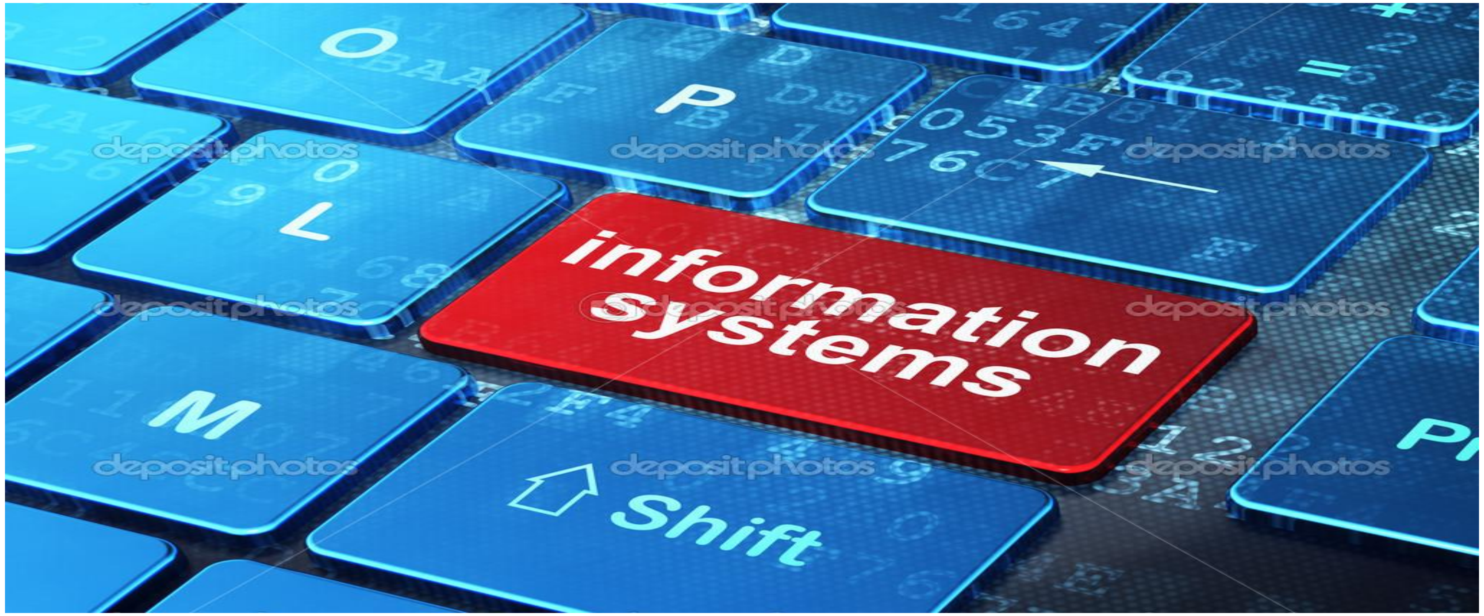
# Advanced Topics in Information systems



## SE204

Dr. Nelly Amer

# Grades

- 60% on :

Attendance

Assignments

Quizzes

Midterm Exam

- Final term Exam: 40 %

# Syllabus

- **Descriptive data analysis and visualization.**
- **Frequent pattern mining.**
- **Cluster and dimensionality reduction.**
- **Machine learning 1: Introduction.**
- **Machine learning 2: Classification.**
- **Machine learning 3: ROC analysis and regression.**
- **Machine learning 4: Deep learning.**

# Descriptive data analysis and visualization.

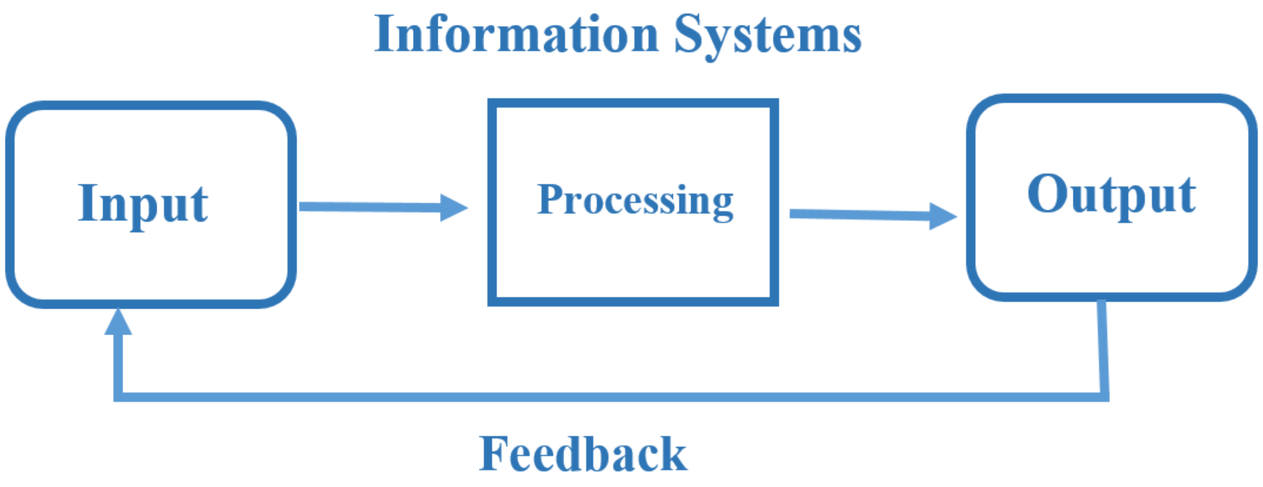
- **Information system definition and elements.**
- **Data, information, and knowledge.**
- **Types of data.**
- **Data analysis.**
- **Data analysis and data analytics.**
- **Data analysis techniques.**
- **Descriptive data analysis methods.**
- **Descriptive data analysis phases.**
- **Data visualization.**
- **Descriptive data analysis and visualization tools.**

# Information system definition and elements

## Information system definition.

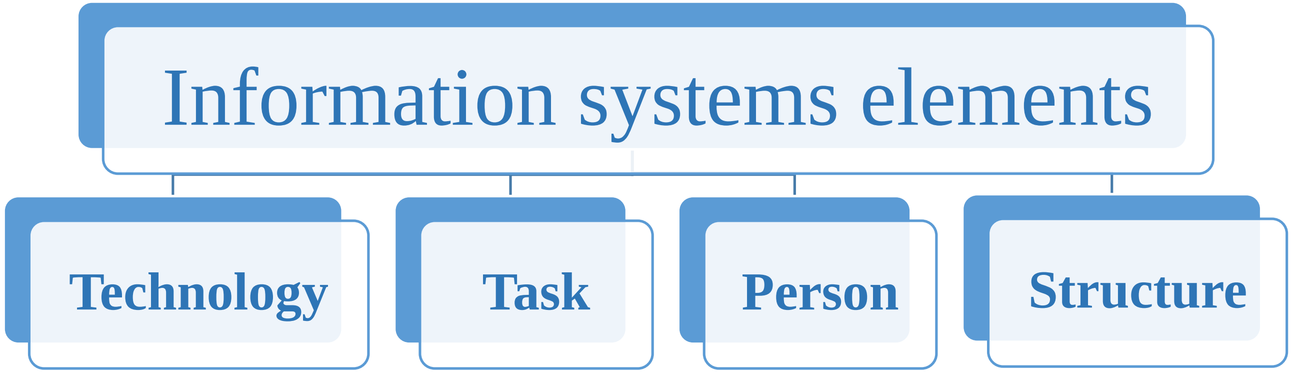
An information system is a set of interrelated components that collect, manipulate, store and disseminate information and provide a feedback mechanism to achieve a goal.

The feedback mechanism helps organizations achieve their goals by increasing profits, revenues, improving customer service, reduce costs, and supporting decision-making and control in organizations [1].



# Information system definition and elements cont.

## Information systems elements.



**Technology**: includes the hardware, software, and telecommunications equipment used to capture, process, store and disseminate information.

**Task**: activities necessary for the production of a good or service. These activities are supported by the flow of material, information, and knowledge between the different participants.

**Person**: The people component of an information system encompasses all the people directly involved in the system. These people include the **managers** who define the goals of the system, the **users**, the **data analysts**, and the **developers**.

**Structure**: the relationship between individuals people components. Thus, it encompasses hierarchical structures, relationships, and systems for evaluating people[1].

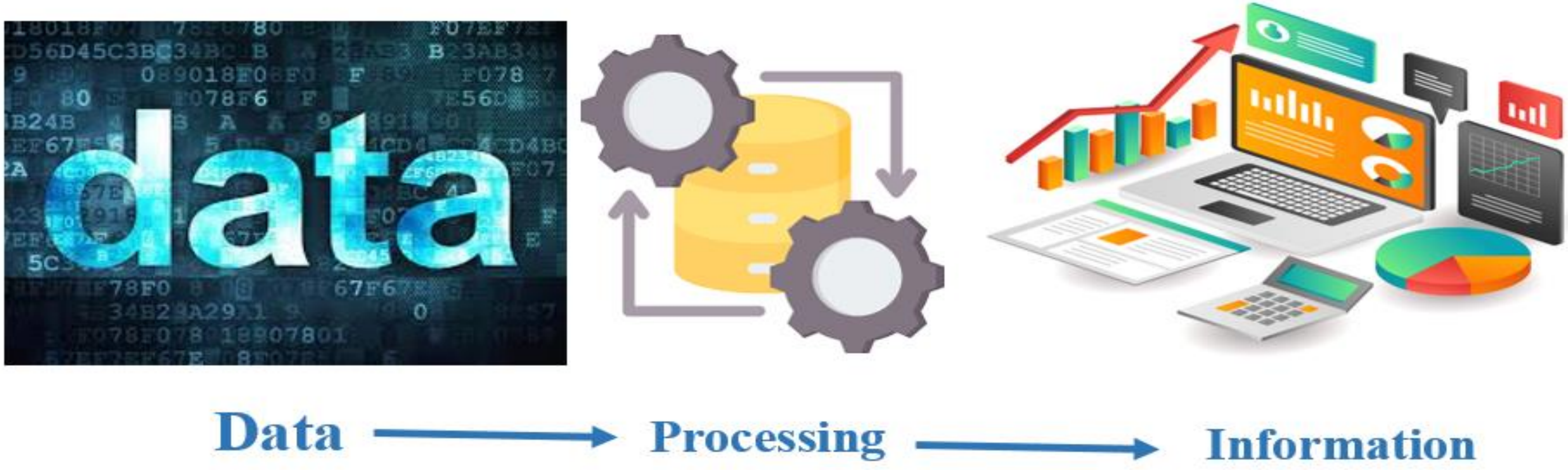
# Data, information, and knowledge.

## Data

Data is a collection of raw, unorganized facts and details like text, observations, figures, symbols and descriptions of things measured in terms of bits and bytes. Data does not carry any specific purpose or significance on its own.

## Information

Data become information  when analyzed, organized or classified, and possibly combined with other data in order to extract meaning, and to provide context, it provides answers to the questions: "who," "what," "where," and "when." it must be timely, accurate, and complete[2].



# Data, information, and knowledge cont.

## Knowledge.

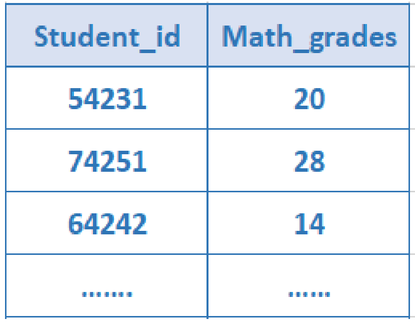
Knowledge can be defined as information combined with experience, context, and interpretation. Knowledge constitutes an additional semantic level derived from information, it provides answers to the question "how".

**Data, information, and Knowledge.**

**Data**: degrees of students in schools in final math exam in 6th grade, primary.

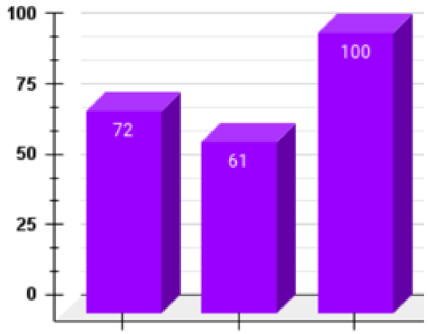
**Information**: percentage of success in math exam for school.

**Knowledge:** Schools with high success rates have qualified teachers in math, and vice versa.



| Student_id | Math_grades |
|---|---|
| 54231 | 20 |
| 74251 | 28 |
| 64242 | 14 |
| ……. | …… |

**Data**



**Information**

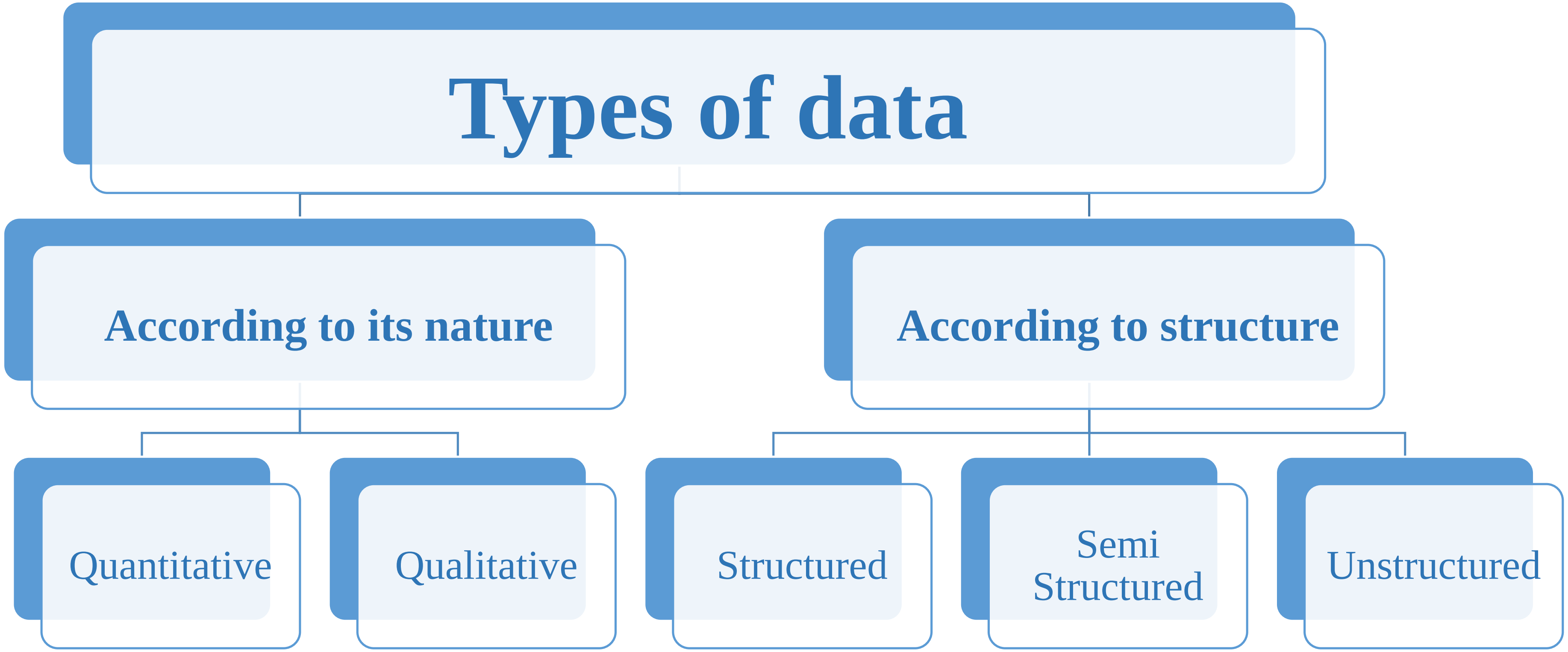
**High grades, how?**

**By qualified teachers**

**Knowledge**

# Types of data.

Data types can be classified according to its nature and according to structure.



Types of data

According to its nature

According to structure

Quantitative

Qualitative

Structured

Semi Structured

Unstructured

# Types of data According to its nature.

Data can be classified according to its nature into : quantitative data, and qualitative data [4].

## Quantitative (numeric)

Data that can be quantified and measured. This kind of data explains a trend or the results of research through numeric values.

This category of data can be subdivided into:

- **Discrete**: Data that consists of whole numbers (0, 1, 2, 3...). For example, the number of children in a family.
- **Continuous**: Data that can take any value within an interval. For example, people’s height (between 60 - 70 inches) or weight (between 90 and 110 pounds).

# Types of data According to its nature cont.

## Qualitative (categorical)

This kind of data is divided into categories based on non-numeric characteristics. It may or may not have a logical order, and it measures qualities and generates categorical answers.

It can be subdivided into :

- **Ordinal**: Meaning it follows an order or sequence. That might be the alphabet or the months of the year.
- **Categorical**: Meaning it follows no fixed order. For example, varieties of products sold.

# Types of data according to structure.

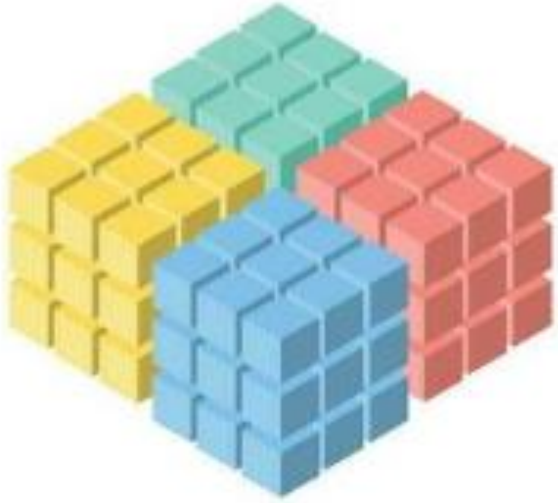
According to structure, data can be classified into : Structured data, semi structured data, and unstructured data.

Structured Data
Semi-Structured Data
Unstructured Data

# Types of data cont.

## Structured data.

- Structured data is data that can be stored in a table (rows and columns), and every instance in the table has the same structure (i.e., set of attributes.)
- Structured data is typically stored in a relational database (RDBMS) such as SQL Server, Oracle, and MySQL.
- Structured data can be managed using SQL (Structured Query Language).
- Structured data is often quantitative data, i.e., it usually consists of numbers or things that can be counted.
- Examples of structured data: consider the demographic data for a population, where each row in the table describes one person and consists of the same set of demographic attributes (name, age, date of birth, address, gender, education level, job status, credit card numbers, etc.) [3].



# Types of data cont.

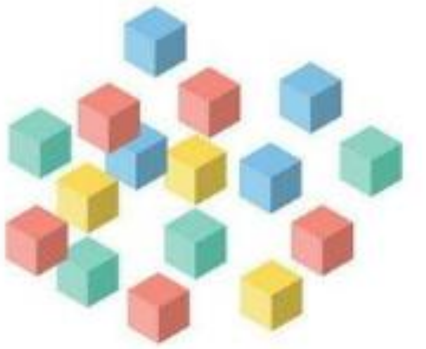
## Semi structured data.

- This is the data which has some structure but there is no a data model, For example, a data set of webpages, with each webpage having a structure but this structure differing from one webpage to another.
- Semi structured data can be stored in NoSQL Databases; for non relational databases, such as MongoDB, Cassandra, and Couchbase.
- Semi structured data can be qualitative and quantitative data.
- Semi structured data examples: collections of human text (emails, tweets, text messages, posts, novels, etc.), XML(Extensible Markup Language) files, and JSON (**J**ava**S**cript **O**bject **N**otation) files[5].

# Types of data cont.

## Unstructured data.

- Unstructured data are data where each instance in the data set may have its own internal structure, and this structure is not necessarily the same in every instance, there is no data model, the data is stored in its native format.
- The unstructured data can be stored in NoSQL Databases; for non relational databases, such as MongoDB, Cassandra, and Couchbase.
- Structured data can be extracted from unstructured data using techniques from artificial intelligence (such as natural language processing and ML), digital signal processing, and computer vision.
- Unstructured data is often qualitative data, it can not be processed using conventional tools and methods.
- Unstructured data examples: sound, image, music, video, and multimedia files,  audio files, and various other formats[5].



# Data analysis

## Data analysis definition.

data analysis is the process of identifying, cleaning, transforming, and modeling data to discover meaningful and useful information. The data is then crafted into a story via reports for analysis to support the critical decision-making process [6].

The two terms: data analysis and data analytics are often used interchangeably and could be confusing, and It’s a common misconception that data analysis and data analytics are the same thing. **No they are not the same thing**, as we will show.

