Data Visualization

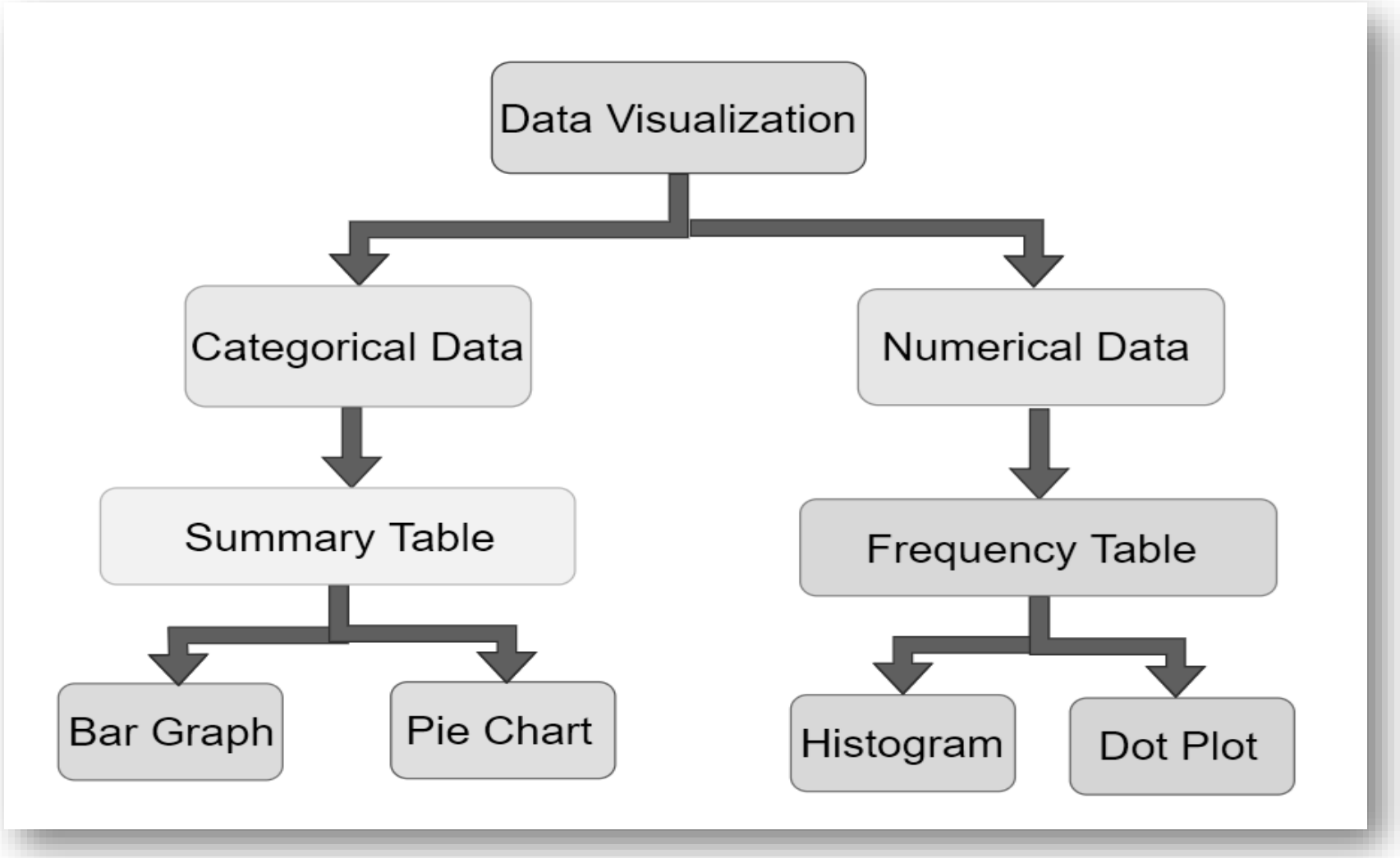
Neba Nfonsang

University of Denver

Data Visualization

- Summary Table
- Bar Graph and Pie Chart
- Frequency Distribution
- Histogram
- Scatter Chart
- Box Plot
- Contingency Tables

Data Visualization



Summary Table

- **A summary table** is a table that contains categorical data and their frequencies.
- **Frequency** is the number of times a data value occurs.
- **Relative frequency** is the frequency divided by the total number of observations in the dataset
- **Percentage frequency** is relative frequency multiplied by 100%

