**Practical AI Phase One**

**Project Report**

**Module Code:** ST1508

**Class:** DAAA/FT/2B/05

**Members:**

Ryan Ueda Teo Shao Ming (2122924)

Chia Wee Leong (2122883)

Nimmagadda Vineesh (2123165)

Table Of Contents

**Introduction:** What Our Project Is About 1

**Background Information:** Dataset 1

**Section 1:** Framework 1

**Section 2:** User Stories 2

**Section 3:** Data Storage 3

**Section 4:** Complex SQL Queries & Data Understanding: 3

Raw Data

**Section 5:** ETL Pipeline 6

**Section 6:** Exploratory Data Analysis 7

**Section 7:** Data Processing & Wrangling 9

**Section 8:** ETL Pipeline For Storage Of New Data 10

**Section 9:** Complex SQL Queries & Data Understanding: 10

Cleansed Data

**Section 10:** Advanced Data Analysis & Visualisation 11

**Section 11:** Insights & Recommendations 14

**Section 12:** References 15

**Introduction**

In Phase One of this project, the objective is to build an intuitive and interactive dashboard to visualize and analyze taxi data, to provide meaningful insights in the bigger picture of the ride-hailing operation.

This includes and is not restricted to: creation of database, storing of data in SQL Server, setting up of data pipeline between database and python, as well as creation of an interactive dashboard for in-depth visualization of data.

**Dataset**

The data for this project consists of 3 tables of data, presented in a csv file format. The first table, sensor data, consists of roughly 7.4 million rows of data and 11 columns. The second table, driver data, consists of almost 150 rows and 6 columns. Lastly, the safety data consists of 20000 rows and 3 columns. In Exploratory Data Analysis, we will dive deeper in exploring and understanding this data.

**Framework**

For this project, we have decided to adopt the CRISP-DM data science process model. CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is an industry-proven way to guide data mining efforts.

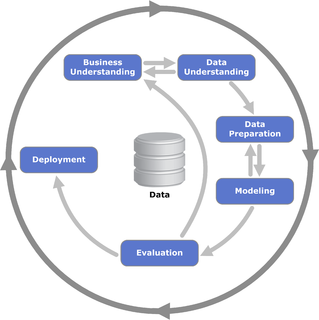


Fig 1: Diagram Of CRISP-DM Process

This framework primarily consists of 6 phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Development.

There are many benefits to adopting this approach, such as its easy-to-implement nature, allowing usage without the need of much training or organizational role changes.

Its flexibility also results in the large plethora of benefits that come along with agile principles and practices. This allows users to cycle and iterate through steps and processes, gaining deeper understanding and knowledge of the problem domain and tasks at hand. It is also heavily generalize-able, and applicable in many instances involving data science, despite the process being designed for data mining.

Iterative CRISP-DM process fulfills the experimental and open-to-new-ideas nature of data science projects, especially at early stages of the project. However, the project is likely to be stuck in between never-ending phases of idea-implementation-testing (Abdollazadeh, 2021).

This brings us to the use of Scrum with CRISP-DM, where each sprint utilizes all CRISP-DM phases but focuses on one or two of them. For example, the first sprint likely focuses on business and data understanding, where the business understands its core business needs, current practices and current performance. The next sprint will then focus on data preparation, where the team is focused on finding the best model specification to solve the specific task in hand. Every sprint has scheduled review and collaboration sessions with relevant stakeholders, which focuses on the evaluation aspect of CRISP-DM. This leads to a final result that is understood, accepted and owned by the project stakeholders, ready for deployment (Thurber, 2020).

**User Stories**

Before starting the project, we first assess and identify the user stories.

User stories are crucial as they help gain a keen sense of the specific project requirements that are to be followed by the team.

This helps simplify the essential core of the project and fundamentals needed to complete the project. At the same time, it allows the team to measure the progress and development of the project and keeps the team focused. User stories align with the agile framework, making planning sprints and organizing tasks with regard to their priority much easier and simpler. Effective and well-written user stories should cover the ‘Who’, ‘What’ and ‘Why’ of the project objectives.

Following the understanding of user stories, we have come up with user stories for the three target audience of our project: Managers, Administrators and Drivers.

**User Stories**

| **Manager** | **Administrator** | **Driver** |
| --- | --- | --- |
| 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. | As an administrator, I want to create a sql database so that I can store data more effectively. | As a taxi driver I want to see how safe my driving is so that I can find out how to improve the riding experience for my clients. |
| As a manager i need to know the safety of my drivers performance so that i can provide appropriate training such that clients can have a safe and comfortable riding experience | As an administrator, I want to create a SQL-Python pipeline so as to allow for efficient data analysis. | As a taxi driver I want to see my driving metrics in an interactive manner so that I can better visualize them. |

**Data Storage**

To store the large amount of data we have, the data will be stored within SQL Server.

Not only is SQL Server an effective way to store large amounts of data, it allows for the execution of complex SQL queries to help gain a better grasp of the data at hand.

In addition it makes it easy for a data pipeline to be established between the database and python, it grants the ability for efficient data processing and further data exploration and analysis, further extending to processes such as machine learning. To do this, we load the data from the CSVs into the three tables in the database. We are then able to execute some complex SQL queries to gain some insight into the data.

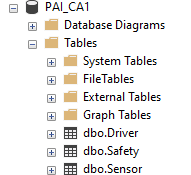


Fig 2: SQL Server Database Overview

**Complex SQL Queries and Data Understanding: Raw Data**

After populating the SQL table with the data from .csv files, we shall write some SQL queries to understand our data better.

