Ramone Martin
ISTA 322
12.10.25

NBA statistics 2015-2016

Data:
The two data sources used are as follows:

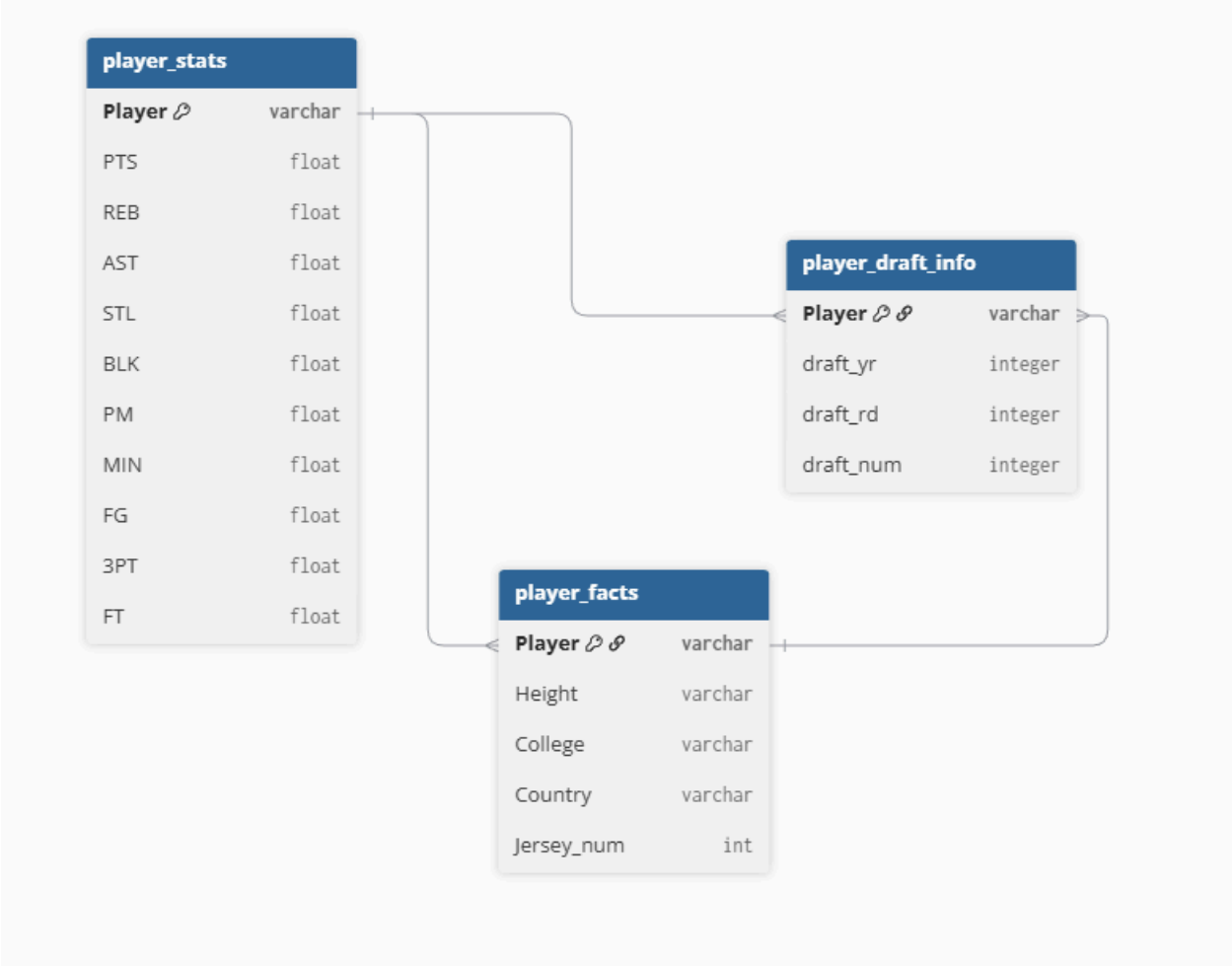
Github:
https://github.com/mvanbommel/scrape_nba_box_scores/blob/master/2015_2016_NBA_box_score_player_data.csv

This dataset is a game by game dataset with over 31,000 rows that has game stats for every NBA player during the 2015-2016 NBA season. This source has only in-game statistics, team affiliations, and other in-game stats such as home and away teams and if the team won or lost. I used the statistical portion of the dataset, so only quantitative data that could be combined and averaged out to show overall season averages. I used the player name as well as a unique identifier. The challenges cleaning this dataset were that it was extremely large to deal with. There were only 475 unique values in the name column so I had to combine like names to average stats, then had to drop duplicates and manually change name values from players that had similar names. I had to change the datatype of some columns in order to have them aggregate properly or load into SQL properly as well.

