Data Wrangling Project Report: MLB Player Performance Analysis

Introduction

Statistics play a crucial role in analyzing player performance, especially in professional sports like baseball. While the performance metrics of seasoned athletes are carefully tracked, many fans may not fully appreciate the various factors that influence these numbers. IN this project, I explore the relationship between MLB player hitting statistics and key variables, including years of experience in the league, position, and playing time. By leveraging fata from RotoWire and ESPN.com, I investigate how these factors contribute to variations in player performance.

Data Sources

RotoWire:

The first dataset comes from RotoWire, a platform that aggregates MLB player statistics for fantasy baseball purposes. It provides a comprehensive range of hitting statistics for all MLB players during that 2024 regular season. I downloaded the data in CSV format and imported it into my Python notebook. To prepare the data for analysis, I removed unnecessary columns and ensured that all data types were correctly formatted.

ESPN.com:

Additionally, I scraped data from ESPN.com, the official source of real-time player information and historical performance data. This allowed me to gather details on players' years of experience in the league, as well as other relevant features for my analysis.

Combining Sources:

To ensure a comprehensive analysis, I combined the RotoWire and ESPN datasets. Since there is overlap in the information from both sources, I standardized the structure of the data frames before merging them on player names. This ensured that I had a unified datast for analysis.

Below is my up to date data dictionary showing the column names, type and description of the variables used in my analysis.

Data Dictionary:

Column	Type	Description
player_name	text	Player full name
age	numeric	Age of player
team	text	Team player plays on
years_in_mlb	numeric	Number of years in major league
batting_average	numeric	Players batting average
home_runs	numeric	Total home runs hit by player
games_played	numeric	Total number of games a player has played
at_bats	numeric	Total number of at bats a player has
on_base_percentage	numeric	Players on base percentage
strike_outs	numeric	Total number of strike outs a player has
runs_batted_in	numeric	Players total runs batted in
position	text	Players position on field

Table 1 Data Dictionary

Analysis

1. **<u>Player Experience vs. Batting Average and Home Runs</u>**
   I wanted to investigate whether years of experience in the MLB affect a player's batting average and home run production. To do this, I created scatterplots to visualize the relationship between years of experience and both statistics
   - **Batting Average vs. Years of Experience**
     In Figure 1, the scatterplot reveals that years of experience do not have a clear impact on batting average. This was surprising, as I expected that more experienced players would have higher batting averages. However, further research suggests that factors such as age, skill development, and natural decline over time might influence batting average more significantly than experience alone. Older players, for example, often experience a decline in performance, which could contribute to this observation
   - **Home Runs vs. Years of Experience**
     Similarly, Figure 2 shows no strong correlation between years of experience and home run production. While I initially assumed that more experienced players would hit more home runs, the reality is that a player's role in the lineup is a more significant factor. Not all players are home run hitters; roles vary from speedsters and contact hitters to power hitters. This is why years of experience doesn't necessarily predict home run totals.

