**NBA team and players data analyze**

**Shuo Wang**

**Project URL：**<https://github.com/Alexander951006/CSE482>

**Abstract：**

This project is named NBA team and players data analyze System. This system is for analyze 5 NBA basketball teams and lots of players come from different basketball teams to track them the whole data of 2019-2020 season. Study the team's performance in reality from the game data.

**1.Introduction：**

This project is based on the visualization of NBA teams and players data.When I browse the statistics of teams and players performance,All data is presented in chart form, And player data and team data are often separate, it’s hard to read sometimes. It is difficult to visually see the relationship between the player's performance and a team and it is difficult to make an accurate analysis. I picked Philadelphia 76ers, Oklahoma City Thunder, Milwaukee Bucks, Brooklyn Nets, Memphis Grizzlies as team test sample and James Harden HOU, Bradley Beal WAS, Giannis Antetokounmpo MIL, Trae Young ATL, Damian Lillard POR as players as test sample. The biggest difficulty in this analysis is that the team and player data are often very complicated. And players' personal and data are often separate from the team's data, sometimes it is difficult to see the connection. There are other factors that affect the game, such as the player's mental state, injuries and other factors

**2.Data:**

I found data form online and I download them.

For teams data:

<https://stats.nba.com/teams/traditional/?sort=W_PCT&dir=-1>

More teams data:

<https://stats.nba.com/teams/>

For players data:

<https://stats.nba.com/leaders/>

More players data：

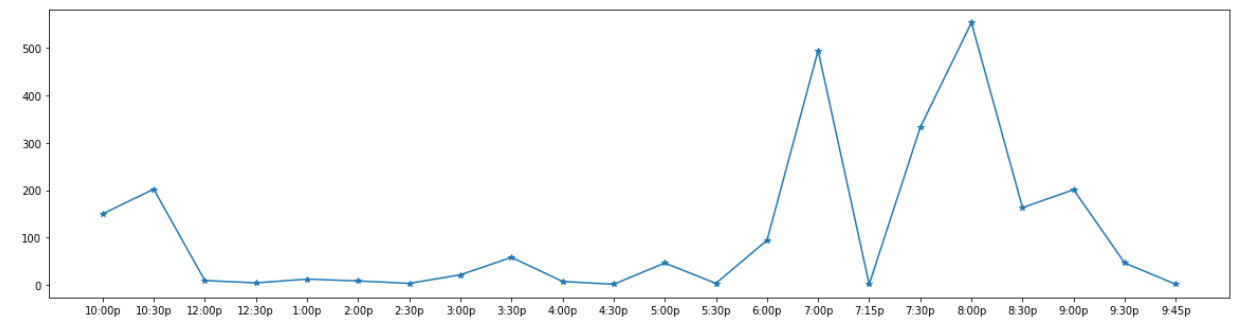
<https://stats.nba.com/players/>

Chris Paul data:

<https://www.basketball-reference.com/players/p/paulch01.html>

All data came from those website above. I collect those I divided these data into team data and player data. I sort the teams according to the date of the match,Sort players by score.The data I collected came from two different seasons. One was from 2019’s schedule and results the other one was from 2020’s schedule and results. I merged them into one file to analyze them. And some data are unnecessary to analyze such as attend notes and time.

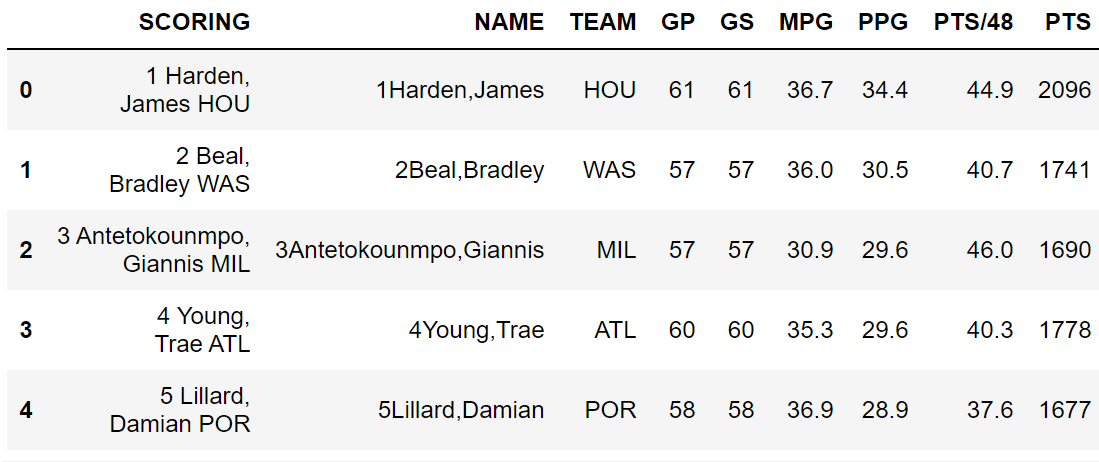
In my preprocessing steps, I collect the data of the game according to the schedule. Then I merged all matches data into one csv file. First I Used matplotlib to read raw data looking for the most time when matches take place.

