How has the racing of Formula 1 changed over time?

Tom Hoad

(tdh1g19@soton.ac.uk)

I. Data Story Summary

My data story aims to explore the history of F1, looking at the dominant periods of drivers, exploring the greatest and lesser-known names, exploring the F1 calendar and trying to find greater meaning behind the technological revolutions of the sport. The audience should aim to guide those who have never seen the sport before and those quite familiar with it. F1 has a long history so even a fanatic may not have been alive for much of the sport.

The main narrative story pattern used is exploration. The desire here was to present the entire history of F1 data to the audience and let them dig deeper into the specific eras or years that interest them. The data story really fits this pattern because F1, as a sport, has a very storied history that most people will not have lived through the entirety so may be more interested in the bits they have not seen than the bits they have and already know about.

II. Dataset Summary

We used three main datasets, with some supporting data sources here and there. The main dataset discussed in the plan was the Ergast API [1] that provided a full dataset of results, lap times, championship standings and more. This made up the bulk of my data story and was used for most of the visualisations. We do however support this dataset with two others: a community made Ergast extension dataset [2] with additional information that Ergast does not provide an overtaking database [4]. We also used Wikipedia entries and the official F1 website [3] to verify these datasets to make sure they were up to date.

We performed a large amount of processing for these visualisations, covering the areas that these datasets do not. We provided the following extra data set tables:

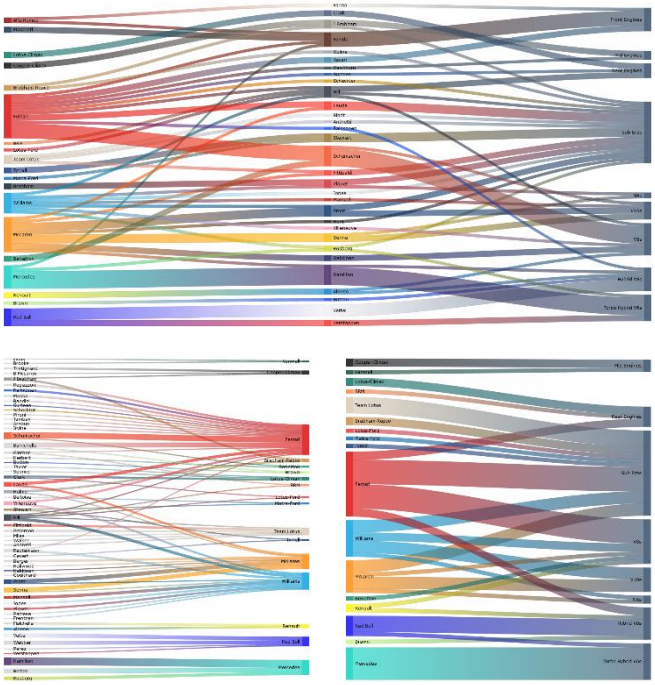
- Constructors and drivers' championships
  - Aggregated the championship standings tables into just the results of each year.
- Normalised results
  - Applied the 2022 points system to all years to make the data more comparable over time.
- Qualifying results
  - Calculated the best laps for each session.
- Race overtakes
  - Merged the Ergast and Overtaking datasets.
- Eras
  - A custom dataset cross referencing official F1 sources showing the years of eras of the sport.

III. Visualisations

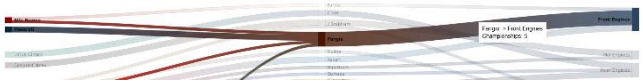
A. The Most Successful Drivers and Constructors Champions

1) Description

The first visualisation I used was a trio of Sankey diagrams representing the championships won by drivers and constructors throughout the history of F1. The visualisation uses one central Sankey diagram to represent the drivers and two supporting Sankey charts to represent the constructors. The diagrams show the number of championships won by the driver branching to the constructors they won them with and the sporting eras they were won in.



The diagram uses specific colour coding for teams and drivers' champions with their nodes being coloured based on either their driver helmet or team colours. The diagrams provide some functionality, nodes can be selected to easily identify their paths and when hovering over a link, more details about the exact number of championships will be provided.



