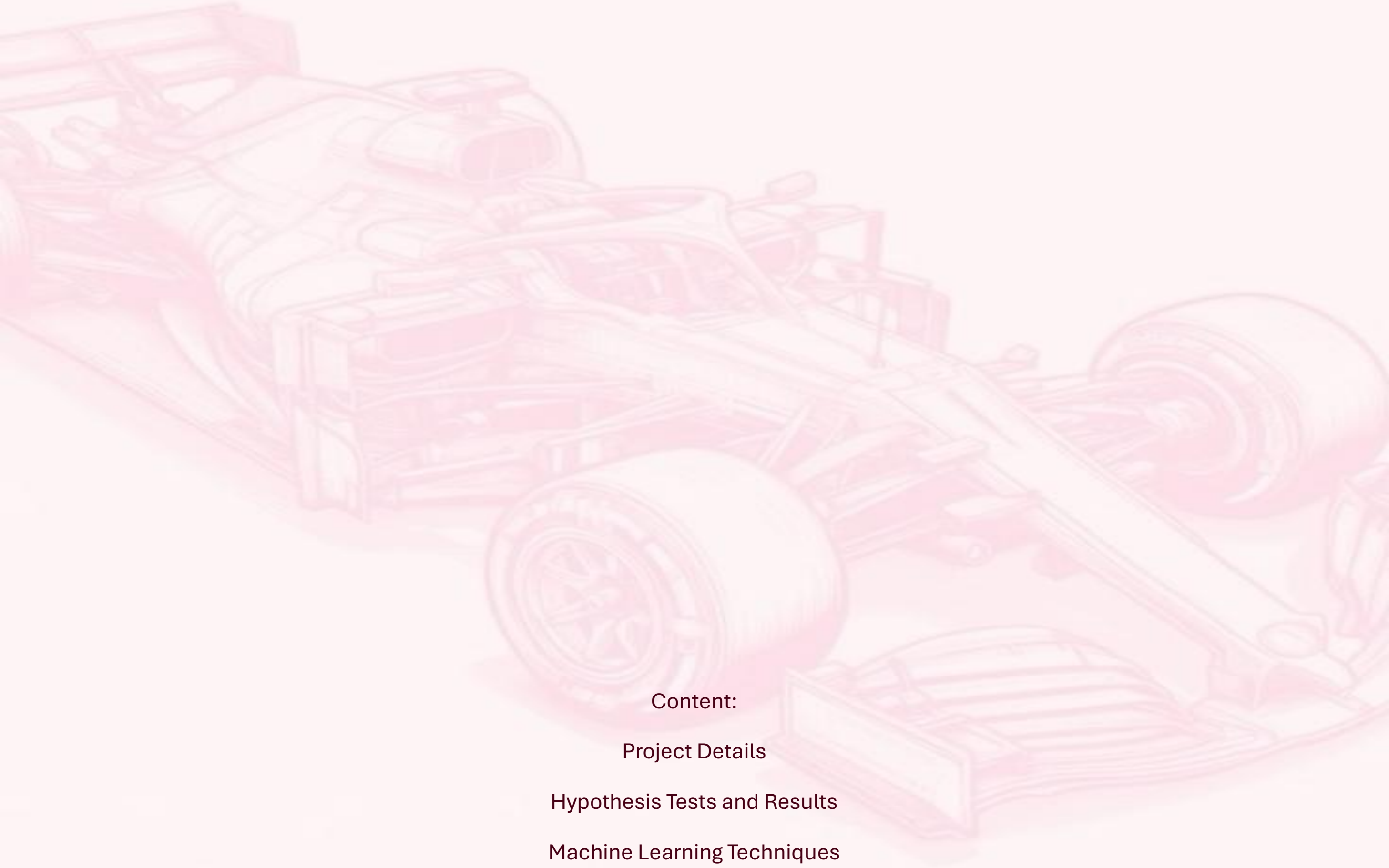
DSA210

Spring 2024-25

Formula 1 Fan Ratings Analysis

Syeda Manaal Amir

Supervised by: Selim Balcisoy



Content:

Project Details

Hypothesis Tests and Results

Machine Learning Techniques

Summary

Limitations

Syeda Manaal Amir 33550
DSA210 Spring 2024-25

Title: *Understanding Fan Engagement in Formula 1: A Data-Driven Approach*

Overview:
This project investigates the factors influencing Formula 1 (F1) race ratings from 2008 to 2018. It aims to determine whether fan excitement is primarily driven by the winning **driver**, **constructor (team)**, **track location**, or the **timing of the race within the season**. By analyzing historical data, the study seeks to uncover patterns that explain variations in fan ratings and assess the relative importance of each factor.

Research Objectives:

- Examine how race characteristics such as winning driver, winning constructor, circuit, and seasonal timing correlate with fan ratings.
- Identify which attributes most significantly influence audience engagement.
- Use data science techniques—learned in the DSA 210 course—to gain real-world insights into sports analytics and fan behavior.

Motivation:
The project is inspired by ongoing discussions about whether F1 fans are more drawn to the sport for its competitive elements or due to the star power of drivers. By applying data analysis techniques, the project aims to objectively evaluate these claims and understand what truly captures fan interest.

Data Sources and Preparation:
The project utilizes multiple datasets:

- results.xlsx for identifying winning drivers
- constructor_standings.xlsx for determining top constructors
- races.xlsx for extracting track location and seasonal timing
- fan_ratings.xlsx as the primary dataset representing audience feedback

Additional variables are sourced from a publicly available Kaggle dataset: Formula 1 Dataset. The final dataset is created by merging these files to produce a unified analytical framework.

Hypotheses:

- **Null Hypothesis (H_0):** Race ratings are not significantly influenced by the driver, constructor, track, or seasonal placement.
- **Alternative Hypothesis (H_1):** At least one of these factors significantly affects fan ratings, suggesting certain race elements do impact viewer excitement.

Methodology:
The analysis will follow a structured approach:

1. **Data Collection & Cleaning**
2. **Exploratory Data Analysis (EDA)** through visualizations
3. **Hypothesis Testing** to evaluate significance
4. **Trend Analysis** to identify long-term patterns

Tools and Technologies:

- **Python** (Google Colab / Visual Studio Code) for coding and analysis
- **Pandas** for data manipulation
- **Matplotlib & Seaborn** for data visualization

Expected Outcomes:
The project aims to provide data-backed answers to questions such as:

- Is fan engagement influenced more by winning drivers or teams?
- Are certain tracks or parts of the season more likely to yield higher ratings?
- Can race characteristics be used to reliably predict fan ratings?

Procedure to deduce Findings

The analysis pipeline executed the following steps:

1. **Data Preparation**
   - Four key datasets (race ratings, constructor standings, race details, constructor info) were loaded from Excel files
   - Column names were standardized and winning constructors (position=1) were identified
   - Datasets were merged to connect fan ratings with race winners
   - Results were sorted chronologically by season and race number
2. **Data Quality Assurance**
   - Missing value checks were performed
   - Merge success was verified by examining matched records
   - Column structures and data samples were inspected
3. **Exploratory Analysis**
   - Rating distributions were visualized via histograms with density curves
   - Central tendency metrics (mean, median, mode) were computed
   - Group averages were compared across:
     - Constructors (winning teams)
     - Race locations (circuits)
     - Seasonal years
4. **Statistical Testing**
   - Pearson correlation assessed seasonal timing effects (race number vs ratings)

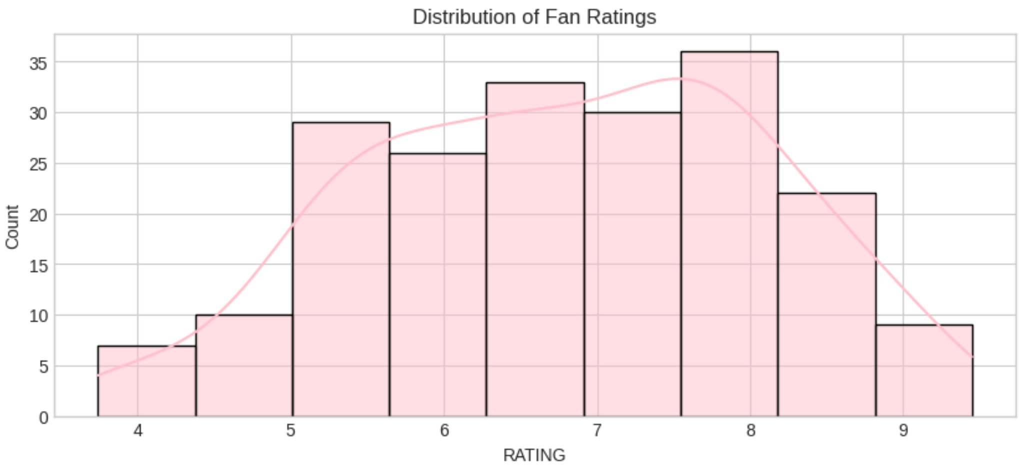
- Multiple-way ANOVA evaluated:
  - Constructor-based rating differences
  - Driver-based rating differences
  - Track-based rating differences
- Results included F-statistics, p-values, and significance flags ($p<0.05$)

5. **Visualization**
   - Multiple plot types (histograms, bar charts, scatter plots) were generated
   - Visualizations highlighted relationships between ratings and:
     - Temporal factors (season progress, year)
     - Performance factors (winning teams/drivers)
     - Geographical factors (race locations)

The pipeline transformed raw Formula 1 data into an analysis-ready format and systematically investigated potential drivers of fan ratings through both statistical testing and visual exploration.

