**RaceCraft Prodigy**

**F1 Winner Predictor**

**GROUP NAME**

Idriss Kadri-Hassani, Jacob Grayzel, Luna Dana, Raphael Barthes

**Columbia University**

ik2552@columbia.edu | jjg2212@columbia.edu | lad2215@columbia.edu | rpb2141@columbia.edu

# **1. Abstract**

# 

Formula 1 involves high-speed races with teams designing and building their own advanced cars. Race weekends include practice, qualifying, and the race itself, with points awarded for finishing positions and additional factors. Each team has two drivers, and strict regulations govern on-track behavior and technical specifications. Predicting race results is crucial for teams, utilizing data science to analyze car performance, driver skills, and strategies. The competitive nature of F1, combined with data-driven decision-making, enhances fan engagement and contributes to technological advancements in motorsport. There are multiple Grand Prix across the globe.

We begin by importing and exploring datasets, ensuring data integrity and identifying key variables. Descriptive statistics reveal correlations between manufacturer performance, grid/finish rankings, and driver nationality's impact on race outcomes. The modeling phase involves data preparation, classification, and regression. Classification algorithms predict race outcomes categorically, while regression techniques forecast specific metrics like finishing positions. The final step compares the strengths and weaknesses of both models. This systematic process aims to uncover patterns, enhance predictive accuracy, and provide valuable insights into Formula 1 race dynamics.

We can look at the notebook for more details and insights.

