**Unveiling the Predictive Power of Individual Sector Times in Motorsport: Assessing their Role in Estimating Overall Lap Performance.**

Abhinav Pegallapati, Meghana Vodnala, Kavyasri Janagam, and Sri Ramya Panja

Indiana University-Purdue University, Indianapolis, IN 46202, USA

[apegalla@iu.edu,](mailto:apegalla@iu.edu,) [mvodnala@iu.edu,](mailto:mvodnala@iu.edu,) [kjanagam@iu.edu,](mailto:kjanagam@iu.edu,) [sripanja@iu.edu](mailto:sripanja@iu.edu)

**Abstract.** In the world of competitive racing, winning relies heavily on analyzing and improving lap times. Traditionally, people have only fo- cused on the total lap time, ignoring the valuable information hidden within each section (sector) that composes a lap. This project, titled ”Unveiling the Predictive Power of Individual Sectors in Motorsports,” aims to address this by exploring how well individual sector times (like sectors 1, 2, and 3) can predict a driver’s overall lap performance. In- spired by research on Formula One pit stops and performance analysis, our project aims to completely change the way races are strategized and how driver performance is optimized. Rigorous data collection, preprocessing, and statistical analysis will be employed to construct predictive models. Techniques like linear regression and exploratory data analysis will identify critical factors influencing lap times, yielding valuable in- insights for strategic optimization. Overall, this project aims to enhance track performance and competitiveness by using the hidden potential of individual sector times.

**Keywords:** Racing, lap times, individual sectors, predictive power, data collection, preprocessing, statistical analysis, predictive models, linear regression, track performance

# Project Scope

## Introduction

In the cutthroat world of motorsports, shaving milliseconds off lap times is the difference between glory and defeat. Teams and drivers meticulously dissect ev- ery scrap of data, relentlessly pursuing marginal gains. However, a critical blind spot exists in traditional analytics that solely rely on overall lap times. This approach overlooks the rich data hidden within individual sectors (s1, s2, s3) that make up a lap. Studies like those by Heine and Thraves.on Formula One pit stop strategies demonstrate the power of analyzing specific race com- ponents beyond final lap times. Similarly, focusing solely on aggregate lap times

in motorsports fails to capture the nuances of driver performance across differ- ent track sections. This limitation hinders teams’ ability to identify a driver’s strengths and weaknesses in areas like cornering efficiency or high-speed stabil- ity. This is where our project, ”Unveiling The predictive Power of Individual Sector Times in Motorsport” comes into play. We believe that by delving deeper into this underutilized data source, we can unlock a wealth of hidden insights. Our analysis of individual sector times has the potential to revolutionize race strategy and propel driver performance optimization to new heights. This aligns with the growing emphasis on data-driven decision-making in motorsports, as evidenced by work from experts like Kumar and Preethi **?**, who explored the role of performance analysis in Formula One.

## Aim

This study aims to predict cumulative lap time in motorsport by analyzing sec- tor durations and the top speed of the drivers. Examine performance within each lap segment. Understand the contributions of individual time components to the overall lap time.

Below are our two hypothesis for our research question

Null Hypothesis (H0): There is no significant relationship between sector times, top speed of the drivers, and overall lap performance in motorsport. In other words, the sector times do not provide predictive power for estimating overall lap times.

Alternative Hypothesis (H1): A significant relationship exists between sector times (s1, s2, s3) and overall lap performance in motorsport. In other words, the sector times provide predictive power for estimating overall lap times.

## Purpose

This study aims to explore the connection between overall lap performance and motorsports sector times. The objective is to create predictive models that can assist teams and drivers in optimizing their race strategies by analyzing data specific to our industry. The goal of this study is to analyze industry-specific data to enhance performance and offer insightful information that will improve track performance and competitiveness. We aim to provide practical insights that can guide strategic decision-making and help teams and drivers achieve better results on the track by thoroughly analyzing sector times and their effect on total lap performance.

