

# 307102

## Descriptive Statistics for Business

### Part 1: Descriptive Statistics

# What is Statistics?

Statistics is a mathematical science that includes methods for collecting, organizing, analyzing and visualizing **data** to extract useful **information** that can be used to **make decisions**.

Statistics help us solve problems, answer questions and make decisions based on data (Data Driven Decisions).

Statistics Helps to in Reduce Uncertainty.

Statistics is the science of making decisions under uncertainty (Savage, The Foundations of Statistics, 1954).



# Statistics Branches

Statistics is generally divided into two main branches:

## 1. Descriptive statistics.

Descriptive statistics refers to the summarizing and explaining data using numbers, tables and charts.

## 2- Inferential statistics.

Inferential statistics refers to drawing conclusions about a population based on a sample data.

A population refers to all members of a specified group (not necessarily people), whereas a sample is a subset of that population.

Population: the collection of all items we are interested in (people, things, plants, stars, trees...etc.), for the purpose of making our decision.

# Example of Useful Descriptive Statistics

## 1. Business Sales:

- A store owner calculates the average daily sales to get a sense about his income.
- The owner also computes the standard deviation of daily sales to understand the variability or consistency in income.

## 2. Classroom Grades:

- A teacher calculates the average grade of a test to understand the overall performance of the class.
- The teacher also determines the highest and lowest scores to understand the spread of the grades.

## 3. Survey Results:

- After a customer satisfaction survey, a company finds that 70% of respondents are "satisfied" or "very satisfied."
- The survey also indicates that the mode response to the question, "What's the most important feature?" is "product durability."

# Examples of Useful Inferential Statistics

## 1. Medicine:

- A research group tests a new drug on 100 patients and finds that 60 of them show significant improvement.
- Using inferential statistics, they can make conclusions about the drug's effectiveness on the entire population.

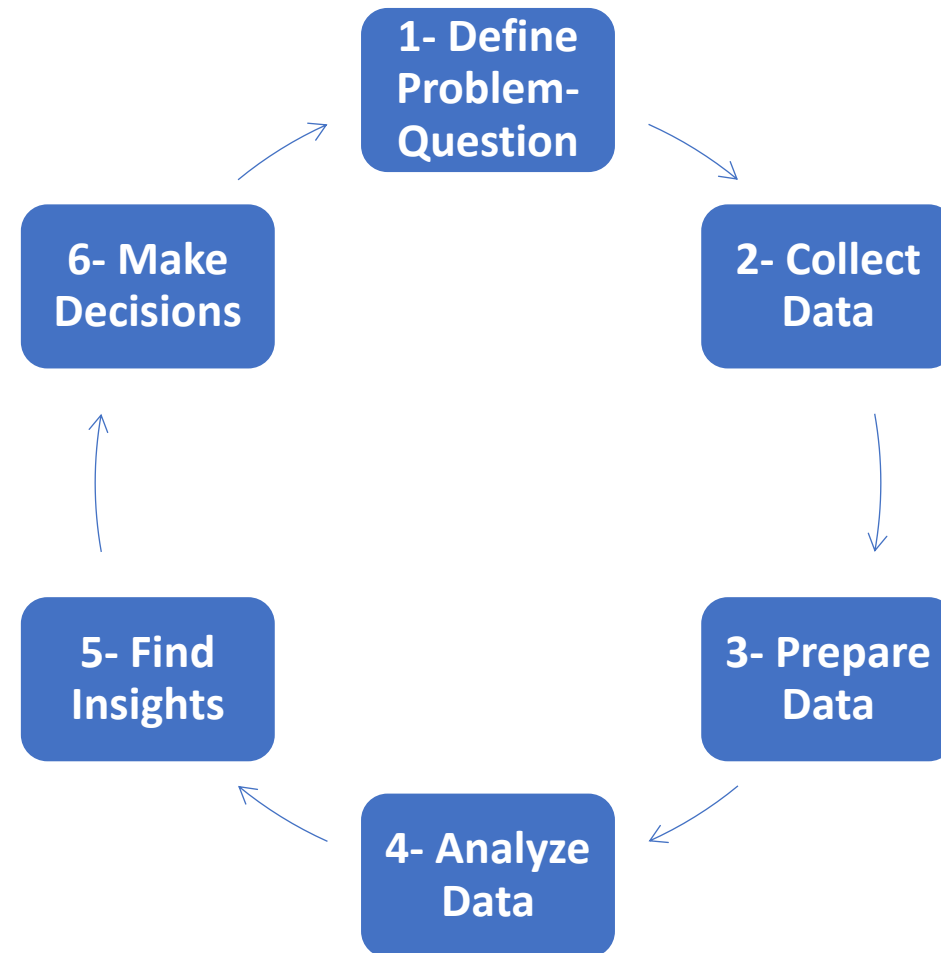
## 2. Agriculture:

- A farmer tests a new fertilizer on a sample plot of land to see if it increases crop yield.
- Inferential statistics can help predict the impact of using the fertilizer on all their plots.

## 3. Marketing:

- A company launches a new ad campaign in a small geographic area.
- They use inferential statistics to determine if rolling out the campaign nationwide would likely produce similar results.

# Statistical Process Life Cycle



# Types of Data & Levels of Measurements

# Categorical Data vs Numerical Data

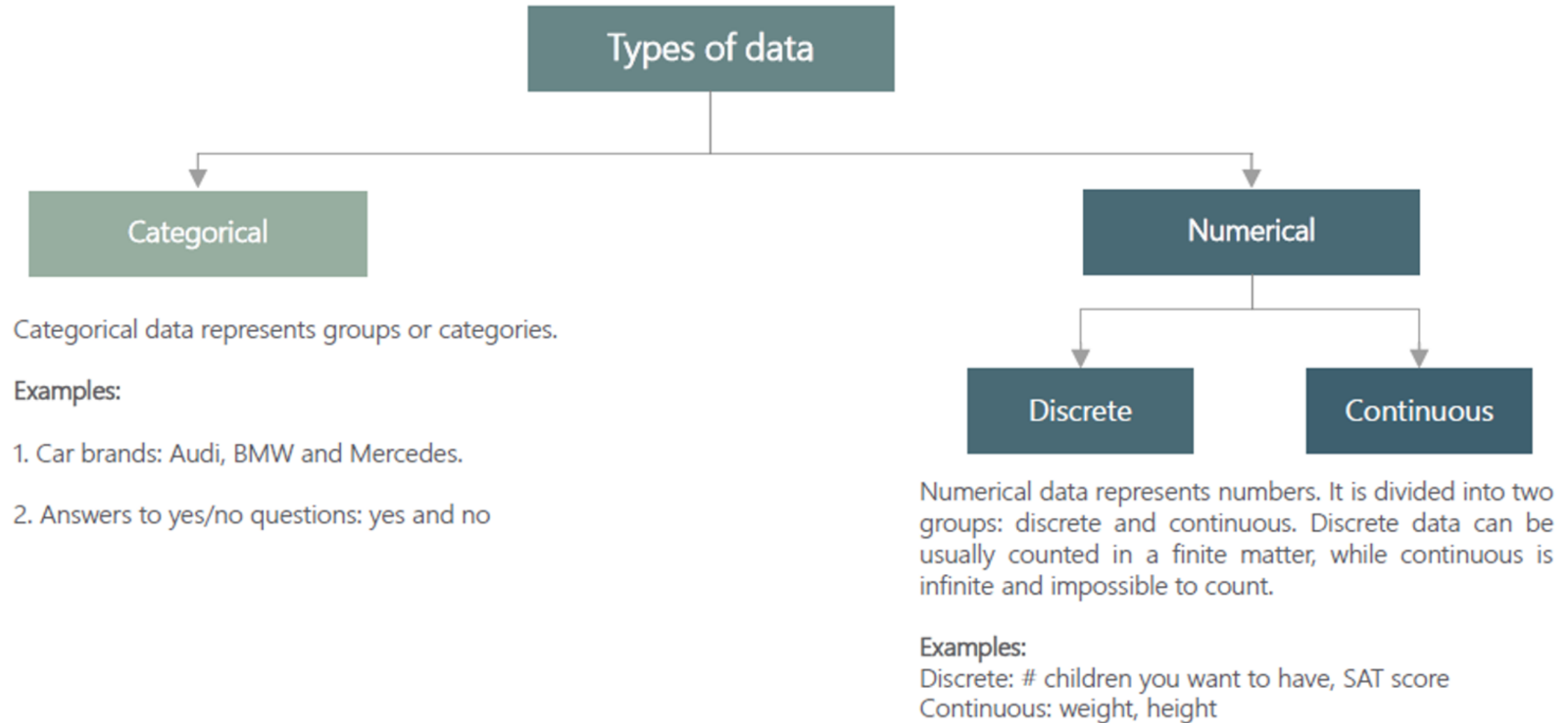
## 1. **Qualitative** or **categorical** (non-numerical) data:

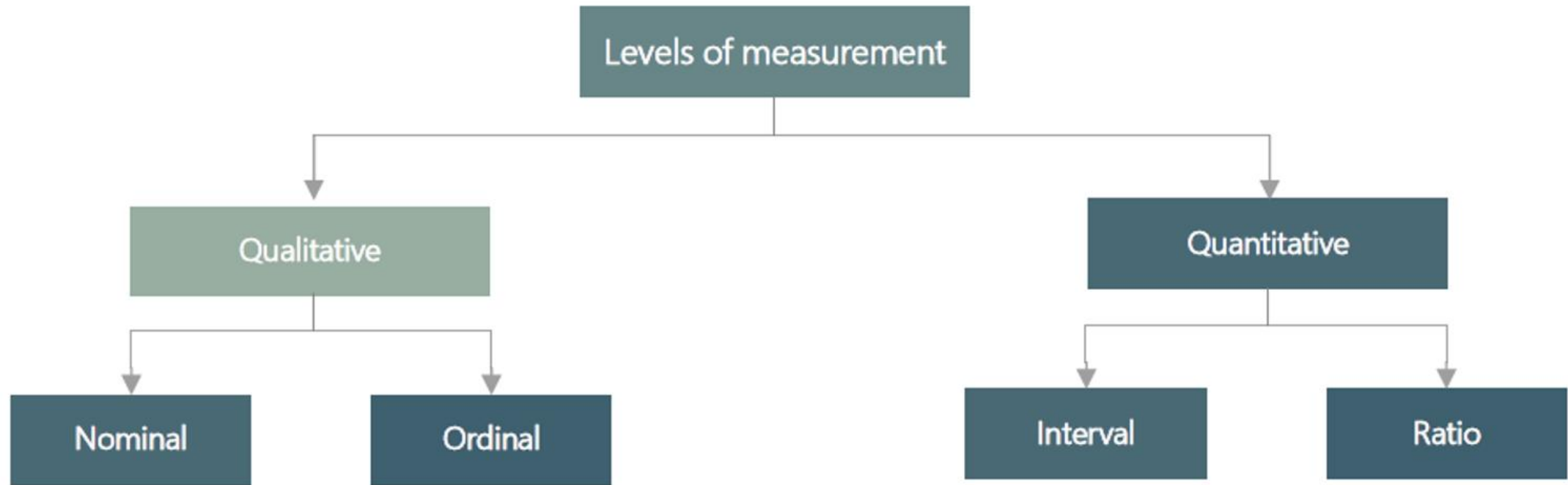
Consists of descriptive information, such as colors, tastes, textures, or any other characteristics that **can be counted**.

## 2. **Quantitative** or **numerical** data:

Includes data that can be **measured** and can be represented numerically, such as height, weight, or temperature.







There are two qualitative levels: nominal and ordinal. The nominal level represents categories that cannot be put in any order, while ordinal represents categories that can be ordered.

**Examples:**

Nominal: four seasons (winter, spring, summer, autumn)

Ordinal: rating your meal (disgusting, unappetizing, neutral, tasty, and delicious)

There are two quantitative levels: interval and ratio. They both represent "numbers", however, ratios **have a true zero**, while intervals don't.

**Examples:**

Interval: degrees Celsius and Fahrenheit

Ratio: degrees Kelvin, length

# Why Types of Data and Levels of Measurements Matter?

1. **Data Visualization**: e.g. Categorical data might be best represented in a bar chart, while continuous data might be better suited for a histogram or scatter plot.
2. **Data Transformation**: e.g. Ordinal data might be converted into interval data under certain conditions, or continuous data might be categorized into ordinal data.
3. **Data Quality**: e.g. If you expect a variable to be discrete and find continuous values, this could indicate a data quality issue.
4. **Interpretation of Results**: e.g. If you have ordinal data, you can make statements about the order of values but not the difference between values.
5. **Appropriate Analysis**: e.g. Nominal data can be analyzed using a Chi-square test, while interval data can be analyzed using a t-test or ANOVA. Using the wrong test can lead to incorrect conclusions.

| Type of attribute    | Qualitative / Categorical          |                                 | Quantitative / Metric                      |   |
|----------------------|------------------------------------|---------------------------------|--|---|
| Scale of measurement | Nominal scale                      | Ordinal scale                   | Interval scale                             | Ratio scale   |
| Examples             | Gender, blood group, rhesus factor | Grades, medical scores          | Temperature in °C, intelligence quotient   | Temperature in Kelvin, body height                  |
| Notes                | Lowest level                       | Order is defined                | Arbitrary zero, distance is defined        | Highest level, natural zero, ratio is defined       |
| Operations           | $A = B, A \neq B$                  | $A = B, A \neq B, A < B, A > B$ | $A = B, A \neq B, A < B, A > B, d = A - B$ | $A = B, A \neq B, A < B, A > B, d = A - B, r = A:B$ |

Figure 2.2: Types of attributes and scales of measurement.

# Descriptive Statistics Tools

# Statistical Tools Used in Descriptive Statistics

1. Tables (Data Tables, Frequency Distribution Tables).
2. Visualization & Charts.
3. Numerical Measures (Measures of Central Location, Dispersion, Shape, Position, Association).



## Tables in Descriptive Statistics

- Tables organize and display data in a structured format.
- They are essential for summarizing large data sets efficiently.

### Types of Tables in Statistics

1. **Frequency Distribution Tables**: Show how often each value in a set of data occurs.
2. **Summary Tables**: Provide key statistics like mean, median, mode, range, etc., at a glance.
3. **Contingency Tables/Cross-tabulation**: Display the multivariate frequency distribution for categorical data.

In the next section, we present tables examples using students dataset.

# Introducing Students Dataset

| #  | stud.id | name         | gender | age | height | weight | country | nc.score | semester | major        | minor        | score1 | score2 | av_score | online.tutorial | graduated | salary   |
|----|---------|--------------|--------|-----|--------|--------|---------|----------|----------|--------------|--------------|--------|--------|----------|-----------------|-----------|----------|
| 1  | 833917  | Gonzales, C  | Female | 19  | 160    | 64.8   | Turkey  | 1.91     | 1st      | Political Sc | Social Scie  | NA     | NA     | 0        | 0               | 0         | NA       |
| 2  | 898539  | Lozano, T    | Female | 19  | 172    | 73     | Other   | 1.56     | 2nd      | Social Scie  | Mathemati    | NA     | NA     | 0        | 0               | 0         | NA       |
| 3  | 379678  | Williams, F  | Female | 22  | 168    | 70.6   | UK      | 1.24     | 3rd      | Social Scie  | Mathemati    | 45     | 46     | 45.5     | 0               | 0         | NA       |
| 4  | 807564  | Nem, Deniz   | Male   | 19  | 183    | 79.7   | Other   | 1.37     | 2nd      | Environme    | Mathemati    | NA     | NA     | 0        | 0               | 0         | NA       |
| 5  | 383291  | Powell, He   | Female | 21  | 175    | 71.4   | Italy   | 1.46     | 1st      | Environme    | Mathemati    | NA     | NA     | 0        | 0               | 0         | NA       |
| 6  | 256074  | Perez, Jadr  | Male   | 19  | 189    | 85.8   | Italy   | 1.34     | 2nd      | Political Sc | Mathemati    | NA     | NA     | 0        | 0               | 0         | NA       |
| 7  | 754591  | Clardy, Ani  | Female | 21  | 156    | 65.9   | UK      | 1.11     | 2nd      | Political Sc | Social Scie  | NA     | NA     | 0        | 0               | 0         | NA       |
| 8  | 146494  | Allen, Rebe  | Female | 21  | 167    | 65.7   | Other   | 2.03     | 3rd      | Political Sc | Economics    | 58     | 62     | 60       | 0               | 0         | NA       |
| 9  | 723584  | Tracy, Rob   | Male   | 18  | 195    | 94.4   | Other   | 1.29     | 3rd      | Economics    | Environme    | 57     | 67     | 62       | 0               | 0         | NA       |
| 10 | 314281  | Nimmons, F   | Female | 18  | 165    | 66     | Russa   | 1.19     | 2nd      | Environme    | Mathemati    | NA     | NA     | 0        | 0               | 0         | NA       |
| 11 | 200803  | Lang, Mack   | Female | 22  | 162    | 66.8   | Other   | 1.04     | 4th      | Economics    | Environme    | 62     | 61     | 61.5     | 1               | 1         | 45254.11 |
| 12 | 444907  | Rodriguez, F | Female | 18  | 172    | 66.8   | Other   | 3.81     | 3rd      | Environme    | Economics    | 76     | 82     | 79       | 0               | 0         | NA       |
| 13 | 354271  | Covar Orer   | Male   | 23  | 185    | 84.6   | Russa   | 1        | 4th      | Environme    | Mathemati    | 71     | 76     | 73.5     | 1               | 1         | 40552.79 |
| 14 | 317812  | Lopez, Mor   | Female | 20  | 158    | 64.4   | Italy   | 2.5      | 6th      | Environme    | Social Scie  | 66     | 70     | 68       | 1               | 1         | 27007.03 |
| 15 | 604115  | Davis, Shağ  | Female | 19  | 157    | 66.3   | Russa   | 1.92     | 2nd      | Economics    | Political Sc | NA     | NA     | 0        | 0               | 0         | NA       |
| 16 | 889551  | Adams, Jos   | Male   | 20  | 172    | 73.9   | Other   | 3.61     | 4th      | Mathemati    | Political Sc | 87     | 91     | 89       | 1               | 0         | NA       |
| 17 | 350040  | Hines, Hail  | Female | 22  | 156    | 61.7   | Other   | 2.27     | 6th      | Political Sc | Biology      | 57     | 54     | 55.5     | 0               | 1         | 33969.16 |
| 18 | 240279  | Daugherty, M | Male   | 22  | 182    | 82.1   | Italy   | 1.42     | 1st      | Economics    | Environme    | NA     | NA     | 0        | 0               | 0         | NA       |
| 19 | 865835  | Roybal, Ebo  | Female | 21  | 162    | 69.2   | Italy   | 1.32     | 3rd      | Political Sc | Environme    | 69     | 46     | 57.5     | 1               | 0         | NA       |
| 20 | 137196  | Baysinger, F | Female | 22  | 168    | 70.9   | UK      | 2.33     | 2nd      | Environme    | Political Sc | NA     | NA     | 0        | 0               | 0         | NA       |
| 21 | 708242  | Phillips, La | Female | 20  | 167    | 68.5   | Other   | 1.79     | 4th      | Biology      | Economics    | 77     | 80     | 78.5     | 1               | 0         | NA       |
| 22 | 499002  | Culbertson   | Male   | 37  | 175    | 70.4   | UK      | 1.97     | 2nd      | Political Sc | Environme    | NA     | NA     | 0        | 0               | 0         | NA       |
| 23 | 873149  | O Reilly, Jo | Male   | 19  | 164    | 70.3   | UK      | 1.68     | 2nd      | Political Sc | Environme    | NA     | NA     | 0        | 0               | 0         | NA       |
| 24 | 807361  | Johnson, S   | Female | 38  | 155    | 67     | Italy   | 2.3      | 2nd      | Environme    | Biology      | NA     | NA     | 0        | 0               | 0         | NA       |



# Frequency Distribution Tables

Used to show the distribution of data across different categories or intervals.

