DSCI 510 Final Project Report

Name and Author: Manasi Thonte

Motivation/ Rationale for the project:

Thirty teams play professional basketball in the National Basketball Association (NBA), of which twenty-nine are from the United States and one is from Canada. It is regarded as the top professional basketball league in the world and is one of the biggest professional sports leagues in the US and Canada. The National Basketball League (NBL) and the Basketball Association of America (BAA), which was created in 1946, combined to form the NBA in 1949. Every team plays 82 games during the October–April NBA season, which is followed by a postseason tournament that ends in the June NBA Finals. Talented athletes, excellent basketball, and a big cultural impact are the league's hallmarks.

This project is driven by a desire to understand the complex relationship between player performance, team salary, and overall team success in the NBA. Understanding the impact of player performance and team salary on success could further help the teams make more informed decisions regarding salary allocation and player recruitment.

Description of data sources:

Data Source 1: https://www.basketball-reference.com/leagues/NBA_2023.html

The data source is a web page which contains statistical data for basketball, particularly focusing on the National Basketball Association (NBA). The source contains tables containing per game stats, total stats tables along with respective parameters. The data from the total stats table would be extracted using web scraping using BeautifulSoup python library.

Data Source 2: https://hoopshype.com/salaries/2022-2023/

The data source is a site that has information on basketball like related news, rumors and salaries data.
The cap salary for NBA teams for the year 2022-2023. The data would be extracted using web scraping using BeautifulSoup python library.

Data Source 3: https://www.kaggle.com/datasets/joebeachcapital/nba-player-statistics/data

The data source is a site that provides a dataset for Basketball player performance statistics for the year 2022-2023. The data would be downloaded from the website in CSV format.

Integrated Data Model:

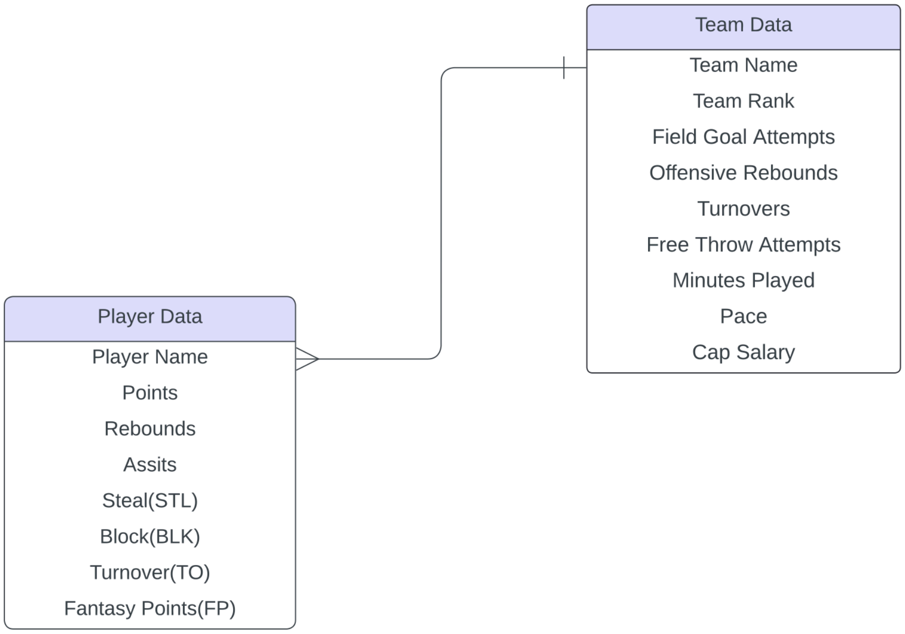
The data model for the NBA dataset typically includes three datasets, each representing different aspects of the NBA ecosystem. Here's a general overview of the main components:

1. Players Table: This table contains information about individual players, including their name, position, age, team affiliation, and statistics such as points scored, assists, rebounds, steals, blocks, turnovers, and more. Each row represents a unique player.

2. Teams Table: The teams table provides details about NBA teams, such as their name, overall performance metrics. Each row represents a unique team.

3. Salaries Table: This table stores salary data for NBA players, including their annual earnings, contract details, and team affiliation. Each row represents a unique player's salary information.

