CS240 - Exploratory Data Analysis

Final Project Report

Abdullah Ihsan Seçer

213972260

College of Engineering

Istanbul Sehir University

03.06.2018

SECTION 1

Before starting brainstorming questions related to basketball, it is indispensable to check which attributes the NBA datasets has. After checking the data, it is found that three plausible questions can be answerable using the data set:

- Does height of a player affect points a player scored?
- Does weight of a player affect three point scores?
- Does age of a player affect his blocking ability?

The first and third questions are somewhat seems me to have obvious answers. Height of a player will probably affect points scored by a player since taller players have more ability to dunk and score from a short distance. Age of a player will probably affect blocking ability since a player will be weaker in terms of body strength when he is older. As a result, it seems more interesting to answer the second question: does weight of a player affect three point scores? My hypothesis is there is a correlation between weight and number of three points he scores. Thus, the null hypothesis is there is no relationship at all.

- $H_0 :\ \rho = 0$
- $H_1 :\ \rho \neq 0$

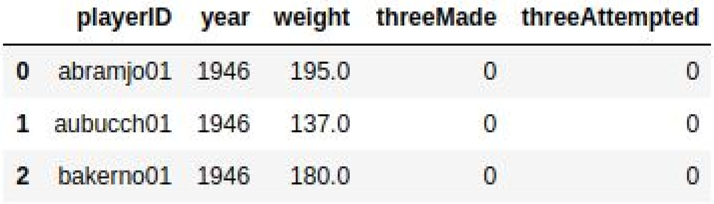
SECTION 2

Two datasets among given data are necessary to conduct analysis: basketball players and basketball master.

Attributes inside basketball_players.csv: **'playerID' 'year'** 'stint' 'tmID' 'lgID' 'GP' 'GS' 'minutes' 'points' 'oRebounds' 'dRebounds' 'rebounds' 'assists' 'steals' 'blocks' 'turnovers' 'PF' 'fgAttempted' 'fgMade' 'ftAttempted' 'ftMade' **'threeAttempted' 'threeMade'** 'PostGP' 'PostGS' 'PostMinutes' 'PostPoints' 'PostoRebounds' 'PostdRebounds' 'PostRebounds' 'PostAssists' 'PostSteals' 'PostBlocks' 'PostTurnovers' 'PostPF' 'PostfgAttempted' 'PostfgMade' 'PostftAttempted' 'PostftMade' 'PostthreeAttempted' 'PostthreeMade' 'note'

Attributes inside basketball_master.csv: **'bioID'** 'useFirst' 'firstName' 'middleName' 'lastName' 'nameGiven' 'fullGivenName' 'nameSuffix' 'nameNick' 'pos' 'firstseason' 'lastseason' 'height' **'weight'** 'college' 'collegeOther' 'birthDate' 'birthCity' 'birthState' 'birthCountry' 'highSchool' 'hsCity' 'hsState' 'hsCountry' 'deathDate' 'race'

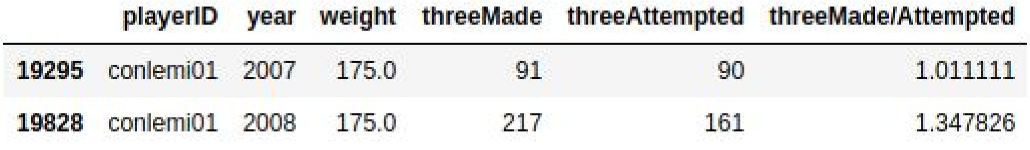
Four attributes inside basketball players and two attributes inside basketball master datasets are used. ‘bioID’ and ‘playerID’ attributes are used to merge two datasets. ‘year’ attributes is used since there is player data for various years. ‘threeMade’ is used to find out number of threes made by a player. 'threeAttempted' is used to find out proportion of numbers of three made and attempts to made it so that I can conduct analysis on success ratio too. ‘weight’ is simply used to get weights of players.



	playerID	year	weight	threeMade	threeAttempted
0	abramjo01	1946	195.0	0	0
1	aubucch01	1946	137.0	0	0
2	bakerno01	1946	180.0	0	0

Table 2.1: Resulting data frame after merge.

