# 1. Introduction

GoBest Cab has commissioned our team to create a portable machine learning application for analysing cab data, focusing on driver safety behaviour. This tool will facilitate in-depth analysis through an interactive dashboard, providing insights into driving patterns for enhanced operational efficiency and customer safety. This report outlines the application's development and the strategic implications of our data-driven findings.

## 2. Business Understanding/ Background Research & Data Understanding

GoBest Cab operates in the competitive ride-hailing market where customer safety and satisfaction are paramount. Recent trends in the industry highlight the importance of leveraging data to improve service quality and ensure passenger safety. Our team's research suggests that advanced analytics can predict and mitigate risks, leading to safer and more reliable services. To this end, we have analysed GoBest Cab's extensive historical sensor data encompassing around 7 million cab journeys, alongside detailed profiles of 500 cab drivers, to decode patterns in driving behaviour.

The sensor datasets provide granular details on the dynamics of each journey, including location accuracy, directional bearing, multi-dimensional acceleration, gyroscope readings, time intervals, and speed.

The drivers' dataset offers a personal dimension, featuring identifiers, experience levels, gender, vehicle details, and performance ratings.

The safety status dataset adds a critical evaluative layer, classifying trips as safe or dangerous based on a binary label. This rich tapestry of data forms the backbone of our analysis, enabling us to construct a nuanced picture of driving behaviours and safety standards within GoBest Cab's operations.

# 3. Database Operations

## 3.1 Data Ingestion

The data was ingested from CSV files into the corresponding tables using the ‘BULK INSERT’ command. Separate CSV files for the sensor data represented different partitions or subsets of the data, which were loaded sequentially.

## 3.2 Driver Table

**Purpose:** To store information about the drivers

**Fields:** Includes ‘driver\_id’, ‘driver\_name’, ‘date\_of\_birth’, ‘driving experience’, ‘gender’, ‘car\_brand’, ‘car\_model\_year’, and ‘driver\_rating’

**Primary Key:** ‘driver\_id’

## 3.3 Safety Status Table

**Purpose:** To keep records of each booking's safety status

**Fields:** ‘bookingID’, ‘driver\_id’, and safety ‘label’

**Primary Key:** ‘bookingID’

**Foreign Key:** ‘driver\_id’ referencing drivers table

## 3.4 Sensor Table

**Purpose:** To collect sensor data related to each booking

**Fields:** ‘sensor\_id’, ‘bookingID’, various sensor readings like ‘accuracy’, ‘bearing’, ‘acc\_x’, ‘acc\_y’, ‘acc\_z’, ‘gyro\_x’, ‘gyro\_y’, ‘gyro\_Z’, ‘sec’, and ‘speed’

**Primary Key:** ‘sensor\_id’

**Foreign Key:** ‘bookingID’ referencing ‘safety\_status’ table

A white pen with green and black text

Description automatically generated

## 3.5 SQL Query 1 (Driver Safety Performance Analysis)

This SQL query aims to analyse driver safety by combining various metrics like sensor data from trips, driver ratings, experience, and age. It provides a multi-faceted view of driver performance and safety.

**Breakdown of query**

**TripAverages Subquery:** This part calculates average values of various sensor readings (like acceleration, gyroscope data, speed, etc.) for each trip (bookingID) where the safety status is labelled as '1' (which might indicate trips considered for safety evaluation)

**danger\_metric Subquery:** It ranks trips based on a composite 'danger level' calculated from normalized differences between the trip's average sensor readings and the minimum and maximum values of those readings across all trips. This normalization scales the values between 0 and 1, and the sum of these normalized values gives a sense of how 'dangerous' the trip is compared to others.

**avg\_danger\_metric\_per\_driver Subquery:** This calculates the average danger level for each driver by averaging the danger levels of all trips they have made.

**rating\_cat Subquery:** It categorizes drivers based on their ratings into various groups like 'Perfect rating', 'Very high rating', etc.

**dangerous\_rate Subquery:** This calculates the percentage of trips for each driver that are marked as 'dangerous' (where ss.label is presumably a flag indicating danger).

**drivers\_age Subquery:** It calculates the current age of each driver based on their date of birth.

**years\_exp\_cat Subquery:** It categorizes drivers into experience categories based on their years of driving experience.

**Final SELECT Statement:** This combines all the above information for each driver, including their name, rating category, experience category, average percentage of dangerous trips, average danger level, and age. The data is grouped by various attributes and ordered by the average danger level in descending order.

## 3.6 SQL Query 2 (Car Brand Safety and Performance Analysis)

This SQL code is designed to analyse various aspects of driver and vehicle performance, focusing on safety, accuracy, and the age category of vehicles, grouped by car brand. The code consists of several subqueries that calculate different metrics, which are then combined in the final SELECT statement.

**Breakdown of query**

**cars\_age Subquery:** Calculates the age of each car by subtracting the car's model year from the current year.

**dangerous\_rate Subquery:** Determines the percentage of trips that are marked as dangerous for each driver. This is done by dividing the number of trips marked as dangerous by the total number of trips and multiplying by 100.

**accuracy\_cat Subquery**: Categorizes each driver's accuracy based on average GPS accuracy readings. Drivers are categorized as 'highly accurate', 'moderately accurate', or 'inaccurate' based on predefined thresholds of GPS accuracy.

**count\_accuracy Subquery:** Counts the number of drivers falling into each accuracy category, grouped by car brand.

**car\_age\_cat Subquery:** Categorizes each car as 'modern', 'classic', or 'vintage' based on its age, determined in the cars\_age subquery.

**count\_car\_age\_cat Subquery:** Counts the number of cars in each age category ('modern', 'classic', 'vintage'), grouped by car brand.

