**Wrangling Project Report**

**MLB Ballpark Factors**

**Tommy Heideman, Jake Mitchell, Joey Matusik**

**1) Introduction**

In Major League Baseball (MLB), the playing fields' diversity is unique, unlike any other major sport. Each baseball park, with its distinct dimensions and environmental factors, weaves a different narrative into the fabric of the game. Our project dives into this distinctive aspect of baseball, focusing on the role of ballpark factors, as detailed in our file, in influencing team and player performances.

Baseball parks vary significantly from one another, a contrast that is starkly different from the uniformity of fields or courts in sports like football, basketball, or soccer. This variety in field dimensions, elevations, and other locale-specific characteristics like weather and altitude, plays a critical role in shaping game outcomes. Our analysis seeks to explore these variations, examining how they impact batting and pitching statistics, and team success.

By melding traditional hitting stats with the diverse ballpark factors, we are not just comparing performances across the league but contextualizing them within these unique environments. This approach allows us to evaluate how a player or team's performance is influenced by the specific characteristics of the ballparks they play in. For instance, a ballpark with shorter outfield walls might favor hitters, while one with larger foul territory could benefit pitchers. Our project utilizes advanced data visualization techniques to bring these insights to life, illustrated through home team performances and ballpark characteristics. Through our analysis, we aim to provide a more nuanced understanding of how the distinct features of each MLB ballpark contribute to the narratives of the game.

In summary, our project is an exploration of one of baseball's most unique elements - its diverse playing fields. By delving into the data, we aim to uncover new perspectives on how these varying ballpark factors influence the outcomes and strategies of MLB games, offering fresh insights into the beloved sport of baseball.

**2) Data**

This project uses two sources of data that come from Kaggle and FanGraph. FanGraph was scraped using a built-in library called baseball-r. Baseball-r code we used for scrapping was found on GitHub page to get the specific tables needed (Petti). Kaggle’s dataset showcases the physical and environmental designs of MLB baseball stadiums (Johnson). FanGraphs dataset pertains to the statistical outcomes for the home teams that have played in their home stadium.

2.1 *Kaggle-MLB Ballparks*

We collected the data from Kaggle by downloading the attached csv file. This data was originally taken from Baseball Savant. This is a baseball website that provides statcast metrics, player matchups, and advanced statistics for many different instances. Further cleaning was not needed as all the data was contained in the correct columns. This data specifically consists of information pertaining to certain specifications that make up the dimensions of each baseball field for all MLB baseball teams. A data frame was created to allow us to implement visualization tools and so we could see what it looked like once loaded in. Our original data frame for this source was called “data” as we loaded it in as a csv file.

2.2 *FanGraph-Park Factors*

The fg\_park function extracts a comprehensive set of park factor data for a specific baseball season from FanGraphs.com. This function constructs a URL for the desired year and retrieves a data page that contains a table of park factors. The key metrics extracted include a variety of park factors such as basic 5-year averages, 3-year, and 1-year factors, along with specific breakdowns for singles, doubles, triples, home runs, strikeouts, unintentional intentional base on balls (UIBB), ground balls (GB), fly balls (FB), line drives (LD), infield fly balls (IFFB), and fielding independent pitching (FIP). It's crucial to note that all these metrics are normalized against a league-wide average set at a value of 100. This means values above 100 indicate a park factor more favorable than the league average for that statistic, while values below 100 indicate a less favorable factor. The function further processes this data by renaming columns for clarity, adding metadata such as the source and retrieval time, and includes error handling for data extraction issues.

Similarly, the fg\_park\_hand function is tailored to scrape park factor data by hitter handedness (left-hand hitters and right-hand hitters) for a given season from FanGraphs.com. It follows a similar procedure in URL construction and data extraction but focuses on handedness-specific park factors. The statistics extracted include park factors for singles, doubles, triples, and home runs, each differentiated by left-hand hitters (LHH) and right-hand hitters (RHH). Like the fg\_park function, these metrics are also normalized to a league-wide average of 100, providing a standardized basis for comparison. Values above 100 suggest a park is more favorable to a particular type of hit or hitter handedness, while values below 100 indicate the opposite. The function ensures the data is user-friendly by renaming columns and enriching it with metadata. Error handling is also a key feature, alerting users to any issues during data scraping. The data frames that were created for this scrapping were called “fangraph\_data” and “extra\_data”. The first data frame fangraph\_data included all fangraph data related specifically to the overall performance outcomes. The extra\_data was all related to splits of the data broken down by left and right-handed hitters.

*2.3 Merging of all data frames*

The merging of the three datasets, data, fangraph\_data, and extra\_data, into a single comprehensive dataset was conducted through a series of methodical steps to ensure accuracy and consistency. The process began with a thorough inspection of the data structures of each dataset to understand the variable types and their formats. Next, for uniformity across all datasets, column names were standardized. The team\_name column in data was expanded from abbreviations to full team names using a predefined mapping, and in both fangraph\_data and extra\_data, the home\_team column was renamed to team\_name.

The datasets were then sorted alphabetically based on team\_name, ensuring alignment of rows corresponding to the same teams, a critical step for accurate merging. To avoid redundancy and irrelevance in the final merged dataset, certain columns were removed. The season column was eliminated from fangraph\_data\_sorted and extra\_data\_sorted, and the team\_name column was removed from these datasets after sorting, as it was already included in data\_sorted.

