Assessment of Formula 1 Drivers and Vehicles to Choose Sponsorship Candidates

Team Melbourne

SMITH SCHOOL OF BUSINESS

Smith SCHOOL OF BUSINESS Queen's University

Global Master of Management Analytics

GMMA 860
Acquisition and Management of Data

Alex Scott

Team Project
May 7, 2019

Team Melbourne:
Anthony Azar, Lorraine Feng, Nicole Hong, Di Wu, Joel McInnis, David Trinh, Busola Daodu

Order of files:

Filename	Pages	Comments and/or Instructions

Additional Comments:

Executive Summary

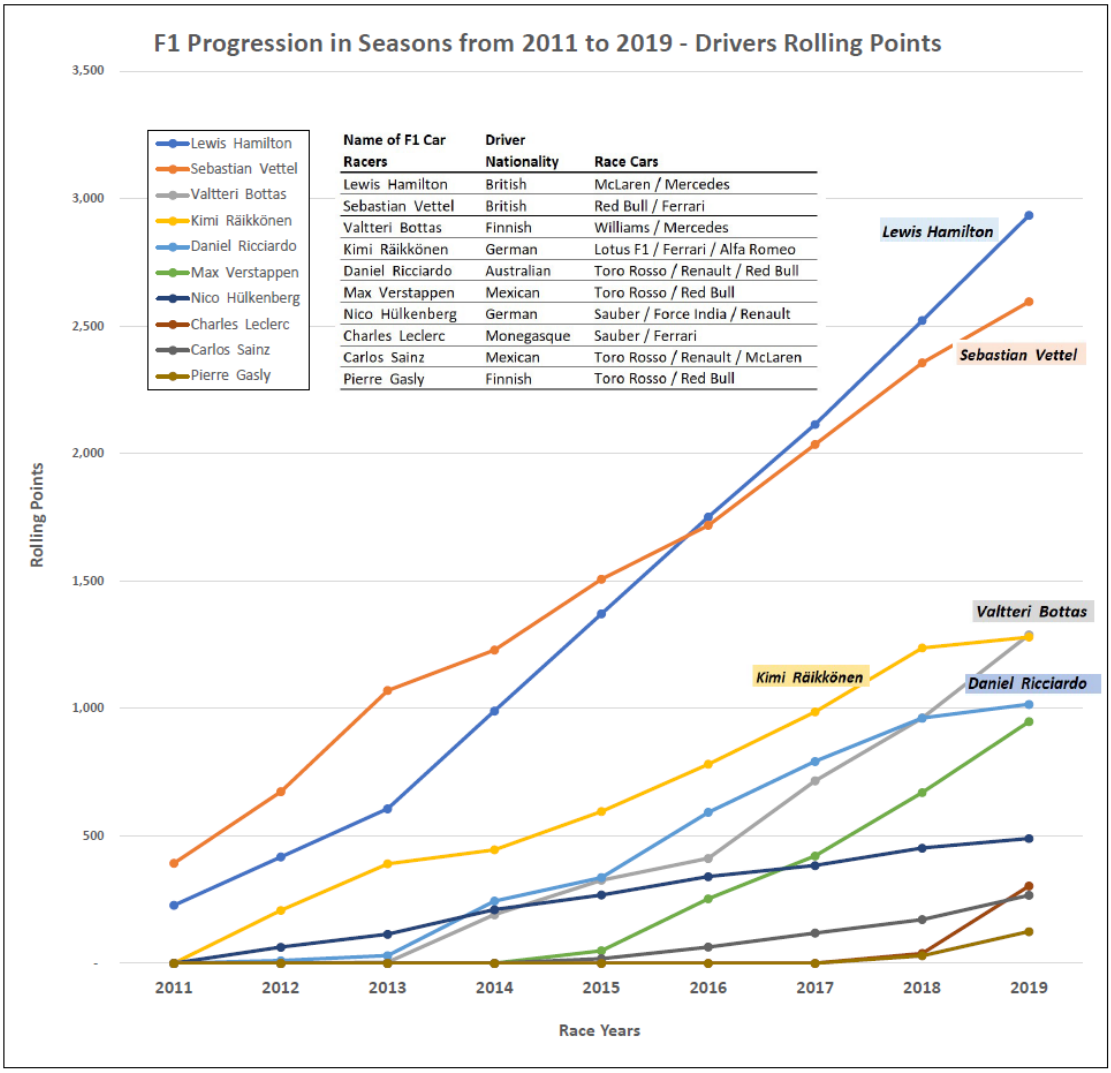
Formula 1 racing is a sport that generates an incredible amount of data around driver performance, car performance and race results. The volume of data is large in both the variables that are measured regularly and the time series of data that dates back decades. Formula 1 data from the last 8 years was joined with weather data and analyzed to recommend potential candidate drivers and constructor brands to buy sponsorship space on for the upcoming races - Canadian Grand Prix on June 14, French Grand Prix on June 28 and Austrian Grand Prix on July 5. Recommendations include a comparison of racer and constructor performance through data visualization as well as regression analysis used to predict the best racer to sponsor.

Historical data was analyzed in a variety of ways in order to make estimates on who might be the most probable winners of the next races (Canadian Grand Prix on June 14, French Grand Prix on June 28 and Austrian Grand Prix on July 5) as well as which car manufacturers are most likely to place.

While our original intent was to predict potential race times to predict winners, this approach did not provide a way to predict future winners easily as the factors that influence race time are primarily race-day based. We shifted to looking at variables that would predict driver point accumulation and found that it has more predictive power for our needs.

Analysis and Findings

Predicting future performance of drivers is based on previous performance. Looking at accumulation of points across the careers of the top drivers, there are clearly two top racers in this sport for the past 9 years: Lewis Hamilton and Sebastian Vettel (Figure 1).



Name of F1 Car Racers	Driver Nationality	Race Cars
Lewis Hamilton	British	McLaren / Mercedes
Sebastian Vettel	British	Red Bull / Ferrari
Valtteri Bottas	Finnish	Williams / Mercedes
Kimi Räikkönen	German	Lotus F1 / Ferrari / Alfa Romeo
Daniel Ricciardo	Australian	Toro Rosso / Renault / Red Bull
Max Verstappen	Mexican	Toro Rosso / Red Bull
Nico Hülkenberg	German	Sauber / Force India / Renault
Charles Leclerc	Monegasque	Sauber / Ferrari
Carlos Sainz	Mexican	Toro Rosso / Renault / McLaren
Pierre Gasly	Finnish	Toro Rosso / Red Bull

Note:

- **N1** Lewis Hamilton: McLaren (until 2012), and Mercedes (2013-present)
- **N2** Sebastian Vettel: Red Bull (2011-2014), and Ferrari (2015-present)
- **N3** Valtteri Bottas: Williams (2013-2016), and Mercedes (2017-present)
- **N4** Kimi Räikkönen: Lotus F1 (2012-2013), Ferrari (2014-2018), and Alfa Romeo (2019-present)
- **N5** Daniel Ricciardo: Toro Rosso (2012-2013), Red Bull (2014-2018), and Renault (2019-present)
- **N6** Max Verstappen: Toro Rosso (2015-2016), and Red Bull (2016-present)
- **N7** Nico Hülkenberg: Sauber (until 2013), Force India (2012-2016), and Renault (2017-present)
- **N8** Charles Leclerc: Sauber (2018) and Ferrari (2019-present)
- **N9** Carlos Sainz: Toro Rosso (2015-2017), Renault (2017-2018), and McLaren (2019-present)
- **N10** Pierre Gasly drove Red Bull (March 2019-August 2019), and Toro Rosso (September 2019-present)

