F1 Constructor Championship Ranking Prediction

Sheng Ang Joy Gao

December 2025

1. Introduction

Formula 1 (F1) has experienced a surge in global popularity in recent years, driven by increased social media engagement, storytelling through documentaries and films, and huge sponsorships behind drivers and teams. As with many sports, there is widespread interest among fans, analysts, and bettors in predicting championship winners. However, predicting winners remains challenging, as race outcomes depend on many complex factors of qualifying performance, race-day execution, strategic decision-making, and historical team performance. This combination of uncertainty and complex performance data makes Formula 1 an ideal setting for applying statistical methods and predictive modeling. The goal of this project is to evaluate historical Formula 1 team-level performance data from the past five seasons (2020-2024) in order to predict which team is most likely to win a race during the 2025 season. Rather than focusing on individual drivers, this analysis is conducted at the constructor (team) level. For example, Ferrari, McLaren, or Red Bull. This modeling choice simplifies the data structure and reduces driver-level variability, while still capturing the strategic and technical differences that distinguish team performance. This project utilizes the public dataset: Formula 1 World Championship data, which contains comprehensive information on races, constructors, drivers, qualifying results, lap times, pit stops, and championship standings. From these data sources, team-level predictors are constructed by aggregating driver-level information within each race. Key predictors include historical constructor standings, average qualifying position, average lap time, pit stop performance, and race grid positions. These variables collectively reflect both long-term team strength and race-specific conditions. An ordinal outcome based on constructor championship position was examined to capture relative team performance across races. Visualizations using ggplot2 and machine learning methods including simple linear regression, k-nearest neighbors (KNN), and random forests were applied and compared using cross-validation to assess predictive accuracy and model interpretability. By applying the modeling techniques learned in class to a real-world and engaging dataset, this project demonstrates how statistical learning methods can be used to uncover performance patterns in competitive sports and to generate data-driven predictions in a complex, high-variability setting.

2. Data Preprocessing

Data preprocessing presented several challenges for this dataset. As described in the Introduction, the data were obtained from Kaggle's Formula 1 World Championship public data. The data

are provided in a relational database format, with each table representing a separate entity. To prepare the dataset for analysis, we merged the tables using primary and foreign keys, introducing additional complexity in maintaining referential integrity.

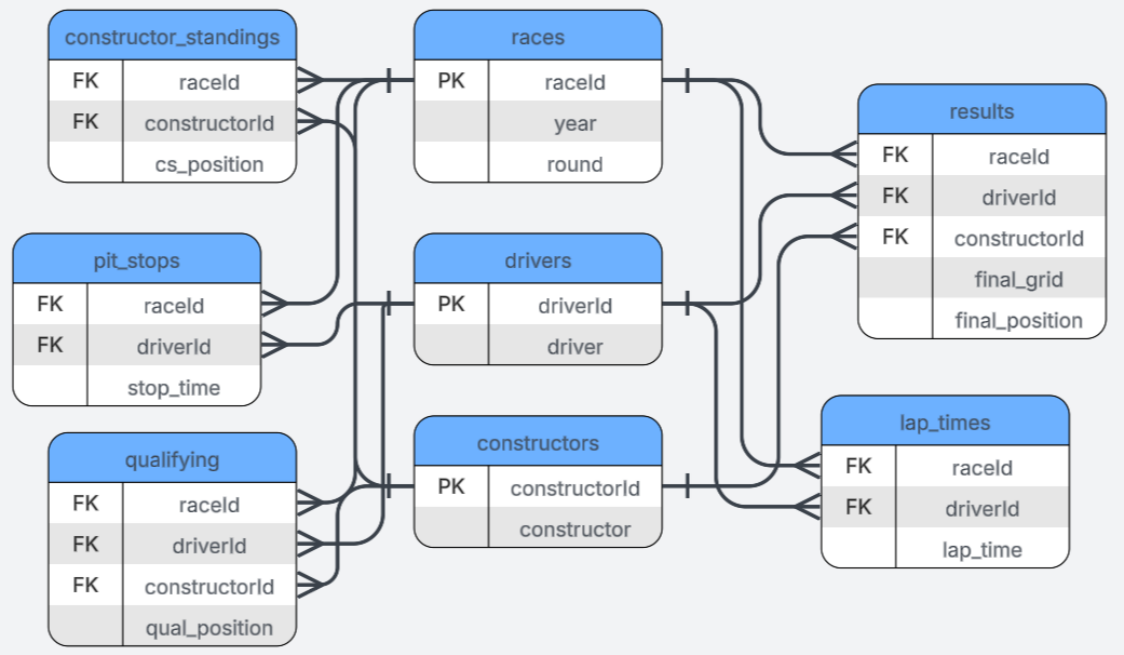

Figure 1: Entity Relationship Diagram of the F1 dataset.

