Formula 1 Driver Style Analysis and Optimal Driver-Car Pairing Predictions Using Telemetry Data

Alison Menezes

Clemson University

CPSC 4300 Applied Data Science

Final Report

Professor Tim Ransom

April 20, 2025

Area 1: Introduction

For my course final project, I will apply the techniques I've learned in this class to address one of the key challenges in Formula 1 – identifying the optimal driver-car pairing. Formula 1 (F1) is the highest class for international racing in single-seater, open-wheel cars. The sport involves twenty teams that compete in a series of races, called Grand Prix (GPs), in countries all over the world. Each team designs and builds its own cars, and has two drivers that compete to win the Constructors' Championship (awarded to the team with the most points) and the Driver's Championship (awarded to the driver with the most points). Every team's car is different, and every driver has a different driving style, making the pairing between driver and team car all the more important. What's especially interesting about this sport is that drivers sign contracts independently and can switch teams between seasons. Drivers can also switch teams or be swapped mid-season, which is why it's so important for teams to be able to assess drivers' performance metrics (and how they perform in their car). A skilled driver paired with a car that matches their driving style is an invaluable combination that can significantly impact a team's performance in the F1 season, but a mismatch in pairing can cost the team tens of millions of dollars in lost points and sponsorships.

This leads to the first question – *Can drivers be categorized into a racing "style" based on their previous racing stats?* Grouping drivers into these profiles can allow teams to track a driver's performance with different team cars and overall determine the drivers whose profile has the best performance with their car. My personal analysis due to my familiarity with the sport shows three general style profiles: cautious, efficient, and aggressive drivers. For example, rookies tend to be more cautious, taking turns slower and being overall more hesitant compared to experienced drivers. In contrast, some of the more experienced drivers fly through corners and

push the car hard, being noticeably more aggressive. Then there are drivers who fall in the middle, experienced and certainly not cautious, but also not pushing the limits with their car.

The second question was inspired by events of the 2025 season. Red Bull, one of the top F1 teams, has had a rather unstable car in the past couple of years. Their 1st seat driver, Max Verstappen, is a talented driver whose aggressive style pairs well with the unstable car; he won the past 4 Driver's Championships with it. However, Red Bull's 2nd seat is the one that's been giving them trouble. The 2024 season was marked with issues, as the 2nd seat driver at the time, Sergio Perez, was having significant difficulties with the car. Perez was replaced by Liam Lawson for the 2025 season; however Lawson also had tremendous difficulty with the car and after only 2 GPs (in both of which he failed to score any points), was swapped with driver Yuki Tsunoda of Red Bull's junior F1 team, Visa Cash App Racing Bulls (a mouthful, which is why many refer to them as Sugar-Free Red Bull, or VCARB). There is also lots of debate around whether Red Bull should even be allowed to have a junior team also competing in F1, but that's a separate discussion.

Anyways, many refer to Red Bull's 2nd seat as being "cursed", and while the car is definitely not optimal to drive, I believe that data science techniques can be used to analyze the styles of each driver, predict which ones would perform well in that seat, and prove that it's not a curse. Therefore the second question is: *Based on a driver's style and telemetry data, can we determine which driver (from the 2025 season) would perform the best in Red Bull's 2nd seat?*. The drivers pool is being limited to the current 20 drivers to make the dataset manageable.

Both of these questions are linked, and there are several stakeholders that would be interested in the results found to answer the questions; for example, the drivers and F1 teams that seek to find the best drivers for their cars that can win the championships. F1 teams have

countless employees that strategize and work to pick the best drivers that fit with their cars, so this analysis would be interesting to all of them. The fans are another group that would be interested in my analysis. Fans want to see their teams perform the best, and everyone has an opinion on the team's strategy. Team sponsors and stakeholders would also be interested in this analysis because it might affect which team they think will perform the best and therefore which team they want to support.

