Data 319 Final Project
Team 2
Mariah Bergquist, Daunte Dash, Carter Timm

Research Questions:

What performance metrics are most strongly associated with the success of NBA teams in the 2023-24 season?

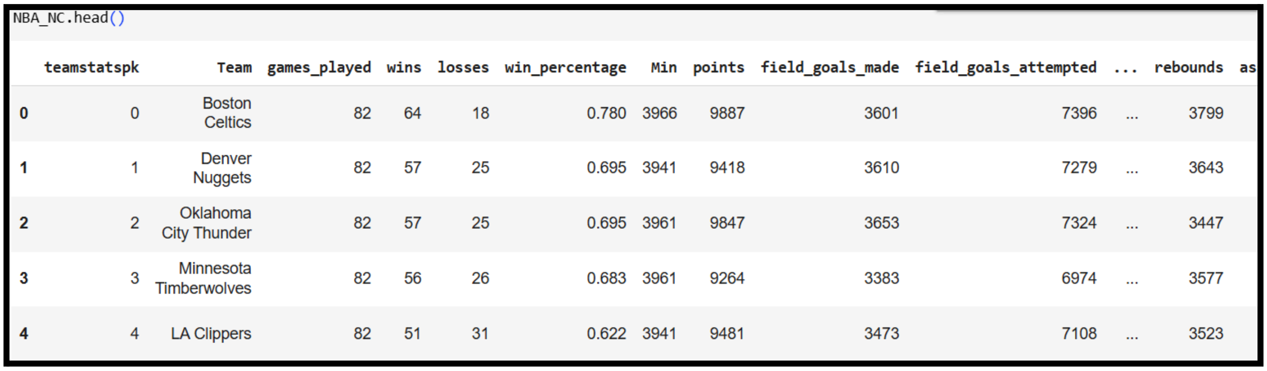
Data Description And Why We Chose Our Data For This Project

We selected our dataset which contains NBA statistics from 2000-2024 and includes performance measures such as total number of games played, points scored, total wins, losses, etc. We chose this dataset for our project because it provides the relevant information needed to answer our research question: which performance measures are most strongly associated with the success of NBA teams during the 2023-2024 season? Our chosen dataset has the proper information and data to help us answer this.

Our dataset has 717 rows and each row represents one team's stats for a particular season. It has 27 columns where one indexes each row starting from zero, one row is team names, another is the year of the season played, and the rest represent different performance measures. The following is a more detailed description of each column in our dataset:

- teamstatspk - indexes each observation in our dataset - INT
- Team - name of the team - STRING
- games_played - number of games played in a season - INT
- wins - number of wins in a season - INT
- losses - number of losses in a season - INT
- win_percentage - the win ratio for each team that season (wins divided by games_played) - FLOAT
- Min - number of minutes played in a season - INT
- points - number of points scored in a season - INT
- field_goals_made - number of field goals made in a season - INT
- field_goals_attempted - number of field goals attempted in a season - INT
- field_goal_percentage - the ratio of field goals made in a season compared to the field goals attempted (field_goals_made divided by field_goals_attempted) - FLOAT
- three_pointers_made - the number of three-pointers made in a season - INT
- three_pointers_attempted - the number of three-pointers attempted in a season - INT
- three_point _percentage - the ratio of three-pointers made in a season compared to the three-pointers made (three-pointers made divided by three-pointers attempted) - FLOAT
- free_throws_made - the number of free throws made in a season - INT
- free_throw_attempted - the number of free throws attempted in a season - INT
- free_throw_percentage - the ratio of free throws made in a season compared to the free throws attempted (free throws made divided by free throws attempted) - FLOAT

- offensive_rebounds - the number of offensive rebounds in a season - INT
- defensive_rebounds - the number of defensive rebounds in a season - INT
- rebounds - the number of overall rebounds in a season - INT
- assists - the number of assists in a season - INT
- turnovers - the number of turnovers in a season - INT
- steals - the number of steals in a season - INT
- blocks - the number of blocks in a season - INT
- blocks_attempted - the number of blocks attempted in a season - INT
- personal_fouls - the number of personal fouls committed in a season - INT
- personal_fouls_drawn - the number of personal fouls drawn in a season - INT
- season - the season that the particular team played in for that observation - STRING



```
NBA_NC.head()
```

