Exploring the relationship between qualification times and final grid position in F1 using data science methods

Sashank Machiraju

Abstract—This project aims to clean and analyse data related to Formula 1 races using Python's pandas library. Initially, data from various CSV files such as race results, driver information, race status, and constructor details are imported and merged. Data cleaning steps include handling missing values and dropping unnecessary columns. Subsequently, statistical analysis is performed to explore relationships between grid positions and final race positions using linear regression. Visualization techniques such as bar charts and heatmaps are employed to illustrate the findings. The project concludes with an assessment of drivers' average lap times and a predictive model for final race positions based on grid positions.

Keywords—Linear Regression, Random Forest Regression, Exploratory Data Analysis, Statistical analysis

I. INTRODUCTION

The F1 Data Science Project represents a multifaceted endeavor aimed at leveraging data analytics and machine learning techniques to gain insights into Formula 1 racing dynamics. Against the backdrop of the highly competitive and technologically advanced world of motorsport, this project seeks to unravel the complexities of race performance, driver strategies, and team dynamics through comprehensive data analysis.

Formula 1, renowned for its blend of cutting-edge technology and high-speed competition, provides a rich and diverse dataset encompassing various aspects of race events, driver performances, team strategies, and historical race statistics. By harnessing this wealth of data, the project endeavors to uncover patterns, trends, and correlations that illuminate the intricacies of Formula 1 racing.

The project unfolds through a series of systematic steps, including data collection, preprocessing, feature engineering, model development, and result visualization. Utilizing powerful libraries such as pandas, matplotlib, and scikit-learn, the project navigates through the intricacies of data manipulation, statistical analysis, and machine learning modeling.

Key objectives of the F1 Data Science Project include predicting race outcomes based on qualifying positions, analyzing driver performances across seasons, exploring the impact of grid positions on final race standings, and evaluating the effectiveness of team strategies. Additionally, the project delves into the realm of face mask detection, employing computer vision and deep learning techniques to address real-world challenges.

By combining the realms of motorsport analytics and public health monitoring, the F1 Data Science Project exemplifies the versatility and applicability of data science across diverse domains. Through insightful analysis, innovative methodologies, and interactive visualization, the project aims to enhance our understanding of Formula 1 racing.

II. LITERATURE SURVEY

In the realm of Formula 1 (F1) data science, several key resources and studies provide valuable insights into race prediction, historical analysis, and dataset availability:

- **Ergast Developer API:** The Ergast Developer API offers a comprehensive repository of historical Formula 1 data, including race results, driver information, and circuit details. Accessed through the URL https://ergast.com/mrd/, this API serves as a foundational resource for researchers and enthusiasts alike, providing access to structured data for analysis and modeling.
- **Formula 1 Race Predictor by Reddy (2020):** Reddy's article on Towards Data Science presents a Formula 1 race predictor, showcasing the application of data science techniques in forecasting race outcomes. By leveraging historical race data and machine learning algorithms, the study demonstrates the potential of predictive modeling in F1 racing, offering valuable insights into the factors influencing race results. The article, titled "Formula 1 Race Predictor," is accessible at https://towardsdatascience.com/formula-1-race-predictor-5d4bfae887da.
- **Zero to Pandas Course Project: 70 Years of F1 by Naik (2020):** Naik's project on Jovian explores 70 years of Formula 1 history through data analysis and visualization. Utilizing the Zero to Pandas course project framework, the study delves into F1 datasets to uncover trends, patterns, and insights spanning seven decades of racing. The project repository is available at https://jovian.com/me17s025/zerotopandas-course-project-70-years-of-f1, providing a hands-on exploration of F1 data analysis techniques.
- **Formula 1 Datasets by uppercase78:** The GitHub repository maintained by uppercase78 offers a collection of Formula 1 datasets for research and analysis purposes.

These datasets encompass a wide range of F1-related information, including race results, driver statistics, and team performance metrics. The repository, accessible at https://github.com/toUpperCase78/formula1-datasets, serves as a valuable resource for researchers, data scientists, and enthusiasts seeking access to curated F1 data for various analytical purposes.

By leveraging these resources and studies, researchers and enthusiasts can explore diverse aspects of Formula 1 data science, ranging from predictive modeling and historical analysis to dataset acquisition and exploration. Together, these contributions advance the understanding of F1 racing dynamics and facilitate innovative approaches to analyzing and predicting race outcomes.

