How to Win a Basketball Game

1) Problem statement:

The Problem:

What are the better predictors in NBA wins: offensive statistics or defensive statistics?

We are going to build several models using only offense-related features, and several models using only defense-related features, and evaluate which models perform better.

Who could benefit from this information?

- NBA management
    - Determine what are the strengths and weaknesses of a particular team.
    - Where to direct focus for team development.
    - How to prepare for a matchup against a particular team.
- Sports analysts/broadcasters
    - More detailed predictive analysis of specific NBA matchups.
    - Objective, numerical measures of team performance.

2) Data Wrangling/Cleaning

For my data, I chose to work with SportRadar's NBA API. The API features a number of different endpoints for obtaining various NBA data, including summary statistics, player profiles, and league information. I chose to analyze summary statistics for every game between 2013 and 2017.

Obtaining the data:

1) Each game has a specific ID associated with it. These IDs are located within the "Schedule" endpoint on SportRadar. For each year from 2013 to 2017, I obtained a JSON with that year's schedule, and from that schedule I obtained each game's ID. These IDs, as well as the year associated with them, were stored in a dictionary.
2) Beginning with the first game of 2013, I called an API request for the summary statistics for every game from 2013 to 2017. Since the json_normalize function tends to fail when the JSON contains null values, I decided to replace all null values with "0," using json.dumps and json.loads. Once the nulls were replaced, I used json_normalize to

create a pandas DataFrame from the first game, and then concatenated it with each subsequent game's DataFrame.
3) I exported this combined DataFrame as a CSV, just in case the connection to Jupyter was lost at some point during my analysis.

Cleaning the data:

1) Of the original columns obtained using this method, I decided to keep only the game ID, home/away rank, and home/away points columns, along with all of the summary statistic columns.
2) Upon exploring the DataFrame, consisting of almost 6200 games, I noticed that many of the remaining columns contained less than half of the data, including the "rank" columns. With 151 columns at this point, I found it reasonable to simply exclude these columns. I restricted the DataFrame to only those columns with 6000+ non-null entries.
3) Of the remaining 75 columns, there were several which I did not find particularly interesting in using as predictive features, such as the number of technical fouls, or total minutes. I dropped these columns from the DataFrame, leaving me with 55 columns.
4) I saw that the remaining columns contained almost no missing values. Rather than imputing these values, I decided to restrict the DataFrame to only those rows which contained no null values. This left me with 55 columns, and almost 6100 records.
5) Since the occurrence of a "0" in any of the statistical categories is highly unlikely, I eliminated any entries where this was the case. This left approximately 5800 entries.
6) Finally, the only other abnormality in the data that could be noted was in the two_points_pct columns for both the home and away teams. Some of the values were listed as decimals (<1), while others were listed as percents (<100). I altered these columns so that all of the values were listed as percents. This should also remedy the problem of having to later scale this column during the machine learning portion of the project.

New columns:

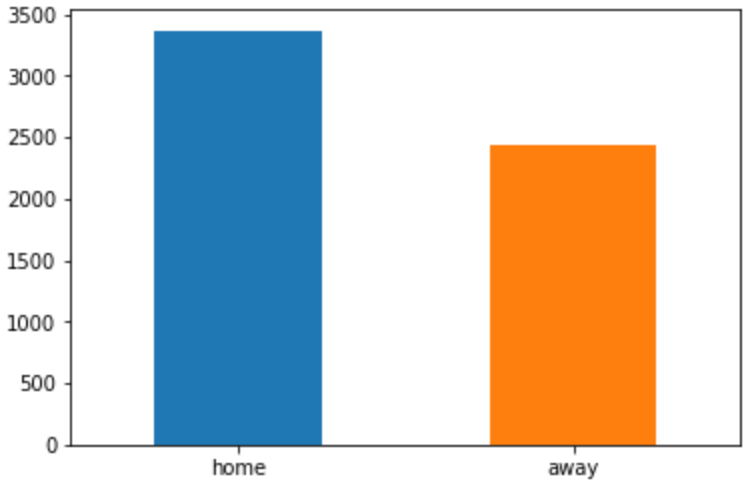
For my analysis, I have decided to created two new columns. The "winner" column will denote which team, home or away, won the game. This column will later be used as our target variable.

The "score_diff" column will denote the difference in total points between the two teams (home minus away).

