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

Denison University

Seminar in Data Analytics

Myunggong Seo

November 2025

Abstract:

This report quantitatively evaluates the impact of the 2022 FIA aerodynamic regulations on Formula 1 race dynamics using machine learning and statistical modeling. By leveraging comprehensive race data from 2019 to 2025 with FastF1 API, the project uses Random Forest Classifiers and Linear Mixed Models to detect distinguishable features and performance shifts between the pre- and post-regulation races. While the linear mixed model faced convergence issues that limited the reliability of results (log likelihood ~ 2029), the Random Forest model achieved high accuracy (~95.5%) and confirmed the existence of significant distinctions of characteristics before and after the overhaul, especially on variables related to the aerodynamic profile of the vehicles. The analysis highlights that track speeds and environmental factors, rather than driver or tires alone, played the largest roles in differentiating race characteristics. These results provide a groundwork for a more quantitative approach to evaluating regulatory changes in motorsport and establish a framework for future research with similar subjects within sports analytics.

Introduction

Formula 1 is recognized as one of the largest and most popular motorsports in the world in terms of 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). FIA routinely overhauls regulations to ensure the competitiveness of the races. The latest change took place in 2022, when the FIA introduced aerodynamic regulations and 18-inch wheels to make races more exciting by reducing the amount of wake from turbulence with less reliance on clean air, which would increase the competitiveness throughout the grid ("7 Key Rule Changes for the 2022 Season | Formula 1" 2022). Existing scholarly articles extensively cover the technical aspects of F1 car design from an engineering perspective (Belgiad 2024, 1) and include statistical analyses of the impact of FIA regulations (Méndez 2023, 2) and forecasting of results using machine learning. However, little work has been done applying ML as a tool to identify and quantify the factors that led to changes in results from regulations. To mitigate this issue, this project will use historical F1 race datasets for pre- and post-2022 periods and develop 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 competition on track will change after the implementation of a major rule overhaul. In short, this project will move beyond simply predicting the likely outcome of 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 Science has evolved to be an inseparable part of Formula 1 operations, such as making strategic decisions, real-time analytics, and predictive modeling for car development, making it an essential foundation for team success in this highly competitive sport (Ambler, 2024). ML in sports analytics has similar intents as other industries, but has several differences due to its nature, given the varying factors that determine sports performance, especially during the training test split stage (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 highlighted by a combination of consistent performance and a fit driver pair (Nagle, 2022). It is also worth noting that computational analysis of the rear wings revealed contradicting results with the intention of regulation, where real-world observation of overtaking actually increased throughout the 2022 season (Méndez, 2023).

Overview of Machine Learning in Sports Analytics

The use of machine learning (ML) has emerged become a central framework in sports analytics, offering a range of tools to model and comprehend the complex dynamics of sports competitions with tremendous variations. 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 refer to the 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, due 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 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 the current race data. The study approaches the analytical challenge as a regression and classification problem, using models such as Random Forest, Gradient Boosting Regressor, and Support Vector Machine to predict a driver's finishing position in each race. A key contribution highlighted is the emphasis on feature engineering. Sicoie enhanced the standard race 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 yielded 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 Bayesian model to analyze race results from the hybrid era (2014-2021). Rather than predicting a single outcome, they used a rank-ordered logit model to estimate skill parameters for both drivers and constructors simultaneously and created independent performance ratings. This approach directly quantifies the extent 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 (~ 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 issue in many sports with heavy reliance on equipment. In fact, some studies, such as from Nagle

(2022), take a step further by clustering different groups of drivers based on driving styles. The study used K-means clustering to uncover that the most successful teams tend to have drivers with consistent performance rather than sporadic wins, and paired 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) 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 Mixed Models in sports analytics is often incomplete, making it difficult to assess the validity of results or to replicate the method. This lack of standardized reporting is a significant methodological gap in the field. Therefore, the present study not only uses a Linear Mixed Model to answer its research question but also tackles the practice of 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 on the race dynamic. Using Computational Fluid Dynamics (CFD), the study simulated and compared the wake turbulence generated by the 2021 and 2022 rear wings. Interestingly, 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, 2023). The discrepancy may be rooted in the study's assumption, where it 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 underscores the necessity of an empirical approach. This literature also creates room for statistical and machine learning models that can be used to validate the real-world impact of a regulatory overhaul.

