CHARLES UNIVERSITY

Advanced Econometrics (JEM217)

Empirical Project

Determinants of Lap Times in Formula 1

This paper analyzes the determinants of Formula 1 lap times using lap-level data from the 2025 season. Employing pooled OLS, fixed effects, and quantile regression models, we examine how race conditions, strategic variables, and tyre dynamics influence performance. The results indicate that unobserved heterogeneity plays a substantial role, rendering fixed effects models more appropriate than pooled estimation. Moreover, quantile regression reveals pronounced heterogeneity across the lap time distribution, particularly for position, weather conditions, and tyre degradation, which is obscured by mean-based approaches. Overall, the findings highlight the importance of high-frequency and distribution-sensitive methods in modeling Formula 1 performance.

Authors: Sebastian Pasz & Vojtěch Dohnal

Date: January 29, 2026

This project builds on a bachelor's thesis previously written by one of the authors. The original analysis has been extended by re-estimating the model parameters using newly available data and by incorporating additional econometric methods covered during the course.

During the preparation of this output, the authors used ChatGPT for code optimization and to improve the writing style. After using this tool, the authors reviewed and edited the content as necessary and take full responsibility for the content of the report.

1. Introduction

Formula 1 represents a uniquely complex performance environment in which lap times emerge from the interaction of advanced engineering, driver behavior, regulatory constraints, and rapidly changing race conditions. Unlike many sports, performance in Formula 1 is generated at very high frequency, with each lap reflecting momentary decisions related to tyre management, pit activity, traffic, weather, and track status. As a result, lap times constitute the most granular and informative measure of on-track performance, serving as a direct link between strategic choices and race outcomes.

This paper investigates the determinants of Formula 1 lap times using lap-level data from the full 2025 season. Building on recent advances in high-frequency motorsport analytics, we estimate pooled OLS, fixed effects, and quantile regression models to disentangle the roles of above mentioned variables. While standard mean-based models provide useful benchmarks, we show that accounting for unobserved heterogeneity and distributional effects is crucial for capturing the complexity of lap time behavior.

The literature reviewed in the Section 2 situates the present study within the broader field of Formula 1 analytics, tracing the evolution from outcome-based performance evaluation toward lap-level and event-level modeling. Particular emphasis is placed on how existing research has approached race strategy, regulatory change, and performance decomposition, and where gaps remain in understanding the determinants of lap times themselves. This review provides the conceptual and empirical foundation for the subsequent analysis and motivates the methodological extensions introduced in this project.

The remainder of the paper is structured as follows. Section 3 describes the data and variable construction. Section 4 outlines the econometric methodology. Section 5 presents the empirical results, and Section 6 concludes.

2. Literature Review

The increasing availability of high-frequency motorsport data over the past decade has enabled a growing body of quantitative research on Formula 1 (F1). Early contributions focused on aggregate outcomes such as race results, championship standings, or win probabilities (e.g. Bell et al. (2016), Kesteren and Bergkamp (2023)), reflecting both data limitations and the traditional emphasis on season-level performance. More recent studies, however, exploit lap-level or event-level data to analyze the microstructure of racing performance, thereby shifting attention toward dynamic, within-race determinants.

A substantial share of the literature concentrates on *race strategy*. Simulation-based approaches model optimal pit stop timing and tyre selection under uncertainty (Bekker and Lotz (2009); Sulsters (2018)), while game-theoretic frameworks formalize strategic interactions between teams (Aguad and Thraves (2023)). With the rise of telemetry and computational power, machine learning methods have become increasingly prominent. Neural-network-based systems (Alexander et al. (2020)) and reinforcement learning models with explainability layers (Thomas et al. (2025)) aim to support real-time decision-making in stochastic race environments. Monte Carlo approaches (Heilmeier et al. (2020); Piccinotti (2019)) further emphasize the probabilistic nature of race outcomes by explicitly simulating safety car deployments, accidents, and pit variability. While these contributions provide valuable strategic insights, lap times typically enter these models as inputs or intermediate outcomes rather than as the primary object of econometric analysis.

