Name   : Diemas Aksya Fachriza

Group   : DS-1

IYKRA Data Science Batch 5 - Statistics Practice Case
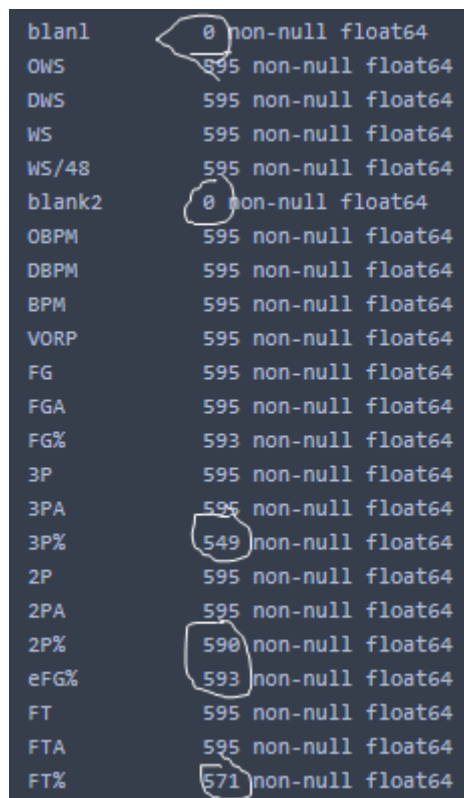
[Click here to open the Code/Notebook](#)

- Case

In this task not all data will be used, only data in 2017. So it is necessary to do filtering at the beginning. Besides that there are some players who make team transfers in the NBA transfer market so that there is duplication of player data. Therefore you can use the df.drop_duplicates() syntax to solve this to produce the same output as the trainer. Delete columns that have as many missing values as the entire row of data. Then you can do additional pre-processing if needed or you can immediately process the data.

Column Descriptions: https://www.basketball-reference.com/about/glossary.html

- Data Preprocessing

There are some columns with missing values or no values at all. For the missing values, it is imputed using its column mean and remove the columns without values.

```
blan1           0 non-null float64
OWS           595 non-null float64
DWS           595 non-null float64
WS            595 non-null float64
WS/48         595 non-null float64
blank2          0 non-null float64
OBPM          595 non-null float64
DBPM          595 non-null float64
BPM           595 non-null float64
VORP          595 non-null float64
FG            595 non-null float64
FGA           595 non-null float64
FG%           593 non-null float64
3P            595 non-null float64
3PA           595 non-null float64
3P%           549 non-null float64
2P            595 non-null float64
2PA           595 non-null float64
2P%           590 non-null float64
eFG%          593 non-null float64
FT            595 non-null float64
FTA           595 non-null float64
FT%           571 non-null float64
```