Figure 1

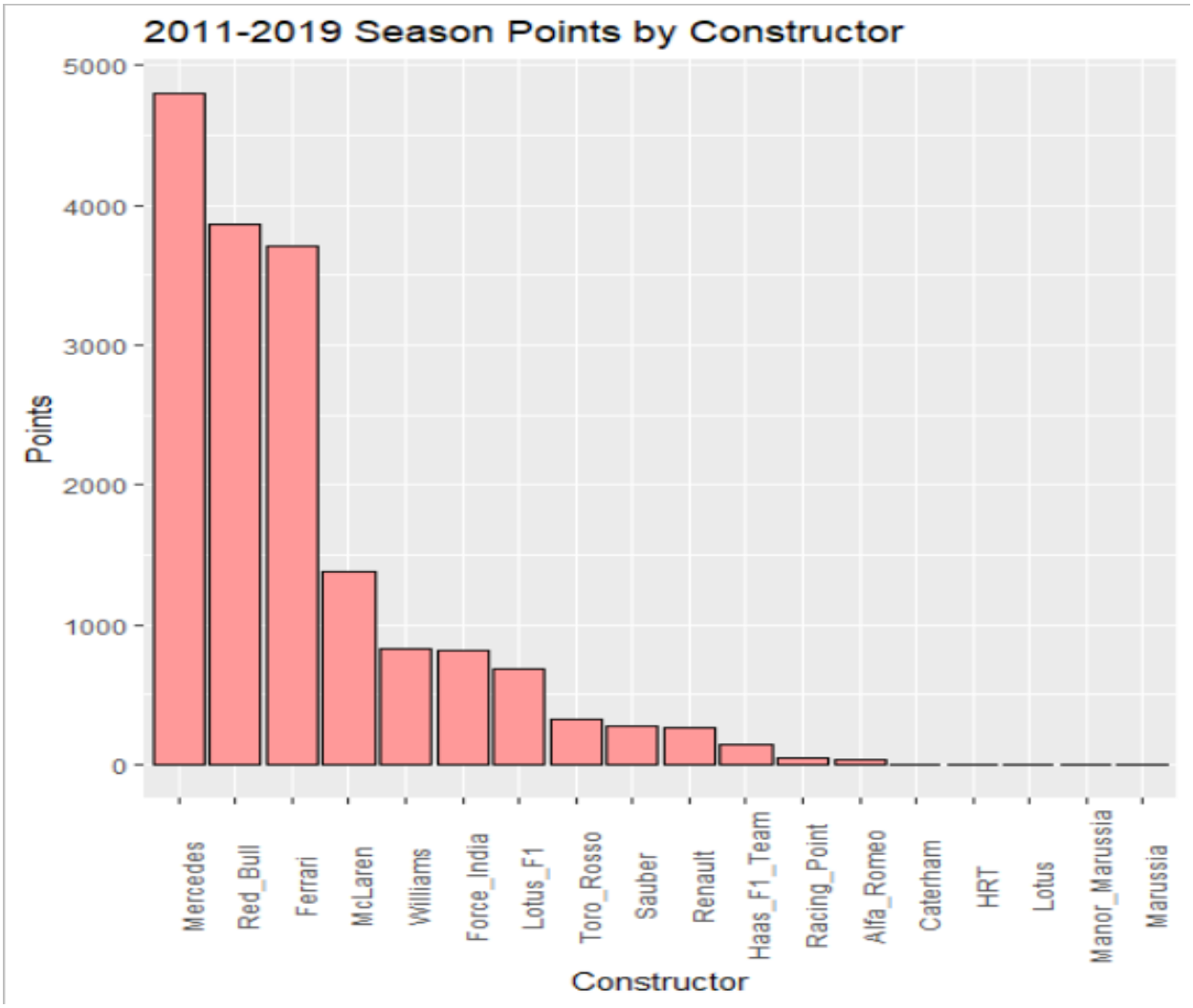


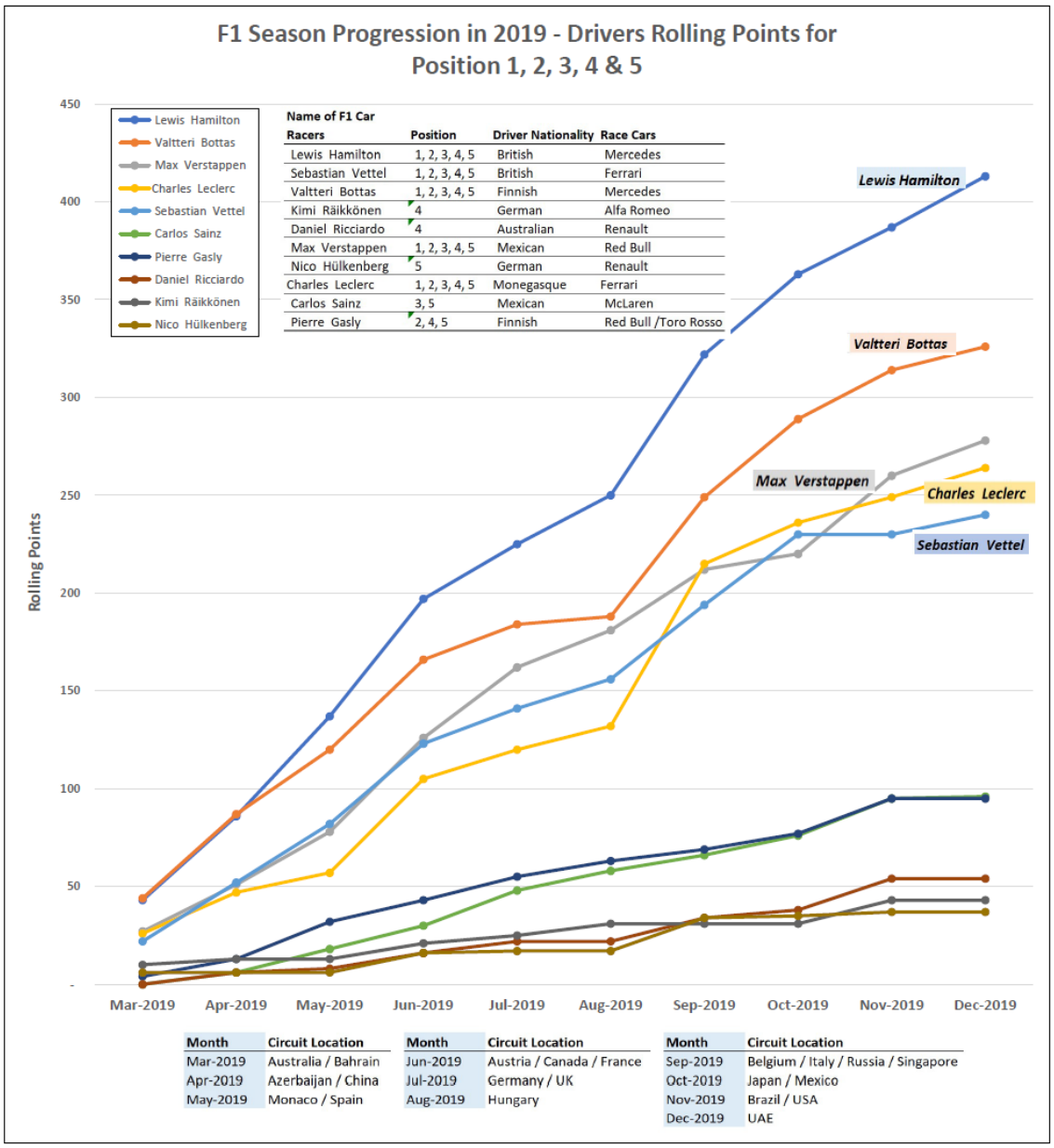
Figure 2

Similarly, there is a trend in the constructor of vehicles that see the most wins: Mercedes leading the pack, with Red Bull and Ferrari competing in a tight contest for second (Figure 2).

