Analyzing the effect of the 2022 FIA regulations on race dynamics using Machine Learning

Seminar in Data Analytics

Myunggong Seo

November 2025

Data Analytics Department

Denison University

Table of Contents

Introduction

Formula 1 is recognized as one of the largest and most popular motorsports in the world by revenue and scale. Governed by the FIA, F1 began in 1950 and has been around for nearly 80 years, drawing more than 827 million fans globally in 2025 and setting race attendance records with over 3.9 million live spectators by mid-season (Courtnee Catnott, 2025).. To ensure the competitiveness of the races, the FIA often overhauls regulations, with the latest change on aerodynamics being made in 2022. The primary aim of this project is to quantitatively assess the impact of the 2022 FIA aerodynamic regulations on the competitiveness dynamics and prediction of Formula 1 races. The aerodynamic regulations were introduced to make racing closer and improve overtaking by drastically reducing the turbulent wake produced by each car. FIA's aimed to make races more exciting, with less reliance on clean air and increased competitiveness throughout the grid ("7 Key Rule Changes for the 2022 Season | Formula 1" 2022). Scholarly literature extensively covers the technical aspects of F1 car design from an engineering perspective (Belgiad 2024, 1) and includes statistical analyses of the impact of FIA regulations (Méndez 2023, 2) and forecasting of results using machine learning. However, little research applies ML as an inferential tool to identify and quantify the factors that drive team performance. The expected outputs of the project are a final research paper and the Python code used for the analysis. The methodology will involve using historical F1 race datasets for pre- and post-2022 periods and then developing a machine learning model to identify the underlying variables that most significantly distinguish between the two periods. The merit of this project lies in the application of a common data analytics and machine learning technique to a question that can be translated into sports analytics and regulatory science. Its broader impact is the potential to provide a more data-driven approach for governing bodies to understand how the characteristics of on-track competition change after the implementation of a major rule overhaul. In short, this project will move beyond simply predicting who will win a race and instead ask, "**What are the distinguishing characteristics of the races that occurred after a major regulatory change compared to those that occurred before?"**

Domain review

Overview of the Sports

Data has evolved to an integral aspect of Formula 1 operations, from strategic decisions and real-time analytics and predictive modeling for car development, making data science an essential foundation for team success in this highly competitive and technologically advanced sport (Ambler, 2024). ML in sports analytics follows a similar path as other industries, but has several differences due to its nature, especially during the training test split, given the varying factors that determine sports performance (Bunker & Thabtah, 2019). Common techniques such as K-means clustering, random forest, and linear regression are employed, but feature engineering varies significantly depending on the goal of the inferential analysis. A common characteristic of a winning team is often characterized by a combination of consistent performance and a fit driver pair (Nagle, 2022). It is also noteworthy that computational analysis of the rear wings revealed contradicting results with the intention of regulation, but real-world observation, which overtaking increased throughout the 2022 season (Méndez, 2023).

Overview of Machine Learning in Sports Analytics

The implementation of machine learning (ML) has emerged as a fundamental framework in contemporary sports analytics, offering a range of tools to model and comprehend the intricate dynamics of athletic competition. As highlighted by Bunker and Thabtah (2019), the application of ML in sports begins with a thorough understanding of the domain and data, followed by processes such as feature extraction, data processing, and model training. Given the nature of sports events, it is crucial to differentiate between match-related and external factors. Match features are those data points generated directly from the competition itself, whereas external features are independent of the match outcomes. Rory and Fadi also emphasize that the conventional 7:3 train-test split may not be appropriate for sports analytics. This is particularly relevant in contexts where past season performance may not strongly correlate with future outcomes, owing to significant yearly changes in team composition and lineups. For instance, a study by Chenjie on baseball prediction employed order-preserving train-test splits, such as rolling predictions by rounds or seasons, as these methods offer a more accurate representation of realistic forecasting (Cao, 2025).

The use of Machine Learning for Formula 1 does not deviate from conventional methods widely utilized in sports analytics. Sicoie’s (2022) and Tejada’s (2023) literature gives a comprehensive overview of machine learning methods and some of the most common variables used for race result and behavior prediction, providing a practical framework for applying machine learning to historical race data. The study approaches the analytical challenge as a regression and classification problem, employing models like Random Forest, Gradient Boosting Regressor, and Support Vector Machine to predict a driver's finishing position in each race. A key methodological contribution highlighted is the emphasis on feature engineering. Sicoie enriched the standard historical data with newly created variables, such as the driver's age at the time of the race and the conversion of finish times into a time difference from the race winner, which proved to be more significant predictors. K-fold cross-validation and randomized parameter search were used for model tuning, and model performance is primarily assessed by ranking correlation metrics such as the Spearman correlation coefficient, ROC curves, and AUC.

