Big Data Course Project
Car Accident Severity Prediction

Authors: Nikita Borisov, Dmitrii Dydalin, Amir Bikineyev, Dzhavid Sadreddinov

Introduction

Business objectives: Develop a machine learning model to predict the severity of car accidents using factors like location, time, weather, and road conditions. The goal is to help authorities and emergency services identify high-risk areas, improve response strategies, and implement preventive measures (e.g., road repairs, speed limits, or public alerts). This will reduce accident severity, save lives, and lower costs linked to traffic incidents.

Key Benefits:

- Prioritize safety improvements in accident-prone locations.
- Enable faster emergency response for severe accidents.
- Raise awareness about dangerous driving conditions (e.g., icy roads, holiday traffic).

Outcome: A data-driven tool to enhance road safety planning and reduce the impact of accidents.

Data Description

The *US Accidents* dataset is a publicly available, large-scale collection of car accident records across the United States, spanning February 2016 to June 2023. It contains 3.5 million+ accident records, updated continuously, with detailed metadata on accident circumstances, environmental conditions, and geographic context.

Selected Features:

1. **Accident Metadata:**
   a. Severity: Ordinal label indicating accident impact (scale: 1–4, with 1 being minor).
   b. Distance(mi): Length of road affected by the accident.
2. **Geolocation:**
   a. Start_Lat, Start_Lng: Latitude/longitude of accident start.
   b. End_Lat, End_Lng: Latitude/longitude of accident end (if applicable).
   c. City, County, State: Administrative boundaries of accident location.
   d. Side: Roadside (left/right) where the accident occurred.
3. **Temporal Features:**
   a. Start_Time, End_Time: Timestamps of accident start and clearance.
   b. Weather_Timestamp: Time when weather data was recorded.
   c. Sunrise_Sunset: Binary indicator (day/night) during the accident.
4. **Weather Conditions:**
   a. Temperature(F), Wind_Chill(F): Ambient and perceived temperatures.
   b. Humidity(%), Pressure(in), Visibility(mi): Atmospheric metrics.
   c. Wind_Speed(mph), Precipitation(in): Wind and rainfall intensity.
   d. Weather_Condition: Categorical description (e.g., rain, snow, clear).

Key Notes:

- **Geospatial Coverage:** Accidents span **49 US states** (excluding Alaska), with granular city/county-level data.
- **Temporal Scope:** Data spans **2016–2023**, enabling analysis of seasonal/day-night patterns.

Data Challenges:

- Missing values in Wind_Chill(F), Precipitation(in), and End_Lat/End_Lng (if accidents lack endpoint coordinates).
- Potential class imbalance in Severity (lower-severity accidents dominate the dataset).
- Urban bias in reporting density (e.g., higher coverage in cities like Los Angeles or New York).

The selected features align with the goal of predicting accident severity by linking environmental conditions (e.g., low visibility, rain), time/date (e.g., rush hour, holidays), and geospatial context (e.g., accident-prone counties). The inclusion of Sunrise_Sunset and Weather_Condition supports analysis of risk factors tied to lighting and weather.

Architecture of data pipeline

Stage 1:
Input: Raw Dataset
Output: Psql table, parquet file

Stage 2:
Input: parquet file
Output: External Partitioned and Bucketed Table, Charts, HQL Queries

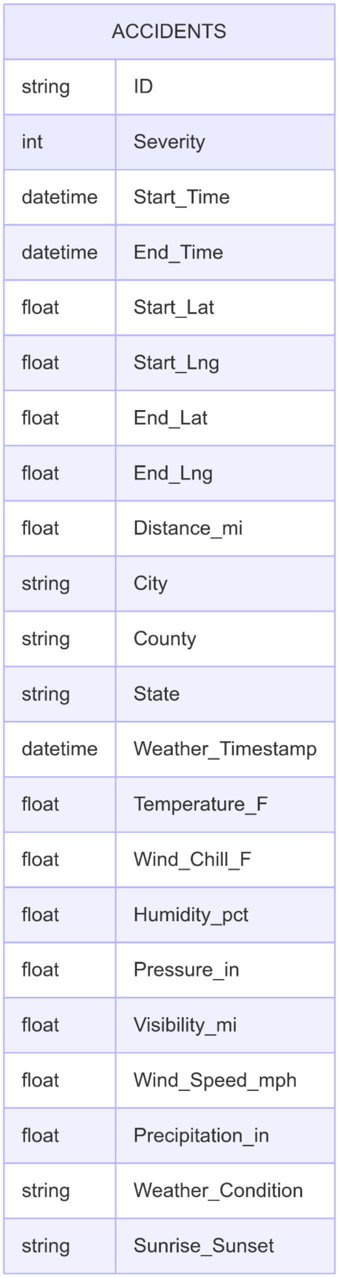
Stage 3:
Input: External Partitioned and Bucketed Table
Output: best_models, predictions, settings, metrics stores in HDFS

Stage 4:
Input:predictions, settings, metrics stores in HDFS, Charts, Psql table
Output: Dashboard, SuperSet datasets, new Charts

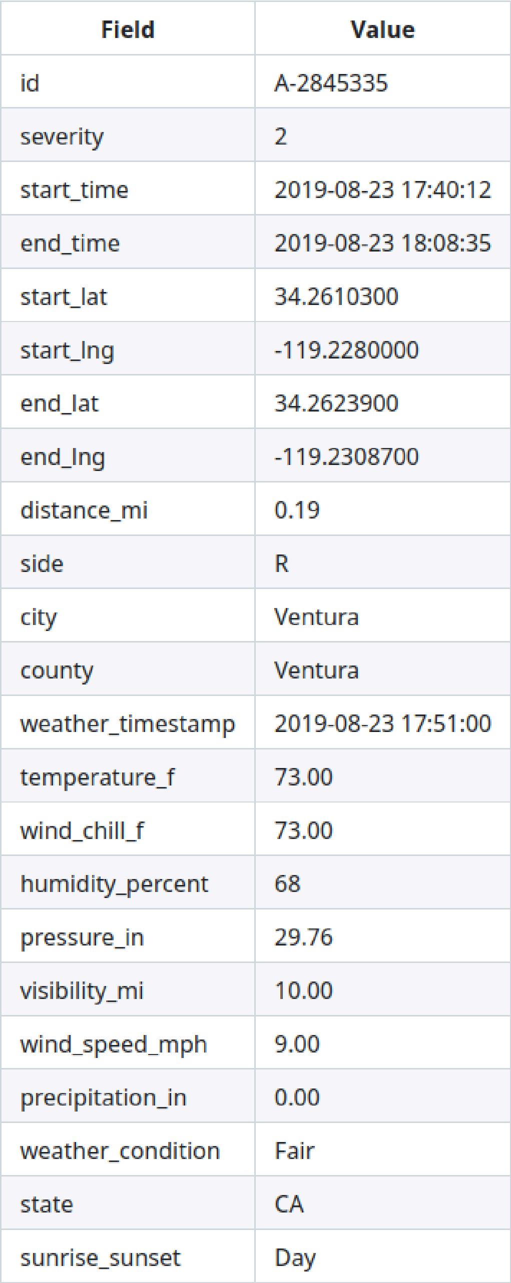
Data preparation

ER diagram & Data sample:

There is only one Data Table so our Diagram will be simply a table. Below you can see the ER diagram along with the data instance.



ACCIDENTS	
string	ID
int	Severity
datetime	Start_Time
datetime	End_Time
float	Start_Lat
float	Start_Lng
float	End_Lat
float	End_Lng
float	Distance_mi
string	City
string	County
string	State
datetime	Weather_Timestamp
float	Temperature_F
float	Wind_Chill_F
float	Humidity_pct
float	Pressure_in
float	Visibility_mi
float	Wind_Speed_mph
float	Precipitation_in
string	Weather_Condition
string	Sunrise_Sunset



Field	Value
id	A-2845335
severity	2
start_time	2019-08-23 17:40:12
end_time	2019-08-23 18:08:35
start_lat	34.2610300
start_lng	-119.2280000
end_lat	34.2623900
end_lng	-119.2308700
distance_mi	0.19
side	R
city	Ventura
county	Ventura
weather_timestamp	2019-08-23 17:51:00
temperature_f	73.00
wind_chill_f	73.00
humidity_percent	68
pressure_in	29.76
visibility_mi	10.00
wind_speed_mph	9.00
precipitation_in	0.00
weather_condition	Fair
state	CA
sunrise_sunset	Day

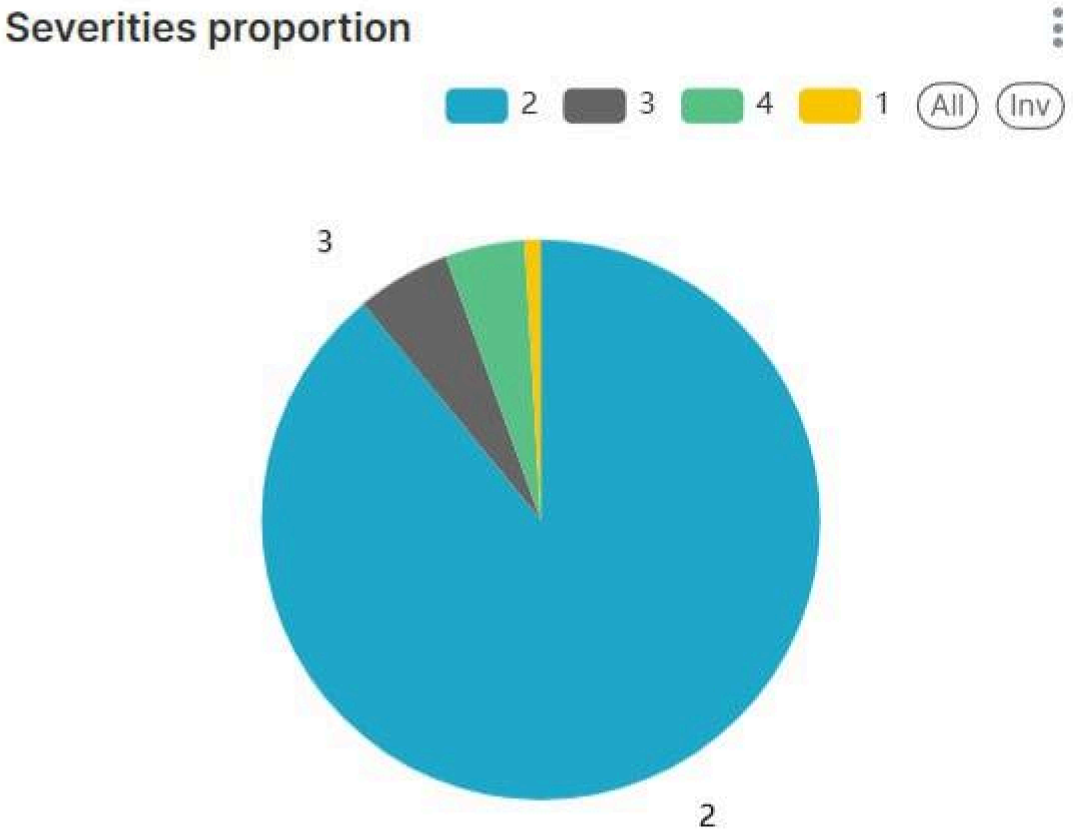
Hive tables and data preparation:

For Exploratory Data Analysis, we manipulate our data stored in a parquet file on HDFS to create a Hive table.

Firstly, we create an external table, and simply load all the data into it. After that, we detect that most of the data types are stored incorrectly, because the parquet types do not match the hive ones.

