Lab 1

Nate Beebe (Worked with Matthew Raitano and Zack Barnes)

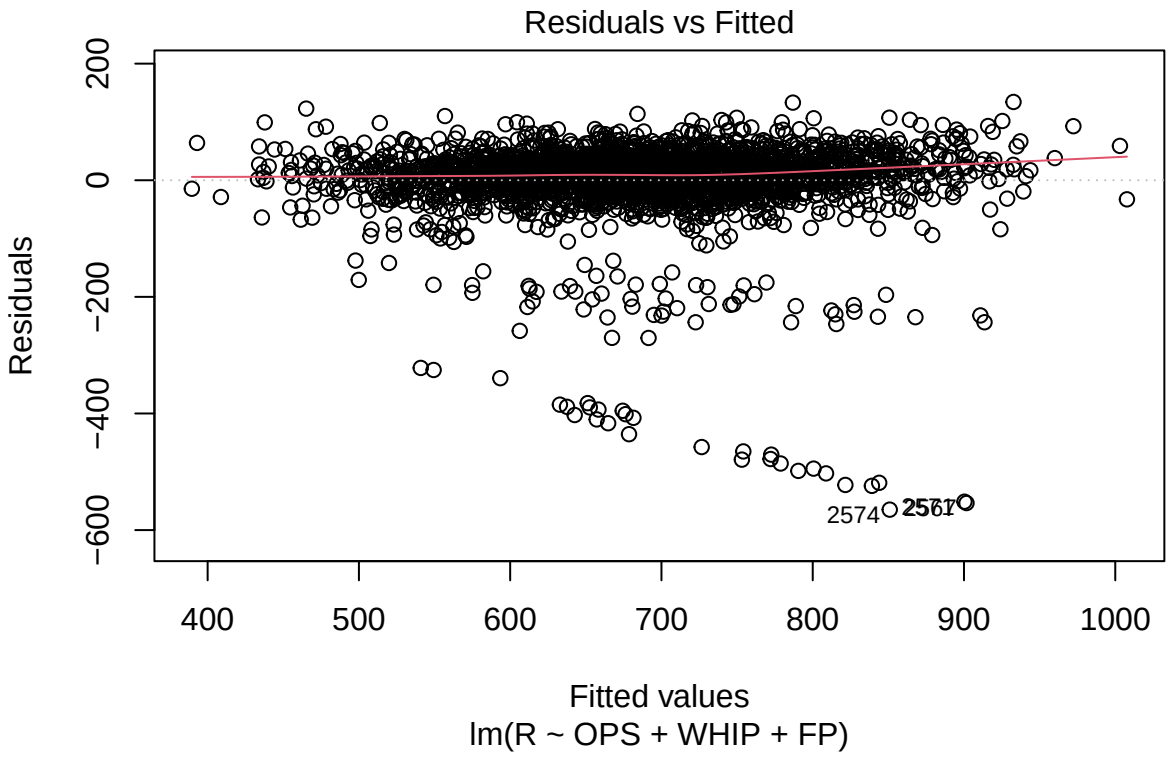
Due on 02/07 at 11:59 PM

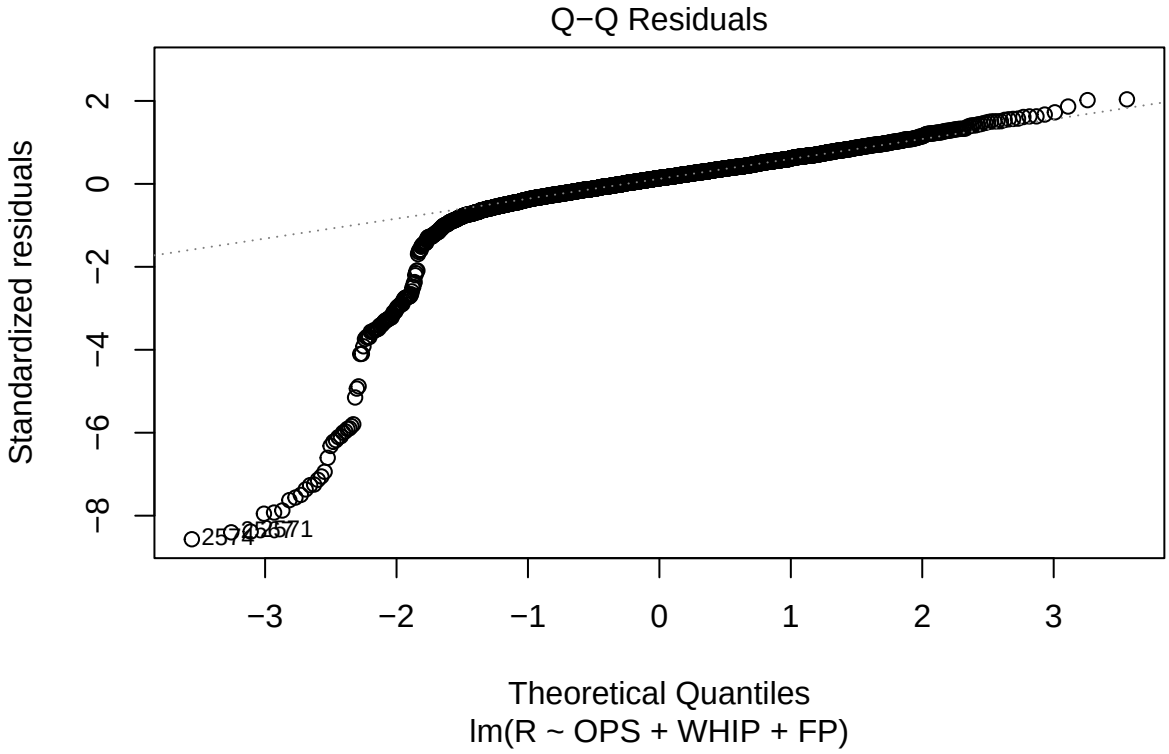
Instructions: This lab report needs to be professional. Only report relevant and finalized code. Your writing should be concise and void of spelling errors. Use code chunk options to hide unnecessary messages/warnings. Your report should be reproducible. Reports that involve simulations need to have the random seed specified so that simulation results are reproducible. You are allowed to work on this lab assignment in groups of 2-3. You still need to submit an individual lab report if you do work in a group, and you need to list your collaborators.

