İSTANBUL
ŞEHİR
ÜNİVERSİTESİ

Final Project Report

Basketball Data Analysis

CS240 – Exploratory Data Analysis

By Mehmet Baysan

215061235

MALEK JAMAL ABDULAH MALKAWI

30.05.2018

Introduction

In this project, a basketball team's dataset containing complete statistics for NBA in 2011-2012 will be used to analyze the relations between the points and assists. Pandas, Numpy, Thinkplot and Thinkstats2 will be used for analyzing process. A 3 different questions will be determined about this statistics and the process will be applied on one of them. The relationship between the points and assists numbers checked through observations and calculations

Section 1

After brainstorming, watching basketball videos and scanning some information. I had three questions in my mind:

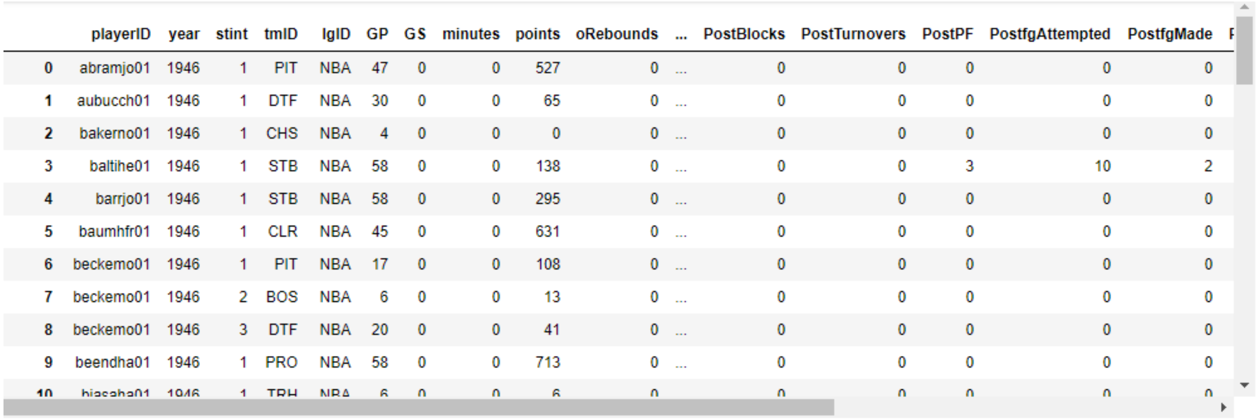
1) What is the relationship between points and assists numbers?
2) Can we analyze the relationship between the years and the points?
3) Do increase the number of assists effect the point's numbers?

I have checked some stats about the game and the rules. I realized that to be able to have a good analysis, I should analyze a very different perspectives separately. But it is hard to look at all of them due to having short time and missing data. That's why I pass the first and second question. That is why I have selected the third question. It is crucial and specific question. In this report I will make some comparisons between different point's numbers with the assists numbers. Thus, my hypothesis is that "Increasing the assists number effect the point's number positively". Also, my null hypothesis would be "Assists number has no effect on point's number"

Section 2

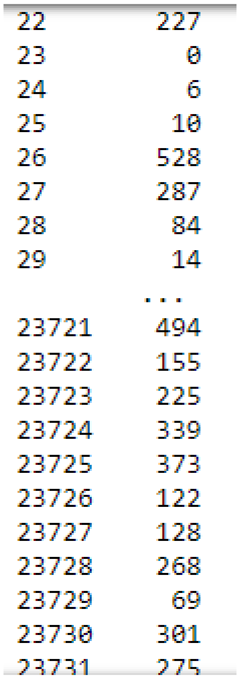
The dataset that I am going to use is "basketball_players.csv". "Assists" and "Points" are the columns to be used. The data is ready, I do not have to clean it or drop any values. I created two variables called 'assists' and 'points'. I am going to find the correlation between them.

