

CMP4011 Big Data and Cloud Computing

Project Report

Team 10

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Sec** | **B.N** | **Code** |
| Ahmed Osama Helmy | 1 | 5 | 9213061 |
| Omar Mahmoud | 1 | 29 | 9210758 |
| Abdallah Ahmed | 1 | 25 | 9210652 |
| Aliaa Gheis | 1 | 27 | 9210694 |

Contents

[Problem Statement: 3](#_Toc196570201)

[Dataset 3](#_Toc196570202)

[Project Pipeline: 4](#_Toc196570203)

[Data Ingestion 4](#_Toc196570204)

[Data Cleaning 4](#_Toc196570205)

[Feature Engineering 5](#_Toc196570206)

[Data Visualization: 6](#_Toc196570207)

[Descriptive Analysis (Insights): 7](#_Toc196570208)

[Predictive Analysis (Insights): 12](#_Toc196570209)

## Problem Statement:

Road safety is a critical concern, and understanding accident patterns can help cities improve traffic management and reduce accident rates. This project aims to analyze accident data to identify high-risk locations, contributing factors, and potential mitigation strategies. By leveraging big data processing, we will extract valuable insights for transportation authorities and urban planners.

## Dataset

**Dataset Name:** US Accidents (2016 - 2023)

**Link:** <https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents>

**Description:**

* Contains 7.7 million accident records with 46 columns
* Main features of accident are (Severity of Accident, Time, Location, Weather, Road Characteristics)

## Project Pipeline:

### Data Ingestion

This stage involves loading the data and specifying the schema of the loaded data which is as follows:

|  |  |  |
| --- | --- | --- |
| Category | Fields | Data Types |
| Incident Info | ID, Source, Severity, Start\_Time, End\_Time, Description | String, Integer, Timestamp |
| Location | Start/End\_Lat/Lng, Distance(mi), Street, City, County, State, Zipcode, Country, Timezone, Airport\_Code | Double, String |
| Weather | Weather\_Timestamp, Temperature(F), Wind\_Chill(F), Humidity(%), Pressure(in), Visibility(mi), Wind\_Direction, Wind\_Speed(mph), Precipitation(in), Weather\_Condition | Timestamp, Double, String |
| Road Features | Amenity, Bump, Crossing, [...] (14 boolean flags) | Boolean |
| Time of Day | Sunrise\_Sunset, Civil/Nautical/Astronomical\_Twilight | String |

### Data Cleaning

#### Handling Missing Values and Nulls

* This was handled by first checking the percentage of missing values in each column and sorting them descending.
* Dropping Columns like End\_Lat & End\_Lng as the percentage of missing values was greater than 40%
* Imputing missing values in numeric columns by inserting the mean value
* Imputing missing values in categorical columns by inserting the mode value

#### Removing Columns with 1 Unique Value

* Features like Country and Turning\_Loop has only one unique value. Thus, they won’t be helpful in analysis

### Feature Engineering

#### Adding Time-Related Features

* The raw timestamp (Start\_Time) was parsed to extract granular temporal components:
  + **Hour of the Day**: Captures the time of day when accidents occur (e.g., morning rush hours, nighttime).
  + **Day of the Week**: Identifies whether accidents are more frequent on weekdays or weekends.
  + **Month**: Highlights seasonal trends in accident occurrences (e.g., higher rates during winter months due to adverse weather conditions).
  + **Year**: Tracks long-term trends in accident frequency over multiple years.
  + **Duration**: Tracks the duration of the accident in minutes by subtracting the start time from the end time
  + **Season**: Determines the Season when the accident happened (Summer, …)

#### Adding Road-Related Features

* A Boolean variable Is\_Complex\_Road was added to interpret whether the road is complex by utilizing the other variables like (Junction, Railway, Crossing)
* This will help in giving insights into the effect of complexity of roads.

