جامعة أم القرى
UMM AL-QURA UNIVERSITY

Applied Statistics for Data Science

SUBMITTED BY:
Nawaf Abdulrhman Alageel
Mohammed Khalid Altufayhi
Abdullah Mansour Habit
Muhammad Saad Al-Sarhani

DEPARTMENT OF (Data Science) COLLEGE OF COMPUTERS UMM AL-QURA UNIVERSITY

2023

Comprehensive Analysis Report on Baseball Data

Source: Kaggle - Baseball Databank

INTRODUCTION:-

Baseball, often referred to as America's pastime, is more than just a sport; it's a chronicle of history, strategy, and evolution. The Baseball Databank encapsulates this narrative, providing a meticulously curated archive of historical baseball data for enthusiasts, analysts, and researchers alike. Sourced from Sean Lahman's website, this databank offers a structured, comprehensive view of baseball's evolution from its early days to the modern era.

While the initial version (v1) of the databank may lack some tables due to file size restrictions, it is by no means incomplete. Future updates are set to include crucial data points like Parks, HomeGames, CollegePlaying, Schools, Appearances, and FieldingPost.

The databank adopts a unique player-centric approach. Each player is assigned a distinct playerID, serving as a link to all their associated information. The MASTER table serves as the foundation of this databank, capturing player names, birthdates, and other biographical details.

Key Tables Include:

MASTER: Player names, birthdates, and biographical data.
Batting: Detailed batting statistics.
Pitching: In-depth pitching records.
Fielding: Defensive statistics.
Additionally, the databank offers supplementary tables providing insights into All-Star appearances, Hall of Fame voting, managerial statistics, team standings, post-season performance, franchise history, and much more.

In essence, the Baseball Databank transcends being just a dataset. It stands as a testament to baseball's history, strategies, and iconic personalities. Systematically organized for exploration and analysis, it caters to fans, statistical enthusiasts, and researchers. This databank not only captures numbers but also the heart and soul of baseball, ensuring that the legacy of America's beloved sport is preserved for generations to come.

Baseball Databank Analysis Report

1. Dataset Acquisition
Datasets AllstarFull.csv and Salaries (1).csv were successfully loaded. Initial checks for missing values in the salaries dataset revealed:

2. Data Preprocessing and Normalization
The data was cleaned by addressing missing values and outliers. For instance, missing gameID entries were removed, and missing values in the GP column were replaced with the median. Outliers in the GP column were identified and managed using the Interquartile Range (IQR) method.

3.Descriptive Statistics

The mean salary represents the average salary of the players in the dataset.
The median salary indicates the middle value when salaries are arranged in ascending order. It helps in understanding the central tendency, especially when the distribution is skewed.
Standard deviation and variance for salary provide insights into the spread or dispersion of the salary data.
The measures for Games Played (GP) follow the same logic. The mean and median provide insights into the average and central tendency of games played by the players, respectively. The standard deviation and variance give an idea about the dispersion in the number of games played.

4. Merging Data
The All-Star and Salaries datasets were merged using the playerID and yearID columns, providing a unified dataset for analysis.

5. Data Exploration and Visualization
Preliminary data exploration was conducted to gain insights:

A preview of the merged data showed key statistics of the first few rows.
A histogram visualized the distribution of games played (GP).
Boxplots depicted the spread of salaries, emphasizing the median and potential outliers.
A scatter plot illustrated the relationship between games played (GP) and salary.
5. Advanced Analysis

Linear Regression: A linear regression model was developed to understand the influence of games played (GP) and startingPos on a player's salary. The model summary can provide insights into significant predictors.

ANOVA Test: An Analysis of Variance (ANOVA) was performed to test for salary differences across different starting positions.

6. Extended Data Analysis

Player Frequency: A list of the top 10 most frequent players in the All-Star dataset was compiled.

Salary Distribution: A time-series line plot showcased the average salary distribution across different years.

Correlation Analysis: Correlations between numeric variables were computed, providing insights into potential relationships or patterns.

Grouped Analysis: The average salaries for different starting positions were calculated.

7. Further Data Analysis

Top Earners: A list of the top 10 players with the highest average salaries was presented.

Performance vs. Salary: A scatter plot illustrated the relationship between a player's average performance (in terms of games played) and their average salary.

Player Longevity: A table showcased players with the longest spans in the dataset.

Position Analysis: Salaries were analyzed based on starting positions, providing insights into which positions tend to earn more on average.

CONCLUSION

The Baseball Databank is an expansive reservoir of baseball history. Through this analysis, we unearthed insights, trends, and patterns, shedding light on the intricacies of the sport. From understanding player salaries and performances to diving deep into the sport's history, this exploration offered a holistic view of baseball's legacy. As with any sport, the numbers only tell part of the story. Yet, through careful analysis, they can provide a fascinating perspective on the game's evolution and its legends.