2) Justification

The visualisation aims to familiarise the audience with key names of successful teams and drivers in the sport, showing which have been the most successful and the eras they were largely dominant in. The colour coding also helps establish a connection to teams and drivers, making it very easy to navigate the visualisation for an audience familiar with the sport already.

The Sankey diagrams are designed to remain as balanced as possible, especially with the constructors Sankey diagrams that began to hide data when connected. I made sure to give the graph enough space to make the links easy to visualise which is supported by the interactivity.

3) Narrative Design Patterns

The main narrative pattern used here is comparison. For each Sankey diagram we can compare the sizes of the nodes where the size corresponds to the number of championships won. We are also using concretisation here to turn several championships into a size and providing visual links to teams and eras that previously come out as tables or text descriptions.

4) Strengths and Weaknesses

Strengths:

- Provides a very overview of the biggest names in the sport.
- Sankey diagram makes for an interesting experience following paths to different drivers and teams.
- Use of colour is very prominent and makes navigation significantly easier.
- A good level of interactivity and can simplify the view nicely.

Weaknesses:

- Viewing the precise numbers is possible but not very clear at all.
- Can be quite overwhelming with excessive colours and too much constructor's information.
- Other than eras, no time frame references making it difficult to highlight when these championships were earnt.

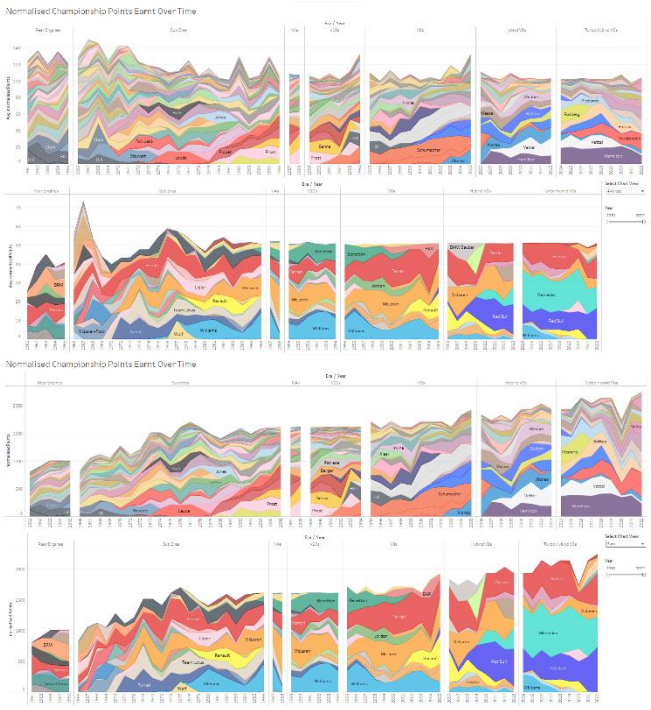
5) Improvements

I would improve this visualisation in two ways. Firstly, I would like to provide some more interactivity that allows the audience to filter by years to make the graphs more readable. Secondly, if I was to expand it even further, I would like to make this part of a circular Sankey diagram. I would have every single driver who has ever raced in the sport in an outer ring connected to every single constructor in the centre ring. I would use the same colouring pattern and mark championship winners with stars by their names.

B. Normalised Championship Points Earnt Over Time

1) Description

The second visualisation breaks down the first by exploring the championship points earnt over time by drivers and constructors. The two charts that make up the visualisation are area charts that show the area based on the number of points earnt in a year. The visualisation comes with a toggle to change between average points and the sum of points in a year. This allows us to see clearly how the number of races has increased over the years and how this gives us a more accurate average for recent seasons.



These graphs use the toggle as a bit of interactivity and come with a year range filter to look closer at different eras. The graph will label more drivers and constructors for shorter time ranges which can be interesting for comparing only a few years.

2) Justification

The aim of this visualisation is to allow the audience to see how dominant periods have occurred over time, where they have ended up, and how many uniquely dominant teams and drivers there are for a given period.

Using an area chart like this was the best way to visualise points data. It makes dominant drivers stand out significantly, and those with much less points get naturally filtered out. It doesn't waste much space on drivers with low points and will only label the most important drivers.