**Query 1: Maximum change in acceleration readings**

For our first query, we shall explore the theory of roll-pitch-yaw.

Roll is the rotation of the car about the longitudinal axis. It is also referred to as the side-to-side motion of a car about the longitudinal axis passing through the center of gravity. When there is a sharp turning maneuver, roll shifts the weight of the car to the left or right, placing a load on the outermost tires. However, if roll is excessive, the car swings into a drift which may cause the vehicle to topple over.

Pitch is the rotation of the car about the transverse axis. It is also referred to the front and rear motion of a car about the transverse axis passing through the center of gravity. When a car decelerates, the weight of the car is transferred onto the front wheels, causing the body of the car to lean forward. Similarly, the weight of the car is transferred onto the rear wheels when a car accelerates.

Yaw is the rotation of a vehicle about the vertical axis. It happens in response to cornering and sometimes in response to side wind. It is also the left to right motion of the nose of a car on its vertical axis, through the center of gravity.

In this query, we shall explore the maximum change in accelerometer reading relative to its previous recorded value, based on the car model and its label (*Beginner's Guide to IMU*).

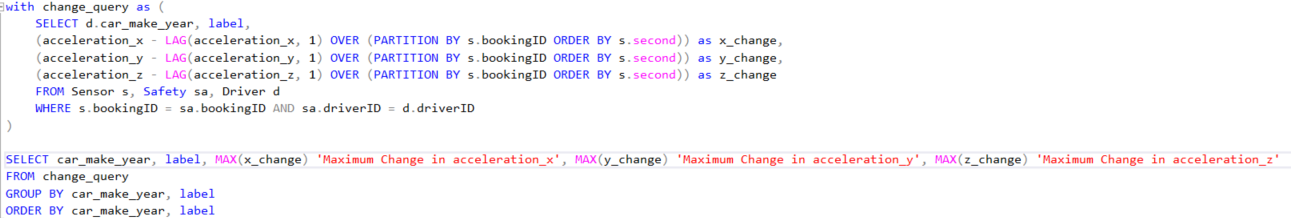


Fig 3: SQL Query 1

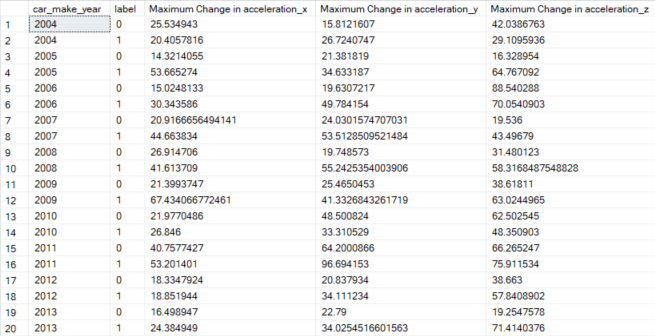


Fig 4: Output of Query 1

As expected, rides that are more likely to be deemed unsafe tend to have a higher maximum change in acceleration in either direction, due to the likelihood of the car crashing or toppling.

Out of the 3 years (2004, 2006, 2010) where the maximum change in acceleration in either angle was greater in safe driving as opposed to unsafe driving, maximum change in acceleration x was lower in safe driving only in 2004. This can be due to the fact that roll is much more dangerous than pitch and yaw, as exerting weight on the area with a larger surface area as compared to pitch has a much greater likelihood for accidents to occur.

**Query 2: Accuracy and rating based on car model and label**

In this query, we shall explore if accuracy and rating is affected by a car’s model and their labeling.

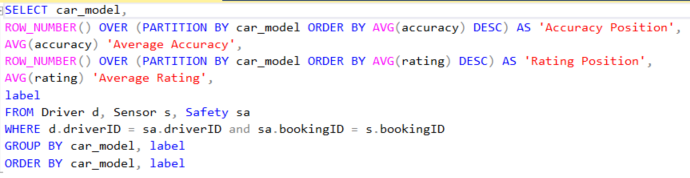


Fig 5: Query 2

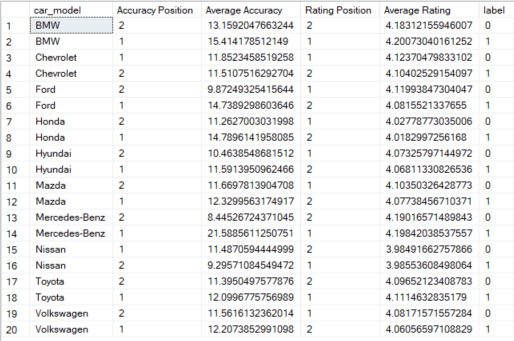


Fig 6: Output of Query 2

Based on query 2’s outputs, almost all safe trips have a lower mean accuracy than their unsafe counterparts, except Nissan and Chevrolet cars. This may be due to the fact that GPS directions may not always lead to the safest routes, and following it too much can lead to driving in areas that are accident-prone ( Dalli, 2022). The reason why Nissan is an anomaly may be due to constant updates by them to find efficient routes that are less dangerous to navigate (Nissan, Why update your navigation system map?).

Out of 10 car models, 4 have a higher rating when the car is driving unsafely. This is expected as safer rides tend to be more comfortable as a passenger, leading to better ratings by passengers. The 4 with a lower rating when the ride is unsave have a relatively small difference of 0.02 as well, and this can be explained by skilled drivers who are able to drive cars well enough that it is unsafe for other drivers.

