Taxi Safety Data

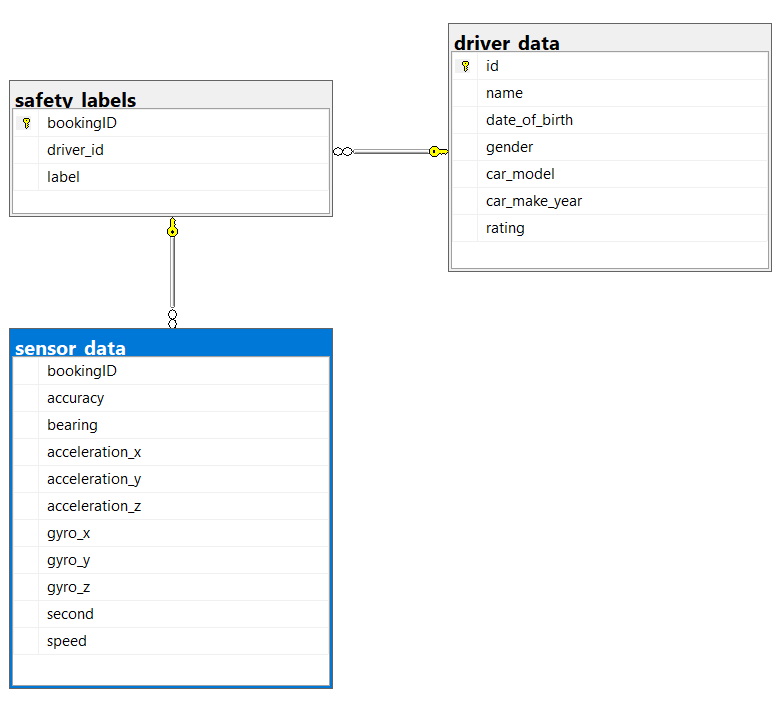
ST1508: PRACTICAL AI

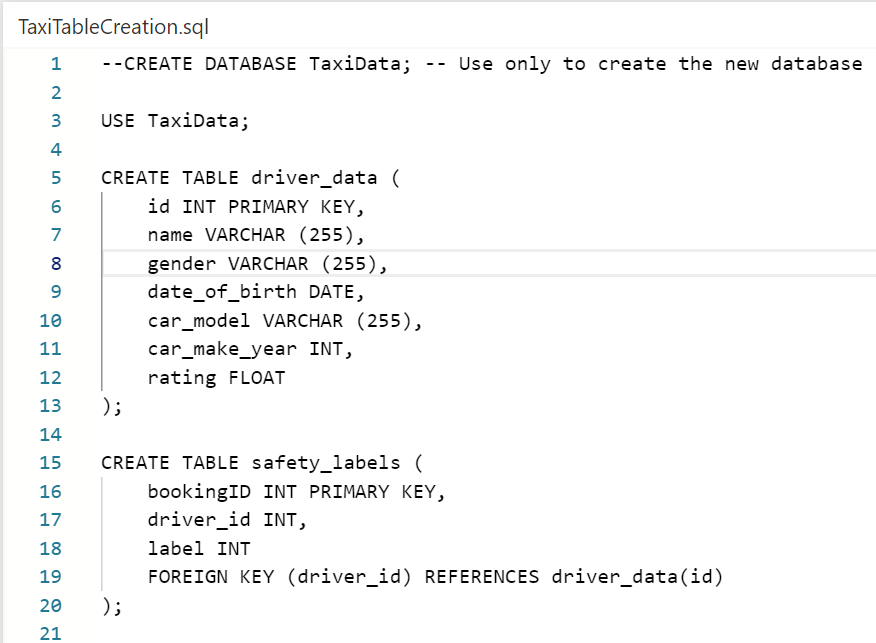
CA1 Assignment

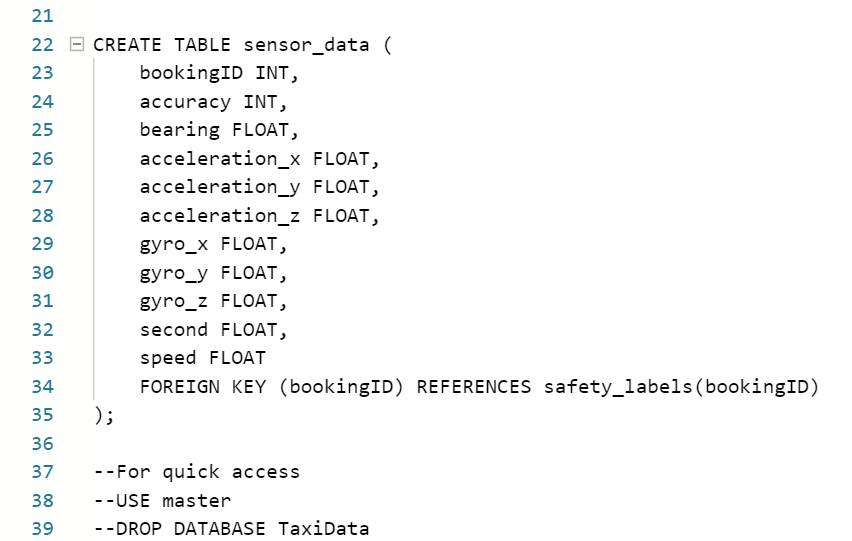
Pratik Ghosh (P2100762)

