Final Project: Option 2

Milla Nielsen

2025-06-10

Introduction

Formula 1 is the number one motorsport in the world generating billions of dollars a year in revenue. Each season only the top 20 best drivers and 10 best constructors complete for the championships. Thus with very few highly competitive spots, understanding what it means to be a winning driver could mean the difference between keeping your position for next season and having to find a new career (1). In the below report I will use SQL in conjunction with R to explore what factors go into being a winning driver. I will investigate the impact of constructors and individual drivers statistics as well other factors to begin to understand why certain outcomes occur in races.

Exploratory Data Analysis

Brief dataset setup:

```r
library(DBI)
```

```
## Warning: package 'DBI' was built under R version 4.4.1
```

```r
library(dplyr, warn.conflicts = FALSE)


con <- dbConnect(RMariaDB::MariaDB(),
    host = "relational.fel.cvut.cz",
    port = 3306,
    username = "guest",
    password = "ctu-relational",
    dbname = "ErgastF1"
)
```

Constructors

In F1 not only do the drivers compete against each other but the car constructors themselves compete as well. Each constructor team is comprised of 2 drivers with the winning constructor being the team with the most aggregated points from it's two drivers across the season (2). Therefore it is important to investigate what role constructors play in driver wins i.e. does the constructor make the race winner. The provided F1 dataset has a table of constructors, constructor standings, races, and season results each race that can be used in this task.

Which constructors won each season?

In this dataset there is a variable called position in the constructorStandings table which we will take advantage of as it reports the position of the constructor at each race. We will look at the wins variable at the very last race as it is already a running total and therefore the last race will contain the entire seasons information under the wins variable.

```
constructor_wins <- dbGetQuery(con, "
  WITH last_race_of_season AS (
  SELECT year, MAX(raceId) AS last_race
  FROM races
  GROUP BY year)
    SELECT seasons.year, constructors.name, wins
    FROM constructorStandings
    INNER JOIN races
    ON constructorStandings.raceId = races.raceId
    INNER JOIN last_race_of_season
    ON races.raceID = last_race_of_season.last_race
    INNER JOIN seasons
    ON races.year = seasons.year
    INNER JOIN constructors
    ON constructorStandings.constructorId = constructors.constructorId
    WHERE constructorStandings.position = 1
    ORDER BY seasons.year;")
head(constructor_wins)
```

```
##   year          name wins
## 1 1958       Vanwall    6
## 2 1959 Cooper-Climax    5
## 3 1960 Cooper-Climax    6
## 4 1961       Ferrari    5
## 5 1962           BRM    4
## 6 1963  Lotus-Climax    7
```

The table above gives the season year, the wining constructor, and the total number of race wins that constructor had. Lets now get a better understanding of which constructors have won the most seasons and their total race wins:

```
table(constructor_wins$name)
```

```
## 
##      Benetton Brabham-Repco         Brawn           BRM Cooper-Climax 
##             1             2             1             1             2 
##       Ferrari  Lotus-Climax    Lotus-Ford    Matra-Ford       McLaren 
##            16             2             1             1             8 
##      Mercedes      Red Bull       Renault    Team Lotus       Tyrrell 
##             3             4             2             4             1 
##       Vanwall      Williams 
##             1             9
```

From the above table we can see that Ferrari has the most seasons won between 1950 and 2017 (the span of the dataset). Thus this gives a clue that perhaps some constructors are more likely to win as compared to others. Let's examine closer.

