Technical Report

Team Kappa

Data pre-processing

Initial exploration of the data sets was carried out in Snowflake where the data is hosted. We started by querying the first few rows of the tables to get an initial impression of the features of the data, such as data types and quality, as well as understanding the structure of the dataset. It was useful to note which columns of data are present in more than one table as these can be used as keys to join tables together. Data visualisations were created using native functionality in Snowflake.

Cleaning Steps

Step	Action Taken	Justification
1	Validate data quality - data types and unexpected null values	Improves efficiency and reliability during analysis
2	Added a column [Fault_Group] in the Status table	Easily group and identify technical failures and retirements.
3	Added a column [Position_Delta] in the Results table, calculated from [GridPosition - RacePosition]	To use as a measure of driver performance during a race
4	Joined tables together and saved it as a view	Improves readability and simplifies querying from the view instead of the raw data

5	Filtered datasets to 2014 onwards	To focus on the hybrid era of F1, for reliable comparison

A note on data types

We found that in general the data types were already in a good format so we did not need to convert data types. While data showing elapsed times, such as race finishing time, were stored as varchar rather than time intervals, it was noted that the data already contained a field which records the duration in milliseconds, stored as an int. Where a time value stored as varchar has the potential to be problematic for time calculations, it is possible to do time calculations using the duration and convert the int value to a time value upon output. Since we could do this, we left the elapsed time data as a varchar because it was easily readable when working with the data.

Exploratory Data Analysis

Total number of unique drivers, races, constructors and circuits which have been analysed in this report:

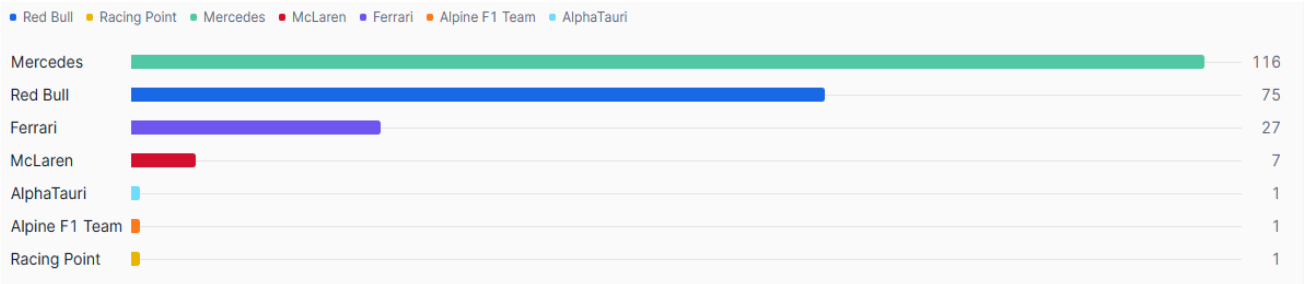
DRIVERS	RACES	CONSTRUCTORS	CIRCUITS
59	228	20	32

Features Identified for Analysis

- Number of race wins
- Number and cause of DNFs
- Average pit stop times
- Correlation between top fastest lap and finish position
- Relationship between pit stop duration and finishing position
- Relationship between qualifying position (grid position) and finishing position

Which team has won the most races?