Understanding the *determinants of lap times* is therefore a prerequisite for both strategy modeling and performance evaluation. The first studies to explicitly examine lap-time evolution focused on regulatory regimes. Guhan and Karthikeyan (2021) analyzed the impact of engine regulation changes across three technological eras (V10, V8, and V6 turbo-hybrid) using polynomial regression. Their results indicate that major regulatory shifts induce immediate performance shocks, followed by relatively rapid convergence as teams adapt. These findings highlight the importance of accounting for structural breaks and non-linear adjustment processes when modeling performance in Formula 1.

Pasz (2025) represents a shift toward high-frequency econometric modeling of Formula 1 performance. Using lap-level panel data from the 2019 season through the early part of 2025, that study applies fixed and random effects panel regressions alongside Generalized Additive Mixed Models (GAMMs) to quantify the impact of tyre life, compound choice, pit stops, weather conditions, and track status. The results indicate that time-varying race-specific factors dominate static characteristics such as driver or team identity. Moreover, the GAMM framework reveals pronounced nonlinearities and interaction effects, particularly in tyre degradation and environmental conditions, that are obscured in linear specifications.

This project directly builds on this literature and can be viewed as a methodological and empirical extension of Pasz (2025). While maintaining the lap-level focus, we restrict attention to the full 2025 season, thereby eliminating cross-season regulatory heterogeneity and enabling a cleaner identification of within-season performance dynamics. Methodologically, we complement standard fixed effects panel models with pooled OLS and *quantile regression*, an approach that has not yet been applied in this context. Unlike mean-based estimators, quantile regression allows the effects of key covariates to differ across the conditional distribution of lap times, thereby capturing heterogeneity between, for example, push laps, management laps, and disrupted laps under adverse conditions.

3. Data

3.1 Discussion of the Dataset

In our project we used dataset obtained using FastF1 Python library concerning F1 races in 2025. Although unofficial the dataset is deemed to be a valid data source as it has already been used in prior academic papers (e.g., Groote (2021)). The period covered by the data contains 24 races with total of 29146 observations in the raw form where each row corresponds to single lap of a F1 driver.

3.2 Data cleaning

Firstly, we converted the variables into proper units, standardized team names across the season and created additional binary indicators such as In/OutPit, TrackStatus. Secondly, we decided to eliminate observations with missing values but the dataset is fully realized with no missing values which we checked in Figure 2.

Then, we also decided to deal with the outliers in the dependent variable. Those are usually caused by special race situations like displayed red flag or safety car on track and therefore controlled for using the variable TrackStatus and kept in our data. On the other hand, outliers that do not correspond to the special situations on the track were filtered out.

Lastly, the correlation matrix in Figure 1 could be found in the Appendix A but as realized values are within expected bounds, no further action was taken.

3.3 Variables

The dependent variable is LapTime, measured in seconds. Numerical explanatory variables include LapNumber, TyreLife, and Position, capturing race progression, tyre degradation, and on-track position. Weather conditions are recorded at the start of each lap and include AirTemp, TrackTemp, Humidity, Pressure, and the binary indicator Rainfall. Categorical variables include Driver, Team, Compound, SessionType, Country, and TrackStatus. The binary indicators InPit, and OutPit capture in pit lane activity. Due to ambiguity caused by the position of the pit box in pit lane timing, a continuous pit time measure is not used.

For numeric variables usual descriptive summary statistics can be found in Figure 3.

4. Methodology

To estimates the determinants we employ three frameworks: Pooled OLS, fixed effects and quantile regression. Pooled OLS provides a baseline estimate of average relationships between lap times and explanatory variables, but does not account for unobserved heterogeneity across drivers, cars, and races. Fixed effects model addresses this limitation by controlling for time-invariant unit specific factors isolating the within unit variation. Finally, quantile regression is used to capture heterogeneity across the conditional distribution of lap times allowing the effects to differ between slower and faster laps.

4.1 Pooled OLS

This approach applies the OLS estimation framework to the full set of observations, treating the panel as a single cross-section and effectively abstracting from the time dimension. Under this specification, all observations are assumed to be independent and identically distributed. For the OLS estimator to yield valid inference, the classical assumptions of homoskedastic errors and absence of serial correlation must hold. However, in the present panel dataset, both assumptions are violated, leading to inconsistent estimates of the standard errors and, consequently, unreliable statistical inference. Methods to address these issues are discussed in a subsequent section.

