Exploratory Data Analysis PROJECT REPORT

NBA STATS

Taj Saleh

Student ID: 214172235

Istanbul Sehir University

Computer Science and engineering

June, 2018

Section 1

After taking a look at basketball glossary in NBA website, scanning and understanding the data we are provided, I had
Few questions on my mind:

1 - Does a player with more attempts have a better success attempts rate and more points?
2- What are the effects of age of a player on his performance, to be exact his successful attempts?
3- Does a player with more GP (games played) necessarily have a better performance?

1- Does a player with more attempts have a better success attempts rate and more points?

I will be choosing this question to work with, after checking the data, I will be analyzing data6 that provides me with Ft and FG hit records.
Taking their total ft and FG attempts and total time played I compared them to their successful attempts, also I will be comparing them to the points they got. After that I will be working with with hits attempted and hits made or with hits attempted and points earned.

2- What are the effects of age of a player on his performance, to be exact his successful attempts?

To check if the age of a player affects his performance I will have to check his birthday and the date of matches for every Player, now this data set have many matches for every player, my approach would be going over all those matches that belong to a certain player, then checking at which age his performance of fg and ft attempts is higher, and i will do the same for the rest Of the players, I will pass over this question for two reasons, one is that the answer for this question can be understood that
The more experienced a player is the better, two is that proving that a certain age is great for players can't always be true.

3- Does a player with more GP (games played) necessarily have a better performance?

To answer this question I will have to make some comparisons between players
With high GP and players with low GP.this comparison will cover amounts games played and success players made in them.
I also choose to skip this question because the way I can measure performance in this data is by points or hits made and We know that naturally the more games played the higher the chance to get points which means better performance If we had better measure for performance it would be better.

My hypothesis:
"There is a significant difference between the performances (hits made) of players with high number of attempts and other players with low number of attempts regrading there performance (hits made and points)"

My null hypothesis:
"There is no difference between the performances (hits made) of players with high number of attempts and other players with low number of attempts regrading there performance (hits made and points)"

Section 2

I will be using "basketball_player_allstar.csv" in this research, from this file I can get FT and FG attempted and made hits for each player and also it will provide me with the points that each player earned.
I will need to combine FT and FG hits to get a total hit attempts and Total made hits, the points are ready to use, and I can create a data frame with the following columns: ID, Total Attempts, Total Made, Points. I believe after that my dataset will be ready to work on.

Before

	player_id	last_name	first_name	season_id	conference	league_id	games_played	minutes	points	o_rebounds	...	steals	blocks	turnovers	personal_fou
0	abdulka01	Abdul-Jabbar	Kareem	1978	West	NBA	1	28	11.0	NaN	...	NaN	NaN	NaN	Na
1	abdulka01	Abdul-Jabbar	Kareem	1969	East	NBA	1	18	10.0	NaN	...	NaN	NaN	NaN	Na
2	abdulka01	Abdul-Jabbar	Kareem	1988	West	NBA	1	13	4.0	NaN	...	NaN	NaN	NaN	Na
3	abdulka01	Abdul-Jabbar	Kareem	1987	West	NBA	1	14	10.0	NaN	...	NaN	NaN	NaN	Na
4	abdulka01	Abdul-Jabbar	Kareem	1986	West	NBA	1	27	10.0	NaN	...	NaN	NaN	NaN	Na

5 rows × 23 columns

After

	fg_attempted	ft_attempted	fg_made	ft_made	points
0	12.0	2.0	5.0	1.0	11.0
1	8.0	2.0	4.0	2.0	10.0
2	6.0	2.0	1.0	2.0	4.0
3	9.0	2.0	4.0	2.0	10.0
4	9.0	2.0	4.0	2.0	10.0
5	15.0	4.0	9.0	3.0	21.0

Section 3

descriptive statistics, Firstly, I found mean, max, min, std, variance, and mode values for Attempts, made attempts and points using built-in functions.