Multiple regression analysis found that drivers showed a positive correlation in earning more points from a variety of indicators: race starting position, age, history of winning, and constructor brand (see technical summary for detailed regression analysis results). The constructors noted in the regression as providing more points were Ferrari, McLaren, Racing Point, and Red Bull. These brands are interesting when you look at the history of performance of the constructors from the nine-year period analyzed. Mercedes is clearly a dominating name, and its curious that Mercedes did not show predictive value in driver points given its high points accumulation over the 9 year period analyzed.

While historic data is important in making an investment choice, the most recent year's performance is also something that should be considered.

2019 season results show that Lewis Hamilton continues to dominate the F1 races. Valterri Bottos who historically places as a far third place managed to convincingly take second place in 2019. Three other drivers in a close race for third place (Figure 3).



Name of F1 Car Racers	Position	Driver Nationality	Race Cars
Lewis Hamilton	1, 2, 3, 4, 5	British	Mercedes
Sebastian Vettel	1, 2, 3, 4, 5	British	Ferrari
Valtteri Bottas	1, 2, 3, 4, 5	Finnish	Mercedes
Kimi Räikkönen	4	German	Alfa Romeo
Daniel Ricciardo	4	Australian	Renault
Max Verstappen	1, 2, 3, 4, 5	Mexican	Red Bull
Nico Hülkenberg	5	German	Renault
Charles Leclerc	1, 2, 3, 4, 5	Monegasque	Ferrari
Carlos Sainz	3, 5	Mexican	McLaren
Pierre Gasly	2, 4, 5	Finnish	Red Bull /Toro Rosso

Month	Circuit Location
Mar-2019	Australia / Bahrain
Apr-2019	Azerbaijan / China
May-2019	Monaco / Spain

Month	Circuit Location
Jun-2019	Austria / Canada / France
Jul-2019	Germany / UK
Aug-2019	Hungary

Month	Circuit Location
Sep-2019	Belgium / Italy / Russia / Singapore
Oct-2019	Japan / Mexico
Nov-2019	Brazil / USA
Dec-2019	UAE

*Note:

N1 Pierre Gasly drove Red Bull from March 2019 to August 2019, and Toro Rosso from September 2019 to December 2019.

Figure 3

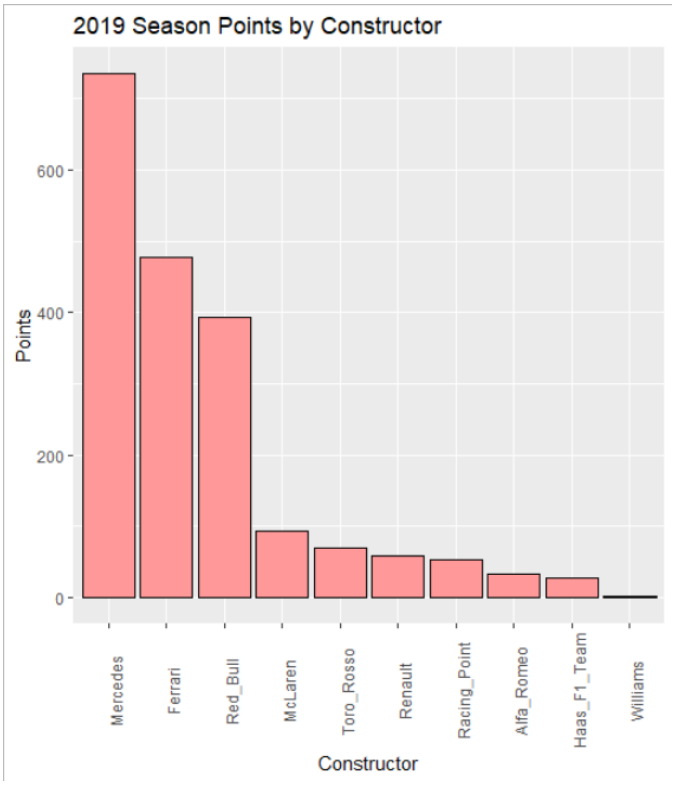


Figure 4

Looking at how the constructors did for 2019, we see that Mercedes continues to dominate the races with Ferrari and Red Bull continuing to battle for second place. 2019 shows Ferrari gaining ground on Red Bull's small lead from the last 8 years (Figure 4).

Recommendations

The question remains, who is the right driver and constructor to sponsor?

High Budget Option: While Lewis Hamilton driving a Mercedes has a dominating lead from of the pack, he will surely be the most sought-after driver to sponsor and thus would have the highest premium on space. Investing in Lewis's car would be the obvious choice, although the return on investment may not be there.

Medium Budget Option: Another potential driver to sponsor would be Sebastian Vettel. While the last season he has not had the strongest performance, he does have a long history of doing well in races and may recover in the future. Predictions based on regression analysis show him taking another second place standing (see below for the points prediction of the next race). He also drives a Ferrari which vies for the number two spot for constructors with Red Bull. Regression analysis on constructors that influence points shows Ferrari having a positive correlation. Sebastian, like Lewis, also has a long history as a successful F1 driver and there is a strong likelihood of him continuing to perform. Given his history, sponsoring him may be expensive.

Low Budget Option: The final driver for consideration would be more of a long-term play. Max Verstappen is a younger driver whose career is on the rise. While he finished in 3rd place in 2019, we project him to also finish next year in 3rd place through regression analysis prediction, and ahead of the number two driver this year. We also know that more wins and older age are predictors for more success in Formula 1, which means Max may have a long career of wins ahead of him. He drives a Red Bull car as well which as stated previous, battles Ferrari regularly for the number two spot for constructors and also has shown through regression analysis as having a positive impact on driver points. Given his shorter history as an F1 driver, and his building performance, we see him as the best deal to pick up for potential long-run value.

	Lewis_Hamilton	Valtteri_Bottas	Max_Verstappen	Charles_Leclerc	Sebastian_Vettel
Points Prediction	19.24623	4.544199	9.539726	5.758261	14.69752

Technical Summary

Data Cleaning

The Formula 1 data set is made up of thirteen tables which were reviewed, and a data dictionary created. Variables were identified in each file and a short description created for each.
Initial data cleaning:

1. Removing columns that were irrelevant to the regression model
2. Substituting "\\N" to NA to make R work easier
3. Removing duplicate columns
4. Converting all lap times into milliseconds to maintain consistency

Data import procedures were expressly coded into R in order to ensure the correct data types and ease of data joining.

Data Joining and Merging

When attempting to predict the outcome of F1 races, there are multiple variables that should be considered. The original data was captured in 13 CSV files each representing an aspect of the race that can be categorized into driver information and race information. A unified data frame was created that included all the information needed to do predictive analytics, visualization, and feature engineering.

