Can algorithms play Moneyball?

Zachary Agrue, Patricia Mills,
Charles VanLeuvan

https://poweredtemplate.com/
00736/0/index.html

Motivation and Problem Statement

- In 2018, the Chicago White Sox lost over 100 games and the Washington Nationals barely had a winning season.
- Teams trade players in the off-season in order to have a better season and stay under the salary cap.

- Can classification algorithms help determine which players to keep and which to trade?

About the Data

- 2018 Game Logs for White Sox and Nationals
  - (n.d.). Retrieved May 3, 2020, from https://www.retrosheet.org/gamelogs/
- 2018 Player Season Stats for White Sox and Nationals
  - 2018 Chicago White Sox Statistics. (n.d.). Retrieved May 6, 2020, from https://www.baseball-reference.com/teams/CHW/2018.shtml
  - 2018 Washington National Statistics. (n.d.). Retrieved May 6, 2020, from https://www.baseball-reference.com/teams/WAS/2018.shtml
- 2018 Salary Data for White Sox and Nationals
  - Cot's Baseball Contracts. (n.d.). Retrieved May 7, 2020, from https://legacy.baseballprospectus.com/compensation/cots/american-league/chicago-white-sox/
  - Cot's Baseball Contracts. (n.d.). Retrieved May 7, 2020, from https://legacy.baseballprospectus.com/compensation/cots/national-league/washington-nationals/

Types of Variables Game Logs

- These datasets contain 42 details for each game played by the respective team, paired down from 161.

Visitor / Home	Abbreviations for Visiting / Home Teams	VisitorScore / HomeScore	Final Score for each team	V_ / H_AtBats	Number of at bats per team
V_ / H_Hits	Hits for each team	V_ / H_HRs	Homeruns for each team	V_ / H_RBI	Runs Batted In for each team
V_ / H_Walks	Walks for each team	V_ / H_Strikeouts	Times a batter for either team struck out.	V_ / H_StolenBases	Bases stolen by either team
W_Pitcher	Name of the Winning Pitcher	L_PitcherName	Name of the Losing Pitcher	S_PitcherName	Name of the Save Pitcher, if there is one
GW_RBIBatterName	Name of the Batter who batted in the winning run	V_ / H_StartingPitcherName	Name of the Starting Pitcher for each team	V_ / H_Batter#Name	Name for each batter in the order for each team and where they are in the

Types of Variables
Player Season Stats – Batting

- These statistics measure a batter's effectiveness.

RK	Rank	**Pos**	Defensive Position	**Player**	Name of the player
Age	Player's Age	**G**	Number of games played in 2018	**PA**	Number of plate appearances in 2018
AB	Number of At Bats in 2018	**R**	Number of Runs Scored in 2018	**H**	Number of Hits in 2018
X2B	Number of Doubles in 2018	**X3B**	Number of Triples in 2018	**HR**	Number of Home Runs in 2018
RBI	Number of Runs Batted In in 2018	**SB**	Number of bases stolen in 2018	**BB**	Number of bases on balls in 2018 (Walks)
SO	Number of times struck out in 2018	**BA**	Batting Average (hits/at bats) in 2018	**OBP**	Percent of At Bats reaching base in 2018
SLG	Slugging Percentage ((Hits + 2*X2B + 3*X3B + 4*HR)/At Bats)	**OPS**	On-Base + Slugging Percentage	**OPS.**	100*[OBP/lg(OBP) + SLG/lg(SLG)-1]

Types of Variables
Player Season Stats – Pitching

- These statistics measure a pitcher's effectiveness.

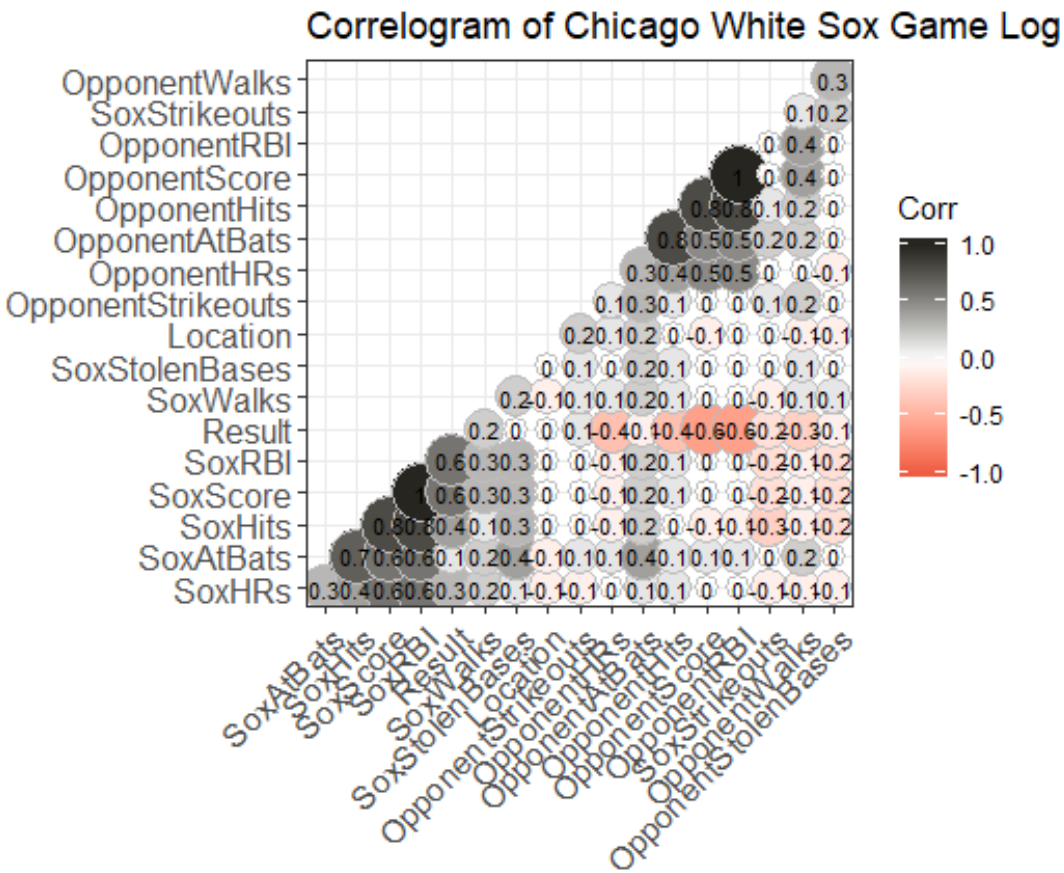
RK	Rank	**Pos**	Defensive Position	**Player**	Name of the player
Age	Player's Age	**W**	Number of games won in 2018	**L**	Number of games lost in 2018
W.L.	Win/Loss percentage in 2018	**ERA**	Earned Runs in 2018	**G**	Number of games played in 2018
GS	Number of games started in 2018	**GF**	Number of games finished in 2018	**CG**	Number of complete games in 2018
SHO	Number of shutouts in 2018	**SV**	Number of games saved in 2018	**IP**	Number of innings pitched in 2018
H	Number of hits allowed in 2018	**R**	Number of runs allowed in 2018	**ER**	Earned Runs in 2018
HR	Number of Home Runs allowed in 2018	**BB**	Number of walks given up in 2018	**IBB**	Number of intentional walks given up in 2018
SO	Number of strikeouts in 2018	**HBP**	Number of times hit batter by pitches in 2018	**BK**	Number of balks in 2018
WP	Number of wild pitches in 2018	**BF**	Number of batters faced in 2018	**FIP**	Fielding Independent Pitching (measures pitcher's effectiveness causing SO, preventing HR, BB, and HBP)

