The Evolution of Speed and Performance in Formula 1

Contents

Project Overview 1

Research Question 1: How have race pace and reliability changed over time across eras? 2

Research Question 2: How Do Drivers' Lap Times Change Over the Course of a Race? 6

Research Question 3: How Does Qualifying Performance Affect Final Race Outcomes? 9

Research Question 4: How Do Pit Stop Durations Influence Race Finishes and Season Rankings? 12

Wrapping up 15

Future Research 15

Author
Leslie Huh, Nick Jitjang, Shria Parupudi, Sohum Walavalkar

Project Overview

Motivation

Formula 1 is a sport where very small differences in pace, reliability, and strategy can completely change the outcome of a race. Fans and teams regularly debate what matters most: how fast the cars are over a stint, how reliable they are over a season, whether qualifying position locks in the result, or whether pit stops and race strategy make the real difference. Since almost every aspect of an F1 weekend is measured (lap times, pit stop durations, grid positions, final results) the sport provides a rich setting to study these questions using real data rather than just opinions.

In this project, we use historical F1 data to take a broader view of performance. Instead of focusing on a single race or driver, we combine information across many seasons to look at how pace, reliability, qualifying, and pit stops all contribute to race and season outcomes. This makes the project interesting both from a sports perspective and a statistical perspective as the dataset is large, structured, and naturally lends itself to methods like time series plots, clustering, and PCA that we have learned in class.

Dataset Description

We use the Formula 1 World Championship dataset from Kaggle, which compiles results from many seasons of Grands Prix. The data are spread across several csv files that link together through IDs like raceId, driverId, constructorId and circuitId. The main file for race outcomes is results.csv, where each row represents one driver in one race and includes their grid position, finishing position, points, team and a status code. Race level details such as the year, round, race name and circuit come races.csv. Driver information (name, 3 letter code, nationality) is stored in driver.csv and team information is in constructors.csv.

For our specific questions, we also use more detailed files. lap_times.csv has one row per driver-lap and gives us lap times so we can study how pace changes over a race. pit_stops.csv has one row per pit stop with the lap and duration of the stop, which we use to look at pit strategy. Finally, status.csv translates the status codes in the results into readable labels such as 'Finished', 'Engine', or 'Accident', which lets us analyze reliability. Together, these files lets us build race-level, driver-season, lap-level, pit stop-level datasets that are easy to work with for our specific research questions.

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

Research Questions

Below is the list of our research questions we will be addressing throughout this report. We will be focusing on different aspects of Formula 1 that influences the performance of the cars/drivers.

RQ1: How have race pace and reliability changed over time across eras?

RQ2: How Do Drivers' Lap Times Change Over the Course of a Race?

RQ3: How Does Qualifying Performance Affect Final Race Outcomes?

RQ4: How Do Pit Stop Durations Influence Race Finishes and Season Rankings?

Research Question 1: How have race pace and reliability changed over time across eras?

Motivation

In Formula 1, both race pace and reliability largely dictate how each driver performs. The faster the race pace, the faster the driver can drive, which increases their chance of winning. On the other hand, reliability is also a critical factor as cars need to be able to finish the race without problems that would interfere with the driving. Therefore, it would be worthwhile for us to investigate these two factor and how they evolve over time as teams adopt new engine regulations and introduce new technologies. Because engine eras are strongly tied to performance and durability, studying how fastest-lap speeds and DNF (Did Not Finish) rates change across eras would help us understand whether newer regulations made cars faster or more reliable. This is especially interesting as F1 is making a transtion from bigger more powerful engine to a smaller engine that features more technonology and more friendly to the environment. Thus, the two graphs below would help us understand how reliability and race has evolved over time.

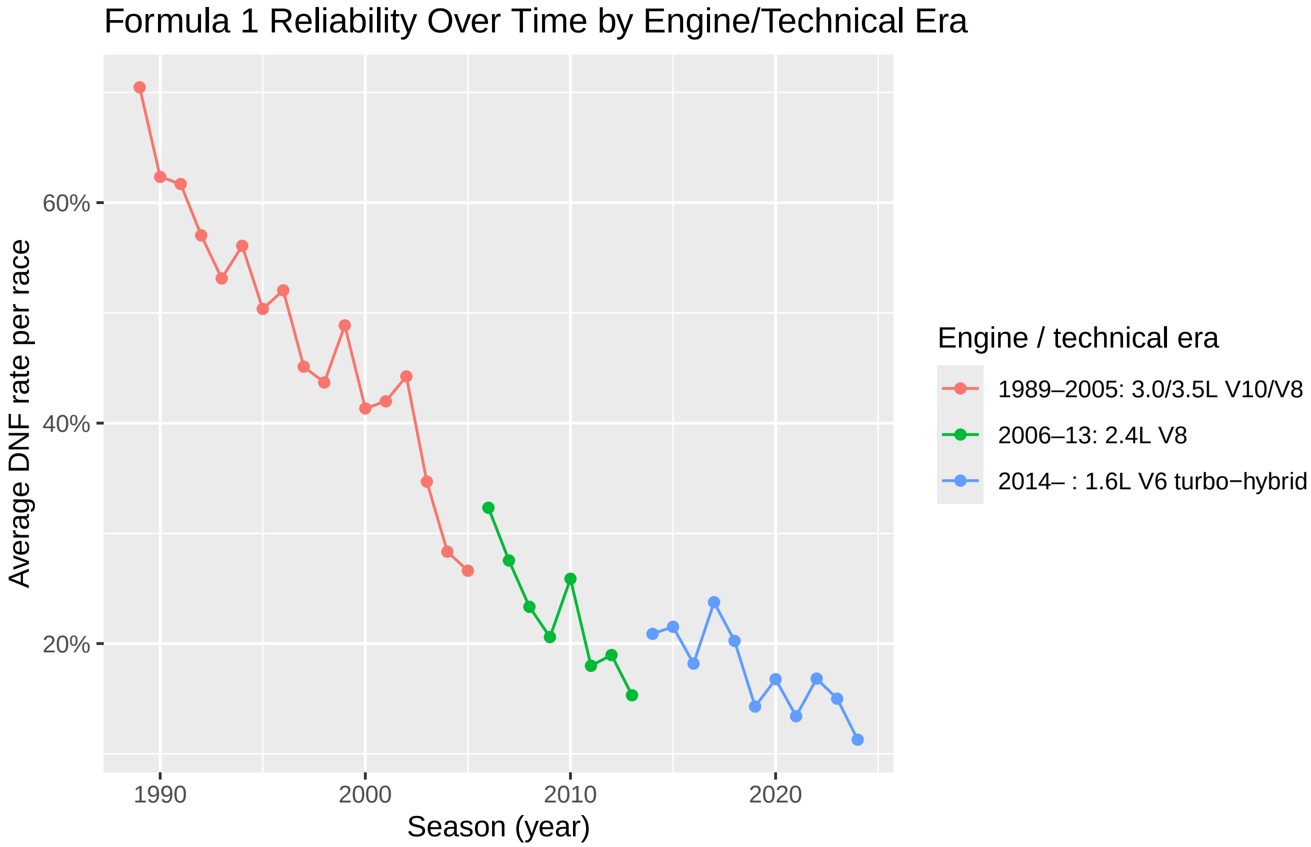
Note: for this question, we will only focus on 1989 - 2024 even though the dataset has information starting from 1950. This is because the data for fastest lap, which is used to interpret the race pace section, does not have data before 1989. Thus, to make the eras consistent for the two graphs, we will focus on 1989-2024 or the "modern era".

