F1 Racer Efficiency

**Dataset Name**: Formula 1 World Championship (1950 - 2024)

**Dataset Link:** <https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020?select=pit_stops.csv>

**Student's Names:** Troy Benner, Kiefer Jenny, Will Moore

### 1. Use Case Description

This project centers around developing a data-driven analytics platform that provides insights into the efficiency and skill of Formula 1 drivers. Using historical data from the Formula 1 World Championship spanning from 1950 to 2024, the platform will analyze individual drivers' performance metrics, including standings, lap times, pitstop efficiency, and qualifying results. The primary objective of this platform is to enable racing analysts, F1 enthusiasts, and potentially betting platforms to access detailed statistics and rankings related to a driver's performance, skills, and efficiencies over time.

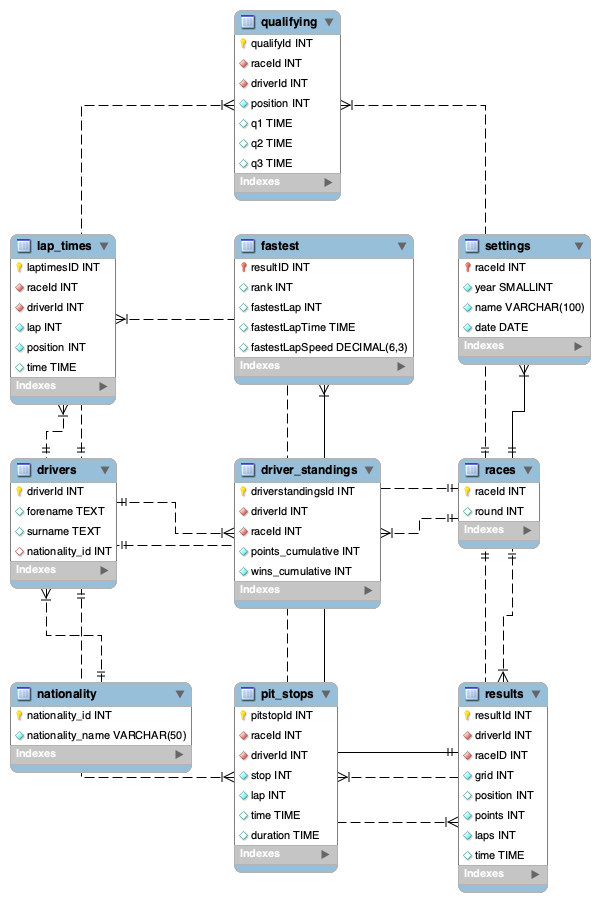
The platform will process and analyze data from the following tables: driver\_standings.csv, driversId.csv, raceId.csv, lap\_times.csv, pit\_stops.csv, qualifying.csv, and results.csv. These datasets contain critical information related to specific driver performance metrics that enable us to structure our platform. Users will view individual drivers and inspect detailed performance statistics, such as average lap times, pit stop efficiency, qualifying positions, and overall race results. The goal is to rank drivers based on individual performance metrics that reflect their consistency, speed, and strategic efficiency in races. This platform will aid enthusiasts and analysts in understanding driver performance trends and provide valuable data points that could be used for predictive insights, such as sports betting.

By focusing on core performance metrics and presenting them in a clear, user-friendly format, this platform provides a valuable tool for anyone looking to understand or analyze nuanced data points of F1 driver efficiency.

### 2. Relationship Rules

* Drivers have one cumulative standing for their careers
  + Each standing belongs to one driver
* Each driver has zero to many lap times for every race
  + Each race has zero to many lap times for every driver
* Each driver can have zero to many pit stops for every race
  + Each pit stop belongs to one driver
* Each driver has one to many results throughout a career
  + Each result belongs to one driver throughout a career
* Each driver has zero to many qualifying positions
  + Each qualifying position belongs to one driver
* Each driver has only one nationality
  + Each nationality can belong to many drivers
* Each race has many results
  + Each result belongs to one race
* One race has zero to many qualifying positions
  + Each qualifying position belongs to one race
* One race has many lap times
  + Each lap time belongs to one race
* One race has many pit stops
  + Each pit stop belongs to one race
* Each race occurs at one location and time
  + One location and time belongs to one race
* Each race awards points to drivers increasing their cumulative statistics
  + A driver’s cumulative standing is influenced by one to many races
* Each result has a fastest lap
  + The fastest lap of each result belongs to one driver
* Each qualifying position has one time for each session
  + Each session time belongs to one qualifying position

### 3. Entity Relationship Diagram (ERD)



### 4. Data Exploration

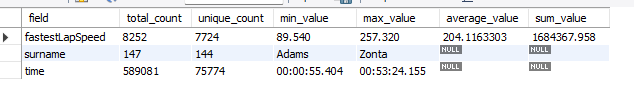
Identify 3 key fields:

**Numeric:** fastestLapSpeed

**String:** surname

**Date/Time:** time

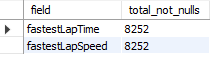
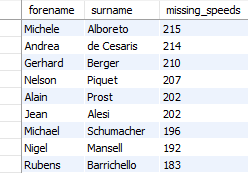
*Union between those fields*



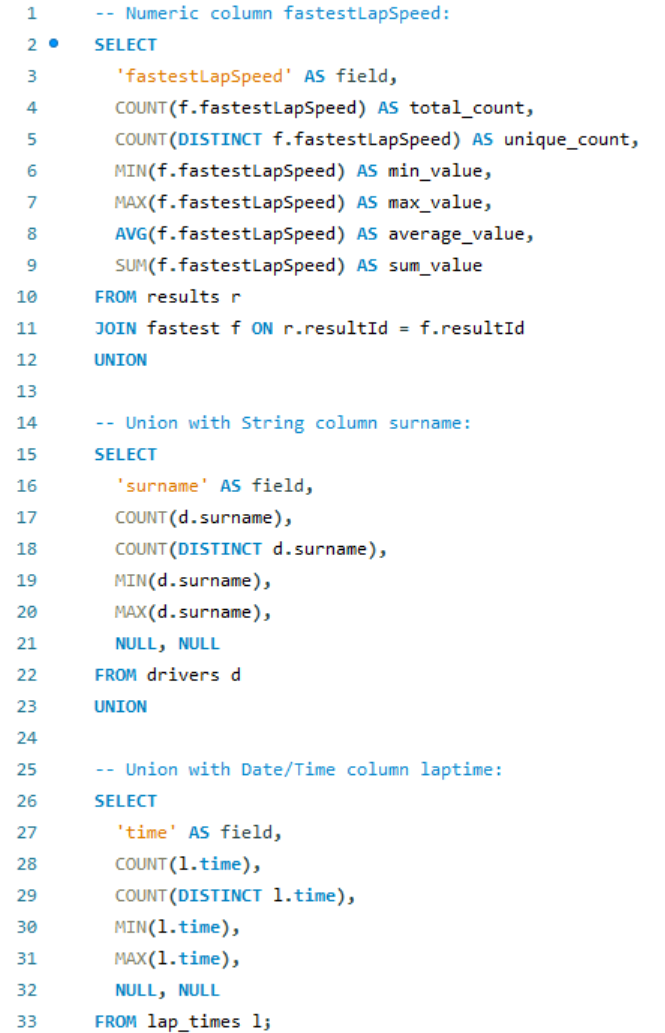
Analysis:

There is a concerning number of NULL values (18,507) in both fastestLapSpeed and fastestLapTime. However, this is explained by the context of the dataset. These NULLs are primarily due to the age of the data; fastest lap data was not included until the 2004 F1 season. This also accounts for data limitations such as incomplete telemetry, recording errors, and situations like driver DNFs (Did Not Finish), where no fastest lap was recorded. For example, Michele Alboreto’s entire F1 career lasted from 1981 to 1994, before the fastest data was included in our dataset. He participated in 215 total races and has 215 null values in fastestLapSpeed and fastestLapTime.

