

The slide features a light gray background with abstract geometric shapes in blue and dark blue at the top and bottom. On the left and right sides, there are thin black lines that resemble circuit traces, ending in small circles.

Travel Stats

Effects of Travel on Team Performance

Jacky Jiang



Problem Statement

The goal of this project is to determine if the travel schedule of MLB teams affects their performance. Key factors include travel distance, consecutive road games, and rest days, e.t.c.

Data Source



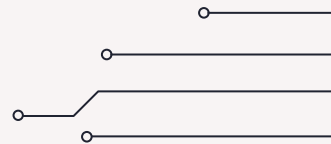
Data Set

The provided data source is a comprehensive dataset of Major League Baseball (MLB) games, comprising 56,775 entries.



Key Features

- Game date
- Teams involved
- Game venue details
- Game type
- Scores
- Performance statistics



Feature Engineering

Location Information

Add geographical coordinates for each MLB game venue to the dataset.

Query **Google Maps API** with the venue and city names, retrieving their geographical coordinates.

Travel Information

The travel distances and time since the last game for each MLB team, both home and away.

Uses the **Haversine** function to compute the travel distance between a team's last game venue and the current game venue.

For each game, record the travel distance and time since the last game for both the home and away teams.

Interactive Features

Create interaction terms and a new variable `travel_dis_diff`

Use the **PolynomialFeatures** object from scikit-learn, which generates interaction terms to capture potential combined effects of different travel distances.

Feature Engineering

Streaks Information

The current winning or losing streaks for both home and away teams.

A positive streak value indicates a winning streak, and a negative value indicates a losing streak.

Aways Information

The number of consecutive days an away team spends on the road.

For each game, the script calculates the number of consecutive days the away team has been on the road. This count is reset at the start of a new road trip or a new year.



Data Cleaning and Wrangling

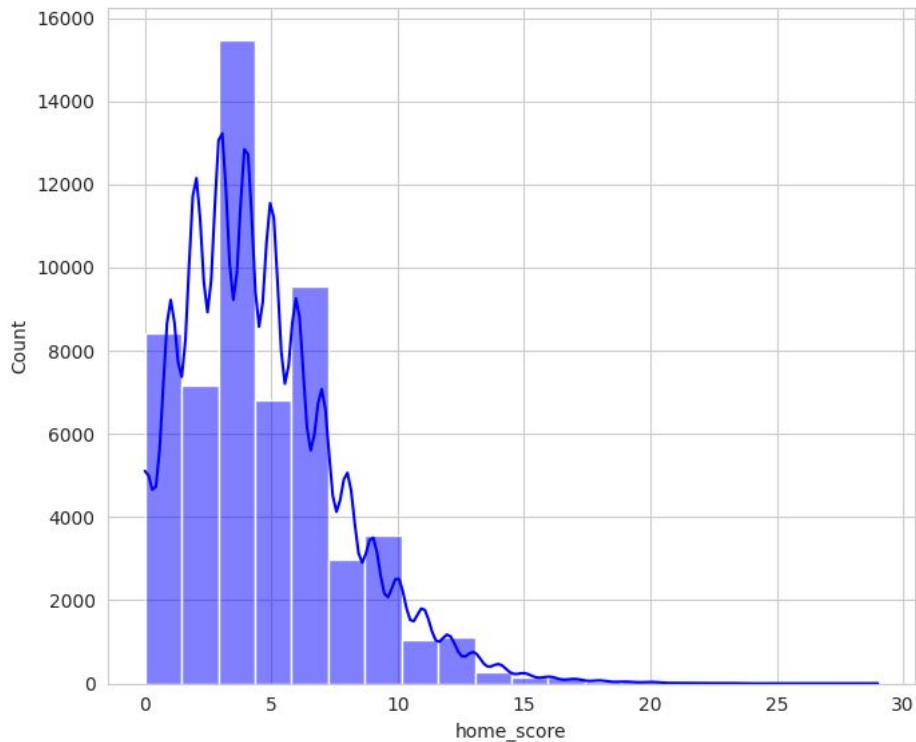
Applied **OneHotEncoder** to transform **categorical variables** like `home_team` and `away_team` are converted into a format suitable for modeling.

Used **StandardScaler** to ensure numerical features contribute equally.

