Predicting NBA Winners

The Problem

- What are more important in predicting NBA wins: offensive statistics or defensive statistics?

Who is interested?

- Sports analysts/broadcasters
  - More accurate predictions
  - More detailed analysis provides greater entertainment
- NBA management
  - Detailed analysis of team strengths/weaknesses
  - Directed focus of team development
  - Preparation for match-ups against particular teams

The Data

- Obtained using SportRadar API
- Acquired game schedule of all games from 2013-2017
- Used game schedule to acquire game summary statistics for each game into combined pandas DataFrame
  - One row per game
- Statistics include:
  - Points, rebounds, assists, turnovers, etc.

Data cleaning

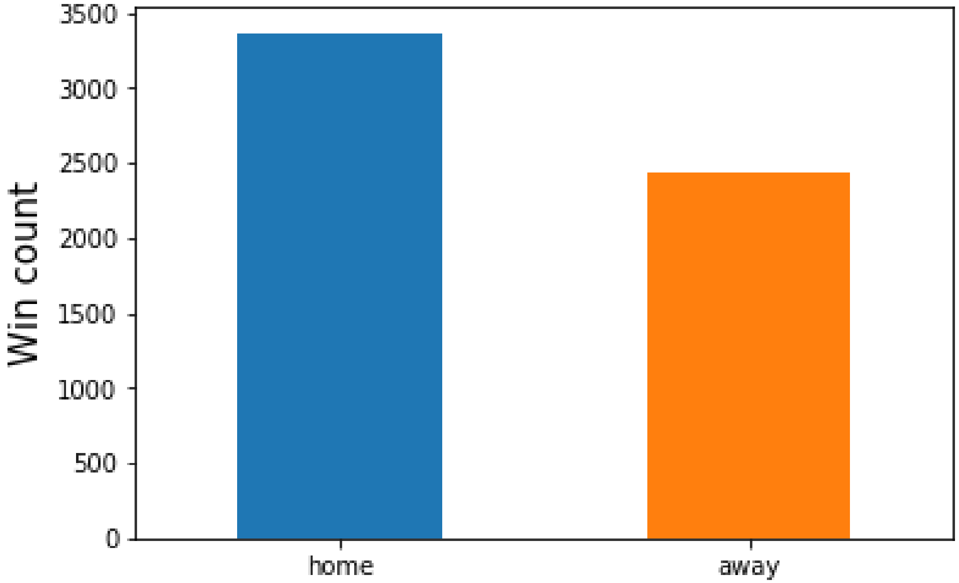
1) Keep only game ID, points, rank, and statistics columns for both teams
2) Eliminate columns with less than 6000 non-null rows
3) Remove unwanted features, such as technical fouls and minutes played
    1) Also remove repeated “points” columns
4) Remaining columns contained very few missing values,so instead of imputing the missing values, restrict the DataFrame to rows with all non-null entries
5) Remove rows containing 0’s (very unlikely in basketball statistics)
6) Convert two_points_pct columns to percents instead of decimals