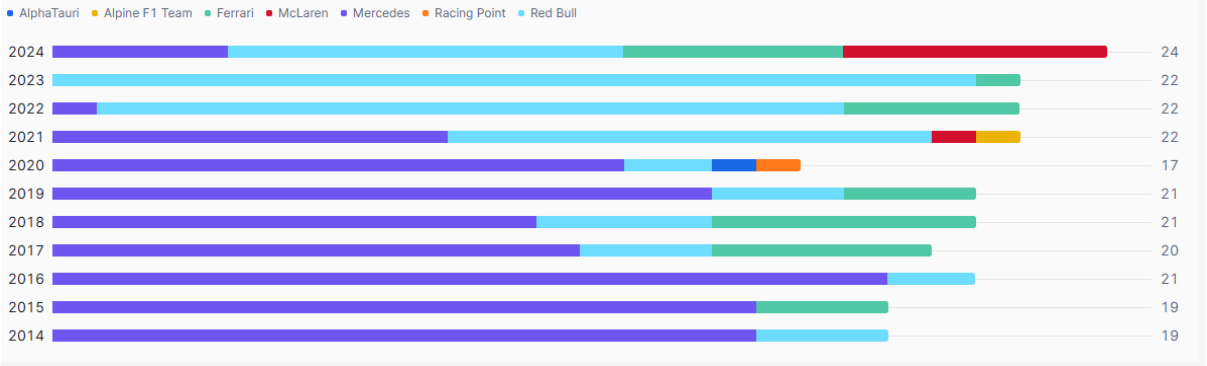
Count of race wins by Constructor



As expected from the project brief, we can see that Mercedes has dominated the competition since the hybrid era began in 2014.

However, if we drill into this further to analyse the number of wins per year since 2014, we find a slightly different story.

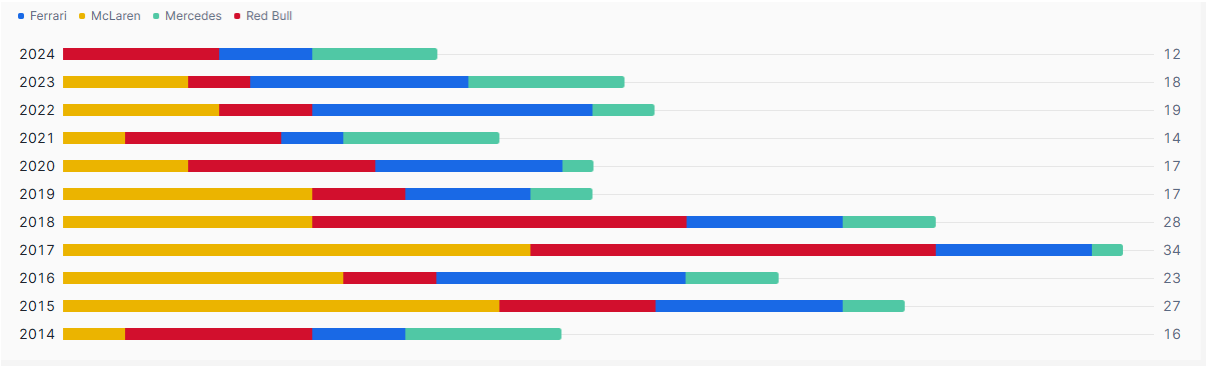
Count of race wins by constructor by year



While it's evident that Mercedes outperformed all other teams from 2014 until 2020, the tipping point came in 2021 when Red Bull secured two more wins than Mercedes, putting them top of the Constructor Championship for the first time. Red Bull have continued to dominate the sport since, winning 77% and 95% of all races in 2022 and 2023 respectively. In 2024, the playing field was more evenly split between Red Bull, Mercedes, Ferrari, and McLaren.

How many times did teams not finish a race?

Number of DNFs by constructor by year (top 4 teams)



