Formula 1 Data Analysis Report

DATASOURCE:

Formula 1 Data (Kaggle) :
https://www.kaggle.com/datasets/thedevastator/formula-one-racing-a-comprehensive-data-analysis?select=constructor_results.csv

DESCRIPTION:

The Formula 1 data set contains the race summary, including details such as races, drivers, teams and results of the races conducted over the years, we have constrained the data to the years of 2012-2022.

PROJECT DESCRIPTION:

The rationale behind this project is to extract good insights from the Formula One dataset. The dataset has multiple csv files with varied data of races from 2012-2022 and each file has multiple types of columns in it. So, this dataset provides feasibility to write and execute multiple complex queries about the Formula One races. Since the amount of data is large, query optimization is also implemented.

Concepts that have been implemented are:

1) DesigninganERdiagramfortheabove-mentioneddatasetandconvertthe same to a schema using PostgreSQL
2) Using normalization during data preparation
3) UsingSQLindexingforbetterperformance
4) WritingandexecutingcomplexSQLqueries
5) Enhanceperformancewithqueryoptimizers

FILES ATTACHED:

- RawFormulaOnedatasetobtainedfromtheresource:csvformat,attached as F1_dataset folder.
- DataPreprocessingcodefiledoneinPython(Data_cleaning.ipynb,Data cleaning.py).
- Processeddatasetusedtoanalyzethedata(modified_dataset)
- Datadefinition(DDL)statementsforthedataset(SQLscripts,textfile)
- DataManipulationLanguage(DML)statementsforthedataset(SQL scripts,text file)
- Indexingstatements(SQLscripts,textfile)
- ThequeriesperformedonthedatabaseattachedasscriptandSQL_text.
- Finalprojectreport.

DATASETTRANSFORMATION:

To make the dataset efficient and compatible with the designed database schema in the PostgreSQL, the original dataset is preprocessed and cleaned on different conditions. All the NULL, duplicate and redundant records which are not significant in the process of analysing the data are dropped.

The original data set contains an overall information of races from 1950 to 2022, which was modified to obtain records of races from 2012 to 2022. The unused columns of the dataset of the analysis were dropped. Also, values in the columns which have data-type format issues are dropped to maintain the integrity of the data.

DATASET:

The FormulaOne database obtained after the data cleaning is imported into the schema which has the below datasets

1. circuits1.csv
2. constructors1.csv
3. races1.csv
4. constructor_standings1.csv
5. drivers1.csv
6. driver_standings1.csv
7. qualifying1.csv
8. lap_times1.csv
9. pit_stops1.csv
10. status1.csv
11. results1.csv

DDL and DML Statements:

The DDL Statements developed to create the database are included in the SQL scripts file

DML statements are written to add the necessary data to the created tables, which are subsequently used to execute various analyses to help comprehend the data.

```
COPY circuits(circuitId ,circuitRef ,name ,location ,country ) FROM
'C:\Users\pmegh\Documents\CIS-556_DB_Sys\Project\cleaned_files\circuits1.csv'
DELIMITER ‘,’ CSV Header;
```

```
COPY constructors( constructorId , constructorRef , name , nationality ) FROM
'C:\Users\pmegh\Documents\CIS-556_DB_Sys\Project\cleaned_files\constructors1.
csv' DELIMITER ‘,’ CSV Header;
```

```
COPY races( raceId , year , round , circuitId , name , r_date , r_time) FROM
'C:\Users\pmegh\Documents\CIS-556_DB_Sys\Project\cleaned_files\races1.csv'
DELIMITER ‘,’ CSV Header;
```

```
COPY constructorstandings( constructorstandingsId , raceId , constructorsId ,
points , position , wins ) FROM
'C:\Users\pmegh\Documents\CIS-556_DB_Sys\Project\cleaned_files\constructor_st
andings1.csv' DELIMITER ‘,’ CSV Header;
```

```
COPY drivers( driverId , driverRef , number , code ,forename , surname , dob
, nationality ) FROM
'C:\Users\pmegh\Documents\CIS-556_DB_Sys\Project\cleaned_files\drivers1.csv'
DELIMITER ‘,’ CSV Header;
```

```
COPY driver_standings(raceId , driverId ,points , position , wins ) FROM
'C:\Users\pmegh\Documents\CIS-556_DB_Sys\Project\cleaned_files\driver_standin
gs1.csv' DELIMITER ‘,’ CSV Header;
```

```
COPY qualifying( raceId , driverId , constructorId , number , position , q1 ,
q2 , q3 ) FROM
'C:\Users\pmegh\Documents\CIS-556_DB_Sys\Project\cleaned_files\qualifying1.cs
v' DELIMITER ‘,’ CSV Header;
```

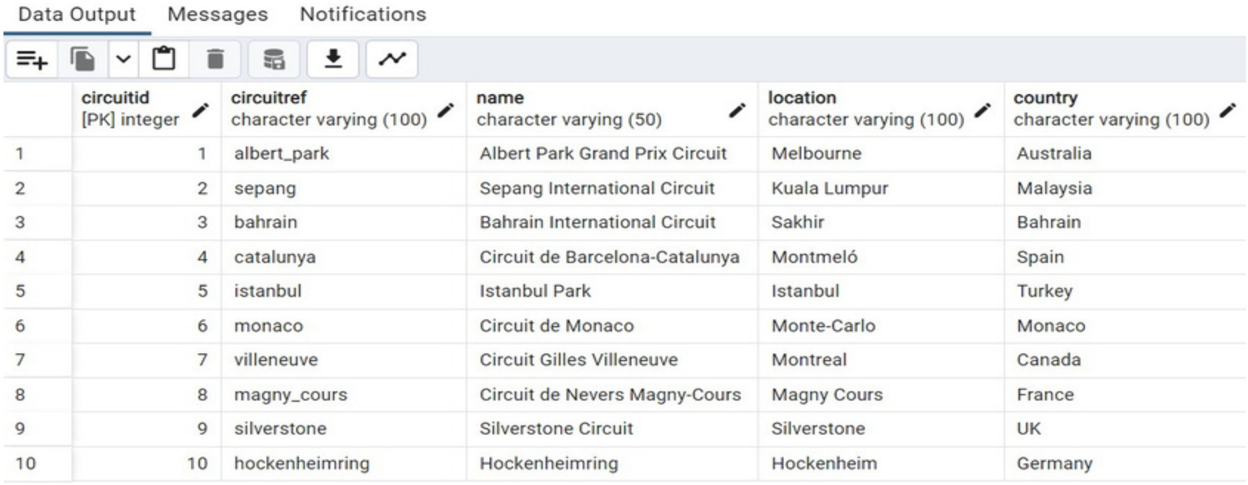
```
COPY laptimes( raceId , driverId , lap , position , l_time , milliseconds )
FROM
'C:\Users\pmegh\Documents\CIS-556_DB_Sys\Project\cleaned_files\lap_times1.csv
' DELIMITER ‘,’ CSV Header;
```

```
COPY pitstops(raceId , driverId , stop , lap , p_time , duration ) FROM
'C:\Users\pmegh\Documents\CIS-556_DB_Sys\Project\cleaned_files\pit_stops1.csv
' DELIMITER ‘,’ CSV Header;
```

```
COPY status(statusId , status ) FROM
'C:\Users\pmegh\Documents\CIS-556_DB_Sys\Project\cleaned_files\status1.csv'
DELIMITER ‘,’ CSV Header;
```

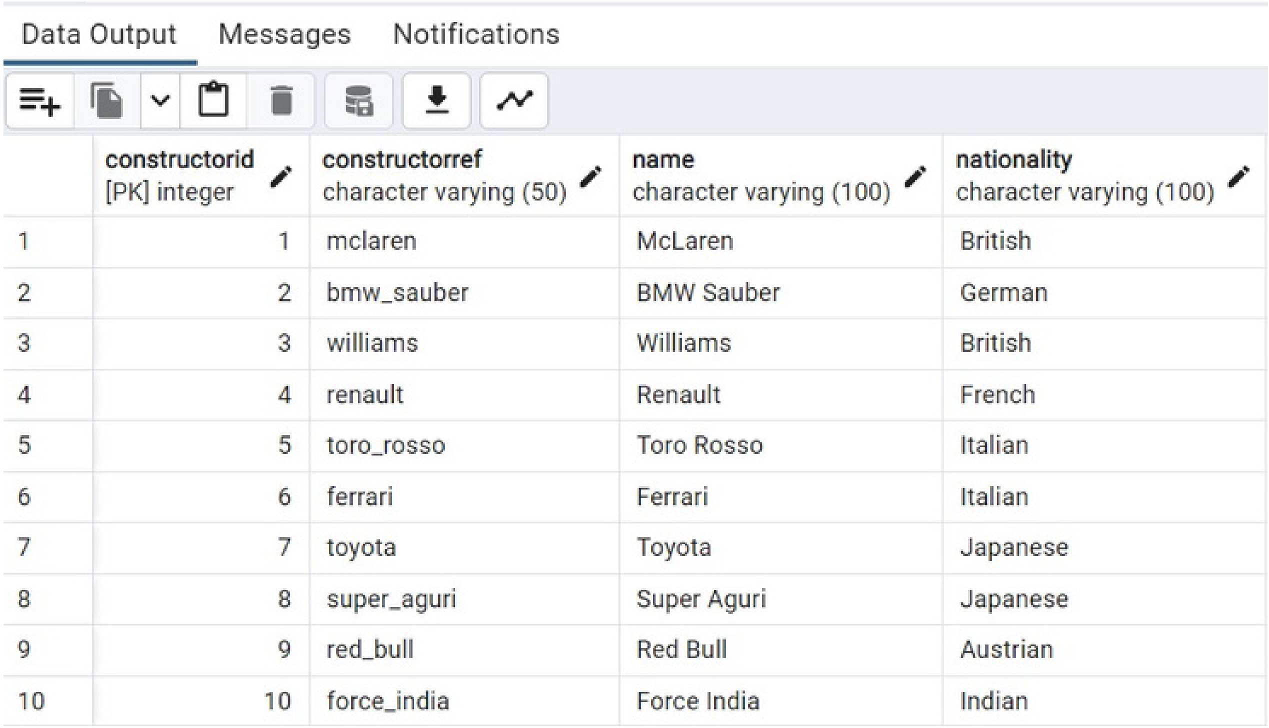
```
COPY results(resultid ,raceid ,driverid ,constructorid ,number ,grid
,position ,points ,laps ,fastest_lap ,rank ,fastest_laptime ,fastest_lapspeed
,statusid ) FROM
'C:\Users\pmegh\Documents\CIS-556_DB_Sys\Project\cleaned_files\results1.csv'
DELIMITER ‘,’ CSV Header;
```