# **2. Data Scraping and Cleaning**

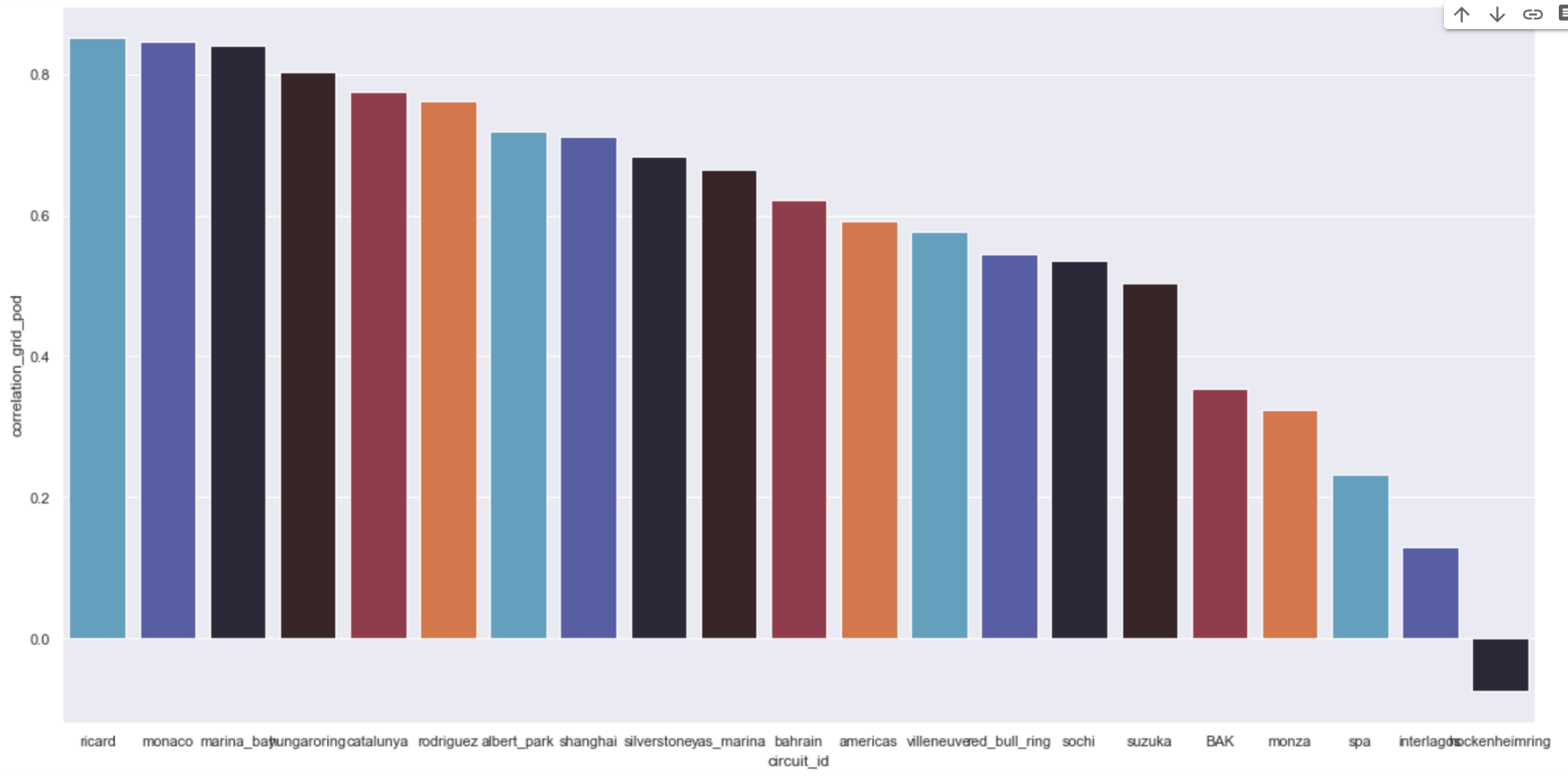
# 

We first fetch Formula 1 race data from the Ergast API for the years 1990 to 2021, extracting relevant information. Here are the different dataset:

| Dataframe | Attributes | Description |
| --- | --- | --- |
| race | season, round, circuit\_id, lat, long, country, date, url | Information about each Formula 1 race, including location details (latitude and longitude), country, date, and circuit information. This dataset is crucial for contextualizing race events, considering geographical factors, and understanding the characteristics of each circuit, all of which contribute to predicting race outcomes and strategies. |
| results | season, round, circuit\_id, driver, date\_of\_birth, nationality, constructor, grid, time, status, points, podium | Detailed race results, including driver and constructor information, grid position, race time, status, points, and podium position, crucial for analyzing individual and team performance. |
| driver\_standings | 'season', 'round', 'driver', 'driver\_points', 'driver\_wins', 'driver\_standings\_pos' | Cumulative driver standings across seasons, providing insights into the consistent performance of drivers over time, which is essential for predicting future race outcomes. |
| constructor\_standings | season, round, constructor, constructor\_points, constructor\_wins, constructor\_standings\_pos | Similar to driver standings, but focused on constructor (team) performance, offering valuable information for predicting team success and analyzing trends in team competitiveness. |
| qualifying\_results | grid, driver\_name, car, qualifying\_time, season, round | Information about drivers' grid positions and qualifying times, providing insights into individual driver performance during the qualifying sessions, which can influence race outcomes. |
| weather\_info | season, round, circuit\_id, weather, weather\_warm, weather\_cold, weather\_dry, weather\_wet, weather\_cloudy | Weather conditions for each race, which can have a significant impact on race strategies and outcomes. Relevant for understanding how weather influences race dynamics and aiding predictions based on weather forecasts. |

