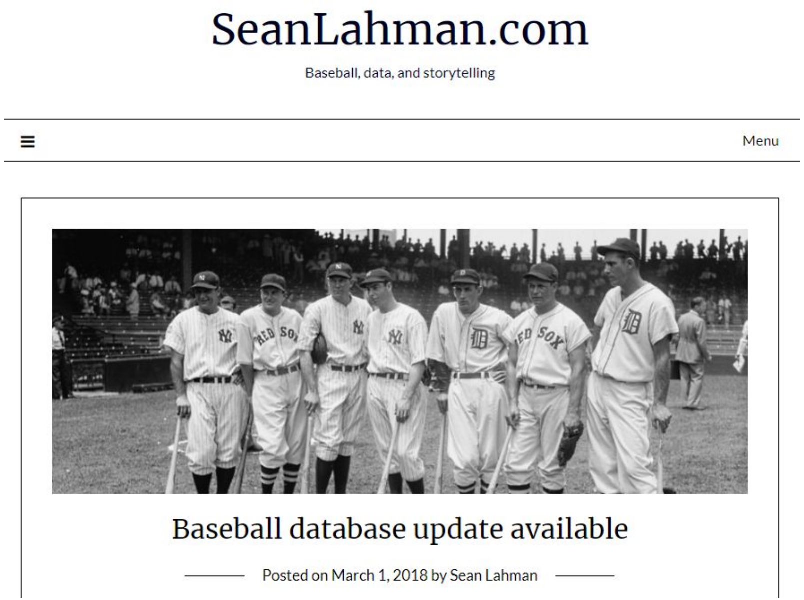
IST 687 – MLB Dataset

Pierre Casco, Parin Patel,
Antonio Llorens, Tim Zalk

Overview of Dataset

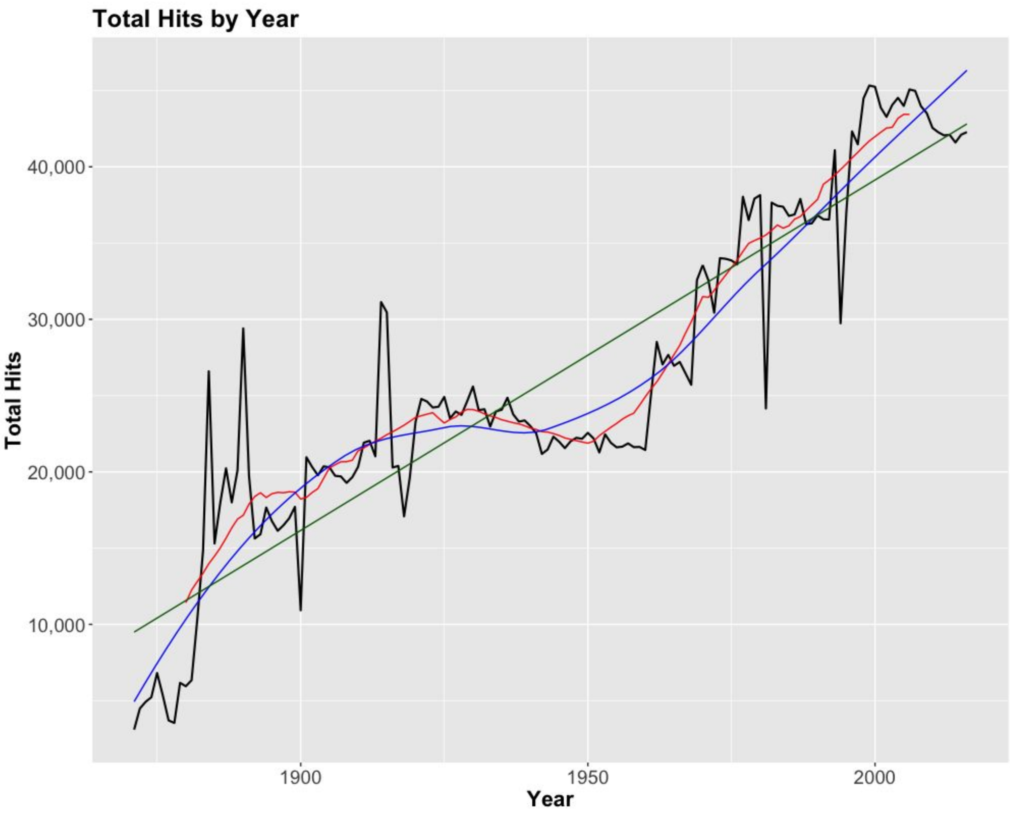
- Major League Baseball dataset spanning from 1871 to 2016 from the `Lahman` package
- The `Lahman` R library contains 27 data frames
  - Dataset includes offense, defense, pitching, salary, all stars, hall of famers, attendance, and team statistics
- Pros:
  - Widely known public database
  - Plenty of resources and discussion boards
  - Has a large, complex volume of data available.
- Cons:
  - Large volume of data:
    - requires extensive cleaning
    - May get “lost in the information”- since there are a lot of variables to look at.
    - Need to be slightly familiar with baseball



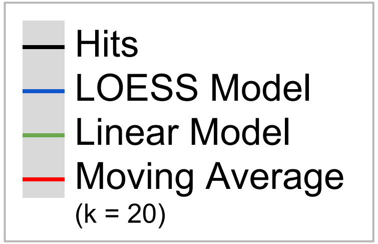
Summary of Project

- Batting Stats – Hits
- Total Games per year
- Batting Stats – Hitsa
- Batting Stats – Runs
- Batting Stats – RBIs
- Batting Stats – Home Runs
- Pitching Stats – ERA
- Total League Annual Salary
- Attendance
- Highest MLB Salary by Year
- Top 5 Average Salaries
- Salary of Hall of Fame Players
- Top 5 Salaries and Team Wins
- Top 5 Salaries and Team Attendance
- Logistic Regression Model for Predicting Hitters’ All Star Appearances
- Logistic Regression Model for Predicting Hitters’ All Star Appearances
- RandomForest Model for Hitters HOF
- Linear Model for Teams Wins
- RandomForest Model for Salary Importance
- World Series Wins using LM()
- Next Steps

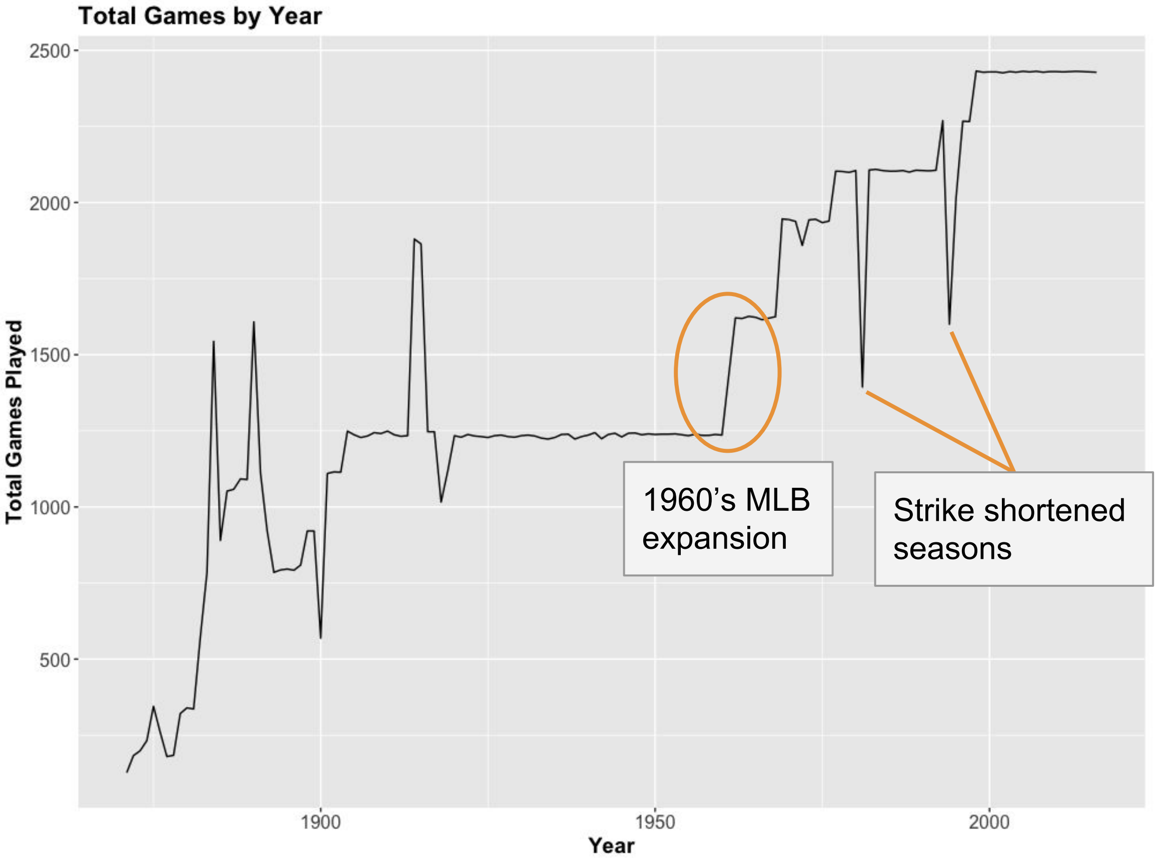
Batting Stats – Hits



- Moving average using `rollmean()` in `zoo` library (k = 20)
- Linear model
  - $R^2 = 0.8241$
  - p-value: $< 2.2e-16$
- LOESS model using `loess()` function in ggplot2 (span = 0.75)

