Pitch Clustering with K-means and Data Visualization with t-SNE

Enrique J. Montanez

University of Florida

MAS 4115 - Linear Algebra for Data Science

Johnathan Bush

April 30, 2024

Pitch Clustering with K-means and Data Visualization with t-SNE

In this project, I clustered different NCAA Division 1 baseball pitchers' pitch types using k-means. After clustering, I used t-SNE to reduce the dimensionality of my clusters, in order to better visualize them in a 2D plot. Then, for each cluster, I calculated the mean values of different performance metrics, and visualized the distribution of performance metric values within each cluster. Lastly, I made predictions on which cluster a specific pitcher's inputted pitch characteristics would fall into. I was inspired to complete this project by a feature on Baseball Savant, which allows one to find similar pitchers based on velocity and movement. The data I used in this research was from the 2023 NCAA baseball season, and was provided to me by 6-4-3 Charts.

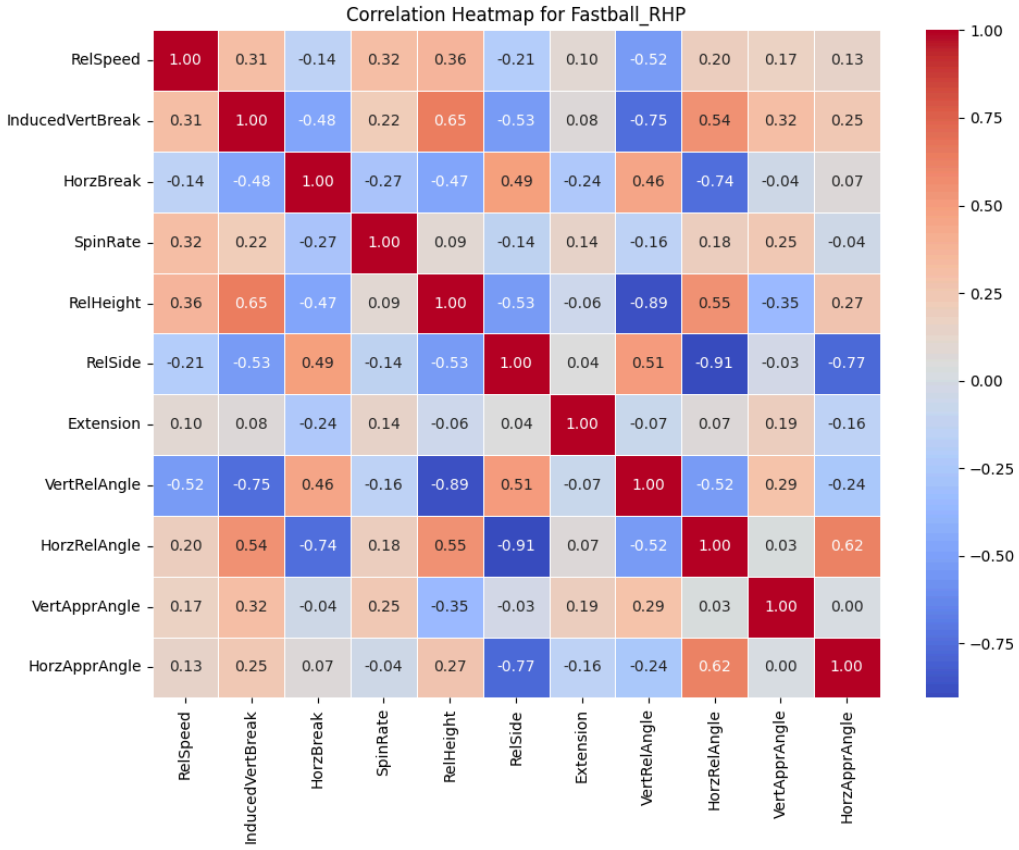
Research Process

Data Cleaning

After loading in my data, I first had to clean it to make sure it was in the proper format for analysis. I used dictionaries called 'merged_RHP_stats' and 'merged_LHP_stats' to store data frames grouped by pitcher name and pitch type, with each row containing each pitch's characteristics, such as velocity, vertical movement and horizontal movement, as well as performance metrics associated with the pitcher for the 2023 season, such as SIERA, FIP, and K-BB%. The data frames in 'merged_RHP_stats' included only right-handed pitchers and the data frames in 'merged_LHP_stats' included only left-handed pitchers. I only included pitchers who threw at least 24 innings in the 2023 season.