DriverID and RaceID were the two key identifiers for majority of the original dataset. Together, they serve as a unique identifier that ties drivers to specific races. We took a phased approach to come up with the unified data frame by leveraging DriverID and RaceID during joining and merging operations:

1. An entity relationship diagram was created for the data set
2. All the driver demographic information was linked to the driverId from the drivers_df data frame. Therefore, anytime that we refer to drivers_df, we would have the accompanying demographic information included.
3. Race information was split between most of the remaining individual CSV files. The same method used for driver information was repeated. In addition to joining the key quantitative variables for to support the regression model, a few descriptive variables were included into the final unified data frame. This created clarity around what the exact race circuit or car manufacture was by name rather than a number. This qualitative information was pulled while also joining on constructorId and raceId.

```
constructorName<-sqldf('SELECT constructors_df.name,results_df.driverId, results_df.raceId from constructors_df,results_df where constructors_df.constructorId = results_df.constructorId ORDER BY results_df.raceId')
```

4. The final joining was to combine the data frames that were created for both categories into one unified data frame. As identified during planning, the DriverID and RaceID were used as identifiers to join the data frames together. The code snippet below shows the columns for unified data frame based on analytics requirements and the joining operation

```
F1_total<-sqldf('SELECT F1_driver.raceId, constructorName.driverId, driverName, driverAge,
nationality, name, grid, points,position, fastestLapTime, fastestLapSpeed, milliseconds, q1, q2, q3,
stop, pitstopTime
        FROM F1_driver JOIN constructorName ON F1_driver.driverId = constructorName.driverId
        AND F1_driver.raceID = constructorName.raceId')
```

Feature Engineering

Business Description

Our feature engineering work was aimed at identifying and developing features to predict which drivers and constructors will finish the 2020 season with the most points: to help drive our sponsorship decision. We interpreted our data based on the applicable underlying theory: the rules and parameters of F1 racing; and identified several potential variables of interest. We highlight a few key ones here:

1. *Constructors*: Constructors are the organizations behind the teams of engineers and drivers. There are currently 10 constructors competing in the 2020 season. Constructor budgets vary widely by team, with leading teams spending as much as $485 million per annum[1]. Sponsorships are a major component of constructor revenue[2]. Constructors tally the points awarded to all drivers in their team, and the constructor with the most points at the end of the racing season is awarded the constructors' championship.
2. *Drivers*: Points are awarded to the first 10 finishers in a race, in the following order: 25,18,15,12,10,8,6,4,2,1. Drivers with the most points at the end of the season are awarded the championship.
3. *Qualifying time*: Drivers run up to 3 qualifying races before the official race. We are looking for drivers with consistently low average qualifying (q) times compared to their peers
4. *Grid*: Driver starting positions in a race are determined by min q; we expect drivers with positions 1-3 in a race have an advantage.
5. *Position*: finishing position in the race
6. *Pit stops*: Pit stops are in-race stops for tire changes and fuel; drivers can make as many as they wish in a race. We expect that number and duration of pit stops in a race will impact finish times and consequently points awarded. Top constructors keep pit stops to a minimum since pit stops increase driver race finish time. Consequently, pit stop durations are a good indicator of constructor team expertise.
7. *Circuits*: There are 22 race circuit locations on the 2020 F1 schedule[3]. Each circuit introduces variations due to weather and track difficulty. We expect that circuit familiarity will be positively correlated with driver performance and this is likely reflected in the number of years a driver has been racing.
8. *Fastest Lap Time*: Drivers complete a specified number of laps around a circuit in completing a race. We expect fastest lap times to provide some indication of race finish positions and therefore points.

---

[1] Smith, C. (2019, 11 26). *Formula One's Most Valuable Teams: Ferrari And Mercedes Gain Ground Amid A Cost-Cutting Tug-Of-War*. Retrieved from forbes.com: https://www.forbes.com/sites/chrissmith/2019/11/26/formula-one-team-values-ferrari-mercedes/#417862e41ddb

[2] Sylt, C. (2019, 04 26). *Revealed: The $285 Million Cost Of Winning The F1 Championship*. Retrieved from forbes.com: https://www.forbes.com/sites/csylt/2019/04/26/revealed-the-285-million-cost-of-winning-the-f1-championship/#767fd3223d81

[3] Formula One World Championship Limited. (2003-2020). *2020 FIA FORMULA ONE WORLD CHAMPIONSHIP™ RACE CALENDAR*. Retrieved from formula1.com: https://www.formula1.com/en/latest/article.f1-schedule-2020-latest-information.3P0b3hJYdFDm9xFieAYqCS.html

Technical Description

1. **Feature Interactions**
    - Pit stops: Originally, observations in our raw data were listed by individual pit stops per race and driver, resulting in multiple lines for each driver in the same race, i.e. 1 line per pit stop. Our first feature engineering step was to sum pit stop times by driver and raceid, so that individual observations are now rolled by driver and race id only. Other pit stop-related interactions we created include average pit stop time and ratio of pit stop to driver race time. In all cases, the smaller the number the better.

```
# Add interaction features to F1_total: average and total pitstop time per driver race, ratio of pitstop to race time

weather_allcircuits_df <- tryCatch ({
  read_csv(paste0(directory_csv, "weather_allcircuits.csv"))
}, error = function(err) {
  read_csv(file.choose())
})

F1_total_FE <- sqldf('SELECT raceId, driverId, driverName, driverAge, nationality, name, position, fastestLapTime, fastestLapSpeed,
milliseconds, q1, q2, q3, stop, SUM(pitstopTime) AS Pitstop_Total_Time, AVG(pitstopTime) AS AvgPitStopTime FROM F1_total GROUP BY
driverid, raceid ORDER BY raceid')
F1_total_FE$ratioPittoTotal <- F1_total_FE$Pitstop_Total_Time/F1_total_FE$milliseconds
```

    - Minimum qualifying times: We calculate minimum qualifying time of up to 3 qualifying races a driver undertakes during qualifying. We anticipate that minimum qualifying time will be a predictor of how fast a driver will finish a race and so how many points they can achieve in a race.
    - Wins: We used wins data to calculate rolling wins for drivers up to specified race date.
    - Races Completed: The more races a driver completes, the more experience they accumulate. We created interaction feature rolling race completed by partitioning and summing driver wins by race date.
    - Points: We created interaction feature rolling points by partitioning and summing points awarded to a driver over their racing career starting from 2009 by race date.

```
F1_master2<-sqldf('select  date, location, raceId, driverId,driver_wins,
                  sum(driver_wins) over (partition by driverId order by date ) as rolling_wins,
                  sum(points) over (partition by driverId order by date ) as rolling_points,
                  count(raceId) over (partition by driverId order by date ) as rolling_race_completed,
                  (sum(points) over (partition by driverId order by date ))/(count(raceId) over (partition by driverId order by
date )) as avg_points_per_race,
                  (sum(driver_wins) over (partition by driverId order by date ))/(count(raceId) over (partition by driverId
order by date )) as avg_wins_per_race
                 from F1_master
                 ')
head(F1_master2)
```

    - Driver age: we were curious if the age of drivers on race day had any predictive value
    - Constructors: to look for any predictive influences from the various constructors, we created dummy variables for each constructor to look for any attribution during regression analysis
    - Nationality: to look for any predictive influences from the various constructors, we created dummy variables for each constructor to look for any attribution during regression analysis

2. **External data**

    Our data did not include weather data for the race circuits. We anticipate that weather variables such as temperature, humidity and precipitation will influence driver performance and possibly car performance (temperature impacts tires for example). We collected weather data using the riem package in R, see R script “Weather”. Riem access weather data from the Iowa Environment Mesonet. We selected this package because of its ease of use relative to other options for weather data collection and importing. External weather data import steps:

- Generate list of riem weather networks. Weather networks generate a list of 237 country weather network codes across the globe.
- Using the circuit cities/countries specified in our F1 data, we selected created weather station list by country from the weather networks. Weather stations are typically 4 letter codes that identify a specific weather station by city, latitude and longitude.
- Using the latitude and longitude data from our circuit cities, we selected the closest weather station codes and proceeded to pull weather data by the desired date range: 2009 to 2019.
- Weather data by city output was presented in daily 24-hour observations. We performed some feature engineering on each city's weather data frame to produce one observation per day instead of 24, creating MAX temp, MIN temp, AVG temp, AVG Rel Humidity, Max Precipitation and MaxFeel.
- We combined the individual city weather data frames into one weather_allcircuits data frame which we then mapped to our F1 data by race date and city. We had a few missing weather dates due to issues with weather stations in certain dates/years. We manually imputed missing weather dates by selecting from alternative online weather data source.

```
install.packages("riem")
library(riem)


# list of weather networks across globe by country
weather_networks<- riem_networks()

# list of weather stations in a given network
australia_stations <- riem_stations("AU__ASOS")

# weather data for a specified weather station by specified date range
### pull weather data by city/country and weather station
### add city, format weather date, select desired weather variables by city
### feature engineer weather data to obtain daily max / min / avg temp, prec, humidty

australia_YMEN <- riem_measures("YMEN","2009-01-01", "2019-12-31")
australia_YMEN$city <- "Melbourne"
australia_YMEN$date <- as.Date(australia_YMEN$valid)
australia_YMEN_weather <- sqldf('SELECT city, station, date, MAX(tmpf), MIN(tmpf), AVG(tmpf), AVG(relh),
                                 MAX(p01i), MAX(feel) FROM australia_YMEN GROUP BY date')

##
##rowbind city weather data frames into one data frame for all circuit cities

weather_allcircuits <- rbind(australia_YMEN_weather,austria_LOWG_weather,azerbaijan_UBBB_weather, bahrain_OBBI_weather,
                             belgium_EBBR_weather,brazil_SBSP_weather, canada_CYUL_weather, china_ZSSS_weather,
                             france_LFMC_weather, germany_EDFM_weather,germany_EDDK_weather,hungary_LHBP_weather,
                             india_VIDD_weather, italy_LIML_weather,japan_RJGG_weather, korea_RKJB_weather,
                             malaysia_WMKK_weather, mexico_MMMX_weather, netherlands_EHAM_weather, russia_URSS_weather,
                             singapore_WSAP_weather,spain_LEBL_weather,turkey_LTFJ_weather, uae_OMAA_weather,
                             uK_EGBB_weather,us_AUS_weather, vietnam_VVNB_weather)
```

Caveats and assumptions

1. In the absence of engine performance data, we assume constructors are a suitable proxy for engine performance. This is reasonable given that constructors typically have long term contractual relationships with engine manufacturers.
2. Tires: we did not have tire data. Constructors chose from a limited selection of tires in each race per F1 rules. The true differentiator is constructor engineering team expertise in determining what tires to use given expected race conditions and at what point in the race to make tire changes. We determined that constructor expertise in tire selection is captured by the constructor feature as well.
3. Chassis: the vehicle base frame design follows F1 rules, with room for differentiating customizations. Again, this expertise resides within constructor organizations and is improved over several years of racing. As such, the constructor feature captures variations due to chassis design:

Regression Analysis

Two models were created to look at assessing racer and constructor wins – one that predicts the average points per race and another that predicts the fastest race time.

Outliers and Missing Values
A box plot and Log transformation on the "milliseconds" variable was used to identify any outliers of this key variable. Several races were found that took much longer to complete than normal and were removed to ensure a normal distribution of the "milliseconds" variable (removing lg_milliseconds greater than 6.85, which is equivalent to 2 hours of race time). Regular Formula one races typically last about 90 mins, with the maximum being 2 hours of race time due to safety for the races. Any time over this has significant stoppage time due to accidents. The records we removed accounted for the top 10% of the longest races in our data set.
All records that had an NA value in "milliseconds" column were removed as this is the value we are looking to predict. The resulting master file had between 2 and 13 missing values for "fastest lap time" "fastest lap speed", and "position". Given the very small number of records in this set, we removed these records completely as well. The final dataset had 1600 records.

Predictive Modeling Steps We Took:

1. Setting Up the Model
   - Test & Train: We have split the whole dataset into 70% train and 30% test. We test our models on the "train" dataset.

2. Run & Refine the Model
   - Check collinearity among variables using VIF values
   - Review heteroscedasticity plots & NCV test
   - Remove variables above 0.05 alpha

3. Validate the Model
   - Compare test and train

Model 1: Predicting Average Points per race

Regression equation: **avg_points_per_race** = -2.38 + (-0.18)grid + (0.26)driverAge_raceday + (-0.0167)rolling_race_completed + (3.467)avg_wins_per_race + (2.40)nationality_Australian + (1.082)nationality_Brazilian + (1.831)nationality_British + (4.289)nationality_Dutch + (2.836)nationality_Finnish + (2.337)nationality_French + (1.186)nationality_German + (1.429)nationality_Mexican + (2.46)nationality_Russian + (3.092)nationality_Spanish + (1.891)name_Ferrari + (1.671)name_McLaren + (1.863)name_Racing_Point + (1.185)name_Red_Bull + (-1.796)name_Toro_Rosso

Avg_points_per_race is run with all X variables. After removing the non-significant X variables, the final model shows the variables that are significant at alpha = 0.05. The regression equation was plotted and the Residual vs Fitted, and Q_Q plots were normal, which indicated no heteroscedasticity in the model. We also performed an NCV test which resulted in a P value of 0.889, signaling again no heteroscedasticity. Collinearity was also tested for and no variables were found that strongly (with VIF values above 10) inflated the coefficient of other variables due to multicollinearity. Validation of the model was completed on the "test" selection of data which resulted in a slightly lower R2, at 0.84, compared to 0.87 from our "train" dataset. This indicated that our model was fairly accurate.

Linear Regression Predictions
Using the regression model, we did a prediction for the upcoming races which showed the following point predictions for the top 5 racers:

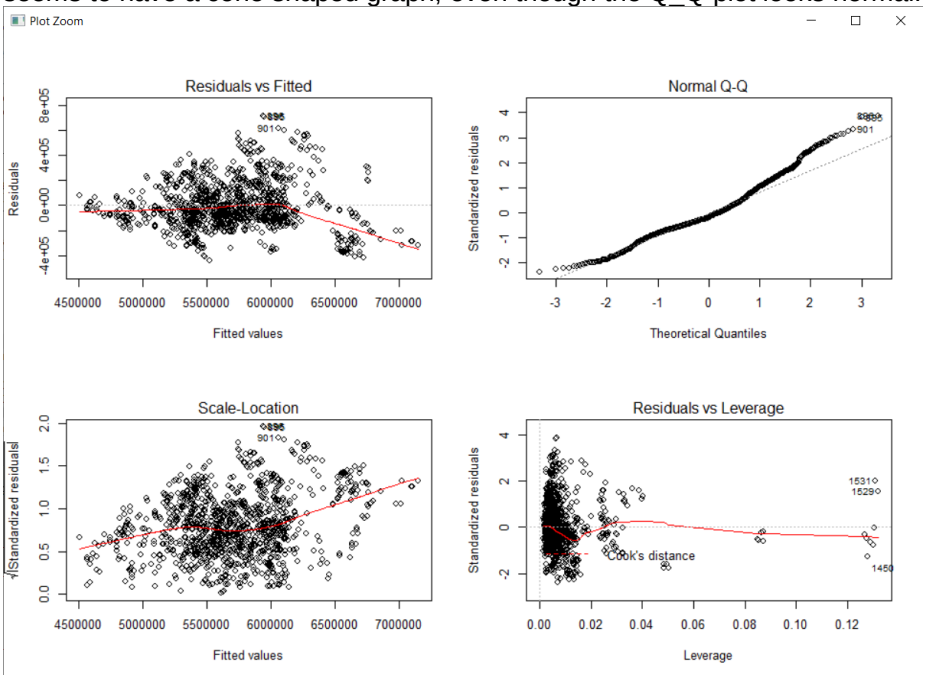
	Lewis_Hamilton	Valtteri_Bottas	Max_Verstappen	Charles_Leclerc	Sebastian_Vettel
Points Prediction	19.24623	4.544199	9.539726	5.758261	14.69752

Model 2: Predicting fastest race speed (lowest milliseconds

Regression equation: **milliseconds** = (9.163e+06) + (3.565e+03)grid + (4.251)fastestLapTime + (-2.152e+04)fastestLapSpeed + (9.995e+04)stop + (1.875)AvgPitStopTime + (6.639e+02)rolling_race_completed + (3.304e+03)AVG_relh + (7.083e+04)circuit_GrandPrix_Name_Canadian_Grand_Prix + (1.680e+05)name_Alfa_Romeo

Milliseconds was run with all X variables. After removing the non-significant X variables, the final model shows the variables that are significant at alpha = 0.05. Running a test for collinearity show that VIF values are within tolerance with no particular variables that strongly inflated the coefficient of other variables due to multicollinearity. Validation of the model was completed on the "test" selection of data and the R2 was actually higher, at 0.869, compared to 0.847 from the "train" dataset, which indicates that our model is better.

