Predicting the circumstances that cause accidents in the United States

Asritha Baddam, Akhila Mitta, AdiLakshmi Meesala, Pitchaiah Lella

Northwest Missouri State University, Maryville MO 64468, USA
Section Number: 44517-02, Team Name: Impact Players

1 Introduction

We have used a nationwide dataset of car accidents that includes data from 49 US states. Using several APIs that offer streaming traffic incident (or event) data, the accident data were gathered from February 2016 to March 2023. These APIs provide traffic data that has been collected by a few organizations, such as state and federal transportation departments, law enforcement organizations, traffic cameras, and traffic sensors installed in road networks. There are currently about 279k accident records in the dataset.

Our goal is to evaluate which factors have the greatest influence on an accident; that is, we aim to figure out the significance of each feature. These may include the driver's location, the time of day, the season, the weather, and nearby attractions. Certain locations are more vulnerable than others.

2 Tools and Technologies

- **Kaggle:** Kaggle is an online platform and community designed for data scientists, machine learning practitioners, researchers, and anyone interested in working with data.

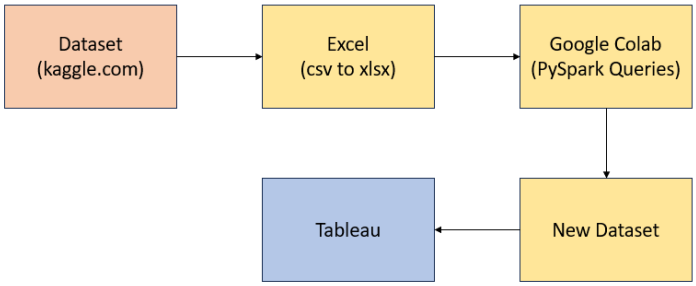
- **Google Colab:** A free cloud-based platform called Google Colab, also known as Google Colaboratory, gives users access to an integrated development environment (IDE) for writing, running, and sharing Python code. It is especially popular among data scientists, machine learning engineers, researchers, and students due to its easy access to computing resources and various libraries, as well as its collaborative features.

- **PySpark:** PySpark is a Python library and part of the Apache Spark ecosystem, designed to enable Python programmers to work with Spark, an open-source distributed computing system. Spark is renowned for its speed and efficiency in processing large-scale data across clusters.

- **Tableau:** Tableau is a powerful and widely used data visualization tool that helps users create interactive and visually appealing representations of data. It allows individuals and organizations to turn raw data into understandable and actionable insights. Tableau's user-friendly interface and robust features make it popular among analysts, businesses, and professionals working with data.

- **Git:** Git is a distributed version control system that facilitates collaboration among developers working on a codebase. Git tracks changes made to files over time, allowing users to revert to previous versions, compare changes, and manage different versions of the codebase.

3 Architecture



4 Architecture Overview

- The first step of the project is to obtain data from Kaggle, which is a csv file. So, initially, we downloaded the US Accidents March23 data set from Kaggle.com.

- Then the data is stored in an excel file. The primary source of data for our project is this Excel file. An extensive analysis involves thorough exploration, interpretation, and presentation of the dataset's various aspects to derive meaningful insights and aid decision-making processes.

- Google Colab is an indispensable tool for data analysis and coding because it allows code execution and collaboration right inside a web browser.

- PySpark's strong capabilities allow for the project's efficient processing and analysis of large datasets of US accidents, allowing for a thorough evaluation of the US accident rate. Here, we write SQL queries to accomplish every goal and reach every objective. Then, we use Tableau to visualize our results.

- Tableau is used to explore the dataset and create visual representations to enhance data comprehension. Through interaction with visualizations, stakeholders can obtain insights more effectively with this approach.

5 Goals

1. **Which states are most prone to accidents?**
   This goal aims to identify the states with the highest number of accidents. By grouping accidents based on their respective states and counting occurrences within each state, it provides an understanding of which states experience more frequent accidents, aiding in resource allocation and targeted safety measures.From the chart it is clear that California has highest number of accidents i.e, 71,784.

