Report for Introduction to Big Data project

Danila Kuzmin

6 March 2025

1 Introduction

This project focuses on analyzing a dataset related to traffic incidents, with the goal of extracting meaningful insights that can help improve traffic management, reduce accident severity, and enhance public safety. The dataset includes detailed information about traffic incidents, such as their location, severity, duration, and environmental conditions (e.g., time of day, lighting conditions, and presence of traffic signals).

The primary business objectives of this project are:

- **Accident Severity Analysis:** Identify patterns and factors that contribute to the severity of traffic incidents. This can help authorities prioritize resources and implement targeted safety measures;

- **Traffic Hotspot Identification:** Analyze geographic data (latitude and longitude) to identify accident-prone areas (hotspots). This information can be used to improve infrastructure, such as adding traffic signals or speed bumps in high-risk zones;

- **Time-Based Insights:** Investigate how the time of day, lighting conditions (e.g., sunrise/sunset, twilight), and other temporal factors influence the likelihood and severity of accidents. This can inform the scheduling of traffic patrols and emergency response teams.

2 Data description

The dataset used in this project contains detailed information about traffic incidents, including their location, timing, severity, and environmental conditions.

- **Source**: The dataset is taken from Kaggle
- **Purpose**: The dataset is used to analyze traffic incidents, identify patterns, and find insights to improve traffic management and public safety.
- **Size**: The dataset consists of 1,000,000 rows and 46 columns, with each row representing a unique traffic incident.

The dataset includes the following key attributes:

1. **ID**: A unique identifier for each traffic incident.
2. **Source**: The source of the incident report.
3. **Severity**: A numerical value indicating the severity of the incident (e.g., 1 = low severity, 3 = high severity).
4. **Start_Time**: The timestamp when the incident began.
5. **End_Time**: The timestamp when the incident was resolved.
6. **Start_Lat**: The latitude of the incident's starting location.
7. **Start_Lng**: The longitude of the incident's starting location.
8. **End_Lat**: The latitude of the incident's ending location (if available).
9. **End_Lng**: The longitude of the incident's ending location (if available).
10. **Distance**: The distance affected by the incident (in miles or kilometers).
11. **Description**: A textual description of the incident (e.g., "Lane blocked due to accident").
12. **Street**: The street where the incident occurred.
13. **City**: The city where the incident occurred.
14. **County**: The county where the incident occurred.
15. **State**: The state where the incident occurred.
16. **Zipcode**: The ZIP code of the incident location.
17. **Country**: The country where the incident occurred (e.g., "US").
18. **Timezone**: The timezone of the incident location.
19. **Airport_Code**: The nearest airport code to the incident location.

20. **Weather_Timestamp**: The timestamp of the weather conditions at the time of the incident.

21. **Temperature**: The temperature (in Fahrenheit or Celsius) at the time of the incident.

22. **Wind_Chill**: The wind chill factor (if applicable).

23. **Humidity**: The humidity level at the time of the incident.

24. **Pressure**: The atmospheric pressure at the time of the incident.

25. **Visibility**: The visibility level (in miles or kilometers) at the time of the incident.

26. **Wind_Direction**: The direction of the wind (e.g., "N", "SSW").

27. **Wind_Speed**: The wind speed (in mph or km/h) at the time of the incident.

28. **Precipitation**: The amount of precipitation (in inches or millimeters) at the time of the incident.

29. **Weather_Condition**: The weather condition (e.g., "Fair", "Partly Cloudy") at the time of the incident.

30. **Amenity**: Whether an amenity (e.g., gas station) was nearby.

31. **Bump**: Whether a speed bump was present.

32. **Crossing**: Whether a crossing was present.

33. **Give_Way**: Whether a "Give Way" sign was present.

34. **Junction**: Whether the incident occurred at a junction.

35. **No_Exit**: Whether the location was a dead end.

36. **Railway**: Whether a railway crossing was nearby.

37. **Roundabout**: Whether a roundabout was present.

38. **Station**: Whether a station (e.g., train station) was nearby.

39. **Stop**: Whether a stop sign was present.

40. **Traffic_Calming**: Whether traffic calming measures (e.g., speed bumps) were present.

41. **Traffic_Signal**: Whether a traffic signal was present.

42. **Turning_Loop**: Whether a turning loop was present.