The tables are interconnected through relationships such as players of various teams, teams with their respective Cap Salaries. Analyzing and visualizing data using this data model helps us gain insights on the most influential player of a respective team, correlation of top 5 players in a team and the team rank as well as team Cap Salary.



Analysis/Visualizations:

The player’s data has various performance metrics available which can be used to compute a composite score for an individual player.

The various advanced metrics that were computed to further compute the composite score are as mentioned below:

a) Fantasy Point:

It is an advanced performance metric used by NBA stat analyst to determine the performance of a player by assigning weightage as mentioned below to following parameters:

- Points = 1.0 fantasy point
- Rebounds = 1.2 fantasy points
- Assists = 1.5 fantasy points
- Steals = 3.0 fantasy points
- Blocks = 3.0 fantasy points
- Turnovers = -1.0 fantasy points

Fantasy Point(FP) is calculated using below formula:

FP= 1*Points + 1.2* Rebounds + 1.5*Assists + 3*Steals + 3*Blocks + (-1)*Turnovers

b) Player Efficiency Rating (PER):

A comprehensive rating that measures a player's overall performance, combining various aspects of the game into a single number.

PER is calculated as:

PER = (1 / PM) * [TP * (2/3) + FGM + (1/2) * (3PM - FGA) + FTM * (1/2) + (REB * (3/4)) + AST + STL + (BLK * (3/2)) - PF - TO]

TP:	Total Points
PM:	Minutes Played
FGM:	Field Goals Made
3PM:	3-Pointers Made
FGA:	Field Goals Attempted
FTM:	Free Throws Made
REB:	Total Rebounds
AST:	Assists

STL:	Steals
BLK:	Blocks
PF:	Personal Fouls
TO:	Turnovers

c) True Shooting Percentage (TS%):

A measure of shooting efficiency that accounts for the value of 3-pointers and free throws.

TS is calculated as:

TS% = PTS / (2 * (FGA + 0.44 * FTA))

PTS: Points
FGA: Field Goals Attempted
FTA: Free Throws Attempted

e) eFG%:

Effective Field Goal Percentage (eFG%), is a metric adjusted for the fact that a 3-point field goal is worth more than a 2-point field goal.

eFG% is calculated as:

eFG%= (FG + 0.5 * 3P) / FGA

FG :	Field goals made
3P:	3-pointers made
FGA:	Field goal attempts

f) Composite Score:

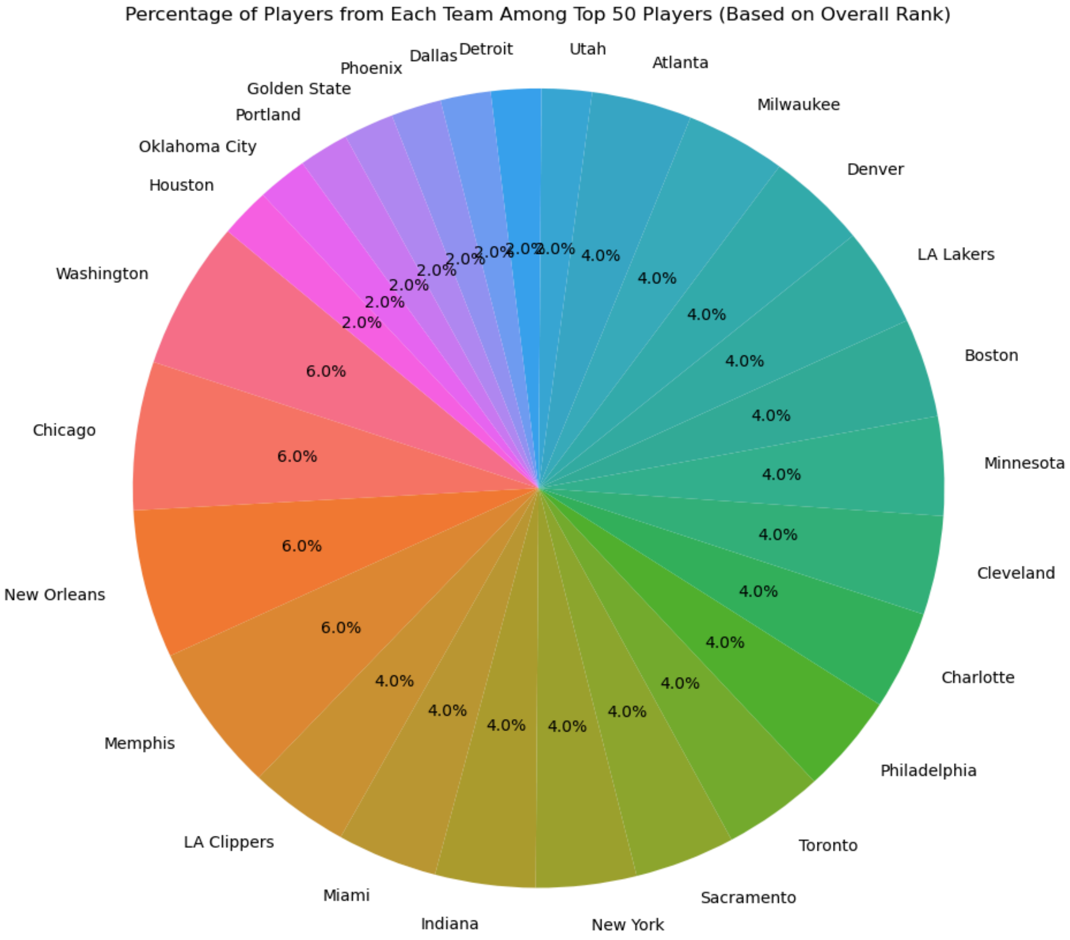
This score is computed using above calculated advanced metrics, giving equal weightage to each metric. The composite score formula is as below:

