Introduction/Background

1. **What is the Plackett_Luce Model?**

Plackett-Luce (PL) model is a probabilistic ranking model that is designed to estimate the likelihood of observing a complete (or partial) ordering of competitors based on latent performance utilities. **Rather than predicting a single outcome or treating final race position as a continuous variable, the PL model directly models the probability of the entire finishing order.** The model views a race outcome as being generated sequentially: first, the winner is selected from all competitors with probability proportional to their latent utility; then, conditional on that choice, the second-place finisher is selected from the remaining drivers, and so on.

Mathematically, if each driver i in race r has a utility $\theta_{i,r}$ the probability of observing a particular ranking (i_1, i_2, … , i_n) is given by

$$P(i_1, \ldots, i_{n_r}) = \prod_{k=1}^{n_r} \frac{\exp(\theta_{i_k,r})}{\sum_{j \in R_{r,k}} \exp(\theta_{j,r})},$$

where $R_{r,k}$ denotes the set of drivers still available at position k.

2. **Assumptions/Restrictions**

The Plackett–Luce model relies on several key assumptions.

   a. First it assumes **independence of irrelevant alternatives (IIA)**, meaning that the relative odds of one driver finishing ahead of another are unaffected by the presence or absence of other competitors. While this assumption is imperfect in Formula 1 due to incidents, safety cars, and strategic interactions, it is commonly accepted as a reasonable approximation for pre-race performance modeling.
   b. Second, the model assumes no ties, which is appropriate for F1 finishing positions.
   c. Third conditional on the modeled utilities, race outcomes are assumed independent across races
   d. Finally the model assumes that all systematic performance differences are captured through the specified predictors, with remaining randomness represented by the probabilistic ranking process.

3. **Python Implementation**

Model parameters were estimated by maximizing the Plackett–Luce log-likelihood using numerical optimization via the BFGS algorithm. **While libraries such as choix provide implementations for basic Plackett–Luce models, directly optimizing the likelihood allows for regression-style predictors, interaction terms, and time-varying covariates, which are required for this analysis.** In choix, utilities are usually: θ_i (one per driver), but in the model we are using, utilities are: $\theta_{i,r} = X_{i,r}^{\top}\beta$. For each race in the dataset, driver

utilities will be computed from the predictor matrix, and the likelihood of the observed finishing order will be evaluated. **Our exact pipeline is Plackett_Luce likelihood + MLE via BFGS = PL model. Just a numerical method for estimating it, not a different model.**
BFGS = Broyden–Fletcher–Goldfarb–Shanno: It is a quasi-newton method and is a gradient-based numerical optimization algorithm used to solve unconstrained nonlinear optimization problems.

Along with this, at the beginning, we had 13 predictors which were, qual position, AVG lap time, fast lap time, weather temp, weather humidity, weather temp var, weather humidity var, num laps, driverform, teamstrength, driver dnf rate, teamdnf rate, and overtake index. **For a PL model, the weather measurements prove to be problematic if included directly as they don't vary by driver within a race**. Including them would violate the relative comparison nature of PL. **Lap-time–based variables were excluded from the Plackett–Luce model, as they are post-race outcomes and undefined for early retirements**. Including such variables would introduce **target leakage and bias relative performance estimates.** If a variable is undefined because the race did not happen for the driver, it is not a predictor — it is an outcome

<u>Part 1</u>

Preprocessing

```
Preprocessing complete.
Missing values by predictor
QualifyingPosition    0
DriverForm            0
TeamStrength          0
DriverDNF_Rate        0
TeamDNF_Rate          0
OvertakeIndex         0
Quali_x_Overtake      0
```

The final predictor set included qualifying position, driver form (rolling 3 race avg, position-based), team strength (rolling, points-based), driver DNF rate, team DNF rate, and overtake difficulty index. This predictor avoids post-race leakage, and ensures table estimation of relative driver utilities. The above screenshot shows that after preprocessing there are no missing values. **The raw dataset is preserved, and all the cleaning and imputation are done programmatically**.

