- **Mean, Variance and Standard Deviation**
    - Mean is average of a given set of data.

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

    - The Variance is average of the squared differences from the Mean. Variance tells you the degree of spread in your data set. The more spread the data, the larger the variance is in relation to the mean.

$$population\ var = \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

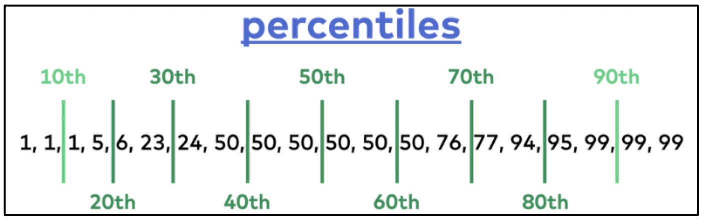
$$sample\ var = s^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N - 1}$$

    - The Standard Deviation is a measure of how spread out numbers are. It is square root of variance.

$$population\ std\ dev = \sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

$$sample\ std\ dev = \ s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N - 1}}$$

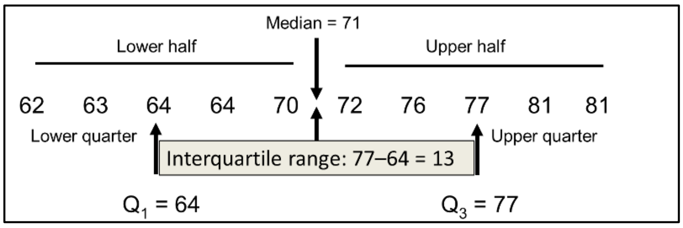
    - With samples, we use $N - 1$ in the formula because using N would give us a biased estimate that consistently underestimates variability. The sample variance would tend to be lower than the real variance of the population.
- **Percentile**
    - If we sort a list containing 100 elements in an ascending order. The n^{th} index specifies n^{th} percentile.
    - A n^{th} percentile specifies that $n\%$ of total dataset are less than n.
    - 25th, 50th, 75th & 100th elements are called as quantiles.
    - E.g., example of amazon (95th and 99th percentile)



- **MAD (Median Absolute Deviation)**
    - 1st we will calculate absolute distance or difference between X_i value and median (basically a median deviation)
    - Then we will find median of all these values
    - This method gives better findings of data over mean and standard deviation as outliers are irrelevant

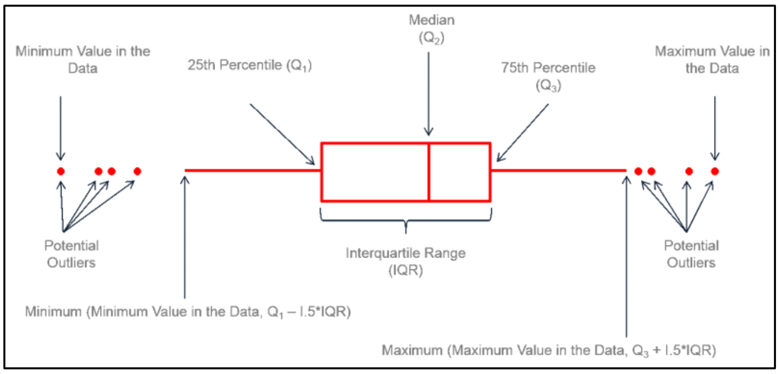
$$MAD = Median(|x_i - \bar{x}|)$$

- **IQR (Inter Quartile Range)**
    - This is a difference between 75th and 25th percentile

