Stage 1: Data Discovery

Formula 1 World Championship from Vopani “Rohanrao”

Sourced from Kaggle

https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020

The dataset chosen by our group contains data on the Formula 1 (F1) World Championship from 1950 to 2024. F1 is a world-renowned auto-racing forum, and is widely considered a premier league for circuit racing. The world championships considered in this dataset are entire seasons appearing in any given year. One season comprises a series of races that takes place on distinct circuits and public roads across the world. The collection of races in sequence is referred to as “Grand Prix.”

The contents of this dataset are measured from the World Championships mentioned before. Specifically, the data pertains to: races, drivers, constructors, qualifying sessions, circuits, lap times, pit stops, and the championships overall. Constructors are the designers and builders of the cars used in races. Some additional supporting entities are driver standings, constructor standings and results, as well as sprint results. Sprints or sprint racers are shorter than circuits and do not require any driver to stop for pit stops. The dataset is split into distinct files, where each file contains an entity set. A last file designated “status” simply carries a mapping of various statuses describing different entities.

Common attributes among the entities are UIDs, names (for circuits, constructors, drivers, and races), and location/nationality (for circuits, constructors, and drivers). Entities also tend to have IDs corresponding to entity entries in other tables (like driver ID, circuit ID, constructor ID, etc) where appropriate (like for races, driver standings, results, etc). The attribute “points” belongs to results, constructor results and standing, as well as driver standings, quantifying number of points accumulated across the seasons. In addition, driver and constructor standings have an attribute describing the number of wins. An attribute regarding position in races and sprints are found in entities for constructor standings, driver standings, lap times, sprint results. Date and/or time attributes belong to majority of the entities aside from the status entity, where the date and/or time attribute may describe data of birth, data of race, year of season, or duration of a race. Driver entities have first and last name attributes. Remaining attributes are uninteresting.

Across all files and entity sets, there is a total number of 701433 records. Roughly 500000 records come from lap times, which aligns with the fact that many races partake in many seasons, all spanning many laps. Aside from that outlier, majority of the records come from constructor results and standings, driver standings, pit stops, qualifying races, and the results, which all have greater than 10000 records. The remaining entity sets of circuits, constructors, drivers, races, seasons, sprint results, and status, have between 70 to 1000 records. Some of the uninteresting attributes contain NULL values, like status of the constructor’s championship results, but some more interesting attributes like race lengths are NULL as well. Hence, there is some cleaning that will need to be done for this dataset, although not too much, and there are still interesting queries that can be considered without cleaning. Omitting the uninteresting attributes entirely is plausible since querying for those is not very insightful, but for total race lengths, it can be derived from the lap times of the last-place driver for any race. This would provide an estimate when the race should end, assuming the race did not end due to time constraints.

[talk about the actual process of cleaning the data]

[maybe remove the mentions of entities, and then replace those with other words so we don’t get clocked for not doing work, might also be worthwhile to specify what will be relationships as opposed to entities, but in a relational database they are all relations anyway]