Code attached contains a Python script that when run produces a five (5) cluster classification containing daily batters within the major leagues. The clusters are the result of 2 sequential k-means Clustering algorithms.

Batting data was taken from the website Baseball Reference:

<https://www.baseball-reference.com/leagues/MLB/2019-standard-batting.shtml>

This data was saved into an Excel file titled “batting\_stats\_2019.xlsx” (also attached)/

The methodology can be described in five (5) distinct sequential phases:

1. Data Acquisition

2. Data Pre-processing

3. k-means Clustering

4. Cluster Identification of half-time to full-time players (played at least 81 games in 2019 season)

5. k-means Clustering of selected clusters of full-time players to determine “value” of player

6. Plotting of the individual data points to two distinct graphs

7. Export of Excel spreadsheet containing the data

From this data, different clusters can be retrieved and then analyzed for different purposes.

The output contains:

1. Two (2) plots:
   1. everday\_clusters.png displays a two-dimensional projection of the final clustering of each player
   2. everyday\_stats.png displays offensive data stats of batting average, homeruns and games played with each point color coded for the cluster to which and belongs
2. An excel spreadsheet with statistics for each player as well as cluster designations, denoted in the final two (2) columns as “Cluster” and “Cluster2”.

To run, place both the Python script and the “batting\_stats\_2019.xlsx” Excel file into the same directory and execute the example Python code file named “JeremyGross\_ExampleCode.xlsx”.