Simulation and Data Visualisation Assignment

Samson Magesh

K21172585

1 Part 1: Analytics

1.1 Exploratory Research Questions Proposed

Q1 – Analyse performances of Formula 1 teams. Are there any detectable trends? This question will use the dataset provided on Kaggle, the Formula 1 World Championship (1950 – 2023). Sub-questions for this section could be, which car constructors (team) are the best based on the points they have won over time. In particular for this question the constructor results, constructor standings, constructors and races data will be analysed.

Q2 – Analysing the impact of pit stop times on a drivers final position. Are there any trends? Formula 1 is a sport where miliseconds has an impact on a drivers' prospects of winning. In such sports everything has to be efficient. In formula 1 one key essential aspect that can have an impact is the pit stop. Every driver must use 1 pit stop in a race [6]. So, this question focuses on whether there is a trend in pit stop times and the drivers' final position. The datasets used would be lap times and pit stops.

Q3 – Analysing how the average speed of cars evolved over time in Formula 1. Are there any trends? It is interesting to see with modern technology advancements how much Formula 1 cars have improved in every aspect with the main one being speed. It will be interesting to see the difference in aspects such as lap times. The datasets used will be results, races and circuits.

1.2 Data Types and Datasets

1.2.1 Constructor Results Dataset

Research question 1 requires the constructors dataset to use for getting the name of the constructor associated to the constructor id given from other datasets such as constructor results. The dataset contains quantitative discrete data and categorical data. The quantitative discrete data are constructorResultsId, raceId, constructorId, and points. The categorical data is status which classifies why a driver did not finish a race, it does not provide numerical values. To answer the exploratory research question the most appropriate dataset is provided by the Kaggle Formula 1 World Championship (1950 – 2023) dataset.

1.2.2 Constructor Standings Dataset

Additionally for research question 1 it requires the constructor standings dataset to be used for analysing the performances of formula 1 teams. The dataset contains only quantitative discrete data. To answer the exploratory research question 1 the most appropriate dataset is provided by the Kaggle Formula 1 World Championship (1950 – 2023) dataset.

1.2.3 Constructors Dataset

Research question 1 requires the constructors dataset to use for getting the name of the constructor associated to the constructor id given from other datasets such as constructor results. The dataset contains categorical data and quantitative discrete data. The quantitative discrete data is constructorId. The categorical data is constructorRef, name and nationality. URL is a qualitative data. To answer the exploratory research question 1 the most appropriate dataset is provided by the Kaggle Formula 1 World Championship (1950 – 2023) dataset.

1.2.4 Races Dataset

Research questions 1 and 3 require the races dataset. Question 1 needs it for. Question 3 requires to be used for calculating the speed of the cars as it contains the year and circuit ID for each race. The dataset contains quantitative discrete data of raceId, year, and round. Contains also quantitative nominal data of circuitId. Contains also qualitative nominal data name. Contains qualitative ordinal data and time as well as qualitative nominal data URL.

1.2.5 Lap Times Dataset

Research question 2 requires the lap times dataset which contains quantitative discrete data on the raceId, the driverId, lap, position and milliseconds. In addition, quantative contionous data is required using the time it took to complete the lap. Breaking down the data provided in this dataset combined some data from driver standings and constructor standings dataset, best constructors can be identified by the driverId which can be used to find the constructor by the constructorId in constructor standings dataset. For this question the lap times dataset provides data from 1950 all the way to 2020, providing a vast range of data from a long period of time. This is plenty of data in itself to allow a good analysis. most datasets related don't provide as much information about the lap times with such a big range of data due to the timeline of data in the dataset. Another dataset found was very similar to the one provided except for containing data up to 2017. Therefore, out of all the datasets found for this section the provided F1 dataset was the best to answer the proposed exploratory question.

