Baseball Metrics and their Determinants

Driving Question:
Does the amount of home games played affect baseball offensive or defensive metrics more?

Micayla Fong

Table of contents

01 Data Description

02 Literature

03 Testing

04 Follow up

05 Testing

06 Results and Conclusion

Our Data

To address our driving question we looked at the past ~120 years of baseball offensive and defensive stats of each baseball team.

Offensive:

- Batting Avg
- Runs
- Hits
- Walks
- Double
- Triples
- etc

Defensive:

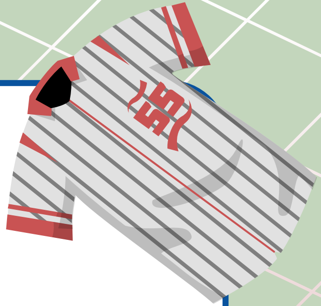
- ERA
- Shutouts
- Hits allowed
- Strikeouts
- Runs allowed
- Walks allowed

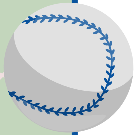
Linear Regression

Each Linear regression was run with a predictor of home games played for the given season. The displayed per game variables are the top statistically significant for both offense and defense out of all the variables tested.

Triples	*Runs Scored*	*Runs allowed*	*ERA*	*Walks allowed*
p-val: 0.042 coeff: 0.554	p-val: 0.061 coeff: 2.46	p-val: 0.004 coeff: -3.87	p-val: 0.017 coeff: -3.41	p-val: 0.034 coeff: -2.23

Interpretation



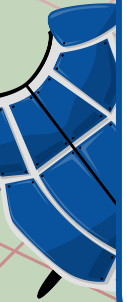


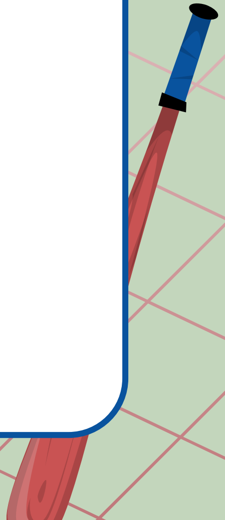
P values:

- Indicate that the home game ratio has more of an affect on defensive metrics than offensive metrics
- Some p values suggest statistically significant relationships between hg ratio and chosen metrics

Coefficients:

- The positive coefficients of triples and runs scored indicate that higher hg ratio tends to mean higher triples and runs
- The negative coefficients of ERA, runs allowed, and walks allowed indicate that higher hg ratio tends to mean lower ERA, runs and walks allowed





Adjusted R-Square:

- While some of the p values suggest statistical significance, the R-squares were low meaning that while the hg ratio may have a relationship, it’s practical effect on these metrics is minimal

Conclusion

(to driving question)

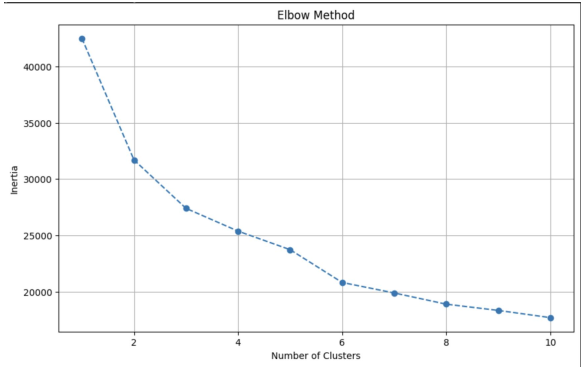
- Although the results suggest relationships between home game ratio and the offensive and defensive metrics (emphasis on defensive), the overall impact on these metrics proved to be small

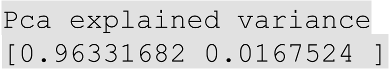
- This small of an effect is not enough to advise any changes in coaching or team strategy to improve either offensive or defensive production

Follow up:

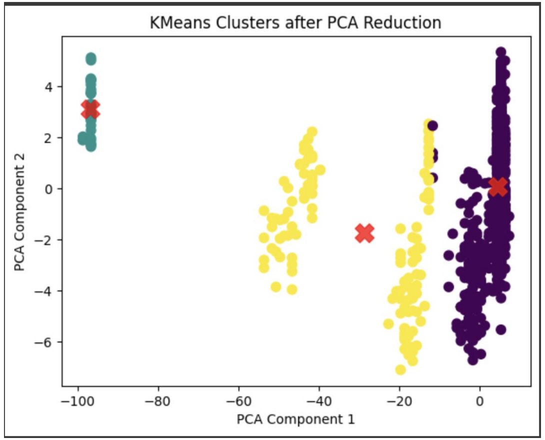
If home game ratio isn't affecting these metrics, what distinct team performance profiles emerge in baseball when analyzing the relationships between offensive production and defensive efficiency?

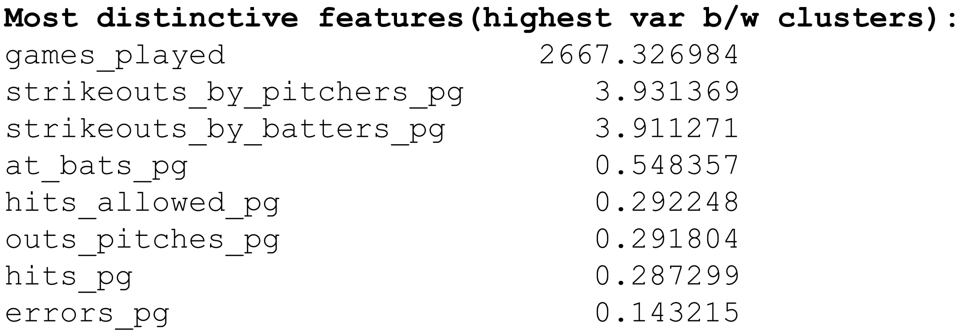
K-means





```
Pca explained variance
[0.96331682 0.0167524 ]
```

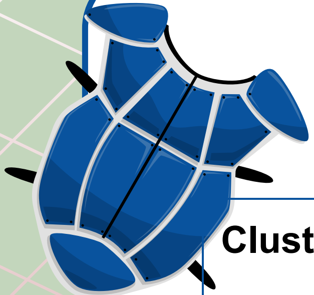


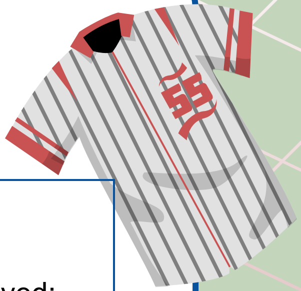


```
Most distinctive features(highest var b/w clusters):
games_played                  2667.326984
strikeouts_by_pitchers_pg        3.931369
strikeouts_by_batters_pg         3.911271
at_bats_pg                       0.548357
hits_allowed_pg                  0.292248
outs_pitches_pg                  0.291804
hits_pg                          0.287299
errors_pg                        0.143215
```

We reduced our features from 27 to 2

Interpretation





Cluster 1 **1418 members**

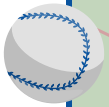
- Most games played: ~161 games
- Moderate strikeout rates by pitchers (~6.1 per game)
- Moderate strikeout rates by batters (~6.1 per game)
- Batting average: .258
- Higher number of at-bats per game (~34.0)

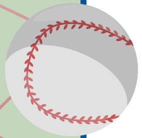
Cluster 2 **30 members**

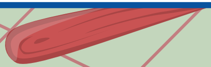
- Very few games played: ~60 games
- Highest strikeout rates by pitchers (~8.7 per game)
- Highest strikeout rates by batters (~8.7 per game)
- Lowest batting average: .244
- Highest home run rates (~1.28 per game)
- Fewest errors per game (~0.58)

Cluster 3 **126 members**

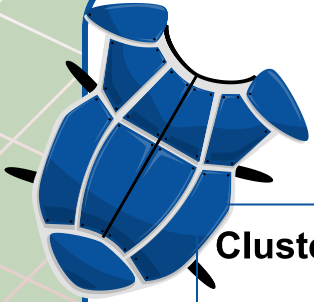
- Moderate games played: ~128 games
- Lowest strikeout rates by pitchers (~4.8 per game)
- Lowest strikeout rates by batters (~4.8 per game)
- Highest batting average: .265
- Most hits per game (~9.1)
- Most errors per game (~1.32)

