The heart of a car is its engine

Visualising Engine Effect on Formula 1 Race Cars

# Abstract

Formula 1 (F1) racing officially started in 1950 and has seen over a hundred different teams compete over the years. F1 has truly become a data driven sport, with current day F1 teams having huge teams of experts monitoring every available car sensor throughout the race to formulate the best winning strategy. The visualisations show that engines have a big part to play in performance and overall success. Engine regulations have had significant effects on car speed across the grid. In addition it’s clear that Mercedes is the engine manufacturer that got the modern day V6-hybrid engine right, dominating the field since 2014.

# Dataset

The main data source of F1 grand prix race statistics since 1950 was provided by Kaggle[[1]](#footnote-1), which as of today consists of 17.53 MB of data from 13 different tables. Additional information to supplement the analysis was scraped from Wikipedia[[2]](#footnote-2).

Velocity and variety are the most prevalent characteristics of big data in the datasets. The data is updated with new race statistics after every F1 grand prix, which take place on average 20 times per year. Variety is present as additional information found on Wikipedia involving course lengths and engine suppliers was merged to help answer the research questions.

The data set is not overly large, hence would not be considered to have substantial volume, though it has a notable amount of complexity attributed to the requirement of joining numerous tables to conduct the analysis.

The content of the Kaggle data set is extensive, the most granular data being individual lap and pit stop times, followed by results and cumulative rankings after each race. This data is then supplemented by multiple tables of qualitative metadata on drivers, constructors, circuits and seasons. Examples of available metadata are names, nationalities, locations and URLs to dedicated Wikipedia pages.

Race lengths were retrieved by using the provided Wikipedia URLs to each individual race and searching for the table on the page with the race metadata on course length. Race lengths can change over the years, hence it would not be sufficient to scrape course lengths from the general page of the circuit. In addition, the engine supplier for each constructor since 1996 was obtained in a similar fashion by searching for the appropriate table from each F1 season Wikipedia URL.

The quantitative data I focused on was individual lap times, course lengths and race results, whilst the descriptive data was provided in the engine, driver, constructor and race metadata tables. Lap times was the largest table used consisting of 491,000 rows of data followed by the results table with 25,000 rows.

# Data Exploration, Processing, Cleaning and/or Integration

#### Data Cleaning and Integration

Race data is available since 1950, however individual lap data has only been recorded since 1996, hence I made the decision to focus the analysis on the 1996-2020 time period.

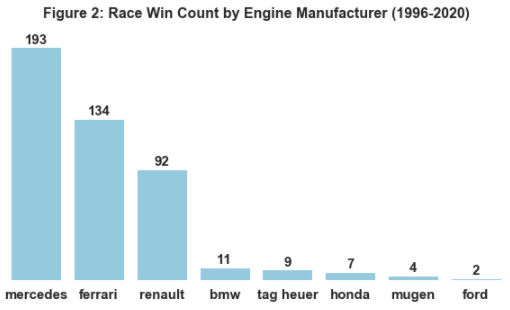
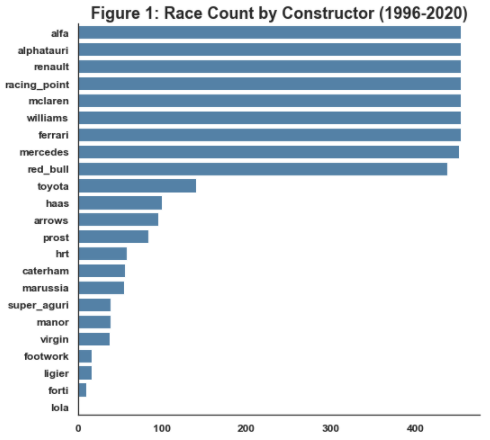
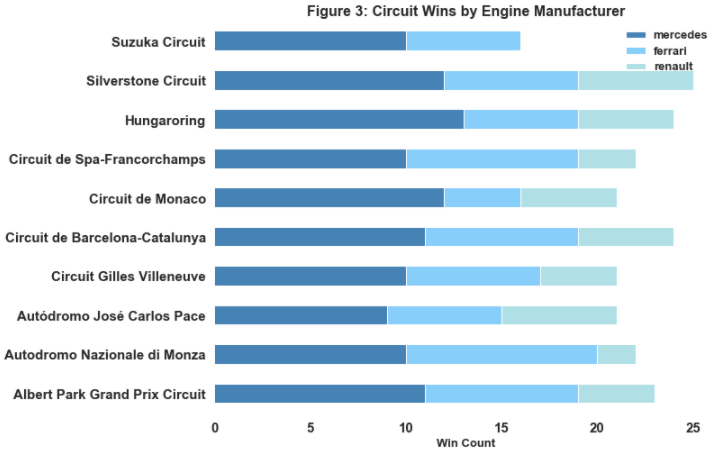
The data scraped from Wikipedia required most of the time in terms of data cleaning, this mainly consisted of splitting, cleaning and removing unwanted text from the scraped tables. The course length data was the simplest to merge with the Kaggle data as a direct link to the race was provided, with only four rows needing to be filled manually due to errors in the Wikipedia links returning inconclusive results. The engine data was trickier to merge as the scraped constructor names needed to be in the same format as the names I had available from Kaggle. I approached this by joining as many constructor names as possible with the available data, then created a small dictionary that could be used to map the constructor names on Wikipedia with the ones I had available from Kaggle.