#### Adding Risk Score for each State

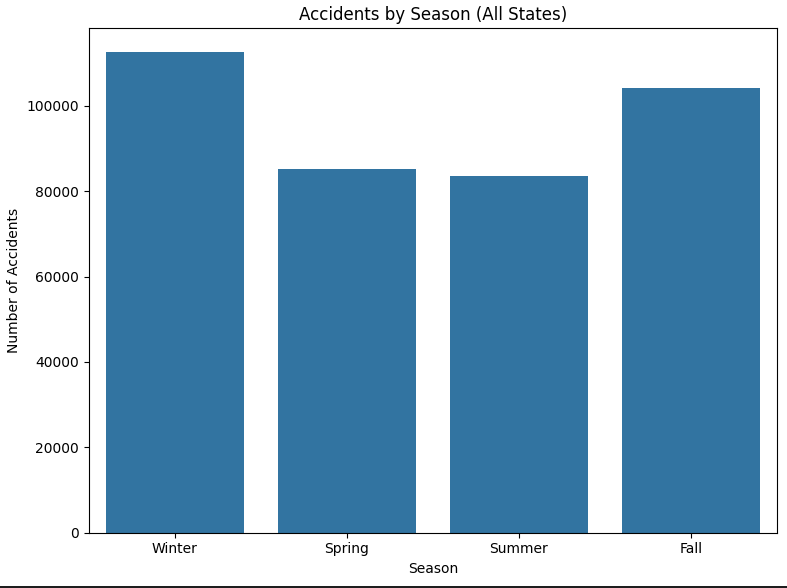
* The dataset was aggregated at the state level to simplify the analysis and provide actionable insights at a regional scale.
* Key metrics such as average accident severity, total accident count, and risk score were computed for each state.
* The **Risk score** was calculated as the product of average severity and accident count, capturing both the frequency and severity of accidents.
* The **Risk score** was then normalized to be from 0 to 1.
* Also, A Boolean variable **Is\_High\_Risk** was added to detect if a state was high risky or not by using the 75th quartile.

## Data Visualization:

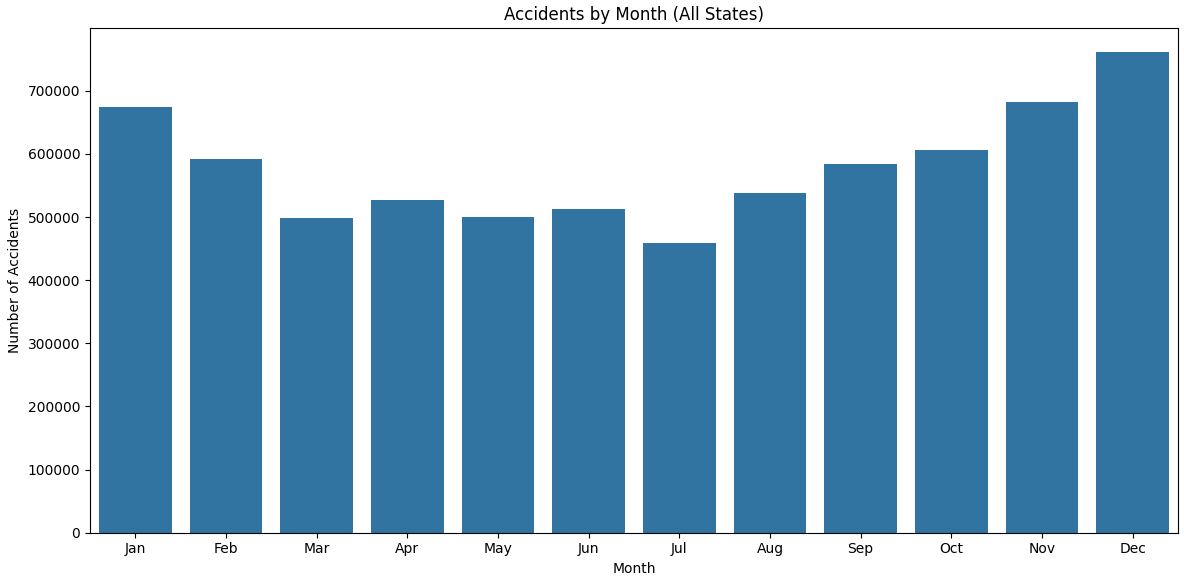
## Descriptive Analysis (Insights):

