlation chart 3 (Normal Correlation)

Figure 13- Pairplot Visualization

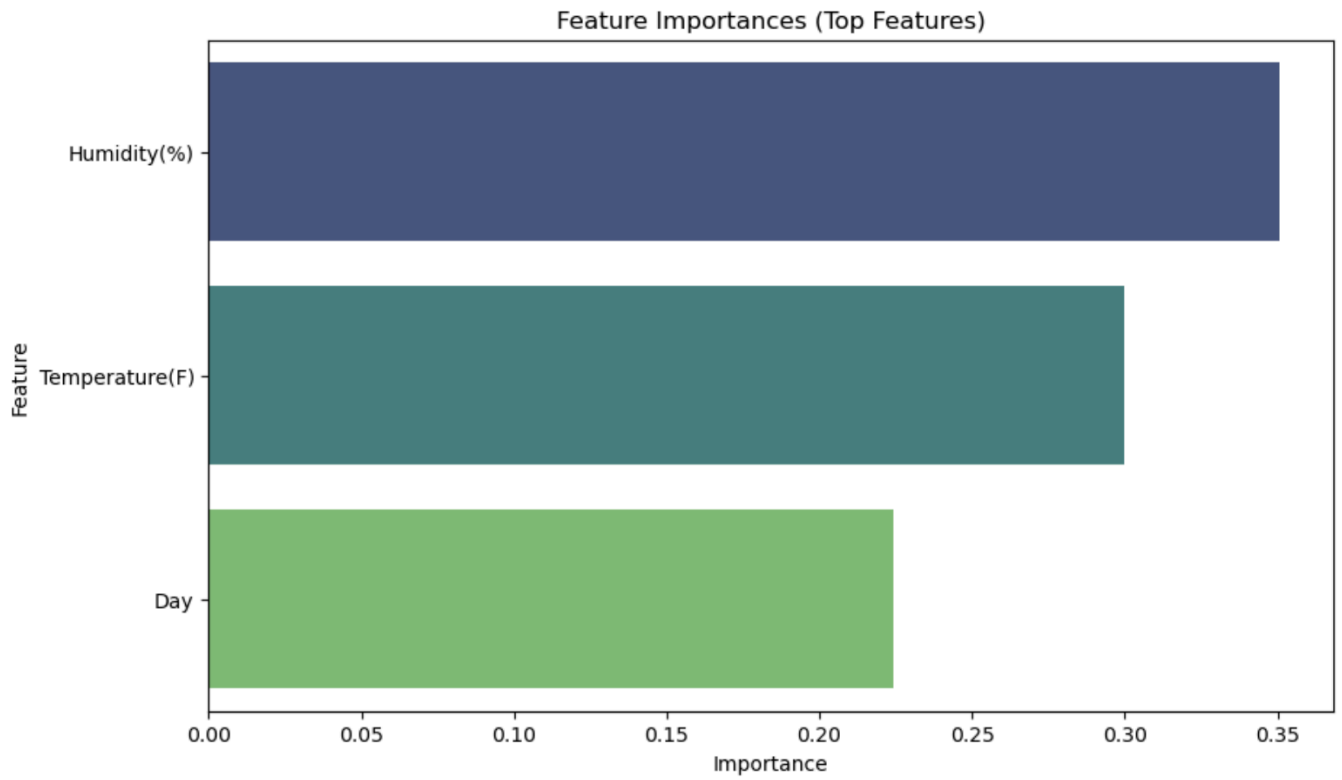

Figure 14- Feature Importance

```
In [42]: population_df_cleaned.head()
```

Out[42]:

	state	pop2023
0	California	38915693
1	Texas	30500280
2	Florida	22661577
3	New York	19496810
4	Pennsylvania	12931957

Figure 15- Data Cleaning

References

1. Moosavi, S. (2023, May 28). *US accidents (2016 - 2023)*. Kaggle. https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents
2. Centers for Disease Control and Prevention. (2023, January 10). *Global Road Safety*. Centers for Disease Controland Prevention. https://www.cdc.gov/injury/features/global-road-safety/index.html
3. Analysis of US accidents and Solutions - California State University. (n.d.-b). https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=2085&context=etd
4. United States - International Transport Forum. (n.d.-b). https://www.itf-oecd.org/sites/default/files/united-states-road-safety.pdf
5. 2023 progress report on the National Roadway Safety Strategy. (n.d.). https://www.transportation.gov/sites/dot.gov/files/2023-02/2023-Progress-Report-National-Roadway-Safety-Strategy.pdf
6. *Johnson v. Phillips*. Legal research tools from Casetext. (2021, May 20). https://casetext.com/case/johnson-v-phillips-9