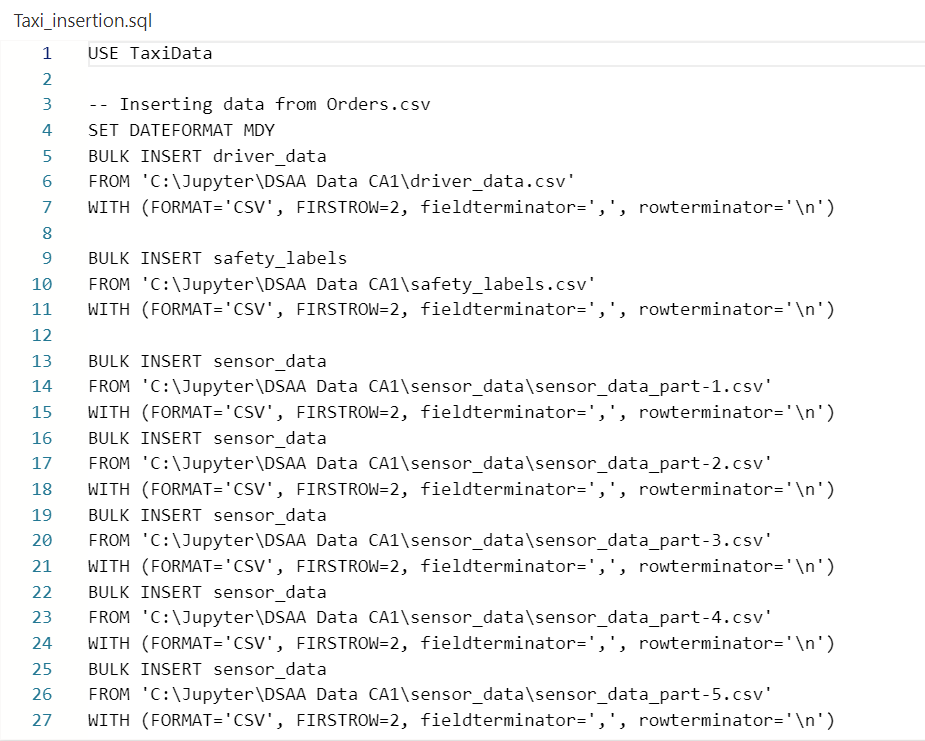
Kok Jeng (P2112985)

Haaris Bin Sulaiman (P2112815)

1. Data engineering (SQL database)
   1. Database Diagram  
      
   2. Database Setup







Note: For our foreign keys, we have to ensure they are the same type: Booking Id, Driver id and id are INT. Safety\_labels acts as a link between driver data and sensor data, because safety\_labels contains both their ids

* 1. SQL Queries

Query 1:

SELECT s.bookingID, ROUND((MAX(second)\*AVG(Speed))/1000, 2) 'Estimated Total Distance based on highest time recorded', ISNULL(AVG(s.acceleration\_x),0) 'Avg Acceleration X for each trip', ISNULL(AVG(s.acceleration\_y),0) 'Avg Acceleration Y for each trip', ISNULL(AVG(s.acceleration\_z),0) 'Avg Acceleration Z for each trip', ISNULL(AVG(s.accuracy),0) 'Avg Accuracy for each trip', DEGREES(ISNULL(AVG(s.gyro\_x),0)) 'Avg Gyro X for each trip', DEGREES(ISNULL(AVG(s.gyro\_y),0)) 'Avg Gyro Y for each trip', DEGREES(ISNULL(AVG(s.gyro\_z),0))'Avg Gyro Z for each trip', (SELECT TOP 1 L.label FROM safety\_labels l WHERE s.bookingID = L.bookingID ORDER BY L.bookingID Asc) 'Labels' FROM sensor\_data s

GROUP BY s.bookingID, s.bearing

Table

Description automatically generatedGraphical user interface, application

Description automatically generatedORDER BY 'Estimated Total Distance based on highest time recorded' desc

**Insights from data:**

The total distance travelled for each trip is found, along with the sensor readings. We see that for readings which are dangerous trips, the acceleration for Y is mostly very high. Studies have shown that speeding is a factor of at least 25% of accidents. The high acceleration readings show that when speeding was high, the taxis seem to be in unsafe trips more.

**Recommendation:**

For unsafe trips with longer distances, the sensors seem to vary drastically compared to the once with shorter distances. This could be due to the driver behavior changing throughout the trip, or the car malfunctioning during the trip which lead to the unsafe trips. As for some of the trips which are long, the sensor readings are not that varying, which means that the cars were working fine or the driver was maintaining the trip pacing.

Query 2:

Table

Description automatically generatedSELECT bookingID, ROUND(Speed,2) 'Speed (m/s)', ROUND(acceleration\_x,2) 'acceleration\_x', ROUND(acceleration\_y,2) 'acceleration\_y', ROUND(acceleration\_z,2) 'acceleration\_z', ROUND(gyro\_x,2) 'gyro\_x', ROUND(gyro\_y,2) 'gyro\_y', ROUND(gyro\_z,2) 'gyro\_z', label FROM(SELECT s.bookingID, Speed, acceleration\_x, acceleration\_y, acceleration\_z, gyro\_x, gyro\_y, gyro\_z, l.label, RANK() OVER (PARTITION BY s.bookingID ORDER BY Speed DESC) AS ranking FROM sensor\_data s, safety\_labels l WHERE s.bookingID=l.bookingID AND l.label=1 GROUP BY s.bookingID, Speed, acceleration\_x, acceleration\_y, acceleration\_z, gyro\_x, gyro\_y, gyro\_z, l.label) AS subquery WHERE subquery.ranking=1

**Insights from data:**

The query shows the average values of speed, acceleration, and gyro for the dangerous trips. It is different from the first query and instead of comparing the trips, it shows us the general trend of the gyro sensors and acceleration readings when the trips are dangerous. The acceleration results above shows that when a vehicle’s speed is at the highest, acceleration x have both negative and positive values while acceleration y and z has mostly positive values. Thus, suggesting acceleration x does not affect speed but acceleration in the forward and upward direction does. For gyro, at high speed, the taxis seem to rotate a lot.

