A New Era of Data Analysis in Baseball

This Project requires that you know your way around Python, pandas, and data visualization. We recommend the following courses as prerequisites:

- **Intermediate Python**
- **Introduction to Data Visualization with Matplotlib**
- **Introduction to Data Visualization with Seaborn**

MLB.com's Statcast glossary (MLB stands for Major League Baseball) may be helpful at various points *throughout* the Project. Through accessible text and video, they explain baseball concepts in more detail than the Project Notebook. Links to specific glossary pages will be provided throughout the Project.

The below are the tasks solved in notebook. Try them on your own before referring the notebook:

Task 1: Instructions

Load the CSV files, which hold the Statcast data for each player, into pandas DataFrames.

- Load `datasets/judge.csv` into a DataFrame and assign it to the variable `judge`.
- Load `datasets/stanton.csv` into a DataFrame and assign it to the variable `stanton`.

Task 2: Instructions

Display the last five rows of the `judge` DataFrame.

- Use pandas' `tail` method to display the last five rows of `judge`.

Task 3: Instructions

Isolate each player's batted ball events for the 2017 season.

- Filter `judge` to include pitches from 2017 only and select the `events` column. Store the result in a variable called `judge_events_2017`.
- Using the `value_counts` method, print out the count of unique values for `judge_events_2017`.

- Filter `stanton` to include pitches from 2017 only and select the `events` column. Store the result in a variable called `stanton_events_2017`.
- Using the `value_counts` method, print out the count of unique values for `stanton_events_2017`

Task 4: Instructions

Isolate each player's home runs then plot exit velocity vs. launch angle.

- Filter the `judge` and `stanton` DataFrames to include home runs only.
- Create a figure using seaborn's `regplot` function with two scatter plots of launch speed vs. launch angle, one for each player's home runs.
- Create a figure using seaborn's `kdeplot` function with two KDE plots of launch speed vs. launch angle, one for each player's home runs.

Task 5: Instructions

Plot the pitch velocities of each player's home runs on box plots.

- Concatenate `judge_hr` and `stanton_hr` using pandas' `concat` function and store the result in a variable called `judge_stanton_hr`.
- Create a boxplot using seaborn's `boxplot` function that describes the pitch velocity of each player's home runs. Make the color argument 'tab:blue'.

Task 6: Instructions

Create a function that returns the x-coordinate of a pitch zone.

- Return the x-coordinate for the left third of strike zone.
- Return the x-coordinate for the middle third of strike zone.
- Return the x-coordinate for the right third of strike zone.

While you should ignore zones 11, 12, 13, and 14 for this plotting task, setting up conditionals to filter these out now isn't necessary. That will come in an upcoming task!

`zone` is the name of the column that holds each pitch's zone data.

Task 7: Instructions

Create a function that returns the y-coordinate of a pitch zone.

- Return the y-coordinate for the upper third of strike zone.
- Return the y-coordinate for the middle third of strike zone.
- Return the y-coordinate for the lower third of strike zone.

While you should ignore zones 11, 12, 13, and 14 for this plotting task, setting up conditionals to filter these out now isn't necessary. That will come in an upcoming task!

`zone` is the name of the column that holds each pitch's zone data.

Task 8: Instructions

Assign Cartesian coordinates to the strike zone and plot pitches that resulted in Judge home runs as a 2D histogram.

- Apply `assign_x_coord` to `judge_strike_hr` to create a new column called `zone_x`.
- Apply `assign_y_coord` to `judge_strike_hr` to create a new column called `zone_y`.
- Plot Judge's home run zone as a 2D histogram (using matplotlib's `hist2d` function) with a colorbar.

Task 9: Instructions

Assign Cartesian coordinates to the strike zone and plot pitches that resulted in Stanton home runs as a 2D histogram.

- Apply `assign_x_coord` to `stanton_strike_hr` to create a new column called `zone_x`.
- Apply `assign_y_coord` to `stanton_strike_hr` to create a new column called `zone_y`.
- Plot Stanton's home run zone as a 2D histogram (using matplotlib's `hist2d` function) with a colorbar.

Task 10: Instructions

A few takeaways:

- Stanton does not hit many home runs on pitches in the upper third of the strike zone.
- Like pretty much every hitter ever, both players love pitches in the horizontal and vertical middle of the plate.

- Judge's least favorite home run pitch appears to be high-away while Stanton's appears to be low-away.
- If we were to describe Stanton's home run zone, it'd be middle-inside. Judge's home run zone is much more spread out.

The grand takeaway from this whole exercise: Aaron Judge and Giancarlo Stanton are not identical despite their superficial similarities. In terms of home runs, their launch profiles, as well as their pitch speed and location preferences, are different.

Should opposing pitchers still be scared

Answer the following question: "Should opposing pitchers be wary of Aaron Judge and Giancarlo Stanton?"

- Store a Boolean value (True or False) in `should_pitchers_be_scared`.