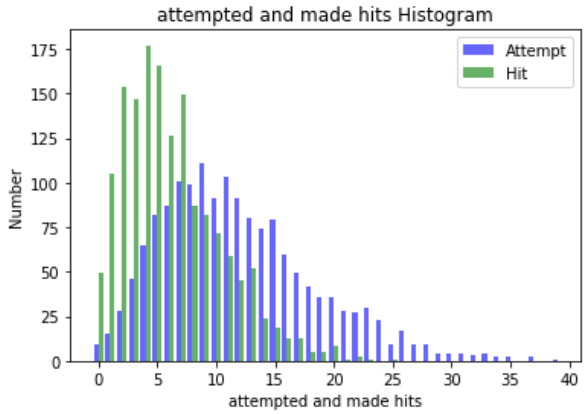
```
Statistics For Total Hits Attempted
Mean
11.9545454545
Median
11.0
Std
6.56743332503
variance
43.1311804787
Mode:
0    9.0
dtype: float64
Min:
0.0
Max:
39.0
```

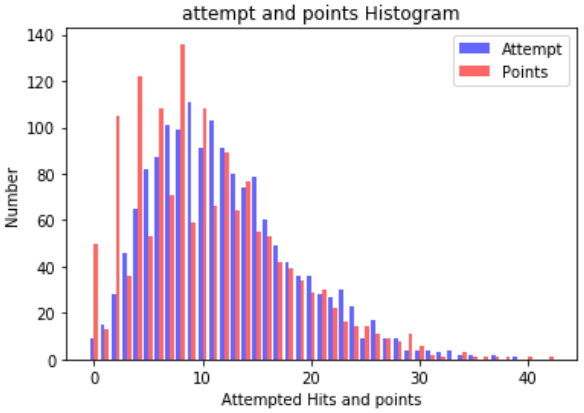
```
Statistics For Hits made
Mean
6.22151088348
Median
5.0
Std
4.19506415335
variance
17.5985632507
Mode:
0    4.0
dtype: float64
Min:
0.0
Max:
25.0
```

```
Statistics For Points
Mean
10.7496798976
Median
10.0
Std
6.96265892976
variance
48.4786193722
Mode:
0    8.0
dtype: float64
Min:
0.0
Max:
42.0
```

Histograms

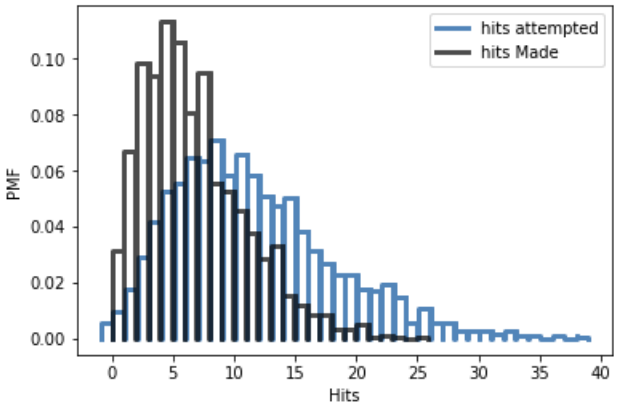
Figure one: green resembles the attempts succeeded while the blue resembles the total attempts.
Figure two: red resembles the points earned while the blue resembles the Total attempts.

