ST1508: Practical AI

**FINAL REPORT**

horizontal line

# 

# 

# Done By: Goh Rui Zhuo, Toh Kien Yu, Wang ZiXin

# 

# 

# 

# 

# 

# 

# 

# 

# Introduction

1. Gobest Cab has hired our data science team to create an intuitive and user-friendly machine learning software application that aims to help them visualise and analyse cab data.
2. The company desires a portable application that can operate on the manager’s laptop. GoBest Cab’s management has requested our data science team to construct an interactive dashboard to give insights into driver behaviour, distinguishing between safe and dangerous drivers.

## Business Understanding

1. Improvement of safety
   1. Enhance safety by identifying and analysiing driver behaviour
   2. Involves distinguishing between safety and dangerous trips and analyse various data such as acceleration in a trip
2. Interactive Dashboard
   1. Important to create an intuitive, user-friendly application
   2. Present information and analytics in a clear and easily understandable format
3. Data-Driven Insights
   1. Focus on utilising cab data collected during trips to extract meaningful insights
4. Provide Business Impact
   1. Utilising the insights to have an actionable steps
   2. Improve on the overall safety standards of GoBest Cabs

## Data Understanding

1. Sensors dataset
   1. In the sensor dataset, it contains the telematics data from speed to accuracy
   2. Dataset is split into 5 different files and with a total of 7 million dataset
   3. This dataset is useful in identifying the trend
2. Drivers Data
   1. Contains the driver information from name to rating
   2. Understand which driver is driving dangerously
   3. Suitable to provide insights of which driver to alert of dangerous driving
3. Safety Labels
   1. Provide the labels of dangerous and safe driving
   2. Contains booking ID for linking of foreign key with the sensor table

## User Stories Creation

### User Story 1

As a **manager,** I need an intuitive application with a user interface so that I can visualise, analyse, and compare cab-driving data such as the telematics data on trip safety and make informed decisions.

1. Acceptance Criteria:
   1. UI should be clean, intuitive, and easy to navigate.
   2. UI should be interactive and able to provide real time visualisation.
   3. Safety metrics must be relevant for the analysis of trip safety.
   4. Real time alerts for unsafe driving behaviour
2. Confirmation
   1. Speed base of each driver can be seen.
   2. Info is well readable by the user.
   3. Able to seamlessly identify dangerous trips.

### User Story 2

As a **cab driver**, I want to have a personalised dashboard where I can view statistics such as my driving ratings and safe driving so that it improves the passenger's experience.

1. Acceptance Criteria:
   1. The dashboard should be updated in a timely manner.
   2. Must be able to maintain privacy support to prevent data leakage.
   3. Drivers can customise the dashboard and should have a user-friendly interface.
2. Confirmation
   1. Drop down menu provided for different options
   2. All the information which appears in dataset can be found in dashboard
   3. The dashboard should be interactive and have user-friendly interface (added colour style)

### User Story 3

As an **administrator**, I want to be able to ensure and efficient access to data related to driver’s performance and safety metrics, so that business analysts can effectively monitor safety and identify trends.

1. Acceptance Criteria:
   1. Ensure high availability of data.
   2. Efficient access to the database.
   3. Able to support customisation.
2. Confirmation
   1. Database contains safety metrics such as safety labels with other information such as speeds.
   2. Databases are well designed with proper primary key and foreign key

### User Story 4

As a **manager**, I want to be able to do a comparison between different drivers so that I can compare between different drivers in visualisation to spot patterns of dangerous driving habits.

1. Acceptance Criteria:
   1. User friendly UI
   2. Able to perform driver selection for deeper analysis.
   3. Anomaly detection
   4. Able to provide historical trends for analysis.
2. Confirmation
   1. User friendly interface for seamless user interaction
   2. Retrieve historical trends.
   3. System capability to detect anomalies in driver behaviour.
   4. Ability to select driver

### User Story 5

As a **manager,** I want to create a model that can predict the safety and unsafety of a trip so that I can notify the driver to beware and be more cautious when driving.

Acceptance Criteria:

1. The accuracy of prediction should be high
2. It should be able to provide real time prediction for trip safety
3. The model should consider a wide range of relevant data that contribute to travel.