The snapshots of the data of all the individual tables which are developed in the postgre are attached below circuits1.csv



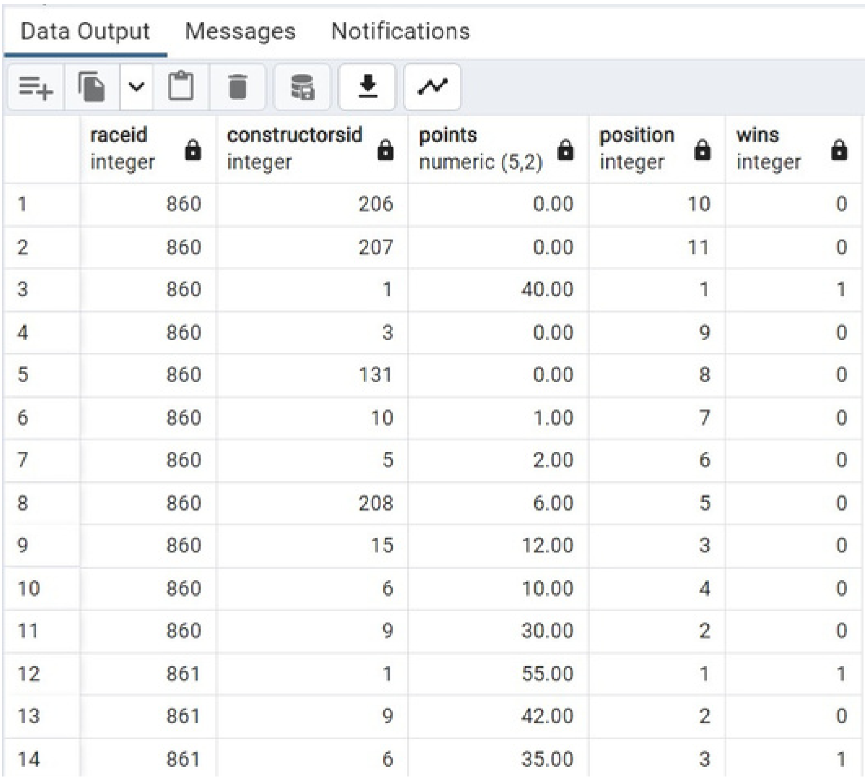
	circuitid [PK] integer	circuitref character varying (100)	name character varying (50)	location character varying (100)	country character varying (100)
1	1	albert_park	Albert Park Grand Prix Circuit	Melbourne	Australia
2	2	sepang	Sepang International Circuit	Kuala Lumpur	Malaysia
3	3	bahrain	Bahrain International Circuit	Sakhir	Bahrain
4	4	catalunya	Circuit de Barcelona-Catalunya	Montmeló	Spain
5	5	istanbul	Istanbul Park	Istanbul	Turkey
6	6	monaco	Circuit de Monaco	Monte-Carlo	Monaco
7	7	villeneuve	Circuit Gilles Villeneuve	Montreal	Canada
8	8	magny_cours	Circuit de Nevers Magny-Cours	Magny Cours	France
9	9	silverstone	Silverstone Circuit	Silverstone	UK
10	10	hockenheimring	Hockenheimring	Hockenheim	Germany

constructors1.csv



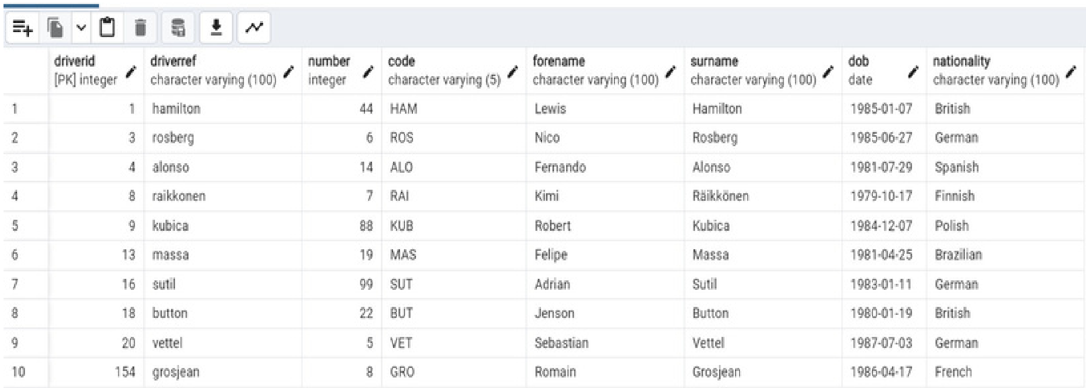
	constructorid [PK] integer	constructorref character varying (50)	name character varying (100)	nationality character varying (100)
1	1	mclaren	McLaren	British
2	2	bmw_sauber	BMW Sauber	German
3	3	williams	Williams	British
4	4	renault	Renault	French
5	5	toro_rosso	Toro Rosso	Italian
6	6	ferrari	Ferrari	Italian
7	7	toyota	Toyota	Japanese
8	8	super_aguri	Super Aguri	Japanese
9	9	red_bull	Red Bull	Austrian
10	10	force_india	Force India	Indian

constructor_standings.csv



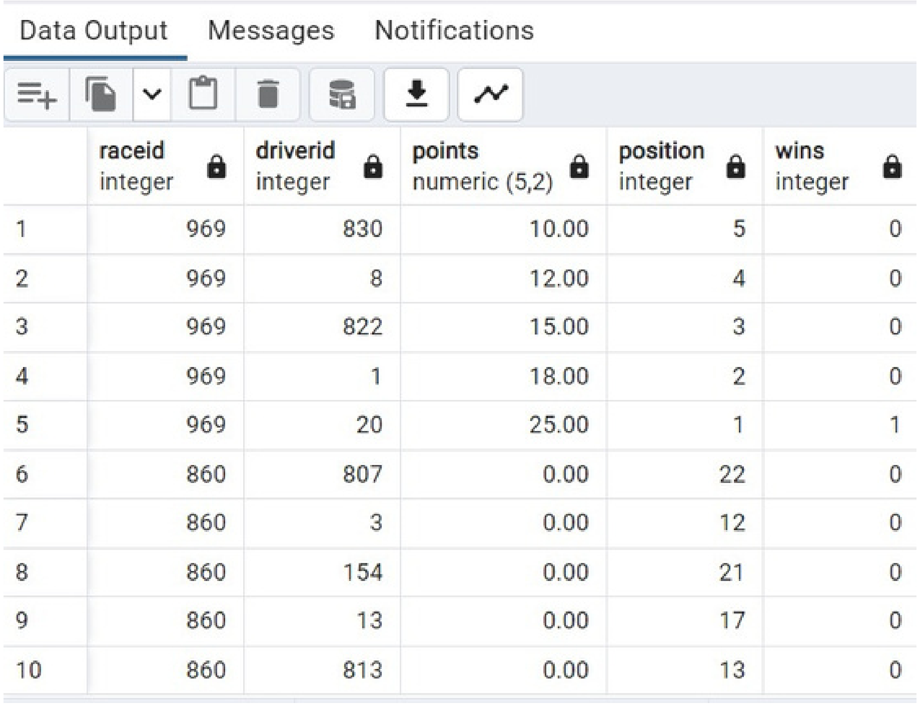
	raceid integer	constructorsid integer	points numeric (5,2)	position integer	wins integer
1	860	206	0.00	10	0
2	860	207	0.00	11	0
3	860	1	40.00	1	1
4	860	3	0.00	9	0
5	860	131	0.00	8	0
6	860	10	1.00	7	0
7	860	5	2.00	6	0
8	860	208	6.00	5	0
9	860	15	12.00	3	0
10	860	6	10.00	4	0
11	860	9	30.00	2	0
12	861	1	55.00	1	1
13	861	9	42.00	2	0
14	861	6	35.00	3	1