The final step involved horizontally merging the datasets using cbind(), appending the columns of fangraph\_data\_sorted and extra\_data\_sorted to data\_sorted. This step was crucial for maintaining data integrity, ensuring that each row across the datasets corresponded to the same team. The merged dataset, now in ballparkfactor\_data, was then saved as a CSV file named "ballparkfactor\_data.csv" for future analysis or sharing. This meticulous process was instrumental in preserving the accuracy and usability of the data in subsequent analyses. The ballparkfactor\_data was 30 total rows since with all 30 MLB teams and then there were 35 columns of data for in the final merged frame.

*Data Dictionary “ballparkfactor\_data”*

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Data Type** |
| **team\_name** | The name of the baseball team (character). | chr |
| **ballpark** | The name of the baseball park or stadium (character). | chr |
| **left\_field** | Distance to left field in feet (integer). | int |
| **center\_field** | Distance to center field in feet (integer). | int |
| **right\_field** | Distance to right field in feet (integer). | int |
| **min\_wall\_height** | Minimum height of the ballpark walls in feet (numeric). | num |
| **max\_wall\_height** | Maximum height of the ballpark walls in feet (integer). | int |
| **hr\_park\_effects** | Home run park effects factor, scaled where 100 is average (integer). | int |
| **extra\_distance** | Total extra distance that a particular ballpark adds to hits. | num |
| **avg\_temp** | Average temperature at the ballpark (numeric). | num |
| **elevation** | Elevation of the ballpark above sea level in feet (integer). | int |
| **roof** | Percent of time the roof is used in the ballpark. | num |
| **daytime** | Percent of time games are played in daytime. | num |
| **basic\_5yr** | Basic 5-year performance factor, scaled where 100 is average (integer). | int |
| **3yr** | 3-year performance factor, scaled where 100 is average (integer). | int |
| **1yr** | 1-year performance factor, scaled where 100 is average (integer). | int |
| **single** | Single hit rate, scaled where 100 is average (integer). | int |
| **double** | Double hit rate, scaled where 100 is average (integer). | int |
| **triple** | Triple hit rate, scaled where 100 is average (integer). | int |
| **hr** | Home run rate, scaled where 100 is average (integer). | int |
| **so** | Strikeout rate, scaled where 100 is average (integer). | int |
| **UIBB** | Unintentional walks rate, scaled where 100 is average (integer). | int |
| **GB** | Ground Ball rate, scaled where 100 is average (integer). | int |
| **FB** | Fly Ball rate, scaled where 100 is average (integer). | int |
| **LD** | Line Drive rate, scaled where 100 is average (integer). | int |
| **IFFB** | Infield Fly Ball rate, scaled where 100 is average (integer). | int |
| **FIP** | Fielding Independent Pitching, scaled where 100 is average (integer). | int |
| **single\_as\_LHH** | Singles rate as Left-Handed Hitter, scaled where 100 is average (integer). | int |
| **single\_as\_RHH** | Singles rate as Right-Handed Hitter, scaled where 100 is average (integer). | int |
| **double\_as\_LHH** | Doubles rate as Left-Handed Hitter, scaled where 100 is average (integer). | int |
| **double\_as\_RHH** | Doubles rate as Right-Handed Hitter, scaled where 100 is average (integer). | int |
| **triple\_as\_LHH** | Triples rate as Left-Handed Hitter, scaled where 100 is average (integer). | int |
| **triple\_as\_RHH** | Triples rate as Right-Handed Hitter, scaled where 100 is average (integer). | int |
| **hr\_as\_LHH** | Home Run rate as Left-Handed Hitter, scaled where 100 is average (integer). | int |
| **hr\_as\_RHH** | Home Run rate as Right-Handed Hitter, scaled where 100 is average (integer). | int |

