MLB Roster Construction Based On Market Size

Malcolm Gaynor and Parker Gibbons

Project Motivation/Past Literature

- **“What Makes a Winning Baseball Team and What Makes a Playoff Team? by Lopez et al (2011)**
  - Data from 1995 to 2009 - Works to determine which statistics can predict winning, runs scored, and playoff appearance
  - Concludes ERA and OPS can be used to predict wins and playoff appearance
- **“Payrolls and Playoff Probabilities in Major League Baseball” by Somberg et al (2012)**
  - Data from 1998 to 2011 - Examines the relationship between team payroll and the probability of playing in the postseason
  - Concludes that teams win and make the playoffs more with higher payroll
- **“Compensation and performance in Major League Baseball: Evidence from salary dispersion and team performance” by Tao et al (2015)**
  - Data from 1985 to 2013 - Looks at whether salary dispersion within teams impacts winning
  - Concludes that salary dispersion is not as impactful as team payroll

- **Articles suggest methods to predict winning regardless of each specific team’s financials**

Our Approach

- We want to determine whether MLB teams of different market sizes depend on different variables to win and have playoff success
  - Hypothesis: Teams of different market sizes depend on different variables to win and have playoff success

- First: Cluster MLB teams by their market size using K-means clustering

- Second: examine models to predict win percentage and playoff success created separately for teams of each market size

Data

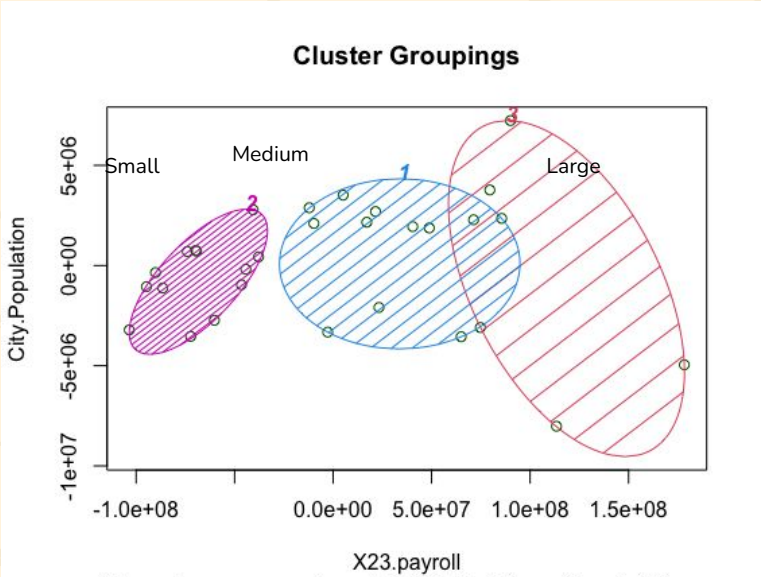
For Clustering:

- Total payroll from the past 4 seasons
- City Population
- Media market size
- Team valuation
- Attendance per game from the past 3 seasons

For Linear Regression and Multivariate Adaptive Regression Splines:

- Response variables:
  - Win percentage and when the team was eliminated (ex. 0 = didn't make playoffs, 5 = won world series)
- Explanatory variables:
  - Total payroll, payroll by position, competitive balance tax threshold
  - Number of players on arbitration, acquired as free agents, acquired by trade, or re-signed or extended, number of rookies
  - Money spent in free agency, money spent re-signing or extending players
  - Average age of batters and average age of pitchers

K-Means Clustering



- Unsupervised non-linear algorithm that clusters data based on their similarity to one another
- Uses a a pre-specified number of clusters
  - We chose three to get our classes
- We want to classify the market size of each MLB team (small, medium, large)
  - Will allow us to later see what variables matter more to each market class and identify similarities/differences

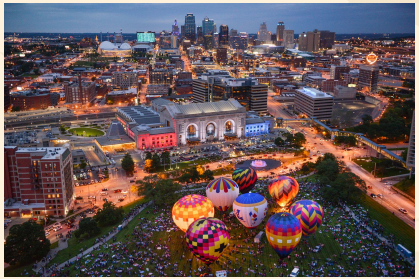
*Each axis scale shows distance from mean

Cluster means:

	X23.payroll	X22.payroll	X21.payroll	X20.payroll	City.Population	media.market.size	Team.valuation	X23.attendence.g	X22.attendence.g	X21.attendence.g
Medium	1 195238927	170387988	150722702	68943348	6575736	2911750	2636.667	32797.83	31270.17	21761.08
Small	2 97365321	92387753	83894109	42645849	3620110	1686462	1445.231	21713.08	18353.31	13474.31
Large	3 272799899	252125385	209895987	93088715	12325086	5079000	3825.000	39515.40	36608.80	24652.80

Conclusions from Clustering

Market Size	Teams
Large	LAD, NYM, NYY, PHI, SDP
Medium	ATL, BOS, CHC, CHW, COL, HOU, LAA, MIN, SFG, STL, TEX, TOR
Small	ARI, BAL, CIN, CLE, DET, KCR, MIA, MIL, OAK, PIT, SEA, TBR, WSH



- Classification of clusters had good accuracy
- Differences in market sizes clearly recognized
- Supports general thinking and our initial expectations
- Could be limited by variable such as team revenue (unavailable)
- Next step: Creating models to determine which financial metrics impact success for each market classification

Linear Regression: Large Markets

- Used backwards elimination and best subsets - Best subsets resulted in models with greater variability explained

- **Model to predict winning percentage (Adjusted R-squared = 0.6159)**
  - **Variables**: Catcher payroll (-), pitcher payroll, CBT space (-), # players resigned/extended, # FA signed (-), money spent on FA (-), average pitcher age (-)
  - **Significant at 99%**: Pitcher payroll, CBT space (-), average pitcher age (-)

- **Model to predict playoff results (Adjusted R-squared = 0.7936)**
  - **Variables**: Infield payroll (-), outfield payroll, pitcher payroll, CBT space (-), # players resigned/extended, money spent on FA (-), average batter age, # of rookies
  - **Significant at 99%**: Infield payroll (-), pitcher payroll, CBT space (-), # players resigned/extended, money spent on FA (-)

Linear Regression: Medium Markets

- **Model to predict winning percentage (Adjusted R-squared = 0.4862)**
  - **Variables**: Total payroll (-), catcher payroll, pitcher payroll, # players resigned/extended (-), money spent on players resigned/extended, average batter age, # of rookies (-)
  - **Significant at 99%**: Pitcher payroll

- **Model to predict playoff results (Adjusted R-squared = 0.3698)**
  - **Variables**: Infield payroll, outfield payroll (-), pitcher payroll, CBT space (-), average pitcher age (-), # of rookies (-)
  - **Significant at 99%**: Outfield payroll (-)

Linear Regression: Small Markets

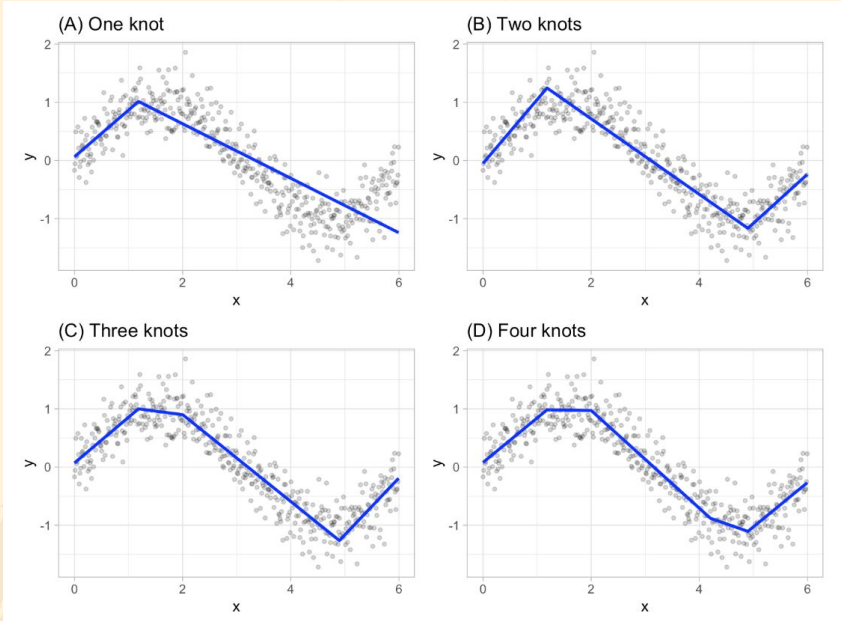
- **Model to predict winning percentage (Adjusted R-squared = 0.4719)**
  - **Variables**: Total payroll (-), Infield payroll, outfield payroll, pitcher payroll, # players acquired via trade, # of rookies (-)
  - **Significant at 99%**: Total payroll (-), Infield payroll, outfield payroll, # of rookies (-)