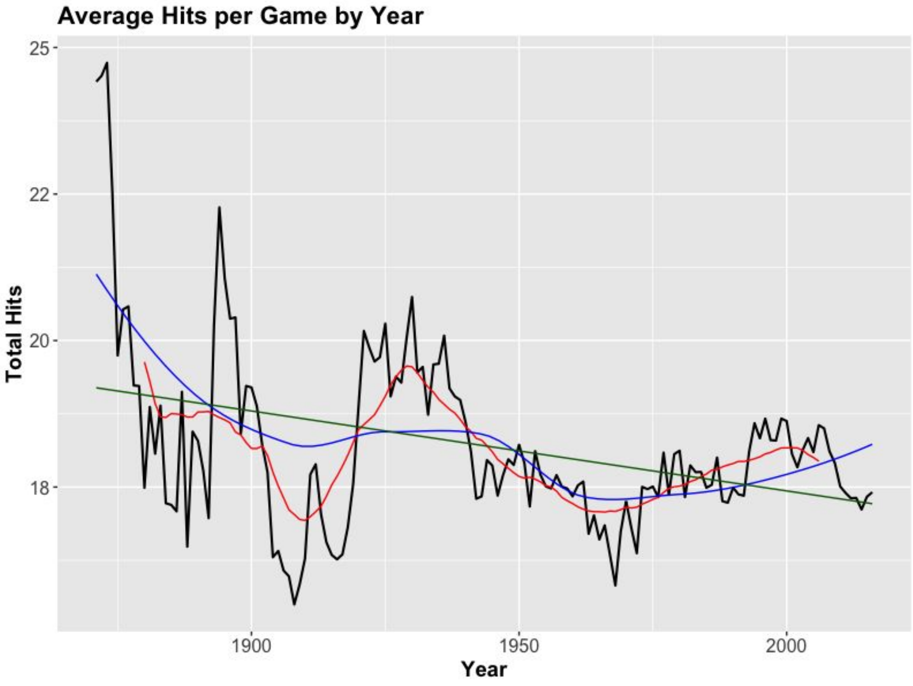


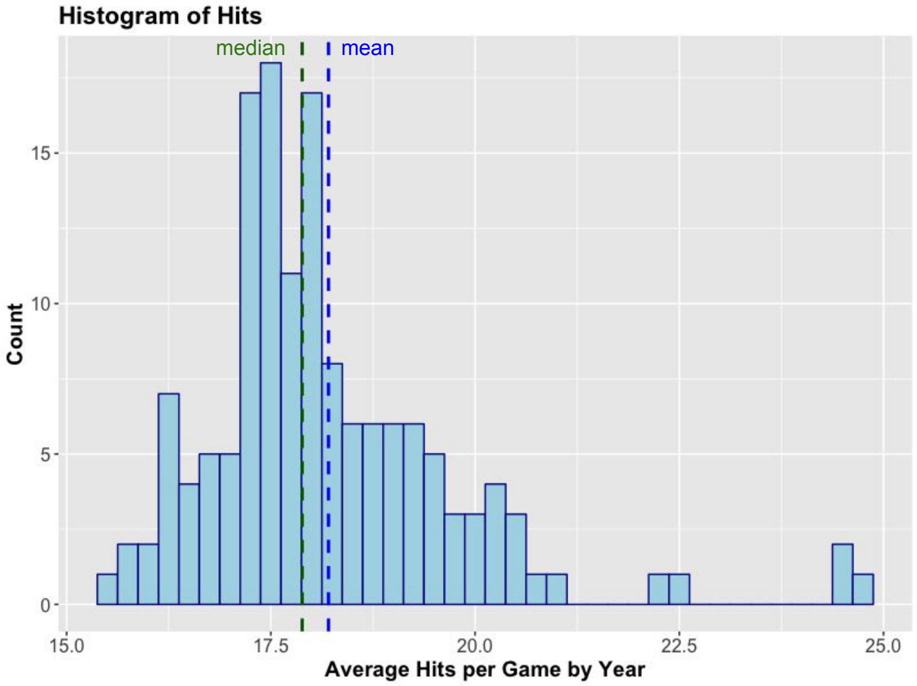
Total Games per year



- MLB currently has 30 teams
  - Only 16 teams in 1903
- Each team has 162 scheduled games per year
  - Some games delayed by weather may be cancelled
  - A 163rd game may be played as a tie-breaker
- 1981 strike cancelled over 700 midseason games
- 1994–1995 strike ended the 1994 season 7 weeks early
- MLB had three leagues in the 1914–1915 seasons, and two since

