CS240 PROJECT

**Name: Malik EKİCİM**

**Student No: 216301933**

Introduction:

In this project I am going to analyze a basketball data which is player properties and match results. We will be analyzing the data by using modules like Pandas, Numpy, ThinkPlot, and Thinkstats2. This modules help to plot the data and analysis.We should provide 3 question about these basketball datas.

**Part 1:**

In my project, I determine 3 different questions

· What is the effect of the awards for the coaches performance ?

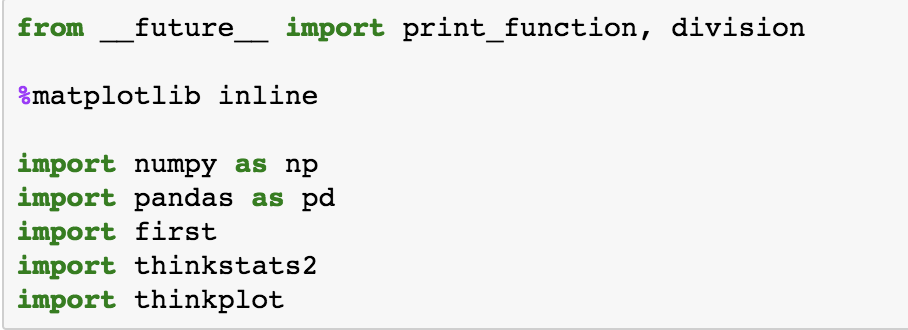
· Does the high of the basketball player affect the amount of time played in all star game ?

· Do basketball team win the match at home when they play away?

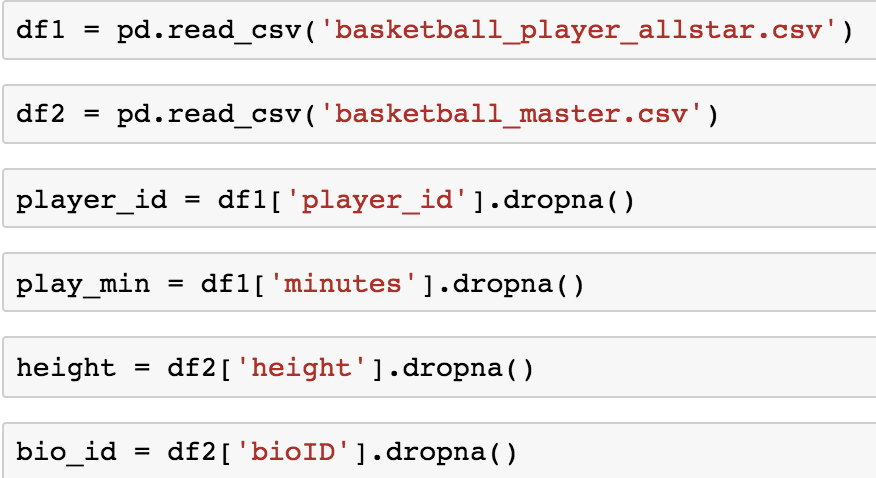
Once I thought of my questions and answered them, I made the hypothesis Does the high of the basketball player affect the amount of time played in all star game ?

**Part 2:**

I am going to use two csv file which are basketball\_player\_allstar.csv and basketball\_master.csv .I will use the playerid,minutes, height and biID. I’m going to observe if there is any relationship with player height and player play time in the match. I am going to observe the similarity between the player height and player play time.

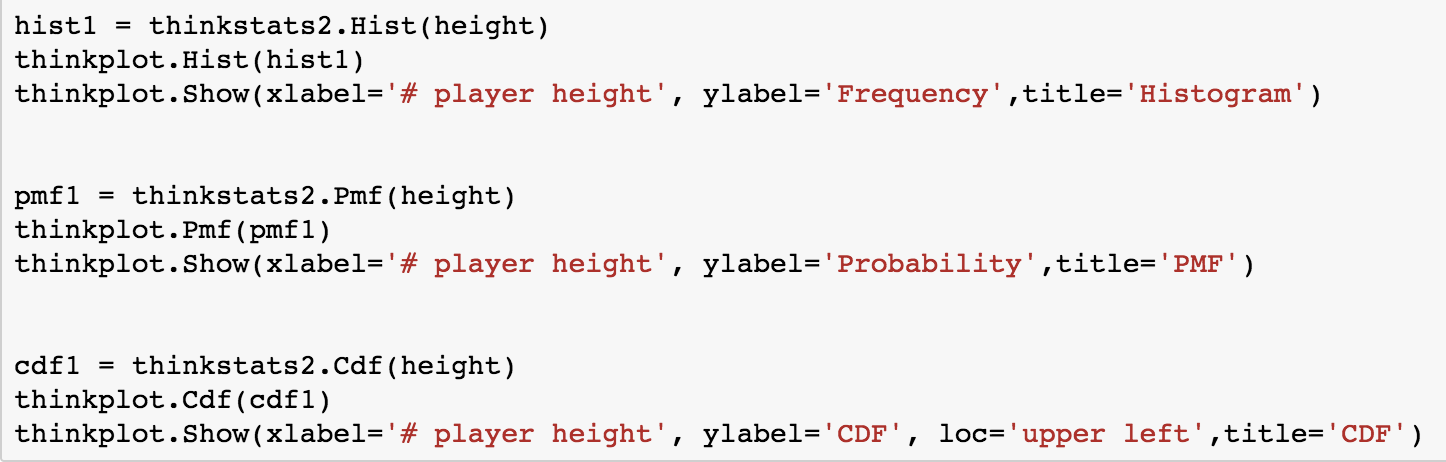


This modules I will use on my project. I will get easier to my job. This is the ready library they have the some modules and some ready code. I will just get the code.



