Project 4 Analysis

Gak Roppongi

Steven Ohms

CSCI 347

5/7/2022

(Link for our presentation)

https://montana.techsmithrelay.com/638p

What problem are you trying to solve or help solve?

For some background information, the Mercedes F1 team has been dominant for the past 8 years. After the regulation changes in 2022, they did not do a good job. To analyze what is wrong with this team this season, we took the data from theOehnrly / fastf1. So the problem that we are trying to solve is exactly why the Mercedes team is performing poorly, and what factors are important to perform well.

Number of Instances in data?

For the number of instances in the timing data we used, the total number of instances is the number of laps in each session for each race. We analyzed three races and for each race there are 5 sessions. The five sessions are free practice 1, 2, 3, qualifying and the race. And we have this data for all 20 drivers. So we had 3\*5\*20 data sets. For each data set, the number of instances is the number of laps the driver drove. So the number of instances for the Bahrain race was 57,the Saudi Arabian race was 50, and the Australian race was 58. However, since the number of laps for free practice and qualifying differs for each driver, they are not the same for all.

Number of Attributes?

In our data set we have 19 attributes. We only used fastf1.api.timing\_data api. Since each of these dataset stores data in an array and has a length of 2 and we only used the first data (which we called timing\_data\_0), we notify there are 19 data attributes in the dataset.

Missing values?

In our data set we do have missing values. We filled the values via mean filling.

Number of Categorical and Numerical Attributes?

In our data we have 2 categorical attributes including a numerical representation of the associated driver number, and a boolean of whether or not the lap was a personal best for the driver. We also have 17 numerical attributes in our data.

What pre-processing techniques were used?

The only preprocessing technique we used was mean filling missing values. We determined that using a mean filling method made the most sense for our data., because eliminating attributes or instances that had missing values would hurt the overall results much more. Aside from this no other preprocessing techniques were necessary. The categorical data of the driver was already associated with a specific driver number, and the personal best was already in the form of a boolean, so no actions were needed.

What data mining techniques did you apply and why?

Data mining techniques we used include DBSCAN, PCA, data visualization, and other statistical analysis. We used DBSCAN to try and find clusters in our data set, to better understand the tiers of racers that were forming. We decided to use DBSCAN over k-means because we did not know the target number of clusters we were searching for. And we also picked up the epsilon and min\_samples values for DBSCAN so that it can get maximal numbers of clulsters. We used PCA on the speed data for 4 speed attributes for each car for each race to determine if we can treat each speed as the same attribute. We implemented data visualization mainly through Scikit Learn in order to visualize the data so we can make sense of it. Finally we used statistical analysis to interpret the data and manipulate the data to find what we are looking for. This includes correlation, standard deviation, variance, covariance, and other statistical methods.

Relevant visualizations and tables summarizing data and findings?

Each color represents a different team or engine. When PU is used, it means we are referring to the team's engine. PU stands for power unit, and represents different engines used. Each color represents the primary color associated with each team

Mercedes: —-----------Teal/green

Ferrari: —--------------Red

Alpine: —--------------Light Blue

Red Bull Racing: —----Dark Blue

Haas: —----------------White

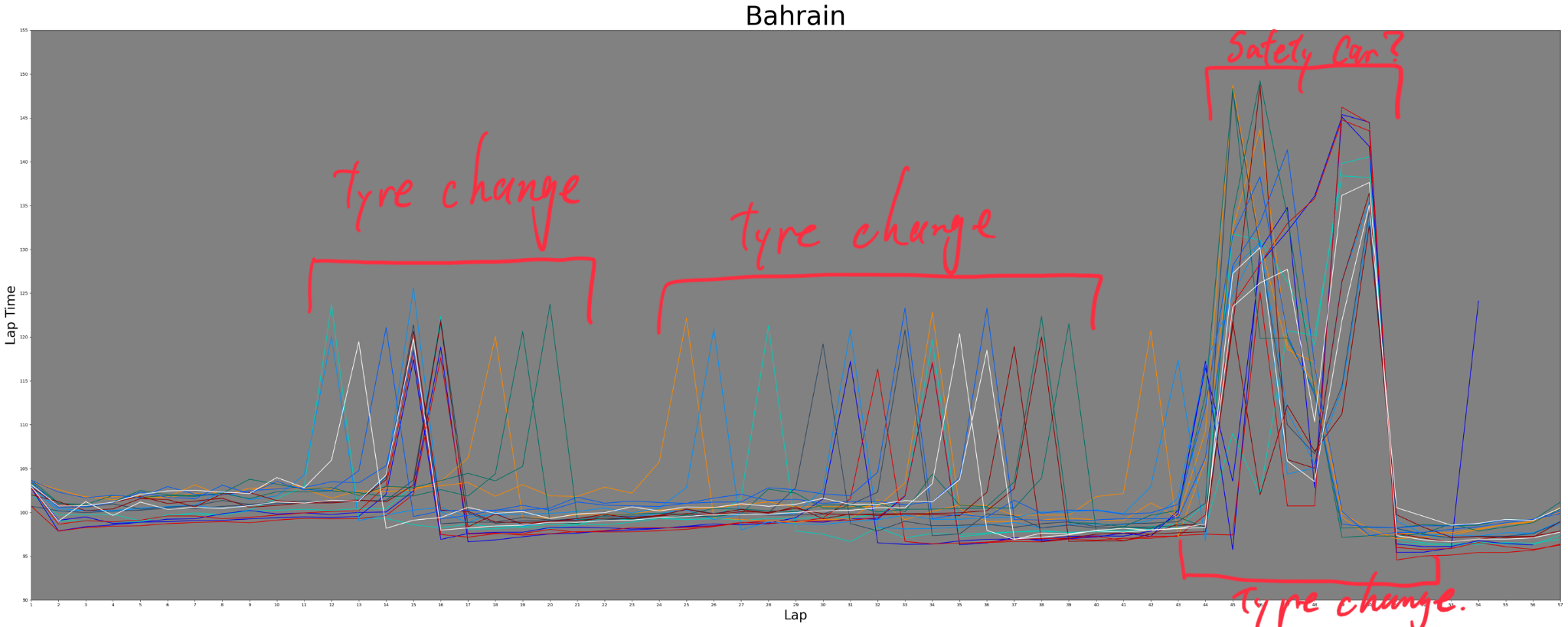
Aston Martin: —-------Dark Blue/Green

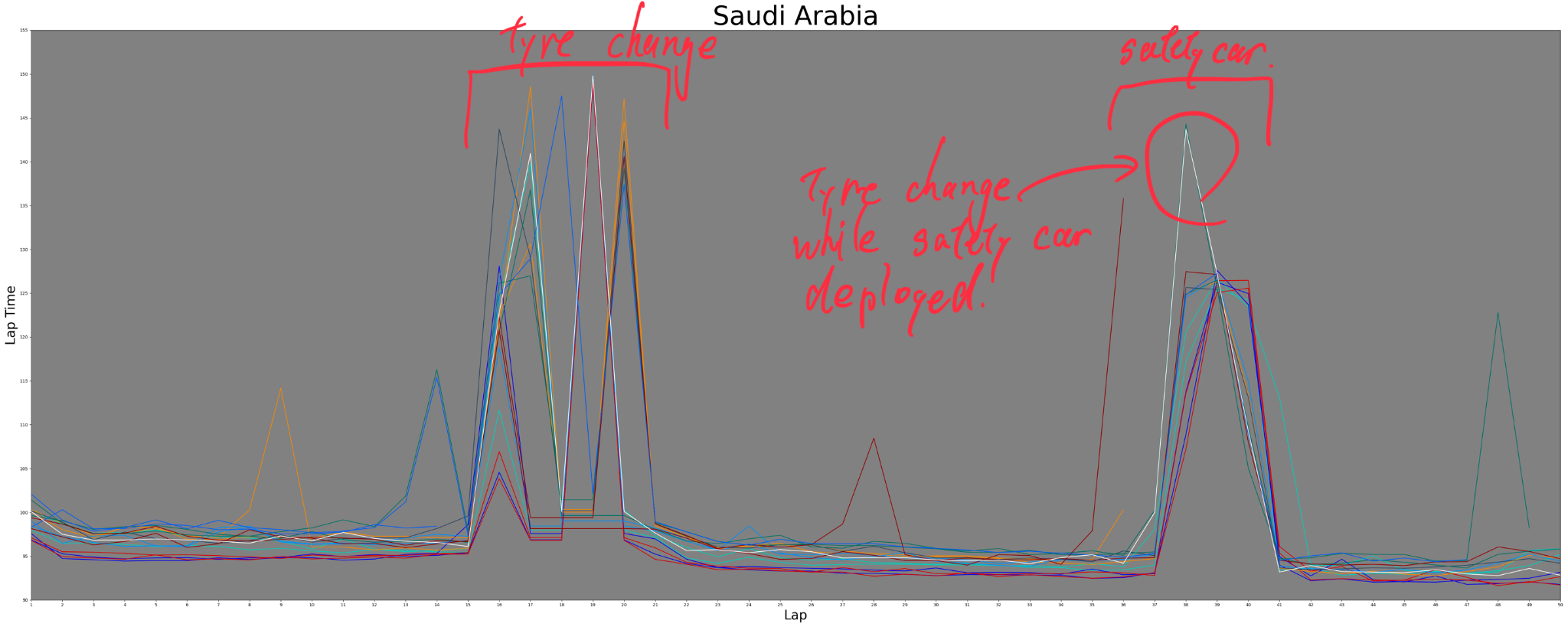
Alpha Tauri: —---------Navy Blue

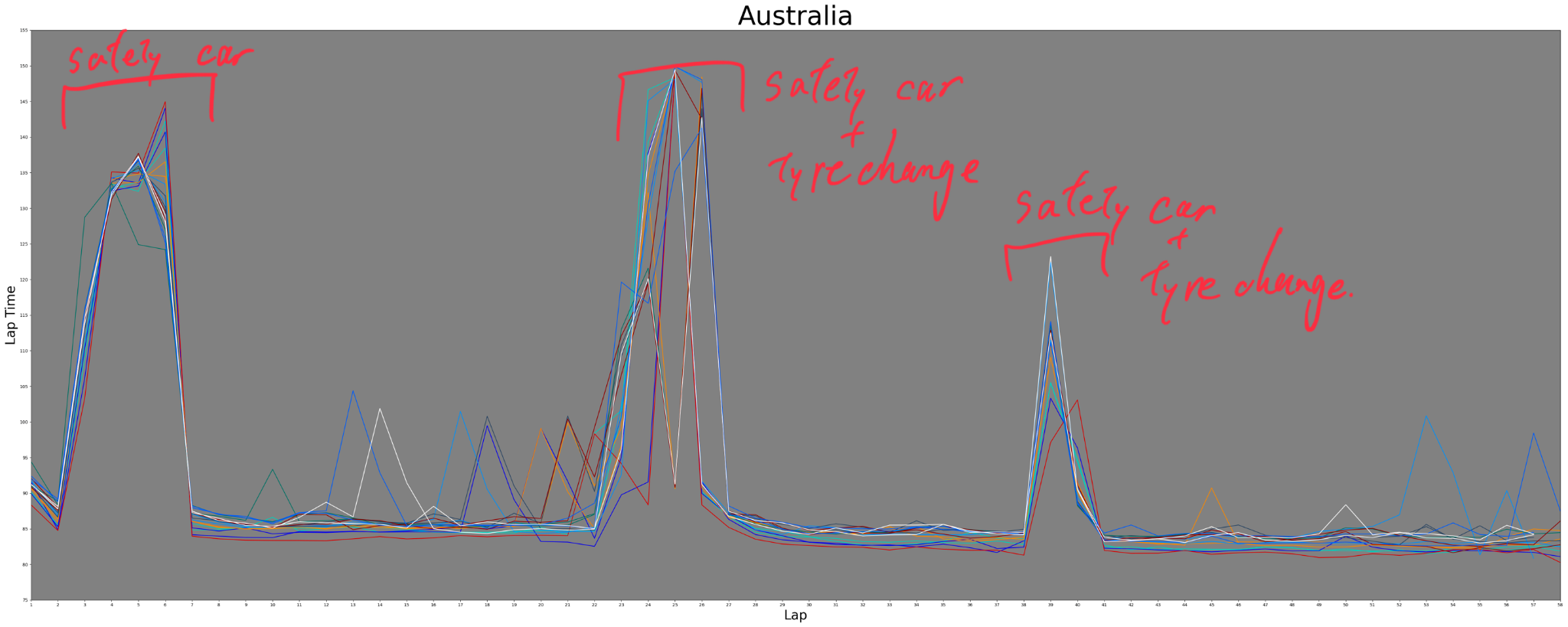
McLaren: —------------Orange

Alfa Romeo Racing: —-Maroon

Williams: —------------Blue

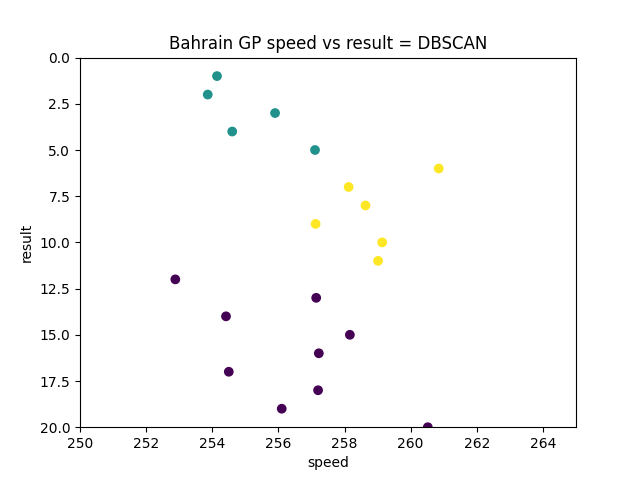
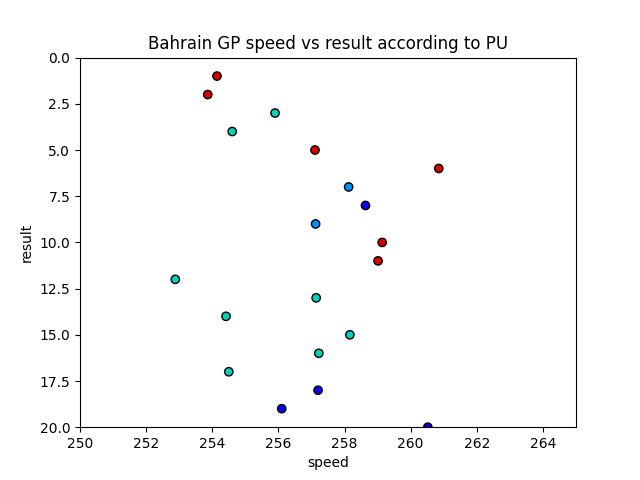