Two new columns

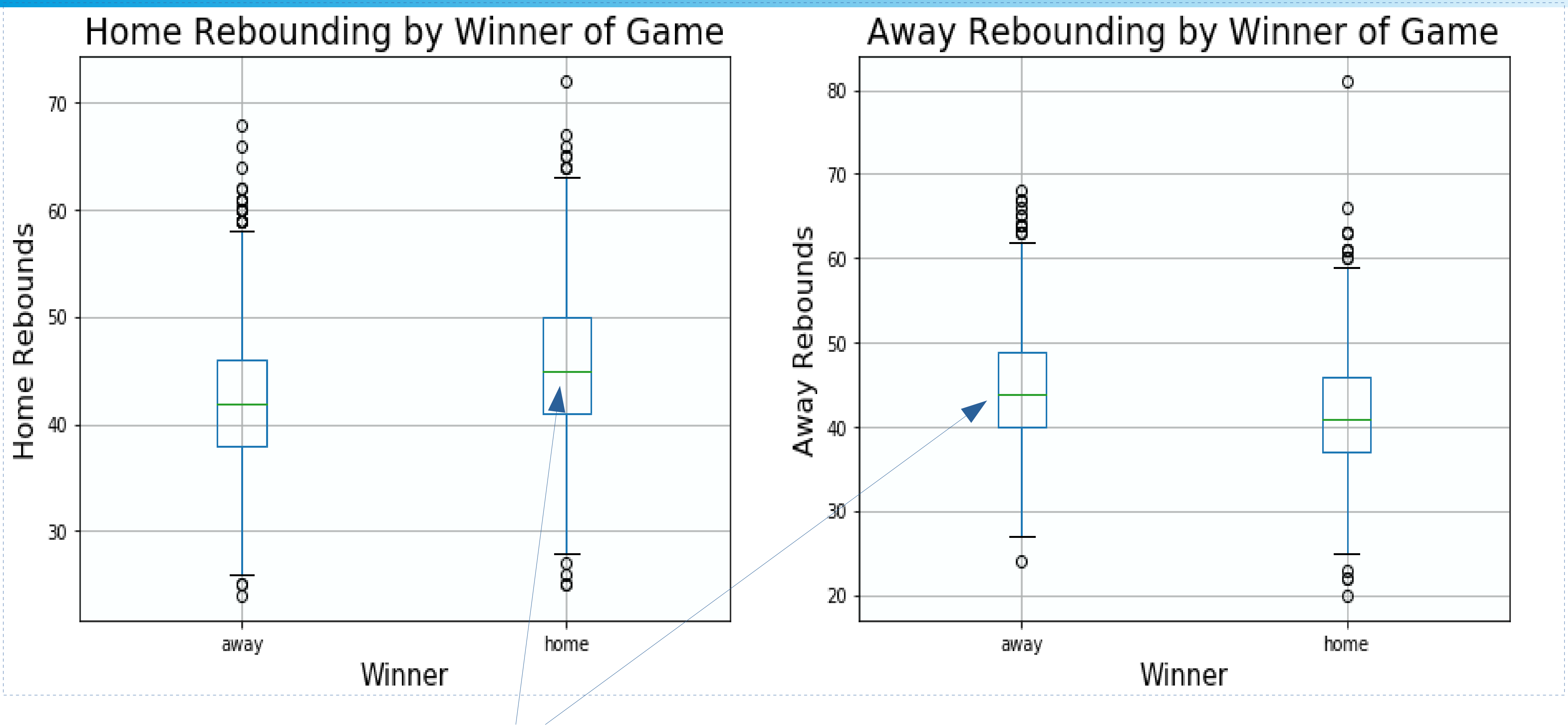
- Create “winner” column, based on which team finished with more points
  - This will later be used as our target variable
- Create “score_diff” column
  - Home points – Away points

Who wins more, home or away?

- Home team wins 58% of the time
  - Avg home win → 12 more pts
  - Avg away win → 10 more pts
- The home crowd tends to play a factor in team morale, which can affect team performance.



Rebounds EDA



Home Rebounding by Winner of Game

Away Rebounding by Winner of Game

- A team is more likely to accrue more rebounds when they win than when they lose.