Figure 1 illustrates the relationships between the tables. While this study focuses primarily on constructor data and race results, driver-level data were also included because some tables reference `driverId` without an associated `constructorId`. The target variable (`cs_position`) represents the constructor's season standing, and the predictors include averaged constructors' lap time, stop time, qualifying positions, grid positions and final standings for each race.

A critical preprocessing step involved accounting for drivers who did not finish (DNF) races. If a driver retired due to crashes, mechanical failures, or critical weather conditions, their recorded `lap_time` might be shorter than drivers who completed the race. To mitigate this bias, we computed the mean `lap_time` and `stop_time` per lap, rather than using the total times for the entire race.

Since each constructor has two drivers, we averaged driver-level metrics to obtain team-level values. For instance, if driver A at McLaren finished first and driver B finished fourth, the constructor's `avg_final_position` would be 2.5. Remaining missing values were then removed. At this stage, the only remaining `NA`s corresponded to two exceptional cases: a double DNF, where both drivers from a constructor failed to complete any laps (one instance), and the 2021 Belgian Grand Prix. The Belgian GP was highly atypical: 44 laps were planned, but the race was stopped on lap three due to extreme wet conditions. According to sporting regulations, results were taken from the end of lap one, with half points awarded to the top ten finishers. Therefore, these instances were removed as anomalies.

Finally, all numeric predictors were normalized, and multicollinearity among predictors was

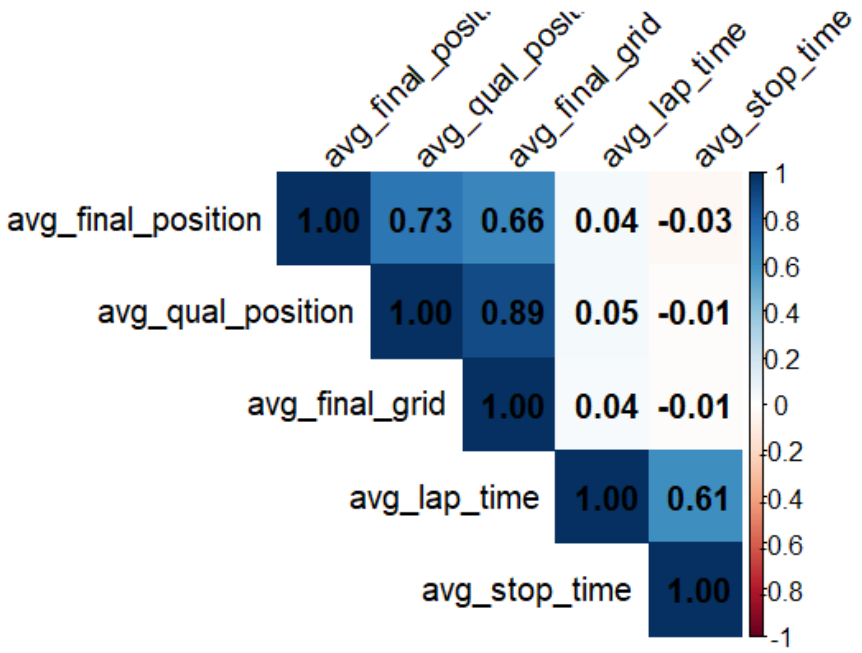

Figure 2: Correlation matrix of predictors.

assessed. Figure 2 shows that `avg_final_position`, `avg_qual_position`, and `avg_final_grid` are highly correlated. Given the limited number of predictors, we opted to retain all columns rather than remove any.