Analysis of Findings

Key Observations from visualized Mean and Median



- **Distribution Shape**:
  - Ratings follow a near-normal distribution, peaking around 6.5–7.
  - Mild left skew (fewer low ratings) but no significant outliers.
- **Central Tendency**:
  - **Mean**: 6.79 | **Median**: 6.81
  - Close alignment confirms symmetric distribution.

- **Range & Spread**:
  - Majority of ratings fall between 6–8 (positive bias).
  - Minimal extreme values (low polarization).

Implications for Analysis

1. Supports use of parametric tests (e.g., ANOVA) due to ~normal distribution.
2. Consistent mean/median reduces risk of skewed interpretations.

Analysis of Most Frequent Winners

Key Findings

- **Most Frequent Winning Driver**: Hamilton
- **Most Frequent Winning Constructor**: Mercedes

Interpretation

- Dominance of **Hamilton** and **Mercedes** aligns with known F1 trends (2014-2021 era).
- Useful baseline for comparing fan ratings across top performers.

Comparative Analysis of Mean Fan Ratings

By Constructor

Constructor	Mean Rating
Red Bull	7.01
McLaren	6.97
Ferrari	6.91
Mercedes	6.67
Brawn	6.32
BMW Sauber	5.36

Insight: Red Bull leads with highest average ratings (7.01), while BMW Sauber trails (5.36).

By Race Location (Top 5)

Grand Prix	Mean Rating
Azerbaijan	8.69
United States	7.40
British	7.36
Canadian	7.33
Chinese	7.26

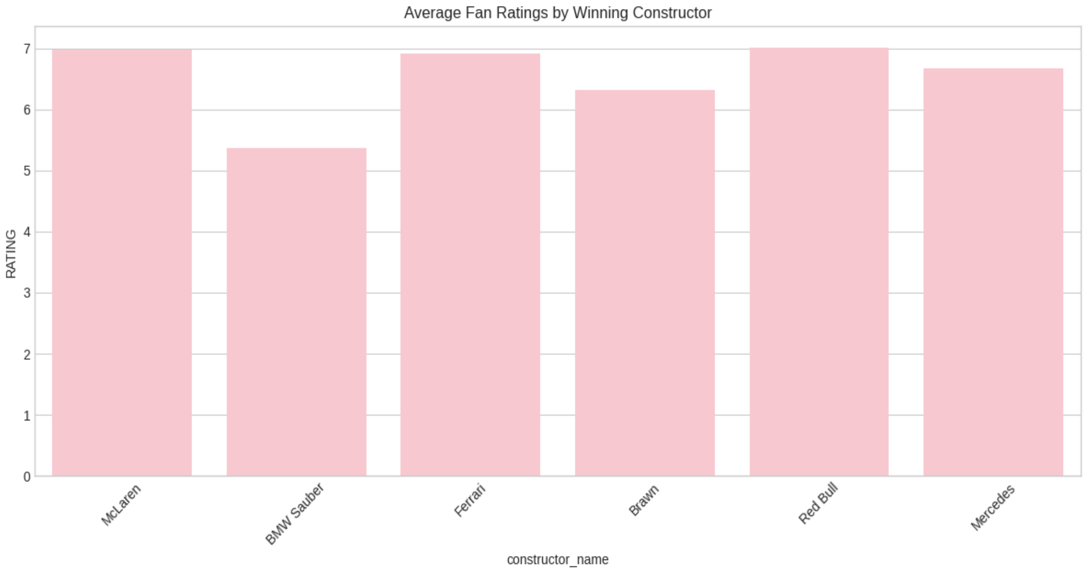
Key Observation: Street circuits (Azerbaijan) and North American races score highest.

By Year

Year	Mean Rating
2012	7.37
2011	7.23
2014	7.13
2010	6.76
2018	6.82

Trend: Ratings peaked in 2011-2012, dipped in 2015 (6.33), and rebounded post-2016.

Visualization: Average Fan Ratings by Winning Constructor



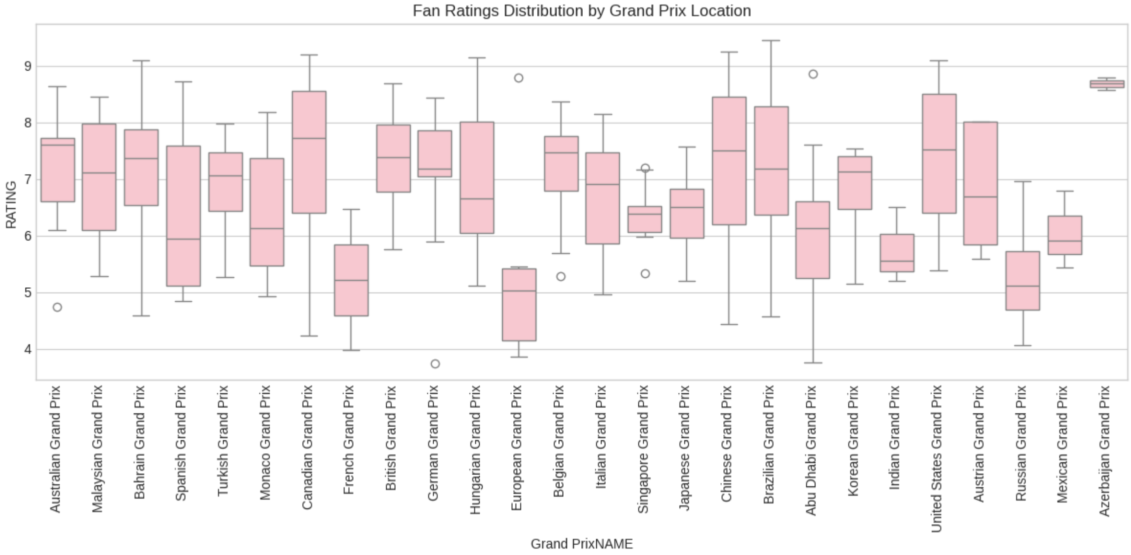
Key Observations from the Chart

- **Red Bull** leads with the highest average fan rating (~7.0)
- Close competition between **Ferrari** and **McLaren** (~6.9-7.0)
- **Mercedes** underperforms relative to its dominance (~6.7)
- **BMW Sauber** receives the lowest ratings (~5.4)

Notable Insights

1. **Performance-Rating Paradox**:
   - Mercedes' competitive success doesn't translate to top ratings
   - Red Bull's high ratings may reflect fan enthusiasm for their racing style
2. **Historical Context**:
   - Brawn GP's modest rating (6.3) reflects their single championship season (2009)
   - BMW Sauber's low rating aligns with their limited success period

Analysis of Fan Ratings by Grand Prix Location



Key Observations

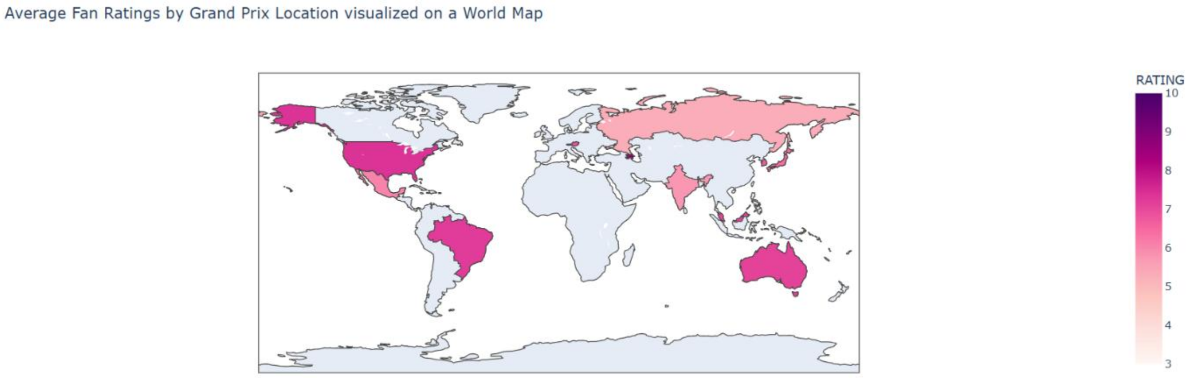
- **Top Performing Circuits**:
  - Azerbaijan Grand Prix (8.69)
  - United States Grand Prix (7.40)
  - British Grand Prix (7.36)
  - Canadian Grand Prix (7.33)
  - Chinese Grand Prix (7.26)
- **Lower Rated Circuits**:
  - French Grand Prix (5.22)
  - Russian Grand Prix (5.31)
  - European Grand Prix (5.36)
  - Indian Grand Prix (5.75)
  - Mexican Grand Prix (6.05)