Some literatures employ a more advanced statistical method to break down the races beyond predictive modeling, which yields more information about the underlying drivers. A significant challenge in Formula 1 analytics is to deconstruct the effects of driver skill and constructor advantage. Van Kesteren and Bergkamp (2023) address this directly by developing a novel Bayesian model to analyze race results from the hybrid era (2014-2021). Rather than predicting a single outcome, their rank-ordered logit model estimates latent skill parameters for both drivers and constructors simultaneously, creating independent performance ratings. This approach allows for a direct quantification of each component's contribution to race outcomes. The study concludes that the constructor advantage accounts for the vast majority of the variance in race results, approximately 88%, while driver skill accounts for the remaining 12% (Erik-Jan van Kesteren & Bergkamp, 2023). This work is methodologically significant as it provides a statistical framework for separating individual talent from equipment advantage, which is a common question in many technology-dependent sports. In fact, some studies, such as from Nagle (2022), take a step further by clustering different groups of drivers. Using k-means clustering, the study revealed that the most successful teams tend to have drivers with consistent performance rather than sporadic wins, and benefit from pairing experienced drivers with younger, rising talent (Nagle, 2022).

A notable characteristic of the formula 1 data is that it is hierarchical, with multiple observations (laps) nested within groups (drivers or teams), which does not fit well with standard linear regressions. Linear Mixed Models (LMMs) and Generalized Linear Mixed Models (GLMMs) are the appropriate statistical tool to properly account for this nested structure (Casals et al, 2025). As highlighted by Casals et al. (2025), the use of GLMMs in sports analytics is often incomplete, making it difficult to assess the validity of results or to replicate findings. This lack of standardized, transparent reporting is a significant methodological gap in the field. Therefore, the present study not only employs a Linear Mixed Model to answer its research question but also to tackle the best-practice reporting guidelines identified by Casals et al.

Technical Aspect of the 2022 FIA Regulation

From a technical perspective, Méndez's (2023) work highlights the complexities of isolating and quantifying the impact of regulation. Using Computational Fluid Dynamics (CFD), the study simulated and compared the wake turbulence generated by the 2021 and 2022 rear wings. Counterintuitively, the simulation found that the 2022 rear wing design generated a higher level of turbulent kinetic energy, failing its intended purpose. However, Méndez notes that this finding contradicts empirical evidence, citing real-world data that shows a significant 30% increase in on-track overtakes during the 2022 season (Méndez, 2022). The discrepancy is attributed to the study's simplifying assumptions, primarily that the simulation modeled the rear wing in isolation, failing to capture its interaction with the car's complete aerodynamic profile. This conclusion is highly relevant as it underscores the limitations of simulation-based analysis and validates the necessity of an empirical approach. This creates room for inferential models that can be used to validate the real-world impact of a regulatory overhaul like the 2022 rule change.

Method

Data Collection & Aggregation

The data for this study will be sourced from the FastF1 Python library (Schaefer, 2025). FastF1 is a publicly available API that provides access to F1 data, which is permissible for use in academic and non-commercial research. The data collection process will involve writing a Python script to iterate through each Grand Prix from the beginning of 2020 through 2025. For each race event, we will load the session data using fastf1.get_session() and use the library's caching feature to store the data in a local folder to export into CSV format to reduce load times on subsequent runs. Data from the results and laps DataFrames will be extracted and combined to form the primary dataset. Results will be aggregated for race-level ML, where each row is dedicated to a driver per race, while lap data has each row dedicated to a single lap per driver in a race. There will be approximately 20 - 24 races per season over 5 seasons. The 2020 season is selected as the starting point as it represents a mature phase of the 2017 aerodynamic regulations, providing a stable and well-established baseline for comparison. Races that were red-flagged and not completed to at least 90% of their original distance, as it would not yield proper results for a fair comparison, and Sprint races, as their shorter format could alter race dynamics. These criteria are established to control for major confounding variables.

