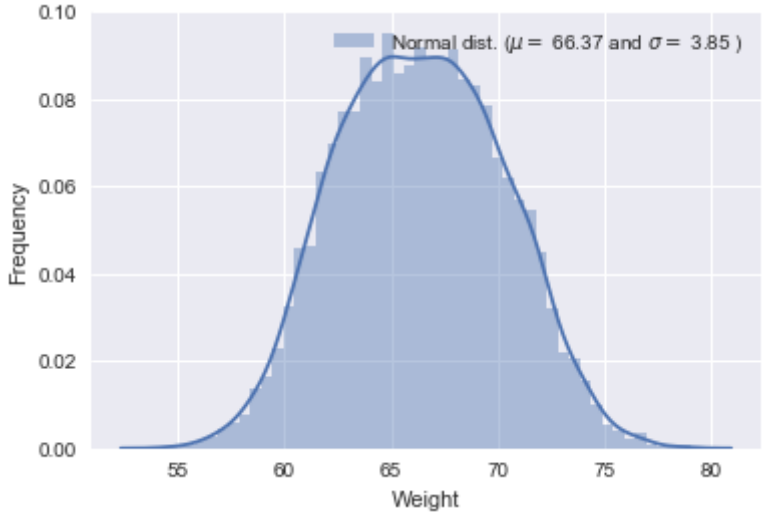
Assignment 1

In our data set we have 3 attributes: Gender, Weight and Height, where weight and height are continuous numerical attributes (these are numbers which can range from negative infinity to positive infinity). Theoretically they may contain any value within some range, range for weight approximately is from 1 kg to 300 kg and for height is from 0.3 meter to 2.5 meter , and gender is nominal categorical attribute (these variables can have a limited set of values, each of which indicate a sub-type) it is either 0 or 1.

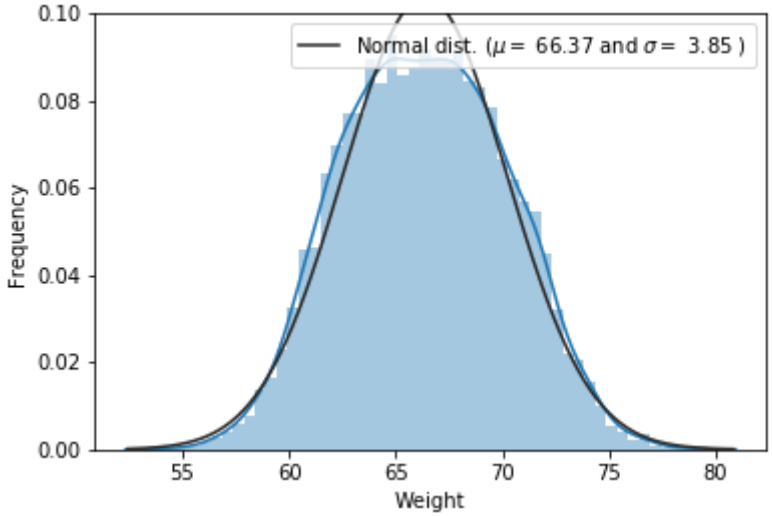
So let's look at the nature of each attribute.

The mean value for weight is 66.36 which is close to median, 66.31. This is hint for us that the graph of distribution can be bell shaped. That's true:

