OM 420 LEC B02 - Winter 2024

Project Presentation

Explanatory Data Analysis & Prediction models

Sheikh Safwan Hossain &
Mohammad Shahriar Hossain

Table of contents

01

Exploratory Data Analysis (EDA)

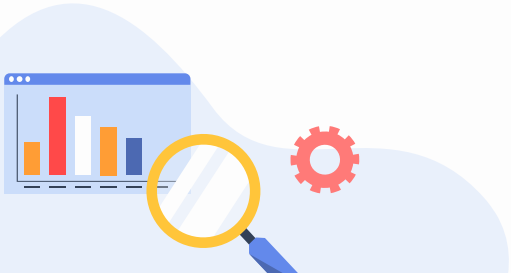
Table: Counts of Vehicle Types

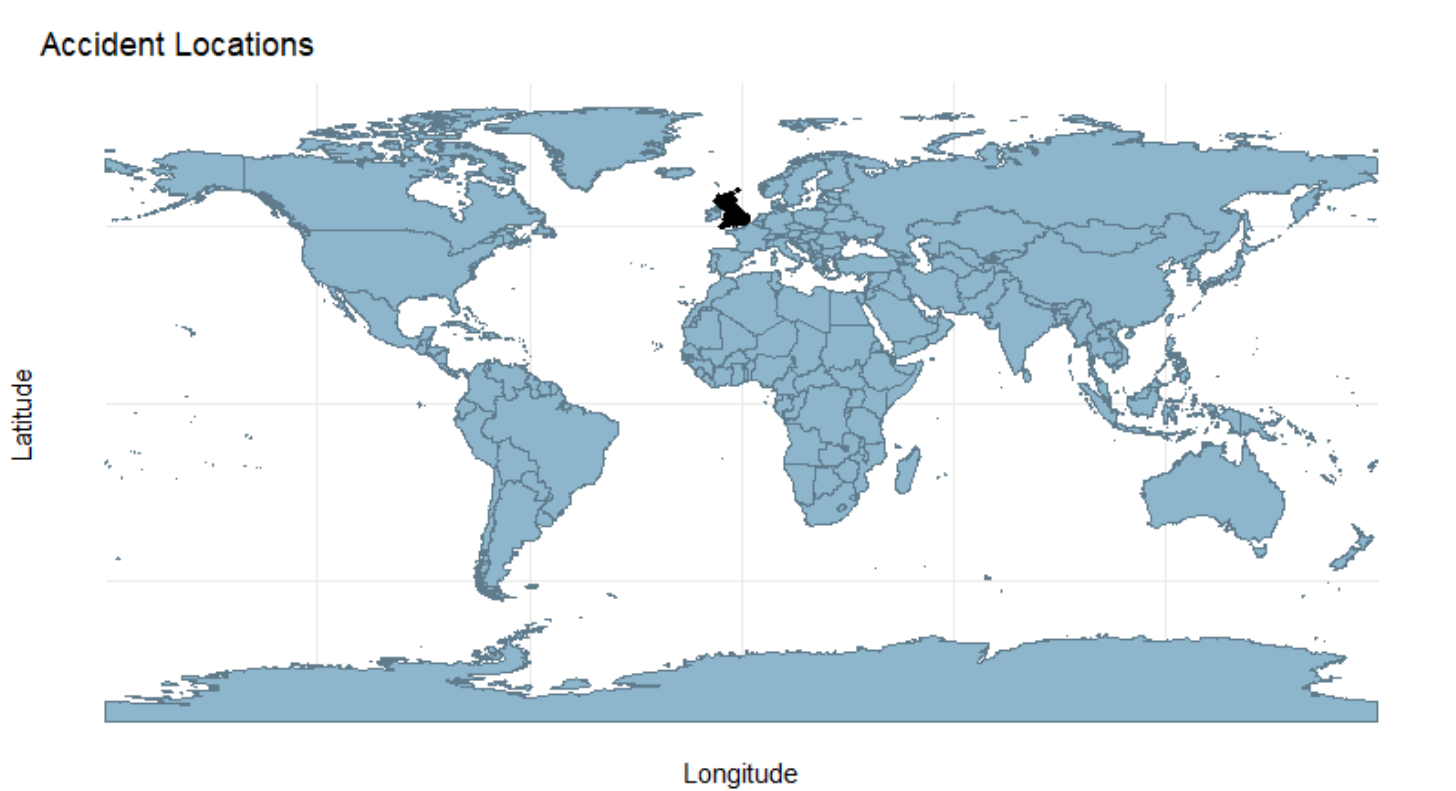
Vehicle Type	Count
PCycle	17437
MC50	1489
MC125	8053
MC500	2119
MC500+	5228
Taxi	4696
Car	152686
Minibus	405
Bus	3902
R.Horse	65
Agri	423
Tram	26
Van	12579
GV7.5	982
GV7.5+	3357
MScooter	250
EMCycle	65
Other	1005
MC-Unknown	415
GV-Unknown	890
Missing	309

- Breakdown of vehicle types in road accidents.
- Cars most frequently involved; followed by taxis, buses.
- Horse riders, trams least involved — lower counts.
- Data informs road safety measures and interventions.

