**analysis of NBA data**

Junyan Chen

**Research Backgrounds:**

The NBA (National Basketball Association) is widely considered to be the premier men’s professional basketball league in the world. This project is trying to analyze NBA player data.

**Dataset Description:**

All of our data is from NBA China official website: <https://www.basketball-reference.com/>

**Data Preparation:**

I use python for doing Data Cleaning, Data analysis and Data Visualization. I used the code in the attachment to collect player information**.**

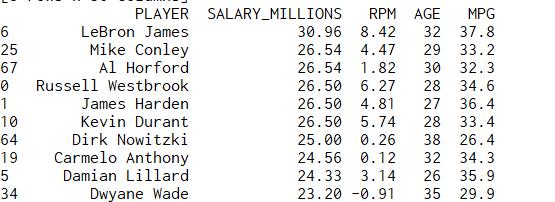
**Methodology**:

Before we load the data, we need to understand what each item means

|  |  |
| --- | --- |
| Rk | Rank |
| Player |  |
| Position |  |
| Age |  |
| MP | Minutes played per game |
| FG | Field goals per game |
| FGA | Field goal attempts per game |
| FG% | Field goal percentage |
| 3P | 3-point field goals per game |
| 3PA | 3-point field goal attempts per game |
| 3P% | 3-point field goal percentage |
| 2P | 2-point field goals per game |
| 2PA | 2-point field goal attempts per game |
| eFG% | Effective field goal percentage |
| FT | Free throws per game |
| FTA | Free throw attempts per game |
| FT% | Free throw percentage |
| ORB | Offensive rebounds per game |
| DRB | Defensive rebounds per game |
| TRB | Total rebounds per game |
| AST | Assists per game |
| STL | Steals per game |
| BLK | Blocks per game |
| TOV | Turnovers per game |
| PF | Personal fouls per game |
| Points | Points per game |
| Team |  |
| GP |  |
| MPG | Minutes per game |
| ORPM | Offensive real plus minus |
| DRPM | Defensive real plus minus |
| RPM | Real Plus minus |
| Wins\_RPM | Winning game real plus minus |
| PIE | shows what % of game events did that player |
|  | or team achieve |
| Pace | used to estimate the number of possessions a |
|  | team has per game |
| W |  |
| Salary\_million |  |

**Player data analysis section**

Top 10 highest-paid players



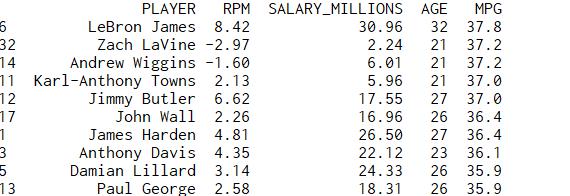
Lebron James was the highest-paid player of the season, and McConley got a big contract, but the star-studded salary list pales in the shade. Also on the list are Westbrook, Harden, Durant and others, Curry was not in the top 10 because his previous contract was too small.

Top 10 highest-efficiency value players



James is the highest paid player in the league, plays without ambiguity, and ranks first in efficiency. Paul and Curry were close behind, with Warriors occupying three of the top 10 spots. It's worth noting that Denver's Jokic and Utah's Gobert, both earning modest salaries, have been among the league's top 10 most efficient players, setting the stage for their next big contracts.