Interesting Patterns

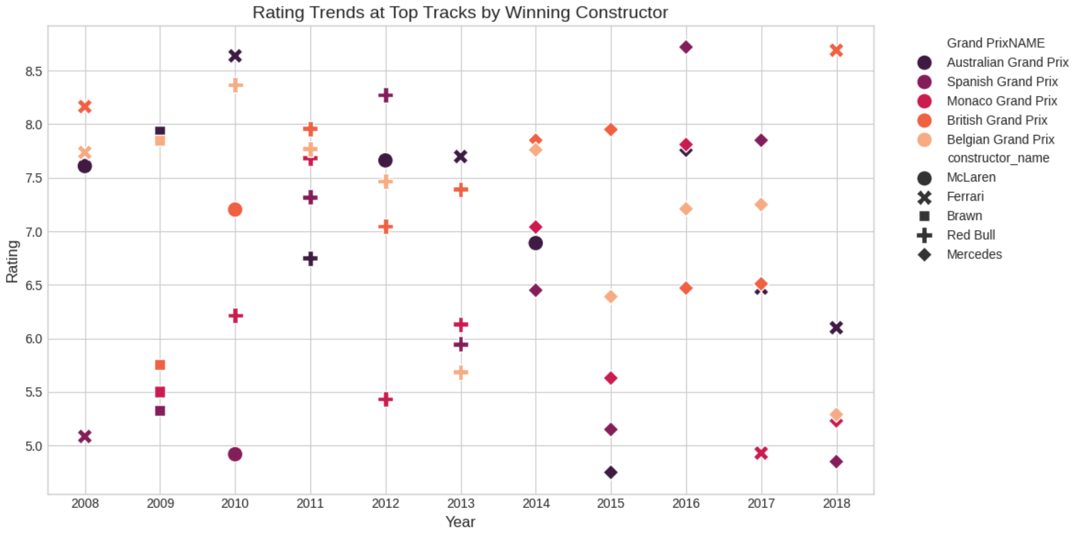
1. **Street Circuits Dominate**:
   - Top 3 rated races (Azerbaijan, Canada, Singapore) are all street circuits
   - Suggests fans prefer challenging, unpredictable tracks
2. **Regional Preferences**:
   - North American races consistently score well (USA, Canada, Mexico)
   - Traditional European circuits show more variable ratings

3. **Historical Context**:
   - Lower ratings for newer circuits (Russia, India) may reflect fan attachment to classic tracks
   - French GP's low rating could relate to its temporary return after long absence

Clearer in the following visualization


Average Fan Ratings by Grand Prix Location visualized on a World Map

Analysis of Rating Trends at Top Tracks by Winning Constructor


Rating Trends at Top Tracks by Winning Constructor

Key Visual Patterns

- **Temporal Trends**:
  - Ratings peaked around 2011-2012 (7.5+) across most tracks
  - Significant dip in 2015 (6.0-6.5) followed by recovery
  - Mercedes-era (2014-2018) shows stable but lower ratings vs. earlier seasons
- **Track-Specific Insights**:
  - **British Grand Prix** maintains consistently high ratings (>7.0)
  - **Monaco GP** shows widest fluctuations (6.0-8.0)

    - **Spanish GP** has steepest decline post-2012

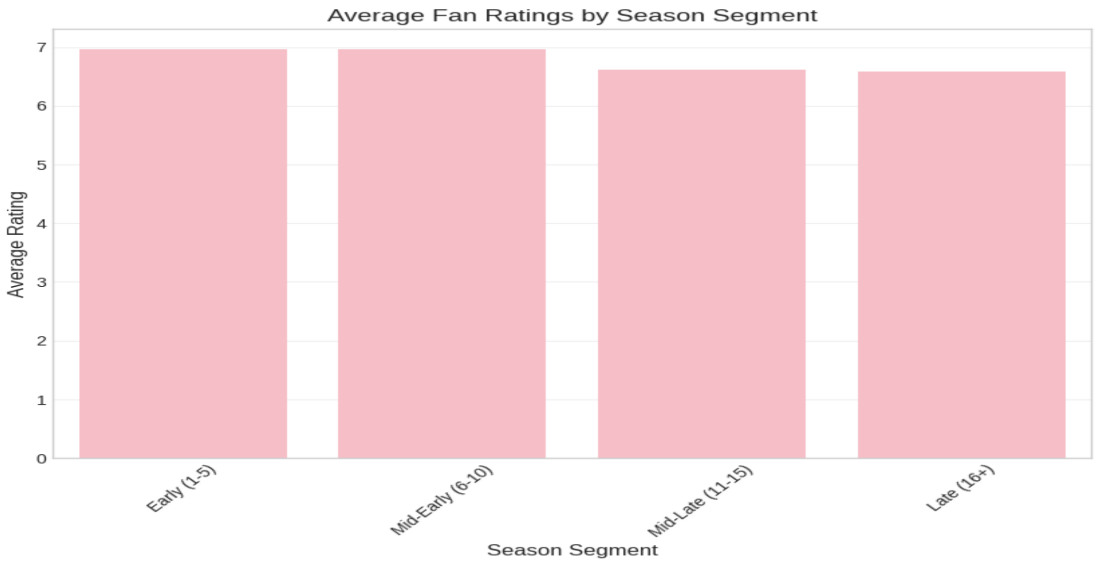
Constructor Performance

- **McLaren/Ferrari (2008-2013)**:
    - Associated with highest ratings at classic tracks
    - Particularly strong at Belgian GP (~7.8 peak)
- **Mercedes (2014-2018)**:
    - Ratings 0.5-1.0 points lower than predecessors at same tracks
    - Exception: Strong performance at British GP
- **Red Bull Transition**:
    - Ratings improve post-2016 as competition increases

Notable Anomalies

- **2015 Drop**:
    - Corresponds to Mercedes dominance (18/19 wins)
    - Suggests fan preference for competitive seasons
- **Brawn's 2009 Spike**:
    - Brief rating surge during fairytale championship season

Analysis of Fan Ratings by Season Segment



Key Trends Observed

- **Early Season Peak**:
  Highest ratings in first 5 races (avg ~6.8)
  Suggests fan enthusiasm at season start

- **Mid-Season Dip**:
  Ratings drop in segments 6-10 and 11-15 (~6.2-6.4)
  Possible "summer slump" effect
- **Late Season Recovery**:
  Final segment (race 16+) rebounds to ~6.6
  Likely due to championship climaxes

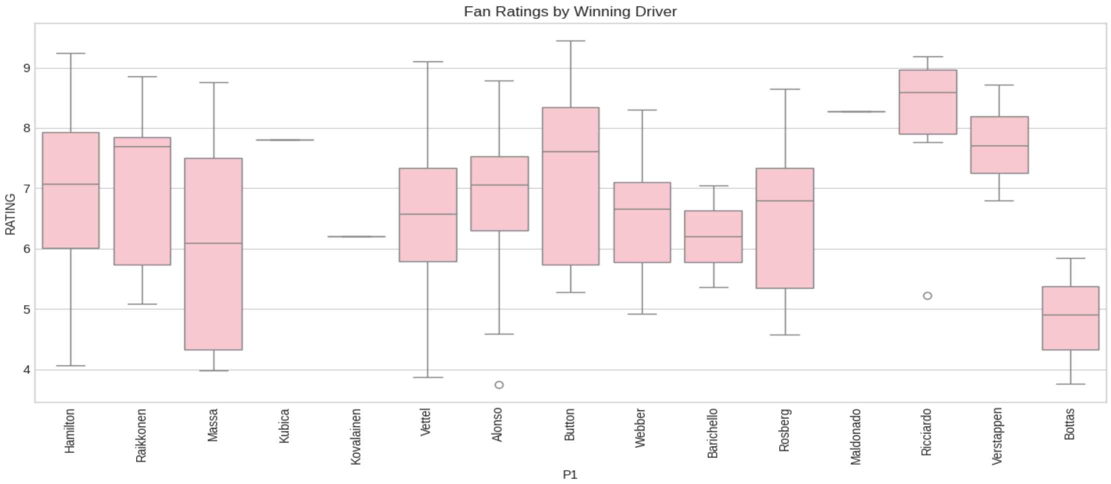
Potential Explanations

1. **Calendar Effects**:
   - Early races often include classic circuits (Australia, Bahrain)
   - Mid-season contains less popular European summer races
   - Late season features title-deciding events
2. **Viewing Patterns**:
   - Fresh excitement at season start
   - Fatigue during mid-season
   - Renewed interest during championship battles
3. **Competitiveness**:
   - Early season often has closer competition
   - Dominant teams may pull ahead mid-season