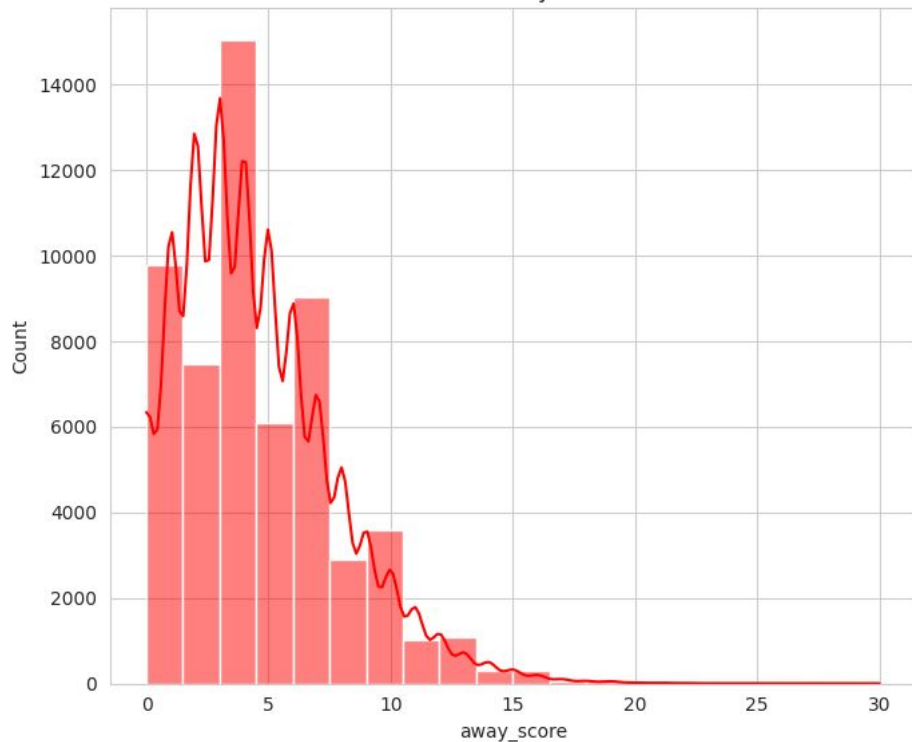
ColumnTransformer facilitated these steps in one go, enhancing our model's ability to learn from both categorical and numerical data effectively.