Analysis

Random Forest Classification

A Random Forest Classifier was employed to predict whether a race lap occurred pre- or post-regulation. The model was trained on the aggregated_laps.csv dataset using a specific subset of features: Driver, LapTime, SpeedI1, SpeedI2, SpeedFL, SpeedST, AirTemp, TrackTemp, WindSpeed, Humidity, and Compound. To prepare the data for modeling, categorical features (Driver, Compound) were transformed into a numerical format using one-hot encoding. The dataset was then split for training and test sets with 80% of the data allocated for training the model and the remaining 20% for validation. Stratification was used during this split to ensure that the proportion of pre- and post-regulation laps was maintained in both the training and testing samples to prevent class imbalance bias.

Generalized Linear Mixed Models

To answer the research question, a Generalized Linear Mixed Model (LMM) was implemented. This model was chosen over a standard linear regression given the nature of the formula 1 data which contains multiple, non-independent observations (laps) from the same drivers. The model will allow us to account for this hierarchical data structure where laps are nested within drivers by treating Driver as a random effect.

The model was constructed to predict the dependent variable, LapTime, using pre/post regulation as the primary fixed effect of interest. To isolate the impact of the regulations and avoid confounding variables, other factors were included as fixed-effect covariates: Compound (categorical, to control for tire differences), Stint (numerical, to control for fuel load and tire wear), and the atmospheric conditions such as air temperature, track temperature, and humidity.

The specific model formula is given by the following equation:

LapTime ~ PostRegulation + Compound + Stint + AirTemp + TrackTemp + Humidity + (1 | Driver)

This equation models lap time as a function of the fixed effects while fitting a random intercept per each driver. This intercept represents a driver's baseline pace, factoring out driver skill and allowing the model to better isolate the true effect of the regulation change and other covariates. Model parameters were estimated using Restricted Maximum Likelihood (REML).

Results

Generalized Linear Models

The Linear Mixed Model successfully ran, but the model failed to converge. This indicates the model failed to find an optimal solution, and therefore the resulting coefficients and p-values may not be reliable.

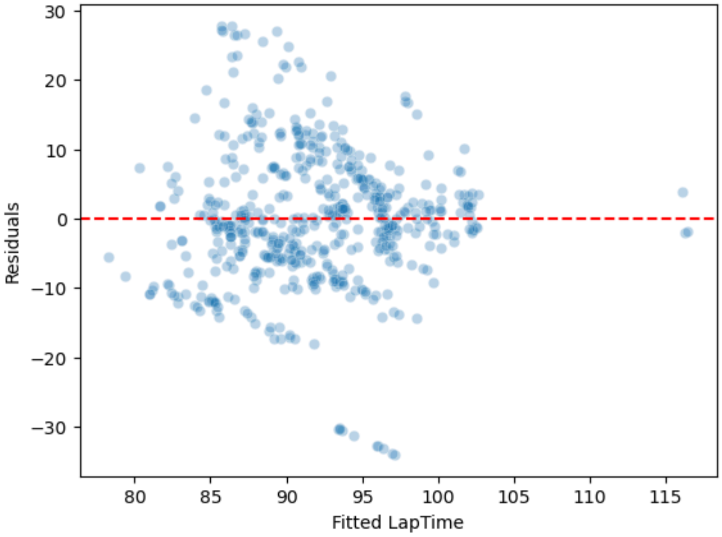


Figure 1. Distribution of of the residuals from the model

Model's Root Mean Square Error (RMSE), was ~9.5 seconds, indicating a serious error in predicting Lap Time. The primary target, PostRegulation, had a positive coefficient of +1.610 (p = 0.067). This means that holding all other factors constant, laps in the post-regulation era were 1.61 seconds slower than those in the pre-regulation era. However, this result is statistically insignificant given the p-value. Several covariates were found to be highly significant predictors of LapTime, such as soft tire compound which suggests soft tires were, on average, 5.89 seconds faster than the baseline hard tire. Air temperature also had a coefficient of +0.783, implying that for every 1-degree increase in air temperature, lap times increased by 0.783 seconds. Track temperature also had a coefficient of -0.665, implying that for every 1-degree increase in track temperature, lap times decreased by 0.665 seconds.

Random Forest Model

Random Forest model performed with high precision. On the test set, the model achieved an overall accuracy of 96.4%.

