**Introduction**  - (3-4 paragraphs all about the area of interest. Do NOT discuss the data in the intro). The intro is non-technical and offers background, objectives, history, value, goals, etc. It must be readable by a person who is 10 or 11.

 "How have different constructors' performances evolved over time in the F1 World Championship, and what factors have contributed to their success or decline?"

The primary objective of this project is to analyze the factors contributing to the success of constructors in Formula 1 racing over the past two decades. Specifically, the aim is to identify key performance indicators (KPIs) that have a significant impact on the constructors' standings. This analysis will help uncover patterns and insights into how constructors can optimize their strategies to improve their performance in future races.

**Analysis**

**Subsection 1: The Data**

*This is where you talk all about the data, the variables, the cleaning, measures of cleanliness, etc.* Even if your data is clean, you must write code to clean it (pretend you cannot see that it is clean).     Minimally: NAs, changing things to factors as needed, outliers, discretization, incorrect values, such as .23 for age, etc.

**Loading the dataset**

**Explanation of Data Loading and Storage Choices**

**Using an "if" Statement for Package Installation**

The "if" statement was used when loading the data to ensure that necessary packages were installed and loaded only if they were not already available. This approach provides several benefits:

1. **Efficiency**: It prevents redundant installations, saving time and computational resources.
2. **Robustness**: It ensures that all required packages are available, minimizing the risk of errors due to missing dependencies.
3. **Seamlessness**: It facilitates smooth execution of the code across different environments, whether packages are pre-installed or not.

**Storing Data in a List**

Storing multiple dataframes in a list instead of having them separately in the environment offers several advantages:

1. **Organization**: It keeps related datasets together in a single object, making it easier to manage and reference them collectively.
2. **Iterative Processing**: It allows for easy application of functions across all datasets using loops or lapply, reducing code duplication and enhancing readability.
3. **Memory Management**: It can be more memory-efficient by avoiding clutter in the global environment, thereby reducing the risk of accidental modifications or deletions.
4. **Dynamic Handling**: It supports dynamic data handling, making it easier to add, remove, or modify datasets within the list without altering the global environment structure.

**Missing Values**

There are no missing values in the cleaned\_circuits dataset, and the data types are appropriate. Extend this check to the rest of the datasets in a single step.

Based on the output, there are no missing values in any of the datasets, and the data types are consistent. Next steps:

Convert Date Columns: Convert any date columns from character to Date type.

Convert Time Columns: Convert any time columns to an appropriate format for analysis.

Handle Categorical Variables: Convert necessary categorical variables to factors.

**Explanation of Variable Changes to Factors**

In the data preparation process, several variables were converted to factors to accurately represent categorical data. This conversion is crucial for the following reasons:

**cleaned\_circuits**

* **circuit\_name**: Converted to a factor to categorize the different circuits used in the races, facilitating analyses such as frequency counts and group-specific computations.
* **circuit\_location**: Converted to a factor to classify the locations of the circuits, aiding in geographical analysis and comparisons.
* **circuit\_country**: Converted to a factor to group the circuits by country, enabling country-specific analyses and visualizations.
* **circuit\_nationality**: Converted to a factor to categorize the nationality of the circuits, supporting analyses related to national distribution and performance.

**cleaned\_constructors**

* **constructor\_name**: Converted to a factor to categorize different constructors, which is essential for comparing their performance and conducting group-wise analyses.
* **constructor\_nationality**: Converted to a factor to group constructors by nationality, facilitating analysis of nationality-based trends and performance.
* **constructor\_home**: Converted to a factor to classify constructors' home locations, aiding in geographical and home-based analyses.

**cleaned\_drivers**

* **driver\_code**: Converted to a factor to categorize drivers uniquely identified by their codes, supporting individual driver analyses.
* **driver\_forename**: Converted to a factor to categorize drivers by their first names, useful for identifying and grouping drivers.
* **driver\_surname**: Converted to a factor to categorize drivers by their last names, aiding in driver-specific analyses.
* **driver\_nationality**: Converted to a factor to group drivers by nationality, enabling nationality-based performance analysis.
* **driver\_home**: Converted to a factor to categorize drivers by their home locations, supporting geographical and home-based analyses.

**cleaned\_status**

* **status**: Converted to a factor to categorize the different status outcomes of races, such as "Finished," "Disqualified," etc. This is essential for analyzing race outcomes and their frequencies.

**Importance of Converting to Factors**

Converting categorical variables to factors is important because:

1. **Categorical Data Representation**: Factors accurately represent categorical data, distinguishing between qualitative differences rather than implying any numeric relationship.
2. **Efficient Memory Usage**: Factors use less memory compared to character variables, as they store data as integer codes with corresponding levels.
3. **Improved Performance**: Many statistical models and functions in R perform better with factors because they can leverage the categorical nature of the data.
4. **Enhanced Analysis**: Factors enable more meaningful statistical analyses and visualizations by treating categories appropriately, facilitating group-specific summaries, comparisons, and visualizations.