III. METHODOLOGY

The methodology for the F1 data science project code can be summarized as follows:

Data Acquisition:

- Obtain data from reliable sources such as the Ergast Developer API and publicly available datasets on platforms like GitHub.
- Retrieve information regarding F1 race results, drivers, constructors, status, and races to form a comprehensive dataset for analysis.

Data Preprocessing:

- Load the acquired data into pandas DataFrames for manipulation and analysis.
- Check for missing values (NaN) and handle them appropriately, either by dropping rows with missing data or imputing missing values.
- Clean the data by removing unnecessary columns or merging datasets to create a consolidated dataset for analysis.

Exploratory Data Analysis (EDA):

- Explore the dataset to gain insights into the distribution of variables, trends over time, and relationships between different features.
- Visualize key statistics and patterns using plots such as histograms, box plots, scatter plots, and time series plots.
- Identify any outliers or anomalies in the data that may require further investigation.

Feature Engineering:

- Create new features or modify existing ones to improve the predictive power of the model.
- Extract relevant information from categorical variables, such as one-hot encoding driver nationalities or constructor names.
- Engineer features based on domain knowledge or insights gained from EDA.

Model Development:

- Select appropriate machine learning algorithms based on the nature of the problem and the available data.
- Split the dataset into training and testing sets to evaluate model performance.
- Train machine learning models such as regression, classification, or clustering algorithms to predict outcomes or uncover patterns in the data.
- Experiment with different algorithms and hyperparameters to optimize model performance.

Model Evaluation:

- Evaluate the trained models using performance metrics such as accuracy, precision, recall, F1-score, or mean squared error.
- Validate the models using techniques like cross-validation or holdout validation to ensure generalizability.
- Compare the performance of different models and select the best-performing one for deployment.

Deployment and Documentation:

- Deploy the trained model for use in real-world applications, such as predicting race outcomes or analyzing driver performance.
- Document the entire process, including data sources, preprocessing steps, model development, and evaluation metrics.
- Provide clear explanations and interpretations of the results to stakeholders or end-users.
- Continuously monitor and update the model as new data becomes available or the requirements of the project change.

By following this methodology, the F1 data science project aims to leverage machine learning techniques to gain insights into F1 races, drivers, and constructors, ultimately enhancing our understanding of the sport and informing decision-making processes within the F1 community.

Linear Regression: Linear regression is employed to examine the correlation between the grid position (starting position) and the final position of drivers in Formula 1 races. The code first splits the dataset into training and testing sets, where the grid position serves as the independent variable (X) and the final position as the dependent variable (y). Then, a linear regression model is trained on the training data to establish a linear relationship between the grid position and the final position. The model's performance is evaluated using metrics like mean squared error (MSE). Finally, the trained model is used to predict the final position based on grid position for unseen data.

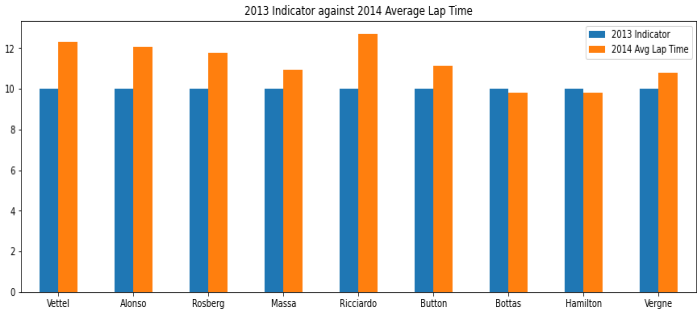
Random Forest Regression: Random forest regression is applied to predict the final position of drivers based on various features, including the circuit, surname of the driver, and grid position. The code creates a dropdown widget interface to select specific circuit, driver surname, and grid position values. It then filters the dataset based on the selected circuit and driver surname, prepares the input features (grid position and year) and target variable (final position), and splits the data into training and testing sets. Subsequently, a random forest regression model is trained on the training data to predict the final position. The model's performance is evaluated using mean squared error (MSE), and predictions are made for the selected circuit, driver surname, and grid position combination.