### 



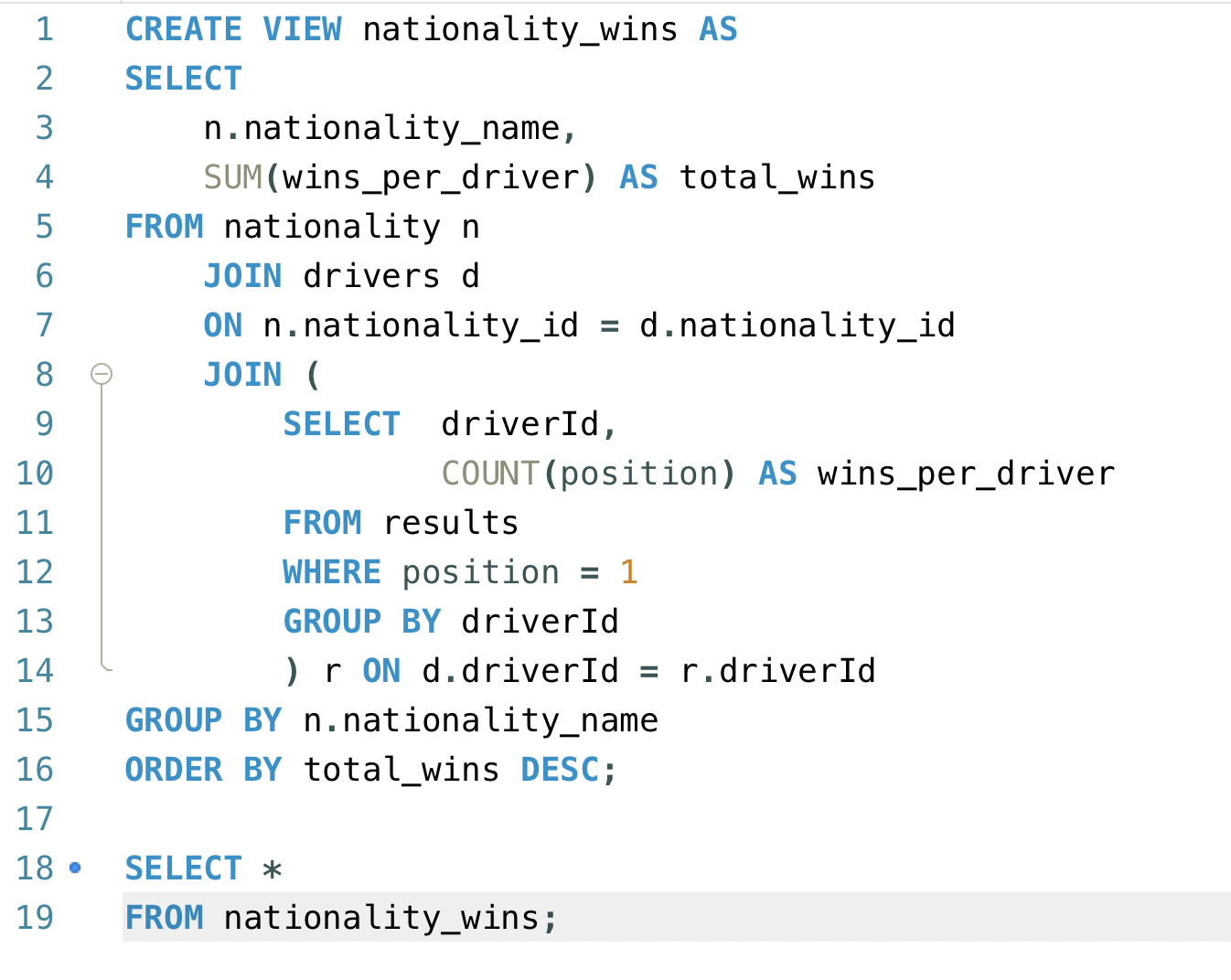
According to our analysis of name data (surname and forename), there are no null values indicated by the total\_problematic column. Additionally, the time field fetched from lap\_times has no null values present in its dataset.



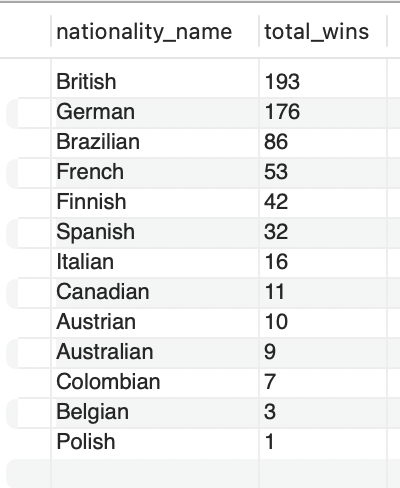
### 5. Data Analysis

Analysis 1:

This view returns the number of First-Place finishes (aliased as total\_wins) that each country has, ordered by the total number of wins from highest to lowest.



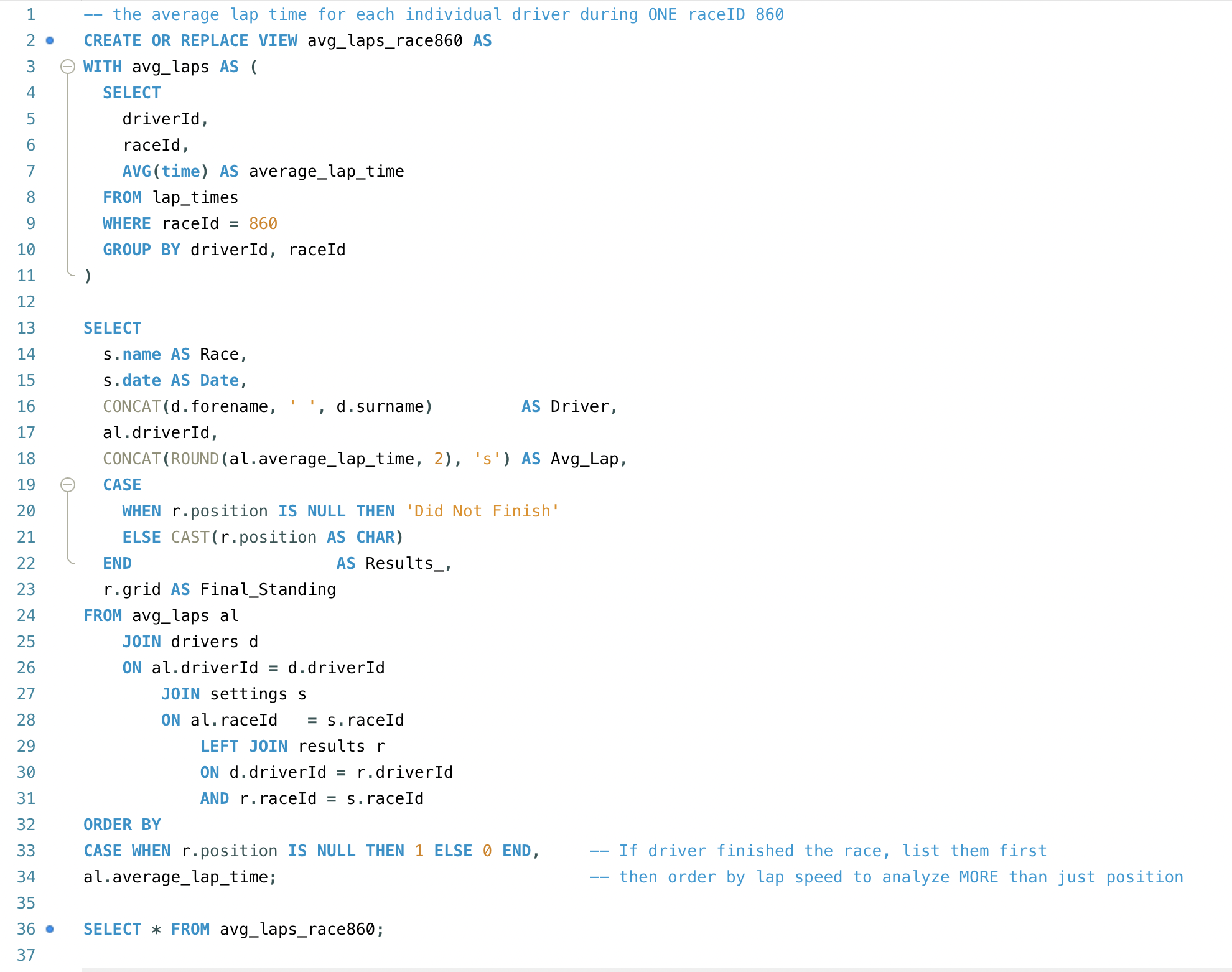
The Nationality’s Total Wins Results:

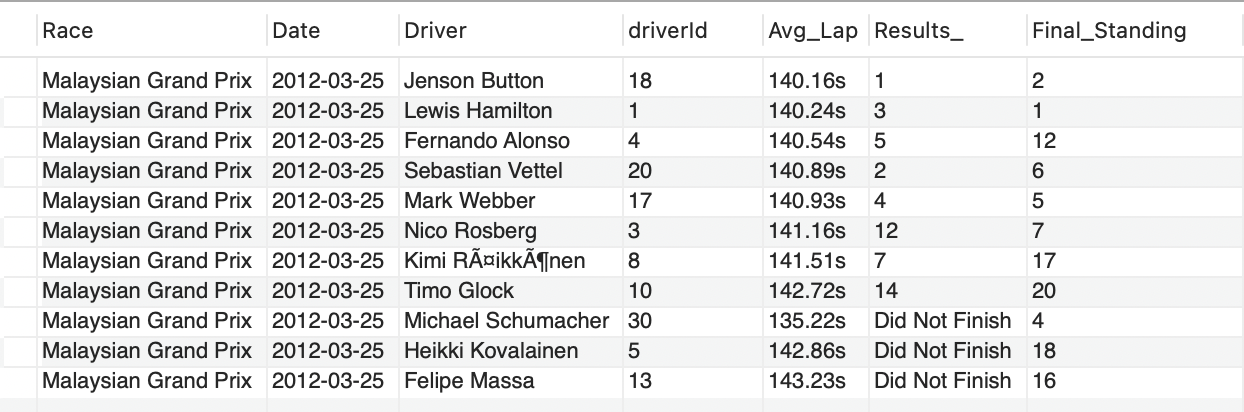


Analysis 2:

This view shows the average lap speed for each driver who raced in the Malaysian Grand Prix on 3/25/2012. The average lap speeds are pulled from a CTE, which looks for results where the *raceId* is 860. The view compares the average lap speed with the finishing place of each F1 driver in race 860. The view shows that the fastest racers, via their average lap speed, are not necessarily the top-placing racers. It also shows how fast the racers who didn’t actually finish the race were completing their laps on average, and what their final position was after their teams points were adjusted.

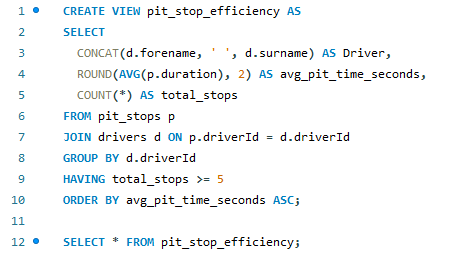
**Important Note:** The F1 Kaggle database is missing 100s of drivers in the driver table, but any given race only has about 24 drivers. Any analysis of data must be understood through this lens.

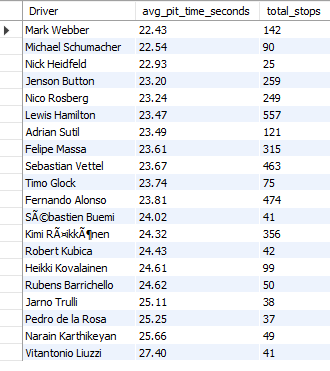




Analysis 3:

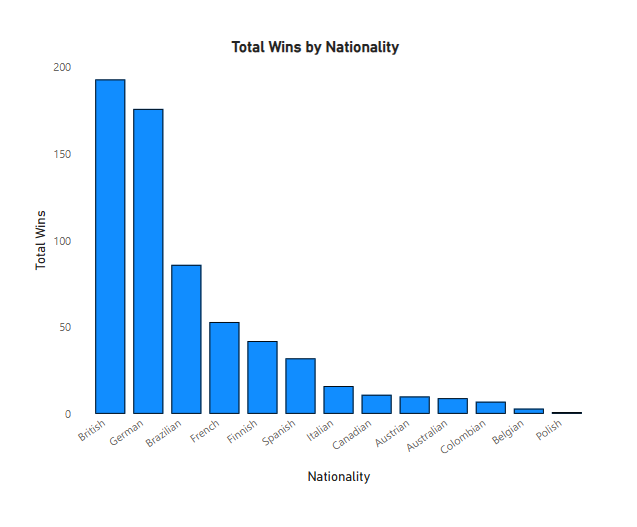
This view returns a rank of Formula 1 drivers based on their average pit stop duration. It includes only those with 5 or more total stops so as to not skew the data. The goal is to identify which drivers (and by extension, teams) had the most efficient pit stops and pit crews.



**Results:**  


### 6. Key Findings

Key Finding 1:

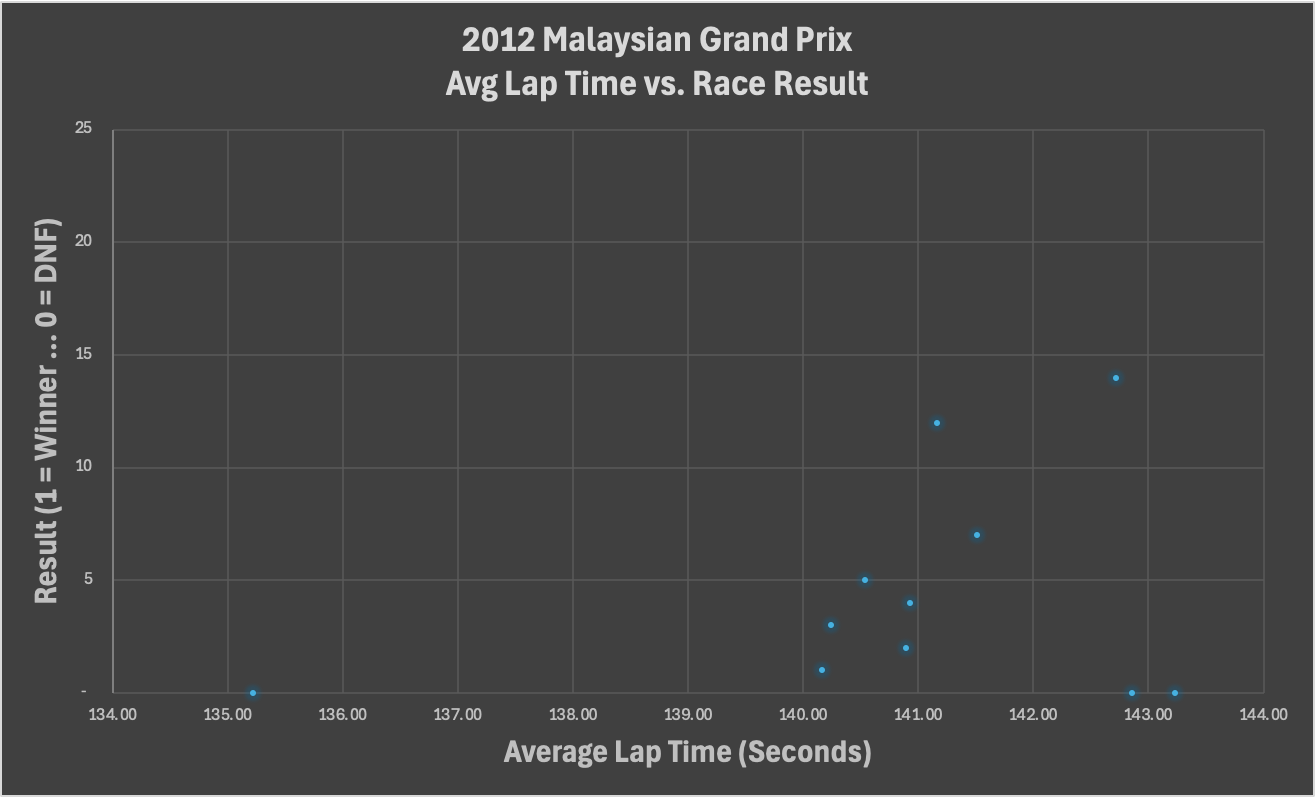


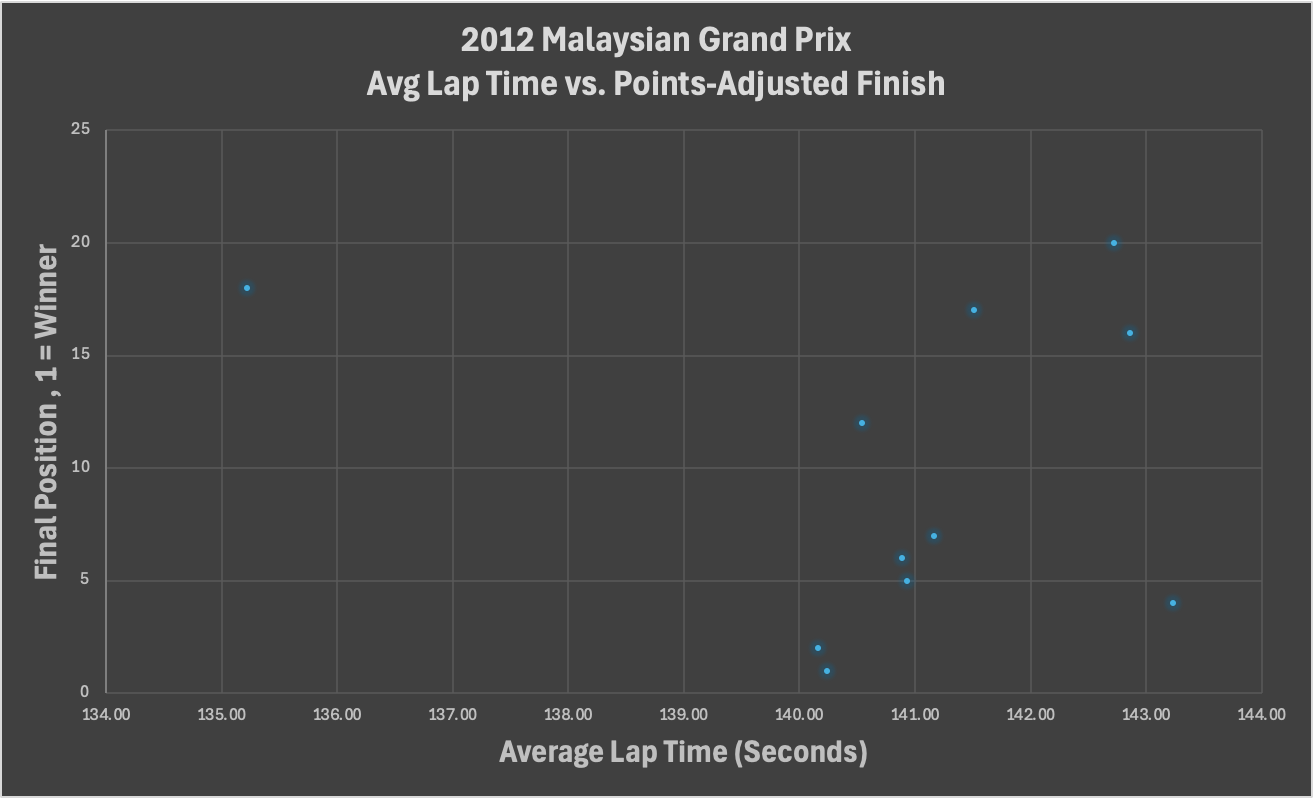
This graph showcases the results from the nationality\_wins view. This reveals which countries have the highest and lowest number of Formula One championship wins. Britain and Germany are the top nations, with Brazil and France behind. Countries not pictured on this graph do not have a Formula One championship win in the database. Interestingly, the United States of America is not pictured on this graph, meaning they have no wins in this Kaggle database. Despite being one of the best nations for international sports, as shown by the Olympics, America struggles in Formula One racing in comparison to smaller countries with Gross Domestic Products that are tiny in comparison. This is also surprising given the United States’ interest in racing, just in the NASCAR format instead of F1.

Key Finding 2:

The *Avg Lap Time VS Race Result* (Pre Points Adjustment)chart depicts the average lap time of each driver against which place they cross the finish line. This doesn’t solely determine whether they win or lose because drivers still must go through an adjustment of “points” before their final time is calculated. For that reason, I’ve added an *Avg Lap Time VS Points-Adjusted Finish* chart.

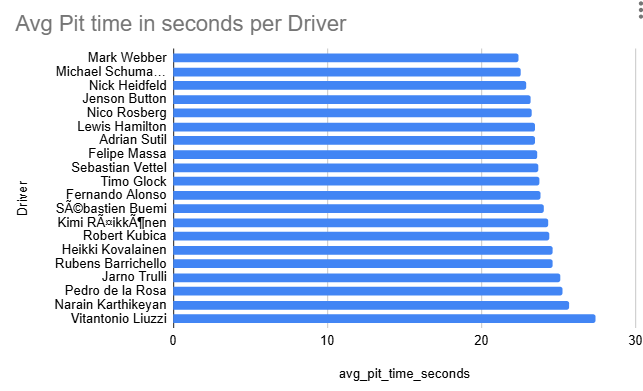
From our *Race Result* chart, we can see that both the fastest driver, with an average lap time of 135.22 seconds, and the slowest driver, with an average lap time of 143.23 seconds, did not finish the race, with “0” place representing that they did not finish. To explain both of these, in our research, we found that there are many reasons why a driver could not finish a race. For example, sometimes drivers crash, but other times, drivers have engine trouble and choose to slow down to preserve their engine.   
 Visually, on our *Race Result* scatter plot, we can see that generally, as a driver’s average lap time goes up, they come in closer to last place. After we factor in points and penalties, the negative trend is largely still in place, suggesting that it’s still a good strategy to finish fast instead of driving cautiously slow. Lastly, outliers that did not finish the race still get a place in the final standings. One outlier placed in the top 5, which suggests that there is still strategy to be had, even though they didn’t finish.