4.2 Fixed effects

Following Wooldridge (1999) Fixed Effects panel data models address unobserved heterogeneity bias by applying a time-demeaning transformation that removes time-invariant individual effects.

$$\underbrace{y_{it} - \overline{y}_i}_{\ddot{y}_{it}} = \underbrace{\beta_0 - \beta_0}_{0} + \beta_1 \underbrace{(x_{it1} - \overline{x}_{i1})}_{\ddot{x}_{it1}} + \beta_2 \underbrace{(x_{it2} - \overline{x}_{i2})}_{\ddot{x}_{it2}} + \ldots + \beta_k \underbrace{(x_{itk} - \overline{x}_{ik})}_{\ddot{x}_{itk}} + \underbrace{u_{it} - \overline{u}_i}_{\ddot{u}_{it}} + \underbrace{a_i - a_i}_{0}$$

They effectively eliminate the individual-specific component a_i, thereby reducing bias when $\mathbb{E}[a_i \mid x_{itk}] \neq 0$. Based on prior literature and the Hausman test, we find strong evidence of heterogeneity bias at the 99% confidence level. Consequently, this paper does not consider Random Effects models, which, although more efficient than Fixed Effects models, are only appropriate when $\mathbb{E}[a_i \mid x_{itk}] = 0$.

4.3 Quantile regression

Quantile regression extends linear regression by estimating conditional quantiles of the dependent variable, allowing covariate effects to vary across different points of the conditional distribution rather than focusing solely on the mean as in OLS. This framework is particularly suitable for skewed or heteroskedastic data and provides a richer characterization of heterogeneous relationships. Practical challenges include potential non uniqueness of solutions, quantile crossing, and more complex inference, which require careful implementation and interpretation.

4.4 Additional tests

Since both heteroskedasticity and serial correlation were found to be present in the data, adjustments were made to account for these issues. Heteroskedasticity was tested using the Breusch–Pagan test, while serial correlation was examined with the Wooldridge test, and both tests produced significant results. As a result, heteroskedasticity and autocorrelation consistent (HAC) standard errors were used in the regression results.
To assess whether OLS or a fixed effects specification is more appropriate, we conducted an F-test for individual effects. The null hypothesis states that all individual-specific intercepts are equal, implying the absence of unobserved time-invariant heterogeneity and therefore the suitability of the pooled OLS model. The alternative hypothesis allows for heterogeneous intercepts across individuals, indicating the presence of significant fixed effects.

Since we rejected the null hypothesis, we conclude that the unobserved individual effects are relevant and the fixed effects model is more appropriate.

4.5 Performance Measures

In addition, we report in-sample Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) as descriptive diagnostics of model fit. These measures are not intended to evaluate predictive performance per se, but rather to provide an auxiliary indication of how well the estimated specification captures the underlying data-generating process while identifying the relevant determinants.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad \text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

5. Results

5.1 Pooled OLS

The pooled OLS estimates reported in Table 1 show that most variables are statistically significant at conventional levels, with coefficient signs that largely correspond to prior expectations. The coefficient on lap number is negative and significant, indicating systematic improvements in lap times as the session progresses, while position is positively related to lap times, suggesting slower laps for cars running further down the order.

Weather conditions, such as air temperature and track temperature, are also statistically significant, although their estimated effects differ in magnitude and in some cases in sign when compared to the fixed effects specification. Track status indicators and pit related variables exhibit large and highly significant effects, reflecting the substantial disruptions associated with safety car periods and pit stop events.

Country, session, and team indicators are mostly significant, highlighting considerable cross sectional variation captured by the pooled model. Overall, the pooled OLS model displays a high level of explanatory power, with an adjusted R^2 of 97.9%.

Nevertheless, given the restrictive assumptions of pooled estimation and the presence of unobserved time invariant heterogeneity, these results should be interpreted with caution and mainly serve as a benchmark for comparison with the fixed effects specification.

Moreover, we evaluated in sample model performance using RMSE and MAE, which were 1.58 and 0.82, respectively, indicating satisfactory predictive accuracy. We do not employ out-of-sample validation techniques, as the primary objective of this study is not to optimize predictive performance but to correctly identify and interpret the underlying determinants. The reported performance measures are therefore used solely to provide an auxiliary assessment of overall model fit.