**Query 3: Deviation of gyroscope reading and speed based on rating**

In this query, we shall explore if passenger rating and safety label has an impact on car’s gyroscopic readings and speed.

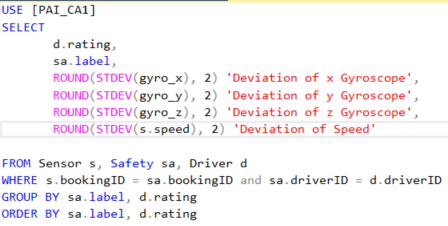
****

Fig 7: Query 3

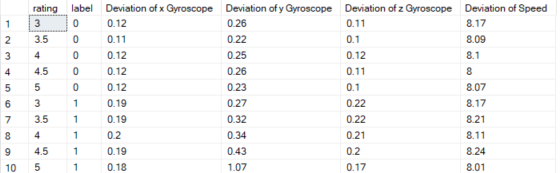


Fig 8: Output of Query 3

Based on query 3’s results, rides that are deemed safe have a higher standard deviation of both gyroscopic readings as well as speed.

This is expected as cars that drive safely tend to have a constant speed and acceleration, so as to reduce the number of G-Forces experienced when a car rotates sharply or has a sudden change in speed.

Within the same labels, cars have similar values for gyroscope deviation and speed deviation. This shows that customer ratings are affected by external factors such as the driver’s friendliness and the time taken to hail a taxi.

**ETL Pipeline**

To handle the data after it is stored in the database, we will follow the Extract-Transform-Load (ETL) process. The ETL process is a three-phase data integration process that combines data from multiple sources into a single consistent data store, allowing for cleansing of data to improve data quality and establish consistency within the data. This includes and is not limited to: Filtering, Cleansing, De-duplicating, Validating, Formatting, Translating and Decrypting.

In order to do this, a data pipeline is established between the SQL Server database and python through the help of SQLAlchemy. After establishing a connection to the database, data can then be extracted into python where it is stored in pandas dataframes, ready to be processed and analyzed.



Fig 3: SQLAlchemy Code For Data Pipeline

**Exploratory Data Analysis**

Upon extracting the data from the database, the tables are stored into three pandas dataframes.

A quick check shows that there are quite a large number of missing values within the sensor data, while the driver and safety data do not have any missing values. In addition, some rows contain up to 4 columns consisting of NULL values.

Furthermore, there appears to be nonsensical values in the sensor data, such as accuracy measures over 2000 and taxi trips lasting over 3000 seconds, or 50 minutes. Plotting simple graphs between the variable(s) help easily visualize the irregularities and anomalies that lie within our data. In alignment with the ETL process, this data will be cleaned up and imputed in the Data Processing phase.

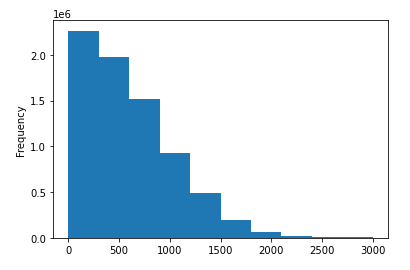
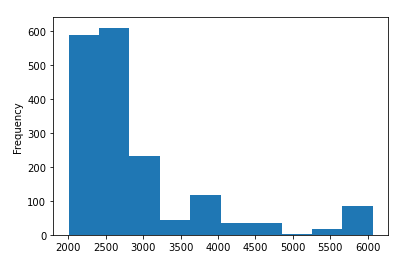


Fig 4: Accuracy Histogram > 2000 Fig 5: Second Histogram > 3000

In addition, there contains data that does not make logical sense, for example, negative speed. If it were velocity, this would make sense. However, since the metric here is speed, which does not have any direction, it is not possible for speed to be negative.

Other plots were also generated in order to help gain more insights into the data.

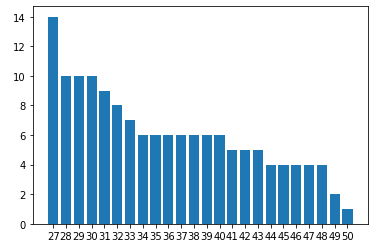
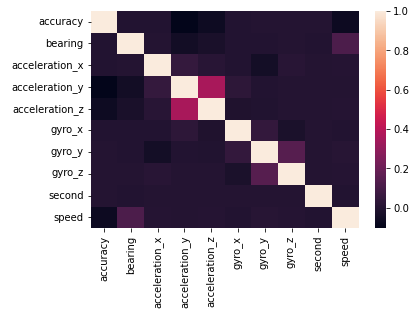


Fig 6: Correlation Heatmap Of Numerical Variables Fig 7: Counts Of Age Of Drivers

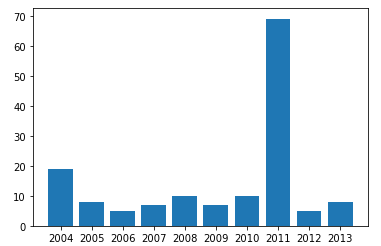
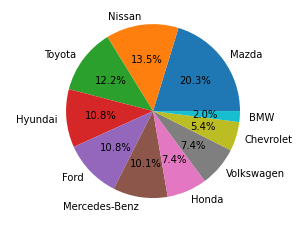


