From Data to Podium: Analytical Approaches to Understanding and Predicting Driver & Strategic Decisions in Formula 1

Hypothesis Testing: The Effect of Second Driver Pit Stop Timing on Race Outcomes.

A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

Work project carried out under the supervision of:

Michail Batikas

20-12-2023

Abstract

This thesis explores Formula 1 pit stop strategies through advanced analytics, with a focus on driver clustering in relation to performance, tactical, and behavioural aspects. Our approach led to the identification of four distinct driver categories, providing a framework to investigate various pit stop strategies. By integrating these driver profiles into hypothesis tests, the study delves into the impact of driver characteristics on team strategy and pit stop efficiency. This research contributes to a more refined understanding of strategic elements in Formula 1, demonstrating the role of tailored analytic methods in optimizing racing tactics and decision-making processes.

Keywords

Data, Data Analytics, Sports Analytics, Formula 1, Strategy, Hypothesis Test, Prediction Model, Machine Learning

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

1. Table of Contents

Tables of Figures

List of Tables

Group Part

2. Introduction

In the competitive world of sports, the strategic use of analytics has revolutionised how competition is understood and won. This transformation is particularly evident in Formula 1, a sport where the smallest decisions can have significant impacts; leveraging data effectively is not merely an advantage but a cornerstone of success. Central to these strategies are the critical decisions made during pit stops, which can significantly influence race outcomes. In this context, the role of advanced analytics has become increasingly prominent, offering new avenues to optimize these crucial decisions.

While research has been conducted on various aspects of Formula 1 strategy, a notable gap exists in exploring driver clustering and its potential impact on strategic decisions as well as the evaluation of analytical approaches. This thesis is motivated by the prospect of harnessing advanced analytics and driver clustering to add additional insights to pit-stop strategies. The aim is to explore how a refined understanding of driver behaviour can enhance race outcomes and the interpretation of analytical outcomes when integrated with data-driven approaches.

This study is guided by the research question: "*How can the application of advanced analytics and driver clustering in Formula 1 enhance the accuracy and effectiveness of strategic decision-making, particularly in the context of optimising pit stop strategies?*" The objectives are twofold: firstly, to develop a robust clustering model based on performance, tactical, and behavioural metrics; and secondly, to use the results of this model within analytical frameworks, like prediction models and hypothesis tests, to assess its impact on pit stop timing, tire selection, and other strategic decisions.

Our methodology centres around leveraging data mainly from the FastF1 API, focusing on recent seasons to construct a clustering model, providing a multifaceted perspective on driver profiles.

Kommentiert [VF1]: Verwenden wir das Clustering wirklich innerhalb der Modelle oder nutzen wir das Clustering, um bestimmte Modelle zu bewerten? Für die Verteidigung müssen wir uns wirklich auf ein bestimmtes Ziel einigen, das wir mit dem Clustering erreichen wollen. Warum war es wichtig, das Clustering durchzuführen? Was war der Vorteil einer Clusterung im Vergleich zu einer fehlenden Clusterung? usw.

The methodology extends to applying this clustering within different analytical models and hypothesis tests, aiming to shed light on the intricacies of pit-stop strategies and potentially add optimisations or additional insights.

This research aims to contribute significantly to the fields of sports analytics and Formula 1 strategic planning. By offering a novel perspective on driver behaviour and its implications for race strategy, the findings of this study could potentially guide teams and drivers towards more informed and effective decision-making. The insights gained could not only enrich academic discourse but also provide practical tools for strategic optimisation in the high-pressure environment of Formula 1 racing.

To lay the foundation for our analysis, we begin with a comprehensive literature review that covers the broader spectrum of sports analytics, with a specific focus on data analytics in Formula 1. This review will also discuss the application of clustering in other sports, noting its relatively nascent presence in Formula 1 literature. Following this, we detail our methodology, including our full data collection process and the clustering of Formula 1 drivers based on their performance metrics. In subsequent chapters, we examine various aspects of pit stop strategy—such as pit timing, tire choice, driver selection, and the impact of safety cars—and explore how these elements interact with the identified driver behaviour clusters.

3. Literature Review

Evolution and Scope of Sports Analytics

Sports analytics refers to the application of scientific techniques for investigating and modelling sports performance. It entails organising historical data in a structured manner, applying predictive analytical models to the data, and utilising information systems to inform decision-makers

(Morgulev, Azar, and Lidor 2018). This practice enables sports organisations to secure a competitive advantage through enhanced player performance analysis, game strategy formulation, health and injury prevention, fan engagement, and operational strategy optimisation (Nadikattu 2020; Tan 2023). Originating in the 1960s with notational analysis in sports like American Football and Basketball (Hughes and MFranks 2004), sports analytics has evolved significantly with technological advancements and the rise of Big Data. Today, its application extends across various sports domains, from individual sports like Tennis to team sports such as Football and has become central in the data-driven world of motorsports like Formula 1. This wide adoption underscores the broad reach and applicability of sports analytics across diverse sporting disciplines (Bai and Bai 2021).

As technology evolved, sports analytics underwent a significant transformation, unveiling new avenues for analysing and improving sports performance. A recent comprehensive review by Ghosh et al. (2023) classifies the technological advancements in sports analytics into three primary research fields: sensors, computer vision, and wireless and mobile-based applications. These areas form the bedrock of modern sports analytics, providing essential methods to collect, analyse, and interpret data. These technological advancements prove particularly pertinent in Formula 1, a sport renowned for its data-driven approach. Incorporating hundreds of sensors in racing cars facilitates real-time data collection, covering various parameters such as speed, temperature, and throttle percentage (Shapiro 2023). Effective analysis of this data offers invaluable insights for making informed decisions on pit stop strategies, car setups, and race strategies. The inclusion of artificial intelligence (AI) and machine learning (ML) algorithms amplifies the potential of these technological advancements, enabling more sophisticated analysis and predictive modelling

(Ghosh et al. 2023; Dindorf et al. 2023). Integrating these novel technologies with sports analytics methodologies has opened and revolutionised research opportunities in Formula 1 racing.

Analytical Approaches in Formula 1 Racing

As established in the preceding section, sports analytics is central to enhancing competitive strategies across various sports disciplines. This becomes particularly evident in Formula 1, a sport where even the minutest decision can significantly alter race outcomes. In F1, where the stakes are high and the financial implications are vast, the synergy between data analytics and elite sporting performance exemplifies the comprehensive capabilities of sports analytics. The body of literature directly engaging with sports analytics in the context of Formula 1 is limited both in terms of the quantity of research and the range of thematic focus. Broadening the search parameters to include 'Circuit Racing Motorsports'—thereby encompassing NASCAR, Formula E, and similar series—yields an expanded body of work. Preliminary examination suggests that while there has been increasing interest in this field, it appears that the field has not yet reached a point of saturation, indicating sufficient opportunity for further research and contribution.

The existing literature can be divided into several overarching categories, however, two are predominant: lap time simulations and race simulations. As Heilmeier (2018) highlights, it is crucial to differentiate between race simulations and the more prevalent lap time simulations. The latter predominates in the literature and typically focuses on the physical or engineering aspects rather than on the holistic view of an entire race. Siegler (2000) identifies three distinct approaches to lap time simulations: Steady State, Quasi-Static, and Transient. Heilmeier (2019) published a study on quasi-static lap time simulation, applying it to both Formula 1 and Formula E to illustrate its utility. Colunga (2014) examined the modelling of transient cornering and suspension dynamics,

along with the investigation of control strategies for an ideal driver within a lap time simulation framework. In a similar vein, Timings (2014) contributed to this body of work by aiming to develop a robust lap time simulation, referring to its comprehensive nature and its resilience in varying conditions.

However, for this thesis, race simulations and their components, specifically those that prioritize pit stop strategies as a key component in modelling or predicting race outcomes, are of greater relevance. Such simulations are instrumental in forecasting final standings by accounting for various factors, including driver interactions, empirical fuel consumption models, tire wear, and probabilistic effects. Bekker (2009) developed one of the earlier holistic race simulations to replicate key on-track activities in Formula 1, such as mechanical failures, overtaking manoeuvres, and pit stops. This model facilitates strategy planning by simulating the mechanical and physical dynamics of a race, thereby offering a team a potential advantage. More recently, Heilmeier (2018) outlined a simulation methodology for circuit motorsport racing strategies, which considers variables such as pit stops, tire choices, and tire degradation. The tool is designed to rapidly simulate races based on discrete lap data and adjustable strategy inputs. Building on this previous work, Heilmeier (2020) introduces a new further improving the simulation. This advanced version surpasses simple optimisation models by providing a comprehensive, automated simulation that responds in real-time to race dynamics. Heilmeier (2020) suggests that the current methodology for optimising pit stop decisions and associated tire compound selections might benefit from additional exploration in the future. Furthermore, he acknowledges the omission of complex strategies, such as the undercut—a tactic where a driver pits and switches to faster tires to gain time on rivals who pit later—from current models. He advocates for the integration of such tactics into subsequent models to enrich the decision-making process.

In addition to comprehensive racing simulations, focused research activities are directed at optimising specific components of the racing domain and simulations themselves, such as pit stops. These efforts aim to refine these aspects to their utmost efficiency. One research exemplifies this by employing machine learning algorithms to aid tire strategy decisions in the NASCAR series. This work utilises predictive analytics, employing historical race data to forecast positional shifts consequent to variables such as tire change frequency and tire lifespan. Extensive feature testing has revealed that support vector regression and LASSO regression yield the highest accuracy in results (Tulabandhula and Rudin 2014). Additionally, Bell (2016) advances the analysis of Formula 1 by conducting a comprehensive analysis of performance determinants in Formula 1, examining the evolving contributions of team dynamics and driver skills over time. The study results in a systematic ranking of drivers, offering insights into the qualifications of the 'potential best' based on a quantifiable set of criteria. Furthermore, Monte Carlo methods and analysis of probabilistic factors play a significant role in simulating the inherent variability present in lap times, pit stops, race incidents, and potential degradation of vehicle parts. The study of Heilmeier (2020) builds upon this foundation, providing a comparative analysis of the seminal works of Bekker (2009), Phillips (2014), and Salminen (2019), thereby extending the understanding of these stochastic elements in race simulations.

The existent body of research on Formula 1 is mainly based on the scope of publicly available data, which has been limited to lap times, and race outcomes. Such data has predominantly been sourced from the Ergast API, a privately maintained database for Formula 1 statistics (Ergast 2009). However, the introduction of the Fast F1 API marks a significant progression in data availability, offering not only the information provided by the Ergast API but also a more comprehensive set of F1 data. This includes telemetry data, official weather statistics, and track information (FastF1

2020). The introduction of the Fast F1 API thus promises to broaden the scope of current literature and models by facilitating the incorporation of mechanical details of the vehicle (such as current gear, RPM, and speed) and more granular data like positional coordinates, distances between drivers, and time specific air and track temperatures. The newly available data points, particularly telemetry data, which have received little attention in the scientific literature to date, present potential opportunities to enhance and broaden current analytical frameworks. The richer and more granular nature of this data allows for a more detailed examination of specific drivers' behaviours based on positional and telemetry information. A prospective methodological approach might include clustering drivers according to their behavioural patterns in various racing scenarios. Such clustering could yield valuable insights into performance differentiators and decision-making processes throughout a race and its strategic decisions.