The variables team_strength, driver_dnf_rate, and team_dnf_rate are defined such that a value of zero has a meaningful outcome. For these variables, **missing values correspond to an absence of prior events rather than unobserved information**. Accordingly, the missing values are replaced with 0. This **zero-filling is applied only where zero has a clear interpretation and isn't applied to the whole dataset**.

Driver form is defined as rolling avg of the past 3 races of finishing of position, where lower values indicate stronger performance. Missing values arise mostly due to mid-season driver replacements with no prior race history, which we see a few times in the 2024 season. Due to zero not being valid or neutral on this scale, missing driver form values were imputed with race-level mean. For replacement drivers with no prior race history, driver form was imputed using a race-level neutral baseline, allowing team performance and qualifying position to drive expected outcomes without imposing unobserved driver effects. Along with this, team performance effects are captured exclusively through team_strength to avoid double-counting of car quality in the driver form variable.

As for overtake difficulty, due it not varying across drivers as they all race the same circuit, any missing values are imputed using the race-level mean. Qualifying position is a vital pre-race performance indicator. Observations with missing qualifying position are excluded, as their expected race performance then cannot be reasonably inferred.

Part 2

Standardization

Right now, our predictors are on very different scales:

Predictors	Typical Scale
Qualifying Position	1-20
Driver Form (Avg Position)	~1-20
Team Strength (points)	0-25
DNF Rates (team and driver)	0-1
Overtake Index	Arbitrary 1-5

The predictors used in the Plackett–Luce model operate on substantially different numeric scales, as seen in the table above. **Directly using these raw values would cause large-scale variables, such as team strength measured in points, to dominate the optimization process numerically and make coefficient magnitudes difficult to interpret. To address this, all predictors are standardized prior to model estimation**.

Standardization is performed by subtracting the mean and dividing by the standard deviation of each predictor, $x_{\text{standardized}} = \frac{x - \mu}{\sigma}$, resulting in variables with mean zero and unit variance. This transformation places all predictors on a common scale, improves numerical stability during likelihood optimization, and allows coefficient magnitudes to be compared directly as the effect of a one–standard-deviation change in each variable on driver utility.

Importantly, standardization preserves the directional interpretation of each predictor. For example, **qualifying position is defined such that lower values indicate stronger performance; after standardization, negative values therefore correspond to better-than-average qualifying results, while positive values indicate worse-than-average performance.** The same logic applies to all predictors, ensuring that their substantive meaning is retained while enabling stable and interpretable estimation within the Plackett–Luce framework.

Part 3

Race-Level Rankings

As mentioned earlier, we know the **PL model operates on ranked sets rather than individual observations.** This means the dataset needs to be **restructured at the race-level.** After preprocessing and standardization, drivers are grouped by race using the (Year, Round) identifiers and sorted by observed finishing position. This produces a complete ordered list of drivers for each race, representing the outcome the model aims to explain.

For each race, a standardized predictor matrix is constructed, with rows corresponding to drivers and columns corresponding to pre-race predictors. The finishing order is encoded implicitly through the sorted driver indices rather than treated as a numerical target. **Each race is stored as a separate ranked observation**, forming the core input to the Plackett–Luce likelihood.

This race-level formulation preserves the ordinal nature of finishing position and ensures that drivers are compared only against others competing in the same race. It also implicitly controls for race-specific conditions, such as track characteristics and strategy, by evaluating outcomes within a shared competitive context.

Part 4

Defining Plackett_luce log-likelihood

Driver performance is modeled through a latent utility function defined as a linear combination of standardized predictors. For each driver i in race r, utility is given by $\theta_{i,r} = X^T_{i,r}\beta$, where β represents the coefficients to be estimated. These utilities determine the probability of observing a particular finishing order.

The Plackett–Luce likelihood is computed sequentially, treating a race outcome as a series of conditional selections. At each finishing position, the probability that the observed driver is selected from the remaining competitors is proportional to the exponential of their utility. The log-likelihood for a race is obtained by summing these contributions across all positions, and the total log-likelihood is the sum over all races.