Fig 8: Pie Chart of Percentage Of Car Brands Fig 9: Number Of Cars by Make Year

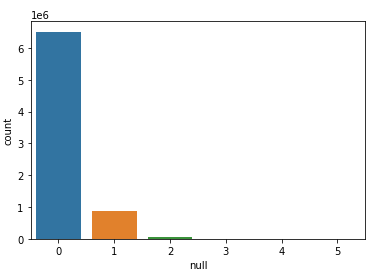
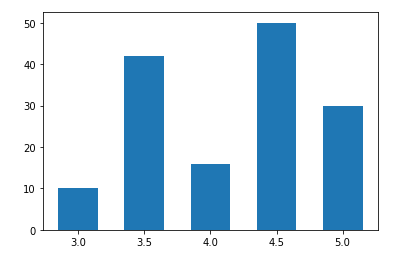


Fig 10: Bar Chart Of Driver Ratings Fig 11: Bar Chart Of Number Of NULL Columns Per Row

**Data Processing & Wrangling**

In the previous section, we have identified irregular and nonsensical behaviors in our data through Exploratory Data Analysis. In this section, we will explore how to process and wrangle this data to present it in an easy to use and functional state.

First, we drop the rows with accuracy values greater than 2000. This approach was adopted as there is an incredibly small percentage of values that lie within this range, in comparison to the full amount of data that we have. Moreover, it does not make much sense to impute these values as they are already nonsensical and make no sense, making the rest of the values in the row less authentic and trustworthy.

Secondly, we drop the rows with negative speed values. Once again, similar as before, this option was adopted as it is nonsensical data. Although speed cannot be negative, there are quite a number of rows containing speeds less than 0. The existence of negative speed values shows that the error lies within faulty speed sensors. This makes it not hard to believe in the chance that other sensors could be possibly malfunctioning as well. As a result, this makes the data in those records highly untrustworthy and unauthentic. Therefore, since it does not offer much value in imputing such values, they will be dropped.

Thirdly, we impute the values of rows consisting of anywhere between one and three NULL values. This method was chosen as the number of NULL values existing in the data is nowhere near as small previously stated. To impute these values, the data is first grouped by bookingID. The NULL values of each group are then imputed using the median of the data per bookingID group. This allows us to have a reasonable level of authenticity of imputed data, at the same time retaining the amount of data we have.

Lastly, we deal with the irregular values in the ‘second’ column. To do this, values over 3000 seconds, or 50 minutes are simply dropped from the dataframe. This is because the number of rows where the accuracy value is greater than 3000 is miniscule when compared against the remainder of the data that still exists. Imputing this column does not particularly make a great deal of sense as well and hence is dropped.

After the imputation of these values, we change the datatypes of columns that exist across multiple tables to be uniform. This is so that tables can be joined to form complex queries in SQL Server later on.

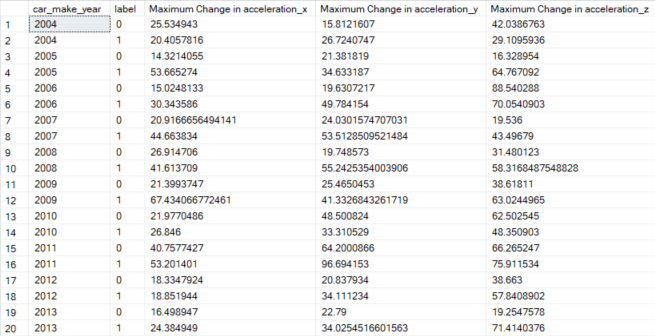
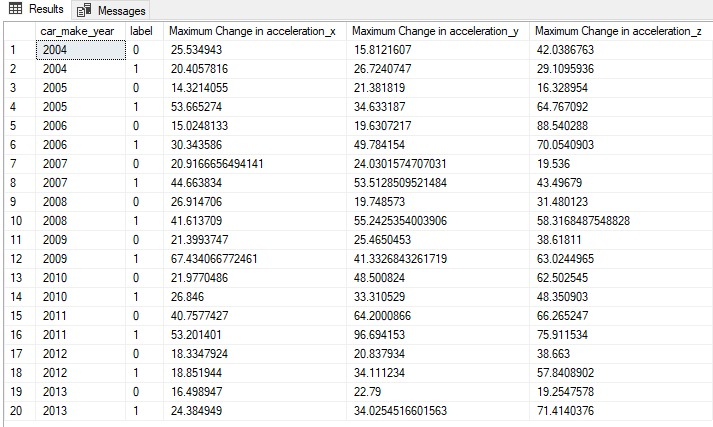
**ETL Pipeline For Storage Of New Data**

Following the ETL process, after the data has been cleansed and imputed to a functional and useful state, we can load the data back into a new SQL Server database.

Once again, this is achieved through the use of SQLAlchemy. Once the new data has been housed in the database, we can perform complex SQL queries to gain an even greater understanding of the data. The data can also be exported to data visualization programs such as Power BI and Tableau, in order to form high-level visualization to provide deep and meaningful insights on the data.