EDA of Score Distributions

Distribution of Home Team Scores

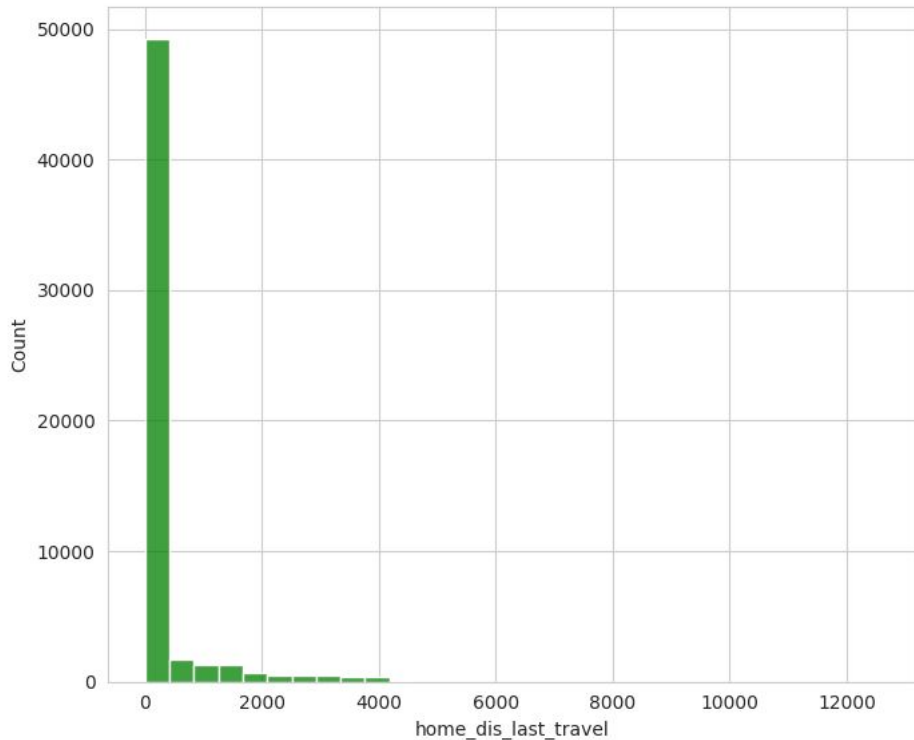


Distribution of Away Team Scores

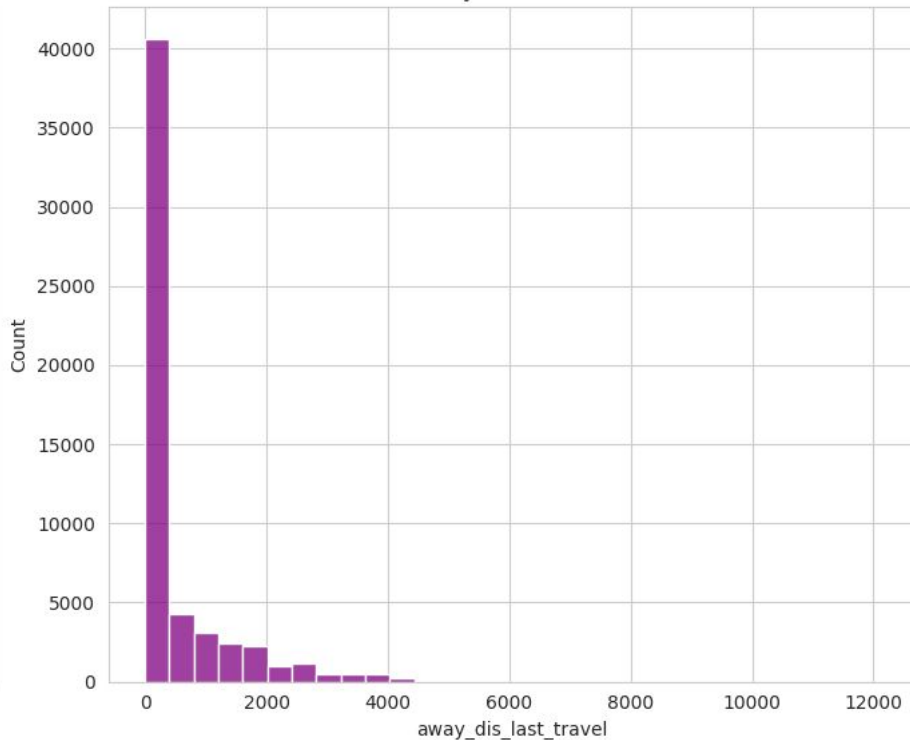


EDA of Home vs. Away Wins

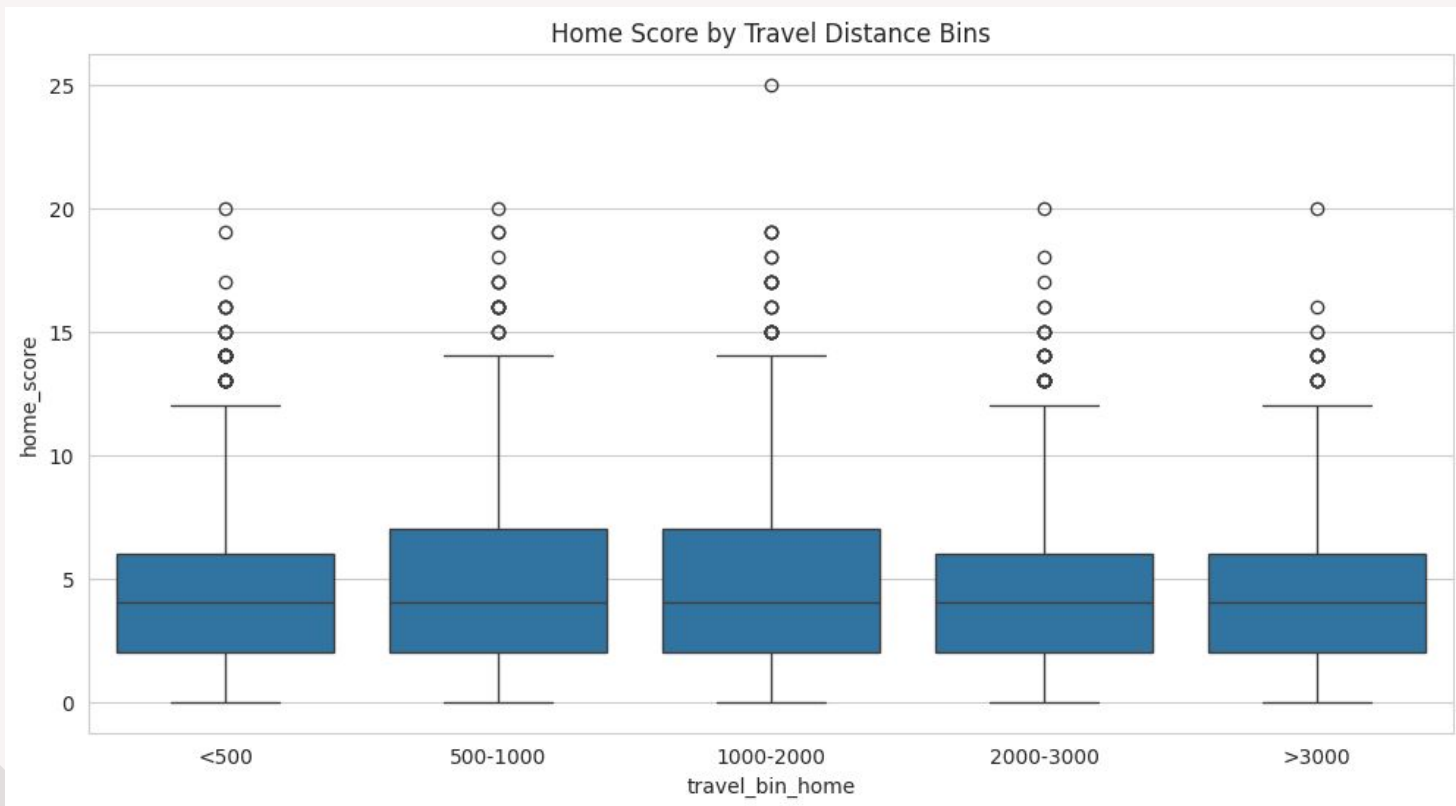
Distribution of Home Team Last Travel Distance



