F1 Data Analysis Project

The Formula 1 Data Analysis project is a comprehensive exploration of the world of Formula 1 racing through the lens of data analytics. This project combines the power of SQL for data manipulation in addition to Deepnote and Tableau for data visualisation to provide insights into the thrilling world of Formula 1 racing from 1953-2020. This project aims to identify candidates for the greatest Formula 1 driver of all time.

Key Components:

Data Collection and Preparation: The project begins by gathering extensive Formula 1 data, including race results, driver statistics, team information, and circuit details. This raw data is then cleaned and organised for analysis.

Database Design: A relational database is created using SQL to efficiently store and manage the Formula 1 data. This step includes defining tables, relationships, and constraints to ensure data integrity.

SQL Data Queries: SQL queries are crafted to extract meaningful insights from the database. This may involve calculating driver and team performance metrics, historical trend analysis, and identifying key patterns in the data.

Tableau Visualization: The SQL-derived insights are visualised using Tableau, a powerful data visualisation tool. Interactive dashboards and reports are designed to provide users with a user-friendly interface to explore the data.

Performance Analysis: Using SQL and Tableau, the project delves into various aspects of Formula 1 racing, such as driver performance across seasons, race track statistics, and the impact of rule changes.

Reporting and Presentation: The project concludes by creating comprehensive reports and presentations that summarise the findings, making it accessible to a wider audience.

Benefits:

Gain a deeper understanding of Formula 1 racing from a data-driven perspective.

Enhance the overall Formula 1 fan experience by making complex data accessible and engaging.

This project offers a unique opportunity to combine the technical skills of SQL with the data visualisation capabilities of Deepnote and Tableau to unlock insights about one of the dynamic sports in the world.

Project Links:

Tableau Dashboard

Github Repository

EDA

We will begin by performing exploratory data analysis to understand the data structure, identify errors, outliers and anomalies, plus uncover patterns and relationships.

General Queries

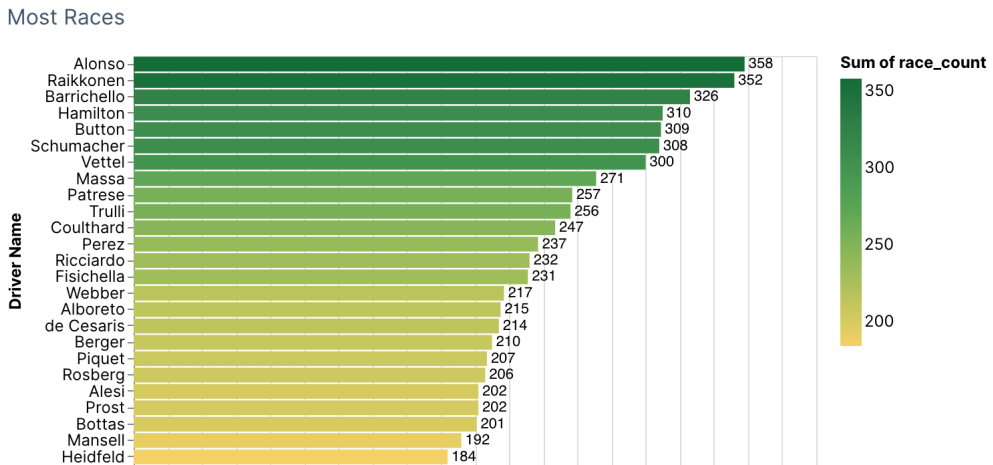
Firstly, let's identify the most experienced driver:

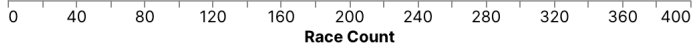
```sql
-- Find the driver who has competed in the most races

SELECT
    r.driverId,
    COUNT(r.RaceId) AS race_count,
    d.surname,
    d.forename
FROM 'results.csv' r
JOIN 'drivers.csv' d ON r.driverId = d.driverId
GROUP BY r.driverId, d.surname, d.forename
ORDER BY race_count DESC;
```

	driverID int64	race_count int64	surname object	forename object
	1 - 856	1 - 358	Taylor 0.6% Fittipaldi 0.5% 794 others 98.9%	John 1.6% Mike 1.6% 469 others 96.7%
0	4	358	Alonso	Fernando
1	8	352	Raikkonen	Kimi
2	22	326	Barrichello	Rubens
3	1	310	Hamilton	Lewis
4	18	309	Button	Jenson
5	30	308	Schumacher	Michael
6	20	300	Vettel	Sebastian
7	13	271	Massa	Felipe
8	119	257	Patrese	Riccardo
9	15	256	Trulli	Jarno

855 rows, showing 10 per page Page 1 of 86

From this query, we can see that the driver who has participated in the most races is **Fernando Alonso (358)**, followed by **Kimi Raikonnen (352).** Let's visualise the top 25 drivers with the most races:





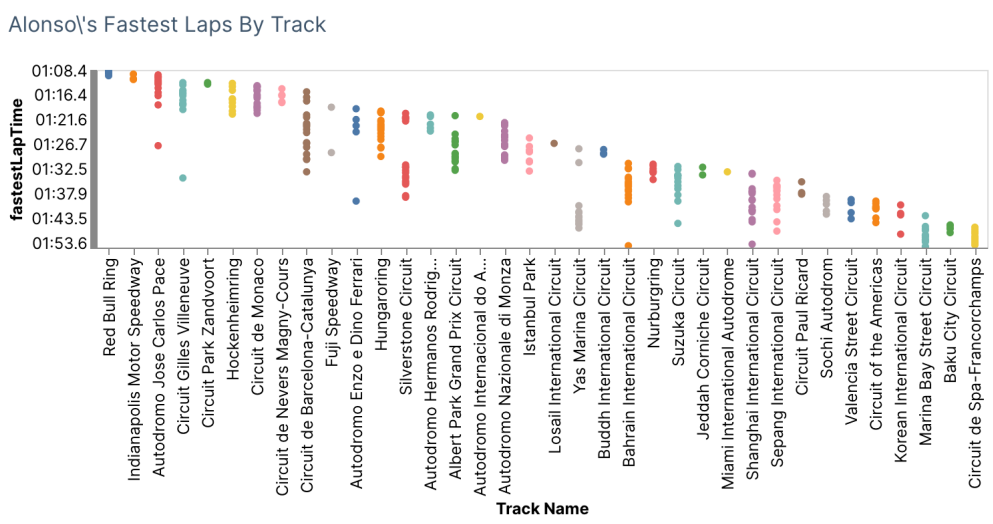
Now let's take a closer look at the stellar career of Fernando Alonso.

```
-- Investigating Alonso's career

SELECT
    r.RaceID,
    r.driverID,
    r.fastestLapTime,
    r.fastestLapSpeed,
    ra.circuitId,
    ra.date,
    c.name
FROM 'results.csv' r
JOIN 'races.csv' ra ON r.RaceId = ra.RaceID
LEFT JOIN 'circuits.csv' c ON ra.circuitId = c.circuitId
WHERE r.driverId = '4' AND r.fastestLapSpeed <> 0
ORDER BY r.fastestLapTime ASC;
```

	RaceID int64 1 - 1096	driverID int64 4 - 4	fastestLapTime ob... 01:22.8 1.3% 01:14.1 1% 229 others 97.8%	fastestLapSpeed fl... 136.503 - 253.874	circuitId int64 1 - 79	date object 2021-07-04 0.3 2022-07-10 0.3 310 others 99.4
0	1060	4	01:08.4	227.246	70	2021-07-04
1	1084	4	01:08.5	226.739	70	2022-07-10
2	997	4	01:08.6	226.399	70	2018-07-01
3	1058	4	01:09.7	222.867	70	2021-06-27
4	956	4	01:11.0	219.284	70	2016-07-03
5	98	4	01:11.2	211.847	19	2004-06-20
6	107	4	01:12.1	215.097	18	2004-10-24
7	907	4	01:12.5	214.527	70	2014-06-22
8	1008	4	01:12.6	213.542	18	2018-11-11
9	87	4	01:12.6	213.513	18	2005-09-25

312 rows, showing 10 per page Page 1 of 32

Visualising Alonso's fastest laps at each circuit:



Now we're going to explore the statistics for all drivers, ordered by fastest ever lap times across all circuits.

```
-- See driver stats ordered by Fastest Lap Time

SELECT
    r.RaceID,
    r.driverID,
    d.forename,
    d.surname,
    r.fastestLapTime,
    r.fastestLapSpeed,
    ra.circuitId,
    ra.date,
    c.name
FROM 'results.csv' r
JOIN 'races.csv' ra ON r.RaceId = ra.RaceID
JOIN 'drivers.csv' d ON r.driverId = d.driverId
LEFT JOIN 'circuits.csv' c ON ra.circuitId = c.circuitId
WHERE r.fastestLapSpeed <> 0
ORDER BY r.fastestLapTime ASC;
```

	RaceID int64 1 - 1096	driverID int64 1 - 856	forename object Nico 5% Fernando 4.2% 89 others 90.8%	surname object Alonso 4.2% Hamilton 4.1% 97 others 91.7%	fastestLapTime ob... 01:21.6 0.4% 01:17.8 0.4% 586 others 99.1%	fastestLapSpeed fl 89.54 - 257.32
0	1046	847	George	Russell	00:55.4	230.21
1	1046	822	Valtteri	Bottas	00:56.5	225.49
2	1046	815	Sergio	Perez	00:56.7	224.59
3	1046	841	Antonio	Giovinazzi	00:56.8	224.21
4	1046	20	Sebastian	Vettel	00:56.9	224.14
5	1046	817	Daniel	Ricciardo	00:56.9	223.8
6	1046	826	Daniil	Kvyat	00:57.0	223.76
7	1046	848	Alexander	Albon	00:57.0	223.54
8	1046	832	Carlos	Sainz	00:57.1	223.12
9	1046	846	Lando	Norris	00:57.2	222.71