3. Exploratory Data Analysis

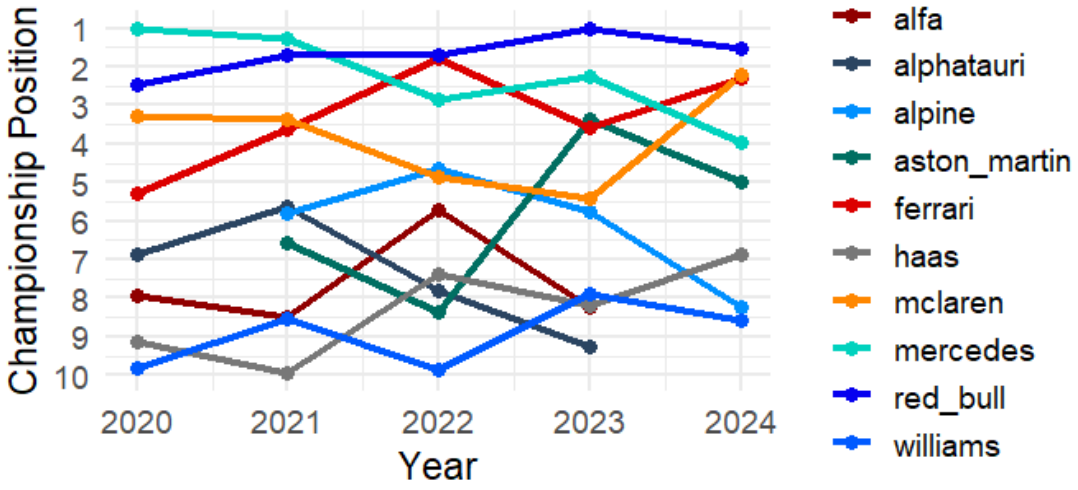

Figure 3: Constructor's championship position over time.

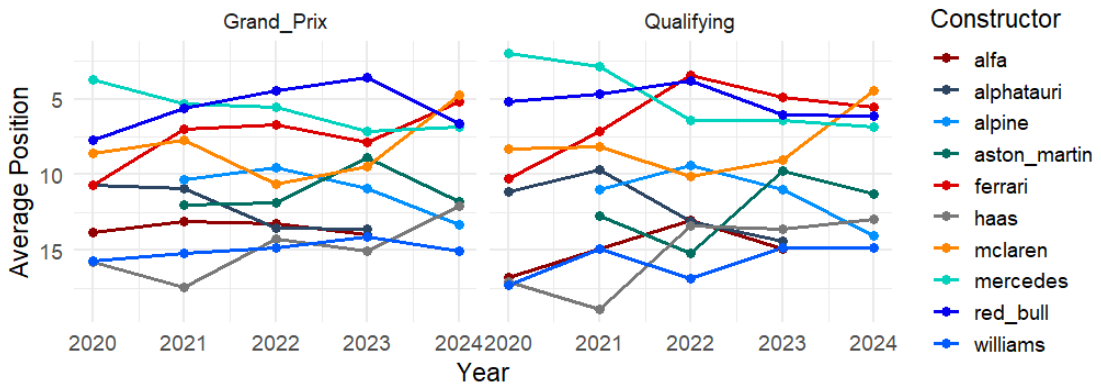

Figure 4: Constructor performance over time: (Left) final race standings and (Right) qualifying positions.

Figure 3 illustrates the constructors' championship positions over the period 2020-2024. Constructors with fewer than two years of data were excluded from the visualization. Red Bull, Mercedes, Ferrari, and McLaren consistently occupy the top positions over the past five seasons. Similarly, Figure 4 shows that Red Bull and Mercedes consistently dominate both qualifying and race performance metrics. Ferrari and McLaren experienced slight setbacks in 2020 and 2021 but recovered in subsequent seasons. Both figures use colors corresponding to each constructor's official team colors.