The Whole Data of "basketball_players.csv"

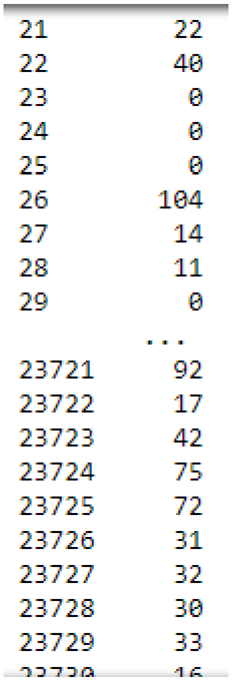


	playerID	year	stint	tmID	lgID	GP	GS	minutes	points	oRebounds	...	PostBlocks	PostTurnovers	PostPF	PostfgAttempted	PostfgMade
0	abramjo01	1946	1	PIT	NBA	47	0	0	527	0	...	0	0	0	0	0
1	aubucch01	1946	1	DTF	NBA	30	0	0	65	0	...	0	0	0	0	0
2	bakerno01	1946	1	CHS	NBA	4	0	0	0	0	...	0	0	0	0	0
3	baltihe01	1946	1	STB	NBA	58	0	0	138	0	...	0	0	3	10	2
4	barrjo01	1946	1	STB	NBA	58	0	0	295	0	...	0	0	0	0	0
5	baumhfr01	1946	1	CLR	NBA	45	0	0	631	0	...	0	0	0	0	0
6	beckemo01	1946	1	PIT	NBA	17	0	0	108	0	...	0	0	0	0	0
7	beckemo01	1946	2	BOS	NBA	6	0	0	13	0	...	0	0	0	0	0
8	beckemo01	1946	3	DTF	NBA	20	0	0	41	0	...	0	0	0	0	0
9	beendha01	1946	1	PRO	NBA	58	0	0	713	0	...	0	0	0	0	0
10	biasaha01	1946	1	TRH	NBA	6	0	0	6	0		0	0	0	0	0

Some of Point Column at left and Assists Column at right



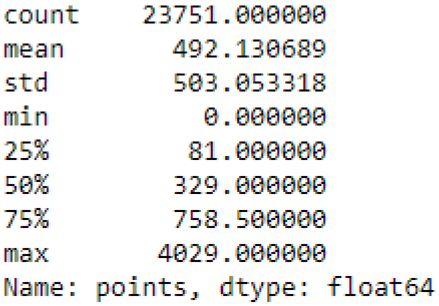
```
22         227
23           0
24           6
25          10
26         528
27         287
28          84
29          14
          ...
23721      494
23722      155
23723      225
23724      339
23725      373
23726      122
23727      128
23728      268
23729       69
23730      301
23731      275
```



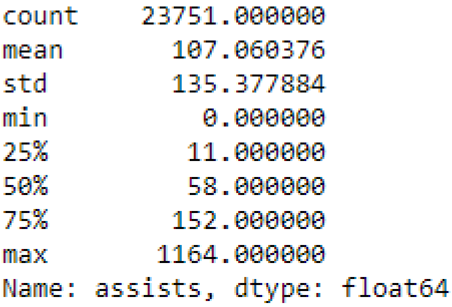
```
21          22
22          40
23           0
24           0
25           0
26         104
27          14
28          11
29           0
          ...
23721       92
23722       17
23723       42
23724       75
23725       72
23726       31
23727       32
23728       30
23729       33
23730       16
```

Section 3

By using describe() which is a built in functions we can have more than 5 descriptive statistics as following about the both. Left one is 'Points' and the right is the 'Assists'.