The goal of the analysis of the first question is to create three groups of values for the features that represent three different styles of driving. Each group will be analyzed and labeled according to the values in it to represent what kind of driving it is, and each driver will be assigned to a group. Since each data row consists of a lap, and each driver will have hundreds of laps, majority values will be used to place each driver into a style profile rather than averages, since the data is gathered over several years and different tracks. The goal of the analysis of the second question is to compare the metrics of each driver to those of Max Verstappen, the current Red Bull 1st seat driver, and evaluate whose driving style is the most similar to Verstappen's. As Verstappen is very successful in the Red Bull car, it makes sense to evaluate potential performance of a driver in Red Bull's 2nd seat by comparing their driving style with Verstappen's.

Achieving this analysis will provide a way for Red Bull to make better informed decisions about which drivers are good candidates for their car and will lead them to victory in the championships. Not only can Red Bull use this analysis, but every F1 team can adapt the analysis to determine how to pick their drivers as well. Team sponsors and fans can look at the results and predict for themselves how they think teams will perform, and thus choose which

team they want to support. Drivers themselves can see predictions of which cars they would be able to best perform in, and which cars they may have to adjust a little to drive well.

The information in this section is taken directly from https://www.formula1.com/. More information about the sport can be found there.

Area 2: Methods

Datasets across the internet were evaluated to see whether they would yield a comprehensive analysis and support the goals of this project; however, no dataset seemed sufficient enough. Most datasets only contained high-level stats like finishing positions and total lap times. Then, I found FastF1, a Python library that wraps publicly available telemetry and timing data directly from Formula1.com. The key feature of this library is that it allows access to telemetry per lap, such as throttle usage, braking behavior, and speed metrics, which will yield a much more detailed result. The telemetry really builds a profile of how the driver interacts with the car. High-level details such as finishing positions and lap times don't really tell you anything about *how* the driver achieved their position.

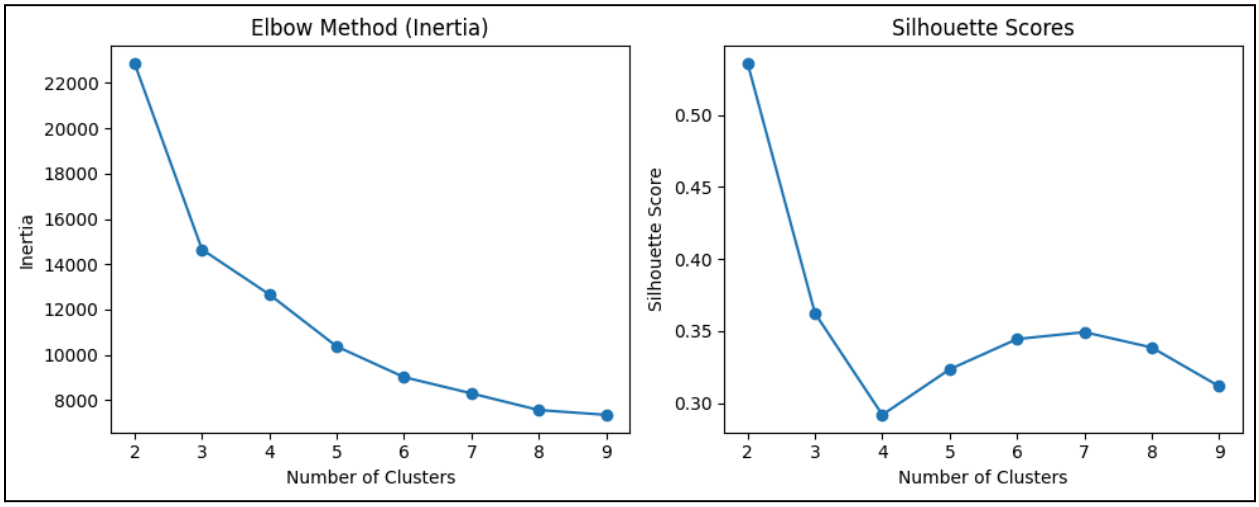
Pulling the data from FastF1 took a while because of the sheer amount of data that was being collected, therefore that code is in a separate notebook. The GitHub and documentation for the library was extremely useful in figuring out how to use the methods defined to access the data. Every aspect of the dataset was considered to ensure the most accurate and helpful result relevant to the questions being asked. To start, the analysis is being performed on the current 2025 lineup of drivers, and the majority of those drivers have been racing in F1 for the past ~5 years. For that reason I decided to use lap data from 2021 to 2025. I also had to account for the 5 rookies taking part this season as they have no F1 lap data. For the established drivers, their top 3

