# Data

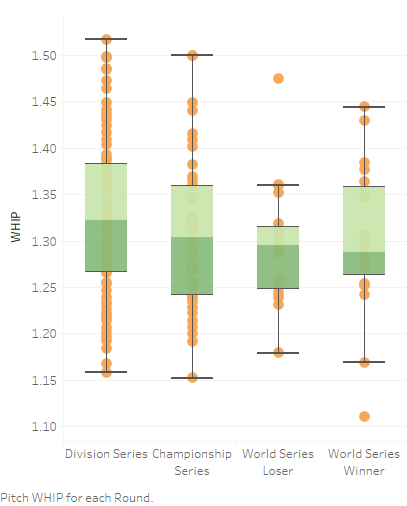
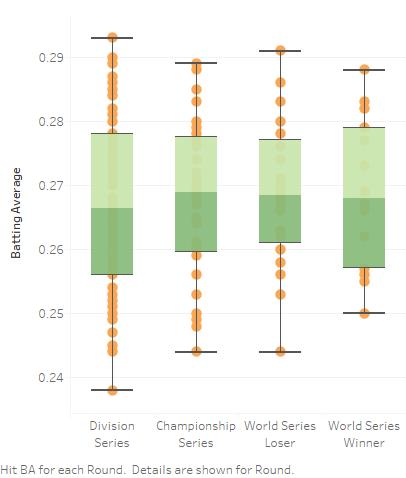
Data for our project consisted of team statistics for each of the eight playoff teams from the 1995 – 2016 major league baseball seasons. Due to a significant rule change after the 1994 season we decided to include seasons after 1995 for consistency. This resulted in 176 rows of data and 63 columns. Two additional columns were manually created for our dataset: 1) a binary column of 1’s and 0’s indicating if a team won a world series in a particular season 2) a categorical column representing how far a team progressed in the playoffs.

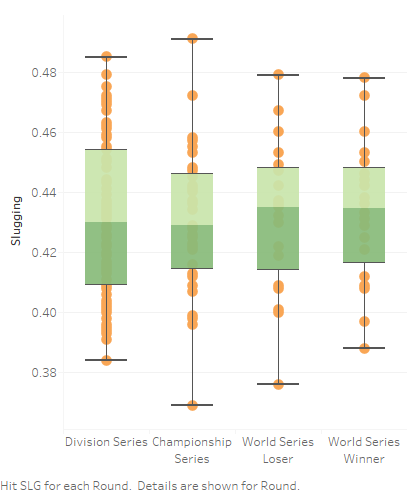
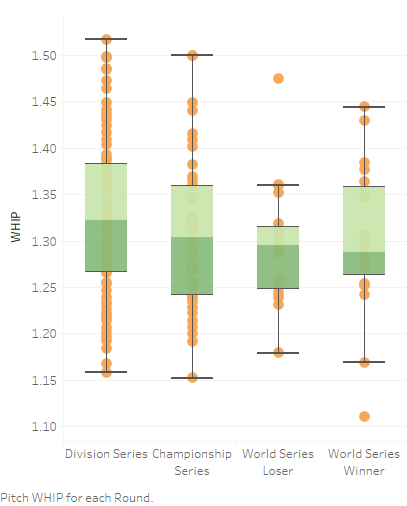
All data were gathered from Baseball-reference.com; a free website which hosts baseball statistics dating back to the 1800’s. Baseball-reference.com is designed for easy data export. All data are stored in downloadable CSV format. CSV files were downloaded for all teams from the 1995 – 2016 seasons and filtered to only include teams which made the playoffs.

Baseball statistics are largely split into two main categories: pitching and hitting. Pitching can be thought of as more “defensive” statistics and hitting equates to “offensive” statistics. Several statistics which we intuitively believed contributed to playoff success were part of our initial data exploration. We created boxplots using Tableau for on base percentage (OBP), batting average (BA), slugging (SLG), and walks + hits by innings pitched (WHIP). Definitions for each statistic are below.

* On base percentage: the number of times that a team reached based divided by the total number of at bats
* Batting Average: the total number of hits for a team divided by their number of at bats
* Slugging: the percent of hits which results in a double, triple, or home run
* Walks + Hits divided by Innings Pitched: the number of walks plus hits divided by the total number of innings pitched

Not surprisingly, the boxplots revealed that hitting and pitching statistics generally improved as teams progressed further in the playoffs.

Once the dataset was loaded into Python, we continued our analysis and created algorithms using packages like Seaborn, Matplotlib, Py4j, Pyspark, Atexit, and Train\_test\_split.

# Conclusions

Overall, we were somewhat satisfied with our model’s results. Given the inherent randomness in sports – and baseball in particular – being able to mathematically predict the world series winner is no small feat.

While we were able to correctly predict the world series winner in a few seasons, our model was not without flaws. In certain cases, the bulk of the model’s accuracy was achieved by predicting every team to lose in the first round of the playoffs. Since 4 out of 8 teams lose in the first round, predicting all 8 teams to lose guarantees an “accuracy” of 50%. This result was typical of the KNN algorithm but not as prevalent with logistic regression. A possible remedy for this situation would be to incorporate some sort of linearity constraint which forces the model to choose a winner and loser in each round of the playoffs.

The next natural step for our model would be to incorporate more data. Team statistics such as SLG, WHIP, and BA are certainly valid predictors for playoff performance, but there are myriad other factors which contribute to a team’s chances of winning a world series. Several extensions for our model include:

* Each team’s expected performance against their opponent
* Strength of schedule during the regular season
* Recent performance for the past ‘n’ regular season games before the playoffs
* Home field advantage
* Individual player analysis

We plan to further investigate each of these factors and improve our model in the coming weeks.