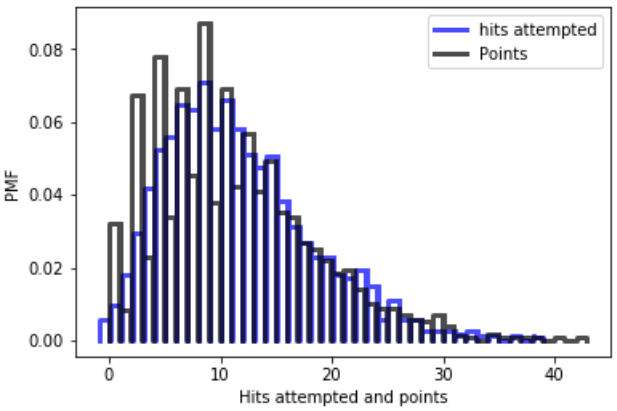




PMF

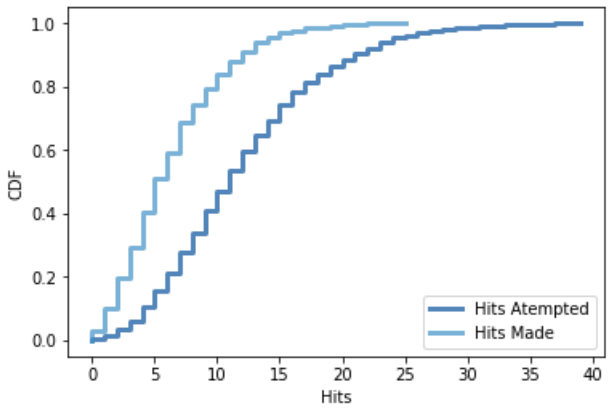
Figure one: black resembles the attempts succeeded while the blue resembles the total attempts.
Figure two: black resembles the points earned while the blue resembles the Total attempts.

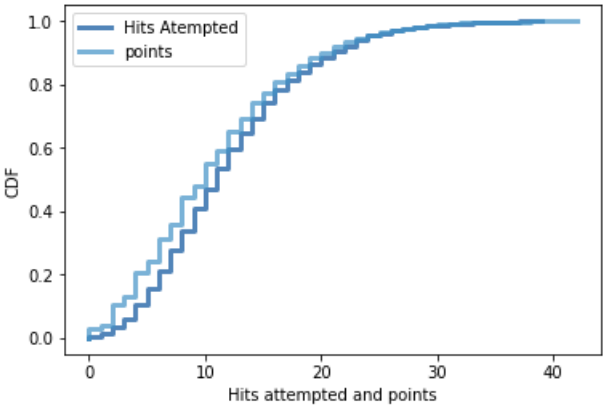




CDF

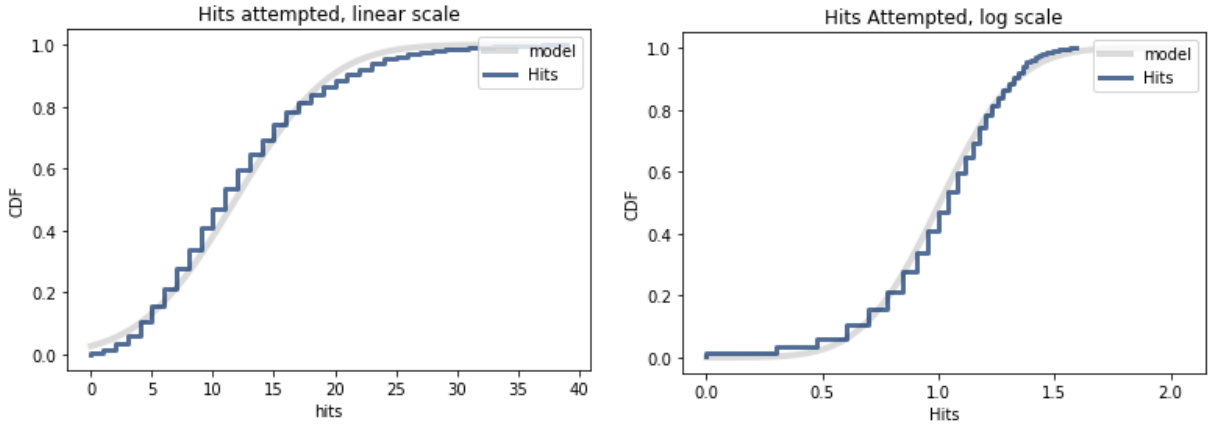
Figure one: light blue resembles the attempts succeeded while the blue resembles the total attempts.
Figure two: light blue resembles the points earned while the blue resembles the Total attempts.