Once again, the user of colour coding is critical for differentiating teams and drivers, especially as I have used the same colour pattern as the first visualisation. Even from a long distance, certain teams and drivers are distinguishable.

3) Narrative Design Patterns

For visualisation B, we mainly use reveal, repetition, and exploration. We do this by showing the slow progress of driver and constructor success over time. We provide two settings for the visualisation (average and sum) to provide two different versions of the same repeated graph to solidify our conclusions through reveal and repetition. We also provide the filters to closer inspect certain eras to provide exploration.

4) Strengths and Weaknesses

Strengths:

- Significantly improves the time references to the driver and team successes.
- Interesting layering can produce interesting patterns and makes it very easy to identify certain entities.
- Similar strengths as visualisation one with the use of colour.
- Interactivity gives a very good insight into season specifics.

Weaknesses:

- Use of normalised points meant I had to eliminate pre-1961 data due to excessive drivers and teams.
- Some calculations don't seem 100 per cent correct making it somewhat inaccurate if you dig into the numbers.
- Use of colour looks worse for entities given default Tableau colours.
- Switching between average and sum mode doesn't give much of a reveal.

5) Improvements

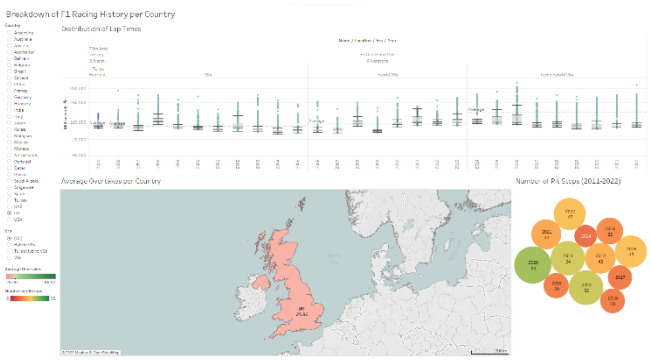
I think there are a few ways to improve this visualisation. Whilst the normalised points have made the graph readable, I would like a better way to show how and when the point systems have changed over time. There are also several events like sprint races that haven't been properly accounted for and could make the more recent years more interesting.

Overall, the biggest change to this visualisation I would like to see would be to increase the space used. Rotating it vertically and having it displayed in a bigger space would make traversing the visualisation more interesting and wouldn't require the years filter to provide more resolution. Most of all, this would allow me to mark more key dates along the timeline that people may find interesting.

C. Breakdown of F1 Racing History per Country

1) Description

My third visualisation is exploring the racing history across the globe. The visualisation heavily relies on the audience's own exploration using the given interactivity to select different countries to view the data provided. For each country, multiple different graphs are provided: a box plot breakdown of lap times over the years, the average number of overtakes per country shown as a map and a bubbles diagram showing the number of pitstops per race.



2) Justification

The aim of this visualisation is to allow the audience to explore the globe and take quite a wide viewpoint and dig deep into the individual races that take place. The key interactivity provided by this visualisation is the country and era filter. This allows the audience to select different countries which will automatically update the three graphs to display the info for that specific country. The box plot allows you to view individual lap times, box plot specifications and colour code tracks to identify changes in location. The map is colour coded to show the number of overtakes compared to the rest of the world. The pitstops are also colour coded by the number of each year.

3) Narrative Design Patterns

The pivotal narrative design pattern here is exploration. The entire visualisation hinges on the audience's use of it as there is no main page to this visualisation, only the exploration the users perform to learn more about the countries.

4) Strengths and Weaknesses

Strengths:

- Interactivity here is essential and will give the audience a lot of time to explore and dig deep into the data.
- Use of colour is also quite useful here and can quickly highlight different tracks in the same country.
- The story this visualisation tells will likely be different from user to user.

Weaknesses:

- Does not provide a global view as it crammed too much into a small space and became unreadable.
- An audience that doesn't use interactivity fully will miss out on a lot of interesting findings.
- Once again, reaching the limit of my data set as very limited years for lap times and pit stops.

