Project 3
Group 6
Project Write Up

Caleb Meinke

Adam Loux

Revano Harahap

Monica Mitry

Intro/Purpose of Project:

For the data story the project team decided to take it out to the ball game to explore the hitting trends and statistical relationships for MLB teams and players from 2019-2022. The original inspiration for choosing the data set was a curiosity on the relationship between stadium field sizes and the number of home runs that occur at the stadium. That curiosity led the team to sliding into team stats and swinging for the top hitters.

Utilizing the Kaggle dataset and merging player statistics with stadium locations the study explored home run leaders, player contributions, statistics correlation, and the impact of geographical location. Through visualizations like bar graphs, scatter plots, and heatmaps, trends and correlations were identified to better understand team offensive data.

The findings reveal fluctuating home run leaders, varying player contributions, and a consistent correlation between power hitting and strikeouts. Contrary to popular belief, altitude had no definitive impact on overall team performance. However, limitations—such as the absence of full team rosters and the shortened 2020 season due to COVID-19—highlight areas for further research. By leveraging data analytics, this project provides valuable insights into modern baseball trends and the evolving nature of offensive strategies.

Data & Data Cleaning:

To achieve the objective of this project, the project team acquired two MLB baseball datasets from the public access Kaggle repository. The first of these datasets, 2019-2022 MLB Hitting Leaders CSV (UTF-8).csv, contains offensive statistics for the top performing hitters on each of the 30 MLB teams for each of the years ranging 2019 to 2022. The second dataset, stadiums.csv, contains information on the geographic locations, via latitude and longitude, of all major league sports arenas in the United States, organized by team and sports league. Initial assessment of these datasets showed largely uniform and consistent data across all columns and rows, which minimized the need for extensive data cleaning.

```
# Data cleaning
df.rename(columns={
    "Team": "team",
    "Team_abv": "team_abv",
    "League": "league",
    "Division": "division",
    "Lat": "latitude",
    "Long": "longitude"
}, inplace=True)

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 151 entries, 0 to 150
Data columns (total 6 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   team       151 non-null    object
 1   team_abv   151 non-null    object
 2   league     151 non-null    object
 3   division   151 non-null    object
 4   latitude   151 non-null    float64
 5   longitude  151 non-null    float64
dtypes: float64(2), object(4)
memory usage: 7.2+ KB
```

As a result, the project team's initial data cleaning efforts were tied to a few specific tasks. First, this effort focused on renaming and reorganizing columns in the Hitting Leaders dataset for dataframe ingestion. Once completed, the project team applied similar data cleaning logic to the Stadiums dataset, which involved correcting a few typographical errors in team names, updating column names, and

preparing the table for merger with the Hitting Leaders dataset.

Following initial data cleaning, the team joined the stadiums data to the Hitting Leaders dataset, leveraging the team abbreviation to bring the datasets together in a single dataframe. From there, the process focused on removing duplicate player entries in the combined table, which ensured the final version of the dataframe contained individual rows for each player in each of the four years contained in the datasets. This finalized version of the dataframe showed uniformity in data and format, making it suitable for saving off to an SQL database to serve as the foundation for the project team's MLB Stats app.