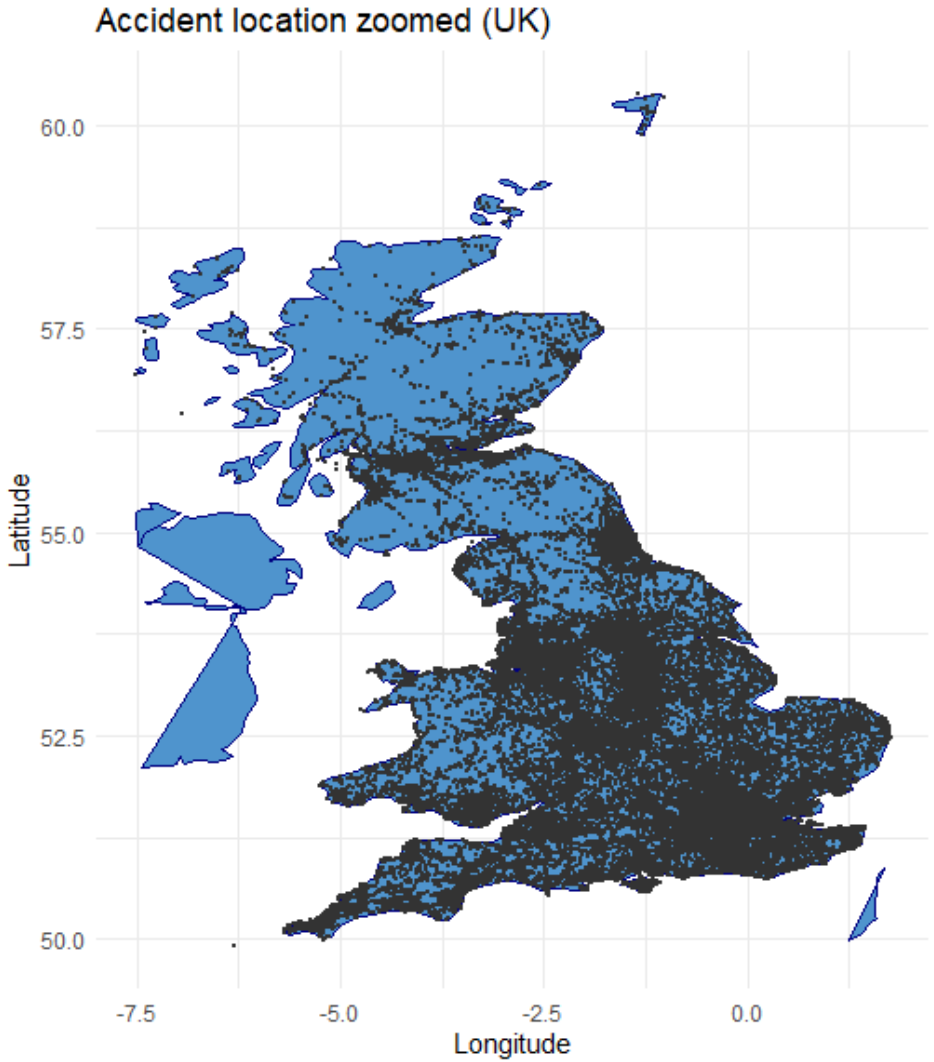
- Severity of road accidents by vehicle type.
- Cars most involved in fatal, serious, and slight injuries.
- High-capacity motorcycles show increased fatal accident rates.
- Highlights need for vehicle-specific safety interventions.

	Severity		
Type	Fatal	Serious	Slight
PCycle	117	3945	13375
MC50	9	380	1100
MC125	56	2125	5872
MC500	40	679	1400
MC500+	279	2293	2656
Taxi	26	680	3990
Car	1833	25692	125161
Minibus	5	71	329
Bus	71	696	3135
R.Horse	0	22	43
Agri	21	137	265
Tram	2	8	16
Van	197	2316	10066
GV7.5	27	182	773
GV7.5+	223	695	2439
MScooter	12	49	189
EMCycle	0	18	47
Other	28	285	692
MC-Unknown	5	156	254
GV-Unknown	18	159	713
Missing	0	50	259