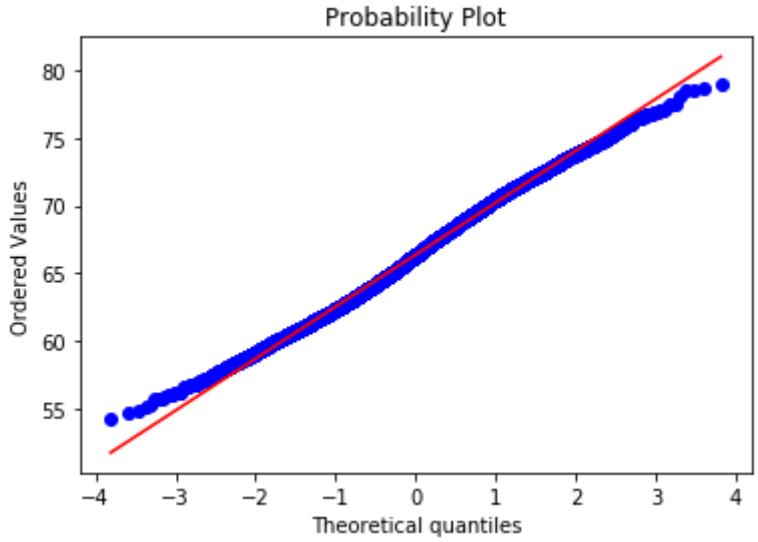


We also can notice that weight has "unimodal" distribution and it is close to normal distribution curve but has heavy tails.

The mode of a data set is the value that appears the most frequently in it. Unimodal distribution is when the data set has a single mode.

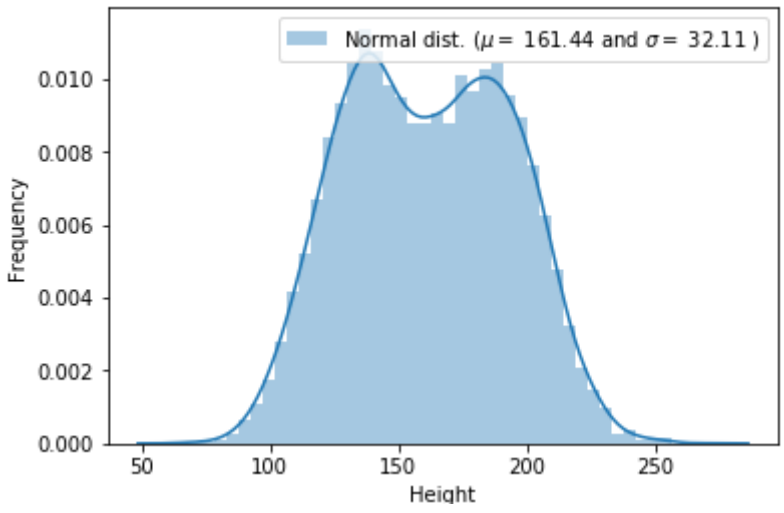


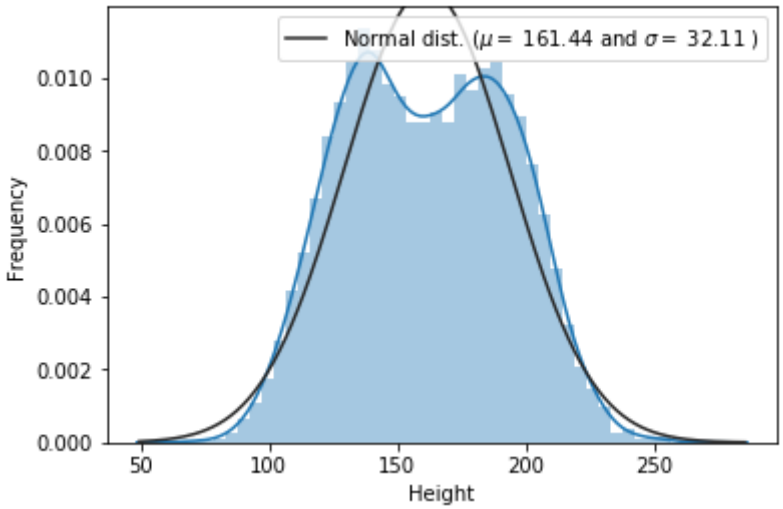
About heavy tails we can know from “Q-Q” plot as well:



If we draw the Q-Q Plot for two distributions A and B, both distributions have right tails (are non-zero in [a, +∞) for some a), and the tails of A are thinner than the tails of B, then the right-hand side of the Q-Q Plot (assuming that the quantiles of A are on the x-axis) will be convex-shaped. The inverse is true for the left-tailed distributions: if the left tail of B is fatter than the left tail of A, then on the left-hand side of the Q-Q Plot (again assuming that the quantiles of A are on the x-axis) we will have a concave-shaped graph.