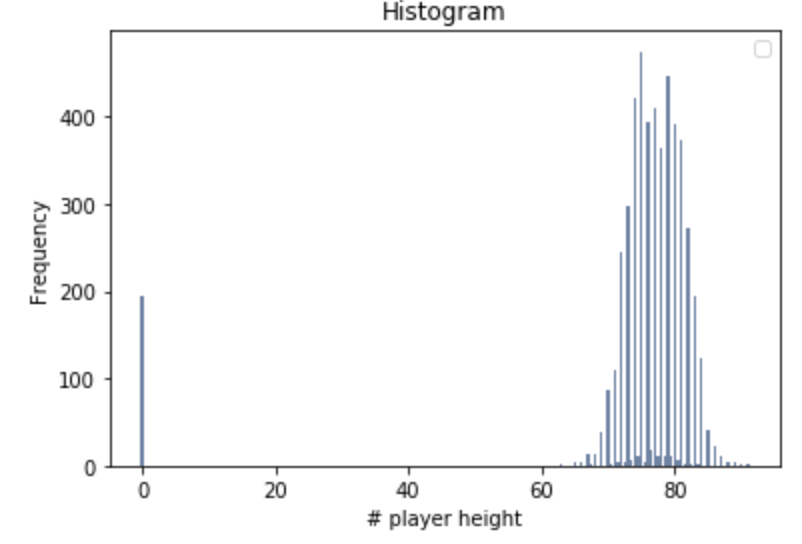
Firstly, I defined my dataframe and I cleaned my data empty cells with dropna.Also I get the my coloums and I defined the name myself.