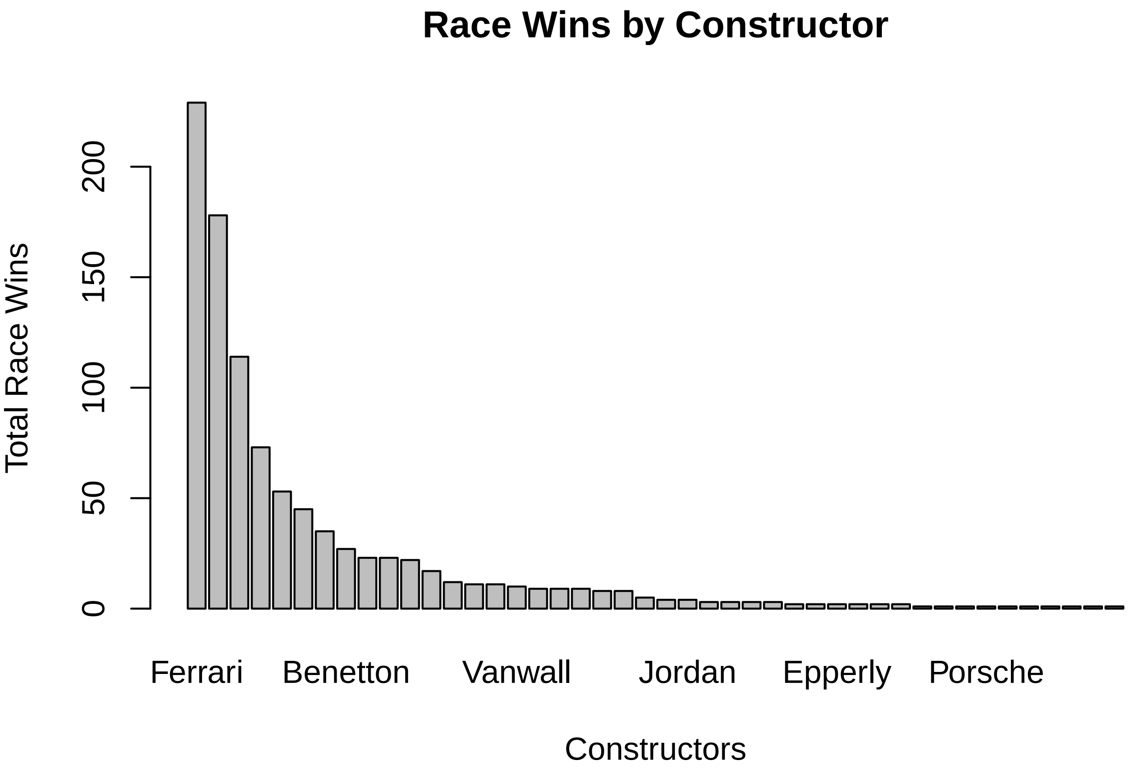
Which constructors have won the most races?

Now that we have a better understanding of the constructors wins seasonally we can break it down further to races, i.e. how many races were won by each constructor.

```
constructor_race_wins <- dbGetQuery(con, "
  SELECT constructors.name, COUNT(*) AS total_race_wins
  FROM results
  INNER JOIN constructors
  ON results.constructorId = constructors.constructorId
  WHERE results.positionOrder = 1
  GROUP BY constructors.name
  ORDER BY total_race_wins DESC;")
head(constructor_race_wins)
```

```
##          name total_race_wins
## 1     Ferrari             229
## 2     McLaren             178
## 3    Williams             114
## 4    Mercedes              73
## 5    Red Bull              53
## 6  Team Lotus              45
```

```
barplot(as.numeric(constructor_race_wins$total_race_wins), names.arg = constructor_race_wins$name,main =
```



As we hypothesized above the constructors that won the most seasons also won the most races thus perhaps constructor could be a good predictor of race winner as it seems some are more likely to win than others. Let's investigate whether the constructors are a good predictor of race winner across all seasons.

Are constructors a good predictor for winning a race?

The below query gives a dataframe with two columns one of the constructor for the driver for the race and a binary variable indicating whether or not that driver won the race. We can use this dataframe to determine whether or not constructor is a signifcant variable in predicting race wins.

```
race_results <- dbGetQuery(con, "
  SELECT constructors.constructorId,
  CASE
```

```
    WHEN results.positionOrder = 1 THEN 1
    ELSE 0
    END AS win
  FROM results
  INNER JOIN constructors
  ON results.constructorId = constructors.constructorId
")
```

```
summary(glm(race_results$win ~race_results$constructorId, family = "binomial"))
```

```
##
## Call:
## glm(formula = race_results$win ~ race_results$constructorId,
##     family = "binomial")
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.9904896  0.0414940 -72.070  < 2e-16 ***
## race_results$constructorId -0.0038395  0.0006925  -5.544 2.95e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8115.2  on 23656  degrees of freedom
## Residual deviance: 8080.5  on 23655  degrees of freedom
## AIC: 8084.5
##
## Number of Fisher Scoring iterations: 6
```

From the above result we can see that the constructorId is a significant predictor in determining whether or not the race was won. The p-value for constructorID is less than 0.05 and therefore significant in predicting win. Thus, we can say that one of the factors for determining who will win a race is their constructor.