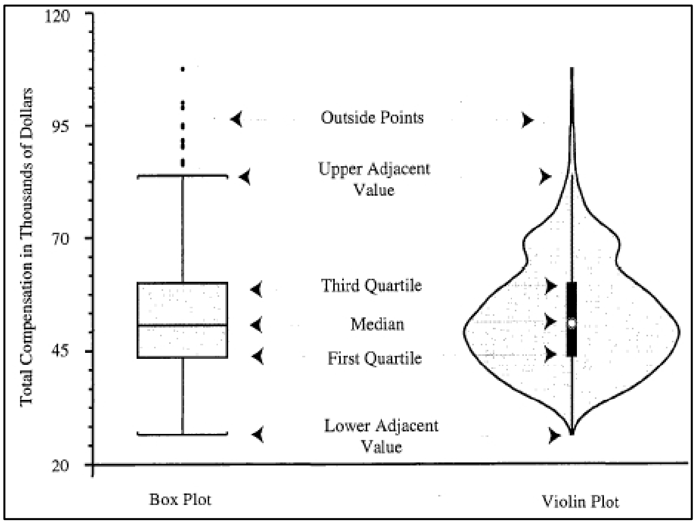


    - An interquartile range is a measure of where the most of the values lie.
    - The IQR is used to build box plots, simple graphical representations of a probability distribution.
    - The IQR is used in businesses as a marker for their income rates.
    - For a symmetric distribution (where the median equals the mid hinge, the average of the first and third quartiles), half the IQR equals the median absolute deviation (MAD).
    - The median is the corresponding measure of central tendency.
    - It can be used to identify outliers
    - The quartile deviation or semi-interquartile range is defined as half the IQR.

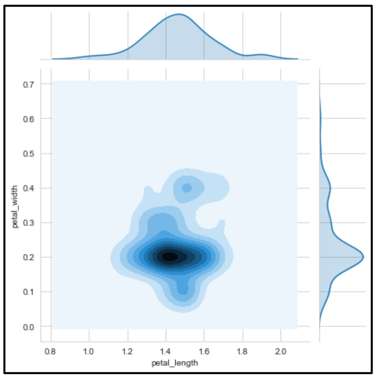
- **Box Plots and whiskers**


Minimum Value in the Data
25th Percentile (Q_1)
Median (Q_2)
75th Percentile (Q_3)
Maximum Value in the Data
Potential Outliers
Interquartile Range (IQR)
Potential Outliers
Minimum (Minimum Value in the Data, Q_1 – 1.5*IQR)
Maximum (Maximum Value in the Data, Q_3 + 1.5*IQR)

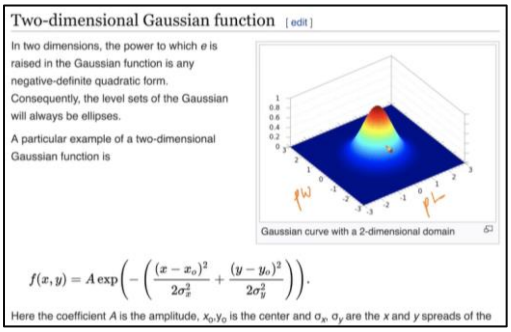
  - Box plots tell us about the skewness of data. If the median cuts it into two equal halves, then we can say that the data is not skewed whereas if they are of unequal size then the data can be positively/negatively skewed.
- **Violin Plots**
  - The curve shows probability density of corresponding data



- **Summarizing Plots, Univariate, Bivariate and Multivariate analysis**
  - Explain your findings/conclusions in plain English to interpret data to other people.
  - Never forget your objective (the problem you are solving). Perform all your EDA aligned with your objectives.
- **Univariate, bivariate and multivariate analysis.**
  - Univariate- depicting data based on single variable. E.g., PDF, CDF, Box Plots, Violin Plots
  - Bivariate- depicting data based on two variables E.g., Pair Plots, Scatter Plots
  - Multivariate- depicting data based on more than 2 variables E.g., 3D Plots
- **Multivariate Analysis**



- This are 2D Density plot.
- In this graph, we can observe 1D Plots for petal width and petal length
- Interpretation of these 2 graphs is a contour plot
- The darker region it gets, it shows us the corresponding probability for particular value.



Two-dimensional Gaussian function [edit]

In two dimensions, the power to which *e* is raised in the Gaussian function is any negative-definite quadratic form. Consequently, the level sets of the Gaussian will always be ellipses.

A particular example of a two-dimensional Gaussian function is

Gaussian curve with a 2-dimensional domain

$$f(x,y) = A\exp\left(-\left(\frac{(x-x_o)^2}{2\sigma_x^2} + \frac{(y-y_o)^2}{2\sigma_y^2}\right)\right).$$

Here the coefficient *A* is the amplitude, x_o,y_o is the center and σ_x, σ_y are the *x* and *y* spreads of the