# Data analysis and data analytics.

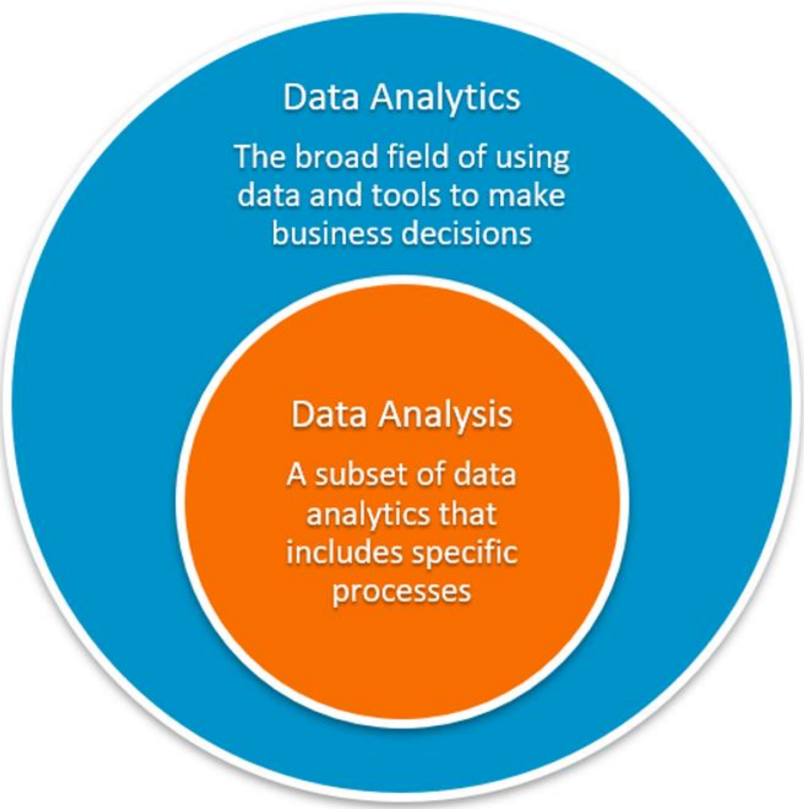
Data analytics and data analysis are closely related processes that involve extracting insights from data to make informed decisions, but there are some  important differences between analysis and analytics [6].

**Data analysis** consists of cleaning, manipulating data, modeling, and questioning data to discover relevant information and gain **insights**, it is a vital part of data analytics.

**Data analytics** refers to a broad range of data-related activities and concepts. It is a process for translating basic facts and figures into specific **actions** by examining raw data assessments and perceptions in the context of organizational problem-solving and **decision-making.**

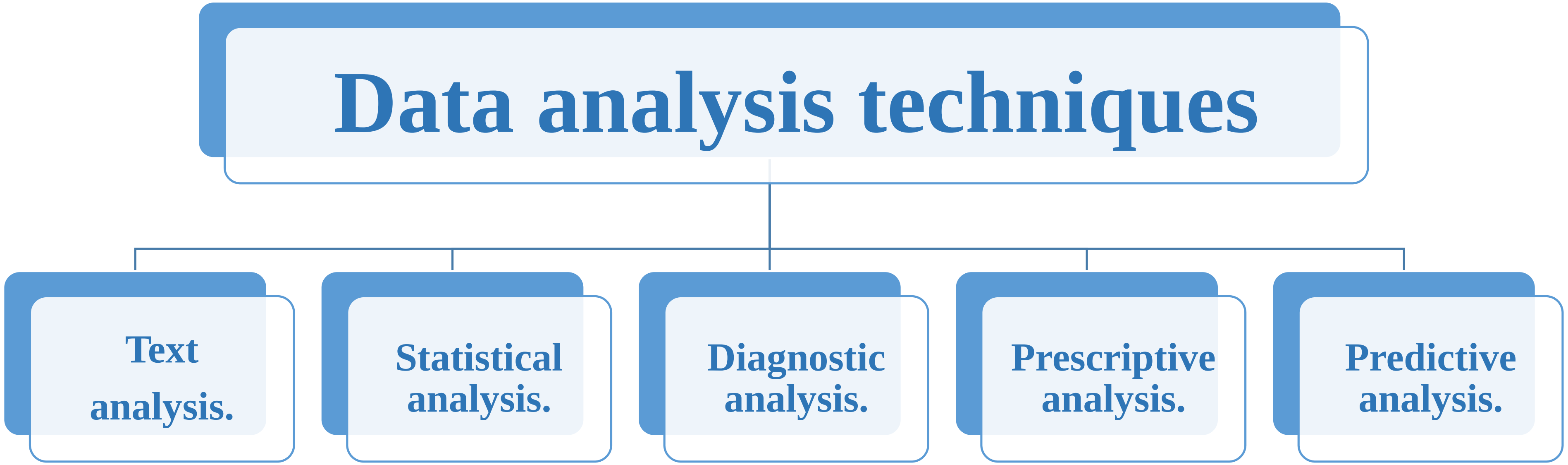
The purpose is to help businesses make better decisions and achieve greater success. Analytics uses data, machine learning, statistical analysis, and computer-based models to gain insight and make better decisions from collected data.

Analysis looks backwards, providing a historical view of what has happened. Analytics, on the other hand, models the future or predicts a result.


Data Analytics
The broad field of using data and tools to make business decisions
Data Analysis
A subset of data analytics that includes specific processes

# Data analysis techniques.

There are several data analysis techniques based on business and technology. The most common data analysis techniques are[7]:



Data analysis techniques

Text analysis.

Statistical analysis.

Diagnostic analysis.

Prescriptive analysis.

Predictive analysis.

# Data analysis techniques cont.

## Text analysis

Text analysis is the process of using computer systems to read and understand human-written text for business insights.

Text analysis **software** can classify, sort, and extract information from text to identify patterns, relationships, and other actionable knowledge.

Text analysis is **important**, where Businesses use text analysis to extract actionable insights from various unstructured data sources as survey responses, emails, call center notes, product reviews, social media posts, product reviews, and any other feedback given in free text.

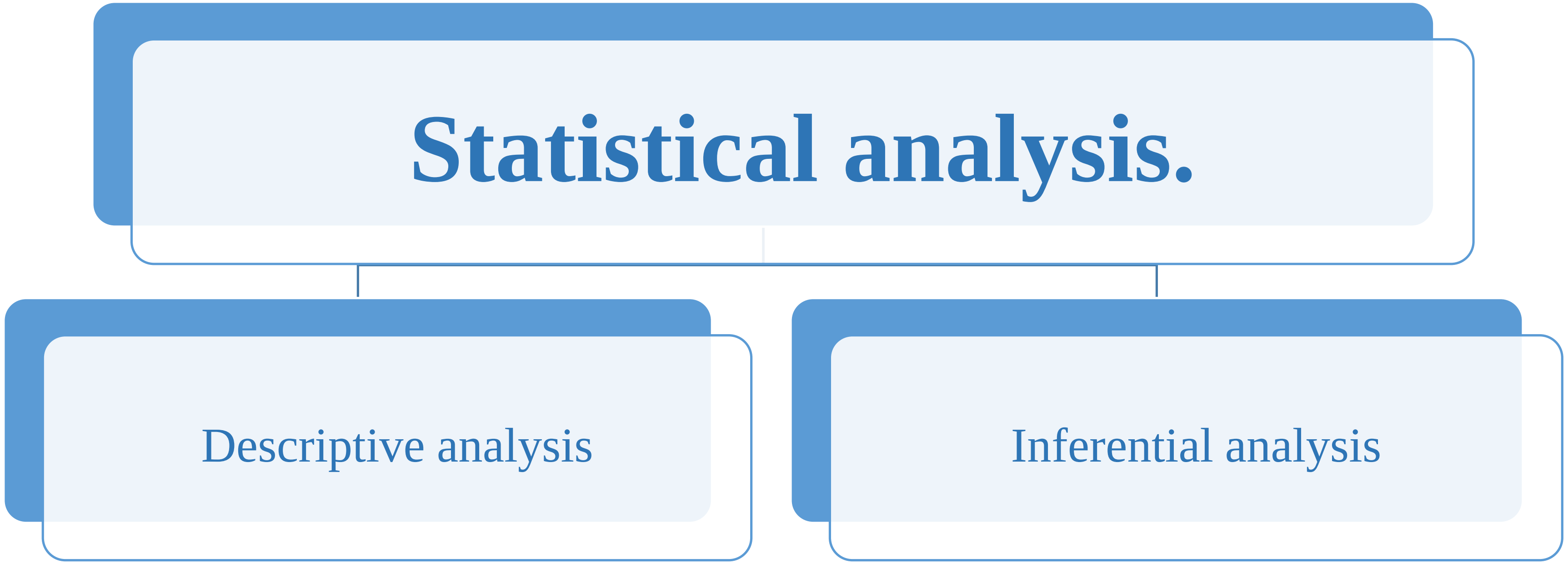
Techni e that comp ers use
to ext act worthwhile inf rmation
to ext act Text wor while
inform tion from the human
languag Analysis in a smart
an efficient mann .

# Data analysis techniques cont.

## Statistical analysis

Statistical Analysis shows "What happened?" by using past data in the form of dashboards. Statistical analysis includes collection, analysis, interpretation, presentation, and modeling of data. **It analyses a set of data or a sample of data.**

There are two categories of Statistical analysis :



Statistical analysis.

Descriptive analysis

Inferential analysis

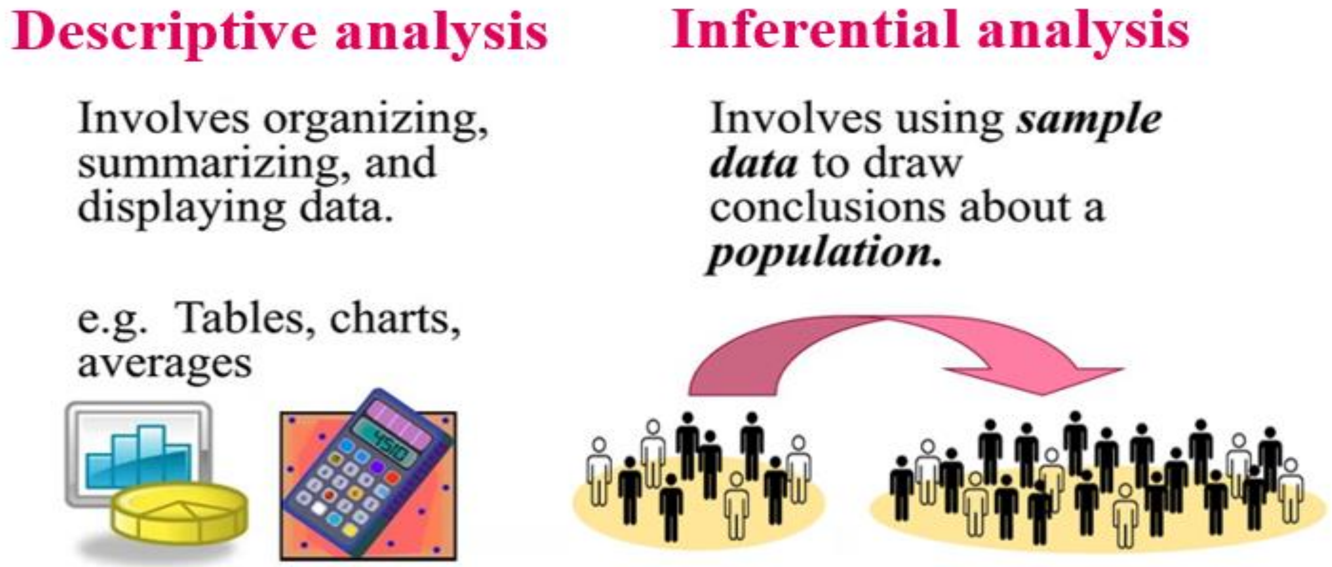
# Statistical Analysis cont.

## Descriptive analysis:

Descriptive analysis answers the question, “What happened?”, **it looks at past data to describe what has happened**. it analyses **complete data or a sample** of summarized numerical data. It shows mean and deviation for continuous data whereas percentage and frequency for categorical data

## Inferential analysis:

It analyses sample from complete data. It means; It uses a small sample to conclude a bigger population.



**Descriptive analysis**

Involves organizing, summarizing, and displaying data.

e.g. Tables, charts, averages

**Inferential analysis**

Involves using ***sample data*** to draw conclusions about a ***population.***

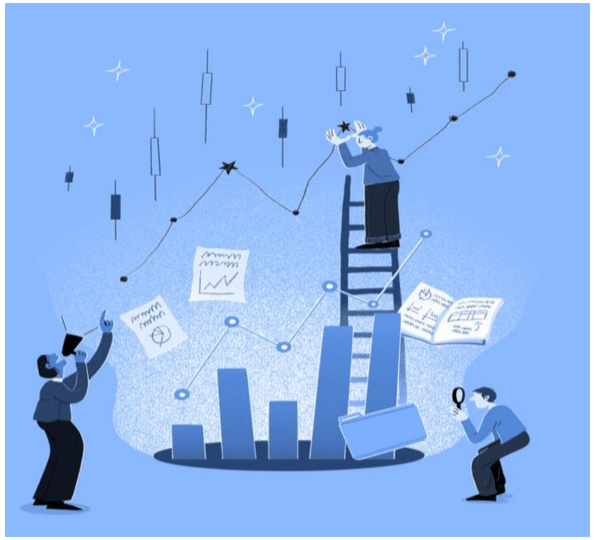
# Data analysis techniques cont.

## Diagnostic analysis

Diagnostic analysis answers the question, **“Why did this happen?”** by finding the cause from the insight found in statistical analysis.

This analysis is useful to identify behavior patterns of data. **If a new problem arrives in your business process**, then you can look into this analysis to find similar patterns of that problem. And it may have chances to use similar prescriptions for the new problems.

## Predictive analysis



Predictive analysis answers the question, “**What might happen in the future?”**

It shows "what is likely to happen" by using previous data, it makes predictions about future outcomes based on current or past data.

# Data analysis techniques cont.

## Prescriptive analysis

Prescriptive analysis answers the question, **“What should we do next?”**, it recommends the best actions to take.

Prescriptive analysis **combines the insight from all previous analysis** to determine **which action to take in a current problem or decision**. Most data-driven companies are utilizing prescriptive analysis because predictive and descriptive analysis are not enough to improve data performance. Based on current situations and problems, they analyze the data and make decisions.

# Data analysis techniques cont.

**Descriptive analysis**

**Example:**

- Looking at sales data to see how sales have changed over time.

**Diagnostic analysis**

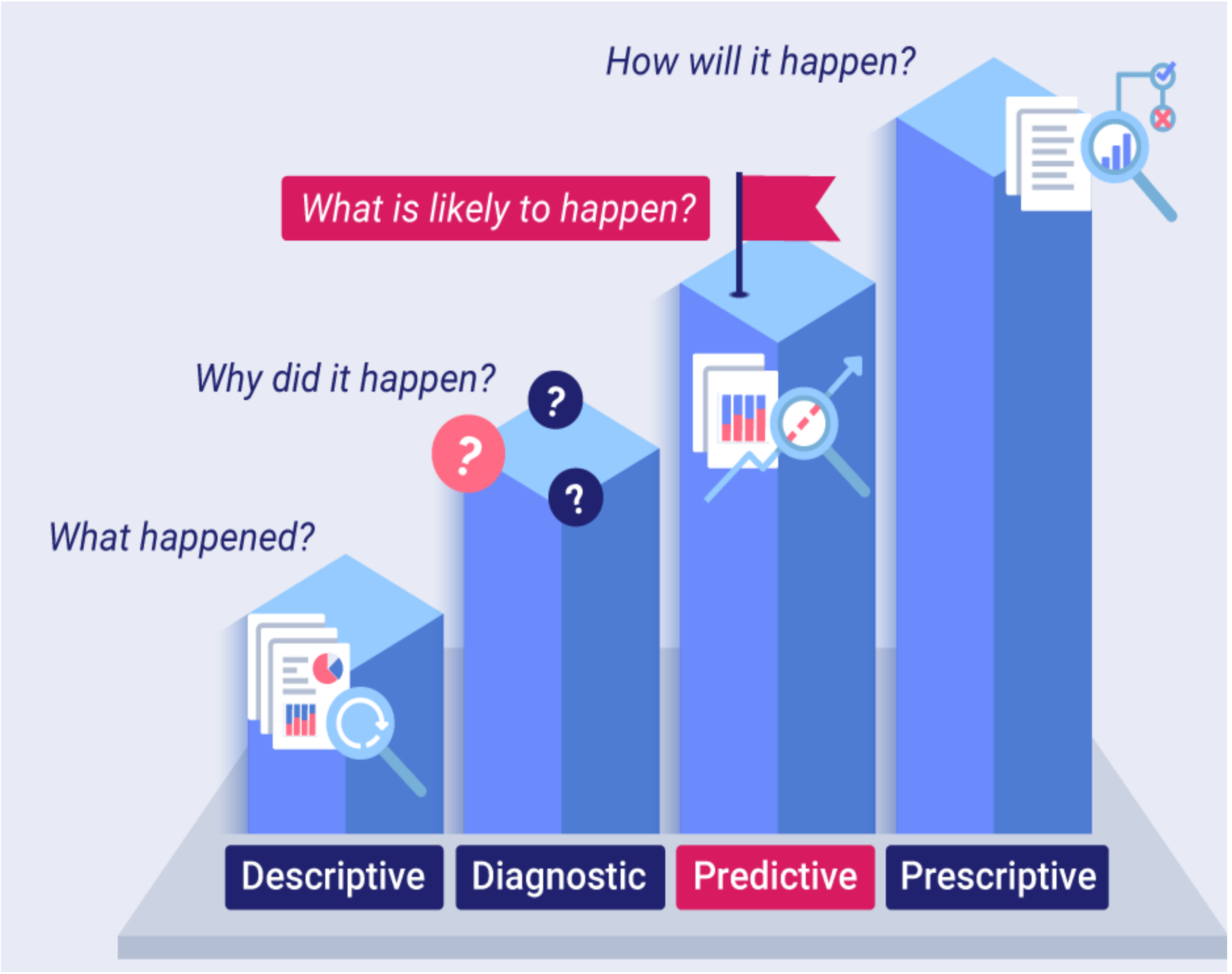
**Example:** Figuring out why sales dropped by looking at different factors.

**Predictive analysis**

**Example:** Predicting next month’s sales based on past trends.

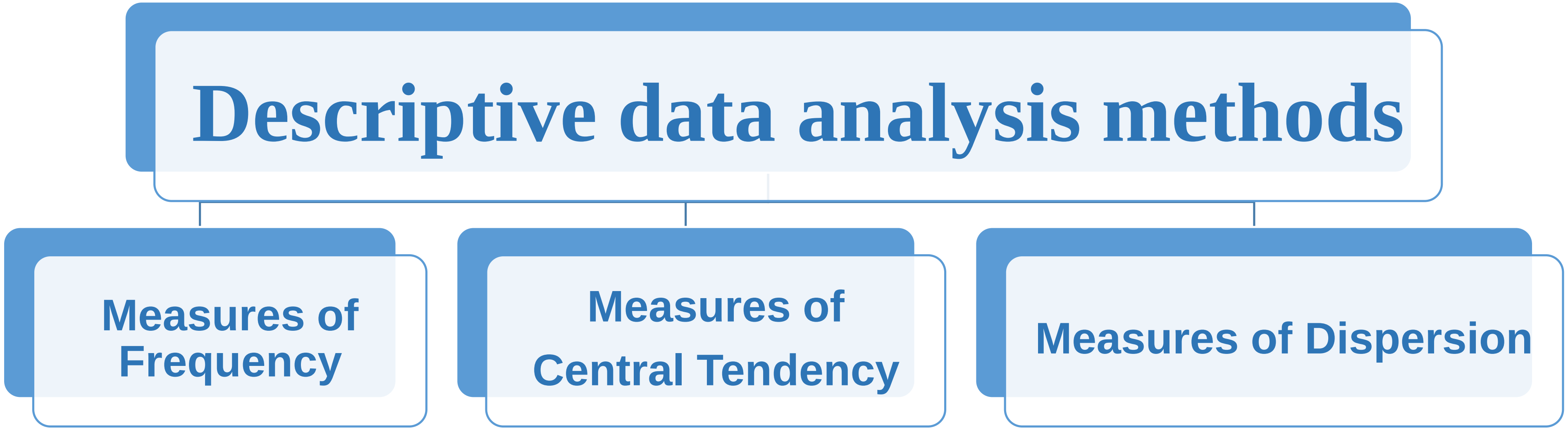
**Prescriptive analysis**

**Example:** Suggesting the best marketing strategy to use for the next campaign.



# Descriptive data analysis methods.

The main descriptive data analysis methods are: measures of frequency, measures of central tendency, and measures of dispersion [6].



Descriptive data analysis methods

Measures of Frequency

Measures of Central Tendency

Measures of Dispersion

# Descriptive data analysis methods cont.

## Measures of Frequency.

The **frequency** of a value is the number of times that value appears in a data set. Measurements of Frequency show how many times each score occurs. The **main goal of frequency** measurements is to provide something like a **count or a percentage.**

**Frequency table** is a tabulation of data values that displays the number of times each value or group of values occurs in the dataset.

Frequency distribution tables can be used for both **categorical and numeric variables**. **Continuous variables** should only be used with class intervals.

Frequency distributions can show either the actual number of observations falling in each range or the percentage of observations. In the latter instance, the distribution is called a **relative frequency distribution**.

