Part 1: Dataset Selection and Justification

Career Relevance

From Kaggle, I chose a baseball data set that contains yearly statistics of MLB teams from 1871 to 2015. This data set was chosen as I have a vested interest in baseball, being on the Richmond baseball team. Additionally, once I graduate, I plan on becoming some sort of analyst in a professional field. Possibly being an analyst for a sports franchise would be an eventual career goal for me if not directly out of school.

Looking at this data can gain possible insights into what makes a team successful over the long term. Obviously, hitting more home runs and striking more people out lead to more wins. However, getting concrete numbers on the precise ranges a team would need to succeed in these variables in order to win more baseball games could be useful. For many professional sports organizations, winning is the key metric to bring in more revenue. More winning generally equates to more ticket and merchandise sales. At the end of the day, for these businesses, their profits are of the utmost importance, so putting a good product on the field is essential.

Data Set Information

URL:
https://www.kaggle.com/datasets/open-source-sports/baseball-databank?resource=download&select=Teams.csv

Creator: Sean Lahman

Data Collection Methodology: Manual data entry, collaboration with researchers, and digitization of historical records.

Limitations: According to Kaggle, the data set is missing a few tables because of a restriction on the number of individual files that can be added.

Data Dictionary(Created by ChatGPT)

Variable Name	Type (Quant/Qual & Continuous/Discrete)	Expected Range / Categories	Business Meaning
yearID	Quantitative – Discrete	1871–2025	The season year that the team's data corresponds to.

lgID	Qualitative – Categorical	e.g., AL, NL, AA, FL	League identifier for the season (American League, National League, etc.).
teamID	Qualitative – Categorical	2–4 letter abbreviation	Unique team code used in the dataset (e.g., BOS, NYY).
franchID	Qualitative – Categorical	Franchise codes	Identifier for the overall franchise across years (e.g., ATL for Braves).
divID	Qualitative – Categorical	E, W, C, or blank	Division within the league (East, West, Central, or missing for early years).
Rank	Quantitative – Discrete	1–10	Team's rank within the division for that season.
G	Quantitative – Discrete	1–162	Total games played in the season.
Ghome	Quantitative – Discrete	0–81	Number of home games played.
W	Quantitative – Discrete	0–162	Total wins during the season.
L	Quantitative – Discrete	0–162	Total losses during the season.
DivWin	Qualitative – Binary	Y / N	Indicates if the team won its division.
WCWin	Qualitative – Binary	Y / N	Indicates if the team won the Wild Card.
LgWin	Qualitative – Binary	Y / N	Indicates if the team won the League Championship.
WSWin	Qualitative – Binary	Y / N	Indicates if the team won the World Series.
R	Quantitative – Discrete	0–1000+	Total runs scored by the team.
AB	Quantitative – Discrete	0–6000+	Total team at-bats.

H	Quantitative – Discrete	0–2000+	Total hits by the team.
2B	Quantitative – Discrete	0–400+	Total doubles hit by the team.
3B	Quantitative – Discrete	0–100+	Total triples hit by the team.
HR	Quantitative – Discrete	0–300+	Total home runs by the team.
BB	Quantitative – Discrete	0–700+	Walks (base on balls) drawn by the team.
SO	Quantitative – Discrete	0–1500+	Strikeouts by the team's hitters.
SB	Quantitative – Discrete	0–300+	Stolen bases recorded by the team.
CS	Quantitative – Discrete	0–150+	Times caught stealing.
HBP	Quantitative – Discrete	0–100+	Batters hit by pitch.
SF	Quantitative – Discrete	0–100+	Sacrifice flies.
RA	Quantitative – Discrete	0–1000+	Runs allowed by the team's pitchers.
ER	Quantitative – Discrete	0–1000+	Earned runs allowed.
ERA	Quantitative – Continuous	0.00–10.00	Team's earned run average.
CG	Quantitative – Discrete	0–100+	Complete games thrown by pitchers.
SHO	Quantitative – Discrete	0–20+	Shutouts thrown.
SV	Quantitative – Discrete	0–60+	Saves recorded by relievers.
IPouts	Quantitative – Discrete	0–4400+	Total outs recorded (each inning = 3 outs).
HA	Quantitative – Discrete	0–2000+	Hits allowed by pitching staff.

