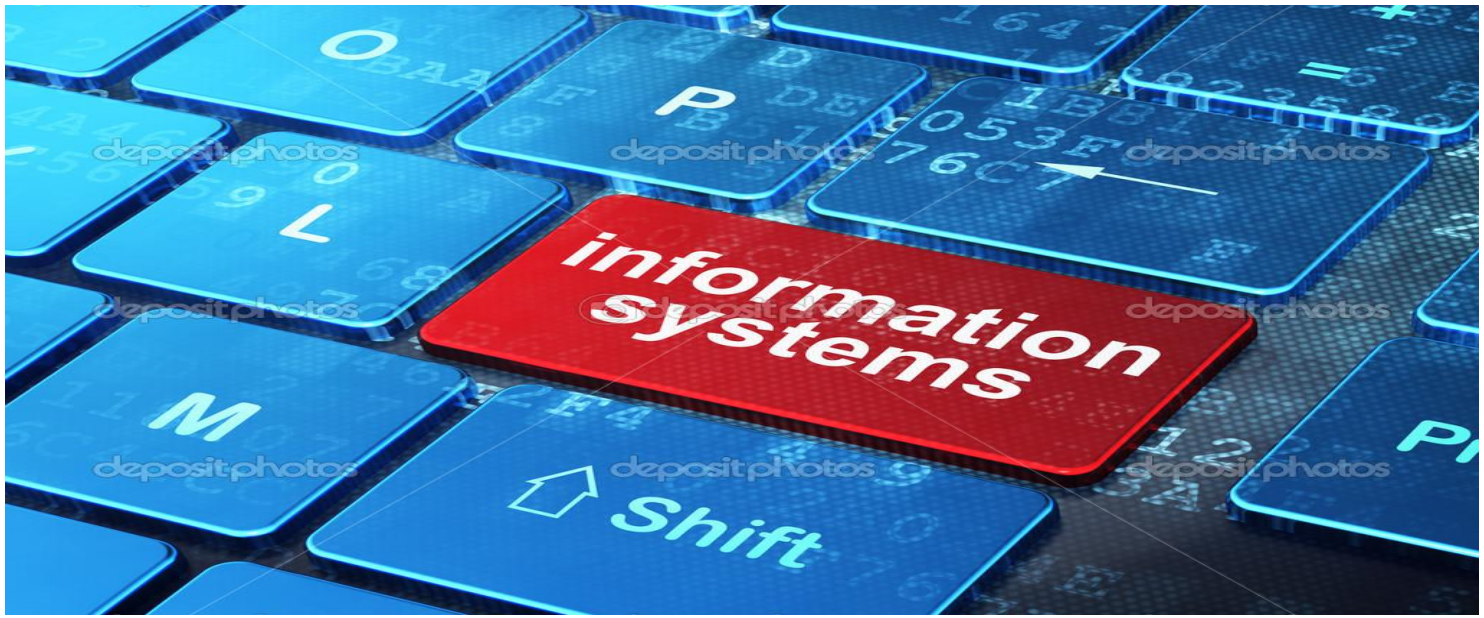
Advanced Topics in Information systems



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:

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

Where $\bar{x}$ 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:

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

Where $\bar{x}$ 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:

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

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

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

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

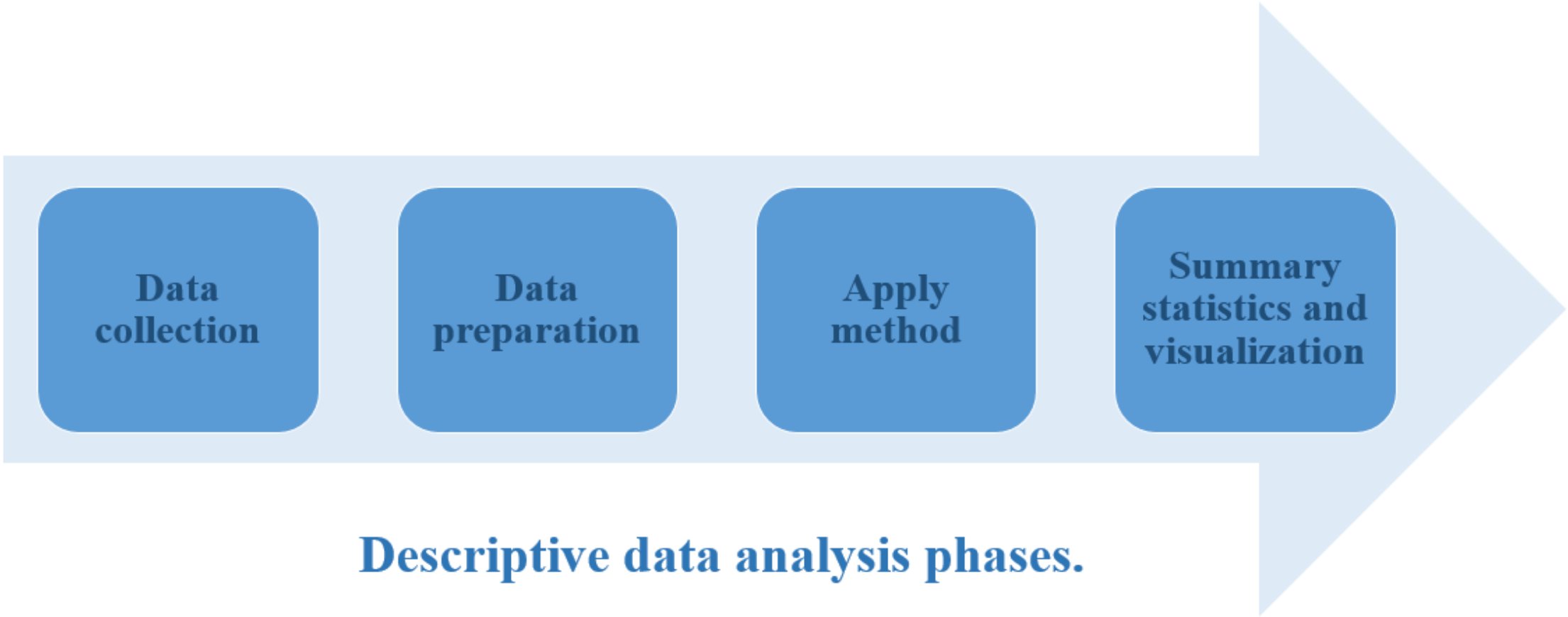
x	$(x-\bar{x})$	$(x-\bar{x})^2$
10	-6	36
11	-5	25
15	-1	1
20	4	16
24	8	64
total	0	142

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

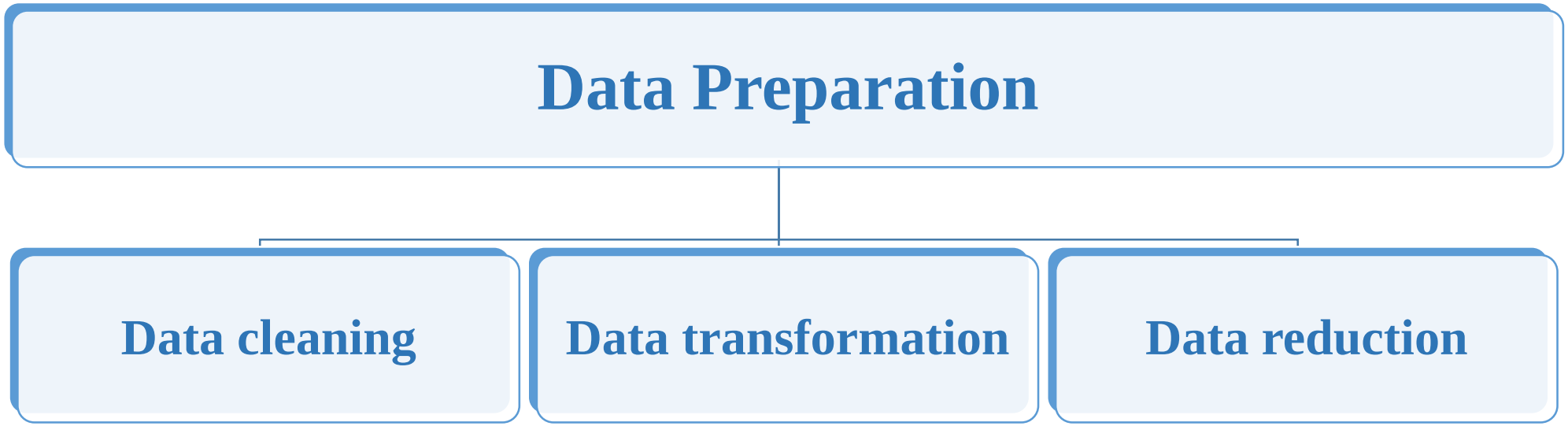
The first phase in descriptive data analysis is collecting relevant data. This process involves identifying data sources, selecting appropriate data-collecting methods, and verifying that the data acquired accurately represents the population or topic of interest.

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

Descriptive data analysis phases cont.