Frequency distributions can be **represented** by a histogram, or a pie chart.

# Measurements of Frequency cont.

## Example 1:

Here are the temperatures at midday for 7 days (in ◦C)

**23, 24, 24, 23, 24, 25, 21**

Represent the above data by the frequency table.

| Temperature | Frequency |
|---|---|
| 21 | 1 |
| 23 | 2 |
| 24 | 3 |
| 25 | 1 |
| total | 7 |

# Measurements of Frequency cont.

## Example 2:

The grades obtained in an English test by a class of 15 students are given below.

**A, C, A, B, B, D, F, D, A, D, F, B, C, D, C**

Represent the above-given grades in a **relative frequency table**.
**Solution:**

| Grades | Frequency | Relative frequency |
|---|---|---|
| A | 3 | 3/15*100=20% |
| B | 3 | 3/15*100=20% |
| C | 3 | 3/15*100=20% |
| D | 4 | 4/15*100=26.67% |
| F | 2 | 2/15*100=13.33% |
|  | 15 | 100% |

# Measurements of Frequency cont.

## Example 3:

The following data represents the age of **25** employees .

**36, 32, 48, 41, 38, 28, 37, 30, 58, 44, 34, 38, 43, 50, 40, 45, 55, 59, 29, 43, 32, 39, 48, 57, 46**

**Make a frequency distribution table.**

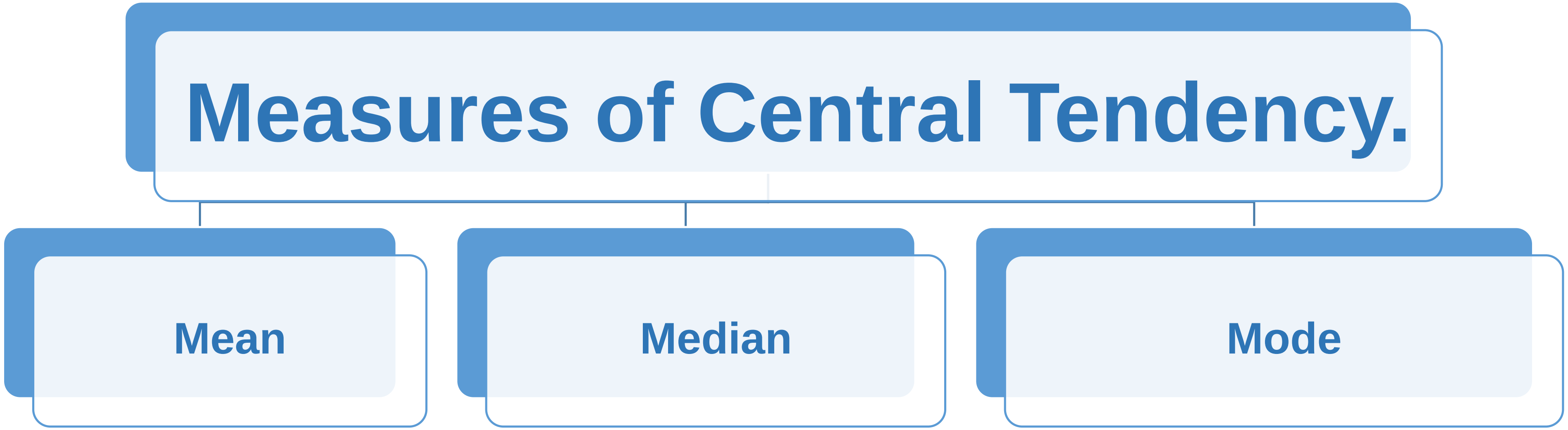
**Solution:**

| Age interval | Frequency |
|---|---|
| 20-30 | 2 |
| 30-40 | 9 |
| 40-50 | 9 |
| 50-60 | 5 |
|  | 25 |

# Measures of Central Tendency.

Finding the central tendency is crucial in descriptive analysis.

Three standards: mean (average), median, and mode are used to calculate central tendency.



Measures of Central Tendency.

Mean

Median

Mode

# Measures of Central Tendency cont.

## Mean.

The mean is commonly known as average, if there are n number of values in a dataset and the values are $x_1$, $x_2$, ..., $x_n$ , then the mean is calculated as:

$$\bar{\boldsymbol{x}} = \frac{\boldsymbol{x_1 + x_2 + x_3 + x_4 \ldots\ldots + x_n}}{\boldsymbol{n}}$$

**Example 1:**

Compute the mean for the following data:

**5, 6, 2, 4, 7, 8, 3, 5, 6, 6**

**Solution:**

**Mean**$= \frac{\mathbf{5+6+2+4+7+8+3+5+6+6}}{\mathbf{10}} = \frac{\mathbf{52}}{\mathbf{10}}$=5.2

**Mean drawback.**

Mean is susceptible to the influence of outliers. Also, mean is only meaningful if the data is normally distributed, or at least close to looking like a normal distribution.

# Measures of Central Tendency cont.

## Mean cont.

## Mean drawback.

Mean is susceptible to the influence of outliers, outliers have a significant impact on the mean, as they can skew it towards their extreme values.

**Example 2**

Here the data for salaries of **10** individuals in month, Compute the mean :

8.000, 8.700, 10.500, 8.000, 9.000, 10.000, 8.500, 10.200, 7.900, **150.000**

**Solution:**

**Mean**= $\frac{\mathbf{8000+8700+10500+8000+9000+10000+8500+10200+7900+150{,}000}}{\mathbf{10}} = \frac{\mathbf{230{,}800}}{\mathbf{10}}$=**23.800**

# Measures of Central Tendency cont.

## Median.

The median is the middle score for a dataset that has been sorted according to the values of the data, it is not affected by the outliers.

### Odd Number of Observations

If the total number of observations given is odd, then the formula to calculate the median is:

Median =$\left(\frac{n+1}{2}\right)^{th}$ $term$, n is , where n is the number of observations.

### Even Number of Observations

If the total number of observation is even, then the median formula is:

Median =$\frac{\left(\frac{n}{2}\right)^{th} term+\left(\frac{n}{2}+1\right)^{th} term}{2}$ , where n is the number of observations.

# Median cont.

**Example 1:**

Find the median of the following:

4, 17, 77, 25, 22, 23, 92, 82, 40, 24, 14, 12, 67, 23, 29

**Solution:**

- Order the numbers ascending, or descending, we have:

4, 12, 14, 17, 22, 23, 23, **24**, 25, 29, 40, 67, 77, 82, 92,

- There are 15 numbers. Our middle is the term no (15+1)/2 , i. e, the eighth term, the median value of this set of numbers is 24.

**Example 2:**

Find the median of the following:

3, 13, 7, 5, 21, 23, 23, 40, 23, 14, 12, 56, 23, 29

**Solution**

- Order the numbers ascending, or descending, we have:

3, 5, 7, 12, 13, 14, **21, 23**, 23, 23, 23, 29, 40, 56

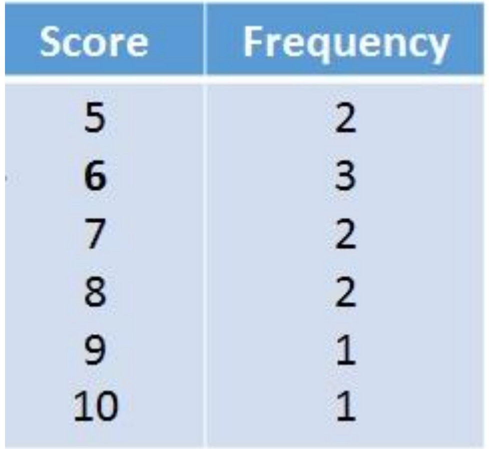
- There are 14 numbers, our middle is the average of the two terms no (14)/2=7, (14)/2+1=8 , i. e, the median value of this set of numbers is (21+23) /2=22.

# Measures of Central Tendency cont.

## Mode.

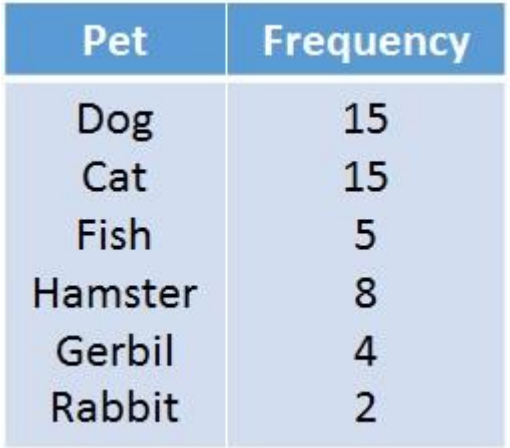
The mode is the most frequently occurring value in a dataset. On a histogram representation, the highest bar denotes the mode of the data. For example the mode for the data 3, 3, 6, 9, 16, 16, 16, 27, 27, 37, 48 is 16

.

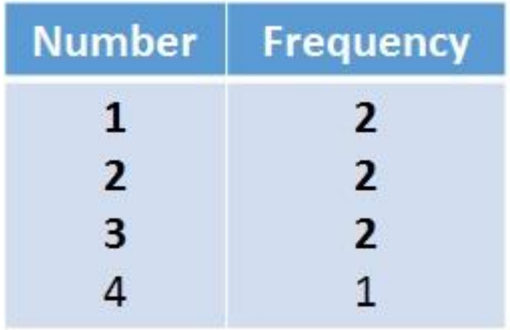


| Score | Frequency |
|---|---|
| 5 | 2 |
| **6** | 3 |
| 7 | 2 |
| 8 | 2 |
| 9 | 1 |
| 10 | 1 |

Uni mode

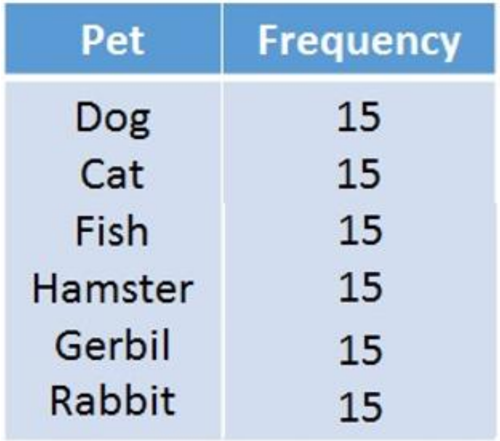


| Pet | Frequency |
|---|---|
| Dog | 15 |
| Cat | 15 |
| Fish | 5 |
| Hamster | 8 |
| Gerbil | 4 |
| Rabbit | 2 |



| Number | Frequency |
|---|---|
| **1** | **2** |
| **2** | **2** |
| **3** | **2** |
| 4 | 1 |

Multi mode



| Pet | Frequency |
|---|---|
| Dog | 15 |
| Cat | 15 |
| Fish | 15 |
| Hamster | 15 |
| Gerbil | 15 |
| Rabbit | 15 |

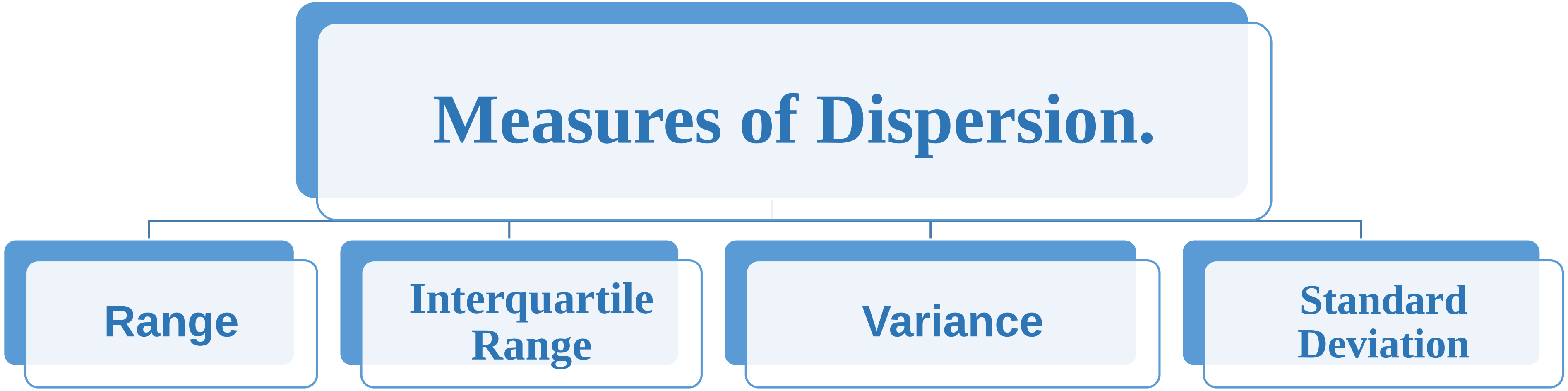
No mode

# Descriptive data analysis methods cont.

## Measures of Dispersion.

Looking at a central point (mean, median, or mode) may not help in understanding the actual shape of a distribution. Therefore, we often look at the spread, or the dispersion, of a distribution.

The most common measures of dispersion is:



Measures of Dispersion.

- Range
- Interquartile Range
- Variance
- Standard Deviation

# Measures of Dispersion cont.

## Range

The range is the difference between largest observation and  smallest observation

**Range = largest observation- smallest observation**

The main **disadvantage** of the range is that it is affected by **extremes or outliers**, resulting in an inaccurate image of the most likely range.

For example for the data  3, 6, 9,16, 27, 27, 37, 48

The range is 48-3=45

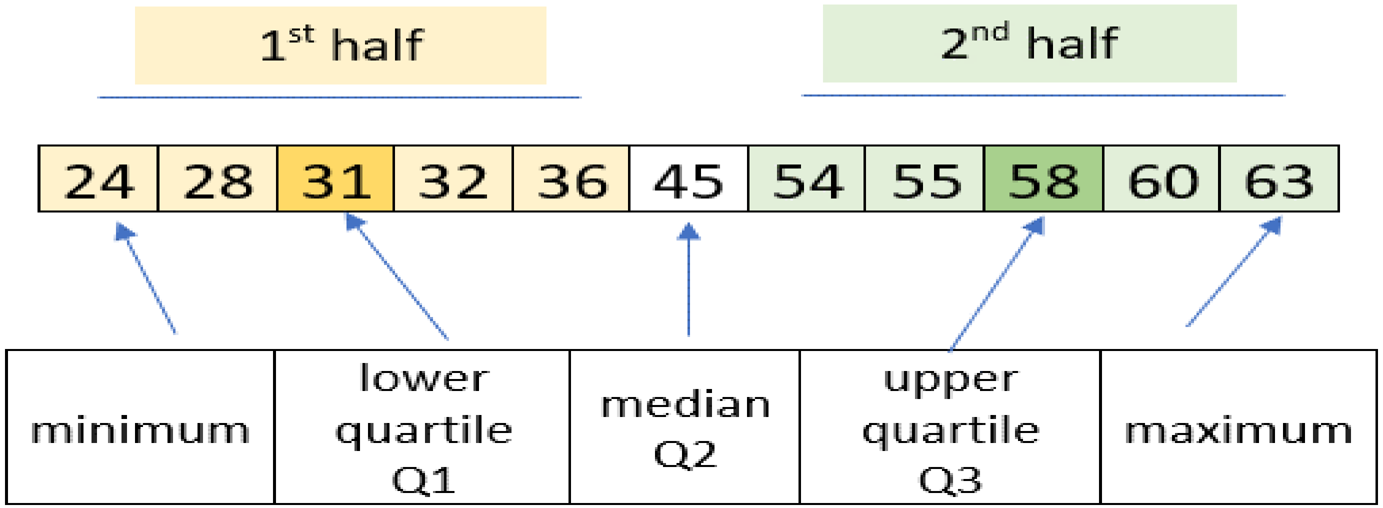
# Measures of Dispersion cont.

## Interquartile Range.

One way around the range’s disadvantage is to calculate it after removing extreme values. One convention is to cut off the top and bottom one-quarter of the data and calculate the range of the remaining middle 50% of the scores.

**Interquartile range = upper quartile − lower quartile,**

**Where** The lower quartile value is the median of the lower half of the data, The upper quartile value is the median of the upper half of the data, and the data should be arranged ascendingly.



1st half

2nd half

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 24 | 28 | 31 | 32 | 36 | 45 | 54 | 55 | 58 | 60 | 63 |

| | | | | |
|---|---|---|---|---|
| minimum | lower quartile Q1 | median Q2 | upper quartile Q3 | maximum |

# Interquartile Range cont.

**Example.1**

Find the interquartile range for the following set of data.

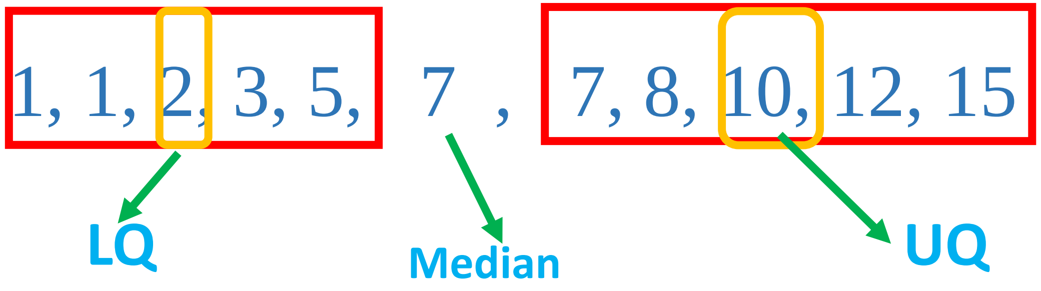
1,10,2,3,15,7,7,8,1,12,5

**Solution**

Arrange the data ascendingly

1, 1, 2, 3, 5, 7, 7, 8, 10, 12, 15

Divide the data to 2 halves



1, 1, 2, 3, 5, 7 , 7, 8, 10, 12, 15

LQ

Median

UQ

**LQ=2**
**UQ=10**
**Interquartile Range=10-2=8**

# Interquartile Range cont.

**Example.2**

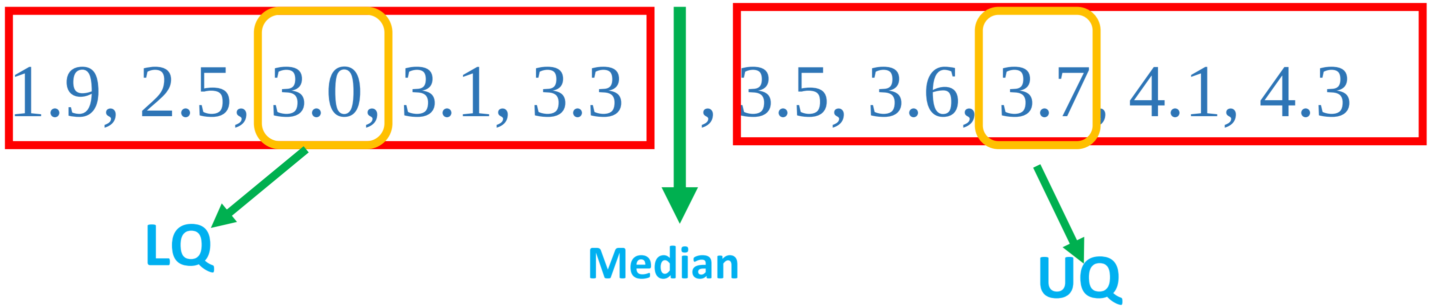
Find the interquartile range for the following set of data.
3.3, 3.7, 2.5, 3.5, 3.0, 4.3, 3.1, 4.1, 1.9, 3.6

**Solution**

Arrange the data ascendingly

1.9, 2.5, 3.0, 3.1, 3.3, 3.5, 3.6, 3.7, 4.1, 4.3

Divide the data to 2 halves



1.9, 2.5, 3.0, 3.1, 3.3 , 3.5, 3.6, 3.7, 4.1, 4.3

LQ=3.0
UQ=3.7
**Interquartile Range= 3.7-3.0= 0.7**

# Measures of Dispersion cont.

## Variance.

The variance is a measure used to indicate how spread out the data points are. To measure the variance, the common method is to pick a center of the distribution, typically the mean, then measure how far each data point is from the center.

The **variance of the population** is defined by the following formula:

$$\sigma^2 = \frac{\sum(x_i - \boldsymbol{\bar{x}})^2}{n}$$

Where $\boldsymbol{\bar{x}}$is the population mean, $x_i$ is the ith element from the population, and n is the number of elements in the population.

The **variance of a sample** is defined by a slightly different formula:

$$S^2 = \frac{\sum(x_i - \boldsymbol{\bar{x}})^2}{n - 1}$$

Where $\boldsymbol{\bar{x}}$ is the sample mean, $x_i$ is the ith element from the sample, and n is the number of elements in the sample.

# Measures of Dispersion cont.

## Standard Deviation.

There is one issue with the variance as a measure. It gives us the measure of spread in units squared. So, for example, if we measure the variance of age (measured in years) of all the students in a class, the measure we will get will be in years squared. However, practically, it would make more sense if we got the measure in years (not years squared). For this reason, we often take the square root of the variance, This measure is known as the standard deviation.