1.2.6 Pit Stops Dataset

Research question 2 requires the pit stops dataset to use for analysing the pit stop times for each constructor team. The dataset contains quantitative discrete data of raceId, driverId, stop and lap and quantitively continuous data of time, the time during the race the pit stop occurred, the duration of the pitstop i.e., how long it took for the pitstop and the pitstop duration in milliseconds. To be able to analyse the pitstop time with the drivers lap time this requires the datasets to be combined. To be able to see the trends of lap time against pitstop time. This is to analyse whether for example a longer pitstop duration increases the driver final position. To answer the exploratory research question the most appropriate dataset is provided by the Kaggle Formula 1 World Championship (1950 – 2023) dataset.

1.2.7 Circuits Dataset

Research question 3 requires the circuits dataset to use for calculating the speed of the cars as it contains information about circuit distance which can be used for the calculation. The dataset contains qualitative nominal data and quantitative continuous data. The qualitative nominal data are circuitId, name, location, country and alt. the qualitative continuous data are latitude and longitude. URL is qualitative nominal data. To answer the exploratory research question the most appropriate dataset is provided by the Kaggle Formula 1 World Championship (1950 – 2023) dataset.

1.2.8 Results Dataset

For the exploratory research question 3 the data required would be a mix of qualitative data such as the team name and race name to analyse whether certain team perform better on certain circuits. Quantitative data is the race year. Quantitative discrete data is the position which helps answer if a team is successful based on the finishing position of the driver for a particular team. The most appropriate dataset found was the dataset provided of Results from the Formula 1 World Championship (1950-2023) from Kaggle. The constructor of each driver is given, along with their grid position, final position, points given for that race based on their final position, the number of laps the driver was able to complete, the time to complete the race, along with fastest lap time and fastest lap. This dataset starts from 1950 and contains data all the way up to the year 2020, meaning the trends detected are valid as it's from lots of data ranging over a long period of time.

1.3 Correlation

Constructor results and constructor standings datasets are correlated where the points columns are proportional to the wins column in constructor standings, where the driver who scores the highest number of points in a race equates to a win. Also, the qualifying dataset and results data can be correlated where the higher the driver finishes in qualifying represented by the column position it can help them to finish higher in the main race as represented by the position column. Additionally, the correlation between constructor_results, constructor_standings, races, constructors, these datasets is primarily based on the constructorId and raceId fields, which are used as keys to connect the relevant information from each dataset. The constructor_standings dataset is enriched with the race date information from the races dataset using the raceId, and constructors' name from the constructors dataset using the constructorId.

2 Part 2: Design and Discussion

The visualisations were created with a focus on user experience. In particular these factors were taken into consideration the use of color, size, and shape to effectively convey the key insights of the visualisation.

- **Colour:** The use of color in the visualisations was thoughtfully considered to effectively convey the data and highlight key insights. A range of colors were used to make the visualisations sound and appropriate as possible. Overall, the goal was to make the visualisations both aesthetically pleasing and easy to read.

- **Size:** The size of the visual elements was carefully considered to ensure that the visualisation would be easily readable. The size of the elements was created to allow for the most important information to be easily distinguishable and readable.

- **Shape:** The selection of shapes was thoughtfully made to enhance the readability of data. For instance, scatterplots were designed using circular shapes to represent data points, as circles are widely recognized as points or dots. Moreover, the size of the shapes was also given importance to make sure data is readable.

2.1 Question 1

For the first proposed research question, the first concept was to create a line chart of the results over a timeline to see the trend of the constructors over time, to see if there is a trend in which constructor has performed the best over the years. Each data point can be viewed to show the constructor, the number of points accumulated up to that date as well as the date of how many points were accumulated up to that point. After a few other designs were considered, the final design chosen was the line chart where the points scored by each constructor can be visualized as shown in Figure 1 in the appendix. The visualisation shows the trend of constructors' performance over time in terms of points. It displays a line chart with date on the x-axis and points on the y-axis. The chart represents the top 10 constructors based on their accumulated points over time. User interaction would allow the user to find out more detail about each specific point. The user could hover their mouse over each data point where the coordinates of the data are displayed alongside the constructors' name. Each line in the chart corresponds to a constructor, and the legend allows users to click on a constructor's name to view its points trend over time. When a constructor is clicked, the chart will only display the trend for that particular constructor. When the user would like to view all the constructors results at once again, they can click on the constructor name again, which will show all the charts. The visualization will generally show the performance trends of the top 10 constructors over time. These trends could reveal periods of high or low performance, steady improvement or decline, or fluctuations in points earned by the constructors throughout the given time frame. By clicking on individual constructors in the legend, users can explore the trends of each constructor and compare their performance over time. This design was chosen to be created instead of the other ones to allow a clearer visualization of the trends to view more clearly the trend of how formula 1 teams have performed over time. This design shows the trends much clearer compared to other designs such the scatter plot. Line charts provide a clearer visualisation of trends within the data to make it more readable to the user. Hence this visualisation would provide a good insight into the possible trends between constructors' performance over time.