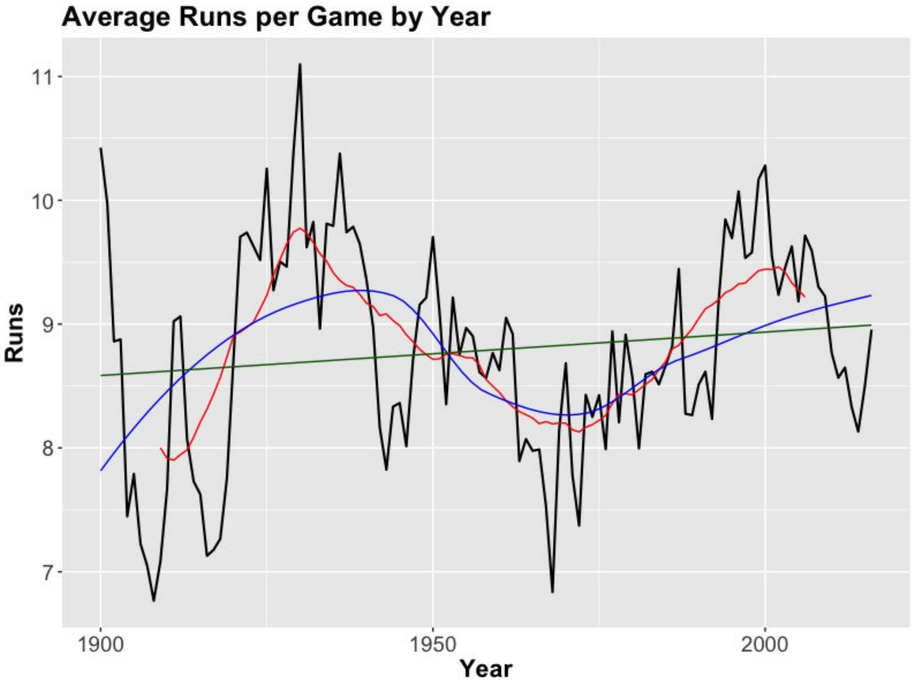
Batting Stats – Hits

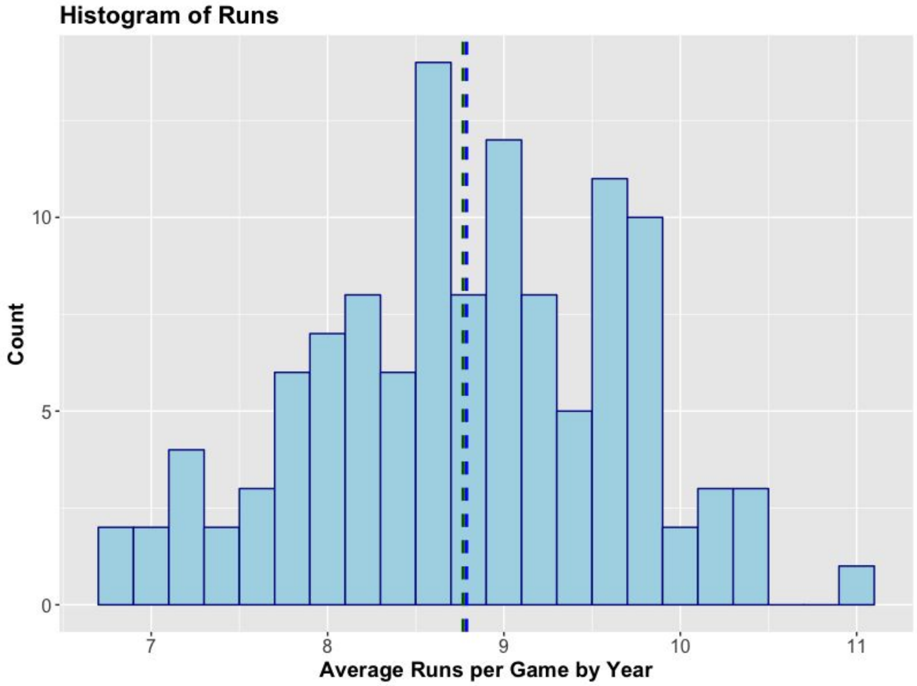




Hits LOESS Model Linear Model Moving Average (k = 20)

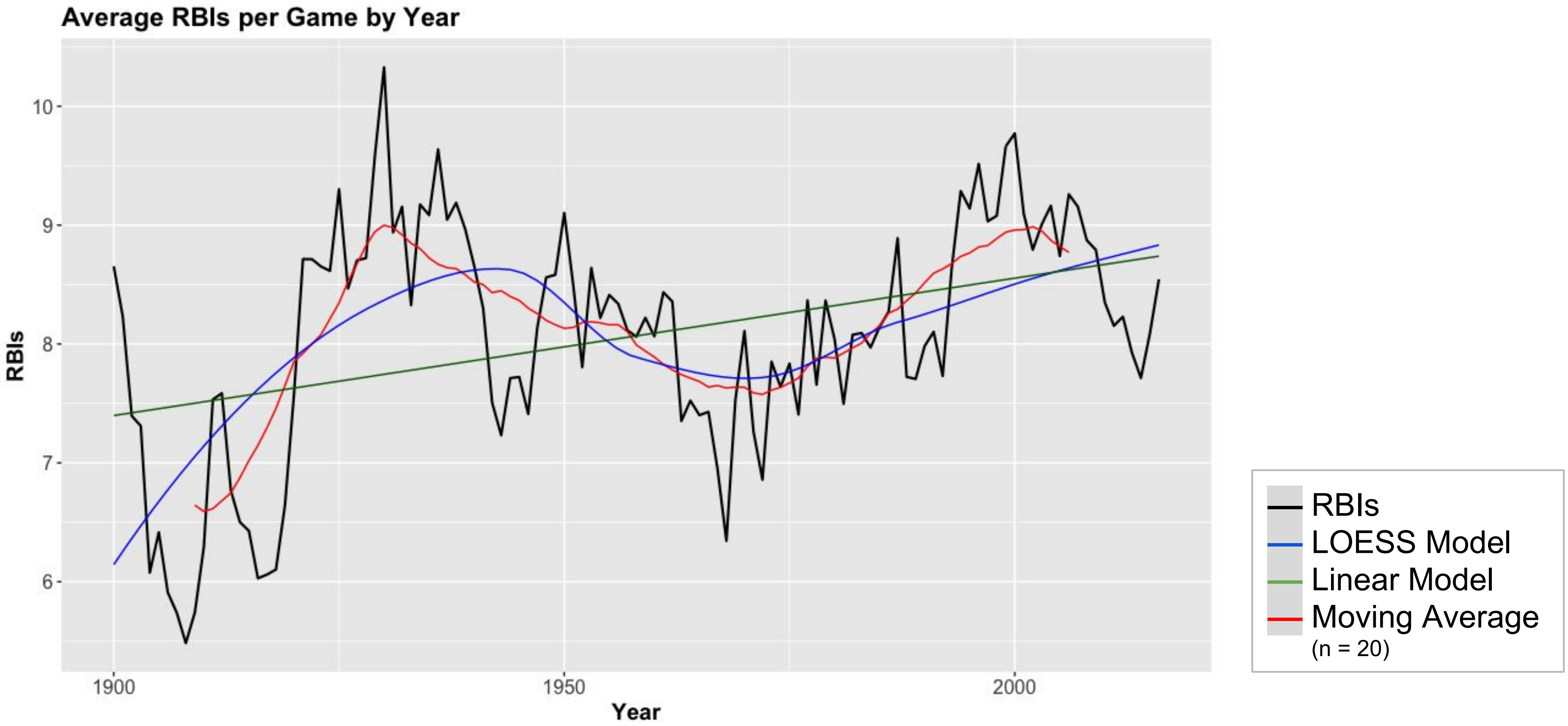
Batting Stats – Runs



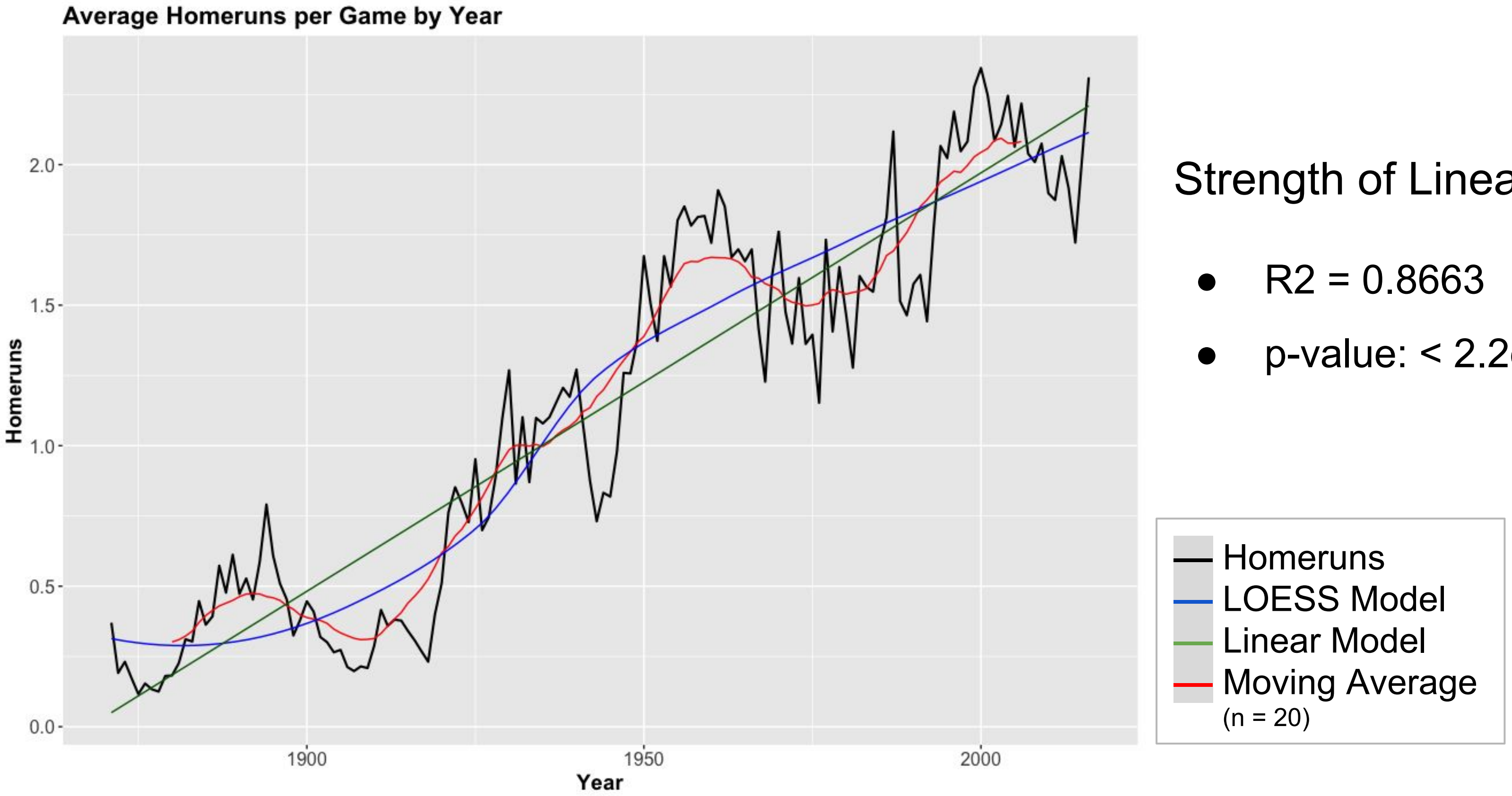


Runs LOESS Model Linear Model Moving Average (k = 20)

