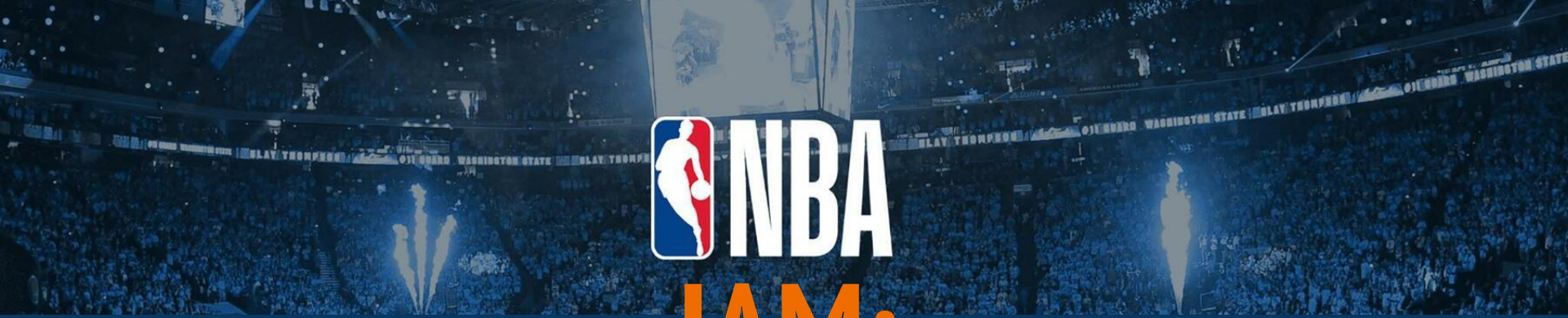


JAM:
just add machine learning

Presented By
Group 6: Lost in Data

Christopher Onubogu
Dan Cabrera
Elsie Pierre-Antoine
Nikoloz Mazmishvili

Using Machine Learning to Predict Game Outcomes

Can we use machine learning to predict which National Basketball Association (NBA) team will win a game?

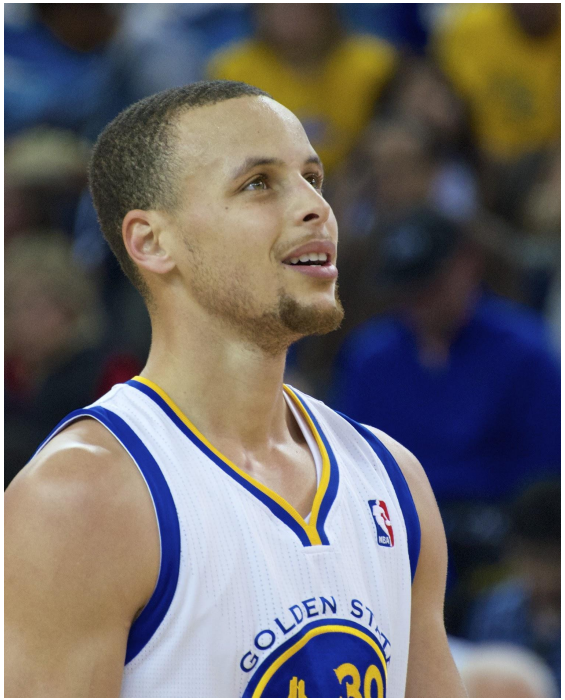
- There are 30 teams in the NBA, and each team plays 82 games in the regular season, with at least 1,230 matchups each regular season.
- There is the potential of an additional 105 games in the postseason.

What makes a winning team? We wanted to see if a machine-learning model could predict game outcomes based on team statistics, such as:

- Points scored at home
- Points scored away
- Free throw percentages
- Field goal (2-pointers and 3-pointers) percentages
- Number of Blocks
- Number of Steals

Source: Basketball Insiders

The NBA Is a Billion-Dollar Business



Steph Curry

With millions of fans, the NBA brings in billions of dollars.

- The NBA's league revenue for the 2022/2023 season was $10.58B
- The average NBA team is worth $3.85B
- NBA players have contracts of more than $200M
  - Steph Curry is the league's most valuable player per year, making more than $51M
  - Nikola Jokic has the largest contract, worth $272M
- Sports betting (on all sports) exceed $10B in 2023
  - Experts estimate it will soon go up to $45B a year

Predicting the outcomes of games could potentially influence the league and its fans in many ways:

- Helping coaches and players know what to focus on in practice and in games
- Owners and agents could use the information to strategically strengthen their positions
- Increase the stakes for online gambling

Sources: Statista, Forbes, SportingNews, Vox

Fundamentals vs. Intangibles

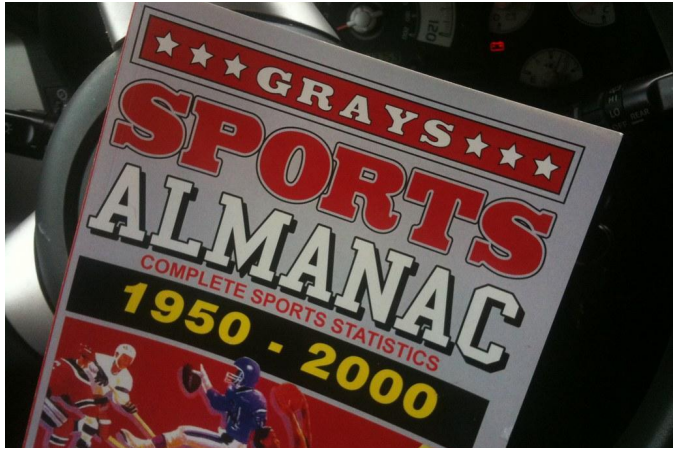
There's no easy way to predict which team will win a game (unless you're a time traveler).

With machine learning, we want to **create a model** and **train it with NBA game statistics** so that it can **accurately predict if a team will win** based on future statistics.