By converting these variables to factors, the data is prepared for accurate and efficient analysis, ensuring that categorical information is handled correctly in subsequent analyses and models.

Have a sub-subsection for each variable and show that you looked at all the criteria noted above. Show AND MEASURE the before and after. For example, if you find that a variable has 10 missing values and you update these with the mean, then the before is the mean before, and the after is the mean after. In this case, you should also include the variance before and after. The measure is \*very\* dependent on what you clean and how you clean it. So this will be for you to think about.

**Subsection 2: EDA: statistical and visual. Create a vis and/or table for each variable in the dataset.**

**Exploratory Data Analysis (EDA)**

**Explanation of Excluded Datasets for EDA**

**Excluded Datasets**

**cleaned\_circuits**:

* This dataset contains details about the circuits, such as names, locations, countries, and nationalities. While important for geographical or location-specific analysis, the primary focus of this EDA is on performance trends and factors affecting constructors and drivers, rather than the specifics of the circuits themselves.

**cleaned\_constructor\_results**:

* This dataset includes specific race results for constructors. Since cleaned\_constructor\_standings provides a more comprehensive view of constructor performance over time, including standings and accumulated points, it is more suitable for EDA related to performance trends.

**cleaned\_lap\_times**:

* This dataset contains detailed lap times for each driver in each race. While valuable for in-depth race performance analysis, the focus of the initial EDA is on broader performance trends and factors over multiple seasons, which can be better captured through standings and results data.

**cleaned\_pit\_stops**:

* This dataset records the details of pit stops during races. Similar to cleaned\_lap\_times, it provides granular race-specific data. The EDA aims to explore higher-level performance trends and factors affecting constructors' and drivers' standings, which are better reflected in the standings and results datasets.

**cleaned\_qualifying**:

* This dataset includes qualifying session results. Although qualifying performance can influence race outcomes, the primary interest in the initial EDA is on overall race results and standings, making cleaned\_results and cleaned\_constructor\_standings more relevant.

**cleaned\_seasons**:

* This dataset contains metadata about each season, including the year and a URL. While useful for contextual information, it does not directly contribute to the performance analysis of constructors and drivers, which is the focus of the EDA.

**Focused EDA Datasets**

**cleaned\_constructor\_standings**:

* Provides comprehensive performance data for constructors, including points, positions, and wins over time.

**cleaned\_drivers**:

* Offers detailed information about drivers, enabling analysis of driver demographics and their impact on performance.

**cleaned\_results**:

* Contains race outcomes, including points, positions, and fastest lap times, essential for examining race-specific performance trends.

**cleaned\_races**:

* Provides metadata about each race, allowing for the contextualization of race results within specific seasons and circuits.

By focusing on these datasets, the EDA can effectively analyze broader performance trends and key factors influencing constructors' and drivers' success in the Formula 1 World Championship.

**cleaned\_constructor\_standings**

**cleaned\_drivers**

**cleaned\_results**

**cleaned\_races**

Basic EDA to understand the data better.

Summary statistics for numerical variables.

Distribution plots for key variables.

Correlation analysis for numerical features.

Visualizing trends over time for constructors' performances.

We'll use the following datasets for EDA:

* cleaned\_constructor\_standings
* cleaned\_drivers
* cleaned\_results
* cleaned\_races

Let's perform these analyses step-by-step.

**Summary Statistics for Numerical Variables**

**Subsections 3 – 6+:**

Here you run all methods learned in this class. One subsection for each. Include tuning and/or different options as applicable - such as different kernels and C for SVM, different k for kmeans, etc.

            Explain as you go. Pretend you are writing a tutorial paper.

**Results**

Subsections 1 – 3+

You will have and will discuss results, issues, and limitations for all the analysis. You will note which ones worked well, and why, which ones did not, and why, etc.

**Conclusions:**

3-4 paragraphs - NON-TECHNICAL. What was the outcome - what did you find, discover, predict, classify? WHy does it matter to humans?

**cleaned\_constructor\_standings**

The cleaned\_constructor\_standings dataset provides valuable information on the performance of constructors in each race, including points, positions, and wins. This data is crucial for analyzing trends in constructors' success over time and identifying patterns that contribute to their overall standings in the championship.

**cleaned\_drivers**

The cleaned\_drivers dataset contains detailed information about the drivers, such as their nationality, date of birth, and career details. This dataset is essential for understanding the demographics and background of the drivers, which can be linked to their performance and the constructors they represent.

**cleaned\_results**

The cleaned\_results dataset records the outcomes of individual races, including positions, points scored, laps completed, and fastest lap times. This dataset is fundamental for examining the performance of drivers and constructors in specific races, enabling detailed analysis of race results and performance metrics.

**cleaned\_races**

The cleaned\_races dataset provides metadata about each race, including the year, round, and circuit details. This data is important for contextualizing the race results and constructor standings within specific seasons and circuits, allowing for a comprehensive analysis of performance trends over time and across different locations.

These datasets collectively offer a robust foundation for conducting exploratory data analysis (EDA) and identifying factors that influence the performance and success of constructors in the Formula 1 World Championship.