2.2 Question 2

For the second exploratory question the idea of using a heatmap was one of the ideas thought of visualizing the research question 2. The design would allow the user to be able to hover a specific data point which represents the correlation between the final position and pitstop time. When hovering over the specific tile the information shown would include what the final position was and the pitstop time along with the correlation between the two variables as well as the number of data points. Ranging from 0 to 1. The darker the tile the stronger the correlation between the final position and pitstop time, the lighter the tile the lower the correlation between the final position and pitstop time. This idea would clearly show the trend between the final position and pitstop time due to the appearance on screen it would be easier to see the trends. Another design idea was using a scatter plot to show the trend between final position and the pitstop time, through this the user interaction would be that they can hover over a data point, and it would display the constructor name, the final position and the pitstop time. Additionally, if the user wanted to know only about a specific constructor, they can click on the legend name and it will highlight the data points relevant to only that team. This would allow the user to be able to compare trends between specific teams. This design would allow the user to be able to see trends and can interpret the data easily.

2.3 Question 3

For the third exploratory question the designs include more creative user interaction-based ideas and generally more interesting visualisations which would be intriguing and fascinating. One of the designs encapsulates this idea through the use of the global map which would allow the user to click on a continent then see which countries within the continent hosted or currently host formula 1 races [5]. The user can then click on specific countries which would display the circuits within the country if applicable. When the user hovers over a circuit it will display the speeds at the circuit from the data available visualised through a line chart. This is to allow the user to visualise the trends of speed change over the years. Additionally, the top speed achieved on the circuit and the lowest speed achieved on the circuit will be displayed. This design takes into consideration many aspects such as colour, size and relevance. The colour scale used would be red, red is the choice here as red is associated to formula 1 such as the colour of the logo. Different hues of red would be used to represent each continent. The hues would vary based on how many circuits a continent has. The more circuits there are the darker the continent and the fewer circuits there are the lighter the colour of the continent. Relevance is considered using the global map, considering formula 1 is an event that takes part all over the world. Trends can be seen through the line graph. This design is both creative making it a fresh way for the user to see the data and enabling the data to be analysed easily. The other visual idea was using an area chart where the area of the chart would represent the overall average speed, each colour on the chart represents a circuit. User interaction here would be by allowing a user the option of scaling the graph to see up to the data they are interested in. for example if a user is interested in analysing the trends from 1950 to 2000 then they can drag the graph from the right side to 2000 where data beyond 2000 disappears and the new graph would be scaled to fit the data of the users' choice. This design allows the user to view the trends easily as it shows data over time.

3 Part 3: Implementation

The data was processed for creating the visualisation in the following steps:

1. Load the data from the relevant CSV files (constructor_results.csv, constructor_standings.csv, races.csv, constructors.csv).
2. Parse the data and filter out relevant columns.
3. Merge the data to get the constructor name for each constructor ID in the constructor_standings.csv file.
4. Group the data by constructor and year.
5. Calculate the total points earned by each constructor in each season.
6. Create scales and define the line generator.
7. Draw the x-axis and y-axis on the chart.
8. Draw a line for each constructor showing their total points earned in each season.
9. Add a legend to the chart to show the name of each constructor and its corresponding color.

