Building a Championship NBA Team

Kyle Lo, Harry Morton, Jyothi Ranjit, & Krystal Sung,

Table of Contents

Topic: Building a Championship NBA Team

What Determines a Championship NBA Team?

Our team wants to **generate a model** that is able to **predict a team's likelihood to win the NBA**.

By training our model on the results of each team in NBA history by season, we can **make a prediction on any given organization** in the future.

Reason For Topic Selection

As fans of the NBA, we wanted to use the skilled we learned in this Boot Camp to analyze a sport we all enjoy watching.

The NBA is a massive franchise that has large data sets readily available.

We take NBA data to create a model that helps us identify successful NBA teams.

Source of Data

We Identified 3 Key Areas for Data Sourcing

1. Player Information

This includes position, historical injuries, aggregated statistics and various salary information

Players are vital in building a championship team!

2. Team Information

Who is the coach? What is the attendance? What is the available salary cap?

Players cannot win the NBA without a sound team as a foundation.

3. NBA Season History

Where did teams placed in the past? What were player stats for each year? Injuries? Other misc. Info

Understanding the rankings gives us training data!

Data Sources: Player Information

Player data includes:

Games played, Points, Field Goals, 3Pts, Free Throws, Steals, Blocks, Turnovers, Rebounds, etc..

Purpose of Data:

Can we determine if a player is on a winning team based on stats?

Can we predict the winning team based on players stats?

How does each feature weigh against each other?

BY YEAR	TEAM	GP	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	TOV	STL	BLK	PF	FP	DD2	T
2020-21	MIL	61	33.0	28.1	10.3	18.0	56.9	1.1	3.6	30.3	6.5	9.5	68.5	1.6	9.4	11.0	5.9	3.4	1.2	1.2	2.8	53.9	41	
2019-20	MIL	63	30.4	29.5	10.9	19.7	55.3	1.4	4.7	30.4	6.3	10.0	63.3	2.2	11.4	13.6	5.6	3.7	1.0	1.0	3.1	56.6	56	
2018-19	MIL	72	32.8	27.7	10.0	17.3	57.8	0.7	2.8	25.6	6.9	9.5	72.9	2.2	10.3	12.5	5.9	3.7	1.3	1.5	3.2	56.2	54	
2017-18	MIL	75	36.7	26.9	9.9	18.7	52.9	0.6	1.9	30.7	6.5	8.5	76.0	2.1	8.0	10.0	4.8	3.0	1.5	1.4	3.1	51.7	42	
2016-17	MIL	80	35.6	22.9	8.2	15.7	52.1	0.6	2.3	27.2	5.9	7.7	77.0	1.8	7.0	8.8	5.4	2.9	1.6	1.9	3.1	49.2	32	
2015-16	MIL	80	35.3	16.9	6.4	12.7	50.6	0.4	1.4	25.7	3.7	5.1	72.4	1.4	6.2	7.7	4.3	2.6	1.2	1.4	3.2	37.7	21	
2014-15	MIL	81	31.4	12.7	4.7	9.6	49.1	0.1	0.5	15.9	3.2	4.3	74.1	1.2	5.5	6.7	2.6	2.1	0.9	1.0	3.1	28.3	10	
2013-14	MIL	77	24.6	6.8	2.2	5.4	41.4	0.5	1.5	34.7	1.8	2.6	68.3	1.0	3.4	4.4	1.9	1.6	0.8	0.8	2.2	18.2	2	

Data Sources: Team Information

Team Info includes:

Team stats, standings, win %, salary,

market cap, attendance, ect...

Purpose of Data:

Does a team stats determine a championship?

Does standing determine which team will win the championship?

Does the money machine go BRRRRR?!

Does popularity of a team attracts better players?

	TEAM	GP	W	L	WIN%	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB
1	Utah Jazz	72	52	20	.722	48.2	116.4	41.3	88.1	46.8	16.7	43.0	38.9	17.2	21.5	79.9	10.6
2	Phoenix Suns	72	51	21	.708	48.6	115.3	43.3	88.3	49.0	13.1	34.6	37.8	15.6	18.7	83.4	8.8
3	Philadelphia 76ers	72	49	23	.681	48.4	113.6	41.4	86.9	47.6	11.3	30.1	37.4	19.6	25.5	76.7	10.0
4	Brooklyn Nets	72	48	24	.667	48.3	118.6	43.1	87.3	49.4	14.2	36.1	39.2	18.1	22.5	80.4	8.9
5	Denver Nuggets	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	43.3	89.2	48.5	12.9	34.2	37.7	15.7	19.5	80.3	10.5
5	LA Clippers	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	41.8	86.7	48.2	14.3	34.7	41.1	16.2	19.3	83.9	9.4
7	Milwaukee Bucks	72	46	26	.639	48.1	120.1	44.7	91.8	48.7	14.4	37.1	38.9	16.2	21.4	76.0	10.3

NBA Champions!!!