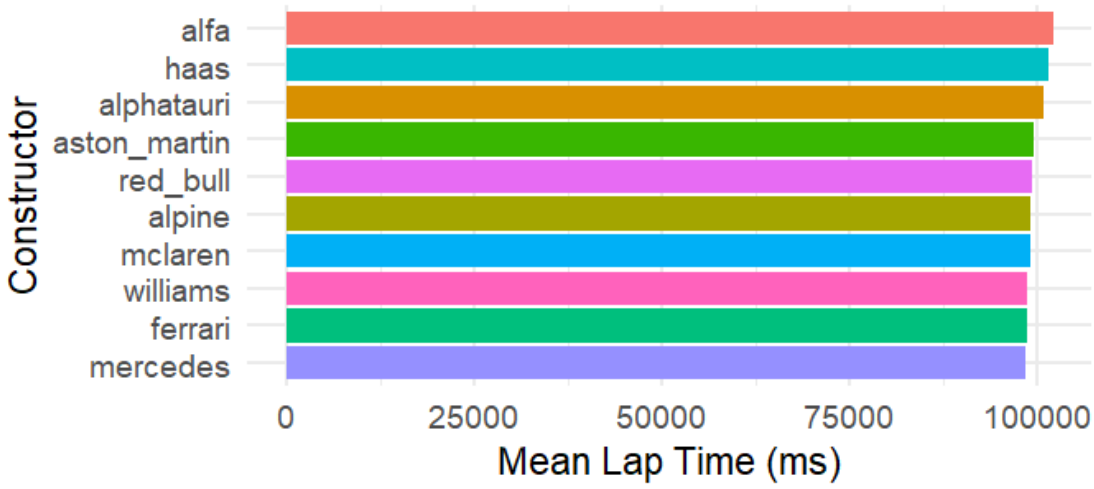

Figure 5: Mean lap time (ms) of constructor across the 2020-2024 period.

Figure 5 presents the mean lap times for each constructor over the five-year period. Differences among constructors remain below 4000 milliseconds. Mercedes exhibits the lowest average lap time overall, while Alfa Romeo records the highest. Interestingly, Red Bull shows a moderate mean lap time despite consistently securing strong positions. This discrepancy arises because Max Verstappen, Red Bull's lead driver, consistently delivers top-tier performances, whereas his

teammates, rotating across various drivers, tend to record higher lap times.

4. Results

F1 2025 winner predictions were generated using 3 approaches: simple linear regression (SLR), KNN, and Random Forest model, using data from the past 5 years. Linear regression was first used because of its simplicity and interepretability. To complement this analysis and capture nonlinear relationships between race-level performance metrics and championship outcomes, we also implemented two machine learning models: K-nearest Neighbors (KNN) and Random Forest. In all three cases, the constructor championship position (cs_position) was treated as a numeric outcome since it has an ordered ranking structure. Additionally, when generating predictions, we applied a weighting scheme that assigned greater importance to more recent seasons, under the assumption that recent performance is more predictive of future outcomes. Table 1 displays the predicted constructor ranks for all three approaches compared to the actual results.

Table 1: Predicted vs. Actual Constructor Rankings for the 2025 F1 Season

Constructor	SLR Predicted	KNN Predicted	RF Predicted	Actual Rank
McLaren	4	4	4	1
Mercedes	3	1	2	2
Red Bull	1	2	1	3
Ferrari	2	3	3	4
Williams	10	10	10	5
RB	1	-	-	6
Aston Martin	5	5	5	7
Haas	9	9	9	8
Sauber	1	-	-	9
Alpine	6	6	6	10
Alphatauri	7	7	7	–
Alfa Romeo	8	8	8	–

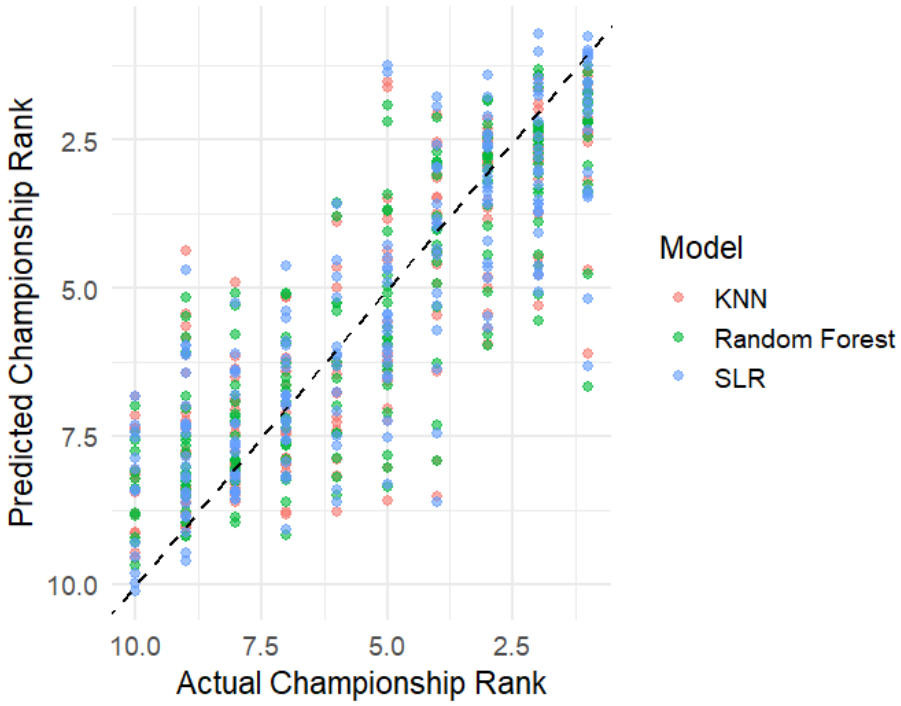

Figure 6: Predicted vs. actual constructor championship rank. The dashed line represents perfect prediction ($y = x$). Points closer to the diagonal indicate more accurate predictions.