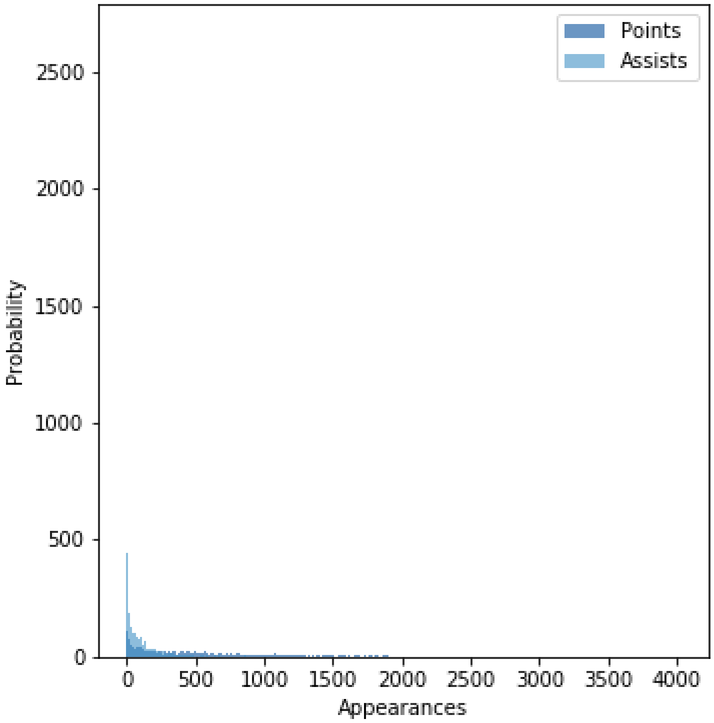


```
count    23751.000000
mean       492.130689
std        503.053318
min          0.000000
25%         81.000000
50%        329.000000
75%        758.500000
max       4029.000000
Name: points, dtype: float64
```

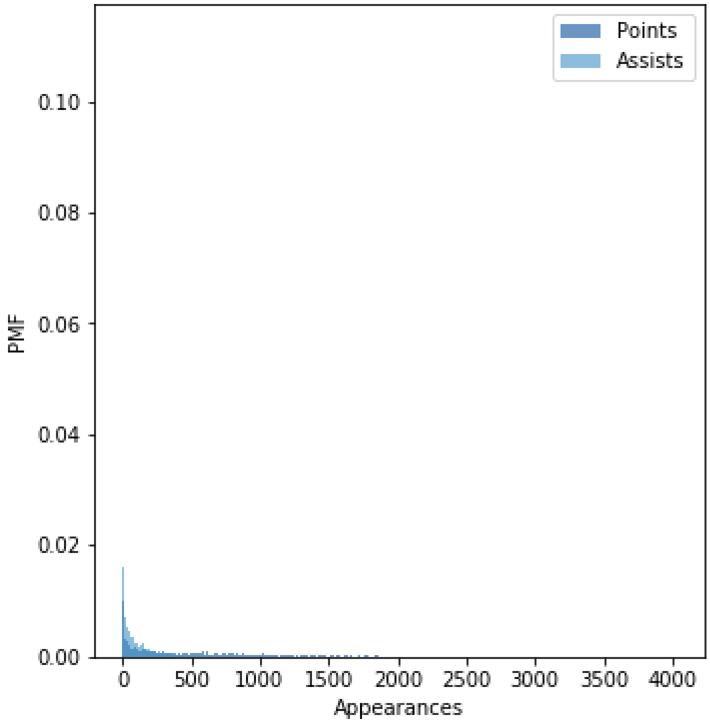


```
count    23751.000000
mean       107.060376
std        135.377884
min          0.000000
25%         11.000000
50%         58.000000
75%        152.000000
max       1164.000000
Name: assists, dtype: float64
```

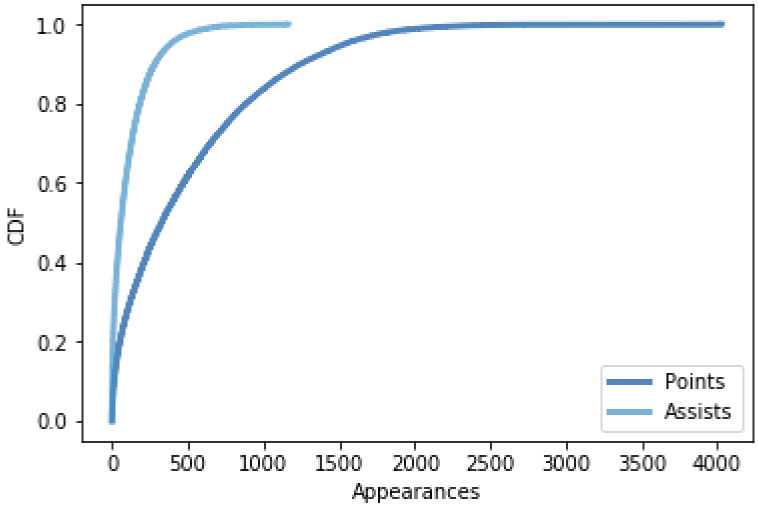
Histogram of the data gives the underlying frequency distribution of the set in magnitude two groups. I used thinkstats2 and preplot for that.