- **Model to predict playoff results (Adjusted R-squared = 0.2783)**
  - **Variables**: Total payroll (-), Infield payroll, outfield payroll, # players acquired via trade, money spent on players resigned/extended, # of rookies (-)
  - **Significant at 99%**: Total payroll (-), # of rookies (-)

- **Wanted to go more in-depth with modeling and explore some possible non-linear relationships and interaction among the predictor variables...**

Intro to MARS (Multivariate Adaptive Regression Splines)

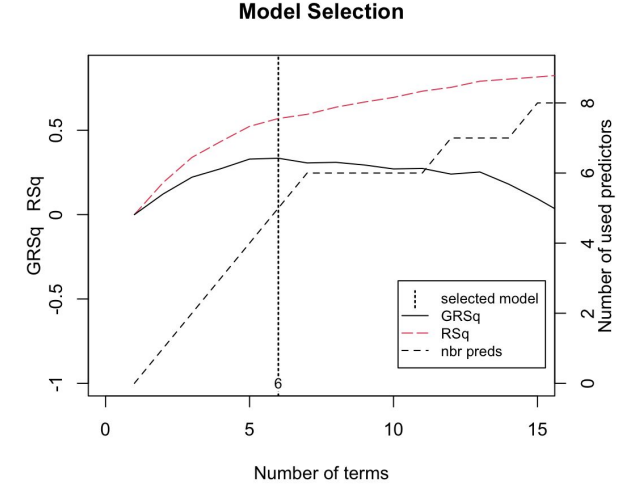
- Piecewise linear regression models
- Explanatory variables put into different “knots”
- Hinge functions
- Important terms:
  - RSS: Residual sum-of-squares
  - RSq: R-squared
  - GCV: Generalized Cross Validation
    - Penalty term per number of knots
    - Goal = minimize
  - GRSq: Generalized R-squared
    - 1- (GCV of model/GCV of intercept-only model)
    - Goal = maximize

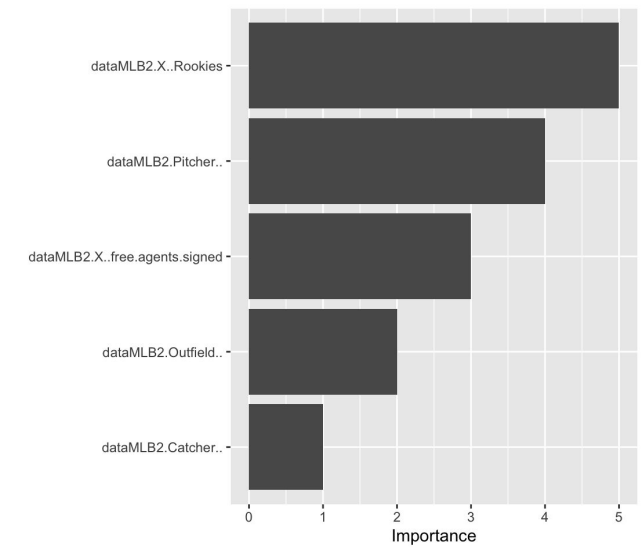


MARS: win% of small market teams

- 6 terms, 5 predictors
- RSq=0.5697135, GRSq = 0.3342206

(Intercept)	h(18-dataMLB2.X..Rookies)	h(1.17122e+07-dataMLB2.Pitcher..)
4.588959e-01	6.325589e-03	-1.414746e-08
h(4-dataMLB2.X..free.agents.signed)	h(6.21898e+06-dataMLB2.Outfield..)	h(dataMLB2.Catcher..-3.325e+06)
2.650714e-02	-1.423724e-08	-5.490049e-09