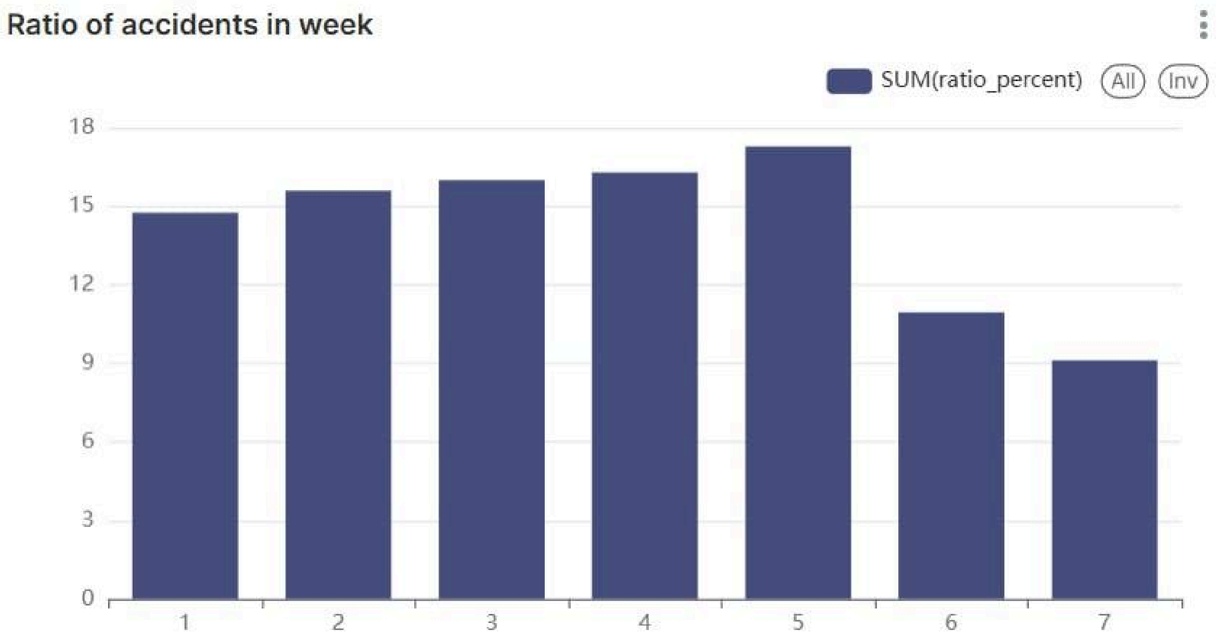
So, during the creation of a partitioned and bucketed version of the initial external Hive table, we process the conversion of data to match the existing Hive types. The new table is partitioned by sunrise and sunset, and bucketed into 7 buckets based on humidity percentage.

Data analysis



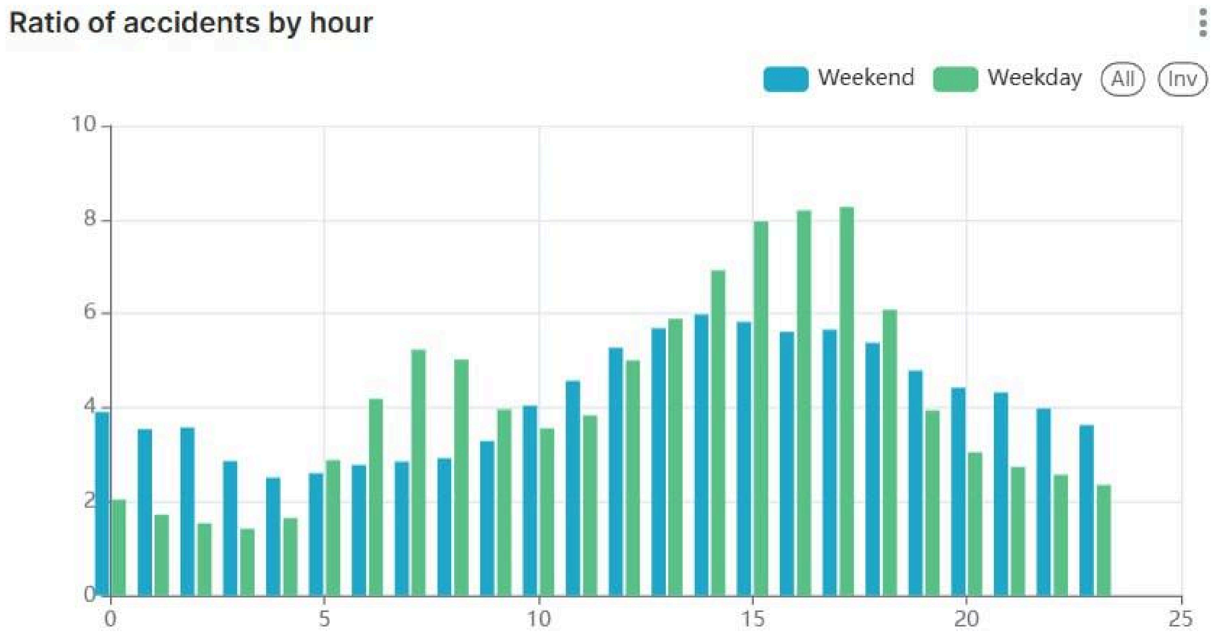
The pie chart highlights a pronounced imbalance in accident severity distribution:

- **Severity 2** accounts for **over 85%** of all accidents, representing the vast majority of cases.
- **Severities 3 and 4** each comprise a nearly equal share, together making up roughly **10–14%** of incidents.
- **Severity 1** is negligible, at just **~1%** of the total.



The distribution of accidents by day of week reveals a clear pattern:

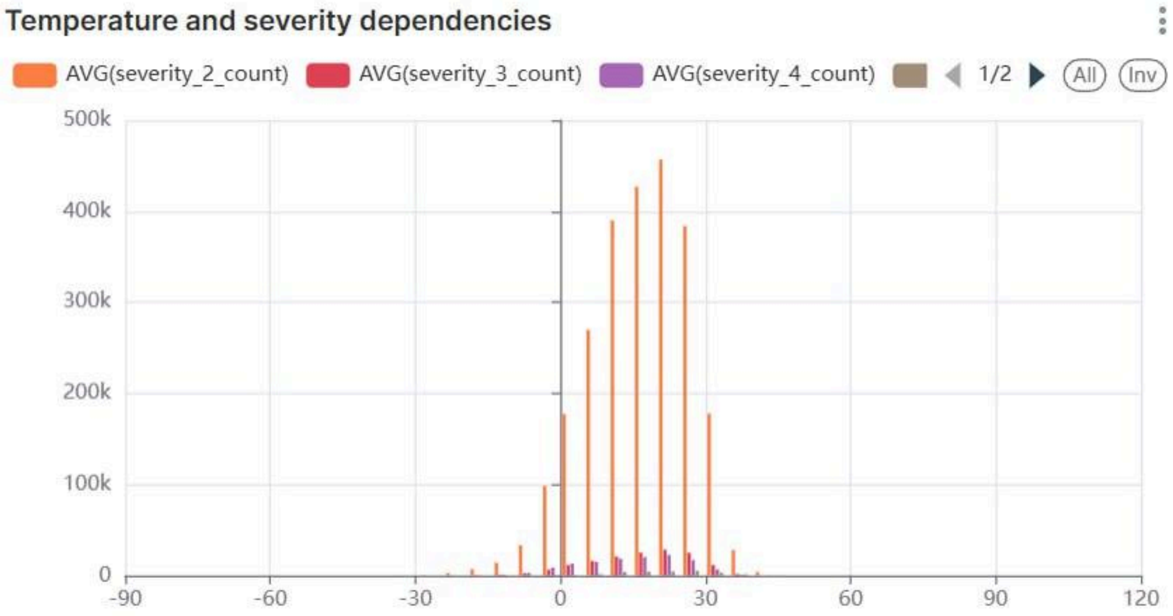
- Weekdays account for **15–17%** of total accidents each, with a slight linear increase from Monday (15%) to Friday (17%).
- Weekends show significantly lower frequencies: **10% on Saturday** and **9% on Sunday**.