Overall, the three methods yielded similar results. Figure 6 demonstrates high performance with predictions close to the diagonal line (meaning high accuracy). Linear regression was used as a baseline for modeling the outcome of championship rank, however, it assumes a linear relationship between predictors and continuous variables which may not capture complex, nonlinear interactions between race-level metrics and overall championship performance. On the other hand, KNN offered a flexible, non-parametric approach that performed relatively well (RMSE = 1.61). Random Forest also allowed complex interactions and nonlinearities while being robust to overfitting. This model performed the best (RMSE = 1.59, $R^2 = 0.705$) with the best overall predictive accuracy with KNN providing similar performance. However, all three models had similar variability in the predictive accuracy meaning that using more complex models didn't improve the predictive power that much.

4.1 Simple Linear Regression (SLR)

Simple Linear Regression predicts constructor championship rank by estimating a linear relationship between the predictors and the target variable, `cs_position`. Each predictor contributes proportionally to the predicted rank, with coefficients learned from historical race data. All available race-level predictors were included in the model, excluding identifier variables (race ID, constructor ID, and constructor name). To ensure valid distance calculations, all numeric predictors were median-imputed and standardized using a pre-processing recipe. Unlike KNN, which relies on local similarity, SLR produces a single linear function across all observations. The model was trained on

five-fold cross-validation to assess generalization, achieving a cross-validated RMSE of 1.51 and a MAE of 1.19. On the test set, the model maintained comparable accuracy (RMSE = 1.57, MAE = 1.19) and explained approximately 69.8% of the variance in constructor rankings ($R^2 = 0.698$). Applying the model to the 2025 season predictions, Red Bull was projected to lead the championship, closely followed by Ferrari and Mercedes, while Williams and Haas were predicted to finish near the bottom of the standings.

4.2 K-Nearest Neighbor (KNN)

KNN regression predicts championship rank by identifying race-level observations with similar performance characteristics, including qualifying position, lap time, grid position, and pit stop duration. Each race (ex. Red Bull race) is a point in the predictor space and measured for similarity using distance. KNN computes the Euclidean distance between that race and every race in the training set so those races that have similar lap pace, qualifying strength, etc. would have a smaller distance. The number of neighbors (k) was selected via 5-fold cross-validation, minimizing the root mean squared error (RMSE). Our cross-validation selected k=25 so the predicted championship rank was an average of 25 comparable races. When evaluated on the test data, the KNN model achieved similar results as SLR, with an RMSE of 1.58 positions and a mean absolute error (MAE) of 1.22 positions, meaning that predicted championship rankings were typically within one to two positions of the true outcome. The model explained approximately 69.6 percent of the variation in championship rankings ($R^2 = 0.696$), suggesting that race-level performance metrics provide reasonable predictive power.

4.3 Random Forest

Random Forest regression is a more flexible machine learning technique that predicts championship rank by averaging results from many decision trees. It is capable of modeling nonlinear effects and interactions among predictors and unlike KNN, Random Forest does not require predictor normalization and is more robust to correlated variables and noise. Key hyperparameters, including the number of variables sampled at each split and the minimum node size, were once again tuned using 5-fold cross-validation with RMSE as the optimization criterion. The cross-validation selected `mtry` = 3 and `min_n` = 10. This means the tree consider three features at every decision node while preventing overfitting by requiring at least 10 data points to justify creating a new split. When evaluated on the test data, the Random Forest model achieved an RMSE of 1.54 positions and a mean absolute error (MAE) of 1.18 positions. The model explained approximately 70.9 percent of the variation in championship rankings, which outperformed KNN and SLR slightly in terms of explainability.

5. Discussion

Using data from the past five seasons with a weighting scheme that emphasizes more recent performance, all models consistently ranked McLaren, Mercedes, Ferrari, and Red Bull among the top four teams (see Table 1). This aligns with their historically strong performance in qualifying, race execution, and overall consistency. Across modeling approaches, Red Bull, Mercedes, and Ferrari consistently occupied the top positions, with no differences in the predicted order of the remaining teams.