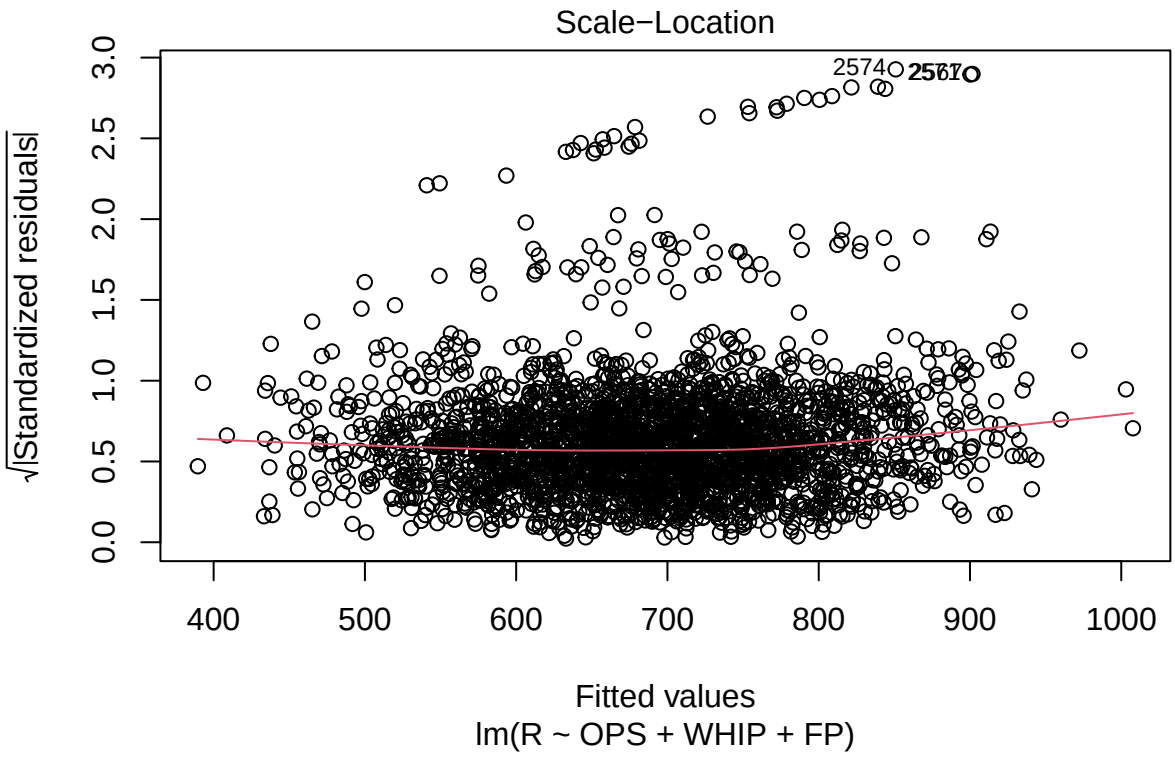
Question 1 In lecture it was demonstrated that baseball is a game of offense, pitching, and defense with a regression model that considered expected run differential as a function of explanatory variables OPS, WHIP, and FP. Do the following:

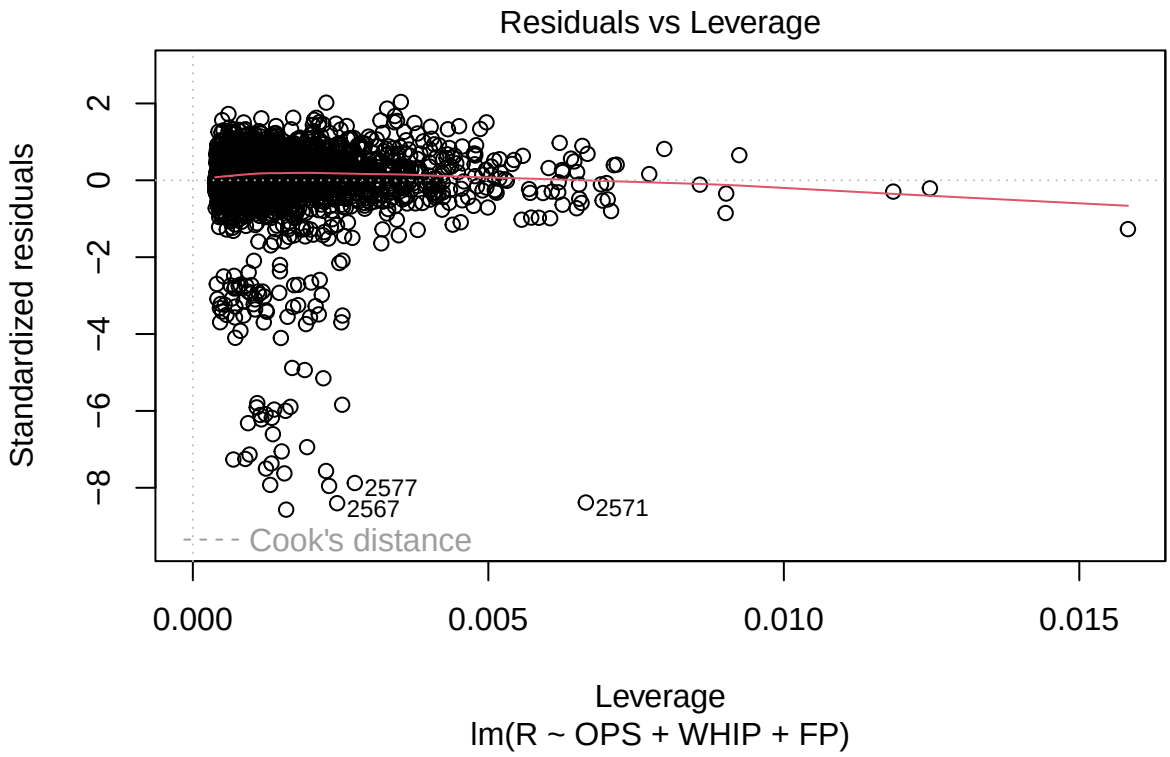
- Fit a similar regression model with runs as the response variable. Report problems with this model. Investigate problematic residuals to discover what went wrong. Fix the problem with this model by adding categorical variable(s) to the list of explanatory variables. Briefly explain what went wrong.

```
##
## Call:
## lm(formula = R ~ OPS + WHIP + FP, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -564.92  -13.31    9.19   29.39  134.26
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1094.19     147.24   7.431 1.44e-13 ***
## OPS          1954.78      29.26  66.799  < 2e-16 ***
## WHIP          -49.20      12.04  -4.087 4.51e-05 ***
## FP          -1778.14     156.64 -11.352  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.98 on 2666 degrees of freedom
## Multiple R-squared:  0.6709, Adjusted R-squared:  0.6705
## F-statistic:  1812 on 3 and 2666 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
Residuals
200
0
−200
−400
−600
400
500
600
700
800
900
1000
2574
Fitted values
lm(R ~ OPS + WHIP + FP)


Q−Q Residuals
Standardized residuals
2
0
−2
−4
−6
−8
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(R ~ OPS + WHIP + FP)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2574
400
500
600
700
800
900
1000
Fitted values
lm(R ~ OPS + WHIP + FP)




```
##    yearID     resid
## 1    2020 -564.9246
## 2    2020 -553.6604
## 3    2020 -551.3131
## 4    2020 -524.1296
## 5    2020 -522.5782
## 6    2020 -518.9455
## 7    2020 -502.7991
## 8    2020 -498.4942
## 9    2020 -494.5779
## 10   2020 -485.7926
## 11   2020 -479.1674
## 12   2020 -478.1305
## 13   2020 -470.5878
## 14   2020 -465.1171
## 15   2020 -457.6092
## 16   2020 -435.5690
## 17   2020 -416.7568
## 18   2020 -410.1996
## 19   2020 -407.4747
## 20   2020 -402.6481
## 21   2020 -401.2635
## 22   2020 -395.4104
## 23   2020 -393.3538
## 24   2020 -389.6302
## 25   2020 -388.5658
```

```
## 26   2020 -384.8624
## 27   2020 -382.2590
## 28   2020 -339.4788
## 29   2020 -325.4069
## 30   2020 -321.8813
## 31   1981 -270.4836
```