Rebounds vs. Points

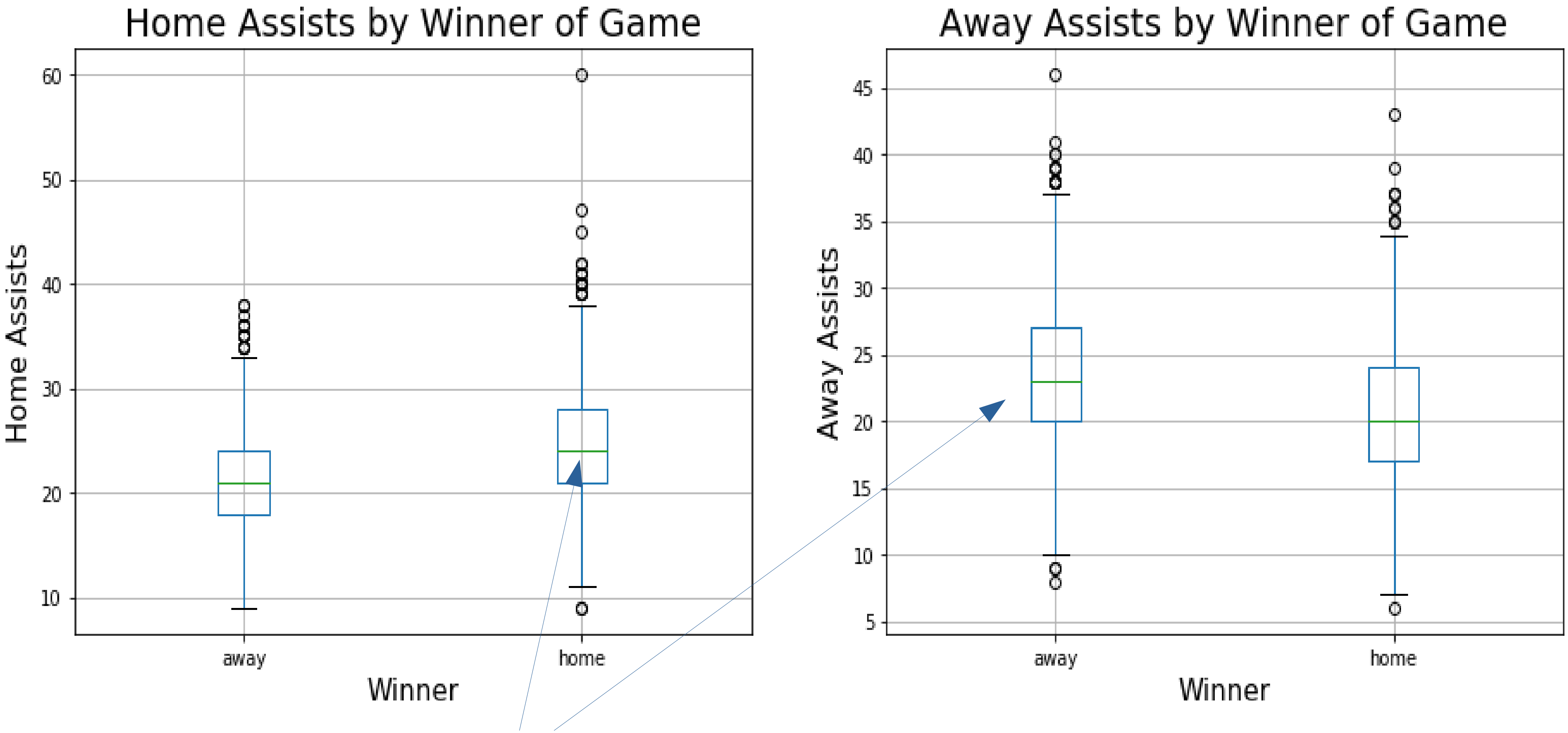
	Home Rebounds	Away Rebounds
Home Points	0.11	-0.23
Away Points	-0.25	0.11

- 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.

Home Offensive Rebound analysis

- Off. Rebounds vs. field goals attempted
  - r = 0.50
  - This makes sense, since an offensive rebound gives the team the opportunity for another shot attempt
- Off. Rebounds vs. points scored
  - r = -0.02
  - Surprisingly, offensive rebounds seem to have almost no correlation with the number of points a team accrues
- Off. Rebounds vs. 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!

