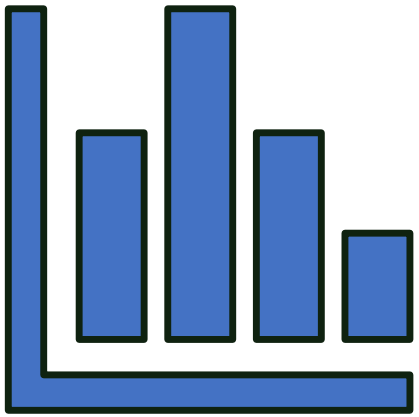
Probability & Statistics

Lecture 03 – Descriptive Statistics

Content



- Descriptive Statistics
  - Numerical Methods
    - Measures of Central Tendency
    - Measures of Dispersion
    - Measures of Range
    - Measures of Shape
  - Graphical Methods

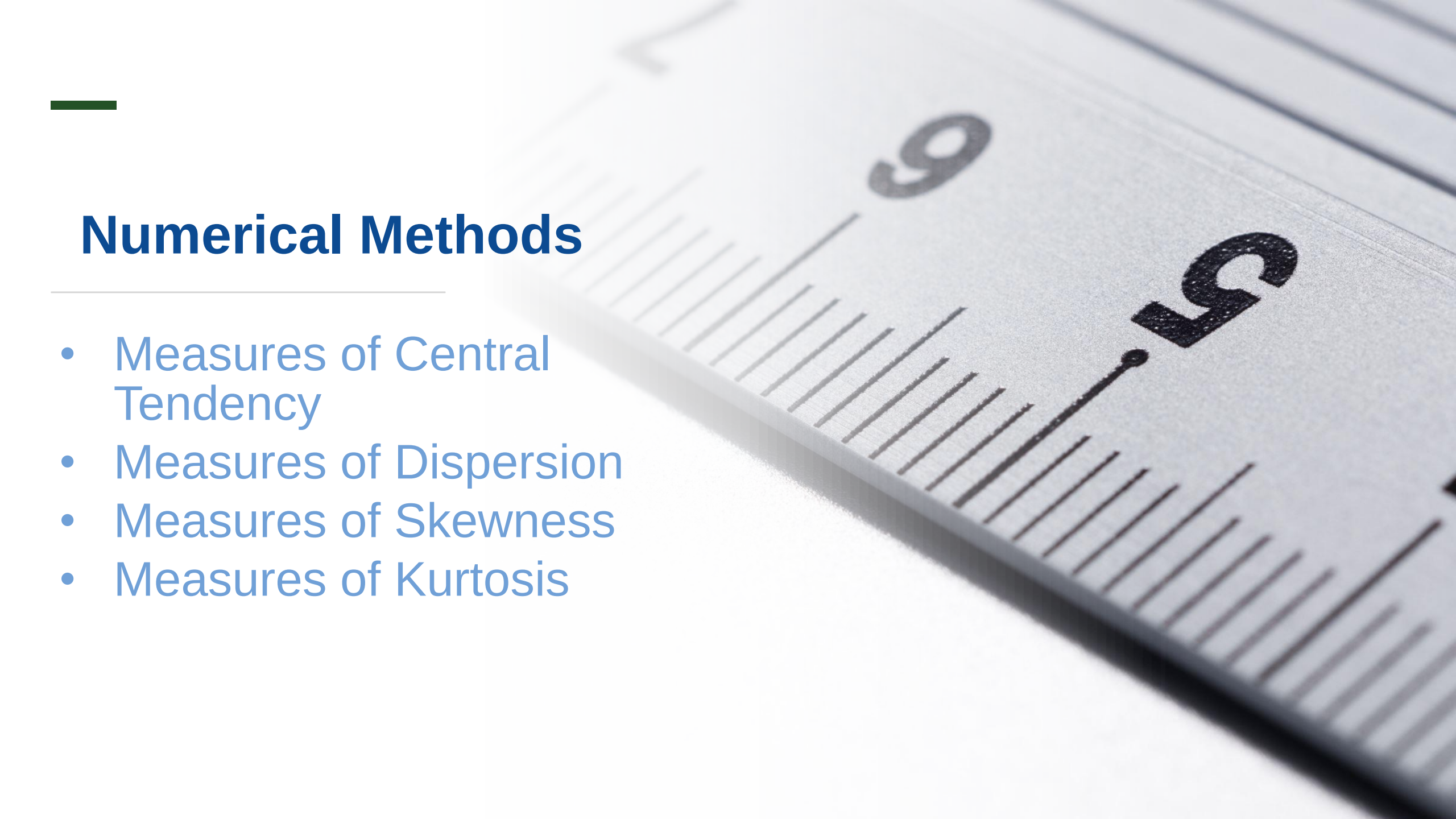
Descriptive Statistics

- Also known as preliminary analysis.
- Describes the behavior of the variables.
- Variables can be described in two ways.
  - Graphical Methods
  - Numerical Methods
- Each method depends on the type of the variable

NUMERICAL METHODS

Numerical Methods

- Measures of Central Tendency
- Measures of Dispersion
- Measures of Skewness
- Measures of Kurtosis



MEASURES OF CENTRAL TENDENCY

Measures of Central Tendency

- Mean
- Median
- Mode

Other Location Measurements

- Percentiles
- Deciles
- Quartiles

Mean

- Mean of n elements $x_1, x_2, \ldots x_n$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample Mean - $\bar{x}$
Population Mean - μ

Example

Example 1.2 (revisited):

Find the median "marks for FCS" of each student at SLIIT Metro.

78	74	82	66	91	71	64	88	55	80
51	74	82	75	16	78	84	79	71	83

Example 1.3 (revisited):

A load of aluminum sheets were purchased to construct a temporary shed. Twenty such sheets were examined for surface flaws. Find the median number of flaws in a sheet.

