Welcome to Week 3

Data Analysis and Visualisation

Session Outline

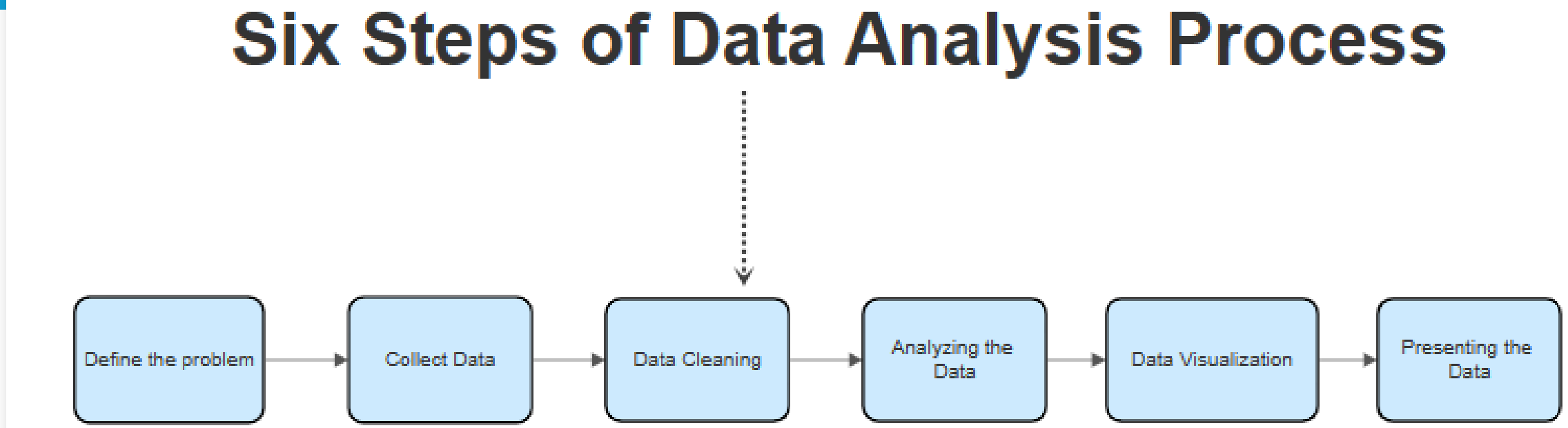
Introduction to Data Analysis

Types of Data

Data Cleaning Basics

Data Exploration – Analysis

Introduction to Data Visualization



Introduction to Data Analysis

The process of inspecting, cleaning, transforming, and modeling data to discover useful information, draw conclusions, and support decision-making.

Key steps

Define Objectives: What are you trying to achieve?

Data Collection: Gather data relevant to your objectives.

Data Cleaning: Handle missing or inconsistent data.