Other than that, there are some duplicate player names in this data. There are 109 duplicates. Therefore, these rows need to be dropped. Until this point, there are 486 rows available, or should I say 486 Players recorded in this dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 486 entries, 0 to 485
Data columns (total 50 columns):
```

Data preprocessing step should be done.

- Tasks
  1. **Who is the youngest and oldest player in the NBA in 2017 for each team (Tm) ?**

	Tm	MaxAge	MinAge	Player Oldest	Player Youngest
0	ATL	32.0	22.0	Gary Neal	DeAndre' Bembry
1	BOS	31.0	20.0	Gerald Green	Jaylen Brown
2	BRK	36.0	21.0	Luis Scola	Isaiah Whitehead
3	CHI	35.0	21.0	Dwyane Wade	Bobby Portis
4	CHO	31.0	21.0	Brian Roberts	Christian Wood
5	CLE	38.0	21.0	Chris Andersen	Kay Felder
6	DAL	38.0	21.0	Dirk Nowitzki	Ben Bentil
7	DEN	36.0	19.0	Mike Miller	Jamal Murray
8	DET	34.0	20.0	Beno Udrih	Henry Ellenson
9	GSW	36.0	20.0	David West	Kevon Looney
10	HOU	34.0	20.0	Nene Hilario	Chinanu Onuaku
11	IND	32.0	20.0	Aaron Brooks	Myles Turner
12	LAC	39.0	19.0	Paul Pierce	Diamond Stone
13	LAL	37.0	19.0	Metta World	Brandon Ingram
14	MEM	40.0	20.0	Vince Carter	Wade Baldwin
15	MIA	36.0	20.0	Udonis Haslem	Justise Winslow
16	MIL	39.0	19.0	Jason Terry	Thon Maker
17	MIN	34.0	20.0	John Lucas	Tyus Jones
18	NOP	33.0	20.0	Jarrett Jack	Cheick Diallo
19	NYK	32.0	21.0	Carmelo Anthony	Kristaps Porzingis
20	OKC	36.0	20.0	Nick Collison	Domantas Sabonis
21	ORL	32.0	20.0	C.J. Watson	Stephen Zimmerman
22	PHI	32.0	21.0	Tiago Splitter	Timothe Luwawu-Cabarrot
23	PHO	34.0	19.0	Leandro Barbosa	Dragan Bender
24	POR	28.0	21.0	Evan Turner	Noah Vonleh
25	SAC	31.0	19.0	Arron Afflalo	Georgios Papagiannis
26	SAS	39.0	20.0	Manu Ginobili	Dejounte Murray
27	TOR	30.0	21.0	DeMarre Carroll	Bruno Caboclo
28	TOT	36.0	21.0	Matt Barnes	Chris McCullough
29	UTA	35.0	21.0	Joe Johnson	Dante Exum
30	WAS	32.0	21.0	Marcin Gortat	Kelly Oubre

According to the above output, player's age are around 19 to 40 years old, while the average age for the whole player in this dataset is 26.4 years old. For example. There's DeAndre' Bembry as the

youngest player and Gary Neal as the oldest player in Atlanta Hawks (ATL). But if we were to look the average of age in this dataset, most players are tend to be young.

2. **Which player has the most minutes played (MP) in each position (Pos)?**

	Pos	MaxMP	Player
0	C	3030.0	Karl-Anthony Towns
1	PF	2803.0	Harrison Barnes
2	PF-C	980.0	Joffrey Lauvergne
3	PG	2947.0	James Harden
4	SF	3048.0	Andrew Wiggins
5	SG	2796.0	C.J. McCollum

According to the above image, Max MP for each position is around 980 and 3048. The above names hold the most minutes played. I assume these players have the most experience as a player in NBA for each position.

3. **Which team has the highest average total rebound percentage (TRB%), assist percentage (AST%), steal percentage (STL%), and block percentage (BLK%)?**