CS=FP + PER + TS + eFG

Visualizations:

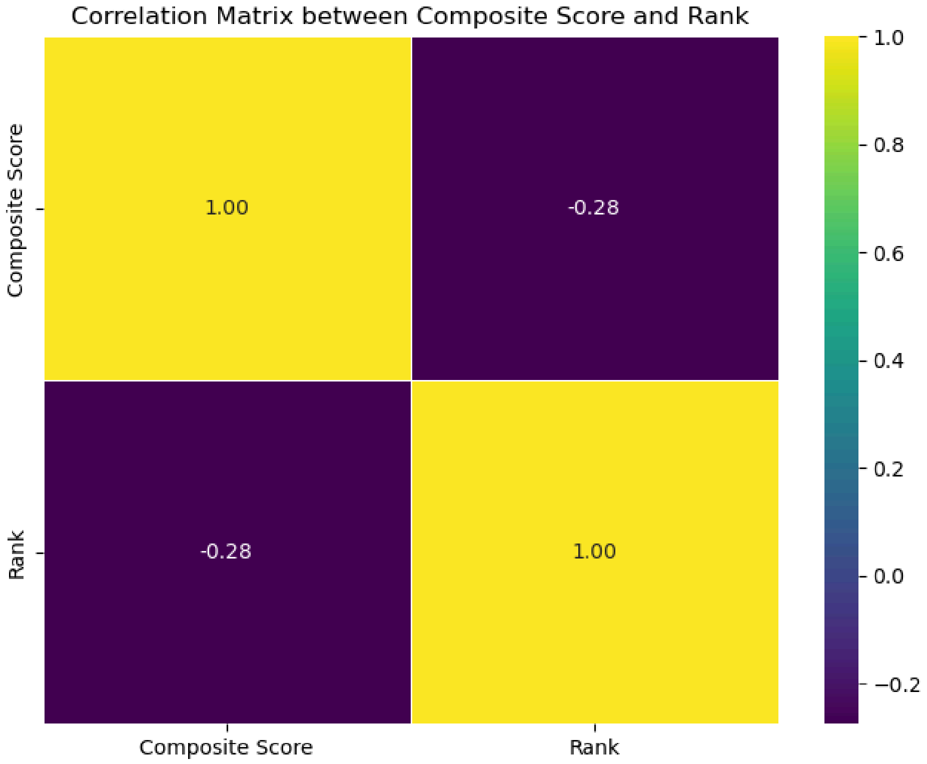
A) Visualization of percentage of contribution from each team in Top 50 Players.