Batting Stats – RBIs



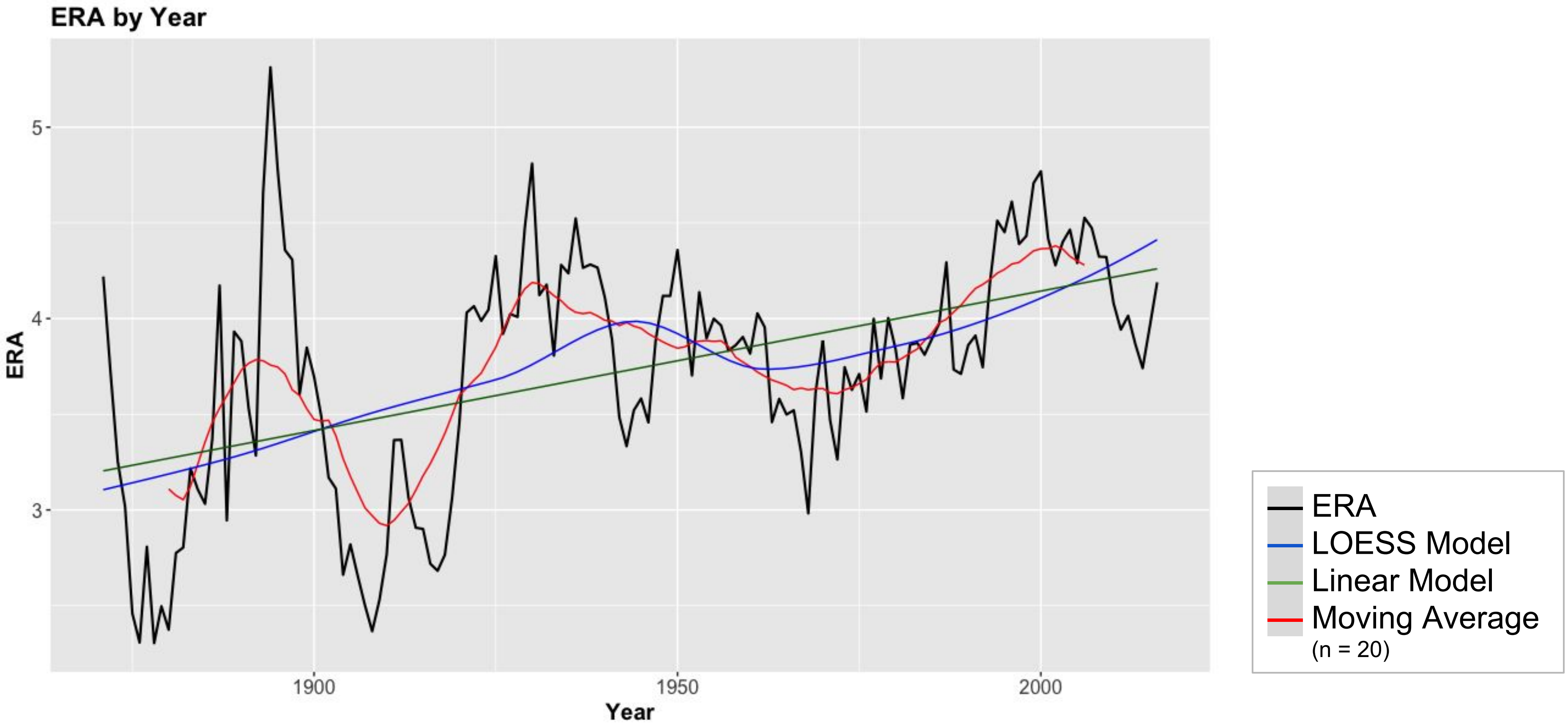
Batting Stats – Home Runs



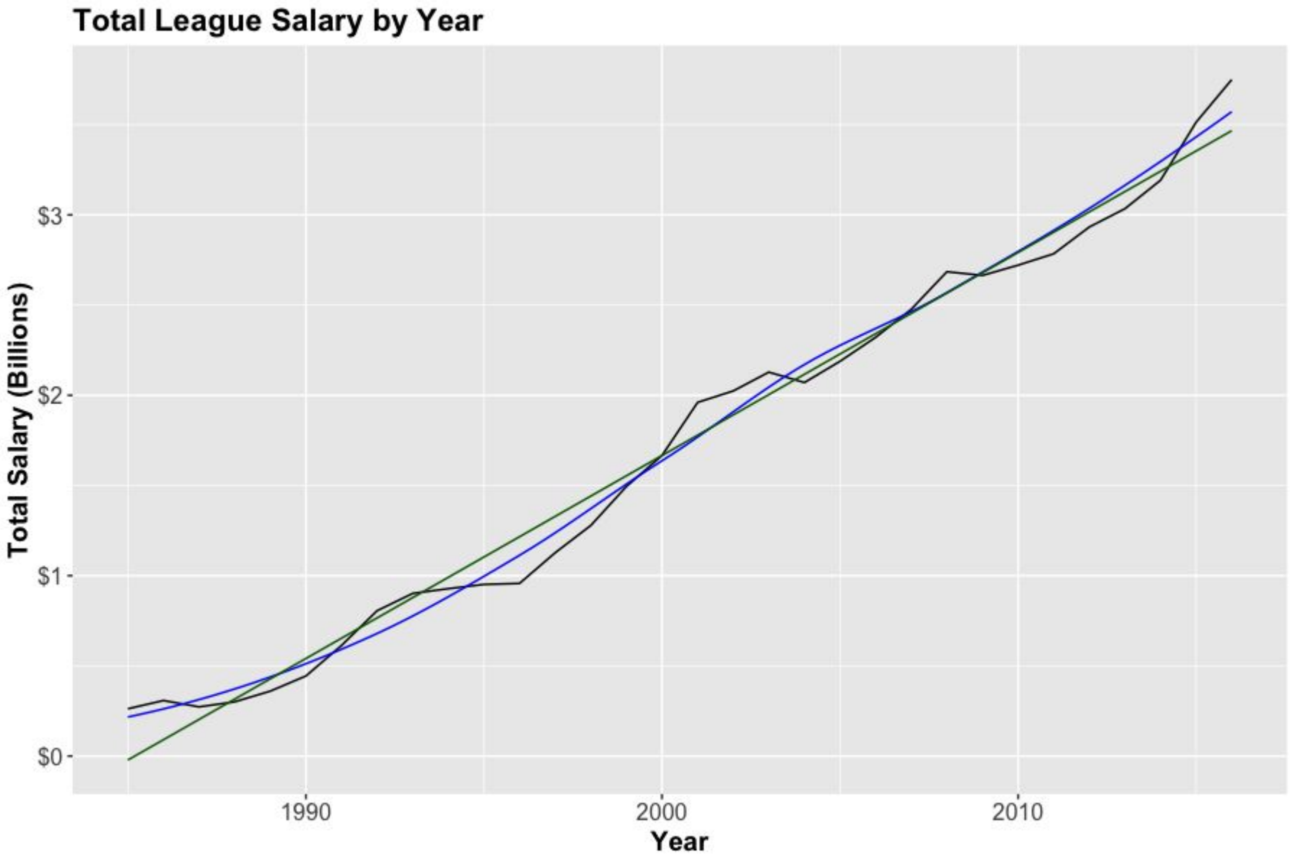
Strength of Linear Model

- R2 = 0.8663
- p-value: < 2.2e-16

Pitching Stats – ERA

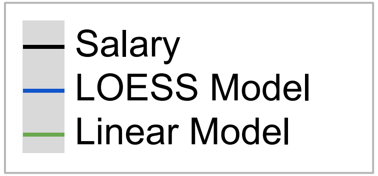


Total League Annual Salary

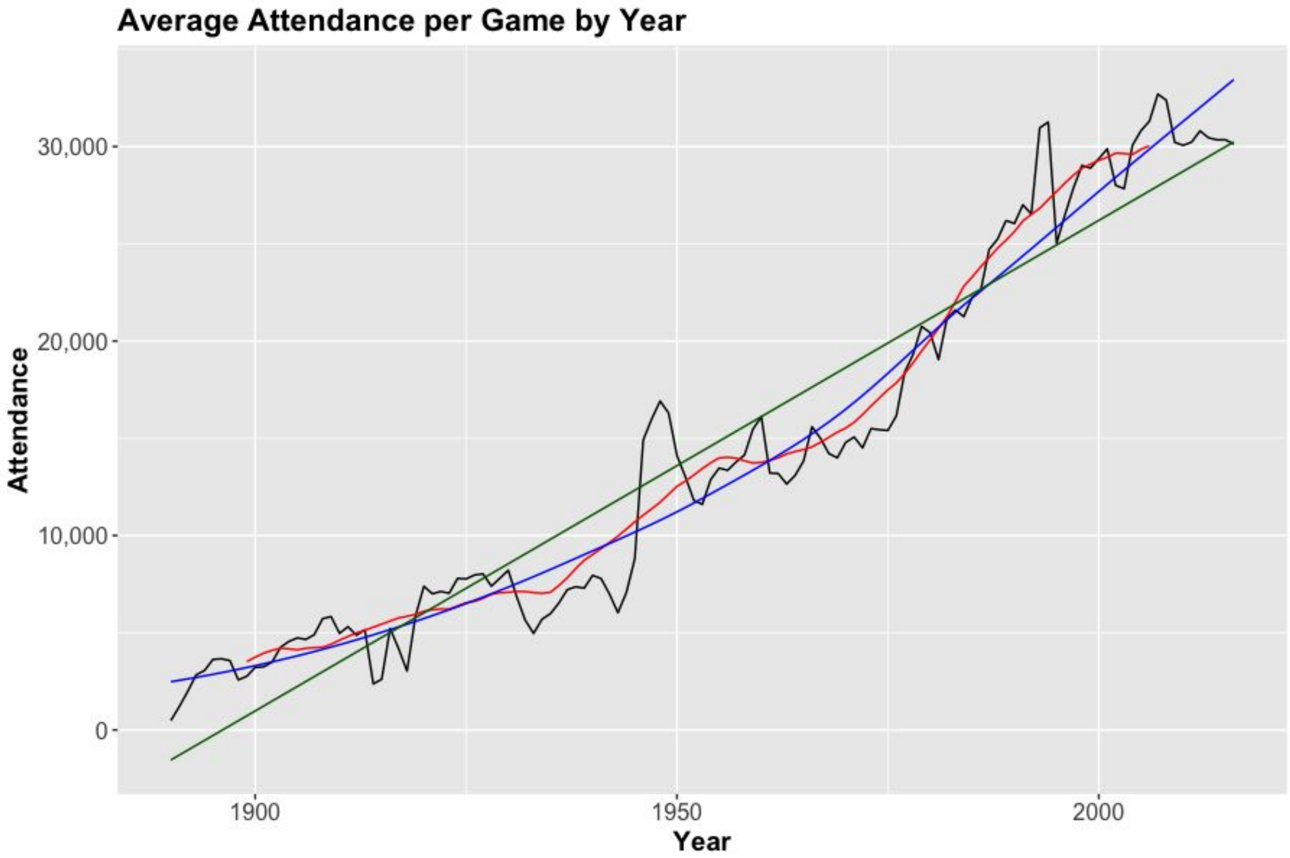


Strength of Linear Model

