Basketball Analytics: Predicting eFG% for the Next 3 Games

Abstract

For me, the goal of this project is to use any of the information available as of 1/4/15, to predict Bradley Beal's shooting efficiencies for the next 3 games. Namely, the goal is to predict his eFG%.

Going forward, it would be fun to do a deeper dive into predicting his 3-Pt FG%, 2-Pt FG% and LU Pct% to see which of the break-downs he over-performed or under-performed each game.

Data Preparation

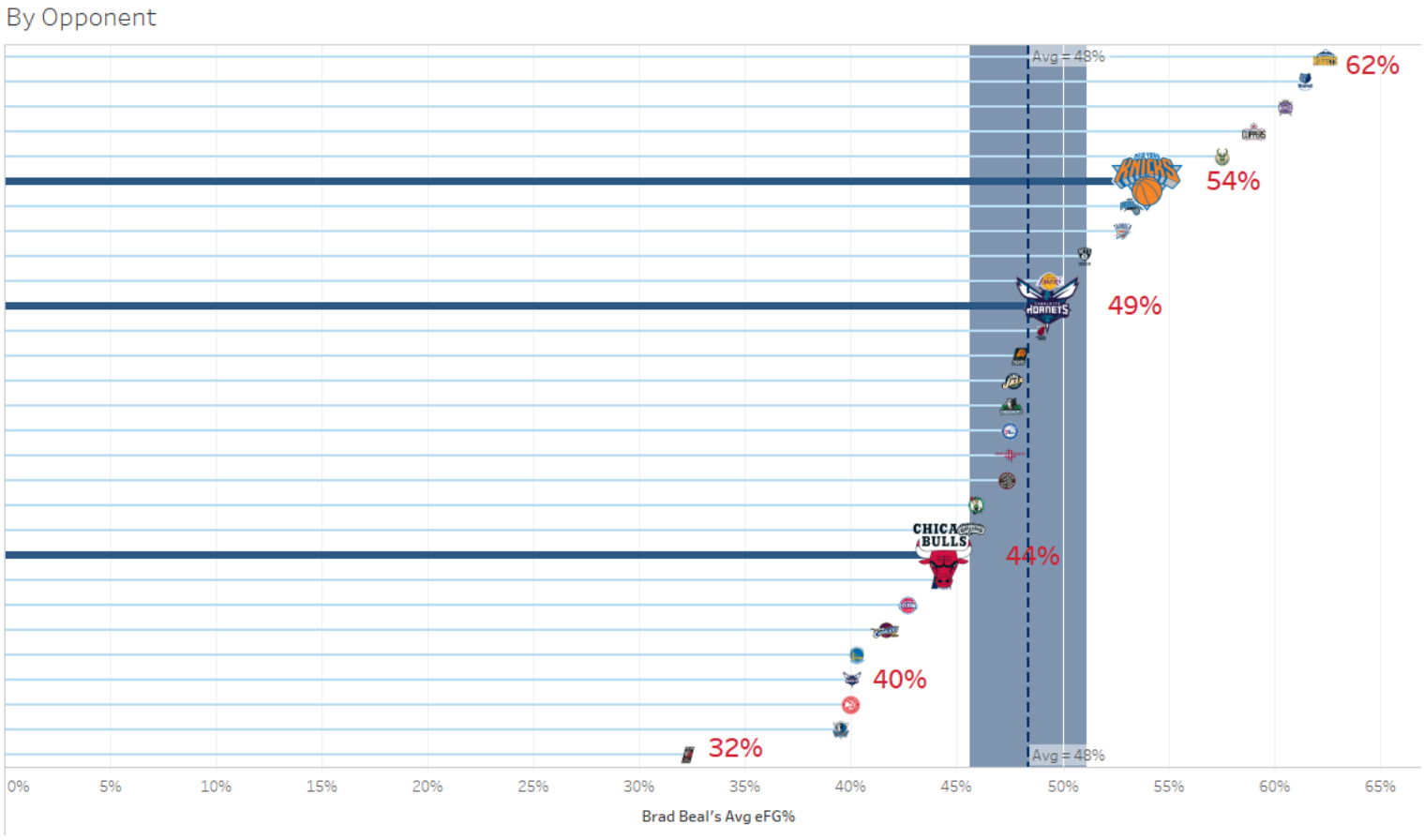
Along with adding the "Advanced" team statistics (basketball-reference.com) and the home and away column, the most time-consuming part of the data preparation was adding recency metrics. First, I calculated the exact eFG% for the last 1 game, 2 games, etc., through the last 10 games. This was done by adding up all the made field goals, 3-pt field goals, and field goal attempts (using the eFG% formula for multiple games combined). The reason I make this distinction, is that the other metrics were calculated a bit differently (and not weighted for the number of shots that were taken in each game). As I wanted a quick (but maybe somewhat "dirty") way to find these recency measures, I built a function in R that calculates what I called L1, L2, …., L10 for 10 different columns of "recent" measures. In all, 40 metrics (including team and individual metrics) were used to create these recent measure columns. The final dataset that was used for evaluation had 458 variables, with 408 variables being variables we could use as predictors in modeling. Therefore, the first thing that had to be done was to calculate the correlations for each of these metrics, so that the dataset could be cut down dramatically before modeling. As we only have 153 games, the goal was to at least cut the number down below 153. As a note, it seems that last 5 games has been widely used as a measure of recent play. But I wanted to try and take a completely scientific approach when possible. Though I thought it should be mentioned before proceeding, so that we don't get bogged down into keeping measures that provide essentially the same information as other variables.

For missing values with the recency metrics, the average of that value for that Season was used. For example, for the Last 10 Games for eFG%, the first 10 games used the average of the final games for the Regular Season. This is called mean imputation and is used regularly in Missing Value analyses. However, it should be remembered in evaluation, as this certainly may have not been the best way to handle these missing values. For example, Usage Rates and Minutes Played likely could have been handled in a better way. But the logic, even for these metrics, is that I did not want these missing values to affect the variable in using them for predictions, but I wanted to use the variables if they had any information for us to help the models.

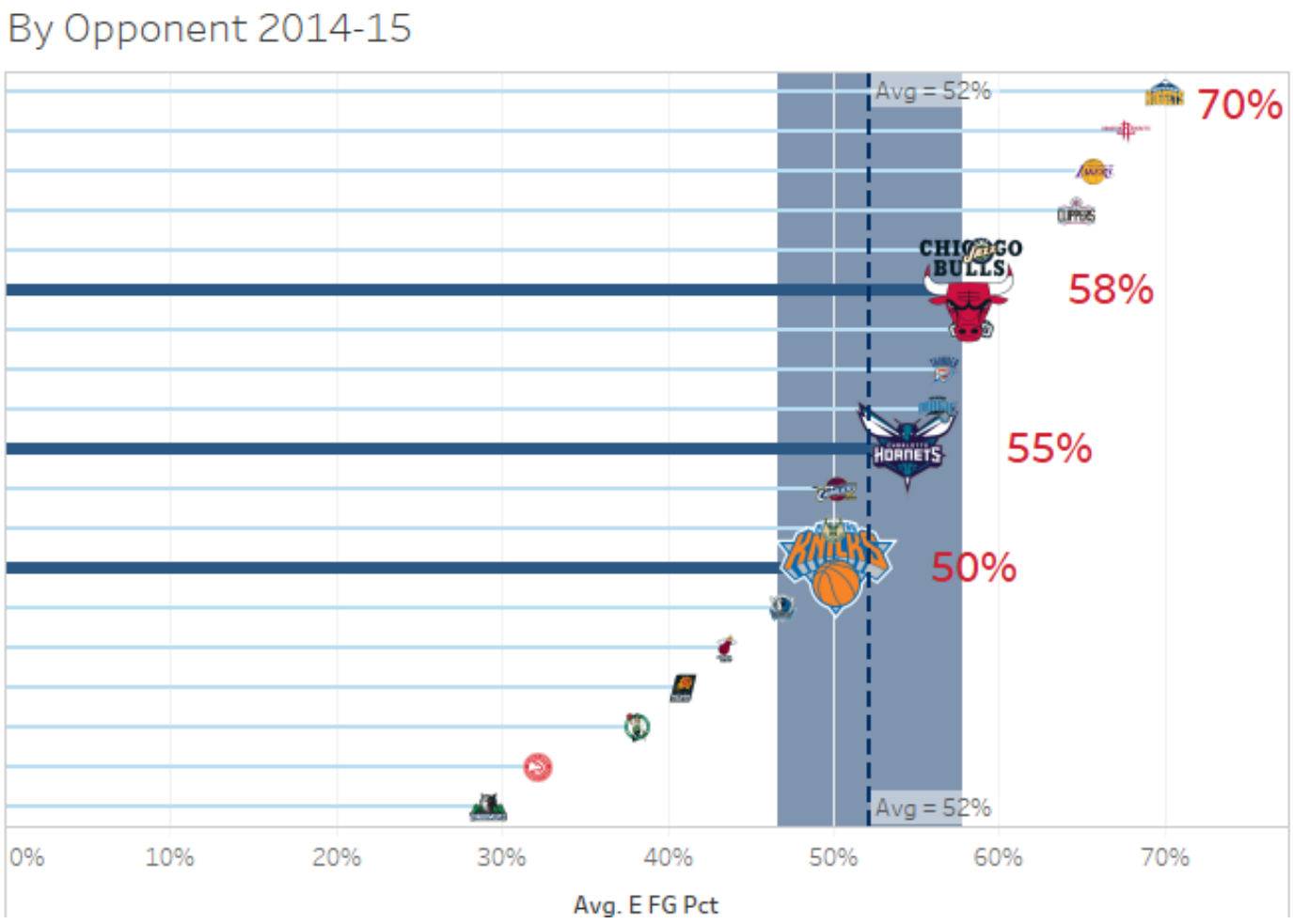
Data Exploration

Keep these in mind as we start the modeling process

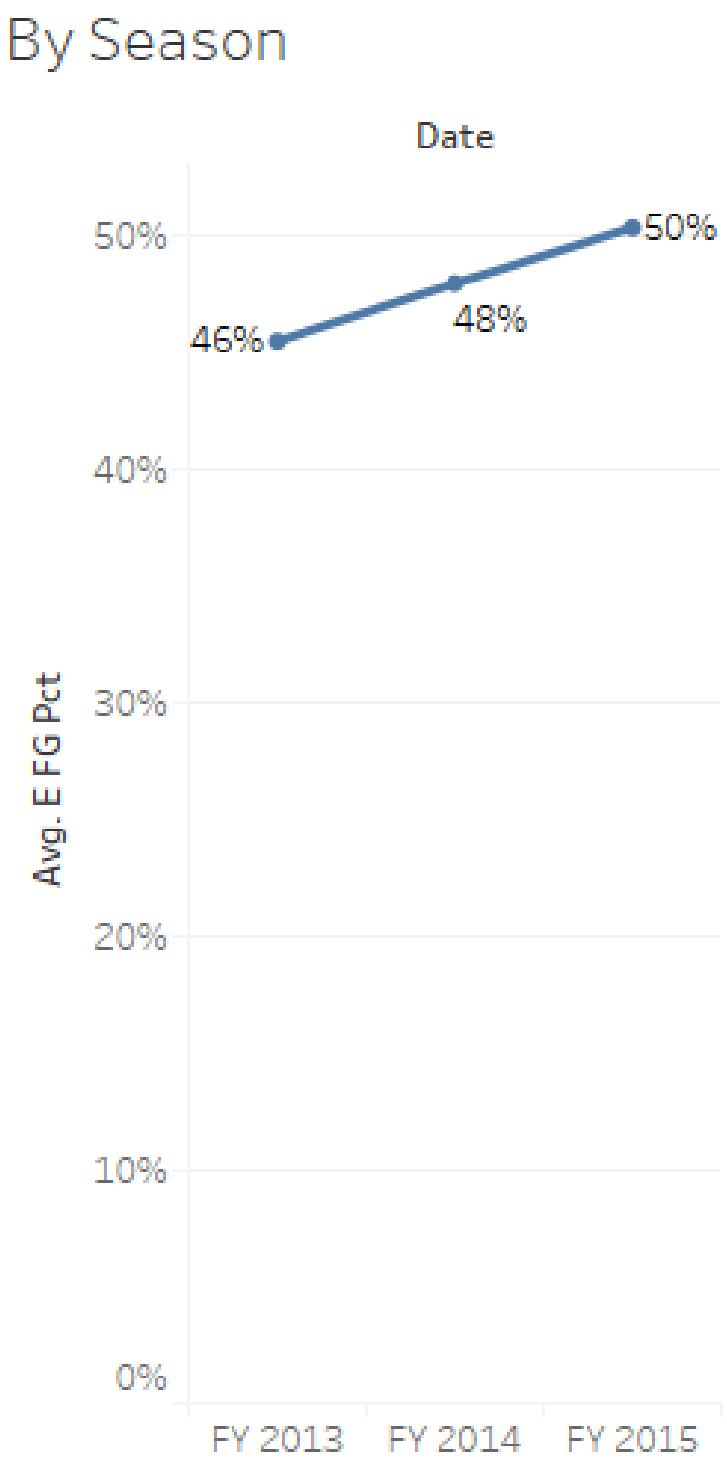
In Brad Beal's career, he is shooting **above the 95 Confidence Interval for his average against the Knicks**. He is **within the 95% Confidence Interval of average against the Hornets** and **below the 95% CI against the Bulls**.