Thankfully, the NBA collects detailed statistics on every game and every player.

For non-NBA fans (and even casual fans), deciphering the stats can be a challenge. On the next slide, we present a glossary of the statistics from our dataset.

There were no stats for "hustle" or "team spirit".



Terminology

ast_away: The total assists by the visiting team

ast_home: The total assists by the home team

blk: Blocks made by the player

fg_pct_away: The field goal percentage of the visiting team

fg_pct_home: The field goal percentage of the home team

fg3_pct_away: The three-point field goal percentage of the visiting team

fg3_pct_home: The three-point field goal percentage of the home team

ft_pct_away: The free throw percentage of the visiting team

ft_pct_home: The free throw percentage of the home team

pf: Personal fouls committed by the player

stl: Steals. The number of times a defensive player legally takes the ball away from an offensive player.

to: Turnover. Any loss of possession of the ball by a team.

Fun Facts

Origin: Basketball was invented in December 1891 by Dr. James Naismith, a Canadian physical education instructor, as an indoor game to keep his students active during the winter months.

First Game: The first official game of basketball was played on December 21, 1891, at the International YMCA Training School in Springfield, Massachusetts. The final score was 1-0.

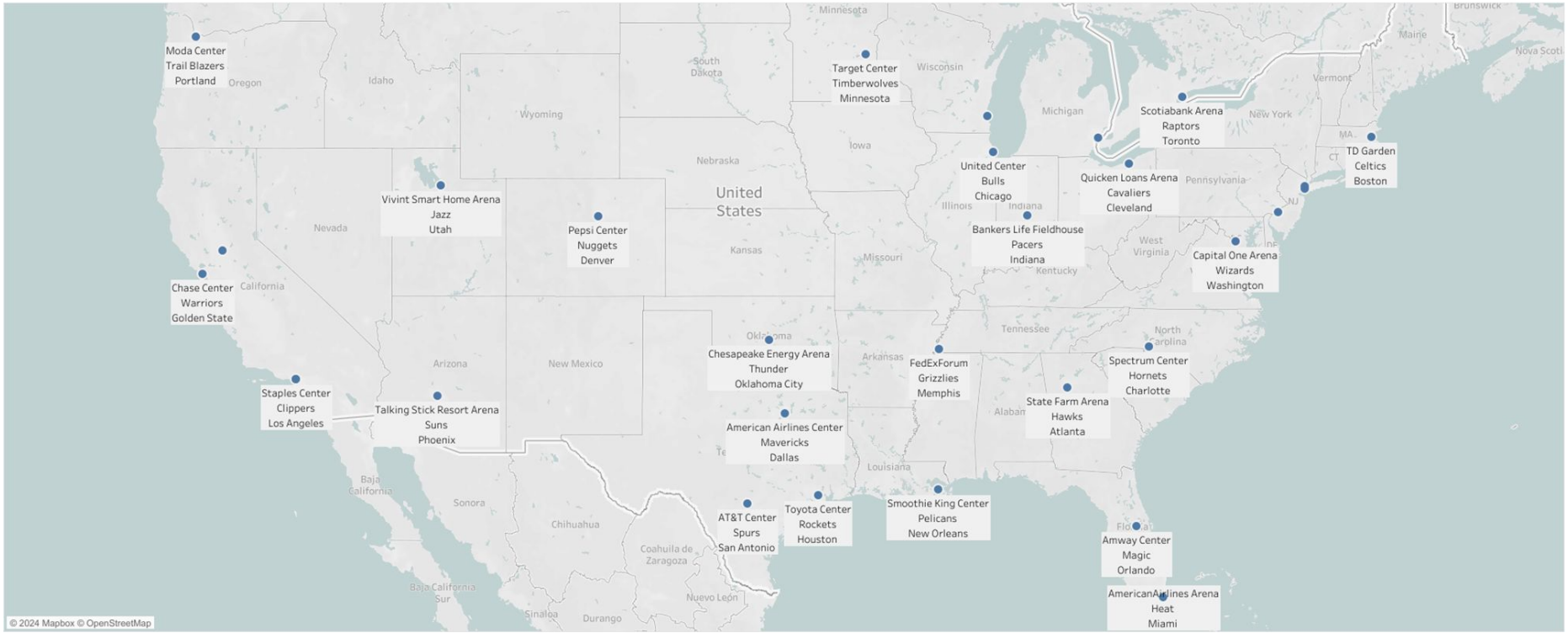
NBA's Humble Beginnings: The National Basketball Association (NBA), the premier professional basketball league in the world, was founded on June 6, 1946, as the Basketball Association of America (BAA). It merged with the National Basketball League (NBL) in 1949 to form the NBA.

A Sport for the Olympics: Basketball has been an Olympic sport since 1936. It is one of the most popular and widely watched sports during the Summer Olympics.

Four-Point Shot: In 2017, the Big3 basketball league introduced a four-point shot, known as the "Big3," which is located 30 feet away from the basket. It adds an extra dimension of excitement to the game.