- $R2 = 0.9845$
- p-value: $< 2.2e-16$

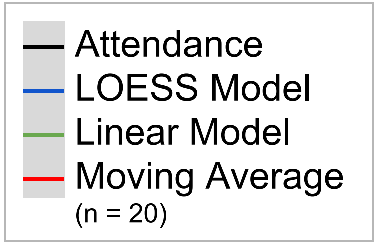


Attendance



Strength of Linear Model

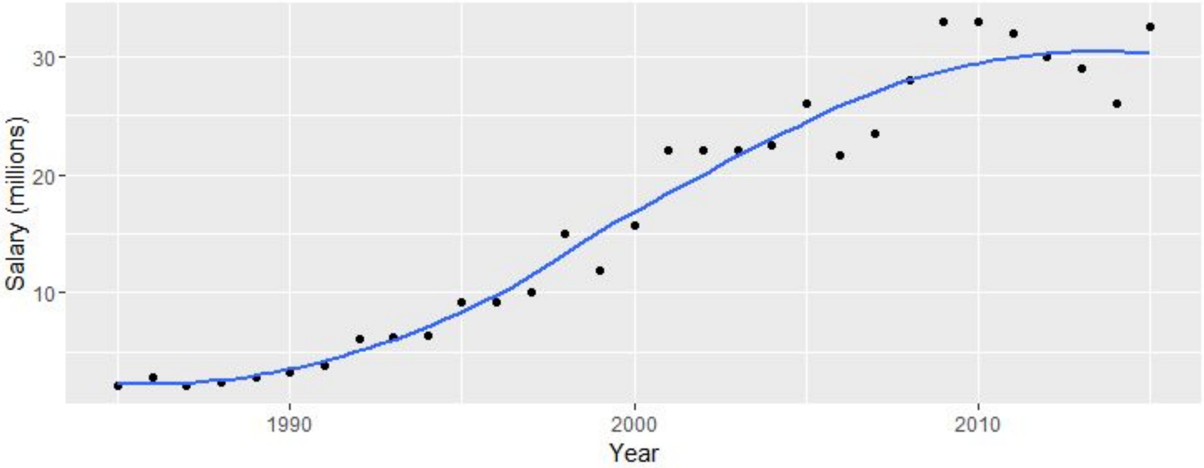
- $R^2 = 0.9181$
- p-value: < 2.2e-16



Highest MLB Salary by Year

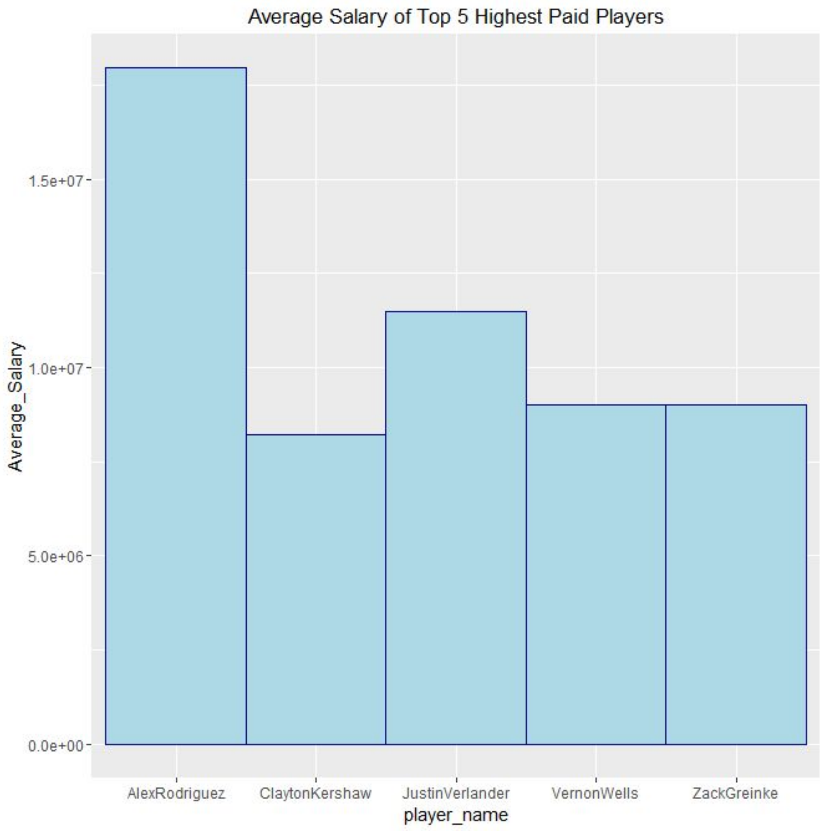
Based on highest player paid per year, grouped by total team salaries