Assists EDA



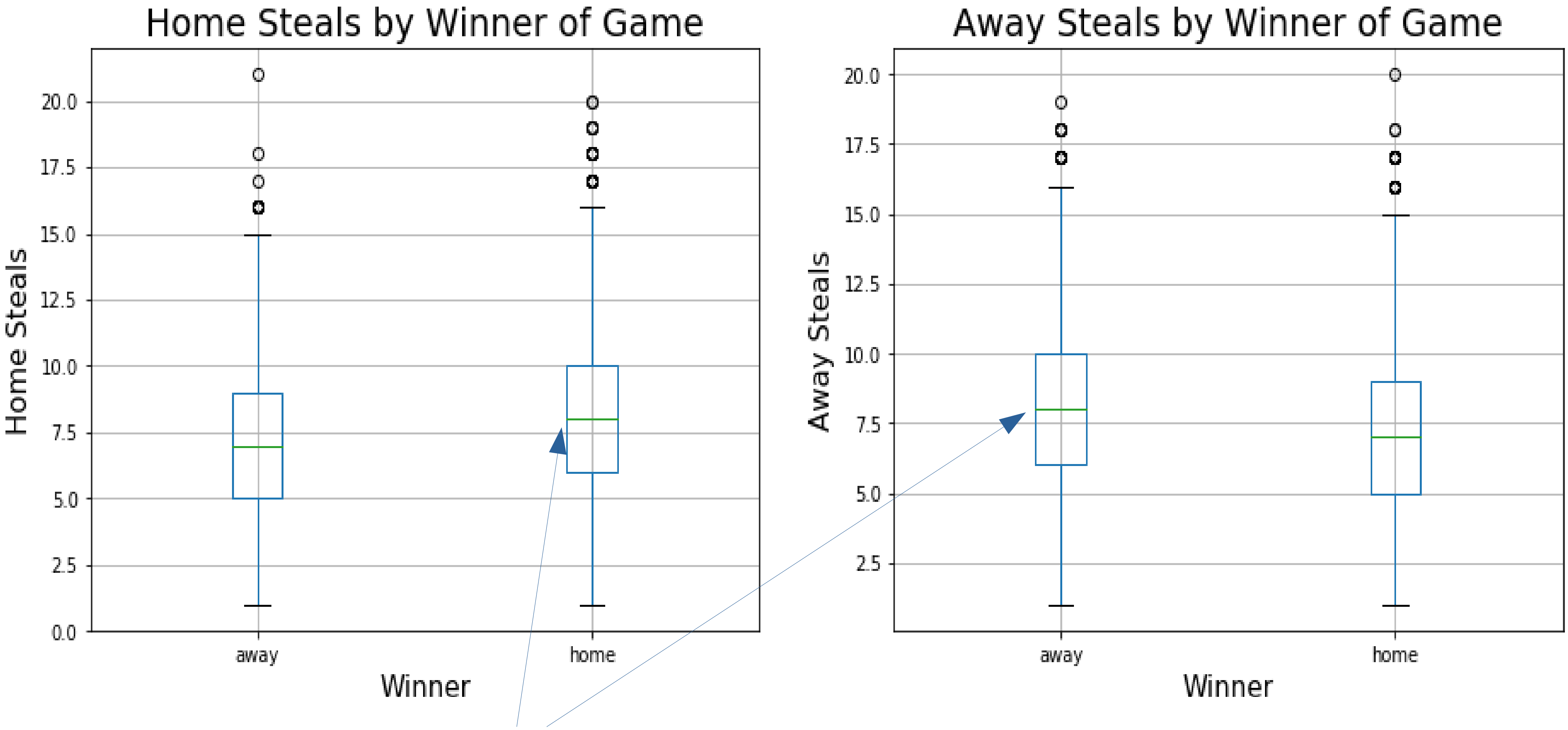
- A team is more likely to accrue more assists when they win than when they lose.

Assists vs. Points

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

- 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.

Steals EDA



- A team is more likely to accrue more steals when they win than when they lose.