Number of flaws	Frequency
0	4
1	3
2	5
3	2
4	4
5	1
6	1

Mode

- Most frequent value
- There can be multiple modes in a data set.
- If all the data values are different, the data set has no mode.

Quartiles

- Divides the entire set of values into 4 equal sections.

$$Position\ of\ Q_1 = \frac{1}{4} \times (n+1)$$

$$Position\ of\ Q_2 = \frac{2}{4} \times (n+1)$$

$$Position\ of\ Q_3 = \frac{3}{4} \times (n+1)$$

- Q2 is also called as the median.

Deciles

- Divides the entire set of values into 10 equal sections.

$$Position\ of\ D_1 = \frac{1}{10} \times (n+1)$$

$$Position\ of\ D_2 = \frac{2}{10} \times (n+1)$$

Percentiles

- Divides the entire set of values into 100 equal sections.

$$Position\ of\ P_1 = \frac{1}{100} \times (n+1)$$

$$Position\ of\ P_2 = \frac{2}{100} \times (n+1)$$

MEASURES OF DISPERSION

Measures of Dispersion

- Variance
- Standard Deviation
- Range
- Inter Quartile Range

Variance

- Describes how the data has dispersed around its mean.
- Not sensitive to outliers.(more robust for outliers).

Variance

Population Variance

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

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

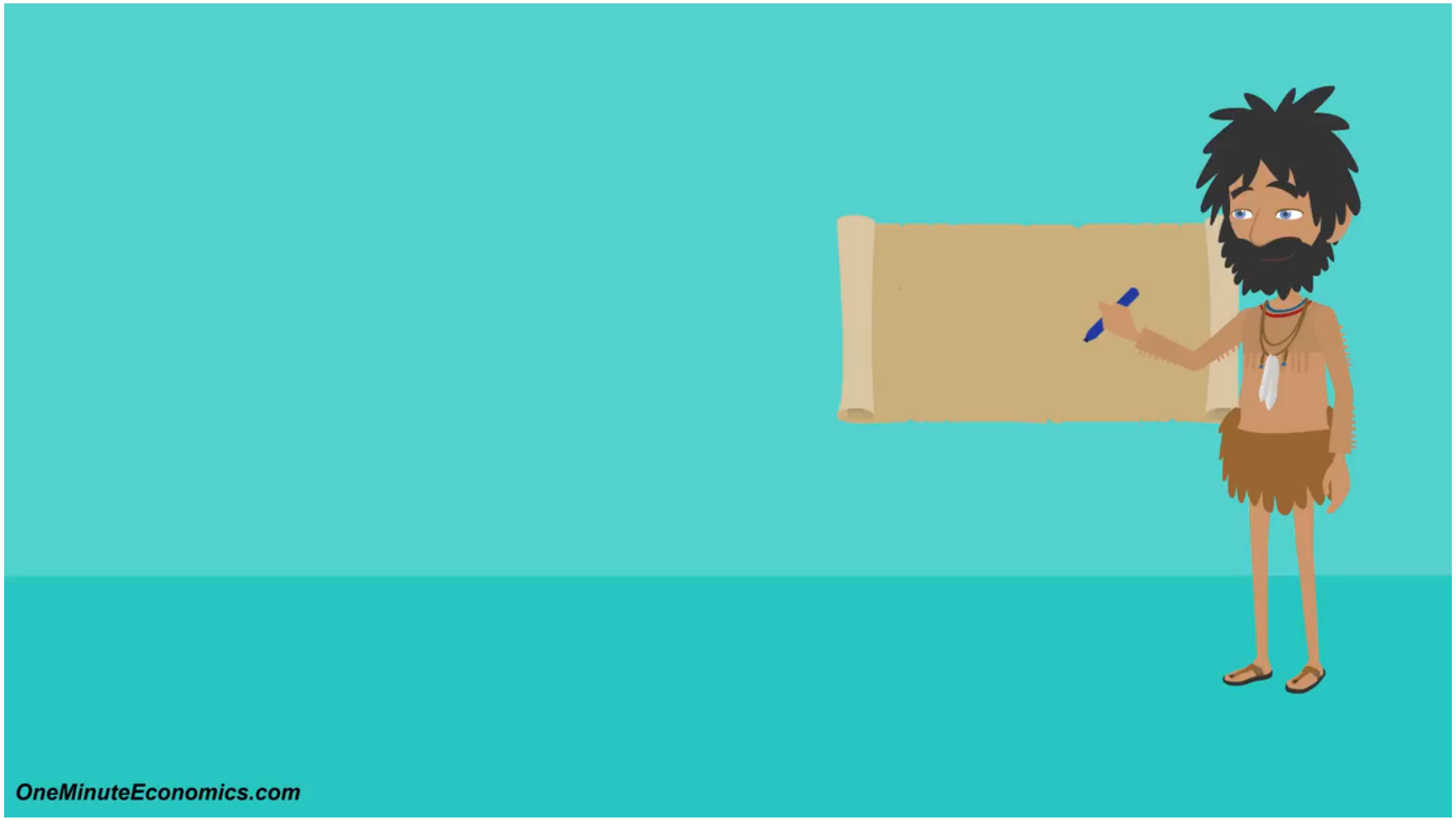
Standard Deviation

- Square root of the variance.