Types of Variables
Salary Data

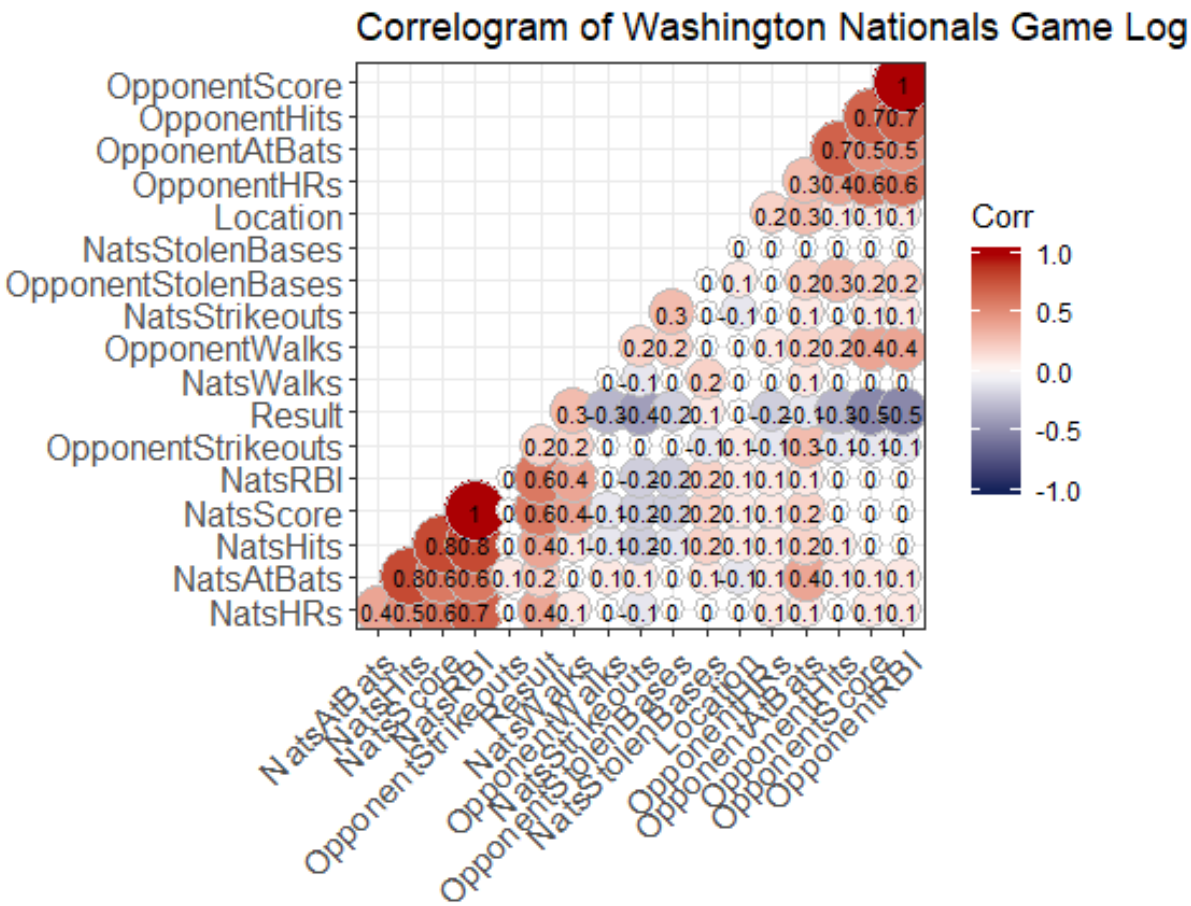
- This data is only available for players who were on the 26-man roster for opening day.

- Player: Player’s Name (Last, First)
- Position: Defensive Position Played
- Length/TotalValue: Length of contract/Total Contracted amount to be paid
- sal2018: Salary for the 2018 season