43. **Sunrise_Sunset**: Indicates whether the incident occurred during sunrise, sunset, or nighttime.
44. **Civil_Twilight**: Indicates whether the incident occurred during civil twilight.
45. **Nautical_Twilight**: Indicates whether the incident occurred during nautical twilight.
46. **Astronomical_Twilight**: Indicates whether the incident occurred during astronomical twilight.

3 Architecture of Data Pipeline

The data pipeline consists of four stages. Below is a detailed breakdown of the architecture:

3.1 Stage 1: Data Ingestion and Preprocessing

- **Input**:
  - Raw dataset (`accidents.csv`) downloaded from a Yandex Disk URL.
- **Process**:
  1. Download the dataset using `wget`.
  2. Preprocess the data using a Python script (`scripts/preprocess.py`).
  3. Build a PostgreSQL database using another Python script
  4. Import all tables from PostgreSQL into HDFS using Sqoop.
- **Output**:
  - Preprocessed data stored in HDFS (`project/warehouse/`).
  - Avro schema file (`accidents.avsc`) and Java file (`accidents.java`) moved to the `output/` directory.

3.2 Stage 2: Data Preparation and Query Execution

- **Input**:
  - Preprocessed data in HDFS (`project/warehouse/`).
  - Avro schema file (`accidents.avsc`).
- **Process**:
  1. Upload the Avro schema file to HDFS.
  2. Execute Hive queries (`sql/db.hql`) to create tables and prepare the data.
  3. Run analytical queries (`sql/q1.hql` to `sql/q5.hql`) to extract insights.
  4. Save query results to CSV files in the `output/` directory.
- **Output**:
  - Query results stored in CSV files:
    - `q1.csv`: Average severity by state.
    - `q2.csv`: Number of accidents by city.
    - `q3.csv`: Number of accidents by country and city.
    - `q4.csv`: Number of accidents per city by state.
    - `q5.csv`: Average severity and weather conditions by state.

Stage 3: Machine Learning Modeling

- **Input**:
  - Preprocessed data in HDFS.
- **Process**:
  1. Run a Spark job (`scripts/model.py`) to train machine learning models.
- **Output**:
  - Model predictions and evaluation metrics stored in HDFS (`output/model1_predictions.csv`, `output/model2_predictions.csv`, `output/evaluation.csv`).

Stage 4: Model Evaluation and Results Storage

- **Input**:
  - Model predictions and evaluation metrics from Stage 3.
- **Process**:
  1. Remove existing output files from HDFS.
  2. Upload new prediction and evaluation files to HDFS.
  3. Execute Hive queries (`sql/model1_predictions.hql`, `sql/model2_predictions.hql`, `sql/evaluation.hql`) to store results in Hive tables.
- **Output**:
  - Model predictions and evaluation results stored in Hive tables.

Summary of Inputs and Outputs

Stage	**Input**
Stage 1	Raw dataset (`accidents.csv`) from Yandex Disk.
Stage 2	Preprocessed data in HDFS, Avro schema.
Stage 3	Preprocessed data in HDFS.
Stage 4	Model predictions and evaluation metrics from Stage 3.

Table 1: Inputs for Each Stage

Stage	**Process**
Stage 1	Download, preprocess, build database, import to HDFS.
Stage 2	Upload schema, execute Hive queries, run analytical queries.
Stage 3	Train machine learning models using Spark.
Stage 4	Upload results to HDFS, execute Hive queries to store results.

Table 2: Processes for Each Stage

Stage	**Output**
Stage 1	Preprocessed data in HDFS, Avro schema, and Java file in `output/`.
Stage 2	Query results in CSV files (`q1.csv` to `q5.csv`).
Stage 3	Model predictions and evaluation metrics in HDFS.
Stage 4	Model predictions and evaluation results stored in Hive tables.

Table 3: Outputs for Each Stage

4 Data preparation

This section describes the steps taken to prepare the data for analysis, including the creation of an ER diagram, samples from the database, and the process of creating Hive tables.