Summary Table

Dataset

Index	Student	Major
1	Sophie	Accounting
2	Abigail	Statistics
3	Daniel	Economics
4	Ryan	Economics
5	Jamie	Statistics
6	James	Accounting
7	Natalie	Economics
8	Daniella	Finance
9	Jack	Int'l Affairs
10	Peter	Int'l Affairs

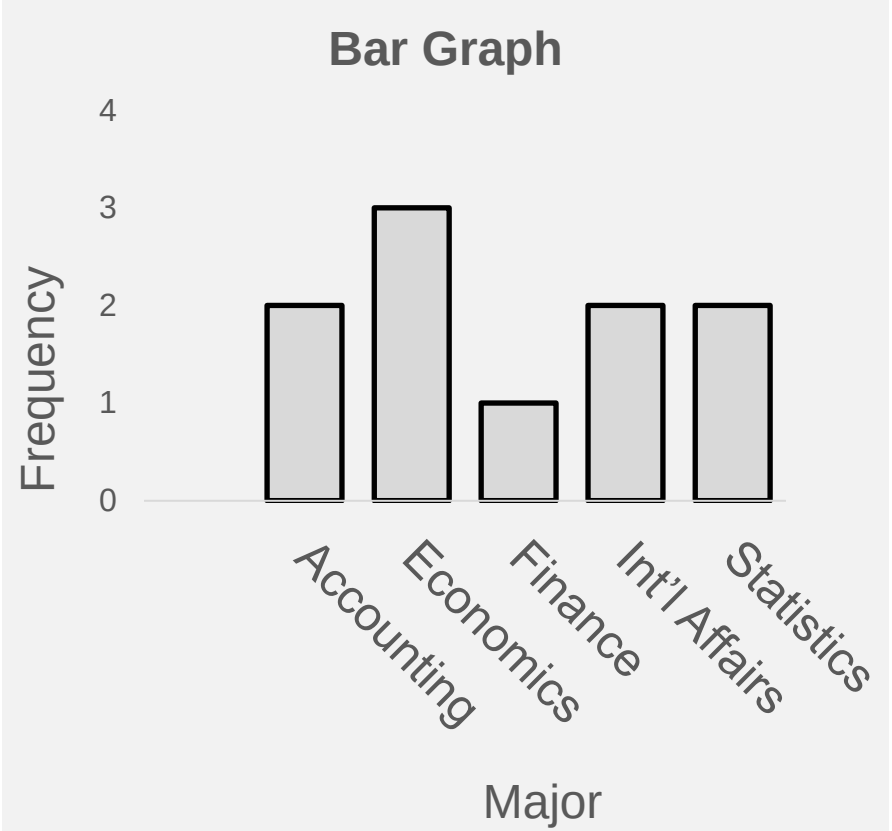
Summary Table

Major	Frequency (or Count)	Relative Frequency	Percentage Frequency
Accounting	2	0.2	20%
Economics	3	0.3	30%
Finance	1	0.1	10%
Int'l Affairs	2	0.2	20%
Statistics	2	0.2	20%

Bar Graph

A bar graph is a plot displaying categorical data and the frequencies of the data.

