Eren Ali Aslangiray Spring 2018

213911362 CS240

**Exploratory Data Analysis Term Project**

**Basketball Data Analysis**

From given database I can clearly ask and get answer the following three question. Which are:

* Does basketball player’s height and their average rebound statistics correlated? If yes, how does it correlated(Positive or Negative)? If you want your players to get more rebound is it wise to get taller people in your team?
* Does being awarded by coach make the team champion at that year and with that coach? How does it correlated? Does it wise to hire a coach with higher probablity to get an award?
* Does player’s origin country affects the player’s career average of points. Which one has better average; USA originated players or rest of the world’s players?

I choose Question 1 and for better understanding and building hypothesis first we need to understand few things, such as; What is rebound?, What value counts as good rebound number for a player? In what percentage of correlation is enough to satisfy basketball standarts for rebound?

In Wikipedia rebound is defined as “Rebound is a statistic awarded to a player who retrieves the ball after a missed field goal or free throw.” So for player to take a rebound he or she needs to be tall than other teams player who contests for that rebound.

And from Foxsports (one of the best data sites for NBA) 5 average rebound per game is accepted good for a player. And the players above 5 rebounds per game are spreaded among all the teams in NBA.

In basketball standarts over %5 change is counted as big in terms of getting rebounds. And for a team’s matches in a season %5 change of rebound rate will affect it performance greatly.

**Building the hypothesis:**

After greater understanding of the topic I can build the null hypothesis. The hypothesis is “Taller basketball players must have better rebound rate. If Average rebound per game statistic is greater than 4, the height data of players must above average height. And If correlation between these two datas are more than %5 a team can get taller players and increase its performance.

The opposite and alternative hypothesis is height is not effective for rebound rate, shorter than average heighted players can achieve average statistic 4. And the correlation of this two variables are less than %5.

**The datasets and variables that will be used during this analysis:**

For required hypothesis validation, I need to find average rebound per game rate for each player that recorded in the database. I also need height of all the players that have rebound statistics.

I will use following 2 datasets for my analysis. basketball\_master.csv (For height values) and basketball\_players.csv (for finding average rebound rate for each player).

From these two datasets I will get several variables which are: PlayerID(stands for player name and its unique), Height, GP(stands for “Games Played” and records that how many matches does a player played during a year. Which we will need to find average rebound rate for each player.), Rebounds(rebound numbers of each player. Yearly statistic.)

So finding the average rebound rate first I will sum up all the games played and rebounds taken for each player and divide total rebound number with total games played. After that I match each players height and their average rebound number.