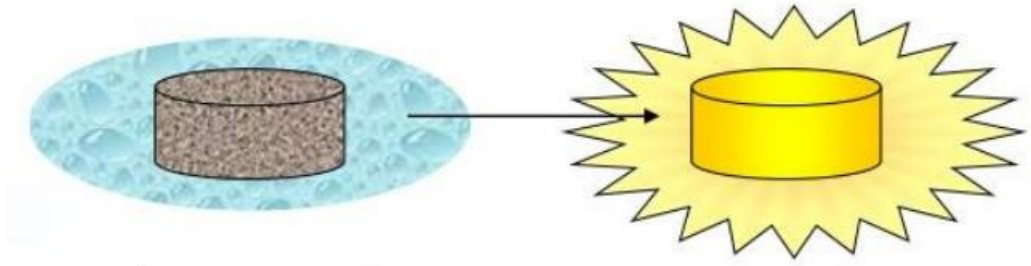
2-Data Preparation

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.

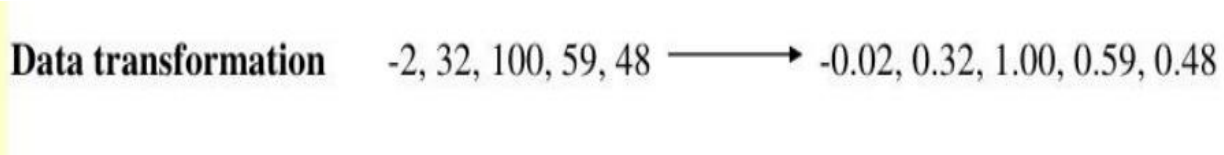
- **Inconsistent**. refers to the lack of uniformity in data format or content, data contains discrepancies.

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

Data Preparation cont.

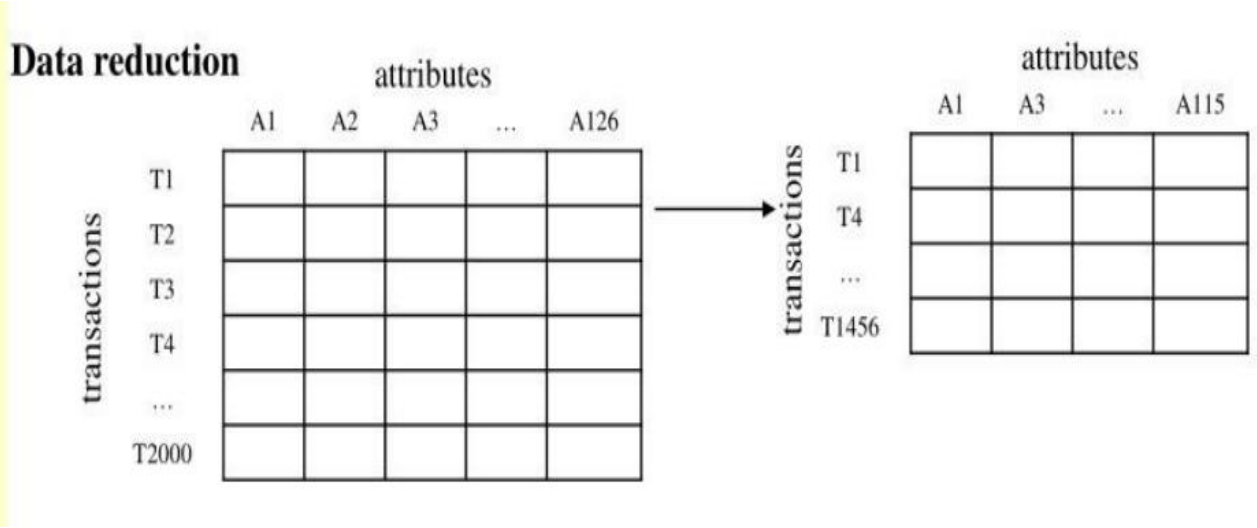
- **Data Transformation:**

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.

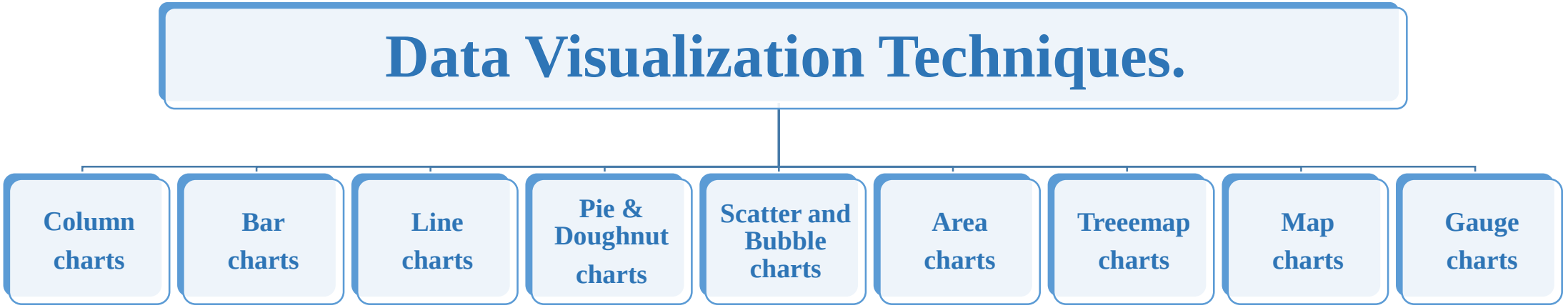
Data visualization.

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

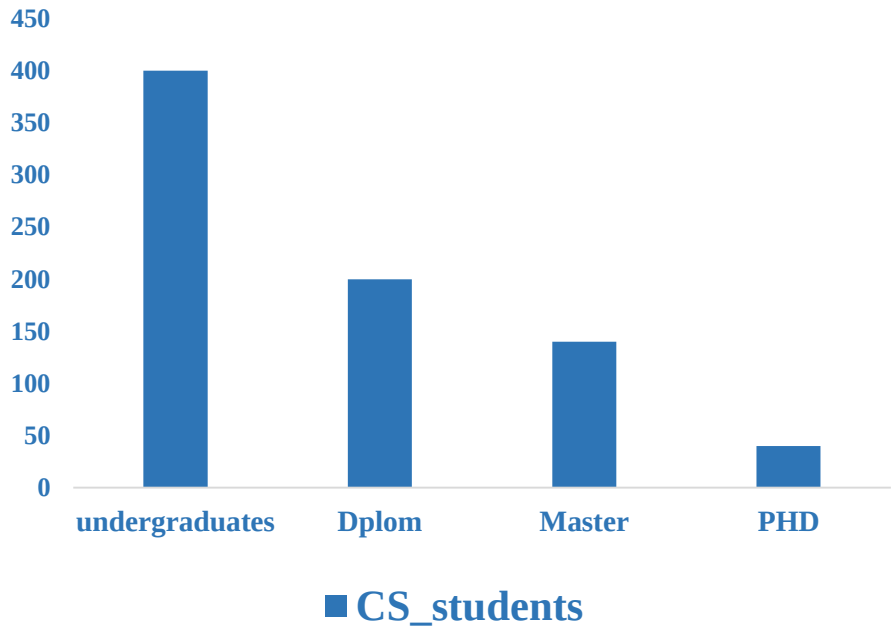
There most common data visualization techniques:

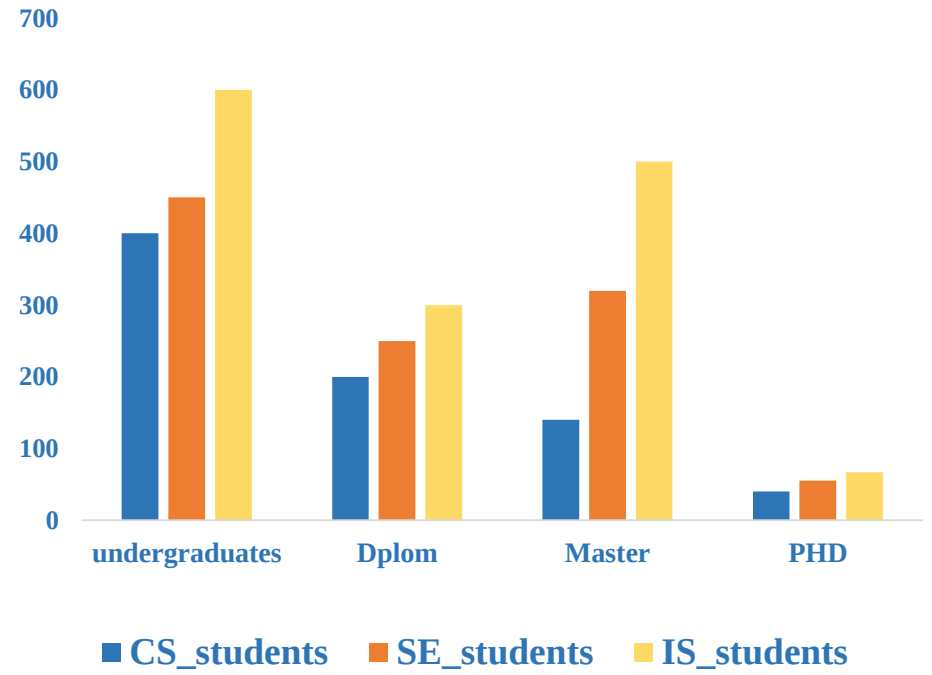


Data Visualization Techniques cont.

Column charts.