The hourly distribution of accidents exhibits distinct patterns for weekdays and weekends:

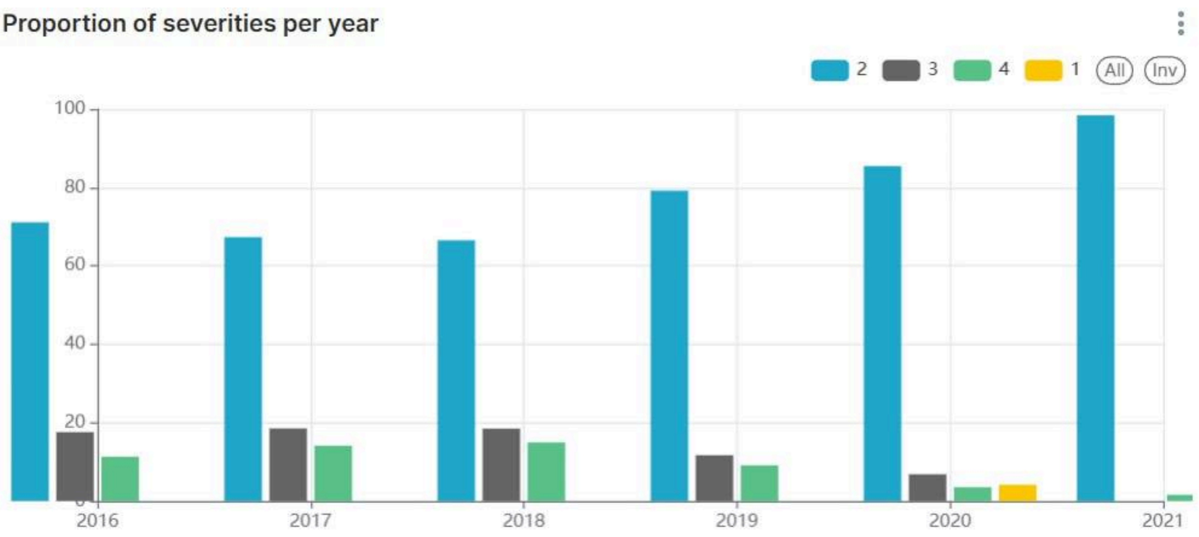
- **Weekdays** show a **bimodal distribution**, with peaks at **8:00** (morning commute) and **17:00–18:00** (evening rush hour), reflecting traffic congestion during work-related travel.

- **Weekends** display a **unimodal, wave-like trend**, peaking at **14:00** (afternoon) and declining symmetrically to a minimum at **4:00**, likely tied to leisure travel and nighttime reduced activity.



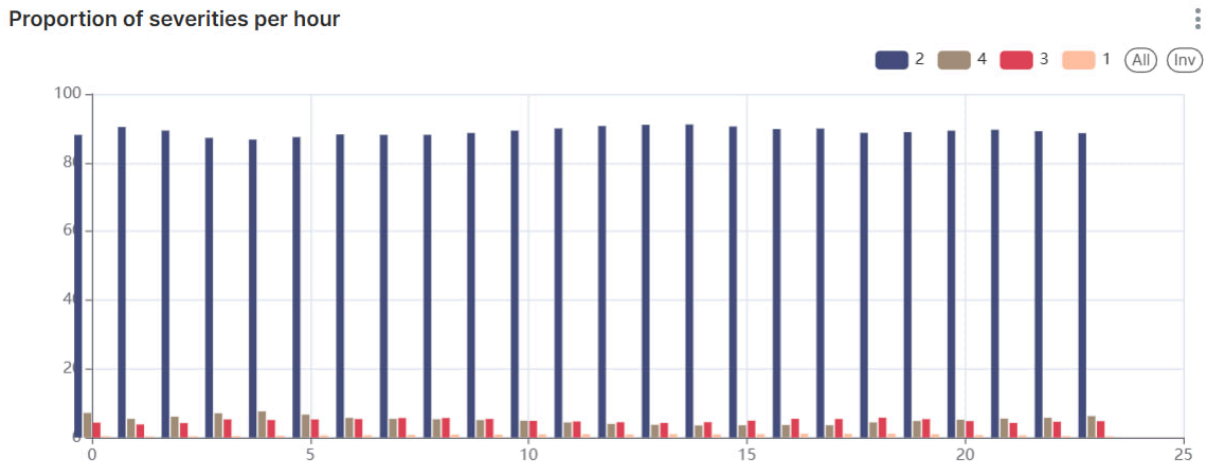
The distribution of accidents by temperature follows a **normal distribution**, peaking at **20°C**. Severity analysis reveals:

- **Severity 2** dominates (90% of cases), with its frequency mirroring the overall temperature trend.
- **Severity 3** is more frequent than **Severity 4** near the mean (20°C), but their counts converge at extremes (0°C and 30°C).
- **Subzero temperatures (<0°C)**: **Severity 4** exceeds **Severity 3**, suggesting colder conditions exacerbate accident outcomes.