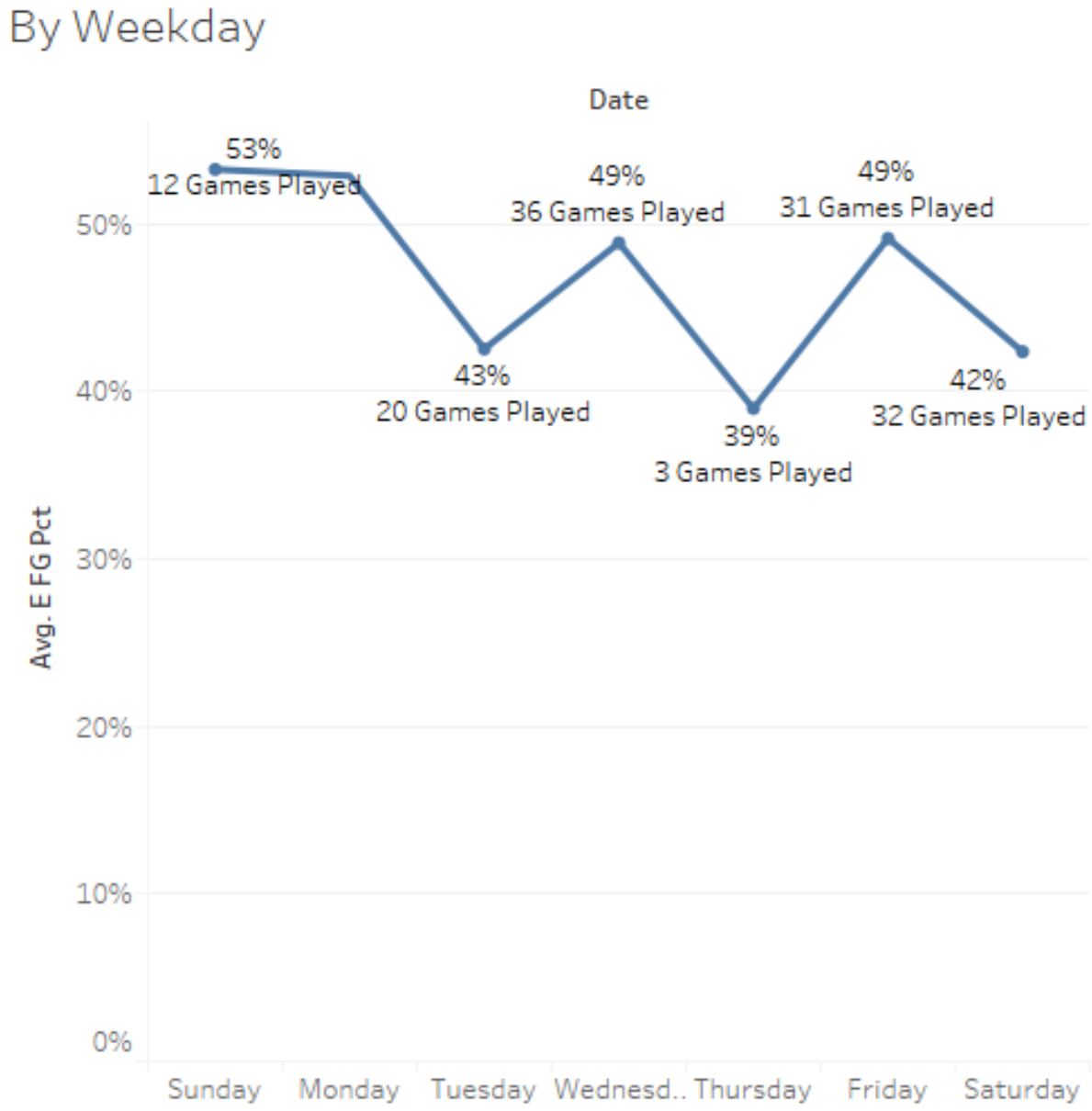
However, for 2014-15 Brad is **slightly above his season average against the Bulls**, **a little above average against the Hornets** (but within the 95% Confidence Interval), and a **little below average against the Knicks** (but also within the 95% Confidence Interval):



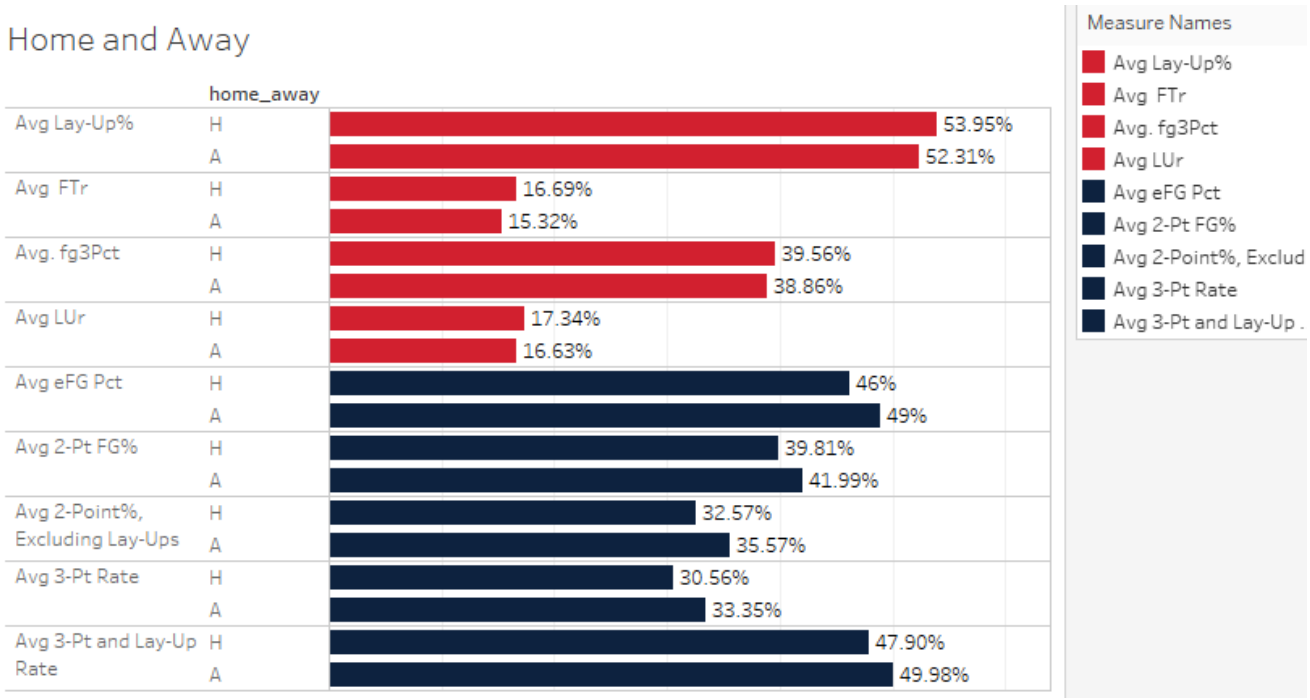
Beal's avg eFG% has gone up each season:



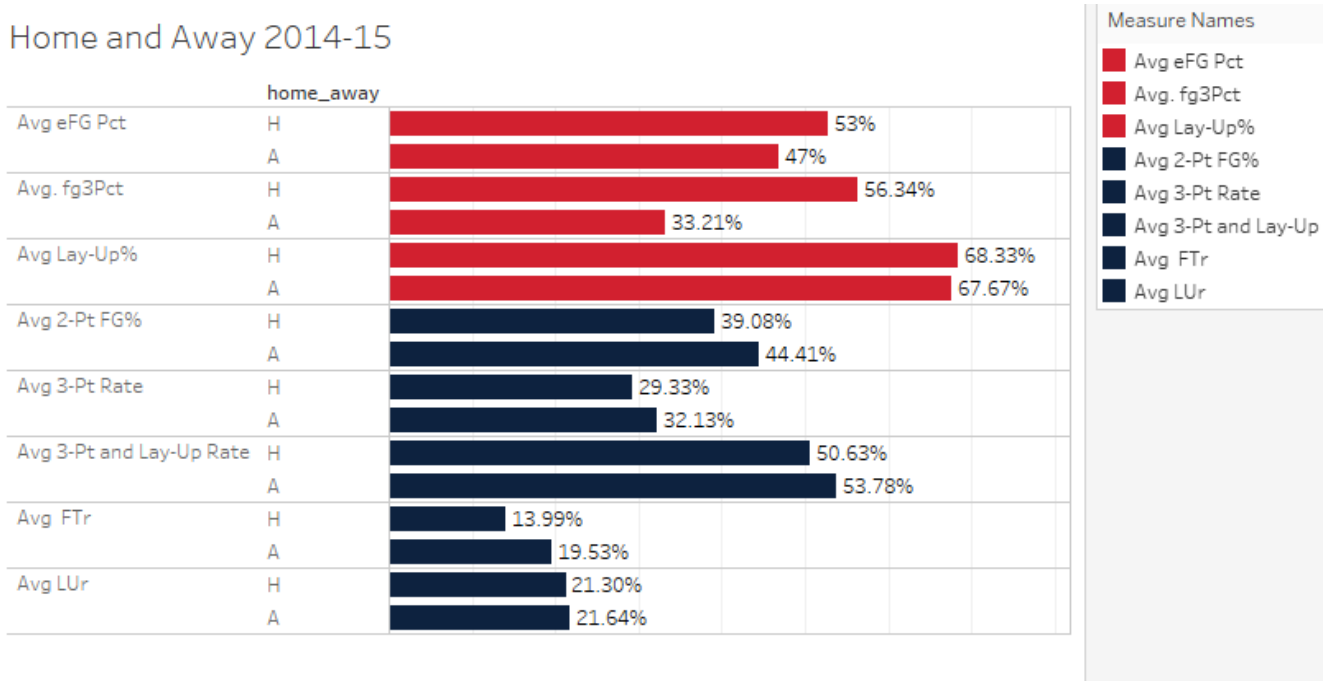
And he shoots slightly worse throughout the week...