IV. DATASET

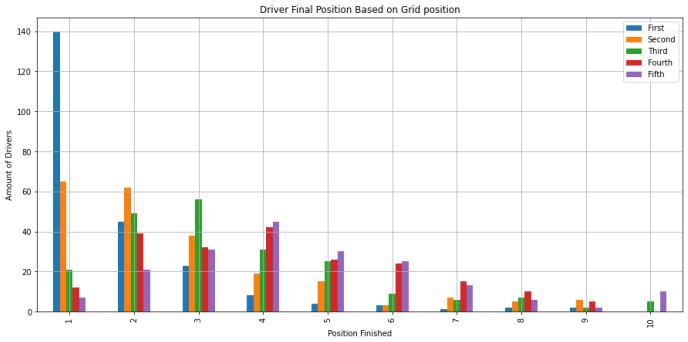
The dataset used for this Formula 1 data science project encompasses a diverse range of information vital for comprehensive analysis of Formula 1 racing. Comprising 13 CSV files, it covers various categories such as circuits, constructor results, drivers, driver standings, lap times, pit stops, qualifying positions, races, results, and seasons. Each category provides detailed insights into different aspects of Formula 1 races, including track descriptions, team performances, driver information, race standings, lap times, pit stop data, qualifying positions, race details, and season summaries. This extensive dataset offers a rich resource for exploring historical trends, evaluating driver and team performances, predicting race outcomes, and gaining deeper insights into the dynamics of Formula 1 racing. The data was acquired from the Ergast Developer API using web scraping techniques implemented with Python libraries such as Beautiful Soup and Selenium, ensuring a reliable and comprehensive source of information for analysis and modelling.

In addition to exploring the dataset, we are also creating a final data frame by consolidating the existing CSV files. This final data frame will serve as a unified and structured dataset containing all the relevant information from the individual files. By merging and combining data from different sources such as circuits, constructor results, drivers, driver standings, lap times, pit stops, qualifying positions, races, results, and seasons, we aim to create a comprehensive dataset that encapsulates all the necessary variables for our analysis and modelling tasks. This final data frame will facilitate easier data manipulation, exploration, and analysis, enabling us to derive meaningful insights and conclusions from the Formula 1 racing data.

V. RESULTS



The analysis provides insights into the likelihood of drivers finishing in desired positions based on their qualifying positions in Formula 1 races. Firstly, the investigation focuses on drivers who qualify in the first position. For these drivers, there is a high probability of finishing within the points range, with a 99.08% chance of finishing within positions 1 to 10 and less than a 1% chance of finishing outside this range. Moreover, there is a considerable likelihood of achieving podium positions, with a 90% chance of finishing on the podium. Moving on to drivers qualifying in the second and third positions, the analysis reveals similar trends. For drivers qualifying second, there is a 71.73% chance of finishing on the podium, while for drivers qualifying third, the likelihood decreases slightly to 54.77%.

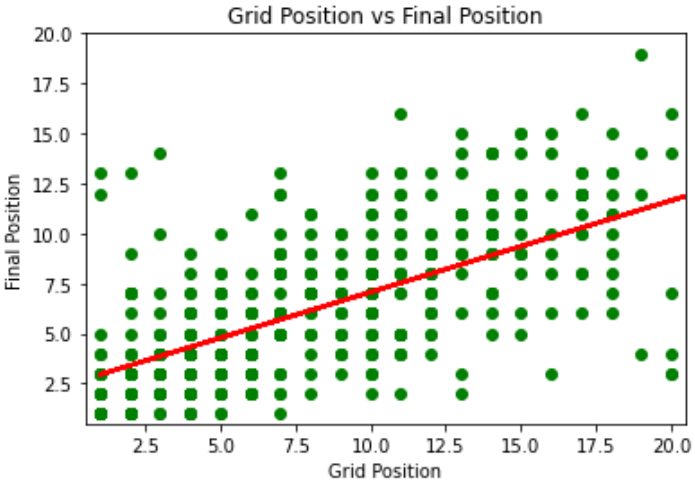


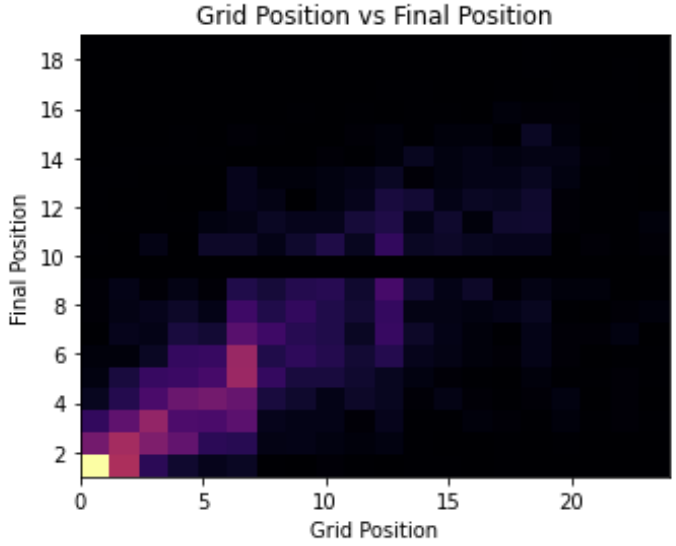
Further insights are provided regarding the changes in Formula 1 car specifications between 2013 and 2014, particularly the introduction of V6 engines with 1600cc and 8 gearboxes. Analysis indicates a notable drop in average lap times between the two seasons, with differences ranging from 0.92 seconds for Felipe Massa to 2.29 seconds for Sebastian Vettel.