Frequency Types:

1. Absolute Frequency: The actual count of occurrences for each category.
2. Relative Frequency: The proportion of the total count represented by each category, often expressed as a percentage.
3. Cumulative Frequency: A running total of frequencies through the categories, showing the number of observations up to a certain category.

| Row Labels ▼       | Number of Students | Relative Frequency | Cumulative Frequency |
|--------------------|--------------------|--------------------|----------------------|
| Italy              | 2797               | 33.95%             | 33.95%               |
| Other              | 2688               | 32.63%             | 66.57%               |
| UK                 | 1839               | 22.32%             | 88.89%               |
| Russa              | 585                | 7.10%              | 95.99%               |
| Turkey             | 330                | 4.01%              | 100.00%              |
| <b>Grand Total</b> | <b>8239</b>        | <b>100.00%</b>     |                      |

**Frequency of Students by Country**

# Converting Numerical Values into Categorical

| Row Labels         | Sum of age    |
|--------------------|---------------|
| <20                | 23415         |
| 20-29              | 144001        |
| 30-39              | 3785          |
| 40-49              | 5433          |
| 50-59              | 5947          |
| 60-70              | 3139          |
| <b>Grand Total</b> | <b>185720</b> |

Converting age into age groups

| Average Category | Description     | Number of Students |
|------------------|-----------------|--------------------|
| Undefined        | value undefined | 3347               |
| F                | < 50            | 587                |
| D                | >= 50 and < 60  | 767                |
| C                | >= 60 and < 70  | 931                |
| B                | >= 70 and < 80  | 1575               |
| A                | >= 80           | 1032               |
|                  | Total           | 8239               |

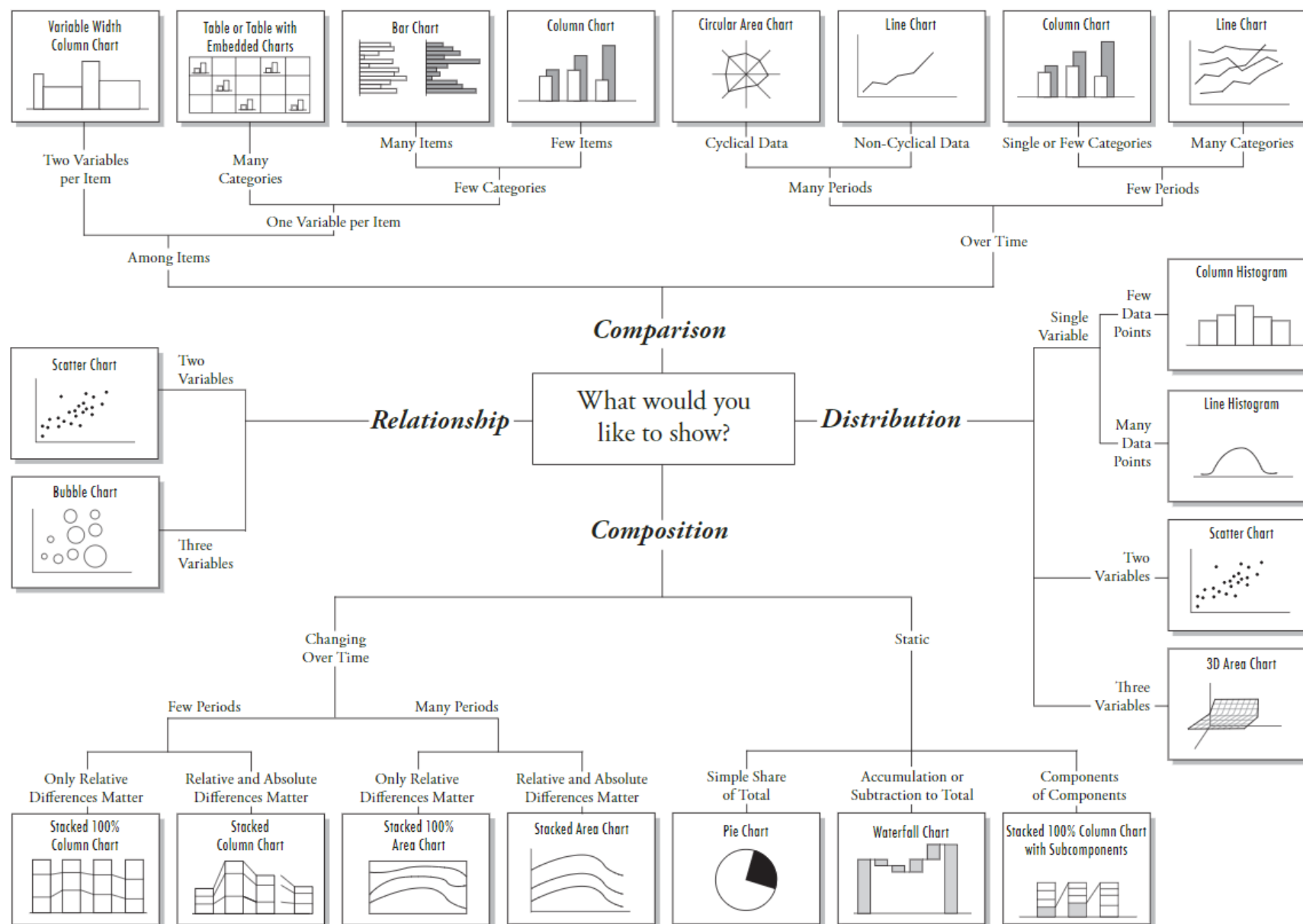
Converting average into average groups

# Cross-tabulation or Contingency Tables

| Count of gender            | Column Labels |      |             |
|----------------------------|---------------|------|-------------|
| Row Labels                 | Female        | Male | Grand Total |
| Biology                    | 959           | 638  | 1597        |
| Economics and Finance      | 461           | 863  | 1324        |
| Environmental Sciences     | 745           | 881  | 1626        |
| Mathematics and Statistics | 276           | 949  | 1225        |
| Political Science          | 978           | 477  | 1455        |
| Social Sciences            | 691           | 321  | 1012        |
| Grand Total                | 4110          | 4129 | 8239        |

Used to show the **relationship between two categorical variables**.  
Major by Gender in our example.

# Data Visualization

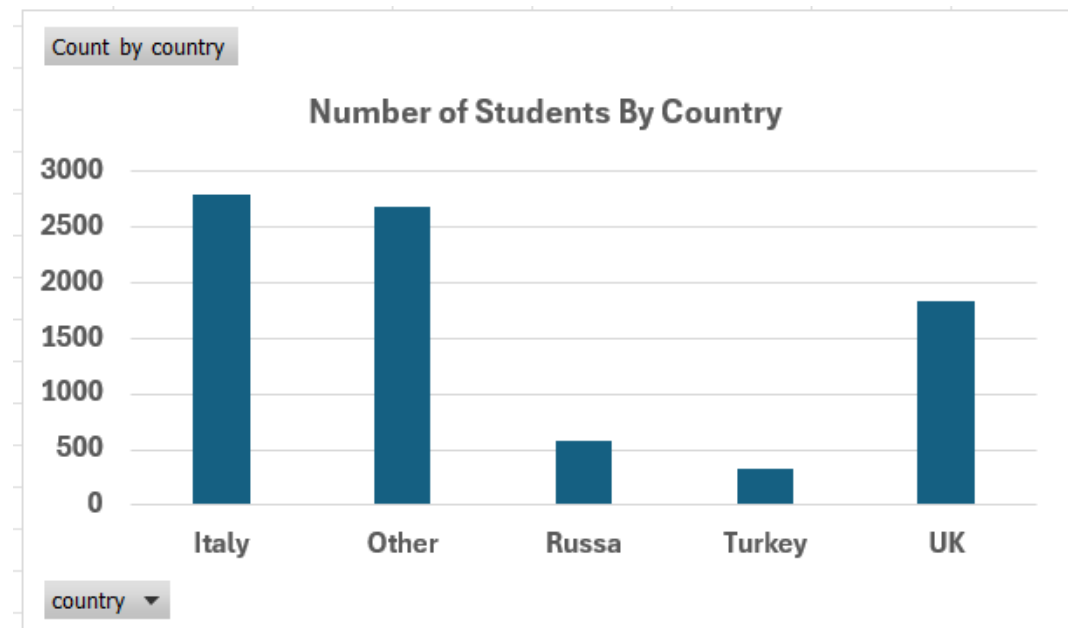


# Visualizing Qualitative Data

Bar Chart, Pie Chart, Pareto Plots, Side by Side Bars Chart

# Bar Charts

- A bar chart is a type of chart that presents data in rectangular bars with lengths proportional to the values they represent.
- Bar charts are commonly used to compare the magnitudes of different categories or groups of data.

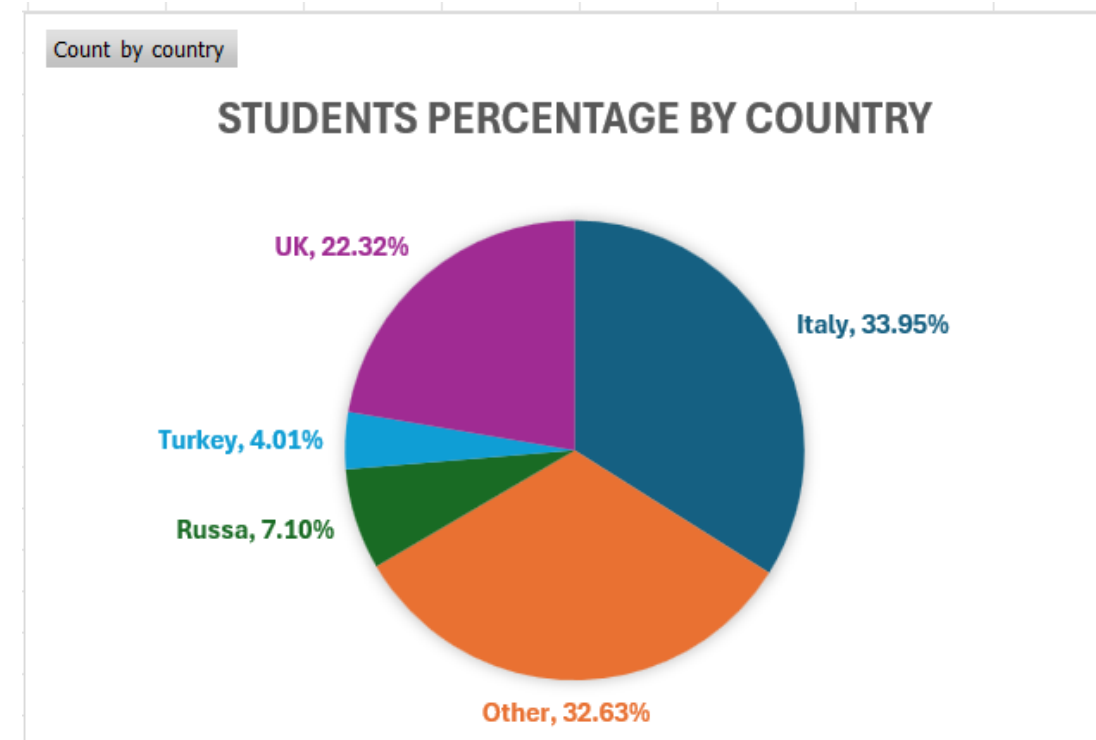


| Row Labels  | Count by country |
|-------------|------------------|
| Italy       | 2797             |
| Other       | 2688             |
| Russa       | 585              |
| Turkey      | 330              |
| UK          | 1839             |
| Grand Total | 8239             |

We need to construct a frequency table first.

## Pie Chart: Visualizing Proportions

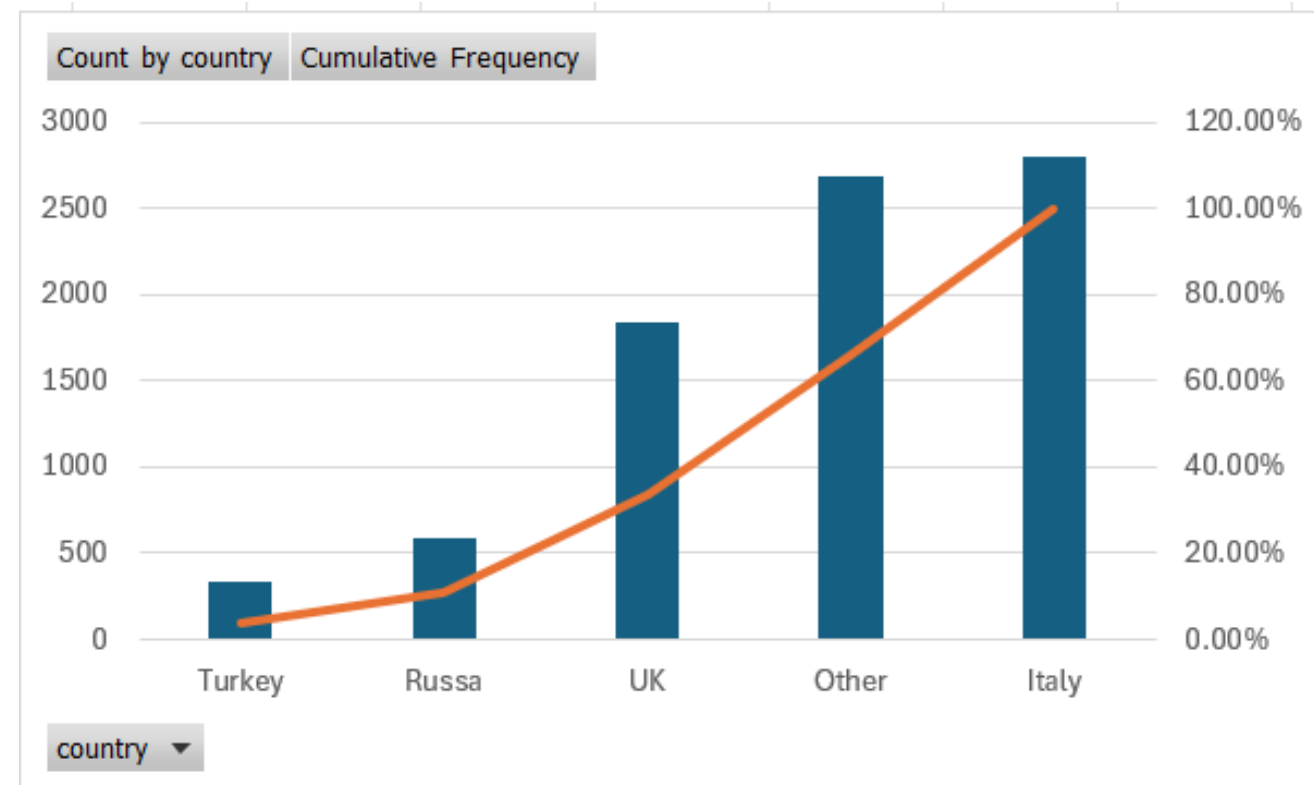
- A pie chart is a circular chart that is divided into segments to represent numerical proportions or percentages.
- Each segment of the pie chart represents a different category or group, and the size of each segment is proportional to the value or percentage it represents.
- Pie charts are commonly used to display data that can be divided into categories, such as market share, demographic distributions, or survey results.





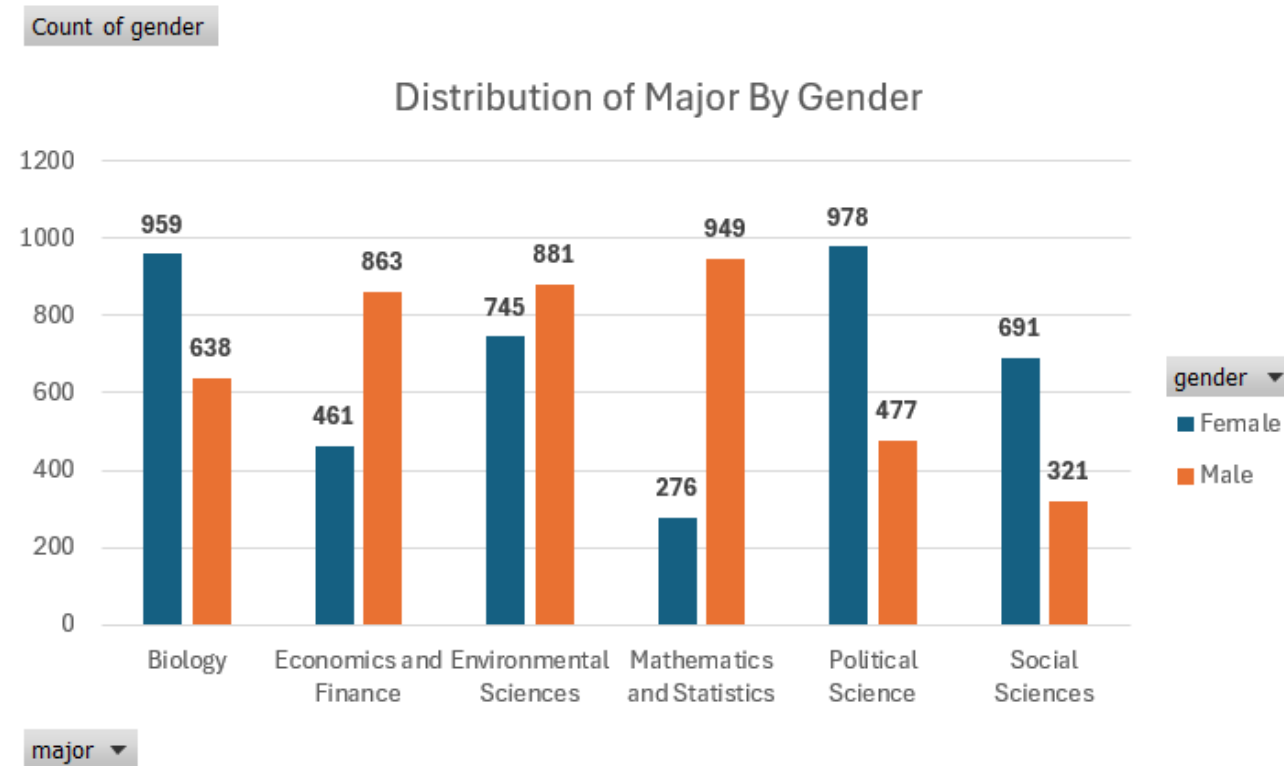
## Pareto Charts: Two Graphs in a Single Chart

- A Pareto chart is a type of chart that **combines** a bar graph with a line graph (Combo Chart).
- It is used to display the relative frequency or size of different categories or groups, and to identify the most important factors affecting a particular outcome or result.