   Code snippet

Listing 1.1. code snippet

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate

from datetime import datetime, date
import pandas as pd
from pyspark.sql import Row
df = spark.createDataFrame([
    Row(a=1, b=2., c='string1', d=date(2000, 1, 1),
        e=datetime(2000, 1, 1, 12, 0)),
    Row(a=2, b=3., c='string2', d=date(2000, 2, 1),
        e=datetime(2000, 1, 2, 12, 0)),
    Row(a=4, b=5., c='string3', d=date(2000, 3, 1),
        e=datetime(2000, 1, 3, 12, 0))
])
df


AccidentsData = spark.read.format('csv').option('
    header', 'true').load('USAccidents.csv')
AccidentsData.createOrReplaceTempView('
    AccidentsDataDetails')
AccidentsData.show()
```

```
18
19 spark.sql("SELECT
20 State, COUNT(*) AS AccidentCount
21 FROM AccidentsDataDetails
22 GROUP BY State
23 ORDER BY AccidentCount DESC").show()
```

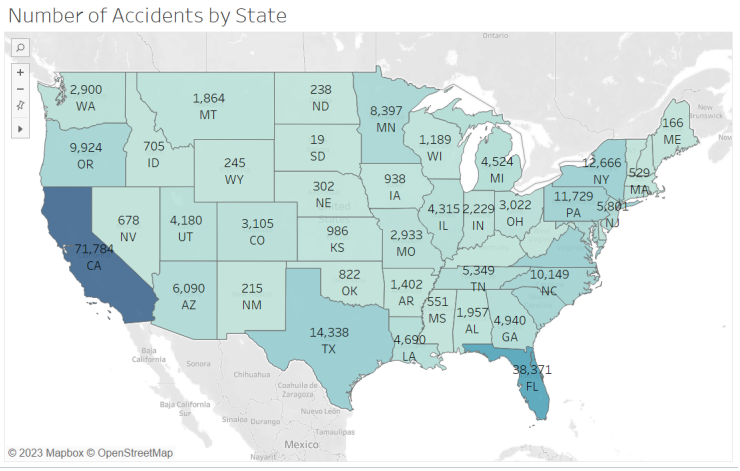


Fig. 1. Number of Accidents by State

2. **What period of the day does the likelihood of accidents occur?**
   This goal categorizes accidents based on the time of day they occur. It segments the day into distinct periods - morning, afternoon, evening, and night - to understand during which timeframe accidents are most likely to happen. This insight assists in focusing attention on specific periods for potential interventions or safety measures. From the chart it is clear that large number of accidents occur during Nautical Twilight condition of a day with an average severity of 2.59.

Code snippet

Listing 1.2. code snippet

```
1 spark.sql('Select
2 Sunrise_Sunset, Civil_Twilight,Nautical_Twilight,
     Astronomical_Twilight, AVG(Severity)
3 FROM AccidentsDataDetails
```

```
4 GROUP BY Sunrise_Sunset, Civil_Twilight,
    Nautical_Twilight, Astronomical_Twilight,
    Severity').show()
```

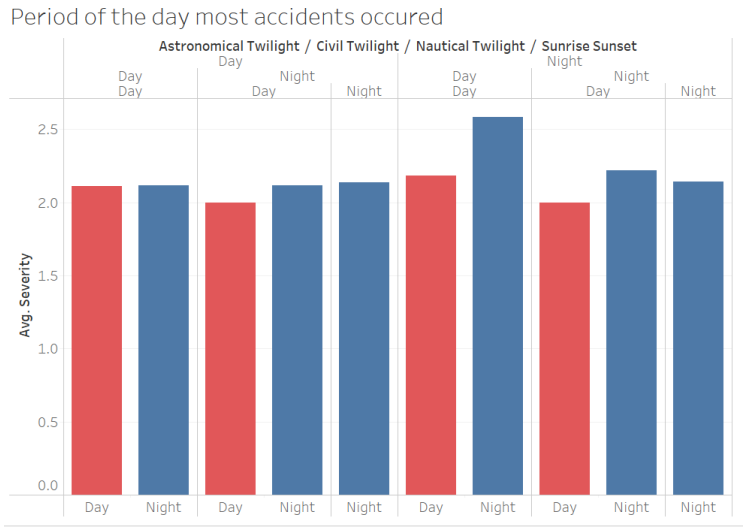


Fig. 2. Period of the day most accidents occured

3. **How many Accidents occurred based on Road Features (Amenity, Crossing, Junction, etc.)**
   By aggregating the counts of accidents related to various road features such as amenities, crossings, junctions, etc., this goal helps in understanding the impact of infrastructure on accident occurrences. It provides an overview of accidents associated with specific road attributes, informing infrastructure planning and safety improvements. From the chart it is clear that most of the accidents occur at Junctions i.e, 25,531.

Code snippet

Listing 1.3. code snippet