Drivers

In Formula 1 history some drivers have far outperformed others. Lewis Hamilton has won 105 races, Michael Schumacher has won 91 and Max Verstappen has won 65 as of 2025 (3). Thus perhaps a good indicator of which driver will win is the driver themselves as some drivers perform consistency much better than others. Additionally how did individual drivers perform compared to others and what makes a good driver. Let's investigate.

Which drivers won each season?

Let's start by looking at the winning driver of each season and then compare that to the results of the previous section.

```
driver_wins <- dbGetQuery(con, "
  WITH last_race_of_season AS (
  SELECT year, MAX(raceId) AS last_race
  FROM races
  GROUP BY year)
  SELECT  races.year, constructors.name AS constructor, drivers.forename,
  drivers.surname
  FROM driverStandings
  INNER JOIN races
  ON driverStandings.raceId = races.raceId
  INNER JOIN last_race_of_season
  ON races.raceId = last_race_of_season.last_race
  INNER JOIN drivers
  ON drivers.driverId = driverStandings.driverId
  INNER JOIN results
  ON results.raceId = races.raceId AND results.driverId = drivers.driverId
  INNER JOIN constructors
  ON results.constructorId = constructors.constructorId
  WHERE driverStandings.position = 1
  ORDER BY races.year;
")
head(driver_wins)
```

```
##   year constructor forename surname
## 1 1950  Alfa Romeo     Nino  Farina
## 2 1951  Alfa Romeo     Juan  Fangio
## 3 1952     Ferrari  Alberto  Ascari
## 4 1953     Ferrari  Alberto  Ascari
## 5 1954    Mercedes     Juan  Fangio
## 6 1955    Mercedes     Juan  Fangio
```

Lets now extract a table of the winning drivers and how many seasons each driver won.

```
table(paste(driver_wins$forename, driver_wins$surname))
```

```
## 
##        Alain Prost         Alan Jones     Alberto Ascari       Ayrton Senna 
##                  4                  1                  2                  3 
##         Damon Hill        Denny Hulme Emerson Fittipaldi    Fernando Alonso 
##                  1                  1                  2                  2 
##        Graham Hill       Jack Brabham     Jackie Stewart Jacques Villeneuve 
##                  2                  3                  3                  1 
##         James Hunt      Jenson Button          Jim Clark     Jody Scheckter 
##                  1                  1                  2                  1 
##       John Surtees        Juan Fangio       Keke Rosberg     Kimi Räikkönen 
##                  1                  6                  1                  1 
##     Lewis Hamilton     Mario Andretti Michael Schumacher      Mika Häkkinen 
##                  3                  1                  7                  2 
##      Mike Hawthorn      Nelson Piquet       Nico Rosberg      Nigel Mansell 
##                  1                  3                  1                  1 
##         Niki Lauda        Nino Farina   Sebastian Vettel 
##                  2                  1                  4
```

As compared to the results from the constructors the max number of season wins is much lower and there is far less variability. It is important to note that there are far more total drivers than constructors as stated above there are 20 drivers per season and only 10 constructors meaning there is not one individual dominating such as with Ferrari. Now lets examine the constructors of the winning drivers and see if the results are the same as the winning constructors.

```
table(driver_wins$constructor)
```

```
## 
##    Alfa Romeo      Benetton       Brabham Brabham-Repco         Brawn 
##             2             2             2             2             1 
##           BRM Cooper-Climax       Ferrari  Lotus-Climax    Lotus-Ford
##             1             2            14             2             1
##      Maserati    Matra-Ford       McLaren      Mercedes      Red Bull
##             1             1            12             5             4
##       Renault    Team Lotus       Tyrrell      Williams
##             2             2             2             7
```

```r
table(constructor_wins$name)
```

```
##
##      Benetton Brabham-Repco         Brawn           BRM Cooper-Climax
##             1             2             1             1             2
##       Ferrari  Lotus-Climax    Lotus-Ford    Matra-Ford       McLaren
##            16             2             1             1             8
##      Mercedes      Red Bull       Renault    Team Lotus       Tyrrell
##             3             4             2             4             1
##       Vanwall      Williams
##             1             9
```

When comparing these two tables we can see that Ferrari has the most driver wins and constructor wins and both McLaren and Williams have a larger number of constructor and driver wins. However there are a few teams that appear in the driver wins table that do not appear in the constructor championship wins table such Alfa Romeo, Brabham, Maserati and Vanwall. These are teams where a driver has won a championship but the constructors did not. We can look further into individual driver race wins and see if there are any similar differences among them.