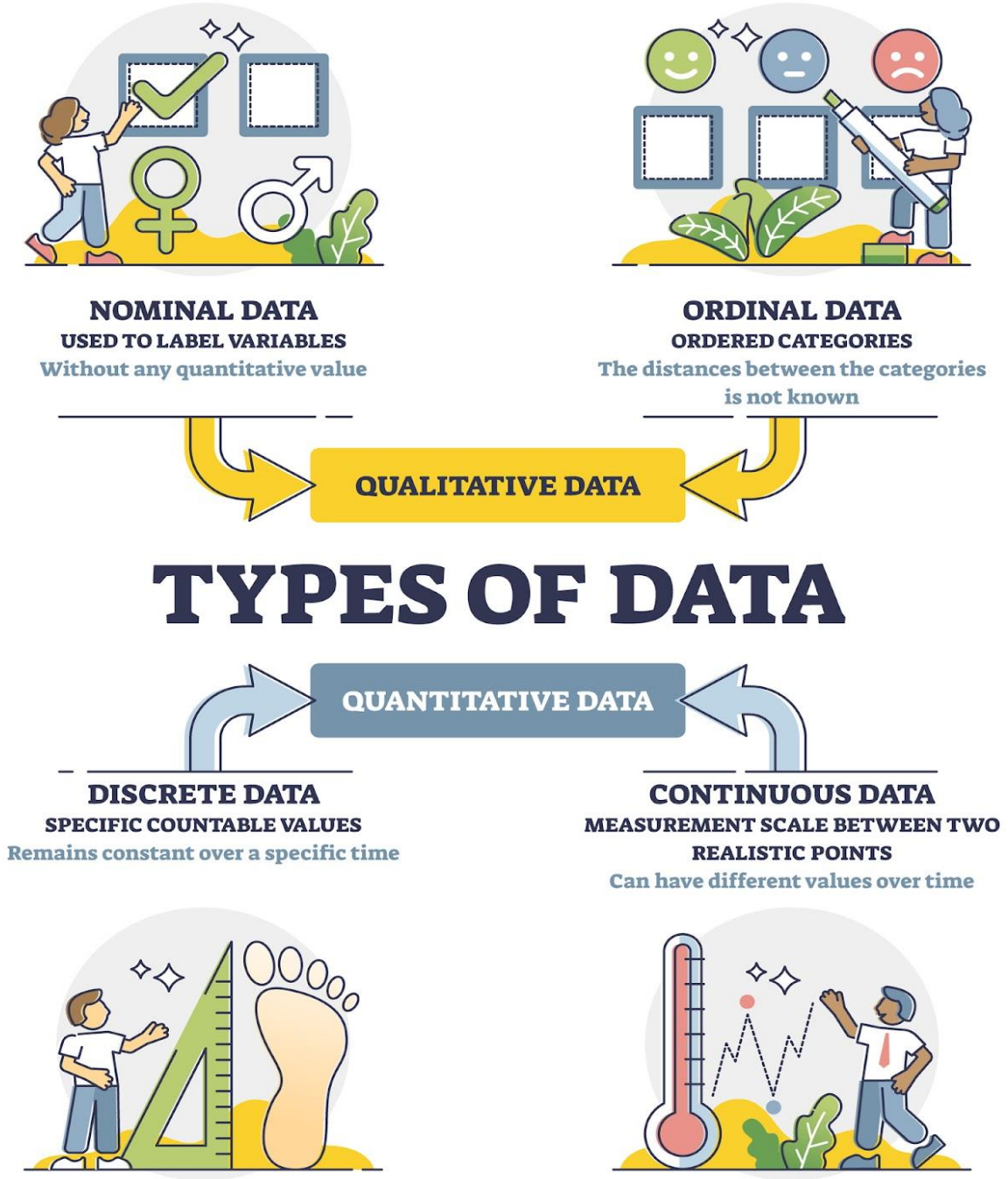
Analysis: Explore and model the data.

