# Analysis of Player Performance by Age

Insights from Logistic Regression and Correlation Analysis

# Introduction

## Research Question

- MLB players are known for having long careers (sometimes 20+ seasons).
- Do MLB batters have unique styles based on their age?

## Goals of Analysis

- Analyze MLB batting metrics that specifically focus on decision-making and style of play.
- Attempt to quantify whether playing styles shift with loss of athleticism.

## Hypothesis

- H1: MLB batting styles are unique to age groups.
- H2: As players get older, they demonstrate better decision-making and judgment.

# Data Overview

- All MLB players with a minimum of 10 at-bats (ABs) per season from 2013-2023.
- Data pulled from Fangraphs via pybaseball module.
- Each season was pulled as individual DataFrame, then merged in Spark.
- The dataset was preprocessed to handle missing values and categorical variables.
- Over 7,500 eligible players/season (row values).
- 1,702,715 total at-bats analyzed.
- 35 columns examined out of 321 total possible.
    - Focus on metrics involving decision-making and style per at-bat.

# Age Group Distribution

**Group 1:** 19-23

**Group 2:** 24-28

**Group 3:** 29-34

**Group 4:** 35-39

**Group 5:** 40-44

# Linear Regression

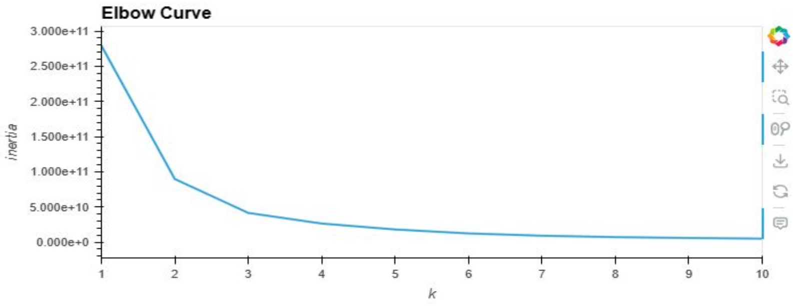
# Linear Regression and MSE

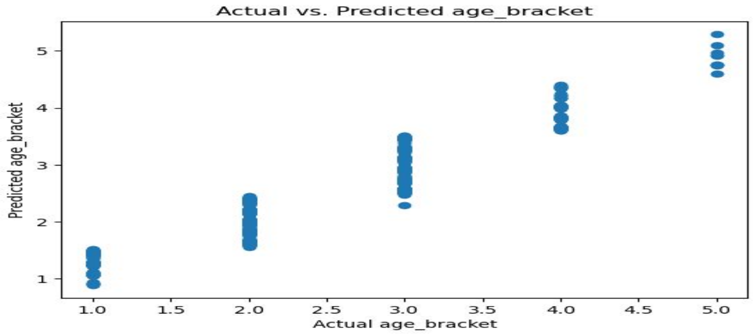
From the elbow curve, it could be noted that the best K-mean value is 3. Plots were done although with the K-mean values of 3 and 4 to compare the output.The variance ratio of the first principal component is approximately 99.76%.

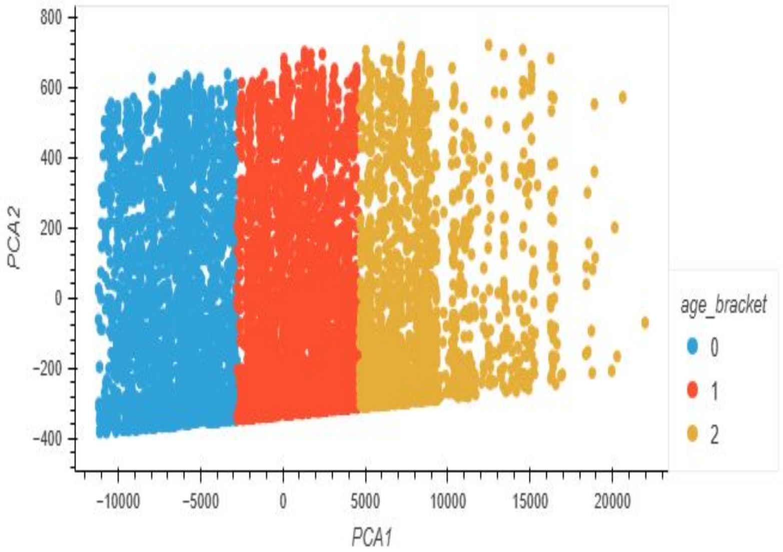
PCA dimensionality reduction on the dataset can be done with just the first principal component since the dataset is highly skewed towards the first principal component indicating a strong underlying patterns or structure of the dataset.

A mean squared error value of 0.0864958441960215 which is lower and closer to zero for this analysis predicting the target variable, which is the age bracket, indicates the model is very accurate.

# Linear Regression analysis using K value of 3 to show three clusters







# Logistic Regression

# Logistic Regression

- Utilized multinomial regression model to account for multiple age groups.
- Model first achieved 50.7% accuracy.
- Various optimization efforts plateaued accuracy at around 53% fit.