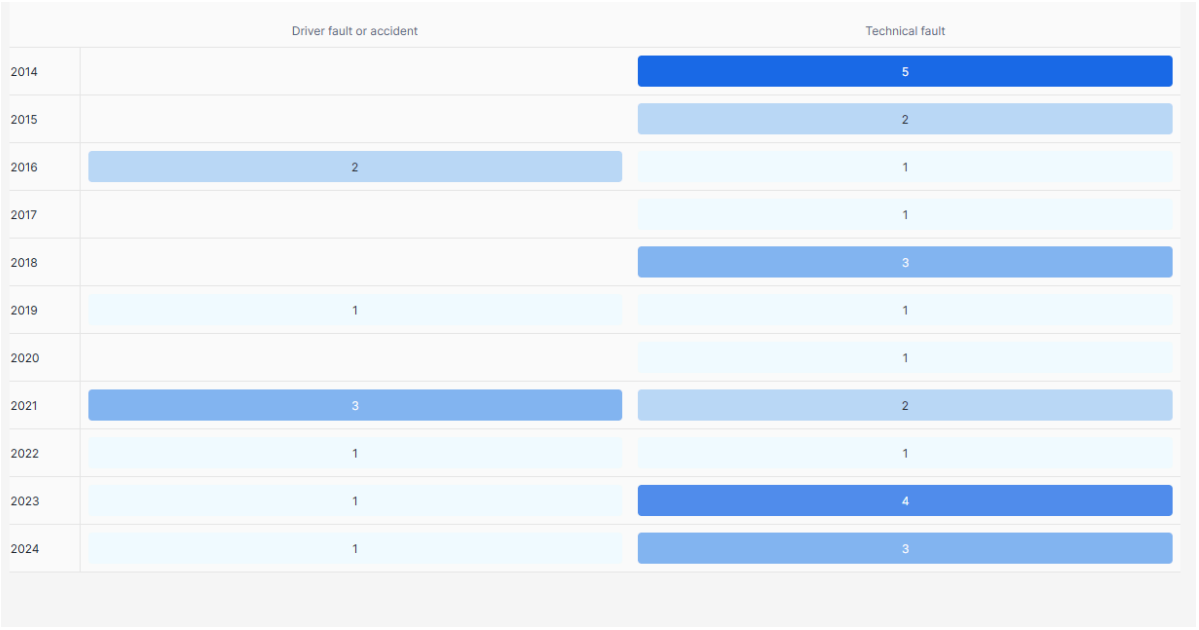
What are the primary causes of not finishing a race?

The primary causes of not finishing a race are technical faults with the car. Driver faults (collisions, accidents) are also a factor but less common.

Cause of DNFs - McLaren

McLaren have consistently improved the reliability of the car by reducing the number of technical faults.

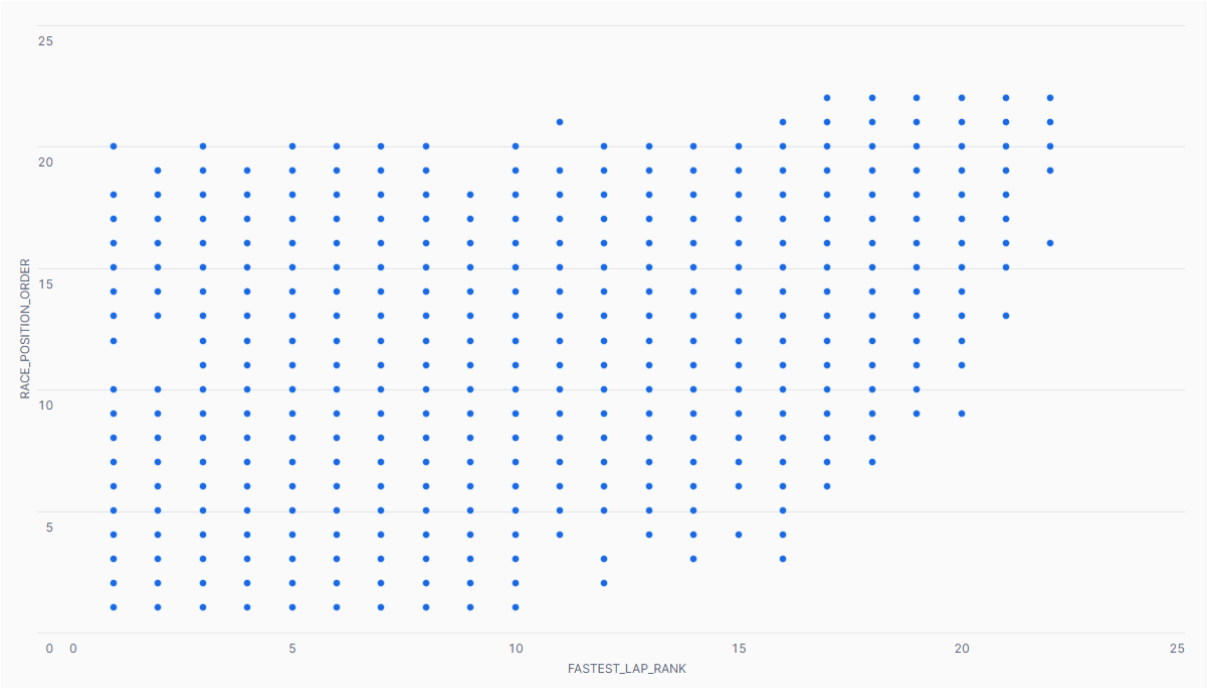
Cause of DNFs - Mercedes



Mercedes had very few technical or driver faults during their dominant years from 2014 - 2020. From 2021 onwards, the higher number of faults correlates with fewer wins and drop in performance in the Constructors Championships.