The mean value for height attribute is 161.44 and the median is 161.21. The distribution of Height and comparison with normal distribution curve are the following:

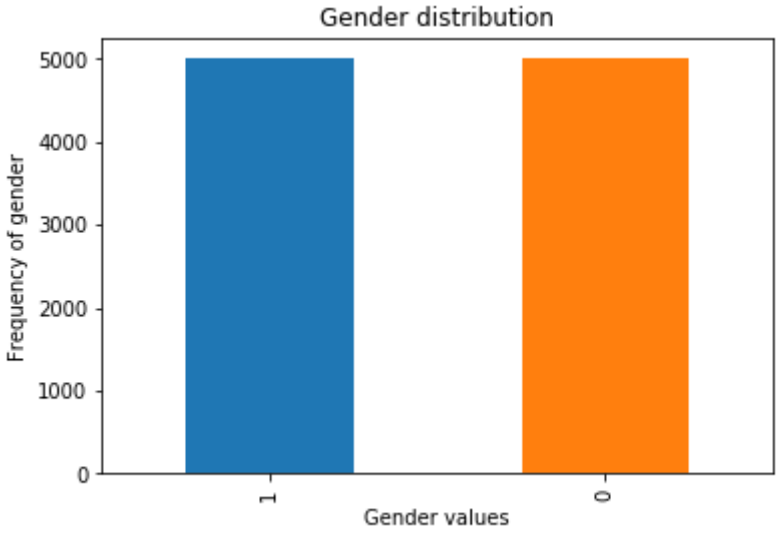




We can see that distribution of height is bimodal (has two different modes).

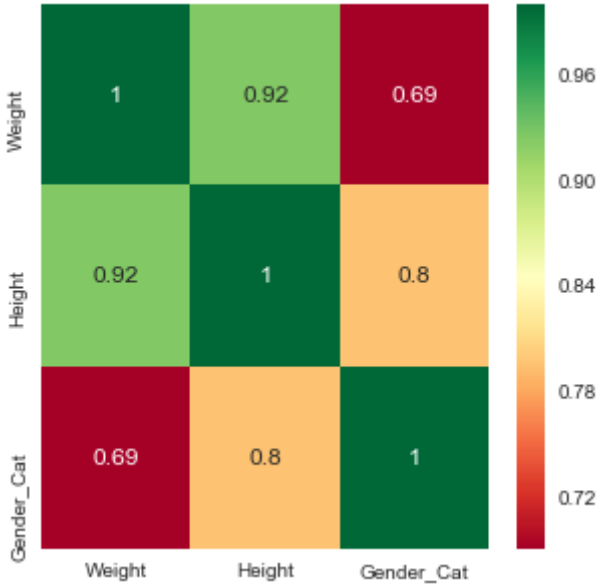
I guess you can ask why we need to know about distribution of attributes or why we are doing comparison between distributions of weight (or height) and normal. I will answer in advance. The matter is that,

Now let's see what is the frequency ratio between two genders:



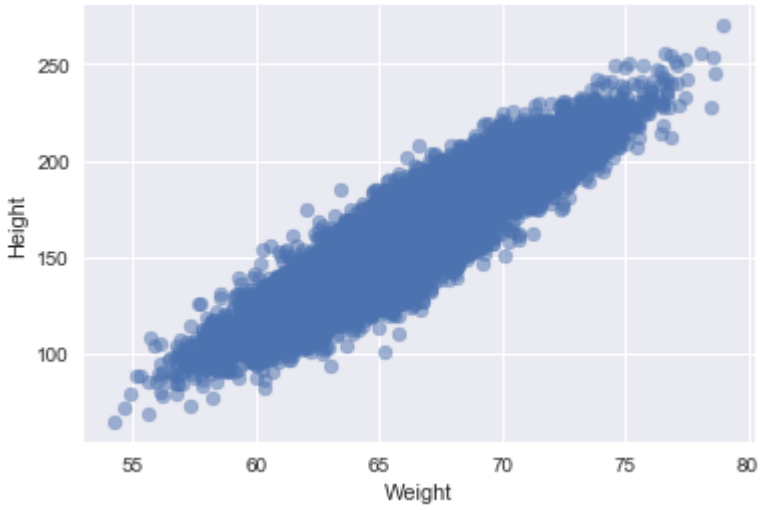
From the above picture we can say, that there are about 5000 male and 5000 female (total number of people in dataset is 10000).