**Final SELECT Statement:** Combines the above metrics for each car brand. It calculates the average age of cars, average net acceleration (using a formula that combines acceleration in three dimensions), and average speed. It also includes the number of drivers in each accuracy category and the average percentage of dangerous driving, along with the count of cars in each age category. These results are grouped by car brand.

## 3.7 SQL Query 3 (Driver Gyroscope Anomaly Analysis)

This SQL code is designed to analyse driver performance based on gyroscope data from their trips, specifically focusing on deviations from average gyroscope readings. It consists of several subqueries and a final SELECT statement to gather and present the data.

**Breakdown of query**

**Declaration of Variables:** Declaring three variables @avg\_gyro\_x, @avg\_gyro\_y, and @avg\_gyro\_z to store average gyroscope readings (x, y, z axes) for each trip.

**Calculating Average Gyroscope Readings:** Calculates the average gyroscope readings for each axis across all trips (bookingID) and stores these averages in the previously declared variables.

**gyro\_anomalies Subquery:** For each trip, it calculates the gyroscope deviation by finding the Euclidean distance (root of sum of squares) between the gyroscope readings for that trip and the average gyroscope readings (stored in variables).

**avg\_gyro\_dev\_per\_trip Subquery:** This calculates the average gyroscope deviation for each trip.

**trip\_count Subquery:** Counts the number of trips made by each driver.

**avg\_gyro\_dev\_per\_driver Subquery:** Averages the gyroscope deviations for each driver across all their trips.

**rating\_cat Subquery:** Categorizes drivers based on their ratings into groups (e.g., 'Perfect rating', 'Very high rating', etc.).

**Final SELECT Statement:** This part of the query combines all the above data. It selects each driver's ID, name, average gyroscope deviation, maximum gyroscope deviation, the number of trips with anomalous gyro readings, the percentage of trips with anomalous gyro readings, and their rating category. A trip's gyroscope reading is considered anomalous if its deviation is more than two standard deviations above the average deviation. This final dataset is ordered by the maximum gyroscope deviation in descending order.

# A diagram of a diagram Description automatically generated4. Data Preparation & Feature Engineering

## 4.1 Data Sources

The Data was extraction from the three tables in the SQL Database: ‘sensor’, ‘safety\_status’, and ‘drivers’

## 4.2 Extraction Process

Data was extracted using SQL queries using SQLAlchemy for database connection management in a Python environment

## 4.3 Data Transformation and Preprocessing

The extracted data underwent several preprocessing steps, including outlier removal, interpolation of missing values, label encoding of categorical variables, and creating of new calculated columns.

Steps Taken

1. **Outlier Removal:** Sensor data was cleaned to remove NaN values and outliers in the ‘sec’ column
2. **Data Interpolation:** Missing sensor values were interpolated linearly
3. **Feature Engineering:** New columns such as ‘hours’ and ‘total\_acceleration’, were created from existing sensor data
4. **Label Encoding:** Driver names were standardized to string data type, categorical variables like ‘gender’ and ‘car\_brand’ were label encoded for uniformity
5. **Age Calculation:** Drivers’ ages were calculated from their birth dates
6. **Safety Data Preprocessing:** Safety data types were converted to Boolean for clearer interpretation

## 4.4 Data Loading

The pre-processed data was saved into .h5 file format as it is more optimized for handling larger volumes data.

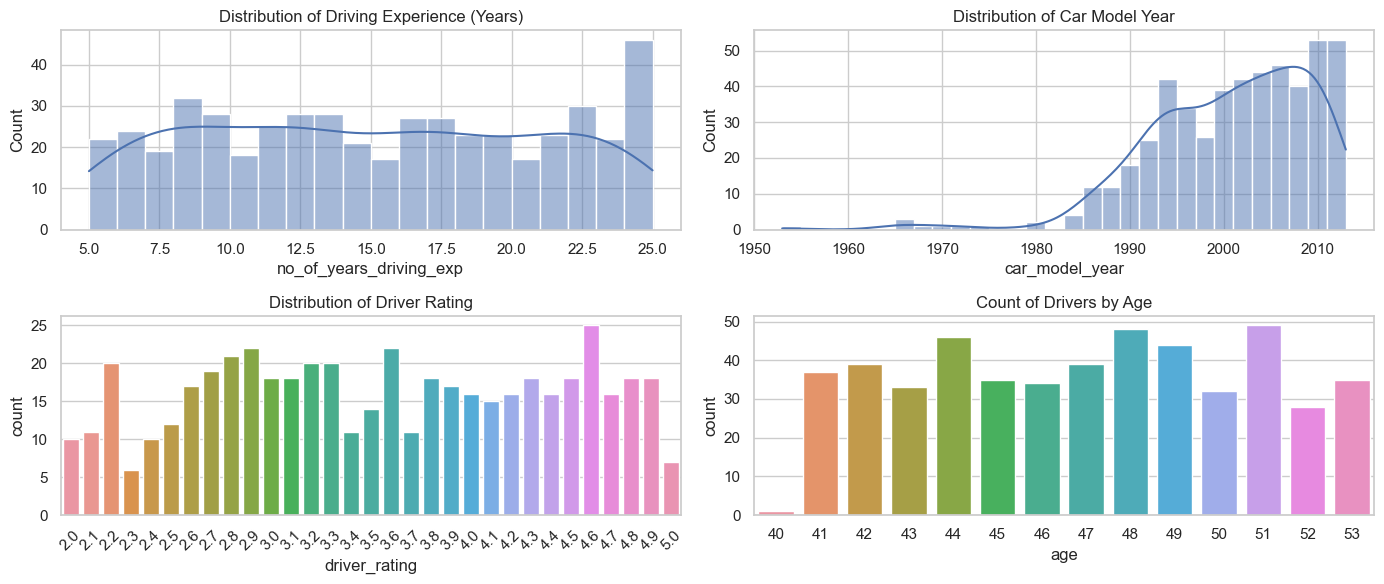
## 4.5 Tools and Libraries Used

The libraries used included Pandas for data manipulation, Dask for parallel processing, SQLAlchemy for database interaction, and scikit-learn for label encoding