Table 1: Model Estimates with robust SEs and p-values

	Fixed Effects			Pooled OLS		
	Estimate	SE	pval	Estimate	SE	pval
Intercept				104.01	0.43	0.00
LapNumber	-0.06	0.00	0.00	-0.06	0.00	0.00
Position	0.07	0.00	0.00	0.08	0.00	0.00
Humidity	0.12	0.04	0.00			
Pressure	0.73	0.21	0.00			
AirTemp	0.98	0.35	0.00	-0.32	0.01	0.00
TrackTemp	-0.35	0.11	0.00	-0.19	0.00	0.00
TrackStatus_2_lag1	0.58	0.26	0.03	0.66	0.09	0.00
TrackStatus_2_lag2	0.15	0.21	0.49	0.26	0.00	0.00
TrackStatus_4_lag1	1.14	0.53	0.03	1.10	0.14	0.00
TrackStatus_4_lag2	0.45	0.61	0.46	0.33	0.10	0.00
TrackStatus_5_lag2	1.57	0.66	0.02	0.94	0.50	0.06
TrackStatus_5_lag3	-0.25	0.38	0.51	-1.01	0.37	0.00
TrackStatus_6_lag1	0.09	0.35	0.79	0.03	0.17	0.82
TrackStatus_6_lag2	-0.45	0.35	0.21	-0.52	0.12	0.00
OutPit_lag1	-0.35	0.12	0.00	-0.35	0.07	0.00
OutPit	7.93	1.85	0.00	8.29	0.19	0.00
InPit	4.01	0.58	0.00	4.08	0.12	0.00
Rainfall	-0.00	0.53	1.00	-0.12	0.14	0.36
as.factor(country)Australia	11.65	4.25	0.01	-3.67	0.22	0.00
... other country proxies ...						
as.factor(Team)Racing Bulls	0.13	0.11	0.22	0.18	0.04	0.00
... other team proxies ...						
as.factor(session)Sprint	-1.24	0.43	0.00	0.29	0.05	0.00
TyreLife:CompoundHARD	0.04	0.01	0.00	0.04	0.00	0.00
TyreLife:CompoundINTERMEDIATE	0.31	0.14	0.02	0.34	0.02	0.00
TyreLife:CompoundMEDIUM	0.05	0.01	0.00	0.06	0.00	0.00
TyreLife:CompoundSOFT	0.04	0.02	0.03	0.06	0.01	0.00
CompoundHARD:I(TyreLife^2)	-0.00	0.00	0.88	0.00	0.00	0.03
CompoundINTERMEDIATE:I(TyreLife^2)	-0.01	0.00	0.00	-0.12	0.00	0.00
CompoundMEDIUM:I(TyreLife^2)	-0.00	0.00	0.71	-0.00	0.00	0.00
CompoundSOFT:I(TyreLife^2)	0.00	0.00	0.13	0.00	0.00	0.32

Fixed Effects: $\overline{R}^2$: 98.0%, F-statistic: 24063.5 on 51 and 24209 DF
Pooled OLS: $\overline{R}^2$: 97.9%, F-statistic: 19760 on 57 and 24223 DF

5.2 Fixed Effects

Turning to the fixed effects estimates reported in columns (2)–(4) of Table 1, the results broadly confirm the qualitative patterns observed under pooled OLS, while accounting for unobserved time-invariant heterogeneity across observational units. Key covariates such as lap number, position, pit-related indicators, track status variables, and tyre dynamics remain statistically significant and retain economically meaningful signs, supporting the robustness of the identified determinants once individual-specific effects are controlled for. Differences in magnitude and, in some cases, sign—most notably for weather-related variables—highlight the extent to which pooled OLS conflates within-unit and between-unit variation, reinforcing the appropriateness of the fixed effects specification for causal interpretation.

Despite the strong explanatory power of the fixed effects model, as reflected by a within adjusted R^2 of 98.0%, the associated in-sample RMSE and MAE are substantially larger, at 87.38 and 87.36 respectively. This outcome should not be interpreted as poor model performance, but rather as a mechanical consequence of the fixed effects transformation. By construction, the fixed effects estimator removes the overall intercept and all time-invariant level differences through within-group demeaning, implying that fitted values are centered around zero deviations rather than absolute lap times. As a result, prediction errors computed on the original outcome scale are inflated, even when the model explains a large share of within unit variation. Consistent with the methodological focus of this study, these diagnostics are therefore reported only as descriptive measures and do not detract from the primary objective of accurately identifying and interpreting the underlying determinants of lap time performance.

