Final Project Data Check point

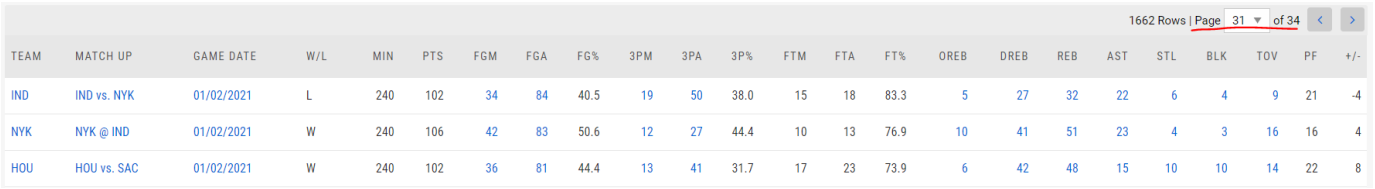
1. Project code:

https://github.com/cwtseng/SI507_final_proj.git

2. Data sources:

A. Origin URL: https://www.nba.com/stats/teams/boxscores/

i. Website data example



1662 Rows | Page 31 of 34

TEAM	MATCH UP	GAME DATE	W/L	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	STL	BLK	TOV	PF	+/-
IND	IND vs. NYK	01/02/2021	L	240	102	34	84	40.5	19	50	38.0	15	18	83.3	5	27	32	22	6	4	9	21	-4
NYK	NYK @ IND	01/02/2021	W	240	106	42	83	50.6	12	27	44.4	10	13	76.9	10	41	51	23	4	3	16	16	4
HOU	HOU vs. SAC	01/02/2021	W	240	102	36	81	44.4	13	41	31.7	17	23	73.9	6	42	48	15	10	10	14	22	8

B. Explanation of data label: https://jr.nba.com/how-to-read-a-box-score/

i. Data label explanation

- **MIN** = Minutes
- **FGM** = Field-goals made
- **FGA** = Field-goals attempted
- **FG%** = Field goal percentage
- **3PM** = 3-pointers made
- **3PA** = 3-pointers attempted
- **3P%** = 3-point percentage
- **FTM** = Free throws made
- **FTA** = Free throws attempted
- **FT%** = Free throw percentage
- **OREB** = Offensive rebounds
- **DREB** = Defensive rebounds
- **REB** = Total rebounds
- **AST** = Assists
- **TOV** = Turnovers
- **STL** = Steals
- **BLK** = Blocked shots
- **PF** = Personal fouls
- **PTS** = Points scored
- **+/-** = Plus/Minus

C. Raw data format: HTML

D. How to access the data?

i. Grab multiple pages raw HTMLs with Selenium(webdriver and Select)

ii. Parse the raw htmls via Beautiful Soup

iii. Organize data with pandas dataframe

iv. Store parsed data in cache and database

E. Summary of data:

i. >400 records available.

ii. >400 records retrieved.

iii. Retrieved all data available on the website, because all of them are useful.

iv. The row of data indicates different teams and different dates result

v. The column of data indicates different key performance result in the team.

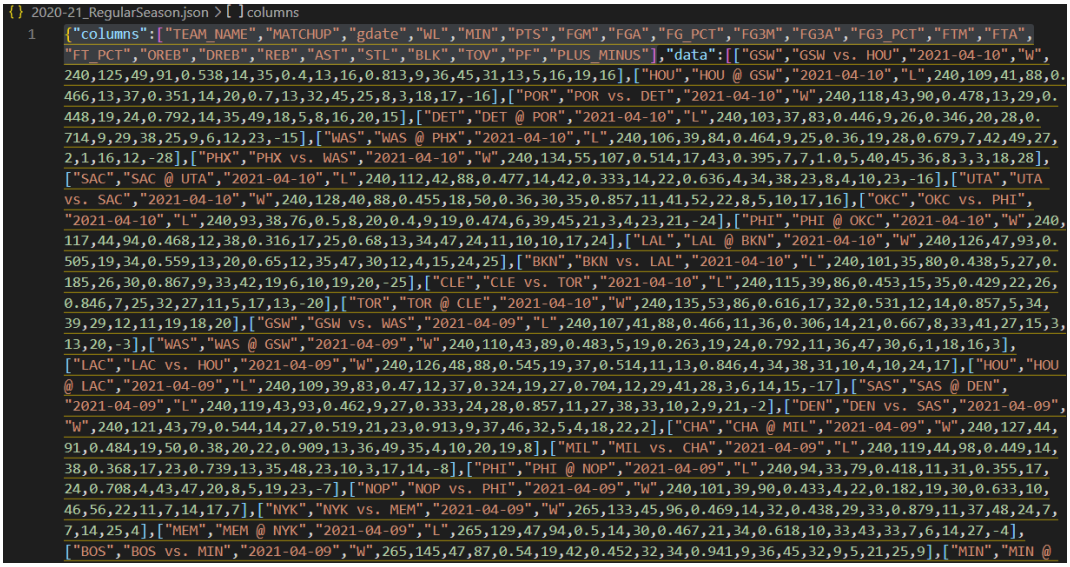
vi. The data can be processing with specified key performance(ex: field goal percent, assist, rebound), teams(ex: Lakers, Clippers), and interesting period(ex: start date and end date)