**Part 3:**

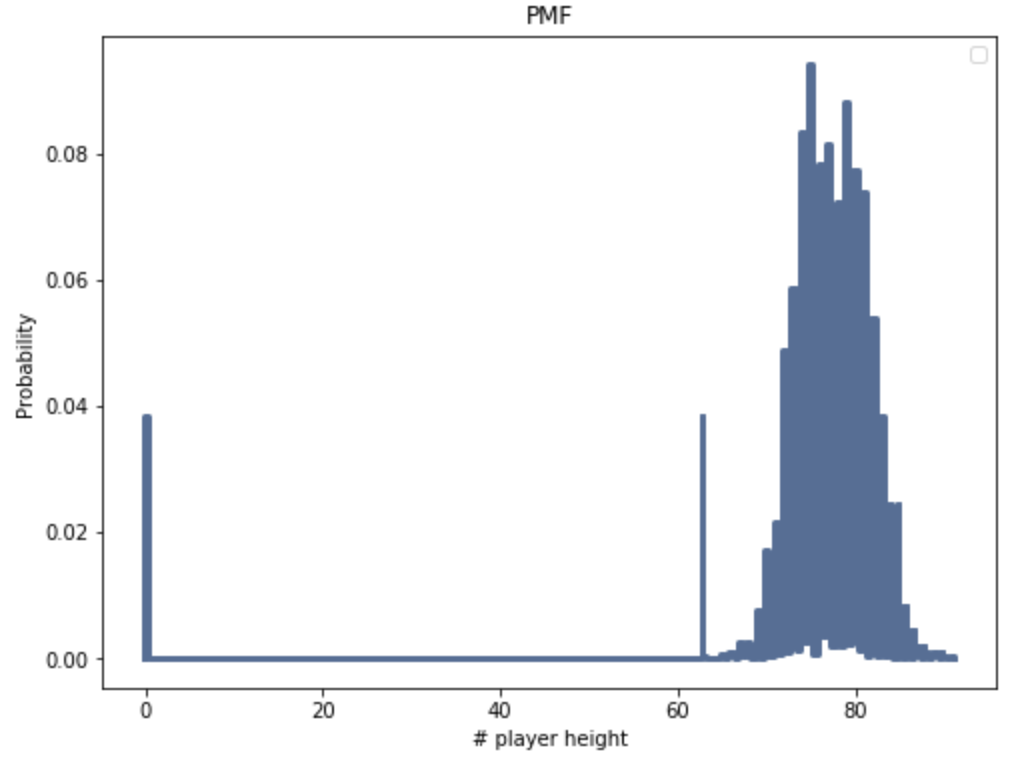


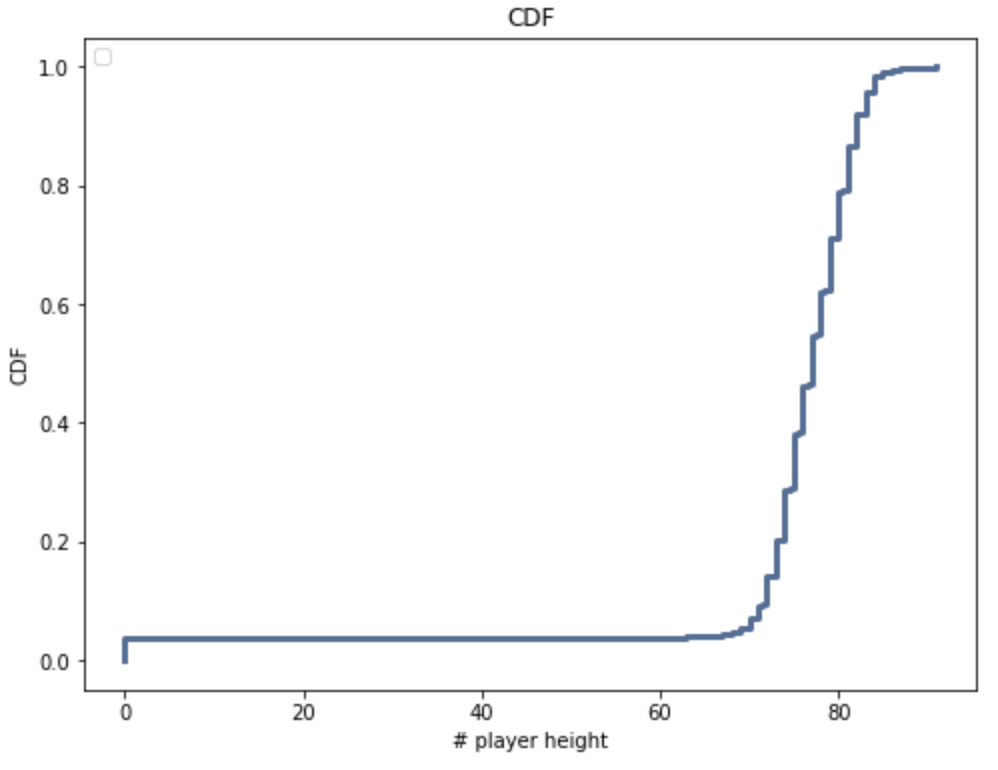
Here is the code for the plot histogram, PMF and CDF for the player height. When I plot the dataframe of course I use thinkplot. I just call the functions and my plots are ready.

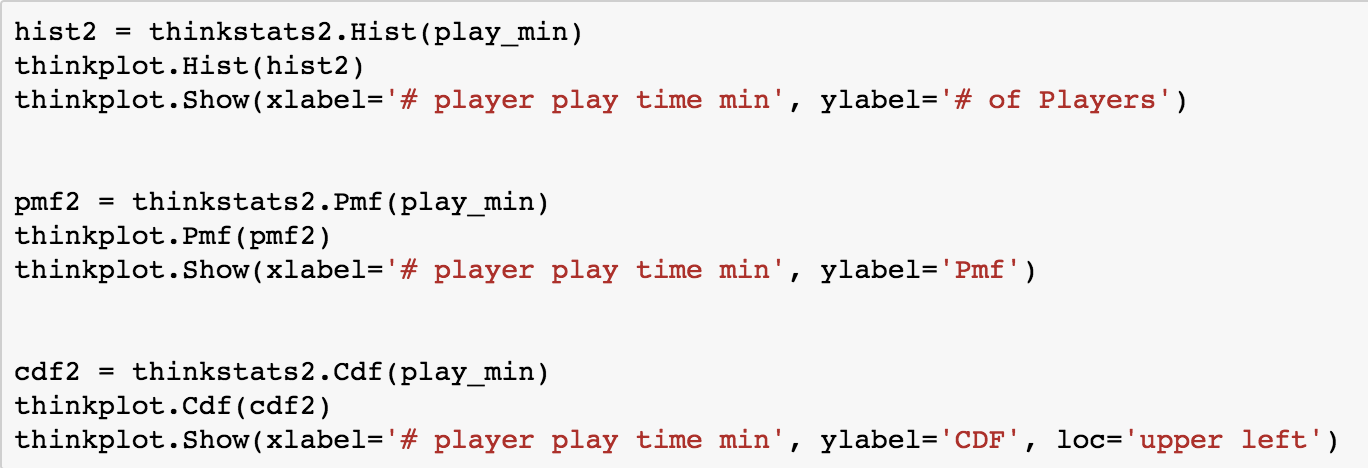
Show of player height histogram



Show of player height PMF.