## Methodology

**Steps of the Project** The methodology of our project involves four stages, which will be explained in more depth later on in the report:

* + 1. **Data Collection**
    2. **Data Extraction and Storage**
    3. **Data Analysis**
    4. **Data Visualization**
  1. **Team Members and Responsibilities**

|  |  |  |
| --- | --- | --- |
| **Name** | **Background** | **Responsibilities** |
| Kavyasri Janagam | Biology | Data collection and storage |
| Sri Ramya Panja | Dentistry | Data extraction and description |
| Abhinav Pegallapati | Dentistry | Statistical analysis and Model de- velopment |
| Meghana Vodnala | Dentistry | Report writing and presentation |

**Table 1.** Team Members and Responsibilities

## Actual Contributions from Team Members

|  |  |  |
| --- | --- | --- |
| **Name** | **Background** | **Responsibilities** |
| Kavyasri Janagam | Biology | Data collection, visualizations, and presentation |
| Sri Ramya Panja | Dentistry | Data visualizations, presentation,  and report writing |
| Abhinav Pegallapati | Dentistry | Statistical analysis, Testing, and presentation |
| Meghana Vodnala | Dentistry | Model Development, presentation,  and report writing |

**Table 2.** Actual Contributions from Team Members

## Project Challenges

* + 1. **Complexity and Dimensionality:** Our dataset’s high dimensionality in- volves numerous features, complicating analysis and visualization. Dimensionality reduction techniques are needed to simplify the dataset, enabling clearer insights by focusing on the most impactful features.
    2. **Data Size:** Handling big data poses computational and storage challenges, necessitating efficient processing techniques. Conversely, smaller data seg ments can undermine statistical power, requiring careful aggregation to ensure robust analysis outcomes.
    3. **Data Quality:** Quality issues such as missing values, inconsistent format- ting, and outliers were addressed through meticulous data cleaning. Stan- dardizing formats and employing statistical methods for outlier detection ensured the reliability and accuracy of our subsequent analyses.

## Data Collection

The data for this analysis was sourced from the FIA World Endurance Cham- pionship lap data available on Kaggle, covering races from 2012 to 2022. The dataset is obtained from Kaggle and is licensed under the Creative Commons Attribution-Non-Commercial-ShareAlike 3.0 Intergovernmental Organization li- cense. Extracted data includes information on race events, lap times, and sector times.

## Data Extraction and Storage

**Data Extraction** We have created a table in the Phpmyadmin involving these columns. (CSV File) The 48 attributes we selected are below:

* s.no
* Number
* Driver number
* Lap number
* Lap time
* Lap improvement
* Crossing finish line in pit
* S1
* S1 improvement
* S2
* S2 improvement
* S3
* S3 improvement
* Kph
* Elapsed
* Hour
* S1 large
* S2 large
* S3 large
* Top speed
* Driver name
* Pit time
* Class
* Group1
* Team
* Manufacturer
* Season
* Circuit
* Round
* Vehicle
* Flag at fl
* Team no
* Lap time ms
* Lap time s
* Engine
* Driver stint no
* Driver stint
* Team stint no
* Team stint
* Elapsed ms
* Position
* Class position
* Interval ms
* Interval1
* Gap
* Class interval
* Class gap
* elapsed s