3) Exploratory Data Analysis Findings

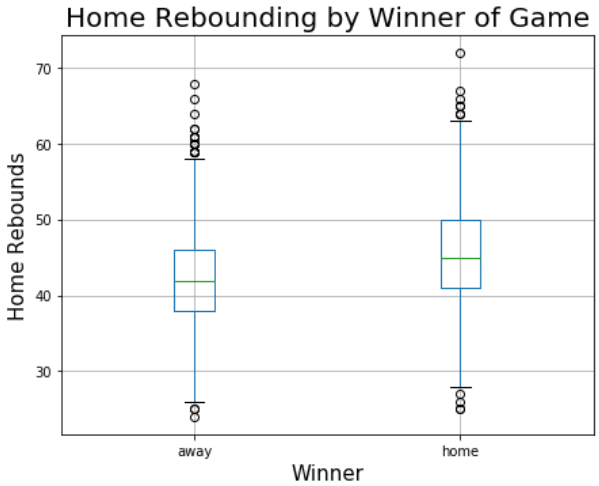
Before we fit our machine learning models, we are interested in exploring any trends in the data to determine which features, offensive or defensive, might be good predictors. Here are some interesting trends I was able to explore over the course of my analysis:

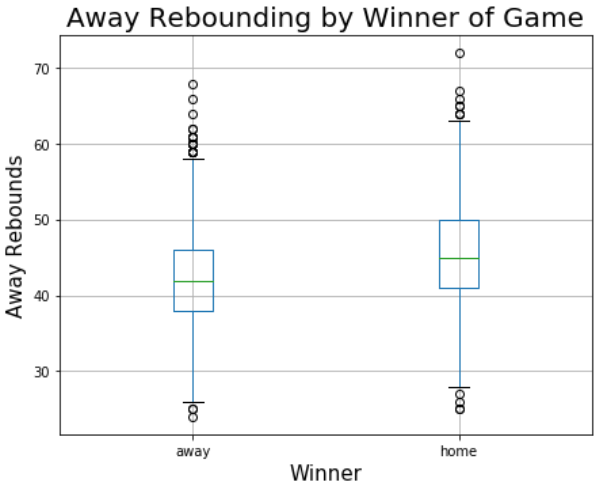
1) Who tends to win more, the home team or the away team?

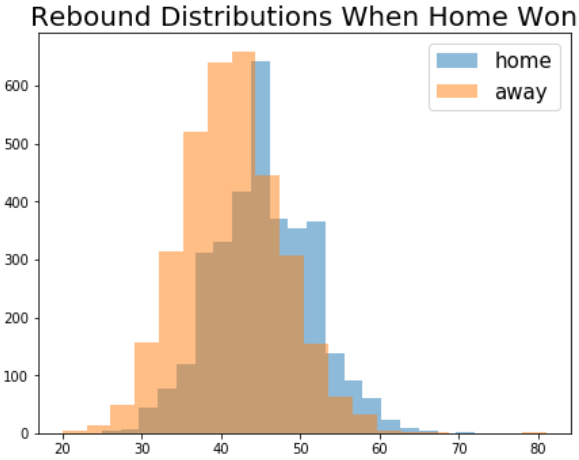


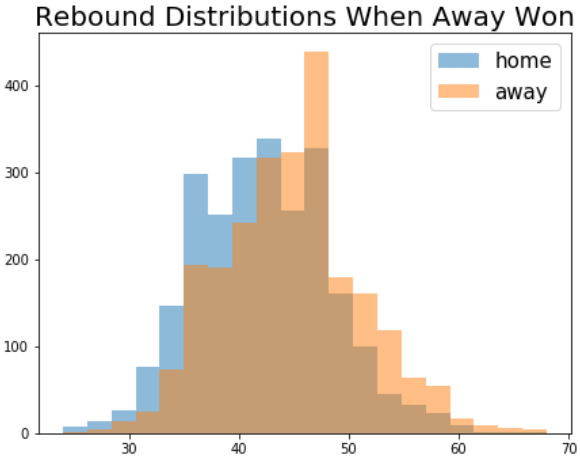
Of the almost 6000 games, approximately 58% were won by the home team. In fact, if we examine the “score_diff” column for games where home won versus games where away won, see that the the average home win is about 12 points, while the average away win is about 10 points. Not only does home tend to win, but when they do win, they tend to win by more.

2) What effect does rebounding have on who wins the game?









From the top two plots, we see that when a team wins, they tend to rebound more than when they lose. When home won, they tended to have more rebounds, and when away won, they tended to have more rebounds.

The bottom two plots are histograms that reflect this trend. The bottom left chart, representing games where home was the winner, shows that the distribution of home rebounds is distinguishably shifted right from the distribution of away rebounds. The bottom right chart, where away won, shows that the away team tended to rebound more during these games.

Here is a table exploring the correlations between rebounds and points scored for each team:

	Home Rebounds	**Away Rebounds**
Home Points	*0.11*	*-0.23*
Away Points	*-0.25*	*0.11*

For both the away and home teams, we see that as the number of rebounds collected by one team increases, the number of points scored by the other team tends to decrease. This makes sense, since most of the time when one team fails to score, the other team picks up the rebound.