laps from 10 of the GPs per season were pulled, giving each established driver around 300 - 400 laps (drivers may not have lap times in certain races if they weren't competing at the time or crashed early). For the rookies, their data was added as the 2025 season progressed; when this project was first assigned, only 2 GPs had happened, and two of the rookies crashed immediately in both races, giving them no data. Currently, there is data from 4 of the 2025 GPs. Around 45 laps per GP was pulled for each of the rookies, and qualifying race data also supplemented their portion (qualifying races happen before the actual GP; however there are very few laps; only about 10 per driver compared to the ~50 laps per driver of a GP).

Each row in the dataset represents one lap. The dataset includes the driver, lap time (in seconds), mean and max speeds (in km/h), throttle percentage, braking percentage, and the number of braking events per lap. After building the dataset, I cleaned it by removing invalid or duplicated laps and checking for missing values. FastF1 pulls directly from Formula1.com, so the data comes straight from the source, but I double-checked to ensure that no anomalies slipped through. When building the dataset, if a lap had missing data, it was skipped entirely, which ensured no missing values. In total, the dataset contains 5920 laps, with between 150 to 400 laps per driver, depending on their availability and history. Generally, the rookies have around 150 - 200 laps each, and established drivers have 300 - 400 laps each.

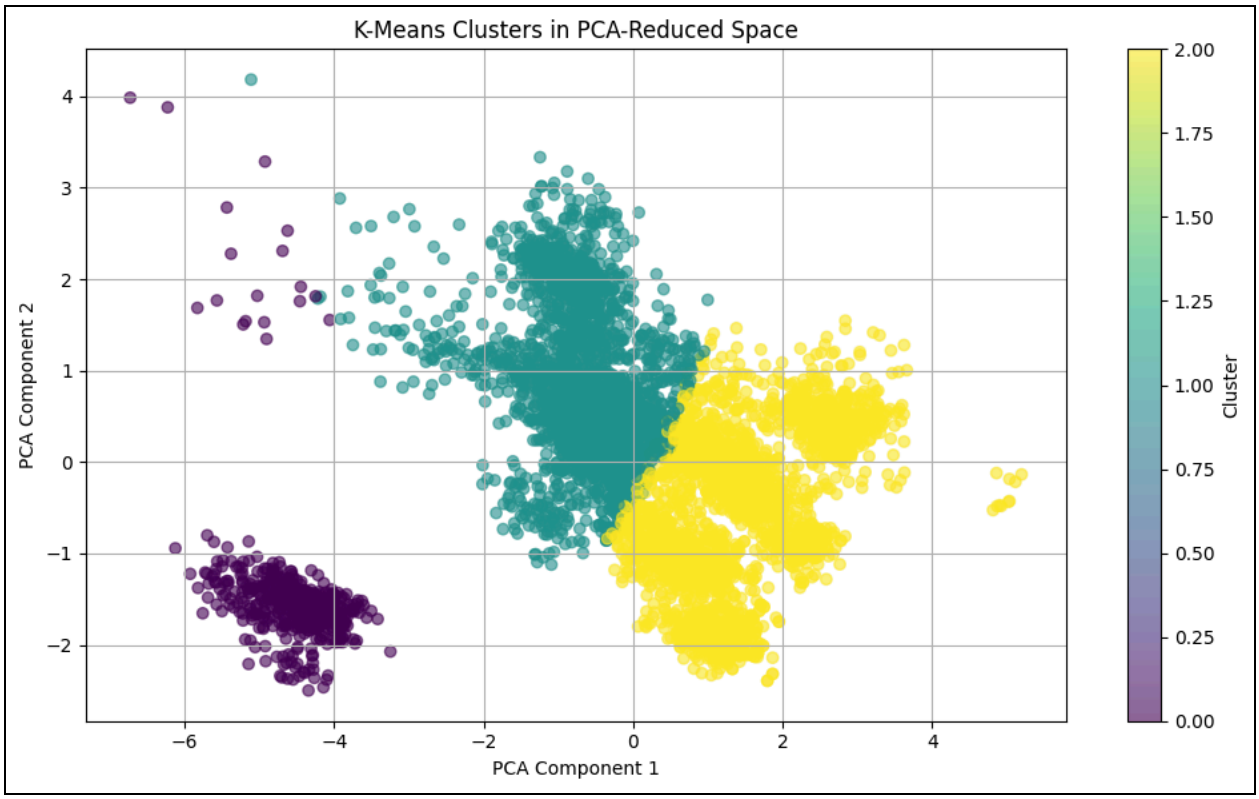
For question 1, since I'd like to cluster drivers based on multiple performance features, a clustering model makes the most sense. I considered both KMeans and DBSCAN, but decided to go with KMeans clustering for several reasons. I'm most familiar with this model, and based on my existing knowledge of the features I've chosen for prediction, there is generally some correlation between them, so I expect the clusters to form more spherical shapes. I also already know that I want 3 clusters, one for each driving style (aggressive, efficient, and cautious), and