Feature Selection

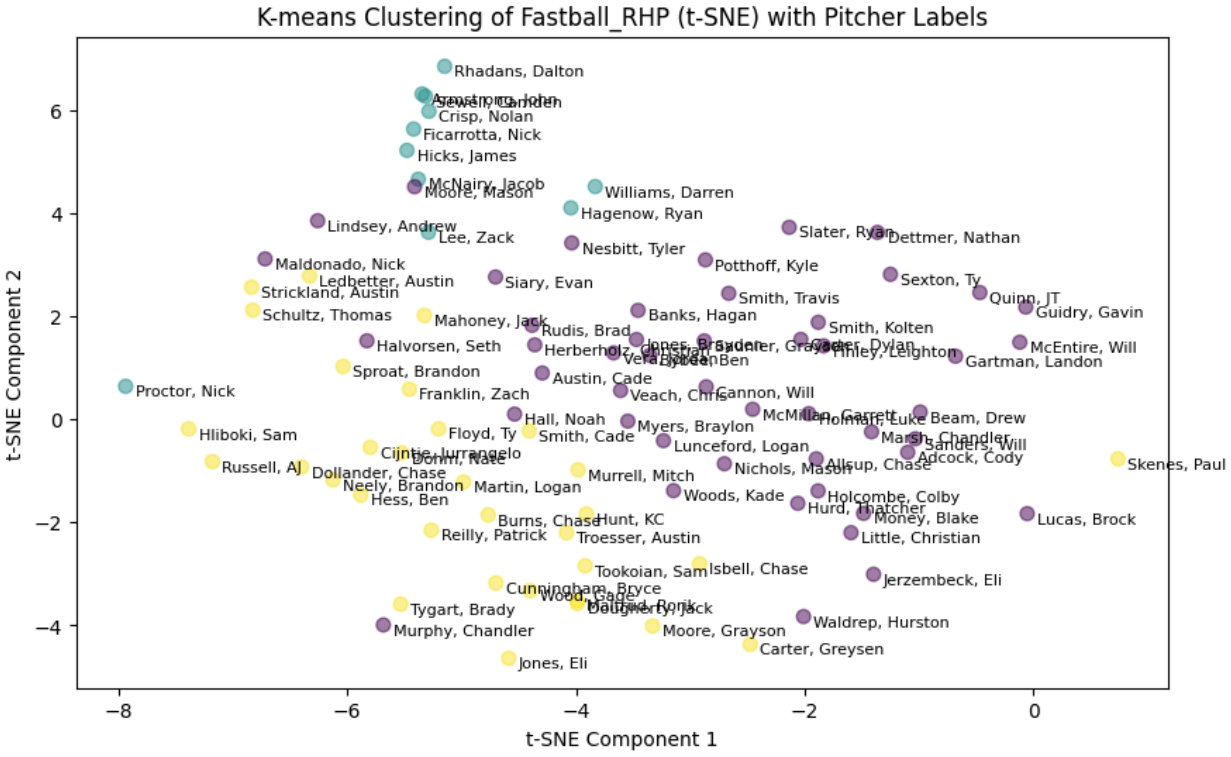
In order to select which features to use for the clustering of each pitch type for each handedness, I created correlation matrix heatmaps to help detect the presence of multicollinearity for all of the dataframes. In order for a feature to be included, the absolute value of its correlation



coefficient with another feature had to be less than 0.7. If two features had a correlation coefficient whose absolute value was 0.7 or greater, then only one of those features would be included. For example, in the figure above, Since the absolute value of the correlation coefficient between InducedVertBreak and VertRelAngle is 0.75, only InducedVertBreak is used in the clustering.