Correlated Variables – White Sox

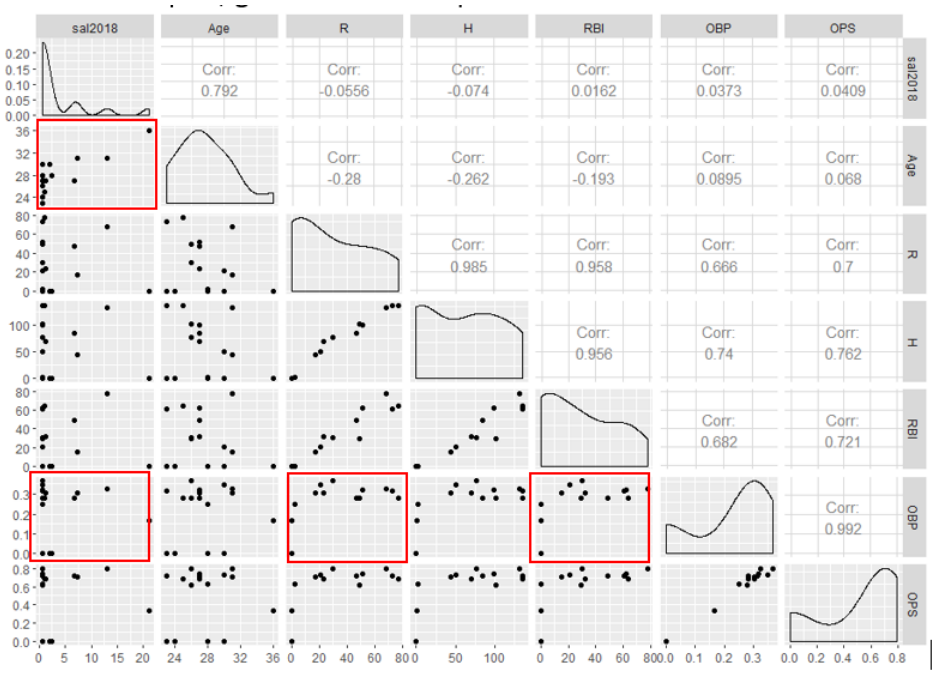


Correlated Variables – Nationals



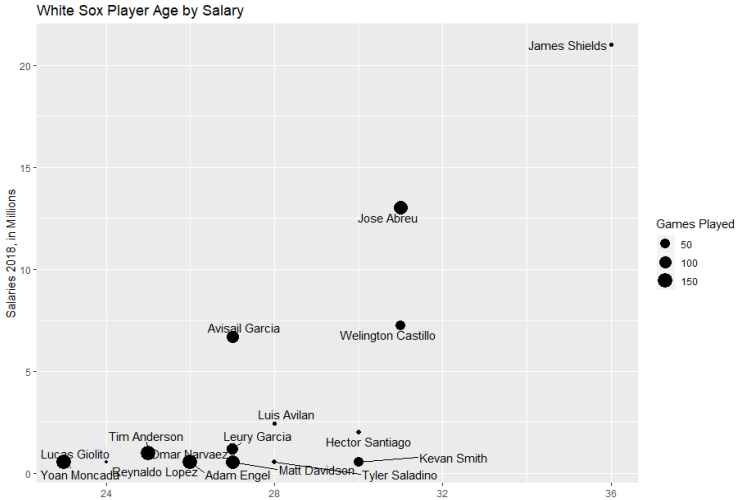
Exploratory Data Analysis – White Sox

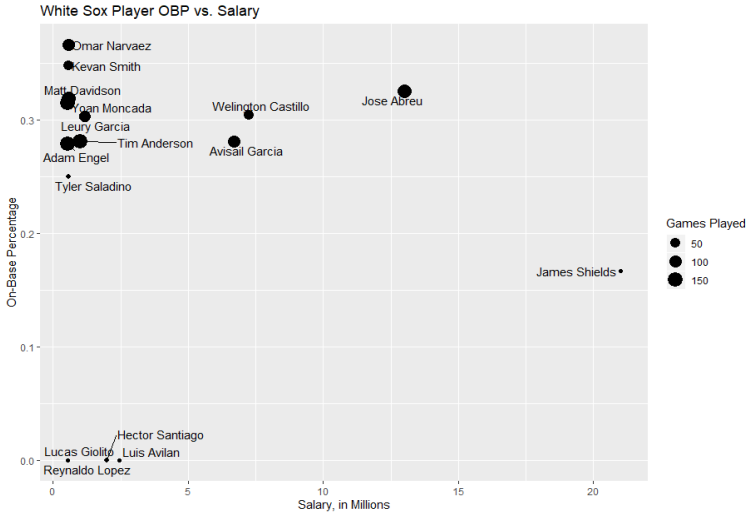
- Used a correlation scatterplot matrix to identify which categories change with salary
- Also useful to see which stats interact with each other



Exploratory Data Analysis – White Sox

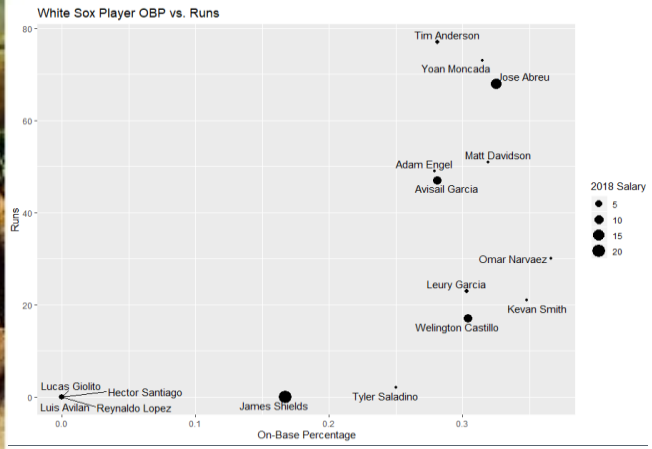
- Older players tend to have a higher salary, likely due to being more established in the league. They can negotiate on prior years' performance
- Age, however, is weakly correlated to performance (R = 0.26)
- Performance vs. Salary→ OBP does not have a big influence on salary

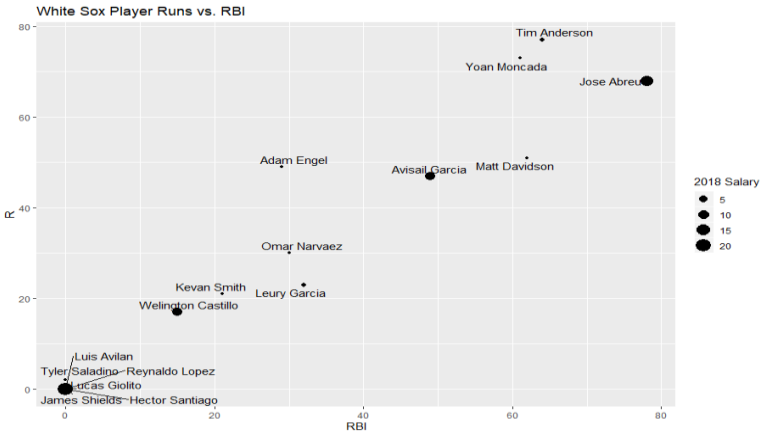




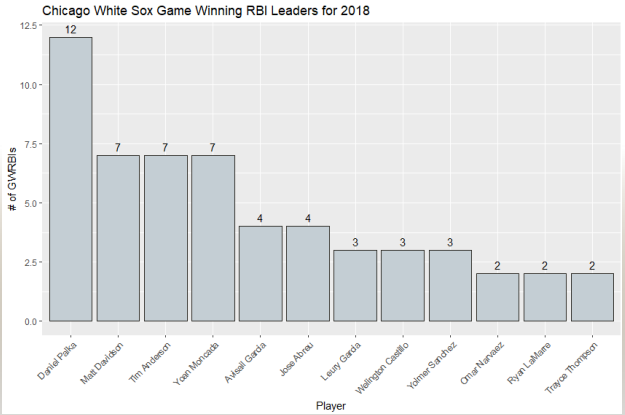
Exploratory Data Analysis – White Sox

- OBP does not necessarily translate to Runs, but RBIs do



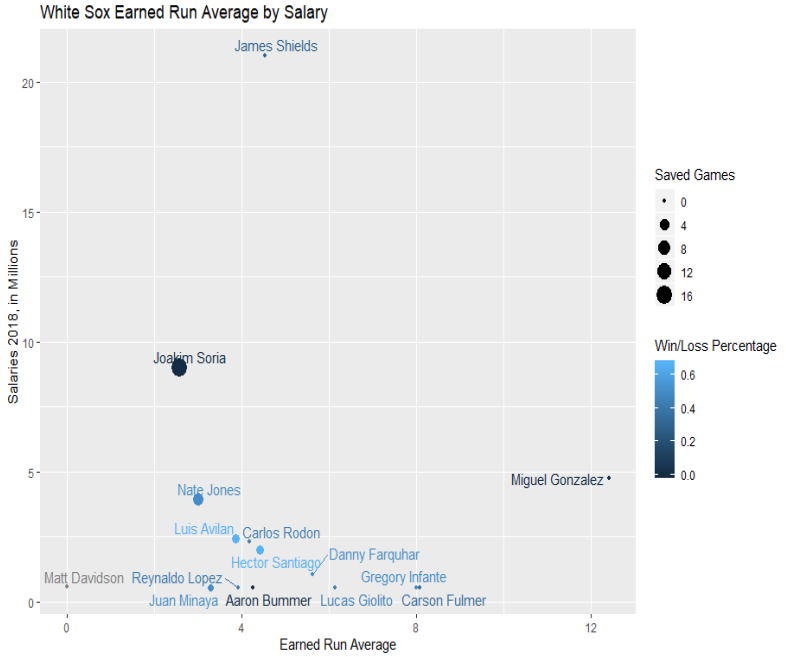


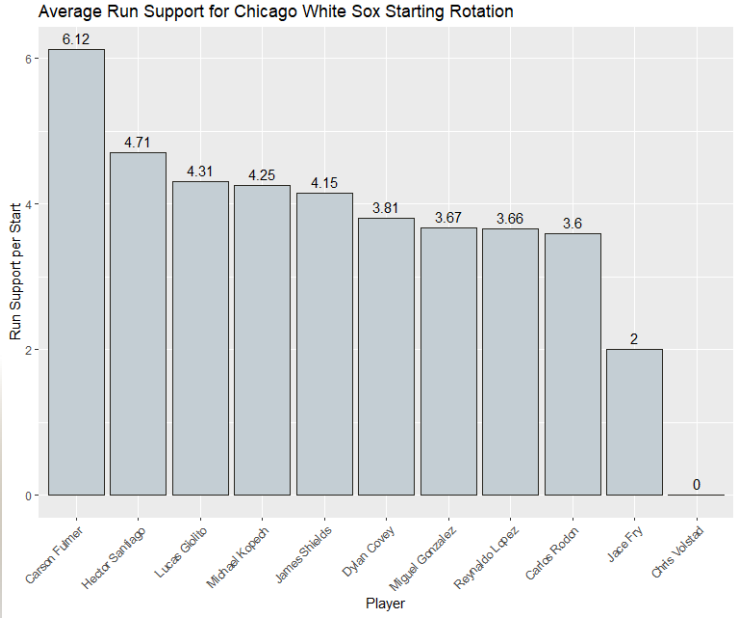
- Clutch factor: Daniel Palka has 40% more game winning RBIs than his closest teammate