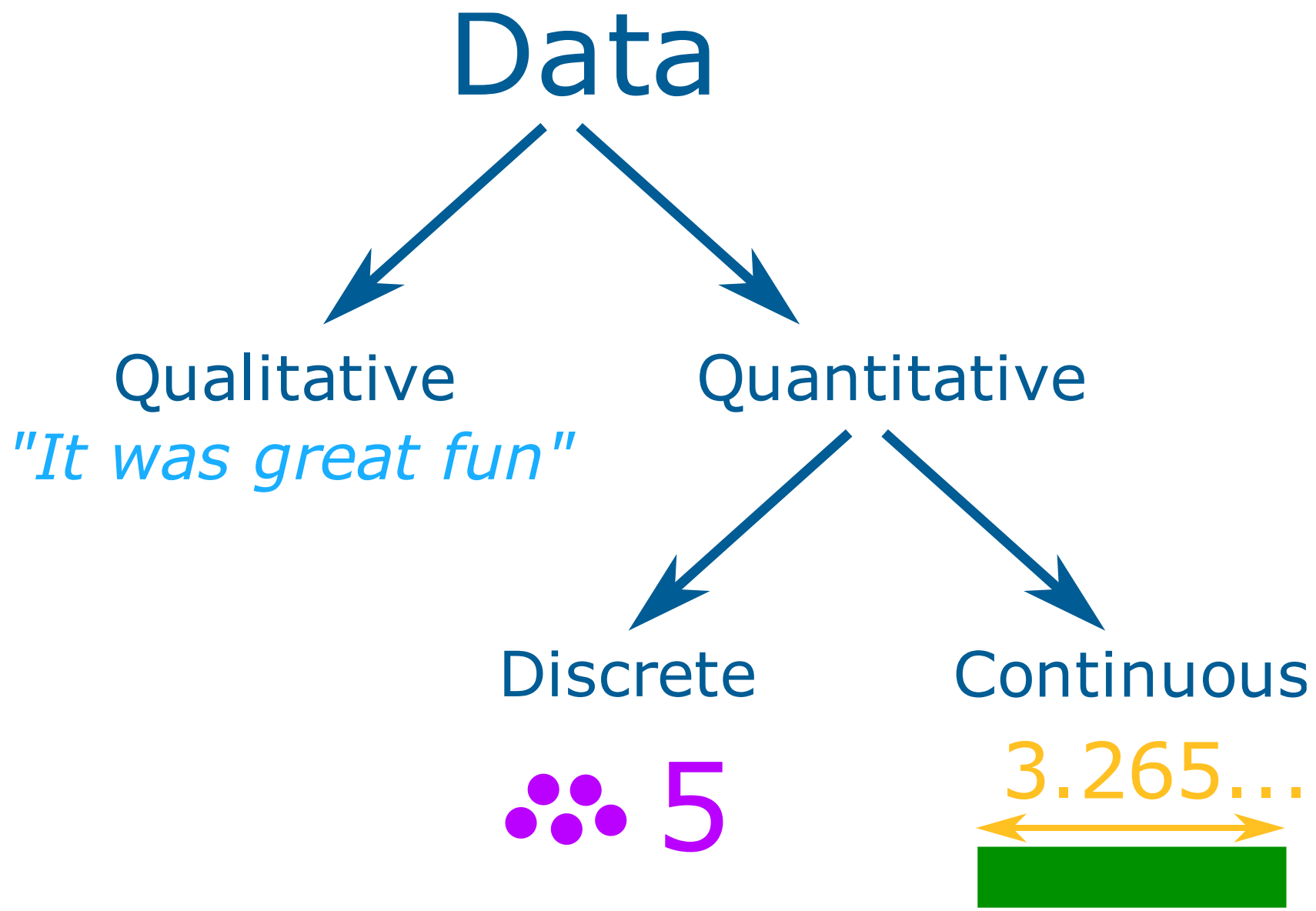
Visualization: Represent data visually for insights.

But why do data analysis?

- Identifies trends, patterns, and anomalies.
- Aids in making informed decisions.

Types of Data




Data
Qualitative
"It was great fun"
Quantitative
Discrete
5
Continuous
3.265...

There are two types of categorical data: nominal and ordinal.