K-Means Clustering

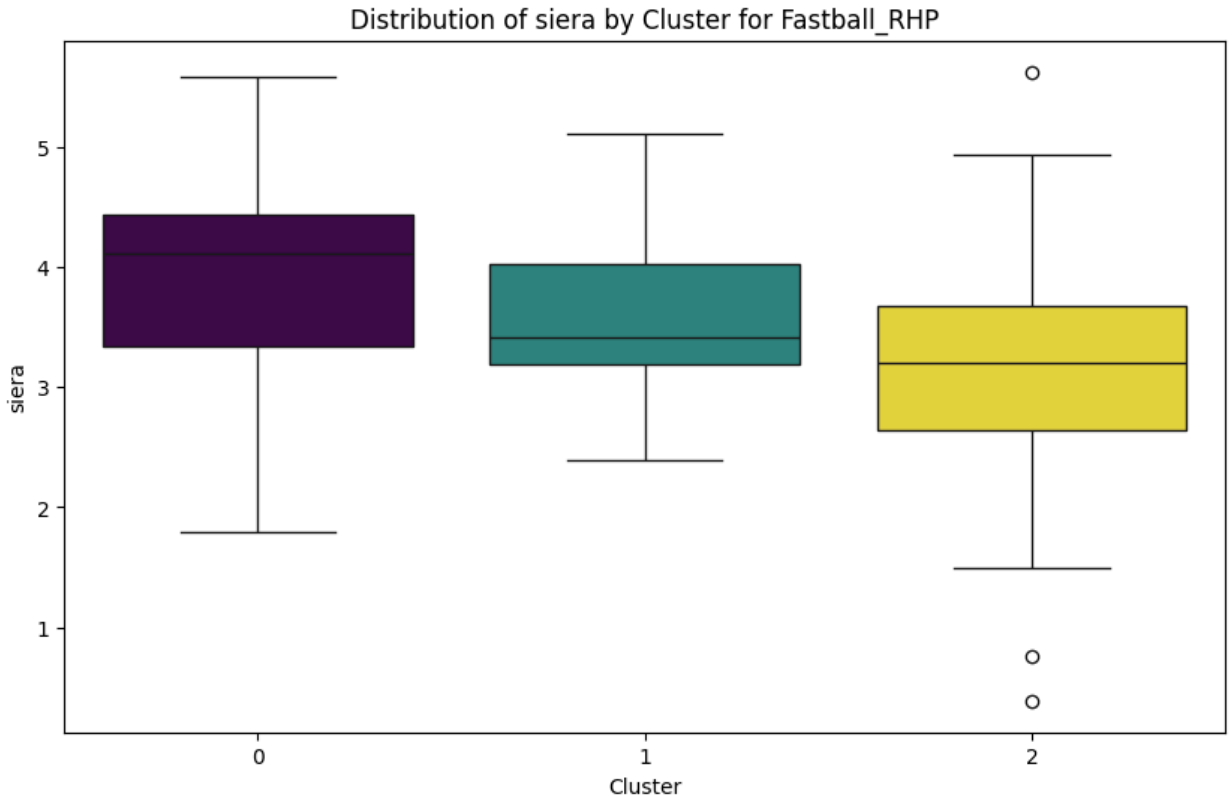
After selecting the features I was going to use, I used K-means to cluster each pitch type for each pitcher's handedness. I selected K to be three by finding the “elbow” in a plot of the Within-Cluster Sum of Squares (WCSS) against corresponding K values between one and ten on which the K-means algorithm was run. Next, 3 initial cluster centers are initially assigned, and pitches are assigned to them based on how close they are. After pitches are initially assigned to clusters, the center of each cluster is recalculated, and pitches are reassigned based on this new center. These steps are repeated until the cluster centers do not change significantly or a certain number of iterations is reached.