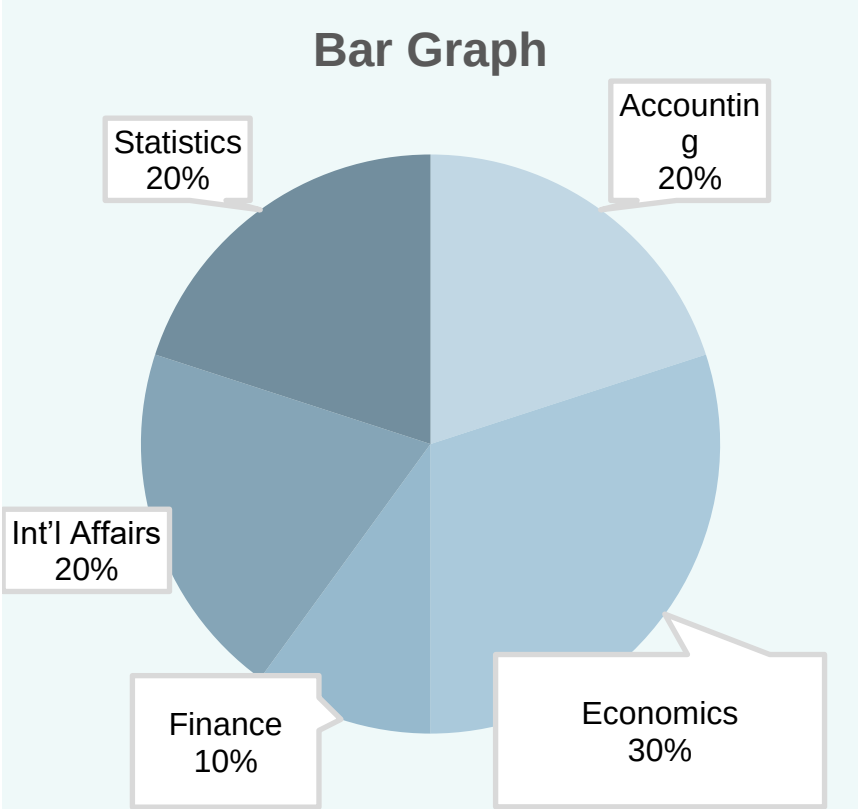
Major	Frequency (or Count)	Relative Frequency	Percentage Frequency
Accounting	2	0.2	20%
Economics	3	0.3	30%
Finance	1	0.1	10%
Int'l Affairs	2	0.2	20%
Statistics	2	0.2	20%



Pie Chart

A pie chart is display of relative frequencies, percentage frequencies or proportions of categorical data using “pie slices”.

Major	Frequency (or Count)	Relative Frequency	Percentage Frequency
Accounting	2	0.2	20%
Economics	3	0.3	30%
Finance	1	0.1	10%
Int’l Affairs	2	0.2	20%
Statistics	2	0.2	20%



Frequency Distribution

Dataset

Index	Student	Age
1	Sophie	26
2	Abigail	19
3	Daniel	25
4	Ryan	30
5	Jamie	21
6	James	30
7	Natalie	34
8	Daniella	40
9	Jack	42
10	Peter	20

Frequency Distribution

Bins	Frequency	Relative Frequency	Percentage Frequency
19 - 24	3	0.3	30%
24 - 29	2	0.2	20%
29 - 34	2	0.2	20%
34 - 39	1	0.1	10%
39 - 44	2	0.2	20%

* We want to have 5 bins for this data.

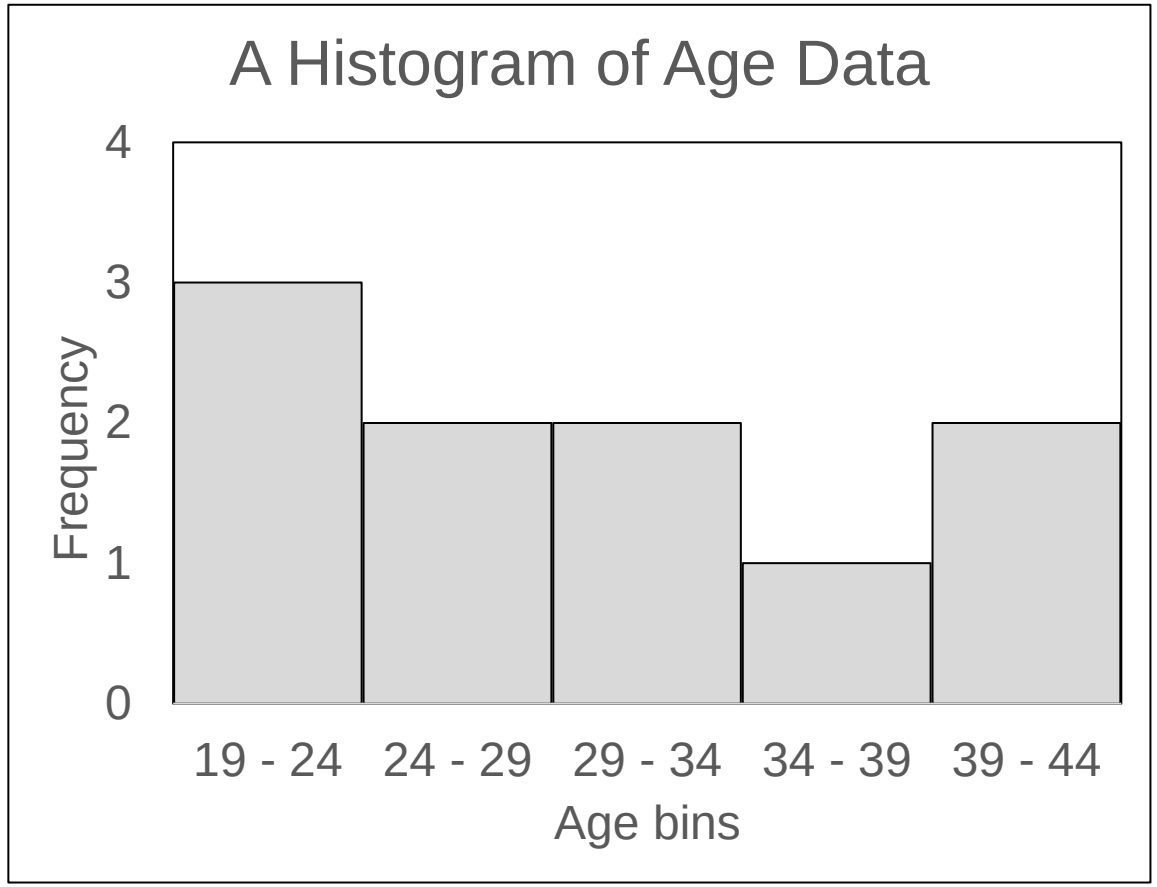
Therefore, Bin Width = $\frac{42-19}{5}$ = 4.6 ≈ 5