Now as we know the distributions of all attributes, let's look at the correlation (Correlation is a statistical technique that can show whether and how strongly pairs of attributes are related. In other words, it shows whether there is a linear relationship between two attributes or not.) between them.



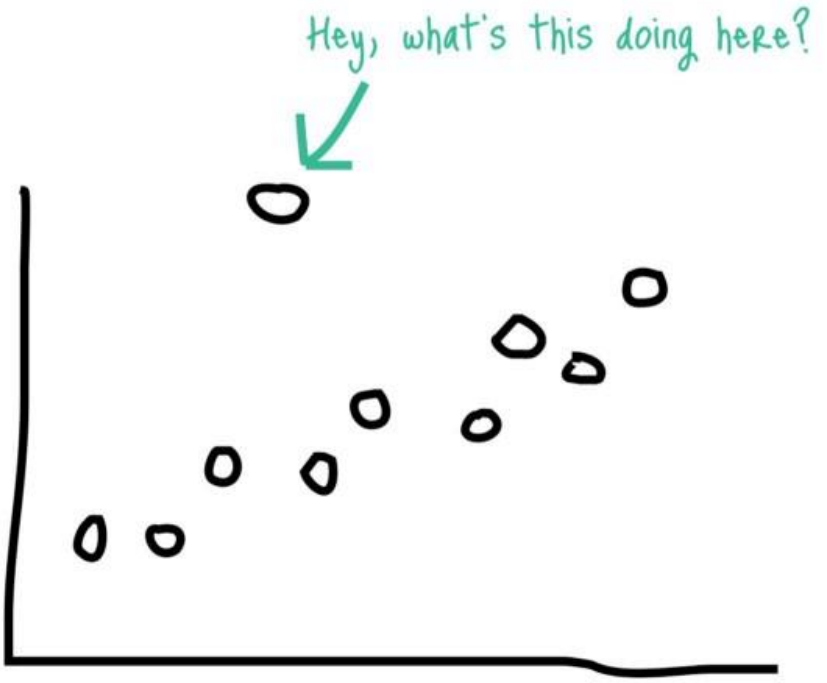
From the grid above, we can say that there is a strong relationship between height and weight and that is logically true (so there is no just correlation but also causation). We also see that there is a relationship between gender and height.

We can check linear relationship in other ways. One of that is scatter plot (A scatter plot is a two-dimensional data visualization that uses dots to represent the values obtained for two different variables - one plotted along the x-axis and the other plotted along the y-axis).



It is important to emphasize outliers if they exists in dataset. I'll use Box plot method to detect outliers. First, I would like to give some explanations about outliers and the IQR method.

An outlier is a data point that is distant from other similar points. Further simplifying an outlier is an observation that lies on abnormal observation amongst the normal observations in a sample set of population.