## Data Dimensions

* Number of columns = 48
* Number of rows = 503680

## Categorical Data

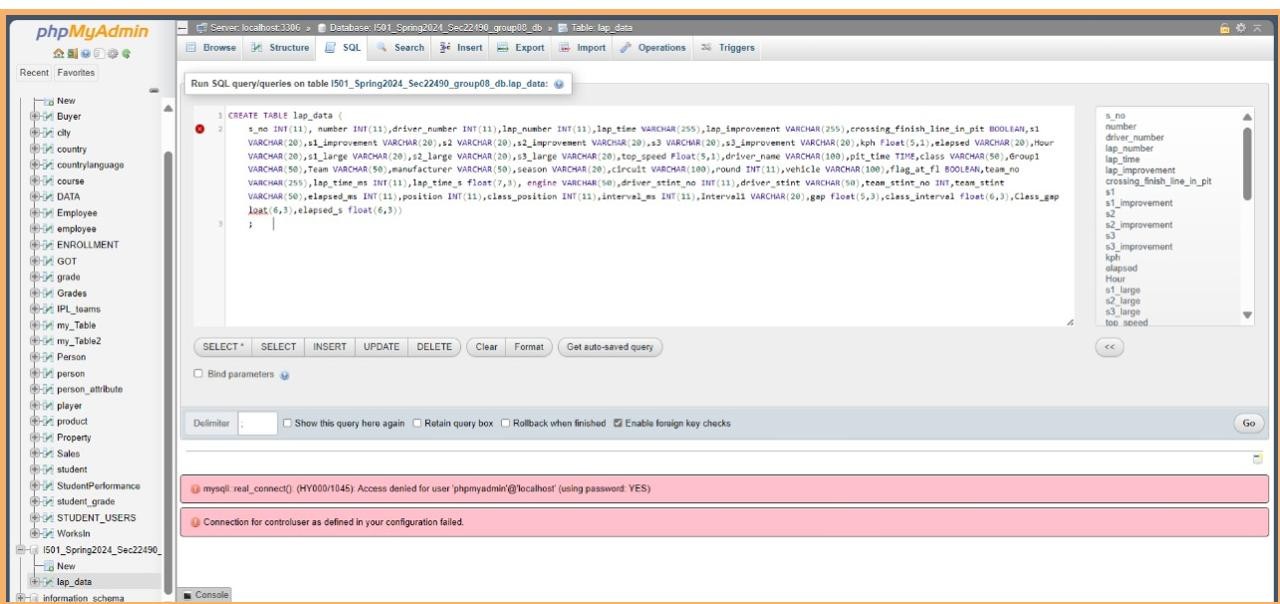
* Driver name, class, team, Group1, manufacturer, circuit, vehicle, team no, engine, driver stint, team stint

## Numerical Data

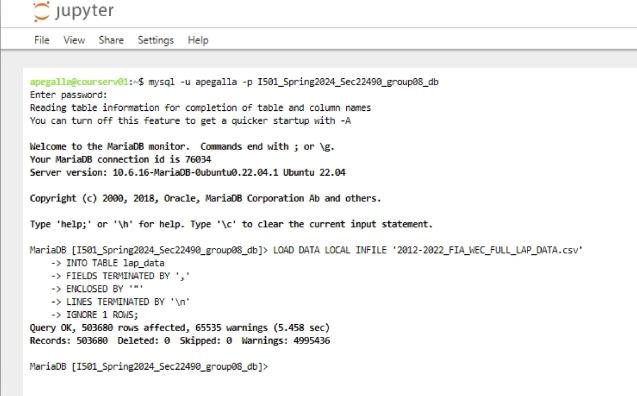
* s.no, driver number, lap number, lap time, lap time, lap\_improvement, s1, s1\_ improvement, s2,s2\_improvement,s3, s3\_improvement, Kph, elapsed, t, Kph, elapsed, hour, top speed, Season, round, lap time ms, lap time s, driver stint no, elapsed ms, position, class position, interval ms, Interval1, gap, class interval, class gap, elapsed s

## 1.1Data Import

We attempted to upload a CSV file into phpMyAdmin but encountered issues due to its large size. Consequently, we imported the CSV file into the Jupyter environment. Using the Jupyter terminal, we established a connection with phpMyAdmin and created a new table. We then uploaded the CSV file into this newly created table in phpMyAdmin, completing the upload process.



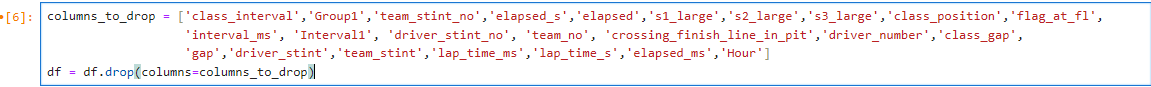
**Fig. 1.** Creating table in PhpMyadmin



**Fig. 2.** Jupiter Terminal Used for Connection

**1.2 Data Cleaning**

The data cleaning process involved reviewing the attributes we selected. This was a very necessary step because our dataset contained a large number of null and missing values. As part of this cleaning process, we dropped the columns that we didn’t want to include in our analysis. We systematically examined our Data Frame to identify and address any instances of missing values.