vii. For a single team, I can plot different key performance over interesting period. Also, user can apply statistics options, ex: average, maximum, minimum, standard deviation within that period.

viii. For multiple teams, the program can show their comparison of key performance in the same plot It can also applied the statistics result as well

F. Evidence of caching:

i. Example of cache file 2020-21_RegularSeason.json, store all the crawling data



3. Database:

A. Database has been used, single table with all team and performance result. The different Id is showing the different date.

B. Only primary key is used in the database

C. Part of snapshot of database

Table: teams Filter in any column

	Id	TEAM_NAME	MATCHUP	gdate	WL	MIN	PTS	FGM	FGA	FG_PCT	FG3M	FG3A	FG3_PCT	FTM	FTA	FT_PCT	OREB	DREB	REB	AST	STL	BLK	TOV
	過濾	過濾	過濾	過濾	過濾	過濾	過濾	過濾	過濾	過濾	過濾	過濾	過濾	過濾	過濾	過濾	過濾	過濾	過濾	過濾	過濾	過濾	過濾
1	1	MIL	MIL @ ATL	2021-04-15	W	240	120	46	95	0.48	16	43	0.37	12	17	0.71	14	37	51	25	8	1	9
2	2	ATL	ATL vs. MIL	2021-04-15	L	240	109	39	86	0.45	14	34	0.41	17	19	0.90	8	31	39	23	5	4	12
3	3	LAL	LAL vs. BOS	2021-04-15	L	240	113	44	98	0.45	12	37	0.32	13	19	0.68	11	25	36	28	12	2	8
4	4	BOS	BOS @ LAL	2021-04-15	W	240	121	48	85	0.57	14	32	0.44	11	13	0.85	7	41	48	33	4	6	21
5	5	SAC	SAC @ PHX	2021-04-15	L	240	114	43	78	0.55	14	30	0.47	14	19	0.74	4	28	32	22	7	4	15
6	6	PHX	PHX vs. SAC	2021-04-15	W	240	122	48	86	0.56	14	38	0.37	12	14	0.86	10	29	39	30	9	4	11
7	7	GSW	GSW @ CLE	2021-04-15	W	240	119	47	92	0.51	17	46	0.37	8	10	0.80	9	38	47	33	9	5	16
8	8	CLE	CLE vs. GSW	2021-04-15	L	240	101	35	81	0.43	10	31	0.32	21	27	0.78	8	31	39	21	11	2	15
9	9	TOR	TOR vs. SAS	2021-04-14	W	240	117	41	88	0.47	14	30	0.47	21	25	0.84	13	41	54	21	8	5	16
10	10	SAS	SAS @ TOR	2021-04-14	L	240	112	38	89	0.43	17	39	0.44	19	27	0.70	12	29	41	26	6	7	13
11	11	MEM	MEM vs. DAL	2021-04-14	L	240	113	42	90	0.47	13	34	0.38	16	21	0.76	11	34	45	27	7	3	7
12	12	DAL	DAL @ MEM	2021-04-14	W	240	114	44	88	0.50	13	40	0.33	13	17	0.77	8	37	45	25	4	3	11

4. Interaction and Presentation:

A. First input message box would ask you to input team names, abbreviations same as official NBA website. Ex: TOR LAC DET …

B. Second input message box would ask you to input options:

i. [MIN| PTS|FGM| …]: the available numerical key performance you are going to display

ii. [none|avg|max|min|std]: statistic options for the data within interesting period

iii. <str1> : starting date of capture, format should be ex:‘2021-04-15’

iv. <str2>: end date of capture, format should be ex:‘2021-04-15’
v. <integer>: how many page number for program to crawl
vi. [none|plot]: whether to plot data or just show table
vii. [none|cache|db]: whether to use cache or db or direct fetch from website