What about offensive rebounding in particular? Offensive rebounds allow a team the opportunity for second chance points after missing the first shot attempt. Examining the correlations between home offensive rebounds and all other numerical columns, we see some expected trends and some unexpected. Here are three of the strongest correlations:

- *Home offensive rebounds vs. home field goals attempted → r = 0.50*

This makes sense, since an offensive rebound gives the team the opportunity for another shot attempt. However, the following correlations are somewhat surprising:

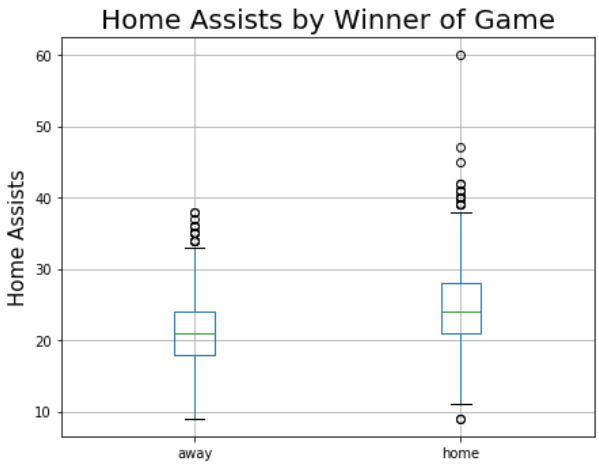
- *Home offensive rebounds vs. home points → r = -0.02*

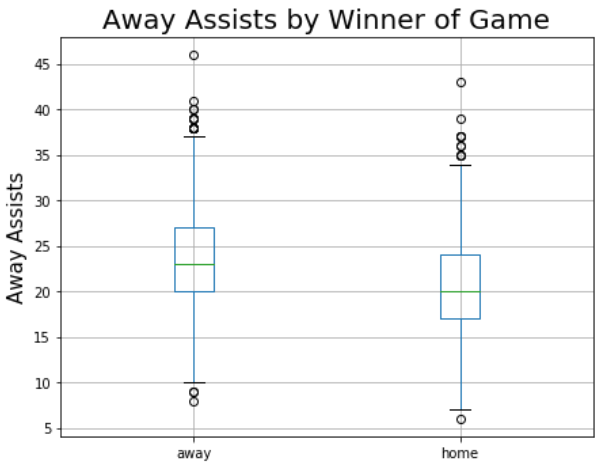
Not only do offensive rebounds have almost no correlation with the number of points a team accrues, but they actually have a somewhat opposite association, as shown by the following correlation:

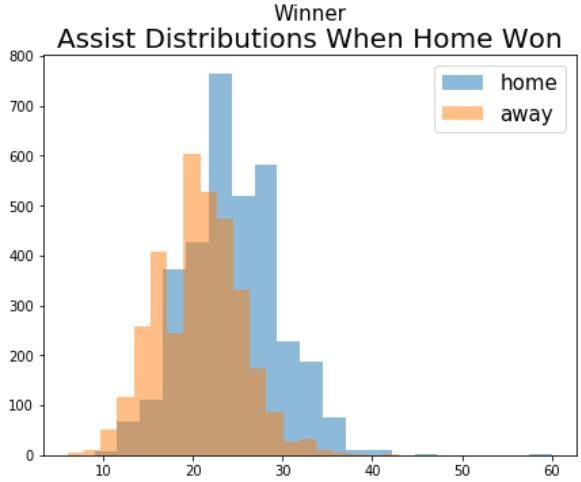
- *Home offensive rebounds vs. home field goal percentage → r = -0.36*

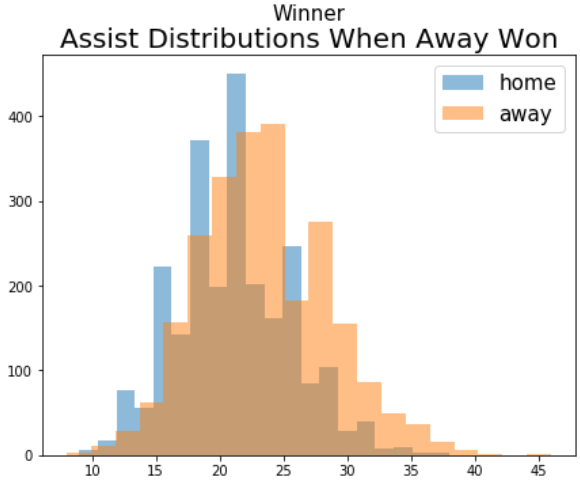
This implies that most of the time, when a team has the opportunity for another shot attempt, they do NOT make the shot!

3) What effect do assists have on who wins the game?









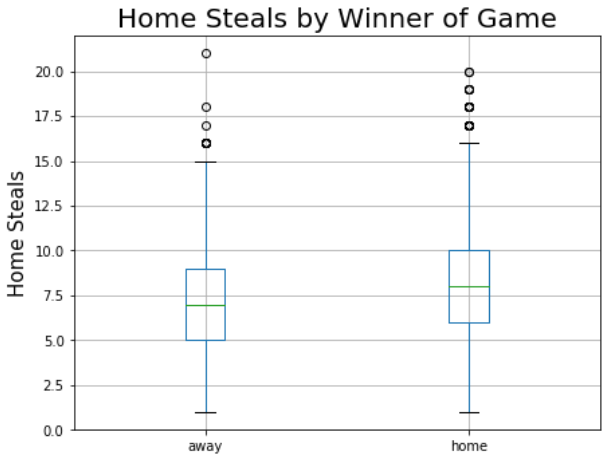
Assists lead to more wins. From the top left plot, we see that when home wins, they tend to have significantly more assists than when they lose. From then top right plot, we see that when away wins, they tend to have significantly more assists.

