**Introduction**

**What is Formula One?**

Formula One (F1) is the highest class auto racing sanctioned by Federation Internationale de l'Automobile (FIA). Since 1950, F1 has attracted hundreds of millions of viewers globally, both on TV and at the race tracks. Fans are drawn to F1 for the roaring engines, rivalry competitions, welcoming atmosphere, and most importantly, the speed. F1 cars can accelerate from 0 to 160 kph (100 MPH) and back down to 0 in just 4 seconds, and they can reach top speed of over 370 kph (230 MPH). Each year, between 10 to 12 teams with 2 drivers each, travel around the world to compete in the prestige sport. There are typically around 20 races each year.

The two most important events during a race weekend are Saturday’s Qualifying Session and Sunday’s Race Session. During the Qualifying Session, drivers compete for Grid position (starting position) for Sunday’s Race. During the Race Session, drivers compete to finish first and the top 10 drivers and their teams receive F1 World Championship points. These points are used to determine the winner at the end of each race year.

**Why are we interested?**

F1 is one of the most watched sports in the world along with soccer and basketball. Each race event attracts over 100,000 people to the race track and millions watch it on TV globally. As a result, the F1 paddock has billions of dollars at stake from team revenues to commercial interests.

We wanted to predict what factors affected a given driver’s performance during a Race. We chose driver’s average speed (total distance travelled / total time) as our performance indicator because it is easily understandable and comparable. We chose not to use driver’s final rank as a factor because it cannot accurately measure performance as compared to another driver. For example, the difference between the 1st and the 2nd driver can be 10 seconds or a minute and their ranks will remain the same. But, average speed can tell us that if the driver finished 1 minute before, he did much better.

**Analysis Setup**

**Factors that might affect a Race**

For understanding the performance of a F1 driver in a race, we took into consideration multiple factors that might influence his performance. We broadly classified these factors into three categories:

**1)** **Driver Attributes**

Driver attributes are the factors that are directly related to the F1 driver himself. These can be further classified into two sub-categories:

**a)** **Driver Demographics**

Demographics, in general, refers to the properties of a population. F1 revolves around the drivers and hence, it is important to get as much information about them, as possible.

**i)** **Driver’s Physique**

The driver’s physique might play an important role in the race. The weight of the driver and his height might affect the performance of his car and hence the maximum speed he can achieve during the race.

**ii)** **Driver’s Age**

Usually, in sports, we see that players tend to improve as they gain experience over years. On the other hand, as they grow older could cause decreased fitness, increased mental pressure due to expectations, slower reflexes etc causing a decrease in their performance. Thus, driver’s age could be an important factor that determines the outcome.

**iii)** **Driver’s Nationality**

The driver’s nationality can affect his upbringing and beliefs, his health, his genes, his commitment to the game etc. For example, we see that, in some nations, like United States, Australia, sports are considered an integral part of the society and are promoted more actively. Thus, nationality might play an important role in determining how well a driver performs.

**iv)** **Number of Dependents of the Driver**

F1 is a very high risk sport. Historically there have been many accidents during the races, some even fatal. Hence, the willingness of a driver to commit to a race, take risks might be critical to his performance. A driver with a family or more number of dependents might be more reluctant to take risks as compared to someone who has no dependents.

**b)** **Driver Performance**

How well the driver has been performing in the recent past could indicate to how he might perform in the upcoming race. We can categorize performance into two factors:

**i)** **Recent Performance**

The driver’s recent performances indicate his form and could be a measure of his confidence. So, depending on how confident a driver is or how good he is feeling about himself, his performance in an upcoming race might improve or plummet. To take this factor into consideration we might look at how well the driver performed in the qualifier.

**ii)** **Driver Ranking**

A driver’s rankings are usually calculated based on how he has been performing over the past few months. A driver with a good ranking might indicate that he has been performing consistently well over the past few months. This might be a better indicator of driver quality than his recent performances. Hence, it is possible that although a good player has not been in great form recently, he might perform well in a race because he is a good player in general.

**2)** **Track Attributes**

A F1 race happens on a racing track and the attributes related to the track might play an important role in the outcome of the race.