```
##
## Call:
## lm(formula = R ~ OPS + WHIP + FP + COVID, data = dat3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -280.455  -16.829    3.655   24.308  131.162
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   707.630    101.938   6.942 4.84e-12 ***
## OPS          1977.001     20.214  97.803  < 2e-16 ***
## WHIP          -62.270      8.318  -7.487 9.55e-14 ***
## FP          -1374.908    108.432 -12.680  < 2e-16 ***
## COVIDTRUE    -454.493      8.404 -54.081  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45.57 on 2665 degrees of freedom
## Multiple R-squared:  0.8431, Adjusted R-squared:  0.8429
## F-statistic:  3580 on 4 and 2665 DF,  p-value: < 2.2e-16
```

The shortened 60 game 2020 season due to COVID-19 did not fit the mold of the model that used seasons with more than 100 games for every other year. The total number of runs was significantly lower in this year than any other. The residuals for each of the 30 teams during that year were the 30 highest absolute value residuals in the entire model (and over 30 units greater than the next closest). Adding the 'COVID' variable allows for a big change in the model when the year was 2020.

- We can significantly improve the regression model in the notes through a principled rescaling of OPS, WHIP, and FP. Split the Teams data frame by `yearID` and, for each year, create variables `OPSscale = OPS/avgOPS`, `WHIPscale = avgWHIP/WHIP`, and `FPscale = FP/avgFP` which require you to first create league average variables `avgOPS`, `avgWHIP`, and `avgFP`. Fit the linear regression model with runs differential as the response and explanatory variables `OPSscale`, `WHIPscale`, and `FPscale`, and report relevant output. Why does this model perform so much better than the model in the notes? Support your answer.

```
##
## Call:
## lm(formula = RD ~ OPSscale + WHIPscale + FPscale, data = scale)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07832 -0.17510  0.00045  0.17746  0.95609
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -38.76526    1.54820  -25.04   <2e-16 ***
## OPSscale      8.97615    0.10516   85.35   <2e-16 ***
## WHIPscale     7.01178    0.08617   81.38   <2e-16 ***
## FPscale      22.75554    1.59649   14.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2659 on 2666 degrees of freedom
## Multiple R-squared:  0.8775, Adjusted R-squared:  0.8774
## F-statistic:  6369 on 3 and 2666 DF,  p-value: < 2.2e-16
```

This model works better because it compares each team relative to their peers. Baseball's offensive output has changed a lot over its lifespan, and as it evolved, so did offensive and defensive output (leading to differences in run differential). A team from 1900 shouldn't be put on the same scale as a team from 2023.

Question 2 Choose 3 batters and 3 pitchers that have played in at least 10 seasons and do the following:

Batters selected: David Ortiz, Salvador Perez, and Lorenzo Cain; Pitchers selected: Madison Bumgarner, Wade Davis, and Jacob deGrom.

- Display the seasonal statistics for these players. The following statistics should be included for batters (derivations of unconventional statistics are in parentheses): year, G, AB, R, H, X2B, X3B, HR, RBI, SB, CS, SBpct (SB / (SB + CS)), BB, SO, OBP, SLG, OPS. The following statistics should be included for pitchers: year, W, L, IPouts, H, ER, HR, BB, HBP, SO, ERA, WHIP, SOper9 (SO / IP * 9), SOperBB (SO / BB). These statistics can be found in or computed from statistics that are found in the `Batting` and `Pitching` dataframes in the `Lahman` package.

```
##    nameFirst nameLast year   G  AB   R   H X2B X3B HR RBI SB CS SBpct  BB  SO
## 1     Lorenzo     Cain 2010  43 147  17  45  11   1  1  13  7  1 0.875   9  28
## 2     Lorenzo     Cain 2011   6  22   4   6   1   0  0   1  0  0   NaN   1   4
## 3     Lorenzo     Cain 2012  61 222  27  59   9   2  7  31 10  0 1.000  15  56
## 4     Lorenzo     Cain 2013 115 399  54 100  21   3  4  46 14  6 0.700  33  90
## 5     Lorenzo     Cain 2014 133 471  55 142  29   4  5  53 28  5 0.848  24 108
## 6     Lorenzo     Cain 2015 140 551 101 169  34   6 16  72 28  6 0.824  37  98
## 7     Lorenzo     Cain 2016 103 397  56 114  19   1  9  56 14  5 0.737  31  84
## 8     Lorenzo     Cain 2017 155 584  86 175  27   5 15  49 26  2 0.929  54 100
## 9     Lorenzo     Cain 2018 141 539  90 166  25   2 10  38 30  7 0.811  71  94
## 10    Lorenzo     Cain 2019 148 562  75 146  30   0 11  48 18  8 0.692  50 106
## 11    Lorenzo     Cain 2020   5  18   4   6   1   0  0   2  0  0   NaN   3   2
## 12    Lorenzo     Cain 2021  78 257  40  66  13   0  8  36 13  2 0.867  26  48
## 13    Lorenzo     Cain 2022  43 145  17  26   5   0  1   9  2  2 0.500   8  36
## 14      David    Ortiz 1997  15  49  10  16   3   0  1   6  0  0   NaN   2  19
## 15      David    Ortiz 1998  86 278  47  77  20   0  9  46  1  0 1.000  39  72
## 16      David    Ortiz 1999  10  20   1   0   0   0  0   0  0  0   NaN   5  12
## 17      David    Ortiz 2000 130 415  59 117  36   1 10  63  1  0 1.000  57  81
## 18      David    Ortiz 2001  89 303  46  71  17   1 18  48  1  0 1.000  40  68
## 19      David    Ortiz 2002 125 412  52 112  32   1 20  75  1  2 0.333  43  87
## 20      David    Ortiz 2003 128 448  79 129  39   2 31 101  0  0   NaN  58  83
## 21      David    Ortiz 2004 150 582  94 175  47   3 41 139  0  0   NaN  75 133
## 22      David    Ortiz 2005 159 601 119 180  40   1 47 148  1  0 1.000 102 124
## 23      David    Ortiz 2006 151 558 115 160  29   2 54 137  1  0 1.000 119 117
## 24      David    Ortiz 2007 149 549 116 182  52   1 35 117  3  1 0.750 111 103
## 25      David    Ortiz 2008 109 416  74 110  30   1 23  89  1  0 1.000  70  74
## 26      David    Ortiz 2009 150 541  77 129  35   1 28  99  0  2 0.000  74 134
```