drivers.csv



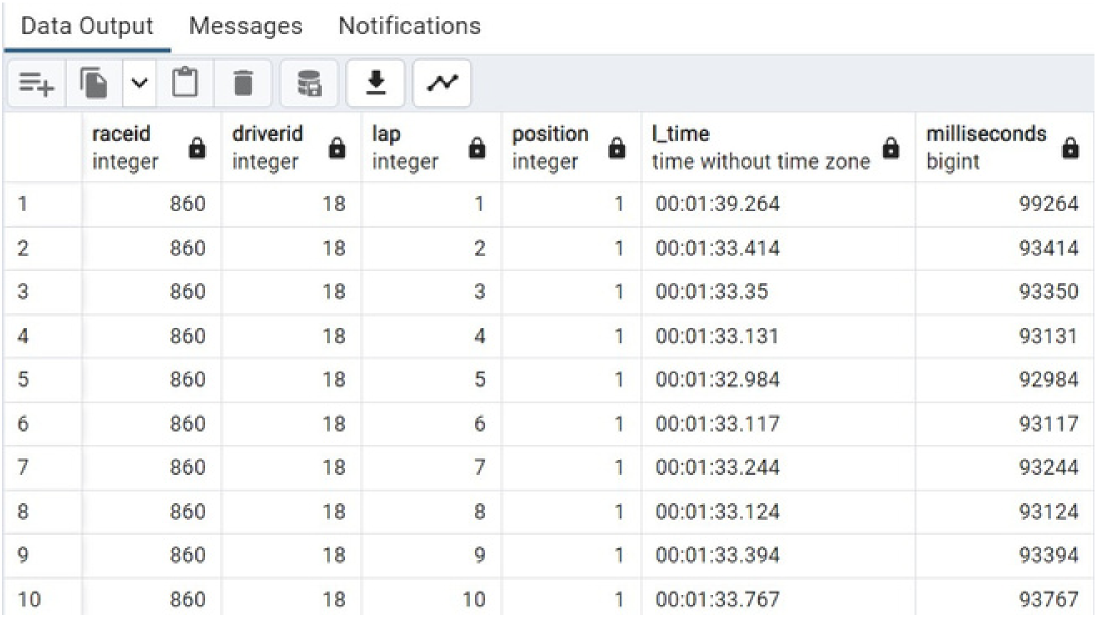
	driverid [PK] integer	driverref character varying (100)	number integer	code character varying (5)	forename character varying (100)	surname character varying (100)	dob date	nationality character varying (100)
1	1	hamilton	44	HAM	Lewis	Hamilton	1985-01-07	British
2	3	rosberg	6	ROS	Nico	Rosberg	1985-06-27	German
3	4	alonso	14	ALO	Fernando	Alonso	1981-07-29	Spanish
4	8	raikkonen	7	RAI	Kimi	Räikkönen	1979-10-17	Finnish
5	9	kubica	88	KUB	Robert	Kubica	1984-12-07	Polish
6	13	massa	19	MAS	Felipe	Massa	1981-04-25	Brazilian
7	16	sutil	99	SUT	Adrian	Sutil	1983-01-11	German
8	18	button	22	BUT	Jenson	Button	1980-01-19	British
9	20	vettel	5	VET	Sebastian	Vettel	1987-07-03	German
10	154	grosjean	8	GRO	Romain	Grosjean	1986-04-17	French

driverstandings.csv



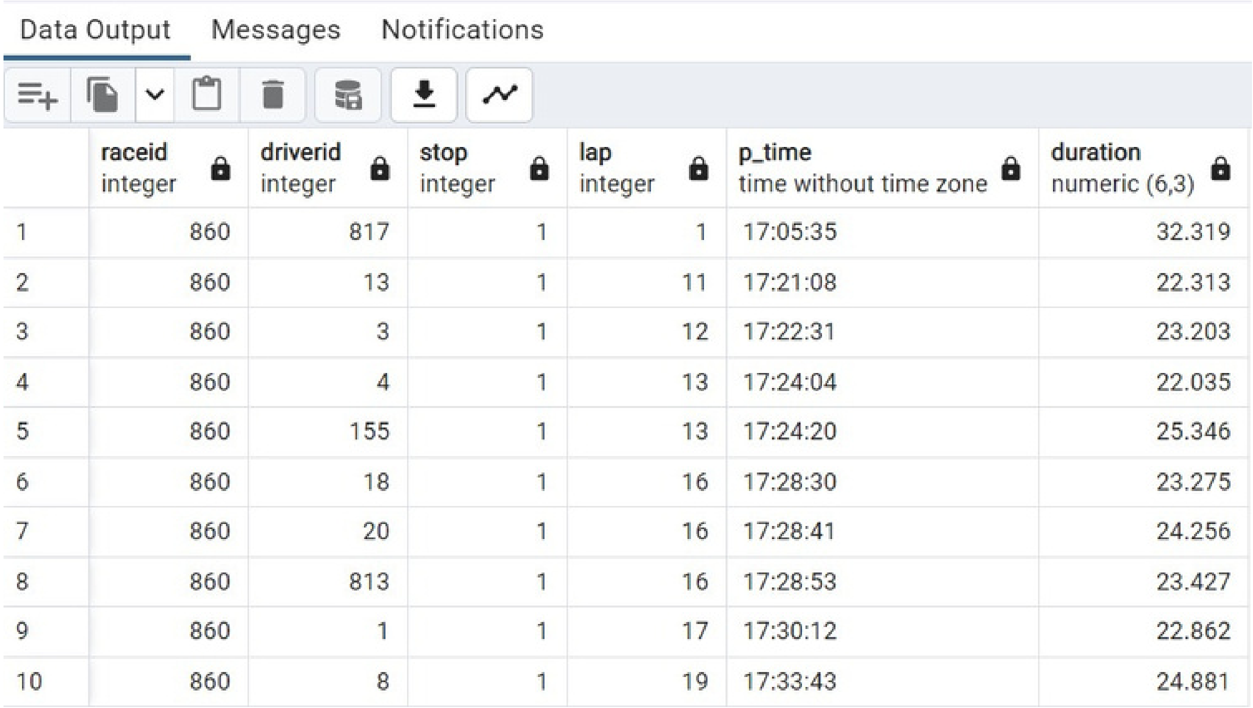
	raceid integer	driverid integer	points numeric (5,2)	position integer	wins integer
1	969	830	10.00	5	0
2	969	8	12.00	4	0
3	969	822	15.00	3	0
4	969	1	18.00	2	0
5	969	20	25.00	1	1
6	860	807	0.00	22	0
7	860	3	0.00	12	0
8	860	154	0.00	21	0
9	860	13	0.00	17	0
10	860	813	0.00	13	0

laptimes.csv



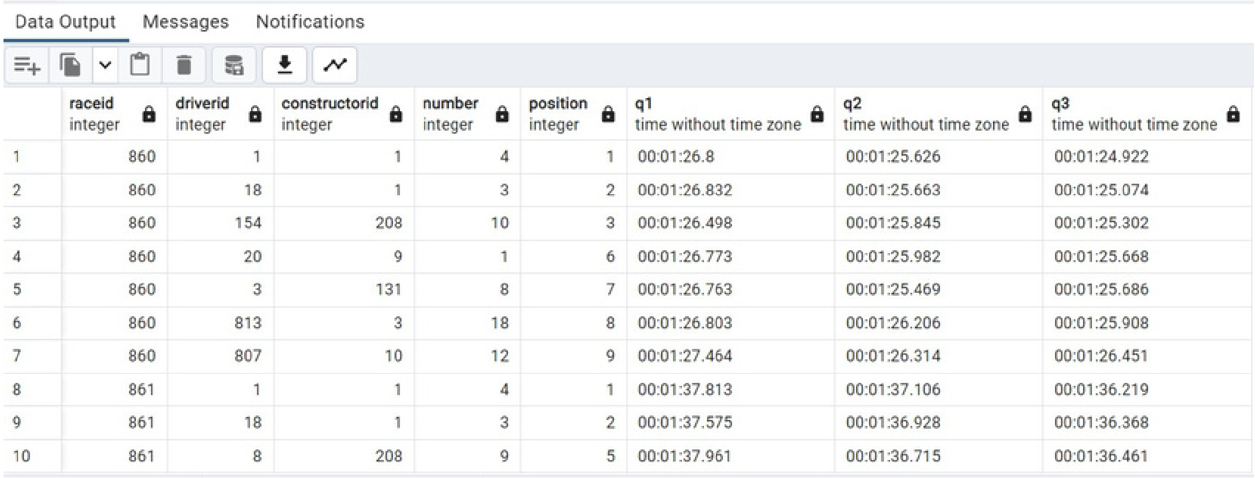
	raceid integer	driverid integer	lap integer	position integer	l_time time without time zone	milliseconds bigint
1	860	18	1	1	00:01:39.264	99264
2	860	18	2	1	00:01:33.414	93414
3	860	18	3	1	00:01:33.35	93350
4	860	18	4	1	00:01:33.131	93131
5	860	18	5	1	00:01:32.984	92984
6	860	18	6	1	00:01:33.117	93117
7	860	18	7	1	00:01:33.244	93244
8	860	18	8	1	00:01:33.124	93124
9	860	18	9	1	00:01:33.394	93394
10	860	18	10	1	00:01:33.767	93767

pitstops.csv



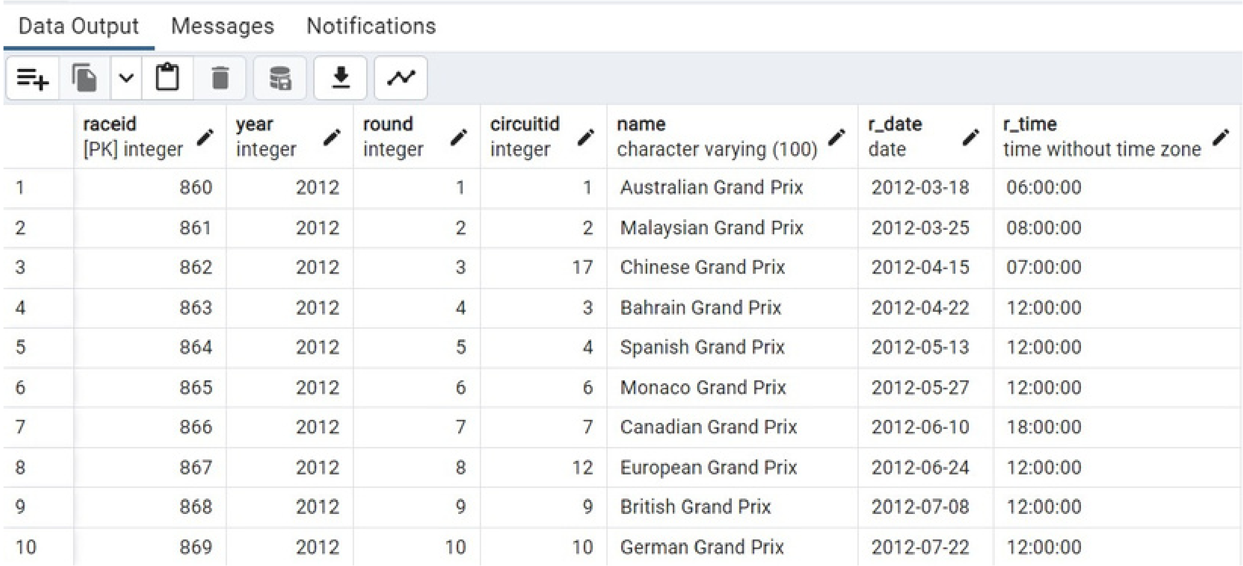
	raceid integer	driverid integer	stop integer	lap integer	p_time time without time zone	duration numeric (6,3)
1	860	817	1	1	17:05:35	32.319
2	860	13	1	11	17:21:08	22.313
3	860	3	1	12	17:22:31	23.203
4	860	4	1	13	17:24:04	22.035
5	860	155	1	13	17:24:20	25.346
6	860	18	1	16	17:28:30	23.275
7	860	20	1	16	17:28:41	24.256
8	860	813	1	16	17:28:53	23.427
9	860	1	1	17	17:30:12	22.862
10	860	8	1	19	17:33:43	24.881