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

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Standard Deviation



Range

- Range = Maximum – Minimum

Inter Quartile Range (IQR)

- IQR = Q3 - Q1

MEASURES OF SKEWNESS

Skewness

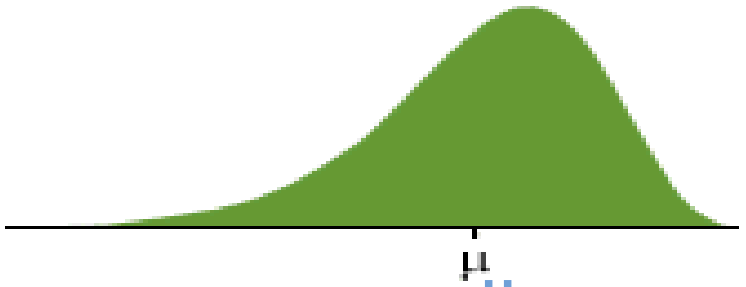
- Skewness is a measure of symmetry of a distribution.
  - Symmetric Distribution
  - Negatively Skewed Distribution
  - Positively Skewed Distribution

$$Skewness = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)s^3}$$

Skewness

Negatively Skewed Distributions

- The left tail is longer
- Mass of the distribution is concentrated on the right.
- The distribution is said to be left-skewed



mean < median < mode

Positively Skewed Distribution

- The right tail is longer.
- Mass of the distribution is concentrated on the left.
- The distribution is said to be right-skewed

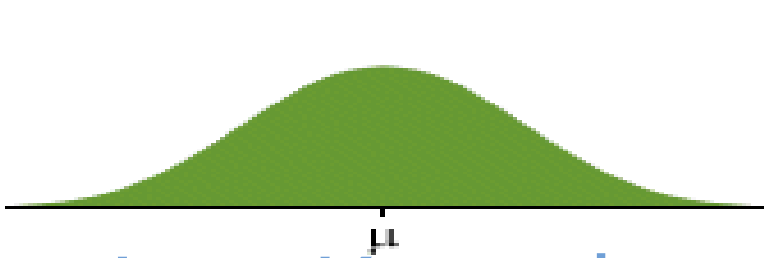
mode < median < mean

MEASURES OF KURTOSIS

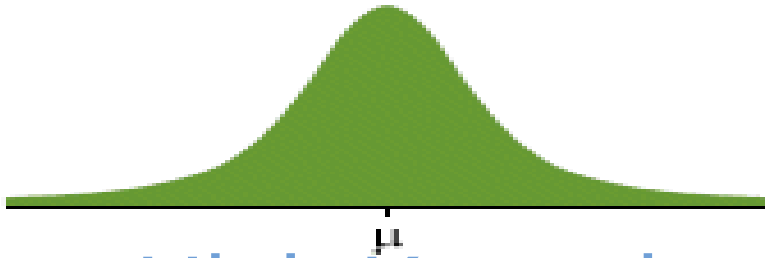
Kurtosis

- It is a measure of the tallness or flatness ("peaked ness") of the distribution.
- It is a measure of whether the data are peaked or flat relative to a normal distribution.

$$Kurtosos = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)s^4}$$



Low Kurtosis



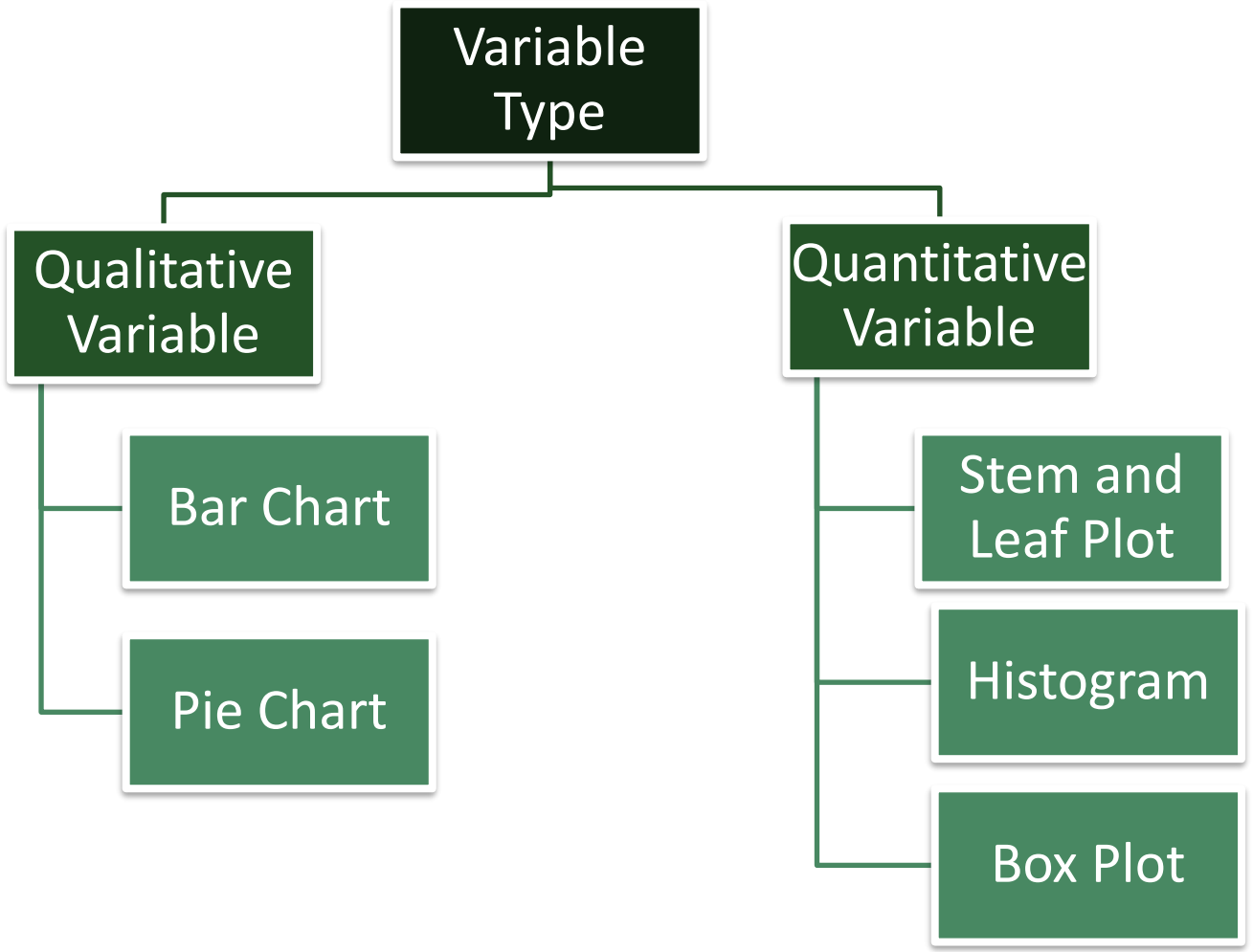
High Kurtosis

GRAPHICAL METHODS

Graphical Methods

- Can be used to analyze both categorical and numerical variables.
- Type of the graph, depends on the type of the variable

