CS 240 Final Project Report

Amer Nour Eddin | 213171245

PART 1

Brainstorm some questions you could answer using the data set you chose, then start answering those questions. Here are some ideas to get you started:

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

• What are the characteristics of baseball players with the highest salaries?

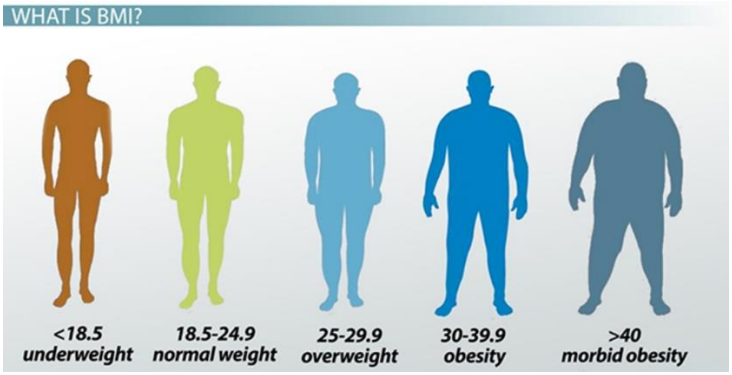
Finally, you must state one question that you want to analyze. Then give your hypothesis for that question to test last part.

I choose the data of the baseball players (Master.csv) which contains a general information about all the players like the (name, Id, birthdate place, weight, height etc.).

Also I choose the (Salaries.csv) data which contains the players Ids and each player salary throughout the years.

My intention is the following:

I will calculate the BMI (Body Mass Index) for each player, which is calculated using the player's weight and height → *BMI = weight (lb.) / (height (in))^2 x 703*



My question is there any relation between the player's BMI and his average salary throughout the years?

So my specific hypothesis is:

"If the player's BMI value is HIGHER than 26 then his salary value will be larger than other players"

PART 2

Show the columns that you are going to use then clean and organize your data to start analysis show your codes and explain it what it does.

Since one of the tow variables that I'm depending on is the BMI values of the players so I needed to use the (Master.csv) file which contains a detailed amount of information about each player including the player weight and height which I used for computing the BMI value, you can see the Filtered Players data frame which I cut from the original file.

And the second variable that I used was the salary for each player so I used the (Salaries.csv) file which contains the salaries for all the players, the salary data frame that I used is showed below also.

```
In [5]: FilteredPlayers
```

Out[5]:

	playerID	nameFirst	nameLast	weight	height
0	aardsda01	David	Aardsma	215.0	75.0
1	aaronha01	Hank	Aaron	180.0	72.0
2	aaronto01	Tommie	Aaron	190.0	75.0
3	aasedo01	Don	Aase	190.0	75.0
4	abadan01	Andy	Abad	184.0	73.0
5	abadfe01	Fernando	Abad	220.0	73.0
6	abadijo01	John	Abadie	192.0	72.0
7	abbated01	Ed	Abbaticchio	170.0	71.0
8	abbeybe01	Bert	Abbey	175.0	71.0
9	abbeych01	Charlie	Abbey	169.0	68.0

```
In [456]: FilteredSalaries
```

Out[456]:

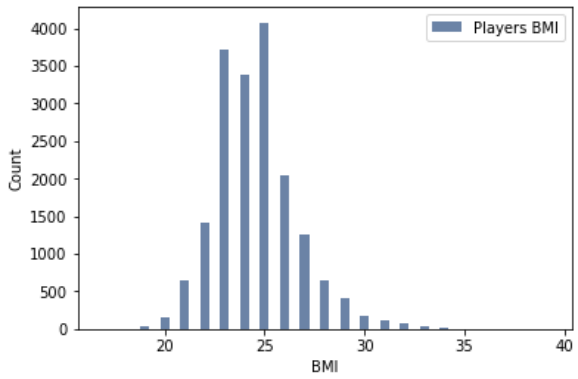
	yearID	playerID	salary
0	1985	barkele01	870000
1	1985	bedrost01	550000
2	1985	benedbr01	545000
3	1985	campri01	633333
4	1985	ceronri01	625000
5	1985	chambch01	800000
6	1985	dedmoje01	150000
7	1985	forstte01	483333
8	1985	garbege01	772000
9	1985	harpete01	250000

PART 3

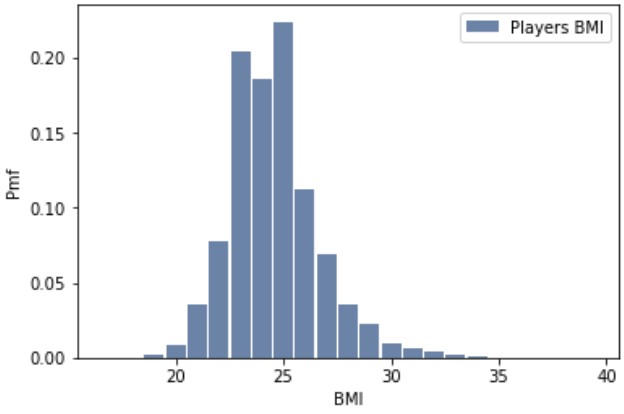
According to your question first show some relevant statistics then plot; 1 Histogram, 1 PMF and 1 CDF show your codes and explain it what it does.