### User Story 6

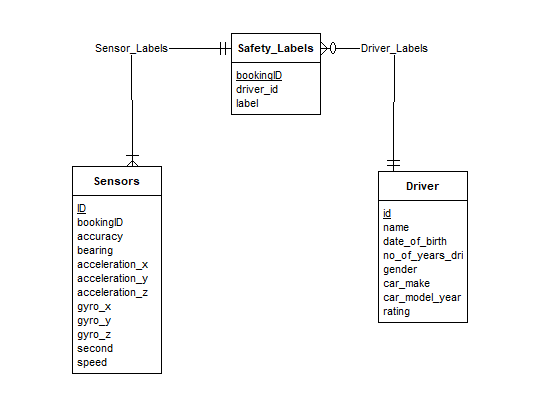
As a **cab driver,** I want a model that can predict the safety and unsafety of a trip so that I can notify the driver to beware and be more cautious when driving.

Acceptance Criteria:

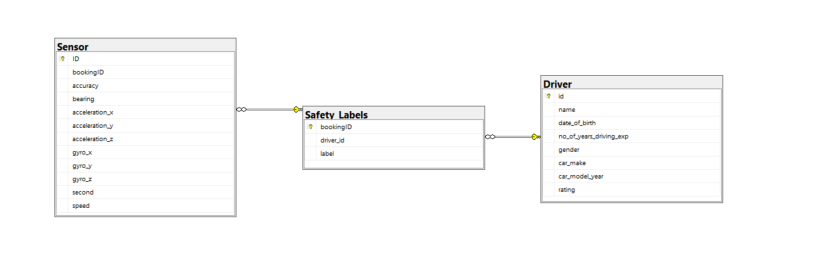
1. The accuracy of prediction should be high
2. It should be able to provide real time prediction for trip safety
3. The model should consider a wide range of relevant data that contribute to travel.

## Database Operation

**Database Design**

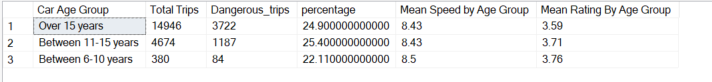
1. With three tables of Sensors, Drivers and Safety Labels, there is a need to link the tables together
2. Sensor will have a separate unique identifier key as its primary key with booking ID as foreign key
3. Drivers will have driver id as the primary key
4. Safety Labels will have the booking ID as the primary key with foreign key driver id
5.  *Figure 1: Database Schema for Sensor, Driver, and Safety Label Tables*

**Database Insertion**

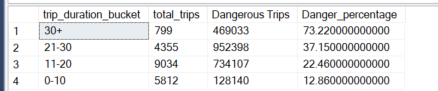
1. Creation of table with the use of create statement in SQL
2. Insert statement was created with the use of bulk insert
3. Database was successfully created with proper primary and foreign key
4. *Figure 2: Database Diagram In SQL Server*

**Query Creation**

**Query 1:** Get the percentage of dangerous trips for each vehicle age group and important information with average speed (Appendix 1 )

1. Idea behind is to provide information on whether each age group of car has a higher percentage of dangerous trips
2. This is so that if higher age group generally has a higher percentage of dangerous trips, we are able to identify some causes and solve the problem
3. *Figure 3: Query 1 Results*
4. Insights
   * 1. From the result, we can see that between 11 and 15 years of car age, it has the highest percentage of dangerous trips at 25.4%
     2. However, the mean sped is actually lower and the mean rating by each group are generally still in the middle
     3. Also, we can see that majority of the trips are operated with majority over 15 years, showing that the cars are at a older age

**Query 2:** Group the total duration by interval, and analyse the danger percentage during that interval itself

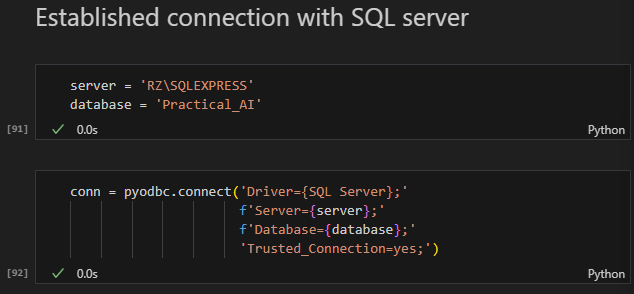
1. Idea behind is to provide information on whether a longer trip has a higher percentage of being Dangerous
2. This is so that if higher trip duration group generally has a higher percentage of dangerous trips, we are able to identify some causes and solve the problem
3.  *Figure 4: Query 2 Results*
4. Insights
   1. From the result, we can see that between 11 and 20 mins, it has the highest total trips at 9034 with 30+mins the leats
   2. However, we can see that in those that are 30 or more mins, it has the highest percentage at 73.2% showing that the this duration has to be look and manage especially for longer trips