# 5. Data insights

## 5.1 EDA Insights

### 5.1.1 Distribution Analysis of Driver Experience and Vehicle Models



**Driving Experience**

The histogram of driving experience shows a relatively uniform distribution with a slight increase in frequency for drivers with 5 to 10 years of experience. This suggests a diverse range of experience levels among the drivers in the dataset.

**Car Model Year**

Vehicle model years are skewed towards newer models, with a peak around the year 2000. This could reflect the company's policy of updating their fleet or a general market trend of drivers preferring newer models.

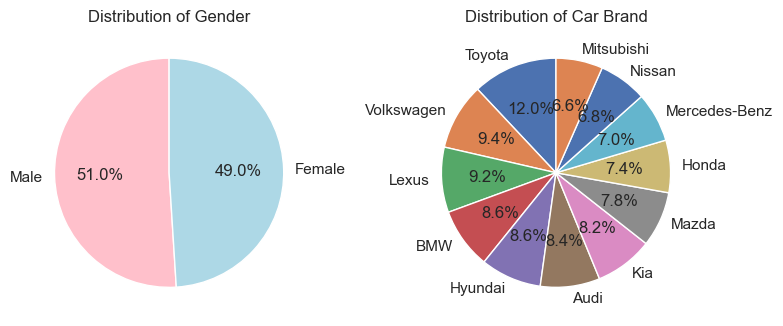
**Driver Rating**

The distribution of driver ratings indicates a multimodal distribution with peaks at ratings of 2, 3, and 4. This could imply a rating system where most drivers are rated above average, with fewer drivers at the extremes of the scale.

**Age Distribution**

It shows the number of drivers across different ages. The distribution across ages appears somewhat uniform with slight variations, indicating a diverse age range among drivers.

### 5.1.2 Gender and Car Brand Distribution Analysis

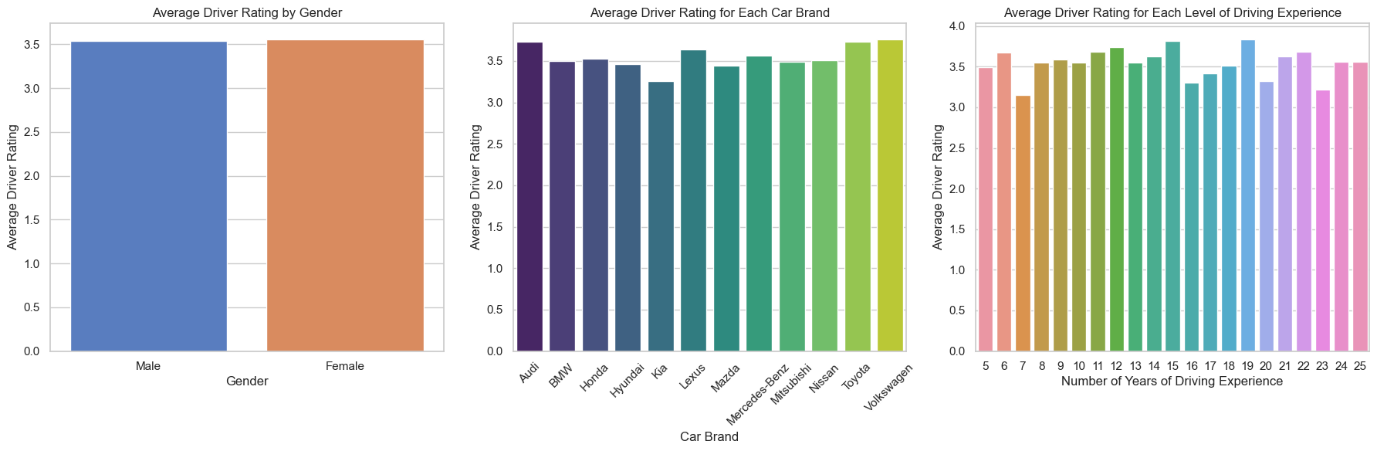


**Gender Distribution**

The pie chart represents the distribution of genders within the driver dataset, with two nearly equal portions indicating a balanced representation of genders. This even distribution suggests that the dataset likely does not contain any gender-related biases that could impact further analysis, such as predictive modelling.

**Car Brand Distribution**

The distribution of car brands is depicted in a pie chart, showing a diverse spread across various brands (labelled as numbers 0 to 11). No single brand dominates the dataset, indicating a heterogeneous vehicle fleet. This variety could be beneficial for subsequent analysis, as it may allow for the investigation of brand-related trends or preferences among drivers.

5.1.3 Analysis of Driver Ratings

**Driver Rating by Gender**

Bar plots were used to assess the average driver rating by gender. The results showed that the ratings are even across the two genders, labelled 'False' and 'True', which indicates a non-biased distribution of ratings across genders.