KMeans requires you to specify the number of clusters beforehand. Validation of the model was performed using inertia and silhouette scores.



The inertia shows the elbow around 3 clusters, indicating the data forms meaningful shapes with 3 clusters. Although the peak silhouette score is at 2 clusters (indicating how well-separated the clusters are), 3 clusters still has a reasonable score and is justified by our preference of 3 style clusters.

After the choice of model and clusters was validated, KMeans clustering was applied with a random state of 42 to the scaled telemetry features. Each lap was assigned to one of the three clusters. Since there are six telemetry features, PCA was applied to reduce the features from 6D to 2D in order to better visualize the clusters. Below is the graph of the clusters in PCA-reduced space:



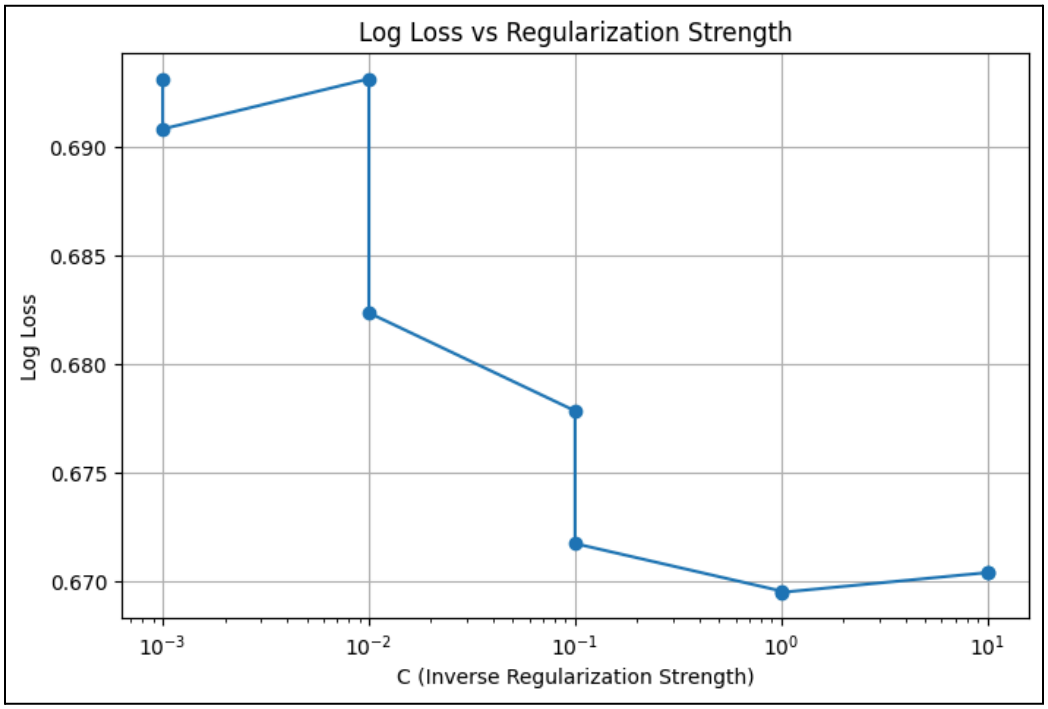
As can be seen from the above plot, the clusters are well-separated and form solid shapes. Now that the clusters are defined, the labels need to be assigned to them based on their values. To reinterpret the scaled clusters, I used the inverse-transform method on the centroids which converted them back to the original feature space so they would be interpretable. The centroid for cluster 0 had the fastest lap time (78.2 s), but the lowest mean speed (154.4 km/h), max speed (283.7 km/h), and throttle mean engagement (49%), and the higher brake percentage (28.6%) and braking events (13.1). Cluster 0 also had the most outliers. The combination of the throttle, speeds, and braking resulted in the quickest laps, suggesting that the drivers in these laps were very good at balancing the throttle with the brakes and weren't just constantly accelerating and then having to brake to get their speed under control at the corners. They got the most out of the car with the least waste of energy. Therefore, I judged this cluster to contain the "efficient" driving styles. The centroid for cluster 1 had the slowest lap time (98.8 s), moderately high