```
1 spark.sql(' select
2 count(case when Amenity = "TRUE" then 1 end) as
    Amenity_Count,
3 count(case when Bump = "TRUE" then 1 end) as
    Bump_Count,
4 count(case when Crossing = "TRUE" then 1 end) as
    Crossing_Count,
```

```
count(case when Junction = "TRUE" then 1 end) as
    Junction_Count,
count(case when No_Exit = "TRUE" then 1 end) as
    No_Exit_Count,
count(case when Railway = "TRUE" then 1 end) as
    Railway_Count,
count(case when Roundabout = "TRUE" then 1 end) as
    Roundabout_Count,
count(case when Station = "TRUE" then 1 end) as
    Station_Count,
count(case when Stop = "TRUE" then 1 end) as
    Stop_Count,
count(case when Traffic_Calming = "TRUE" then 1 end)
     as Traffic_Calming_Count,
count(case when Traffic_Signal = "TRUE" then 1 end)
    as Traffic_Signal_Count,
count(case when Turning_Loop = "TRUE" then 1 end) as
     Turning_Loop_Count
from AccidentsDataDetails').show()
```

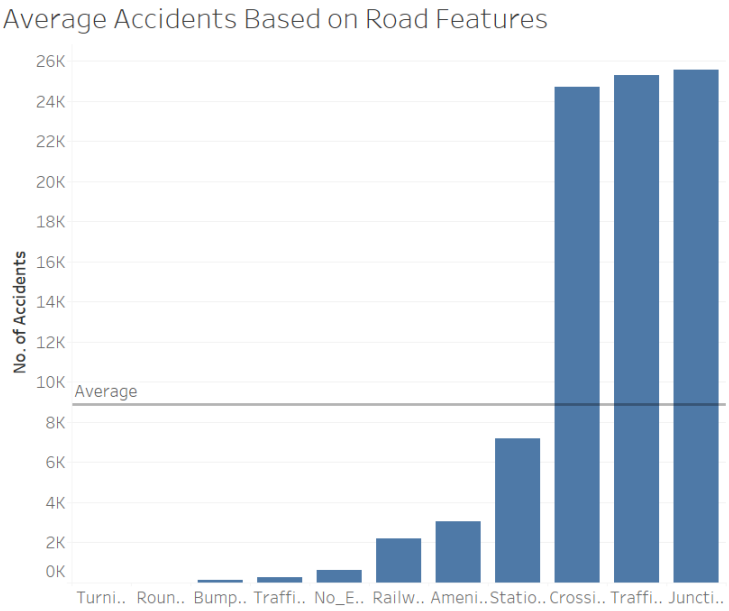


Fig. 3. Average Accidents Based on Road Features

4. **Which weather patterns are most typical for accidents?**
   This goal categorizes accidents based on different weather conditions to determine which weather patterns are commonly associated with accidents. Understanding these patterns helps in adapting driving practices and implementing safety measures tailored to specific weather conditions to reduce accident rates. The accidents mostly occur during Fair conditions i.e, 116,888.

Code snippet

Listing 1.4. code snippet