**a)** **Track Location**

Players across sports tend to prefer playing in some locations over others because those locations suit their style of play. It is also possible that F1 drivers might prefer racing on certain tracks depending on the weather conditions of the location. For example, the temperature, the air pressure, the wind speed, the humidity levels etc. of a location might suit certain players’ driving style over others’.

**b)** **Track Difficulty**

We have also seen that some sportspersons tend to perform much better when they are put under difficult situations. They have a higher threshold to absorb pressure. If we apply the same analogy to F1, some players might be able to perform much better on more difficult tracks as compared to others. The major factor which might determine the difficulty of a track is the number of turns the track has.

**3)** **Team Attributes**

Team attributes refer to the properties of the team to which the driver belongs

**a)** **Team Car**

In F1, the team to which a player belongs is usually also the manufacturer of its team cars. Even if they do not manufacture their own cars, all their cars are manufactured by the same maker. Assuming that the team builds similar performing cars for both of its team drivers, we can say that a driver’s performance might also depend on which team he belongs to because a driver’s car is crucial to his speed.

**b)** **Team Score**

The team score indicates how well the team has been performing over the past few months. The team performance could be an indicator of how good the team’s car maker is and might reflect on the performance of the driver.

**Analysis Results** (Appendix C.2-B)

Through our analysis we were able to narrow down the factors that affect a driver’s performance in a F1 race by eliminating the ones that do not. Surprisingly, driver attributes such as their physique (height and weight) or their age and family size do not affect their performance in a race or as we referred to it as “their willingness to take risk”. Driver’s nationality and rank before the race, do not contribute to their speed in an upcoming race either.

We considered willingness to take risk as a potential factor affecting the performance and speed, since F1 is a very high risk sport. But family size, which we chose as an indicator for this factor might not be of much significance anymore. This is because car manufacturers and F1 organizers consider safety of drivers as their highest priority. In fact, the number of fatalities during a F1 race have gone down drastically in the last two decades.

We were also able to conclude that a car’s make has no effect on a driver’s performance during the race. Probably because a race is more about the current technology in the car which keeps on changing too frequently to make a lasting impact on driver performance.

Through our analysis we found evidence that the following factors affect the speed of a driver in an upcoming race:

1. Driver’s qualifying position:
   1. Higher grid position implies better performance and there is evidence that the better a driver performs in his qualifying session (i.e. the higher his qualifying position) better are his chances in the main race
   2. On an average, a driver’s average speed is 0.49 KPH (0.30 MPH) more than a driver who finishes one position below him in the qualifying session. With every increase in grid position, a driver’s average speed for the race decreases by 0.49 kilometer per hour (0.30 MPH) given all the other affecting factors remain constant
   3. A driver’s grid position is his rank in the qualifying session that happens the day before the actual race and is a deciding factor in the grid that he’ll be assigned in the race. This metric is a good representation of the driver’s current form and seems to have a significant impact on his performance in the actual race
2. Team’s previous year’s performance:
   1. Team’s previous year’s ending score has a positive impact on a driver’s speed, that is, a driver belonging to a team with a higher score at previous year’s end has a higher chance of performing better than a similar driver belonging to a team with a lower score
   2. We found evidence that with every unit increase in a team’s score, the race speed of a driver goes up by approximately 0.0019 kilometer per hour (0.0012 MPH) given other factors remain constant
   3. Though this looks like a small increment, we need to consider that the increase stated above is for every additional point that a team had earned in the previous year when compared to another team
3. Track location:
   1. We found evidence that different tracks affect drivers differently during a race, that is, all tracks cannot be considered identical when predicting a driver’s speed
   2. This might be the case because the quality of certain race tracks must be different from others. Also, each location has a different climate when compared to others
   3. Track locations might also affect a driver’s performance during a race due to the inherent human nature of personal preferences and comfort
   4. Our analysis also showed us that given all other affecting factors are kept constant, a driver would be fastest on a North American track and slowest on a South American track.