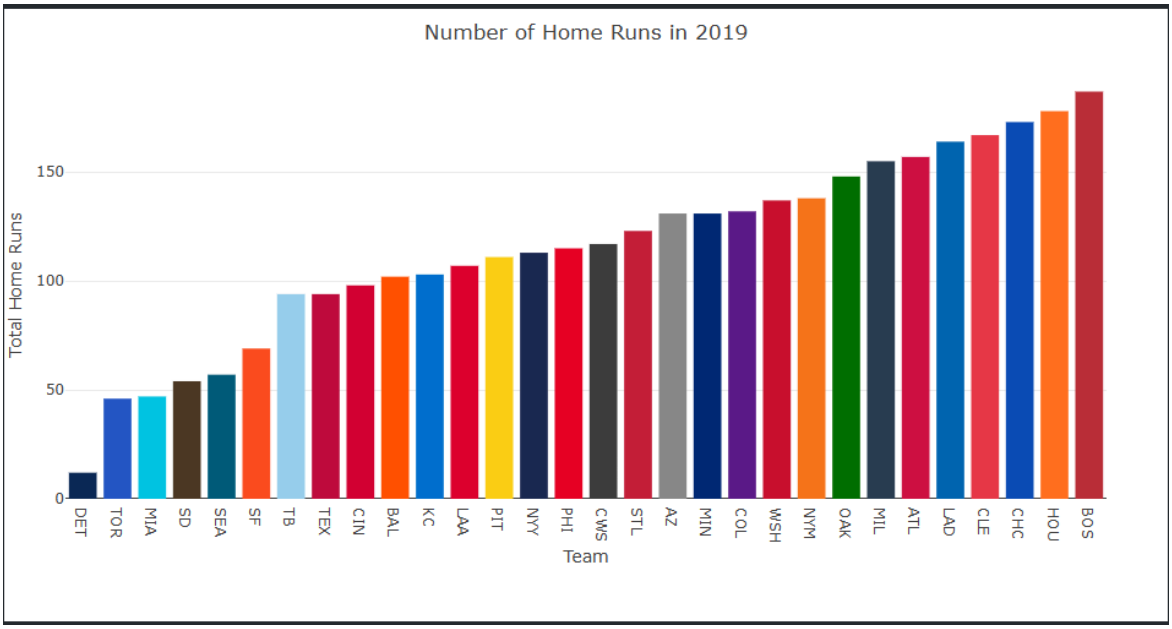
```
# Table join
hitting_leaders = pd.read_csv("Resources/hitting_leaders.csv")
stadiums = pd.read_csv("Resources/stadiums_mlb.csv")

# Merge on team_abv
df2 = hitting_leaders.merge(stadiums, on='team_abv', how='left')
df2
```

	year	player_name	player_position	team_abv	G	AB	R	H	2B	3B	...	CS	AVG	OBP	SLG	OPS
0	2022	Aaron Judge	CF	NYY	157	570	133	177	28	0	...	3	0.311	0.425	0.686	1.111
1	2022	Yordan Alvarez	DH	HOU	135	470	95	144	29	2	...	1	0.306	0.406	0.613	1.019
2	2022	Paul Goldschmidt	1B	STL	151	561	106	178	41	0	...	0	0.317	0.404	0.578	0.982

Major Questions:

Assessing the columns and information the first major question that came to mind was which team had the most home runs per year. For this it was necessary to aggregate all the players per team and sum the total of home runs each player contributed to the overall team statistics per year. Utilizing a bar graph to visualize this data, the project team developed a SQL query that extracted the year, team abbreviation, and sum of home runs data from the MLB database. Leveraging an api call to the SQL query, javascript code formatted this data into a custom bar chart that arranged annual home run data, by team, in ascending order using a custom team color Javascript object, hover display layout, and Plotly structure to display the data.

```
// Function to update bar chart
function updateBarChart(filteredData, year) {
    let teams = filteredData.map(d => d.team_abv);
    let homeRuns = filteredData.map(d => +d.total_home_runs);

    let trace = {
        x: teams,
        y: homeRuns,
        type: "bar",
        marker: { color: teams.map(team => team_colors[team] || "#888888") }
    };

    let layout = {
        title: `Number of Home Runs in ${year}`,
        xaxis: { title: "Team" },
        yaxis: { title: "Total Home Runs" },
        height: 500,
        width: 950,
        margin: { t: 50, l: 50, r: 50, b: 100 }
    };

    Plotly.newPlot("plot", [trace], layout);
}
```



Analysis of this data showed that in 2019 the top team with home runs was the Boston Red Sox, in 2020 it was a tie between the Atlanta Braves and the Los Angeles Dodgers, in 2021 it was the Toronto Blue Jays, and in 2022 it was the Houston Astros. It was clear that throughout the four years there was no dynasty team that held the top for total home runs. However, among the top three, the Boston Red Sox were in the top three for two of the four years. In 2022 and 2020 the top team for home runs was also the team that won the World Series (ESPN, 2025).

Total home runs lead to a deeper dive of player contribution. The inquiry was whether a certain player had a higher influence on total home runs or was it a fairly even team distribution. To investigate this further, the project team developed an interactive player table and team statistics heatmap, both filtered by year. The table permits the direct query of individual player metrics of all database statistics, and the heatmap provides a comparative view of team performance data to assess the top performing teams, by year. The data visualizations provided insightful information on individual and team performance.

