ASSIGNMENT 1
Exploratory Data Analysis

The dataset used in this exercise is **2021-2022 NBA Player Stats**

Exploratory Data Analysis (EDA), also known as Data Exploration, is a step in the Data Analysis Process, where several techniques are used to better understand the dataset being used.

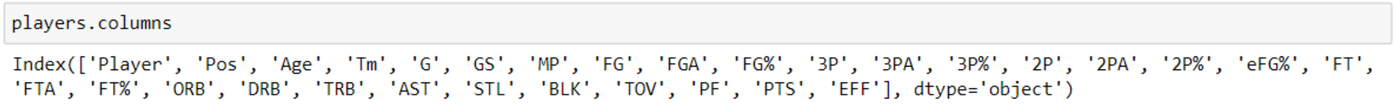
About Dataset:

We are using the 2021-2022 NBA Player Stats - Regular.csv file here. This dataset contains data on several NBA player stats for every game of the 2021-2022 regular season.
These are the variables/columns in the data set:

Rk *- Rank*
Player*- Player's name*
Pos *- Position*
Age*- Player's age*
Tm*- Team*
G *- Games played*
GS *- Games started.*
MP*- Minutes played per game.*
FG *- Field goals per game*
FGA *- Field goal attempts per game*
FG%- Field goal percentage
3P *- 3-point field goals per game*
3PA*- 3-point field goal attempts per game*
3P%- *3-point field goal percentage*
2P*- 2-point field goals per game*
2PA *- 2-point field goal attempts per game*
2P% *- 2-point field goal percentage*
eFG% *- Effective field goal percentage*
FT *- Free throws per game*
FTA *- Free throw attempts per game*
FT% *-Free throw percentage*
ORB *- Offensive rebounds per game*
DRB *- Defensive rebounds per game*
TRB *- Total rebounds per game*
AST *- Assists per game*
STL *- Steals per game*
BLK *- Blocks per game*
TOV*- Turnovers per game*
PF*- Personal fouls per game*
PTS *- Points per game*

Understanding the Dataset & variables

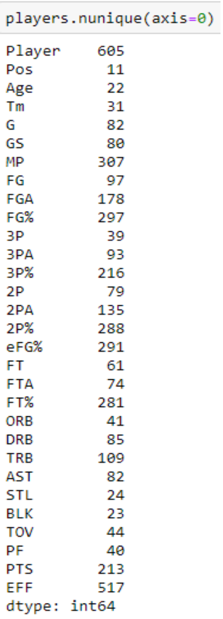
There are (812,30) that's Rows and Columns in the data set.
The names of the columns are as shown below in and are also referenced in the 'About Dataset' section



```
players.columns
```

```
Index(['Player', 'Pos', 'Age', 'Tm', 'G', 'GS', 'MP', 'FG', 'FGA', 'FG%', '3P', '3PA', '3P%', '2P', '2PA', '2P%', 'eFG%', 'FT',
'FTA', 'FT%', 'ORB', 'DRB', 'TRB', 'AST', 'STL', 'BLK', 'TOV', 'PF', 'PTS', 'EFF'], dtype='object')
```

Figure 1



```
players.nunique(axis=0)
```

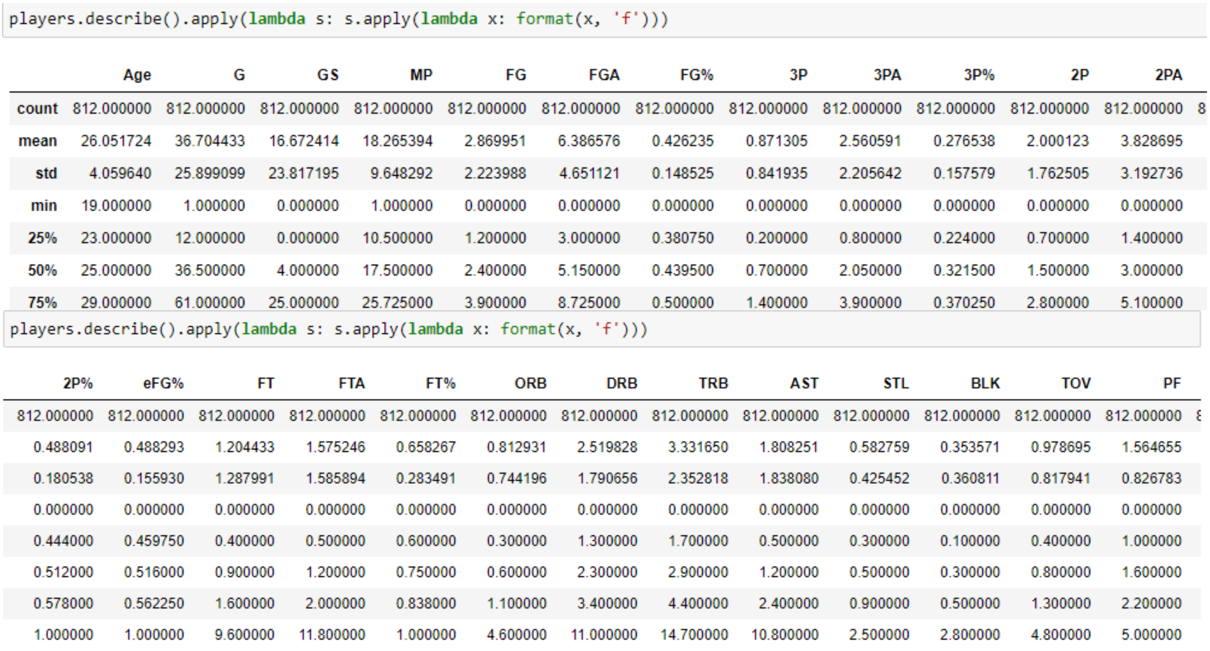
```
Player    605
Pos        11
Age        22
Tm         31
G          82
GS         80
MP        307
FG         97
FGA       178
FG%       297
3P         39
3PA        93
3P%       216
2P         79
2PA       135
2P%       288
eFG%      291
FT         61
FTA        74
FT%       281
ORB        41
DRB        85
TRB       109
AST        82
STL        24
BLK        23
TOV        44
PF         40
PTS       213
EFF       517
dtype: int64
```