What proportion of races each driver entered did they win?

Now we can examine total race wins and race win ratio across drivers to see if certain drivers outperform others and the margins to begin to understand if driver can be used predict race outcomes.

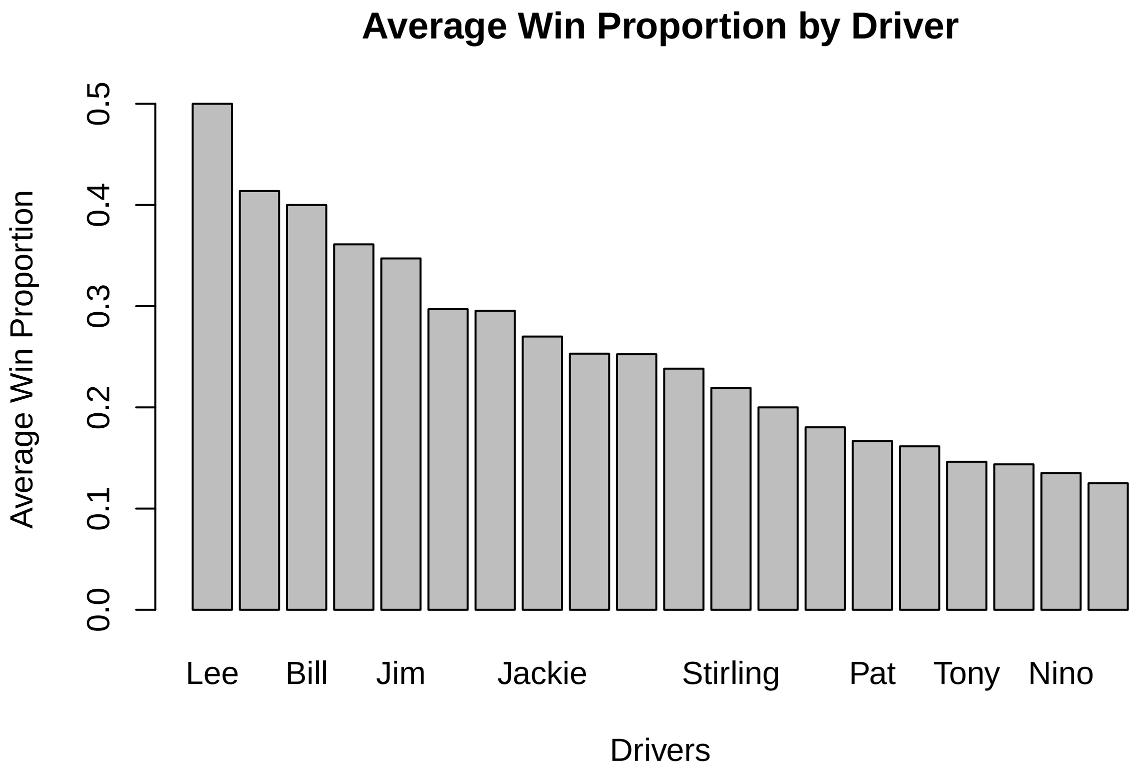
```r
driver_wins_ratio <- dbGetQuery(con, "
  WITH driver_results AS (SELECT results.resultId,
  drivers.driverId, drivers.forename, drivers.surname,
  CASE
  WHEN results.position = 1 THEN 1
  ELSE 0
  END AS wins
  FROM results
  JOIN drivers
  ON results.driverId = drivers.driverId),
  driver_stats AS (SELECT driverId, forename,surname,
  COUNT(*) OVER (PARTITION BY driverId) AS races_entered,
  SUM(wins) OVER (PARTITION BY driverId) AS wins
  FROM driver_results)
  SELECT DISTINCT driverId, forename, surname, races_entered, wins,
  wins /races_entered AS win_proportion
  FROM driver_stats
  ORDER BY win_proportion DESC;
")
head(driver_wins_ratio, 20)
```

```
##    driverId  forename    surname races_entered wins win_proportion
## 1       766       Lee    Wallard             2    1         0.5000
## 2       579      Juan     Fangio            58   24         0.4138
## 3       657      Bill   Vukovich             5    2         0.4000
## 4       647   Alberto     Ascari              36   13          0.3611
## 5       373       Jim      Clark              72   25          0.3472
## 6         1     Lewis   Hamilton             202   60          0.2970
## 7        30   Michael Schumacher             308   91          0.2955
## 8       328    Jackie    Stewart             100   27          0.2700
## 9       102    Ayrton      Senna             162   41          0.2531
## 10      117     Alain      Prost             202   51          0.2525
## 11       20 Sebastian     Vettel             193   46          0.2383
## 12      475  Stirling       Moss              73   16          0.2192
## 13      628       Bob   Sweikert               5    1          0.2000
## 14       71     Damon       Hill             122   22          0.1803
## 15      559       Pat   Flaherty               6    1          0.1667
## 16       95     Nigel    Mansell             192   31          0.1615
## 17      479      Tony     Brooks              41    6          0.1463
## 18      182      Niki      Lauda             174   25          0.1437
## 19      642      Nino     Farina              37    5          0.1351
## 20      786     Luigi    Fagioli               8    1          0.1250
```