Below is a detailed comparison between the tracks in our data as compared to South American track(s):

i. A driver’s average speed is 36.45 KPH (22.65 MPH) more on a North American track

ii. An Asian race track’s average speed is 33.47 KPH (20.8 MPH) higher

iii. An European race track’s average speed is 28.29 KPH (17.58 MPH) higher

iv. An Australian race track’s average speed is 25.86 KPH (16.07 MPH) higher

4. Track difficulty:

1. A driver’s average speed during a race is likely to go down as the difficulty level for a race increases
2. With every additional turn in a race, a driver’s average speed reduces by 0.08 KPH (0.05 MPH)

**Analysis Limitations**

1. We cannot predict the speed for drivers (existing or new) in a new team
2. Data used for this analysis had information only for races in Asia, Australia, Europe, North America and South America. We cannot predict for any races in Africa and Antarctica
3. Our analysis is limited to current design of tracks where F1 has happened in the last three years from 2014 through 2016. If F1 adds an entirely new track with extreme diversions from current tracks in terms of design, quality, number of turns or hurdles our analysis may not be able predict driver performance for such a track

**Conclusion**

From our analyses above, we conclude that brands can now make data-driven business investment decision based on our model, especially in the three focus areas below.

1. Driver and team sponsoring

Businesses can now use our model to determine which driver is most likely to have the highest average speed during a race (most likely to win a race), then they can decide to sponsor those drivers that fit within the their budget. This is valuable to the advertisers because the fastest drivers are more likely to be featured in F1’s official broadcast, website, blogs, social media and other media outlets. What’s more, Brands that wants to invest in F1 generally would want to associate themselves with high speed and winning. Investing in the fastest driver and team as determined by our model will further enhance their message to end consumers.

1. Track side and TV advertising

Our model also includes location as a factor, which will help advertisers determine which track they should invest money in. In general, F1 tracks with the higher speeds also attract more viewers both at the track and on TV. If an advertiser had limited budget, its budget will be best spent at tracks with higher average speed according to our model.

1. Betting

Our model will also help enthusiastic F1 fans in Europe, where betting is generally legal. Most of the betting services are established in small European legal shelters and tax havens like Gibraltar and Cyprus, and they target customers across the world. F1 betting companies typically host bets on winner of a race and winner of a season. Our model considers various factors to predict the race performance of a given driver. Instead of placing a bet on someone based on intuition, participants can now make data driven bets with much more confidence.

To sum up, one can cautiously use our model to predict performance of a given F1 driver in a race.

**Appendix**

**(A) Data File:**

**(A.1) Data extraction from various sources:**

Data was extracted from various sources as listed below:

Dependent Variable:

Average Speed = Length of Track x Number of Laps/Total Time to Complete the Race: <http://ergast.com/mrd/db/>

Independent Variables:

1. Driver attributes:
   1. Age[2] : <https://en.wikipedia.org>
   2. Nationality[2]: <http://www.statsf1.com/en/default.aspx>
   3. Dependents[3]: <http://www.f1fanatic.co.uk/2016-f1-season/statistics/race-data/>
   4. Weight[3]: <https://www.auto123.com/en/racing-news/f1-height-and-weight-of-the-current-formula-1-drivers?artid=176733>
   5. Height[3]: <https://www.auto123.com/en/racing-news/f1-height-and-weight-of-the-current-formula-1-drivers?artid=176733>
   6. Driver’s rank before the race[2]: <http://www.statsf1.com/en/default.aspx>
2. Car attributes:
   1. Manufacturer used[4]: <http://www.statsf1.com/en/default.aspx>
   2. Constructors score : <https://en.wikipedia.org>
3. Race attributes:
   1. Track (distance, turns, laps, track country)[1]: <http://www.f1fanatic.co.uk/2016-f1-season/statistics/race-data/>
   2. Qualifier position for each driver in the race[4]: <http://www.bbc.com/sport/formula1/2016/results>

**(A.2) Data collation:**

Data sources as listed above in section A.1 was collected and collated in a file on excel. We collected the data for three years: 2014, 2015 and 2016. Few variables that were included in the data were derived and transformed as listed below.

