Baseball Stats

Anya Lee, Kathryn Stewart, Annika Ström

Background Information

- Since its conception, baseball has grown in popularity, and is important to America’s history
- Before football, baseball was known as America’s sport
- Generated about $11.34 billion in 2023 alone
  - $378 million made for each team

Dataset Overview

Data from 1871 to 2015

- Downloaded dataset on Kaggle that has 20 total files, but selected the ones below

Batting.csv

- 101,332

Pitching.csv

- 44,139

Fielding.csv

- 170,526

Salaries.csv

- 25,575

Data cleaning performed on situation basis (ie. log transformations, BB%, SO%, other ratios)

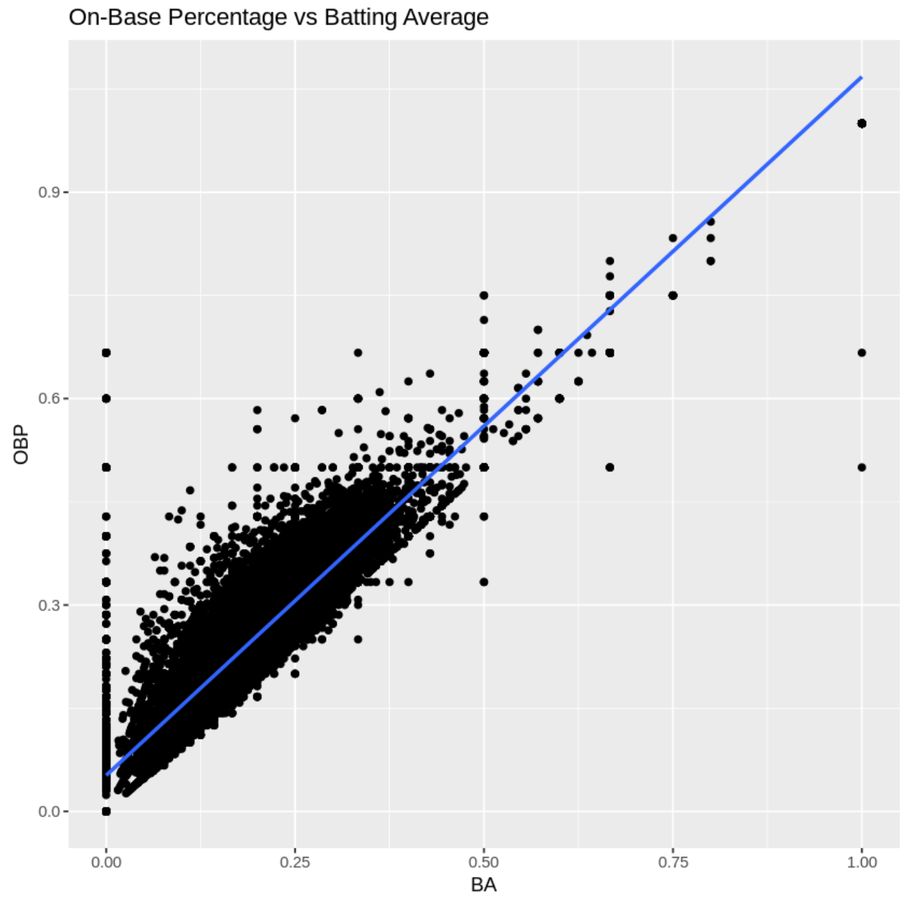
- Depending on research question

Research Questions

- Batting
  - Which predictors are most relevant to batters striking out?
- Pitching
  - Which predictors are most relevant to pitching ERA?
  - Which predictors are most relevant to pitchers getting strikeouts?
- Salary, Pitching, and Fielding
  - What position makes the most?
  - What are the most relevant predictors for pitching salary?
  - What model best predicts the salaries based on hitting/batting?

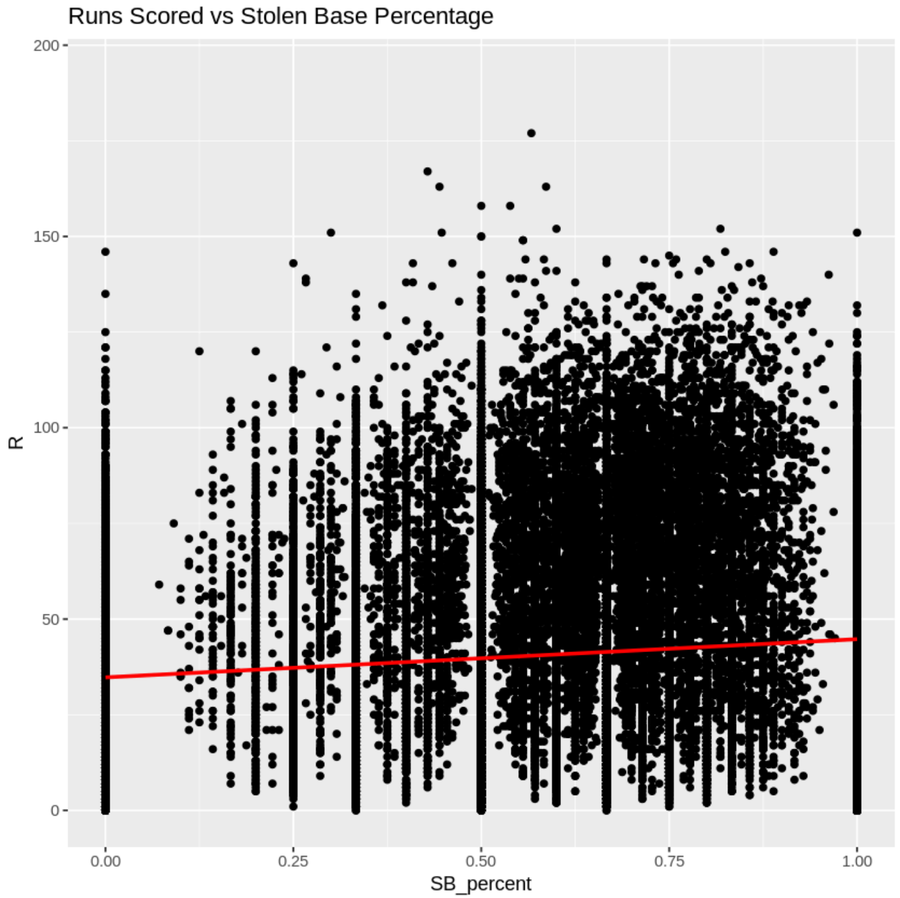
Batting

Batting Average (BA) vs On-Base-Percentage (OBP)