5) Improvements

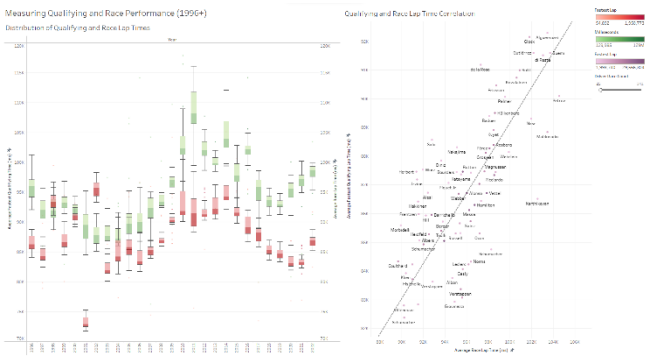
This visualisation was the one I was happiest with, but there are still a few things that could be improved. I would like to use some of the space taken up by the map and provide a

couple of interesting stats such as the most common winner, the most common pole sitter, the number of safety car incidents and the number of fatalities. I think this could provide a unique look into track safety as well as help identify the years that were raced at the tracks. By far the largest improvement that could be made would be to provide a global view option. This may have to be simplified to reduce the clutter on the page but could be better for comparing countries easier.

D. Measuring Qualifying and Race Performance

1) Description

The fourth visualisation dives deep into the raw qualifying and race lap times, using two charts to make up the visualisation. The first is a composite of two box plot graphs that measure the lap times from 1996+ onwards. The second chart shows a scatter graph of the average race lap time vs the average qualifying lap time for each of the drivers.



The second graph provides zoom controls to further explore the data easier and make more labels visible at closer resolutions. Also, there is a slider that will allow the audience to filter out drivers that haven't raced X number of races which the audience will find interesting in seeing how the line of best-fit changes.

The aim of this visualisation is to compare the performance of lap times to expose the raw difference in car performance over time as well as to use this data to further show the dominance of certain drivers.

2) Justification

Whilst my final visualisation was not created to show the viewing figure data as I initially intended, I feel as if the messages conveyed using the same graph are still well conveyed. The use of dual axis box plots meant I could start to merge the data I had to make comparison even easier. We hide most of the data beneath the box plot itself but still display a few outliers of interest.

The line graph requires a lot of exploration to fully appreciate but is the best measure of comparing drivers. We start the graph with the race filter set at 25 so most of the null data is ignored but can be viewed if those are interested. This filter can heavily influence how the graph appears.

3) Narrative Design Patterns

The final visualisation uses exploration and reveal once again for exploring the differences in drivers and cars. This is shown in how the box plot graph can be traversed and how the line graph can be moved and modified as well.

4) Strengths and Weaknesses

Strengths:

- The closest visualisation to providing some sort of deeper meaning showing a definite correlation/pattern over time.
- Once again allows the user to explore using the interactive graph and filters.

Weaknesses:

- Lack of audience figures and attendance records makes this visualisation a lot more generic than I had hoped and planned.
- The resulting message of the story is more a guide through time than a larger message.
- The data for lap times is once again heavily constricted by a lack of data. If lap times had been recorded since 1950 then these graphs would be significantly more interesting.

5) Improvements

My final visualisation went in a different direction than I had planned and provided the beginnings of what I hoped would be a larger discovery. The biggest change I would do to this visualisation is to incorporate track attendance and TV viewing figures. Unfortunately, getting this data is inconsistent, unreliable, inaccurate and has so much missing data. F1 does not provide these metrics and am going off random numbers provided by biased articles that often round up and down. I believe this would give me a much more interesting conclusion to the story if the data was made publicly available.

IV. References

[1] Ergast Developer API. (2023). Retrieved from: http://ergast.com/mrd/db/ (Accessed: November 18, 2022)

[2] Formula 1 Race Events, Kaggle (2023) Available at: https://www.kaggle.com/datasets/jtrotman/formula-1-race-events (Accessed: November 18, 2022).