Figure 2

Fig2 shows the number of unique values for each column/variable in our dataset.

To get a better understanding of our variables and its values we use the '.describe()' function to get the statistics such as Mean, Min, Max, Standard Deviation, Quartiles ,Count for each variable in the dataset as shown in Fig3.

Exploring the only discrete variable we have we can see that in fig4 we have 31 different NBA teams' data in our data set.



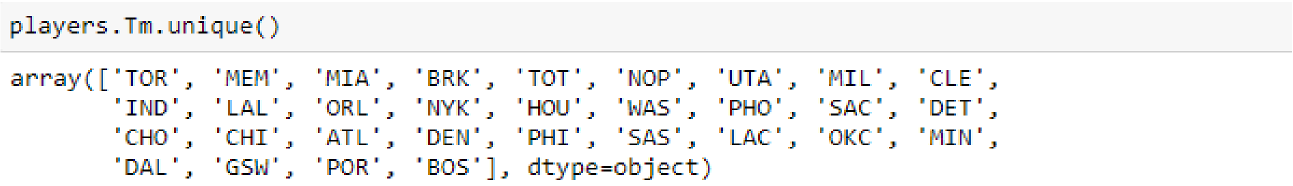
```
players.describe().apply(lambda s: s.apply(lambda x: format(x, 'f')))
```

	Age	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA
count	812.000000	812.000000	812.000000	812.000000	812.000000	812.000000	812.000000	812.000000	812.000000	812.000000	812.000000	812.000000
mean	26.051724	36.704433	16.672414	18.265394	2.869951	6.386576	0.426235	0.871305	2.560591	0.276538	2.000123	3.828695
std	4.059640	25.899099	23.817195	9.648292	2.223988	4.651121	0.148525	0.841935	2.205642	0.157579	1.762505	3.192736
min	19.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	23.000000	12.000000	0.000000	10.500000	1.200000	3.000000	0.380750	0.200000	0.800000	0.224000	0.700000	1.400000
50%	25.000000	36.500000	4.000000	17.500000	2.400000	5.150000	0.439500	0.700000	2.050000	0.321500	1.500000	3.000000
75%	29.000000	61.000000	25.000000	25.725000	3.900000	8.725000	0.500000	1.400000	3.900000	0.370250	2.800000	5.100000

```
players.describe().apply(lambda s: s.apply(lambda x: format(x, 'f')))
```

2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF
812.000000	812.000000	812.000000	812.000000	812.000000	812.000000	812.000000	812.000000	812.000000	812.000000	812.000000	812.000000	812.000000
0.488091	0.488293	1.204433	1.575246	0.658267	0.812931	2.519828	3.331650	1.808251	0.582759	0.353571	0.978695	1.564655
0.180538	0.155930	1.287991	1.585894	0.283491	0.744196	1.790656	2.352818	1.838080	0.425452	0.360811	0.817941	0.826783
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.444000	0.459750	0.400000	0.500000	0.600000	0.300000	1.300000	1.700000	0.500000	0.300000	0.100000	0.400000	1.000000
0.512000	0.516000	0.900000	1.200000	0.750000	0.600000	2.300000	2.900000	1.200000	0.500000	0.300000	0.800000	1.600000
0.578000	0.562250	1.600000	2.000000	0.838000	1.100000	3.400000	4.400000	2.400000	0.900000	0.500000	1.300000	2.200000
1.000000	1.000000	9.600000	11.800000	1.000000	4.600000	11.000000	14.700000	10.800000	2.500000	2.800000	4.800000	5.000000

Figure 3



```
players.Tm.unique()
```

```
array(['TOR', 'MEM', 'MIA', 'BRK', 'TOT', 'NOP', 'UTA', 'MIL', 'CLE',
       'IND', 'LAL', 'ORL', 'NYK', 'HOU', 'WAS', 'PHO', 'SAC', 'DET',
       'CHO', 'CHI', 'ATL', 'DEN', 'PHI', 'SAS', 'LAC', 'OKC', 'MIN',
       'DAL', 'GSW', 'POR', 'BOS'], dtype=object)
```