At this point, all variables seemed significant; however, the Residual vs Fitted plot of the regression seems to have a cone shaped graph, even though the Q_Q plot looks normal.



This means that heteroskedasticity might be present. An NCV test was run to assess heteroskedasticity and it proved that there was heteroskedasticity present in the model. Therefore, the variables in the model may not be as powerful in predicting "milliseconds" as previously indicated. A Robust Linear Regression was run using the "HCCME" method to correct their estimates.

```
Coefficients:
                                          Estimate Std. Error  t value  Pr(>|t|)   CI Lower   CI Upper   DF
(Intercept)                              9.163e+06  1.024e+05  89.4732 0.000e+00  8.962e+06  9.363e+06 1119
grid                                     3.426e+03  1.131e+03   3.0309 2.495e-03  1.208e+03  5.645e+03 1119
fastestLapTime                           4.309e+00  5.901e-01   7.3031 5.327e-13  3.152e+00  5.467e+00 1119
fastestLapSpeed                         -2.154e+04  3.707e+02 -58.0936 0.000e+00 -2.226e+04 -2.081e+04 1119
stop                                     9.881e+04  6.928e+03  14.2632 1.559e-42  8.522e+04  1.124e+05 1119
AvgPitStopTime                           1.877e+00  1.093e-01  17.1781 7.000e-59  1.663e+00  2.092e+00 1119
rolling_race_completed                   6.637e+02  1.404e+02   4.7264 2.576e-06  3.882e+02  9.392e+02 1119
AVG_relh                                 3.262e+03  5.056e+02   6.4517 1.643e-10  2.270e+03  4.254e+03 1119
circuit_GrandPrix_Name_Canadian_Grand_Prix 7.084e+04  2.329e+04   3.0413 2.411e-03  2.514e+04  1.165e+05 1119
nationality_Polish                       3.925e+05  2.097e+06   0.1871 8.516e-01 -3.723e+06  4.508e+06 1119
name_Alfa_Romeo                          1.702e+05  7.676e+04   2.2174 2.680e-02  1.960e+04  3.208e+05 1119

Multiple R-squared:  0.8473 ,   Adjusted R-squared:  0.8459
F-statistic:   617 on 10 and 1119 DF,  p-value: < 2.2e-16
```

This turns out that the nationality_Polish variable was no longer significant and was removed from the final model. Below is the R code for our final regression model:

```
rob_reg_train2<- lm_robust(milliseconds~

                             grid+
                             fastestLapTime+
                             fastestLapSpeed+

                             stop+
                             AvgPitStopTime+

                             rolling_race_completed+
                             AVG_relh+
                             circuit_GrandPrix_Name_Canadian_Grand_Prix+

                             name_Alfa_Romeo,
                           train,
                           se_type="HC3")

summary(rob_reg_train2)
```

```
Coefficients:
                                          Estimate Std. Error t value  Pr(>|t|)   CI Lower   CI Upper   DF
(Intercept)                              9.160e+06  1.024e+05  89.453 0.000e+00  8.959e+06  9.361e+06 1120
grid                                     3.565e+03  1.135e+03   3.141 1.725e-03  1.338e+03  5.792e+03 1120
fastestLapTime                           4.251e+00  5.909e-01   7.194 1.152e-12  3.091e+00  5.410e+00 1120
fastestLapSpeed                         -2.152e+04  3.708e+02 -58.057 0.000e+00 -2.225e+04 -2.080e+04 1120
stop                                     9.995e+04  6.928e+03  14.426 2.123e-43  8.635e+04  1.135e+05 1120
AvgPitStopTime                           1.875e+00  1.091e-01  17.177 6.976e-59  1.661e+00  2.089e+00 1120
rolling_race_completed                   6.639e+02  1.403e+02   4.733 2.499e-06  3.887e+02  9.392e+02 1120
AVG_relh                                 3.304e+03  5.058e+02   6.531 9.883e-11  2.311e+03  4.296e+03 1120
circuit_GrandPrix_Name_Canadian_Grand_Prix 7.083e+04  2.334e+04   3.034 2.465e-03  2.503e+04  1.166e+05 1120
name_Alfa_Romeo                          1.680e+05  7.623e+04   2.204 2.772e-02  1.845e+04  3.176e+05 1120

Multiple R-squared:  0.8467 ,   Adjusted R-squared:  0.8455
F-statistic: 685.5 on 9 and 1120 DF,  p-value: < 2.2e-16
```

Data Visualization
We created visualizations for this project with a combination of Tableau, R Studio and Excel. Our data visualization work focuses on the sponsorship perspective and the interpretation of our dataset that is aligned with the identification of the best performing drivers and constructors, the best performing records, and any potential impact of the weather condition and pitstop time on those performances, given that the sponsorship interest lies in the top 3 to 5 positions and the circuit locations related to the next three races taking place (Canadian Grand Prix on June 14, French Grand Prix on June 28 and Austrian Grand Prix on July 5).

Based on the data visualization presented in tables and charts, the key highlights of the performance of drivers and constructors and their records and other factors affecting their performance are provided as below:

Rolling points for drivers – Completed in excel
F1 Line Charts The line charts were created in Excel based on the F1 Season Progression tables below.

F1 Progression in Seasons 2011-2019 Table
10 drivers within the positions from 1 to 5 were selected based on their race attendances and performance. For each driver, the latest rolling points for each season from 2011 to 2019 were obtained from the datasets, and summarized in table for each season, which is sorted by the highest to the lowest rolling points of 10 drivers.

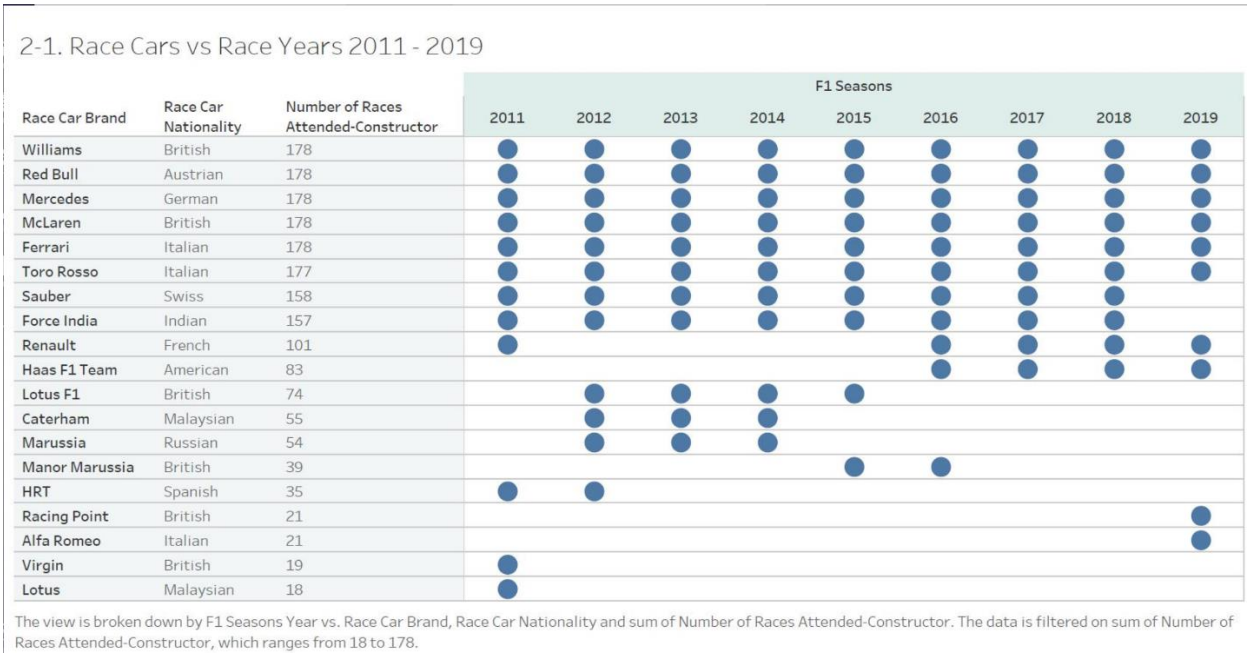
F1 Progression Season 2019 Table
10 drivers in the F1 Progression in Seasons from 2011 to 2019 were also selected for this chart, which breaks down the season 2019 to each month and the respective circuit locations for the F1 races. For each driver, the points for each month from March to December 2019 were computed and rolling points for each month were calculated based on points in each month. These data were summarized in table, which is sorted by the highest to the lowest rolling points of 10 drivers.