Data Sources: NBA Season History

History Data includes:

Past championships, past stats, injuries, games played per season, past standings, finals appearances, ect...

Purpose of Data:

Training data to train our model on winning teams.

How does injury affect a team's chance of winning a championship?

Determining probability to win championship

Likelihood to place in the top 8 teams or bottom 8 teams

Our 4 Key Areas for Data Sourcing

Kaggle

TGFK: Thank goodness for Kaggle!

Contains massive amount of clean datasets that fits our needs.

Contains previous analysis we can build upon

NBA-API (nba.stats.com)

Used to extract more recent data to build upon existing Kaggle data

Can obtain raw form of data that will need wrangling

fivethirtyeight

Contains RAPTOR data that is a new metric to measure basketball stats.

Contains analysis for dataset like kaggle.

espn.com

Web Scraping needed

Contains salaries

Has a simple layout of leaderboards with stats of players/teams

Questions Answered

Analysis with The SIX

Who will win the championship? The Machine Learning model will predict based on players stats, who is going to win the championship. To win the game of basketball is simple. Whoever has the most points at the end of the time limit is the winner. We can do a quick analysis and assume that the teams that scores the most points per game are more likely to win.

What makes a championship team? We know that a player with good stats will make a team good, but teams are limited with cap space. So does a players salary determine how good they are? With this knowledge we know that teams are willing to pay individuals a lot more money to have them on their team.

When: Have the attributes that predict a championship winning team changed over time? We can answer this by aggregating various statistics by year for our ideal population

Where did the team place last year in the standings? We know that every season is different, but understanding where teams placed in the standings give us insight in their future investment. In contrast teams usually place depending on where star players move.

Why does a team win? Does offense or defense win a game? Does injury play a bigger role than having a team with good stats? From this analysis alone we can determine that if a star player is injured then it is unlikely for your team to win. Since your star player takes a majority of the salary cap, a team will have worser players on the bench. Looking at the standings will only tell a person how a team performed during the regular season. When playoffs come around many players are usually injured.

How does playoff game stats differ from the regular game stats? When analyzing the different game stats, it is best to differentiate the regular season stats with playoff stats. In every game there is an unknown factor called "clutch." How clutch is a player? A team or a player can have amazing regular season stats but chokes in the playoffs when it matters.

Data Exploration

The Data Exploration Process

What are we trying to explore?

Does good stats make a good player/team?

Does a bigger market cap entice better players?

Does a player’s salary affect a teams chance of winning?

Can we predict this years winner based on past championship game stats?

Our 4 Main Data Sources

From our four main data sources, we had to adjust our analysis based off the data we were able to obtain.

At first we did not find any easily available salary data. We had to web scrape data to obtain a vital part of our analysis.

The data exploration phrase was a critical portion of our final project.

Data Analysis

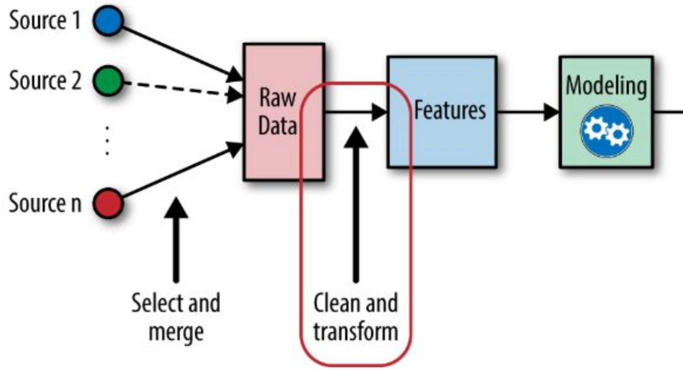
Data Analysis: Preprocessing

For data preprocessing, we needed to find a way to utilize the statistics we had collected in a way that was most fair when modeling at the team level

- As our statistics increased in to the hundreds and thousands for some teams, we needed to scale our data to ensure larger values would not affect our model

- As we had data from 1946, we chose to drop rows of training data where we saw nulls (due to old data)