The formula to compute the standard deviation of the population is:

$$\sigma = \sqrt{\frac{\sum(x_i - \boldsymbol{\overline{x}})^2}{n}}$$

The formula to compute the standard deviation of the sample is:

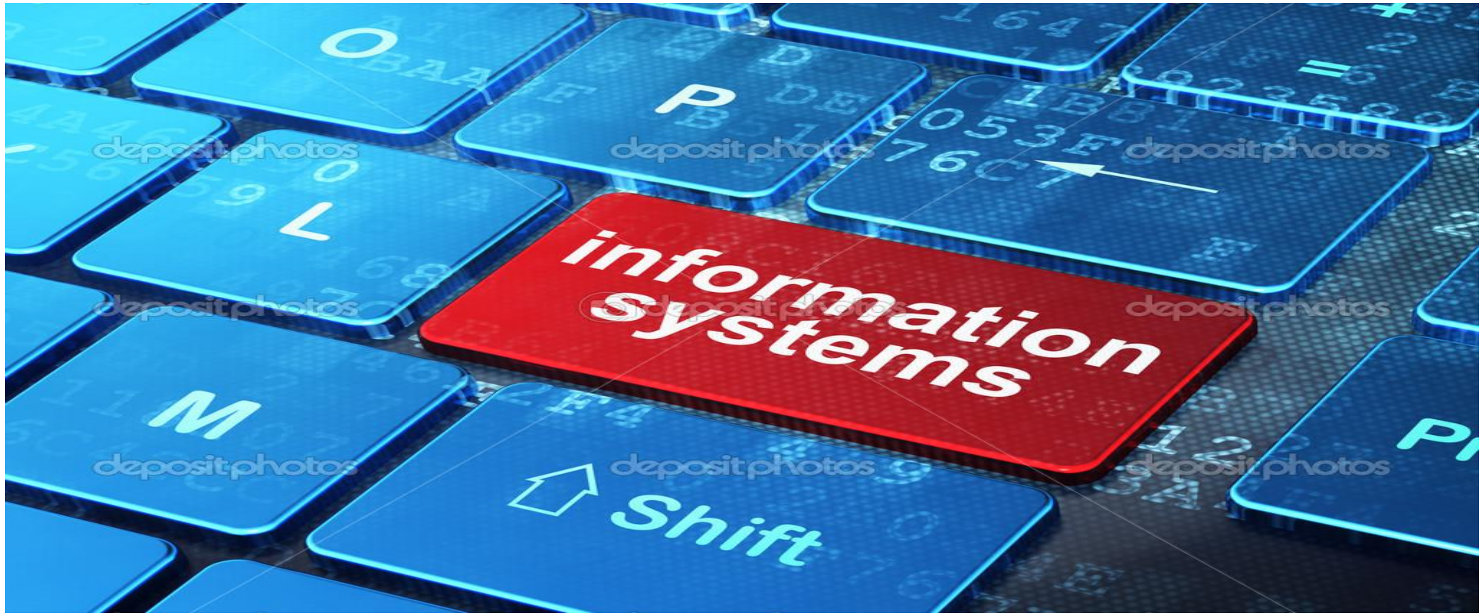
$$S = \sqrt{\frac{\sum(x_i - \boldsymbol{\overline{x}})^2}{n-1}}$$

# Assignments # 1

1- Give an example of data, information and knowledge?

2- Mention the difference between a data model and a schema.

# Advanced Topics in Information systems



## SE204 lecture_2

Dr. Nelly Amer

# Measures of Dispersion cont.

## **Variance**.

The variance is a measure used to indicate how spread out the data points are. To measure the variance, the common method is to pick a center of the distribution, typically the mean, then measure how far each data point is from the center.

The **variance of the population** is defined by the following formula:

$$\sigma^2 = \frac{\sum(x_i - \boldsymbol{\bar{x}})^2}{n}$$

Where $\boldsymbol{\bar{x}}$is the population mean, $x_i$ is the ith element from the population, and n is the number of elements in the population.

The **variance of a sample** is defined by a slightly different formula:

$$S^2 = \frac{\sum(x_i - \boldsymbol{\bar{x}})^2}{n-1}$$

Where $\boldsymbol{\bar{x}}$ is the sample mean, $x_i$ is the ith element from the sample, and n is the number of elements in the sample.

# Measures of Dispersion cont.

## Standard Deviation.

There is one issue with the variance as a measure. It gives us the measure of spread in units squared. So, for example, if we measure the variance of age (measured in years) of all the students in a class, the measure we will get will be in years squared. However, practically, it would make more sense if we got the measure in years (not years squared). For this reason, we often take the square root of the variance, This measure is known as the standard deviation.

The formula to compute the standard deviation of the population is:

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{\boldsymbol{x}})^2}{n}}$$

The formula to compute the standard deviation of the sample is:

$$S = \sqrt{\frac{\sum(x_i - \bar{\boldsymbol{x}})^2}{n-1}}$$

# Measures of Dispersion cont.

## Example 1

**Calculate the variance and standard deviation of the sample data:**

**10, 11, 15, 20, 24.**

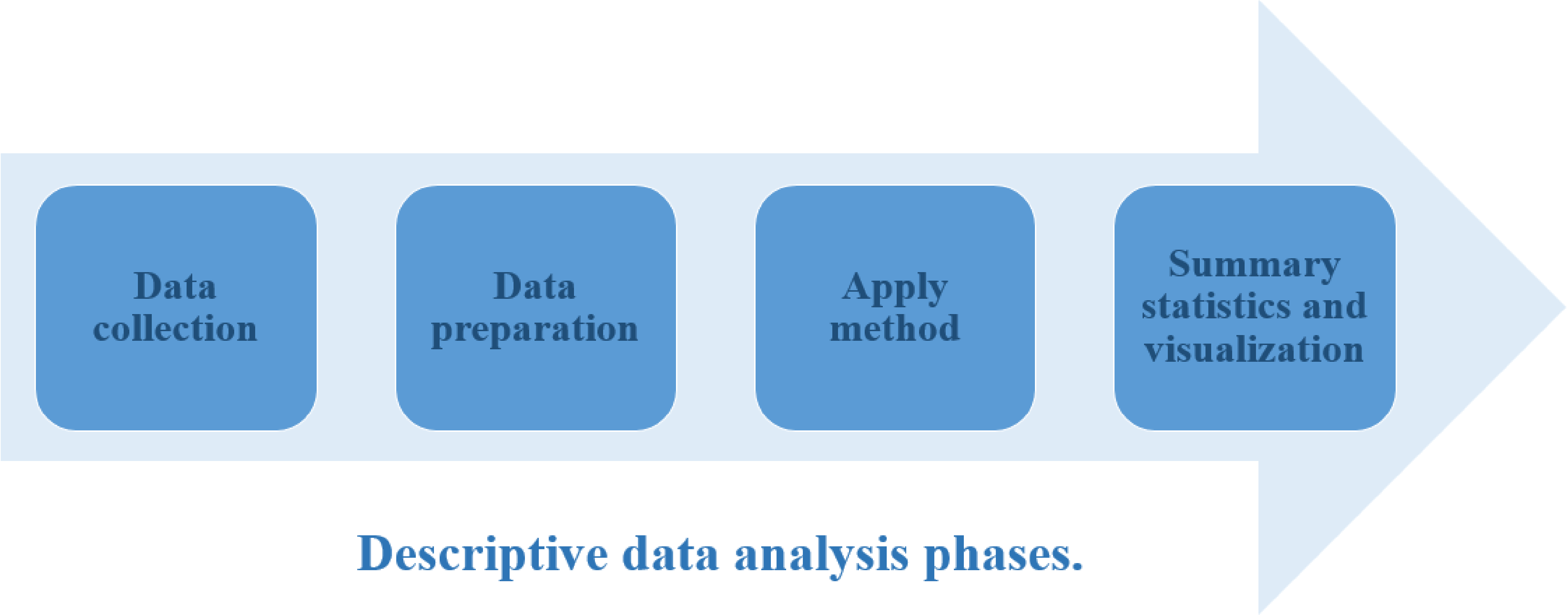
**Solution**
**Mean = (10+11+15+20+24)/5= 16**

| x | $(x\text{-}\bar{x})$ | $(x-\bar{x})^2$ |
|---|---|---|
| 10 | -6 | 36 |
| 11 | -5 | 25 |
| 15 | -1 | 1 |
| 20 | 4 | 16 |
| 24 | 8 | 64 |
| total | 0 | 142 |

$S^2$= 142/ 4 =35.5
$S$ = 5.958

# Descriptive data analysis phases.

Conducting a descriptive analysis entails several critical phases [7].



Descriptive data analysis phases.

# Descriptive data analysis phases cont.

## 1- Data Collection

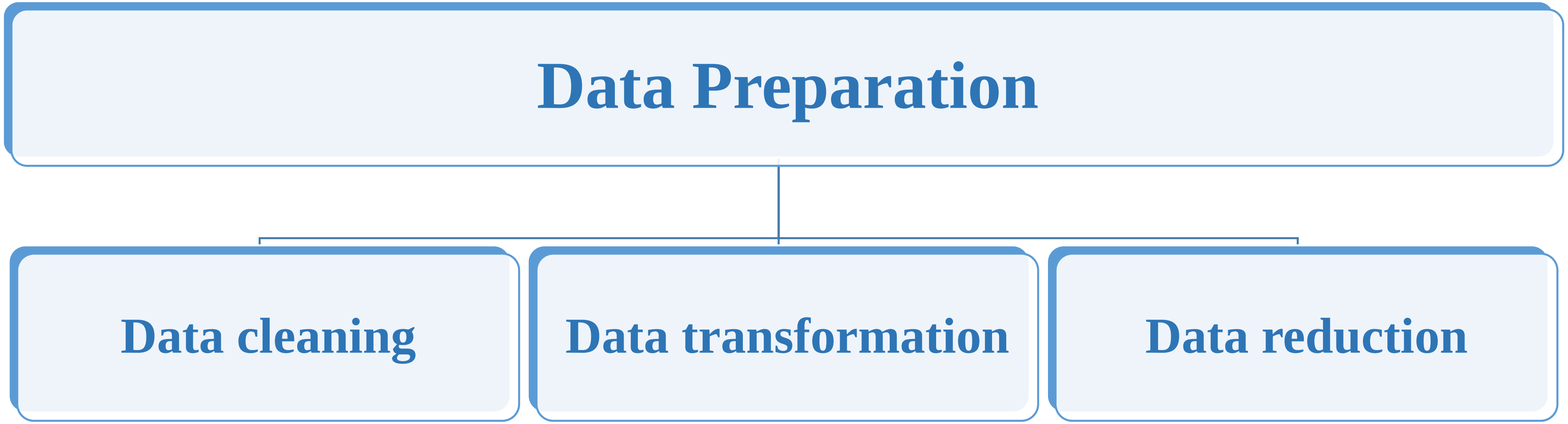
The first phase in descriptive data analysis is collecting relevant data. This process involves identifying data sources, selecting appropriate data-collecting methods, and verifying that the data acquired accurately represents the population or topic of interest.

You can collect data through surveys, experiments, observations, existing databases, or other data collection methods.

# Descriptive data analysis phases cont.

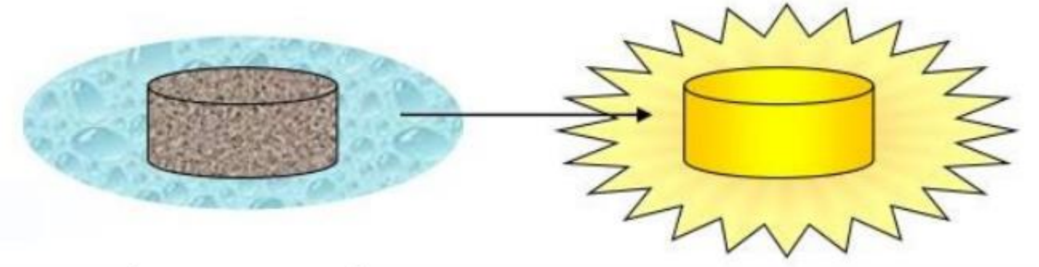
## 2-Data Preparation

Data preparation is crucial for ensuring the dataset is clean, consistent, and ready for analysis. This step covers the following tasks:



# Data Preparation cont.

Data preparation is crucial for ensuring the dataset is clean, consistent, and ready for analysis. This step covers the following tasks:



- **Data Cleaning:**

Data in the real world is often uncleaned ; that is, it is in need of being cleaned up before it can be used for a desired purpose. This is often called data pre-processing.

## Factors that indicate that data is not clean:

- **Incomplete**. When some of the attribute values are missed.

- **Noisy**. When data contains errors or outliers.

For example, some of the data points in a dataset may contain extreme values that can severely affect the dataset’s range.

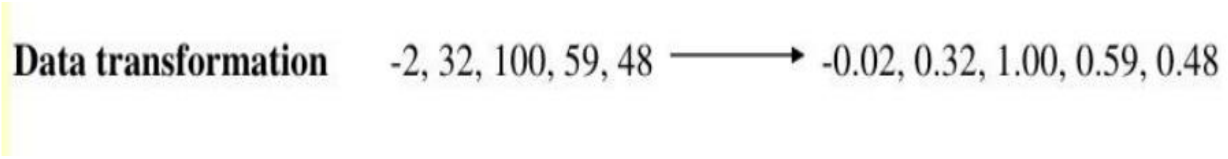
- **Inconsistent**. refers to the lack of uniformity in data format or content, data contains discrepancies.

For example if records do not start with a capital letter, discrepancies are present.

# Data Preparation cont.

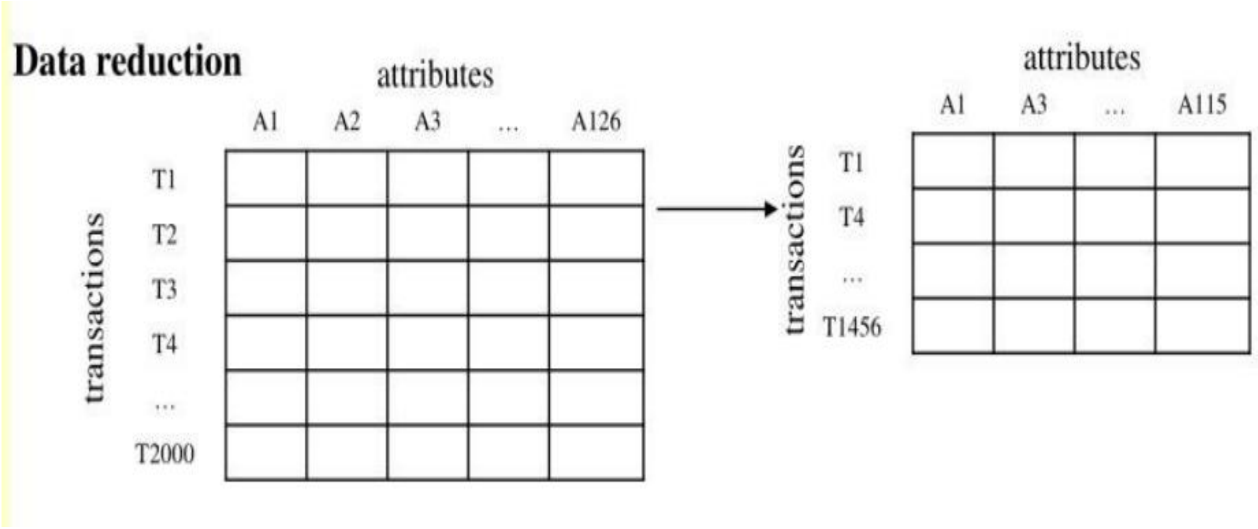
- **Data Transformation:**

Convert data into an appropriate format. Examples of this are changing data types, encoding categorical variables.



- **Data Reduction:**

For large datasets, try reducing their size by sampling or aggregation to make the analysis more manageable.



# Descriptive data analysis phases cont.

## 3- Apply Methods

In this step, Identify which variables are important to your descriptive analysis and research questions, then  analyze and describe the data using descriptive data analysis, which are frequency measures, central tendency measures, and Dispersion measures

After the data set has been analyzed, researchers may interpret the findings in light of the goals. The analysis was successful if the conclusions were what was anticipated. Otherwise, they must search for weaknesses in their strategy and repeat these processes to get better outcomes.

## 4- Summary Statistics and Visualization

- **Summary Statistics:** Summarize your findings clearly and concisely.
- **Data Visualization:** Use various charts and plots to visualize the data. Create histograms,  scatter plots, or line charts for numerical data. Use bar charts, pie charts, or stacked bar charts for categorical data.

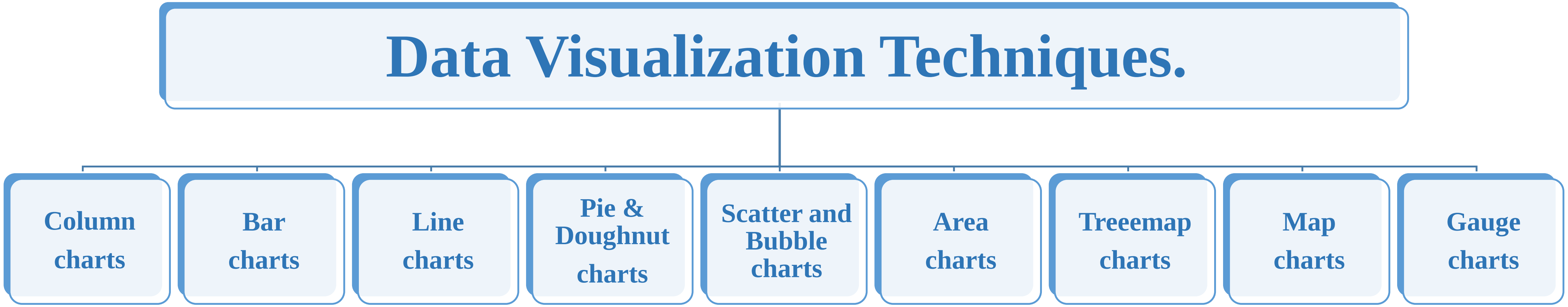
# Data visualization.

Data visualization is a powerful tool for enhancing understanding and communication of complex data. It involves representing data in a graphical or pictorial form, making it easier to gain insight into their structure and patterns , understand and interpret [8].

There are several types of data visualization techniques, The choice of data visualization technique will depend on the data type being analyzed, the insights being sought, and the target audience. Effective data visualization involves choosing the right technique for the data and the message being conveyed and presenting the data clearly and visually appealingly.

# Data Visualization Techniques

There most common data visualization techniques:



Data Visualization Techniques.