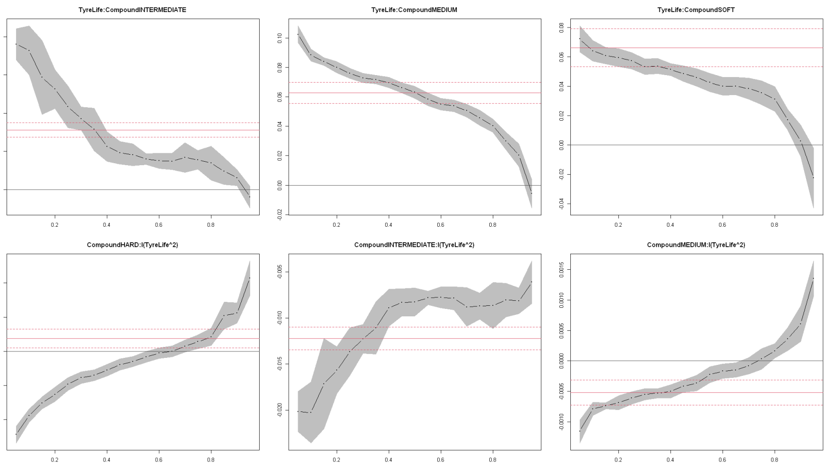
5.3 Quantile Regression

The quantile regression results show important heterogeneity in the effects of several key variables across the distribution of lap times. In particular, weather variables such as air temperature and track temperature exhibit noticeably different magnitudes and in some cases different signs across quantiles, while pooled OLS constrains these effects.

Similarly, pit related variables also display asymmetric effects across quantiles, with larger disruptions occurring in slower laps that are not fully captured by mean based estimation. The effect of position varies substantially across quantiles, with a noticeably stronger impact in the upper quantiles of the lap time distribution, indicating that running further down the order is particularly detrimental during slower laps.

Moreover, both the linear and non-linear tyre degradation effects (see Figure below) differ across quantiles, suggesting that performance decay plays an important role for certain parts of the lap time distribution, which is a pattern that the pooled OLS does not reflect.

Overall, these findings indicate that pooled OLS provides an incomplete representation of the underlying relationships by focusing solely on average effects and ignoring distributional heterogeneity that is crucial to lap time dynamics.



5.4 Limitations and Interpretation

Several limitations apply to the empirical models employed in this study and should be acknowledged to avoid an overly causal interpretation of the results. First, the pooled OLS specification relies on strong assumptions regarding the exogeneity of regressors and the absence of unobserved heterogeneity. Although it provides a useful benchmark and captures substantial cross-sectional variation, its estimates may be biased due to omitted time-invariant factors correlated with the included covariates. Consequently, pooled OLS coefficients should be interpreted as descriptive associations rather than causal effects.

Second, while the fixed effects model mitigates bias arising from unobserved time-invariant heterogeneity by exploiting within-unit variation, it does not address endogeneity stemming from time-varying omitted variables, reverse causality, or measurement error. Moreover, the fixed effects transformation eliminates all between-unit information and level effects, restricting inference to within-unit changes. As a result, the estimated coefficients reflect conditional correlations within observational units and should not be interpreted as structural causal parameters.

Finally, the quantile regression framework, although informative about heterogeneous effects across the conditional distribution of lap times, is subject to its own limitations. In particular, quantile regression estimates characterize conditional quantiles rather than mean effects and are sensitive to distributional features such as heteroskedasticity and tail behavior. Without a fully specified structural model or valid instruments, quantile-specific coefficients cannot be interpreted as causal effects at different points of the outcome distribution, but rather as descriptive measures of conditional heterogeneity.

Taken together, all three modeling approaches serve a complementary and primarily explanatory role. The results are intended to shed light on robust correlates and patterns in lap time performance, rather than to establish definitive causal relationships.

6. Summary

This project examines the determinants of Formula 1 lap times using lap level data from the full 2025 season. Employing pooled OLS, fixed effects, and quantile regression models, the study analyzes how race conditions, strategic variables, and tyre dynamics influence track performance. While pooled OLS provides a useful benchmark, formal tests indicate that fixed effects are more appropriate due to unobserved heterogeneity. The quantile regression results reveal substantial variation in covariate effects across the distribution of lap times, particularly for position, weather conditions, and tyre degradation, which are not captured by mean based models. Overall, the findings highlight the importance of accounting for both unobserved heterogeneity and distributional effects when modeling high frequency F1 performance.