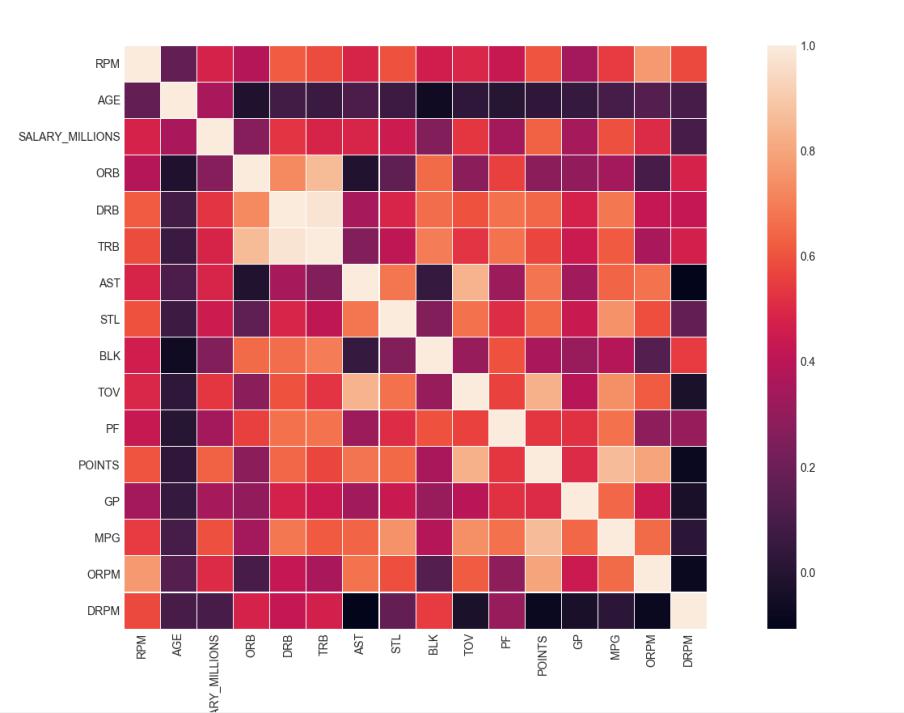
Top 10 highest-playing time player



James rank the 1st in list again, followed by Lavine and Wiggins, with the Timberwolves making up three of the top 10.

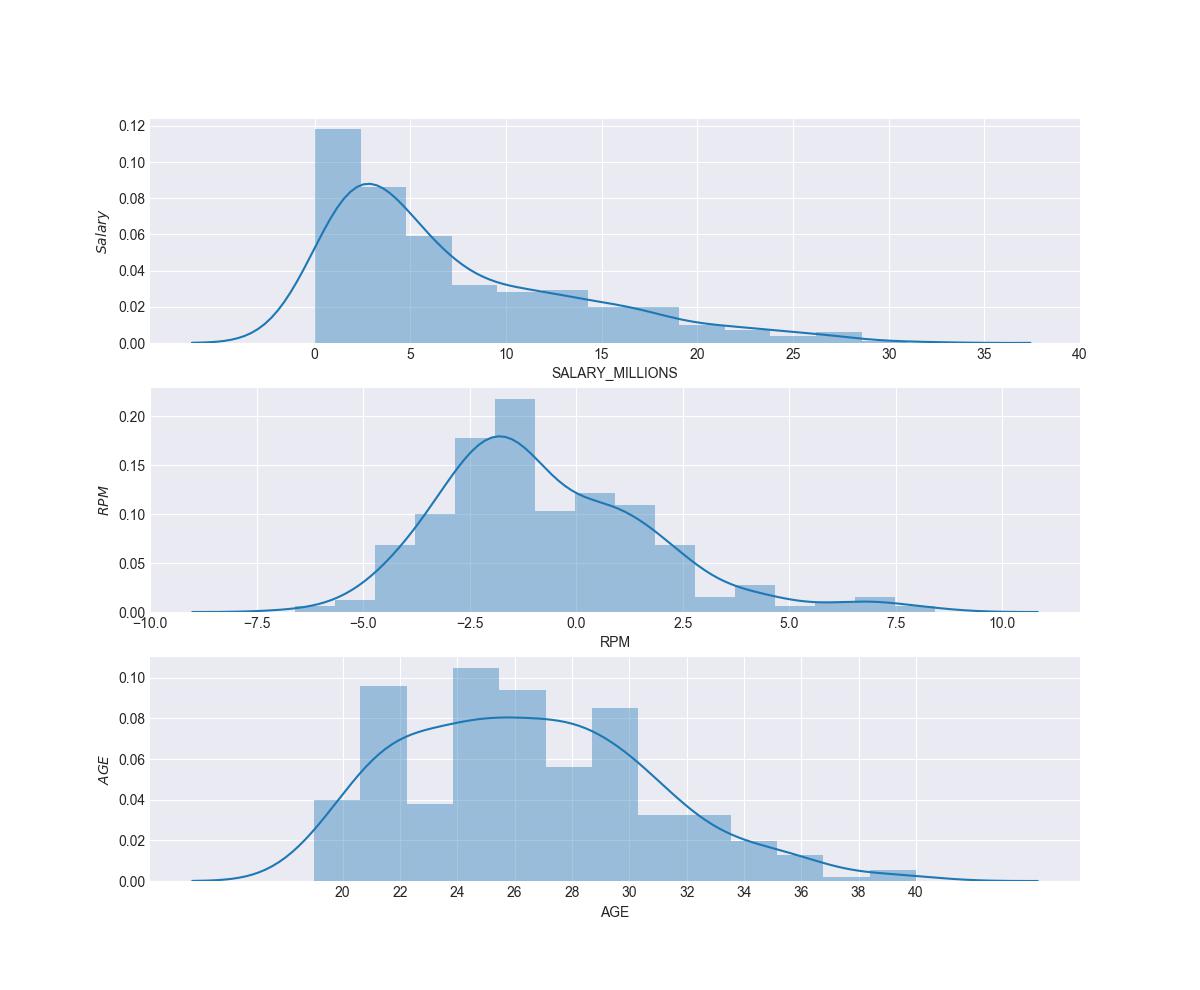
**Correlation analysis of efficiency value**