7379 rows, showing 10 per page Page 1 of 738

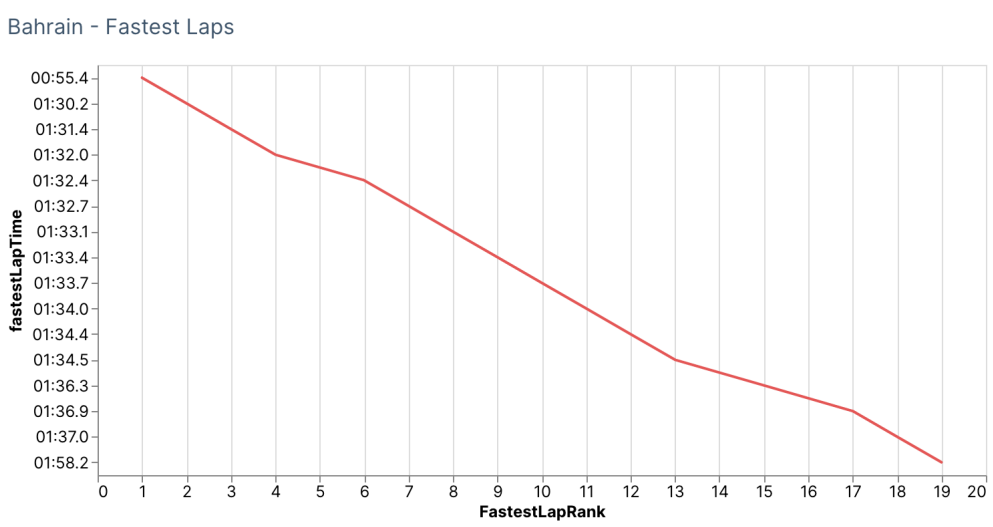
The previous query shows some unusually fast lap times for RaceID 1046 (under 60 seconds), which was at Bahrain 2020. Let's investigate further by looking at the fastest laps for each race at Bahrain:

```
-- What are the fastest lap times for Bahrain?

WITH BahrainRaces AS (
    SELECT
        r.RaceID,
        r.fastestLapTime,
        ra.circuitId,
        ra.date,
        c.name,
        ROW_NUMBER() OVER (PARTITION BY r.RaceID ORDER BY r.fastestLapTime ASC) as LapRank
    FROM 'results.csv' r
    JOIN 'races.csv' ra ON r.RaceId = ra.RaceID
```

```
    LEFT JOIN 'circuits.csv' c ON ra.circuitId = c.circuitId
    WHERE r.fastestLapSpeed <> 0 AND c.name = 'Bahrain International Circuit'
),
FastestLaps AS (
    SELECT
        RaceID,
        fastestLapTime,
        circuitId,
        date,
        name
    FROM BahrainRaces
    WHERE LapRank = 1
)
SELECT
    RaceID,
    fastestLapTime,
    circuitId,
    date,
    name,
    RANK() OVER (ORDER BY fastestLapTime ASC) as FastestLapRank
FROM FastestLaps
ORDER BY FastestLapRank;
```

	RaceID int64 4 - 1074	fastestLapTime ob... 01:32.0 10.5% 01:34.5 10.5% 14 others 78.9%	circuitId int64 3 - 3	date object 2020-12-06 5.3% 2004-04-04 5.3% 17 others 89.5%	name object Bahrain Inte... 100%	FastestLapRank in. 1 - 19
0	1046	00:55.4	3	2020-12-06	Bahrain Internatio...	
1	92	01:30.2	3	2004-04-04	Bahrain Internatio...	
2	73	01:31.4	3	2005-04-03	Bahrain Internatio...	
3	1045	01:32.0	3	2020-11-29	Bahrain Internatio...	
4	1052	01:32.0	3	2021-03-28	Bahrain Internatio...	
5	53	01:32.4	3	2006-03-12	Bahrain Internatio...	
6	971	01:32.7	3	2017-04-16	Bahrain Internatio...	
7	20	01:33.1	3	2008-04-06	Bahrain Internatio...	
8	1011	01:33.4	3	2019-03-31	Bahrain Internatio...	
9	990	01:33.7	3	2018-04-08	Bahrain Internatio...	1

19 rows, showing 10 per page Page 1 of 2



The results show that discounting RaceID 1046, the next fastest times are over 30 seconds slower, meaning we likely have an anomaly. Further investigation shows this race was a one time, reformatted race due to COVID. Therefore, it shall be discounted from this query to avoid skewing the results. Let's run the query without data from RaceID 1046:

```
-- View driver stats ordered by Fastest Lap Time excluding RaceID 1046

SELECT
    r.RaceID,
    r.driverID,
    d.forename,
    d.surname,
    r.fastestLapTime,
    r.fastestLapSpeed,
    ra.circuitId,
    ra.date,
    c.name
FROM 'results.csv' r
JOIN 'races.csv' ra ON r.RaceId = ra.RaceID
JOIN 'drivers.csv' d ON r.driverId = d.driverId
LEFT JOIN 'circuits.csv' c ON ra.circuitId = c.circuitId
WHERE r.fastestLapSpeed <> 0 AND r.RaceID != '1046'
ORDER BY r.fastestLapTime ASC;
```

	RaceID int64 1 - 1096	driverID int64 1 - 856	forename object Nico 5% Fernando 4.2% 88 others 90.8%	surname object Alonso 4.2% Hamilton 4.1% 96 others 91.6%	fastestLapTime ob... 01:21.6 0.4% 01:17.8 0.4% 574 others 99.1%	fastestLapSpeed fl 89.54 - 257.32
0	1032	832	Carlos	Sainz	01:05.6	236.89
1	1032	830	Max	Verstappen	01:06.1	235.0
2	1060	830	Max	Verstappen	01:06.2	234.81
3	1032	1	Lewis	Hamilton	01:06.7	232.98
4	997	8	Kimi	Raikkonen	01:06.9	232.1
5	1058	1	Lewis	Hamilton	01:07.0	231.8
6	997	20	Sebastian	Vettel	01:07.0	231.72
7	1032	815	Sergio	Perez	01:07.1	231.36
8	1032	846	Lando	Norris	01:07.1	231.34
9	1084	830	Max	Verstappen	01:07.2	231.06

7361 rows, showing 10 per page Page 1 of 737

The above query shows that the fastest ever racing lap was recorded by **Carlos Sainz at Red Bull Ring (2020-07-12)**, with a time of **01:05.6**. Next, we're going to look at the fastest laps ever recorded at each track:

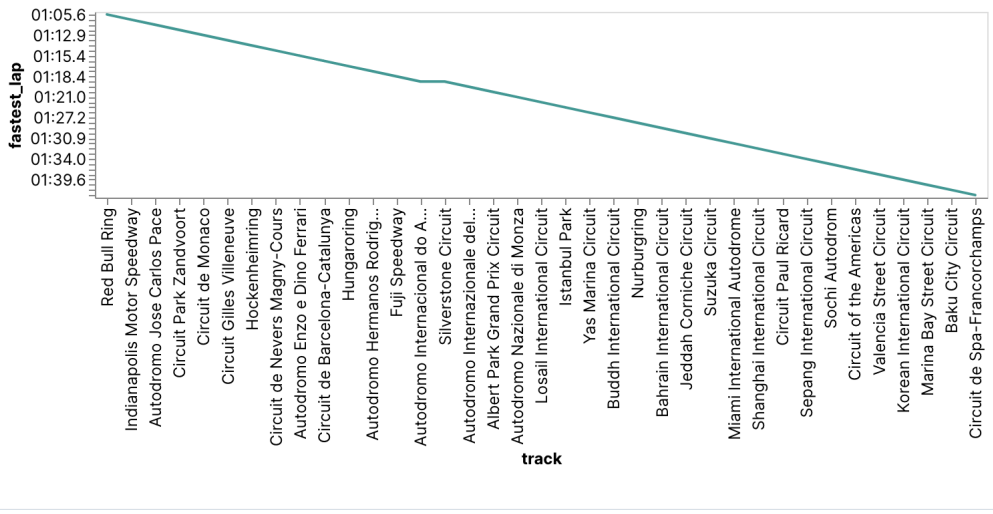
```
-- Fastest Lap by Track

SELECT
    c.name AS track,
    MIN(r.fastestLapTime) AS fastest_lap
FROM 'results.csv' r
JOIN 'races.csv' ra ON r.RaceId = ra.RaceID
JOIN 'drivers.csv' d ON r.driverId = d.driverId
LEFT JOIN 'circuits.csv' c ON ra.circuitId = c.circuitId
WHERE r.fastestLapTime <> '0' AND c.name IS NOT NULL AND r.RaceID != '1046'
```

```
GROUP BY c.name
ORDER BY fastest_lap ASC;
```

	track object	fastest_lap object
	Red Bull Ring ..... 2.7% Indianapolis ... . 2.7% 35 others 94.6%	01:18.7 ..... 5.4% 01:05.6 ..... 2.7% 34 others 91.9%
0	Red Bull Ring	01:05.6
1	Indianapolis Moto...	01:10.3
2	Autodromo Jose ...	01:10.5
3	Circuit Park Zand...	01:11.0
4	Circuit de Monaco	01:12.9
5	Circuit Gilles Ville...	01:13.0
6	Hockenheimring	01:13.7
7	Circuit de Nevers...	01:15.3
8	Autodromo Enzo ...	01:15.4
9	Circuit de Barcelo...	01:15.6

37 rows, showing 10 per page Page 1 of 4

Let's visualise these findings in a line chart, with lap times in ascending order:



As previously discovered, the fastest ever lap was recorded at Red Bull Ring. The track with the slowest time is the iconic Circuit de Spa-Francorchamps in Belgium.

Driver Performance Analysis

Now we're going to dig deeper into driver performance analysis. Let's begin by finding the driver with the most career points:

```
-- Find career points leader

SELECT
    r.driverId,
    SUM(r.points) AS total_points,
    d.forename,
    d.surname
FROM 'results.csv' r
JOIN 'drivers.csv' d ON r.driverId = d.driverId
GROUP BY r.driverId, d.forename, d.surname
ORDER BY total_points DESC;
```

	driverID int64 1 - 856	total_points float64 0.0 - 4396.5	forename object John ..... 1.6% Mike ..... 1.6% 469 others 96.7%	surname object Taylor ..... 0.6% Fittipaldi ..... 0.5% 794 others 98.9%
0	1	4396.5	Lewis	Hamilton
1	20	3098	Sebastian	Vettel
2	4	2061	Fernando	Alonso
3	830	1983.5	Max	Verstappen
4	8	1873	Kimi	Raikkonen
5	822	1778	Valtteri	Bottas
6	3	1594.5	Nico	Rosberg
7	30	1566	Michael	Schumacher
8	817	1307	Daniel	Ricciardo
9	18	1235	Jenson	Button

855 rows, showing 10 per page Page 1 of 86

The driver with the most career points is **Lewis Hamilton,** with **4396,** followed by **Sebastian Vettel** with **3098.**

Next, identifying the driver with the highest average points per race:

```
-- Find highest avg points per race

SELECT
    r.driverId,
    AVG(r.points) AS average_points,
    COUNT(r.RaceId) AS total_races,
    d.surname,
    d.forename
FROM 'results.csv' r
JOIN 'drivers.csv' d ON r.driverId = d.driverId
GROUP BY r.driverId, d.surname, d.forename
ORDER BY average_points DESC;
```

	driverID int64 1 - 856	average_points flo... 0.0 - 14.182258064...	total_races int64 1 - 358	surname object Taylor ..... 0.6% Fittipaldi ..... 0.5% 794 others 98.9%	forename object Mike ..... 1.6% John ..... 1.6% 469 others 96.7%
0	1	14.18225806	310	Hamilton	Lewis
1	830	12.16871166	163	Verstappen	Max
2	20	10.32666667	300	Vettel	Sebastian
3	822	8.845771144	201	Bottas	Valtteri
4	844	8.262135922	103	Leclerc	Charles
5	3	7.740291262	206	Rosberg	Nico
6	591	6	1	Amick	George
7	4	5.75698324	358	Alonso	Fernando
8	817	5.63362069	232	Ricciardo	Daniel
9	8	5.321022727	352	Raikkonen	Kimi