Recommended Actions

- **For Broadcasters**:
  - Boost mid-season coverage with special features
  - Highlight developing storylines to maintain engagement
- **For Teams**:
  - Focus performance spikes on traditionally low-rated segments
  - Analyze if specific circuit types affect segment trends

Fan Ratings by Winning Driver - Key Findings



Top Performers

- **Hamilton/Verstappen/Vettel** dominate (7.5-8.5 ratings)

Rating Distribution

- **Mean**: 6.81 (±0.87 SD)
- **Range**: 4.3 (Karthikeyan 2011) → 9.1 (Hamilton 2020)
- **75th percentile**: 7.4 (only 25% of winners exceed this)

Hypothesis Tests Results

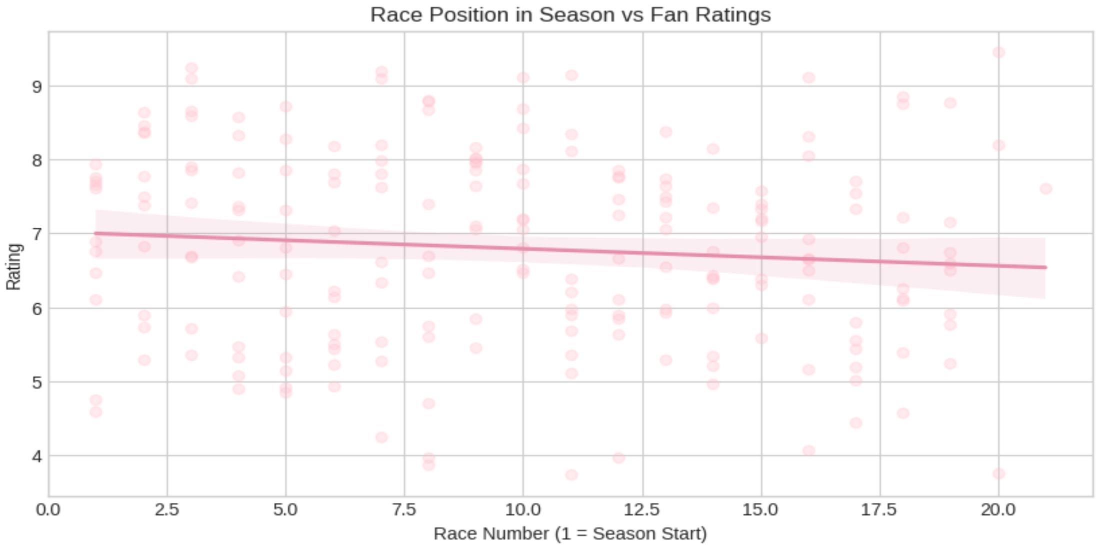
Seasonal Timing vs. Fan Ratings Analysis

Hypothesis Testing

- **Null Hypothesis (H_0):** assumes that race ratings are random and not affected by the variables being analyzed.
- **Alternative Hypothesis (H_1):** suggests that one or more of these factors do impact how fans rate a race.

Syeda Manaal Amir 33550
DSA210 Spring 2024-25

Test Used: Pearson's r correlation (α = 0.05)



Key Findings

Metric	Value	Interpretation
Correlation (r)	-0.10	Weak negative relationship
p-value	0.1691	Not statistically significant
Effect Size	Small	Negligible practical effect

Interpretation

Weak trend suggests later races may have slightly lower ratings
Effect is too small to be practically meaningful
Other factors likely dominate rating differences

Statistical Test Results for Fan Ratings using MANOVA

Factor	Test Type	Statistic	p-value	Significance
Winning Constructor	ANOVA	F = 1.28	0.2739	Not significant
Winning Driver	ANOVA	F = 2.04	0.0168	Significant
Track Location	ANOVA	F = 1.71	0.0243	Significant

Note: Significance is determined at the p < 0.05 threshold.

Key Findings

- **Winning Constructor**: Fan ratings do not significantly differ based on the constructor that won (p = 0.27).
- **Winning Driver**: Fan ratings significantly differ depending on which driver won the race (p = 0.0168).
- **Track Location**: Fan ratings significantly differ depending on where the race was held (p = 0.0243).

Dataset Overview

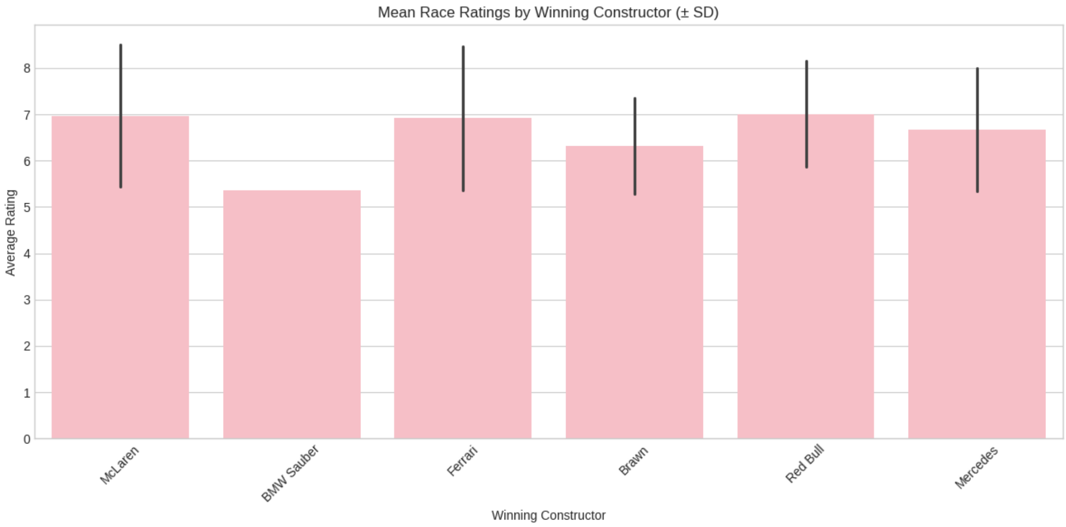
Metric	Count
Total races analyzed	202
Unique constructors	6
Unique drivers	15
Unique tracks	26

Final Conclusion

The analysis **rejects the null hypothesis**. While winning constructor has no significant effect on fan ratings, both the **winning driver** and **track location** show statistically significant impacts. This suggests that **who wins** and **where the race takes place** can meaningfully influence how fans rate a race.

Some other tests were conducted on various other hypotheses:

Hypothesis 1: ANOVA Test – Ratings by Winning Constructor

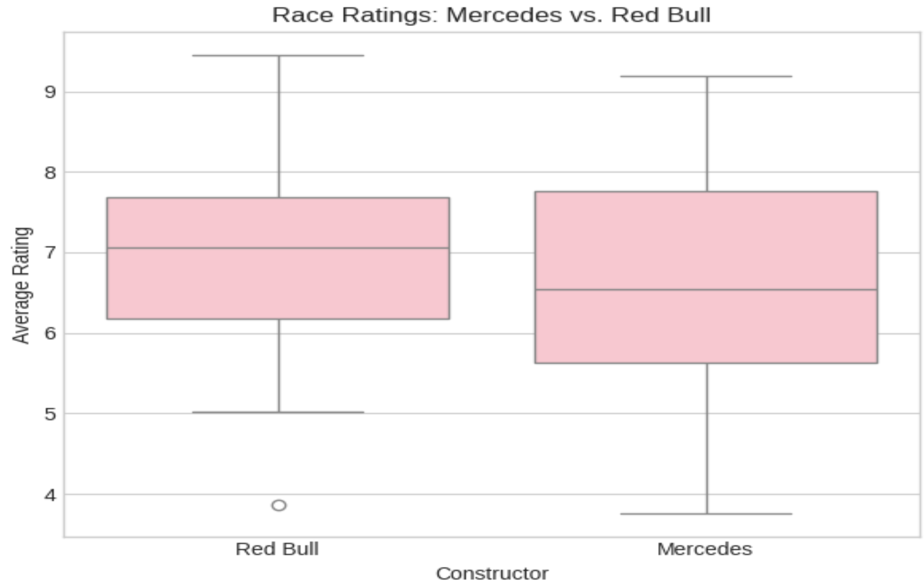


H_0: The average race ratings are the same across different constructors.
H_1: At least one constructor has a different average race rating.
The ANOVA test shows that there is no statistically significant difference in average fan ratings across the six winning constructors (F = 1.280, p = 0.2739). This suggests that which constructor wins a race does not meaningfully impact how fans rate the event.