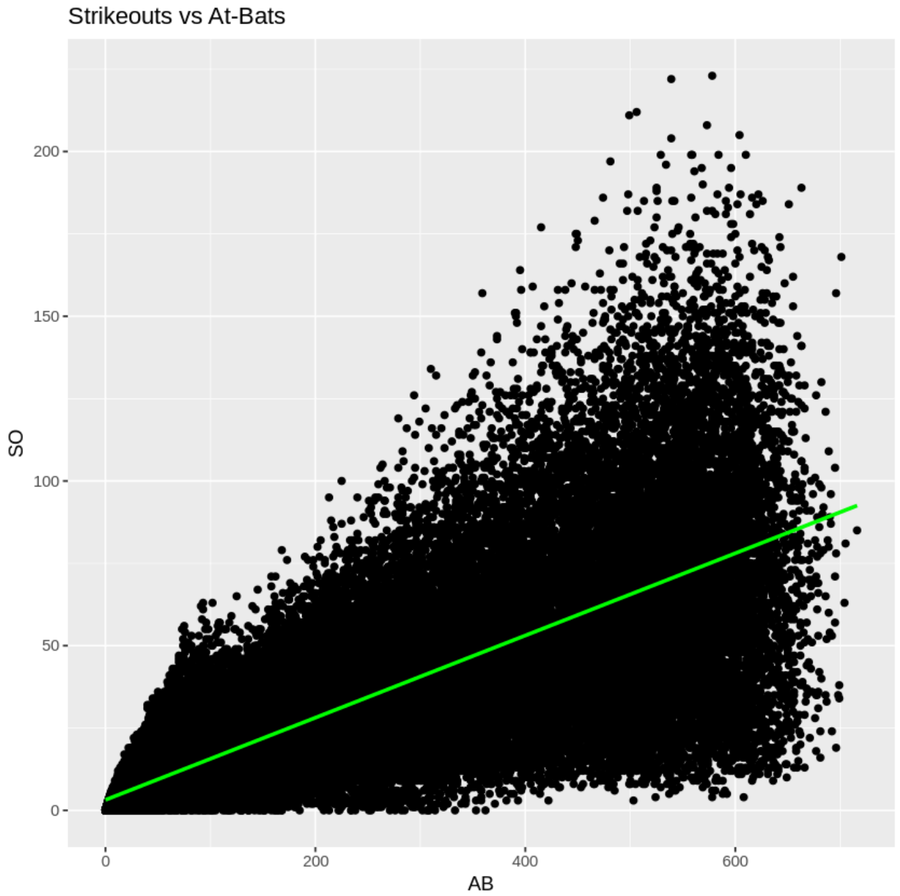
- Positive linear relationship between OBP and BA represented by the line OBP = 1.015544*BA + 0.052297
- Reasonable with Adjusted R-Squared value of 0.8393, which is pretty high
- Interpretation:
  - For every one unit increase in BA, the OBP of a batter increases by 1.015544
- This is what we expect!

Stolen Base Percentage (SB) and Runs Scored (R)



- Data has a significantly higher variance than the previous graph
- A linear model does not fit the data well since it is so scattered (high variance!)
- Hitters with a relatively higher stolen base percentage (< 50%) also scored more runs since many of the data points are clustered between 50% and 100%

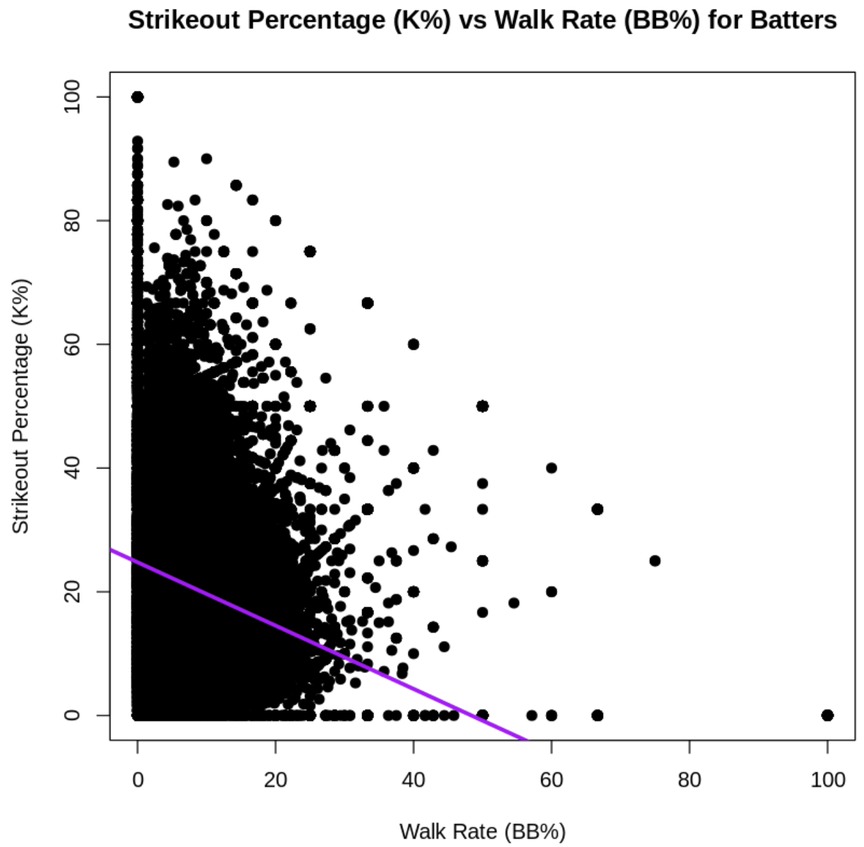
Strikeouts (SO) and At-Bats (AB)



Does the number of At-Bats have an effect on number of Strikeouts?

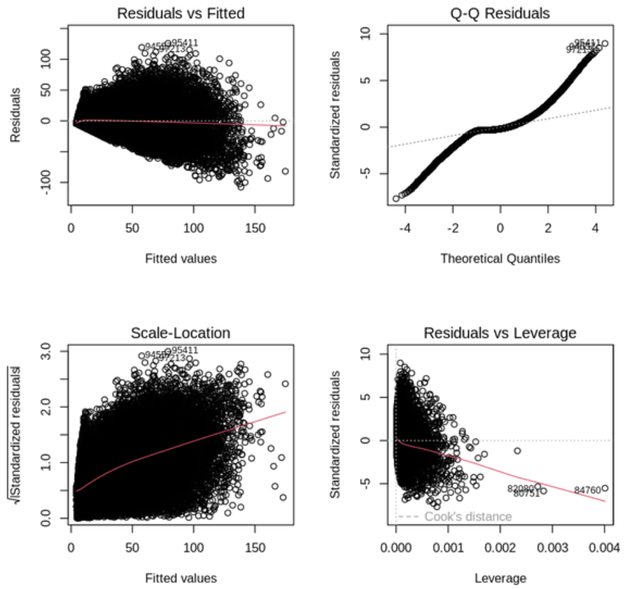
- Data points funnel out, which indicates a larger variance
  - Adjusted R-Squared value of 0.6709 with a positive trend
- If the linear model was an accurate representation, we can interpret this as:
  - For every one increase in At-Bats, the number of Strikeouts a batter will have will increase by 0.1247909
- This makes sense to us - more ABs means more opportunity to SO