We set the lower bin limit of the first bin to be the minimum value in the data. Then, we increment by 5 to get the exclusive upper limit. This upper limit is the inclusive lower limit of the next bin. We keep incrementing by 5 until the entire data range is covered.

Histogram

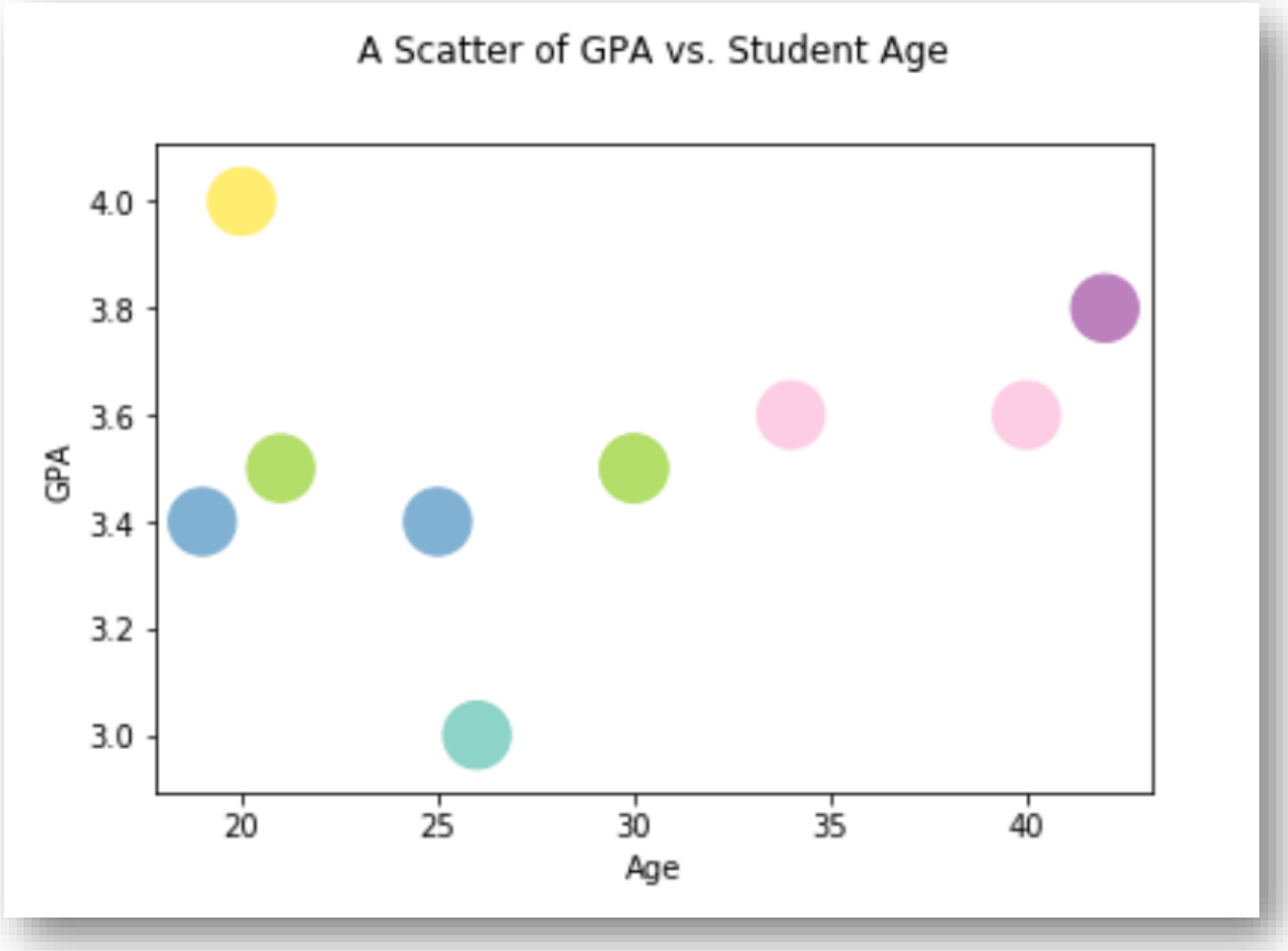
A histogram is a graphical display of frequencies of numerical data.

