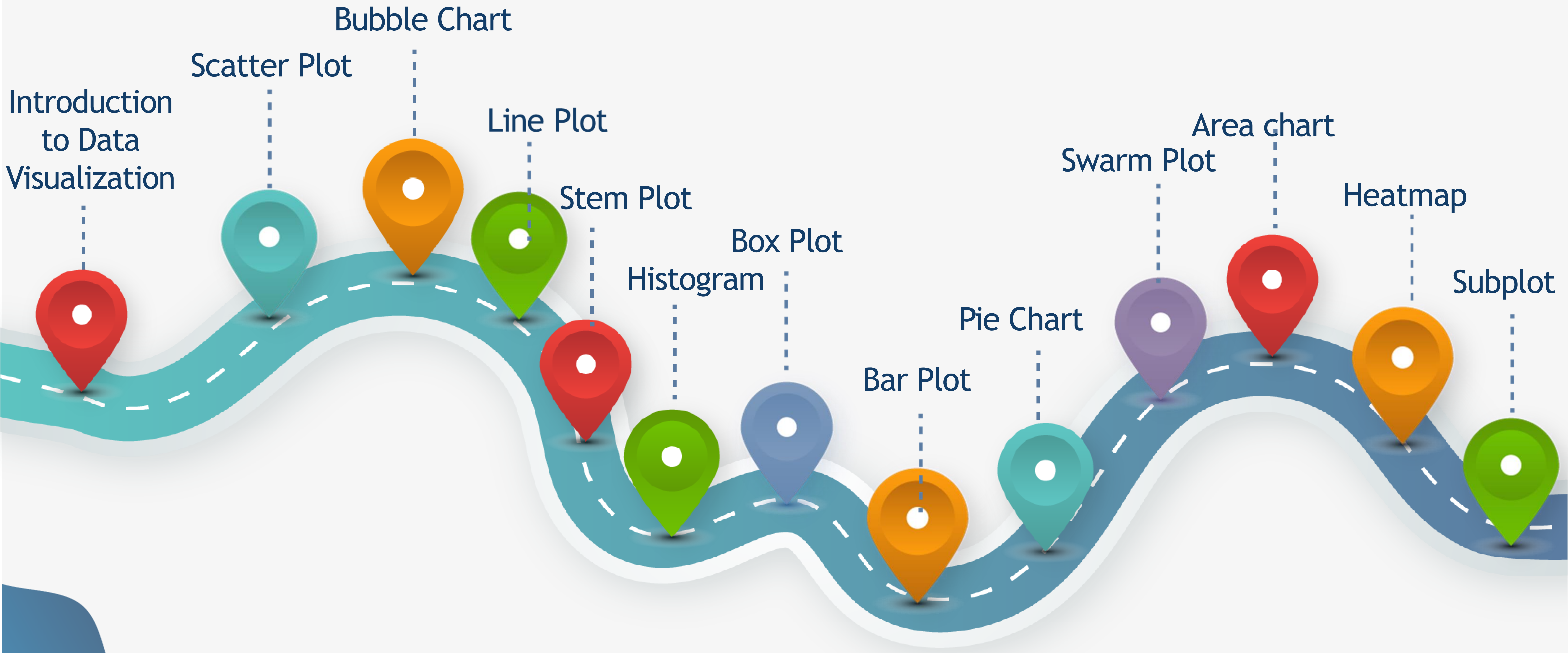
The Learning Hub

Data Science Advanced

Data Visualization

Education and Training Solutions

2023



Introduction to Data Visualization

Scatter Plot

Bubble Chart

Line Plot

Stem Plot

Histogram

Box Plot

Bar Plot

Pie Chart

Swarm Plot

Area chart

Heatmap

Subplot

Introduction to Data Visualization

Data visualization

- **Data visualization** can be defined as the graphical representation of information and data.
  - Using visual elements such as graphs, charts, maps, and data visualization tools are very important to provide an attainable way to imagine and comprehend trends, outliers, and patterns in data.
  - In the world of big data, data visualization tools and techniques are essential for analyzing vast amounts of information and making data-driven decisions.

Data visualization

❖ **Data Visualization** Sits Right in the Middle of Analysis and Visual Storytelling.

❖ **The most popular plotting libraries:**

➢ **Matplotlib**: Low level, provides a lot of freedom.

➢ **Pandas visualization**: open-source, easy-to-use interface, built on Matplotlib, high-performance.

➢ **Seaborn**: High-level interface, cool virtual styles.

Main Concepts of Data Visualization

❖A Good Graphic Tells a Story:

When Making a Graphic, it is Important to Understand What the Graphic is for.

❖Everything Should be Made as Simple as Possible, but not Simpler:

A Reasonable Representation Should be Followed When Designing a Chart in Terms of Colour, Font Size, and Explanation on the Chart/Graph Itself.

Main Concepts of Data Visualization

❖**Use the right tool for the job.**

Choose the Correct Type of Representation According to the Number of Variables and their Type (Continuous, Discrete/Categorical).

❖**Ink is cheap. Electrons are even cheaper.**

Split Charts and Graphs into More Graphs in Multivariable Visualization Tasks.

How to Tell Story ?

❖Two Main Concepts to Think About When Designing a Graph:

1 -Expressiveness:

A set of Facts is Expressible in a Visual Language if the Sentences(i.e., the Visualizations) in the Language Express all the Facts in the Set of Data, and Only the Facts in the Data. (Tell the Truth and Nothing but the Truth).

How to Tell Story ?

❖Two Main Concepts to Think About When Designing a Graph:

2-Effectiveness:

A Visualization is More Effective than Another Visualization if the Information Conveyed by one Visualization is More Readily Perceived than the Information in the other Visualization. (Use Encodings that People Decode Better).

How to Choose the Suitable Graph

- Determine the aim of presenting the data.
- Find out which features could help you to achieve your goal.
- Collect your data.
- Select the suitable type of graph or chart.

How to Choose the Suitable Graph

- **Determine the aim of presenting the data.**
- A graph can help you in case you need to compare different features, understand how could they affect the whole, and analyze patterns.
- graphs can be used to recognize outliers and help you see correlations between groups.
- Clarify your aims, to guide your chart selection.

How to Choose the Suitable Graph

❖ **Select the suitable type of graph or chart.**

- Choosing the incorrect visual aid or defaulting to the most familiar type of data visualization could cause confusion guide to incorrect data analysis. But a graph is only useful to you and your business if it communicates your point easily and effectively.
- To help find the right map or graph type, ask yourself the questions below.

How to Choose the Suitable Graph

- To help find the right map or graph type, ask yourself the questions below.

1. **Do you aim to compare values?**
2. **Do you aim to show the composition of something?**
3. **Do you aim to understand the distribution of your data?**
4. **Are you interested in studying patterns in your data set?**
5. **Do you aim to better understand the relationship between value sets?**

How to Choose the Suitable Graph

1. Do you aim to compare values?

Charts and graphs are useful for comparing many value sets and can easily show the min and max values in the data sets. To create a comparison chart, the following types of graphs can be used:

- **Column**
- **Bar**
- **Pie**
- **Line**
- **Scatter Plot**

How to Choose the Suitable Graph

2. Do you aim to show the composition of something?

This type of chart shows how singular parts make up the whole of something.

To show composition, use the following plotting types:

- **Pie**
- **Stacked Bar**
- **Area**

How to Choose the Suitable Graph

3. Do you aim to understand the distribution of your data?

Distribution charts are perfect for understanding the normal tendency and range of information and detecting outliers in the dataset.

Use these charts to show distribution:

- **Scatter Plot**
- **Line**
- **Column**
- **Bar**
- **Histogram**

How to Choose the Suitable Graph

4. Are you interested in studying patterns in your dataset?

If you want to analyse the trends and know more information about how some variables performed during a specific period, you should choose a:

- **Line**
- **Dual-Axis Line**
- **Column**

How to Choose the Suitable Graph

5. **Do you aim to better understand the relationship between value sets?**

   If you want to know how one variable relates to one or many different variables. You could use this to show how something positively affects, has no effect, or negatively affects another variable.

○ Use these charts to show the relationship:

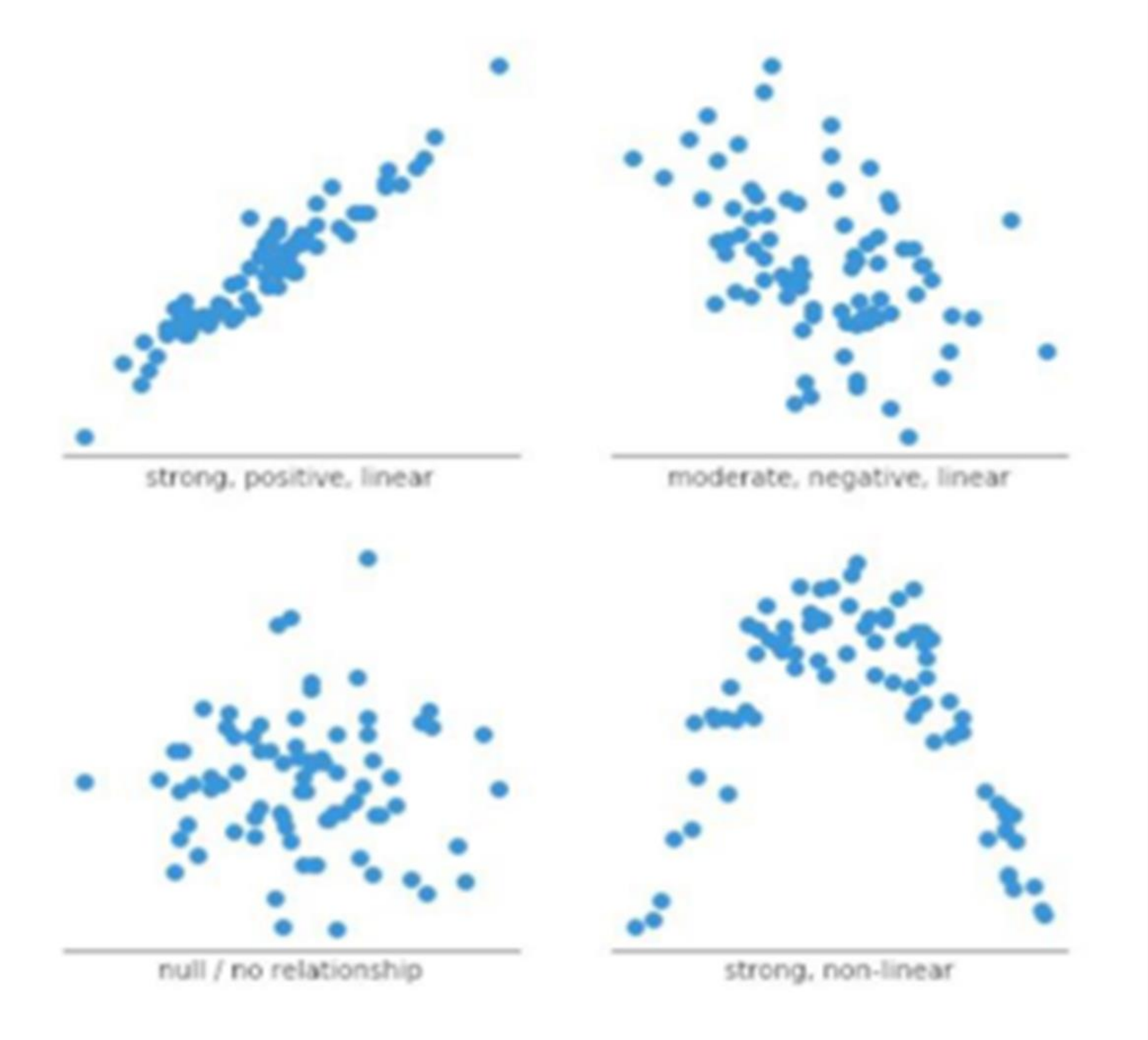
- **Scatter Plot**
- **Bubble**
- **Line**
- **Heatmap**