To find out success rates of three score attempts, I divide ‘threeMade’ to ‘threeAttempted’ to create a new column ‘threeMade/Attempted’. After creating it, I found out there are values which are higher than 1, which should not happen normally. I assumed that one of columns are wrongly counted. As a result I decided to equalize ‘threeMade’ to ‘threeAttempted’ where ‘threeMade/Attempted’ is higher than 1 and equalize ‘threeMade/Attempted’ to 1. There might be other assumptions held in this case. For instance, the one entering data might enter ‘threeMade’ to ‘threeAttempted’ and vice versa. However, we cannot figure out it from existing datasets.



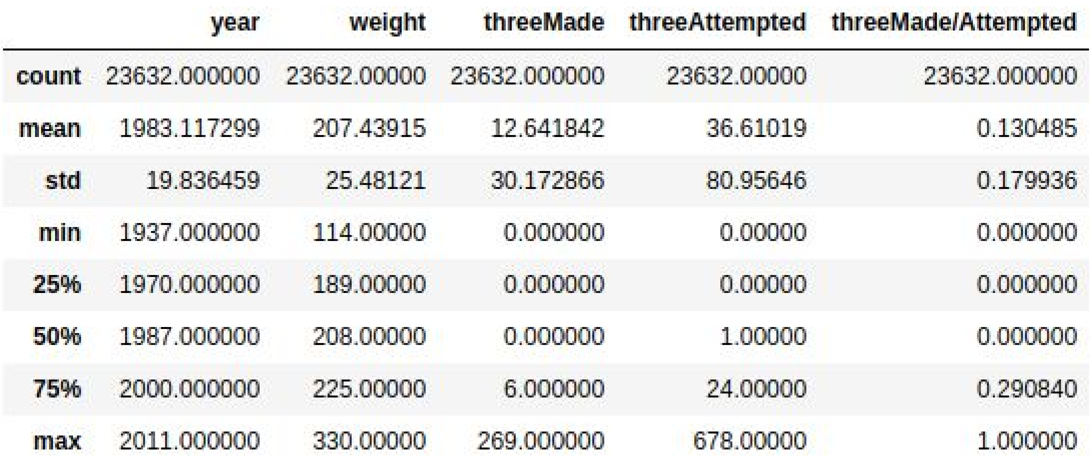
	playerID	year	weight	threeMade	threeAttempted	threeMade/Attempted
19295	conlemi01	2007	175.0	91	90	1.011111
19828	conlemi01	2008	175.0	217	161	1.347826

Table 2.2: Rows where ‘threeMade/Attempted’ is higher than 1

Another significant thing to consider is that one can have ‘threeMade/Attempted’ value equal to 1 while having a few number of three attempts. In the data frame, there are 164 players with ‘threeMade/Attempted’ equal to 1 and 162 of them are attempted to three score less than four times. Thus, both ‘threeMade’ and ‘threeMade/Attempted’ are important in this analysis.

SECTION 3

Calculating descriptive statistics resulted in realization of another factor which can cause noise. There are rows where weight value is equal to 0. To get rid of noise caused by these rows, I dropped off these rows and then started to analyze the data frame.



	year	weight	threeMade	threeAttempted	threeMade/Attempted
count	23632.000000	23632.00000	23632.000000	23632.00000	23632.000000
mean	1983.117299	207.43915	12.641842	36.61019	0.130485
std	19.836459	25.48121	30.172866	80.95646	0.179936
min	1937.000000	114.00000	0.000000	0.00000	0.000000
25%	1970.000000	189.00000	0.000000	0.00000	0.000000
50%	1987.000000	208.00000	0.000000	1.00000	0.000000
75%	2000.000000	225.00000	6.000000	24.00000	0.290840
max	2011.000000	330.00000	269.000000	678.00000	1.000000

Table 3.1: Descriptive statistics

Looking at Table 3.1, we can analyze general characteristics of the cleaned data which has 23642 rows in it. Year represents a year of a season which a player is attended. The minimum year is 1937 while the maximum is 2011. From percentiles, we can deduce that most of data belongs to years ranging from 1970 to 2000. The mean is 1983, which indicates that there are more instances belongs to recent seasons than older ones considerings minimum and maximum values.

