Aya Tarist
ANOP330
Prof. Bailey

Executive Summary: Exploratory Data Analysis for Formula1 PitStops Prediction

Project Overview:

- In this project, the aim is to develop a machine learning model to predict pit stops in Formula 1 races, enhancing race strategies and performance.
- Through exploratory data analysis (EDA), we delve into factors like tyre wear and race conditions, guiding data cleaning and revealing key predictors for our model.
- This foundation aids in crafting a predictive tool that, by learning from historical patterns, forecasts pit stops, offering insights into F1 racing strategies and decision-making.

Dataset Description:

- *Source:* **fastf1 Python package**, a specialized library for accessing and analyzing Formula 1 data. This comprehensive resource outlines how to fetch telemetry, timing, and session data, which are pivotal for conducting in-depth analyses of races, teams, and drivers' performances.

- The dataset encompasses 792 instances across 37 attributes, derived from the 2022 Belgian Grand Prix - for this specific project.

- The primary objective is to predict the occurrence of a pit stop in a given lap in this race, indicated by the target variable PitStopOccurred.

→To facilitate our analysis, particularly for the classification task of predicting pit stops, I have introduced a custom variable, PitStopOccurred. This binary indicator was not directly available in the raw data provided by fastf1. Instead, I derived it by examining the PitInTime and PitOutTime fields for each lap. If either field is not null for a given lap, indicating that a car entered or left the pit lane, PitStopOccurred is marked as 1; otherwise, it's marked as 0. This process of feature engineering transforms raw timing data into actionable insights for predictive modeling, turning an indirect indication of pit stops into a clear target variable for our machine learning project.

- Among the varied features, key attributes include:

→ LapTime (float64): Measures how long a driver takes to complete a lap, crucial for assessing performance and strategy.

→ Driver (object) & DriverNumber (object): Identifies the driver, essential for analyzing individual performance and strategies.

→ Stint (float64): Indicates the sequence of laps on the same set of tyres, key for understanding race strategy and tyre management.

Aya Tarist
ANOP330
Prof. Bailey

→ PitOutTime & PitInTime (timedelta64[ns]): Record times of leaving and entering pits; basis for creating PitStopOccurred, predicting strategic stops.

→ Sector Times (Sector1Time, Sector2Time, Sector3Time as float64): Show performance across different track sections, useful for pinpointing where races are won or lost.

→ SpeedI1, SpeedI2, SpeedFL, SpeedST (float64): Capture speed metrics at different track points, vital for evaluating car performance and aerodynamic efficiency.

→ Compound (object): The tyre compound used, affecting grip and wear, critical for strategy on different circuits.

→ TyreLife (float64): How many laps tyres have been used, directly influencing pit stop timing and strategy.

→ Team (object): The racing team provides context for strategic decisions and technology level.

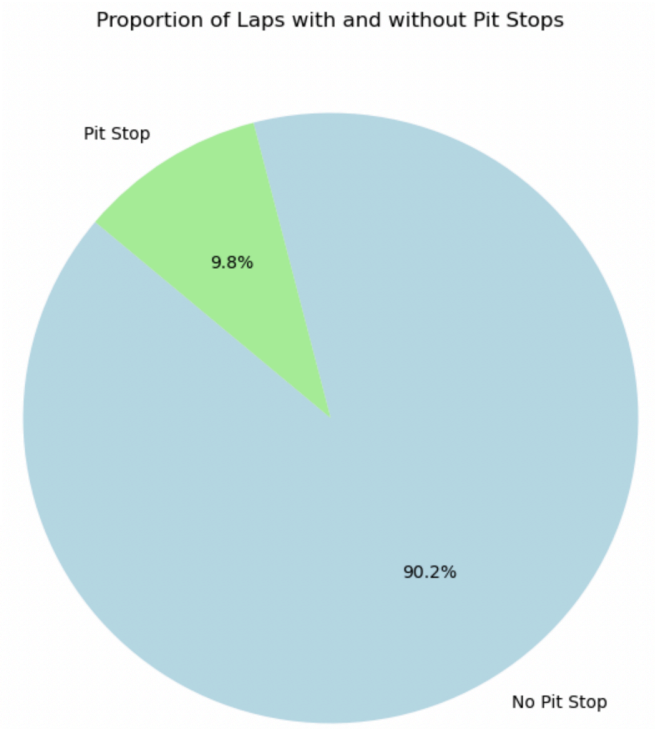
→ PitStopOccurred (int64): Engineered feature indicating if a pit stop happened, the primary target for prediction models to improve strategic planning.