Data Analysis: Feature Engineering



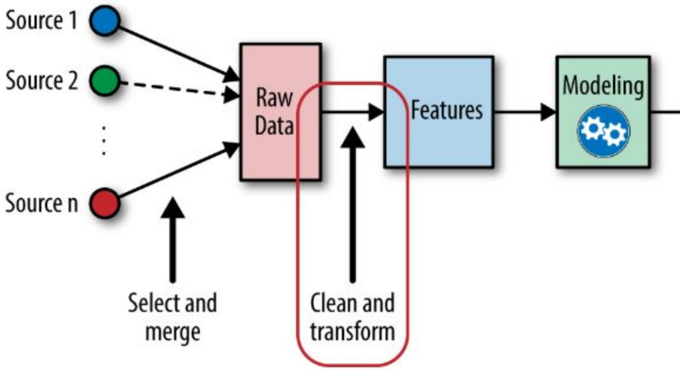
We needed to find a way to utilize all of the statistics we had gathered at an individual level to represent performance of the team

- Here we chose to aggregate statistics by position to get a sense of how winning teams looked across the roster
- We used a rolling lookback window of 3 years, to get a sense of how a player had performed and get a sense of 'form' vs. trusting career stats

Data Analysis: Feature Eng Cont.

We elected to use certain statistics for our model, to reduce the noise that comes with too many features that can be correlated. These were:

- Total Assists
- Total Steals
- Total Blocks
- Total Turnovers
- Personal Fouls



Data Analysis: Splitting to Train

When splitting our training data, we wanted to give as many training data points to our model as we could

- When training the model, we split the training size as 82%
- This allowed us to utilize the training data for our model and evaluation

Data Analysis: Choosing the Model

We decided to utilize decision trees within our classification model, for a few different reasons:

- They are simple, and there is logic in not overcomplicating
- Data preprocessing is relatively simple
- We could potentially train without normalizing / scaling

We understood that they require a lot of memory / time for the mathematical calculation, but given the size of our dataset this was not a concern

Data Analysis: Training the Model

We trained our model on 2 different subsets:

- Ideal: Teams that have won the NBA in recent history each season
- Non Ideal: The 2 teams that have historically had the worst win percentage in the NBA each season

These two have enough differentiation where we can predict ideal / non ideal effectively

- We played with including more teams, but it becomes harder to model a winning team the farther from the top of the table you get
- They all start to look similar!

Data Analysis: Additional Training

We have an abundance of data at our fingertips. We may try to include the following in our training:

- Reducing the lookback window to 2 years vs. 3 years (to get a better sense of immediate form)
- Join salary aggregations in as a feature
    - This could potentially be by position
    - Could also just be observed salary on the team

Data Analysis: Accuracy

Our current model predicts with 82% accuracy. We are happy with this as a baseline model for a few reasons:

- Our model is effective at identifying teams that will not win the NBA based on the recent performance of players on the roster
- 33% recall for ideal prediction is pretty good, because there can be noise in player statistics as we get closer to the top of the table

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	13	1
Actual 1	2	1

Accuracy Score : 0.8235294117647058

Classification Report

	precision	recall	f1-score	support
0.0	0.87	0.93	0.90	14
1.0	0.50	0.33	0.40	3
accuracy			0.82	17
macro avg	0.68	0.63	0.65	17
weighted avg	0.80	0.82	0.81	17

Data Analysis: Summary

All in all we are pretty satisfied with the model performance thus far, as we are able to distinguish teams that are not likely to win the NBA based on player performance!