The unit of weight is pound (Lbs). To contrast with the unit of kilogram, 1 Lbs is approximately equal to 0,45 kg. Mean of weight is around 207 while median is 208. Thus, we can assume that there are not many number of outliers. The maximum value is 330 while the minimum is 114. In addition to these values, after looking at ranges of percentiles, we can expect that the weights are somewhat evenly distributed.

The descriptive statistics of other three variables can be analyzed together. The number of threes attempted are higher than number of successful ones as usual. Respectively, their means are 12 and 36. The standard deviation of threes attempted are almost three times higher than number of threes made. This can be because of the difficulty of a successful three score. 25 and 50 percentiles are 0 for ‘threesMade’. 25 percentile is again 0 for ‘threesAttempted’ while 50 percentile is equal to 1. From this statistics, we can also deduce that half of players do not attempted to make a three score or attempted at most one time. The remaining meaningful statistics for ‘threeMade/Attempted’ column are mean, standard deviation and 75 percentile. Looking at 75 percentile, we can deduce that, 75% of player’s success in three scores are 29%. Here it is important to note that if a player does not attempted to make a three score his success is count as 0. The mean of ‘threeMade/Attempted’ is 0.13, which is very low. On the other hand, considering the possible interval between 0 and 1, the standard deviation is high. As a result, we can deduce that there are players which are significantly successful to made three scores.