Graphical Methods



Graphical Methods

BAR CHARTS

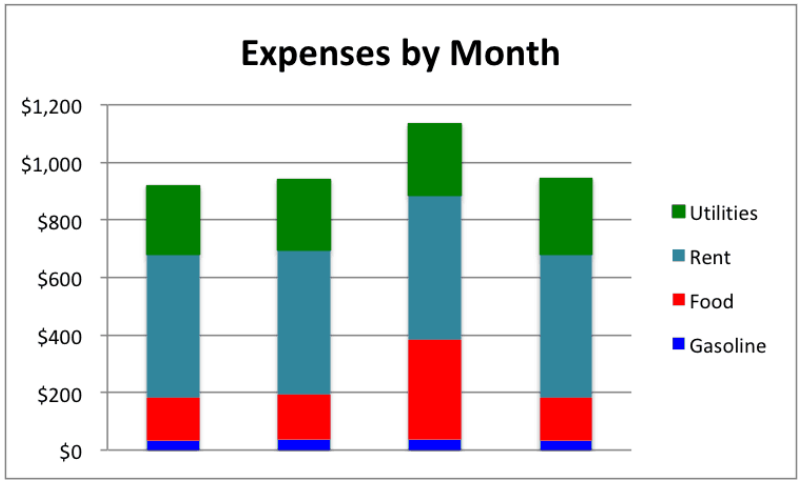
Bar Charts

- Bars can be drawn vertically or horizontally.
- Y axis can represent Frequency, cumulative frequency or percentages
- X axis represents the categorical variable.

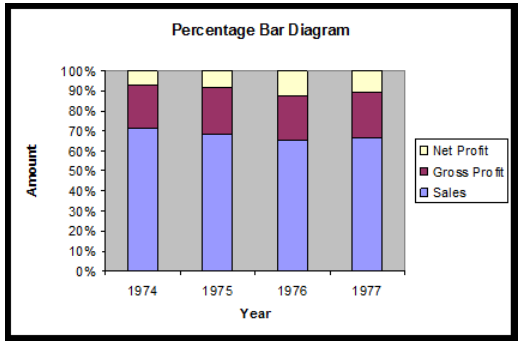
Bar Charts

- Different Types of bar charts
  - Simple Bar Charts
  - Component Bar Charts/Stacked Bar Charts
  - Percentage Component Bar Charts
  - Multiple Bar Charts