# **3. Exploratory Data Analysis: study of the link between our variables and victory**

# **3.1 Correlation between Start Grid and Finish Ranking**

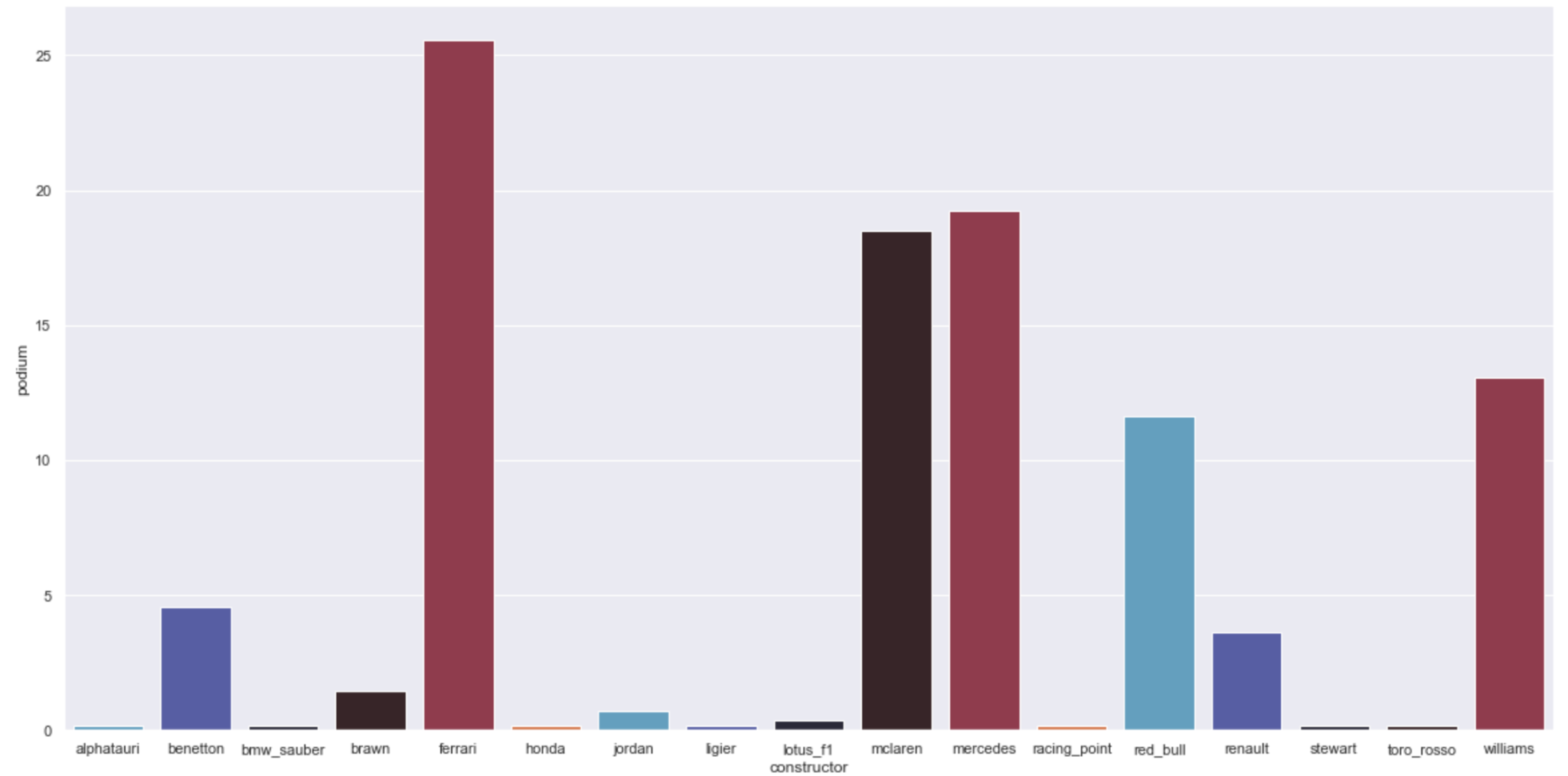


Formula 1 race data initially showed no clear correlation between start grid positions and podium standings across all years, likely due to dataset complexity. A focused exploration in 2020 revealed a subtle linear trend, leading to the creation of the corr\_grid\_podium function. Illustrated with 2019 data, the function systematically calculates and displays start grid–podium correlations for each circuit. The resulting bar plot provides insights into a notable relationship, with correlations exceeding 80% in certain circuits.

# **3.2 Correlation between manufacturer and grid and finish ranking**

Here is a summarized table showcasing the average podium and grid positions for each Formula 1 manufacturer, providing insights into their performance relative to starting positions and final podium outcomes.