Here are his averages at home and away:



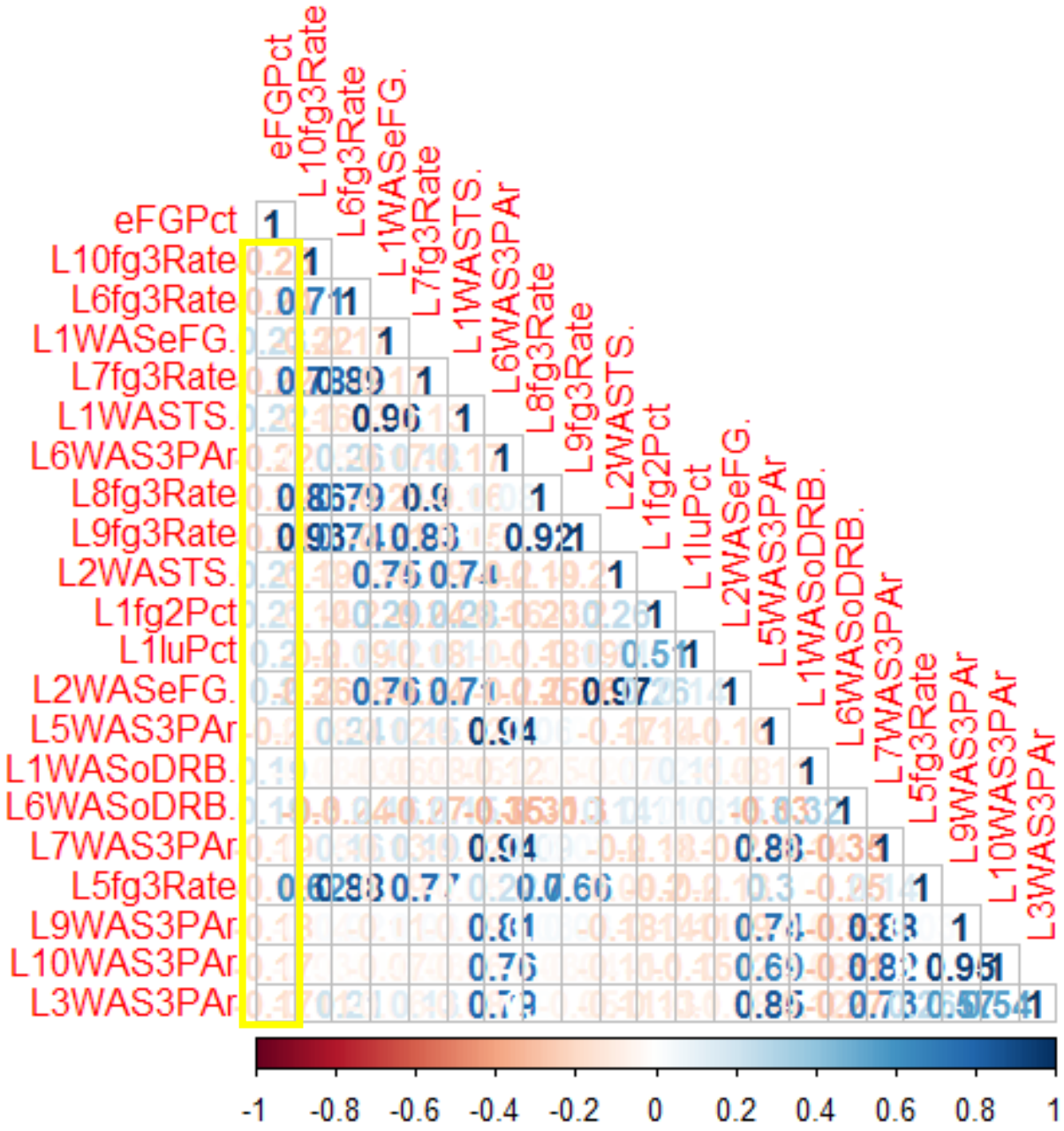
And also for 2014-15:



Higher at Home; Higher on the Road

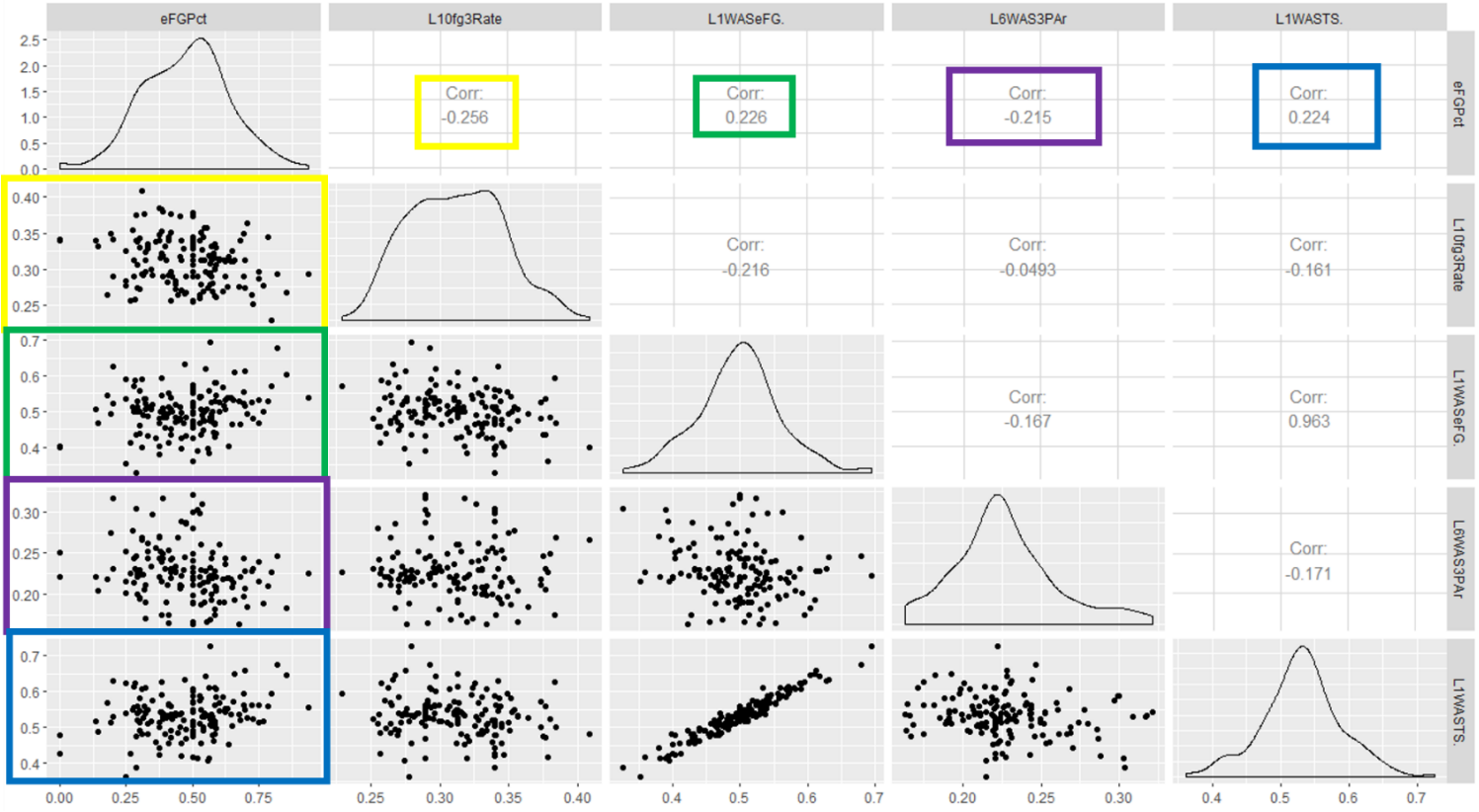
All Correlations whose Absolute Value is greater than 15% to Brad's eFG% Next Game

We will get to what all of these variables mean, but they are individual and team stats. Any L1, L2,…,L10 metrics represent these same stats in the Last 1, Last 2,…, Last 10 games

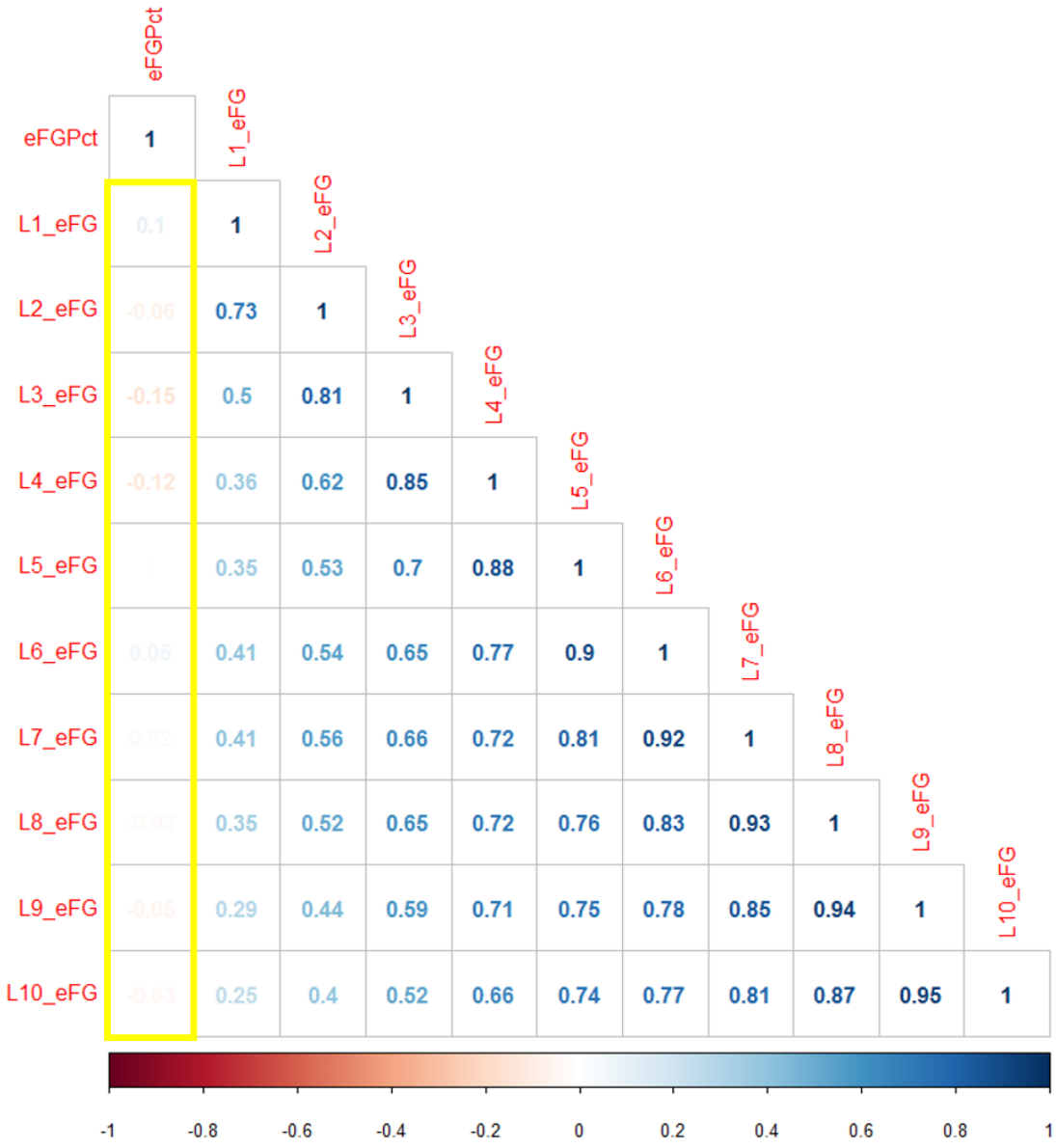


Here is my favorite exploration visualization. We can see the scatterplot, along with the correlations. Here is a closer look at 4 of our highest correlated, recent game metrics to Brad Beal's next game eFG%:

1. A higher 3-Pt Attempt Rate for Brad in the last 10 games (being higher) indicates a lower eFG% next game (makes sense, since he has tended to shoot a higher amount of 3's lately)
    a. Note the approximate bell curve for eFG%, since we assume normal distributions
2. Washington's eFG% last game (being higher) indicates a tendency for Brad to have a higher eFG% next game
3. Washington's 3-Pt Rate in their last 6 games (being higher) indicates a lower eFG% for Brad next game (High correlation to number 1, but could give us a tad different information too)
4. Washington's Assist% last game (being higher) indicates a higher eFG% for Brad next game



Correlations for Brad Beal's eFG% in the Last 10 Games against his current Game: 2012-13 thru 2014-15