* Teams Previous Year Score: We considered Team Score for a team corresponding to the previous year and not the year in which the race occurred. Using current year team score would have led to ‘data leakage’
* Age: The age of every driver is calculated as of end of the previous year
* Average Speed: We divided the total distance of a track with the time taken by the driver to finish the race to calculate the column Speed
* Total Turns : This variable is calculated by multiplying the turns per lap with total number of laps in every race

All the variables from A.1 and A.2 were collated into a final master data file.

**(A.3) Document** #placeholder to add the excel data file object that was used for the modeling exercise

**(B) Descriptive Statistics**

**(B.1) Data Investigation & Cleaning**

* Outliers: There were few records where driver had extremely high speed because these drivers had very less time recorded since they had not competed the race. So we did an outlier treatment by dropping records where a driver had not finished all the laps in any given race
* We have taken a driver’s age for a race as of the first day of the year of the race’s
* Missing data:
  + - Track temperature is very important component for F1 car’s speed. We obtained this data for 2014 but we could not find data for year 2015 and 2016
    - Missing Count of dependents were replaced with zeros
    - “Team’s Previous Year Score” variable had missing values since there were new teams with no previous score being added every year. These missing values were replaced with zeroes.

**(B.2)** Variable Distribution:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Weight | Height | Age | Speed | TotalTurns | TeamPrevYrScore | Grid |
| Min | 55 | 166 | 17 | 101 | 568 | 0 | 0 |
| 1st Qu | 64 | 173 | 24 | 170 | 864 | 70 | 3 |
| Median | 66 | 175 | 27 | 184 | 973 | 84 | 6 |
| Mean | 66 | 176 | 27 | 182 | 988 | 82.6 | 7.626 |
| 3rd Qu | 68 | 180 | 30 | 199 | 1120 | 98 | 11 |
| Max | 78 | 186 | 36 | 238 | 1482 | 161 | 22 |

Here the range of driver attributes is less than that of Speed, Total turns and Teams Previous Year Score attributes. Weight, Height, Age and Speed are normally distributed with mean and median very close to each other, But Total Turns and Grid are right skewed since these have median lesser than the mean. Teams Previous Year Score is left skewed.

**(C) Inferential Statistics**

**(C.1)Relationships of interest:**

Correlation matrix was constructed to understand the relationship between the variables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Height | Age | Speed | TotalTurns | TeamPrevYrScore | Grid |
| Height | 1.00 | -0.31 | -0.08 | 0.02 | -0.58 | 0.06 |
| Age | -0.31 | 1.00 | 0.10 | -0.06 | 0.12 | -0.17 |
| Speed | -0.08 | 0.10 | 1.00 | -0.66 | 0.07 | -0.16 |
| TotalTurns | 0.02 | -0.06 | -0.66 | 1.00 | 0.01 | 0.06 |
| TeamPrevYrScore | -0.58 | 0.12 | 0.07 | 0.01 | 1.00 | -0.04 |
| Grid | 0.06 | -0.17 | -0.16 | 0.06 | -0.04 | 1.00 |

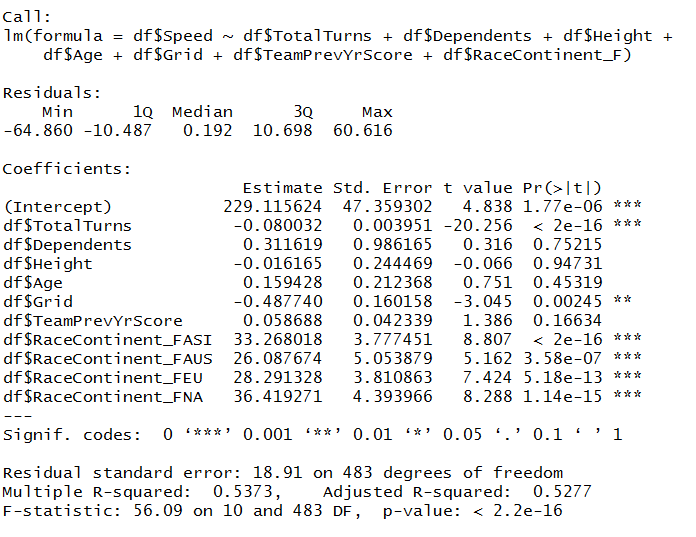
There is no high collinearity issue between our predictor variables since none of the correlation is greater than 0.7