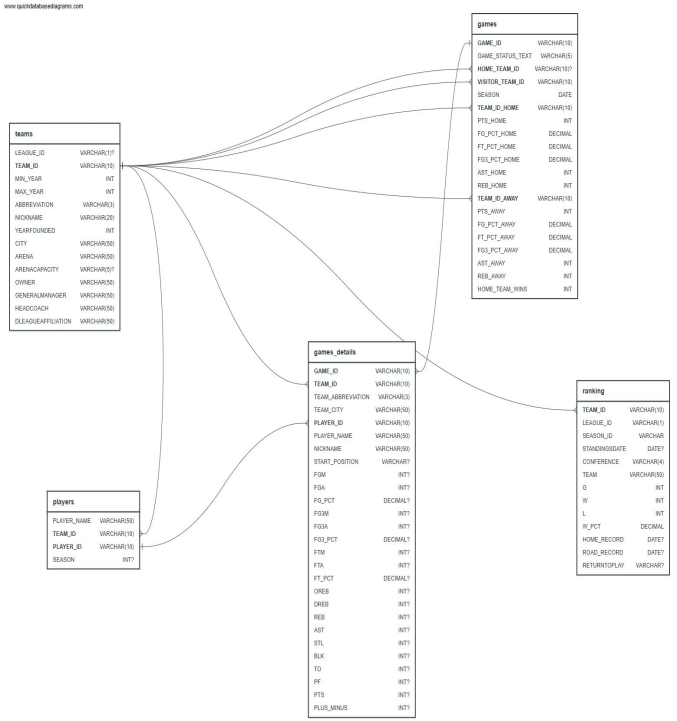
The primary professional basketball league in the US is the National Basketball Association (NBA), which consists of 30 teams. These teams are spread across the country and compete at the highest level of professional basketball in the US.

US NBA TEAMS


Moda Center
Trail Blazers
Portland
Target Center
Timberwolves
Minnesota
Scotiabank Arena
Raptors
Toronto
TD Garden
Celtics
Boston
United Center
Bulls
Chicago
Quicken Loans Arena
Cavaliers
Cleveland
Vivint Smart Home Arena
Jazz
Utah
Pepsi Center
Nuggets
Denver
Bankers Life Fieldhouse
Pacers
Indiana
Capital One Arena
Wizards
Washington
Chase Center
Warriors
Golden State
Chesapeake Energy Arena
Thunder
Oklahoma City
FedExForum
Grizzlies
Memphis
Spectrum Center
Hornets
Charlotte
Staples Center
Clippers
Los Angeles
Talking Stick Resort Arena
Suns
Phoenix
American Airlines Center
Mavericks
Dallas
State Farm Arena
Hawks
Atlanta
AT&T Center
Spurs
San Antonio
Toyota Center
Rockets
Houston
Smoothie King Center
Pelicans
New Orleans
Amway Center
Magic
Orlando
AmericanAirlines Arena
Heat
Miami
United States
© 2024 Mapbox © OpenStreetMap

Data Processing Workflow

1. Data Acquisition:
   - Obtained data from Kaggle.
   - https://www.kaggle.com/datasets/nathanlauga/nba-games?select=games.csv
   - https://www.kaggle.com/datasets/logandonaldson/sports-stadium-locations
   - Consisted of 5 CSV files: rankings, teams, games, games_details, and players.
2. Initial Challenges:
   - Difficulty importing data into SQL directly.
   - Leveraged Excel for initial data cleaning to resolve import issues.
3. Data Integration:
   - Merged selected files using SQL to create a unified dataset.
4. Data Refinement with Jupyter:
   - Utilized Jupyter for further data cleaning.
   - Dropped columns with missing or irrelevant data to streamline analysis.
5. Outcome:
   - Achieved a refined dataset ready for analysis and insights extraction.



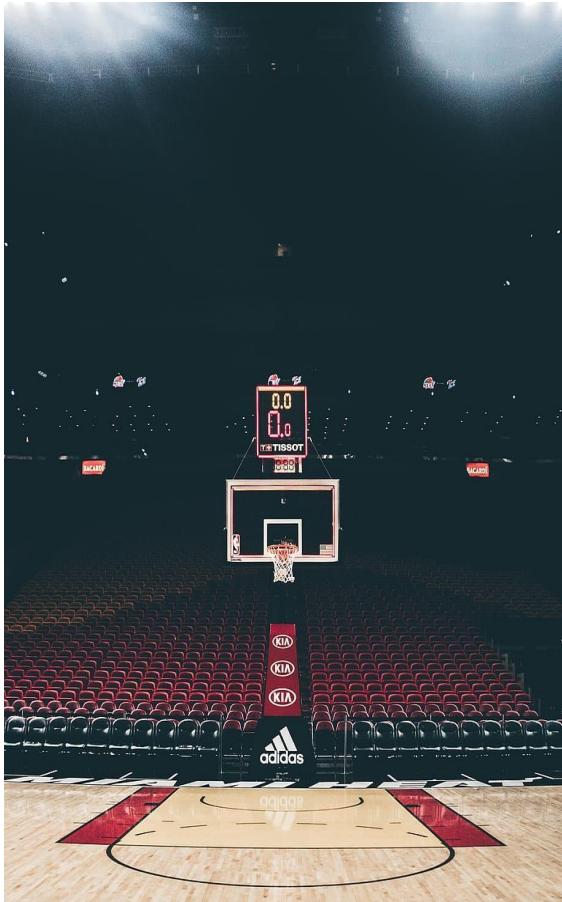
Model Preparation & Results

Random Forest

- Random Forest is a machine learning model based on decision trees.
- Utilizes multiple decision trees to improve predictive performance and reduce overfitting.
- Decided to use Random Forest because:
  - It is good at handling large datasets with multiple dimensions.
  - It can be used to generate feature importance which is useful for further analysis.