```
## 27     David    Ortiz 2010 145 518  86 140  36   1 32 102  0  1 0.000  82 145
## 28     David    Ortiz 2011 146 525  84 162  40   1 29  96  1  1 0.500  78  83
## 29     David    Ortiz 2012  90 324  65 103  26   0 23  60  0  1 0.000  56  51
## 30     David    Ortiz 2013 137 518  84 160  38   2 30 103  4  0 1.000  76  88
## 31     David    Ortiz 2014 142 518  59 136  27   0 35 104  0  0   NaN  75  95
## 32     David    Ortiz 2015 146 528  73 144  37   0 37 108  0  1 0.000  77  95
## 33     David    Ortiz 2016 151 537  79 169  48   1 38 127  2  0 1.000  80  86
## 34  Salvador    Perez 2011  39 148  20  49   8   2  3  21  0  0   NaN   7  20
## 35  Salvador    Perez 2012  76 289  38  87  16   0 11  39  0  0   NaN  12  27
## 36  Salvador    Perez 2013 138 496  48 145  25   3 13  79  0  0   NaN  21  63
## 37  Salvador    Perez 2014 150 578  57 150  28   2 17  70  1  0 1.000  22  85
## 38  Salvador    Perez 2015 142 531  52 138  25   0 21  70  1  0 1.000  13  82
## 39  Salvador    Perez 2016 139 514  57 127  28   2 22  64  0  0   NaN  22 119
## 40  Salvador    Perez 2017 129 471  57 126  24   1 27  80  1  0 1.000  17  95
## 41  Salvador    Perez 2018 129 510  52 120  23   0 27  80  1  1 0.500  17 108
## 42  Salvador    Perez 2020  37 150  22  50  12   0 11  32  1  0 1.000   3  36
## 43  Salvador    Perez 2021 161 620  88 169  24   0 48 121  1  0 1.000  28 170
## 44  Salvador    Perez 2022 114 445  48 113  23   1 23  76  0  0   NaN  18 109
## 45  Salvador    Perez 2023 140 538  59 137  21   0 23  80  0  0   NaN  19 135
##       OBP   SLG   OPS
## 1   0.348 0.415 0.763
## 2   0.304 0.318 0.622
## 3   0.316 0.419 0.735
## 4   0.310 0.348 0.658
## 5   0.339 0.412 0.751
## 6   0.361 0.477 0.838
## 7   0.339 0.408 0.747
## 8   0.363 0.440 0.803
## 9   0.395 0.417 0.812
## 10 0.325 0.372 0.697
## 11 0.429 0.389 0.818
## 12 0.329 0.401 0.730
## 13 0.231 0.234 0.465
## 14 0.353 0.449 0.802
## 15 0.371 0.446 0.817
## 16 0.200 0.000 0.200
## 17 0.364 0.446 0.810
## 18 0.324 0.475 0.799
## 19 0.339 0.500 0.839
## 20 0.369 0.592 0.961
## 21 0.380 0.603 0.983
## 22 0.397 0.604 1.001
## 23 0.413 0.636 1.049
## 24 0.445 0.621 1.066
## 25 0.369 0.507 0.876
## 26 0.332 0.462 0.794
## 27 0.370 0.529 0.899
## 28 0.398 0.554 0.952
## 29 0.415 0.611 1.026
## 30 0.395 0.564 0.959
## 31 0.355 0.517 0.872
## 32 0.360 0.553 0.913
## 33 0.401 0.620 1.021
## 34 0.361 0.473 0.834
```

```
## 35 0.328 0.471 0.799
## 36 0.323 0.433 0.756
## 37 0.289 0.403 0.692
## 38 0.280 0.426 0.706
## 39 0.288 0.438 0.726
## 40 0.297 0.495 0.792
## 41 0.274 0.439 0.713
## 42 0.353 0.633 0.986
## 43 0.316 0.544 0.860
## 44 0.292 0.465 0.757
## 45 0.292 0.422 0.714
```