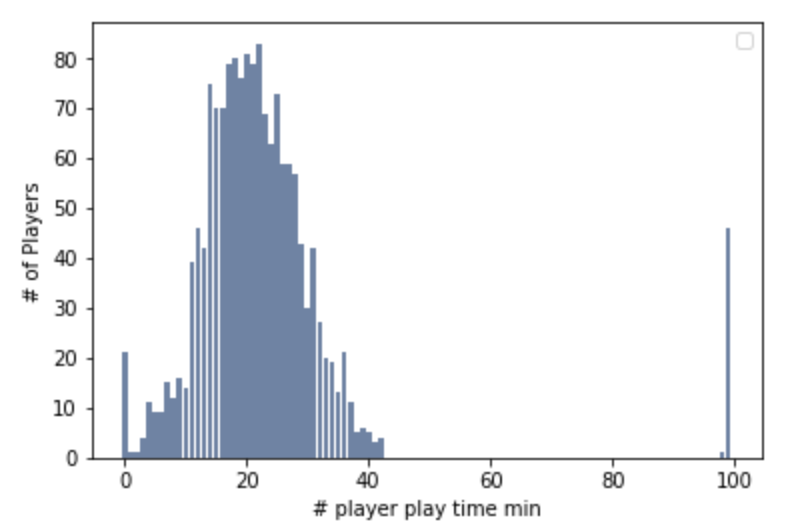


Show of player height CDF

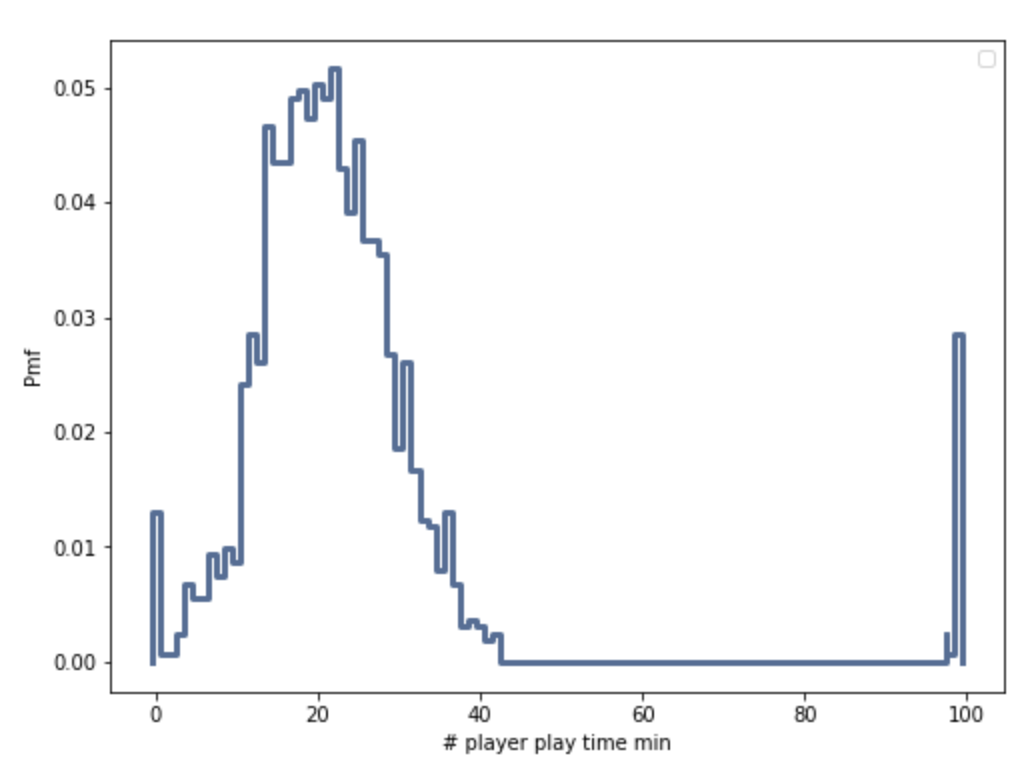


Here is the code for the plot histogram, PMF and CDF for the player play time on the match. When I plot the dataframe of course I use thinkplot. I just call the functions and my plots are ready.