throttle (63.8%), and relatively high braking events (9.8) and percentage (20.3%). The combination of these telemetry features suggests that during these laps, drivers were fast but also tended to take corners cautiously (most braking events in F1 occur around corners and turns). This is actually a trend that can be seen with a lot of rookies since they are less familiar with the tracks; they take the straights very fast but are less experienced with the corners, so they tend to brake earlier and more often. Therefore, I labeled this cluster as “cautious”. The centroid for cluster 2 also had a very fast lap time (81.7 s), high throttle engagement (71.7%), and low braking events (7.5) and percentage (16%). This combination of features shows high-speed laps with minimal braking, indicating a very aggressive driving style; therefore this cluster was labeled “aggressive”. After these cluster labels were defined, each driver was assigned to a style based on which label the majority of their laps fell into.

For question 2, we want to predict which driver would perform the best in Red Bull’s 2nd seat using Max Verstappen’s driving style as the comparison. Deciding which model to use in this case took a bit more research; in the end, I elected to use logistic regression with probability estimates to rank drivers by their similarity to Max Verstappen’s driving style. Essentially, the model will display the probability that a lap belongs to Max Verstappen. Logistic regression is a well-understood technique that can handle technical/numerical features and will be able to output interpretable probabilities, which is better for comparing similarity of drivers. Rather than having a binary label such as “drives like Verstappen” or “doesn’t drive like Verstappen”, probabilities allow a spectrum of similarity that is more interpretable.

