CS-240 Final Project Report

Nurcan Tüylüoğlu -213951823

Part-1

• What is the relationship between different performance metrics? Do any have a strong negative or positive relationship?

• Could it be any strong negative or positive relationship between variables?

•What are the characteristics of variables in datasets?

Finally, you must create at least 3 question that you can analyze. Then select one of among those three questions, create your hypothesis, and explain why you choose that question for hypothesis. That is necessary for the testing at last part.

I choose the data of the basketball players(Master.csv) which includes information about the players like(name,ıd,birtdayplace,weight,heigt..).

Also,I choose the(Players.csv) data which contains the players Ids and each players threepoint rate throughout the seasons.In this study,we will try to aswer the following question,Is there a relation between height of players and three point accuracy? Firstly,I want to give informaiton about three point accuracy. It is called 3 point .Estimating because the team member provides their pessimistic,optimistic and best guess estimates for their deliverable.In addition, the 3 point estimating technique gives you better data because you're explicitly considering risks.I will calculate the correlation for each player which is calculated ;

$$r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (X_i - \bar{Y})^2}}$$

Correlation: The mathematical expression of the relationship between two variables.It is a measure giving information about the strength and direction of the relationship between variables. It does not tell us whether there is a causal relationship or not. The correlation coefficient is between -1 and +1. If the sign of the correlation coefficient is positive, the value of one of the variables increases (decreases) while the value of the other increases (decreases).If the sign of the correlation coefficient is negative, the value of one of the variables increases (decreases) while the value of one of the variables increases (increases). So there is an opposite relationship.

Calculation result = -0.1.

As we can see there is a negatif relationship.Therefore,when the players' height increase, the shooting rate is reduced.

Part-2

According to your hypothesis, show the variables and datasets that you are going to use then clean and organize your data to start analysis, interpret your results and explain what it means in a clear way.

One file,basketball_master.csv.contains player profiles.Only bioID and height columns are used in the analysis.

Other file has seasonal statistics for each player.threeMade and threeAttempted columns are used to calculate three point rate and player ID column is use to link datasets.

Data Cleaning and Processing

NaN values are dropped

Dataset has some abnormal values for height variable. According to the dataset, some players have height of zero. These values are regarded as NaN values because we do not expect basketball players to have 0 height. Such rows are removed.

Dataset does not have 3 point accuracy value but it contains 3 point attemps and 3 point scores. By using these two column, 3 point accuracy calculated as *accuracy=points/attempts*accuracy=points/attempts.

To get meaningful accuracies, players who do not have at least 3 attemps, are removed from analysis. This step solves divide by zero problem as well.

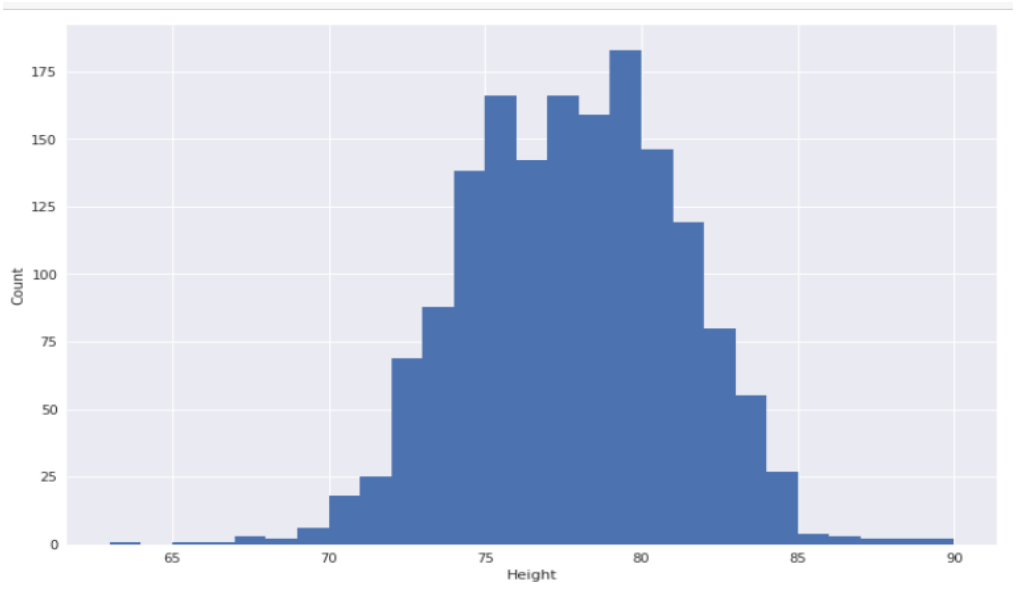
	playerID	year	stint	tmID	lgID	GP	GS	minutes	points	oRebounds	...	PostBlocks	PostTurnovers	PostPF	PostfgAttempted	PostfgMade
0	abramjo01	1946	1	PIT	NBA	47	0	0	527	0	...	0	0	0	0	0
1	aubucch01	1946	1	DTF	NBA	30	0	0	65	0	...	0	0	0	0	0
2	bakerno01	1946	1	CHS	NBA	4	0	0	0	0	...	0	0	0	0	0
3	baltihe01	1946	1	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