The table above shows us the top 20 f1 drivers in terms of the proportion of the races that they have won. We can see that there are a few outliers that have entered relatively few races but had victories. I wanted to investigate this a bit further. The top driver in terms of win_proportion was Lee Wallard however he had only entered 2 races and won one of them so his win proportion was considered the highest at 0.5. Upon further research it seems that in the 1950's the Indianapolis 500 was a part of the Formula 1 world championships and therefore drivers that were not a part of the grid all year were able to get points and wins in F1 (4). Outliers such as this case could potentially have an impact on modeling results if for example win_proportion was used as a predictor of wins.

We can also examine a histogram of the top 30 drivers by race win ratio

```
barplot(as.numeric(driver_wins_ratio$win_proportion[1:20]), names.arg = driver_wins_ratio$forename[1:20
```



We can see that average wins steadily decrease with no specific chunk of drivers ourperforming others.

Which circuit does each driver perform the best on?

Beyond the win proportion, does the race track make a difference on race outcomes, i.e. do certain drivers perform better on certain tracks.

```
driver_best_circuit <- dbGetQuery(con, "
  WITH driver_circuit_avg AS (
  SELECT drivers.driverId, drivers.forename, drivers.surname,circuits.name,
  AVG(results.positionOrder) AS avg_position
  FROM results
  INNER JOIN drivers
  ON results.driverId = drivers.driverId
  INNER JOIN races
  ON results.raceId = races.raceId
  INNER JOIN circuits
  ON races.circuitId = circuits.circuitId
  WHERE results.positionOrder IS NOT NULL
  GROUP BY drivers.driverId, circuits.name),
  ranked_circuits AS (SELECT *, ROW_NUMBER()
  OVER (PARTITION BY driverId ORDER BY avg_position) AS rank
  FROM driver_circuit_avg)
  SELECT driverId, forename, surname, name, avg_position
  FROM ranked_circuits
  WHERE rank = 1
  ORDER BY avg_position;
")
head(driver_best_circuit, 10)
```

```
##    driverId  forename    surname                          name avg_position
## 1         1     Lewis   Hamilton   Indianapolis Motor Speedway            1
## 2         3      Nico    Rosberg              Baku City Circuit            1
## 3        20 Sebastian     Vettel   Buddh International Circuit            1
## 4        30   Michael Schumacher Okayama International Circuit            1
## 5        35   Jacques Villeneuve           Autódromo do Estoril            1
## 6       102    Ayrton      Senna                Donington Park            1
## 7       177      Keke    Rosberg                     Fair Park            1
## 8       178      Alan      Jones      Las Vegas Street Circuit            1
## 9       200    Jochen       Mass                      Montjuïc            1
## 10      224   Emerson Fittipaldi              Nivelles-Baulers            1
```

```
frequencies <- data.frame(table(driver_best_circuit$name))
frequencies[frequencies[,2] > 20,]
```

```
##                             Var1 Freq
## 11 Autódromo Juan y Oscar Gálvez   34
## 12  Autodromo Nazionale di Monza   58
## 21             Circuit de Monaco   35
## 24  Circuit de Spa-Francorchamps   32
## 28         Circuit Park Zandvoort   23
## 38   Indianapolis Motor Speedway  114
## 51                   Nürburgring   57
## 63           Silverstone Circuit   45
## 67                  Watkins Glen   23
```