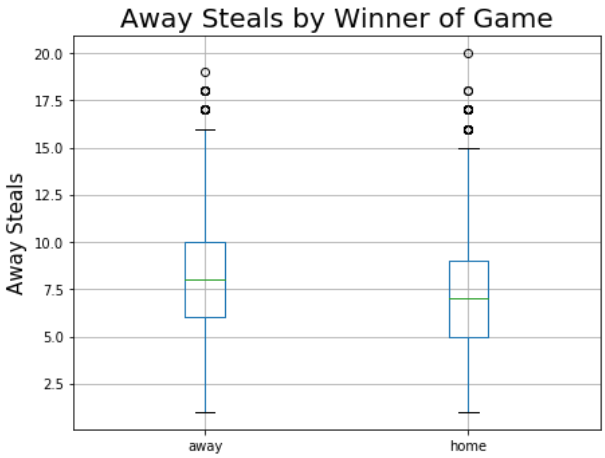
Here is a similar table as the one displayed above, but for assists instead of rebounds:

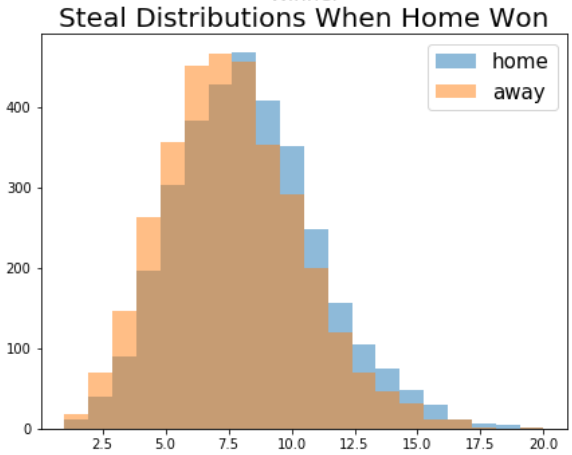
	Home Assists	**Away Assists**
Home Points	*0.58*	*0.16*
Away Points	*0.11*	*0.56*

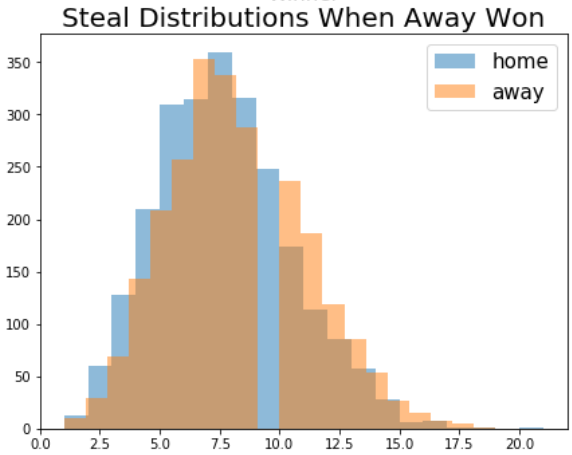
There is a strong correlation between the number of assists a team accrues and the number of points they score. This is fairly expected, since every assist results in points for that team. It is possible that some teams that tend to play isolation, one-on-one basketball tend to win as well, but these plots and correlations show that overall, assisted scoring is valuable in winning. Assists will lead to not only more wins, but more points.

4) What effect do steals have on who wins the game?









Although the histogram plots for steals when home wins versus when away wins are not markedly different, it is clear from the boxplots that when a team wins, they tend to have more steals than when they lose.

Here are the correlations for steals versus points:

	Home Steals	**Away Steals**
Home Points	*0.14*	*-0.05*
Away Points	*-0.08*	*0.10*

