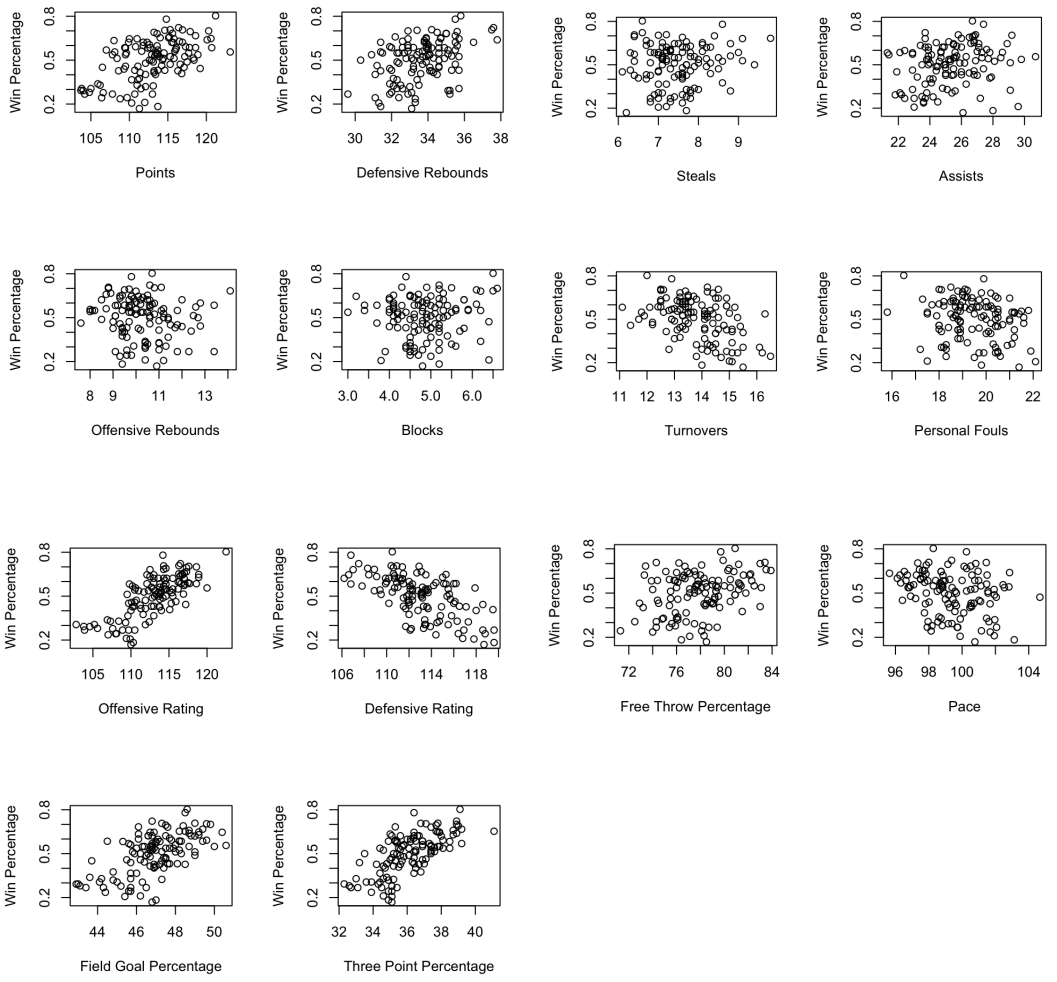
Bayes Project Daily Report

Day 1 - Looking at Data and First Regression Model

- Looked through data set and determined all variables that I perceived to be factors that I wanted to test for significance
- Plotted all predictor variables against response variable (win%) separately to get an idea of general trends and what to expect