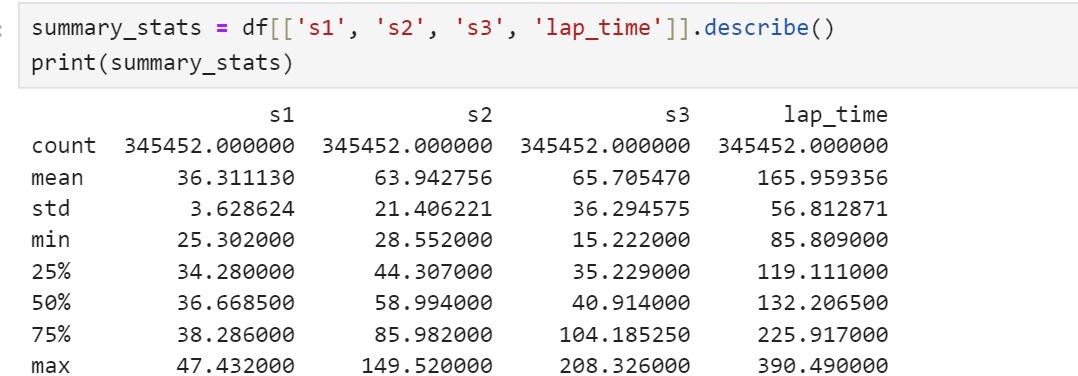


**Fig.3.** Dropped Columns

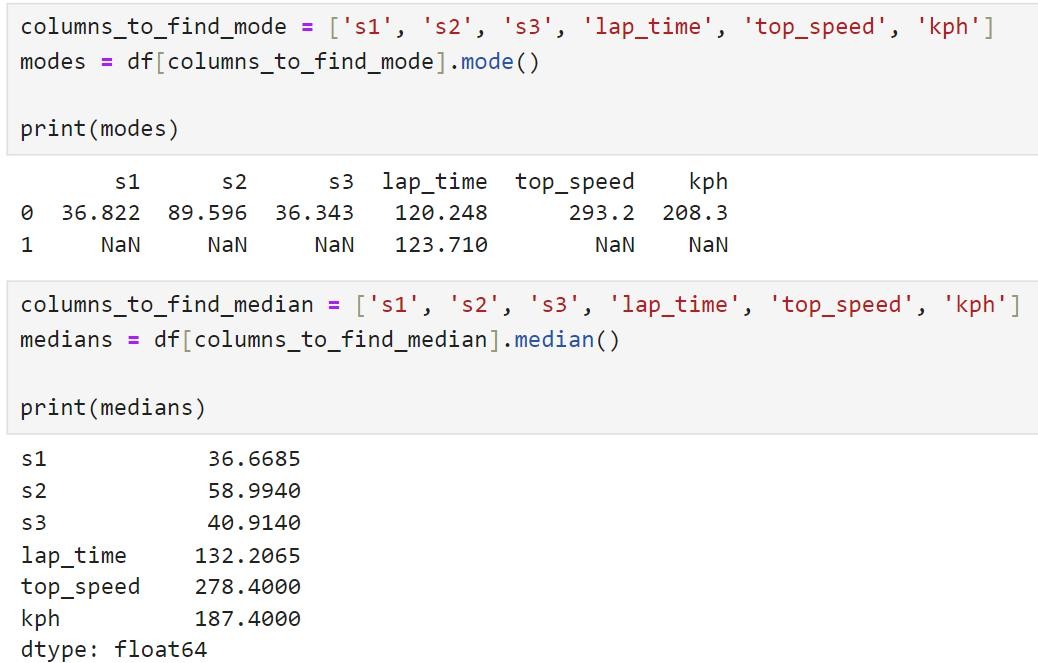
We have found missing values that were dropped before our analysis. We used Pandas functions to calculate descriptive statistics using df.describe() and additionally computed the mode and median. We identified the outliers for each variable through the visualization of box plots. The outliers were dropped to maintain data quality for analysis.



