Alan Gutierrez and Aaron Zhang
Prof. Hope
Data Analytics & Visualization

Final Project Technical Report

Data Collection and Processing

In our project, we used data from FanGraphs as they provide various baseball statistics that are useful for our analysis. Specifically, for our data collection, we exported batting and pitching player data from the 2010-2024 MLB seasons, excluding the 2020 season because that was a shortened season due to COVID-19. We had to export two types of datasets for both batters and pitchers from FanGraphs, as they split up certain statistics based on category. So, if you were to click the FanGraphs link, we exported the data from the "Dashboard" and "Win Probability" stat presets that they had. To make the analysis less confusing for us, we separated the regular season and postseason datasets. We then merged the data for all regular seasons from 2010 to 2024 into one dataset and did the same for the postseason data, grouping each by player names and calculating the averages of numeric columns to summarize each player's performance across seasons. This process allowed us to create two comprehensive datasets for each player type (batter and pitcher), one reflecting overall regular season performance and the other reflecting overall postseason performance. We then filtered the regular season data to only include players who met our criteria for being considered a “star player” based on their Win Above Replacement (WAR) statistic. For batters, we included those who averaged 3.75 or more WAR per season, and for pitchers, we included those who averaged 2.75 or more WAR per season. Additionally, for relief pitchers, we included those who averaged 20 or more saves. Finally, we filtered the postseason datasets to include only players who were identified as "star players" in the filtered regular season dataset so that the names in both datasets matched (Fig. 1).

```
[ ]  star_pitcher_rs = result_pitching_rs.query('WAR >= 2.75')

     star_pitcher_rs.to_csv('/content/drive/MyDrive/Data_AV_Batting_Stars/pitching_stars_regular_season.csv', index=False)

[ ]  star_reliever_rs = result_pitching_rs.query('SV >= 20')

     star_reliever_rs.to_csv('/content/drive/MyDrive/Data_AV_Batting_Stars/pitching_stars_reliever_regular_season.csv', index=False)
```

```
[ ]  # Filter result_df_postseason to only include players in star_df_reg_szn
     star_pitcher_ps = result_pitching_ps[result_pitching_ps['Name'].isin(star_pitcher_rs['Name'])]

     # Display the filtered DataFrame
     star_pitcher_ps.head()

     star_pitcher_ps.to_csv('/content/drive/MyDrive/Data_AV_Batting_Stars/pitching_stars_postseason.csv', index=False)
```

```
[ ]  star_reliever_ps = result_pitching_ps[result_pitching_ps['Name'].isin(star_reliever_rs['Name'])]

     star_reliever_ps.to_csv('/content/drive/MyDrive/Data_AV_Batting_Stars/reliever_stars_postseason.csv', index=False)

     star_reliever_ps
```

Figure 1: Definition and querying of star pitcher and relievers during the regular season

Visualization Choices & Alignment with Goals

Our goals when it came to this project were to:

- Quantify difference in performance between both batters and pitchers between the regular season and the postseason
- Analyze how external factors such as competition and pressure could potentially be attributed to difference in performance
- See if a trend we observe for one statistic is the same for other statistics

To do this, we created several visualizations using techniques learned from the course.

Scatter Plots

Scatter plots are useful for our analysis because they allow us to visualize individual data points, observe the distribution, variability, and relationships between two variables. We use a scatter plot to look at slugging percentage (SLG) vs. on-base percentage (OBP) of star players for the regular season and postseason. In our scatter plots, we include 95% confidence intervals for both the regular season and postseason in order to show the region in which the true mean of SLG and OBP is likely to fall for each season. We also include the actual mean of SLG and OBP with a red dot to emphasize the differing values between the regular season and postseason, highlighting the difference in offensive performance metrics. An alternative we considered was plotting the SLG vs. OBP as a linear regression and seeing how well the data fit. If there was a good fit, this could potentially allow us to predict SLG vs. OBP. However, we chose not to do this because it didn't really align with proving our hypothesis and goals of quantifying the difference in performance between the regular season and postseason. Moreover, the visualizations we chose to do instead align with our project's objective of analyzing how star players' performance changes between the regular season and postseason.