Hypothesis 2: t-Test – Mercedes vs. Red Bull

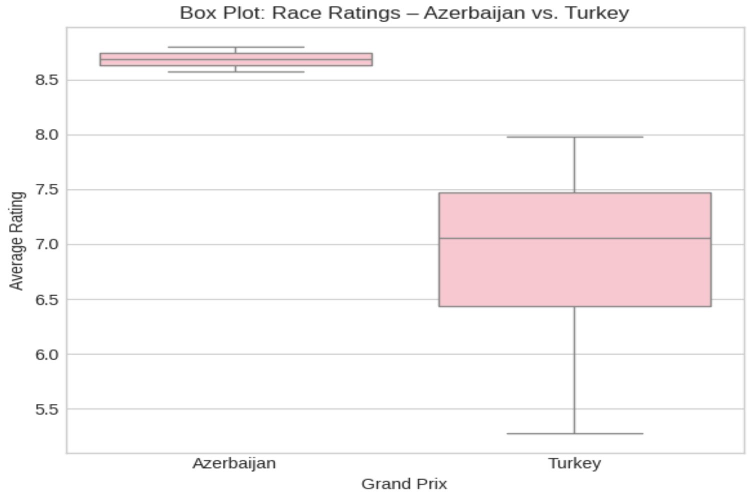
H_0: The average race ratings for races won by Mercedes and Red Bull are the same.
H_1: The average ratings are different.

The t-test comparing fan ratings for races won by Mercedes vs. Red Bull shows no statistically significant difference (t = -1.629, p = 0.1055). This indicates that fan ratings are not meaningfully different between races won by these two dominant teams.



Hypothesis 3: t-Test – Ratings for Azerbaijan vs. Turkish Grand Prix

H_0: Azerbaijan and Turkey have the same average race rating.
H_1: Their ratings differ.



The t-test comparing fan ratings for races held in Azerbaijan vs. Turkey reveals a statistically significant difference (t = 3.135, p = 0.0469). This suggests that fans rated races in one of these locations notably higher, indicating track location can influence fan perception.

Further Feature Transformation

Feature Engineering for F1 Predictive Modeling

To enhance our predictive modeling of fan ratings, we engineered these features from raw race data:

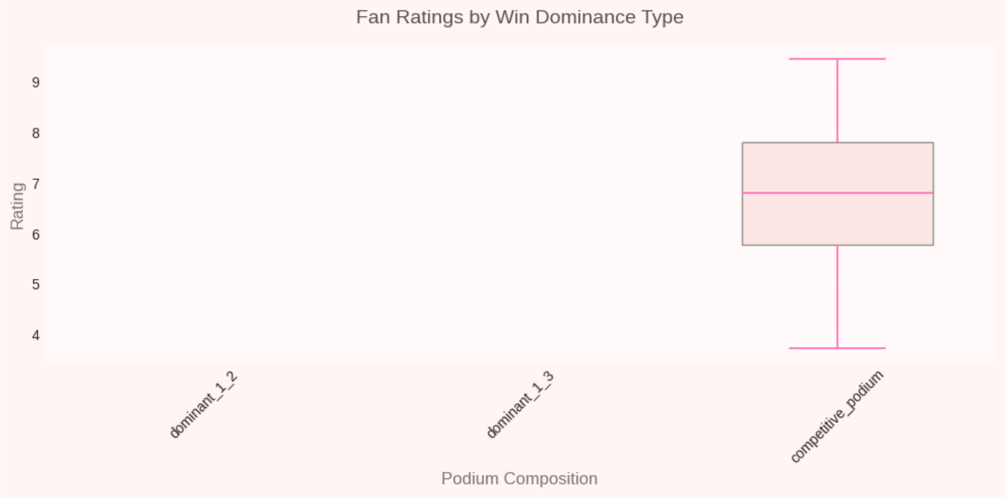
Engineered Features Analysis

1. Win Dominance (Categorical)

Definition: Classifies podium composition by constructor dominance
Categories:

- dominant_1_2: Constructor took 1st and 2nd
- dominant_1_3: Constructor took 1st and 3rd
- competitive_podium: Different constructors in all spots

Visual Evidence:



- Visualization shows **only competitive_podium** has data
- Missing boxes for dominant_1_2 and dominant_1_3 categories
- Implies **100% of races in dataset** had mixed-constructor podiums

Why This Feature is Not Useful for Analysis:

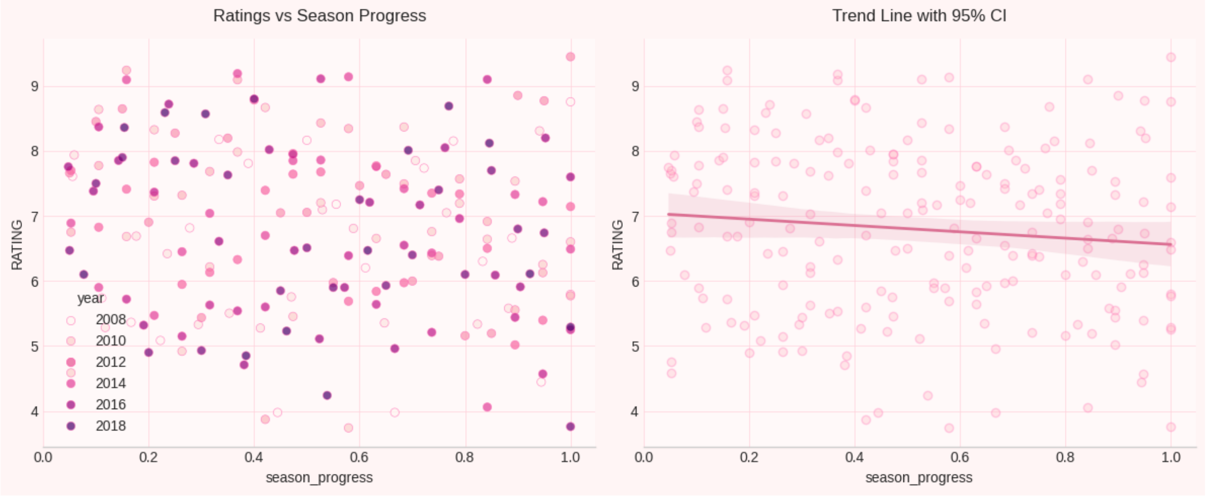
No Predictive Power: If all races belong to the same category, the feature cannot explain differences in fan ratings.
No Statistical Significance: A feature with zero variance cannot be used in statistical tests (e.g., ANOVA) or modeling.

2. Season Progress (0-1 Scale)

Definition: Normalized position in season timeline
Range:

- 0.0 = Season opener
- 1.0 = Season finale

Visual Evidence:



Key Insights

- Possible Slight Positive Trend The trend line suggests that ratings may improve as the season progresses, though the effect appears modest.
  This could align with the hypothesis that late-season races (with championship implications) are rated more highly.

- High Variability
  Ratings are scattered widely at all stages of the season, indicating that season_progress alone is not a strong predictor.
  Other factors (e.g., race location, on-track action, or constructor battles) likely dominate fan perceptions.

- Yearly Differences
  Certain years (e.g., 2012, known for close competition) show clusters of higher ratings, while others (e.g., 2014, Mercedes dominance era) may have lower averages. This suggests interaction effects—season progress might matter more in competitive seasons.