# Side by Side Bar Charts: : Visualizing Two Variables

- A side-by-side chart is used to compare two or more variables or categories.
- The chart displays the data side-by-side in separate columns or bars, making it easy to compare the values of each variable or category.
- Side-by-side chart allows viewers to quickly and easily compare the values of different variables and identify patterns or trends in the data.



# Visualizing Quantitative Data, Frequency Distributions and Histograms

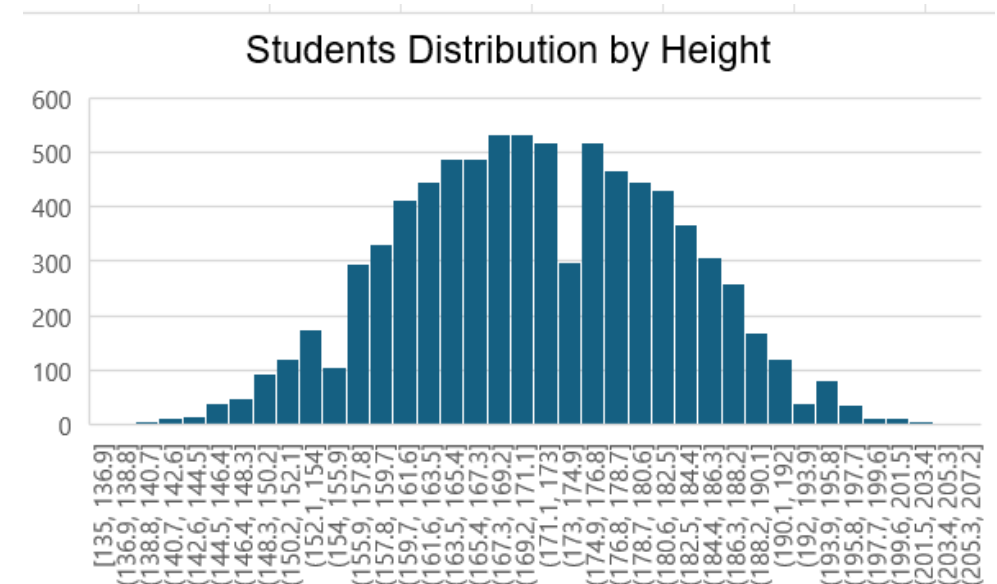
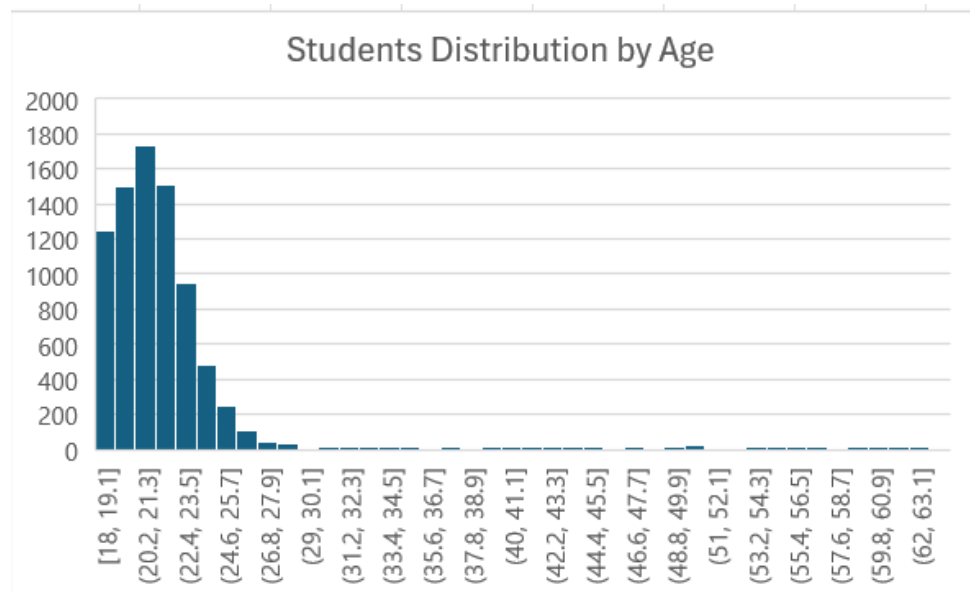
# Histograms

- Histograms provide a way to visualize the distribution of a numeric variable.
- Histograms group numeric responses into bins and display the frequency of responses in each.
- The x-axis of a histogram reflects the range of values of a numeric variable, while the y-axis can reflect either the frequencies (counts) or relative frequencies.

# Histograms

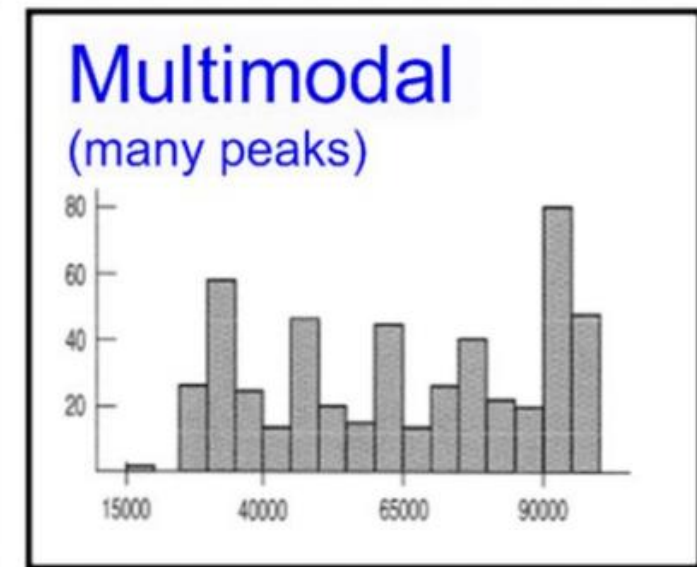
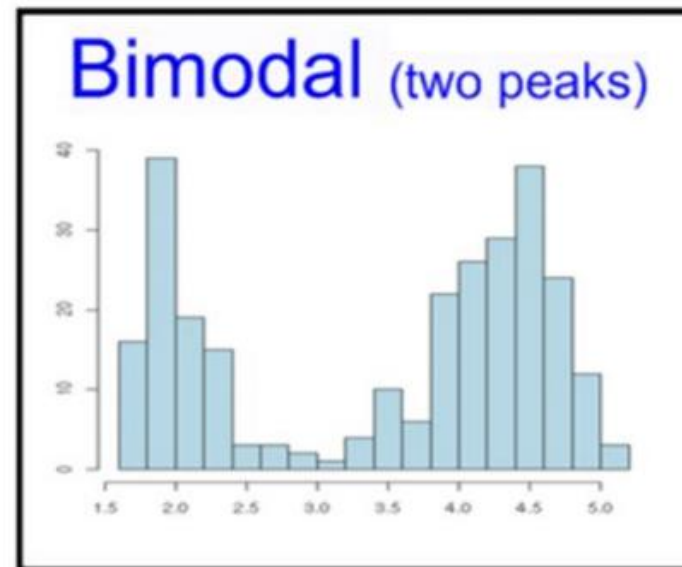
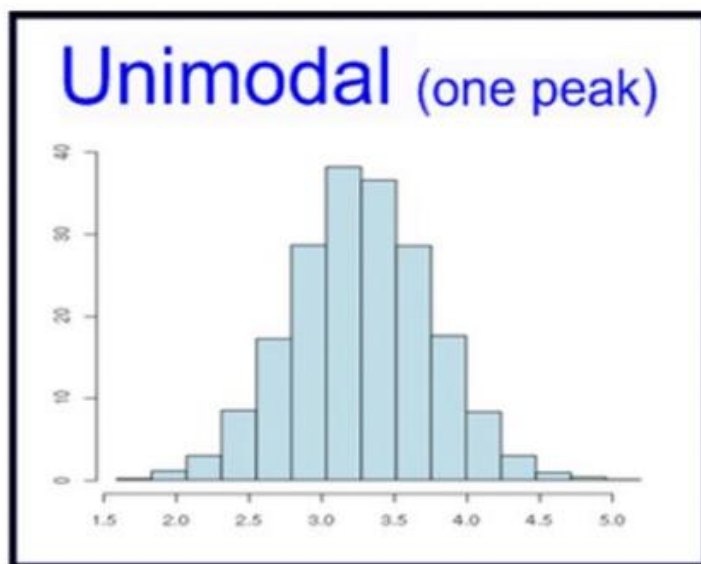
Histograms tell us the following about our data:

1. Shape – what is the shape of the distribution?
2. Center – what is an “average” value?
3. Spread – how far away from the center do values tend to fall?
4. Unique features – are there any outliers?



# Number of Modes

- The first distinguishing feature apparent in a histogram is the number of modes, or, in the distribution.
- A unimodal distribution only has one peak in the distribution, a bimodal distribution has two peaks, and a multimodal distribution has three or more peaks.

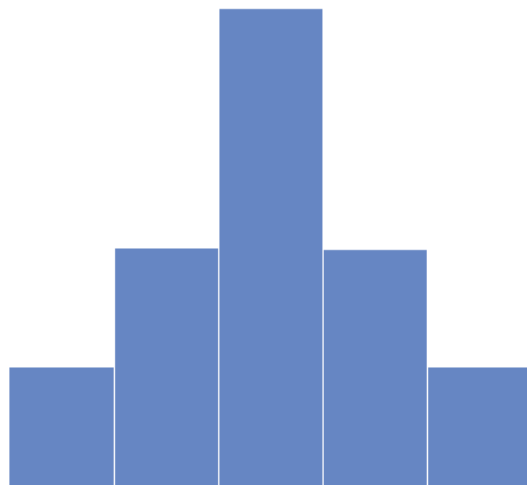


# Skewness

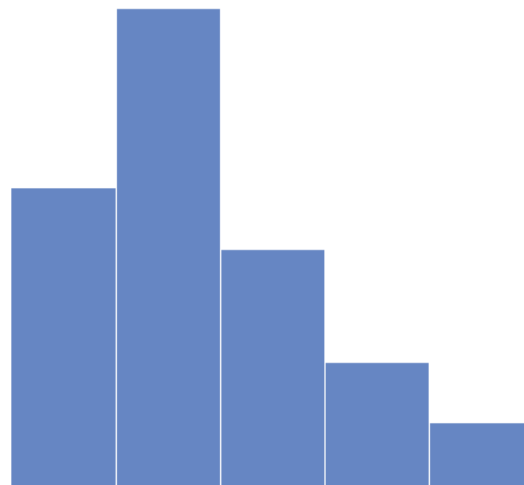
- Histograms can show us if the data is **skewed** or **symmetric**.
- Symmetric data should look nearly identical if folded in half at the center point of the distribution.
- Skewed data indicates that there is a large portion of the data collected on one side of the chart and only a small portion on the other side.

**FIGURE** Histograms with differing shapes

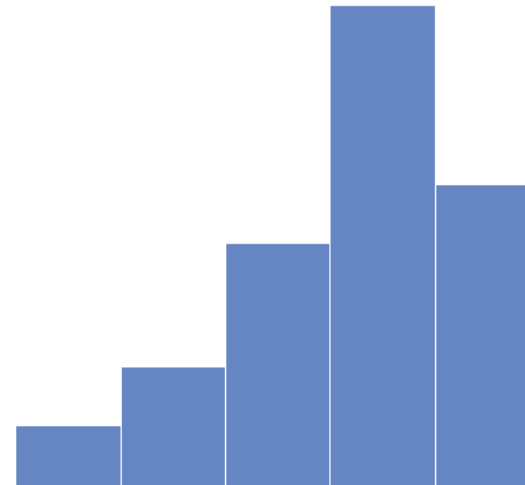
(a) Symmetric distribution



(b) Positively skewed distribution

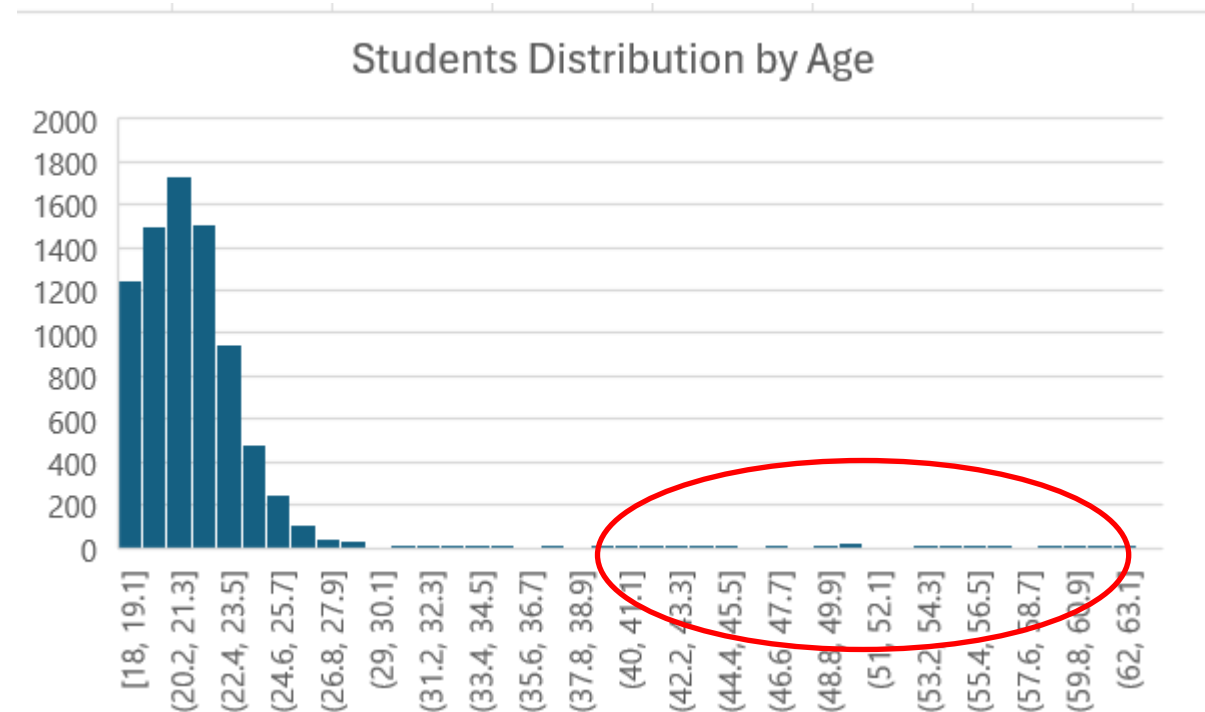


(c) Negatively skewed distribution



# Outliers

- Outliers are responses that fall well away from the rest of the values.
- Histograms can be useful in identifying outliers in our data.
- Defining an observation as an outlier is subjective and should lead to an investigation of that value (not an automatic removal from the dataset).



