# MGS613 Final Project: Group 6

*Harshul Srivastava, Jathin Suresh, Swaroop Ravindranath, Piyush Ghule, Avinash Venugopal*

## 1. Introduction

***Races are won at the track. Championships are won at the factory.*** - Mercedes AMG [1]

Formula 1 is an exciting motorsport competition featuring the best drivers and constructors in the world competing in more than 20 circuits spread across 6 continents. Founded in 1950 with just seven events in its first year, a typical season today consists of over 20 championship races, each with its own set of qualifying trials to determine the starting position of cars on the grid.

Each race, called a Grand Prix, can be up to 2 hours long, with drivers completing distances of approximately 190 miles. During the race, drivers often take strategic pit stops to change tires and perform other light maintenance. As such, the order of the drivers on the track keeps changing, which is monitored closely by the race stewards and shared live on TV using status indicators and a 3-letter reference name (eg. HAM for Lewis Hamilton). The race ends when the drivers either cross the line with a finishing time, or fail to do so entirely, resulting in a DNF status.

F1 is enjoyed by fans around the world, including members of our team, which provided the primary motivation for this project. This investigation aims to use analytical modeling and SQL to develop a robust, well-connected set of tables in order to answer important questions about driver and team performance in each season of Formula 1; these are detailed further in Section 4 of this document.

## 2. Data

The dataset selected is from an online web service called Ergast and contains information about various entities such as drivers, circuits, constructors and races from the start of F1 in 1950 to the 2024 season. The full dataset contains 14 tables, but not all of them are especially relevant for our use case, so only 7 entities have been chosen for the initial database implementation.

Applicable Business rules (defined by Federation Internationale de l'Automobile):

- One driver may only be associated with one constructor team in a single season
- One constructor team may have only two drivers in a single season that race for them
- Only 20 drivers may be allowed in a single race in a single venue
- A driver can use up to 20 sets of tires over a season (13 dry, 4 intermediate and 3 wet)
- A maximum of 10 teams may compete in a single World Championship season
- Points are awarded by finishing position (eg. 25 pts for 1st place, 18 for 2nd, 15 for 3rd and so on; a constructor team receives the aggregate points between both their drivers
- A circuit may host zero or many races, but a race can only occur at one circuit at a time

- A race may have an unlimited number of pit stops, but it may have none as well.

Source files and documentation:

- Full dataset: https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020?select=status.csv
- Ergast API (deprecated): http://ergast.com/mrd/
- Link to full set of tables as CSVs: http://ergast.com/downloads/f1db_csv.zip

Data Dictionary:

| Column Name | Data Type | Description |
|---|---|---|
| driverId | Numeric | Unique ID for a driver |
| driverNumber | Numeric | Driver Car Number |
| code | Text | 3-letter alphabet code of driver |
| forename | Text | First name of Driver |
| surname | Text | Last name of Driver |
| dob | Date | Driver's Date of Birth |
| nationality | Text | Driver's Country of Nationality |
| url | Text | Wikipedia URL |
| raceID | Numeric | Id of the race |
| name | Text | Name of the race, e.g. British Grand Prix |
| year | Numeric | Year in which the race was held |
| date | Date | Date on which the race was held |
| contructorID | Numeric | Unique ID of Constructor |
| name | Text | Actual Name of Constructor (Manufacturer) |
| nationality | Text | Country |
| circuitID | Numeric | Unique ID of Circuit |
| name | Text | Actual name of circuit |
| location | Text | Circuit City |

| | | |
|---|---|---|
| country | Text | Circuit Country |
| stop | Numeric | No. of pit stops |
| lap | Numeric | Lap Number |
| duration | Text | Duration of pitstop |
| positionOrder | Numeric | Final Rank |
| points | Numeric | Points |
| fastestLap | Text | Lap number of fastest lap |
| fastestLapTime | Text | Lap time of fastest lap |
| fastestLapSpeed | Text | Top speed of fastest lap |
| statusID | Numeric | ID |
| status | Text | Race Status |

Table 1: Data dictionary

ER Diagram:

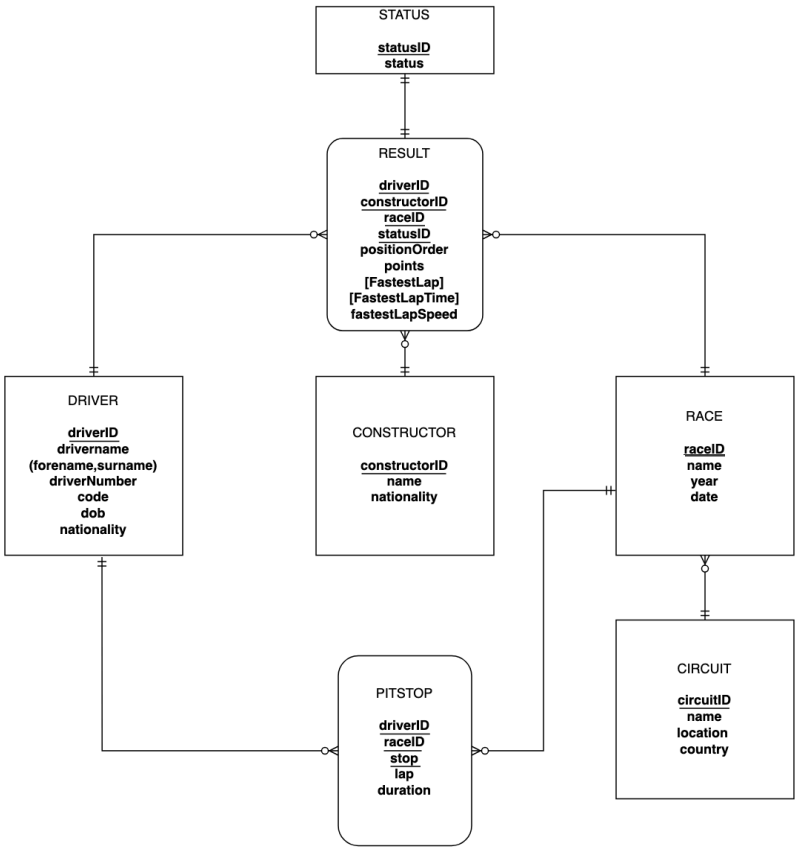



Figure 1: ER Diagram

Relational Schema:

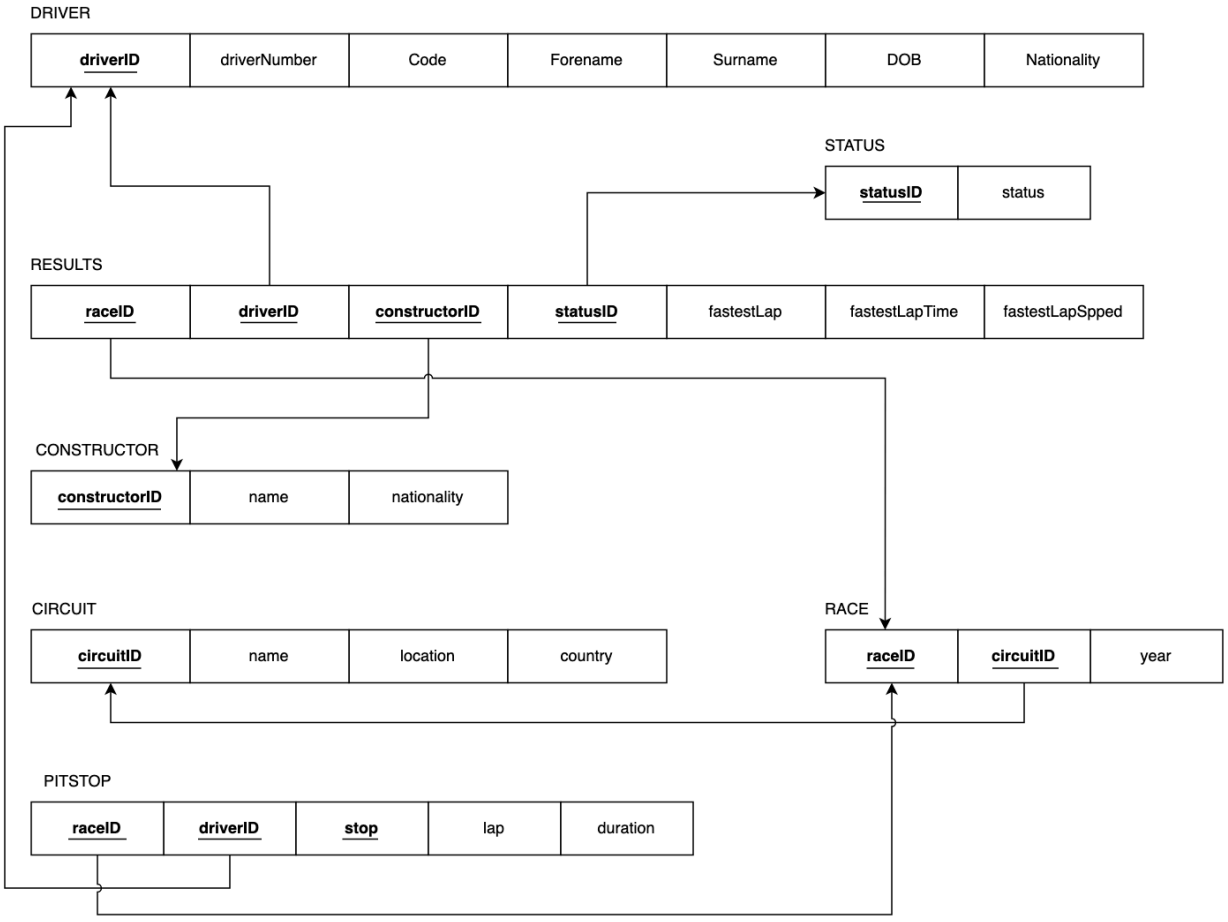



Figure 2: R-Schema Diagram

## 3. Database Implementation

Data Cleaning:

- Several tables in the source data had data quality issues, such as invalid NULL values for the driverNumber field in the DRIVER table. In order to comply with the RSchema, data was modified by making it blank to work with the data upload tool.
- Some field names were unsuitable for use as a column name in a database table, because they were reserved keywords in Oracle SQL (eg. number). These columns were renamed where appropriate (eg. number → driverNumber).

The primary tables in the implemented database are the DRIVER, CONSTRUCTOR, CIRCUIT, RACE tables. The PK for these tables are called *driverId, constructorId, circuitId* and *raceId* respectively. The associative RESULT entity relates DRIVER, CONSTRUCTOR and CIRCUIT instances with a race status; the PITSTOP and STATUS tables also contain race and result-related information that may be used for analytics and research. No weak entities exist as each either has its own primary key or has a composite primary key to facilitate associative relationships.

DDL and INSERT statements:

| Entity | DDL and Sample INSERT statement |
|---|---|
| CONSTRUCTOR | Create table CONSTRUCTOR (<br>constructorId NUMBER(3,0) NOT NULL,<br>name VARCHAR2(50) NOT NULL,<br>nationality VARCHAR2(50) NOT NULL,<br>CONSTRAINT CONSTRUCTOR_PK PRIMARY KEY (constructorID)<br>);<br><br>*INSERT INTO CONSTRUCTOR VALUES(1,'McLaren','British');* |
| CIRCUIT | Create table CIRCUIT (<br>circuitId NUMBER(2,0) NOT NULL,<br>name VARCHAR2(100) NOT NULL,<br>location VARCHAR2(50) NOT NULL,<br>country VARCHAR2(50) NOT NULL,<br>CONSTRAINT CIRCUIT_PK PRIMARY KEY (circuitID)<br>);<br><br>*INSERT INTO CIRCUIT VALUES (1,'Albert Park Grand Prix Circuit','Melbourne','Australia');* |
| DRIVER | CREATE TABLE DRIVER (<br>driverId NUMBER(3,0) NOT NULL,<br>driverNumber NUMBER(3,0),<br>code CHAR(3),<br>forename VARCHAR2(25),<br>surname VARCHAR2(25),<br>dob DATE,<br>nationality VARCHAR2(25),<br>CONSTRAINT driver_pk PRIMARY KEY (driverID)<br>);<br><br>*INSERT INTO DRIVER VALUES (1,44,'HAM','Lewis','Hamilton',TO_DATE('01/07/1985'), 'British');* |
| RACE | CREATE TABLE RACE (<br>raceId NUMBER(4,0) NOT NULL,<br>year NUMBER(4,0) NOT NULL,<br>circuitId NUMBER(2,0) NOT NULL,<br>name VARCHAR2(30) NOT NULL,<br>rdate DATE NOT NULL,<br>CONSTRAINT RACE_PK PRIMARY KEY (raceId),<br>CONSTRAINT RACE_FK FOREIGN KEY (circuitId) REFERENCES CIRCUIT(circuitId)<br>);<br><br>*INSERT INTO RACE VALUES (1,2009,1,'Australian Grand Prix', TO_DATE('03/29/2009', 'MM-DD-YYYY'));* |
| PITSTOP | CREATE TABLE PITSTOP(<br>raceId NUMBER(4,0) NOT NULL,<br>driverId NUMBER(3,0) NOT NULL,<br>stop NUMBER(2,0) NOT NULL,<br>lap NUMBER(2,0) NOT NULL,<br>duration NUMBER(10,3) NOT NULL,<br>CONSTRAINT PITSTOP_PK PRIMARY KEY (driverId, raceId, stop),<br>CONSTRAINT PITSTOP_FK1 FOREIGN KEY (driverId) REFERENCES DRIVER (driverId),<br>CONSTRAINT PITSTOP_FK2 FOREIGN KEY (raceId) REFERENCES RACE (raceId)<br>);<br><br>*INSERT INTO PITSTOP VALUES('841','153', '1','1','26.898');* |
| STATUS | CREATE TABLE STATUS(<br>statusId NUMBER(3,0) NOT NULL, |

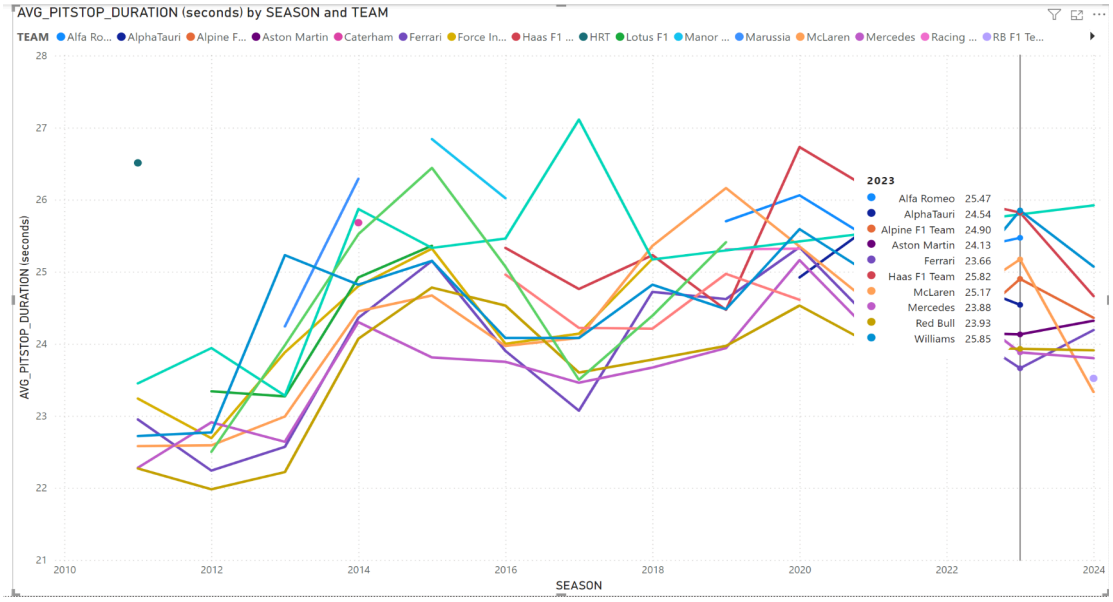
| | |
|---|---|
| | status VARCHAR2(20) NOT NULL,<br>CONSTRAINT STATUS_PK PRIMARY KEY (statusId)<br>);<br><br>*INSERT INTO STATUS VALUES('1','Finished');* |
| RESULT | CREATE TABLE RESULT (<br>raceID NUMBER(4,0) NOT NULL,<br>driverID NUMBER(3,0) NOT NULL,<br>constructorId NUMBER(3,0) NOT NULL,<br>statusID NUMBER(3,0) NOT NULL,<br>fastestLap NUMBER(5,0),<br>fastestLapTime VARCHAR2(50),<br>fastestLapSpeed NUMBER(6,3),<br>points NUMBER(3,0),<br>CONSTRAINT RESULT_FK1 FOREIGN KEY (driverId) REFERENCES DRIVER (driverId),<br>CONSTRAINT RESULT_FK2 FOREIGN KEY (raceId) REFERENCES RACE (raceId),<br>CONSTRAINT RESULT_FK3 FOREIGN KEY (constructorId) REFERENCES CONSTRUCTOR (constructorId),<br>CONSTRAINT RESULT_FK4 FOREIGN KEY (statusId) REFERENCES STATUS (statusId)<br>);<br><br>*INSERT INTO RESULT VALUES(18,1,1,1,39,'1:29.36',218.3);* |

