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

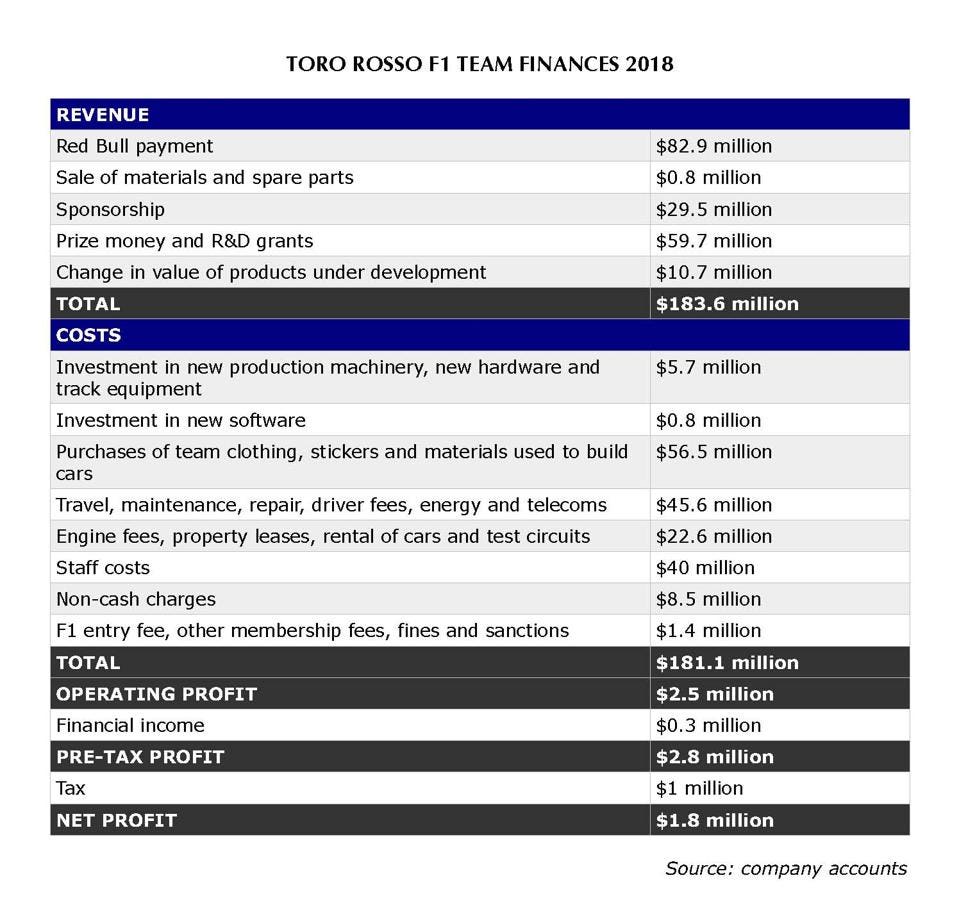
**Task 2**

**A1. Research Question and Organizational Need**

For this study, our research question will be: *“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 driver’s skill in this question refers to the ability of the man or woman behind the wheel. This would include reaction time, mental and physical endurance, decision making, and technique. A skilled driver can extract the maximum amount of performance from a car, even if the car is not necessarily the best-performing vehicle. The constructor’s design refers to the engineering and craftsmanship involved in the production of a Formula One car. Factors would include the aerodynamics, engine configuration, and suspension setup of the vehicle. A good constructor produces a car that is capable of great acceleration, speed, maneuverability, grip, and reliability throughout a race. The purpose of this question is to determine if there is a significant difference in impact between the driver and the constructor, and if so, which one is more impactful. This information, if definitive, could be valuable to a team’s decision-making in allocating scarce resources.