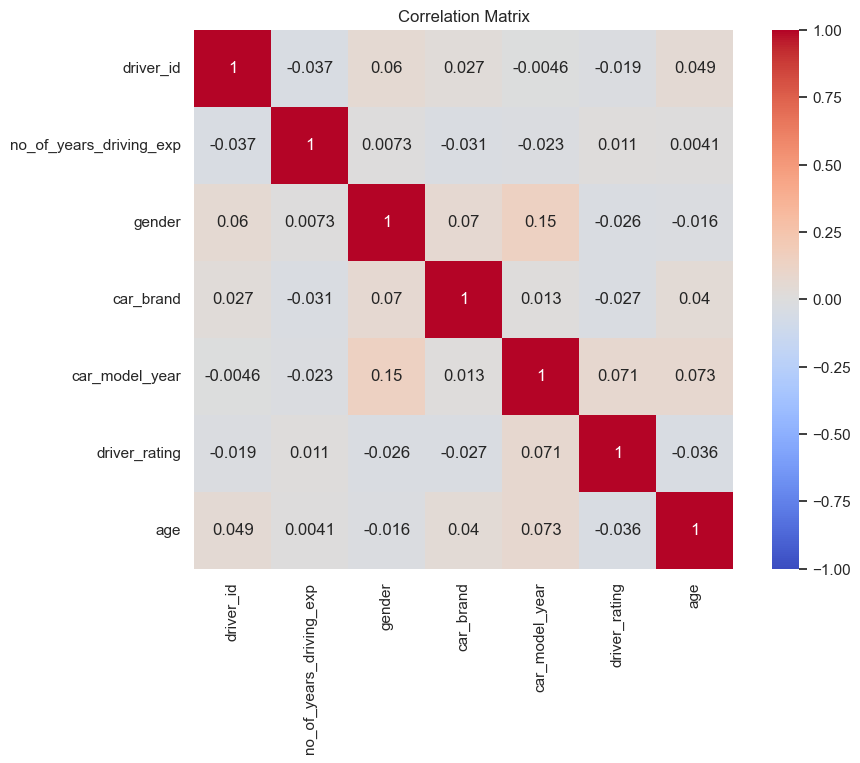
**Driver Rating by Car Brand**

The bar plot of average driver ratings across different car brands (numerically labelled from 0 to 11) indicated slight variations in ratings between brands. This suggests that certain car brands may correlate with slightly higher or lower driver ratings. Understanding these nuances could assist in evaluating brand-related performance metrics.

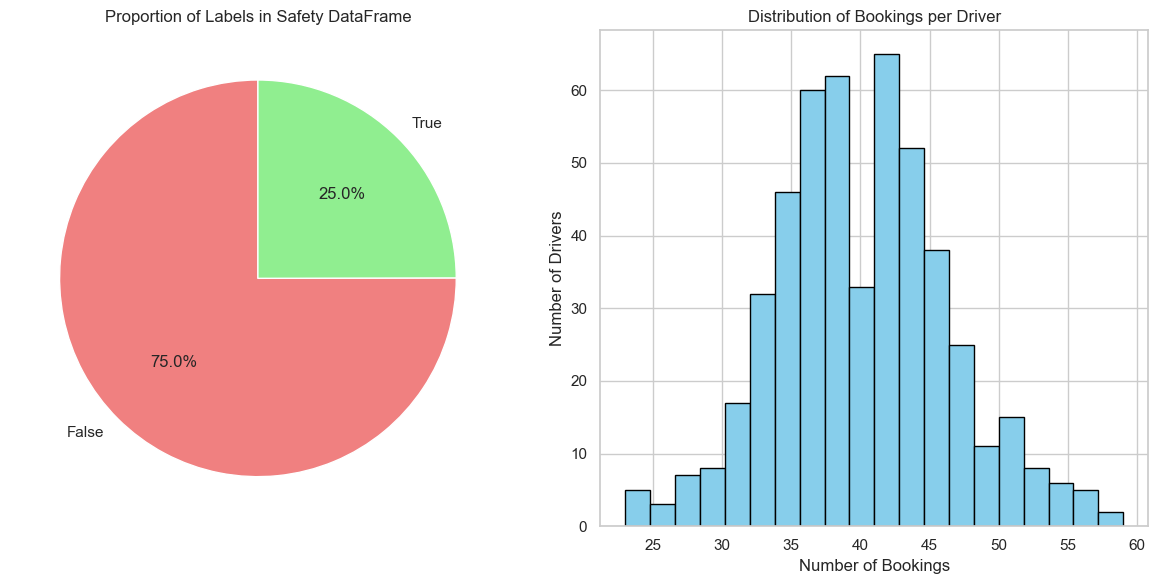
**Driver Rating by Level of Driving Experience**

Investigating the relationship between driving experience and driver ratings revealed a diverse pattern, with no clear trend suggesting a direct correlation between years of experience and the average rating. This indicates that experience alone is not a defining factor for a driver's rating within the dataset.

### 5.1.4 Driver Data Correlation Analysis



The absence of strong correlations implies that there is no single factor that significantly influences driver ratings. This could suggest that driver performance, as rated in this dataset, is multifactorial and not heavily dependent on the measured variables such as age, car brand, or years of experience.