**5 descriptive statistics for each variable that will be used:**

Average Height : 198.3 cm

Median of Height : 198.1 cm

Variance of Heigth : 81.2 cm

Standart deviation of Height : 9

Mode of Height : 200.66 cm

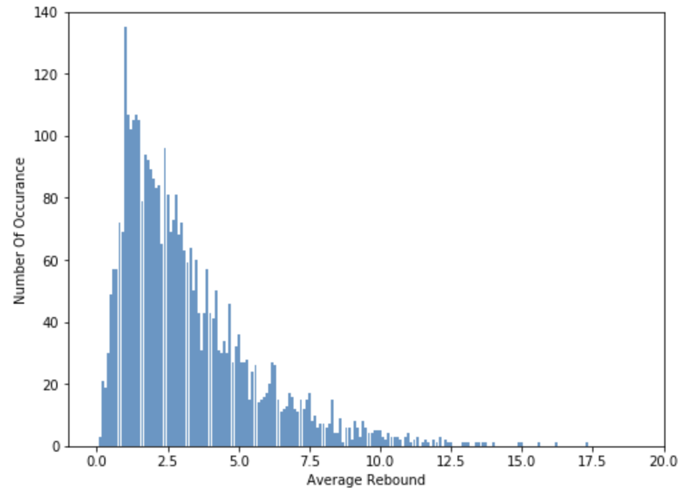