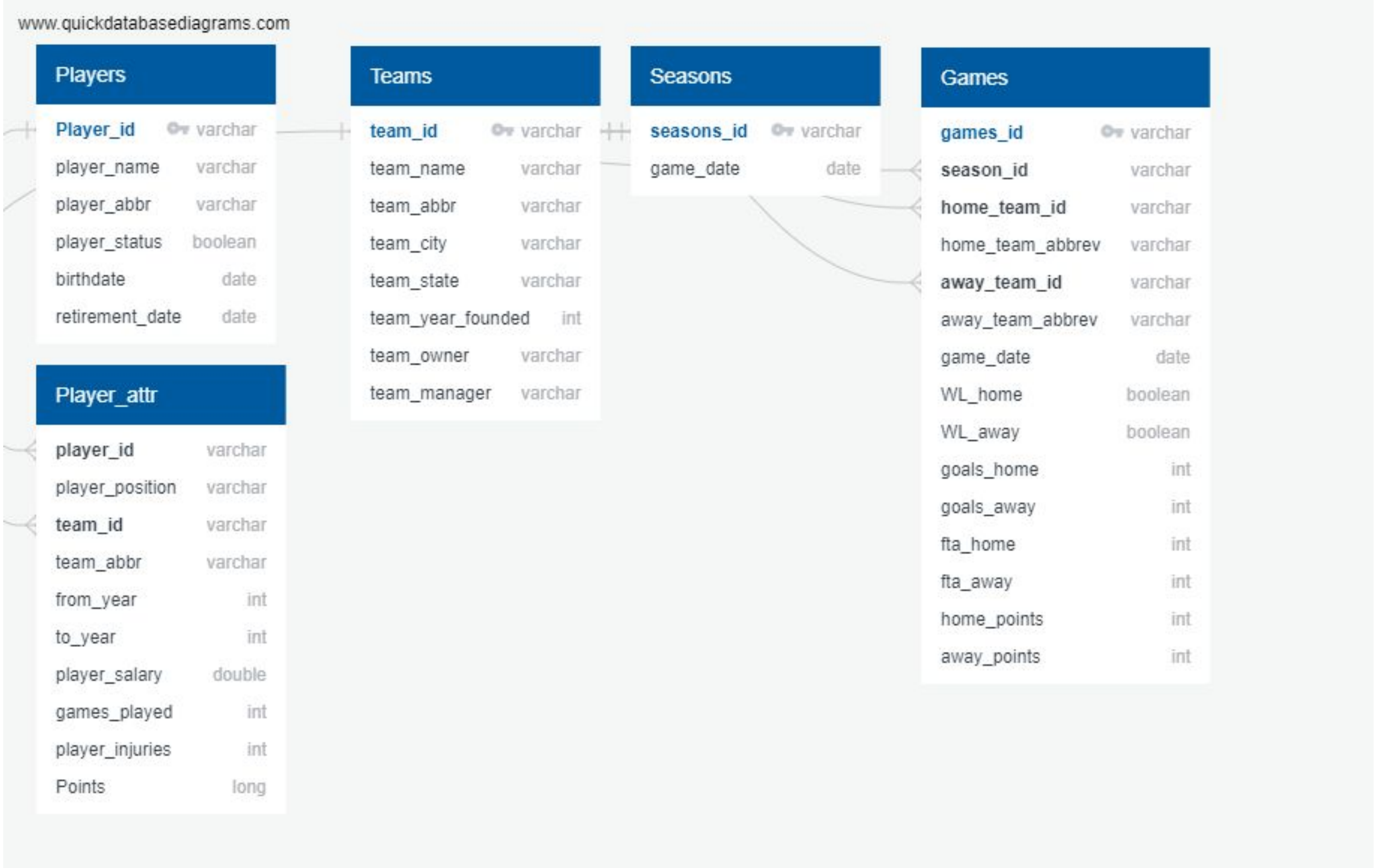
- These predictions were based on real, historical NBA seasons and the statistics for those players at that point in time!

Layering in salary will be another great feature towards answering the questions that we have laid out in the slides above

Resources Used Throughout the Project

Leveraging SQL & Postgres for our Database

- AWS database used to store all of our exploratory data so anyone on the team can access it at anytime.
- We are using Postgres (pgAdmin) to connect to the AWS database.
- Cleaning will be done outside of Postgres
- Database is for clean final exploratory data only


www.quickdatabasediagrams.com
Players
Player_id varchar
player_name varchar
player_abbr varchar
player_status boolean
birthdate date
retirement_date date
Player_attr
player_id varchar
player_position varchar
team_id varchar
team_abbr varchar
from_year int
to_year int
player_salary double
games_played int
player_injuries int
Points long
Teams
team_id varchar
team_name varchar
team_abbr varchar
team_city varchar
team_state varchar
team_year_founded int
team_owner varchar
team_manager varchar
Seasons
seasons_id varchar
game_date date
Games
games_id varchar
season_id varchar
home_team_id varchar
home_team_abbrev varchar
away_team_id varchar
away_team_abbrev varchar
game_date date
WL_home boolean
WL_away boolean
goals_home int
goals_away int
fta_home int
fta_away int
home_points int
away_points int

Final Dashboard Tool: Tableau Public

We will use Tableau Public to create our final dashboard.

The dashboard will be an interactive platform to help users understand what builds a championship NBA team.