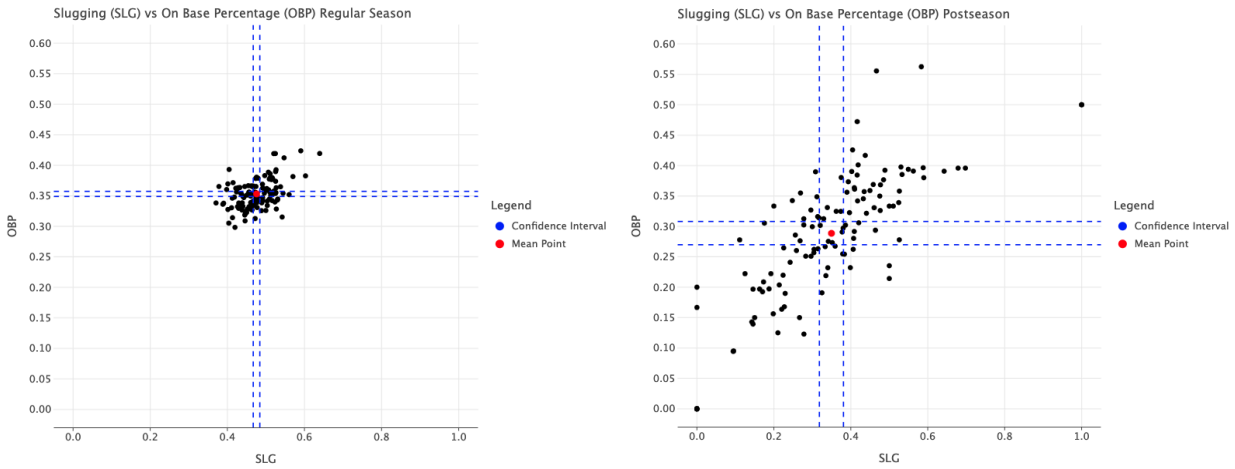

Figure 2: Comparison of star players SLG and OBP in the Regular Season and Postseason

Bar Plots

We chose bar plots as a visualization because they effectively highlight differences across groups and provide a clear, intuitive representation of numerical values and trends across distinct categories. By grouping data into meaningful categories, bar plots allow for straightforward side-by-side comparisons. Making them ideal for analyzing changes in performance across different contexts such as leverage levels, player roles, and seasons.

For the first bar plot (Fig. 3), we visualized on-base percentage (OBP) values across different player leverage index (pLI) levels (low and high) and seasons (regular vs. postseason). To make the comparison clearer, we used color to distinguish between the regular season (red) and postseason (blue), making it easy to identify trends and differences. Before creating the visualization, we had to group both star regular season and postseason datasets into a combined dataset and categorize the data by player leverage index (pLI) levels. First, we added a ‘Season’ column to each dataset to differentiate between the regular season and postseason data. We then combined the two datasets into a single dataframe. We then defined our leverage levels by creating bins to categorize player leverage index (pLI) into two distinct levels: "Low" for pLI values less than or equal to 1.0 and "High" for pLI values greater than 1.0. Finally, we grouped the combined data by season and pLI level and calculated the mean on-base percentage (OBP) for each group. This visualization aligned with our goal in wanting to analyze how pressure impacts star batting performance.

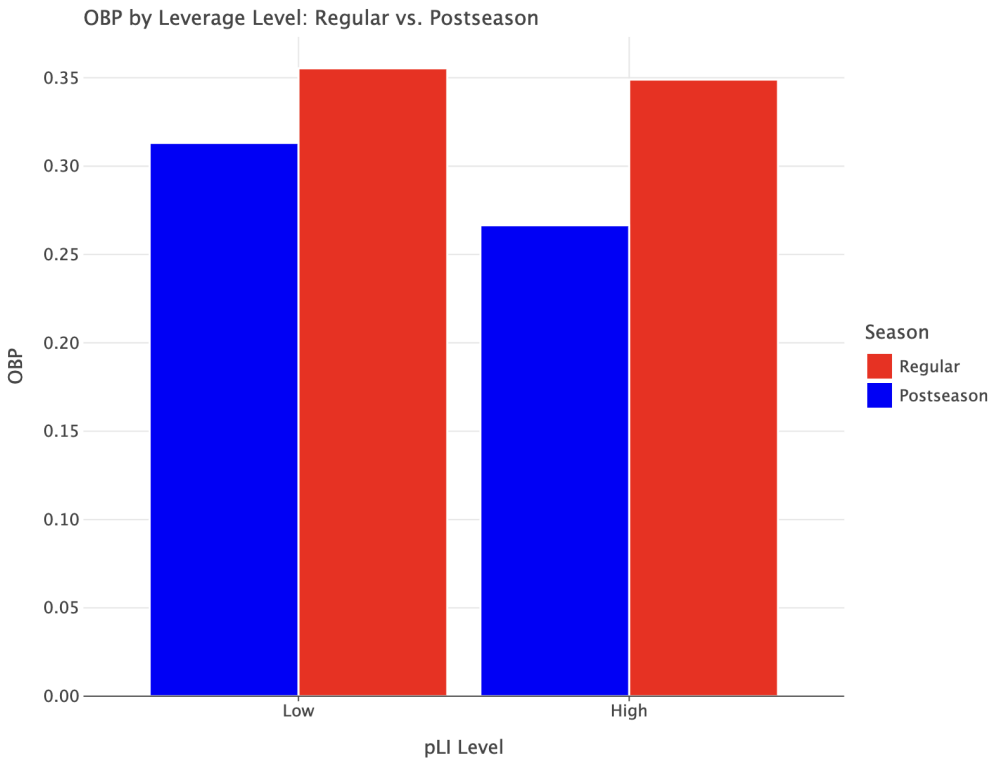

Figure 3: Comparison of star players OBP by pLI in the Regular Season and Postseason.