Column charts

Bar charts

Line charts

Pie & Doughnut charts

Scatter and Bubble charts

Area charts

Treeemap charts

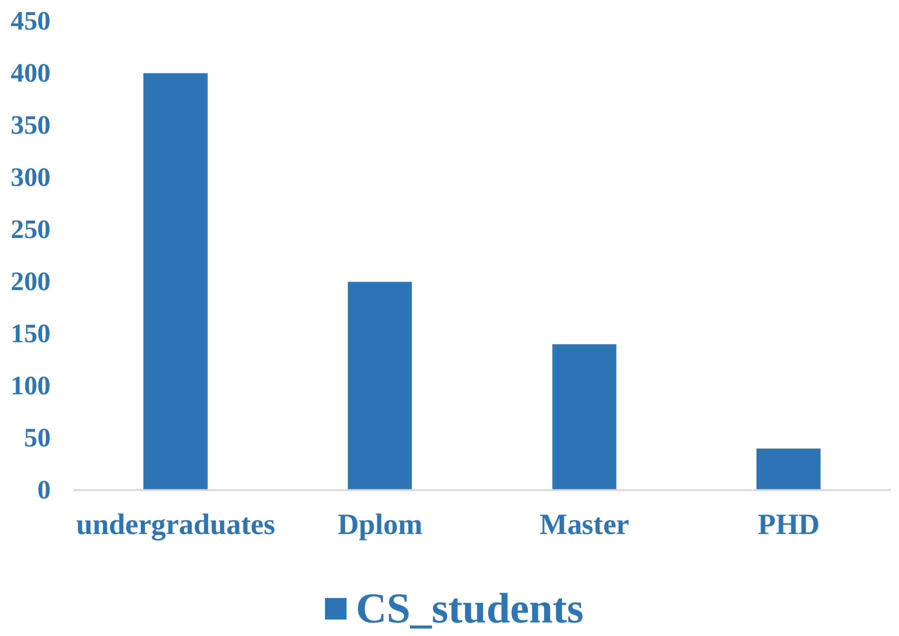
Map charts

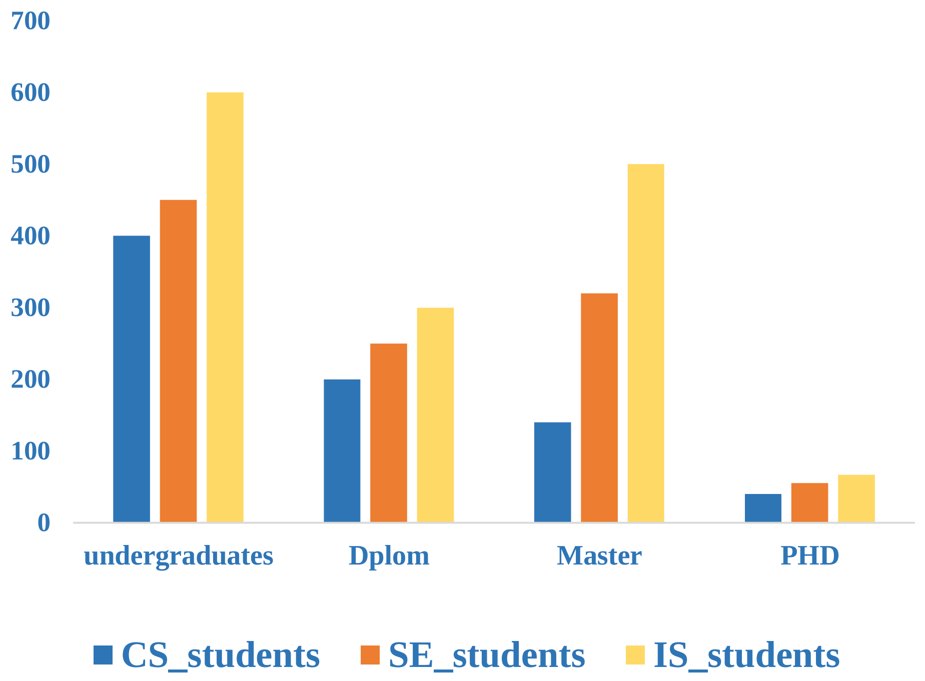
Gauge charts

# Data Visualization Techniques cont.

## Column charts.

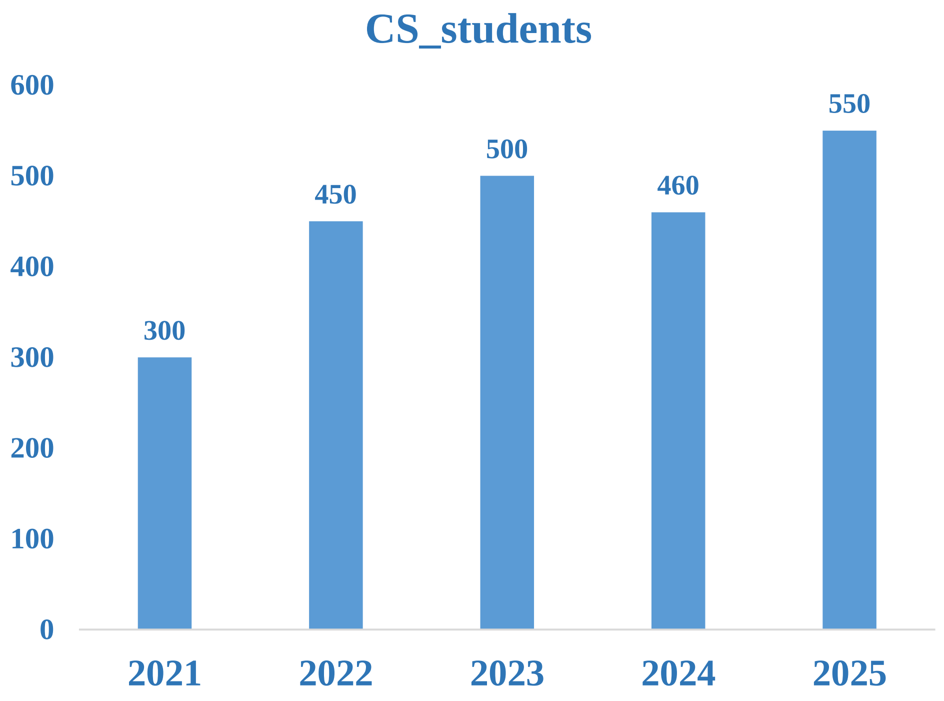
The column chart is the most used chart type, With column charts you could compare values for different categories or compare value changes over a period of time for a single category.



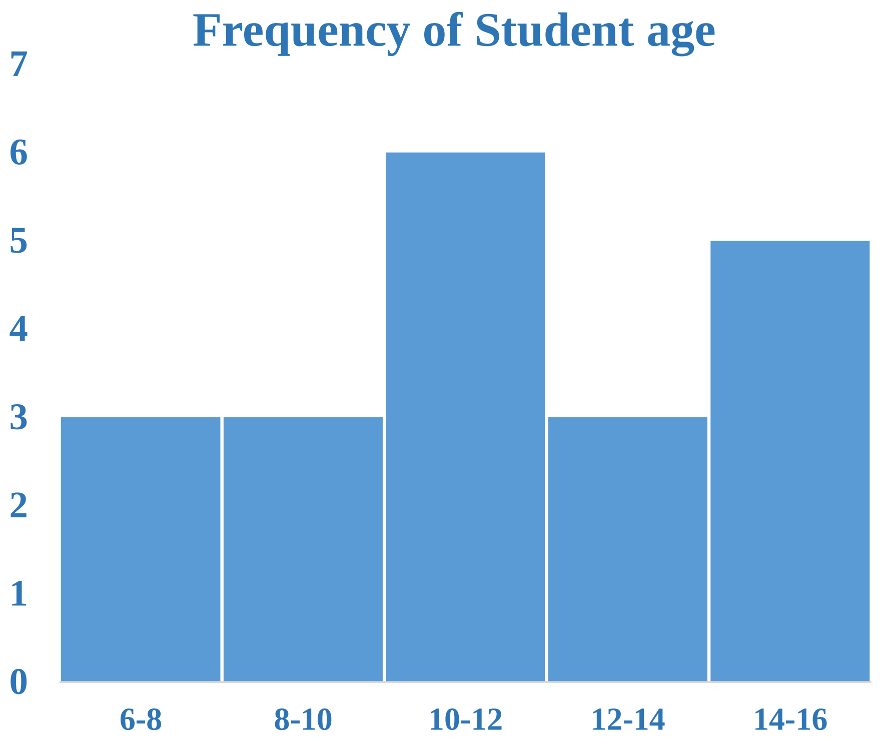


Column chart compare values for different categories.

# Column charts cont.

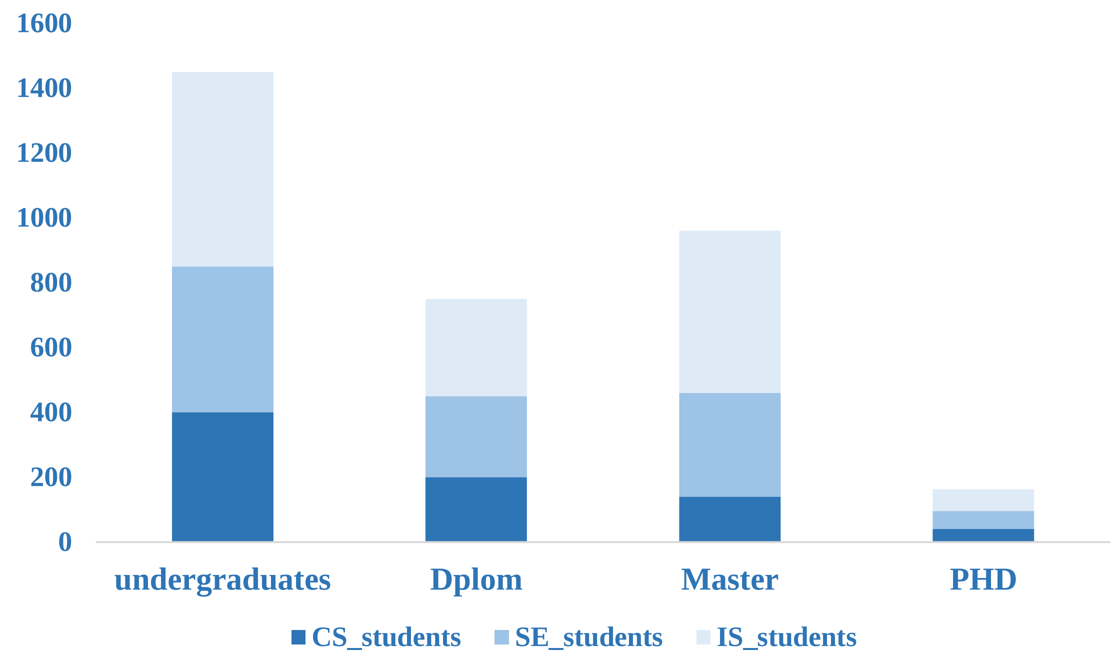


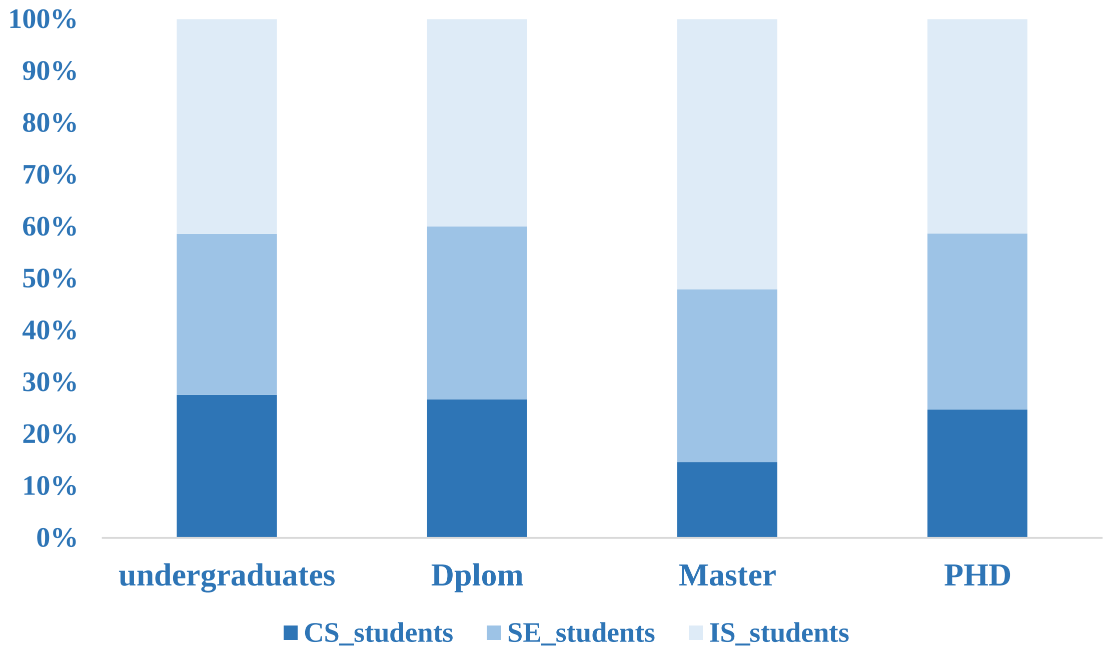
Column chart compares value changes over a period of time .



Column chart is used to represent Histogram.

# Column charts cont.





## Stacked column chart.

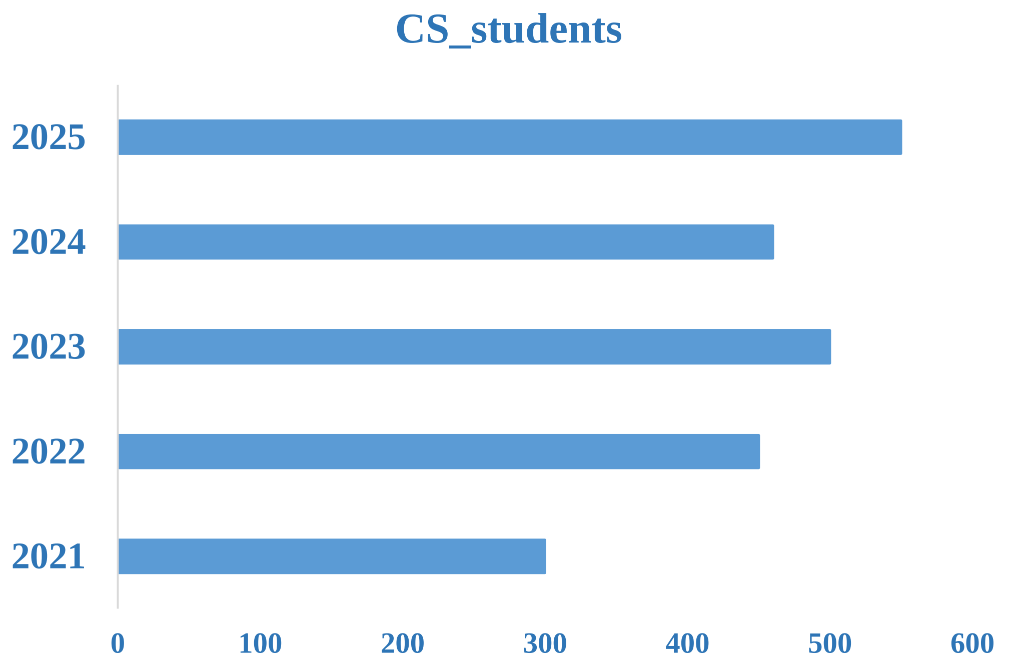
It compares parts of whole

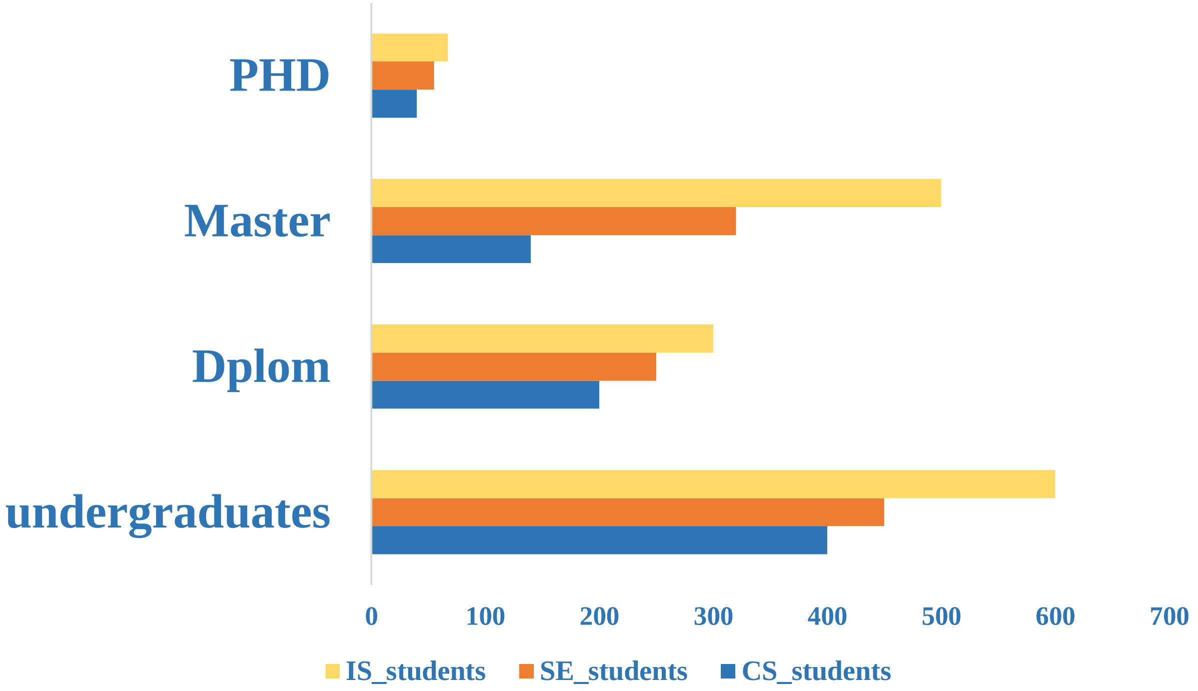
## Fully stacked column chart.

It compares the percentage that each value contributes to a total.

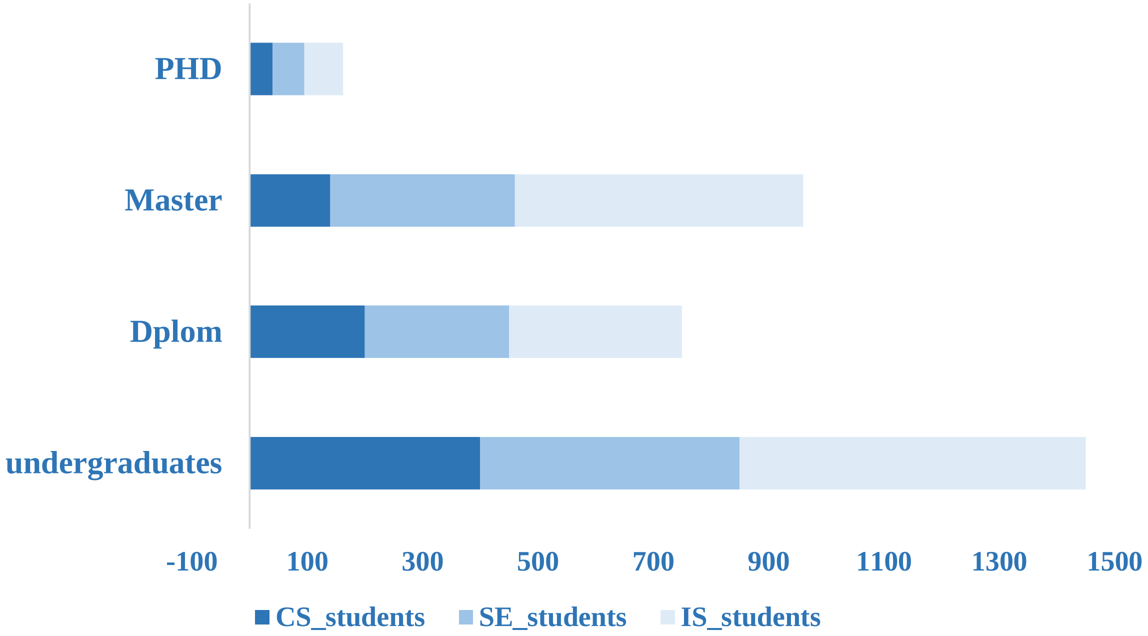
# Bar charts.

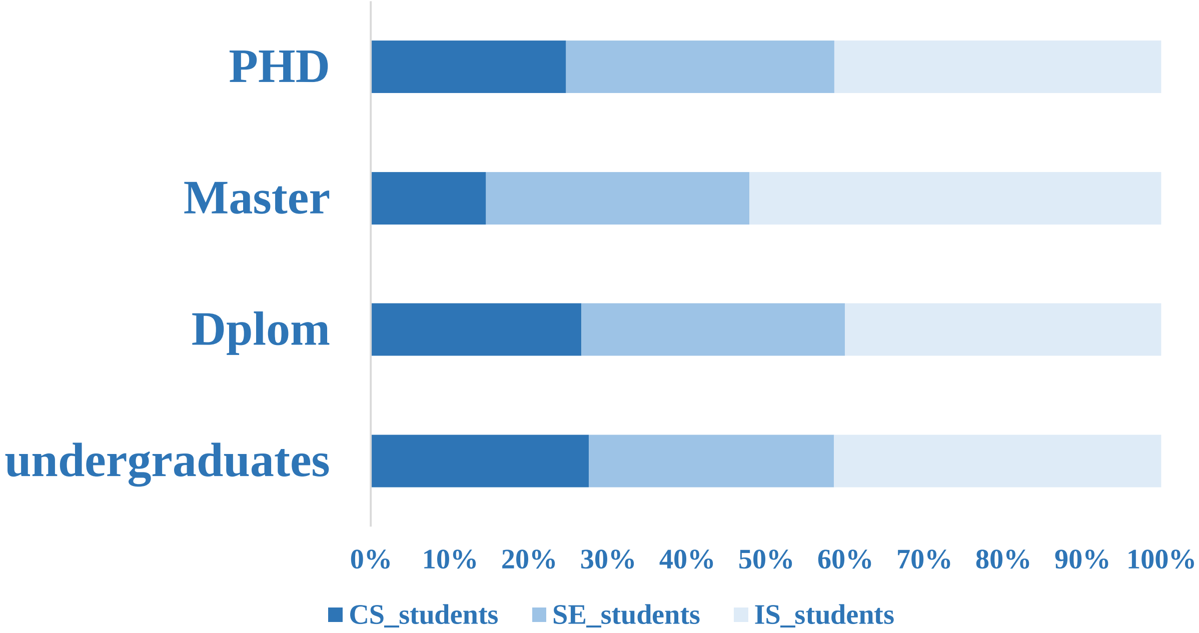
The bar chart is used to visually compare values across a few categories when the category text is long





# Bar charts cont.





## Stacked bar chart.

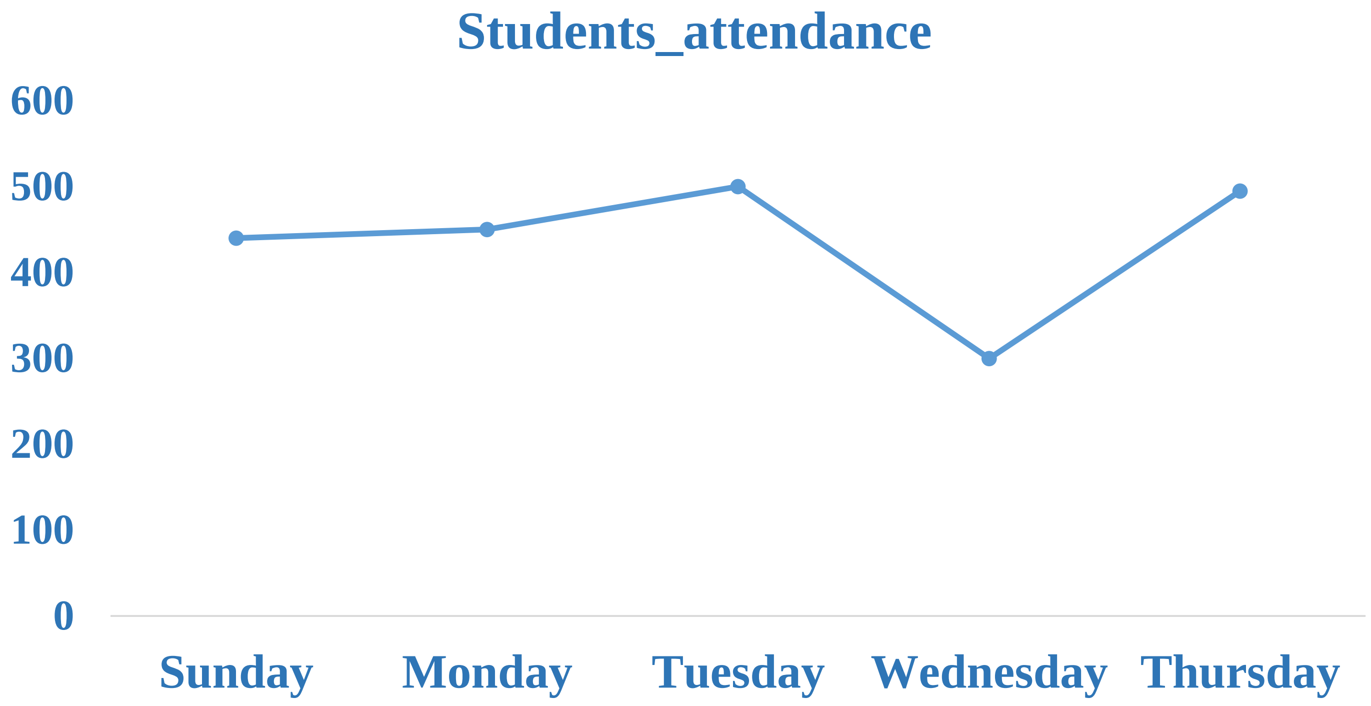
It compares parts of whole

## Fully stacked bar chart.

It compares the percentage that each value contributes to a total.