5.1.5 Safety Data and Bookings Analysis

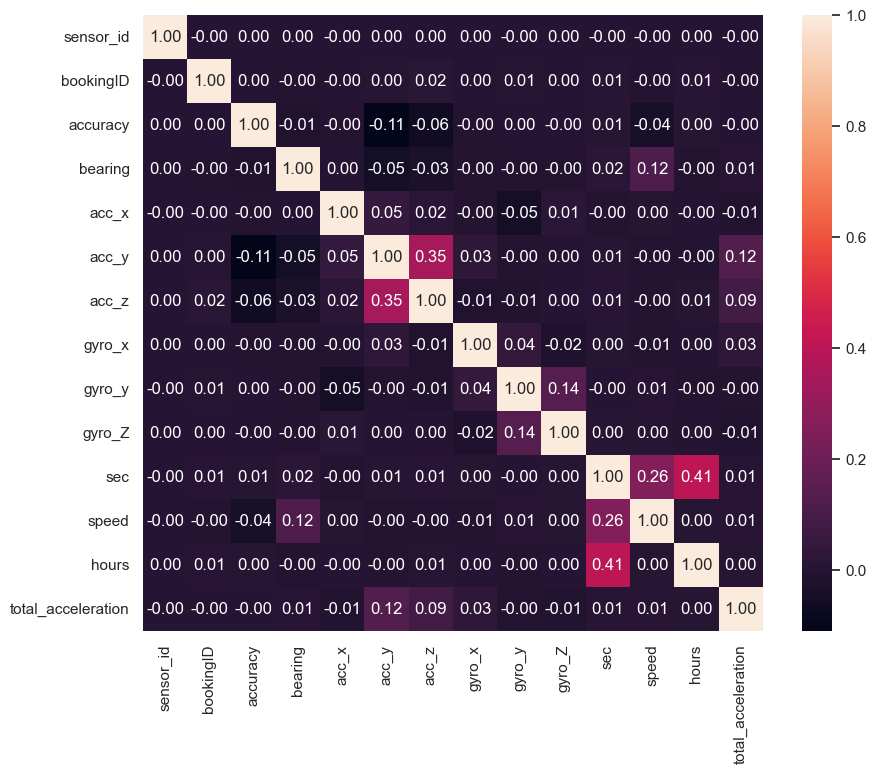
**Proportion of Safety Labels**

A pie chart was generated to visualize the proportion of safety-related labels within the dataset. It showed that 75% of the instances were labelled as 'False', indicating no safety concerns, while 25% were labelled as 'True', which may indicate some safety incidents or issues. This significant difference suggests that a large majority of the bookings are considered safe, which is a positive indication of overall safety standards.

**Distribution of Bookings Per Driver**

The histogram of bookings per driver reveals the frequency distribution of the number of bookings handled by each driver. Most drivers have between 30 and 50 bookings, with the distribution peaking around this range. This indicates that there is a relatively balanced workload distribution among drivers, with no apparent overloading or underutilization.

### 5.1.6 Sensor Data Correlation Analysis



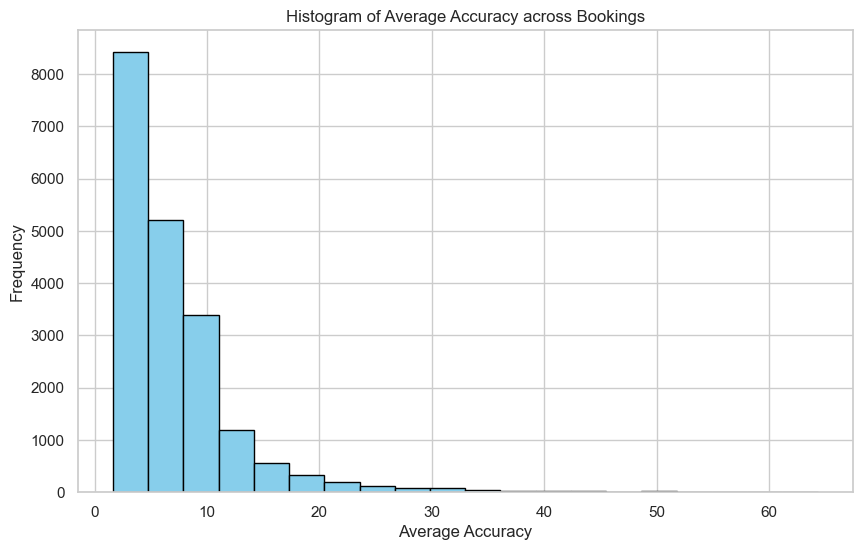
**Acceleration Axes:** A notable positive correlation is observed between ‘acc\_y’ and ‘acc\_z’ (0.35), which could indicate that these axes of acceleration are often influenced together during driving events.

**Gyroscopic Axes:** There is a moderate positive correlation between ‘gyro\_y’ and ‘gyro\_z’ (0.14), suggesting a relationship between these rotational movements, possibly during turns or manoeuvres.

**Time and Speed:** The ‘sec’ (time) feature exhibits a positive correlation with ‘speed’ (0.26) and ‘hours’ (0.41), indicating that longer trips might have higher speed readings, possibly due to highway travel or less congested conditions.

**Total Acceleration:** Interestingly, ‘total\_acceleration’ shows a positive correlation with ‘speed’ (0.12), which aligns with the understanding that higher speeds are typically associated with higher dynamic forces.

### 5.1.7 Analysis of GPS Accuracy Across Bookings



**Histogram Findings**

Most bookings have an average accuracy within the 0-10 range, which likely indicates a high degree of precision in the GPS data.