Bins	Frequency
19 - 24	3
24 - 29	2
29 - 34	2
34 - 39	1
39 - 44	2

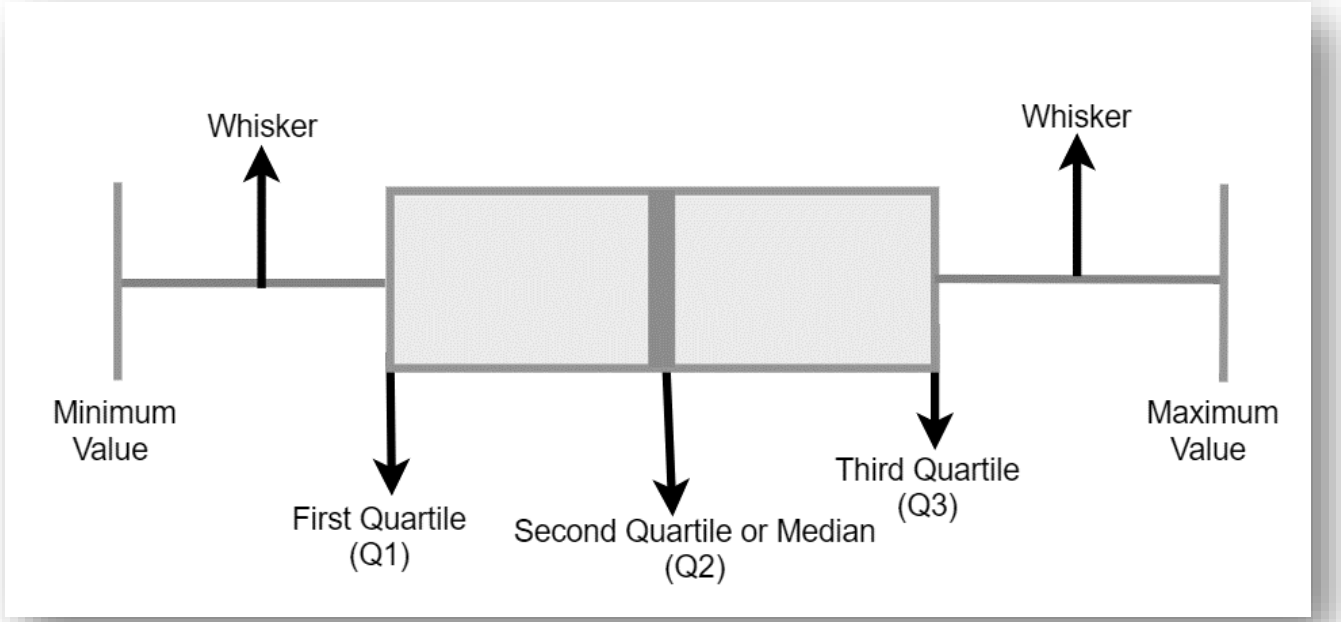


Scatter Plot

Index	Student	Age	GPA
1	Sophie	26	3.0
2	Abigail	19	3.4
3	Daniel	25	3.4
4	Ryan	30	3.5
5	Jamie	21	3.5
6	James	30	3.5
7	Natalie	34	3.6
8	Daniella	40	3.6
9	Jack	42	3.8
10	Peter	20	4.0