855 rows, showing 10 per page Page 1 of 86

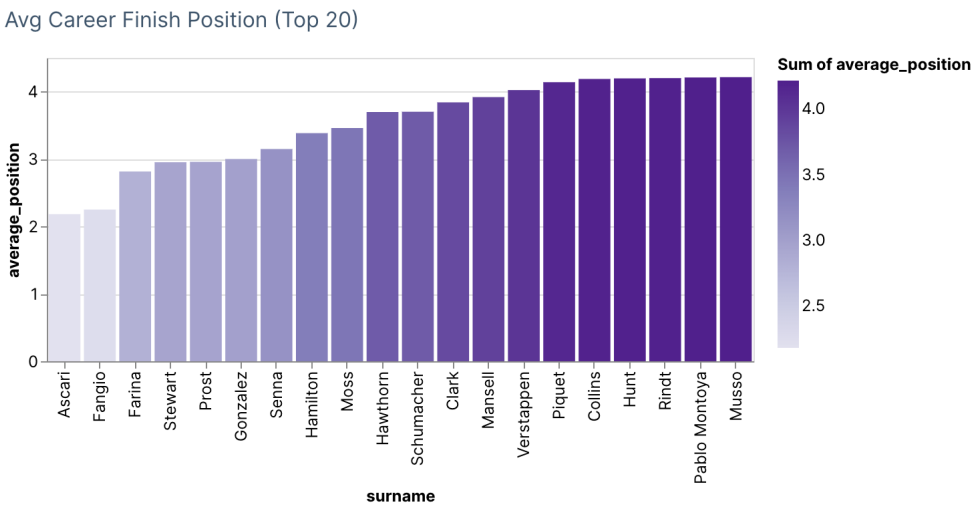
Lewis Hamilton has the highest average points per race with **14.18,** which comes as no surprise given his unprecedented success at the pinnacle of motor racing. He is followed by the Red Bull prodigy himself, **Max Verstappen**, with an average of **12.16.**

Which driver has the highest average career finish? We will set a minimum threshold of 20 races to discount any outliers:

```sql
-- Find the driver with the highest average career finish (min 20 races)

SELECT
    r.driverId,
    AVG(NULLIF(CAST(r.position AS INT), 0)) AS average_position,
    COUNT(r.RaceId) AS total_races,
    d.surname,
    d.forename
FROM 'results.csv' r
JOIN 'drivers.csv' d ON r.driverId = d.driverId
GROUP BY r.driverId, d.surname, d.forename
HAVING COUNT(r.RaceId) > 20
ORDER BY average_position ASC;
```

	driverID int64 1 - 855	average_position fl... 2.181818181818181...	total_races int64 21 - 358	surname object Schumacher 1.2% Hill 1.2% 239 others 97.6%	forename object Mike 1.6% Pedro 1.6% 198 others 96.9%
0	647	2.181818182	36	Ascari	Alberto
1	579	2.25	58	Fangio	Juan
2	642	2.814814815	37	Farina	Nino
3	328	2.952380952	100	Stewart	Jackie
4	117	2.958041958	202	Prost	Alain
5	498	3	29	Gonzalez	Jose Froilan
6	102	3.148148148	162	Senna	Ayrton
7	1	3.383802817	310	Hamilton	Lewis
8	475	3.459459459	73	Moss	Stirling
9	578	3.696969697	48	Hawthorn	Mike

255 rows, showing 10 per page Page 1 of 26



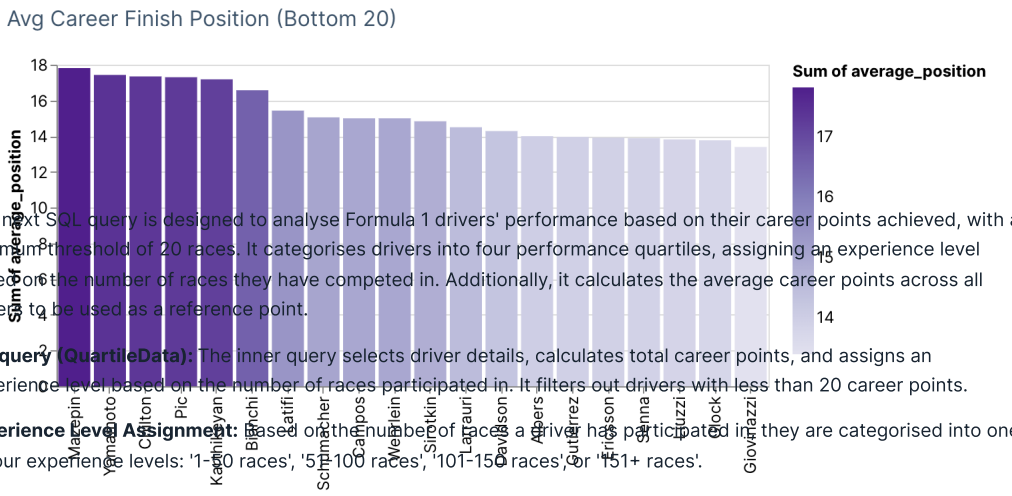
Here we see that the driver with the highest average finishing position with a minimum of 20 career races is **Alberto Ascari**, with an incredible average of **2.18**, followed by **Juan Fangio at 2.25.** Next, the drivers with the lowest average career finish:

```sql
-- Find the driver with the lowest average career finish (min 20 races)

SELECT
    r.driverId,
    AVG(NULLIF(CAST(r.position AS INT), 0)) AS average_position,
    COUNT(r.RaceId) AS total_races,
    d.surname,
    d.forename
FROM 'results.csv' r
JOIN 'drivers.csv' d ON r.driverId = d.driverId
GROUP BY r.driverId, d.surname, d.forename
HAVING COUNT(r.RaceId) > 20
ORDER BY average_position DESC;
```

