CS 240 PROJECT REPORT

Ahmet Salih Kaycioglu / 216570446

Section - 1

Question 1 : Do players who have more turnovers, play in less games?

Question 2: Do players who play in more games, score more points?

Question 3: Do teams which steal more balls, score more points?

For analysis i am going to choose the 2nd question. Because it aims to show the relation between points and games. Of course it has other factors like the position of the player but aside from this in this report we are going to analyze the relationship between the player points and number of games the player played in. So my null hypothesis is going to be “The players who play in more games does not score more points.”

Section – 2

For my hypothesis i am going to use the “basketball_players.csv” file which contains the data about the players. I assigned the points and the games played in their respective variables ready to use and analyse.

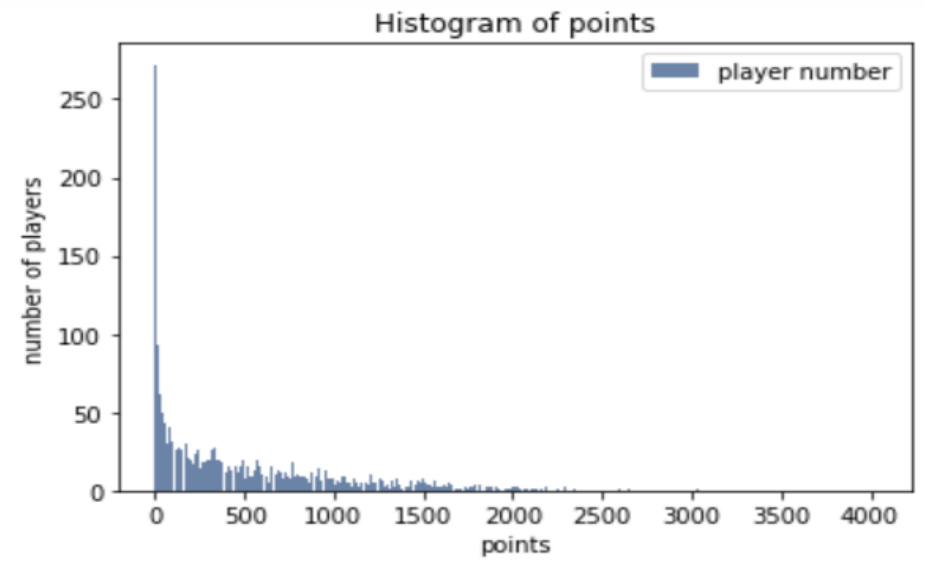
Section – 3

First i collected some statistics about the variable i am going to use which is points data. I used several functions to gather:

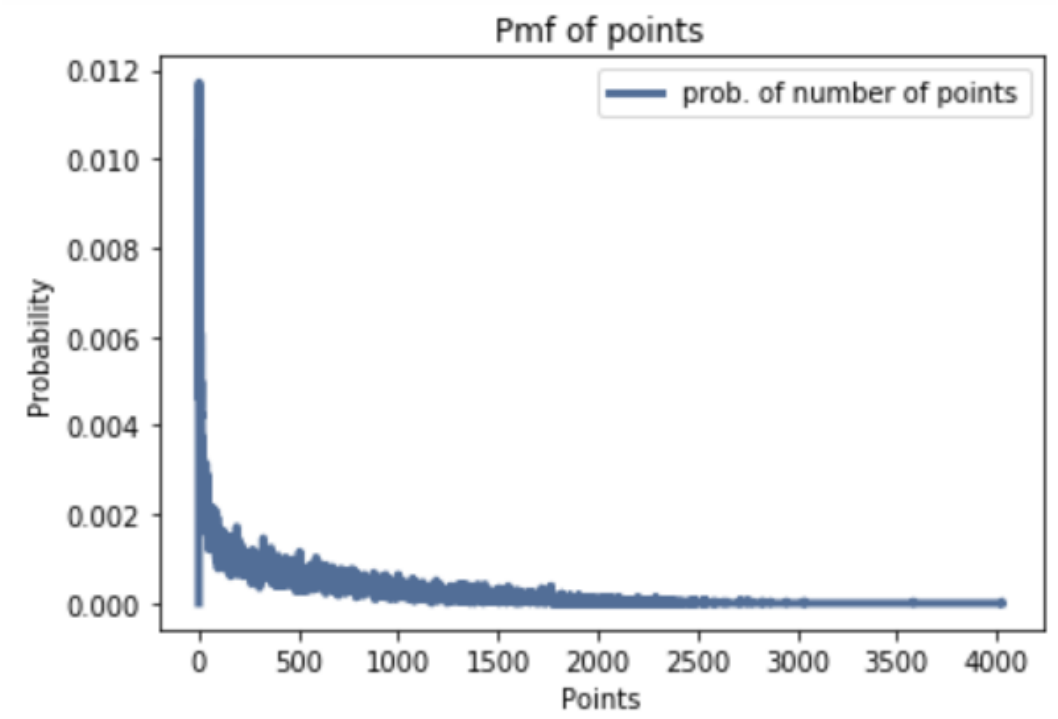
- mean:503
- variance:253000
- standard deviance:502
- median:344
- mode:2