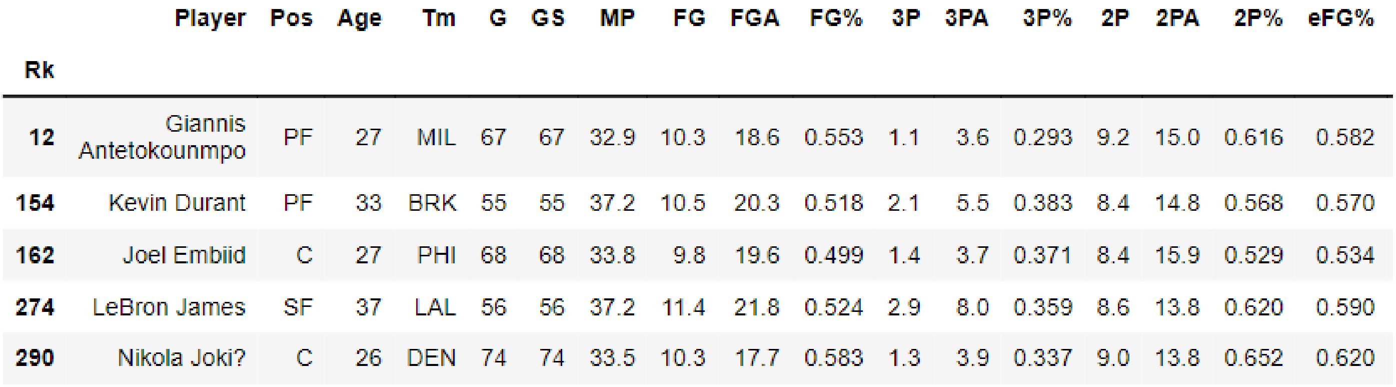
Figure 4

Q1. What are the important variables and how can it be interpreted in a better way?

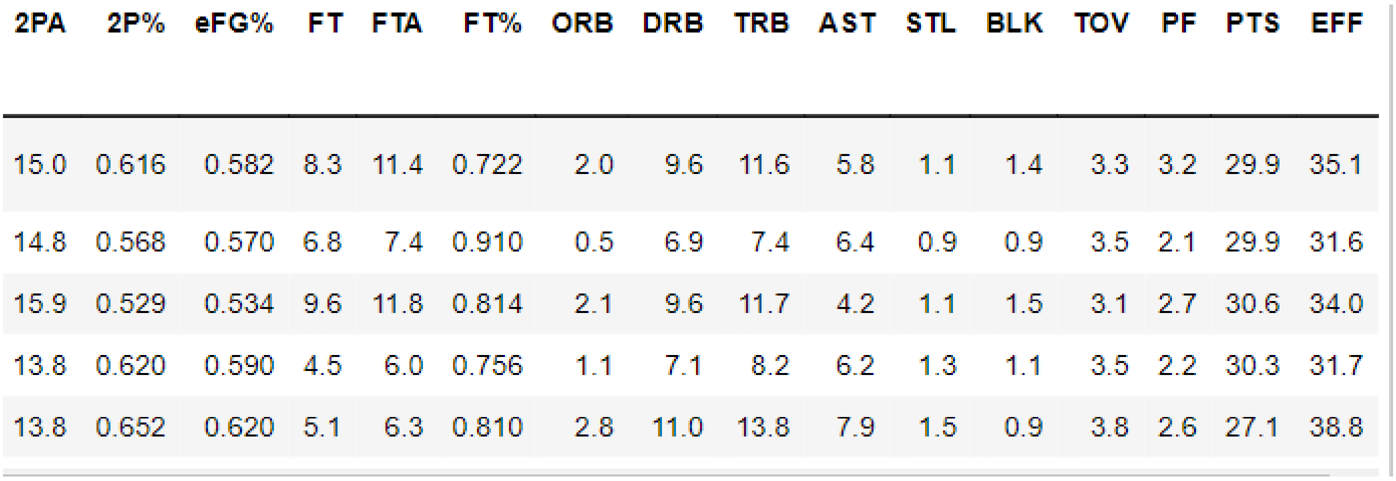
Now that we have a basic understanding about the dataset we realise that there are some variables with almost the same information as that of the others for example 'FG'(*Field goals per game*) and 'FGA'(*Field goals attempts per game*) when taken ratio of gives us 'FG%'(*Field goals percentage*), similarly there are variables such as '3P', '3PA', '2P', '2PA', 'FT', and 'FTA' that can be considered for discarding. But before we disregard these variables without knowing much we can use these values to create a variable that would make better sense to us when we try to look at the bigger picture, for that we created a new variable called **Efficiency (EFF)** in our dataset for which the formula is as follows:

Efficiency = Points +Total rebounds +Assists +Steals +Blocks - (Field Goal Attempts – Field Goal) - (Free Throw Attempts – Free Throws) - Turnovers

Where Points , Total rebounds ,Assists ,Steals are desirable qualities in a player and whereas Field Goal Attempts – Field Goal i.e. failure to make a field goal/wasted attempt and similarly wasted free throw attempts and turnover are undesirable qualities in a player , and so the formula gives us an estimate of the overall performance of the player in the 2021-2022 NBA season. We can use this to **compare** the **overall performance** of each player as in fig.



Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%
12	Giannis Antetokounmpo	PF	27	MIL	67	67	32.9	10.3	18.6	0.553	1.1	3.6	0.293	9.2	15.0	0.616	0.582
154	Kevin Durant	PF	33	BRK	55	55	37.2	10.5	20.3	0.518	2.1	5.5	0.383	8.4	14.8	0.568	0.570
162	Joel Embiid	C	27	PHI	68	68	33.8	9.8	19.6	0.499	1.4	3.7	0.371	8.4	15.9	0.529	0.534
274	LeBron James	SF	37	LAL	56	56	37.2	11.4	21.8	0.524	2.9	8.0	0.359	8.6	13.8	0.620	0.590
290	Nikola Joki?	C	26	DEN	74	74	33.5	10.3	17.7	0.583	1.3	3.9	0.337	9.0	13.8	0.652	0.620