```
spark.sql("SELECT
Weather_Condition, COUNT(*) AS AccidentCount
FROM AccidentsDataDetails
GROUP BY Weather_Condition
ORDER BY AccidentCount DESC").show()
```

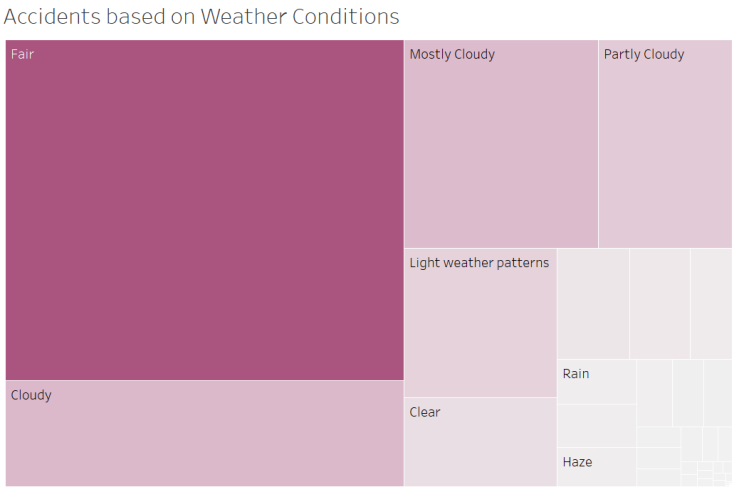


Fig. 4. Accidents Based on Weather Conditions

5. **Which time of the day(Sunrise/Sunset) does most of the accidents occur?**
   This goal categorizes accidents based on sunrise or sunset to understand when most accidents occur in relation to these specific times of the day. It aids in comprehending accident trends concerning changes in natural light conditions. From the chart it is clear that most accidents occur during day time i.e, 182,222.

Code snippet

Listing 1.5. code snippet

```
spark.sql("SELECT Sunrise_Sunset,COUNT(*) AS
    AccidentCount
FROM AccidentsDataDetails
GROUP BY Sunrise_Sunset
ORDER BY AccidentCount DESC").show()
```

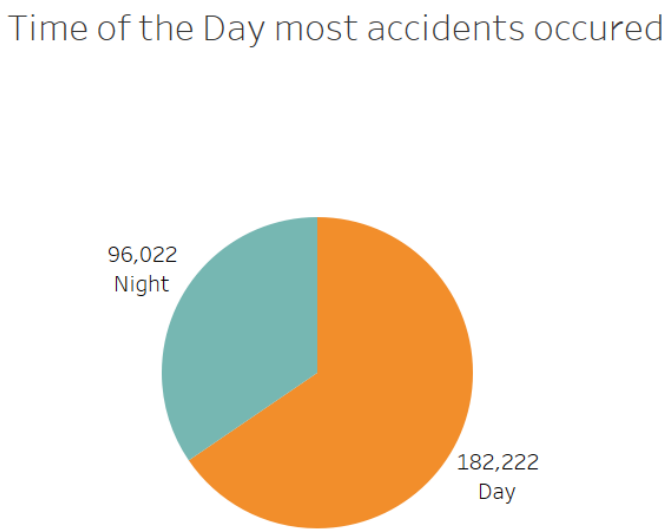


Fig. 5. Time of the day(Day/Night) most accidents occured

6. **How many Accidents occur based on the Severity?**
   Categorizing accidents by severity levels (e.g., low, medium, high) provides insights into the distribution and frequency of accidents concerning their severity. This information is crucial for understanding the impact of accidents and guiding policy decisions related to safety measures. From the below chart it is clear that nearly 250k accidents occur for severity 2.

Code snippet

Listing 1.6. code snippet

