**Machine vs Man: An analysis of Formula One Cars and Drivers**

**Task 3**

**A1. Research Question Summary**

Research Question: *“Is the performance of a Formula One car, concerning time behind the leader, impacted more by the driver's skill or the constructor’s design?”*

The purpose of this study is to compare two specific factors, the driver and the car, and determine which has a greater impact on the success of a Formula One team. The driver utilizes technique, reaction time, and skill to extract maximum performance from the car. The constructor designs a car that is reliable, fast, and agile. Both of these factors work together to provide a win as one is useless without the other. With the high budget a team uses for each Formula One season, answering the research question will help teams better allocate their scarce resources.

**A2. Scope Summary**

The scope of this project is to gather data, analyze it, build a model testing a null hypothesis derived from the research question, and compile our findings into a report. Some aspects of the analysis will be included in the report to provide context for the data analysis. The primary feature of the report is to show the conclusion derived from the analysis on whether or not the car or the driver has a greater impact on a quantifiable success metric. The report will be presented to key stakeholders.

**A3. Solution, Tools, and Methodologies Summary**

The developer environment will be managed with Anaconda. The workspace used will be a Jupyter Notebook. The data required for the solution will be downloaded and imported into datasets using Python and the Pandas library. The Pandas library will be used for cleaning and organizing the data. Pandas will also be used to create attributes required to test our null hypothesis, such as the key metric: time behind the lead racer (or tbl). Once the data is cleaned, organized, and has all attributes needed, tools such as the Matplotlib library and the Seaborn library will be used to construct charts showcasing the analysis used to contextualize the data. The model used to test our null hypothesis will be a mixed effects model. The language R has the library lme4, which will allow for the generation of this model. Using R, we will import a dataframe containing all relevant data needed for the model and generate the model. A summary of the model will be used to view the results. After the model is generated, the R library lattice will generate caterpillar plots showcasing the variances of the cars and drivers' impacts on tbl. Markdown will be used to store all of our findings in a report.

**B1. Project Plan Summary**

The initial plan to gather the data, organize the data, answer the research question, and summarize the conclusions was well followed in the execution. Each step proceeded in the order it was outlined, with little variation. While answering the research question, changes needed to be made in the data organization and the tools used to generate the research question for the model. This revision of the data organization was an expected feature of the project.

**B2. Project Methodology Summary**

The agile methodology allowed flexibility in execution. There was ample room to address the issue whenever an obstacle was encountered, such as the need for a change in the data organization. The development phase of the agile process took the longest of all processes. Paired with the test phase, once the development phase was completed, the testing phase allowed for the discovery of mistakes or areas needing improvement. After which, the development phase would begin again, addressing the problems discovered. Once the testing phase was completed, the rest of the project was finished easily.

**B3. Project Timeline and Milestones Summary**

Revised Timeline

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Milestone | Expected Time | Actual Time | Start Date | End Date |
| Setup and Preparation of Developer Environment | 1 Day | 0.5 Days | 5/7/2025 | 5/7/2025 |
| Import and Clean Data | 1 Day | 0.5 Days | 5/7/2025 | 5/8/2025 |
| Explore and Analyze Data and Generate Charts | 2 Days | 2 Days | 5/8/2025 | 5/10/2025 |
| Conduct Statistical Analysis | 2 Days | 3 Days | 5/10/2025 | 5/13/2025 |
| Generate Report/Summarize Findings | 1 Day | 1 Day | 5/13/2025 | 5/14/2025 |
| Present Report | 1 Day | 1 Day | 5/14/2025 | 5/15/2025 |

**C1. Actual Data Collection Process**

The actual data collection process did not differ from the planned data collection. The dataset was downloaded from the source and imported using Pandas. All the data needed was already identified, so there was no need to revisit the data collection process. Most of the process involved finding and writing the functions needed to best organize the data for analysis.

**C2. Obstacles**

The dataset chosen was of great quality and didn’t present many hurdles in the development process. There were a few attributes that did not have the correct format. These included the date, position, milliseconds, fastest lap, and fastest lap speed. Since they were in strings and had a “\N” value, they had to be correct to a Pandas null value so they could be converted to their proper formats (dates, floats, and integers). Once the data was ready for analysis, it was discovered that the Python package we were using was not sufficient for the analysis, thus, R with the lme4 library was chosen instead. After the data was ready and the right tools were selected, it was a matter of testing different values to ensure the model was working correctly. Due to unfamiliarity with the model type, the mixed effects model was tested with driver and constructor IDs and then retested with driver and constructor names to ensure the results were the same. Since the IDs were stored as integers, we needed to ensure they were tested as categorical values and not as quantitative values. The model was also tested with all tbl (time behind lead racer) values that were null, and again with all null values removed to ensure the model was not accounting for racers who did not finish the race and therefore had a null tbl. The was no difference in both results. The null values were kept out of the dataset to ensure the processing was faster.