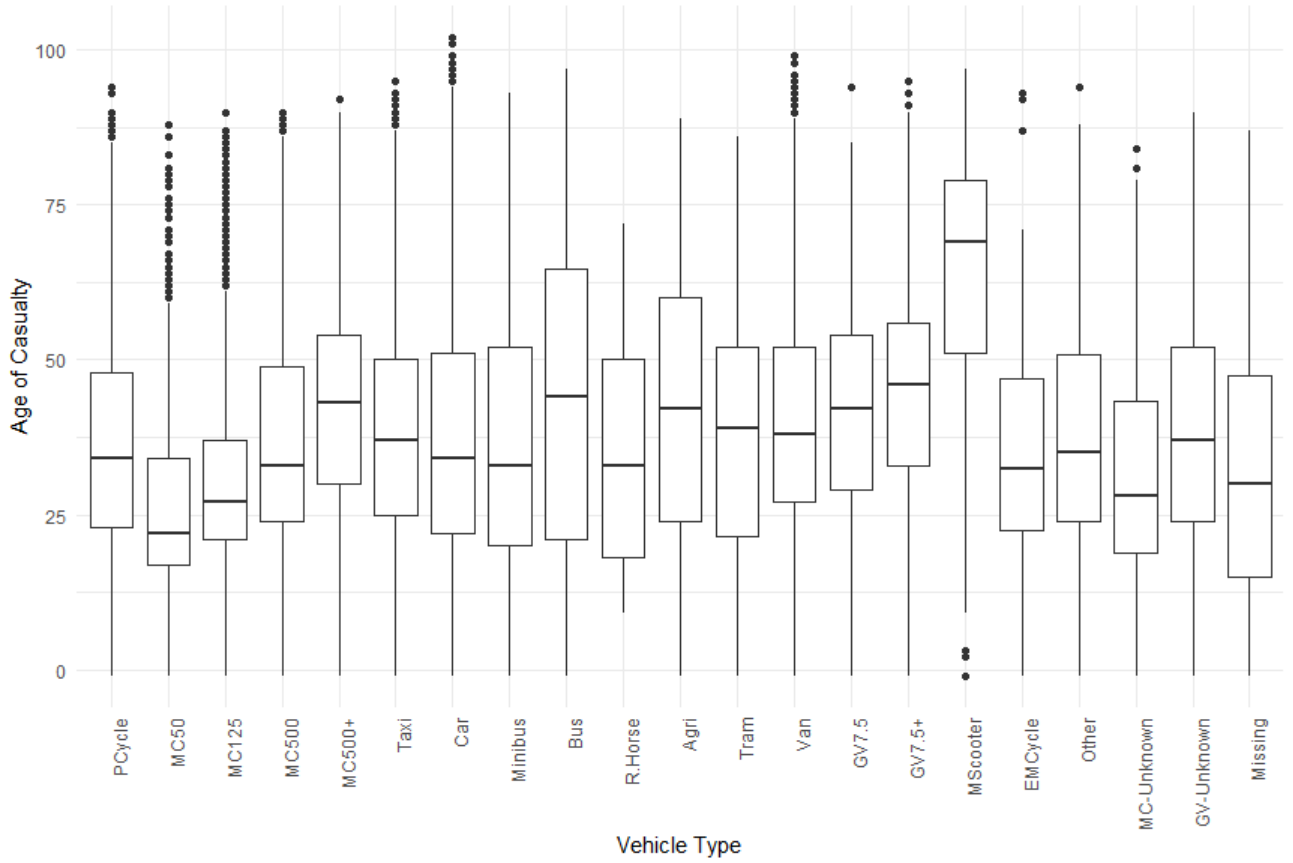




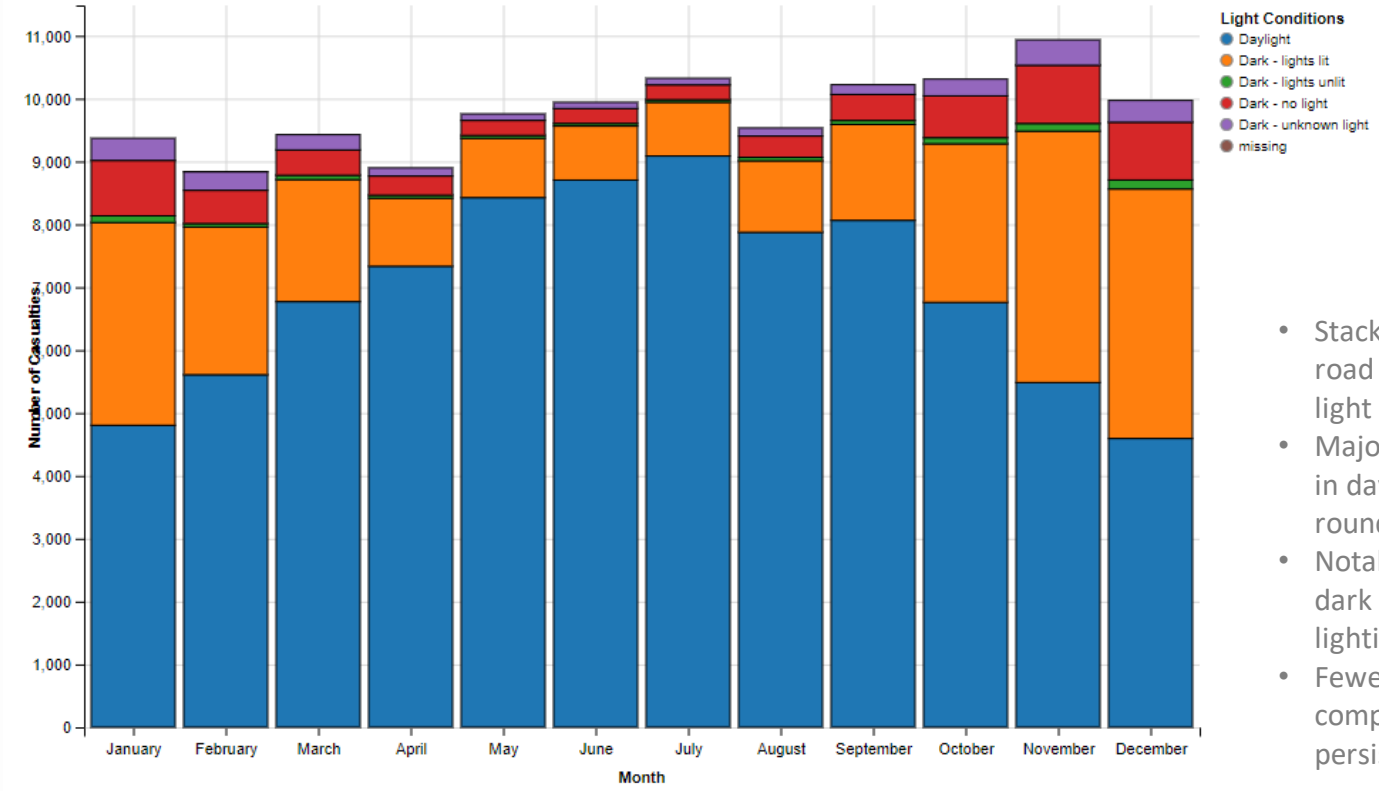
This world map displays the geographic distribution of road accident locations, with a concentrated cluster of accidents depicted in a specific region. The visualization highlights spatial patterns in accident occurrences, which could be crucial for regional safety assessments and the development of location-specific traffic safety strategies.