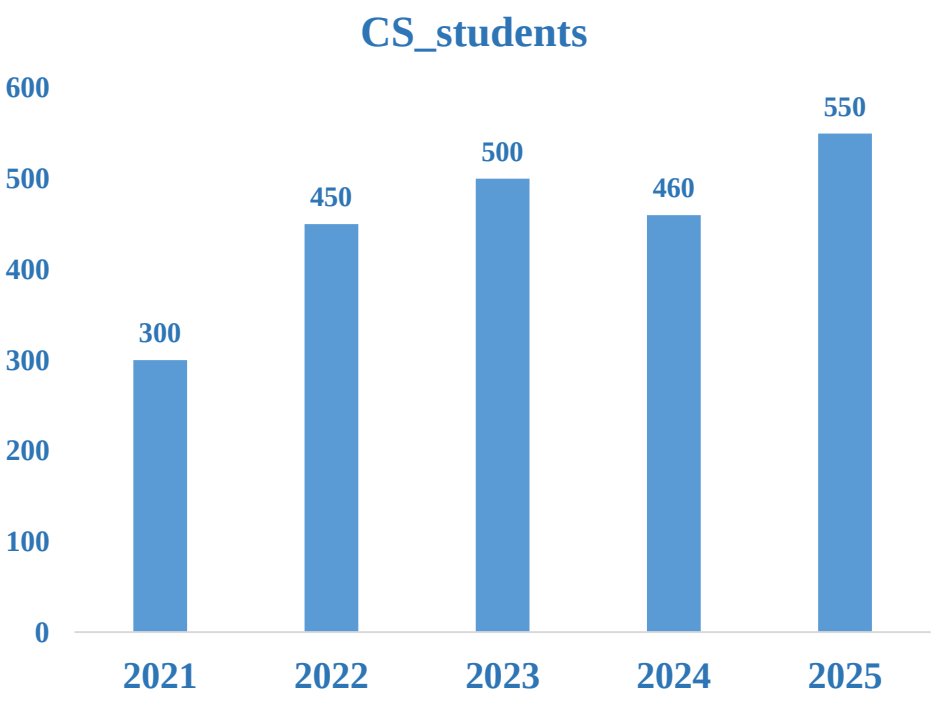
The column chart is the most used chart type, With column charts you could compare values for different categories or compare value changes over a period of time for a single category.



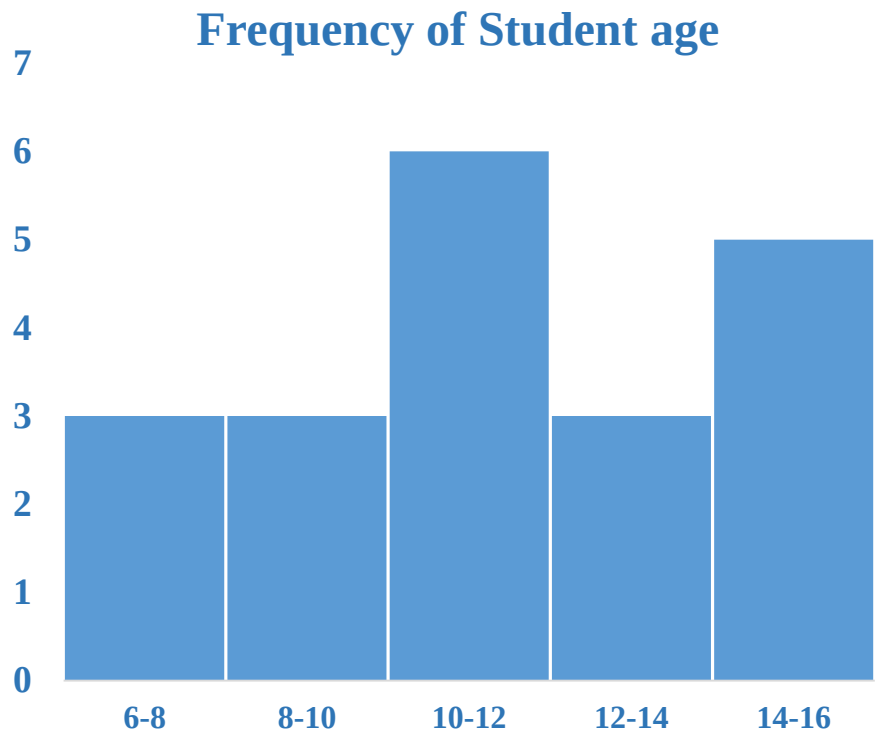


Column chart compare values for different categories.

Column charts cont.

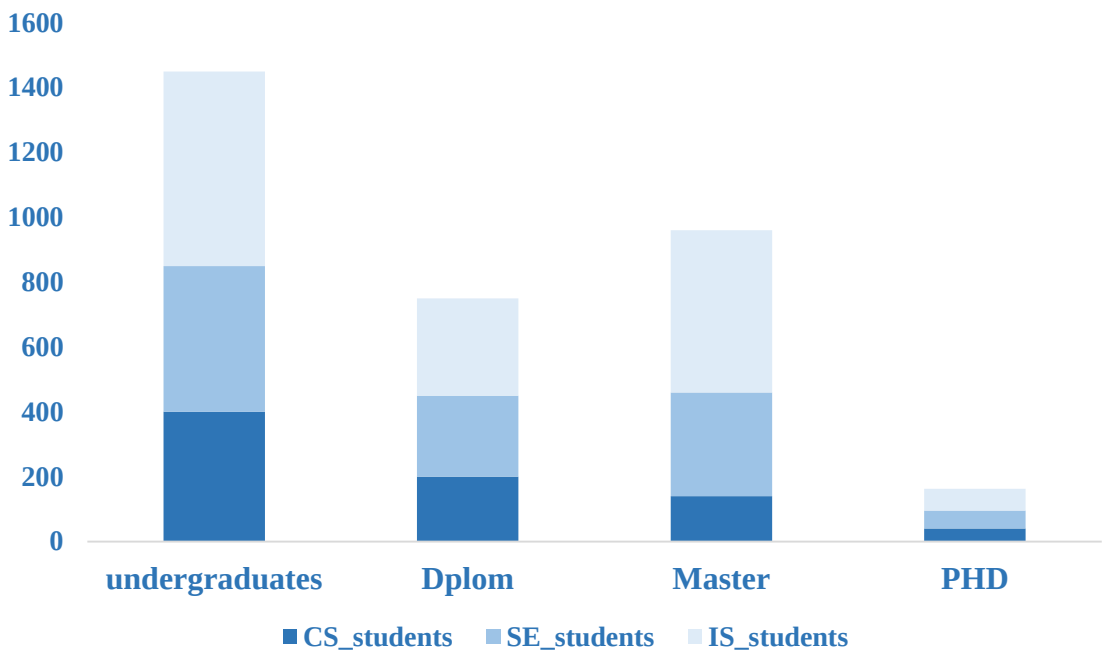


Column chart compares value changes over a period of time .



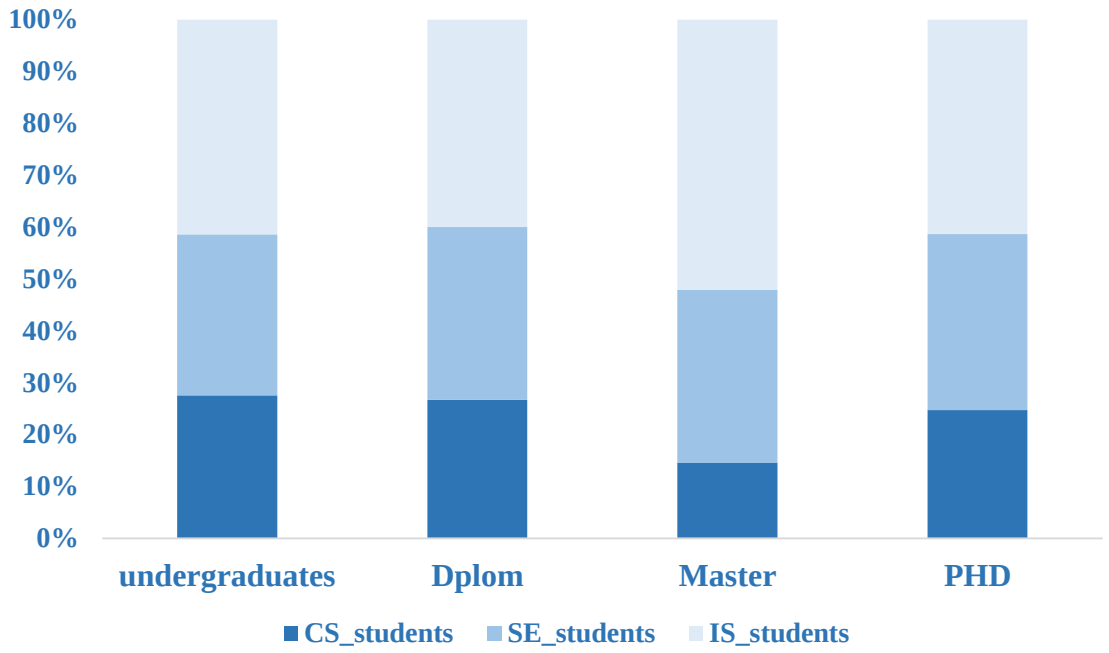
Column chart is used to represent Histogram.

Column charts cont.



Stacked column chart.