There are a few circuits that have been general performance across drivers. Such that more than 20 drivers perform the very best at those circuits. Indianapolis Motor Speedway is the top performing track across drivers however this may also be due to the discrepancy discovered in the previous section where the Indy 500 races used to be a part of F1 in the 50s and there were many drivers who only drove a few races for F1 specifically only at that location.

```
mean(driver_best_circuit$avg_position)
```

```
## [1] 12.4014
```

The overall average position each drivers best circuit is 12.4014. This means there are likely many drivers with relatively low average positions even on their best tracks.

Is driver a signifcant predictor for race win?

So far we've discovered driver wins by season, proportion of races won for each driver and best circuit for each driver. We've seen that amongst drivers there is a lot of variability and now we can ask the question of if we can use that variability to predict the outcomes of races.

```
race_results_drivers <- dbGetQuery(con, "
  SELECT drivers.driverID,
  CASE
    WHEN results.positionOrder = 1 THEN 1
    ELSE 0
    END AS win
  FROM results
  INNER JOIN drivers
  ON results.driverID = drivers.driverID
")
```

```
summary(glm(race_results_drivers$win ~ race_results_drivers$driverID, family = "binomial"))
```

```
##
## Call:
## glm(formula = race_results_drivers$win ~ race_results_drivers$driverID,
##     family = "binomial")
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.872357   0.044535 -64.497  < 2e-16 ***
## race_results_drivers$driverID -0.001422   0.000179  -7.946 1.93e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8115.2  on 23656  degrees of freedom
## Residual deviance: 8041.0  on 23655  degrees of freedom
## AIC: 8045
##
## Number of Fisher Scoring iterations: 6
```

From the logistic regression model above we can see that the driver is a significant predictor for predicting race outcome as the p-value of 1.93e-15 is less than 0.05. This means that some of the variability we observed in the previous EDA questions can explain race outcome. Thus some drivers win significantly more frequently than others. Further analysis for this section could mean fitting a model with both constructor and driver Id values in order to see if a model could more accurately predict using both variables.

Basic Patterns Among Race Winners

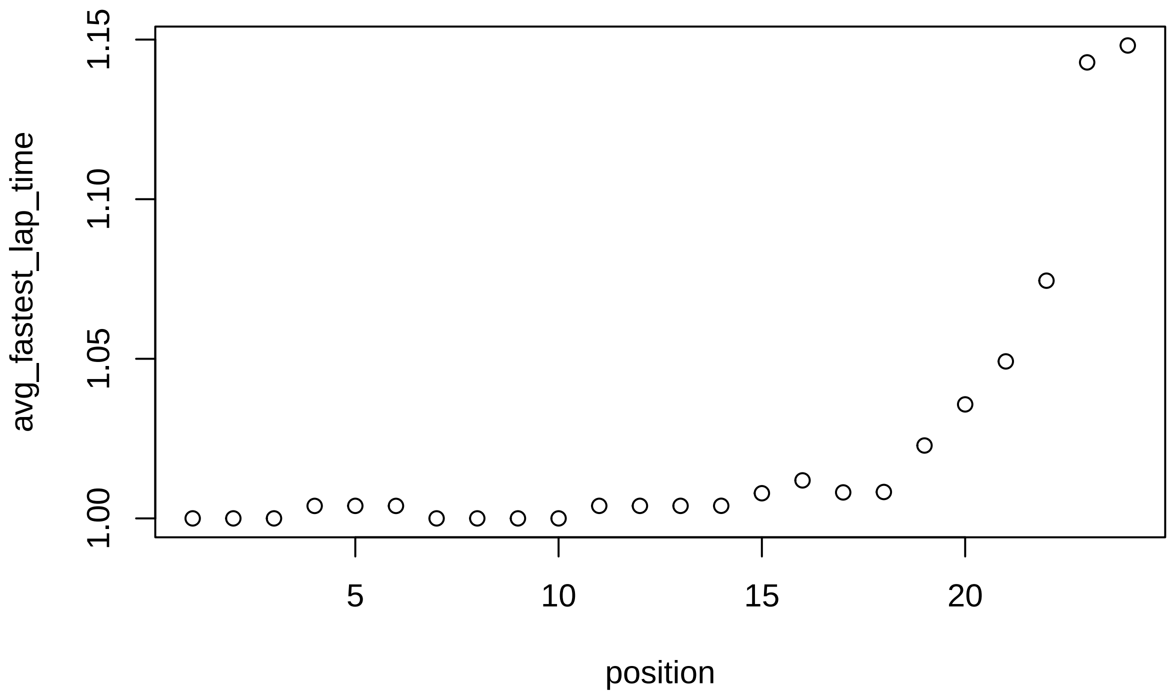
Now that we've discovered that a race's outcome can be predicted by both constructor and driver let's look a little further into what actually makes a race winner by looking at other variables too see what effect those have.