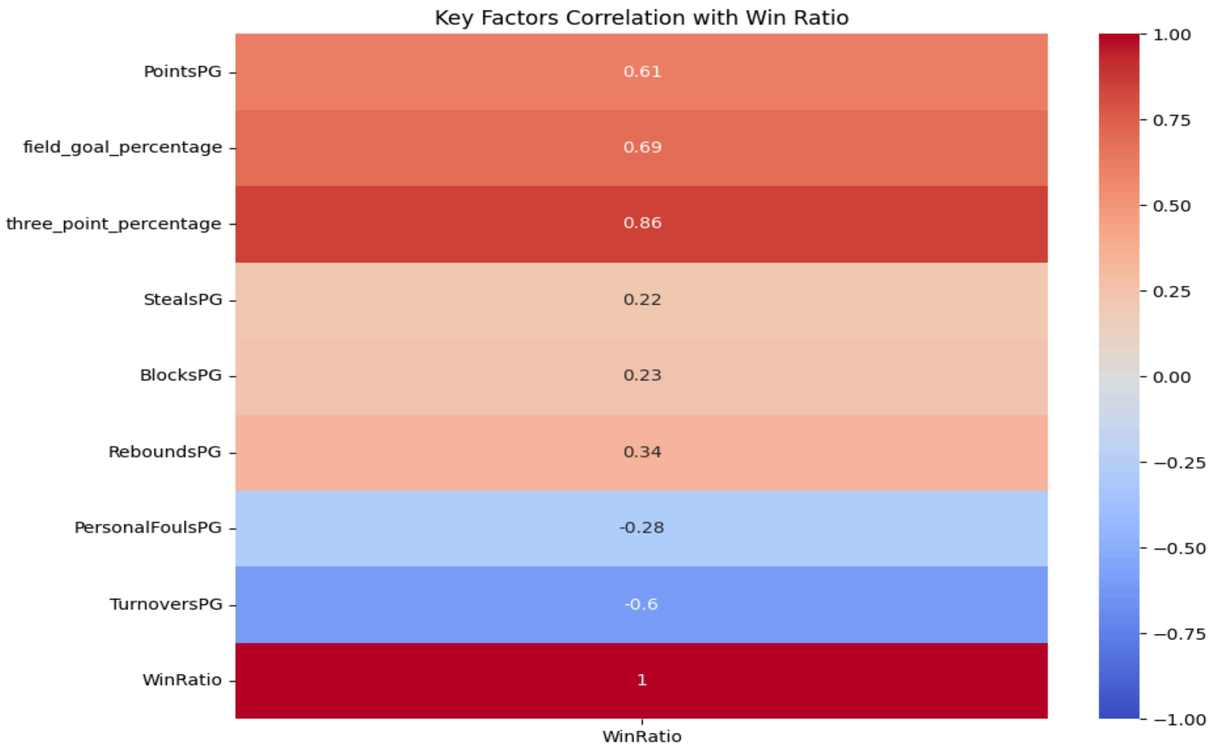
	teamstatspk	Team	games_played	wins	losses	win_percentage	Min	points	field_goals_made	field_goals_attempted	...	rebounds	as
0	0	Boston Celtics	82	64	18	0.780	3966	9887	3601	7396	...	3799	
1	1	Denver Nuggets	82	57	25	0.695	3941	9418	3610	7279	...	3643	
2	2	Oklahoma City Thunder	82	57	25	0.695	3961	9847	3653	7324	...	3447	
3	3	Minnesota Timberwolves	82	56	26	0.683	3961	9264	3383	6974	...	3577	
4	4	LA Clippers	82	51	31	0.622	3941	9481	3473	7108	...	3523	

Processing Problems We Had With Our Data

Our dataset was already partially clean when we first downloaded and read it into our code. One of the problems was that we wanted to do our analysis using “per game” statistics and not the “season total” statistics. This is because using per-game statistics to compare teams/players in the NBA is much more common, so we wanted to incorporate this into our analysis. Another problem we had was that our dataset had team statistics from 2000-2024, and we only needed the most recent season, not the 24 most recent seasons. To fix this, we just filtered out everything except for the 2023-24 NBA season.

Exploratory Analysis Results And Preliminary Conclusions

We selected columns that represent the key statistics used to compare and represent NBA teams. Offensive metrics include scoring statistics such as 3-point percentage, field goal percentage, and points per game which all have a pretty strong positive correlation with win ratio. Turnovers per game have a moderately strong negative correlation with win ratio. Defensive metrics include steals per game, blocks per game, and rebounds per game which all have a moderately positive correlation with win ratio.