It compares parts of whole

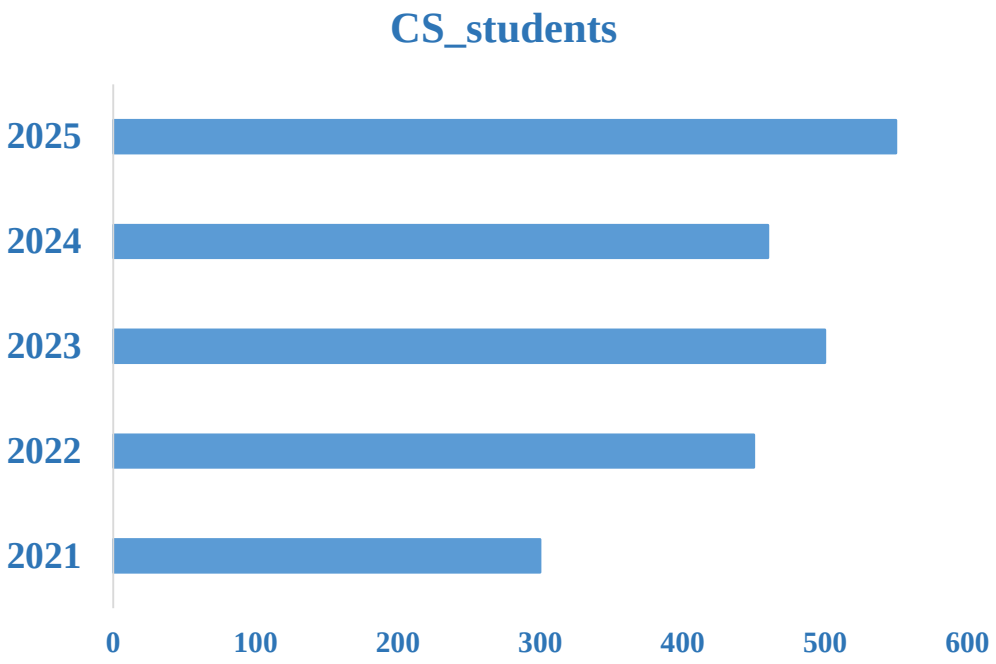


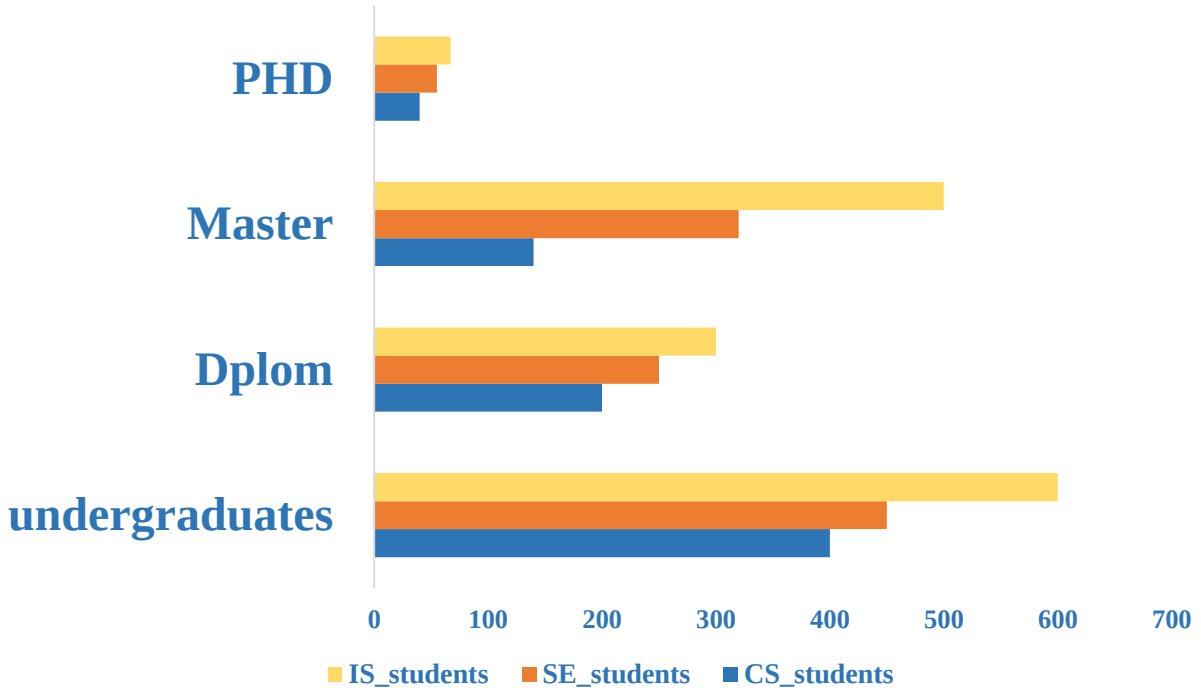
Fully stacked column chart.

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

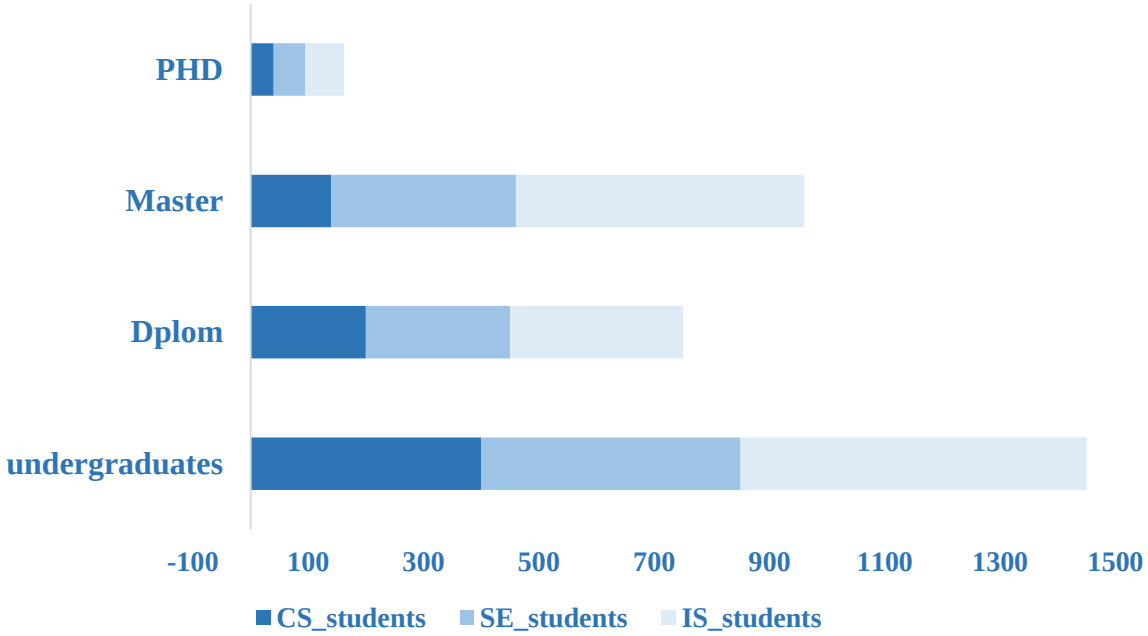
Bar charts.

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



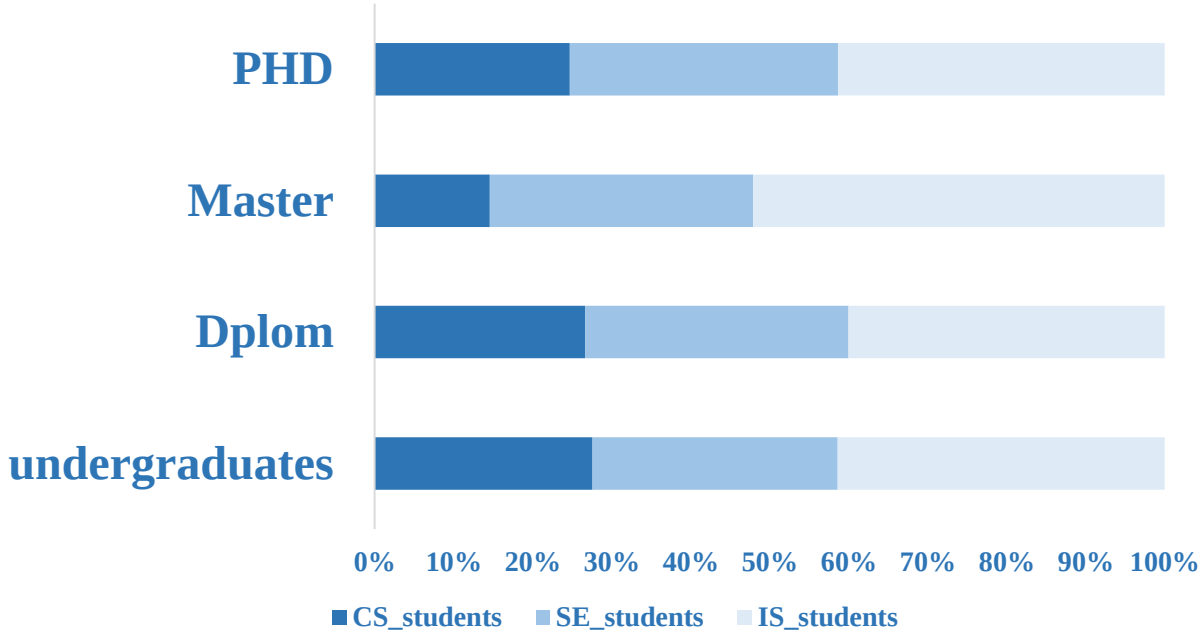


Bar charts cont.