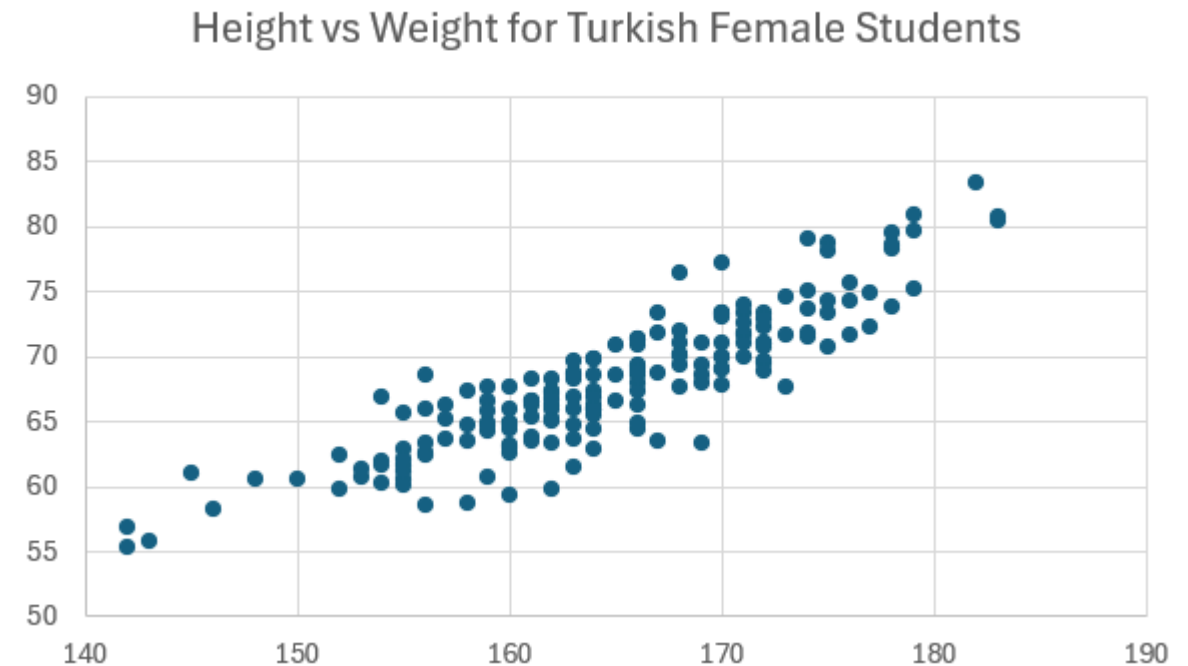


# Scatter Plots

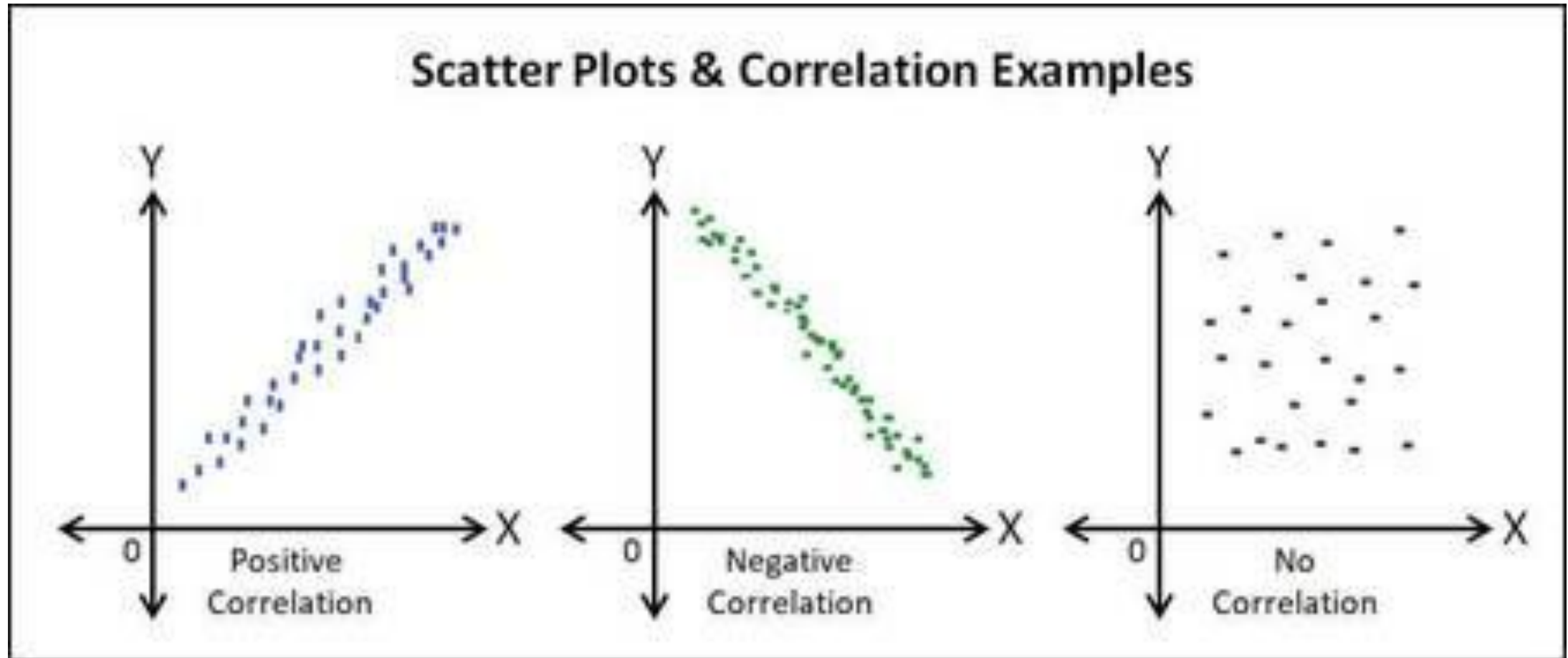
Scatter plots can be used to examine if there is a relation between two variables.

Consider for instance, how:

- Incomes vary with education.
- Sales vary with advertising expenditures.
- Stock prices vary with corporate profits.
- Crop yields vary with the use of fertilizer.
- Cholesterol levels vary with dietary intake.
- Price varies with reliability.



# Relations between Two Variables



# Numerical Measures in Descriptive Statistics

# Numerical Measures in Descriptive Statistics

- Numerical measures in descriptive statistics are used to summarize and describe the main features of a dataset through numbers.
- These measures provide insights into the distribution, central tendency, dispersion, and shape of the data.
- Understanding these measures helps in interpreting the data more effectively.

## Examples

1. Rates in a Society (unemployment rate, crime rate, divorce rate).
2. Average Income, Maximum age, Minimum score...etc.

# Measures of Central Location

Mean, Median and Mode

## The Mean

- The **Arithmetic Mean** is the primary measure of central location.
- Generally, we refer to the arithmetic mean as simply the **Mean** or the **Average**.
- In order to calculate the mean of a data set, we simply add up the observations and divide by the number of observations in the population or sample.
- However, the mean is sensitive to **Outliers**. Consider the following sample dataset:

**Data = (\$1.00, \$2.00, \$3.00, \$3.00, \$3.00, \$1.00, \$100.00)**

- Most of the data values are around \$3, yet their average is **\$16.14**

# The Median

## MEASURE OF CENTRAL LOCATION: THE MEDIAN

The median is the middle value of a data set. The data are arranged in ascending order (smallest to largest) and the median is calculated as

- The middle value if the number of observations is odd, or
- The average of the two middle values if the number of observations is even.

The median is especially useful when outliers are present.

$$\text{Med}(X) = \begin{cases} X[\frac{n+1}{2}] & \text{if } n \text{ is odd} \\ \frac{X[\frac{n}{2}] + X[\frac{n}{2} + 1]}{2} & \text{if } n \text{ is even} \end{cases}$$

$X$  = ordered list of values in data set

$n$  = number of values in data set

1, 3, 3, **6**, 7, 8, 9

Median = **6**

1, 2, 3, **4**, **5**, 6, 8, 9

Median =  $(4 + 5) \div 2$   
= **4.5**

# The Mode

## MEASURE OF CENTRAL LOCATION: THE MODE

The mode is the most frequently occurring value in a data set. A data set may have no mode or more than one mode.



# Excel Functions for Measuring Central Location

Mean

=AVERAGE(array)

Median

=MEDIAN(array)

Mode

=MODE(array)

Minimum

=MIN(array)

Maximum

=MAX(array)

# Practical Exercise

Use the students.xlsx file:

- Compare the Mean, Median for the “age” column for “Italy” and “Russia”.
- Find the Mode “age” column for “UK”.

Note: In Excel you can filter out rows by converting a Range to a Table and using the filtering list in each header of a column.

|    | A       | B            | C      | D   | E      | F      | G       | H |
|----|---------|--------------|--------|-----|--------|--------|---------|---|
| #  | stud.id | name         | gender | age | height | weight | country |   |
| 1  | 833917  | Gonzales, T  | Female | 19  | 160    | 64.8   | Turkey  |   |
| 2  | 898539  | Lozano, T    | Female | 19  | 172    | 73     | Other   |   |
| 3  | 379678  | Williams, F  | Female | 22  | 168    | 70.6   | UK      |   |
| 4  | 807564  | Nem, Denz    | Male   | 19  | 183    | 79.7   | Other   |   |
| 5  | 383291  | Powell, He   | Female | 21  | 175    | 71.4   | Italy   |   |
| 6  | 256074  | Perez, Jadr  | Male   | 19  | 189    | 85.8   | Italy   |   |
| 7  | 754591  | Clardy, Ani  | Female | 21  | 156    | 65.9   | UK      |   |
| 8  | 146494  | Allen, Rebe  | Female | 21  | 167    | 65.7   | Other   |   |
| 9  | 723584  | Tracy, Rob   | Male   | 18  | 195    | 94.4   | Other   |   |
| 10 | 314281  | Nimmons, J   | Female | 18  | 165    | 66     | Russia  |   |
| 11 | 200803  | Lang, Mac    | Female | 22  | 162    | 66.8   | Other   |   |
| 12 | 444907  | Rodriguez, J | Female | 18  | 172    | 66.8   | Other   |   |
| 13 | 354271  | Covar Orer   | Male   | 23  | 185    | 84.6   | Russia  |   |
| 14 | 317812  | Lopez, Mor   | Female | 20  | 158    | 64.4   | Italy   |   |
| 15 | 604115  | Davis, Sha   | Female | 19  | 157    | 66.3   | Russia  |   |
| 16 | 889551  | Adams, Jos   | Male   | 20  | 172    | 73.9   | Other   |   |
| 17 | 350040  | Hines, Hail  | Female | 22  | 156    | 61.7   | Other   |   |
| 18 | 240279  | Daugherty, M | Male   | 22  | 182    | 82.1   | Italy   |   |
| 19 | 865835  | Roybal, Ebr  | Female | 21  | 162    | 69.2   | Italy   |   |
| 20 | 137196  | Baysinger, J | Female | 22  | 168    | 70.9   | UK      |   |
| 21 | 708242  | Phillips, La | Female | 20  | 167    | 68.5   | Other   |   |
| 22 | 499002  | Culbertson   | Male   | 37  | 175    | 70.4   | UK      |   |
| 23 | 873149  | O Reilly, Jo | Male   | 19  | 164    | 70.3   | UK      |   |
| 24 | 807361  | Johnson, S   | Female | 38  | 155    | 67     | Italy   |   |

# Measures of Dispersion

Range, Mean Absolution Difference, Variance and Standard Deviation

## Measures of Dispersion

- Average alone is not sufficient to summarize and describe our data.
- Consider the following 5 sample datasets.
- Although the data values are very different, they all have the same average.

|                | <b>S1</b>     | <b>S2</b>     | <b>S3</b>     | <b>S4</b>     | <b>S5</b>     |
|----------------|---------------|---------------|---------------|---------------|---------------|
|                | 100           | 200           | 0             | -600          | -2000         |
|                | 200           | 200           | 600           | 600           | 2500          |
|                | 300           | 200           | 0             | 600           | 100           |
|                |               |               |               |               |               |
| <b>Average</b> | <b>200.00</b> | <b>200.00</b> | <b>200.00</b> | <b>200.00</b> | <b>200.00</b> |

- The **Range** is the simplest measure of dispersion, it is the difference between the maximum value and the minimum value in a data set:

$$\text{Range} = \text{Maximum Value} - \text{Minimum Value}$$

- Going back to the previous sample datasets, we can see that they have the same average, but their ranges are very different.

|         | S1     | S2     | S3     | S4      | S5      |
|---------|--------|--------|--------|---------|---------|
|         | 100    | 200    | 0      | -600    | -2000   |
|         | 200    | 200    | 600    | 600     | 2500    |
|         | 300    | 200    | 0      | 600     | 100     |
|         |        |        |        |         |         |
| Average | 200.00 | 200.00 | 200.00 | 200.00  | 200.00  |
| Range   | 200.00 | 0.00   | 600.00 | 1200.00 | 4500.00 |

- Therefore, using the **Range** with the **Average** helped us to better describe and summarize our data (Better Understand Our Data).
- However, the Range is sensitive to Outliers. Consider the following sample dataset:**  
Data = (\$1.00, \$2.00, \$3.00, \$3.00, \$3.00, \$1.00, \$100.00)
- Most of the data values are around \$3, yet their Range is **\$99.00**

# Measures of Dispersion - The Mean Absolute Deviation

- A good measure of dispersion should consider differences of all observations from the mean.
- If we simply average all differences from the mean, the positives and the negatives will cancel out, even though they both contribute to dispersion, and the resulting average will equal zero.
- The **mean absolute deviation** (MAD) is an average of the absolute differences between the observations and the mean.

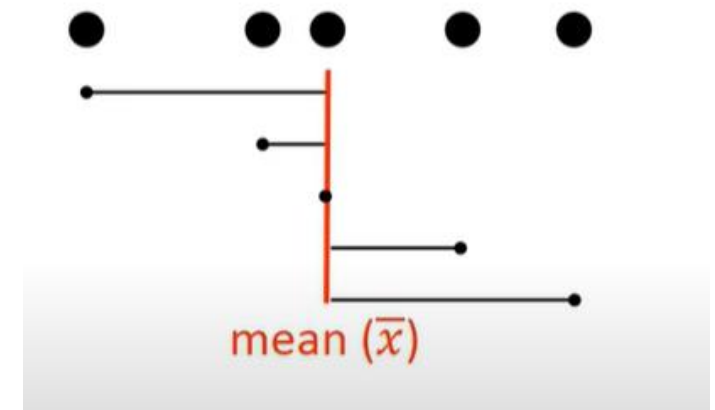
## MEASURE OF DISPERSION: THE MEAN ABSOLUTE DEVIATION (MAD)

For sample values,  $x_1, x_2, \dots, x_n$ , the sample MAD is computed as

$$\text{Sample MAD} = \frac{\sum |x_i - \bar{x}|}{n}$$

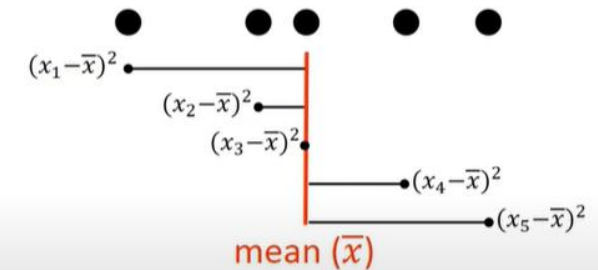
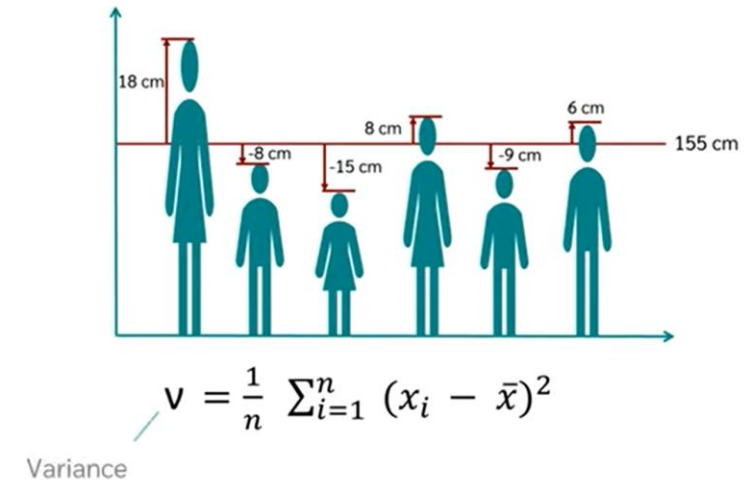
For population values,  $x_1, x_2, \dots, x_N$ , the population MAD is computed as

$$\text{Population MAD} = \frac{\sum |x_i - \mu|}{N}$$



# Measures of Dispersion - The Variance and the Standard Deviation

- MAD weights large and small differences **equally**.
- Better measures are the **variance** and the **standard deviation**.
- Instead of calculating the average of the absolute differences from the mean, as in MAD, we calculate the average of the squared differences from the mean.
- The squaring of differences from the mean **emphasizes larger differences more than smaller ones**.
- The variance has squared units, therefore, to return the values to their original units of measurement, we compute the square root of the variance, which gives us the standard deviation.



# Measures of Dispersion - The Variance and the Standard Deviation

## MEASURES OF DISPERSION: THE VARIANCE AND THE STANDARD DEVIATION

For **sample** values  $x_1, x_2, \dots, x_n$ , the sample variance  $s^2$  and the sample standard deviation  $s$  are computed as

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} \quad \text{and} \quad s = \sqrt{s^2}.$$

For **population** values  $x_1, x_2, \dots, x_N$ , the population variance  $\sigma^2$  (the Greek letter sigma, squared) and the population standard deviation  $\sigma$  are computed as

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \quad \text{and} \quad \sigma = \sqrt{\sigma^2}.$$

*Note:* The sample variance uses  $n - 1$  rather than  $n$  in the denominator to ensure that the sample variance is an unbiased estimator for the population variance, a topic discussed in Chapter 8.



# Excel Functions for Measuring Dispersion

Range

=MAX(array)-MIN(array)

Mean Absolute Deviation

=AVEDEV(array)

Sample Variance

=VAR.S(array)

Sample Standard Deviation

=STDEV.S(array)

Population Variance

=VAR.P(array)

Population Standard Deviation

=STDEV.P(array)

# Practical Exercise

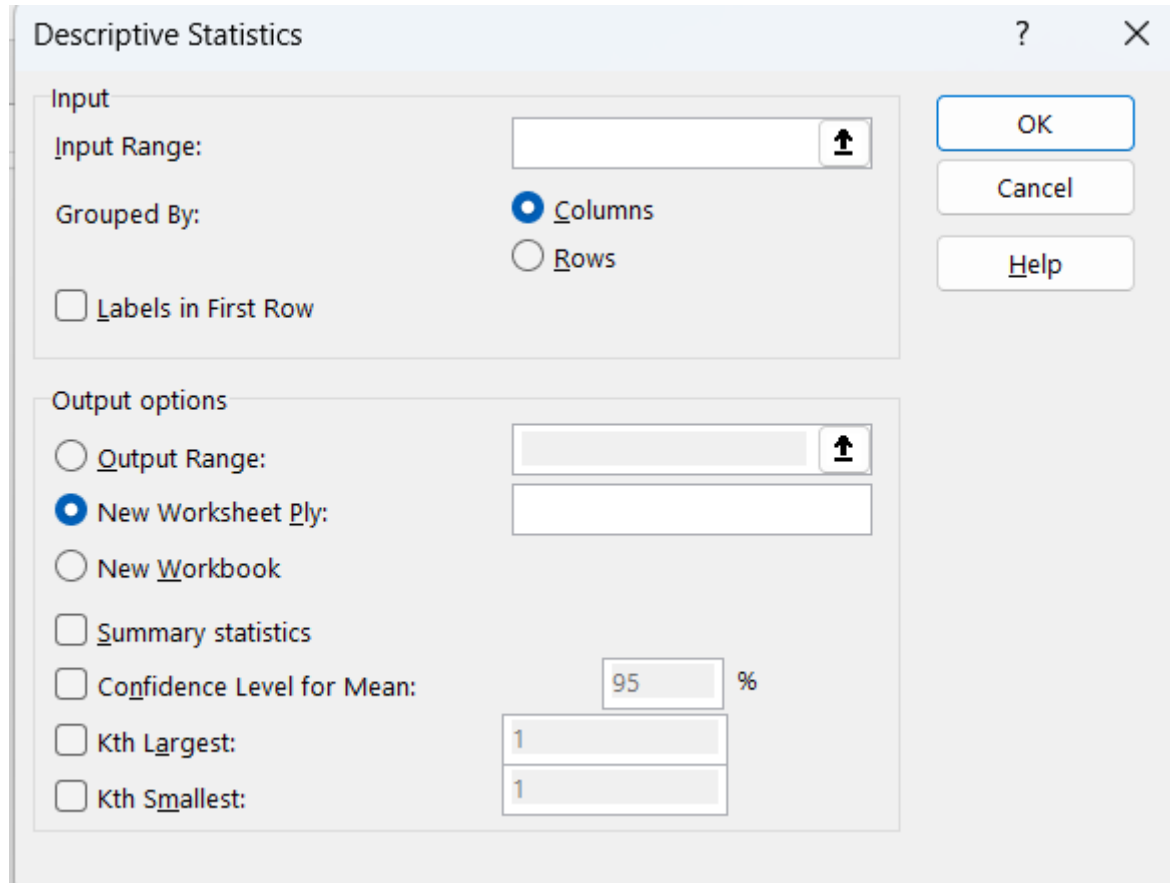
Use the students.xlsx file to find the Range, MAD, Variance and Standard Deviation for the age column.