	Precision	Recall	F1-score	Support
Pre-regulation	0.96	0.98	0.97	66
Post-regulation	0.98	0.93	0.95	45

The confusion matrix also validated this performance, showing only three misclassifications out of 111 test samples:

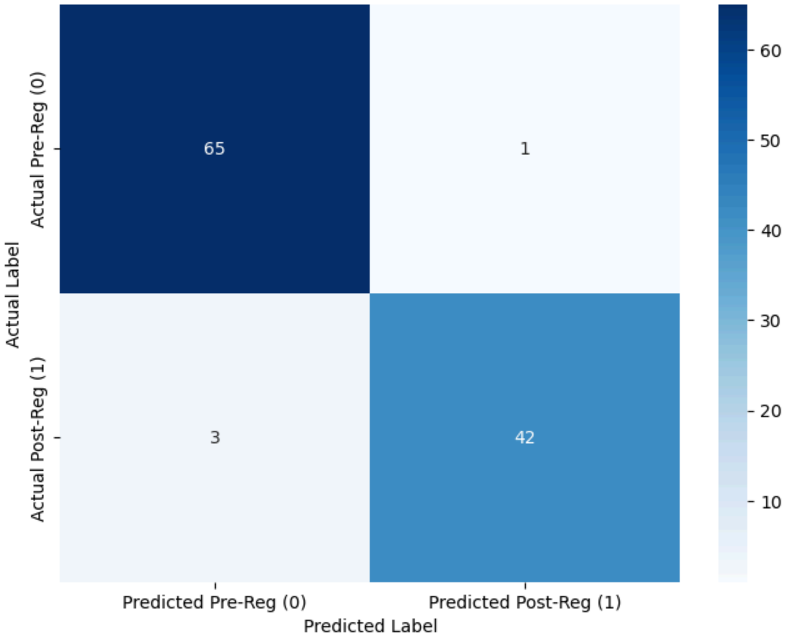


Figure 2: Confusion matrix of the Random Forest Model

The significance level of the of the covariates were as following:

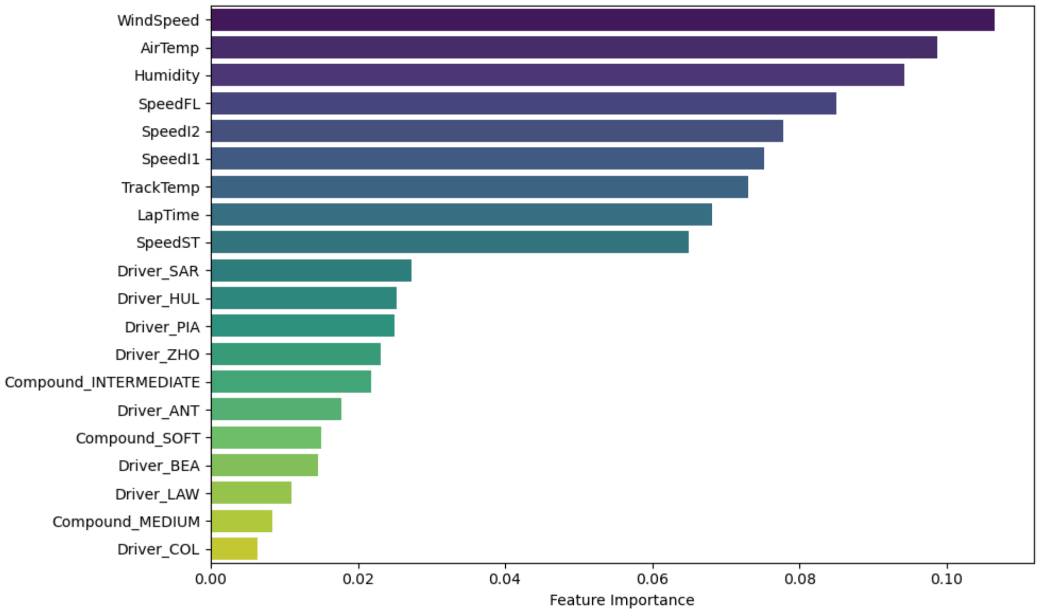


Figure 3: Visualization of the importance of the features. Weather attributes and average track speeds accounted for the most significant factors.

All the individual Drivers and Compound features had very low importance. This doesn't mean dismiss a Driver being insignificant, but rather implies that no single driver was a key predictor.

Conclusion

Discussion

The primary finding from this analysis is that the model failed to converge, rendering all results unreliable. This non-convergence is the most important outcome and must be addressed before any conclusions can be drawn about the research question. The likely reason for the outcome is the unknown compound category. Its extremely large coefficient and high significance suggest these laps are outliers that are fundamentally different from normal racing laps. If the convergence warning and p-value is to be ignored, the model would suggest that the post-regulation cars are 1.61 seconds slower than their pre-regulation counterparts, a finding that contradicts the goal behind regulation, meaning it would imply the regulations have slowed the cars down.