Correlated with Win Ratio

Scoring Efficiency (Strong Positive Correlation)

- Field Goal Percentage
- 3-Point Percentage
- Points Per Game

Teams that score frequently and efficiently tend to win more games. This is logical when simplified: to win a basketball game, a team must score more points than its opponent. Based on this correlation analysis, scoring efficiency measures appear to play a significant role in determining an NBA team's success.

Rebounds Per Game (Moderately Positive Correlation)

Rebounding, whether on offense or defense, plays a crucial role in a team's success. Offensive rebounds provide additional scoring opportunities, while defensive rebounds allow a team to regain possession and transition to offense. Our data shows a moderately positive correlation between rebounding and win ratio, suggesting that effective rebounding measures may contribute to determining an NBA team's success.

Defense (Low Positive Correlation)

- Steals Per Game
- Blocks Per Game

There is a common saying that “defense wins championships”. It may be important to know that this quote originated from football and not basketball because our correlation analysis doesn’t appear to fully support this. While we do see a positive correlation between steals per game and blocks per game, it's pretty low suggesting that defensive measures may not contribute to determining an NBA team's success

Turnovers Per Game (Strong Negative Correlation)
Personal Fouls Per Game (Moderate Negative Correlation)
It appears that both turnovers per game and personal fouls per game have negative correlations with win ratio. Given the moderate and strong negative correlations, our data suggests that minimizing turnovers and personal fouls may significantly contribute to an NBA team's success.

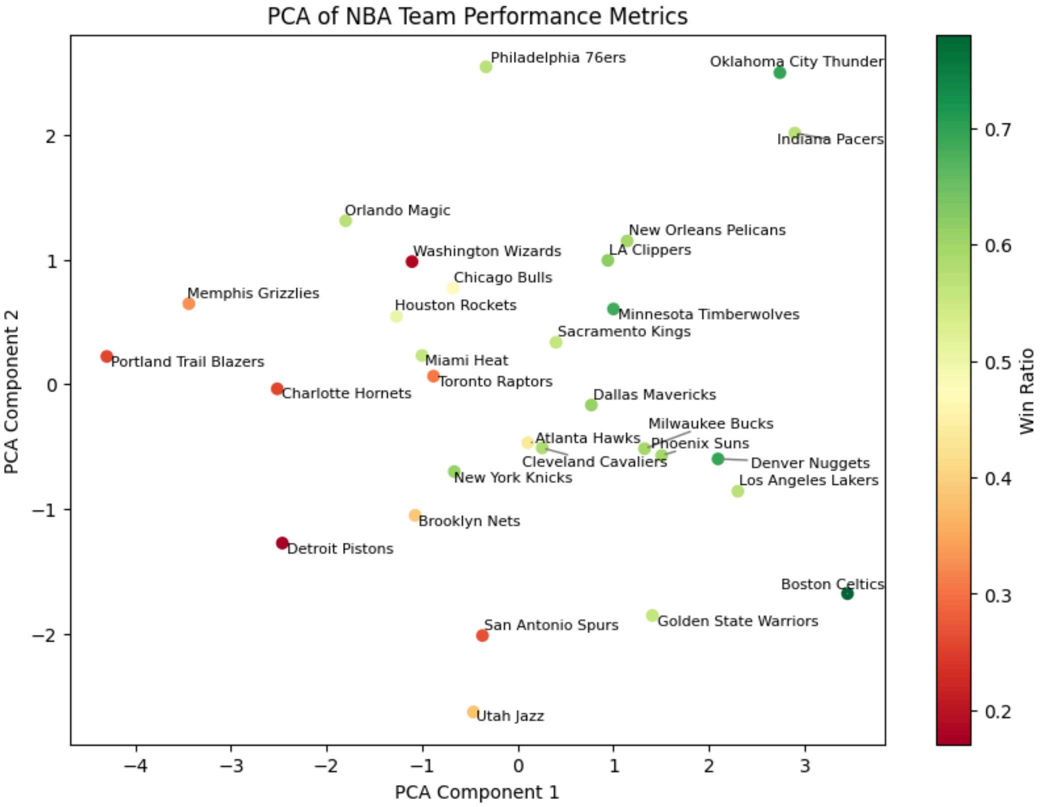
Describing How We Determined Our Methodology For Our Analysis

Our dataset was pretty easy to clean and transform into a usable form for our analysis. We did, however, run into problems regarding our analyses. For example, instead of using PCA, we were going to use MDS to try and capture what performance metrics were most strongly associated with the success of NBA teams. The thought process behind that was we could measure the similarities between teams, find groups if there were any, of the teams with better records, and look at the performance metrics they succeeded in. This would help us answer our research question, but it didn’t go as planned. It went pretty well until we had to find the performance metrics that made the better teams succeed. To mitigate this issue, we decided to use PCA. This switch allowed us to be able to use the loadings to determine the performance metrics that made the better teams succeed and because of this, we were able to better answer our research question.