#### Accident Trends by Season.

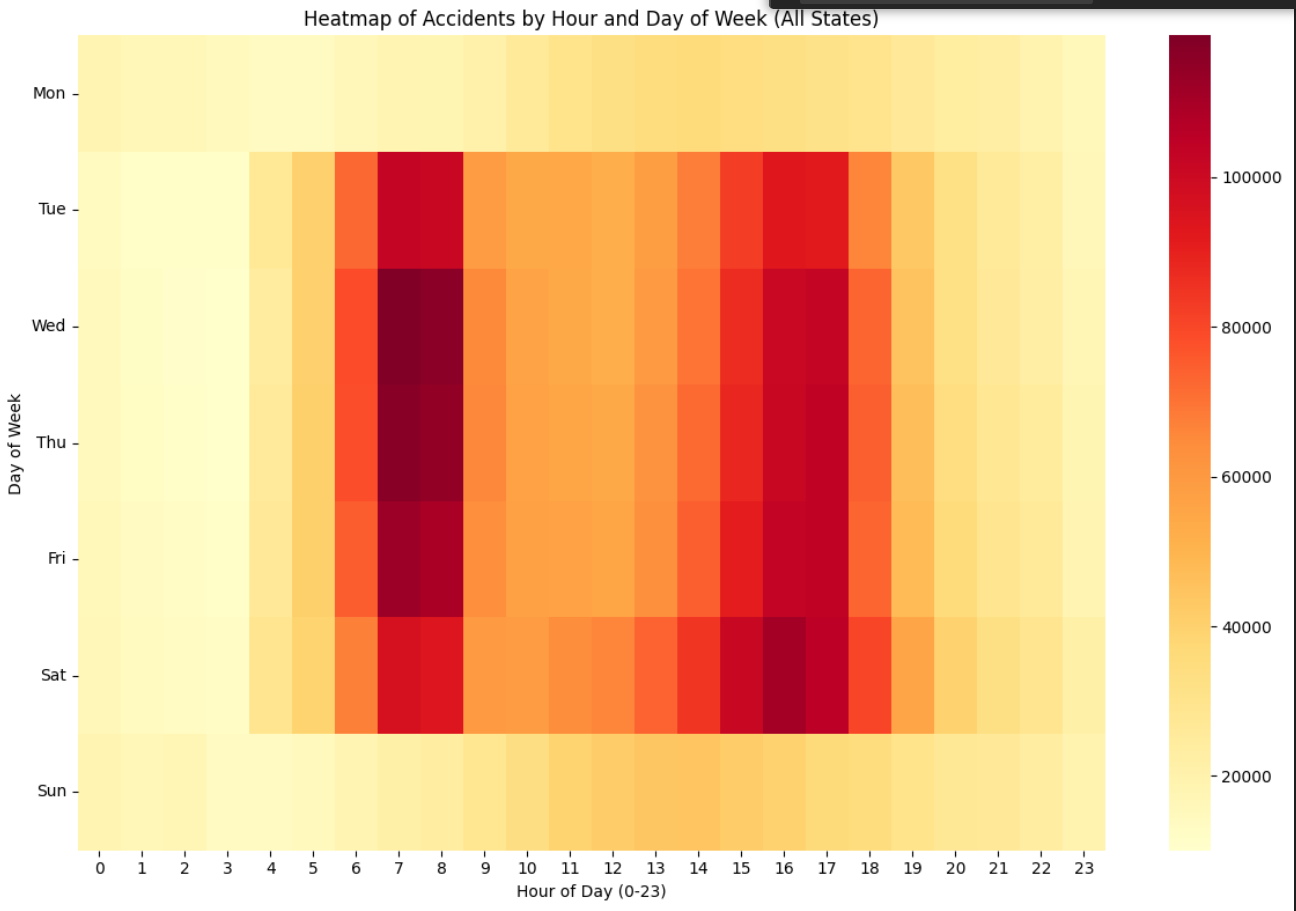
* Winter and Fall experience more accidents due to hazardous weather, such as snow, ice, and reduced daylight.



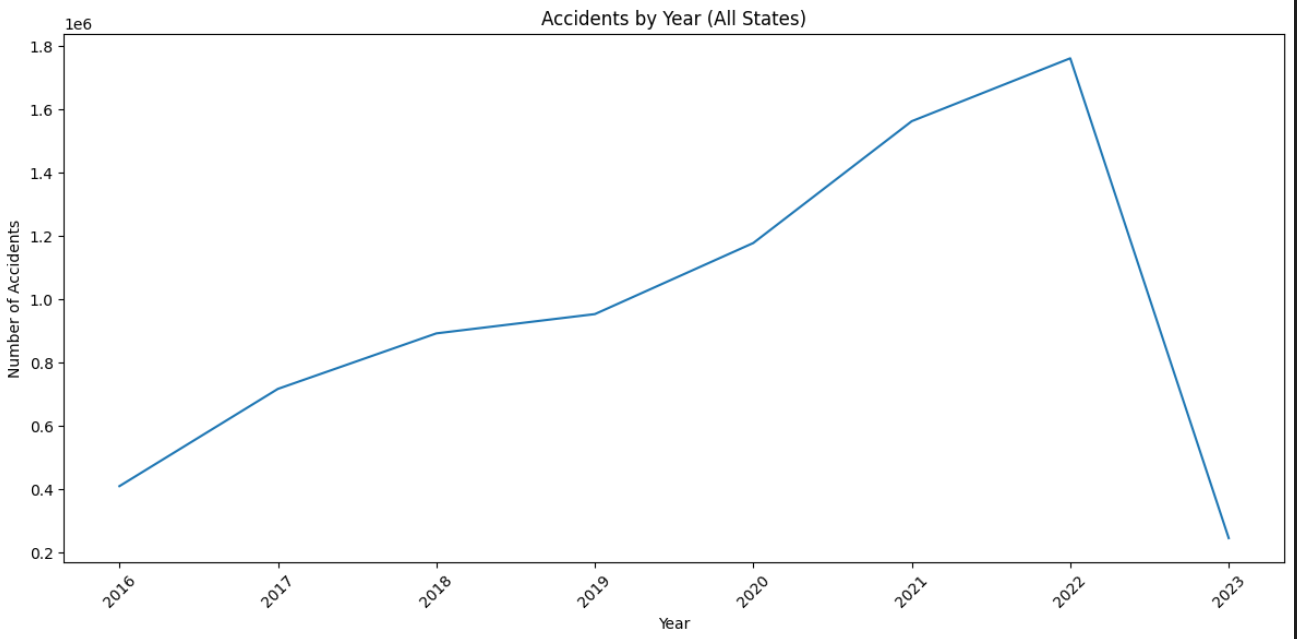
#### Accident Trends by Month.