**Fig. 4.** Number of null values in each column



**Fig. 5.** Descriptive Statistics



**Fig. 6.** Descriptive Statistics Containing median and mode

A group of white squares with red lines

Description automatically generated

**Fig.7.** Boxplot representation of outliers

A group of white squares with red dots

Description automatically generated

**Fig.8.** Boxplot representation of outliers

A screenshot of a graph

Description automatically generated

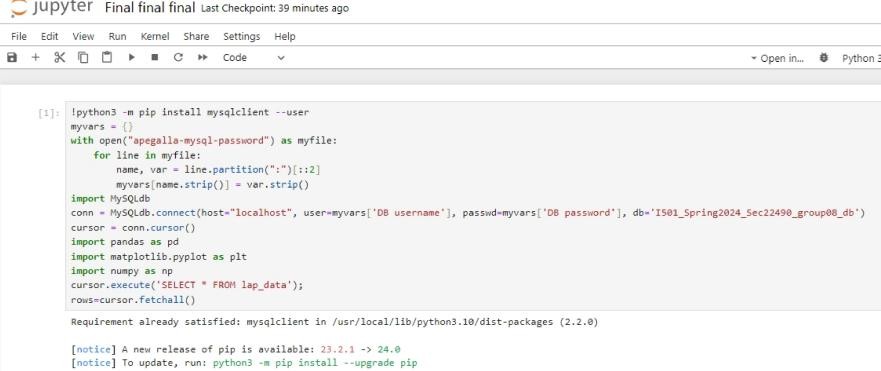
**Fig.9.** Boxplot representation of outliers

## Data Analysis

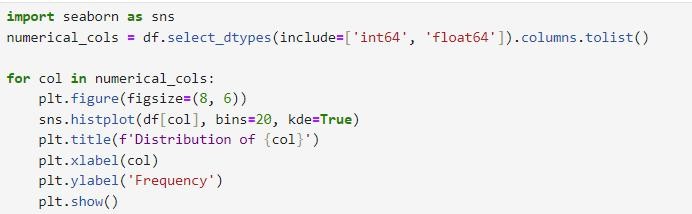
Python was used for data analysis, employing linear regression to analyze the data. Additionally, we implemented predictive models including polynomial re- gression and the gradient boosting technique. Below are some of the steps we took to effectively analyze the data:

**Steps to Test the Normality of the Data Step 1:** For the analysis, we connected phpMyAdmin SQL to the Python Jupyter notebook and imported ”Dataset.csv” into it. The image below shows the data for the independent vari- ables, confirming a successful connection between the database and the Jupyter notebook.

**Step 2:** After establishing a successful connection between the database and the Python Jupyter notebook, we tested whether the data is normally dis- tributed. To do this, we utilized Python’s Seaborn package to plot histograms and pairplots. Initially, we assessed the data distribution by examining histograms to determine if it followed a normal distribution.



**Fig. 10.** Importing dataset into Jupyter from phpMyAdmin



**Fig. 11.** Checking for the distribution

A graph of a distribution of lap time

Description automatically generated

**Fig.12.**Distribution of Lap time

A graph of a distribution of a number

Description automatically generated

**Fig.13.**Distribution of s1

A graph of a distribution of data

Description automatically generated with medium confidence

**Fig.14.**Distribution of s2

A graph of a distribution of data

Description automatically generated

**Fig.15.**Distribution of s3

## Correlation Among Attributes

Our analysis revealed a robust correlation between lap time and sectors S2 and S3, underscoring their critical role in driving lap performance. Sector S1 displayed a modest correlation with lap time, indicating its relatively lower influence on total lap times in our dataset. We observed a notable negative correlation

A computer code with many colorful text

Description automatically generated with medium confidence

**Fig.16.**Code to check correlation

A screenshot of a graph

Description automatically generated

**Fig.17.**Spearman Correlation Heatmap

between sector S1 and average speed, suggesting that higher speeds in S1 do not necessarily contribute to faster lap times. Average speed (kph) showed a strong positive correlation with sector S2, linking higher speeds in this sector to better overall lap performance.

## Linear Regression Model

Our project aimed to enhance lap time prediction for race cars through linear regression analysis. Utilizing sensor data (s1, s3), top speed, and kilometers per hour (kph), we trained a model using sklearn in Python, employing an 80:20 split for training and testing datasets. Our results revealed a highly accurate model, as evidenced by a mean squared error (MSE) of 51.26 and an impressive R-squared value of 0.98. The coefficients elucidated the influence of each feature on lap time prediction, with s1 and s3 sensors demonstrating substantial im- pacts. This project not only provides a reliable tool for lap time prediction but also offers valuable insights for optimizing race strategies and enhancing vehicle performance in the domain of motorsports analytics. Future work may explore additional features or advanced machine learning techniques to further refine predictive capabilities.

