STATISTICS FOR DATA SCIENCE PART - 4

Exploring the Data Distributions:

- The topics related to the location and variability are covered. Now it is time to understand the distribution of the data.
- Data distribution is very useful in many cases and machine learning models and statistical modelling depends on this distribution.
- To understand the data distribution different plots and mathematical expressions are used.

Key terms in exploring the Data Distributions:

- The key terms used in data distributions are

  Boxplot:A plot introduced as a quick way to visualize the distribution of data. Also known as Box,whiskers plot.

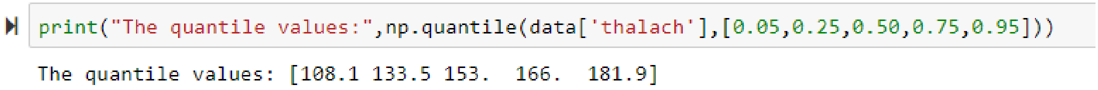
  Frequency table:A tally of the count of numeric data values that fall into a set of intervals (bins).

  Histogram:A plot of the frequency table with the bins on the x-axis and the count (or proportion) on the y- axis.

  Density plot:A smoothed version of the histogram, often based on a *kernal density estimate*.

Percentiles and Boxplots:

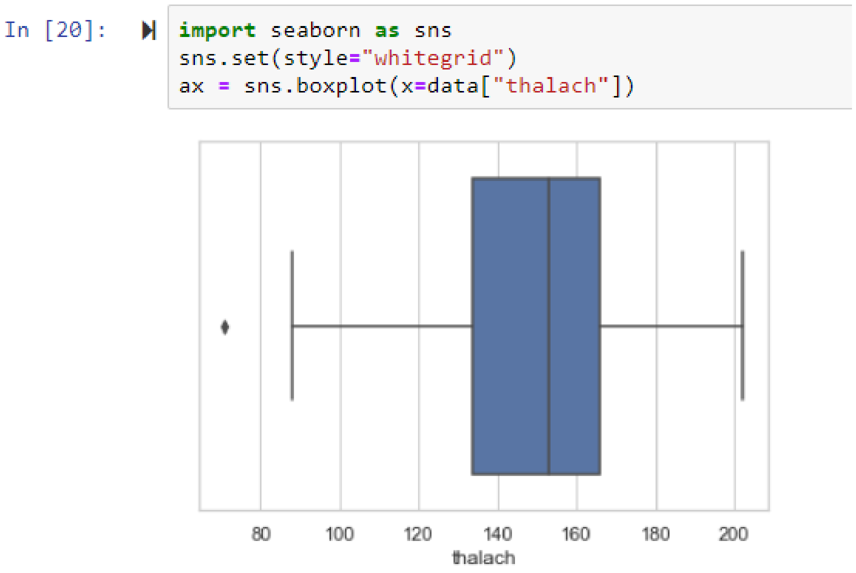
- Percentiles are also useful in summarizing the data distribution. Percentiles are mainly useful to understand the tail(outer range) data distribution.
- The quantile values of *thalacha* can be shown as



```
print("The quantile values:",np.quantile(data['thalach'],[0.05,0.25,0.50,0.75,0.95]))
```

```
The quantile values: [108.1 133.5 153.  166.  181.9]
```

- This shows that the 5th, 25th, 50th, 75th, and 95th percentiles are 108.1,133.5,153.0, 166.0 and 181.9 respectively.
- The boxplot for this can be drawn as