Strikeout Percentage (K%) and Walk Rate (BB%) for Batters

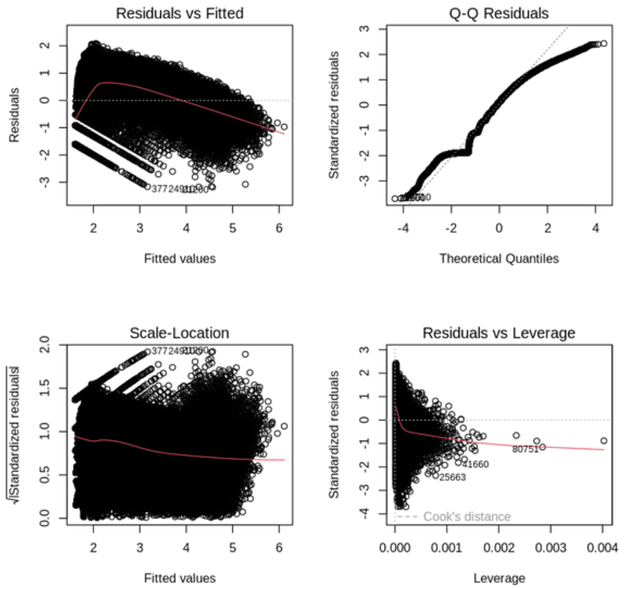


- From the graph, we see a negative trend.
  - The more a batter strikes out, the less opportunities a batter has to achieve a walk
- This is logical
  - Batters that strike out more become more picky and selective at the plate when they hit
  - Correlation does not imply causation
- There are other confounding variables to consider here.
  - At-Bats can result in a hit, which is neither a strikeout or a walk

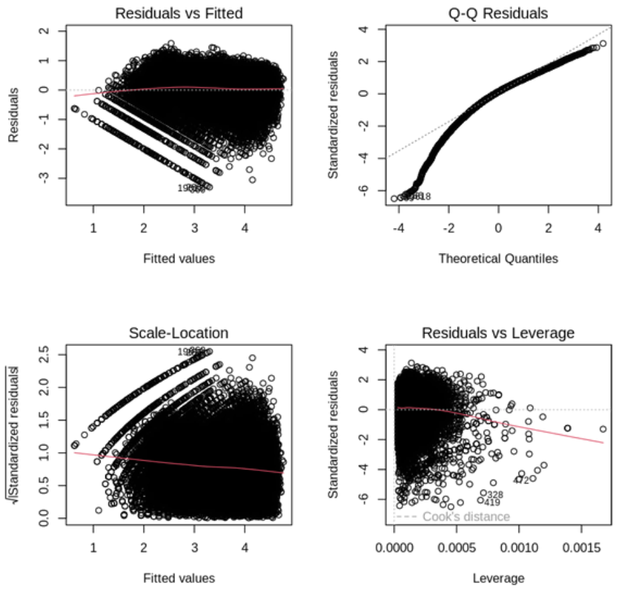
Predicting Strikeouts by a Batter: Checking linear regression assumptions



Original



Log transform response



Log transform response & predictors

We tried fitting the following MLR model

SO = beta_0 + beta_1*AB + beta_2*HR + beta_3*BB + error_term

Predicting Strikeouts by a Batter

A data.frame: 10 × 4

Predictors	AIC	BIC	Adj_R2
<chr>	<dbl>	<dbl>	<dbl>
SO, AB	744007.9	744036.1	0.6709030
SO, AB, HR	717409.3	717446.8	0.7564642
SO, AB, HR, H	690856.0	690903.0	0.8196878
SO, AB, HR, X2B, H	688275.4	688331.7	0.8248806
SO, AB, HR, X2B, H, GIDP	545449.3	545513.4	0.8492198
SO, AB, HR, X2B, RBI, H, GIDP	544697.2	544770.5	0.8508326
SO, AB, G, HR, X2B, RBI, H, GIDP	544564.4	544646.8	0.8511175
SO, AB, G, BB, HR, X2B, RBI, H, GIDP	544539.9	544631.4	0.8511716
SO, AB, G, BB, HR, X2B, X3B, RBI, H, GIDP	544532.5	544633.2	0.8511894
SO, AB, G, BB, HR, R, X2B, X3B, RBI, H, GIDP	544523.0	544632.9	0.8512118

544522.980179923

A data.frame: 1 × 4

Predictors	AIC	BIC	Adj_R2
<chr>	<dbl>	<dbl>	<dbl>
10 SO, AB, G, BB, HR, R, X2B, X3B, RBI, H, GIDP	544523	544632.9	0.8512118

544631.448698064

A data.frame: 1 × 4

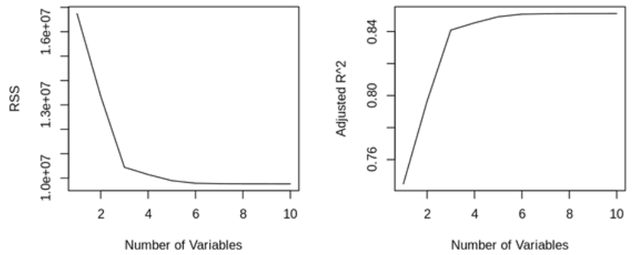
Predictors	AIC	BIC	Adj_R2
<chr>	<dbl>	<dbl>	<dbl>
8 SO, AB, G, BB, HR, X2B, RBI, H, GIDP	544539.9	544631.4	0.8511716

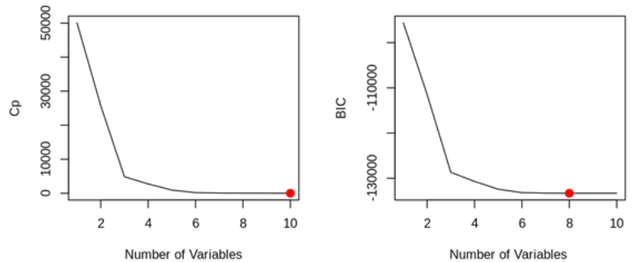
0.851211813297913

A data.frame: 1 × 4

Predictors	AIC	BIC	Adj_R2
<chr>	<dbl>	<dbl>	<dbl>
10 SO, AB, G, BB, HR, R, X2B, X3B, RBI, H, GIDP	544523	544632.9	0.8512118

- Started with hitting predictors:
  - Games Played, Games Started, Outs Pitched, Batters Faced by Pitcher, Earned Run Average, Batting Average Against, Homeruns Allowed, Walks Allowed, Complete Games, Shutouts

