Stage 1: Data Discovery

09/26/25

Group:

- Ayesha Qadir
- Kriza Cyrene del Moro
- Jatinder Sahota

Dataset Selection: 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.”

Data contained in the file(s)

The contents of this dataset are measured from the World Championships mentioned before. Specifically, the data consists of tables pertaining to the following:

- races
- drivers
- constructors
- qualifying sessions
- circuits
- lap times
- pit stops
- the championships overall

Constructors are the designers and builders of the cars used in races. Some additional supporting tables are driver standings, constructor standings and results, as well as sprint results. Sprints or sprint races are shorter than circuits and do not require any driver to stop for pit stops. The dataset is split into distinct files consisting of tables for the above data, and so we may prescribe entities using those tables. Then, much of the relevant data from the tables can be considered attributes, which we describe below. A last file designated “status” simply carries a mapping of various statuses describing different entities.

Defining entities using the dataset

We treat the tables as an entity set to establish our entities. Common attributes shared among these entities are: UIDs, which are assigned to all entities; names (for circuits, constructors, drivers, and races); and location/nationality (for circuits, constructors, and drivers). Entities also have IDs—such as driver ID, circuit ID, constructor ID, etc.—that reference entries in other tables like 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. The entities constructor standings, driver standings, lap times, and sprint results, are assigned an attribute related to position in races and sprints. Majority of the entities except for the status entity, possess date and/or time attributes—which may describe date of birth, date of race, year of season, or duration of race. Driver entities have distinct first and last name attributes. The remaining attributes will be excluded from this model.

On the size of the dataset, and cleaning efforts

Across all files and entity sets, there is a total number of 701,433 records. The entity, lap times, accounts for approximately 500,000 records—reflecting the extensive number of tracked laps across many races and seasons. Excluding the quantity of records contributed by lap times, the remaining records come from constructor results and standings, driver standings, pit stops, qualifying races, and the results, with each having greater than 10,000 records respectively. The remaining entity sets—circuits, constructors, drivers, races, seasons, sprint results, and status—each have between 70 to 1,000 records. Dataset cleaning will be necessary due to the presence of NULL values in primary attributes, such as race lengths. We can omit less-essential attributes since querying for such offers limited insight, but total race lengths 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. Fortunately, there is a lot of data to work with in this dataset—conveniently allowing us to remove any attributes that are riddled with NULL values and no sensible default value to replace them with. To this end, we may write a python script to simply remove any columns with NULL values. Otherwise, for attributes that can support a working default value, we can simply replace any NULL values with our chosen default. Creating a script to do this is straightforward. In order to start with the process, we will create a separate script to discover attributes that contain NULL values. Upon the collection of such attributes, we will decide on an individual basis whether or not we should omit the attribute or substitute in a valid default value—if available.

Group Timeline

Last updated: September 25th, 2025

Our timeline revolves around the deadline for each stage and weekly meetings that we hold. We intend to submit work for all optional deadlines. All our shared work is stored in a git repository.

Part A: Designing a Database

Stage 1: due September 26

Dataset Selection & Group Timeline (3%)

Completed before this timeline was established

Stage 2: due October 3rd

Optional ER Diagram Check-In (0%)

Before the meeting:

- every member will create a draft of an ER diagram and point-form justifications before our next meeting
- KC will summarize our initial write-up into one paragraph (ideally based on Stage 1 feedback, if possible)

Meeting on October 2nd at 1pm:

- discuss feedback from Stage 1
- finalize entities and relationships
- finalize participation and cardinality constraint

After the meeting:

- Ayesha will draw out the diagram by the end of the day on October 2nd
- every member needs to approve all the work to be submitted (paragraph, ER diagram)

Ayesha will submit the final copies of everything

Stage 3 & Reflection #1: due October 10th

Database Design (ER/EER model & relational model) (8%), Reflection (3%)

Before the meeting:

- every member will consider how to modify the model based on Stage 1 and 2 feedback, and how to incorporate EER elements
- Ayesha will modify the Stage 1 write up based on the feedback (as needed)
- Ayesha will create a draft for the updated timeline

- Jatinder will begin the process of moving from ER diagram to relational model (without any merging or normalizing)

Meeting on October 7th at 1pm:

- finalize the ER/EER diagram
- finish merging in the relational model
- discuss and begin normalizing the relational model
- discuss reflection and create bullet point answers to all questions
- assign tasks in updated timeline

After the meeting:

- every member is responsible for writing and submitting their personal reflections
- Ayesha will finish normalizing and finalize the relational model
- KC will turn bullet points from meeting into 3-5 paragraph group reflection
- every member needs to approve all the work to be submitted (updated write-up, ER/EER diagram, relational model, reflection)

KC will submit the final copies of everything

Part B: Query & Interface Design

Stage 4: due October 24th

Optional Query Check-In (0%)

Before the meeting:

- every member will consider (and perhaps draft up) potential queries for the database

Meeting:

- discuss all the queries and come up with a finalized list

After the meeting:

- every member needs to approve all the work to be submitted (ER diagram, potential queries)

Jatinder will submit the final copies of everything

Stage 5: due November 7th

Query Design (5%)

Before the 1st meeting:

- every member will review the feedback from Stage 4 and consider how to improve our queries

Meeting 1:

- discuss strategies for query improvement

- record any modifications/new queries

Before the 2nd meeting:

- potentially attend office hours for advice, if needed

Meeting 2:

- implement any feedback from office hours/Stage 4 feedback
- come up with a new finalized list of queries (if necessary)
- potentially begin work on Stage 6

After the meetings:

- every member needs to approve all the work to be submitted (ER diagram, potential queries)

Ayesha will submit the final copies of everything

Stage 6 & Reflection #2: due November 21st

Interface Design (5%), Reflection (3%)

Could potentially have 2 weekly meetings, though one would fall during Reading Week. Ideally, we delegate tasks before Reading Week and are able to complete them on our during the break.

Part C: The Database and its Interface

This includes…

- Project Demonstration (10%) – December 1 - 5
- Final Project Submission (35%) – December 5
- Final Report (25%) – December 5
- Reflection #3 (3%): Individual and Group Reflection – December 5