First, Verstappen’s laps were separated from the rest of the drivers. Because the dataset was initially built with approximately equal numbers of laps per driver but comparison to Verstappen implies that he should have more laps than the rest, I downsampled the rest of the

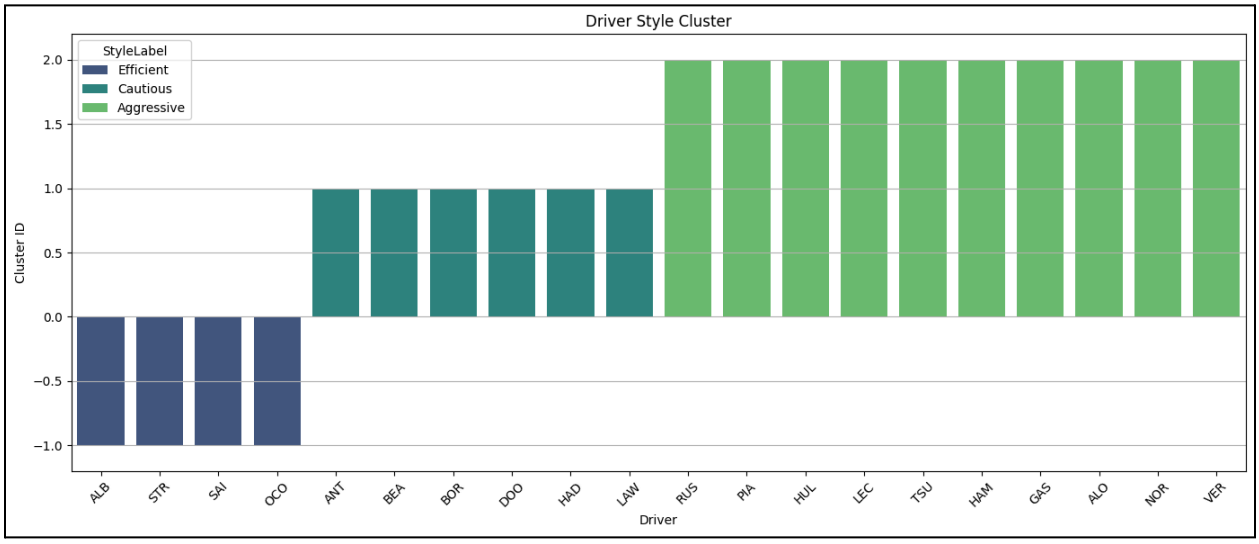
drivers' laps to balance the training dataset and ensure that the model wouldn't be biased towards the majority (non-Verstappen laps). Verstappen had 383 laps, and I randomly sampled from the non-Verstappen dataset to add 30 laps per other driver, resulting in a training dataset size of 953 (570 non-Verstappen and 383 Verstappen). Hyperparameter tuning was performed using GridSearchCV with 5-fold cross-validation and log loss as the scoring metric. I chose log loss as the scoring metric because it's traditionally used with probability-based analysis rather than mean squared error (what we typically used in class). The best-performing metrics were C=1, penalty=□l2', and solver=□liblinear', with a minimum log loss of 0.6.



Lastly, the trained model was applied to the rest of the laps from the original dataset to generate the similarity probabilities averaged by driver.