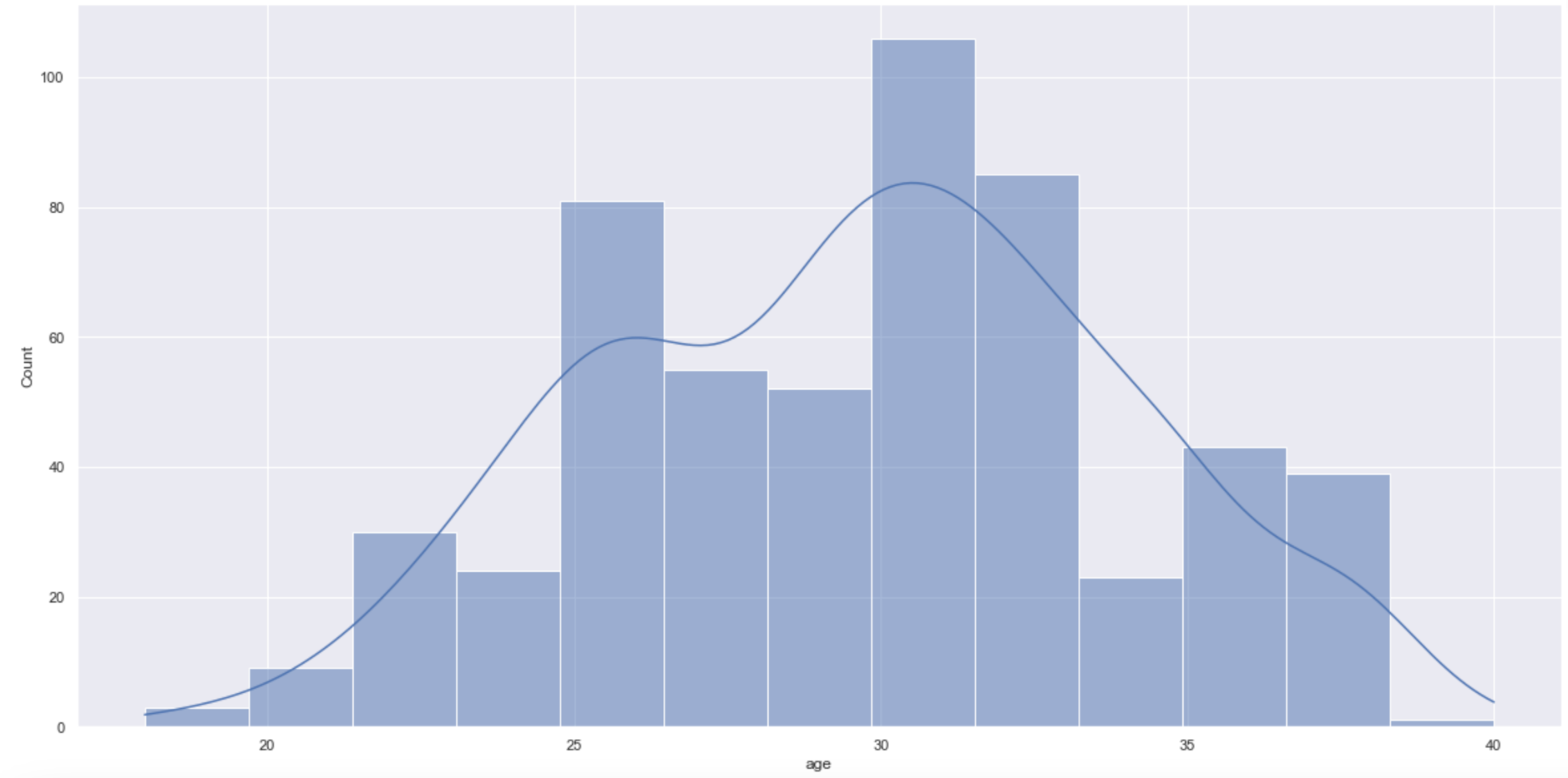
| Constructor | Average Podium | Average Grid |  | Constructor | Average Podium | Average Grid |
| --- | --- | --- | --- | --- | --- | --- |
| Mercedes | 5.77 | 4.61 |  | Lotus Racing | 17.97 | 18.77 |
| Ferrari | 7.43 | 5.72 |  | Virgin | 19.0000 | 21.22 |
| Red Bull | 8.00 | 6.98 |  | Lola | 19.67 | 17.89 |



Here again, without much surprise, some manufacturers stand out from the crowd, such as Ferrari, McLaren and Mercedes. There's a huge disparity between the manufacturers, so each one contains a signal that can be used to predict the final victory

# 

# **3.3 Correlation between driver and final victory**

We then turn to the link between nationality and victory. We're going to find out whether this variable can explain or predict final victory. We'll simply plot the percentage of victories by nationality. Unsurprisingly, the drivers with the most Grand Prix titles are German (Vettel, Schumacher, Rosberg), English (Hamilton), Brazilian (Senna) and Finnish (Räikkönen). There is a disparity between nationalities, which means that it is an important factor.

Moreover, there's a slight negative correlation between age and ranking (-0.055): in other words, the older you are, the lower your ranking and the closer you are to victory.

Moreover, the number of victories per age is plotted, along with an approximation of the density which shows an interesting trend of driver consistency over time or effect of age on victory.

# **3.4 Weather analysis**

Since there are 363 different types of tense, we look at the most common ones and analyze the correlation between grid and podium on wet and dry circuits. When the circuit is dry, like in Bahrain and the Hockenheimring, there's a strong correlation between grid position and final podium finish. On these dry circuits, a good grid position virtually guarantees a final podium finish, and this correlation is stronger than when climate is not taken into account. On wet circuits, we observe the same pattern.