Visualization 1

This graph displays how race reliability evolves over time by plotting:

- x-axis: year
- y-axis: mean DNF rate per race in that season
- color: engine/technical era (The F1 eras aren't given to us directly from the dataset, but we are able to make that up using public information on what kind of engines were use during each year to creating grouping.)

A time-series line plot is appropriate here as we are tracking how a different engines eras changes over chronological time and what impact that those has on the average DNF rates for each race aggregated across the different years. The plot provides a good concrete performance metric to assess the reliability of F1 cars over time. Since we can't really measure "reliability" directly, the graph helps us interpret this information by coming up with a measurement of DNF per season that highlights how reliability is increasing or decreasing. Thus, we are able to answer the original question for the reliability part that the reliability have been increasing over the time and eras of F1.



Interpretation

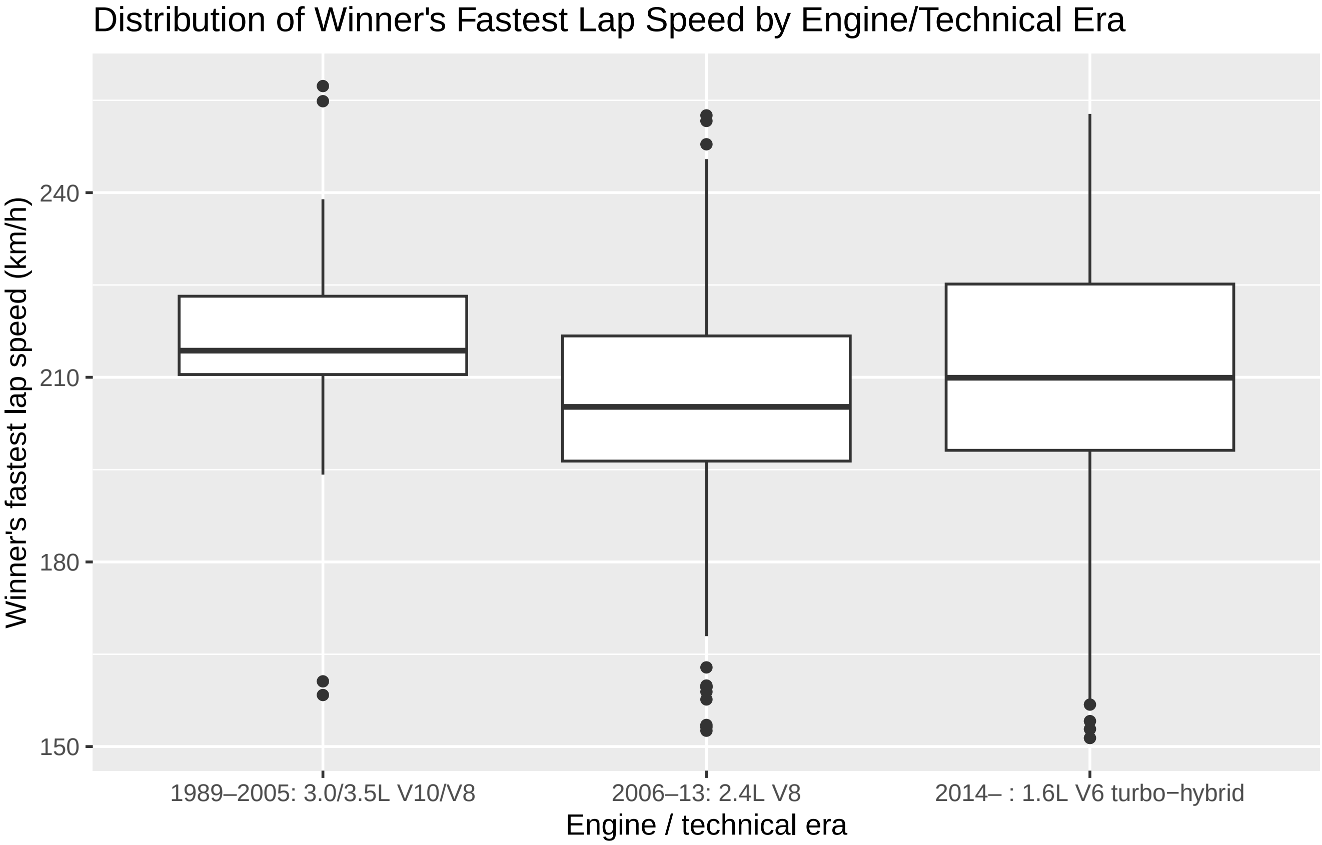
We can observe a clear downward trend in DNF rates across eras, showing that the reliability of the F1 cars have increased significantly over the years as they are switching to a more reliable engine and the DNF per season is going down. The most interesting era is the 1989-2005 era when the average DNF goes down drastically over the year when F1 cars were using v8 and v10 engines. This is a very important time of F1 as F1 was still focusing largely on naturally aspirated engines. Then, in the 2.4L V8 era from 2006-2013, the DNF rates jumps to a higher level compared to the last year of the previous era, but starts decreasing to lower level than the previous era after that. The 1.6L hybrid era from 2014 - present (2024) shows a similar trend where the DNF rate jumps to a higher level initially and decreases, showing the lowest DNF levels overall. Thus, we see a trend here where the DNF rates are decreasing for each era as time progresses, and the DNF rate can spike a little when there is a transition. However, a newer and more technological advanced engine still achieves a lower DNF rate than the previous era. Therefore, the steady decline across eras suggests that engineering, safety, and component durability have all improved substantially.