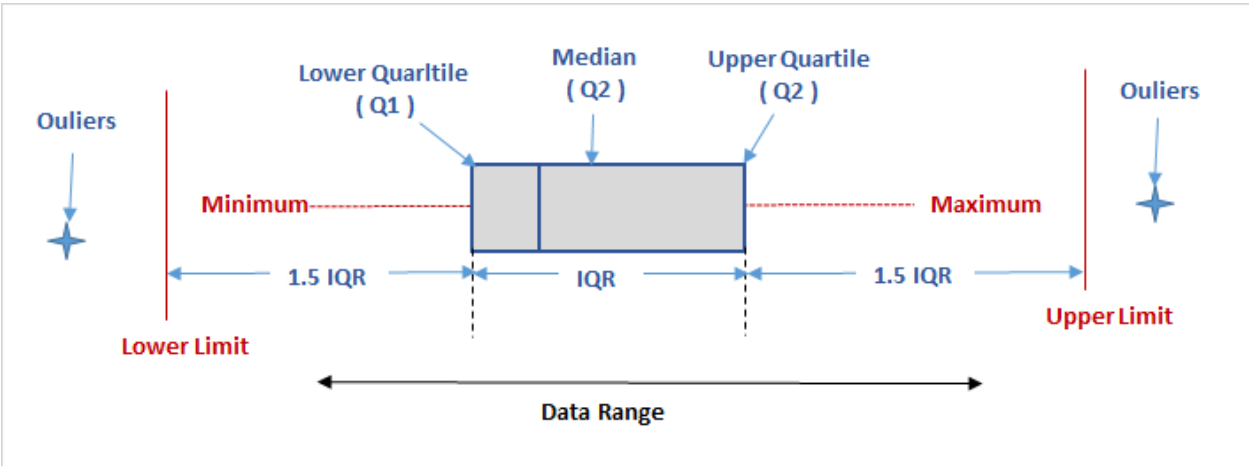


In statistics, an outlier is an observation point that is distant from other observations.

IQR is one of the robust methods for labeling outliers is the IQR (interquartile range) method of outlier detection developed by John Tukey, the pioneer of exploratory data analysis. This was in the days of calculation and plotting by hand, so the datasets involved were typically small, and the emphasis was on understanding the story the data told. If you've seen a box-and-whisker plot (also a Tukey contribution), you've seen this method in action.1

A box-and-whisker plot uses quartiles (points that divide the data into four groups of equal size) to plot the shape of the data. The box represents the 1st and 3rd quartiles, which are equal to the 25th and 75th percentiles. The line inside the box represents the 2nd quartile, which is the median.

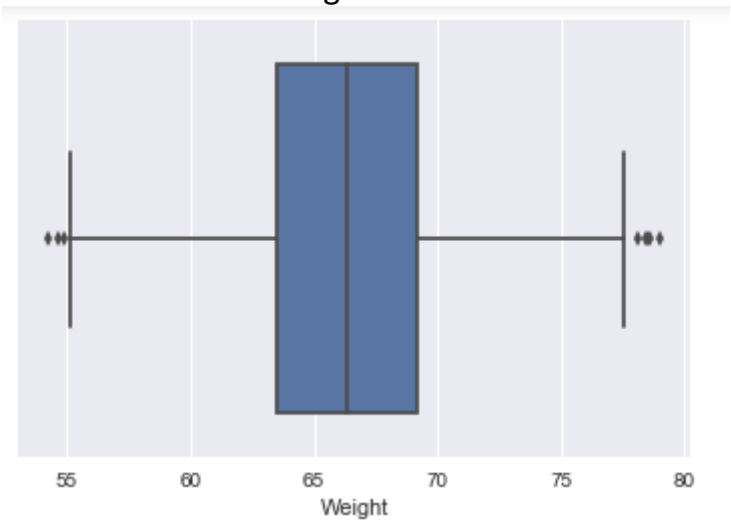
The interquartile range, which gives this method of outlier detection its name, is the range between the first and the third quartiles (the edges of the box). Tukey considered any data point that fell outside of either 1.5 times the IQR below the first – or 1.5 times the IQR above the third – quartile to be "outside" or "far out". In a classic box-and-whisker plot, the 'whiskers' extend up to the last data point that is not "outside".



In our case we have 8 outliers:

Gender	Weight	Height
Male	78.095867	255.690835
Male	78.462053	227.342565
Male	78.998742	269.989699
Male	78.528210	253.889004
Male	78.621374	245.733783
Female	54.616858	71.393749
Female	54.873728	78.606670
Female	54.263133	64.700127

Another way to find outliers is graphically, box plot.
For weight attribute:



And for height attribute: