Project 2 Summary:
Predicting Daily NBA Fantasy Points

Linda Ju

Motivation

Daily Fantasy Sports (DFS) is a $7 billion industry with millions of users who spend thousands of dollars every day placing bets. I am one of those users. I play fantasy basketball and fantasy football, the former of which is the subject of my second project. I performed analysis to predict a basketball player's fantasy production in the hopes of making me a more successful DFS player.

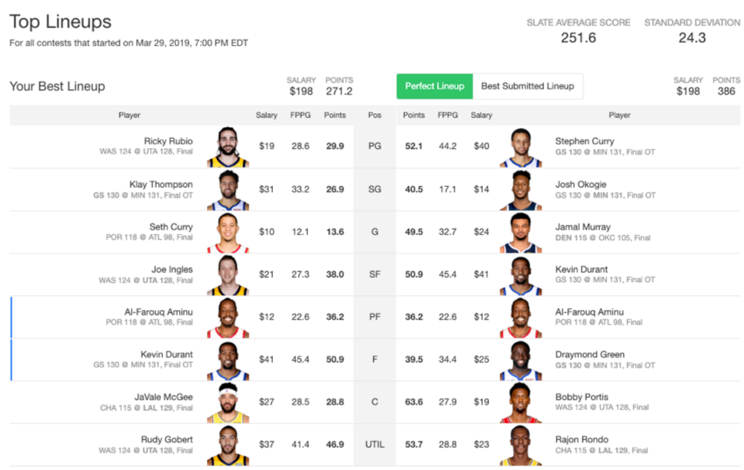
What is daily fantasy basketball?

The objective of daily fantasy basketball is to create a team of NBA basketball players who score more fantasy points than your opponents' teams. Each player has a price associated with them to add to your fantasy team. Generally, the more productive players have higher price tags and vice versa. Each team manager is allotted a budget to spend on any player from any team playing that night in order to add them to their fantasy team.

As the night progresses and the NBA players rack up their stats, such as points, rebounds, assists, steals, blocks, and turnovers, your team's total fantasy points increase with positive production (for example points scored and number of rebounds) and decrease with negative production (for example, turnovers). If your players play well that night, they will score more fantasy points for your team. If your players play poorly, you'll have fewer fantasy points.

<u>Example</u>

The figure to the right is a screenshot of my daily fantasy results on March 29. The eight players on the left side of the screenshot from Ricky Rubio to Rudy Gobert are the players I chose for my team. I paid $198 out of my $200 budget to create my team. They scored 271.2 points. The eight players on the right from Stephen Curry to Rajon Rondo scored the highest points possible that night (386 points) within the $200 budget. The salary for each player are displayed, their average fantasy points per games is shown under 'FPPG' (Fantasy Points Per Game), and their actual points scored is provided under 'Points'. That night, my lineup came in 4th place compared to 19 other lineups in the contest and I won some money. The slate average score and standard deviation shown at the top right of the screenshot are the average and standard deviation of the scores in my contest. Finally, the middle column in gray is a player's position (point guard, shooting guard, center, etc.) Each team is required to choose players that fill these positions. For example, you cannot have a team of all point guards.



Top Lineups

For all contests that started on Mar 29, 2019, 7:00 PM EDT

SLATE AVERAGE SCORE 251.6 | STANDARD DEVIATION 24.3

Your Best Lineup — SALARY $198 POINTS 271.2 | Perfect Lineup | Best Submitted Lineup — SALARY $198 POINTS 386

Player	Salary	FPPG	Points	Pos	Points	FPPG	Salary	Player
Ricky Rubio WAS 124 @ UTA 128, Final	$19	28.6	29.9	PG	52.1	44.2	$40	Stephen Curry GS 130 @ MIN 131, Final OT
Klay Thompson GS 130 @ MIN 131, Final OT	$31	33.2	26.9	SG	40.5	17.1	$14	Josh Okogie GS 130 @ MIN 131, Final OT
Seth Curry POR 118 @ ATL 98, Final	$10	12.1	13.6	G	49.5	32.7	$24	Jamal Murray DEN 115 @ OKC 105, Final
Joe Ingles WAS 124 @ UTA 128, Final	$21	27.3	38.0	SF	50.9	45.4	$41	Kevin Durant GS 130 @ MIN 131, Final OT
Al-Farouq Aminu POR 118 @ ATL 98, Final	$12	22.6	36.2	PF	36.2	22.6	$12	Al-Farouq Aminu POR 118 @ ATL 98, Final
Kevin Durant GS 130 @ MIN 131, Final OT	$41	45.4	50.9	F	39.5	34.4	$25	Draymond Green GS 130 @ MIN 131, Final OT
JaVale McGee CHA 115 @ LAL 129, Final	$27	28.5	28.8	C	63.6	27.9	$19	Bobby Portis WAS 124 @ UTA 128, Final
Rudy Gobert WAS 124 @ UTA 128, Final	$37	41.4	46.9	UTIL	53.7	28.8	$23	Rajon Rondo CHA 115 @ LAL 129, Final

Data
In order to limit the scope of my analysis, I focused my modeling on the top 100 NBA guards from the 2018-2019 regular season. The fantasy points scoring method I chose to predict is Yahoo's:

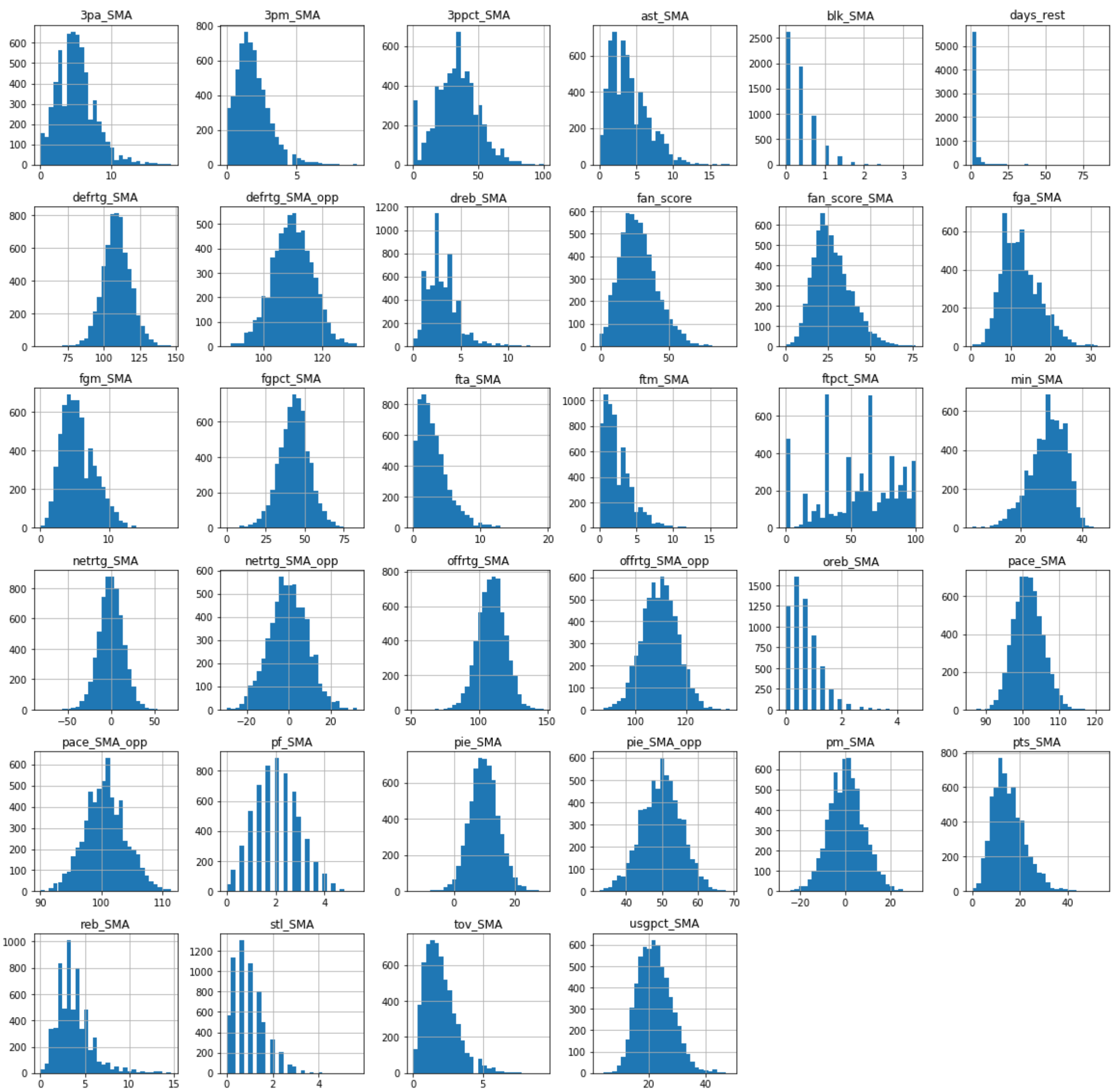
$$\text{Fantasy Points} = \text{points} + 1.2 \times \text{rebounds} + 1.5 \times \text{assists} + 3 \times \text{steals} + 3 \times \text{blocks} - \text{turnovers}$$