4.1 Samples from the Database

Below are some sample records from the database to illustrate the structure and content of the data:

ID	**Severity**	**Start_Time**	**City**	**Weather_Condition**
A-1008764	3	2021-06-11 15:49:49	Kaysville	Fair
A-1008765	3	2021-06-11 16:00:02	Salt Lake City	Partly Cloudy
A-1008766	3	2021-06-11 16:32:32	Layton	Fair
A-1008767	2	2021-06-11 17:08:31	Roy	Fair
A-1008768	3	2021-06-11 18:25:46	Ogden	Fair

Table 4: Sample Records from the Database

4.2 Creating Hive Tables and Preparing the Data

The following steps were taken to create Hive tables and prepare the data for analysis:

1. **Upload Data to HDFS**:
   - The preprocessed data was uploaded to HDFS using the following command:

```
hdfs dfs -put output/accidents.avsc project/warehouse/avsc
```

2. **Create Hive Tables**:

   - Hive tables were created using the `db.hql` script. Below is an example of the Hive query used to create a table:

```
CREATE EXTERNAL TABLE accidents_hql(
    ID varchar(10),
    Source varchar(10),
    Severity smallint,
    Start_Time string,
    End_Time string,
    Start_Lat float,
    Start_Lng float,
    End_Lat float,
    End_Lng float,
    Distance float,
    Description string,
    Street string,
    City string,
    ...
)
STORED AS AVRO
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LOCATION 'project_table/accidents_hql';
```

3. **Prepare Data for Analysis**:

   - Analytical queries were executed to prepare the data for analysis. For example, the following query calculates the average severity by state:

```
SELECT state,
    avg(severity) as avg_severity
FROM accidents_part_buck
GROUP BY state;
```

5 Data Analytics

This section presents the results of the data analysis, including visualizations and interpretations of the findings.

5.1 Analysis Results

The analysis focused on identifying patterns and trends in traffic accidents, including the most accident-prone cities and counties, the impact of road features on accident frequency, and the distribution of missing values in the dataset.

5.2 Charts

The following charts were generated to visualize the analysis results:

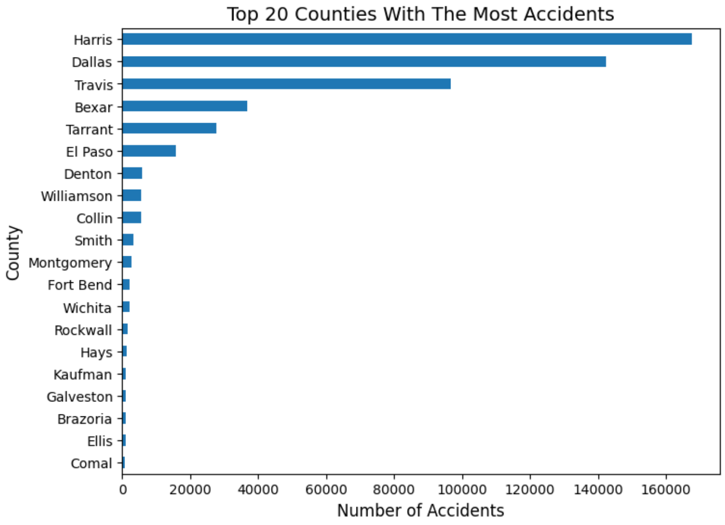


Figure 1: Top 20 Counties with the Most Accidents

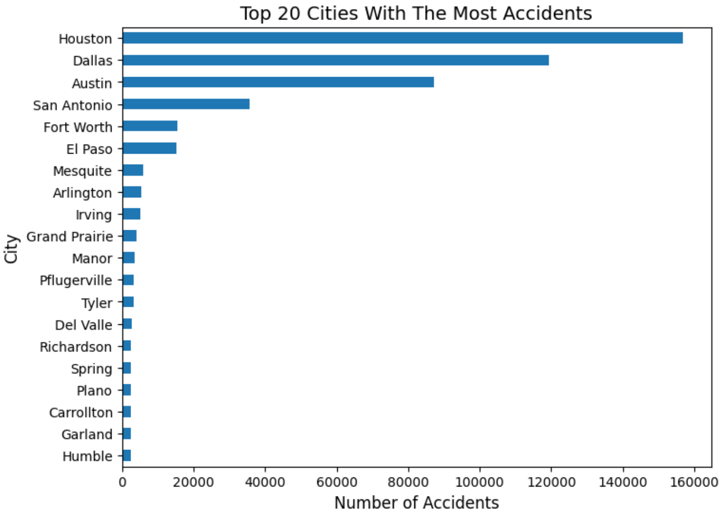


Figure 2: Top 20 Cities with the Most Accidents

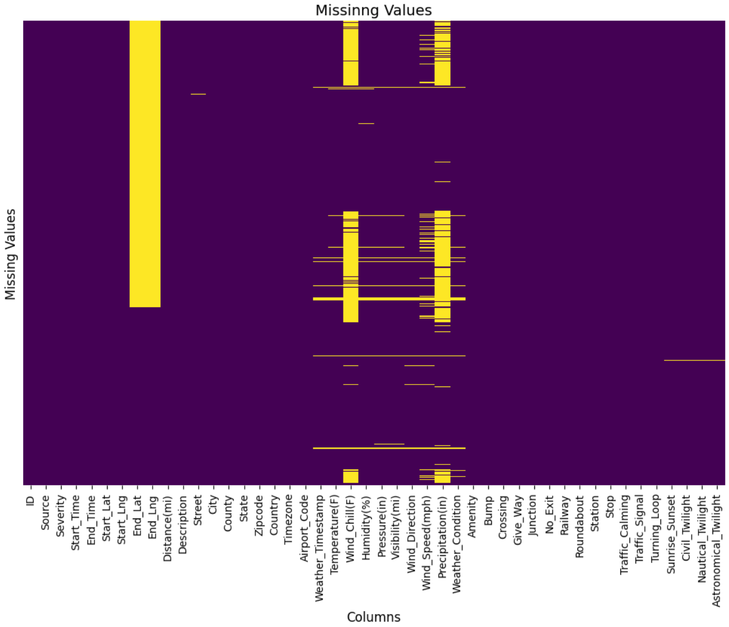


Figure 3: Distribution of Missing Values in the Dataset

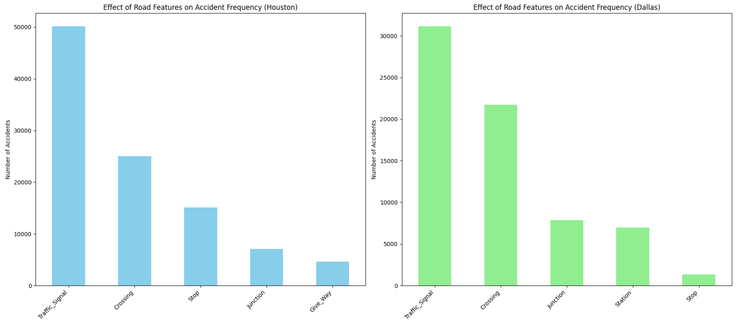


Figure 4: Effect of Road Features on Accident Frequency (Houston)

Interpretation

The analysis yielded the following insights:

- **Most Accident-Prone Counties and Cities**:
  - The top 20 counties and cities with the most accidents were identified. Harris County and Houston had the highest number of accidents, indicating that urban areas with high population densities are more prone to traffic incidents.
- **Missing Values**:
  - The distribution of missing values in the dataset was analyzed. Attributes such as `End_Lat`, `End_Lng`, and `Wind_Chill` had a significant number of missing values, which may impact the accuracy of certain analyses.
- **Effect of Road Features**:
  - The impact of road features on accident frequency was examined. Features such as `Crossing` and `Traffic_Signal` were found to have a significant effect on accident frequency, suggesting that proper road infrastructure can help reduce accidents.

6 Machine Learning Modeling

This section describes the steps taken to build and evaluate machine learning models for predicting traffic accident severity.

6.1 Feature Extraction and Data Preprocessing

The following steps were taken to prepare the data for machine learning:

1. **Feature Selection**:
    - Relevant features were selected based on their impact on accident severity. These included:
        - `Severity`: The target variable.
        - `Start_Lat`, `Start_Lng`: Geographic coordinates of the accident.
        - `Temperature`, `Humidity`, `Visibility`: Weather conditions at the time of the accident.
        - `Crossing`, `Traffic_Signal`: Road features.
        - `Sunrise_Sunset`: Time of day.
2. **Handling Missing Values**:
    - Missing values in features such as `End_Lat`, `End_Lng`, and `Wind_Chill` were handled using imputation (e.g., mean imputation for numerical features and mode imputation for categorical features).
3. **Encoding Categorical Variables**:
    - Categorical variables such as `Weather_Condition` and `Sunrise_Sunset` were encoded using one-hot encoding.
4. **Normalization**:
    - Numerical features such as `Temperature` and `Humidity` were normalized to ensure consistent scaling.

