STATISTICS FOR DATA SCIENCE PART - 2

Estimates of Location:

- A variable has many values. For example age is a variable which can take values from 0 to 125 years(world record is 122).
- So a typical value for each feature can be handy for exploration in this kind of situation.
- One of the best ways to select that typical value is by choosing the value where most of the values are located.
- This can be expressed by using one word *Central Tendency.*

Key Terms in Estimates of Location:

- The key terms used in central tendency are

  Mean: The sum of all values divided by total number of values. Also known as average.

  Weighted Mean: The sum of all values multiplied by a weight divided by sum of all weights. Also known as weighted average.

  Trimmed Mean: The sum of all values divided by total number of values after removing some values. Also known as truncated mean.

  Median: The value such that half of the data lies above and the remaining half lies below. Also known as 50th percentile.

  Weighted Median: The value such that half of the sum of weights lies above and the other half lies below.

  Outlier: A data value which is very different from most of data.Also known as extreme value.

  Robust: Data that is not effected by extreme values. Also known as resistant.

Mean:

- The most basic estimate of location.
- Consider the numbers 2 3 4 9 6 then the mean of these numbers is (2+3+4+9+6)/5 which is 5.
- Mean x = $(x_1+x_2+\ldots+x_n)/n$.

- Here *n* refers to the total number of records.
- The convention followed for total number of records is *n* if we are dealing with a sample drawn from a population and *N* if we are dealing with population.
- In case of trimmed mean we sort the values and remove first *p* and last *p* values so that the mean is not sensitive for outliers.
- In many cases trimmed mean is more preferable than the normal mean.
- The third case is a weighted mean where each data element x_i is multiplied with some weight w_i and these are summed up then divided by sum of all weights $w_{i.}$

  Weighted Mean x_w = $(x_1w_1+x_2w_2+\ldots\ldots.+x_nw_n)/(w_1+w_2+\ldots+w_n)$.
- This weighted mean is useful in cases like some sensors are more accurate and some are less accurate.

Median and Robust Estimates:

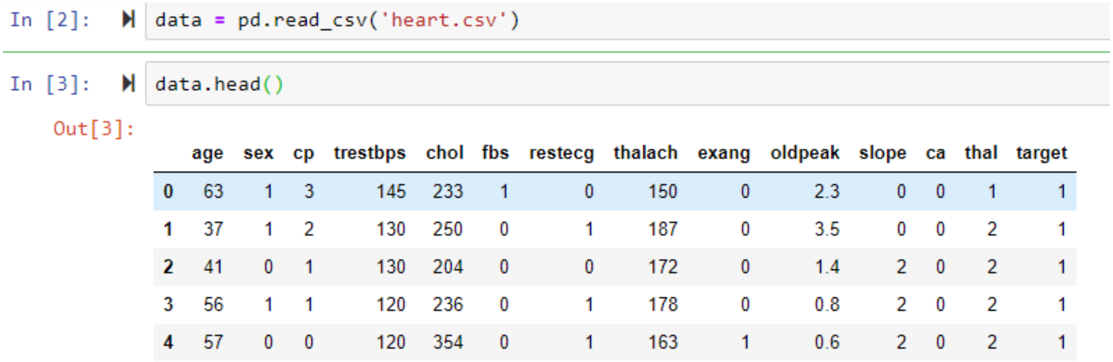
- Simply median can be defined as the middle value of a sorted list of elements.
- For example consider the numbers 2,3,5,1,8 then their median is 3.
- The major difference between mean and median is that mean depends on all the values of the feature whereas the median depends on the middle values.
- In some cases median works better than mean.
- For example if we take the average income of hundred houses that includes Bill Gates house then the mean is not a good estimate but median can tell the average income as it doesn't depend on Bill Gates income.
- It is also possible to calculate the weighted median.
- First the elements are multiplied by their respective weights and then sorted.
- The weighted median is a number such that the lower half and the upper half have the same weight sums.
- Median and Weighted medians are robust to outliers.

Outliers:

- An outlier is a value that is distant from any other value in the data.
- The major reason for outliers is either the bad observation (bad sensor) or the usage of a wrong unit (grams instead of kilograms).
- Mean is sensitive to these outliers whereas median is not effected by them.
- Trimmed mean is also robust for outliers but requires more data to correctly locate the value.
- Thus trimmed mean can be treated as a compromise between mean and median.

Python Implementation:

- Now It is time for practical implementation. In this section implementation of mean in Python is discussed.
- Consider the data frame *Heart.csv*



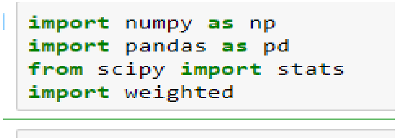
```
In [2]:  data = pd.read_csv('heart.csv')
```

```
In [3]:  data.head()
```

Out[3]:

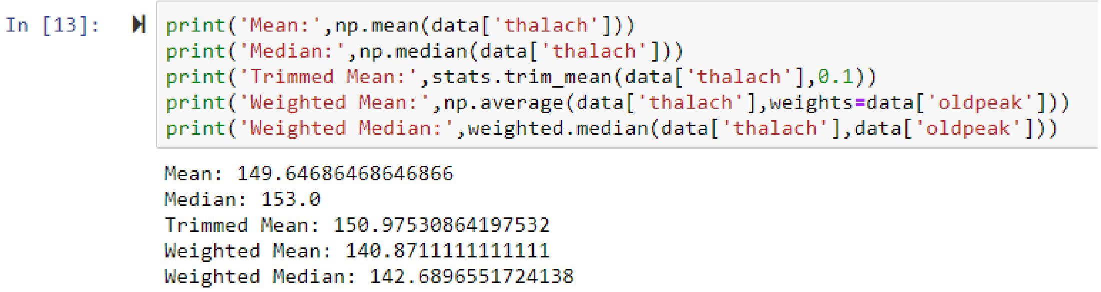
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

- The Libraries to be imported are



```
import numpy as np
import pandas as pd
from scipy import stats
import weighted
```

- The code snippet and output are



```
In [13]:  print('Mean:',np.mean(data['thalach']))
          print('Median:',np.median(data['thalach']))
          print('Trimmed Mean:',stats.trim_mean(data['thalach'],0.1))
          print('Weighted Mean:',np.average(data['thalach'],weights=data['oldpeak']))
          print('Weighted Median:',weighted.median(data['thalach'],data['oldpeak']))
```

```
Mean: 149.64686468646866
Median: 153.0
Trimmed Mean: 150.97530864197532
Weighted Mean: 140.8711111111111
Weighted Median: 142.6896551724138
```