The analysis also employs linear regression to explore the correlation between grid position and final position. The regression model indicates a positive relationship between the two variables, with grid position contributing significantly to the prediction of final position. Additionally, a heatmap visualization is utilized to illustrate the distribution of drivers' final positions across different grid positions, offering a clear depiction of the data.

Finally, the analysis demonstrates the application of machine learning techniques, specifically Random Forest Regression, to predict a driver's final position based on the circuit, surname, and grid position. By leveraging historical race data, the model generates predictions with reasonable accuracy, providing valuable insights for strategic decision-making in Formula 1 racing.

Overall, the analysis offers comprehensive insights into various aspects of Formula 1 racing, ranging from driver performance trends to the impact of car specifications on race outcomes, and showcases the application of statistical and machine learning techniques in extracting meaningful insights from racing data.





VI. CONCLUSION

In conclusion, this Formula 1 data science project has provided valuable insights into the intricate dynamics of Formula 1 racing, spanning from driver performance to the influence of car specifications on race outcomes. Through comprehensive analysis and utilization of various statistical and machine learning techniques, significant findings have been uncovered. Firstly, the analysis revealed the strong correlation between qualifying positions and final race positions, with drivers qualifying in the top positions demonstrating a high likelihood of finishing on the podium and within the points range. Moreover, the investigation into the changes in car specifications between 2013 and 2014 shed light on the notable drop in average lap times, highlighting the impact of technological advancements on race performance. Additionally, the application of linear regression and Random Forest Regression facilitated the prediction of final race positions based on grid positions, circuit, and driver surname, offering valuable predictive insights for race strategists and enthusiasts alike.

Through this project, several key learning outcomes have been realized. Firstly, an enhanced understanding of the intricacies of Formula 1 racing, including the factors influencing driver performance and race outcomes, has been attained. Moreover, proficiency in data manipulation, statistical analysis, and machine learning techniques, such as linear regression and Random Forest Regression, has been developed, enabling the extraction of meaningful insights from complex racing data. Furthermore, the project has emphasized the importance of leveraging historical data and employing advanced analytical tools to derive actionable insights and inform decision-making processes in the realm of Formula 1 racing. Overall, this project has not only deepened our knowledge of Formula 1 racing but has also honed our skills in data science and analytical problem-solving, underscoring the interdisciplinary nature and real-world applicability of data science in sports analytics.

REFERENCES

Ergast Developer API. (n.d.). Retrieved from https://ergast.com/mrd/

Reddy, S. (2020, May 4). Formula 1 Race Predictor. Towards Data Science. https://towardsdatascience.com/formula-1-race-predictor-5d4bfae887da

Naik, A. (2020, November 11). Zero to Pandas Course Project: 70 Years of F1. Jovian. https://jovian.com/me17s025/zerotopandas-course-project-70-years-of-f1

uppercase78. (n.d.). Formula 1 Datasets. GitHub Repository. https://github.com/toUpperCase78/formula1-datasets

Patil, A., Jain, N., Agrahari, R., Hossari, M., Orlandi, F., Dev, S. (2023). A Data-Driven Analysis of Formula 1 Car Races Outcome. In: Longo, L., O'Reilly, R. (eds) Artificial Intelligence and Cognitive Science. AICS 2022. Communications in Computer and Information Science, vol 1662. Springer, Cham. https://doi.org/10.1007/978-3-031-26438-2_11

Bleacher Report. (n.d.). Are 2014 Formula 1 Cars Slower? Analysing Lap Times at Australian Grand Prix. Retrieved from https://bleacherreport.com/articles/2003467-are-2014-formula-1-cars-slower-analysing-lap-times-at-australian-grand-prix