**Group 2 Detailed Project Report**

Members: Jayden (p2112790), Wee Loon (p2112802), Johnnie (p2007476)

**WEEK 1**

**Background Research**

Our team was hired by Just Taxi, a ride-hailing service company. Our data science team is required to develop an intuitive machine learning software app for analysis and visualisation of their Taxi data.

Singapore is no stranger to road accidents, and Taxi drivers are no exception to this. According to ScienceDirect, taxi drivers in Singapore are involved in more than 14% of all reported road traffic crashes despite there only being 100,000+ drivers in 2016. Hence, Taxi driver safety is a huge priority for these companies.

The purpose of the data and therefore our app, is to give an understanding of the safety of the Taxi drivers’ when they are driving on the road. Our team decided that we need to App needs to be portable and able to be run on the manager’s laptop. It was decided that we would develop an interactive dashboard for data visualisation using Tableau.

**User Stories**

Before trying to help our users, we must understand our users. Hence, we conducted extensive secondary research on each of our user groups.

* Taxi Drivers
  + To understand our stakeholders (taxi drivers) even deeper, Johnnie hopped on to a taxi ride and asked the driver how AI and Machine Learning could help them for safety.
  + Found out that they wanted a way to keep track of how safe their driving is
* Taxi Company Manager
* IT Administrators

**User Stories board produced:**

**Diagram

Description automatically generated**

**WEEK 2**

**Product Backlog**

The user stories done by Johnnie allowed Jayden to create a clear and concise Product Backlog. It includes 8 items ordered based on priority with much detail so that everyone on the team knew exactly what needed to be done. In addition, information on man-hours was included to get a good idea of how long each product would take. Product Backlog is also made to accommodate the needs of our user stories

**Final Product Backlog:**

Graphical user interface, text, application, email

Description automatically generated

**SQL Database Design**

Diagram

Description automatically generated

The metadata was first viewed by Jayden, and subsequently a simple ERD with relations was designed using ER Assistant. Wee Loon was then able to write an SQL script in MSSQL to create the database, schemas, tables and also import the data. Tables **drivers, safety\_labels and sensor** were made. As the data in the csv file is not of the correct datatype, all the columns in the tables will initially be of VARCHAR