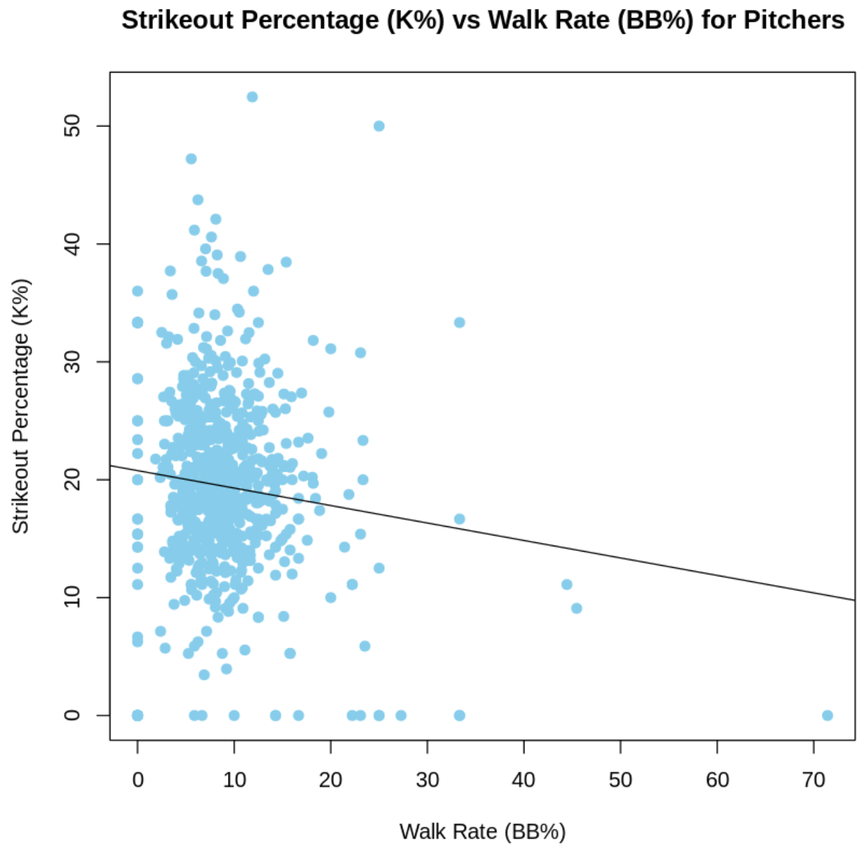




- Used regsubsets
- Calculated AIC, BIC, and Adjusted R-Squared for each model
- Results
  - Model 10 for AIC and Adjusted R-Squared
    - Contained all 10 predictors
  - Model 8 for BIC

Pitching

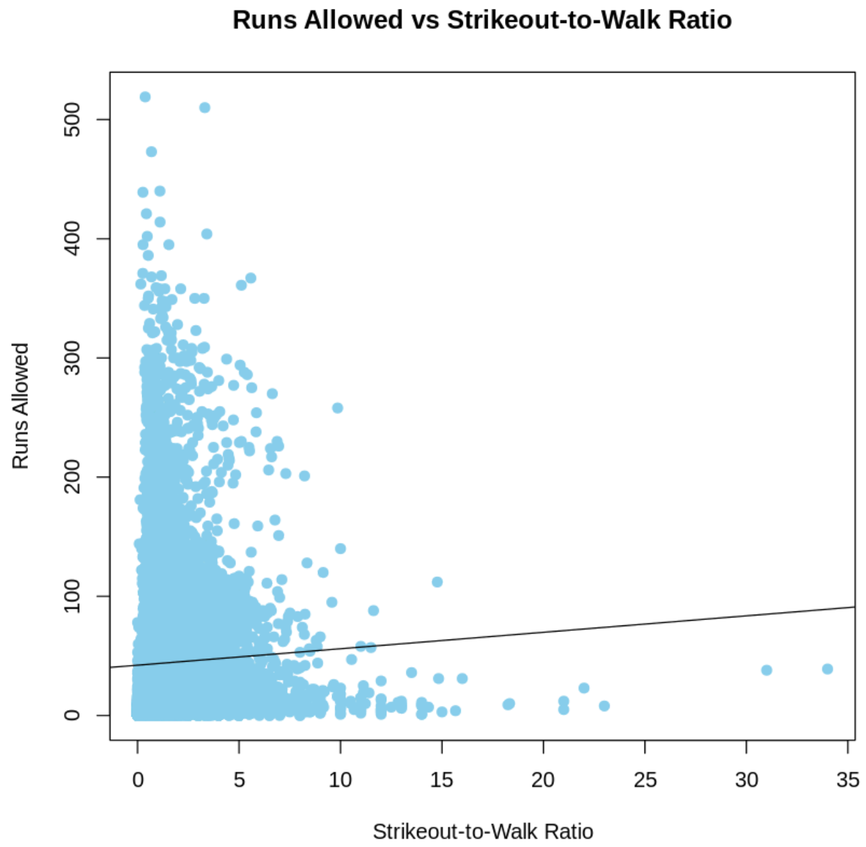
Strikeout Percentage and Walk Rate



- Observe a cluster of points
  - Not best represented using a linear model with extremely low Adjusted R-Squared of 0.01079
  - Model here only captures 1% of the variance in the data (after cleaning)

- Most data points are clustered
  - Walk Rate between 0% and 20%
  - Strikeout Rate between 10% and 30%
  - Cannot conclude there is correlation between Walk Rate and Strikeout Rate

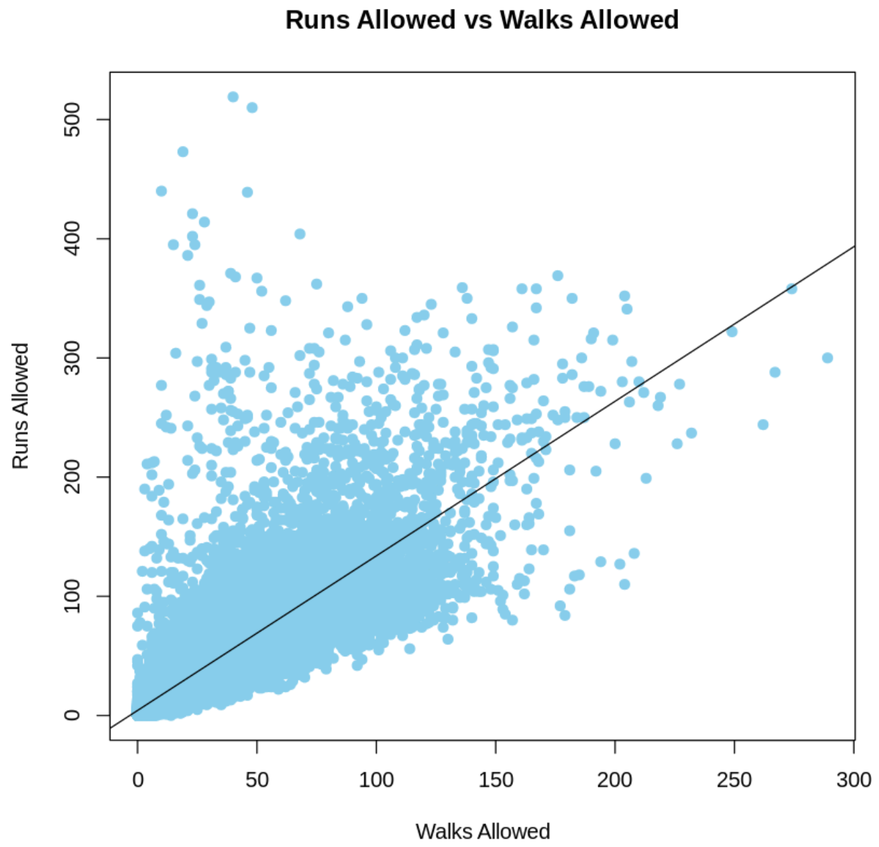
Strikeout to Walk Ratio Relative to Runs Allowed