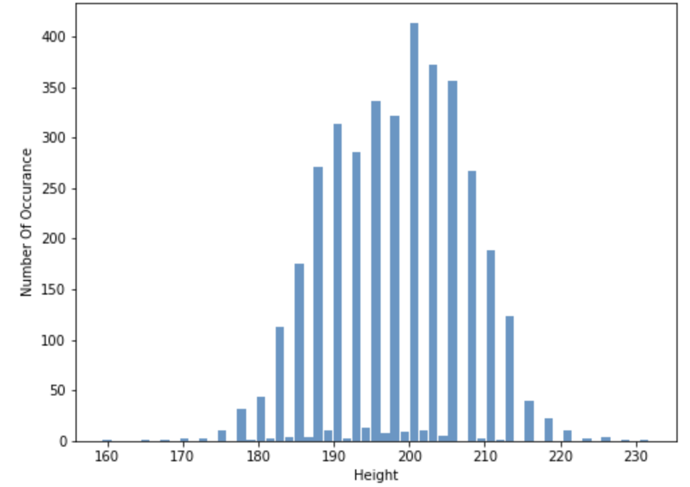
Average of Average Rebound Number : 3.1

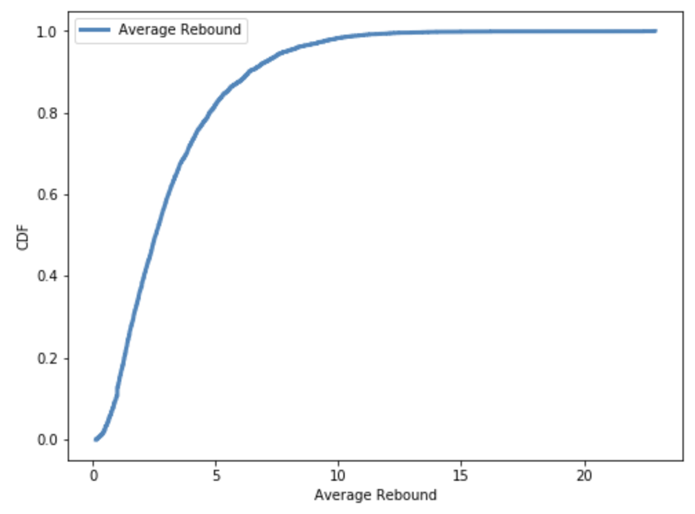
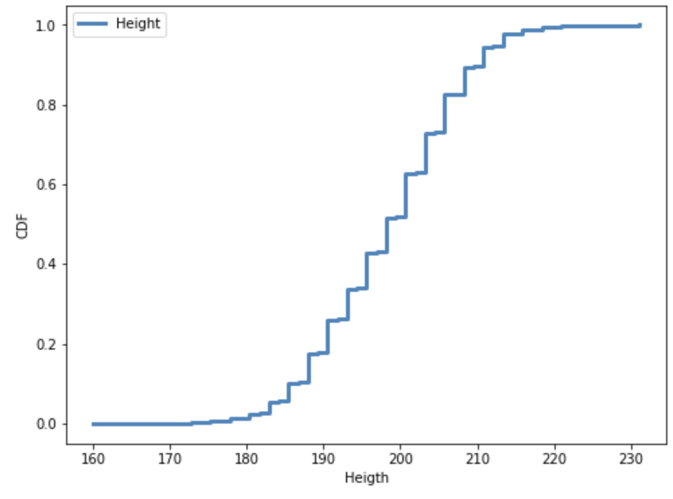
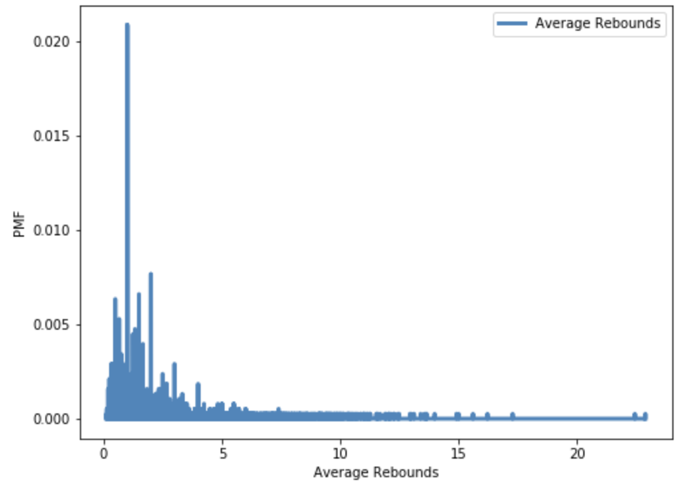
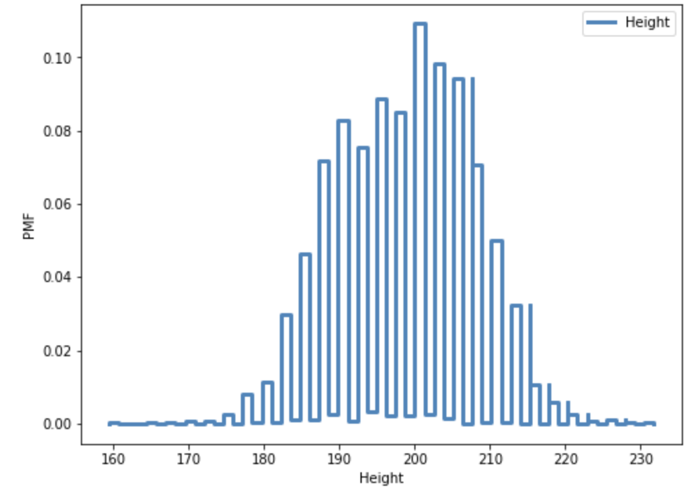
Median of Average Rebound Number : 2.5