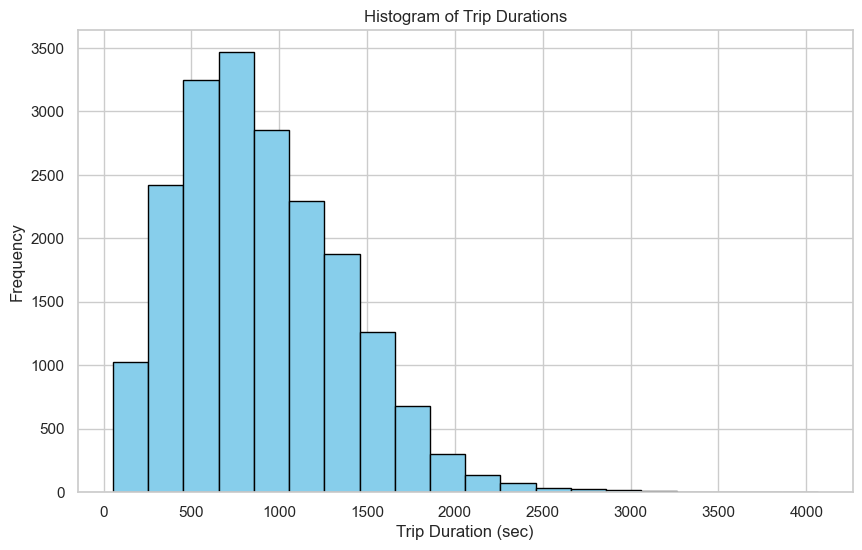
There is a rapid decline in frequency as the average accuracy value increases, which is consistent with expectations for GPS data, where most readings are expected to be precise with fewer instances of poor accuracy.

**Implications**

The concentration of data in the lower range of average accuracy suggests that the GPS equipment used in the bookings is generally reliable.

The distribution's skewness towards lower accuracy values implies that the instances of poor accuracy are relatively infrequent, but they do exist and may need to be addressed.

### 5.1.8 Analysis of Trip Duration Across Bookings



**Histogram Findings**

A wide range of trip durations, with most bookings having an average duration of less than 500 seconds.

The frequency of bookings decreases as the duration increases, suggesting that longer trips are less common than shorter ones.

**Implications**

The concentration of shorter trips could indicate the service's urban setting, where trips are often brief due to close proximities.

The presence of longer-duration bookings, though less frequent, may point to a subset of the service catering to longer-distance travel.

## 5.2 Tableau Insights

### 5.2.1 Dashboard 1 (Cab Driver)

A close-up of a graph

Description automatically generated

This dashboard is useful for a fleet manager or safety coordinator to monitor and evaluate driver performance, identify trends in driving behaviour, and take necessary actions to ensure safety and maintain high service standards.

**Driver Ratings**

This bar chart shows the individual driver ratings, with the colour coding likely indicating whether they are above, below, or at the average rating level. The chart clearly identifies top-performing drivers, which can be useful for recognizing and rewarding good performance, as well as identifying drivers who may need additional training or support.

**Proportion of Dangerous Trips**

The bar chart quantifies the proportion of trips that are considered dangerous for each driver. Drivers with a higher proportion of dangerous trips may require targeted interventions such as additional training, re-evaluation of their routes, or even disciplinary action depending on the cause of these incidents.

**Driver Average Net Acceleration**

This bar chart displays the average net acceleration for each driver, which can be an indicator of aggressive driving behaviour. Drivers with higher net acceleration might be contributing to the overall danger level and could benefit from safety training focused on maintaining a steadier pace while driving.