Exploratory Data Analysis – White Sox

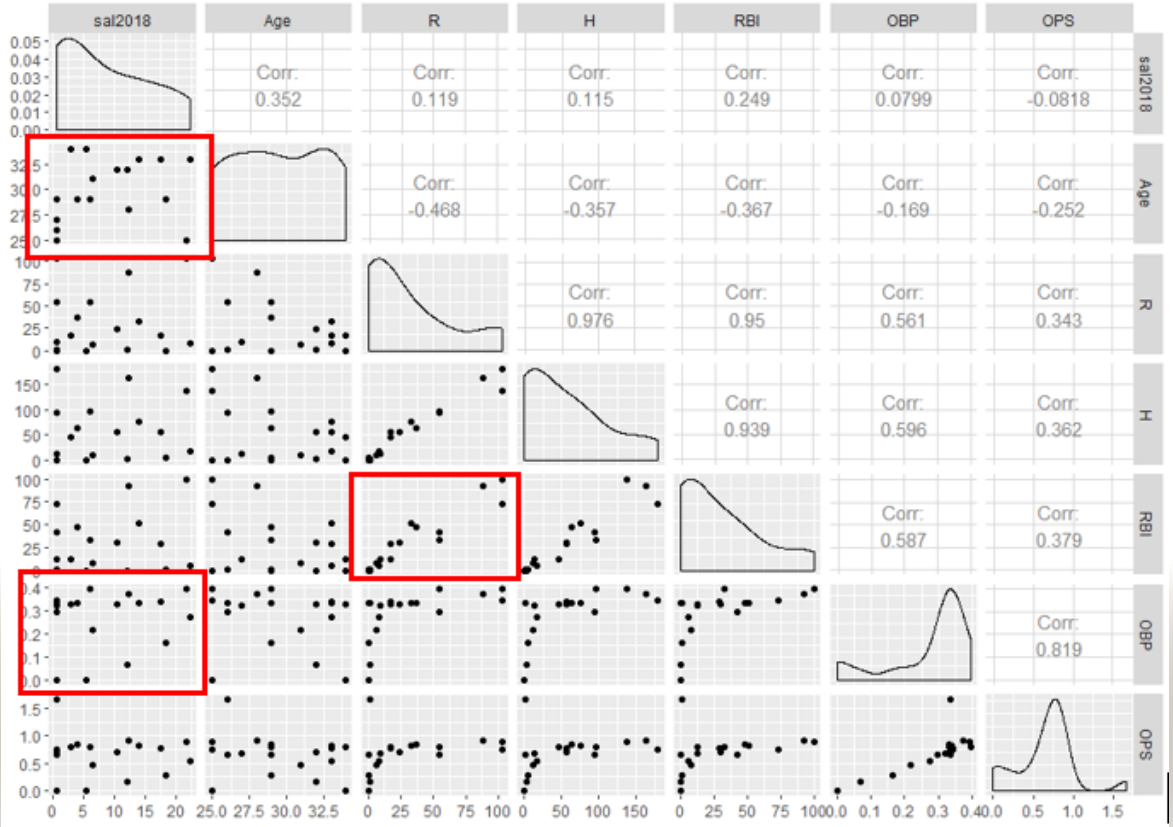
- ERA is the gold standard for pitcher comparison
- Lower ERA pitcher have higher salaries.
- For closers, saves has the same effect as ERA. The highest salary has a Win/Loss percentage of 0 but the most saves.
- The run support is there but it doesn't counter-balance many of the ERAs that are greater than 4.





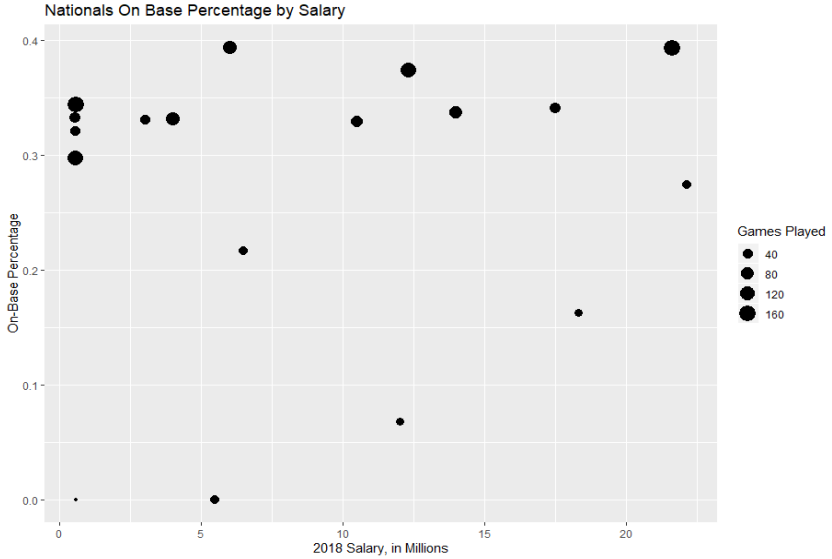
Exploratory Data Analysis – Nationals

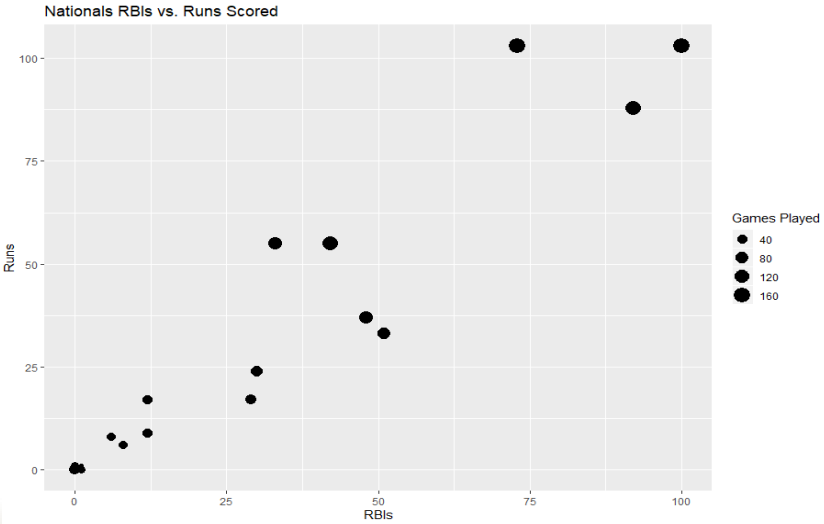
- Nationals batters correlations are similar to the White sox
  - **Calculated stats are less correlated with salary than raw stats**