Steals vs. Points

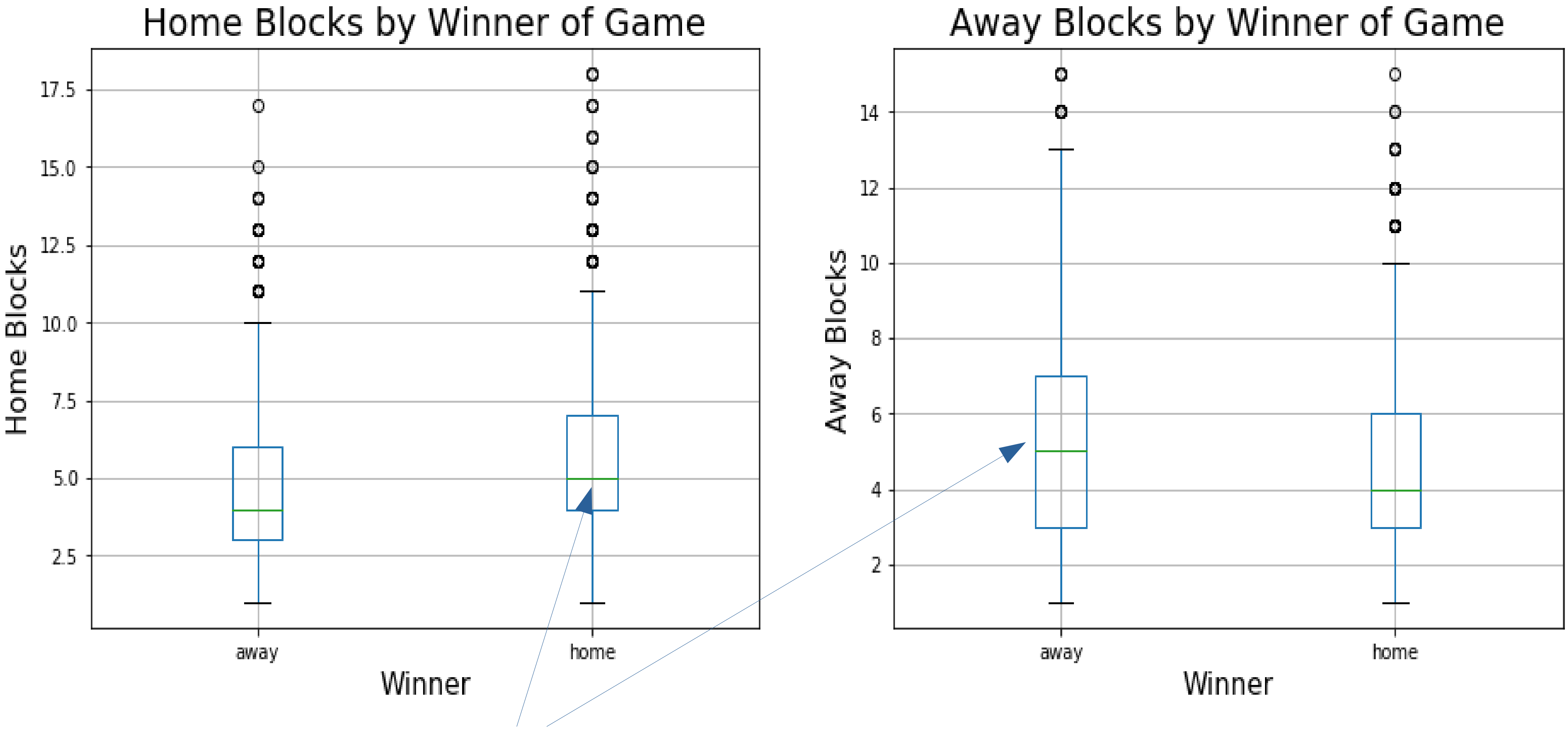
	Home Steals	Away Steals
Home Points	0.14	-0.05
Away Points	-0.08	0.10

- 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.

Further steals analysis

- 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.
  - 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 games.

Blocks EDA



- A team is more likely to accrue more blocks when they win than when they lose.