qualifying.csv



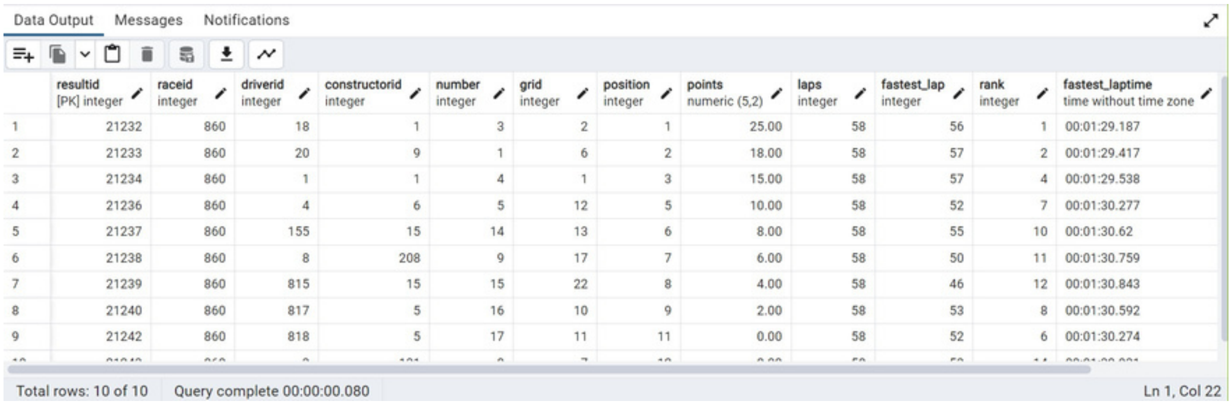
	raceid integer	driverid integer	constructorid integer	number integer	position integer	q1 time without time zone	q2 time without time zone	q3 time without time zone
1	860	1	1	4	1	00:01:26.8	00:01:25.626	00:01:24.922
2	860	18	1	3	2	00:01:26.832	00:01:25.663	00:01:25.074
3	860	154	208	10	3	00:01:26.498	00:01:25.845	00:01:25.302
4	860	20	9	1	6	00:01:26.773	00:01:25.982	00:01:25.668
5	860	3	131	8	7	00:01:26.763	00:01:25.469	00:01:25.686
6	860	813	3	18	8	00:01:26.803	00:01:26.206	00:01:25.908
7	860	807	10	12	9	00:01:27.464	00:01:26.314	00:01:26.451
8	861	1	1	4	1	00:01:37.813	00:01:37.106	00:01:36.219
9	861	18	1	3	2	00:01:37.575	00:01:36.928	00:01:36.368
10	861	8	208	9	5	00:01:37.961	00:01:36.715	00:01:36.461

races.csv



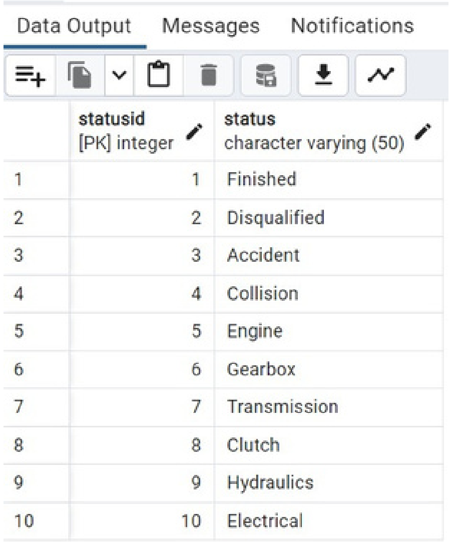
	raceid [PK] integer	year integer	round integer	circuitid integer	name character varying (100)	r_date date	r_time time without time zone
1	860	2012	1	1	Australian Grand Prix	2012-03-18	06:00:00
2	861	2012	2	2	Malaysian Grand Prix	2012-03-25	08:00:00
3	862	2012	3	17	Chinese Grand Prix	2012-04-15	07:00:00
4	863	2012	4	3	Bahrain Grand Prix	2012-04-22	12:00:00
5	864	2012	5	4	Spanish Grand Prix	2012-05-13	12:00:00
6	865	2012	6	6	Monaco Grand Prix	2012-05-27	12:00:00
7	866	2012	7	7	Canadian Grand Prix	2012-06-10	18:00:00
8	867	2012	8	12	European Grand Prix	2012-06-24	12:00:00
9	868	2012	9	9	British Grand Prix	2012-07-08	12:00:00
10	869	2012	10	10	German Grand Prix	2012-07-22	12:00:00

results.csv



	resultid [PK] integer	raceid integer	driverid integer	constructorid integer	number integer	grid integer	position integer	points numeric (5,2)	laps integer	fastest_lap integer	rank integer	fastest_laptime time without time zone
1	21232	860	18	1	3	2	1	25.00	58	56	1	00:01:29.187
2	21233	860	20	9	1	6	2	18.00	58	57	2	00:01:29.417
3	21234	860	1	1	4	1	3	15.00	58	57	4	00:01:29.538
4	21236	860	4	6	5	12	5	10.00	58	52	7	00:01:30.277
5	21237	860	155	15	14	13	6	8.00	58	55	10	00:01:30.62
6	21238	860	8	208	9	17	7	6.00	58	50	11	00:01:30.759
7	21239	860	815	15	15	22	8	4.00	58	46	12	00:01:30.843
8	21240	860	817	5	16	10	9	2.00	58	53	8	00:01:30.592
9	21242	860	818	5	17	11	11	0.00	58	52	6	00:01:30.274

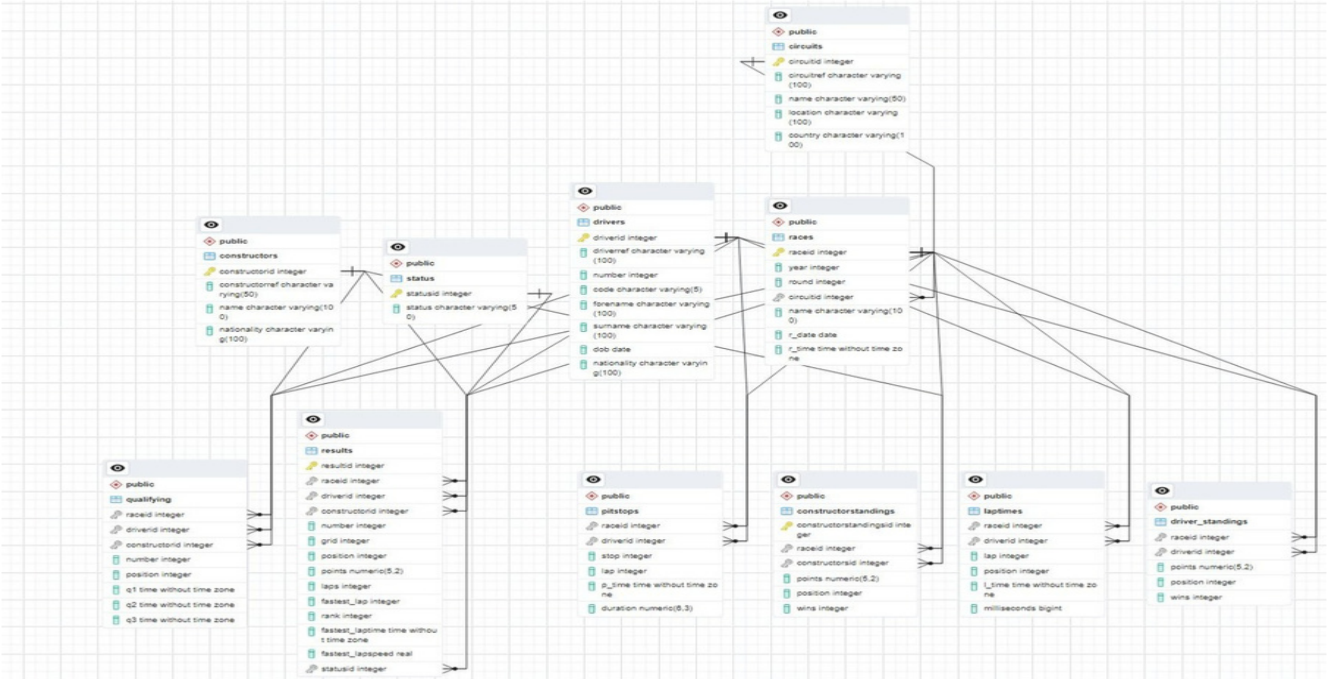
Total rows: 10 of 10   Query complete 00:00:00.080   Ln 1, Col 22 status.csv