Exploratory Data Analysis – Nationals

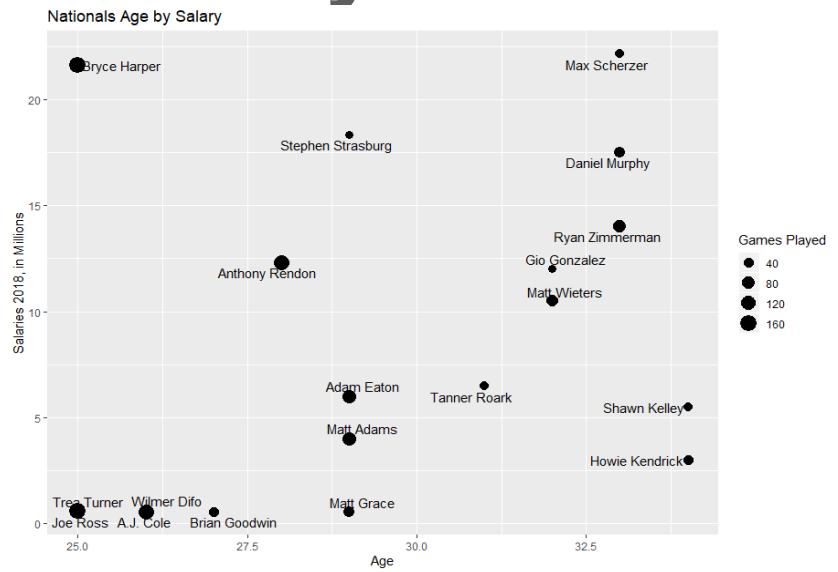
- Nationals have higher OBP across the salary range compared to the White Sox since they were a better offensive team
- Runs directly contribute to a teams score (as opposed to OBP). Players that bat in runs also score runs (R = 0.95)
- Harper, Rendon, and Turner are the top offensive players

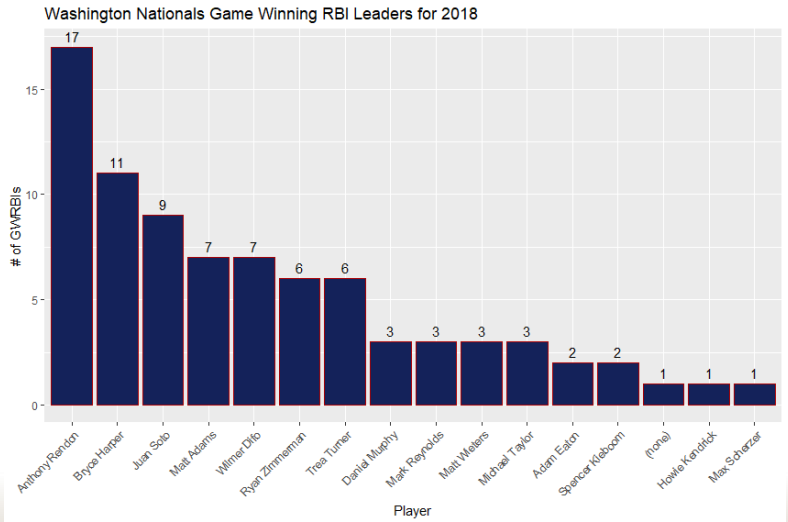




Exploratory Data Analysis – Nationals

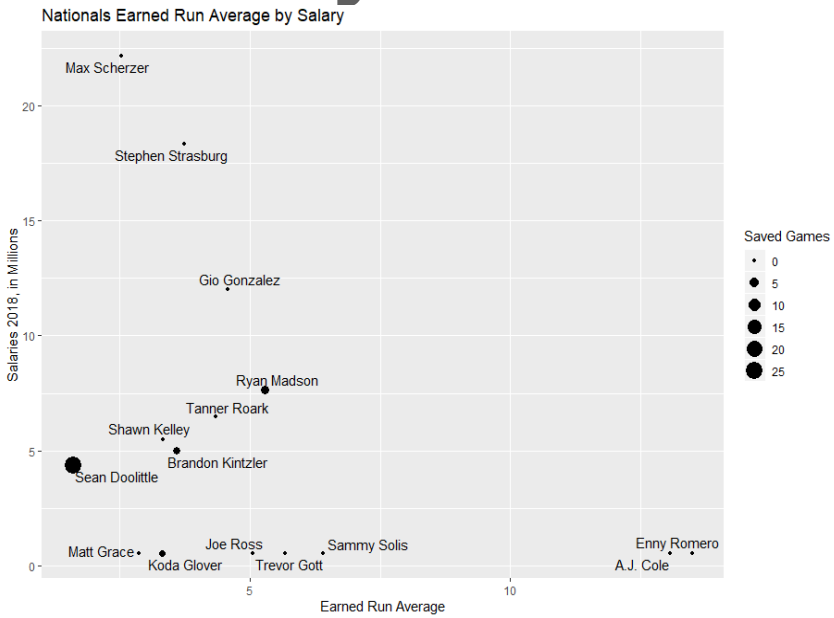
- Consistent with the White Sox, older players have higher salaries.
- However, older =/= better performance
  - **Trea turner performed similar to Harper, but is paid 20x less**
- Clutch Factor: Looking at it from a salary perspective, Rendon has 2.5x more game wins than Daniel Murphy but is paid $10 million less

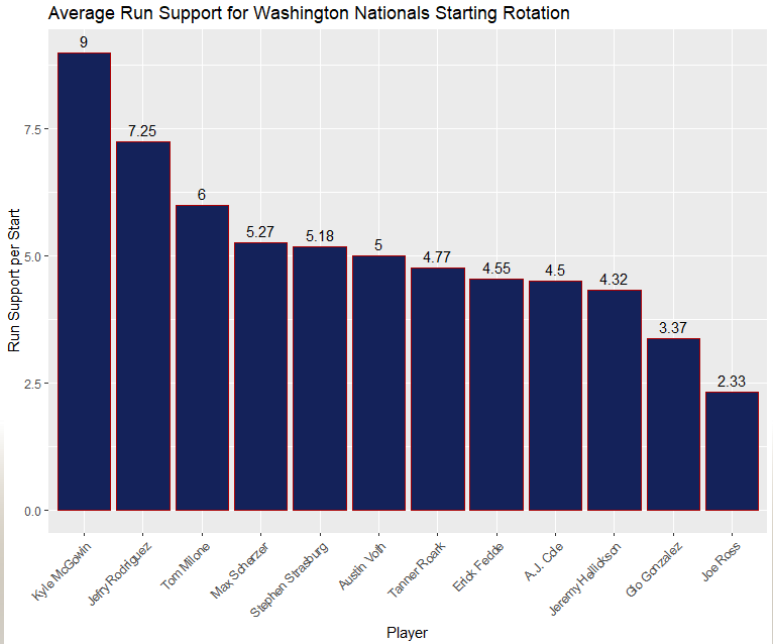




Exploratory Data Analysis – Nationals

- Nationals starting rotation clearly show the lower ERA = higher salary
- Saves also scales well with salary.
- As most pitchers' ERA was under 5, the run support is enough to win games





Data Transformation

- Salaries: Transformed by dividing by 1,000,000
- Game Logs for each team:
  - Added column for result WIN/LOSS
    - Changed to 1 or 0 when creating the correlogram
  - Added column for HOME/AWAY
    - Changed to 1 or 0 when creating the correlogram
- Added a column to classify salaries as low, mid, high for each team
  - Mid range based on average salaries.