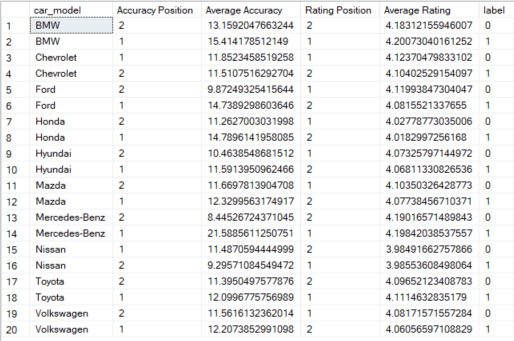
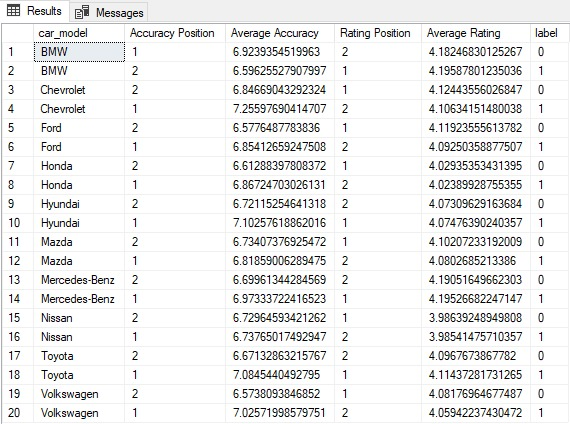
**Complex SQL Queries & Data Understanding: Cleansed Data**

Let us compare of SQL Query results before and after data cleansing

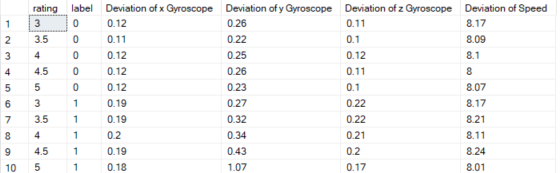
Query 1: Before Query 1: After

There seems to be no difference in output before and after data cleansing.

Query 2: Before Query 2: After

After cleansing the data, the number of car models whose accuracy is higher for normal trips dropped from 2 to 1. The number of car models where normal trips received higher rating remained the same at 6.

Query 3: Before Query 3: After

There is no difference for query 3 before and after data cleansing.

**Advanced Data Visualisation & Insights**

To make interactive dashboards and visualizations that meet our users’ needs and specifications we have opted to use Tableau as our data visualization tool of choice.

Tableau is an excellent data visualization andbusiness intelligence tool used for reporting and analyzing vast volumes of data. It is an American company that started in 2003—in June 2019, Salesforce acquired Tableau. It helps users create different charts, graphs, maps, dashboards, and stories for visualizing and analyzing data, to help in making business decisions.Tableau has a lot of unique, exciting features that make it one of the most popular tools in business intelligence (BI)

According to our user stories, the main objective of our users is to determine the safety of the ride and some metrics which are presented in an interactive manner to drivers to allow them to improve their driving. To accomplish this first we need to determine what differentiates a safe trip from an unsafe one. We plotted a line graph plotting the average accelerations per second across all bookings versus each second of the trip which is differentiated by the safety label . This allows us to very closely examine the patterns of acceleration increases and decreases at a micro level to determine the differences between a safe trip and an unsafe one.