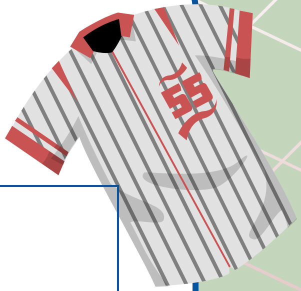




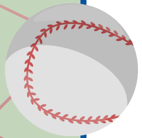


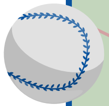
Interpretation

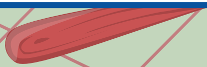




Cluster 1	Cluster 2	Cluster 3
● These represent regular full-season teams with typical baseball statistics ● Largest cluster (1418 members) suggesting this is the "norm"	● Characterized by extremely high strikeout and home run rates ● Very small cluster (30 members) suggests these are outliers ● Less games played might explain abnormally high rates	● Highest batting average, more hits, fewer strikeouts and home runs ● More errors ● Represents a more contact-oriented style of baseball







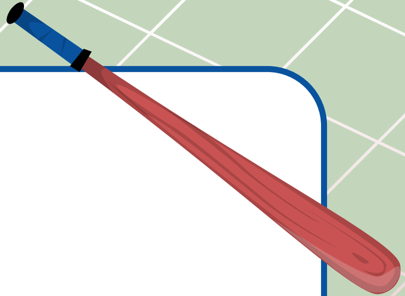
Team Profiles

- One common theme we have seen between clusters is that strikeouts and home runs have a positive relationship when it comes to offensive production.

- Another relationship we see between clusters is that strikeouts by pitchers and strikeouts by batters have a positive relationship.

- We also see that a higher batting average is characterized by less home runs and less strikeouts.

- Lastly we see that the teams with the highest batting averages tend to have the highest error rates in the field.

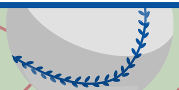


Conclusion

Causes of Profile Differences:

- We attribute some differences to how playing style may have changed over time due to the wide range of years of the selected data,
- Changes in coaching styles/ differences in team strategy

- **Strikeouts/batting average vs home runs**: this relationship could be due to the fact that batters trying to hit homeruns, often have bigger swings and are less focused on contact, therefore leading to more missed pitches and therefore strikeouts, however this is pretty standard knowledge when it comes to offensive strategy in baseball.



Offensive on Defensive

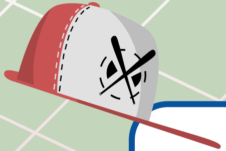
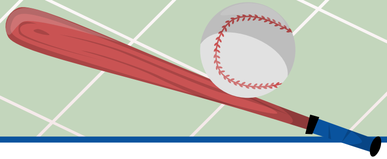
- **Strikeouts**: this offensive vs defensive metric relationship is common in all 3 of the clusters and would be an area of further research to determine the cause of this relationship so teams may be able to use this to their advantage

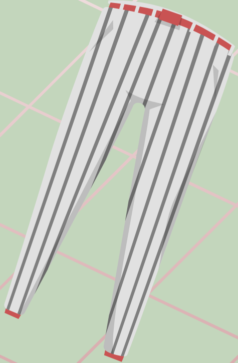
- **Batting average and error rates:** we also see a positive relationship of batting average and error rate in all three clusters. Another area of further research for teams to improve error rates and maybe shift focus from their bp to defensive training or vice versa

Anova

We used anova to identify which features show statistically significant differences between the clusters found in k-means. The higher f-statistics indicate greater variances between the groups. The lower p-values indicates stronger statistical significance

```
ANOVA Results (most significant differences):
                            F-statistic       p-value
fielding_percentage_pg     18736.816379  0.000000e+00
games_played                8871.794288  0.000000e+00
outs_pitches_pg              181.691770  1.040376e-71
errors_pg                    114.747947  3.061054e-47
strikeouts_by_batters_pg     111.933802  3.580371e-46
strikeouts_by_pitchers_pg    111.346058  5.989883e-46
triples_pg                    83.404598  3.756219e-35
```



Thank you!