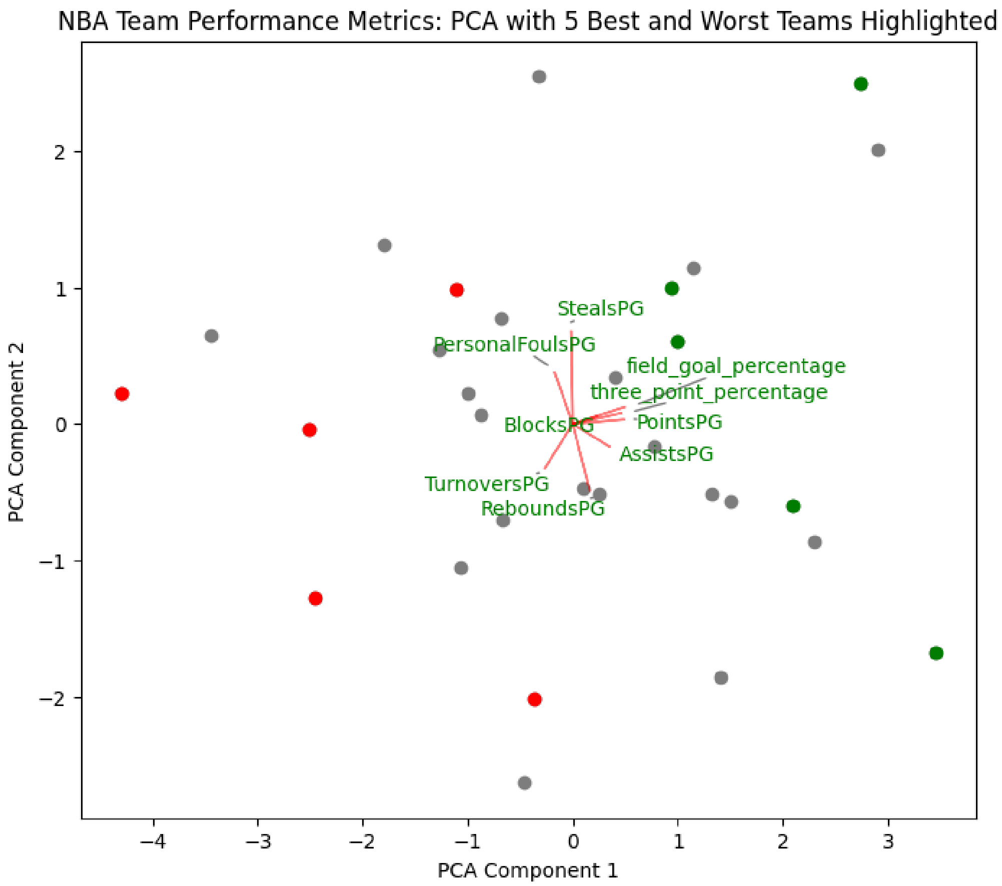
Principal Component Analysis can be really useful for reducing the dimensionality of our data which allows for our clustering to be more efficient. However, PCA does have its downsides like how it always assumes linearity. Assuming linearity between variables can possibly lead to ignoring non-linear patterns in the data which would cause us to lose extremely valuable information in our analysis. It also doesn’t explain all of the variance of the original variables chosen. This is the tradeoff when conducting PCA, so while it does perform dimension reduction, it will explain less of the variance in the data.

While completing our hierarchical clustering analysis, selecting the best features to use was important. At first, we started with rebounds and points per game. These two features had a low correlation of 0.32. I figured this would be good to analyze because low correlation won’t generalize the data, and having two features allows us to interpret and visualize our results more easily. I also wanted to make sure before I did any analysis to standardize the data. I wanted to make sure I used the standardized data to ensure there was an equal contribution from each variable. Hierarchical clustering has some pros including easy visualization using a dendrogram, and not having to specify the amount of clusters in our analysis. However there are some downsides including a lack of efficiency with large datasets, and it can be hard to determine the number of clusters in our analysis. Luckily our dataset is not large, making hierarchical clustering a good option for our analysis.