To create the "Constructors' Performance over Time" visualization, the following columns from each dataset were used:
constructor_results.csv: constructorId, raceId, points
constructor_standings.csv: constructorId, raceId, points, position
races.csv: raceId, year
constructors.csv: constructorId, name

In constructor_results.csv, the constructorId column is used to link each result to the corresponding constructor in constructors.csv. The raceId and points columns are used to calculate the total points earned by each constructor in each race.
In constructor_standings.csv, the constructorId column is used to link each standing to the corresponding constructor in constructors.csv. The raceId, points, and position columns are used to calculate the total points earned and final position of each constructor in each season.
In races.csv, the raceId column is used to link each race to the corresponding season. The year column is used to group the data by year and to create the x-axis of the chart.
In constructors.csv, the constructorId column is used to link each constructor to the corresponding results and standings in the other datasets. The name column is used to display the name of each constructor in the legend.

References

[1] C. Viau, "The Big List of D3.js Examples," [Online]. Available: https://christopheviau.com/d3list/. [Accessed 2 April 2023].

[2] I. Martínez, "formula-one-viewer," 16 November 2019. [Online]. Available: https://github.com/imartinezl/formula-one-viewer.

[3] M. Bostock, "D3 Gallery," [Online]. Available: https://observablehq.com/@d3/gallery.

[4] world_mapped_, "Maps on the Web," 22 Sunday 2022. [Online]. Available: https://mapsontheweb.zoom-maps.com/post/684961764333469696/location-of-every-single-f1-circuit-ever-raced-on.

[5] S. Sealy, "F1 pit stops: rules of the pit lane and stop, crew roles - who had the longest pitstop in Formula 1 history?," 28 March 2023. [Online]. Available: https://www.nationalworld.com/sport/formula-1/f1-pit-stops-rules-pit-lane-stop-crew-roles-longest-pitstop-formula-1-history-4076807.

[6] M. Bostock, "Home," 4 July 2021. [Online]. Available: https://github.com/d3/d3/wiki. [Accessed 16 March 2023].

[7] "Line chart with cursor showing exact value," [Online]. Available: https://d3-graph-gallery.com/graph/line_cursor.html. [Accessed 18 March 2023].

Appendix

Link to dataset
https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020?select=races.csv

Designs for Part 2: Design and Discussion

Q1

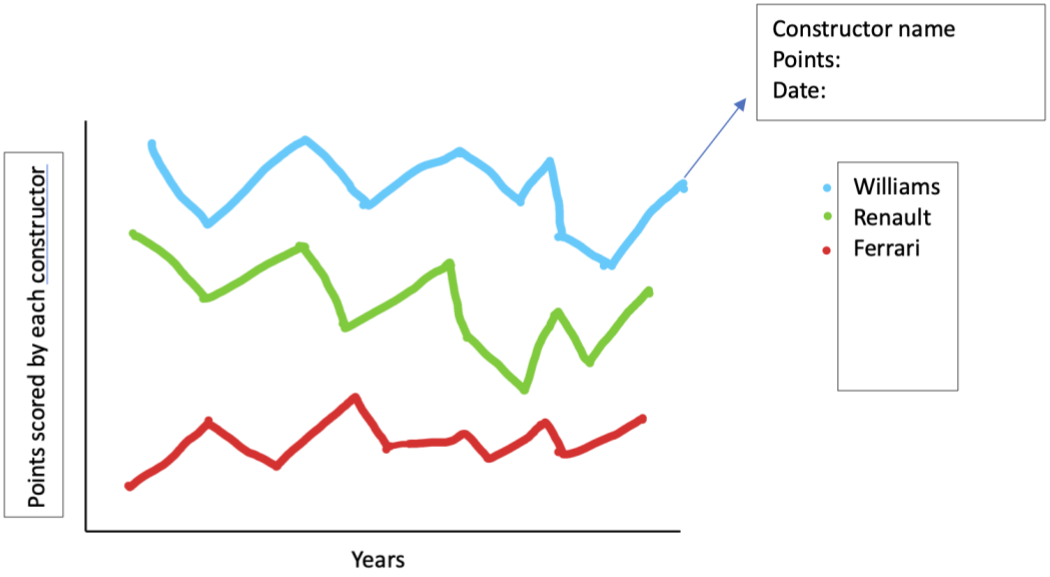


Figure 1: Line graph design showing trend between points scored by each constructor over the years for Q1