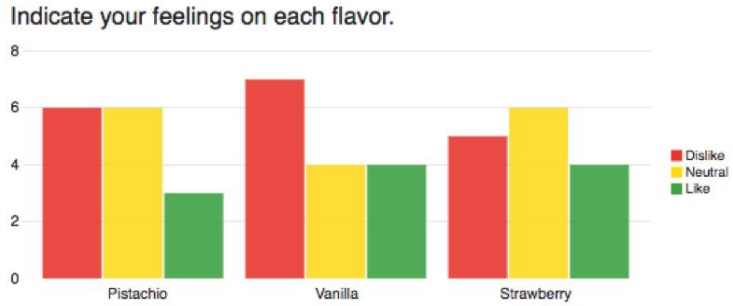
Examples



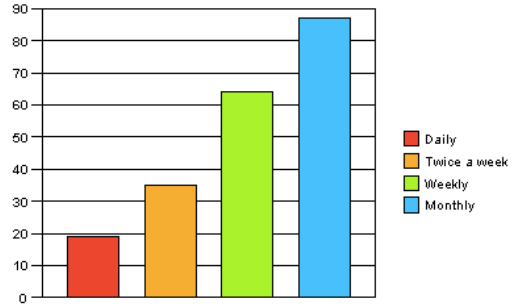
Stacked Bar Chart



Percentage Component Bar Chart



Multiple Bar Chart

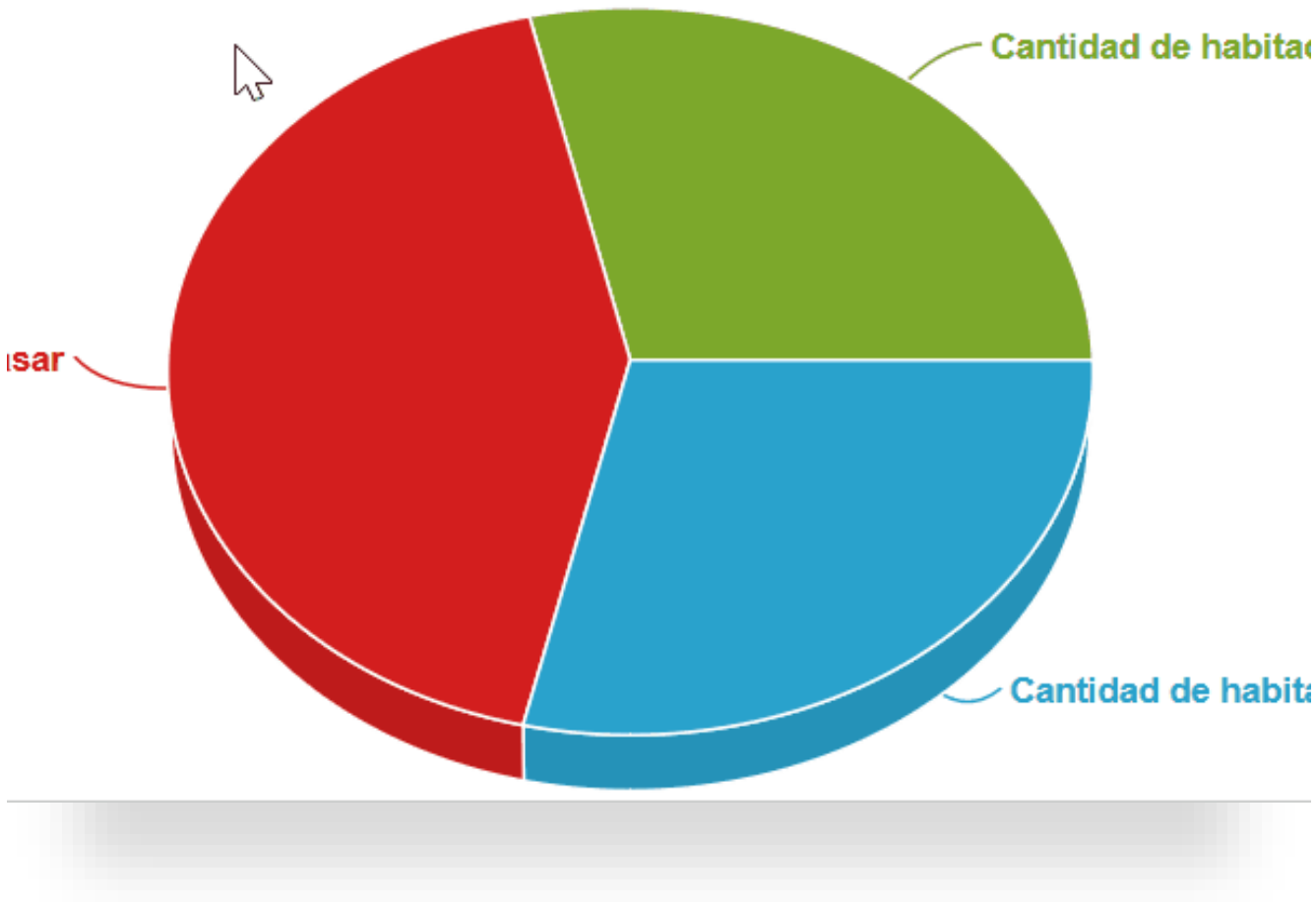


Simple Bar Chart

PIE CHARTS

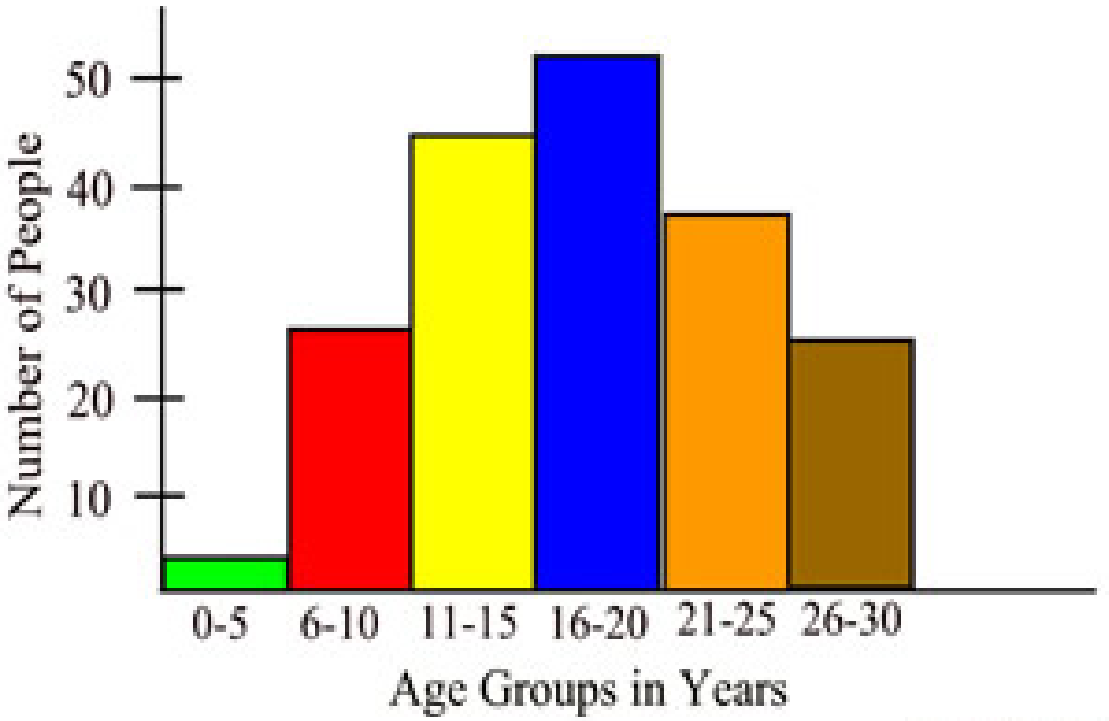
Pie Charts

- Use to analyze one categorical variable
- Area of each sector is proportional to the value of the category
- Appropriate, when there are **few number of categories** or when **value of each category is varying widely**



Example

HISTOGRAMS


Number of People
50
40
30
20
10
0-5
6-10
11-15
16-20
21-25
26-30
Age Groups in Years

Histograms

- Divide the given data set into suitable number of classes (intervals/categories) which have the same width.
- Frequency, relative frequency or percentages can be used for the y axis while x axis will represent the classes of the variable.
- Classes with their frequencies (counts) is called a frequency distribution.
- Bar represents a class and length of the bar is proportional to the frequency of respective class.
- Bars are adjacent to each other (No gaps between two bars)

Example

- Draw the histogram for the following data set.