From the NBA Advanced Stats website, I scraped 34 features for over 6,000 observations. The features are as follows:

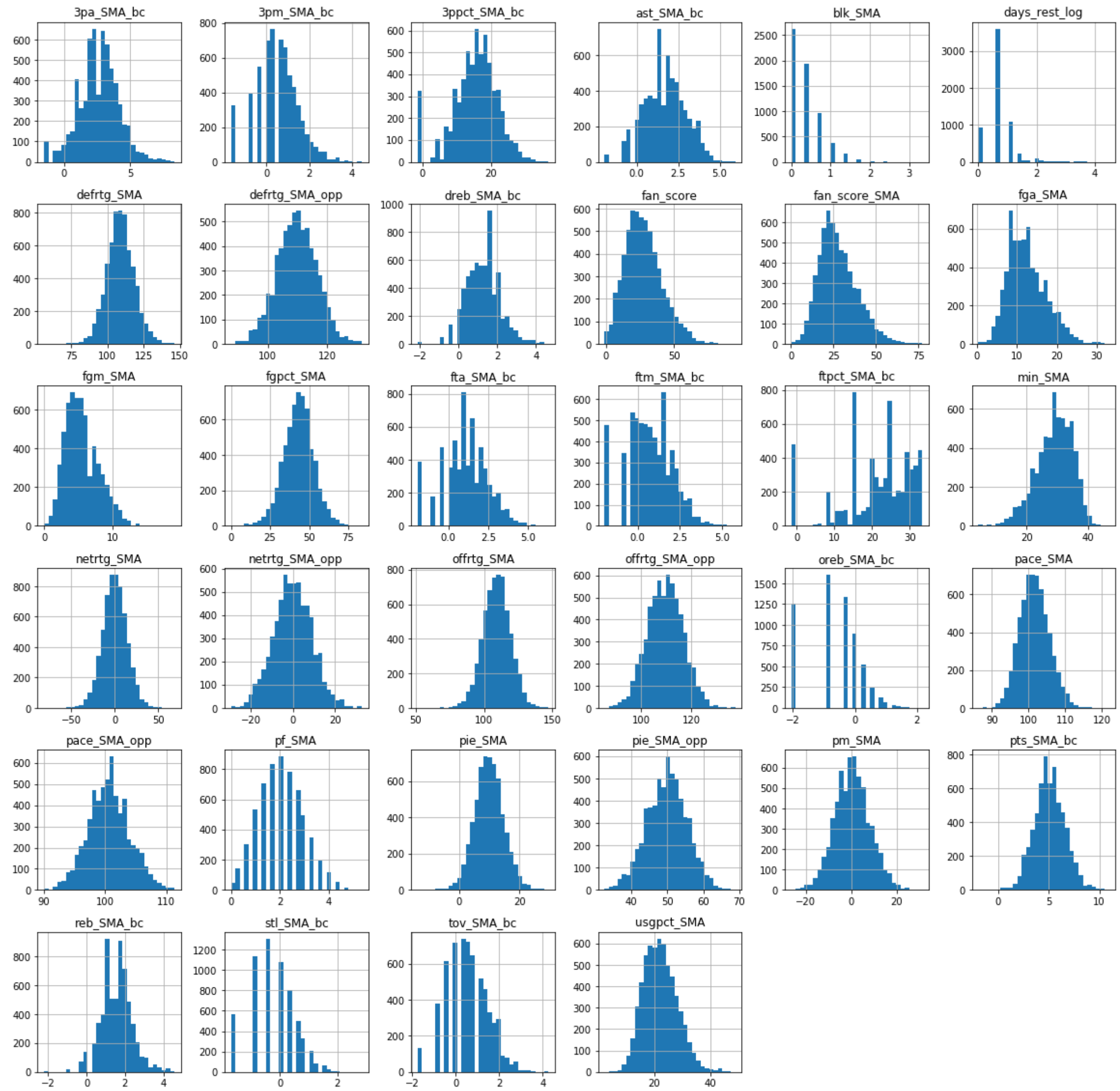
Source	**Example**	**Features**
Player's Traditional Game Log	James Harden	• Simple avg of the past three games: ○ Minutes Played ○ Points ○ Field Goals Made ○ Field Goals Attempted ○ Field Goal Percentage ○ 3 Point Field Goals Made ○ 3 Point Field Goals Attempted ○ 3 Point Field Goals Percentage ○ Free Throws Made ○ Free Throws Attempted ○ Free Throw Percentage ○ Offensive Rebounds ○ Defensive Rebounds ○ Rebounds ○ Assists ○ Steals ○ Blocks ○ Turnovers ○ Personal Fouls ○ Plus-Minus ○ Fantasy points • Home or away game • Number of days since last game
Player's Advanced Stats	James Harden	• Simple avg of the past three games: ○ Offensive Rating ○ Defensive Rating ○ Net Rating ○ Usage Percentage ○ Pace ○ Player Impact Estimate
Opponent's Advanced Stats	Teams Advanced Box Scores	• Simple avg of the past three games: ○ Offensive Rating ○ Defensive Rating ○ Net Rating ○ Usage Percentage ○ Pace