2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	EFF
15.0	0.616	0.582	8.3	11.4	0.722	2.0	9.6	11.6	5.8	1.1	1.4	3.3	3.2	29.9	35.1
14.8	0.568	0.570	6.8	7.4	0.910	0.5	6.9	7.4	6.4	0.9	0.9	3.5	2.1	29.9	31.6
15.9	0.529	0.534	9.6	11.8	0.814	2.1	9.6	11.7	4.2	1.1	1.5	3.1	2.7	30.6	34.0
13.8	0.620	0.590	4.5	6.0	0.756	1.1	7.1	8.2	6.2	1.3	1.1	3.5	2.2	30.3	31.7
13.8	0.652	0.620	5.1	6.3	0.810	2.8	11.0	13.8	7.9	1.5	0.9	3.8	2.6	27.1	38.8

Figure 5

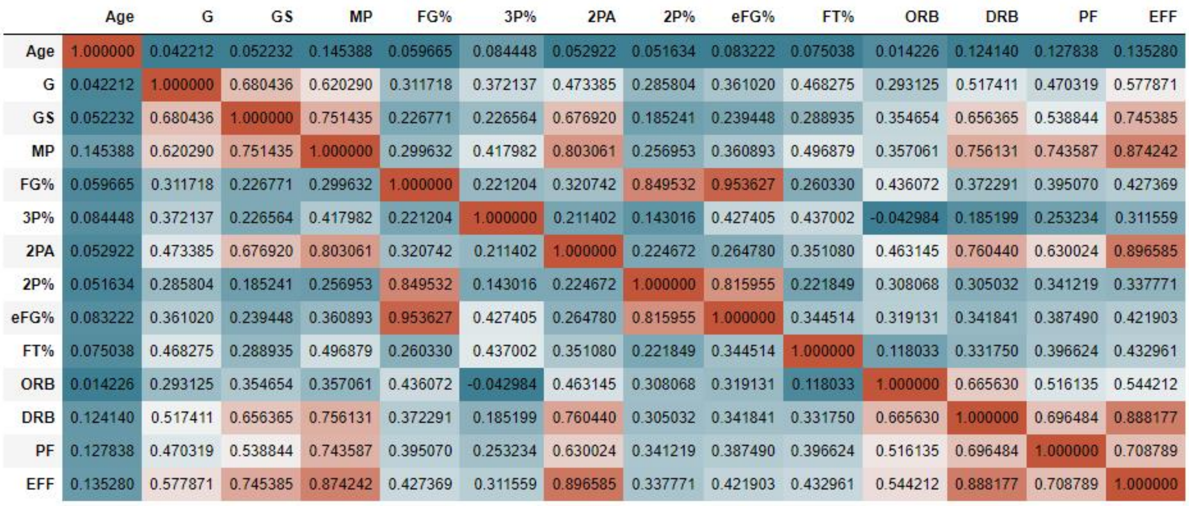
Q2. How are the variables correlated?

Now to see how the variables are correlated we make a correlation matrix for the required variables as discussed earlier. The numbers in the matrix represent correlation coefficients, which measure the strength and direction of the linear relationship between two variables. Correlation coefficients range from -1 to 1, with values closer to -1 or 1 indicating a stronger relationship, and values closer to 0 indicating a weaker relationship.

Here in the plot we can see that MP (minutes played) has a reasonably significant positive association with G (games played) and GS (games started), which makes sense because players who play and start more games are more likely to collect minutes on the court.

EFF (player efficiency rating) shows a reasonably strong positive association with G, GS, and MP, as would be expected given that players who play more games, start more games, and collect more minutes are more likely to have a higher efficiency rating.

Except for 3P% (three-point field goal percentage), which has a weak negative association with FG%, FG% has a weak positive correlation with other variables. This shows that there may be some trade-off between shooting accuracy at various distances and players who excel at one type of shot may not necessarily excel at the other.



	Age	G	GS	MP	FG%	3P%	2PA	2P%	eFG%	FT%	ORB	DRB	PF	EFF
Age	1.000000	0.042212	0.052232	0.145388	0.059665	0.084448	0.052922	0.051634	0.083222	0.075038	0.014226	0.124140	0.127838	0.135280
G	0.042212	1.000000	0.680436	0.620290	0.311718	0.372137	0.473385	0.285804	0.361020	0.468275	0.293125	0.517411	0.470319	0.577871
GS	0.052232	0.680436	1.000000	0.751435	0.226771	0.226564	0.676920	0.185241	0.239448	0.288935	0.354654	0.656365	0.538844	0.745385
MP	0.145388	0.620290	0.751435	1.000000	0.299632	0.417982	0.803061	0.256953	0.360893	0.496879	0.357061	0.756131	0.743587	0.874242
FG%	0.059665	0.311718	0.226771	0.299632	1.000000	0.221204	0.320742	0.849532	0.953627	0.260330	0.436072	0.372291	0.395070	0.427369
3P%	0.084448	0.372137	0.226564	0.417982	0.221204	1.000000	0.211402	0.143016	0.427405	0.437002	-0.042984	0.185199	0.253234	0.311559
2PA	0.052922	0.473385	0.676920	0.803061	0.320742	0.211402	1.000000	0.224672	0.264780	0.351080	0.463145	0.760440	0.630024	0.896585
2P%	0.051634	0.285804	0.185241	0.256953	0.849532	0.143016	0.224672	1.000000	0.815955	0.221849	0.308068	0.305032	0.341219	0.337771
eFG%	0.083222	0.361020	0.239448	0.360893	0.953627	0.427405	0.264780	0.815955	1.000000	0.344514	0.319131	0.341841	0.387490	0.421903
FT%	0.075038	0.468275	0.288935	0.496879	0.260330	0.437002	0.351080	0.221849	0.344514	1.000000	0.118033	0.331750	0.396624	0.432961
ORB	0.014226	0.293125	0.354654	0.357061	0.436072	-0.042984	0.463145	0.308068	0.319131	0.118033	1.000000	0.665630	0.516135	0.544212
DRB	0.124140	0.517411	0.656365	0.756131	0.372291	0.185199	0.760440	0.305032	0.341841	0.331750	0.665630	1.000000	0.696484	0.888177
PF	0.127838	0.470319	0.538844	0.743587	0.395070	0.253234	0.630024	0.341219	0.387490	0.396624	0.516135	0.696484	1.000000	0.708789
EFF	0.135280	0.577871	0.745385	0.874242	0.427369	0.311559	0.896585	0.337771	0.421903	0.432961	0.544212	0.888177	0.708789	1.000000