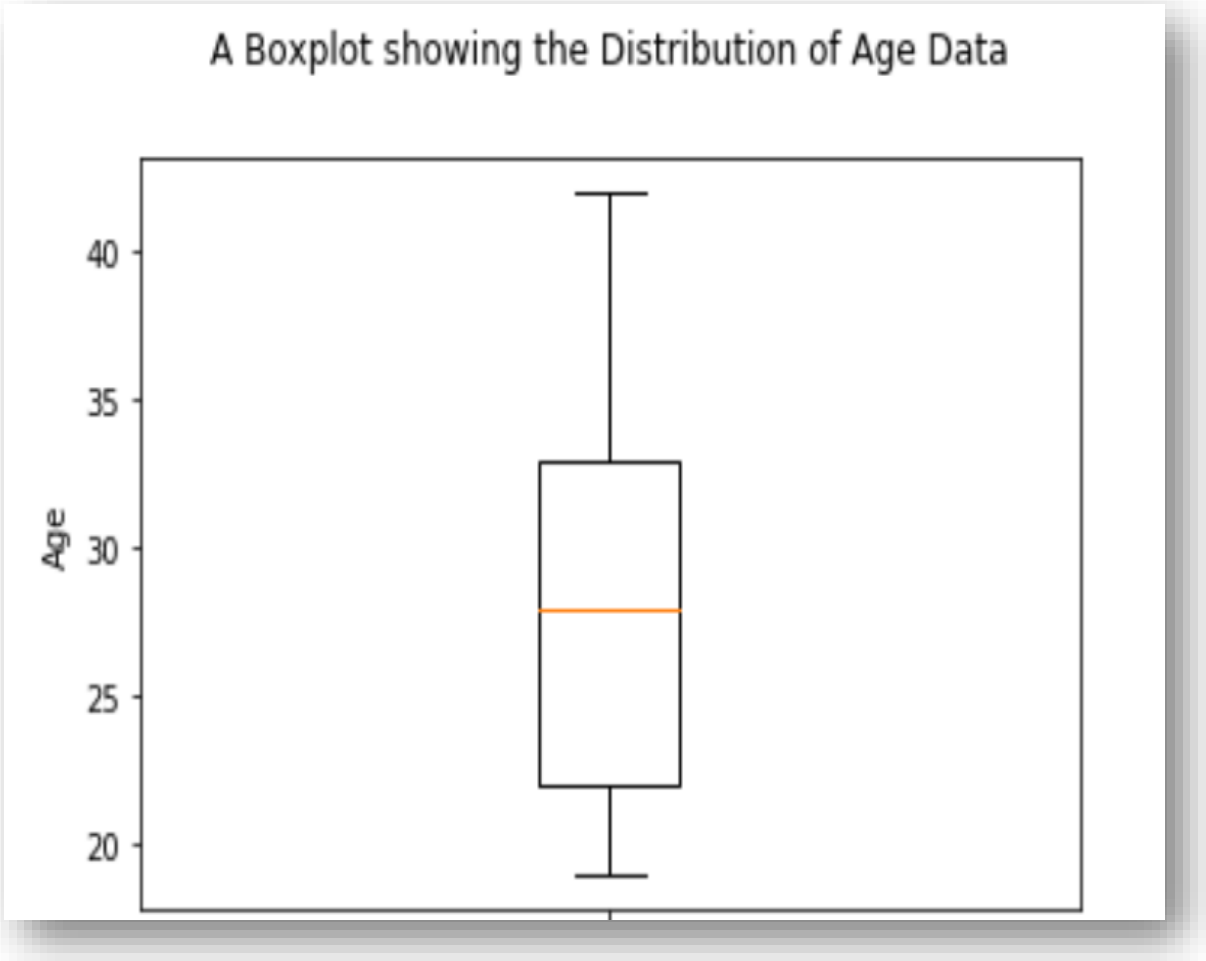
A Box Plot



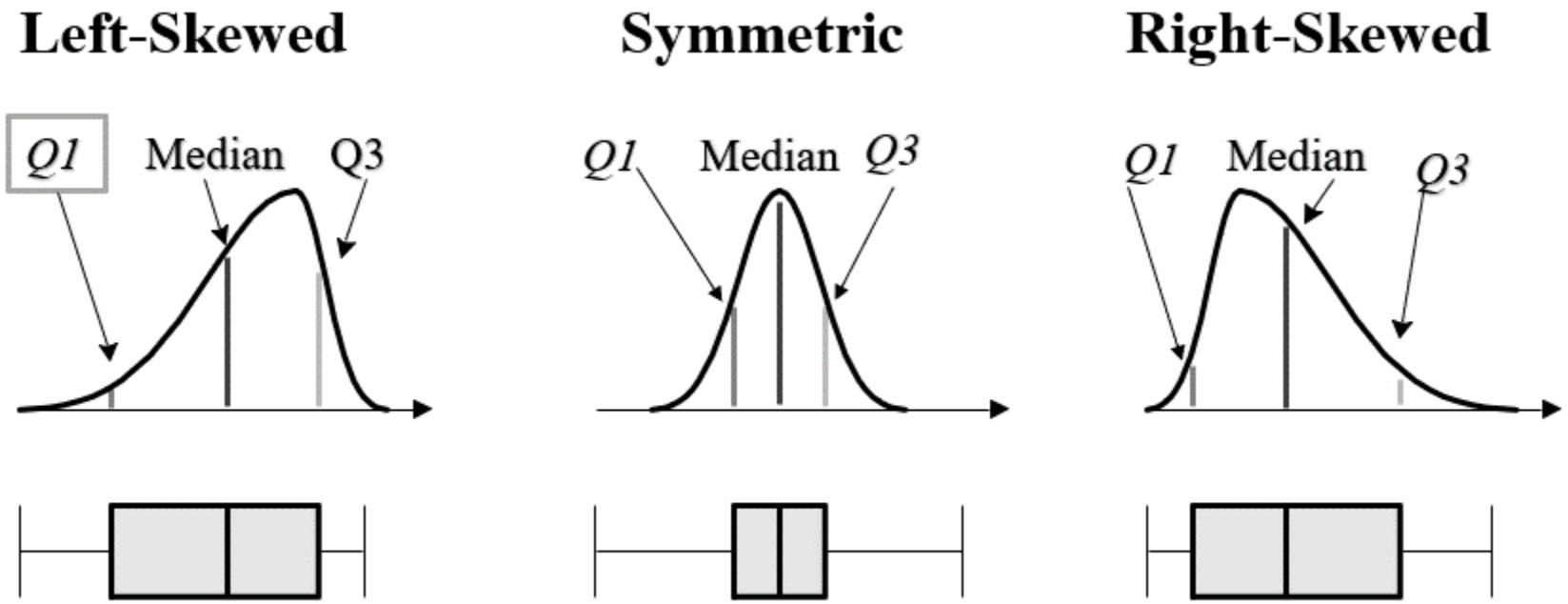
- A box plot is a five-summary plot used to display the distribution of data.
- The boxplot contains the minimum value, maximum value, first quartile, second quartile and third quartile.
- First quartile (25^{th} percentile) is the data point below which 25% percent of the sorted data lies.
- Second quartile (50^{th} percentile) is the median. 50% percent of the sorted data lies below this data point.
- The third quartile (75^{th} percentile) is the point below which 75 percent of the sorted data lies.
- Minimum value is the lowest data point and maximum value is the highest data point.

A Box Plot Showing Age Distribution

Index	Student	Age
1	Sophie	26
2	Abigail	19
3	Daniel	25
4	Ryan	30
5	Jamie	21
6	James	30
7	Natalie	34
8	Daniella	40
9	Jack	42
10	Peter	20



Box Plot and Shape of Distribution



Contingency Table

- A contingency table is also called cross tabulation.
- This is a two-way table used to display the frequencies of two variables (typically categorical variables).
- The levels of one variable are the columns and the levels of the other variable are the rows.

		Ethnicity		
Gender		Native American	Black	White
	Male	10	5	20
	Female	30	20	10