## Part A

1. **https://www.basketball-reference.com/players/a/**. For each player listed on this page, my crawler will enter his detail page and acquire more information if his last year is larger than 2009. For each table in player’s detail page, I set one specific crawler to get his information using the method aforementioned. **https://www.basketball-reference.com/teams/**. For each team listed on this page, my crawler will enter his detail page and acquire more information. For each table in team’s detail page, I set one specific crawler to get his information using the method aforementioned.
2. Here is a part of my code, you can see it in detail under folder **/spider/NBA/spiders*.***



1. For each player, I parse his basic information, performance per game, total performance, performance per 36 minutes, performance per 100 poss, advanced performance, adjusted shooting data, play-by-play data, shooting data, game highs, playoff series and his salaries. There are 12 tables in total. You can find them in fold **/data.**
2. For each team, I parse basic team information, index of seasonal statistics, basic seasonal statistics, basic seasonal statistics of opponent, seasonal statistics ranks, seasonal statistics ranks of opponent, seasonal statistics year by year, seasonal statistics year by year of opponent. There are 9 tables in total. You can find them in fold **/data.**
3. Salary cap information is useful when evaluating players’ salary(https://www.basketball- reference.com/contracts/salary-cap-history.html). The reason is that salary cap is not only an indicator of upper limit of salary, it also shows the average level or the development of the league and nation’s economy. If salary cap is high, it means that players earn more as a whole. The code is showed below which is in **salarycap\_spider.py**.

## Part B

1. Data clean and transfrom

Data dictionary

Name

Born

Is\_playoff

Season

Age

Team

Lg

Pos

G

GS

MP

FG

FGA

FG%

3P

3PA

3P%

2P

2PA

2P%

FT

FTA

FT%

ORB

DRB

TRB

AST

STL

BLK

TOV

PF

PTS

Salary

TRB\_p100p

AST\_p100p

PTS\_p100p

TRB\_pergame

AST\_pergame

PTS\_pergame

PER

TS%

WS

VORP

W/L%

SRS

Playoffs

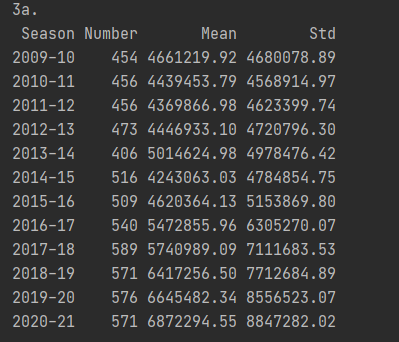
1. a) 571 (Number of players that appears in salary table in 2020-21 season)

b) SG: 199, PG: 139, C: 133, PF: 178, SF: 177

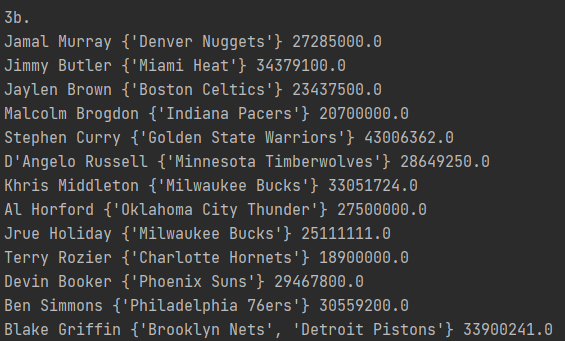
c) average age: 26.20, average weight: 217.16lb, average experience: 4.12 years, average salary: 6872294.55.

d) 22647206.79

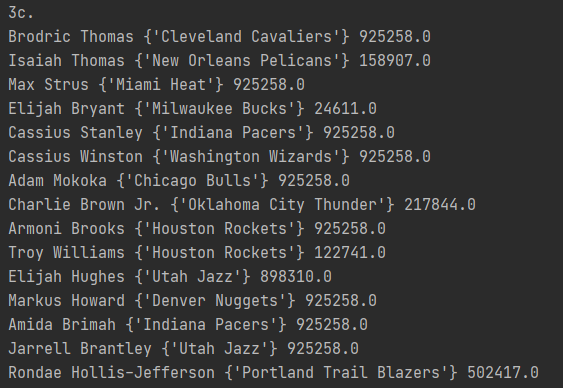
1. a) The result is showed below.



