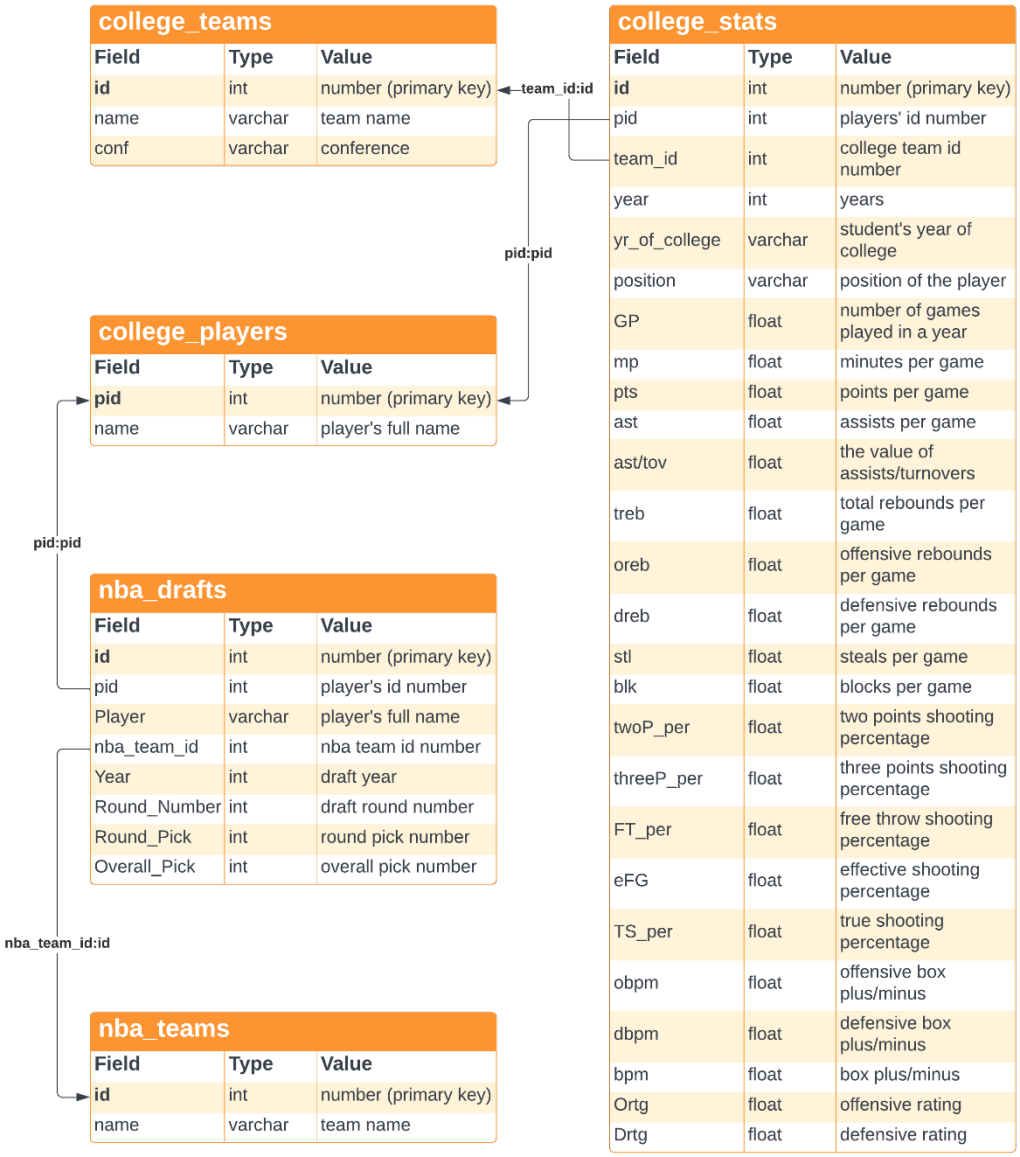
SI 564 Final Project

Tianyi Xiong

Uniqname: xiongty

1. Documentation + ERD

There are five tables in my database college_basketball. Table college_stats includes most of the data with the stats of men's college basketball players from 2009-2021. Table college_players is a connection table which allow me to track the player's name and NBA draft results result using pid. Tables college_teams and nba_teams store the information of the college teams and nba teams.



2. Letter from Raj

Hi,

I'm interested in college basketball stats and the NBA drafts. The database includes the stats of over 4500 US college players from 2009-2021, as well as the NBA draft results in that time period. I wish you can explore the tables and answer the following questions. Unfortunately, international players are not recorded in the college_stats table. Those players without player ids are international players or those who didn't go to college.