Highest:
Alex Rodriguez- $33,000,000.00

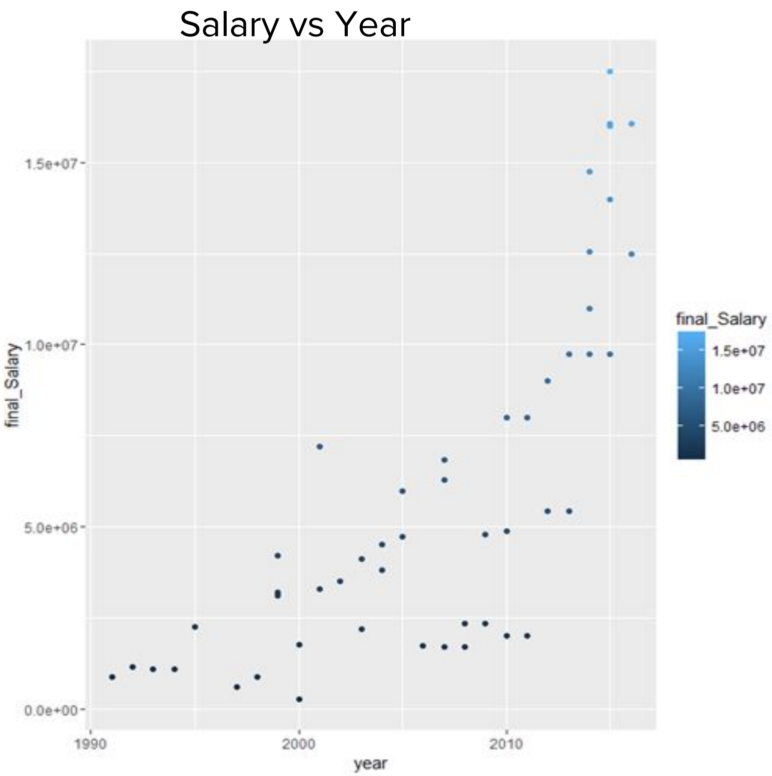


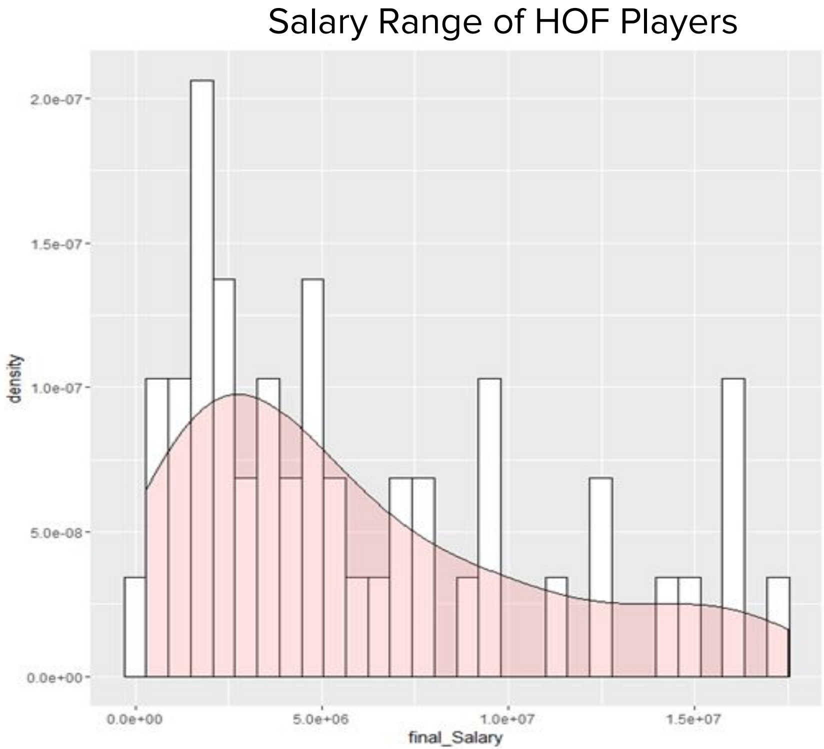
Top 5 Average Salaries

- Alex Rodriguez
- Clayton Kershaw
- Justin Verlander
- Vernon Wells
- Zack Greinke