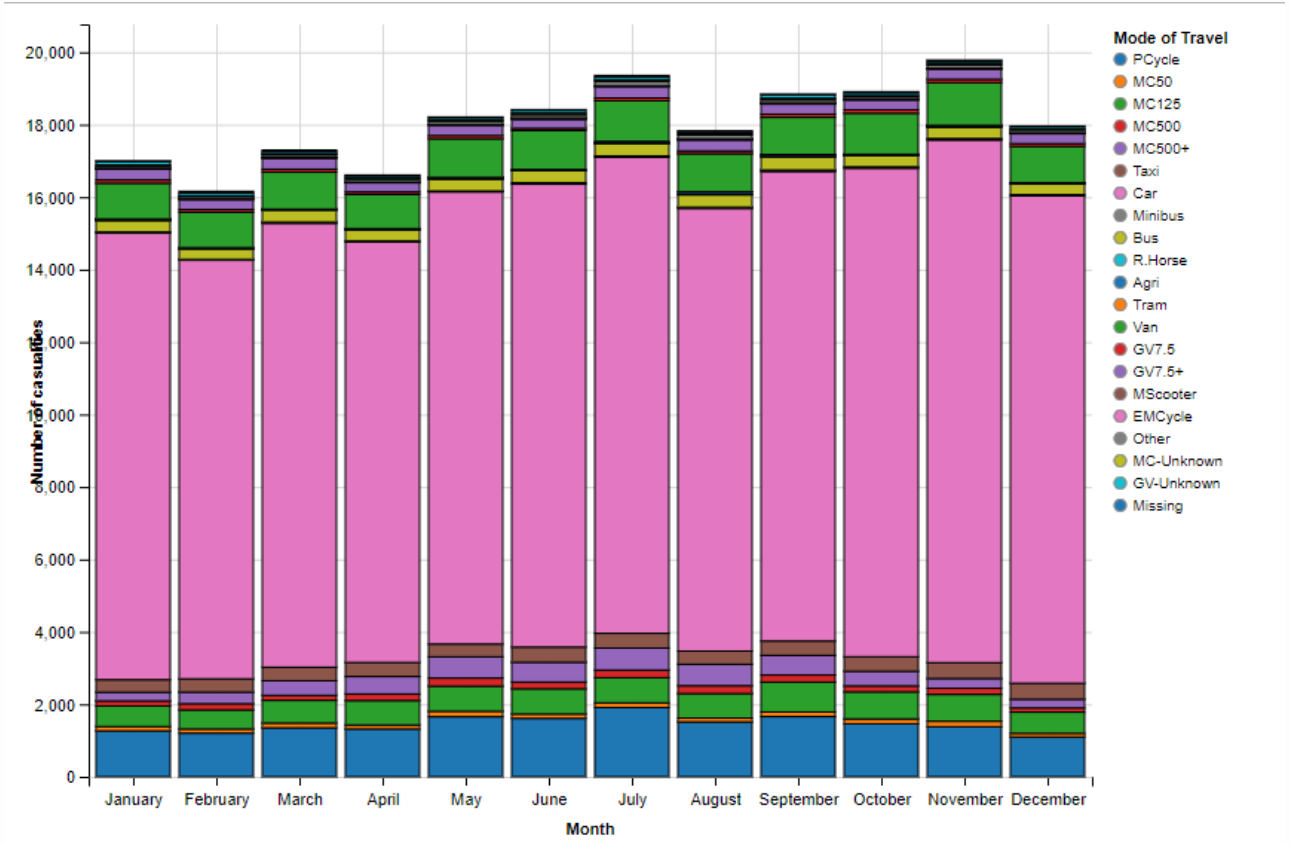
- Detailed map of UK road accident locations.
- Dense clusters indicate accident hotspots.
- Can inform local traffic safety and planning.
- Aids in addressing high-risk areas.



The boxplot displays age distribution of road accident casualties by vehicle type, revealing the median, range, and age outliers. Cars show a wide casualty age range, while motorcycles predominantly affect younger individuals. This aids in directing safety campaigns towards the most vulnerable age groups per vehicle type.



- Stacked bar chart of road casualties by light conditions.
- Majority of accidents in daylight, year-round.
- Notable accidents in dark with artificial lighting.
- Fewer accidents in complete darkness, a persistent issue.



Bar chart shows road casualties peak in summer, particularly for cars and motorcycles, suggesting seasonal influences on travel behavior.