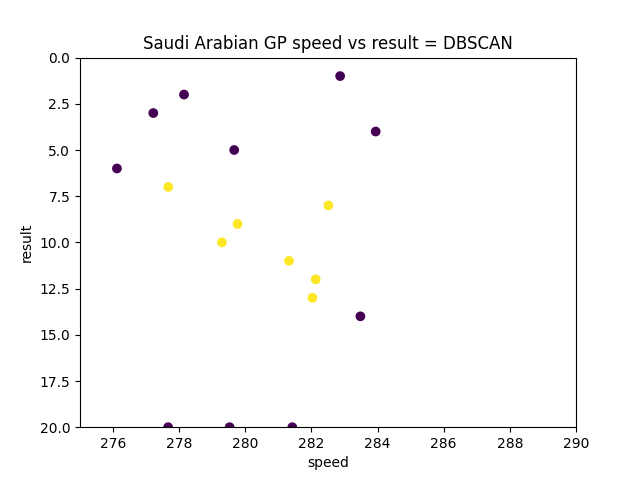
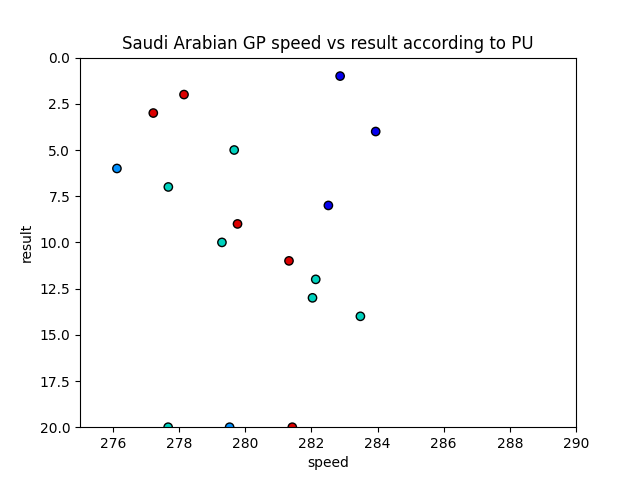


(plot shows different race events)

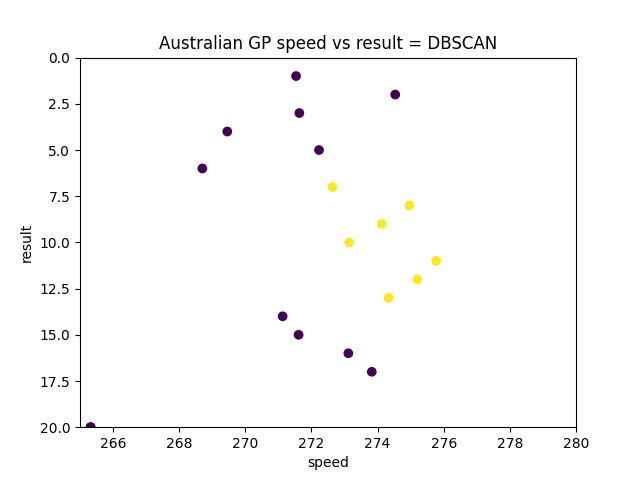
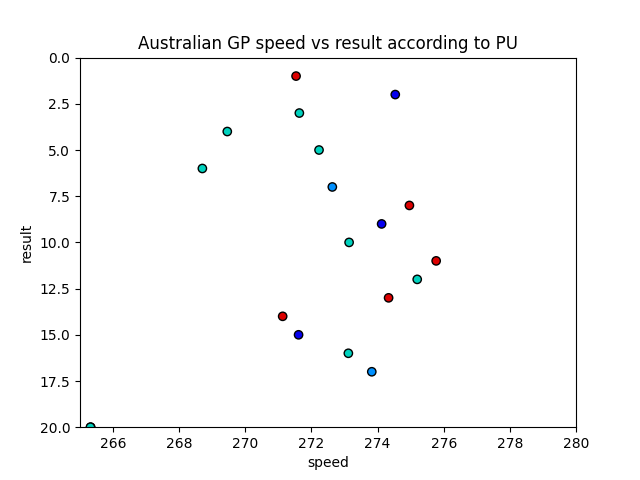
The tyre changing time is different for each team. For each race, they have 3 different tyres with different grip levels and durability. Tyre strategy is often one of the most important keys in the race. When looking at the lap time plot, the tyre change time is not one specific lap in a race, meaning that some teams are on a different strategy. Therefore , we hypothesized that the tyre strategy should be key for the performance.



We used the DBSCAN for the plot analysis because the number of clusters was hard to detect. The DBSCAN was done with eps = 3, min\_samples = 5. The green and the yellow points are the two clusters and the purple points are the noise. There are only two Mercedes cars in the green cluster with the good result, and everyone else is in the noise cluster with the bad result. This result might not be accurate in terms of the speed and result correlation because at the end of the race, two red bull cars retired because of engine trouble. Their mean speeds were very good and running in the top 3 for a long time in the race, but they resisted. So the clustering data must have been different if they finished the race, and the Mercedes car performed more poorly.

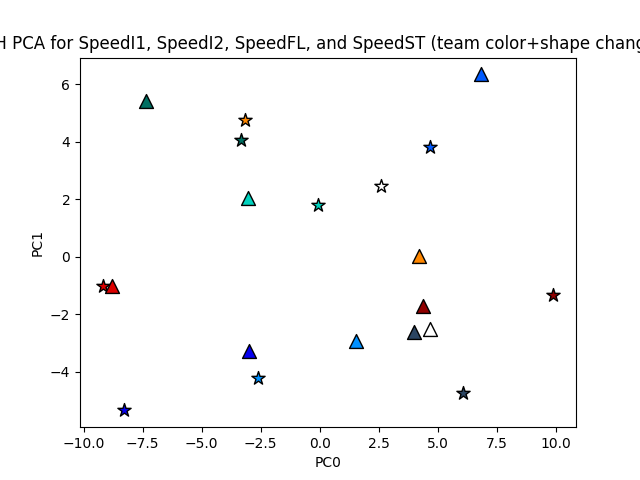


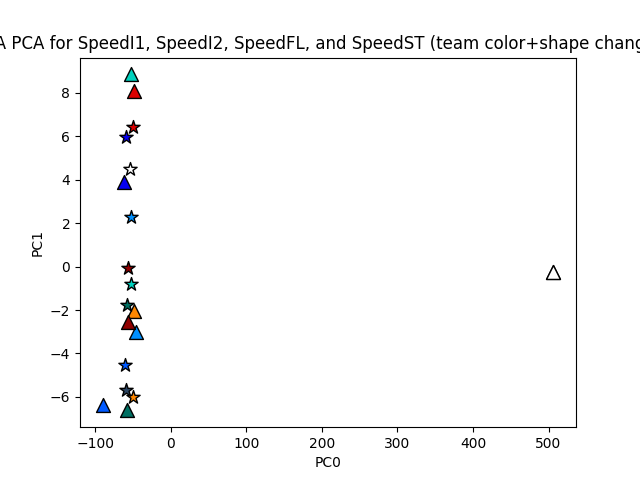
We used the DBSCAN for the plot analysis because the number of clusters was hard to detect. The DBSCAN was done with eps = 3, min\_samples = 5. We only see one cluster in this data in the middle, which contains 4 out of 8 Mercedes powered cars. This time red bull power trains cars are performing well. However, we clearly see that Mercedes cars did not do a great job.

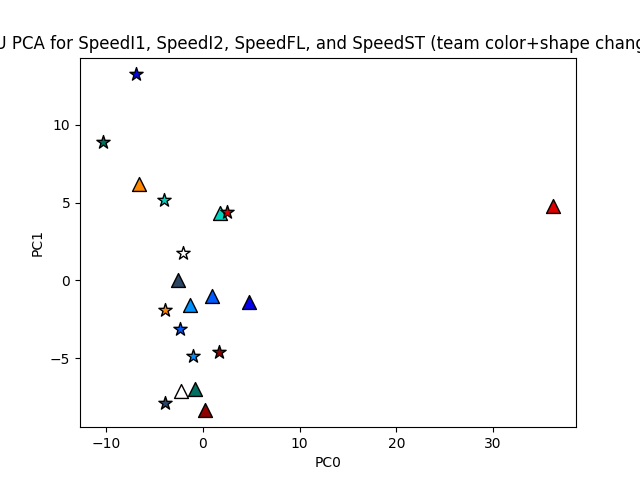


We used the DBSCAN for the plot analysis because the number of clusters was hard to detect. The DBSCAN was done with eps = 3, min\_samples = 5. This time, there is only one cluster the same as the previous race. However, the Mercedes powered cars are performing better than the last time. We see only 3 cars in the cluster and the noise with worse results. The noise with better results contains 4 Mercedes powered cars.

We used the DBSCAN for the plot analysis because the number of clusters was hard to detect. The DBSCAN was done with eps = 3, min\_samples = 5.







The values below the plot show the ratio of variance caught by each principal component:

BH: '0.65062', '0.25733', '0.05225', '0.03980'

SA: '0.99862', '0.00080', '0.00047', '0.00011'

AU: '0.63515', '0.26515', '0.06927', '0.03044'

Next, we got the PCA of the speed data for 4 speed datas for each car for each race since we wanted to know if we can treat each speeds as the same attribute. The PCA plots are Bahrain, Saudi Arabia and Australia from top to bottom. We plotted the speed in Km/h for each sector and the longest straight and performed the PCA. The PCA for each did not catch the variance well. However, since there is an outlier, the PCA must catch the variance a little bit better. There two cars retired early in the race before they set a single speed data for the race.