Variance of Average Rebound Number : 5.6

Standart deviation of Average Rebound Number : 2.3

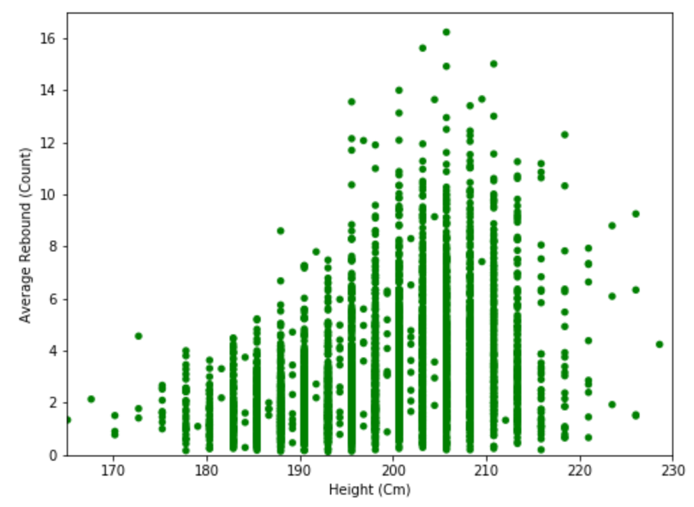
Mode of Average Rebound Number : 1



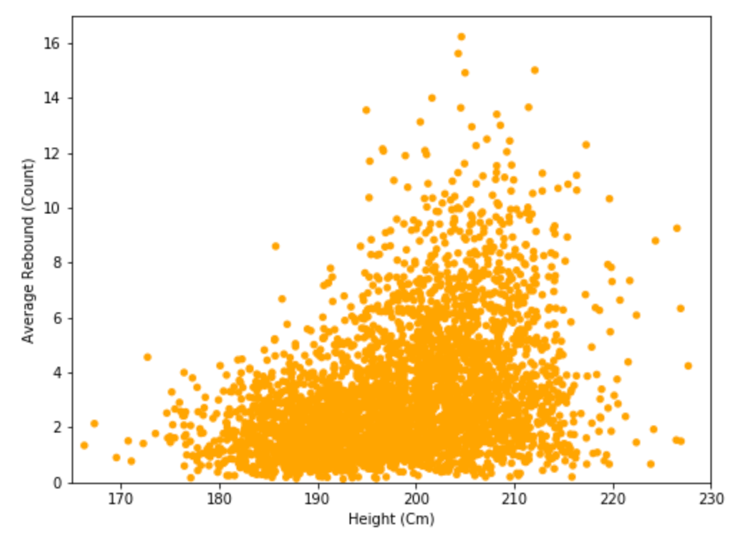
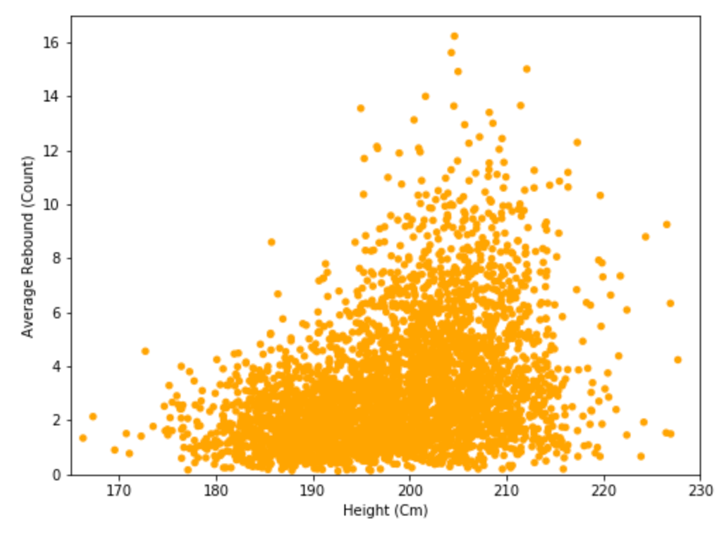


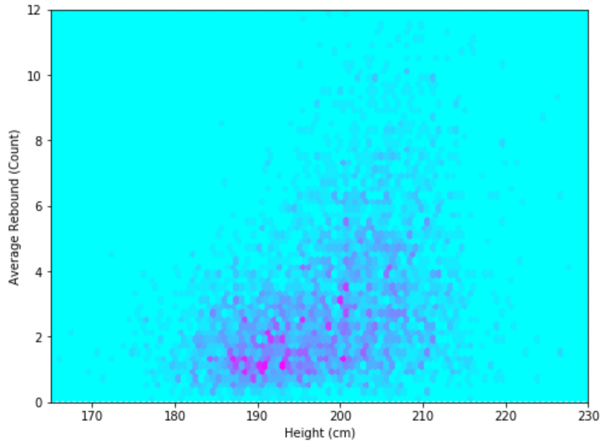
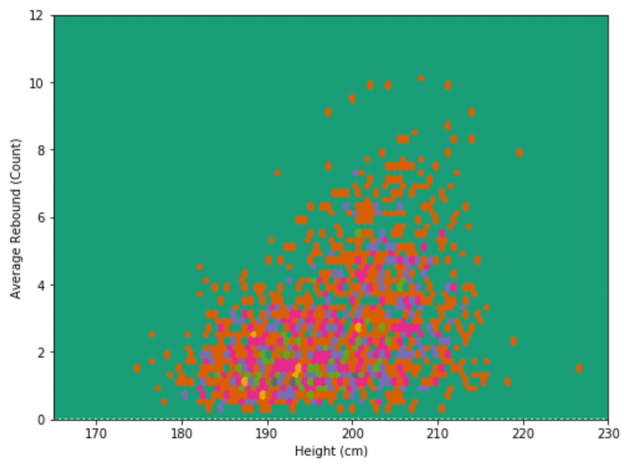
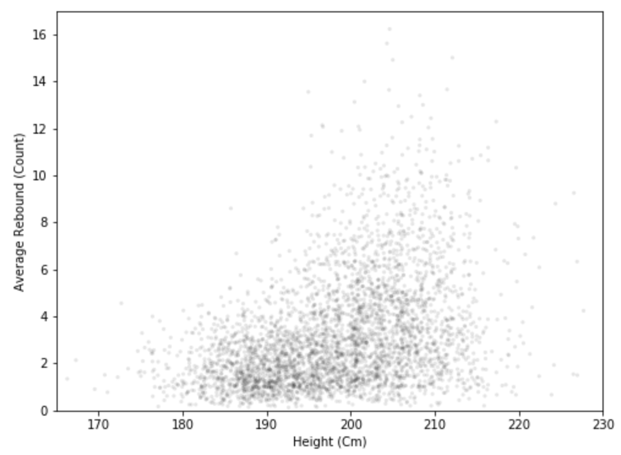
When we see the Height’s graphs we can say that classic height human error is valid. This is because when you ask, and record such as human height, weight, etc. Humans anser closest sum. So the data is faulty. In deeper analysis we will need Jittering (Aka. Adding Noise)