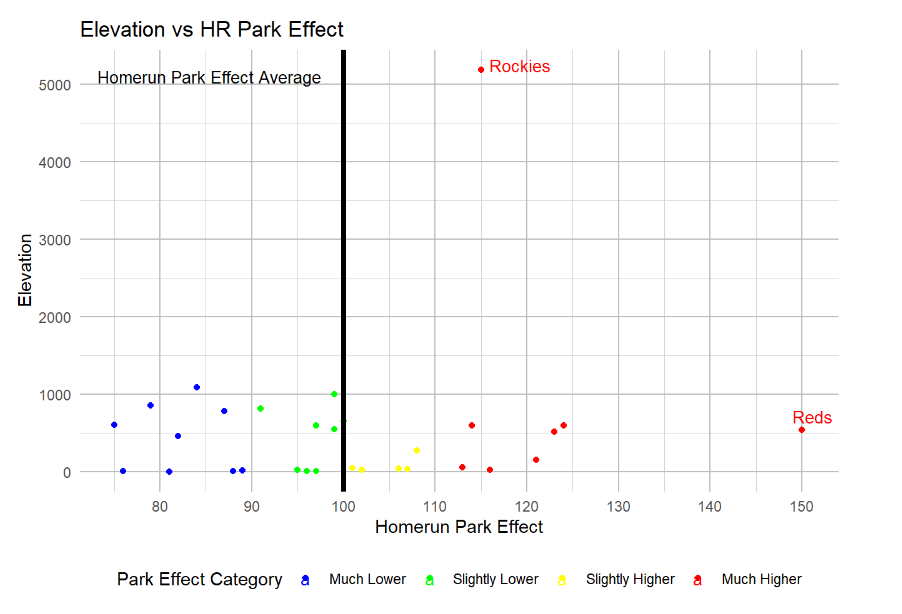
**3) Analysis**

*3.1 Environmental Factors and Homerun Effect*

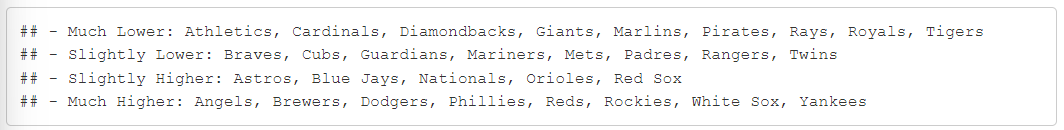
The scatter plot graph titled "Elevation vs HR Park Effect" analyzes the impact of baseball park elevation on the frequency of home runs. The graph highlights the Colorado Rockies' stadium as a significant outlier, suggesting that its unique elevation greatly affects home run probability. While the data points spread across elevations and home run effects, there is no definitive pattern indicating a strong correlation between the two variables, with a correlation of 0.13. This spread demonstrates that while elevation may be a factor in home run probability, it is not the sole determinant, as parks at similar elevations show a wide range of effects on home runs. The average effect line serves as a benchmark, with several parks falling above and below this line, indicating varying degrees of “hitter friendliness” or “pitcher friendliness”. The Cincinnati Reds’ Park, for example, is at a lower elevation but has a home run effect above the average, suggesting other factors at play that could influence the park's impact on home runs.

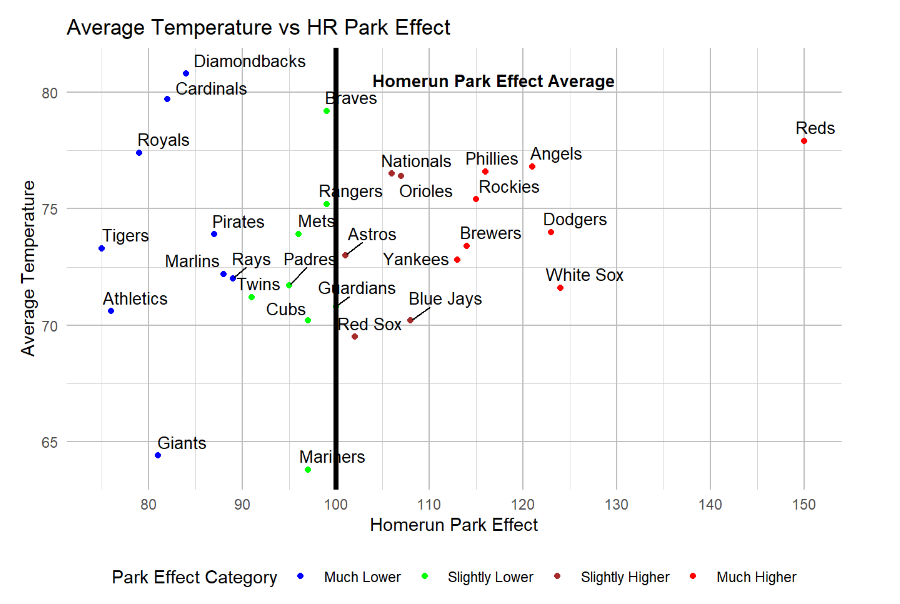
The next scatter plot "Average Temperature vs HR Park Effect", attempts to show a relationship between a stadium's average temperature and its propensity to allow home runs. With a noted correlation coefficient of 0.18, there is a very weak positive correlation, suggesting that temperature has a minimal impact on home run probability. Teams are spread across the graph without a discernible pattern, indicating that other factors may have a more significant influence on home run frequencies than temperature alone. For instance, both the Diamondbacks and Cardinals show a tendency towards higher temperatures and home run park effects, whereas the Reds, despite cooler temperatures, show a higher park effect. This distribution struggles to show the complexity of the factors affecting home run rates and suggests that temperature is not a strong predictor on its own.

Lastly, the "Daytime vs HR Park Effect" scatter plot suggests a weak negative correlation between the proportion of daytime games and the effect of the park on home runs, with a correlation coefficient of -0.21. This weak correlation suggests that as the number of daytime games increases, there might be a slight decrease in the park's home run effect, or vice versa. However, the spread of teams across the graph indicates that the time of day is not a strong predictor of a park's home run effect. The Reds again appear as an anomaly with a high home run park effect, while the Cubs, often associated with daytime baseball, sit near the average line, reflecting a moderate home run park effect. This data could imply that factors other than game timing may have a more significant impact on home run statistics.

