BIG DATA AND BUSINESS INTELLIGENCE MODULE

CIS4008-N-BF1-2022

Formula One World Championship Analysis

Module Leader : Dr Annalisa Occhipinti

Lab Tutor : Ms. Nissy Mathews

Submitted By : Ajay Kuruvilla Johnson (W9521662)

Submission Date: 11/01/2023

Contents

A.Executive Summary:

Formula One with its rich history is becoming a more global sport. The report focuses on how accidents and finishes vary over the years from (1950 – 2022).

We attempt to find certain characteristics and KPIs for a driver to be successful and using AI we try to forecast driver wins for the next 3 years.

We also try to find out what it takes to become a world champion and how different teams have been dominant in different eras of F1.

Findings and Conclusions

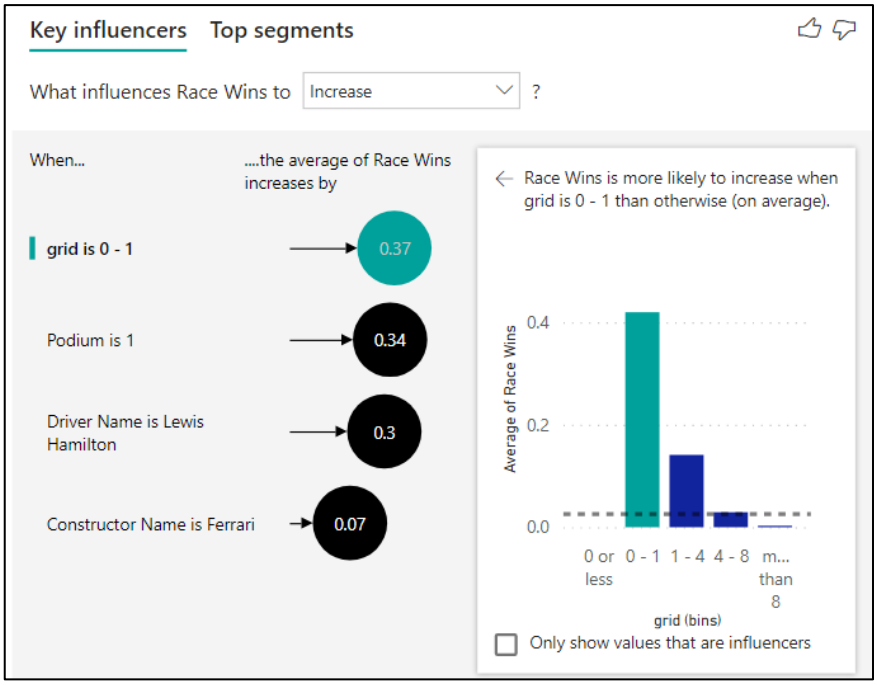
1. The most successful driver is Lewis Hamilton with 103 wins to his name and is a **seven time** world championship winner, tieing the record with Micheal Schumacher.
2. The most successful team is Ferrari because they have been in F1 since the beginning. Mercedes and RedBull are dominant in the Hybrid era.
3. The accident percentage has drastically reduced from 6.42% in 1950s it has reduced to 1.82% in 2022.
4. The finished percentage has drastically increased from 17.2% in 1950 to 63.41% in 2022.
5. In the past 10 years, Monaco has been the track with the highest accident percentage with 7.14% . On the other hand, Spa is the race track which has a high relative finish rate of 67.39%

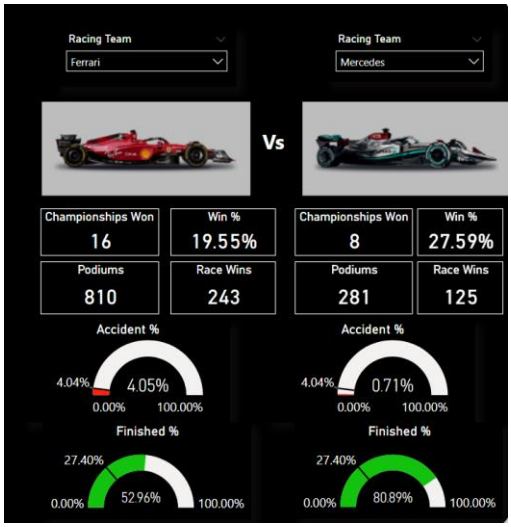
The graphs below gives a quick understanding of the type of analysis done in this report.

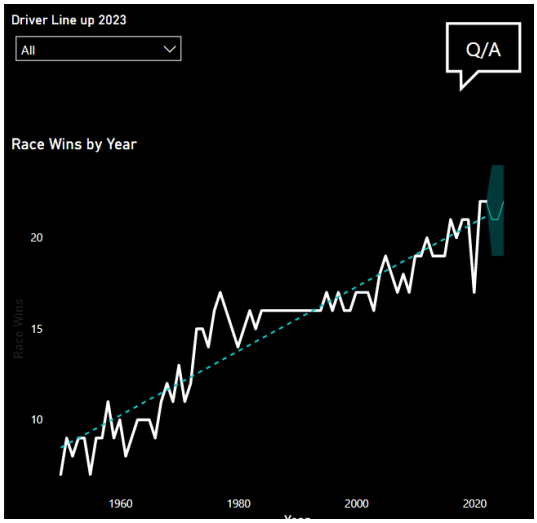
Due to time constraints I have only chosen the top 20 driver to analyse in detail. And the top 5 constructors

Apart from this we try to use AI to forecast driver wins for the year 2023

- **Some Graphs from the Main Dashboard**







Recommendations:

1. Teams with vast resources like Ferrari and McLaren are not able to perform at the top level. The number of (DNFs) – Did Not finish is higher due to car issues compared to Mercedes and Red Bull. The cars built should be reliable and hence top teams need to invest in the future to get their winning strategy right.
2. Winning Races is paramount to getting the prized Drivers and Constructors Championships and hence the main factors to look into that is to qualify in top 3 grid position in qualifying to give a racer maximum chances of winning on race day.
3. The Average age of a championship winning driver is 31.85 years. This means that the driver has been nurtured within a team and has gotten comfortable in order for him to give his best. Teams and senior management play a vital role in therefore recruiting bright talent and nurturing them so that they can become future world champions. This strategy has worked out well for red bull with Max Verstappen.

B.Body

Introduction:

Formula One is one of the biggest global sports and has a rich history originating in the 1950's. Winning races consistently leads to achieving the prestigious Driver's and Constructor's Championship titles. Hence racing teams are always on the lookout to understand and analyse factors which result in race wins. Using this dashboard, we can dive deep into some metrics which makes drivers win and also reasons for not winning as well.

Formula One is considered to be one of the dangerous sports with drivers at times having fatal consequences. We will explore the trend on accident rates and how this could help drivers be more efficient and consistent in their performance.

The steps of data pre-processing, data modelling and data visualisation with some emphasis on AI and analytics would have been showcased in this dashboard.

- ## Data Source :

  The dataset is taken from Kaggle from the link : Formula 1 World Championship (1950 - 2022) | Kaggle.

  The Championship Data was taken from Wikipedia : List of Formula One World Drivers' Champions - Wikipedia and List of Formula One World Constructors' Champions - Wikipedia

  The dataset from kaggle has **14** tables namely:

  1. **Circuits**: circuitId,circuitRef,name,location,country,lat,lng,alt,url.
  2. **Constructor_results:** constructorResultsId,raceId,constructorId,points,status.
  3. **Constructor_standings:** constructorStandingsId,raceId,constructorId,points,position,positionText, wins.
  4. **Constructors:** constructorId,constructorRef,name,nationality,url.
  5. **Driver_standings:** driverStandingsId,raceId,driverId,points,position,positionText,wins.
  6. **Drivers:** driverId,driverRef,number,code,forename,surname,dob,nationality,url.
  7. **Lap Times** : raceId,driverId,lap,position,time,milliseconds.
  8. **Pit stops** : raceId,driverId,stop,lap,time,duration,milliseconds.
  9. **Qualifying :** qualifyId,raceId,driverId,constructorId,number,position,q1,q2,q3
  10. **Races :** raceId,year,round,circuitId,name,date,time,url,fp1_date,fp1_time,fp2_date,fp2_time,fp3_date,fp3_time,quali_date,quali_time,sprint_date,sprint_time

11. **Results :**
    resultId,raceId,driverId,constructorId,number,grid,position,positionText,positionOrder,points,laps,time,milliseconds,fastestLap,rank,fastestLapTime,fastestLapSpeed,statusId
12. **Seasons:** year,url
13. **Sprint results:**
    resultId,raceId,driverId,constructorId,number,grid,position,positionText,positionOrder,points,laps,time,milliseconds,fastestLap, fastestLapTime , statusid.

14 **status:** statusid, status

The extra Championships data is taken from Wikipedia and has 2 tables:

1. Driver Championships
2. Constructor Championships

- ## The reason for choosing this dataset:

I am an ardent fan of Formula One and I know the sport quite well since I watch the races regularly. The possibility of finding some exciting insights in the data was quite motivating. Hence, I was excited on the opportunity to do an in depth analysis in PowerBI with this dataset.

- ## Will this dataset help in developing specific business skills?

I believe by doing this Big Data analysis, I got to understand how to analyse a huge dataset and how to break it down into different sections. The importance of pre-processing and data modelling became much more apparent to me.

The skills I have gained such as looking for KPI's and Performance metrics is of high importance and can be applied to my further career aspiration as a business analyst or data analyst.

- ## Business Questions/Problems:

This is a very important part of my project. The business questions I have taken are given below:

1. Have race accidents decreased over the years (1950 -2022)?
2. How has the finished % of drivers changed over the years (1950 -2022)?
3. Does Qualifying P1 (Position 1) have any impact on winning on race day?
4. Which racetrack has the highest risk of accidents on average and average finishes?
5. Can we forecast the number of race wins for drivers in 2023?
6. Which driver has been the most successful and is there any defining reason why he has been so successful?
7. Among the constructors, which teams are performing the best in the recent turbo era (2014 -2021)?
8. The Formula One sport was initially a European sport, how has that changed in recent years?
9. What is the average age for winning a driver's championship?
10. Which driver and team have the best winning percentage?

These questions are answered in the key findings section

- ### Key User Groups :

This dashboard is mainly intended for senior management and team principles who can use the historical data of drivers and teams and run useful analysis on every performance metric.

- ### Specific Features and it's need:

The main features I will be looking at is the safety of cars and how accidents have reduced or increased over time and what are the reasons for the same. This has massive applications as F1 has seen some fatal accidents in the past and making the more safe and fun is the objective.

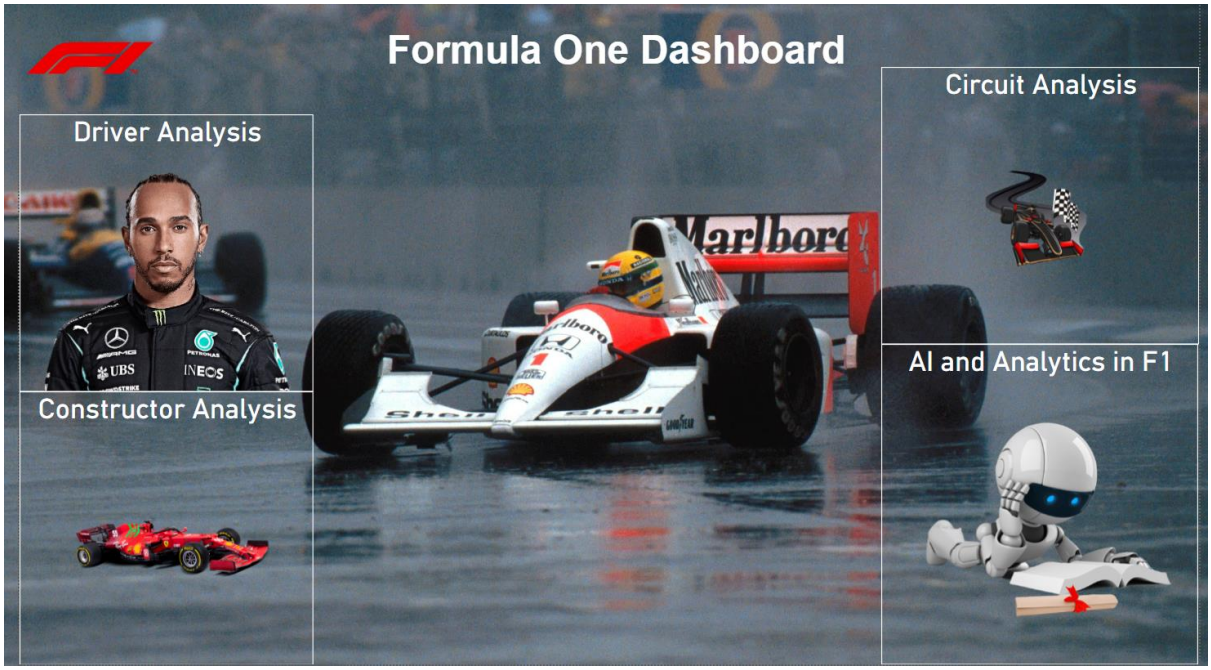
- ### Most Important KPI

The main KPI would be **race wins** and **win%.** As a racing driver, the sport is extremely competitive and the goal is to secure a championship. That's where the money and fame lies. To get the championship one has to win and get as many points as possible in a given season.

C. Finding based on analysis and evaluation

This is the most important section of the project and I have explained with screenshots of the various visuals used in my dashboard.

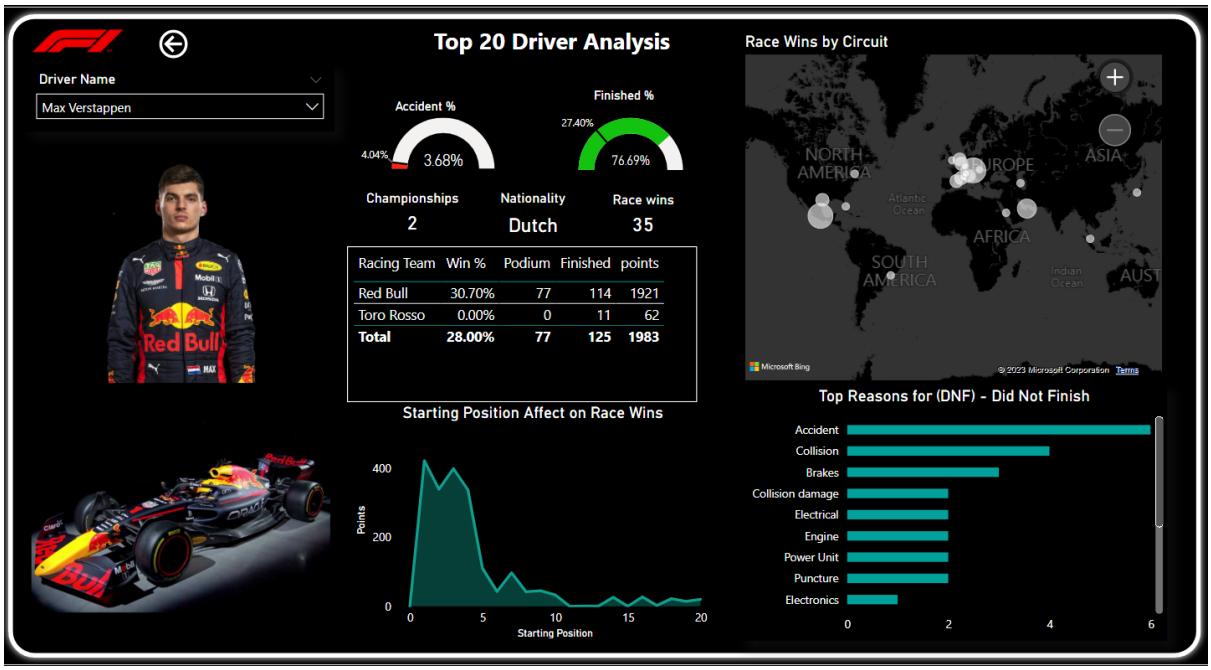
- **HOME Page:**



- This is a home page which use different buttons for navigation.
- A background image of F1 legend ,'Ayrton Senna' is kept as an homage to his legacy.
- Each card is a navigation button with a background image given to make it visually pleasing. The same is applied for the other buttons as well.

Navigation Button

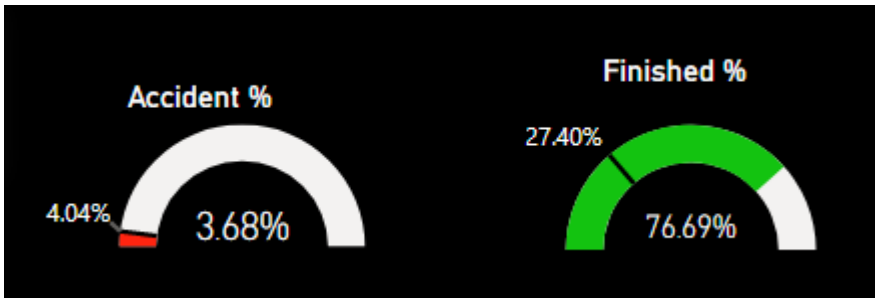
- **DRIVER Analysis:**



Overview

- Used here is the ‘Image by CloudScope’ visual.
- Using this we can make a dynamic image change according to the selection in the slicer.
- As the slicer selects another driver the image and their car will change dynamically.
- I had to get the image URL which I have put in a separate table called ‘Driver and Car Images’ and then made a 1:1 data modelling relationship to get this visual working.
- Only the top 20 Drivers were selected as it woud have been a herculean task to get the images for all 848 drivers!

- **Gauge Visual:**



This is nice visual to show the threshold value, the minimum and maximum value.

My main business question and finding is based on the accident and finished % of drivers, constructors and the circuits.

DAX measures have been used to create this visual:

```
1 Accidents =
2         IF(Results[Status] = "Accident", 1, 0)
```

aceId	grid	position	points	Status	Race Wins	Accidents

```
1 Finished = if(Results[Status] = "Finished", 1 , 0)
```

aceId	grid	position	points	Status	Race Wins	Accidents	Finished
29	4	0	0	Engine	0	0	0

Key Findings:

1. Able to calculate Accident % and finished % of Drivers.

- **Table visual:**

Racing Team	Win %	Podium	Finished	points
Red Bull	30.70%	77	114	1921
Toro Rosso	0.00%	0	11	62
Total	**28.00%**	**77**	**125**	**1983**

This table is chosen as it easy to show multiple multiple columns In an elegant without me having to add a lot of cards to the page.

Key Findings:

1. Able to see the Win % of Drivers with different teams.

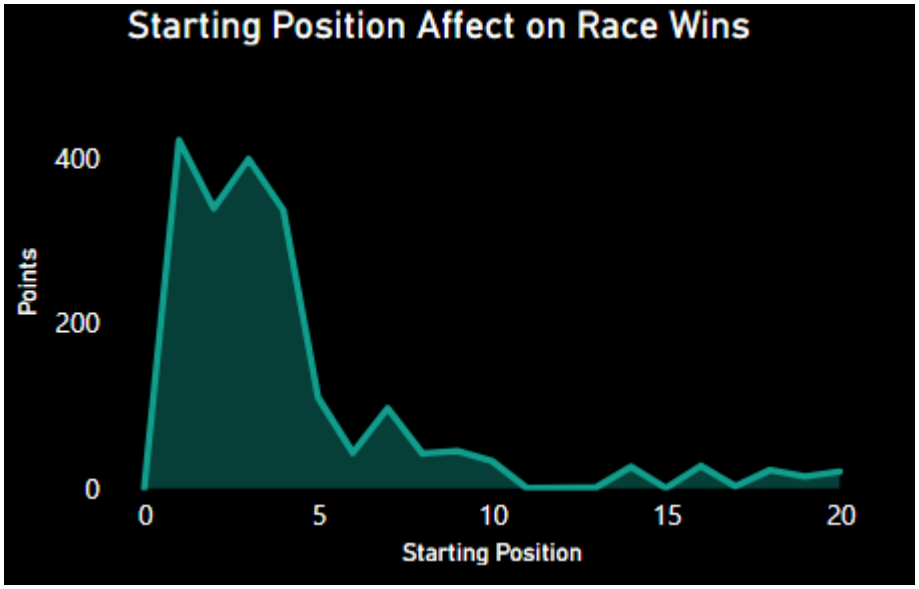
The Win% , Podium and Finished Columns use DAX Measures and Columns.

```
Win % = (sum(Results[Race Wins])/sum(Results[Finished]))
```

```
Podium = IF
        (OR(if(Results[position] = 1,1,
            IF(Results[position] = 2,1,
                IF(Results[position]=3,1,0))),0),1,0)
```

ceId	grid	position	points	Status	Race Wins	Accidents	Finished	Podium
29	4	0	0	Engine	0	0	0	0

- **Area Chart visual:**

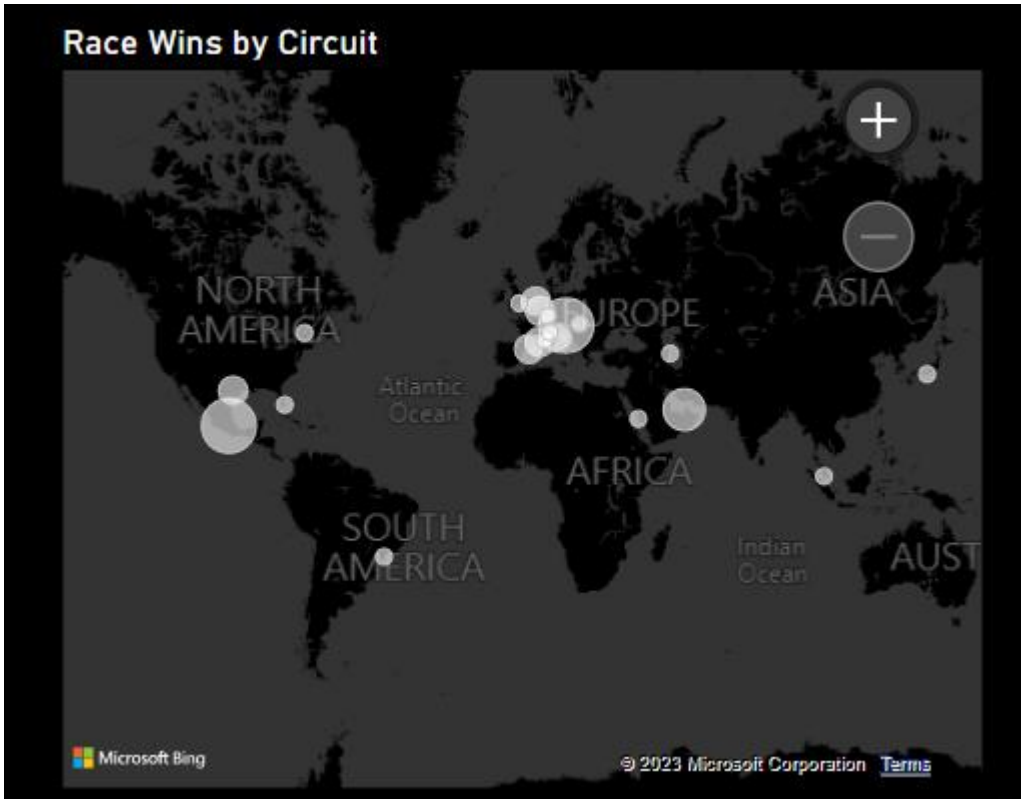


Having the y-axis as the points column from results and in the x-axis we have the 'grid' column from results table. This chart is chosen as it is easy to interpret and visually pleasing to look at.