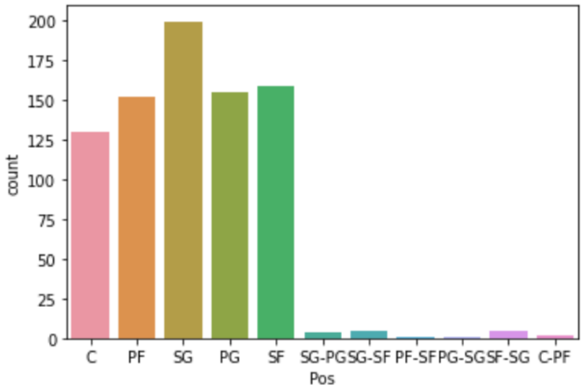
ORB (offensive rebounds) has a weak positive relationship with G and DRB (defensive rebounds), but not with most of the other factors. This could imply that offensive rebounding ability is independent of other skills and is more dependent on effort and positioning.

Personal fouls (PF) have a somewhat significant positive link with G, GS, and MP, which is understandable given that players who play more games and minutes are more prone to commit fouls. Yet, no clear relationships exist between PF and shooting percentages or efficiency rating, implying that fouls are not always symptomatic of poor performance or a lack of competence.

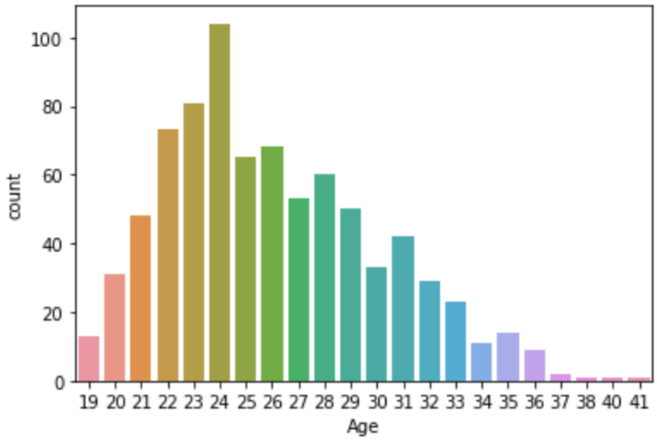
Strong positive connections exist between various shooting percentages (FG%, 2P%, eFG%, FT%) while weak to moderate negative correlations exist between shooting percentages and foul avoidance are important skills for basketball players, and players who excel at these skills tend to have higher efficiency ratings.

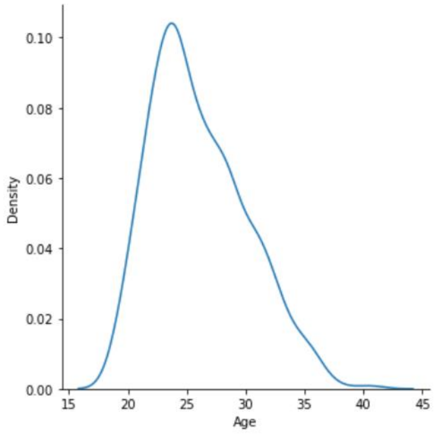
Q3. How many players are there in each position class?

Further trying to explore the data we can see that in plot here that even 'positions' variable had 11 unique values most of the players are clustered in the original 5 position groups (C , PF, SG, PG, SF) and the rest are the hybrid ones that only a few players are a part of and is made of the combinations of the original 5 position groups.

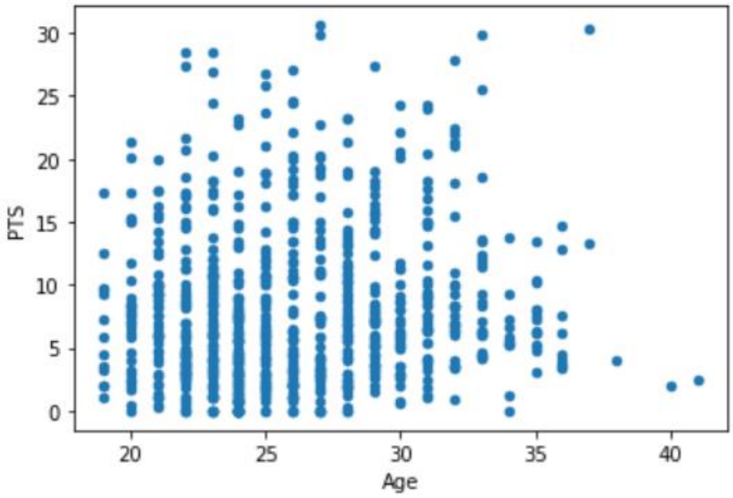


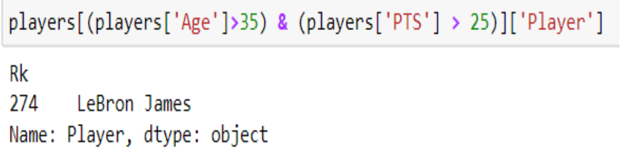
Q4. How many players are there in each age group?
The graph here shows the number of players in an age group , here we also realize that there is a preference for players in the range of 22-26 years as they are the majority in the dataset .