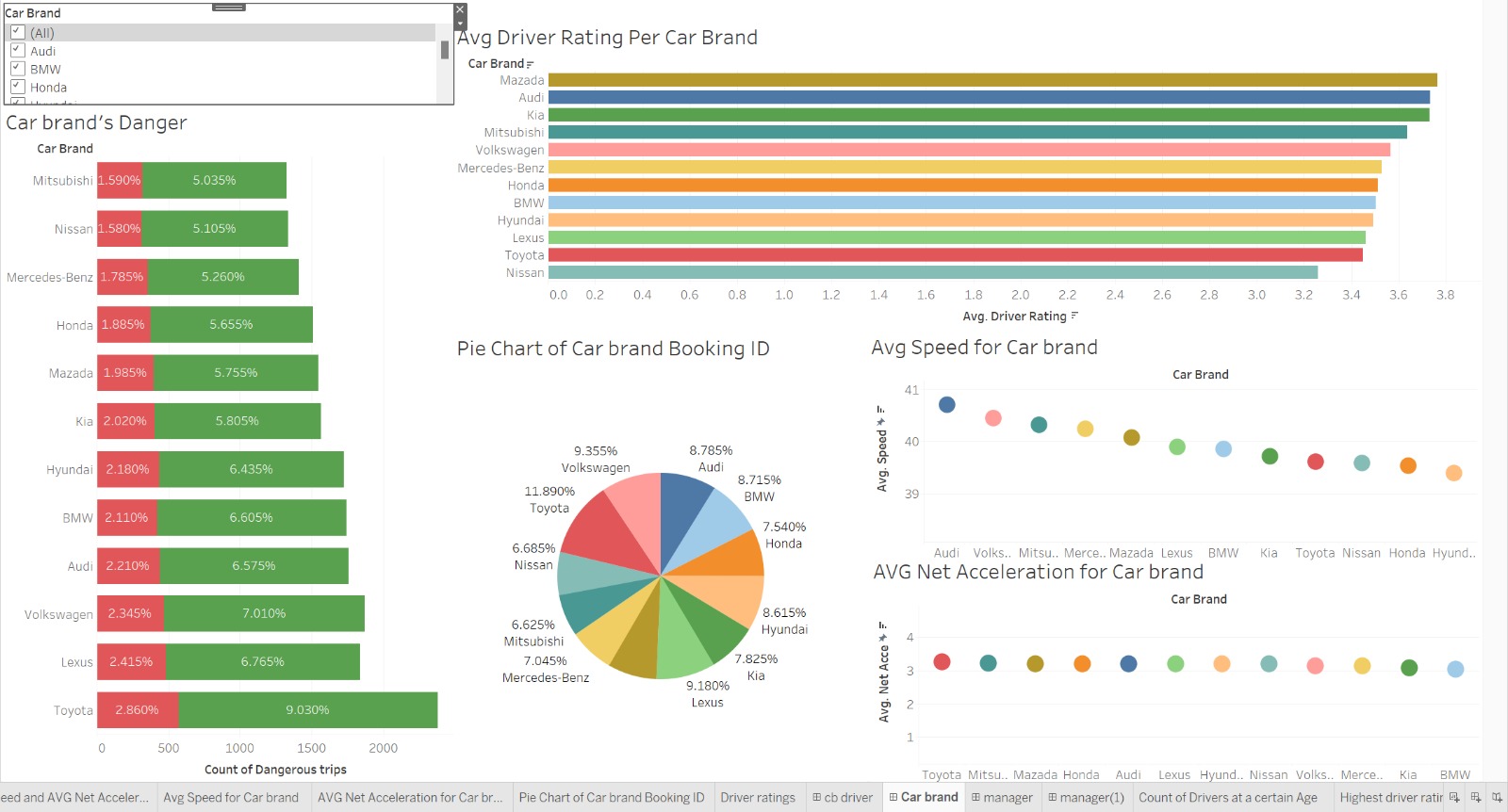
**Acceleration Danger Level**

This bar chart categorizes average net acceleration into different danger levels. It shows the distribution of drivers across these categories, providing a clear visualization of how many drivers are engaging in potentially risky driving behaviours based on their acceleration patterns.

**Extreme Acceleration per Booking ID**

This bar chart seems to represent the frequency of extreme acceleration events for individual bookings. It could be used to pinpoint specific instances where intervention might be necessary, such as a particular route that consistently leads to dangerous driving or identifying specific trips for further investigation.

### 5.2.2 Dashboard 2 (Car Brand)



This dashboard is for assessing car brand safety and performance as measured by driver ratings, trip danger levels, and driving metrics such as speed and acceleration.

**Car Brand’s Danger**

This bar chart displays two types of danger metrics associated with different car brands. The green bars likely represent a lower level of danger, and the red bars a higher level. The percentages may indicate the proportion of trips that fell into each danger category, with certain car brands showing higher proportions of dangerous trips.

**Average Driver Rating Per Car Brand**

The bar chart shows the count of dangerous trips associated with different car brands. The green and red bars might represent different severity levels of danger. Brands with higher counts of dangerous trips may require further investigation into whether these are due to vehicle performance issues or other factors. This data could influence fleet composition with a focus on safety.

**Pie Chart of Car Brand Booking ID**

This pie chart illustrates the proportion of bookings for each car brand. The data could be used to identify the most popular or most utilized car brands within the service. This might reflect customer preferences or the availability of certain car brands in the fleet.

**Average Speed for Car Brand**

The scatter plot shows average speed by car brand. There is a spread of average speeds across various brands. Car brands that tend to have higher average speeds might be preferred for certain types of trips, or it could indicate which vehicles are driven more on highways versus urban areas.

**Average Net Acceleration for Car Brand**

Similar to average speed, this scatter plot represents average net acceleration for each car brand. Brands that show higher average net acceleration could potentially indicate a tendency for more aggressive driving, or it might reflect the performance capabilities of those car brands. This could have implications for driver training and vehicle maintenance schedules.

### 5.2.3 Dashboard 3 (Manager)



These visualizations collectively offer a comprehensive look at driver demographics, experience, and performance.

**Highest Driver Ratings**

This horizontal bar chart lists drivers and their corresponding ratings out of a possible 5 points. The graph shows a range of ratings, with some drivers achieving near the top score. This indicates that there are standout performers who may have best practices that could be shared with other drivers. It also suggests the possibility of a reward system for high achievers to encourage continued excellence.

**Proportion of Dangerous Trips per Driver**

This bar chart displays the proportion of dangerous trips for each driver. Drivers are listed on the y-axis, and the proportion of dangerous trips is on the x-axis. Some drivers have a higher proportion of dangerous trips. This could be an area for targeted safety training or further investigation into the reasons behind these statistics.

**Average Age for Gender**

The pie chart shows the average age of drivers, divided by gender. It appears to be a near 50/50 split between male and female drivers, indicating gender parity within the driving team. The average ages are also very close, suggesting there isn't a significant age gap between male and female drivers.