Average constructor rolling points – Complete in R

Average constructor points for 2011-2019 and just 2019
Visualizations were completed in R using the GGPLOT package. Data was subsetted from the master data source, aggregated by constructor and visualized.

Some additional Tableau Visualizations
Some of our data visualization work was done in Tableau which assisted in the analysis of the data set.

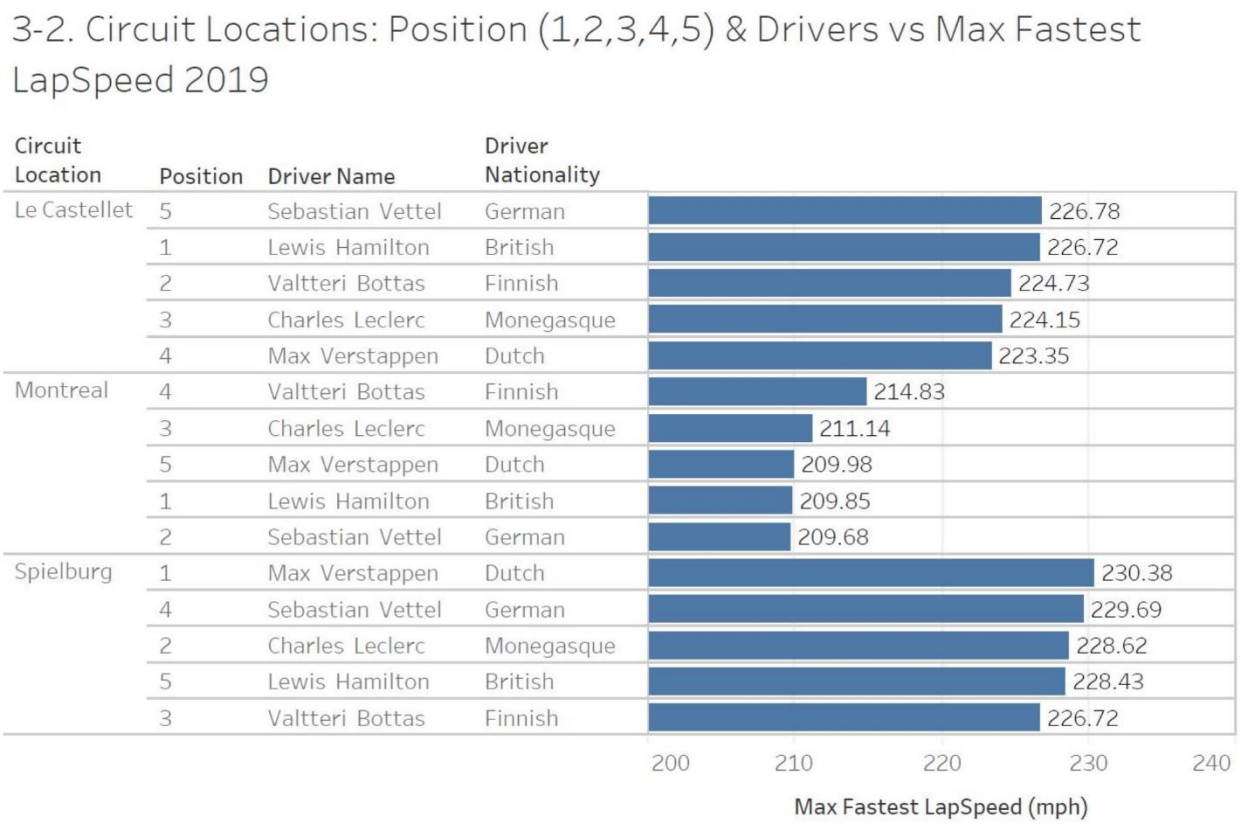
Constructer attendance for past 8 years



The view is broken down by F1 Seasons, constructors and the sum of the number of Races Attended-Constructor. The data is filtered on the sum of Number of Races Attended-Constructor, ranging from 18 to 178.

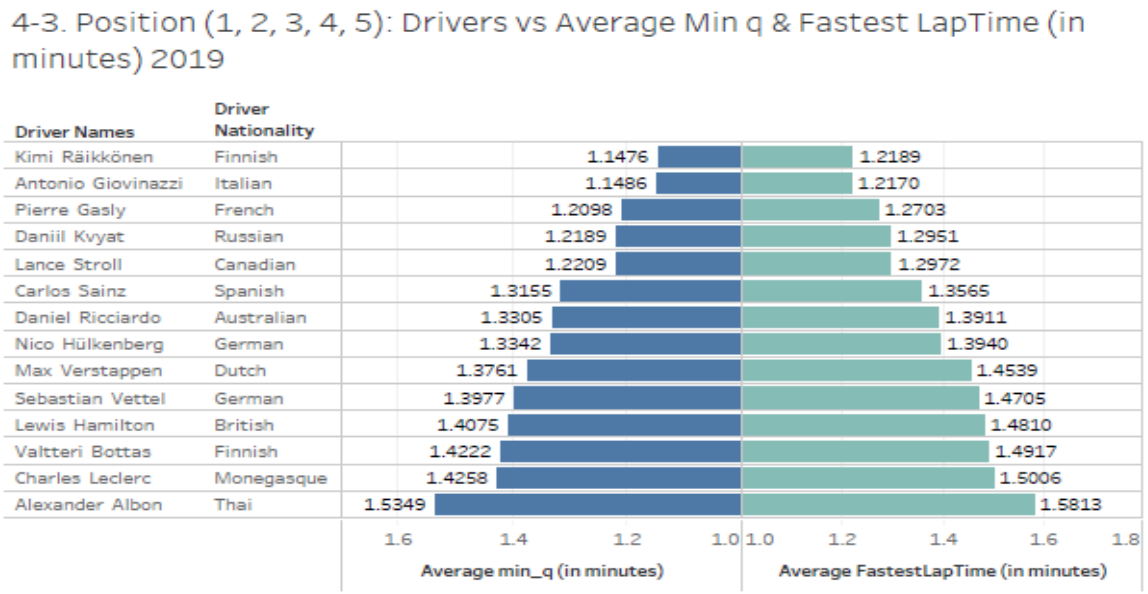
Under the measure value, we created the calculated field, "Number of Races Attendance-Constructor", and used the following codes ({fixed [Race Car Brand]:countd([raceId])})to calculate the number of race attendances throughout the F1 seasons from 2011 to 2019.

Top speed for three circuits by driver



Maximum of the fastest lap speed is broken down by the circuit location, position and driver names, where data is filtered on the season 2019 and the view is filtered on position and circuit location. The position filter keeps 1 to 5, and the circuit location filter keeps Montreal, Le Castellet and Spielburg.