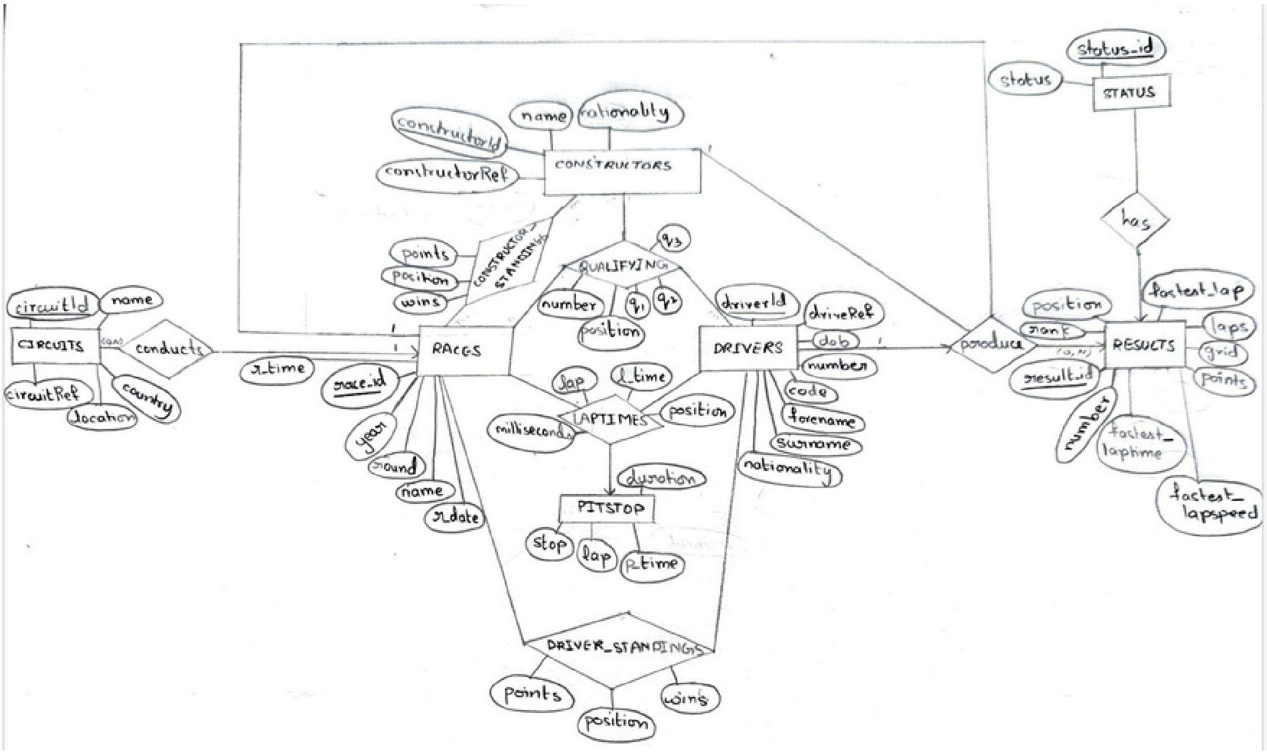
	statusid [PK] integer	status character varying (50)
1	1	Finished
2	2	Disqualified
3	3	Accident
4	4	Collision
5	5	Engine
6	6	Gearbox
7	7	Transmission
8	8	Clutch
9	9	Hydraulics
10	10	Electrical

CONCEPTUAL DESIGN:

The Schema is developed from the imported dataset, from which we can get the idea of how the data is organized and the relationship between all the entities

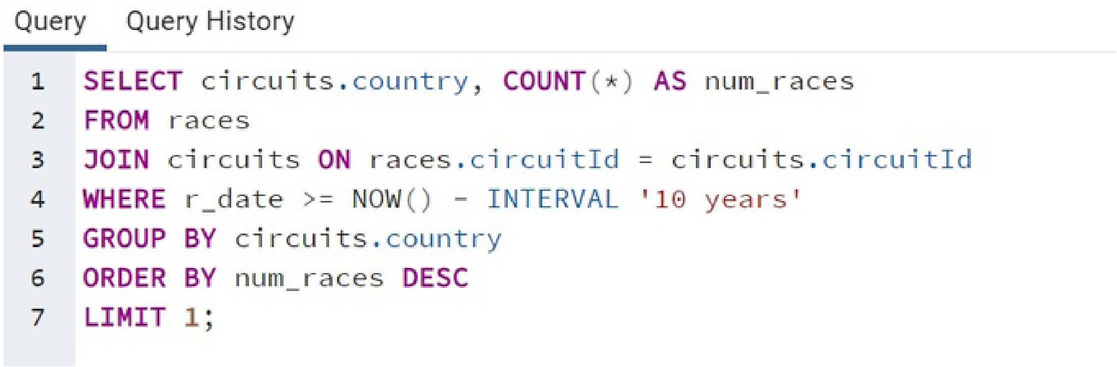


The Entity-relationship helped us gain better insight on the Formula One database, the E-R diagram obtained from the schema is drawn below:

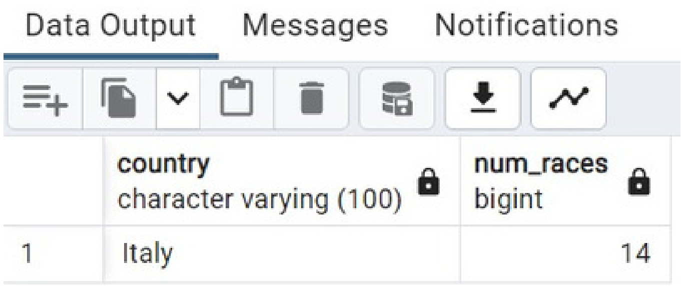


QUERIES:

1. Which country hosted the most number of races from the past 10 years

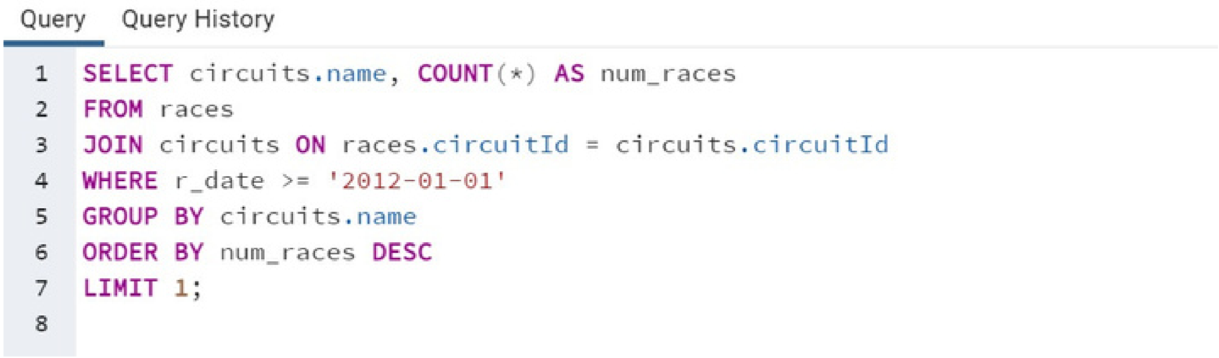


```
SELECT circuits.country, COUNT(*) AS num_races
FROM races
JOIN circuits ON races.circuitId = circuits.circuitId
WHERE r_date >= NOW() - INTERVAL '10 years'
GROUP BY circuits.country
ORDER BY num_races DESC
LIMIT 1;
```



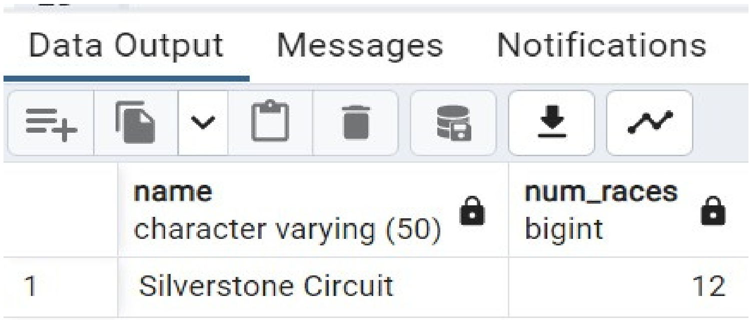
	country character varying (100)	num_races bigint
1	Italy	14

2. Whichcircuithasthehighestnumberofracesfrom2012



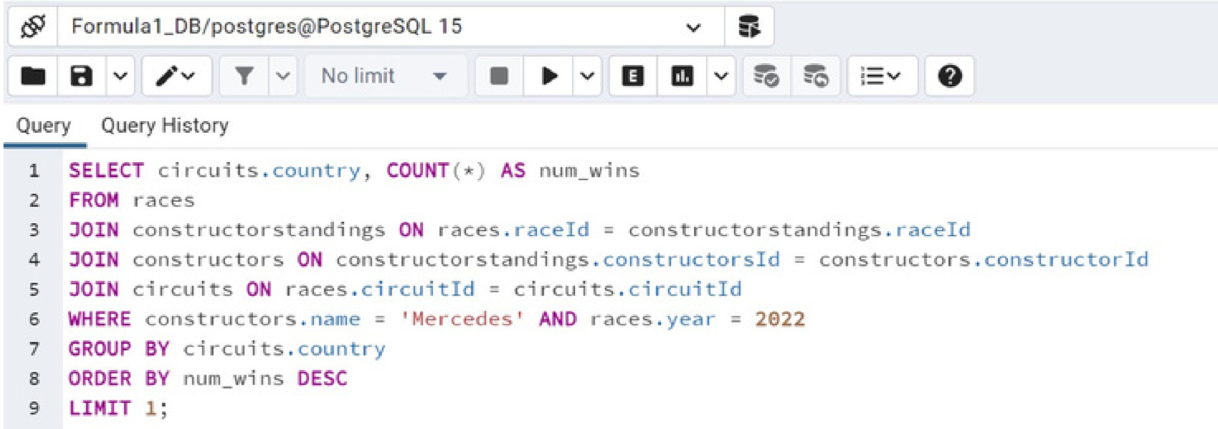
```sql
SELECT circuits.name, COUNT(*) AS num_races
FROM races
JOIN circuits ON races.circuitId = circuits.circuitId
WHERE r_date >= '2012-01-01'
GROUP BY circuits.name
ORDER BY num_races DESC
LIMIT 1;