#### Data Transformations

One of the main challenges that arose once starting data exploration was that constructor names change regularly in F1 which makes analysing constructor performance over time difficult. Hence I used another dictionary that mapped any old constructor names to their current name using team history names from Wikipedia[[3]](#footnote-3).

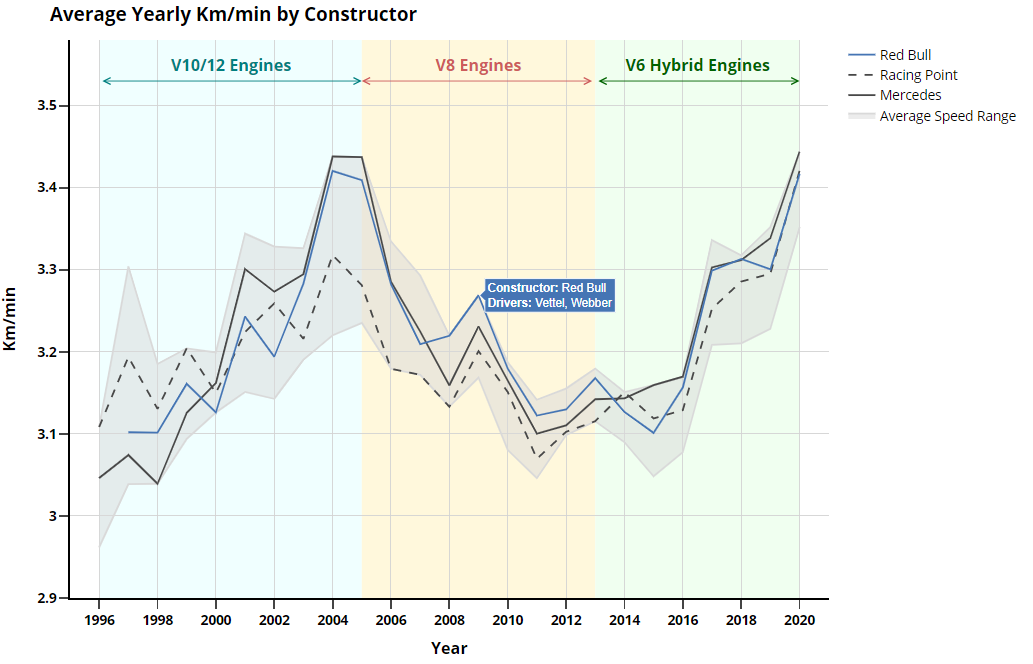
#### Data Exploration

There isn’t a limit on the number of F1 teams in a given year, the 2020 season consisted of 10 constructors but there have been others since 1996. Since my intentions are to analyse team performance over the 25 year period, I want to focus on the teams with a consistent amount of data over this time. From figure 1 it is evident that 9 out of 10 of the current F1 teams have taken part in the most races since 1996. Haas is the only current team with significantly fewer races since it made its debut in 2016. Hence, for time comparisons I choose to focus just on the ten current F1 teams.