HRA	Quantitative – Discrete	0–300+	Home runs allowed.
BBA	Quantitative – Discrete	0–700+	Walks allowed by pitchers.
SOA	Quantitative – Discrete	0–1500+	Strikeouts recorded by pitchers.
E	Quantitative – Discrete	0–200+	Total errors committed by the team.
DP	Quantitative – Discrete	0–200+	Double plays turned by the defense.
FP	Quantitative – Continuous	0.800–1.000	Fielding percentage (measure of defensive efficiency).
name	Qualitative – Categorical	Team names	Full name of the team for that season.
park	Qualitative – Categorical	Stadium names	Primary ballpark where home games were played.
attendance	Quantitative – Discrete	0–5,000,000+	Total home attendance for the season.
BPF	Quantitative – Continuous	~80–120	Batting park factor (effect of park on run scoring).
PPF	Quantitative – Continuous	~80–120	Pitching park factor (effect of park on pitching results).
teamIDBR	Qualitative – Categorical	2–4 letter code	Baseball Reference team identifier.
teamIDlahman45	Qualitative – Categorical	Code	Lahman database team identifier.
teamIDretro	Qualitative – Categorical	Code	Retrosheet team identifier.

Variable	Type	Measurement Level	Expected Range / Categories	Business Meaning

OffEff	Quantitative	Continuous	0 – ∞	Measures how efficiently a team converts hits into runs. High values indicate strong offensive efficiency and effective lineup management.
PitchControl	Quantitative	Continuous	0 – ∞	Strikeouts per walk. Evaluates pitcher control and command. High values suggest disciplined pitching and reduced baserunners.
AttendPerWin	Quantitative	Continuous	0 – ∞	Attendance divided by wins. Represents fan engagement and financial return relative to team success. Higher numbers imply strong fan loyalty or larger markets.
PerformanceTier	Qualitative	Ordinal	{Low Performing, Average, Above Average, Elite}	Categorizes teams based on winning percentage. Used for comparing overall team quality and success level across seasons.

Part 2: Data Quality Assessment and Documentation

Completeness

After reviewing the dataset, several variables were found to contain missing values. The most noticeable were the variables HBP (Hit By Pitch), SF (Sacrifice Flies), and WCWin (Wild Card Win).

Both HBP and SF are missing roughly 83% of records because these statistics were not officially tracked in the early years of professional baseball. Hit-by-pitch data was inconsistently recorded before the 1900s, and sacrifice flies didn't become an official statistic until 1954.

WCWin is also missing for a large portion of records since the Wild Card system wasn't introduced until 1995.

In general, there are many missing records, with over 2000 rows missing at least one. These missing values are not random but reflect the evolution of how baseball statistics were recorded. As a result, analyses involving these variables should be limited to modern seasons or adjusted to account for the historical gaps.

Consistency

When checking for duplicates no exact duplicates were found but a small number of potential duplicates were identified based on the combination of yearID and teamID, especially in older seasons where a single franchise may have changed leagues or names. These do not seem to be true errors, but they can lead to double counting if not reviewed before summarizing data by team or franchise.

In terms of categorical consistency, the data seems to be consistent. The binary columns such as DivWin, WCWin, and WSWin use a mix of uppercase all use capital Y and N with no lower case making analysis easier.

Accuracy and Validity

Outlier detection using the IQR method revealed several numerical anomalies. Variables such as ERA, attendance, and R (runs) contain values that fall far outside the typical range. Most of these outliers appear to be legitimate extreme performances (for example, historically high win totals or ERA values from shortened seasons) rather than data entry errors. Still, their presence means that statistical averages can be skewed if outliers are not addressed.

All quantitative fields are stored correctly as numeric data types. However, a few categorical fields like lgID and divID are stored as strings but could be encoded as categories for more efficient analysis.