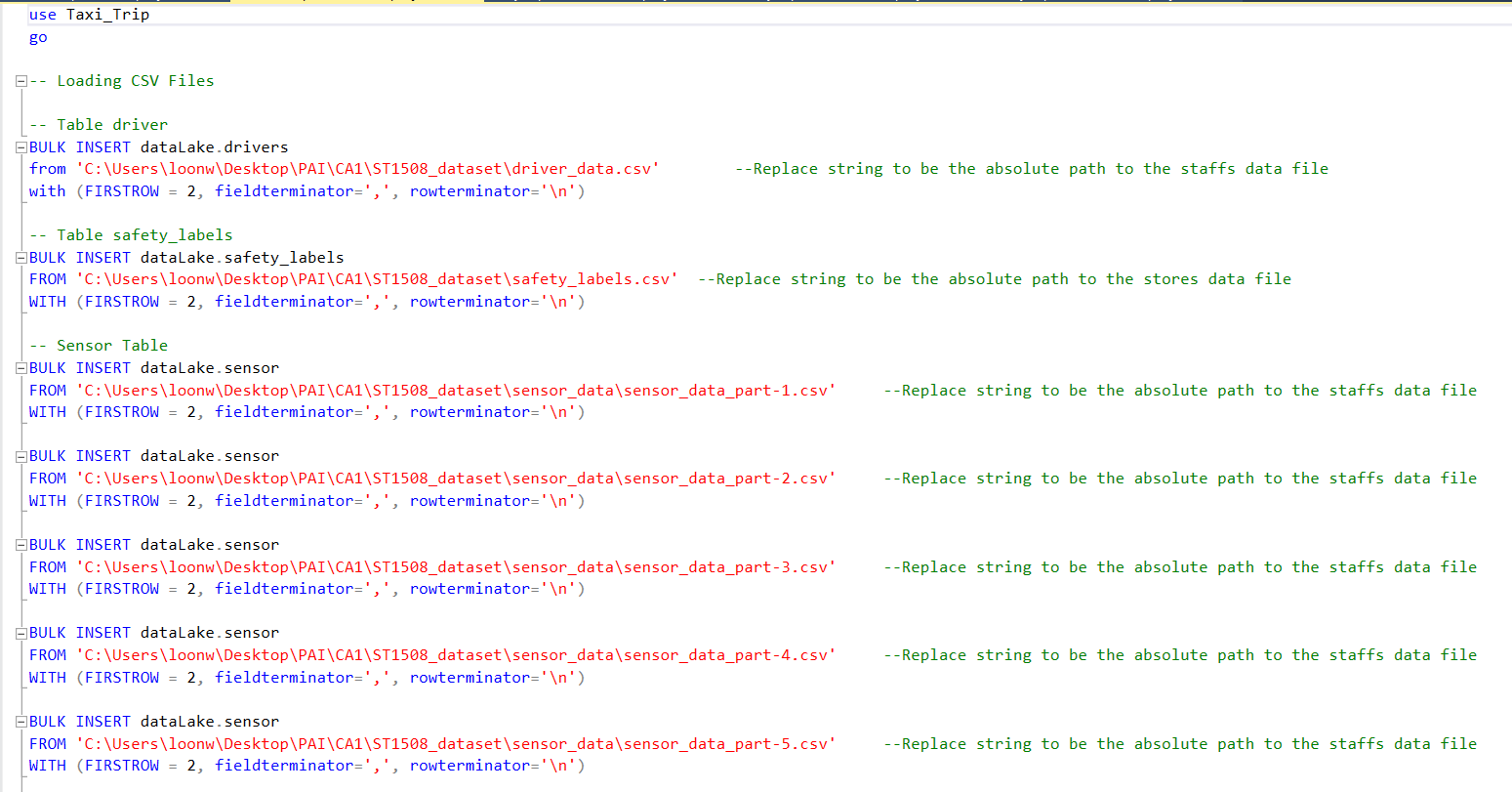
**SQL Queries used to create table:**

Text

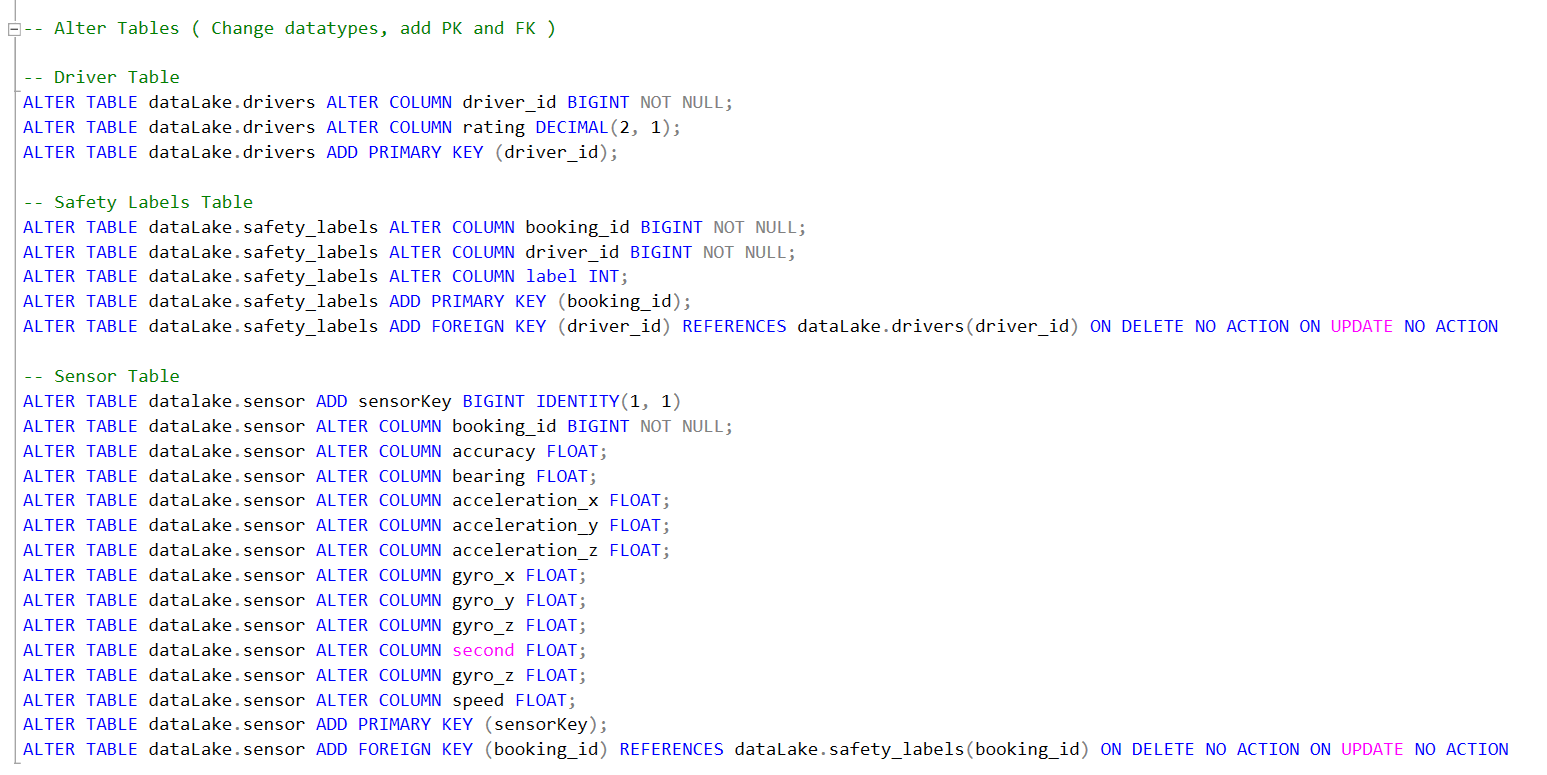
Description automatically generated

**SQL Queries used to import data:**

Bulk insert is used to import the data from the csv files into the tables.

****

After insertion of data, the datatype and relations were established.

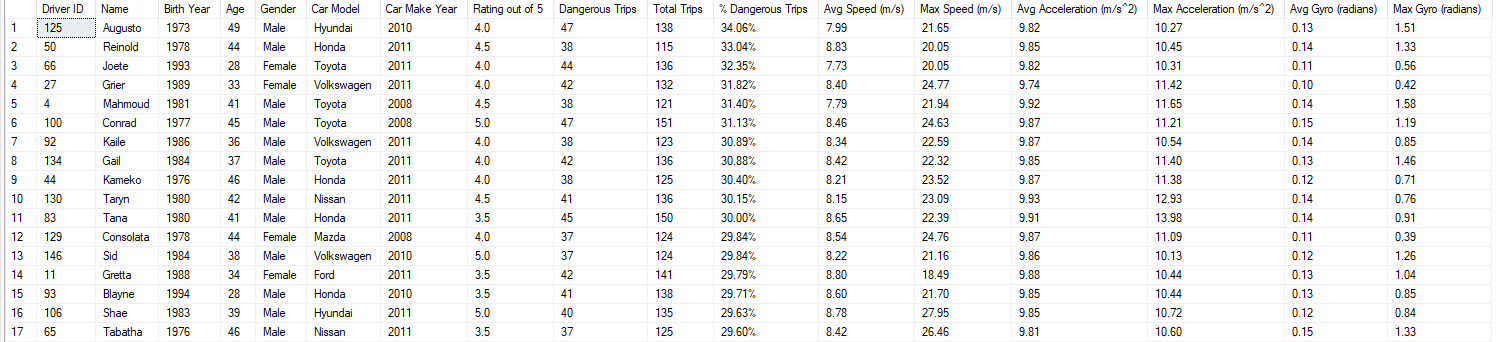
****

**WEEK 3**

**SQL Raw Data Complex Queries**

Jayden created 3 complex queries that were able to give insightful knowledge on the taxi trips data. He made use of sub-queries, partitions (ranking) and other methods to achieve our results. Sub queries are important as we might display the wrong number of labels if we simply join all our tables. E.g., if we were to simply join all tables it would show 7.3 million labels when we only have 0.02 million labels. All the 3 complex queries are as shown:

**Query 1:**



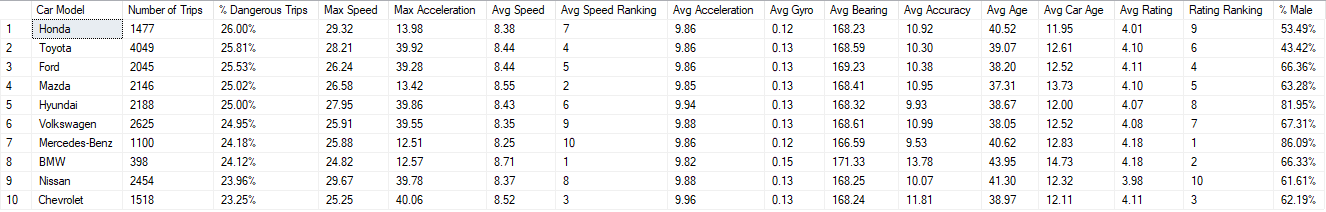
Query 1 display top dangerous drivers by percentage of dangerous labels and their information (DriverID, Name, Birth Year, Age, Gender, Car Model, Car Make Year, Avg Rating, No of Dangerous Trips, No of Total Trips, % of Dangerous Trips) with their car speeds information (Avg Speed, Max Speed, Avg Acceleration, Max Acceleration, Avg Gyro, Max Gyro) as well.

**Query 2:**

****

Query 2 display information regarding the safety labels such as, the number of Safe/Dangerous Count, Avg Speed, Max Speed, Avg Acceleration, Avg Acceleration (x,y,z), Max Acceleration, Avg Rating, Avg Gyro, Avg Gyro (x,y,z) and Max Gryo.

**Query 3:**

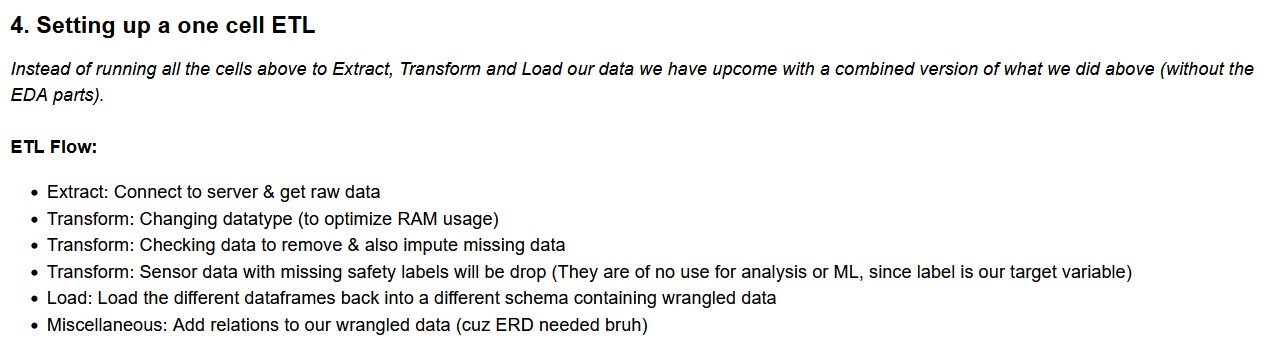
****

Query 3 display information regarding the different car models such as their Number of Trips, % Of Dangerous Trips, Max Speed, Max Acceleration, Avg Speed, Avg Acceleration, Avg Gyro, Avg Accuracy, Avg Age, Avg Car Age, Avg Rating, Ranking of Ratings, % Of Males.