78	74	82	66	91	71	64	88	55	80
51	74	82	75	16	78	84	79	71	83

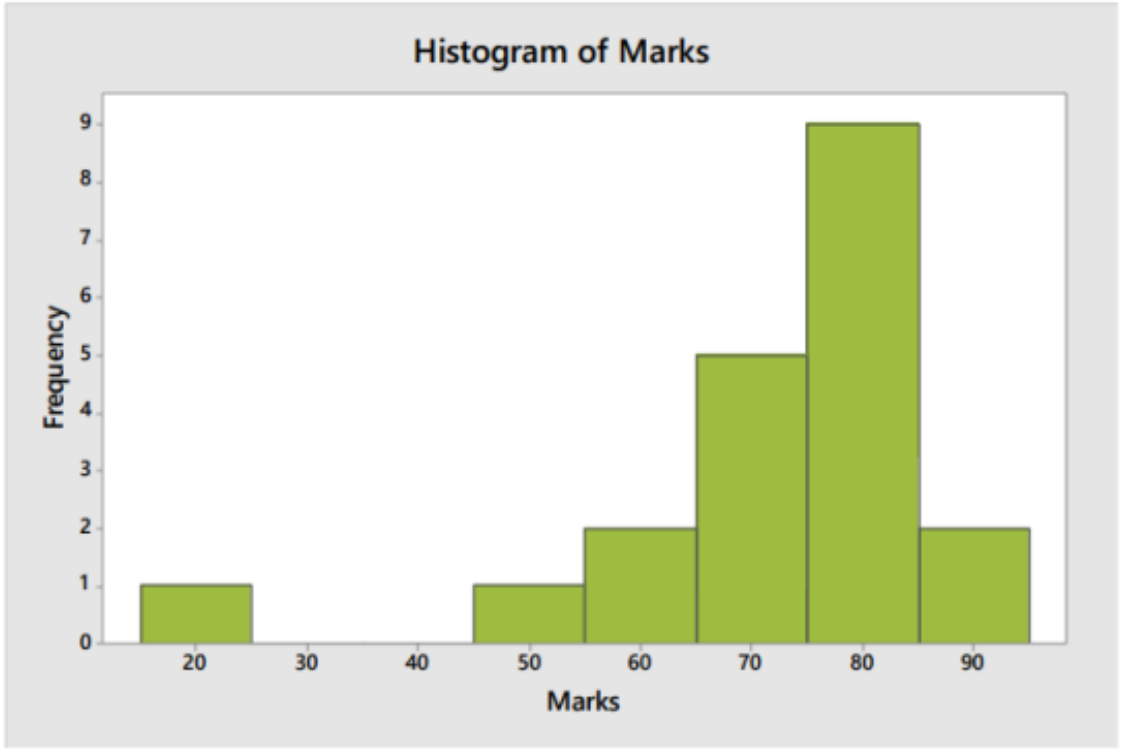
$$Range = Maximum - Minimum = 91 - 16 = 75$$

$$Class\ Width = \frac{Range}{Number\ of\ classes}$$

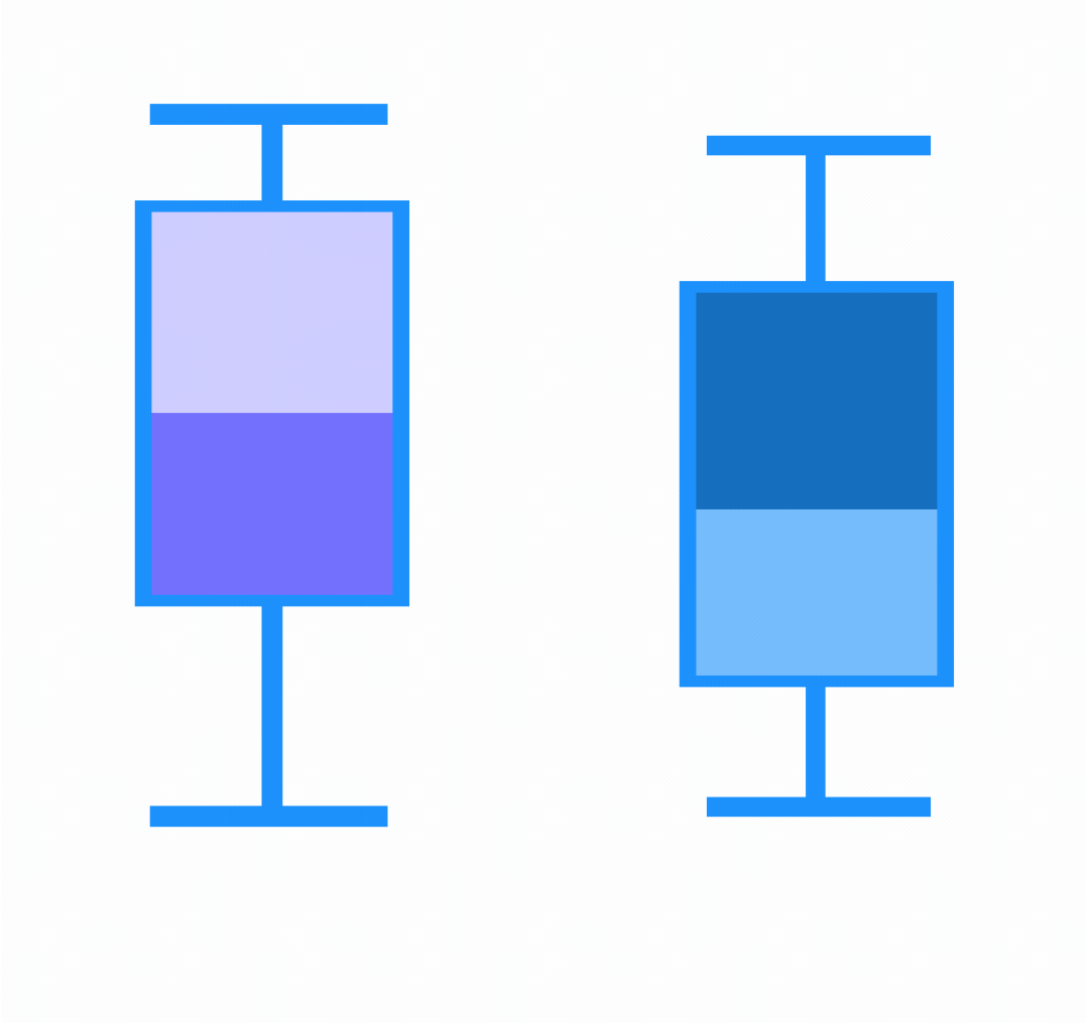
- Classes can be selected by fixing the class width also