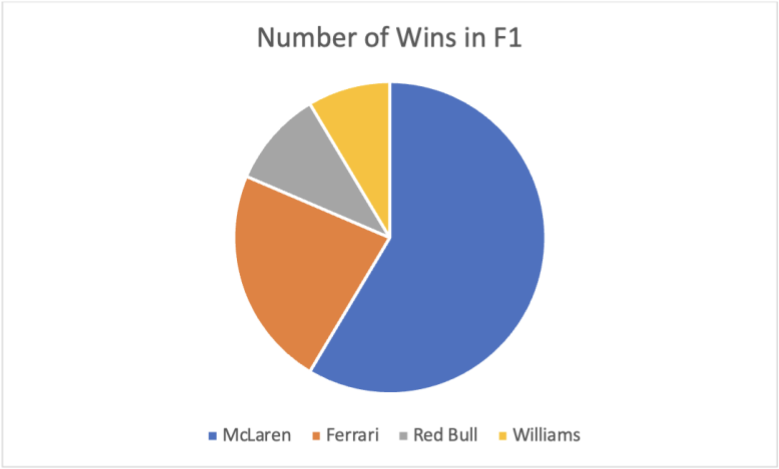


Figure 2: Pie chart design showing trend of most wins in formula of each team for Q1

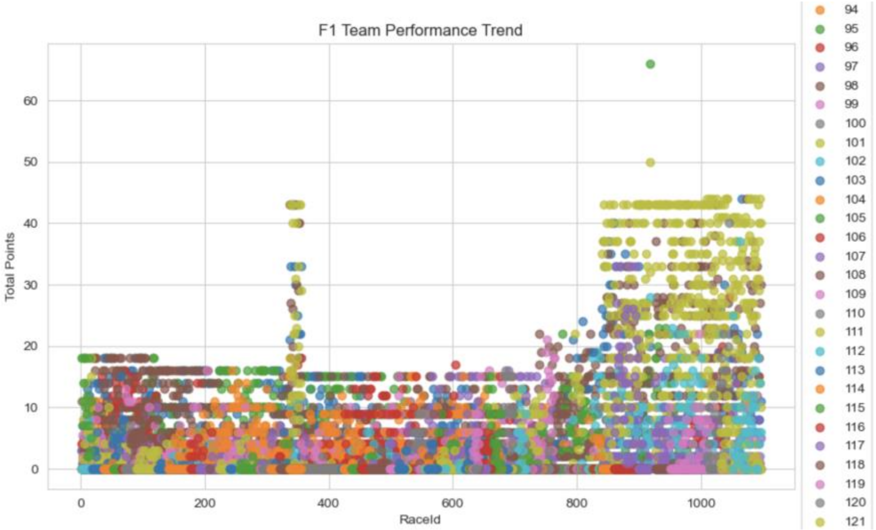


Figure 3: Scatter plot design showing trend between total points and raceId for each constructor for Q1