Do drivers with the fastest lap time finish higher in the race?

Correlation between fastest lap and race position using rank



The scatter plot shows the fastest lap rank against race position order. The bottom-left quadrant shows drivers with fast laps and top finishes. The top-right quadrant shows drivers with slow laps and slow finishes. However there is no correlation in the data, indicating that many drivers with faster laps don't necessarily win (and vice versa). This suggests other factors, such as strategy, pit stops, faults or collisions have a strong impact on race outcomes.

Modelling

For the prediction model, the objective was to determine what position within the Top 3 would McLaren finish in during the upcoming Formula 1 races. The machine learning model used to make the prediction was Decision Tree Classifier, combined with Random Forest Classifier to identify the most influential factors behind the top 3 podium finishes.

From the combined dataset "VW_ALLRACESTATS"., the following key variables were chosen for the prediction model based on analysis done using SQL:

Feature Type	**Columns**
Driver Details	DRIVER_NAME, DRIVERID, DRIVER_POSITION_STANDINGS, DRIVER_POINTS_STANDINGS, WINS
Constructor Information	CONSTRUCTOR_NAME, CONSTRUCTOR_POSITION_STANDINGS
Race Statistics	GRID_POSITION, RACE_POSITION_ORDER,

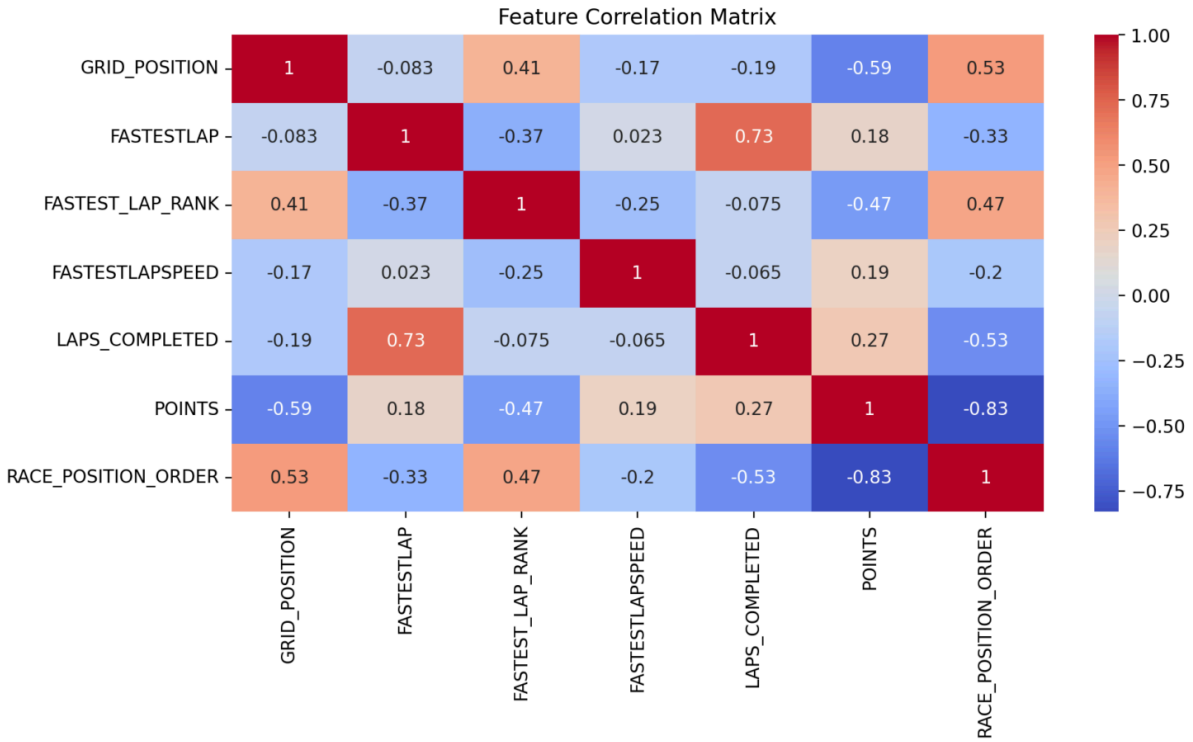
	FASTESTLAPSPEED, TIME_BEHIND_LEADER
Qualifying Results	QUALI_POSITION, Q1, Q2, Q3