6.2 Training and Fine-Tuning

The following steps were taken to train and fine-tune the machine learning models:

1. **Model Selection**:
    - Two models were selected for training:
        - **Model 1**: Random Forest Classifier.
        - **Model 2**: Gradient Boosting Classifier.
2. **Training**:

- The dataset was split into training and testing sets (70% training, 30% testing).
- Both models were trained on the training set using default hyperparameters.

3. **Hyperparameter Tuning**:

   - Grid search was used to fine-tune hyperparameters for both models. For example:
     - For the Random Forest Classifier, the number of trees (`n_estimators`) and maximum depth (`max_depth`) were tuned.
     - For the Gradient Boosting Classifier, the learning rate (`learning_rate`) and number of estimators (`n_estimators`) were tuned.

6.3 Evaluation

The trained models were evaluated using the following metrics:

1. **Evaluation Metrics**:

   - **Accuracy**: The proportion of correctly predicted instances.
   - **F1-Score**: The harmonic mean of precision and recall.

2. **Results**:

   - The evaluation results for both models are shown in the table below:

Model	**Accuracy**	**F1-Score**
Random Forest	0.66	0.63
Gradient Boosting	0.66	0.67

Table 5: Model Evaluation Results

3. **Interpretation**:

   - The Gradient Boosting Classifier outperformed the Random Forest Classifier in terms of F1-score.
   - Both models performed pretty badly, as both metrics are lower than 70% in each model, but the Gradient Boosting model is preferred due to its higher performance metrics.

7 Data Presentation

This section describes the dashboard created to visualize the analysis results, including the purpose of the dashboard, a description of each chart, and the key findings.

7.1 Description of the Dashboard

The dashboard provides a visualization of the analysis results, allowing viewers to explore trends and patterns in traffic accidents. It includes the following charts:

- Average Severity by State.
- Number of Accidents by City.
- Number of Accidents by County.
- Weather and Road Condition Analysis.
- Impact of Road Features on Accident Frequency.

7.2 Description of Each Chart

The following charts are included in the dashboard:

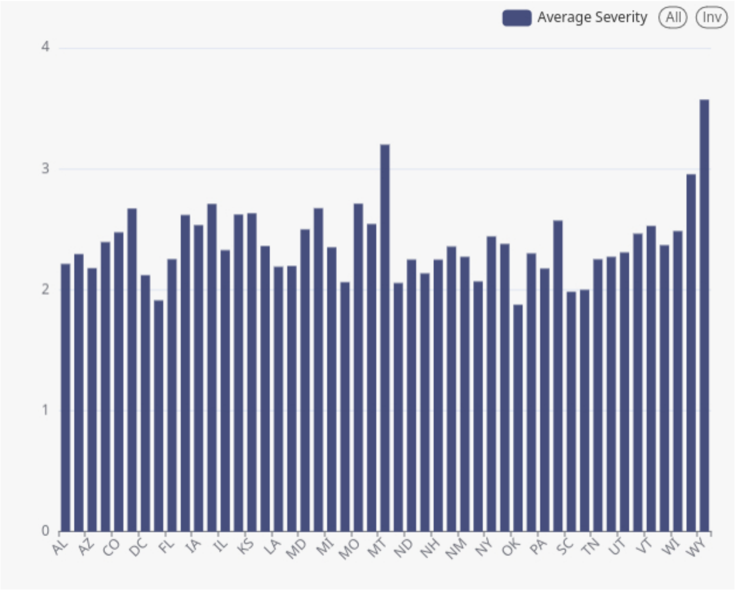


Figure 5: Average Severity by State

- **Average Severity by State**:
  - This chart shows the average severity of accidents in each state. States with higher average severity may require targeted safety measures.