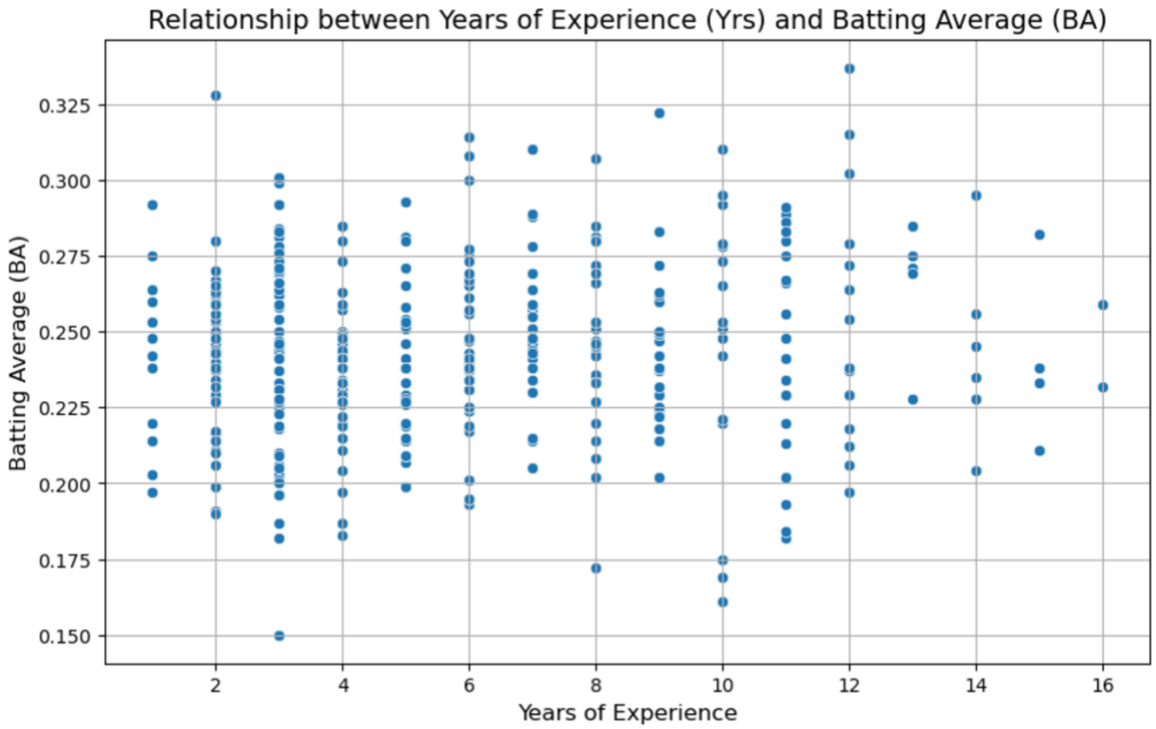

Figure 1 Years of Experience v. Batting Average

Figure 2 Years of Experience v. Home Runs

2. **<u>Player Performance Metrics by Position</u>**

   Next, I examined how player performance varies by position on the field. Different positions demand different skill sets, and this can influence performance metrics like batting average and home runs. I calculated the mean values for several performance metrics, including batting average, home runs, slugging percentage, on-base percentage, OPS (on-base plus slugging), runs, and hits.

   - **Batting Average by Position**

     In Figure 3, we see that the Designated Hitter (DH) has the highest batting average, which aligns with expectations. The DH is typically one of the best hitters on the team, and their primary role is to provide offensive production
   - **Home Runs by Position**

     In Figure 4, the DH also leads in home runs, with first basemen coming in second. This is consistent with the profile of a first baseman, who is often a power hitter with a strong build. Power hitting is a key part of their role in the lineup, which is reflected in the higher home run totals