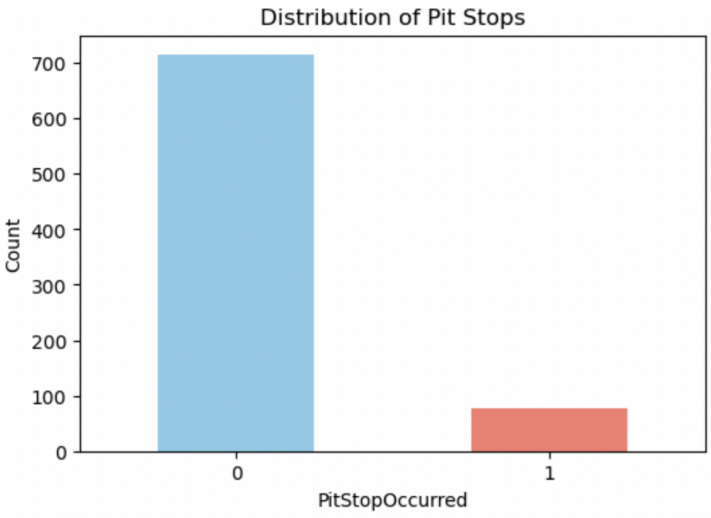
Key Findings:

1. Data Distribution:

- The distribution of the target variable indicates that a significant portion of instances belong to Class 0, representing laps without pit stops (714), while Class 1, representing laps with pit stops, constitutes a smaller proportion of the data (78 laps).

- Class Imbalance: Class imbalance is observed, with 90.15% of instances belonging to class 0 and 9.85% belonging to class 1.



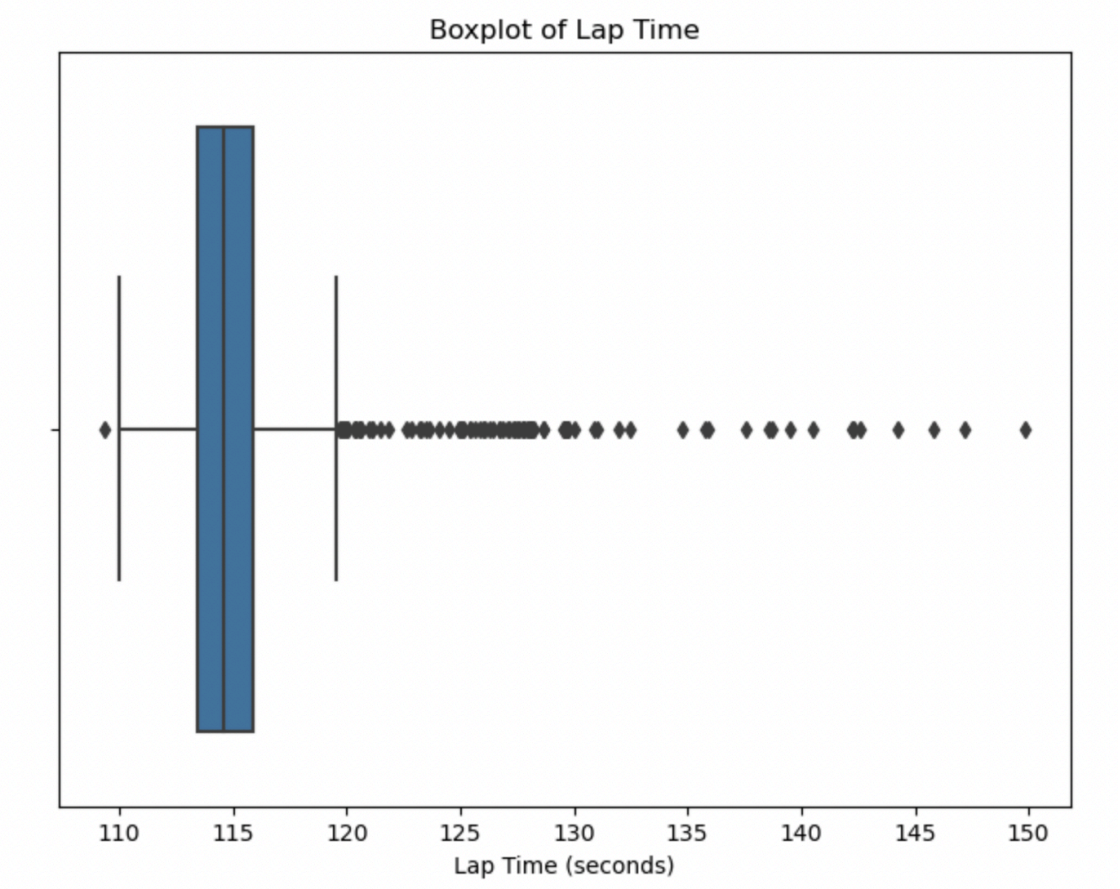


Aya Tarist
ANOP330
Prof. Bailey

2. Feature Analysis:

Distribution of Lap Time:

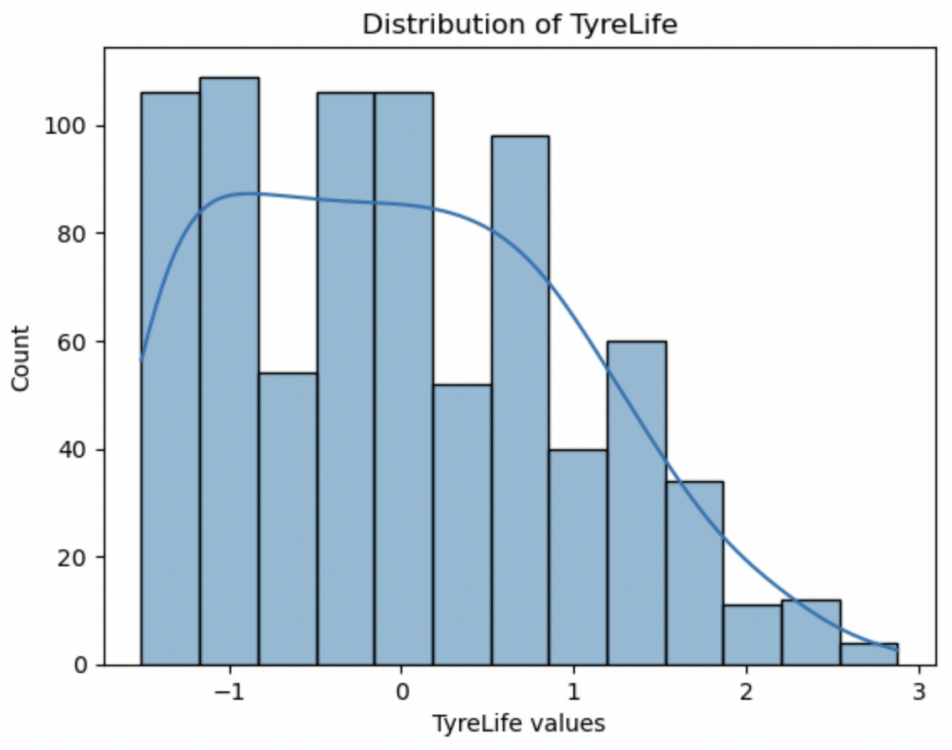
- The box plot displays a range of lap times with a median roughly around 125 seconds.
- There are several outliers on both the lower and higher ends, indicating laps that were significantly faster or slower than the average. These could be due to various factors such as safety car periods, accidents, or exceptionally good lap performance.
- The interquartile range is narrow, suggesting that lap times are generally consistent aside from the outliers.



Distribution of TyreLife:

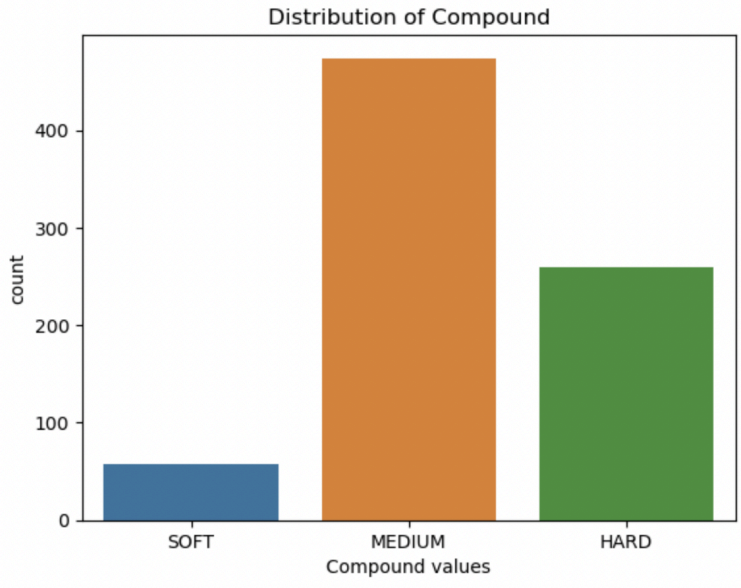
- This histogram shows a right-skewed distribution of TyreLife values.
- Most of the data points are clustered at the lower end of the scale, suggesting that tyres are often changed before they have been used for an extended number of laps.
- There are fewer instances of tyres that have been used for a greater number of laps, as indicated by the tail on the right side of the histogram.