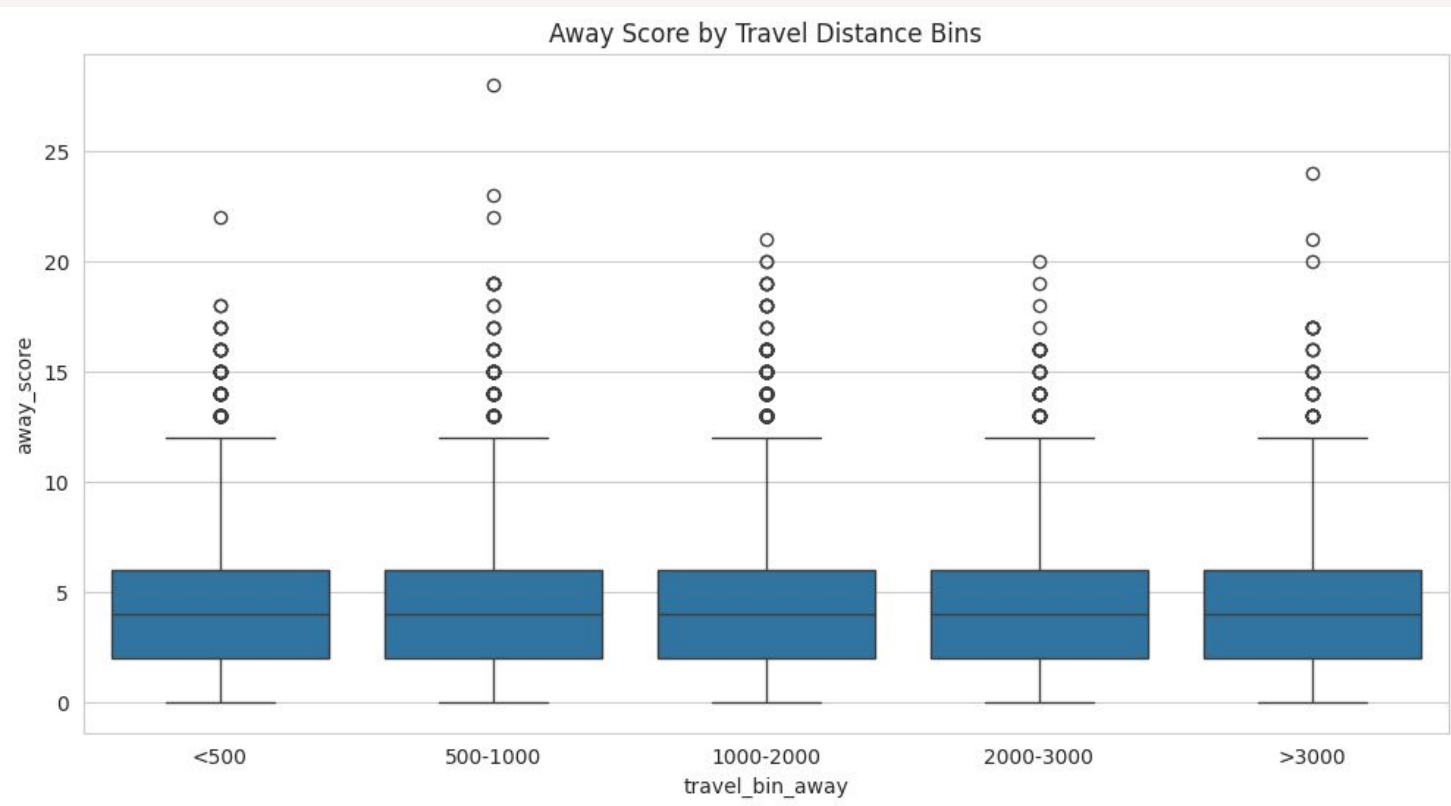
Distribution of Away Team Last Travel Distance