	driverID int64 1 - 855	average_position fl... 2.181818181818181...	total_races int64 21 - 358	surname object Schumacher 1.2% Fittipaldi 1.2% 239 others 97.6%	forename object Pedro 1.6% Mike 1.6% 198 others 96.9%
0	853	17.8125	22	Mazepin	Nikita
1	29	17.42857143	21	Yamamoto	Sakon
2	820	17.34375	35	Chilton	Max
3	819	17.3	39	Pic	Charles
4	39	17.18181818	48	Karthikeyan	Narain
5	824	16.57142857	34	Bianchi	Jules
6	849	15.43137255	61	Latifi	Nicholas
7	854	15.05263158	44	Schumacher	Mick
8	167	15	21	Campos	Adrian
9	836	15	39	Wehrlein	Pascal

255 rows, showing 10 per page Page 1 of 26

The data shows us that the driver with the lowest average finishing position with a minimum of 20 career races is **Nikita Mazepin**, with an average of **17.8**, followed by **Sakon Yamamoto with 17.4**. See below for visualisation:



The next SQL query is designed to analyse Formula 1 drivers' performance based on their career points achieved, with a minimum threshold of 20 races. It categorises drivers into four performance quartiles, assigning an experience level based on the number of races they have competed in. Additionally, it calculates the average career points across all drivers to be used as a reference point.

Subquery (QuartileData): The inner query selects driver details, calculates total career points, and assigns an experience level based on the number of races participated in. It filters out drivers with less than 20 career points.

Experience Level Assignment: Based on the number of races a driver has participated in, they are categorised into one of four experience levels: '1-50 races', '51-100 races', '101-150 races', or '151+ races'.

Performance Quartile Calculation: Using the NTILE window function, drivers are divided into four performance quartiles based on their total career points, with the first quartile representing the top performers.

Average Career Points Calculation: The query calculates the average career points across all drivers using the AVG window function.

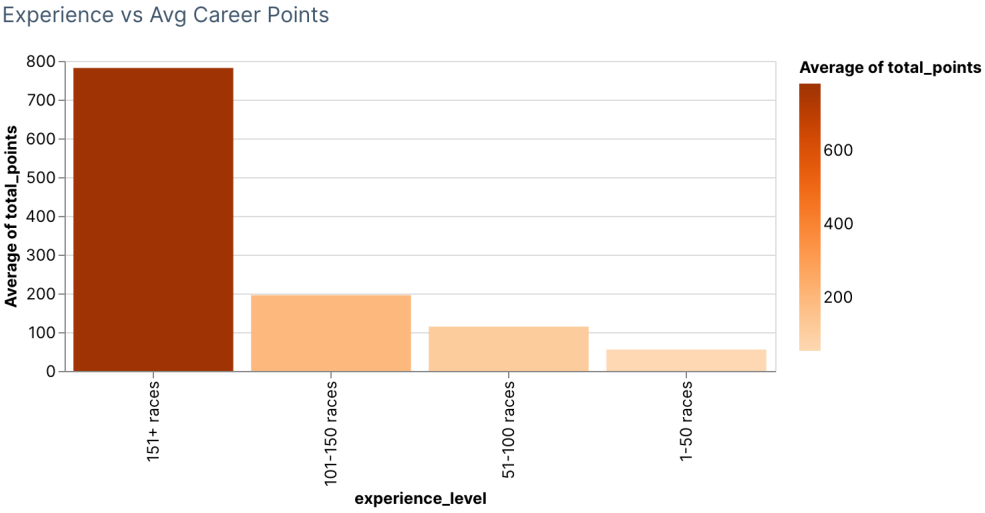
Final Selection: The outer query selects the driver's ID, surname, forename, total career points, experience level, performance quartile, and average career points across all drivers.

Ordering: Finally, the results are ordered in descending order based on total career points, showcasing the drivers with the highest points at the top.

```sql
-- Divide drivers into quartiles based on avg career points (min 20 races)

SELECT
    driverId,
    surname,
    forename,
    total_points,
    experience_level,
    NTILE(4) OVER (ORDER BY total_points DESC) AS performance_quartile,
    AVG(total_points) OVER () AS avg_career_points
FROM (
    SELECT
        r.driverId,
        d.surname,
        d.forename,
        SUM(r.points) AS total_points,
        CASE
            WHEN COUNT(r.RaceId) <= 50 THEN '1-50 races'
            WHEN COUNT(r.RaceId) <= 100 THEN '51-100 races'
            WHEN COUNT(r.RaceId) <= 150 THEN '101-150 races'
            ELSE '151+ races'
        END as experience_level
    FROM 'results.csv' r
    JOIN 'drivers.csv' d ON r.driverId = d.driverId
    GROUP BY r.driverId, d.surname, d.forename
    HAVING SUM(r.points) > 20
) AS QuartileData
ORDER BY total_points DESC;
```

	driverID int64 1 - 852	surname object Hill 2.1% Rosberg 1.4% 136 others 96.5%	forename object Carlos 2.1% Mike 2.1% 117 others 95.8%	total_points float64 21.0 - 4396.5	experience_level o... 151+ races 30.6% 51-100 races 29.9% 2 others 39.6%	performance_quar.. 1 - 4
0	1	Hamilton	Lewis	4396.5	151+ races	
1	20	Vettel	Sebastian	3098	151+ races	
2	4	Alonso	Fernando	2061	151+ races	
3	830	Verstappen	Max	1983.5	151+ races	
4	8	Raikkonen	Kimi	1873	151+ races	
5	822	Bottas	Valtteri	1778	151+ races	
6	3	Rosberg	Nico	1594.5	151+ races	
7	30	Schumacher	Michael	1566	151+ races	
8	817	Ricciardo	Daniel	1307	151+ races	
9	18	Button	Jenson	1235	151+ races	

144 rows, showing 10 per page Page 1 of 15



As expected, more races drastically increases the average of a driver's total career points. But what effect does experience level have on other performance metrics?