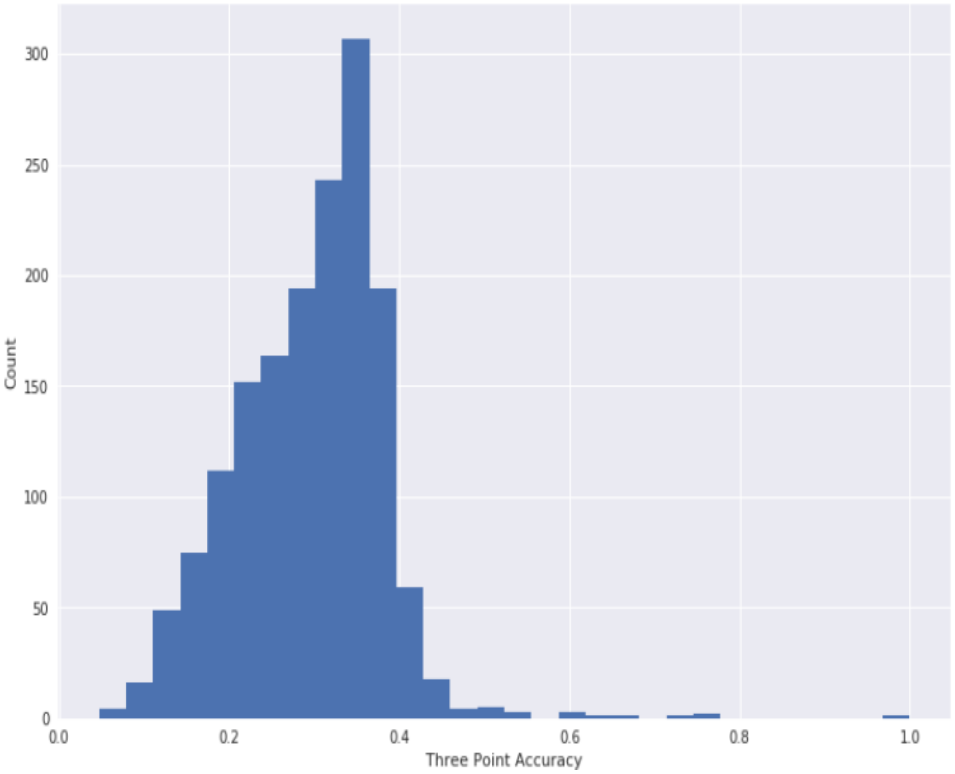
ds	...	PostBlocks	PostTurnovers	PostPF	PostfgAttempted	PostfgMade	PostftAttempted	PostftMade	PostthreeAttempted	PostthreeMade	note
	...	0	0	0	0	0	0	0	0	0	NaN
	...	0	0	0	0	0	0	0	0	0	NaN
	...	0	0	0	0	0	0	0	0	0	NaN
	...	0	0	3	10	2	1	0	0	0	NaN
	...	0	0	0	0	0	0	0	0	0	NaN