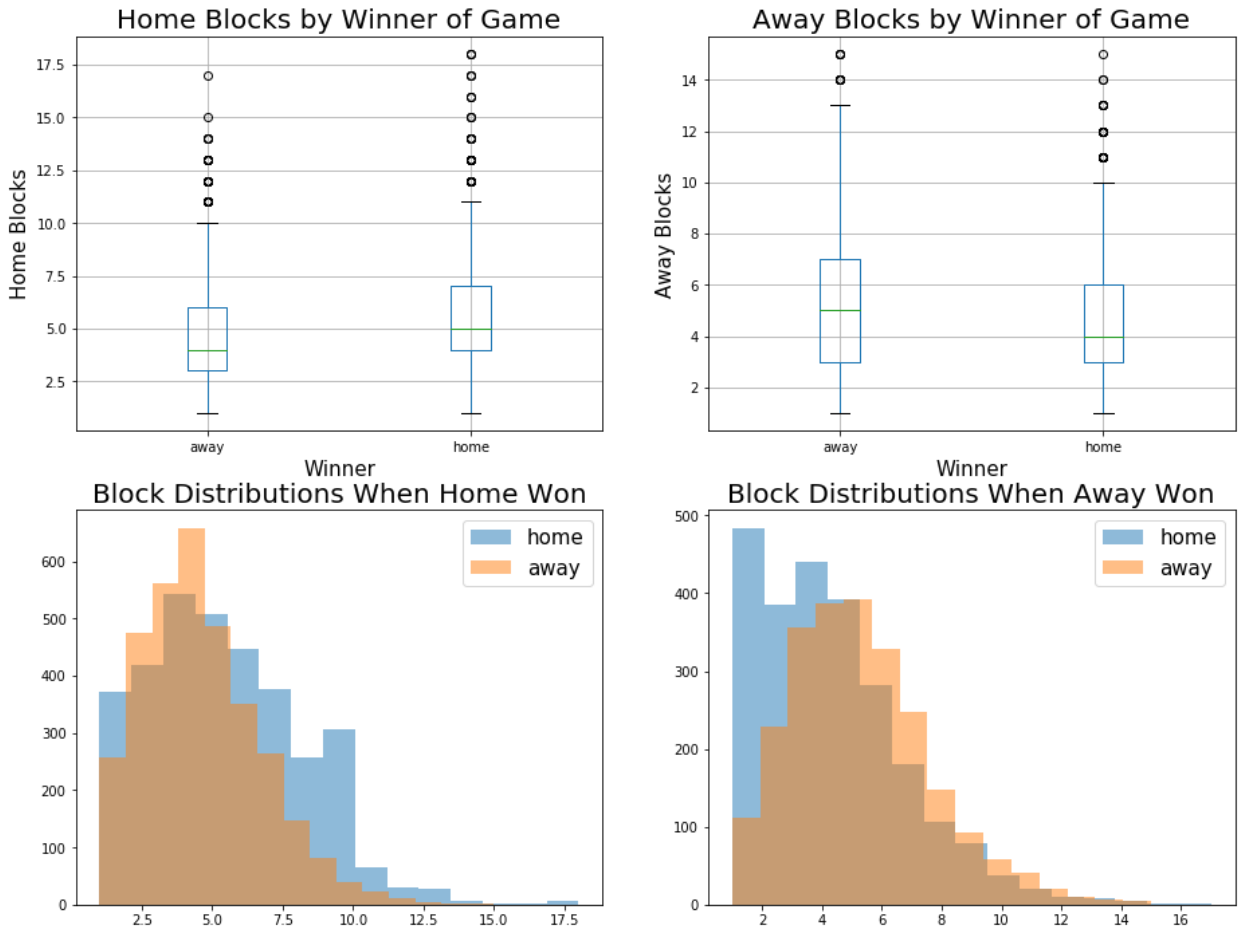
None of these correlations are particularly statistically significant. The strongest correlations show that when a team accrues more steals, they tend to score more points. In particular, we might expect that steals lead to more fast break points, since the defense is often

out of position once they have the ball stolen from them. The following two correlations support this notion:

- *Home steals vs. home fast break points → r = 0.30*
- *Away steals vs. away fast break points → r = 0.31*

While steals overall do not necessarily strongly correlate with more points for one team and less points for the other overall, they do tend to provide more fast break points, which are valuable in winning the game.

5) What effect do blocks have on who wins the game?



The boxplots clearly show that when a team wins, they tend to have more blocks than when they lose. We might expect that more blocks, which are the result of a denied shot attempt, also results in fewer points for the opponent. Let's examine the blocks vs. points correlations:

	Home Blocks	Away Blocks
Home Points	*0.06*	*-0.14*
Away Points	*-0.14*	*0.06*

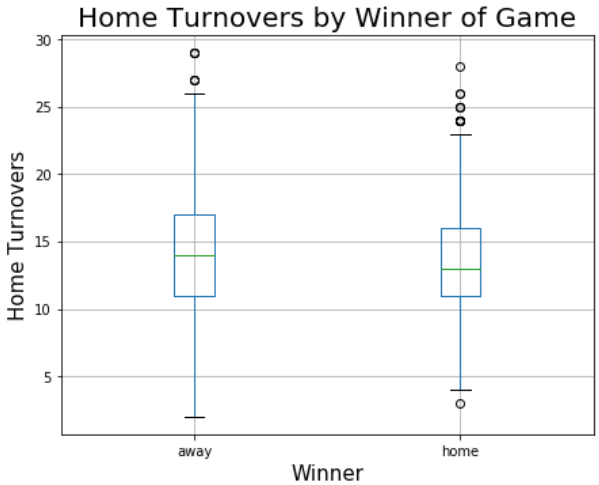
Although the correlations are not particularly strong, we do see that as the number of blocks of one team increases, the number of points of the other team decreases.