Q5. How are age and points scored related?
Here in the scatter plot we can notice that as the age is increasing the overall points scored decreases towards the end, that means that the younger players are the ones that are scoring most of the points which explains the previous result of why younger players are preferred overall in NBA. Here we also notice that there is only one player that is above the age of 35 and scores more than 30 points. That is none other than LeBron James.



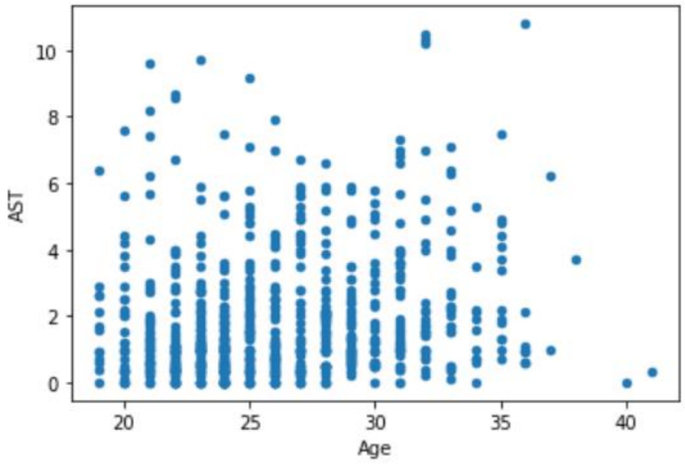


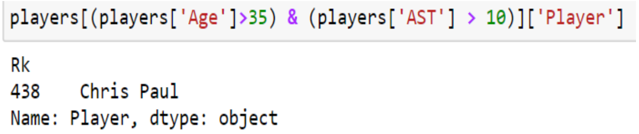
```
players[(players['Age']>35) & (players['PTS'] > 25)]['Player']
```

```
Rk
274    LeBron James
Name: Player, dtype: object
```

Q6. How are age and assists related?

Here in the scatter plot we can notice that as the age is increasing, we see that the number of assists are doesn't have a rapid fall like in the case of points i,e it is more or less consistent (considering till only age 35 because most players retire by that age), the reason for this can be that as the players get old they are prone to more injuries and so the players tend to shift their gameplay towards assisting rather than scoring points because it involves less rough contact with the opponents. Here again we have one player with age above 35 and assists above 10. That player is none other than Chris Paul.

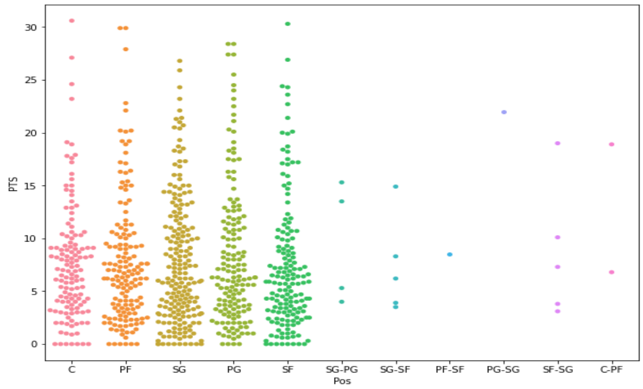


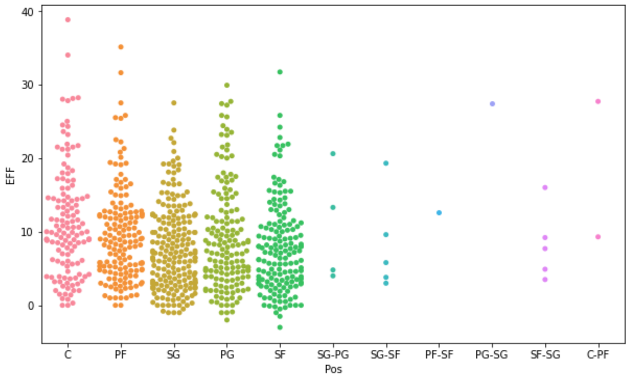


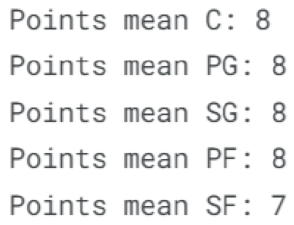
```
players[(players['Age']>35) & (players['AST'] > 10)]['Player']

Rk
438    Chris Paul
Name: Player, dtype: object
```

Q7. Player positions and points distribution and also efficiency distribution?