While the Linear Mixed Model failed to yield significant results, the Random Forest classification model achieved a high accuracy of 96.4%, confirming that a distinct difference exists between the pre- and post-regulation data. A feature importance analysis of this model revealed that environmental factors were, by a large margin, the most significant predictor followed by average speed throughout the laps. This finding does validate the central premise that the regulation change had a measurable and consistent impact on on-track performance to some extent. The results strongly indicate that the selected features, combining driver, compound, lap performance, and atmospheric conditions, are highly predictive of the regulation era. The high precision and recall for both classes are particularly significant. The model is not only accurate overall but is also reliable in identifying both 'Pre' and 'Post' regulation instances, avoiding a class bias This implies that the regulation changes had a consistent and discernible impact on the combination of features measured. The low number of false negatives (2) and false positives (1) in the test set suggests the model is robust.

Future work

In regards to the mixed model, we can first address the issue of non-convergence. The model must be re-run after filtering out all laps where Compound is 'UNKNOWN'. These data points are affecting the analysis and are the most likely cause of the convergence failure. Only after a converged model is achieved can we have a solid answer about the distinguishing characteristics of the regulation change. At the same time, we could possibly include additional factors into account and change the equations.

For the Random Forest Model, To better understand the distinguishing characteristics other than the lap time itself, the next step would be to re-run the classification model after removing several environment factors from the feature set. This would force the model to identify which combination such as tire compounds, or driver-independent speed trap data are the next-best predictors of the regulation era. Additionally, we could attempt to perform hyperparameter tuning by manipulating the n_estimators to see whether any improvements can take place along with additional model validation.

References

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

Schaefer, P. (2025). FastF1, version 3.6.1, MacOS GitHub. Retrieved from https://github.com/theOehrly/Fast-F1

Python Software Foundation. (2025). Python Language Reference, version 3.13.7 [Computer software]. Retrieved from https://docs.python.org/3/reference/index.html

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). DOI: 10.1038/s41586-020-2649-2.

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Data structures for statistical computing in Python, McKinney, Proceedings of the 9th Python in Science Conference, Volume 445, 2010.

F1 and the weather. (2011, September 22). RMetS. https://www.rmets.org/metmatters/f1-and-weather

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

Casals, M., Fernández, D., Zumeta-Olaskoaga, L., Sánchez, A., & Zuccolotto, P. (2025). Reporting of generalized linear mixed models (GLMM) in sports sciences: A scoping review. Journal of Sports Analytics. https://doi.org/10.1177_22150218251384557

Courtnee Catnott. (2025, August 28). Formula 1 2025 season – a half year review | Formula One World Championship Limited. Formula1.com. https://corp.formula1.com/formula-1-2025-season-a-half-year-review/

Ambler, W. (2024, February 13). How Data Analysis Transforms F1 Race Performance | Catapult. Catapult. https://www.catapult.com/blog/f1-data-analysis-transforming-performance

Bunker, R. P., & Fadi Thabtah. (2017). A machine learning framework for sport result prediction. Applied Computing and Informatics, 15(1), 27–33. https://doi.org/10.1016/j.aci.2017.09.005

Tejada, G. (2023, August 21). Applying Machine Learning to Forecast Formula 1 Race Outcomes. Aalto.fi. https://aaltodoc.aalto.fi/items/5848c100-478d-45dd-b2e8-5caf3a3114fb

Nagle, D. (2022). Racing Your Rival: Cluster Analysis of Formula 1 Drivers. SURFACE at Syracuse University. https://surface.syr.edu/honors_capstone/1607/

van Kesteren, E. J., & Bergkamp, T. (2023). Bayesian analysis of Formula One race results: disentangling driver skill and constructor advantage. Journal of quantitative analysis in sports, 19(4), 273–293. https://doi.org/10.1515/jqas-2022-0021

Cao, C. (2025). Sports Data Mining Technology Used in Basketball Outcome Prediction. ARROW@TU Dublin. https://arrow.tudublin.ie/scschcomdis/39/

Casals, M., Fernández, D., Zumeta-Olaskoaga, L., Sánchez, A., & Zuccolotto, P. (2025). Reporting of generalized linear mixed models (GLMM) in sports sciences: A scoping review. Journal of Sports Analytics, 11, 22150218251384557.