```

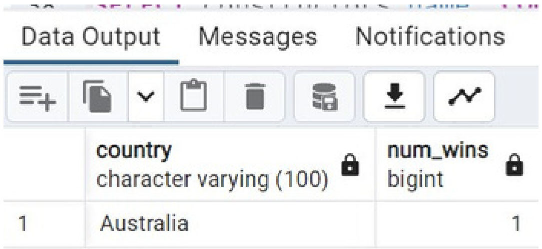


	name character varying (50)	num_races bigint
1	Silverstone Circuit	12

3. QuerytoretrievethecountryinwhichtheconstructorMercedeshasthemost
number of wins in the year 2012

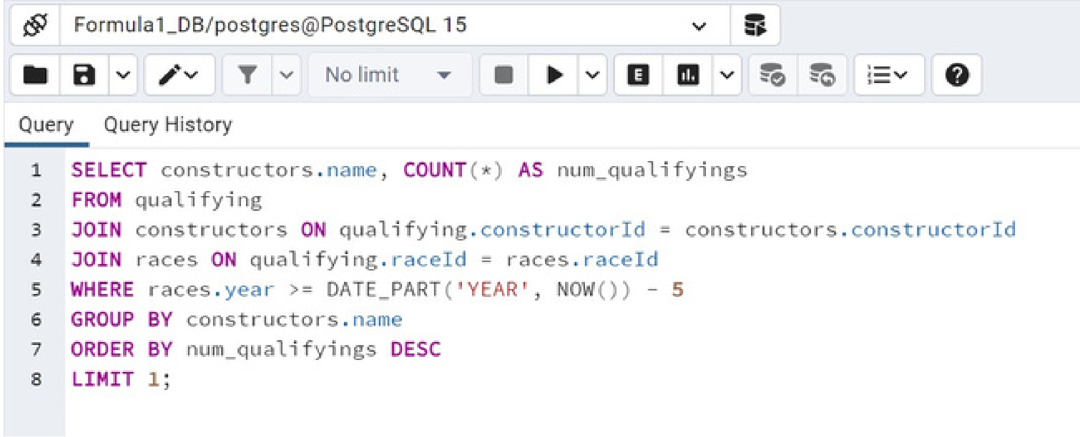


```sql
SELECT circuits.country, COUNT(*) AS num_wins
FROM races
JOIN constructorstandings ON races.raceId = constructorstandings.raceId
JOIN constructors ON constructorstandings.constructorsId = constructors.constructorId
JOIN circuits ON races.circuitId = circuits.circuitId
WHERE constructors.name = 'Mercedes' AND races.year = 2022
GROUP BY circuits.country
ORDER BY num_wins DESC
LIMIT 1;
```

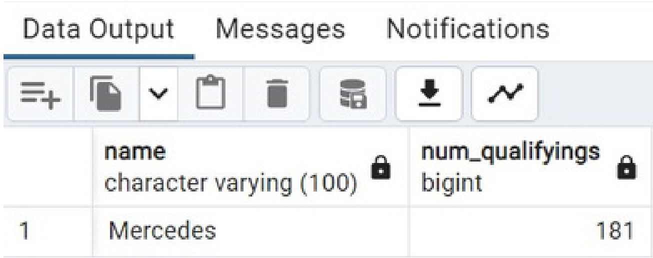


	country character varying (100)	num_wins bigint
1	Australia	1

4. Querytoretrievetheconstructorthathasthemostnumberofqualifiersinthe
past 5 years

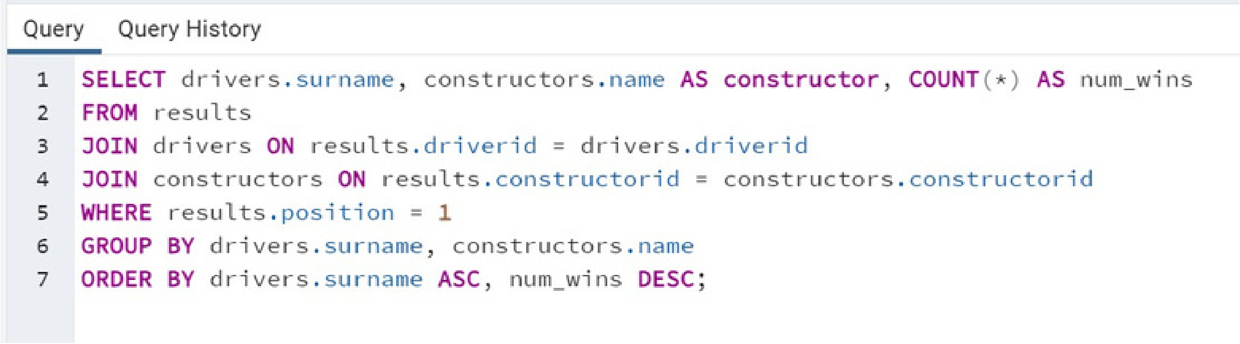


```sql
SELECT constructors.name, COUNT(*) AS num_qualifyings
FROM qualifying
JOIN constructors ON qualifying.constructorId = constructors.constructorId
JOIN races ON qualifying.raceId = races.raceId
WHERE races.year >= DATE_PART('YEAR', NOW()) - 5
GROUP BY constructors.name
ORDER BY num_qualifyings DESC
LIMIT 1;
```

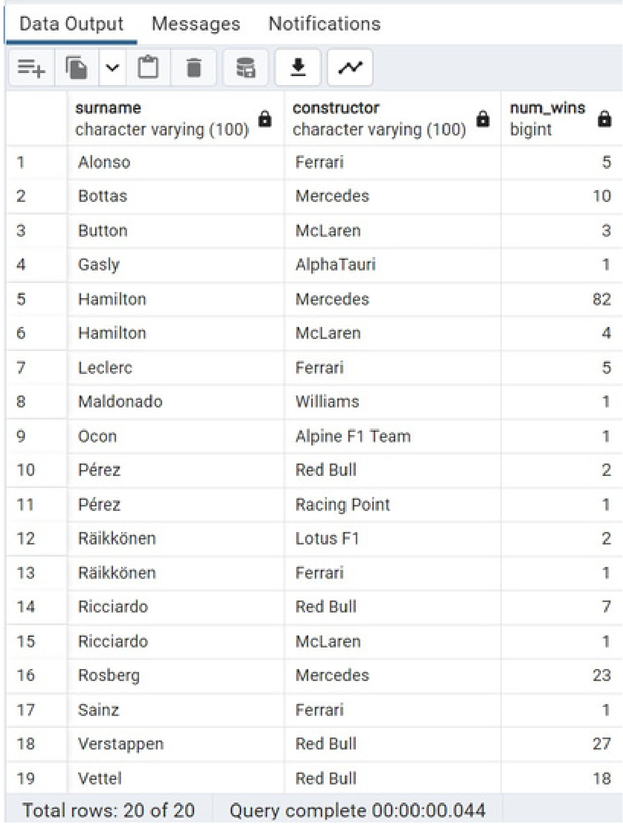


	name character varying (100)	num_qualifyings bigint
1	Mercedes	181

5. Querytoretrievethenumberoftimesadriverhaswonforaparticular constructor



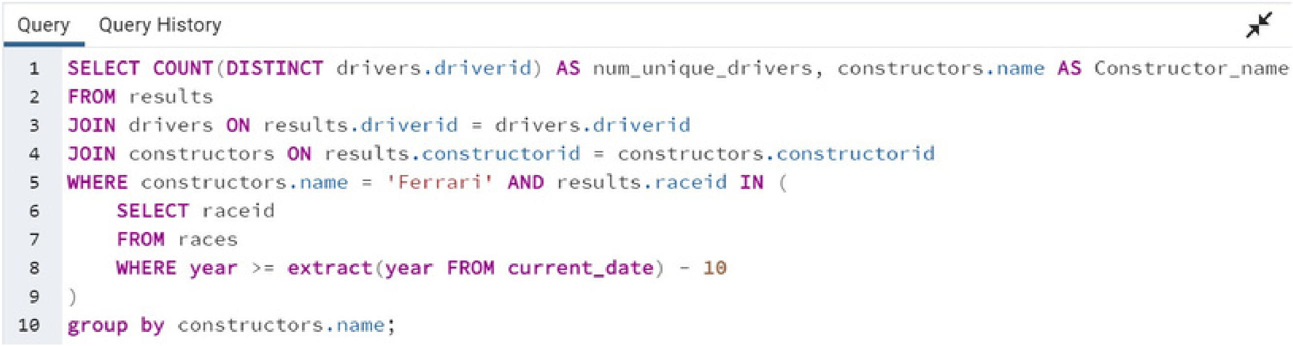
```sql
SELECT drivers.surname, constructors.name AS constructor, COUNT(*) AS num_wins
FROM results
JOIN drivers ON results.driverid = drivers.driverid
JOIN constructors ON results.constructorid = constructors.constructorid
WHERE results.position = 1
GROUP BY drivers.surname, constructors.name
ORDER BY drivers.surname ASC, num_wins DESC;
```