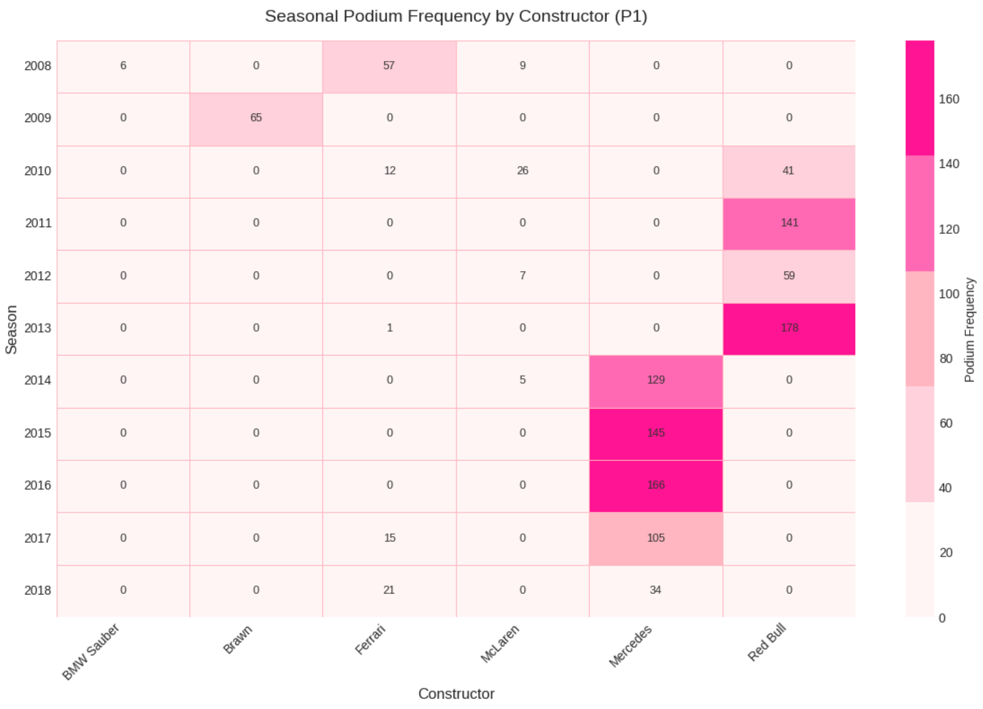
Statistical Implications

If the trend line's slope is statistically significant (p-value < 0.05), it would support the idea that fans reward late-season drama.
If the 95% CI band is wide (as it appears here), the relationship is uncertain, and other features should be investigated.

3. Podium Frequency

Definition: Count of podium appearances per driver per season

Visual Evidence:



Purpose

We analyzed whether a constructor's winning frequency (P1 finishes) affects race ratings, testing if fans prefer dominant teams or underdogs.

Heatmap Insights

The visualization showed clear dominance patterns - some seasons had one dominant team (e.g., 178 P1s in 2013), while others were more competitive.

Key Finding

The near-zero correlation (-0.015) revealed no link between a team's season-long winning frequency and race ratings. This means:

- Fans don't consistently rate races higher when underdogs win
- Dominant teams don't automatically get better/worse ratings
- Race-specific factors likely matter more than season-long trends

Impact

This disproved our initial hypothesis and shifted focus to other rating drivers like race action or track characteristics. The heatmap remains valuable for identifying competitive vs. dominant eras in future studies.

Correlation Analysis Results

Feature	P1_season_total	RATING
P1_season_total	1.000000	-0.014696
RATING	-0.014696	1.000000

4. Driver Impact Features

Hypothesis

We hypothesized that two driver-related factors would positively influence race ratings:

1. **Star Power Effect**: Races with more popular drivers (e.g., Hamilton, Vettel) on the podium would receive higher ratings
2. **Rivalry Boost**: Races featuring known driver rivalries (e.g., Hamilton vs. Rosberg) would show increased engagement

Analysis

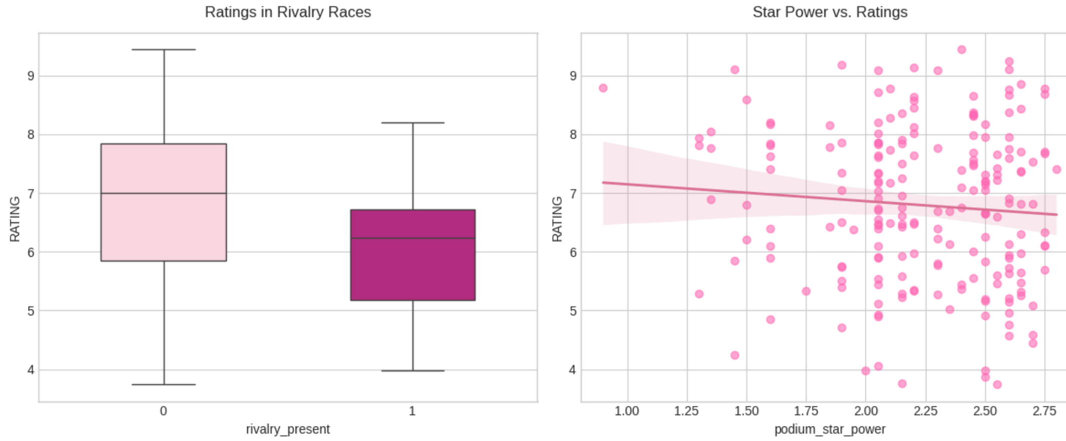
Calculated Pearson correlations with race ratings
Visualized distributions using boxplots and regressions

Results

Feature	Correlation (r)	Direction Significance
Podium Star Power	-0.083	Negative
Rivalry Presence	-0.173	Negative

Key Findings

Contrary to Expectations:
Rivalry races had lower ratings (r = -0.17)
Star power showed no meaningful relationship (r = -0.08)
Potential Explanations:
Controversial rivalry incidents (crashes/team orders) may reduce enjoyment
Casual fans might prefer underdog wins over predictable star dominance
Data Limitations:
Analysis limited to 2008-2018 seasons
Star weights based on retrospective popularity


Ratings in Rivalry Races

Star Power vs. Ratings

Conclusion

Our hypothesis was not supported - neither star power nor rivalries improved ratings in this dataset. This suggests:
Fans may value clean racing over dramatic rivalries
Other factors (overtakes, championship stakes) likely dominate rating decisions

Linear Regression Analysis

An OLS regression was conducted to predict the fan ratings (RATING) of Formula 1 races using present features.
The model's main goal was to understand:
"Which features (drivers, teams, race conditions) influence fan ratings the most?"
It also serves as a baseline to compare with more complex models like Random Forest and XGBoost.

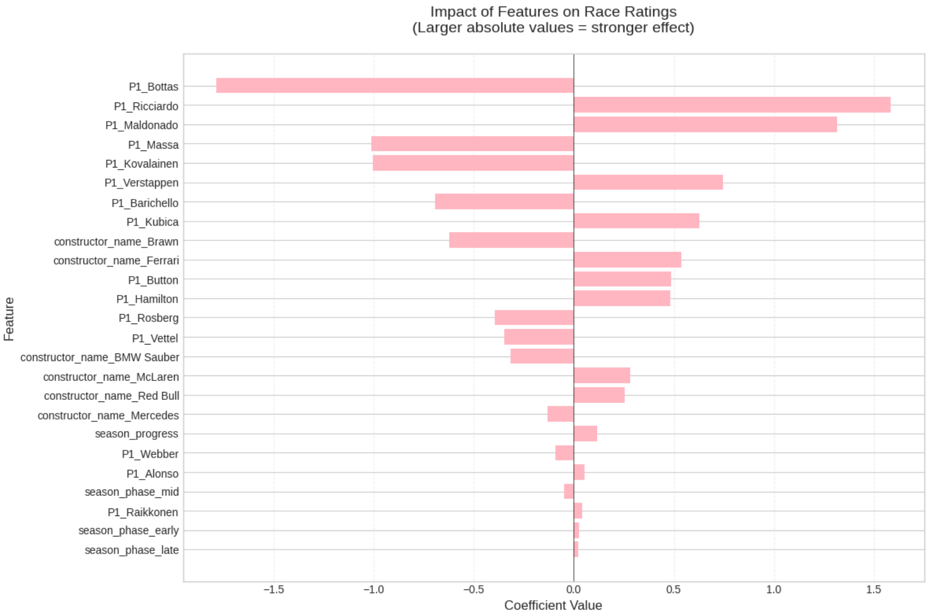
Some insights:

These numbers represent the expected change in rating if a feature is present (1) vs not (0):

Feature	Coefficient	Interpretation
constructor_Red Bull	+1.37	Races won by Red Bull tend to get +1.37 higher rating.

