Major League Baseball (MLB) has a problem: they spend too much money on players because of a historical flawed assumption that buying (i.e., offering contracts) players who are perceived "better" will buy the team more wins. Proven by the Oakland A’s and shown in the movie Moneyball, buying runs (i.e., exclusively looking at the players who have historically made a high number of runs in comparison to other plays) will buy the team wins. With this background in mind, this paper looks to answer the question: which Major League Baseball team stats impacts team wins the most?

To answer this, a 2016-2020 dataset of MLB Team Batting statistics was analyzed. No outliers were found because over a 5 year span, because no team will be truly bad or truly good (outliers would only be seen in single 1 season). A PMF comparing number of home runs in the MLB American and National Leagues showed that there is very little difference in the number of home runs between the American and National Leagues. A CDF of Team Wins showed the mode (374) is clearly visible, about 10% of the Teams have less than 300 wins, about 10% have more than 420 wins, and about 90% have less than 420 wins over 5 years. The CDF could not answer the term paper question as there needs to be a comparison between Wins to another variable in the dataset. The two analytical distribution plots (Normal Distribution, Normal Probability Plot) showed that the dataset is normally distributed which is to be expected as this was a large dataset (Central Limit Theorem). In looking at the scatterplots, the relationship between Wins and Bases on Balls/Walks is linear. Pearson's correlation was calculated at 0.86, which means there's a positive correlation between Team Wins and Bases on Balls/Walks. In terms of causation, BB (Bases on Balls/Walks) is a way to get runs, and runs leads to more Team Wins. A hypothesis test of the difference in Wins between the American and National Leagues, showed a p-value of 0.877 or 87.7% was found which "means that we expect to see a difference as big as the observed effect about 87.7% of the time. So, this effect is not statistically significant" (Downey, 2021). 87% is bigger than 5%, which is the threshold of statistical significance, so this effect is not statistically significant. A single (Wins to BB) and multiple (Wins to BB and Doubles) regression analysis resulted in an R2 increase from 0.75 to 0.82 respectively, suggesting that the apparent difference in Team Wins being may be explained by the number of Bases on Walks/Balls and Doubles (strong correlation).

In terms of what was missed during this analysis, outliers were not found as the dataset is over a 5 year period (2016-2020). This was also a challenge, and to overcome it I learned how to make a boxplot in Python to validate that there were no real outliers in Team Wins. In a subsequent analysis, I would look at 1 current season to find outliers in the data, even if that means there is less data to analyze. Pitching statistics were not included in this dataset (only batting statistics), and they could have added to determine which of all the baseball stat categories improved team wins. I assumed normal distribution in the dataset which was later verified, however I did not have any assumption that were found incorrect after the analysis. In the end, I believe that a logistic regression model would have been a better fit and I would look to understand how to do that in a future analysis.

**References**

* 1. Downey, A. (2011). Think stats: Probability and statistics for programmers. O'Reilly.