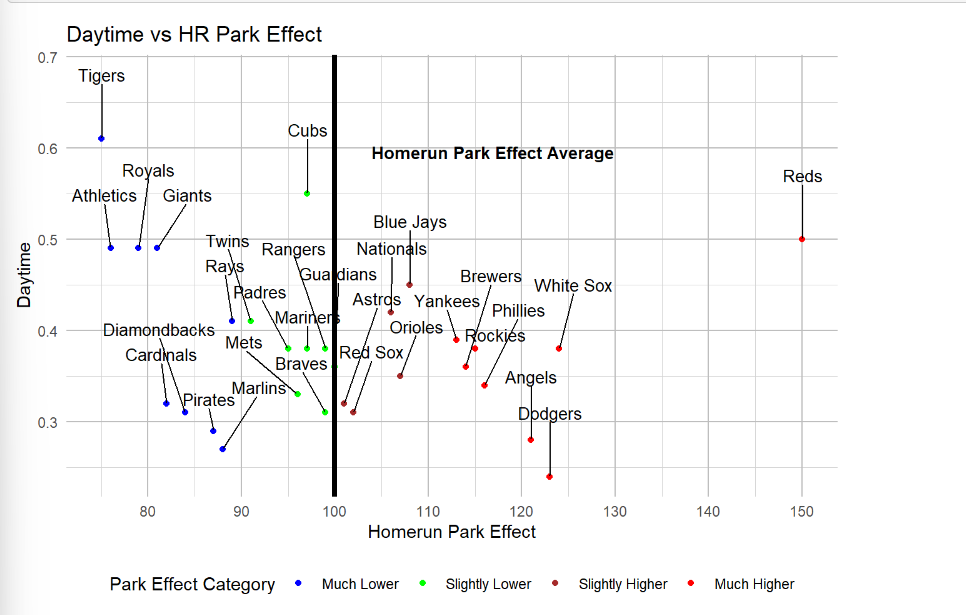




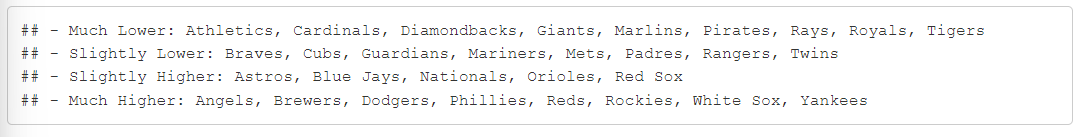












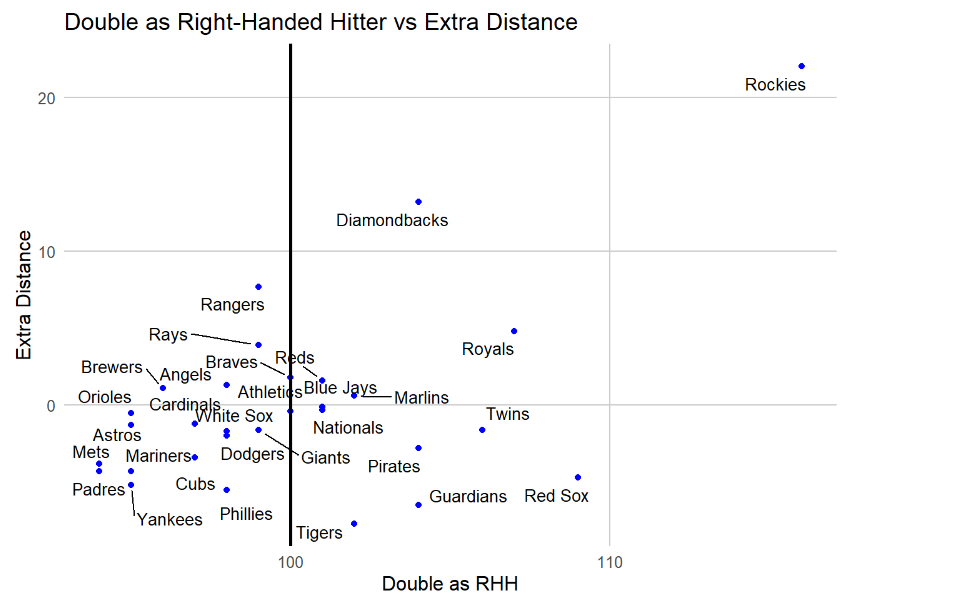
*3.2 Extra base hits vs extra distance in Stadiums*

Next, we will focus on the factors that affect the extra distance a hitter gains due to certain ballpark specific factors. The plot "Double as Right-Handed Hitter vs Extra Distance" reveals a pattern where most baseball teams are grouped together, suggesting that there is typically not a significant difference in the extra distance of doubles hit by right-handed hitters across various parks. A coefficient of 0.54 supports that there is some correlation but not enough to definitive resolution. However, the Colorado Rockies are a notable exception, indicating that doubles at their park travel a considerable extra distance, which can be attributed to the unique conditions at Coors Field, such as its altitude and spacious outfield. The Arizona Diamondbacks also show a trend of more frequent doubles with a moderate increase in extra distance, suggesting park factors that might favor right-handed hitters. The data implies that while certain parks like these can influence the distance of hits, for most teams, the ballpark does not drastically change the extra distance of doubles by RHH.