Visualization 2

This plot shows how the race pace of winning drivers changes across three modern engine eras by plotting:

- x-axis: engine/technical era
- y-axis: winner's fastest lap speed

A boxplot is used here because it summarizes the full distribution of lap speeds within each era (median, quartiles, and variability). By examining the fastest-lap speed, we can measure how race pace, how variable of interest, changes over the time as faster lap speed corresponds to faster race pace. Since pace is a continuous measurement and eras are categorical, boxplots allow us to compare of how pace distributions shift across eras to help us answer our research question for the race pace part.



Interpretation

The boxplot shows an interesting insight of the race pace across eras. First, we can see that peak lap speeds were highest in the 1989–2005 V10/V8 era, reflecting powerful naturally-aspirated engines shown by the outlines. We can also see that the median speed for this V10/V8 era is also the highest among the three eras, and that the distribution seems to be the least variable as the box and line are not wide compared to the other eras, suggesting that the bigger engines seem to allow for faster and more consistent lap speed. Looking at the 2.4L V8 era, the median fastest lap speed seems to be the lowest among the three eras as F1 makes a transition to a smaller engine. This is further supported by looking at the lower quartile where this era has the lowest quartile lap speed overall. However, within the hybrid era, the upper whiskers, upper quartile, and median rise again, indicating that hybrid technology has steadily improved over time. We can also see that the upper quartile is higher than the V10/V8 era in addition to the upper whisker being taller. Thus, we can see that the race pace is improving within this era again, although it may not be as consistent as the old big naturally aspirated V10/V8 era. Therefore, this supports the idea that engine regulation changes meaningfully affect race pace.

Conclusion

The two visuals collectively show that race pace and reliability were definitely influence by the engines and technical eras of Formula 1. As we make more technological advancement and engineering changes from bigger engines to a smaller and more environmentally friendly engine, we see a fluctuations in performance and reliability, affecting how the sport needs to adapt to changing rule and condition. First, We can see that the overall trend for reliability has been on a positive side where the engines are becoming more and more reliable with each passing year as DNF rate has been decreasing throughout with some fluctuations transitioning from era to era. Even though we get a more reliable engine, we can also see that the smaller engine seems to be slowing down the race pace over the year as there is more variability in the fastest lap speed from the two latest eras. However, the most recent hybrid era seems to be catching up with the old big engine as the sport continues to make engineering advancements. Therefore, the engine eras positively affects the reliability as it has been improving over time while the race pace struggled initally due to smaller engine, but are improving with the recent hybrid advancement.

Further Research

While our analysis provides an interesting insight, it is still limited due to only using a few factor to consider a large question. In other words, our work only focuses on DNF to measure reliability and fastest lap speed to consider race pace. Future work could explore several unanswered questions about race pace and reliability and go in more detail for each. For instance, while reliability generally improves across eras, variability within periods like the early V10/V8 era suggests underlying technical or circuit-specific factors not captured in our dataset. We could analyze the weather, tire strategies, or also examining the cause of DNF (which can be caused by a crash rather than the car's fault) to fully unfold how each affect race pace and reliability as team makes different decisions to try to win the race. Additionally, our current analysis focuses only on winners' pace while richer insights could come from examining all-driver pace distributions (which is not captured by the dataset) to make sure that the data is more representative of the full era. These directions were beyond the scope of this project due to data limitations and the need for more advanced statistical methods, but they could lead to a more complete understanding of how F1 technology and regulations shape on-track performance.

Research Question 2: How Do Drivers' Lap Times Change Over the Course of a Race?

Motivation

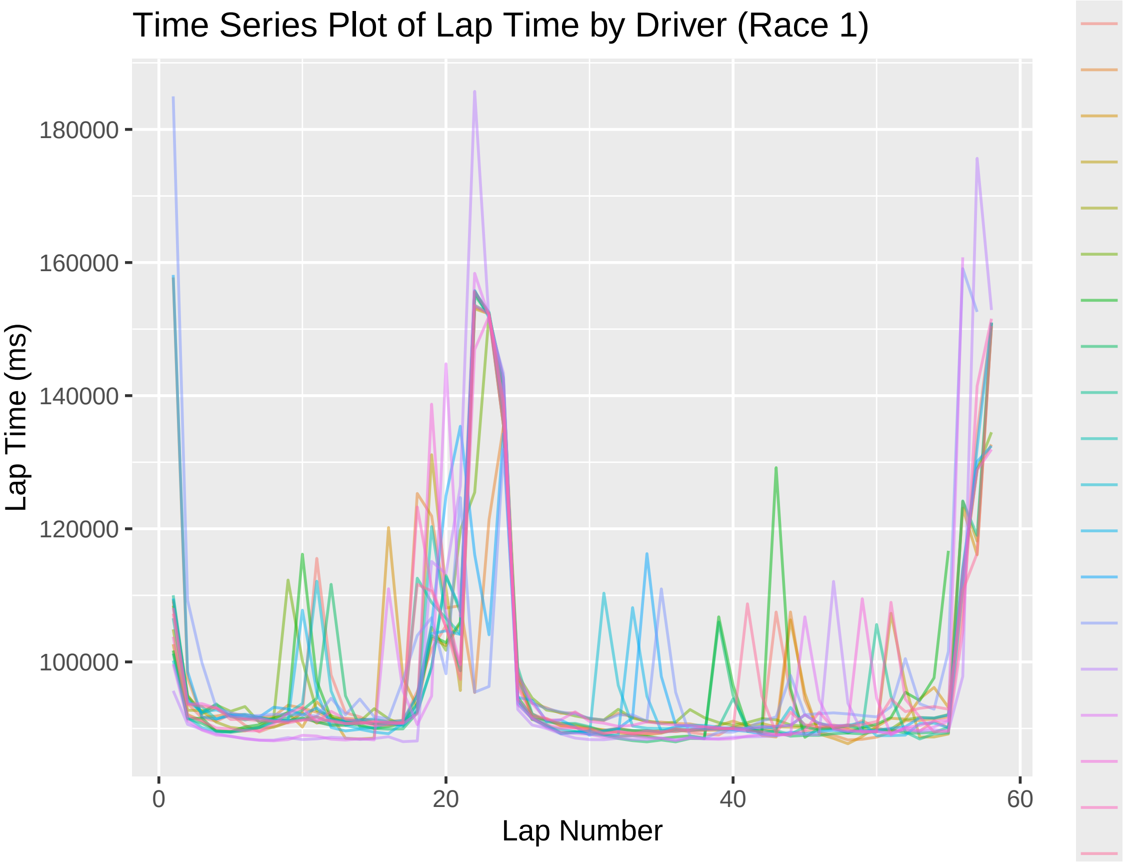
In Formula 1 Races, it's clear that a driver's lap time is a direct measurement of their performance. Drivers need to obtain the lowest lap times while balancing competitors' effects, fuel efficiency, pit stops, and other dependencies that influence their performance. Lap times may vary during the course of a race as a byproduct of these covariants. For example, a driver may have one or two really slow laps due to stopping for a pit stop in the middle of their race. Understanding overall how a driver's lap times change over the course of a race is valuable as it allows for dive more in-depth analysis than just analyzing final times and positions allows for significant real-world implications; for example, enabling drivers to analyze their distinct races in efforts to optimize performance in the future.