Exploratory Data Analysis

Before the Analysis, EDA was performed to ensure multicollinearity and outliers were not integrated into the model. EDA focused on the numeric features and identified that the general trend of LapTime stayed mostly within the interval between 75 to 150 seconds across 7 seasons. Most of the outliers were concentrated in 2021 and 2025. Interestingly, it can be seen that the general trend of Lap Time marginally increases before and after 2022.

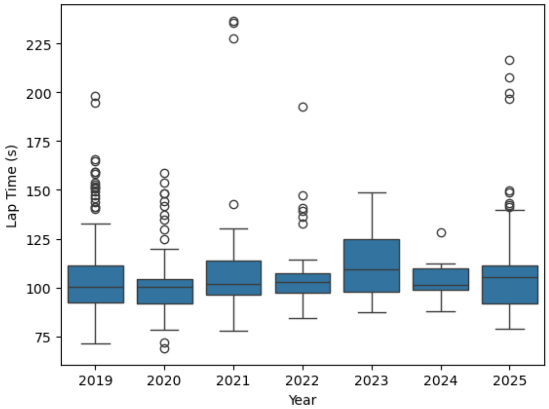


Figure 1. Boxplot of the range of Lap Time across 5 seasons

Checking the distribution of the frequency across the board also once again revealed that, except for a few outliers, Lap Time follows a normal distribution, as can be observed below. This ensured that our data would be fit for a Linear Mixed Model. Additionally, important weather covariates such as humidity and Air Temperature also followed a roughly normal distribution.

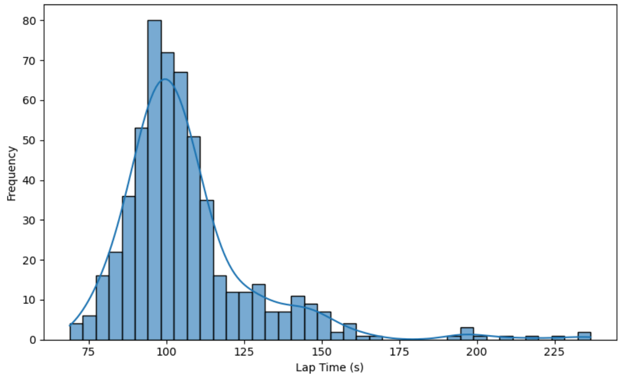


Figure 2. Distribution of the residuals from the model

Lastly, the correlation between all numerical columns revealed that there was no serious multicollinearity between covariates that were thought to be highly correlated, such as the speed between different sectors and the Lap Time. Most of the correlation coefficients remained below the threshold of 0.5.

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 code to iterate through each race from the beginning of 2019 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 lap per driver in a race. There will be approximately 20 - 24 races per season over 5 seasons. The 2019 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 were removed as they would fail to yield proper results for a fair comparison. Sprint races were similarly omitted 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 lap 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 through 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 lapses was maintained in both the training and testing samples to prevent class imbalance.

Generalized Linear Mixed Models

Linear Mixed Model (LMM) 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- and 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 for 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 interpretation may not be reliable. It is worth noting that VIF scores for all of the covariates were below 5, which is considered a threshold.