* July seems the safest month
* Jan & Dec having high records of accidents probably due to holidays

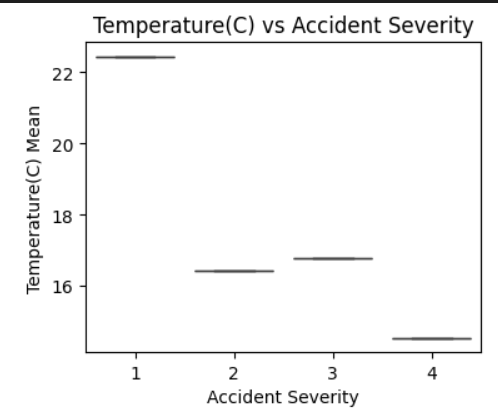
#### Accident Trends by Hour.

* Rush hours (7-9 AM and 3-6 PM) have the highest accident frequency due to increased traffic.

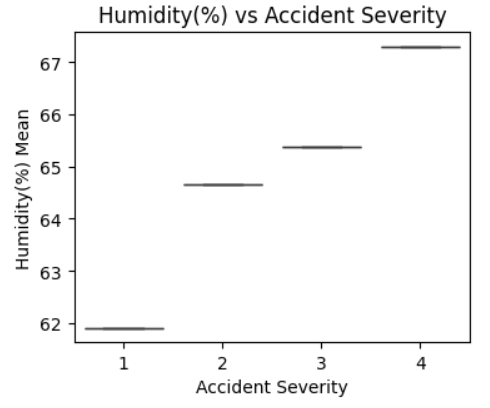
#### Accident Trends by Year.

* Accident counts steadily grew from 2016 to 2022, nearly quadrupling in this period, indicating an increase in accident frequency.
* The sharp drop in 2023 accidents is likely due to incomplete data (up to March), not a true decline.

#### Relation between Temperature and Accident Severity.

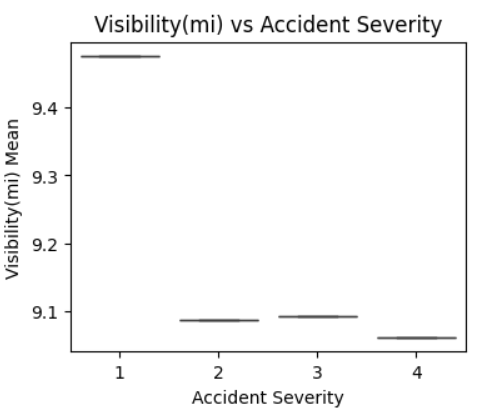
* The data suggests that accidents occurring at lower temperatures tend to be associated with greater severity.
* Probably, Icy conditions or other cold-weather hazards could play a significant role in the seriousness of these incidents

#### Relation between Humidity and Accident Severity.



* Accidents happening with higher humidity levels appear to correlate with increased severity.
* Perhaps rain or other moisture-related factors contribute to more impactful collisions.

#### Relation between Visibility and Accident Severity.

* Lower visibility is strongly linked to more severe accidents.
* This underscores the critical impact of clear sight on road safety and the potential dangers of driving in compromised visual conditions.

#### Top Cities contributing with Accidents.

* A graph of blue rectangular bars

  AI-generated content may be incorrect.Miami, Houston & Los Angeles are the top cities contributing to accidents

#### Top States contributing with Accidents.

* California, Florida & Texas are leading the Accidents contribution

A graph of a number of states with the highest accident counts

AI-generated content may be incorrect.

## Predictive Analysis (Insights):

#### Predicting Accident Severity using Random Forest.

* The **Random Forest** model achieved an 81% accuracy in predicting the severity of accidents.
* **Weather** conditions (rain, snow, fog, etc.) have the greatest impact on accident severity, with Importance 59.60% highlighting the need for businesses to consider weather when making operational decisions.
* Actionable Insight: Adjust transportation operations based on weather forecasts to mitigate risk, particularly during adverse conditio

A graph of blue rectangular bars

AI-generated content may be incorrect.

Random Forest : 0.81

A screenshot of a computer

AI-generated content may be incorrect. A screen shot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.