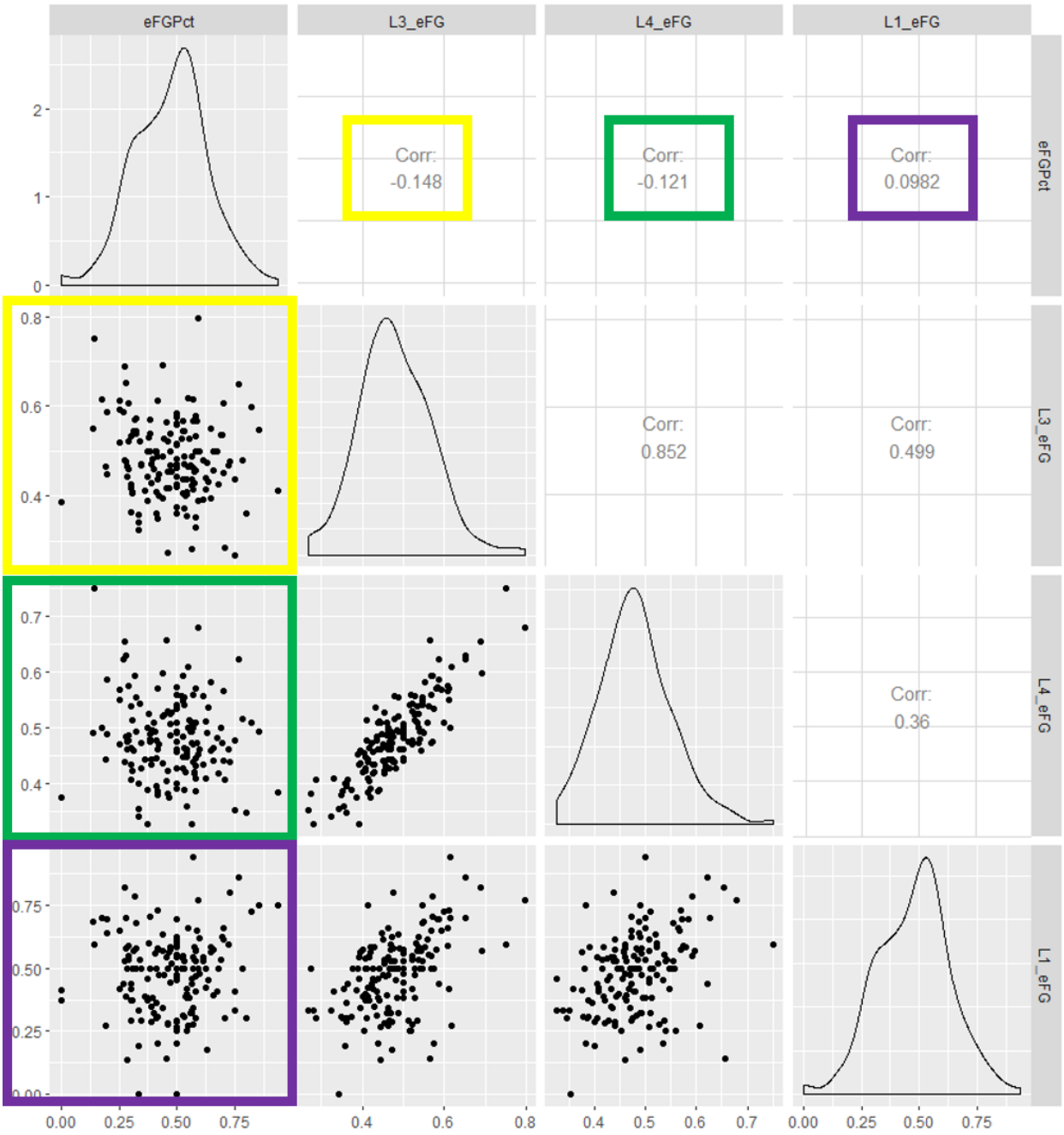


*L1_eFG is Brad's eFG% in the last game, L2_eFG is his eFG% in his last 2 games, and so on

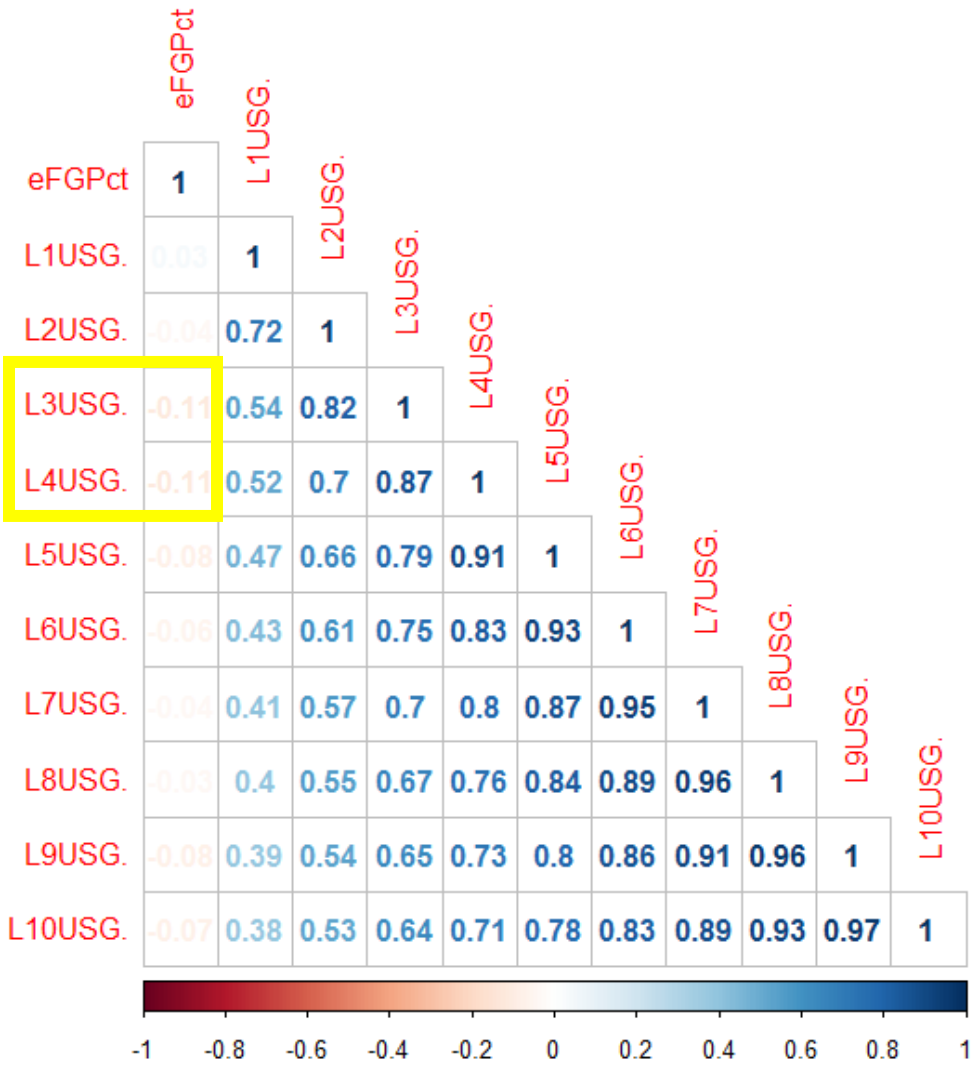
- There is a 10% correlation between his last game and his current game
- There is a -15% and -12% correlation to his eFG% from his last 3 and 4 games, respectively, and his current game
    - This suggests the "law of averages" that many shooters use to justify their next shot after missing multiple in row is somewhat correct!

These aren't strong, but it is more information we can have in making a prediction, so this could account for some of the variability that the higher correlated metrics could not account for

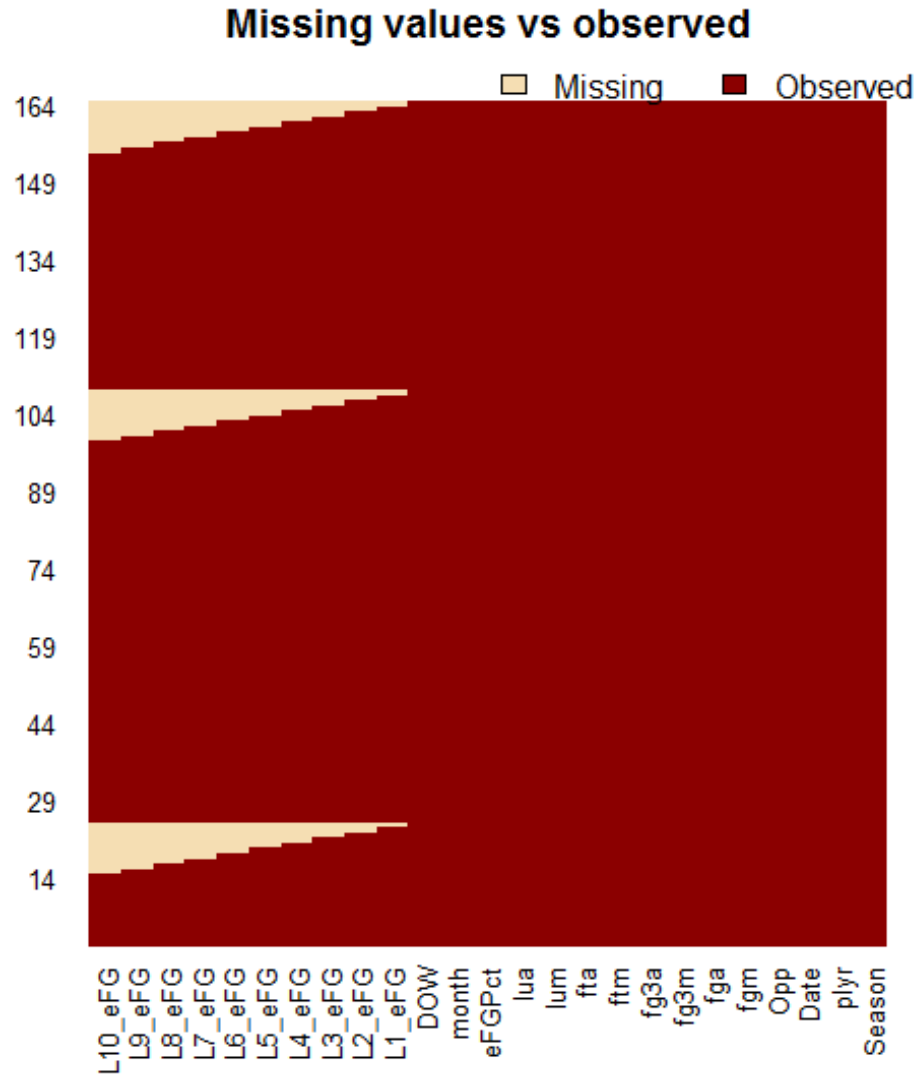
We can see these slight relationships and confirm a near normal distribution in the data (though slightly skewed):



Though not used in the final models, we see that his Avg Usage the last 3 and 4 games is negatively related to his eFG% next game at 11%:



We know we do not have the 1st game of the season for L1_eFG (no previous game that year), the 1st two games of the season for L2_eFG (no previous two games), up through L10_eFG. But we will have to deal with these missing values so that we can create our models. Here is our **visual of what these missing values look like**:



"Imputing" missing values means that we take a "best estimate", based on a mixture of stats and subject area expertise or use case for this specific data.