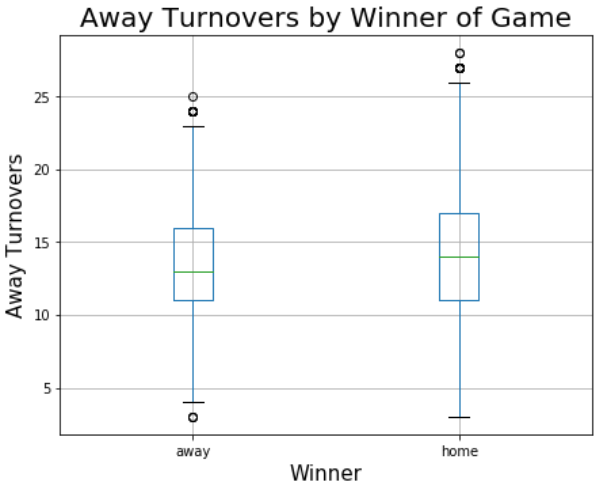
Shot blocking is an important deterrent in the opposing team's scoring. In particular, we would expect that with more shots blocked, the opposing team's field goal percentage decreases. The following correlations support this idea:

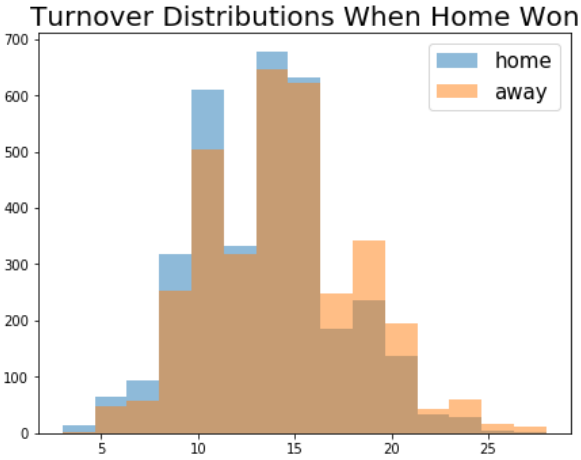
- *Home blocks vs. away field goal percentage → -0.32*
- *Away blocks vs. home field goal percentage → -0.33*

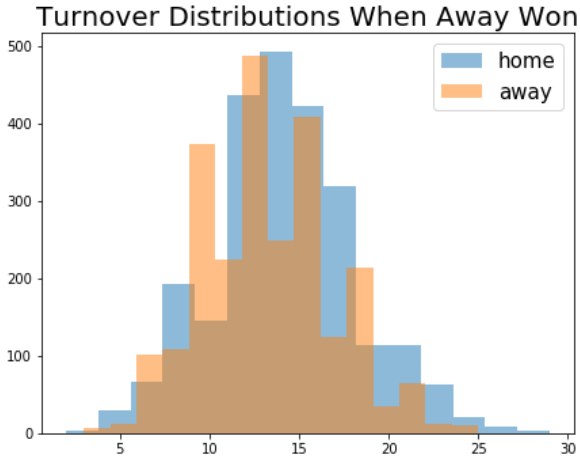
More blocked shots implies higher defensive efficacy, which is an important factor in winning basketball games.

6) What effect do turnovers have on who wins the game?









The boxplots above show another somewhat expect trend: a team accrues fewer turnovers when they win than when they lose. This implies higher offensive efficiency. Here are the correlations between turnovers and points.

	Home Turnovers	**Away Turnovers**
Home Points	*-0.08*	*0.05*
Away Points	*0.03*	*-0.11*

None of the correlations between turnovers and points are particularly strong, but we do see that as a team accrues more turnovers, they tend to score fewer points. With more turnovers, there are fewer chances for a team to attempt a field goal resulting in points.

Turnovers, when resulting in steals for the opposing team, also allow the opposing team to get out in transition for the opportunity at fast break points. Fast break points tend to be “easy” points, since there is much less opposing defense. This expectation is supported by the following correlations:

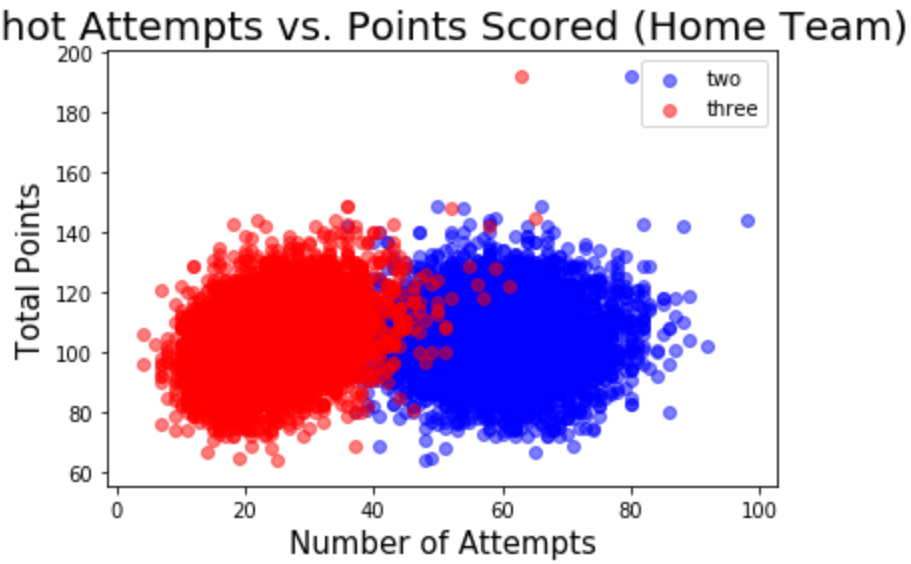
- *Home turnovers vs. away steals → r = 0.76*
- *Home turnovers vs. away fast break points → r = 0.22*