Next, we will examine Formula 1 drivers based on their average career finish positions, using the same format and criteria as above:

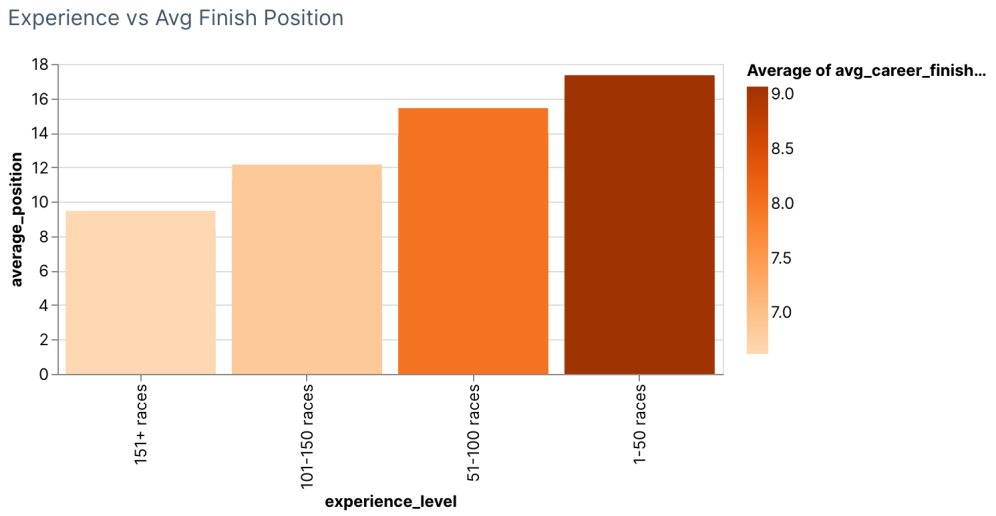
```sql
-- Divide drivers into quartiles based on avg career finish position (min 20 races)

SELECT
    driverId,
    surname,
    forename,
    average_position,
    total_races,
    experience_level,
    NTILE(4) OVER (PARTITION BY experience_level ORDER BY average_position) AS performance_quartile,
    AVG(average_position) OVER (PARTITION BY experience_level) AS avg_career_finish_position
FROM (
    SELECT
```

```
        r.driverId,
        d.surname,
        d.forename,
        AVG(NULLIF(CAST(r.position AS SIGNED), 0)) AS average_position,
        COUNT(r.RaceId) AS total_races,
        CASE
            WHEN COUNT(r.RaceId) <= 50 THEN '1-50 races'
            WHEN COUNT(r.RaceId) <= 100 THEN '51-100 races'
            WHEN COUNT(r.RaceId) <= 150 THEN '101-150 races'
            ELSE '151+ races'
        END as experience_level
    FROM 'results.csv' r
    JOIN 'drivers.csv' d ON r.driverId = d.driverId
    WHERE r.position <> '0'
    GROUP BY r.driverId, d.surname, d.forename
    HAVING total_races > 20
) AS QuartileData
ORDER BY experience_level, average_position ASC;
```

	driverID int64 1 - 854	surname object Hill 1.7% Schumacher 1.7% 163 others 96.6%	forename object Pedro 1.7% Carlos 1.7% 147 others 96.6%	average_position fl... 2.181818181818181...	total_races int64 21 - 291	experience_level o... 1-50 races 46.6 51-100 races 34.1 2 others 19.3
0	647	Ascari	Alberto	2.181818182	22	1-50 races
1	579	Fangio	Juan	2.25	44	1-50 races
2	642	Farina	Nino	2.814814815	27	1-50 races
3	475	Moss	Stirling	3.459459459	37	1-50 races
4	578	Hawthorn	Mike	3.696969697	33	1-50 races
5	373	Clark	Jim	3.84	50	1-50 races
6	231	Hunt	James	4.195652174	46	1-50 races
7	358	Rindt	Jochen	4.2	25	1-50 races
8	554	Behra	Jean	4.92	25	1-50 races
9	327	Cevert	Francois	5.034482759	29	1-50 races

176 rows, showing 10 per page Page 1 of 18

For our final driver performance chart, let's examine the effect of experience on average finish position:



The data shows that drivers with more experience have a higher average career finishing position. This is logical as drivers performing poorly won't get the opportunity to keep their seat for long, thus not allowing them to participate in a high number of races. Formula 1 teams are notoriously ruthless with who they employ, and there is always a plethora of driving talent waiting for their opportunity.

