**Chapter 1. Exploratory Data Analysis**

**Elements of Structured Data**

Data types:

1. Numeric: continuous & discrete

2. Categorical

3. Ordinal: categorical data that has orders

**Estimates of Location**

Mean, Trimmed Mean and Weighted Mean:

Mean is sensitive to outliers, trimmed mean eliminates the influence of extreme values because it drops a fixed number of sorted values at each end and then take an average of remaining values.

Weighted mean is calculated by multiplying each data value by a weight and dividing their sum by the sum of the weights. Advantages: Some values are intrinsically more variable than others, and highly variable observations are given a lower weight; The data collected does not equally represent the different groups that we are interested in measuring. To correct that, we can give a higher weight to the values from the groups that were underrepresented.

Median

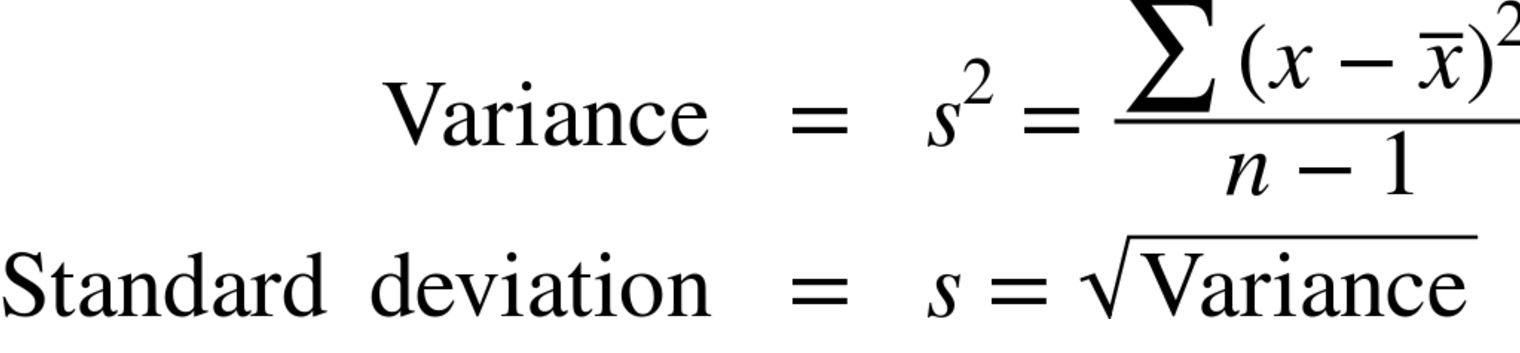
The *median* is the middle number on a sorted list of the data. It is referred to as a *robust* estimate of location since it is not influenced by *outliers* (extreme cases) that could skew the results.

**Estimates of Variability**

Variability measures if the data values are tightly clustered or spread out.

Deviation: The difference between the observed values and the estimated value.





**Order statistics**

Range: the difference between the largest and smallest number.

Percentile: Sorted the value and get the percentage values from the smallest value.

**Exploring the Data Distribution**

Percentiles and boxplot

Frequency table and Histogram

Density Plot

**Correlation**

Range from [-1, 1]. If the association is not linear, correlation coefficient may not be a useful matric.

**Exploring Two or More Variables**

Hexagonal binning

Contingency table

Boxplot