- *Away turnovers vs. home steals → r = 0.75*
- *Away turnovers vs. home fast break points → r = 0.21*

Turnovers tend to lead to more steals and fast break points for the opposing team.

7) What is more important in securing wins: two point attempts or three point attempts?

We are interested in exploring which offensive focus is more effective: mid-range / “down_low” offense (two-pointers), or perimeter three-point shooting (three-pointers).

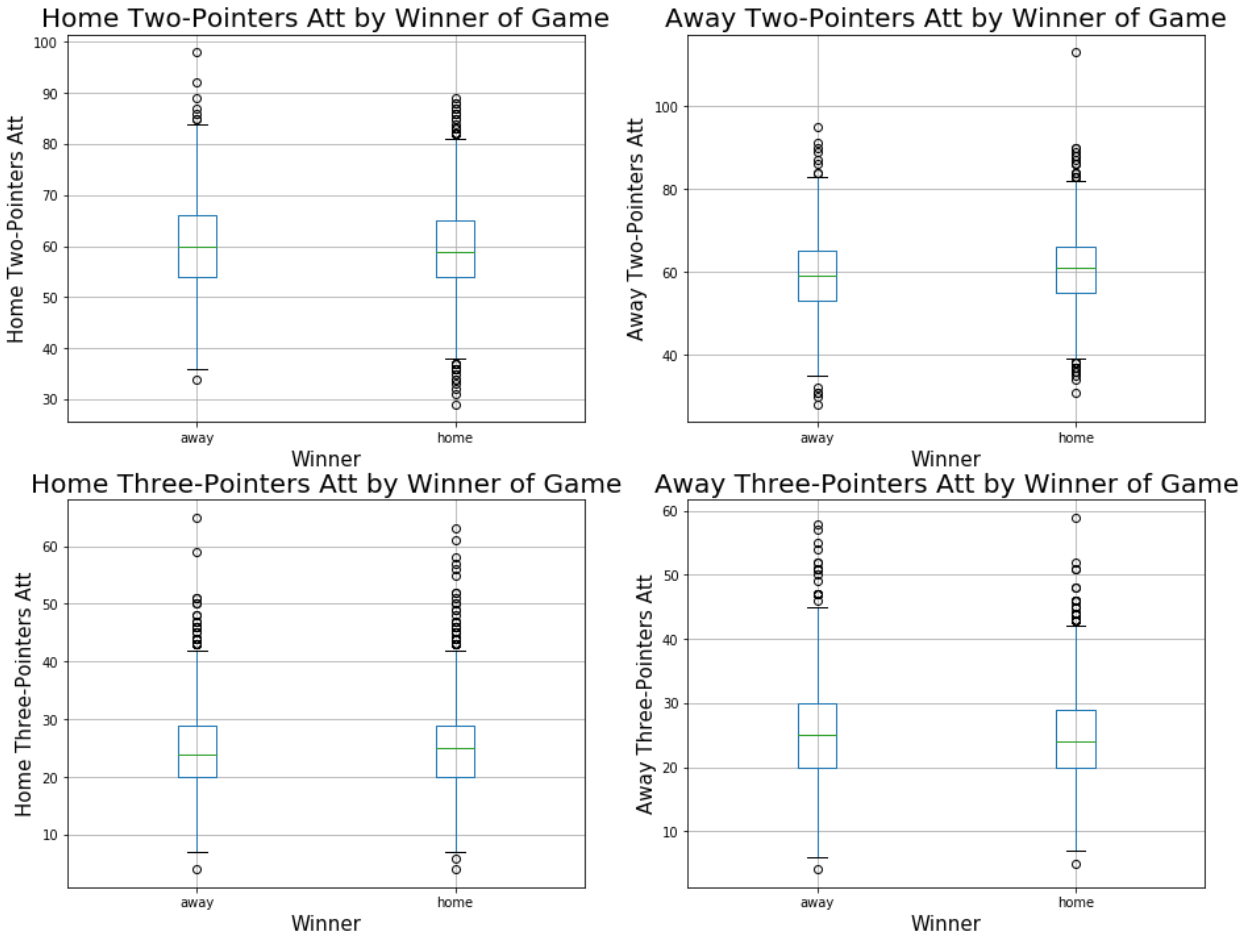


It is difficult to tell which is more strongly correlated with points based on this scatterplot. The only thing we see for sure is that there tend to be many more two-pointers attempted than three-pointers. let us examine the correlations:

- *Home two pointers attempted vs. home points → r = 0.02*
- *Away two pointers attempted vs. home points → r = 0.01*

- *Home three pointers attempted vs. home points → r = 0.29*
- *Away three pointers attempted vs. home points → r = 0.30*

We see here that three-pointers attempted are much more strongly correlated with the total number of points scored for a given team. However, this might simply be because three-pointers are worth more. Do winning teams tend to take many more twos or many more threes?



