CS 240 Exploratory Data Analysis Project

Ahmet Ensar Köprülü
215241187

June 1,2018

Section 1:

Possible questions are :

- Relationship between offensive rebounds and points.
- Relationship between attendance and winning.
- Relationship between race and height.

Main question of this project is what is the relationship between basketball players' offensive rebounds and their points. The reason behind the choice of question is that it is more observable from other options.

The test statistic is relationship between points of basketball player and offensive rebounds of basketball players.

The hypothesis is that there is no relation between points of basketball players and offensive rebounds of basketball players.

Section 2:

The variables in project are "df" as main data frame, "filteredDf" as a sub data from of "df" contains points and rebound columns and does not contains inapplicable values or outliers. "Points" as a column which contains data of points that basketball players have. "oRebounds" as a column that contains offensive rebounds data of basketball players.

In this project firstly i will go over points variable and analysis in different distributions and graphs. After that i will begin to analyze relation between offensive rebound numbers and points of basketball player. In order to achieve this firstly "df" had a lot of null and 0 values which will affect whole project and analysis. Rather than dropping all the null values, i decide the save them by replacing with the mean value of "points". Secondly i handled the 0 values by filter the "df" in between 1 and max value of points. Thus, all datas in project prepared to analyze.

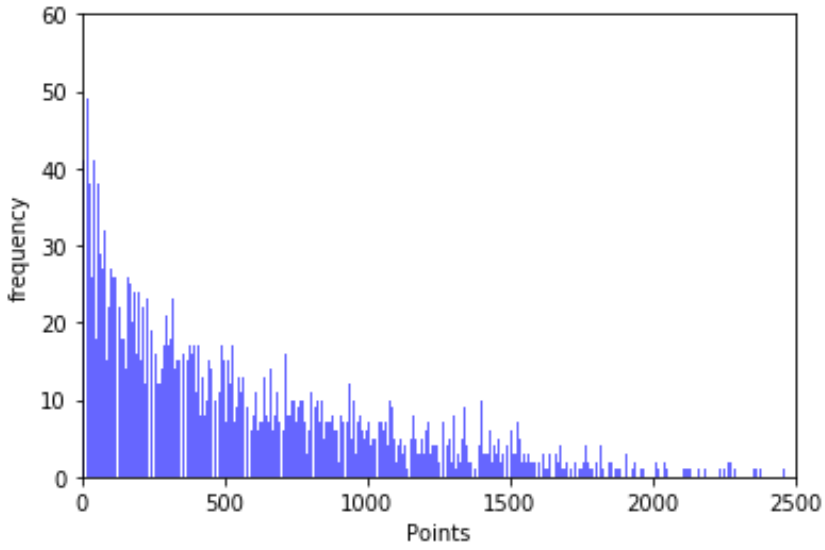
Section 3:

Some statistics about hypothesis of project:

- Mean value of basketball players point is 542.68

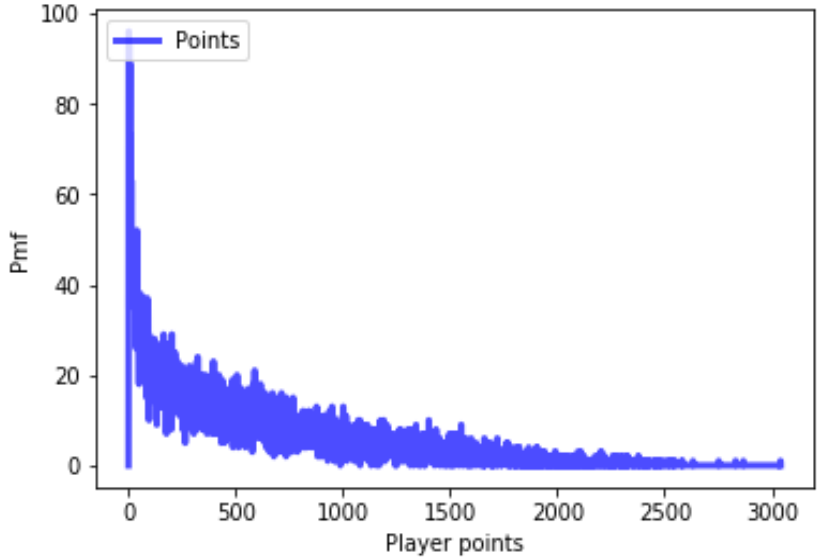
- Variance value of basketball players point is 253870.62
- Standard Deviation of basketball players point is 503.85
- Median value of basketball players points is 396.0
- Mode value basketball players point is 2