Key Findings:

1. We see from the visual that the starting position of 1 , 2 and 3 have the most chances of winning maximum points and this significantly reduces if the starting position is 5th and above.
2. The chart will be different for different racers. Some racers prefer to lead while others perform well during a chase or race battle.

- **Map Chart:**



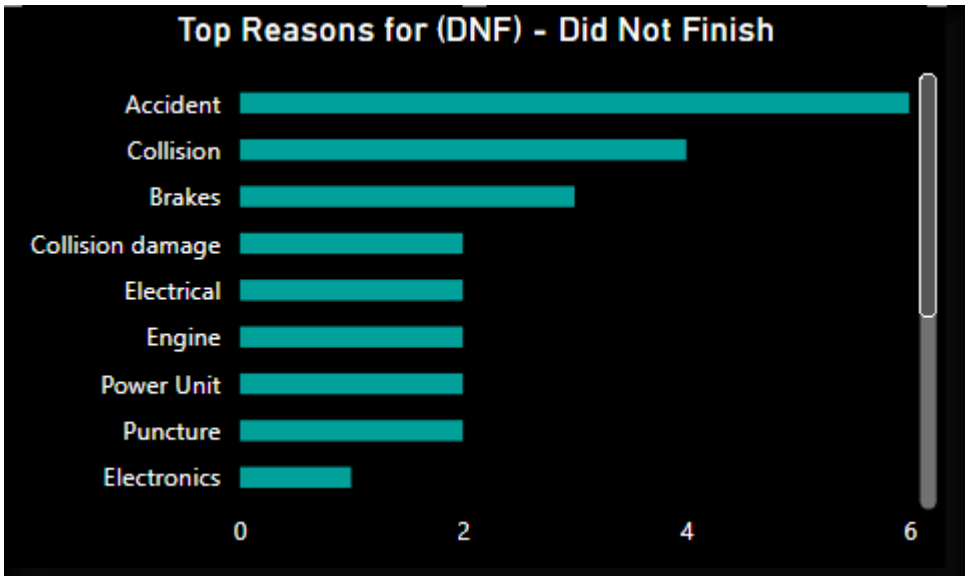
This chart is taken specifically to showcase clearly which **circuits** are a driver's strong point. The size of the **bubble** denotes the number of race wins by a driver. The map is interactive and moves accordingly to the driver selected.

A gray scale was chosen with white bubbles was chosen to give a contrast to the grayish background.

Key Findings:

1. Most racers are dominant in their home tracks but the drivers who are successful in historically proven tracks such as Monza and Silverstone are generally classified as elite racers.

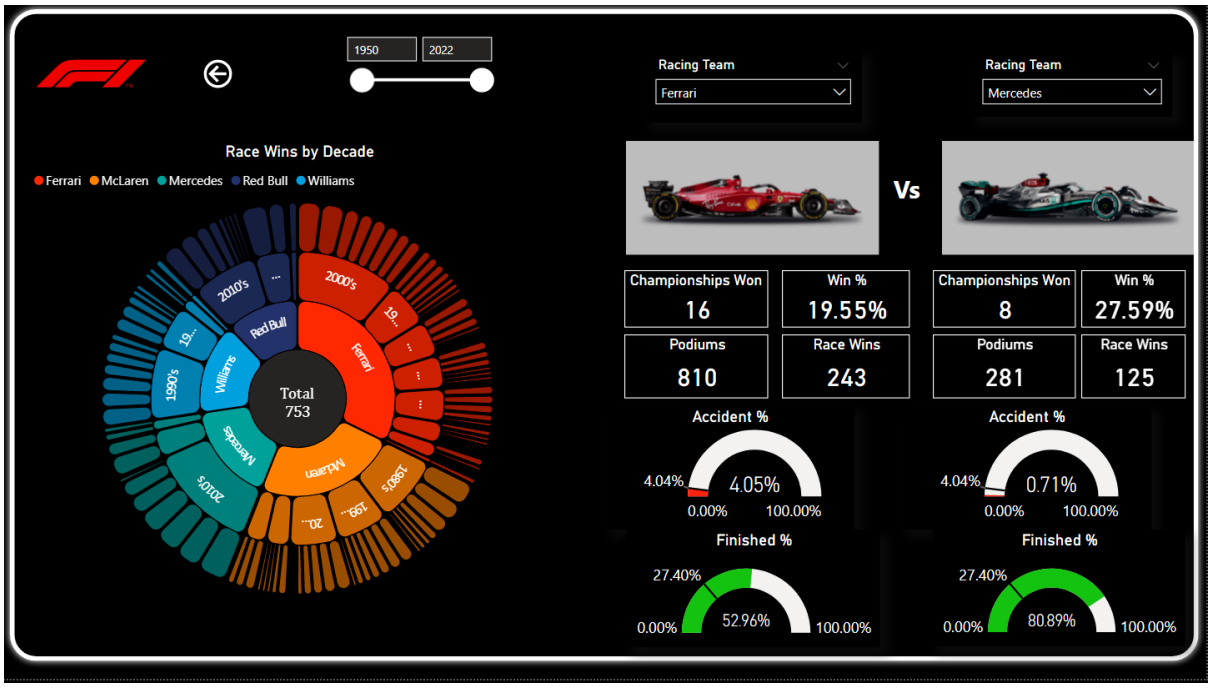
- **Horizontal Bar Chart:** This is a standard bar graph. It is chosen specifically to understand which are main reasons for Driver (DNF) – Did not finish.



Key Findings:

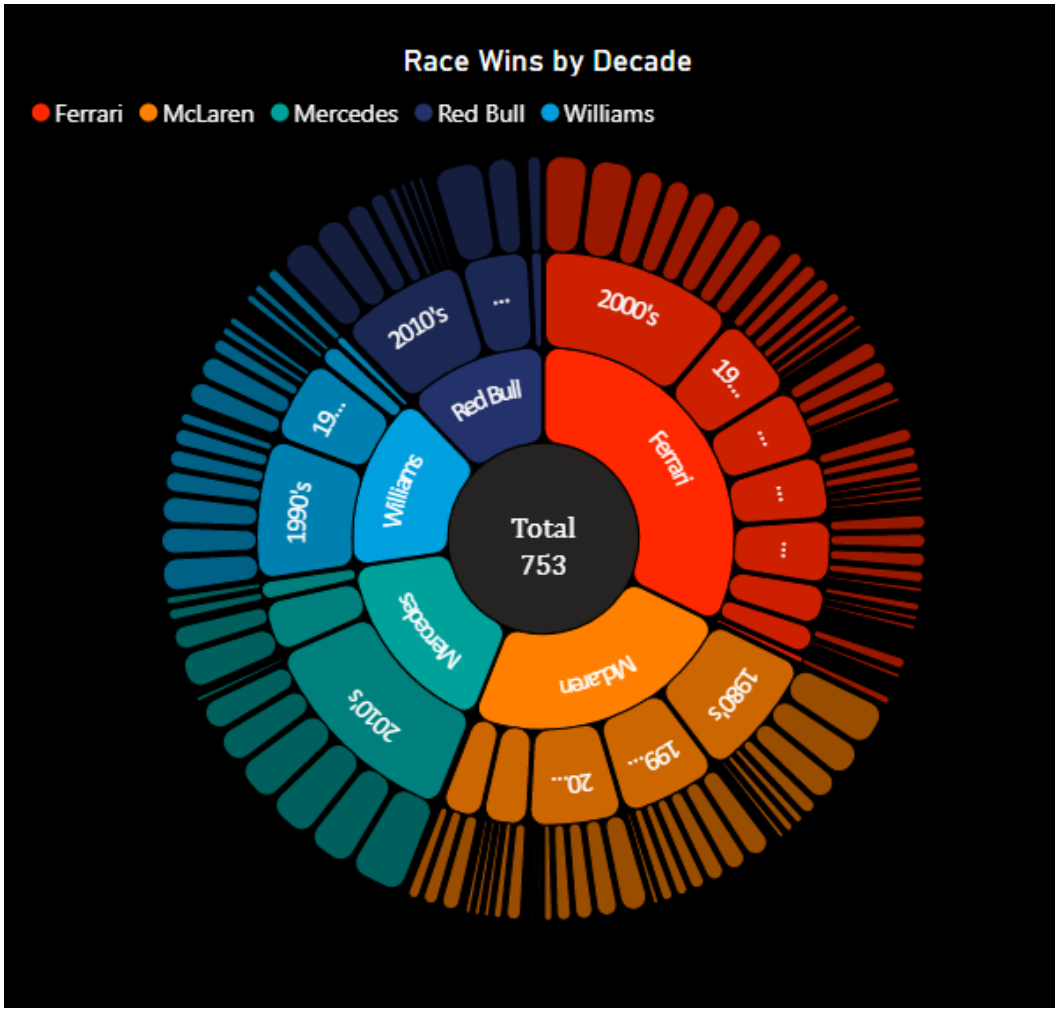
1. This is an important visual as we can see have an insight into their driving style and also how efficient their cars are.
2. When the reason is ‘Engine’ or ‘Electrical’ failure at the top, we can understand the car is not fully reliable.
3. On Driver traits, ‘Lewis Hamilton’ has very few (DNF’s) and are mostly due to collision by another racer. Hence it can be said he drives much more safely than for example, ‘Max Verstappen’ who is more of an aggressive driver by comparison.

- **CONSTRUCTOR Analysis (Page 2):**



Overview

- **The Zoomable Sunburst chart: (Not Covered In Lesson)**



This is a zoomable sunburst chart and this is the external chart which I have chosen. I have chosen this particular graph as I wanted to display the dominance of certain teams over the 72 years of F1 history.

It has two levels and it shows the race wins by Constructors over the decades.

Key Findings:

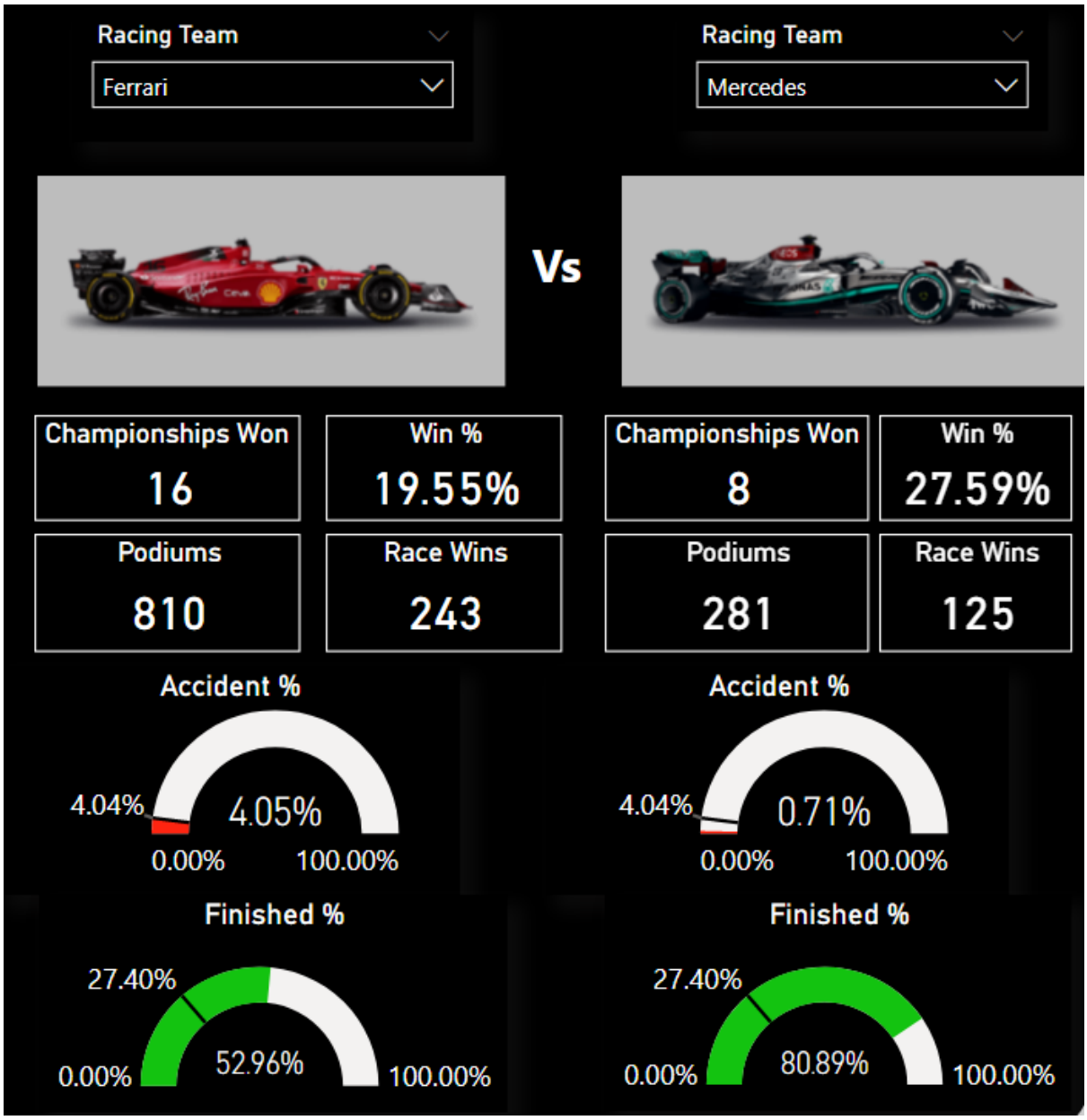
1. From this visual we can see that in the Hybrid Era (2014-2021) Mercedes have been completely dominant with Lewis Hamilton breaking records.
2. The years between (2010 – 2013) belonged to Red Bull.
3. And the years between (2000 – 2004) and some would argue till 2008 was dominated by 'Ferrari'.

Hence this insight can be displayed nicely using the sunburst chart.

For this a DAX Column **'Decade'** has been used to create this visual.

```
Decade = IF(Results[Year] < 1960, "1950's",
            IF(AND(Results[Year] < 1970, Results[Year] >= 1960), "1960's",
               IF(AND(Results[Year] < 1980, Results[Year] >= 1970),"1970's",
                  IF(AND(Results[Year] < 1990, Results[Year] >= 1980), "1980's",
                     IF(AND(Results[Year] < 2000, Results[Year] >= 1990), "1990's",
                        IF(AND(Results[Year] < 2010, Results[Year] >= 2000), "2000's",
                           IF(AND(Results[Year] < 2020, Results[Year] >= 2010), "2010's",
                             IF(AND(Results[Year] < 2030, Results[Year] >= 2020), "2020's", "Unknown")))))))
```

- **Comparison using Cards, Gauge axis and and Image visual:**



This is used to compare the stats and figures of different teams. (Only the Top 5 teams are used).

So this is designed to see which teams dominated in which year. All of this can be controlled by a year slicer to show data of particular years.

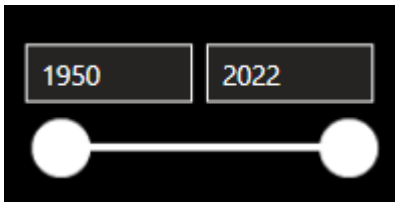
This visual using multiple visuals to portray a story. This is particularly useful to understand which teams are more safe and more accident prone. Which team has a better winning %.

Key Findings:

1. We see that Mercedes in recent year have a very high Win % and their finished Races % is much higher than any other which means their Accident % is also extremely low.
2. Red Bull and Mercedes in the past 10 years have been more successful than all the other teams.
3. Ferrari has fallen short in recent years and their last drivers championship was in 2008.

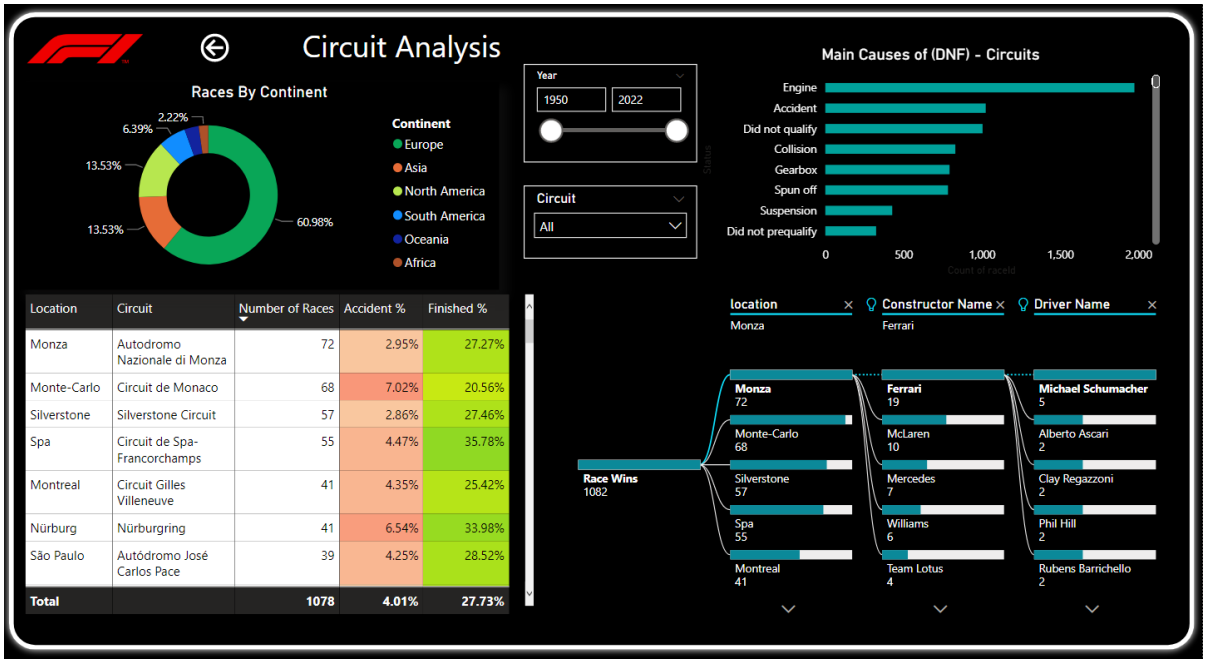
Images of the cars and the URL was used to put in a table called 'Driver and Car Images' and a 1:1 relationship is done with Drivers Data and Constructors data.

- Slicer



Standard Slicer with the seasons (Race Year) are given so as to drilldown on a particular season or maybe an certain time period or decade.

- CIRCUIT ANALYSIS: (Page 3)



Overview

- **Table Visual with Conditional Formatting:**

This visual is best to display multiple columns of data side by side for several rows. The conditional formatting is done so that we can sort according to their descending or ascending values.