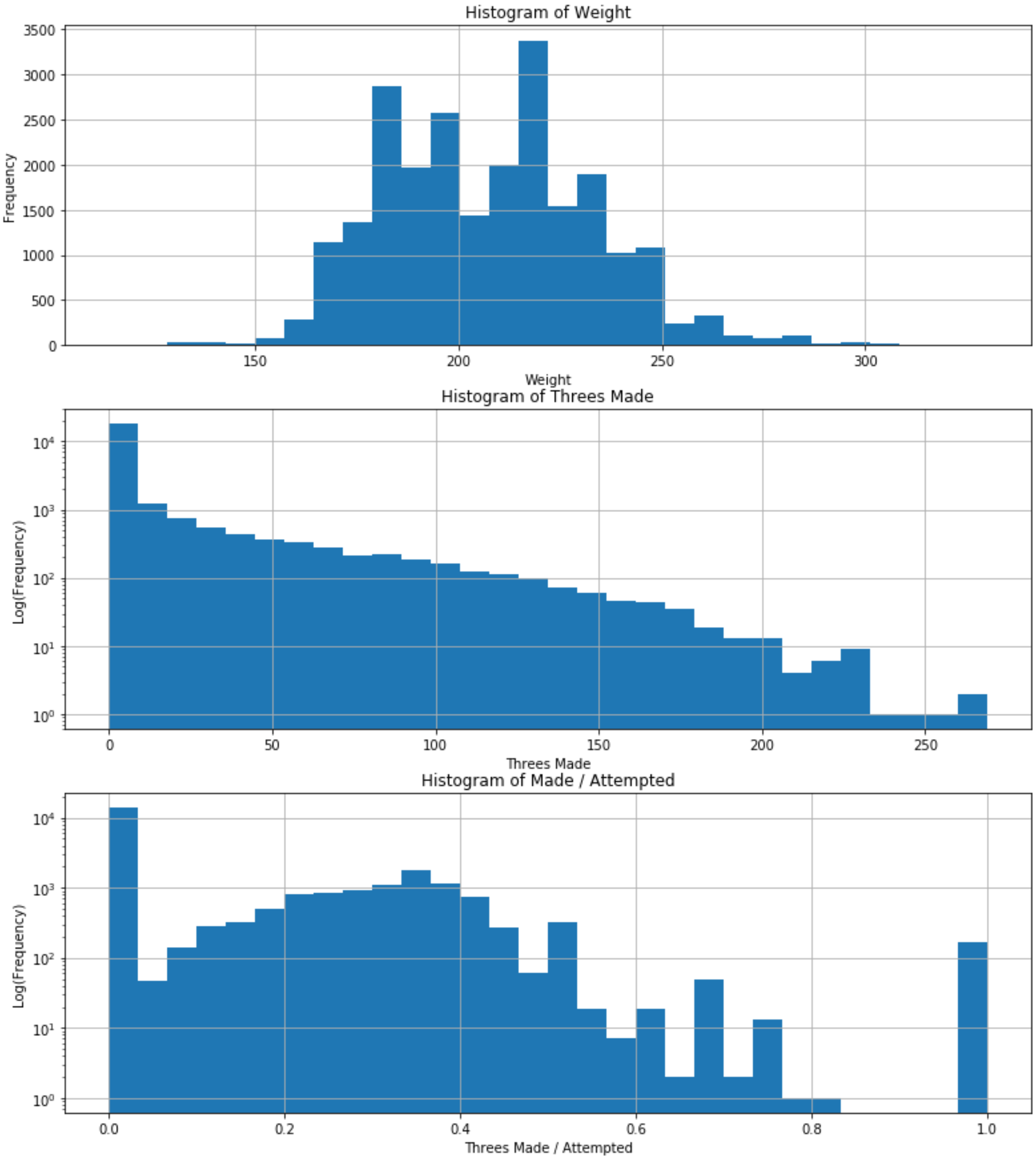


Figure 3.1: Histogram of Weight, Three Made and Three Made / Attempted

By looking at the first histogram in Figure 3.1, we can observe a normal distribution in weight of players. The threes made and threes made / attempted column’s histograms are not meaningful in the normal scale. Thus, their histograms are created using logarithms of their frequencies. By contrasting these two histograms, we can see that while threes made are decreasing almost continuously while there is an increase from 0 (excluded) to 0.4 in threes made / threes attempted. I did not plot threes attempted since it is not directly related to the selected question.

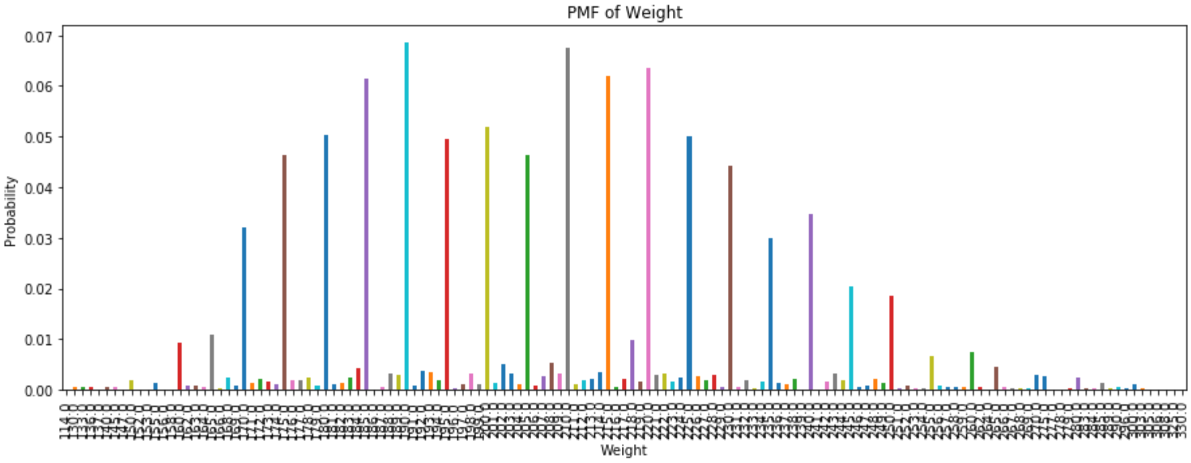


Figure 3.2: PMF of Weight

PMF plots of these three variables does not account for much information. The three score related columns did not plotted due to scale of distributions. These plots are meaningless in the normal scale. Since the weight data has continuous variables it is not easy to interpret its PMF plot. However, we can find out most probable weights through it. For instance, the two most probable weights are 190 and 210 Lbs. Between 0.02 and 0.07, there are 17 weight values. However, most of the values have probability less than 0.1. We can also observe a pattern similar to normal distribution in this plot as usual.