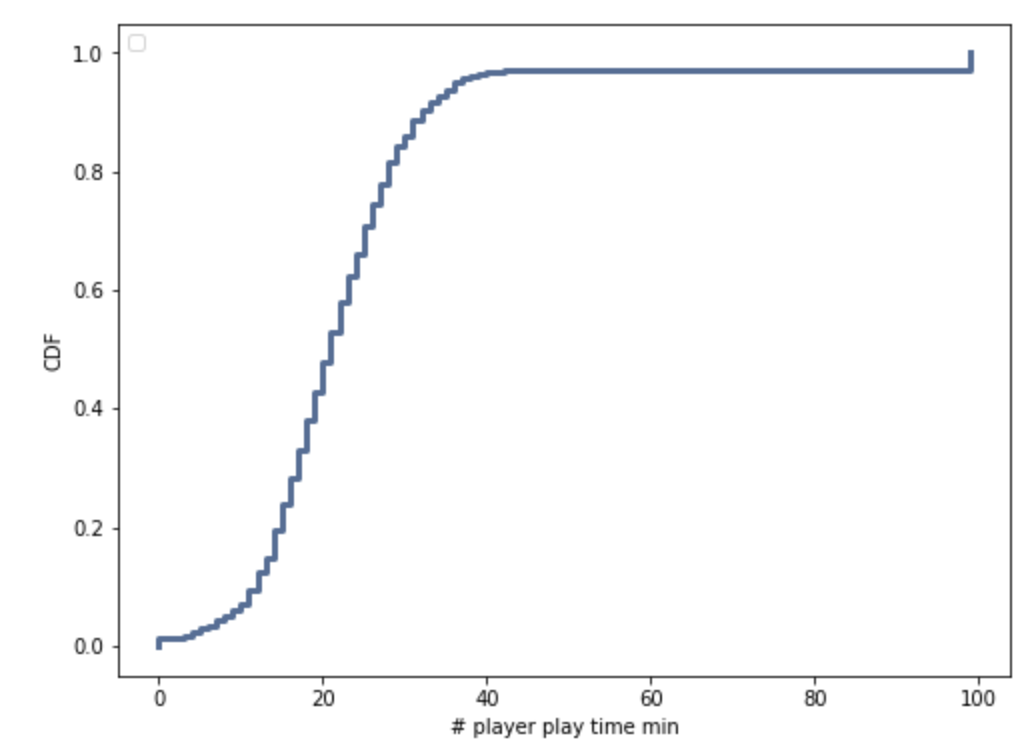
Show of player play time histogram

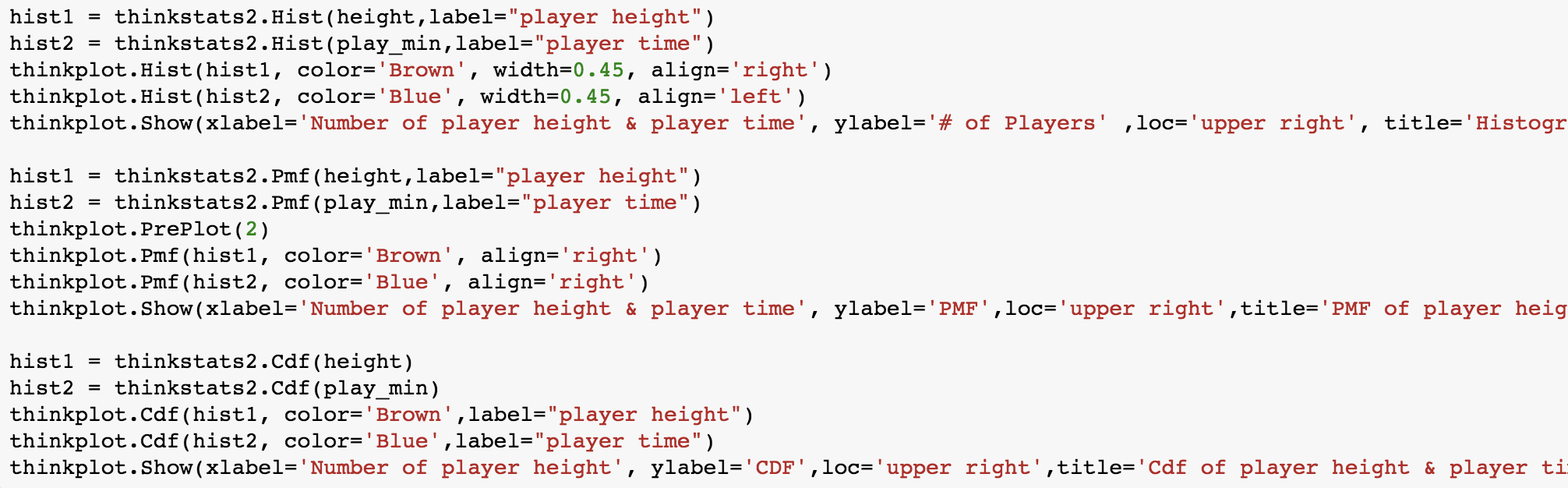


Show of player play time PMF

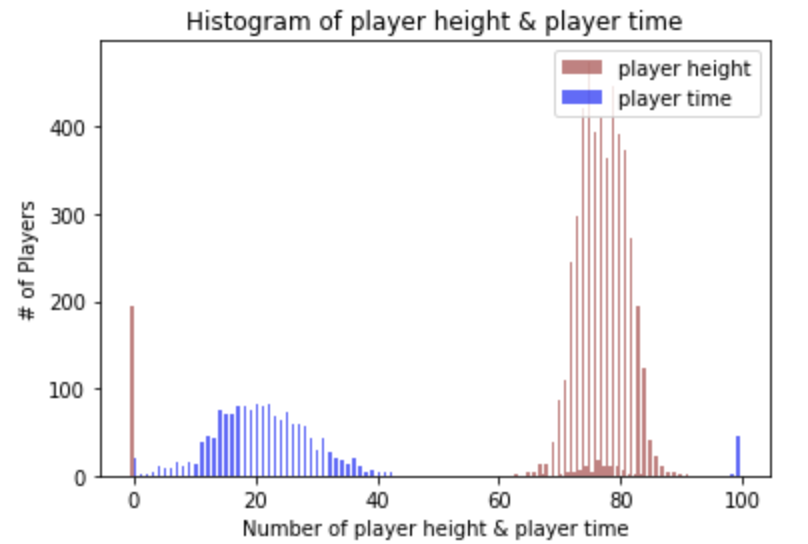


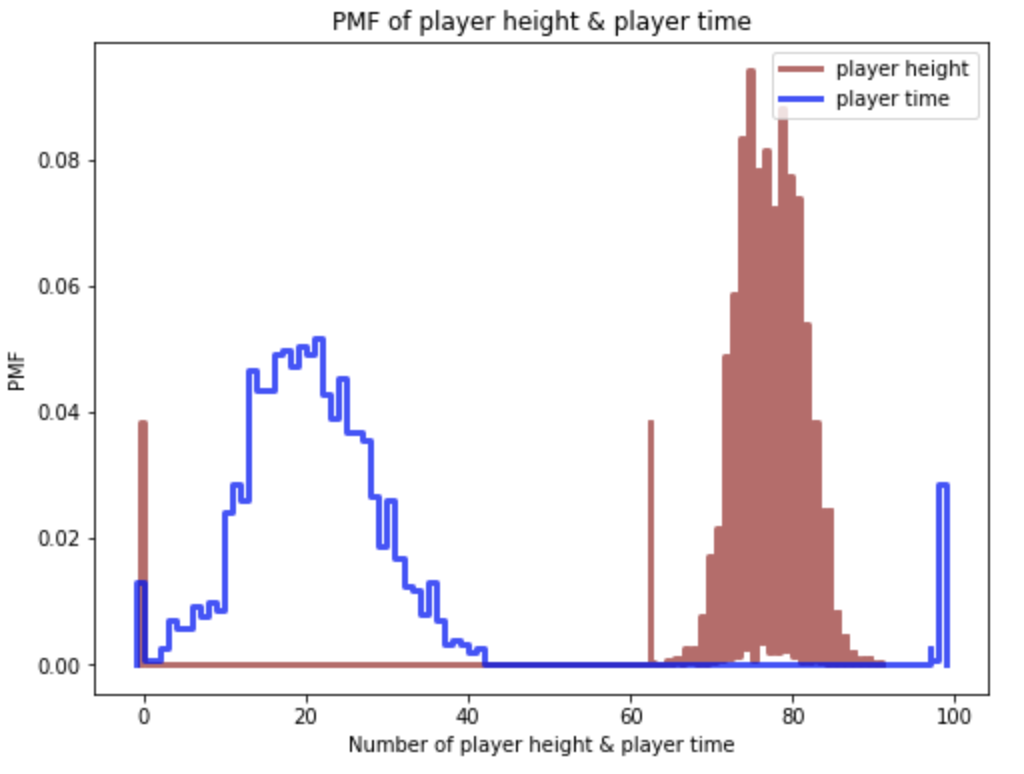
Show of player player play time CDF