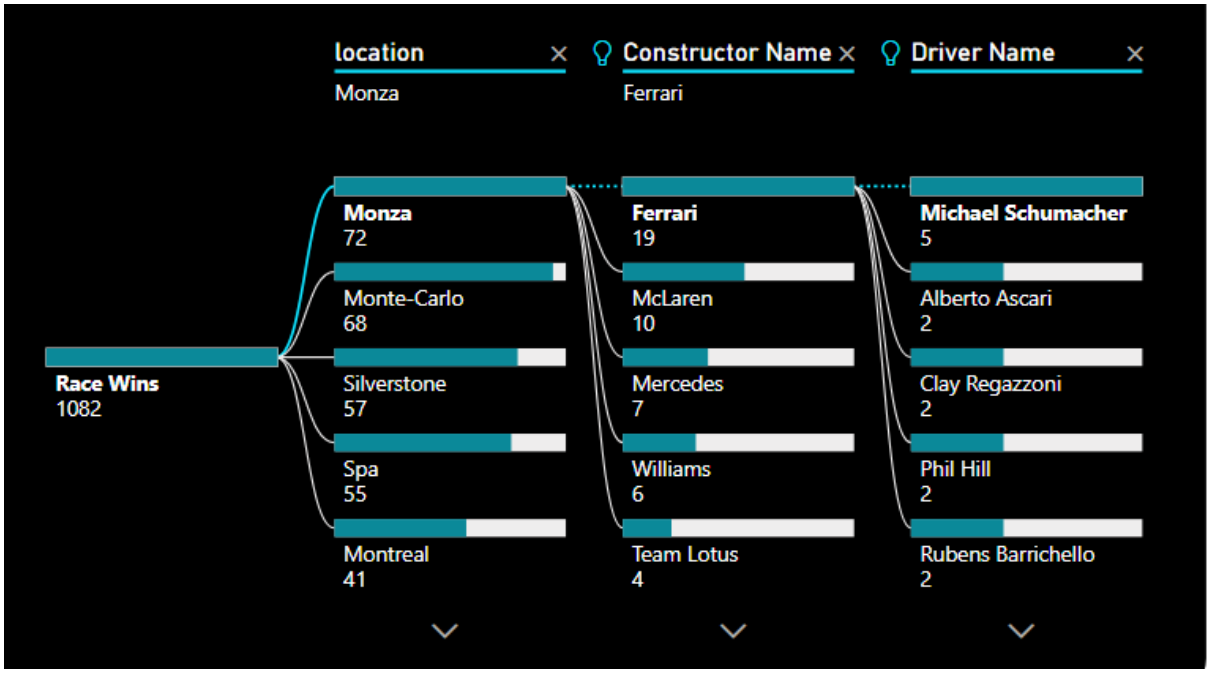
Location	Circuit	Number of Races	Accident %	Finished %
Casablanca	Ain Diab	1	20.00%	20.00%
Pescara	Pescara Circuit	1	12.50%	25.00%
Heusden-Zolder	Zolder	10	12.41%	15.96%
California	Long Beach	8	12.27%	15.91%
Barcelona	Montjuïc	4	11.90%	13.10%
Indianapolis	Indianapolis Motor Speedway	19	11.20%	31.66%
Detroit	Detroit Street Circuit	7	10.47%	15.18%
Styria	Zeltweg	1	10.00%	10.00%
Total		**1078**	**4.01%**	**27.73%**

Since accident % and finished % were quite important to me, this visual made it super easy and makes it easy to understand for the viewer as well to understand the circuits with high accident rates and finished rates.

Key Findings:

1. If we take data from the past 10 years, we see that the most dangerous track is monaco's circuit with a Accident % of 7.26%
2. The track with most average finishes is Spa with 66.14%.

- **Decomposition tree visual:**



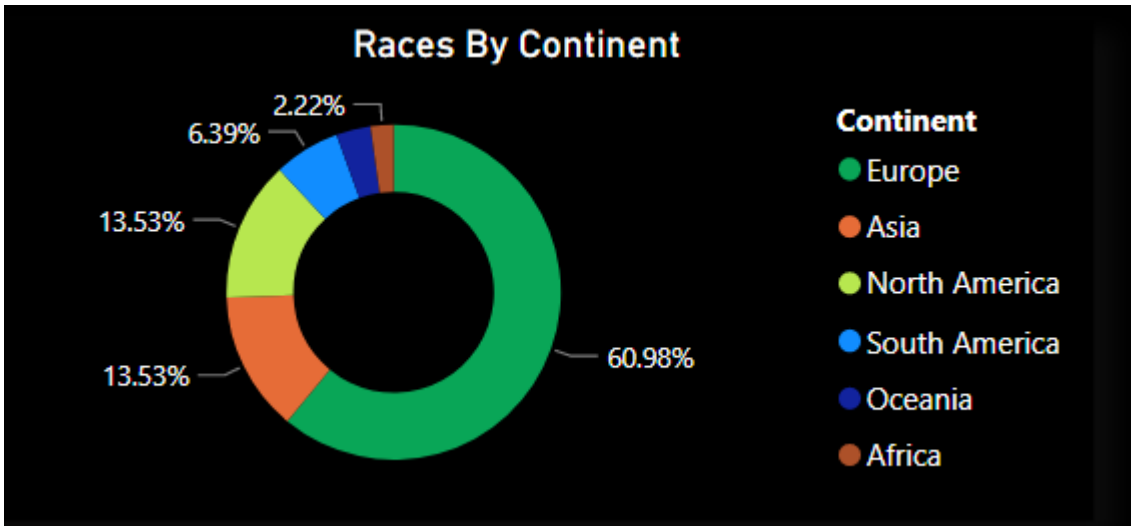
This visual has been taken as it displays relative value for the Constructors and Drivers who have been most successful in a certain circuit. The 'bulb' next to the Constructor and driver Name heading signifies that using AI 'high value' or 'Low value', power BI automatically detects the different input and gives an output, which, in this case is of high value.

This visual is also helpful to see how other drivers or constructors compare to the top value.

Key Findings:

1. Silverstone which is the track in Great Britain has most wins by Lewis Hamilton. And the UK is the home base for Mercedes.
2. Monza has Michael Schumacher in race wins which is the home track of Ferrari.

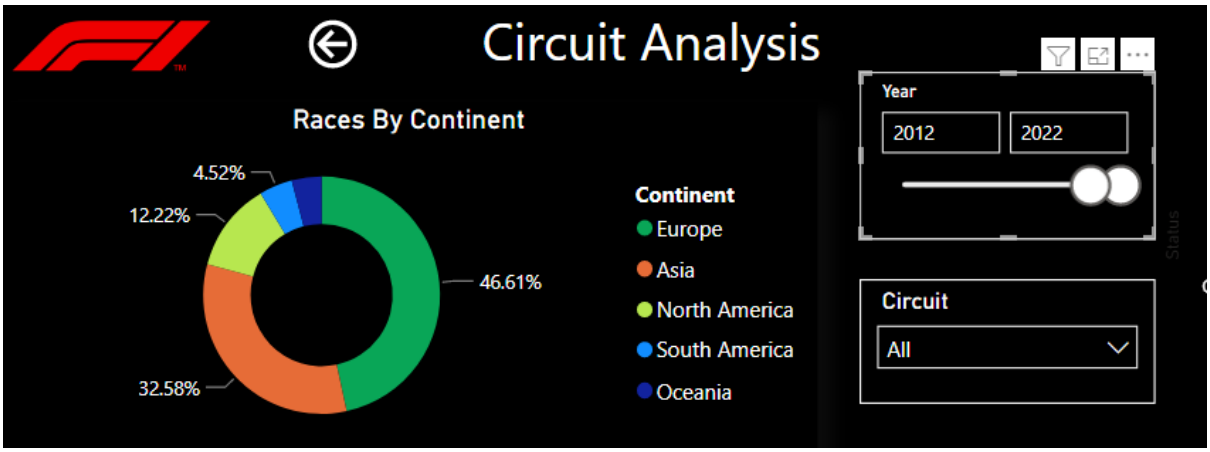
- **Pie chart:**



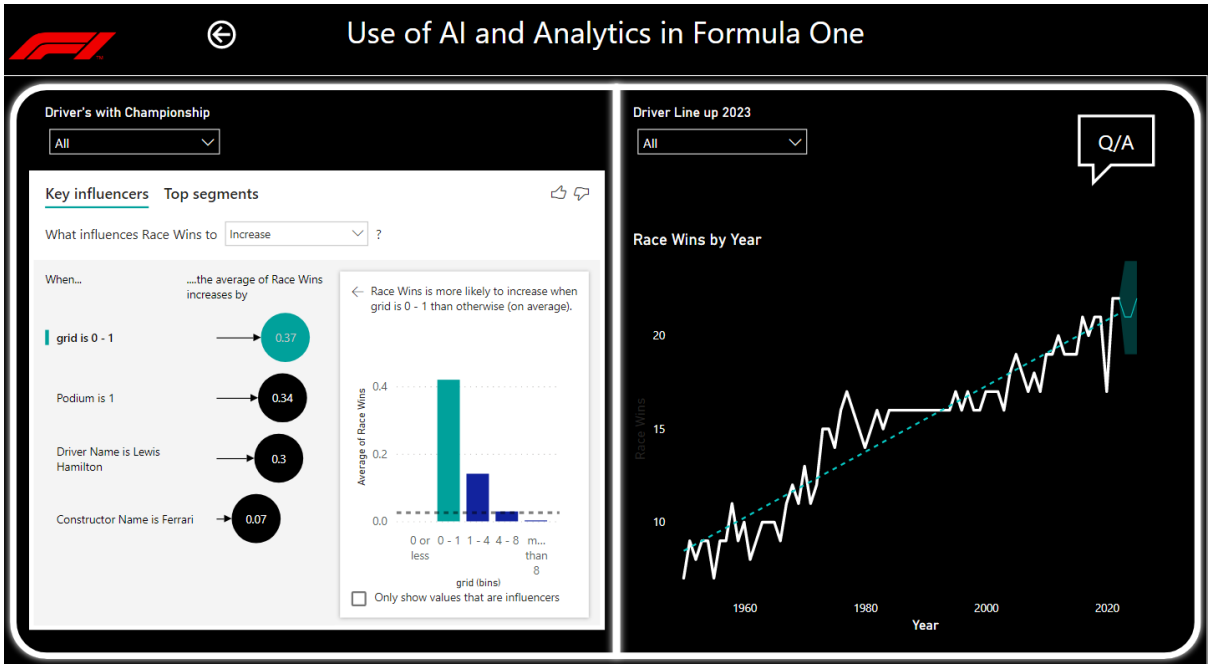
This is a standard pie chart visualised that gives an idea of which continent hosts more of the F1 races. This chart helped me answer the question of how in recent years, the sport is become much more global.

Key Findings:

1. In past 10 years there has been more races happening in the Asian continent and some in North America making F1 truly a global sport.



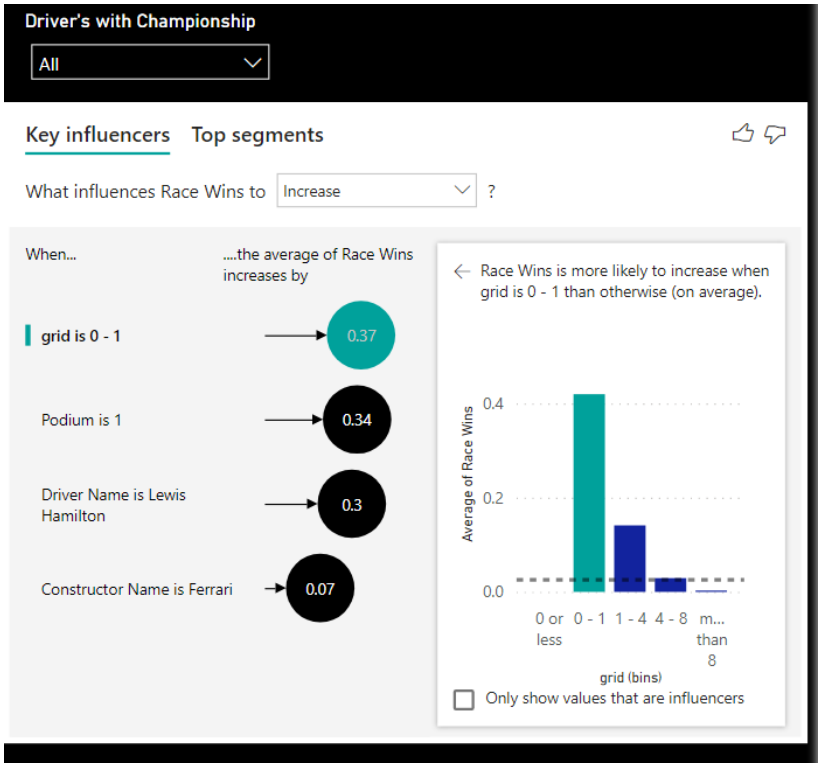
- **Use of AI and Analytics in Formula One (Page 4)**



Overview

I have kept this page mainly only for charts which have AI being used.

- **KPI Influencers Visual:**

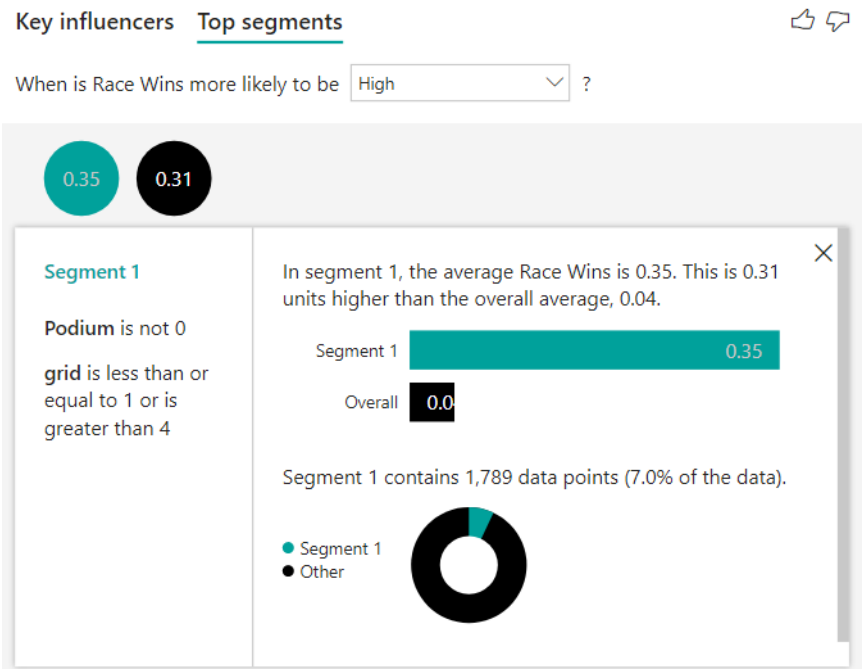


This visual helps us understand what are the factors that can help increase Race wins.

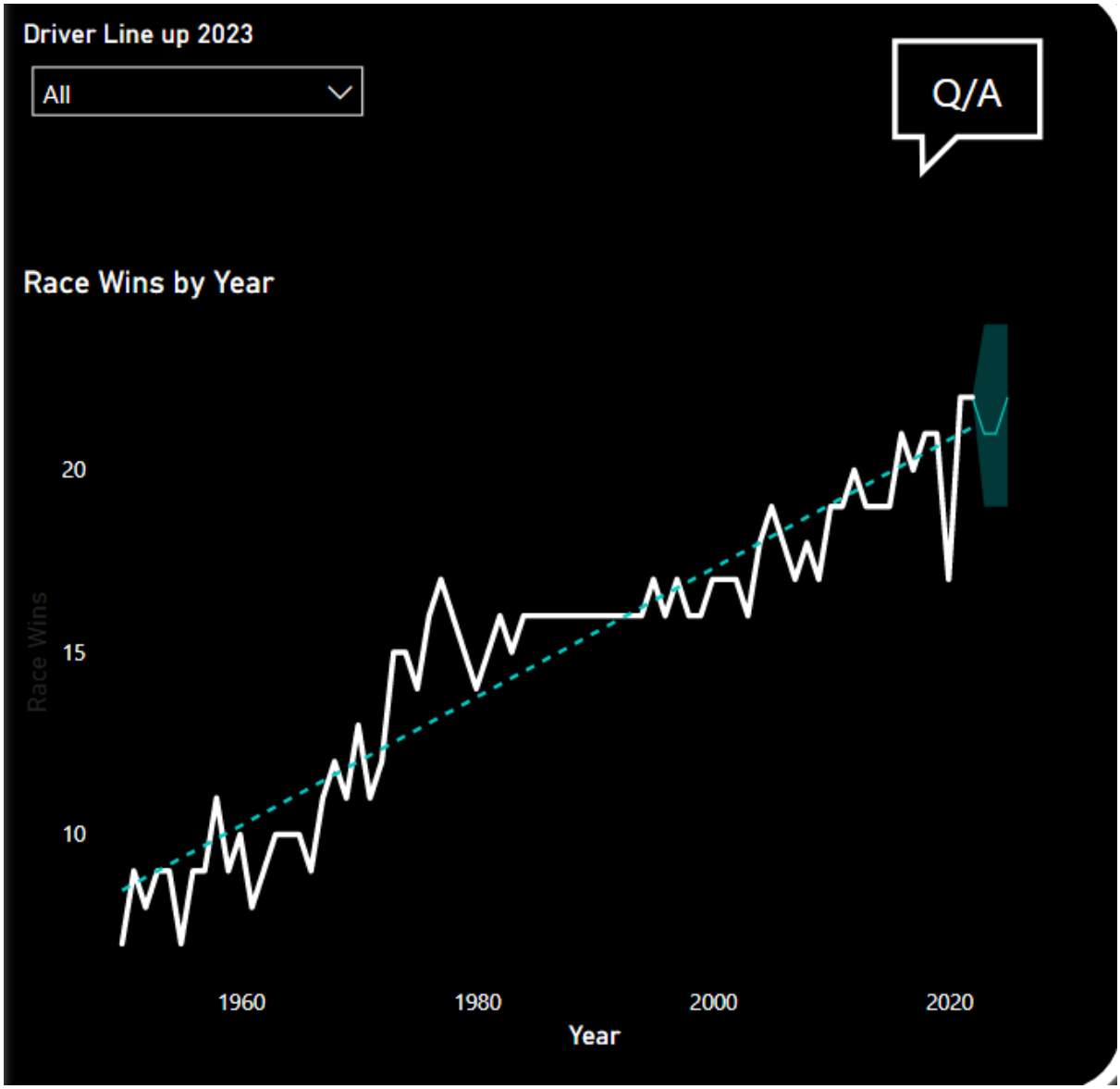
Key Findings:

1. If the starting grid position is 1, the chances of winning are much higher. This is especially true for circuits like Monaco.
2. The chances of winning a championship is by accumulating most amount of points. And the way to do that is to get is by either winning the race or having a podium (1,2,3 finish)

This visual also gives us Top Segments which are very helpful for analysis.



- **Line Chart with Forecasting:**



The aread at the end of the line chart is the forecasted part. I have forecasted race wins for 3 more from 2022. It is fun to have a look into the future, although this not very accurate, it gives a finding based on the previous historical data.

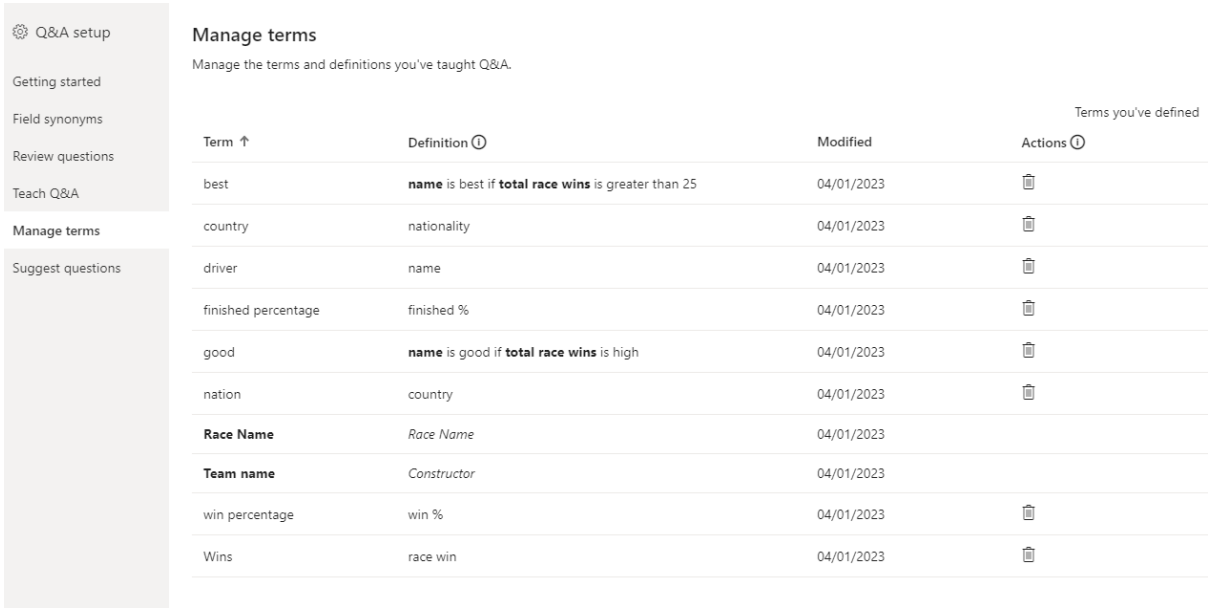
Key Findings:

1. Lewis Hamilton is forecasted to win only 2 races in 2023.

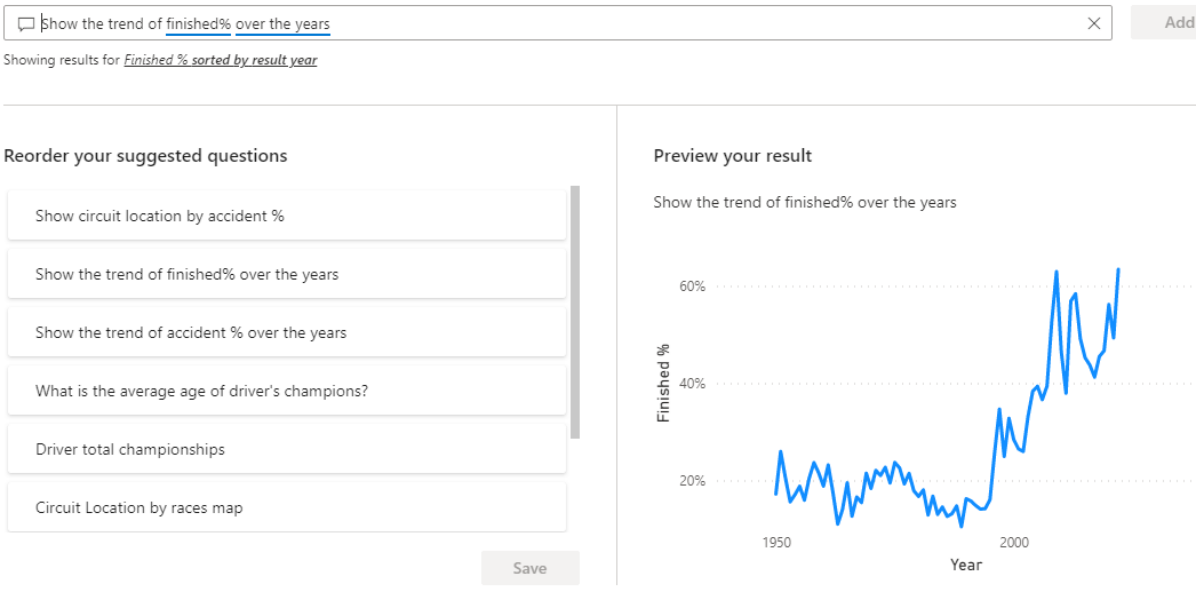
- **The Q/A Visual:**

I have kept the Q/A as a button to conserve space.