Interestingly enough, the top two plots show that when a team wins, they actually tend to shoot FEWER two-pointers on average than when they lose! By contrast, the bottom two plots show that when a team wins, on average, they tend to shoot more three-pointers than when they lose.

4) In-Depth Analysis: Machine Learning

The problem posed asks up which factors are a better predictor of NBA wins: offensive statistics or defensive statistics. With this goal in mind, we will create three machine learning models using only the offensive statistics analyzed above, and three using only the defensive statistics.

We will fit several machine learning algorithms to a portion of our data, the training set, and use the model to predict the winner of the game on a holdout set (test set). Here are our models of interest:

- Random Forest
- Logistic Regression
- Support Vector Classifier

1) I will be using, for both the home and away teams, the seven columns described in the EDA section:
    a) Assists
    b) Blocks
    c) Rebounds (broken out into defensive and offensive rebounds)
    d) Steals
    e) Three-pointers attempted
    f) Two-pointers attempted
    g) Turnovers
2) I have also decided to include free-throws attempted, as well as personal fouls. Being able to draw fouls and get to the free-throw line can be a huge advantage for winning teams. In a similar vein, if a team gets into foul trouble and sends their opponent to the line too much, they might be less likely to win.
3) I decided to exclude several statistical columns that explicitly describe points being scored, such as field goals made, and points in paint. Knowledge of these metrics provides an unfair and unrealistic advantage for the model in predicting who wins each game. I kept only metrics that are correlated with wins and points being scored, without measuring those points being scored directly.
4) Our offense feature space, X_train_off and X_test_off, will consist of 12 total offense-related columns, 6 for each team:
    a) Assists
    b) Offensive rebounds
    c) Three-pointers attempted
    d) Two-pointers attempted
    e) Free-throws attempted
    f) Turnovers
5) Our defense feature space, X_train_def and X_test_def, will consist of 8 total defense-related columns, 4 for each team:
    a) Blocks
    b) Steals
    c) Defensive rebounds
    d) Personal fouls
6) Our target variable, y, is a binary variable, with 1 denoting a home win, and 0 denoting an away win.
7) I split the data into training and test sets (30% split), and then split the “X” further into the offense/defense feature spaces described above.
8) I then used GridSearchCV to train and fit each of the three models, tuning the following hyperparameters:
    a) For the random forest, the n_estimators parameter.

b) For the logistic regression, the C parameter.
c) For the SVM, the C, kernel, and gamma parameters.

Summary of Machine Learning Findings

Here is a table of the accuracy and F1 scores for the six models:

	Offensive model			**Defensive model**		
Model	*Accuracy*	*Away F1 Score*	*Home F1 Score*	*Accuracy*	*Away F1 Score*	*Home F1 Score*
Random Forest	74.08%	0.67	0.78	78.79%	0.74	0.82
Logistic Regression	75.98%	0.70	0.80	81.03%	0.77	0.84
Support Vector Classifier	76.15%	0.70	0.80	80.92%	0.77	0.84

For all three models, the defense-focused version always had better accuracy, away F1 score, and home F1 score! Overall, the models were better at predicting home victories. However, the disparity between the F1 scores in the defensive models was less.

Overall, the best results were achieved using a logistic regression classifier using defense-related features. This model returned an accuracy of 81%, and average F1 scores of 0.81

6) Conclusion: Final Thoughts and Further Research

We were interested in determining which kind of statistical features, offensive or defensive, were the best predictors in NBA basketball wins. We found that the home team had a 58% advantage in winning. In addition, we discovered that when a team wins, whether home or away, they tend to rebound the ball more, accrue more assists, more steals, more blocks, fewer turnovers, and shoot more three pointers on average.

After exploring these trends, we fit three of the same models to two different sets of feature data: offensive features and defensive features. We found that for all three models--Random Forest, Logistic Regression, and Support Vector Machine--the one using the defensive features performed better. Of the six models, the best model overall was a Logistic

Regression classifier using defense-related features, with an accuracy of 81% and average F1 score of 0.80.

What else might be interesting to explore/include?

1) Can we alter our model to predict a different target, such as the continuous variable of total points scored for each team?
2) How could we incorporate more detailed game-level analysis to improve our models, such as the location of shot attempts, or quarter-by-quarter statistical analysis?
3) Could the knowledge of which particular teams are involved in each game help us predict the winner? What about the knowledge of particular players on those teams?