Bibliography

Aguad, F. and C. Thraves (2023). “Optimizing Pit Stops Strategies with Competition in a Zero-Sum Feedback Stackelberg Game”. In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.4470115. URL: http://dx.doi.org/10.2139/ssrn.4470115.

Alexander, H., A. Thomaser, M. Graf, and J. Betz (2020). “Virtual strategy engineer: Using artificial neural networks for making race strategy decisions in circuit motorsport”. In: *Applied Sciences* 10.12, p. 4229.

Bekker, J. and W. Lotz (2009). “Planning Formula One race strategies using discrete-event simulation”. In: *Journal of the Operational Research Society* 60.7, pp. 952–961. DOI: 10.1057/palgrave.jors.2602626. URL: http://dx.doi.org/10.1057/palgrave.jors.2602626.

Bell, A., J. Smith, C. E. Sabel, and K. Jones (2016). “Formula for success: Multilevel modelling of Formula One Driver and Constructor performance, 1950–2014”. In: *Journal of Quantitative Analysis in Sports* 12.2, pp. 99–112.

Groote, J. de (2021). “Overtaking in Formula 1 during the Pirelli era: A driver-level analysis”. In: *Journal of Sports Analytics* 7.2, pp. 119–137. DOI: 10.3233/JSA-200466. eprint: https://doi.org/10.3233/JSA-200466. URL: https://doi.org/10.3233/JSA-200466.

Guhan, T. and K. Karthikeyan (2021). “Analysis of Lap Times in Formula-1 Motorsport due to Regulation Changes Using Polynomial Regression”. In: *La Pensée* 51, pp. 892–905.

Heilmeier, A., M. Graf, J. Betz, and M. Lienkamp (2020). “Application of Monte Carlo Methods to Consider Probabilistic Effects in a Race Simulation for Circuit Motorsport”. In: *Applied Sciences* 10.12. DOI: 10.3390/app10124229. URL: https://www.mdpi.com/2076-3417/10/12/4229.

Kesteren, E.-J. van and T. Bergkamp (2023). “Bayesian analysis of Formula One race results: disentangling driver skill and constructor advantage”. In: *Journal of Quantitative Analysis in Sports* 19.4, pp. 273–293. DOI: 10.1515/jqas-2022-0021. URL: http://dx.doi.org/10.1515/jqas-2022-0021.

Pasz, S. (2025). “Determinants of Lap Times in Formula 1: An Econometric Analysis of the 2019–2025 Seasons”. Bachelor’s Thesis. Charles University.

Piccinotti, D. (2019). “Open loop planning for formula 1 race strategy identification”. Master’s Thesis. Politecnico di Milano.

Sulsters, C. (2018). “Simulating formula one race strategies”. MSc Dissertation. Vrije Universiteit Amsterdam.

Thomas, D. et al. (2025). “Explainable Reinforcement Learning for Formula One Race Strategy”. In: *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, pp. 1090–1097.

Wooldridge, J. M. (1999). *Introductory Econometrics: A Modern Approach.* 1st. Cincinnati, OH: South-Western College Publishing.