Here we can see that the importance of positions can't be found in fig1 as the contribution of all positions seem equal and its mean is also almost the same but in fig2 we can see that all position means of efficiency has a variation of 10% and are ordered as C, PF, PG, SG, SF (in descending order)

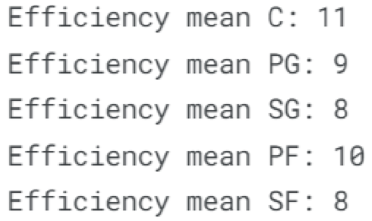






```
Points mean C: 8
Points mean PG: 8
Points mean SG: 8
Points mean PF: 8
Points mean SF: 7
```

Mean of points distribution 1

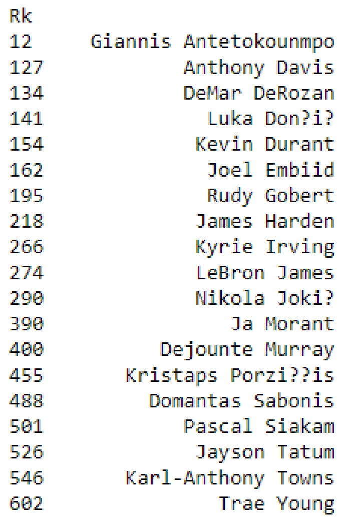


```
Efficiency mean C: 11
Efficiency mean PG: 9
Efficiency mean SG: 8
Efficiency mean PF: 10
Efficiency mean SF: 8
```

Mean of efficiency distribution

Q8. Top 20 players in their prime in their prime?

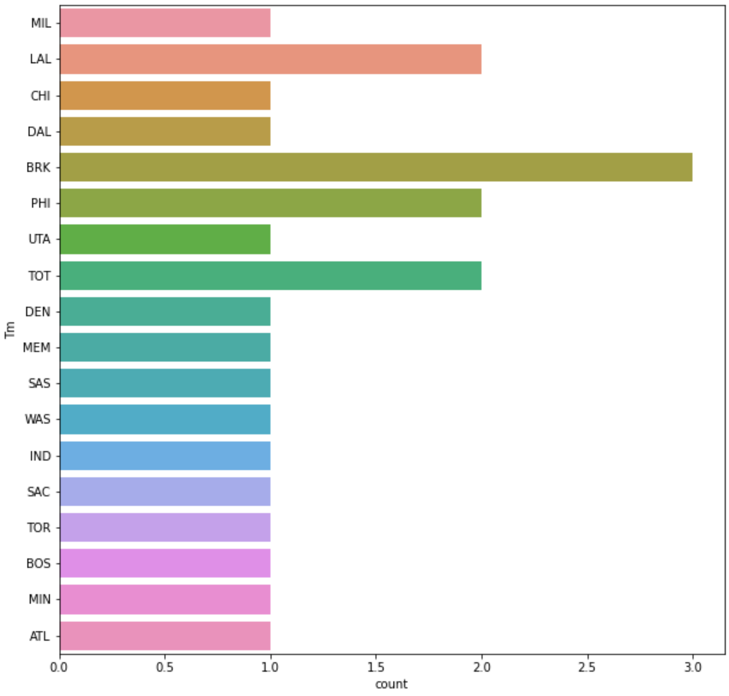
Here we have considered players to be in their prime if they have an efficiency greater than 25, and the list here shows most of the NBA superstars (familiar names) . The figure shows the top 20 most desirable and best performing players in this 2021-2022 season.



```
Rk
12      Giannis Antetokounmpo
127             Anthony Davis
134             DeMar DeRozan
141               Luka Don?i?
154              Kevin Durant
162               Joel Embiid
195               Rudy Gobert
218              James Harden
266              Kyrie Irving
274              LeBron James
290               Nikola Joki?
390                 Ja Morant
400           Dejounte Murray
455        Kristaps Porzi??is
488         Domantas Sabonis
501             Pascal Siakam
526              Jayson Tatum
546        Karl-Anthony Towns
602                Trae Young
```

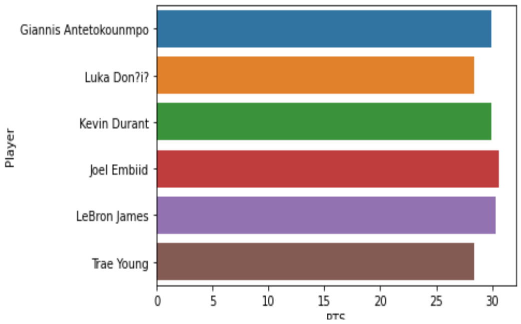
Q9. Teams with the most players in their prime?

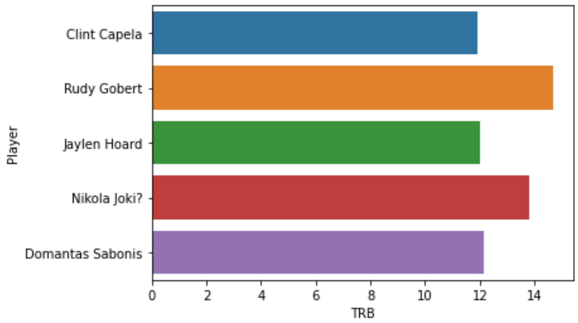
Earlier we found out the players in their prime and from that data we have found out the teams with most of their players in their prime (frequency of prime players in each team). And the top 4 teams are Brooklyn Nets, LA Lakers, Philadelphia Sixers , Toronto Raptors making them prime candidates for winning the league. Also we can interpret here that these are the top 4 teams spending a lot of money to acquire top players.