This is a really cool visual. I have taught the AI some Q/A and it have come up with some suggested graphs.

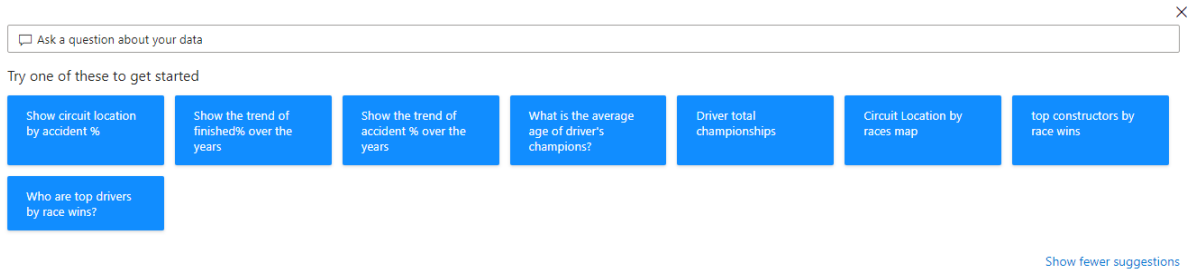


Teaching some terms to AI



Suggested and trained questions

Choosing the right columns to be displayed so that the graphs are displayed accordingly

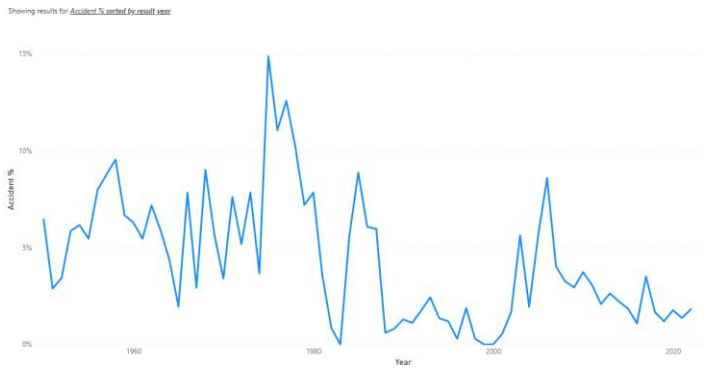


Suggested and trained questions

D. Conclusions with Business Questions and Answers

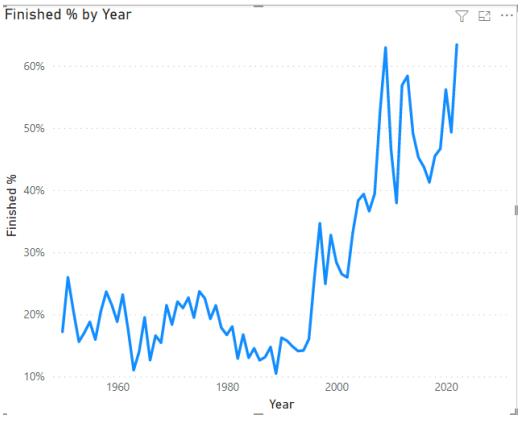
1. Have race accidents decreased over the years (1950 -2022)?

   Answer : Yes they hav, I was able to find this from the Q/A graph. From 6.42% in 1950s it has reduced to 1.82% in 2022.



2. How has the finished % of drivers changed over the years (1950 -2022)?

   From 17.2% in 1950 to 63.41% in 2022.



3. Does Qualifying P1 (Position 1) have any impact on winning on race day?

   Yes qualifying P1 puts a significant advantage for the driver to lead the race.

4. Which racetrack has the highest risk of accidents on average and average finishes?

   In the past 10 years:

   Highest Accident % - Monaco

   Highest Finish % - Spa

5. Can we forecast the number of race wins for drivers in 2023?

   Yes, Lewis Hamilton – 2 wins

   Max Verstappen – 14 wins

6. Which driver has been the most successful and is there any defining reason why he has been so successful?

   Lewis Hamilton as per the KPI Influencer Chart. His accident % over his career is 0.32% and finished % is 87.10%. Main reasons of DNF are Collision and Gearbox issues which shows how good his driving calibre is and only if a car issues comes he does not finish.
7. Among the constructors, which teams are performing the best in the recent turbo era (2014 -2021)?

   Mercedes with 8 Consecutive Constructors' Championships.
8. The Formula One sport was initially a European sport, how has that changed in recent years?

   In the past 10 years, Asia hosts 32,8% of the races compared to only 13.53% in the past.
9. What is the average age for winning a driver's championship?

   31.85 years
10. Which driver and team have the best winning percentage?

    Mercedes with 27.59% and Hamilton with 38.15%

Recommendations

1. Ferrari and McLaren, two teams with vast resources, are unable to compete at the highest level. Compared to Mercedes and Red Bull, there are more drivers who (DNF) - Did Not Finish - the race because of car problems. The cars constructed should be dependable, thus top teams must invest in the future to develop a winning strategy.
2. The major consideration is to qualify in the top 3 grid positions in qualifying to provide a racer the best chance of victory on race day. Winning races is essential to obtaining the coveted Drivers and Constructors Championships.
3. A championship-winning driver is 31.85 years old on average. This indicates that the driver has grown up in a team environment and is now at ease enough to give it his all. Teams and senior management are crucial in finding and developing talented individuals who have the potential to become future world champions. Red Bull has had success with this tactic while using Max Verstappen.
4. The circuit at Monaco has the highest average risk of having accidents or collisions at 7.14% in the past 10 years. This is due to street style circuit.

Although this has been the circuit layout for more than 50 years a change in the track should be considered as there is always some sort of collision bound to happen due to its narrow roads. Safety of the drivers is the main priority and other newer circuits such as the Red Bull ring, Baharain International Circuit and the Abu Dhabi Yas Marina circuit have accident percentages namely 0%, 0.8% and 0.88% respectively. Newer safety principles are considered while building these race tracks.

5. It would be nice to see more races happening in North America and Asia and other parts of the world. Asia is slowly climbing up in the percentages of races hosted and the same would be nice to see other parts of the world.
6. Drivers need to mindful of rash and aggressive racing because as per the data, the chances of not finishing the race is quite high and you need to be in the race to get points for the Drivers Championships. Taking the example of Lewis Hamilton: He has one of the lowest accident ratios and has a the highest finished percentages. He is the most successful driver in F1 after Michael Schumacher.

SECTION 2 : APPENDIX

ICA – Appendix: BI Design

A. Data Pre-Processing or Data Cleansing:

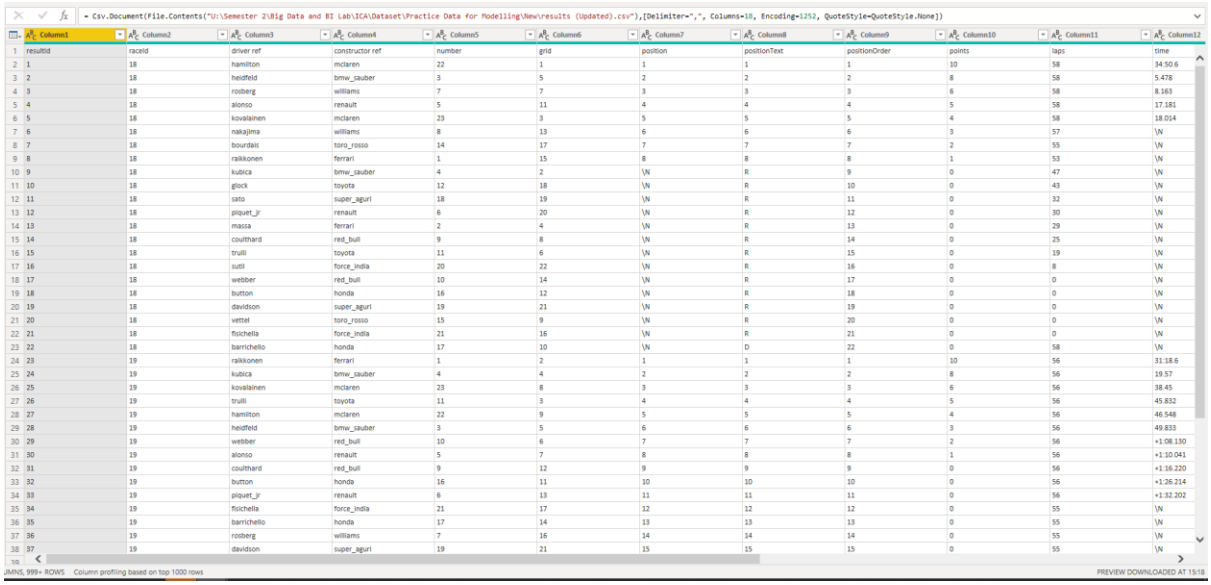
Data Pre-Processing and Data Cleansing is an integral part of any analysis and visualisation. From the Formula One dataset, I have taken only selected tables so that it's relevant to my business questions. So once the data has been uploaded to Power BI, the steps to rectify such as removing /N values, null values, renaming and removing columns, changing data types, rectifying errors, merging tables etc.

Demonstrated below are all the pre-processing steps with the **'result'** column taken as an example

The steps are given below:

- **Replacing '/N' values:**

Before Pic:



	Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9	Column10	Column11	Column12
1	resultId	raceId	driver ref	constructor ref	number	grid	position	positionText	positionOrder	points	laps	time
2	1	18	hamilton	mclaren	22	1	1	1	1	10	58	34:50.6
3	2	18	heidfeld	bmw_sauber	3	5	2	2	2	8	58	5.478
4	3	18	rosberg	williams	7	7	3	3	3	6	58	8.163
5	4	18	alonso	renault	5	11	4	4	4	5	58	17.181
6	5	18	kovalainen	mclaren	23	3	5	5	5	4	58	18.014
7	6	18	nakajima	williams	8	13	6	6	6	3	57	\N
8	7	18	bourdais	toro_rosso	14	17	7	7	7	2	55	\N
9	8	18	raikkonen	ferrari	1	15	8	8	8	1	53	\N
10	9	18	kubica	bmw_sauber	4	2	\N	R	9	0	47	\N
11	10	18	glock	toyota	12	18	\N	R	10	0	43	\N
12	11	18	sato	super_aguri	18	19	\N	R	11	0	32	\N
13	12	18	piquet_jr	renault	6	20	\N	R	12	0	30	\N
14	13	18	massa	ferrari	2	4	\N	R	13	0	29	\N
15	14	18	coulthard	red_bull	9	8	\N	R	14	0	25	\N
16	15	18	trulli	toyota	11	6	\N	R	15	0	19	\N
17	16	18	sutil	force_india	20	22	\N	R	16	0	8	\N
18	17	18	webber	red_bull	10	14	\N	R	17	0	0	\N
19	18	18	button	honda	16	12	\N	R	18	0	0	\N
20	19	18	davidson	super_aguri	19	21	\N	R	19	0	0	\N
21	20	18	vettel	toro_rosso	15	9	\N	R	20	0	0	\N
22	21	18	fisichella	force_india	21	16	\N	R	21	0	0	\N
23	22	18	barrichello	honda	17	10	\N	D	22	0	58	\N
24	23	19	raikkonen	ferrari	1	2	1	1	1	10	56	31:18.6
25	24	19	kubica	bmw_sauber	4	4	2	2	2	8	56	19.57
26	25	19	kovalainen	mclaren	23	8	3	3	3	6	56	38.45
27	26	19	trulli	toyota	11	3	4	4	4	5	56	45.832
28	27	19	hamilton	mclaren	22	9	5	5	5	4	56	46.548
29	28	19	heidfeld	bmw_sauber	3	5	6	6	6	3	56	49.833
30	29	19	webber	red_bull	10	6	7	7	7	2	56	+1:08.130
31	30	19	alonso	renault	5	7	8	8	8	1	56	+1:10.041
32	31	19	coulthard	red_bull	9	12	9	9	9	0	56	+1:16.220
33	32	19	button	honda	16	11	10	10	10	0	56	+1:26.214
34	33	19	piquet_jr	renault	6	13	11	11	11	0	56	+1:32.202
35	34	19	fisichella	force_india	21	17	12	12	12	0	55	\N
36	35	19	barrichello	honda	17	14	13	13	13	0	55	\N
37	36	19	rosberg	williams	7	16	14	14	14	0	55	\N
38	37	19	davidson	super_aguri	19	21	15	15	15	0	55	\N

Fig 1: Shows an image of the 'results' table before replacing '/N' values

Replacing '/N' values in column 'fastestLap' and 'rank' with '0' so that the null values will not having any meaning while doing aggregate functions.

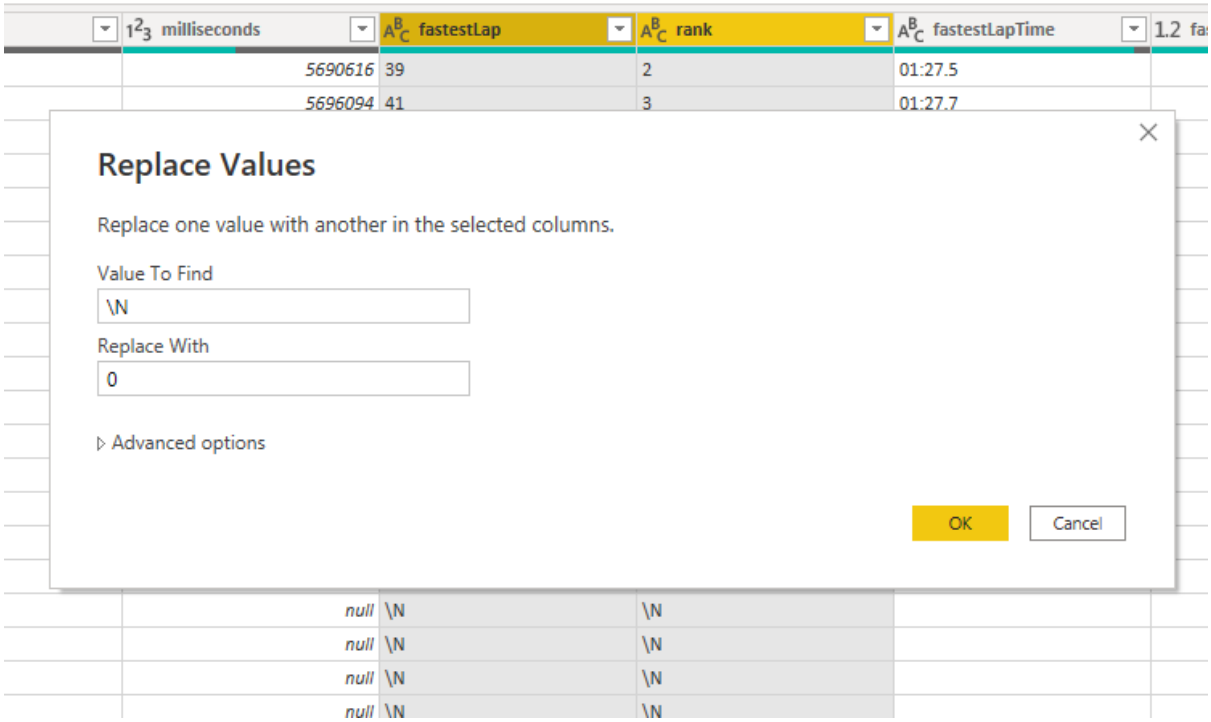


Fig 2: An image showing how to '/N' values are replaced with 0

The same process has been done for other columns in results,

- **Renaming columns:**

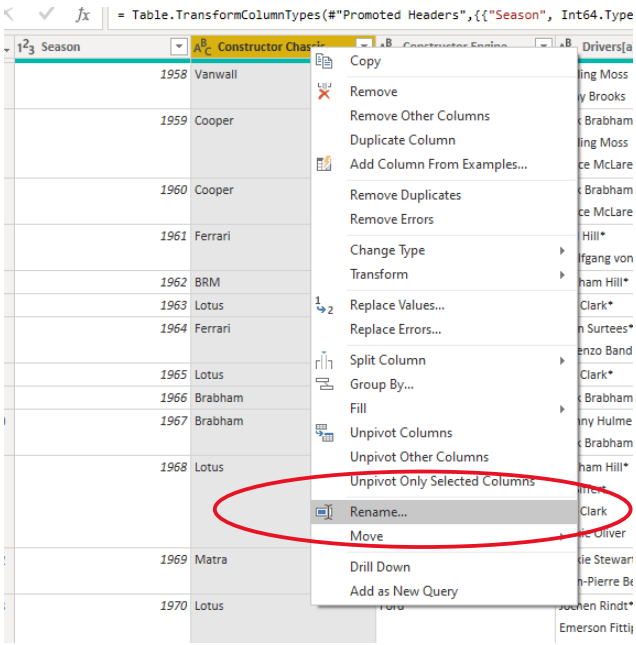


Fig 3: The'Constructor Chassis' column has been renamed to 'Constructor Champions'

- **Changing the data type:**

Changing the data type from text to the appropriate format of whole number

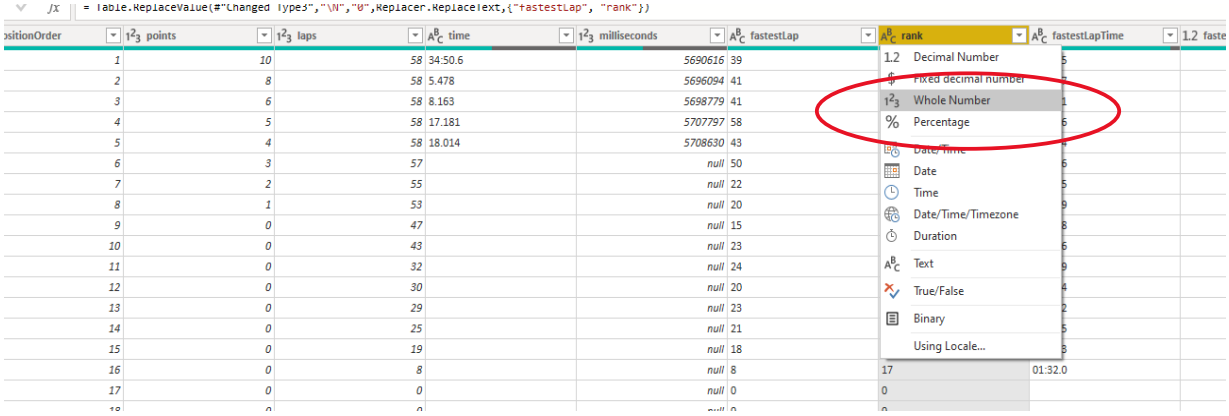


Fig 4 : A image showing how to change the data type

- **Rectifying Errors:**

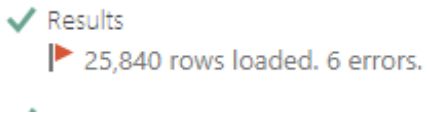


Fig 4: An image showing the errors while loading the dataset

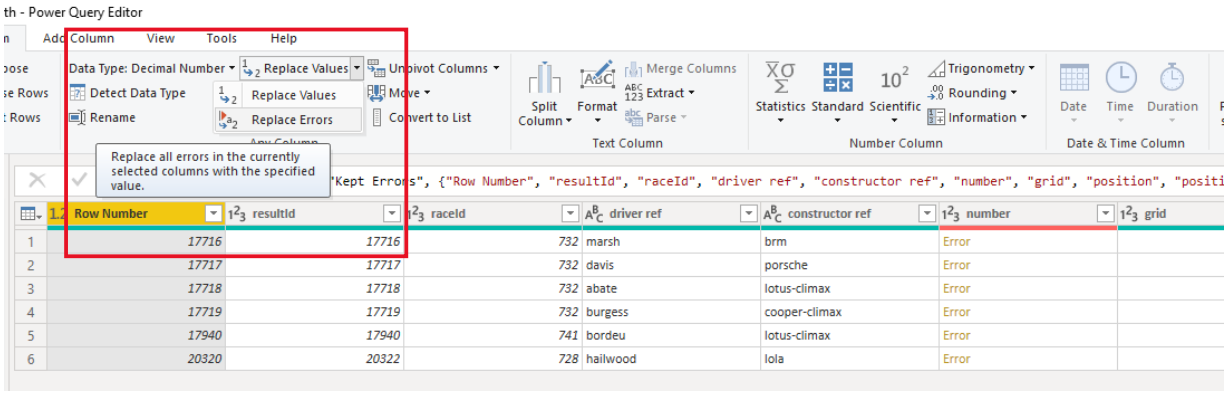


Fig 5: An image showing error rectification in the results column

Replace Errors with '0' as the values causing the error '\N' and since it's a number data type it shows up an error. Hence replacing it with 0 so that the errors are removed.