EDA of Travel Distances



EDA of Travel Distances





Choosing the Right Model: XGBoost

We selected the eXtreme Gradient Boosting (XGBoost) model for its high performance in tabular data prediction tasks.

XGBoost is an ensemble learning method renowned for its speed and accuracy, making it ideal for our goal to assess the impact of travel schedules on MLB team performance.



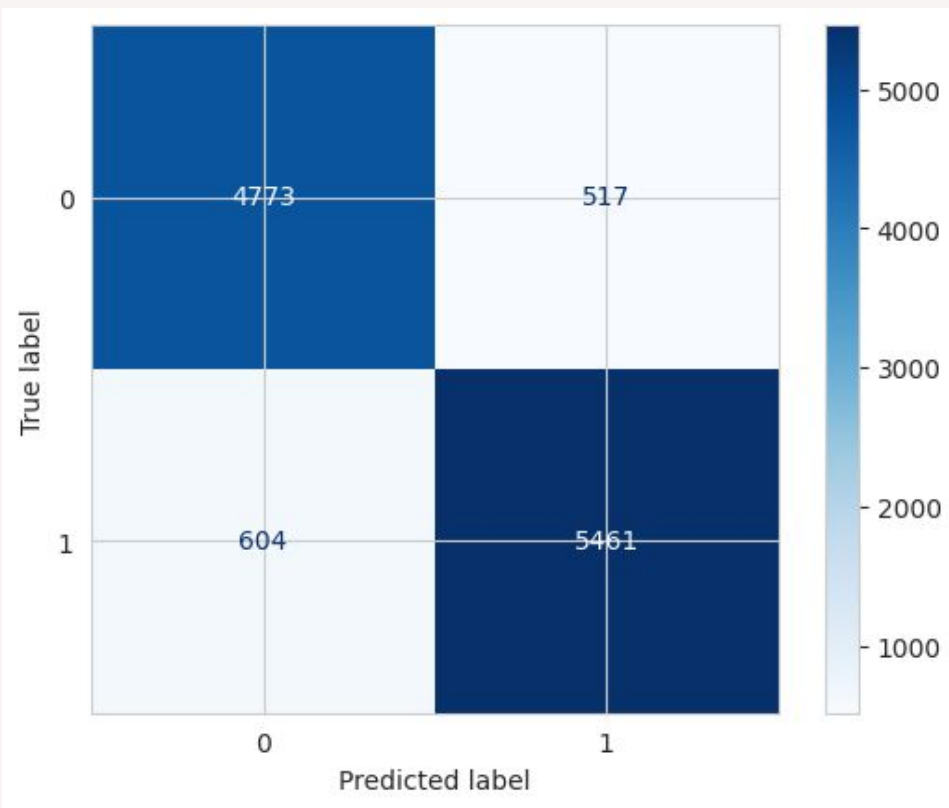
Hyperparameter Tuning with XGBoost

We fine-tuned our XGBoost model using a **RandomizedSearchCV** approach, optimizing over a wide range of hyperparameters to find the best combination.