Aya Tarist
ANOP330
Prof. Bailey



Distribution of Compound:

- This bar chart illustrates the count of different tyre compounds used.
- The ‘MEDIUM’ compound is the most frequently used, followed by ‘HARD’, with ‘SOFT’ being the least used among the three.
- This distribution could impact pit stop strategies, as different compounds have different performance characteristics and lifespans.

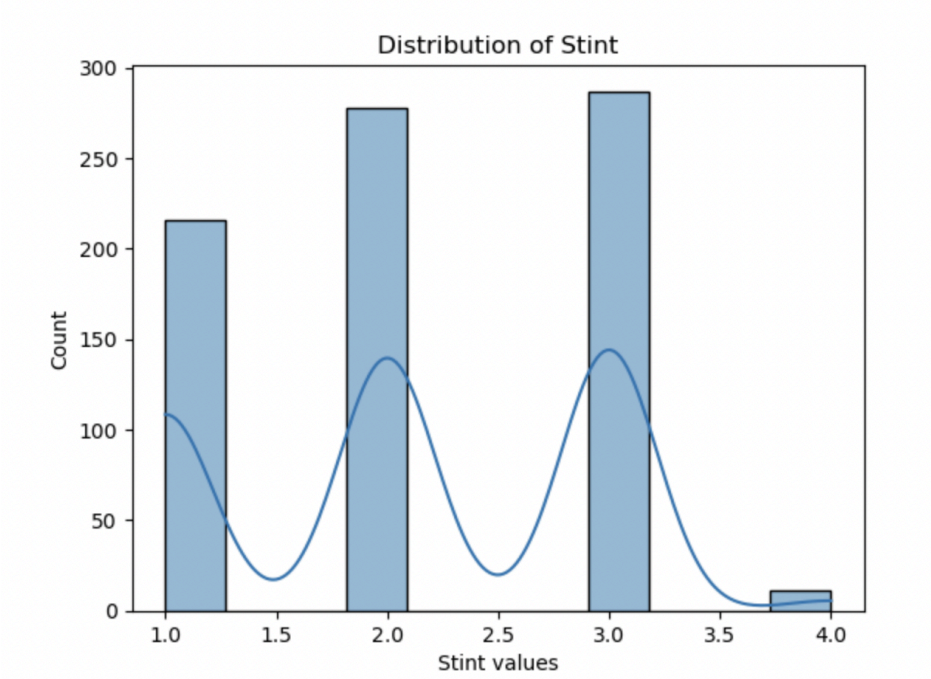


Distribution of Stint:

- This histogram, complemented by a kernel density estimate (KDE), shows the distribution of stint lengths.
- The data appears to be multimodal, with peaks at stint lengths of around 1.5, 2.5, and 3.5. These could correspond to standard stint lengths in a race strategy.

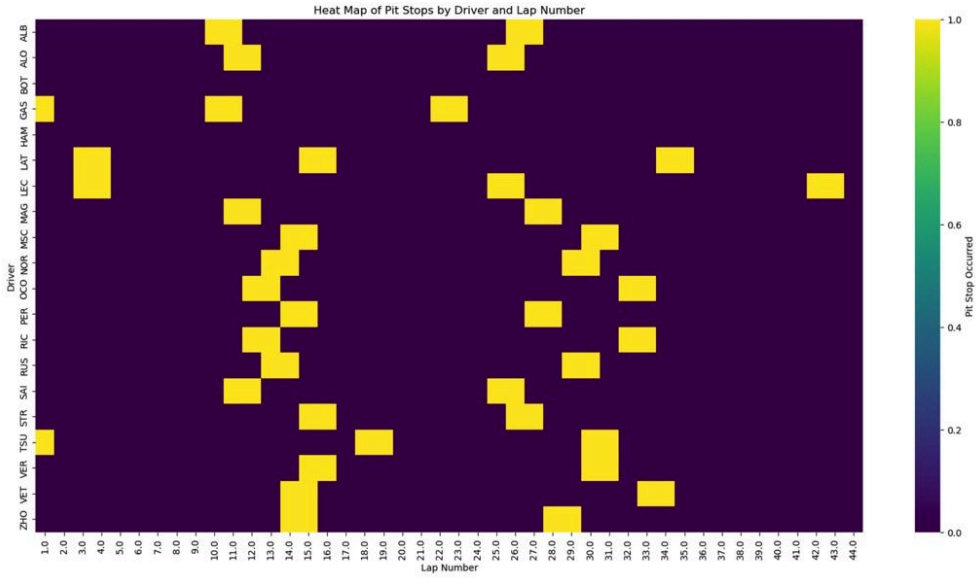
Aya Tarist
ANOP330
Prof. Bailey

- There is a very small count for stint lengths of 4, which might indicate a less common strategy or unique race conditions.



