Hassan Zind El Hadid

215460795

CS 240: Project Report

- ## Section 1:

After looking at the datasets, I came up with 3 questions that can be answered after some analysis. These 3 questions are:

1. Do the players experience a decline in their general performance as they age?

2. Do the weights and heights of players have an effect on their shooting abilities?

3. Are the height of a player and his average number of blocks per game positively correlated? In other words, do taller players have better average number of blocks per game than shorter ones?

I choose the third question to work on in this project.

My hypothesis: I think that taller players have better average number of blocks per game than shorter ones in general. So height and average number of blocks per game are positively correlated.

- Section 2:

According to my hypothesis, I need to access the “height” variable in the “basketball_master” dataset and the “blocks” and “GP” variables in the “basketball_players” dataset.

The data in “basketball_players” is collected on a large span of years. But the variable “blocks” is recorded consistently starting from the year 1973. So I decided to consider the data starting from the year 1973. I also decided to consider the data where “blocks” variable is more than zero. I then cleaned this dataset from the missing values.

Then, I added up the total number of blocks made through the years for each player. I also summed up the total number of games played “GP” over the years for each player. I was left with a cleaned dataset, called “df”, comprised of 2579 players. I calculated the average number of blocks per game by dividing the total number of blocks by the total number of games.

After that, I used the “playerID” variable to get the heights of these 2579 players from the “basketball_master” dataset. Then I took those heights and added them to the “Height” column in my data frame “df”. By doing so, I managed to create a data frame of cleaned data showing the players, their total number of game played, their total blocks, the average number of blocks per game, and their heights.

- Section 3:

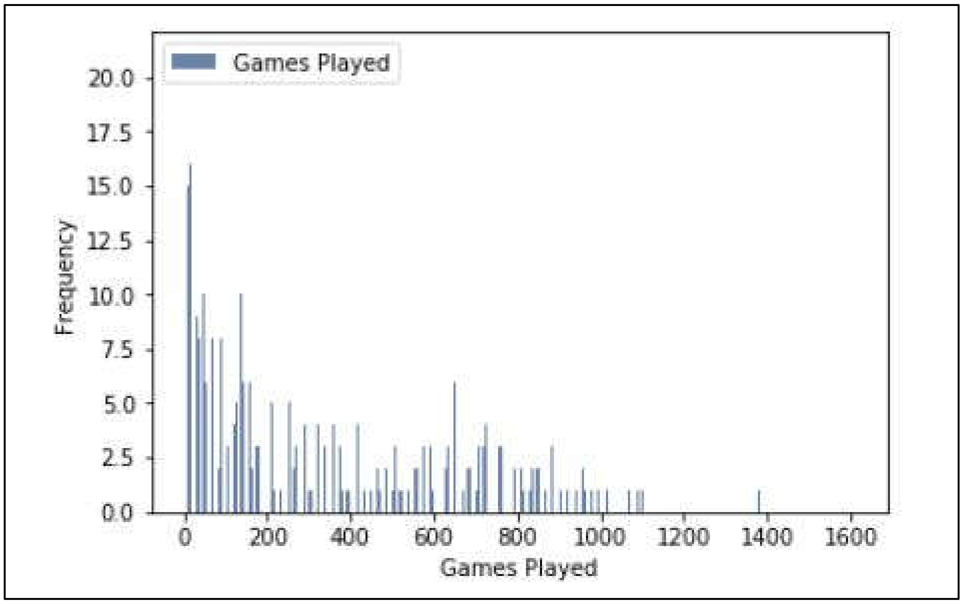
For the variable “Blocks per game”:

- Mean = 0.372
- Standard deviation = 0.403
- Max = 3.502
- Min = 0.012

For the variable “Height”:

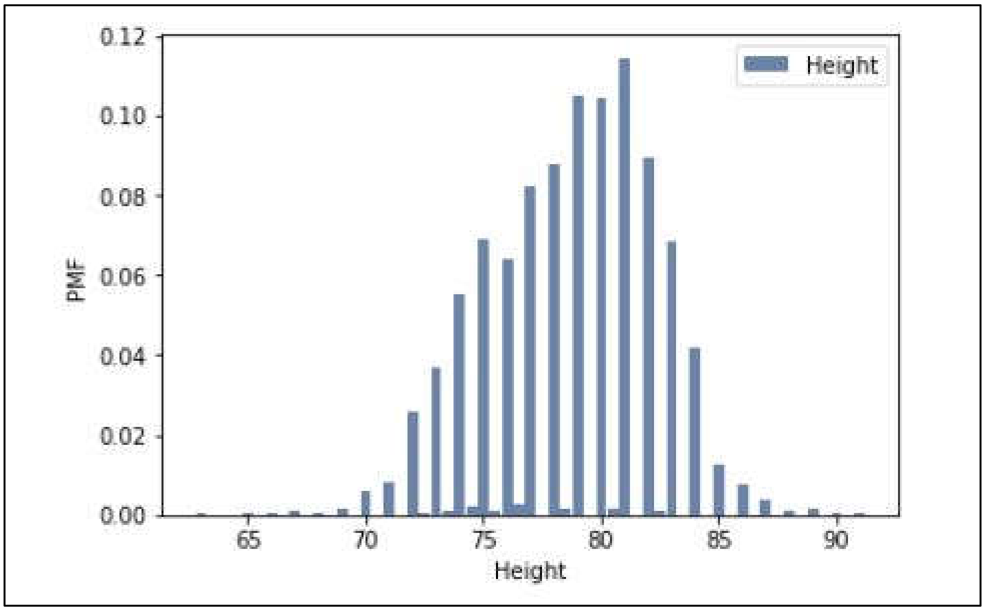
- Mean = 78.69
- Standard deviation = 3.55

Histogram of games played:



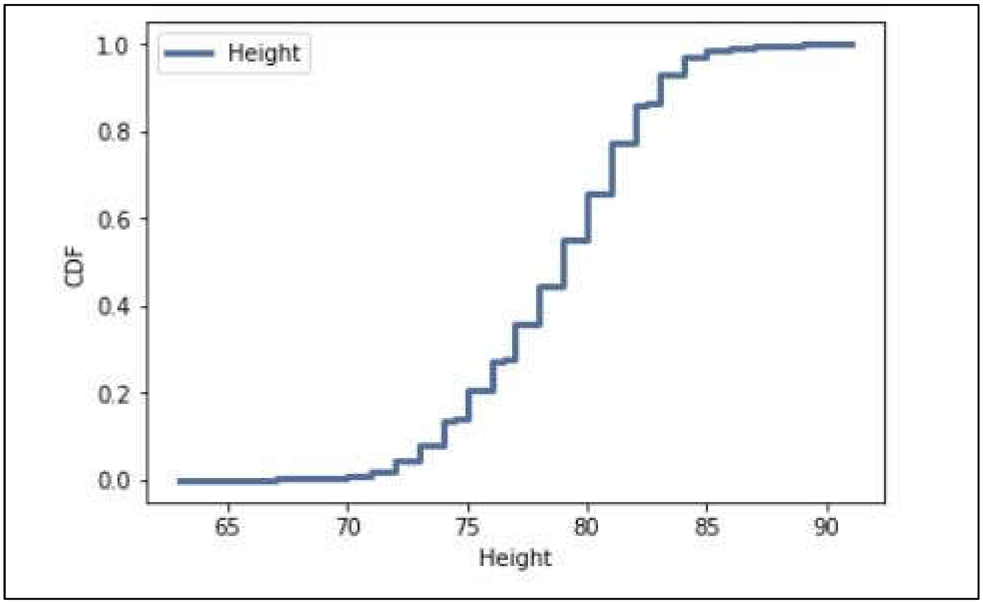
From this histogram, it is evident that not many players have a high total number of games played. The frequency of players with less than 200 games played is also really high. The frequency generally drops as the number of games played increases.

PMF of heights:



Here, a PMF of the heights of players is constructed. It shows a distribution of heights among the players.

CDF of heights:



A construction of a CDF for the heights. It shows how many players are in the ranges of height.