**WEEK 4**

**ETL Pipeline**

Boxplots as well as a function were made to return and remove outliers given a lower/upper fence magnitude, this way we can easily remove outliers in our ETL. Instead of running multiple cells to do our ETL, we have compressed all our ETL into a single cell (final cell of our notebook). Some techniques were used to also optimize our ETL process, such as changing the floating point of our datatype or using batch sizes to load our cleansed data, which will be shown below:

****

**ETL Code:**

Diagram Legend:

0 - Import modules & Configuration

1 - Extract

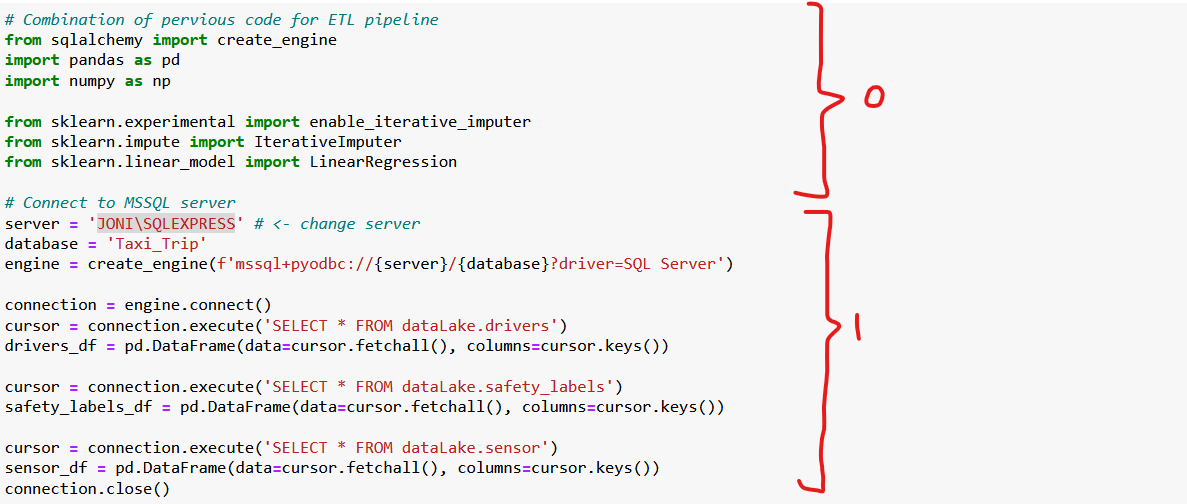
2 - Transformation: Changing Datatypes (Optimized Datatypes used) + Feature Engineering (Creating age, car age columns)