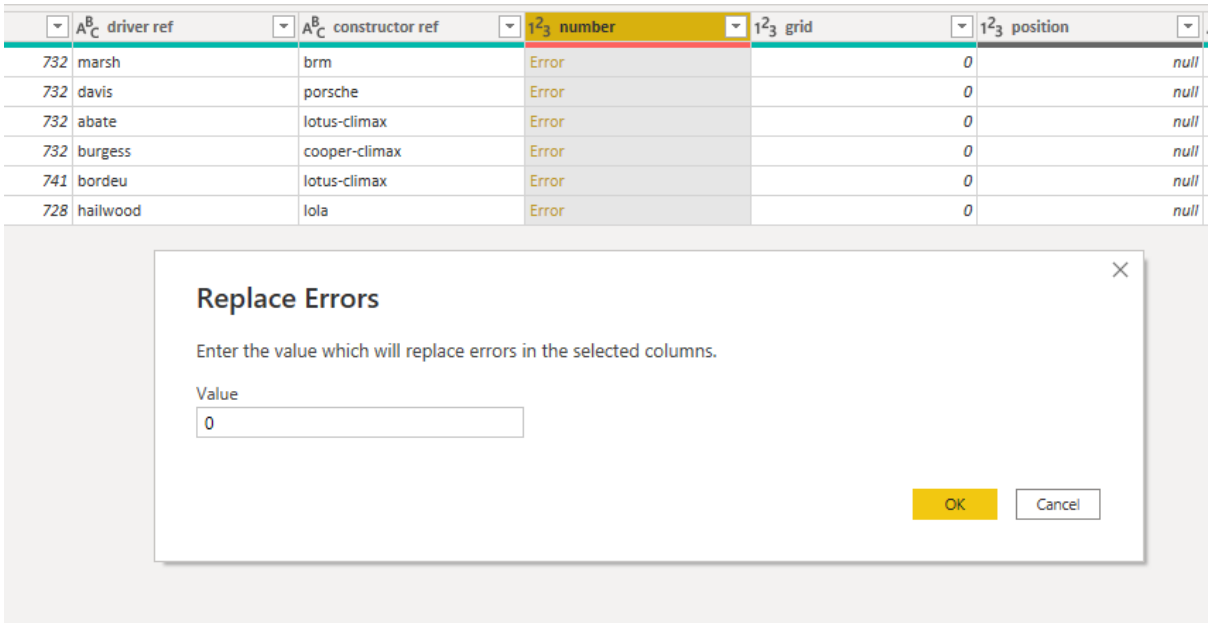


Fig 6: Replacing Error with 0 as the number column shows

After Error Correction:

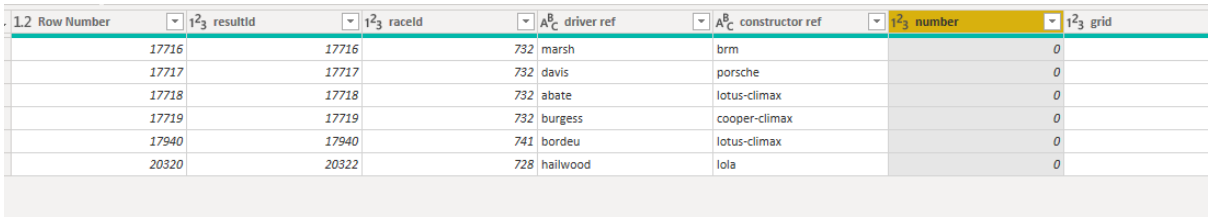


Fig 7: Image after error correction

- **Removing columns:**

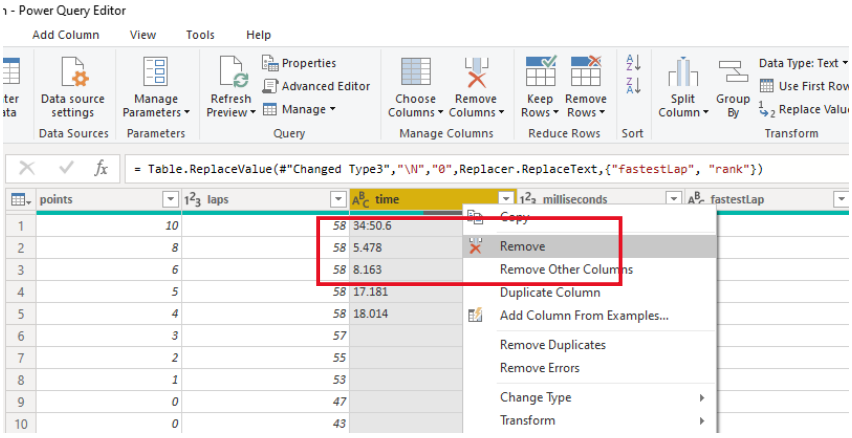


Fig 8: Image showing how to remove columns

- **Merged Queries :**

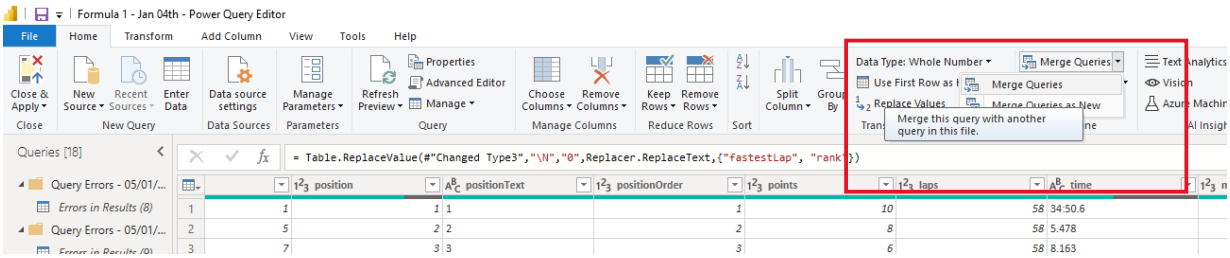


Fig 9: An Image showing the steps of merging

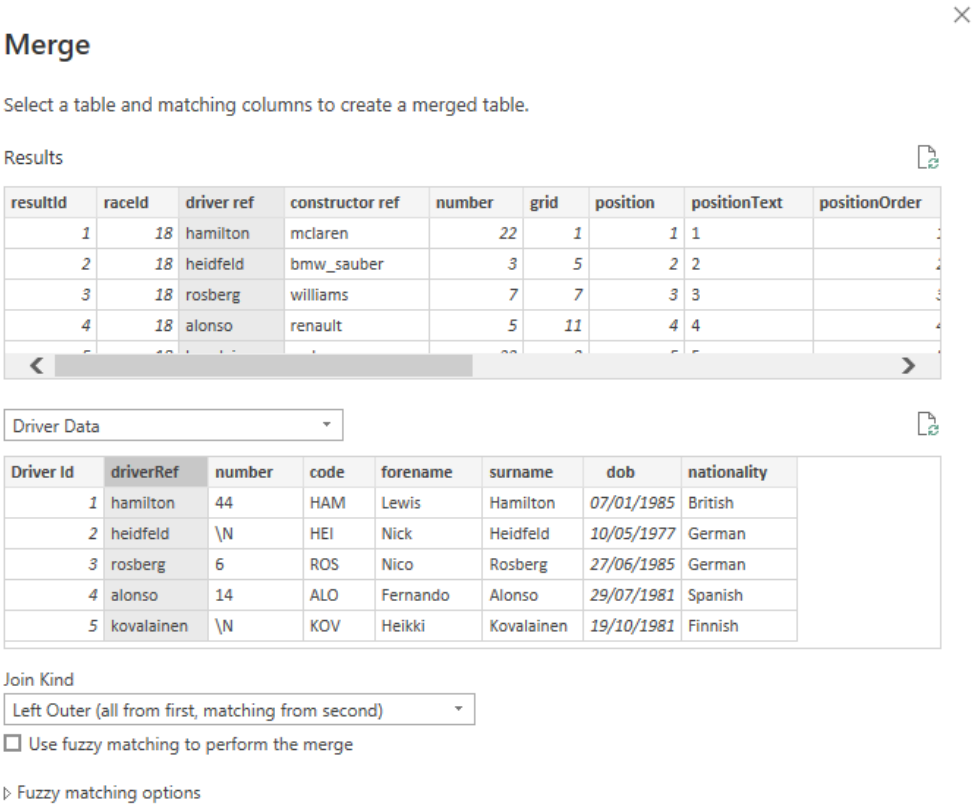


Fig 10: Columns with the same name needs to be chosen to complete merging

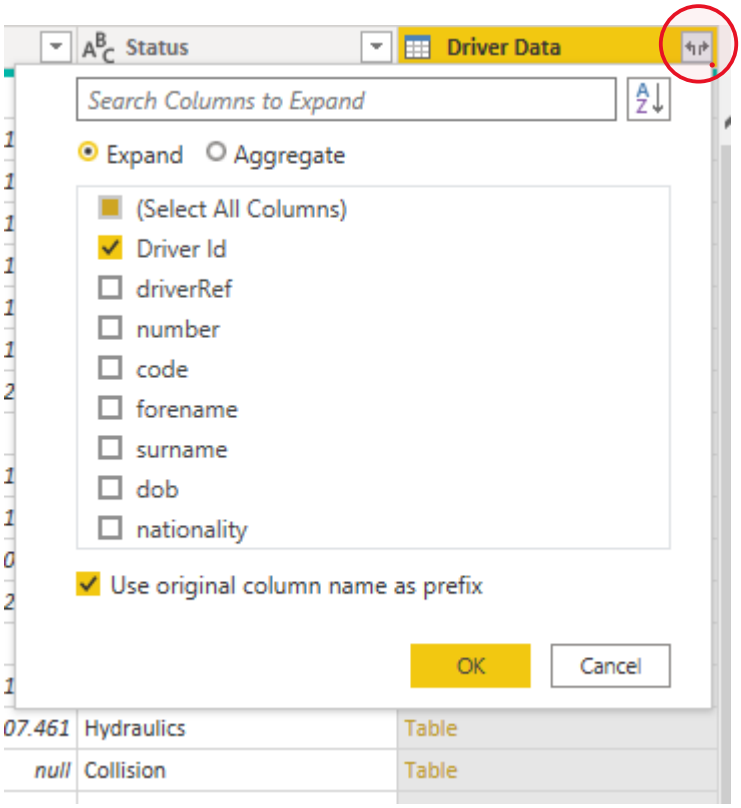
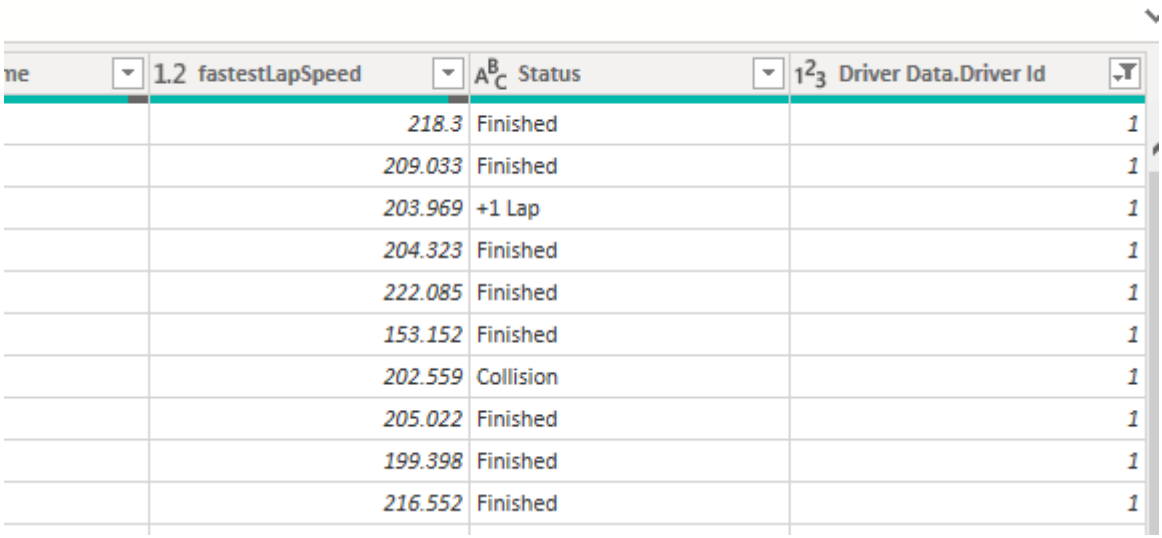


Fig 11 : Driver Id is the column we need and is expanded to get the required values



ne	1.2 fastestLapSpeed	Status	Driver Data.Driver Id
	218.3	Finished	1
	209.033	Finished	1
	203.969	+1 Lap	1
	204.323	Finished	1
	222.085	Finished	1
	153.152	Finished	1
	202.559	Collision	1
	205.022	Finished	1
	199.398	Finished	1
	216.552	Finished	1

Fig 12 : Image after Merged Query is completed

- **Results Table**

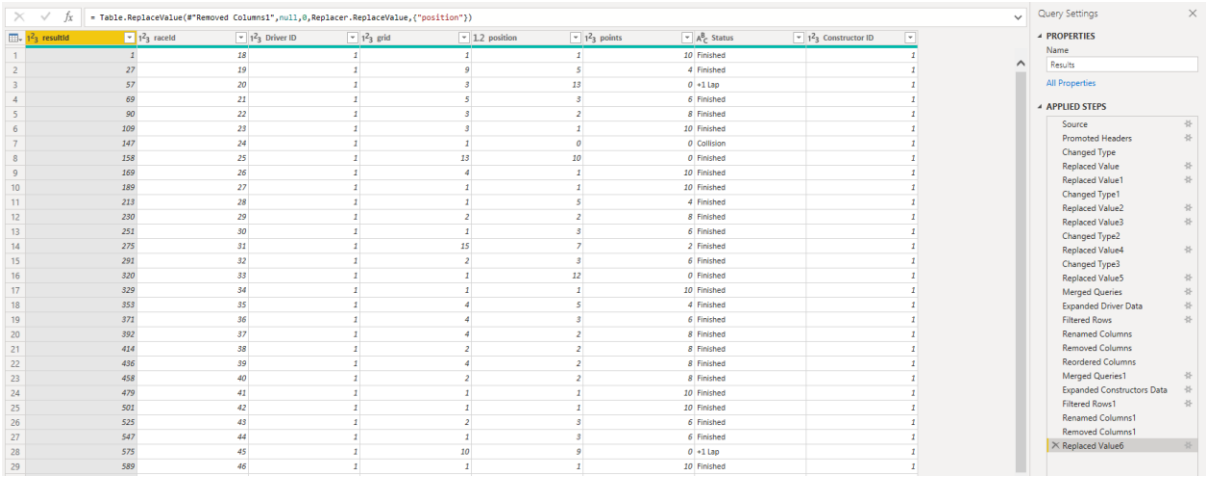
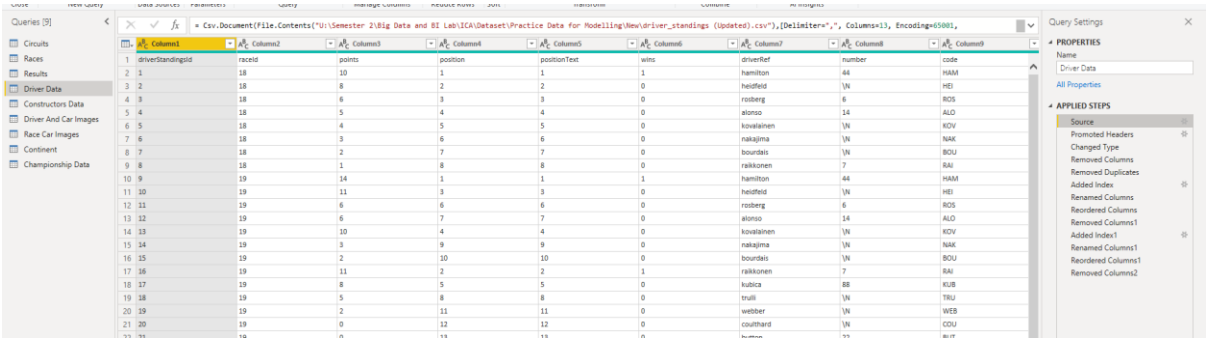


Fig 13: Results Table after pre-processing with the 'applied steps' image:

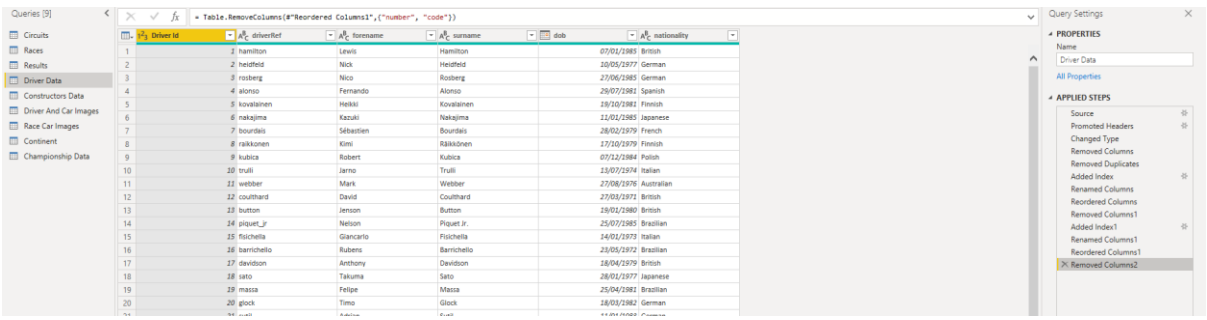
In a similar way, all the steps of pre-processing have been done for the other columns in the dataset.

- **Drivers Data Table:**

Before:

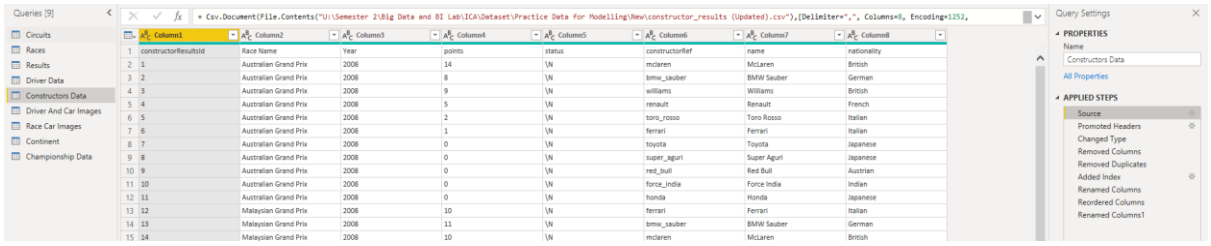


After:

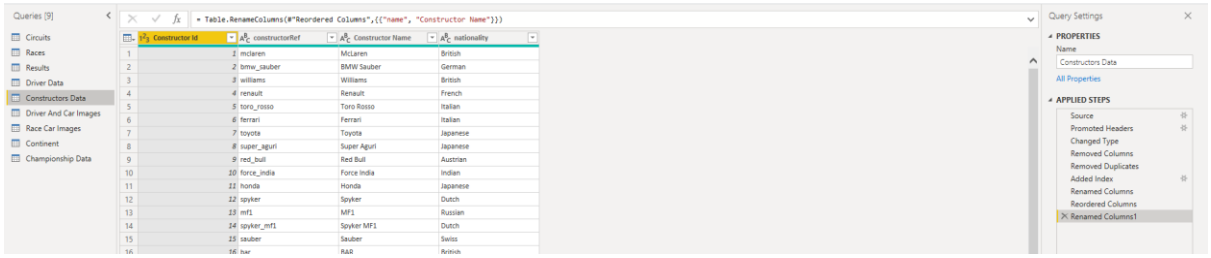


- **Constructors Data Table:**

Before:

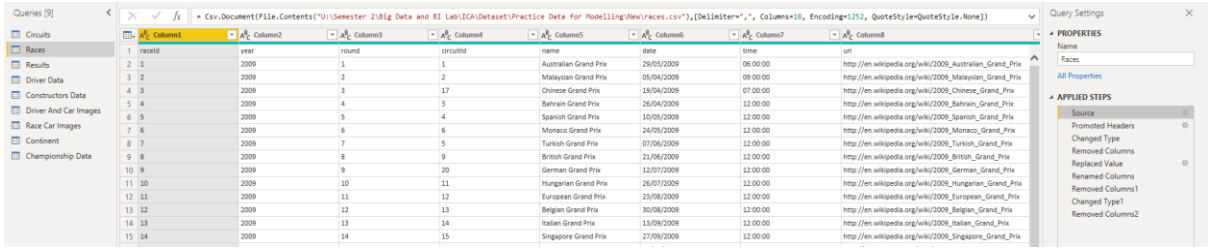


After:

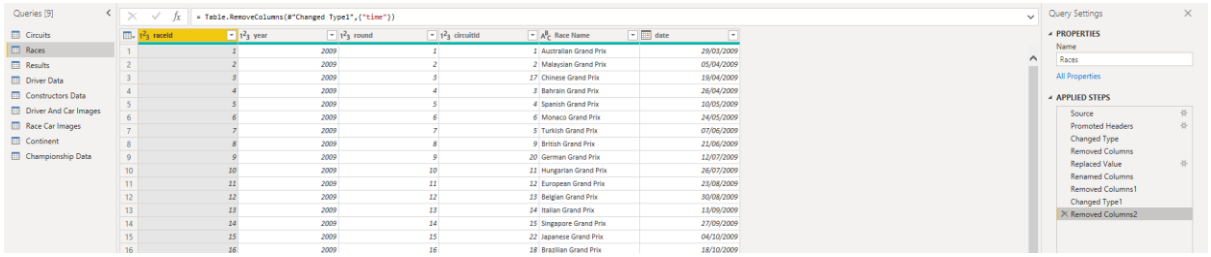


- **Races Table:**

Before:

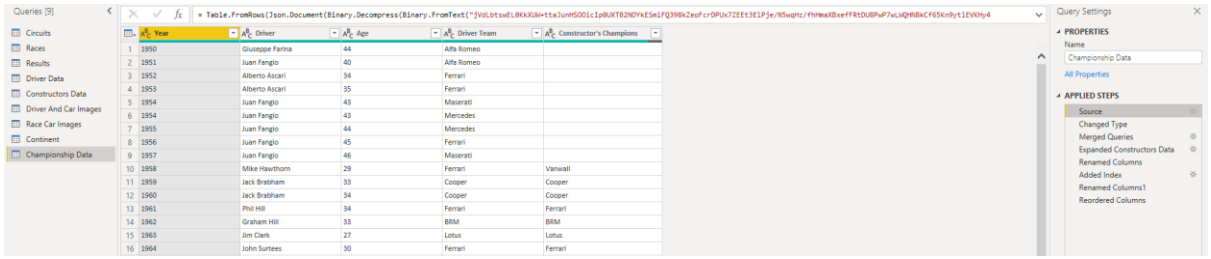


After:

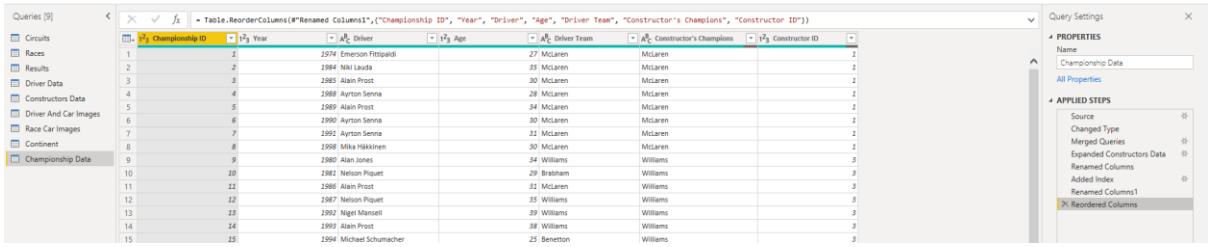


- **Championship Data**

Before:

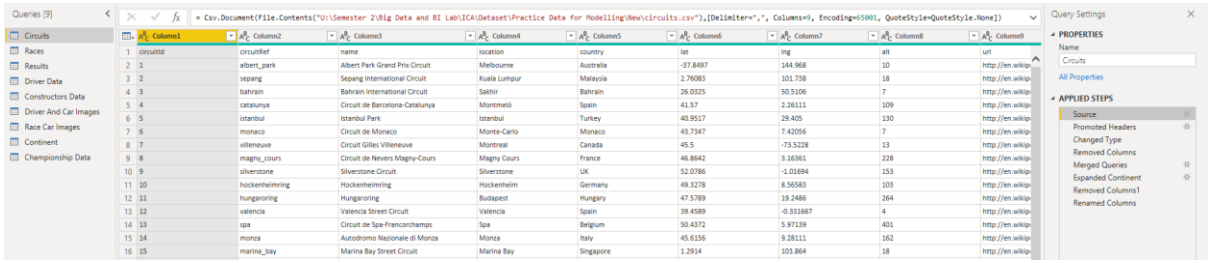


After:

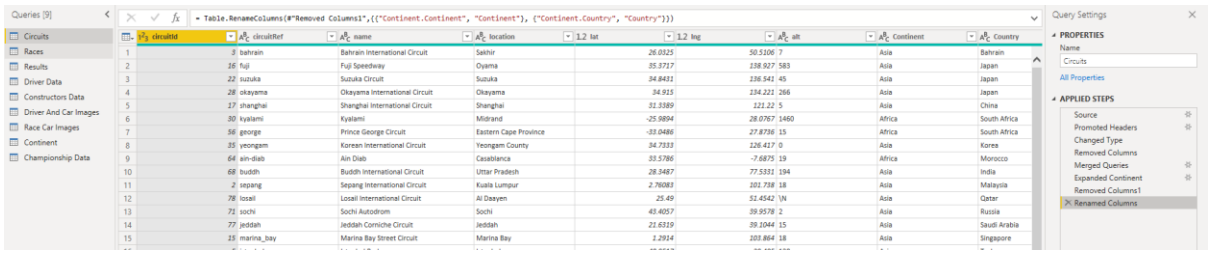


- **Circuit Data**

Before :



After:



B. Data Modeling via Star Schema:

Among the 16 tables provided from Kaggle source, I have chosen only selected tables.

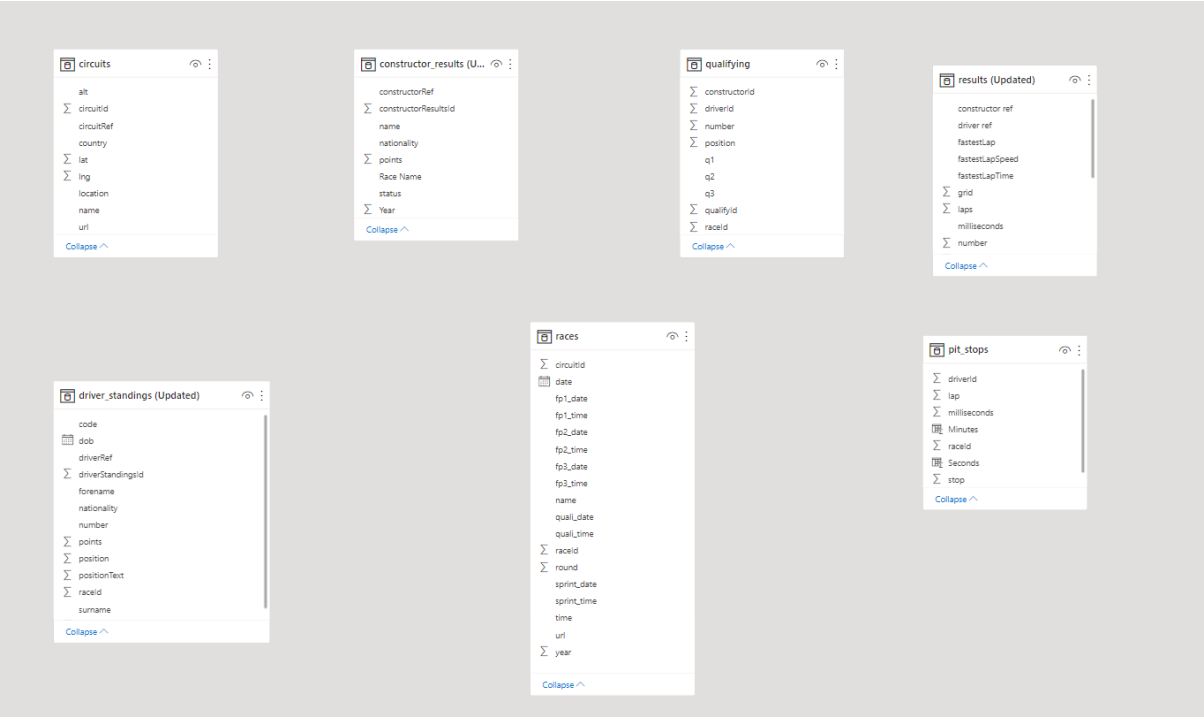


Fig 14: Before Data Modeling Pic:

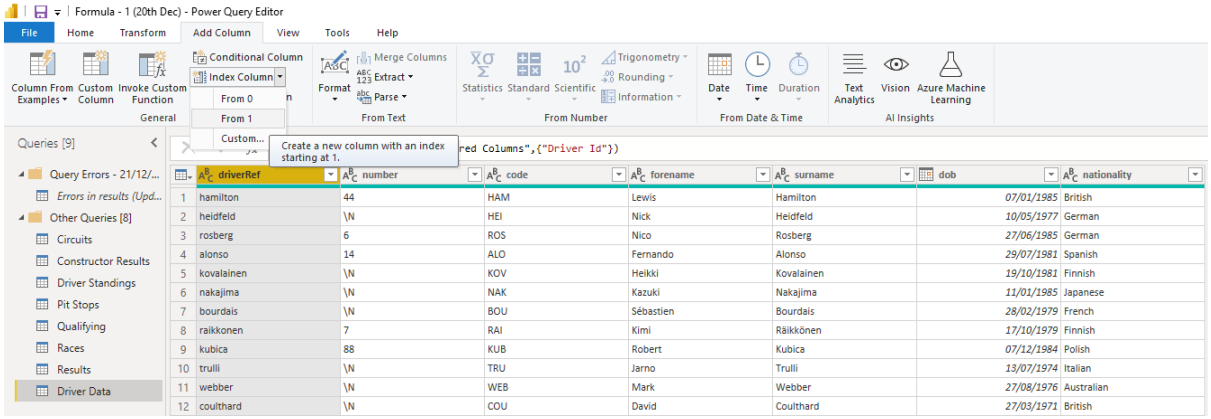
Since the dataset already has some structure, I have re-created some tables such as drivers data and constructors data to display the normalisation technique. I have then created an index column for each of these two tables and then did a merge query to connect these new columns. I have also discarded the 'pitstops' and 'qualifying' tables as I am not using them for my visualisation.

The steps for normalisation and creating new relationships is given below:

1. Normalisation

Creating a new table called ‘Drivers Data’

1. Duplicating the ‘Driver standings’ data and removing unwanted columns



	driverRef	number	code	forename	surname	dob	nationality
1	hamilton	44	HAM	Lewis	Hamilton	07/01/1985	British
2	heidfeld	\N	HEI	Nick	Heidfeld	10/05/1977	German
3	rosberg	6	ROS	Nico	Rosberg	27/06/1985	German
4	alonso	14	ALO	Fernando	Alonso	29/07/1981	Spanish
5	kovalainen	\N	KOV	Heikki	Kovalainen	19/10/1981	Finnish
6	nakajima	\N	NAK	Kazuki	Nakajima	11/01/1985	Japanese
7	bourdais	\N	BOU	Sébastien	Bourdais	28/02/1979	French
8	raikkonen	7	RAI	Kimi	Räikkönen	17/10/1979	Finnish
9	kubica	88	KUB	Robert	Kubica	07/12/1984	Polish
10	trulli	\N	TRU	Jarno	Trulli	13/07/1974	Italian
11	webber	\N	WEB	Mark	Webber	27/08/1976	Australian
12	coulthard	\N	COU	David	Coulthard	27/03/1971	British

Fig 15: Before Pic of ‘Driver Standings’ table:

2. Removing unwanted columns in Driver Standings

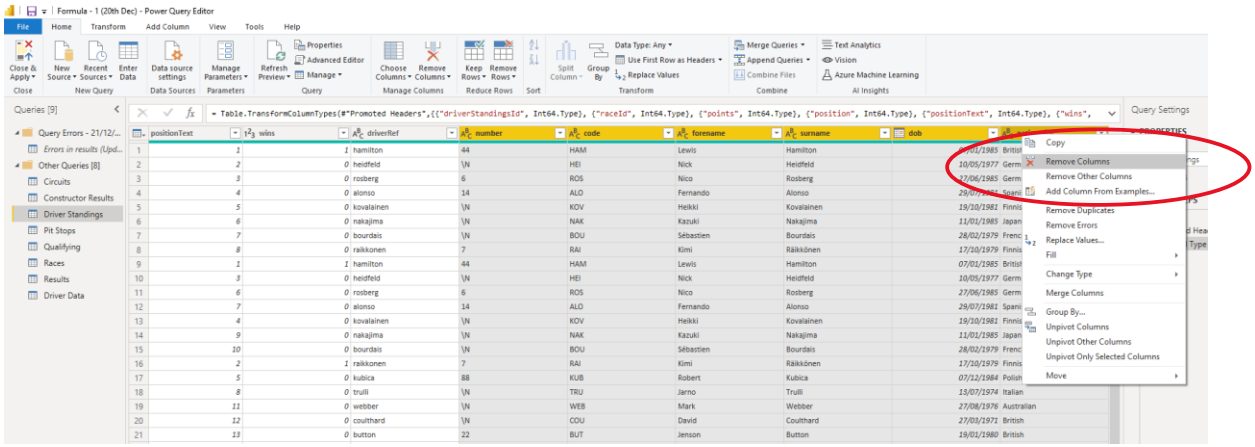


Fig 16: Removing unnecessary columns in ‘Driver Standing’s table

3. Creating Index column driver Id:

We are selecting the index column from 1

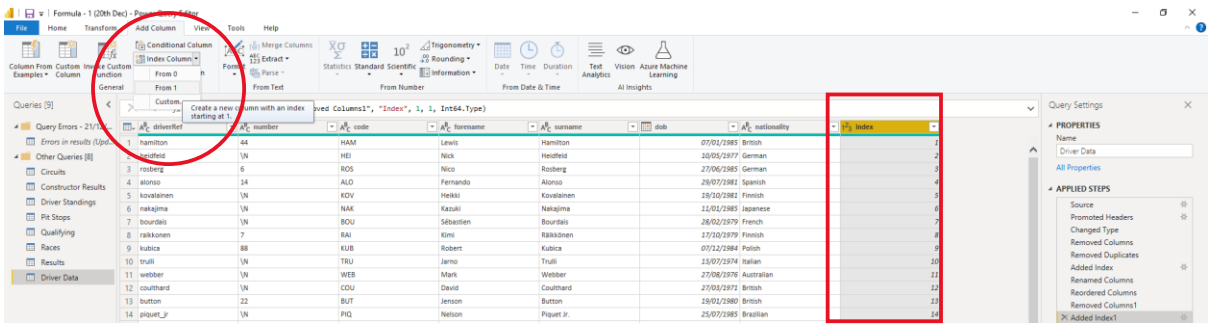
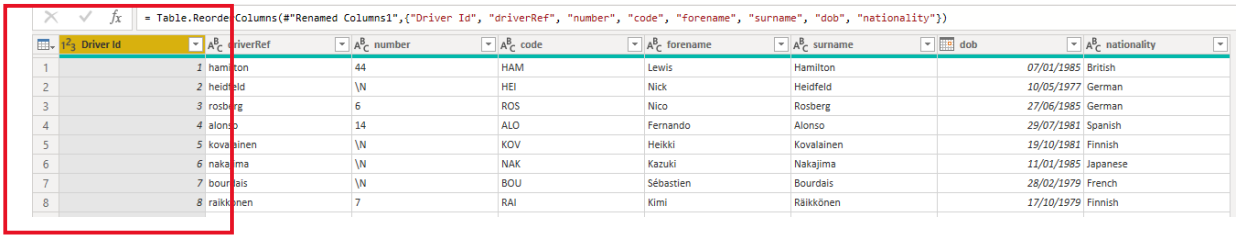


Fig 17:Image showing the main options to choose for creating an index column

After pic:



= Table.ReorderColumns(#"Renamed Columns1",{"Driver Id", "driverRef", "number", "code", "forename", "surname", "dob", "nationality"})

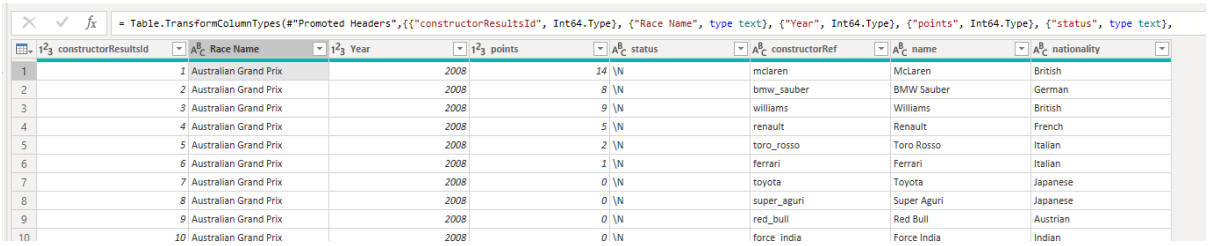
	Driver Id	driverRef	number	code	forename	surname	dob	nationality
1	1	hamilton	44	HAM	Lewis	Hamilton	07/01/1985	British
2	2	heidfeld	\N	HEI	Nick	Heidfeld	10/05/1977	German
3	3	rosberg	6	ROS	Nico	Rosberg	27/06/1985	German
4	4	alonso	14	ALO	Fernando	Alonso	29/07/1981	Spanish
5	5	kovalainen	\N	KOV	Heikki	Kovalainen	19/10/1981	Finnish
6	6	nakajima	\N	NAK	Kazuki	Nakajima	11/01/1985	Japanese
7	7	bourdais	\N	BOU	Sébastien	Bourdais	28/02/1979	French
8	8	raikkonen	7	RAI	Kimi	Räikkönen	17/10/1979	Finnish

Fig 18:And after pic of the new 'Drivers ID' column

Creating a new Constructors Data Table:

The same steps shown in creating the 'Drivers ID' table has been used to create the Constructors Data table.

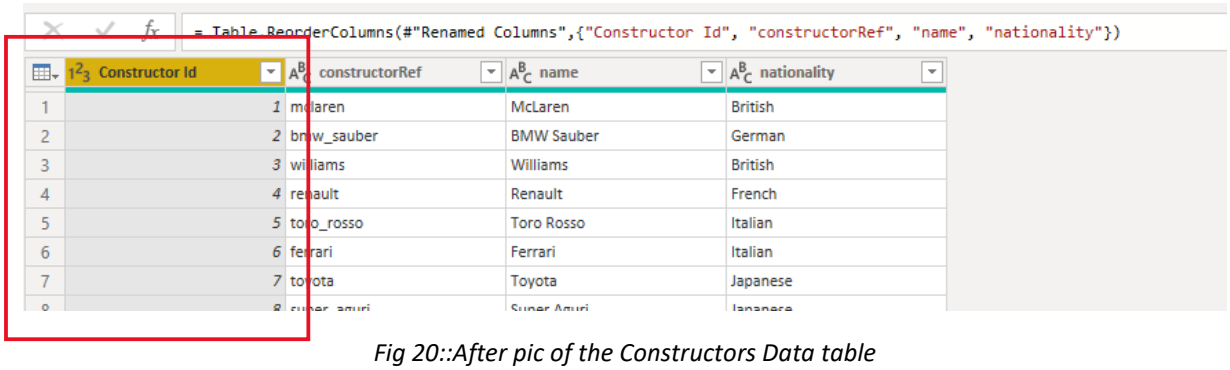
The constructors data table has been created from the **constructors results** table shown below.