Blocks vs. Points

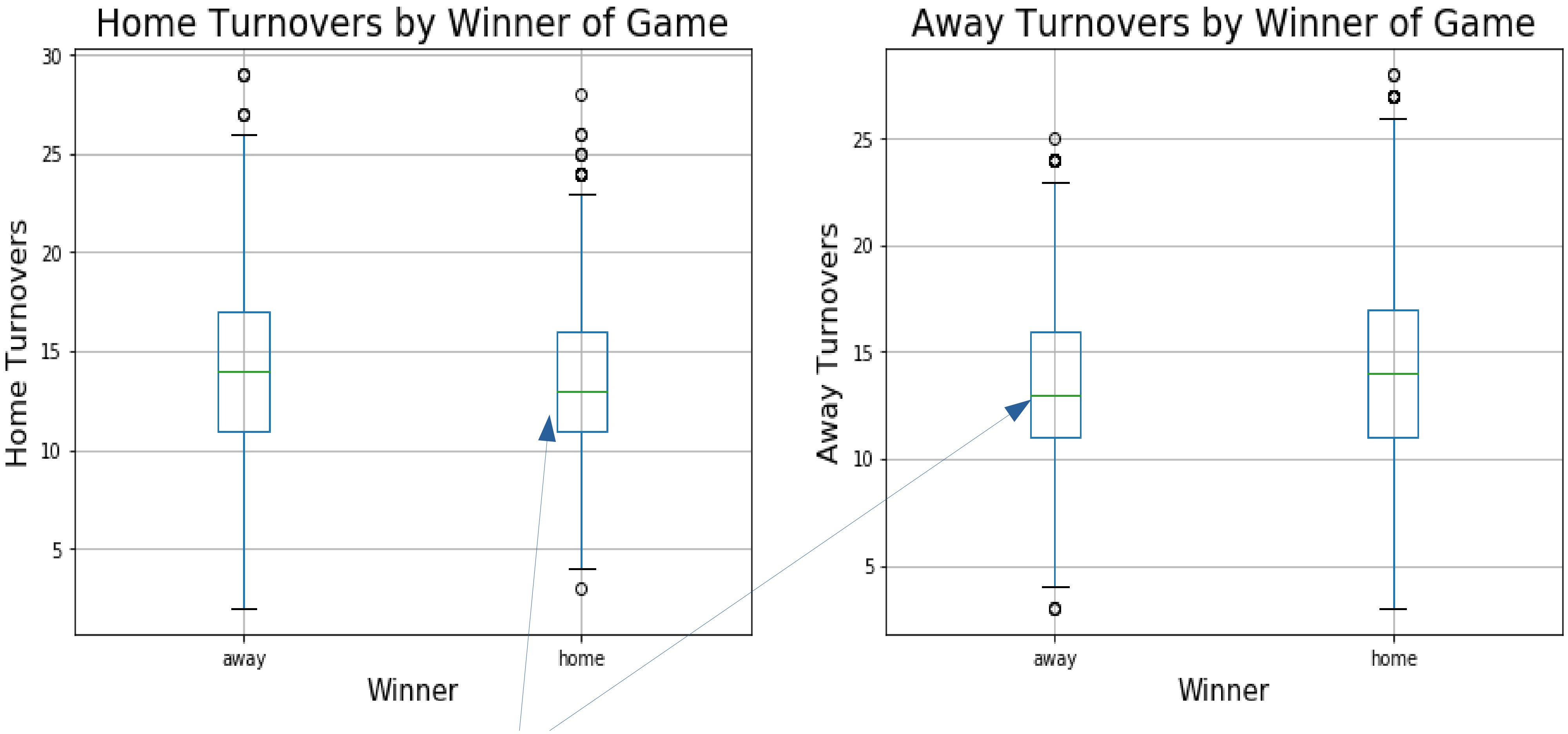
	Home Blocks	Away Blocks
Home Points	0.06	-0.14
Away Points	-0.14	0.06

- 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.

Further blocks analysis

- 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.

Turnovers EDA



- A team is more likely to accrue fewer turnovers when they win than when they lose.