**A2. Context and Background**

Formula One is one of the most competitive motorsports in the world. Teams focus their resources on constructing a car that performs efficiently and reliably throughout a race and training a driver with the skill and endurance to win. Formula One is a very expensive sport. The costs for a season run over $100 million. By regulation, each team is given a limited budget for each season to create a level playing field for teams. Before the implementation of the budget, the gap between team spending was quite large. In the 2019 season, the biggest spender, Mercedes, spent around $484 million, while the smallest spender, Haas, spent $173 million (George, 2019). In 2021, the first budget cap of $145 million was applied to all teams and has since been adjusted to account for inflation. The budget cap does not apply to driver salaries. The reason information about the budget cap is included in this paper is to illustrate how much teams spend. The true cost of a Formula One season is deeper than the budget cap demonstrates. Red Bull revealed their 2018 season balance sheet in a Forbes article. As this was before the budget cap, they spent $181.1 million and received $183.6 million in revenue. All are broken down in the following image from the article.



Sylt, C. (2022, October 12). *Red Bull reveals how much it really costs to run an F1 team*. Forbes. [https://www.forbes.com/sites/csylt/2020/01/14/red-bull-reveals-how-much-it-really-costs-to-run-an-f1-team/](https://www.forbes.com/sites/csylt/2020/01/14/red-bull-reveals-how-much-it-really-costs-to-run-an-f1-team/%20)

Formula One has been around since 1950. The sport has continued to evolve over the decades thanks to advancements in technology. Cars have become increasingly computerized and thus more expensive.

**A3. Published Works Summary**

*"Race to the Podium: Separating and Conjoining the Car and Driver in F1 Racing"* by Duane Rockerbie and Stephen Easton explores the impacts of the driver vs the car in Formula One race results. The authors analyzed race results and financial data from the 2012 through 2019 Formula One seasons, applying a regression to each season, and combining driver and team effects. The use of these methods was to account for season-specific factors. Throughout Formula One's history, teams have applied what is called the 80/20 rule as a heuristic. The rule simply states that 80% of the odds of success are placed on the car and the remaining 20% is placed on the driver. Rockerbie and Easton challenge the efficacy of the 80/20 rule, suggesting there is a more intricate interaction between the car and driver. Specifically, they suggest that 15% is the driver, 20% is the car, and 30 to 40% of the odds can be placed on the relationship between the driver and the car. They suggest that synergy plays a key role in winning. Their financial analysis suggests there is a cooling effect or diminishing returns when it comes to the budget spent on the car and the driver. After a certain point in spending, it becomes evident that more money is being spent on increasingly smaller gains in performance. How does this apply to our study? The key takeaway for us is that the game of Formula One is more elaborate than what a simple analysis can bring us. It suggests that a Formula One team would be mistaken if it used the 80/20 rule in its competitive philosophy. We can observe whether or not the data we analyze suggests that 80% of the odds rest on the car.

*"When Success is Rare and Competitive: Learning from Others' Success and My Failure at the Speed of Formula One"* by Michael Lapré and Candace Cravey uses Formula One as a case study for describing how organizations can learn from mistakes. We care more about their Formula One findings, but the principles are universal. There is only one winner in a Formula One race. This adds to the highly complex nature of the sport. Exploring the learning process a driver and a car undertake is crucial to improving at the sport. How does a car learn? In this specific case, we are talking about the team behind the car. Repeatedly addressing maintenance problems when they arise throughout the race requires careful collaboration and planning for a Formula One team. If a driver notices the car is not turning as it should, the mechanics need to be able to understand what the driver is referring to and how they can best rectify it to keep the car performing optimally. Lapré and Cravey found that teams learn the best from their mistakes and their competitors' wins. The quicker a team learns and adapts, the better they perform. Lapré and Cravey also discovered that drivers are less likely to learn from their own mistakes. This relates to our research question of whether or not the driver or the car has a larger impact. If the team behind the car responds better to failure than the driver, then it could be concluded that the car has a greater impact than the driver over multiple races.