Among the numerous data, there is an item named "RPM", which indicates the efficiency value of players. This data reflects the contribution of players to the team's victory in the game when they are on the spot, and it can best reflect the comprehensive strength of players. Let's see how it relates to other data:



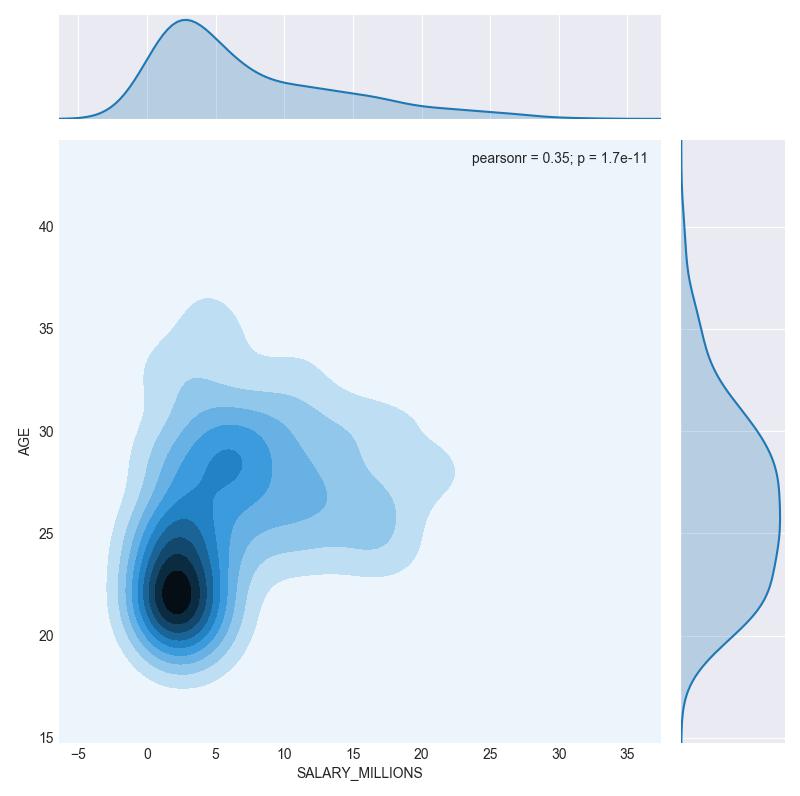
It can be seen from the heatmap of the correlation analysis that the RPM value has the weakest correlation with age, and it has the strongest correlation with the game technical data such as "attack efficiency value", "average score per game" and "average steals per game".

**Seaborn method analyzes a single attribute**



It can be seen that the age and efficiency values are more in line with the normal distribution, while the salary of players is more like a skewed distribution, with a smaller proportion of high-paid players.