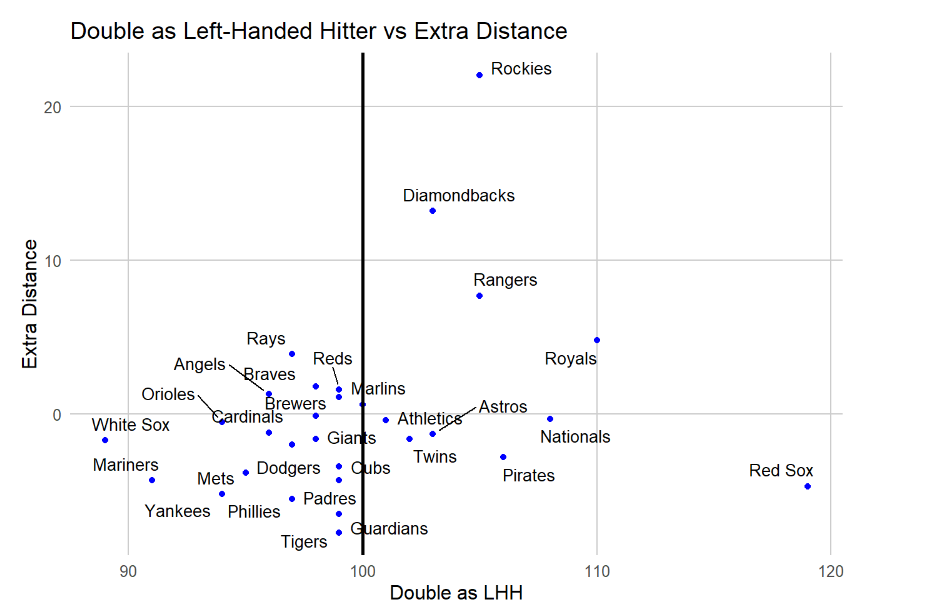
The "Double as Left-Handed Hitter vs Extra Distance" graph also provides a visual representation of the relationship between the number of doubles hit by hitters, this time left-handed, and the extra distance these hits travel at different baseball parks. The Colorado Rockies once again stand out with a significant extra distance, suggesting that Coors Field's unique environment dramatically affects the distance of hits, a phenomenon consistent for both right and left-handed hitters. The clustering of most teams within a central range suggests that for left-handed hitters, as with right-handed hitters, the extra distance of doubles is relatively consistent across different parks. The Boston Red Sox show a high frequency of doubles, yet the extra distance is less remarkable, indicating other park factors may contribute to the high number of doubles. With a coefficient of 0.25, the notion of little correlation is supported. These observations suggest that while park factors can influence the distance of hits, the effect is again, not universally drastic and may vary between parks.

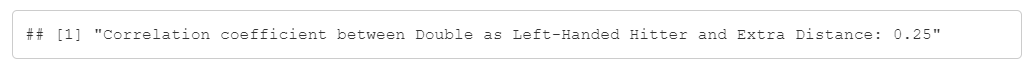
"Triple as Right-Handed Hitter vs Extra Distance" provides the same analysis with extra distances, but this time with triples. It is immediately apparent that the Colorado Rockies' stadium is an outlier again, with a significantly higher extra distance, echoing the trend seen in other hit types and likely attributable to the park's high altitude and spacious dimensions. The Arizona Diamondbacks also show a tendency for longer-distance triples, yet to a lesser extent. Most teams cluster at a lower range of extra distances, indicating a relative uniformity in park dimensions or atmospheric conditions affecting triples and a coefficient of 0.41 shows little correlation. The Detroit Tigers show a higher frequency of triples, suggesting Comerica Park's layout is favorable for three-base hits, though not necessarily for additional distance. The graph underscores the unique characteristics of certain ballparks that can influence the dynamics of the game, particularly with respect to extra-base hits by right-handed hitters.

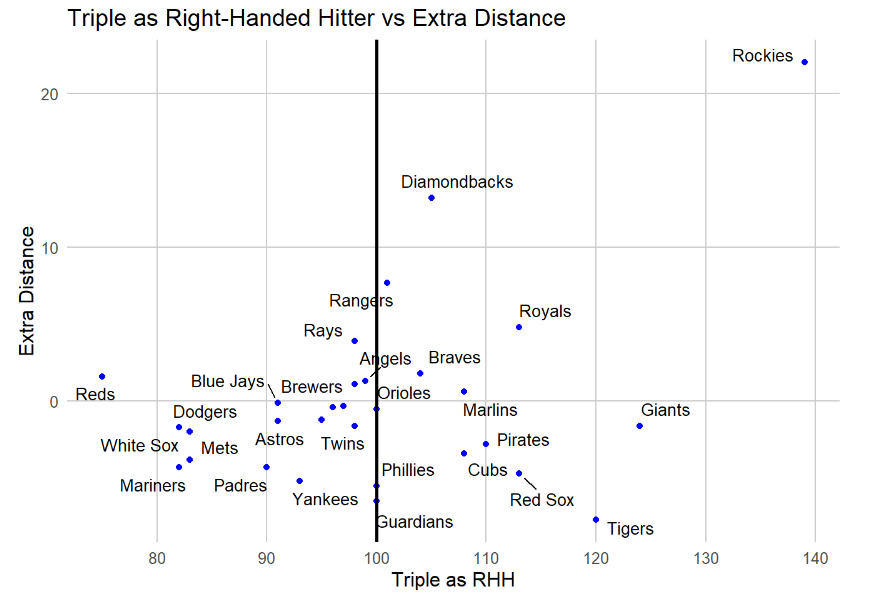
Lastly, "Triple as Left-Handed Hitter vs Extra Distance" shows a striking outlier with the Colorado Rockies, which is seen to be a trend, where triples by LHH travel a notably longer distance, reinforcing the unique impact of their ballpark's environment. The Arizona Diamondbacks also demonstrate a likelihood for longer triples by LHH, though less pronounced than the Rockies. Most teams are spread out within a central band, which suggests a relative consistency in the performance of triples by LHH in terms of extra distance across the league. The Detroit Tigers, while having a higher number of triples by LHH again, do not show a corresponding increase in extra distance, implying that their ballpark dimensions are more conducive to hitting triples regardless of the hit distance. A 0.2 coefficient shows a small correlation.