Formatting and Data Type Validity

Most data types are correctly assigned. Quantitative variables such as W, L, R, and ERA are stored numerically, while team identifiers and league codes are strings. However, some fields that look numeric (like yearID) might need to be treated as time-based variables for certain types of analysis.

Minor text inconsistencies were also observed in categorical fields such as park, where stadium names occasionally differ by formatting ("South End Grounds I" vs. "South End Grounds 1"). These formatting differences do not affect the dataset's usability but should be standardized before merging with other sources or performing text-based grouping.

Part 3: Comprehensive Data Cleaning

Handling Missing Data

Based on the completeness assessment, the variables HBP, SF, and WCWin had the highest levels of missing data marks due to historical record-keeping gaps. Following Module 3's imputation framework, numeric fields were filled using the median to preserve central tendency, and categorical variables used the mode. For high-missing historical fields, the value "Not Available" was inserted to retain completeness.

I chose imputation over deletion to avoid losing valid records that still contained useful information in other columns. Deletion would have reduced the dataset's size and analytical scope, while mean imputation was rejected because it is more sensitive to outliers and could distort historical data distributions. Keeping the variables instead of removing them entirely allows for partial analysis in the time frames where those stats were actually recorded.

Duplicate Removal

Exact duplicate records were dropped using the drop_duplicates() function. Potential duplicates were minimal and represented overlapping historical franchises. Removing them ensures each (yearID, teamID) pair appears only once.

This method was chosen over fuzzy matching or manual merging since duplicates were exact and limited. Using a simple automated approach preserved data integrity without the risk of false merges.

Data Type Correction

Identifier variables such as teamID, lgID, and divID were converted to uppercase strings for consistency, and numeric columns were coerced to numeric types. This aligns with Module 4's emphasis on ensuring data type validity and categorical uniformity.

This approach was chosen instead of manually redefining each column's data type since automated coercion with errors='coerce' safely converts invalid entries to NaN for further review, reducing the chance of manual input errors.

Outlier Treatment

The Interquartile Range (IQR) method was applied to ERA, attendance, and R to cap values beyond 1.5 × IQR. This approach was selected over deletion to preserve extreme but legitimate performances, reflecting Module 4's principle of "preserve analytical value while minimizing distortion."

I chose IQR over Z-score detection because the dataset includes historical values with non-normal distributions. IQR works better for skewed or bounded data like ERA or attendance, where Z-scores could incorrectly flag valid outliers from early baseball eras as errors.

Text and Category Standardization

Team and park names were converted to Title Case and stripped of whitespace. League and division codes were standardized to uppercase. This ensures uniform categorical representation, consistent with the “Data Standardization” section in Module 4.

Normalization of Continuous Variables

Continuous features (ERA, attendance, and R) were normalized using Min-Max scaling (range 0–1). This step was drawn directly from Module 4’s discussion on scaling and normalization to improve feature comparability and support downstream statistical or machine-learning analyses.

Min-Max scaling was chosen over standardization (Z-score) because it maintains the original distribution’s shape and confines all values to a consistent range, which is more appropriate for interpretability and comparison between metrics on different scales (e.g., runs vs. attendance).

Part 4: Strategic Feature Engineering

Variable 1 – Offensive Efficiency (OffEff)

The OffEff variable measures how efficiently a team converts hits into runs, calculated as runs divided by hits (R ÷ H). It reflects a team’s ability to maximize offensive opportunities and produce runs without relying solely on volume. From a business standpoint, this provides valuable insight into roster construction, helping analysts determine whether a team’s scoring success stems from player skill or sheer frequency of at-bats. This variable was chosen over runs per game because OffEff normalizes output by hit volume, allowing for fairer comparisons across seasons with different offensive environments.

Variable 2 – Pitching Control Ratio (PitchControl)

The PitchControl variable evaluates a pitching staff’s command through strikeouts per walk (SO ÷ BB). A higher value indicates stronger control and discipline on the mound, both of which correlate with long-term success and lower earned run averages. This metric gives front offices a cleaner look at pitching efficiency independent of team defense or park factors. It was selected over metrics such as ERA or WHIP because those depend on fielding and ballpark effects, while

PitchControl isolates a fundamental skill that teams can directly target when recruiting or developing pitchers.