**Query 3:** Analyse the pitch, roll and yaw by the safety labels

1. Idea behind is to provide information of the pitch , roll and yaw of the driver
2. This is so that we are able to analyse the different statistics for pitch, row and raw and determined is there any major difference between the labels
3. Pitch is the rotation around the side to side axis, Roll is the rotation around the front to back axis and yaw is the rotation around the vertical axis
4. *Figure 5: Query 3 Results*
5. Insights
   1. Average pitch degrees of normal are higher compared to dangerous at 3.71 compared to 3.5.
   2. Average roll degrees of normal are higher compared to dangerous at 34.84 compared to 32.04.
   3. Average yaw degrees of normal are lower compared to dangerous at -0.02 compared to -0.01.
   4. Across all metrics, it seems that generally the raw pitch and yaw did not exactly resulted in the classification of safety labels

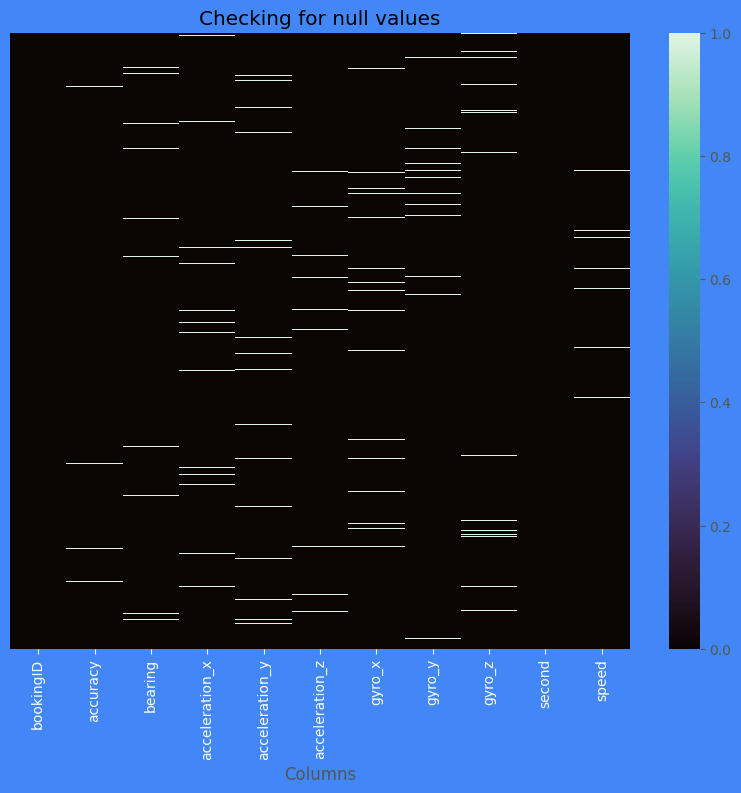
**ETL Pipeline**

**Connect to SQL server**

1. Utilising pyodbc to integrate the dataset into python jupyter notebook
2.  *Figure 6: Code for the ETL pipeline*

**Data Preprocessing - Part 1**

**Check for null values**

1. Null values was check for each dataset with the use of pandas
   1. Sensor dataset contains null values while the other two dataset did not contain null values
2. Additional check for sensor dataset
   * 1.  *Figure 7: HeatMap Visualisation*
     2. Insights
        1. For sensor dataset, majority of the null values can be seen in almost all columns and occurs at random with no observable pattern

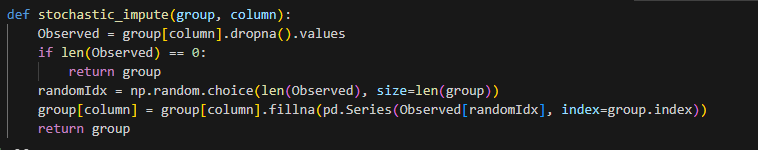
**Check for Correlation between each variable to determine imputation methods**

1. Insights
   1. For both of the datasets, there is no clear correlation between the variables and the best is just a moderate correlation

**Drop rows where seconds is null**

1. Because of the nature seconds is, it is not viable to impute seconds as we do not know the correct order of it
2. Dropping those rows is the best option here

**Imputation of all columns**

1. Stochastic Imputation was used
   1. Aims to reduce the bias by an extra step of augmenting the predicted score with a residual term
   2. Residual term is normally distributed with mean zero and variance equal to residual variance
   3. Useful when missing values occurs at random
   4. *Figure 8: Stochastic Imputation Code*

**Exploratory Data Analysis**

**Analysing on the acceleration column**

1. Utilised Plotly Go to plot out the graph
2. Insights
   1. For the acceleration data, we can see that x is symmetrical, y is negatively skewed and z is symmetrical, all contains outliers
   2. However, outliers may not be able to drop due to the fact that it can affect the overall data