Clustering Techniques in Sports Analytics and Formula 1

Clustering, a core technique in unsupervised machine learning, groups data points into distinct categories based on common attributes without predefined labels (Pedregosa et al. 2011). In sports analytics, clustering is a key tool, enabling teams and coaches to decode complex patterns and detect subtle correlations. This method can help identify performance trends and effective team compositions, which, in turn, can be leveraged to improve predictive models that forecast future sports outcomes based on the grouping of player or play characteristics. Clustering may uncover non-intuitive groupings or strategies, providing a strategic advantage in the competitive world of professional sports. Its utility and adaptability across various sports disciplines are well-documented, with a significant body of literature.

In Basketball analytics, player assessment and categorisation have evolved well beyond the confines of traditional position labels. One study by Duman, Sennaroğlu, and Tuzkaya (2021) applies hierarchical cluster analysis to a rich dataset of game-related statistics from 15 NBA seasons, uncovering four to six distinct playing styles within each traditional position. The clusters, characterised by unique attributes and skill sets, offer a multifaceted perspective on player capabilities, yielding strategic insights for player placement and team composition. Muniz and Flamand (2022) introduce an advanced clustering technique based on weighted networks. The methodology starts with k-means clustering to form preliminary groupings, which then inform a network where players are interconnected by weighted edges reflecting performance similarities. Employing the Louvain method for community detection, the study identifies eight player archetypes, surpassing the insight offered by the five traditional positions. They further enhance this method by using tracking data, adding a layer of depth to the analysis with precise measurements of player movements and interactions. For Football, a fuzzy clustering model that can handle mixed data types has been introduced. This model assigns objective weights to attributes such as player performance metrics, positional data, and physical characteristics, thereby uncovering clusters that the complexity of mixed-attribute data might hide. Such detailed analysis facilitates the identification of player clusters according to their on-field roles, skill sets, and physical profiles, which is instrumental in tactical team structuring and player market valuation (D'Urso, De Giovanni, and Vitale 2023). A study in Tennis clustered 1188 Grand Slam players by analysing their physical and play style data, revealing four distinct profiles. Through two-step cluster analysis and further MANOVA and discriminant analysis, the research identified how factors like height and handedness correlated with performance, notably in serving and net play (Cui et al. 2019). Clustering in sports analytics reaches beyond mainstream sports, extending its application to disciplines like Badminton. Sinadia and Murwantara (2022) apply k-means and

Kommentiert [DP2]: spelling?

hierarchical agglomerative clustering to analyse Badminton athletes' performances, identifying clusters based on game results and consecutive scoring. K-Means clustering discerns four distinct performance clusters, with one particularly strong cluster associated with high scores and consecutive points, supported by similar results from hierarchical clustering.

Collectively, these studies showcase the breadth and depth that clustering techniques bring to sports analytics. They enable a sophisticated segmentation of athletes, offering a granular understanding that can guide coaching decisions, athlete development, and competitive strategy formulation. In the data-rich context of Formula 1, the application of clustering techniques to group drivers by various data-driven similarities could potentially yield insights that extend beyond traditional performance metrics. While such methods have provided detailed understandings in other sports disciplines, the extent to which they have been applied to driver analysis in Formula 1 remains limited. Most existing studies on driver performance within this sport have focused on race results without a significant exploration of specific clustering techniques that have been beneficial in broader sports analytics.

For instance, one pioneering study by Phillips (2014) presents a statistical model that quantifies the contributions of drivers and teams to performance, using championship points as the metric. In another notable work by Rockerbie and Easton (2021), they employ an econometric method to dissect the relative impact of driver skill versus car technology on race results, discussing the "80-20 rule" but primarily concentrating on race positions and outcomes without delving into driver-specific skill and performance metrics. Similarly, a recent study utilises a Bayesian multilevel rank-ordered logit model to analyse historical ranking data, aiming to distinguish between driver skill and constructor efficiency (Van Kesteren and Bergkamp 2023). However, like its predecessors, it does not extensively investigate drivers' characteristics or behaviour.

While providing valuable insights, the discussed studies often centre around broader performance metrics and race outcomes without specifically exploring the clustering of drivers based on their attributes or behaviours. This gap indicates an opportunity for further research into the application of driver clustering techniques within Formula 1. Exploring driver clustering, especially by using detailed telemetry data, expands the analytical horizon and could establish a foundation for in-depth examination of strategic components, such as pit stop tactics and other decision-making processes in the sport. Furthermore, the comprehensive use of telemetry data remains underutilised in existing models. Incorporating telemetry data into clustering analyses could reveal more profound insights into driver behaviour, driving styles, and performance metrics, which go beyond the conventional race outcomes and rankings.

4. Data

F1DataFetcher

The publicly accessible API from FastF1 and the associated Python library can be used to generate the required data. In contrast to the previously common ERGAST API, the FastF1 API also includes interesting aspects such as weather, position and telemetry information in addition to the usual Formula 1 data. Telemetry information includes speed, revolutions per minute and much more. Position data not only informs about the exact position of a car on the racetrack but also provides information about direct duels by showing which driver is driving in front of a driver and how far away he is. This data can be a great asset and a clear advantage over existing literature. That's why the decision on FastF1 as the main source of data was made quickly. Using this API is relatively simple and works without any problems. Interested readers are referred to the detailed documentation provided by the operators (FastF1 3.1.6). However, the API has one limitation for efficient data collecting purposes: the user can only ever load one event of a session, be it

Kommentiert [VF3]: Irgendwie fehlt uns hier eine Überleitung wofür die “required data” überhaupt required ist. Auch später hinsichtlich der Data Limitations wird nicht wirklich klar, für welchen Teil der Arbeit diese Limitations relevant sind (eigentlich nur das Clustering oder?)

qualifying, race, free practice, etc., from the API. This is not due to the user's authorisation but merely to the logic of the API. To get around this, a new logic was created.

The Python class we developed is called F1DataFetcher. This allows us to pull data for multiple races simultaneously from the API into a notebook, automatically creates the required data frames in the required structure and outputs them as Panda's data frames. In addition to the challenge of being able to receive multiple data at once, it was also important to be able to save the data so that local access without the need of reloading data every time was granted. Therefore, we created a relational database in PostgreSQL. This allows to save the individual data frames that are collected by the API via the F1DataFetcher in a meaningful, coherent, and comprehensible way and to keep them ready for further use. The database was created manually in pgAdmin4, and a connection was established via the Python library 'sqlalchemy'. This package allows users to send SQL commands in Python to SQL-based databases and receive pandas' data frames. The methods of the FastF1 Data Fetcher were enhanced by the capability of sending collected data automatically to a specified PostgreSQL database. The final relational database contains a total of 10 tables. The structure and relations are composed as shown in Figure 1.

Kommentiert [DP4]: same spelling as above

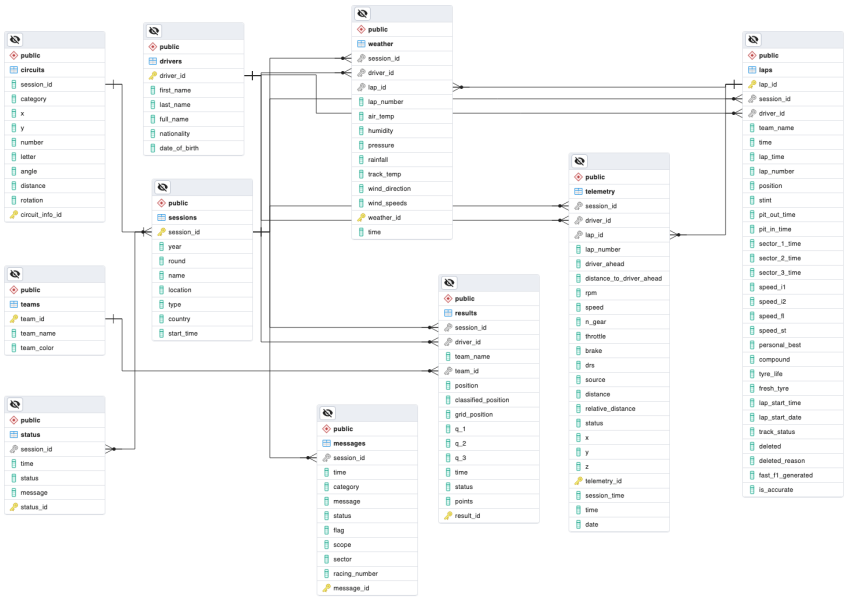

Figure 1: Simplified Schema of Our Own Data Base

The final F1DataFetcher class therefore contains ten methods plus one more to collect different data frames from FastF1 (and ERGAST API) and optionally store them in a database with the above schema. The data frames are on different levels which will be explained in the following along with the methods used to collect this data.

The first group of methods collect general information, such as drivers or teams. Although this data is collected on session level, the information is overarching and not tied to individual sessions. A comprehensive list of the methods used to collect this data is shown in Table 1.

Table 1: F1DataFetcher Methods - Group 1

Method	Description
`get_multiple_sessions()`	Load detailed information on grand prix events, optionally store to database
`get_unique_drivers()`	Load detailed information on unique drivers, optionally store to database
`get_unique_teams()`	Load detailed information on unique teams, optionally store to database

The second group of data is the one where data is on session level. The methods that collect these kinds of data are listed in Table 2.

Table 2: F1DataFetcher Methods - Group 2

Method	Description
`get_race_results()`	Load detailed information on grand prix results, optionally store to database
`get_race_control_messages()`	Load detailed information on race control messages, optionally store to database
`get_track_status()`	Load detailed information on flags and statuses, optionally store to database
`get_circuit_info()`	Load detailed information on track characteristics, optionally store to database
`get_weather()`	Load detailed information on weather characteristics, optionally store to database

The final set of methods collect data that is highly granular, meaning collected individually on lap basis for each driver. A full overview is provided in Table 3.

Kommentiert [VF5]: Spelling Mistake in Store in the Table.

Table 3 F1DataFetcher Methods - Group 3

Method	Description
`get_laps()`	Load detailed information on laps, optionally store to database
`get_telemetry()`	Load detailed information on telemetry, highly granular
`store_to_postgresql()`	Store highly granular telemetry data to database

The methods of the F1DataFetcher provide a complete, easy to use way of collecting large amounts of data from FastF1 API and automatically store in PostgreSQL. They were then applied to set up a full database that served as the input data for the complete work.

Data Limitations

For feature engineering, data pre-processing and normalization, data was limited for better performance and other reasons. The process of data filtering was conducted with the following steps:

- **Exclusion of races during rain**: Grand Prix events in rainy weather conditions were excluded. The reason behind this was that in the complete database, there is a lack of data for events in the rain. Since not all drivers participated in all the races, and metrics can be significantly different on wet tracks, the decision was made to eliminate these data points from the data that serves as input for feature engineering.
- **Removal from retired drivers**: Although there is some information in the data about why they retired, there is no sufficient way to know if it was their fault, possibly some other driver's fault, a tactical decision or simply a problem with the car. Therefore, the decision was made to remove all data of drivers who retired on a particular track and only look at those data points where a driver successfully drove all laps of the race.
- **Removal of interrupted laps**: Sometimes races get interrupted due to accidents, weather conditions or other reasons. In these cases, the data is noisy, leading to lap- and pit stop times of several hours, which makes no sense and only distorts some of our engineered features. To avoid these kinds of complications, the data for laps in which a red flag was hissed, meaning the race was interrupted, was omitted from the data.