Stacked bar chart.

It compares parts of whole

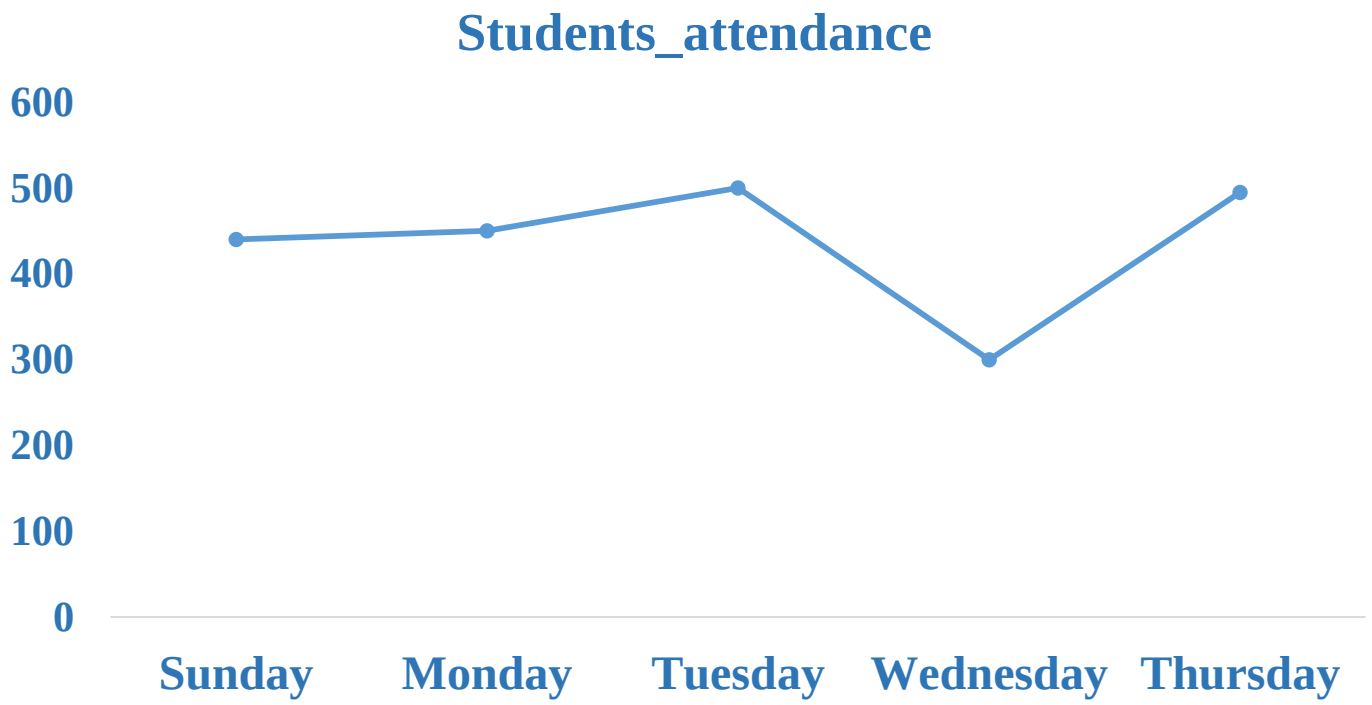


Fully stacked bar chart.

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

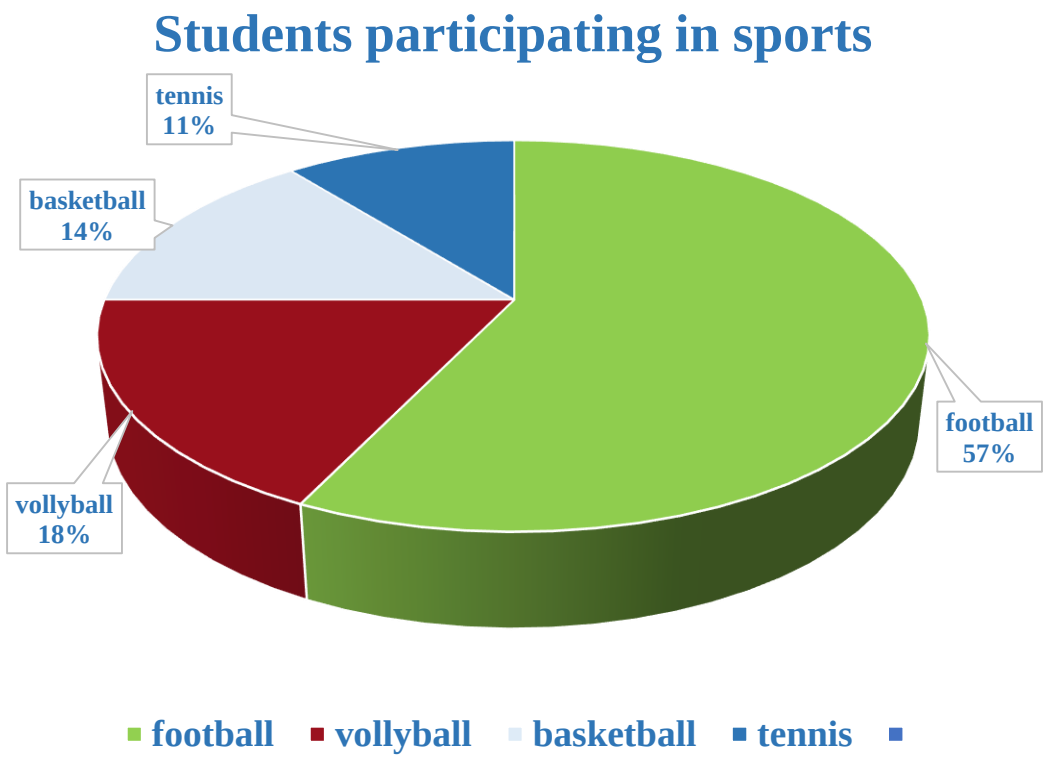
Line charts.

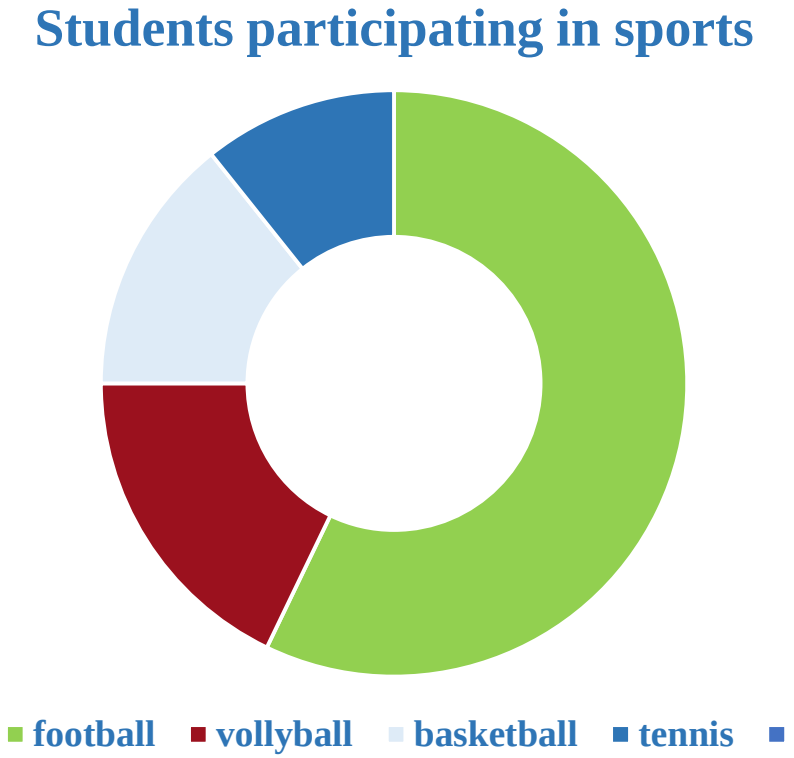
The line chart is used to show trends over time (years, months, and days) or categories