Scatter Plot

Scatter Plot

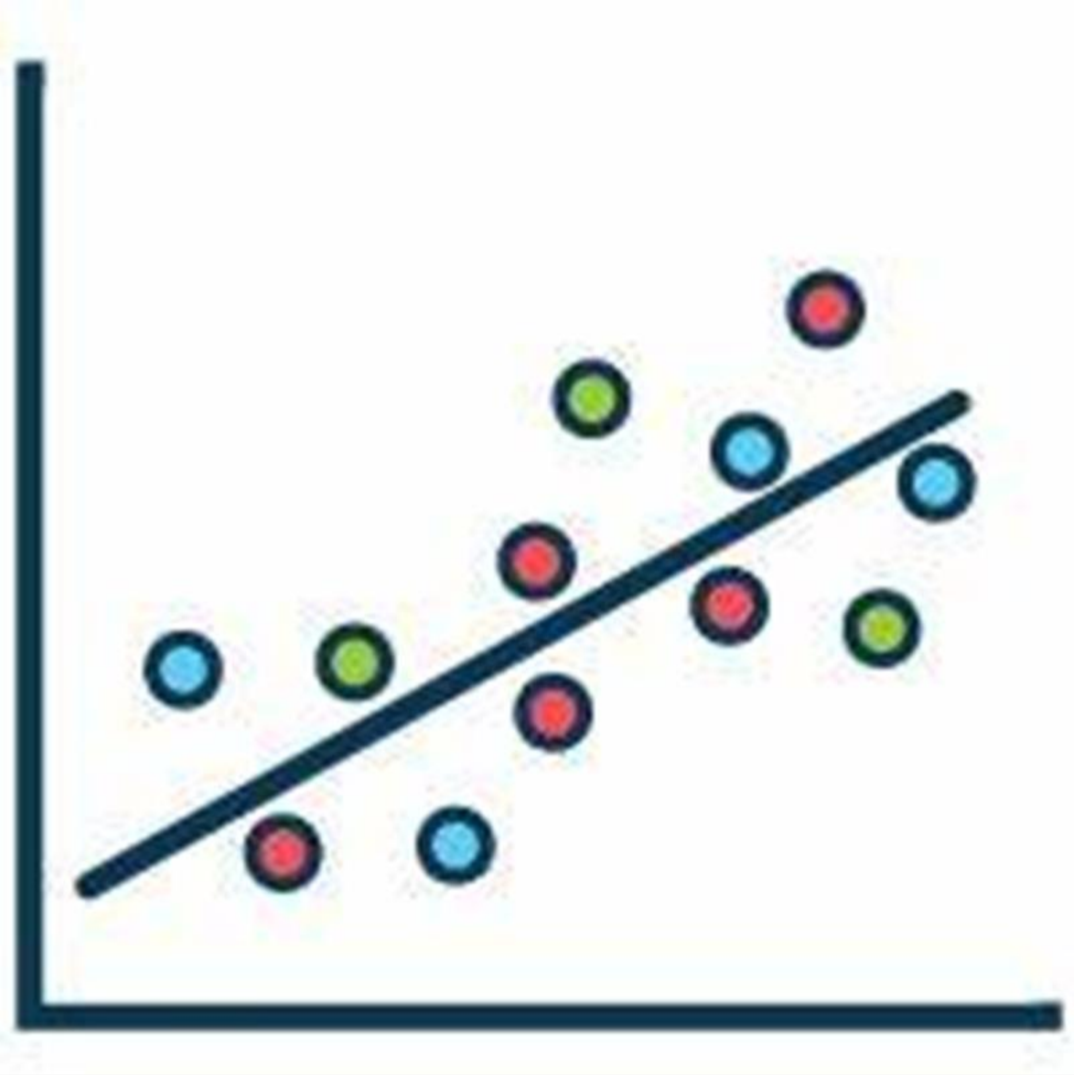
- A Scatterplot (2D) is a plot that Uses Dots to Represent Values for Two Different Numeric Variables. it is useful for comparing variables against each other
- Scatter Plots are Used to Observe **Relationships** Between Variables
- Scatter Plots are Used to Identify the **Correlation** Relationships Between Variables

Scatter Plot



Scatter Plot

- The data in a scatter plot is considered to express a trend. With further analysis using tools like regression, we can mathematically calculate this relationship and use it to predict trends outside the dataset.

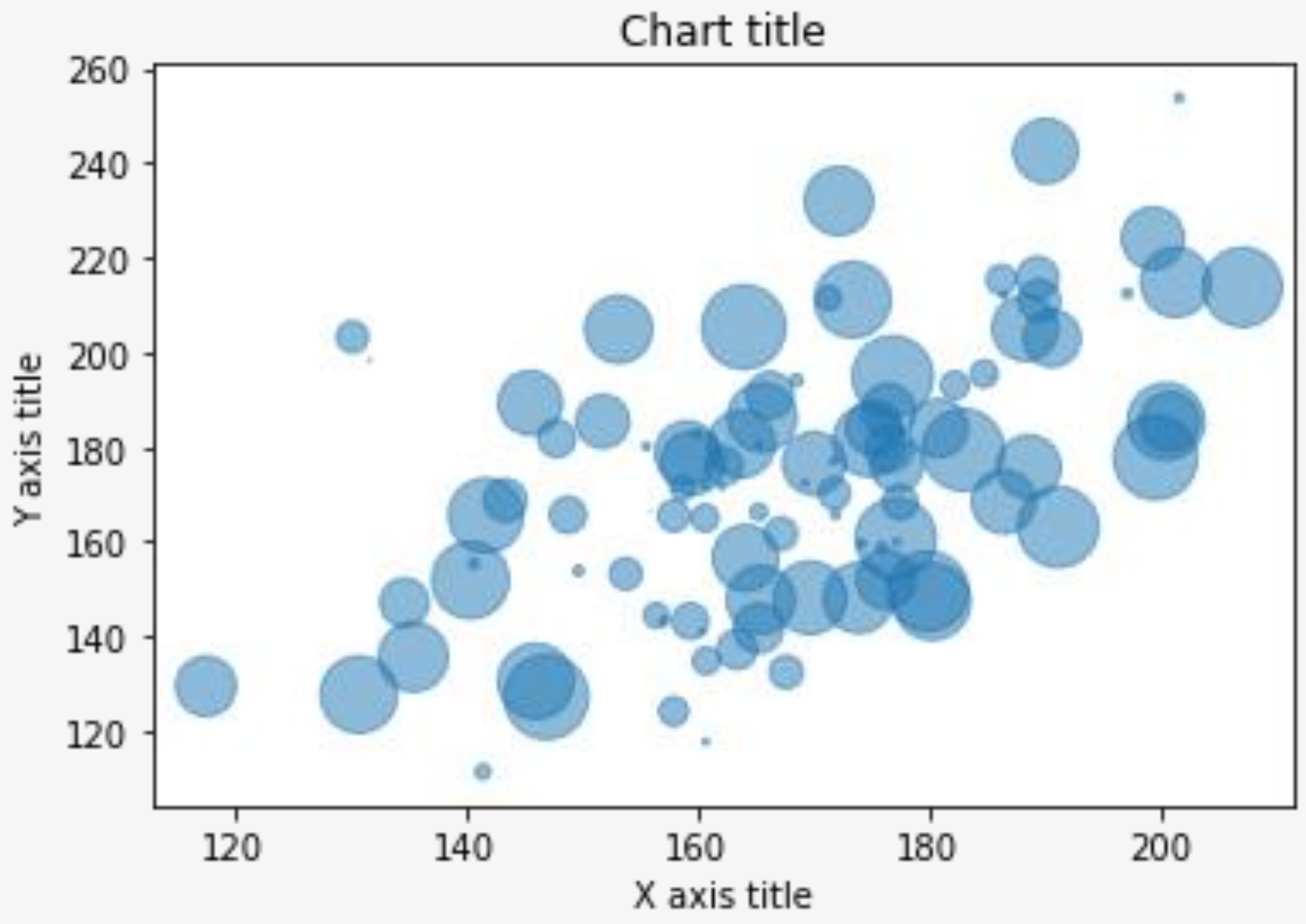


Bubble Chart

Bubble Chart

- ❖A bubble chart is similar to a scatter plot, it can show distribution or relationship.

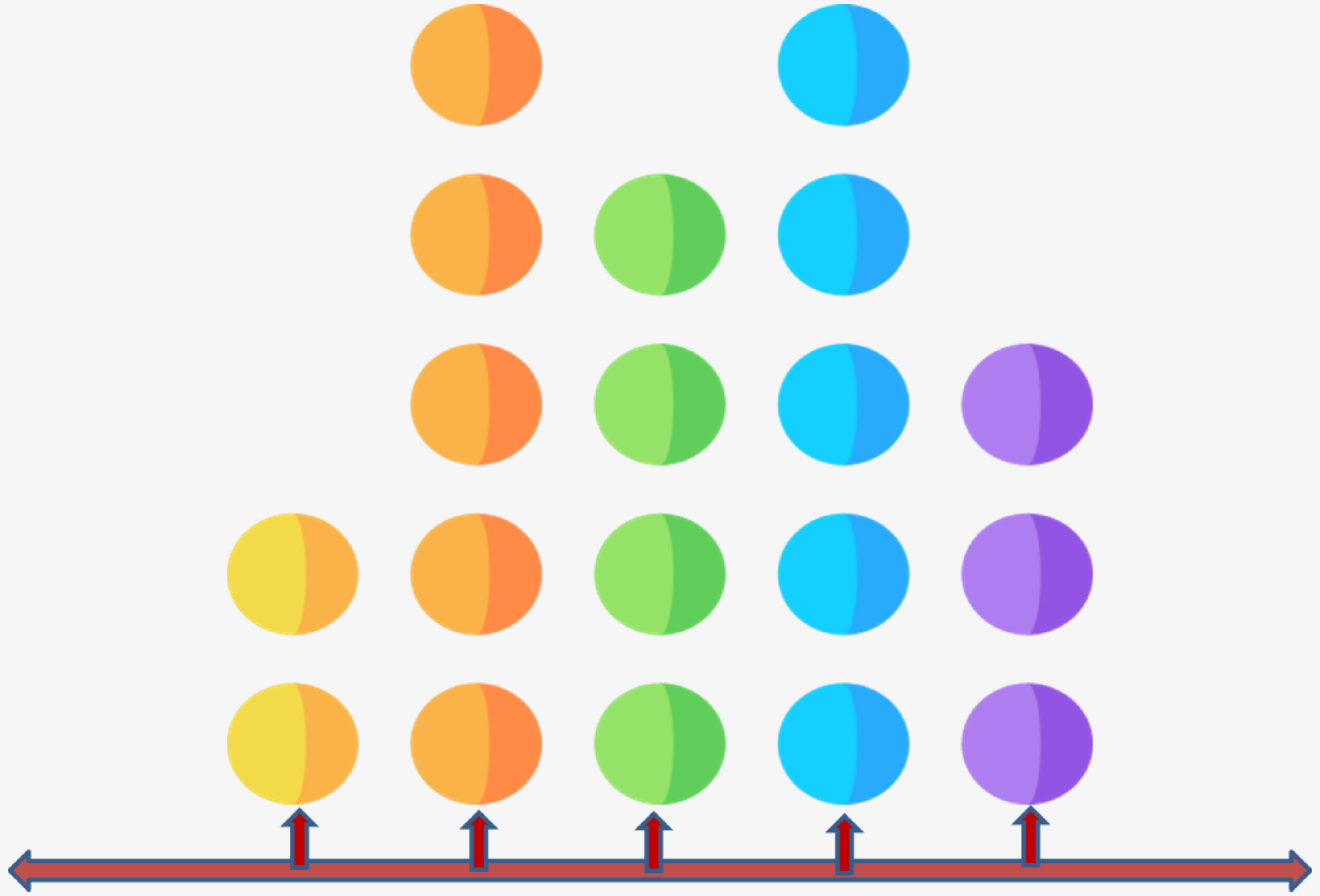
- ❖The difference is a third variable is shown by the size of the bubble or circle



Line Plot

Line Plot

❖ A **Line Plot** Can be Defined as a Graph that Displays Data as Points or Checks Marks Above a Number Line, Showing the Frequency of Each Value



Line Plot

❖The methods used in different libraries to create a line plot

❖Matplotlib ⟶ plot()

❖Pandas ⟶ [dataframe].plot.line().