Kommentiert [VF6]: We mention features but it is not yet mentioned what we are aiming to do and what we need features for.

5. Clustering

Introduction to Clustering in Racing Analysis

Transitioning to clustering, a principal method in unsupervised machine learning, we leverage the refined dataset to explore and interpret the complex dynamics of Formula 1 racing.

Clustering is essential in the analysis and interpretation of complex data sets. This technique, as elucidated by Jain (2010), involves the organization of a collection of patterns into clusters, based on similarity. Patterns within a cluster exhibit a higher degree of similarity to each other than to those in different clusters, thereby revealing the intrinsic structure of the data. The significance of clustering extends beyond mere data categorization. Jain (2010) emphasizes its role in exploratory data analysis, summarization, and compression, as well as its utility as a tool for hypothesis generation and testing. In diverse fields such as image and pattern recognition, bioinformatics, and information retrieval, clustering is instrumental in uncovering hidden structures, identifying anomalies, and simplifying complex data landscapes.

In the realm of sports analytics, particularly in Formula 1 racing the application of clustering takes on a more defined and critical role. In the high-octane world of Formula 1 racing, understanding the nuances of driver behaviour is essential for gaining a competitive edge. This study employs a clustering methodology to delve into the intricacies of driving behaviour, aiming to achieve a set of specific objectives that enhance our comprehension of what differentiates top performers on the track. A central goal of this research is the characterization of driver profiles through clustering. This involves categorizing drivers into distinct groups based on a comprehensive array of performance metrics and observed behaviours on the track. The aim is to construct detailed profiles that capture the essence of each driver's approach to racing, including their performance, behaviour, and tactics. These profiles are expected to provide a holistic view of each driver's strengths and areas for improvement, offering insights into the diverse strategies and techniques employed in

Kommentiert [DP7]: we also introduce clustering in the literature review …

Kommentiert [VF8R7]: Lets try to bring the source to the literature review?

Kommentiert [VF9]: profiles or behaviour?

Kommentiert [SG10R9]: synonym

Kommentiert [VF11]: We mention improvement but this is to the scope of the thesis i guess.

Kommentiert [VF12R11]: @Simon Leonard Grube Was war hermit gemeint?

Kommentiert [SG13R11]: dass man wenn man die einzelnen charactersitiken der Fahrer weis natürlich auch weiss wo ihre Schwächen liegen im vergleich zu den anderen Fahrern, an denen sie noch arbeiten können.

Formula 1 racing. Furthermore, the study engages in a comparative analysis across the derived clusters. This analysis is crucial for identifying the unique characteristics that distinguish each cluster, thereby shedding light on what sets apart the most successful drivers. This comparative approach is anticipated to reveal valuable insights into the factors that contribute to effective driving strategies and overall race performance.

Finally, the clustering analysis aims to provide practical strategic implications for teams and drivers. By understanding the distinct behavioural clusters, teams can tailor their training programs, strategy development, and in-race decision-making to better align with the identified strengths and weaknesses of their drivers. This objective is predicated on the belief that data-driven, personalized strategies can significantly enhance performance and competitiveness in Formula 1 racing. Aligning training and strategies with the insights gained from clustering analysis, teams can optimize their approach to each race, maximizing their potential for success in this dynamic and challenging sport.

Kommentiert [VF14]: Hier müssen wir vorsichtig sein, da die Teams sicherlich genau das schon machen.

Feature Engineering

Before being able to cluster driver behaviour and performance, it is vital to define metrics and input features that capture the essence of what is tried to analyse. Since Formula 1 is a complex competition with various influential factors, we decided to limit it down to three main aspects: The performance metrics employed to evaluate a driver's performance, behavioural metrics that illustrate drivers' conduct during driving, and tactical metrics, reflecting the strategic choices made by a driver or team throughout a race.

Kommentiert [DP15]: here it sounds like that there are only two aspects: behav. and tactial

Kommentiert [VF16R15]: I think it is fine, due to the fact that we mention the word metrics exactly three times.

Kommentiert [VF17R15]: changed the wording slightly. Now it seems clearer.

In the following section, an overview of the features created for the clustering of Formula 1 drivers is represented, detailing the data utilized for their calculation.

Performance Metrics

Despite the significant dependency of performance on the car's capabilities, for which we have insufficient data, our objective was to quantify specific dimensions of driver performance. We defined three main aspects of driver performance: lap time, position gain, and consistency in driving. The first two metrics are self-explanatory and do not require additional clarification. The third metric, consistency, aims to quantify a driver's uniformity on the track, specifically regarding lap time variations. These three metrics effectively represent a comprehensive assessment of a driver's performance.

Table 4: Clustering - Performance Metrics

Variable	Name	Description
Lap time	`avg_lap_time`	Average lap time for each driver on all races
Position gain	`position_gain`	Average number of positions gained (or lost)
Consistency	`consistency`	Standard deviation in average lap time

Behavioural Metrics

More complicated than defining metrics to see a driver's performance is mapping various aspects of driving behaviour to input features. The focus was primarily on aggressiveness and a driver's behaviour towards other drivers, especially the driver ahead of him. To analyse behaviour in direct competitive scenarios, we quantified various gaps and distances relative to the leading driver, thereby attempting to delineate patterns of conduct in these head-to-head duels. After all, the five new input features for capturing driver behaviour were defined as the *distance to the driver ahead*, the *time within range*, *persistence*, *overtakes*, and *defensive actions*.

Table 5: Clustering - Behavioural Metrics

Variable	Name	Description
Distance to driver ahead	`avg_distance`	Average distance to the opponent ahead
Time within range	`time_within_range`	Average time a driver spends in a range of 50 meters behind his opponent
Persistence	`persistence`	Standard deviation of the distance to the driver ahead
Overtakes	`overtakes`	Average number of successful overtakes
Defensive actions	`position_maintained`	Average number of laps without losing a position

Kommentiert [VF18]: We have two times persistence as a variable name in the table.

Kommentiert [VF19R18]: @Max Rafael Leischner lets change the variable name

Tactical Metrics

As tactical decisions that are made by team leads and the driver himself during constant communication play an essential role in Formula 1 and can impact a driver's final position and overall performance significantly, capturing differences in strategic decisions was the goal that was aimed for. Two main aspects of Formula 1 tactics are pit stop timing and compound choice. Deciding when to pit stop, which tires to wear and how many pit stops a driver does can drastically change the outcome of a race. To capture the strategic decisions for each driver, a new set of features was created from the data to capture the pit stop timing and the compound strategies.

Table 6: Clustering - Tactical Metrics

Variable	Name	Description
Pit time	`pit_time`	Average time a pit stop takes
Pit window: Early	`early`	Percentage of pit stops in first 25% of the laps
Pit window: Mid	`mid`	Percentage of pit stops in mid 50% of the laps
Pit window: Late	`late`	Percentage of pit stops in last 25% of the laps
Laps on SOFT	`laps_on_soft_percentage`	Percentage of laps driven on SOFT compound
Laps on MEDIUM	`laps_on_medium_percentage`	Percentage of laps driven on MEDIUM compound
Laps on HARD	`laps_on_hard_percentage`	Percentage of laps driven on HARD compound

Kommentiert [VF20]: is the pit time really called avg_distance in the clustering?

Kommentiert [ML21R20]: nope

Normalization

Normalization is a common technique to scale data to a standard range, eliminate redundant or noisy data, and improve algorithm performance significantly (Patel and Mehta 2011). Literature shows that it is common practice in machine learning to apply normalization techniques to data before using it as input data (Gopal, Patro, and Kumar Sahu 2015).

Kommentiert [DP22]: bold on purpose?

Since most clustering algorithms, such as K-means and others, rely on scaled data, we needed to normalize our data points. Another reason to follow that approach is that in motorsports in general and Formula 1 in particular, the information on the actual car build is highly confidential and due to this not publicly accessible. The cars have different characteristics not shared by the teams participating in Formula 1. There might be cars performing better in general than other cars. Since

Kommentiert [DP23]: wir brauchen eine einheitliche schreibweise

the focus was merely on the driving capabilities of the individual drivers only, a need to reduce the potential bias that could influence our clustering algorithms was urged.

Besides that, some input features can be significantly influenced by track characteristics, such as the number of corners on a track and their angle. Since data from various Grand Prixes was used, results achieved on multiple tracks were compared. To make them comparable, again, there is a need for data normalization.

There are several ways of scaling and normalizing data, the most common being z-score scaling and min-max normalization. For this purpose, the decision was made to apply the latter to our input data. Min-max normalization is a technique that keeps the original relationship among the data (Gopal, Patro, and Kumar Sahu 2015). It follows a simple calculation:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

The Python library 'scikit-learn' (Pedregosa et al. 2011) provides a standard pre-processing algorithm for this kind of normalization that we applied to our data.

Selection of Clustering Algorithm

After normalizing our data to ensure comparability and reduce potential biases, as detailed previously, we move to the next crucial step in our analysis: the selection of an appropriate clustering algorithm. The choice of an algorithm is crucial, as it must align with the nature of our normalized data and the specific objectives of our study. The selection of the K-Means clustering algorithm for the analysis of driver behaviour in Formula 1 racing is underpinned by several compelling reasons that align with the specific nature of the data involved. A primary factor is the simplicity and computational efficiency of K-Means, as highlighted in the study of Jain (2010).

This characteristic is particularly advantageous given the large volume and complexity of data generated in Formula 1. K-Means offers an efficient and straightforward approach, which is essential for the initial exploratory analysis where a fundamental understanding of data grouping is sought. Additionally, the suitability of K-Means for clustering numerical data, as noted by Arthur and Vassilvitskii (2007), aligns well with the mostly numerical nature of our engineered data. This alignment ensures that K-Means is well-equipped to handle the specific types of data encountered in this research. In addition, the efficiency of K-Means in processing large datasets is a significant advantage, as it allows for the swift and effective handling of the extensive data typical in Formula 1 racing. This efficiency, coupled with the ease of interpretation of its results, as Jain (2010) notes, makes K-Means a practical choice for researchers and analysts who require clear and actionable insights from their data.

However, it must also be noted that K-Means has certain limitations in its application. One challenge is that it operates under specific assumptions about the nature of the clusters it forms. Typically, the algorithm assumes that clusters are spherical and of similar size. However, as Arthur and Vassilvitskii (2007) point out, these assumptions may not always hold true in real-world data scenarios, including those encountered in Formula 1 racing. This limitation necessitates an understanding that the clusters formed may not perfectly represent the complex and varied nature of the data.