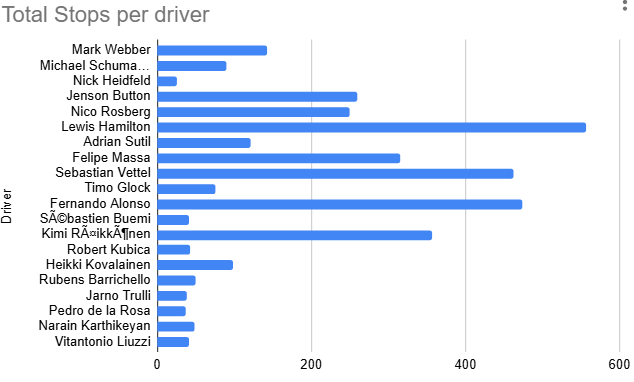
Key Finding 3:

The two charts that were shown outline the average duration of pit stops per driver against the sum of all the stops made by each driver. According to this data, Mark Webber benefits the most out of the drivers, with the fastest average pit stops at 22.43 seconds, followed by Michael Schumacher and Nick Heidfeld. It seems clear that these drivers had high-functioning pit crews. On the other end, Vitantonio Liuzzi and Narain Karthikeyan had the slowest average pit times above 25 seconds. This will likely yield negative results for the racers regarding their overall race positions. Surprisingly, some drivers maintained fast averages despite a high volume of pit stops. For instance, Lewis Hamilton and Sebastian Vettel each recorded well over 400 stops yet somehow managed to average below 24 seconds. The total stops graph also reveals the tendency of some other high-caliber drivers, such as Fernando Alonso, Kimi Räikkönen, and Jenson Button, to frequently go into the pits, perhaps due to age or particular racing strategies. In any event, the data would appear to demonstrate that as an athlete’s workload increases, the performance consistency and speed during the pit stops become more refined.



X Axis = The average time spent during a Pit Stop in seconds

Y Axis = Top 20 Drivers with regard to Pit Stop time average



X Axis = Number of pit stops recorded by the data set

Y Axis = Top 20 Drivers with regard to Pit Stop time average

### 7.1 Data Dictionary

### 

### Table: driver\_standings

| **Attribute** | **Data Type** | **Key** | **Description** |
| --- | --- | --- | --- |
| driverstandingsId | INT | PK | Unique identifier for each driver standings record. Used to track a driver’s cumulative performance over the season. |
| driverId | INT | FK referencing drivers(driverId) | Identifier of the driver whose standings are being recorded. |
| raceId | INT | FK referencing races(raceId) | Identifier of the race after which the standings are updated. |
| points\_cumulative | INT |  | Total championship points the driver has accumulated up to the current race. |
| wins\_cumulative | INT |  | Total number of race wins the driver has achieved in the season. |

————————

### Table: drivers

| **Attribute** | **Data Type** | **Key** | **Description** |
| --- | --- | --- | --- |
| driverId | INT | PK | Unique identifier for each driver. |
| forename | TEXT |  | Driver’s first name. |
| surname | TEXT |  | Driver’s last name. |
| nationality\_id | INT | FK referencing nationality(nationality\_id) | Identifier for the driver’s nationality. Important for team and fan identities in F1. |

————————

### Table: fastest

| **Attribute** | **Data Type** | **Key** | **Description** |
| --- | --- | --- | --- |
| resultID | INT | PK, FK referencing results(resultId) | Unique identifier for the fastest lap record. Also links to the overall race result. Cascading updates/deletes ensure referential integrity. |
| rank | INT |  | Rank of the lap time among competitors (e.g., fastest, second fastest). |
| fastestLap | INT |  | The lap number on which the fastest lap was set. |
| fastestLapTime | TIME |  | The recorded fastest lap time during the race. |
| fastestLapSpeed | DECIMAL(6,3) |  | The average speed (typically in km/h) achieved on the fastest lap. |

————————

### Table: lap\_times

| **Attribute** | **Data Type** | **Key** | **Description** |
| --- | --- | --- | --- |
| laptimesID | INT | PK (AUTO\_INCREMENT) | Unique identifier for each lap time record. |
| raceId | INT | FK referencing races(raceId) | Identifier of the race during which the lap was recorded. |
| driverId | INT | FK referencing drivers(driverId) | Identifier of the driver who completed the lap. |
| lap | INT |  | The lap number within the race. |
| position | INT |  | The driver’s position at the end of the specific lap. |
| time | TIME(3) |  | The lap time achieved by the driver, including millisecond precision. |

————————

### Table: nationality

| **Attribute** | **Data Type** | **Key** | **Description** |
| --- | --- | --- | --- |
| nationality\_id | INT | PK (AUTO\_INCREMENT) | Unique identifier for each nationality. |
| nationality\_name | VARCHAR(50) | Unique | Name of the nationality (e.g., British, Italian, German). |

————————

### Table: pit\_stops

| **Attribute** | **Data Type** | **Key** | **Description** |
| --- | --- | --- | --- |
| pitstopId | INT | PK (AUTO\_INCREMENT) | Unique identifier for each pit stop record. |
| raceId | INT | FK referencing races(raceId) | Identifier of the race in which the pit stop occurred. |
| driverId | INT | FK referencing drivers(driverId) | Identifier of the driver making the pit stop. |
| stop | INT |  | Sequential number of the pit stop during the race (e.g., 1 for the first stop, 2 for the second). |
| lap | INT |  | The lap number during which the pit stop took place. |
| time | TIME |  | The clock time at which the pit stop was initiated. In F1, this can indicate race strategy and timing. |
| duration | TIME(3) |  | The duration of the pit stop, measured with millisecond precision, reflecting how quickly the stop was completed. |

————————

### Table: qualifying

| **Attribute** | **Data Type** | **Key** | **Description** |
| --- | --- | --- | --- |
| qualifyId | INT | PK | Unique identifier for each qualifying session record. |
| raceId | INT | FK referencing races(raceId) | Identifier of the race for which the qualifying session is held. |
| driverId | INT | FK referencing drivers(driverId) | Identifier of the driver participating in the qualifying session. |
| position | INT |  | The grid position earned by the driver based on their qualifying performance. |
| q1 | TIME(3) |  | Lap time achieved during the first qualifying session (Q1). |
| q2 | TIME(3) |  | Lap time achieved during the second qualifying session (Q2), if applicable. |
| q3 | TIME(3) |  | Lap time achieved during the third qualifying session (Q3), if applicable. |

————————

### Table: races

| **Attribute** | **Data Type** | **Key** | **Description** |
| --- | --- | --- | --- |
| raceId | INT | PK | Unique identifier for each race. |
| round | INT |  | The round number of the race within the current F1 championship season. Often indicates sequence in the season. |

————————

### Table: results

| **Attribute** | **Data Type** | **Key** | **Description** |
| --- | --- | --- | --- |
| resultId | INT | PK | Unique identifier for each race result record. |
| driverId | INT | FK referencing drivers(driverId) | Identifier of the driver whose result is being recorded. |
| raceID | INT | FK referencing races(raceID) | Identifier of the race corresponding to the result record. Note: Ensure consistency in column naming (raceId vs. raceID). |
| grid | INT | – | The starting grid position of the driver at the beginning of the race. |
| position | INT | – | The final finishing position of the driver at the end of the race. |
| points | INT | – | Championship points awarded to the driver based on their finishing position. |
| laps | INT | – | The total number of laps the driver completed during the race. |
| time | TIME | – | The overall race time or time difference relative to the winner (if applicable). |