Variable 3 – Attendance per Win (AttendPerWin)

The AttendPerWin variable connects athletic performance to business outcomes by dividing fan attendance by total wins (Attendance ÷ Wins). This measures how much fan turnout corresponds to team success, offering insight into market engagement and financial return on winning seasons. This variable is particularly valuable for front-office executives and marketing teams to forecast revenue potential and assess the impact of performance on fan loyalty. It was chosen instead of total attendance because AttendPerWin adjusts for season length and team quality, isolating the business value of each win rather than aggregate volume.

Variable 4 – Team Performance Tier (PerformanceTier)

The PerformanceTier variable groups teams into four categories based on winning percentage: Low Performing, Average, Above Average, and Elite. This variable was created using binning, which converts continuous data into interpretable categories. By categorizing teams rather than analyzing win percentages as raw numbers, the data becomes easier to visualize and discuss in strategic contexts. This transformation is especially useful for historical comparisons, allowing decision-makers to assess how often teams reach “elite” levels over decades. Binning was chosen over continuous measures because executives and analysts typically evaluate team performance in tiers or milestones (e.g., playoff-caliber vs. rebuilding) rather than by precise decimal values.

Part 5: Thorough Exploratory Data Analysis

Explanations and analysis of visualizations can be seen in the output of the code.

Part 6: Business Insights Report

Executive Summary

This analysis examined Major League Baseball team performance data from 1871 to 2015 to uncover factors that strongly drive winning and fan engagement. Analysis of the dataset revealed consistent patterns linking offensive efficiency, pitching control, and team success. Beyond performance outcomes, the analysis also explored business-oriented variables such as attendance, providing a holistic view of both on-field and financial performance.

The findings confirm that offensive efficiency, measured as runs per hit, is one of the most reliable indicators of team success. Teams with higher conversion rates between hits and runs

tend to win more consistently, reflecting the importance of extra base hits in lineup construction. Additionally, strong pitching control, defined as strikeouts relative to walks, shows a negative relationship with earned run average (ERA), indicating that pitchers who limit free passes and maintain command effectively reduce scoring by opponents.

Attendance data show that successful teams attract larger fan bases, linking athletic performance directly to business outcomes. Historical trend analysis also highlights distinct eras in offensive production, from the low-scoring dead-ball era to the surge of the late 1990s steroid era, emphasizing how external factors like rule changes and player conditioning shape long-term performance metrics.

For baseball executives or sports analytics professionals, these insights reinforce the strategic value of balancing lineup efficiency with disciplined pitching. Efficient run production and control-oriented pitching staffs not only translate to wins but also drive fan interest and revenue potential. The historical perspective provides context for evaluating modern teams, helping decision-makers benchmark success across eras while identifying the variables most predictive of sustained competitiveness.

Detailed Findings

The strongest and most consistent relationship in the dataset lies between offensive efficiency and win percentage. Scatterplot analysis revealed a clear positive correlation: as teams convert a greater percentage of their hits into runs, their likelihood of winning rises sharply. This suggests that optimizing quality at-bats and situational hitting may yield higher returns than raw hit totals. From a roster-construction standpoint, general managers should prioritize players with high slugging and on-base efficiency rather than merely high batting averages. The box-and-whisker analysis by performance tier reinforces this finding: elite teams exhibit higher median offensive efficiency than all other tiers, with minimal overlap between lower and upper quartiles.

Offensive efficiency also reflects managerial and strategic decision-making, such as effective lineup sequencing, base-running aggressiveness, and situational awareness. Teams with comparable hit totals but more runs scored likely demonstrate stronger coaching execution and advanced analytics integration. These findings align with sabermetric principles popularized by Moneyball, affirming that the ability to “turn contact into scoring” remains one of the most valuable assets in baseball.