Additionally, K-Means is known to be sensitive to the initial choice of centroids. This sensitivity can lead to variability in results across different runs of the algorithm, as observed by Celebi, Kingravi, and Vela (2013). This characteristic requires careful consideration and potentially multiple iterations to ensure robustness in the clustering outcomes. This need for precise selection

Kommentiert [VF24]: klingt etwas so als würde es “nur” für initial exploratory analysis vorteilhaft sein

Kommentiert [VF25R24]: @Simon Leonard Grube Was war der Grund hier exploratory data analysis zu erwähnen?

Kommentiert [VF26]: Lasst uns eine Überleitung einbauen.

of starting points leads directly into the next crucial step of our analysis: determining the optimal number of clusters.

Algorithm Parameters and Tuning - Determining the Number of Clusters

In defining the number of clusters for our k-means analysis, we employed a diverse approach to ensure methodological rigor. Initially, we utilized the elbow method, widely recognized for its effectiveness in cluster analysis (Kodinariya and Makwana, 2013). This method involved plotting the explained variance against the number of clusters and identifying an 'elbow' point where the rate of decrease in variance sharply changes. Our analysis indicated an elbow point between three and four clusters, suggesting that further increasing the number of clusters would result in marginal improvements in explained variance.

To complement this quantitative method, domain knowledge regarding the performance of drivers was also incorporated. This additional layer of analysis provided valuable context in guiding the decision-making process. Specifically, the decision to opt for four clusters was influenced by the distinct characteristics of drivers Russel and Latifi, who formed a separate fourth cluster, in contrast to the other three clusters. Given their different tire choices compared to the other drivers and their overall race positioning at the lower end of the grid, it was deemed appropriate to assign them to a separate cluster. This decision underscores the importance of integrating quantitative methods with domain-specific knowledge for more insightful and informed analyses in complex situations.

While the silhouette score, a measure of cluster cohesion and separation, was also calculated as part of our evaluation, it was the combination of all three together that ultimately led us to choose four clusters. This number of clusters appeared to offer a reasonable balance, ensuring that the clusters were distinct and meaningful without introducing overfitting or unnecessary complexity.

Kommentiert [ML27]: Silhouette Score. Müssen wir eventuell auf Nachfrage providen.

The decision to opt for four clusters was made with the goal of achieving a segmentation that would allow for an insightful analysis of driver behaviour.

Results

Applying the outlined clustering methodology, drivers in the Formula 1 season of 2019, 2020, and 2021 were effectively segregated into four distinct clusters, which can be seen in the scatterplot in Figure 2. This is based on the presented combination of performance, tactical, and behavioural metrics.

The allocation of drivers across these clusters is as follows:

- **Cluster 0:** Albon, Raikkönen, Perez, and Magnussen.
- **Cluster 1:** Hamilton, Verstappen, and Bottas.
- **Cluster 2:** Grosjean, Kvyat, Leclerc, Norris, Ocon, Gasly, Ricciardo, Sainz, Stroll, Giovinazzi, and Vettel.
- **Cluster 3:** Latifi and Russell.

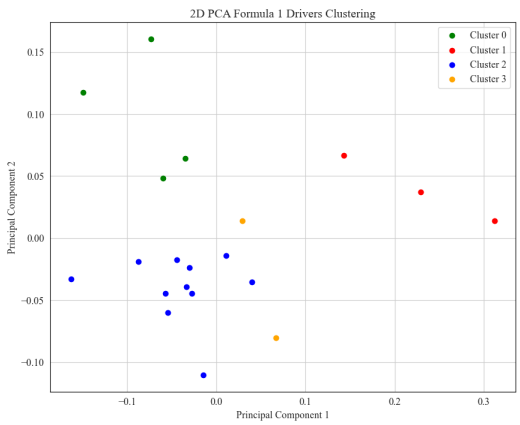

Figure 2: Scatterplot Illustrating Driver Clusters in 2D PCA Space

Kommentiert [ML28]: Ist nicht erklärt davor, warum wir diese Zeitspanne wählen.

Kommentiert [DP29]: klein?

Kommentiert [VF30]: Sind das alle Fahrer? Oder haben wir welche ausgeschlossen basierend auf Annahmen? Falls ja, erwähnen wir das nicht oder?

Kommentiert [VF31R30]: @Simon Leonard Grube Kannst du das noch einmal raussuchen?

An initial analysis reveals a reasonable distribution reflective of known performance tiers within the sport. The top performers, typically dominating the front of the pack, are distinctly grouped in Cluster 1, while those frequently at the rear form Cluster 3. The midfield drivers, recognized through domain knowledge as a diverse group in terms of tactics and performance, are split between Clusters 0 and 2.

To describe the clusters in more detail and to interpret the classification, the underlying performance, tactical and behavioural metrics, which were defined as input features, need to be put into perspective. The different degrees of these features can also be seen in Figure 3. Here, the mean of the features per cluster are shown in the form of a parallel plot enabling a visual comparison between the clusters.

Kommentiert [VF32]: Average of the features = Cluster Centroids?

Interpretation

When interpreting the clustering results, it is important to recall that in Formula 1 the competition involves the top 20 drivers in the world. This elite group ensures a high degree of skill parity, as illustrated by our parallel plot analysis. The analysis underscores the close competition and high level of competence inherent in this sport. However, upon closer examination, subtle but meaningful variations in performance can still be observed. It's crucial to emphasize that while these differences are partly due to the varying budgets of the teams, and consequently the performance of the cars, the vehicle is not the sole determinant of a driver's overall performance. Teams with larger financial resources often have an advantage in terms of car performance, but the skills and decisions of the driver are also key to success. Teams with limited budgets may face challenges in vehicle performance, yet the individual performance and strategy of the driver can

mitigate these disadvantages to some extent. This interplay between car performance and driver skill is a key element in analysing the competitive dynamics of Formula 1.

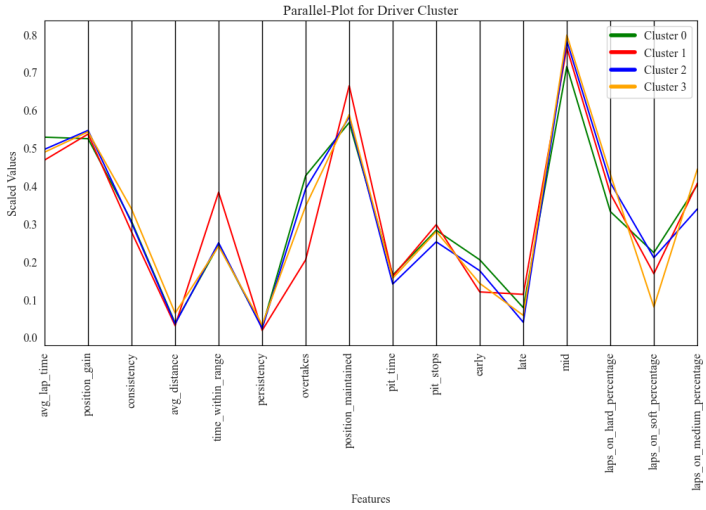

Figure 3: Parallel Plot Showing Mean Features of Driver Cluster

In our subsequent analysis, we will explore the specific characteristics of the clusters, bearing in mind the influence of vehicle differences, to provide a comprehensive understanding of the factors driving performance in Formula 1. This detailed examination starts with Cluster 0 and is based on the mean of each feature for each cluster.

Kommentiert [VF33]: Again mean = average = cluster centroid? Lasst uns auf einen begriff einigen.

Drivers in Cluster 0 are characterized by above-average lap times suggesting a focus on endurance and tactical positioning rather than outright speed. Their notable position gains reflect effective racing behaviour, with a high number of overtakes suggesting skill in dynamic racing scenarios. The moderate consistency of their performance can potentially be attributed to the positioning of the drivers in the midfield. Interactions between the drivers are much more frequent here, which has a greater impact on lap times. These interactions are also reflected in the frequent overtaking

manoeuvres. The cluster's pit strategy, marked by average pit times and a preference for early stops, points to a tactical edge in race management. This is complemented by their significant usage of hard tires, implying a strategy centred on durability and long-term race planning over the immediate speed offered by soft tires.

In essence, drivers in Cluster 0 excel in strategic manoeuvring and consistency, prioritizing race longevity and tactical positioning over peak lap speed. This approach aligns with drivers who may not have the fastest cars but utilize strategic race management to optimize their overall standings.

Moving on to Cluster 1, which is characterised by the lowest average lap times, indicating superior performance. Their excellent position maintenance skills, combined with the highest average amount of pit stops, point to strategic control of both race pace and pit strategy. The frequent pit stops in this strategy are likely a tactical decision to optimize tire performance, considering there is usually sufficient time for an additional pit stop towards the end of the race without losing positions. This approach aligns with the balanced tire usage observed, where a high percentage of laps are run on both medium and hard compounds. Their exceptional consistency underscores a notable ability to deliver stable, reliable performances across various races. Additionally, their highest time within range of the driver in front, paired with a lower number of overtakes, suggests they often lead races and if not remain close to the driver in front, efficiently managing their positions and indicating an aggressive approach to overtaking manoeuvres. In essence, Cluster 1's drivers demonstrate a blend of speed, strategic expertise, and consistent performance, marking them as elite competitors in the Formula 1 field.

Next is Cluster 2, which presents a profile of competent race management with a focus on strategic tire usage. This cluster's average lap times and position gains, along with a consistency level akin to Cluster 0, indicate a balanced approach to race performance, matching speed with strategic positioning. The drivers in this cluster exhibit a lower frequency of overtakes compared to Cluster 0, yet they maintain their positions effectively during races. This suggests a focus on consistent lap performance and strategic placement rather than aggressive overtaking manoeuvres. Their ability to hold positions is indicative of skilful ability, particularly in defending against competitors and capitalizing on track opportunities. A notable aspect of Cluster 2 is their pit strategy. These drivers tend to pit the earliest among all clusters and have the lowest average number of pit stops. The fewer pit stops suggest a preference for longer stints on track. This aligns with their unique tire usage pattern, as Cluster 2 is the only group to perform more laps on hard tires compared to medium ones. This preference indicates a strategy leaning towards tire durability and longer race stints, possibly to maintain consistent lap times and reduce the time lost in pit stops. The hard tire's longevity would be particularly advantageous in races with high tire degradation or where preserving tire life is crucial for race strategy.

Kommentiert [VF34]: Or always on the lower end, not having anyone in front to overtake or behind to defend against.

In summary, Cluster 2 drivers excel in strategic racing behaviour and tire management, combining average lap times with effective position holding and a distinctive early pitting strategy, highlighting a focus on endurance and tactical adaptability in the Formula 1.

Lastly, Cluster 3 presents a unique set of characteristics. Their lap times and position gains are comparable to those in Cluster 2, suggesting a similar approach in terms of speed and race advancement. However, the distinctive feature of this cluster is the lowest consistency among all groups, as indicated by the highest standard deviation in average lap times. This variability could

be attributed to a range of factors such as adapting to different race conditions, varying strategic choices, or fluctuations in both driver and car performance. Another notable aspect of Cluster 3 is their racing style, which seems to focus on endurance and stability. This is inferred from their longest average distance to the driver ahead. Such characteristics may imply a tendency towards a more calculated and defensive racing approach, possibly prioritizing maintaining a steady pace and avoiding risks. The tire usage pattern in this cluster further supports this interpretation. The drivers in Cluster 3 have the lowest percentage of laps on soft tires, which are typically used for aggressive, short-term speed advantages. Their reluctance to use soft tires might indicate a preference for strategies that emphasize tire longevity and sustained performance over short bursts of speed. This approach is often seen in races where managing tire degradation and fuel consumption is critical for achieving a favourable outcome.