**Recommended:**

When there is a lot of fluctuation in the gyroscope readings and when acceleration y and z is high, the taxi might have chances of leading to an unsafe trip. Therefore, the company could take this into account when classifying a taxi trip as dangerous or safe.

Query 3:

Graphical user interface

Description automatically generated with medium confidenceselect\*from (select\* , Rank() over (Partition BY d.car\_model ORDER BY d.[No.of Dangerous trip] DESC ) AS Rank from (select s.driver\_id, name,d.gender,d.car\_model, d.rating, count(\*) as "No.of Dangerous trip", rs.[Dangerous Trips Per Car Model] FROM driver\_data d, safety\_labels s, (SELECT d.car\_model, COUNT(\*) as "Dangerous Trips Per Car Model", Rank() over (Partition BY d.car\_model ORDER BY COUNT(\*) DESC ) AS Rank FROM driver\_data d, safety\_labels s WHERE s.label = 1 and d.id=s.driver\_id GROUP BY d.car\_model) rs WHERE s.driver\_id = d.id and s.label=1 and rs.car\_model = d.car\_model and RANK <= 1 GROUP BY s.driver\_id, d.car\_model,d.name,d.gender, d.rating,rs.car\_model,rs.[Dangerous Trips Per Car Model] --order by "Dangerous Trips Per Car Model" desc, "No.of Dangerous trip" desc) d) e WHERE e.Rank <= 3 order by e.[Dangerous Trips Per Car Model] desc,e.[No.of Dangerous trip] desc, e.rank;

**Insights from data:**

In this query, we compare the top three/four drivers for the top five companies which have dangerous trips. This way, we can see which brand or drivers are more generally unsafe. Here, we see that Toyota seems to be a company with the highest number of dangerous trips. The drivers with this car brand seem to be having relatively high number of dangerous trips as well.

**Recommendation:**

Hence, the business company can take into consideration whether these drivers should be sent for retraining, and if the specific car brand company should have their car safety regulations reinspected, while they can consider to continue using the ”safer” brands like Nissan or BMW, since they have drivers with high ratings and also low dangerous trip rates.

* 1. User Stories
     1. Manager  
        As a manager, i want to see which people and cars are the safest for our company to have, so I can have a better business and reputation.  
          
        As a manager, I need an intuitive application with a user interface so that I can visualize, analyze and compare taxi-driving data such as the telematics data on trip safety and make informed decisions.
     2. Administrator  
        As an Administrator, I want to be able to make insights regarding data so that I know which features of the data are important to focus on.  
          
        As an Administrator, I want to see clean and precise data in my database, so that I will be able to navigate, explore and make insights freely.
     3. Taxi Driver  
        As a Taxi Driver, I want to know which car company brand is safest for me to drive in, so that my passenger and I can have less worries about crashing.

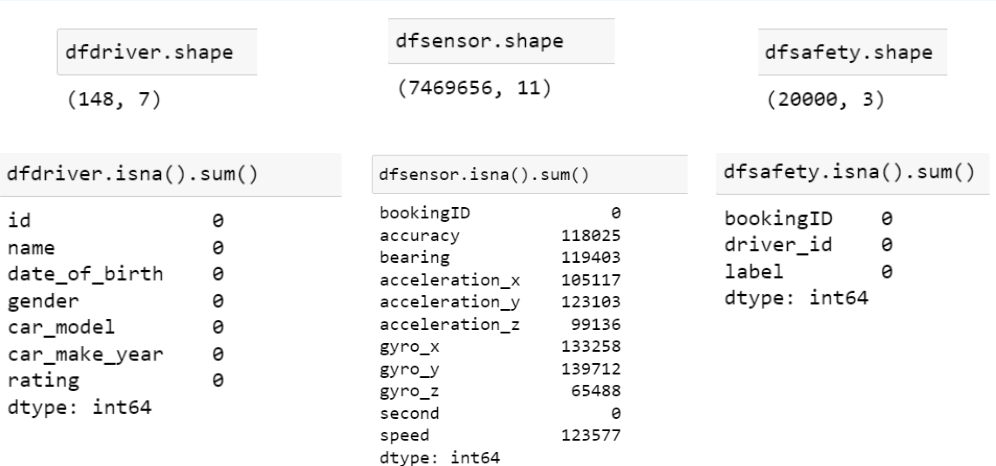
1. ETL Pipeline
   1. Transfer Data to Jupyter Notebook

In the jupyter notebook, we make use of SQLAlchemy to fetch the data for driver\_data, safety\_labels and sensor\_data from MySQL. A pandas DataFrame is made for each of them so that we can use for them for data visualization later on. We assign a simple dataframe name for each of the dataset. dfdriver:driver\_data, dfsafety:safety\_labels, dfsensor: sensor\_data 

* 1. Data Handling

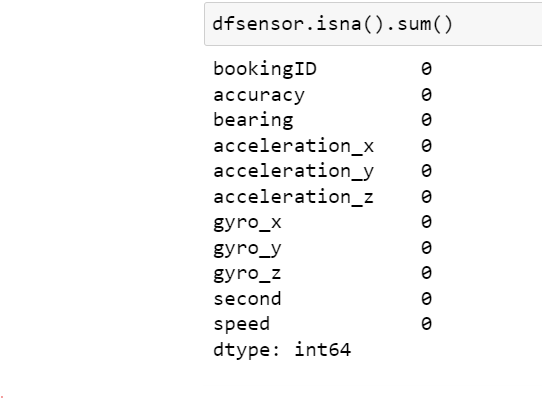
2.2.1 Data Exploration

Now that we have DataFrames, we explore all 3 DataFrames to know more about and understand the data that we are handling better. We can see that dfdriver and dfsafety have no null values whereas dfsensor has so many null values and will require cleaning. Here we also see the number of rows and columns for each of the dataframes and we can see that dfsensor is a really huge dataframe and such huge datasets will expectedly have null values.