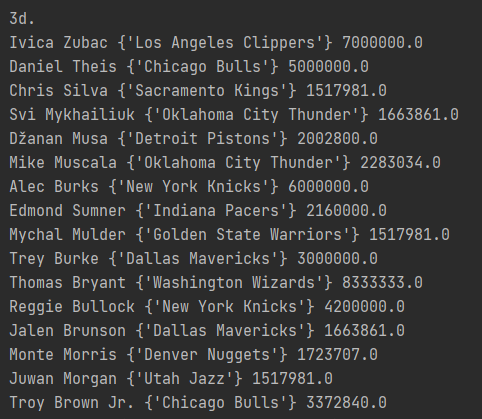
b) Part of the result is showed below. You can see the full answer in **partb.py**.



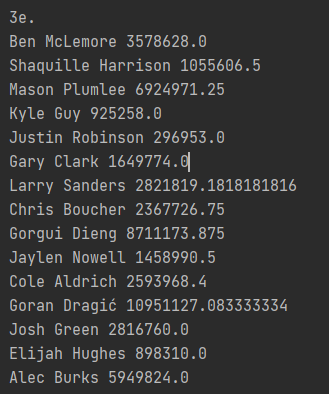
c) Part of the result is showed below. You can see the full answer in **partb.py**. Some of these players have the salary less than minimum. I change it to 925258 according to Internet.



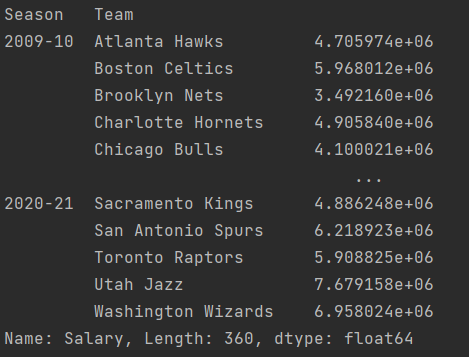
d) Part of the result is showed below. You can see the full answer in **partb.py**.



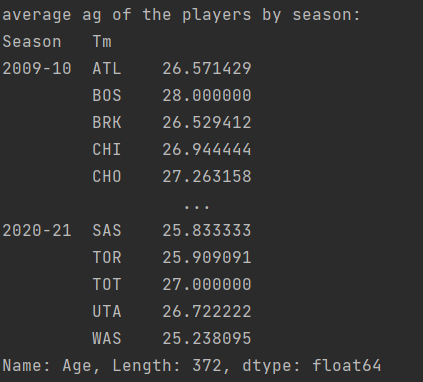
e) Part of the result is showed below. You can see the full answer in **partb.py**.



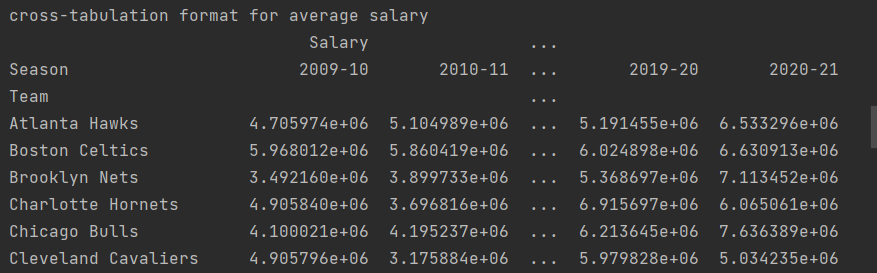
1. a) Part of the result is showed below. You can see the full answer in **partb.py**. And the result file saved in **/output/avg\_salary.csv.**



b) Part of the result is showed below. You can see the full answer in **partb.py**. And the result file saved in **/output/avg\_age.csv** and **/output/avg\_var\_exp.csv**.



c) Part of the result is showed below. You can see the full answer in **partb.py**. And the result file saved in **/output/avg\_salary\_ct.csv, /output/avg\_age\_ct.csv, /output/avg\_ exp\_ct.csv** and **/output/var\_exp\_ct.csv**.



1. Salary cap. The table of summary statistics is showed below.