Turnovers vs. Points

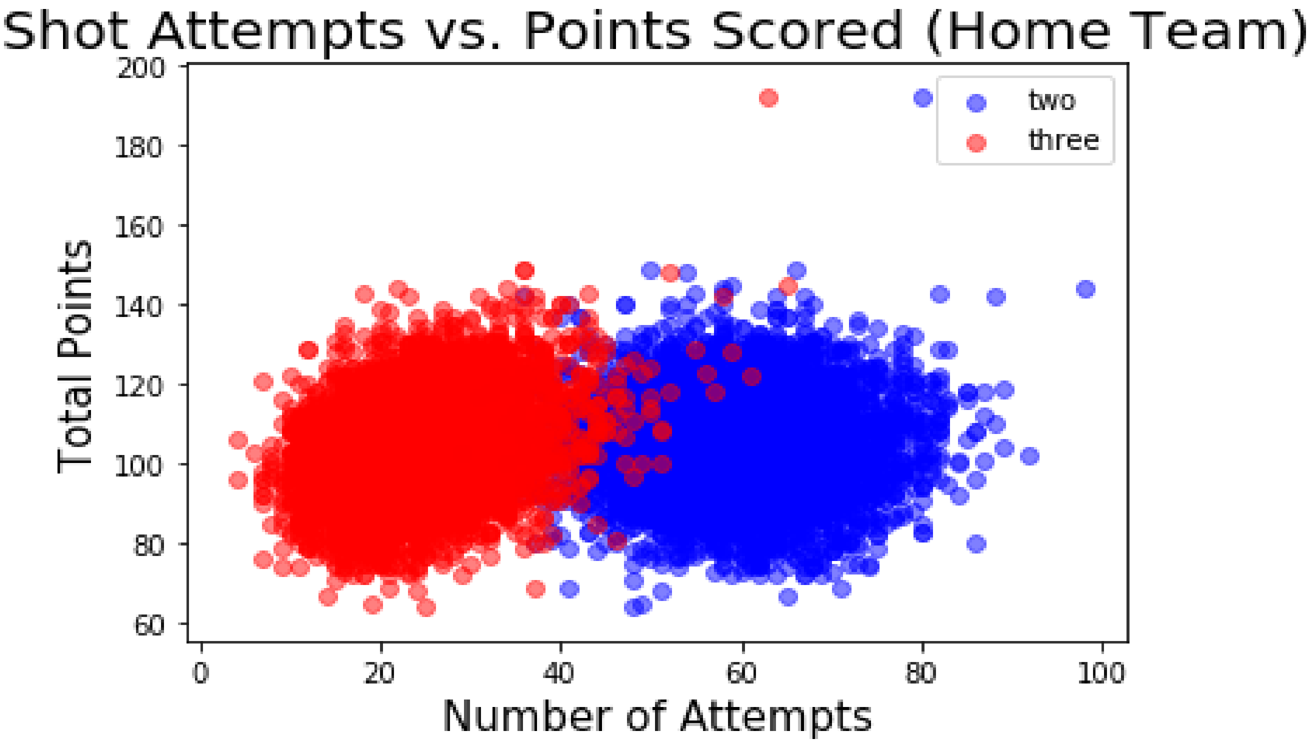
	Home Turnovers	Away Turnovers
Home Points	-0.08	0.05
Away Points	0.03	-0.11

- 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.

Further turnovers analysis

- 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.

