How Different Variables Affect F1 Race Performance

Group 28 DS 3000, Phase 2

Alexander McMillan, Audrey Romanik, Matthew Chang, Jacob Wu-Chen

Here is our dataset: Formula 1 Dataset

Our first step in this data analysis project is to conduct a data cleaning process. We begin by using the.info() method to identify if the data types assigned to each column are appropriate. Sometimes, numeric data may erroneously be represented as object (string) data, or dates may be stored as strings instead of date-time objects. We want to identify and correct these issues to ensure our data is accurately represented. The data type of a column can influence the choice of imputation method for missing values. For numeric data, mean or median imputation might be suitable, while for categorical data, mode imputation or a separate category for missing values might be more appropriate.

Next, we check for missing data by invoking the .isnull() method. If there are missing values, we need to assess their impact on the dataset. If the percentage of missing values is less than 10%, we use imputation techniques to fill in the missing values. If there is a significant portion of the data missing (greater than 30%), deletion of the rows should be considered. If the decision is made to impute missing values, various techniques can be applied depending on the data type and distribution. For numerical data, common imputation methods include mean, median, or mode imputation. For categorical data, the missing values can be replaced with the most frequent category. Alternatively, if the decision is made to delete missing data, rows containing missing values can be dropped using the .dropna() function. If there are no missing values, great! We can continue with our exploratory data analysis.

The following step is to check for duplicate values using the .duplicated() function. Duplicate values can hamper the accuracy of our data set with Machine learning models. Additionally, duplicate values can be misleading and lead us to wrong patterns and conclusions. Then, we will check if the values for each variable in the dataset are appropriate or not. If the values don’t make sense, we will drop it.

For our last step in this data analysis, we have to prepare the data by turning the nationality of drivers into the name of the country. For example, if a driver's nationality is ‘American’, we want to change it to ‘United States’, which is consistent with the countries in circuits.csv This allows us to compare the home country of drivers with the countries which the races take place in. We will do this by using a dataset, demonyms.csv, which contains the country that every nationality is associated with.

Related Works

related pit stop analysis, race winner predictions, related pit analysis 2

During our research into our dataset, we discovered several studies that utilized the same dataset as us to do analysis work on pit stops and race-winner predictions. These studies though primarily focused on explorative analysis, aiming to find patterns in pit stop times among constructors across different circuits over the years. While our exploratory analysis may have similarities to these pit stop comparisons and race predictions, our approach differs in its ultimate objective. Unlike these previous projects, our goal is to use the insights gained from our exploratory analysis of both constructors' performance and pit stop data, which includes timing and frequency on each circuit, to develop predictive models for when to pit and eventual winners on each circuit.

Reviewing these works has provided us with a valuable perspective on how we can approach our analysis and build upon the existing insights others have gathered to draw our conclusions. Moving forward, it is important for us as a team to differentiate our exploratory analysis from these previous efforts and use more recent data to gain new insights.

https://github.com/knowitall/chunkedextractor/blob/master/src/main/resources/edu/knowitall/chunkedextractor/demonyms.csv