Area 3: Results

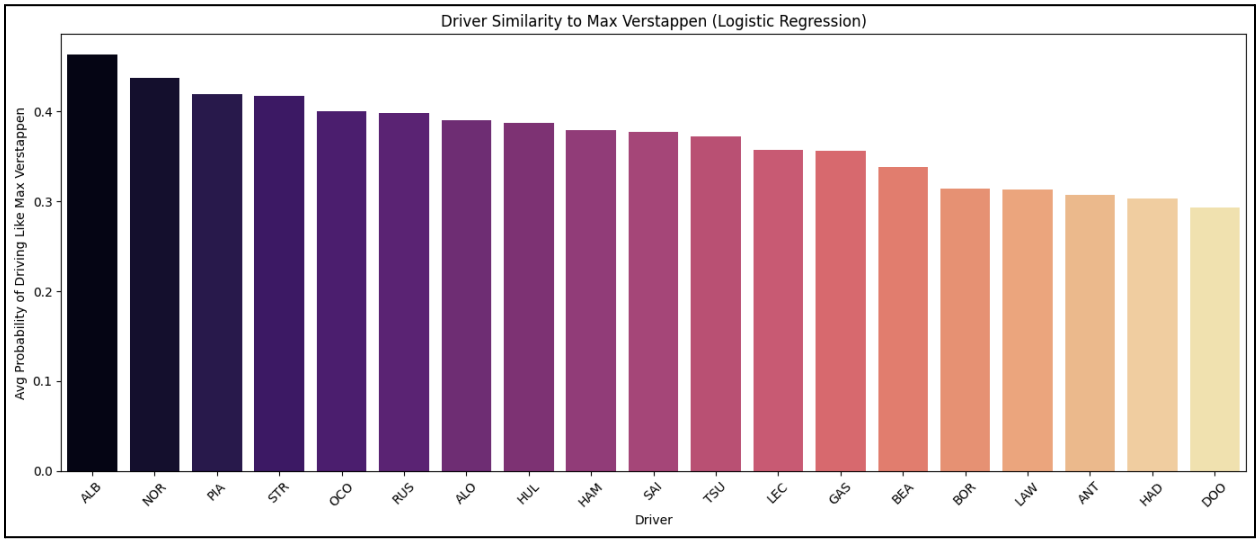
For question 1, I created a bar chart showing the dominant style of each driver. The KMeans model with 3 clusters had an inertia of 10376 and a silhouette score of 0.32, which shows relatively well-separated and compact clusters.



As we can see, all five rookies' dominant racing style is "cautious", and this can be seen by watching any GP and observing them in comparison to the rest of the grid. As they get more familiar with the tracks their style may change, but as of right now their stats fall squarely into the cautious category. Albon, Stroll, Sainz, and Ocon have a dominant efficient driving style, and from my familiarity with the sport that makes sense based on their performance this past season so far. Their team cars are not as fast as the top teams' cars, and overall more responsive to the driver. The rest of the drivers have an aggressive driving style, and all of these drivers are the more experienced / established ones on the grid, so this result is supported in reality.

For question 2, I created a bar chart ranking the overall similarity of each driver to Max Verstappen. The total training data size was 953, with 383 Verstappen laps and 570 non-Verstappen laps, and the trained model was evaluated on the rest of the laps from the

original dataset. The model had a minimum log loss of 0.67, meaning that it was pretty good at evaluating a lap's probability of belonging to Verstappen.



Surprisingly, Alex Albon has the most similar driving style to Max Verstappen, and when we look back at the style profiles, Albon actually has a rather efficient driving style compared to Verstappen's aggressive style. Interestingly enough, Albon was Verstappen's teammate at Red Bull during the 2019 and 2020 seasons; however that was before Red Bull upgraded their car to the more unstable version it has been for the past couple seasons. I'd love to see how Albon, who is now at Williams with a slower car, performs in the 2025 Red Bull car.

Area 4: Discussion & Conclusion

The results for the KMeans clustering are overall well-supported by my personal experience with the sport. The rookies and Lawson all having cautious driving styles can be seen in any race. Many people were surprised and skeptical about the decision to place Liam Lawson in the Red Bull 2nd seat for the 2025 season because of how unstable the car is and how relatively new he is to the sport compared to other drivers, and especially because they were choosing between him and Yuki Tsunoda. Many were proven right after Lawson failed to score

any points in the first two races and actually crashed during the first one, and Red Bull swapped him with Tsunoda, who has so far been performing well in the Red Bull. A keen observer during the first two GPs can spot that Lawson is not used to the car's issues, resulting in him being overly careful as he drives, ultimately slowing his pace down. Tsunoda definitely has a rougher style and adapts well to the car's instability.