3 - Transformation: Removing outliers + Saving outliers

4 - Transformation: Imputation

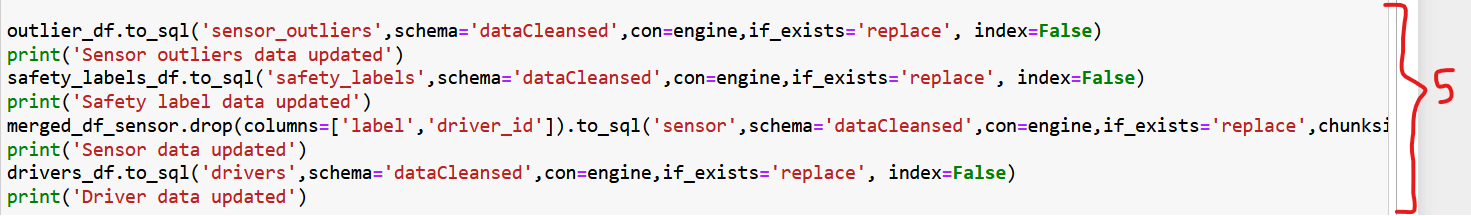
5: Load @ chunksize = 500000

6: Miscellaneous: Creating relations in cleansed schema

****

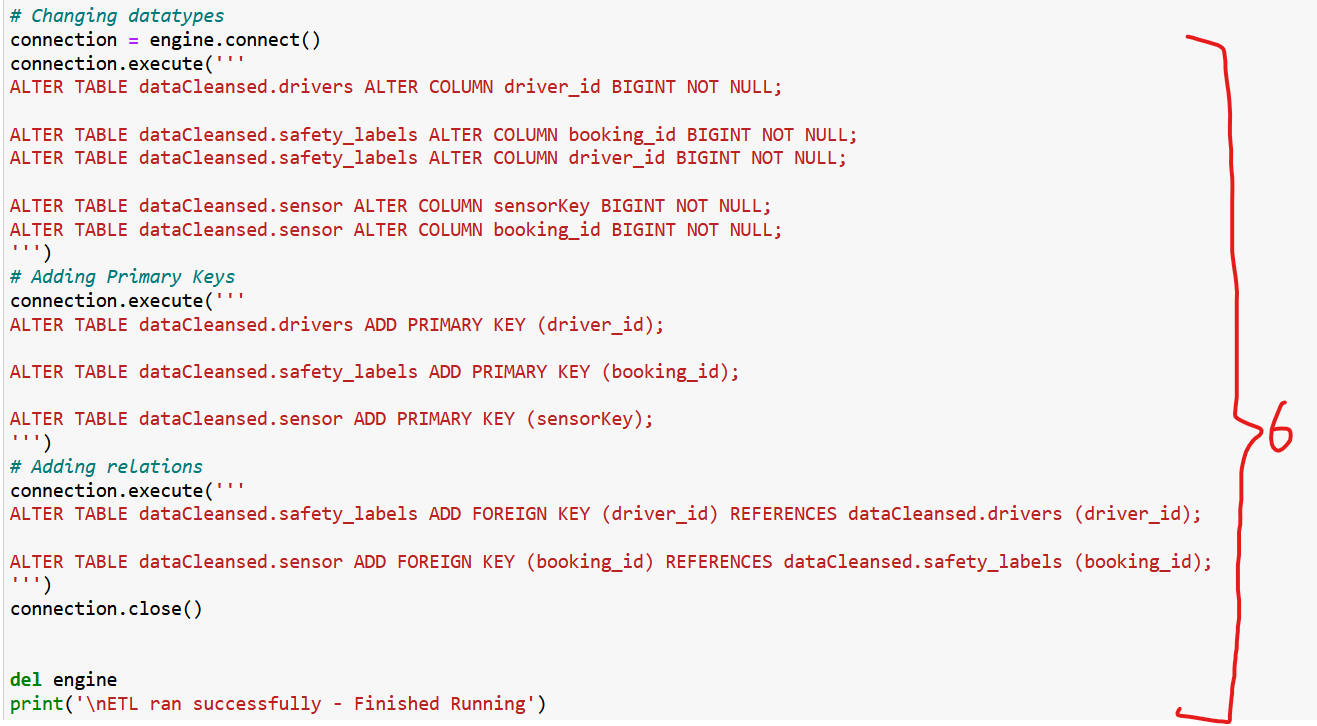
****

****

****

Chunk size for loading = 500000, otherwise it will attempt to load all 7.3 million sensor data into the DB, and our computers catches on fire.

****

****

**Final ERD of our DB:**

Graphical user interface, diagram

Description automatically generated

**WEEK 5**

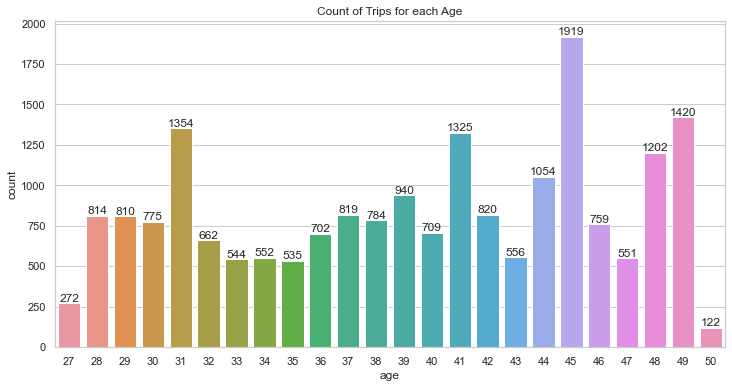
**Python EDA**

Wee Loon was tasked to conduct exploratory data analysis to gain more insights on the data. The more insightful charts that Wee Loon generated were the following:

Chart, bar chart

Description automatically generated

The total number of dangerous trips in this dataset is 4,993 and the number of safe trips is 15,007.



The age with the greatest number of trips made is 45 years old, having 1,919 total trips. Whereas the age with the least number of trips made is 50 years old, having only 122 total trips.

Chart, bar chart

Description automatically generated

The age group with the least percentage of dangerous trips made is 50 years old, having only 18% of the trips being marked as dangerous. The reason could be because of the lower number of total trips made by this age group. On the other hand, the age group with the highest percentage of dangerous trips is 33 years old, having 28% of the trips being marked as dangerous.