Exploratory Data Analysis
Prior to modeling, I performed EDA to get a feel for my data. To clean the data, I tried normalizing the features and removing outliers in the target. However, neither of these data cleansing methods increased the quality of my analysis. Therefore, I ended up not transforming or removing outliers in my final model.

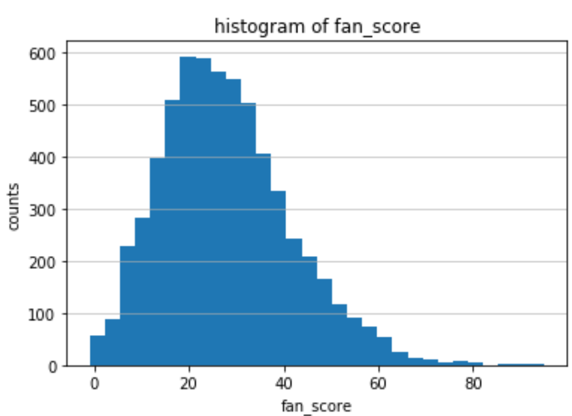
As part of the EDA, I plotted a histogram for each feature (with the exception of home vs. away) and the target:

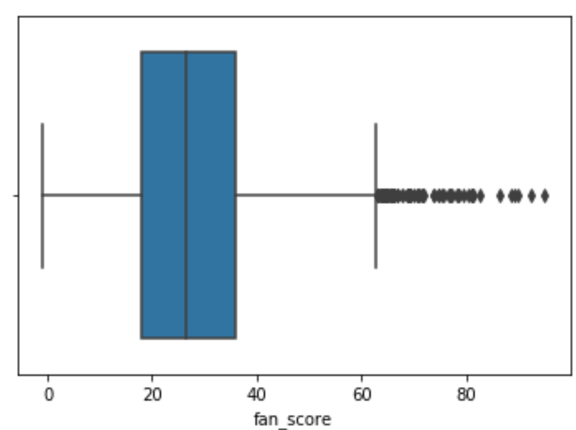


Some of the features, such as assists and defensive rebounds, did not look “normal”. Therefore, I applied box-cox and log transformations which resulted in the following distributions.

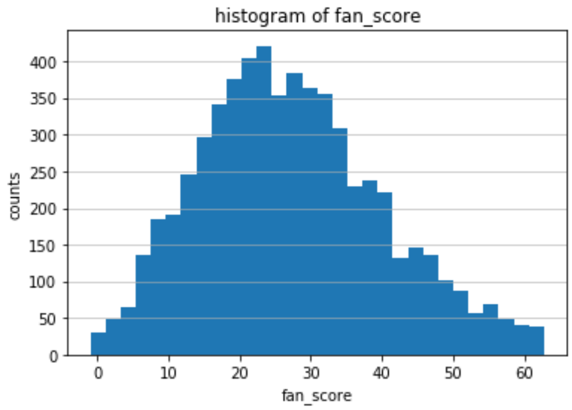


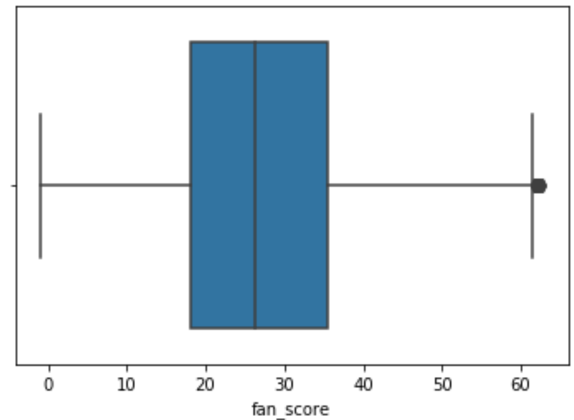
Additionally, I noticed the target distribution is slightly right-skewed:





Therefore, I tried removing the outliers:





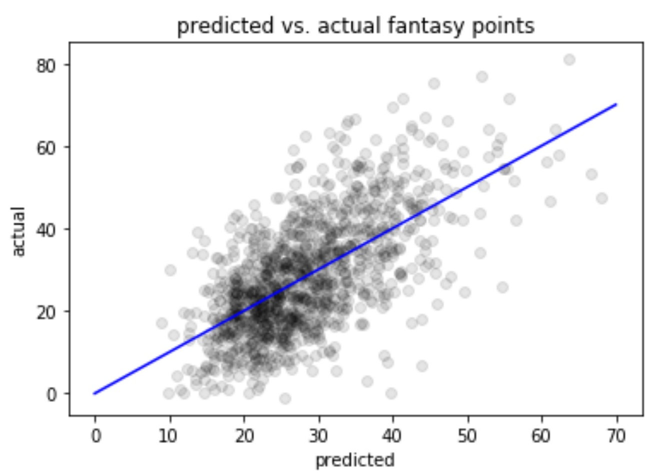
As mentioned, cleaning the data did not significantly improve my model results, therefore my final model used the original data with no outliers removed.

Modeling Process

I used linear regression to predict fantasy points scored from the 34 features. To reduce the complexity of my model, I standardized the features and used lasso/ridge regularization and cross validation in order to find the best lambda. Finally, I compared the test data R^2 of the two regularization methods in order to choose the best one.

Results

Lasso regularization gave slightly better results than ridge regularization, with R^2 of approximately 41% and mean absolute error approximately 8.4. The best lambda was about 0.086.