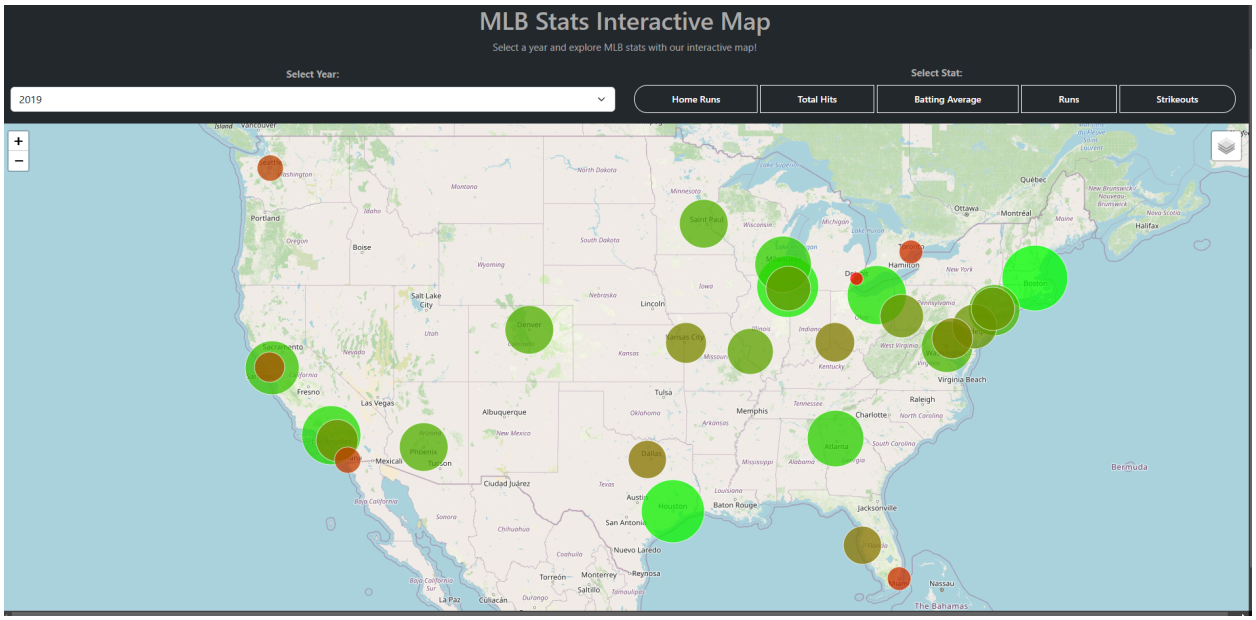
```
// Function to update heatmap
function updateHeatmap(filteredData, year) {
    if (!filteredData || filteredData.length === 0) {
        console.error(`Heatmap data is missing or empty for year ${year}.`);
        return;
    }

    let teams = filteredData.map(d => d.team_abv);
    let stats = ["AVG", "HR", "R", "SO"];

    // Ensure numeric values, default to 0 if undefined
    let zValues = stats.map(stat => filteredData.map(d => d[stat] !== null ? +d[stat] : 0));

    console.log("Heatmap Data:", { teams, stats, zValues });

    let trace = {
        z: zValues,
        x: teams,
        y: stats,
        type: "heatmap",
        showscale: true,
        xgap: 1,
        ygap: 1,
        colorscale: [
            [0, "blue"],
            [0.5, "white"],
            [1, "red"],
        ],
        hovertemplate: "Team: %{x}<br>Stat: %{y}<br>Result: %{z}<extra></extra>"
    };

    let layout = {
        title: `MLB Team Statistics Heatmap (${year})`,
        xaxis: { title: "Teams", tickangle: -45 },
        yaxis: { title: "Statistics" },
        height: 500,
        width: 950,
        margin: { t: 50, l: 100, r: 50, b: 100 }
    };
```

Specifically related to the comparison of individual and team performance on annual home runs outcomes, analysis of the data showed instances of both occurrences. Some teams such as the Texas Rangers in 2022 were the four players all had a total of home runs between 26-33. However, in the same year Aaron Judge with the New York Yankees had contributed to 40% of the total home runs for the Yankees and Anthony Rizzo with the next highest contribution of

20%. This could have led to a further analysis of which teams had outshining players versus overall team effort.