This is a pie chart that demonstrates the percentage of players from each team among top 50 players. This is calculated based on the overall rank, that is assigning a rank to each player based on their composite score(as mentioned previously in the report).



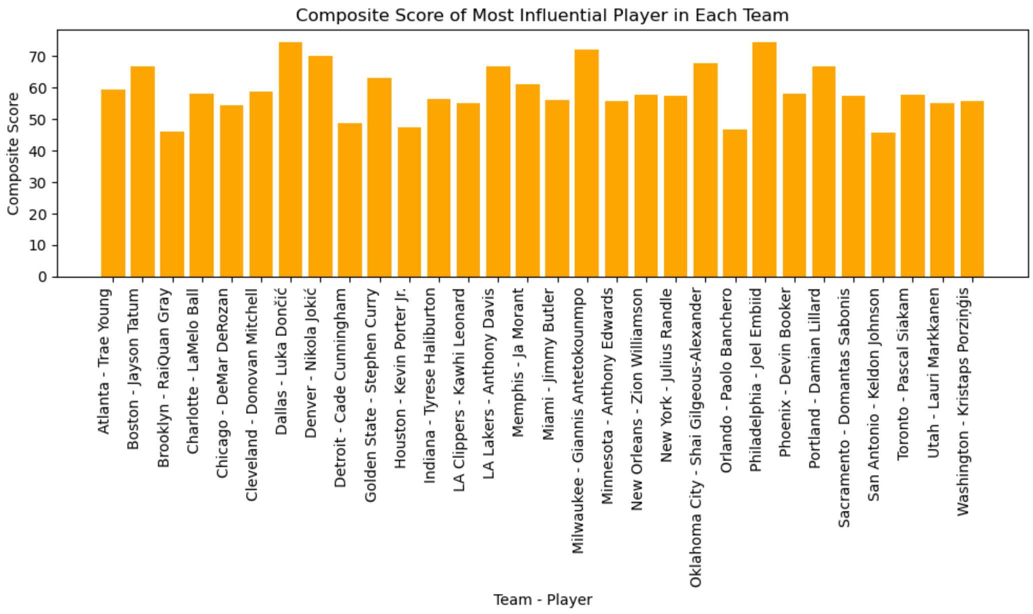
B) Correlation between Composite score and Rank:

A correlation matrix is a table showing correlation coefficients between variables. In this case the variables are the Cumulative Composite Score of a team (this is calculated by taking the sum of composite scores of top 5 players in the team) and team rank.



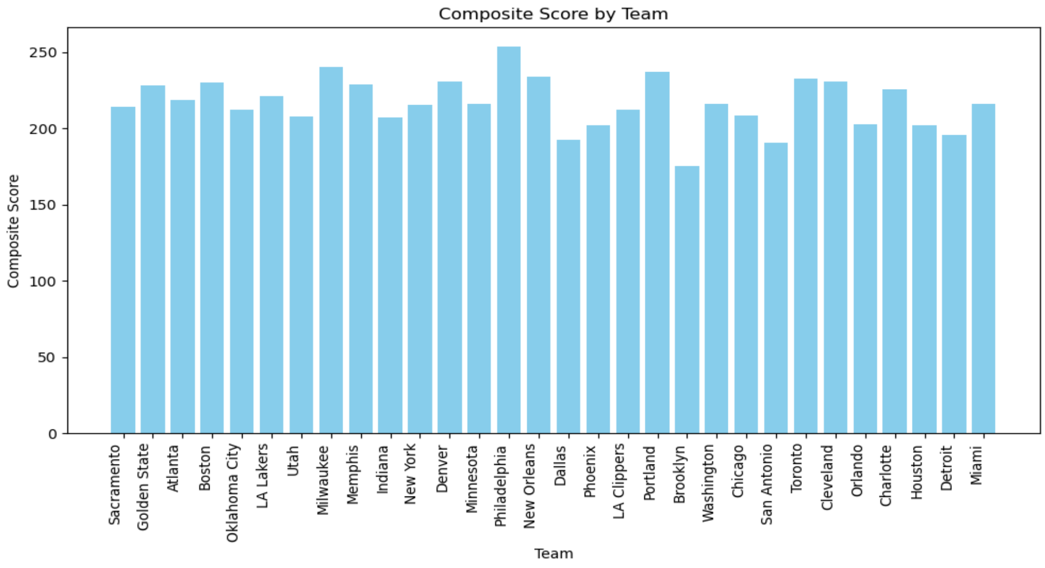
C) Bar Plot of Composite Score of Most Influential Player in Each Team.