```
Team with the highest TRB% is WAS with 13.45 TRB%
Team with the highest AST% is DEN with 15.86 AST%
Team with the highest STL% is MIN with 2.37 STL%
Team with the highest BLK% is GSW with 2.74 BLK%
```

For TRB%, Washington Wizards (WAS) holds the highest with 13.45 TRB%. For AST%, Denver Nuggets holds the highest with 15.86%. For STL%, Minnesota Timberwolves (MIN) holds the highest with 2.37%. For BLK%, Golden State Warriors (GSW) holds the highest with 2.75%. These teams have the highest average of the specified variables.

4. **Who is the best player in your opinion based on his record stats? note: you can refer to variables point (PTS), assists, rebounds, or anything else. A combination of several variables would be nice.**

The Player Efficiency Rating (PER) is a per-minute rating developed by ESPN.com columnist John Hollinger. In John's words, "The PER sums up all a player's positive accomplishments, subtracts the negative accomplishments, and returns a per-minute rating of a player's performance." It appears from his books that John's database only goes back to the 1988-89 season.

The calculation of PER itself involves a lot of other variables, such as MP, AST, FGA, FG, FTA, FT, DRB&, TRB, ORB, STL, BLK, and much more. Therefore, this analysis would only find the player with the highest PER.

Link to PER formula explanation : https://www.basketball-reference.com/about/per.html

	Player	PER	G	PTS	MP
413	Jarnell Stokes	31.5	2.0	3.0	7.0

According to the above output, Jamell Stokes has the highest PER status. But, there's something wront. His G (Games), PTS (Points), and MP (Minutes Played) seems too low compared to the other legends I know, such as L. James, T. Duncan, and M. Jordan. It seems Jamell Stokes is a new player in NBA and the PER calculation isn't that credible for new players. Therefore, I can't choose Jamell Stokes as the best player.

I dig deeper into this data by sorting each player statis by PER, and here is the output.

	Player	PER	G	PTS	MP
413	Jarnell Stokes	31.5	2.0	3.0	7.0
216	Demetrius Jackson	30.8	5.0	10.0	17.0
457	Russell Westbrook	30.6	81.0	2558.0	2802.0
280	Boban Marjanovic	29.6	35.0	191.0	293.0
118	Kevin Durant	27.6	62.0	1555.0	2070.0
260	Kawhi Leonard	27.5	74.0	1888.0	2474.0
99	Anthony Davis	27.5	75.0	2099.0	2708.0
171	James Harden	27.3	81.0	2356.0	2947.0
219	LeBron James	27.0	74.0	1954.0	2794.0
423	Isaiah Thomas	26.5	76.0	2199.0	2569.0

Demetrius Jackson has the second highest PER, but his G, PTS, and MP seems too low compared to the other players below his name. Meanwhile, Russell Westbrook holds the third highest PER with a relatively high G, PTS, and MP values as well. According to this link, He is a nine-time NBA All-Star and earned the NBA Most Valuable Player Award for the 2016–17 season. Therefore, I choose him as the best player in NBA (2017).

5. **Which team has the best average stat record of their players? Note: you can refer to points, assists, rebounds, or anything else. A combination of several variables would be nice**

For this Analysis, I would like to use TS%, FT%, eFG%, DRB%, and BLK% columns (For the column details, please take a look at the Glossary link at the top page of this notebook). These columns would be averaged for each player. And then, player's averaged score would be grouped by team to find the best team. Here is the initial statuses of each players.

	TS%	FT%	eFG%	DRB%	BLK%
0	0.560	0.898	0.531	7.1	0.6
1	0.565	0.750	0.521	18.0	2.0
2	0.589	0.611	0.571	15.5	2.6
3	0.559	0.892	0.514	8.4	0.4
4	0.529	0.725	0.500	23.8	3.1
...	...	...	...	...	...
481	0.604	0.679	0.571	17.3	3.0
482	0.508	0.564	0.494	17.0	3.3
483	0.346	0.600	0.323	24.9	3.7
484	0.503	0.775	0.473	14.2	1.5
485	0.547	0.653	0.529	21.9	4.4

The above values has different scale. To make it much more fair, the data need to be normalized first using MinMaxScaler. Here is the data after rescaling.

	TS%	FT%	eFG%	DRB%	BLK%
0	0.700876	0.898	0.531	0.071	0.034682
1	0.707134	0.750	0.521	0.180	0.115607
2	0.737171	0.611	0.571	0.155	0.150289
3	0.699625	0.892	0.514	0.084	0.023121
4	0.662078	0.725	0.500	0.238	0.179191
...	...	...	...	...	...
481	0.755945	0.679	0.571	0.173	0.173410
482	0.635795	0.564	0.494	0.170	0.190751
483	0.433041	0.600	0.323	0.249	0.213873
484	0.629537	0.775	0.473	0.142	0.086705
485	0.684606	0.653	0.529	0.219	0.254335

Then we can calculate each player's average status.

	TS%	FT%	eFG%	DRB%	BLK%	avgStat%
0	0.700876	0.898	0.531	0.071	0.034682	0.447112
1	0.707134	0.750	0.521	0.180	0.115607	0.454748
2	0.737171	0.611	0.571	0.155	0.150289	0.444892
3	0.699625	0.892	0.514	0.084	0.023121	0.442549
4	0.662078	0.725	0.500	0.238	0.179191	0.460854
...	...	...	...	...	...	...
481	0.755945	0.679	0.571	0.173	0.173410	0.470471
482	0.635795	0.564	0.494	0.170	0.190751	0.410909
483	0.433041	0.600	0.323	0.249	0.213873	0.363783
484	0.629537	0.775	0.473	0.142	0.086705	0.421248
485	0.684606	0.653	0.529	0.219	0.254335	0.467988

Finally, we can group the average of avgStat% variable by team names. Here is the result.

	Tm	avgStat%
26	SAS	0.454016
9	GSW	0.452604
7	DEN	0.446155
12	LAC	0.443542
25	SAC	0.442405
29	UTA	0.441357
1	BOS	0.438883
22	PHI	0.437931
30	WAS	0.437288
10	HOU	0.436645

Therefore, I choose San Antonio Spurs (SAS) is the best team according to the above variables with the average status is 0.454016.

Hopefully, this analysis would give you more new insights about NBA Players.