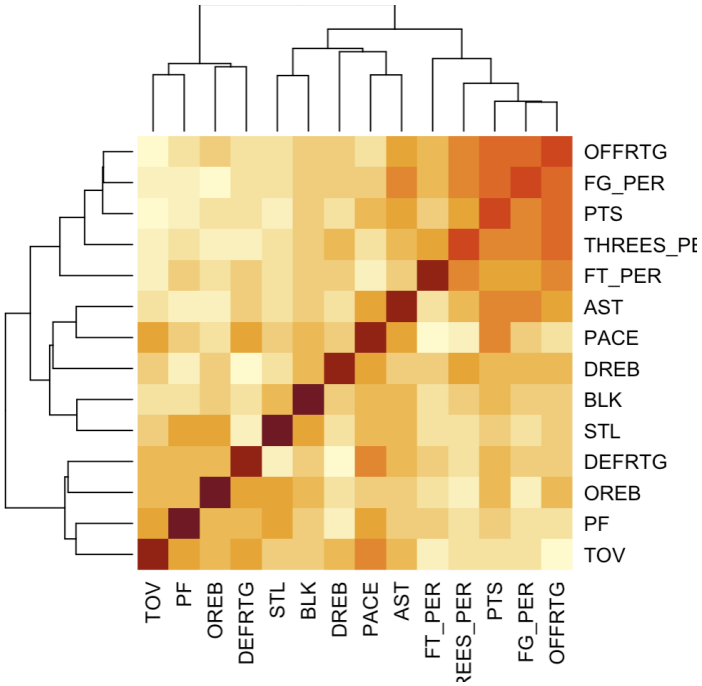


<u>We can analyze the general trends from these plots:</u>

- Clear Pos/Neg Trend:
  - Points (P)
  - Offensive Rating (P)
  - Defensive Rating (N)
  - Field Goal Percentage (P)

    - Three Point Percentage (P)
- Somewhat Pos/Neg Trend:
    - Defensive Rebounds (P)
    - Steals (P)
    - Assists (P)
    - Turnovers (N)
    - Pace (N)
- Unclear Trend:
    - Offensive Rebounds
    - Blocks
    - Personal Fouls
    - Free Throw Percentage

- After this, I created a correlation matrix and heat map to get an idea beforehand of how bad certain correlations could be (something like points and offensive rating with undoubtedly be very correlated)



The heatmap shows some concerning correlations between:

OFFRTG and FG_PER
OFFRTG and PTS
OFFRTG and THREES_PER
FG_PER and PTS

<u>Priors:</u>

- Based off my general knowledge of basketball statistics, my priors for the beta coefficients is:

PTS ~ N(2, 5)	DREB ~ N(1,5)	OREB ~ N(1,5)
BLK ~ N(1,5)	STL ~ N(1,5)	AST ~ N(1,5)
TOV ~ N(-1,5)	PF ~ N(-1,5)	OFFRTG ~ N(2,5)
DEFRTG ~ N(2,5)	FG_PER ~ N(2,5)	PACE ~ N(1,5)
FT_PER ~ N(1,5)	THREES_PER ~ N(2,5)	

- My prior on the intercept is N(0,15) to be non-informative

Regression:

- Summary with a 95% and 80% confidence intervals

```
> tidy(nba_model, conf.int = TRUE, conf.level = 0.95)
# A tibble: 15 × 5
   term          estimate std.error conf.low conf.high
   <chr>            <dbl>     <dbl>    <dbl>     <dbl>
 1 (Intercept)  0.335       1.18    -1.98      2.64
 2 PTS         -0.0000716   0.00992 -0.0196    0.0197
 3 DREB         0.00540     0.00528 -0.00519   0.0160
 4 OREB        -0.00305     0.00538 -0.0137    0.00747
 5 BLK         -0.00998     0.00575 -0.0214    0.00155
 6 STL          0.0101      0.00852 -0.00673   0.0268
 7 AST         -0.00155     0.00284 -0.00711   0.00404
 8 TOV         -0.00344     0.00527 -0.0138    0.00726
 9 PF          -0.00418     0.00405 -0.0121    0.00374
10 OFFRTG       0.0250      0.0102   0.00451   0.0451
11 DEFRTG      -0.0245      0.00289 -0.0303   -0.0189
12 FG_PER       0.00718     0.00617 -0.00497   0.0194
13 THREES_PER   0.00444     0.00425 -0.00397   0.0126
14 FT_PER      -0.00106     0.00175 -0.00451   0.00243
15 PACE        -0.00283     0.0115  -0.0258    0.0200
```

```
> tidy(nba_model, conf.int = TRUE, conf.level = 0.8)
# A tibble: 15 × 5
   term          estimate std.error  conf.low conf.high
   <chr>            <dbl>     <dbl>     <dbl>     <dbl>
 1 (Intercept)  0.335       1.18    -1.19       1.85
 2 PTS         -0.0000716   0.00992 -0.0128     0.0128
 3 DREB         0.00540     0.00528 -0.00164    0.0123
 4 OREB        -0.00305     0.00538 -0.00998    0.00384
 5 BLK         -0.00998     0.00575 -0.0173    -0.00247
 6 STL          0.0101      0.00852 -0.000821   0.0210
 7 AST         -0.00155     0.00284 -0.00518    0.00213
 8 TOV         -0.00344     0.00527 -0.0103     0.00338
 9 PF          -0.00418     0.00405 -0.00927    0.00105
10 OFFRTG       0.0250      0.0102   0.0116     0.0380
11 DEFRTG      -0.0245      0.00289 -0.0283    -0.0209
12 FG_PER       0.00718     0.00617 -0.000736   0.0151
13 THREES_PER   0.00444     0.00425 -0.00104    0.00975
14 FT_PER      -0.00106     0.00175 -0.00330    0.00121
15 PACE        -0.00283     0.0115  -0.0179     0.0121
```