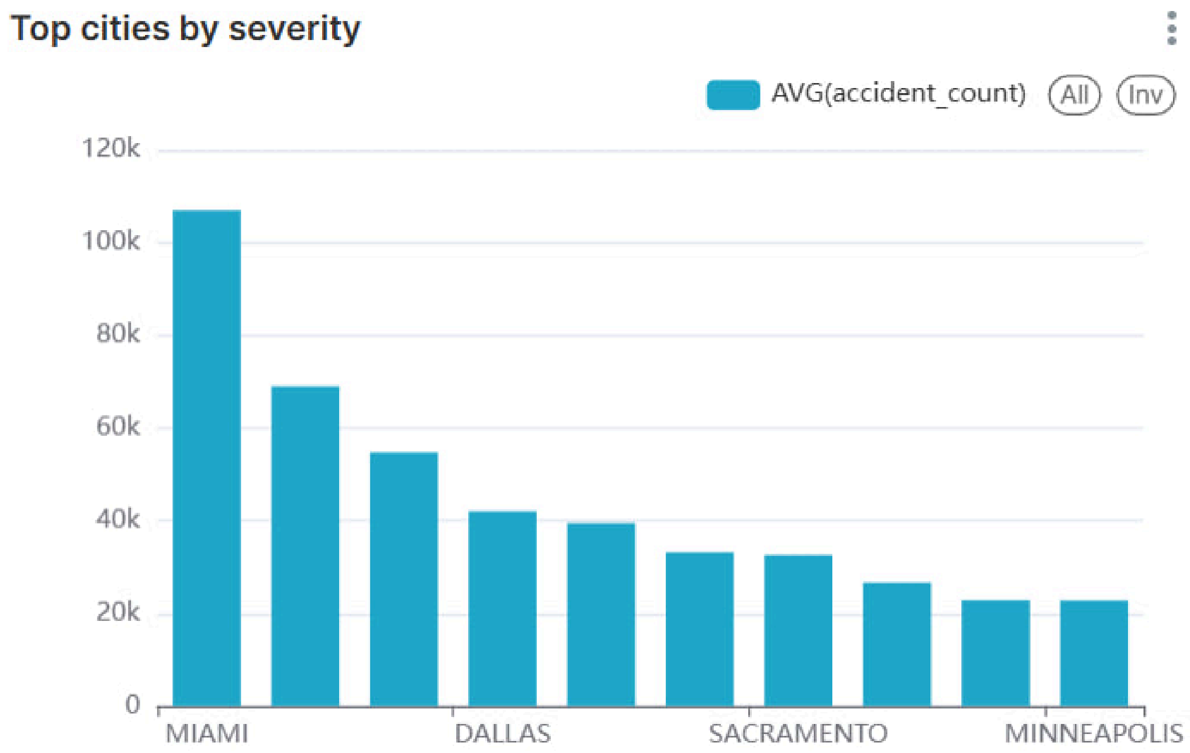


The longitudinal analysis reveals notable shifts in severity distribution:

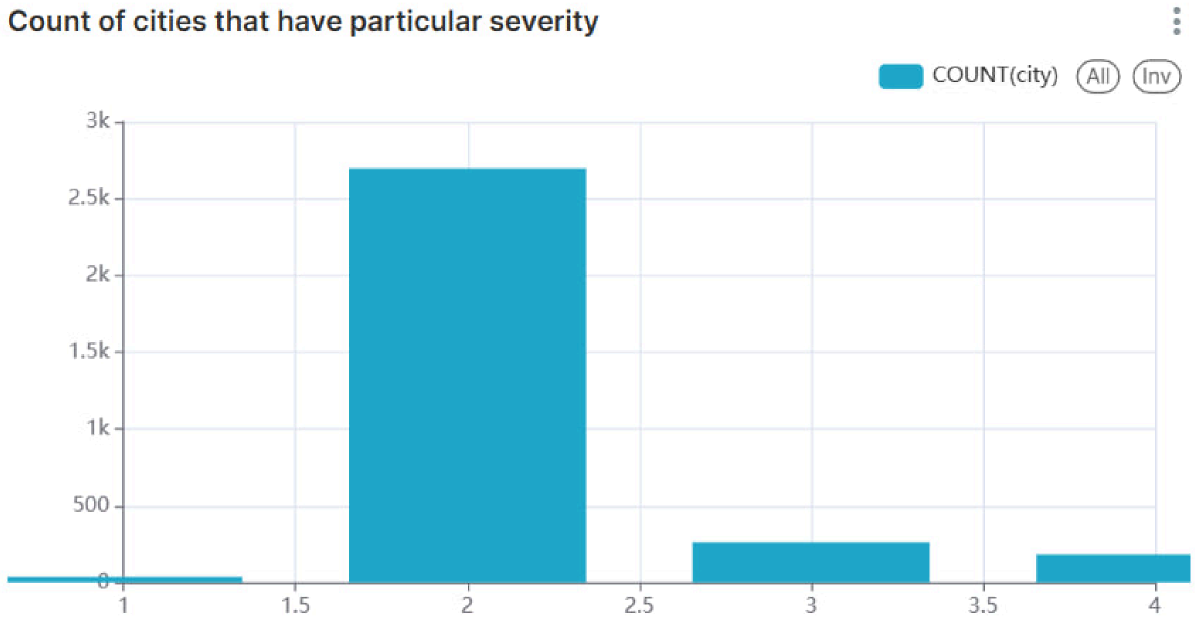
- **Severity 2** demonstrates a consistent **upward trend**, increasing its dominance over time.
- **Severities 3 and 4** show a **declining trajectory**:
  - Severity 3 decreased from **20%** to **5%**
  - Severity 4 dropped from **15%** to **3%**
- **Severity 1** appears only in **2020** (minimal representation)**.**



