Project Phase 1

Edmund Lo, Zexi Lv

Domain

The domain for our project is the NBA 2020-2021 regular and playoff season. We are interested in the statistics for teams, players, coaches and games.

Datasets

Teams Datasets:

Per Game Stats from: 2020-21 NBA Standings | Basketball-Reference.com

Conference Standings from: 2020-21 NBA Season Summary | Basketball-Reference.com

Players Datasets:

Player Per Game from: 2020-21 NBA Player Stats: Per Game | Basketball-Reference.com

Coaches Datasets:

2020-2021 NBA Coaches from: 2020-21 NBA Coaches | Basketball-Reference.com

Games Datasets:

Schedules for each month from: 2020-21 NBA Schedule and Results | Basketball-Reference.com

These datasets contain information about the NBA 2020-2021 regular and playoff seasons. It contains information about the players, teams, coaches, and games. While all of the information was relevant to our domain, not all of it was necessary to answer our investigative questions. To answer the first question, the relevant information was the statistics about a team's performance in their games and their win/loss percentage. The relevant information for the second question was the statistics about each player's performance in their games and how many points they scored. Finally, the relevant information for our third question was the amount of experience the coaches had (how many seasons they had coached their team) and how well their team performed.

We may have to learn more about what strategies players and teams employ. The playstyle and practices of different players, or teams can have a dramatic effect on their statistics and their chances of winning against other teams, which is relevant for our first two questions since we are studying what factors make a team successful. Finally, the data is very clean and structured as is, so we do not need to clean the data. To present the data, we may have to make the attribute names more descriptive and suitable for an audience that is not very familiar with basketball terminology.

Investigative Questions

What team stats have the greatest effect on a team's win loss percentage?

What player stats have the greatest effect on a player's points per game?

What is the relationship between a coach's experience and their team's success during the season?

Schema

- Teams(tID, city, name, wins, losses, W/L%, FG, FGA, FG%, 3P, 3PA, 3P%, 2P, 2PA, 2P%, FT, FTA, FT%, ORB, DRB, TRB, AST, STL, BLK, TOV, PF, PTS)

  A tuple in this relation represents a team and its stats during the games the team played.

- Players(pID, firstName, lastName, tID, position, age, G, GS, MP, FG, FGA, FG%, 3P, 3PA, 3P%, 2P, 2PA, 2P%, eFG%, FT, FTA, FT%, ORB, DRB, TRB, AST, STL, BLK, TOV, PF, PTS)

  A tuple in this relation represents a player. It contains information about the player such as their name and which team they belong to, as well as some of their game stats such as the number of points they scored.

- Coaches(cID, firstName, lastName, tID, seasonsFranch, seasonsOverall, regGames, regWins, regLosses, playoffGames, playoffWins, playoffLosses)

  A tuple in this relation represents a coach. It contains information about the coach such as their name, which team they coached in the 2020-2021 season, and their win/loss records in the season.

- Games(gID, date, tIDHome, pointsHome, tIDAway, pointsAway)

  A tuple in this relation represents a game and contains information on the teams which played in the game and the number of points they scored in the game, as well as when the game was played.

Data Dictionary

Teams

Attribute	Description	Type	Required	Default
tID	The ID of a team	INT	YES	
city	The city of a team	TEXT	YES	
name	The name of a team	TEXT	YES	
wins	The wins of a team in the 2020-2021 season	INT	YES	
losses	The losses of a team in the 2020-2021 season	INT	YES	
W/L%	The percentage of games a team won in the 2020-2021 season	INT	YES	
FG	The number of field goal makes per game	INT	YES	
FGA	The number of field goal attempts per game	INT	YES	
FG%	The percentage of field goal makes per game	INT	YES	
3P	The number of three point makes per game	INT	YES	
3PA	The number of three point attempts per game	INT	YES	
3P%	The percentage of three point makes per game	INT	YES	
2P	The number of two point makes per game	INT	YES	
2PA	The number of two point attempts per game	INT	YES	
2P%	The percentage of two point makes per game	INT	YES	
FT	The number of free throw makes per game	INT	YES	
FTA	The number of free throw attempts per game	INT	YES	
FT%	The percentage of free throw makes per game	INT	YES	
ORB	The number of offensive rebounds per game	INT	YES	
DRB	The number of defensive rebounds per game	INT	YES	
TRB	The total number of rebounds per game	INT	YES	
AST	The number of assists per game	INT	YES	
STL	The number of steal per game	INT	YES	

BLK	The number of blocks per game	INT	YES	
TOV	The number of turnovers per game	INT	YES	
PF	The number of personal fouls per game	INT	YES	
PTS	The total points per game	INT	YES	