This is a bar plot of the most influential player in each team and their composite score. The influential player is determined by its rank in the team.



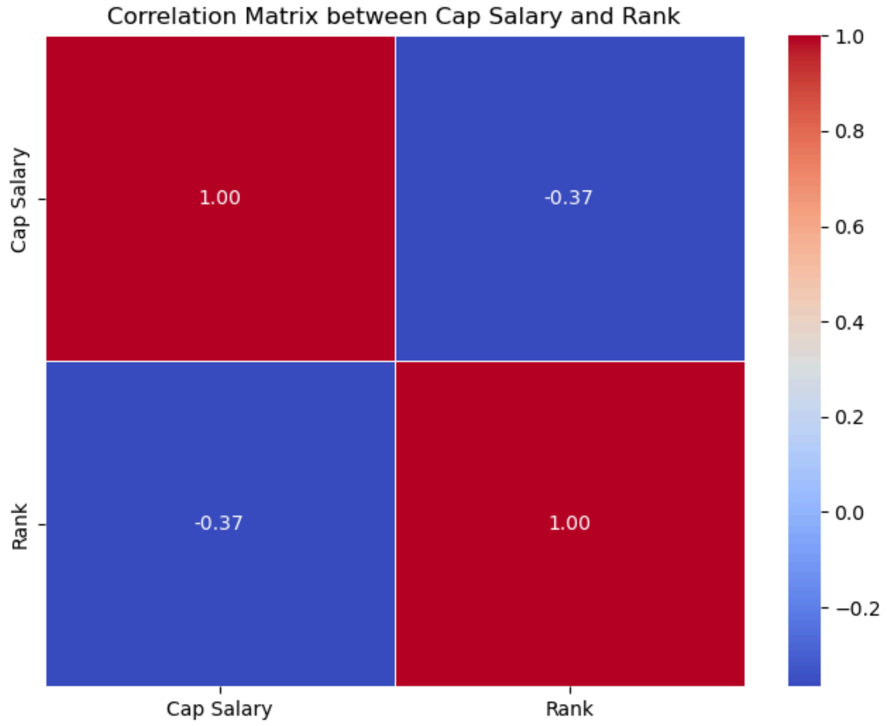
D) Bar Plot of Composite Score of Team.

This is a bar plot of the teams with their respective composite score (his is calculated by taking the sum of composite scores of top 5 players in the team).



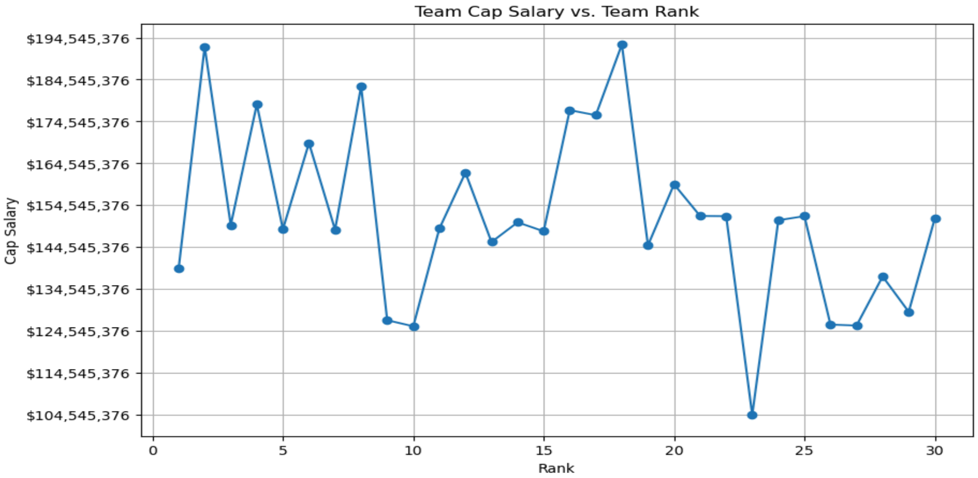
E) Correlation between Composite score and Rank:

A correlation matrix is a table showing correlation coefficients between variables. In this case the variables are the Cap Salary of a team and team rank.