Q2

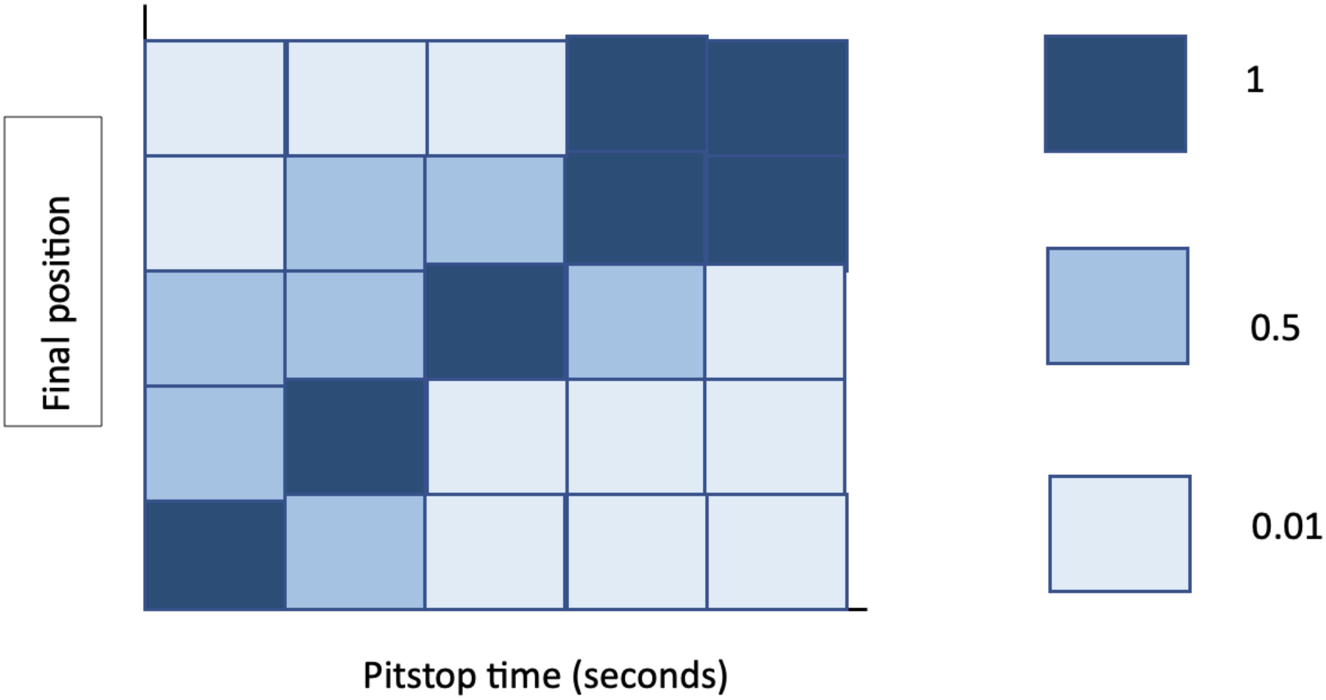


Figure 4: Heatmap design showing correlation between final position and pitstop time for Q2