Though there are multiple F1 teams, they don’t all manufacturer their own engines. Often teams share engine manufacturer, yet performance can vary a lot. Hence, I wanted to examine engine manufacturer performance over the years, but from figure 2 it is clear since 1996 there are really only three engines worth comparing in terms of winning results. Mercedes, Ferrari and Renault have produced by far the most winning engines since 1996, hence I just focus on these three in my analysis.

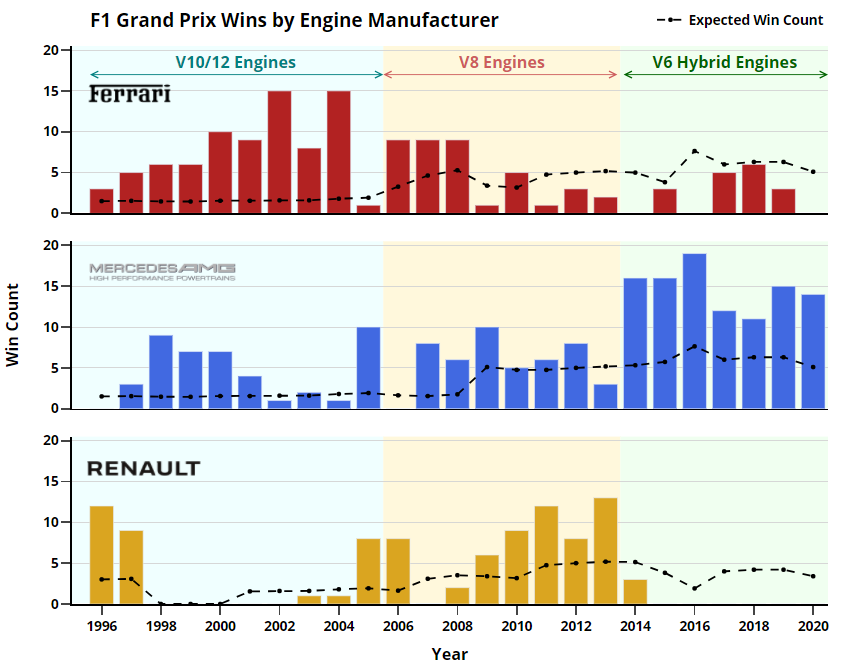
I was also interested in seeing if certain engines perform better in certain circuits, but focusing on the circuits which have occurred at least twenty time in the 25 year period, the distribution of win count between the three engine manufacturers didn’t appear too interesting, with Mercedes always performing the best, usually followed by Ferrari (Figure 3). Hence, I decided not to analyse circuits further.

# Visualisations

For the visualisations I decided to focus on average lap speed since 1996 for the ten constructors. When initially plotting the change in speed over time, the result was surprising as my general instinct would be that cars are getting faster not slower. The big dip in speed between 2005 and 2013 was very interesting, even more so the fact all teams follow a similar trajectory over the entire time period. A bit of research showed changes in engine regulations occurred exactly during the period the trends in speed changed. To really emphasise this I shaded the backgrounds associated with the time period for each specific engine regulation.

I wanted the viewer to be able to explore the visualisation further, hence I introduced interactivity using Plotly and Dash. Plotly allows for the option to hover over data points and reveal more information, in this case I wanted the user to see who the drivers were for a given constructor and year. Furthermore, it is possible to zoom in on certain areas of the plot, making small discrepancies in speed easier to see. In addition I incorporated filters with Dash so the user could analyse specific teams of interest, and see which teams were faster on average each year. Having all the teams on the plot at the same time makes it difficult to examine, hence the Dash web page will open up displaying results for the top three teams of 2020, as shown in the image, with the option to add or remove constructors as desired. I included the shaded region for range in max and min average speed because I found the fact all teams have similar trajectories quite interesting, hence this shaded region shows that without having all the lines on top of each other. In addition it makes it easy to compare teams to the max and min over time.

I used a combination of line types and colours here to make the data easier to compare. The colouring system is based on the fact there are often engine-manufacturer teams and engine customer teams. For example Williams and Racing Point are both customer teams of Mercedes. Hence manufacturer teams are assigned solid lines whilst customer teams have dashed or dotted lines in the same colour as the manufacturer team.