Chart, bar chart

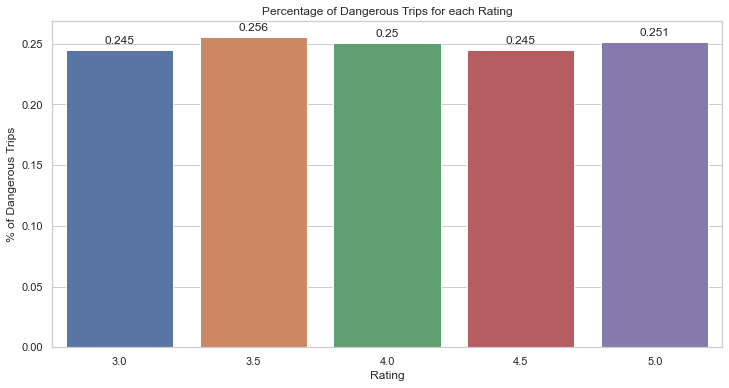
Description automatically generated

The car model with the greatest number of trips made is Toyota, having a total of 3,004 safe trips and 1,045 dangerous trips. The car model with the least number of trips made is BMW, having a total of 302 safe trips and 96 dangerous trips.

Chart, bar chart

Description automatically generated

Ratings that each driver has ranges from 3 to 5 in increments of 0.5. Most of the drivers (50) have a rating of 4. Whereas there are only 10 drivers with the lowest rating of 3 and 13 drivers with the highest rating of 5.



There seems to not be any pattern or relation between rating and percentage of dangerous trips. All ratings seem to have very similar percentages of dangerous trips of around 24.5% to 25.6%.

**WEEK 6**

**Tableau Sheets**

Wee Loon brainstormed and came up with general ideas of suitable charts that would be relevant for the overview of trips information. He then connected the SQL database to Tableau, after which he established the joins between the tables. The charts he generated were the following:

Graphical user interface, application

Description automatically generated

Two pie charts which displays the total percentage of dangerous trips overall. The second pie chart has filters applied to it, such that comparisons can be made between the two. The total percentage of dangerous trips made is 24.97% and the total percentage of safe trips made is 75.03%.

Chart, bar chart

Description automatically generated

A bar chart that displayed the percentage of dangerous trips for each car model, highlighting the top three car models with the highest percentage of dangerous trips. The average (24.79%) line shows that are 6 car models that have above average percentage of dangerous trips, while the other 4 car models are below average. The top 3 car models with the highest percentage of dangerous trips are Honda, Toyota, and Ford, with 26%, 25.8% and 25.5% respectively. On the other hand, the car model with the lowest percentage of dangerous trips is Chevrolet.

Chart

Description automatically generated with medium confidence

Upon deeper inspection of the information about Chevrolet, the total number of trips made by this car (Total Car Count) is 1,515, with 354 of them being dangerous trips. On average, this car achieves an acceleration of 9.767 and a speed of 9.1418.

Chart

Description automatically generated

Whereas for Honda, the total number of trips made is 1,477 and 385 of them are labelled as dangerous trips. On average, this car achieves an acceleration of 9.771 and a speed of 9.037.

Chart, histogram

Description automatically generated

A line graph across each second, that compares the selected information on the y-axis on both labels. This is done with a dynamic filter that allows users to choose the desired information they wish to display (e.g., avg. speed). Also allows for the user to filter down into specified driver information.

Background pattern

Description automatically generated

A column chart was made, with the help of custom SQL, to highlight the drivers with the highest percentage of dangerous trips. The top 3 drivers with the highest percentage of dangerous trips (34.06%) are Augusto, Reinoid (33.04%) and Joete (32.35%).

Chart

Description automatically generated with low confidence

Reinoid (Second highest % of dangerous trips) is of age 50 years old, driving a Honda car model that is 11 years old and has a rating of 4.5.