Visualization 1

This graph displays how drivers' lap times change throughout Race 1 of the dataset by plotting:

- x-axis: Lap Number
- y-axis: Lap Time (in milliseconds)
- color: Driver ID

A Time Series Plot is appropriate for answering the question: in race 1 in the dataset, how does drivers' lap times change over the laps in the race? Here, each lap number presents a chronological time period within a race, and lap time is the performance metric that changes over the course of a race.



Interpretation

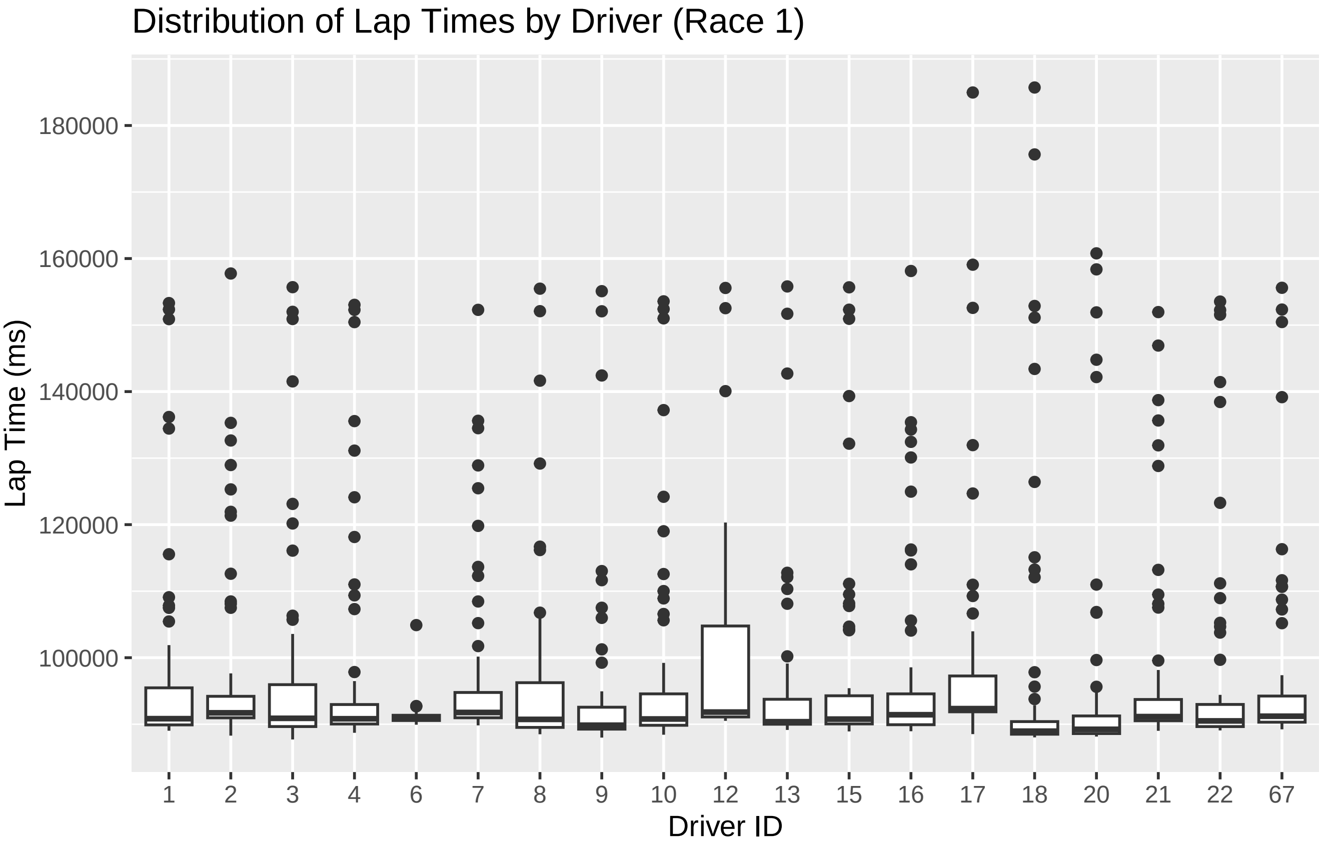
Here, we can analyze how performance changes during a specific race and by driver. To analyze this, we can plot a time series plot where each driver's lap time is plotted for each lap number. From the graph, we can see that lap times in general are relatively fast for most drivers, followed by two spikes with a huge spike around lap 21; these spikes may mean that drivers get pit stops at roughly the same lap. Following that, lap times seem to stabilize back to really fast, with a few drivers having spikes. To summarize, across most drivers, the curve of their lap time follows the same pattern; perhaps they have the same strategy. This plot directly answers the question I aimed to address since it shows all 3 variables effectively to show the change in lap times in a race: a time series effectively captures that. Lap number here is the time, lap time measures performance, and the plot is colored by driver ID.

Visualization 2

This graph shows how lap time performance varies across drivers within Race 1 by plotting:

- x-axis: Driver ID
- y-axis: Lap Time (in milliseconds)

A Box Plot is appropriate here because we can analyze the distribution of lap times for each driver in a single race. This allows us to evaluate how median, quartiles, and variation of lap times change over the course of a race. This allows to get granular in analyzing how lap-time performance differs between drivers. It's a direct complement of the previous visualization that provided insight into how lap times change over time. This allows us to determine which drivers were the most consistent across a race and what outliers exist among drivers.



Interpretation

From the plot, we can see that most drivers share a similar median lap time, roughly around 50,0000 ms. However, there exist many outliers amongst all drivers. These outliers are sharp increases from the average lap time of each driver and may be associated with pit stops or race accidents. One driver, driver 12, has significant variation among his lap times relative to their competitors. This suggests that driver 12 may have had a race accident, enabling heavy variation in his lap times. However, further research is needed to investigate this correlation. Some drivers also have relatively less variance than competitors: drivers 6, 18, and 20 maintain consistent lap times across the course of race 1. Overall, while lap times were fairly consistent across drivers, each driver had significant outliers.