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

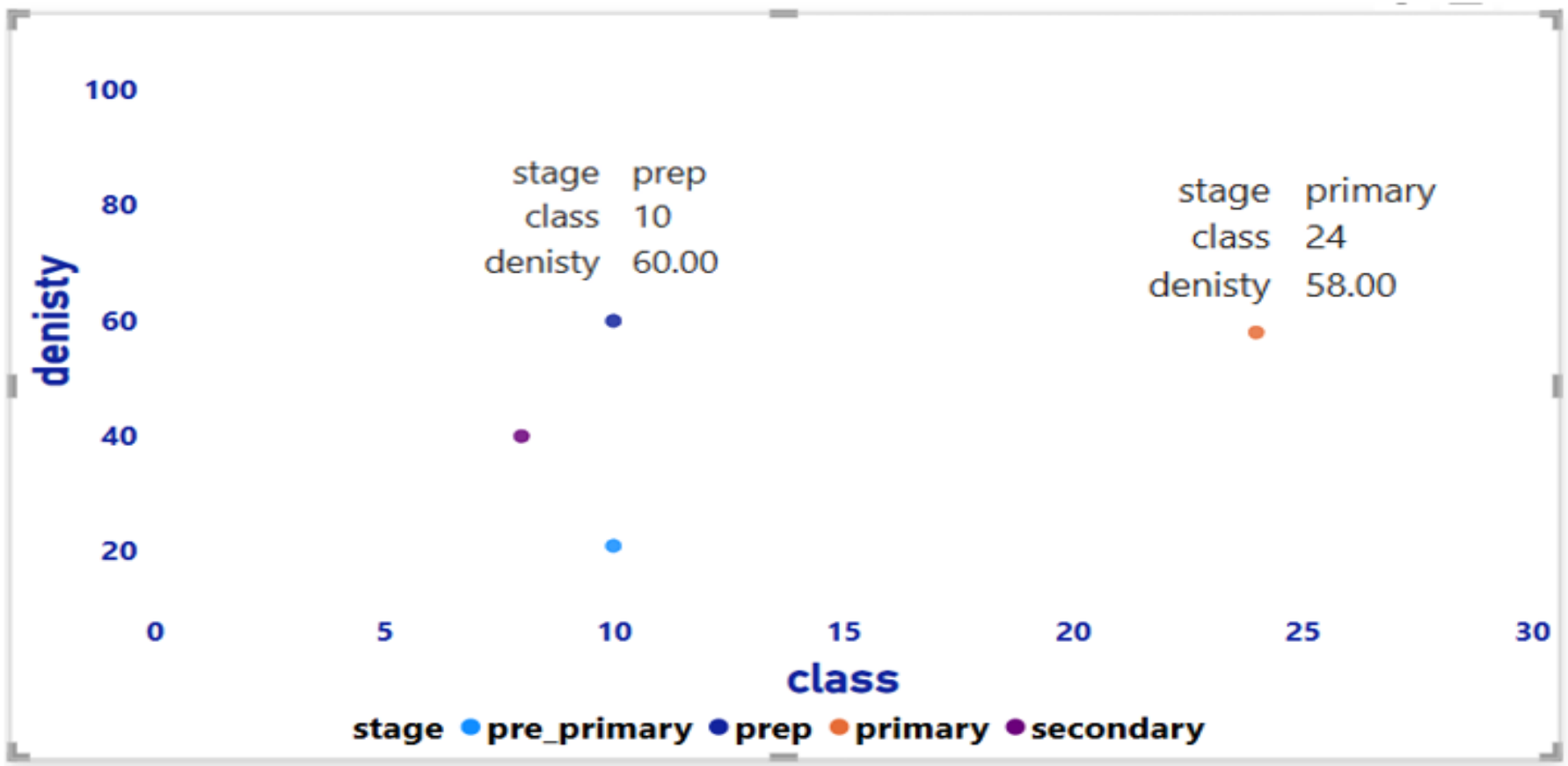




Scatter and Bubble charts

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.

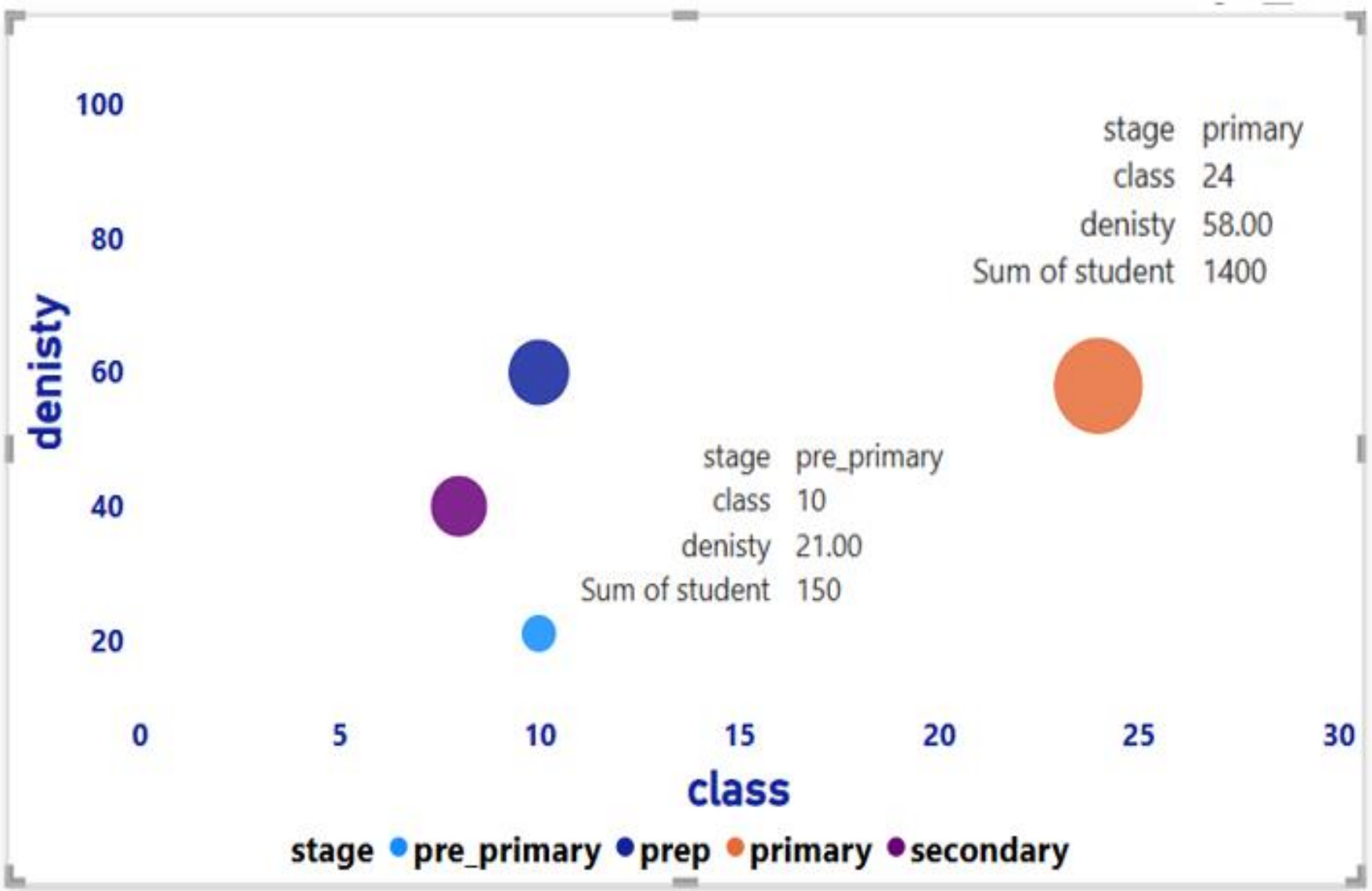
Scatter chart example



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.