F) Line Plot of Team Cap Salary vs Team Rank

This plot is plotted using the two variables that are the Team Cap salary and Team Rank.

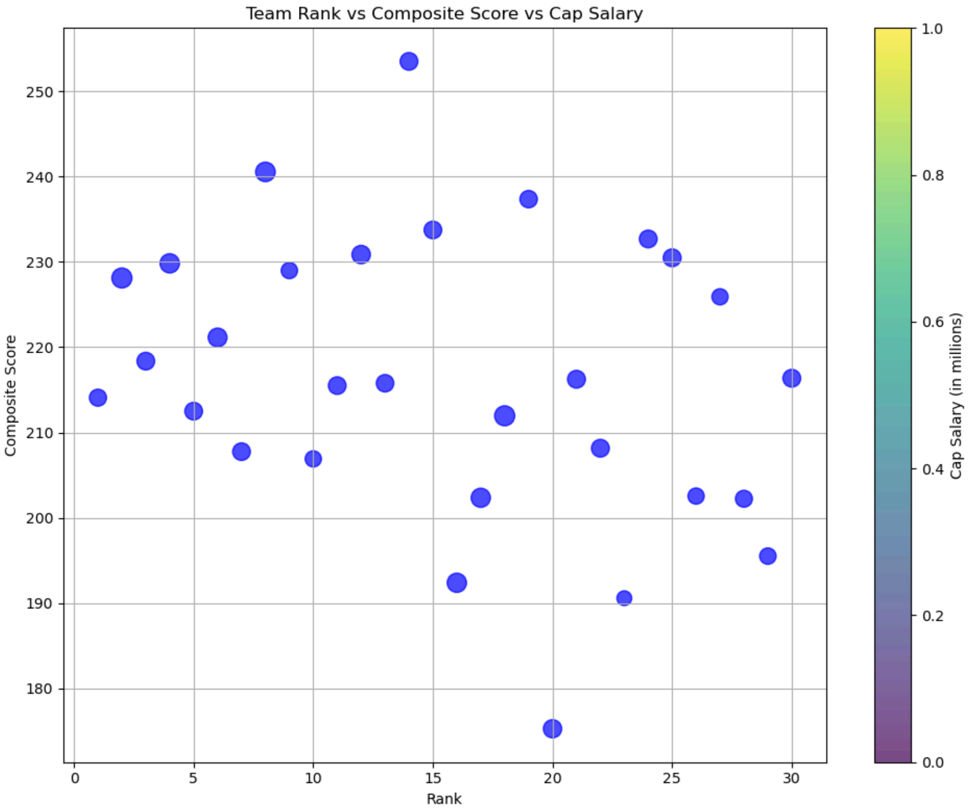


G) Scatter Plot

This is a scatter plot between the three variables that is team rank, composite score of team and Cap Salary.

- The x-axis represents the team's rank.
- The y-axis represents the composite score.
- Each point on the plot represents a team.
- The size of each point corresponds to the team's cap salary, with larger circles indicating higher cap salaries.

This plot visually displays the relationship between team rank, composite score, and cap salary, allowing for an understanding of how these variables are related across different teams.



Conclusions:

The analysis of player performance based on computation of advanced performance metric and then eventually calculating composite score for each player led to determining the most influential player in each team. Analysis of correlation between two parameters that is the cumulative composite score of team and team rank suggest that there is a moderate negative correlation between the parameters. Similarly, correlation between Cap salary and team rank also indicates a moderate negative correlation which in turn means that as the team rank is high, the Cap salary would also be higher. Line plot and scatter plot are used for visualizing the same relations.

Future Work:

Given more time, I would incorporate more advanced metrics in determining player performance. I would also build a prediction model to determine the success rate of individual teams for upcoming NBA seasons.