The full set of source files tables may be found here:
https://drive.google.com/drive/folders/1FGrL8EDqq7tqdZiBUcB98Vg4Uf4EzRy3?usp=sharing
\

## 4. Analysis

1. Average pit stop durations by team and season, to assess which teams have the fastest and most efficient pit crews per season.

```
SELECT
   CONSTRUCTOR.name AS TEAM,
   RACE.year AS Season,
   ROUND(AVG(PITSTOP.duration),2) AS Avg_PitStop_Duration
FROM
   PITSTOP
   JOIN RESULTS ON PITSTOP.driverID = RESULTS.driverID AND PITSTOP.raceID = RESULTS.raceID
   JOIN CONSTRUCTOR ON RESULTS.constructorID = CONSTRUCTOR.constructorID
   JOIN  RACE ON PITSTOP.raceID = RACE.raceID
GROUP BY
   CONSTRUCTOR.name, RACE.year
ORDER BY
   RACE.year, AvgPitStopDuration ASC;
```




| TEAM | SEASON | AVG_PITSTOP_DURATION |
|---|---|---|
| Red Bull | 2011 | 22.27 |
| Mercedes | 2011 | 22.28 |
| McLaren | 2011 | 22.58 |
| Williams | 2011 | 22.72 |
| Ferrari | 2011 | 22.95 |
| Force India | 2011 | 23.24 |
| Sauber | 2011 | 23.45 |

The data shows that Red Bull and Mercedes were the most successful teams in terms of having the fastest pit stop timings, which was reflected in the number of race championships they won over the years. On the other hand, Williams needs to improve their pit stop timings in order to compete with and win against other top teams.

2. Seasonal and circuit wise pit-stop frequencies for each driver, to determine if drivers get better at pit stops over time on the same circuit:

```
SELECT
   year as season,
   C.name,
   P.driverID,
   D.forename || ' ' || D.surname as driverName,
   count(*)
FROM
   RACE
   LEFT JOIN CIRCUIT C ON (C.circuitId = RACE.circuitId)
   LEFT JOIN PITSTOP P ON (P.raceId = RACE.raceID)
   LEFT JOIN DRIVER D ON (P.driverid = D.driverid)
WHERE C.name is not null
GROUP BY year, C.name, P.driverID, D.forename || ' ' || D.surname
```