BallDontLie API - https://docs.balldontlie.io/#get-a-specific-player

This dataset was an API that contains information on every NBA player since 1946. This dataset contained over 5000 rows and contained many duplicates, as well as the same players with like names as well. I had to manually create a Player column from the name columns to match the github dataset. This dataset contained information such as year drafted, country of origin, height, weight, and other non game related statistics. I used the player name and then used niche columns that would be interesting to see after querying such as jersey number, height, draft position, and college affiliation as well. The challenges cleaning this dataset were similar to the other dataset, there were lots of nefarious rows that had to be removed as well as unnecessary columns. I had to change the datatype of some columns in order to have them aggregate properly or load into SQL properly as well.

RDB Schema :


player_stats
Player varchar
PTS float
REB float
AST float
STL float
BLK float
PM float
MIN float
FG float
3PT float
FT float
player_draft_info
Player varchar
draft_yr integer
draft_rd integer
draft_num integer
player_facts
Player varchar
Height varchar
College varchar
Country varchar
Jersey_num int

Queries:
I ran 5 test queries.

1) Top 10 players in PTS who went to UCLA

```
run_query(""" SELECT t1.Player, PTS, College
FROM player_facts AS t1
INNER JOIN player_stats AS t2 ON t1.Player = t2.Player
WHERE College = 'UCLA'
ORDER BY t2.PTS DESC
LIMIT 10; """)
```

	Player	PTS	College
0	R. Westbrook	23.5	UCLA
1	Jr. Holiday	16.8	UCLA
2	K. Love	16.0	UCLA
3	D. Collison	14.0	UCLA
4	Z. LaVine	14.0	UCLA
5	A. Afflalo	12.8	UCLA
6	T. Ariza	12.7	UCLA
7	S. Muhammad	10.5	UCLA
8	M. Barnes	10.0	UCLA
9	J. Farmar	9.2	UCLA

2) Top 10 players in Rebounds who are 6'3

```
run_query(""" SELECT t1.Player, REB, Height
FROM player_facts AS t1
INNER JOIN player_stats AS t2 ON t1.Player = t2.Player
WHERE Height = '6-3'
ORDER BY t2.REB DESC
LIMIT 10; """)
```

	Player	REB	Height
0	J. Wall	4.9	6-3
1	V. Oladipo	4.8	6-3
2	M. Smart	4.2	6-3
3	J. Jack	4.2	6-3
4	E. Payton	3.6	6-3
5	B. Wright	3.6	6-3
6	E. Mudiay	3.4	6-3
7	D. Rose	3.4	6-3
8	D. Russell	3.4	6-3
9	M. Ellis	3.3	6-3

3) Top 10 players in PTS drafted 1st overall

```
run_query(""" SELECT t1.Player, PTS, draft_yr, draft_num
FROM player_draft_info AS t1
INNER JOIN player_stats AS t2 ON t1.Player = t2.Player
WHERE draft_num = 1
ORDER BY t2.PTS DESC
LIMIT 10; """)
```

	Player	PTS	draft_yr	draft_num
0	L. James	25.3	2003	1
1	A. Davis	24.3	2012	1
2	B. Griffin	21.4	2009	1
3	A. Wiggins	20.7	2014	1
4	J. Wall	19.9	2010	1
5	K. Irving	19.6	2011	1
6	K. Towns	18.3	2015	1
7	D. Rose	16.4	2008	1
8	D. Howard	13.8	2004	1
9	T. Duncan	8.6	1997	1

4) Top 10 players in PTS from the USA

```
run_query(""" SELECT t1.Player, PTS, Country
FROM player_facts AS t1
INNER JOIN player_stats AS t2 ON t1.Player = t2.Player
WHERE Country = 'USA'
ORDER BY t2.PTS DESC
LIMIT 10; """)
```

	Player	PTS	Country
0	S. Curry	30.1	USA
1	J. Harden	29.0	USA
2	K. Durant	28.2	USA
3	D. Cousins	26.9	USA
4	L. James	25.3	USA
5	D. Lillard	25.1	USA
6	A. Davis	24.3	USA
7	D. DeRozan	23.5	USA
8	R. Westbrook	23.5	USA
9	P. George	23.1	USA