Note: In Excel you can filter out rows by converting a Range to a Table and using the filtering list in each header of a column.

|    | A | B         | C            | D        | E     | F        | G        | H         |
|----|---|-----------|--------------|----------|-------|----------|----------|-----------|
| #  | ▼ | stud.id ▼ | name ▼       | gender ▼ | age ▼ | height ▼ | weight ▼ | country ▼ |
| 1  |   | 833917    | Gonzales, t  | Female   | 19    | 160      | 64.8     | Turkey    |
| 2  |   | 898539    | Lozano, T    | Female   | 19    | 172      | 73       | Other     |
| 3  |   | 379678    | Williams, f  | Female   | 22    | 168      | 70.6     | UK        |
| 4  |   | 807564    | Nem, Denz    | Male     | 19    | 183      | 79.7     | Other     |
| 5  |   | 383291    | Powell, He   | Female   | 21    | 175      | 71.4     | Italy     |
| 6  |   | 256074    | Perez, Jadr  | Male     | 19    | 189      | 85.8     | Italy     |
| 7  |   | 754591    | Clardy, Ani  | Female   | 21    | 156      | 65.9     | UK        |
| 8  |   | 146494    | Allen, Rebe  | Female   | 21    | 167      | 65.7     | Other     |
| 9  |   | 723584    | Tracy, Rob   | Male     | 18    | 195      | 94.4     | Other     |
| 10 |   | 314281    | Nimmons,     | Female   | 18    | 165      | 66       | Russa     |
| 11 |   | 200803    | Lang, Mac    | Female   | 22    | 162      | 66.8     | Other     |
| 12 |   | 444907    | Rodriguez,   | Female   | 18    | 172      | 66.8     | Other     |
| 13 |   | 354271    | Covar Ozer   | Male     | 23    | 185      | 84.6     | Russa     |
| 14 |   | 317812    | Lopez, Mor   | Female   | 20    | 158      | 64.4     | Italy     |
| 15 |   | 604115    | Davis, Sha   | Female   | 19    | 157      | 66.3     | Russa     |
| 16 |   | 889551    | Adams, Jos   | Male     | 20    | 172      | 73.9     | Other     |
| 17 |   | 350040    | Hines, Hail  | Female   | 22    | 156      | 61.7     | Other     |
| 18 |   | 240279    | Daugherty,   | Male     | 22    | 182      | 82.1     | Italy     |
| 19 |   | 865835    | Roybal, Ebr  | Female   | 21    | 162      | 69.2     | Italy     |
| 20 |   | 137196    | Baysinger,   | Female   | 22    | 168      | 70.9     | UK        |
| 21 |   | 708242    | Phillips, La | Female   | 20    | 167      | 68.5     | Other     |
| 22 |   | 499002    | Culbertson   | Male     | 37    | 175      | 70.4     | UK        |
| 23 |   | 873149    | O Reilly, Jo | Male     | 19    | 164      | 70.3     | UK        |
| 24 |   | 807361    | Johnson, S   | Female   | 38    | 155      | 67       | Italy     |

# Practical Exercise: using Excel's data analysis tools option


Select: Data > Data Analysis > Descriptive Statistics > OK.



The image shows the 'Descriptive Statistics' dialog box in Microsoft Excel. The dialog is titled 'Descriptive Statistics' and has a question mark icon and a close button (X) in the top right corner. It is divided into two main sections: 'Input' and 'Output options'. In the 'Input' section, the 'Input Range' field is empty, followed by a selection icon. The 'Grouped By' section has two radio buttons: 'Columns' (which is selected) and 'Rows'. There is also a checkbox for 'Labels in First Row' which is currently unchecked. In the 'Output options' section, there are three radio buttons: 'Output Range:' (unchecked), 'New Worksheet Ply:' (selected), and 'New Workbook' (unchecked). The 'Output Range:' field is empty with a selection icon. The 'New Worksheet Ply:' field is also empty. Below these, there are four checkboxes: 'Summary statistics' (unchecked), 'Confidence Level for Mean:' (unchecked), 'Kth Largest:' (unchecked), and 'Kth Smallest:' (unchecked). The 'Confidence Level for Mean:' field has a text box containing '95' and a '%' symbol. The 'Kth Largest:' field has a text box containing '1'. The 'Kth Smallest:' field has a text box containing '1'. On the right side of the dialog, there are three buttons: 'OK', 'Cancel', and 'Help'.

Descriptive Statistics


Input

Input Range:  

Grouped By: ☒ Columns ☐ Rows

☐ Labels in First Row

Output options

☐ Output Range:  

☒ New Worksheet Ply:

☐ New Workbook

☐ Summary statistics

☐ Confidence Level for Mean:  %

☐ Kth Largest:

☐ Kth Smallest:

OK Cancel Help

# The Coefficient of Variation

## The Coefficient of Variation

- In some instances, we need to compare the variability of two or more data sets that have **different means or different units of measurement**.
- Consider the following example:

|                           | <b>DS1</b>   | <b>DS2</b>    |
|---------------------------|--------------|---------------|
|                           | \$10         | \$1010        |
|                           | \$20         | \$1020        |
|                           | \$30         | \$1030        |
|                           | \$40         | \$1040        |
|                           | \$50         | \$1050        |
| <b>Average</b>            | <b>\$30</b>  | <b>\$1030</b> |
| <b>Standard Deviation</b> | <b>15.81</b> | <b>15.81</b>  |

- We cannot tell which dataset has more variance in its values since both datasets have the same standard deviation.

## The Coefficient of Variation

- For such cases, we can use the Coefficient of Variation Measure which serves as a relative measure of dispersion.
- It can be calculated by dividing a data set's standard deviation by its mean:

$$CV = \frac{\sigma}{\mu}$$

- When comparing two datasets, the dataset that have larger coefficient of variation indicates that its values varies more than the other dataset. Therefore, it is less stable or less uniform.

## The Coefficient of Variation

- Going back to the previous example, we can calculate the coefficient of variation for both sets to help us in determining which dataset has more variability in its values.

|                                 | DS1            | DS2            |
|---------------------------------|----------------|----------------|
|                                 | \$10           | \$1010         |
|                                 | \$20           | \$1020         |
|                                 | \$30           | \$1030         |
|                                 | \$40           | \$1040         |
|                                 | \$50           | \$1050         |
| <b>Average</b>                  | <b>\$30</b>    | <b>\$1030</b>  |
| <b>Standard Deviation</b>       | <b>\$15.81</b> | <b>\$15.81</b> |
| <b>Coefficient of Variation</b> | <b>0.430</b>   | <b>0.013</b>   |

- Using the Coefficient of Variation, we can tell that DS1 varies more than DS2, and that DS2 is more stable than DS1.

- In finance, the coefficient of variation is **used to measure the risk per unit of return.**
- For example, assume that the average monthly return on stock A is 2980 with a standard deviation of 111.06, and the average monthly return on stock B is 9.8 with a standard deviation of 1.3
- We can use the coefficient of variation to determine which stock is more stable:

$$\text{Coefficient of Variation for Stock A} = \frac{111.6}{2980} = 0.037$$

$$\text{Coefficient of Variation for Stock B} = \frac{1.3}{9.8} = 0.134$$

The dispersion per unit monthly return for Stock B is larger than that of A.

Therefore, investing in Stock B is riskier than an investing in stock A.



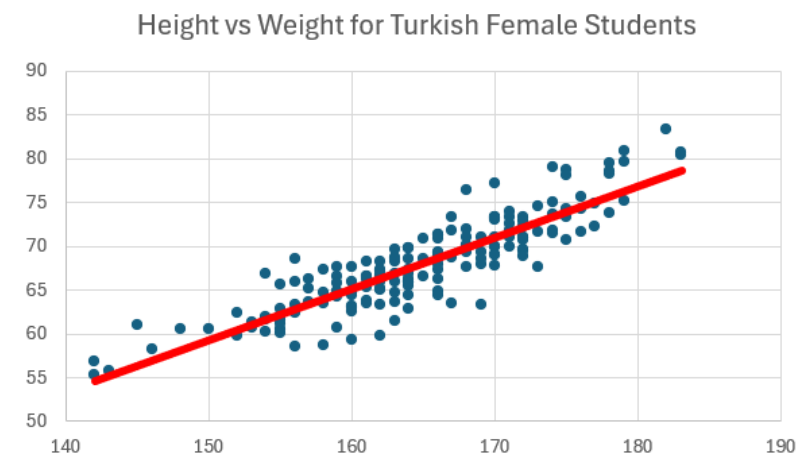
# Measures of Association

Covariance and Coefficient of Correlation

## Measures of Association

- We previously introduced the idea of a **scatter plots** to visually assess whether two variables had some type of linear relationship.
- In this section, we present two numerical measures of association that quantify the direction and strength of the linear relationship between two variables,  $x$  and  $y$ .
- These two numerical values are the **Covariance** and the **Correlation**.

Note: Covariance and Correlation works when the relation between the two variables is LINEAR.

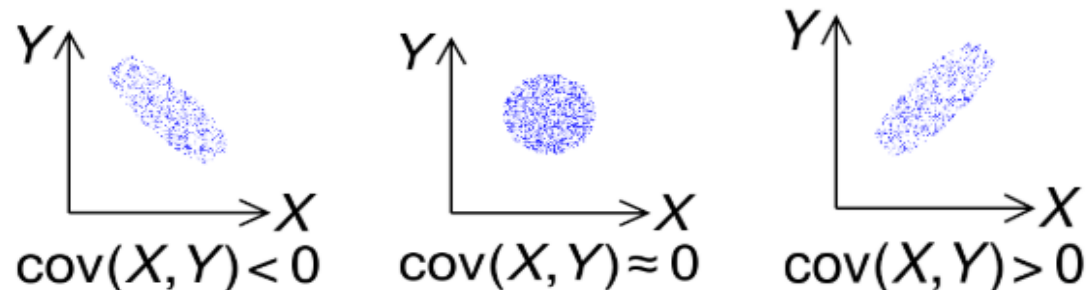


# Covariance

- Covariance measures the joint variability of two random variables.
- Indicates the direction of the linear relationship between variables.

$$\text{cov}_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

- Positive Covariance: X and Y move in the same direction.
- Negative Covariance: X and Y move in opposite directions.
- Zero Covariance: No linear relationship between X and Y.



## Limitations of Covariance

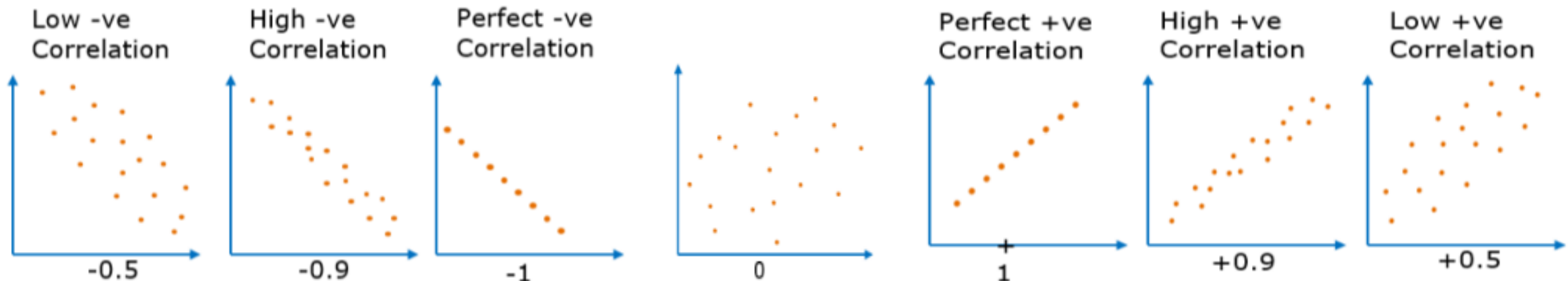
Covariance is sensitive to the units of X and Y.

Changing units changes the covariance value, making direct comparisons difficult.

|                    | Height m   | Height cm  | Weight KG                   |
|--------------------|------------|------------|-----------------------------|
|                    | 1.6        | 160        | 55                          |
|                    | 1.7        | 170        | 68                          |
|                    | 1.8        | 180        | 74                          |
| <b>Covariance</b>  | <b>.95</b> | <b>95</b>  | <b>Which one is bigger?</b> |
| <b>Correlation</b> | <b>.98</b> | <b>.98</b> |                             |

# The Correlation Coefficient

- Correlation coefficient can be used to determine the **Direction** and the **Strength** of a relationship between two numerical variables.
- Correlation Formula 
$$r = \frac{cov(x, y)}{\sigma_x \cdot \sigma_y}$$
- The correlation coefficient is a dimensionless metric and its value ranges from -1 to +1.
- The closer it is to +1 or -1, the more closely the two variables are related.



| Covariance   | Correlation  |
|--|--|
| Covariance indicates the direction of the linear relationship between variables. | Correlation measures both the strength and the direction of the linear relationship between two variables. |
| Covariance value can vary between $-\infty$ and $+\infty$                        | Correlation value ranges between -1 and +1   |
| Covariance unit is based on the product of the units of the two variables.       | Correlation is dimensionless, i.e. It's a unit-free measure of the relationship between variables.         |

# Practical Exercise

Use the students.xlsx file:

- Compute the covariance and correlation between heights and weights of the students.
- Compute the covariance and correlation between heights and ages of the students.

| A | B       | C           | D      | E   | F      | G      | H | I   |
|---|---------|-------------|--------|-----|--------|--------|---|---|
| # | stud.id | name        | gender | age | height | weight |   |   |
| 1 | 833917  | Gonzales, C | Female | 19  | 160    | 64.8   |   |   |
| 2 | 898539  | Lozano, T   | Female | 19  | 172    | 73     |   | =CORREL(Table1[age],Table1[height])       |
| 3 | 379678  | Williams, F | Female | 22  | 168    | 70.6   |   | =COVARIANCE.S(Table1[age],Table1[height]) |
| 4 | 807564  | Nem, Deniz  | Male   | 19  | 183    | 79.7   |   | =COVARIANCE.P(Table1[age],Table1[height]) |
| 5 | 383291  | Powell, He  | Female | 21  | 175    | 71.4   |   |   |
| 6 | 256074  | Perez, Jadr | Male   | 19  | 189    | 85.8   |   |   |
| 7 | 754591  | Clardy, Ani | Female | 21  | 156    | 65.9   |   |   |
| 8 | 146494  | Allen, Rebe | Female | 21  | 167    | 65.7   |   |   |
| 9 | 723584  | Tracy, Rob  | Male   | 18  | 195    | 94.4   |   |   |

# Practical Exercise

We can also use the Data Analysis add-in for this purpose.

|    | A  | B       | C           | D      | E   | F      | G      | H | I | J | K |
|----|----|---------|-------------|--------|-----|--------|--------|---|---|---|---|
| 1  | #  | stud.id | name        | gender | age | height | weight |   |   |   |   |
| 2  | 1  | 833917  | Gonzales, I | Female | 19  | 160    | 64.8   |   |   |   |   |
| 3  | 2  | 898539  | Lozano, T   | Female | 19  | 172    | 73     |   |   |   |   |
| 4  | 3  | 379678  | Williams, I | Female | 22  | 168    | 70.6   |   |   |   |   |
| 5  | 4  | 807564  | Nem, Deniz  | Male   | 19  | 183    | 79.7   |   |   |   |   |
| 6  | 5  | 383291  | Powell, He  | Female | 21  | 175    | 71.4   |   |   |   |   |
| 7  | 6  | 256074  | Perez, Jadr | Male   | 19  | 189    | 85.8   |   |   |   |   |
| 8  | 7  | 754591  | Clardy, Ani | Female | 21  | 156    | 65.9   |   |   |   |   |
| 9  | 8  | 146494  | Allen, Rebe | Female | 21  | 167    | 65.7   |   |   |   |   |
| 10 | 9  | 723584  | Tracy, Rob  | Male   | 18  | 195    | 94.4   |   |   |   |   |
| 11 | 10 | 314281  | Nimmons,    | Female | 18  | 165    | 66     |   |   |   |   |
| 12 | 11 | 200803  | Lang, Mack  | Female | 22  | 162    | 66.8   |   |   |   |   |
| 13 | 12 | 444907  | Rodriguez,  | Female | 18  | 172    | 66.8   |   |   |   |   |
| 14 | 13 | 354271  | Covar Orer  | Male   | 23  | 185    | 84.6   |   |   |   |   |
| 15 | 14 | 217812  | Long, Mar   | Female | 20  | 158    | 64.4   |   |   |   |   |

## Data Analysis

### Analysis Tools

Anova: Two-Factor With Replication  
Anova: Two-Factor Without Replication  
**Correlation**  
Covariance  
Descriptive Statistics  
Exponential Smoothing  
F-Test Two-Sample for Variances  
Fourier Analysis  
Histogram  
Moving Average