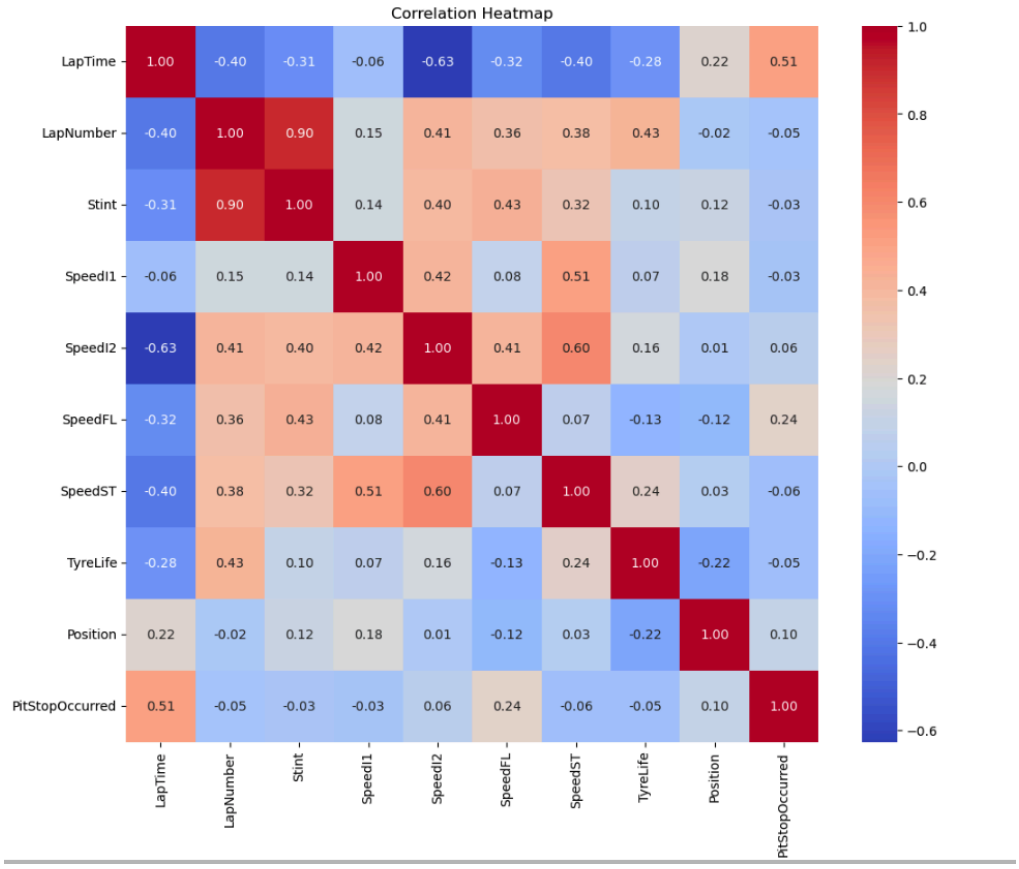
3. Correlation Analysis:

- This heat map displays the occurrence of pit stops across drivers and lap numbers.
- The pattern suggests that pit stops are not uniformly distributed across all laps or drivers.
- Certain lap numbers have higher frequencies of pit stops, possibly indicating common strategic points in the race.