Questions:

1) I'm a fan of Los Angeles Lakers, could you give me a list of the player names and draft years of all the players that the Lakers drafted from 2009-2021?
2) Could you tell me the players of which college get the highest average offensive rating and defensive rating in 2015?
3) I'm looking for all-around college players, could you give me a list of players whose average points are greater than 15, average assists greater than 5 and average rebounds greater than 5 during his whole college career?
4) Could you give me a list of all the players drafted by the NBA that are from Michigan? I'm also wondering which college had most students drafted by the NBA from 2009-2021?
5) It seems that younger players are more likely to be drafted at higher picks. Could you show me the proportion of freshman that are drafted among the top 3 picks?
6) I want to compare the top 3 pick players' advanced stats (True shooting percentage, box plus minus, offensive box plus minus, defensive box plus minus) with the average stats and max stats in his draft year. Could you give me a list including all these stats from 2009-2021? I want the average and max data of those main players of each team, so please only choose those who played more than 10 games a year and played more than 20 minutes per game.

Best,

Raj

3. Answers

Hi Raj,

Thanks for reaching out to me. I'm also a basketball fan! And I'm glad to provide the answers.

1) I'm a fan of Los Angeles Lakers, could you give me a list of the player names and draft years all the players that the Lakers drafted from 2009-2021?

   I use the query **select nd.Player, nd.Year from nba_drafts nd join nba_teams nt on nd.nba_team_id = nt.id where nt.name = "Los Angeles Lakers";** and I get a list of 24 players drafted by the Lakers.

```
mysql> select nd.Player, nd.Year from nba_drafts nd join nba_teams nt on nd.nba_team_id = nt.id where n
t.name = "Los Angeles Lakers";
+-----------------------+------+
| Player                | Year |
+-----------------------+------+
| Isaiah Jackson        | 2021 |
| Jaden McDaniels       | 2020 |
| De'Andre Hunter       | 2019 |
| Moritz Wagner         | 2018 |
| Svi Mykhailiuk        | 2018 |
| Lonzo Ball            | 2017 |
| Tony Bradley          | 2017 |
| Brandon Ingram        | 2016 |
| Ivica Zubac           | 2016 |
| D'Angelo Russell      | 2015 |
| Larry Nance Jr.       | 2015 |
| Anthony Brown         | 2015 |
| Julius Randle         | 2014 |
| Ryan Kelly            | 2013 |
| Robert Sacre          | 2012 |
| Darius Morris         | 2011 |
| Andrew Goudelock      | 2011 |
| Chukwudiebere Maduabum | 2011 |
| Ater Majok            | 2011 |
| Devin Ebanks          | 2010 |
| Derrick Caracter      | 2010 |
| Toney Douglas         | 2009 |
| Patrick Beverley      | 2009 |
| Chinemelu Elonu       | 2009 |
+-----------------------+------+
24 rows in set (0.04 sec)
```

2) Could you tell me the players of which college get the highest average offensive rating and defensive rating in 2015?

   I use the queries

   select cs.year, avg(cs.Ortg), ct.name from college_stats cs join college_teams ct

on cs.team_id = ct.id where cs.year =2015 group by cs.team_id order by avg(cs.Ortg) desc limit 1;

select cs.year, avg(cs.Drtg), ct.name from college_stats cs join college_teams ct on cs.team_id = ct.id where cs.year =2015 group by cs.team_id order by avg(cs.Drtg) desc limit 1;

As a result, in 2015, the players of Indiana get the highest average offensive rating, which is 118.34, while the players of the Citadel get the highest average defensive rating, which is 117.85.

```
mysql> select cs.year, avg(cs.Ortg), ct.name from college_stats cs join college_teams ct on cs.team_id = ct.
id where cs.year =2015 group by cs.team_id order by avg(cs.Ortg) desc limit 1;
+------+-------------------+---------+
| year | avg(cs.Ortg)      | name    |
+------+-------------------+---------+
| 2015 | 118.34000066121419 | Indiana |
+------+-------------------+---------+
1 row in set (0.08 sec)

mysql> select cs.year, avg(cs.Drtg), ct.name from college_stats cs join college_teams ct on cs.team_id = ct.
id where cs.year =2015 group by cs.team_id order by avg(cs.Drtg) desc limit 1;
+------+-------------------+-------------+
| year | avg(cs.Drtg)      | name        |
+------+-------------------+-------------+
| 2015 | 117.8509091463956 | The Citadel |
+------+-------------------+-------------+
1 row in set (0.07 sec)
```

3) I'm looking for all-around college players, could you give me a list of players whose average points are greater than 15, average assists greater than 5 and average rebounds greater than 5 during his whole college career?