**4. Modeling**

# **4.1 Data Preparation**

In this phase, we focused on preparing our data for the modeling process. Given the complexity and variability of Formula 1 data, this step was crucial. We utilized Python libraries such as Pandas for data manipulation and Scikit-learn for preprocessing.

We transformed raw data into features suitable for modeling. This involved creating new variables that could better capture the nuances of race dynamics, such as aggregated historical performance metrics for drivers and constructors. We employed StandardScaler and MinMaxScaler for normalizing our data. This was essential to ensure that features with larger scales did not unduly influence the model's performance. The dataset was split into training and testing sets, maintaining a balance between exposing the models to enough data and retaining an unbiased subset for evaluation.

Our approach aimed to create a comprehensive dataframe containing all variables deemed useful for modeling, along with our target variable, i.e., the race winner or a proxy such as the race ranking. This process involved several critical steps:

**Merging Dataframes:** We prioritized the results dataframe as the foundation for merging due to its comprehensiveness. Our strategy was to sequentially merge other dataframes, like qualifying\_results, constructor\_standings, and driver\_standings, with results. This approach ensured a holistic view of each race, incorporating various aspects like driver performance, constructor standings, and qualifying results.

**Handling Missing Values:** Given the nature of our datasets, dealing with missing values was essential. We employed techniques such as imputation and exclusion of certain records where necessary, ensuring the integrity and usability of our data.

Our data contained numerous categorical variables, such as driver names and constructor teams. We transformed these into a format suitable for modeling, including encoding techniques to convert categorical data into numerical formats that our models could process effectively.

**Creating New Variables:** An essential part of our data preparation was the creation of new variables that could provide additional predictive power. We calculated the age of drivers to the nearest year, considering the potential impact of experience and physical condition on race performance. The qualifying\_time variable was created to capture the cumulative difference in qualifying times from the pole position. This variable aimed to quantify the advantage or disadvantage a driver had at the start of the race based on their qualifying performance.

# **4.2 Model Selection and Implementation**

Our choice of models was guided by the nature of our predictive tasks – both categorical (classification) and continuous (regression).

**Neural Network Models:**.For classification, we opted to use a Support Vector Machine (SVM) model, a popular choice for solving classification problems. To optimize the model, we experimented with different parameter combinations, despite the computational intensity of this process.

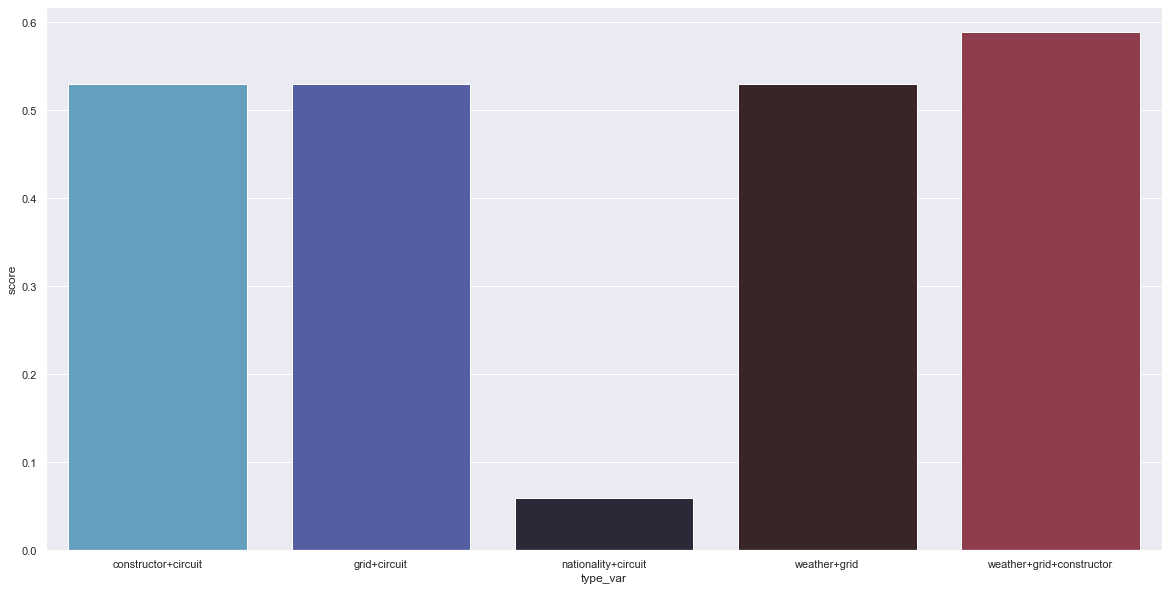
**Linear Regression Models:** Considering race ranking as a continuous variable, we treated our problem as a regression task, aiming to predict the winner as the one with the lowest ranking.We implemented linear regression for its suitability in modeling continuous outcomes and interpreting relationships between variables.