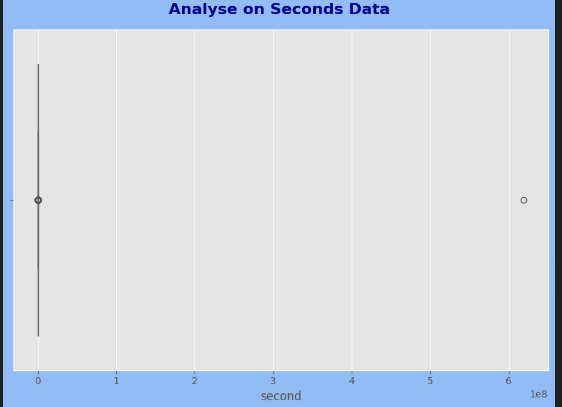
**Analysing on the gyro columns**

1. Insights
   1. For the gyro data, we can see that x is symmetrical, y is negatively skewed and z is symmetrical, all contains outliers
   2. However, outliers may not be able to drop due to the fact that it can affect the overall data

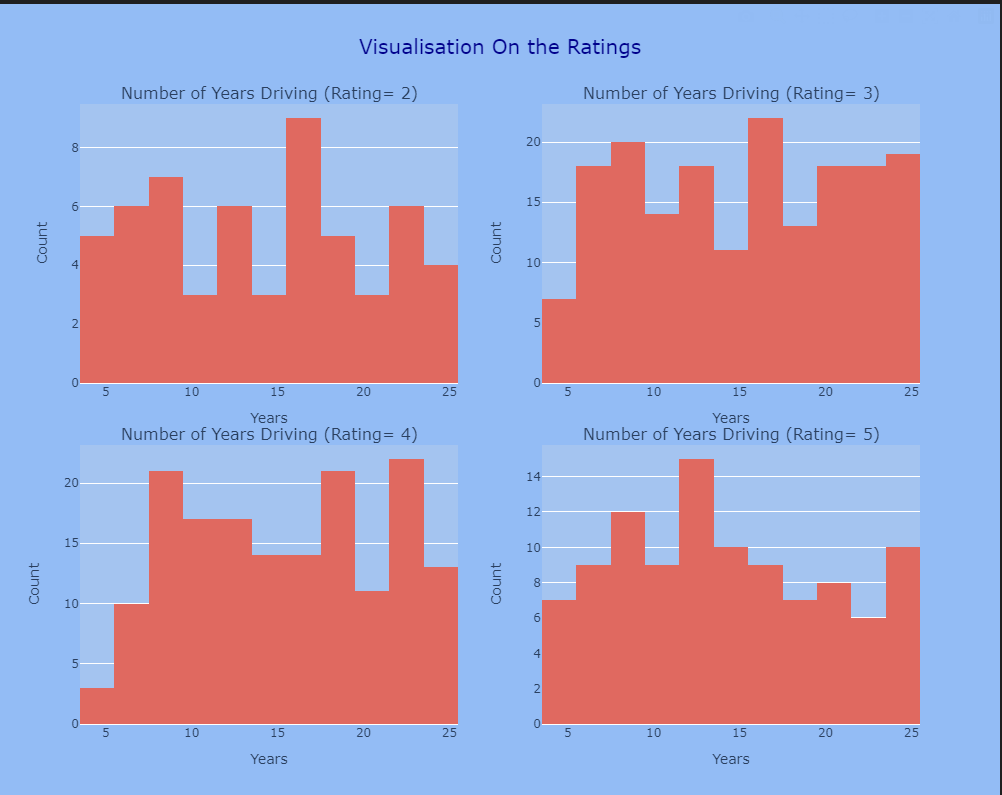
**Analysing on the other columns**

1. Insights
   1. For the other columns, we can see that bearing is symmetrical, accuracy is negatively skewed and speed is symmetrical, all contains outliers
   2. However, outliers may not be able to drop due to the fact that it can affect the overall data

**Analysing on the second column**

1. The seconds column is separated out to analyse on the second dataset due to the nature of the data
2.  *Figure 9:Speed outlier testing*
3. Insights
   1. For the speed, we can see that there is a extreme outlier and therefore we need to have a data preprocessing part two to remove
   2. Does not make sense to have 500 days of trip