On the PCA graph, the red triangle on the right in the middle was retired early in the AU race as well as the white triangle in the SA race.

| BH stdev | SpeedI1 | SpeedI2 | SpeedFL | SpeedST | SpeedAll |
| --- | --- | --- | --- | --- | --- |
| Stdev | 2.5874839893194794 | 3.2742544193591385 | 2.2971933394878716 | 4.986885268512842 | 2.1947984548258725 |

| BH correlation | SpeedI1 | SpeedI2 | SpeedFL | SpeedST | SpeedAll |
| --- | --- | --- | --- | --- | --- |
| SpeedI1 |  | 0.3178591101497587 | 0.3581615906666482 | 0.5634241230685136 | 0.8270387480557837 |
| SpeedI2 |  |  | -0.33706051645778073 | -0.20553422804617866 | 0.26169156731106924 |
| SpeedFL |  |  |  | 0.7468207710937579 | 0.6657351981521912 |
| SpeedST |  |  |  |  | 0.8528523993084601 |
| SpeedAll |  |  |  |  |  |

| SA stdev | SpeedI1 | SpeedI2 | SpeedFL | SpeedST | SpeedAll |
| --- | --- | --- | --- | --- | --- |
| Stdev | 80.31385462512482 | 85.80668523742499 | 85.69575834536069 | 86.40805077406814 | 84.49703882930763 |

| SA correlation | SpeedI1 | SpeedI2 | SpeedFL | SpeedST | SpeedAll |
| --- | --- | --- | --- | --- | --- |
| SpeedI1 |  | 0.9982051092614944 | 0.9975413014097588 | 0.9968072363963087 | 0.9988035069474032 |
| SpeedI2 |  |  | 0.9984447924913998 | 0.999310265369064 | 0.9997016633886415 |
| SpeedFL |  |  |  | 0.9983983166713888 | 0.999310273368223 |
| SpeedST |  |  |  |  | 0.9993589396740359 |
| SpeedAll |  |  |  |  |  |

| AU stdev | SpeedI1 | SpeedI2 | SpeedFL | SpeedST | SpeedAll |
| --- | --- | --- | --- | --- | --- |
| Stdev | 4.873625340943086 | 7.564821931840256 | 3.28423154873576 | 5.976069604633701 | 3.5282589494922965 |

| AU correlation | SpeedI1 | SpeedI2 | SpeedFL | SpeedST | SpeedAll |
| --- | --- | --- | --- | --- | --- |
| SpeedI1 |  | 0.7326531562575563 | 0.512659260128899 | -0.18501189917710412 | 0.7790006269319643 |
| SpeedI2 |  |  | 0.644123310937835 | -0.30087720130076384 | 0.8115111508173503 |
| SpeedFL |  |  |  | 0.18444274081660197 | 0.8331064758187957 |
| SpeedST |  |  |  |  | 0.24119978084746493 |
| SpeedAll |  |  |  |  |  |

| BH | SpeedI1 | SpeedI2 | SpeedFL | SpeedST |
| --- | --- | --- | --- | --- |
| result | -0.02462007387254157 | 0.07673552208775115 | 0.4440338236522099 | 0.13401598767627507 |

| SA | SpeedI1 | SpeedI2 | SpeedFL | SpeedST |
| --- | --- | --- | --- | --- |
| result | -0.4451766229464749 | -0.41473200118580467 | -0.4164523384431444 | -0.40886518597009835 |

| AU | SpeedI1 | SpeedI2 | SpeedFL | SpeedST |
| --- | --- | --- | --- | --- |
| result | -0.31547391537362535 | 0.25738965383349166 | 0.13251521495766067 | -0.1976635677825166 |

| BH | SpeedI1 | SpeedI2 | SpeedFL | SpeedST |
| --- | --- | --- | --- | --- |
| Result without retire | -0.17733279932430016 | -0.012743617914929038 | 0.4017459091928089 | 0.044538886980315995 |

