|  |  |
| --- | --- |
| **Programme** | BSc (Hons) Data Scientist |
| **Module name** | Data Science Professional Practice |
| **Schedule Term** | Jun 2024 |
| **Student Reference Number (SRN)** | BP0290098 |
| **Report/Assignment Title** | Evaluating Formula 1 Drivers; The Impact of Data Science Within Stantec |
| **Date of Submission**  ***(Please attach the confirmation of any extension received)*** | 28/08/2024 |
|  | |
|  | |
| Declaration of Original Work:  I hereby declare that I have read and understood BPP’s regulations on plagiarism and that this is my original work, researched, undertaken, completed and submitted in accordance with the requirements of BPP School of Technology.  The word count, excluding contents table, bibliography and appendices, is [2868] words.  Student Reference Number: BP0290098 Date: 28/08/2024 | |
| By submitting this coursework you agree to all rules and regulations of BPP regarding assessments and awards for programmes. **Please note, submission is your declaration you are fit to sit.**  BPP University reserves the right to use all submitted work for educational purposes and may request that work be published for a wider audience.  **BPP School of Technology** | |

## Contents

[Contents 2](#_Toc175760457)

[Evaluating Formula 1 Drivers 3](#_Toc175760458)

[Executive Summary 3](#_Toc175760459)

[Infrastructure and Tools 3](#_Toc175760460)

[Engineering 4](#_Toc175760461)

[Visualisation and Dashboards 4](#_Toc175760462)

[Analytics 5](#_Toc175760463)

[Evaluation 5](#_Toc175760464)

[Documentation 6](#_Toc175760465)

[References 7](#_Toc175760466)

[The Impact of Data Science Within Stantec 8](#_Toc175760467)

[Executive Summary and Project Background 8](#_Toc175760468)

[Analysis 8](#_Toc175760469)

[Evaluation 9](#_Toc175760470)

[Documentation 10](#_Toc175760471)

# Evaluating Formula 1 Drivers

## Executive Summary

Formula 1 is a complex and competitive sport with a long and rich history. Since the first Grand Prix at Silverstone in 1950, there have been 115 Grand Prix winners of the 776 total drivers in the 1,116 FIA official Grands Prix that have been held. (Contributors, 2024) As a franchise, the sport is worth billions, with numerous billion-dollar companies sponsoring the various teams that compete. (Ozanian & Knight, 2023) The teams themselves compete for a multi-million-euro prize pool while they themselves gain recognition through the marketing of their own brands. While the teams and mechanical engineering play arguably the most important part in winning races and championships, the drivers are imperative in getting the car across the finish line as fast as possible; naturally, this presents an interesting question: who is the greatest driver? This becomes an incredibly important postulation for teams as the more competitive they are, the greater both their prize money and their attractiveness to potential sponsors – an obvious business incentive. Unfortunately, this is far from an easy exercise, as it is incredibly difficult to measure the pure skill of a driver when there are so many other factors at play: the speed of the car compared to the others of the time; the speed of the car compared to the cars of different eras; the comparative speed of the driver’s teammate; the number of races a driver partook in throughout their career; etcetera. The aim of this project is to use various metrics to determine which driver throughout the sport’s history is the greatest, as well as which driver’s currently competing in the sport are the best among their peers. “Best” is a highly subjective quality, which makes this exercise somewhat difficult; one could define “best” as the driver with the most wins (which would be a far more trivial question: Lewis Hamilton has 105, more than any other driver in the history of the sport) (Donaldson, 2023) but I believe a far better definition would be the driver with the greatest capacity to win a championship, as this is the ultimate sporting goal of any given team, and disregards the bias of a faster car.