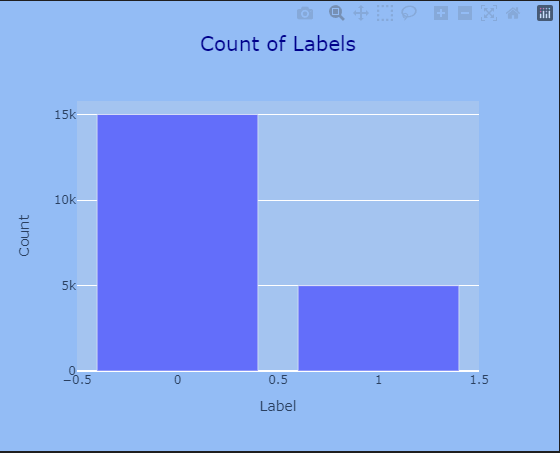
**Analyse the Years Of Experience with respect to Rating**

1. This is to analyse the distribution of experience base on the rating the driver has
2.  *Figure 10: Rating Distribution Graph*
3. Insights
   1. For the rating 2, we can see that 15 to 16 years has the highest count
   2. For the rating 3, we can see that 15 to 16 years has the highest count
   3. For the rating 4, we can see that 23 to 24 years has the highest count
   4. For the rating 5, we can see that 12 to 13 years has the highest count

**Analyse the Years Of Experience with respect to Rating**

1. This is to analyse the dataset in terms of car makes and the gender
2. Insights
   1. For the car make, we can see that Toyota has the highest count and Mitsubishi has the lowest count
   2. For the gender, the male has a higher count than the female

**Analyse the count of each label**

1. This is so that we know how the dataset looks like in terms of number of counts for each class
   1. *Figure 11: Count of labels for each class*
   2. Insights
      1. We can see that there is more 0 labels as compared to 1 label in the dataset, showing an imbalance of class

**Analyse the Distribution by labels**

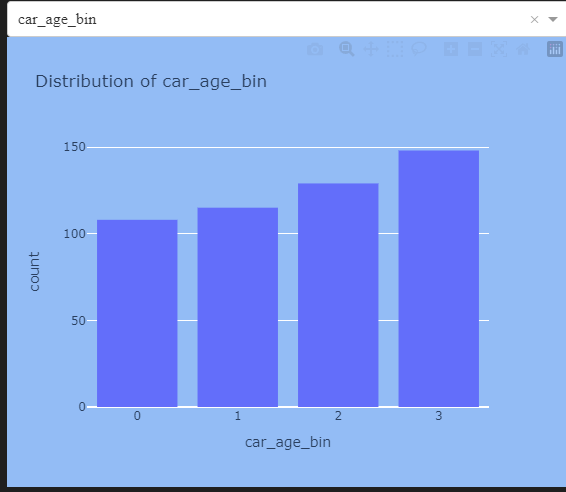
1. This is to analyse on all the columns with respect to the labels
2. Insights
   1. Distribution across all labels and columns are about symmetrical with the count around the same

**Analyse on top 5 with the most dangerous trips**

1. This is to analyse the top 5 most dangerous trips by the driver id of the driver
2. Insights
   1. Driver id 252 441 133 150 479 are part of the top 5 which has the most dangerous trips

**Advanced Exploratory Data Analysis**

**Full Analysis of Driver dataset with Plotly Dash**

1. This is to analyse the top 5 most dangerous trips by the driver id of the driver
2. Provide seamless and effective use of analysing the drivers dataset
3.  *Figure 12: Plotly Dash with Dropdown Menu*

**Data Preprocessing - Part 2**

**Detect outlier of second data**

1. This is to detect the outlier in the seconds data to determine how much to drop
2. 1 day is the limit for the trips and therefore anything more than that will be remvoed

**Feature Engineering**

**Dropping speed which is negative**

1. The negative speed are being drop off due to it being unrealistic to have negative speed

**Feature Extraction the new columns base on speed**

1. New columns base on speed will be created to provide more possible columns for visualisation in tableau
2. Mean, Median, Standard deviation , max and distance will be feature extracted