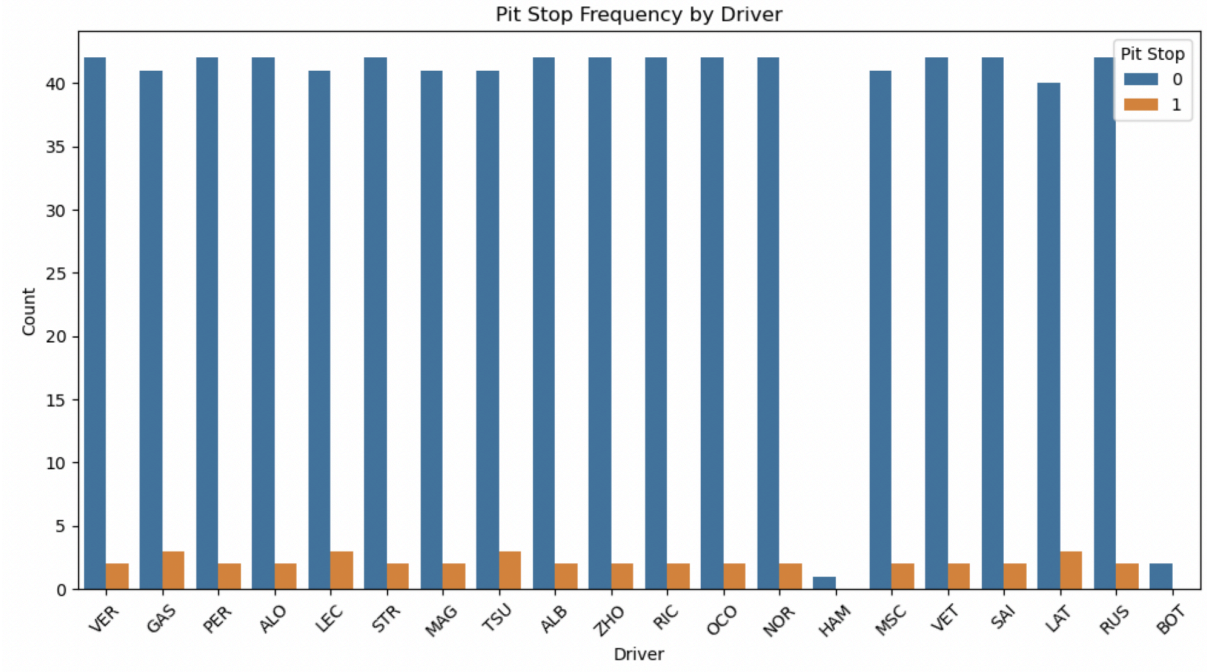
Aya Tarist
ANOP330
Prof. Bailey

- LapTime has a moderately positive correlation with PitStopOccurred (0.51). This indicates that laps with pit stops are likely to have longer lap times, which makes sense as pit stops would add time to a given lap.
- LapNumber and Stint show very little correlation with PitStopOccurred. This suggests that the mere progression of the race (in terms of lap numbers) or the length of the current stint alone are not strong indicators of when a pit stop will occur.
- SpeedI2 shows a strong negative correlation with PitStopOccurred. This implies that higher speeds are less likely to coincide with pit stops, possibly because drivers do not pit while they are at a high speed in this part of the track.
- TyreLife also doesn't show a very strong correlation with PitStopOccurred.I expected tyre life to play a more substantial role in determining pit stops, so this low correlation might be an area to investigate further
- Position has weak correlation of when the PitStopOccurred. This suggests that a car's race position is not a very strong predictor of its pit stop strategy, which could indicate that teams decide on pit stops based on the car's condition and other strategic elements rather than its position alone throughout the race

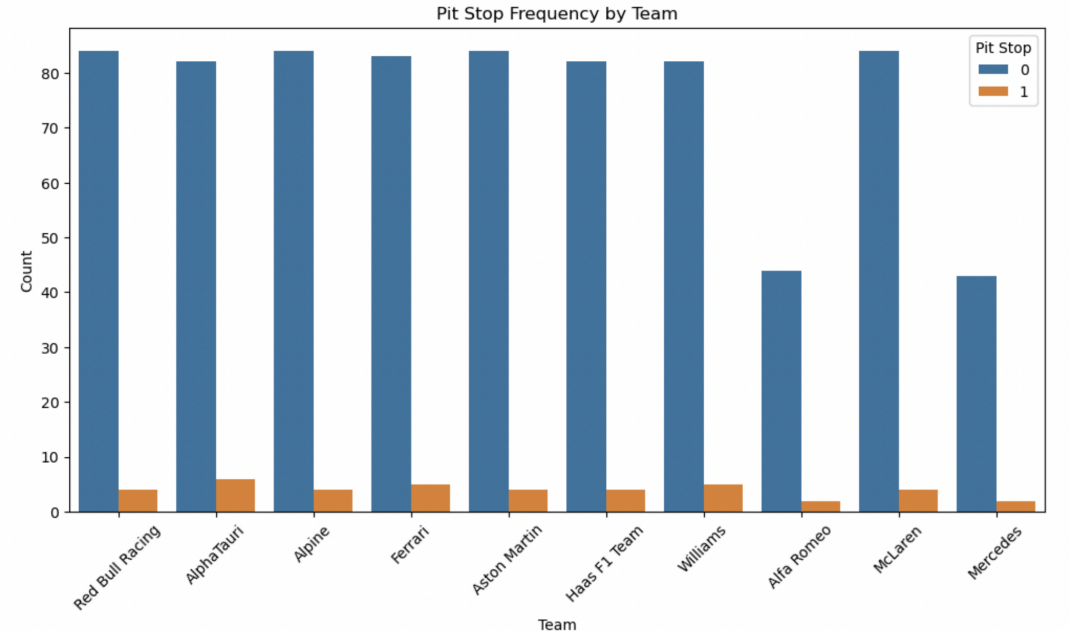


Aya Tarist
ANOP330
Prof. Bailey

- Most drivers have a high number of zero pit stops and a smaller number of one pit stop. This might indicate that most races require only one pit stop per driver, or that the data represents a specific race where most drivers only needed one pit stop.