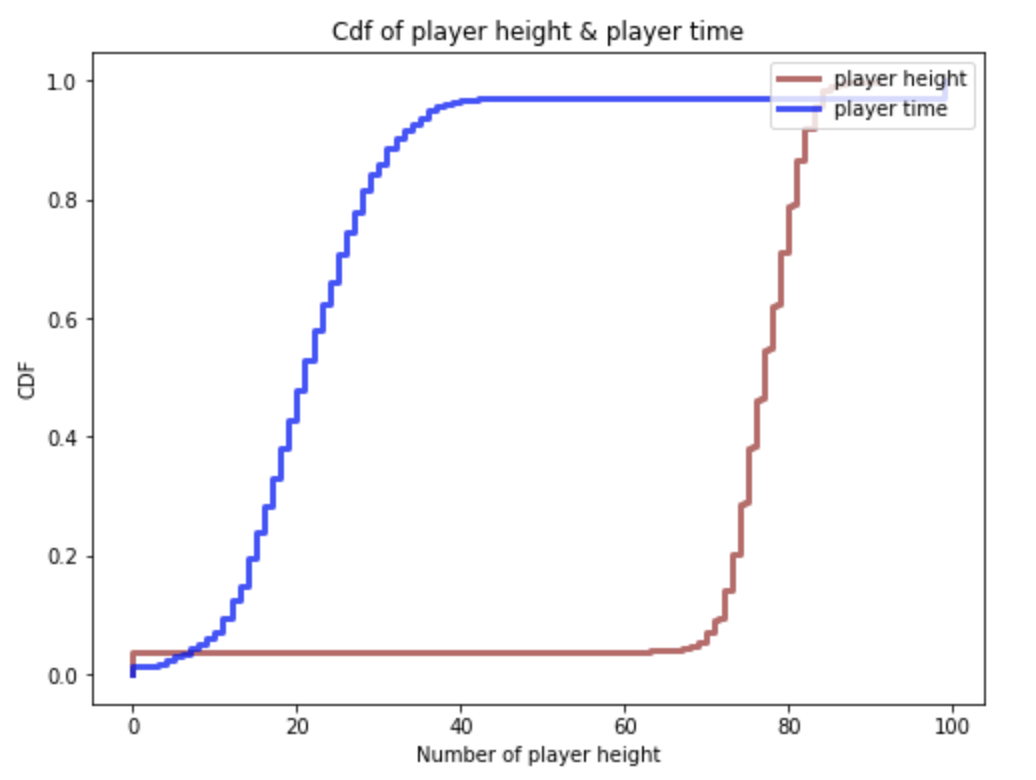




This code shows the how to combine to two different plots with the different values. Also there are histogram,pmf and cdf.ı will show the all plots.

Histogram of the player height and player play time in the match.