## Infrastructure and Tools

The dataset chosen was well-maintained, freely available bundle of CSV files on Kaggle, each file having a slightly different structure to portray different data; one such file provided results for every driver for every race, with the total championship points of each driver after the race result, meaning the data did not explicitly say how many points each driver achieved after a given race. The formatting of this data makes a tool like Power Query rather difficult to use and complete analysis with, as metrics such as the points gained over a teammate in a given race become very difficult to calculate in that language. As such, the primary engineering and analysis tool used was Python, more specifically the Pandas library, which allows for more intricate analysis and the creation of metrics specific to the dataset to be calculated, as well as an output formatted in whatever way proves to be most convenient for later use. Its fast processing speeds, readability, and ease of use for someone with basic programming experience proves Pandas to be an ideal platform for this project. Later visualisations and formal presentation of findings can be done in Power BI, using that tool purely for its dashboarding capabilities, rather than for real data engineering. Given the in-built CSV-handling methods in the Pandas library, as well as the smaller size of the data, it would be unnecessary if not a little redundant to store this data in any more advanced or sophisticated manner, such as a relational database. Were the data not publicly available (let alone easily accessible and, in some cases, widely known), a more sophisticated approach may be justifiable for the sake of data governance; however, the data used in this project is open-source and not at all sensitive, and therefore does not need such security. Given this, the most convenient format for my analysis, the original CSV formatting, was used.

## Engineering

As previously stated, the dataset used is well-maintained and presented as a collection of CSV files, each with a unique yet consistent structure. Initial investigations into the relationships between the various files of the dataset, a general overview of contents and scope, as well as a quick check for accuracy, were done in Power BI, which allowed for quick establishments of said relationships and digestible visual looks at the dataset as a whole. This method made clear the structure of the CSV files as well as the completeness of the entire dataset; for some historical drivers, fields like Racing Number and Abbreviation (the three characters used to identify each driver for the audience during a race) were mostly scarce, as this was only available and relevant for more modern drivers, although this incompleteness only means data is missing on rather cosmetic fields, and is not relevant for our purposes. The most significant field containing incomplete data was Milliseconds (the number of milliseconds a driver took to finish a race from its beginning), which only contains data for drivers that finished the race on the leading lap; those who may have finished a race one or more laps behind would have misleading Milliseconds values which would imply that said drivers finished in a higher position than they did, and so the creator of this dataset seems to have elected to omit this data as it would not be useful to include. Otherwise, the neat structure of the dataset made the cleaning process a more manageable task than if it were a less tidy dataset, as the only real issue was the formatting of null values; Pandas processes empty values in a CSV as the Numpy (another Python library) Not a Number (NaN) value. This seems trivial at first, especially since Pandas is not incorrect when making this assumption, until it comes to iterating over the dataset and performing arithmetic calculations with these NaN values, which produce cascading errors. By replacing all empty values with the Python Null value, the iterations instead simply skip over the empty values in an elegant way that avoids compounding errors in order to produce the desired result.

## Visualisation and Dashboards

Given that the ultimate goal of this project is to score each driver in Formula 1’s history and rank them from “best” to “worst,” the visualisations required for the final presentation of the output data are not complex. A bar graph showing the top few drivers’ scores would neatly display the best drivers and how they compare to one another in a way that is easy to read and provides an intuitive understanding of how the scores differ between drivers of similar calibre. A histogram, boxplot, or other similar visualisation to portray distribution would communicate the variation in scoring nicely and may provide some holistic insight into the output data that is not immediately apparent when viewing the raw results. Slicers would also help to give a more user-oriented experience that tailors the results shown to their needs, such as the ability to filter to just Grand Prix winners or perhaps champions, as well as filtering by only modern-day drivers. This would be beneficial for the business case of finding the best current driver for a given team and allows for the data to help cater to the different needs of said team.