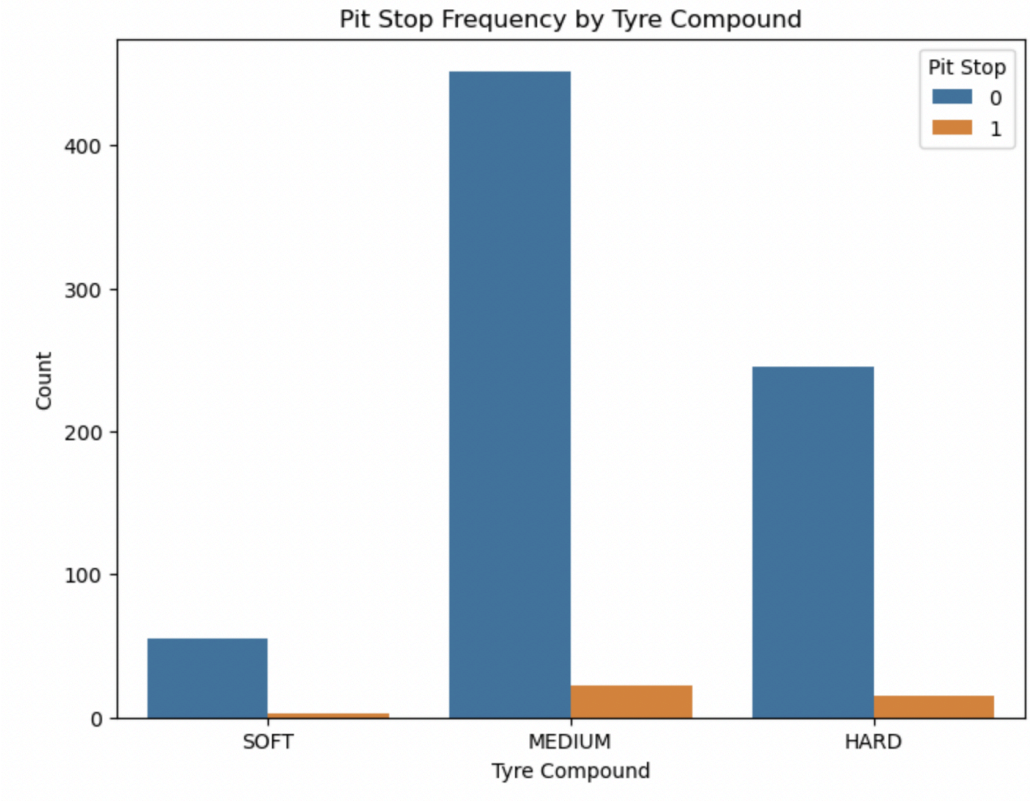


- The two-color scheme is consistent, and it's clear that some teams have more pit stops than others. This could indicate differences in strategy or reliability. Teams with fewer pit stops might have more reliable cars or are better at managing tire wear.

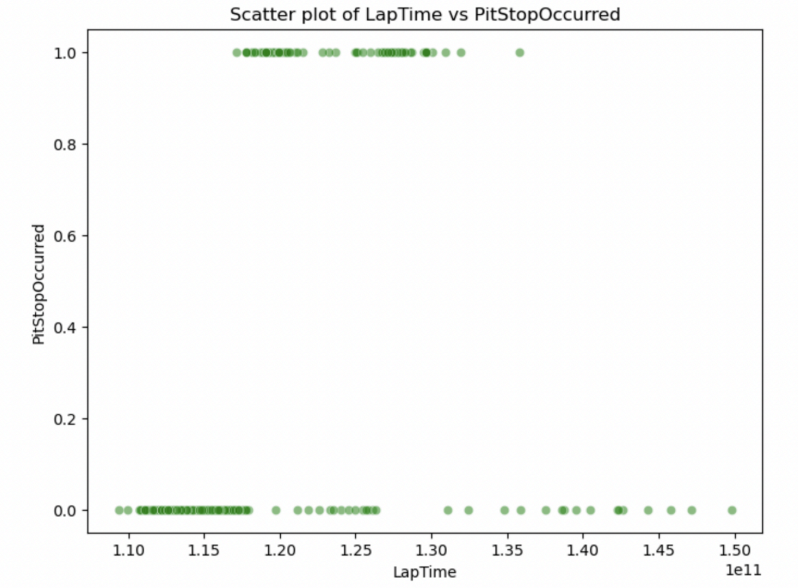


Aya Tarist
ANOP330
Prof. Bailey

- The frequency of pit stops is higher with medium compound tires, which may wear out faster but provide better grip, leading to more frequent changes.