Part 5

Fitting Model

Model parameters are estimated using maximum likelihood estimation (MLE) by numerically minimizing the negative Plackett–Luce log-likelihood. Because the likelihood does not have a closed-form solution for regression-based utilities, **optimization is performed using the BFGS algorithm via scipy.optimize.minimize**. Initial coefficient values are set to zero, corresponding to equal baseline utilities across drivers prior to observing the data.

At each optimization step, the algorithm evaluates how well a given set of coefficients explains the observed race-level rankings by computing the likelihood across all races. The optimizer iteratively updates the coefficient vector to improve the fit until convergence is achieved. Standard errors are obtained from the inverse Hessian matrix returned by the optimizer, allowing for inference on coefficient significance.

Part 6

Interpret Coefficients

Constructed 23 race rankings.

	Predictor	Coefficient	StdError	z_value	p_value
0	QualifyingPosition	-1.416345	0.855899	-1.654804	9.796433e-02
1	DriverForm	-0.247034	0.105673	-2.337729	1.940129e-02
2	TeamStrength	0.398467	0.071200	5.596424	2.188184e-08
3	DriverDNF_Rate	-0.028588	0.065228	-0.438282	6.611817e-01
4	TeamDNF_Rate	0.148109	0.051587	2.871027	4.091407e-03
5	OvertakeIndex	-0.000062	1.692447	-0.000037	9.999709e-01
6	Quali_x_Overtake	0.821243	0.890227	0.922509	3.562630e-01

The figure above shows us the estimated coefficients of the Plackett-Luce model along with their standard errors, z-statistics, and two-sided p_values. Statistical significance is assessed using asymptotic z-tests, where coefficients with p-values below conventional thresholds (5% or 1%) are considered statistically distinguishable from zero. Because all predictors are standardized, coefficient magnitudes reflect the relative importance of each variable in determining race outcomes.

Qualifying position has a large negative coefficient (−1.42) and is marginally significant at the 10% level ($p \approx 0.098$). While the magnitude confirms the dominant role of grid position, the **associated standard error reflects increased uncertainty due to the inclusion of interaction terms and inherent race-level variability. Team strength exhibits a strong positive effect (0.40) with a highly significant p-value ($p < 10^{-7}$),** indicating that **car performance is one of the most consistent and reliable predictors of finishing order across races. Driver form also shows a statistically significant negative effect ($p \approx 0.014$), suggesting that recent performance contributes meaningfully to outcomes beyond qualifying position and team quality.**

Reliability-related variables capture additional sources of uncertainty. **Team-level DNF rate is statistically significant ($p \approx 0.005$)**, highlighting the importance of organizational reliability in race outcomes, **while driver-level DNF history is not statistically significant once team effects are accounted for**. The **overtake difficulty index does not exhibit a significant main effect, and the interaction between qualifying position and overtake difficulty is also not statistically significant ($p \approx 0.36$)**.

This suggests that, within the **2024 season, overtaking characteristics do not consistently moderate the effect of grid position**, and that a **substantial portion of race-to-race variability arises from unobserved factors in race factors such as strategy, safety cars, and incidents on track.**

Part 7:

Model Evaluation

```
Log-likelihood: -829.5992745753305
Mean Spearman: 0.7755475645635828
Mean Kendall: 0.6352402745995424
Top-3 accuracy: 0.6811594202898551
```

Model performance is evaluated using ranking-appropriate metrics that assess how well the predicted driver order aligns with observed race outcomes. **Traditional regression or classification measures such as R^2 or AUC are not suitable in this context, as the Plackett–Luce model predicts full rankings rather than individual outcomes**. Instead, evaluation focuses on likelihood-based and rank-correlation measures.

The **log-likelihood (−829.6)** summarizes how well the model explains the observed race rankings under the assumed probabilistic structure. Higher (less negative) values indicate a better fit, and this metric is primarily used for comparing alternative model specifications rather than as a standalone measure of accuracy. In this case, the log-likelihood confirms that the model assigns relatively high probability to the observed race outcomes given the predictors.