Part-3

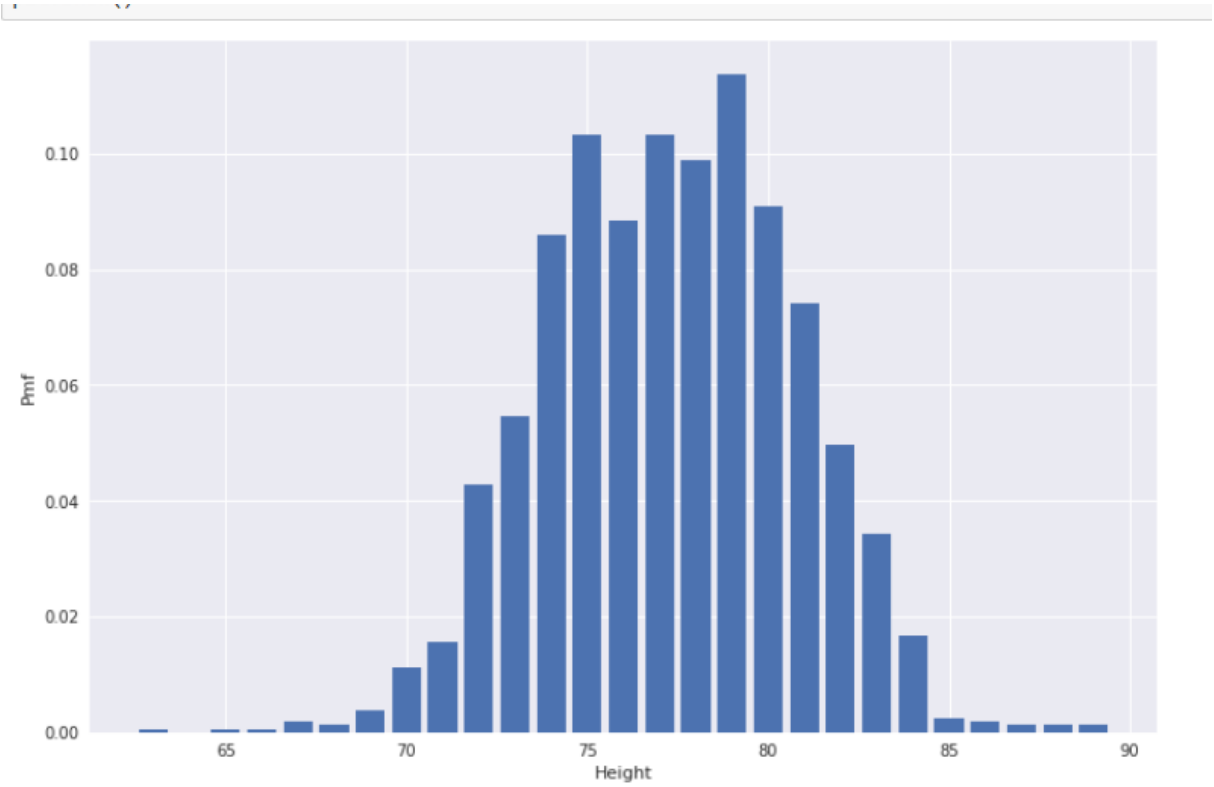
I computed the players' height and add it to the players threepoint accuracy.