- Expected a higher Strikeout-to-Walk Ratio to result in a lower number of Runs Allowed by pitchers
  - Some data points that do follow this intuition

- Data is concentrated within a Strikeout-to-Walk Ratio of 0 and 5, where there are higher numbers of Runs Allowed
  - Follows the inverse of our expectation. More people allowed on base means a higher chance for base runners to score a run

Runs Allowed (RA) vs Walks Allowed (WA)



- Linear model: RA=4.266236+1.296938 *WA
  - Captures the positive linear trend pretty well
  - Adjusted R-Squared value of 0.7134

- Reasonable to say that number of Walks Allowed increases the number of Runs Allowed by a pitcher
  - This makes sense why pitchers want to avoid walking batters

Predicting Strikeouts by a Pitcher

379271.884605165

A data.frame: 1 × 4

	Predictors	AIC	BIC	Adj_R2
	<chr>	<dbl>	<dbl>	<dbl>
10	SO, G, GS, IPouts, BFP, ERA, BAOpp, HR, BB, CG, SHO	379271.9	379375.8	0.814935

379375.778209785

A data.frame: 1 × 4

	Predictors	AIC	BIC	Adj_R2
	<chr>	<dbl>	<dbl>	<dbl>
10	SO, G, GS, IPouts, BFP, ERA, BAOpp, HR, BB, CG, SHO	379271.9	379375.8	0.814935

0.81494307632271

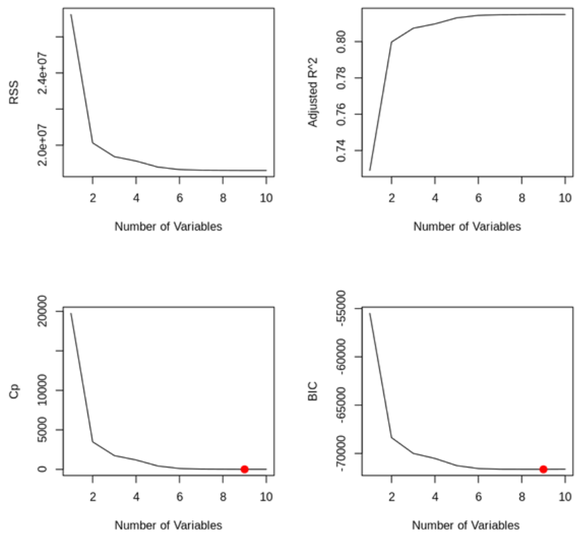
A data.frame: 1 × 4

	Predictors	AIC	BIC	Adj_R2
	<chr>	<dbl>	<dbl>	<dbl>
9	SO, G, GS, IPouts, ERA, BAOpp, HR, BB, CG, SHO	379345.8	379441	0.8149431

A data.frame: 10 × 4

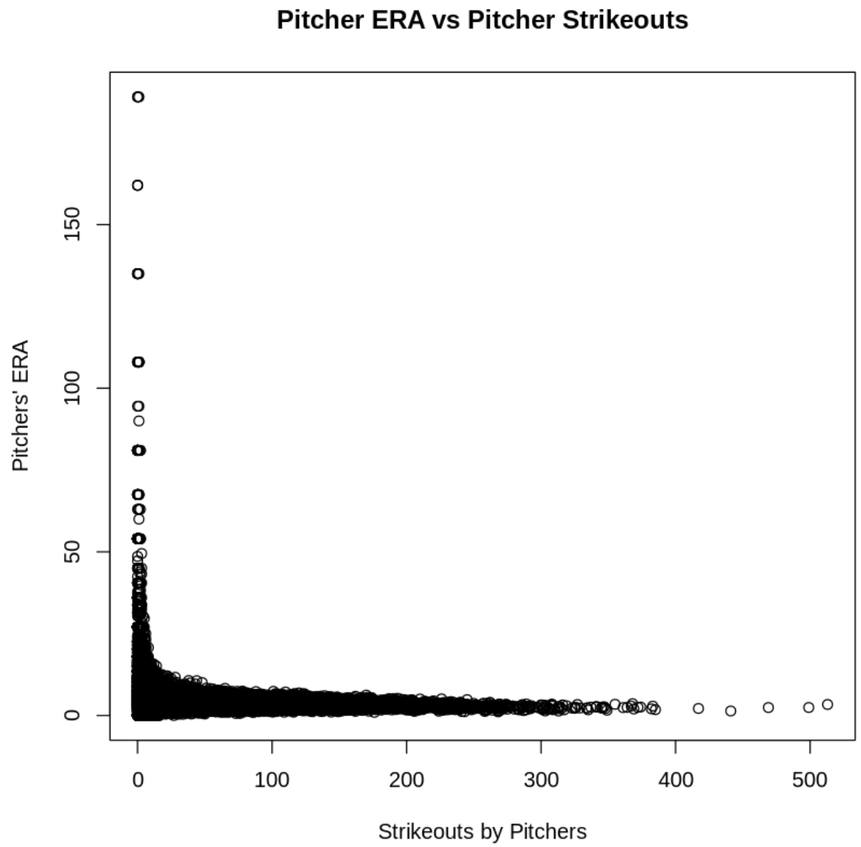
Predictors	AIC	BIC	Adj_R2
<chr>	<dbl>	<dbl>	<dbl>
SO, IPouts	413059.8	413085.9	0.7209725
SO, IPouts, CG	399817.3	399852.1	0.7933012
SO, IPouts, CG, SHO	398377.6	398421.0	0.7999392
SO, G, IPouts, CG, SHO	397831.0	397883.2	0.8024057
SO, G, IPouts, HR, CG, SHO	397068.6	397129.5	0.8057939
SO, G, GS, IPouts, HR, CG, SHO	396742.4	396811.9	0.8072284
SO, G, GS, IPouts, HR, BB, CG, SHO	396541.5	396619.7	0.8081082
SO, G, GS, IPouts, ERA, HR, BB, CG, SHO	395814.0	395901.0	0.8078407
SO, G, GS, IPouts, ERA, BAOpp, HR, BB, CG, SHO	379345.8	379441.0	0.8149431
SO, G, GS, IPouts, BFP, ERA, BAOpp, HR, BB, CG, SHO	379271.9	379375.8	0.8149350

- Started with pitching predictors:
  - Games Played, Games Started, Outs Pitched, Batters Faced by Pitcher, Earned Run Average, Batting Average Against, Home runs Allowed, Walks Allowed, Complete Games, Shutouts



- Used regsubsets
- Calculated AIC, BIC, and Adjusted R-Squared for each model
- Results:
  - Model 10 for AIC and BIC
    - Contained all 10 predictors
  - Model 9 for Adjusted R-Squared

Strikeouts vs ERA plot



- A high number of Strikeouts (by pitchers) is expected to indicate a low Earned Run Average (ERA), which is what we see in the graph!