Histogram Example

Class	Frequency
14.5 – 24.5	1
24.5 – 34.5	0
34.5 – 44.5	0
44.5 – 54.5	1
54.5 – 64.5	2
64.5 – 74.5	5
74.5 - 84.5	9
84.5 – 94.5	2



BOX PLOT



Box Plot

- Five Number Summary is used to draw a box plot.
- Five Number Summary includes:
  - Minimum
  - Q1 (First Quartile)
  - Q2 (Second Quartile)
  - Q3 (Third Quartile)
  - Maximum

Example

- Draw the boxplot for the following data set.

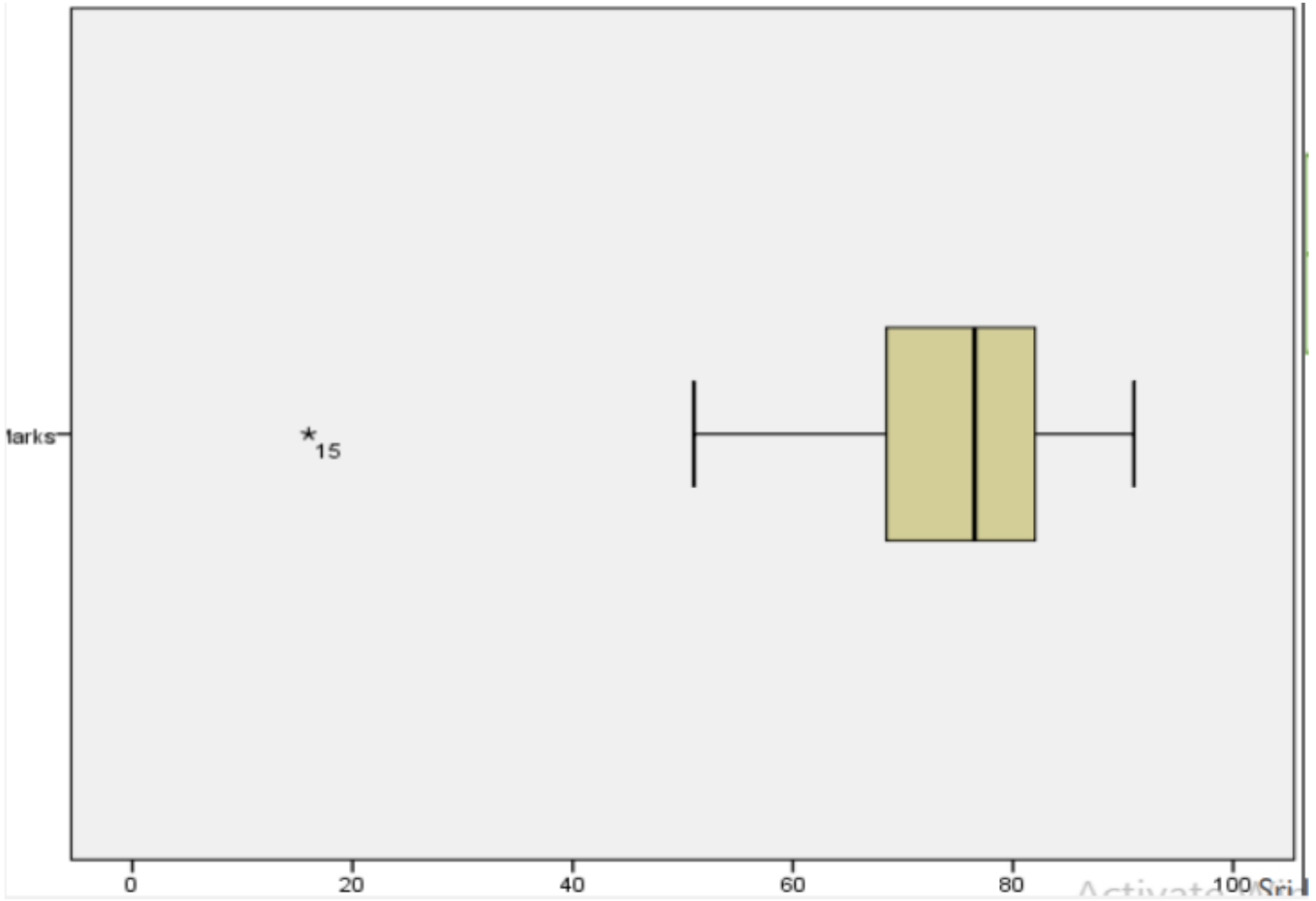
78	74	82	66	91	71	64	88	55	80
51	74	82	75	16	78	84	79	71	83