Here is the histogram chart for players' height and three point accuracy

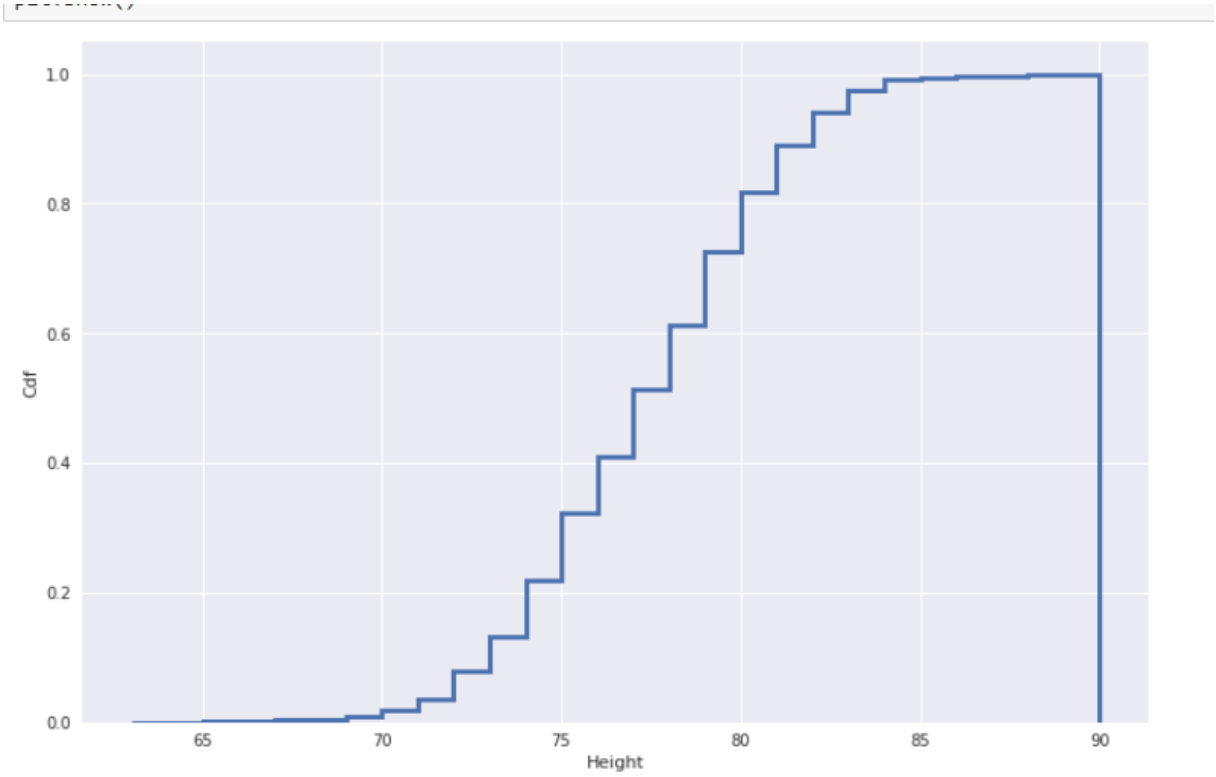