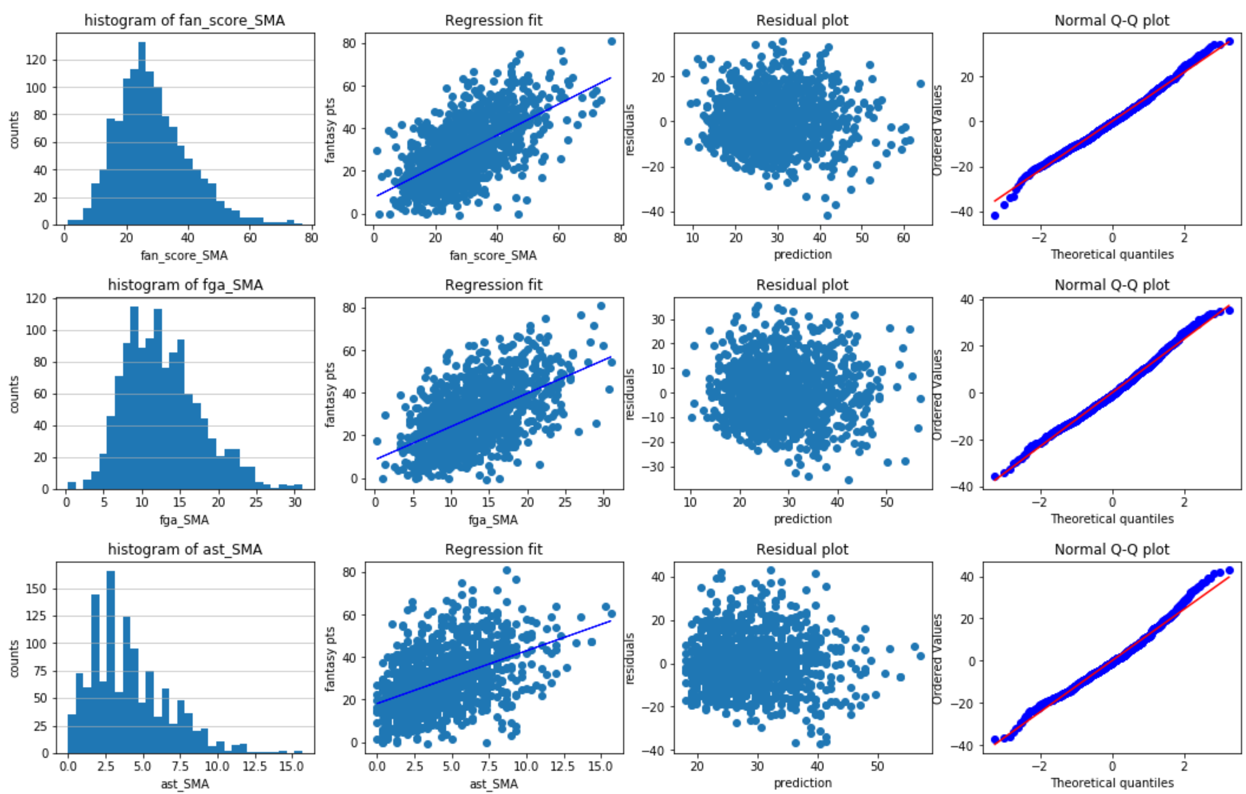


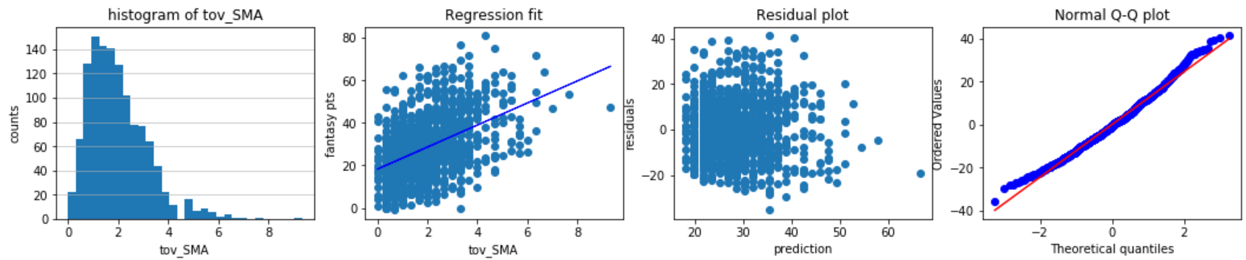
The features with non-zero coefficients are shown below in decreasing order of absolute magnitude:

1. Avg fantasy points in the past three games: $\beta \approx 2.92$
2. Avg field goal attempts in the past three games: $\beta \approx 2.07$
3. Avg number of assists in the past three games: $\beta \approx 1.66$

4. Avg number of turnovers in the past three games: β ≈ 1.57
5. Avg number of defensive rebounds in the past three games: β ≈ 0.83
6. Avg free throws made in the past three games: β ≈ 0.73
7. Avg 3-pt percentage in the past three games: β ≈ -0.48
8. Avg opponent pace in the past three games: β ≈ 0.39
9. Avg free throw attempts in the past three games: β ≈ 0.36
10. Avg player impact estimate in the past three games: β ≈ -0.31
11. Number of rest days: β ≈ -0.28
12. Avg opponent defensive rating in the past three games: β ≈ 0.21
13. Avg number of blocks in the past three games: β ≈ 0.19
14. Avg number of steals in the past three games: β ≈ 0.19
15. Avg opponent player impact estimate in the past three games: β ≈ -0.18
16. Avg defensive rating in the past three games: β ≈ 0.17
17. Avg 3-pt attempts in the past three games: β ≈ 0.12
18. Avg offensive rating in the past three games: β ≈ 0.07
19. Home vs. away game: β ≈ 0.05
20. Avg pace in the past three games: β ≈ -0.05
21. Avg free throw percentage in the past three games: β ≈ -0.01

An additional step I took after I obtained my results was checking the residuals for any patterns. Shown below are the residuals of the features with the highest coefficients. The residuals seem random.





Conclusion
According to the resulting coefficients, players who on average scored the most fantasy points in the past three games, attempted the most shots, passed the ball the most often, and got the most rebounds are likely to score the most fantasy points.

Russell Westbrook fits this description perfectly. He shoots often, makes many assists, and grabs rebounds. He hasn't won any real championships yet, but he's been on many winning fantasy teams.

Future Work
In the future, I'd like to test out more features such as assist to turnover ratio, playoff standing relative to time to playoffs, and stats allowed by opponents to improve prediction power.

I'd also like to test different ways of representing past performance. For example, instead of taking the average of the past three games, perhaps it's better to take the average of the past five or ten games. Another alternative is taking a weighted average of past performance, with more weight given to more recent performance.

Finally, in order to pick an entire team with multiple positions, I'll need to extend this analysis to forwards and centers. In addition, the model is set up to predict how many points a player may score in their next game, however, the most valuable information would be to know which players are the most undervalued based on their salary requirement.