Project Documentation

Data Source

This project utilizes data obtained from the MLB-StatsAPI, a comprehensive source for Major League Baseball statistics. The API documentation can be found here: MLB-StatsAPI.

Data Retrieval and Cleaning

To facilitate data analysis, the pybaseball Python package was used. This package provides access to a wide array of baseball data, including statcast data, pitching stats, batting stats, division standings/team records, and awards data. The package scrapes data from popular sources such as Baseball Reference, Baseball Savant, and FanGraphs, offering data at the individual pitch level, season level, and custom time periods. Detailed information on the package can be found here: pybaseball.

Data Cleaning Process

The following code outlines the process of fetching and cleaning the data. This process involved downloading yearly data, saving it to CSV files, and then combining these files into a single dataset for further analysis:

Summary of Data Cleaning Steps:

1. **Fetching Data:** Data was fetched for each year from 2018 to 2024 using the pybaseball package.
2. **Saving Data:** Each year's data was saved separately into CSV files to manage the large dataset size effectively.
3. **Combining Data:** All yearly CSV files were read and concatenated into a single DataFrame.
4. **Saving Combined Data:** The combined DataFrame was saved as statcast_data_2018_2024.csv for further analysis.

This process was necessary due to the large size of the dataset (over 2 gigabytes). By breaking down the data into yearly chunks and then combining them, it ensured efficient handling and processing.

Tableau Dashboards

Link to dashboard:
https://public.tableau.com/app/profile/vanellsa.acha/viz/Baseball_17216957934700/RunExpectancy

The cleaned and combined data was used to create interactive visualizations in Tableau. The following dashboards were created to provide insights into various aspects of baseball performance:

1. **Run Expectancy by Pitch Type and Team:**
    - **Objective:** To analyze how different pitch types affect run expectancy for various teams.
    - **Features:**
        - Heatmap displaying run expectancy changes by pitch type for each team.
        - Filters to select specific teams and pitch types for detailed analysis.

        - Detailed charts showing average delta run expectancy for selected teams and pitch types over time.

2. **Actual Batting Average vs. Expected Batting Average:**
    - **Objective:** To compare the actual batting average of players against their expected batting average.
    - **Features:**
        - Scatter plot showing the relationship between expected and actual batting averages.
        - Detailed charts breaking down performance by batter/pitcher pairs.
        - Filters to select specific game years and teams for targeted analysis.
3. **Pitching Effectiveness Metrics:**
    - **Objective:** To evaluate various pitching metrics for different teams.
    - **Features:**
        - Scatter plot comparing multiple pitching metrics (e.g., BB/9, ERA, FIP) across teams.
        - Line chart showing trends of key pitching metrics over multiple seasons.
        - Parameter control to select different metrics for comparison.
4. **Team Pitching Metrics Over Time:**
    - **Objective:** To observe how pitching metrics have evolved over time for different teams.
    - **Features:**
        - Line charts displaying trends for metrics such as ERA, FIP, and K/9 over the years.
        - Filters to select specific teams and seasons for focused analysis.
5. **Data Dictionary:** A data dictionary was created to provide detailed descriptions and definitions of the metrics used in the analysis. This dashboard serves as a reference for understanding the various baseball metrics, ensuring that users can interpret the visualizations accurately.

Data Dictionary Features:

- **Metric Definitions:** Detailed explanations of each metric used in the analysis, including how they are calculated and what they represent.
- **Categories of Metrics:** The metrics are categorized into different groups for easy navigation and reference. Categories include Batting Metrics, Pitching Metrics, Fielding Metrics, and Advanced Metrics.
- **Search Functionality:** Users can search for specific metrics to quickly find their definitions.

- **Interactive Elements:** Filters and dropdowns allow users to select and view different categories of metrics.

Conclusion

This project utilized the MLB-StatsAPI and the pybaseball package to gather comprehensive baseball data. Through effective data cleaning and processing, the data was prepared for visualization in Tableau. The resulting dashboards provide valuable insights into run expectancy, batting averages, and pitching effectiveness, aiding analysts in making informed decisions based on the data.