OK

Cancel

Help

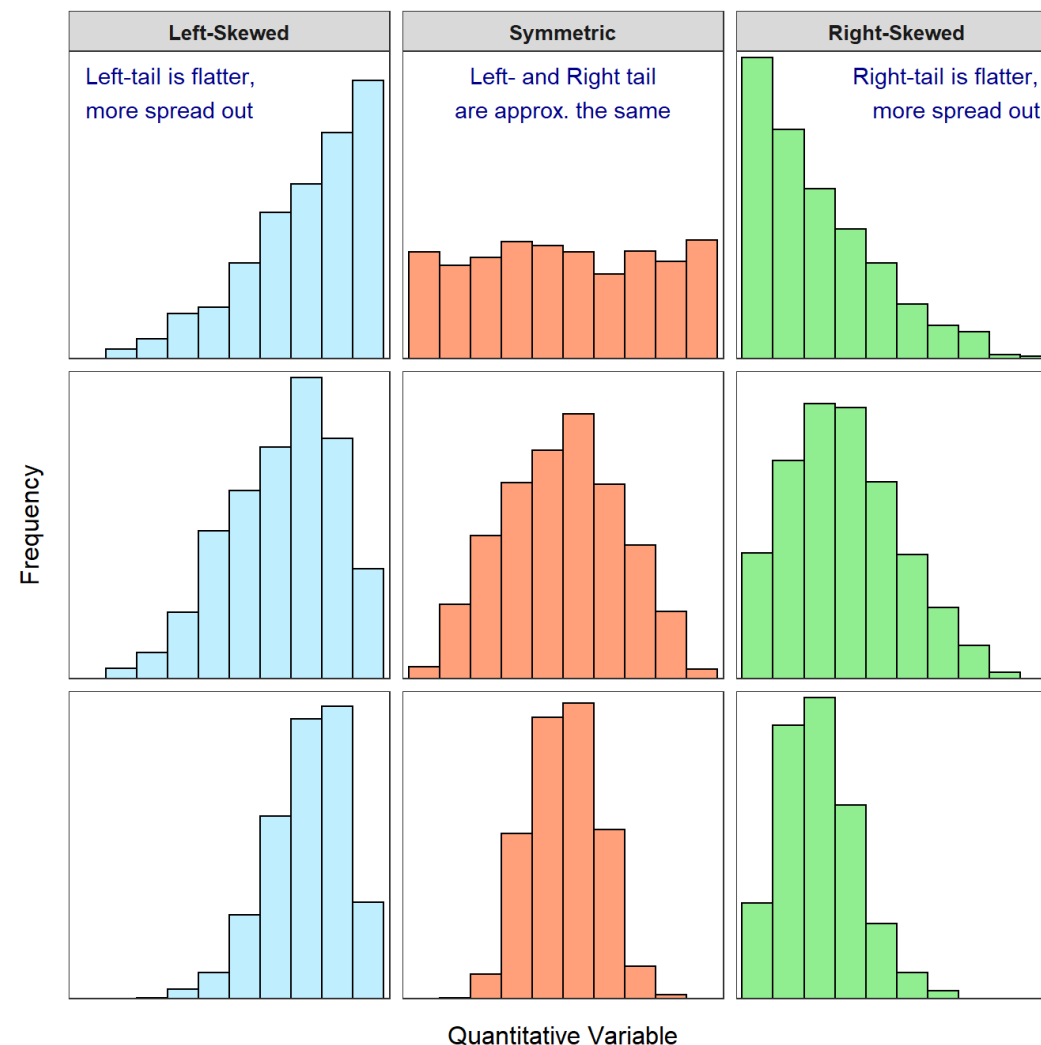
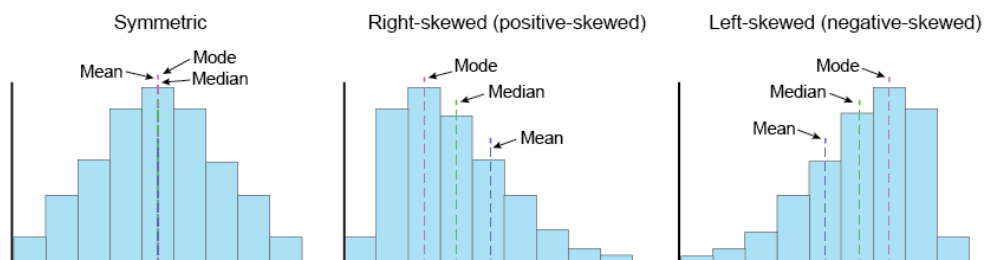


# Measures of Distribution

Symmetry and Skewness

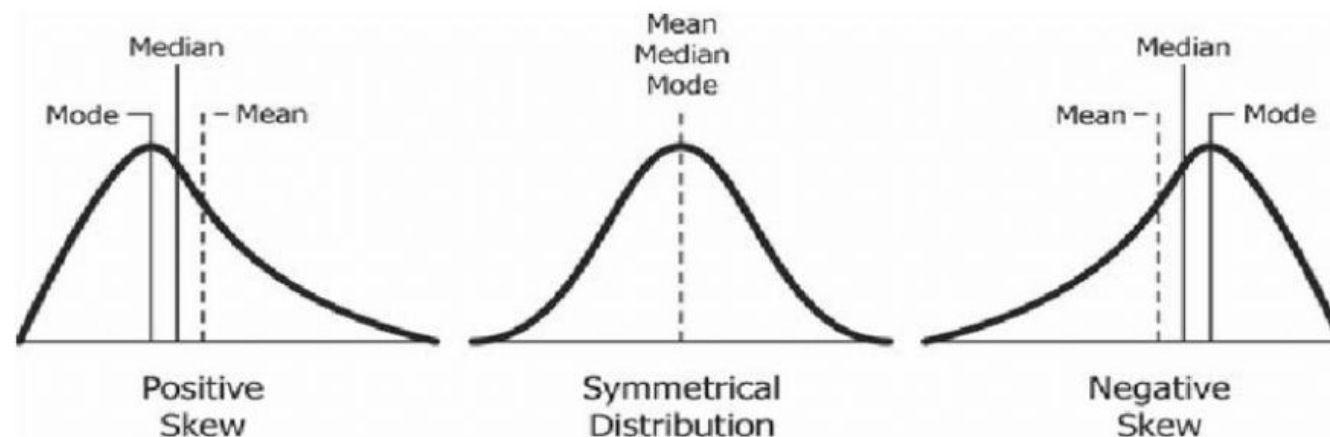
# Skewness

- **Definition:** Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.
- **Key Point:** Indicates whether the data are spread out more to one side of the mean.
- **Positive Skewness (Right-skewed):** Tail is longer on the right side of the distribution. Mean is greater than the median.
- **Negative Skewness (Left-skewed):** Tail is longer on the left side of the distribution. Mean is less than the median.
- Symmetrical Distribution: Skewness is around 0, indicating no skew.
- Use histograms or density plots to illustrate examples of positively skewed, negatively skewed, and symmetrical distributions.



# Skewness Coefficient

- A **positive skewness** coefficient implies that extreme values are concentrated in the right tail of the distribution, pulling the mean up, relative to the median and the bulk of values lie to the left of the mean.
- Similarly, a **negative skewness** coefficient implies that extreme values are concentrated in the left tail of the distribution, pulling the mean down, relative to the median, and the bulk of values lie to the right of the mean.

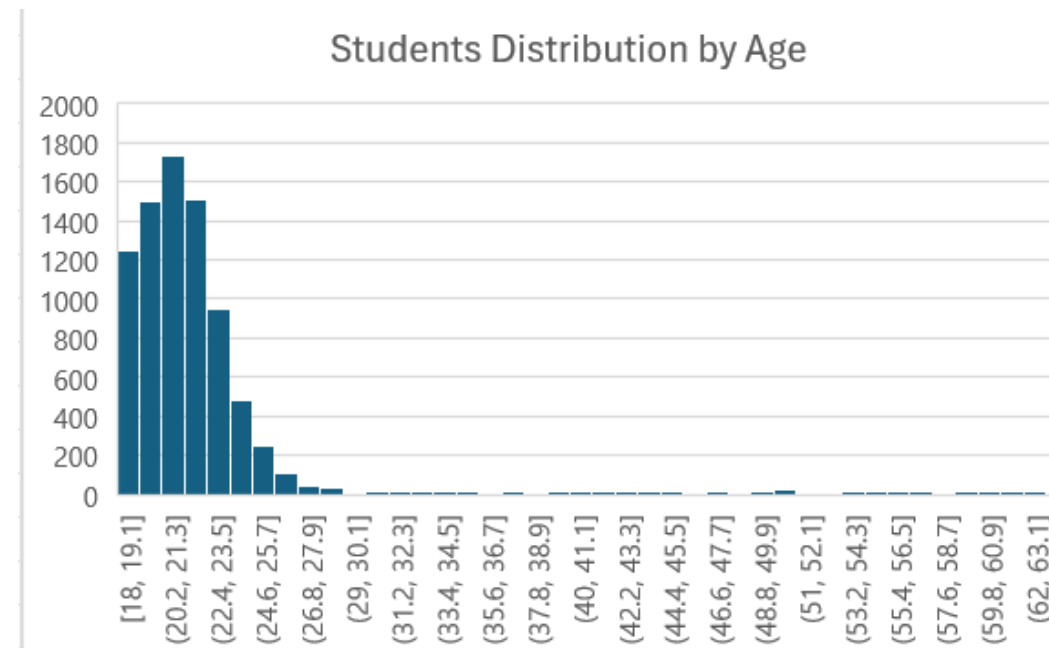


$$\text{Formula: } Skewness = \frac{1}{n} \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^3$$

- A skewness value greater than 1 or less than -1 indicates a highly skewed distribution.
- A value between 0.5 and 1 or -0.5 and -1 is moderately skewed.
- A value between -0.5 and 0.5 indicates that the distribution is fairly symmetrical.
- A skewness coefficient of zero indicates that the values are evenly distributed on both sides of the mean.

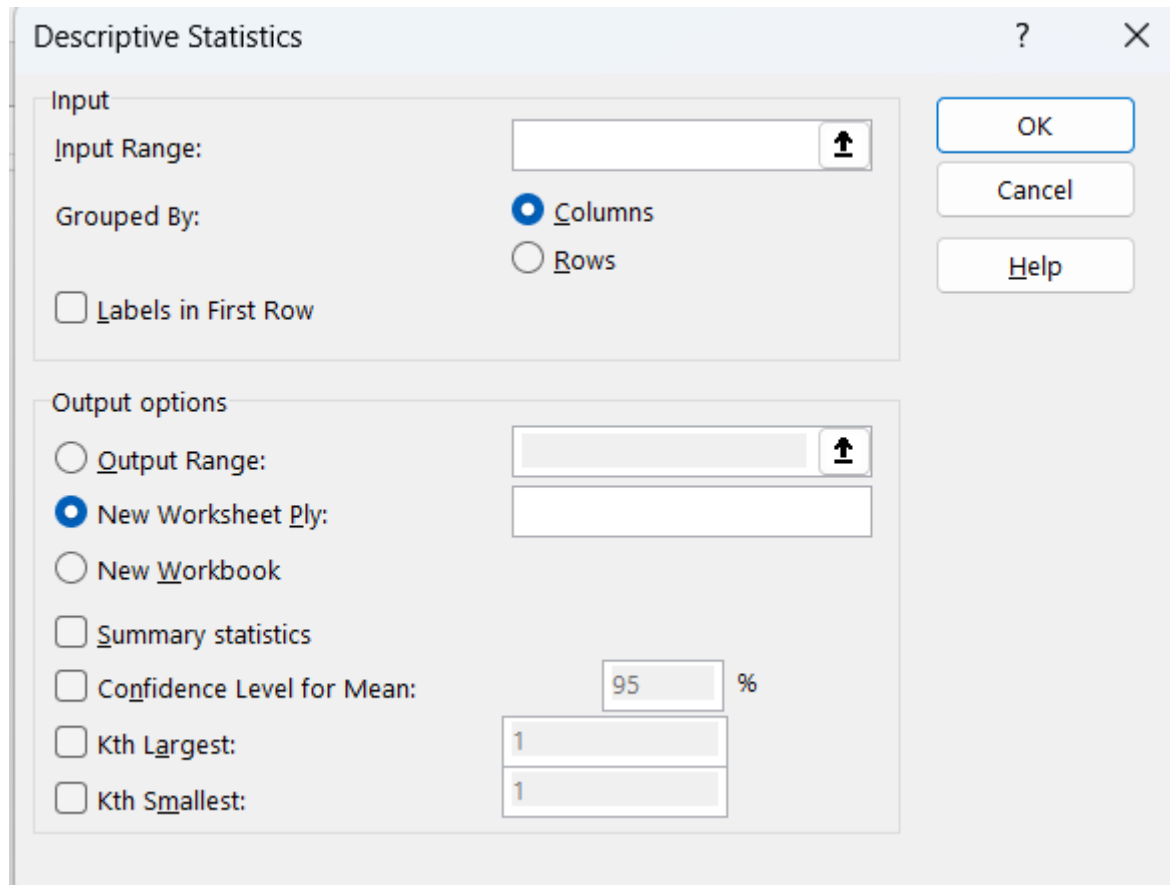
## Practical Example

- Using the students.xlsx file, find the Skewness value for the “age” column.
- You can use Excel’s skew() function to calculate the skewness for a sample, and the skew.p() function to calculate the skewness for a population.



# Practical Exercise: using Excel's data analysis tools option


Select: Data > Data Analysis > Descriptive Statistics > OK.



The image shows the 'Descriptive Statistics' dialog box in Microsoft Excel. The dialog is titled 'Descriptive Statistics' and has a question mark icon and a close button (X) in the top right corner. It is divided into two main sections: 'Input' and 'Output options'. In the 'Input' section, the 'Input Range' field is empty, followed by a selection icon. The 'Grouped By' section has two radio buttons: 'Columns' (which is selected) and 'Rows'. There is also a checkbox for 'Labels in First Row' which is currently unchecked. In the 'Output options' section, there are three radio buttons: 'Output Range' (unchecked), 'New Worksheet Ply:' (selected), and 'New Workbook' (unchecked). The 'Output Range' field is empty, followed by a selection icon. Below these, there are four checkboxes: 'Summary statistics' (unchecked), 'Confidence Level for Mean:' (unchecked), 'Kth Largest:' (unchecked), and 'Kth Smallest:' (unchecked). The 'Confidence Level for Mean' field has a value of '95' and a '%' symbol. The 'Kth Largest' and 'Kth Smallest' fields both have a value of '1'. On the right side of the dialog, there are three buttons: 'OK', 'Cancel', and 'Help'.

Descriptive Statistics


Input

Input Range:  

Grouped By: ☒ Columns ☐ Rows

☐ Labels in First Row

Output options

☐ Output Range:  

☒ New Worksheet Ply:

☐ New Workbook

☐ Summary statistics

☐ Confidence Level for Mean:  %

☐ Kth Largest:

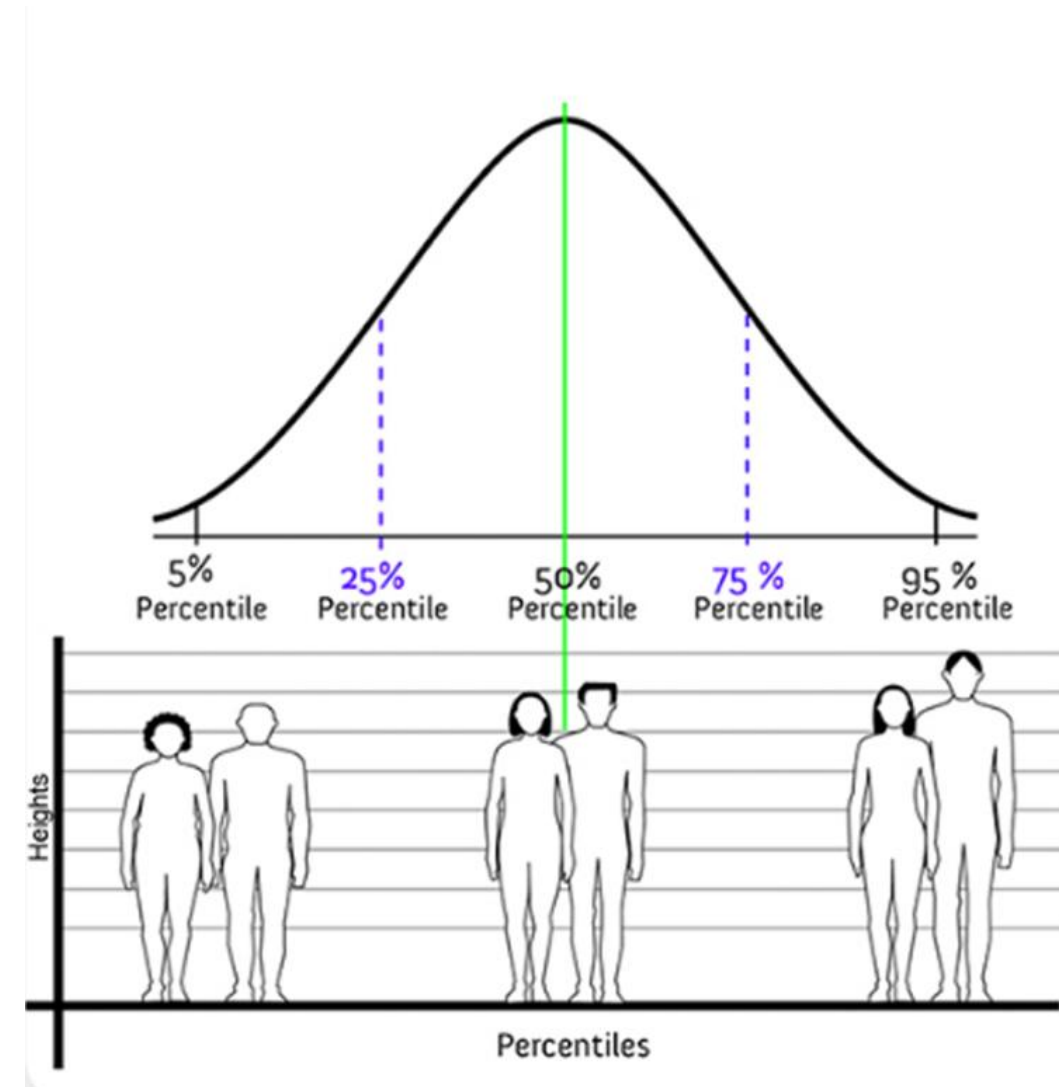
☐ Kth Smallest:

OK Cancel Help

# Percentiles, Quartiles and Boxplots

# Percentiles

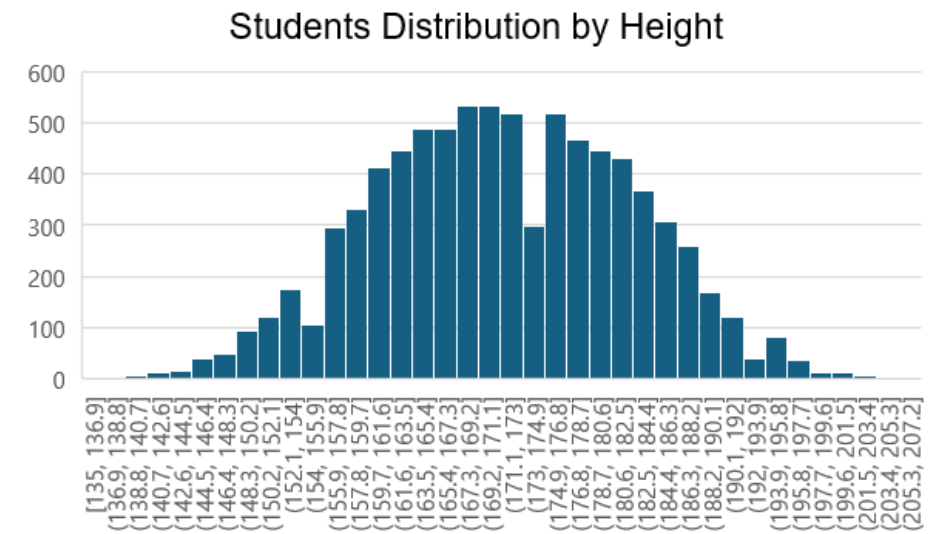
- A **percentile** is a measure that helps us understand the position or relative standing of a particular value within a dataset.
- Percentiles help us divide the data into equal-sized groups and determine the position of a value within those groups.
- For example, if someone's score is at the 80th percentile, it means they performed better than 80% of the people in the dataset.
- Percentiles provide insights into how a specific value compares to others, allowing us to **assess performance, rankings, or distributions**.
- They are commonly used in various fields, such as education, healthcare, and finance, to understand and communicate data in a meaningful way.