```
spark.sql("SELECT Severity, COUNT(*) AS
    AccidentCount
FROM AccidentsDataDetails
GROUP BY Severity
ORDER BY AccidentCount DESC").show()
```

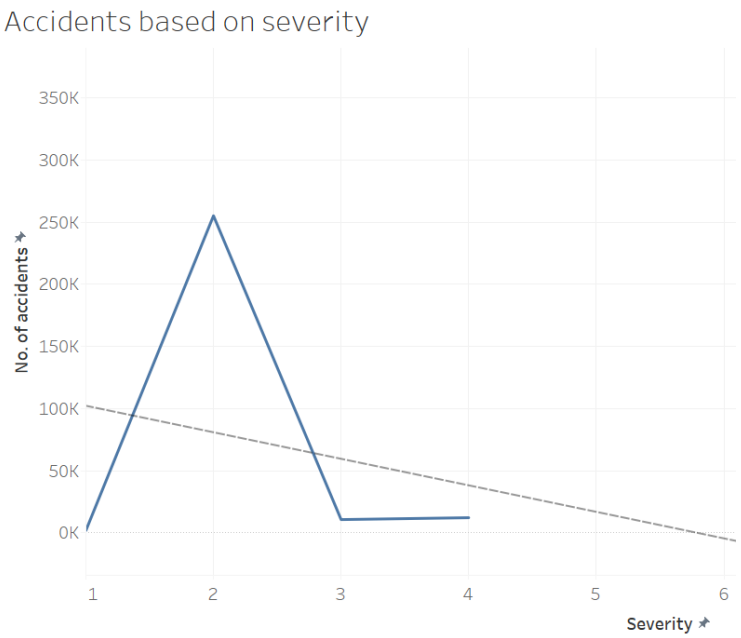


Fig. 6. Accidents Based on Severity

7. **Analysis based on distance accidents occured?**
   This goal information provides a quick overview of the distribution of accidents based on distance. This analysis could be valuable for decision-makers, traffic planners, or safety officers to understand the frequency of accidents at different distance intervals.

Code snippet

Listing 1.7. code snippet

```
spark.sql('select
count(case when distance < 1 then 1 end) as short,
count(case when distance between 1 and 5 then 1 end)
    as mid, count(case when distance > 5 then 1 end)
    as Long
from AccidentsDataDetails').show()
```

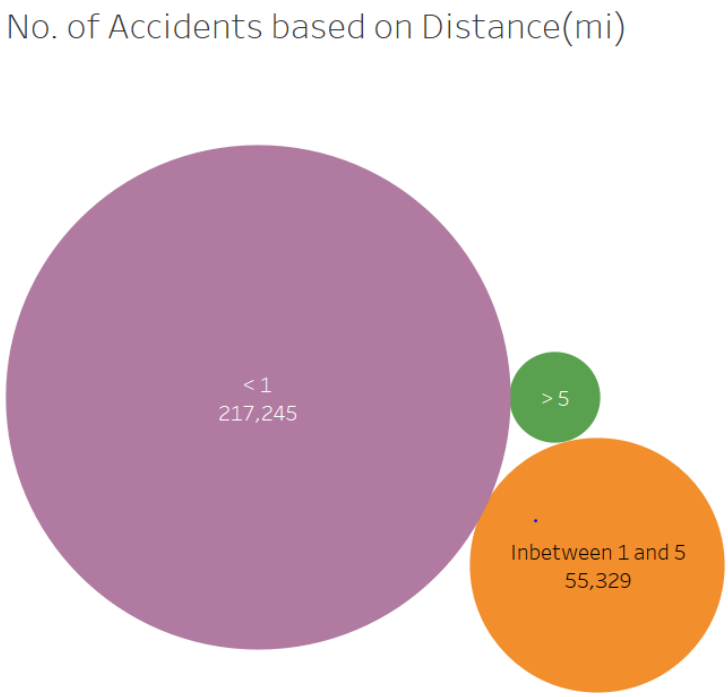


Fig. 7. Number of accidents based on Distance(mi)

8. **Which counties have the greater number of accidents based on Severity?**
   By categorizing accidents based on both county and severity levels, this goal provides a detailed breakdown of accidents within each county concerning severity. It helps in identifying areas that might require more targeted safety interventions based on severity levels in specific counties. From the chart it is clear that most accidents occur at a distance of less than 2 miles i.e, 217,245.

Code snippet

Listing 1.8. code snippet