*"A Research Study on Formula One: Determining the Effectiveness of Drivers Based on Their Experience"* by Ignacio Felix-Padilla examines the value of a driver's experience in Formula One racing. Ignacio's dataset includes racing results from 2005 to 2019 and contains data on points scored by drivers and driver salaries. He classifies the drivers into 3 groups: rookies, experienced, and veterans, by their years of experience. The number of points scored by the driver is divided by their salary to determine the effectiveness of the money spent. Ignacio asserts that rookie drivers can punch above their weight, so to speak, when it comes to performance (points scored) vs salaries. Rookies outperformed 30% of experienced drivers and 62% of veteran drivers when it came to cost efficiency. The most cost-efficient drivers were experienced drivers (with 6-13 years of experience). Ignacio's conclusion is that experience does lead to better drivers, but they are severely overpaid. How does it relate to our research question? This study asserts that there is a point at which a driver's experience is beneficial to the team's overall performance, but they can also be overpaid. Since we are trying to observe the impact of car vs driver for budget allocation purposes, we can see what experience level for a driver leads to a worthwhile investment. If the driver has a higher impact on performance, a team might be more willing to incur a higher cost to obtain an experienced driver. Ignacio's study only considers drivers, which doesn't give a complete picture, but is still useful.

**A4. Deliverables**

To best address 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?" we must provide a detailed summary of Formula One data. The summary will be organized into a markdown file with charts and graphs showcasing insights from the data. A combination of Python with the pandas toolkit will be used to clean and analyze data, the matplotlib library for generating charts, as well as R for modeling and statistical analysis. The report will address the research question while also contextualizing the data. It will show why the data is a good fit for addressing the research question and what was and wasn't included in the final analysis. The goal of the report is to provide a clean perspective to a Formula One team seeking to improve their results over their competition by best allocating their resources. The target audience for this report will be people who are familiar with the data analytics process, as well as the background behind Formula One racing.

**A5. Analytical Solution**

This analytical solution will help the prospective Formula One team best allocate its scarce budget. They will be given insights on which makes a greater difference and thus requires a larger share of the budget proportionate to its current cost. The key purpose of this study should be to act as a lead to point a team's decision makers in the direction of further experimentation. In a sport as varied and complex as Formula One, there are no easy solutions. The findings of this will be a starting point upon which to build. Another benefit this project offers is the ability for teams to better understand the intricate relationship between the car and the driver. Observing historical data related to different racers racing in different cars will provide teams with an understanding of the conditions and combinations that create an exceptional performance on the track, concerning time, or even a win.

**B1. Goals, Objectives, Deliverables**

The primary goal of this project is to generate a report containing our findings from the data analysis of Formula One race results data. The report will address the research question of whether the car or the driver has a larger impact on the total time behind the lead racer of a Formula One team.

Our first objective will be to summarize our data. The deliverables will include tables and charts such as bar charts, line charts, and scatter plots. These could depict racing statistics, how the sport has changed, and the relationships between differing data points. This will add necessary context for understanding our conclusions.

Our second objective will be to answer the research question. We will use the data we have to test whether there is a stronger correlation between the car and the driver in influencing the driver’s total time behind the lead racer. The deliverables will include statistical models used to answer the research question.

Our final objective will be to summarize our conclusions. The deliverables will include a written analysis of our findings along with any relevant charts.

**B2. Scope**

This project's scope will include finding and importing a viable dataset containing Formula One race results data. The data will be cleaned in a Jupyter notebook using pandas. Whatever data we group together will be stored in a .csv file for easy access. Once we have all of the data we need and in a workable form, we will begin our analysis. This will include aggregating the data into valuable insights. Examples might include drivers with the most wins, constructors with the most wins, average win rate by driver, best lap speed by constructor, etc. Most importantly, we want to include the time behind the lead racer, as it is the metric we will test. Useful insights will be used to generate charts using Matplotlib and Seaborn. The charts will be saved in SVG format for the report. All Python code used to clean and analyze the data will be saved in Jupyter notebooks. For statistical analysis tools like Scikit-learn with Python and lme4 with R may be used. All statistical analysis and charts will be included in the final report, which will be written in a markdown document and saved as a PDF. Features that are out of scope for this current project would include a dynamic data storage system. Formula One is an ongoing sport, and data from current seasons could be gathered and stored in a SQL database and used to generate dynamic reports in Tableau. Machine learning will also be excluded from the scope of this project. Machine learning could be used to predict potential outcomes by building models from existing data.

**B3. Project Planning Methodology**

For this project, we will use an Agile project management methodology. Agile is a great methodology for its flexibility with changing requirements. With data analysis, there are a lot of unknown variables encountered, and being able to adapt quickly is crucial. A plan will be developed for analysis, but contingencies need to exist for whatever obstacles are encountered. To provide the right information to key stakeholders, there needs to be communication to see what they want and if those wants have changed as the project begins to take shape. Our project will follow these steps: Gather Requirements, Design, Develop, Test, Deploy, and Review.