- Pitchers that do not strikeout batters often either gave up a hit, walk, or home run.
  - Giving these up results in a higher Earned Run Average (ERA), which is also shown in the graph!

- A linear model would not fit here
  - Non-linear model such as exponential decay would likely fit the data better here

Predicting Pitchers Earned Run Average (ERA)

A anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	44044	1178409	NA	NA	NA	NA
2	44041	1131325	3	47084.73	610.9824	0

AIC: 269789.9
BIC: 269842
Adjusted R-squared: 0.05977688

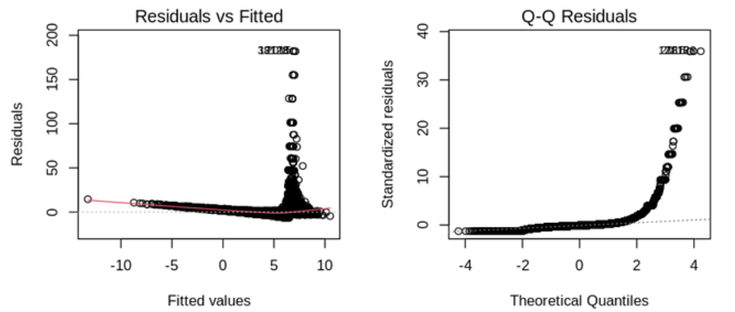
- Full model: lm(ERA ~ SO + BB + HR + IPouts + ER + H + GS)
- Reduced model: lm(ERA ~ SO + BB + HR + IPouts)

- Reject the null hypothesis (reduced model) since the ANOVA has a large F-value of 610.9824, which means the additional predictors are significant

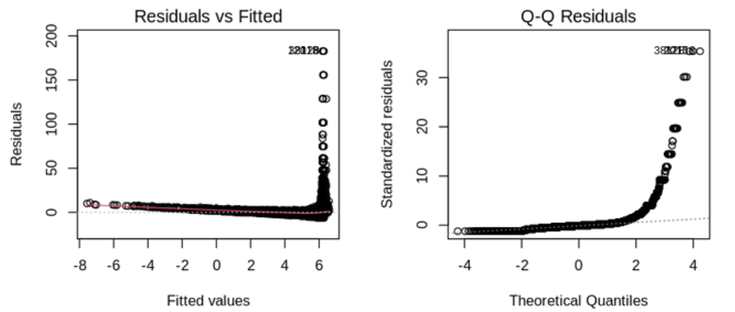
- All predictors are significant to the response variable, ERA

- Adjusted R-Squared values for both models are very low, meaning they do not understand the variance well within the data

Predicting Pitchers Earned Run Average (ERA)



Full Model



Reduced Model

Full Model: lm(ERA ~ SO + BB + HR + IPouts + ER + H + GS)

Reduced Model: lm(ERA ~ SO + BB + HR + IPouts)

Predicting Pitchers Earned Run Average (ERA)

A data.frame: 7 × 4

Predictors	AIC	BIC	Adj_R2
<chr>	<dbl>	<dbl>	<dbl>
ERA, IPouts	270121.3	270147.3	0.05261283
ERA, IPouts, ER	268919.0	268953.8	0.07814053
ERA, IPouts, ER, GS	268579.4	268622.9	0.08524143
ERA, HR, IPouts, ER, GS	268257.6	268309.8	0.09192041
ERA, BB, HR, IPouts, ER, GS	268087.3	268148.1	0.09544621
ERA, BB, HR, IPouts, ER, H, GS	268009.2	268078.7	0.09706934
ERA, SO, BB, HR, IPouts, ER, H, GS	267999.7	268078.0	0.09728311

267999.732773535

A data.frame: 1 × 4

	Predictors	AIC	BIC	Adj_R2
	<chr>	<dbl>	<dbl>	<dbl>
7	ERA, SO, BB, HR, IPouts, ER, H, GS	267999.7	268078	0.09728311

268077.970294902

A data.frame: 1 × 4

	Predictors	AIC	BIC	Adj_R2
	<chr>	<dbl>	<dbl>	<dbl>
7	ERA, SO, BB, HR, IPouts, ER, H, GS	267999.7	268078	0.09728311

0.0972831111824469

A data.frame: 1 × 4

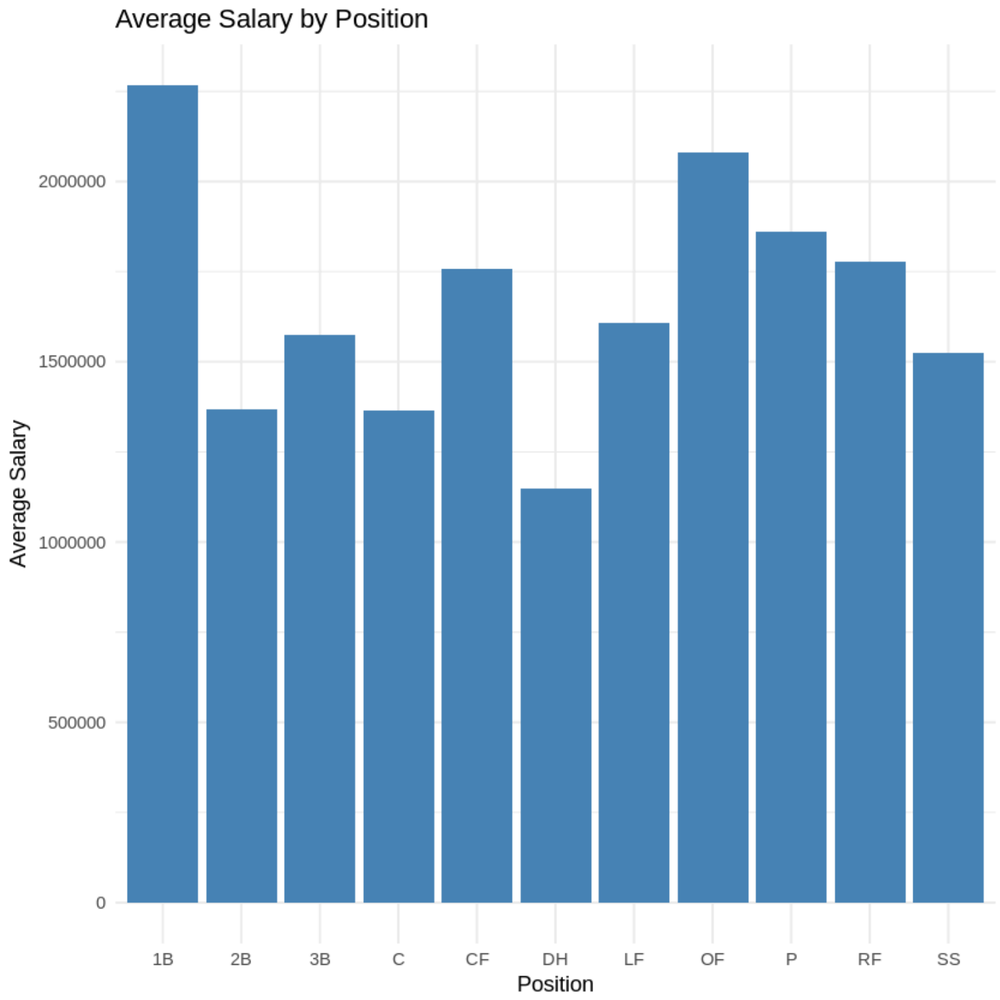
	Predictors	AIC	BIC	Adj_R2
	<chr>	<dbl>	<dbl>	<dbl>
7	ERA, SO, BB, HR, IPouts, ER, H, GS	267999.7	268078	0.09728311

