**Analysis of Formula 1 Racing**

**2014 to Present**

Alison Sturge

May 9, 2022

ABSTRACT

Formula 1 racing is an international motorsport that combines the talents of car builders and their drivers. It began in 1950 and is the most prestigious motor racing competition in the world. We know the big constructor names like Ferrari, Alfa Romeo, Mercedes, BMW and McLaren, but there are over 200 companies in competition to be called the best and fastest. These teams race in as few as 15 and as many as this year’s record of 23, with a minimum distance of 305km in each race. The number of laps depends on the length of those laps, and drivers must complete as many as it takes to reach that distance.

THE DATA

The data was pulled from Kaggle https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020), but was originally compiled from <http://ergast.com/mrd/>.

Table

Description automatically generated

I will be evaluating the data with races from 2014 to present. 2014 is considered the beginning of the current technology in racing engines, therefore is a relatively homogenous dataset. I combined several csv files and converted milliseconds to minutes for more comprehensible numbers.

FIRST LOOK AT THE DATA

Finishing Time

Chart, histogram

Description automatically generated Chart

Description automatically generated

Our first graph shows a distribution skewed slightly to the right. Looking at a boxplot of the same data, this time by race, tells us we may have had an incident in the British Grand Prix that caused a number of drivers to finish much more slowly than the rest, most likely a crash. Without these outliers, we would have a more normal distribution of the data. Fans would likely argue that these outliers are what make the sport so exciting.

Pitstop Time

Chart, histogram

Description automatically generatedOur next graph shows us the total time each driver spends on pitstops in each race. Since pitstops are primarily for taking on fuel and new tires, it should come as no surprise that each team uses consumables at a similar rate and we have a lot of density in the low end of the distribution. Teams with more pitstop time will have had mechanical issues that cost them vital seconds in the race.

Qualifying Time

Chart, histogram

Description automatically generatedOur final graph, which shows the average of 3 qualifying lap times, shows most of the density being evenly distributed between 1.1 and 1.8 minute laps. This suggests that most of the racers exert a similar amount of effort to qualify for a grid position, with very few using extra effort for a better time, and another small portion that barely qualifies.

CONTINUOUS TO CATEGORICAL VARIABLE

Table

Description automatically generatedTable

Description automatically generated

Our first look at the final standing variable gives us a mean. Given that the nature of the data in this column is a rank, the mean and standard deviation are effectively meaningless. When looking at the results from PROC FREQ, we see that the Top Ten finishes account for more than 75% of the results. With 20 drivers in each race, it is surprising to see such a small number finishing in the bottom ten. This points to a large number of DNFs (did not finish).

Table

Description automatically generated with medium confidence

A quick look at PROC FREQ for all final standings shows us that in fact we occasionally have more than 20 drivers in a race, with the number of finishers at the bottom of the table get smaller.

PRELIMINARY SCATTERPLOT

Chart, scatter chart

Description automatically generated

A scatter plot of total pitstop time and finishing time suggests there could be a weak correlation between the two. We also see a concentration of pitstop times just under 30 seconds in this plot.

SUMMARY STATISTICS

Table

Description automatically generatedMeasures of Center and Spread

Finishing time has a wide range from maximum to minimum. The mean is much closer to the minimum, which reflects the concentration we saw earlier in the lower end of the histogram. Standard deviation of 10.48 is as expected with a large variance in times.

With pitstop times, we also see a reflection of the earlier histogram, with the mean weighted toward the lower end of the distribution. Standard deviation is also large relative to the given times.

In qualification times, we find the mean much closer to the mode of the distribution. We also have a moderate to small standard deviation.

LINEAR REGRESSION

Table

Description automatically generated

The equation for the Least Squares Line for Finishing Time as a function of Pitstop Time

ŷ= 87.297 + 10.287x

A t Value of 15.15 with a p-value that is 0 at 4 decimal places indicate that we can state that there is enough evidence to support the hypothesis that finishing time is dependent on pitstop time.

We have a decreasing variance in the residuals, but it is centered at zero, so with further work and possibly transformation, we could arrive at a model that fits the data .

Chart, scatter chart

Description automatically generated

Chart, histogram

Description automatically generated

Barring the handful of outliers from the British Grand Prix, the residuals on the QQ plot follow a more or less normal distribution.

Chart, scatter chart

Description automatically generated

A fit plot for finishing time finds that most observations fall within two standard deviations of the regression line. However the model only accounts for 14% of the variance in the data.

CONCLUSION

From the data, we have found a likely correlation between pitstop time and finishing time, as well as a smaller but not insignificant correlation between qualifying time and finishing time. Further investigation of the very large original dataset would likely find even more correlations that did not fit withing the scope of this project.

REFERENCES

One Stop Racing:

<https://onestopracing.com/how-long-is-an-f1-race/>

TutorialsPoint:

<https://www.tutorialspoint.com/formula_one/formula_one_quick_guide.html>

Kaggle:

<https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020>)

Ergast Developer API:

<http://ergast.com/mrd/>