Players

Attribute	Description	Type	Required	Default
pID	The ID of a player	INT	YES	
firstName	The first name of a player	TEXT	YES	
lastName	The last name of a player	TEXT	YES	
tID	The ID of the team, a player plays for	INT	YES	
position	The position of a player	TEXT	YES	
age	The age of a player	INT	YES	
G	The number of games played	INT	YES	
GS	The number of games started	INT	YES	
MP	The number of minutes played per game	INT	YES	
FG	The number of field goal makes per game	INT	YES	
FGA	The number of field goal attempts per game	INT	YES	
FG%	The percentage of field goal makes per game	INT	YES	
3P	The number of three point makes per game	INT	YES	
3PA	The number of three point attempts per game	INT	YES	
3P%	The percentage of three point makes per game	INT	YES	
2P	The number of two point makes per game	INT	YES	
2PA	The number of two point attempts per game	INT	YES	
2P%	The percentage of two point makes per game	INT	YES	

eFG%	The effective field goal percentage	INT	YES	
FT	The number of free throw makes per game	INT	YES	
FTA	The number of free throw attempts per game	INT	YES	
FT%	The percentage of free throw makes per game	INT	YES	
ORB	The number of offensive rebounds per game	INT	YES	
DRB	The number of defensive rebounds per game	INT	YES	
TRB	The total number of rebounds per game	INT	YES	
AST	The number of assists per game	INT	YES	
STL	The number of steal per game	INT	YES	
BLK	The number of blocks per game	INT	YES	
TOV	The number of turnovers per game	INT	YES	
PF	The number of personal fouls per game	INT	YES	
FG	The number of field goal makes per game	INT	YES	
FGA	The number of field goal attempts per game	INT	YES	
PTS	The total points per game	INT	YES	

Coaches

Attribute	Description	Type	Required	Default
cID	ID of the coach	INT	YES	
firstName	First name of the coach	TEXT	YES	
lastName	Last name of the coach	TEXT	YES	
tID	ID of the team the coach belongs to	INT	YES	
seasonsFranch	Number of seasons the coach has been with the team	INT	YES	
seasonsOverall	Number of seasons the coach	INT	YES	
regGames	Number of games the coach has been with the team in the 2020-2021 regular	INT	YES	

	season			
regWins	Number of games the coach has won with the team in the 2020-2021 regular season	INT	YES	
regLosses	Number of games the coach has lost with the team in the 2020-2021 regular season	INT	YES	
playoffGames	Number of games the coach has been with the team in the 2020-2021 playoff season	INT	YES	
playoffWins	Number of games the coach has won with the team in the 2020-2021 playoff season	INT	YES	
playoffLosses	Number of games the coach has lost with the team in the 2020-2021 playoff season	INT	YES	

Games

Attribute	Description	Type	Required	Default
gID	ID of the game	INT	YES	
date	Date the game was played	DATE	YES	
tIDHome	Team ID of the home team	INT	YES	
pointsHome	Number of points the home team scored	INT	YES	
tIDAway	Team ID of the away team	INT	YES	
pointsAway	Number of points the away team scored	INT	YES	

Integrity Constraints

- Players[tID] ⊆ Teams[tID]
- Coaches[tID] ⊆ Teams[tID]
- Games[tIDHome] ⊆ Teams[tID]
- Games[tIDAway] ⊆ Teams[tID]
- For a tuple in Teams, wins plus losses must equal 72, since there are 72 games in the 2020-2021 season
- Players[G] must be less than 72
- Players[position] ⊆ {“PG”, “SG”, “SF”, “PF”, “C”} or a combination of any two
- For a tuple in Games, tIDHome must not equal tIDAway

Justification of Design

For our Teams relation, we combined data from two different datasets. We took the Team Name, Wins, Losses and Win/Loss Percentage attributes from the Conference Standings dataset and combined it with the statistics about a team's performance per game from the Per Game Stats dataset. This would allow us to investigate the relationship between a team's per game statistics like field goal percentage, points per game or assists per game and their success during the 2020-2021 season which can be defined by their win/loss percentage.

For our Teams relation, we also split apart team names into its city and its name because they should be two different attributes. This led us to create a new attribute called tID (team id) which is a key to uniquely identify each team. The key, tID, would help reduce any redundancy in our other relations. Instead of a reference to both the team's city and the team's name in our Player, Coach and Games relation, we can now use the team's unique id.

For our Players relation, we took the full Players Per Game dataset, minus the Rank of the player. In the Players Per Game dataset, the same player can have multiple rows because of their games and statistics on different teams in the same season. We decided to treat these as completely different players. Thus, we created a key called pID (player id), where each id corresponds to a certain player on a certain team (the same player has different pIDs for when they were on different teams). By keeping the statistics for a certain player's time on different teams separate, we can see if there is any relationship between the player's performance on a specific team and his team's performance.

For our Coaches relation, we used data from the 2020-2021 NBA Coaches dataset, taking the Seasons with Franchise and Seasons Overall attributes as well as all other attributes relating to the current season (2020-2021 regular season record and playoff record). Since all of our other data is related to the 2020-2021 season, we felt that these were the only necessary attributes. We also added a key called cID (coach id) to uniquely identify each coach.

For our Games relation, we used data from the Schedules datasets. We also added a key called gID (game id) to uniquely identify each game. We did not need to change the structure of this dataset, since it already has a good design where each row is a different game and has information on the home and away teams and the points for the home and away team.