Bellevue University

DSC530 Data Exploration and Analysis (T302)

Professor Matthew Metzger

Summary by Jake Meyer

MLB Salaries and Player Performance

My final project focused on Major League Baseball (MLB). The question I analyzed was, “Do MLB players with high salaries tend to perform better than MLB players with lower salaries?” The salary scale varies substantially throughout the league with some players making over $20 million and others making under $200 thousand. Organizations must invest in players based on their skillset and potential to get better. My intentions were to understand if the players paid the most also perform the best. Some additional research questions to support this hypothesis are discussed below:

Statistical/Hypothetical Question(s)

* Main Question: Do MLB Players with high salaries tend to perform better than MLB players with lower salaries?
  + Does position play a role in salary? (i.e. do pitchers make more than position players)
  + Do both offensive and defensive performance influence salary?
  + Are there players that are at the top of the MLB pay scale that play below average?
  + Are there players at the lower end of the MLB pay scale that play above average?

Outcome of the EDA

The exploratory data analysis was performed on 5 variables for both pitcher and non-pitcher data sets. The 5 variables considered for pitchers were salary, strikeouts, earned run average, wins and losses. The 5 variables considered for field position players were salary, batting average, homeruns, RBI’s (runs batted in), and errors. A histogram was generated for each of the variables to get an understanding for how the values were distributed. The salary data for both sets showed a heavy positive skew of roughly 2.4. Later in the analysis, it was found that a lognormal transformation helped reduce this skew to 0.36 (pitchers) and 0.40 (field position players). A box plot was generated with the 75th percentile of player salary separated from each respective data set. This showed that majority of the players that who were paid high received similar salaries, however a select few from each category get paid extremely well (above $20 million). Probability Mass Function (PMF) and Cumulative Distribution Function (CDF) plots were generated next to compare the pitcher salary and non-pitcher salary distributions. The plots showed minimal differences. Scatter plots were generated for salary compared to each variable considered for the analysis. A difference of means test was performed to determine if pitcher salary and field position player salary were significantly different. Regression analysis was performed to understand which variables may be more influential for salary to construct a model. The last step was to characterize the relationship between salary and the other variables within each data set. This was completed by plotting the percentiles of player salary with respect to the different variables. The summary of results is provided below:

* Players with higher salaries did perform better in the variables considered in the analysis. (pitchers- strikeouts, ERA, Wins, Losses and non-pitchers- batting average, homeruns, RBI’s, and fielding errors)
* Pitchers make, on average, a little less than field position players. However, the difference is not significant when considering the salaries for the players (~$51,000 less).
* Players paid in the top 25% in the MLB tend to play better than average.
* Players paid in the bottom 25% range from performing below average to excelling above average.

What was missed during the analysis?

The topics required for the analysis were covered. Time was not considered for whether a player made more over a career. For Hypothesis Testing, the Difference in Means for pitcher salary and non-pitcher salary was compared. However, other Hypothesis Testing such as testing a correlation or testing proportions was not included in the analysis.

Were any variables that could have helped in the analysis?

Yes, there were many variables not included in this analysis that could have been beneficial for the statistical questions. The variables chosen for the analysis did provide insights into the questions, but some additional variables might help further clarify some findings. For example, some additional fielding variables would have been beneficial to include in the analysis for the non-pitchers.

Were there any assumptions made that appeared incorrect?

My final data sets consisted of data from the year 2000 and forward. In addition, the pitcher or field position player had to have a minimum of 40 batters faced or 40 at bats. This helped filter out the individuals that only played in a handful of games for a season. No outliers were removed from this data set once this criterion was implemented. In addition, there were no missing data values for salary.

What challenges did were faced and what did I not fully understand?

The biggest challenge I faced during the analysis was cleaning the data set. This portion of the project took quite a bit of time to learn and execute. Eventually, I was able to get to my final data sets for pitchers and non-pitchers, but this was by far my biggest challenge. Another challenge faced during the project was bouncing back and forth between the thinkstats.py code and thinkplot.py code or just using the available packages. I tried both methods in my final project, but probably should have stuck with one.

References:

Downey, Allen B. (2015). *Think Stats Exploratory Data Analysis*. [2nd Ed.]. O-Reilly Media, Inc

Lahman, S. (2021). *Download Lahman’s Baseball Database*. [Dataset]. [Download Lahman’s Baseball Database – SeanLahman.com](http://www.seanlahman.com/baseball-archive/statistics/)