# Using a Similarity Algorithm to Visualize And Compare Baseball Players

Clint Appleseth (cappleseth3), Daniel Jiang (djiang49), John Welt (jwelt3), Mike Wolfgram (mwolfgram3)

## Introduction

Sabermetrics, the study of baseball analytics and statistics, has been around for a long time. However, there is no publicly available tool equipped with a great user-friendly interface and visualization. Also, most methods of comparing players require the user to manually select players, then compare their stats side-by-side in a tabular format. Our approach uses principal component analysis, a dimensionality reduction technique, and three similarity measures to allow for easier comparisons among multiple players as well as a nice visualization of key player stats.

## Data

We used R to acquire and clean data from several sources:

- **Chadwick Baseball Bureau** to get player IDs.
- **MLB API** for player attributes for all players who have played in MLB in the past 10 years.
- **Fangraphs** for batting statistics.

After downloading and cleaning data for 1986-2019, we created a SQL Server database on AWS. With the addition of a table containing all similarity measures, the database has 6 tables with a total of 868 columns and 410,685 rows.

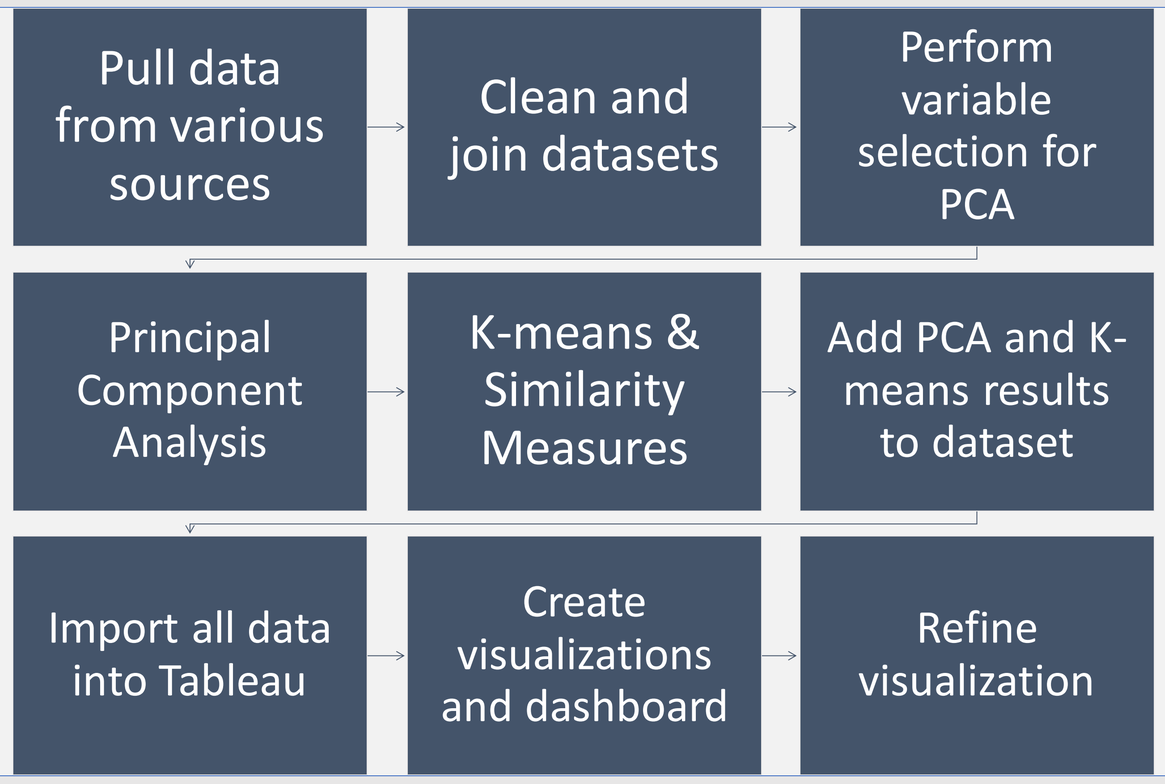
## Approach

Our approach centered around using Principal Component Analysis to reduce multiple offensive statistics down to two principal components and then create three similarity measures between each player. This approach improves over some of the current similarity algorithms by taking in multiple metrics instead of just comparing one or two stats. Our approach also focused on stats that are good predictors of a player's offensive potential rather than traditional statistics that can be influenced by factors outside a player's control.

After acquiring and cleaning the data, we built a database with 6 tables joined by player IDs. We then chose 22 statistics that centered around the player's offensive profile to use for Principal Component Analysis and subsequent K-means clustering.

Next, we created similarity measures between each player based on the distance between the two principal components. We included three similarity measures between each point: Euclidean Distance, Manhattan Distance, and Cosine Similarity.

K-means was used to group the data points into similar clusters for visualization purposes. The last step was to import the data into Tableau to create an interface that allows users to select a player and compare that player against 5 similar players suggested by the algorithm.




## Visualization and UI

The visualization has four major components, which are labeled in the below screenshot in red:

**1** – Control panel, which lets the user select the baseball season, minimum number of plate appearances, position filter, similarity measure, and the cluster legend.

**2** – Similarity plot, which charts the results from the PCA algorithm and gives a visual of the next closest/similar players. The user can hover over each data point to see the player's name and position. Selecting a player will filter the dashboard for that player, and the most similar players to him, ranked in order from most similar to least.