From these statistics i understand that in my data which has over 23.000 entries, the mean points a player scored in a season is 503 with a standard deviance of 502. However if we look at the actual data we can see player scores that are much higher than 1000. The mode of my

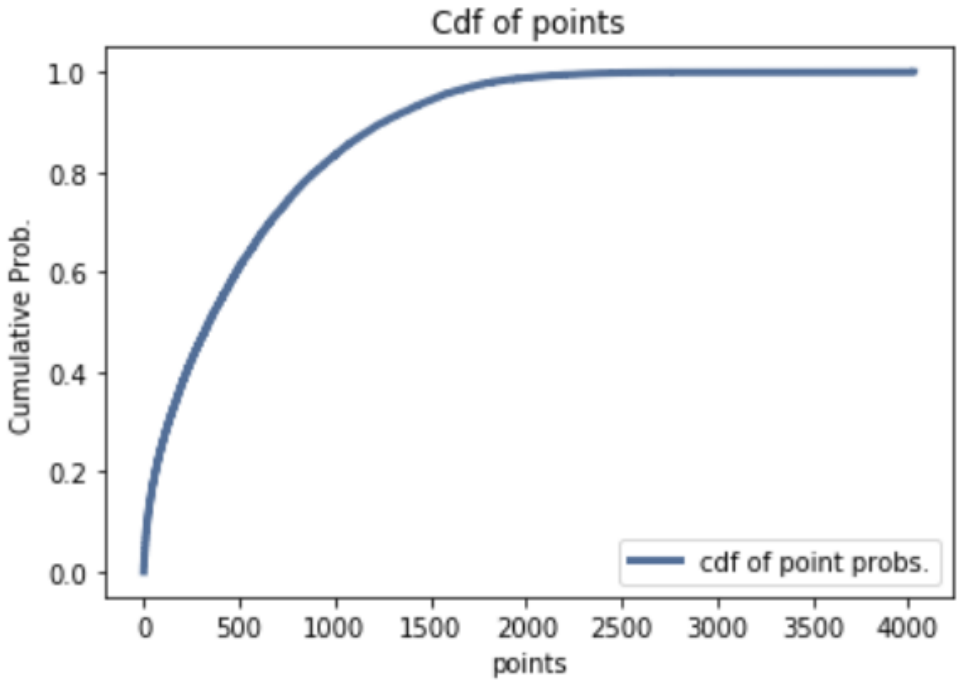
data is 2 which is understandable as many players may get less game time or play in a position that is less likely to take shots.



- As we can see histogram of the total points the players made in a season kind of resembles half of a normal distribution plot. The mode of my data was 2 which can be seen clearly in this histogram as the single highest column in the beginning of the plot.



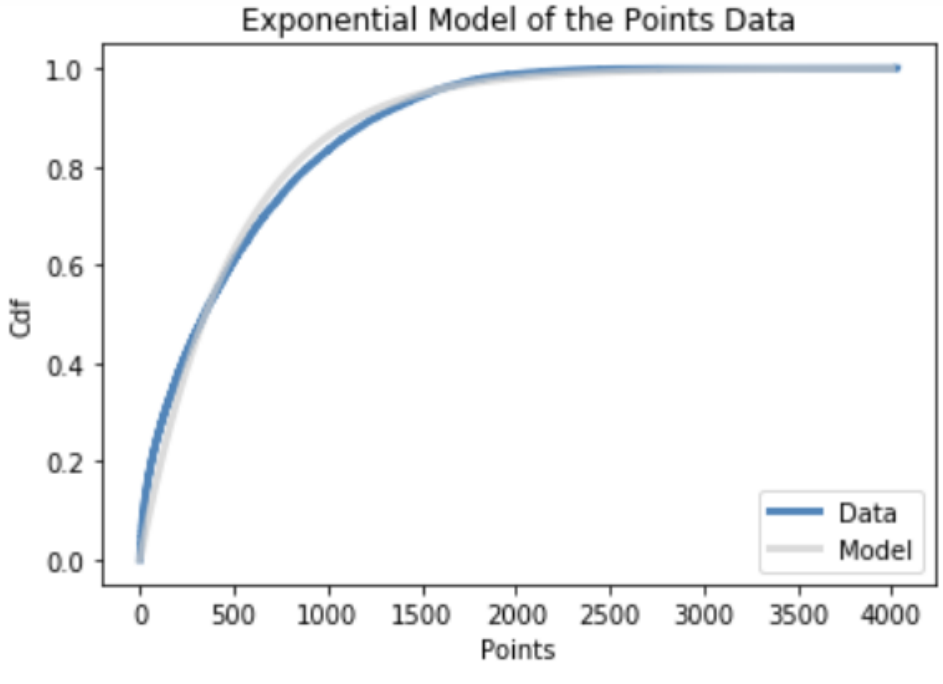
- We can see from the probability mass function the probability of each points made in a season.