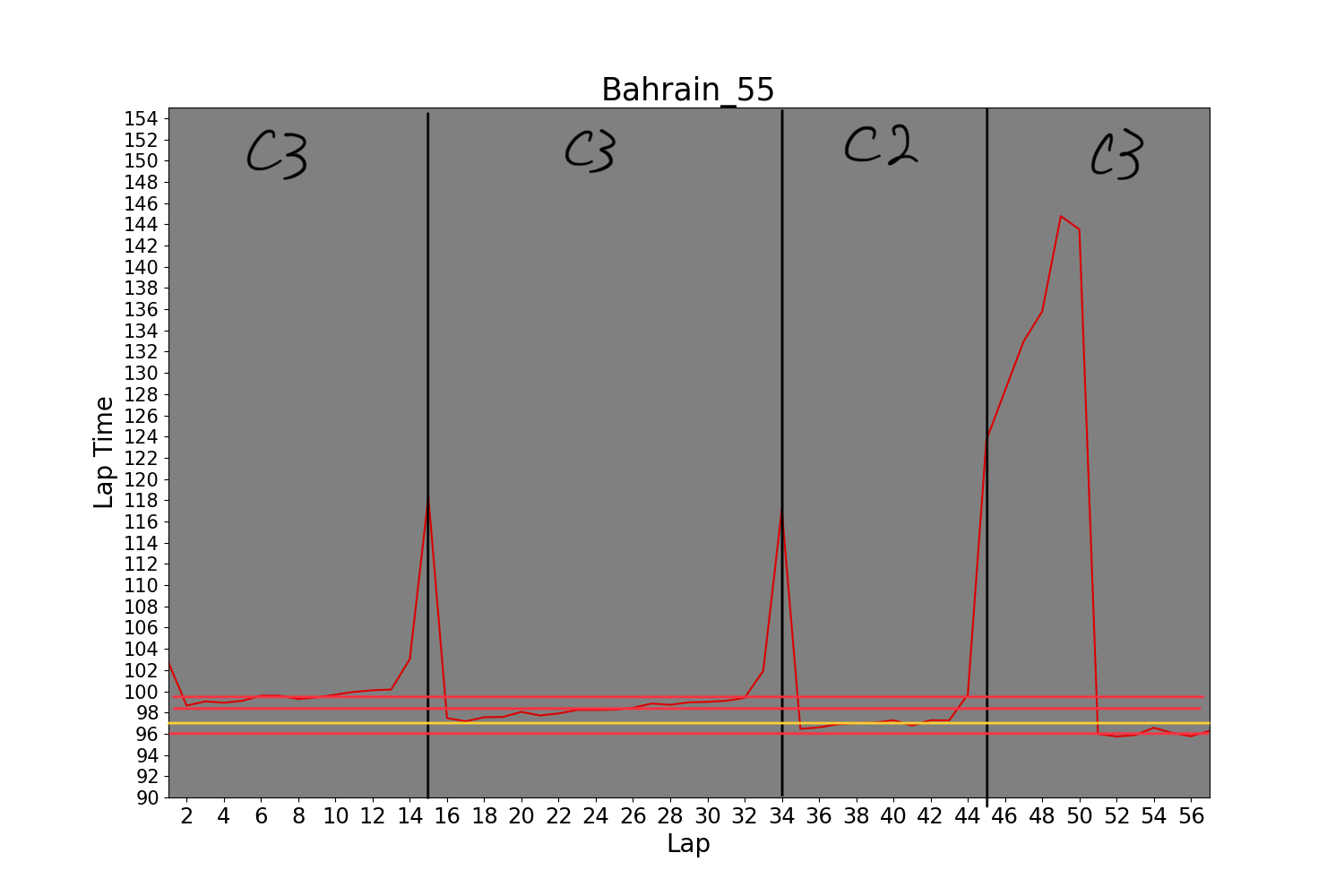
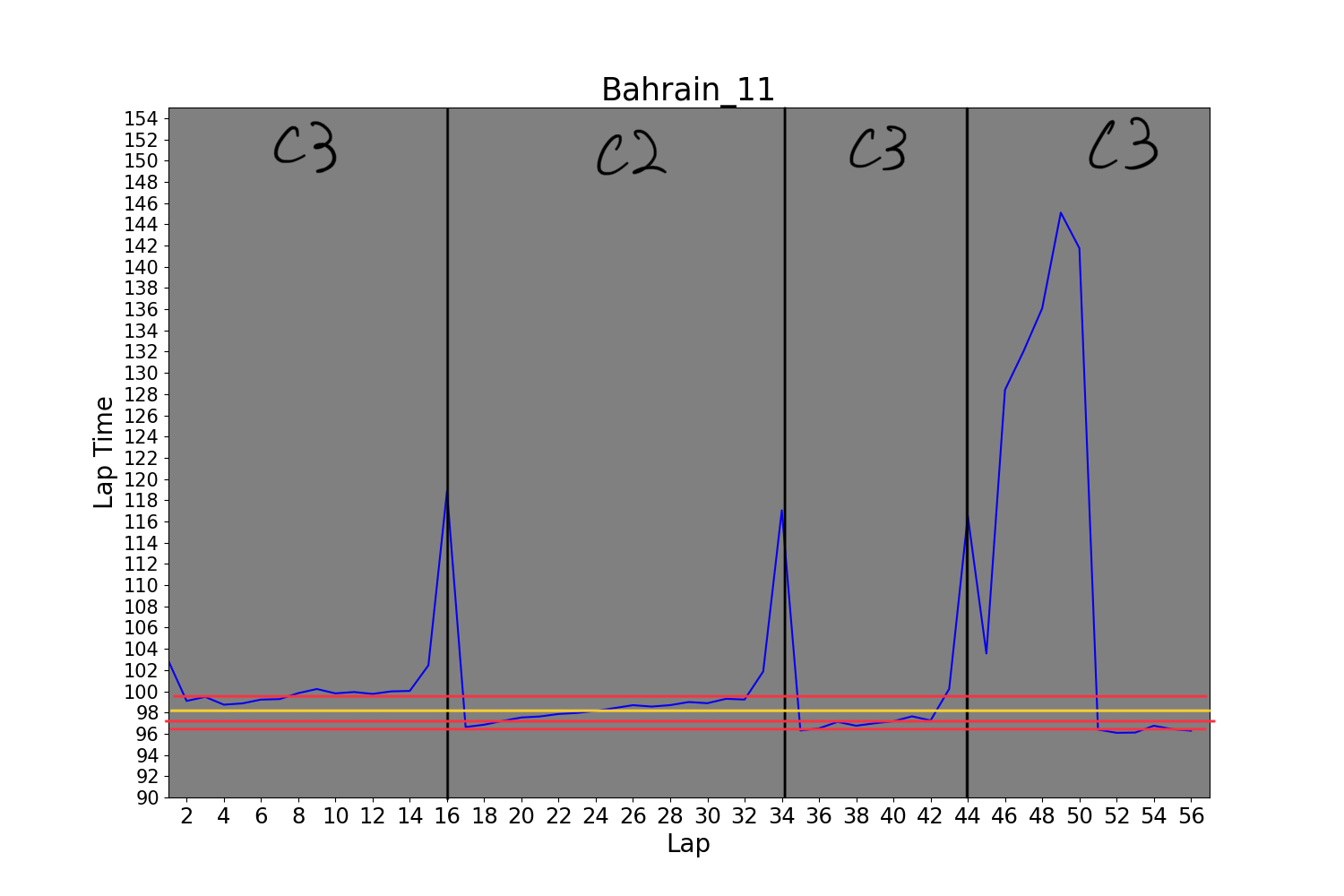
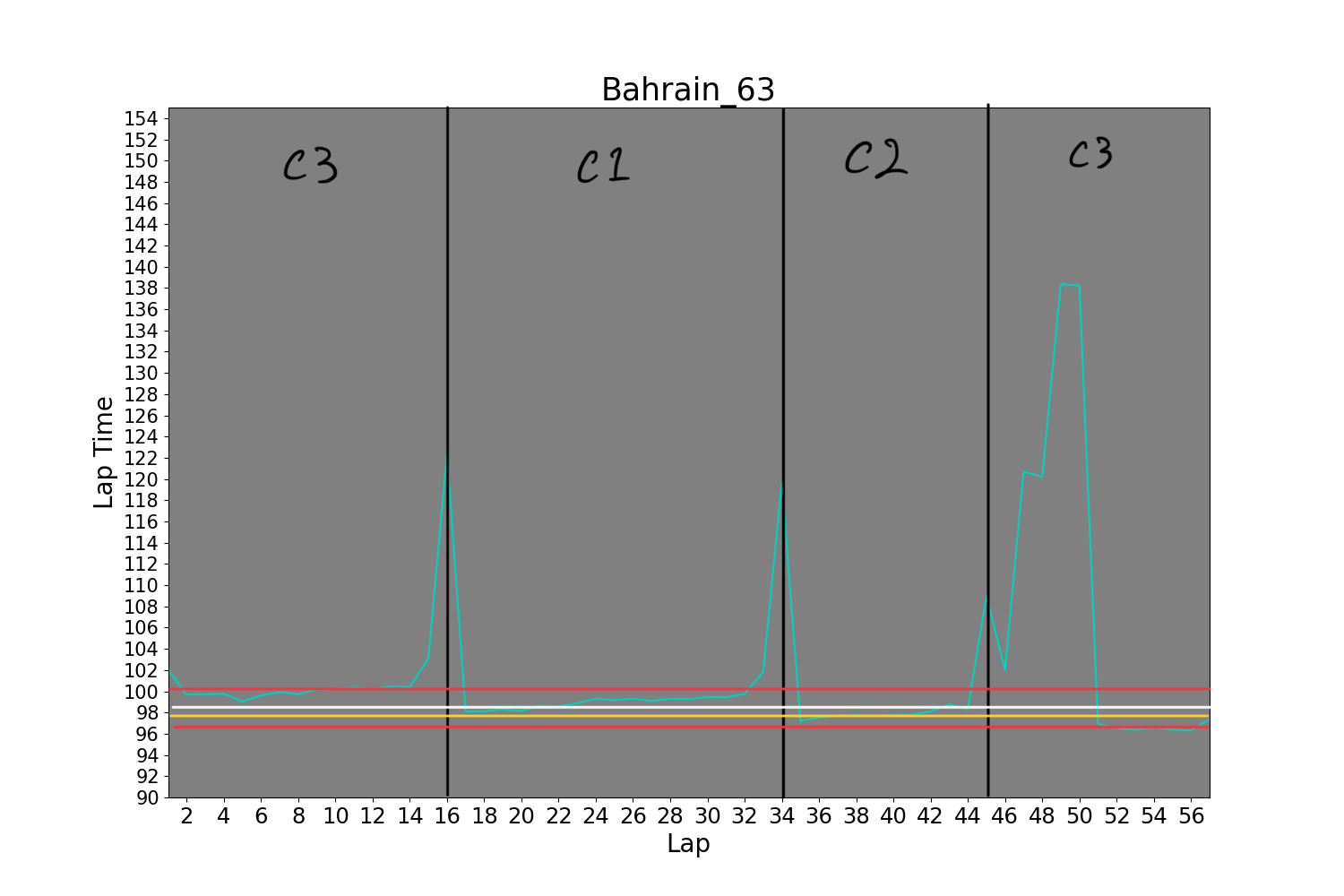
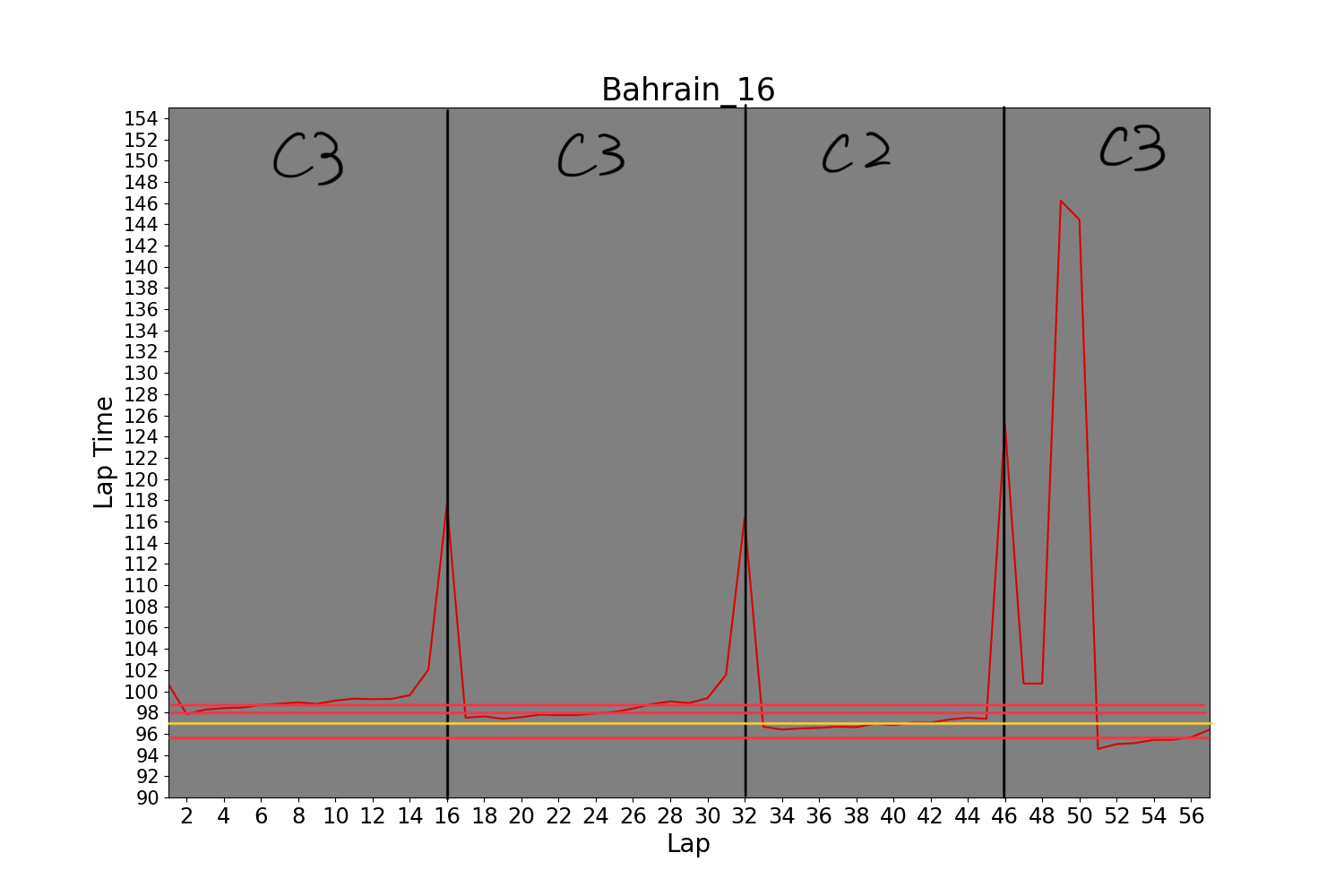
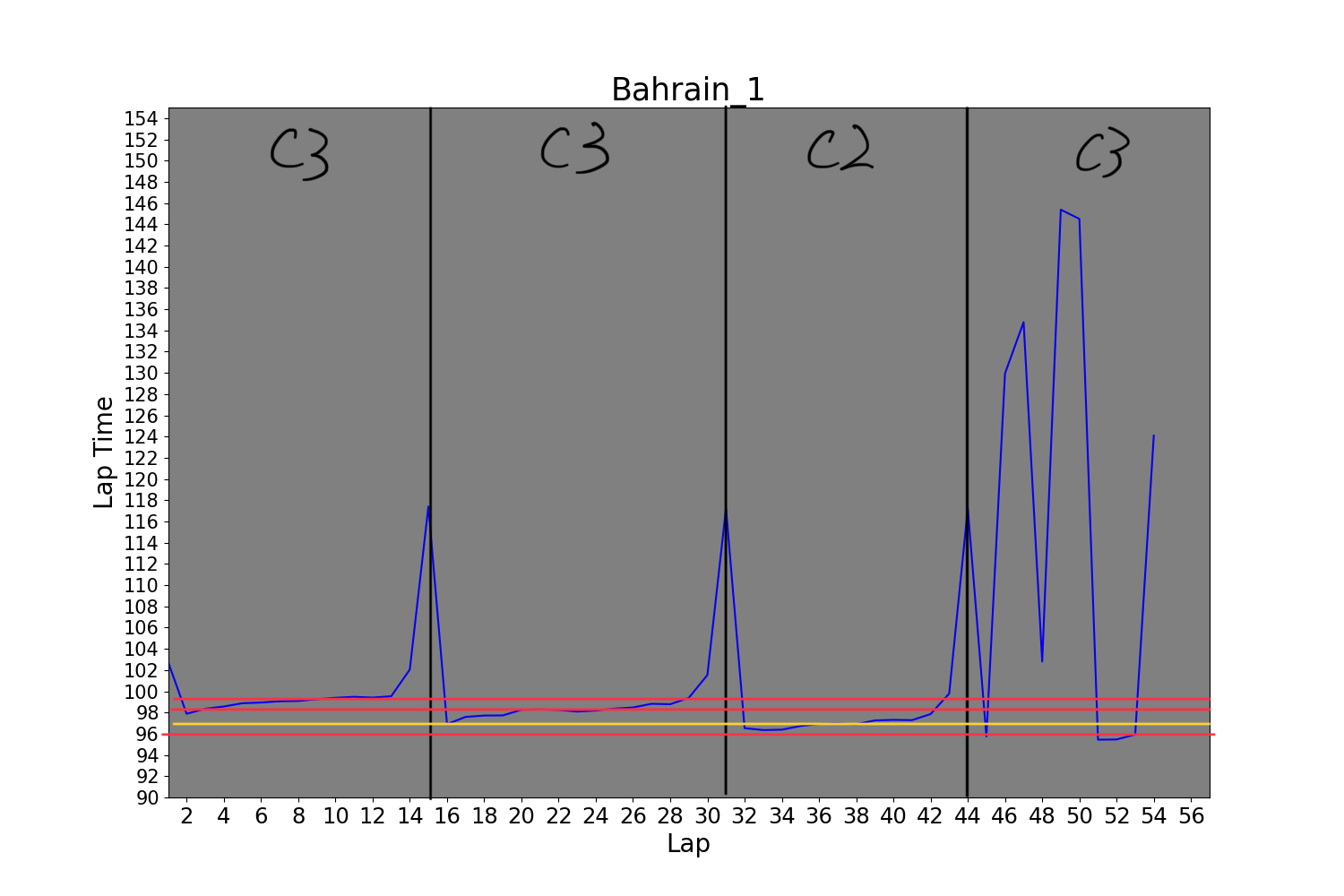
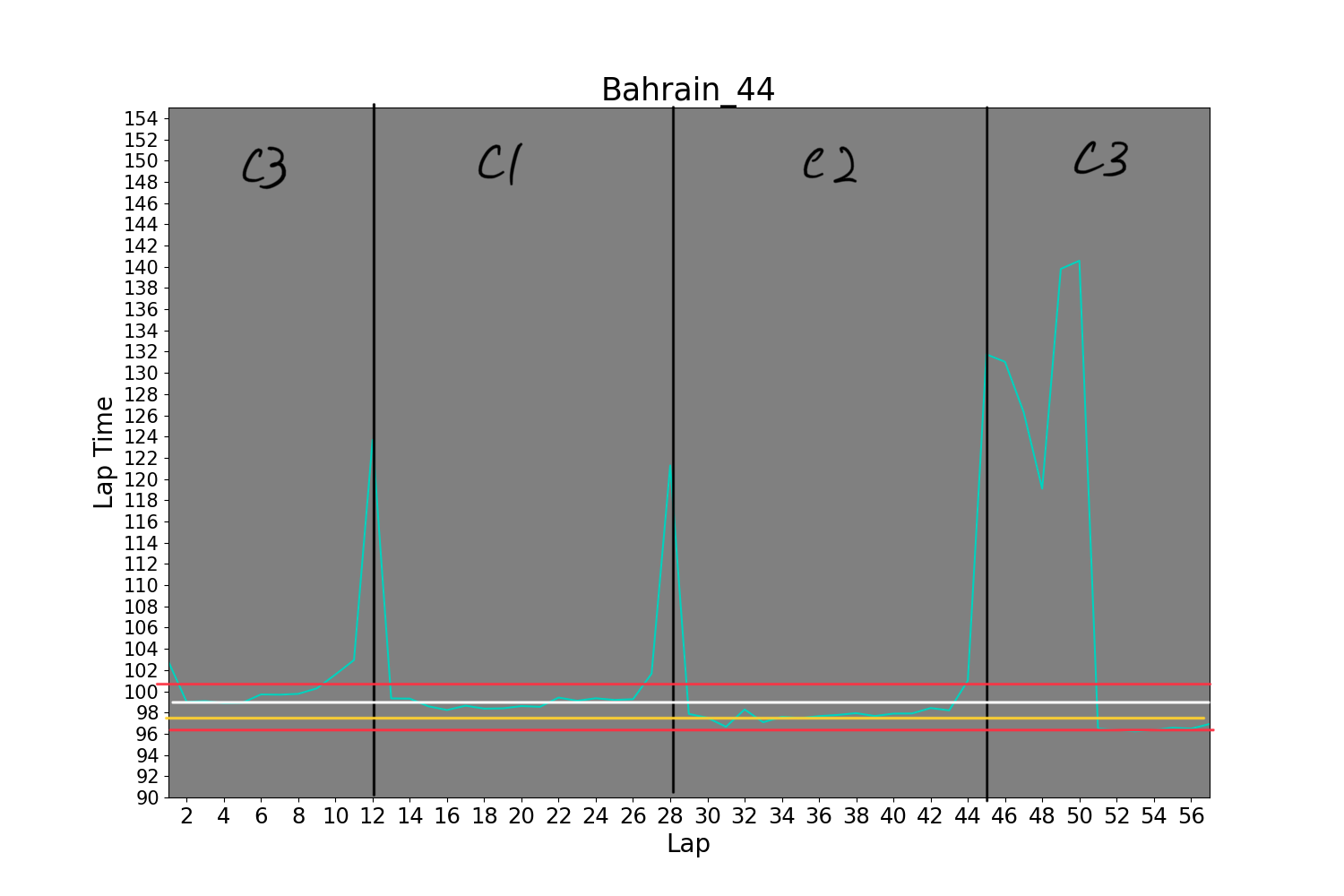
| SA | SpeedI1 | SpeedI2 | SpeedFL | SpeedST |
| --- | --- | --- | --- | --- |
| Result without retire | -0.48959880165839254 | 0.1672859881860141 | 0.7420572558551509 | 0.675735466193413 |

| AU | SpeedI1 | SpeedI2 | SpeedFL | SpeedST |
| --- | --- | --- | --- | --- |
| Result without retire | -0.7132236085453163 | 0.19701791570992647 | 0.34877137772023353 | 0.5238330669940736 |