```
##    nameFirst  nameLast year  W  L IPouts   H ER HR BB HBP  SO   ERA  WHIP
## 1    Madison Bumgarner 2009  0  0     30   8  2  2  3   0  10  1.80 1.100
## 2    Madison Bumgarner 2010  7  6    333 119 37 11 26   5  86  3.00 1.306
## 3    Madison Bumgarner 2011 13 13    614 202 73 12 46   5 191  3.21 1.212
## 4    Madison Bumgarner 2012 16 11    625 183 78 23 49   7 191  3.37 1.114
## 5    Madison Bumgarner 2013 13  9    604 146 62 15 62   6 199  2.77 1.033
## 6    Madison Bumgarner 2014 18 10    652 194 72 21 43   6 219  2.98 1.091
## 7    Madison Bumgarner 2015 18  9    655 181 71 21 39   7 234  2.93 1.008
## 8    Madison Bumgarner 2016 15  9    680 179 69 26 54   8 251  2.74 1.028
## 9    Madison Bumgarner 2017  4  9    333 101 41 17 20   3 101  3.32 1.090
## 10   Madison Bumgarner 2018  6  7    389 118 47 14 43   5 109  3.26 1.242
## 11   Madison Bumgarner 2019  9  9    623 191 90 30 43  10 203  3.90 1.127
## 12   Madison Bumgarner 2020  1  4    125  47 30 13 13   6  30  6.48 1.440
## 13   Madison Bumgarner 2021  7 10    439 134 76 24 39  11 124  4.67 1.182
## 14   Madison Bumgarner 2022  7 15    476 179 86 25 49   9 112  4.88 1.437
## 15   Madison Bumgarner 2023  0  3     50  25 19  4 15   1  10 10.26 2.400
## 16      Wade     Davis 2009  2  2    109  33 15  2 13   0  36  3.72 1.266
## 17      Wade     Davis 2010 12 10    504 165 76 24 62   5 113  4.07 1.351
## 18      Wade     Davis 2011 11 10    552 190 91 23 63   8 105  4.45 1.375
## 19      Wade     Davis 2012  3  0    211  48 19  5 29   0  87  2.43 1.095
## 20      Wade     Davis 2013  8 11    406 169 80 15 58   4 114  5.32 1.677
## 21      Wade     Davis 2014  9  2    216  38  8  0 23   3 109  1.00 0.847
## 22      Wade     Davis 2015  8  1    202  33  7  3 20   0  78  0.94 0.787
## 23      Wade     Davis 2016  2  1    130  33  9  0 16   3  47  1.87 1.131
## 24      Wade     Davis 2017  4  2    176  39 15  6 28   3  79  2.30 1.142
## 25      Wade     Davis 2018  3  6    196  43 30  8 26   2  78  4.13 1.056
## 26      Wade     Davis 2019  1  6    128  51 41  7 29   2  42  8.65 1.875
## 27      Wade     Davis 2020  0  1     13   9 10  3  3   0   3 20.77 2.771
## 28      Wade     Davis 2021  0  3    128  44 32  8 19   2  38  6.75 1.476
## 29     Jacob    deGrom 2014  9  6    421 117 42  7 43   1 144  2.69 1.140
## 30     Jacob    deGrom 2015 14  8    573 149 54 16 38   2 205  2.54 0.979
## 31     Jacob    deGrom 2016  7  8    444 142 50 15 36   3 143  3.04 1.203
## 32     Jacob    deGrom 2017 15 10    604 180 79 28 59   2 239  3.53 1.187
## 33     Jacob    deGrom 2018 10  9    651 152 41 10 46   5 269  1.70 0.912
## 34     Jacob    deGrom 2019 11  8    612 154 55 19 44   7 255  2.43 0.971
## 35     Jacob    deGrom 2020  4  2    204  47 18  7 18   0 104  2.38 0.956
## 36     Jacob    deGrom 2021  7  2    276  40 11  6 11   1 146  1.08 0.554
## 37     Jacob    deGrom 2022  5  4    193  40 22  9  8   0 102  3.08 0.746
## 38     Jacob    deGrom 2023  2  0     91  19  9  2  4   0  45  2.67 0.758
##    SOper9 SOperBB
## 1   9.000   3.333
## 2   6.973   3.308
```