Two-pointers vs. three-pointers

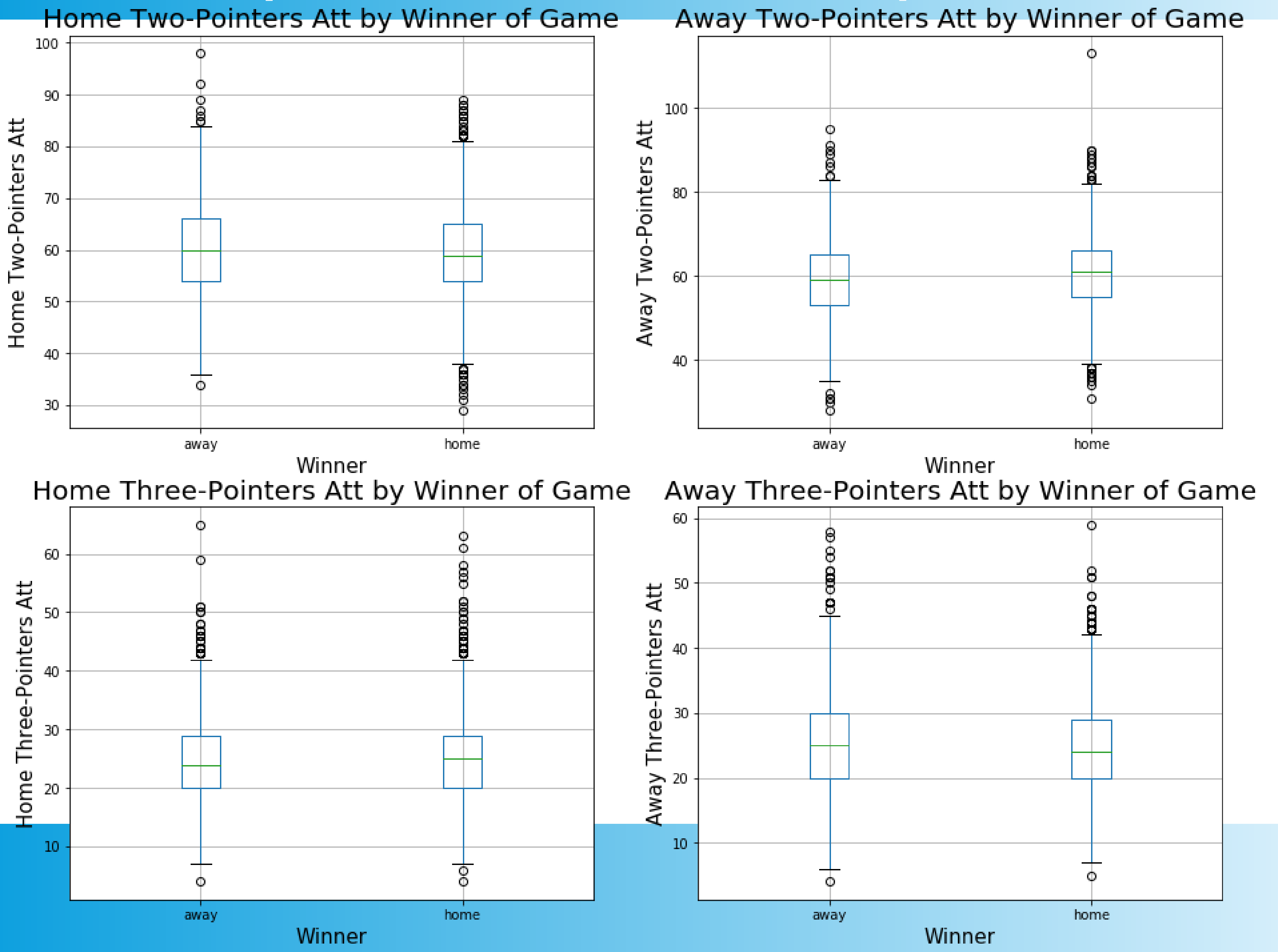


- It is difficult to tell which is more strongly correlated with points, two-pointers or three-pointers, based on this scatterplot.
- The only thing we see for sure is that there tend to be many more two-pointers attempted than three-pointers.

Two-pointers vs. three-pointers

- Let’s examine how two and three-pointers correlate with points scored:
  - Home two pointers attempted vs. home points → r = 0.02
  - Away two pointers attempted vs. away points → r = 0.01
  - Home three pointers attempted vs. home points → r = 0.29
  - Away three pointers attempted vs. away 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?

Two-pointers vs. three-pointers



Two-pointers vs. three-pointers

- Interestingly enough, the top two plots show that when a team wins, they actually tend to shoot FEWER two-pointers on average!
- 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.

Predicting the winner

- It is time to create two kinds of model to predict the winner of the match. Which features will perform better? The offensive statistics or defensive statistics?
- Three candidate models:
  - Random Forest
  - Logistic Regression
  - Support Vector Classifier
- Each model will be fit using either just the offense-related features or just the defense-related features, for a total of six models

Candidate features

- I will be using, for both the home and away teams, the seven columns described in the EDA section:
    - Assists, Blocks, Rebounds (broken out into defensive and offensive rebounds), Steals, Three-pointers attempted, Two-pointers attempted, and Turnovers
- 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.

Additional note on candidate features

- 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.

Offense features

- Our offense feature space, X_train_off and X_test_off, will consist of 12 total offense-related columns, 6 for each team:
    - Assists
    - Offensive rebounds
    - Three-pointers attempted
    - Two-pointers attempted
    - Free-throws attempted
    - Turnovers

Defense features

- Our defense feature space, X_train_def and X_test_def, will consist of 8 total defense-related columns, 4 for each team:
  - Blocks
  - Steals
  - Defensive rebounds
  - Personal fouls

Model Comparison

	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 performed better!
- Overall, the models were better at predicting home victories. However, this disparity was less severe for the defensive models.
- Overall, the best results were achieved using a logistic regression classifier with defense-related features.

Summary

- Which features, offensive or defensive, are the best predictors in NBA basketball wins?
  - 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.
- Three of the same models using 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.

Further research

- What else can we explore?
  - Can we alter our model to predict a different target, such as the continuous variable of total points scored for each team?
  - 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?
  - 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?