Conclusions from our analysis:



PCA – From our PCA analysis of NBA teams in the 2023-24 season, we found that high-performing and efficient offenses were crucial for team success. The principal components explained 54% of the variance in the data. While this means that our principal components do not capture 46% of the variance, the analysis still provides significant insights into the factors that contribute to team performance.

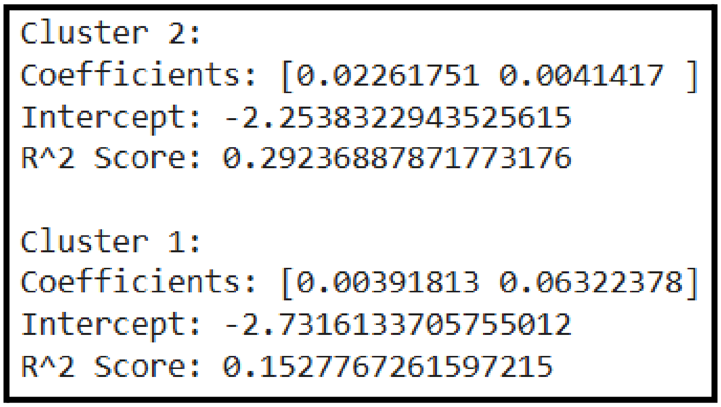


After analyzing the loadings from PCA, the top-performing teams were farther to the right on the horizontal axis while the worst-performing teams were farther to the left. This was primarily because of four performance measures: field goal percentage, three-point percentage, points per game, and assists per game. These performance measures represent offensive performance and efficiency, and their loadings for the principal components largely affected where teams were plotted on the horizontal axis

Defensive performance measures didn't play as large of a role in determining the team's success. The top-performing and bottom-performing teams varied quite a bit on the vertical axis and there didn't appear to be any clear difference between these teams on the vertical axis. The loadings that affected the vertical axis the most were steals per game, rebounds per game, personal fouls per game, and turnovers per game. While these statistics intuitively are important in basketball, they aren't performance measures that are strongly associated with the success of NBA teams in the 2023-24 season.

Hierarchical Clustering – From our hierarchical clustering analysis, we concluded that teams with a higher win ratio also found more success with a more efficient offense. The points and rebounds per game explain about 44% of the variance in win ratio. About 66% of the variance in win ratio is not explained by rebounds per game and points per game. While these factors don’t explain a huge amount of the variance, they still explain some of the variance which allows us to conclude their relationship with winning games.

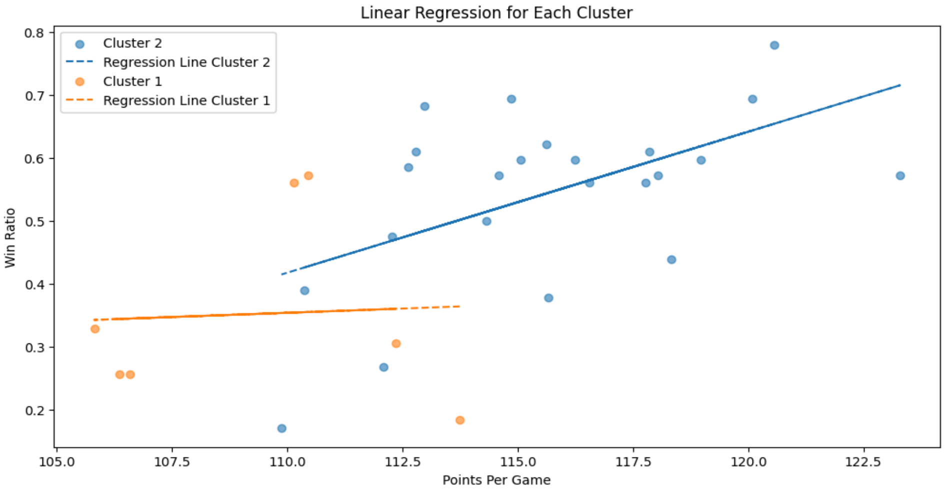
We started with creating a dendrogram that grouped teams based on similar rebounds per game and points per game statistics. We decided to cut our dendrogram off at two given that it would be easier to visualize relationships with only two features. From here we created two clusters and analyzed each cluster's win ratio. Cluster One had seven teams while Cluster Two had 23 teams. This may have slightly skewed our win ratios, but Cluster One appeared to have a win ratio of 35%, while Cluster Two had a win ratio of 54%. We graphed our clusters on a scatter plot that showed their relationship between points per game and rebounds per game. This scatterplot showed that Cluster One teams tended to not have many points per game or rebounds per game whereas Cluster Two tended to have higher points per game and rebounds per game with higher win ratios.