The way I decided to replace these, was with the column for that year's average. For example, for the first 10 games of 2015, L10_eFG was replaced with the L10_eFG average of 49.6%. (This is also the most extreme case of imputing missing values in this dataset)

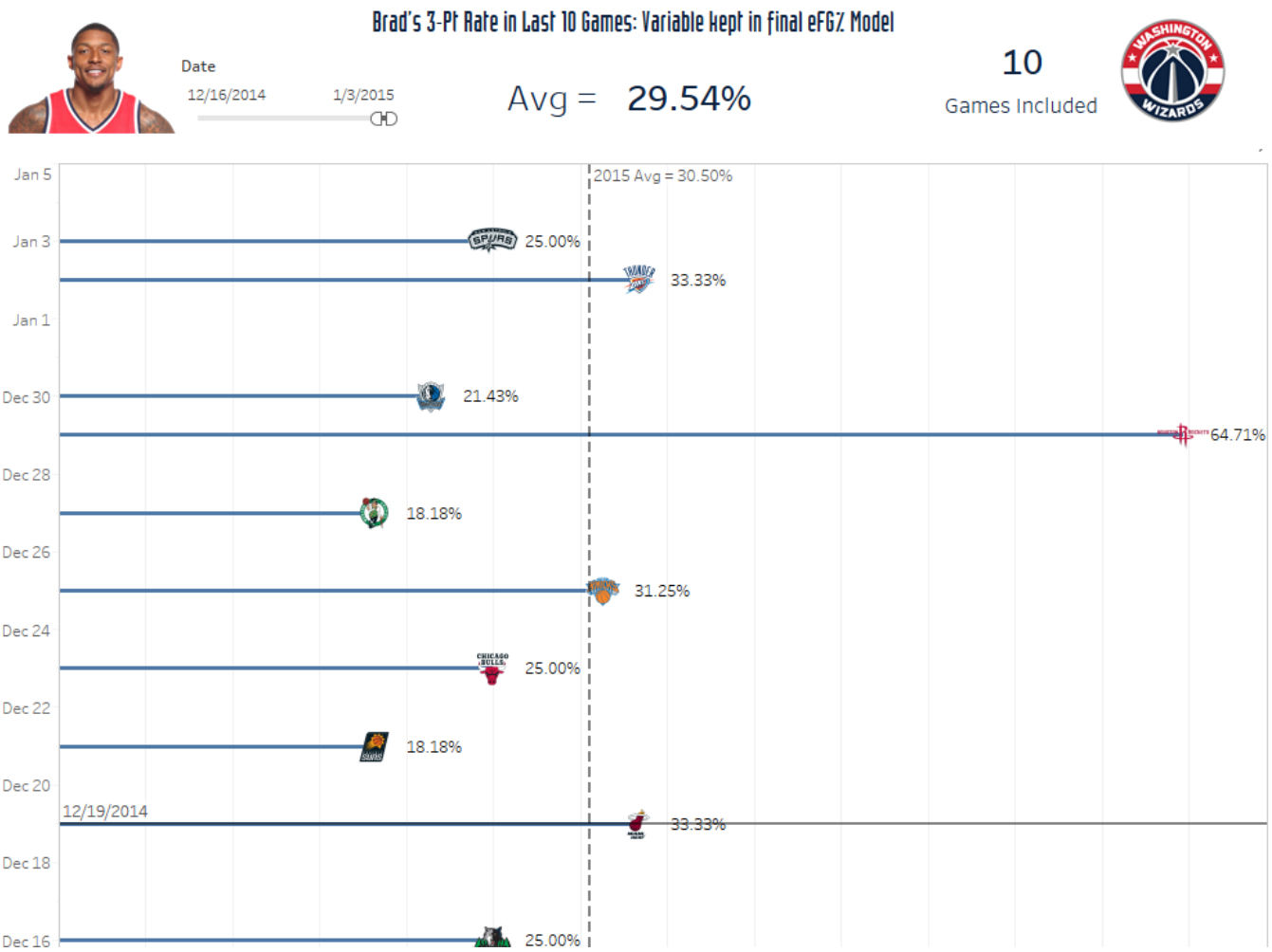
Modeling

I kept 106 variables to predict eFG% before going through an algorithm to select the best ones. I kept the variables that had at least a 10% correlation with eFG%.

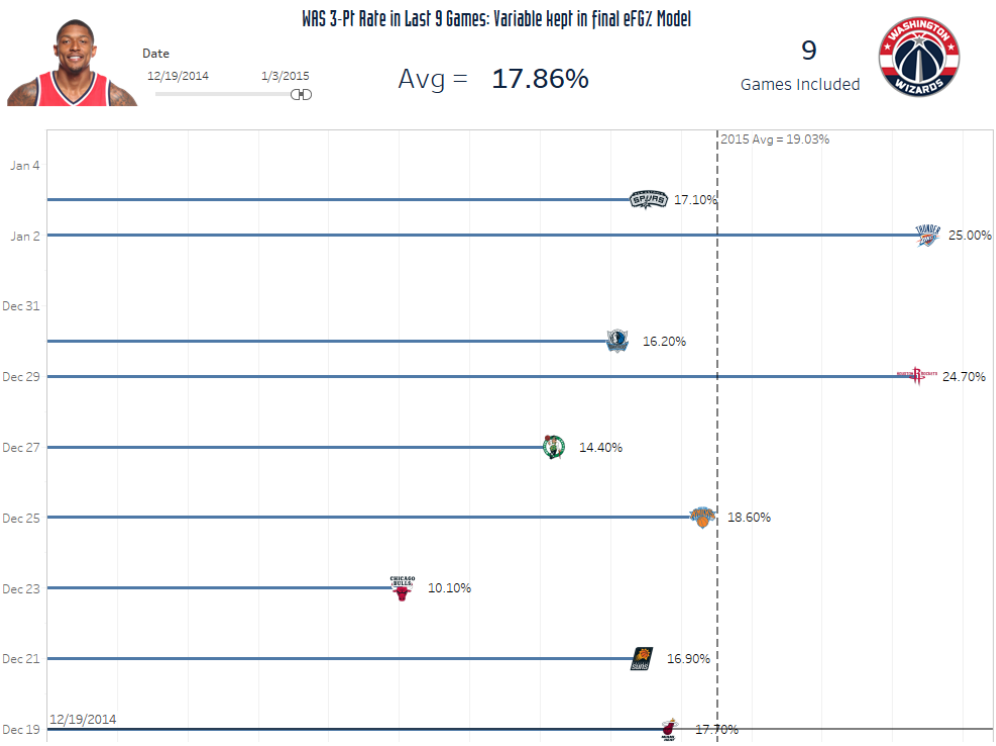
The full AIC model model predicted 40.0% eFG%. However, this model did not keep the Season's Opponent or Season, which seem to be paramount in knowing how Brad will shoot (answering the question "How good is he this year?" and "How good is he against this opponent, this year?")

In an attempt to incorporate these factors, before running the AIC algorithm, I forced the varaibles that were kept in the original model of **Brad Beal's 3-Pt Rate in his last 10 games, Washington Wizard's 3-Pt Rate in their last 9 games, Brad Beal's eFG% in his last 4 games, and Washington's Block% in their last 7 games.** To go along with these, I forced **Season** and **Season.Opp** to be in the final models.

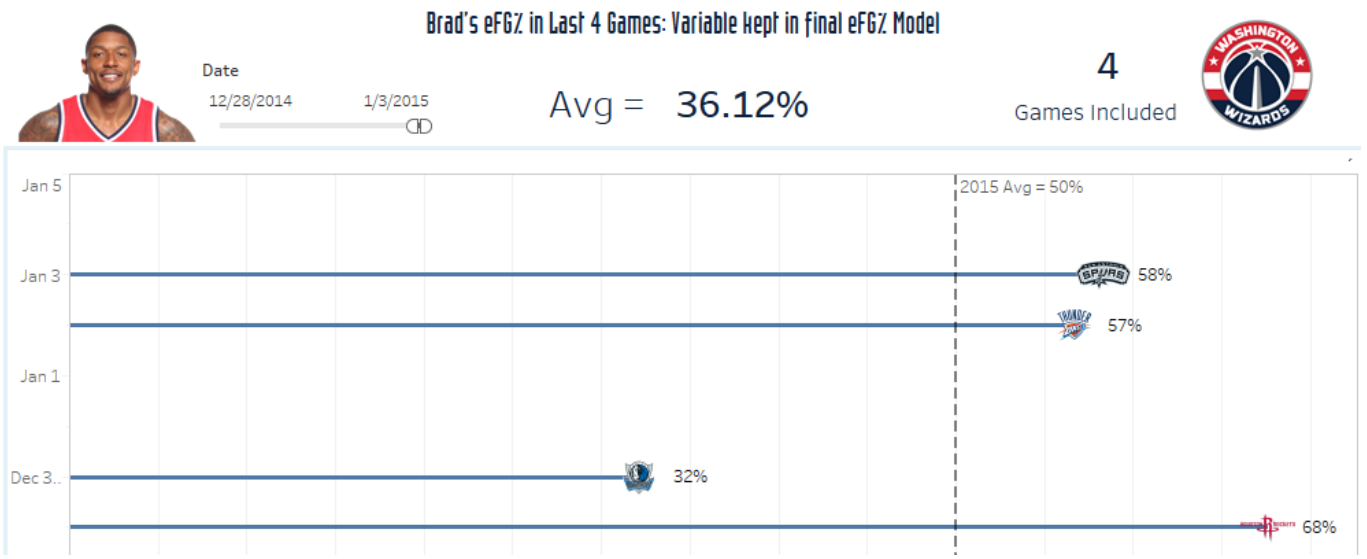
Last 10 Game 3-Pt Rate is about average for 2015:



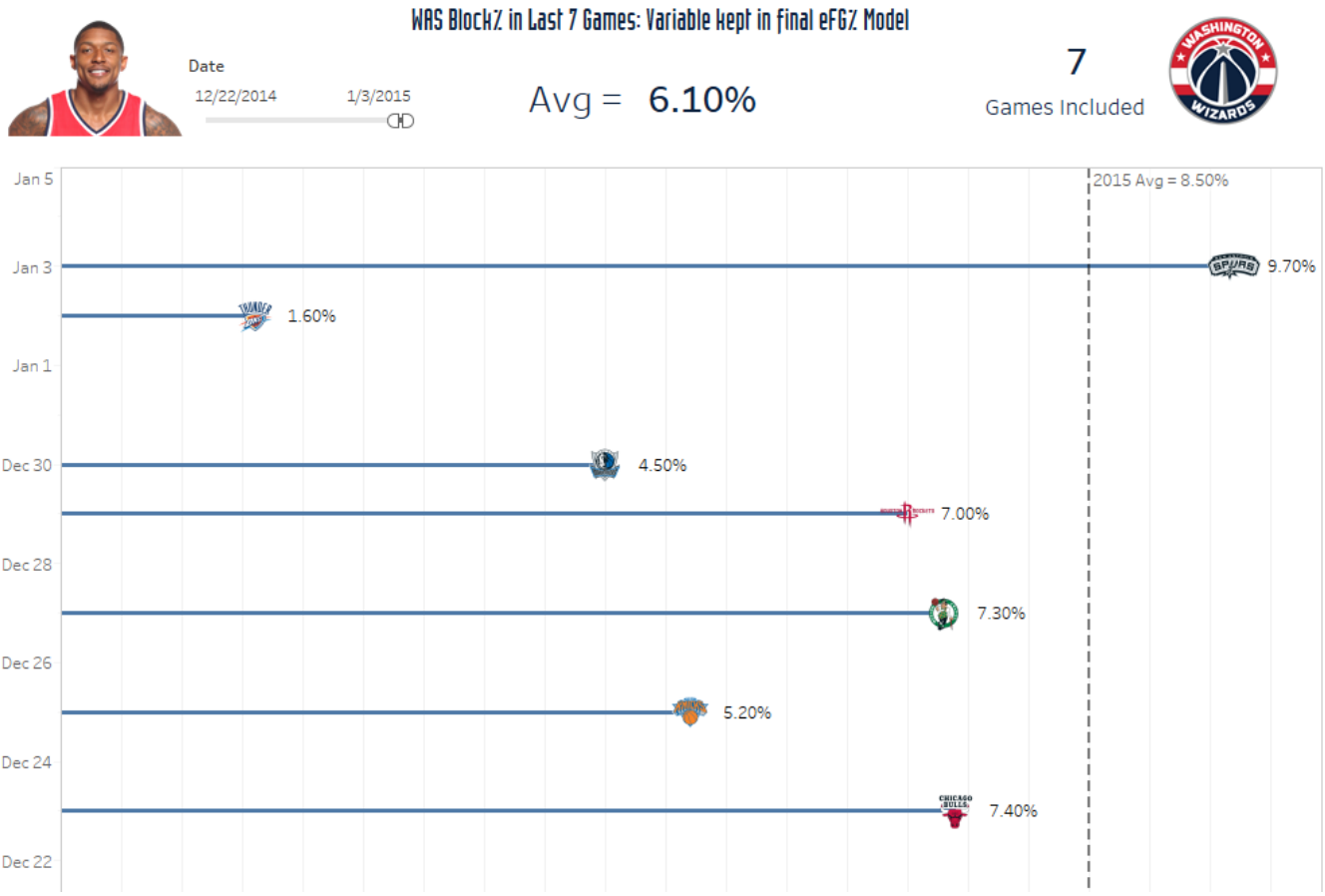
Washington's 3-Pt Rate is slightly below the season average:



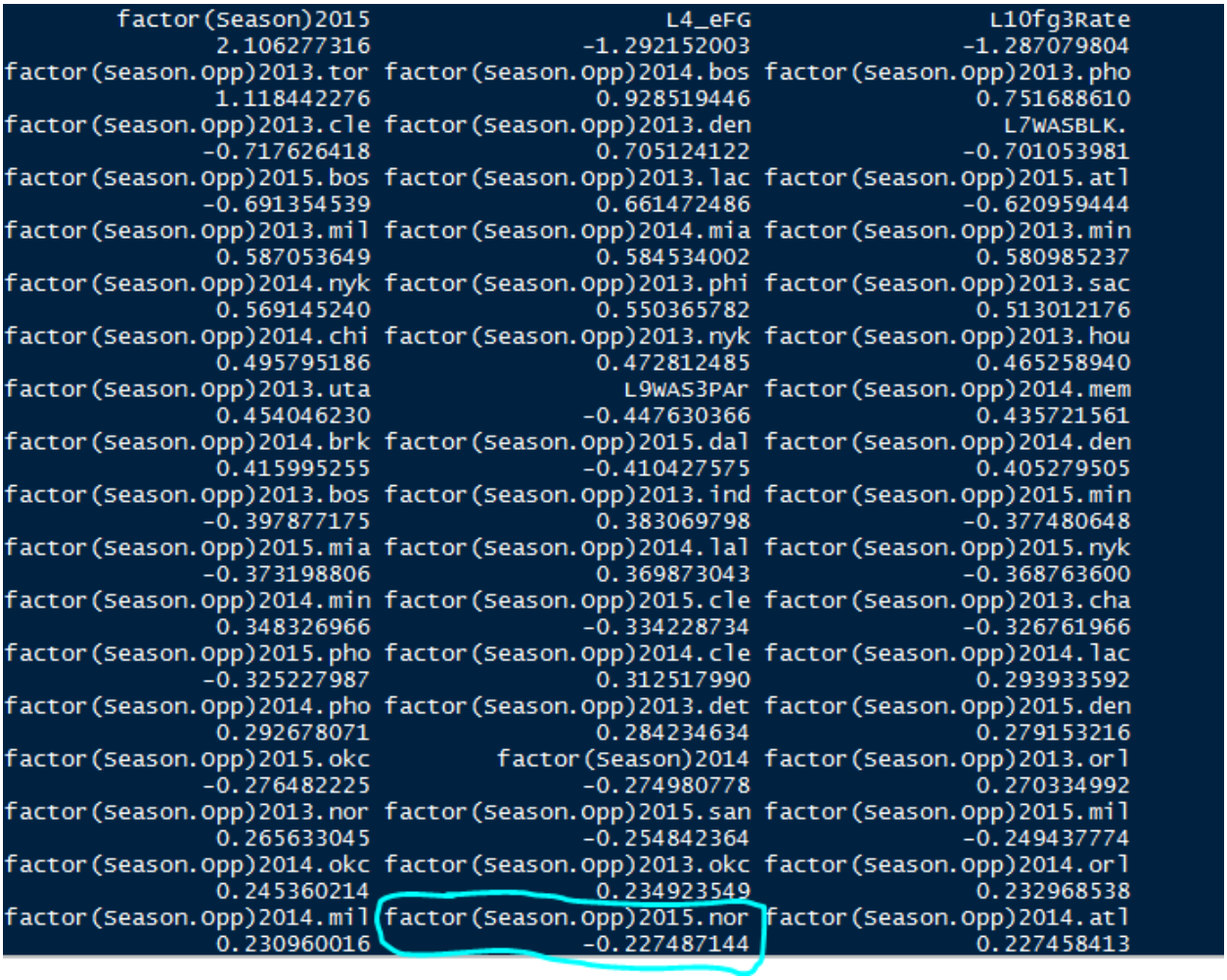
Brad's eFG% was lower than his season (and career) average in the last 4 games, which was lowered by a poor shooting game on 12/30 vs DAL:



Lastly, to go along with the Season's Opponent adjustment and Season adjustment, was Washington's Block% (estimate of Opponent's two-point FG blocked) in the last 7 games:



Here is an R Printout of the Final Model coefficient's, in order of **most important to least important**:

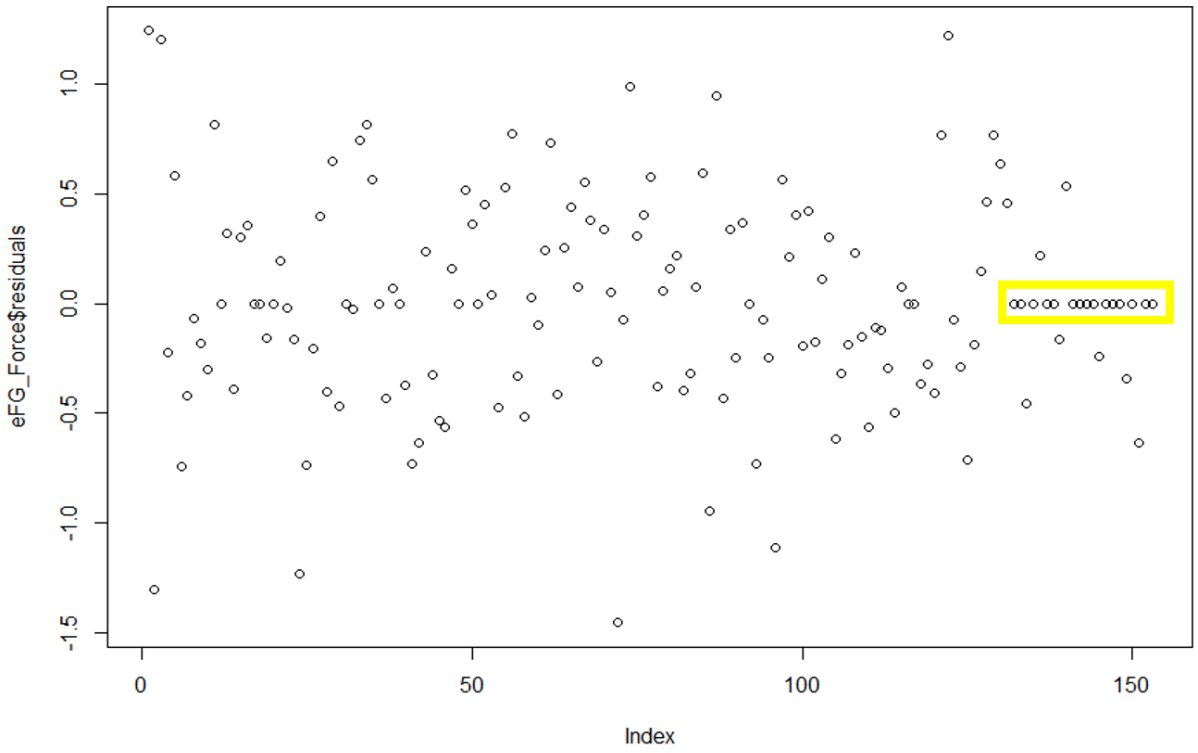


factor(Season)2015	L4_eFG	L10fg3Rate
2.106277316	-1.292152003	-1.287079804
factor(Season.Opp)2013.tor	factor(Season.Opp)2014.bos	factor(Season.Opp)2013.pho
1.118442276	0.928519446	0.751688610
factor(Season.Opp)2013.cle	factor(Season.Opp)2013.den	L7WASBLK.
-0.717626418	0.705124122	-0.701053981
factor(Season.Opp)2015.bos	factor(Season.Opp)2013.lac	factor(Season.Opp)2015.atl
-0.691354539	0.661472486	-0.620959444
factor(Season.Opp)2013.mil	factor(Season.Opp)2014.mia	factor(Season.Opp)2013.min
0.587053649	0.584534002	0.580985237
factor(Season.Opp)2014.nyk	factor(Season.Opp)2013.phi	factor(Season.Opp)2013.sac
0.569145240	0.550365782	0.513012176
factor(Season.Opp)2014.chi	factor(Season.Opp)2013.nyk	factor(Season.Opp)2013.hou
0.495795186	0.472812485	0.465258940
factor(Season.Opp)2013.uta	L9WAS3PAr	factor(Season.Opp)2014.mem
0.454046230	-0.447630366	0.435721561
factor(Season.Opp)2014.brk	factor(Season.Opp)2015.dal	factor(Season.Opp)2014.den
0.415995255	-0.410427575	0.405279505
factor(Season.Opp)2013.bos	factor(Season.Opp)2013.ind	factor(Season.Opp)2015.min
-0.397877175	0.383069798	-0.377480648
factor(Season.Opp)2015.mia	factor(Season.Opp)2014.lal	factor(Season.Opp)2015.nyk
-0.373198806	0.369873043	-0.368763600
factor(Season.Opp)2014.min	factor(Season.Opp)2015.cle	factor(Season.Opp)2013.cha
0.348326966	-0.334228734	-0.326761966
factor(Season.Opp)2015.pho	factor(Season.Opp)2014.cle	factor(Season.Opp)2014.lac
-0.325227987	0.312517990	0.293933592
factor(Season.Opp)2014.pho	factor(Season.Opp)2013.det	factor(Season.Opp)2015.den
0.292678071	0.284234634	0.279153216
factor(Season.Opp)2015.okc	factor(Season)2014	factor(Season.Opp)2013.orl
-0.276482225	-0.274980778	0.270334992
factor(Season.Opp)2013.nor	factor(Season.Opp)2015.san	factor(Season.Opp)2015.mil
0.265633045	-0.254842364	-0.249437774
factor(Season.Opp)2014.okc	factor(Season.Opp)2013.okc	factor(Season.Opp)2014.orl
0.245360214	0.234923549	0.232968538
factor(Season.Opp)2014.mil	factor(Season.Opp)2015.nor	factor(Season.Opp)2014.atl
0.230960016	-0.227487144	0.227458413

The variable that seems to drop the prediction is the Season.Opp = 2015.NOR. Keeping in mind that a higher-variable model may not make intuitive sense right away, if at all (since Brad Beal has shot a bit higher than average this year against New Orleans), playing the Pelicans in 2014-15 drops the prediction for Brad's eFG%.

However, though we know in retrospect this model was near precise in its prediction, I would have never suggested to use this model before the game, because the model overfit recent games leading up to the game:

Residuals for the Model that Predicted 29.6% eFG% for Next Game



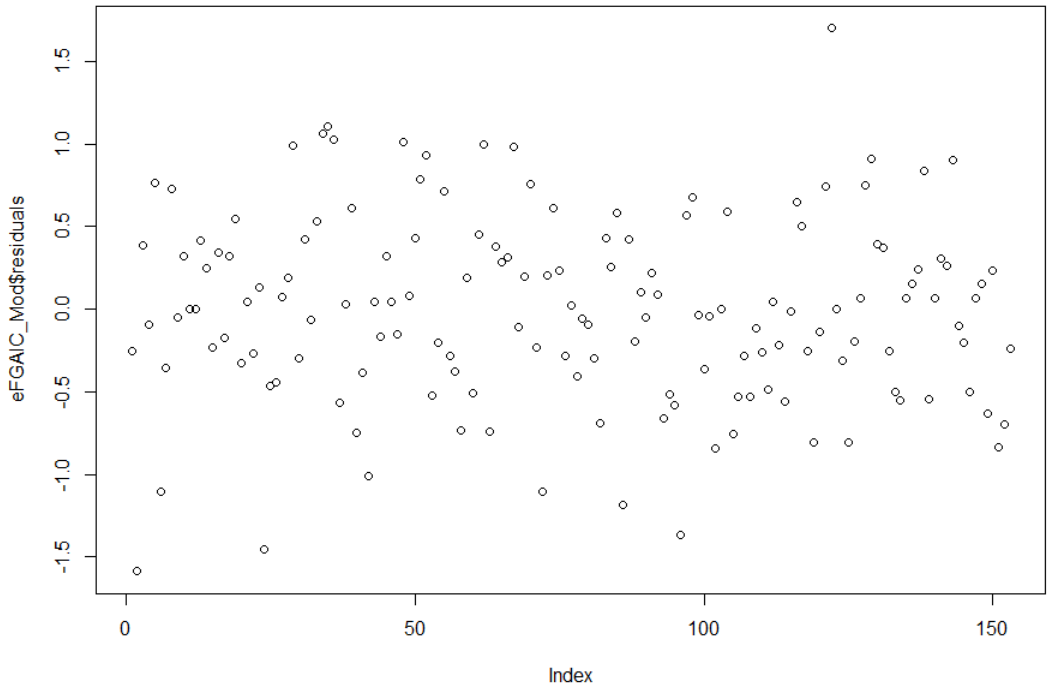
So, here are the **standardized coefficient for the model that I ended up going with in order of most important to least important**:

1. Brad's eFG% in last 3 games (-1.37)
2. Brad's 3-Pt Rate in Last 10 Games (-1.15)
3. Washington's Assist% in Last 2 Games (1.00)
4. Washington's 3-Pt Rate in last 9 Games (-0.97)
5. Washington's Block % in Last 7 Games (-0.90)
6. Washington's Opponent's Defensive Rebound% Last Game (0.81)
7. Washington's Turnover% Last Game (0.60)

The "intercept" term accounts for how well Brad shoots in general, and adjust using the above metrics

There are two main reasons I like this model the best.