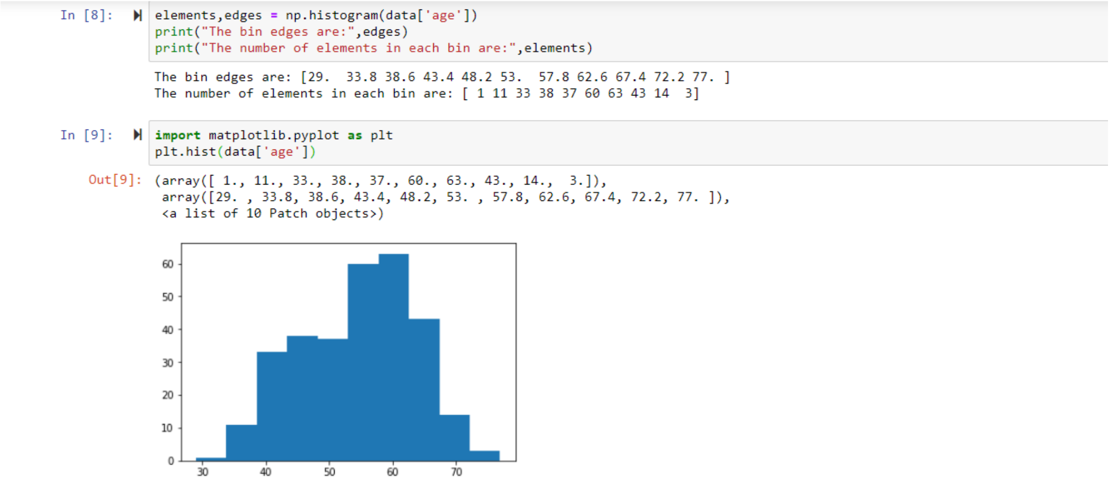


```
In [20]:  import seaborn as sns
          sns.set(style="whitegrid")
          ax = sns.boxplot(x=data["thalach"])
```

- The left and right ends are 25th percentile and 75th percentile.The middle vertical line shows the median. The left most and right most vertical bars also known as whiskers tell about the range of data.

Frequency Table and Histograms:

- A frequency table of a variable divides up the variable range into equally spaced segments.
- Consider the same *Heart.csv* file. Now the *age* in this file is divided into bins. By default the number of bins is 10 in python.
- The edges of a histogram are the starting point of a new bin. The difference between the edges of the bin define the bin width.
- The number of elements in each bin define how many elements are present in each interval.
- Histograms can be drawn to these bins. The bin edges are taken on the X-axis and the number of elements in each bin are taken on the Y-axis.
- Python implementation can be shown as

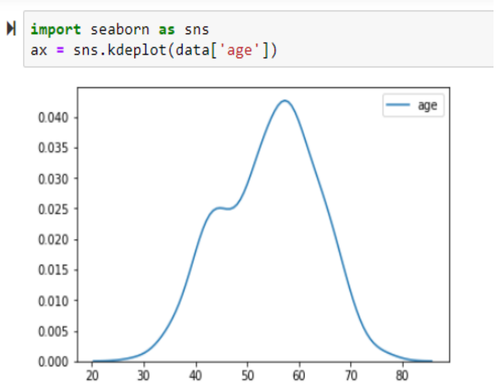


```
In [8]:  elements,edges = np.histogram(data['age'])
         print("The bin edges are:",edges)
         print("The number of elements in each bin are:",elements)

         The bin edges are: [29.  33.8 38.6 43.4 48.2 53.  57.8 62.6 67.4 72.2 77. ]
         The number of elements in each bin are: [ 1 11 33 38 37 60 63 43 14  3]

In [9]:  import matplotlib.pyplot as plt
         plt.hist(data['age'])

Out[9]:  (array([ 1., 11., 33., 38., 37., 60., 63., 43., 14.,  3.]),
          array([29. , 33.8, 38.6, 43.4, 48.2, 53. , 57.8, 62.6, 67.4, 72.2, 77. ]),
          <a list of 10 Patch objects>)
```

- The histograms are plotted such that the bins are of equal width, the number of bins are of user defined.
- In histograms the empty bins are also considered and the bars are contiguous.
- Location and variability are the first and second moments of statistics.
- The third moment of statistics are skewness which refers to whether the data is skewed to large,small values.
- The fourth moment of statistics is the kurtosis which is defined as the propensity of the extreme values.
- These two moments don’t have any units and can be understood by using visualization techniques.

Density Estimates:

- Density plot show the distribution of the data in a contiguous line. It can be treated as smoothed histogram.
- It is also known as Kernal Density Estimation (KDE).
- The python implementation for the same *age* feature can be shown as



```
import seaborn as sns
ax = sns.kdeplot(data['age'])
```