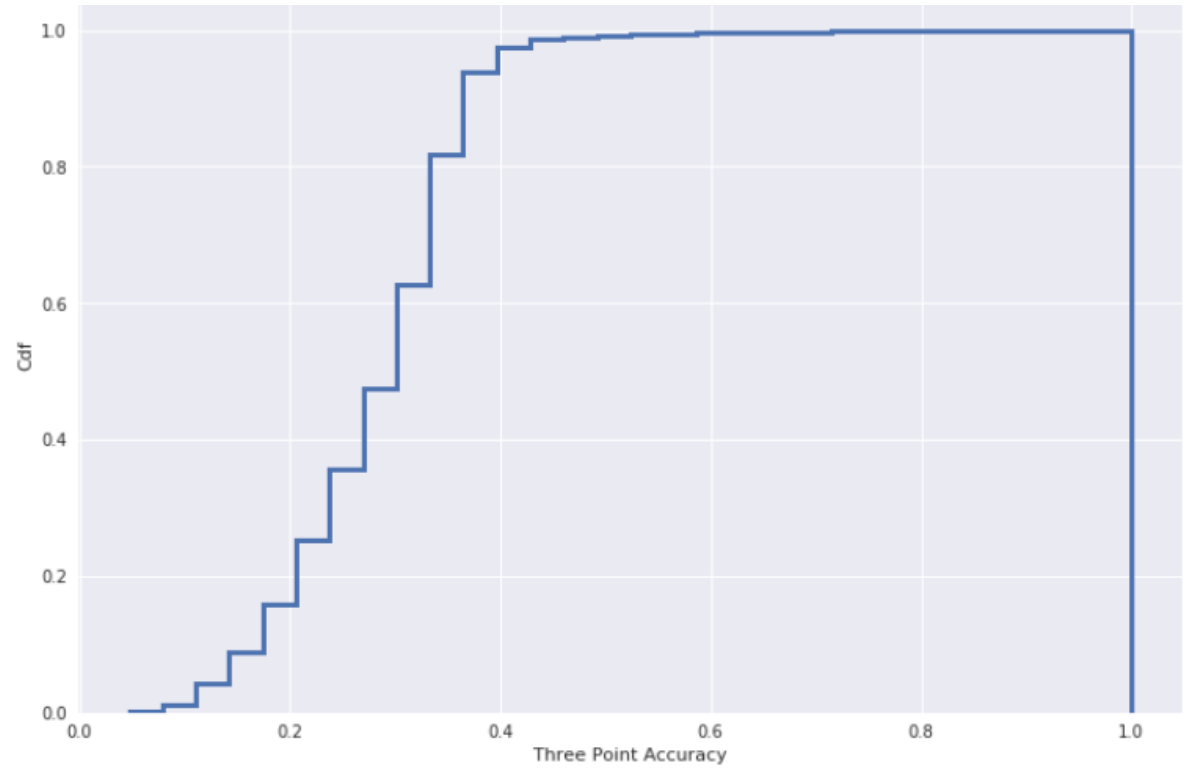


Here is the Probability Mass Function(PMF



And here is the Cumulative Mass Function(CDF)