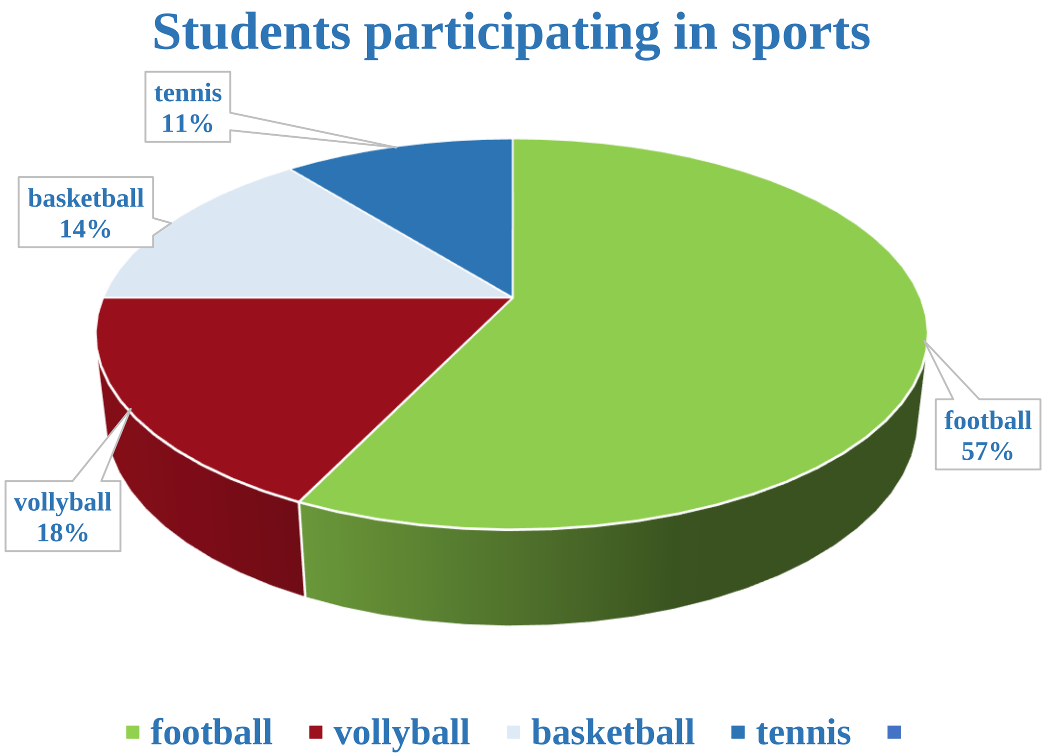
# Line charts.

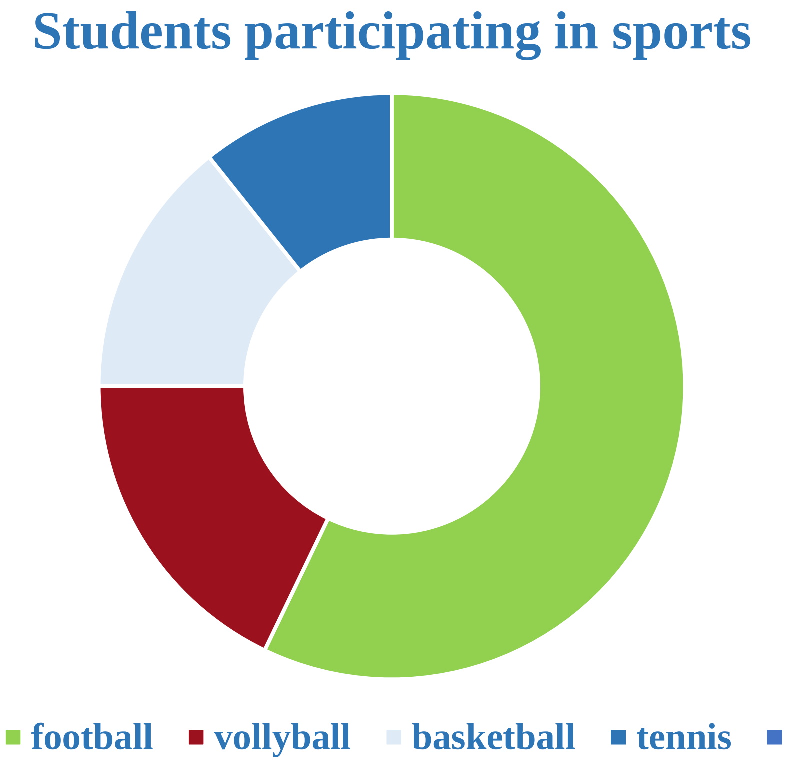
The line chart is used to show trends over time (years, months, and days ) or categories



# Pie & Doughnut charts.

The pie or doughnut charts are used to show proportions of a whole, use it when the totals of your numbers is 100%, and the categories is few, where many categories make the angles hard to estimate

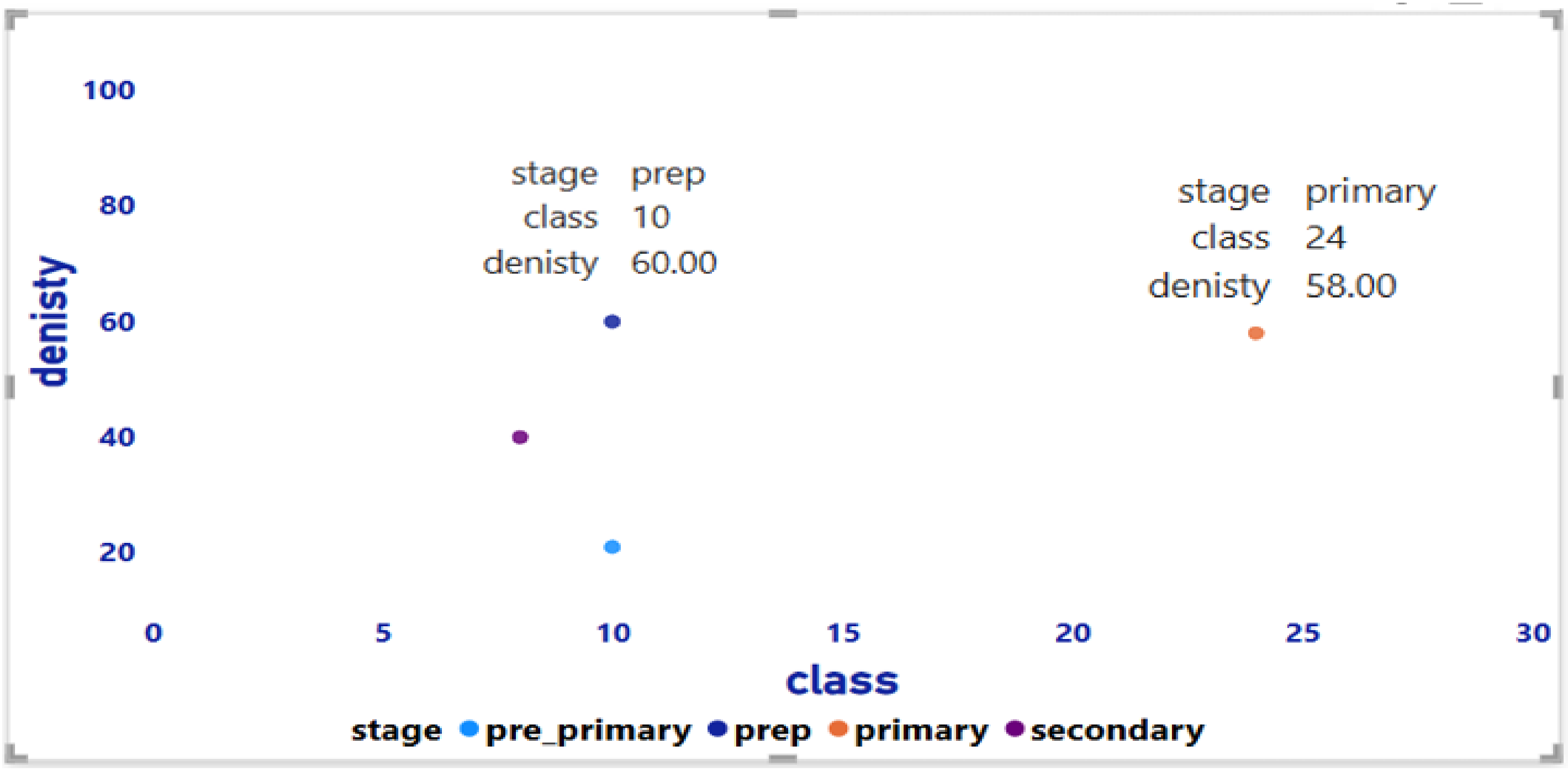




# Scatter and Bubble charts

The scatter chart is used to  compare at least two sets of values, and show relationship between them,  **Bubble** charts act as scatter charts with adding  a bubble size as a third dimension.

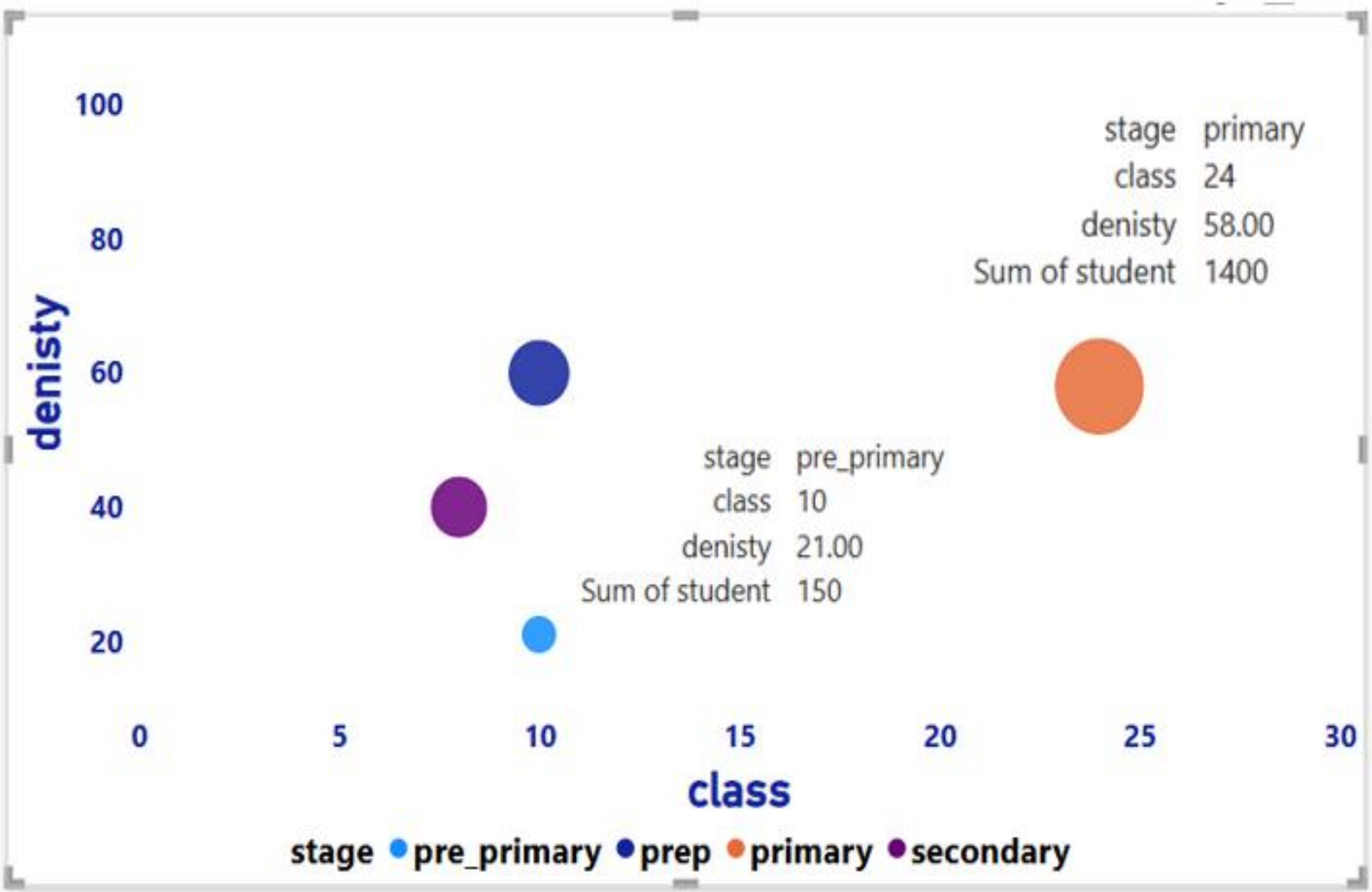
## Scatter chart example



| Stage | student | class | density |
|---|---|---|---|
| Pre primary | 150 | 10 | 21 |
| primary | 1400 | 24 | 58 |
| Prep | 600 | 10 | 60 |
| secondary | 500 | 8 | 43 |

# Scatter and Bubble charts cont.

## Bubble charts example



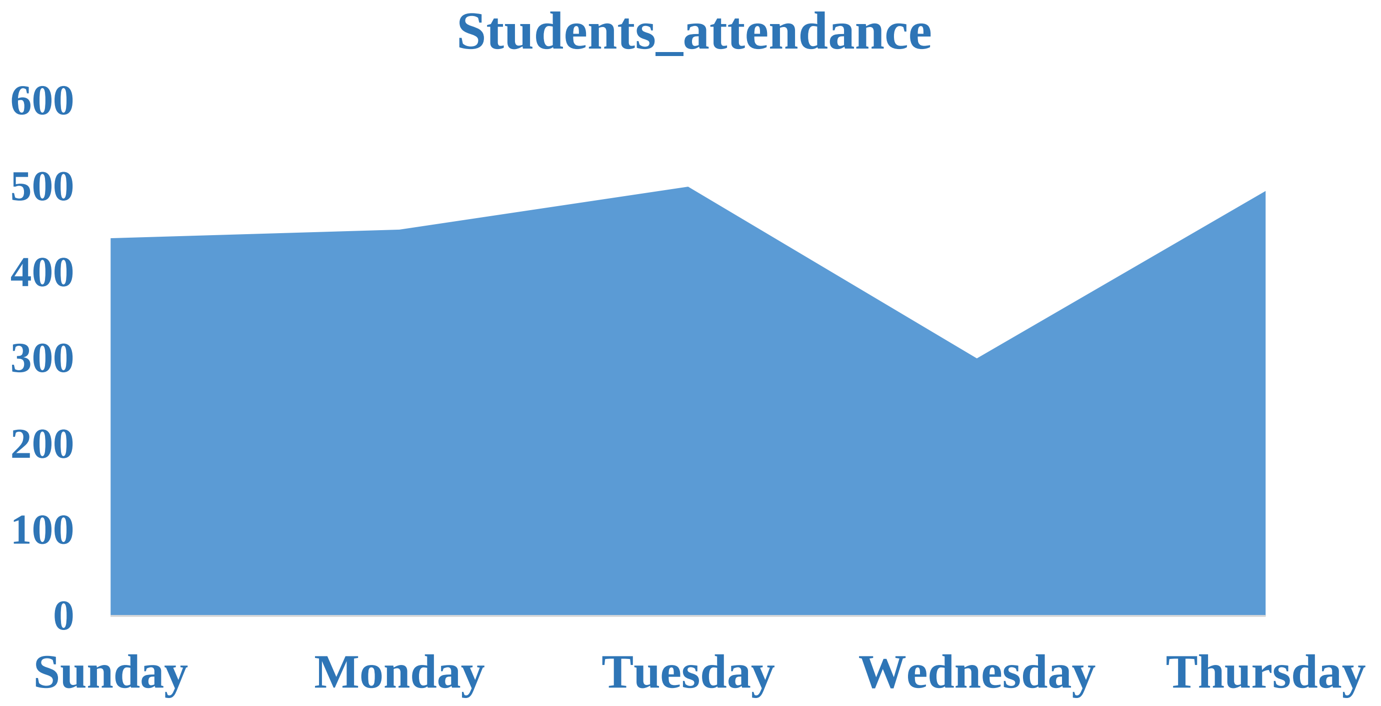
stage primary
class 24
denisty 58.00
Sum of student 1400

stage pre_primary
class 10
denisty 21.00
Sum of student 150

| Stage | student | class | density |
|---|---|---|---|
| Pre primary | 150 | 10 | 21 |
| primary | 1400 | 24 | 58 |
| Prep | 600 | 10 | 60 |
| secondary | 500 | 8 | 43 |

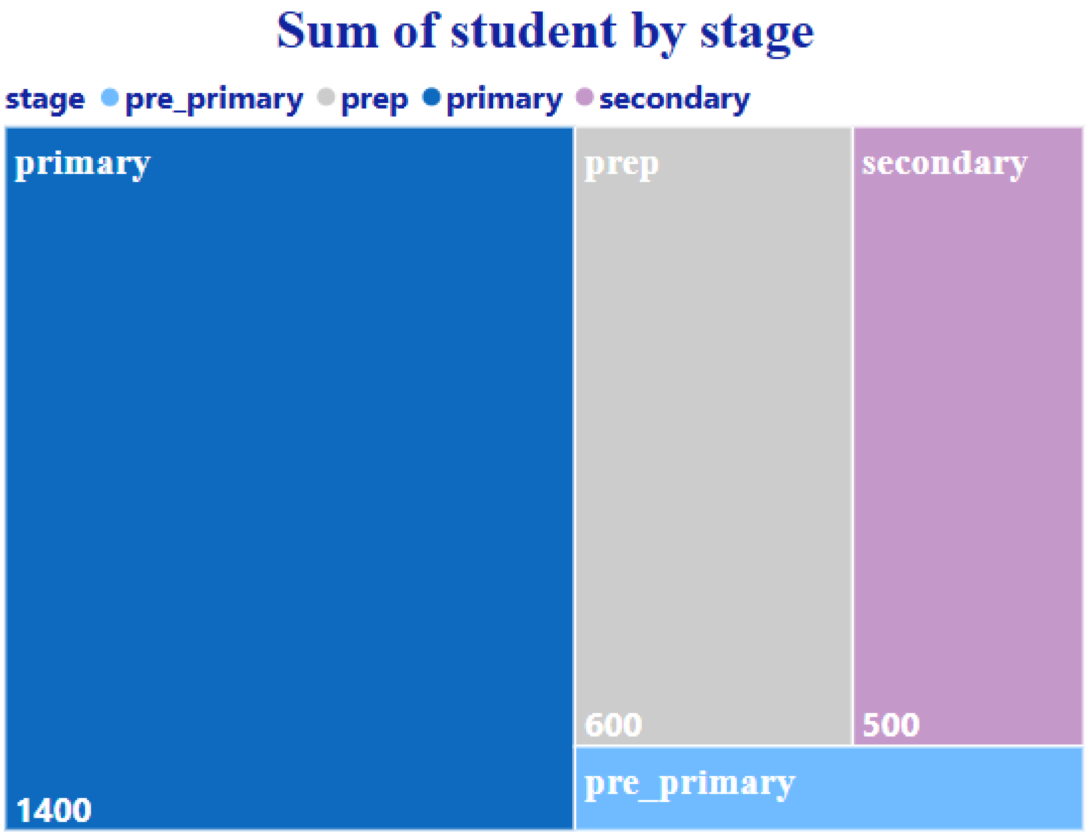
# Area charts.