2.2.2 Data Cleaning

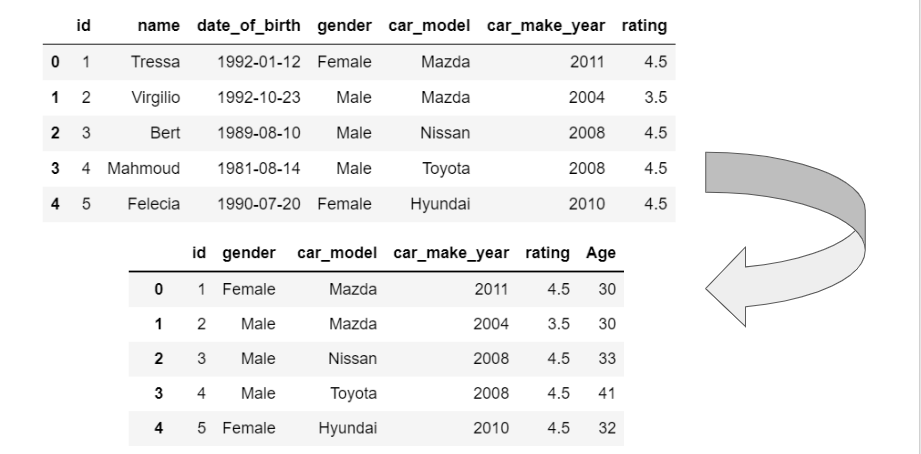
We filled all the null values with the mean value for each column in the dataframe. Now we longer have any null values and can proceed to data preprocessing.



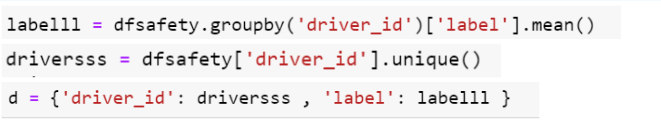
2.2.3 Data Preprocessing

We now have to add and drop columns and make new dataframes so that we can have dataframes that we can actually utilize to produce data visualization later on. Firstly, for dfsafety we will use datetimr and make a function to make a column for the age of the driver so that we can drop the ‘date\_of\_birth” column. Having a column with integers is much easier and efficient to use for data visualization



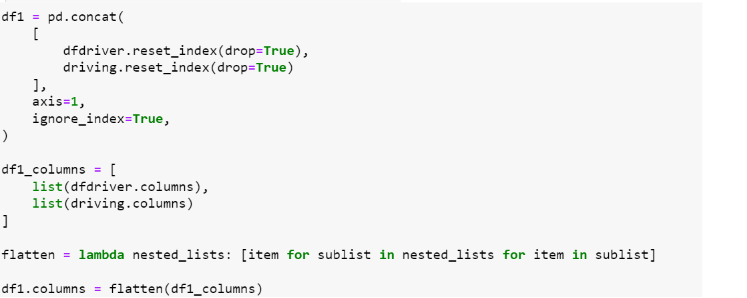


dfsafety contains many repeated driver\_id and so we make a dataframe called driving where we only store the unique driver\_id and group them by the label.

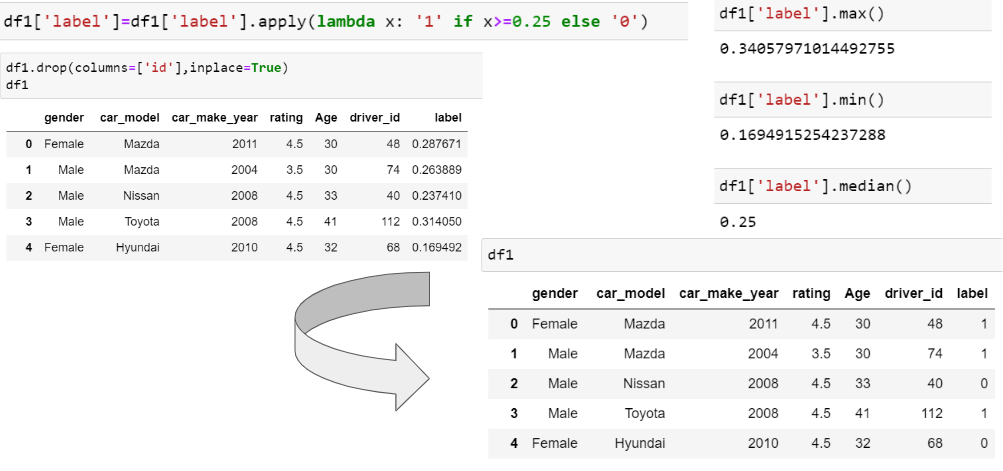




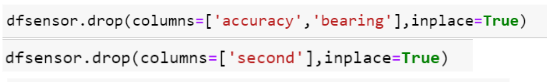
Then we concat driving and dfdriver to make a dataframe called df1



We have an issue. Since we grouped by mean, the label will be in decimal form instead of integer and so we use a lambda function to check how to assign the label. Since the median is 0.25, any label bigger than or equals to 0.25 will be assigned as ‘1’ and any label smaller than 0.25 will be assigned as 0. The labels in this context is how dangerous the trip was with 1 being dangerous and 0 being a normal trip. Hence, it makes sense to assign trips with a decimal value higher than the median as ‘1’. df1 is now ready. The purpose of df1 is so that we can now see the label and compare and use it for visualization later.