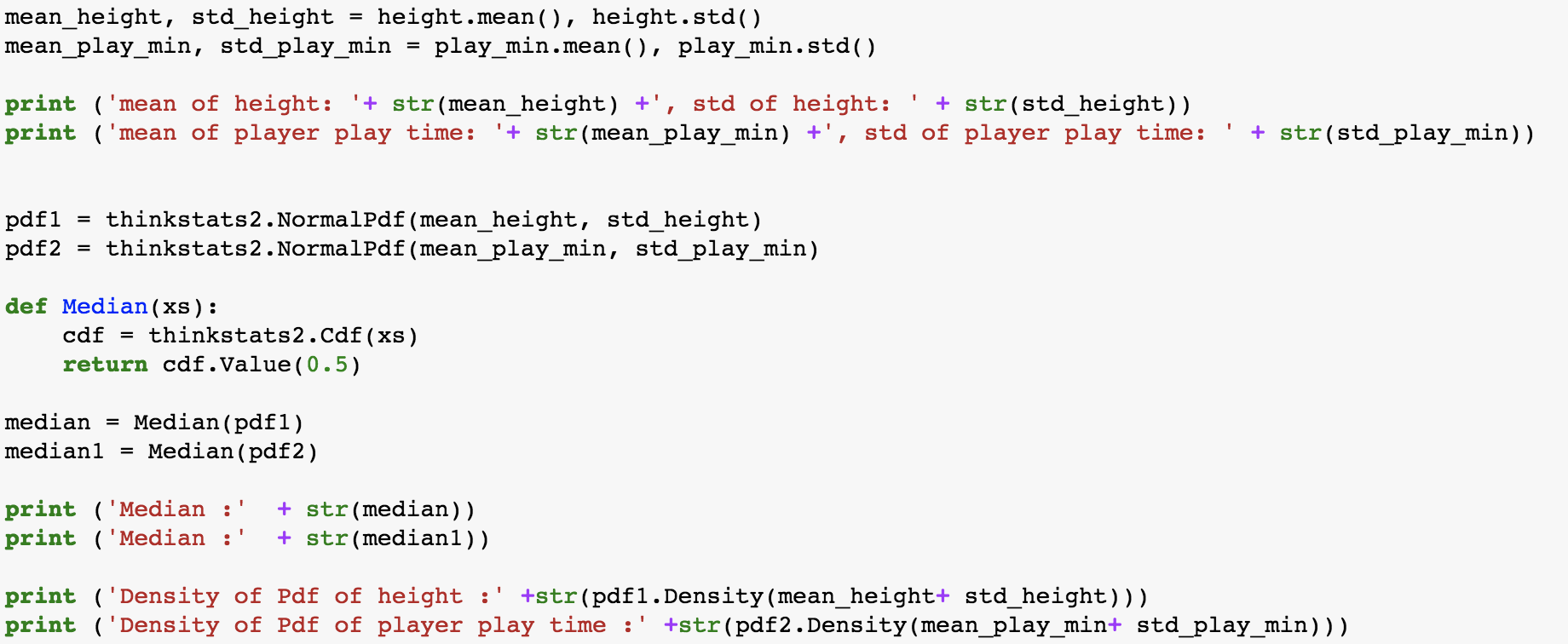
PMF of the player height and player play time in the match.

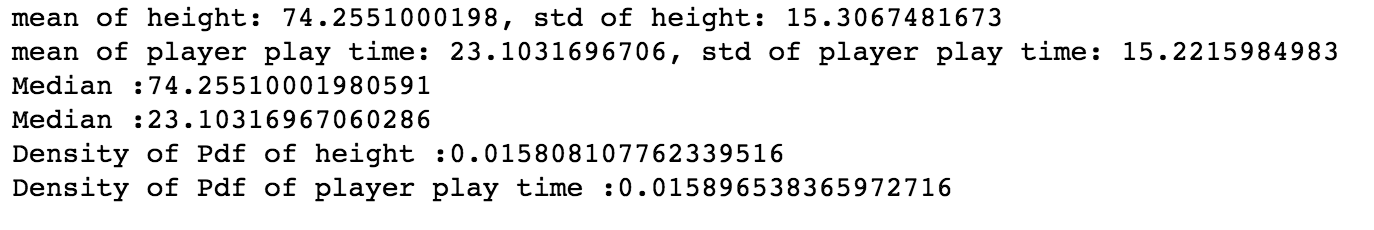
CDF of the player height and player play time in the match.

As you can see, this figure shows you the number of player frequency by number of the our players height and players playing time in the basketball match.There is a similarity between them.