There are a few things to note about the analysis; for one, F1 tracks vary in their distance, so the lap time ranges are all slightly different per track. I tried to include an equal number of laps from each race to account for the lap time variation, but it's not exact. If drivers crash early in the race, they don't have data. Also, the lineup changes every year; drivers may participate one year, leave, then come back another year. Some drivers were rookies last year so only have data for that past year. And the most important factor to consider – drivers switch teams very often! Lewis Hamilton drives for Ferrari this season, but he was at Mercedes for the past *12 years*, meaning most of his data comes from his performance in their car. Carlos Sainz has raced for Toro Rosso, Renault, McLaren, Ferrari, and Williams in his F1 career. All of those cars are completely different, and a driver's performance varies based on the car they're in.

However, that is the nature of the sport, so by not separating out those features and leaving them in for the model to account for, the analysis is more realistic. It would be extremely difficult to categorize every single driver's style by individually analysing their performance record based on each team they've been on, and would certainly take longer than I have for this final project. Also, it's possible for a driver to drive well in different car setups; Alex Albon is a good example of this. He has an efficient driving style according to the analysis, and that style works well on the Williams team where he's been for the past few seasons. However he also drove for Red Bull alongside Verstappen, and though the car wasn't as unstable back then as it is

now, Albon himself has said that the car is difficult to drive. Despite these comments he was able to manage the car, showing his adaptability to different car setups.

Overall, the KMeans clustering for driving styles is a good metric for stakeholders to evaluate and predict the performance of their drivers in different cars. The analysis gives a categorical label to what teams, fans, and sponsors observe during a race. The dataset I built is just one way to do the analysis; it can be done with any lineup of drivers, any range of years, and any amount of tracks depending on the question being asked. To refine my particular analysis, I would probably change the way I build the dataset. I would pick the one or two most recent years and use fewer tracks, but ensure that most if not all of the drivers have data for those tracks, and increase the number of laps I pull from each track. That way, I keep the volume of the dataset but the feature values are much more consistent across all the drivers. More lap data from fewer tracks over a short range of years would show a driver's style more consistently than fewer laps across a large timespan.

The results for the logistic regression with probability estimates are very interesting. I've heard several people remark that Albon would perform well in the Red Bull. It's kind of funny how all the rookies and Lawson are at the bottom of the ranking for similarity to Verstappen, but Lawson was still chosen over Tsunoda to fill the 2nd seat at the start of the season despite having less similarity to Verstappen. That decision was a controversial one, and this graph supports the outcome of that choice. Teams, fans, and sponsors can tailor this model to draw comparisons to the driver of their choice, and it may influence big decisions such as replacement drivers.

Looking back, I would make the same changes to the dataset as I outlined above, just to make the data more consistent. I would also add more Verstappen data. The same dataset was used for both analyses, but one analysis required equal amounts of data while the other needed

skewed amounts. Because of that I had to downsample some of the data to eliminate bias, but that also shrunk the amount of data I had to work with. In the future, two datasets can be used or Verstappen data can be appended to the dataset after the first analysis is complete. Because the two models are linked in terms of the results, there isn't one that I recommend over the other if a well-rounded analysis is preferred, but the logistic regression with probability estimates is the more useful of the two if trying to make driver-specific decisions.

This project demonstrates how data science can be used to draw meaningful insights from Formula 1 data. We were able to successfully cluster drivers into style profiles using telemetry data and quantify driver similarity to a distinct driver, which supported mismatches that actually happened in the F1 world while this project was being built. The results of these models can be used by drivers, teams, sponsors, and fans to make better-informed decisions about driver selection and team strategy. There are limitations due to the nature of the sport, such as driver turnover and track variability, but this general analysis works with the dynamic nature of the sport rather than overcomplicating it. The model can be fine-tuned even more by adjusting how the dataset is built and tailoring the dataset to the type of analysis being performed, which will increase the value of the results. The questions I focused on for this analysis are extremely relevant in the F1 world, and data science offers a modern way of answering them.

(a meme poking fun at Red Bull's 2nd seat drama)