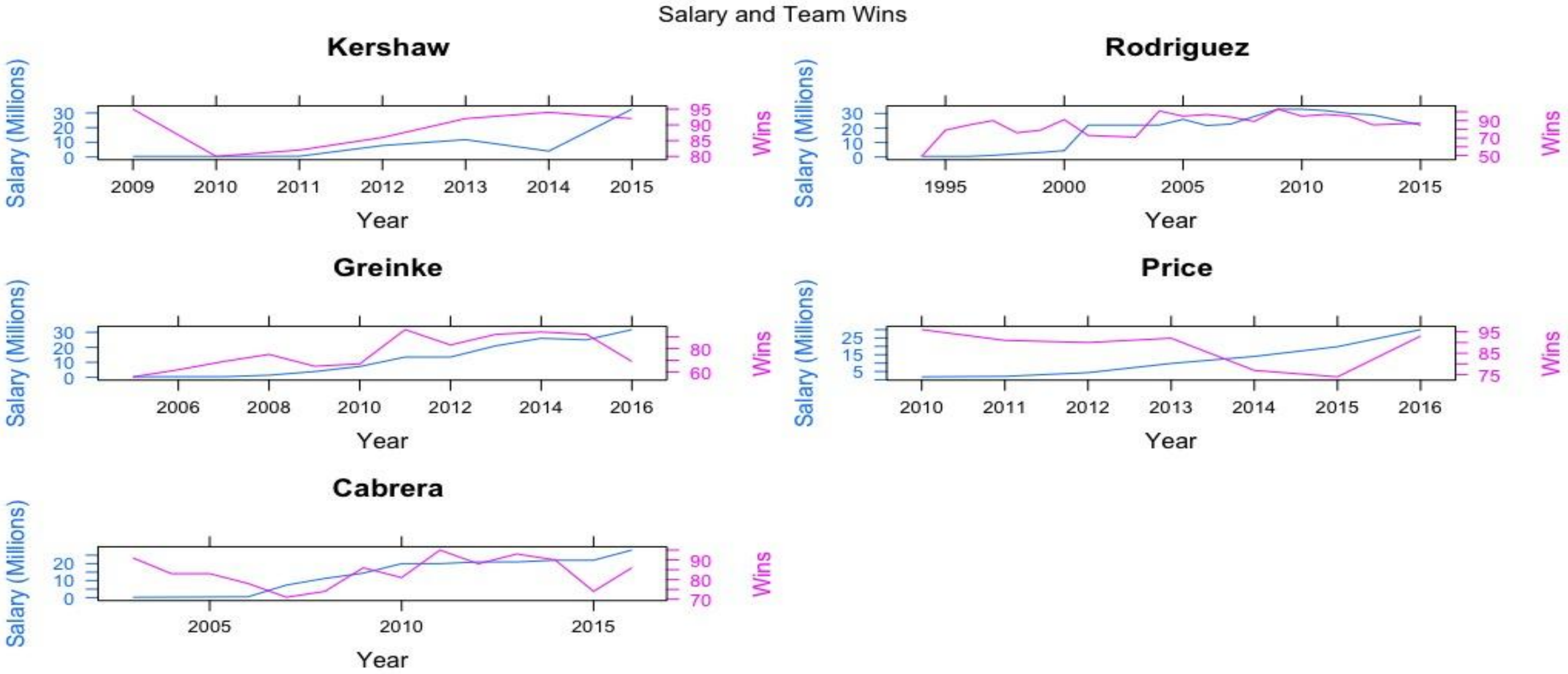


Salary of Hall of Fame Players

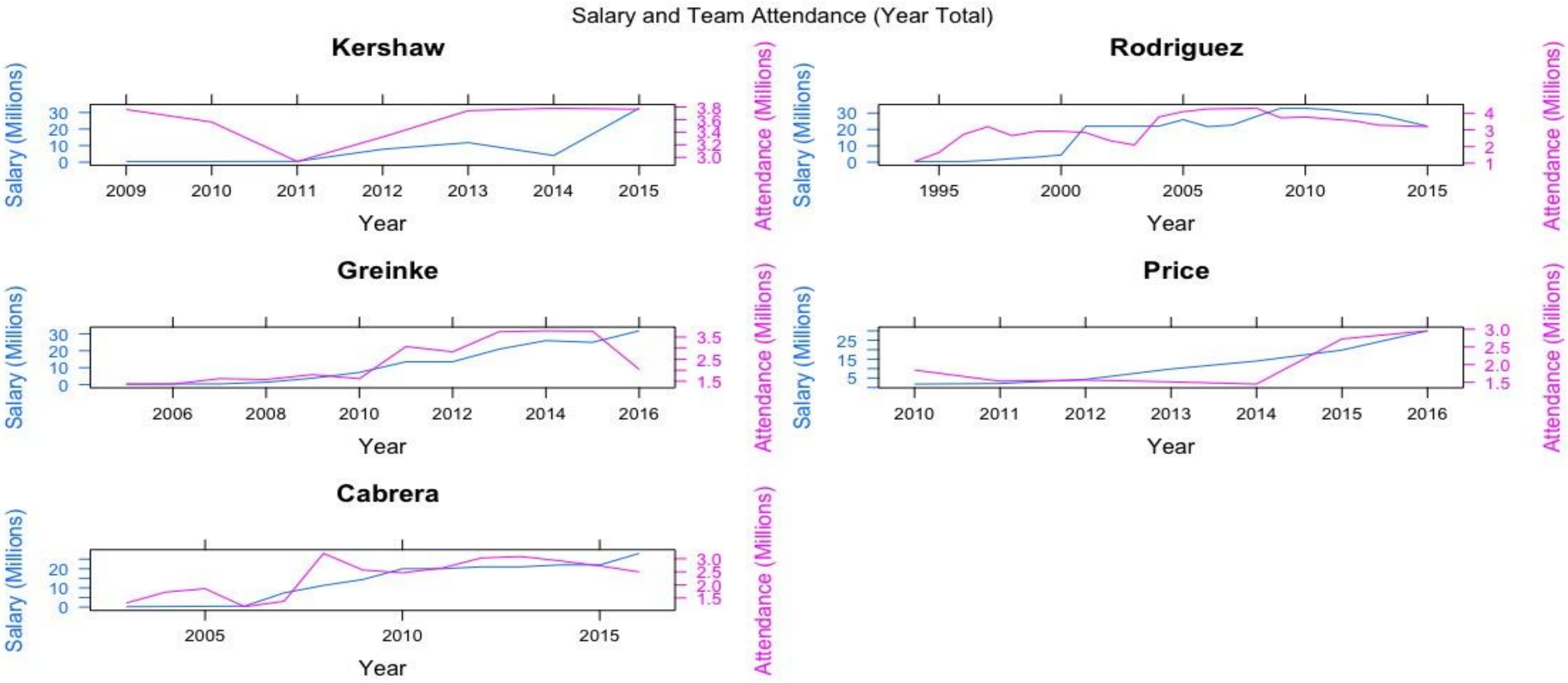




Top 5 Salaries and Team Wins



Top 5 Salaries and Team Attendance



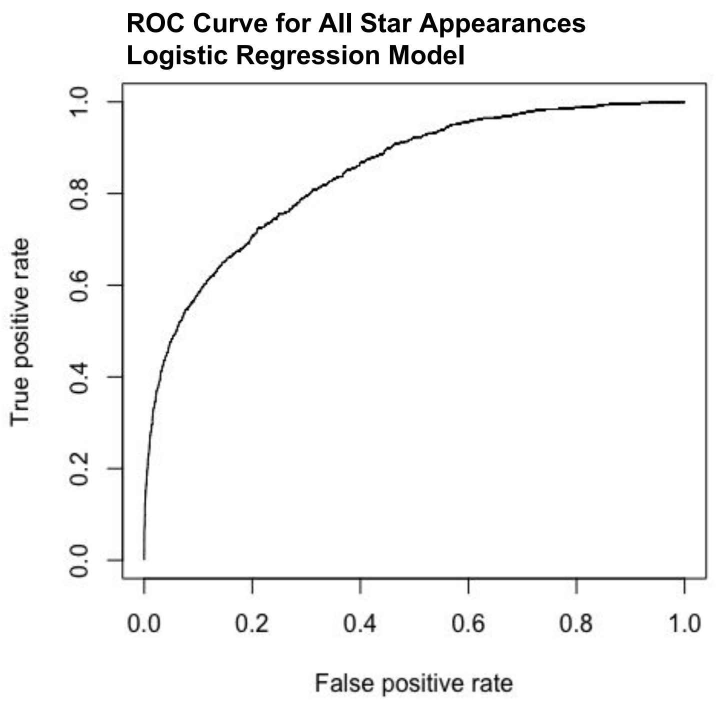
Model for Predicting Hitters' All Star Appearances

- Use only non-categorical variables
- Shuffled rows to randomize order of data, and split data into training and test set (80% train, 20% test)
- Generate the model with `glm()` and a logit model
  - Estimates the likelihood a player will have an All Star appearance based on their stats
- The model determined that most stats were significant in predicting All Star appearances, with the exception of doubles, triples, hit-by-pitch, sacrifice fly, and grounded into double play

```
Call:
glm(formula = asAppearance ~ ., family = binomial(link =
"logit"), data = as.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5490  -0.2799  -0.2135  -0.1714   3.1304
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) 19.7349588  2.7757849   7.110 1.16e-12 ***
birthYear   -0.0159859  0.0013787 -11.595  < 2e-16 ***
weight       0.0091644  0.0013651   6.713 1.90e-11 ***
height       0.0762822  0.0120807   6.314 2.71e-10 ***
G            0.0151576  0.0013901  10.904  < 2e-16 ***
AB          -0.0169273  0.0008828 -19.176  < 2e-16 ***
R            0.0182937  0.0034740   5.266 1.40e-07 ***
H            0.0453672  0.0030362  14.942  < 2e-16 ***
X2B         -0.0046418  0.0050015  -0.928  0.35336
X3B         -0.0110108  0.0119935  -0.918  0.35858
HR           0.0551743  0.0069754   7.910 2.58e-15 ***
RBI          0.0094905  0.0031036   3.058  0.00223 **
SB           0.0243994  0.0036063   6.766 1.33e-11 ***
CS          -0.0474472  0.0105630  -4.492 7.06e-06 ***
BB          -0.0053448  0.0018166  -2.942  0.00326 **
SO          -0.0074015  0.0013097  -5.651 1.59e-08 ***
IBB          0.0792122  0.0073409  10.790  < 2e-16 ***
HBP         -0.0054817  0.0080496  -0.681  0.49588
SH           0.1632589  0.0070484  23.162  < 2e-16 ***
SF           0.0162876  0.0129292   1.260  0.20776
GIDP         0.0112144  0.0065215   1.720  0.08550 .

    Null deviance: 24820  on 53020  degrees of freedom
Residual deviance: 17976  on 53000  degrees of freedom
AIC: 18018

Number of Fisher Scoring iterations: 6
```