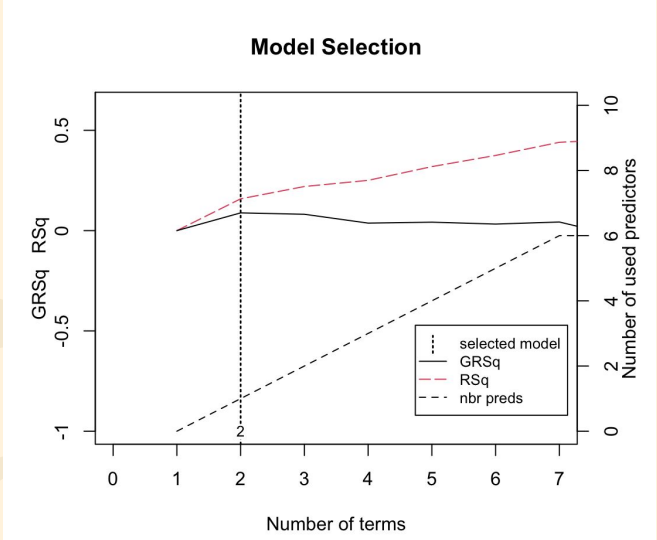




MARS: playoff results of small market teams

- 2 terms, 1 predictor
- RSq=0.78431, GRSq = 0.5042049

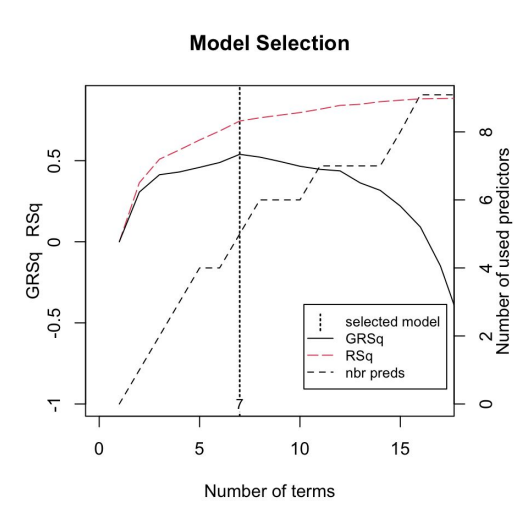
```
(Intercept) h(18-dataMLB2.X..Rookies)
 0.06343837                0.07883805
```

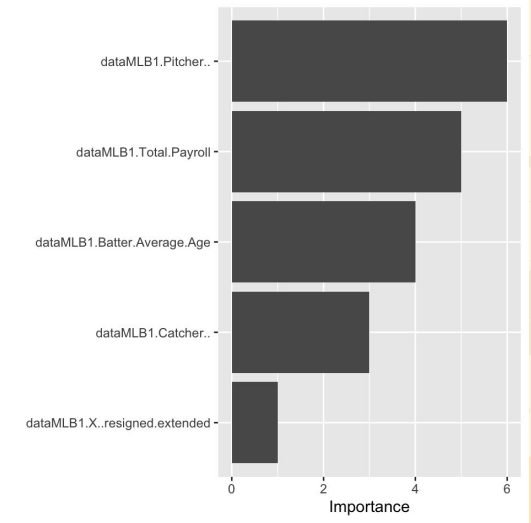


MARS: win% of medium market teams

- 7 terms, 5 predictors
- RSq=0.7443721, GRSq = 0.5390351

(Intercept)	h(5.74756e+07-dataMLB1.Pitcher..)	h(8.65962e+07-dataMLB1.Total.Payroll)
6.249644e-01	-2.802152e-09	4.036541e-09
h(8.66707e+06-dataMLB1.Catcher..)	h(dataMLB1.Batter.Average.Age-28.5)	h(14-dataMLB1.X..resigned.extended)
-1.863056e-08	7.545049e-02	5.016475e-03
h(dataMLB1.Catcher..-1.88593e+06)		
-8.743005e-09		