## Analytics

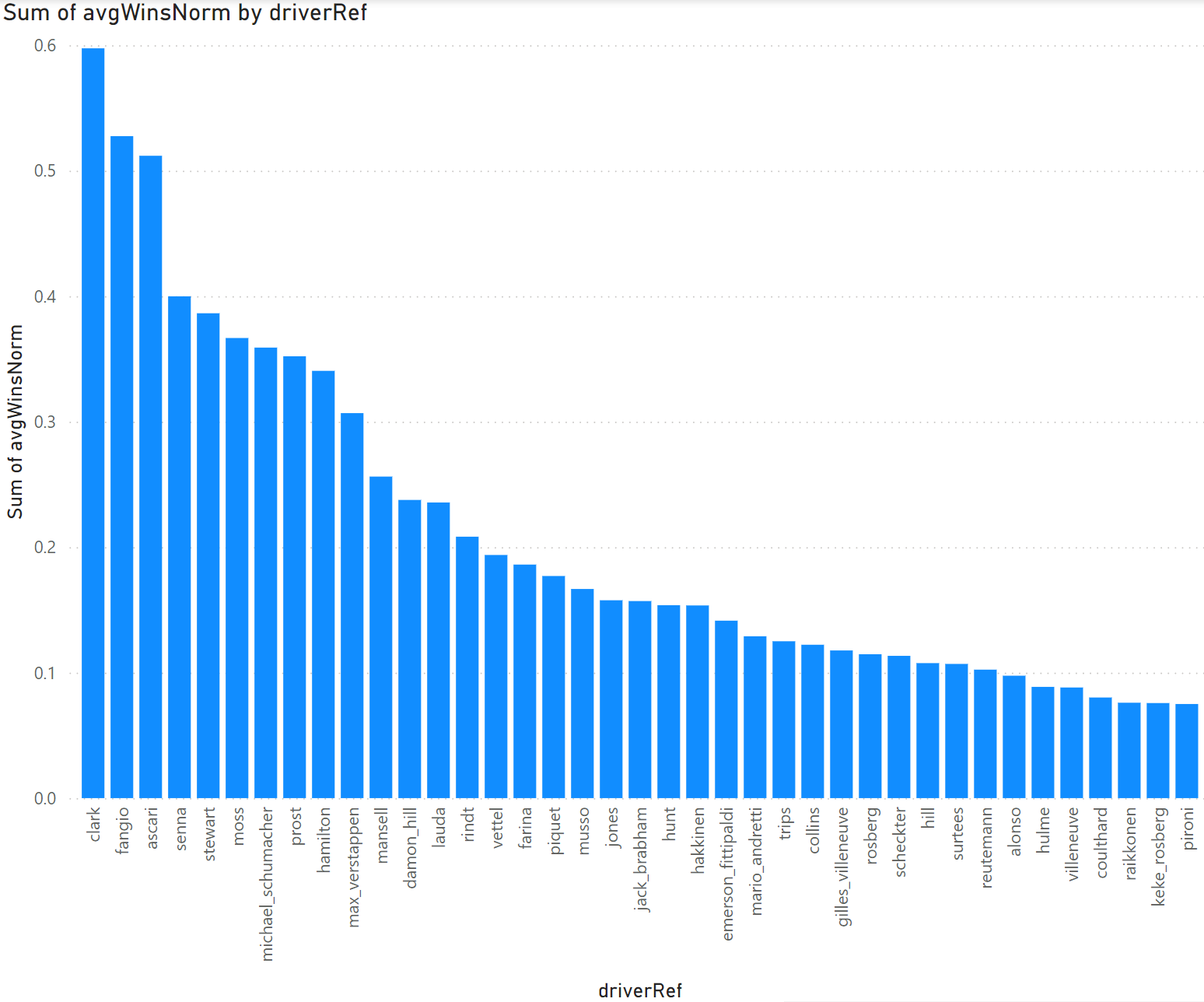
With a dataset as unique and non-traditional as that of the complete collection of Formula 1 results, it is difficult to find an analytical tool that fits the needs of this project. When attempting to apply a linear regression to the dataset, you first have to choose which two variables to compare in the first place, which is not a trivial task. If comparing total wins to total championships, you will find some correlation (most datapoints sitting at (0, 0), while those with the most wins tend to have the most championships), but this is much to be expected, as race wins are all but required to win a Formula 1 championship, which introduces a direct causality into the relationship that provides no new insight. Instead, as someone with existing knowledge of the sport, I chose to apply what I believed to be the best metrics for determining the greatest drivers: the average delta between the finishing time of a driver and their teammate normalised by the best finishing time of that race (this removes both bias from the relative speed of the car compared to others of the same time by comparing only to the driver’s teammate, as well as relative speed of the car compared to others of different eras of racing by normalising to a consistent benchmark relevant to that timeframe); average wins per race, counting only the races that the driver finished (this removes the bias of unreliable cars or other factors that may cause a driver not to finish a race that are out of their control as well as eliminating drivers who may have only entered one or two races and beaten their teammate, but fails to account for some drivers having faster cars than others); average points per race, applying the current points scoring system employed by the FIA and discounting races where the driver did not finish (accounts for drivers consistently scoring points while perhaps having less wins, which is important for a championship).