A screenshot of a computer program

Description automatically generated

**Fig.18.**Code for Linear Regression

A graph with a red line

Description automatically generated

**Fig.19.**Linear Regression

**Polynomial Regression Model** In our project, we explored Polynomial Regression as a predictive model for lap time estimation, incorporating sector times, top speed, and kilometers per hour (kph) as features. After splitting our dataset into training and testing sets, we trained the Polynomial Regression model with a degree of 2 using sci-kit-learn’s PolynomialFeatures and LinearRegression modules.

Upon evaluation, the model exhibited a Mean Squared Error of 37.66 on the test dataset, reflecting the average squared difference between the predicted and actual lap times. Despite this relatively higher error compared to previous models, the model achieved a commendable R-squared value of 0.988, indicating that approximately 98.8% of the variance in lap times was captured by our model’s predictions.

Overall, while Polynomial Regression showcased strong predictive performance, it’s essential to consider its trade-offs, such as potential overfitting with higher-degree polynomials, in the context of our specific project goals and dataset characteristics.

A screenshot of a computer program

Description automatically generated

**Fig.20.**Code for Polynomial Regressor

A graph with blue dots and a red line

Description automatically generated

**Fig.21.** Polynomial Regression

**Gradient Boosting Model** In our project, we utilized the Gradient Boosting Regressor to predict lap time using sector times as features. With 100 estimators and a learning rate of 0.1, the model achieved a Mean Squared Error of 10.93 on our test dataset. This error represents the average squared difference between the lap times predicted by our model and the actual lap times recorded during testing, indicating the model’s proficiency in capturing the variability in lap times based on sector times.

Furthermore, the high R-squared value of 0.997 suggests that approximately 99.7% of the variance in lap times is explained by our model’s predictions, un- derscoring its robust predictive accuracy and its ability to effectively discern patterns in the sector time data.

These findings highlight the effectiveness of the Gradient Boosting Regres- sor in accurately predicting lap times based on sector times, offering valuable insights for optimizing race strategies and enhancing performance in motorsport competitions.

A screen shot of a computer code

Description automatically generated

**Fig.22.**Code for Gradient Boosting Machine

A graph with a red line

Description automatically generated

**Fig.23.**Gradient Boosting Regressor

**Cross Validation** In our project, we employed K-Fold Cross-Validation, with *k* = 5, to evaluate the performance of the Decision Tree Regressor model for lap time prediction. Utilizing numpy and scikit-learn libraries, we split our training dataset into five folds, ensuring each fold serves as both a training and validation set during iterations.

Upon fitting the Decision Tree Regressor model to each fold, we calculated the Mean Squared Error (MSE) for lap time predictions on the validation sets. The average MSE across all folds was approximately 3.18, with a standard deviation of approximately 0.27.

These results underscore the model’s consistency and reliability in predicting lap times across different subsets of the training data, validating its robustness and generalization capability. By implementing K-Fold Cross-Validation, we ensured a thorough assessment of our model’s performance, enhancing the credibility and applicability of our findings in real-world motorsport scenarios.

A screenshot of a computer program

Description automatically generated

**Fig.24.**Code for cross validation

A graph of a graph of a number of different colored bars

Description automatically generated with medium confidence

**Fig.25.**Train- Test data distribution

**The Shapiro-Wilk Test** The Shapiro-Wilk tests revealed that variables *s*1, *s*2, *s*3, *lap time*, *kph*, and *top speed* exhibit non-normal distributions, as indicated by extremely low p-values (0*.*0), leading to the rejection of the null hypothesis. Consequently, nonparametric tests may be more appropriate for analyzing these variables.

A screenshot of a computer code

Description automatically generated

**Fig.26.** Shapiro-Wilk Test to check the normality of data

**Levene’s Test for Homogeneity of Variance** The results of the test indicate that the variances across the groups (*s*1, *s*2, *s*3, *lap time*, *top speed*, and *kph*) are not homogeneous, as evidenced by a Levene’s test statistic of 142575*.*99362368893 and a p-value of 0*.*0, leading to the rejection of the null hypothesis.