1. The **residuals** are beautifully, **randomly and uniformly distributed about 0** (a MUST for a model that does not overfit or bias the predictions):



2. The variables account for LOTS of information. I really like being able to account for as much information as possible, even by including some variables that may not be as highly correlated. We see we account for Brad's shooting and 3-point rate first (the highest correlated variables), then Washington's Assist%, 3-Pt Rate, Block%, Opponent's Defensive Rebound%, and Turnover%. I'd say we accounted for lots more information than a model that forces you to account for every single team coefficient for every year (which is presumably how the model that was so close overfit to the most recent games).

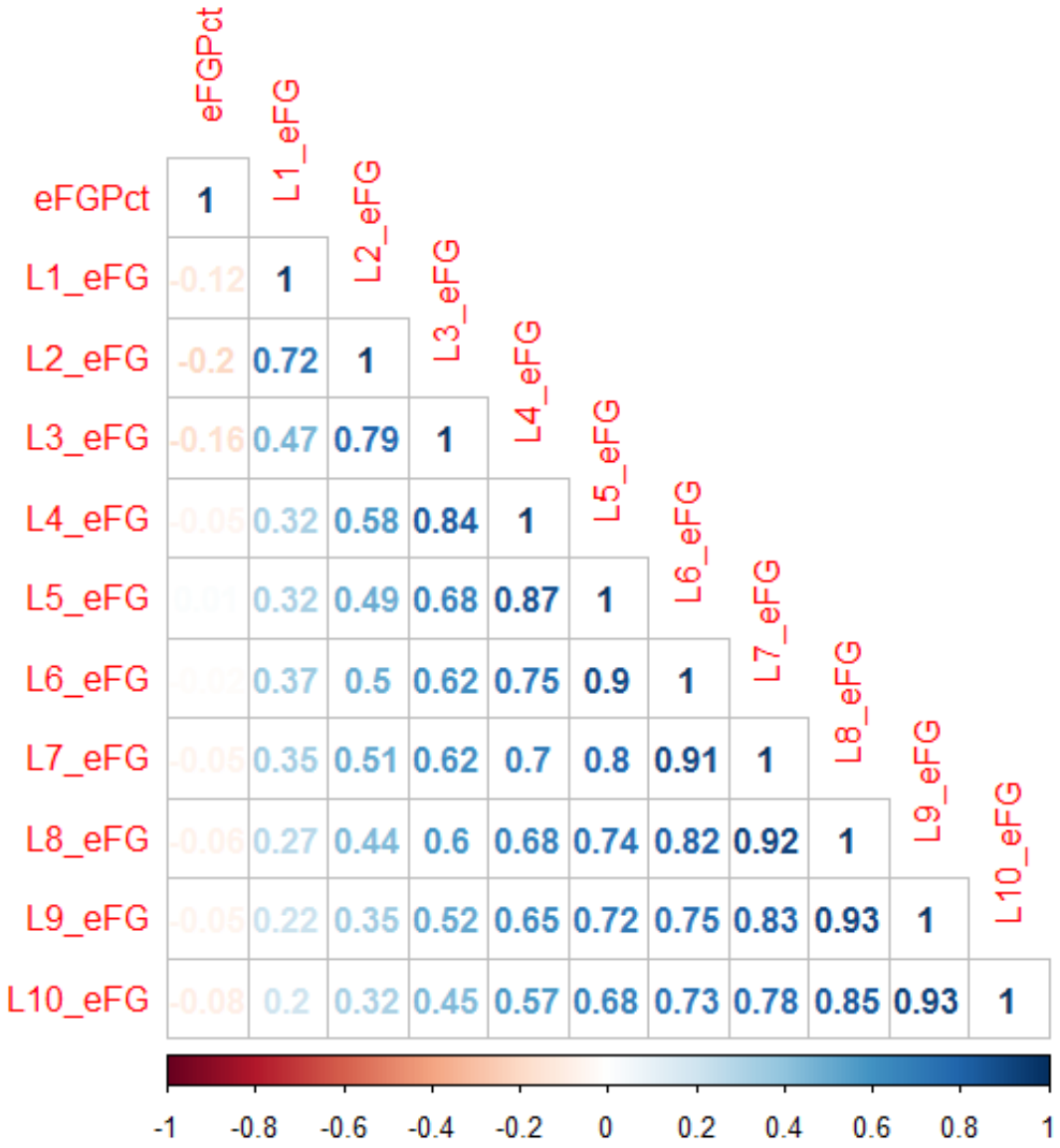
As a note, in practice (and without having hindsight to be able to help gauge in any way), I would do far more diagnostics of each model and potentially go with a mixture of multiple models. However, looking at the residuals and seeing, intuitively, how much information we covered with this model, I feel good about using this model for this project.

It would also be good practice to do group mean testing on our categorical variables such as Opponent, Season, Day of Week.

Predicting 1/7/15 vs NYK and 1/9/15 vs NOP: 2 Games Away and 3 Games Away

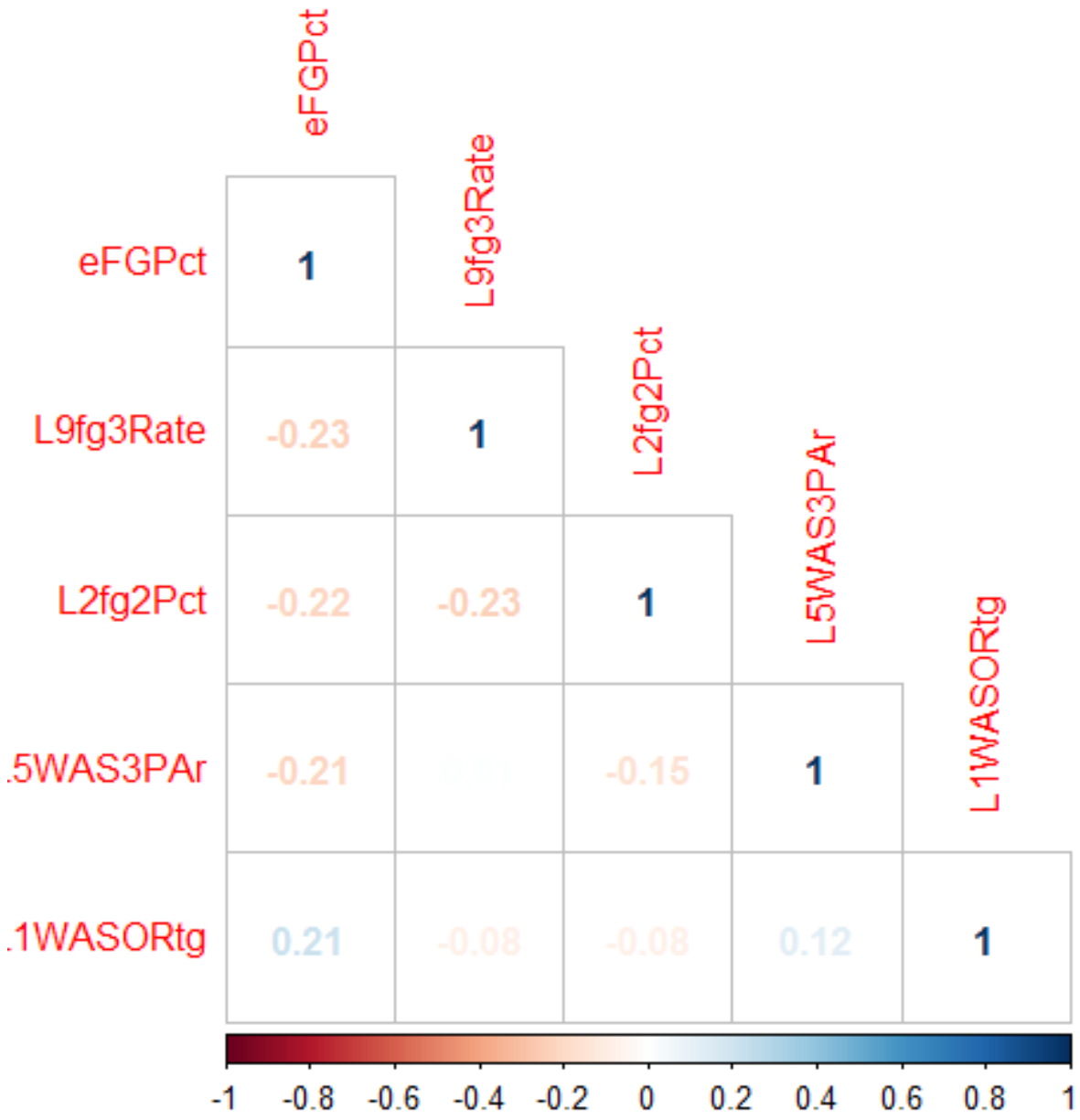
The methodology and process were exactly the same as listed above for predicting 2 games away and three games away. The only difference is, the game info such as Date, Oponent, and eFG% (which we are predicting) were moved up one row in excel. That is, we use the most current information as of 1/4/15 to predict two games away for 1/7/15 and three games away for 1/9/15.

Brad's last 2 games eFG% are related to his eFG% 2 games from now at -20%



Highest Correlations to Brad's eFG% **Two Games from Now**

Very similarly to before (not surprisingly), the biggest correlations are his 3-Pt Attempt Rate in his last 9 games (it was 10 before) at -23%. Then was his 2-Pt FG% in his last two games at -22%, Washington's 3-Pt Rate in their last 5 games at -21% and Washington's Offensive Rating last game at 21%.



Again, similar to predicting the next game, I started with building a full model based on the variables that were correlated at least 10% with eFG% for 2 games from now. Then I started with a base model for force Season and Season Opponent in the final model.

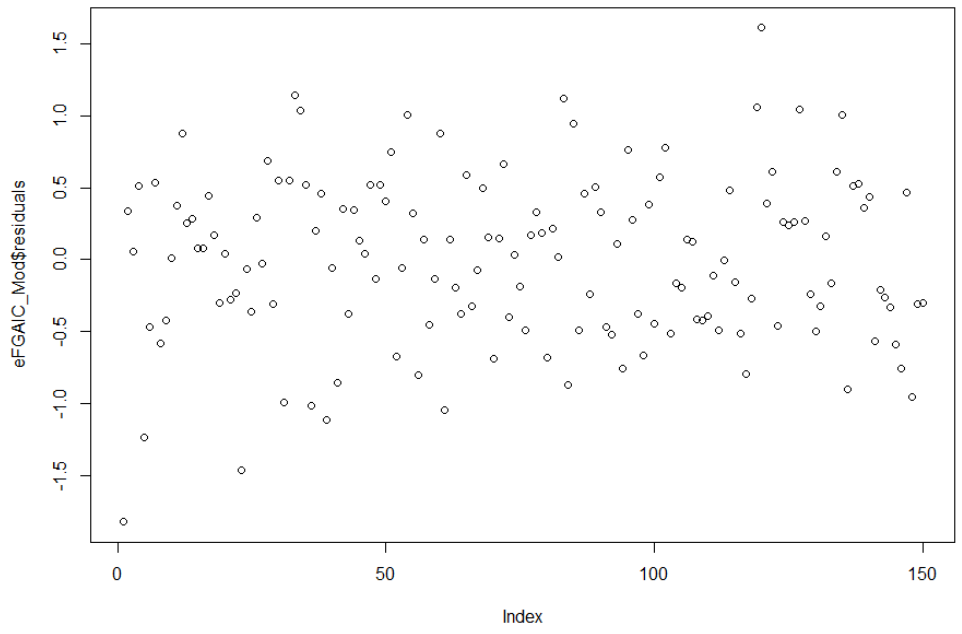
Illustrating the danger of over-fitting, the model that we did not go with before (which overfit the data based on the residual plot above), did not make any sense, as it "predicted" a negative number.