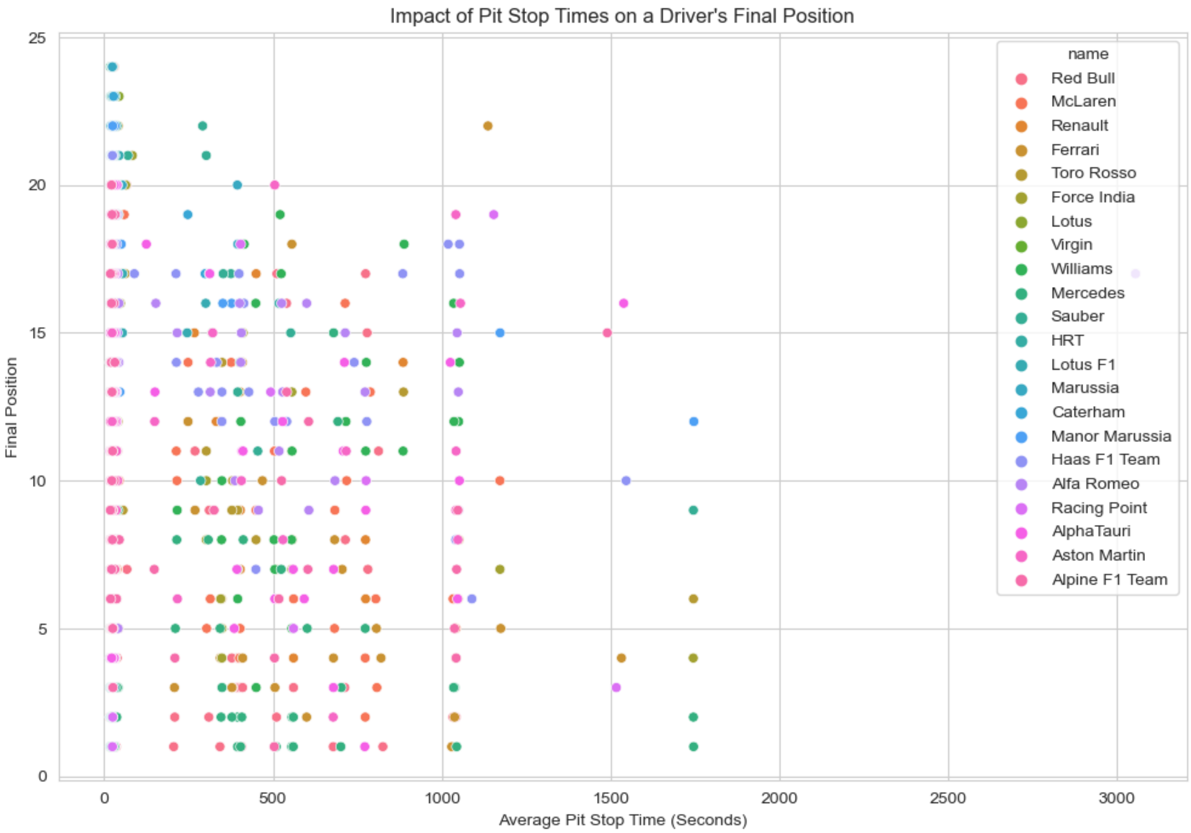


Figure 5: Scatter plot showing trend between final position and average pitstop time for Q2