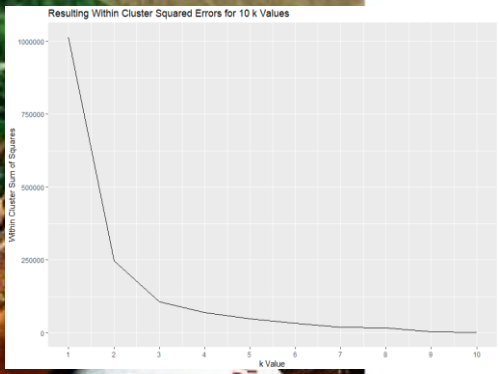
	CWS Batters	CWS Pitchers	WAS Batters	WAS Pitchers
Low	< $2 mil	< $2 mil	< $3 mil	< $2 mil
Mid	$2 mil - $6.99 mil	$2 mil - $6.99 mil	$3.1 mil – $10.99 mil	$2 mil - $6.99 mil
High	> $7 mil	> $7 mil	> $11 mil	> $7 mil

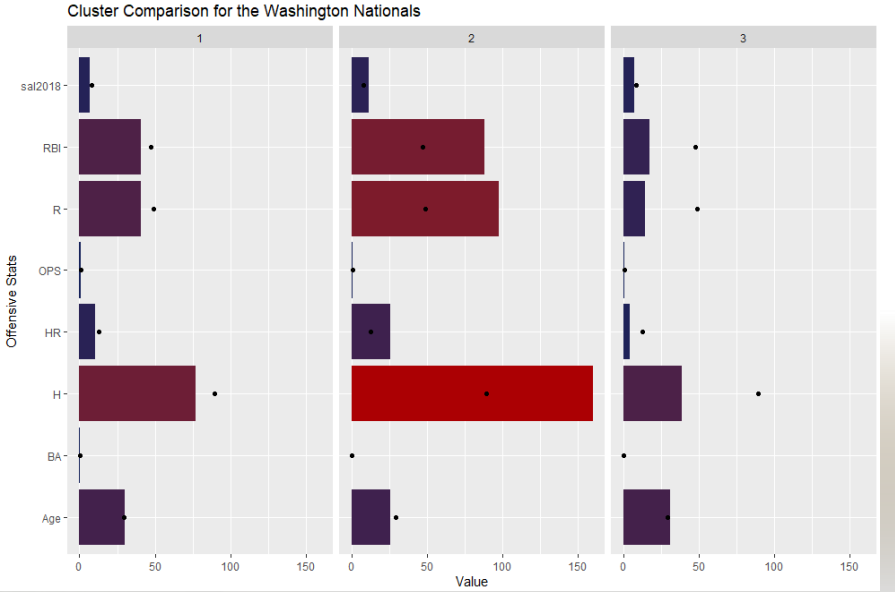
Association Rules

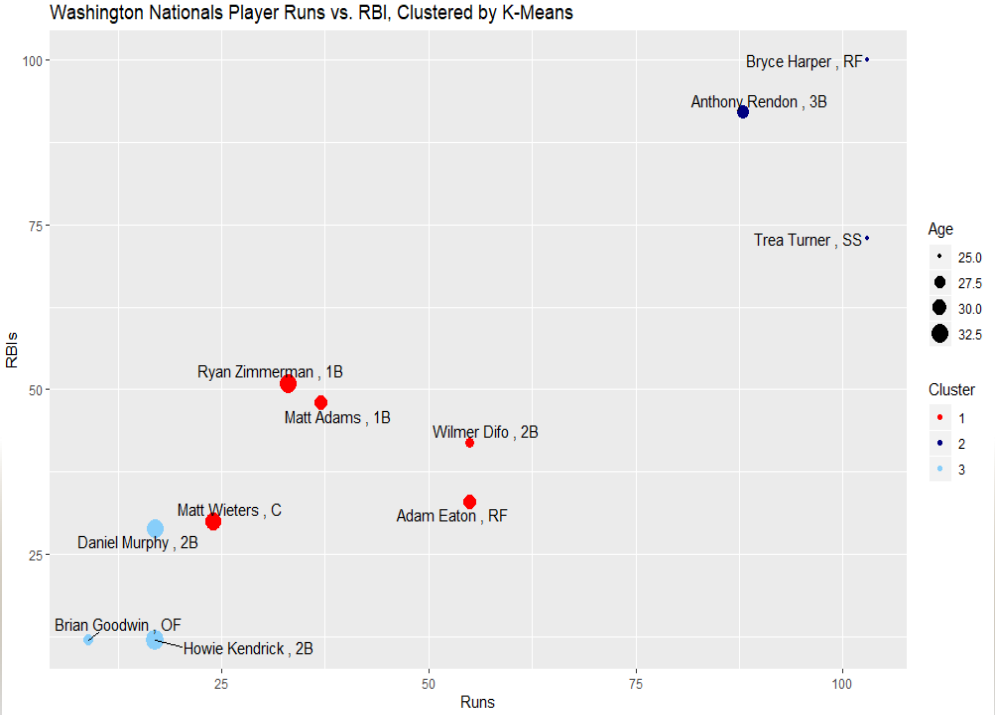
- Chicago White Sox General
  - If the White Sox play at night and the opponent has no home runs, they are likely to win.
- Chicago White Sox Winning Batting Line Ups
  - Four of the top 20 rules show Adam Engel in the 9th spot in the line up.
    - Engel was on a 1 year, $552,000 contract.
- Washington Nationals General
  - If the Nationals win, the opponent is likely to not have any stolen bases.
    - This rule appeared if they were are home or if they were away.
- Washington Nationals Winning Batting Line Ups
  - Trea Turner appears in 12 of the top 20 rules where he is either batting first or second in the line up.
    - Turner was on a 1 year, $577,000 contract.
  - Juan Soto appears in 8 of the top 20 rules as batting 5th in the line up.
    - Soto was not on the opening day roster so his salary for 2018 is unknown.
    - He started the year with the Hagerstown Suns in the Class A South Atlantic League.

Clustering

- Do the higher paid players perform similarly to each other?
  - Answer → Some do
  - **Nationals:** Bryce Harper (highest paid) and Anthony Rendon (4th highest paid) are clustered together due offensive power (OPS) along with Trea Turner
    - Trea Turner groups with Harper and Rendon, despite being on a 1 year, *low value contract*. He had the most ABs, Hits, and runs for the Nationals in 2018
    - Turner's OPS is far smaller than Harper and Rendon due to his more ABs
    - OPS = SLG + OBP. Raw stats have more influence.
    - Elbow Method for k value. As k increases, players form single clusters