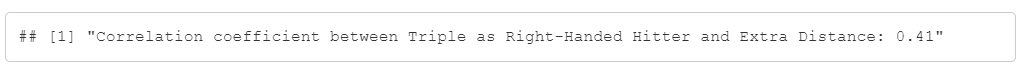


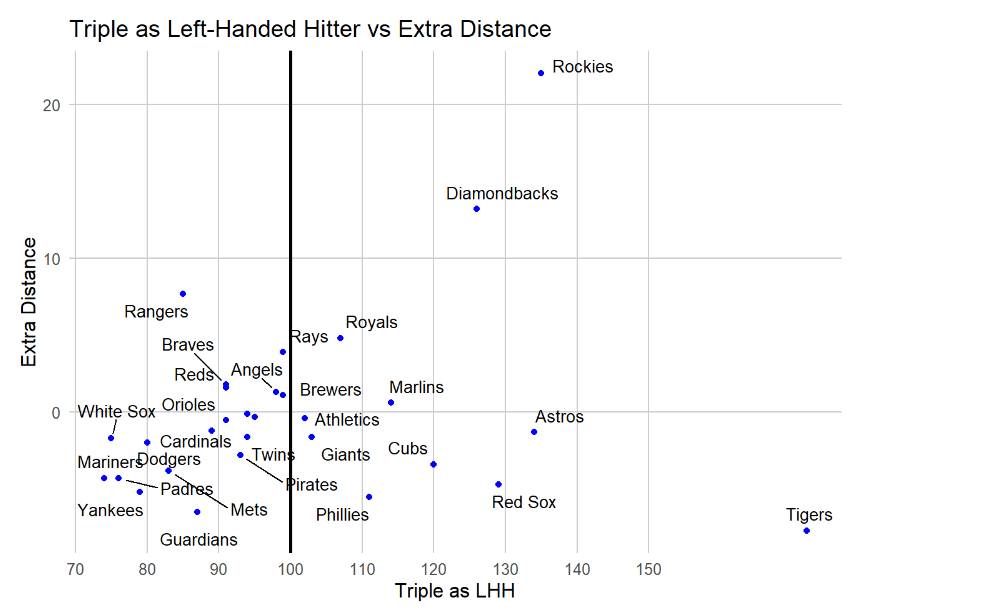


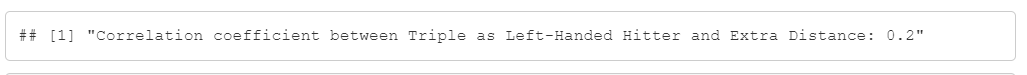












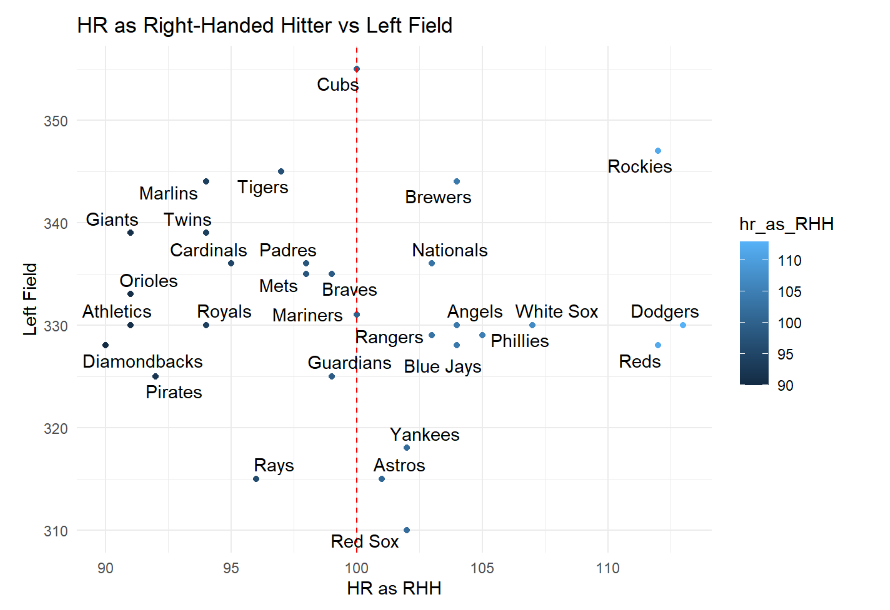
*3.3 Homeruns analysis for right and left-handed hitters in relation to pull side distance (right field for LHH and left field for RHH)*