PMF is another way to represent distribution. PMF gives us the probabilities for the discrete random variables.



CDF gives us a clearer picture just by displaying the curvature they form.

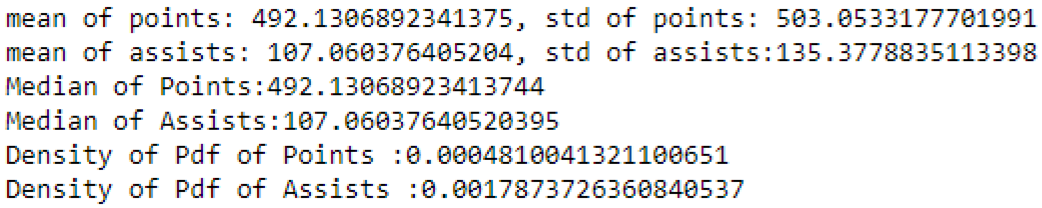


These graphs tell us the relation between the two groups. Overall, they have an average relation

Section 4

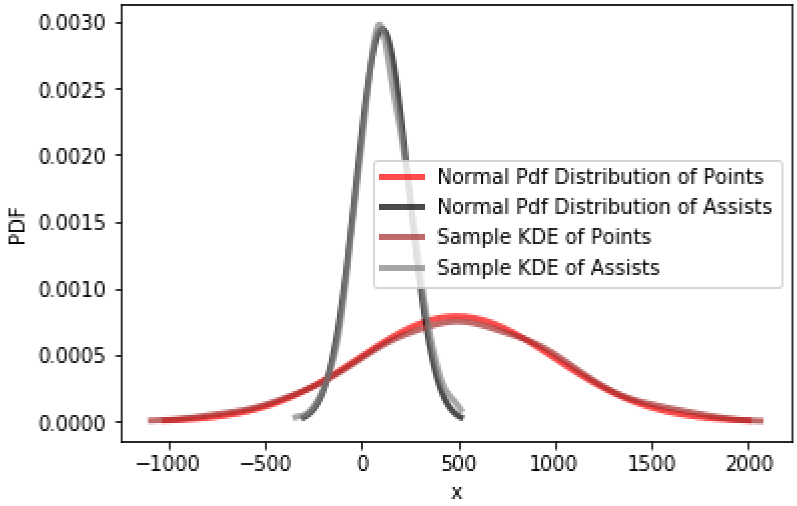
I used Normal PDF distribution for this data to be able to see the density of the distribution.

Firstly, I calculated the mean and standard deviation to be able to plot the PDF distribution



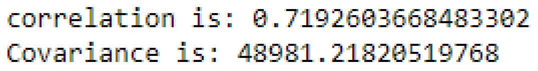
```
mean of points: 492.1306892341375, std of points: 503.0533177701991
mean of assists: 107.060376405204, std of assists:135.3778835113398
Median of Points:492.13068923413744
Median of Assists:107.06037640520395
Density of Pdf of Points :0.0004810041321100651
Density of Pdf of Assists :0.0017873726360840537
```

I calculated them by in-built functions. After that I found the kernel density estimation by randomize the means and standard deviation and iterate it 1000. The figure below shows the results