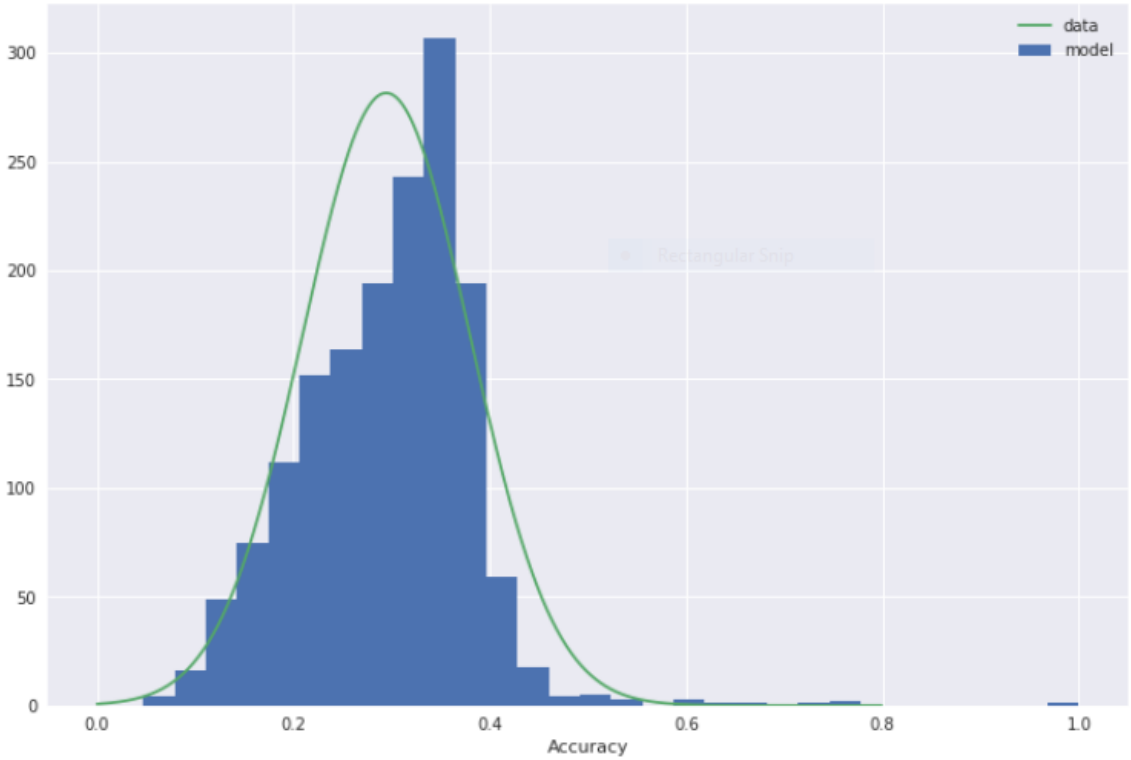


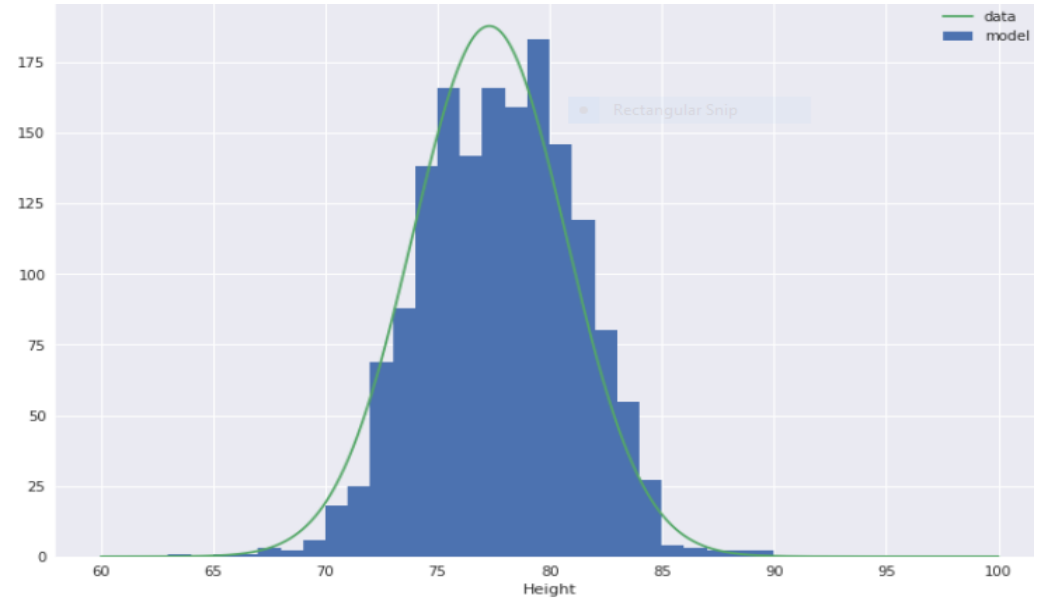


Part 4

According to your hypothesis, use one of modelling distributions, model your data and interpret your results (how your model fits to your data etc.) and explain what it means in a clear way).

Histogram of both height and three point accuracy hints that normal distribution can fits the data well.


data
model
300
250
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0
Accuracy


data
model
175
150
125
100
75
50
25
0
60
65
70
75
80
85
90
95
100
Height

What is normal distribution?

A normal distribution of data is one in which the majority of data points are relatively similar, occurring within a small range of values, while there are fewer outliers on the higher and lower ends of the range of data.