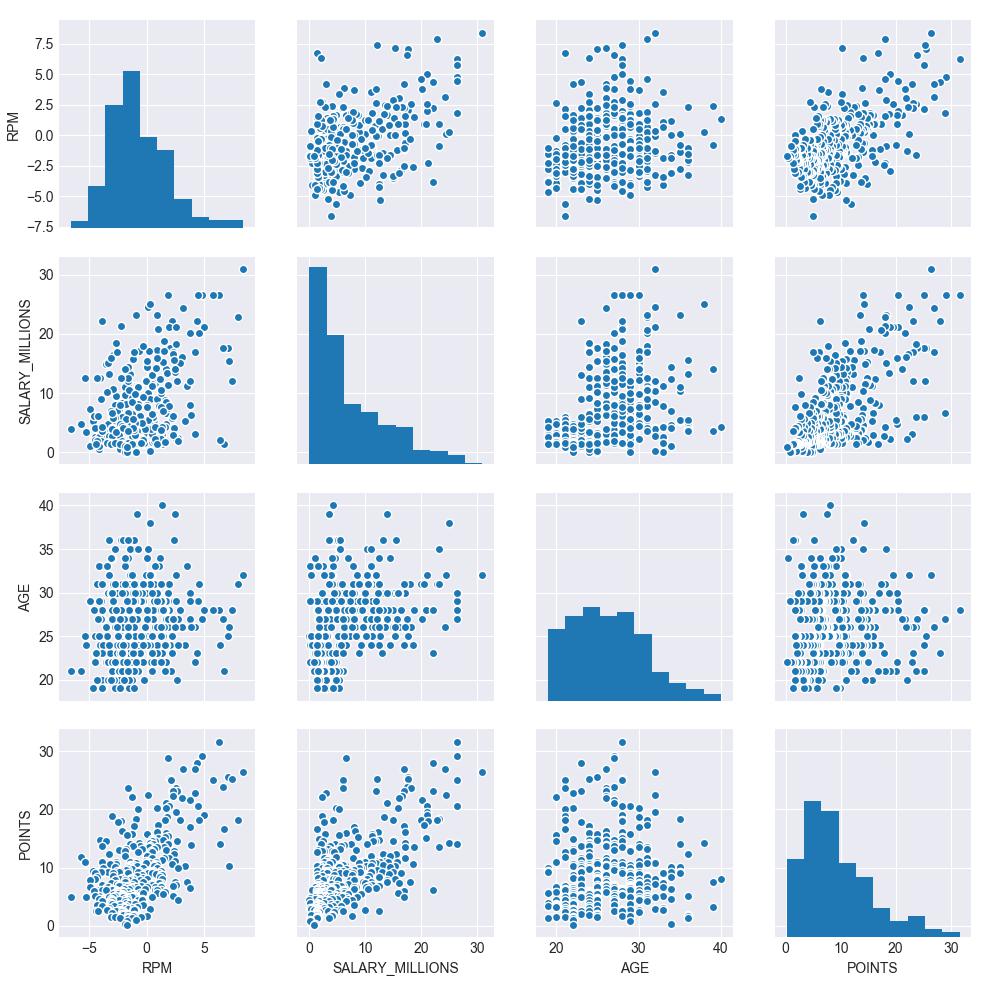
**Seaborn method analyzes two attributes**



The figure above shows the relationship between players' salary and age. We can generally feel the concentrated characteristics of age and salary.