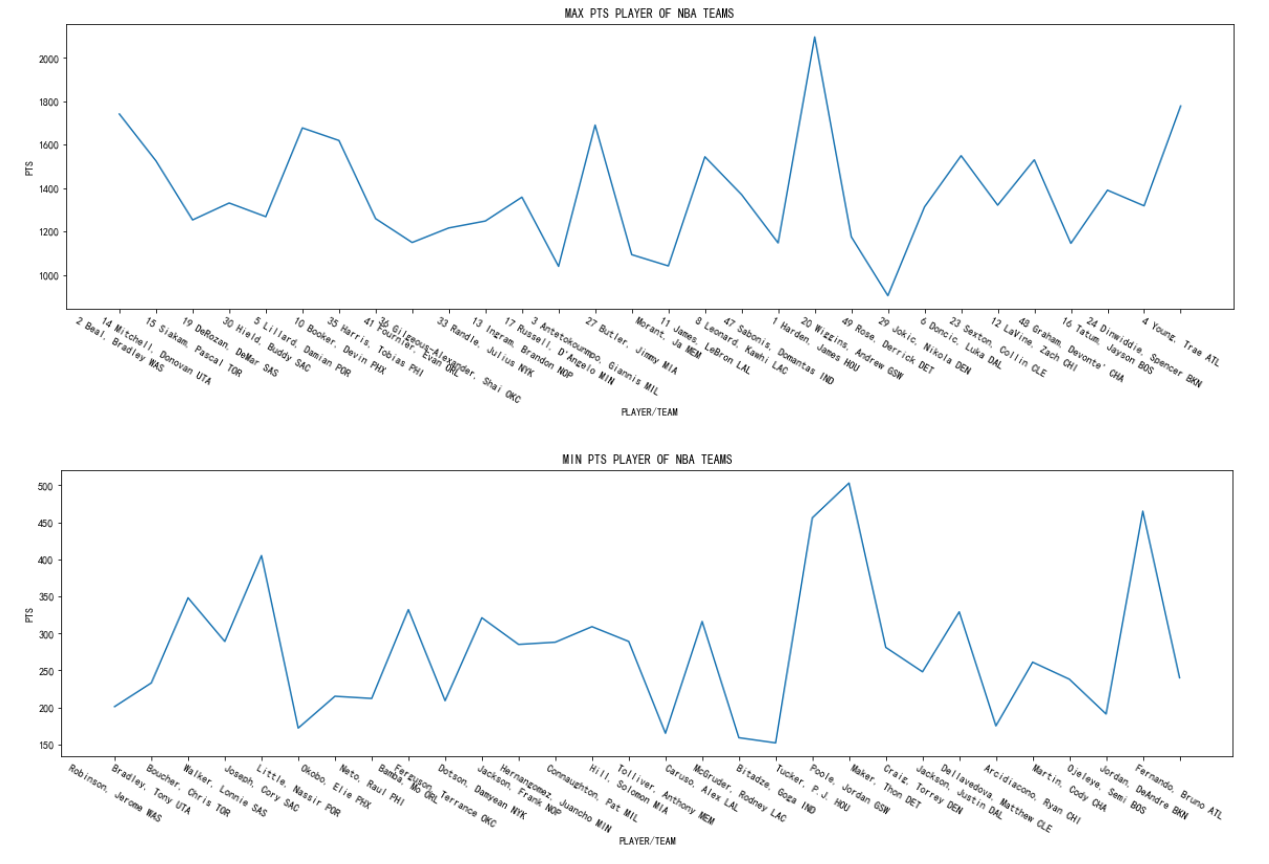


I got that the first peak is at 7pm. Around 500 matches happen at this time. And the highest peak is at 8pm. Over 500 matches took place.