❖Seaborn ⟶ sns.lineplot()

Line Plot

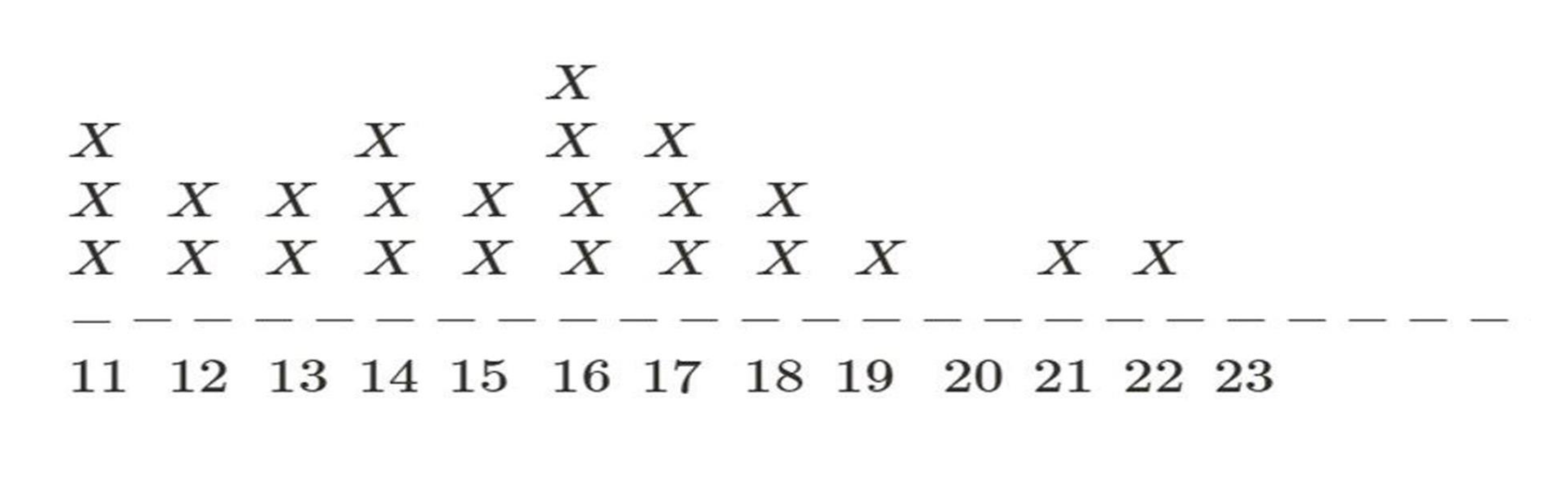
- ❖We can plot multiple columns in a single graph, by looping through the columns we want and plotting each column on the same axis.
- ❖in Matplotlib, looping over every column that wanted to be plotted is a necessary step.
- ❖Pandas automatically plots all available numeric columns (at least if we don't define a specific column/s).

Line Plot

❖Example

- The Following Numbers are the Result Of a Test Taken by a Class of 24 Students:

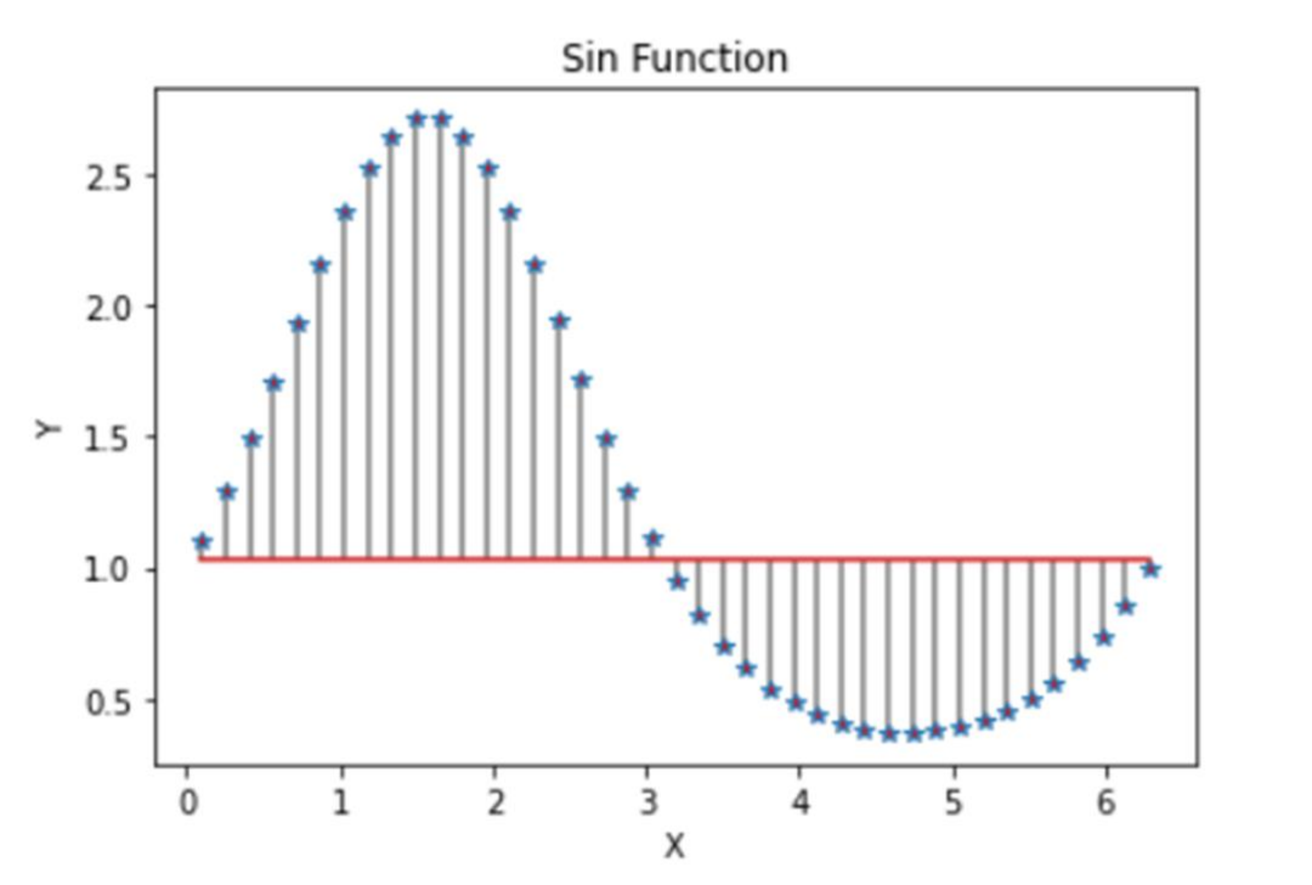
16,14,17,11,14,19,11,17,12,21,22,18,11,16,15,14,18,12,13,16,17,15,13,17



