**Data Analytics for NBA statistic and Tweet**

**Background:**

NBA has a great influence in the sports field, and the performance, salaries of players are always the very hot topics discussed in our daily life. As fans, we are all like to know more about players' careers after analyzing statistics. For team managers, they would like to know whether a single penny they spend is worthwhile. So, we try to find the relation between players' performance and their salaries, and of course, the subtle influence of their tweets to their performance on the pitch.

**Data Science Problem:**

(We will choose 2 or 3 questions for our project from here.)

1. The GOAT

The greatest of all time based only on statistics on the pitch.

2. The most/least cost-effective player

The ratio of their salaries to their performances.

Who is the one earning less but providing a solid performance?

And who is the one having a huge contract but help the team very little?

3. Career turning point

As age increases, a player's performance inevitably goes downhill.

We'd like to use the player's statistics to predict from when the statistics of Lebron James, James Harden or Stephen Curry will go downstairs.

4. Tweet and NBA

When posting something on Tweet, some players try to show their life from various aspects, and others focus more on their jobs like training and matches.

Is there any connection between the tweets' contents and their performance on the pitch?

Will these players focus more on training have a better performance than those who have a colorful life?

**Data Collecting:**

Data source:

www.basketball-reference.com

www.twitter.com

**Data Set:**

1. Player statistic

We try to get all the players in NBA history and their seasonal/career statistics.

1.1 Player Seasonal statistic

Raw record QTY: 29,620

Attributes: 31

|  |  |  |
| --- | --- | --- |
| Name | Season | Age |
| Team | League | Position |
| Game Played | Game Started | Minutes Per Game |
| Field Goals Per Game | Field Goal Attempts Per Game | Field Goal Percentage |
| 3-Point Field Goals Per Game | 3-Point Field Goal Attempts Per Game | 3-Point Field Goal Percentage |
| 2-Point Field Goals Per Game | 2-Point Field Goal Attempts Per Game | 2-Point Field Goal Percentage |
| Effective Field Goal Percentage | Free Throws Per Game | Free Throw Attempts Per Game |
| Free Throw Attempts Per Game | Offensive Rebounds Per Game | Defensive Rebounds Per Game |
| Total Rebounds Per Game | Assists Per Game | Steals Per Game |
| Blocks Per Game | Turnovers Per Game | Personal Fouls Per Game |
| Points Per Game |  |  |

1.2 Player career statistic

Raw record QTY: 18,695

Attributes: 31

|  |  |  |
| --- | --- | --- |
| Name | Seasons Played | Team |
| League | Game Played | Game Started |
| Minutes Per Game | Field Goals Per Game | Field Goal Attempts Per Game |
| Field Goal Percentage | 3-Point Field Goals Per Game | 3-Point Field Goal Attempts Per Game |
| 3-Point Field Goal Percentage | 2-Point Field Goals Per Game | 2-Point Field Goal Attempts Per Game |
| 2-Point Field Goal Percentage | Effective Field Goal Percentage | Free Throws Per Game |
| Free Throw Attempts Per Game | Free Throw Attempts Per Game | Offensive Rebounds Per Game |
| Defensive Rebounds Per Game | Total Rebounds Per Game | Assists Per Game |
| Steals Per Game | Blocks Per Game | Turnovers Per Game |
| Personal Fouls Per Game | Points Per Game |  |

3. Player salary

Players' annual salaries

QTY: 14,932

Attributes: Player's name, Season, Team, League, Salary

4. Tweets

This dataset includes both tweets posted by players' and teams' official accounts.

4.1. Timeline dataset:

QTY: 45,484 (Including 8 players and 9 teams.)

Attributes: screen\_name, time, hashtags, tweets text, user\_mentions, re-tweet count, favorite count.

4.2. Keywords dataset:

QTY: 15,160

Attributes: time, key\_words, hashtags, tweet text, user\_mentions, re-tweet count, favorite count.

**Data cleanliness and cleaning:**

1. Player Statistics

1.1 Player Seasonal Statistic

Remove the records with a missing value.

Error record QTY: 10,007

Raw record QTY: 29,620

Error rate: 33.8%

1.2 Player Career Statistic

Since we need one player's statistic of his entire career, we remove the career statistic of serving specific teams.

Error record QTY: 15,721

Raw record QTY: 18,695

Error rate: 84.1%

2. Player Salaries dataset

Some players' salaries are shown '<Minimum' or 'None', we take them as error data and remove them from our records.

Error record QTY: 118

Total record QTY: 14,932

Error rate: 0.8%

3. Tweet dataset

Using Tweet API to fetch the data, so there is no None value.

a. Timeline dataset:



b. Keyword dataset:



**Group members and roles in project phase1:**

|  |  |
| --- | --- |
| Member | Roles |
| Lu Sun (ls1377) | Tweet timeline and keyword datasets |
| Yijun Gan (yg270) | Players' statistics and salaries datasets. |
| Shuyang Yu (sy614) |
| Bo Zhou (bz166) | Document integration |