Test

To further test if all drivers have the same mean lap time, we can utilize an ANOVA test. Here we have:

- Null Hypothesis: All drivers have the same mean lap time
- Alternate Hypothesis: At least one driver's mean lap time is significantly different

```
##                             Df    Sum Sq   Mean Sq F value Pr(>F)
## as.factor(race1$driverId)  18 1.556e+09  86467688   0.328  0.996
## Residuals                 986 2.596e+11 263312406
```

Interpretation

Here, we have an F value of 0.328 with a p-value of 0.996. As the p-value is greater than an alpha of 0.05, we fail to reject the null hypothesis. Thus, we can conclude that there's no statistically significant difference in drivers' mean lap times in Race 1. This reaffirms our intuition that drivers shared similar lap times in Race 1. Although every driver had outliers for their lap times, overall, they were pretty consistent relative to each other across the race.

Conclusion

The visualizations allow us to determine that drivers' lap times change in a structured and predictable way over the course of a race. From the time series plot, we can see that ap times in general are relatively fast for most drivers, followed by two spikes with a huge spike around lap 21; these spikes may mean that drivers get pit stops at roughly the same lap. Spikes in lap times are represented by the outliers in our second visualization; here, we can see that most drivers maintain similar lap times, but each has outliers. Driver 12 also displays high variance in their lap times from the box plot. It is supported by large spikes in lap times for Driver 12, roughly around lap 21 and again from laps 50 to the end of the race. To reaffirm our intuition that drivers shared similar lap times in Race 1, we conducted an ANOVA test with the null hypothesis that all drivers have the same mean lap time. We determined that there's no statistically significant difference in drivers' mean lap times in Race 1. Thus, overall lap times are relatively consistent across drivers and change in predictable ways, signalling dependencies on race strategy, pit stops, and race accidents.

Further Research

This study is limited in a couple of ways: it focuses on Race 1 of the dataset and may not be generalizable to all F1 races; further research can extend our analysis to multiple races, tracks, locations, and seasons. Speaking of seasons, this study doesn't account for external variables that affect lap times, like tire differences and weather conditions. Further research can include these as variables. More research will enable a more granular understanding of how performance evolves throughout a Formula 1 race.

Research Question 3: How Does Qualifying Performance Affect Final Race Outcomes?

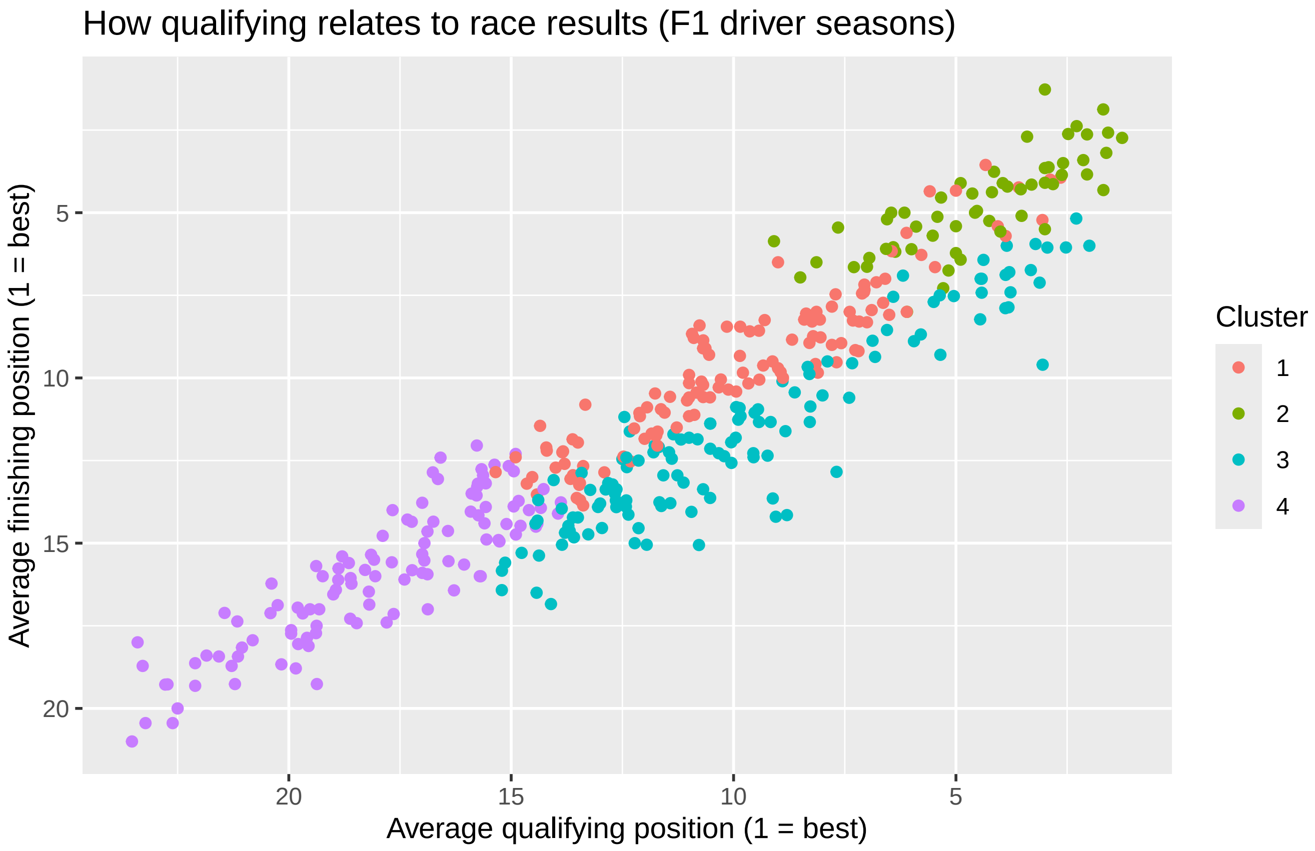
Motivation

Qualifying is often described as “half the race” in Formula 1, especially on tracks where overtaking is difficult. Teams invest huge effort into one-lap pace and grid position, but on race day we still see drivers moving forward or dropping back. For this research question, I focus on quantifying how strongly qualifying performance is tied to final race outcomes at the season level. In other words, if a driver generally starts near the front, do they reliably finish near the front, and how different is the experience for drivers who usually start in the midfield or at the back?