In conclusion, Cluster 3's drivers appear to adopt a strategic approach focused on endurance and stability, with less emphasis on aggressive manoeuvres and high-speed performance. Their lower consistency might reflect a more adaptive or variable strategy, responding to the specific demands of each race. This approach, combined with a cautious tire strategy, suggests a focus on long-term race management rather than immediate position gains.

In summary, the clustering analysis of Formula 1 drivers reveals distinct strategic profiles across the four groups. Cluster 0 and 2 both emphasize endurance and tactical positioning, but Cluster 2 distinguishes itself with an earlier pitting strategy and a greater emphasis on hard tires. Cluster 1 stands out for its combination of high speed, strategic pit stops, and consistent performance, showcasing the traits of front-running drivers. In contrast, Cluster 3 is defined by its variability in performance and a more conservative tire strategy, indicating a focus on stability and endurance.

This comparative analysis highlights the diverse approaches in Formula 1, from aggressive speed and tactical prowess in Cluster 1 to the more calculated and endurance-focused strategies in Clusters 0, 2, and 3.

Individual Part

6. Hypothesis Testing: The Effect of Second Driver Pit Stop Timing on Race Outcomes

Introduction

In the world of motorsport, Formula 1 is characterized by an interplay of technological advancement and strategic finesse. While drivers and their cars are in the spotlight, the team decisions behind the scenes often influence the race's outcome. The research gap addressed by this thesis lies in the under-researched question of whether and how the order in which a team's drivers are called in for a pit stop influences the race result. Previous literature has focused primarily on optimizing pit stop timing and tire choice but has yet to consider the internal team hierarchy's impact systematically. This leads us to the following two hypotheses:

Kommentiert [VF35]: Dadurch das hier Thesis genutzt wird, Wirt es so als würde die gesamte Arbeit anstatt ausschließlich dieser Teil diese Frage behandeln.

H0: The order of pit stops does not significantly affect the normalized race result of a team's second-placed driver.

H1: The order of pit stops significantly influences the normalized race result of a team's second-placed driver.

In order to investigate these questions, a comprehensive data set from several racing seasons is used and subjected to statistical analysis. By applying quantitative methods and a deep understanding of the sport, the aim is to gain new insights into the factors influencing race performance and to understand whether the teams' strategic decisions regarding pit stops systematically impact the race outcome. The study is designed to be methodologically based on

robust statistical procedures while considering the specific circumstances and challenges of Formula 1. The results of this study will enrich the scientific discussion on the tactical component in Formula 1 and derive practical recommendations for the teams to strengthen their competitive position.

A detailed description of the data basis used follows this introductory section. The selection criteria for the data, the methodology used to collect and process the data, and the statistical methods used to test the hypotheses are specified.

Data & Methodology

Data Pre-processing and Selection

The data we used for the hypothesis test is gathered from the same data basis used in the clustering. We look at the cross-sectional data for the 2019 - 2021 seasons, including the laps, results, circuits, and weather tables from the database. In addition, we use the information resulting from clustering, which tells us which drivers belong to which cluster. To declare the driver's position within a team, meaning whether he is seen as the first or second driver, we have added the driver's salaries to our existing data (Madison Pearce 2020; Dieter Recken 2021; Andreas Reiners 2019). We only selected the entries of the drivers who finished the race, as we can only analyse the influence on the race result if the driver has finished the race. Furthermore, only races where it did not rain were taken into account to align with the data used in the clustering. In addition, we only included these entries where both team drivers started with the same tire. Thus, we ensure that the drivers start with the same strategy, which implies that their tires should be worn out simultaneously and consequently the drivers should be pitted at the same time. This ensures that the pit order does not depend on tire durability. Looking at the tire strategy implies that we only look at team entries where both drivers participate in a race. In addition, both drivers of a team must make at least one pit stop so that we can elaborate which driver comes into the pit first. If a driver leaves the race before the first pitstop

or does not continue after his first pitstop, the team will not be included in this race. This way, we can ensure that the pit order was a deliberate decision by the team management and was not caused by technical problems. This results in a data set of 422 entries with 47 different races. The variables used in the regression analysis are stated in the following part.

Feature Engineering

In the scope of feature engineering, relevant variables were developed from the extensive database, which includes lap times, results, racetrack data, and weather conditions, in order to be able to depict the race outcome precisely. The aim was to consider both the influence of internal team dynamics and external conditions. The following main characteristics were constructed for the regression analysis.

Dependent Variable

The null hypothesis (H_0) of this research asserts that the strategic decision to pit the second driver first exerts no statistically significant influence on normalized race outcomes. The term 'race outcome' serves as the dependent variable in this analysis, though its definition is subject to variability. The most straightforward approach might involve using the actual race results. However, considering the diversity in resources and starting conditions among different teams and drivers, a mere reliance on raw outcomes may not reflect an equitable assessment of performance. To account for these disparities, normalization of results is proposed in the following. This approach enables the evaluation of a driver's performance relative to their specific baseline standards, thus fostering a more equitable comparison.

The dependent variable is conceptualized as the net gain or loss in position relative to the driver’s average seasonal position.

$$position\_change = average\_position - actual\_position \quad (3)$$

This operationalization facilitates a quantifiable analysis of the impact of strategic decisions on the race performance of the second driver. By anchoring the variable to the driver's average seasonal performance, this approach allows for a more nuanced examination of strategy effectiveness, controlling for individual driver consistency and performance trends over the season.

Independent Variable

In this study, the independent variables are those factors hypothesized to exert influence on the dependent variable. Specifically, the research model incorporates two independent variables. These variables are critical in exploring their respective impacts on the dependent variable, as delineated in the various scenarios of this analysis.

Intra Team Position

In the context of Formula 1 racing, each of the ten teams fields two drivers, typically categorized as a first driver and a second driver. Our regression analysis focuses on these second drivers. While teams generally do not publicly declare the hierarchy between their drivers, in certain cases, it is relatively apparent. For instance, within the Mercedes team, it is widely recognized that Lewis Hamilton holds the position of the primary driver. To systematically determine the intra-team driver hierarchy for the purpose of this analysis, the annual salary is used as the deciding criterion. The driver with the lower salary in comparison to his teammate is designated as the second driver in the team.

Pit Order

The hypothesis concerns the impact of the order in which drivers are called into the pit lane during a race. To analyze this, drivers are classified into two categories: first-pitted and second-pitted. This classification is derived from the recorded pit-in times of the drivers. By examining these times, we can ascertain the sequence of pit stops for each team and thus determine whether a driver was

hat gelöscht: $position\ change = average\ position - actual\ position$ ()¶

hat gelöscht:)¶

Kommentiert [VF36]: @Simon Leonard Grube Lets use the same format for the whole document.

the first or the second to pit during a race. A constraint for these features is that both drivers must pit at least once and continue the race after the pit stop.

In addition, the regression model includes an *interaction* term between the two independent variables, which analyses the impact of both independent variables together. This is essential to capture any interaction effects that may be present, as the joint consideration of these variables is crucial for analyzing the hypothesized effect. The interaction term is calculated the following:

$$interaction = intra\_team\_position * pit\_order \quad (4)$$

Control Variables

These variables are integrated into the model to control for the effects of confounding variables and to enable a more accurate estimation of the impact of the independent variables. Although they are independent variables in their own right, the focus of the analysis is not on their specific influence but rather on how they modulate the relationship between the primary independent variables and the dependent variable. The following control variables have been included in the regression models:

- *Position Gain*: Reflects the difference end position relative to their starting position.
- *Pit Time and Number of Pit Stops*: These variables provide insights into the pit stop strategy and its execution.
- *Number of Corners*: Considers the complexity and technical demands of the circuit.
- *Grid position:* Represents the start position of a driver, resulting from the qualifying.
- *Tire Choice*: Different tire compounds affect performance under varying track and weather conditions. Examining the tire choice in the lap following a pit stop reveals how the tire selection strategy aligns with the actual race conditions.

Multiple Linear Regression

Following the meticulous selection and definition of variables for our analysis, we have opted for multiple linear regression, precisely Ordinary Least Squares (OLS), as our primary analytical tool. This decision is anchored in the continuous nature of our dependent variable, as well as the two independent variables, rendering multiple linear regression an apt model. Multiple linear regression is particularly well-suited to examining the impact of our identified independent and control variables on driver performance in a quantifiable and interpretable manner. In our regression model, we incorporated both the main independent variables as well as an array of control variables. A critical component of our model is the inclusion of an interaction term between 'Intra Team Position' and 'Pit Position' to capture potential synergistic or antagonistic effects between these variables. This approach allows us to gain a deeper understanding of how the combination of these factors might influence race performance. In addition to the main effects, control variables such as performance metrics, and tire choices are integrated to account for the impacts of other pertinent factors. Through this meticulous model specification, we ensure that our model adequately represents the complexity and various factors influencing race performance. The next chapter involves testing the models against our data, with a particular focus on validating the models' assumptions to ensure the accuracy and reliability of our findings. First, we are going to conduct a regression analysis with all drivers, followed by regression analysis including only drivers of each cluster. Thus, we are able to compare weather there is an impact on pitting the second driver first throughout different driver behaviour cluster.

In preparation for the regression analysis, it is essential to check the data regarding compliance with the basic assumptions and conditions (Poole and O'farrell 1970). Firstly, the linearity assumption was considered. This is fulfilled by definition for binary variables. Furthermore, a thorough examination of correlations between all variables was meticulously undertaken. This

Kommentiert [VF37]: New favourite word? ;).

critical step ensured the appropriateness of each variable for inclusion in the linear regression model, mitigating concerns of multicollinearity which could otherwise skew the results. The correlation analysis provided a foundational understanding of the interrelationships among variables, reinforcing the robustness and validity of the subsequent regression analysis. Next, the homoscedasticity of the residuals was analysed. Figure 8 illustrates that the variance of the residuals remains constant and that this condition is, therefore, fulfilled. Another crucial aspect is the normal distribution of the residuals. The distribution was checked using a histogram and a Q-Q plot shown in Figure 9 and 10. The histogram shows that the residuals are approximately normally distributed. The Q-Q plot confirms this, except for some slight deviations in certain races, which are visible on the bottom left in the *Normality of Residuals: Q-Q Plot* in Figure 10. These outliers represent driver who have very unusual position change between the start position and the end position, like Hamilton loosing 13 positions. We deliberately leave outliers in the data because this reflects the reality of Formula 1 racing. The multicollinearity check revealed that the independent variables have a Variance Inflation Factor (VIF) values below three, indicating acceptable multicollinearity. Only values above five indicate high multicollinearity. The tire compound variables are an exception, exhibiting infinitely high VIF values due to their mutual exclusivity. This high value is expected in the context of Formula 1, as a driver can only drive one type of tire at a time. The last assumption to be tested concerns the independence of the residuals, which was evaluated using the Durbin-Watson coefficient. This coefficient usually varies between zero and four, with a value of about two indicating the absence of autocorrelation, while values close to 0 or four indicate positive or negative autocorrelation, respectively. Our analysis's calculated Durbin-Watson value of 1.531 indicates a slightly positive autocorrelation, which is quite common in real live data. Now that all assumptions are fulfilled, we proceed to the regression analysis in the following section.