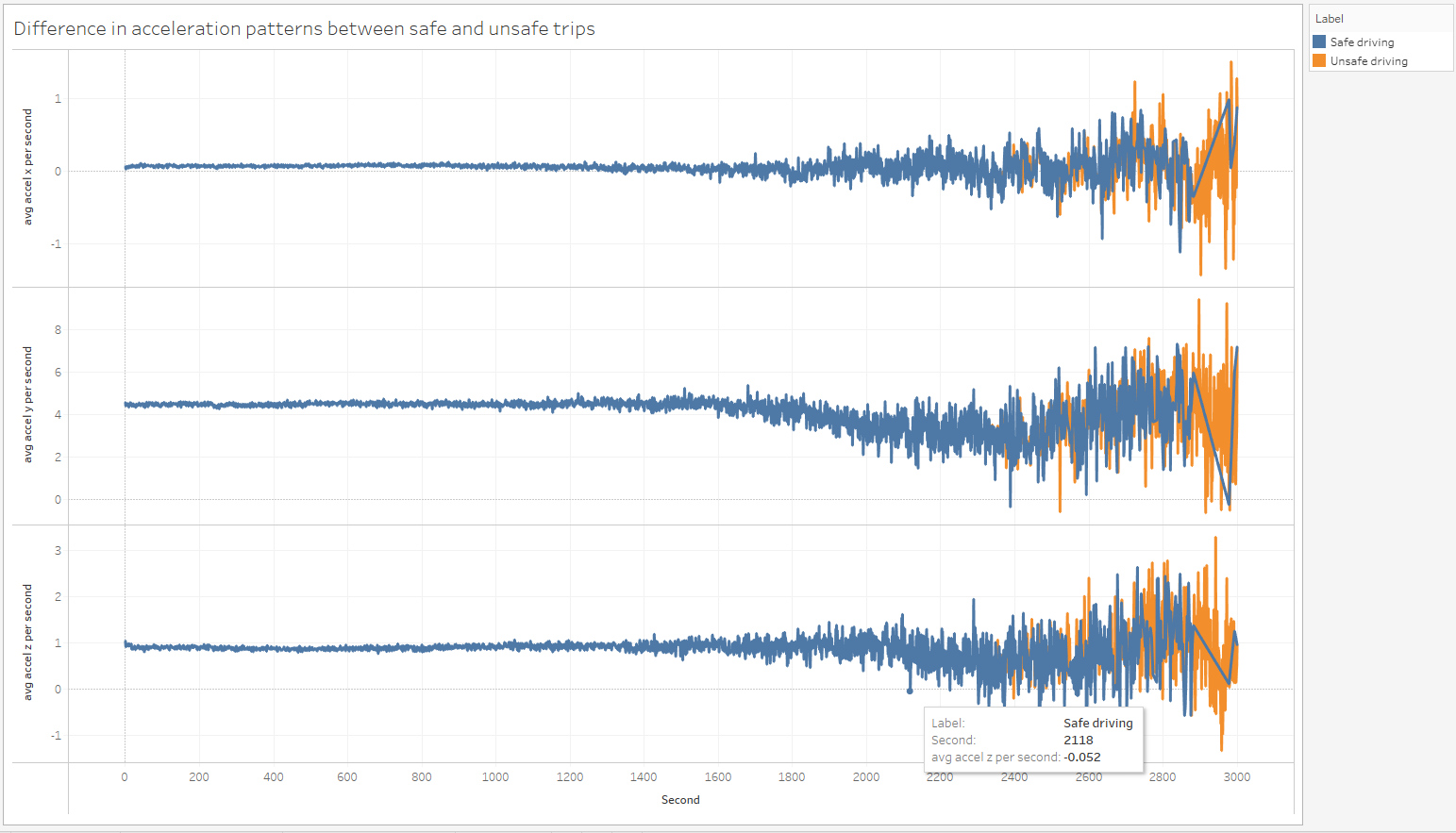


Fig 12: Difference in acceleration patterns between safe and unsafe trips

Looking at Fig 8, we see that generally there are not many differences between a safe and unsafe trip for the first 1800 seconds. However, beyond the 1800 second mark we see that the acceleration patterns are much more volatile. Furthermore, unsafe trips appear to have greater volatility in the acceleration patterns than safe trips. This implies that unsafe trips are largely characterized by the volatility of the acceleration patterns towards the end of the trip. Based on our background research to measure volatility in a given variable variance or standard deviation is used with standard deviation being used more widely (Hayes 2003).

Hence from our preliminary graph we see that the volatility i.e. standard deviation of acceleration from 1800 seconds is related to the safety of a trip. To more accurately measure this relationship we plotted a scatter plot between the standard deviation of speed. Standard deviation of speed is calculated rather than acceleration as speed represents the sum of the different accelerations. Hence, it can be considered a holistic representation of the three acceleration metrics.

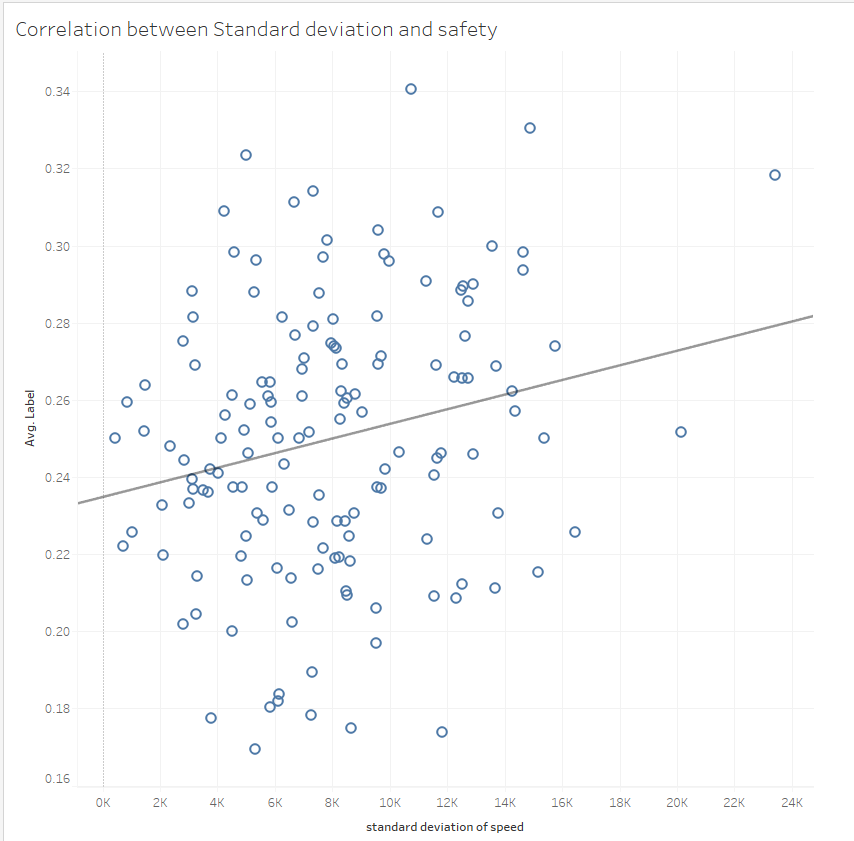


Fig 13 : Correlation between Standard deviation and safety

From Fig 9 we can see that as initially suspected standard deviation of speed is indeed correlated with the average safety label. The p-value that measures the significance of the correlation observed is 0.008 which is much lower than the threshold of 0.05 wherein a p-value that is lower than the threshold implies that the result observed is significant and not just due to chance. This shows that standard deviation of the speed after the first 1800 seconds of a trip can be measured to determine the safety of a trip.

Now that we have discovered a metric to determine safety by, it is time to provide the users an interactive dashboard to measure the safety of their driving by. To achieve this we shall use a packed bubble chart. A packed bubble chart uses the size of the bubbles to visualize the variable. This is more interactive and intuitive than a simple bar or line graph.