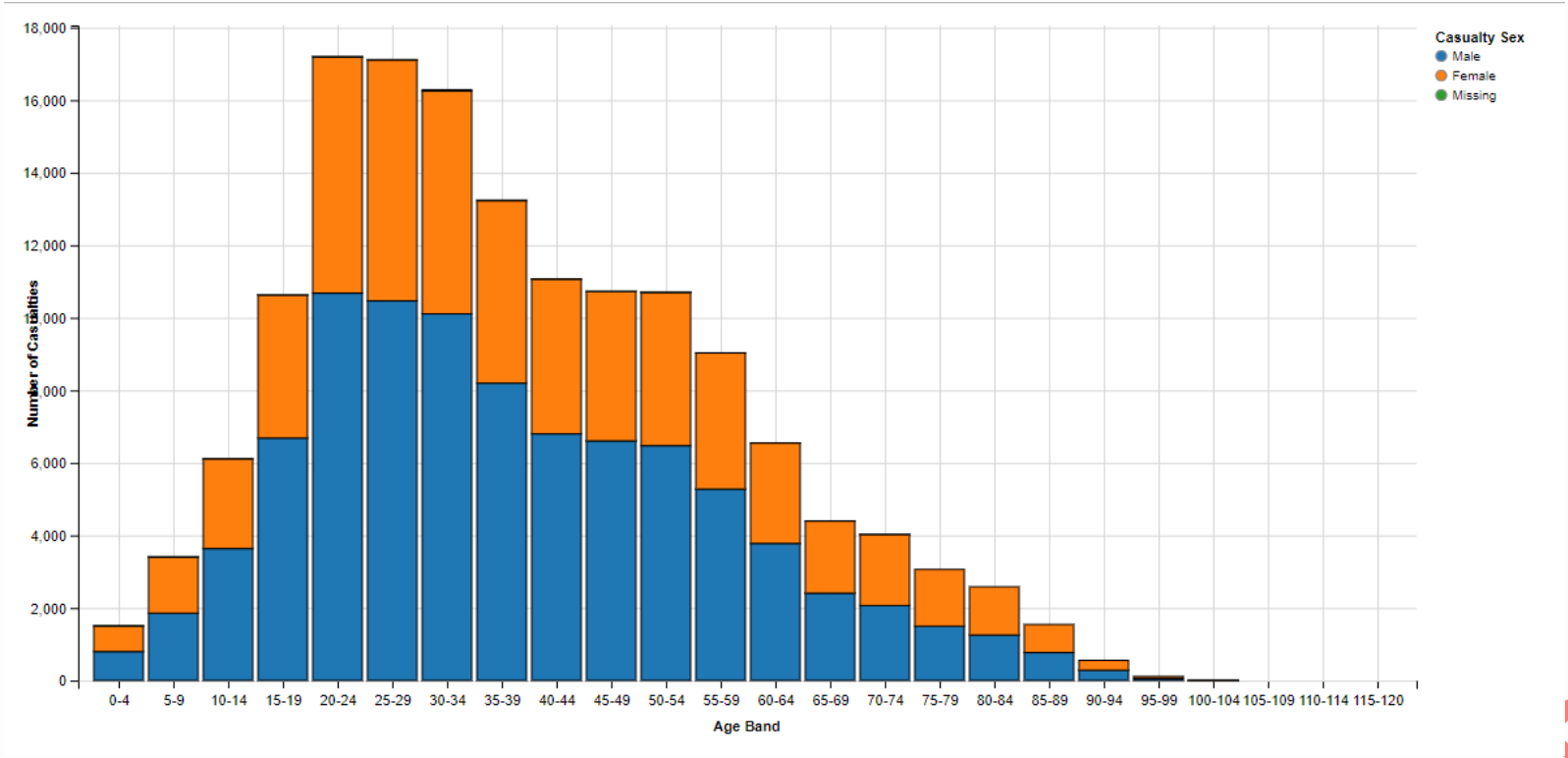
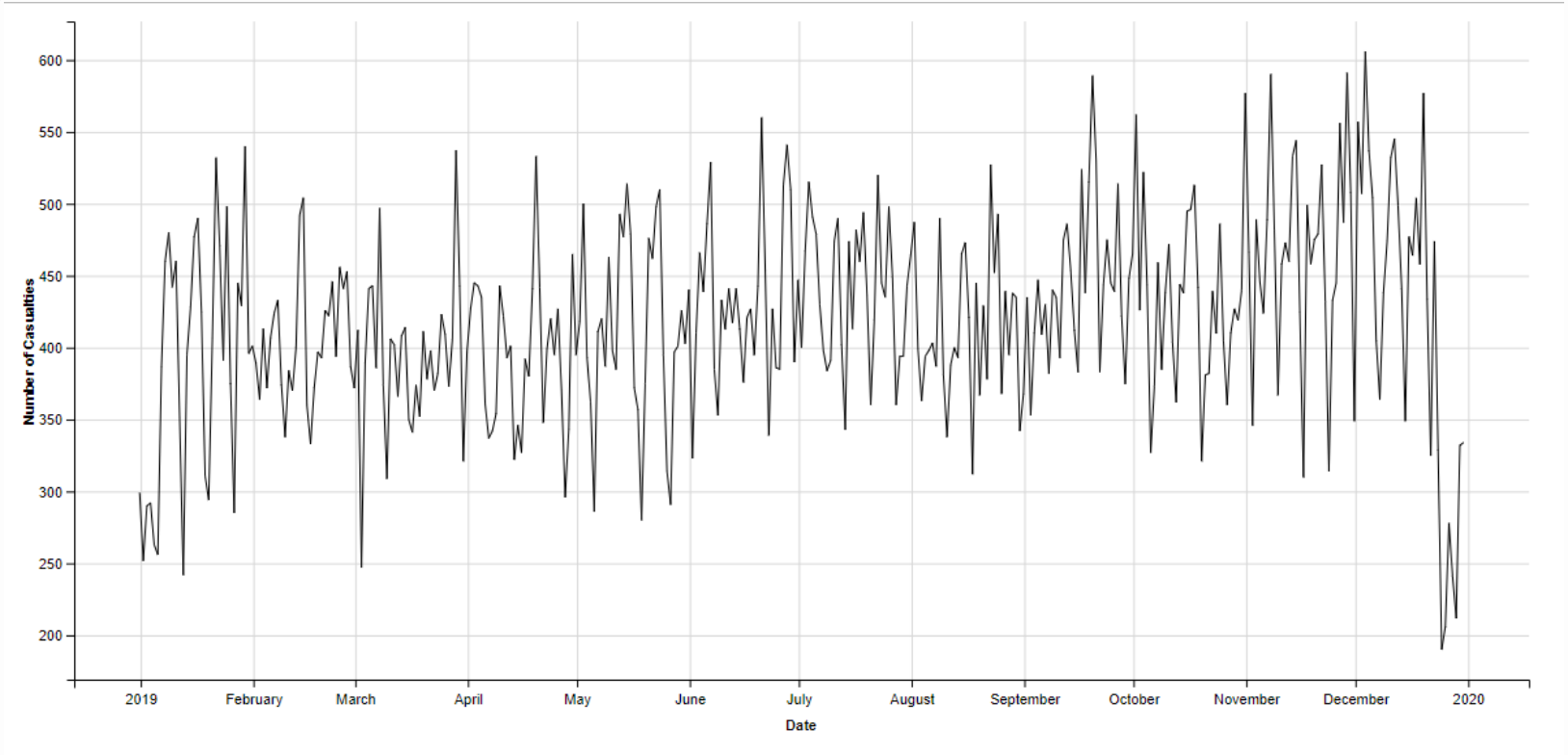