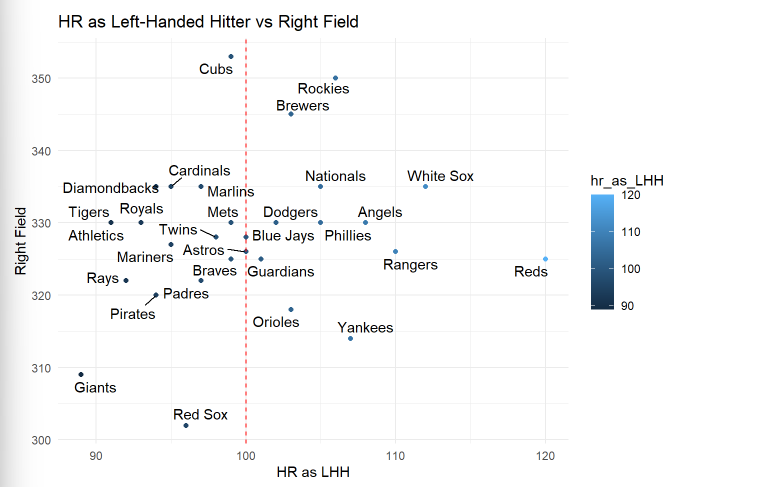
"HR as Right-Handed Hitter vs Left Field" scatter plot suggests a varied relationship between the distance to left field and the number of home runs hit by right-handed hitters in major league baseball parks. The Colorado Rockies' placement at the top of the graph reflects their park's tendency to favor hitters. On the other hand, teams such as the Red Sox, Yankees, and Astros show a higher frequency of RHH home runs despite shorter left field distances, possibly indicating hitter-friendly park dimensions. Conversely, the Cubs and Brewers show that a greater left field distance does not necessarily suppress the number of home runs hit by RHH. This data highlights the complex interplay between park dimensions and home run frequencies, with each park offering a unique environment that can affect hitting outcomes.

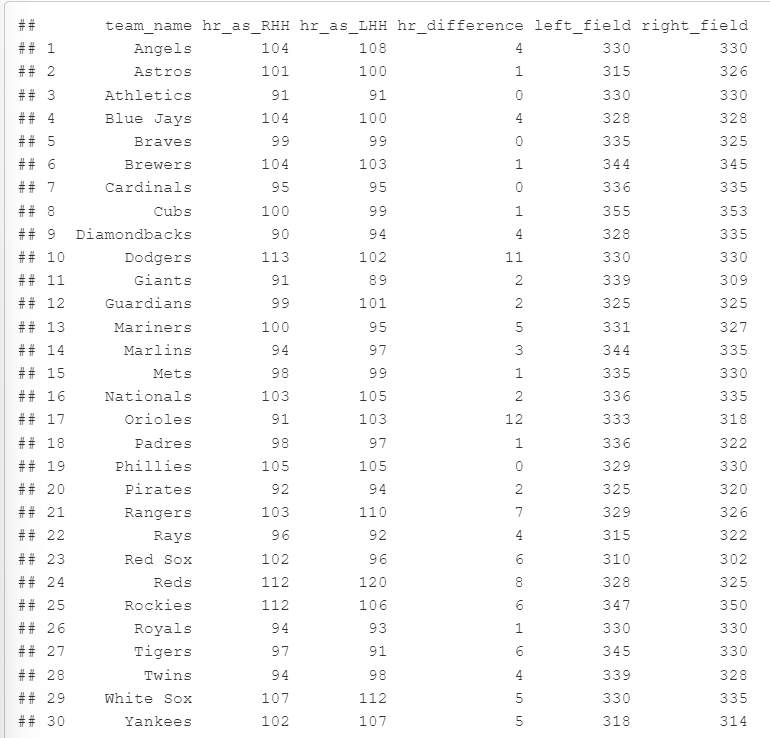
The scatter plot titled "HR as Left-Handed Hitter vs Right Field" shows the relationship between left-handed hitters' home run rates and the distance to right field across various baseball teams. The plot indicates no strong correlation between the distance to right field and home run rates for left-handed hitters. For example, teams like the Cubs, Rockies, and Brewers have high home run rates despite differing distances to right field. Conversely, the Giants and Red Sox have lower home run rates even with shorter distances to right field, suggesting that factors other than distance may play a role. The gradient of blue shades on the plot corresponds to the home run rates, with darker shades indicating higher rates. The dashed red line might represent the average home run rate, dividing teams with above-average rates from those with below-average rates. In summary, the plot suggests that right field distances alone do not dictate home run rates for left-handed hitters; other variables such as stadium design and environmental factors may also have an impact.