Kommentiert [VF38]: Lasst uns das hier noch einmal checken. Ich habe teilweise andere Grenzen gelesen. Final sollten wir uns einig bezüglich der Thresholds sein, um potentielle Fragen einheitlich beantworten zu können. Eine Quelle wäre gut.

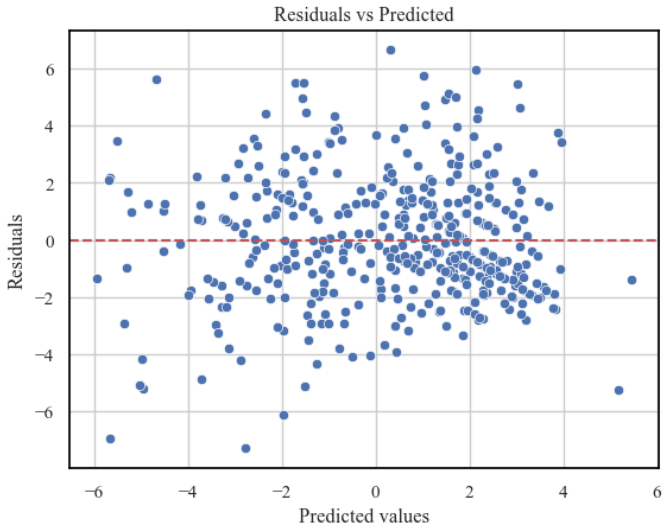

Figure 4: Homoscedasticity Check: Residuals vs Predicted

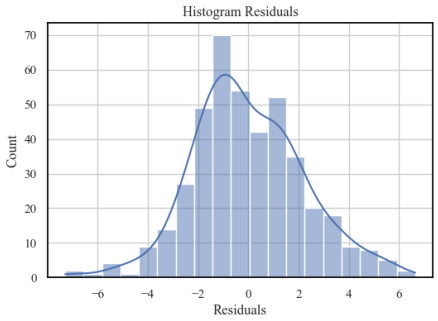

Figure 5: Normality of Residuals: Histogram

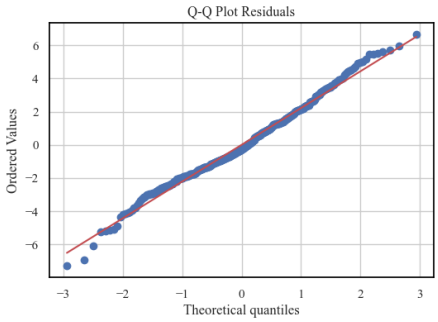

Figure 6: Normality of Residuals: Q-Q Plot

Primary Regression Model

In the OLS regression carried out to investigate the influence of the order of the pit stops *(pit_order*) and the internal team position (*team_position*) - both modelled as dummy variables - on the position change, an R-squared of 0.514 was obtained, indicating that the model can explain around 51.4% of the variance in the dependent variable (see Table 13). In the context of Formula 1, where a multitude of variables impact race outcomes, achieving a comprehensive dataset is challenging. Given the absence of detailed car-specific data and other critical factors, reaching a higher R-squared value with the available dataset is improbable. This limitation highlights the complexity of accurately modelling performance in such a technologically intricate sport. The dummy variable *pit_order* showed a significant effect on the dependent variable with a coefficient of 0.6702 and a p-value of 0.034. This indicates that the driver called into the box first gains an average of 0.67 places compared to his season average. This advantage may be related to the fact that the driver got new tires earlier and, therefore, drives less time on worn tires, which are usually slower. The *team_position* variable was also significant, with a coefficient of 0.9857 and a p-value of 0.002. That means that a driver who is categorised as the second driver in the team experiences on average, a greater change in position compared to the first driver in the team. It should be noted that those classified as second drivers are usually further back in the field than their team partner. As a result, these riders generally have a better chance of gaining positions than those who are automatically very far in front of the field. However, this only applies to teams where one of the two drivers is always placed far forward in the field. The *interaction* between *pit_order* and *team_position* was significant with a coefficient of -1.0235, and a p-value of 0.024, indicating that the complex dynamics within teams do significantly influence race outcomes. It follows that being the second driver to be called into the pit first has an impact on the dependent variable, meaning that when the

Kommentiert [VF39]: Das ist nur zutreffend für die Zeit in der der andere Fahrer noch nicht gepitted hat oder? Sollten die Fahrer gleich oft pitten fährt derjenige der zuerst bittet doch zwingend mit den reifen zu einem anderen Zeitpunkt länger

Kommentiert [SG40]: @Valentin Fedor Hettmann glaubst du ich kann das so staten. Oder zu weit hergeholt ?

second driver is pitted first, he will lose around one position compared to his seasonal average. Thus, we can reject H0, stating the order of pit stops significantly influences the normalized race result of a team's second-placed driver. Although the result is significant, it is essential to remember that correlation does not mean causation. It may be that the team management is trying out strategies with the second driver, which can lead to some strategies not proving favourable and the driver performing worse than usual. However, losing positions can also depend on many other factors and less on pitstop prioritisation, like for example driving skills, the car factor or unpredictable circumstances during the race. Ultimately, the regression results lead to the following equation.

$$y = 0.6702 * pit\_order + 0.9857 * team\_position + \beta * var_{control} + 2.4715 \quad (5)$$

All control variables are represented in a vector summarised as *var control*. Among the control variables *grid_position*, is significant with a negative coefficient of -0.3745 and a p-value of 0.001. This result means that drivers who start the race from worse grid positions are likely to experience less variance in their race positions. Any increase in starting position correlates with a decrease in position compared to the annual average position, indicating the importance of qualifying performance for race outcomes since the performance in the qualifying determines the grid position. In addition, *position_gain*, with a coefficient of 0.5794 and a p-value of 0.001, indicates that position gains during the race are strongly associated with the overall *position_change*. This result suggests that a driver's ability to gain positions during the race is a crucial determinant of their overall position change. It is important to emphasis that although *position_gain* and *position_change* sound similar, they measure different aspects of the race. While *position_gain* reflects a rider's immediate performance in the race by looking at the change from starting to finishing position, *position_change* represents the deviation of a rider's finishing position in a

particular race from their average performance over the season. This distinction is crucial as it allows to analyse the immediate race procedures and their impact on the final result, compared to the seasonal average. The remaining variables, such as *pitstop_duration*, *pit_stops*, *num_corner*, and various compound types, did not show statistically significant effects. This lack of significance could be due to several factors, including the possibility that these aspects are less influential in determining position changes or that the data available for these variables did not capture their impact sufficiently.

Table 7: OLS Regression Result

	(1)	(2)	(3)	(4)	(5)
	ALL	**CL0**	**CL1**	**CL2**	**CL3**
Variable	N=422	N=77	N=106	N=211	N=28
const	2.4715*	4.3631*	1.9043*	5.2821*	10.8032*
position_gain	0.5794*	0.6352*	1.0002*	0.7676*	1.0000*
pitstop_duration	-0.0033	-0.0337	0.0008	-0.0121	-1.596e-16
pit_stops	-0.2010	0.7729	-0.0018	-0.0478	-3.775e-15
num_corner	-0.0088	0.0356	0.0002	-0.0188	9.714e-17
grid_position	-0.3745*	-0.6145*	-0.9960*	-0.6341*	-1.0000*
pit_order	0.6702*	0.6295	0.0286	0.5308	3.719e-15
team_position	0.9857*	0.9086	1.4899*	0.3119	1.1119*
compound_HARD	0.1930	0.4164	0.4717*	0.9359*	3.6011*
compound_INTERMEDIATE	0.8574	1.7378	0.4755*	1.0521	3.245e-16
compound_MEDIUM	0.8347	0.6859	0.4801*	1.3548*	3.6011*
compound_SOFT	0.5561	1.5230	0.4770*	1.1367*	3.6011*
interaction	-1.0235*	-0.4070	-0.0223	-0.9925	-9.992e-16
R-squared	0.514	0.677	1.000	0.745	1.000
Adj. R-squared	0.499	0.617	1.000	0.729	1.000
F-statistic	33.21	11.19	2.348e+04	44.36	3.479e+28
Prob (F-statistic)	5.59e-56	1.29e-11	3.14e-156	2.33e-51	1.36e-238

p values are shown as: * $p < 0.05$

Cluster based Regression Model

Following the main regression analysis, which provided a broad overview of factors impacting race performance, we now shift our focus to cluster-specific analyses. These analyses aim to discern whether different driving behaviors, identified and grouped into clusters, exert varying influences on race outcomes (see Table 13). As we transition from this broader analysis to a more focused cluster-based approach, considerations extend beyond sample size to include statistical robustness and reliability. The main regression analysis benefits from the comprehensive nature of the dataset, but the cluster-specific regressions are limited by the smaller data subsets they represent. Cluster 0, with a relatively large sample size, offers a valuable foundation for regression analysis. However, Cluster 1, despite its larger sample size, presents challenges that compromise the validity of its regression model. These challenges include signs of overfitting, issues with autocorrelation, and high multicollinearity in certain variables. Such statistical concerns necessitate the exclusion of Cluster 1 from further detailed analysis to ensure the integrity of our findings. Cluster 2, as the largest subset, provides a robust sample for analysis, yet it cannot fully replicate the diversity and complexity captured in the complete dataset. The most striking sample size limitation is observed in Cluster 3, which, due to its significantly smaller size, raises concerns regarding the statistical robustness and reliability of any regression analysis. The reduced number of 28 observations in Cluster 3, alongside significant multicollinearity and signs of autocorrelation, increases the susceptibility to anomalies. This potentially skews the results and fails to represent the broader patterns inherent in Formula 1 racing. Consequently, to uphold the analytical rigor, Clusters 1 and 3 were excluded from the detailed regression analysis.

In Cluster 0, focusing on endurance and tactical positioning, the regression model reveals an R-squared of 0.677. The significant predictors here are *position_gain* and *grid_position*. However, *team_position* does not emerge as a significant factor, suggesting that whether a driver is

considered the first or second within their team does not markedly impact their position changes. Similarly, *pit_order* and the *interaction* between *pit_order* and *team_position* do not show significant influence, indicating that the strategies around pit stop order, in combination with the driver's status in the team, are not key determinants of race outcomes for these drivers. In Cluster 2, where drivers exhibit a balance between performance and strategy, the model shows an R-squared of 0.745. Significant predictors include *position_gain and grid_position*. The absence of significance in team*_position* implies that the internal team hierarchy does not have a considerable impact on these drivers' abilities to change positions during a race. Moreover, the non-significance of *pit_order* and the *interaction* term suggests that the combined effect of pit stop order and team position is not a primary influencer of race outcomes for this cluster.