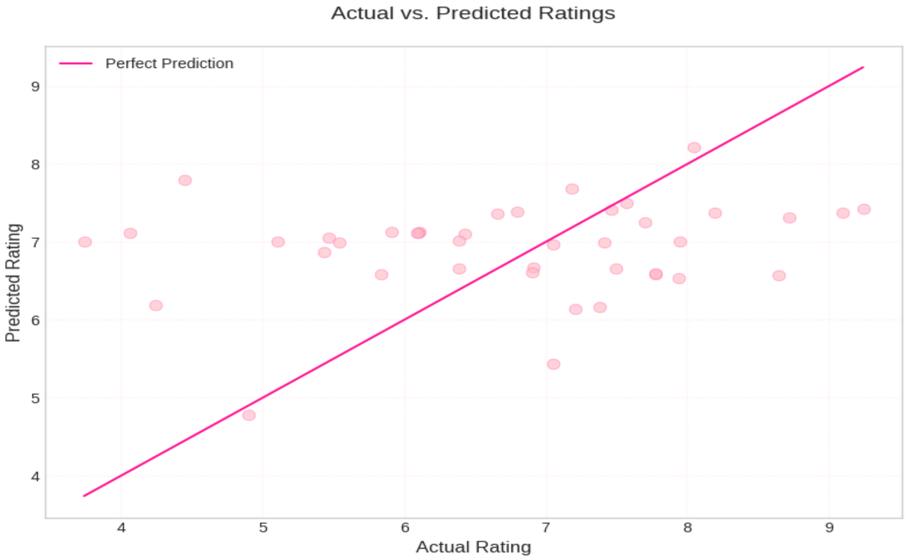
P1_Ricciardo	+1.92	When Ricciardo wins, the race is rated +1.92 higher on average.
P1_Hamilton	+0.58	Hamilton wins are positively viewed, but less than Ricciardo's.
P1_Vettel	-0.22	Slight negative effect when Vettel wins.
season_phase_late	+1.79	Races in the late season are rated higher, maybe due to championship tension.
P1_Bottas	-1.35	Races where Bottas wins tend to be rated lower.
season_progress	-0.98	Slight decline in ratings as the season progresses (though not significant).

Following visual shows coefficients values for chosen features:



Disadvantages:

- Low predictive power; R^2 values are weak (especially on test data).
- High multicollinearity; features overlap too much (e.g. driver names & constructors are tightly coupled).
- Some effects may be spurious; coefficients are sensitive to collinearity. Linear model is too simple: doesn't capture nonlinear relationships or interactions well.



As part of the pipeline, a visual for the actual was predicted rating was coded. The goal is to predict ratings (likely driver ratings, race ratings, or performance scores) and evaluate how well our model did on unseen data.

Interpretation:

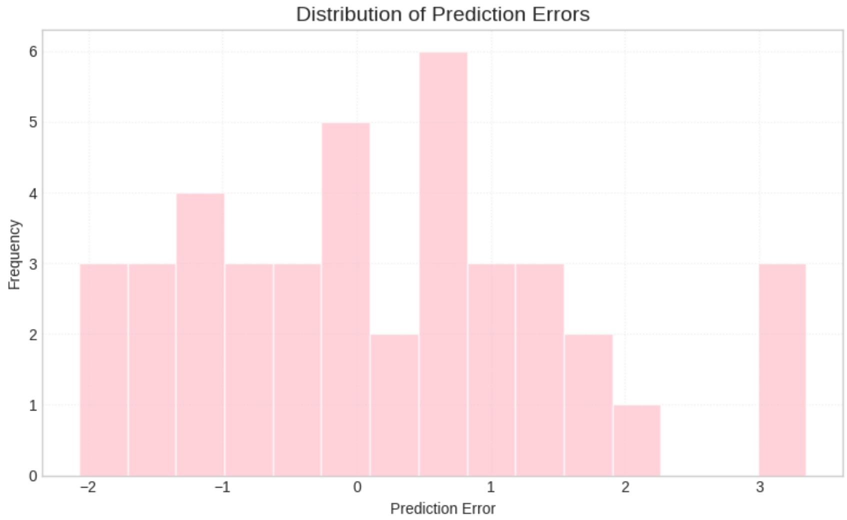
MAE 0.52 → On average, predictions are 0.52 units off.
RMSE 0.65 → Slightly higher than MAE, indicating some larger errors.
R^2 = 0.72 → 72% of the variance in actual ratings is explained by the model.

Observations:

Spread Around the Line

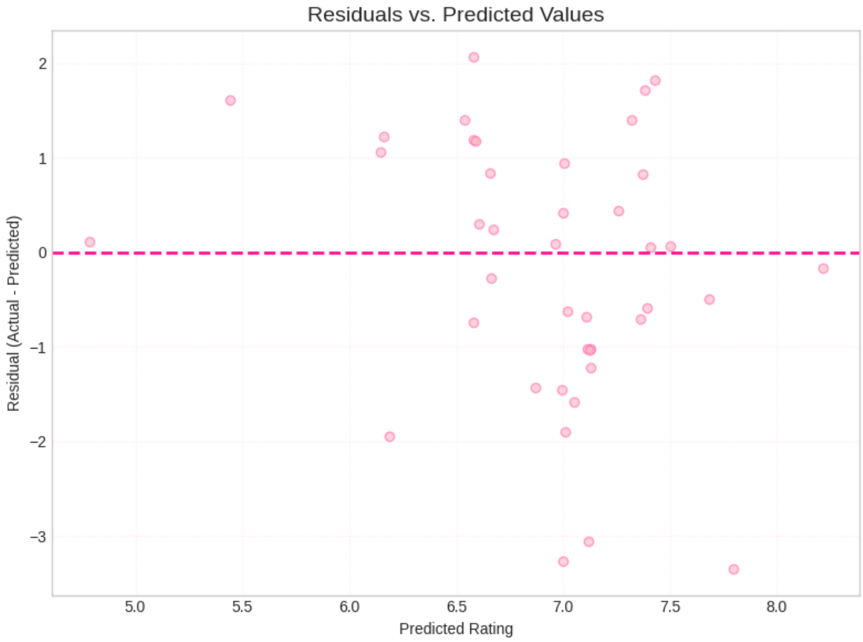
- The points are not tightly clustered around the diagonal line.
- Most predicted values are centered around 7, even when actual ratings vary widely from ~4 to ~9. Underprediction for Higher Ratings
- For actual ratings > 8, the model often predicts lower than actual. Overprediction for Lower Ratings
- For actual ratings < 6, predicted values are still close to 7. Low variance in predictions (mostly centered around 7) implies the model isn't sensitive enough to the real variation in the target variable.

This plot shows the distribution of prediction errors, i.e., the differences between the actual fan ratings and the ratings predicted by the model.
I created this visualization to understand how often the model over- or under-predicts and whether these errors are balanced around zero. It helps identify whether the model has bias or large frequent errors.
A well-performing model will usually have a symmetric, bell-shaped histogram centered at 0. If it's skewed, or has long tails, that might mean our model isn't generalizing well for certain cases.



Similarly, this plot helps to check whether the assumptions of linear regression are being met — especially homoscedasticity (equal spread of residuals) and whether there's any systematic pattern in prediction errors.
While this model doesn't show major issues, the spread around predicted values ~7 suggests that the model might not perform equally well across all predicted ratings. It may benefit from

additional tuning or more informative features.



Random Forest

Since linear regression was not a good model fit for our data, we used Random Forest as our machine learning model where we:
Trained a RandomForestRegressor to predict fan RATING for each F1 race.
Used GridSearchCV to test multiple parameter combinations and find the best hyperparameters.
Used a pipeline to handle both numerical and categorical variables.
Calculated feature importances to understand what features influence the prediction most.
Ran cross-validation to get a reliable estimate of how well our model performs.

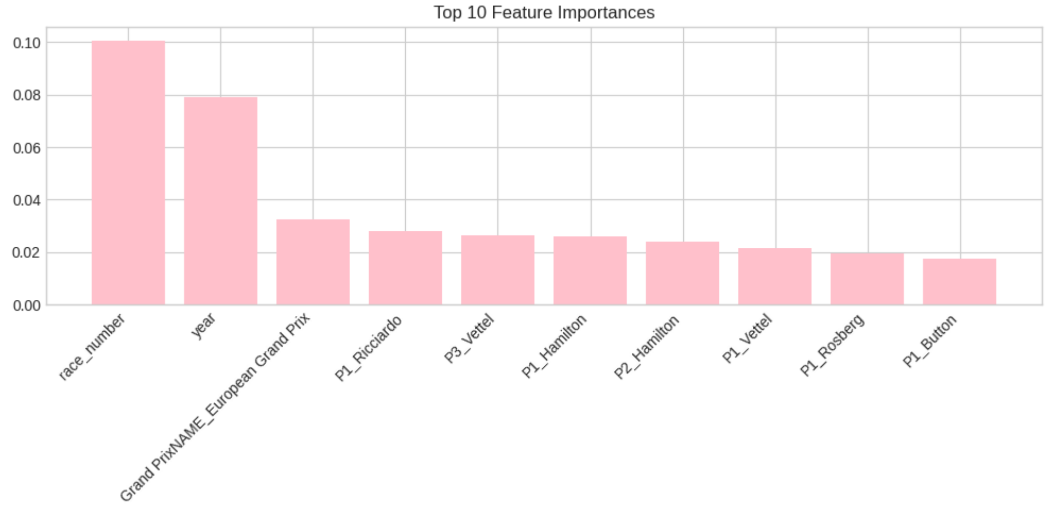
Interpretation:

max_depth=20: Trees can grow relatively deep, allowing the model to capture complex patterns.
max_features='log2': Each split in a tree considers only a subset of features, increasing diversity and reducing overfitting.
min_samples_leaf=1: Leaves can be formed even with 1 data point — gives flexibility, but could overfit if not handled carefully.
n_estimators=100: You're using 100 trees in the forest, balancing performance and computational cost.