```
## 3   8.399   4.152
## 4   8.251   3.898
## 5   8.896   3.210
## 6   9.069   5.093
## 7   9.646   6.000
## 8   9.966   4.648
## 9   8.189   5.050
## 10  7.565   2.535
## 11  8.798   4.721
## 12  6.479   2.308
## 13  7.627   3.179
## 14  6.353   2.286
## 15  5.399   0.667
## 16  8.918   2.769
## 17  6.054   1.823
## 18  5.136   1.667
## 19 11.133   3.000
## 20  7.581   1.966
## 21 13.625   4.739
## 22 10.426   3.900
## 23  9.762   2.938
## 24 12.119   2.821
## 25 10.745   3.000
## 26  8.859   1.448
## 27  6.236   1.000
## 28  8.015   2.000
## 29  9.235   3.349
## 30  9.660   5.395
## 31  8.696   3.972
## 32 10.684   4.051
## 33 11.157   5.848
## 34 11.250   5.795
## 35 13.765   5.778
## 36 14.283  13.273
## 37 14.270  12.750
## 38 13.353  11.250
```

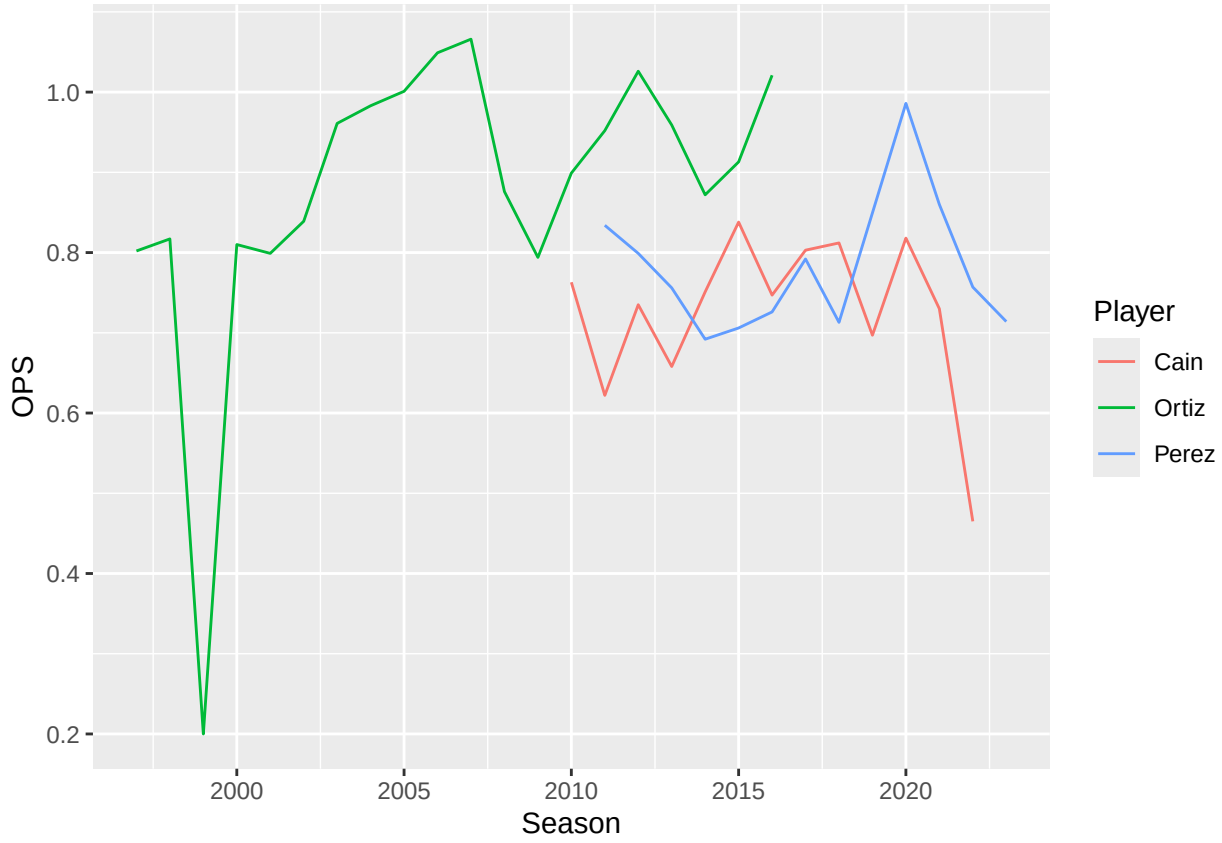
- Create career stat lines for each of the players that you selected. Be careful about how these statistics are calculated.