Visualization 1

For the first visualization, I made a scatterplot where each point represents one driver-season. The x axis shows that driver’s average qualifying position for the season, and the y axis shows their average finishing position, with both axes reversed so that smaller numbers (better positions) appear toward the top-right. I color the points using the cluster labels from hierarchical clustering so that seasons with similar qualifying and race performance are grouped visually. This graph directly addresses my question by showing how starting positions and finishing positions move together across seasons, and whether there are clear patterns such as front-running, midfield, and backmarker seasons.

```
##
##   1   2   3   4
## 123  56 129 112
```

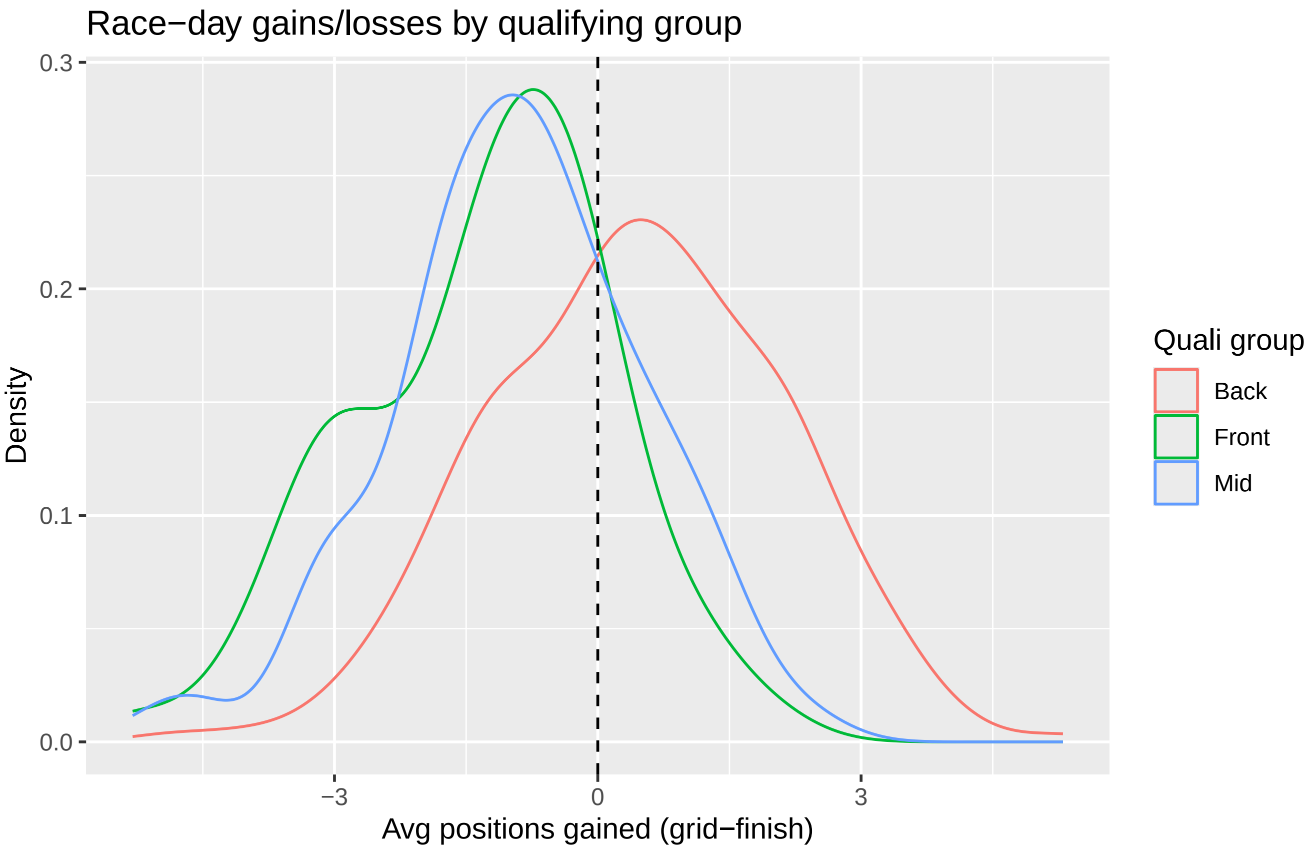


Interpretation

For the first scatterplot, each dot is one driver in one season. On the x-axis I plotted the driver's average qualifying position for that year, and on the y-axis I plotted their average finishing position. Both axes are reversed so that 1 (best) is at the top-right. I used the cluster labels from my hierarchical clustering to colour the points. The cloud of points lies roughly along a diagonal line: seasons where a driver usually qualifies near the front also tend to be seasons where they finish near the front, and seasons with poor qualifying are usually seasons with poor results. The different colours just break this overall pattern into groups. One cluster is mostly "front-running" seasons, where drivers qualify well and also finish well. Another cluster contains "backmarker" seasons, where drivers tend to start and finish near the back. The other clusters are more like midfield seasons. You can also see that some points sit a little above or below the main diagonal, which are seasons where drivers slightly under or over perform their usual qualifying position on race day.

Visualization 2

For the second visualization, I plot the density of average positions gained, defined as grid minus finish, for three qualifying groups: Front (average qualifying positions 1 to 5), Mid (6 to 10), and Back (11 or worse). Each curve shows the distribution of race-day gains or losses for that group on the same x axis, and I add a dashed vertical line at zero to mark no net change in position. This density plot is useful because it lets me compare how much drivers from different qualifying groups typically move through the field over a season, rather than just comparing single averages, and it complements the first plot by focusing specifically on raceday position changes.



Interpretation

For the second graph, I wanted to look more directly at how much drivers move up or down in races depending on how they usually qualify. I first created a simple variable that puts each driver-season into one of three qualifying groups based on their average qualifying position: “Front” (positions 1 to 5), “Mid” (6 to 10), and “Back” (11+). Then I plotted the density of average positions gained (grid - finish) for each group on the same x axis, and added a dashed vertical line at 0. Values to the right of 0 mean the driver tends to gain places; values to the left mean they usually lose places. The curves show that ‘back of the grid’ drivers are generally shifted to the right of 0, so they tend to make up a couple of spots during the race. The front and mid groups are more centred around 0 or slightly to the left, which means that when you already start near the front, you do not usually gain a lot of positions and sometimes lose a bit instead. Together, these two graphs show that qualifying is very strongly related to where you finish overall, and that drivers who start further back have more room to gain spots, but they still usually do not end up matching the results of the consistently strong qualifiers

Conclusion

Taken together, my two graphs show that qualifying performance is very strongly connected to final race outcomes, but not in a completely deterministic way. The clustered scatterplot demonstrates a clear upward relationship between average qualifying position and average finishing position: seasons where drivers usually start near the front also tend to be seasons where they finish near the front, and seasons with poor qualifying are usually seasons with poor results. The clusters highlight distinct groups of seasons that match F1 intuition: front-running seasons, midfield seasons, and backmarker seasons.