= Table.TransformColumnTypes(#"Promoted Headers",{{"constructorResultsId", Int64.Type}, {"Race Name", type text}, {"Year", Int64.Type}, {"points", Int64.Type}, {"status", type text},

	constructorResultsId	Race Name	Year	points	status	constructorRef	name	nationality
1	1	Australian Grand Prix	2008	14	\N	mclaren	McLaren	British
2	2	Australian Grand Prix	2008	8	\N	bmw_sauber	BMW Sauber	German
3	3	Australian Grand Prix	2008	9	\N	williams	Williams	British
4	4	Australian Grand Prix	2008	5	\N	renault	Renault	French
5	5	Australian Grand Prix	2008	2	\N	toro_rosso	Toro Rosso	Italian
6	6	Australian Grand Prix	2008	1	\N	ferrari	Ferrari	Italian
7	7	Australian Grand Prix	2008	0	\N	toyota	Toyota	Japanese
8	8	Australian Grand Prix	2008	0	\N	super_aguri	Super Aguri	Japanese
9	9	Australian Grand Prix	2008	0	\N	red_bull	Red Bull	Austrian
10	10	Australian Grand Prix	2008	0	\N	force_india	Force India	Indian

Fig 19:Before pic of the Constructor Results table



= Table.ReorderColumns(#"Renamed Columns",{"Constructor Id", "constructorRef", "name", "nationality"})

	Constructor Id	constructorRef	name	nationality
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

Fig 20::After pic of the Constructors Data table

2. Creating New relationships

- Showing data relationship between Drivers Data and Results table

Step 1: Select 'Transform Data'

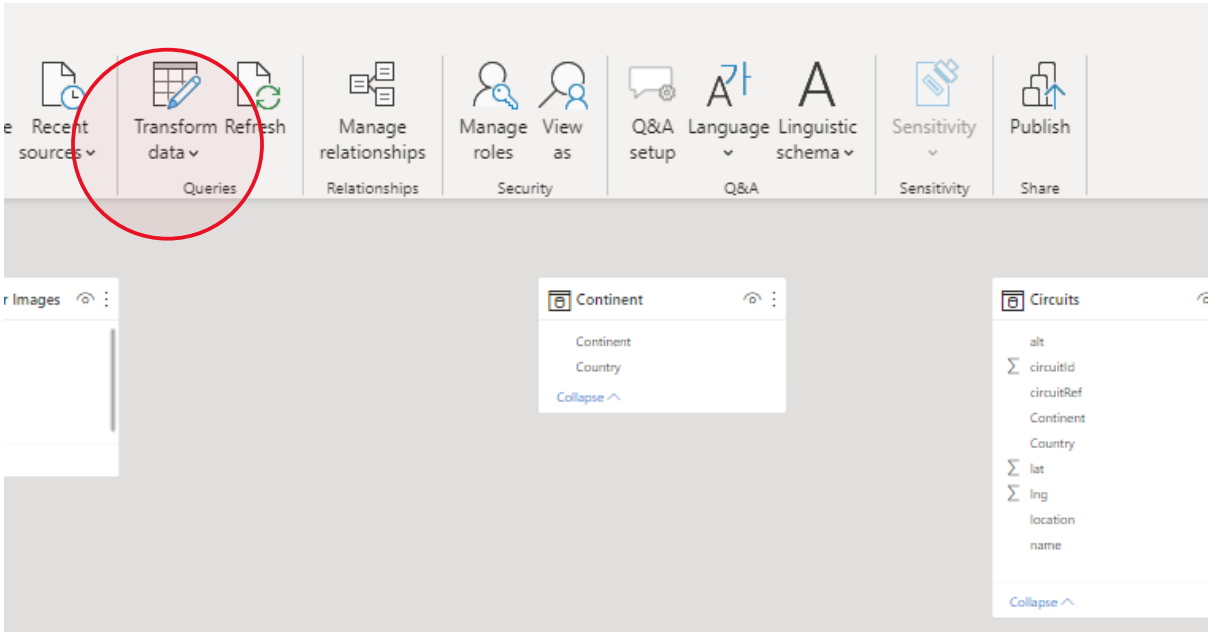
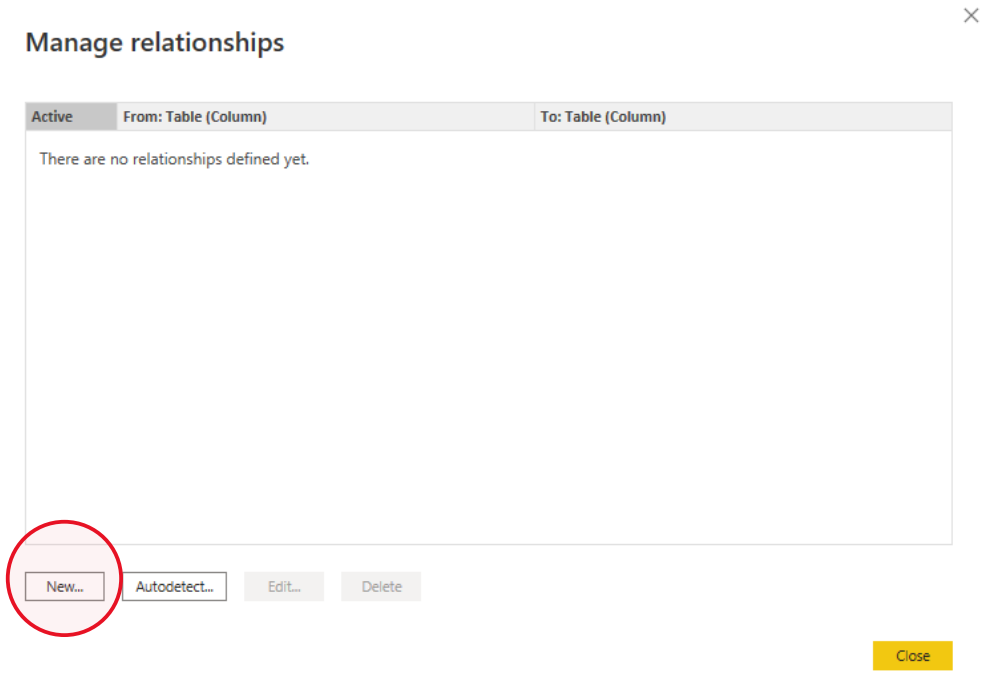


Fig 21:First step is to select Transform data to create a new relationship

Step 2: Click on 'New'



Step 3:

Select the key column with which you would connect the two tables. Creating relationship with Drivers Data and Results table.

We choose the cardinality as One to Many 0r Many to One and the cross filter direction as 'Single'.

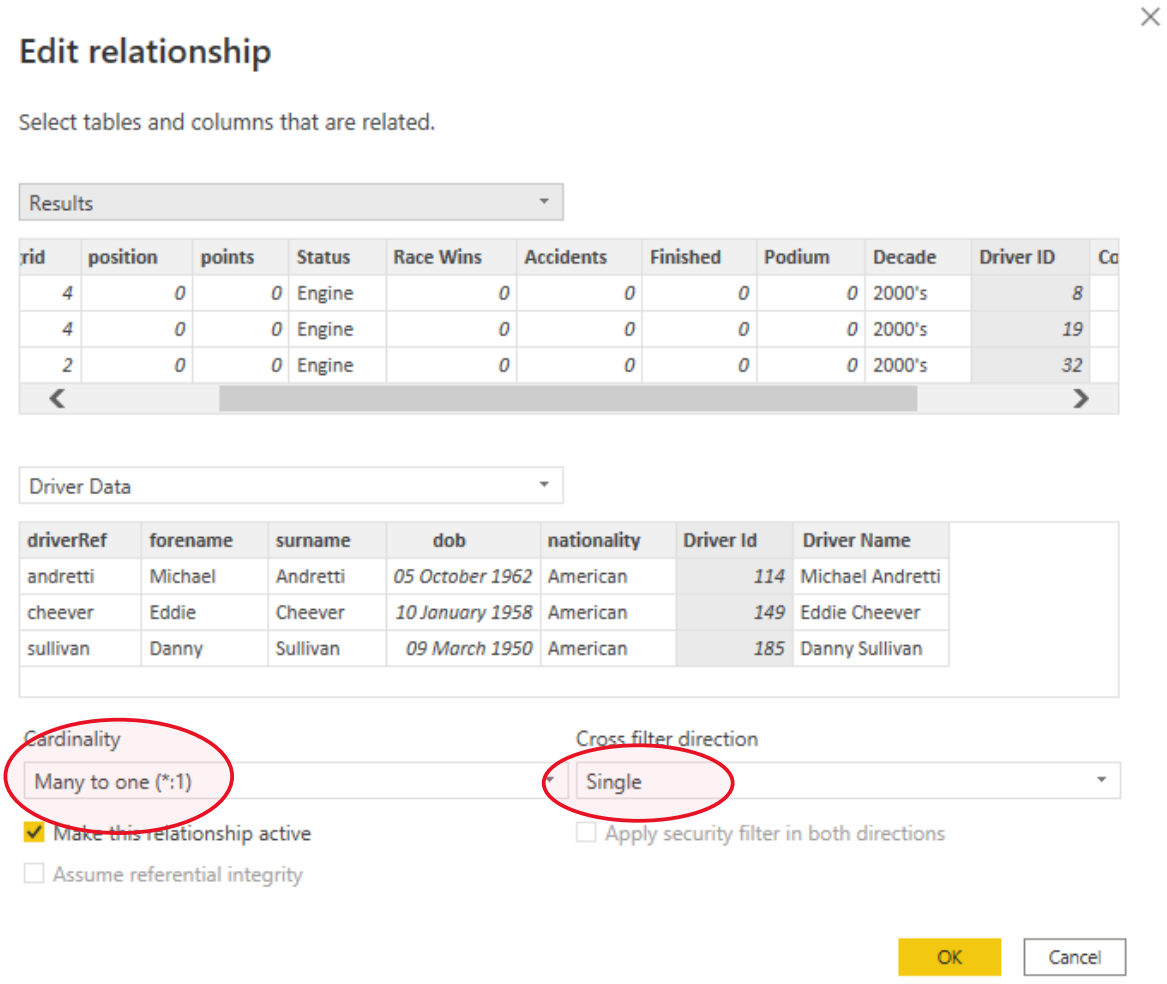


Fig 22:'Driver ID' is the key which connects these two tables

We choose the Cardinality as 'Many to One' or 'One to Many' because we want data to connect from a dimension table such as 'Driver Data' which has all unique values (Single) to the table having ('Many').

We choose the cross filter to be 'Single' because I have two or more tables that also have lookup tables and the data modelling is a Snowflake Schema.

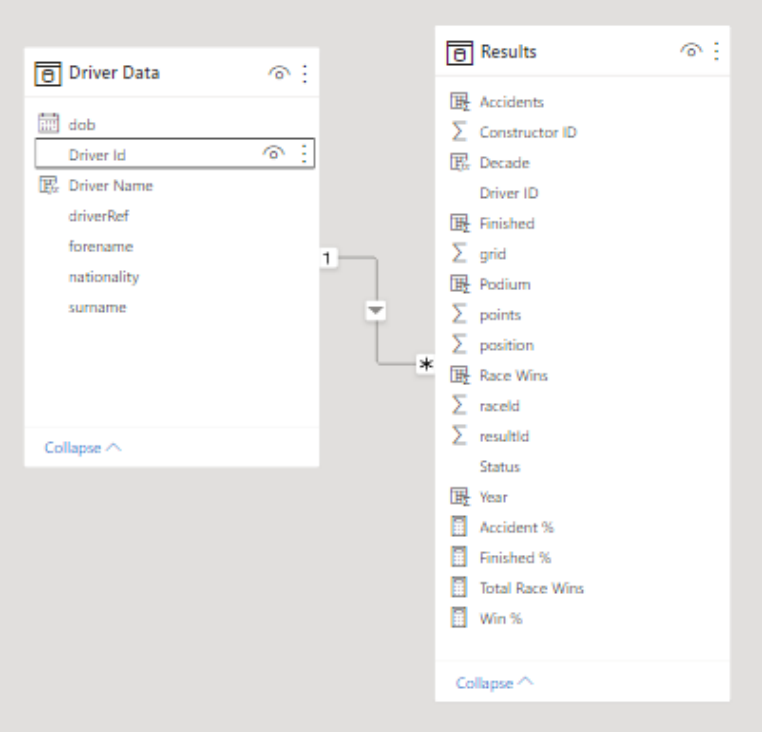


Fig 23:An after pic showing the relationship between 'Driver Data' and 'Results' table

- **Creating the relationship between Constructors Data and Results table:**

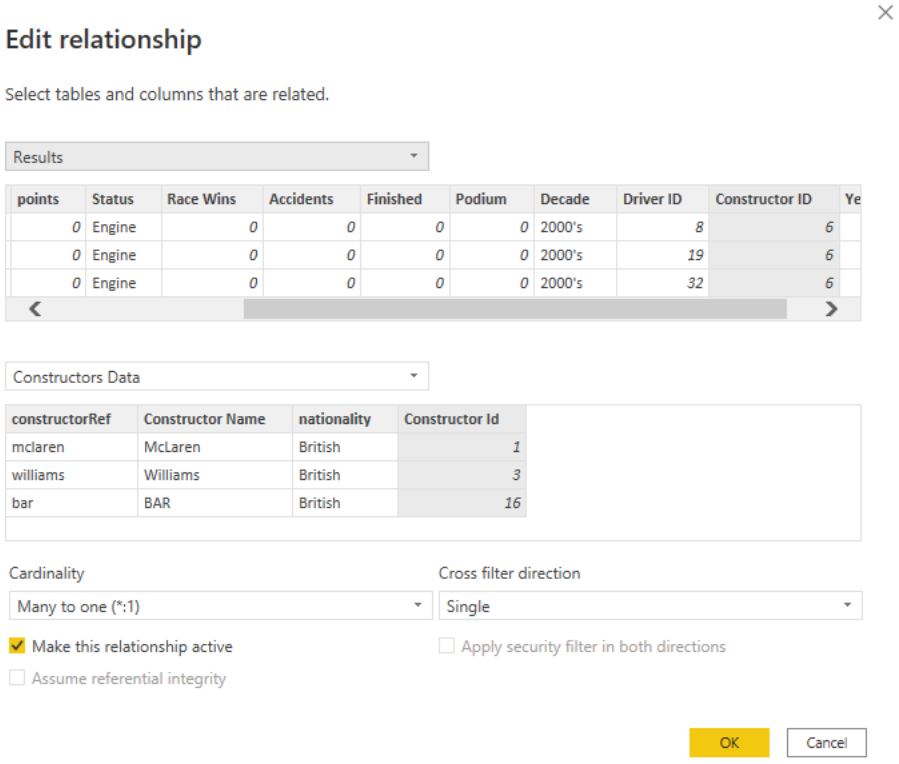
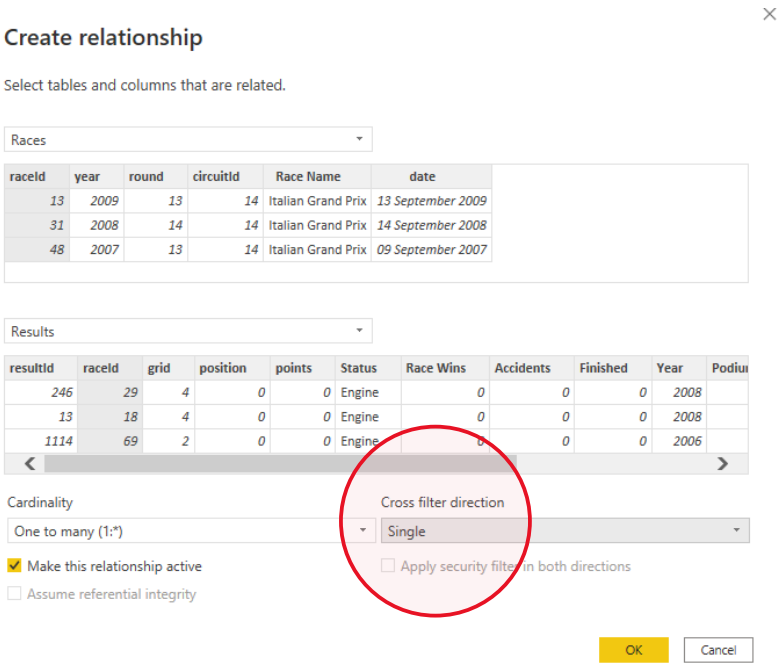


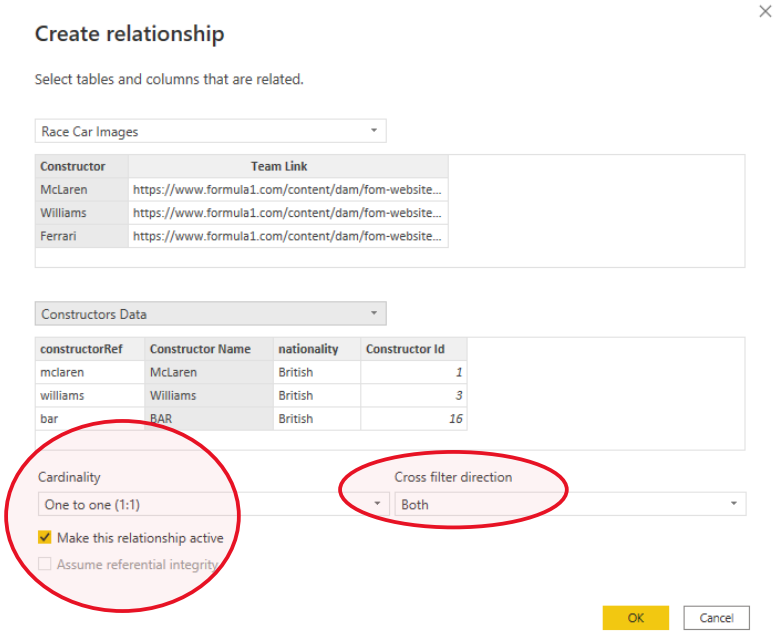
Fig 23:'Constructor ID' is the key which connects these two tables

'Constructor ID' is the key which connects these two tables



'raceid' is the key which connects these two tables

- Creating Relationship between custom columns created for 'Race Car Images' to 'Constructors Data'



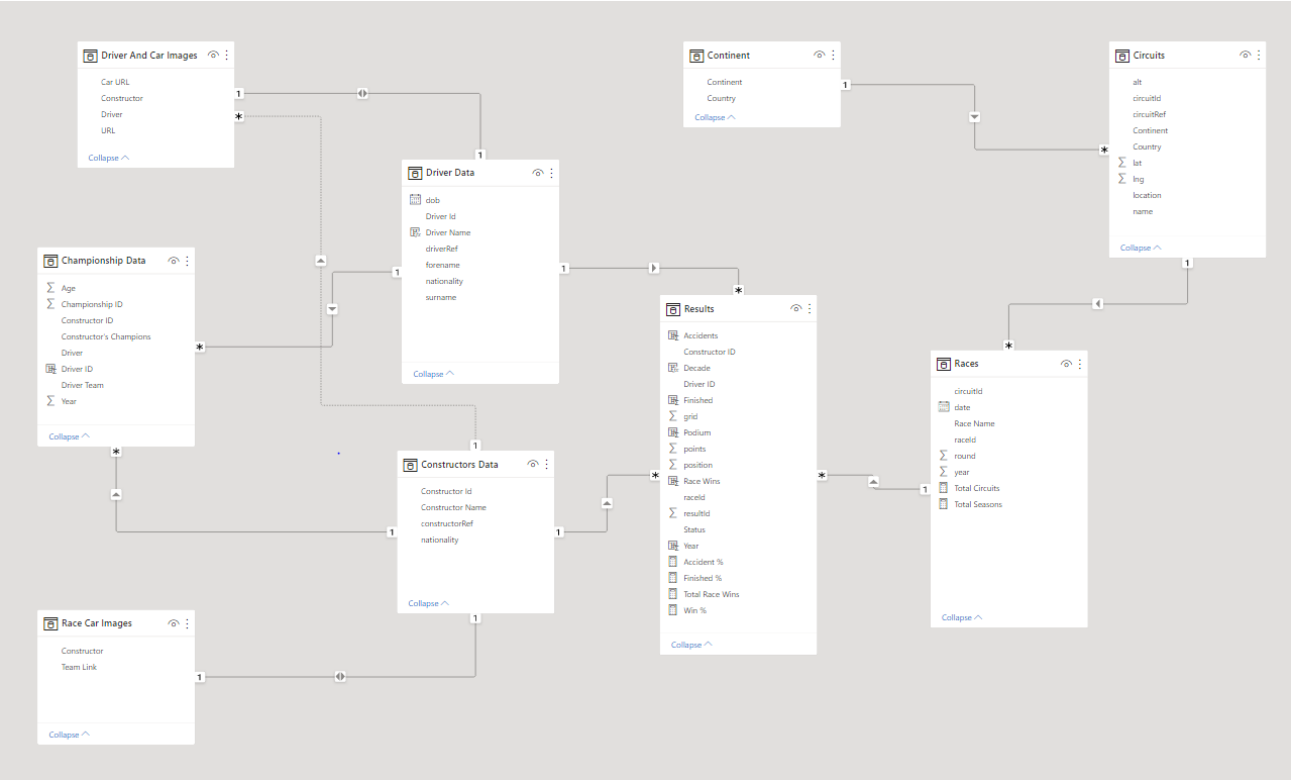
In this case we have a One to One (1:1) relationship because the Race Car Images is a unique column with a purpose only for the images and the Cross filter direction is 'Both' so that the information can flow both ways.

Hence using the same method, all the relationships have been created

All the relationships and their keys:

1. Driver Data to Results Data — **Key : Driver ID:**
   Cardinality : One to Many
   Cross Filter Direction : Single
2. Constructors Data to Results Data — **Key : Constructor ID:**
   Cardinality : One to Many
   Cross Filter Direction : Single
3. Races to Results Data — **Key : raceid**
   Cardinality : One to Many
   Cross Filter Direction : Single
4. Races to Results Data — **Key : raceid**
   Cardinality : One to Many
   Cross Filter Direction : Single
5. Circuits to Races Data — **Key : circuitid**
   Cardinality : One to Many
   Cross Filter Direction : Single
6. Championship Data to Driver Data — **Key : Driver ID**
   Cardinality : Many to One
   Cross Filter Direction : Single
7. Championship Data to Constructors Data — **Key : Constructor ID**
   Cardinality : Many to One
   Cross Filter Direction : Single
8. Race Car Images to Constructor Data — **Key : Driver Name**
   Cardinality : One to One
   Cross Filter Direction : Both
9. Driver and Car Images to Constructor Data — **Key : Constructor Name**
   Cardinality : Many to One
   Cross Filter Direction : Single
10. Driver and Car Images to Driver Data — **Key : Driver Name**
    Cardinality : One to One
    Cross Filter Direction : Both

Final Screenshot after creating the relationships:



C. DAX and M Language

DAX:

DAX and M language has been used to create custom columns and measures so as to get the right visualisation.