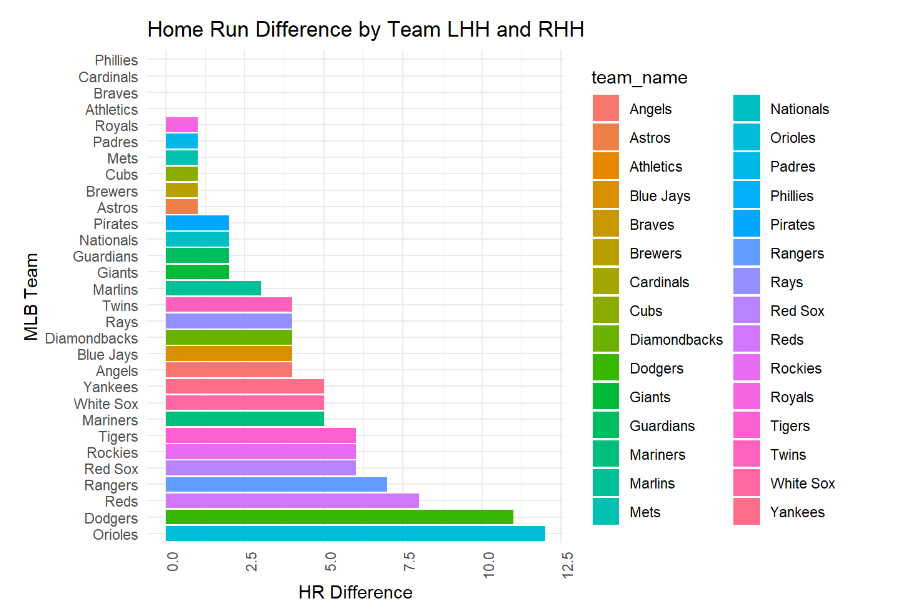
The provided table and bar graph offer insightful data on the variances in home run rates between left-handed hitters (LHH) and right-handed hitters (RHH) among the MLB teams. This analysis is crucial for understanding park factors and how they might influence player performance. it lists MLB teams with corresponding home run rates for LHH and RHH, indicating the differential between these figures, along with the ballpark's left and right field distances. This data is essential for assessing how ballpark dimensions may impact home run likelihood for hitters depending on their handedness.

The accompanying bar graph visually conveys the differential in home run rates between LHH and RHH for each team, with the MLB teams arrayed along the y-axis and the home run difference plotted along the x-axis. The graph's color coding facilitates the identification and comparison of teams at a glance. From the graph, it's observable that there's a significant variation in home run differentials across the league. Some teams show a pronounced advantage for LHH, such as the Dodgers, suggesting that their ballpark may be more conducive to left-handed power. In contrast, teams like the Yankees display a negative differential, hinting at a possible advantage for RHH in their home stadium. The detailed figures in the table enable a more granular examination. One could correlate the reported distances to left and right field with the home run rates for hitters of different handedness. For instance, a shorter right field may not always correspond with increased home runs for LHH, as other factors like the height of outfield walls also come into play. It's evident that ballpark dimensions might have a notable impact on home run numbers for hitters based on handedness. However, it's also clear that these dimensions are just one piece of a larger puzzle.







 **4) Conclusion**

In this project, we evaluated three factors that pertain to average results in a ballpark for any game that has occurred in the 2022 season. These consisted of environmental factors, extra base hits, and home runs. In summary, from our analysis questions that were displayed in our proposal, we found the following results:

1.) How do environmental factors impact the homerun effect?

The analysis of elevation indicates that while it plays a role, it's not the sole determinant, as shown by the diverse effects of parks at similar elevations. Similarly, the weak correlations observed in temperature and daytime games suggest that these factors alone have minimal predictive power on home run probabilities, highlighting the complexity of the interplay between environmental variables and their impact on home run rates in baseball.

2.) To what degree do extra base hits occur in a stadium that has longer average field dimensions?

The analysis of extra base hits, specifically doubles and triples by both right and left-handed hitters, indicates that certain ballparks, notably Coors Field for the Colorado Rockies, exhibit significantly longer distances for these hits. While there is some correlation between ballpark-specific factors and the extra distance of hits, the overall patterns suggest that such effects are not universally drastic across different stadiums. The occurrence of longer extra base hits appears to be influenced by unique environmental conditions and dimensions in specific parks rather than a consistent trend related to longer average field dimensions across the league.

3.) How does the number of home runs vary between right-handed and left-handed hitters in correlation with the distance to their respective pull sides?

The data analysis reveals a nuanced relationship between home run frequencies and the distance to pull sides for both right-handed and left-handed hitters in Major League Baseball parks. While certain teams, like the Colorado Rockies, exhibit a positive correlation between right-handed hitters' home runs and shorter left field distances, others such as the Red Sox, Yankees, and Astros demonstrate higher home run frequencies for right-handed hitters despite shorter distances. Conversely, the scatter plot for left-handed hitters in relation to right field distances indicates no strong correlation, emphasizing that factors beyond distance, such as stadium design and environmental conditions, play a crucial role in shaping home run outcomes for both handedness. Overall, the analysis underscores the complexity of the interplay between park dimensions and home run rates, necessitating a comprehensive understanding of various factors influencing hitter performance.

This project has some limitations that were present. Included in these limitations is that FanGraph data only pertains to the year 2022 which could potentially showcase bias amongst the data. MLB stadiums also undergo renovations and change outfield dimensions, which could lead to either positive or negative impacts when it comes to hitting performances. Further work that could be done on this project could include incorporating more statistical information such as pitcher performances and implementing the time of day, weather type, and season. Using these new metrics would create a more robust dataset and allow more insights to be had to see whether or not there are new correlations being made. We would also be able to evaluate if any hitting statistics would become altered due to the new features joining the dataset.

**Worked Cited**

Johnson, Paul R. "MLB Ballparks." Kaggle, 8 Dec. 2023, www.kaggle.com/datasets/paulrjohnson/mlb-ballparks.

Petti, Bill. "Baseballr." 8 Dec. 2023, billpetti.github.io/baseballr/