The density plot adds more nuance by looking at how many positions drivers typically gain or lose on race day. Drivers who usually qualify at the back tend to gain a few places on average, while front and mid-field qualifiers are centred around zero or slightly negative changes, meaning they mostly hold position or lose a little. This supports the idea that qualifying matters a lot: starting ahead gives a strong baseline advantage, and even though drivers starting further back can move forward, they rarely ‘erase’ the gap to consistently strong qualifiers over a full season.

Further Research

Further research on this question could look at qualifying and race outcomes with more formal models rather than just summaries and plots. For example, we could fit a regression model with finishing position or points as the response and grid position as a predictor, while also controlling for team, circuit, and season to separate the effect of qualifying from car and track differences. Another extension would be to move from season averages to race by race data, to see how the qualifying-finish relationship changes across different circuits or eras. These ideas would likely require more detailed modelling and additional variables than we used in this project, so we treat them as directions for future work rather than part of the current analysis.

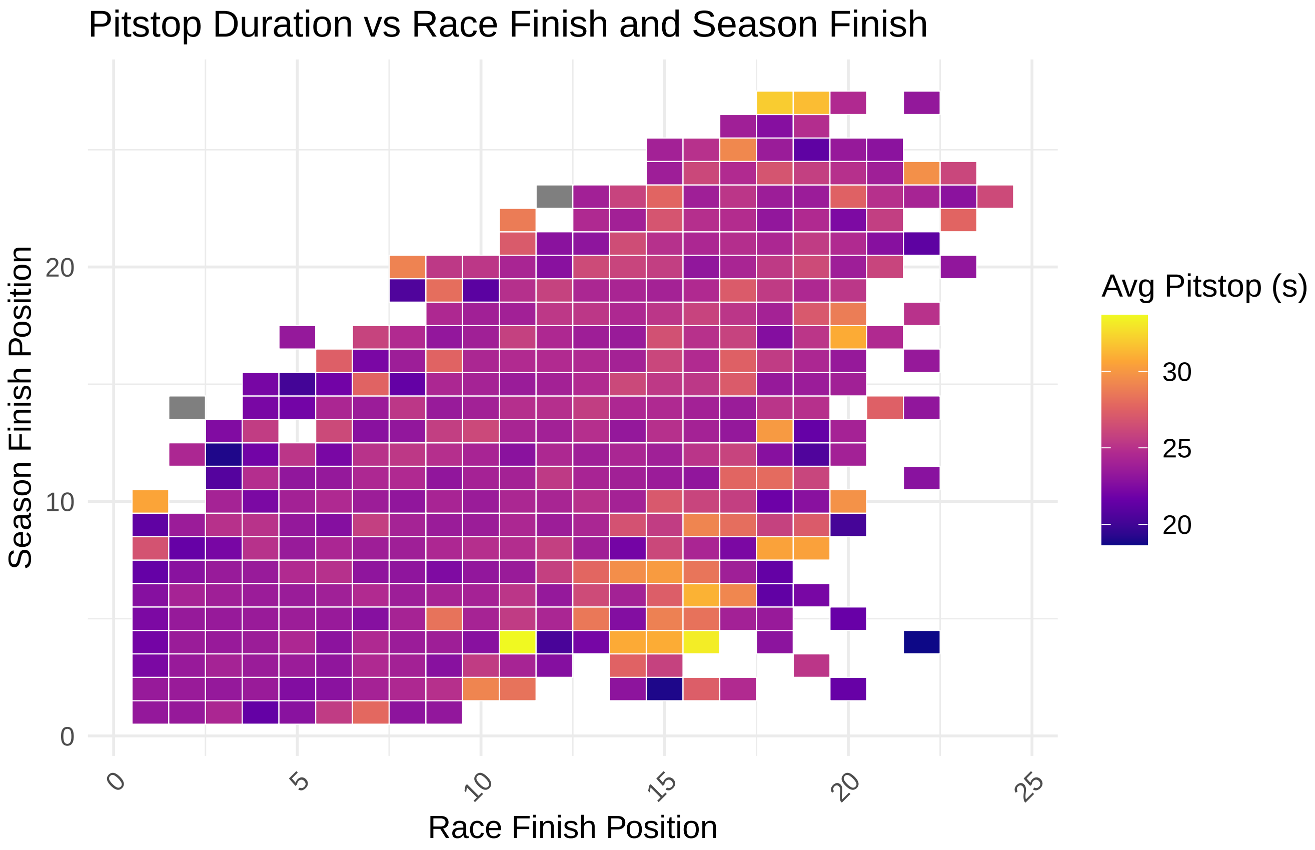
Research Question 4: How Do Pit Stop Durations Influence Race Finishes and Season Rankings?

Motivation

In a sport where everything is down to extreme precisions, one of the sources of error that cannheot fully be accounted for is the influence of pitstops. No matter how much a pit crew practices, errors are bound to happen and can cause race-changing events. The main goal of this question is to see if long pit-stops can end up influencing the outcomes of not only the race, but the season’s outcome.

Visualization 1

This is a heatmap that looks at the results that drivers got in individual races and in the seasons. The y axis represents the driver’s season finish while the x axis examines the driver’s race finish. The colors indicate the average pitstop times for each of the drivers final finish. This visualization is more of a summary visualization that gives more of the broader picture overall rather than specific details.