Random Forest Parameters

- **n_estimators:** 100
- **random_state:** 42



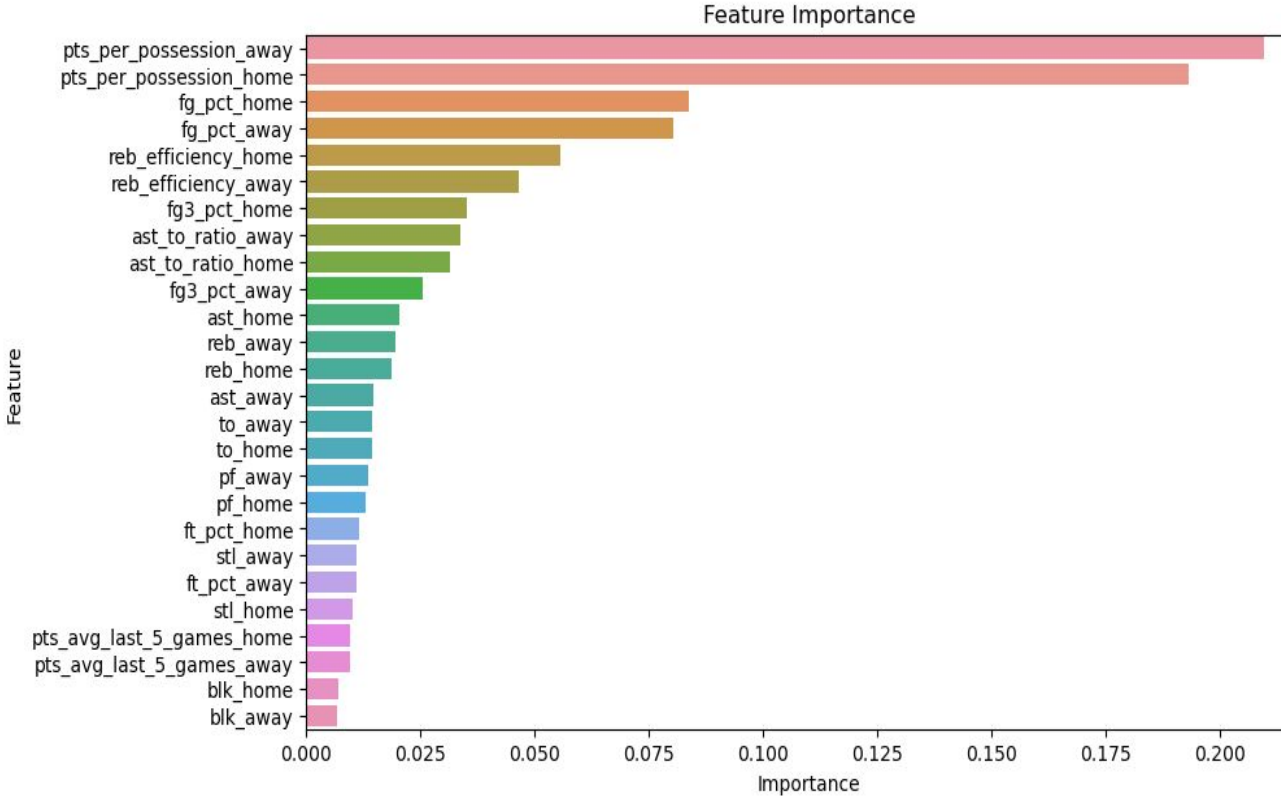
Classification Report

```
Confusion Matrix:
[[2038  166]
 [ 134 2967]]


Accuracy: 94.345%


Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.92      0.93      2204
           1       0.95      0.96      0.95      3101

    accuracy                           0.94      5305
   macro avg       0.94      0.94      0.94      5305
weighted avg       0.94      0.94      0.94      5305
```

Results Interpretation

- The model accuracy of **94.35%** shows a highly reliable model which is robust to overfitting.
- The precision and recall values are high for both home team wins and losses. The model succeeds at making predictions and could identity the majority of instances in both classes.
- The F1-score, macro average, and weighted average are all high, indicating that the model performs well overall. The high weighted average, in particular, shows that the model maintains its accuracy even when considering the support (the number of actual occurrences) of each class.