[3] Formula 1 Results (2023) Available at: https://www.formula1.com/en/results.html (Accessed: November 18, 2022)

[4] Overtaking Database (2023) https://docs.google.com/spreadsheets/d/1XueNI7ZawEX0R

LDq5dAGVqsEb1-DBOK2kUWGwM1OMKs/edit#gid=0 (Accessed: December 28th, 2022)

V. Appendix

A. Molly - 18

1) Persona

Molly is 18 years old. She has lived in Bournemouth almost all her life and has grown up with three older brothers. They have always loved watching F1 growing up and Molly has only watched the occasional race with them. Now Molly's brothers have gone to work/university so likes to watch F1 as it reminds her of spending time with them. She's interested in studying geography at university next year but is worried about leaving home.

2) Scenario

Molly's brothers are coming home from university and work for Christmas, and she wants to impress them with her knowledge of the latest F1 season. However, she is worried they are going to start comparing it to previous seasons and doesn't want to get lost in the conversation. Molly is looking for a visualisation to find out who the most successful drivers and teams are.

3) Use Case

"Hey sis, have you enjoyed watching the F1 whilst we've been away?" "Yes, I have, I've been trying to learn a lot about it whilst you've been away?" *"How've you done that? It's quite boring to read through lots of Wikipedia pages?"* "Well, I watched a few races, but I also found this great infographic that showed a network of championships earnt by drivers and constructors. It had a full breakdown all the way from 1950 onwards!" *"Wow that's really cool, did it help you learn about the current drivers a bit more then?"* "It did indeed. I might even know a thing or two more than you!"

B. Graham - 74

1) Persona

Graham is 74 years old. He grew up in West Bromwich and moved to Sheffield in the 90s. He married his wife Margery in 1971 and they have lived happily together ever since. Graham used to work as a car mechanic before he retired a decade ago. He now likes to tinker with his own classic Mini whilst his wife owns a brand new Electric mini. Graham and Margery love to sit down on a Sunday afternoon and watch the Grand Prix. They have both watched F1 for much of their life but are struggling to understand all the new technology in the sport.

2) Scenario

Graham is going to visit his son in a few weeks for a Sunday roast and is keen to watch the F1 race happening that weekend with him. The two of them love to tease each other and get into heated debates about F1 as they used to when Graham's son was younger. However, Graham doesn't understand all the technical insights of the sport and is hoping to impress his son by knowing about all the technical eras of the sport and having fun telling him that racing was better in the 80s.

3) Use Case

"Hey son, it's time for the GP!" *"Okay, on my way! Who's starting on pole?"* "Err, Lewis, I think... I miss the days of Senna; the tracks and racing were so much better." *"You're kidding right, the racing this year has been so much better than before."* "They are today, but since you were born, the racing has only gone downhill, and the tracks have moved for money over racing." *"How do you know that?"* "Well, I found this cool graph on the web, it gave a really thorough breakdown of the racing over time, and it shows how the gaps between cars have been getting bigger!" *"Wow I didn't realise, it's hard to tell because we don't really compare that much detail of the grid to the past."* "I'll send you a link, I know you'll like it."

C. Tony - 31

1) Persona

Tony is 31 years old. He grew up in Norwich and has recently moved to London for work. He lives with his new girlfriend Olivia who works with Tony as a lawyer. They work very long days and don't get much time during the week to hang out together. They relish their weekends when they get a few sparse minutes to watch the Grand Prix, which they started watching about 10 years ago and have been Red Bull fans since.

2) Scenario

Tony has recently started an F1 fantasy league with his mates and wants the inside scoop on who is most likely to win in which car. He wants to fully understand the best teams in F1 and calculate how easy it is for drivers to overtake in each race to determine his fantasy team.

3) Use Case

"Great meal Tony, your Sunday roasts are the best!" "You're welcome, nothing better than watching a Grand Prix then having a roast!" *"But wait Tony, I have a question. How is your F1 fantasy team doing with your friends?"* "I've done very well this week. I got prepared for the Dutch GP by doing some proper research. I found several useful visualisations that gave me the edge in selecting my team." *"What did the infographics show you?"* Well, there was a really great breakdown of how many overtakes happen in the Netherlands and I used that to rely on those cars with great race pace who I knew would get overtakes." *"Wow that's very clever. Maybe I should start a fantasy league with my friends and use it to beat them."*