**C3. Unplanned Data Governance Issues**

All data governance issues were addressed before the project began. The data acquired was open source and did not legally require any attribution to its source, but a source was still included. Only one person operated on the data, and the audience does not need any special permission to view the report.

**C4. Dataset Advantages and Limitations**

Advantages: The dataset provided us with all of the information needed to answer the research question and test our null hypothesis. It had <26,000 different racer results from different races spanning from 1950 to 2024. Of these results, there were many attributes to choose from, such as the fastest lap, overall time in milliseconds, and finishing position. The dataset was very easy to download and import. Joining datasets together was also very easy, as the schema was built correctly with primary keys and foreign keys.

Limitations: The dataset used is static. The Ergast API is now deprecated. The data can be downloaded, but it will not be updated. This means any future Formula One races will not be added to this dataset by the provider. Another limitation comes with some missing information regarding top lap speed and best lap time for each season before 2004. You can still analyze these factors, but not for older seasons. This metric was not used for the research question and did not affect the study.

**D1. Data Extraction and Data Preparation Process**

Data was downloaded from the source. Each relevant CSV file within the downloaded folder was imported into a Jupyter notebook using Pandas. Each dataset was stored in a Pandas dataframe and was investigated. The main dataset used was the results dataset, which contained data on racing teams’ results for each race. Several attributes were string type and needed to be converted to a numeric format for analytical use. Null values were converted into Pandas null values using the Pandas replace function. After the null values were fixed, the attributes could be converted into their proper data type. Next, the Pandas merge function was used to perform left joins on the results dataset. Constructors, drivers, and races datasets were joined to the results dataset. Only the required attributes from each dataset were included in the mergers. The rest of the operations involved data manipulation. The driver's forename and surname columns were put together into one name and then dropped. The fastest lap time and date were taken from the races dataset and converted to the proper types. The tbl attribute was created by using the group by function to subtract all racers’ times from the lead racer’s time. Columns were rearranged into a preferred order, and the set was saved into a CSV file for analysis. The Pandas library was crucial to data processing. It allowed for easy and accurate manipulation of the data.

**E1. Analytical Methods**

The main analytical method used in this project is the mixed effects model. As our dataset includes many repeating and nested combinations of racers and constructors, this model will allow us to compare the two as random effects in their impact on time behind the lead racer. The value we wish to extract from this model is the variance. A greater variance means that a particular effect has a greater impact.

**E2. Analytical Tools and Techniques: Advantages and Limitations**

Mixed effects models are great for repeated measures of data, modelling fixed effects and random effects, handling unbalanced data, and partial pooling. The fixed and random effects feature allows for accounting of group and individual variability. They can handle unbalanced data or data containing differing numbers of observations for a given subject. In the data, there are wide gaps between the total races constructors and drivers have participated. Partial pooling reduces overfitting by producing stable estimates from groups and individuals through shrinking outliers towards the mean and smaller clusters.

Limitations of mixed effects models lie in their added complexity, lack of certainty found in other linear models, and added computation. Compute is not a large factor in this project, thus, mixed effects models are not a problem. These models are more complex than traditional models and are less straightforward in the results they provide. Mixed effects models have difficulty in hypothesis testing for random effects because of the boundary constraints of variances.

**E3. Analytical Method Process**

A dataset containing the race ID, driver ID, constructor ID, and tbl was imported into a Jupyter notebook using R. Using the lme4 library, a model was generated with the lmer function. For the response, tbl ~ 1 was used as it was the variable we were testing the impact of. There were no fixed effects in this model, so the constructor and driver were used as separate random effects. The data was the imported dataset. The entire function looked like this:

*tbl\_model\_combined <- lmer(tbl ~ 1 + (1 | constructorId) + (1 | driverId), data = f1data)*

Using the summary function, a summary was generated for the model. The variances related to the driver, vehicle, and residual noise were obtained, and we could conclude which had the greatest impact on performance. The variances were added together, and then each variance was divided by the total to obtain the total proportion explained. The variances were copied into a Python Jupyter notebook, and a bar chart was generated showing the differences. In R, the lattice library was used for its dotplot function to generate a caterpillar plot showcasing the variances of drivers and constructors.

**F1. Data Analytics Solution**

The variances returned from the mixed effects model were as follows:

|  |  |  |
| --- | --- | --- |
| Groups | Variance | % |
| Drivers | 109,107² | 59.93% |
| Constructors | 55,157² | 15.32% |
| Residual | 70,118² | 24.75% |

Drivers accounted for the greatest amount of variance at 60%, and constructors accounted for 15%, while the rest was unexplained residual noise at 25%. The drivers’ impact was nearly quadruple that of the constructors on the time behind the lead racer in this model. The null hypothesis of the constructor has a greater impact on time behind the lead is disproven, as the driver accounts for more variation in time behind the lead racer.

