Formula 1 race strategy is a complex combination of data analysis on many datasets. Factors such as the car's downforce, aerodynamics, engine performance, drivers cornering preference, braking strategy, etc. Tire condition and pit stop strategy are two factors that can dramatically change a range and these are reliant on other conditions. In order to simplify the tire stint choice and tire saving versus push strategy I attempted to predict the mean number of tires that in a Formula 1 Race weekend given a number of features that are commonly varied such as weather conditions, Pirelli weekend tire allocation, number of laps under a safety car, etc. Outside of tire strategy these features can have a huge impact on race strategy and I desired to explain their impact on tire degradation and the tire needs for these changing conditions to supplement the team's knowledge for a better race finish result.

Unfortunately I was not able to apply features such as curb size, downforce pressure on tires, impact of racing under dirty air, etc. as most of these datasets would be protected by the teams as they would give an edge to the competitors if released. Thus while my predictor does take in common knowledge it does not specify for cars, constructors, car setups etc.

In order to do this prediction I used three datasets that were available through a user on kaggle called diana. She worked on predicting a drivers race status (finished, crashed, retired, technical problems etc.) through various datasets. I choose to take Pirelli Tire Categories, Tyre Strategy Clusters and Weather Safety Car from her page and collate the features I desired.

Pirelli\_Tyre\_Categories.csv is a dataset that gives data on Pirelli’s (Formula 1 tire partner) allotted tires per race weekend. I felt this feature was important because if Pirelli allotted a set of softer tires per the weekend more tire degradation would be expected and as such more tires would be needed for the race. Tyre\_strategy\_clusters.csv is a dataset that has each driver's tires used during the race. This data was used to find the mean per a race weekend aka our Y variables. Weather\_SafetyCar.csv dataset contains information on the weather of a race weekend and the number of laps that were under a safety. Weather is important as it can dictate how much tire degradation there is. Hotter weather equals more degradation while cooler weather would allow tires to last longer. As well, weather can impact changes of tires from slicks to wets, this could increase the number of tires used during a race. Laps under safety cars are a major factor as these laps historically have a tire lower degradation due to the decreased speeds.

In this project I manipulated the dataframes to extract the features I felt were necessary. This involved matching indexes, cutting out slices of data, getting dummy variables, label encoding etc.

I choose classifier models in order to do the prediction on my dataset included Random Forest, SVM, KNeighbors and Decision Tree. I found that the best model for this dataset was the Random Forest Classifier.  This makes sense as it is regarded as one of the best models currently in terms of accuracy. Although its accuracy was only 65% which is concerning and could point to other factors(features) that influence the mean number of tires used. As well it could mean that you can not thoroughly predict the mean number of tires for all drivers but a more specific analysis and prediction on individual drivers could be done. This would make sense as individual drivers have specific driving styles. For example Sergio Perez is known for making his tyres last no matter the conditions whereas a rookie driver may have a harder time accomplishing the same number of laps on a set of tyres. Also analysis on individual drivers with the teams data on tire degradation and the impact of the teams cars on tires could improve the model.

Overall I feel that 65% accuracy is a good result for this model, with the features selected.There are outliers in which the model did not even get close such as the Canadian Grand Prix of 2017 in which one tyre was predicted when four tyres was the mean. The limited amount of specific data available did lend itself to a more generalized analysis in my opinion. Without more detailed and specific datasets a superb accuracy is more than likely not possible.