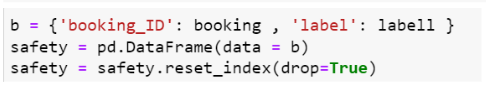


For dfsensor, we drop the ‘accuracy’, ‘bearing’ and ‘second’ columns as we don’t need them and they are redundant.



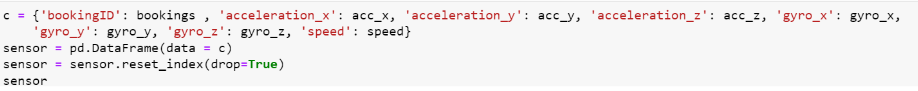
We made a dataframe calleed safety where for the unique booking ids and the labels for the boking ids.



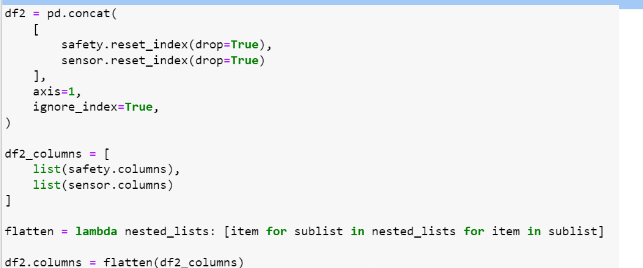


We make another dataframe called sensor which stores the unique booking ids and the mean value of each column grouped by the unique booking ids.





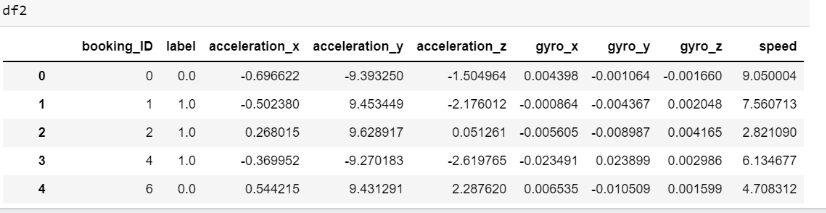
Then, we concat sensor and safety to make a new dataframe called df2



We drop bookingID because there are 2 of them.



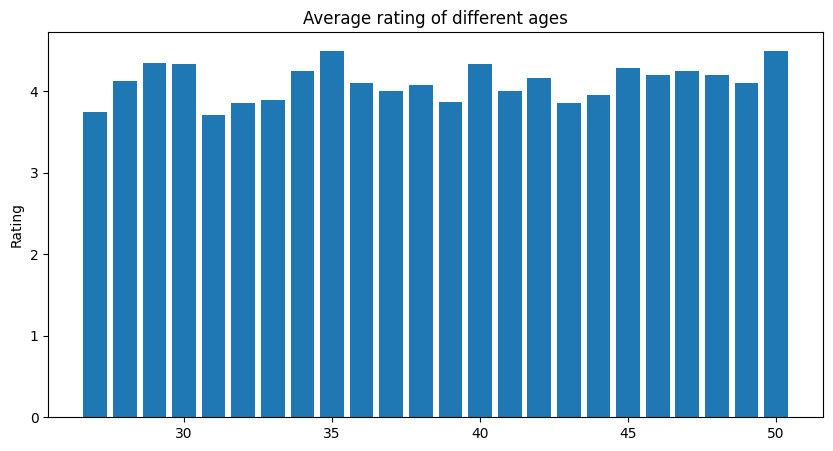
Df2 is now ready. The purpose of df2 is so that we can make sense to we how the different dynamics can affect the label for each trip, making it a dangerous or normal trip.

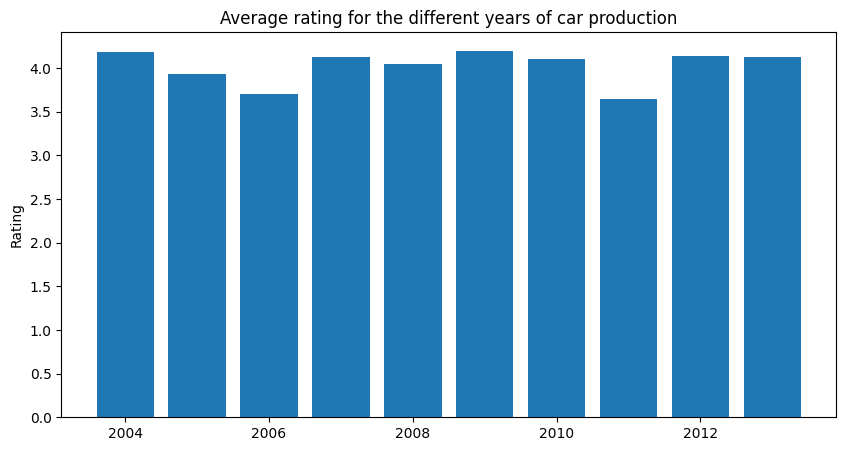


Now we will only focus on df1 and df2 because they are all we need to do visualizations.