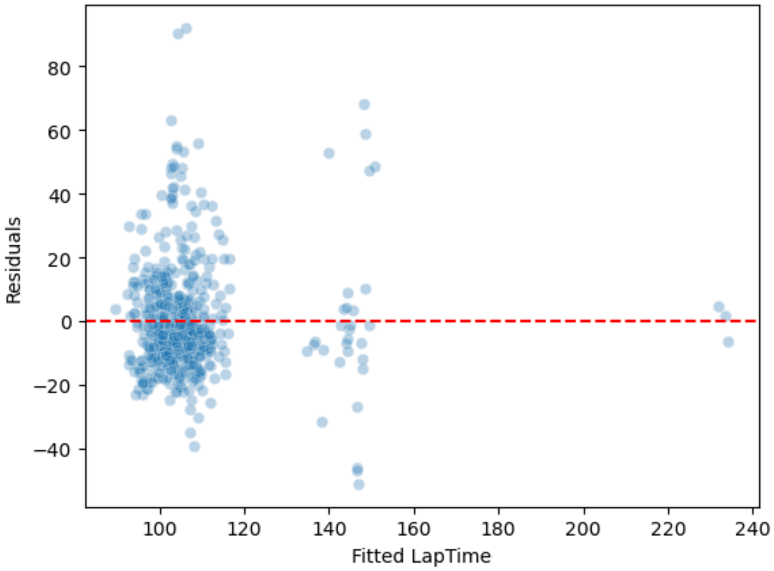


Figure 3. Distribution of the residuals from the model

The model's Root Mean Square Error (RMSE) was ~17 seconds, indicating a serious error in predicting Lap Time. The primary target, PostRegulation, had a positive coefficient of +5.414 (p = 0.03). This means that holding all other factors constant, laps in the post-regulation era were 5.414 seconds slower than those in the pre-regulation era, given the p-value. Track temperature also had a coefficient of -4.217, implying that for every 1-degree increase in track temperature, lap times decreased by 4.217 seconds. However, even some of the significant results with low p-value, such as the intermediate tire compound, had a coefficient of +32.184, which is significantly large compared to other variables. In fact, the coefficient of the intermediate compound was 128, which is an unlikely result.

Random Forest Model

Random Forest model performed with high precision. On the test set, the model achieved an overall accuracy of 9.550% with specific scores in the table below.

Table 1: Model Performance of the Random Forest Classification

	Precision	Recall	F1-score	Support
Pre-regulation	0.95	0.99	0.97	78
Post-regulation	0.97	0.88	0.92	33

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

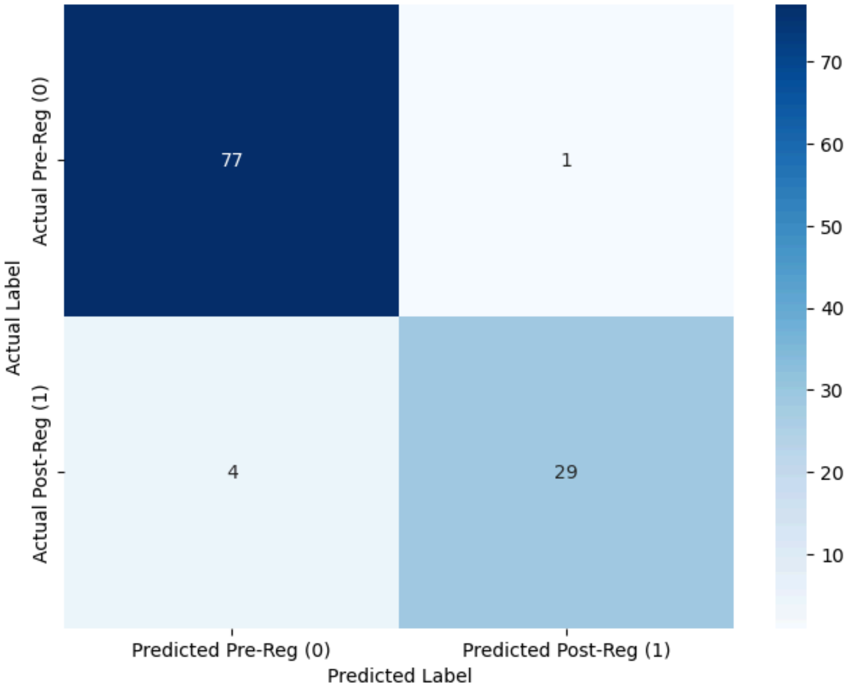


Figure 4: Confusion matrix of the Random Forest Model. The model performed well with a high number of TN, TP

The significance level of the covariates was as follows:

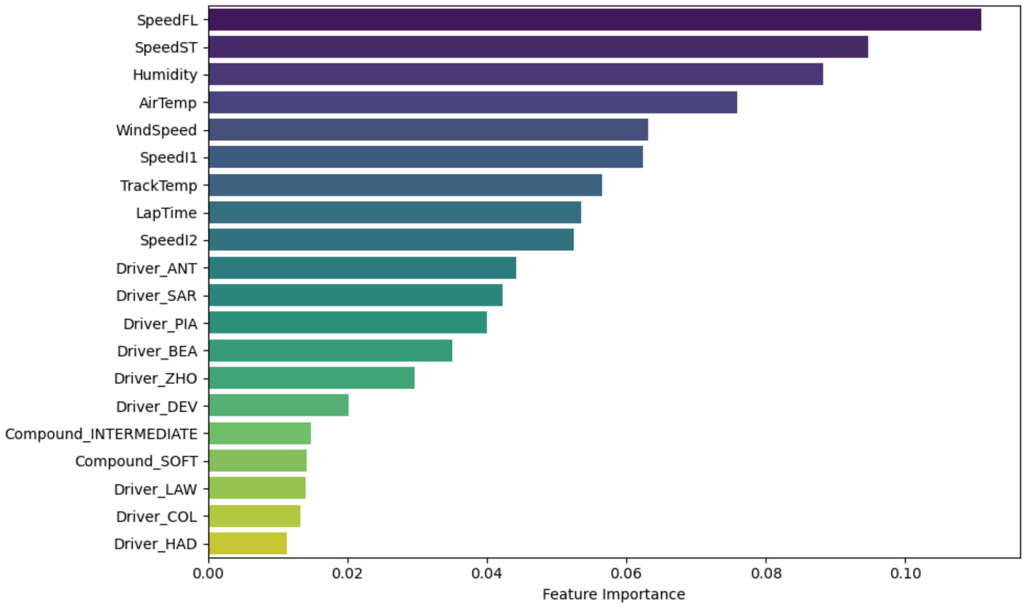


Figure 5: Visualization of the importance of the features. Weather attributes and average track speeds accounted for the most significant factors.

All the individual Drivers and Compound features had very low importance. This doesn't mean to dismiss a Driver being insignificant, but rather implies that no single driver was a key predictor. On the other hand, covariates directly related to the speed of the vehicles and the environment (humidity, weather, etc.) accounted for the most significant drivers behind whether a race was pre- or post-regulation.

Conclusion

Discussion

The primary finding from the Linear Mixed Model analysis is that the model failed to converge, rendering most results unreliable. This failure 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 are 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 95.550%, confirming that there exists a marked difference between the pre- and post-regulation races. This 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 and false positives in the test set suggests the model is robust. Visualization of the importance of the features revealed that race speeds and environmental factors were, by a large margin, the most significant predictors, followed by some driver factors and tire compounds, relatively lower. This finding does validate the central premise that the regulation change had a measurable 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. At the same time, the importance of speed as a feature suggests that regulation may have altered the aerodynamic and mechanical profile of the cars (Ménez, 2023).

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 compount is 'UNKNOWN'. This feature is affecting the analysis and is 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 include additional factors into account and change the equation. Building a new relevant feature, such as “Number of Overtakes” or “Changes in Max/Min Velocity” could also help us obtain meaningful results, as witnessed by previous work by Sicoie and Tejada. Another mitigation is to normalize numeric columns. For instance, humidity, temperature, and Lap Time all have different scales when it comes to being recorded. Therefore, minimizing the discrepancy between numerical variables may help us obtain a more sensible result.

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 environmental factors from the feature set. This would force the model to identify which combination, such as tire compounds or driver-independent speed trap data, is the next-best predictor of whether a race was held prior to or after regulation. 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. One possible change to validate the significance of the results is to conduct the analysis on a single selected track (e.g., the Spanish Grand Prix) across seven seasons and expand the analysis to a lap-level instead of a race-level. This approach would help control for environmental factors and allow for a deeper investigation, as the model could be run on a more granular scale that takes into account lap-by-lap variations.

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.

Méndez, Luis A. "Quantifying the Impact of the 2022 Formula One Technical Regulations on Wake Turbulence: A Numerical Analysis." (2023).