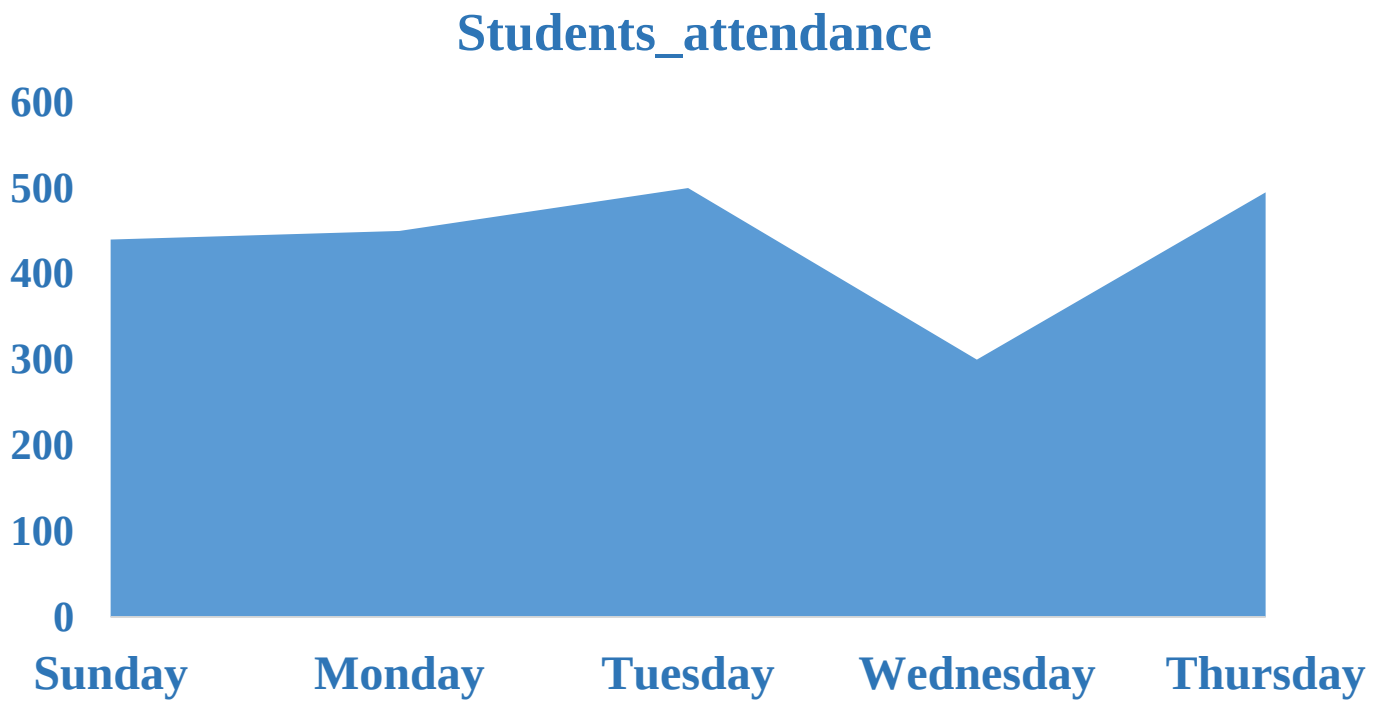
Bubble charts example



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

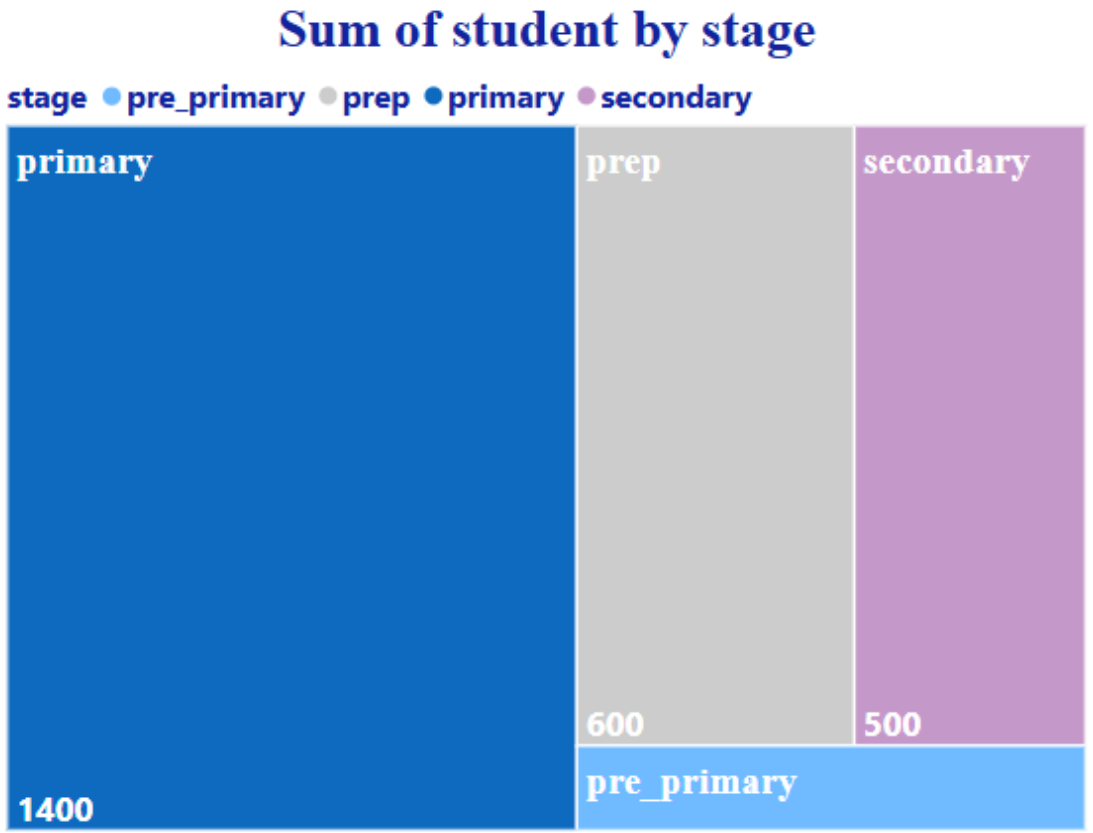
Area charts.

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.



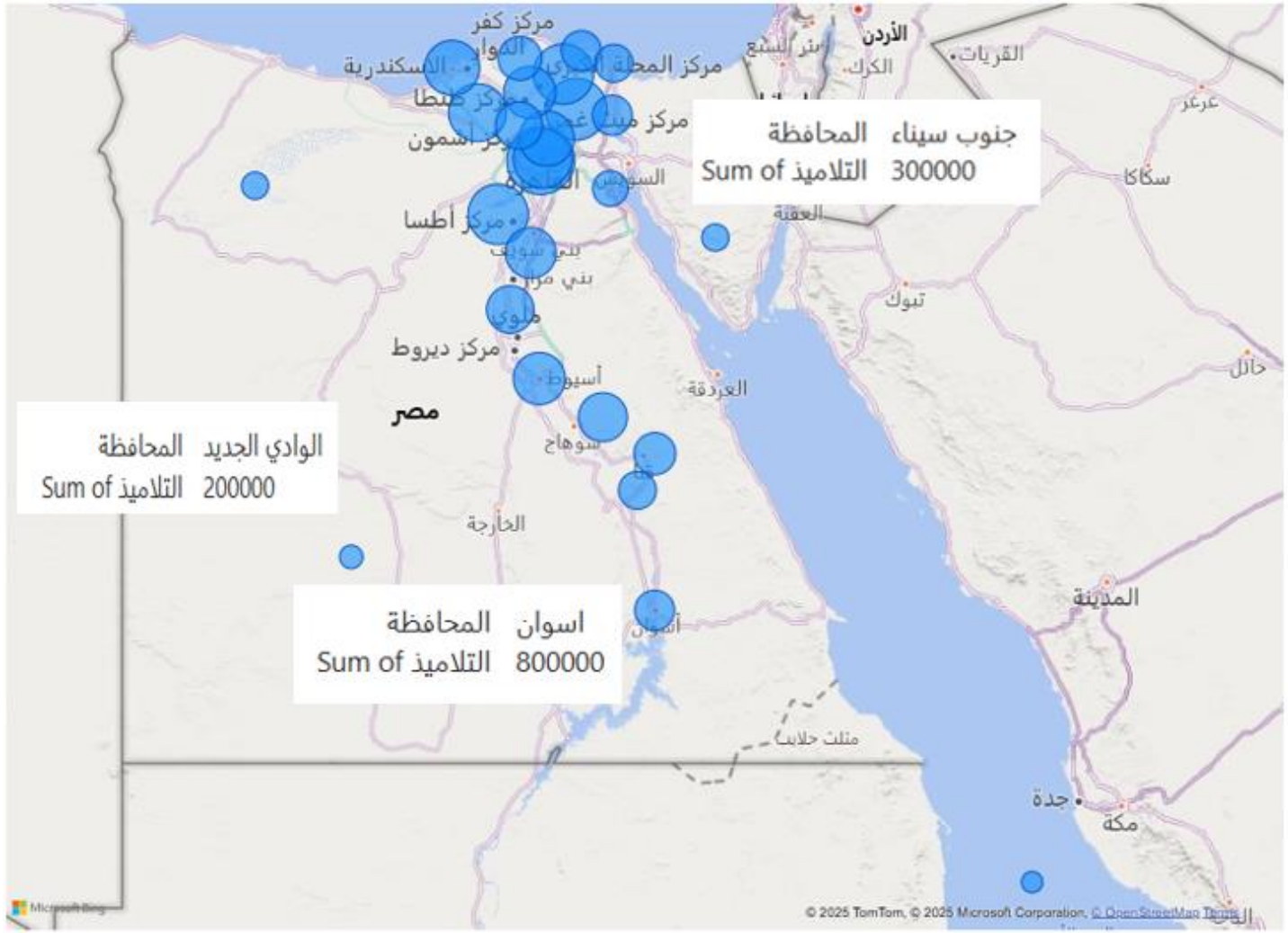
Treemap charts.

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.



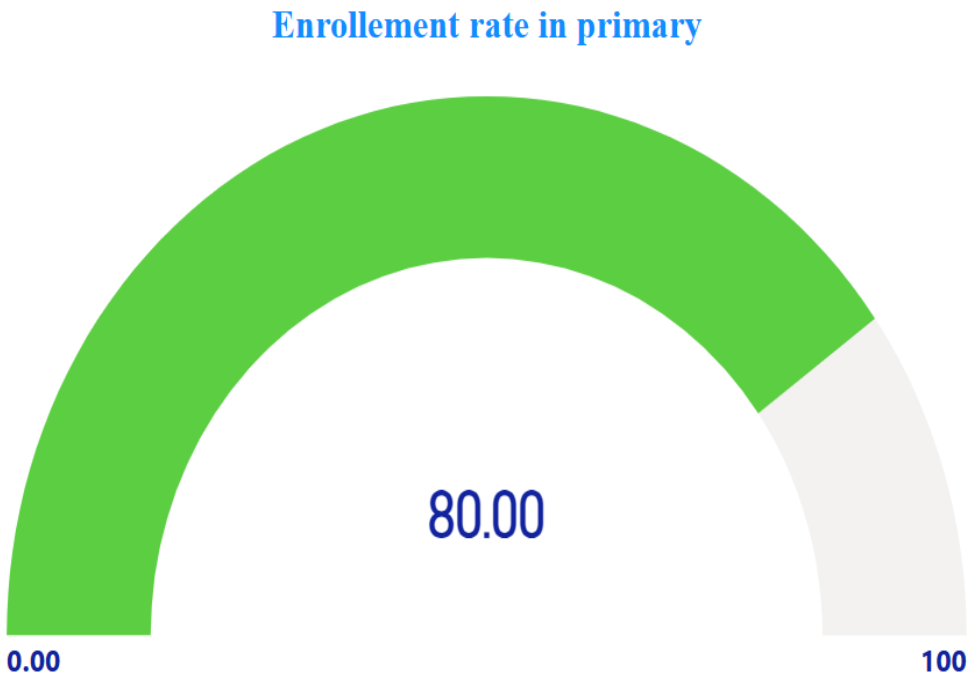
Map charts.