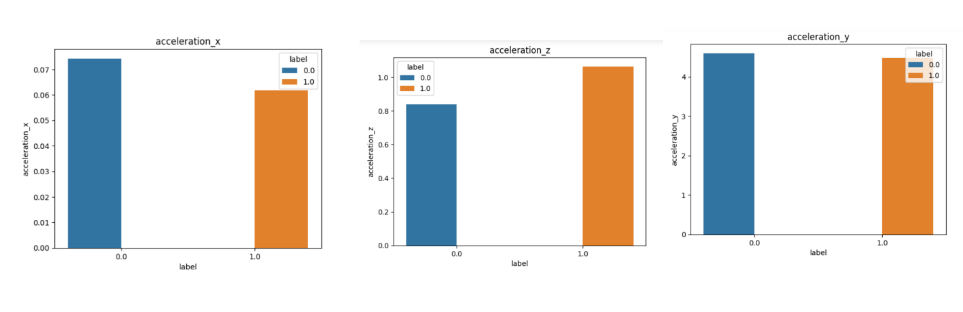
1. Data Visualization
   1. Python EDA

These are some of the graphs that we did. The graph below shows the average rating of different ages. This was done using df1. We can see that there is actually no trend between the rating and the age of the driver as well as between the year of car production and the rating since the bars are fluctuating.



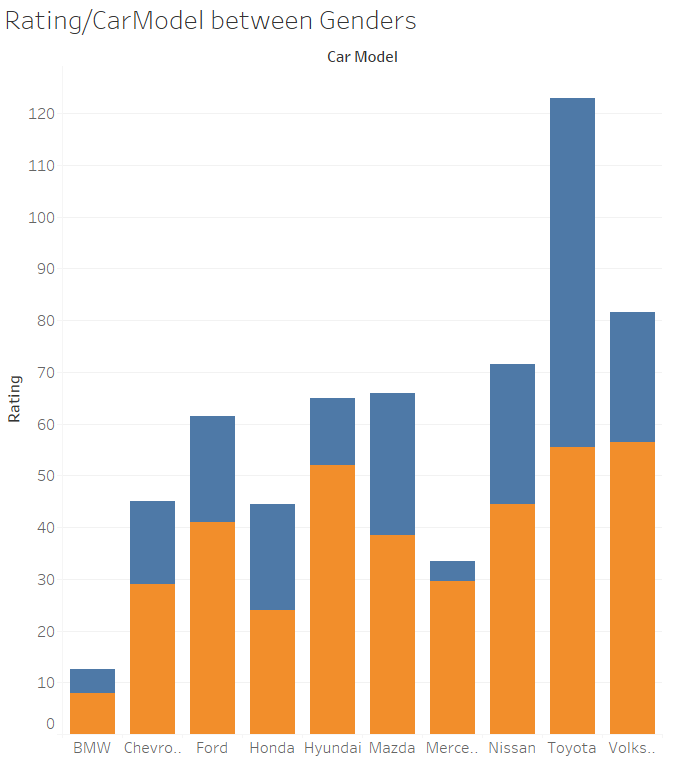


We can also see that the dynamics like the acceleration for each axis affects the label meaning that the acceleration for each axis affects how dangerous a trip is.

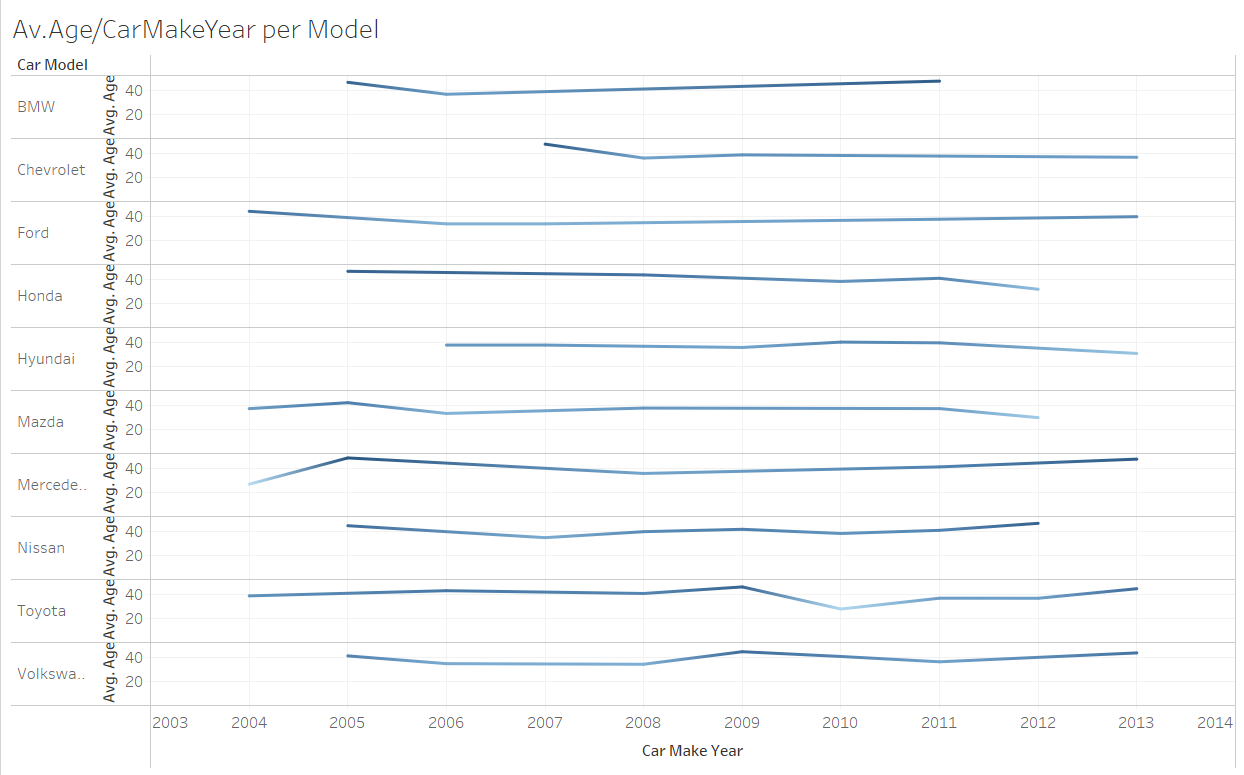


Now we will proceed to Tableau where we can do better and more detailed data visualizations. The dataframes df1 and df2 are used for tableau. We exported and renamed df1 and df2 as driver.csv and sensor.csv respectively.