Nominal data

NOMINAL DATA

DEFINITION

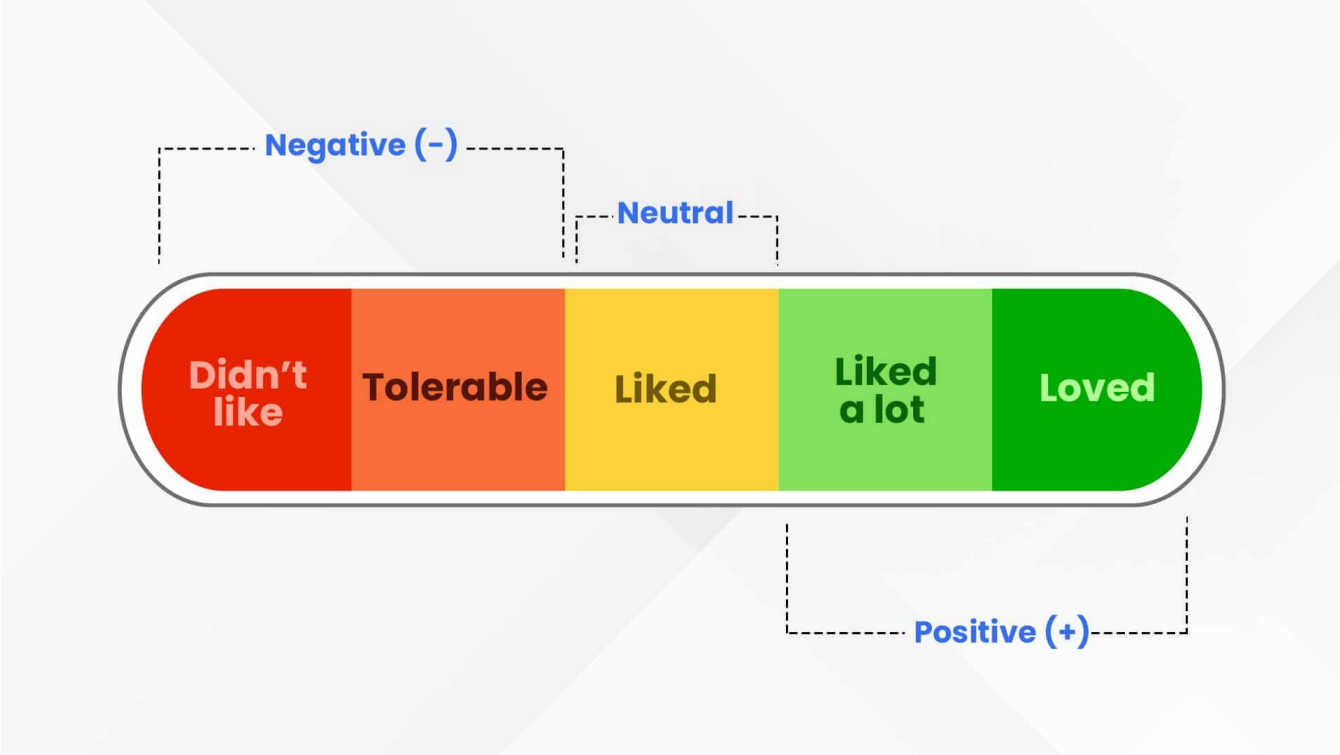
Nominal data refers to data that can be categorized but not ranked or measured in a specific order. It is purely descriptive and does not have an inherent numerical value.

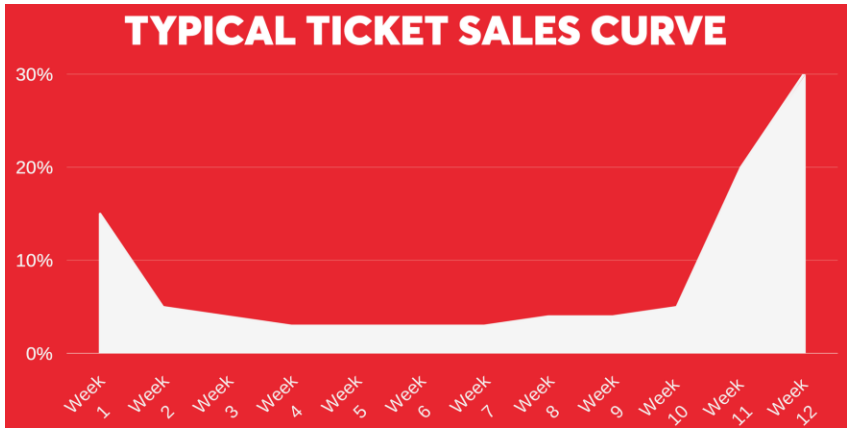
EXAMPLES

- **The colors of a rainbow:** red, orange, yellow, green, blue, indigo, and violet.
- **Types of pets:** dog, cat, fish, bird, and hamster.
- **Names of countries:** USA, France, Brazil, and Japan.