Notably, as I am an avid enjoyer of motorsport and, particularly, Formula 1, the results are inevitably influenced by my own inherent bias. I maintain personal opinions on which drivers I believe to be better than others, and when these opinions are not supported by my analytical outputs, I may adapt the outputs to better suit my assumptions. This has its benefits and drawbacks, as it allows me to see when an output is quite obviously erroneous, such as if a driver who has only completed 3 races and placed no higher than fifth is the greatest racer in history according to my algorithms. This approach is imperfect, however, as it benefits the drivers whom I personally prefer. The ideal scenario is for someone familiar with the sport, or at least with its rules and the dataset, but with no bias or vested interests in the results to undertake this project so as to produce results free of unfair interference on the side of the data scientist.

## Evaluation

The final results are imperfect. According to the analysis completed, drivers who have primarily, if not only, competed in the Indianapolis 500, a race that used to be part of the Formula 1 calendar during the ‘50s, are the best Formula 1 drivers in history. This is because this race is almost an entirely different sport, and so the metrics used favour drivers who did exceptionally well in this race, even if they have not competed in any other Formula 1 event. If these drivers are to be excluded, the next best racer is Alberto Ascari, a 2-time champion who raced in the early years of the sport and is considered to be one of the all-time greats. (Donaldson, n.d.) This again highlights a bias in the methodology used towards earlier competitors, when the margins in the sport were march larger and people tended to win by significantly greater amounts compared to the races of today. Were I to complete this exercise again, I would seek to find a better metric to compare these drivers such that every era of the sport has an equal opportunity to produce a racer that could be considered the greatest of all time.

## Documentation



A quick visualisation of the output results, this bar graph shows the average wins of each driver normalised by the number of races that they finished, then filtered to only show drivers who have led or won a Formula 1 championship during their career. According to this graph, Jim Clark won significantly more of the races that he finished than other results.

# References

Contributors, W., 2024. *List of Formula One Drivers.* [Online]   
Available at: https://en.wikipedia.org/wiki/List\_of\_Formula\_One\_drivers  
[Accessed 15 August 2024].

Donaldson, G., 2023. *Hall of Fame: Lewis Hamilton.* [Online]   
Available at: https://www.formula1.com/en/information/drivers-hall-of-fame-lewis-hamilton.6TbSwzJA2miTUwtnXUHkjk  
[Accessed 15 August 2024].

Donaldson, G., n.d. *Hall of Fame: Alberto Ascari.* [Online]   
Available at: https://www.formula1.com/en/information/drivers-hall-of-fame-alberto-ascari.6zgWYwDIjoQmYSiz48FAZD  
[Accessed 17 August 2024].

Ozanian, M. & Knight, B., 2023. *Formula 1's Most Valuable Teams.* [Online]   
Available at: https://www.forbes.com/sites/mikeozanian/2023/07/19/formula-1s-most-valuable-teams-2023/  
[Accessed 15 August 2024].