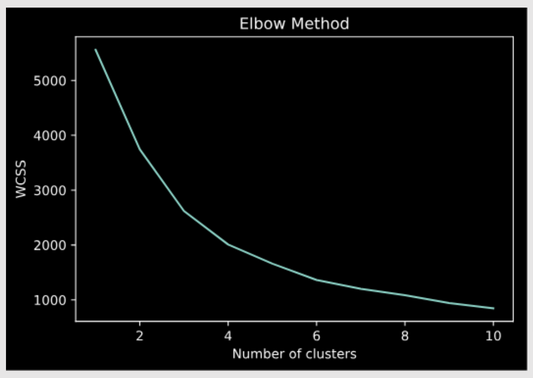
**3** – A panel which lists the selected player, pulls the player's image from the MLB website and displays general stats for that player and 5 similar players.

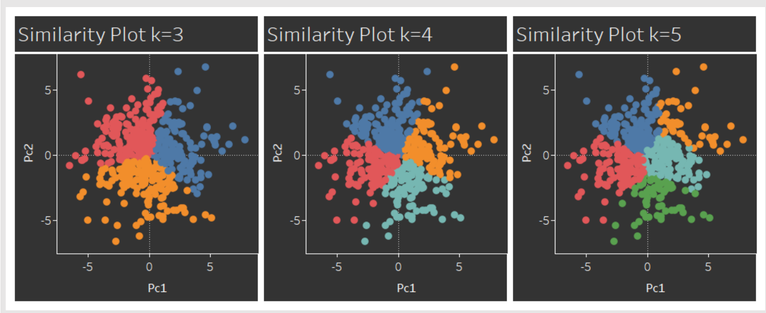
**4** – The fourth section displays the offensive profile and plate discipline stats for each of the players from the third section. This allows for a quick visual comparison of all statistics, allowing the user to easily compare multiple similar players at once.

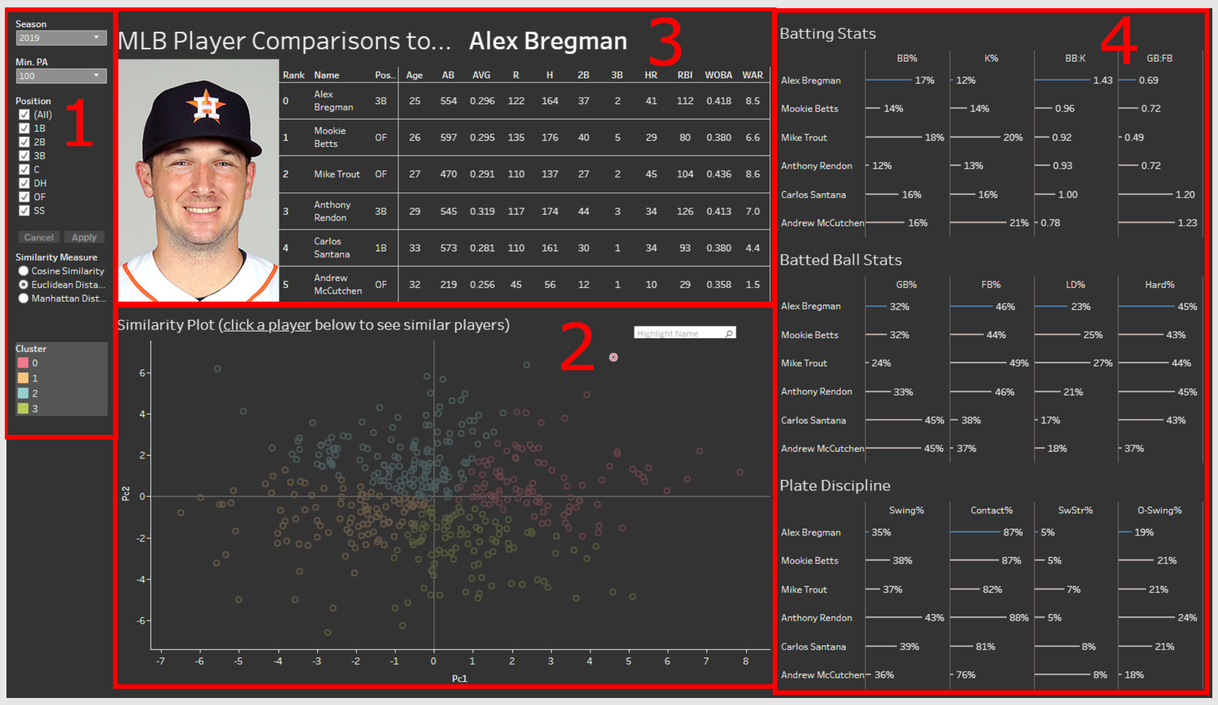
## Experiments and Results

**Determining the number of clusters with k-means**

To select the value *k* with the k-means algorithm, we used a scree plot to identify clusters with a low Within Center Sum of Squares (WCSS) that did not overfit the data. Using the scree plot, we tested 3, 4, and 5 clusters and decided that *k=4* provided the best visualization.




**Power User Questionnaire**

We gave 3 fantasy-baseball users access to the tool and asked them to experiment with it and complete a questionnaire. Each participant has 10+ years of experience managing fantasy baseball teams. Below is a summary of their feedback:

- **The filtering functionality could be streamlined to improve experience**
- Colors used in scatterplot are not color-blind friendly
- **Provide an explanation of Principal Components Analysis and the different similarity measures in layman's terms**
- **Each user preferred a different similarity measure, and they liked having the option to use different similarity measures**
- Allow the user to compare players across multiple seasons
- Show more than 5 comparable players
- Display some additional batting stats (wRC+, BABIP, OBP, ISO) and remove WAR because it's fielding related
- Shows the actual similarity value so you can see how much more similar player 2 is than player 3
- **Overall, users believe the tool could be useful for Fantasy Baseball**