```
## # A tibble: 3 x 20
##   nameLast     G    AB     R     H   X2B   X3B    HR   RBI    SF    SB    CS
##   <chr>    <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int>
## 1 Cain      1171  4314   626  1220   225    24    87   454    32   190    44
## 2 Ortiz     2408  8640  1419  2472   632    19   541  1768    92    17     9
## 3 Perez     1394  5290   598  1411   257    11   246   812    46     6     1
## # i 8 more variables: BB <int>, HBP <int>, SO <int>, OBP <dbl>, X1B <int>,
## #   SLG <dbl>, OPS <dbl>, SBpct <dbl>
```

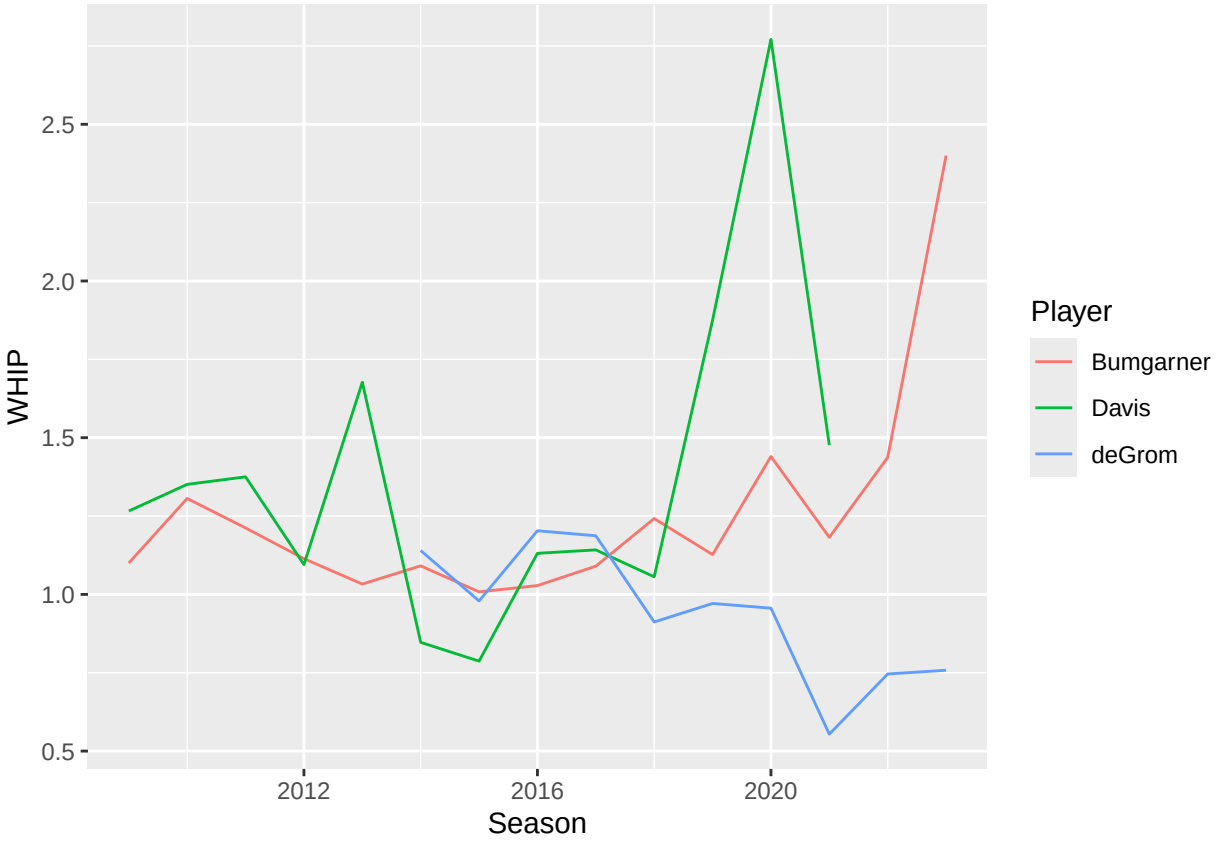
```
## # A tibble: 3 x 11
##   nameLast      W     L    IP     H    ER    HR    BB   HBP    SO   ERA
##   <chr>     <int> <int> <dbl> <int> <int> <int> <int> <int> <int> <dbl>
## 1 Bumgarner   134   124 2209.  2007   853   258   544    89  2070  3.47
```

```
## 2 Davis        63    55  990.   895   433   104   389    32   929  3.94
## 3 deGrom       84    57 1356.  1040   381   119   307    21  1652  2.53
```

- Provide a plot for career trajectories for one batting and one pitching statistic of your choice. These are two separate graphics, one for the batters and one for the pitchers. The graphics that you produce should display the trajectories of the 3 batters and the 3 pitchers. Provide interesting commentary on your graphic.



David Ortiz started his career by far the earliest of these three players, and he had the best OPS of the three for the entire time they overlapped. He is known as a power hitter, and having the highest OPS of the three reflects that. Lorenzo Cain was far from a power hitter, but he did finish third in AL MVP voting in 2015 (where his OPS was the highest of his career). Perez is considered a power hitter, but he has had some issues reaching base in his career which is reflected in OPS. In 2020 he was honored with All-MLB First-Team honors at catcher, and in 2021 he hit 48 home runs. Those two seasons are the peak and the next highest point in his OPS graph, respectively.



Wade Davis had a stretch from 2013-2015 where he was arguably the best relief pitcher the game had ever seen. That is reflected here where his WHIP is the lowest of his career. Madison Bumgarner had a lengthy career and he was known to be a much better pitcher in the playoffs (playoff data is not included in this chart). His WHIP got noticably worse as he reached the tail end of his career. Jacob deGrom is known as one of the best pitchers in baseball when he is healthy. His WHIP from 2018 onward does not go above 1.00 which is incredibly impressive.

Question 3 Exercise 1.2 in the online version of Analyzing Baseball Data with R. Exercise 2 on page 27 of the textbook.

```
##   GS CG CompletedRatio
## 1 34 28      0.8235294
```

```
##    SO BB    KBB
## 1 268 62 4.3226
```

```
##       IP
## 1 304.67
```

```
##    WHIP
## 1 0.853
```

a. 28/34 games were completed by Gibson (0.8235).
b. Gibson's strikeout to walk ratio was 268/62 or 4.3226.
c. He pitched 304 and 2/3 innings.
d. Gibson had a WHIP of 0.853.

Question 4 Exercise 1.3 in the online version of Analyzing Baseball Data with R. Exercise 3 on page 27 of the textbook.

a. The game was 2 hours and 19 minutes long.
b. Because it is the second game of a doubleheader and the attendance was logged in the first game.
c. They had 3 extra-base hits.
d. It was 0.375 (12/32).