- These are not good results however that could be expected with the amount of variables in the model
- The only significant variables are:
  - Offensive Rating
  - Defensive Rating
  - Blocks (only in 80% interval)

- Offensive Rating and Defensive Rating seem to be the most significant variables in the whole data set according to our regression and plots

- Blocks seems odd here since our plot before regression showed blocks had a very unclear relationship with win percentage

Day 2 - Selecting Variables

ELPD Method:

- There are too many variables to test all possible models (especially if interaction is ever included)
- First approach is to use ELPD to test:
    - Model 1 - all variables
    - Model 2 - all variables without OREB, BLK, PF, FT_PER
        - This is made from getting rid of all variables that had an unclear trend in the original plots
    - Model 3 - all variables without OREB, PF, FT_PER
        - Blocks appeared significant in our 80% confidence interval therefore we should test if should be kept or not

Model 1

	Estimate	SE
elpd_loo	201.11747	7.756636
p_loo	15.34136	1.935285
looic	-402.23494	15.513272

Model 2

	Estimate	SE
elpd_loo	203.43321	8.064443
p_loo	11.33626	1.566554
looic	-406.86641	16.128886

Model 3

	Estimate	SE
elpd_loo	203.96689	7.945430
p_loo	12.19391	1.609104
looic	-407.93378	15.890860

- As we can see, the ELPD value of Model 3 is the highest overall value
- Therefore, we can safely assume that we can eliminate OREB, PF, and FT_PER based on the plots, initial regression, and ELPD values

- Despite blocks appearing insignificant in the plot and initial regression's 95% confidence interval, the 80% confidence interval of the initial regression along with the ELPD value of model 3 makes blocks appear to be a significant variable worth keeping

Coefficients in Model 3:

```
> tidy(nba_model.3, conf.int = TRUE, conf.level = 0.95)
# A tibble: 12 × 5
   term        estimate std.error conf.low conf.high
   <chr>          <dbl>     <dbl>    <dbl>     <dbl>
 1 (Intercept) -0.0580    1.14    -2.29      2.18
 2 PTS         -0.00203   0.00984 -0.0213    0.0170
 3 DREB         0.00761   0.00459 -0.00144   0.0166
 4 BLK         -0.00955   0.00567 -0.0207    0.00162
 5 STL          0.00961   0.00791 -0.00605   0.0252
 6 AST         -0.00120   0.00276 -0.00661   0.00416
 7 TOV         -0.00564   0.00453 -0.0148    0.00333
 8 OFFRTG       0.0249    0.0102   0.00513   0.0449
 9 DEFRTG      -0.0239    0.00264 -0.0292   -0.0187
10 FG_PER       0.0106    0.00464  0.00143   0.0197
11 THREES_PER   0.00476   0.00361 -0.00238   0.0119
12 PACE        -0.00174   0.0114  -0.0239    0.0206
> tidy(nba_model.3, conf.int = TRUE, conf.level = 0.8)
# A tibble: 12 × 5
   term        estimate std.error   conf.low conf.high
   <chr>          <dbl>     <dbl>      <dbl>     <dbl>
 1 (Intercept) -0.0580    1.14    -1.51       1.39
 2 PTS         -0.00203   0.00984 -0.0147     0.0105
 3 DREB         0.00761   0.00459  0.00175    0.0135
 4 BLK         -0.00955   0.00567 -0.0168    -0.00231
 5 STL          0.00961   0.00791 -0.000676   0.0196
 6 AST         -0.00120   0.00276 -0.00476    0.00230
 7 TOV         -0.00564   0.00453 -0.0115     0.000170
 8 OFFRTG       0.0249    0.0102   0.0120     0.0381
 9 DEFRTG      -0.0239    0.00264 -0.0273    -0.0205
10 FG_PER       0.0106    0.00464  0.00463    0.0166
11 THREES_PER   0.00476   0.00361  0.0000976  0.00935
12 PACE        -0.00174   0.0114  -0.0164     0.0127
```