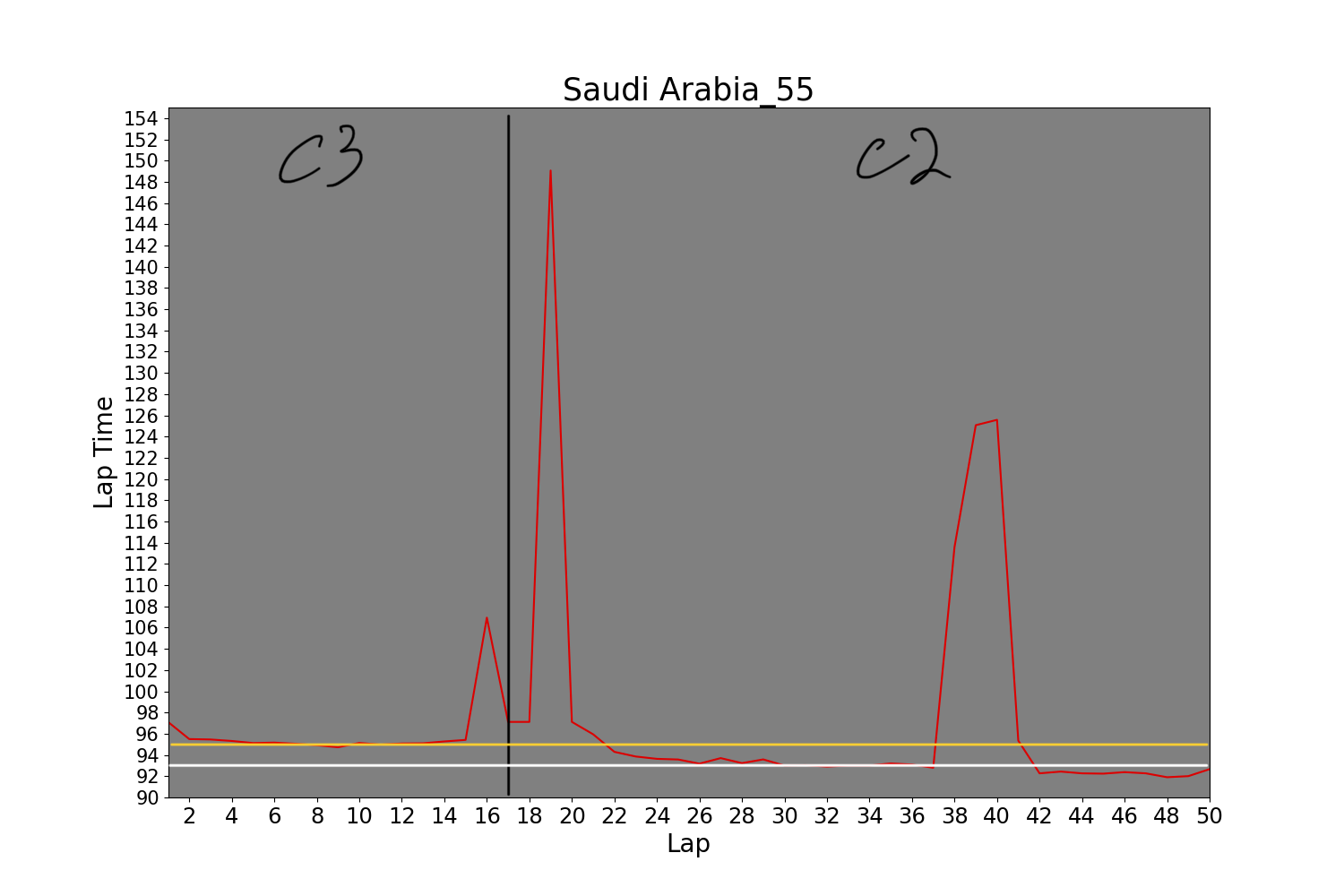
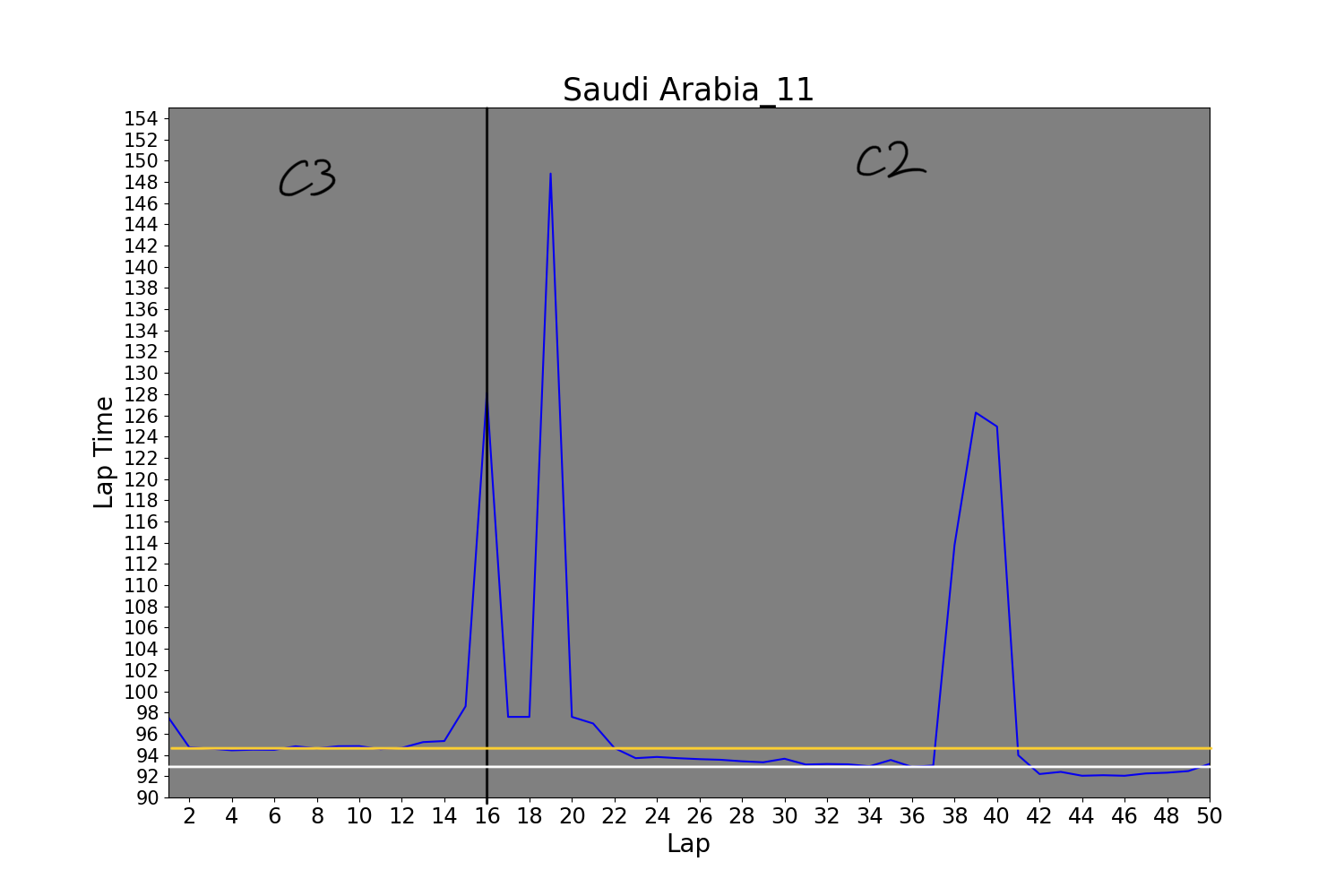
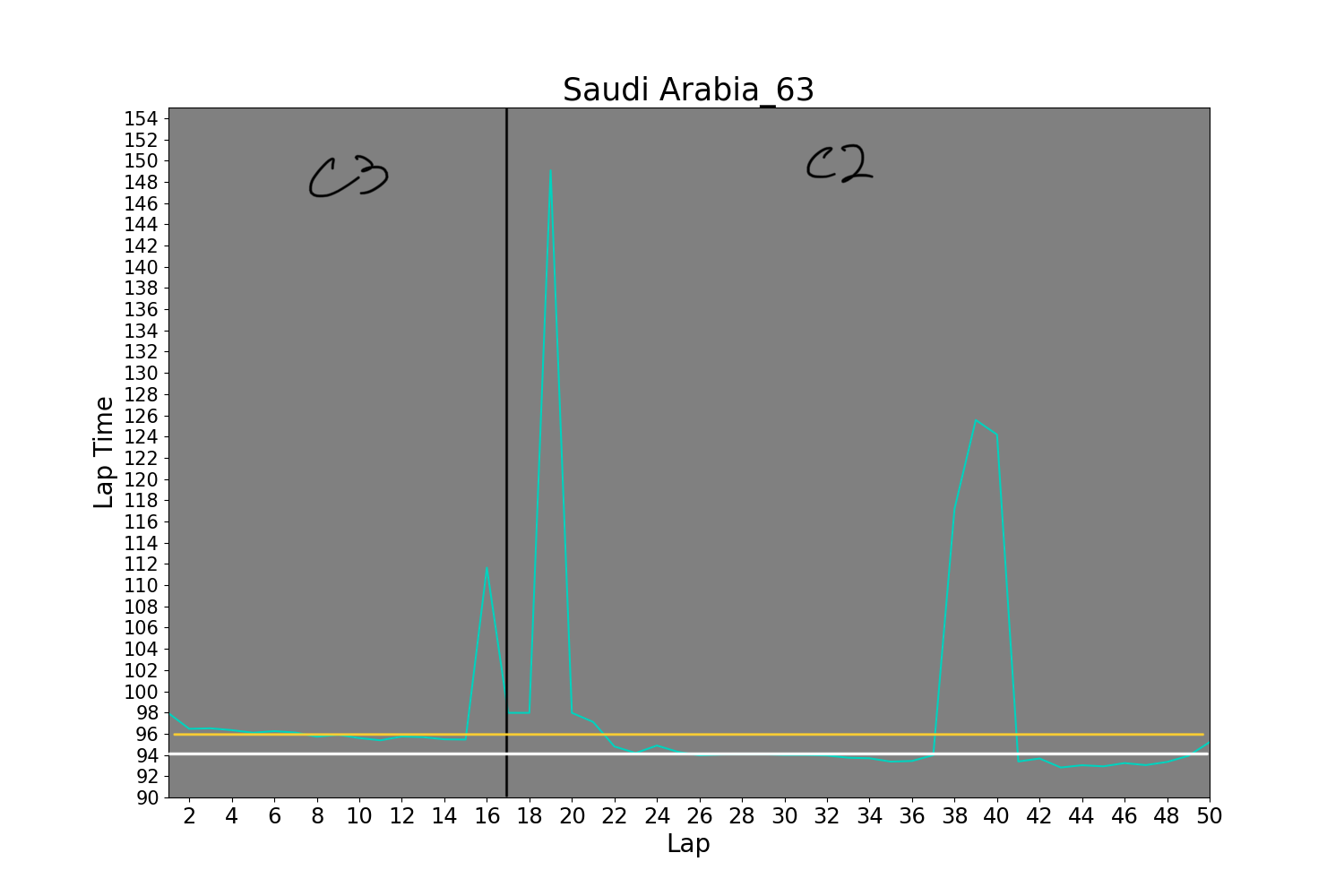
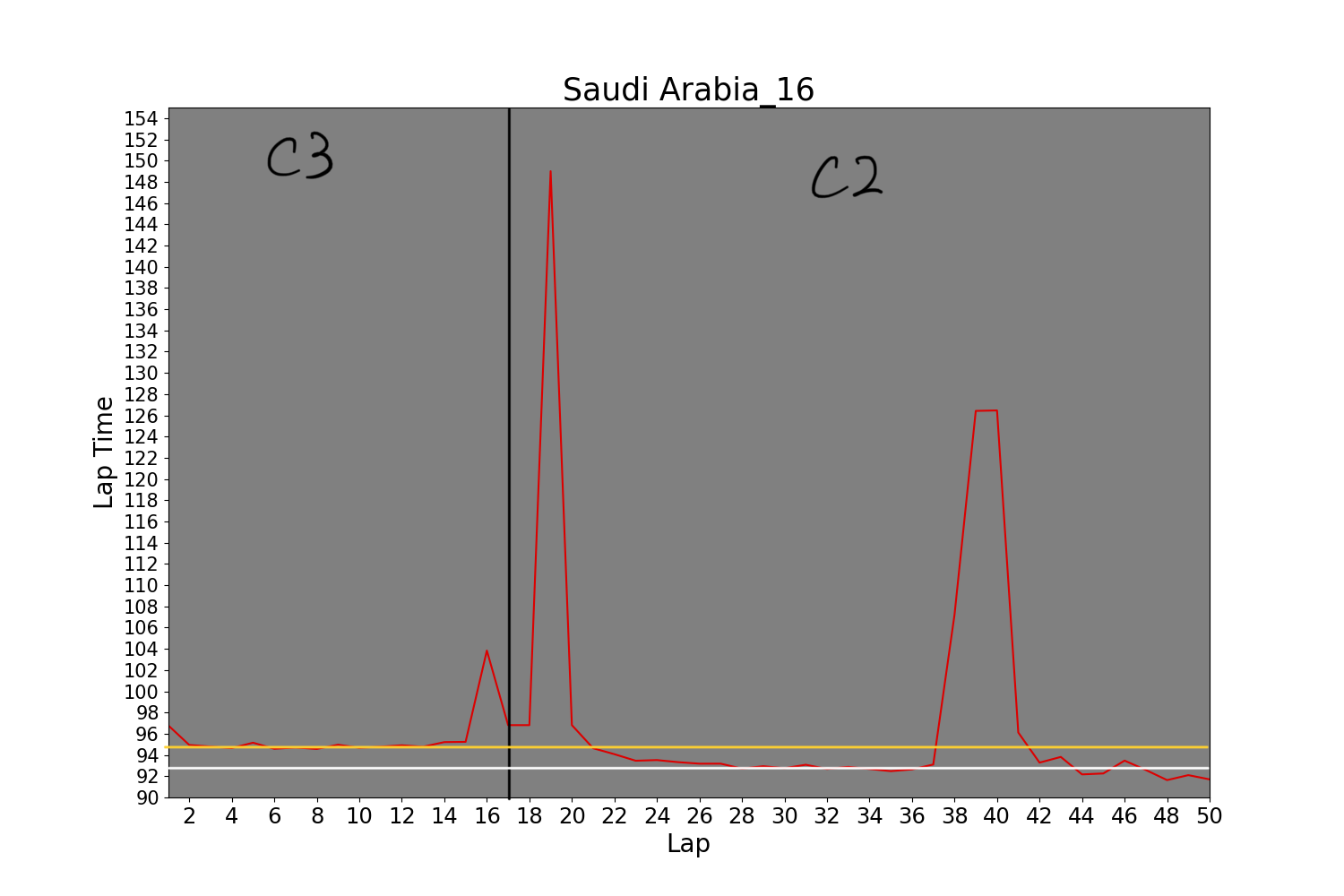
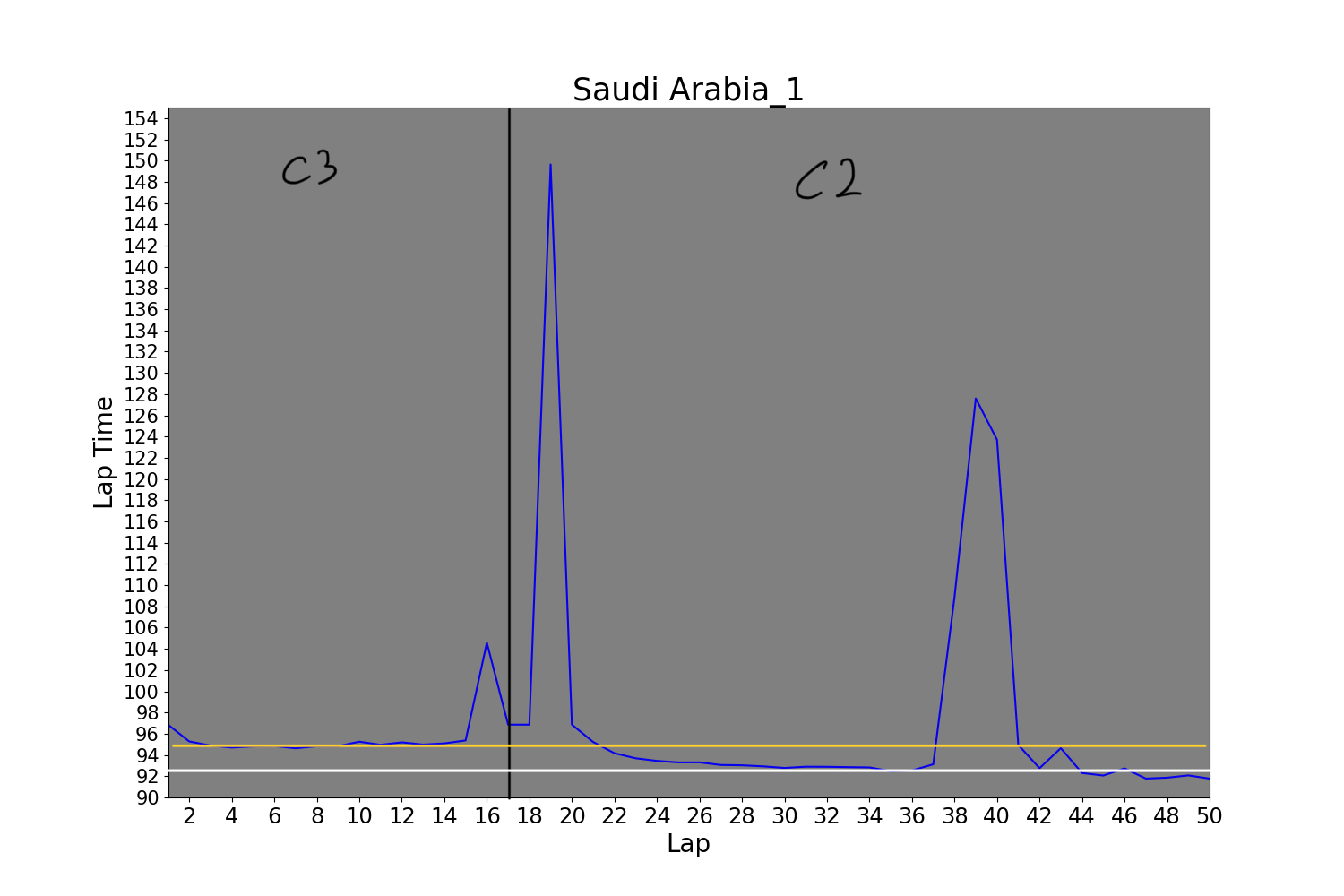
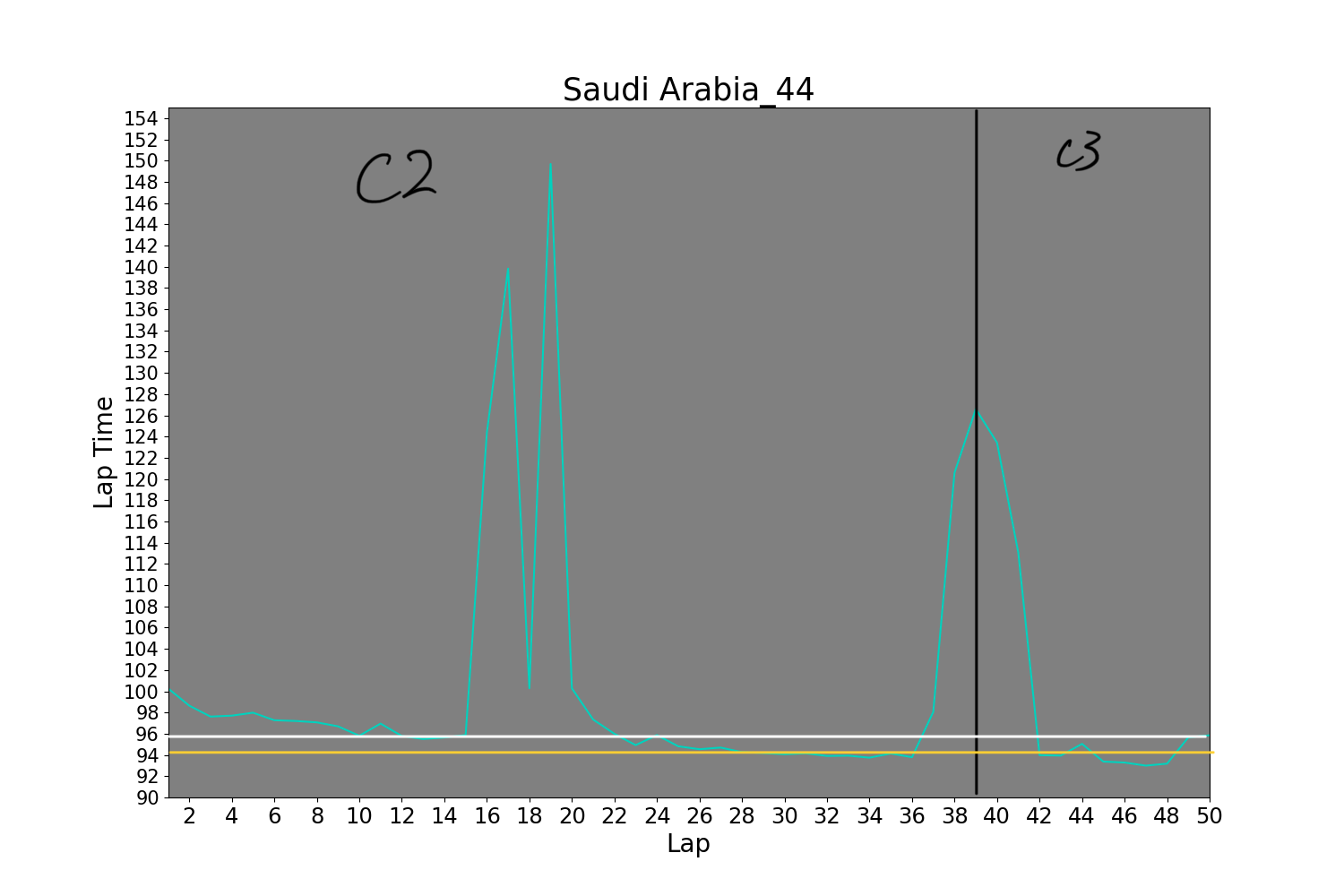
We took the standard deviation of the speed data to see if there is a huge difference in the speed among the cars. The data says that the deviation of the speed for Bahrain and Australia are pretty much the same, but since Saudi Arabia is more a high-speed circuit, the standard deviation is more large for this race.