Random Forest Findings

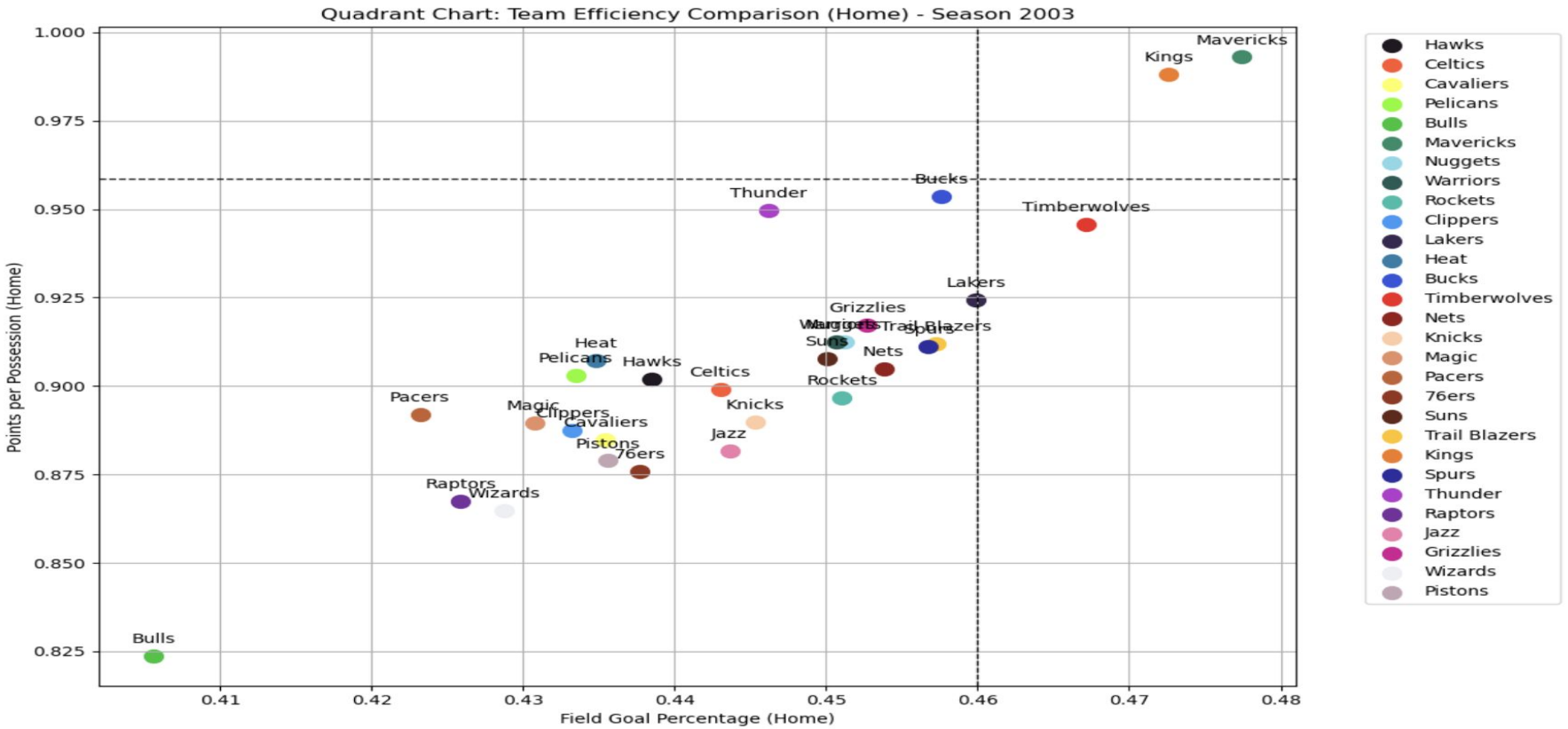


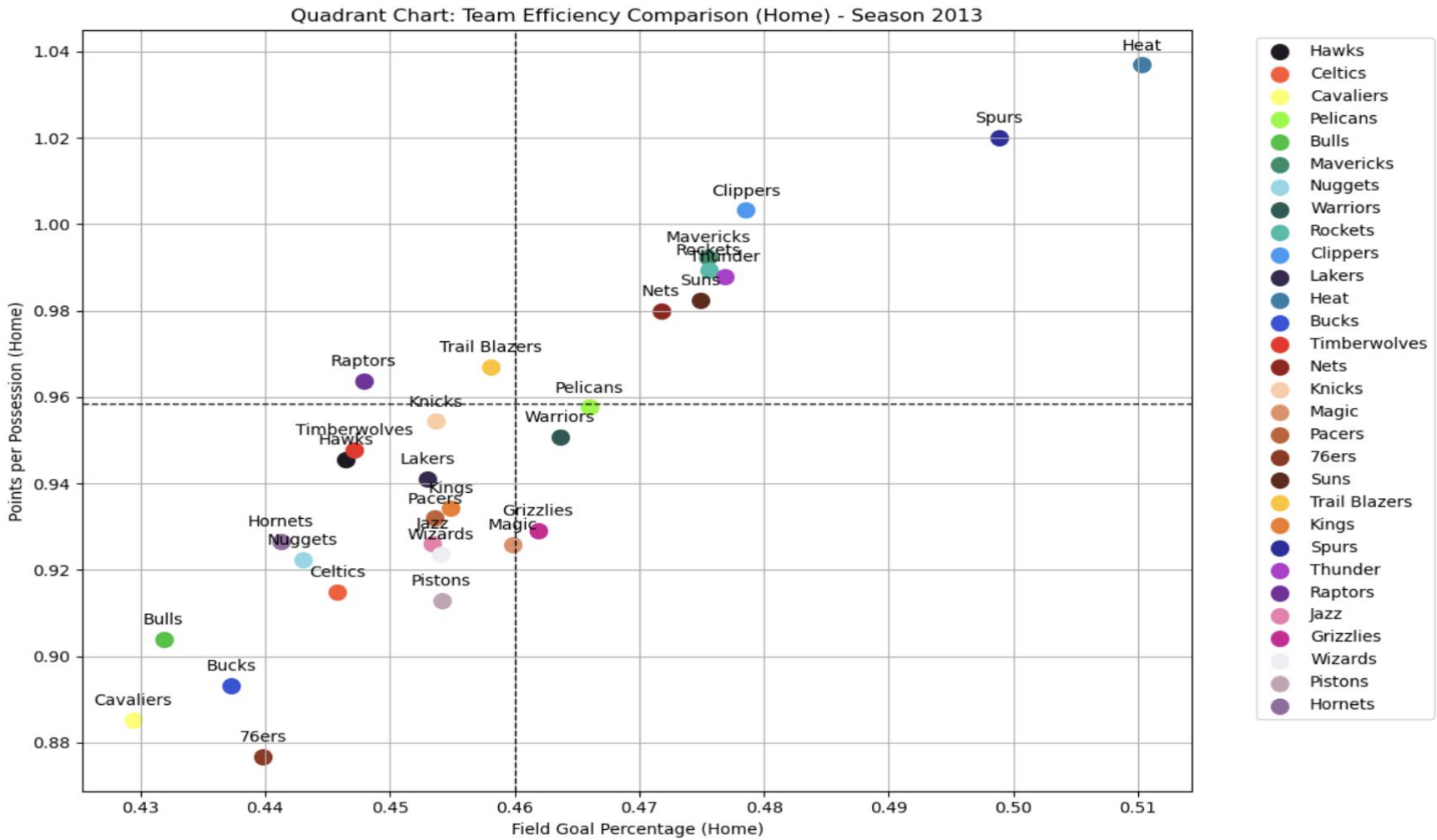
Using Random Forest, we found that the most important stats in predicting a winner was the **points per possession metric** for each team. This was a calculated feature which measures how efficient a team is an a possession by possession basis.