**Gather Requirements**: This phase will involve developing criteria for what data we want to collect. Collecting the right attributes to conduct our analysis is important. We will determine how large we want our dataset to be. We might set a date range for what seasons we want to include in our analysis. We will need to determine what information is and isn't relevant to our study. We will need to work with key stakeholders closely throughout this phase. We will need to understand their perspectives and what information they think we serve them best. We will need to outline the scope of the project and set reasonable expectations as to what we can accomplish in the timeframe given. Key dates will need to be established for when the next phase will start and when it can be expected to be finished. A budget will also need to be established for the project. We will need to determine what tools will be needed to accomplish the project and how much they will cost.

**Design**: This phase will involve high-level architecture. We will have our requirements outlined from the previous stage and the tools we need to fulfil them. Now, it's time to develop our plan. We will begin by outlining a process for what data we will gather and the process by which we will begin to clean and aggregate it. Once we have the dataset cleaned, we will need to plan what analysis we will run first and where we can go from there. We will also need to determine what format we want to use for this data. All of this must be done to meet the stakeholders' requirements.

**Develop**: Once we have an idea of what we want to do, it is time to execute. We will begin by finding and importing a dataset. Once we have the data, we will write Python pandas functions to clean and organize the data into a better readable format. If we feel that the data does not meet the stakeholders' requirements, then we may need to look at obtaining another dataset. There must be close communication throughout this process to ensure requirements are fulfilled. When we have the right dataset, we will continue to analyze it by creating charts with Matplotlib and Seaborn. Statistical modeling can also begin. We can utilize tools such as Scikit-learn and lme4 for modeling. We will continue this until we develop a report.

**Test**: Once we have finished our initial report, it is time to perform tests on our data and analysis to ensure it was processed correctly. This would require other user input or functions used to ensure data integrity. The goal of this phase is to replicate results. If results cannot be replicated, it means there may be an error in the data or functions that need to be addressed. The testing process should be repeated until we are confident in the project we plan to deliver.

**Deploy**: This phase involves releasing the finalized report to key stakeholders. This could include a presentation of findings and distributing the report to key stakeholders, whether by PDF documents or paper copies. We will need to be ready to answer any questions stakeholders may have about the report and the data.

**Review**: After deployment, we must collaborate with stakeholders and developers to examine how the project went. This will help with future iterations and analysis. A retrospective on the process will help smooth future processes and ensure improvement in analytics.

**B4. Timeline**

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone | Expected Time | Start Date | End Date |
| Setup and Preparation of Developer Environment | 1 Day | 5/7/2025 | 5/8/2025 |
| Import and Clean Data | 1 Day | 5/8/2025 | 5/9/2025 |
| Explore and Analyze Data and Generate Charts | 2 Days | 5/9/2025 | 5/11/2025 |
| Conduct Statistical Analysis | 2 Days | 5/11/2025 | 5/13/2025 |
| Generate Report/Summarize Findings | 1 Day | 5/13/2025 | 5/14/2025 |
| Present Report | 1 Day | 5/14/2025 | 5/15/2025 |

**B5. Resources and Associated Costs**

|  |  |
| --- | --- |
| Item | Cost |
| Laptop with Windows 11 OS | **$800.00** |
| Visual Studio Code | **$0.00** |
| Anaconda | **$0.00** |
| Formula One Data | **$0.00** |
| Employee Salary | **$0.00** |
| Total | **$800.00** |

**B6. Success Criteria**

For this project to succeed, the following criteria must be met.

* A cleaned dataset organized from a reliable data source will be provided for analysis and modeling.
* A mixed effects model that demonstrates the differences in variance between constructors and drivers’ impacts on time behind the lead racer will need to be generated.
* A report containing a summary of the analysis, models, and key findings needs to be provided and presented to key stakeholders.

**C1. Hypothesis**

Null Hypothesis: The car's impact is more significant than the driver's on determining the time behind the lead racer.