A binary target was created indicating whether McLaren finished in the Top 3 in the races recorded in the data. To proceed with creating the dataset, rows which were not relevant to McLaren, i.e., rows containing other Constructor performances, were dropped from the table and the new table was stored as a new dataframe variable df_mclaren.

Next step was feature engineering which was done using correlation matrix to find the appropriate features with high correlation to McLaren winning. The correlation matrix showed high correlation between:

- Grid position and Fastest Lap Rank
- Grid position and Race Position Order
- Fastest Laps and Laps completed
- Fastest Lap Rank and Race Position Order

Points feature had less correlation with the other features, therefore it was removed from the training data.



After feature cleaning and engineering, the following features were chosen for the training data:

```
features = [
   "GRID_POSITION", "FASTEST_LAP_RANK", "FASTESTLAPSPEED",
```

```
    "DRIVER_POINTS_STANDINGS", "DRIVER_POSITION_STANDINGS",
    "CONSTRUCTOR_POINTS_STANDINGS",
"CONSTRUCTOR_POSITION_STANDINGS",
    "QUALI_POSITION"
]
```

These features capture the starting performance, position standings and lap-based metrics, all which have shown to have higher correlation with successful race performance.

For the train/test split, the features and target variable were identified as follows with test size as 0.2:

```
X = df_mclaren[["GRID_POSITION", "QUALI_POSITION", "POINTS",
"FASTEST_LAP_RANK",
    "FASTESTLAPSPEED", "DRIVER_POINTS_STANDINGS",
"DRIVER_POSITION_STANDINGS",
    "CONSTRUCTOR_POINTS_STANDINGS",
"CONSTRUCTOR_POSITION_STANDINGS", "WINS"
]]

y = df_mclaren["MCLAREN_TOP3"]
```

Standard Scaler was used to ensure input data points have a balanced scale for the machine learning model.

Grid Search was used to conduct Hyperparameter tuning to find the optimal parameters to train the Random Forest Classifier model. The parameters chosen for the grid search are as below:

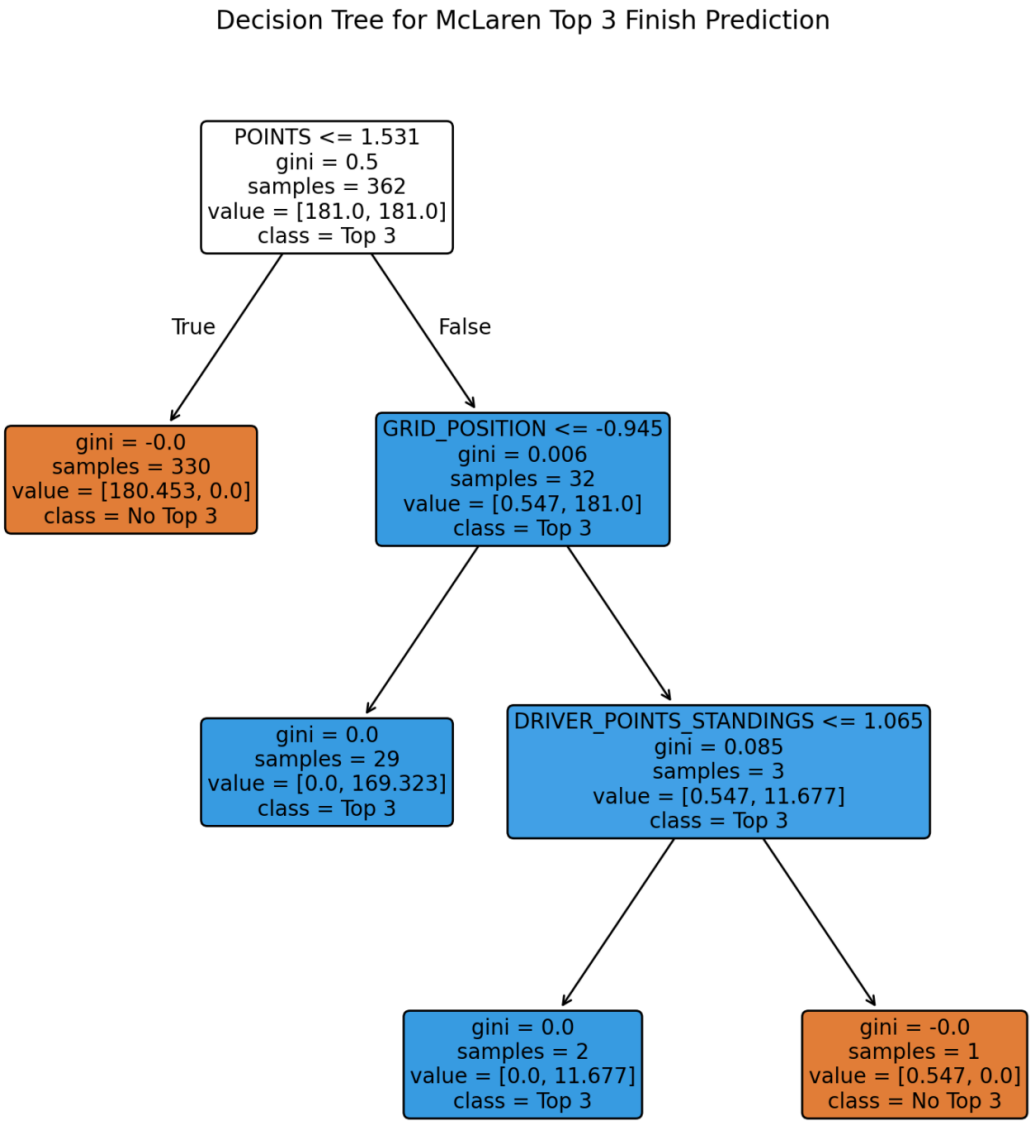
```
param_grid = {
    "n_estimators": [100, 200],
    "max_depth": [5, 10, None],
    "min_samples_split": [2, 5],
}
```

Model Evaluation after fitting the data in Random Forest Classifier showed 97% accuracy:

```
Model Performance:
              precision    recall  f1-score   support
           0       0.98      1.00      0.99        81
           1       1.00      0.80      0.89        10
```

```
    accuracy                           0.98        91
   macro avg       0.99      0.90      0.94        91
weighted avg       0.98      0.98      0.98        91
Accuracy: 0.978021978021978
```

Decision Tree Classifier Output:



From the decision tree output, it was identified that McLaren has a best chance of occupying Top 3 positions if the Grid Position <= -0.945 and DRIVER_POINTS_STANDINGS <= 1.065.

Accuracy of Decision Tree Classifier = 97%

Prediction:

Based on the output from the decision tree, the model predicted True for McLaren winning a future race within the Top 3 podium standings given the following scores:

"GRID_POSITION": 0.945,
"QUALI_POSITION": 8,
"POINTS": 80,
"FASTEST_LAP_RANK": 50,
"FASTESTLAPSPEED": 130,
"DRIVER_POINTS_STANDINGS" : 1.065,
"DRIVER_POSITION_STANDINGS": 120,
"CONSTRUCTOR_POINTS_STANDINGS": 120,
"CONSTRUCTOR_POSITION_STANDINGS":190,
"WINS": 40

Decision:

- Grid position is critical predictor of McLaren finishing at Top 3
- Grid position and Driver Points Standing are indicators of actual race potential.

Conclusion

Overall, the findings demonstrate the complex and multifaceted nature of formula 1 racing. Each race outcome is influenced by a wide combination of performance metrics. The report highlights that while raw speed alone is not enough to be the race winner, that grid position, driver points standings (race history) and technical reliability are the most important factors in securing podium finishes.