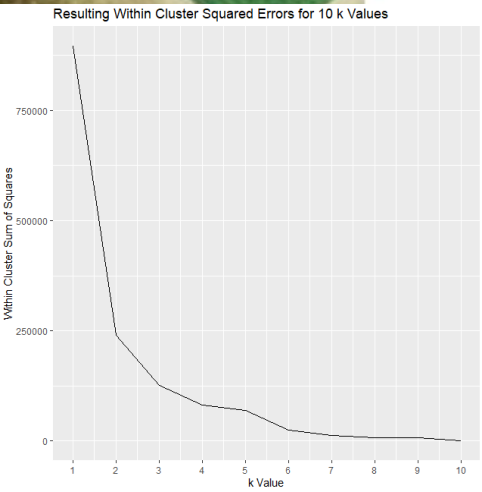


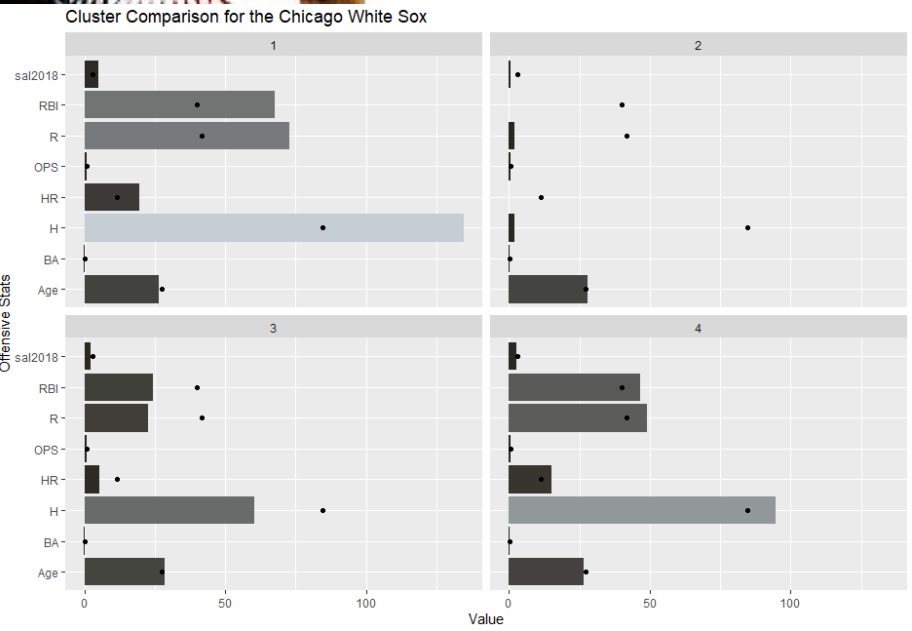


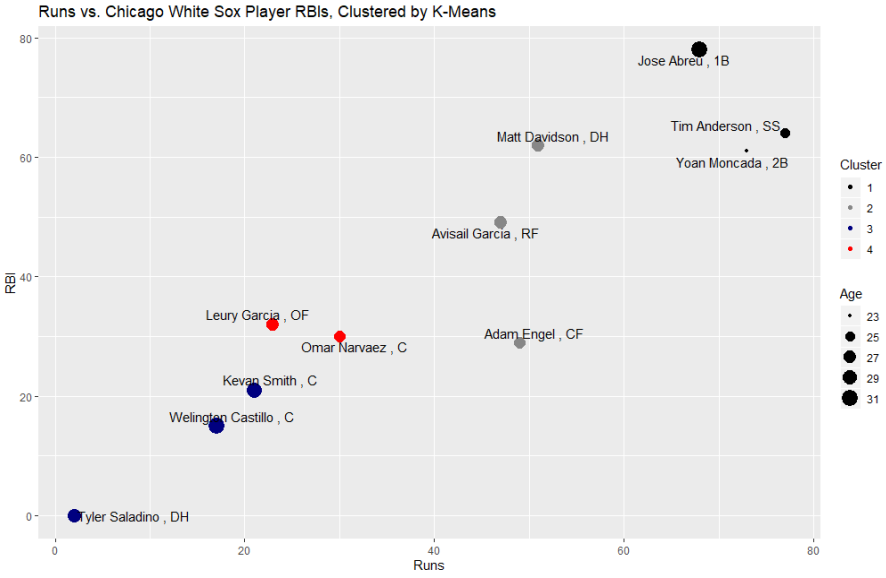


Clustering

- Do the higher paid players perform similarly to each other?
  - Answer → Not really
  - **White Sox**
    - Jose Abreu is the highest paid player and is a member of the cluster that is above all team offensive averages
    - The other three multi-million $ paid players (Garcias and Castillo) do not match Abreu in hits, total bases, or runs. However, they did play far fewer games
      - This supports the observation that raw stats have more influence
    - Lower paid players Tim Anderson, SS, and Yoan Moncada, 2B, cluster with Abreu in the top group, and they match his performance for the 2018 season in hits, runs, and total bases. (again, raw stats)
    - Elbow method was more ambiguous with this team. K 2:6 could be valid, optimal use was found with k=4







kNN Clustering

- Can a player's salary be predicted?
  - **Use clustering to predict a player's salary range**
    - Nearest neighbors, in terms of performance, should indicate what the player's salary should be.
    - Use Case: Identifying players that are overvalued or undervalued
  - **K = 5 for batters, K = 3 for pitchers**
  - Nationals Batters
    - Accuracy → 48.9%
  - Nationals Pitchers
    - Accuracy → 60%
  - White Sox Batters
    - Accuracy → 58%
  - White Sox Pitchers
    - Accuracy → 42%

Model Process

1. Factor player salaries into "High","Mid","Low"
2. Convert to numeric

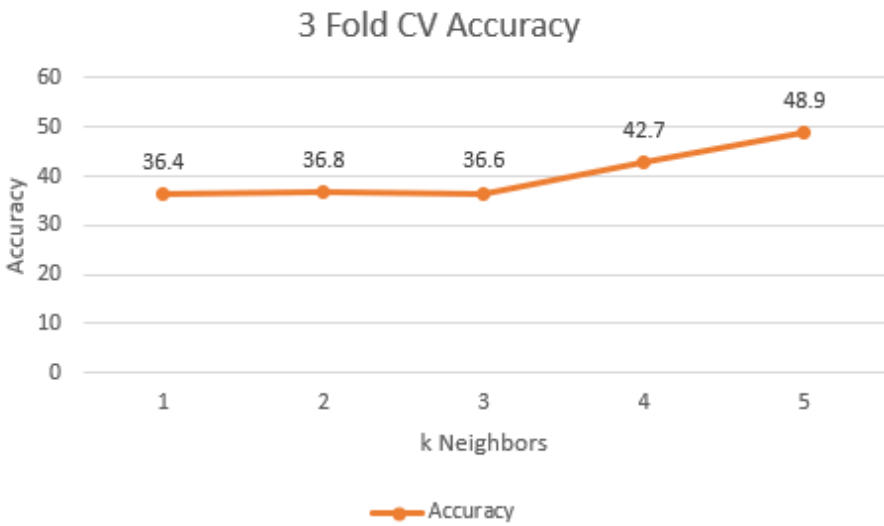
kNN →

	Predicted		
Actual	**High**	**Low**	**Mid**
High	x		
Low		x	
Mid			x

1. 3-fold CV
2. Accuracy averaged over the CVs

kNN Clustering

- Model Details
  - **Training**
    - Sampling with replacement to generate data
    - 19 samples in players data set, 12 samples in pitchers
    - Results are highly variable with repetition
  - **Testing**
    - 3-fold cross validation
    - Tested for accuracy
      - True Positive + True Negative / All predictions
  - **K values**
    - Iterated from 1 to sqrt(rows)
    - For batters, k = 5 had highest 3-fold CV accuracy throughout the iterations



Can random forest predict batter salaries based on performance?

- Chicago White Sox
-

	high	low	mid
high	0	3	0
low	1	6	3
mid	0	3	0

- 37.5% accuracy
  - Most of the team's batters are low paid (10 out of 16).

Can random forest predict batter salaries based on performance?

- Washington Nationals
-

	high	low	mid
high	2	3	2
low	2	2	3
mid	4	1	0

- 21.05% accuracy

Can random forest predict pitcher salaries based on performance?

- Chicago White Sox
-

	high	low	mid
high	0	2	0
low	0	6	2
mid	0	4	1

- 46.67% accuracy
  - A slight majority of the team's pitchers is low paid (8 ou t of 15).

Can random forest predict pitcher salaries based on performance?

- Washington Nationals
-

	high	low	mid
high	2	0	2
low	0	5	2
mid	1	2	1

- 53.33% accuracy
  - 7 out of 15 of the team's pitchers are low paid.