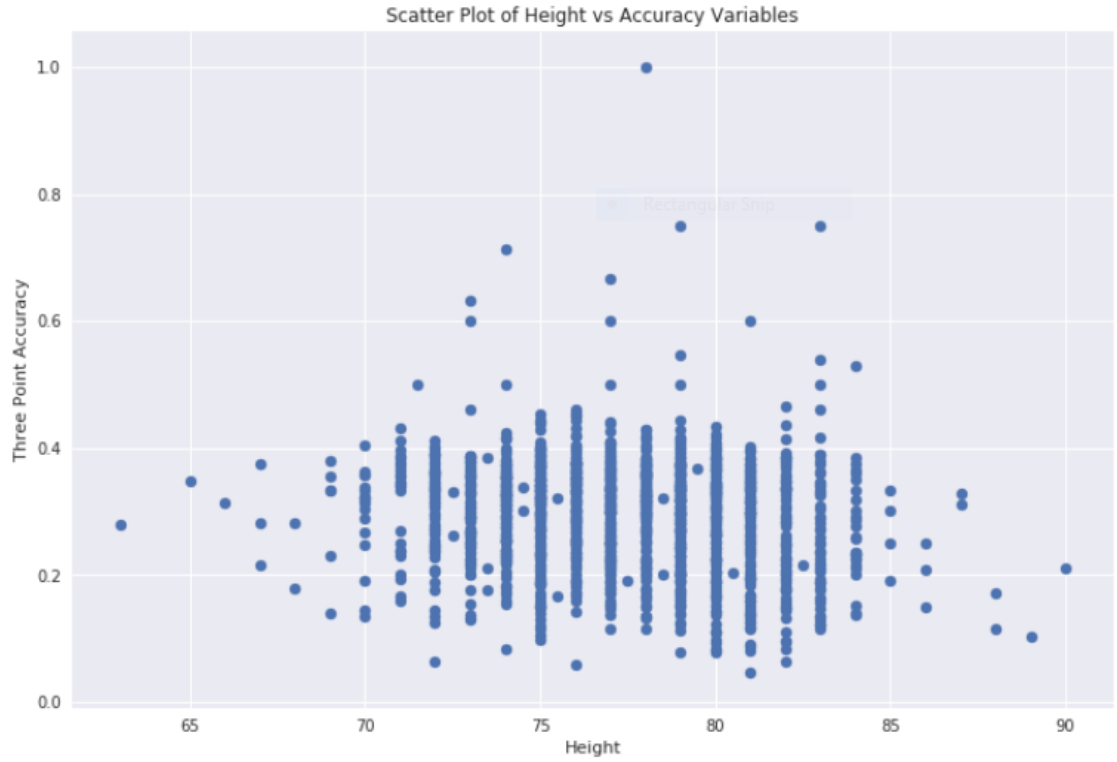
As we can see histogram, most of the players' heights are in the range of 70-85.

Part-5

Built one relationship according to your hypothesis and choose 2 variables in your data explain and show their correlation then visualize this correlation. Also, interpret your results and explain what it means in a clear way.

I planned to show the correlation between 2 variables which are players' height values and three point accuracy and I used the scatter plot in order to visualize this correlation. Scatter plot does not say much about relation between height of the player and hist three point accuracy. Therefore, Correlation is calculated as -0.13. This results implies that there might be a weak negative correlation between variables. In other words, accuracy of three points decreases as height of the player increases.statistical tests should be applied (Section 6) to show that it is not just some spurious correlation but an actual relation.

Below we can see the graph




```
Correlation between height and accuracy : -0.13154889056643443
```

Part-6

Test your hypothesis step by step, show your steps and explain why you need that step and what needed for that step. Interpret your results (according to the p-value) at the end and explain what it means in a clear way .

First of all ,I want to give information that The P-value is the probability that you would have found the current result if the correlation coefficient were in fact zero (null hypothesis). If this probability is lower than the conventional 5% (P<0.05) the correlation coefficient is called statistically significant.

It is, however, important not to confuse correlation with causation. When two variables are correlated, there may or may not be a causative connection, and this connection may moreover be indirect. Correlation can only be interpreted in terms of causation if the variables under investigation provide a logical (biological) basis for such interpretation.95% confidence interval (CI) for the correlation coefficient: this is the range of values that contains with a 95% confidence the 'true' correlation coefficient.As we can see,

Null hypothesis : There is no statistically significant relationship between height variable and three point accuracy variable.Since 0.05 threshold value is commonly used, we also use this value to test significance of this relation.After applying test, p value is calculated as 1.197*x*10−71.197x10−7 .This value is much smaller than 0.05 therefore we can say that we reject null hypothesis, so there is a relationship between height and three point accuracy.

Part-7

Write a conclusion that describes your analysis and what you get end of the analysis such as the result of hypothesis testing and some important findings out of your whole analysis.

We tried to find statistically relation between height and 3 point accuracy of NBA basketball players. Some analysis on data show that these values are distributed normally. Scatter plot and correlation coefficient show that there might be a weak relation between these variables. Correlation value is calculated as -0.13.

To show that this correlation is not some random occurence, null hypothesis test is applied. Null hypothesis is contructed as "There is no statistically significant relationship between height of player and three point accuracy, assuming that the player has made at least 3 attemps". Low p value we calculated in section 6 proof that it is unlikely to observe such relation in two uncorrelated variable. In other words, our sample provides enough evidence to reject null hypothesis. So, there is a statistically significant relation between two variable although this relation is not strong.