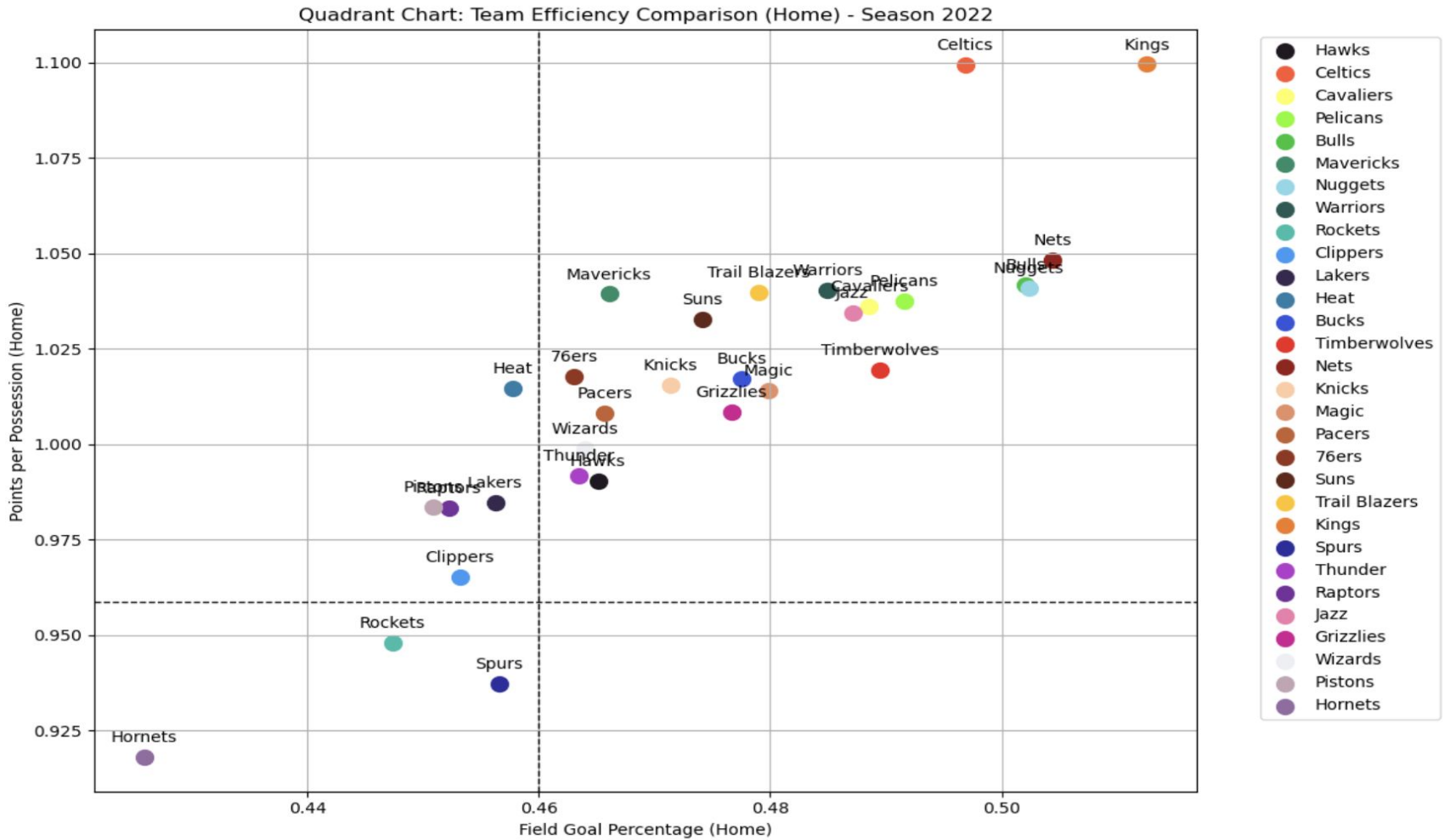
Rebound efficiency, which measures the proportion of available rebounds secured by a team, emerged as a significant indicator of game outcomes. This metric highlights the importance of controlling the boards, as teams with higher rebound efficiency are more likely to win games. Emphasizing improvements in this area could provide teams with a strategic advantage, suggesting that focusing on rebounding can be crucial for success

Blocks were the least important feature, suggesting it is more important for teams to make their own baskets than prevent the other team from doing so.