**Total Duration of Each Trip**

1. Total duration of the trip was got through group by and get the total duration
2. Then it is converted to mins

**Car Age Bin and Age Bin**

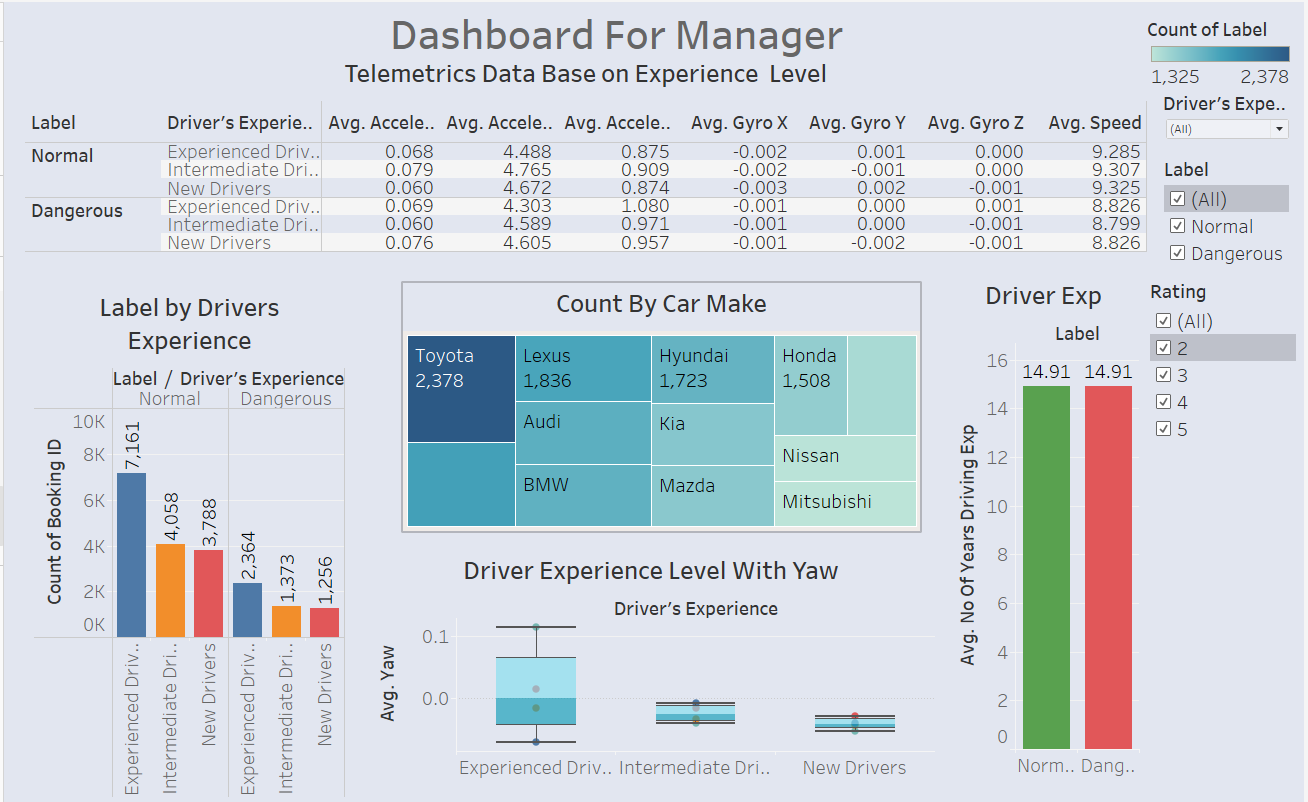
1. With the use of KBinsDiscretizer, the age and car age bin was bin into 5 different bins and then categorise into low and high etc
2. This is to provide analysis on age category for visualisation and not just the age

**Get the Pitch, Roll and Yaw**

1. Pitch: Rotation around the vehicle side to side axis, it indicates the up or down angle of the nose, experience when accelerating or braking
2. Roll: Refers to rotation around the vehicle front to back axis, felt through turning
3. Yaw: Rotation of the vehicle around the vertical axis which runs from top to bottom the vehicle