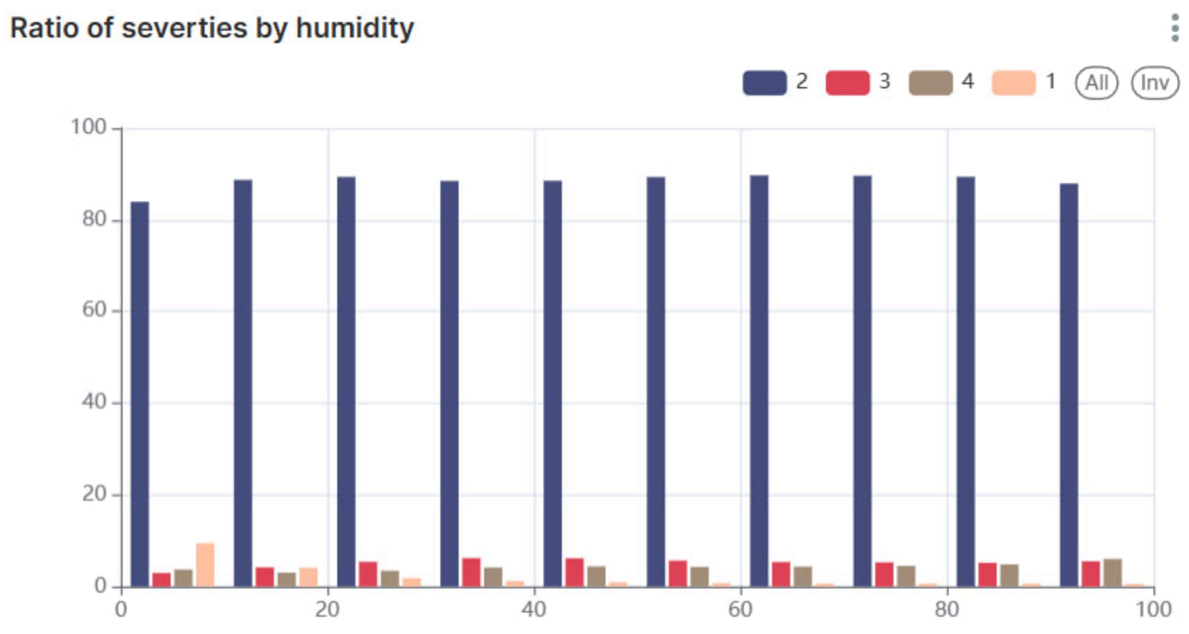
The data reveals a smooth wave-like pattern in the distribution of each severity level, indicating that the likelihood of a particular severity fluctuates predictably throughout the day.



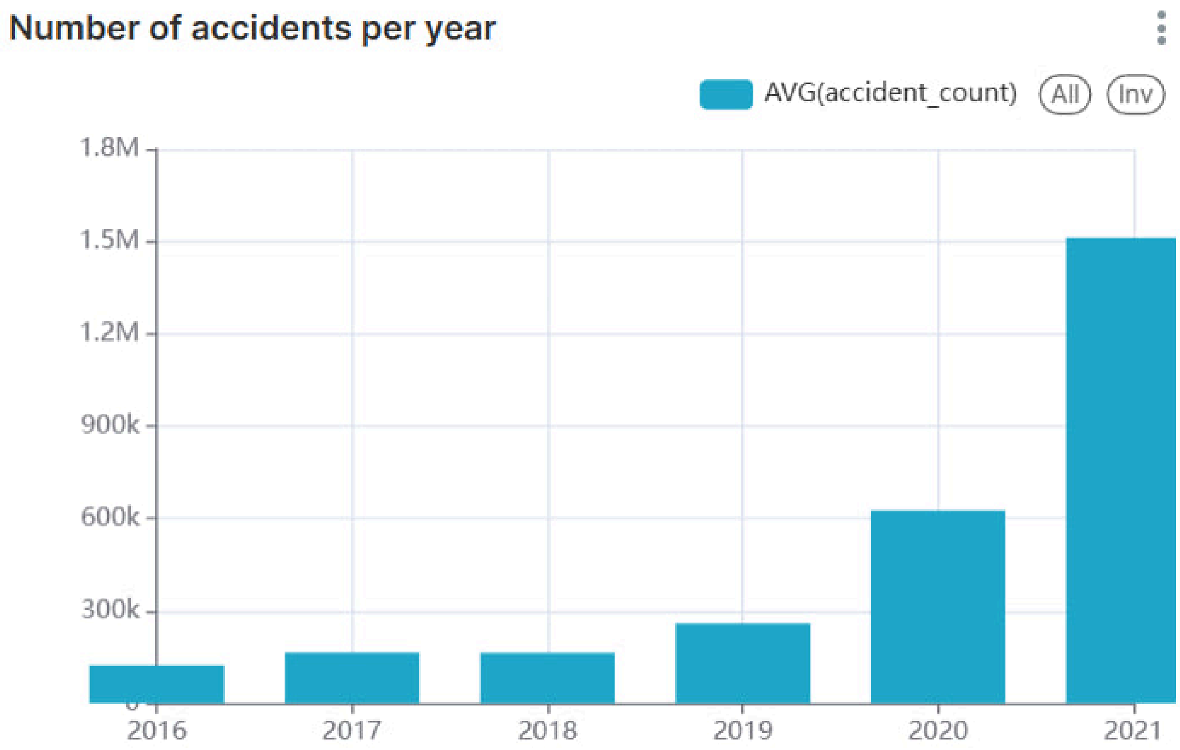
This chart shows that each city has a distinct distribution of severity counts. When linked to the state map, it reveals an interesting trend: even if a city has the highest severity count, its state may not rank as the largest overall in terms of total severity incidents.



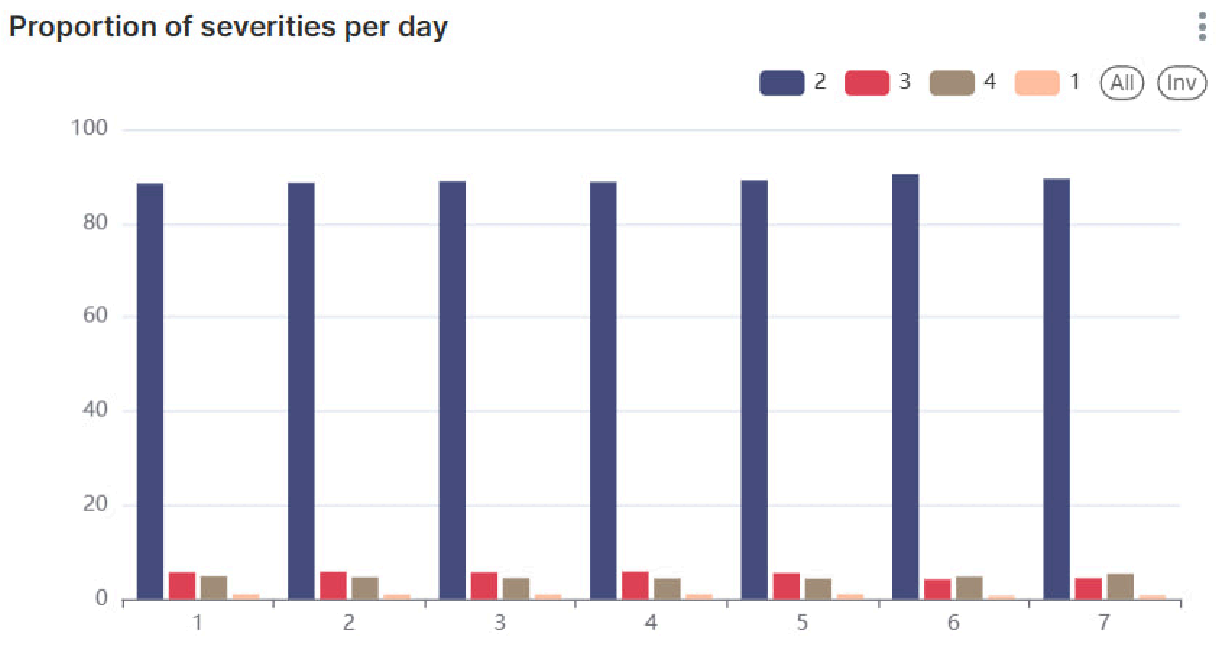
The data shows an uneven distribution of incident severities across cities. Notably, some locations report only a subset of severity levels, indicating that not all cities experience the full range of possible incidents.