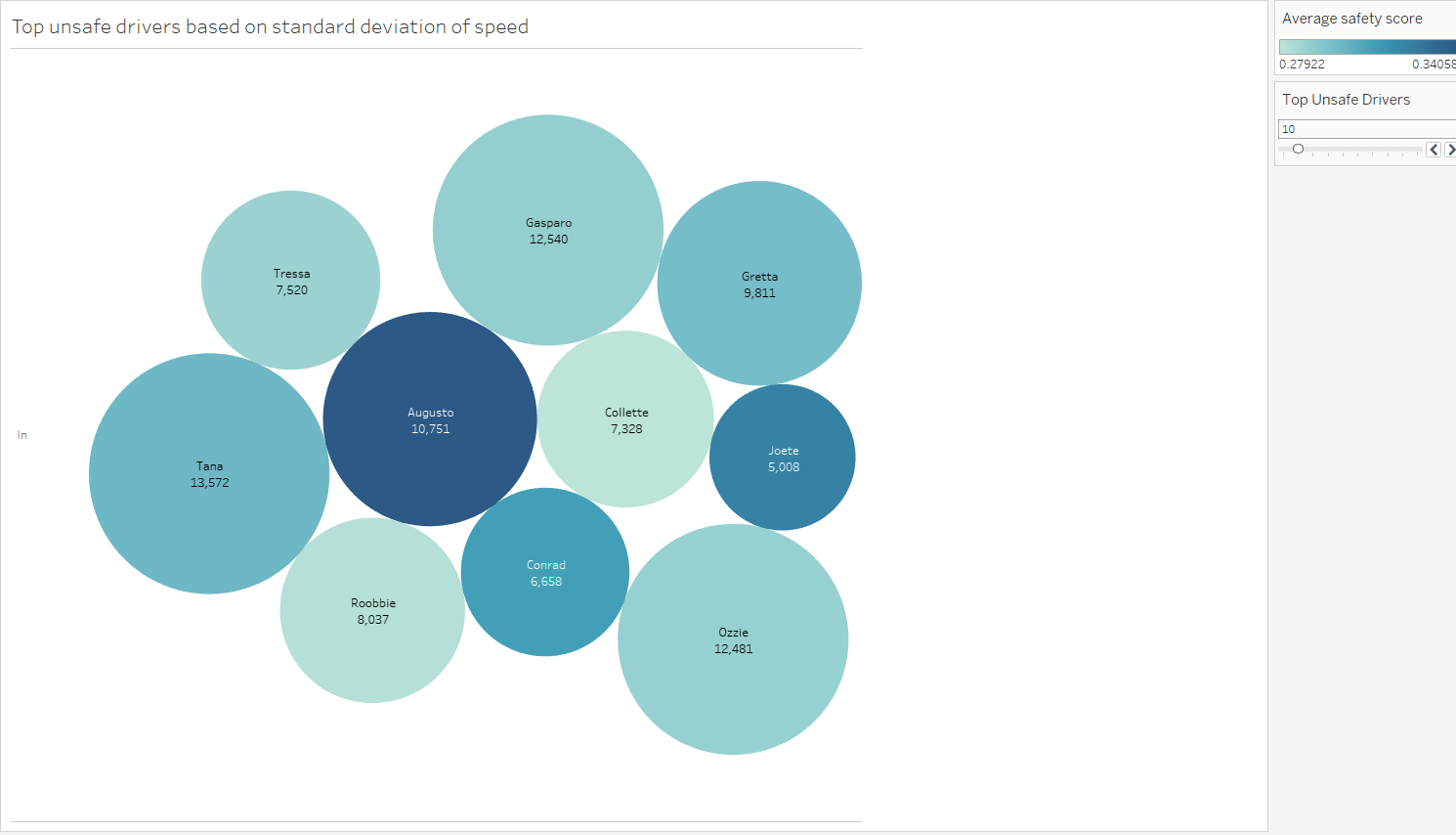


Fig 14: Top unsafe drivers based on standard deviation of speed

The colors of the bubble refer to the average safety label of the driver while the size of each bubble refers to the standard deviation of speed. As we can see most of the drivers with high standard deviation of speed also have high average safety label scores. Another requirement from the users is for the visualization to be interactive. To achieve this the parameter called Top Unsafe Drivers was used.. The user can enter the number of top unsafe drivers they wish to see and the graph would change based on user input.

We also have the details of each car such as car model and car make year provided in the dataset. Different car models have varying levels of safety features and driving features. Volvo for example was one of the first car manufacturers to pilot the three point seat belt which is now a common requirement. Therefore, it is worth visualizing the safety of each car by its model. For this , we used a bar graph in which a table calculation is applied to calculate the rank of each car model based on the average safety label. Different car models also have different target audiences. Some car models like Dodge which fall under the class of muscle cars cater almost exclusively to male audiences. This also means that the features of the car are made with their target audience in mind. Therefore, it is also worth exploring the differences between the genders.