```
spark.sql("SELECT
County, Severity, COUNT(*) AS AccidentCount
FROM AccidentsDataDetails
GROUP BY County, Severity
ORDER BY County, Severity").show()
```

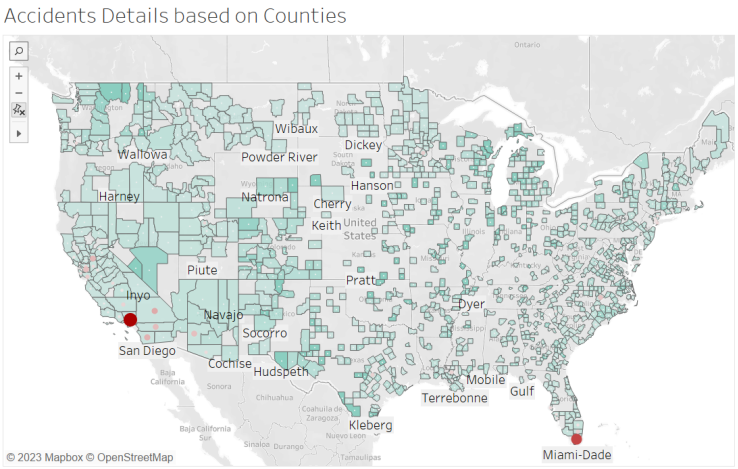


Fig. 8. Accidents based on Counties

6 The 5Vs of accidents in the United States

Volume:

The term "Big Data" refers to a massive amount of data. Volume is one of the factors that must be addressed while dealing with Big Data solutions. We took into account the most recent US Accidents Dec21 data. There are 42,769 rows and 47 columns in our data collection. As a result, our data is content with the volume.

Velocity:

From February 2016 to December 2021, this dataset was produced in real-time utilising numerous real-time traffic incident (or event) data APIs. This collection now has over 2.8 million accident recordings, which meets the velocity requirement.

Veracity:

It contains a large number of null values and doubtful values that can be cleaned. Here's how it works: In a recent study, Verisk Protection Arrangements discovered that the frequency of protection claims for engine vehicle disasters in the United States in 2019 was 4.4 percent higher than in 2018. The average cost of such claims was 10.6 percent greater. The increase in recurrence is the highest since 2014. According to the study, the average fetched of a real harm obligation claim in 2019 was 26.2 percent more than in 2018. The average cost of a property damage risk claim was 14.0 percent higher. A collision claim would normally garner 13.0 percent more. The average fetched for a personal injury security claim was 9.7 percent greater. The average cost of a medical installment claim was 8.8 percent greater. Montana saw the highest increase in recurrence, at 16.4 percent. The average cost of a claim also increased the most in Montana, to 17.9 percent. Nevada had the lowest increase in recurrence, at 1.3 percent. In Nevada, the average fetched of a claim did not increase entirely.

Variety:

This collection includes many datatypes such as integers, float values, date format, strings, and negative values. Furthermore, it provides a wide range of parameters from which to choose, each of which has a different impact on accidents. Here's how it works:Human error is the leading cause of accidents and crashes. We are describing some common human behaviours that contribute to accidents. Numerous domestic and international research have discovered that this is the most common type of road driver behaviour that leads to collisions.

- Driving while texting is prohibited.
- Using a GPS gadget or an app.
- I was driving and eating.
- Talking with passengers or on the phone.
- Changing the climate control.
- Daydreaming.

  And certain types of mishaps are caused by:
- Data on traffic accidents based on vehicle make and model
- Data on traffic accidents depending on route type and condition
- Data on traffic accidents depending on location
- Data about traffic accidents dependent on time of day/year
- Data on traffic accidents based on long-term climate and fluctuating weather circumstances

Value:

The value of big data is determined by how useful the information obtained is to your firm. No matter how much data there is, it rarely has much value on its own; for it to be valuable, it must be turned into insights or knowledge, which is where data processing comes in. This dataset on accidents in the United States includes data from all 49 states, providing more than enough information to complete the project. The accident data was gathered using several APIs that provide streaming traffic incident data from February 2016 to December 2021. Only a few of the organisations are the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors embedded in road networks.

7 Discussions around Relavent Metrics

Completeness:

Completeness refers to the extent to which all required data is present in a dataset.Check if there are any missing values in any of the columns.

Accuracy:

Accuracy is the degree to which data reflects the true values or reality.Evaluate the accuracy of numerical values like Temperature.Ensure that these values are within reasonable and valid ranges. Check for any discrepancies or errors in the data.

Consistency:

Consistency refers to the uniformity and coherence of data across different sources or within the same dataset. Look for inconsistencies in the data. For instance, check if the same location is represented consistently across different columns.

Relevance:

Relevance measures the significance of the data in relation to the goals and requirements of the analysis or task. Assess the relevance of each column to the purpose of your analysis. may or may not be relevant depending on the analysis goals. Identify and filter out irrelevant columns for your specific use case. By examining the dataset through the lens of completeness, accuracy, consistency, and relevance, you can ensure that the data is of high quality and suitable for your analytical or modeling needs.

8 Latency and Processing Time for accidents in the United States