| SEASON | NAME | DRIVERID | DRIVERNAME | COUNT(*) |
|---|---|---|---|---|
| 2011 | Circuit de Barcelona-Catalunya | 39 | Narain Karthikeyan | 3 |
| 2011 | Circuit de Monaco | 10 | Timo Glock | 1 |
| 2011 | Circuit de Monaco | 15 | Jarno Trulli | 2 |
| 2011 | Circuit de Monaco | 155 | Kamui Kobayashi | 1 |
| 2011 | Circuit Gilles Villeneuve | 18 | Jenson Button | 6 |
| 2011 | Circuit Gilles Villeneuve | 4 | Fernando Alonso | 3 |
| 2011 | Circuit Gilles Villeneuve | 816 | JÃ©rÃ´me d'Ambrosio | 5 |
| 2011 | Circuit Gilles Villeneuve | 15 | Jarno Trulli | 3 |
| 2011 | Valencia Street Circuit | 1 | Lewis Hamilton | 3 |
| 2011 | Valencia Street Circuit | 30 | Michael Schumacher | 3 |

This result is important to see driver trends over time; however, it is unclear from the available data whether there is a statistically significant trend available. Some drivers show a cyclic trend with pitstop frequency first increasing and then decreasing YoY, but for others the values hardly change which may indicate they are more seasoned or consistent.

3. Statistical measures of driver wins for a given season and racetrack using Box-and-Whisker plots:

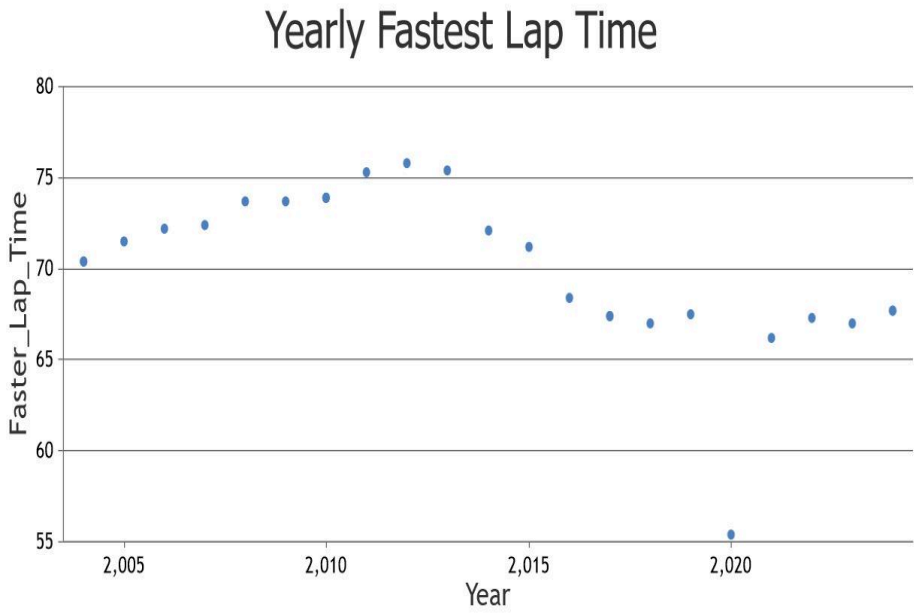
```
SELECT
   D.driverid,
   D.forename || ' ' || D.surname as driverName,
   year as season,
   min(points),
   PERCENTILE_CONT(0.25) WITHIN GROUP (ORDER BY points) as pct25,
   MEDIAN(points),
   PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY points) as pct75,
   max(points)
FROM
   RESULT RSLT JOIN RACE ON RSLT.raceId = RACE.raceId
   LEFT JOIN DRIVER D ON (RSLT.driverid = D.driverid)
GROUP BY D.driverid, D.forename || ' ' || D.surname, year
```

| DRIVERID | DRIVERNAME | SEASON | MIN(POINTS) | PCT25 | MEDIAN(POINTS) | PCT75 | MAX(POINTS) |
|---|---|---|---|---|---|---|---|
| 1 | Lewis Hamilton | 2007 | 0 | 5 | 8 | 8 | 10 |
| 1 | Lewis Hamilton | 2008 | 0 | 2.5 | 6 | 9.5 | 10 |
| 1 | Lewis Hamilton | 2009 | 0 | 0 | .5 | 6 | 10 |
| 1 | Lewis Hamilton | 2010 | 0 | 8.5 | 13.5 | 18 | 25 |
| 1 | Lewis Hamilton | 2011 | 0 | 7 | 12 | 18 | 25 |
| 1 | Lewis Hamilton | 2012 | 0 | .5 | 10 | 15 | 25 |

From the results above, we can see a large variability in driver performance between even consecutive seasons; for example, Lewis Hamilton's median score went up by a staggering 13 points from the 2009 - 2010 season, which can largely be attributed to the performance increases in the car he was driving.