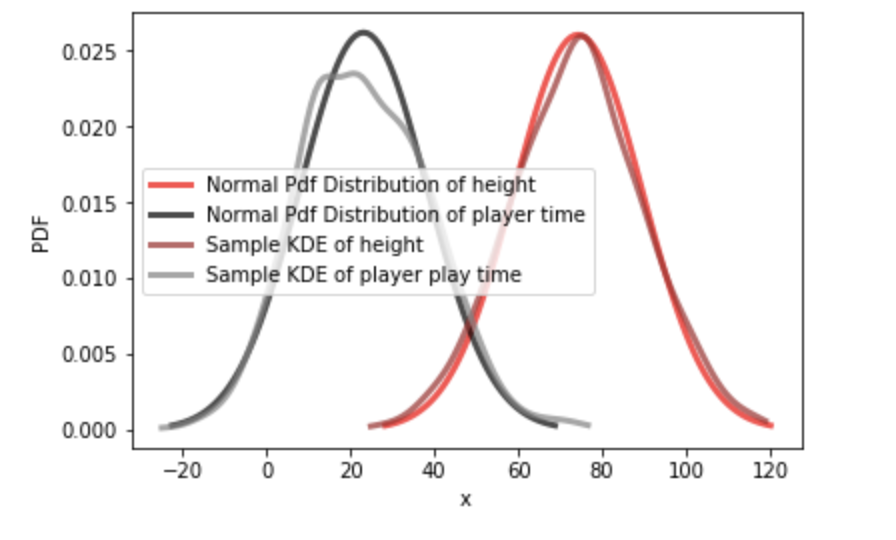
**Part 4:**

Here is the my probablity density function code. This code , first calculated the mean and standard deviation by using “.mean ()” and “.std ()” respectively because pdf distribution is drawn based on the mean and standard deviation.used “thinkstats2.NormalPdf ()” and inserted the “mean\_of\_height” and “std\_of\_player play time” variables as parameters and created a Normal Pdf variable called “pdf1”. Then I did the same thing for the runs and used “thinkstats2.NormalPdf ()” and inserted the “mean\_of\_height” and “std\_of\_player play time” variables as parameters and created a Normal Pdf variable called “pdf2”.

****

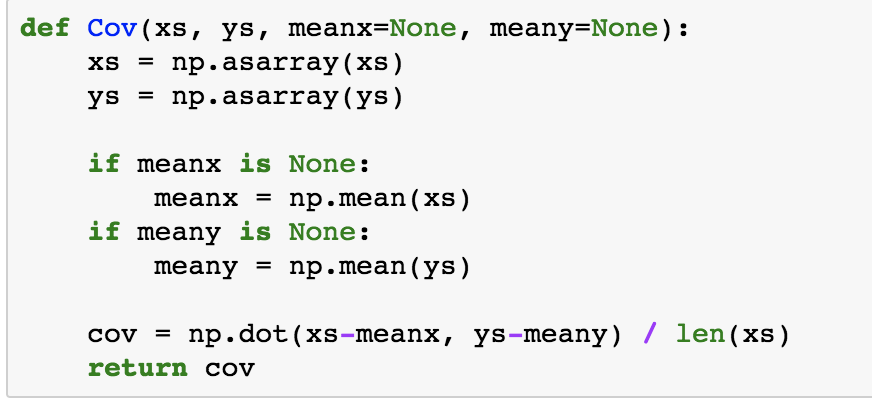


Here is the my result. With my result I plot the PDF plot.



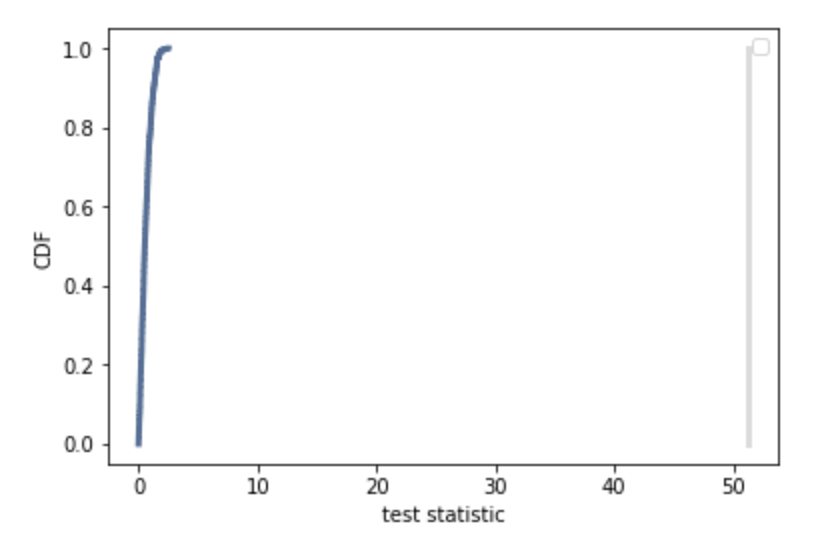
**Part 5:**

My question was Does the high of the basketball player affect the amount of time played in all star game ?Is there a positive relationship or negative relationship? In order to answer this question we will have to use the covariance test which will give us the tendency of two variables to vary together and correlation test which will give us the strength of the relationship between two relationships.



In this code computes deviations from the sample means, or you can provide known means. If “xs” and “ys” are Python sequences, “np.asarray” converts them to NumPy arrays. If they are already NumPy arrays, “np.asarray” does nothing. Pearson’s correlation is always between -1 and +1 (including both). If *r* is positive, we say that the correlation is positive, which means that when one variable is high, the other tends to be high. If *r* is negative, the correlation is negative, so when one variable is high, the other is low.

**Part 6:**



In my test statistic p-value is 0, so this means that it is statically significant.

**Part 7:**In a conclusion part in a second hypothesis was, “Does the high of the basketball player affect the amount of time played in all star game ?” We compare with the histogram,pmf,cdf. Combine the all plot with two value.When we check the covariance and correlation we can conclude that the similarity is pretty high. So It means there is a correlation between them.

REFERENCES

Thinkstat