I use the query

select cp.name, avg(cs.pts), avg(cs.ast), avg(cs.treb) from college_stats cs join college_players cp on cs.pid = cp.pid group by cs.pid having avg(cs.pts)>15 and avg(cs.ast)>5 and avg(cs.treb)>5;

And I get a list of 8 players that meet the requirement.

```
mysql> select cp.name, avg(cs.pts), avg(cs.ast), avg(cs.treb) from college_stats cs join college_players cp
on cs.pid = cp.pid group by cs.pid having avg(cs.pts)>15 and avg(cs.ast)>5 and avg(cs.treb)>5;
+------------------+-------------------+------------------+------------------+
| name             | avg(cs.pts)       | avg(cs.ast)      | avg(cs.treb)     |
+------------------+-------------------+------------------+------------------+
| Greivis Vasquez  | 18.770000457763672 |  5.68500018119812 | 5.059999942779541 |
| Aaron Jackson    |  19.31999969482422 | 5.710000038146973 |  5.46999979019165 |
| Alex Renfroe     |  16.18000030517578 | 5.119999885559082 | 7.150000095367432 |
| Nick Calathes    | 17.170000076293945 | 6.420000076293945 | 5.329999923706055 |
| Courtney Fortson | 16.385000228881836 | 5.824999809265137 | 5.325000047683716 |
| Delon Wright     | 15.175000190734863 | 5.140000104904175 | 5.644999980926514 |
| Markelle Fultz   |  23.15999984741211 | 5.920000076293945 |  5.71999979019165 |
| Ja Morant        |  18.59999990463257 | 8.109999895095825 | 6.099999904632568 |
+------------------+-------------------+------------------+------------------+
8 rows in set (0.69 sec)
```

4) Could you give me a list of all the players drafted by the NBA that are from Michigan? I'm also wondering which college had most students drafted by the NBA from 2009-2021?

I first use the query **select nd.Year, nd.Player from nba_drafts nd join college_stats cs on nd.pid = cs.pid join college_teams ct on cs.team_id = ct.id where nd.Year = cs.year and ct.name = "Michigan";** so as to get a list of drafted players from Michigan.

Then I use the query **select count(nd.Player), ct.name from nba_drafts nd join college_stats cs on nd.pid = cs.pid join college_teams ct on cs.team_id = ct.id where nd.Year = cs.year group by cs.team_id order by count(nd.Player) desc limit 1;** As a result, I find that Kentucky had 39 students drafted by the NBA from 2009-2021, which is the most.

```
mysql> select nd.Year, nd.Player from nba_drafts nd join college_stats cs on nd.pid = cs.pid join college_tea
ms ct on cs.team_id = ct.id where nd.Year = cs.year and ct.name = "Michigan";
+------+-------------------+
| Year | Player            |
+------+-------------------+
| 2021 | Franz Wagner      |
| 2021 | Isaiah Livers     |
| 2019 | Jordan Poole      |
| 2019 | Ignas Brazdeikis  |
| 2018 | Moritz Wagner     |
| 2017 | D.J. Wilson       |
| 2016 | Caris LeVert      |
| 2014 | Nik Stauskas      |
| 2014 | Mitch McGary      |
| 2014 | Glenn Robinson III |
| 2013 | Trey Burke        |
| 2013 | Tim Hardaway Jr.  |
| 2011 | Darius Morris     |
+------+-------------------+
13 rows in set (0.07 sec)
```

```
mysql> select count(nd.Player), ct.name from nba_drafts nd join college_stats cs on nd.pid = cs.pid join college_teams ct
 on cs.team_id = ct.id where nd.Year = cs.year group by cs.team_id order by count(nd.Player) desc limit 1;
+-------------------+----------+
| count(nd.Player) | name     |
+-------------------+----------+
|               39 | Kentucky |
+-------------------+----------+
1 row in set (0.05 sec)
```

5) It seems that younger players are more likely to be drafted at higher picks. Could you show me the proportion of freshman that are drafted among the top 3 picks?

   By applying the query **select count(1) as num_of_all_top3_picks from college_stats cs join nba_drafts nd on cs.pid = nd.pid where cs.year = nd.Year and nd.Overall_pick <=3;** I get that there are 35 college players drafted at top 3 picks during 2009-2021.