Appendix A

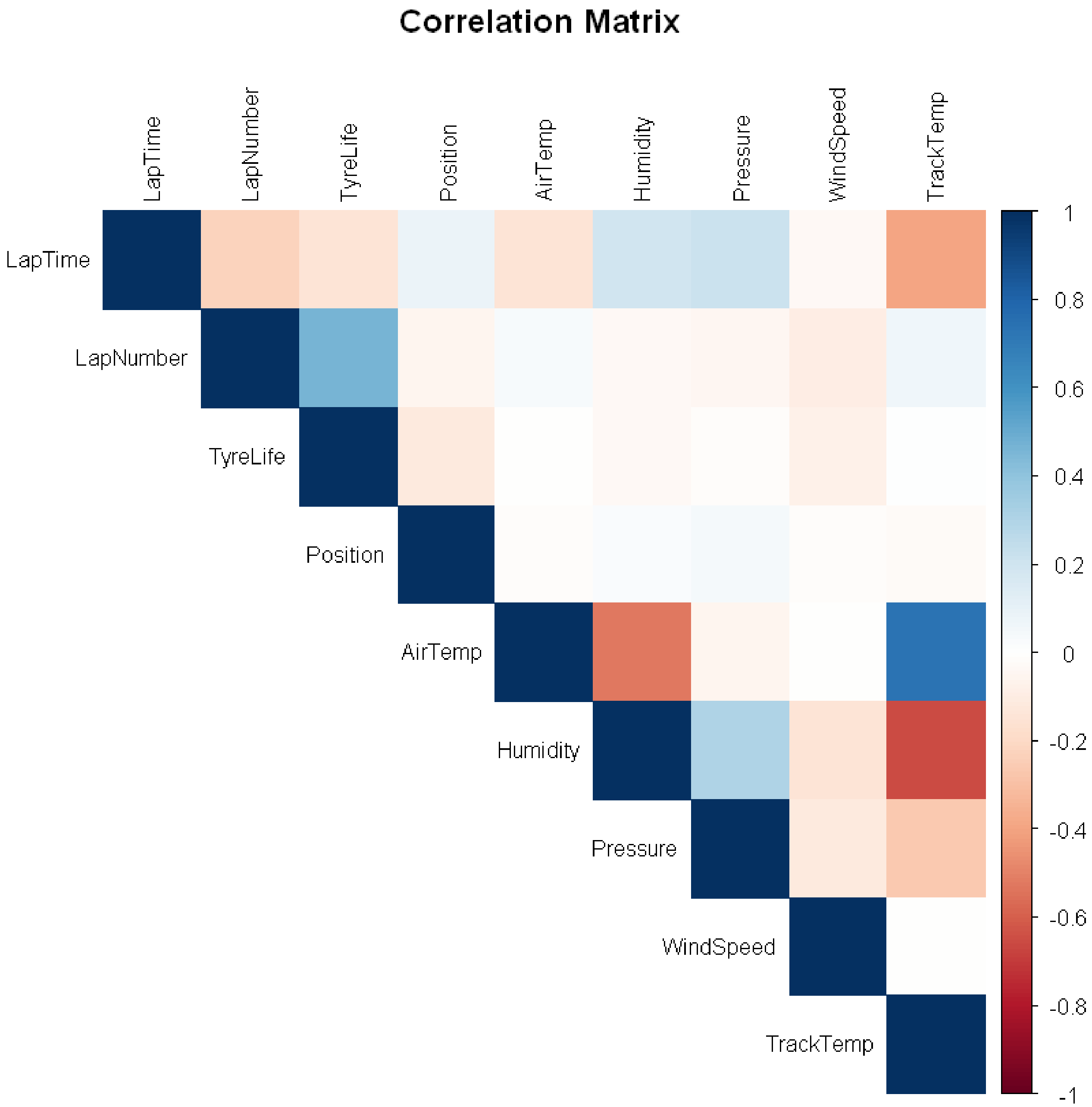


Figure 1: Correlation Matrix

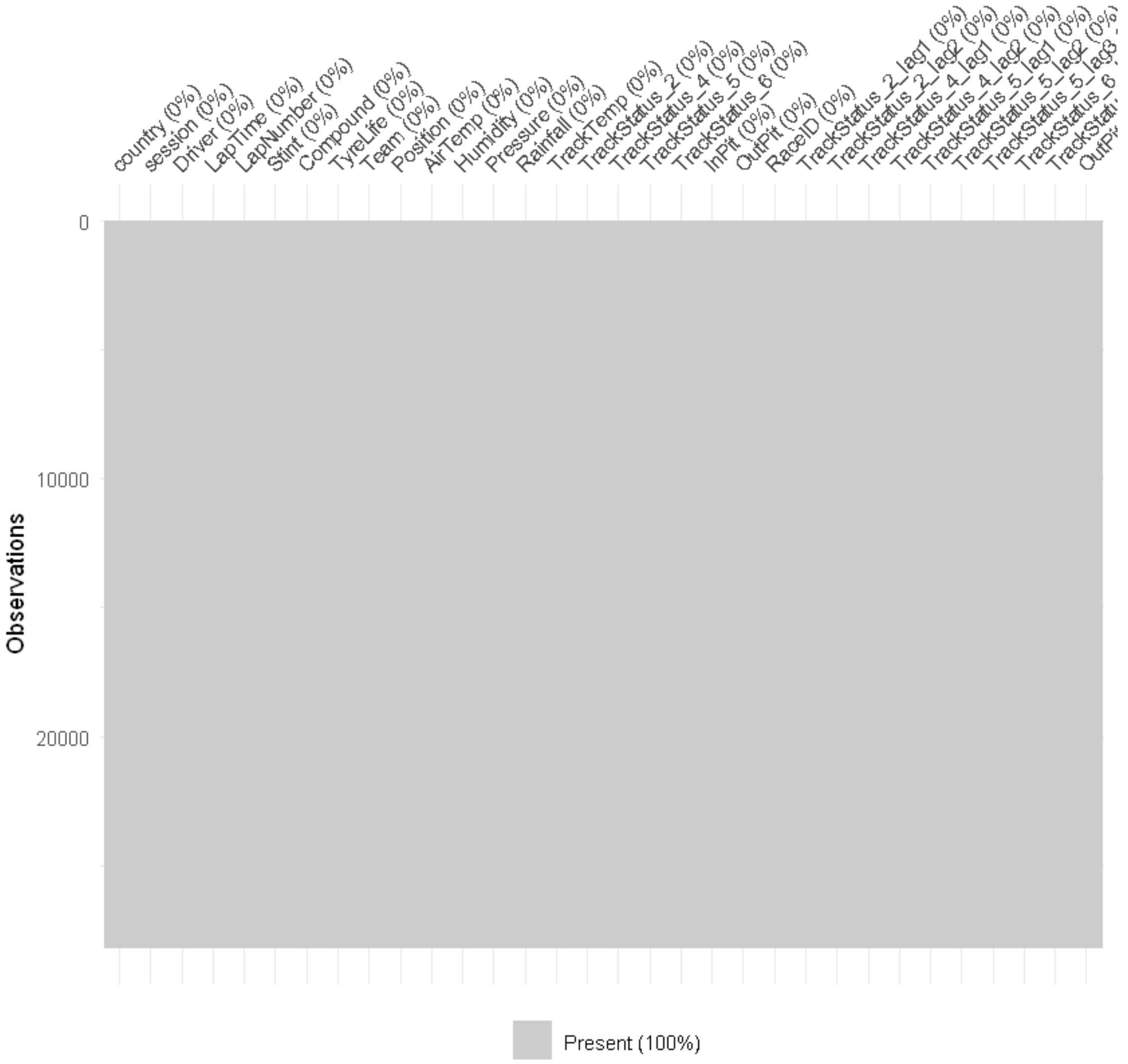
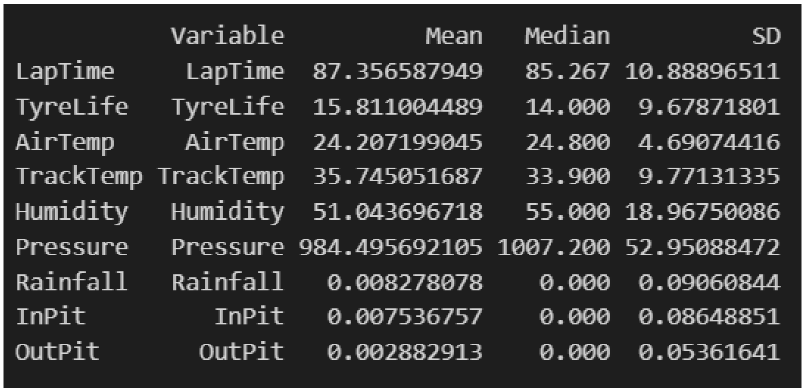


Figure 2: Missingness map



```
          Variable          Mean   Median          SD
LapTime    LapTime   87.356587949   85.267 10.88896511
TyreLife   TyreLife  15.811004489   14.000  9.67871801
AirTemp    AirTemp   24.207199045   24.800  4.69074416
TrackTemp TrackTemp  35.745051687   33.900  9.77131335
Humidity   Humidity  51.043696718   55.000 18.96750086
Pressure   Pressure 984.495692105 1007.200 52.95088472
Rainfall   Rainfall   0.008278078    0.000  0.09060844
InPit         InPit   0.007536757    0.000  0.08648851
OutPit       OutPit   0.002882913    0.000  0.05361641
```

Figure 3: Descriptive Statistics