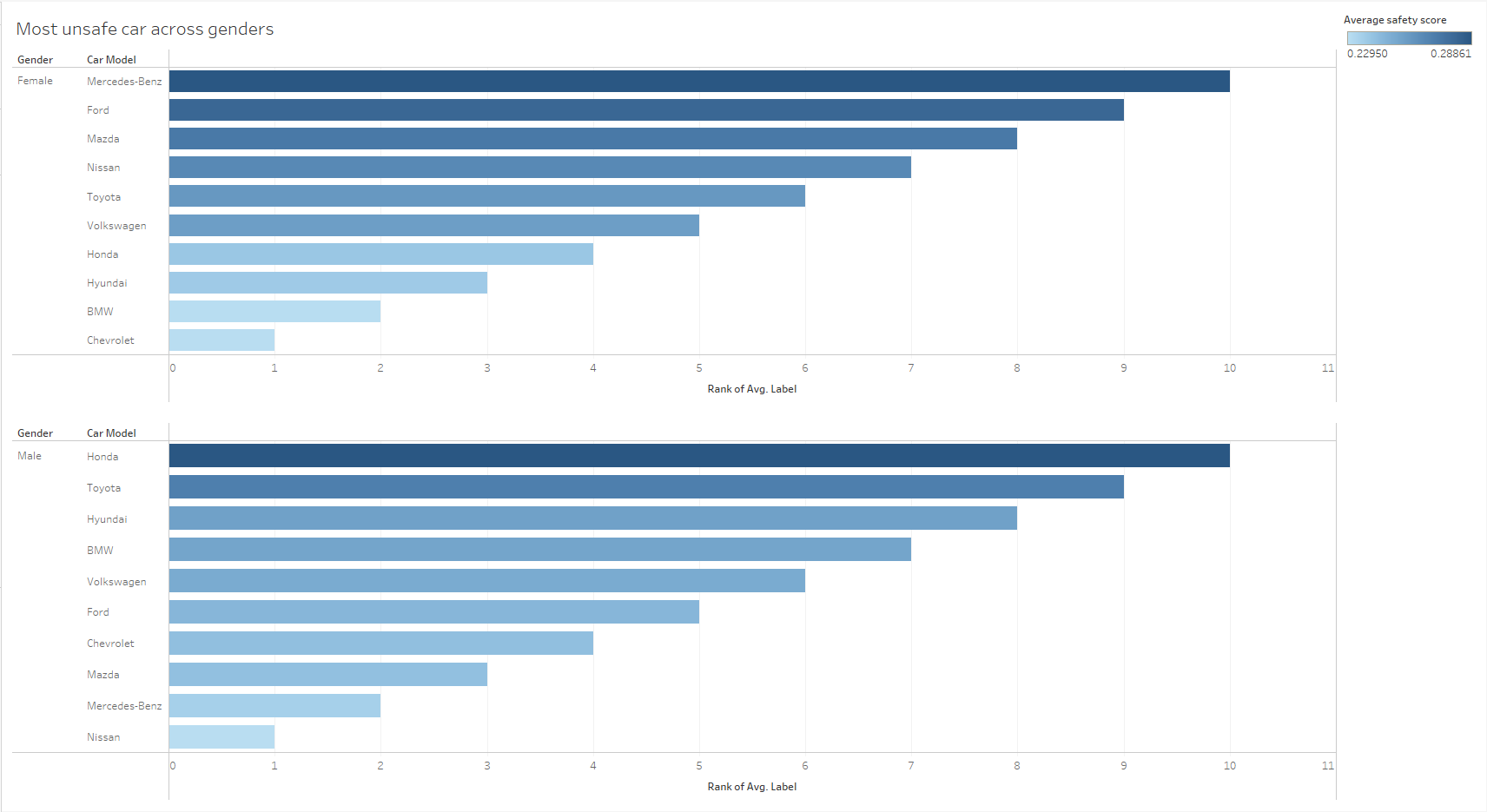


Fig 15: Most unsafe car model across genders

As suspected the rank of car brands vary across genders. The worst car in terms of safety is the Mercedes Benz model for women while the worst car in terms of safety for men is the Honda model. In general we see that the Chevrolet model is a safe car for both genders.

**Insights & Recommendations**

In conclusion we see that drivers with a higher standard deviation of speed generally drive less safely. This is probably because sudden changes in speed is dangerous and might not give enough time for other drivers to respond appropriately. Perhaps appropriate training can be provided to train drivers to avoid such sudden changes in speed thereby reducing their standard deviation and making them safer drivers.

Next appropriate cars can be assigned to people based on their genders. For example, since a Mercedes Benc is the most safe car for men and the least safe car for women, more men can be assigned to a Mercedes Benz instead of women. Generally the taxi agency can stock on more Chevrolet model cars as these seem to be appropriate for both genders.

**References**

Abdollazadeh, A. (2021) Crisp-DM and agile-scrum methodology for Data Science Project Delivery, LinkedIn. Available at: <https://www.linkedin.com/pulse/crisp-dm-agile-scrum-methodology-data-science-project-abdollazadeh> (Accessed: December 2, 2022).

Beginner's Guide to IMU (no date) Beginner's Guide to IMU | Robotics Club. Available at: <https://students.iitk.ac.in/roboclub/2017/12/21/Beginners-Guide-to-IMU.html>(Accessed: December 2, 2022).

Dalli, K. (2022) GPS directions may not always lead you to the safest routes, study finds, ConsumerAffairs. Available at: <https://www.consumeraffairs.com/news/gps-directions-may-not-always-lead-you-to-the-safest-routes-study-finds-030122.html> (Accessed: December 2, 2022).

Hayes, A., (2003). Volatility: Meaning In Finance and How it Works with Stocks [online]. *Investopedia*. [Viewed 2 December 2022]. Available from:<https://www.investopedia.com/terms/v/volatility.asp>

Nissan (no date) Why update your Nissan Navigation System Map? Available at: <https://nissan.navigation.com/cms/page.WhyUpdate/en_US/NissanNA/USD> (Accessed: December 2, 2022).

Thurber, M. (2020) A holistic framework for managing data analytics projects, Elder Research. Available at: <https://www.elderresearch.com/blog/a-holistic-framework-for-managing-data-analytics-projects/> (Accessed: December 2, 2022).

Alhyari, Salah. (2016). Re: What is the acceptable R-squared in the information system research? Can you provide some references?. Retrieved from: <https://www.researchgate.net/post/What-is-the-acceptable-R-squared-in-the-information-system-research-Can-you-provide-some-references/569cf1be6225ff80618b4578/citation/download.>