Trade Recommendations: Chicago White Sox Batters

```
##                Player2 orig pred Rk  AB RBI    BA   OBP   SLG   OPS
## 1          Abreu, Jose high  low  2 499  78 0.265 0.325 0.473 0.798
## 2        Anderson, Tim  low  mid  4 567  64 0.240 0.281 0.406 0.687
## 3         Avilan, Luis  mid  low 28   1   0 0.000 0.000 0.000 0.000
## 4  Castillo, Welington high  low 13 170  15 0.259 0.304 0.406 0.710
## 5      Garcia, Avisail  mid  low  8 356  49 0.236 0.281 0.438 0.719
## 6       Giolito, Lucas  low  mid 24   6   0 0.000 0.000 0.000 0.000
## 7      Lopez, Reynaldo  low  mid 29   1   0 0.000 0.000 0.000 0.000
## 8      Saladino, Tyler  low high 21   8   0 0.250 0.250 0.375 0.625
## 9     Santiago, Hector  mid  low 25   4   0 0.000 0.000 0.000 0.000
## 10      Shields, James high  low 23   6   0 0.167 0.167 0.167 0.333
```

```
##          Player2 orig pred Rk Pos  AB  R RBI    BA   OBP   SLG   OPS
## 1 Davidson, Matt  low  low  9  DH 434 51  62 0.228 0.319 0.419 0.738
## 2    Engel, Adam  low  low  7  CF 429 49  29 0.235 0.279 0.336 0.614
## 3  Garcia, Leury  low  low 11  OF 258 23  32 0.271 0.303 0.376 0.679
## 4  Moncada, Yoan  low  low  3  2B 578 73  61 0.235 0.315 0.400 0.714
## 5  Narvaez, Omar  low  low  1   C 280 30  30 0.275 0.366 0.429 0.794
## 6   Smith, Kevan  low  low 12   C 171 21  21 0.292 0.348 0.380 0.728
```

- Adam Engel
  - OPS is the only good stat he has.
- Omar Naraez
  - Batting Average and OPS good but lackluster RBIs compared to other batters.
- Wellington Castillo
  - High paid and decent player but other players are better for less money.

Trade Recommendations: Washington Nationals Batters

```
##            Player2 orig pred Rk  AB RBI    BA   OBP   SLG   OPS
## 1      Adams, Matt  mid high 10 249  48 0.257 0.332 0.510 0.842
## 2     Difo, Wilmer  low  mid  3 408  42 0.230 0.298 0.350 0.649
## 3      Eaton, Adam  mid high  9 319  33 0.301 0.394 0.411 0.805
## 4    Gonzalez, Gio high  low 27  44   0 0.068 0.068 0.091 0.159
## 5   Goodwin, Brian  low  mid 17  65  12 0.200 0.321 0.354 0.674
## 6    Harper, Bryce high  low  8 550 100 0.249 0.393 0.496 0.889
## 7    Kelley, Shawn  mid  low 38   1   0 0.000 0.000 0.000 0.000
## 8  Kendrick, Howie  low high 14 152  12 0.303 0.331 0.474 0.805
## 9   Murphy, Daniel high  low 13 190  29 0.300 0.341 0.442 0.784
## 10   Roark, Tanner  mid high 25  58   8 0.190 0.217 0.259 0.475
## 11       Ross, Joe  low  mid 32   5   0 0.000 0.000 0.000 0.000
## 12   Scherzer, Max high  mid 24  70   6 0.243 0.274 0.271 0.545
## 13    Turner, Trea  low high  4 664  73 0.271 0.344 0.416 0.760
## 14   Wieters, Matt  mid high  1 235  30 0.238 0.330 0.374 0.704
## 15 Zimmerman, Ryan high  mid  2 288  51 0.264 0.337 0.486 0.824
```

```
##              Player2 orig pred Rk  AB RBI    BA   OBP   SLG   OPS
## 1         Cole, A.J.  low  low 33   3   1 0.333 0.333 1.333 1.667
## 2        Grace, Matt  low  low 35   3   0 0.333 0.333 0.333 0.667
## 3    Rendon, Anthony high high  5 529  92 0.308 0.374 0.535 0.909
## 4 Strasburg, Stephen high high 26  41   1 0.122 0.163 0.122 0.285
```

- Wilmer Difo: low batting average, mediocre OBP, low slugging, okay OPS
- Brian Goodwin: low batting average, low slugging
- Matt Wieters: low RBIs, low batting average.

Trade Recommendations: Chicago White Sox Pitchers

```
##            Player2 orig pred Rk  W.L.   ERA SV  FIP
## 1     Avilan, Luis  mid  low 10 0.667  3.86  2 2.71
## 2    Bummer, Aaron  low  mid 13 0.000  4.26  0 2.40
## 3 Gonzalez, Miguel  mid  low 21 0.000 12.41  0 8.02
## 4      Jones, Nate  mid  low 14 0.500  3.00  5 4.56
## 5     Minaya, Juan  low  mid  9 0.500  3.28  1 3.57
## 6    Rodon, Carlos  mid  low  5 0.429  4.18  0 4.95
## 7   Shields, James high  low  1 0.304  4.53  0 5.09
## 8    Soria, Joakim high  low  6 0.000  2.56 16 2.15
```

```
##            Player2 orig pred Rk  W.L.  ERA SV  FIP
## 1   Davidson, Matt  low  low 30 0.000 0.00  0 2.83
## 2  Farquhar, Danny  low  low 25 0.500 5.63  0 5.79
## 3   Fulmer, Carson  low  low 12 0.333 8.07  0 7.27
## 4   Giolito, Lucas  low  low  3 0.435 6.13  0 5.56
## 5 Infante, Gregory  low  low 23 0.500 8.00  0 4.49
## 6  Lopez, Reynaldo  low  low  2 0.412 3.91  0 4.63
## 7 Santiago, Hector  mid  mid  7 0.667 4.41  2 5.09
```

- Miguel Gonzalez
  - High FIP and High ERA
- Carlos Rodon
  - Bad win/loss record, higher ERA and FIP
- Carson Fulmer
  - All around awful record
- Lucas Giolito
  - All around BAD record

Trade Recommendations: Washington Nationals Pitchers

```
##             Player2 orig pred Rk  W.L.  ERA SV  FIP
## 1     Gonzalez, Gio high  mid  3 0.389 4.57  0 4.25
## 2       Grace, Matt  low  mid  8 0.500 2.87  0 3.39
## 3     Kelley, Shawn  mid  low 15 1.000 3.34  0 4.55
## 4 Kintzler, Brandon  mid  low 14 0.333 3.59  2 3.44
## 5      Madson, Ryan high  mid 10 0.286 5.28  4 4.36
## 6     Roark, Tanner  mid high  2 0.375 4.34  0 4.27
## 7      Solis, Sammy  low  mid 11 0.333 6.41  0 4.91

##              Player2 orig pred Rk  W.L.   ERA SV   FIP
## 1         Cole, A.J.  low  low 25 0.500 13.06  0 10.51
## 2    Doolittle, Sean  mid  mid  7 0.500  1.60 25  1.89
## 3       Glover, Koda  low  low 22 0.250  3.31  1  4.69
## 4       Gott, Trevor  low  low 19 0.000  5.68  0  6.21
## 5       Romero, Enny  low  low 29 0.000 13.50  0 10.66
## 6          Ross, Joe  low  low 23 0.000  5.06  0  5.85
## 7      Scherzer, Max high high  1 0.720  2.53  0  2.65
## 8 Strasburg, Stephen high high  4 0.588  3.74  0  3.62
```

- Brandon Kintzler: bad win-loss, okay ERA, okay FIP, 2 saves
- Ryan Madson: bad win-loss, high ERA, high FIP, 4 saves
- Tanner Roark: bad win-loss, high ERA, high FIP
- Sammy Solis: bad win-loss, high ERA, high FIP
- AJ Cole: 2nd highest ERA, 2nd highest FIP
- Koda Glover: bad win-loss, mediocre ERA, high FIP
- Gio Gonzalez: bad win-loss, high ERA, high FIP
- Trevor Gott: bad win-loss, high ERA, high FIP
- Enny Romero: bad win-loss, highest ERA, highest FIP
- Joe Ross: bad win-loss, high ERA, high FIP