Stem Plot

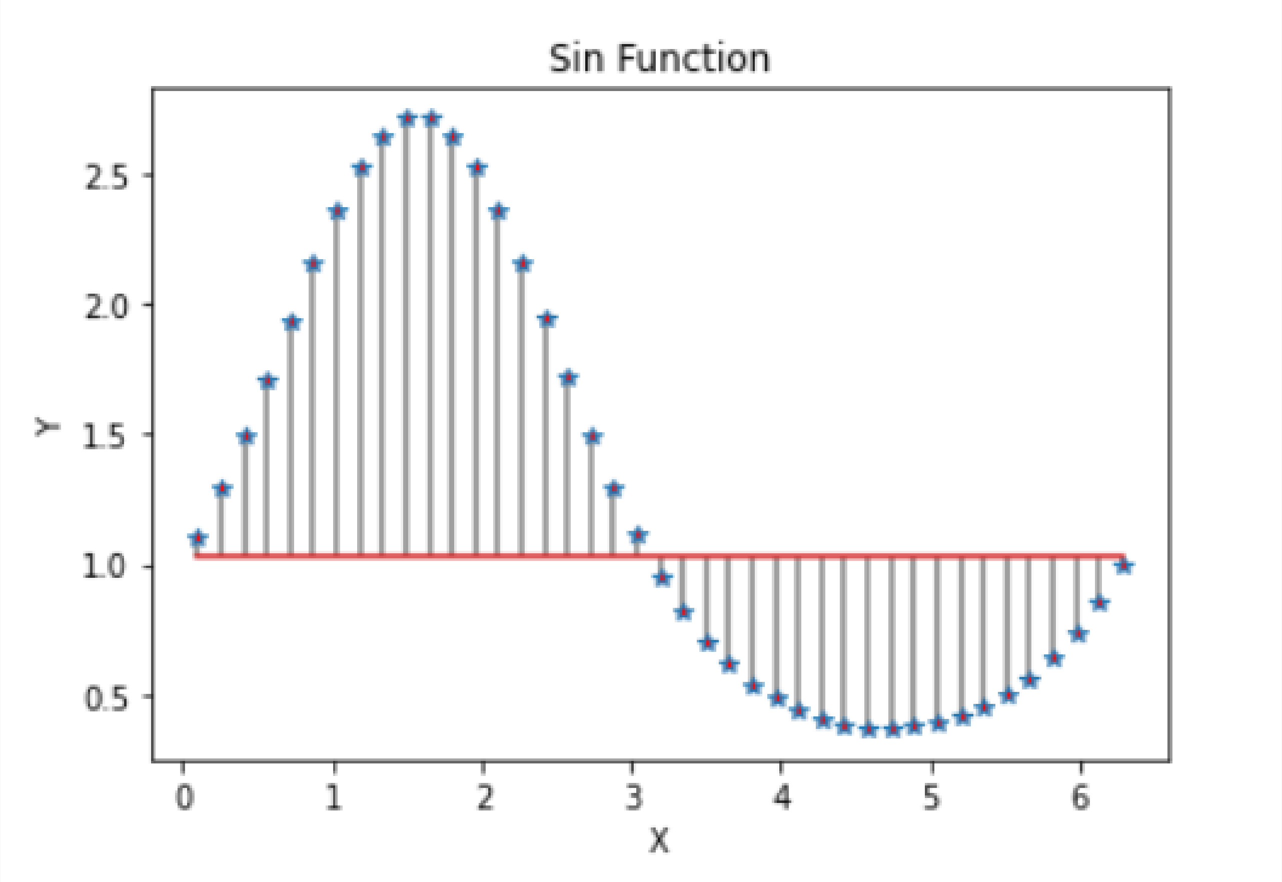
Stem Plot

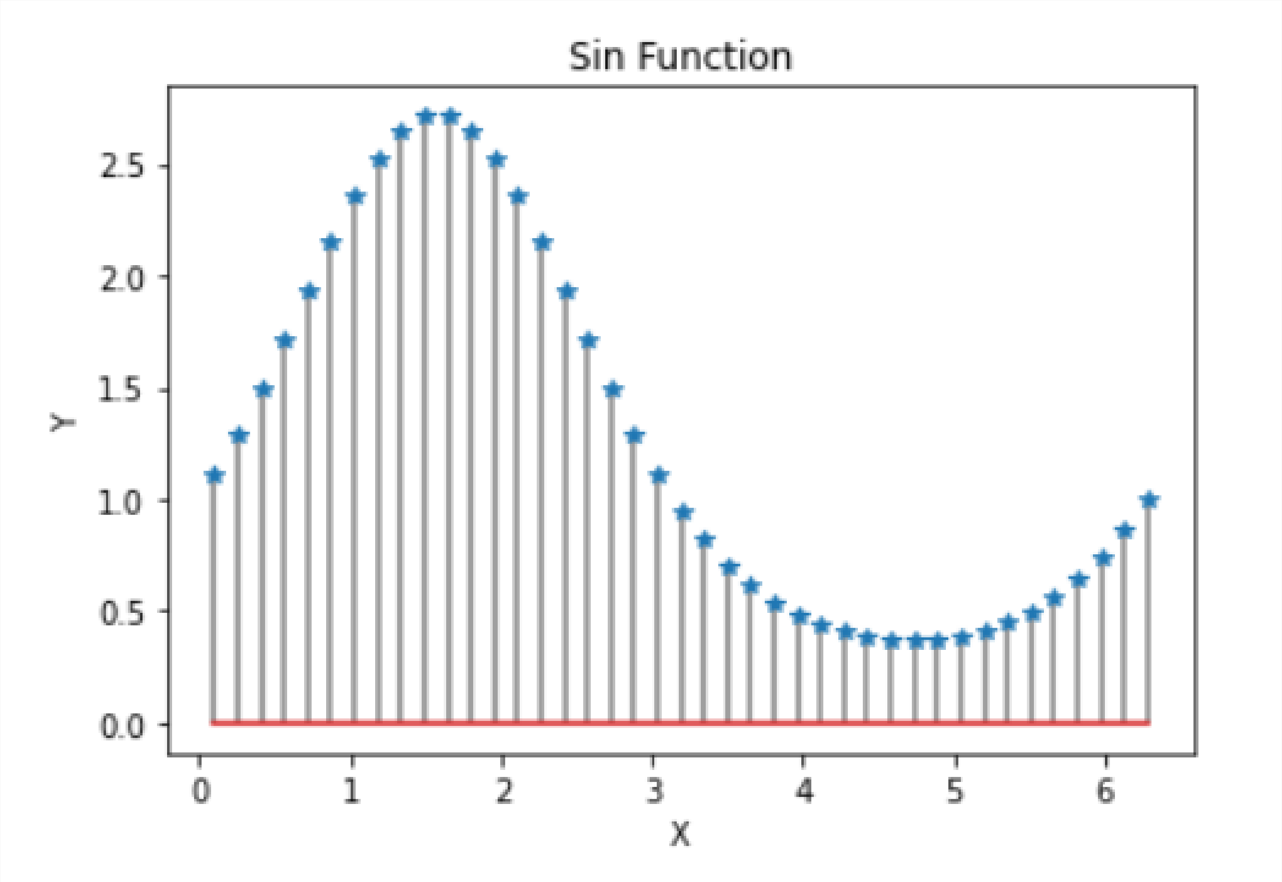
- It is a plot that represents vertical lines from a baseline to the Y-coordinate and places a marker at the tip.
- **stem()** method can be used to plot a stem plot in Matplotlib.



Stem Plot

- **Scatter plots** are important for visualizing data points in two dimensions, and great for displaying correlations and groupings in data .

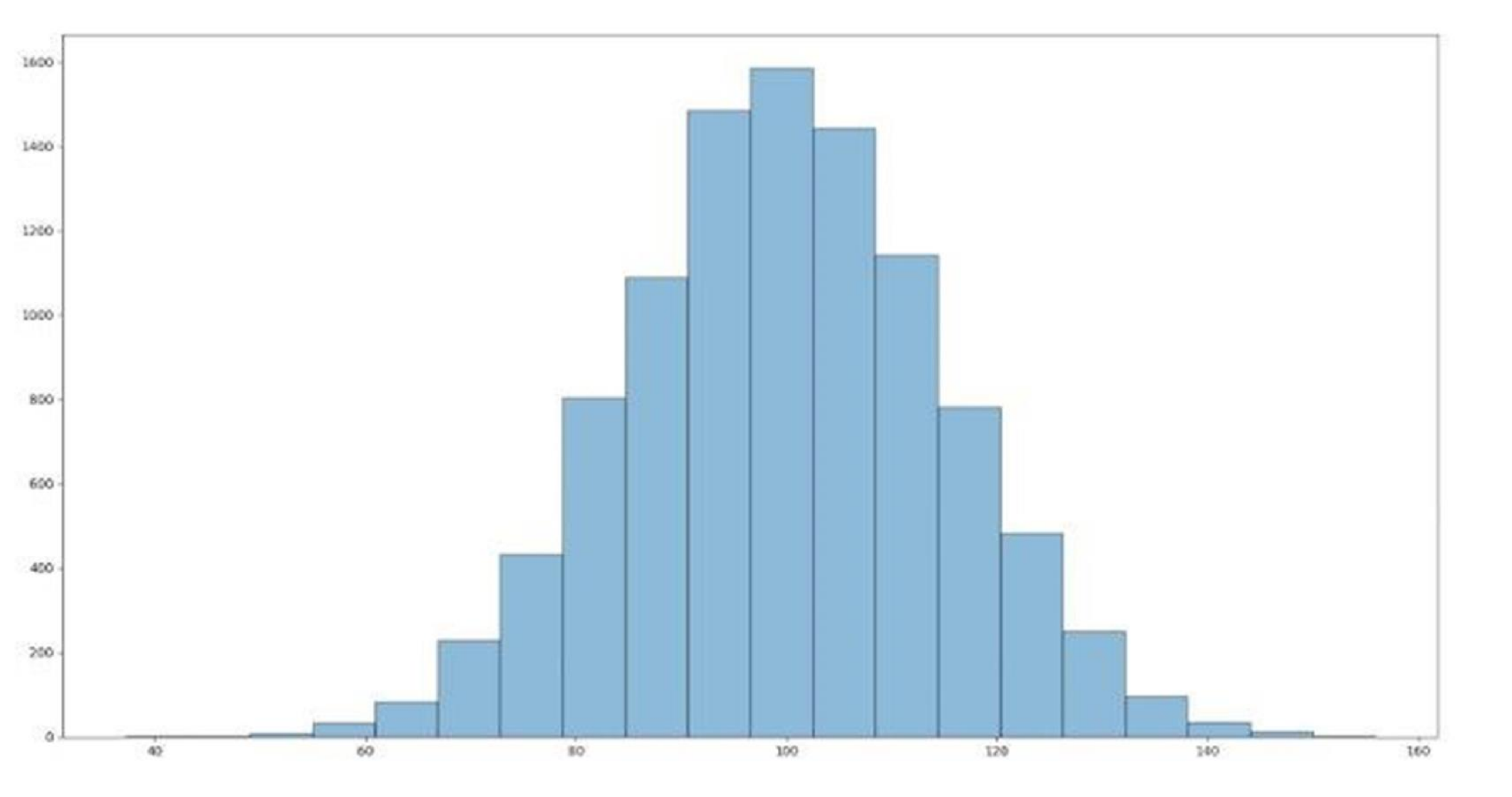




Histogram

Histogram

- A **Histogram** is a Graphical Representation of a Grouped Frequency Distribution with Continuous Classes



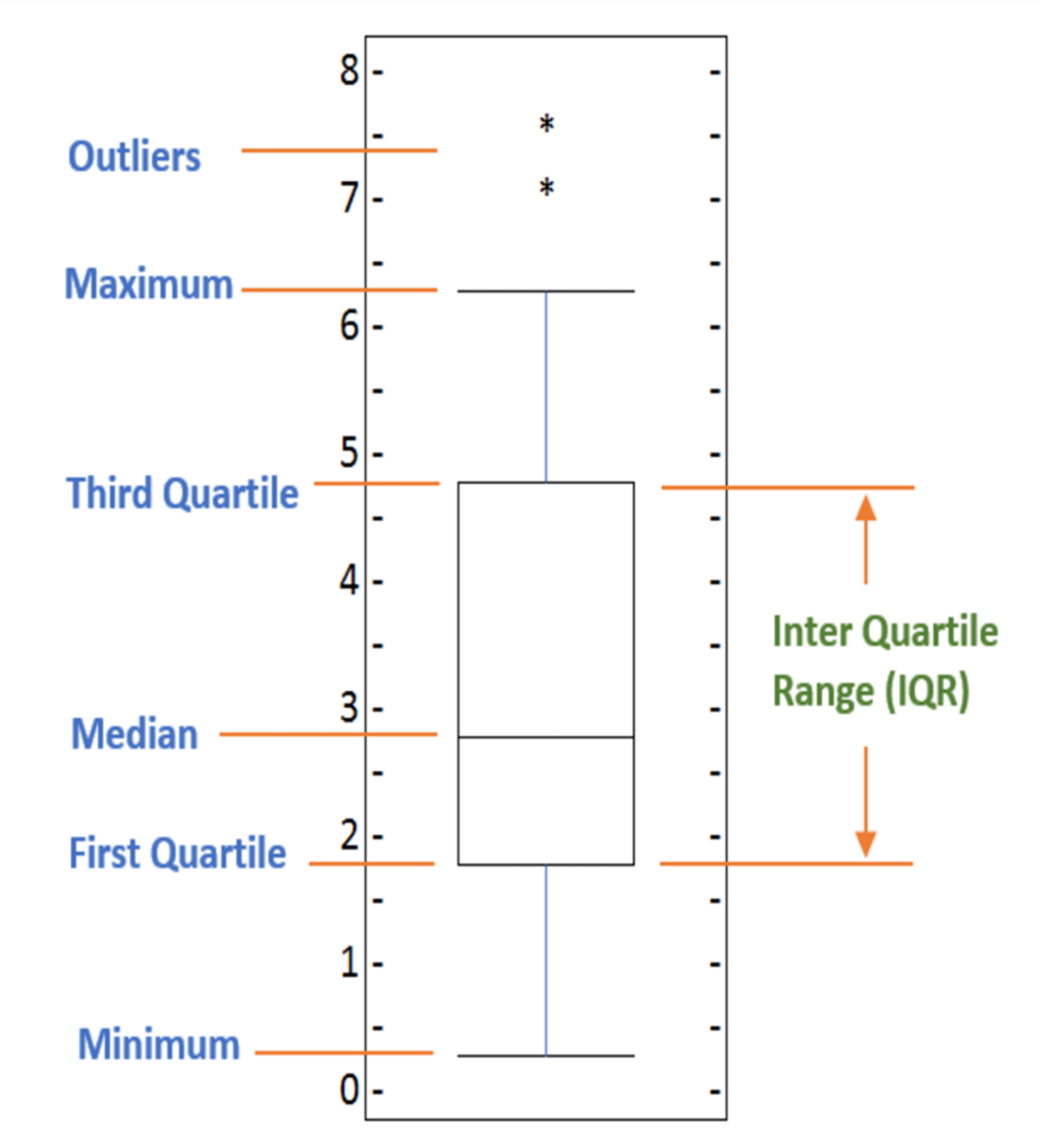
Box Plot

Box Plot

❖ A **box plot** is a way of statistically representing the distribution of the data through five main dimensions:

- **Minimun**: Smallest number in the dataset.
- **First quartile**: Middle number between the minimum and the median.
- **Second quartile (Median):** Middle number of the (sorted) dataset.
- **Third quartile**: Middle number between median and maximum.
- **Maximum**: Highest number in the dataset.