# Practical Exercise

- Using the students.xlsx file, find the 90% Percentile value for the “height” column.
- You can use percentile.inc or percentile.exc to compute percentiles in Excel.

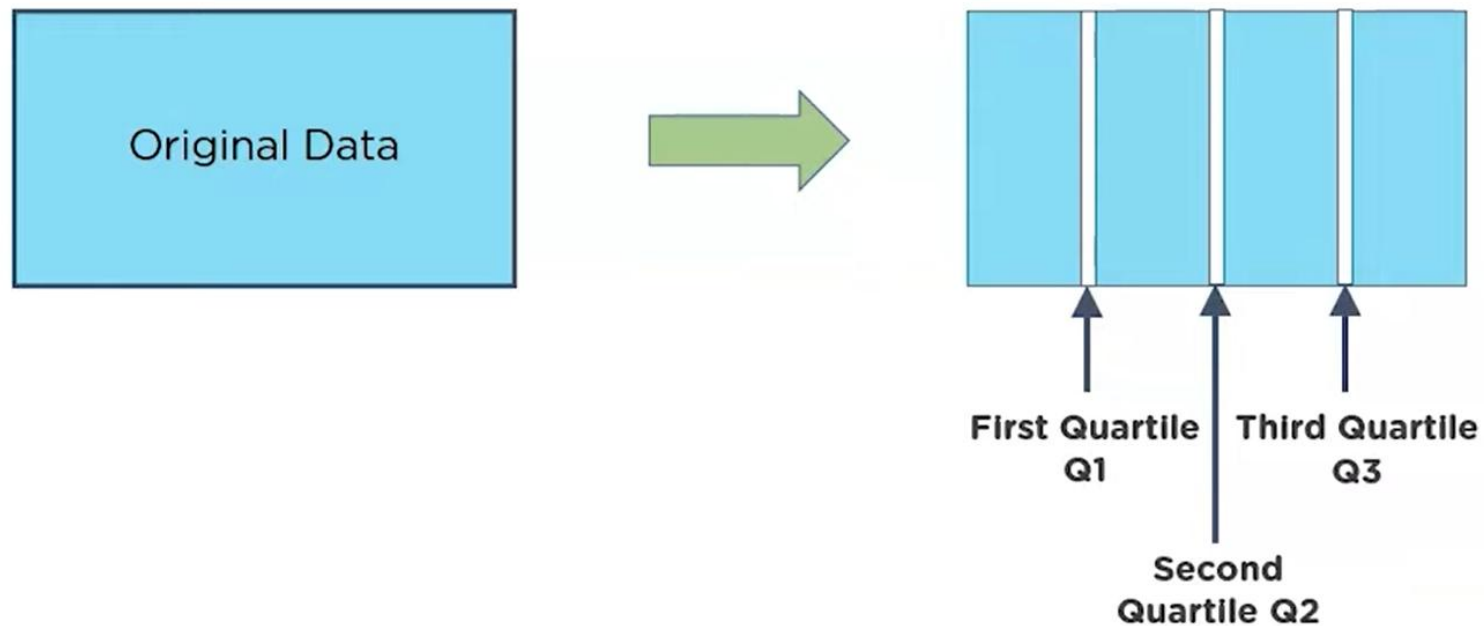
|    | A   | B      | C      | D | E              | F   | G   |
|----|-----|--------|--------|---|----------------|---|-----|
| 1  | age | height | weight |   |                |   |     |
| 2  | 19  | 160    | 64.8   |   |                |   |     |
| 3  | 19  | 172    | 73     |   |                |   |     |
| 4  | 22  | 168    | 70.6   |   |                | <b>90% Percentile for Students Height</b> |     |
| 5  | 19  | 183    | 79.7   |   | percentile.inc | =PERCENTILE.INC(Table2[height],0.9)       | 186 |
| 6  | 21  | 175    | 71.4   |   | percentile.ex  | =PERCENTILE.EXC(Table2[height],0.9)       | 186 |
| 7  | 19  | 189    | 85.8   |   |                |   |     |
| 8  | 21  | 156    | 65.9   |   |                |   |     |
| 9  | 21  | 167    | 65.7   |   |                |   |     |
| 10 | 18  | 195    | 94.4   |   |                |   |     |





# Quartiles

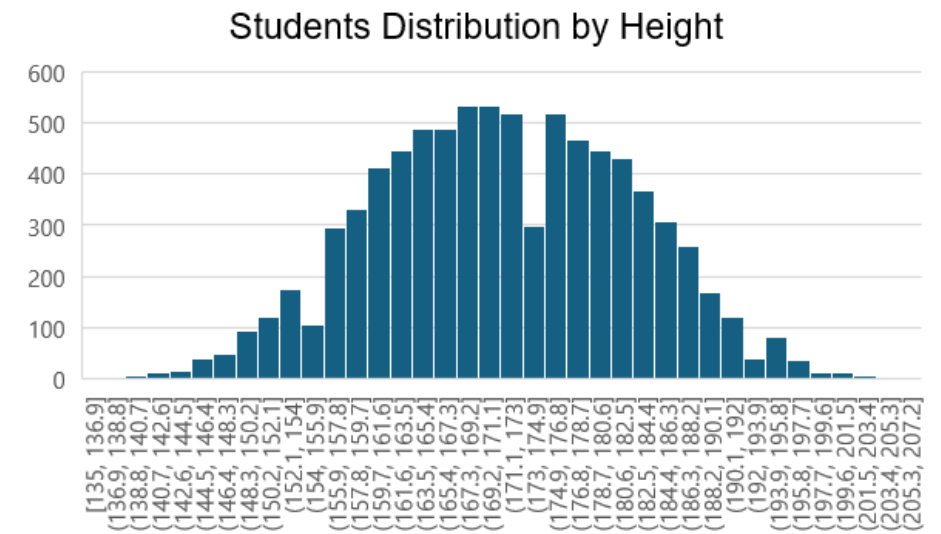
A Quartile divides a **sorted** data set into 4 equal parts, so that each part represents  $\frac{1}{4}$  of the data set.



# Practical Exercise

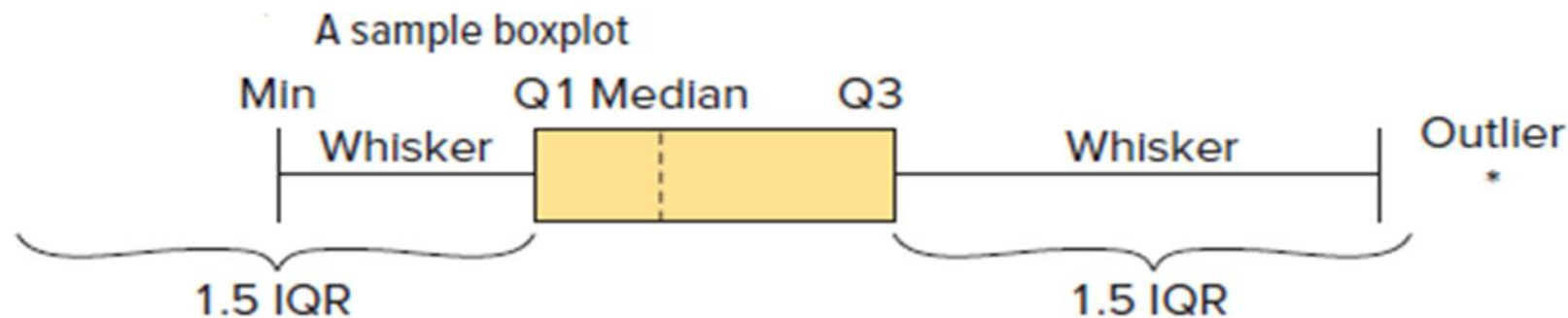
- Using the students.xlsx file, find the 1<sup>st</sup> and the 2<sup>nd</sup> quartiles for the “height” column.
- You can use quartile.inc or quartile.exc to compute percentiles in Excel.

|   | A      | B                | C                         | D | E   |
|---|--------|------------------|---------------------------|---|-----|
| 1 | height |                  |                           |   |     |
| 2 | 160    | 1st Quartile INC | =QUARTILE.INC(A2:A8240,1) |   | 163 |
| 3 | 172    | 2nd Quartile EXC | =QUARTILE.EXC(A3:A8241,2) |   | 171 |
| 4 | 168    |                  |                           |   |     |
| 5 | 183    |                  |                           |   |     |



## Boxplots – Box and Whiskers Plots

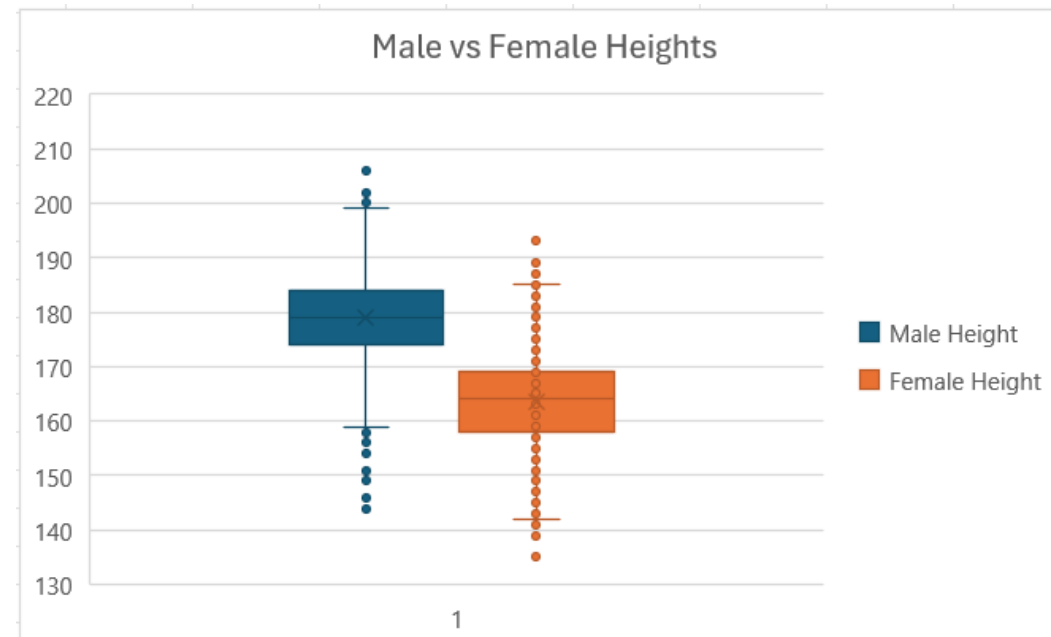
- A boxplot (box-and-whisker plot) is a standardized way of displaying the distribution of data based on a five number summary.
- The five number summary include the minimum value (Min), quartiles (Q1, Q2, and Q3), and the maximum value (Max).
- Boxplots are useful in showing us the range and the skewness of our data as well as outliers.



# Practical Exercise

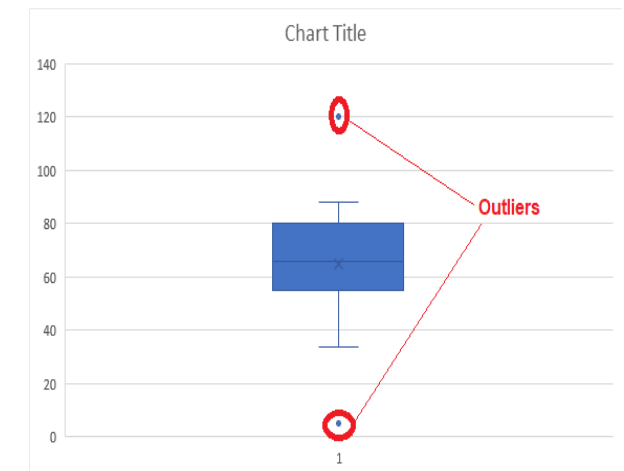
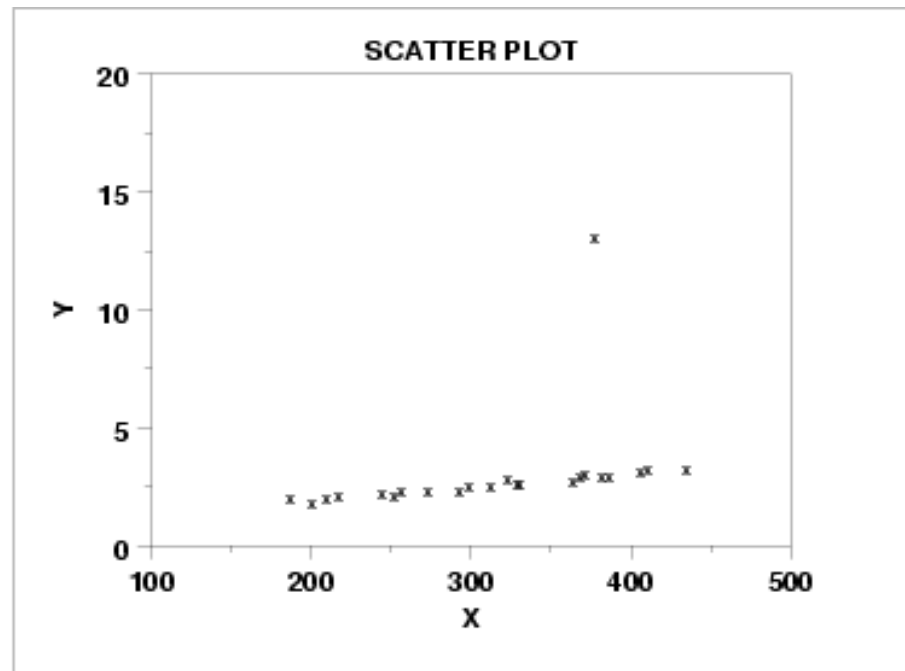
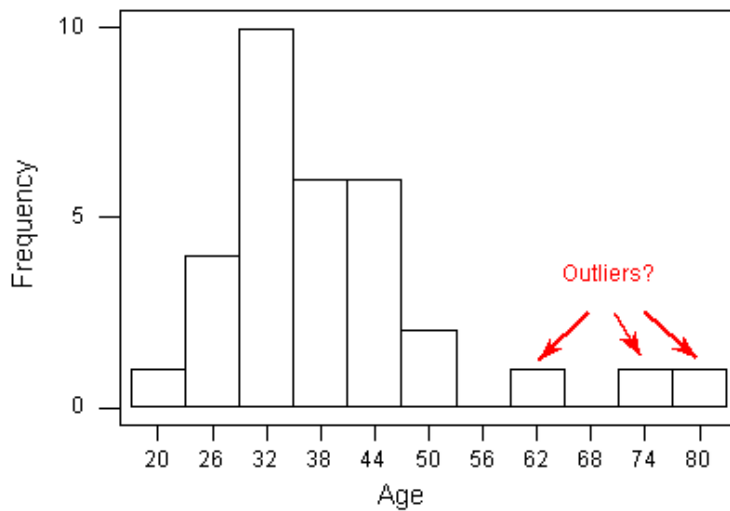
Use the students.xlsx file:

- Create a box plot that compares the height of male students vs female students.
- Create a box plot that compares the height of Turkish male students vs Russian male students.



# Outliers

- An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.
- We can identify potential outliers using Histograms, scatter plots, boxplots...etc.