But, using the exact same algorithm with the highest correlated variables, here is the model that was used for two games away:

Standardized **coefficient for the model in order of most important to least important**:

1. Washington's 3-Pt Rate in last 9 Games (-1.51)
2. Washington's Offensive Rating last game (1.05)
3. Washington's Turnover% in the last 4 Games (-1.04)
4. Average Days of Rest in the Last 9 Games; Days between games (-0.89)
    a. Note: This metric is updated through the game date since we already know the date of the games and the games that were played before; it assumes Brad played next game
5. Brad's average seconds played in the last 4 Games (0.86)
    a. However, this is only updated through 1/4/15 (we do not have 1/5/15 vs NOP yet)
6. Brad's Offensive Rating in the last 2 games (-0.86)
7. Brad's average percent of shots taken that were 3's or Lay-ups the last 3 Games (-0.85)
8. Washington's Block % in Last 6 Games (-0.72)

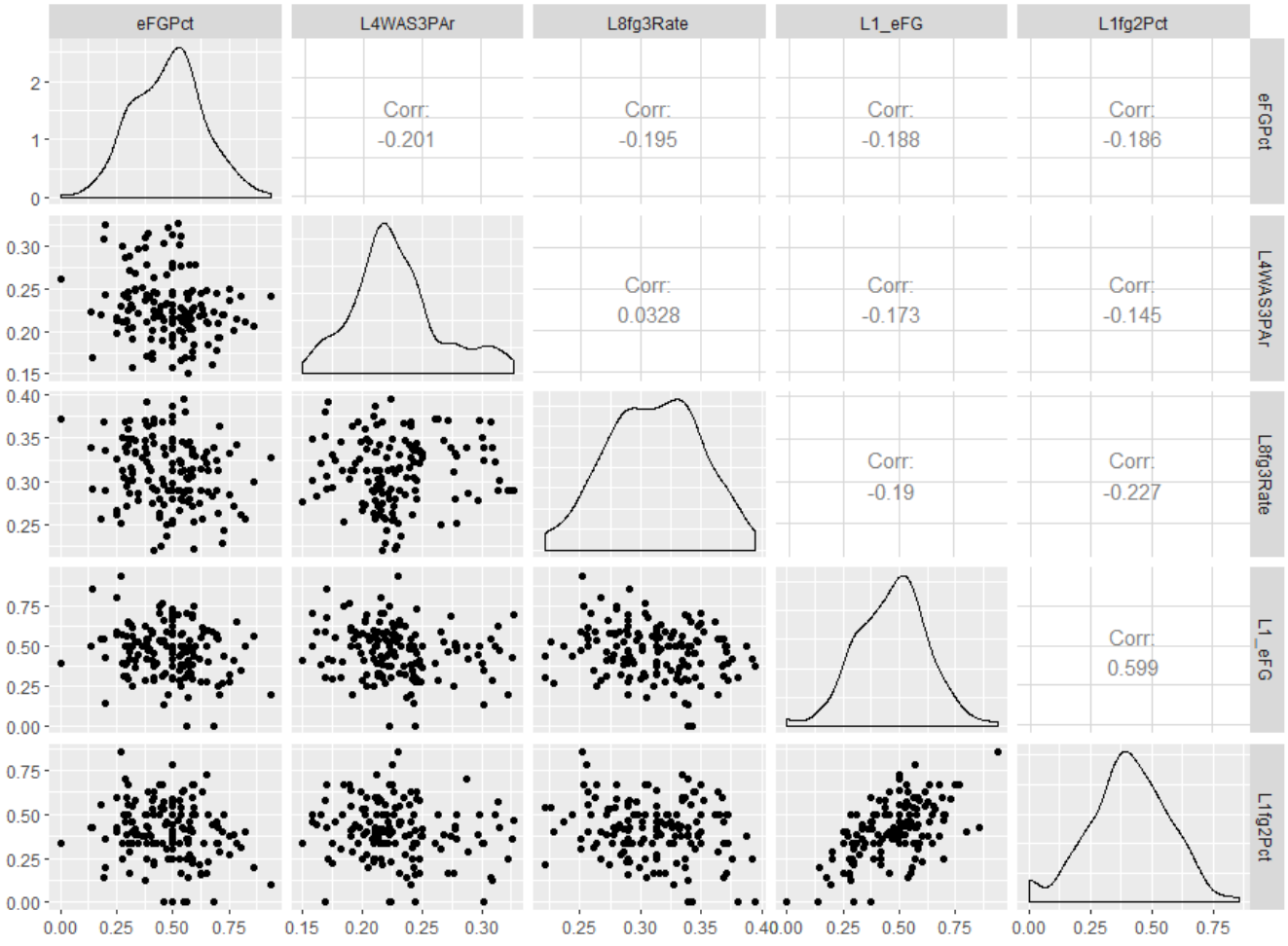
Model Residuals: Randomly/Uniformly Distributed about 0



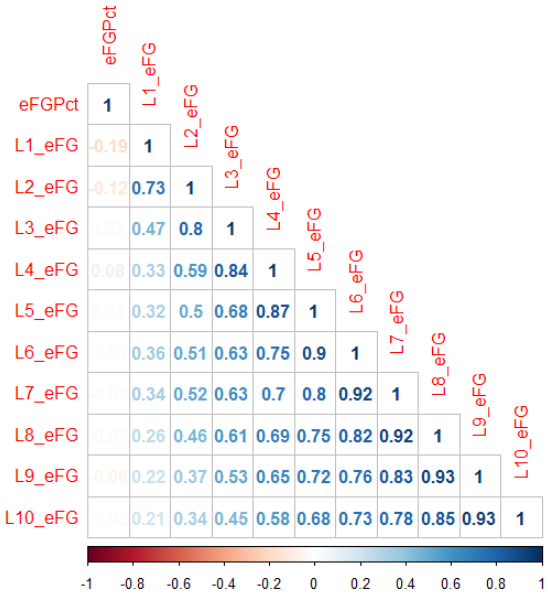
And finally, the exact same methodology was repeated for games that are 3 days out from now to predict 1/9/15 vs CHI.

Highest Correlations to Brad's eFG% **Three Games from Now**

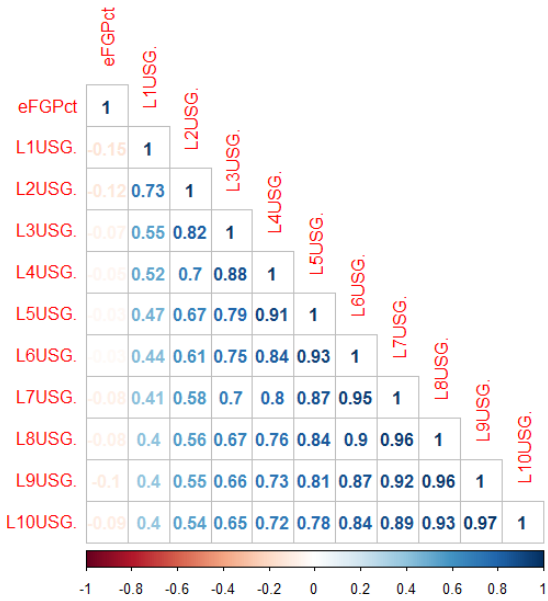
1. Washington's 3-Pt Rate the Last 4 Games (-20.1%)
2. Brad's 3-Pt Rate the last 8 games (-19.5%)
3. Brad's eFG% last game (-18.8%)
4. Brad's 2-Pt% last game (-18.6%)



Not surprisingly, the highest correlated, individual recent eFG% was only the last game to 3 games from now (predicting 1/9/15) at -19%:



And Brad's Usage% last game at -15%:

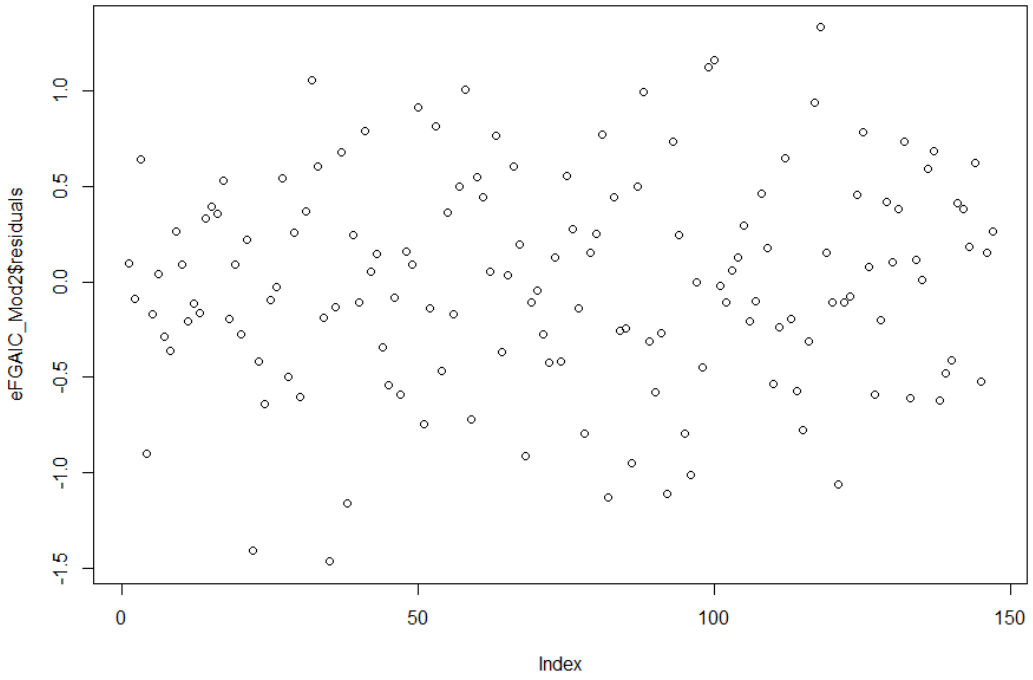


And by the exact same logic, using the top correlated variables and AIC selection, here are the models and residuals used for our prediction for 1/9/15 vs CHI:

Standardized **coefficient for the model in order of most important to least important**:

1. Washington's 3-Pt Rate in last 9 Games (3.6)
2. Brad's 3-Pt Rate in the Last 9 Games (-2.56)
3. Brad's 3-Pt Rate in the Last 7 Games (2.56)
4. Washington's 3-Pt Rate in the last 7 Games (-2.43)
5. Washington's Opponent's eFG% in the Last 9 Games (-1.93)
6. Washington's Opponent's Defensive Rebound% in the Last 9 Games (1.73)
7. Washington's Offensive Rating in the Last 10 Games (1.53)
8. Washington's Opponent's Points in the Last 10 Games (-1.49)
9. Washington's Block% in the Last 5 Games (-1.14)
10. Brad's 3-Pt Rate in the Last 4 Games (-1.00)
11. Washington's Offensive Rebound% in the Last 4 Games (-0.94)
12. Brad's Lay-up Percent last game (-0.93)

Model Residuals: Randomly/Uniformly Distributed about 0



Predicting 3 games out was a little more difficult. The other two predictions were pretty quick after the algorithm and made intuitive sense. I do not like that the final model for three days out used the same variable twice, but simply with different numbers of days included—i.e., team 3-pt Rate and individual 3-pt rate. There was a different model that predicted 37%, and another more simple, 2-variable model that predicted 49.1%.

Changes Going Forward

With more time, I would like to test Lasso for finding coefficients. More importantly, I would like to test and incorporate season and Opponent+Season. Further, I would like to take an average of the best scientific models, to get a bit better version of the middle value of what previous data suggests. Further, I would do more diagnostics to test the models. And certainly, as with any modeling, getting better and more granular data could open up new insights into models that cannot be found without either more data or more feature engineering (finding new variables). Lastly, there should be cross-validation and better means of generalizing these models.