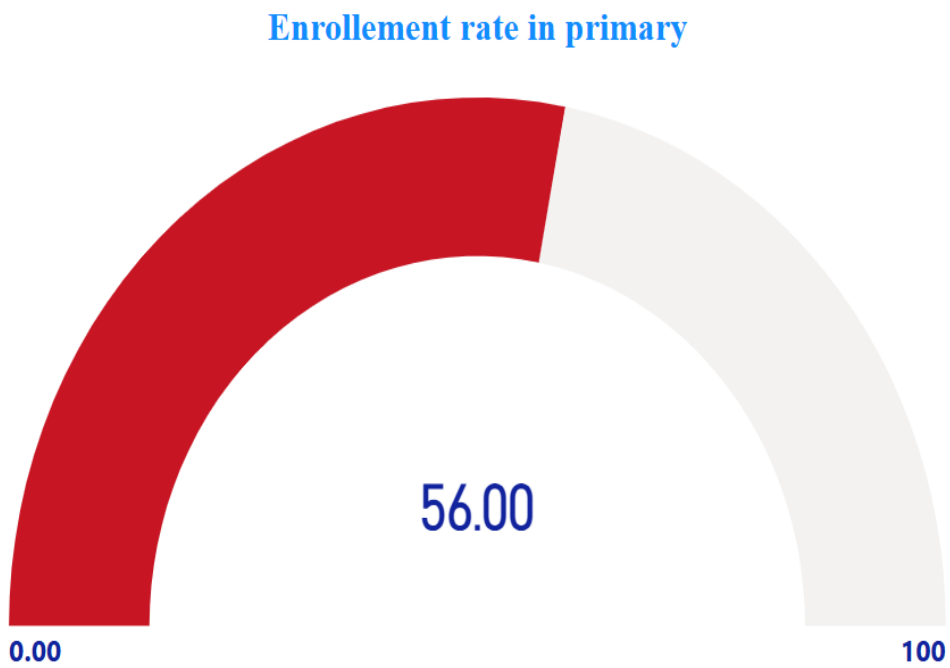
The map chart is used to visualize the geographic distribution of your data.



Gauge charts.

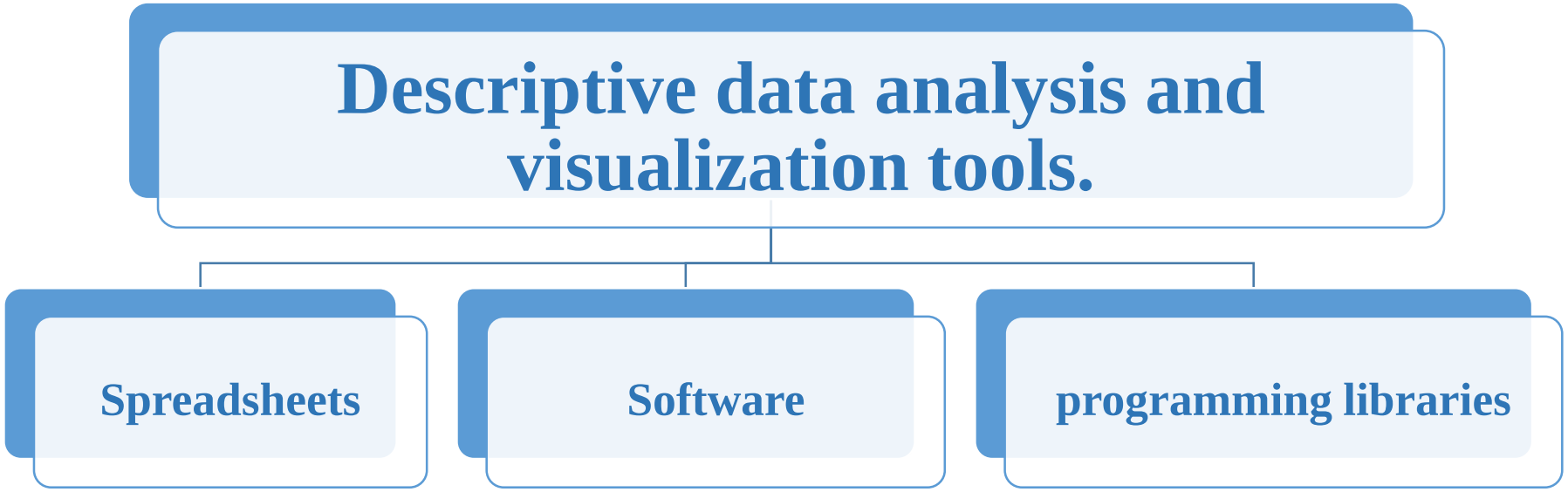
Gauge charts are good for displaying KPIs (Key Performance Indicators). They typically display a single key value, comparing it to a color-coded performance level indicator, typically showing green for “good” and red for “trouble.” Gauges are a great choice to show progress toward a goal, and Represent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

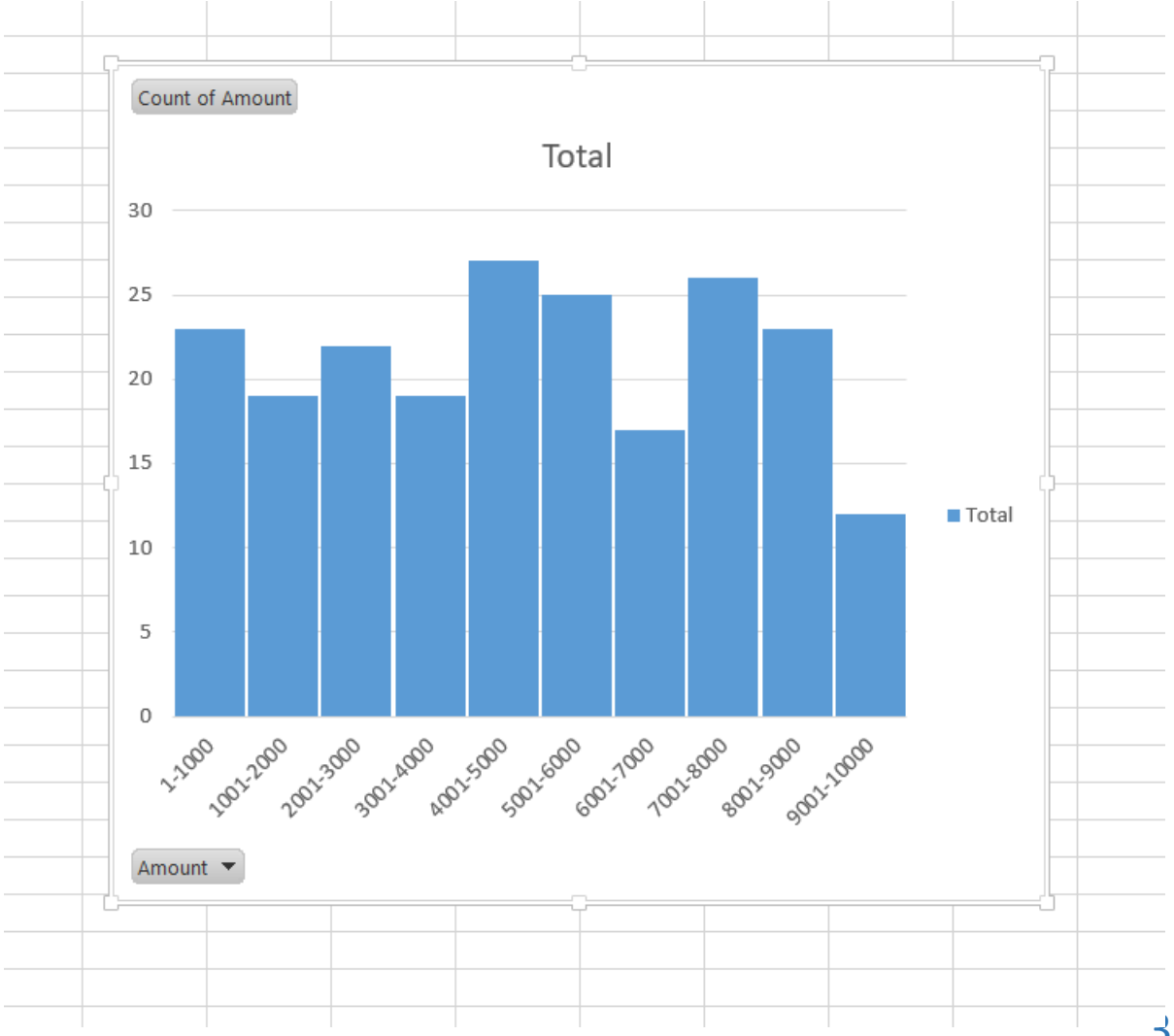
Pivot tables example

A	B	C	D	E	F	G	H	I	J
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

Row Labels	Count of Amount
1-1000	23
1001-2000	19
2001-3000	22
3001-4000	19
4001-5000	27
5001-6000	25
6001-7000	17
7001-8000	26
8001-9000	23
9001-10000	12
Grand Total	**213**



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.