For player the first thing to do is split scoring into two columns: name and team

and first 5 players to read if spliting worked.:

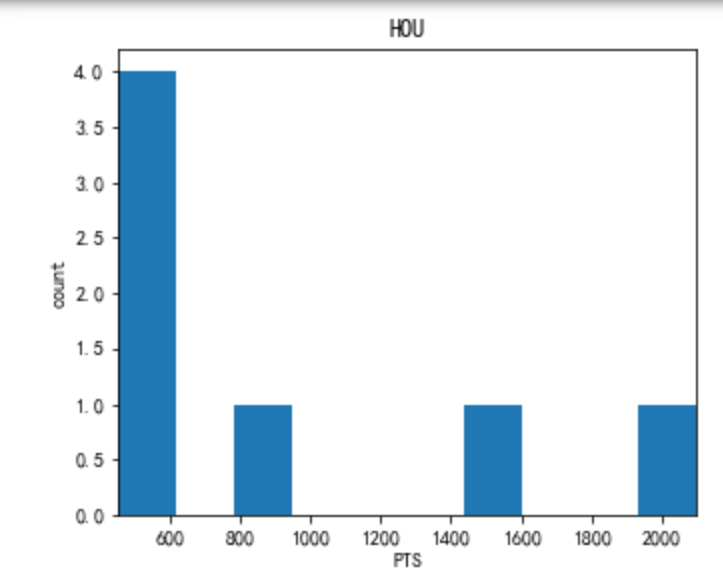


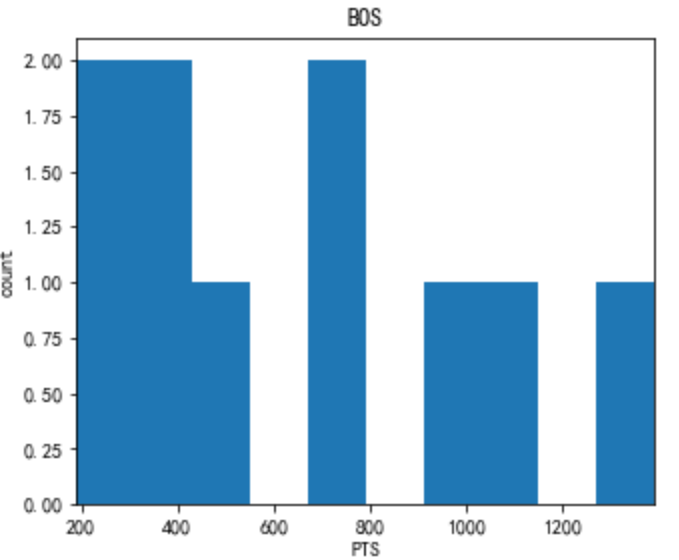
After we know the split works we compared of the highest-scoring player scores of each team. And compared of the lowest scoring player scores of each team.:

**3.METHODOKOGY:**

This is a data visualization analysis, So I collected as much data as possible to make the image.

I used bar charts and line charts to show the division and changes of data more intuitively.Player data can often determine the data of a whole team. For example the James Harden in Huston Rockets.The more players with better data, the stronger the team. Like the Golden State Warriors and Boston Celtics.





This is the scoring distribution of Boston Celtics and Houston Rockets players.Because Harden scored more than 2000 points,So the Rockets are strong.But the Celtics are more balanced so the Celtics are also strong and more stable than the Rockets.

I used 3 files to complete this project. One is main\_download\_data.jpynb this is for download the data of NBA teams. For analyzing the data of NBA teams I used main\_mdl.jpynb. The file named main\_eda.jpynb I used this for player’ data. The file named main\_arima.ipynb is I use it to predict Chris Paul’s data to 2024.