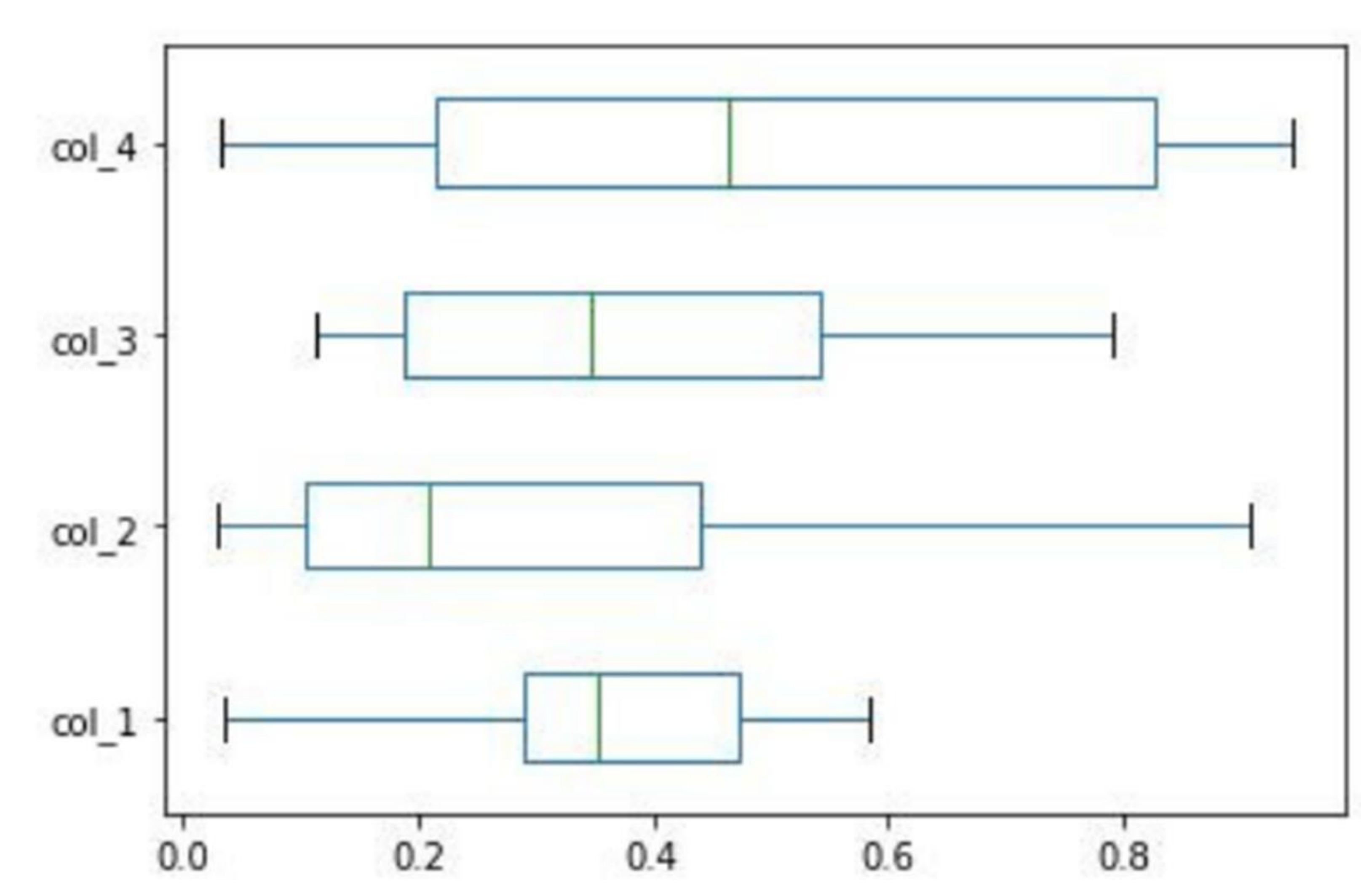
Box Plot


8
7
6
5
4
3
2
1
0
Outliers
Maximum
Third Quartile
Median
First Quartile
Minimum
Inter Quartile
Range (IQR)

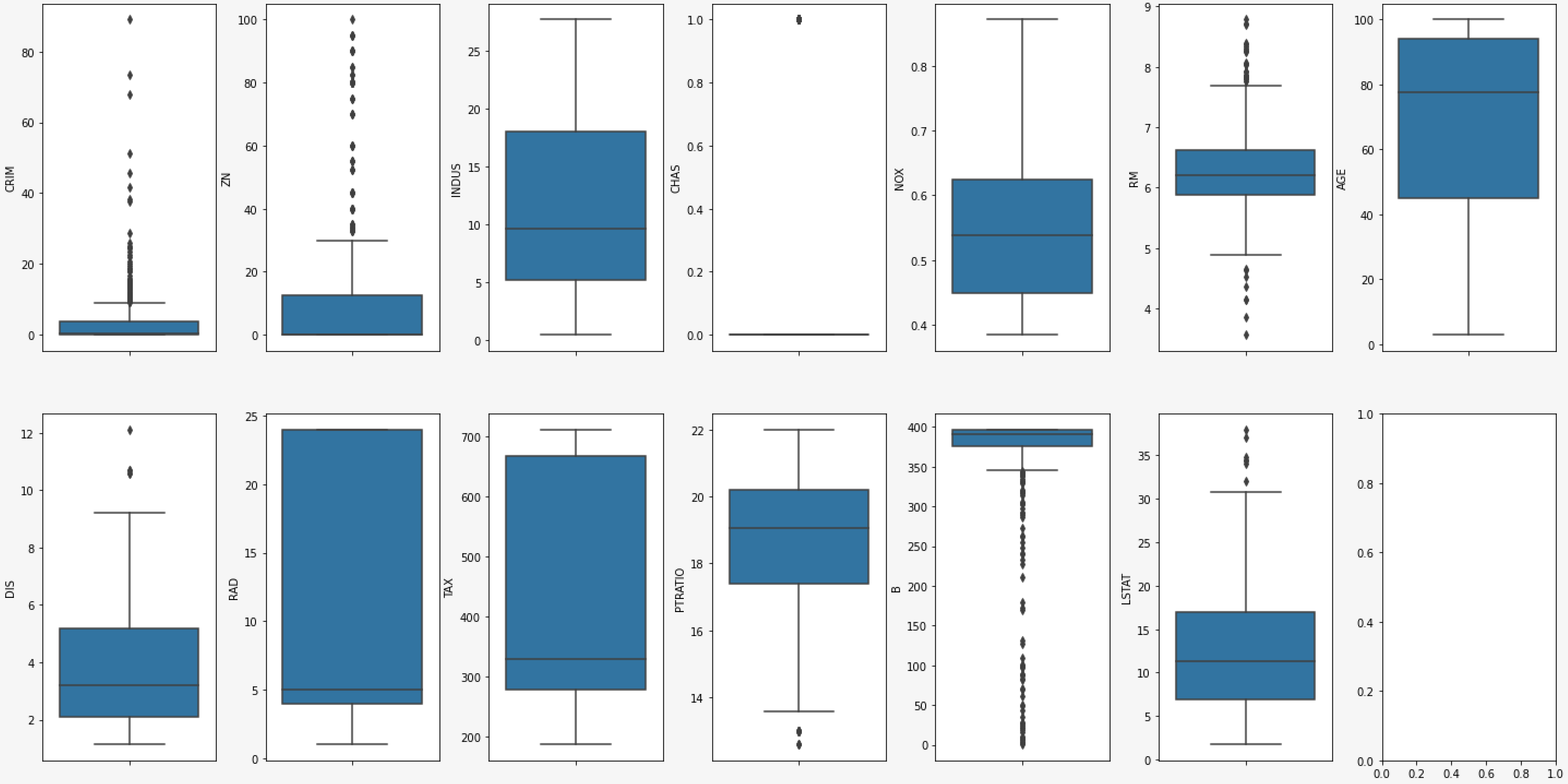
Box Plot

- **Generating the box plot in a horizontal form.**

"dataframe.plot.box(vert=False, positions=[1, 2, 3, 4]);"



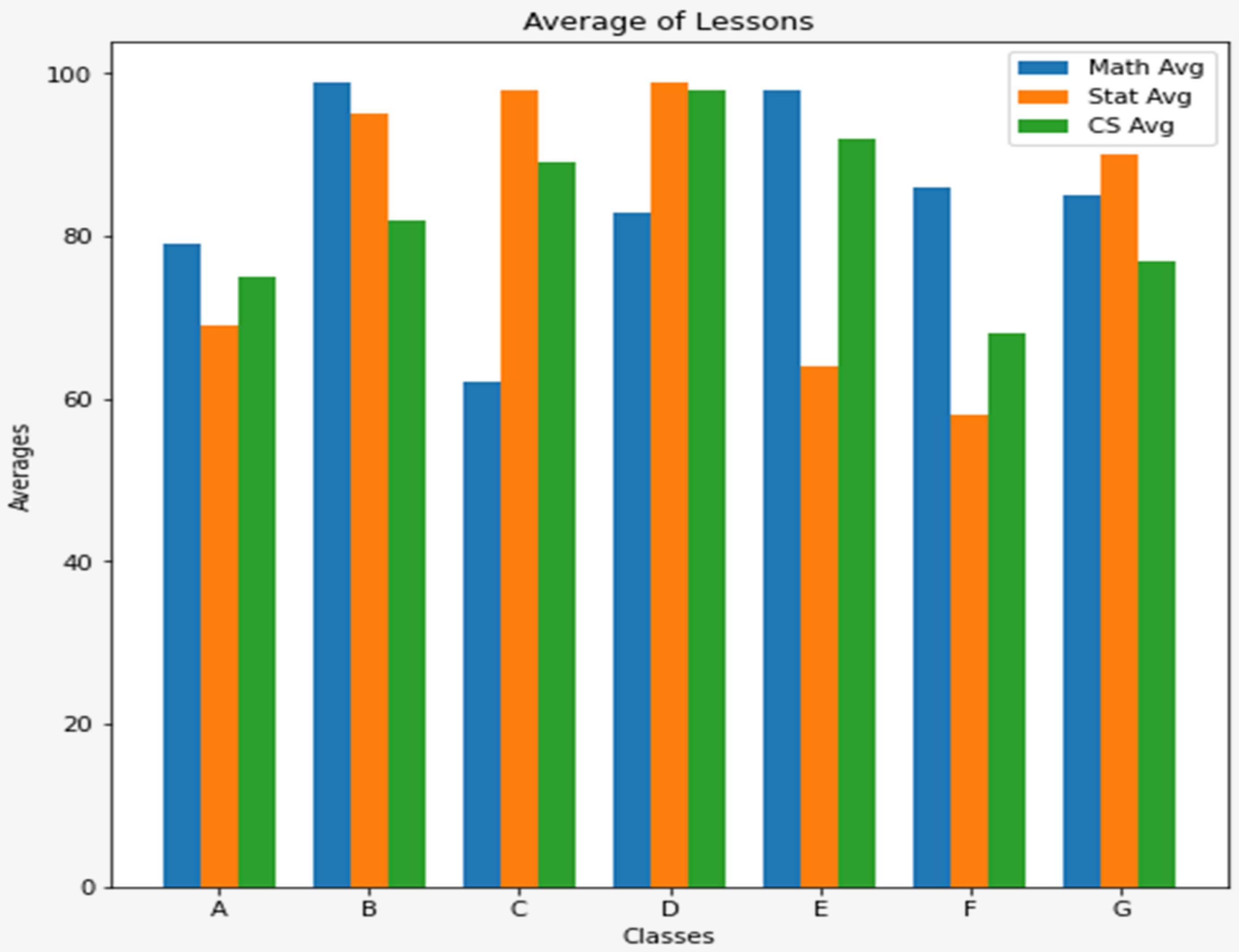
Box Plot


CRIM
0
20
40
60
80
ZN
0
20
40
60
80
100
INDUS
0
5
10
15
20
25
CHAS
0.0
0.2
0.4
0.6
0.8
1.0
NOX
0.4
0.5
0.6
0.7
0.8
RM
4
5
6
7
8
9
AGE
0
20
40
60
80
100
DIS
2
4
6
8
10
12
RAD
0
5
10
15
20
25
TAX
200
300
400
500
600
700
PTRATIO
14
16
18
20
22
B
0
50
100
150
200
250
300
350
400
LSTAT
0
5
10
15
20
25
30
35
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

Bar Plot

Bar Plot

- A **Bar Plot** is a Plot that Presents Categorical Data with Rectangular Bars with Lengths Proportional to the Values that They Represent