What is the average fastest lap time of each final position (1,2,3 etc.)?

In Formula 1 the ending position is crucial to understanding a race outcome i.e. which driver placed in each position. Fastest lap time is a variable given for each driver for each race along with the final position they had in the race. Does the fastest lap time increase for drivers finishing in lower places?

```
fastest_laps <- dbGetQuery(con, "SELECT positionOrder AS position, AVG(fastestLapTime) AS avg_fastest_la
FROM results
WHERE fastestLapTime IS NOT NULL AND positionOrder IS NOT NULL
GROUP BY positionOrder
ORDER BY positionOrder;")
```

```
plot(fastest_laps)
```



As to be expected as the position increases so does the average fastest lap time. There is a significant increase for positions beyond 20 as in this data set these positionOrder assigns a number to each driver regardless of if they finished the race or not so the drivers in those final places likely would have very fast lap times if they did not complete the whole race. The fastest average lap does remain relatively steady until around position 15. This is because the drivers are doing many laps so even if they are able to get a relatively quick lap on one of the laps there is still much room for mistakes. It also seems that a lap on average does not go much quicker than around 1 min even for top finishing positions.

What is the percentage of race winners qualifying 1st?

F1 has a qualifying day before the actual Grand Prix which is used to determine race starting order (1). This starting order can change the outcomes of races and thus begs the question are most race winners starting in first at the begining of the race?

```
pole_postion_winners <- dbGetQuery(con, "WITH race_winners AS (
SELECT raceId, driverId, grid, COUNT(*) OVER (PARTITION BY positionOrder) AS total_winners
FROM results
WHERE positionOrder = 1),
pole_to_win AS (
SELECT *, COUNT(*) AS pole_towin_count
FROM race_winners
WHERE grid = 1)
SELECT 100.0*pole_towin_count/total_winners AS percent
FROM pole_to_win;")
```

```
pole_postion_winners
```

```
##    percent
## 1 41.52107
```

Not a majority but 41.5% of drivers who started a race in first place ended it that way. This means that qualifying likely plays a large role into the outcomes of races and those drivers that qualify high consistently likely have a greater chance of winning races.