We have more significant variables in Model 3:

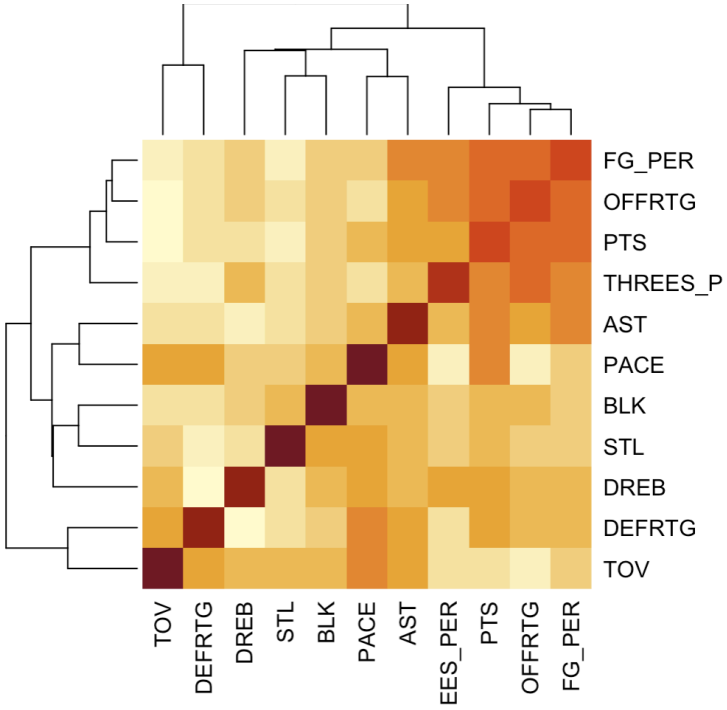
95%:
- OFFRTG
- DEFRTG
- FG_PER

80%:
- DREB
- BLK
- OFFRTG
- DEFRTG
- FG_PER
- THREES_PER

Day 3 - Model Selection

- We have successfully eliminated 3 variables from our regression model and confidently believe that the regression model has benefitted from it
- However, we still have insignificant variables
- I believe the next approach should be to address the correlation issues. If two variables are very correlated, we should consider one over the other

Correlation:



Concerning Correlations:

PTS and OFFRTG
PTS and FG_PER
OFFRTG and FG_PER
OFFRTG and THREES_PER

Test Models Based on Correlations:

Model 1 - no points
Model 2 - no fg%
Model 3 - no points and fg%
Model 4 - no offensive rating
Model 5 - no 3p%

Current Model

	Estimate	SE
elpd_loo	203.96689	7.945430
p_loo	12.19391	1.609104
looic	-407.93378	15.890860

Model 1

	Estimate	SE
elpd_loo	204.91172	7.977884
p_loo	11.43071	1.566116
looic	-409.82344	15.955769

Model 2

	Estimate	SE
elpd_loo	202.10308	7.633925
p_loo	11.23561	1.482854
looic	-404.20616	15.267849

Model 3

	Estimate	SE
elpd_loo	203.18060	7.695557
p_loo	10.44711	1.434629
looic	-406.36119	15.391115

Model 4

	Estimate	SE
elpd_loo	201.41593	7.922206
p_loo	11.52665	1.534147
looic	-402.83186	15.844413

Model 5

	Estimate	SE
elpd_loo	204.08597	7.879051
p_loo	11.33657	1.535737
looic	-408.17194	15.758103

- All these ELPD values are relatively close, however here are the rankings:

1. Model 1 - no points
2. Model 5 - no 3p%
3. Current Model
4. Model 3 - no points and fg%
5. Model 2 - no fg%
6. Model 4 - no offensive rating

- It feels odd to consider dropping points as a predictor variable, but its correlation with offensive rating is concerning (obviously they are very similar). Offensive Rating seems to definitely be a much better predictor for win percentage than points.