* 1. Tableau Dashboard
     1. Driver Data

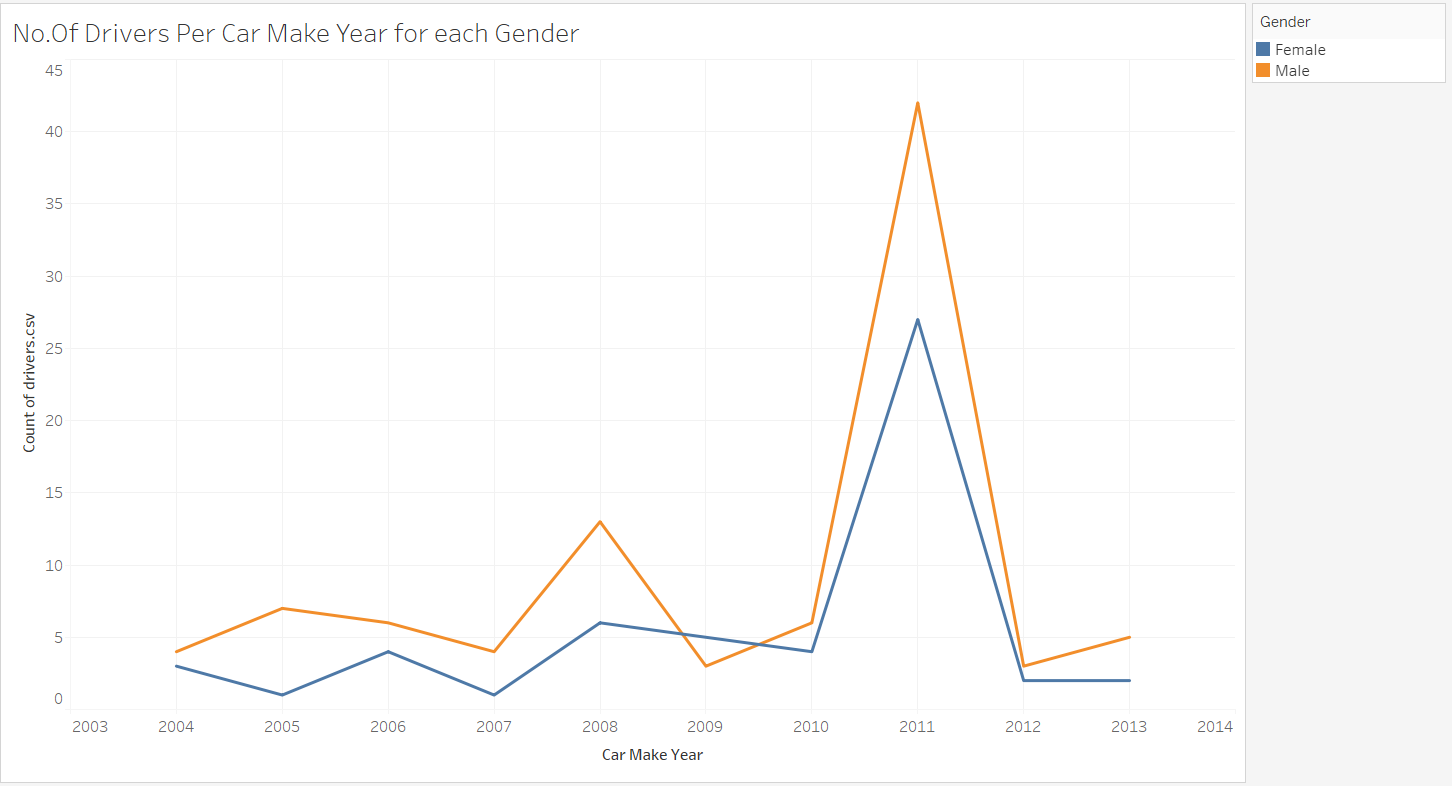


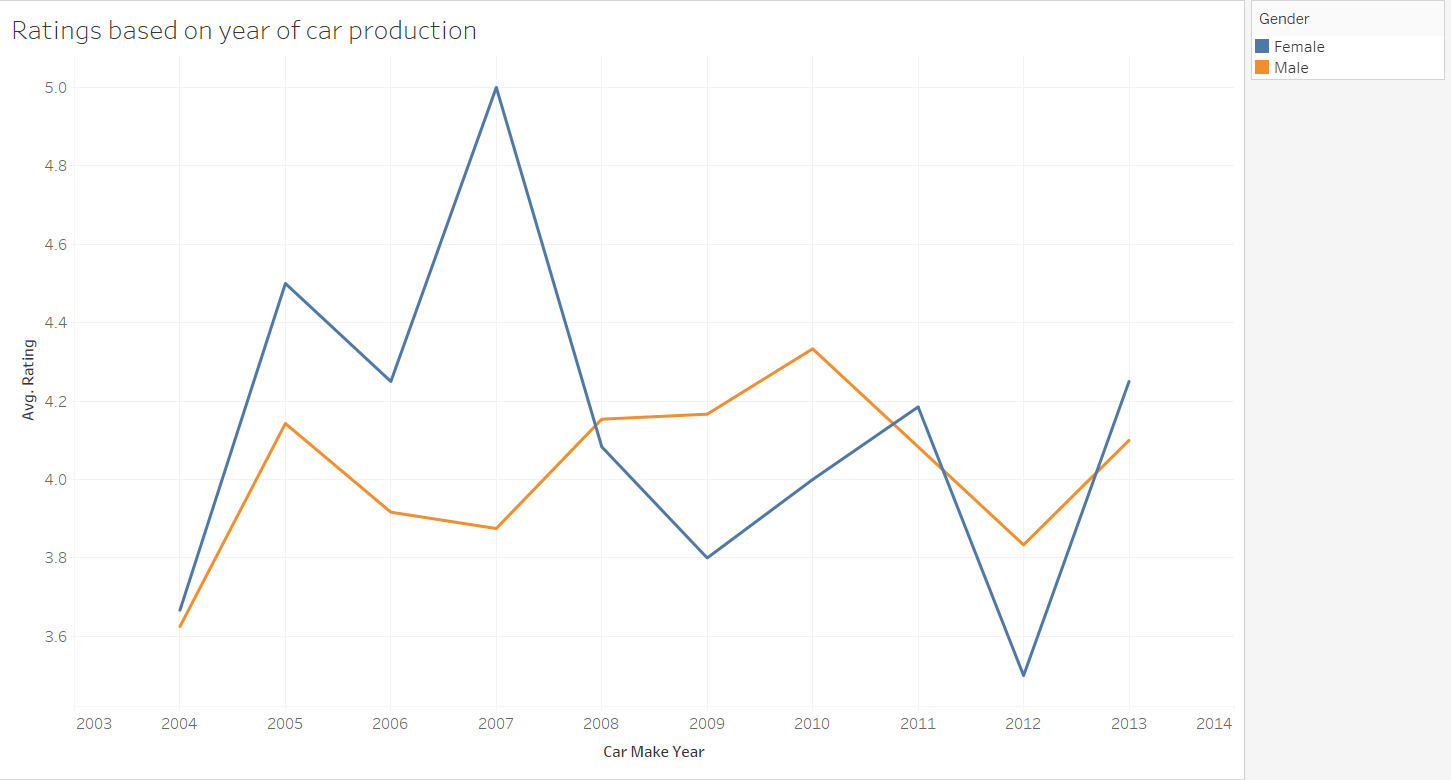
Insights: This graph shows us the average rating of each car brand, and the proportion of it by gender. We can say that Toyota has the highest average rating out of all the other car brands. Along with how the male gender has higher overall ratings compared to females.



Insights:

This chart shows us the average age of the drivers throughout the Years their car was made, for each brand. From here we can deduce that the Mercedes, Toyota and Honda brand has the old users(Past 40), which can mean that their performance may be affected, leading to higher or lower ratings as we compare to the other graphs, in the previous graphs, the mentioned brands seem to have a lower overall rating.



Insights:  
Here we can see the overall count of drivers for their respective car make year. We can see here that in 2011, there was a sudden rise in the count, meaning that drivers that were recorded tend to buy cars made in 2011. Additionally, we see that there tend to be more male buyers than female buyers for all years except 2009.  
Insights:  
Here we have the average rating of each gender for their car make year specifically. Following up from the previous graph we can look at their correlation, the highest average rating was 2007 with a rating of 5, however, we see from the previous graph that this is just caused by the low amount of female drivers that bought a 2007 car. But another type of reasoning can be used for the 2012 lowest rating case, whereby there was a huge drop in people buying 2012 cars, perhaps we can say that there was a defect or design flaw that made 2012 cars less likable. We can say the same for the 2004, whereby the rating may be caused by the age of the car and its defects.

* + 1. Chart, scatter chart