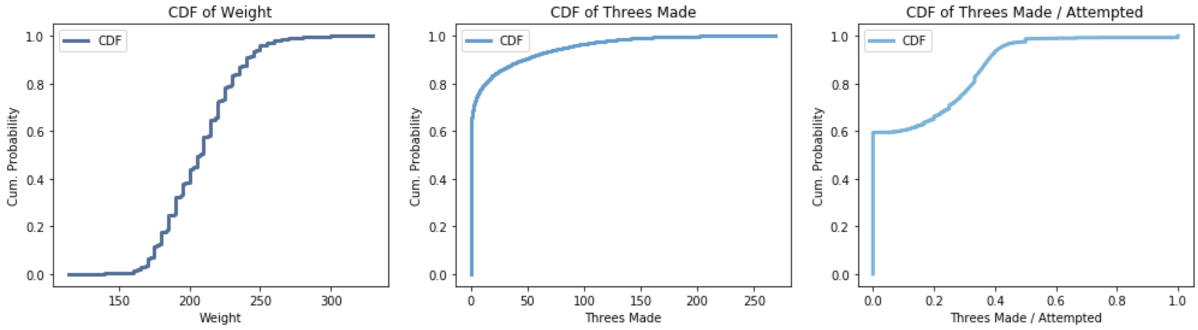


Figure 3.3: CDF of Weight, Three Made and Three Made / Attempted

In CDF plots, we can observe cumulative probabilities for the three columns. The CDF of weight mostly changes between 150 and 250. It means that the most of the data remains in this interval. The increasing steepness in the plot around 220 means that the majority of instances are in this region. Normality of weight can also observed from the CDF plot. CDF of threes made shows that most instances are accumulated at 0. So most of players did not made a three score. The steepness is slowly decreasing, which means that as number of threes made increasing, number of players are decreasing. By looking at, CDF of threes made / attempted, we can observe a gap between 0.0 and 0.2. It means that the number of players are not so much in this interval. Around 0.4, the steepness is peaked, meaning that there are high number of players in this region in contrast to the other regions.

SECTION 4

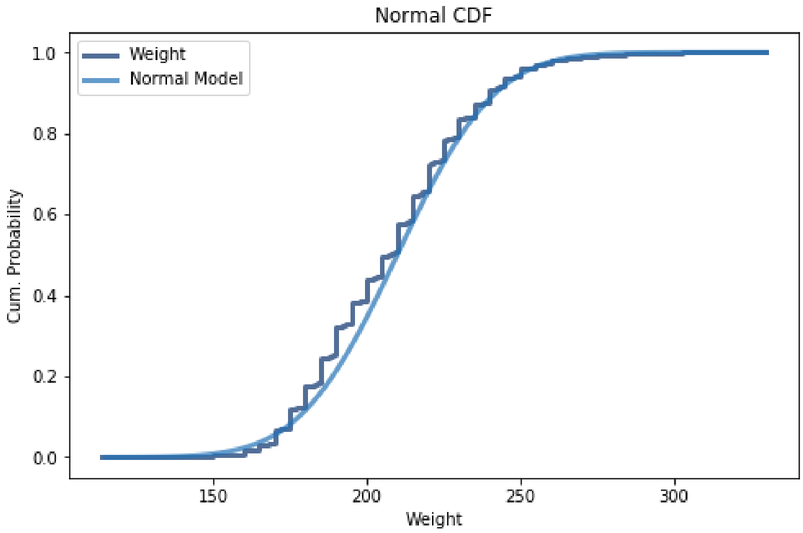


Figure 4.1: CDF of normal model fitted and weight

As seen in the plot, the weight distribution mostly fits to a normal distribution. In case we assume that the distribution is perfectly normal, the 68% of instances are between mean plus standard deviation and mean minus standard deviation. Thus, we can assume that majority of players have weights between 182 and 232 Lbs. However, the data is not perfectly fitting the normal distribution. There are tails which are not evenly distributed and also around mean there are abnormalities on distribution of weight. However, the given interval might be still account for majority of people as seen from the plot.