The parameters included the **number of trees**, **learning rate**, **tree depth**, and **subsampling rates** for rows and columns, tailored to boost our model's predictive accuracy.

The best-performing model achieved a cross-validation accuracy of **89.9%**, with a standard deviation of **0.5%**, indicating robustness in its predictions.

Confusion Matrix





Feature Importance

We analyzed feature importances to understand what drives predictions in our model.

Notably, **in-game statistics** such as bases on balls (home_bb) and doubles (home_2b and away_2b) emerged as top predictors.

Travel-related features, while not at the very top, still showed a significant impact on the model's outcomes, supporting the hypothesis that travel schedules influence team performance.

Feature Importance

Feature Importances with Emphasis on Travel-Related Features

Feature	Importance Score (approx.)
home_bb	0.175
home_2b	0.095
away_2b	0.075
home_hr	0.065
home_pa	0.065
is_day_game	0.065
home_3b	0.060
home_fo	0.055
away_team	0.035
home_1b	0.030
away_pa	0.030
away_1b	0.025
home_hbp	0.020
away_3b	0.018
away_hr	0.018
home_so	0.018
away_dis_last_travel	0.010
travel_dis_diff	0.010
away_fo	0.010
away_time_last_game	0.010
home_dis_last_travel away_dis_last_travel	0.010
away_days_onroad	0.010
away_hbp	0.010
away_so	0.010
away_dis_last_travel home_dis_total_travel	0.010
away_dis_total_travel	0.010
away_bb	0.010
home_time_last_game	0.010
home_dis_last_travel	0.010
home_dis_last_travel away_dis_total_travel	0.010
home_dis_last_travel home_dis_total_travel	0.010
home_dis_total_travel away_dis_total_travel	0.010
away_dis_last_travel away_dis_total_travel	0.010
home_dis_total_travel	0.010
home_team	0.010



Comparing Travel vs. Non-Travel Models

We constructed two models: **one with travel-related features** and **one without**, to directly assess the travel impact.

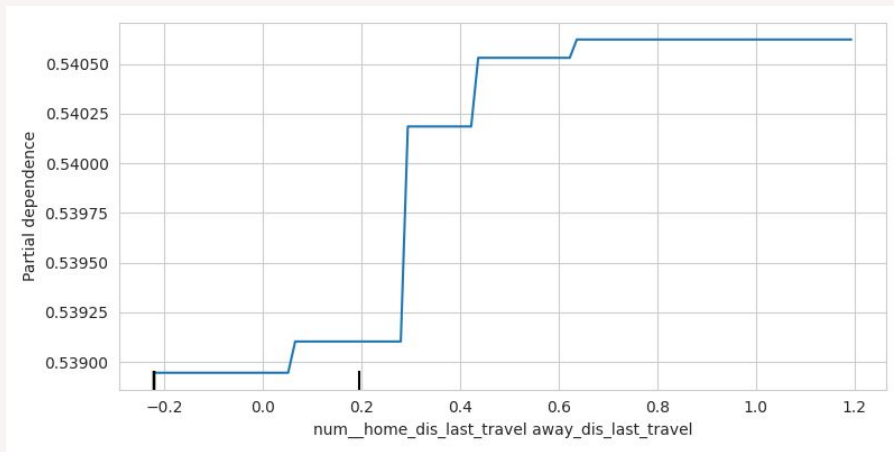
The comparison revealed only a **performance drop** of **15%** when **excluding** travel metrics, a paired t-test yielded a **p-value** of **0.006**.

This statistical test suggests that while travel features are **not the most dominant**, their influence is **significant enough** to affect model performance and, by extension, team performance predictions.

Visualize the Effect of Travel

PDP analysis indicates that certain travel distances correlate with changes in win probability.

Sharp changes in plots like `pdp_num__home_dis_last_travel` `away_dis_last_travel.png` demonstrate that specific travel scenarios can have a notable impact on win probability.






The Verdict on Travel Schedules

Our analysis indicates that travel schedules are a **relevant** but **not overriding** factor of MLB team performance.

The statistically significant result from the paired t-test reinforces that **travel has a measurable albeit modest impact.**



The background features a light gray field with decorative elements in the corners. The top-left and bottom-right corners contain large, overlapping geometric shapes in bright blue and dark blue. The top-right and bottom-left corners feature thin, dark gray lines that mimic the layout of a circuit board, ending in small open circles.

Thanks!