Rank correlation metrics provide an interpretable measure of predictive alignment. The **mean Spearman correlation (0.78) indicates a strong monotonic relationship between predicted and observed finishing orders, while the mean Kendall's τ (0.65) reflects substantial agreement across driver pairwise comparisons. Finally, top-3 accuracy (0.68) measures how often the model correctly identifies podium contenders, capturing practical relevance in a highly stochastic racing environment.** Together, these results demonstrate that the Plackett–Luce model captures a meaningful portion of the systematic structure in Formula 1 race outcomes while appropriately accounting for inherent race-to-race variability.

<u>Part 8:</u>

PL Race Level Sheet

This excel sheet doesn't show one row per race, but it shows one row per driver per race. For each 23 races we have 20 rows, each containing what the Plackett-Luce model assigns to that driver in that specific race.

The first 3 columns are descriptive columns with year, round, and driver. We then have the **true finishing position** from the race. This is **what the model is evaluated against**.

The next column is the **predicted rank**. This is the **model's most likely ranking based on estimated utilities**. It is computed by sorting drivers' predicted utility: θ = Xβ, where **lower is equal to predicting to finish higher**. It **isn't probability** its the **ranking mode**, and the interpretation would be **if the model had to pick a single finishing order, this is where it would place the driver**.

The next column is the **PL_Win_Probability**, which is the **PL marginal probability that the driver wins the race**:

$$P(\text{driver wins}) = \frac{e^{\theta_i}}{\sum_j e^{\theta_j}}$$

Probabilities sum to 1 within each race and even **dominant drivers rarely exceed 0.4-0.5** in a chaotic race. Interpretation would be **given to all the drivers in the race, how likely is this driver to win according to the model.**

The next column is the **Race_LogLikelihood**. This is the log-likelihood contribution of the entire ranking. **Same value for all drivers** in the race and **summarized how well the model explains the race outcome**. Interpretation would be, **less negative** means **race outcome aligns better with model expectations,** while **more negative** means the **race was harder to predict (incidents, chaos, in race factors)**.

The last column is the **Approx_Pl_Top3_Probability**. This is an **approximate measure of podium contention and is derived from scaled win probabilities**. Interpreted as h**ow likely the model believes this driver is to be in podium contention, relative to others**.

Part 9:

Conclusion

This study applied a Plackett–Luce ranking model to Formula 1 race outcomes in the 2024 season, modeling finishing order as a probabilistic ranking driven by pre-race performance indicators. **The results highlight qualifying position and team strength as the dominant determinants of race outcomes, with recent driver form providing additional but smaller explanatory power.** Reliability at the team level also emerged as a meaningful factor, while driver-specific reliability and overtaking difficulty showed limited direct influence once other predictors were accounted for. These findings align closely with domain knowledge in F1, reinforcing the central role of grid position and car performance in shaping race results.

In terms of predictive performance, the model demonstrated strong ranking accuracy given the inherent stochasticity of racing. A **mean Spearman correlation of approximately 0.78 and Kendall's τ of 0.65 indicate substantial agreement between predicted and observed finishing orders, while a top-3 accuracy of 0.68 shows that the model reliably identifies podium contenders.** These results suggest that the Plackett–Luce framework effectively captures the systematic structure in race outcomes while appropriately allowing for randomness due to strategy, incidents, and race dynamics.

Future work could extend this framework by incorporating in-race information, such as safety car deployments, pit stop timing, tire strategies, and lap-by-lap position changes, which are not captured by pre-race predictors alone. More flexible models, including dynamic or time-varying Plackett–Luce models, state-space approaches, or survival-based hazard models for DNFs, may further improve predictive accuracy by explicitly modeling race evolution. Additionally, machine learning–based ranking models or hybrid approaches combining probabilistic rankings with telemetry data could offer promising avenues for capturing complex, non-linear race dynamics.