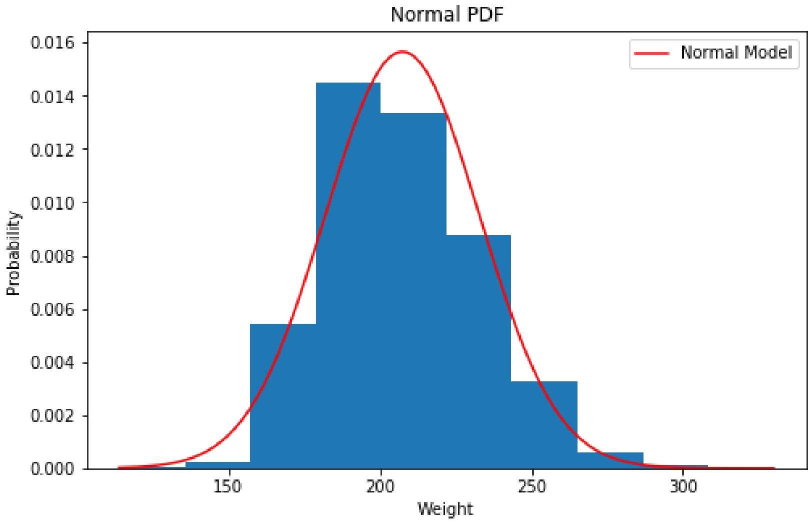


Figure 4.2: PDF of normal model fitted and weight

SECTION 5

There are two correlations to calculate related to my hypothesis. One correlation is the correlation between ‘weight’ and ‘threesMade’. Other one is between ‘weight’ and ‘threesMade/Attempted’. First one does not show the success of players to made a three score. Thus, we need the other correlation score. However, the ‘threesMade/Attempted’ fails to consider the ones with no attempt and the ones with a very few attempt with 100% success. Two metrics are used to calculate correlations:

- Correlation between ‘weight’ and ‘threesMade’
    - Pearson: - 0.10
    - Spearman: - 0.15
- Correlation between ‘weight’ and ‘threesMade/Attempted’
    - Pearson: - 0.10
    - Spearman: - 0.12

It seems that both two correlations are negative. So while the weight of a player is increasing the number of threes made and the proportion of threes made / attempted are decreasing and vice versa. However, the correlations are not strong considering that the scores are between - 0.1 and - 0.15.

Pearson’s correlation score is affected by outliers and skewness of the data. So, taking these into account, the calculated pearson correlation score is - 0.1 for each variables. On the other hand, spearman’s correlation score is not affected by outliers and skewness. It is actually calculated using pearson correlation formula but with percentile ranks. The spearman correlation score for weight and and threes made is 0.15. For weight and threes made / attempted it is 0.12.

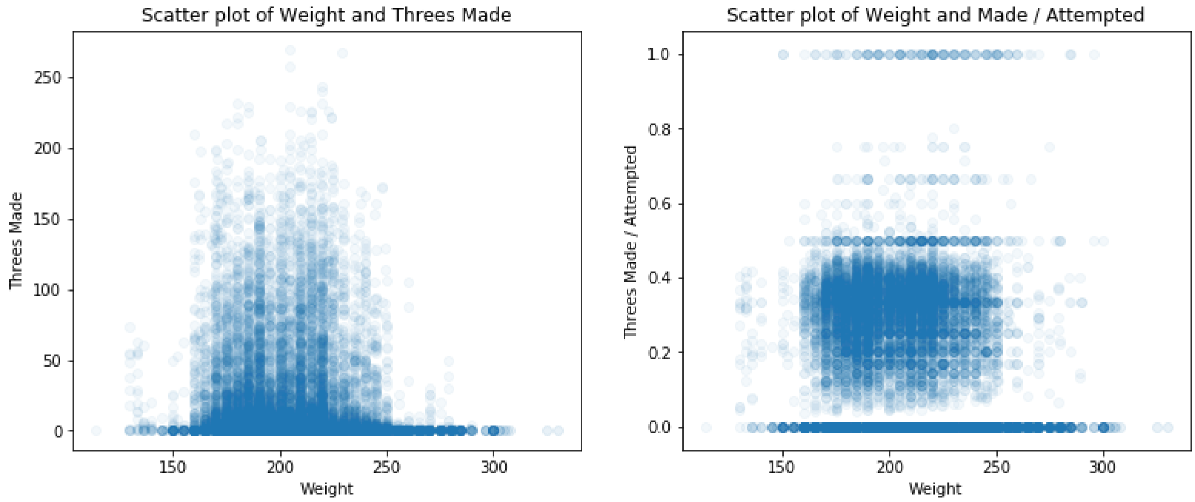


Figure 5.1: Scatter plots of Weight and Three Made, Three Made / Attempted

By looking at scatter plots, we cannot come up with any deduction related to a linear relationship. However we can observe number of threes made and the success rate are higher between 150 and 250 Lbs. However, it is probably because of the normal distribution.

