CSE 564 Visualization - Project Proposal

Driven to Break: Visualizing Failures and Reliability in Formula 1

Team 20

Harsh Vivek Londhekar ID: #116647641
Aitik Dandapat ID: #116626453

📊 Dataset Description

Source: https://github.com/f1db/f1db

The dataset chosen for this project is the **F1DB (Formula 1 Database)**, an **open-source** and **community-maintained** resource that compiles historical data from Formula 1 racing. It spans several decades, capturing detailed records from the early days of Formula 1 up to the most recent seasons. The dataset is structured in a **relational format** and is designed to reflect the multifaceted nature of the sport, including information about races, drivers, teams, and various race-related events.

This dataset is particularly valuable because of the depth and breadth of data it offers, allowing for both high-level trend analysis and granular, race-by-race investigation. It enables users to explore patterns in performance, reliability, safety, and strategy over time. The dataset is frequently used in motorsport analytics and academic research, making it a reliable foundation for building data visualizations and deriving insights.

This project will leverage the dataset to examine the reliability and safety aspects of Formula 1, drawing connections between mechanical failures, driver performance, and track characteristics. Its well-maintained and organized structure makes it ideal for integrating with visualization tools and developing interactive dashboards that highlight key trends in the sport's history.

⚙ Attributes:

1. **reasonRetired**: Indicates the reason a driver did not finish a race, helping categorize failures like mechanical issues, crashes, or engine problems.

2. **laps**: Represents the number of laps completed by a driver, allowing calculation of the lap on which failure occurred or the race completion percentage.

3. **raceName:** Race name, useful for linking data across tables and performing race-specific analysis.

4. **Year**: Denotes the season year of the race, enabling temporal trend analysis and year-over-year comparisons.

5. **round:** Indicates the sequence of the race in a season, which helps in analyzing reliability or performance changes across the season.

6. **driverName:** Driver name, allowing driver-wise aggregation and comparison of performance and reliability.

7. **constructorName:** Represents the team or constructor, enabling analysis of team-based trends in failures or successes.

8. **gridPositionNumber:** Shows the starting position of a driver on the grid, useful for assessing the relationship between start position and race outcome or failure.

9. **positionsGained:** Measures how many positions a driver gained during the race, helping explore aggressive driving and its potential link to failures.

10. **polePosition:** Indicates whether the driver started from pole position, often used to relate initial advantage with race completion or failure.

11. **fastestLap:** Flags whether the driver achieved the fastest lap, offering a performance metric that can be analyzed against race reliability.

12. **driverOfTheDay:** Recognizes the most outstanding driver in a race as voted by fans, useful for correlating fan-favorite performances with technical outcomes.

13. **pitStops:** Records pit stop details (count and timing), valuable for studying race strategies and their relation to failures or performance.

14. **tyreManufacturerId:** Identifies the tyre manufacturer used by a driver/team, enabling analysis of tyre brand impact on race performance and failure rates.

15. **sharedCar:** Indicates if a car was shared between drivers in a race, which may impact or correlate with reliability and race outcome.

16. **engineManufacturerId:** Represents the engine supplier, allowing comparison of engine reliability and performance across different teams.

17. **circuitName:** The name of the racing circuit where the event is held; useful for identifying and labeling circuits in visualizations or geographic plots.

18. **circuitType***:* Describes the nature of the circuit—whether it is a permanent race track (road) or a temporary street circuit—helpful in comparing failure or incident rates across

track types.

19. **Longitude:** The longitudinal coordinate of the circuit's location, used for plotting races on a global map or analyzing regional patterns.

20. **Latitude:** The latitudinal coordinate of the circuit's location, complements longitude for accurate geographic visualizations and spatial analyses.

🤔Problem Statement:

In the high-stakes world of **Formula 1 racing**, **reliability** and **performance** are critical to a team's success. Despite cutting-edge engineering and strategy, races are often influenced or lost due to **mechanical failures**, **strategic missteps**, or **unpredictable factors**. Understanding the patterns behind these **failures**, **retirements**, and **performance deviations** can provide valuable insights into what affects a race outcome. However, due to the complex nature of the sport and the multitude of influencing variables, these insights are often buried in large, historical datasets.

This project aims to perform a **visual analytical study** of **Formula 1 race data** to explore **failure trends**, **driver** and **constructor reliability**, and **performance-impacting factors** over time. By leveraging various attributes such as reasons for retirement, grid positions, tyre and engine manufacturers, pit stops, and geographical circuit data, the project seeks to uncover meaningful patterns and correlations. These insights can help fans, analysts, and even teams better understand the underlying causes of race outcomes and reliability issues across different seasons and circuit types.