**Scatter Plotting The Data:**

 After getting the 5 descriptive statistics for each variable that I will work on. I made a scatter plot for seeing 2 variables together.

Here on the left we can see first step of the plotting. Now we can clearly see that height data needs Jitter. For making it more readable and geting more healthier conclution. Since height value is in cm we can jitter the data at value 1. So in the next step the data will be like more round up.



**** On top-left we can see the data after jittering. For analysing it at top-right you can see blue circle, all the basketball players can have average of 4 rebound after 185cm and 5 after 195cm and also in the red circle we can see only height of 195cm and more, can only achieve the average rebound rate that reaches 5 and more. So for even better reading of data and see the heatmap and more.

On top 3 graphs I did different and useful filtering to understand where are data stacked and more frequent. And in conclution of graphs there is obvious correlation of these two data by seeing only third graph and a line on top of it.

**Testing Hypothesis:**

After the graph analysis I did find Pearson Correlation.

P-value = 0.37 Which is there is positive correlation between this two variables that says that taller the player, higher the rebound chance.

R-squared = 0.14 = %14 Which is there is not much but still a correlation between these variables.

So If we remember our hypothesis It asks three things. First is taller the player, more rebound. And from the values I get, we can clearly say that this is valid. Second is average height and more must have at least 4 rebound. From the graphs and datas, we can say 200cm and more height has value of average 4.07 rebound. And last point of our hypothesis we assumed that correlation must have at least %5 for increase the teams rebound performance but its even better, our correlation value is %14 so it means for increase the teams rebound performance we can hire taller players.

Our hypothesis is **Valid**.

**Conclution:**

After testing the hypothesis and validate it, I can say that I am a little bit suprised what I find. Even though my personal guess was more then %25 correlation in the end there was only %14. For basketball statistics and sport %14 is enough but in the end I learned that shorter players have higher jump rate so they can contest the rebound with taller player so that is what makes rebound statistics and correlation low. One of the important findings was median and average height is also same and that was also suprising. In the end building and testing hypothesis and came out with outcome of valid made me to understand data science application and its magic.

References:

* <https://en.wikipedia.org/wiki/Rebound_(basketball)>
* <https://www.foxsports.com/nba/stats?season=2017&category=REBOUNDING&group=1&sort=8&time=0&pos=0&team=0&qual=1&sortOrder=0&opp=0&page=1>
* <https://squared2020.com/2017/07/31/rebounding-rates-good-for-teams-bad-for-players/>