Similarly, we created a bar plot to visualize the meltdown (MD) per game rate between the regular season and postseason for star pitchers. We again used color to distinguish between the regular season (red) and postseason (blue). Before creating the visualization, we had to take the ratio between MD and games a pitcher played. We did this because the postseason has fewer games compared to the regular season, and standardizing the MD rate allowed for a fair comparison of performance across the two contexts while accounting for differences in sample sizes. We then separated the visualization into starting pitchers and relievers to gain insight into the type of star pitcher that tends to experience performance differences across both seasons. Finally, we visualized their meltdown (MD) per game rates for both the regular season and postseason (Fig. 4). Overall, this bar plot aligns with our goal of analyzing the performance difference between different types of star pitchers in the regular season and postseason.

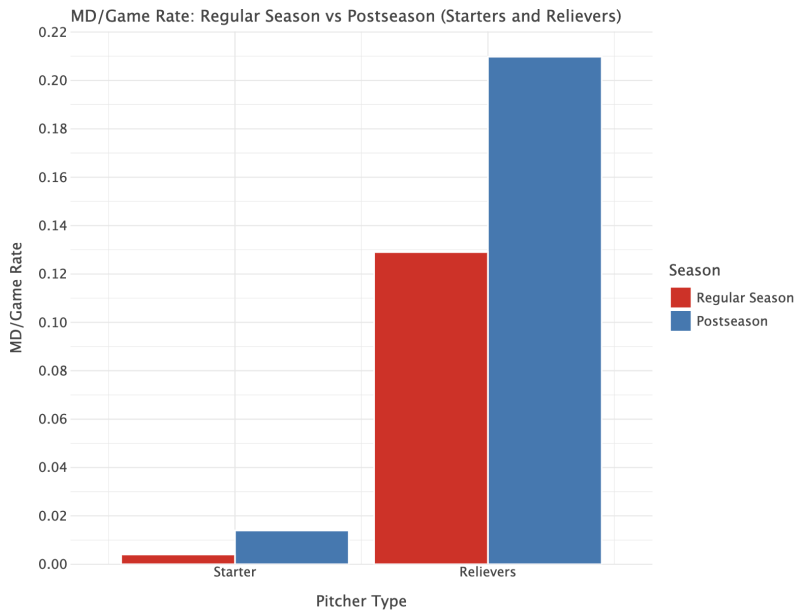

Figure 4: Comparison of Meltdowns/Game between regular season and postseason for Starters (left) and Relievers (right)

Histogram

We chose histograms as another type of visualization since they can provide an idea of the distribution of a certain statistic as well as the density of certain values of the statistic being evaluated. Additionally, with histograms, we are able to detect outliers, which may impact the average of certain statistics. We visualized the distribution of pitcher ERA by density for both the regular season and postseason and overlaid the two histograms on top of each other. This allowed us to once again compare the regular season and postseason datasets. Specifically, it allowed us to compare the distribution of ERA among pitchers between the regular season and postseason (Fig. 5). This showed us that the regular season ERA data seems to follow a normal distribution while the postseason ERA data seems to follow a left skewed distribution. However, we also observe a higher density of outliers for postseason pitchers. One potential reason for this is that the postseason is a much smaller sample size than the regular season. Pitchers only have a few opportunities to pitch during the postseason. Thus, one good performance or one bad performance could heavily affect a pitcher's ERA average, hence why we see a greater density of pitchers with lower and higher ERAs in the postseason relative to the regular season. This visualization aligns with our goal of quantifying differences in performance between the regular season and postseason for pitchers since we see the difference in distribution between the two datasets. This tells us that there will probably be a difference in ERA summary statistics, which we later confirmed was true. We considered using a box

plot instead of histogram to visualize the frequency of ERA along with summary statistics such as median, first and third quartiles, and range. However, we decided that these statistics were not as valuable as visualizing the distribution of ERA since the summary statistics could easily be calculated and displayed as a table as we did in our article.