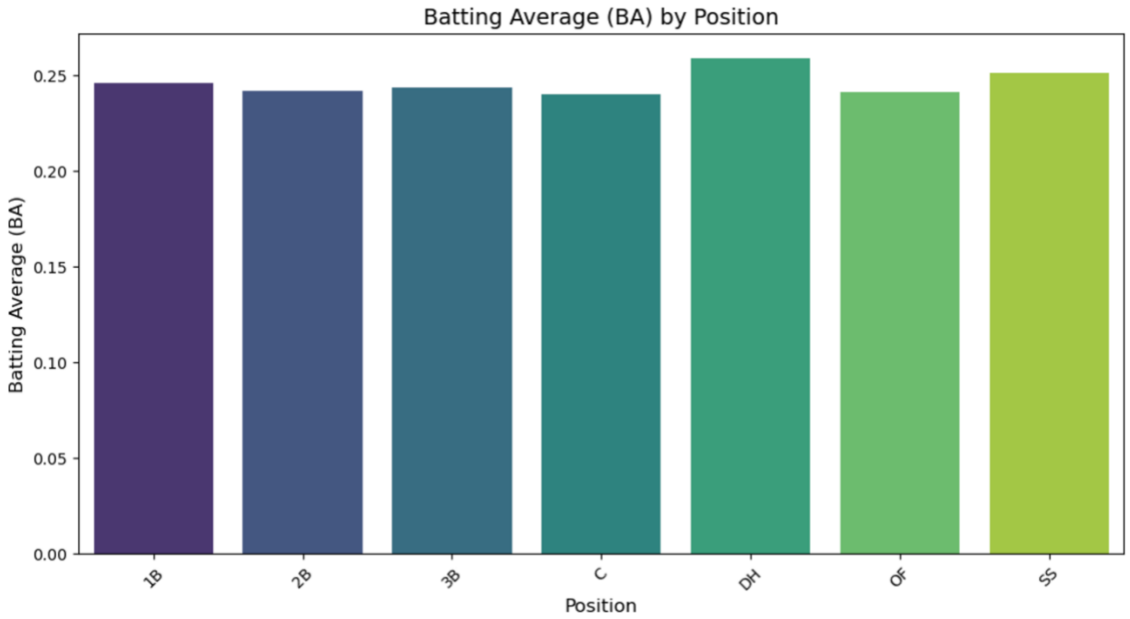

Figure 3 Position v. Batting Average

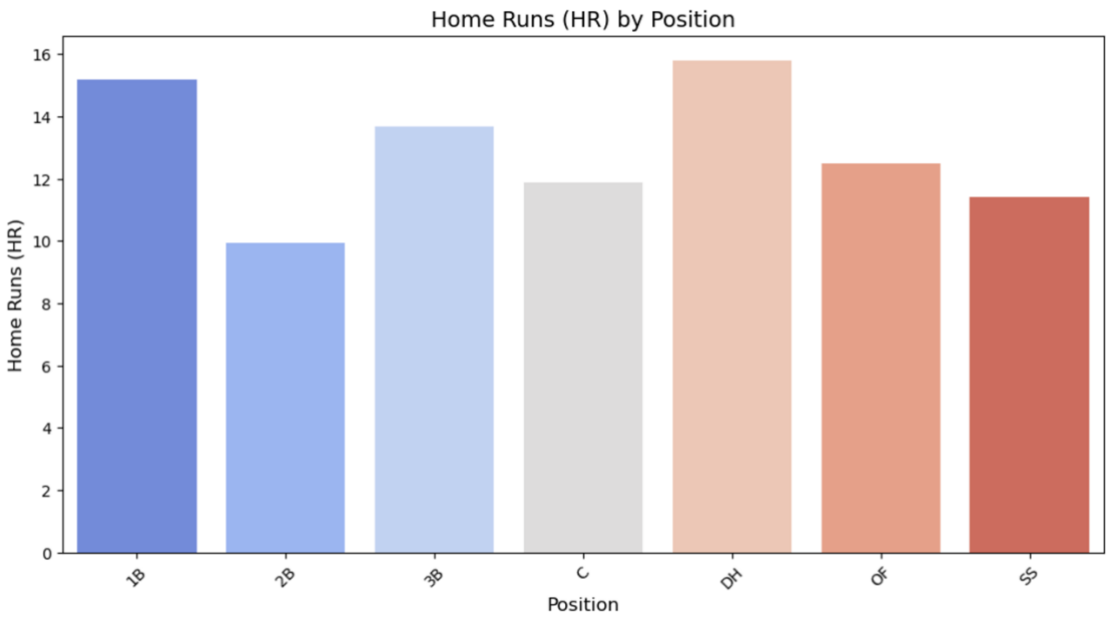

Figure 4 Position v. Home Runs

3. **<u>Playing Time v. Batting Average and On Base Percentage</u>**

   Finally, I explored how playing time, measured by the number of games played, affects a player's batting average and on-base percentage. Intuitively, it makes sense that as players accumulate more game experience, they would improve their performance.

   - **Batting Average vs. Games Played**

     In Figure 5, the scatterplot shows a slight positive trend between the number of games played and batting average. Players who spend more time on the field tend to improve their batting consistency over time, as they gain more experience and confidence.
   - **On-Base Percentage vs. Games Played**

     Figure 6 shows a similar trend, where the on-base percentage improves with the number of games played. This likely reflects players' increasing familiarity with opposing pitchers and their ability to adjust over the course of a season.