4. Which drivers had the fastest laps across a single season (year)?

```
SELECT
   DRIVER.forename || ' ' || DRIVER.surname AS DRIVERNAME,
   DRIVER.nationality AS NATIONALITY,
   RACE.year AS SEASON,
   RSLT.fastestLapTime AS FASTEST_LAP_TIME
FROM RESULT RSLT
JOIN RACE ON RSLT.raceId = RACE.raceId
JOIN DRIVER ON RSLT.driverId = DRIVER.driverId
WHERE (RACE.year, RSLT.fastestLapTime) IN (
   SELECT
   RACE.year,
   MIN(RSLT.fastestLapTime)
   FROM RESULT RSLT
   JOIN RACE ON RSLT.raceId = RACE.raceId
  WHERE RSLT.fastestLapTime IS NOT NULL
  GROUP BY RACE.year
)
ORDER BY RACE.year;
```




**Graph for yearly fastest laps over years**

| DRIVERNAME | NATIONALITY | SEASON | FASTEST_LAP_TIME |
|---|---|---|---|
| Rubens Barrichello | Brazilian | 2004 | 01:10.4 |
| Michael Schumacher | German | 2004 | 01:10.4 |
| Michael Schumacher | German | 2005 | 01:11.5 |
| Michael Schumacher | German | 2006 | 01:12.2 |
| Kimi Räikkönen | Finnish | 2007 | 01:12.4 |
| Felipe Massa | Brazilian | 2008 | 01:13.7 |
| Mark Webber | Australian | 2009 | 01:13.7 |
| Lewis Hamilton | British | 2010 | 01:13.9 |
| Fernando Alonso | Spanish | 2010 | 01:13.9 |
| Jenson Button | British | 2010 | 01:13.9 |
| Mark Webber | Australian | 2011 | 01:15.3 |

The results here speak no different than the history, early 2000s were indeed the Years of the "GOAT" Michael Schumacher whose name appears three times in a row for fastest lap of the year….We really wish for your speedy recovery Champ !!

5. Point distribution across all constructors in their entire racing careers.

```
SELECT
   CONSTRUCTOR.constructorID,
   CONSTRUCTOR.name,
   SUM(RESULT.points) AS TotalPoints
FROM
   RESULT
   JOIN CONSTRUCTOR ON (RESULT.constructorId = CONSTRUCTOR.constructorId)
GROUP BY CONSTRUCTOR.constructorID, CONSTRUCTOR.name
ORDER BY SUM(RESULT.points) DESC;
```

| CONSTRUCTORID | NAME | TOTALPOINTS |
|---|---|---|
| 9 | Red Bull | 7477 |
| 131 | Mercedes | 7384 |
| 6 | Ferrari | 6785 |
| 1 | McLaren | 3974 |
| 4 | Renault | 1354 |
| 3 | Williams | 1269 |
| 10 | Force India | 1098 |
| 208 | Lotus F1 | 706 |
| 5 | Toro Rosso | 500 |
| 117 | Aston Martin | 466 |
| 214 | Alpine F1 Team | 442 |

From the above data, it can be concluded that Red Bull and Mercedes were the most consistent of teams over the years, which is indicated by the accumulated points. The likes of Alpine and Aston Martin have some serious catching up to do!

**References:**


1. "Insight: The Trackside Engineers - Mercedes-AMG Petronas F1 Team." *Mercedes*, 6 Nov. 2018, www.mercedesamgf1.com/news/insight-the-trackside-engineers.