The analysis of pitching control (strikeouts per walk) against ERA underscores the importance of disciplined pitching staffs. The negative correlation found confirms that as control improves, run prevention strengthens. This aligns with established performance metrics like Fielding Independent Pitching (FIP), which isolates a pitcher’s control of outcomes. Teams that develop or acquire pitchers with superior command reduce variability in defensive dependency and improve overall consistency across long seasons.

In practical terms, pitching analytics teams could use these findings to inform scouting models , emphasizing walk rate reduction and first-pitch strike percentage as leading indicators of success. Additionally, coaching programs can prioritize mechanical efficiency and situational pitching strategies to lower ERA. Over long periods, even small improvements in control metrics compound significantly across innings pitched.

While the correlation between attendance and win percentage is not perfectly linear, the broader pattern is unmistakable: success breeds fan engagement. Teams in higher performance tiers consistently record greater average attendance, demonstrating a direct link between athletic outcomes and financial performance. For front-office executives, this insight translates into a clear business case for investment in player development and performance analytics, both of which contribute to sustained winning and, consequently, revenue stability.

The historical correlation matrix strengthens this point, showing that offensive and pitching performance variables not only drive wins but also correspond with higher attendance figures. In modern baseball, where digital viewership and merchandising complement stadium sales, these insights can guide marketing and operations teams to align promotional efforts with peak performance windows.

The line chart of league-wide offensive efficiency over time provides valuable historical perspective. Offensive output has fluctuated dramatically, with visible troughs during the early “dead-ball” years and peaks during the 1990s and early 2000s. Such variation can be attributed to evolving player training, rule modifications, equipment improvements, and external factors such as the steroid era. For modern analysts, contextualizing current team statistics within historical baselines prevents over- or under-valuation of player performance. A .750 team OPS in 1975, for instance, carries a different relative meaning than the same metric in 2000.

For researchers and historians, this reinforces the importance of era-adjusted metrics like OPS+ or ERA+, which normalize performance relative to league averages. For team analysts, these era patterns serve as benchmarks when forecasting future offensive trends, particularly amid discussions of rule changes or pitch-clock enforcement affecting offensive balance.

Recommendations

1. **Prioritize Offensive Efficiency Metrics in Scouting and Player Evaluation.**
   Player acquisition and development should focus on improving run-conversion rates. Metrics like weighted runs created (wRC+), on-base plus slugging (OPS), and base-running efficiency provide deeper insights than traditional hit totals.

2. **Invest in Pitching Command Analytics.**
   Organizations should allocate resources to refine strike-to-walk ratios through biomechanical analysis, video-based feedback, and data-driven training. These

investments will reduce ERA variance and enhance overall staff reliability.

3. **Integrate Performance Analytics with Fan Engagement Strategy.**
   Marketing departments should coordinate promotional campaigns around strong performance stretches to capitalize on attendance surges. Analytics teams can model attendance elasticity relative to win streaks to optimize pricing and scheduling.

4. **Adopt Era-Adjusted Performance Contexts for Decision-Making.**
   Front-office analysts should incorporate historical baselines into modern evaluations, ensuring player comparisons and projections remain valid across different offensive environments.

By uniting performance analytics with business operations, teams can achieve both competitive and financial sustainability. Data-driven decisions in player development and fan engagement will continue to define the next generation of successful franchises.

Limitations and Next Steps

While the dataset provides broad historical coverage, several limitations constrain the scope of inference. First, the time span from 1871 to 2015 includes periods with inconsistent record-keeping and differing league structures, which may affect comparability across eras. Variables like hit-by-pitch and sacrifice flies were missing in early seasons, limiting complete statistical alignment. Additionally, the analysis is correlational rather than causal, while offensive efficiency and pitching control relate strongly to success, underlying factors such as managerial strategy, player health, and opponent quality remain unobserved.

The analysis done here was fairly baseline. For the future, it would be beneficial to dig deeper into more modern statistical data points not present in the data set. Advanced sabermetric indicators (e.g., wOBA, BABIP, FIP) can add even more value to the analysis. Expanding the business analysis dimension, future studies could integrate ticket pricing, regional demographics, and media revenue data to estimate the true financial elasticity of performance. Such extensions would strengthen both analytical precision and business relevance for real-world decision-making.