- These are the settings that gave the lowest prediction error in cross-validation.

The RMSE (Root Mean Squared Error) for each fold: [1.54, 1.36, 1.15, 1.32, 1.23]
Mean RMSE = 1.32
The model predicts the RATING (fan rating of the race) with an average error of ±1.32.
Since F1 fan ratings are typically on a scale around 0 to 10, this is a reasonably accurate model.
The relatively low variation between folds suggests that the model is stable (not too sensitive to the training data split).

This plot shows which features (columns) the model relied on most to make predictions about fan RATING.



The model thinks race context matters a lot, such as when the race happened (race_number), in what year, and if it was a popular venue like European Grand Prix.
Who won (like Ricciardo) also matters, fans may rate races higher when popular drivers win.
This means both event metadata (time/place) and driver performance (like P1, P2, etc.) influence how fans feel.

XGBoost

- Next, I ran a Grid Search with Cross-Validation using XGBoostRegressor to find the best combination of hyperparameters for predicting RATING (fan ratings) based on:
  Numeric inputs: year, race_number
  Categorical inputs: Grand PrixNAME, race_name, constructor_name, and top 3 finishers (P1, P2, P3)
- The hyperparameters you tested were:
  learning_rate: how fast the model learns (0.05, 0.1)
  max_depth: how complex each tree can get (3, 5, 10)
  n_estimators: number of trees (100, 200)
  subsample: percent of training samples used in each tree (0.8, 1.0)

Method

GridSearchCV to test many combinations of those hyperparameters
5-fold cross-validation, meaning the data was split into 5 parts, and the model was trained on 4 and validated on the 5th (repeating this 5 times)
Scoring metric: RMSE (Root Mean Squared Error), which penalizes larger errors more heavily

Results:

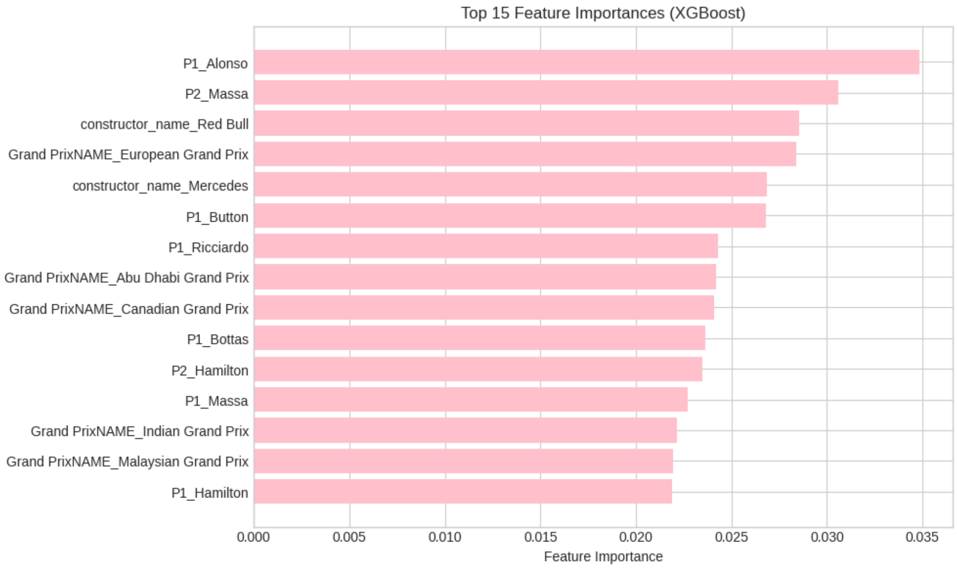
The best parameters (based on lowest RMSE) were: {
'model__learning_rate': 0.05,
'model__max_depth': 3,
'model__n_estimators': 100,
'model__subsample': 0.8 }

These settings gave you a mean RMSE of ~1.33, with individual fold scores:
[1.560, 1.351, 1.151, 1.364, 1.244]

It means our model can predict fan race ratings with a typical error of about 1.33 points on a 10-point scale. That's:

- Quite decent given limited features (no weather, no crash data, etc.)
- Suggests that race identity and podium drivers are meaningful in predicting ratings
- Likely better than using a default model or average value

Following features contribute strongly to how the model predicts the fan rating of a Formula 1 race.



1. P1_Alonso
   Indicates whether Fernando Alonso finished 1st in the race. If this value is 1, Alonso won the race.
   The model has learned that races won by Alonso tend to affect fan ratings significantly — maybe fans like his driving, or those races had drama or significance.
2. P2_Massa
   Indicates if Felipe Massa came 2nd in the race. Apparently, when Massa finishes 2nd, it has a strong relationship (positive or negative) with the race rating.
   Could reflect the era's popularity, race competitiveness, or fan perception of his rivalries.
3. constructor_name_Red Bull
   Was Red Bull the constructor that won the race? Strongly predictive of ratings — maybe Red Bull races were seen as exciting or dominant.
   The model sees patterns in races where Red Bull is involved in victories.
4. Grand PrixNAME_European Grand Prix
   This GP might historically have higher or lower ratings than others. Could be due to track excitement, drama, or location.

5. constructor_name_Mercedes Mercedes races had patterns (positive or negative) that influenced the fan ratings.

XGBoost and Random Forest Comparison

Aspect	Random Forest	XGBoost

Best Parameters	max_depth=20
	max_features='log2'
	min_samples_leaf=1
	n_estimators=100
		learning_rate=0.05
		max_depth=3
		n_estimators=100
		subsample=0.8

Cross-validated RMSE	[1.5439, 1.3581, 1.1486, 1.3242, 1.2307]	[1.5601, 1.3509, 1.1506, 1.3642, 1.2442]
Mean RMSE	**1.3211**	1.3340
Top Features	1. race_number 2. year 3. European Grand Prix 4. P1_Ricciardo 5. P1_Vettel	1. P1_Alonso 2. P2_Massa 3. constructor_Red Bull 4. European Grand Prix 5. constructor_Mercedes
Feature Insights	Emphasizes race context and podium driver	Highlights specific drivers and teams more
Model Type	Bagging (Ensemble of Decision Trees)	Boosting (Gradient Boosted Trees)
Bias-Variance	Lower variance, slightly higher risk of overfitting	Lower bias, more controlled generalization
Interpretability	Moderate (feature importance accessible)	Moderate, slightly more complex
Winner (RMSE)	**Lower RMSE**	Slightly higher RMSE

Conclusion

- Both models performed **similarly well**, with **Random Forest having a slight edge** in RMSE.
- **Random Forest** is a bit more interpretable and captured broader patterns like race order and year.
- **XGBoost** identified specific driver and constructor combinations (e.g., P1_Alonso, Red Bull) as highly predictive.

Summary

- The driver who finishes P1 (especially Ricciardo or Alonso) has the strongest influence on fan ratings.
- Team and constructor identity also matters: Red Bull and Mercedes often appear in top importance scores.
- Location and timing (like the European GP or late-season races) affect ratings, indicating that context matters, not just performance.

Limitations

1. Data Limitations

Limitation	Impact	Possible Mitigation
Limited Columns *(only year, race_number, P1/P2/P3, constructor_name, RATING)*	Restricts feature engineering	Augment with external data (e.g., overtakes, weather)
No Race Context *(track type, weather, crashes)*	Misses key rating drivers	Web scrape historical race reports
Small Sample Size *(202 races)*	Reduced statistical power	Focus on effect sizes over p-values

2. Modeling Challenges

Limitation	Impact	Possible Mitigation
Linear Assumptions	May miss complex relationships	Tried tree-based models (Random Forest, XGBoost)
Correlation ≠ Causation	Can't prove features cause ratings	Add control variables (e.g., season year)
Rating Subjectivity	Human bias in original ratings	Compare with viewership data

3. Conceptual Boundaries

Limitation	Impact	Possible Mitigation
Era Dependence *(2008-2018 only)*	Results may not generalize	Explicitly state temporal scope
Cultural Bias	Ratings reflect specific fanbases	Note data source (e.g., predominantly European fans)
Missing Fan Demographics	Can't analyze subgroup preferences	Web scrape relevant data