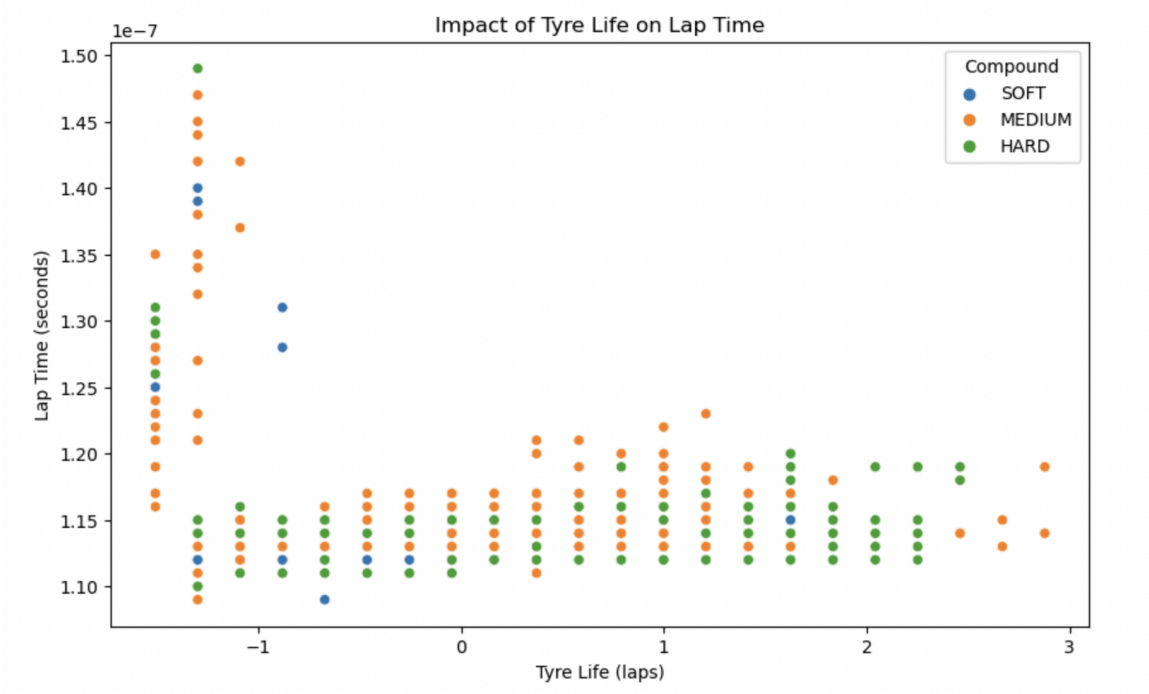


- The absence of any trend suggests that lap time by itself may not be a good predictor of whether a pit stop will happen.



Aya Tarist
ANOP330
Prof. Bailey

- This final scatter plot shows a clear trend where lap times (in seconds) increase with the tire life. Different tire compounds are color-coded. As we may expect, older tires (higher 'Tyre Life' values) generally correspond to slower lap times, reflecting tire wear.
- The spread in lap times for a given tire life indicates different driver skills, car performances, or track conditions.



Next Steps/Conclusion:

The exploratory data analysis provided valuable insights into the factors that could influence pit stop timing in Formula 1 races. There is significant class imbalance with most laps not resulting in pit stops.

The custom feature 'PitStopOccurred' emerged as a key variable.

Weather data, which is crucial in race strategy, is no longer available from Ergast. Alternative methods that I have seen other models use is to merge the weather data of when and where the race happened. This adjustment could be an addition to the EDA before I get into building the model. This is an attempt at having models with more accurate predictions.

Moving forward, the focus will be on building the model using the relevant features (Tyrelife, Speed, Weather).

Aya Tarist
ANOP330
Prof. Bailey

References:

FastF1 Api: https://theoehrly.github.io/Fast-F1/
Github Repo: https://github.com/SpencerStaub/Capstone