# The Impact of Data Science Within Stantec

## Executive Summary and Project Background

Working in the clean water industry for a consultancy, a common type of project we are asked to do by our various clients (the clean water suppliers) is to model a water distribution network so as to optimise it for their customers. To do this, flow and pressure data is taken from large areas of the network at key points every 15 minutes to give an idea of the general change in clean water demand over time, and how this changes over a single day, over an entire year, or even over many years. These demand patterns are then used to create the model and are compared against what the model is showing to ensure that the digitised water network is accurate to the real one. In particular, a “peak day” and an “average day” are used to show how the network behaves with the maximum demand in a year and the normal demand in a year respectively. Given that a modelled area may have hundreds or even thousands of data loggers, each recording pressure and flow every 15 minutes for several years, finding these days through the massive quantities of data is not a simple task, especially not using traditional methods. In the past, this work would be done with numerous Excel spreadsheets that are years old themselves, running many slow and tedious calculations with a lot of requirements for user input. Through the creation of a Power BI dashboard and with the help of tools such as DAX Query, I was able to speed up this process in a way that will be repeatable for future projects, saving both time and money.

## Analysis

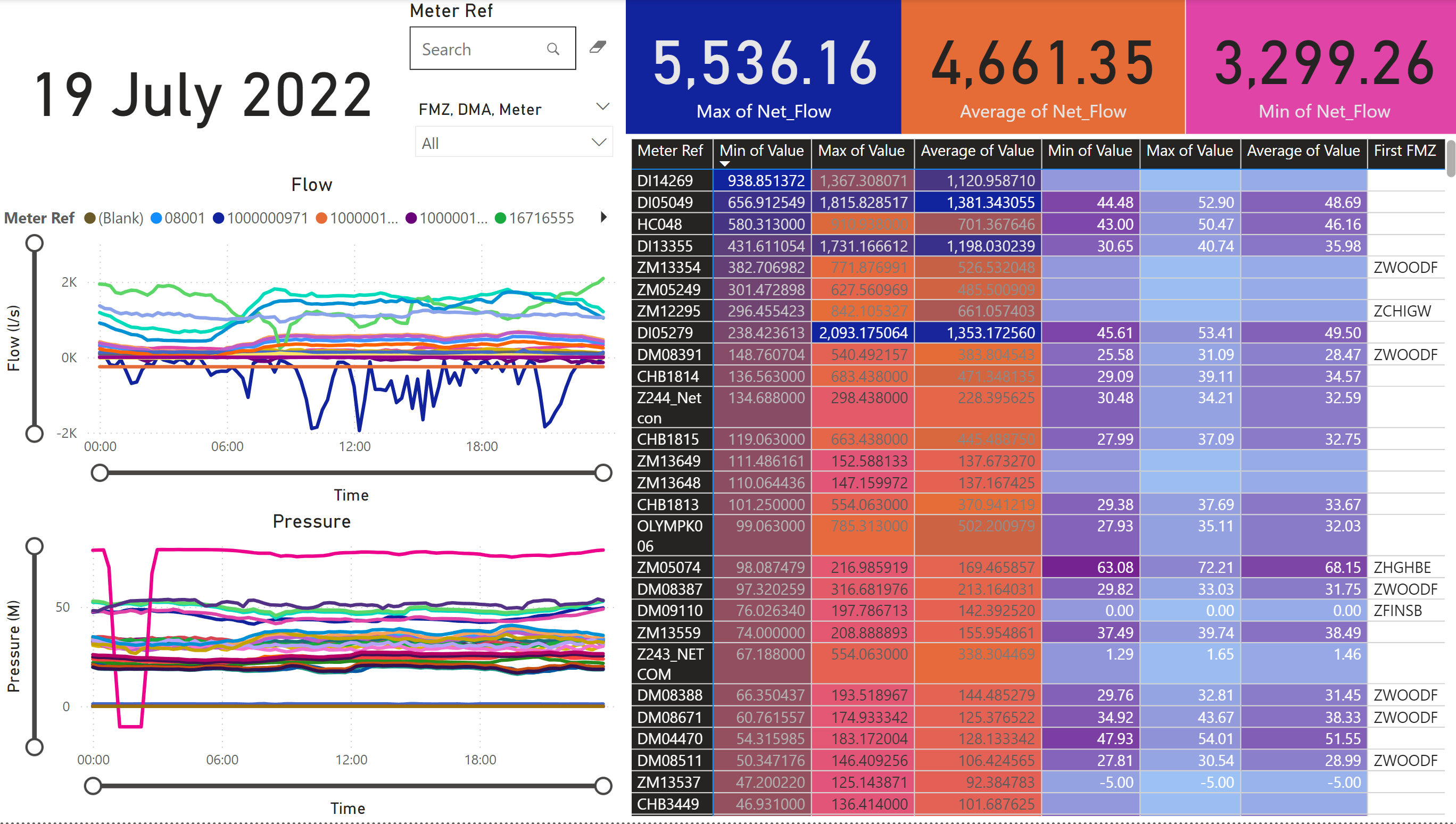
The bulk of the project was time-series analysis, as each water network meter logs data at consistent intervals to show a change over time in the telemetry data. The scope of the data is from the start of 2018 through to the end of 2023, such that the change in demand over lockdown can be measured and accounted for, covering the entire area supplied by the water treatment works that is being modelled. To find the demand for an area as large as this one when there are thousands of data points to account for every 15 minutes, the net flow is calculated by taking the balance of a few key meters at pipes serving the borders of the area in order to measure how much water is entering/leaving the area at any given time interval. Then, by aggregating this data to give daily averages, the peak day can easily be found: the day with the greatest average flow is the peak. The Summer of 2022 was particularly hot, and therefore had a higher water demand on average than other times, making this Summer the ideal candidate for a peak day demand profile – the exact day chosen was July 19th. Then, the data for each meter only for that day is extracted from the larger dataset, giving a detailed demand profile of the peak day.

