Proposal

Angela Lin, Haochen Yu, Jicong Zhang, Lanlan Qing

2025-11-05

Group member

Angela Lin(al4925), Haochen Yu(hy2951), Jicong Zhang(jz3899), Lanlan Qing(lq2250)

Tentative Project Title

Statistical Analysis of Accident-Prone Factors in Formula 1 World Championship - Insights from Formula 1 Race Data

Motivation

Formula 1 is one of the most technologically advanced and data-driven sports in the world, yet it remains inherently dangerous due to the extreme speeds and physical limits involved. Understanding the factors that contribute to car accidents in Formula 1 is essential not only for improving driver safety but also for enhancing race strategy, track design, and vehicle engineering. In this project, we would like to know how variables such as driver age, city and manufacturers are related to the occurrence of Formula 1 accidents. Understanding these relationships can help identify patterns—for example, whether younger or older drivers are more prone to accidents, whether certain tracks or cities have higher crash rates, or whether specific cars tend to be involved in more incidents. Our goal is to better understand the interplay between human, environmental, and technical factors in Formula 1 safety, contributing to both improved race conditions and more informed data analysis practices.

Intended final products

1. Exploratory Data Analysis
2. Data visualizations

Anticipated data source

1. Formula 1 Race Safety data (complements) from Kaggle
2. Formula 1 Race Events data from Kaggle

Planned Analyses/ Visualizations

1. EDA: intervention by season/era, circuit, street vs. permanent; histograms of SC/VSC counts, red flag frequency over time. Distribution of car accidents by driver age, city and manufacturers.
2. Bivariate analysis: Correlation analysis between car accidents and these variables.
3. Visualization analysis: Accident heatmap by city. Histogram of accident per car manufacturer.

Coding challenges

One challenge in our project is that the dataset spans many years, which results in inconsistent records and missing values. In addition, the data sources are broken into multiple detailed tables (e.g., race location, driver profiles, car team information, etc.), so we need to carefully identify, filter, and merge only the datasets that are relevant to our analysis. To tackle these challenges, we'll need to focus on writing efficient code and keeping our documentation clear, so that our work is easy to reproduce and follow.

Planned Timeline

November 1-8: Data collection, cleaning, and preliminary exploration.

November 10-14: Start analysis of the correlations between each factor (driver age, race location, cars, etc) and incident rates/race results.

November 15-25: Conduct additional outcomes and/or interactive visualization.

November 26-30: Finalize all analyses and begin drafting the report.

December 1-7: Complete the draft report with visualizations and interpretation.

December 8-12: Review and finalize the project for submission.