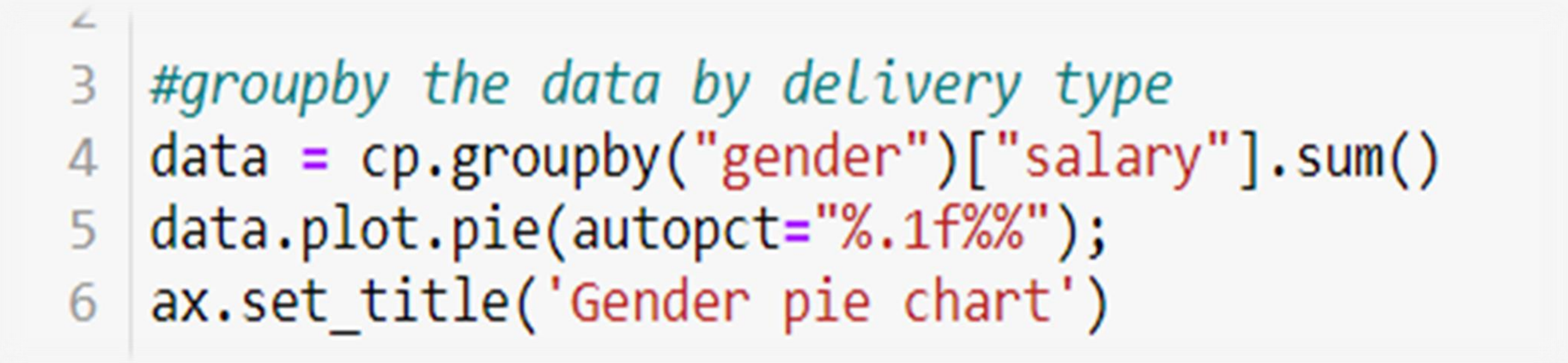
Pie Chart

Pie Chart

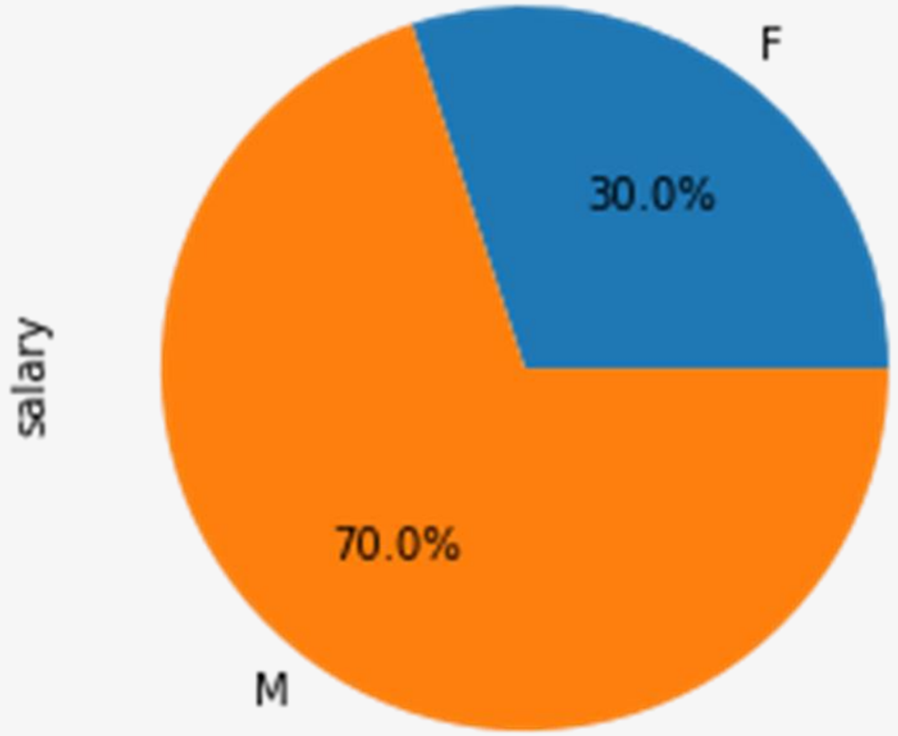
❖ A Pie Chart is a Circular Statistical Plot that Can Display Only One Series of Data.

Pie Chart

❖ Pie plot using pandas

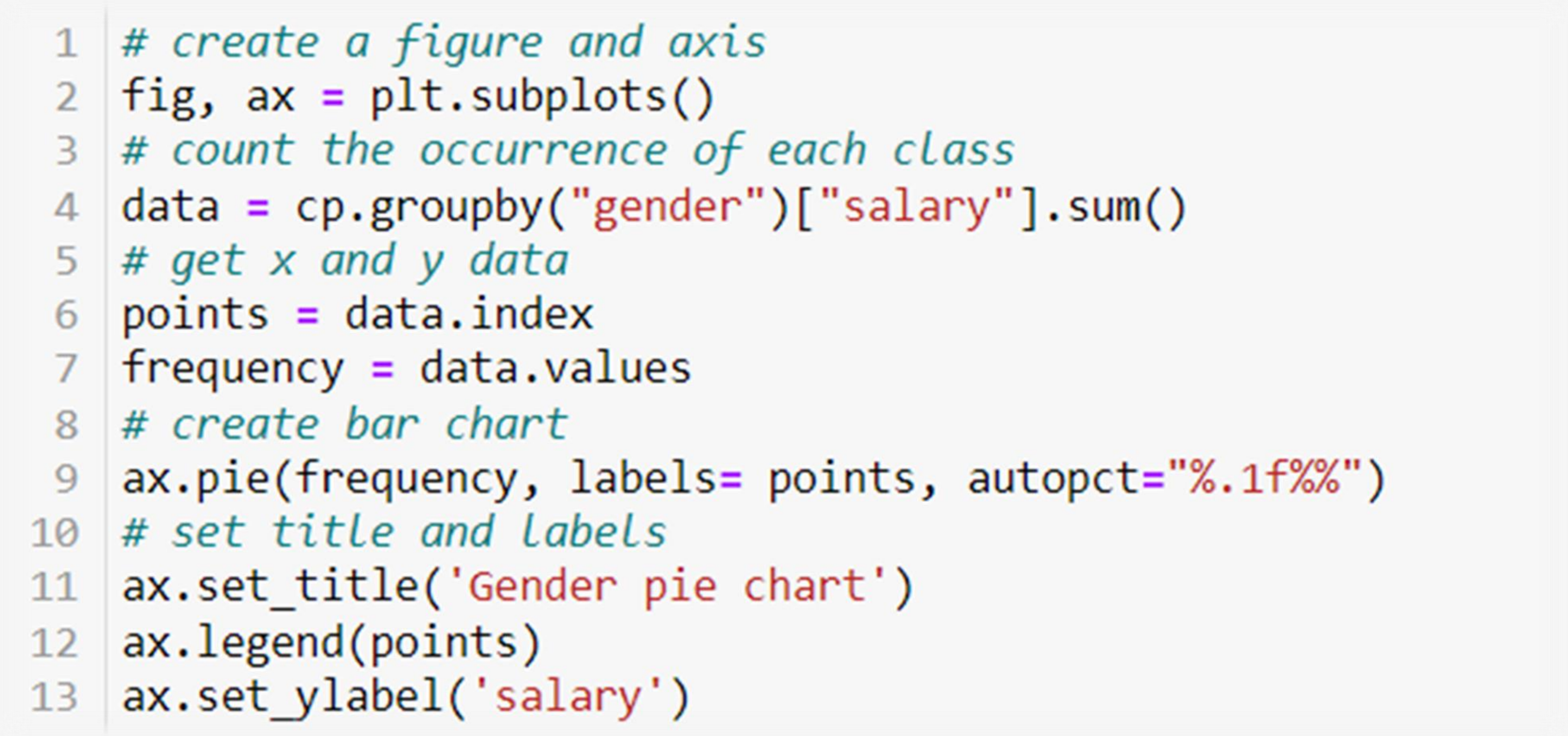


```
#groupby the data by delivery type
data = cp.groupby("gender")["salary"].sum()
data.plot.pie(autopct="%.1f%%");
ax.set_title('Gender pie chart')
```

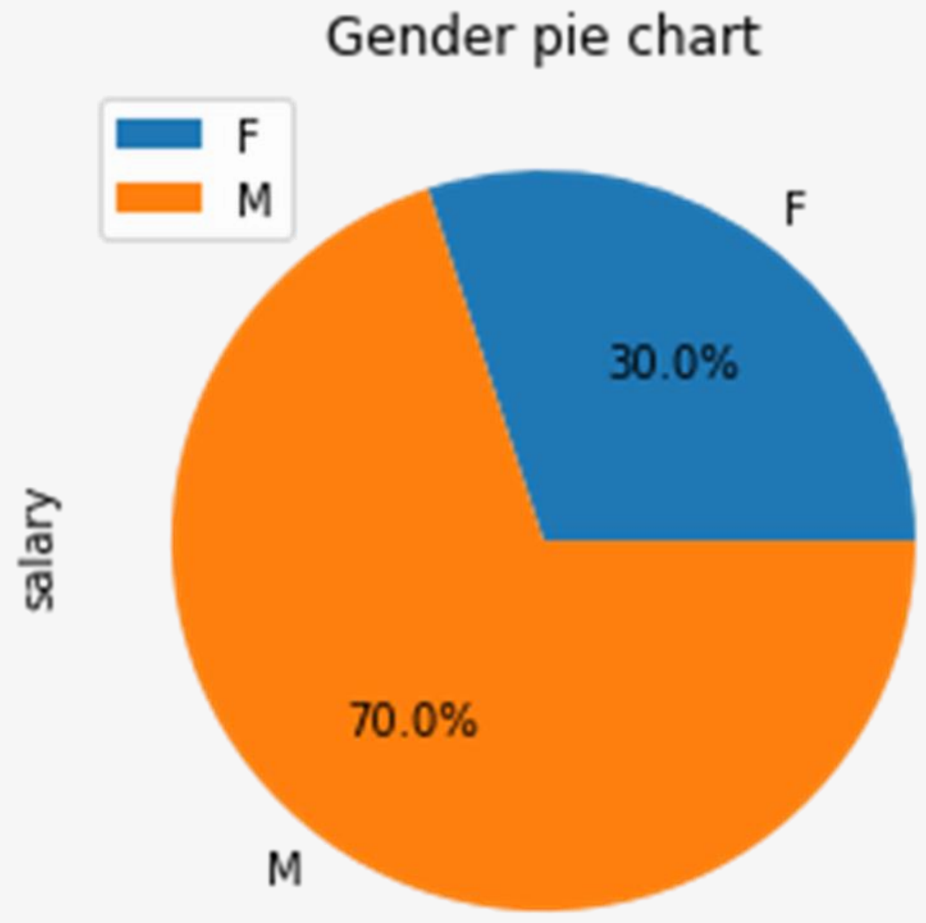


Pie Chart

❖ Pie plot using Matplotlib



```
# create a figure and axis
fig, ax = plt.subplots()
# count the occurrence of each class
data = cp.groupby("gender")["salary"].sum()
# get x and y data
points = data.index
frequency = data.values
# create bar chart
ax.pie(frequency, labels= points, autopct="%.1f%%")
# set title and labels
ax.set_title('Gender pie chart')
ax.legend(points)
ax.set_ylabel('salary')
```