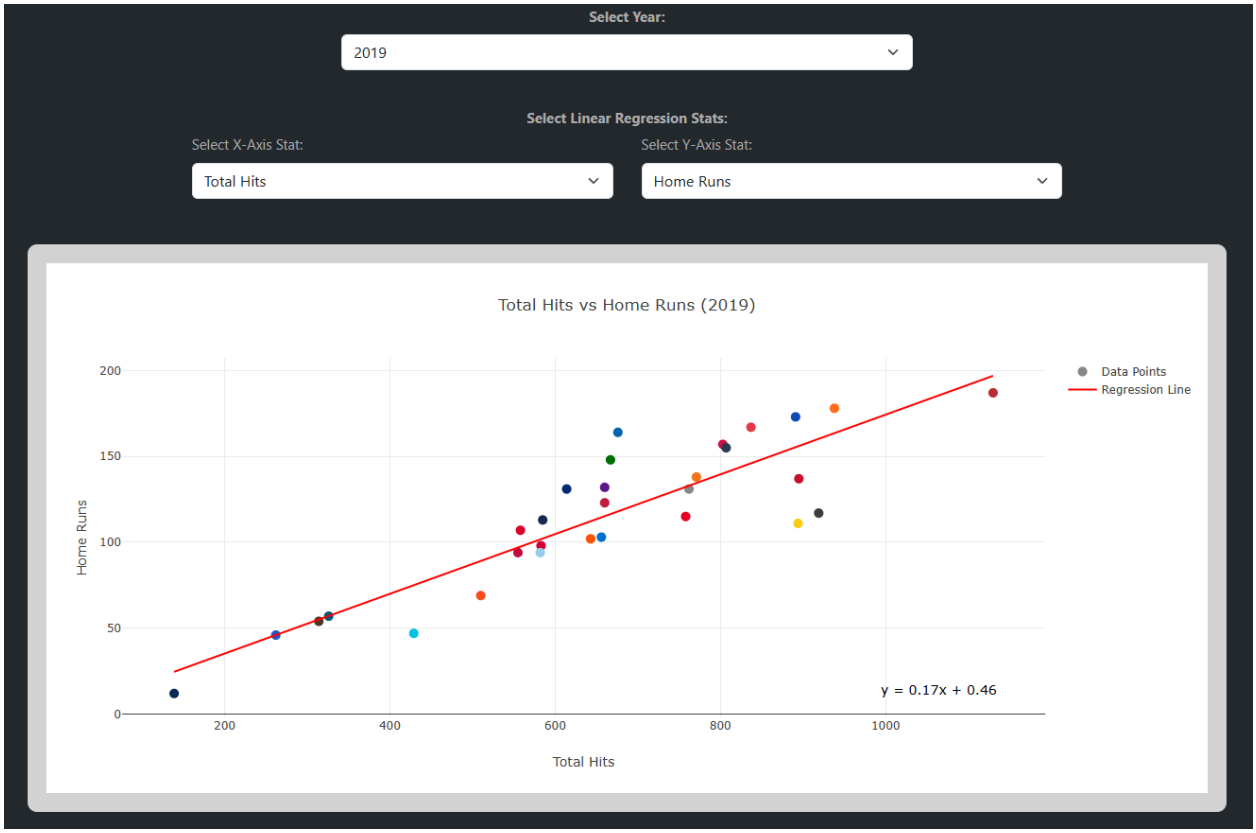
Among the data we wanted to see if there is a correlation between strikeouts and home runs. There seemed to be a trend of increasing the focus on home runs with a correlation of strikeouts and batting average. Teams with high strikeout counts also were power hitting teams. This remained fairly consistent over the four years. Based on the heatmap it was clear that a team's overall offensive performance between their home runs and total runs were directly correlated.

There has often been a discussion regarding altitude and elevation having an impact on ball movement in baseball. With higher altitudes, such as in Denver where the Colorado Rockies play, potentially contributing to increased home runs and altered pitch movement due to thinner air resistance. To assess this potential impact on performance outcomes, the project team constructed a geographic plot for annual team stats including home runs, total hits, batting average, runs, and strikeouts. This data was plotted by team location, with each plot scaling to overall statistical size and the gradient shading of the plots reflecting the league rank of each statistic.