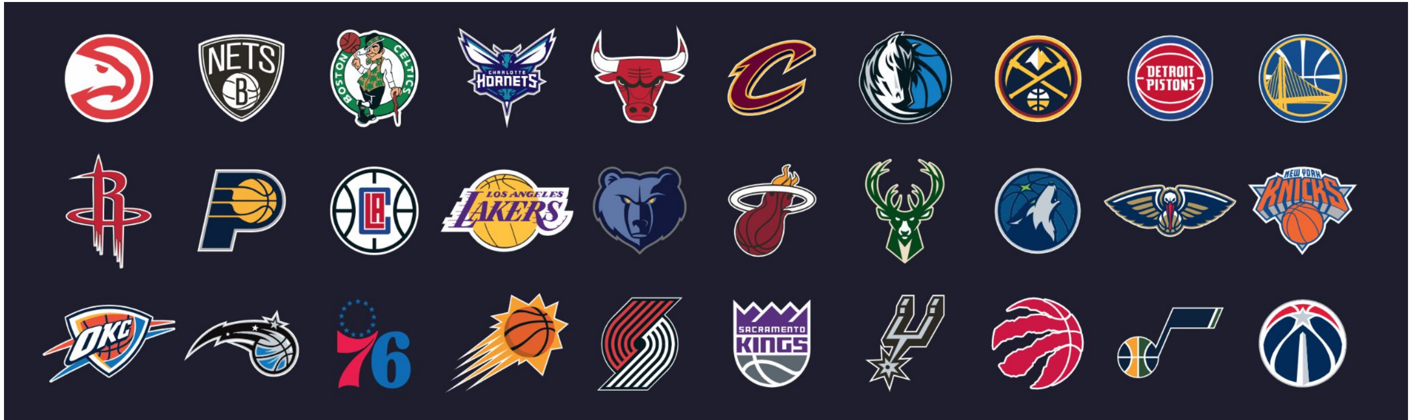
Building a Championship NBA Team

Final Dashboard

For our Final Dashboard, we will create an interactive dashboard by NBA Team and NBA Player.

Our Final Dashboard will include statistics as well as information on salaries for teams and players.

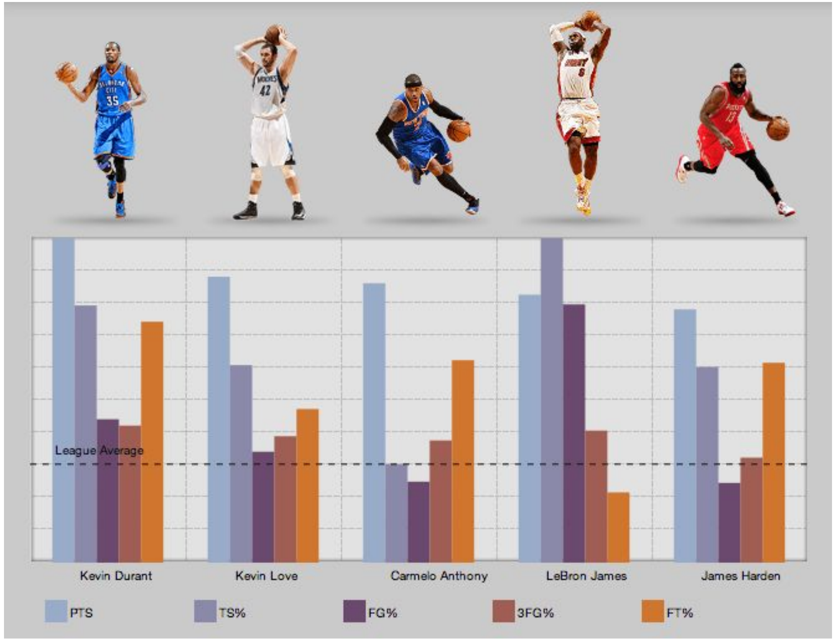
The Final Dashboard will help users determine what it takes to be a championship NBA team.



Final Dashboard: Statistics

Player information, team information, and NBA ranking history are all vital in creating a championship NBA team.

Based on NBA statistics, we will provide a rank of best performing NBA Teams and NBA Players.



Final Dashboard: Salary

A NBA team's salary cap plays a vital role in building a team and the success of the season.

Based off the players salaries from espn.com, we will provide information on a team's salary cap and player's salary in our Final Dashboard.

We analyze how teams utilize their salary cap and the success of the players the team acquires.

What does it take to win an NBA Championship?

The final dashboard will include an interactive model where users can create their own NBA team.

This estimation will be calculated by understanding the players statistics, injury history, salary, and more!

Thank You.