**WEEK 7**

**Tableau Dashboards**

Jayden designed the layout of the dashboard focusing on interactivity, multi granularity visualization, accurate values for our stakeholders (taxi drivers and managers). The completed dashboard is presented as shown:

Graphical user interface, chart, Excel

Description automatically generated

This is the final tableau dashboard. It consists of 5 main visualisations. (From top left to bottom right)

1. **Most Dangerous Drivers (by % of trips that are dangerous)**

**Graphical user interface, application

Description automatically generated**

1. **Most Dangerous Car Models (by % of trips that are dangerous)**

**Chart, bar chart

Description automatically generated**

1. **Dynamic Filter by Driver and Booking ID**

**Graphical user interface, text, application, email

Description automatically generated**This dynamic filter allows the user to click on a driver’s name, id or a booking ID to filter multiple graphs in the dashboard at once. It also allows the user to see the datapoints where booking ID and driver match. (eg. Clicking Tabatha would filter the other graphs to show records of that driver only)

1. **Overview of Trips over Time (Either by speed, acceleration, accuracy or gyro)**

**Chart, histogram

Description automatically generated**

1. **Pie Chart of Dangerous/Safe trips (One version affected by filters, other one is unaffected)**

**Chart, bubble chart

Description automatically generated**

**For the focus on interactivity, graphs 1-3 are made to be dynamic filters. This means when you click on an item in the graph, it filters the other graphs by that driver/car model/booking ID as well.**

**Insights from dashboard:**

**Chart, histogram

Description automatically generated**

**Graphical user interface, application

Description automatically generatedAugusto Most Dangerous Driver:**

From clicking the top most dangerous driver, we learn his name is Augusto, he drives a Hyundai car model and is of older age (49). 34% of his trips are dangerous. The car is also 12 years old. Despite this, his rating remains quite high at 4/5.

Chart, histogram

Description automatically generated **Overview of Trips in Speed: (All Drivers)**

We can see their average speed in their ‘Dangerous’ trips are peaking higher than the safe trips more often than not.

**Graphical user interface, application, bar chart

Description automatically generated**

Chart, line chart, histogram

Description automatically generated

Filtering by top 3 most dangerous by clicking on the top 3 drivers

**Overview of Trips in Speed: (Top 3 Dangerous)**

When we filter by the top 3 most dangerous drivers, we can see that in contrast to the comparison with all drivers, the differences between ‘safe’ and ‘dangerous’ trips are much smaller. This means that the drivers’ behaviours are more consistently dangerous and the ‘safe’ trips are actually very close to being ‘dangerous’ ones.

**Chart, line chart

Description automatically generated**

**Overview of Trips in Acceleration:**

With filter for top 3 dangerous drivers again, we can see that average acceleration (in total magnitude) had higher peaks and troughs. Indicates that acceleration can be a significant factor too. This is because the drivers are either increasing or decreasing their speed at a faster rate. If decreasing speed, this could mean the drivers are jam-breaking, which is a very dangerous move that has been shown to lead to accidents in many situations. [2]

Chart, histogram

Description automatically generated **From Overview of Trips over time: (Gyro)​**

This time we filter by Top 3 most dangerous car models (brands)​

Magnitude of gyro much higher on average for dangerous trips. This means the drivers are rotating at a very fast rate. This can lead to accidents from losing control of the car when turning or by crashing into others.

Diagram

Description automatically generatedChart, bubble chart

Description automatically generatedWhen filtering for drivers aged 37-50​, % of dangerous trips increased from 24.97% to 33.60% (almost 10%)!​

This could mean that more senior drivers are driving more recklessly​

**Conclusion**

In conclusion, our team used SCRUM techniques to understand our users as well as the end products they require. We used MSSQL to create a database that can be easily maintained by Just Taxi’s IT administrators. We created an ETL pipeline to clean our data and we used Tableau to visualize big data with an interactive dashboard. This powerful dashboard will allow Just Taxi’s managers to seamlessly analyse their taxi data with no downtime, the pipeline will ensure that future data is able to be cleaned and visualized easily too. With this, they can make more informed decisions to ensure the safety of taxi drivers and passengers.

**Project Documentation** **End (Phase 1)**

References:

1. <https://www.sciencedirect.com/science/article/pii/S2185556022000025>
2. <https://mothership.sg/2022/03/yishun-driver-swerve-cyclists/>