## The Interquartile Range (IQR)

- The IQR is like the range, and it shows how the data is spread about the median.
- It is less susceptible than the range to outliers and can, therefore, be more helpful.
- The IQR is calculated according to the following formula:

$$\text{IQR} = Q3 - Q1.$$

# Using the IQR to Find Outliers

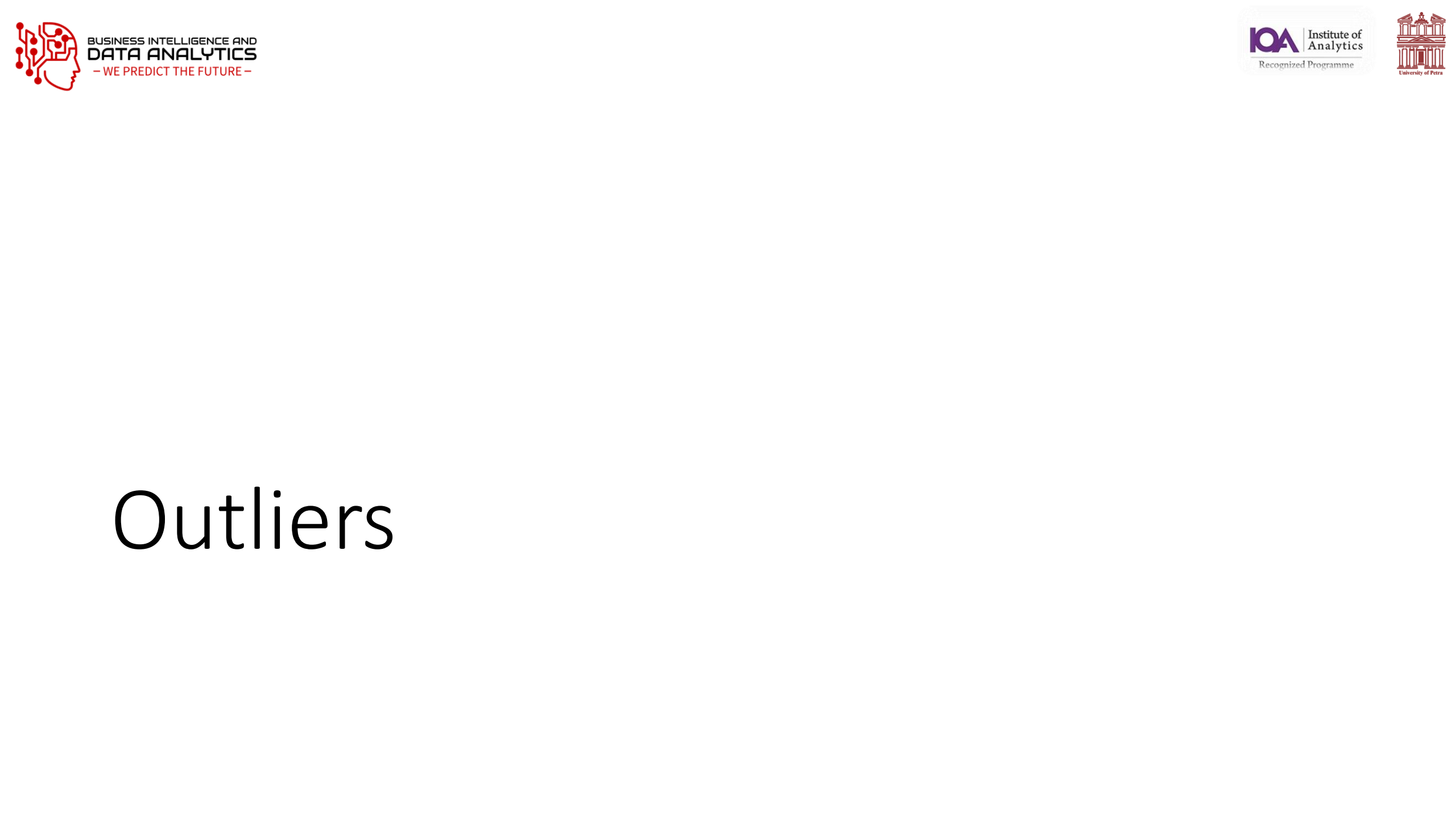
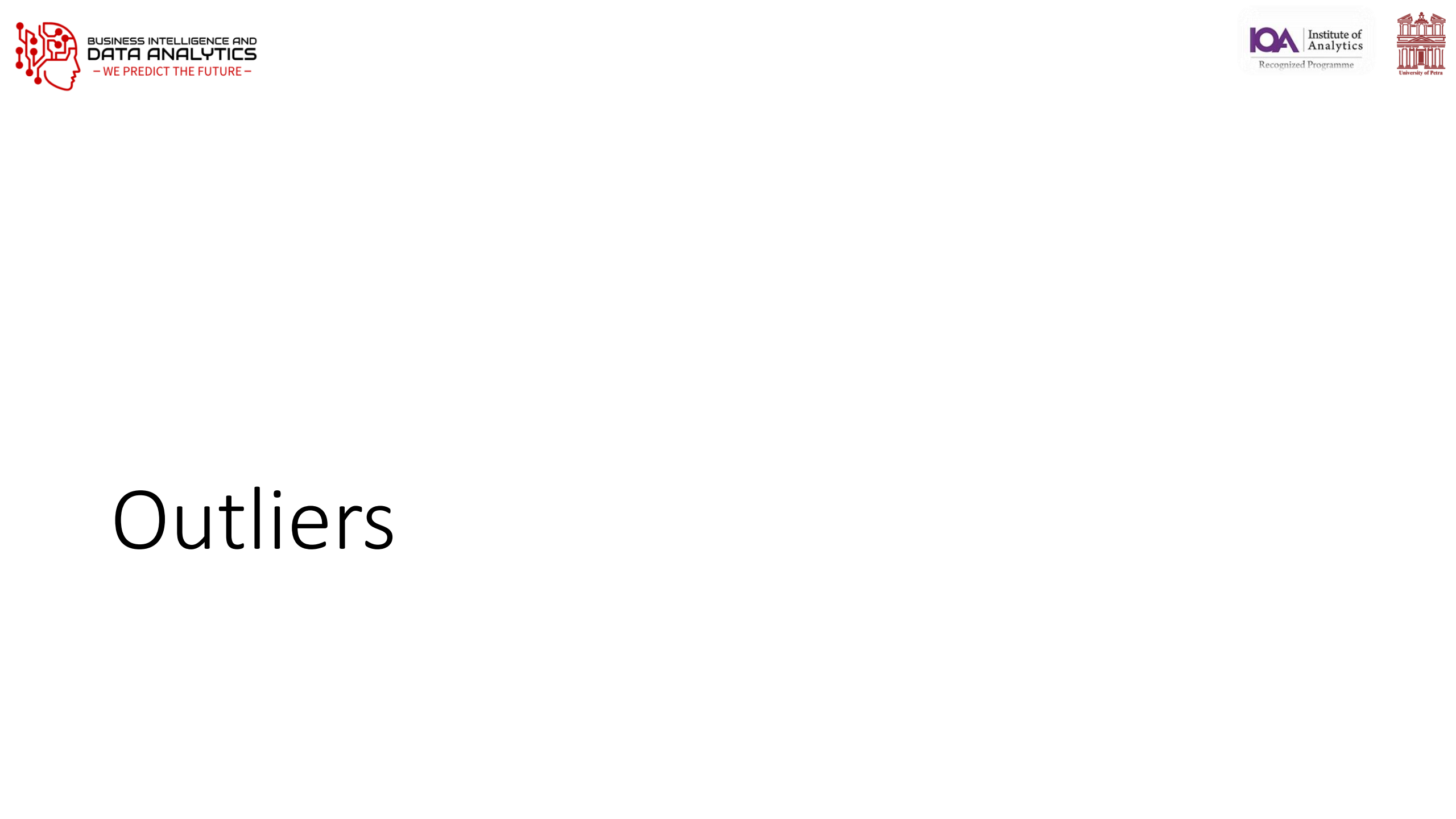
Using the IQR to find potential outliers:

1. Calculate the IQR for the data.
2. Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers).
3. Add the  $(1.5 \times \text{IQR})$  to the third quartile, any number greater than this is a suspected outlier.
4. Subtract the  $(1.5 \times \text{IQR})$  from the first quartile, any number less than this is a suspected outlier.

However, any potential outlier obtained by the interquartile method should be examined in the context of the entire set of data.

We should always follow up our outlier analysis by studying the resulting outliers to see if they make sense.

|   | A      | B      | C | D   | E   | F   |
|---|--------|--------|---|---|---|-----|
| 1 | gender | height |   |   |   |     |
| 2 | Male   | 183    |   | Height of Male Students                           |   |     |
| 3 | Male   | 189    |   | Q1  | =QUARTILE.INC(Table3[ [#All],[height]],1) | 174 |
| 4 | Male   | 195    |   | Q3  | =QUARTILE.EXC(Table3[ [#All],[height]],3) | 184 |
| 5 | Male   | 185    |   | IQR = Q3 - Q1                                     | =F4-F3                                    | 10  |
| 6 | Male   | 172    |   | Upper Bound = $1.5 \times \text{IQR} + \text{Q3}$ | =1.5*F5+F4                                | 199 |
| 7 | Male   | 182    |   | Lower Bound = $\text{Q1} - 1.5 \times \text{IQR}$ | =F3-1.5*F5                                | 159 |
| 8 | Male   | 175    |   |   |   |     |
| 9 | Male   | 164    |   |   |   |     |



# Outliers



# Impacts of Outliers

## 1. Negative Impacts

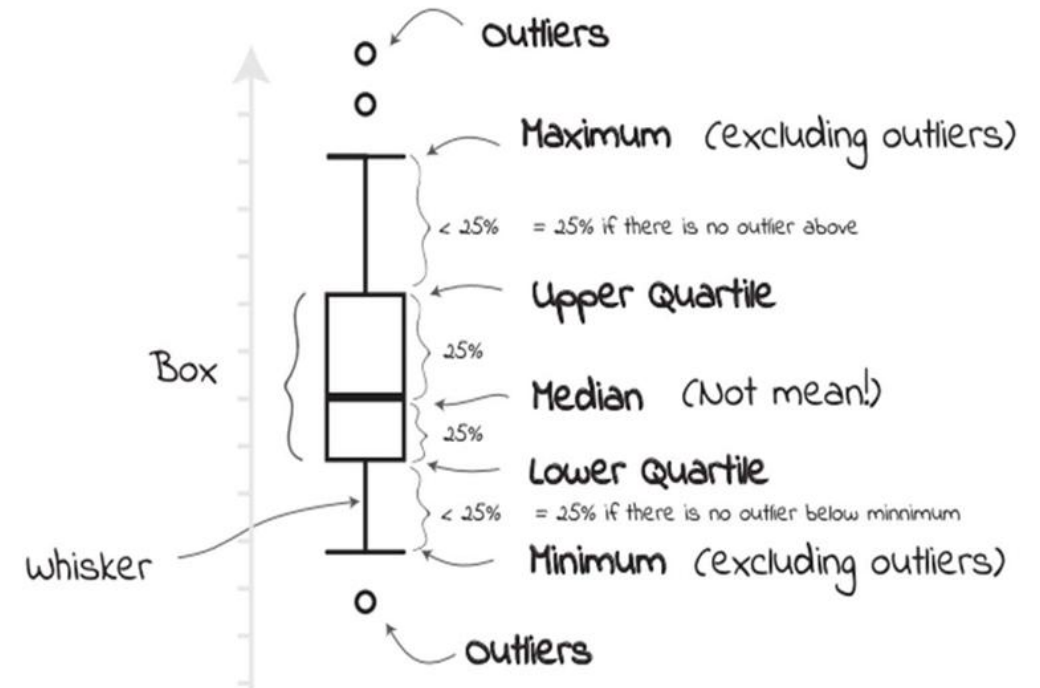
- Influence on Measures of Central Tendency: A single extreme outlier can pull the mean in its direction, making it unrepresentative of the majority of the data.
- Impact on Dispersion Measures: The presence of outliers can inflate the measures like the standard deviation and the interquartile range (IQR), making them larger than they would be without outliers.
- Skewing Data Distributions: Positive outliers can result in right-skewed distributions, while negative outliers can result in left-skewed distributions. This can affect the interpretation of the data.
- Misleading Summary Statistics: Outliers can distort the interpretation of summary statistics.
- Impact on Hypothesis Testing: Outliers can affect the results of hypothesis tests. They can lead to incorrect conclusions, such as detecting significant differences when none exist or failing to detect real differences when outliers mask them.

## 2. Positive Impacts

- Detection of Anomalies: Outliers can signal the presence of anomalies or rare events in a dataset. Identifying these anomalies can be valuable in various fields, including fraud detection, quality control, and outlier detection in scientific experiments.
- Robust Modeling: In some cases, outliers can be genuine observations that are important to model. For example, in financial modeling, extreme stock price movements may contain valuable information for predicting market trends.

# Methods to screen for outliers in a dataset

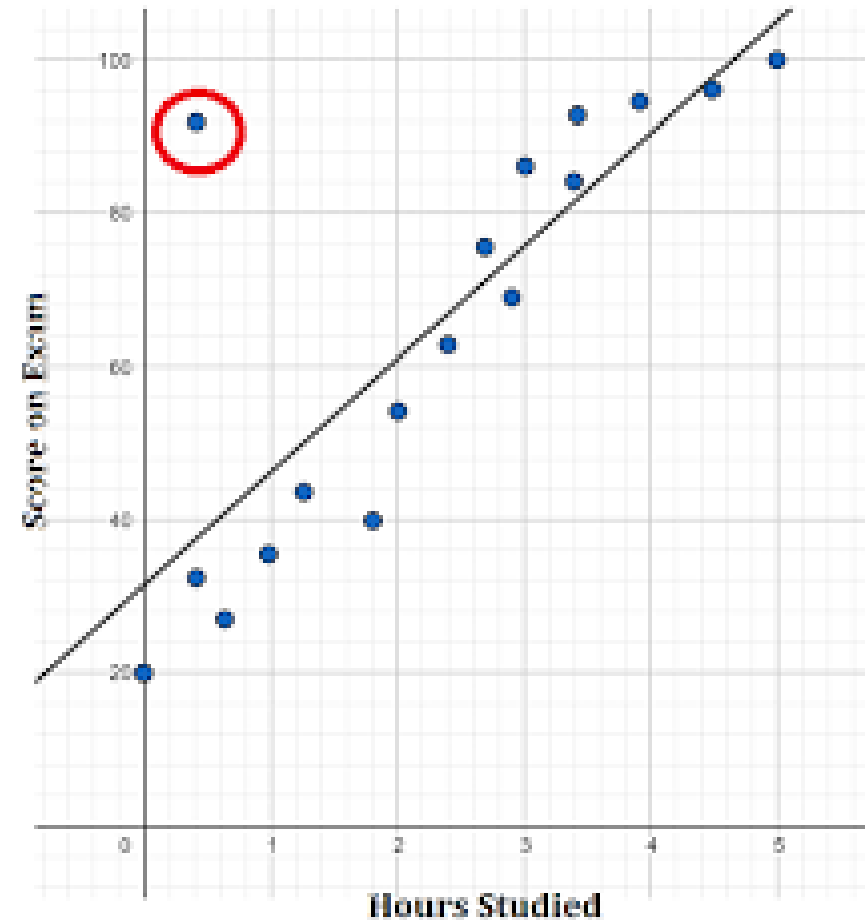
- There are several methods to screen for outliers in a dataset, ranging from graphical techniques to statistical tests. Here are some commonly used methods:
- Box Plots (Box-and-Whisker Plots):  
Box plots provide a visual representation of the distribution of data, including the identification of potential outliers. In a box plot, outliers are typically shown as individual data points beyond the whiskers of the plot.



# Methods to screen for outliers in a dataset

## Scatterplots:

- Scatterplots are particularly useful for identifying outliers in bivariate or multivariate data.
- Outliers can appear as data points that are far from the main cluster of points in the scatterplot.



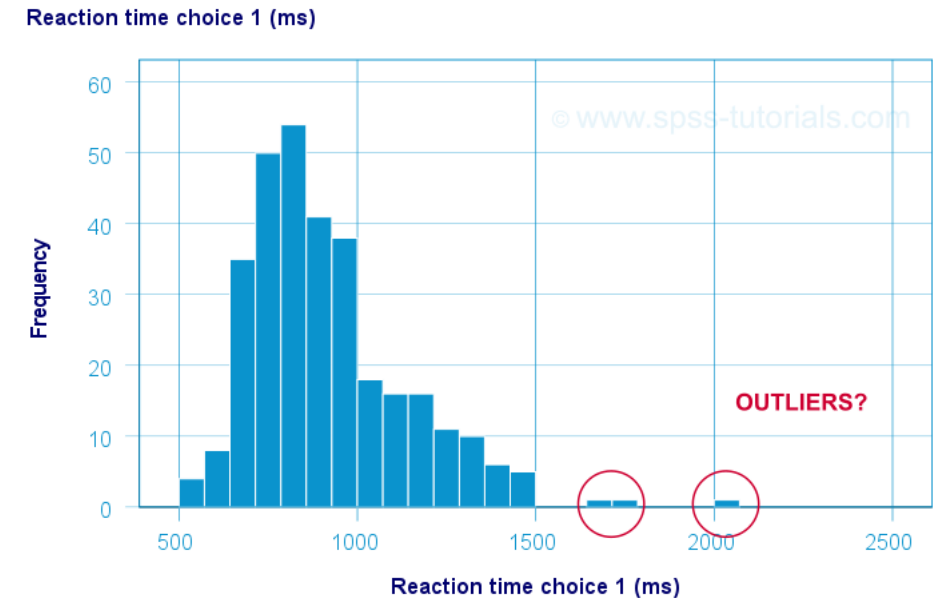
# Methods to screen for outliers in a dataset

- Z-Scores:  
Z-scores (standard scores) measure how many standard deviations a data point is away from the mean. Data points with high absolute Z-scores (typically greater than 2 or 3) are often considered potential outliers.
- IQR (Interquartile Range) Method:  
The IQR method involves calculating the interquartile range ( $IQR = Q3 - Q1$ ) and then identifying values that fall below  $Q1 - 1.5 \times IQR$  or above  $Q3 + 1.5 \times IQR$  as potential outliers.

# Methods to screen for outliers in a dataset

## Visual Inspection:

- Sometimes, simple visual inspection of the data through histograms, QQ plots (quantile-quantile plots), or other visualization techniques can reveal the presence of outliers.



# Handling Outliers in Datasets

- Handling outliers is an important step in data preprocessing to ensure they do not unduly influence the results of analysis or modeling. The approach depends on the nature of the data, the context of the analysis, and the specific objectives. Here are several methods for handling outliers:
- **Data Truncation or Removal:**
  - Simply remove outliers from the dataset.
  - Should be done cautiously, especially if the outliers represent valid and important observations.
  - Appropriate when outliers are likely due to data entry or measurement errors.
- **Data Transformation:**
  - Transforming the data can mitigate the impact of outliers.
  - Common transformations: logarithmic, square root, or inverse.
  - These compress the range of extreme values.
- **Winsorization:**
  - Involves capping or limiting extreme values by replacing them with a specified percentile value.
  - Example: Replace values above the 95th percentile with the value at the 95th percentile.
- **Imputation:**
  - For missing values (not extreme outliers), impute using methods like:
    - Mean imputation
    - Median imputation
    - Advanced methods (e.g., regression imputation).

# Handling Outliers in Datasets

- **Robust Statistics:**

- Use statistical methods less sensitive to outliers.
- Examples:
  - Replace mean with median.
  - Use interquartile range (IQR) instead of standard deviation.

- **Model-Based Approaches:**

- In predictive modeling, use algorithms less sensitive to outliers.
- Examples: robust regression methods, ensemble methods (e.g., random forests).

- **Domain Knowledge:**

- Rely on domain expertise to interpret outliers.
- What looks like an outlier may actually be an important and valid data point.

- **Reporting and Transparency:**

- Clearly document how outliers were handled.
- Ensures reproducibility and interpretability of results.