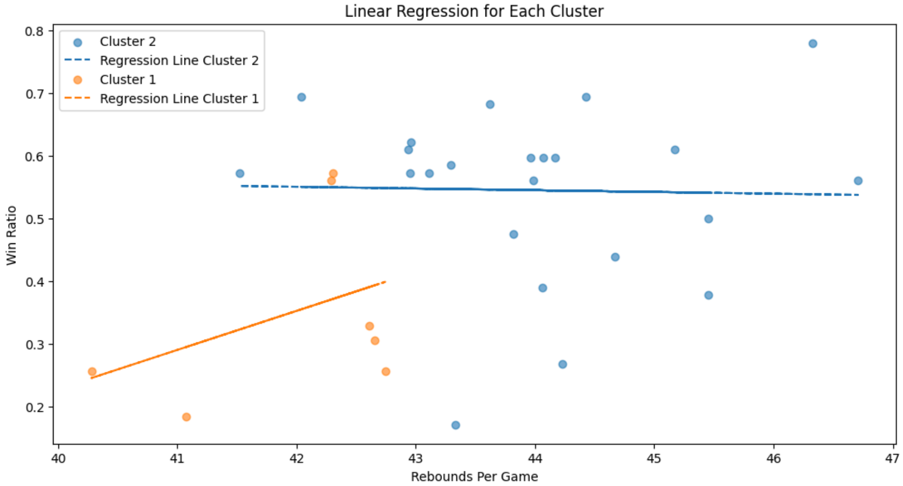
```
Cluster 2:
Coefficients: [0.02261751 0.0041417 ]
Intercept: -2.2538322943525615
R^2 Score: 0.29236887871773176

Cluster 1:
Coefficients: [0.00391813 0.06322378]
Intercept: -2.7316133705755012
R^2 Score: 0.1527767261597215
```

We wanted to conduct further analysis to understand the relationships these clusters had so we created a linear regression model. This model's results gave us the coefficient of points per game (0.0039) and the coefficient of rebounds per game (0.0632) for Cluster One. These coefficients tell us that rebounds per game have a stronger relationship with the win ratio than points per game. For Cluster Two our coefficients were [0.02261751, 0.0041417]. This tells us that our points per game have a stronger relationship with win ratio in Cluster Two than in Cluster One.



After looking at the linear regression model results, we thought it would be a good idea to graph the results for each feature and look at each one separately. This graph showcases our models' results for how effective points per game are for each team's win rate. Cluster 1 has a flatter regression line meaning scoring points does not reliably predict these teams winning their games. Cluster 2 has a steep positive slope meaning that points per game play a big role in why these teams win.



This graph showcases our models' results for how effective rebounds per game are for each team's win rate. Cluster 1 has a positive slope meaning they may rely mainly on defense to win games. Cluster 2 has a primarily flat regression line which means getting rebounds per game does not reliably predict these teams winning their games.

Additional Analyses We Would Have Liked To Carry Out

It would have been interesting to extend our analysis to multiple seasons. We only analyzed the 2023-24 season because of how fast the NBA changes, so we wanted the most recent data to conduct our analysis. If we were to extend our analysis to the past five years, we would have many more data points, which could appear a little overcrowded. However, if our stipulation that the NBA changes too quickly to conduct this type of analysis is wrong, then this would be a good analysis to do. We do have the data for this as well. Our data included NBA team statistics from 2000-2024, so we wouldn't have needed any additional data.

Another compelling analysis that we would have liked to carry out would be a more in-depth one with the same research question by using team statistics per 100 possessions instead of per game. This would account for the different tempos that teams play with. Some NBA teams move the ball around a lot on offense and use up most of the shot clock to find the best shot, as well as wearing the other team's defense out. Other teams may have a quicker playstyle to catch defenses off guard. Whatever it may be, NBA teams play at different paces, and that isn't captured in our data. The additional data we would've needed for this would be the same team statistics, but transformed into "per 100 possessions."