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 from EDA

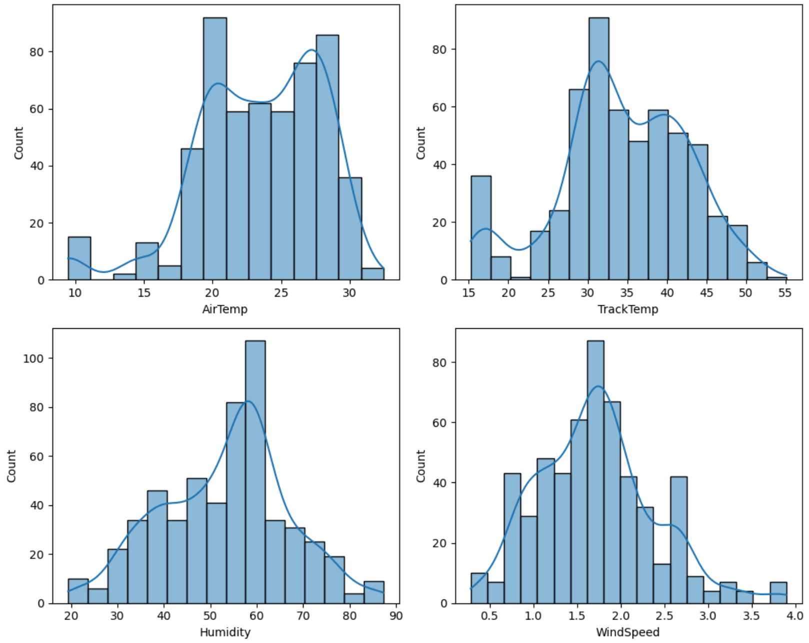


Figure 1. Visualization of the distribution of the weather features

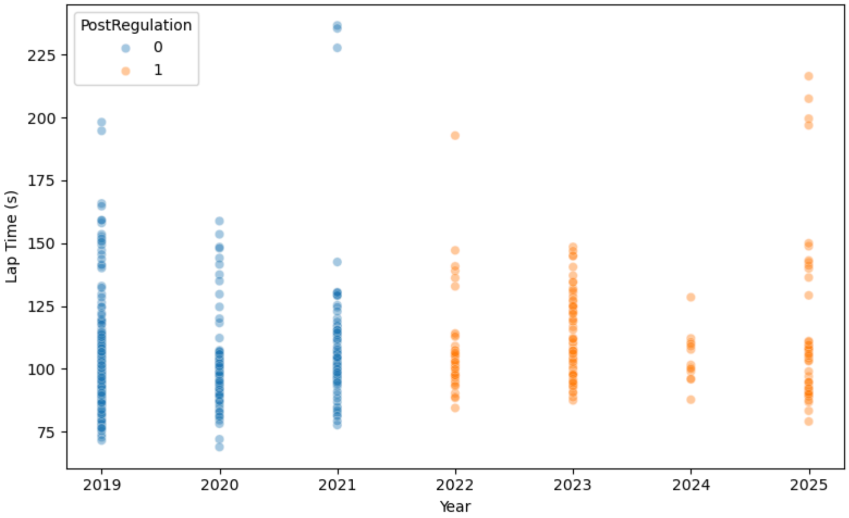


Figure 2. Scatterplot of lap times from 2020-2025

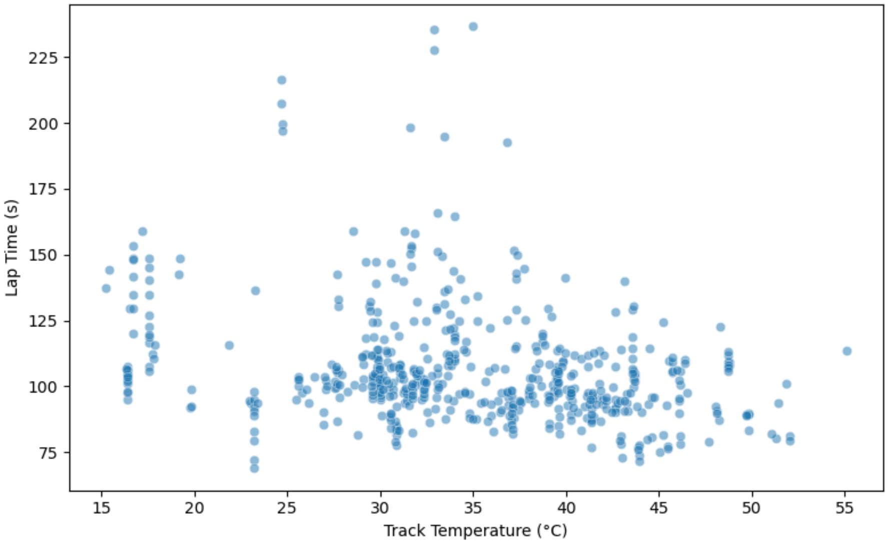
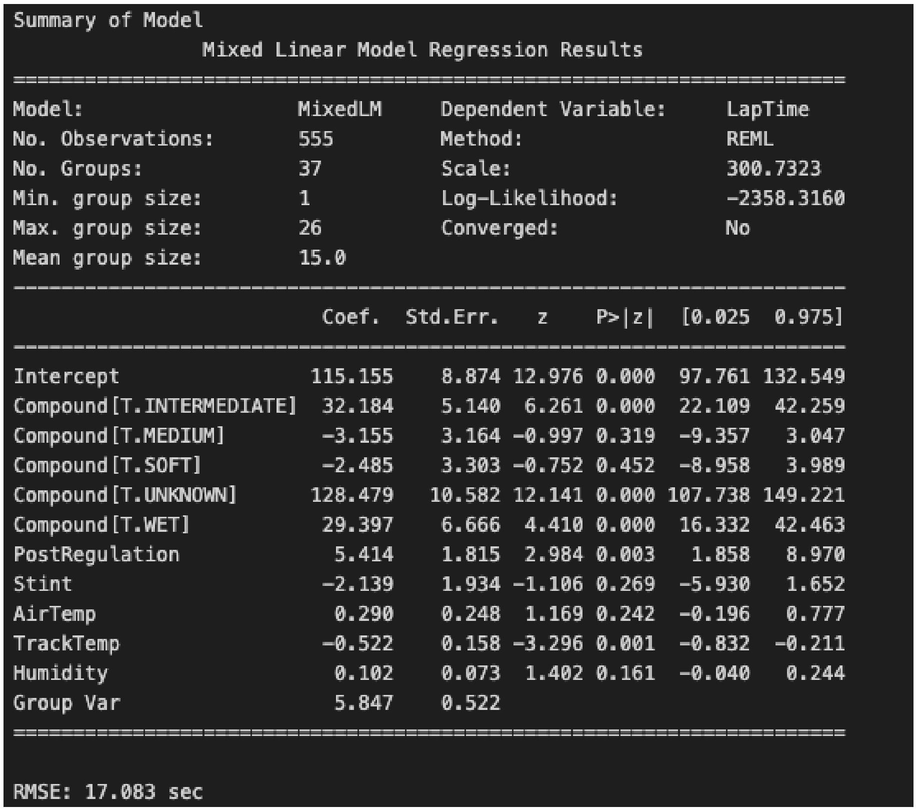


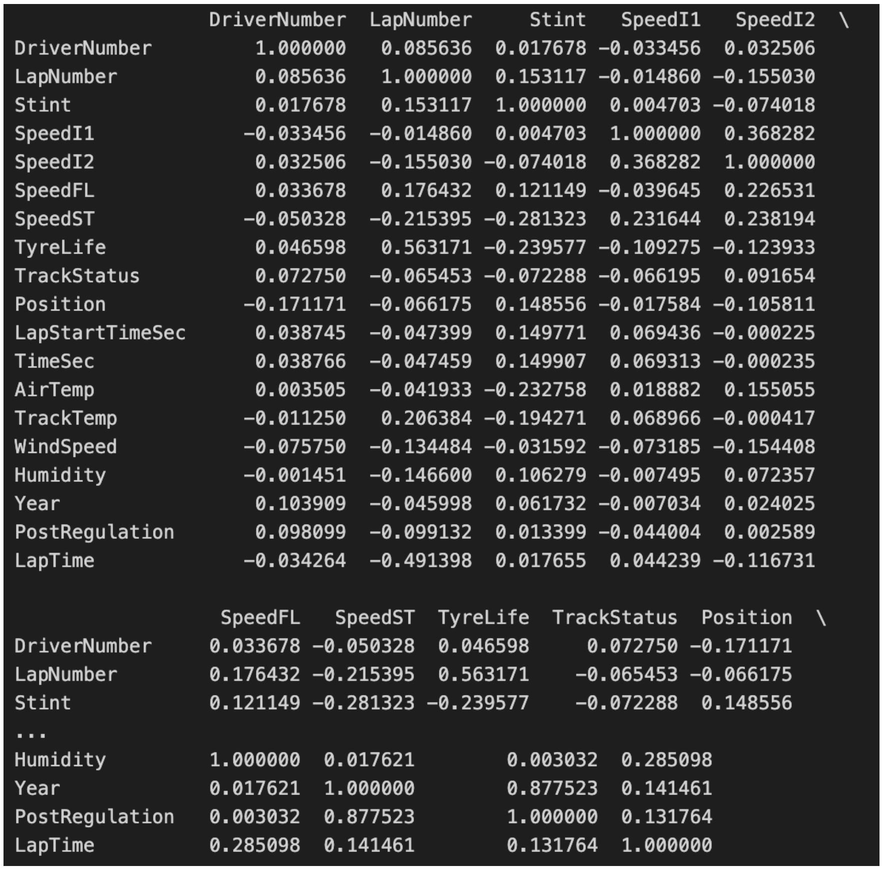
Figure 3. Scatterplot of lap time vs Track Temperature



```
Summary of Model
               Mixed Linear Model Regression Results
===================================================================
Model:                   MixedLM     Dependent Variable:     LapTime
No. Observations:        555         Method:                 REML
No. Groups:              37          Scale:                  300.7323
Min. group size:         1           Log-Likelihood:         -2358.3160
Max. group size:         26          Converged:              No
Mean group size:         15.0
-------------------------------------------------------------------
                           Coef.  Std.Err.   z    P>|z|  [0.025  0.975]
-------------------------------------------------------------------
Intercept                 115.155    8.874 12.976 0.000  97.761 132.549
Compound[T.INTERMEDIATE]   32.184    5.140  6.261 0.000  22.109  42.259
Compound[T.MEDIUM]         -3.155    3.164 -0.997 0.319  -9.357   3.047
Compound[T.SOFT]           -2.485    3.303 -0.752 0.452  -8.958   3.989
Compound[T.UNKNOWN]       128.479   10.582 12.141 0.000 107.738 149.221
Compound[T.WET]            29.397    6.666  4.410 0.000  16.332  42.463
PostRegulation              5.414    1.815  2.984 0.003   1.858   8.970
Stint                      -2.139    1.934 -1.106 0.269  -5.930   1.652
AirTemp                     0.290    0.248  1.169 0.242  -0.196   0.777
TrackTemp                  -0.522    0.158 -3.296 0.001  -0.832  -0.211
Humidity                    0.102    0.073  1.402 0.161  -0.040   0.244
Group Var                   5.847    0.522
===================================================================

RMSE: 17.083 sec
```