- For all metrics, AIC, BIC, and Adjusted R-Squared
  - Model 7 is the best (all seven predictors)

- As seen before from the violations in assumptions, a non-linear model would be a better fit (via GLMs or GAMs) since the Adjusted R-Squared for model 7 is extremely low to be selected as the best model (0.09728)

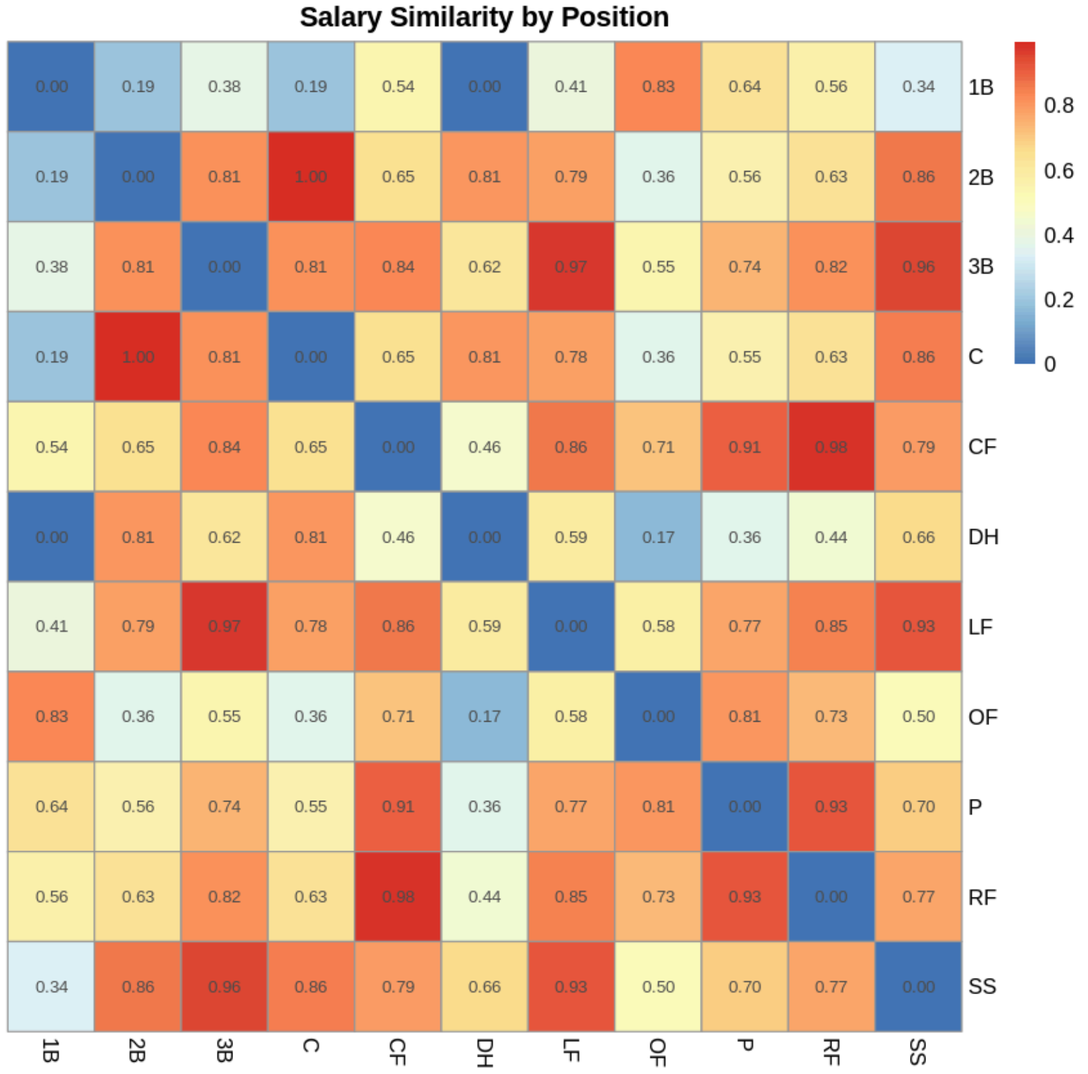
Fielding, Batting, and Salaries

Salary Difference in Positions



- Highest paid position is First Base (1B) with an average salary of $2,266,841
- Lowest paid position is Designated Hitter (DH) with an average salary of $1,149,819.
- Keep this in mind for the next slide!

Similarities between Average Salary and Fielding Position



- Most surprised that Second Base position and Catcher have a similarity of 1

- Relative to other mean differences in salary…
  - High similarity: a mean difference of $49,728 is low (as shown by 3B and SS)
  - Low similarity: a mean difference of $901,753 is large (shown by 1B and C)

- Remember the last slide?
  - The highest and lowest paid positions are 1B and DH, relatively
  - This makes sense why their similarity score is zero! They have a very large difference in mean salary of $1,117,022 (crazy!)

Predicting Salary for Pitchers via ANOVA

A anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	6204	7.237506e+16	NA	NA	NA	NA
2	6194	7.215480e+16	10	2.202621e+14	1.890801	0.04166401

- Data Cleaning
  - 11537 --> 6214 data points
- Full model: lm(salary ~ W + L + G + GS + CG + SHO + IPouts + H + ER + HR + BB + SO + BAOpp + ERA + IBB + HBP + R + SH + SF)
- Reduced model: lm(salary ~ L + GS + IPouts + ER + HR + BB + SO + IBB + R)
- Reject the null hypothesis that the reduced model is sufficient
  - Low p-value of 0.04166401 & F-statistic of 1.890801
  - These linear models are probably not the best to fit the data
    - Adjusted R-Squared value is still significantly low for the better full model (0.1829)

Predicting Salary based on Hitting/Batting (excluding pitchers)

A anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	35333	2.557933e+17	NA	NA	NA	NA
2	35329	2.557605e+17	4	3.278722e+13	1.132251	0.3391415