Chart reveals more males are casualties in road accidents, especially ages 20 to 44, highlighting a key demographic for safety interventions.



Line graph indicates a seasonal trend in road casualties, with higher numbers in summer and lower in winter, reflecting potential changes in driving habits and conditions.

Some interesting findings!

01

Prevalence in Passenger Vehicles

Cars are the leading contributors to road accidents and casualties, indicating a significant area of focus for traffic safety improvements and driver education programs.

02

Illumination and Incidents

Daylight sees the highest occurrence of road accidents, suggesting that despite good visibility, other factors such as increased traffic volume during daylight hours may influence accident rates.

03

Grading the Gravitas

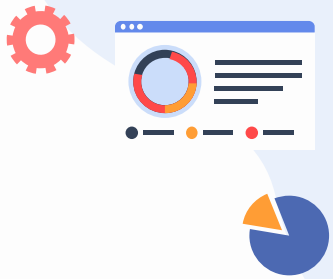
The majority of accident outcomes are classified as slight, with a smaller yet noteworthy proportion resulting in serious injuries, and fatal outcomes being the least common yet most critical concern.

02

Variables and Data Preparation

Data Preparation for Predictive Modeling

1. Target Variable Creation
2. Streamlining Data
3. Data Cleansing
4. Data Integration
5. Handling NA Values Post-Join
6. Categorization of Variables
7. Reproducibility and Sampling
8. Dataset Division for Model Assessment



1 Target Variable Creation

A binary 'Severity_Target' was created to distinguish between accidents of high severity (fatal or serious) and others, facilitating a more focused analysis on severe accidents. Streamlining Data:

2 Streamlining Data

Unnecessary columns were removed to streamline datasets, ensuring that only the most relevant information is retained for model building.

3 Data Cleansing

Rows with missing values (NA) or placeholder values (-1) were removed to enhance data quality, which is crucial for the accuracy of predictive models.

4 Data Integration

The cleaned accident, vehicle, and casualty datasets were merged using 'Accident_Index' as a key, creating a comprehensive dataset for analysis.

5 Handling NA Values Post-Join

NA rows that emerged after joining datasets were omitted to maintain the integrity of the dataset.