Note: variance is equal to standard deviation squared and was depicted as such to reduce the size of the number in the table.

**F2. Practical Significance**

The results of this study are incredibly conclusive that the driver has a massive impact on a team’s time in a Formula One race. While the driver has a larger impact on time, it shouldn’t be assumed that the constructor’s role is diminished. The study does not explain why a driver has a larger impact, but some reasons can be easily inferred. A constructor can make the fastest Formula One car in the world, but if the driver behind the wheel cannot extract that performance, then it is not as impactful as the driver. The study could show that a Formula One team cannot afford a bad driver as much as they can afford a bad vehicle.

**F3. Overall Success and Effectiveness**

The high variance of the driver over all other variances provides great confidence when answering the research question. Formula One teams can use this to understand the importance of having a solidly skilled driver on their team. If a Formula One team were forced to make a cost decision on whether or not to spend more money on a better driver or an improvement to the design of the car, they could use these findings to justify choosing the driver.

The findings provide information on one facet of a Formula One team’s success. As the datasets used showed, there are many other variables associated with success in a Formula One race. Many cars do not finish races. If a team is struggling to finish a race, this study likely wouldn’t help them. The study also doesn’t provide answers as to what a successful driver looks like. A team can use this as a foundation for their competitive philosophy, but should experiment more to find other avenues.

**G1. Conclusions Summary**

The mixed effects model demonstrated that the driver accounted for 60% of the variance in time behind the lead racer, while the constructor accounted for 15% of the variance. This was definitive enough proof to reject the null hypothesis, the constructor has a greater impact on the time behind the lead. The research question *“Is the performance of a Formula One car, concerning time behind the leader, impacted more by the driver's skill or the constructor’s design?”* was answered as affirmative for the driver.

**G2. Tools and Graphical Representations**

* Bar Chart: Used three times in the report, twice to depict the total races, finished races, and wins for the top 5 constructors and drivers, and once to show the differences in variances for the mixed effects model. Bar charts are preferable for comparing the size of variables. For the top 5 racers and constructors, it is evident that for all of the races they participate in, they only win a fraction. The differences in variances chart shows just how much more impact the driver has on tbl than the constructor.
* Line Chart: Used twice in the report to demonstrate how the sport has changed over time concerning total races per year, average laps per race per year, total drivers in a year, and total constructors in a year. Line charts are excellent for displaying chronological data. It is evident that the number of racers participating in races has shrunk over the history of Formula One, and now fewer teams compete each season, however much more competitively.
* Scatterplot Chart: Used twice to show the correlation between the total unique constructors a driver has vs their win rate and vice versa for constructors. The constructor correlation between more unique drivers and win rate is stronger than the driver correlation between more unique constructors and win rate. These plots include trendlines demonstrating the strength of the correlation. While they do not tell us the reason for the relationship, they do show how strong the relationship is between two observations.
* Caterpillar Plot: The most unique plot used in this project, the caterpillar plot is used twice to demonstrate the variances of constructors and drivers’ effects on time behind the lead racer. Each item in the dataset is stacked in order of its estimated magnitude. Each point contains an error bar depicting its 95% confidence interval. The charts used demonstrate that the drivers account for a higher degree of variance than the constructors. These models are unique to mixed effects models as they are required to demonstrate the estimated variance repeated across different groups.

**G3. Recommended Courses of Action**

* A Formula One team should invest more in their driver. Using this study, a team should value the driver over the construction of the car when predicting outcomes for races. If a financial decision is being made and it comes down to spending more money on the driver or the car, they should tend towards the driver. When troubleshooting why the car is underperforming, in its lap times, the driver is more likely to be the cause than the car. This should be adopted into the team’s competitive philosophy.
* Given that the driver has a higher impact on time, a Formula One team should conduct further research on what constitutes a highly skilled driver. The best Formula One drivers by total wins, according to the data, are Lewis Hamilton (105 wins), Michael Schumacher (91 wins), and Max Verstappen (63 wins). Research focusing on the interplay between the driver and the car should also be conducted so that the constructor can design the car to the needs of the driver. While the study was conclusive, a team should be aware of the causes so as to not make an error in assigning causation to the driver when the constructor is at fault. The dataset provided contains plenty of other relevant Formula One race data for other research questions.

**H1. Panopto Video Link**

[https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=26e537ec-0fef-4a41-aa57-b2e0016095e6#](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=26e537ec-0fef-4a41-aa57-b2e0016095e6%23)

**Appendices**

**I1. Code**

Code is contained in the following Jupyter Notebook saved as html files:

* f1-exploration.html
* f1-analysis.html
* mixed-effects-analysis.html

**I2. Data Source**

* [https://ergast.com/mrd/db/#csv](https://ergast.com/mrd/db/%23csv)

**I3. Other Relevant Deliverables**

* summary.md contains a markdown report of all findings and charts. I also included it as a pdf to ensure the images are present, though the spacing between pages is off.