The data indicates that Severity Type 1 incidents occur more frequently under lower humidity conditions.



The data reveals a concerning upward trend in the annual number of incidents, with counts increasing each year.



The daily distribution of incidents remains relatively consistent overall. However, a notable deviation occurs on weekends, where Level 4 incidents surpass Level 3 in severity frequency. This weekend pattern suggests distinct incident dynamics that warrant further investigation.

Analysis results

1. While most accidents are moderate (Severity 2), the near-equal proportions of high-severity cases (3 and 4) suggest non-trivial risks requiring targeted mitigation, especially given their disproportionate societal impact.
2. Trend suggests higher accident risk during weekdays, likely due to increased traffic volume from work commutes and commercial activity. The gradual rise toward Friday may reflect cumulative fatigue or heightened traffic before the weekend.
3. Patterns highlight the influence of commuting schedules on accident occurrence, with weekday risks concentrated around rush hours and weekend risks aligned with daytime activity.
4. Moderate temperatures correlate with lower-severity accidents, while extreme cold (<0°C) disproportionately increases high-severity (Severity 4) risks.
5. The substantial reduction in high-severity accidents (3 & 4) may reflect improvements in vehicle safety standards, road infrastructure, or emergency response effectiveness.

ML modeling

Feature extraction and data preprocessing

1. **Numerical**

   Most of the data is our dataset is numerical and the our pipeline for preprocessing numerical features if following:
   - Data imputing based on the mean of current state
   - Data Scaling
   - Vector Assembling
   - Drop features that was copied and transformed
2. **Categorical**

   Categorical features are harder to input. For side, county, state, and Weather_Condition we can use simple mode. But to impute the city we need to track the state. So for city imputing we take the state mode. For encoding either String Indexing + One-Hot either frequency encoding was applied.

3. **Geospatial**

   We preprocess geographic features by converting latitude, longitude, and altitude from geographic (WGS-84) coordinates to Earth-Centered, Earth-Fixed (ECEF) coordinates using a well-known geodetic formula. A UDF is used to apply the conversion row-wise, returning ECEF x, y, z values as new columns in the dataset.

4. **Timestamp**

   Firstly, we extract data from timestamps by separating them into day of week, hour, lit, minute, month, second, and year.

   Secondly, to encode timestamp features we use Cyclic Encoding. This technique separates each time or date feature to the sin and cos term.

Training and fine-tuning

We decided to train Logistic Regression and Decision Tree, since our task is multiclass classification. Logistic Regression gives good estimates of probabilities and works well with straight-line relationships. Decision Trees naturally divide the data into different sections, allowing for flexible and understandable predictions for multiple outcomes without needing complicated changes. These models provide a good balance between being easy to understand and being effective, which is important when analyzing factors that contribute to car accident severity.

Both models were fine-tuned with cross validation on 3 folds. Tuned parameters are the follows:

- For Logistic Regression: regParam & elasticNetParam
- For Decision Tree: maxDepth & impurity

Evaluation

For grid search evaluation we prioritized F1 score. However, the F1 score itself will not show us a picture for all classes. So also we collected accuracy for the whole test and recall and precision for each class separately.

Data presentation

The description of the dashboard

The dashboard is organized into three core sections: Data Description, Data Insights, and ML Modeling.

Data Description

This section provides essential background about the dataset, beginning with a brief introduction to its scope and purpose along with a direct access link. A structured table outlines the dataset's schema, including column names, data types, and total record count. A representative data sample is also included to showcase the dataset's actual content and formatting.

Data Insights

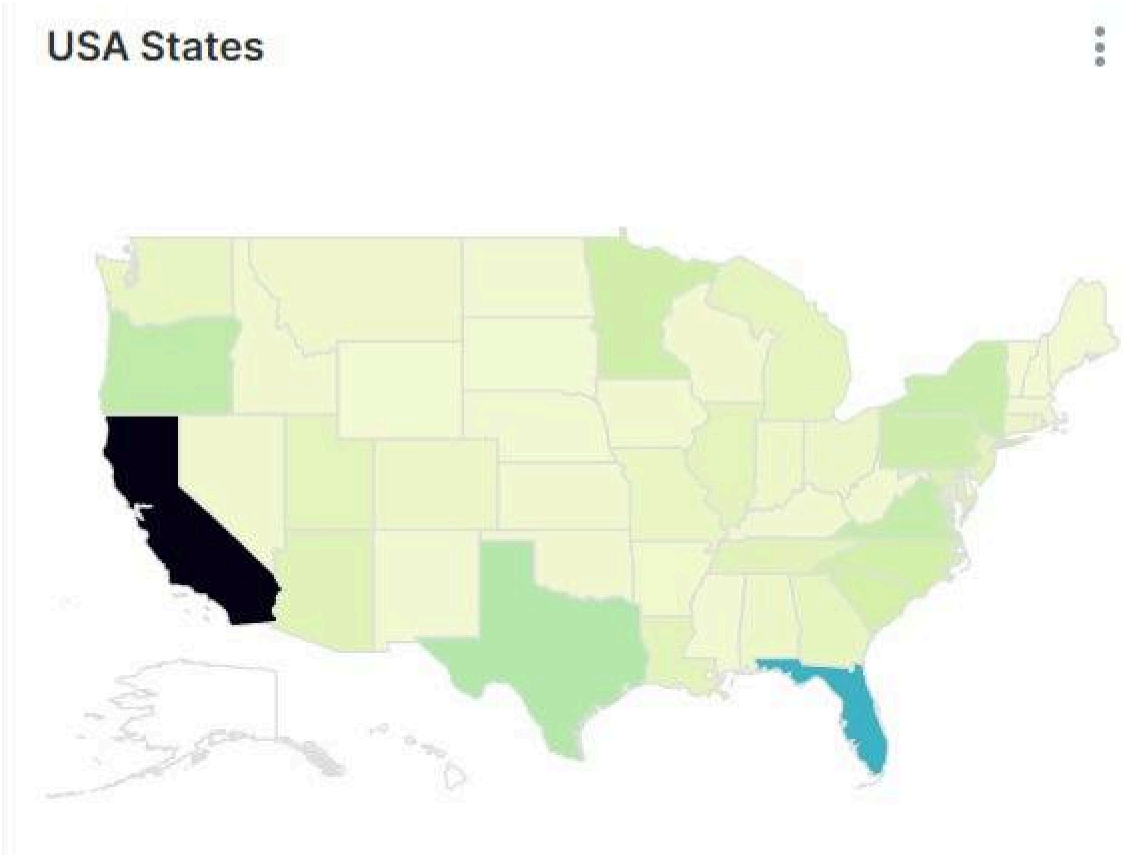
This section presents key findings from Exploratory Data Analysis (EDA) through 12 visualizations. Each chart examines relationships between the target variable and different features, revealing critical patterns and dependencies in the data.