Coefficients for New Model:

```
> tidy(nba_model.no_pts, conf.int = TRUE, conf.level = 0.95)
# A tibble: 11 × 5
   term        estimate std.error conf.low conf.high
   <chr>          <dbl>     <dbl>    <dbl>     <dbl>
 1 (Intercept)  0.169     0.380   -0.580     0.914
 2 DREB         0.00768   0.00455 -0.00134   0.0166
 3 BLK         -0.00953   0.00567 -0.0208    0.00164
 4 STL          0.00967   0.00795 -0.00587   0.0254
 5 AST         -0.00117   0.00276 -0.00658   0.00435
 6 TOV         -0.00578   0.00450 -0.0146    0.00300
 7 OFFRTG       0.0229    0.00219  0.0186    0.0272
 8 DEFRTG      -0.0239    0.00264 -0.0291   -0.0187
 9 FG_PER       0.0105    0.00457  0.00146   0.0196
10 THREES_PER   0.00495   0.00370 -0.00216   0.0120
11 PACE        -0.00408   0.00345 -0.0107    0.00257
> tidy(nba_model.no_pts, conf.int = TRUE, conf.level = 0.8)
# A tibble: 11 × 5
   term        estimate std.error  conf.low  conf.high
   <chr>          <dbl>     <dbl>     <dbl>      <dbl>
 1 (Intercept)  0.169     0.380   -0.319     0.653
 2 DREB         0.00768   0.00455  0.00176   0.0135
 3 BLK         -0.00953   0.00567 -0.0168   -0.00228
 4 STL          0.00967   0.00795 -0.000558  0.0198
 5 AST         -0.00117   0.00276 -0.00470   0.00237
 6 TOV         -0.00578   0.00450 -0.0114    0.0000183
 7 OFFRTG       0.0229    0.00219  0.0201    0.0257
 8 DEFRTG      -0.0239    0.00264 -0.0272   -0.0205
 9 FG_PER       0.0105    0.00457  0.00463   0.0164
10 THREES_PER   0.00495   0.00370  0.000196  0.00961
11 PACE        -0.00408   0.00345 -0.00845   0.000350
```

Significant Predictors:

95%:

- OFFRTG
- DEFRTG
- FG_PER

80%:

- DREB
- BLK
- OFFRTG
- DEFRTG
- FG_PER
- THREES_PER

- The same significant variables as the last model so was this worth it? Probably not especially since the coefficient values didn't even change that much if at all

Test Models of Significant Variables Only:

- Model 1 - significant variables at 95%
- Model 2 - significant variables at 80%

Model 1

	Estimate	SE
elpd_loo	206.515327	8.0765847
p_loo	5.267298	0.8746655
looic	-413.030655	16.1531695

Model 2

	Estimate	SE
elpd_loo	206.568463	7.562575
p_loo	7.851342	1.117860
looic	-413.136927	15.125150

- Model 2 is better here meaning it is better to include all significant variables at the 80% confidence level as opposed to only the significant variables at the 95% confidence level

Day 4 - Final Model Selection

- For the fun of it, let's use the gibbs sampler on the 10 remaining variables

Current Model of 10 Variables

												z.probs
[1,]	0	0	0	0	0	0	1	1	0	0	0	0.1997
[2,]	0	1	0	0	0	0	1	1	0	0	0	0.0883
[3,]	1	0	0	0	0	0	1	1	0	0	0	0.0797
[4,]	0	0	0	0	0	0	1	1	1	0	0	0.0578
[5,]	0	0	0	0	0	0	1	1	0	1	0	0.0524
[6,]	0	0	0	0	0	0	1	1	0	0	1	0.0286
[7,]	0	0	1	0	0	0	1	1	0	0	0	0.0285
[8,]	0	0	0	1	0	0	1	1	0	0	0	0.0202
[9,]	0	0	0	0	1	0	1	1	0	0	0	0.0177
[10,]	0	0	0	0	0	1	1	1	0	0	0	0.0175

Model of All Variables

															z.probs
[1,]	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0.1127
[2,]	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0.0545
[3,]	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0.0527
[4,]	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0.0392
[5,]	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0.0307
[6,]	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0.0250
[7,]	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0.0172
[8,]	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0.0161
[9,]	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0.0134
[10,]	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0.0128