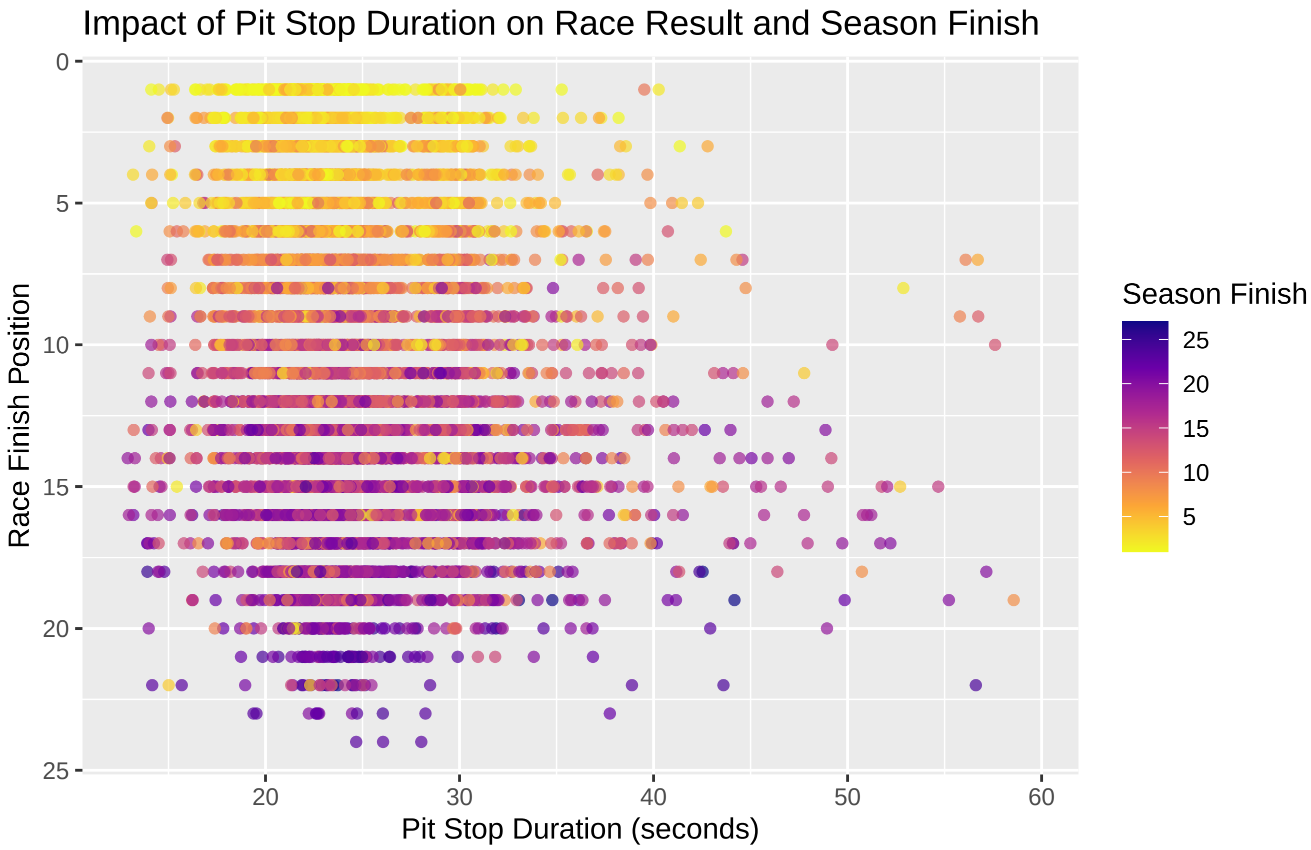


Interpretation

The main takeaway from this graph is that there is a relationship between pitstop times and race finishes. Additionally, there seems to be somewhat of a weak correlation between the overall season finish and the race finish position as most of the occurences of the lowest pitstop time averages typically occured for teams that finished 10th place or above. Additionally, it is notable that average pitstop times are relatively consistent as shown by the fact that most of the field is within the same volet to purple color range, indicating a sub 25 second pitstop.

Visualization 2

This is a scatter plot that looks at the same variables as the previous graph. However, since this is a scatterplot, the details are a lot more specific. The x axis represents that pitstop dxuration, the y axis represents the overall season finish of the driver, and the color of the point indicates how the driver finished that season.



Interpretation

Overall, we can gain similar insights from this visualizaation. However, the main difference is that it is more specific, given that specific data points of the pitstop times are graphed rather than being binned. From this graph, we can see a more definitive positive relationship between race finish and pitstop time. This is to be expected. However, by comparison, we can see that the influence of pitstop times on the overall season finish of a driver is a lot weaker. For the most part, the pitstop times are spread relatively equally when looking at it from the perspective of the y - axis while there is an obvious vertial spread that correlates race finish and pit stop time.

Conclusion

Overall, while there is definitely a relationship between pitstop times and race finish, there is little to no relationship between pitstop times and season finishes. The scatterplot gives a better view on the relationship between pitstop times and race finishes while the heat map gives a better idea of the potential relationship between pitstop times and season finishes. However, since the relationship is so weak, it suggests that is typically not likely for pitstops to have a lasting impact on the rest of the overall season as other factors seem to outweigh performance hindrances that come from slow pitstops when looking at the season as a whole.

Further Research

Something that could potentially help answer this question in the future is looking at the relationship between pitstop times and season finish on constructor by constructor basis rather than a driver by driver basis. Additionally, since so much of performance is associated by the money and budget of teams, including pitstop performances, a possible future inquiry could be looking at the relationship between a team's budget and their continuous pitstop performance. This could explain whether the weak relationship between pitstop times and overall season finish is more indicative of a team's budget being reflected through their pitstop times or if it is actually the root cause of the influence.

Wrapping up

To recap, the following were our research questions:

1. How have race pace and reliability changed over time across eras?
2. How do drivers' lap time change over the course of a race?
3. Howe does qualifying performance affect final race outcomes?
4. How do pitstop duration influence race finishes and season rankings?

We found that while there has been an overall increase in reliability over time in F1, the newer, smaller engines lack in race pace compared to older era engines that were larger.

Secondly, drivers' lap times in Race 1 are broadly similar and change in predictable patterns, with brief spikes likely due to pit stops or incidents, and no statistically significant differences in mean lap times across drivers.

Additionally, there is a strong relationship between qualifying and final race finish. Not only is there is a clear relationship between how drivers end up finishing based on their qualifying position, but the general trend is that backmarkers typically overtake a few positions, but the vast majority of leading and midfield cars gain little to no position, some losing a position or two.

Finally, there is a clear relationship between pitstop times and final race finishes. However, whether there is a relationship between pitstop times and overall season finish is less clear.

Future Research

Future research across these projects could benefit from expanding the scope of the data, incorporating additional variables, and applying more advanced modeling techniques. While our analyses provide useful initial insight, each is limited by narrow datasets—such as focusing only on DNF counts and fastest laps for reliability and race pace, examining only Race 1, relying on winner-only pace data, or evaluating pit stops at the driver rather than constructor level. Future work could incorporate richer contextual factors such as weather, tire strategy, circuit characteristics, detailed DNF causes, and team spending.Additional inquiry into team-level factors, including budgets and resource allocation, could also clarify whether observed patterns in pit stop performance and season results stem from operational execution or underlying financial disparities.