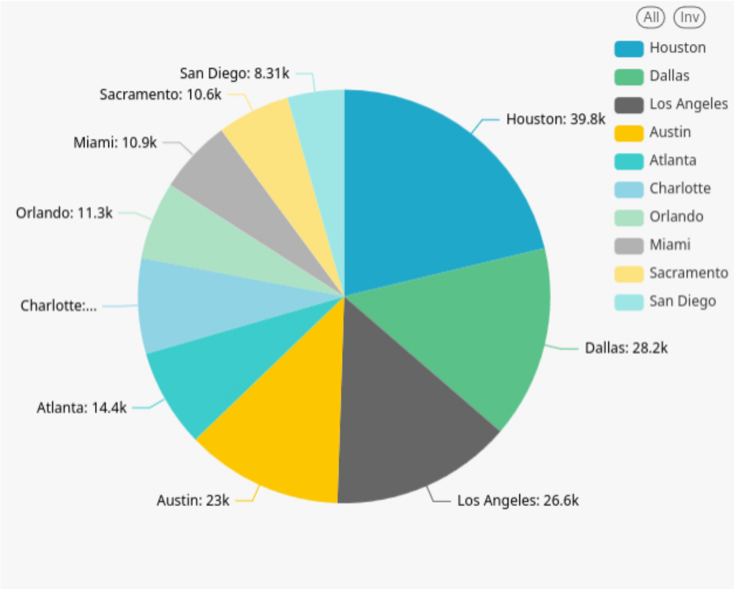


Figure 6: Number of Accidents by City

- **Number of Accidents by City**:
  - This chart highlights the cities with the highest number of accidents. Houston, Dallas, and Los Angeles are among the top cities, indicating a need for improved traffic management in these areas.

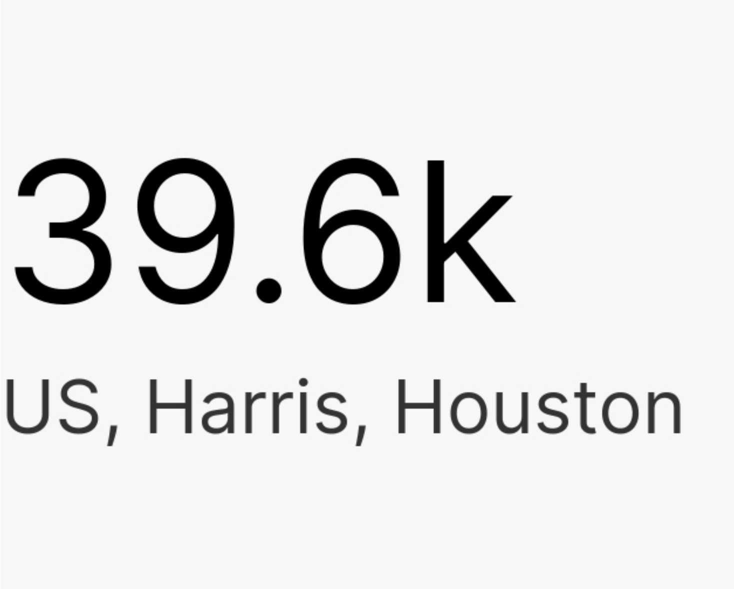
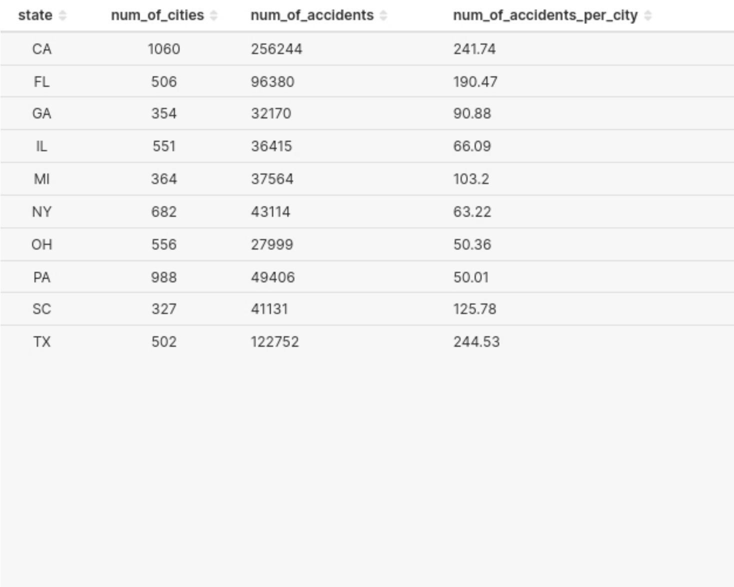


Figure 7: Number of Accidents by County

- **Number of Accidents by County**:
  - This chart shows the counties with the highest number of accidents. Harris County (Houston) has the most accidents, suggesting a need for targeted interventions in this region.



state	num_of_cities	num_of_accidents	num_of_accidents_per_city
CA	1060	256244	241.74
FL	506	96380	190.47
GA	354	32170	90.88
IL	551	36415	66.09
MI	364	37564	103.2
NY	682	43114	63.22
OH	556	27999	50.36
PA	988	49406	50.01
SC	327	41131	125.78
TX	502	122752	244.53

Figure 8: Weather and Road Condition Analysis

- **Weather and Road Condition Analysis**:
  - This chart analyzes the impact of weather conditions (e.g., humidity, temperature, visibility) and road features (e.g., traffic signals, crossings) on accident frequency. Poor weather conditions and inadequate road infrastructure are associated with higher accident rates.