**C2. Analytical Methods**

Diagnostic analytics will be used for this project. Diagnostic analytics are focused on discovering causes from historical observations. Descriptive analytics summarize data and are not a good fit for answering our research question. The interplay between the driver and the car is a key aspect of Formula One. Diagnostic analytics allows us to compare drivers and their relationship with cars built by different constructors. By comparing two categorical variables (driver and car) against a quantitative variable (time), we can diagnose which is more impactful or if there is even a significant difference in impacts between the two.

The specific analytical method we will use for this is a mixed effects model. Mixed effects models will allow us to explore both fixed and random effects. The mixed effects model best suits this project because the observations being analyzed are nested and repeated, as there are different combinations of car and driver repeated throughout the data. The mixed effects model will provide the respective variance for the effects tested. If one effect explains a higher variance than the other, then it can be assumed that the effect (constructor or driver) has a greater impact on the variable being tested (time behind lead).

**C3. Tools and Environments**

For our tools and environments, we will be using the following:

- VS Code: For writing, editing, and testing code

- Python: primary language used for data manipulation and cleaning.

- Pandas: Python package for manipulating and cleaning data

- NumPy: Python package for numerical computing

- Matplotlib: Python package for generating statistical charts

- Seaborn: Python package used in tandem with Matplotlib for sharper chart generation

- Statsmodels: API for statistical modeling in Python

- R: a programming language for data analytics

- tidyr, dplyr, lubridate: R packages for easier data manipulation

- lme4: R package for the generation of linear mixed-effects models

- Jupyter Notebooks: a tool for testing Python and R code in segments

- Anaconda: an environment management service for handling Python and R packages

- Markdown: a file format the report will be structured in

**C4. Methods and Metrics**

To test the null hypothesis, we will be using a mixed effects model. This model type allows for the incorporation of both random and fixed effects. Fixed effects are factors that are repeatable and consistent throughout the data. Random effects are factors that are inconsistent and non-repeatable. Mixed effects models are applicable for analyzing data with recurring variables. This is useful for our purposes because we are dealing with the same car constructors and the drivers repeating with various combinations of the two. For this model, we will be comparing the car and driver as unique categorical variables and the total time behind the lead. We are trying to determine which explains a greater amount of variance in the time behind the lead. The mixed effect model will give us the variances associated with the driver, car, and residual noise or variance that is not explained by the car or driver.

**C5. Practical Significance**

If the variance caused by the driver is significantly higher than the variance caused by the car, then we will reject the null hypothesis. A higher driver-related variance means that the driver has a bigger impact on the race and therefore should be placed higher on a team’s list of priorities. This does not mean the driver’s salary should be higher than the cost of the car, but when the team is looking to gain an advantage, they might look at the driver over the car and vice versa if the study fails to reject the null hypothesis. The results could impact the way teams look at certain concepts in Formula One, such as the previously mentioned 80/20 rule. If the car explains four times the variance as the driver, then the 80/20 rule would likely have merit.

**C6. Tools and Graphical Representations**

Using Python tools, Matplotlib and Seaborn, several bar charts and tables will be generated to showcase the data used and provide context for the statistical model. A chart showcasing the number of unique drivers per car constructor and vice versa will be used to demonstrate that there are sufficient numbers of combinations of cars and drivers. The variances gathered will be showcased in a table and bar chart. Using the R tool, lattice, A caterpillar plot will be used to depict the differences in variances in impact between the driver and the constructor. Caterpillar plots are unique to mixed effects models. They allow for the demonstration of the uncertainty of random effects. A caterpillar plot will be provided for each random effect or group, in this case, drivers and constructors.

**D1. Source of Data**

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

**D2. Why this Data?**

This data source provides multiple CSV files. The files used will include:

Constructors: 212 rows x 5 columns. Attributes used are constructorId and name. This will provide us with a list of all the constructors. The ID will allow for the joining of other datasets.

Drivers: 860 rows x 9 columns. Attributes used are driverId, forename, and surname. The forename and surname will be merged into one full name. This will give us a list of all drivers to join with other datasets.

Races: 1125 rows x 18 columns. The attributes used are raceId and date. This will allow us to aggregate by race as well as when the race happened. The date will be further broken down into day, month, and year.