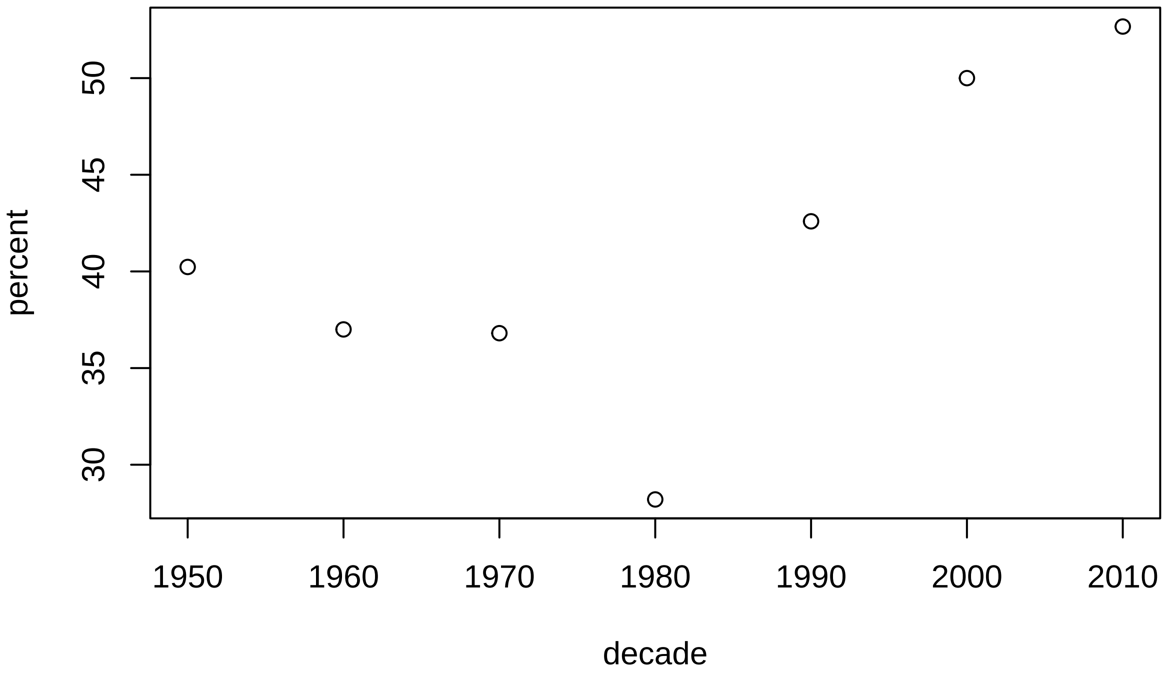
How has the proportion of wins from pole position changed over time?

Now let's examine how this breaks down across decades. More specifically, were you more likely to win from pole position in the 1950's than you are now? Let's get this percentage across decades.

```
decades <- dbGetQuery(con, "
WITH race_winners AS ( SELECT results.raceId, results.driverId, results.grid,
FLOOR(races.year /10)*10 AS decade
FROM results
JOIN races
ON results.raceId = races.raceId
WHERE results.positionOrder = 1),
winners_per_decade AS (SELECT decade, COUNT(*) AS total_winners
FROM race_winners
GROUP BY decade),
pole_winners_per_decade AS (SELECT decade, COUNT(*) AS pole_winners
FROM race_winners
WHERE grid = 1
GROUP BY decade),

joined AS (SELECT winners_per_decade.decade, winners_per_decade.total_winners,
pole_winners_per_decade.pole_winners
FROM winners_per_decade
LEFT JOIN pole_winners_per_decade
ON winners_per_decade.decade = pole_winners_per_decade.decade
)
SELECT decade,100.0 * pole_winners/total_winners AS percent
FROM joined
ORDER BY decade;
")
```

```
plot(decades)
```



It seems that in recent decades the percent of winners starting at pole position has increased from previous decases will decades from 1990 all having percent winners starting on pole above 40%.

Conclusion

In conclusion there are many factors that contribute to race outcomes. I started by examining constructors and seeing what roles they play race outcomes. From that analysis it was shown that Ferrari and Williams and McLaren were the top performing teams seasonal and in total races won for that team. There was much variation across different constructors with some having won only one championship and others winning many. When used as a predictive variable for race outcome, constructor was statistically significant and therefore in a larger model would likely be a good feature to include. Next I examined drivers themselves. It was shown through this EDA that the is a lot of variation amongst winning drivers as well with some winning many and other with very few race/championship wins. Through this analysis outliers were also discovered as F1 has changed which races it includes in it's championships throughout the years and there is a large variation amongst how many races each driver has entered. The average position on each drivers best course was approximately 12. Driver was then used as a variable to predict race outsomes and it was discovered to be statistically significant as well. Finally some trends among race winners were examined and from these a few conclusions were drawn. Namely that the percentage of race winners starting in pole position is approximately 41% overall and that percentage has increased in recent decades. There are many more variables that could be explored and other questions to be answered. If I were to do this project again I would like to look further at other factors leading to race wins and perhaps try to fit an accurate model at predicting this variable.

Sources

1. https://f1chronicle.com/a-beginners-guide-to-formula-1/
2. https://www.formula1.com/en/latest/article/the-beginners-guide-to-the-f1-constructors-championship.66nTfWSqrUYv3bnbosPkHV

3. https://www.autosport.com/f1/news/whos-the-best-formula-1-driver-schumacher-hamilton-senna-more-4983210/4983210/

4. http://en.espn.co.uk/f1/motorsport/story/12047.html