**Dashboard Insights**

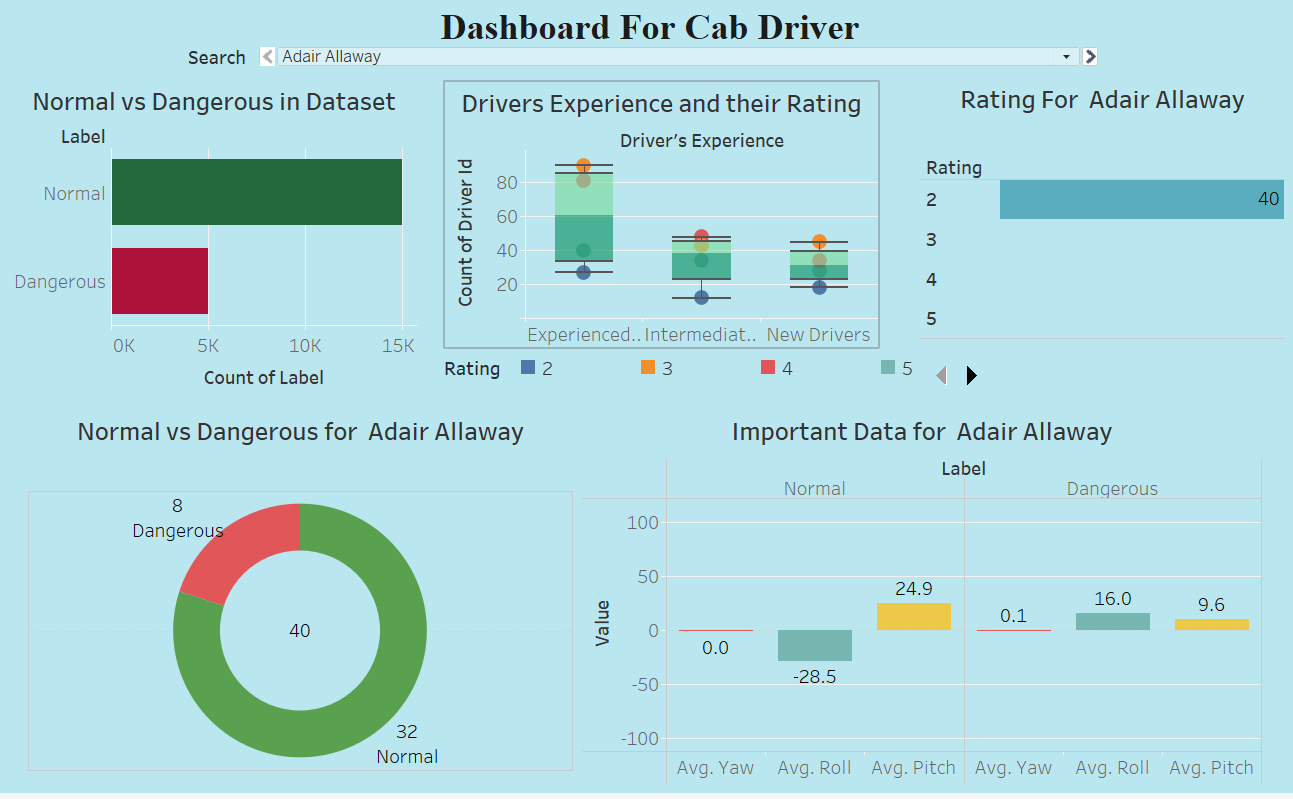
**Dashboard 1**

*****Figure 13: Dashboard 1*

**Dashboard 1 Analysis**

* Dashboard 1 allows managers to visualise, analyse, and compare cab-driving data such as the telematics data on trip safety and make informed decisions.
* Manager is able to customise cab driver’s based on car make, safety label and driver’s experience for analysis.
* It can be seen that most drivers are experienced and drive Toyota.
* A box and whiskers plot was made to analyse the yaw experience based on their driving experience. The newer drivers are more conservative have a smaller yaw which means smaller rotations between turns whereas the more experienced drivers are more ‘daring’, having a larger yaw which suggests a difference in driving behaviour.
* In addition there is a bar plot to compare between the average years of driving experience