$$Q_1 = ?$$
$$Q_2 = ?$$
$$Q_3 = ?$$

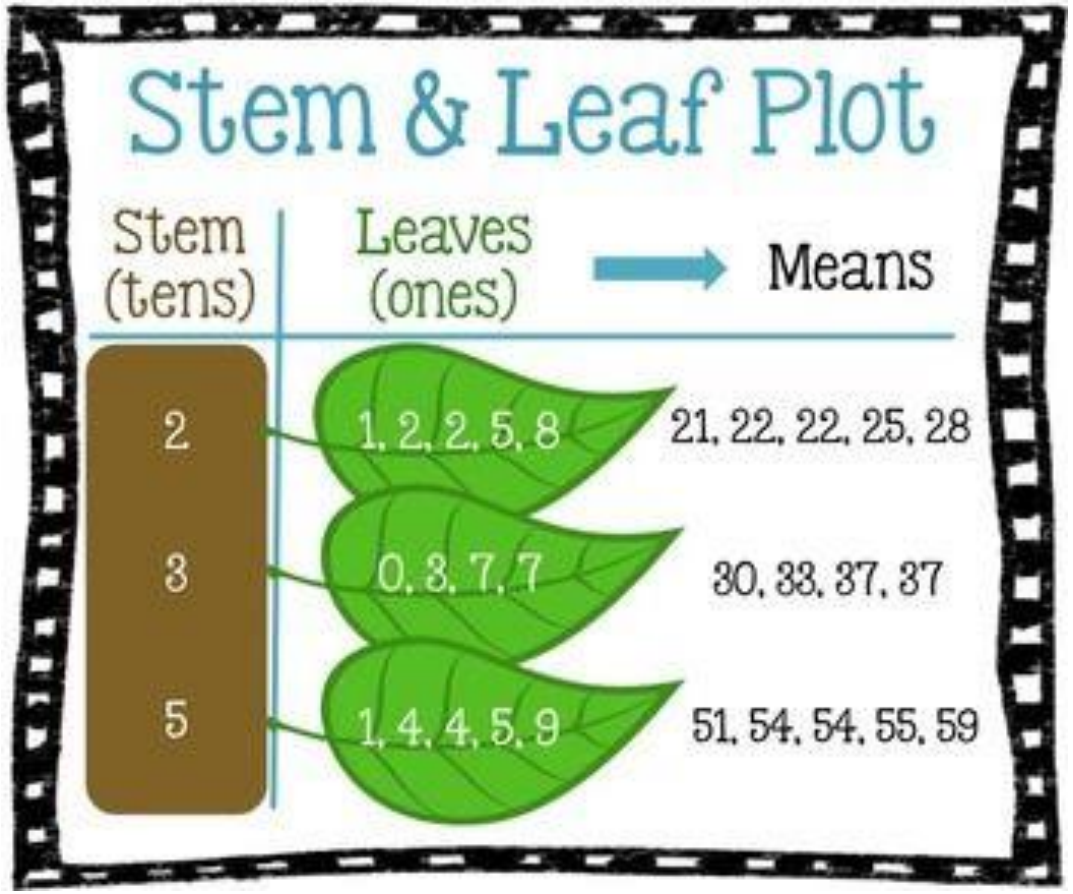
Outliers

- Outliers are the data points (x_i) satisfy the following.
  - $x_i > Q_3 + 1.5 \times IQR$
  - $x_i < Q_1 - 1.5 \times IQR$

Box Plot

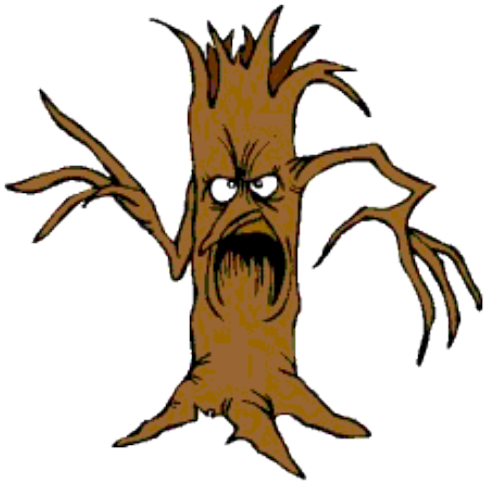


STEM AND LEAF PLOT



Stem and Leaf Plot

- Useful when the data set is very small.
- First the data set is sorted in ascending order.
- Then, each data value is split into two parts known as stem and leaf.
- The “leaf” is usually the last digit of the number.
- The other digits to the left of the “leaf” form the “stem”.

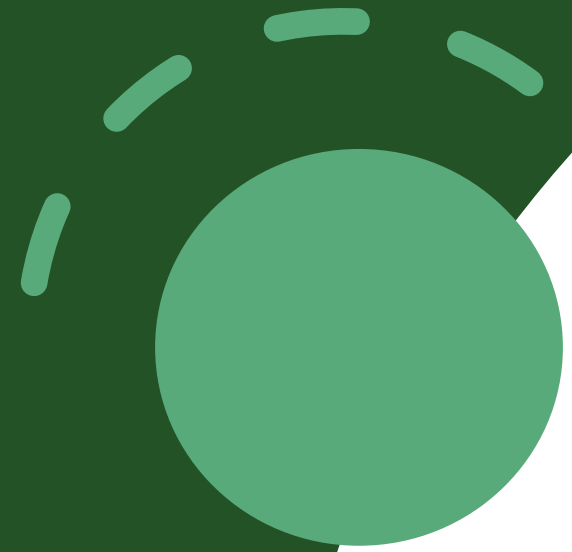


You did WHAT
with my leaves???

Stem	Leaves
1	6
2	
3	
4	
5	1 5
6	4 6
7	1 1 4 4 5 8 8 9
8	0 2 2 3 4 8
9	1

Key: 1|6 → 16

Stem and Leaf Plot



Thank You

Questions?