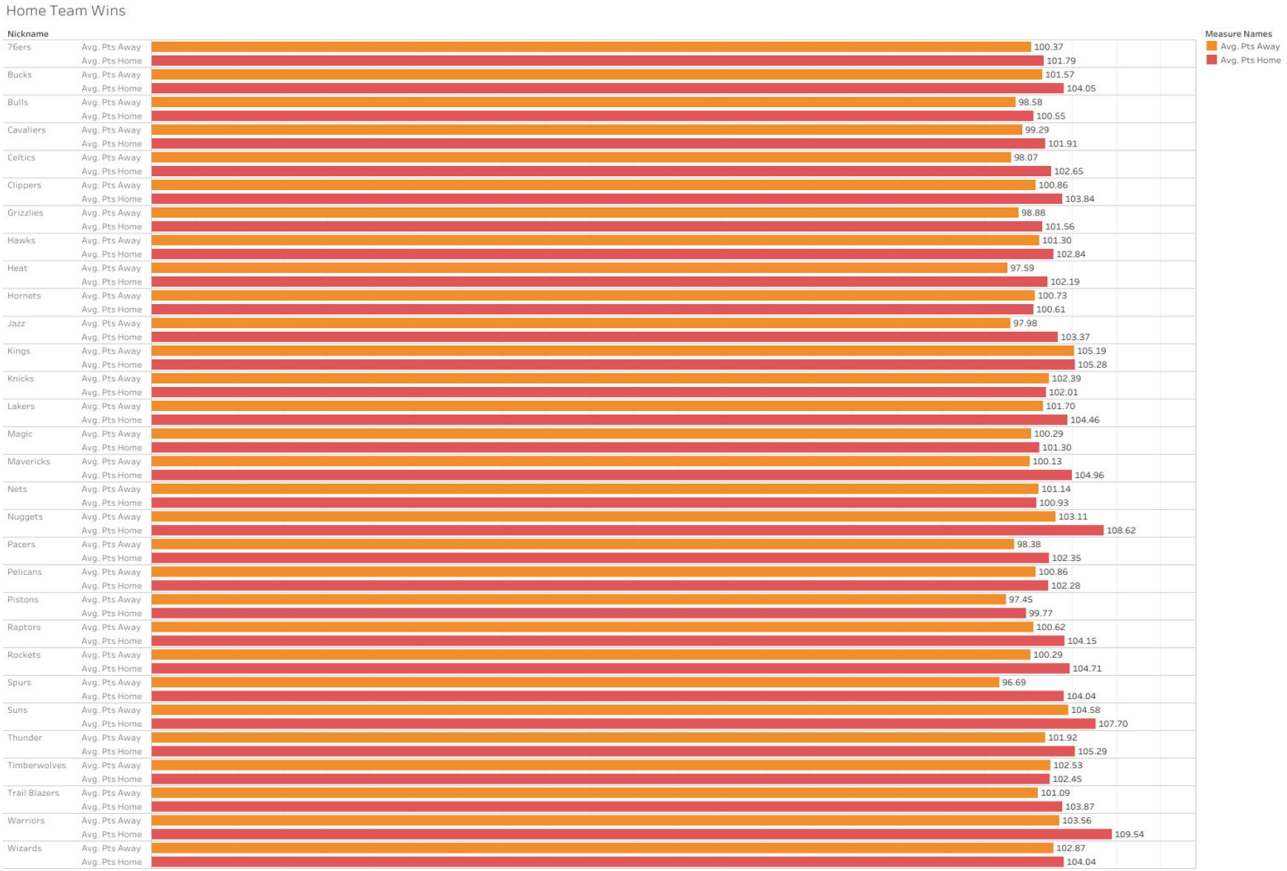
Further Analysis: FG% v PPP




Quadrant Chart: Team Efficiency Comparison (Home) - Season 2013
Points per Possession (Home)
Field Goal Percentage (Home)
1.04
1.02
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.43
0.44
0.45
0.46
0.47
0.48
0.49
0.50
0.51
Heat
Spurs
Clippers
Mavericks
Rockets
Thunder
Suns
Nets
Trail Blazers
Raptors
Pelicans
Knicks
Warriors
Timberwolves
Hawks
Lakers
Kings
Pacers
Grizzlies
Hornets
Jazz
Magic
Wizards
Nuggets
Celtics
Pistons
Bulls
Bucks
Cavaliers
76ers
Hawks
Celtics
Cavaliers
Pelicans
Bulls
Mavericks
Nuggets
Warriors
Rockets
Clippers
Lakers
Heat
Bucks
Timberwolves
Nets
Knicks
Magic
Pacers
76ers
Suns
Trail Blazers
Kings
Spurs
Thunder
Raptors
Jazz
Grizzlies
Wizards
Pistons
Hornets


Quadrant Chart: Team Efficiency Comparison (Home) - Season 2022
Points per Possession (Home)
Field Goal Percentage (Home)
1.100
1.075
1.050
1.025
1.000
0.975
0.950
0.925
0.44
0.46
0.48
0.50
Celtics
Kings
Nets
Mavericks
Trail Blazers
Warriors
Pelicans
Cavaliers
Jazz
Suns
Timberwolves
76ers
Knicks
Bucks
Magic
Heat
Pacers
Grizzlies
Wizards
Thunder
Hawks
Lakers
Clippers
Rockets
Spurs
Hornets
Hawks
Celtics
Cavaliers
Pelicans
Bulls
Mavericks
Nuggets
Warriors
Rockets
Clippers
Lakers
Heat
Bucks
Timberwolves
Nets
Knicks
Magic
Pacers
76ers
Suns
Trail Blazers
Kings
Spurs
Thunder
Raptors
Jazz
Grizzlies
Wizards
Pistons
Hornets

Home vs. Away Games



Null Hypothesis: There is no difference in performance between playing at home and playing away.

Alternative Hypothesis: There is a difference in performance between playing at home and playing away.