We also calculated the correlation between each speed data and the race result to see if a speed from a particular sector contributed more to the result. Since the retired cars negatively affects the correlation, we removed the data of the retired cars here. However, we see lots of negative correlation between the speed and result. This is because the modern tracks have a lot of corners and the time cars spending in corners is much longer than the time cars spending on the straight. Therefore, modern formula cars have huge down force to gain the grips to make the cornering time faster. Usually this balance between the cornering time and the straight speed is a trade-off and straight speed is not always the dominating factor of the race pace. Therefore, here, we concluded that top speed on the straight is not a factor of their poor performance in the race.

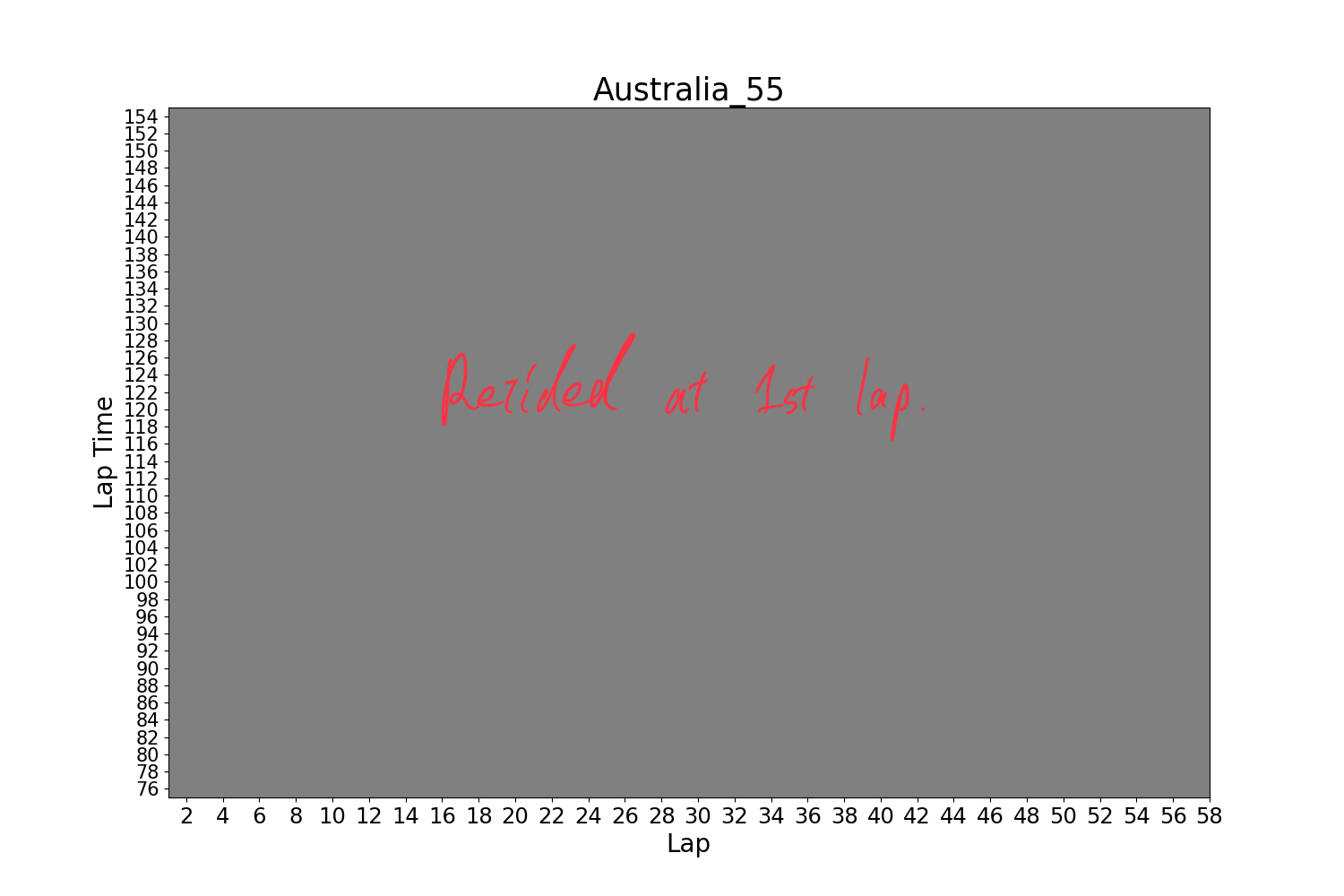
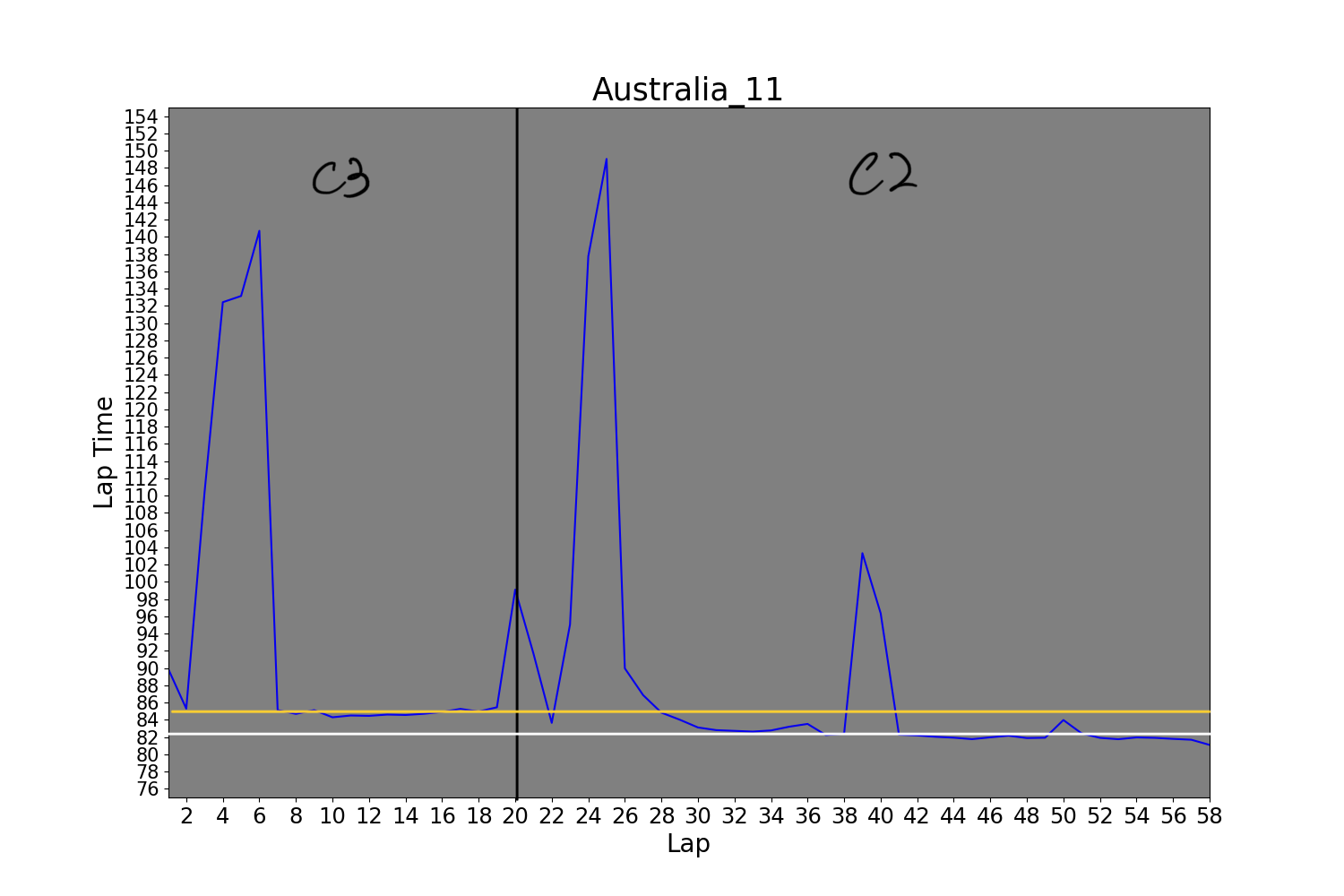
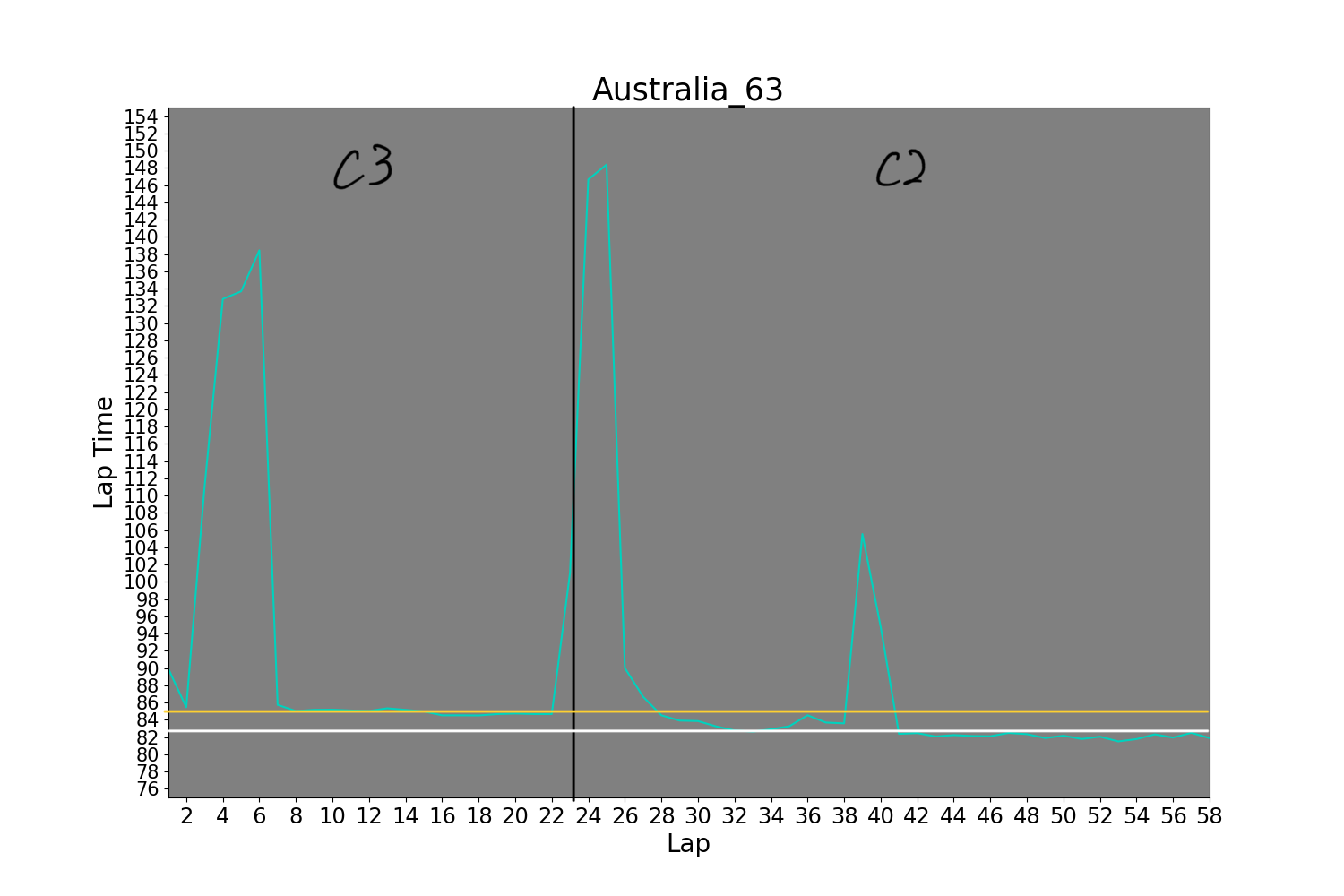
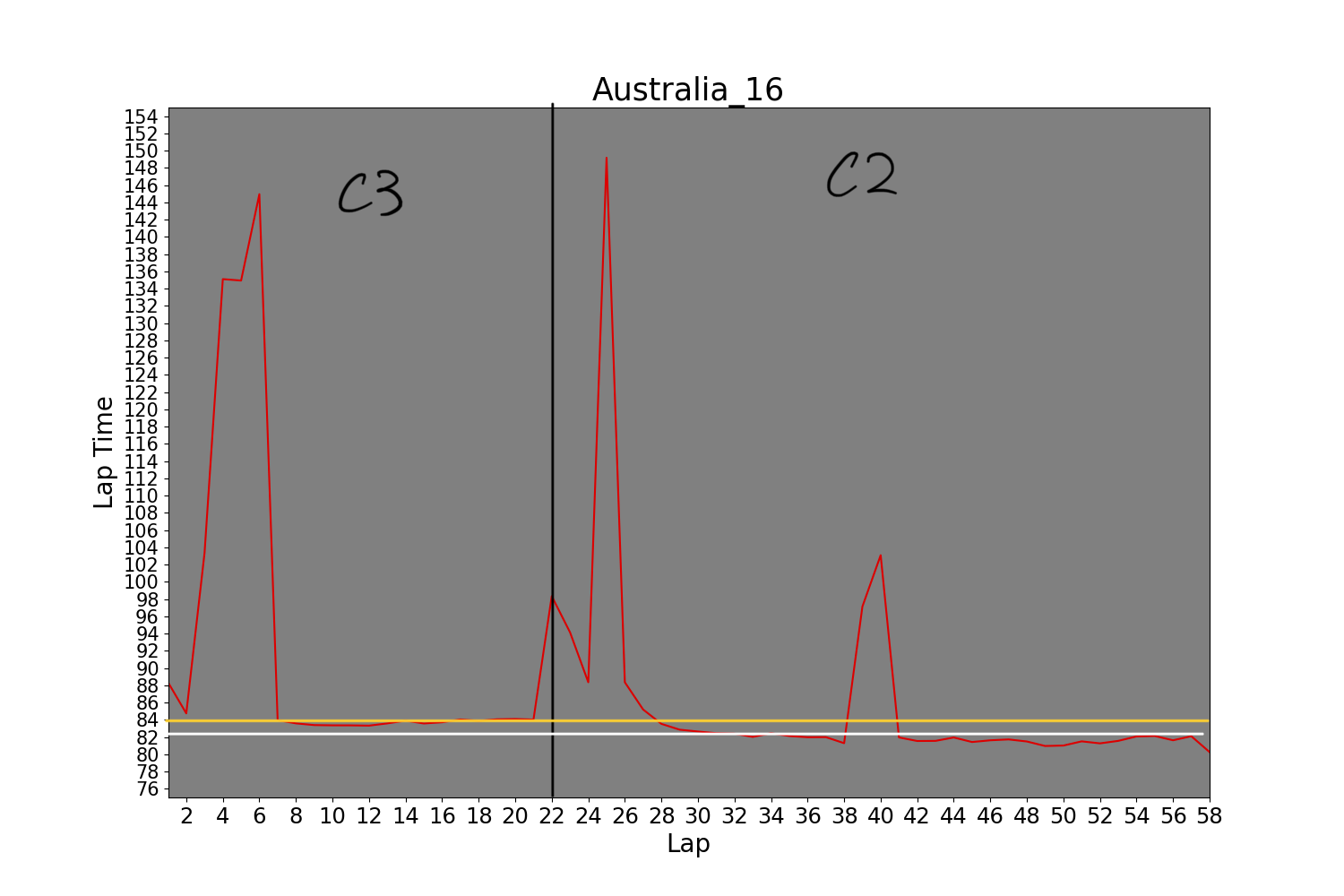
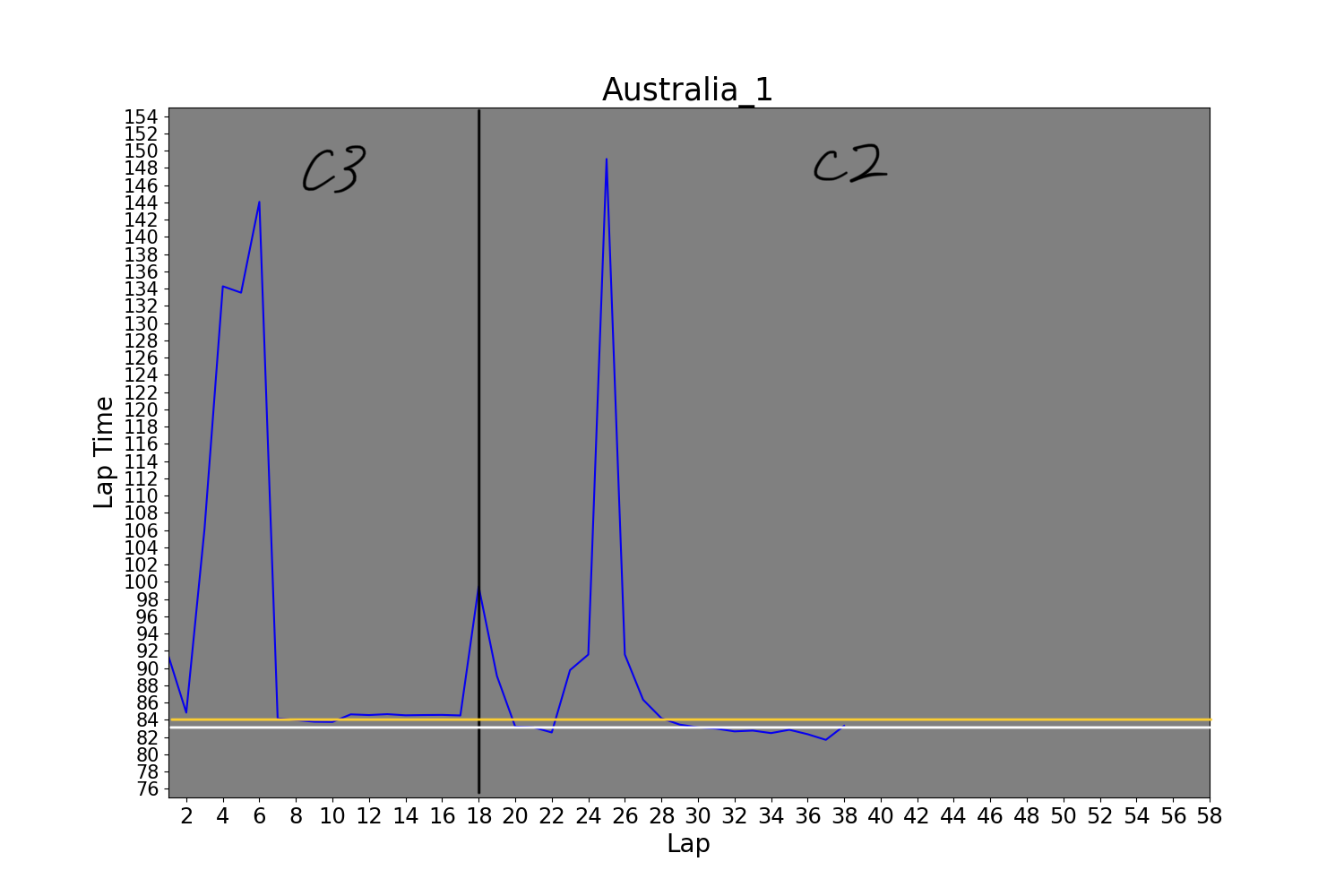
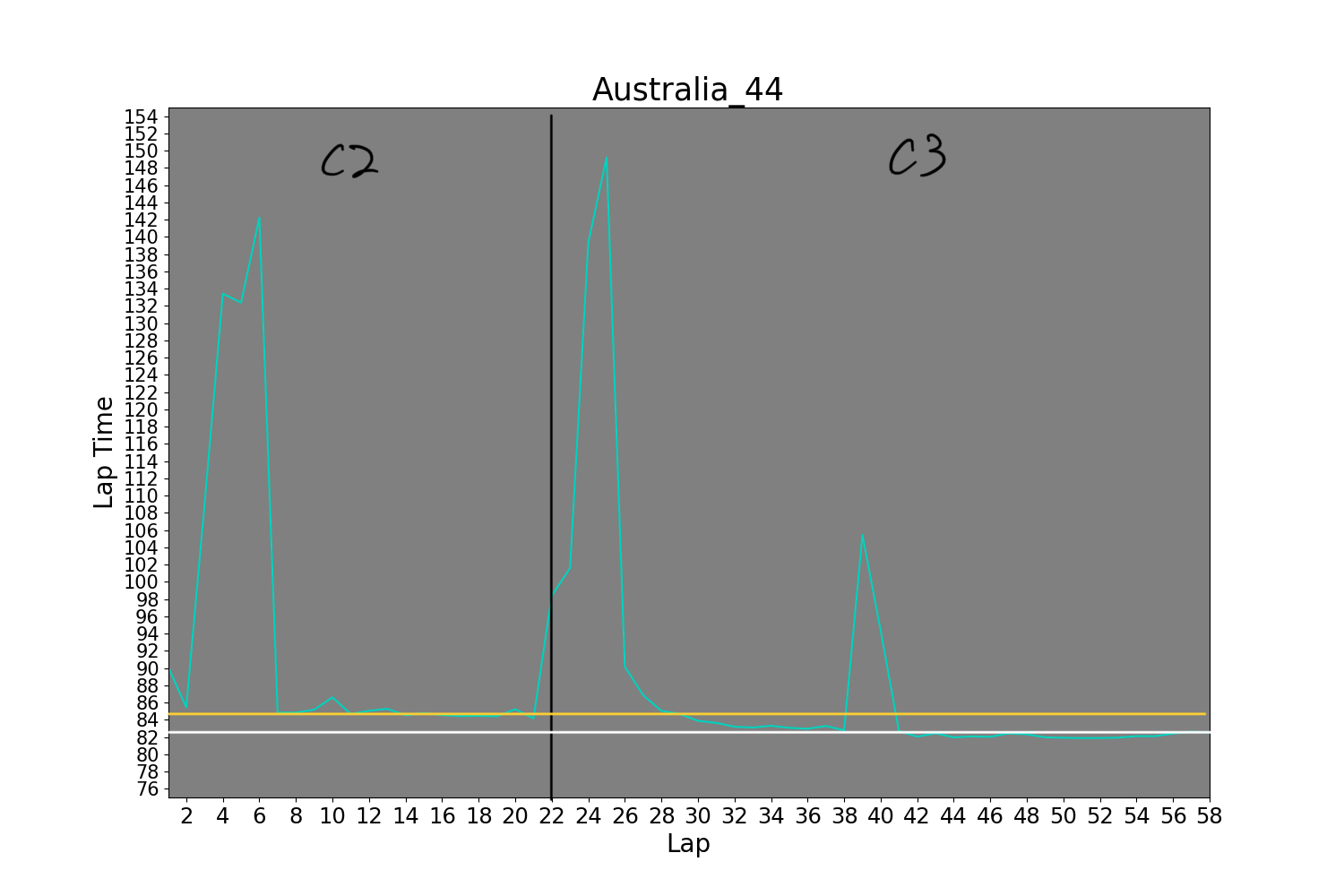
The quali lap time from Saudi Arabian GP for Max and Carlos for lap 11 were replaced for NaN because they were too fast (61.871 sec, 73.67 sec). They are probably mistaken data. Australian GP quali vs race plots are with and without the outlier. This outlier (Lance Stroll = car number 18) got in an accident right after he went out for the outlap for the first flying lap with Nicholas Latifi (car number 6).