Taking a look at the overall geographical location of each team expressed that there was no true correlation between altitude change and the overall performance. Teams that often play in higher altitudes, such as The Colorado Rockies, did not dominate the leaderboards in key statistics, suggesting that altitude had little influence. Other factors such as team strategy, player skill, and ballpark dimensions played a more significant role in the team performances.

Lasty, the relationships between various baseball statistics were examined using scatter plots to identify trends and correlations. Coding for these plots was constructed to permit an interactive visualization to filter data by year and to permit the user to select the statistical inputs for the x-axis and y-axis of the linear regression (ChatGPT.com, 2025). This provides a custom view of each regression, which permits the user to explore the relationships of various statistics for each year.



With the given sample size there was a positive correlation between each statistics per year. This indicated that teams that excelled in certain areas, such as total hits, also performed well in other areas, such as total runs. Analyzing total hits versus total runs there was a strong positive correlation among the four years. Taking a look at 2019 specifically, there was the widest spread among data points suggesting there may have been a strategic shift in following years that led to a more balanced approach offensively.

To further this analysis, the project team also constructed a Sunburst chart of aggregate MLB data, which permits hierarchical investigation of the various statistics from the league level down to the individual player level. This provides a unique view of this data and information via an interactive custom visualization that organizes players and teams by team color (ChatGPT.com, 2025). For the purposes of the project team's analysis, this feature permitted the team to drill down on data points observed in the regression plot, which helped better understand how the data and statistics were organized across teams and divisions. While this visualization did not provide any additional insights or conclusions, it helps create a more diverse user experience in

the dashboards built for the project team's public facing website, which was deployed at https://calebmeinke.pythonanywhere.com/

Data Limitations & Bias:

Assessment of the MLB data included in the original datasets revealed several biases and limitations. First, the data containing the Hitting Leaders dataset contained a limited number of top performing players from each MLB team. The absence of performance data for the full rosters of these teams created a direct impact and bias in the statistical analysis of the performance data. This is most observable in the regression plots associated with the MLB database, which show positive correlations between all analyzed stats. In example, an analysis of strikeouts against home runs shows a positive correlation in this data; however, assessment of broader MLB performance data shows a negative relationship between these statistics when full team rosters are considered (MLB.com, 2025).

Additionally, the data analyzed has implicit limitations due to the restricted sample size of players from each team. While the full dataset does contain over 500 entries of player stats over the four year period included, player counts range from one to six players per team. As such, the ability to pull meaningful team performance data from the database is limited. This greatly swings the team performance metrics from one year to the next, which causes observable volatility in the team rank and team performance for each year.

Finally, the MLB data in this database is further limited by the inclusion of 2020 MLB performance data. Due to the COVID-19 pandemic the 2020 baseball season was abbreviated, which resulted in much fewer instances of each statistic measured. This has a notable impact on the data analysis of individual statistics and their relationships during the 2020 season.

These biases and limitations will need to be addressed in future iterations of the MLB Data app to ensure it adequately considers broader MLB metrics. Doing so will ensure an accurate and complete assessment of team stats and the statistical relationships of these performance measures. Such a change will make the app a more interactive and dynamic resource for users, which will provide deeper and more meaningful insights into the expansive connections between MLB statistics.

Conclusion:

In conclusion the data between 2019 and 2022 expressed a strong positive correlation between various baseball metrics, especially total hits vs. total runs and home runs vs. strikeouts. This was mostly due to the sample size of the data set consisting of top hitters per team. A more complete data set with full team rosters would likely showcase some negative correlations between certain statistics offering a more accurate picture of overall team performances.

Examining home run totals across the four years showed no single dynasty dominating team. While the Boston Red Sox appeared in the top three in two of the four years, the top home run producing team varied season to season.

Looking at the geographical location of each team and considering altitude change it was clear that there is not a strong influence of geographical location or altitude affecting the overall total home runs per team. This suggested that factors such as team strategy, player skill, and potentially ballpark dimensions play a more significant role in the overall team performance.

Future analysis would be best to incorporate a more recent data set that does not include a pandemic year and includes a full team roster. This will help provide a more accurate representation of league-wide trends. Expanding the dataset can also assist with predictive analytics for future team performance and potential championship outcomes.

References

- ChatGPT Coding Assistant, https://chatgpt.com/, 2025
- ESPN, https://www.espn.com/mlb/worldseries/history/winners, 2025
- MLB, https://www.mlb.com/, 2025