The initial motivation for my second visualisation came from realising the effect engines can have on performance. I wanted to compare engine manufacturer performance over time, especially to see if the certain time period regulations seemed to have an effect. I contemplated for a while the best metric to use for measuring performance, one option is using points; points are awarded to the top ten finishers of each race, the higher the rank, the higher the points. However points will reward

situations when a particular engine manufacturer has more engines on the grid compared to others, resulting in more chance to score points. Hence, this led me to believe race win count would be a better measure of performance, plus expected win count would be much easier to calculate than expected point count.

I think one of the main messages from this visualisation is that Mercedes engines are not only dominating since 2014 (same period as hybrid era) but very rarely perform worse than expected. In addition there is a middle time period where the three manufacturers seem to have somewhat equal performance, this makes me think the V8 engines might have levelled the playing field.

It’s obvious from this plot that Ferrari has been struggling significantly since 2010, never performing above expected. Expected win count was calculated by dividing the sum of all starters with a particular engine across all races by the total number of starters in a year, then multiplying this by the number of Grand Prixs that year. It is essentially giving an equal opportunity to every engine to win an amount of races proportional to the number of cars on the grid with that engine.

I split this visualisation into 3 separate bar charts because having them side by side or stacked would be very difficult to interpret given the time period is quite long. I added sub titles to each individual subplot to remove the need for a legend, making it much easier for the viewer to cross reference plots. Though doing this meant separate colours for each plot aren’t needed, I thought it would help the viewer compare all three engines at once. A simple black dotted line is used to show how performance compares in terms of actual win count vs. expected win count. This visualisation is also interactive, where hovering over a data point highlights specific information. I added the number of teams with a specific engine to the expected win count line to add a bit more context.

For both visualisations the axes and text were kept simple and bold. Grid lines were used to help the viewer guide their eye to axes values. Colour pallets were chosen to allow viewers with colour blindness to distinguish between shades. I put both visualisations on the dash plot so it was easy to compare both side by side. Since they both have the same x-axis for year, I made sure they were in line with each other to keep the dashboard looking tidy.

#### Tools and Libraries

PyCharm was used to do the brunt of the data scraping, cleaning and Dash app design. Jupyter notebooks were used to guide data exploration and visualisations.

Plotly and Dash were used for both visualisations.

# Conclusion

There are truly a lot of different metrics and metadata about F1 teams that can be visualised; my biggest struggle was figuring out what the most important information was and how to deliver it without overwhelming the viewer. The line plot was the most difficult in terms of data clutter, as there are quite a few messages being displayed with the various shades and lines. I would have liked to add the option to deselect or select all constructors in the Dash plot but could not figure out how to get this functionality or if it’s possible with Dash. In addition I don’t like having the symbol for average speed range in the same legend as the constructors but could not find a better way to display this.

The bar charts are simple, but clear and display a lot of information. My main concern is the way expected value is calculated may be a bit obscure to a viewer with a non-technical background. I would have liked to do a more unique visualisation other than line and bar charts but their simplicity is the reason they are so popular for data visualisations. Overall I am particularly happy with the results and exploring this data was very interesting, each plot making me more curious about another aspect to explore. In addition it was ideal the World championship for 2020 ended a week before the assignment submission, allowing me to update the graphs with the latest data for the year.

# References

#### [Kaggle Formula 1 data](https://www.kaggle.com/rohanrao/formula-1-world-championship-1950-2020)

#### [List of Formula 1 constructors and old team names](https://en.wikipedia.org/wiki/List_of_Formula_One_constructors)

#### [History of Formula 1 engines](https://en.wikipedia.org/wiki/Formula_One_engines)

#### [Resource for catering to colour blindness](https://towardsdatascience.com/two-simple-steps-to-create-colorblind-friendly-data-visualizations-2ed781a167ec)

#### [Dash interactive graph tutorial](https://dash.plotly.com/interactive-graphing)

#### [Plotly tutorials](https://plotly.com/python/)

1. <https://www.kaggle.com/rohanrao/formula-1-world-championship-1950-2020> [↑](#footnote-ref-1)
2. <https://www.wikipedia.org/> [↑](#footnote-ref-2)
3. <https://en.wikipedia.org/wiki/List_of_Formula_One_constructors> [↑](#footnote-ref-3)