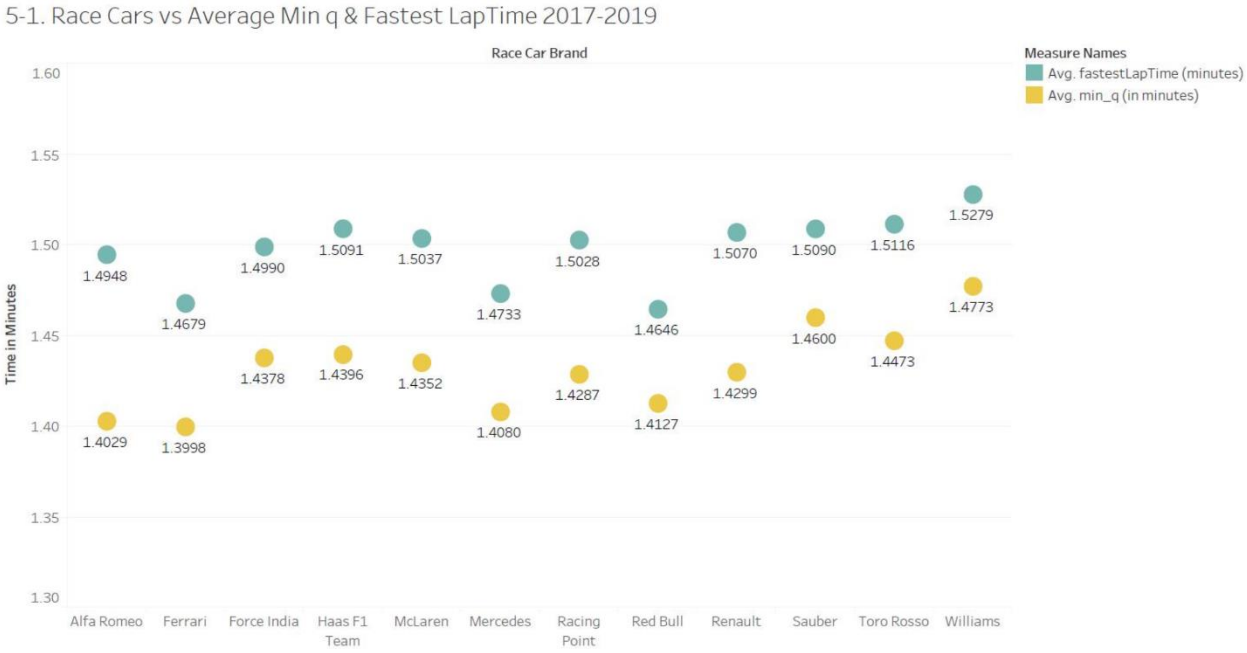
Qualifying and fastest lap time for top drivers



Average of min_q (in minutes) and average of fastestLapTime (minutes) for each Driver Nationality broken down by Driver Names. The data is filtered on year and Position. The year filter ranges from 2019 to 2019. The Position filter keeps 1, 2, 3, 4 and 5.

Average of the minimum qualifying time and fastest lap time is broken down by drivers' names, where data is filtered on the season 2019 and the position that keeps from 1 to 5.

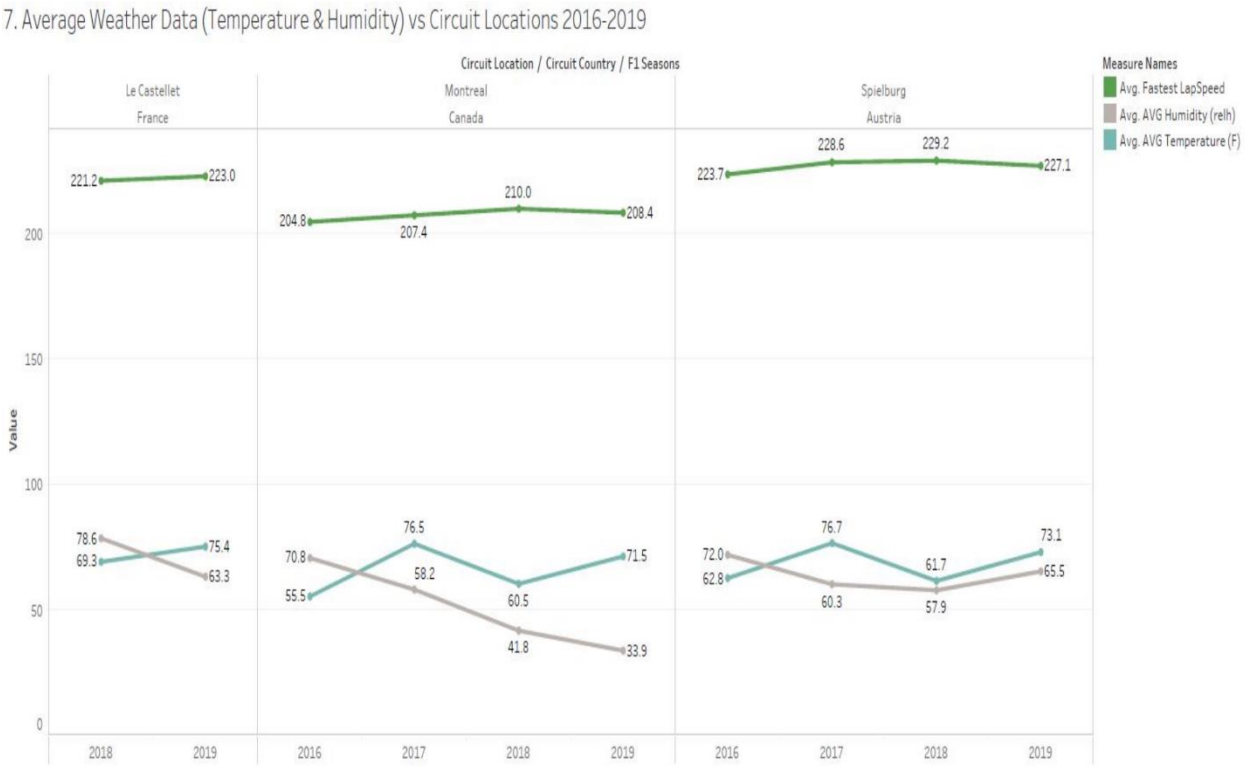
Correlation between qualifying and fastest lap time



It shows the average values of the fastest lap time and minimum qualifying time in minutes for each constructor. Color shows details about fastest lap time and minimum qualifying time, where the yellow

represents the average values of the minimum qualifying time and green, the average values of the fastest lap time. Data is filtered on the season 2019 and the view on 19 constructors.

Weather information for three circuit locations



It shows the trends of the average fastest lap speed, average of the humidity (relh) and temperature (F) throughout the F1 seasons broken down by the circuit locations. The color shows details about the average of the fastest lap speed in green, the average of the humidity in grey, and the average of the temperature in blue. The view is filtered on the circuit location of which filter keeps Le Castellet, Montreal and Spielburg, and the F1 seasons ranging from 2016 to 2019.

Caveats and assumptions

- As for the line chart in 7 above, the data related to the temperature and humidity (weather data) were obtained from the weather_allcircuits data frame mapped to the F1_master data frame. There was an issue related to the missing weather data in the F1_master data frame. This issue was dealt with by restricting the visualization to the three circuit locations and the three seasons from 2016 to 2019 only.
- We removed drivers' data from our visualization, which showed no race records before 2016, so that the overall data visualization can reflect the most recent seasons.
- The circuit locations were limited to Montreal, Le Castellet, and Spielburg based on the next three races taking place (Canadian Grand Prix on June 14, French Grand Prix on June 28 and Austrian Grand Prix on July 5).
- For the charts and tables with the position filter, the positions were limited based on the assumption that the sponsorship interest lies in the top 3 to 5 positions.