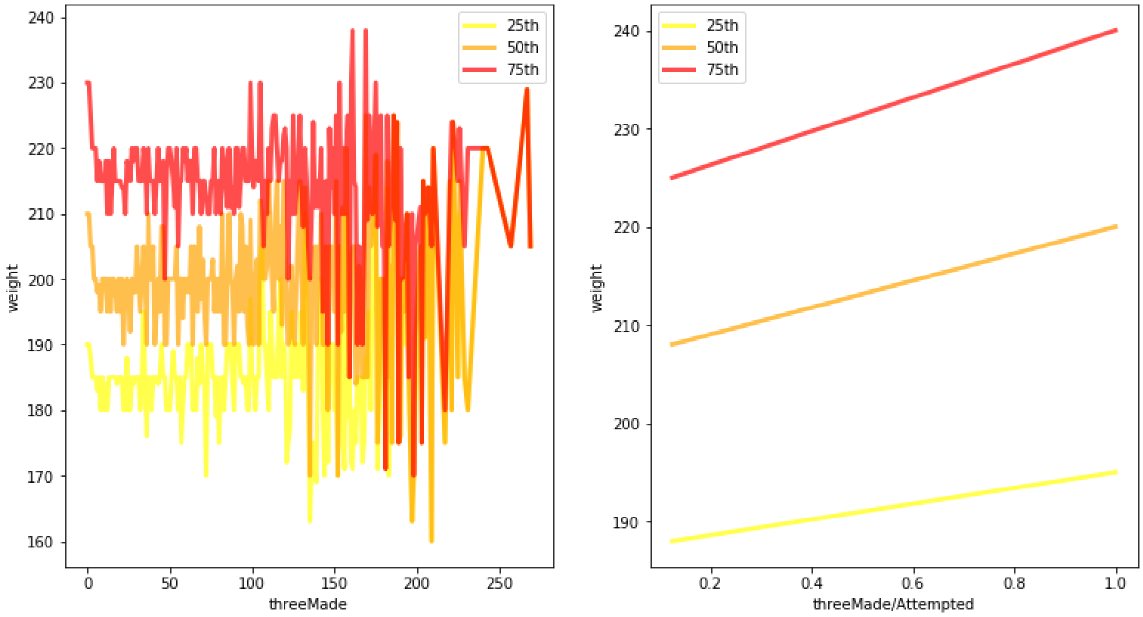


Figure 5.2: Percentile plots of Weight with respect to Three Made and Three Made / Attempted

By looking at the fist plot of Figure 5.2, we can deduce that players with weight between 180 and 230 tend to make three scores more. As number of threes made decrease, gaps between percentiles becomes more normal. The second plot indicates that when success rate of three attempts increases the height percentiles increase.

SECTION 6

To test if the correlation found in the sample is by chance or not, we can conduct hypothesis testing using permutations. We need three typical functions for our test. Firstly, we need a function (RunModel) to run a simulation which we will use to create different permutations of data. Then, we need function (TestStatistic) to calculate a test statistics in each simulation. In our case our test statistics is the absolute value of correlation score. Before all of them, we need to initialize the data and then calculate the test statistic using the original form of the data. Lastly, we need a function (PValue) which calculates p-value using all of the defined functions. In this function, we call RunModel function number of times to create different permutations. Then, for each run we call TestStatistics to calculate our test statistic in each permutation. After these steps, we check how many times the test statistic is not holding the actual test statistic calculated using the original form of the data. The p-value is the number of times the test statistics differ divided by total number of simulations. After all steps, we calculated the proportion that we mistakenly assume that the alternative hypothesis is true, in other words the p-value. In our case, the calculated probability in a thousand simulation is 0.0 for each of four correlation calculation. If we increase number of simulations using a powerful computer, the p-value might increase in a sense. However, we can reject the null hypothesis according to our tests.

SECTION 7

All in all, we can deduce that there is a possible negative correlation between threes made and number of successful threes with weight of a player. However, they are not strong correlations. For number of times we can observe that players with weight around 180 and 230 are the player with most three scores made. However, it is because most players are in this region. We observed that number of players that attempted to make a three score is very low. Furthermore, half of the player did not achieve to make a three score or achieved to make one three score. Among players who tried to make a three score, many players have a success rate between 0.2 and 0.4. There are also players which are fully successful in their three score attempts but most of them attempted to make a three score at most 3 times.