6 Categorization of Variables

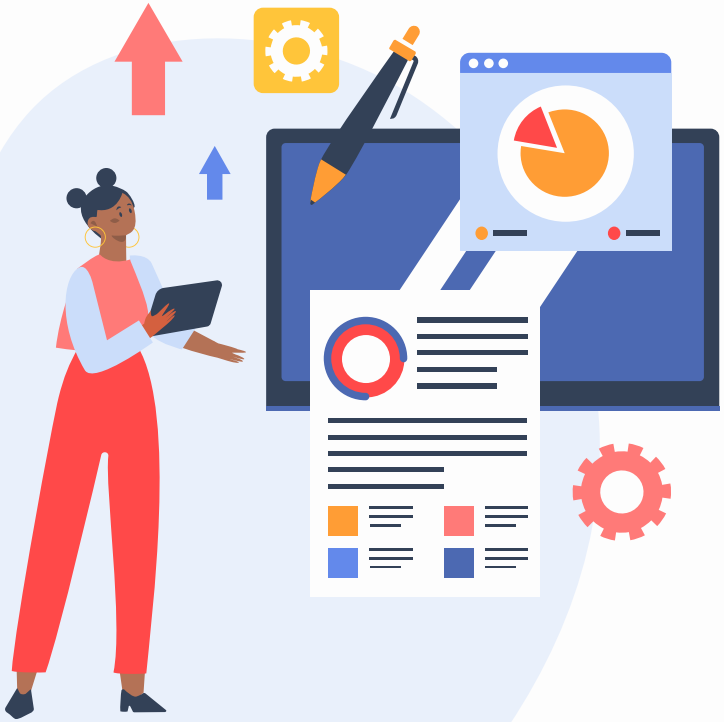
Selected columns were converted into factors to accurately represent categorical data within the modeling process.

7 Reproducibility and Sampling

A random seed was set for reproducibility, and the dataset was split into training, validation, and testing subsets to evaluate model performance.

8 Dataset Division for Model Assessment

The data was proportionately split into 70% training, 15% validation, and 15% testing sets. This separation is crucial for not only building the model but also for fine-tuning and assessing its generalizability to new, unseen data.



03

Prediction Model Development

K-Nearest Neighbors (KNN)

Applied KNN using two sets of 9 predictors from 18 total, to leverage diverse aspects of the data in predicting accident severity.

Linear Discriminant Analysis (LDA)

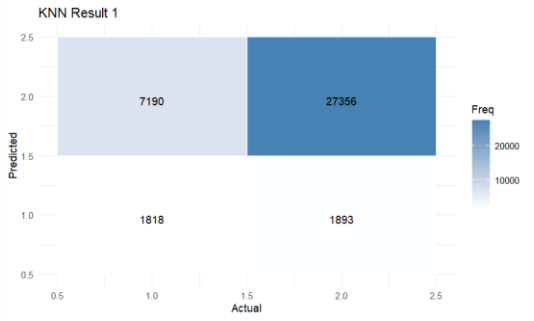
Conducted LDA with different combinations of 9 predictors, aiming to find the linear combination that best separates high severity accidents from others.

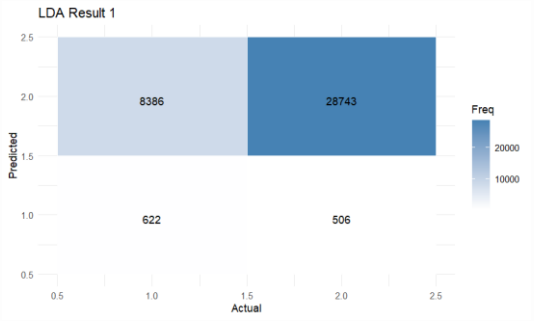
Decision Trees

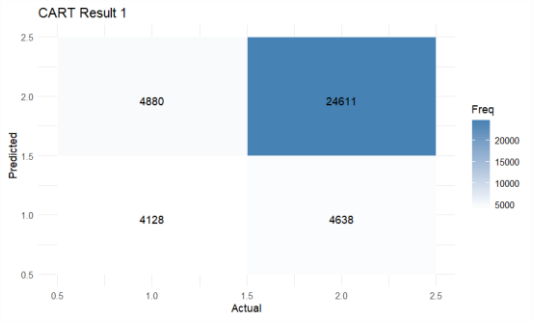
Developed decision tree models using distinct sets of 9 predictors, providing a transparent and interpretable approach to classify the severity of road accidents.