However, the model predictions deviated from reality due to substantial changes in drivers and car performance. McLaren emerged as the 2025 season leader due to major car upgrades that were not captured in the historical data. Similarly, Williams delivered an unexpectedly strong performance, largely attributable to the addition of Carlos Sainz Jr., a consistently high-performing driver, to their lineup.

Improved feature engineering, incorporating variables such as car specifications (e.g., tires, engines, aerodynamics) or contextual factors like track layout and weather conditions, could potentially enhance predictive accuracy. However, such data are not publicly available and would require access to proprietary sources, such as the official F1 API.

This experiment highlights the inherent challenges of predicting dynamic, real-world outcomes. Even robust machine learning models can be limited when unexpected changes occur in the system being modeled.

6. Appendix

A. Contribution

- Joy Gao: EDA Visualizations, KNN, Random Forest, Results
- Sheng Ang: Data preprocessing, SLR, Discussion and Conclusion

B. Additional Results

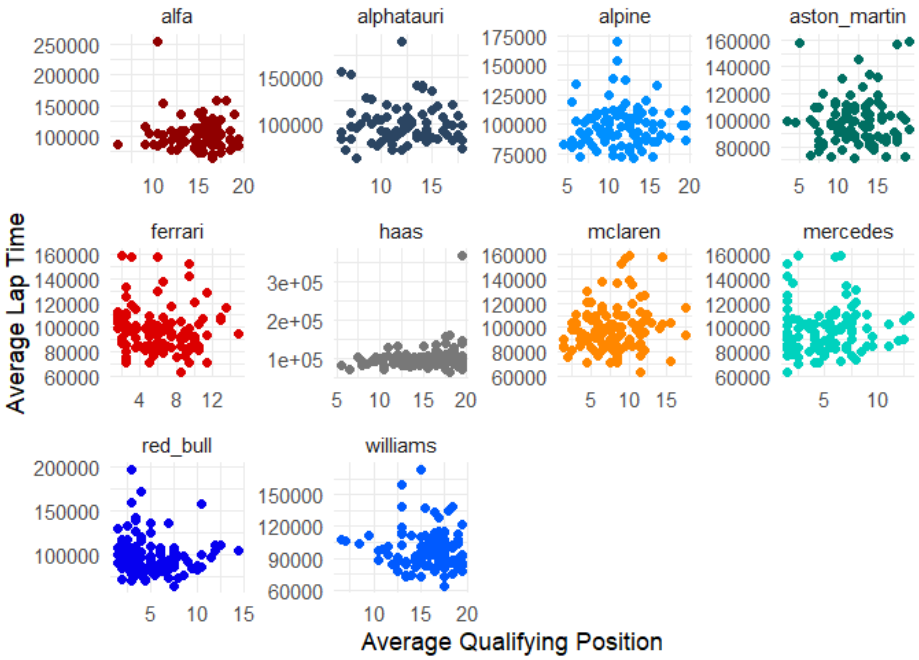

Figure 7: Scatterplot of average lap times vs. qualifying position across teams.

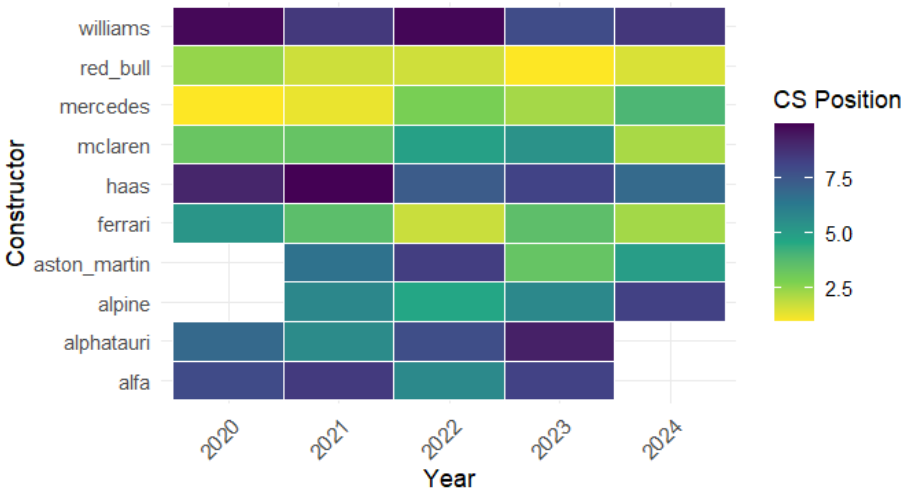

Figure 8: Heatmap of `cs_position`