————————

### Table: settings

| **Attribute** | **Data Type** | **Key** | **Description** |
| --- | --- | --- | --- |
| raceId | INT | PK, FK referencing races(raceId) | Unique identifier for the race settings record; ties settings directly to the race. |
| year | SMALLINT | – | The calendar year in which the race took place. |
| name | VARCHAR(100) | – | The official name of the race (e.g., “British Grand Prix”). |
| date | DATE | – | The date on which the race was held. |

### 7.2 Screenshots of All Tables and Relationships

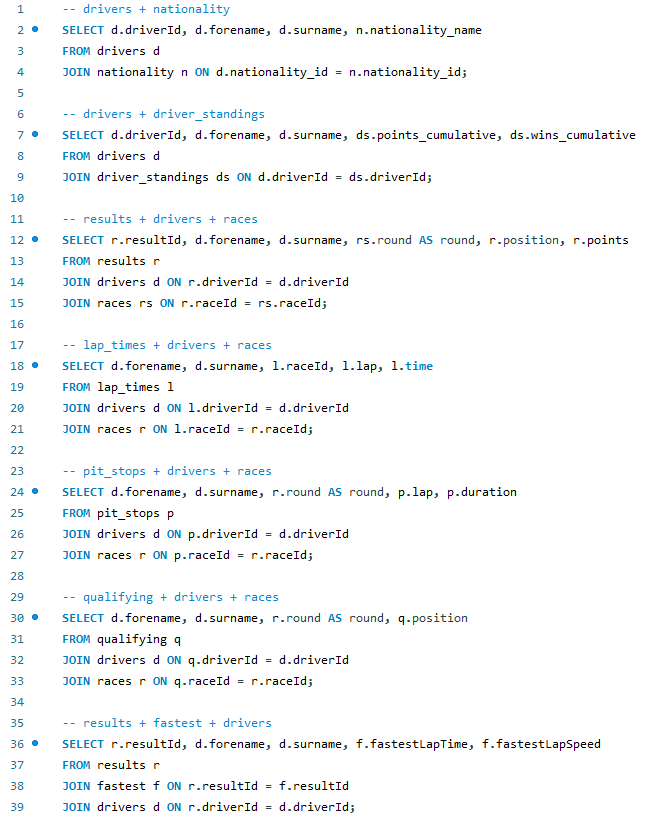
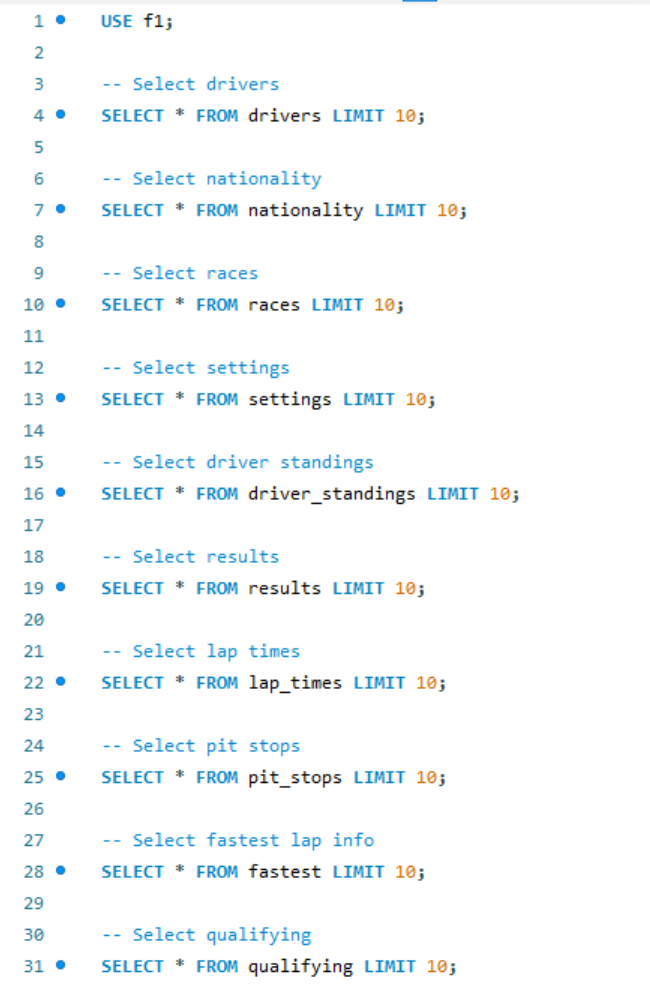
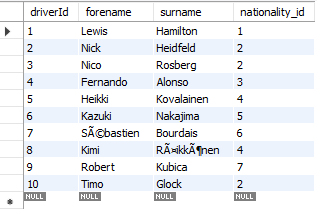
Join Code:  


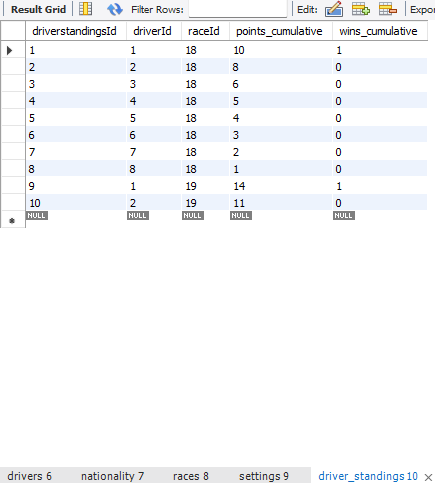
Table Selects:



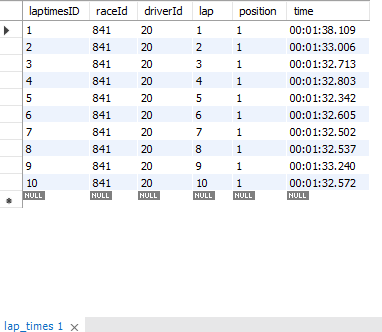
Drivers:



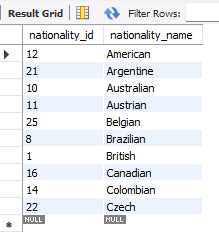
DriverStandings:



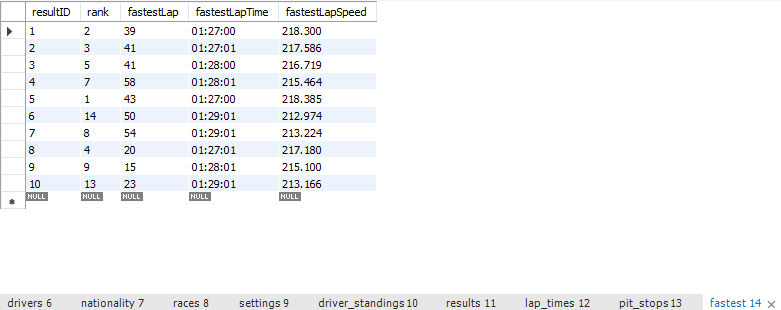
Laptime:



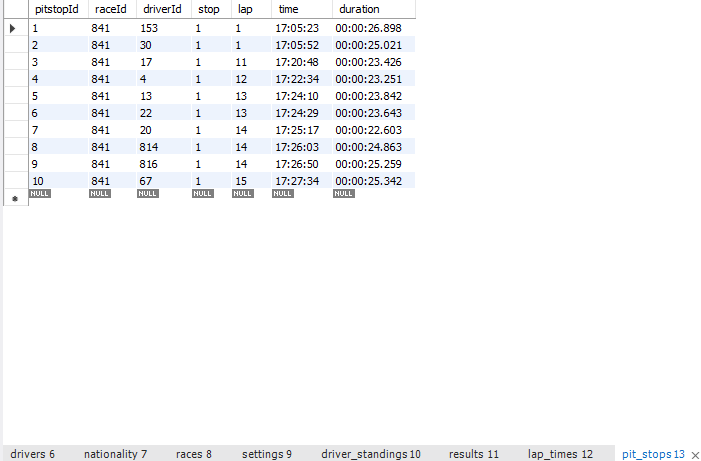
Nationality:



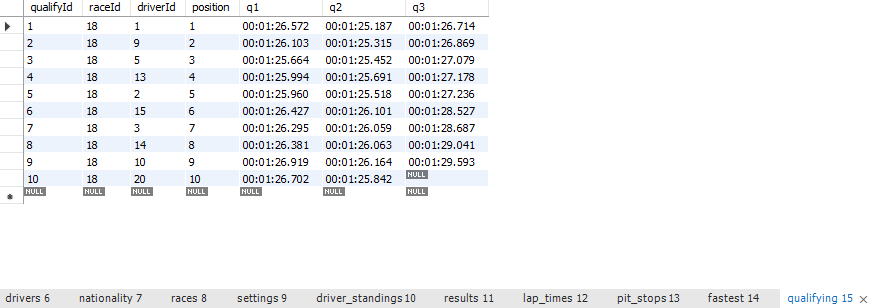
Fastest:



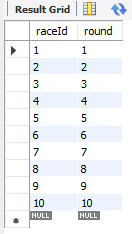
PitStops:



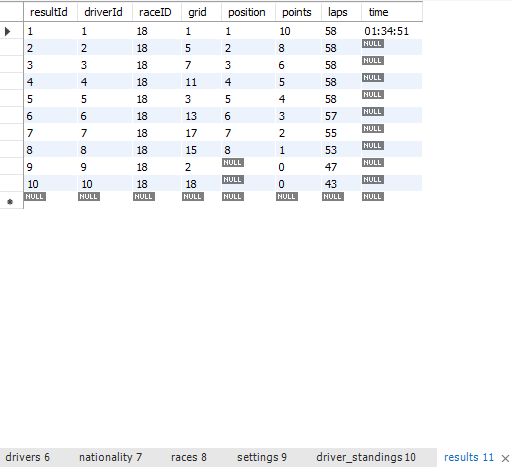
Qualifying:



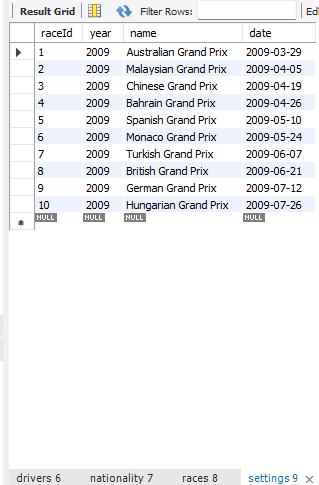
Races:



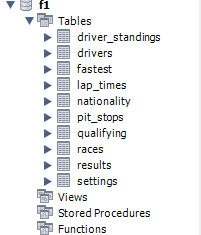
Results:

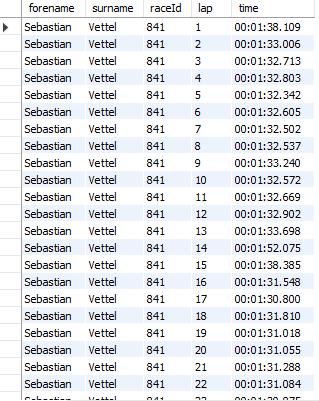


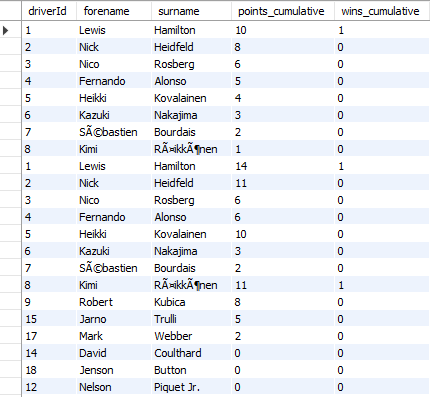
Settings:



Tables:

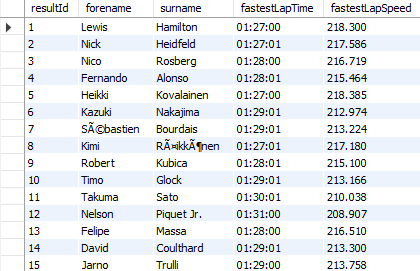
  
Join between Driver and LapTime:

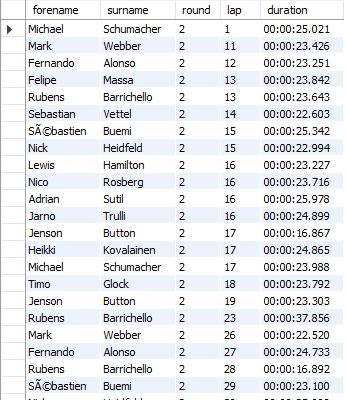


Join between Driver and Races:  
  
Join between Driver and Nationality:

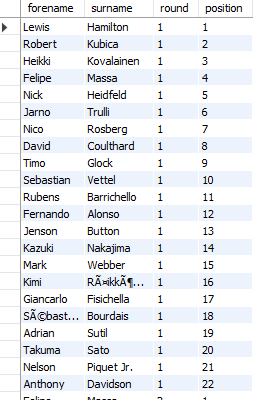


Join between Results and Fastest:

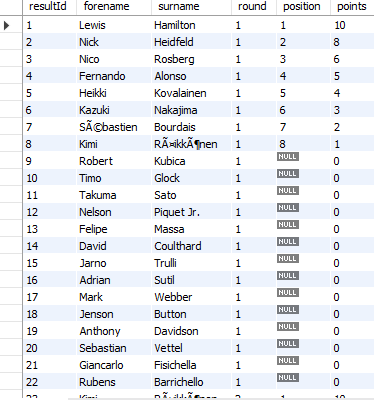


Join between Driver, Races and PitStops:  


Join between Driver, Races and Qualifying:



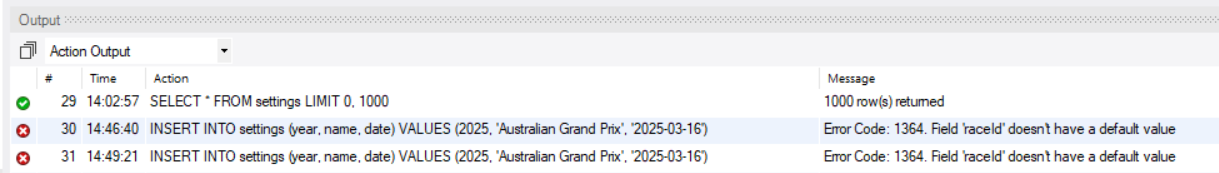
Join between Driver, Races and Results:



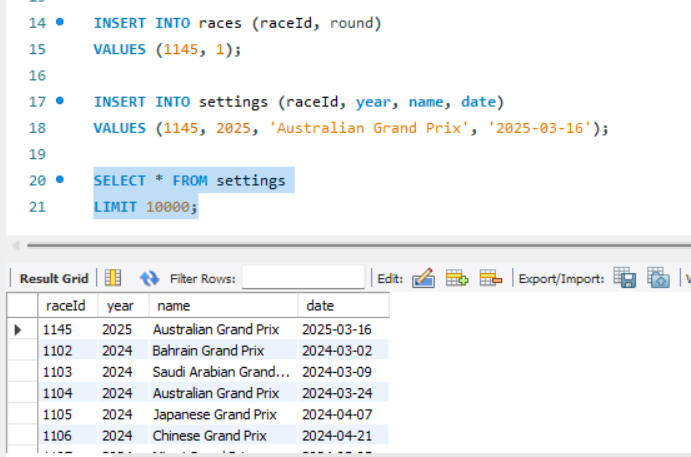
### 7.3 Data Modification

Insert Data:

We are inserting a new Grand Prix into the races and settings table. It is necessary for both tables to be updated as they are related to each other, and the *setting* table’s primary key is a foreign key in the races table.