Q3

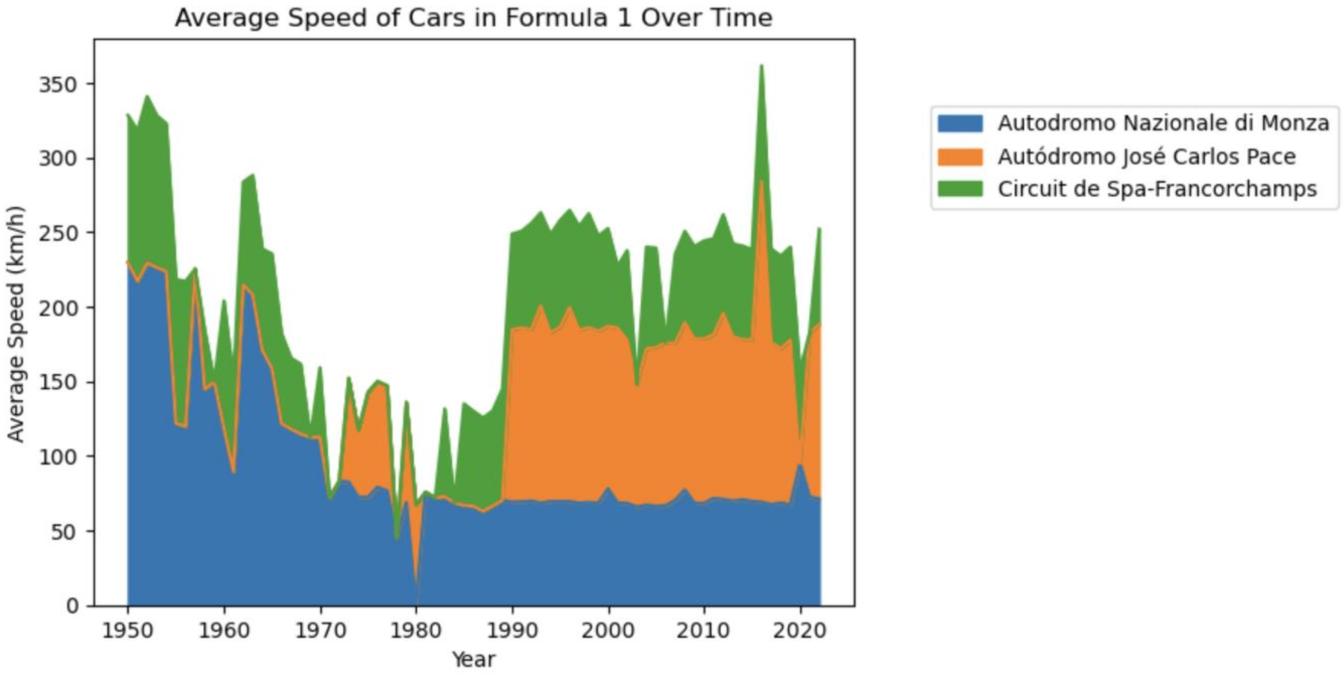
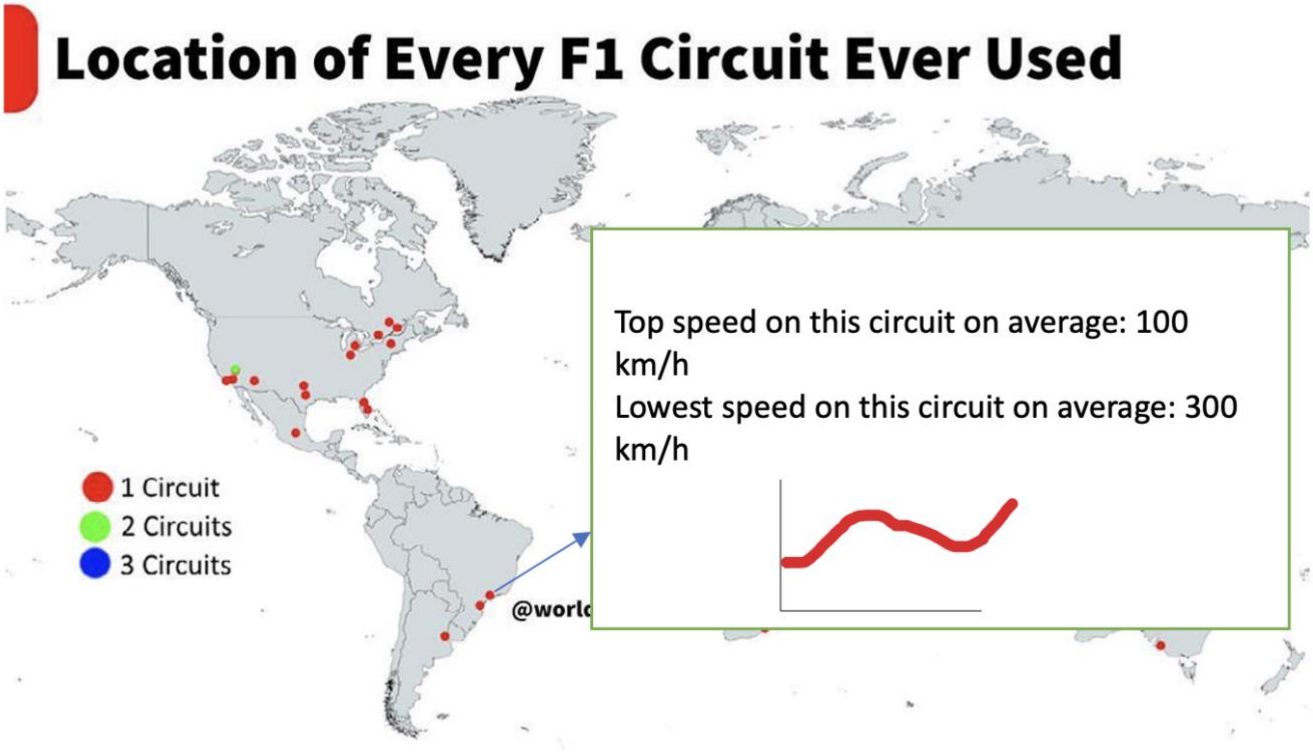


Figure 6: Area Chart showing trend of average speed at circuits over the years for Q3



Location of Every F1 Circuit Ever Used

Top speed on this circuit on average: 100 km/h

Lowest speed on this circuit on average: 300 km/h

Figure 7: Design idea for Q3[4]