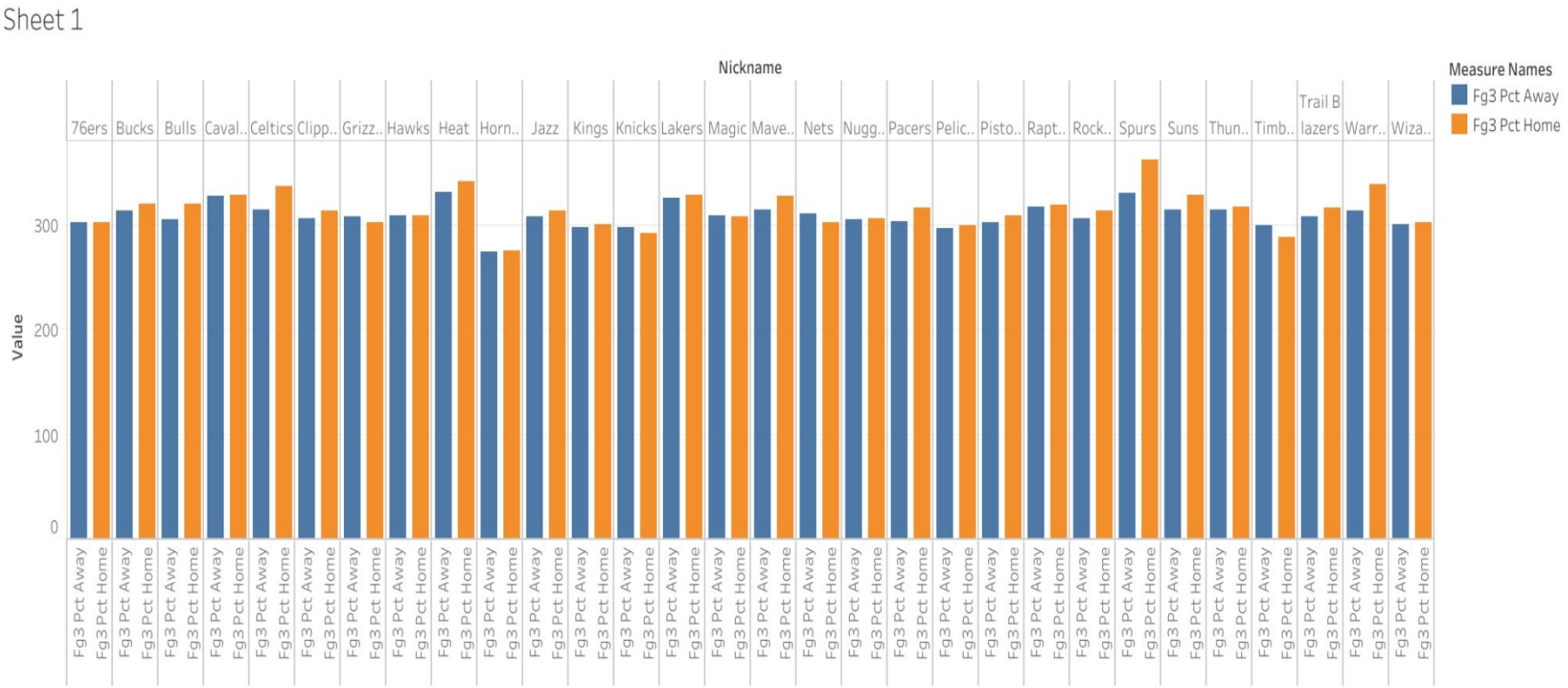
Paired t-test results:

t-statistic: 33.98416142035553

p-value: 7.907332620332287e-248

Reject the null hypothesis: There *is* a significant difference in performance between home and away games.

Home or Away Court Advantage?



Statistical Analysis of Three-point Shooting Performance

Home

Mean: 0.3560521811258153

Median: 0.357

Standard Deviation: 0.11117998835908582

Away

Mean: 0.34946676469479315

Median: 0.35

Standard Deviation: 0.10945199157913627

Three-point field goal percentage (FG3%) of teams

The visualization shows that the difference in three-point field goal percentage between home and away is generally insignificant, teams maintain similar three-point field goal percentage regardless of game location.

Contrary to common belief, the difference in FG3% suggests that there is no substantial advantage or disadvantage associated with playing at home or away.

The data underscores that while there are slight differences, they are not substantial enough to indicate a significant home court advantage in three-point shooting efficiency. This insight challenges conventional wisdom and highlights the importance of factors beyond game location in perimeter shooting performance.

Future Models

- Generally, our features focused on team-wide stats, but focusing on **player performance and team roster** could help strengthen future models.
- Future models could **use more or other stats**, such as court position of field goals, to see if the model's accuracy could improve. Likewise, other stats may be more important features than ones we used.
- Using a **neural network** could potentially increase the accuracy as well.
- Our model used mostly quantitative data with some qualitative data. However, all our data were hard, objective facts. It would be interesting to **incorporate subjective analysis** such as "team morale" or "crowd energy."

Glossary of Features

abbreviation: The abbreviated name of the team

arena: The name of the arena where the team plays its home games

arenacapacity: The seating capacity of the team's home arena

ast: Assists made by the player

ast_away: The total assists by the visiting team

ast_home: The total assists by the home team

asst_to_ratio: The ratio of a teams assists to the number of turnovers committed by the team.

blk: Blocks made by the player

city: The city where the team is located

comment: Any additional commentary or notes related to the player or game

dleagueaffiliation: The affiliation of the team with the NBA G League (if applicable).

dreb: Defensive rebounds grabbed by the player

fg_pct: Field goal percentage, calculated as (fgm / fga) * 100

fg_pct_away: The field goal percentage of the visiting team

fg_pct_home: The field goal percentage of the home team

fg3_pct: Three-point field goal percentage, calculated as (fg3m / fg3a) * 100

fg3_pct_away: The three-point field goal percentage of the visiting team

fg3_pct_home: The three-point field goal percentage of the home team

fg3a: Three-point field goals attempted by the player

fg3m: Three-point field goals made by the player

fga: Field goals attempted by the player

fgm: Field goals made by the player

ft_pct: Free throw percentage, calculated as (ftm / fta) * 100

ft_pct_away: The free throw percentage of the visiting team

ft_pct_home: The free throw percentage of the home team

fta: Free throws attempted by the player

ftm: Free throws made by the player

game_date_est: The date of the game in Eastern Standard Time (EST).

game_id: A unique identifier for each basketball game

game_id: A unique identifier for each game in a basketball season

game_status_text: Text indicating the status of the game (e.g., "Final", "Scheduled")

generalmanager: The general manager of the team

headcoach: The head coach of the team

home_team_id: The unique identifier for the home team

home_team_wins: Indicates whether the home team won the game

min: Minutes played by the player in the game

nickname: An informal name or alias for the player, if applicable

nickname: The nickname or mascot of the team

oreb: Offensive rebounds grabbed by the player

owner: The owner(s) of the team

pf: Personal fouls committed by the player

player_id: A unique identifier for each player in a basketball league

player_name: The name of the player

points_per_possession: Measure of the average number of points a team scores per possession. It is calculated by dividing the total points scored by the number of possessions

pts: Total points scored by the player

pts_away: The total points scored by the visiting team in the game

reb_efficency: The percentage of available rebounds secured by a team.

stl: Steals. The number of times a defensive player legally takes the ball away from an offensive player.

to: Turnover. Any loss of possession of the ball by a team

Questions?