Lap Times: 588,048 rows x 6 columns. Attributes used are raceId, driverId, lap, and milliseconds. The key information we are trying to extract from this dataset is the fastest lap time in milliseconds in a given race by a given driver. The results dataset has a fastest lap attribute, with which we will pull the time in milliseconds from this dataset.

Results: 26,739 rows x 18 columns. Attributes used are resultId, raceId, constructorId, position, overall time in milliseconds, fastest lap speed, position, total laps, and whether or not they finished the race. This is the main dataset we will join with the previously mentioned datasets. The stats we will be analyzing will mostly come from this set. Time behind lead, the metric we are investigating in our mixed effects model, will be created by subtracting every racer’s time from the first-place racer’s time.

There is a lot of data here. The attributes we have chosen are there to help contextualize and answer the research question for stakeholders. Data such as fastest lap time and fastest lap speed will be included in the analysis to show stakeholders how the sport has evolved or stayed the same. The data related to the research question will be the IDs for the constructors and drivers, and the time behind the lead racer for each racer over every race. Racers who did not finish will be given a null value for time behind the lead and will not be included in the study. The data selected is sufficiently sized and contains all of the details needed to test the hypothesis.

**D3. Data Collection Methods**

The process used to collect this data simply involved downloading a zip file containing all of the CSV files from the source provided. From there, pandas was used to import and organize the data into dataframes.

**D4. Observations of the Quality of the Data**

The quality of the data was good. Certain integer values were stored as text, and the null values needed to be changed to a format that would allow the value type of the attribute to be changed to integer. Aside from changing columns to integer format, another column needed to be changed into date format so month, day, and year values could be extracted. Forename and surname were taken from the driver and combined into one full name for brevity. The bulk of the work needed to be done on the data was organizing the datasets into one results dataset for analysis. It could be said that there was too much data within the dataset, however, the case could easily be made so that other useful studies could be performed from the data. It was easy to locate and wrangle the data needed for this project.

**D5. Data Governance**

The source of the data, Ergast, does not require attribution but appreciates it. They ask for a reference to Ergast and a link to their site to be included if the report is being shared with a technical audience. Since the data is open source, security measures are not required for the data itself. If the stakeholders do not want the analysis provided to be open information, then certain data governance measures should be applied. A list of stakeholders authorized to view the report could be generated to limit who can view the report. If physical copies of the report are made, they could be collected after the presentation. If electronic copies of the report are shared, they could be encrypted, and each authorized viewer would be issued a key to access the report. There is no incriminating or sensitive information contained in the data itself, so there can be no repercussions for handling the data.

**References**

George, D. (2019, December 20). *Amid budget cap for 2021, how much is the current budget of teams in 2019?*. EssentiallySports. [https://www.essentiallysports.com/what-are-the-budgets-for-all-10-formula-one-teams-2019/](https://www.essentiallysports.com/what-are-the-budgets-for-all-10-formula-one-teams-2019/%20)

Sylt, C. (2022, October 12). *Red Bull reveals how much it really costs to run an F1 team*. Forbes. [https://www.forbes.com/sites/csylt/2020/01/14/red-bull-reveals-how-much-it-really-costs-to-run-an-f1-team/](https://www.forbes.com/sites/csylt/2020/01/14/red-bull-reveals-how-much-it-really-costs-to-run-an-f1-team/%20)

Rockerbie, D. W., & Easton, S. T. (2022). Race to the podium: Separating and conjoining the car and driver in F1 Racing. *Applied Economics*, *54*(54), 6272–6285. <https://doi.org/10.1080/00036846.2022.2083068>

Lapré, M. A., & Cravey, C. (2022). When success is rare and competitive: Learning from others’ success and my failure at the speed of Formula One. *Management Science*, *68*(12), 8741–8756. [https://doi.org/10.1287/mnsc.2022.4324](https://doi.org/10.1287/mnsc.2022.4324%20)

Felix-Padilla, I. (2023). A Research Study on Formula One: Determining the Effectiveness of Drivers Based on Their Experience. *Wharton Sports Analytics Journal*. <https://wsb.wharton.upenn.edu/wp-content/uploads/2023/12/Padilla_2023_N.pdf>