ML Modeling

The modeling section details the data preparation process, covering the treatment of numeric, categorical, and spatio-temporal features, as well as null value handling and feature preprocessing. It evaluates the performance of logistic regression and decision tree models during training, concluding with key metrics and final observations on model effectiveness..

Description of each chart

Each chart of the data presentation tab is the same as we make in the Data Analysis stage.
The only additional chart that was added is



The choropleth map illustrates significant regional disparities in accident frequency:

- **California** (black) dominates with the highest accident count, reflecting its dense population and heavy traffic volume.
- **Florida** (blue) and **Texas** (green) follow as high-risk states, consistent with major urban centers and extensive highway networks.
- **Oregon** (green) shows moderately elevated counts, likely due to key transport corridors.
- Remaining states appear in **light green shades**, with no standout hotspots beyond the aforementioned regions.
- Accident prevalence strongly correlates with population density and transportation infrastructure, with coastal and southern states exhibiting the highest risk.

Our findings

The analysis reveals several critical patterns in incident data:

1. **Temporal Trends**

   Incidents peak during early mornings and evenings on weekdays, likely tied to commuting patterns.

   A distinct weekend dynamic emerges: Level 4 severity incidents disproportionately increase compared to Level 3, suggesting altered risk factors during non-workdays.

A concerning upward trend in annual incident counts highlights the need for proactive interventions.

2. **Geographic Distribution**
   Incidents peak during early mornings and evenings on weekdays, likely tied to commuting patterns.

   A distinct weekend dynamic emerges: Level 4 severity incidents disproportionately increase compared to Level 3, suggesting altered risk factors during non-workdays.
   A concerning upward trend in annual incident counts highlights the need for proactive interventions.

   Incident severities are unevenly distributed across cities, with many locations experiencing only a subset of severity levels. Notably, a city with high severity counts does not necessarily correlate with its state having the highest overall severity burden.

3. **Data Anomalies**
   Level 1 severity incidents appear exclusively in 2020, contributing to its underrepresentation in the dataset.

   Severity distributions across temperature levels are remarkably consistent, with only minor deviations for Level 1.

4. **Patterns in Severity**
   Hourly severity distributions follow smooth wave-like patterns, indicating predictable fluctuations in risk throughout the day.

   The dataset exhibits significant class imbalance in crash severities, a critical consideration for future modeling efforts.

Conclusion

We present a machine learning project aimed at predicting the severity of car accidents using data such as location, time, weather, and road conditions. The dataset spans from 2016 to 2023 and contains over 2.8 million accident records across 49 U.S. states. Exploratory data analysis revealed patterns such as higher accident frequency on weekdays, during rush hours, and in certain regions like California and Florida. The data showed significant class imbalance, with Severity 2 (moderate) accidents dominating the dataset. Various preprocessing techniques were applied, including handling missing values, feature encoding, geospatial transformation, and temporal feature extraction. Logistic Regression and Decision Tree models were trained and evaluated using cross-validation and performance metrics like F1 score, accuracy, precision, and recall. A dashboard was developed to visualize data insights and model results, supporting data-driven decisions for improving road safety.

Reflections

Challenges and difficulties

1. Data Type Incompatibility: There was an issue with Parquet and Hive data type mismatches, causing integration problems and additional overhead during type conversion.
2. Resource Constraints on the Hadoop Cluster: The Hadoop cluster occasionally experienced failures due to resource limitations.
3. Superset Limitations: Superset's sorting functionality is limited to alphabetical order, preventing us from displaying weekday labels in the correct chronological order in one of our visualizations.
4. Lack of Training Progress Feedback in Spark: Spark does not have a built-in progress bar or time estimator for model training, making it challenging to monitor progress or estimate completion times.

Recommendations

1. Use of Avro for Improved Compatibility: Switching to the Avro format could enhance compatibility with Hive and reduce time spent on type conversions.
2. Test-First Development: Writing unit tests before development could help minimize debugging time and improve code reliability.

The table of contributions of each team member

Project Tasks	**Task description**	**Dmitry Dydalin**	**Amir Bikineyev**	**Dzhavid Sadreddinov**	**Nikita Borisov**	**deliverables**	**Average hours spent**
Data ingestion and preparation	Collect the data, cleaned column names, uploaded to postgres	100%	0%	0%	0%	data.csv Postgres table accidents	5
Creating the hive table	Create a hive table out of the parquet file stored in hdfs	0%	0%	0%	100%	Partitioned and Bucketed Hive table, HQL script for table creation	5
Data preprocessing & encoding	Working with spark dataset, handling the null values and encoding the features	50%	0%	0%	50%	test.csv train.csv	1.5
Model training	Train the models, optimize hyperparameters, evaluate and compare the best versions	50%	0%	0%	50%	model1 model2 predictions evaluation.csv	7
Report & presentation	Write and format report and slides	25%	25%	25%	25%	report.pdf slides.pptx	2
Dashboard Creation	Aggregate results is one board	0%	66%	33%	0%	Publicly available Dashboard in Superset	4.5
EDA	Extract Data Insights	0%	50%	50%	0%	qx.csv Charts with data insights	7