Failed Attempt: Error Code 1364 due to inserting the data in an incorrect order

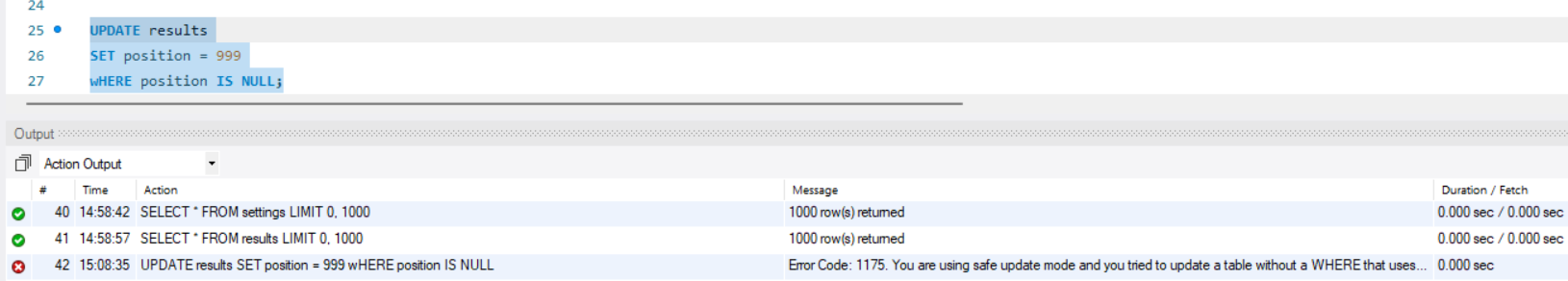
Successful Attempt: Inserted the raceId and round into values first so I had a raceId value to insert into settings



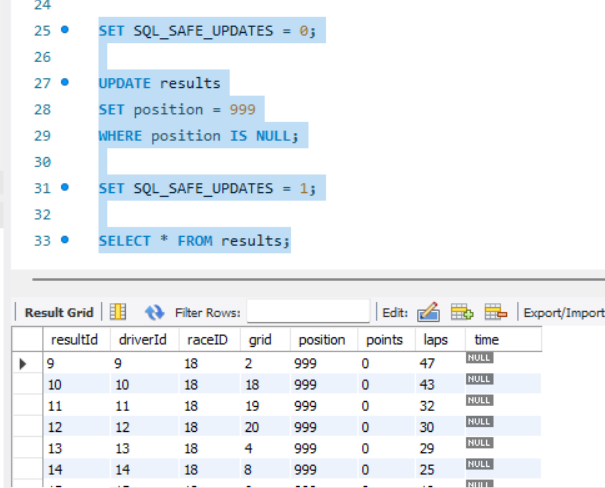
Update data:

We updated the data by setting null values in the position column of the results table to 999. This makes the viewer aware that this was an incomplete time and position, while not being null, to avoid any issues. However, they must be made aware that 999 represents null for this column.

Failed Attempt: Error Code 1175. Safe mode prevents updating without a where clause using a primary key.



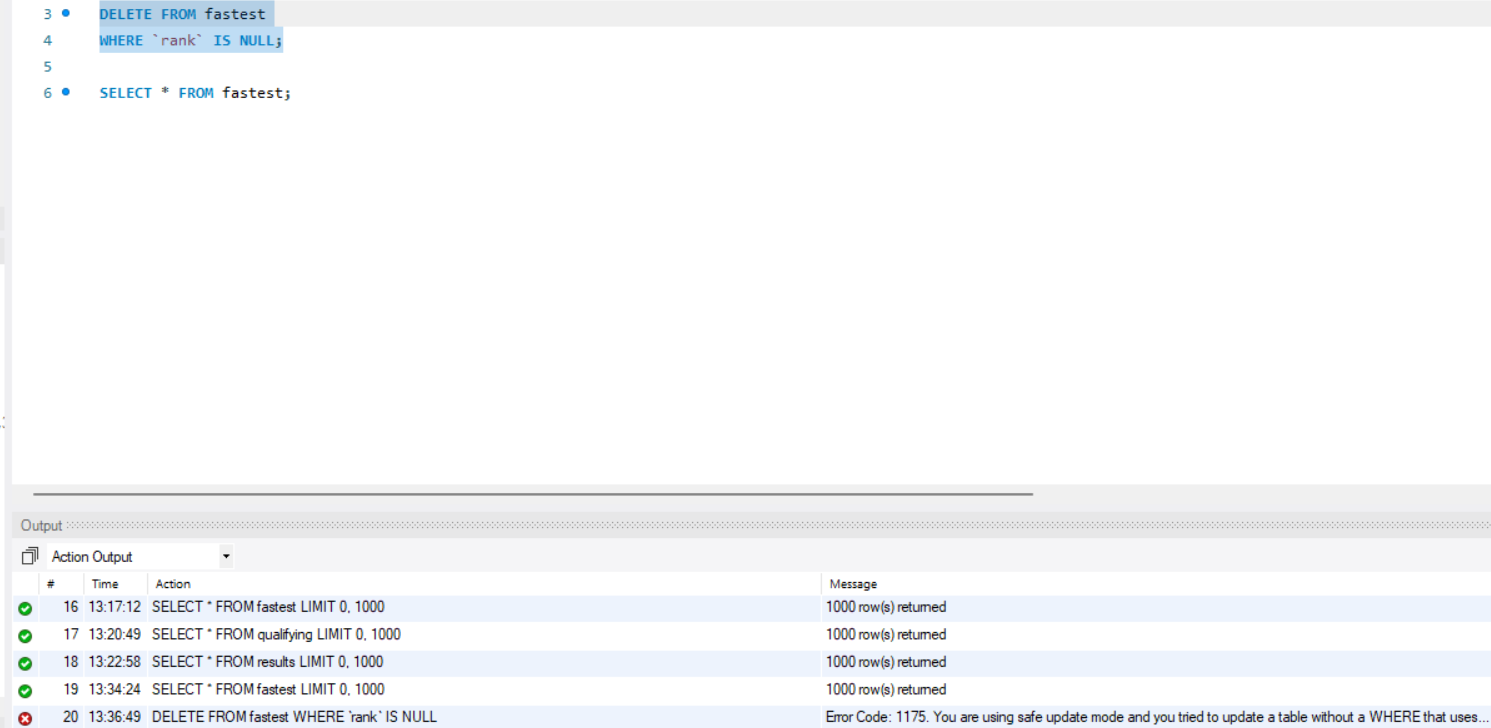
Successful Attempt: Disabled safe mode temporarily to edit the data, and then turned safe mode back on.



Delete data:

Deleting rows in the fastest table that have null values for rank, etc. This is due to this statistic not being tracked until 1995.

Failed Attempt: Error Code 1175. Safe mode prevents updating without a where clause using a primary key. Same problem I ran into during the updating process.



Successful Attempt: Updated Safe Mode Temporarily 