Constructor Performance Analysis

To wrap things up, we're going to briefly examine constructor performance.

```
-- Find the constructor with the highest average finish (min 20 races)

SELECT
    r.constructorId,
    AVG(NULLIF(CAST(r.position AS INT), 0)) AS average_position,
    COUNT(r.RaceId) AS total_races,
FROM 'results.csv' r
JOIN 'constructor_results.csv' cr ON r.constructorId = cr.constructorId
GROUP BY r.constructorId
HAVING COUNT(r.RaceId) > 20
ORDER BY average_position ASC;
```

	constructorId int64 1 - 214	average_position fl... 3.65517241379310...	total_races int64 25 - 2347000
0	196	3.655172414	920
1	191	3.871794872	2541
2	23	4.0625	578
3	131	4.153225806	145040
4	118	4.185185185	781
5	190	4.333333333	357
6	180	4.610169492	5632
7	195	4.625	252
8	6	4.648600357	2347000
9	187	4.82	3393

126 rows, showing 10 per page Page 1 of 13

The data shows that **Constructor ID #196** has highest average career finish at **3.65**, followed by **Constructor ID #191 at 3.87.**

```
-- Find the constructor with the lowest average finish (min 20 races)

SELECT
    r.constructorId,
    AVG(NULLIF(CAST(r.position AS INT), 0)) AS average_position,
    COUNT(r.RaceId) AS total_races,
FROM 'results.csv' r
JOIN 'constructor_results.csv' cr ON r.constructorId = cr.constructorId
GROUP BY r.constructorId
HAVING COUNT(r.RaceId) > 20
ORDER BY average_position DESC;
```

	constructorId int64	average_position fl...	total_races int64

	1 - 214	3.65517241379310...	25 - 2347000	
0	164	18.98734177	6612	
1	166	17.87755102	2888	
2	206	17.19148936	6431	
3	207	16.67045455	6272	
4	205	16.37735849	2888	
5	209	16.33333333	3042	
6	12	15.9	578	
7	14	15.4	32	
8	97	15	49	
9	8	14.5	3042	

126 rows, showing 10 per page Page 1 of 13

Constructor ID #164 has lowest average career finish at **18.98**, followed by **Constructor ID #166 at 3.87.**

Insights & Conclusion

Recap

Let's recap our findings of the F1 analysis from 1953-2020:

- Most races: Fernando Alonso (358)
- Most career points: Lewis Hamilton (4396)
- Fastest ever lap: Carlos Sainz (Red Bull Ring, 07-12-2020)
- Highest average driver finish position (min 20 races): Alberto Ascari (1.33)
- Lowest average driver finish position (min 20 races): Nikita Mazepin (15.8)
- Highest average constructor finish position (min 20 races): Constructor #106
- Lowest average constructor finish position (min 20 races): Constructor #206

Who is the Greatest Driver of All Time?

Determining an answer to this question in an objective manner is, realistically, a near impossible task. As fans of the sport will naturally understand, there are simply too many variables at play to be able to accurately conclude a definitive list through statistics alone.

The Formula 1 points system has undergone numerous changes over the years, evolving to adapt to the sport's shifting dynamics and competitive landscape. These alterations in the distribution of points have significant implications when attempting to compare drivers across different eras.

In earlier years, only a handful of top finishers were awarded points, whereas in more recent formats, points are distributed more broadly down the grid. This means that drivers in the modern era have the potential to accumulate points more consistently, even if they are not consistently finishing in the top positions.

Additionally, advancements in technology have dramatically transformed Formula 1 cars, making them faster, safer, and more reliable. These technological improvements, coupled with changes in regulations and car design, have resulted in substantial disparities in car performance across different periods.

As a result, comparing drivers solely based on their points tally or race finishes can be misleading, as it does not account for the varying levels of competition, car performance, and points distribution systems in place at different times. Therefore, while statistics provide valuable insights, they may not fully capture a driver's skill, impact, or the context of their achievements, making it challenging to definitively conclude who the better drivers are across different Formula 1 eras.

It's also equally critical to consider which constructors the drivers are representing. The disparity in Formula 1 team budgets plays a pivotal role in influencing driver performance, creating a significant divide between the teams at the top and those further down the grid. Teams with larger budgets, often backed by major automotive manufacturers or wealthy entities, have the financial resources to invest in cutting-edge technology, top-tier engineering talent, and extensive research and development.

This investment translates into faster, more reliable cars, providing their drivers with a competitive advantage on the track. In contrast, teams with smaller budgets face limitations in their ability to develop and maintain high-performance vehicles, often resulting in less competitive machinery. This budgetary constraint can hinder a driver's ability to consistently compete at the front of the grid, regardless of their individual skill and talent.

Additionally, well-funded teams can offer more comprehensive support, including advanced simulation tools, extensive data analysis, and superior pit strategy, all of which contribute to optimising a driver's performance. Consequently, while driver skill is undeniably a crucial factor in Formula 1 success, the financial capabilities of their team play a substantial role in determining the extent to which a driver can realize their potential and achieve top results.

Conclusion

This Formula 1 Data Analysis project has helped us breakdown the intricate dynamics of Formula 1 racing, providing valuable insights into driver performance and statistics. The robust capabilities of SQL allowed us to delve deep into data manipulation, uncovering patterns, trends, and performance metrics that shed light on the iconic drivers, teams, and circuits that have defined the sport for decades.

The integration of data visualisations via Deepnote and Tableau brought our findings to life, transforming complex datasets into interactive and engaging charts. This makes the data more accessible, allowing enthusiasts and analysts alike to interact with the data in a meaningful way.

We have also established that while statistics provide valuable insights, they may not fully capture a driver's skill, impact, or the context of their achievements, making it challenging to definitively conclude who the better drivers are across different Formula 1 eras. Through this performance analysis and comprehensive reporting, we have provided a data-driven lens through which the world of Formula 1 can be understood and appreciated with a unique and insightful perspective.

Thank you for reading! Who gets your vote for greatest driver of all time?

`BestDriverEver`

Lewis Hamilton