C1, C2, and C3 show the type tyres. C1 is the slowest with longest durability and C3 is the fastest with the shortest life. The car number 1, 11, 16 and 55 were dominating the race even car 1 and 11 retired just before the race ended. All these were on the same strategy. However, the Mercedes team was on a different strategy. They are the only team that used C1, the slowest tyre.



In saudi arabian race, car number 1, 11, 16, 55 and 63 followed the same strategy. However, 44 did not. As a matter of fact, these car finished the top 5th, but 44 finished in 10th with a different tyre strategy.



Every car other than driver 44 in the top 6 used the same tyre strategy.

Numbers are rounded up to 3 decimal places

| Bahrain | Stint 1 | Stint 2 | Stint 3 | Stint 4 |
| --- | --- | --- | --- | --- |
| 44 | 99.984 (C3 for 11) | 98.880 (C1 for 15) | 97.730 (C2 for 16) | 96.485 (C3 for 12) |
| 63 | 99.948 (C3 for 15) | 98.930 (C1 for 17) | 97.874 (C2 for 10) | 96.647 (C3 for 12) |
| 1 | 98.937 (C3 for 14) | 98.203 (C3 for 15) | 97.188 (C2 for 12) | 95.721 (C3 for 12) |
| 11 | 99.554 (C3 for 15) | 98.162 (C2 for 17) | 96.968 (C3 for 9) | 96.345 (C3 for 12) |
| 16 | 98.839 (C3 for 15) | 98.134 (C3 for 15) | 98.493 (C3 for 13) | 95.375 (C3 for 11) |
| 55 | 99.457 (C3 for 14) | 98.277 (C3 for 18) | 96.953 (C2 for 10) | 96.041 (C3 for 12) |

For the Bahrain race, car number 44 and 63 were on a different strategy for the stint 2. Their average lap times for stint 2 were 98.880 for 15 laps and 98.930 for 17 laps. Since the lengths of the stint were similar for all of them, 0.796 ~ 0.653 sec of delay for each lap stuck for this stint.

| Jeddah stint average lap time | Stint 1 | Stint 2 |
| --- | --- | --- |
| 44 | 95.747 (C2 for 38) | 94.149 (C3 for 11) |
| 63 | 95.909 (C3 for 16) | 93.837(C2 for 33) |
| 1 | 94.975 (C3 for 16) | 92.763 (C2 for 33) |
| 11 | 94.737 (C3 for 15) | 93.076 (C2 for 34) |
| 16 | 94.855 (C3 for 16) | 93.026 (C2 for 33) |
| 55 | 95.162 (C3 for 16) | 93.028 (C2 for 33) |

For the Saudi Arabian race, car number 44 took a different strategy. However, the his first stint average lap time is faster than the ones on the faster tyres. However, everyone else changed the tyre at the same time, which they deployed the safety car for there was an accident at lap 15. On the safety car laps, the cars have to go slower. Therefore, the lap time is slower and the drivers get benefit if they can pit it to change tyre for this time. However, car number 44 missed this time to change the tyre because the pit lane entry was closed for the safety reason. Therefore he lost time when he changed the tyre at lap 45.

| Australian stint average lap time | Stint 1 | Stint 2 |
| --- | --- | --- |
| 44 | 85.161 (C2 for 21) | 85.566 (C3 for 36) |
| 63 | 85.196 (C3 for 22) | 85.479 (C2 for 34) |
| 1 | 84.583 (C3 for 17) | 83.390 (C2 for 20, retired) |
| 11 | 85.055 (C3 for 19) | 83.036 (C2 for 40) |
| 16 | 84.235 (C3 for 21) | 84.646 (C2 for 36) |
| 55 | retired at 1st lap | retired at 1st lap |

For the Australian race, only the car number 44 was on a different tyre strategy than everyone else on the table. His first stint was on a slower tyre and second stint was on the fastest tyre. He was supposed to drive a lot more longer on the slower tyre, or the other cars on the faster tyre had to drive for shorter. However, their tyre change timings were not significantly different. Looking at the second stint for car number 44, even though he was on the fastest tyre, he was the slowest of all. This might be because the tyre was gone quicker than the other cars and he had to drive slower, or the race track was too hot to degrade his tyres a lot quicker and thus the cars on the slower tyre benefited from being on the more durable tyre.

What did you learn through your analysis?

There were a couple key things we learned through our analysis of the 2022 formula 1 races. First, the speed on the track was not a huge difference among the top 6 cars that competed in the last part. In the first two races, the Mercedes team did not perform well in their top speed, but they caught up with the other teams in the third race. Next, the qualifying session's fastest time and the race pace did not show a significant difference. Sometimes, these are trade-off and the fastest cars on the qualifying session can go slower in the race. However, we confirmed the strong correlation between the qualifying result and the race pace for all races. Finally we learned that the tire pattern mattered. The Mercedes team tended to have a different strategy than the other two teams. And when they take a different strategy, they perform poorly. Moreover, car number 63 was on the same strategy as the others on the saudi arabian and australian race, and he finished in top 5 for both. However, the car number 44, who drives the same car as 63, was on a different strategy than the car number 63 in the Saudi Arabian and Australian race. He did not finish in a good position for these two races as well.

Was anything about your results surprising or unexpected?

One thing I thought was surprising was that the tire strategy seemed to impact the results the most compared to other attributes. Another thing I was surprised about was how well the speed pca attribute worked with the data. According to some formula 1 analysts, they hypothesized the race result corresponds to the engine supplier. And it is true regarding the result plot, but the top speed was not the factor in the case of the top 6 cars. I was surprised because there must be other race result factors that are affected by the engine and I did not know that.

How will your work help with understanding the problem you set out to solve?

I think our work helped better understand the problem as to why Mercedes was underperforming and what factors are important in having a higher overall result. Specifically I think the results we had about the different tire strategy patterns best answers the two questions, but more importantly this sheds light on how important tire strategy is, and why more teams should strategize about it.

What else would you do if you had more time?

First we would compare the 2022 formula 1 data to previous years to increase our sample size. Next if we had more time we could dive into more specific data, like how fast/slow each turn is taken and seeing if there are certain turns that give a harder time than others, or data on the car specific aerodynamics and if that has any major impact on speed and handling. Another data point we would try to create is to create graphs on rpm and horsepower and compare it to other factors. Another technique we would implement is a line of best fit to better interpret the data. Finally we would wait until the end of the current season to implement data analysis on the final circuit races.