The Learning Hub

Swarm Plot

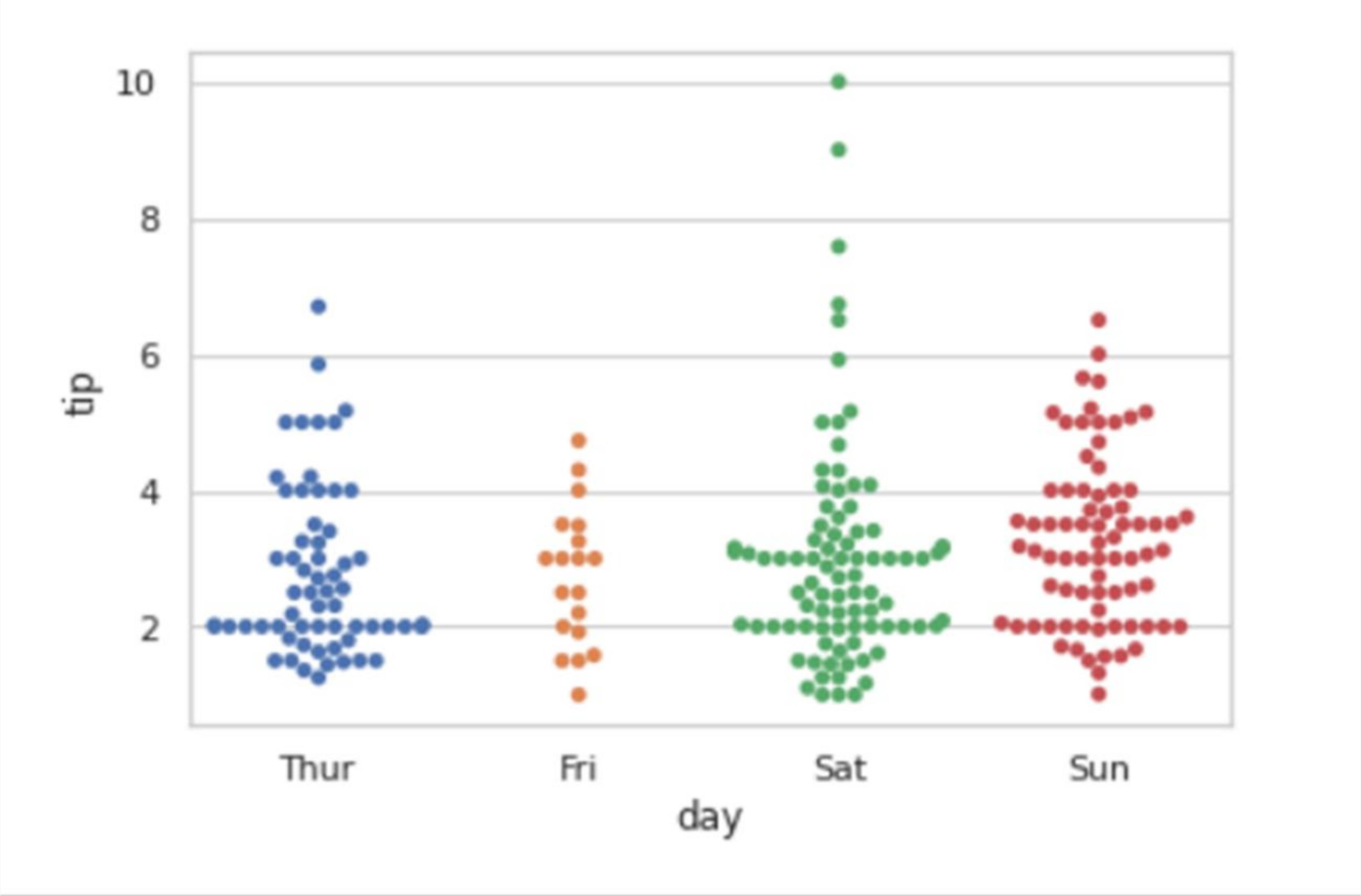
Swarm Plot

- A **Swarm Plot** is another method or technique to plot the distribution of a feature or the joint distribution of a couple of features.

- A **Swarm Plot** helps observe how the data is distributed across a categorical feature and how the continuous variable varies within a category.

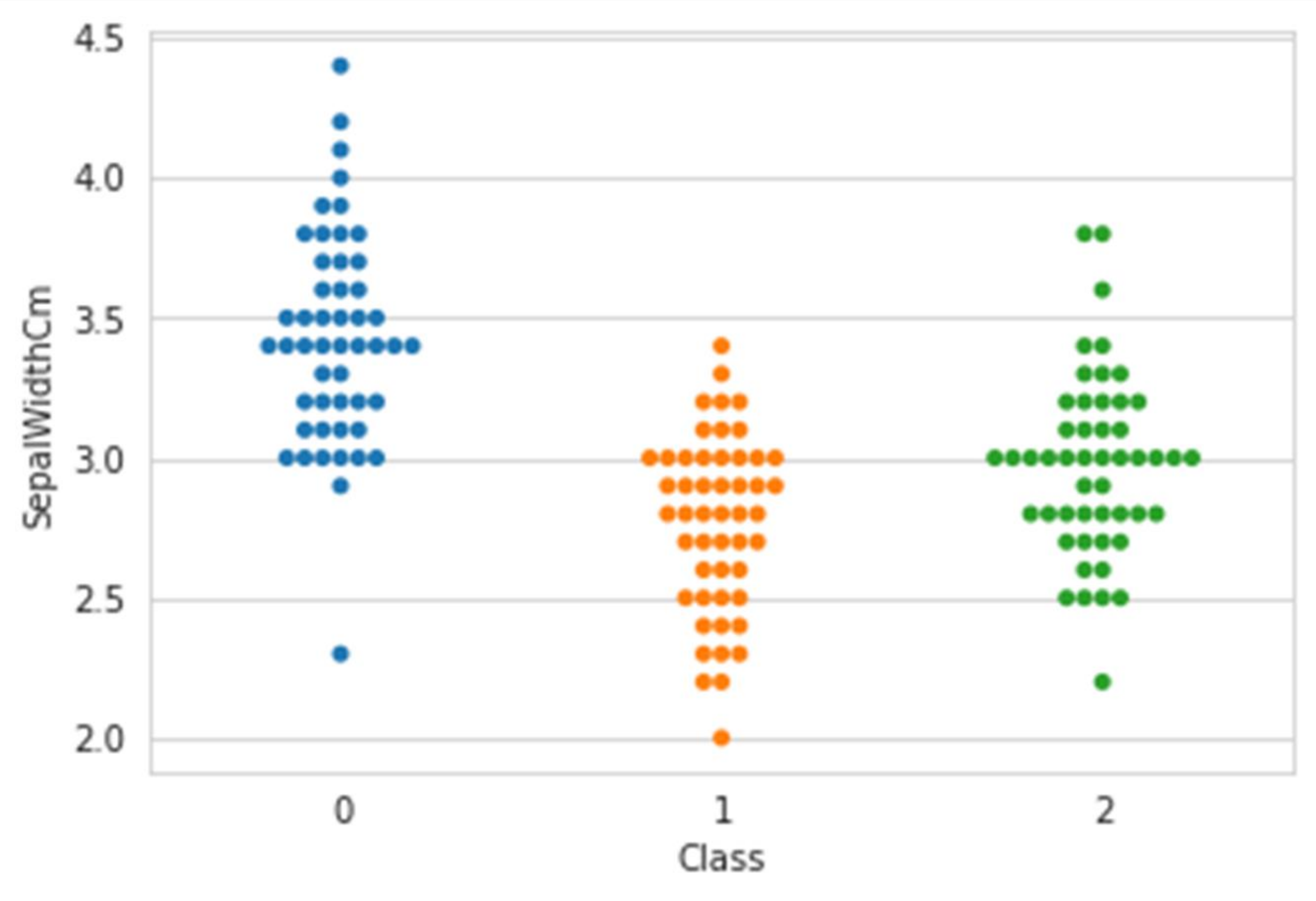
Swarm Plot

- A **swarm plot** is not useful in the case of huge datasets or datasets that are not scaled properly. (can't scale data)



Swarm Plot

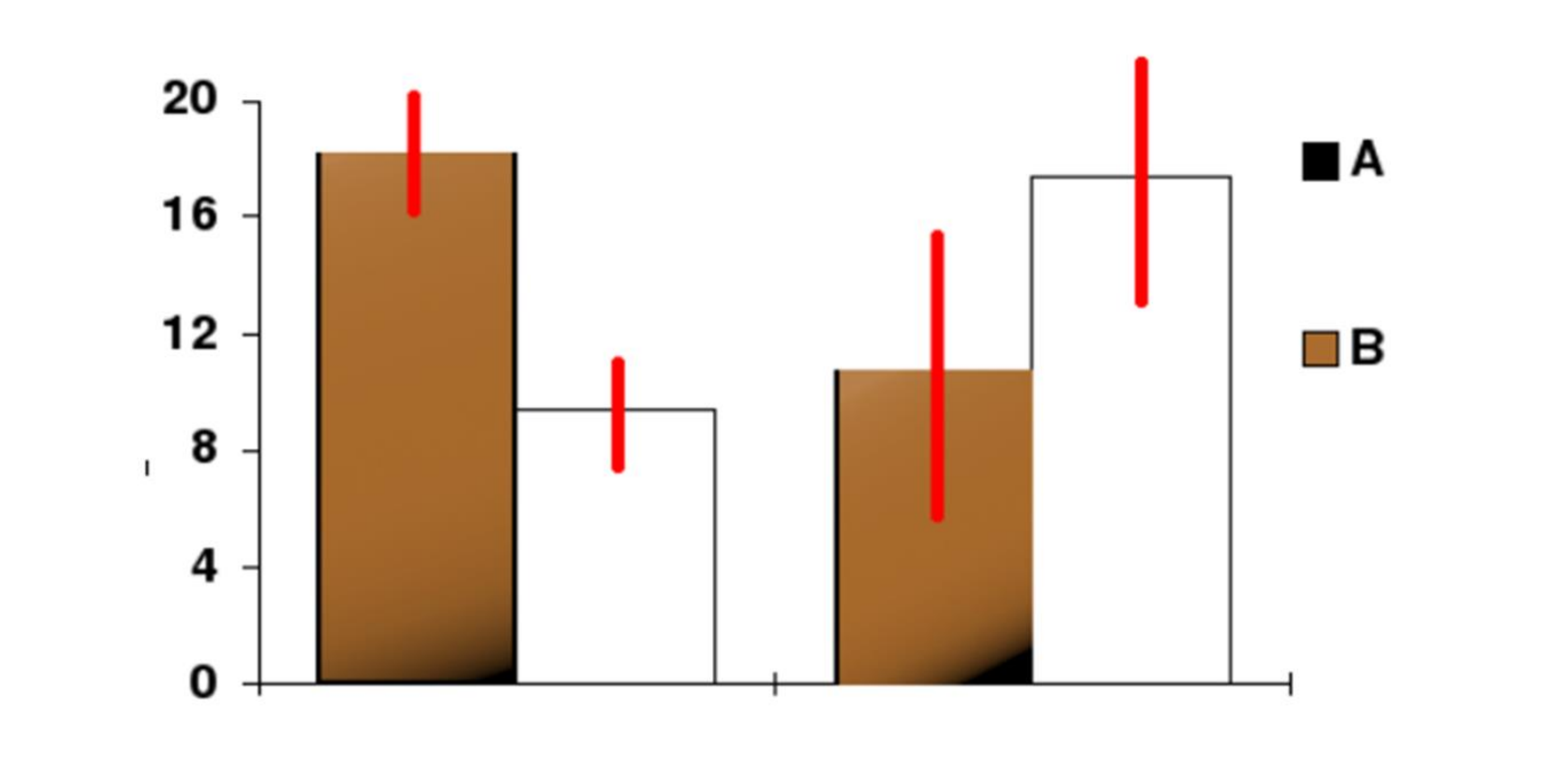
- ❖ sns.swarmplot(x="Class" , y="SepalWidthCm" , data=iris_pd)



Error Plot

Error Plot

- A line through a point on a graph, parallel to one of the axes, which represents the variability or uncertainty in a reported measurement.
- The standard deviation of a dataset is often presented as **error bars**.