Kommentiert [VF41]: suggest anstatt suggests

The previous findings indicate that in the primary regression, the order of pit stops and the internal team position significantly influence race position. However, this finding is not corroborated in the cluster-based regressions, where neither the order of pit stops nor team position shows a significant impact on race outcomes. The acceptance of the null hypothesis in each cluster-based regression suggests that pitting the second driver first does not significantly influence the race results. This insignificance could be attributed to the smaller sample size of the clusters. Smaller samples are more prone to statistical anomalies and can lead to less accurate results. This suggests that further research with larger samples is necessary to draw definitive conclusions.

An important aspect to consider in interpreting the results is that correlation does not imply causation. Although significant relationships between certain variables and race performance were identified, this does not necessarily mean that one variable directly influences the other. In a sport as complex and dynamic as Formula 1, many different factors individual driver skill, car

performance, and race-day conditions influence race outcomes. Therefore, caution is advised before drawing direct causal conclusions from the correlations.

Despite these limitations, the results of the regressions could be of practical use to Formula 1 teams. Insights into the influence of pit stop order and internal team position can help teams refine their strategies and potentially improve their race performance. In particular, considering the internal team dynamics and the optimal pit stop strategy for the second driver could lead to an overall improved team performance.

That leads us directly to the future research. It would be beneficial to extend the scope of our analysis beyond the impact on the second driver. Investigating whether similar influences affect the performance of the first driver or the overall team results would provide a more holistic understanding of Formula 1 dynamics. This expanded focus could reveal important distinctions in how strategy variables influence different roles within a team.

7. Discussion

In the initial phase of our study, we explored the application of advanced analytics in Formula 1, focusing on how driver clustering might inform pit-stop strategies. Our methodical approach to clustering revealed four distinct driver categories within the dataset. This classification emerged from a comprehensive analysis incorporating three key dimensions: performance, tactical, and behavioural patterns. These dimensions encompassed a variety of metrics, each contributing to a deeper understanding of driver profiles.

Despite the high-performance level uniformly exhibited by the drivers, which led to some similarity in cluster characteristics, our analysis succeeded in distinguishing between nuanced aspects of

driver behaviour and strategy. The resulting profiles offered four differentiated interpretations of driver profiles and decision-making styles.

Subsequently, these clusters were instrumental in underpinning statistical analyses related to pit stop strategies. Our investigation spanned the critical area - driver prioritization. Through this analysis, we established associations between the divergent driver profiles and the outcomes of our analytical models. These connections allowed us to attribute specific strategic decisions and outcomes to identifiable cluster characteristics, thereby providing a deeper insight into the interplay between driver behaviour and pit stop strategy.

The research findings underscore the central role of driver clustering in enhancing the efficacy of analytical models within Formula 1. Notably, the application of driver clusters in our pit stop prediction models provided a more detailed understanding, enabling us to trace and explain strategic decisions in relation to specific driver characteristics. Similarly, the incorporation of these clusters into hypothesis tests - particularly those examining strategies for pitting the first versus the second driver, and decisions made under safety car conditions - brought a different perspective to our statistical analysis. This approach yielded insights that would likely remain obscured under a more generalized analytical framework.

Kommentiert [DP42]: ich änder das in die schreibweise: pit stop

These results from our study imply the critical importance of tailoring analytical models to reflect the distinct characteristics of drivers and teams in Formula 1. This customization is vital for capturing essential features and characteristics that might otherwise be overlooked in a broader, more generalized approach. Our findings suggest that the success of strategic models in Formula 1 hinges on their ability to account for the unique attributes and behaviours of individual drivers and teams.

This study, while thorough, encounters several limitations that should be considered when interpreting the results. Primarily, the availability of telemetry data through the FastF1 API, limited to the 2019 season onwards, restricts the temporal scope of our analysis of driver clusters and its evolution. Although future seasons will incrementally enrich our dataset, the absence of pre-2019 data constrains our historical analysis.

Additionally, a significant unknown in our study is the detailed information on car characteristics, which likely influences driver performance and tactics and ultimately the resulting clusters. The unavailability of these specifics, typically kept confidential by teams, may limit our understanding of the technical factors impacting driver behaviour. We attempted to mitigate this through feature engineering and normalization techniques but acknowledge that this is an approximation.

Another deliberate exclusion for our initial cluster analysis was races affected by rain. This decision was based on the disproportionate representation of wet races in our dataset, which could potentially skew the results. While this aids in maintaining data consistency, it omits the distinct strategic dynamics and behaviours prevalent in wet conditions. Furthermore, the challenge of quantifying track difficulty, despite having data on corners and marshal lights since 2019, presented another limitation. Our approach to normalising this data was a method to address this issue, but it remains uncertain.

These limitations underscore the need for cautious interpretation of our findings, particularly in their application to strategic decision-making in Formula 1. The clustering methodology applied for the 2019 to 2021 seasons and the resulting application on our advanced analytical models demonstrates promise for future applicability, yet it's important to consider these constraints and assumptions. Despite these challenges, we believe our methodology is robust and adaptable for future seasons.

Taking into account the dynamic nature of motorsport in general and Formula 1 in particular, which is subject to constantly evolving strategies, technologies and rules, our work is not intended to be a final solution, but rather food for thought for future work and analysis. As the scope of the data grows over time, there is the potential to deepen and develop our methodology. This promises even deeper and more differentiated insights into the strategic complexity of Formula 1 and motorsports. One key area is the expansion of the dataset. With the FastF1 API continually updating, incorporating data from subsequent seasons would not only improve the robustness of our clustering model but also facilitate longitudinal analyses. Such studies could track the evolution of driver strategies and team tactics over time. If telemetry data for seasons before 2019 becomes accessible, it would offer valuable historical insights, enriching our understanding of the sport's strategic development in response to technological and regulatory shifts.

Rain-affected races, excluded from our current analysis, represent a distinct strategic element in Formula 1. Future research could delve into these scenarios, perhaps through specialized models or by incorporating new features into existing frameworks. This focus could unveil how teams and drivers adapt to the challenges posed by variable weather conditions.

Another promising opportunity is the integration of track characteristics into the analysis. Understanding the influence of different track designs and surface conditions on driver performance and pit stop strategies would add considerable depth to our model. Additionally, while data on specific car builds is largely confidential, future collaborations with Formula 1 teams or utilization of public data could shed light on the relationship between car technology, driver skills, and strategic choices. Also, the field of data analytics and machine learning is rapidly advancing, offering the potential for refining our clustering model with more sophisticated algorithms. These

advancements could handle larger and more complex datasets, providing finer-grained insights into driver performance and pit strategies.

As Formula 1 continues to evolve, so does the opportunity for deeper analytical exploration. The advancements in data collection and analysis present fertile ground for enriching our understanding of this complex sport. By adapting to these changes and refining our methodologies, we can contribute not only to academic discourse but also to the practical strategic toolkit of teams and drivers. This research represents a step towards a more data-driven and detailed comprehension of Formula 1 racing, setting the stage for future scholarly and practical advancements in the field.

Group Part

8. Conclusion

This thesis has delved into the realm of Formula 1 pit stop strategies, employing advanced analytics to understand how driver clustering informs strategic decisions. We discovered four distinct driver categories based on performance, tactical, and behavioural dimensions, providing a nuanced view of driver profiles within the sport. This classification served as a foundation for examining the influence of different pit strategies. The research illustrates the impact of driver characteristics on pit stop strategies in Formula 1.

This study enhances our comprehension of how individual behaviours and decisions intertwine with team strategies by linking divergent driver profiles to strategic outcomes. Using driver clusters in statistical analyses has shed light on the complexities of strategic decision-making in this high-speed sport. This thesis contributes to the understanding of Formula 1 racing by offering a data-informed perspective on the strategic elements of the sport. It underscores the value of bespoke

strategies that consider the distinct qualities of drivers and teams, highlighting the potential of analytics in refining racing tactics.

In summary, this thesis provides an insightful exploration of the strategic dimensions of Formula 1 pit stops. It offers a clearer picture of how data analytics can be applied to decode the intricacies of racing strategies, enriching our understanding of this dynamic and technologically advanced sport.

References

Andreas Reiners. 2019. “Lewis Hamilton Top-Verdiener, Talente Mit Mini-Gehalt: Das Bekommen Die Formel-1-Stars.” 2019. https://web.de/magazine/sport/formel-1/lewis-hamilton-top-verdiener-talente-mini-gehalt-bekommen-formel-1-stars-36379892.

Anıl Duman, Eyüp, Bahar Sennaroğlu, and Gülfem Tuzkaya. 2021. “A Cluster Analysis of Basketball Players for Each of the Five Traditionally Defined Positions.” *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*. https://doi.org/10.1177/17543371211062064.

Arthur, David, and Sergei Vassilvitskii. 2007a. *K-Means++: The Advantages of Careful Seeding*. *Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms*. Vol. 8. https://doi.org/10.1145/1283383.1283494.

———. 2007b. *K-Means++: The Advantages of Careful Seeding*. *Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms*. Vol. 8. https://doi.org/10.1145/1283383.1283494.

Bai, Zhongbo, and Xiaomei Bai. 2021. “Sports Big Data: Management, Analysis, Applications, and Challenges.” *Complexity* 2021: 1–11. https://doi.org/10.1155/2021/6676297.

Bekker, J., and W. Lotz. 2009. “Planning Formula One Race Strategies Using Discrete-Event Simulation.” *Journal of the Operational Research Society* 60 (7): 952–61. https://doi.org/10.1057/palgrave.jors.2602626.

Bell, Andrew, James Smith, Clive E. Sabel, and Kelvyn Jones. 2016. “Formula for Success: Multilevel Modelling of Formula One Driver and Constructor Performance, 1950-2014.” *Journal of Quantitative Analysis in Sports*. https://doi.org/10.1515/jqas-2015-0050.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45: 5–32. https://doi.org/http://dx.doi.org/10.1023/A:1010933404324.

Celebi, M. Emre, Hassan A. Kingravi, and Patricio A. Vela. 2013. “A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm.” *Expert Systems with Applications* 40 (1): 200–210. https://doi.org/10.1016/J.ESWA.2012.07.021.

Chen, Rung Ching, Christine Dewi, Su Wen Huang, and Rezzy Eko Caraka. 2020. “Selecting Critical Features for Data Classification Based on Machine Learning Methods.” *Journal of Big Data* 7 (1): 1–26. https://doi.org/10.1186/s40537-020-00327-4.

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August-2016:785–94. Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785.

Colunga, Ivan Fernandez, and Andrew Bradley. 2014. “Modelling of Transient Cornering and Suspension Dynamics, and Investigation into the Control Strategies for an Ideal Driver in a Lap Time Simulator.” *Proceedings of the*

Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering 228 (10): 1185–99. https://doi.org/10.1177/0954407014525362.

Cui, Yixiong, Miguel Ángel Gómez, Bruno Gonçalves, and Jaime Sampaio. 2019. "Clustering Tennis Players' Anthropometric and Individual Features Helps to Reveal Performance Fingerprints." *European Journal of Sport Science* 19 (8): 1032–44. https://doi.org/10.1080/17461391.2019.1577494.

Daoud, Jamal I. 2018. "Multicollinearity and Regression Analysis." In *Journal of Physics: Conference Series*. Vol. 949. Institute of Physics Publishing. https://doi.org/10.1088/1742-6596/949/1/012009.