Data Visualization using t-SNE. Since these clusters were calculated based on multiple features, it would be impossible to visualize them without some sort of dimensionality reduction technique. For this reason, I used t-SNE to reduce the dimensionality of the clusters to 2D. The resulting t-SNE components do not have direct physical meaning like the original features.

Instead, the t-SNE components represent abstract coordinates in the lower-dimensional space where similar points are grouped together. The distances and relationships between points in the t-SNE embedding reflect their similarities in the high-dimensional space.

Evaluating Clusters with Performance Metrics. In order to learn more about what the clusters meant, I calculated the mean value of different performance metrics for each cluster and made box plots to visualize the distribution of the performance metric values across the three clusters for each data frame.



SIERA is a metric that tries to encapsulate pitcher performance, where the lower the value is, the more successful the pitcher has been. In the figure above, we can see that right-handed fastballs from cluster 2, the yellow group, tend to have lower SIERA values.

Making Cluster Predictions. After gaining some more insight on the meaning of each of the clusters, I wanted to be able to input any pitcher's pitch characteristics to see what cluster he would fall into.

Case Study: Bryce Miller. Bryce Miller is a right-handed pitcher for the Seattle Mariners with an electric fastball, so I wanted to see if he would fall into cluster two, the cluster which tended to have lower SIERA values, based on his fastball characteristics.

```
# Input features for Fastball_RHP
bryce_miller = [[95.0, 18.4, 6.1, 2438.0, 5.7, 1.3, 6.4, -4.2]]

# Use the previously trained scaler to scale the input features
scaled_input_features = scalers_RHP['Fastball_RHP'].transform(bryce_miller)

# Use the previously trained KMeans model to predict the cluster
predicted_cluster = kmeans_models_RHP['Fastball_RHP'].predict(scaled_input_features)

# Print the predicted cluster
print("Predicted cluster for Fastball_RHP:", predicted_cluster)
```

```
Predicted cluster for Fastball_RHP: [2]
```

As shown in the figure above, Miller did indeed fall into cluster two, which helped support the integrity of my clusters, as his fastball does perform very well in the major leagues.

Results

After strenuous data cleaning and multiple attempts at using different methods to find similarities in pitches, I am pleased with the possible insights I can gain from these clusters and can think of two use cases for this research.

Use Case 1

First, these clusters can be used in player acquisition and scouting. Scouts, coaches, and analysts can see which clusters each of their possible recruits fall into based on their pitch types and handedness, and can use the information to decide which pitchers to target.

Use Case 2

Second, teams can use these clusters for game preparation. If a team is facing a pitcher who they have never seen before, they can use his pitches and pitch characteristics to find similar pitchers of the same handedness who they have faced before in order to have a better idea of what to expect at the plate.

Next Steps

I would like to continue working on this project to develop some type of application where teams can input a pitcher's pitch characteristics for any handedness and pitch type and see which cluster he would fall into. I would like to create my clusters on more data in past years to improve them, and allow users to see, for example, the top five most similar pitches to the input pitch based on their similar characteristics and place in the cluster.

Click here for Google Colab Link.

References

Baseball Savant: Statcast, trending MLB players and visualizations. baseballsavant.com. (n.d.). https://baseballsavant.mlb.com/

Chamberlain, A. (n.d.). *Pitch Leaderboard v6*. Public.tableau.com. https://public.tableau.com/app/profile/chamb117/viz/PitchLeaderboardv6/Dashboard

Starmer, J. (2017, September 18). *StatQuest: T-sne, clearly explained*. YouTube. https://youtu.be/NEaUSP4YerM?si=ynX3LbPuCBaTOQLz

Starmer, J. (2018, May 23). *StatQuest: K-means clustering*. YouTube. https://youtu.be/4b5d3muPQmA?si=roHU8got_I-AWikH