- This gibbs sampler doesn't tell us a lot
- The intercept is rarely included which wouldn't be possible
- OFFRTG and DEFRTG are definitely significant variables and other variables like BLK, DREB, FG% and 3P% appear sometimes too

Coefficients of 3 and 6 Variable Models

3 Variables

```
# A tibble: 4 × 5
  term        estimate std.error  conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)  0.459     0.194    0.0790      0.845
2 OFFRTG       0.0258    0.00169  0.0225      0.0291
3 DEFRTG      -0.0285    0.00129 -0.0311     -0.0259
4 FG_PER       0.00741   0.00384 -0.000146    0.0148
```

6 Variables

```
# A tibble: 7 × 5
  term        estimate std.error conf.low conf.high
  <chr>          <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept)  0.287     0.273   -0.265     0.836
2 DREB         0.00239   0.00311 -0.00376   0.00855
3 BLK         -0.0104    0.00542 -0.0212    0.000449
4 OFFRTG       0.0251    0.00190  0.0213    0.0288
5 DEFRTG      -0.0274    0.00160 -0.0306   -0.0243
6 FG_PER       0.00629   0.00389 -0.00133   0.0139
7 THREES_PER   0.00424   0.00354 -0.00291   0.0113
```

- Coefficients are very similar for these two models as well as with the current model (10 variables)

Final Decision with ELPD

All 10 Variables

	Estimate	SE
elpd_loo	204.91172	7.977884
p_loo	11.43071	1.566116
looic	-409.82344	15.955769

6 Variables

	Estimate	SE
elpd_loo	206.568463	7.562575
p_loo	7.851342	1.117860
looic	-413.136927	15.125150

3 Variables

	Estimate	SE
elpd_loo	206.515327	8.0765847
p_loo	5.267298	0.8746655
looic	-413.030655	16.1531695

- Of these 3 models, the 3 and 6 variable models are the best and have a very close ELPD
- 6 variables is higher so I believe we should go with this

Day 5 - Model Interpretation

Coefficients

A tibble: 7 × 5

	term	estimate	std.error	conf.low	conf.high
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	0.287	0.273	-0.265	0.836
2	DREB	0.00239	0.00311	-0.00376	0.00855
3	BLK	-0.0104	0.00542	-0.0212	0.000449
4	OFFRTG	0.0251	0.00190	0.0213	0.0288
5	DEFRTG	-0.0274	0.00160	-0.0306	-0.0243
6	FG_PER	0.00629	0.00389	-0.00133	0.0139
7	THREES_PER	0.00424	0.00354	-0.00291	0.0113

Regression Equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

$$Y = 0.287 + 0.002X_1 - 0.01X_2 + 0.025X_3 - 0.027X_4 + 0.006X_5 + 0.004X_6$$

Prediction

23-24 Houston Rockets

- What is the predicted win percentage of the 23-24 Houston Rockets
  - 34 defensive rebounds
  - 6 blocks
  - 113.7 offensive rating

    - 112.8 defensive rating
    - 45.9 fg%
    - 35.2 3P%

- According to our equation, the Rockets should have had a winning percentage of **0.5521** and won **42.81** games. The Rockets won **41** games with a winning percentage of **0.5** this season making this prediction very close

18-19 Toronto Raptors

- What about a team from before the data?
- The 18-19 Raptors averaged over the season
    - 35.6 defensive rebounds
    - 5.3 blocks
    - 112.6 offensive rating
    - 106.8 defensive rating
    - 47.4 fg%
    - 36.6 3P%

- The equation predicts the Raptors to have a win percentage of **0.667** and win **54.73** games. That season they won **58** games and had a winning percentage of **0.707**
- This prediction was surprisingly close despite the shift the NBA has made in recent years towards more frequent scoring (aka our data shouldn't be the best predictor for years before but it did well here)

97-98 Chicago Bulls

- What about a team that is years far before our sample
- The 18-19 Raptors averaged over the season
    - 29.7 defensive rebounds
    - 4.3 blocks
    - 106.2 offensive rating
    - 98.4 defensive rating
    - 45.1 fg%
    - 32.3 3P%

- The equation predicts that the bulls should have had a winning percentage of **0.701** and won **57.51** games. This is also somewhat close since the Bulls won **62** games with a winning percentage of **0.756** that season