- Confusion matrix and classification reports were produced.
- Model performs strongest in identifying players aged 24-28.
- Struggles with other age brackets, particularly older players.

# Attempts to Improve Model Accuracy

## Change Age Ranges

**Decreased Accuracy**

- Consolidated age brackets to 4.
- Reorganized distribution so that 3 of the brackets. were nearly equal
- Accuracy of the model dropped to 36%.

## Increase Iterations

**Marginally Increased Accuracy**

- Increasing max iterations improved the accuracy of the model marginally.
- Tops out at around 1000 iterations.

## Normalization and PCA

**No Impact on Accuracy**

- Only improved model accuracy <2%.
- Reducing to 20, 15, 10, and 5 components yielded nearly identical results.

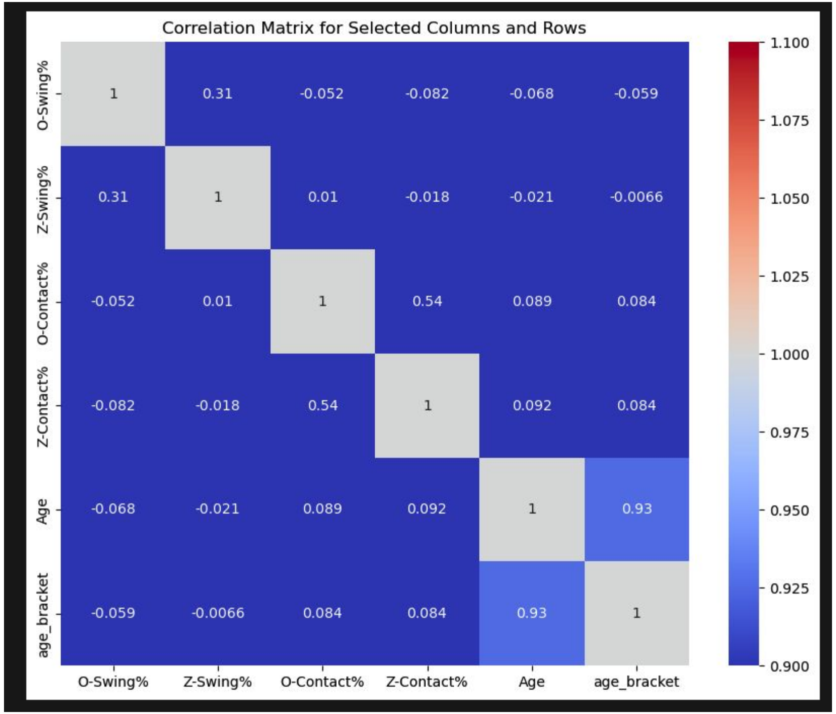
# Random Forest

# Random Forest

- Initial random forest model was able to achieve an accuracy score of 52%.
- Optimization Attempts:
  - Added feature columns that were previously dropped
  - Adjusted the number of estimators, increased by 300.
- After attempting to optimize the random forest model, we were able to obtain an accuracy score of 54%
- Similar to the other models we built, the random forest performs strongest in identifying players aged 24-28.
- Struggles with other age brackets, particularly older players.

# Correlation Analysis

Measuring the strength on ‘Age’ and ‘age_bracket’ between other variables in the dataframe



Correlation Matrix for Selected Columns and Rows

| | O-Swing% | Z-Swing% | O-Contact% | Z-Contact% | Age | age_bracket |
|---|---|---|---|---|---|---|
| O-Swing% | 1 | 0.31 | -0.052 | -0.082 | -0.068 | -0.059 |
| Z-Swing% | 0.31 | 1 | 0.01 | -0.018 | -0.021 | -0.0066 |
| O-Contact% | -0.052 | 0.01 | 1 | 0.54 | 0.089 | 0.084 |
| Z-Contact% | -0.082 | -0.018 | 0.54 | 1 | 0.092 | 0.084 |
| Age | -0.068 | -0.021 | 0.089 | 0.092 | 1 | 0.93 |
| age_bracket | -0.059 | -0.0066 | 0.084 | 0.084 | 0.93 | 1 |

- Shows some negative correlation on some key metrics.
  - From this we can deduce that players actually do make fewer questionable decisions as they get older

- Although we could make speculations based on the data, the correlations were so weak that reducing the data down to those columns wouldn’t make a significant impact.

# Conclusions

- Logistic regression model achieved around 53% maximum accuracy.
- Attempts to rearrange age groups and apply PCA/normalization did not significantly improve accuracy.
- Random forest analysis yielded similar accuracy results.
- Weak negative correlations observed in key metrics like decision-making with age, but not impactful enough for classification.
- Linear regressions and low mean square error indicate strong predictive value in overall player contribution (WAR), but limited classification value.

# Q&A

Ian Summers, Felix Ologo-Gyan, Jesse Olivarez & Christopher McCormick