- Full model: lm(salary ~ AB + R + H + X2B + X3B + HR + RBI + SB + CS + BB + SO + IBB + HBP + SH + SF + GIDP)
- Reduced model: lm(salary ~ AB + R + X3B + HR + RBI + SB + CS + BB + IBB + HBP + SH + GIDP
- F-statistic = 1.132251 & high p-value of 0.33914
  - Fail to reject the null hypothesis that the reduced model is sufficient
- After concluding that the reduced model is better, we are interested in seeing if there is a different combination of predictors that provide a higher Adjusted R-Squared than the reduced model already provides (0.2103)

Forward, Backward and Stepwise Selection Results

```
Predictors from Forward Selection
 [1] "HR"   "BB"   "CS"   "GIDP" "SH"   "R"    "X3B"  "IBB"  "SB"   "HBP"
[11] "AB"   "RBI"  "SO"
Predictors from Backward Selection
 [1] "AB"   "R"    "X3B"  "HR"   "RBI"  "SB"   "CS"   "BB"   "SO"   "IBB"
[11] "HBP"  "SH"   "GIDP"
Predictors from Stepwise Selection
 [1] "AB"   "R"    "X3B"  "HR"   "RBI"  "SB"   "CS"   "BB"   "SO"   "IBB"
[11] "HBP"  "SH"   "GIDP"
```

- Summary output of the model after each selection method has an Adjusted R-Squared of 0.2104, which is higher by only 0.0001
- We can see there is a very small improvement from the reduced model in the ANOVA test above

From the three selection methods, Backward and Stepwise Selection have exactly the same 13 final predictors:

1. At-Bats (AB)
2. Runs Scored (R)
3. Triples (X3B)
4. Homeruns (HR)
5. Runs Batting In (RBI)
6. Stolen Bases (SB)
7. Caught Stealing (CS)
8. Walks (BB)
9. Strikeouts (SO)
10. Intentional Walks (IBB)
11. Hit By Pitch (HBP)
12. Sacrifice Hit (SH)
13. Grounded Into Double Play (GIDP)

Conclusion & Future Improvements

- Many of the models we found are insufficient for predicting the outcomes we are interested in
  - Why? real world data isn’t perfect, nor linear
  - Maybe look at interactions terms?
  - Curse of dimensionality
  - GLMs & GAMs

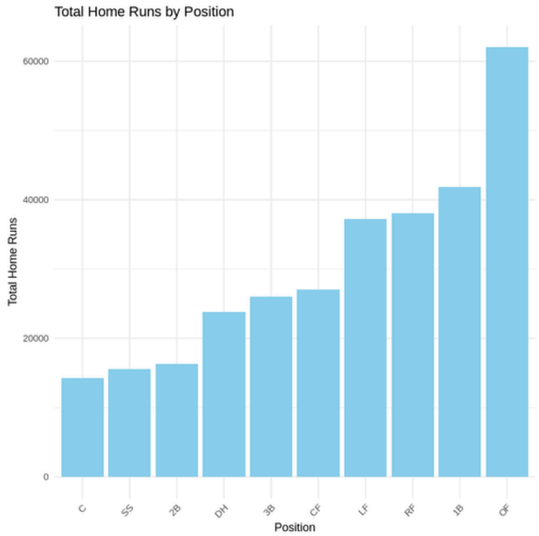
- Future work should focus on a set length of years to narrow and improve our research questions

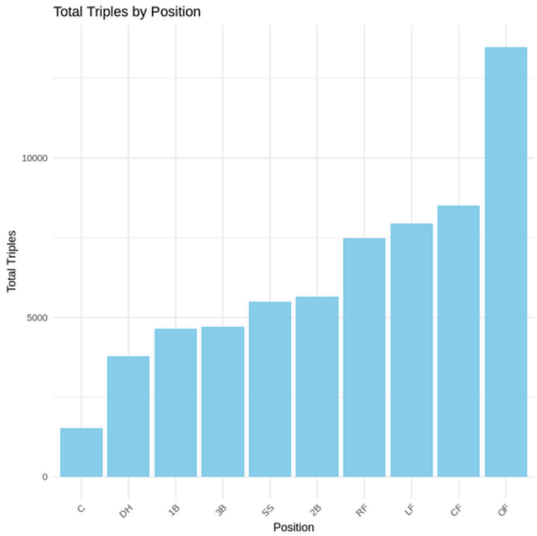
- Make sure to test models (split train/test/valid. sets)
  - Find MSPE as a metric to evaluate models

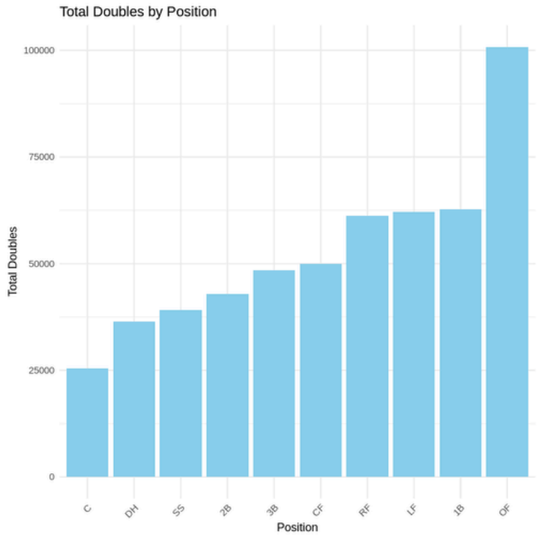
Questions??

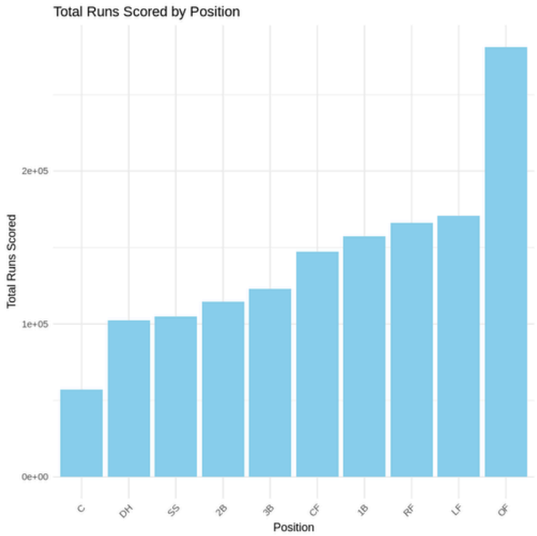
We have fun facts (about hitting and position) if you want to ask what fun facts we have!

Fun Facts and Graphs









- Total Home Runs by position
  - 1B, RF, LF
- Total Triples by position
  - CF, LF, RF
- Total Doubles by position
  - 1B, LF, RF
- Total runs scored by position
  - LF, RF, 1B

Citations

- https://www.loc.gov/collections/jackie-robinson-baseball/articles-and-essays/baseball-the-color-line-and-jackie-robinson/1860s-to-1890s/#:~:text=Americans%20began%20playing%20baseball%20on,as%20America%27s%20%22national%20pastime.%22
- https://www.statista.com/statistics/193466/total-league-revenue-of-the-mlb-since-2005/
- https://www.kaggle.com/datasets/open-source-sports/baseball-databank
- https://www.kaggle.com/