**(C.2)Modelling:**

We started with modeling exercise after data extraction, data cleaning, data collation, missing value, outlier treatment, and data distribution understanding. Race location had many levels so we bucketed countries into continents, so as to make it more interpretable after factorizing the variable and using it in the model

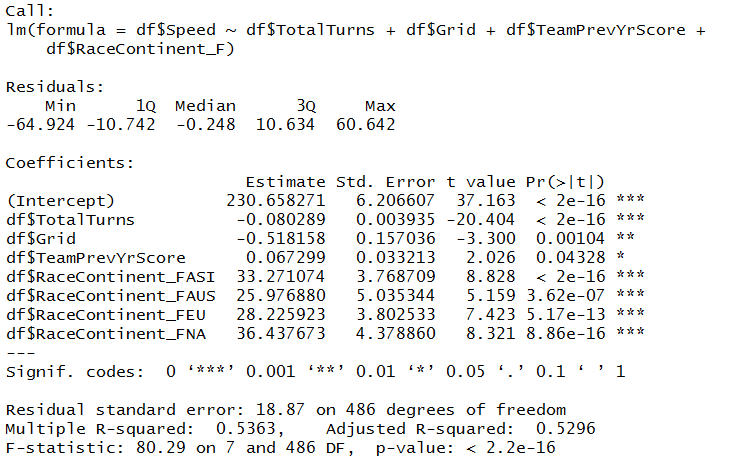
1. Output of an initial iteration of modeling exercise:

This is an output from one of our initial iterations and some potential predictors that might not affect driver’s performance in a race significantly in addition to those that do



1. Output of Final Iteration of modeling exercise:

This section has the R-output for the final model of our analysis. All data in results result to this section.



* **Summary of all the iterations**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Change Description | Iteration | Residual Standard Error | R Square | Highest P-value among predictors | Model P-Value |
| Considering all variables | 1 | 18.91 | 0.537 | 0.95 | 2.2e-16 |
| Height removed | 2 | 18.89 | 0.537 | 0.75 | 2.2e-16 |
| Dependents removed | 3 | 18.87 | 0.537 | 0.31 | 2.2e-16 |
| Age Removed | 4 | 18.87 | 0.537 | 0.04 | 2.2e-16 |

Our final multiple regression model has 3 interval variables - Total turns, Grid and Team Previous Year Score- and one nominal variable - Race Continent - that significantly explains our dependent variable speed

**(C.3) Assumption Checks:**

We made the following assumptions while creating the regression model:

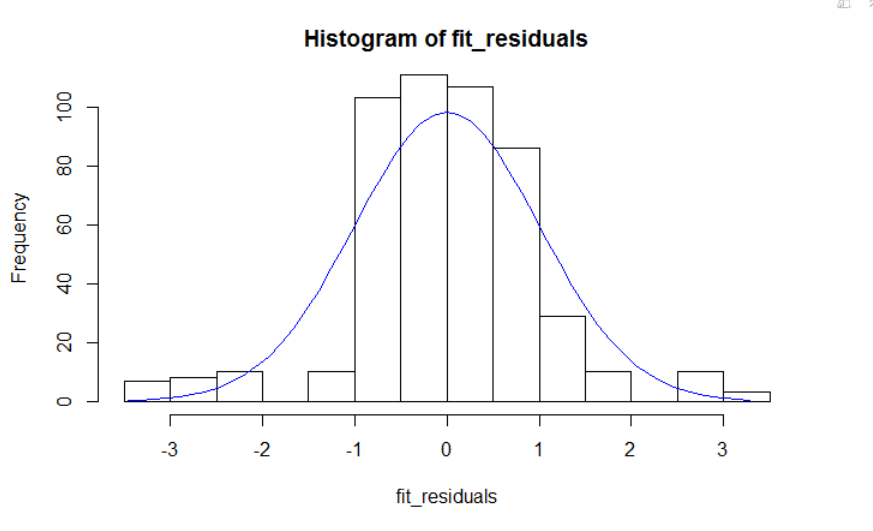
**Random Sampling:**

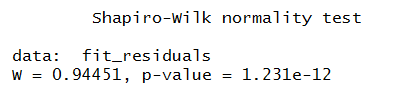
We chose all the data for the past three years since we wanted to take into account recent performances in our model. We also restricted the scope of our model to certain tracks. However, we did not pick up a sample out of the data that we had selected, we took the whole dataset.