   Then, I use the query **select count(1) as num_of_Fr_top3_picks from college_stats cs join nba_drafts nd on cs.pid = nd.pid where cs.year = nd.Year and nd.Overall_pick <=3 and cs.yr_of_college = "Fr";** So as to get the number of freshman that are drafted at top 3 picks, and I get the number 27.

   At last, in order to calculate the porportion I use **select 27/35 as pro_of_Fr from college_stats limit 1;** As a result, 77.14% of the players drafted at top 3 picks are freshman.

   It can be inferred that NBA teams prefer to choose younger players at higher picks.

```
mysql> select count(1) as num_of_all_top3_picks from college_stats cs join nba_drafts nd on cs.pid = nd.pid where cs.year = nd.Year and nd.Overall_pick <=3;
+----------------------+
| num_of_all_top3_picks |
+----------------------+
|                   35 |
+----------------------+
1 row in set (0.04 sec)

mysql> select count(1) as num_of_Fr_top3_picks from college_stats cs join nba_drafts nd on cs.pid = nd.pid where cs.year = nd.Year and nd.Overall_pick <=3 and cs.yr_of_college = "Fr";
+----------------------+
| num_of_Fr_top3_picks |
+----------------------+
|                   27 |
+----------------------+
1 row in set (0.05 sec)

mysql> select 27/35 as pro_of_Fr from college_stats limit 1;
+-----------+
| pro_of_Fr |
+-----------+
|    0.7714 |
+-----------+
1 row in set (0.05 sec)
```

6) I want to compare the top 3 pick players' advanced stats (True shooting percentage, box plus minus, offensive box plus minus, defensive box plus minus) with the average stats and max stats in his draft year. Could you give me a list of all these stats from 2009-2021? I want the average and max data of those main

players of each team, so please only choose those who played more than 10 games a year and played more than 20 minutes per game.

To get the required list, I need to create two temporary tables and join them to combine the data. I first use the command

Create table temp1 as (select cp.name, cs.year, cs.TS_per, cs.bpm, cs.obpm, cs.dbpm from college_stats cs join college_players cp on cs.pid = cp.pid join nba_drafts nd on nd.pid = cp.pid where cp.pid in (select pid from nba_drafts where Overall_Pick<=3) and nd.Year = cs.year order by cs.year);

And I create a temp table named temp1 that include the advanced stats of all top 3 picked players.

```
mysql> create table temp1 as (select cp.name, cs.year, cs.TS_per, cs.bpm, cs.obpm, cs.dbpm from college_stats cs join college_players cp on cs.pid = cp.pid join nba_drafts nd on nd.pid = cp.pi
d where cp.pid in (select pid from nba_drafts where Overall_Pick<=3) and nd.Year = cs.year order by cs.year);
Query OK, 35 rows affected (0.09 sec)
Records: 35  Duplicates: 0  Warnings: 0

mysql> select * from temp1;
+----------------------+------+--------+-------+-------+-------+
| name                 | year | TS_per | bpm   | obpm  | dbpm  |
+----------------------+------+--------+-------+-------+-------+
| James Harden         | 2009 |  60.66 | 10.37 |  7.17 |   3.2 |
| Hasheem Thabeet      | 2009 |  64.84 | 11.51 |  3.53 |  7.98 |
| Blake Griffin        | 2009 |  64.77 | 11.34 |  6.91 |  4.43 |
| Derrick Favors       | 2010 |  62.31 |  8.92 |  2.85 |  6.07 |
| Evan Turner          | 2010 |  58.15 | 11.19 |  5.67 |  5.52 |
| John Wall            | 2010 |  56.18 |  7.09 |   4.4 |  2.69 |
| Derrick Williams     | 2011 |  68.97 |  9.35 |   7.3 |  2.04 |
| Kyrie Irving         | 2011 |   69.7 | 12.19 |  9.41 |  2.79 |
| Bradley Beal         | 2012 |  57.45 |  8.03 |  5.02 |  3.01 |
| Anthony Davis        | 2012 |  65.41 | 17.67 |  7.55 | 10.12 |
| Michael Kidd-Gilchrist | 2012 |  57.05 |  8.19 |     5 |  3.18 |
| Otto Porter          | 2013 |     59 | 12.41 |  6.49 |  5.93 |
| Victor Oladipo       | 2013 |  67.12 | 15.28 |   8.8 |  6.48 |
| Anthony Bennett      | 2013 |  60.36 |  6.61 |  3.59 |  3.02 |
| Joel Embiid          | 2014 |   65.5 | 12.74 |  3.38 |  9.36 |
| Jabari Parker        | 2014 |  55.77 |  6.23 |  4.14 |  2.09 |
| Andrew Wiggins       | 2014 |  56.34 |  7.15 |  4.67 |  2.48 |
| Jahlil Okafor        | 2015 |  64.14 |  9.34 |  5.73 |  3.61 |
| D'Angelo Russell     | 2015 |  57.29 |  9.16 |  6.89 |  2.28 |
| Karl-Anthony Towns   | 2015 |  62.72 | 14.88 |  5.64 |  9.24 |
| Jaylen Brown         | 2016 |  51.76 |  2.65 |  0.01 |  2.64 |
| Brandon Ingram       | 2016 |  55.25 |  6.31 |  4.07 |  2.23 |
| Ben Simmons          | 2016 |  59.95 | 10.29 |  5.25 |  5.04 |
| Markelle Fultz       | 2017 |   55.8 |  7.58 |  6.55 |  1.04 |
| Lonzo Ball           | 2017 |  67.26 | 11.04 |  8.18 |  2.86 |
| Jayson Tatum         | 2017 |  56.61 |  6.48 |  3.18 |   3.3 |
| Marvin Bagley III    | 2018 |  64.35 |  8.66 |  6.31 |  2.35 |
| Deandre Ayton        | 2018 |  64.98 |   8.9 |  5.47 |  3.43 |
| R.J. Barrett         | 2019 |  53.19 |  5.93 |  4.59 |  1.34 |
| Ja Morant            | 2019 |  60.77 |  8.98 |  6.94 |  2.04 |
| Zion Williamson      | 2019 |  70.19 | 16.46 | 10.16 |   6.3 |
| James Wiseman        | 2020 |  75.98 | 13.81 |  8.53 |  5.28 |
| Anthony Edwards      | 2020 |  51.73 |  3.97 |  3.54 |  0.42 |
| Cade Cunningham      | 2021 |  57.46 |  6.29 |   3.2 |  3.09 |
| Evan Mobley          | 2021 |   62.4 |    13 |  5.53 |  7.47 |
+----------------------+------+--------+-------+-------+-------+
35 rows in set (0.04 sec)
```

Then, I use the command

create table temp2 as (select year, round(avg(TS_per), 2) as avg_TS_per, max(TS_per) as max_TS_per, round(avg(bpm), 2) as avg_bpm, max(bpm) as max_bpm, round(avg(obpm), 2) as avg_obpm,max(obpm) as max_obpm, round(avg(dbpm), 2) as avg_dbpm, max(dbpm) as max_dbpm from college_stats where GP > 10 and mp > 20 group by year);

And I get the average and max data of every year.

```
mysql> create table temp2 as (select year, round(avg(TS_per), 2) as avg_TS_per, max(TS_per) as max_TS_per, round(avg(bpm), 2) as avg
_bpm, max(bpm) as max_bpm, round(avg(obpm), 2) as avg_obpm,max(obpm) as max_obpm, round(avg(dbpm), 2) as avg_dbpm, max(dbpm) as max_
dbpm from college_stats where GP > 10 and mp > 20 group by year);
Query OK, 13 rows affected (0.24 sec)
Records: 13  Duplicates: 0  Warnings: 0

mysql> select * from temp2;
+------+------------+------------+---------+---------+----------+----------+----------+----------+
| year | avg_TS_per | max_TS_per | avg_bpm | max_bpm | avg_obpm | max_obpm | avg_dbpm | max_dbpm |
+------+------------+------------+---------+---------+----------+----------+----------+----------+
| 2009 |      53.01 |      72.32 |    0.69 |   14.07 |     0.61 |    10.85 |     0.09 |     8.31 |
| 2010 |      53.03 |      70.35 |    0.81 |   12.82 |      0.7 |    10.28 |     0.11 |    10.33 |
| 2011 |      53.25 |      72.33 |    0.77 |   12.41 |     0.65 |     9.97 |     0.12 |        9 |
| 2012 |      53.01 |      70.39 |     0.7 |   17.67 |     0.59 |      8.8 |     0.11 |    10.36 |
| 2013 |      52.67 |      72.94 |    0.72 |   15.28 |     0.58 |     9.56 |     0.13 |    11.68 |
| 2014 |      54.01 |      77.76 |    0.73 |   13.64 |      0.6 |     9.22 |     0.13 |    10.21 |
| 2015 |      53.24 |      71.49 |     0.8 |   14.88 |     0.66 |      8.7 |     0.14 |      9.6 |
| 2016 |      53.94 |      73.84 |    0.77 |   14.39 |     0.65 |    11.22 |     0.12 |     9.47 |
| 2017 |      54.38 |      76.13 |     0.8 |   13.57 |     0.68 |     8.59 |     0.11 |     8.85 |
| 2018 |      54.88 |      71.78 |    0.75 |   13.71 |      0.6 |     9.53 |     0.15 |     9.37 |
| 2019 |      54.37 |      74.94 |     0.6 |    16.6 |     0.55 |    10.16 |     0.05 |     9.24 |
| 2020 |      53.38 |      72.33 |    0.62 |   12.82 |     0.52 |     8.75 |     0.09 |     9.31 |
| 2021 |      53.87 |      78.73 |    0.71 |      13 |     0.58 |     8.58 |     0.12 |    10.76 |
+------+------------+------------+---------+---------+----------+----------+----------+----------+
13 rows in set (0.04 sec)
```

At last, I use the command

select t1.name, t1.year, t1.Ts_per, t2.avg_Ts_per, t2.max_Ts_per, t1.bpm, t2.avg_bpm, t2.max_bpm, t1.obpm, t2.avg_obpm, t2.max_obpm, t1.dbpm, t2.avg_dbpm, t2.max_dbpm from temp1 t1 join temp2 t2 on t1.year = t2.year;

So as to combine the data together.

As a result, the top 3 picked players all have amazingly high bpms, some of them have the highest bpm in the draft year. It seems that box plus minus is an important stat for NBA teams to draft players.

```
mysql> select t1.name, t1.year, t1.Ts_per, t2.avg_Ts_per, t2.max_Ts_per, t1.bpm, t2.avg_bpm, t2.max_bpm, t1.obpm, t2.avg_obpm, t2.max_obpm, t1.dbpm, t2.avg_dbpm, t2.max_dbpm from temp1 t
1 join temp2 t2 on t1.year = t2.year;
+----------------------+------+--------+------------+------------+-------+----------+----------+-------+------------+------------+------+----------+----------+
| name                 | year | Ts_per | avg_Ts_per | max_Ts_per | bpm   | avg_bpm  | max_bpm  | obpm  | avg_obpm   | max_obpm   | dbpm | avg_dbpm | max_dbpm |
+----------------------+------+--------+------------+------------+-------+----------+----------+-------+------------+------------+------+----------+----------+
| James Harden         | 2009 |  60.66 |      53.01 |      72.32 | 10.37 |     0.69 |    14.07 |  7.17 |       0.61 |      10.85 |  3.2 |     0.09 |     8.31 |
| Hasheem Thabeet      | 2009 |  64.84 |      53.01 |      72.32 | 11.51 |     0.69 |    14.07 |  3.53 |       0.61 |      10.85 | 7.98 |     0.09 |     8.31 |
| Blake Griffin        | 2009 |  64.77 |      53.01 |      72.32 | 11.34 |     0.69 |    14.07 |  6.91 |       0.61 |      10.85 | 4.43 |     0.09 |     8.31 |
| Derrick Favors       | 2010 |  62.31 |      53.03 |      70.35 |  8.92 |     0.81 |    12.82 |  2.85 |        0.7 |      10.28 | 6.07 |     0.11 |    10.33 |
| Evan Turner          | 2010 |  58.15 |      53.03 |      70.35 | 11.19 |     0.81 |    12.82 |  5.67 |        0.7 |      10.28 | 5.52 |     0.11 |    10.33 |
| John Wall            | 2010 |  56.18 |      53.03 |      70.35 |  7.09 |     0.81 |    12.82 |   4.4 |        0.7 |      10.28 | 2.69 |     0.11 |    10.33 |
| Derrick Williams     | 2011 |  68.97 |      53.25 |      72.33 |  9.35 |     0.77 |    12.41 |   7.3 |       0.65 |       9.97 | 2.04 |     0.12 |        9 |
| Kyrie Irving         | 2011 |   69.7 |      53.25 |      72.33 | 12.19 |     0.77 |    12.41 |  9.41 |       0.65 |       9.97 | 2.79 |     0.12 |        9 |
| Bradley Beal         | 2012 |  57.45 |      53.01 |      70.39 |  8.03 |      0.7 |    17.67 |  5.02 |       0.59 |        8.8 | 3.01 |     0.11 |    10.36 |
| Anthony Davis        | 2012 |  65.41 |      53.01 |      70.39 | 17.67 |      0.7 |    17.67 |  7.55 |       0.59 |        8.8 | 10.12 |     0.11 |    10.36 |
| Michael Kidd-Gilchrist | 2012 |  57.05 |      53.01 |      70.39 |  8.19 |      0.7 |    17.67 |     5 |       0.59 |        8.8 | 3.18 |     0.11 |    10.36 |
| Otto Porter          | 2013 |     59 |      52.67 |      72.94 | 12.41 |     0.72 |    15.28 |  6.49 |       0.58 |       9.56 | 5.93 |     0.13 |    11.68 |
| Victor Oladipo       | 2013 |  67.12 |      52.67 |      72.94 | 15.28 |     0.72 |    15.28 |   8.8 |       0.58 |       9.56 | 6.48 |     0.13 |    11.68 |
| Anthony Bennett      | 2013 |  60.36 |      52.67 |      72.94 |  6.61 |     0.72 |    15.28 |  3.59 |       0.58 |       9.56 | 3.02 |     0.13 |    11.68 |
| Joel Embiid          | 2014 |   65.5 |      54.01 |      77.76 | 12.74 |     0.73 |    13.64 |  3.38 |        0.6 |       9.22 | 9.36 |     0.13 |    10.21 |
| Jabari Parker        | 2014 |  55.77 |      54.01 |      77.76 |  6.23 |     0.73 |    13.64 |  4.14 |        0.6 |       9.22 | 2.09 |     0.13 |    10.21 |
| Andrew Wiggins       | 2014 |  56.34 |      54.01 |      77.76 |  7.15 |     0.73 |    13.64 |  4.67 |        0.6 |       9.22 | 2.48 |     0.13 |    10.21 |
| Jahlil Okafor        | 2015 |  64.14 |      53.24 |      71.49 |  9.34 |      0.8 |    14.88 |  5.73 |       0.66 |        8.7 | 3.61 |     0.14 |      9.6 |
| D'Angelo Russell     | 2015 |  57.29 |      53.24 |      71.49 |  9.16 |      0.8 |    14.88 |  6.89 |       0.66 |        8.7 | 2.28 |     0.14 |      9.6 |
| Karl-Anthony Towns   | 2015 |  62.72 |      53.24 |      71.49 | 14.88 |      0.8 |    14.88 |  5.64 |       0.66 |        8.7 | 9.24 |     0.14 |      9.6 |
| Jaylen Brown         | 2016 |  51.76 |      53.94 |      73.84 |  2.65 |     0.77 |    14.39 |  0.01 |       0.65 |      11.22 | 2.64 |     0.12 |     9.47 |
| Brandon Ingram       | 2016 |  55.25 |      53.94 |      73.84 |  6.31 |     0.77 |    14.39 |  4.07 |       0.65 |      11.22 | 2.23 |     0.12 |     9.47 |
| Ben Simmons          | 2016 |  59.95 |      53.94 |      73.84 | 10.29 |     0.77 |    14.39 |  5.25 |       0.65 |      11.22 | 5.04 |     0.12 |     9.47 |
| Markelle Fultz       | 2017 |   55.8 |      54.38 |      76.13 |  7.58 |      0.8 |    13.57 |  6.55 |       0.68 |       8.59 | 1.04 |     0.11 |     8.85 |
| Lonzo Ball           | 2017 |  67.26 |      54.38 |      76.13 | 11.04 |      0.8 |    13.57 |  8.18 |       0.68 |       8.59 | 2.86 |     0.11 |     8.85 |
| Jayson Tatum         | 2017 |  56.61 |      54.38 |      76.13 |  6.48 |      0.8 |    13.57 |  3.18 |       0.68 |       8.59 |  3.3 |     0.11 |     8.85 |
| Marvin Bagley III    | 2018 |  64.35 |      54.88 |      71.78 |  8.66 |     0.75 |    13.71 |  6.31 |        0.6 |       9.53 | 2.35 |     0.15 |     9.37 |
| Deandre Ayton        | 2018 |  64.98 |      54.88 |      71.78 |   8.9 |     0.75 |    13.71 |  5.47 |        0.6 |       9.53 | 3.43 |     0.15 |     9.37 |
| R.J. Barrett         | 2019 |  53.19 |      54.37 |      74.94 |  5.93 |      0.6 |     16.6 |  4.59 |       0.55 |      10.16 | 1.34 |     0.05 |     9.24 |
| Ja Morant            | 2019 |  60.77 |      54.37 |      74.94 |  8.98 |      0.6 |     16.6 |  6.94 |       0.55 |      10.16 | 2.04 |     0.05 |     9.24 |
| Zion Williamson      | 2019 |  70.19 |      54.37 |      74.94 | 16.46 |      0.6 |     16.6 | 10.16 |       0.55 |      10.16 |  6.3 |     0.05 |     9.24 |
| James Wiseman        | 2020 |  75.98 |      53.38 |      72.33 | 13.81 |     0.62 |    12.82 |  8.53 |       0.52 |       8.75 | 5.28 |     0.09 |     9.31 |
| Anthony Edwards      | 2020 |  51.73 |      53.38 |      72.33 |  3.97 |     0.62 |    12.82 |  3.54 |       0.52 |       8.75 | 0.42 |     0.09 |     9.31 |
| Cade Cunningham      | 2021 |  57.46 |      53.87 |      78.73 |  6.29 |     0.71 |       13 |   3.2 |       0.58 |       8.58 | 3.09 |     0.12 |    10.76 |
| Evan Mobley          | 2021 |   62.4 |      53.87 |      78.73 |    13 |     0.71 |       13 |  5.53 |       0.58 |       8.58 | 7.47 |     0.12 |    10.76 |
+----------------------+------+--------+------------+------------+-------+----------+----------+-------+------------+------------+------+----------+----------+
35 rows in set (0.04 sec)
```

After getting the list, I use **drop table temp1** and **drop table temp2** to get rid of the temporary tables.

```
mysql> drop table temp1;
Query OK, 0 rows affected (0.07 sec)

mysql> drop table temp2;
Query OK, 0 rows affected (0.06 sec)

mysql> show tables;
+------------------------------+
| Tables_in_college_basketball |
+------------------------------+
| college_players              |
| college_stats                |
| college_teams                |
| nba_drafts                   |
| nba_teams                    |
+------------------------------+
5 rows in set (0.04 sec)
```

I've also exported the college_basketball database by using the command **mysqldump --set-gtid-purged=OFF -h 34.71.12.223 --port 11045 -u xiongty-rw -p college_basketball > college_basketball.sql**

I've attached the exported sql file to this email, please feel free to review it.

```
labsuser@host:~$ mysqldump --set-gtid-purged=OFF -h 34.71.12.223 --port 11045 -u xiongty-rw -p college_basketball > college_basketball.sql
Enter password:
```

Hope this helps! Please feel free to ask me further questions, I'm glad to explore more about the database!

Best,

Tianyi

4. Outline

By doing the final project, I intended to explore the college basketball stats, as well as research about the preference of NBA teams on college players. In other words, I wanted to find out what stat is important for college players to be drafted.

My database is based on the data from Kaggle

https://www.kaggle.com/datasets/adityak2003/college-basketball-players-20092021?select=CollegeBasketballPlayers2009-2021.csv

I used 2 tables as my original data, one including all the U.S. college men's basketball players' stats from 2009-2021, one including the results of the NBA drafts each year. Apparently, I imported the college_stats as my primary table. However, there are data that appeared repeatedly. Players' names and pids appear recursively each year since many players play more than one year in college. And the college team names also appeared repeatedly. Thus, in order to normalize the table, I created connection tables college_players and college_teams to store the names of the players and college teams. When creating the table college_players, I extracted the pid and name fields from the original table, and used Excel to delete the replicate pids, so that I could set pid as a primary key for the table. The table college_players also allowed me to track the player's draft results in the nba_drafts table using pid. However, since international players were not recorded in the college_stats table, their pids were recorded as NULL, and those who didn't attend college also didn't have pids in the nba_drafts table. Similarly, I also created the nba_teams table to store the names of NBA teams.

Here are some of my screenshots of my normalization and commands to connect the tables.

Connect college_stats.team_id to college_teams.id

```
update college_stats cs join college_teams ct on cs.team = ct.name set cs.team_id = ct.id
```

Connect nba_drafts.team_id to nba_teams.id

```
insert into nba_teams (name) select distinct Team from nba_drafts
update nba_drafts nd inner join nba_teams nt on nd.Team = nt.name set nd.team_id = nt.id
```

Connect nba_drafts.pid to college_players.pid

Since there are players with different pids but have same names, I need to join the tables with college_stats by the corresponding draft year, so as to avoid choosing the wrong player.

```
update nba_drafts nd join college_players cp on nd.Player = cp.name join college_stats cs on nd.Year = cs.year set nd.pid = cp.pid
```

This is my whole process of creating the database. There are still bugs in the tables since some of the names in the table college_players are expressed differently in the table nba_drafts. For example, names with “J.R” may differ from “J.R.”. I’ve fixed some of them but there might still be unmatched names. If I have more time, I may complete the database to make sure all players’ names are matched between tables. And there is another table from Kaggle which includes the advanced NBA stats of the players, maybe I can compare the NBA stats with the college stats to track the college players’ performance in the NBA.