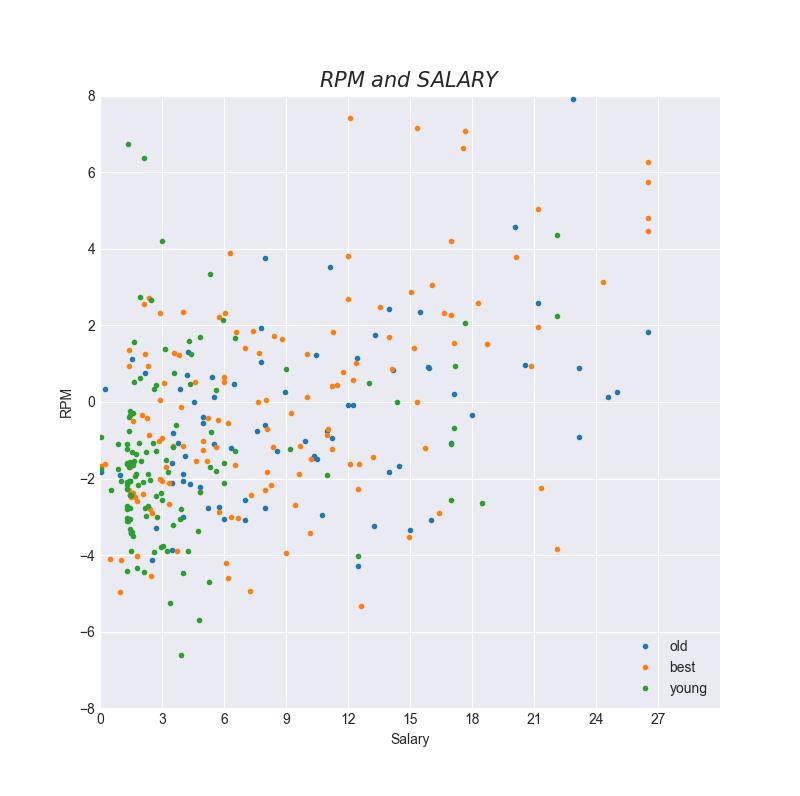
**Seaborn method analyzes multivariate attributes**

Let's first use the distplot plot in seaborn to see the distribution of player salary, efficiency value and age



The figure above shows the pin-wise correlation between the four variables of player salary, efficiency value, age and average score. The diagonals show the distribution diagram of players themselves, and we can see the correlation degree of different features from the trend of scatter. On the whole, the correlation of all dimensions is not very strong. Positive and negative values have a weak positive correlation with salary and field average score, while age has a weak correlation with other variables.

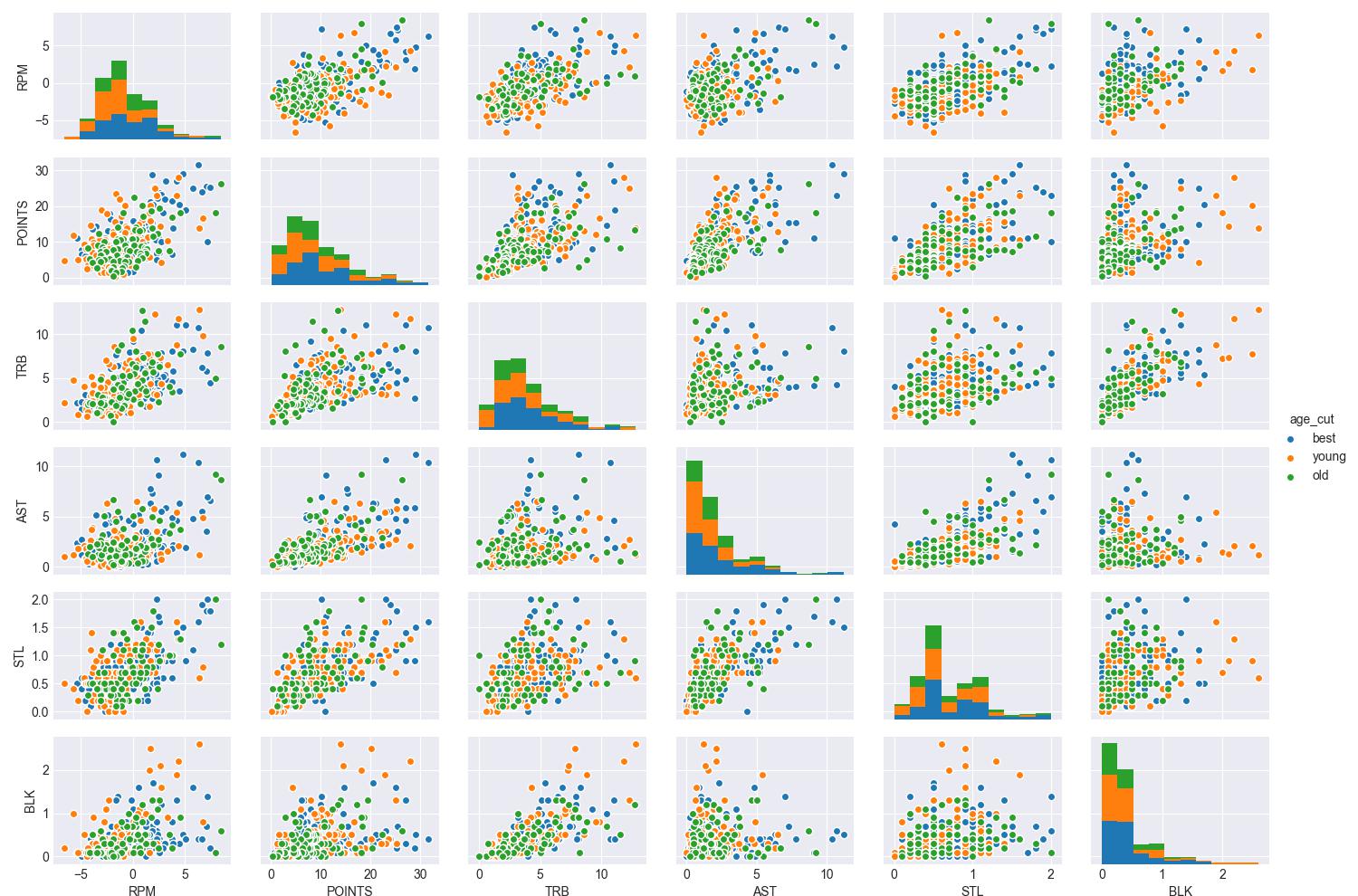
Then I divided the players into three generations according to their ages. Now that we have obtained the label of three generations of senior, middle and young players, let's take a look at the relationship between RPM (positive and negative) of players of different ages and salary before:



The x of the graph is player salary, and the y is efficiency value. It can be observed that:

* The vast majority of young players are paid less and the data is very concentrated. There are two outliers.
* The data of best age players and old players are relatively divergent, and the positive correlation between best age players' salary and efficiency value is stronger.
* Older players are older and less active, and slightly more "high-paid, low-efficiency" players.

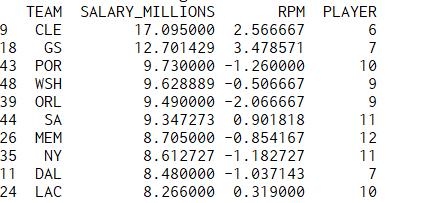
Then I Use the method of the previous to see the distribution of the three generations of technology statistics:



**Team data analysis**

***Team salary ranking***

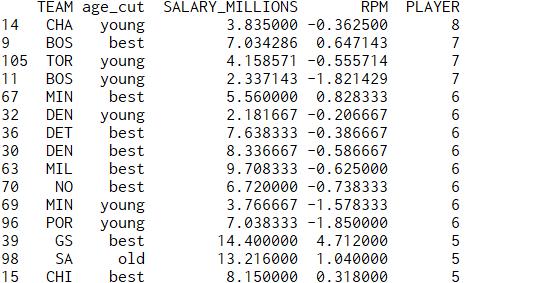
Group the data into teams, rank the average salary in descending order, and take a look at the top ten richest teams in the league:



* The cavaliers and warriors, who are both in the top two on this list for their high salaries, have been star-studded enough to fight their way through the playoffs to the division finals.
* The blazers, who finished third with 10 players on the list, are a dynamic addition. The health of the team's salary structure is crucial to the development of the team.

***Age structure of the team***

I ranked the players on the list in descending order according to the age group of the team. If the number of players on the list is the same, the players on the list will be ranked in descending order according to the efficiency value.



* The Hornets, at the top of the table, have eight young players but low efficiency.
* Boston Celtics is gorgeous, the best age players and young players a total of 14, who have high efficiency value.
* The young Timberwolves have six best age players and the old spurs have five older players

**Conclusion**

With the progress of science and technology, we can better record and analyze the data of basketball games, which enables us to better understand basketball and players, combining our professional knowledge and interests, and better enjoy the infinite charm of basketball games.