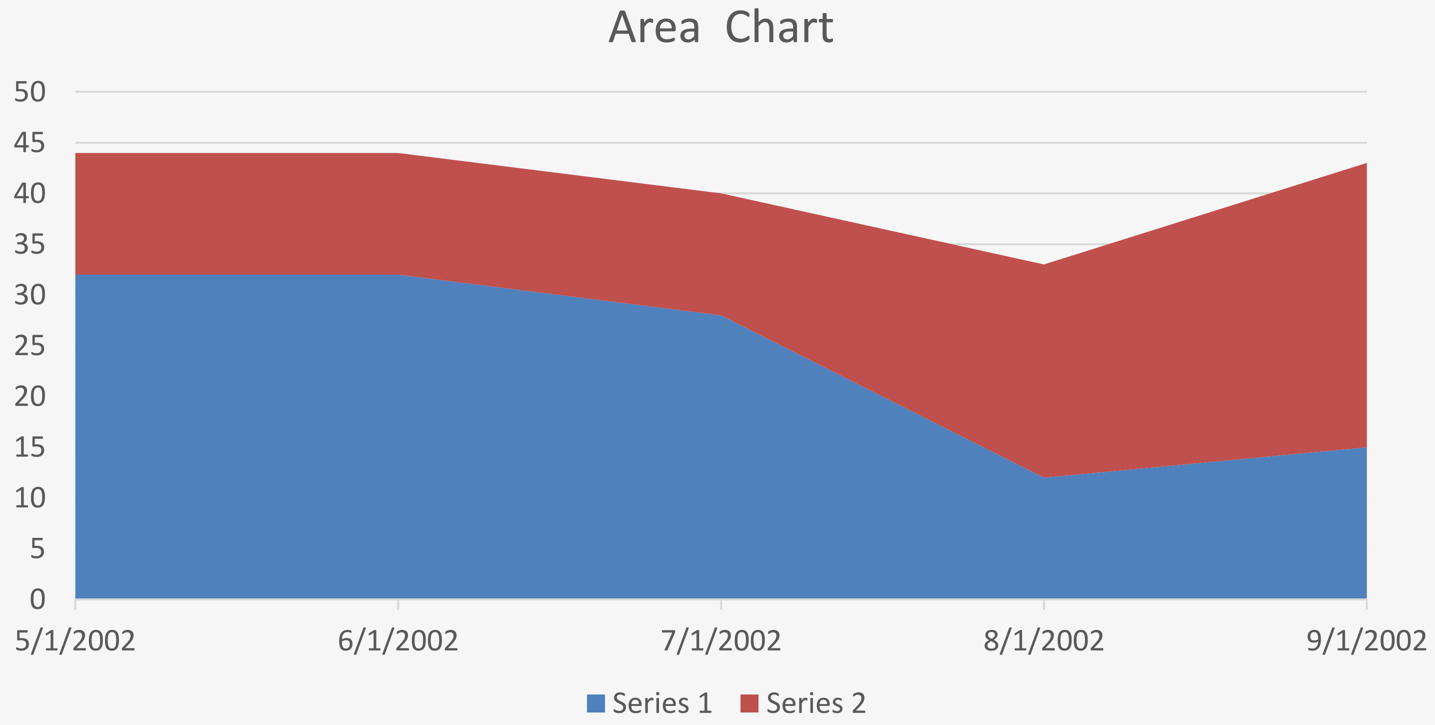


The Learning Hub

Area Chart

Area Chart

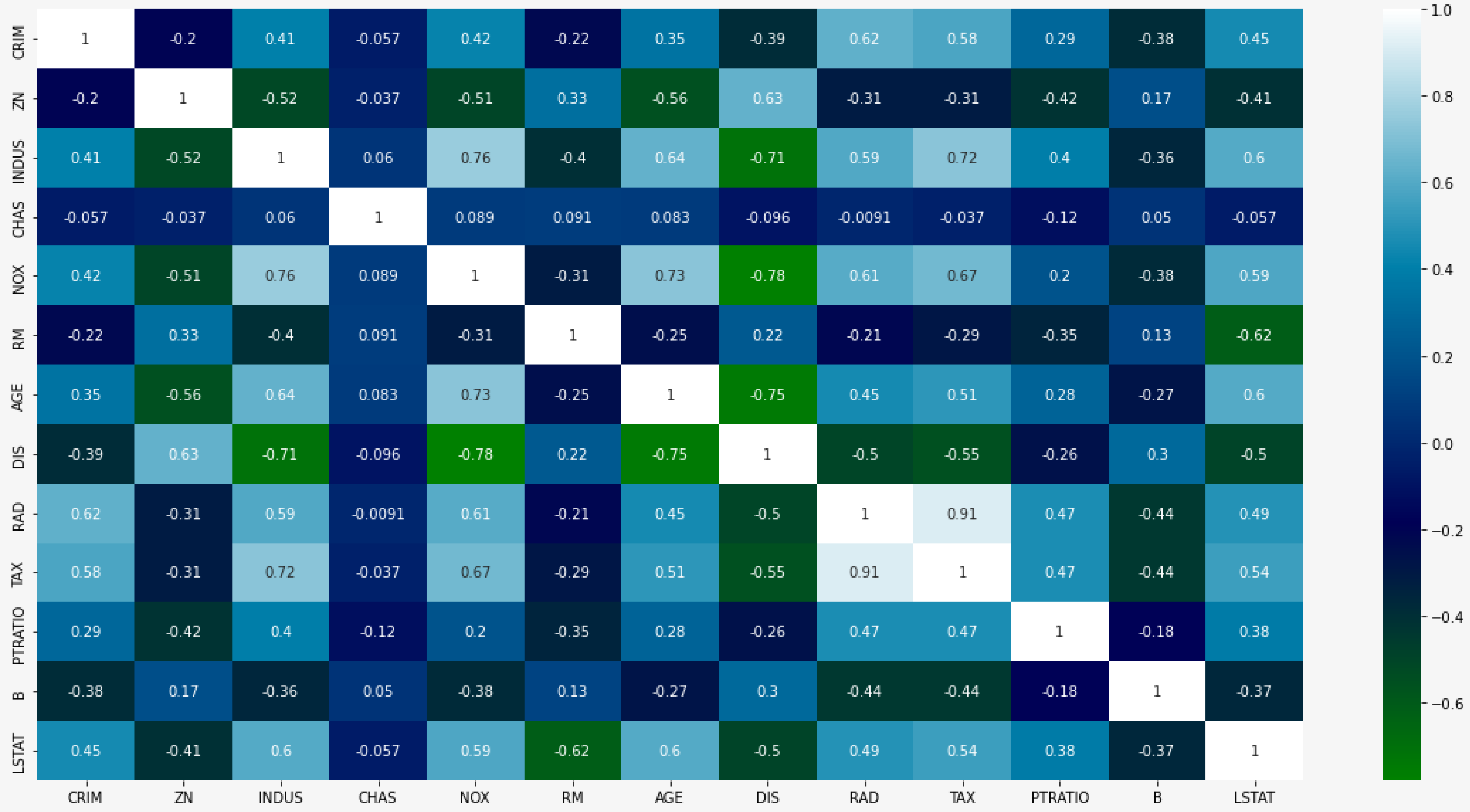
- An area chart is basically a line chart, but the area below the line is coloured in or shaded. It is useful for comparing values between groups, showing changes over time, and the data overlapping.
- It is also useful for analyzing both overall and individual trend information.



Heat Map

Heat Map

- ❖ Heat Map is a data visualization technique that displays the magnitude of a phenomenon using varying colours or saturation.
- ❖ It can be used a lot to show correlations between variables.



Subplot

Subplots

- In the case of plotting multiple plots within the same figure are needed. It is a useful technique to perform a side-by-side comparison of two or more plots

- To visualize multiple plots together, we can create a figure (overall canvas) and divide it into subplots, each containing a plot. With subplots.

Subplots

- Typical syntax is :

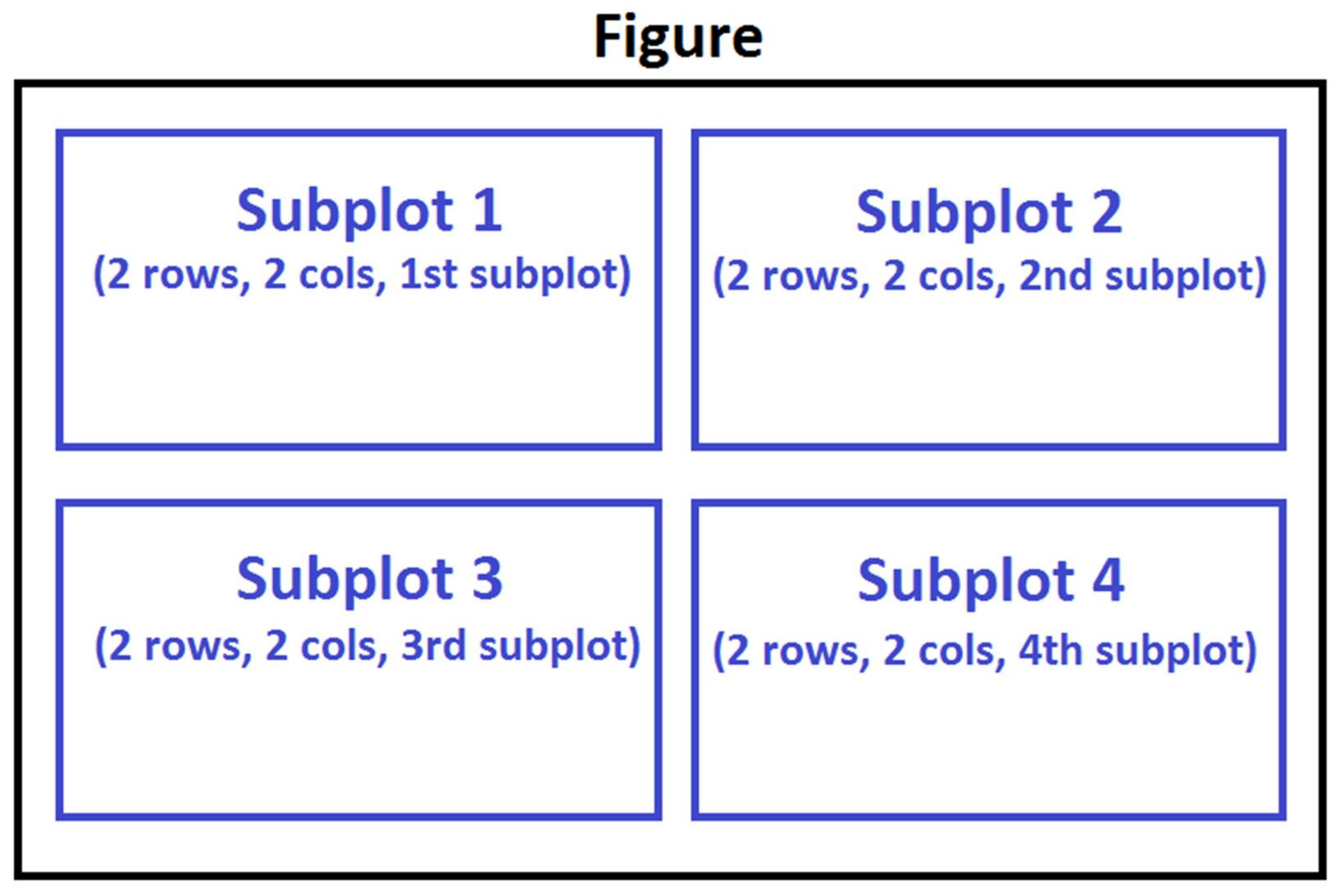
```
fig = plt.figure()                                    # create figure
ax = fig.add_subplot(nrows, ncols, plot_number)       # create subplots
```

Where,

- **nrows** and **ncols** : parameters used to split the figure into (nrows * ncols) sub-axes,
- **plot_number** is used to identify the particular subplot that this function is to create within the national grid.

Subplots

❖ **plot_number** starts at 1, increments across rows first and has a maximum of nrows * ncols as shown below



Figure

Subplot 1
(2 rows, 2 cols, 1st subplot)

Subplot 2
(2 rows, 2 cols, 2nd subplot)

Subplot 3
(2 rows, 2 cols, 3rd subplot)

Subplot 4
(2 rows, 2 cols, 4th subplot)

References

1. Oetting, J. (2022) 14 best types of charts and graphs for data visualization [+ guide], HubSpot Blog. HubSpot. Available at: https://blog.hubspot.com/marketing/types-of-graphs-for-data-visualization (Accessed: January 26, 2023).

2. Bubble chart (no date) Bubble Chart - Learn about this chart and tools to create it. Available at: https://datavizcatalogue.com/methods/bubble_chart.html (Accessed: January 26, 2023).

3. What is an area chart? (no date) TIBCO Software. Available at: https://www.tibco.com/reference-center/what-is-an-area-chart (Accessed: January 26, 2023).

4. Ajaykumar, A. (2022) Data Visualization in python: Data visualization for beginners, Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/02/an-intuitive-guide-to-visualization-in-python/ (Accessed: January 26, 2023).

THANK YOU