I computed the BMI values and add it to the players data frame

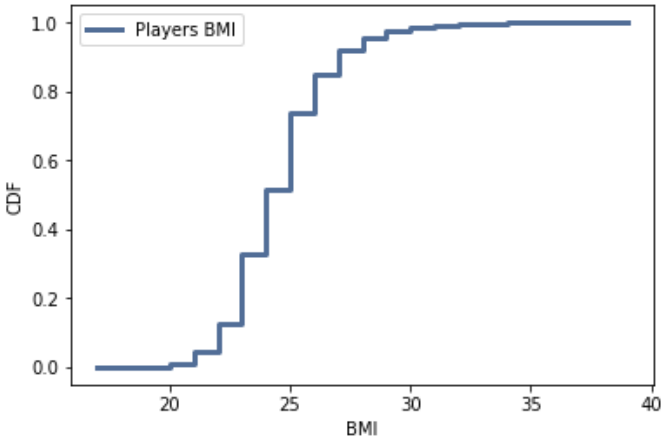
Here is the histogram chart for the BMI values distribution for all the players.



And here is the Probability Mass Function (PMF)



And here is Cumulative Distribution Function (CDF)



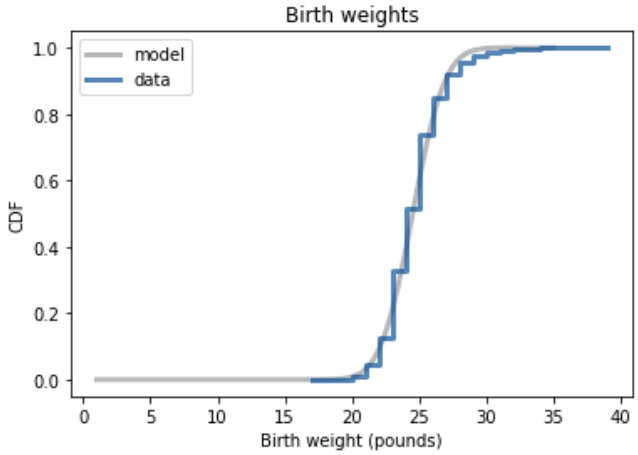
PART 4

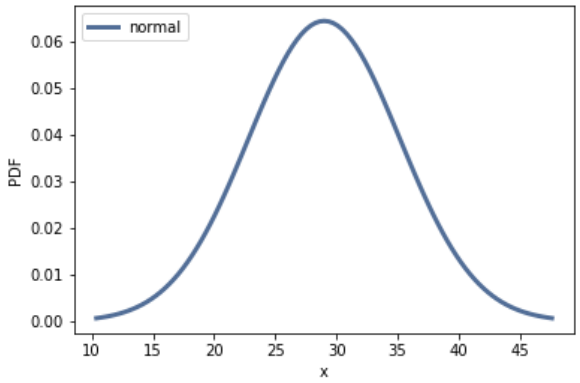
Use one of modelling distributions, model your data and explain it why you used that and what it does explain with showing also your codes

I used the Normal distribution for modeling as I saw that it is the most that fits the data shape.

As we can see here's the observed CDF and the model. The model nearly fits the data except in the left tail.

I also plotted the normal PDF below.



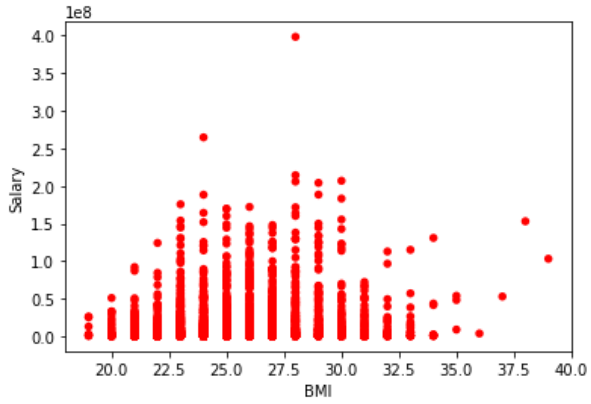


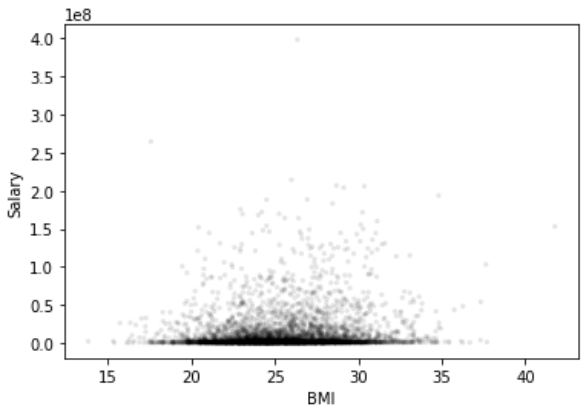
PART 5

Built one relationship according to your question and choose 2 variables in your data explain and show their correlation then visualize this correlation. Also, show your codes and explain what it does.