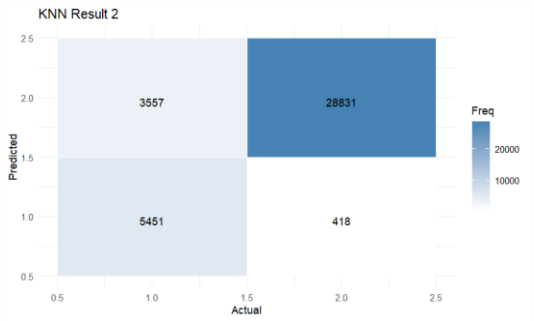
04

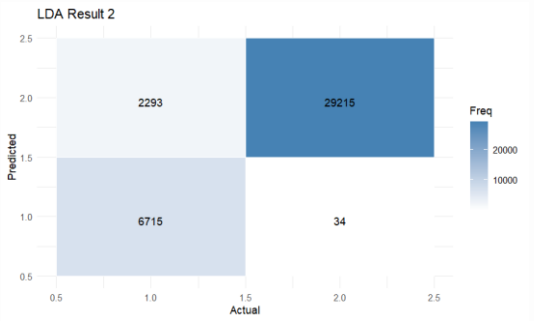
Model Comparison and Evaluation

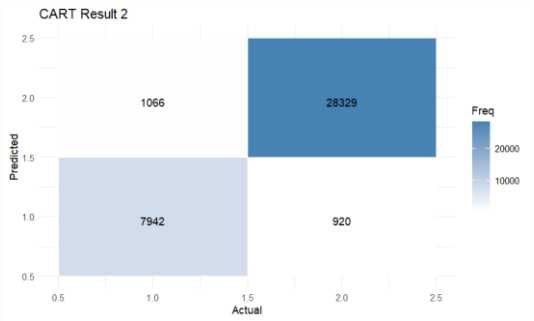

KNN Result 1
7190
27356
1818
1893
Predicted
Actual
Freq
20000
10000

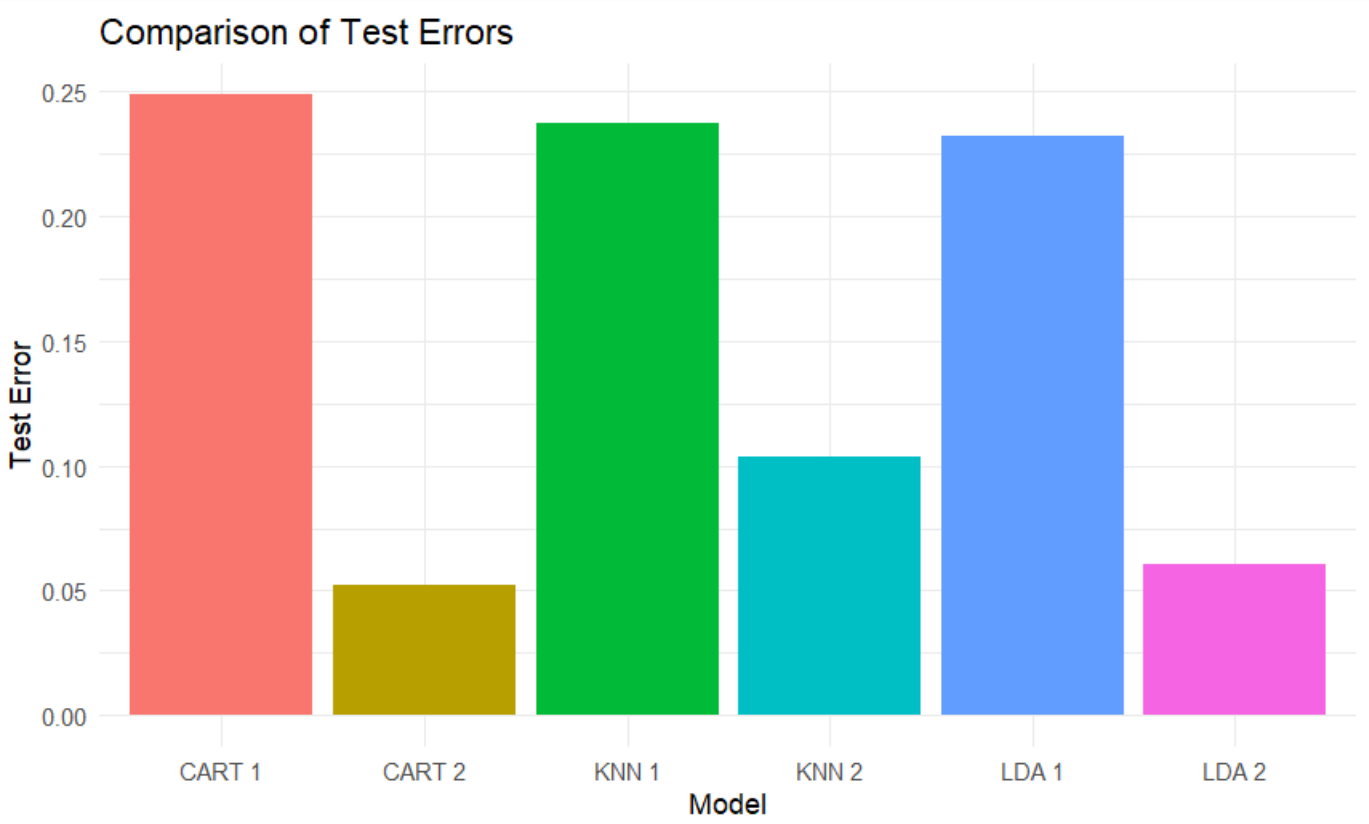

LDA Result 1
8386
28743
622
506
Predicted
Actual
Freq
20000
10000


CART Result 1
4880
24611
4128
4638
Predicted
Actual
Freq
20000
15000
10000
5000


KNN Result 2
3557
28831
5451
418
Predicted
Actual
Freq
20000
10000


LDA Result 2
2293
29215
6715
34
Predicted
Actual
Freq
20000
10000


CART Result 2
1066
28329
7942
920
Predicted
Actual
Freq
20000
10000


Comparison of Test Errors
Test Error
0.25
0.20
0.15
0.10
0.05
0.00
CART 1
CART 2
KNN 1
KNN 2
LDA 1
LDA 2
Model

Optimal Model Selection: Efficacy Proven by Error Rate

- Rigorous error rate comparisons guide our model selection process.
- Both LDA models outperform their KNN and CART counterparts, with LDA Model 2 showing the lowest error rate overall.
- LDA Model 1 also presents a lower first error rate compared to the corresponding KNN and CART models.
- Based on the compelling low error rates, LDA is chosen for its superior accuracy in predicting accident severity.

Thanks!

Do you have any questions?