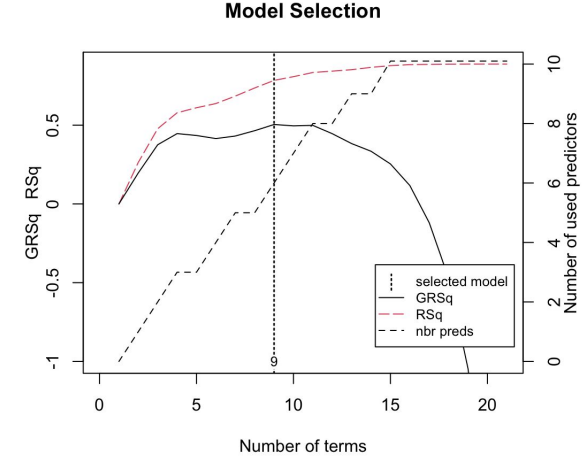


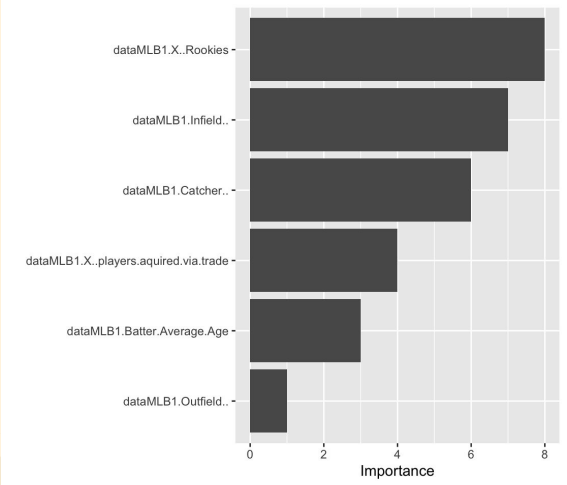


MARS: playoff results of medium market teams

- 9 terms, 6 predictors
- RSq=0.78431, GRSq = 0.5042049

(Intercept)	h(dataMLB1.Catcher..-7e+06)
1.431522e+00	-1.335468e-07
h(7e+06-dataMLB1.Catcher..)	h(dataMLB1.X..Rookies-10)
-3.194314e-07	2.051860e-01
h(10-dataMLB1.X..Rookies)	h(dataMLB1.Infield..-4.68933e+07)
3.832674e-01	6.651180e-08
h(dataMLB1.Outfield..-1.58115e+07)	h(dataMLB1.Batter.Average.Age-28.8)
-3.369155e-08	1.505552e+00
h(dataMLB1.X..players.aquired.via.trade-11)	
-2.382604e-01	

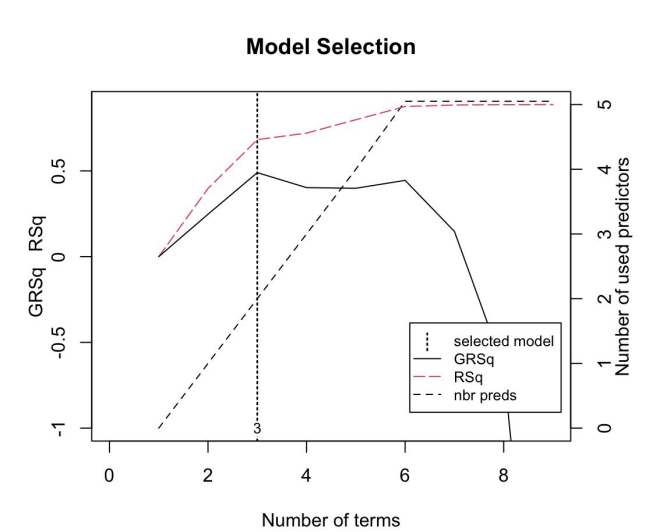


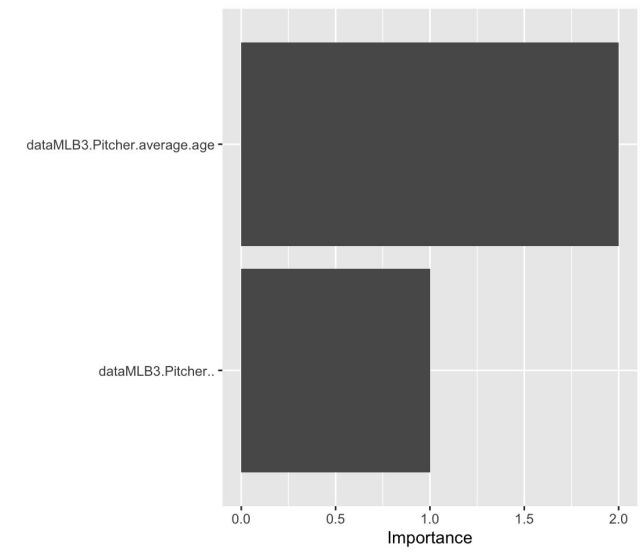


MARS: win% of large market teams

- 3 terms, 2 predictors
- RSq=0.6821049, GRSq = 0.489955

```
 (Intercept) h(29.2-dataMLB3.Pitcher.average.age)    h(6.72196e+07-dataMLB3.Pitcher..)
5.559686e-01                         1.480844e-01                        -2.548250e-09
```