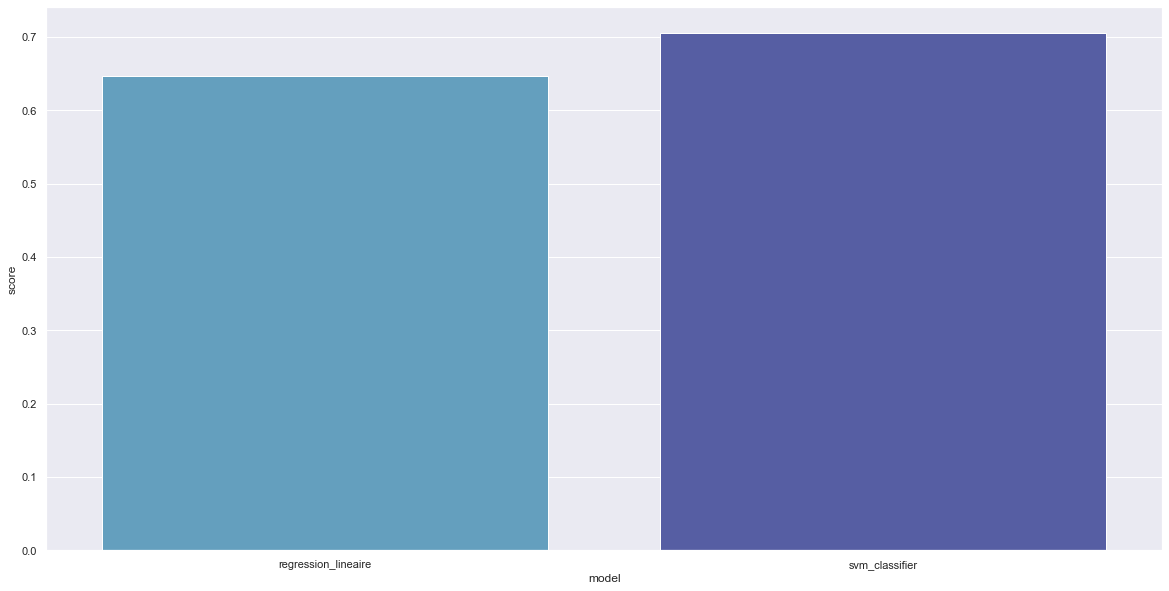
# **4.3 Training and Validation**

We trained our models using the prepared datasets, ensuring that they learned to predict race outcomes from a wide range of input variables. We employed techniques like batch normalization and dropout for neural networks to improve convergence and reduce overfitting. We employed cross-validation techniques to assess the generalizability and robustness of our models.

# **4.4 Model Evaluation**

For classification tasks, we used metrics such as accuracy, precision, recall, and F1-score. For regression tasks, we employed mean squared error (MSE) and R-squared.To evaluate the performance of our model, we decided to train our model on the years prior to 2020 and measure the performance of our model as the percentage of races "well" predicted in 2020. There are other approaches, but we chose the latter because it allows us to intuitively assess the relevance or otherwise of the model over a recent year.

**5. Results**

With this project, we want to predict the winner of an F1 grand prix. We proceeded by classification (SVM classifier), considering that the winner is in category 1 and the loser in category 0, and by regression (linear regression) the ranking can be seen as a continuous variable, so the winner will be the one with the lowest ranking prediction. We will regress on different combinations of variables to understand which types of variables best predict the race winner (Some combinations of variables are better at predicting the final result. For example, the constructor+circuit combination scores 52%, while the nationality+circuit combination scores 5%.).

We note that linear regression scores 65% versus 70% for the SVM classifier.

However, we note that linear regression runs much faster than the SVM classifier.