Figure 4. Snapshot of the results from the Mixed Linear regression



```
                 DriverNumber  LapNumber     Stint   SpeedI1   SpeedI2  \
DriverNumber         1.000000   0.085636  0.017678 -0.033456  0.032506
LapNumber            0.085636   1.000000  0.153117 -0.014860 -0.155030
Stint                0.017678   0.153117  1.000000  0.004703 -0.074018
SpeedI1             -0.033456  -0.014860  0.004703  1.000000  0.368282
SpeedI2              0.032506  -0.155030 -0.074018  0.368282  1.000000
SpeedFL              0.033678   0.176432  0.121149 -0.039645  0.226531
SpeedST             -0.050328  -0.215395 -0.281323  0.231644  0.238194
TyreLife             0.046598   0.563171 -0.239577 -0.109275 -0.123933
TrackStatus          0.072750  -0.065453 -0.072288 -0.066195  0.091654
Position            -0.171171  -0.066175  0.148556 -0.017584 -0.105811
LapStartTimeSec      0.038745  -0.047399  0.149771  0.069436 -0.000225
TimeSec              0.038766  -0.047459  0.149907  0.069313 -0.000235
AirTemp              0.003505  -0.041933 -0.232758  0.018882  0.155055
TrackTemp           -0.011250   0.206384 -0.194271  0.068966 -0.000417
WindSpeed           -0.075750  -0.134484 -0.031592 -0.073185 -0.154408
Humidity            -0.001451  -0.146600  0.106279 -0.007495  0.072357
Year                 0.103909  -0.045998  0.061732 -0.007034  0.024025
PostRegulation       0.098099  -0.099132  0.013399 -0.044004  0.002589
LapTime             -0.034264  -0.491398  0.017655  0.044239 -0.116731

                  SpeedFL   SpeedST  TyreLife  TrackStatus  Position  \
DriverNumber     0.033678 -0.050328  0.046598     0.072750 -0.171171
LapNumber        0.176432 -0.215395  0.563171    -0.065453 -0.066175
Stint            0.121149 -0.281323 -0.239577    -0.072288  0.148556
...
Humidity         1.000000  0.017621        0.003032  0.285098
Year             0.017621  1.000000        0.877523  0.141461
PostRegulation   0.003032  0.877523        1.000000  0.131764
LapTime          0.285098  0.141461        0.131764  1.000000
```

Figure 5. Snapshot of the correlation coefficient between different columns