**Simulation and Data Visualization Assignment**

Zheshan Cao

K1923587

King’s College London

**ACADEMIC HONESTY INTEGRITY**

Students at King’s are part of an academic community that values

trust, fairness, and respect and actively encourages students to act

with honesty and integrity. It is a College policy that students take

responsibility for their work and comply with the university’s standards and requirements. Online proctoring/invigilation will not be used for our online assessments. By submitting their answers

students will be confirming that the work submitted is completely

their own. Misconduct regulations remain in place during this period and students can familiarize themselves with the procedures on the College website. I agree to abide by the expectations as to my conduct, as described in the academic honesty and integrity statement.

**1 PART 1: ANALYTICS**

**1.1 Exploratory Research Questions Proposed**

**Q1- Analyze the performances of Formula 1 teams. Are there any detectable**

**trends?**

In this question, the focus should be the datasets that implicitly or explicitly show the performance of the Formula 1 teams such as the combination of result.csv and constructor\_standing.csv. And determining which attributes from the dataset are significant in affecting the performance of the Formula 1 teams.

**Q2- How does the age of Formula 1 drivers correlate with their win rate and overall performance throughout their career?**

Like other sports, age is an important attribute of the competition. It is worth analyzing the relationship between age and drivers’ performances. To achieves it, it should combine drivers.csv and races.csv to calculate the age of drivers when they were in the races. And compare the age with the results.csv.

**Q3- Do certain circuits have a higher frequency of accidents and collisions? Based on the average frequency of dangerous accidents.**

Life is always the most important. This question investigates the relationship between some characteristics of Formula 1 circuits and the frequency of accidents such as collisions. By analyzing the number of accidents and collisions at each circuit, we can identify any circuits that have a higher incidence of such events.

**1.2 Data Types and Datasets**

**Q1:**

Type of data necessary: continuous data, the points scored by the constructor as the performance. Integer data, the wins of the constructor.

Constructor\_results.csv: It provides information on constructors' points accumulation and race-by-race performance. This dataset includes raceId, constructorId, points, and status.

Appropriateness: This dataset tracks constructors' performance over time by points accumulation.

constructor\_standing.csv: it includes constructor standings by the end of each season, with details such as raceID, constructorId, points, position, and wins.

Appropriateness: This dataset helps to analyze the overall performance and wins of constructors across seasons, which is useful for detecting trends in championship rankings.

There are still many important datasets left in this question, it is too complex to analyze them one by one. It will be more comprehensive if we consider the question “Are there any detectable trends based on the performances?” to “What attribute/feature from the dataset affect the performance?”. And then use all datasets in PCA algorithm. Therefore, all datasets provided is necessary.

**Q2:**

Type of data necessary: Date type, the date of birth of driver, the date of the competition. Continuous type, the point ownd by driver as the performance.

drivers.csv: it has information on driverId, driverRef, forename, surname, dob, and nationality. Having the date of birth (dob) is crucial for calculating drivers' age during races.

Appropriateness: This dataset provides the necessary information to track drivers' age.

races.csv: it contains the time for each race, including year, round, circuitID, name, date, time.

Appropriateness: This dataset can be used to analyze drivers' age when he/she was taking part in the competition.

Results.csv: This dataset includes driverID, raceID and the performance of the driver in that race.

Appropriateness: the performance of driver in this race can be used to analyzed this question.

**Q3:**

Type of data necessary: continuous data and categorical data that describe the circuits, latitude and longitude of circuits, name and country of the circuit. Categorial data that represent different status.

Circuit.csv: it includes circuitId, name, location, country, latitude and longitude.

Appropriateness: This dataset provides information on each circuit's characteristics, which can be used to investigate factors contributing to accidents and collisions.

Races.csv and results.csv: these datasets contain data such as raceId, circuitId, and statusId.

Appropriateness: These datasets can be used to analyze the frequency of accidents and collisions at each circuit by the statusId.

**1.3 Correlation**

In question 1, Formula 1 team performances can be detected by correlating the datasets based on constructorId and raceId. It is able to correlate Constructor\_results.csv with Constructor\_standings.csv using constructorId, and both of these can be further correlated with Races.csv and Results.csv using raceId. In question 2, the Age of drivers and their performance can be analyzed by correlating the Drivers.csv dataset with the Results.csv dataset using the common field driverId. Additionally, linking the Races.csv dataset by raceId to gather information about each race, such as the date. There is one limitation, Drivers' age is not a direct field in the dataset, so it is necessary to calculate it using the date of birth (dob) field in Drivers.csv and the date field in Races.csv. In question 3, to analyze circuits with a higher frequency of accidents and collisions, correlating the Circuits.csv dataset with the Races.csv dataset using the circuitId field is important. Additionally, Results.csv dataset can be correlated with the Status.csv dataset using the statusId field to identify the string representation of statuses such as accidents and collisions.