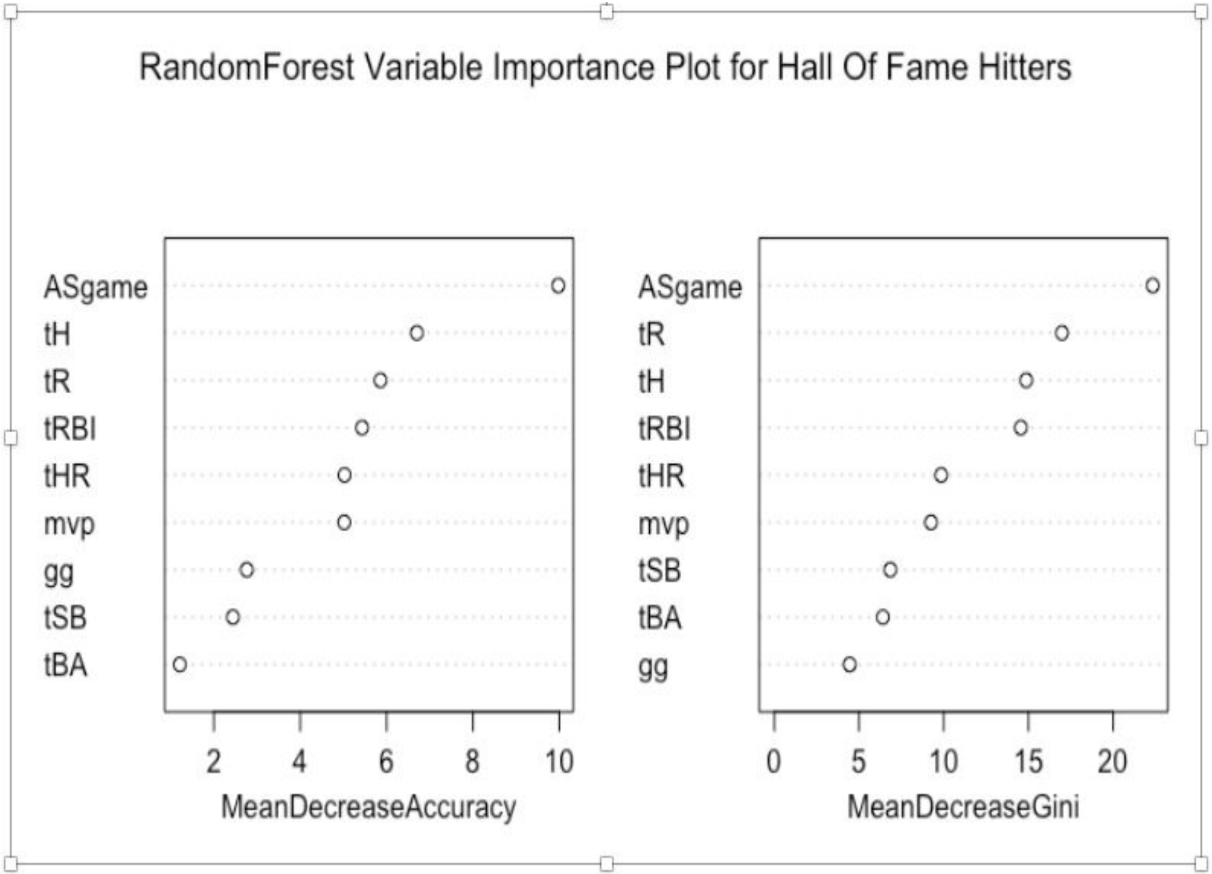
Predicting Hitters' All Star Appearances

- Calculated model accuracy by predicting values for the test data, and comparing to the actual all star appearance each year for each player
- Accuracy for this model was 0.954375
- Plotted the true positive vs true negative of the model using `ROCR` library's `prediction` and `performance` objects
  - Looking for a curve towards high true positive and low false positive
- Calculated area under curve to be 0.8446
  - Looking for a number closer to 1 than to 0.5



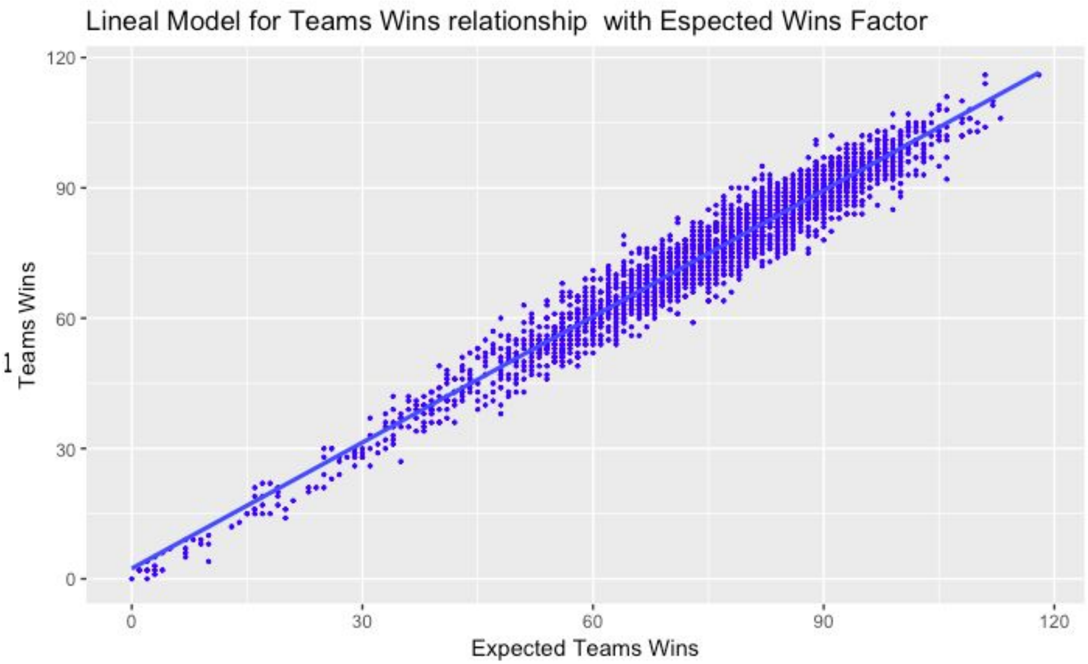
RandomForest Model for Hitters HOF

- All-Star Appearance (ASgame) seem to be picking up a large portion of the variation in Hall of Fame induction for Hitters
- Hits, RBIs,and runs also are significant predictors and best than HRs
- Stolen Bases(SB) and base accepted (BA)don’t have much predictive ability
- Type of random forest: classification. Number of trees;100, Variables each split: 2



Linear Model for Teams Wins

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.353917   0.328461   7.167 9.77e-13 ***
Expwin      0.967597   0.004269 226.681  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.022 on 2833 degrees of freedom
Multiple R-squared:  0.9477,    Adjusted R-squared:  0.9477
F-statistic: 5.138e+04 on 1 and 2833 DF,  p-value: < 2.2e-16
```



Salary Importance using RF

```
> allSalary.rf

Call:
 randomForest(formula = salary ~ teamID + G + AB + R + H + HR +      RBI, data = allSalary, ntree = 100, mtry = 2)
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 2

          Mean of squared residuals: 9.390217e+12
                    % Var explained: 19.27
```

```
> importance(allSalary.rf)
       IncNodePurity
teamID  5.265702e+16
G       3.413666e+16
AB      3.644148e+16
R       2.491169e+16
H       2.589516e+16
HR      2.346352e+16
RBI     2.734937e+16
```

- Used RF to identify factors that may lead to a higher salary
- The model identified team as the most important factor to high salaries followed by at bats and games played
- Teams make sense in baseball as there is no salary cap and teams with more money and known for overpaying players (Yankees, Red Sox, Cubs)

World Series Wins - Importance using LM

- Used `lm()` to look at the total list of World Series winners to determine what factors throughout the season influenced their win
- The model confirmed that in order to be great throughout the season and win a World Series, you need to score runs, and not let the opposition score
- It would confirm the theory that if you score more runs than the other team, you will win 100% of the time

```
Call:
lm(formula = W ~ R + H + X2B + X3B + HR + AB + BB + SO + SB +
    RA + ER + ERA + attendance, data = WSWinners)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7459 -2.3395 -0.2046  2.2947  7.6467

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.058e+02  2.313e+01   4.573 1.53e-05 ***
R            9.640e-02  1.505e-02   6.403 6.75e-09 ***
H           -1.630e-02  1.101e-02  -1.480 0.142241
X2B         -1.935e-02  1.279e-02  -1.513 0.133824
X3B         -2.326e-02  3.350e-02  -0.694 0.489268
HR          -2.653e-02  1.759e-02  -1.508 0.134994
AB          -5.511e-04  5.471e-03  -0.101 0.919981
BB          -8.345e-03  7.003e-03  -1.192 0.236530
SO          -2.803e-03  3.074e-03  -0.912 0.364353
SB          -8.770e-03  9.447e-03  -0.928 0.355723
RA          -6.917e-02  2.071e-02  -3.340 0.001220 **
ER           2.439e-01  6.237e-02   3.911 0.000178 ***
ERA         -3.802e+01  8.372e+00  -4.541 1.73e-05 ***
attendance   5.009e-07  5.782e-07   0.866 0.388578
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.176 on 90 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.8129,    Adjusted R-squared:  0.7859
F-statistic: 30.08 on 13 and 90 DF,  p-value: < 2.2e-16
```

Next Steps...

- Look at pitching and defensive statistics
- Plotting coordinate heatmaps of hit and homerun locations
- Plotting coordinate heatmaps of pitch locations
- Finding correlation with players pre-MLB history to their MLB performance (college, nationality, minor league, etc.)
- SVM model to predict players’ future salaries