I planned to show the correlation between 2 variables which are the BMI values and the salaries of the players and I used the scatter plot in order to visualize this correlation.

Below we can see the each BMI value and its corresponding salary on the y axis, surly there are some outliers but we can see a normal distrbution shape for these tow variables.





PART 6

Test your hypothesis step by step, show your codes, and explain what it does.

Here I tested my hypothesis that “the players with BMI above 26 have higher salary than others”

Using the hypothesis testing methods provided by thinkstats2.

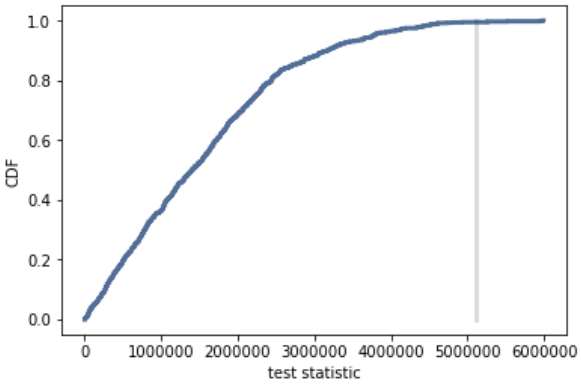
My test statistic was taking the average of the 2 groups (salaries with BMI above 26 and below 26) and taking the absolute difference

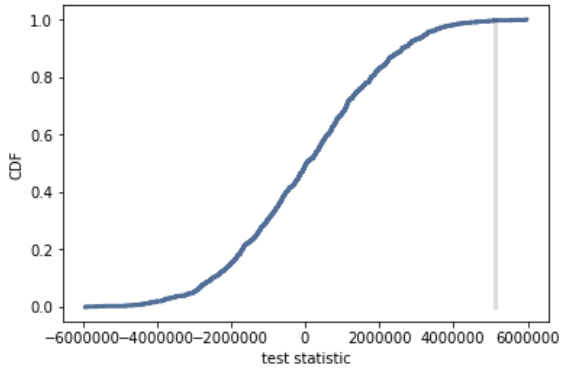
The results was as the following:

For null hypothesis my p value was 0.005 which is not statistically significant.

But also the p value for my hypothesis was not statically significant too!

That made me draw my final conclusion which you can see it in the last part





PART 7

Write a conclusion that describe your analysis and what you get end of the analysis.

My final conclusion was:

There is no real relation or cause and effect in my proposed hypothesis, so I couldn't prove that players with BMI higher than 26 will have higher salary.

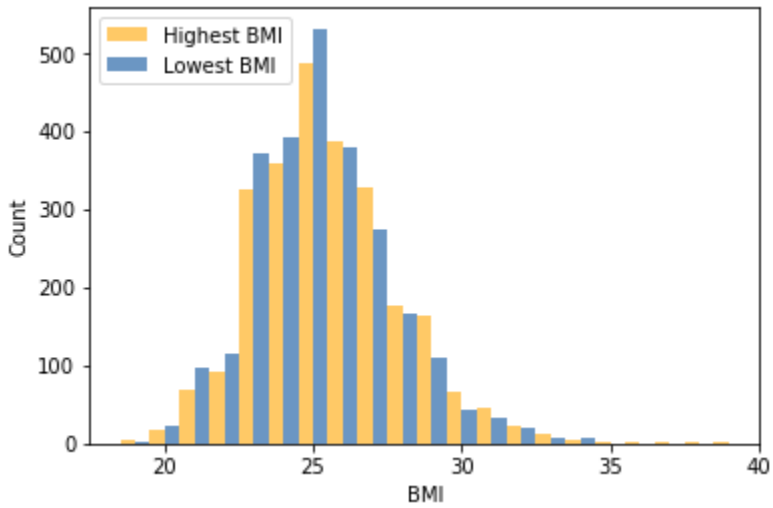
And I've done the following to show that:

And I affirmed that the hypothesis that I made is not true by the following:

I sorted all the salaries by ascending order and cut them in halves (the x highest and the x lowest), and I looked at the BMI values of the highest salaries and plotted their histogram, similarly I looked at the BMI values for the lowest salaries and plotted their histogram.

And I did that to see IF THE HIGHEST SALARIES HAS A SPECIAL DISTIRBUTION FOR THE BMI VALUES.

Here is the resulting plot:



And as we can see from the previous plot that there is no significant difference between the BMI values for the highest and the lowest salaries.

So there is no relation between BMI and salary that may cause the salary to increase.