Histogram for points of basketball players:



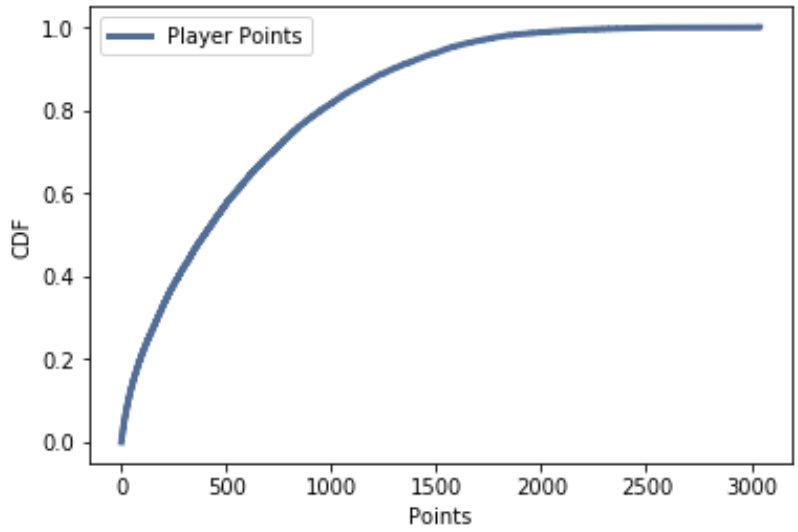
According to histogram we can say median of basketball players points is around 500-600 and it placed on the left of the graph it shows that graph is distributed asymmetric. Moreover, most of the players has points between 1 and 200 especially overwhelming majority of basketball player has only 1 points.

Pmf for points of basketball players:



According to pmf we can say Probability of players who had 1 point is much more than others. We can expect if new player join the NBA league, most probably he will had point between 1 and 50. There is huge difference of probability between maximum and minimum probability of points.

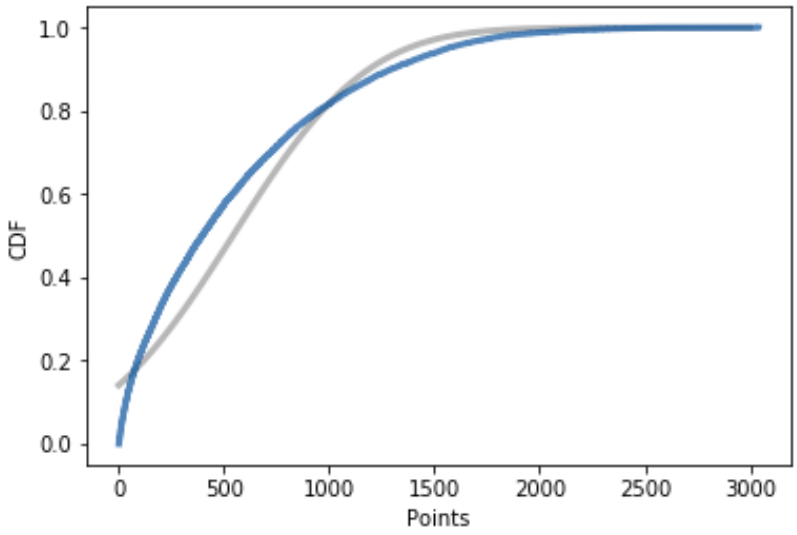
Cdf for points of basketball players:



According to cdf we can say most of the player had less than 2000 points. However, especially players who had points between 1 and 1000 creating eighty percent of NBA league. Moreover, only nearly less than 1% of basketball players had 2000 and 3000.