Section 4

Here I tried two model to try fit the data, Linear model (figure one) and lognormal model (figure two).
The first distribution of hits attempted and a normal model, which is not a very good fit.
The second distribution of attempted hits and a lognormal model, plotted on a log-x scale, this model is a better fit for the data.

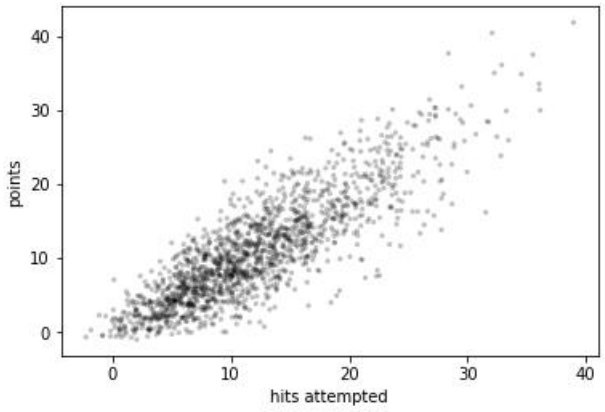


Section 5

First I started by plotting the scatter plot, i noticed many things:

- the data fall in obvious columns because they were rounded off.
- We can reduce this visual artifact by adding some random noice to the data.
- The columns are gone, but now we have a different problem: **saturation**. Where there are many overlapping points.
- We can usually solve the saturation problem by adjusting alpha and the size of the markers, s.

Finaly i ended up with this plot



İn this part correlation and covariance found. Covariance gives tendency and correlation give how much strength relationship there is between two variables.
I used 'Total Attempted Hits' and 'Points' as parameters in cov function. The covariance came up as high as 39.8541787918 which means the points tend to together.
Then I found correlation by using Pearson's correlation, 0.8620358874528079.

Usually if the result is positive then the correlation is positive which means if one variable is high the other is high too.but, when it is negative it means when one variable is high the other is low.
By multiplying the correlation by 100 i can get the percentage of the correlation 86.20%.

Section 6

To test my hypothesis I need to do the following:

- Define a null hypothesis.
- Compute a p-value, which is the probability of seeing the apparent effect if the null hypothesis is true.
- Interpret the result. If the p-value is low, the effect is said to be statistically significant, which means that it is unlikely to have occurred by chance.

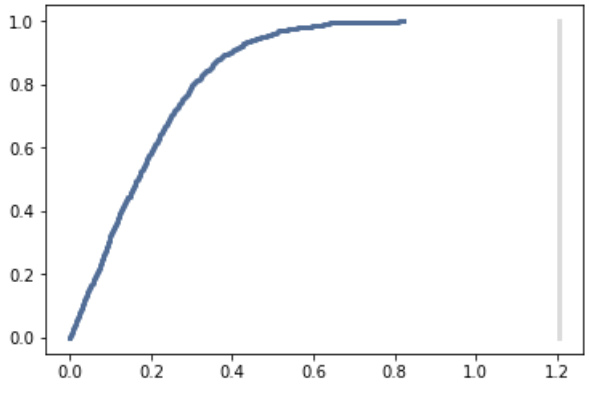
I used “HypothesisTest” to test my hypothesis.

It is called “Test Statistics” if there is a high relationship between win a game and having the total of lose get less. Nevertheless, it is called “Null Hypothesis” if there is no relationship.

Here, I used Hypothesis Test in order to Test mine:

If we run the analysis with total attempts and points, the computed p-value is 0;
after 1000 attempts, the simulation never yields an effect as big as the observed difference.

İf we only tried 100 attempt only from the data,we get a p-value of 0.208, the observed effect about 20% of the time. So this effect is **statistically significant**.



the plot is for total attempts

But considering all, I would report that $p < 0.001$.

Section 7

In conclusion, it seems that number of attempts is connected to the points earned, as we showed above, we need to conclude that overall there is a significant difference between the two as our p-value proved to us, I used histogram, pmf, cdf to help see if there is a relationship, also I found covariance and correlation, and ended by testing my hypothesis and output was there is no relation between high hit attempts and points earned in a game.