- As we can see from the cumulative distribution function, even if there are more than 3000 entries that scored more than 1000 points, our data reaches the total probability around 1500-2000 points.

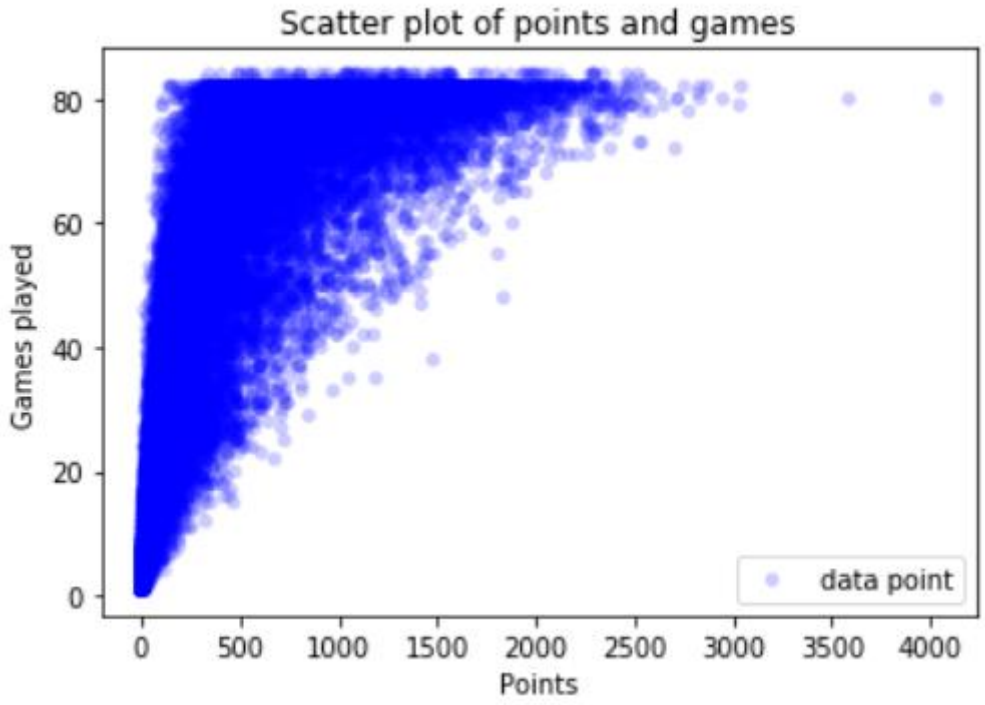
Section – 4

- After some tries, i decided that the best distribution that represented my data was exponential. The reason was players cannot score less than 0 points so normal distribution had to fail at some point. Exponential distribution very much resembles my data because the high scoring players are much more rare than the usual players whose total points are around the mean of the data.



- The model struggles in the beginning part and the part after the mean. The beginning part can be explained by players who scored less points than average in a season is higher than the model expected. And the latter part can be explained by the lack of people caused by the beginning part. Aside from these the model represents my data around the mean and towards end pretty accurately and overall the model somewhat fits my data well.

Section – 5



- Here we can see the relation between points and games played variables in a scatter plot. From this plot we can guess that the pearson and the spearman correlation results will be positive. We can get this conclusion from the positive kind of slope the whole picture has. Now we calculate the spearman and the pearson correlation coefficients.

Pearson Correlation : 0.7396324101123817

Spearman Correlation : 0.8703939701785037

- We can see that the points players made and the games players played are highly correlated. Spearman is a little bit higher because of it's slightly higher tolerance to the more scattered data.

Section – 6

In this part i tested the correlation between the variables to see if it happened by chance or not. For my test statistic i used the actual pearson correlation between the variables. And for testing i changed the order of the points data and tested the correlation again.

Actual Correlation: 0.7396324101123817
P-Value: 0.0

I was not convinced by the p value being so low so i manually tried the procedure for finding p value. In all of my manual trials none of the results were even close to the actual correlation. In the end i was convinced that my p value is really that low so my hypothesis must be wrong.

Section – 7

In this analysis i reached the conclusion that the players points scored is highly related to the games player played. I analysed the points data and saw that in a season majority of players scored around 500 points. Then i made a model for points data. I decided that an exponential distribution model represented my data the best because my data had no entries that is less than 0 and the number of players that scored higher points got less towards the high point marks. Then i tested my hypothesis whether the scores and the number of games played are related to each other or not. The p value i got in the end suggested that the variables were highly related with each other