5) Top 10 players in PTS with the jersey number 23

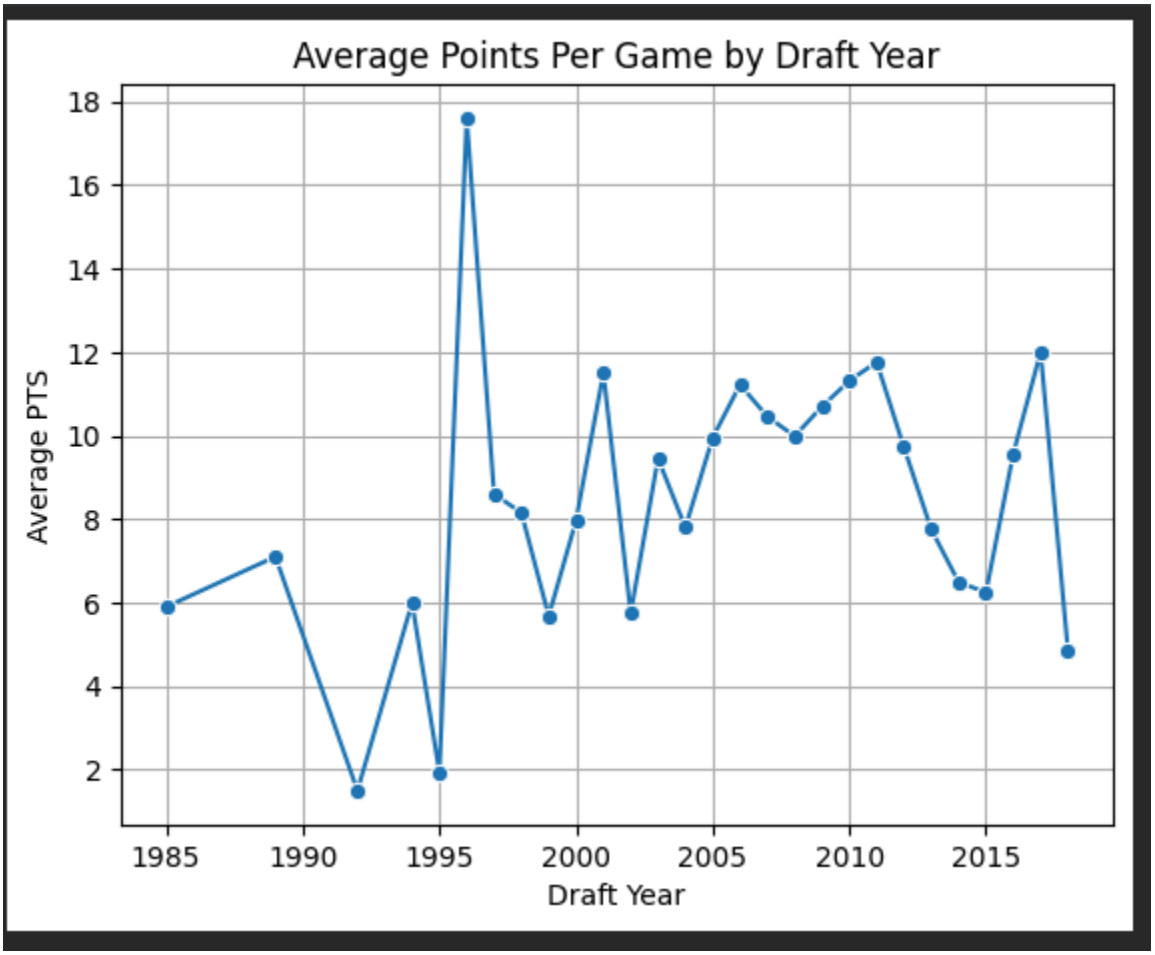
```
run_query(""" SELECT t1.Player, PTS, Jersey_num
FROM player_facts AS t1
JOIN player_stats AS t2 ON t1.Player = t2.Player
WHERE Jersey_num = 23
ORDER BY t2.PTS DESC
LIMIT 10; """)
```

	Player	PTS	Jersey_num
0	L. James	25.3	23
1	D. Rose	16.4	23
2	E. Gordon	15.2	23
3	Dr. Green	14.0	23
4	B. McLemore	7.8	23

Plots:

I created two plots.

1) A plot showing the average scoring for players drafted by year drafted



2) PPG avg by country for the top 10 scoring countries