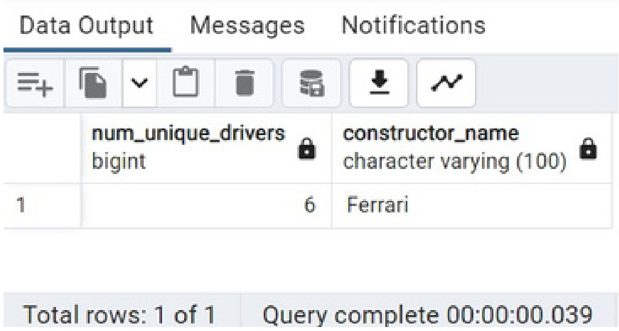
	surname character varying (100)	constructor character varying (100)	num_wins bigint
1	Alonso	Ferrari	5
2	Bottas	Mercedes	10
3	Button	McLaren	3
4	Gasly	AlphaTauri	1
5	Hamilton	Mercedes	82
6	Hamilton	McLaren	4
7	Leclerc	Ferrari	5
8	Maldonado	Williams	1
9	Ocon	Alpine F1 Team	1
10	Pérez	Red Bull	2
11	Pérez	Racing Point	1
12	Räikkönen	Lotus F1	2
13	Räikkönen	Ferrari	1
14	Ricciardo	Red Bull	7
15	Ricciardo	McLaren	1
16	Rosberg	Mercedes	23
17	Sainz	Ferrari	1
18	Verstappen	Red Bull	27
19	Vettel	Red Bull	18

Total rows: 20 of 20 Query complete 00:00:00.044

6.Query to find the number of drivers ferrari has produced in the last 10 years



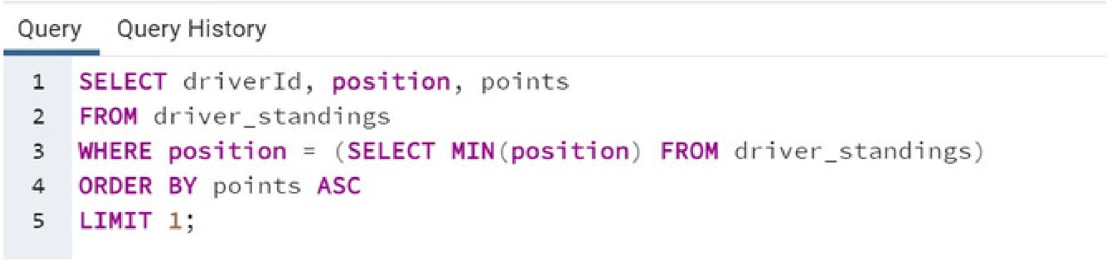
```sql
SELECT COUNT(DISTINCT drivers.driverid) AS num_unique_drivers, constructors.name AS Constructor_name
FROM results
JOIN drivers ON results.driverid = drivers.driverid
JOIN constructors ON results.constructorid = constructors.constructorid
WHERE constructors.name = 'Ferrari' AND results.raceid IN (
    SELECT raceid
    FROM races
    WHERE year >= extract(year FROM current_date) - 10
)
group by constructors.name;
```



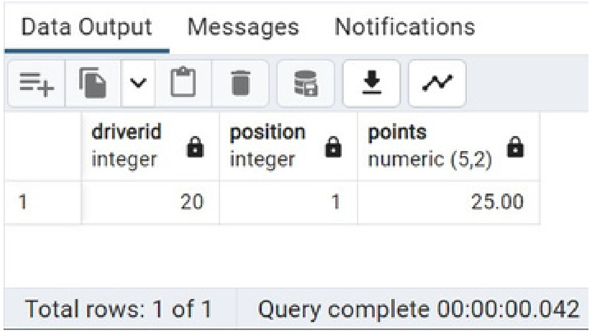
	num_unique_drivers bigint	constructor_name character varying (100)
1	6	Ferrari

Total rows: 1 of 1 Query complete 00:00:00.039

7. Query to find the driver that has the least number of points and his position in the race.



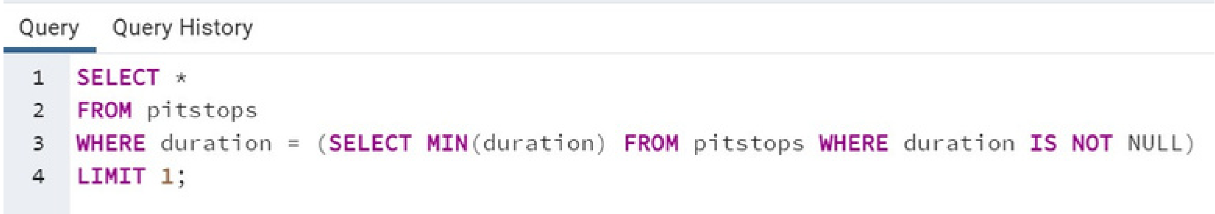
```sql
SELECT driverId, position, points
FROM driver_standings
WHERE position = (SELECT MIN(position) FROM driver_standings)
ORDER BY points ASC
LIMIT 1;
```



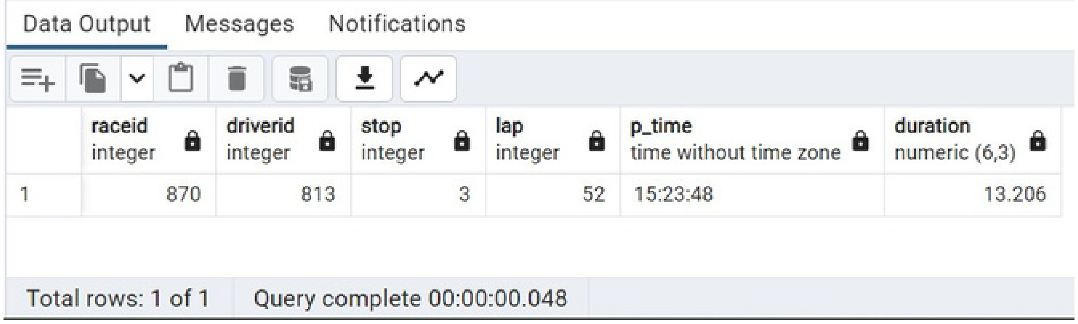
	driverid integer	position integer	points numeric (5,2)
1	20	1	25.00

Total rows: 1 of 1 Query complete 00:00:00.042

8. Query to retrieve the driver details of fastest pit stop made in the past 10 years .



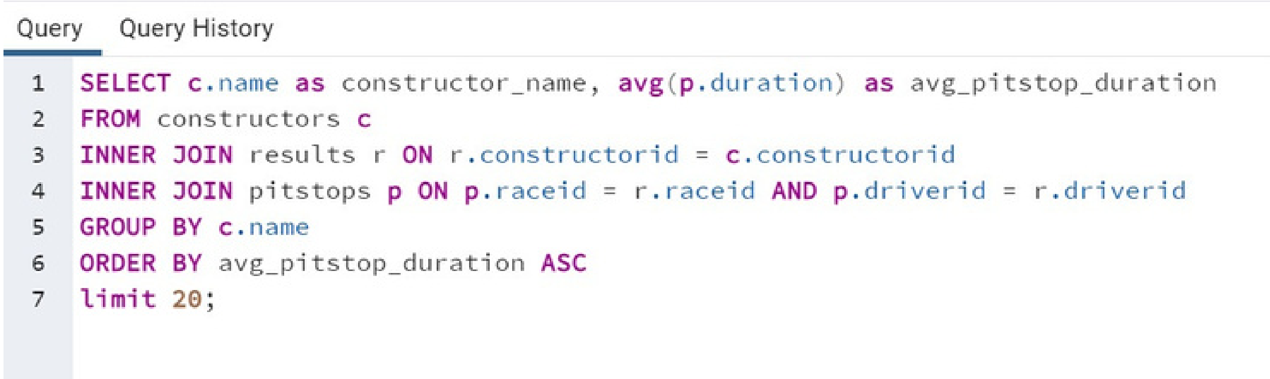
```sql
SELECT *
FROM pitstops
WHERE duration = (SELECT MIN(duration) FROM pitstops WHERE duration IS NOT NULL)
LIMIT 1;
```



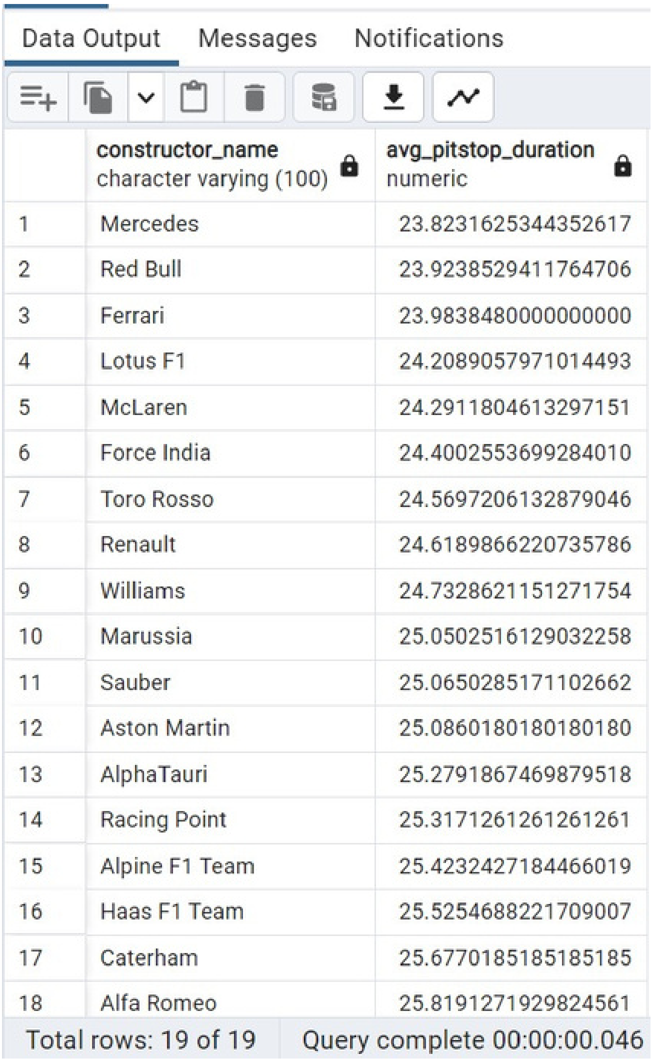
	raceid integer	driverid integer	stop integer	lap integer	p_time time without time zone	duration numeric (6,3)
1	870	813	3	52	15:23:48	13.206