# 4.EXPERIMENTAL EVALUATION

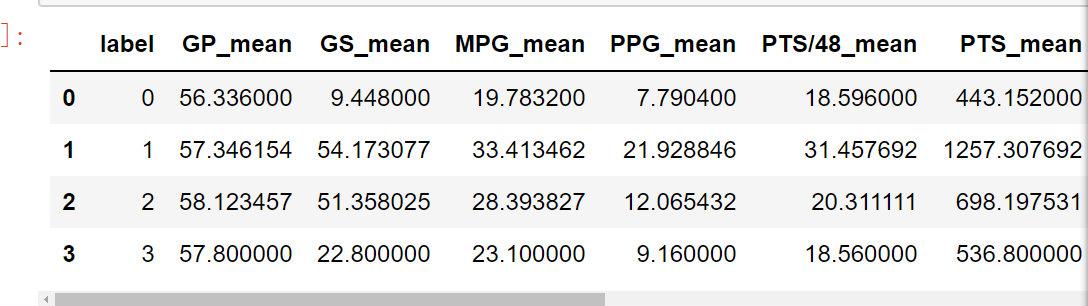
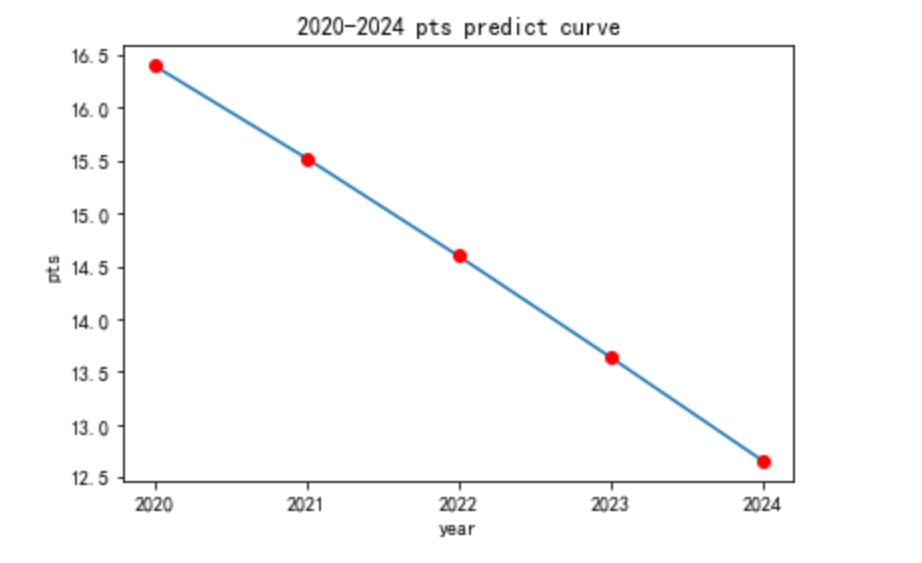
4.1 Experimental setup

I used jupyter notebook to write all work. I am using the accuracy as result and k—SSE curve to evaluate this project.

4.2 Experimental Results

I am using predict curve to predict Chris Paul from now to 2024. We can tell his data

His data is showing a downward trend. For players I am using clusters to divide the average data in the alliance. For Teams I got Performance of the prediction set. The AUC:0.505524225012414 and ACC:0.5957193816884662.



# CONCLUSIONS

In the process of doing this project, I realized that the accuracy and authenticity of the data are often very important.Back to the subject,The influence of the player team is very important.Some teams have very good star players so this team is very strong such as James Harden in HOU. But for some teams, the data for each player belongs to the upper middle level.So this team is also strong and more stable.However, there are often effects on the team that cannot be analyzed. Such as player injuries,

Team influence in history and so on. For players, as his age increases, his data will show a downward trend.

1. **REFERENCES**
2. ”NBA STATS SEASON LEADER”,26 Apr. 2020,https://stats.nba.com/teams/.
3. ”NBA STATS Advanced Stats”,26 Apr. 2020,https://stats.nba.com/teams/traditional/?sort=W\_PCT&dir=-1
4. ”NBA STATS SEASON LEADER”,26 Apr. 2020,https://stats.nba.com/player/.
5. ”NBA STATS Advanced Stats”,26 Apr. https://stats.nba.com/leaders/
6. “BASKETBALL REFERENCE”26 Apr, <https://www.basketball-reference.com/players/p/paulch01.html>

Grading criteria

Note that the project accounts for 10% of your final grade. The project will be graded based on the following criteria:

1. Presentation - structure/organization and clarity of writing (including tables and figures).
2. Technical - Correctness and thoroughness of the analysis performed. What are the challenges faced and how well did you address them? How do you evaluate the performance of the method you'd applied to the data? How much detailed discussion you provide to explain the results you'd obtained (e.g., discussion about why the method works or didn't work on the data)?
3. Difficulty level - How large is the dataset used? How much effort you had to spend to collect, integrate, preprocess, and analyze the data? Are you implementing the project on a cluster or a single machine? What tools did you use (do you have to implement them or are you simply using existing libraries)?