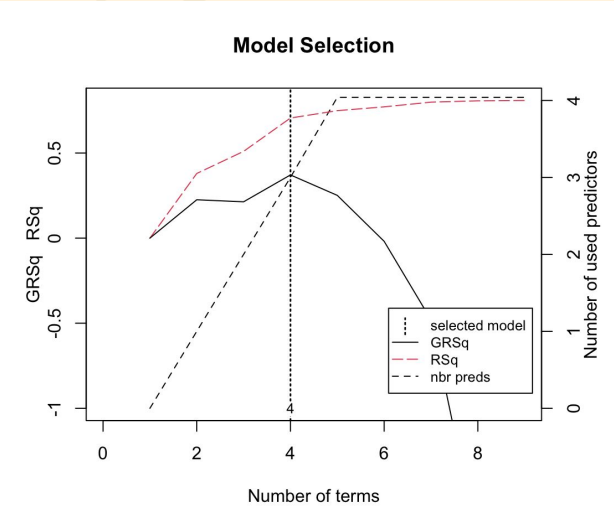


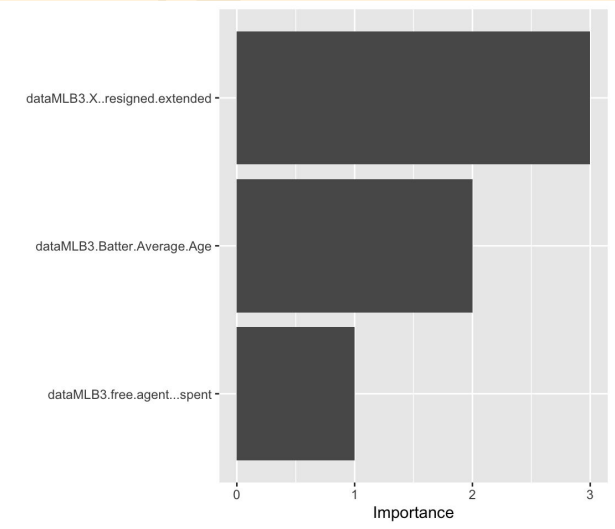


MARS: playoff results of large market teams

- 4 terms, 3 predictors
- RSq=0.7056753, GRSq = 0.3712946

(Intercept)	h(dataMLB3.X..resigned.extended-9)
4.814695e-01	6.556653e-01
h(28.5-dataMLB3.Batter.Average.Age)	h(1.3435e+08-dataMLB3.free.agent...spent)
-3.088181e+00	1.780545e-08





Conclusions

High Winning Percentage:

- Small market teams:
  - **Decrease rookies,** increase pitcher payroll, increase outfield payroll
- Medium market teams:
  - Increase pitcher payroll, decrease total payroll, increase batter average age, increase catcher payroll, decrease number of players re-signed or extended
- Large market teams:
  - Decrease pitcher average age, increase pitcher payroll

Playoff Success:

- Small market teams:
  - **Decrease rookies**
- Medium market teams:
  - Decrease rookies, increase infield payroll, decrease outfield payroll
- Large market teams:
  - Increase number of free agents, increase batter average age, decrease free agent money spent

Future work

- Other clustering techniques (Hierarchical)
- More variables (e.g. team revenue, contract lengths, merchandise sales, ticket prices)
- Look into player performance statistics
- Examine specific world case studies (Royals, Rangers, Dodgers)
- Partial Dependence Plots for better understanding relationships between MARS model features
- Degree 2 MARS for interaction terms
- K-nearest neighbors

Sources

- Spotrac: https://www.spotrac.com/mlb/payroll/2023/, https://www.spotrac.com/mlb/arbitration/, https://www.spotrac.com/mlb/tools/offseason/
- Baseball Reference https://www.baseball-reference.com/leagues/majors/2023-misc.shtml, https://www.baseball-reference.com/leagues/majors/2023-rookies.shtml
- Fangraphs: https://www.baseball-reference.com/postseason/
- MLB Trade Rumors: https://www.mlbtraderumors.com/arbtracker2021
- US Census: https://www.census.gov/content/dam/Census/newsroom/stories/baseball/stories-baseball.pdf
- Sports Media Watch: https://www.sportsmediawatch.com/nba-market-size-nfl-mlb-nhl-nielsen-ratings/
- Wikipedia: https://en.wikipedia.org/wiki/Forbes_list_of_the_most_valuable_MLB_clubs
- Textbook: https://bradleyboehmke.github.io/HOML/mars.html
- MARS documentation: https://cran.r-project.org/web/packages/earth/earth.pdf
- Images:
    - https://www.nbcsports.com/mlb/news/jackson-chourio-gets-8-year-82-million-deal-with-brewers-largest-before-a-players-big-league-debut,
    - https://goldengrizzlies.com/sports/baseball/roster/ronnie-krsolovic/6235
    - https://www.nationalgeographic.com/travel/destination/new-york-city
    - https://www.kcur.org/arts-life/2016-01-21/new-to-kansas-city-guide-things-to-do-best-food