The average day is less simple. While it would make sense to assume that the methodology is largely the same – taking the day with the median net flow – this approach does not work perfectly for a handful of reasons. One such reason is that the real-world water network is constantly changing, with new pipes being installed and old ones becoming redundant, and valves being opened or closed largely at the client’s discretion. To create a consistent model, the average day has to occur closely enough to the peak day such that the network at large is more or less the same; this means that only a day in the same fiscal year (at the client’s request) can be chosen as the average day. On top of this, different areas within the larger area will have different average days, and so an approach as broadly-reaching as taking the net flow for the entire area is not suitable. Instead, new net flows are calculated for smaller areas known as Flow Monitoring Zones (FMZs), which are then aggregated by day. Then, a delta is taken from the mean average day for each FMZ to show how each day in the 2022/23 fiscal year compares to the rest of the year. From here, these deltas are summed to give an insight into what the average day must be for the entire zone, allowing a data scientist to inspect these days with a greater level of fidelity. The day ultimately chosen using this method was October 10th.

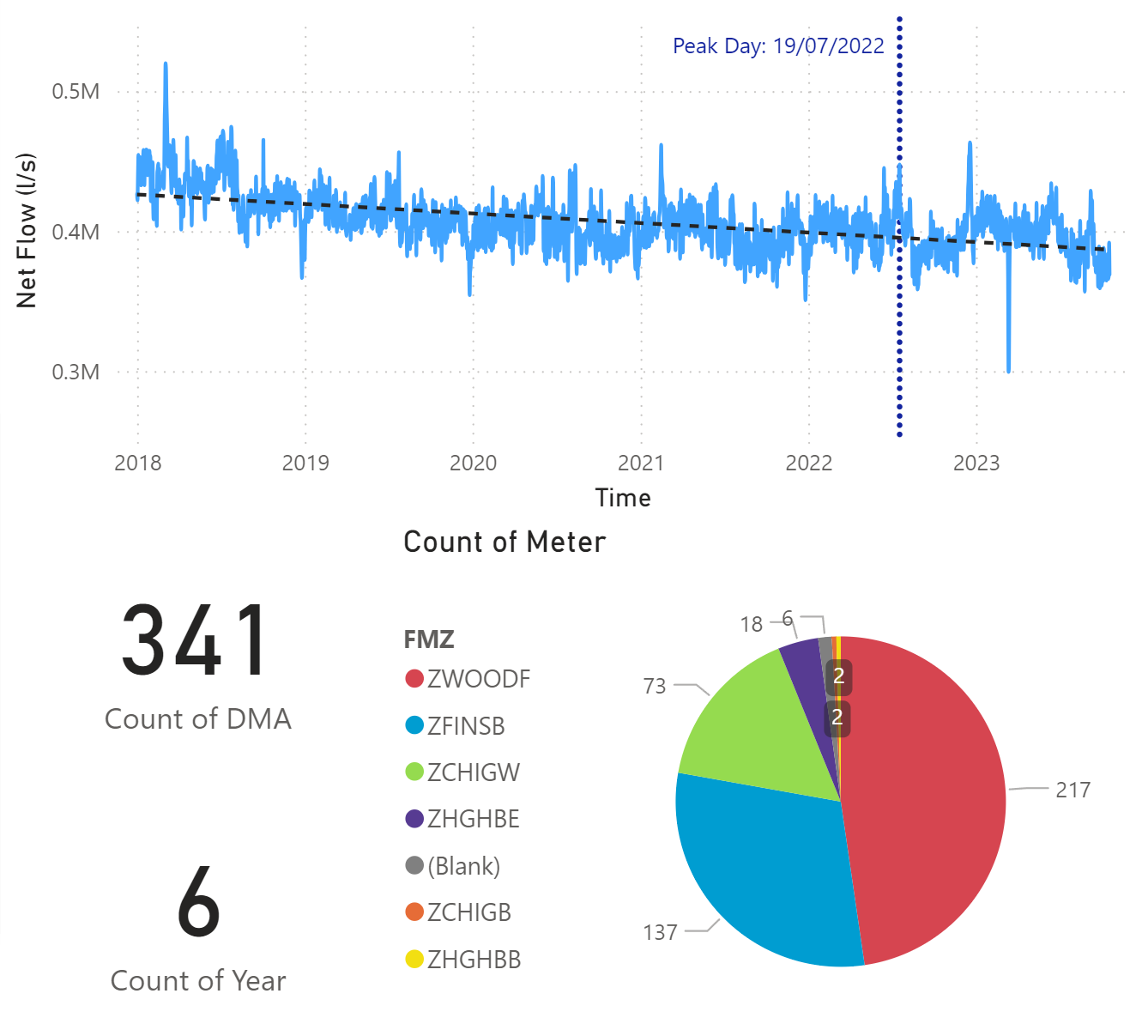
## Evaluation

I have been told that these tasks at times take multiple weeks of deliberation between numerous stakeholders just to agree on the selection of two different days. Using the methods stated above with more advanced tools such as Power BI and DAX Query, this project took myself alone only a single week. The time saved allowed my teammates to begin the modelling work far sooner, leading to an overall better project deliverable for the client as they had more time to focus on the actual modelling process. This task is now also able to completed more quickly for future modelling projects, and therefore clients are more likely to approach our consultancy for such work, boosting Stantec’s reputation and market share within this sector. Were I to undertake this project again, I would likely aim to improve the communication process between myself and the relevant stakeholders so as to provide a peak and average day more suited to their needs.

## Documentation



A page of the Power BI dashboard dedicated to analysis of each meter in the area for the peak day. This page can be modified for similar analysis for the average day.



A page of the dashboard dedicated to providing a holistic view of the entire dataset. To the right is a map showing the location of each meter colour-coded by FMZ but this has been cropped for security reasons.