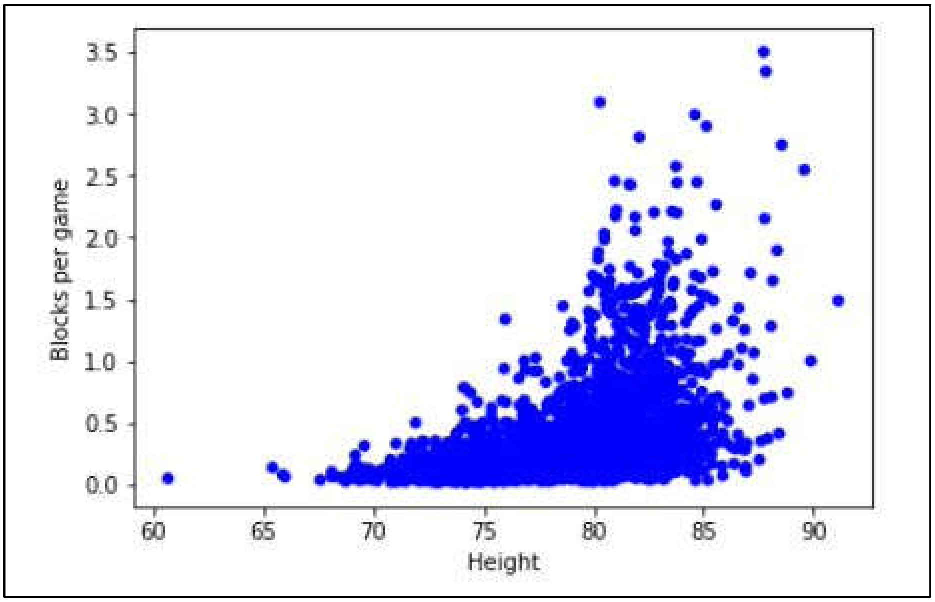
- Section 4:

- Section 5:

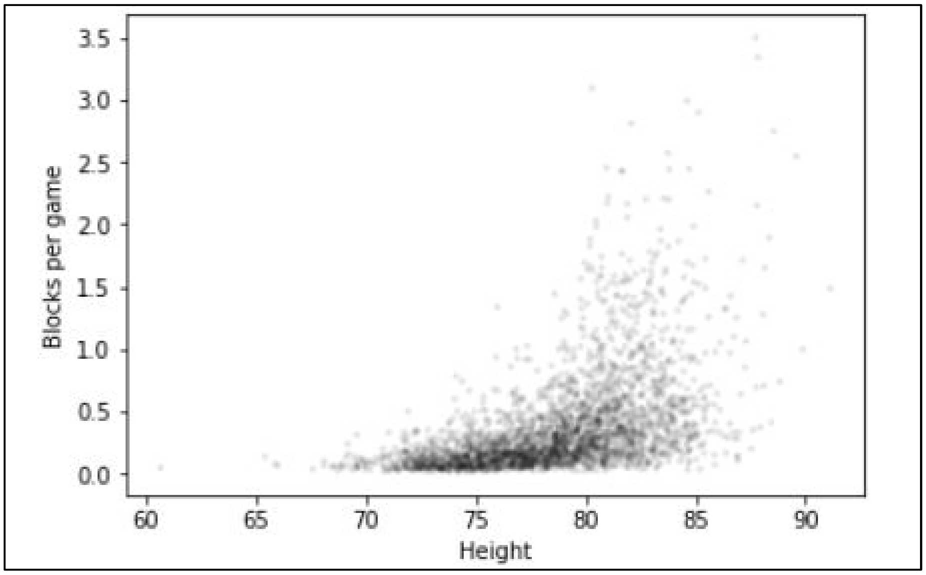
Calculations of the Pearson correlation and the Spearman correlation were made between the two variables, “Height” and “Blocks per game”. The results are:

- Pearson correlation = 0.527
- Spearman correlation = 0.627

Visualization of the correlation between “Height” and “Blocks per game” using scatter plot:



Here's another scatter plot of the same variables but with changing alpha and s in order to show the distribution of data points and the relation in a better way.



From the results and visualizations, we can conclude that there is indeed a positive correlation between heights and the average number of blocks per game.

- Section 6:

Alternative Hypothesis:

H_a: There is a positive correlation between the height of players and their average number of blocks per game.

Null Hypothesis:

H_o: There is no correlation between the height of players and their average number of blocks per game.

To to perform the hypothesis test, I used the HypothesisTest class in thinkstats2. A permutation test is run. The test statistic used is the correlation coefficient between “Height” and “Blocks per game”. After running the model for 1000 iterations, a p-value of 0.0 is calculated. Which means that in 1000 trials no correlation under the null hypothesis exceeded the observed correlation. The p-value is not exactly zero, but is less than 1/1000.

Since the p-value found is less than 0.05, the null hypothesis is rejected. Thus, the alternative hypothesis is accepted.

- Section 7:

As seen from the analysis, a positive correlation was indeed found by calculating the spearman and pearson correlation coefficiants. And after doing the hypothesis testing, this was proven to be true. However, since many players didn’t have a high number of games played, the results might not be very accurate.