An area chart is essentially a line chart, it is used to show trends over time (years, months, and days ) or categories, it highlights the magnitude of change over time.



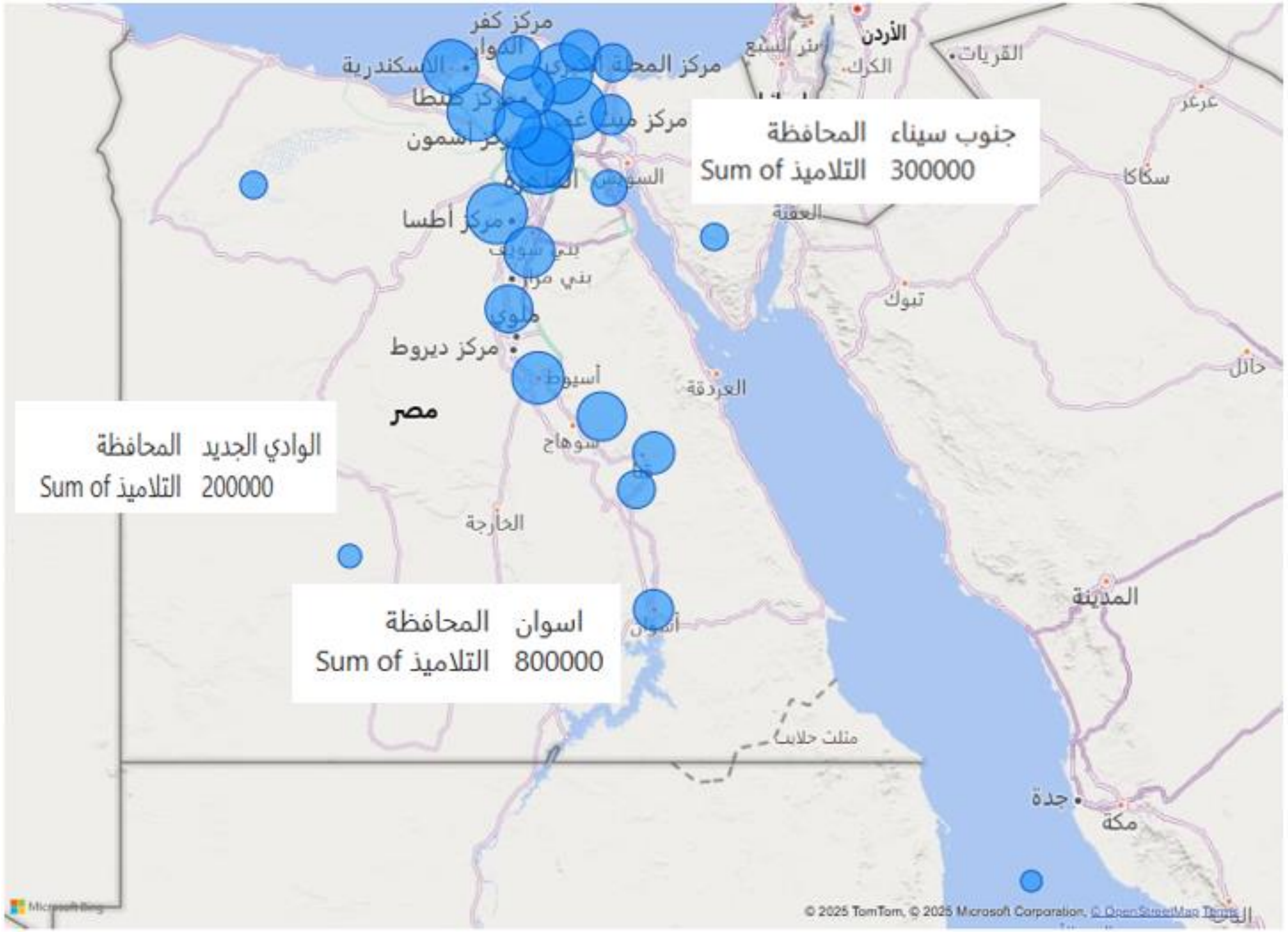
# Treemap charts.

Treemap charts show the relationship of parts to the whole by dividing the data into segments, These charts are best suited for illustrating percentages, such as the top five sales by product or country, or any other available categories.



# Map charts.

The map chart is used to  visualize the geographic distribution of your data.



المحافظة جنوب سيناء
Sum of التلاميذ 300000

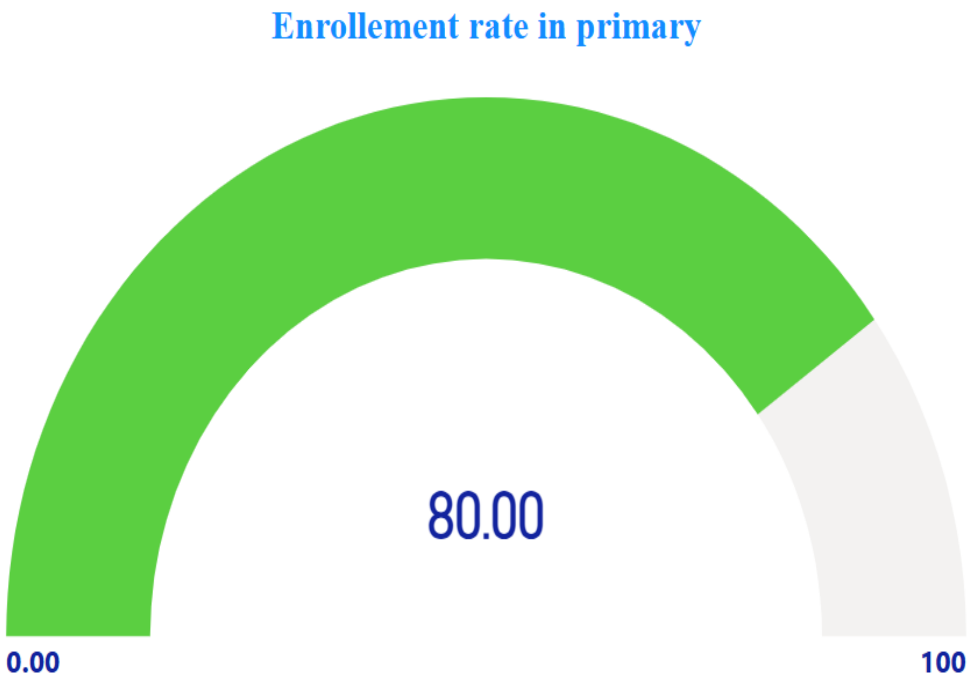
المحافظة الوادي الجديد
Sum of التلاميذ 200000

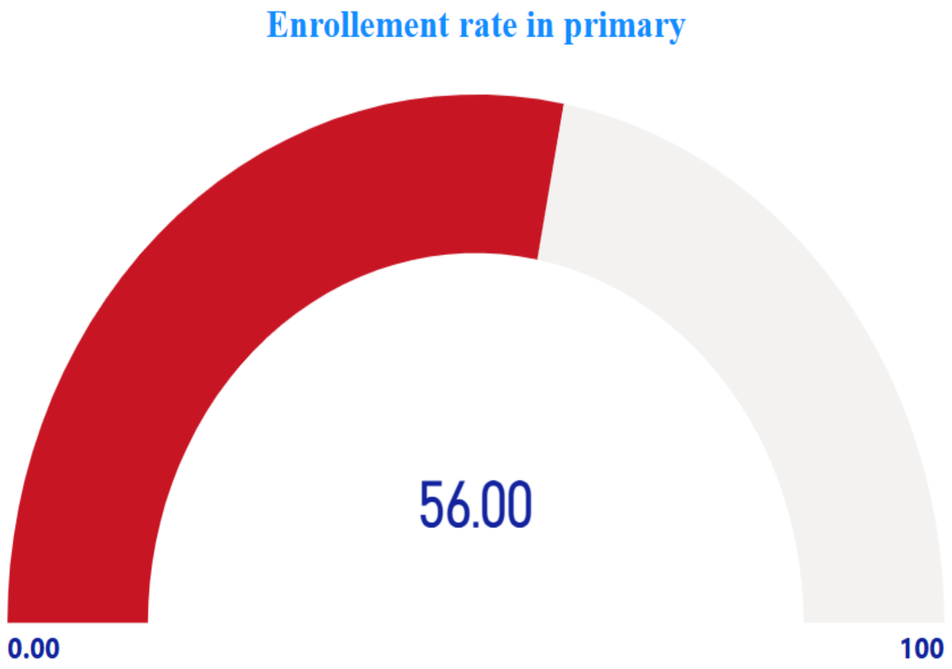
المحافظة اسوان
Sum of التلاميذ 800000

© 2025 TomTom, © 2025 Microsoft Corporation, © OpenStreetMap Terms

# Gauge charts.

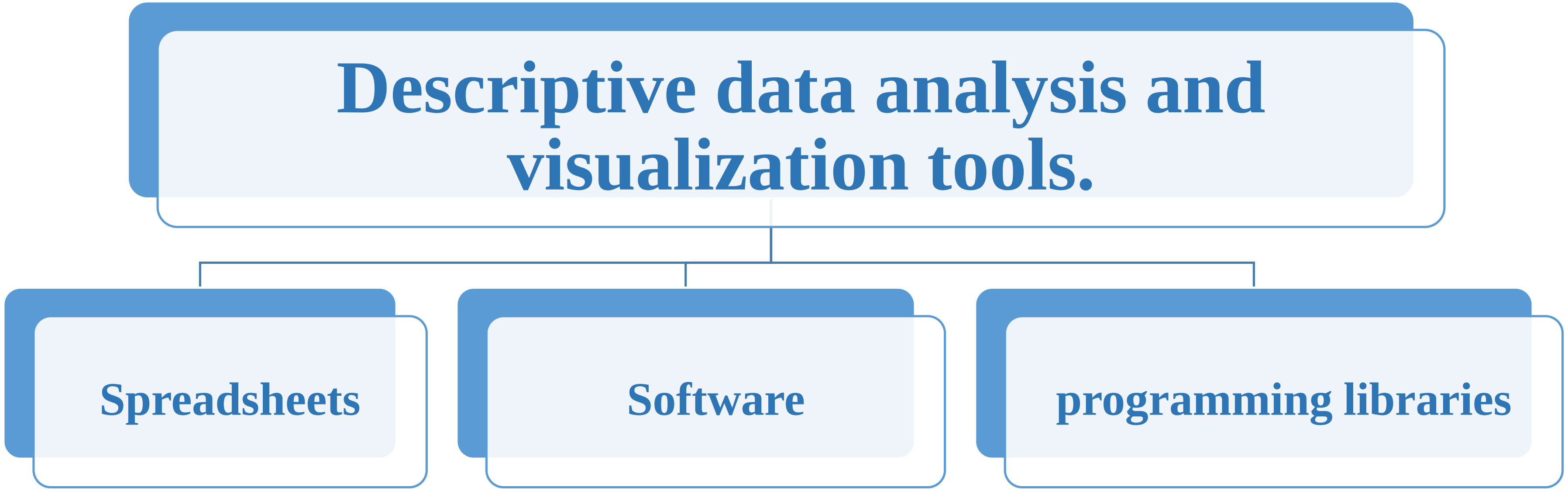
Gauge charts are good for displaying KPIs  (Key Performance Indicators). They typically display a single key value, comparing it to a color-coded performance level indicator, typically showing green for  “good” and red for “trouble.” Gauges are a great choice to show progress toward a goal, and Represent a percentile measure, like a KPI.





# Descriptive data analysis and visualization tools.

Data visualization tools can be classified into three categories [8]:



# Data analysis and visualization tools cont.

## Spreadsheets

Spreadsheets, such as Microsoft Excel and Google Sheets, are one of the most common data visualization tools used in various domains. They provide basic data visualization capabilities, such as bar charts, line graphs, and scatter plots.

## Data Visualization Software

Data visualization software is a specialized tool designed for data visualization and analysis. Examples of data visualization software include Tableau, QlikView, and Power BI. These tools provide advanced data visualization capabilities, including interactive dashboards, heat maps, and network diagrams.

## Programming libraries

Programming libraries, such as Matplotlib  (in python), ggplot2 ( in R), and D3.js in (Java), are a type of data visualization tool that can be used to create custom data visualizations. They provide a more flexible and customizable approach to data visualization but require a higher level of technical expertise

# Applications

## Pivot tables

Pivot tables are one of Excel's most powerful features. A pivot table allows you to extract the significance from a large, detailed data set.

Our data set consists of 213 records and 6 fields. Order ID, Product, Category, Amount, Date and Country.

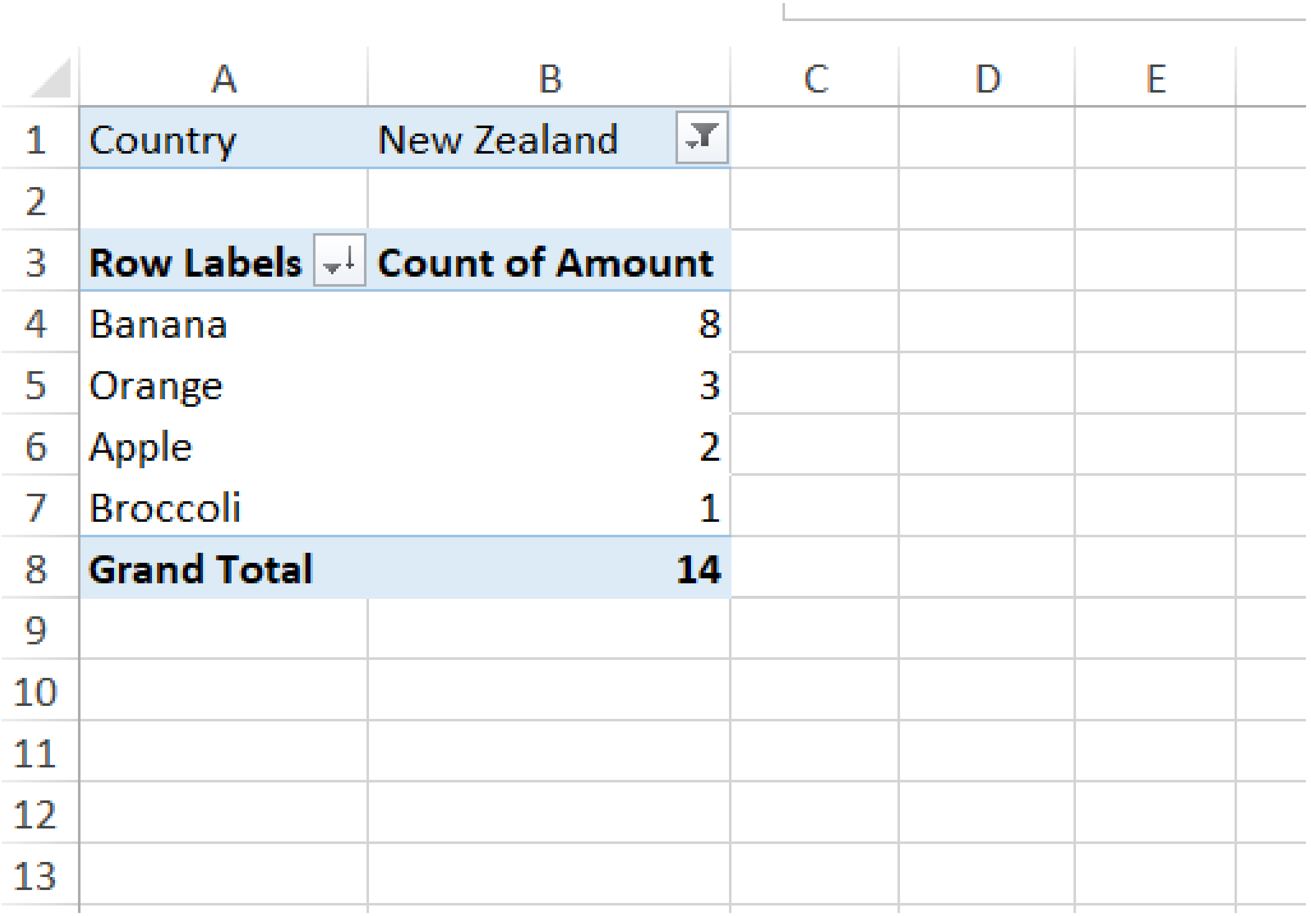
# Applications

## Pivot tables

- Add a PivotTable icon in Excel.
- Insert a PivotTable in Excel.
- Create a crosstab table using Excel PivotTable.
- Change null value to specific value.
- Make a histogram using Excel PivotTable.

# Applications

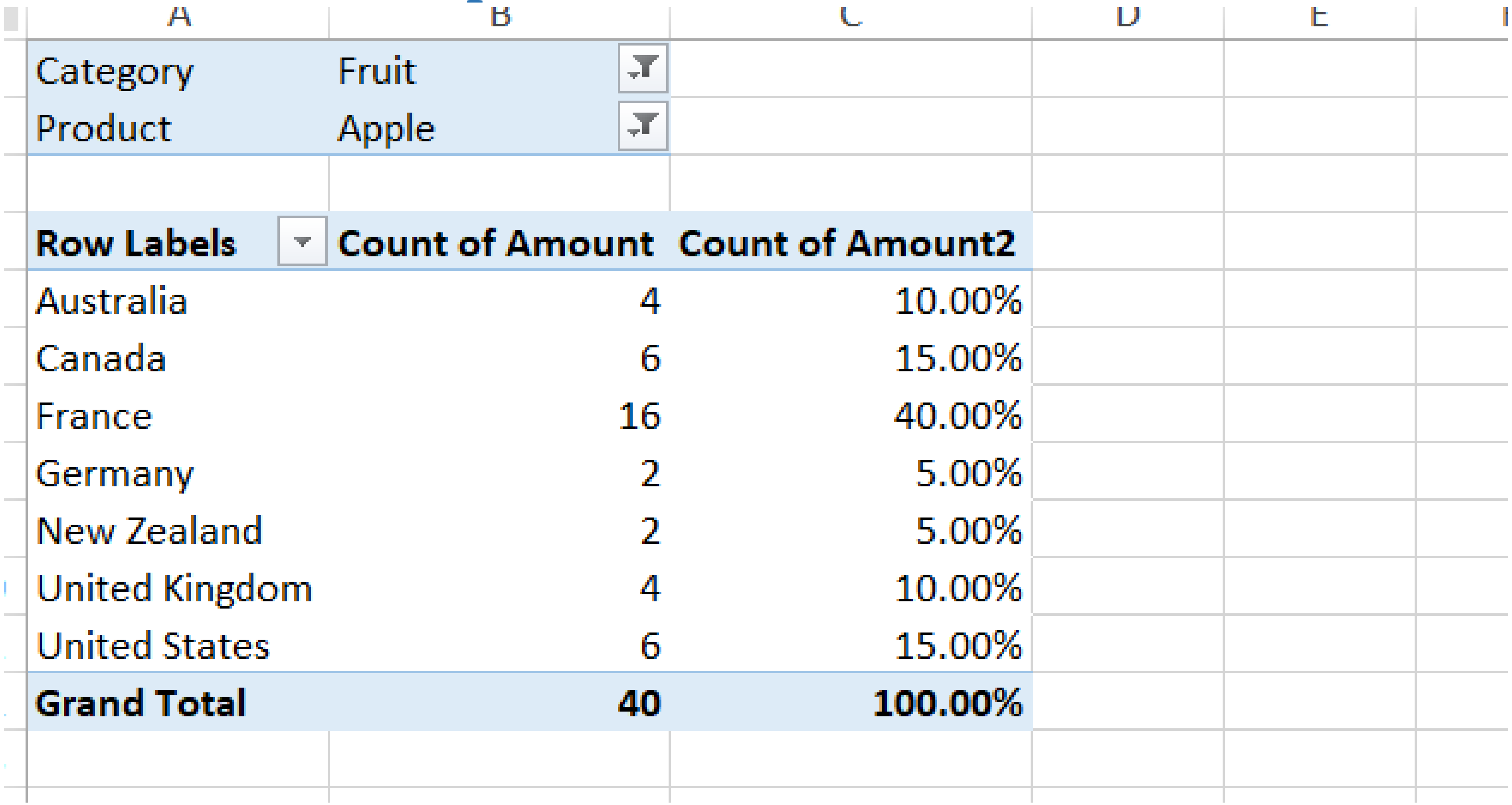
## Pivot tables example



|  | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Country | New Zealand |  |  |  |
| 2 |  |  |  |  |  |
| 3 | **Row Labels** | **Count of Amount** |  |  |  |
| 4 | Banana | 8 |  |  |  |
| 5 | Orange | 3 |  |  |  |
| 6 | Apple | 2 |  |  |  |
| 7 | Broccoli | 1 |  |  |  |
| 8 | **Grand Total** | **14** |  |  |  |
| 9 |  |  |  |  |  |
| 10 |  |  |  |  |  |
| 11 |  |  |  |  |  |
| 12 |  |  |  |  |  |
| 13 |  |  |  |  |  |

