# **Project Report: Investigating a Baseball Dataset**

The analysis investigated the **baseball data set**. The dataset contains batting and pitching statistics from 1871 to 2014, and all the supplementary files including fielding, teams data, and the master data with personal player records.

A baseball game is played between two teams, each composed of nine players, that take turns playing offense (batting and baserunning) and defence (pitching and fielding). The goal of the game is to score more points (runs) than the other team. Pandas and Numpy will be the tools to use in conducting the analysis.

This investigation explored the following **questions**:

1. Which were the 3 best teams in the 2rd millenium era(2001-todate)
2. What are the above average features that contribute to Wins? (assumption here is that Wins is equivalent to Winning)

**Process:** To start the exercise, the tools were imported into the jupyter working environment. Since the exercise required us to use numpy and pandas packages, these were imported. For visualisation matplotlib.pyplot was imported as plt, and seaborn as sns. % matplotlib inline was used within the cell environment for displaying plots in the document.

As part of the data wrangling phase, the data was loaded into the working environment using the Pandas library read.csv() function method which allows the URLs of the data files to be passed in. A .csv type was used because the data file were in a comma separated format.

With the tools and data available, the data **wrangling phase** started. Data wrangling is a process that cleans or pre-processes data before its explored and analysed. This step involves understand the data and the subject of the data - in this case the baseball game. Looking into sources like Wikipedia, <https://en.wikipedia.org/wiki/Baseball> and discussions with subject matter experts is important here.

The major files of the dataset like Batting, Pitching, fielding, Master and Team files were visually inspected using the head(), tail() or .sample() functions of the Pandas library to provide an overview of the contents of the files and to detect errors or data problems like name errors, missing data, outliers, duplicates, untidy data, and wrong dtypes among others. The shape and column attributes were used too. Preliminary investigation showed that the data sets had different shapes and different contents although they shared some variables. Inspection also showed that the data had missing values and that the disparate files needed to be combined for better analysis of the data.

**Summary statistics** were calculated using .value\_counts() for the non-categorical values. This shown the proportion of the Player's batting hand over the years. 6.6% of the players play with both the left and right hand in batting, while throwing happens only with either the right or left hand.

Summary stats of the numeric variables revealed that the mean weight of the players is 186 pounds compared to a min of 65 and a max of 320. This high variance also shown by other variables in the datasets. In the Batting file H(hits) has a mean of 37 with a max of 262, RBI has a mean of 17 with a max of 191. R, SB, IPouts are some of the other variables with high variance, indicating the presence of outliers in the dataset. Exploratory data analysis looked at summary stats using .describe() and the .corr() method on the dataframe. Based on this IPouts, H\_x Hits, Earned runs ER, Walks BB\_x, Strikeouts SO were seen to be strongly correlated with Wins, W ie 0.93, 0.89, 0.81, 0.79, 0.82.

The exercise identified missing data, separate files and outliers as the problem issues in the datasets. The main data sets Pitching, batting and Masters were combined. Missing values were isolated using the .isnull() method. The intention was to handle the missing values by dropping them or filling them with the mean or with zeros depending on whether they were more than 7% or 5% or less than 5%. Based on this the mean was used to fill in missing values for weight and height. Bats and throws were filled with 0. And other columns with bigger values of missing data were dropped.

Data visualisation was done with a line plot and scatter Plots using Pandas and seaborn pairplot. The line plot had no clear relation. The pair plots showed clear linear relationship with wins is IPouts a Pitching variable, followed by AB in Batting and Hits. RBI though dropped earlier due to missing values also has a linear relation as shown by the correlation method.

## **Analysis of Questions**

**1. List the 3 best teams in the 2rd millenium era(2001-todate)**

The teams dataset was sorted by Rank and then grouped by rank using the groupby method. The .value\_counts() method was then called on the series object to find the best teams.

New York Yankees and st. Louis Cardinals managed to reach position 1 eight times during this period and Los Angeles Dodgers have done so seven times.

**2. What are the above average features that contribute to Wins? (assumption here is that Wins is equivalent to Winning)**

Correlation, scatter plots and boxplots were used to analyse this question. The exercise found that the above average values of IPouts, AB and Hits will improve winnings and so will players below a height of 75 inches and 225 pounds.

## **Conclusion**

Investigating the Baseball dataset has been a pleasant challenge. The exercise involved importing the necessary tools, loading data into the working environment, understanding through investigation, identifying errors, cleaning and making the data tidy.This was followed with exploration and visualisation before finally analysing the questions.

The limitations to this exercise was lack of subject knowledge to clarify the contents of the datasets, information to understand why the observations had missing values. There was not edplanation about the trends and patterns in the missing data. Therefor variables like SH, SF, GIDP, IBB, BAopp with huge missing values like 32924, 13510, 32902, 14575, and 1534 respectively were dropped. The the missing values in weight and height were filled out with the mean values while 0 was used to fill out the missing values in bats and throws. The presence of outliers in most of the variables also could not be explained. thus inevitably affecting the results of the exercise.

Question 1 was to were to find out the best 3 teams in the millenium era. The exercises finds the New York Yankees and Louis Cardinals and Los Angeles Dodgers the best teams in the millenium period.

Question 2 was to find out "What the above average features are that contribute to Winnings? Based on the analysis which was is actually fully conclusive, the above average values of IPouts, AB and Hits will improve winnings and so will players below a height of 75 inches and 225 pounds.

Not fully conclusive means that this exercise could go further to carry out a hypothesis test to confirm or deny the claim that IPouts in the range of 650-500, AB values in the range of 72-45, hits in the range of 12-6, a weight of 225 pounds and a height of 75 inches are vital in determining winnings. Such test are not definative in their conclusions but can simply emphasize the research hypothesis that optimisizing the factors this analysis has found improves the result to a certain degree as compared to the status quo of the null hypothesis.