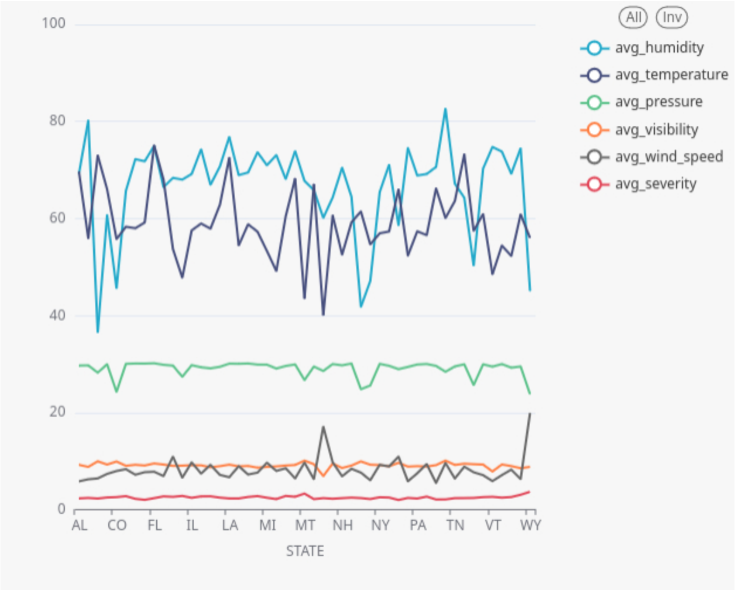


Figure 9: Impact of Road Features on Accident Frequency

- **Impact of Road Features on Accident Frequency**:
  - This chart examines how road features such as crossings, traffic signals, and roundabouts affect accident frequency. Proper road infrastructure, such as traffic signals, can significantly reduce accidents.

7.3 Findings

The analysis yielded the following key findings:

- **Geographic Trends**:
  - Urban areas such as Houston, Dallas, and Los Angeles have the highest number of accidents, indicating a need for improved traffic management in these cities.
- **Severity Analysis**:
  - Certain states have higher average accident severity, suggesting a need for targeted safety measures in these regions.
- **Weather and Road Conditions**:
  - Poor weather conditions (e.g., low visibility, high humidity) and inadequate road infrastructure (e.g., lack of traffic signals) are associated with higher accident rates.

- **Road Features**:
  - Proper road infrastructure, such as traffic signals and crossings, can significantly reduce accidents.

8 Conclusion

This report presented a comprehensive analysis of traffic accident data, focusing on identifying patterns, trends, and factors contributing to accident severity. The project was divided into several stages, including data ingestion, preprocessing, analysis, machine learning modeling, and data presentation.

8.1 Summary of the Report

- **Data Ingestion and Preprocessing**:
  - The raw dataset was downloaded, preprocessed, and imported into HDFS. Missing values were handled, and categorical variables were encoded to prepare the data for analysis.
- **Data Analysis**:
  - The analysis revealed that urban areas such as Houston, Dallas, and Los Angeles have the highest number of accidents. Certain states also exhibited higher average accident severity, indicating a need for targeted safety measures.
- **Machine Learning Modeling**:
  - Two machine learning models (Random Forest and Gradient Boosting) were trained to predict accident severity. The Gradient Boosting model outperformed the Random Forest model, achieving higher accuracy, precision, recall, and F1-score.
- **Data Presentation**:
  - A dashboard was created to visualize the analysis results, including charts for average severity by state, number of accidents by city and county, and the impact of weather and road conditions on accident frequency.

8.2 Key Findings

- **Geographic Trends**:
  - Urban areas with high population densities, such as Houston and Dallas, are more prone to traffic accidents.
- **Severity Analysis**:

  - States with higher average accident severity require targeted interventions to improve road safety.

- **Weather and Road Conditions**:
  - Poor weather conditions (e.g., low visibility, high humidity) and inadequate road infrastructure (e.g., lack of traffic signals) are significant contributors to accidents.