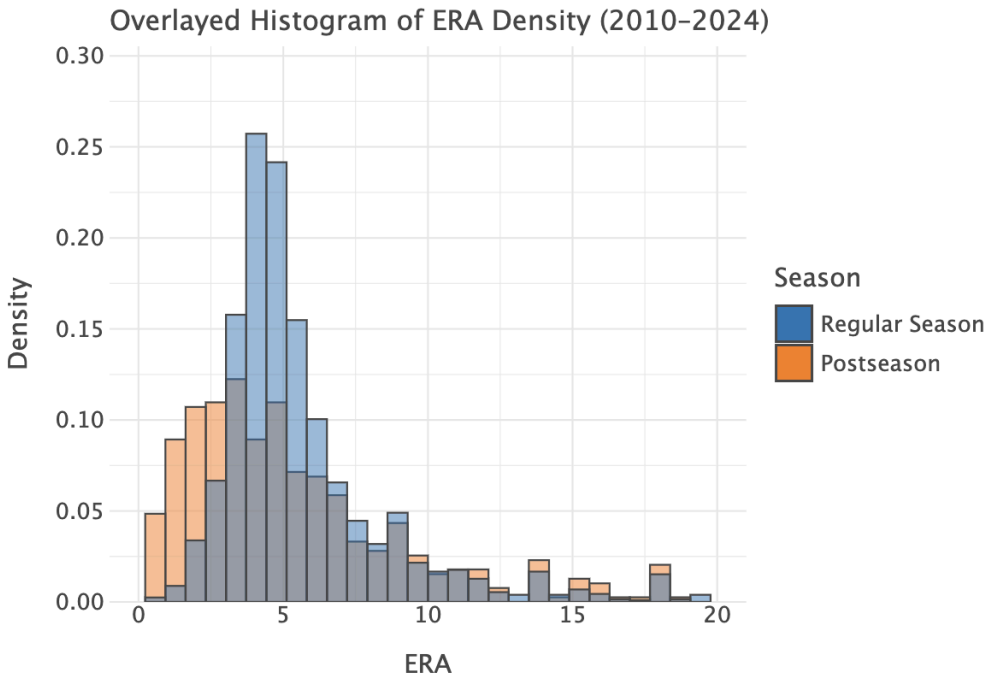

Figure 5: Comparison of density distributions between pitcher ERA during the regular season and postseason.

Statistical Models and Techniques

Descriptive statistics

Overall, we utilized many visualization techniques that help describe or quantify trends or differences in certain statistics between different datasets. We believe that this is justified because sports and more specifically player performance is heavily reliant on making comparisons for certain statistics. This helps determine factors such as a player's value in free agency or how well they fit within a team scheme philosophy. Such visualizations include scatter plots, bar plots, and histograms. Scatter plots were used to visualize the relationship between SLGS and OBP, the spread of individual data points, and more. Bar plots were used to directly and easily compare player performance between the regular season and

postseason. This was done specifically with OBP for batters and Meltdowns/game for pitchers. Through the use of Leverage Index we were also able to visualize differences in OBP for batters in low or high leverage situations. Finally, histograms were used to visualize the distribution of data, specifically for ERA. This required further combining of the datasets and creating a new column variable, ‘Season Type’, to indicate the regular season or postseason.

Bootstrapping

In the postseason, the smaller sample size introduces greater variability among players' statistics, making traditional methods for estimating averages less reliable. Since a main component of our project is analyzing postseason statistics, we had to implement bootstrapping to compensate for this issue with our data. Bootstrapping allows us to address this issue by resampling the data repeatedly to create a distribution of possible means. We used bootstrapping to provide us an estimate of the true mean for slugging percentage, on-base percentage, earned run average, and fielding independent pitching, ensuring that our comparisons between regular season and postseason performance were robust despite the smaller postseason sample size. Specifically, for both the regular season and postseason, we resampled each statistic 10,000 times with a confidence level of 95%. Therefore, by resampling we ensured that our results were not overly influenced by the variability in smaller datasets and accounted for the limited sample sizes in the postseason compared to the regular season.

Conclusion

Our analysis shows a consistent decline in performance for star batters and pitchers (starters and relievers) during the postseason compared to the regular season. This was consistent with the different metrics explored including SLG, OBP, ERA, FIP, and MD rates. In summary, we utilized many different visualization techniques (bootstrapping, descriptive statistics, etc.) to help accomplish our goals of quantifying the difference in performance between both batters and pitchers between the regular season and the postseason. This led to key insights for higher competition and increased pressure. We saw that postseason opponents exhibit stronger performance metrics than the average level of difficulty for star players during the regular season. We also saw that high-leverage scenarios and heightened stakes appear to negatively impact consistency and performance.