Appendix

Appendix A: List of variables from the dataset from FastF1 API documentation:

- Time (pandas.Timedelta): Session time when the lap time was set (end of lap)
- Driver (str): Three letter driver identifier
- DriverNumber (str): Driver number
- LapTime (pandas.Timedelta): Recorded lap time. To see if a lap time was deleted, check the Deleted column.
- LapNumber (float): Recorded lap number
- Stint (float): Stint number
- PitOutTime (pandas.Timedelta): Session time when car exited the pit
- PitInTime (pandas.Timedelta): Session time when car entered the pit
- Sector1Time (pandas.Timedelta): Sector 1 recorded time
- Sector2Time (pandas.Timedelta): Sector 2 recorded time
- Sector3Time (pandas.Timedelta): Sector 3 recorded time
- Sector1SessionTime (pandas.Timedelta): Session time when the Sector 1 time was set
- Sector2SessionTime (pandas.Timedelta): Session time when the Sector 2 time was set
- Sector3SessionTime (pandas.Timedelta): Session time when the Sector 3 time was set
- SpeedI1 (float): Speedtrap sector 1 [km/h]
- SpeedI2 (float): Speedtrap sector 2 [km/h]
- SpeedFL (float): Speedtrap at finish line [km/h]
- SpeedST (float): Speedtrap on longest straight (Not sure) [km/h]
- IsPersonalBest (bool): Flag that indicates whether this lap is the official personal best lap of a driver. If any lap of a driver is quicker than their respective personal best lap, this means that the quicker lap is invalid and not counted. For example, this can happen if the track limits were exceeded.
- Compound (str): Tyres event specific compound name: SOFT, MEDIUM, HARD, INTERMEDIATE, WET, TEST_UNKNOWN, UNKNOWN. The actual underlying compounds C1 to C5 are not differentiated. TEST_UNKNOWN compounds can appear in the data during pre-season testing and in-season Pirelli tyre tests.
- TyreLife (float): Laps driven on this tire (includes laps in other sessions for used sets of tires)
- FreshTyre (bool): Tyre had TyreLife=0 at stint start, i.e. was a new tire
- Team (str): Team name
- LapStartTime (pandas.Timedelta): Session time at the start of the lap
- LapStartDate (pandas.Timestamp): Timestamp at the start of the lap
- TrackStatus (str): A string that contains track status numbers for all track status that occurred during this lap. The meaning of the track status numbers is explained in fastf1.api.track_status_data(). For filtering laps by track status, you may want to use Laps.pick_track_status().

- Position (float): Position of the driver at the end of each lap. This value is NaN for FP1, FP2, FP3, Sprint Shootout, and Qualifying as well as for crash laps.
- Deleted (Optional[bool]): Indicates that a lap was deleted by the stewards, for example because of a track limits violation. This data is only available when race control messages are loaded.
- IsAccurate (bool): Indicates that the lap start and end time are synced correctly with other laps. Do not confuse this with the accuracy of the lap time or sector times. They are always considered to be accurate if they exist! If this value is True, the lap has passed as basic accuracy check for timing data. This does not guarantee accuracy but laps marked as inaccurate need to be handled with caution. They might contain errors which can not be spotted easily. Laps need to satisfy the following criteria to be marked as accurate:

Appendix B: Additional Figures

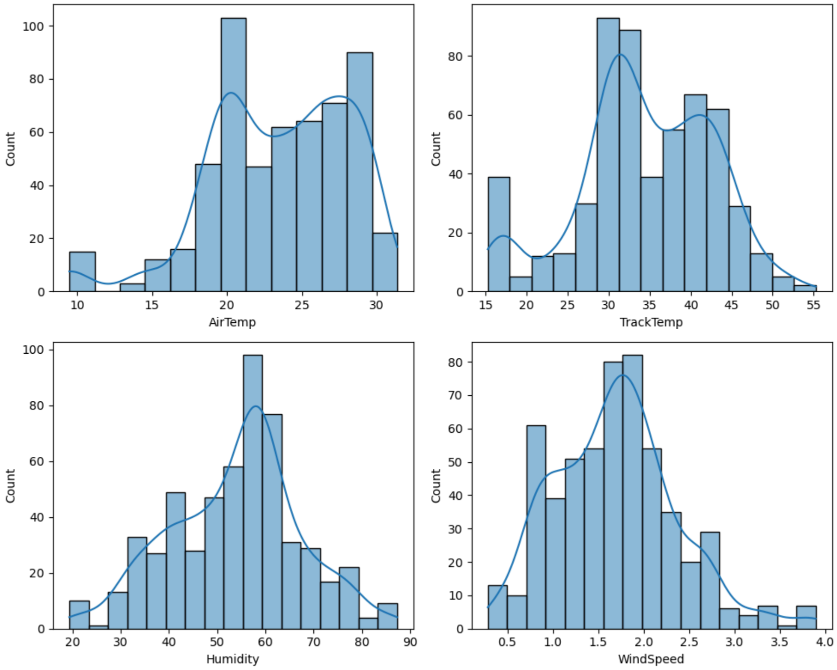


Figure 1. Distribution of the weather data

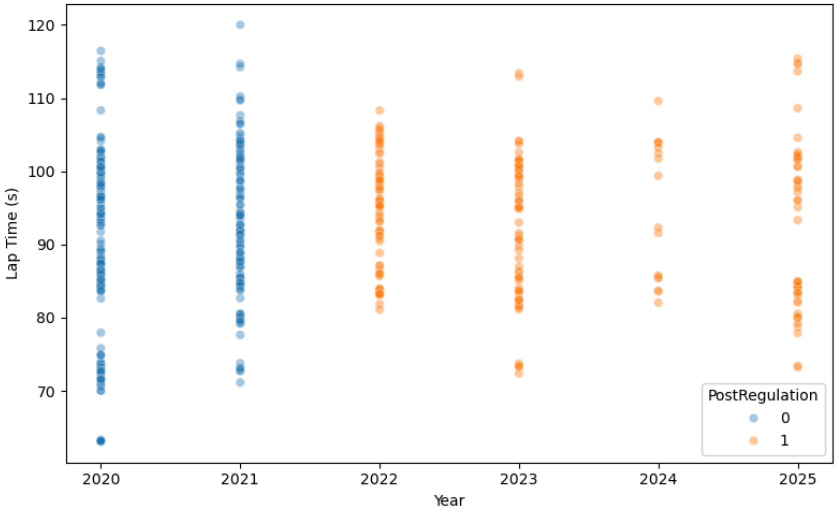


Figure 2. Scatterplot of lap times from 2020-2025

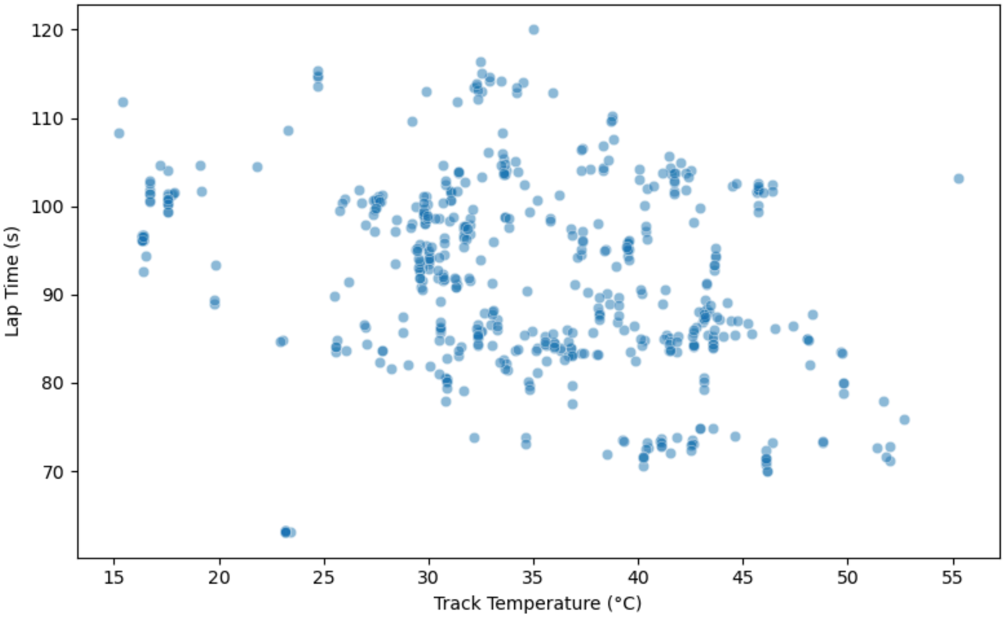
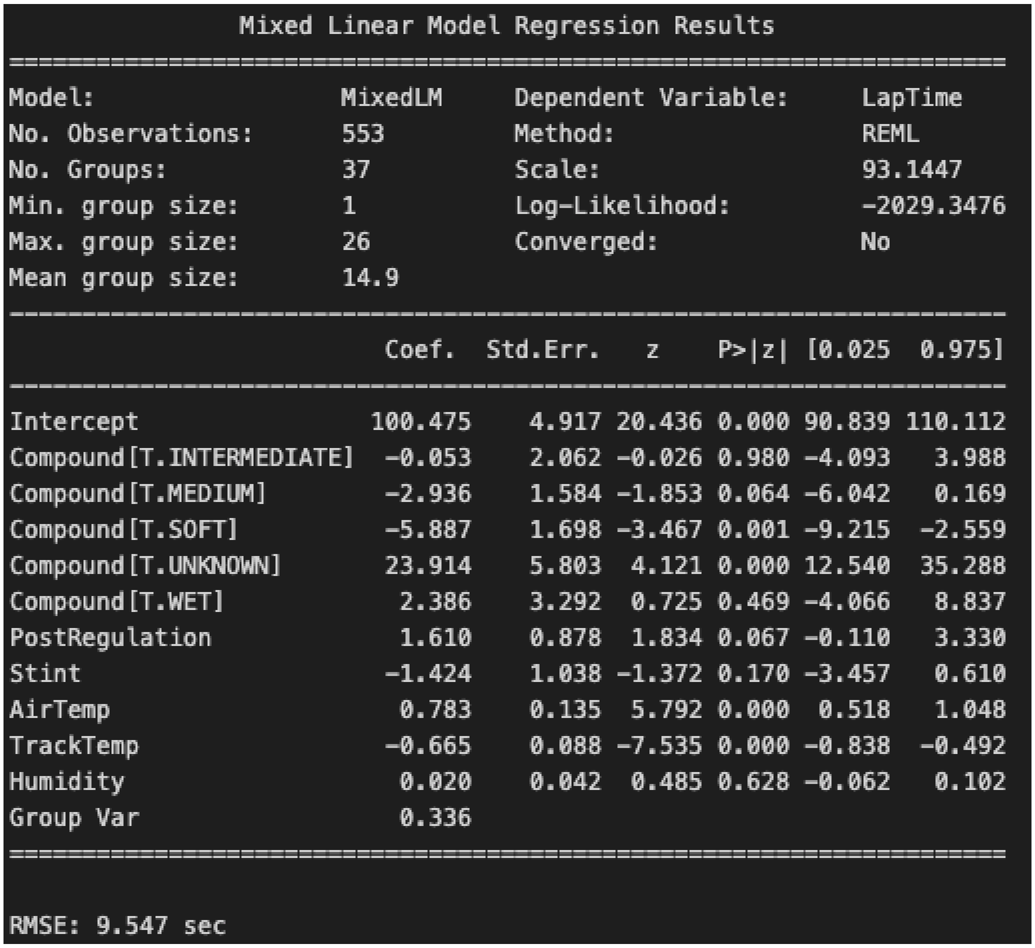


Figure 3. Scatterplot of lap time vs Track Temperature



```
                Mixed Linear Model Regression Results
=====================================================================
Model:                 MixedLM     Dependent Variable:     LapTime
No. Observations:      553         Method:                 REML
No. Groups:            37          Scale:                  93.1447
Min. group size:       1           Log-Likelihood:         -2029.3476
Max. group size:       26          Converged:              No
Mean group size:       14.9
---------------------------------------------------------------------
                          Coef.  Std.Err.   z    P>|z| [0.025  0.975]
---------------------------------------------------------------------
Intercept                100.475    4.917 20.436 0.000 90.839 110.112
Compound[T.INTERMEDIATE]  -0.053    2.062 -0.026 0.980 -4.093   3.988
Compound[T.MEDIUM]        -2.936    1.584 -1.853 0.064 -6.042   0.169
Compound[T.SOFT]          -5.887    1.698 -3.467 0.001 -9.215  -2.559
Compound[T.UNKNOWN]       23.914    5.803  4.121 0.000 12.540  35.288
Compound[T.WET]            2.386    3.292  0.725 0.469 -4.066   8.837
PostRegulation             1.610    0.878  1.834 0.067 -0.110   3.330
Stint                     -1.424    1.038 -1.372 0.170 -3.457   0.610
AirTemp                    0.783    0.135  5.792 0.000  0.518   1.048
TrackTemp                 -0.665    0.088 -7.535 0.000 -0.838  -0.492
Humidity                   0.020    0.042  0.485 0.628 -0.062   0.102
Group Var                  0.336
=====================================================================


RMSE: 9.547 sec
```

Figure 3. Snapshot of the results from the Mixed Linear regression.