A screenshot of a computer

Description automatically generated

**Fig.27.** Levene’s Test to check Homogeneity

**The Kruskal-Wallis H-Test** The Kruskal-Wallis H-test yielded a test statistic of 1771109*.*2103688098 with a p-value of 0*.*0, leading to the rejection of the null hypothesis of equal distributions. This indicates potential relationships between sector times and lap performance, prompting further analysis for predictive modeling to optimize lap time.

A screenshot of a computer code

Description automatically generated

**Fig.28.**Kruskal-Wallis Test

# Summary of Findings

After thoroughly examining our methodology encompassing data cleaning, analysis, and visualization, we have derived numerous insights pertinent to our research inquiries. Below are some of the significant findings categorized by research question.

## Research Question One Findings

**Research Question:** How do the individual sector durations and top speed influence the prediction of cumulative lap time in motorsport?

**Answer:** Our project aimed to enhance lap time prediction for race cars through various regression analyses, incorporating sensor data, top speed, and kilometers per hour (kph) as features. Initially, we employed linear regression, achieving a mean squared error (MSE) of 51.26 and an R-squared value of 0.98, indicating strong predictive accuracy. Subsequently, we explored Polynomial regression, which yielded an MSE of 37.66 and an R-squared value of 0.988, showcasing robust performance despite potential overfitting concerns. Additionally, we leveraged the Gradient Boosting Regressor, achieving an MSE of 10.93 and an impressive R-squared value of 0.997, underscoring its efficacy in accurately

predicting lap times based on sector times. Our analysis highlighted the critical role of sector times, particularly S2 and S3, in driving lap performance, with S1 displaying a more modest correlation. Notably, we observed a notable negative correlation between sector S1 and average speed, suggesting nuanced relationships between speed and lap times across sectors. Moreover, our findings emphasized the importance of top speed in predicting lap performance, reaffirming its significance in our performance model. These insights offer valuable guidance for optimizing race strategies and enhancing vehicle performance in motorsports competitions.

## Research Question Two Findings

**Research Question:** How do top speed and kilometers per hour (kph) individually and collectively influence the prediction of cumulative lap time in motorsport?

**Answer:** Our project aimed to explore the predictive power of top speed and kph in conjunction with other factors like sector durations in estimating cumulative lap time. Through regression analyses, we found that while sector durations play a significant role in lap time prediction, top speed, and kph also exhibit strong correlations with overall lap performance. Specifically, higher top speeds and kph were associated with improved lap times, indicating their importance in enhancing race performance. Moreover, our findings underscore the need to consider both individual and collective effects of top speed and kph in optimizing race strategies and improving vehicle performance in motorsport competitions.

## Limitations

Dealing with large datasets posed significant challenges throughout our project. Firstly, ensuring the accuracy and reliability of such voluminous data was a daunting task. Additionally, the complexity and high dimensionality of the dataset made analysis and visualization particularly arduous, hampering our ability to effectively interpret results. Moreover, the practical aspect of loading such large data into a database presented a formidable obstacle, especially given the file size limitations we encountered. Overcoming these challenges required extensive support from our professor and teaching assistants (TAs). Furthermore, as a team primarily composed of members from non-technical backgrounds, navigating through the project while simultaneously learning new concepts presented its own set of challenges. However, through collaboration, perseverance, and guidance from our instructors, we were able to overcome these limitations and successfully complete our project.

**Bibliography**

Choo, C. L. W. (2015). Real-time decision making in motorsports: Analytics for improving professional car race strategy <https://dspace.mit.edu/handle/1721.1/100310>

FIA World Endurance Championship lap data 2012-22. (2022, June 23). Kaggle. https://www.kaggle.com/datasets/tristenterracciano/fia-wec-lap-data- 20122022?resource=download

Heilmeier, A., Thomaser, A., Graf, M., & Betz, J. (2020). Virtual strategy engineer: Using artificial neural networks for making race strategy decisions in circuit motorsport. Applied Sciences, 10(21), 7805. <https://doi.org/10.3390/app10217805>

Heine, O. F. C., & Thraves, C. (2022). On the optimization of pit stop strategies via dynamic programming. Central European Journal of Operations Research, 31(1), 239–268. <https://doi.org/10.1007/s10100-022-00806-4>

Kumar, M. K., & Preethi, N. (2023). Formula one race analysis using machine learning in lecture notes in networks and systems (pp. 533–540). <https://doi.org/10.1007/978-981-19-6088-8_47>