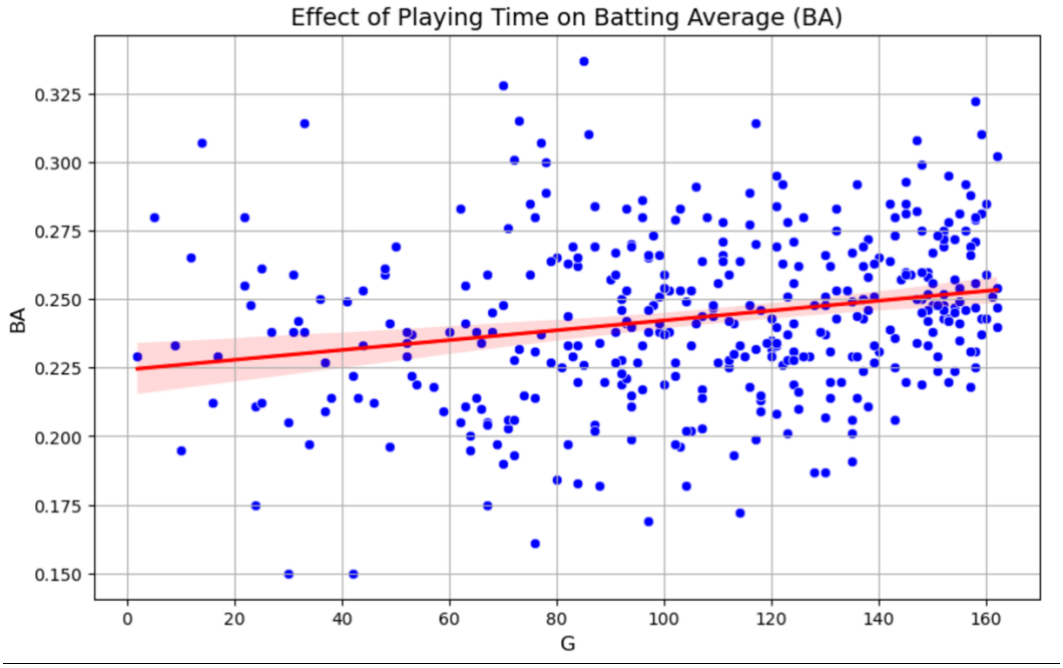

Figure 5 Number of Games v. Batting Average

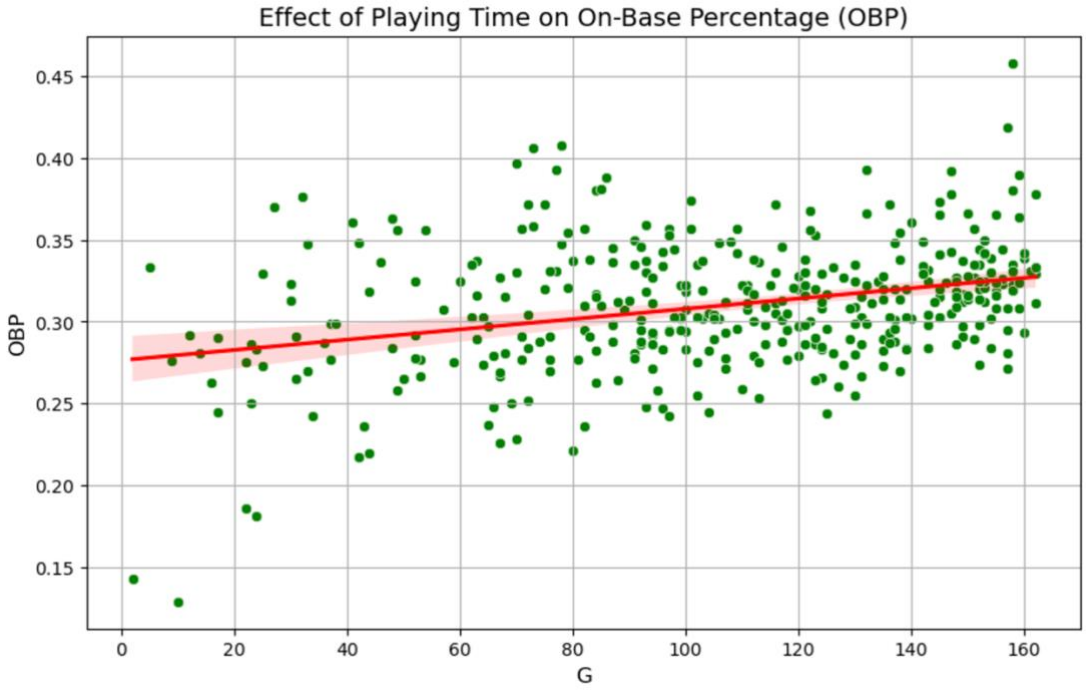

Figure 6 Number of Games v. On Base Percentage

Conclusion

In this project, I analyzed the impact of several key factors on player performance in Major League Baseball (MLB) during the 2024 season, focusing on player experience, position, and playing time. The main findings are as follows:

1. How does player experience (years in MLB) affect batting average and home run production?
    a. Years of experience in the MLB did not appear to significantly affect a player's batting average or home run totals. This suggests that other factors, such as age, role in the lineup, and individual skill development, play a more important role in determining these metrics

2. How do player performance metrics vary by position?
    a. Designated hitters and first basemen generally had the highest home run totals, with the DH also leading in batting average (.258). This reflects the different roles players occupy within a team, with power hitters (such as DHs and first basemen) naturally excelling in these categories.

3. What is the effect of playing time on players batting average and on base percentage
    a. There was a positive relationship between the number of games played and both batting average and on-base percentage. This indicates that players tend to improve over time as they gain more experience and settle into a rhythm throughout the season.

While the analysis provides valuable insights, there were several limitations to consider. Key factors, such as player age and injury history, were not included in the analysis, and these can significantly influence a player's performance. Additionally, the scope of the project was limited to the 2024 season, and trends observed in this year may not necessarily apply to other seasons due to changes in player talent, performance, and external factors.

Future Work

For further analysis, expanding the dataset to include multiple seasons and additional variables (such as player age, injuries, and advanced performance metrics like WAR) would provide a more comprehensive understanding of the factors that influence player performance in MLB.