# Applications

## Pivot tables example

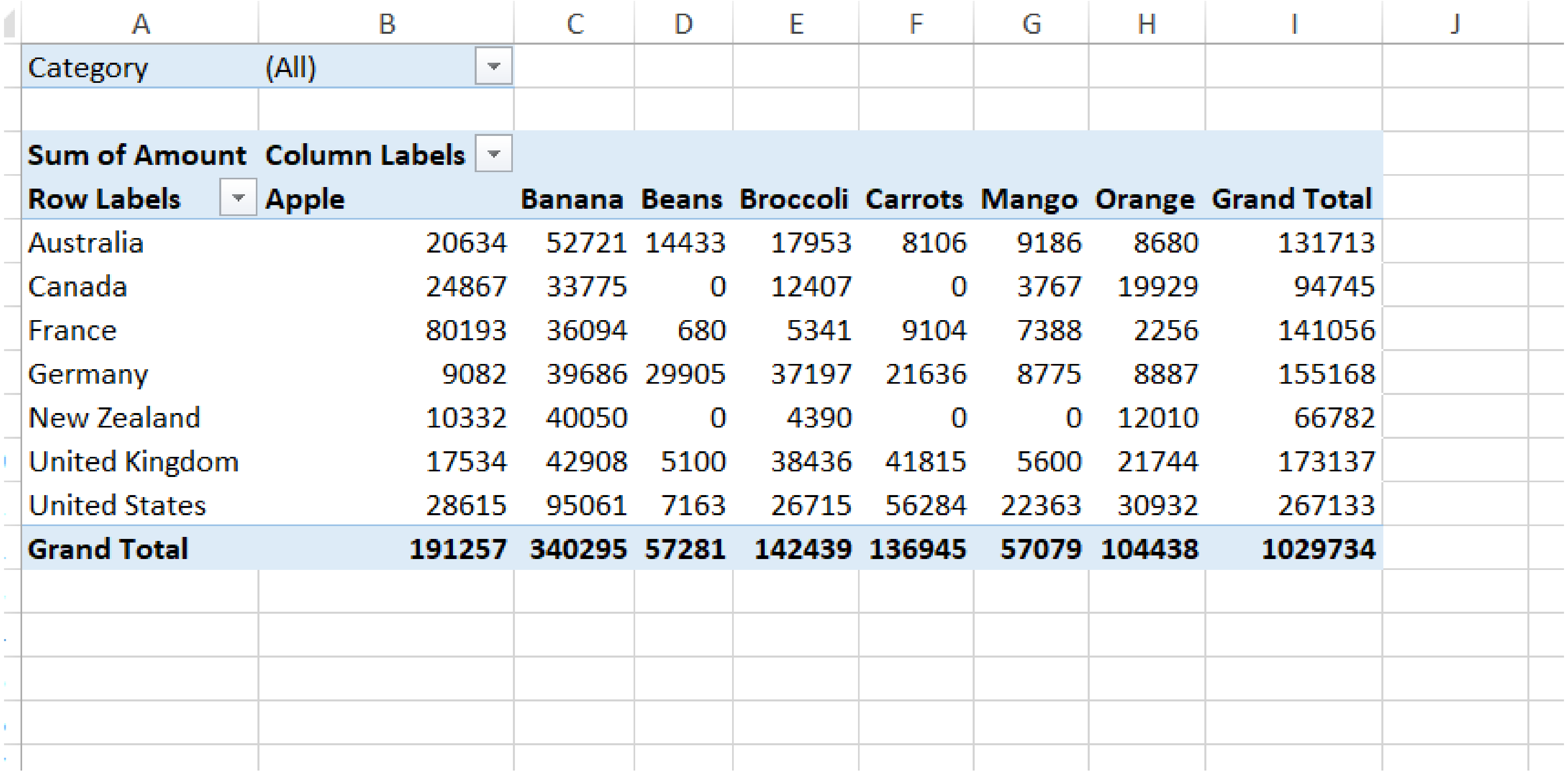


| | |
|---|---|
| Category | Fruit |
| Product | Apple |

| Row Labels | Count of Amount | Count of Amount2 |
|---|---|---|
| Australia | 4 | 10.00% |
| Canada | 6 | 15.00% |
| France | 16 | 40.00% |
| Germany | 2 | 5.00% |
| New Zealand | 2 | 5.00% |
| United Kingdom | 4 | 10.00% |
| United States | 6 | 15.00% |
| **Grand Total** | **40** | **100.00%** |

# Applications

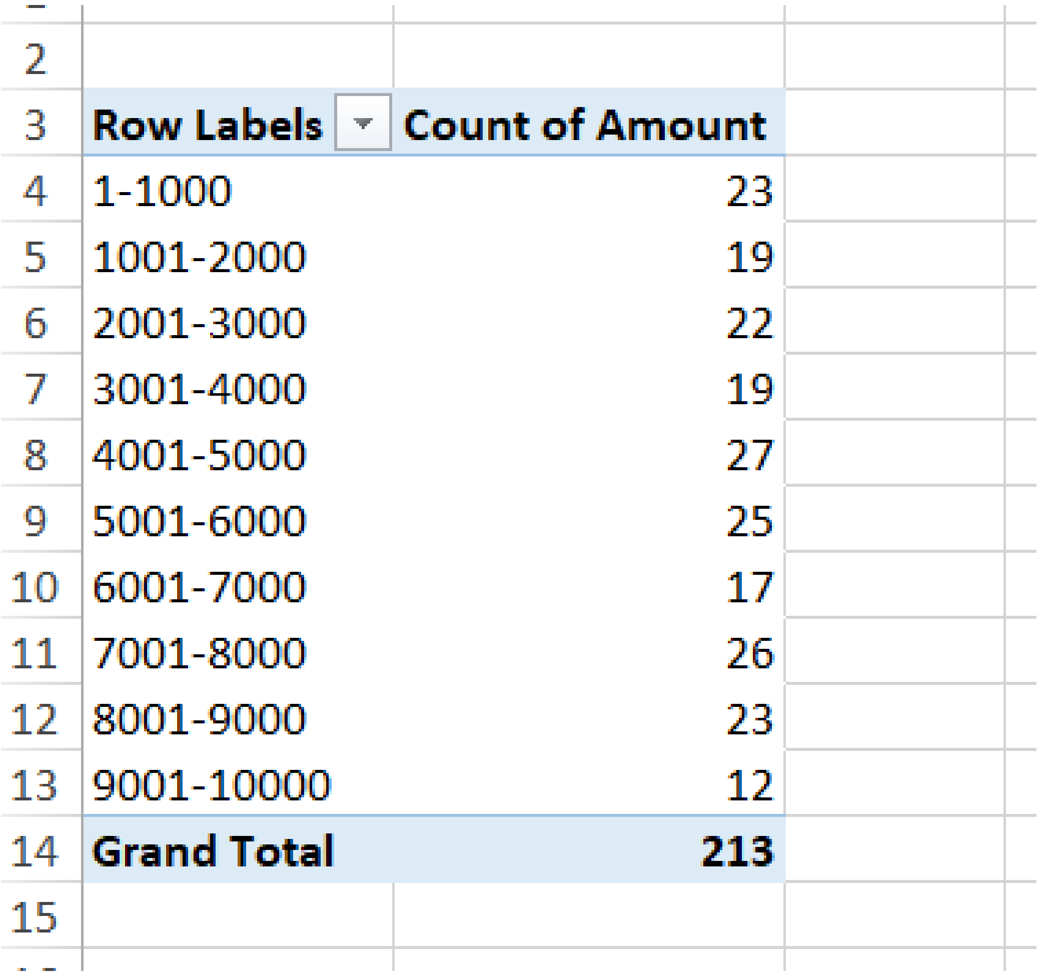
## Pivot tables example



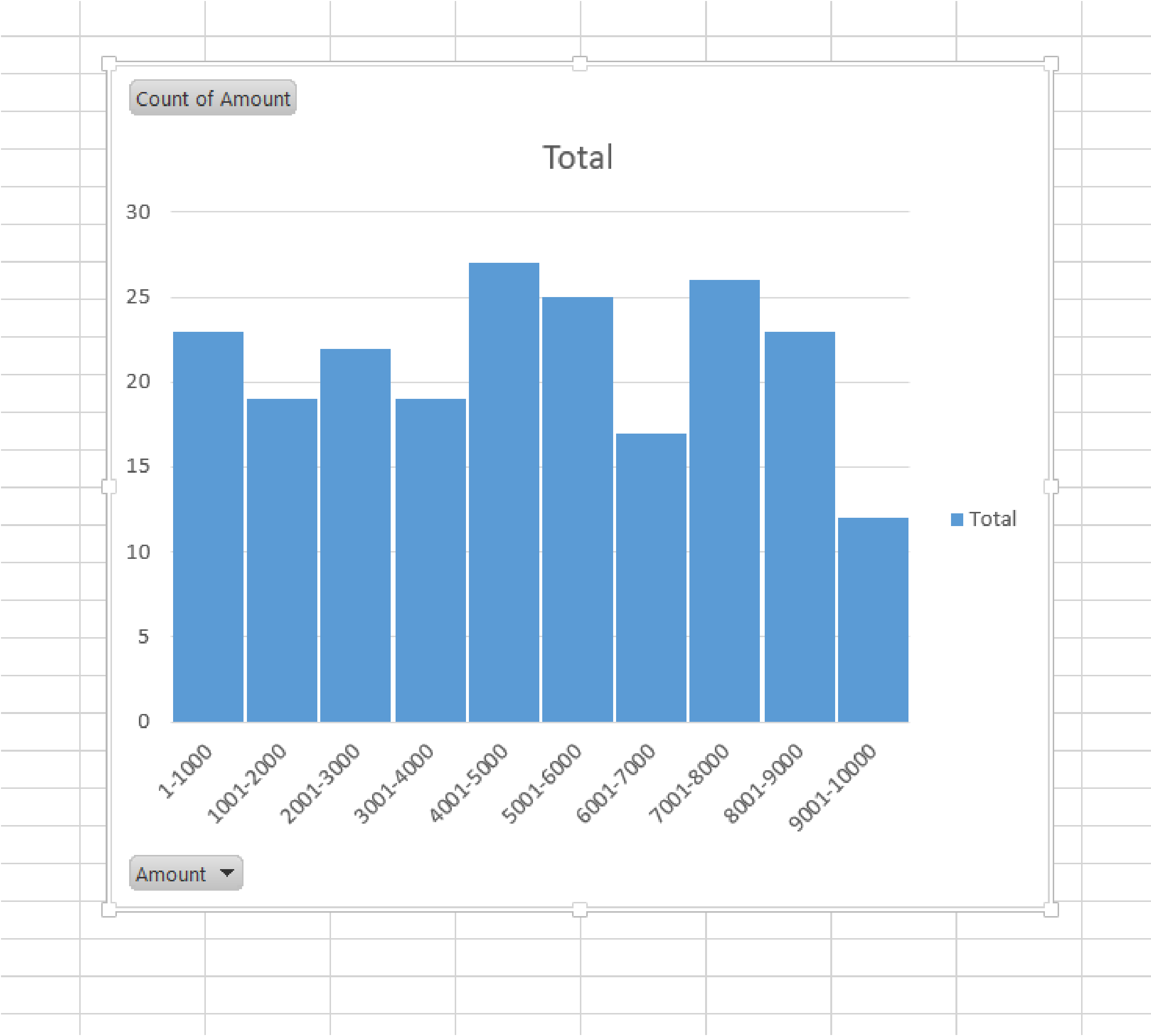
| Category | (All) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |
| **Sum of Amount** | **Column Labels** | | | | | | | |
| **Row Labels** | **Apple** | **Banana** | **Beans** | **Broccoli** | **Carrots** | **Mango** | **Orange** | **Grand Total** |
| Australia | 20634 | 52721 | 14433 | 17953 | 8106 | 9186 | 8680 | 131713 |
| Canada | 24867 | 33775 | 0 | 12407 | 0 | 3767 | 19929 | 94745 |
| France | 80193 | 36094 | 680 | 5341 | 9104 | 7388 | 2256 | 141056 |
| Germany | 9082 | 39686 | 29905 | 37197 | 21636 | 8775 | 8887 | 155168 |
| New Zealand | 10332 | 40050 | 0 | 4390 | 0 | 0 | 12010 | 66782 |
| United Kingdom | 17534 | 42908 | 5100 | 38436 | 41815 | 5600 | 21744 | 173137 |
| United States | 28615 | 95061 | 7163 | 26715 | 56284 | 22363 | 30932 | 267133 |
| **Grand Total** | **191257** | **340295** | **57281** | **142439** | **136945** | **57079** | **104438** | **1029734** |

# Applications

## Pivot tables example

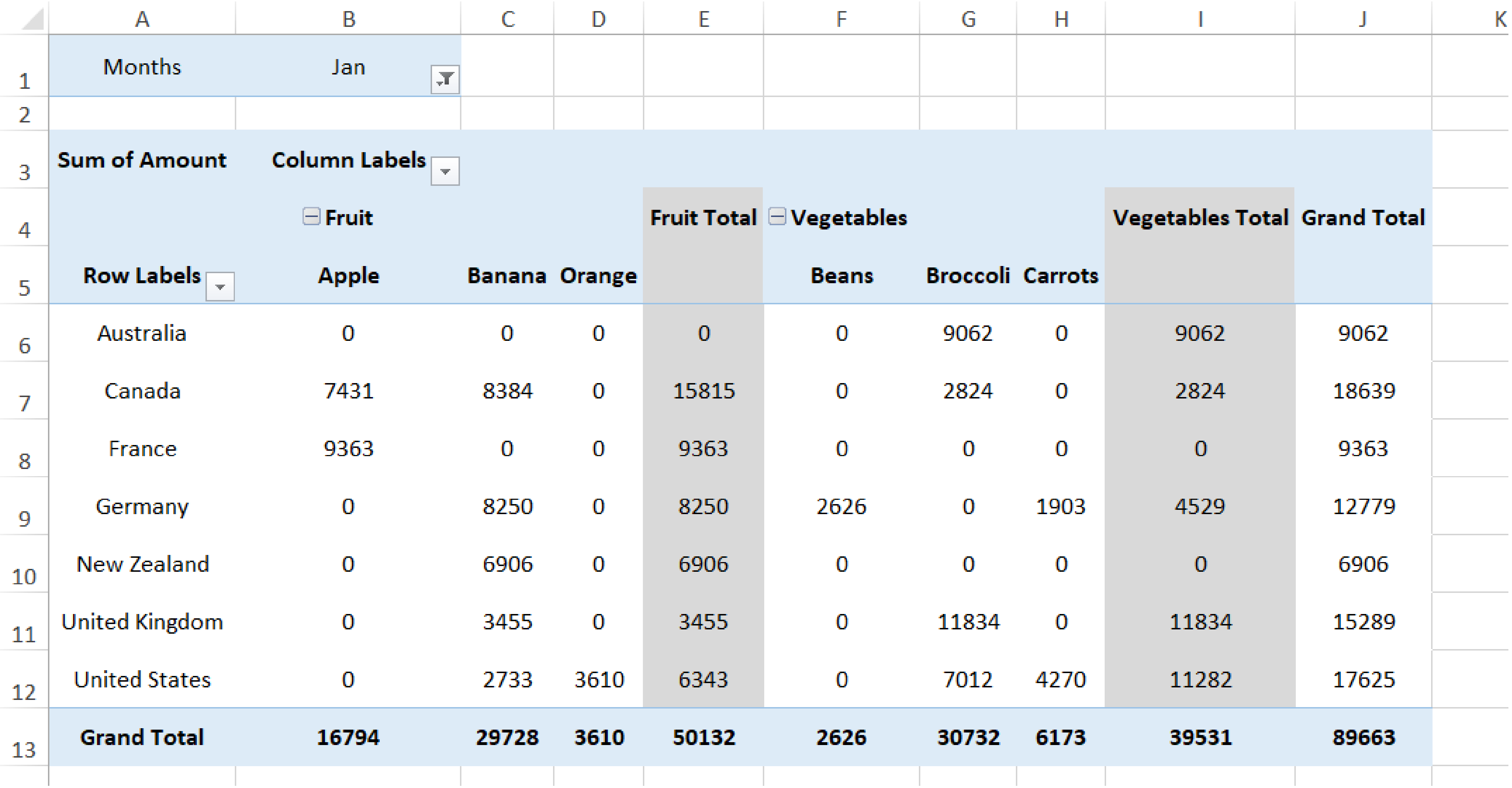


| | | |
|---|---|---|
| 2 | | |
| 3 | Row Labels | Count of Amount |
| 4 | 1-1000 | 23 |
| 5 | 1001-2000 | 19 |
| 6 | 2001-3000 | 22 |
| 7 | 3001-4000 | 19 |
| 8 | 4001-5000 | 27 |
| 9 | 5001-6000 | 25 |
| 10 | 6001-7000 | 17 |
| 11 | 7001-8000 | 26 |
| 12 | 8001-9000 | 23 |
| 13 | 9001-10000 | 12 |
| 14 | Grand Total | 213 |
| 15 | | |



# Assignments # 2

1- Based on the given data used in the pivot table examples design the following pivot table.



|  | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Months | Jan |  |  |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |  |  |  |  |
| 3 | **Sum of Amount** | **Column Labels** |  |  |  |  |  |  |  |  |  |
| 4 |  | **⊟ Fruit** |  |  | **Fruit Total** | **⊟ Vegetables** |  |  | **Vegetables Total** | **Grand Total** |  |
| 5 | **Row Labels** | **Apple** | **Banana** | **Orange** |  | **Beans** | **Broccoli** | **Carrots** |  |  |  |
| 6 | Australia | 0 | 0 | 0 | 0 | 0 | 9062 | 0 | 9062 | 9062 |  |
| 7 | Canada | 7431 | 8384 | 0 | 15815 | 0 | 2824 | 0 | 2824 | 18639 |  |
| 8 | France | 9363 | 0 | 0 | 9363 | 0 | 0 | 0 | 0 | 9363 |  |
| 9 | Germany | 0 | 8250 | 0 | 8250 | 2626 | 0 | 1903 | 4529 | 12779 |  |
| 10 | New Zealand | 0 | 6906 | 0 | 6906 | 0 | 0 | 0 | 0 | 6906 |  |
| 11 | United Kingdom | 0 | 3455 | 0 | 3455 | 0 | 11834 | 0 | 11834 | 15289 |  |
| 12 | United States | 0 | 2733 | 3610 | 6343 | 0 | 7012 | 4270 | 11282 | 17625 |  |
| 13 | **Grand Total** | **16794** | **29728** | **3610** | **50132** | **2626** | **30732** | **6173** | **39531** | **89663** |  |

# References.

[1] L. Zemmouchi-Ghomari, “Basic concepts of information systems,” In: “ Contemporary Issues in Information Systems - A Global Perspective,” 2021.

[2] M. Krčál and M. Kubiš , “Differences between Knowledge and Information Management Practices: Empirical Investigation.”, 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016), vol. 3, pp. 190-198, 2016.

[3] J. D. Kelleher and B. Tierney, “Data science”, The MIT Pres, Cambridge, Massachusetts, 2018.

[4] S. D. Köseoğlu, W. M. Ead, and M. M. Abbassy, “Basics of Financial Data Analytics”, In: Financial Data Analytics, S. D. Köseoğlu, pp. 23-57, 2022.

[5] S. Praveen, U. Chandra, “ Influence of Structured, Semi- Structured, Unstructured data on various data models”, - International Journal of Scientific and Engineering, vol. 8, pp. 67-69, 2020.

[6] C. Shah, “ A Hands-On Introduction to Data Science.”, Cambridge University Press, 2020.

[7] M. Islam, “Data Analysis: Types, Process, Methods, Techniques and Tools, ” International Journal on Data Science and Technology, vol. 6, no. 1, pp. 10-15, 2020.

[8] D. Srivastava, “An Introduction to Data Visualization Tools and Techniques in Various Domains,” International Journal of Computer Trends and Technology, vol. 71, pp. 125-130, 2023.