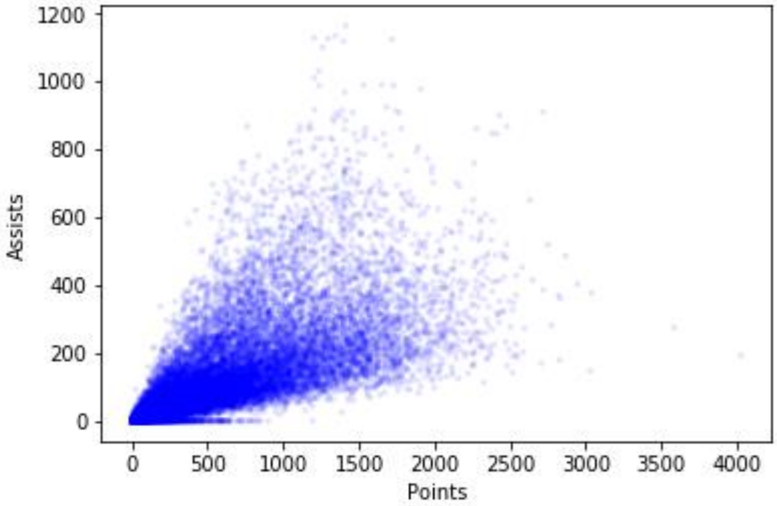


Section 5

At this part, i will use covariance test to have tendency of two variables to vary together and correlation test to have the strength of the relation between them. I calculated the correlation which has a range -1 to 1. If the r is closer to 1 or -1, it means they are closely related and the opposite is true and zero means no relation. Due to the value we have we can say that they are closely related to each other. There is a high and positive relation.



```
correlation is: 0.7192603668483302
Covariance is: 48981.21820519768
```



The figure above shows the visualization of the relation by scatter plot. It shows that they have a high similarity

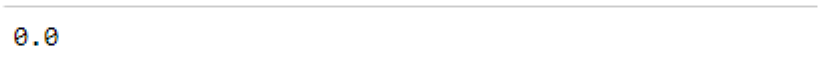
Section 6

At this section, I will apply a hypothesis test to answer the question and know is it true or not.

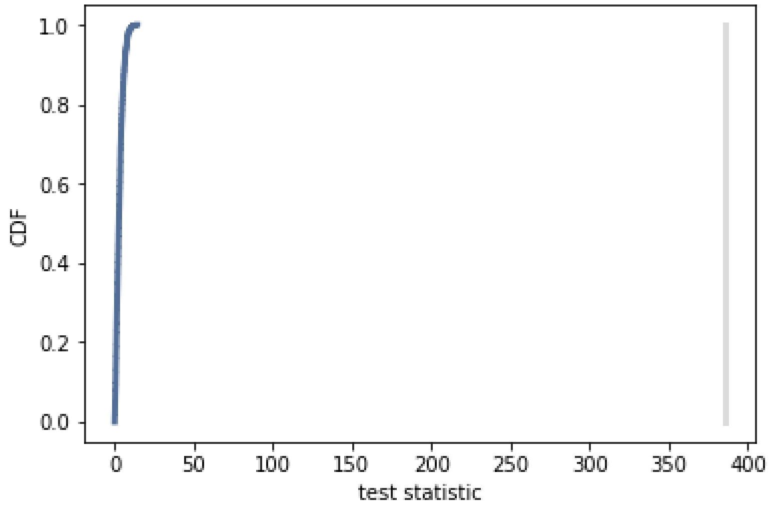
Test Statistics: There is a high relationship between increasing the number of assists and the points. If they increase it, they get more points.

Null Hypothesis: There is no relations between assists numbers and points.

I found the P-Value ZERO which means it is statically significant.



```
0.0
```



A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis. Thus, we reject the null hypothesis. As a result, i reject my null hypothesis.

Section 7

To take everything into consideration, I found a different statistical information about the data to be able to answer the question and make sure that my hypothesis is true. I applied my hypothesis test and I had a very strong values shows that it is true. I found my p value as 0 which means statistically significant and I find the correlation of them that ensure the correlation. Then, I test my hypothesis and I could reject the null because they were correlated closely and my p value is less than 0.05. Therefore if the number of assists increases, the number of points that they can get will increase. The relationship is very strong.

References

Think Stats: Probability and Statistics for Programmers