Analyzing Formula 1 World Championship Data Using SQL

Dataset – Kaggle Formula 1 World Championship Dataset

Tools – MySQL Workbench, DrawSQL.app

Author – Isuri Perera Balasooriya

Date – 23rd Feb 2025

City – Colombo, Sri Lanka

Contents

1 Introduction

This project explores how SQL can be applied to extract insights from a real-world sports dataset specifically, Formula 1 race data. By asking a set of well-defined questions and using SQL queries to answer them, the goal was to sharpen technical skills while exploring patterns in driver performance, qualifying stats, and race outcomes.

Before importing the data into MySQL the following basic data cleaning steps were taken,

- Removed the tables not required for the project
- Removed nulls from key columns like primary keys
- Ensured data types matched expected values (no text in numeric columns)

1.1 Project Goal

Carry our advanced SQL analytics on Formula 1 racing data

1.2 Project Objectives

- Design and implement a normalized relational schema for F1 data.
- Write complex SQL queries to answer strategic and performance-based questions.
- Produce visual-ready data outputs that can be extended into dashboards.

To guide the analysis, 10 questions were prepared in advance. This not only gave direction to my querying but also made the analysis more purposeful.

1. Who are the most consistent drivers?
2. Which drivers convert entries into points the most efficiently?
3. Which drivers converts pole position to a win on race day?
4. Who are the fastest one-lap drivers in Q3?
5. Who overtakes the most positions on average?
6. Which constructors have the highest pole to podium ratio?

7. Which constructors have dominance at specific circuits?
8. Which teams have the fastest pit crew on average?
9. What are the head-to-head race stats between Lewis Hamilton and Max Verstappen
10. Who holds the most fastest lap record at each circuit?

2 Methodology

- **Database System Used:** MySQL
- **Skills Applied:**
  - Inner and conditional joins
  - Aggregation (SUM, COUNT, AVG)
  - Filtering with WHERE, HAVING, and REGEXP
  - Data transformation using CAST, SUBSTRING_INDEX, and time conversion logic

3 Analysis

The full analysis including relational schema, SQL queries and results are included in the following GitHub folder.

GitHub Folder Location

The key insights discovered from this analysis,

- Drivers with the lowest average Q3 times (with more than 50 appearances) were identified, revealing consistent performance across races.
- Some drivers showed high win counts despite not starting at the front highlighting racecraft over qualifying speed.

- Pole-to-win conversion rates helped evaluate how well teams capitalize on strong qualifying performance.

4 Key Learnings

- Query design is just as important as writing syntax. Framing the right questions matters.
- Data cleaning is not a side task — it directly impacts accuracy.
- Real-world datasets are messy, and transforming them often requires creative use of SQL functions.
- Even a domain like Formula 1 can be used to demonstrate fundamental data skills in a compelling way.

5 Future Work

- Include **seasonal trends** or **year-on-year** comparisons
- Merge with **telemetry or weather** data to analyze conditions
- Build **visual dashboards** using tools like Tableau or Power BI
- Try predictive models using Python + SQL pipeline

6 Conclusion

This project served as a practical application of SQL to derive actionable insights from a complex real-world dataset. By using Formula 1 data, we were able to formulate and answer a series of analytical questions, demonstrating how structured query language can be

employed to uncover patterns, evaluate performance metrics, and generate statistically relevant conclusions.

Throughout the process, key SQL concepts such as joins, aggregation, conditional logic, and string manipulation were explored in depth. Beyond technical proficiency, this exercise reinforced the importance of thoughtful question design, data cleaning, and logical consistency when working with large datasets.

From an academic perspective, this project lays the groundwork for more advanced exploratory and predictive analyses. Future directions may include integrating this dataset with telemetry or weather data, applying machine learning models to forecast race outcomes, or building interactive dashboards for dynamic visual storytelling.

In essence, this project highlights the intersection of data science, sports analytics, and database querying and underscores how domain-specific enthusiasm, when coupled with analytical rigor, can lead to impactful insights.