Section 4:

In this project normal probability plot used in order to modelling:



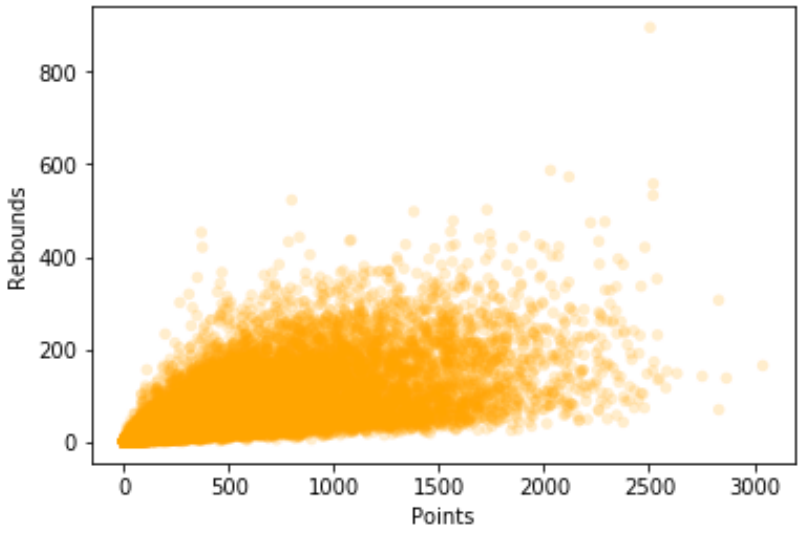
According to the model plot our cdf plot of data is not completely fitting with our cdf plot of model. According to plots percentage of basketball players who have points between 1000 and 2000 are more than our actual plot. However, percentage of basketball player who have points between 1 and 1000 are less in model plot than actual plot.

Section 5:

At this part correlation between points and offensive rebounds found with two methods:

- Spearman rank correlation is found 0,77
- Pearson correlation is found 0,64

Scatter plot of points and offensive rebound numbers:



According to scatter plot since spearman rank correlation bigger than pearson correlation, as we expected there are few outliers for example players who had 2500 points more offensive rebounds number than players who have 3000 points. At this point we can expect player who have more points has more offensive points as we can see on most of the players clusters since correlation scores more close to one.

Section 6:

To check whether an observed correlation is statistically significant, i will run a permutation test. For the permutation test firstly test statistic should be found and it will be the absolute value of correlation between points and offensive rebounds columns.

I need to samples to simulate the experiment while calculating p value. In order to obtain samples i will create model of sample and it will be the permutation of points column and actual

offensive rebounds. Thus, every time i simulate the experiment i will have shuffled points column and constant offensive rebounds column.

In order to find p-value, i will simulate the experiment 1000 time and calculate the test statistic value of each experiment. It will allow me to compare each experiments result with the actual test statistic value. I need number of test statistics which is greater or equal to actual test statistic. After all p-value will be equal to number of test statistics which is greater or equal to actual test statistic divided by number of number of experiment simulated.

Lastly i will compare p-value with the threshold in order to decide our null hypothesis statistically significant or not. There two possible result, my p-value will be greater than threshold which is 0.05 for this test or p-value less or equal to threshold. At the second outcome i can say our hypothesis statistically significant which means the event cannot be occurred by chance.

In my test the p-value was zero or too small number and threshold was 0.05. Thus, our hypothesis is statistically significant.

Section 7:

Everything into consideration, my hypothesis is statistically significant since p-value found as a result of test is lesser than 0.05. Hypothesis of project marked at the start was the that there is no relation between points of basketball players and offensive rebounds of basketball players. Since hypothesis found statistically significant we claimed there is relationship between basketball players and their offensive rebound numbers. Therefore, correlation scores which found close to 1 confirmed.