Total rows: 1 of 1   Query complete 00:00:00.048

9. Query to retrieve the constructor details that has the best pitstop timing average in a race.



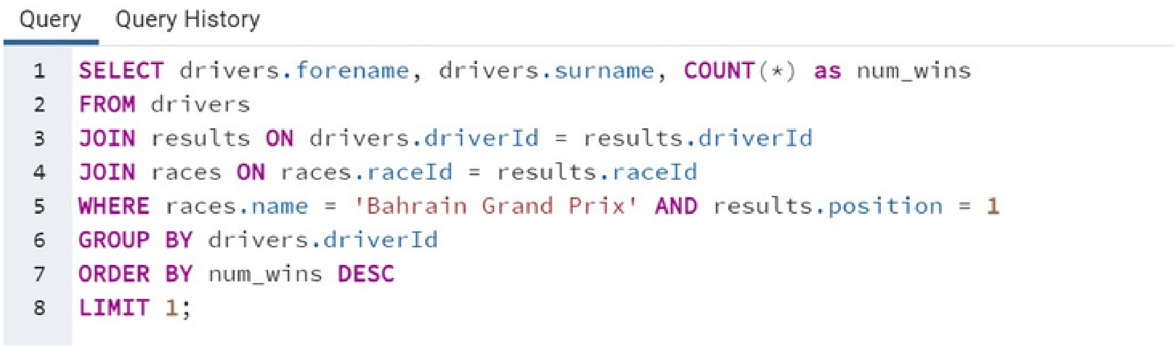
```sql
SELECT c.name as constructor_name, avg(p.duration) as avg_pitstop_duration
FROM constructors c
INNER JOIN results r ON r.constructorid = c.constructorid
INNER JOIN pitstops p ON p.raceid = r.raceid AND p.driverid = r.driverid
GROUP BY c.name
ORDER BY avg_pitstop_duration ASC
limit 20;
```



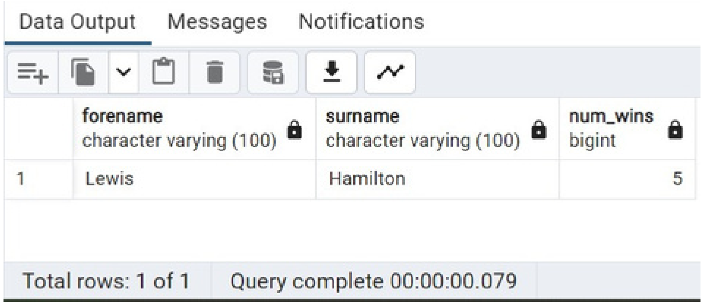
	constructor_name character varying (100)	avg_pitstop_duration numeric
1	Mercedes	23.8231625344352617
2	Red Bull	23.9238529411764706
3	Ferrari	23.9838480000000000
4	Lotus F1	24.2089057971014493
5	McLaren	24.2911804613297151
6	Force India	24.4002553699284010
7	Toro Rosso	24.5697206132879046
8	Renault	24.6189866220735786
9	Williams	24.7328621151271754
10	Marussia	25.0502516129032258
11	Sauber	25.0650285171102662
12	Aston Martin	25.0860180180180180
13	AlphaTauri	25.2791867469879518
14	Racing Point	25.3171261261261261
15	Alpine F1 Team	25.4232427184466019
16	Haas F1 Team	25.5254688221709007
17	Caterham	25.6770185185185185
18	Alfa Romeo	25.8191271929824561

Total rows: 19 of 19 Query complete 00:00:00.046

10. Query to retrieve the most number of titles the “Bahrain Grand Prix”.



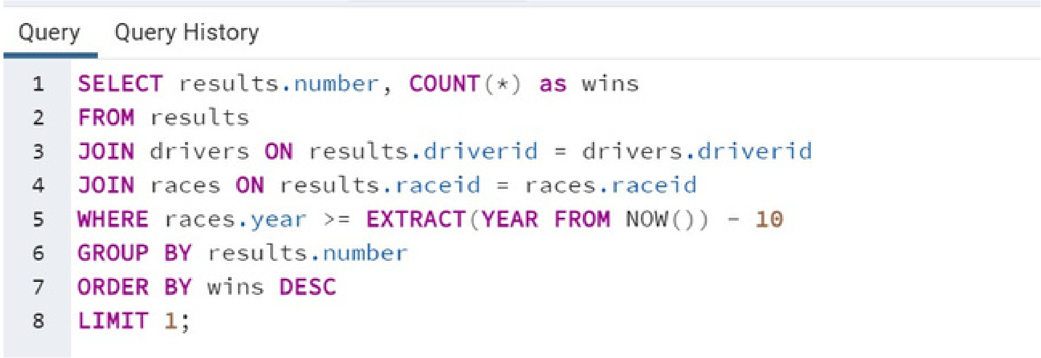
```sql
SELECT drivers.forename, drivers.surname, COUNT(*) as num_wins
FROM drivers
JOIN results ON drivers.driverId = results.driverId
JOIN races ON races.raceId = results.raceId
WHERE races.name = 'Bahrain Grand Prix' AND results.position = 1
GROUP BY drivers.driverId
ORDER BY num_wins DESC
LIMIT 1;
```



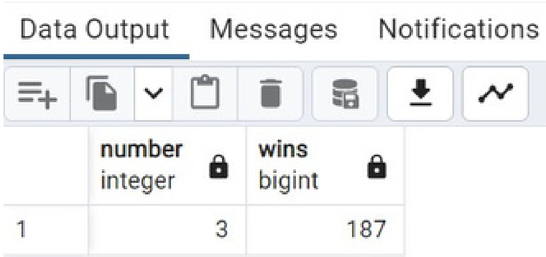
	forename character varying (100)	surname character varying (100)	num_wins bigint
1	Lewis	Hamilton	5

Total rows: 1 of 1   Query complete 00:00:00.079

11. Query to retrieve the driver details with the highest wins in the past 10 years



```
SELECT results.number, COUNT(*) as wins
FROM results
JOIN drivers ON results.driverid = drivers.driverid
JOIN races ON results.raceid = races.raceid
WHERE races.year >= EXTRACT(YEAR FROM NOW()) - 10
GROUP BY results.number
ORDER BY wins DESC
LIMIT 1;
```



	number integer	wins bigint
1	3	187

QUERY OPTIMIZATION:

To check the performance of the query we are using the below query which returns the raceId, driverId of drivers from races who made laps above the average time of laps made of all drivers in all races.

select distinct raceid, driverid
from laptimes
where milliseconds>(select avg(milliseconds) from laptimes)
group by raceid,driverid;

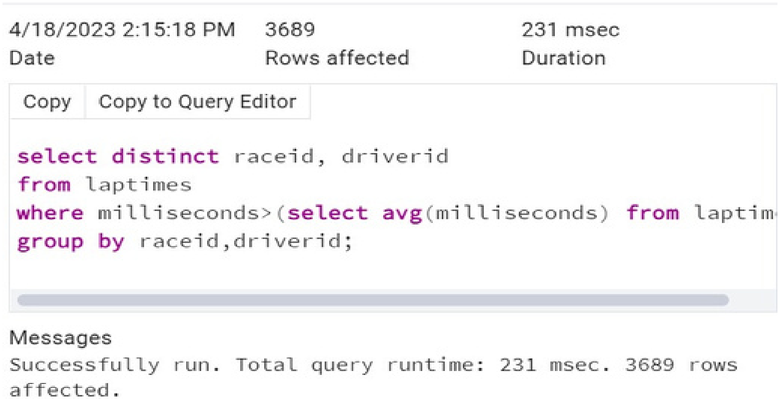
We first executed the query without any optimization, to check the execution time.
We use the concept of indexing for optimizing the performance.

INDEXING:

To improve the performance of the query we create an index column.

create index l_index on laptimes(milliseconds);

Performance before:



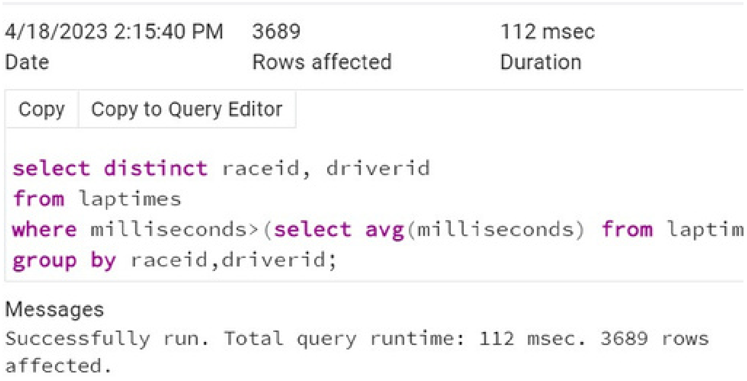
4/18/2023 2:15:18 PM 3689 231 msec
Date Rows affected Duration

```
select distinct raceid, driverid
from laptimes
where milliseconds>(select avg(milliseconds) from laptim
group by raceid,driverid;
```

Messages

```
Successfully run. Total query runtime: 231 msec. 3689 rows
affected.
```

Performance after:



4/18/2023 2:15:40 PM 3689 112 msec
Date Rows affected Duration

```
select distinct raceid, driverid
from laptimes
where milliseconds>(select avg(milliseconds) from laptim
group by raceid,driverid;
```

Messages

```
Successfully run. Total query runtime: 112 msec. 3689 rows
affected.
```

Hence, we achieve better execution time by indexing the table with l_index.

FINDINGS:

From all the queries executed, we got an insight to the important features of the data which are,

- The top most country that hosted most number of races in the past 10years.
- Some interesting insights based on drivers and constructors.
- Further,we found the insights based on circuits, which conducted highest number of races
- The insights about number of titles won in a particular race.