       Description automatically generatedSensor Data

Insights:

This chart displays the distribution of all the booking trips, labelled by the safety (orange is unsafe, while blue is safe). The chart shows that the unsafe trips are generally anomalies for some of the sensors, such as the gyro Z and the acceleration Y. This tells us information that the driver may have trouble with steering, as the data points are for maneuvering.

Chart, box and whisker chart

Description automatically generated

Insights: The next chart gives us information about the average speed for the booking trips, again labelled by safety as we compare the speed for these. We see that, the safe trips have a higher average range of speed, which shows that maybe speeding is not the only factor of the accidents. This could mean that it was other factors such as the driver maneuvering incorrectly, or external factors not in their hands.

A picture containing chart

Description automatically generated

Insights:

This chart compares the acceleration data, and it shows that the unsafe trips generally had higher acceleration for the Z axis. From this detail, the we see that the Z-axis also had a significant role for the Gyro sensing, which means the companies can look into the problems arising with this high Z-axis readings.

Chart, waterfall chart

Description automatically generated

Insights:

This chart compares the average Gyro sensing for the trips. We see that the Gyro X had a significantly high absolute value compared to the other gyro sensings, both for non-dangerous and dangerous trips. From this we can infer that the gyro sensings have a role to play in determining the safety of the trip, as the Gyro Y sensing for the unsafe trips had a higher absolute value than those for safe trips.

# **References**

[1] <https://iopscience.iop.org/article/10.1088/1757-899X/100/1/012017/pdf>