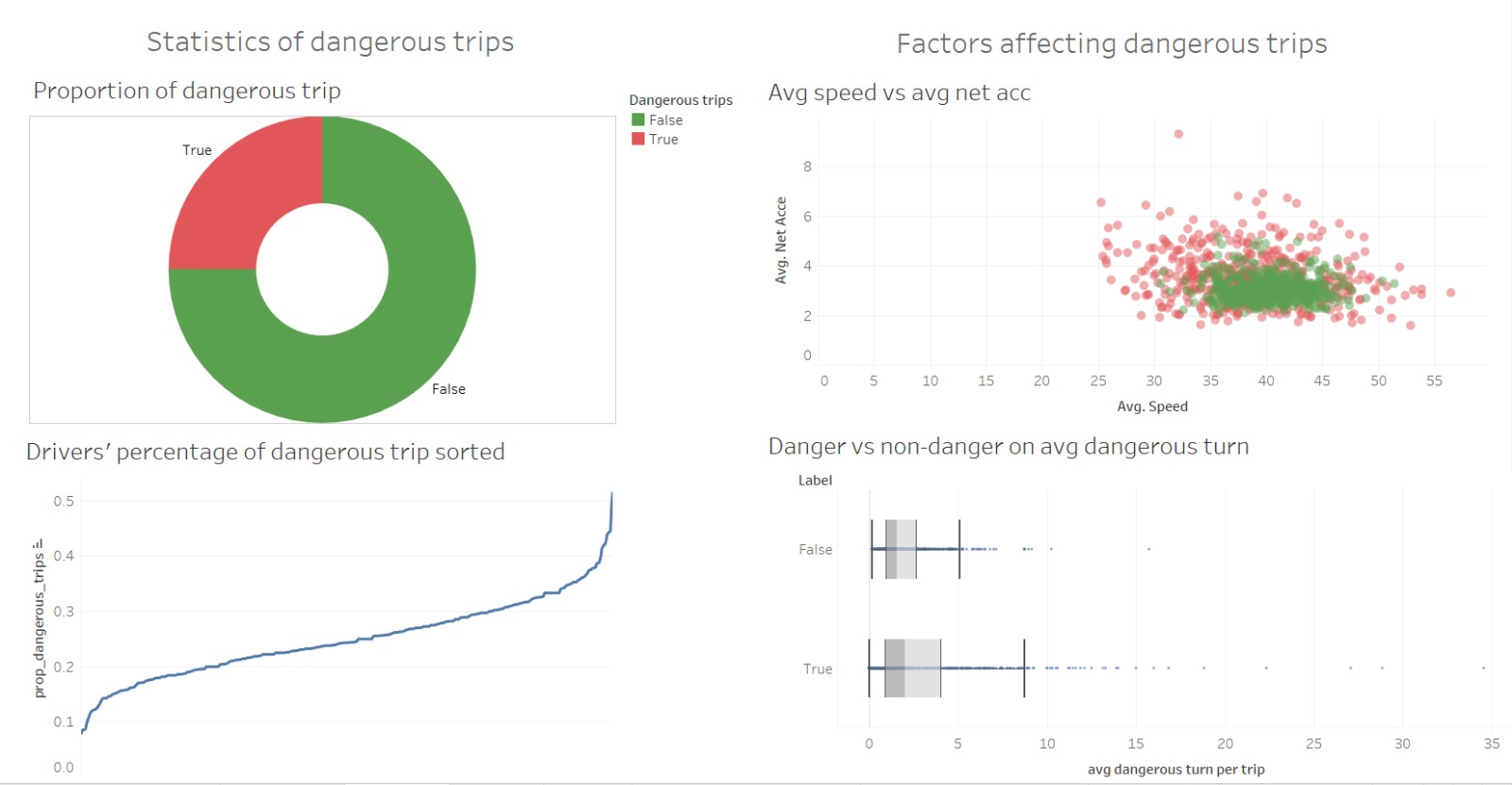
**Count of Drivers at a Certain Age**

This histogram shows the distribution of drivers across various age ranges. The peaks of the distribution could indicate common ages for hiring or the ages at which drivers are most available in the job market. This information could be used for demographic targeting in hiring campaigns or predicting turnover due to retirement.

**Number of Years in Driving Experience**

Another histogram displays the number of drivers according to their years of driving experience. There's a wide range in the level of experience, with fewer drivers at the very high end of the scale. This may indicate a young workforce or high turnover. It could also suggest the need for a structured experience or mentorship program to foster skill development among less experienced drivers.

### 5.2.4 Dashboard 4 (Danger)



These visualizations would be valuable for identifying and targeting areas for improvement in driving behaviour, potentially through training or policy changes

**Proportion of Dangerous Trips**

This donut chart shows the breakdown of trips into two categories: 'Dangerous' (True) and 'Not Dangerous' (False). The almost equal proportions suggest that the number of trips classified as dangerous is significant. This indicates that safety measures could be improved, and it is essential to investigate what factors contribute to a trip being classified as dangerous.

**Drivers’ Percentage of Dangerous Trip Sorted**

The line graph depicts the cumulative proportion of dangerous trips for drivers, sorted in ascending order. The curve's steepness towards the right suggests that a smaller number of drivers have a disproportionately higher percentage of dangerous trips. This could imply that specific drivers or driving behaviours contribute more significantly to the overall danger metrics and should be the focus of targeted training or intervention programs.

**Average Speed vs Average Net Acceleration**

The scatter plot compares trips based on average speed and net acceleration, with dangerous trips marked in red and non-dangerous in green. The concentration of green in the middle suggests that most trips occur at moderate speeds and accelerations. The red points scattered, particularly at higher speeds and accelerations, indicate that these factors may increase the likelihood of a trip being dangerous. This suggests that implementing speed and acceleration controls could mitigate risk.

**Danger vs Non-danger on Average Dangerous Turn**

This set of box plots compares the average number of dangerous turns per trip between non-dangerous and dangerous trips. Dangerous trips have a visibly higher median of dangerous turns, indicating a strong association between the frequency of dangerous turns and trip danger classification. This insight could be used to develop driver assistance or alert systems to reduce the occurrence of dangerous turns, thereby improving overall trip safety.

# 6. Conclusion

GoBest Cab's comprehensive analysis of journey, driver, and safety datasets unveils key insights into driver behaviour and trip safety. Machine learning and interactive dashboards reveal that while most journeys are secure, a notable fleet segment exhibits riskier behaviour, prompting targeted safety interventions.

Correlations between sensor data and safety incidents, particularly in acceleration and gyroscope readings, suggest opportunities for enhanced driver training and vehicle maintenance. Analysing driver demographics indicates potential for mentorship programs to foster skill development. Variability in vehicle model years and booking distribution highlights fleet management streamlining possibilities, influencing procurement strategy. Driver ratings by brand can guide targeted improvements and policies.

As GoBest Cab prioritizes safety and customer satisfaction, insights from this project can shape strategic decisions. Continuous use of the developed application for real-time monitoring is recommended. Future efforts may include integrating more real-time data, advanced predictive analytics, and expanding analysis to include customer feedback for a holistic approach to safety and service quality improvements.