**2 PART 2: DESIGN AND DISCUSSION**

Some concepts that all these visualizations design contain.

**Color:** Hues were specifically chosen for categorical filtering, as they are ideal for representing categories without inherent ordering. To ensure clarity and avoid confusion, no more than five different colors were used for categorical representation. Avoiding using high luminance colors to minimize eye strain, and colors chosen for differentiating between two categorical types had a relative difference in accordance with Weber's Law.

**Interaction:** Interaction: Interactive elements were incorporated to support the exploration of complex datasets, allowing users to better understand the data. Transitions were used when changing the shape or order of objects, helping users maintain focus on the relevant elements. Multiple views were also utilized, enabling users to explore the data from different perspectives and gain a comprehensive understanding of the information being presented.

**Aesthetics:** All designs should not exceed the size of the screen. The golden ratio (0.618:1) was employed to determine the size of individual visualizations and to divide the area between multiple views, ensuring a harmonious and visually appealing layout.

**Q1**

For question 1, the initial concept is to create a visualization that combines a bar chart and a line chart. The bar chart displays the significance of each feature based on the PCA results, while the line chart shows how the performance (points) changes with the selected feature's value as Figure 1 shows.

The bar chart displays the features ordered by their significance based on the PCA loadings. The y-axis represents the importance of each feature, and the x-axis represents the features themselves. Users can see the most significant features affecting the performance of Formula 1 teams.

Users can select a feature from the bar chart by clicking on the corresponding bar. This action triggers the display of a line chart that shows how the performance (points) changes with the selected feature's value. The line chart has the selected feature's values on the x-axis and the performance (points) on the y-axis. Users can observe the relationship between the selected feature and the performance over time. The line chart should be accompanied by a label indicating the selected feature.

This visualization will allow users to explore the most important features affecting Formula 1 teams' performance and understand how these features influence performance.

**Q2**

For the second research question, It is a 3D visualization that combines a cube plot and a line chart to analyze the correlation between the age of Formula 1 drivers, and overall performance throughout their careers as Figure 2 shows. Here's a description of the visualization:

The 3D cube has three dimensions: the x-axis represents the age of drivers, the y-axis represents the average performance at a fixed age for a specific driver (e.g., Nico Rosberg), and the z-axis represents different drivers, the z-axis is ordered by their average performance throughout their careers. Each layer in the z-axis corresponds to one driver, resulting in multiple layers since there are 858 drivers in the dataset.

For example, Calculate the age of each driver based on their date of birth and the date of the race for the x-axis. Compute the average performance at fixed ages for each driver, such as 27 years old for the y-axis.

For more interaction, Users can zoom the cube and select a specific driver by choosing a layer from the z-axis. This action triggers the display of a line chart that shows the relationship between the driver's age and performance.

The line chart has the driver's age on the x-axis and their performance on the y-axis. Users can observe the performance trend for the selected driver as they age. The line chart should include a label indicating the selected driver's name.

Why not just have a two-axis chart with age and overall performance for all drivers at that age? it is easy to give different drivers the same weight and add their points.

Because Formula 1 drivers have varying levels of performance and distinct career timelines. So the age that has more sample data from the dataset is more precise than other ages, which causes the result of this visualization having bias.

This visualization allows users to explore the correlation between the age of Formula 1 drivers and their performance throughout their careers while considering individual differences among drivers.

**Q3**

For question 3, the visualization effectively displays the frequency of accidents and collisions at different circuits as Figure 3 shows. The visualization will consist of bubble charts and pie charts to provide insights into the data.

Before showing the chart, it is necessary to manually define which status belongs to dangerous accidents because there are 140 different statuses. In default, the chart chooses 12 cases as dangerous status. They are {3Accident, 4Collision, 20Spun off, 56Engine fire, 66Fire, 73Injured, 82Injury, 100Driver unwell, 107Eye injury, 137Damage, 138Debris, 139Illness} where the number is the id of status.

The main visualization will be a bubble chart where each bubble represents a different circuit. The size of the bubbles will indicate the average number of accidents per race at each circuit. By default, the bubbles will be randomly ordered on the screen, influenced by a simulated force.

Users can choose to order the bubbles in two ways: by average accidents per race or by total accidents count. Users can also filter bubbles, bubbles can be filtered by country, with circuits from a specific country displayed in red.

When a user clicks on a bubble, additional information about the circuit and a pie chart will be displayed. The pie chart will show the distribution of different dangerous situations at the selected circuit. This will help users understand the types of accidents and collisions that are most common at each circuit.