Dieter Recken. 2021. "F1 Driver Salaries 2021." February 2021. https://racingnews365.com/formula-1-driver-salaries-2021.

Dindorf, Carlo, Eva Bartaguiz, Freya Gassmann, and Michael Fröhlich. 2023. "Conceptual Structure and Current Trends in Artificial Intelligence, Machine Learning, and Deep Learning Research in Sports: A Bibliometric Review." *International Journal of Environmental Research and Public Health* 20 (1). https://doi.org/10.3390/ijerph20010173.

D'Urso, Pierpaolo, Livia De Giovanni, and Vincenzina Vitale. 2023. "A Robust Method for Clustering Football Players with Mixed Attributes." *Annals of Operations Research* 325 (1): 9–36. https://doi.org/10.1007/s10479-022-04558-x.

Ergast. 2009. "Ergast Developer API." 2009. http://ergast.com/mrd/.

"FastF1 3.1.6." n.d. Accessed December 18, 2023. https://docs.fastf1.dev/.

FIA. 2011. “FIA Formula One World Championship Power Unit Regulations.” FEDERATION INTERNATIONAL DE L’AUTOMOBILE.

Ghosh, Indrajeet, Sreenivasan Ramasamy Ramamurthy, Avijoy Chakma, and Nirmalya Roy. 2023a. “Sports Analytics Review: Artificial Intelligence Applications, Emerging Technologies, and Algorithmic Perspective.” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. John Wiley and Sons Inc. https://doi.org/10.1002/widm.1496.

———. 2023b. “Sports Analytics Review: Artificial Intelligence Applications, Emerging Technologies, and Algorithmic Perspective.” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 13 (5). https://doi.org/10.1002/widm.1496.

Gopal, S, Krishna Patro, and Kishore Kumar Sahu. 2015. “Normalization: A Preprocessing Stage.”

Guyon, Isabelle, and André Elisseeff. 2003. “An Introduction to Variable and Feature Selection André Elisseeff.” *Journal of Machine Learning Research* 3: 1157–82.

Heilmeier, Alexander. 2020. “F1-Timing-Database: SQLite Database Containing Formula 1 Lap and Race Timing Information for the Seasons 2014 - 2019.” 2020. https://github.com/TUMFTM/f1-timing-database.

Heilmeier, Alexander, Maximilian Geisslinger, and Johannes Betz. 2019. “A Quasi-Steady-State Lap Time Simulation for Electrified Race Cars.” In *2019 Fourteenth International Conference on Ecological Vehicles and Renewable Energies (EVER)*. IEEE. https://doi.org/10.1109/EVER.2019.8813646.

Heilmeier, Alexander, Michael Graf, Johannes Betz, and Markus Lienkamp. 2020a. "Application of Monte Carlo Methods to Consider Probabilistic Effects in a Race Simulation for Circuit Motorsport." *Applied Sciences (Switzerland)* 10 (12). https://doi.org/10.3390/app10124229.

———. 2020b. "Application of Monte Carlo Methods to Consider Probabilistic Effects in a Race Simulation for Circuit Motorsport." *Applied Sciences (Switzerland)* 10 (12). https://doi.org/10.3390/app10124229.

Heilmeier, Alexander, Michael Graf, and Markus Lienkamp. 2018a. "A Race Simulation for Strategy Decisions in Circuit Motorsports." In *2018 21st International Conference on Intelligent Transportation Systems (ITSC) Maui, Hawaii, USA, November 4-7, 2018*. Maui.

———. 2018b. "A Race Simulation for Strategy Decisions in Circuit Motorsports." In *21st International Conference on Intelligent Transportation Systems*, 2986–93. Maui.

———. 2018c. "A Race Simulation for Strategy Decisions in Circuit Motorsports." In . Maui.

Heilmeier, Alexander, André Thomaser, Michael Graf, and Johannes Betz. 2020a. "Virtual Strategy Engineer: Using Artificial Neural Networks for Making Race Strategy Decisions in Circuit Motorsport." *Applied Sciences (Switzerland)* 10 (21): 1–32. https://doi.org/10.3390/app10217805.

———. 2020b. "Virtual Strategy Engineer: Using Artificial Neural Networks for Making Race Strategy Decisions in Circuit Motorsport." *Applied Sciences (Switzerland)* 10 (21): 1–32. https://doi.org/10.3390/app10217805.

Hughes, Mike, and Ian MFranks. 2004. “Notational Analysis of Sport Second Edition: Systems for Better Coaching and Performance in Sport.”

Jain, Anil K. 2010. “Data Clustering: 50 Years beyond K-Means.” *Pattern Recognition Letters* 31 (8): 651–66. https://doi.org/10.1016/J.PATREC.2009.09.011.

Karsmakers, Peter, Kristiaan Pelckmans, and Johan A.K. Suykens. 2007. “Proceedings of International Joint Conference on Neural Networks.” In . Orlando, Florida, USA. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4371223&casa_token=j51q_Wvzg0oAAAAA:4E-IrAi9llQKvU_QX31TxpN7fFWL4oKEdJnZP_hYxGDnzI6RJ_WPezJmooxcnG0NeOt5jetuRw&tag=1.

Kesteren, Erik Jan Van, and Tom Bergkamp. 2023. “Bayesian Analysis of Formula One Race Results: Disentangling Driver Skill and Constructor Advantage.” *Journal of Quantitative Analysis in Sports*. https://doi.org/10.1515/jqas-2022-0021.

Kulkarni, Vrushali Y, Pradeep K Sinha, and Manisha C Petare. 2013. “Weighted Hybrid Decision Tree Model for Random Forest Classifier.” *The Institution of Engineers* 97 (2): 209–17. https://doi.org/10.1007/s40031-014-0176-y.

Li, Jundong, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. 2017. “Feature Selection: A Data Perspective.” *ACM Comput. Surv* 50 (6): 94–139. https://doi.org/10.1145/3136625.

Madison Pearce. 2020. “F1 2020: Driver Salaries.” Fox Sports. 2020. https://www.foxsports.com.au/motorsport/formula-one/f1-2020-driver-salaries-how-much-does-daniel-ricciardo-earn-lewis-hamilton-money-wages-richest-f1-drivers/news-story/e1ef97e9eb4fcf8dcf2792c168d64d7e.

Morgulev, Elia, Ofer H. Azar, and Ronnie Lidor. 2018. “Sports Analytics and the Big-Data Era.” *International Journal of Data Science and Analytics* 5 (4): 213–22. https://doi.org/10.1007/s41060-017-0093-7.

Muniz, Megan, and Tulay Flamand. 2022. “A Weighted Network Clustering Approach in the NBA.” *Journal of Sports Analytics* 8 (4): 251–75. https://doi.org/10.3233/jsa-220584.

Nadikattu, Rahul Reddy. 2020. “IMPLEMENTATION OF NEW WAYS OF ARTIFICIAL INTELLIGENCE IN SPORTS.” *Journal of Xidian University* 14 (5). https://doi.org/10.37896/jxu14.5/649.

Patel, Vaishali R., and Rupa G. Mehta. 2011. “Impact of Outlier Removal and Normalization Approach in Modified K-Means Clustering Algorithm.” *IJCSI International Journal of Computer Science Issues* 8 (5): 331–36.

Pedregosa, Fabian, Vincent Michel, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Jake Vanderplas, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12: 2825–30. http://scikit-learn.sourceforge.net.

Pedregosa FABIANPEDREGOSA, Fabian, Vincent Michel, Olivier Grisel OLIVIERGRISEL, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Jake Vanderplas, et al. 2011. “Scikit-Learn: Machine Learning in Python Gaël

Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot." *Journal of Machine Learning Research* 12: 2825–30. http://scikit-learn.sourceforge.net.

Phillips, A. 2014. "Building a Race Simulator." https://f1metrics.wordpress.com/2014/10/03/building-a-race-simulator/.

Phillips, Andrew J.K. 2014. "Uncovering Formula One Driver Performances from 1950 to 2013 by Adjusting for Team and Competition Effects." *Journal of Quantitative Analysis in Sports* 10 (2): 261–78. https://doi.org/10.1515/jqas-2013-0031.

Poole, Michael A, and Patrick N O'farrell. 1970. "The Assumptions of the Linear Regression Model."

Rockerbie Duane, and Easton Stephen. 2021. "Race to the Podium: Separating and Conjoining the Car and Driver in F1 Racing."

Salminen, Tomi. 2019. "Race Simulator: Downloadable R Program Code." 2019. https://f1strategyblog.wordpress.com/2019/05/10/race-simulator-downloadable-r-program-code/.

Shapiro, Joel. 2023. "Data Driven at 200 MPH: How Transforms Formula One Racing." January 26, 2023. https://www.forbes.com/sites/joelshapiro/2023/01/26/data-driven-at-200-mph-how-analytics-transforms-formula-one-racing/.

Siegler, Blake, Andrew Deakin, and David Crolla. 2000. "Lap Time Simulation: Comparison of Steady State, Quasi-Static and Transient Racing Car Cornering

Strategies." In *SAE Motorsports Engineering Conference & Exposition*. https://doi.org/10.4271/2000-01-3563.

Sinadia, Herbie Ewaldo, and I. Made Murwantara. 2022. "Sports Analytics: A Comparison of Machine Learning Perfomance for Profiling Badminton Athlete." In *Proceedings - 2022 1st International Conference on Technology Innovation and Its Applications, ICTIIA 2022*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICTIIA54654.2022.9935852.

Tan, Xiaomeng. 2023. "Enhanced Sports Predictions: A Comprehensive Analysis of the Role and Performance of Predictive Analytics in the Sports Sector." *Wireless Personal Communications*, October. https://doi.org/10.1007/s11277-023-10585-z.

theOehrly. 2020. "FastF1." 2020. https://docs.fastf1.dev/index.html.

Timings, Julian, and David Cole. 2014. "Robust Lap-Time Simulation." In *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 228:1200–1216. SAGE Publications Ltd. https://doi.org/10.1177/0954407013516102.

Tulabandhula, Theja, and Cynthia Rudin. 2014. "Tire Changes, Fresh Air, And Yellow Flags: Challenges in Predictive Analytics For Professional Racing." *Big Data* 2 (2): 97–112. https://doi.org/10.1089/big.2014.0018.

Zaidi, Abdelhamid, and Asamh Saleh M. Al Luhayb. 2023. "Two Statistical Approaches to Justify the Use of the Logistic Function in Binary Logistic

Regression." *Mathematical Problems in Engineering* 2023 (April): 1–11. https://doi.org/10.1155/2023/5525675.

Zhang, Yongli. 2012. "Support Vector Machine Classification Algorithm and Its Application." *Communications in Computer and Information Science* 308 CCIS (PART 2): 179–86. https://doi.org/10.1007/978-3-642-34041-3_27/COVER.

Zhao, Zibin. 2023. "Transforming ECG Diagnosis:An In-Depth Review of Transformer-Based DeepLearning Models in Cardiovascular Disease Detection." *Department of Chemical and Biological Engineering*, June. http://arxiv.org/abs/2306.01249.