How has the racing of Formula 1 changed over time?

Tom Hoad

(tdh1g19@soton.ac.uk)

# 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.

# 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 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 with additional information that Ergast does not provide an an overtaking database. We also used Wikipedia entries and the official F1 website 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 final 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 particular eras of the sport.

# Visualisations

## The Most Successful Drivers and Constructors Champions

### 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.

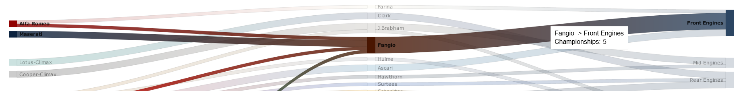
A screenshot of a computer

Description automatically generated with low confidence

Diagram

Description automatically generated

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.



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.

### Justification

The design of this visualisation is to really start the story by showing where the connections between drivers and constructors. Whilst they are separate entities, periods of the sport have been dominated by a specific driver and/or constructor. It's also good for directly comparing drivers from different periods. Whilst I am using the story to show how racing over time has changed, I need to also identify the key characters in the sport and compare the recent champions to those of the past. Attaching eras to the diagram was also useful to help familiarise the audience with the different eras, which ones had the most unique winners and their relative lengths of them.

The colouring of drivers and teams was a very important part of this visualisation to make certain teams such as Ferrari stand out. Even for those who haven't ever watched before, the association of the red is instantly recognisable and will help set the tone for the rest of the story.

I decided on a Sankey diagram because the interactivity of following a path and seeing where it leads is very interesting and lets the audience self-explore beyond what may be necessary for the story's aim.

### Narrative Design Patterns

### 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.

### 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 a number of stars by their names. I would remove the eras from the original Sankey and use the year's filter to show eras and a toggle to show only drivers and constructors champions.

## Normalised Championship Points Earnt Over Time

### 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.

Chart, diagram

Description automatically generated

Chart, diagram

Description automatically generated

An important calculation that I had to do for this visualisation was to normalise the championship points. Over time, the points system has evolved over time and hence meant that the graph ended up being quite skewed and was hard to compare progression over time. I computed all the past results using the current modern-day points system from 1961 onwards.

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. Another important part of this visualisation is the continuing use of colour, matching some of the constructors and drivers to the colours used by their helmets (and by extension the same colours as in visualisation 1.)

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.

### Justification

### Narrative Design Patterns

### 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.

### 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. I could also expand the timeline to show points earnt at every single race over time rather than at every year's end.

## Breakdown of F1 Racing History per Country

### 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.

Map

Description automatically generated with low confidence

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.

### Justification

### Narrative Design Patterns

### 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.

### 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.

## Measuring Qualifying and Race Performance

### 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.

Chart

Description automatically generated

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.

### Justification

### Narrative Design Patterns

### 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.

### 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.