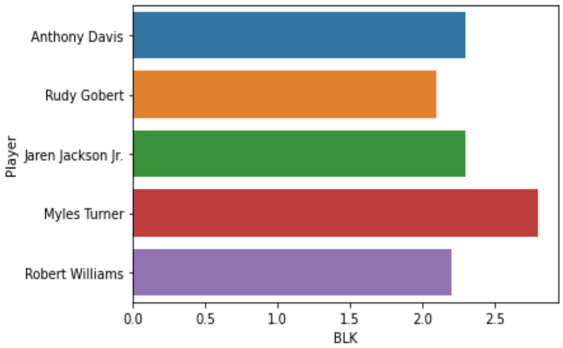


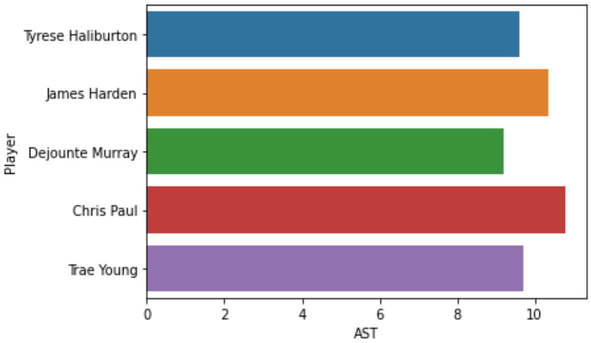
Q10.Top 5 players in Points, Total rebounds, Blocks, Assists, Games played and finally efficiency?

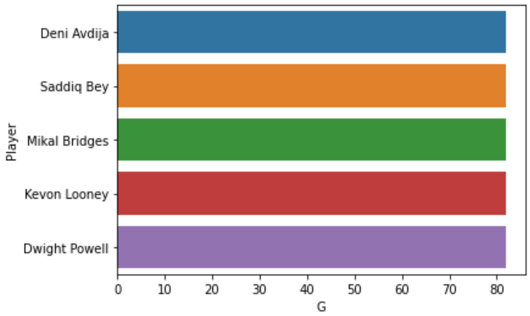
Here in the plots, we can see how different variables have its own different top 5 players from efficiency and so we can say that the player must be good at all the 5 departments to be considered as efficient and valuable so that they can contribute to all aspects for the team to win matches.

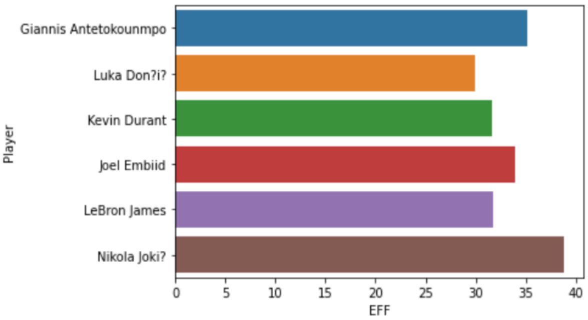








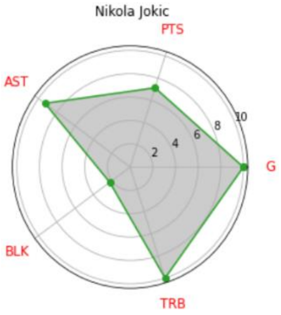


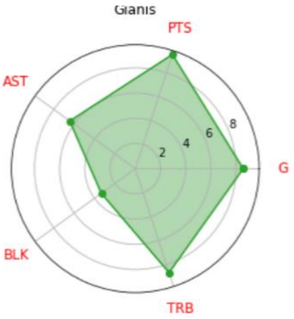


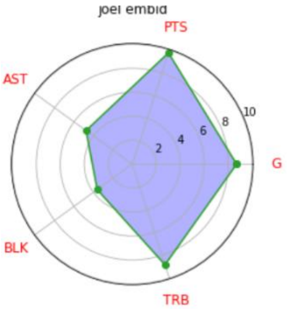
Q11.How good are the individual players from the top 5 efficiency category in the previously stated aspects (Points, Total rebounds, Blocks, Assists, Games played)?

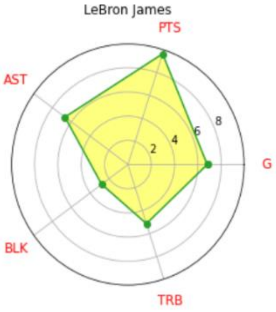
The spider plots are arranged in descending order of efficiency here and we can notice the areas covered by the spider plots also seem in a descending manner, that is because the players with higher efficiency have points further out in the circle in turn covering larger areas.

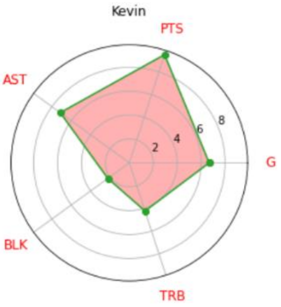
One thing that we can notice here is that LeBron James and Kevin Durant in this case got lesser games to play, but still they managed to score points equivalent to the other players with more games, this could be an indication that the team as a whole couldn't perform well as justified by the low game appearances as they might have been eliminated earlier in the season than the others but still they managed to perform well individually. So in turn we could conclude that if the team as a whole would have performed well then maybe LeBron James and Kevin Durant would have been at the top of the efficiency table.











Q12.How can this analysis be used to create a predictive analysis?

All the above analysis until now was based on the relation/correlation of the variables with each other in the dataset, so the variables which we feel were the most correlated to other variables or basically had influence on other variables (for e.g., Games Played impacted LeBron James's Efficiency rating in turn his overall performance) such variables can be used for the prediction model to increase our model's accuracy and efficiency.

Data Source:
https://www.kaggle.com/datasets/vivovinco/nba-player-stats?resource=download