📈Data Preparation and Feature Engineering:

1. **Data Merging**
   Relevant tables were merged using keys like raceId, driverId, and statusId. The final master table combines race metadata, driver and constructor details, finishing status, and failure causes. This unified view enables comprehensive cross-sectional analysis of incidents and DNFs by race, team, and driver.
2. **Data Cleaning**
   Cleaning involved removing duplicates, handling missing values, and normalizing categorical fields. Inconsistent labels for retirement reasons were grouped under standardized failure types. Only complete race entries with valid results were retained for accurate analysis.
3. **Data Sampling Strategy**
   Balanced sampling was used to ensure fair representation of DNF vs. completed races. Stratified sampling by decade and team helped highlight temporal and constructor-level trends. This ensures diverse and unbiased coverage across the dataset**.**

4. **Feature Selection**
   Key features selected include year, circuitId, driverId, constructorId, gridPosition, finishPosition, laps, and reasonRetired. Derived features like DNF_flag and FailureType are central to our analysis of safety patterns and reliability indicators.

📉Types of Visualizations:

1. **DNFs Over Time (Line Plot)**
   This plot will show the number of DNFs (Did Not Finish) per year, giving us a temporal trend of failures in F1 history. By plotting the count of non-finished races using the reasonRetired attribute over the years, we can understand if F1 races have become more or less reliable over time.
   Interaction: In the DNFs Over Time (Line Plot), users can hover over data points to view the total DNFs and top failure type for each year. Filters will allow selection by team, driver, or failure category. Zoom and pan features help focus on specific time periods.

2. **Failure Cause Breakdown (Treemap)**
   We will visualize the distribution of different failure causes using a treemap. This will highlight the most frequent reasons for DNFs and help categorize them (e.g., mechanical issues, crashes, technical faults). It provides insight into which factors dominate reliability issues in races.
   Interaction: The Failure Cause Breakdown (Treemap) will show proportions of different failure reasons with hover tooltips displaying counts and percentages. A season dropdown will update the view accordingly. Clicking a cause will filter the entire dashboard by that failure type.

3. **Team Reliability Ranking (Stacked Bar)**
   Using constructorName, we will compare the number of DNFs for each team per season. This will help assess which teams are most reliable and which tend to suffer the most retirements. A stacked bar will allow seasonal comparison as well as overall performance
   Interaction: In the Team Reliability Ranking (Stacked Bar Chart), users can compare DNF counts by team across seasons. Hovering reveals team-wise stats, and filters allow team of season selection. This helps identify reliable vs. failure-prone constructors.

4. **Tyre vs Engine Failure (Grouped Bar)**
   This visualization will compare failures attributed to tyre and engine issues using tyreManufacturer and engineManufacturer. It helps in identifying which manufacturers had more reliability issues over time.
   Interaction: The Tyre vs Engine Failure (Grouped Bar Chart) compares failure counts by tyre and engine manufacturers. Users can toggle between the two and filter by season. Hover tooltips show detailed failure data for each manufacturer.

5. **Driver Experience vs DNFs (Scatter Plot)**
   This plot will compare a driver’s experience (measured by total number of races) to their DNF ratio. It can help assess if experience plays a role in avoiding retirements or if failures are mostly team or machine-dependent.
   Interaction: For the Driver Experience vs DNFs (Scatter Plot), users can see how DNF ratios vary with race experience. Hovering shows driver stats, and a brush tool helps highlight clusters. A team filter enables intra-team comparisons.

6. **Circuit Risk Index (GeoMap area chart)**
   The **Geo Map Area Chart** will visualize each race circuit on a world map using its GPS coordinates. The size or color of each point will indicate DNF frequency, helping us spot high-risk circuits geographically. We’ll also use color to differentiate between circuit types (road vs. street) and add tooltips showing circuit name, country, and top failure reasons.
   Interaction: The Circuit Risk Index (Geo Map Area Chart) maps each circuit using its location, with size/color indicating DNF frequency. Users can toggle by circuit type and click to zoom into high-risk tracks. Tooltips show circuit name, country, and common failures.

7. **Factors Influencing DNFs: Multivariable Analysis (Parallel Coordinate Plot)**
   Display attributes like grid position, laps, pit stops, fastest lap speed, and a DNF flag to explore multi-variable patterns. It helps us see how combinations of factors influence DNFs. We can highlight or filter by DNFs to find trends, clusters, or outliers in driver or team performance.
   Interaction: The Parallel Coordinate Plot shows patterns across variables like laps, pit stops, and grid position. Users can hover over lines to explore race details and use sliders to filter ranges. A toggle highlights DNF-related patterns for deeper analysis.