**Dashboard 2**

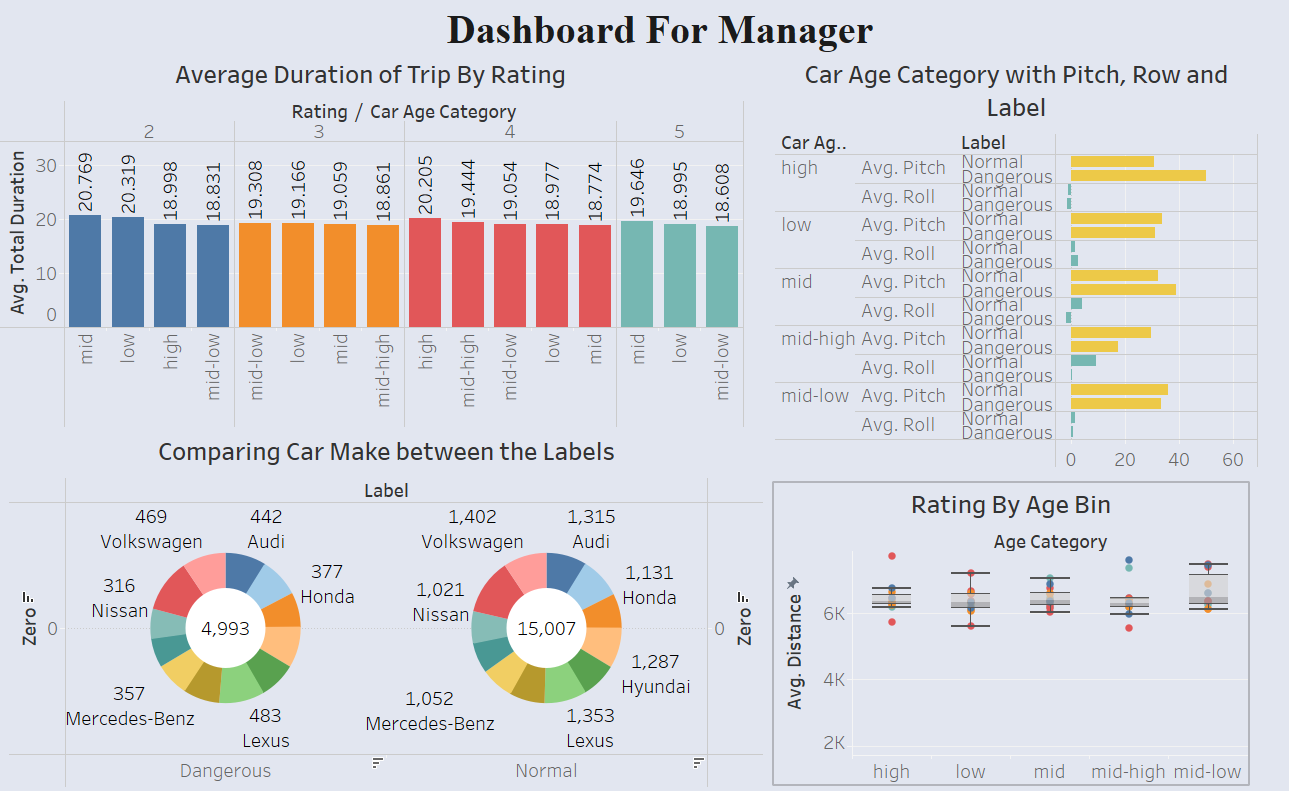
****

*Figure 14: Dashboard 2*

**Dashboard 2 Analysis**

* This dashboard shows the information about the rating of the driver and the trip he takes.
* It provides a comparison between normal and dangerous in the dataset for all the drivers, we can see that there is a higher count of label 0 which is Normal
* Driver can check his experience level , rating , percentage of safe and dangerous trip and some important data by put his name in dropdown menu, this give insights on how the experiences of drivers might correlate with satisfaction of the customer
* Even the driver with the rating of 2 , 80% of the trip ate safe
* The pitch, row and yaw shows the comparison between the labels of Adair Allaway and we can see that for normal trip, he has a lower yaw as compared to for dangerous trips
* In addition , the roll is lower in Normal as compared to Dangerous. However yaw is a little higher in normal as compared to dangerous
* Therefore, from this plot, Adai Allaway can consider having less sharp turn to improve passenger experience to improve the safeness of the trip

**Dashboard 3**

****

*Figure 15: Dashboard 3*

**Dashboard 3 Analysis**

* Dashboard 3 provides more in depth analysis on the different car make for managers to delve deeper into.
* The average duration of trips was also compared based on the cab driver’s rating
* Average duration of trip with respect to rating to provide how the age of the car might affect the length of trips and the rating receives
* A donut chart was plotted out to sieve out the count of different car makes between safe and dangerous drivers which helps the manager to see which one are more associated with each safety category and to make a suggestion to the upper management
* Grouped horizontal bar chart displays the average pitch and roll by car age category that is further break down into normal and dangerous which provide insights that certain cart ages are associated with riskier driving patterns
* Rating with age bin that provides the idea of distribution of average distance by age category which helps us to determine if there is a correlation between the cart age and the distance its driven and ratings it receives
* The trip duration does not exactly affect the rating, however, we can see that in each group such as rating 2, the mid age group has the highest total duration
* We can see that Volksawagen has the highest in normal while Lexus has the highest count in the dangerous label
* In the average roll bar chart, we can see that there is a higher roll in dangerous as compared to normal labels in the dataset

**Conclusion**

In conclusion, our team has developed 3 dashboards for GoBest Cab for managers and cab drivers to utilise based on their different needs. We placed ourselves in the shoes of these users when utilising our product to ensure high standards of our work.

Through this project, we learn about the different processes of a data science project. At the start, we learn about database designing, creation of query and pipeline followed by doing Data cleansing and data preprocessing. In addition, additional analysis with Plotly was tried and created

Overall, the dashboard provided a good analysis to fulfil the user story of users manager and drivers with the database design fulfilling the database administrator.