Initial Latency Assessment:

Initial latency assessment involves evaluating the time delay or lag between the occurrence of an accident and the recording of relevant data. In the context of the provided dataset, this would involve analyzing how quickly accident information is recorded and made available for further processing. It could include assessing the time elapsed between the actual accident occurrence and the timestamp recorded in the dataset.

Processing Time for Data Analysis:

Processing time for data analysis refers to the duration taken to perform analytical tasks, such as extracting insights, generating reports, or running algorithms on the dataset. Evaluate the efficiency of the data analysis processes applied to the accident dataset. This could include assessing the time required to perform tasks like aggregating statistics, identifying patterns, or generating visualizations based on the accident data.

Latency in Transmission Handling:

Latency in transmission handling measures the delay introduced when transmitting data from one system or location to another. In the context of the dataset, it involves assessing the time delay introduced during the transmission of accident data. This could include the time taken to send, receive, and process data between different systems or entities involved in handling accident information.

Optimizing Query Response Time:

Optimizing query response time involves improving the speed at which queries or requests for information are processed and returned.Evaluate and implement strategies to reduce the time it takes to retrieve specific information from the dataset. This could include optimizing database queries, indexing relevant columns, or utilizing caching mechanisms to enhance the responsiveness of queries related to accident data.

9 Resource Utilization, Security, and Cost

Resource Utilization:

Resource Utilization refers to the efficient allocation and usage of computing resources such as CPU, memory, storage, and network bandwidth in

handling and processing the dataset. Assess how well the systems and infrastructure managing the accident dataset utilize resources. This involves monitoring and optimizing the usage of computational resources to ensure that data processing tasks are performed efficiently. For instance, optimizing database queries, employing parallel processing, and using appropriate hardware configurations can enhance resource utilization.

Security:

Security involves safeguarding the dataset against unauthorized access, data breaches, and ensuring the confidentiality, integrity, and availability of the information. Evaluate the security measures in place to protect the accident dataset. This includes access controls, encryption, and other security protocols to prevent unauthorized access. Additionally, assess data integrity to ensure that the dataset remains accurate and unaltered. Regular audits and compliance with data protection regulations are also critical aspects of ensuring dataset security.

Cost:

Cost refers to the financial expenditure associated with managing, processing, and maintaining the accident dataset, including infrastructure costs, software licensing, and personnel expenses. Analyze the cost-effectiveness of the systems and processes involved in handling the dataset. Considerations include infrastructure costs, licensing fees for software tools, and expenses related to personnel involved in managing and analyzing the data. Cost optimization strategies may involve adopting efficient technologies, utilizing cloud services judiciously, and optimizing workflows to reduce overall operational expenses.

10 Conclusion

Considered optimizing query response times and transmission handling to improve real-time incident reporting. Implement strategies to enhance resource utilization, ensuring efficient processing and analysis. Strengthen security measures to protect the confidentiality and integrity of the dataset. Explore cost-effective solutions, including cloud services and streamlined workflows. The dataset comprises diverse incidents, including accidents, road hazards, and other traffic-related issues. Incidents vary in severity, providing a comprehensive view of the traffic conditions. Temporal data, including start and end times, provides insights into the timing of incidents. Trends related to specific times of the day, days of the week, or seasons can be identified. Weather-related information such as temperature, wind speed, and precipitation allows for understanding the influence of weather on traffic incidents. Conditions like rain, snow, or fog might contribute to accidents.Latency and

processing times in incident reporting and handling may impact real-time decision-making. Opportunities exist for optimizing resource utilization, enhancing security measures, and managing costs effectively.

11 References

https://jupyter.org/
https://pandas.pydata.org/
https://spark.apache.org/
https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents
https://github.com/S556508/Impact-Players