Gustavo Cesar, Justin Gill, Jack Griffin,
Simon Naylor, Jared Pursiano
MSBX 5405
12/10/2021

SQL Database Report: Formula 1

Project Summary

Formula One has a massive worldwide following. Up until 2019, this was not the case for the United States. With little representation in racing drivers and race tracks on the schedule, they needed to do something to make a splash in the market. They decided on making a Hard Knocks-style documentary that followed teams and drivers throughout the season named “Drive to Survive”. This worked spectacularly and now with a massive rise in popularity of Formula One in the US market, For made deciding on F1 data as our project a simple decision. Formula One Media (FOM) needs to continue to make smart business decisions. Within two years of the release of their docu-series, US race viewership has nearly doubled (Smith, 2021). How do you get fans to pick teams and drivers, and how can you easily show the history of the sport in informatics? “Drive to Survive” only takes place in the current day, should there be a historical adaptation to it as well? One that would capitalize on the current US craze?

These cars are the fastest in the world, which means they have to be advanced. With advances in racing technology, mountains of data have to be collected. Due to the sport being very transparent, they also make data about the racing very transparent as well. Having a multitude of data combined with a great starting question, we knew this data was right for us. Having easy access to data allowed us to think of any question, and have the answer buried somewhere in our database. Our group wanted to be able to make a representation of the history of Formula 1 and convey it in a digestible way. By making a historical representation of “Drive to Survive”, we think that FOM could capitalize on the current craze to not only educate new fans, but to build strong allegiances to teams and races.

It is extremely hard to find sports data open to the public. Our group broadly enjoys sports, but sports data is extremely hard to find. Formula One is quite new to us all. We have all started following F1 to some degree within the past 5 years, so to make a project on it would also further educate ourselves. As a group, we could also dictate what went into our database in terms of how many years we wanted to reach back, lap times, pit times, and even broader information such as Constructor’s and Driver’s championships. Due to this, each group member could easily find a different place to explore within our sprawling database.

This data consists of numerous tables with a plethora of information. The first table is the constructors table, consisting of the name of the racing teams, an auto generated id, their country of origin, an abbreviation for reference, and a web address to their online site. That table connects to the constructor results table that combines the constructor ids with racing results by race, that is points scored, the status of if the racing team finished the race, crashed, disqualified, or had other mechanical issues. Both of these tables relate to the constructors standings table which, similar to the constructors results table, keep track of points and finishing position per race.

The next important table is the circuits table. This table contains all of the track data. An id for each circuit, the name, location, country, latitude, longitude, altitude for each track. After that comes the drivers table. This table houses all of the driver's information. An id for each driver, their first and last names, their date of birth, country of origin, their number, and a website to a wikipedia page of more detail about them. The drivers table, just like the constructors table, is connected to the driver standings table that combines the driver information with the finishing position of the drivers, their total points, and number of wins.

In our database, races and results are pivotal tables when querying data. The races table connects to virtually every other table and includes an id for each race, the year, date, and time of the race, as well as the id of the circuit relating to the circuits table. In addition, the results table is the largest table in our database. It contains information regarding the specific race id, driver id, and constructor id all connecting to their respective tables. In addition to the position each driver/constructor finished per race, their time in milliseconds, their fastest lap number, fastest lap time, fastest lap speed, and the status of if the racing team finished the race, crashed, disqualified, had other mechanical issues, etc.

Finally in relation to the races and results tables, the lap times, qualifying, and pit stops tables also play a vital role in further specificity of our data. In the lap times tables it has the position and time in milliseconds per driver and race determined by the driver and race ids. Much like the previous table the qualifying table shows the position and qualifying time per driver, race, and constructors categorized by their respectives ids. Lastly, the pit stops table shows the amount of pit visits, their duration in milliseconds, and the lap they were taken based on the specific race and driver ids.

These tables consist of valuable data that drives our queries and provides a key structure to the network of our data. The most important data to our group was driver's and constructor's IDs. These were the base of almost every query because our questions and ideas were all based around fandoms. We had a couple queries that relied on nationality as opposed to driver or team allegiances. This is due to countries like Brazil treating their drivers as heroes and the associated advertisements and engagements surrounding cultures as well as country-specific products and companies. Another would be driver or constructor points. Teams and drivers are granted points based on their finishing positions in races. The sum of these at the end of every season determine which driver and constructor wins the championships. Points are great for describing how fast a car is as well as a driver.

Wins was another important column found in constructor and driver data. Constructor wins, like points, are the sum of the production of the drivers on the team. These were separate and very useful to use. Individual wins are the product of being a fast driver in a fast car. In a constructor format, these can be put next to other drivers on the team, to figure out if a certain driver is faster or not with the same machinery. We did not focus on this too much, because we were showcasing drivers and not their rivalries with teammates. Such could be done, but it is not within the scope and size of this project.

Transformations were very light as our data was precise and neat. We had to transform milliseconds to seconds for lap times as well as pitstops. Another that needed modifying was wins in certain circumstances where cumulative values were not necessary. This was dealt with by lagging or just using the MAX() function in mySQL and using a simple Tableau query for visualizations. For these reasons, and the variable's dual use, we decided to not permanently alter it, and transform in views and subqueries as necessary.

Our data was normalized from the start. Formula 1 has a great data stream that has pre-cleaned data, and is just mere steps and a few scripts away from putting the data into a database. Some tables such as 'results' and 'constructor results' were very similar and could be merged, but that would then omit quickly accessible data that could be misinterpreted or lost in a combination. The same goes for tables like 'lap times' and 'races' where the data is similar, but the tables can be used in vastly different ways.

"'Drive to Survive' on Netflix Has Ignited Formula 1." *The New York Times*, 2021, www.nytimes.com/2021/07/16/sports/autoracing/drive-to-survive-netflix-formula-one.html. Accessed 7 Dec. 2021.