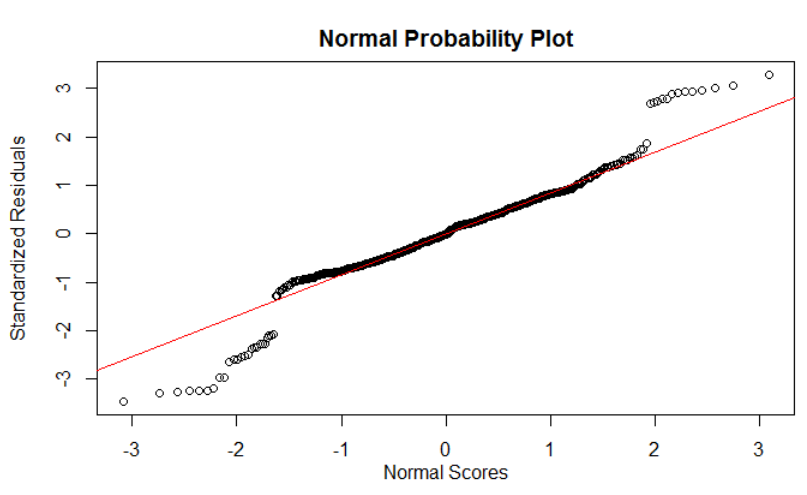
**Normality**

Normality tests on standard residuals









Shapiro-Wilk test resulted in a very small p-value. Thus, in-spite of having a good R-Square of 0.5363 and Residual standard error of 18.87, the residuals are not normally distributed. The histogram, the Q-Q plot and the standardized residual plot reinforce this findings.

**Mean of residuals is 0 and SD is fixed across predictors**

The standardized residual plot given above is not symmetric about the 0 line. This indicates that the mean of residuals is not 0. It also suggests that the SD of residuals is not fixed across predictors and there is a possible heteroscedasticity.

**(D)Scope of Inference:**

* Result of our model can be inferred only if the following conditions are met
  + The race is part of the following list of continents: Europe, South America, Australia, Asia and North America
  + Rules of the game do not change
* The model can be used for directional inference but we should avoid/be cautious while making predictions from this model because the model fails to meet the assumptions that are require validation for a linear regression model (Appendix C.3). The p-value of Shapiro-Wilk test, Q-Q plot and Normality Plot shows that assumption of residuals should be normal is not being met by the model

**References and data sources**

1. About Formula 1. (2015, September 22). Retrieved from <http://www.espn.co.uk/f1/story/_/id/13715980/about-formula-1>

##### Database Images<http://ergast.com/mrd/db/>

1. Stats F1. (n.d.). Retrieved from <http://www.statsf1.com/en/pilotes.aspx>
2. F1 race data. (n.d.). Retrieved August 3, 2017, from <http://www.f1fanatic.co.uk/2016-f1-season/statistics/race-data/>
3. F1 race data. (n.d.). Retrieved August 3, 2017, from <http://www.f1fanatic.co.uk/2016-f1-season/statistics/race-data/>
4. 2014 Formula 1 Results. (n.d.). Retrieved from [http://www.bbc.com/sport/formula1/2014/results](http://www.bbc.com/sport/formula1/2016/results)
5. 2015 Formula 1 Results. (n.d.). Retrieved from [http://www.bbc.com/sport/formula1/2015/results](http://www.bbc.com/sport/formula1/2016/results)
6. 2016 Formula 1 Results. (n.d.). Retrieved from <http://www.bbc.com/sport/formula1/2016/results>

Data sources