- **Decade**: A column named '**Decade'** has been created which is used in the zoomable sunburst chart so to understand which teams are dominant and also to keep the chart clean instead of having multiple years scattered throughout.

```
Decade = IF(Results[Year] < 1960, "1950's",
            IF(AND(Results[Year] < 1970, Results[Year] >= 1960), "1960's",
               IF(AND(Results[Year] < 1980, Results[Year] >= 1970),"1970's",
                  IF(AND(Results[Year] < 1990, Results[Year] >= 1980), "1980's",
                     IF(AND(Results[Year] < 2000, Results[Year] >= 1990), "1990's",
                        IF(AND(Results[Year] < 2010, Results[Year] >= 2000), "2000's",
                           IF(AND(Results[Year] < 2020, Results[Year] >= 2010), "2010's",
                              IF(AND(Results[Year] < 2030, Results[Year] >= 2020), "2020's", "Unknown")))))))
```

ceId	grid	position	points	Status	Race Wins	Accidents	Finished	Year	Podium	Decade	Driver ID	Constructor ID
29	4	0	0	Engine	0	0	0	2008	0	2000's	8	6
18	4	0	0	Engine	0	0	0	2008	0	2000's	19	6

- **Podium**: A column named '**Podium'** has been created to understand how many podiums drivers achieve. This is especially useful in the 'KPI Influencers' visual.

ture | Formatting | Properties | Sort

```
Podium = IF
         (OR(if(Results[position] = 1,1,
            IF(Results[position] = 2,1,
               IF(Results[position]=3,1,0))),0),1,0)
```

raceId	grid	position	points	Status	Race Wins	Accidents	Finished	Year	Podium	De
29	4	0	0	Engine	0	0	0	2008	0	20
18	4	0	0	Engine	0	0	0	2008	0	20

- **Race Wins:** A column named **'Race wins'** is created to understand the total wins by drivers and constructors. This column has been extremely useful is creating many visual for the driver and constructor analysis.

```
Race Wins =
    IF(Results[position] = 1, 1 ,0)
```

aceId	grid	position	points	Status	Race Wins	A
29	4	0	0	Engine	0	
18	4	0	0	Engine	0	

- **Accidents:** A column named **'Accidents'** is created to understand how the average accident percentage over the years in formula one history.

```
1 Accidents =
2         IF(Results[Status] = "Accident", 1, 0)
```

aceId	grid	position	points	Status	Race Wins	Accidents
29	4	0	0	Engine	0	0

- Driver name: A column named **'Driver Name'** is created to join driver forename and surname together.

```
1 Driver Name = 'Driver Data'[forename] & " " & 'Driver Data'[surname]
```

forename	surname	dob	nationality	Driver Id	Driver Name
Michael	Andretti	05 October 1962	American	114	Michael Andretti
Eddie	Cheever	10 January 1958	American	149	Eddie Cheever
Danny	Sullivan	09 March 1950	American	185	Danny Sullivan

- **Driver ID**

  A column named **'Driver ID** is created by using the 'LOOKUPVALUE' function.

```
1 Driver ID =
2         LOOKUPVALUE('Driver Data'[Driver Id],'Driver Data'[Driver Name],'Championship Data'[Driver])
```

Driver	Age	Driver Team	Constructor's Champions	Constructor ID	Championship ID	Driver ID
rson Fittipaldi	27	McLaren	McLaren	1	1	214
Lauda	35	McLaren	McLaren	1	2	174

Measures:

- **Win %**

```
1  Win % = (sum(Results[Race Wins])/sum(Results[Finished]))
```

- **Total Race Wins**

```
Total Race Wins = SUM(Results[Race Wins])
```

- **Total Circuits**

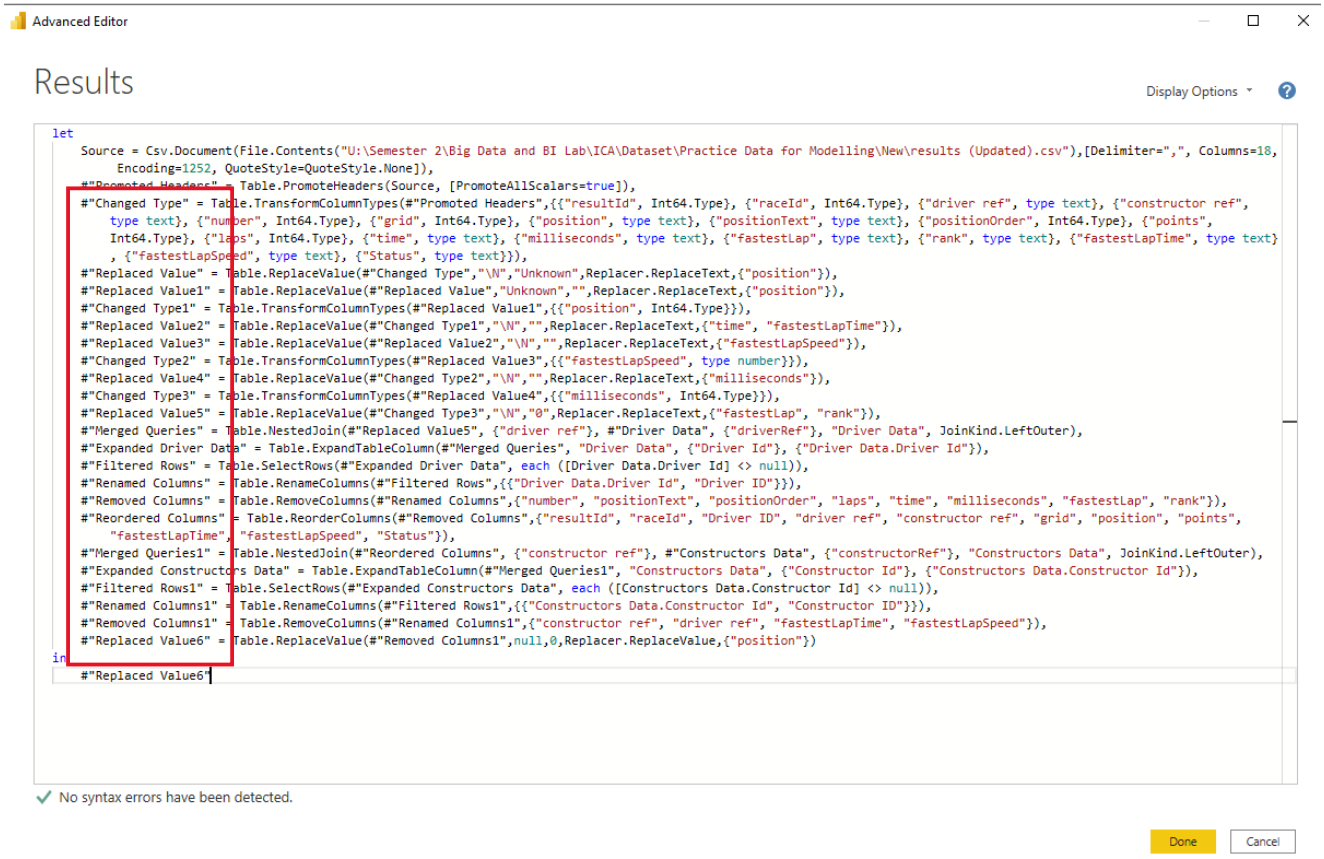
```
Total Circuits = DISTINCTCOUNT(Circuits[circuitId])
```

- **Total Seasons**

```
Total Seasons = DISTINCTCOUNT(Races[year])
```

M Language

- **Results Table:**



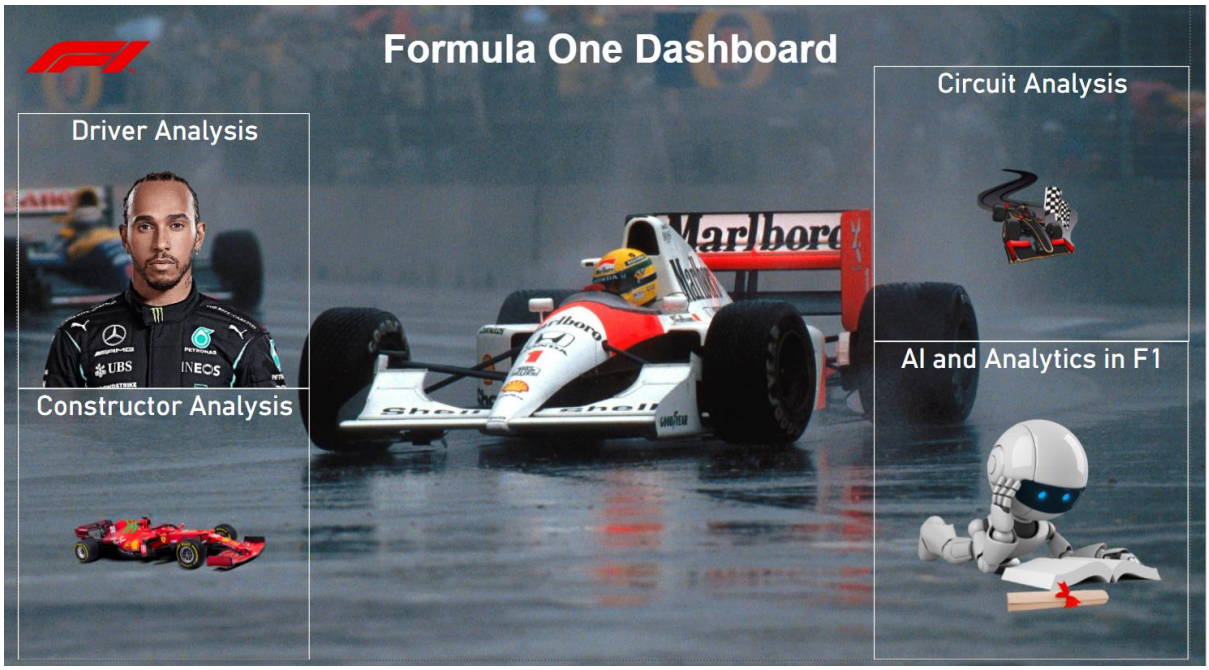
The highlighted region shows the pre-processing steps which are done in m-language this includes Replace Values, Fixing errors and Merged Queries etc.

In the same way we can see how m-language is used in other tables as well.

D. Dashboard

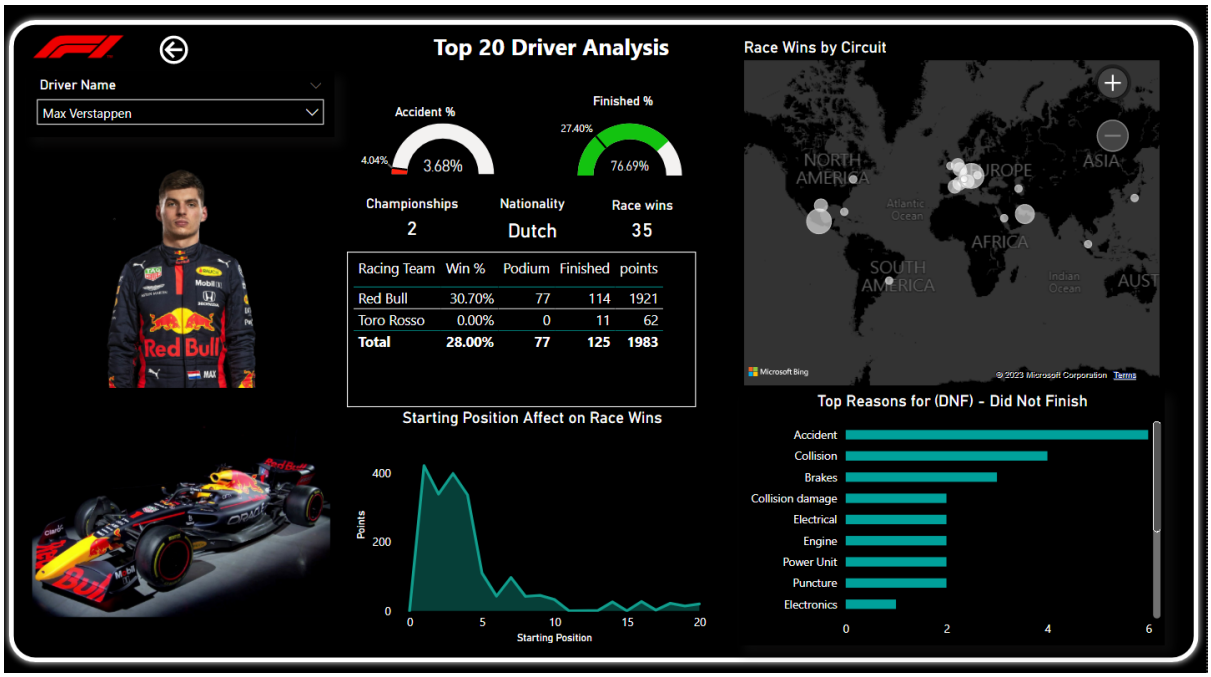
- Power BI dashboard (full collection of visuals) and how the content of the Power BI pages is organised.

Home Page :



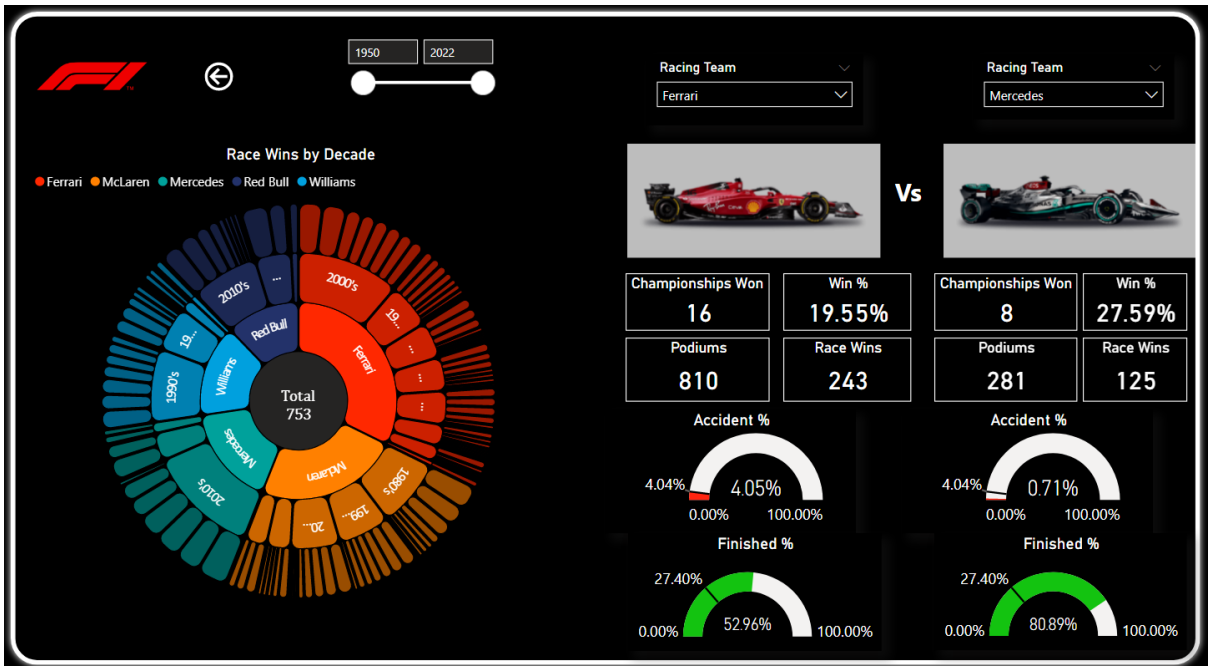
- This is the home page of the dashboard. All the 4 pages of the dashboard is linked to the home page.
- I have taken first the Driver Analysis to be my first page because the sport mainly revolves around the drivers.
- The second page is Constructor Analysis. The teams of each of these drivers is analysed here.
- The third page is circuit analysis of all the different race tracks in F1 and analysis based on track location is done here.
- AI and Analytics in F1 is the last page. It uses AI related charts to forecast and see what factors influences certain performance metrics like race wins.

Driver Analysis:



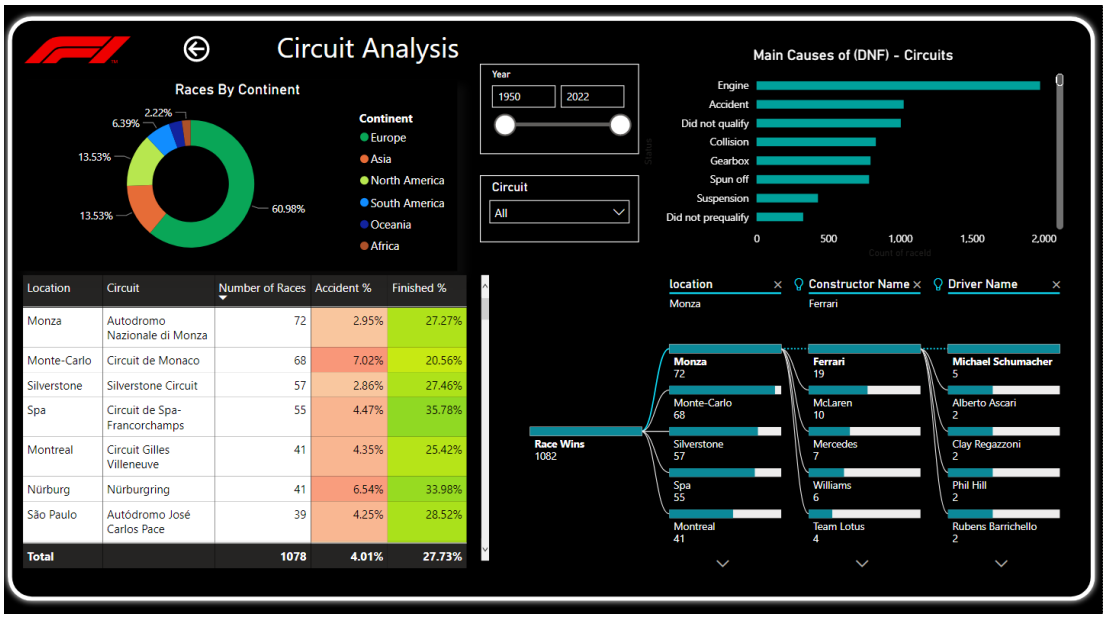
- This is the first page of my dashboard.
- It gives an in-depth analysis of the top 20 Drivers sorted by the race wins.
- The design is done with the consideration of normal human eye movement going from the left side of the screen to the right.
- Hence the driver image and their respective cars are to the left with the slicer.
- Then the gauge axis is kept in the middle which shows accident % and finished % along with the cards and table showing individual driver stats. An area chart is kept beneath it as well to show the influence of starting points on race wins.
- To the right we have the map chart which shows the circuits each driver has won a race.
- The graph to the bottom right is a horizontal bar chart showing top reasons for (DNF) – Did Not Finish.

Constructor Analysis:



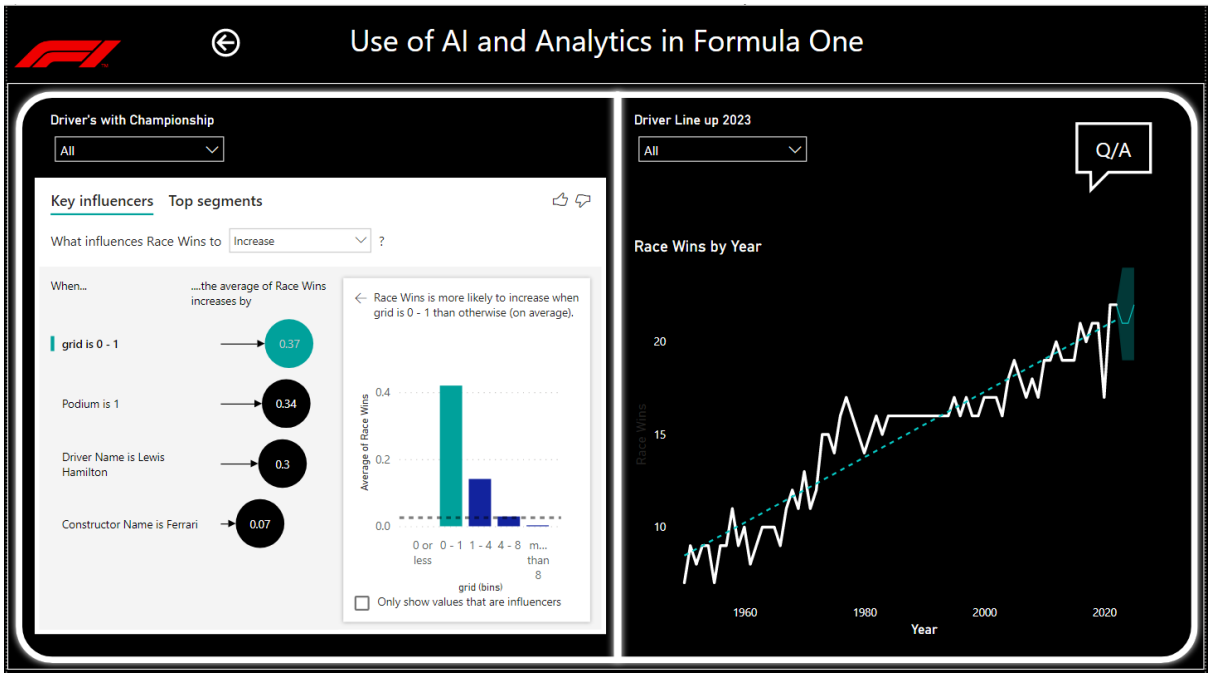
- This is the page analysing the Constructors or the team analysis.
- The animated sunburst is kept to the left as it is the most important chart in this page.
- To the right I have made a comparison between any two teams. (Only top 5 teams are selected).
- The cards and gauge axis move with respect to the input given.

Circuit analysis:



- The Circuit analysis is the 3rd page of the dashboard. The table and the decomposition tree chart are the most import visuals in this graph.
- The bar chart on the top right shows different causes of DNF are linked to the year and circuit.
- The pie chart shows the circuits by continent level.

Use of AI and Analytics in F1:



- Graphs in which AI plays a major role is kept in this page.
- It has been ordered with no specific organising technique in mind.

E. Self- Assessment

Use the table below to self-assess your work. This will help reflect on your work. You must keep this table in your report.

Report Section	**Description**	**Grade your work from 0 to 100**
Report Structure	The report is well-written, and it contains all the relevant sections	95
Data Pre-processing and Data Modelling	Many pre-processing steps have been applied. The data model is well-structured	100
Dax and M language	Both DAX and M Language have been **extensively** used in the report	100
Dashboard Design	The dashboard contains a variety of charts, including advanced ones not covered in the module.	95
Average		**Add below the average of the four cells above: 97.5**