Ordinal data

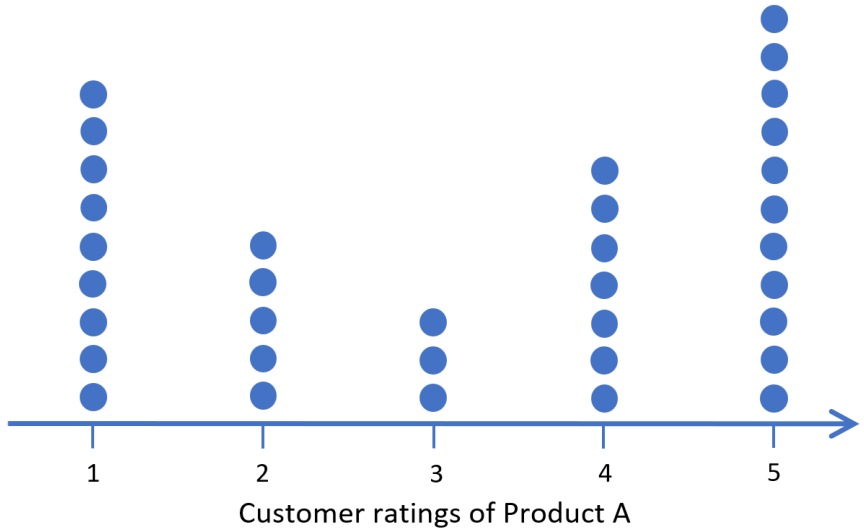
In social scientific research, ordinal variables often include ratings about opinions or perceptions, or demographic factors that are categorised into levels or brackets (such as social status or income).





Discrete data

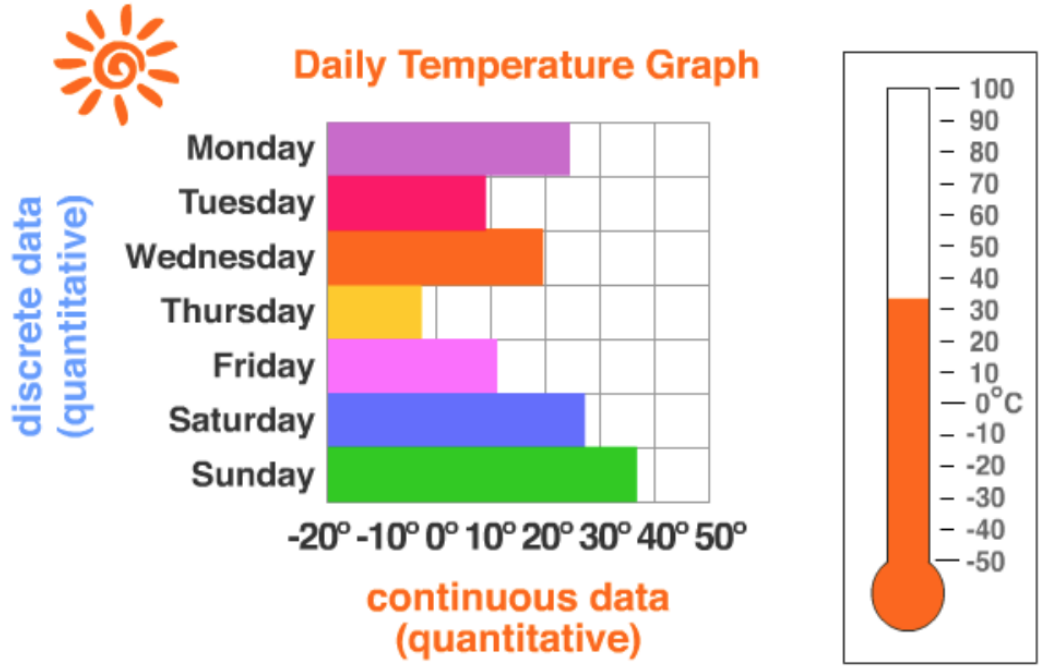
- Ticket sales : The number of ticket sales on one day is a common example of discrete data.



- Number of product reviews: The number of reviews a company's product receives in a specific time frame, such as one week, is another example of discrete data.

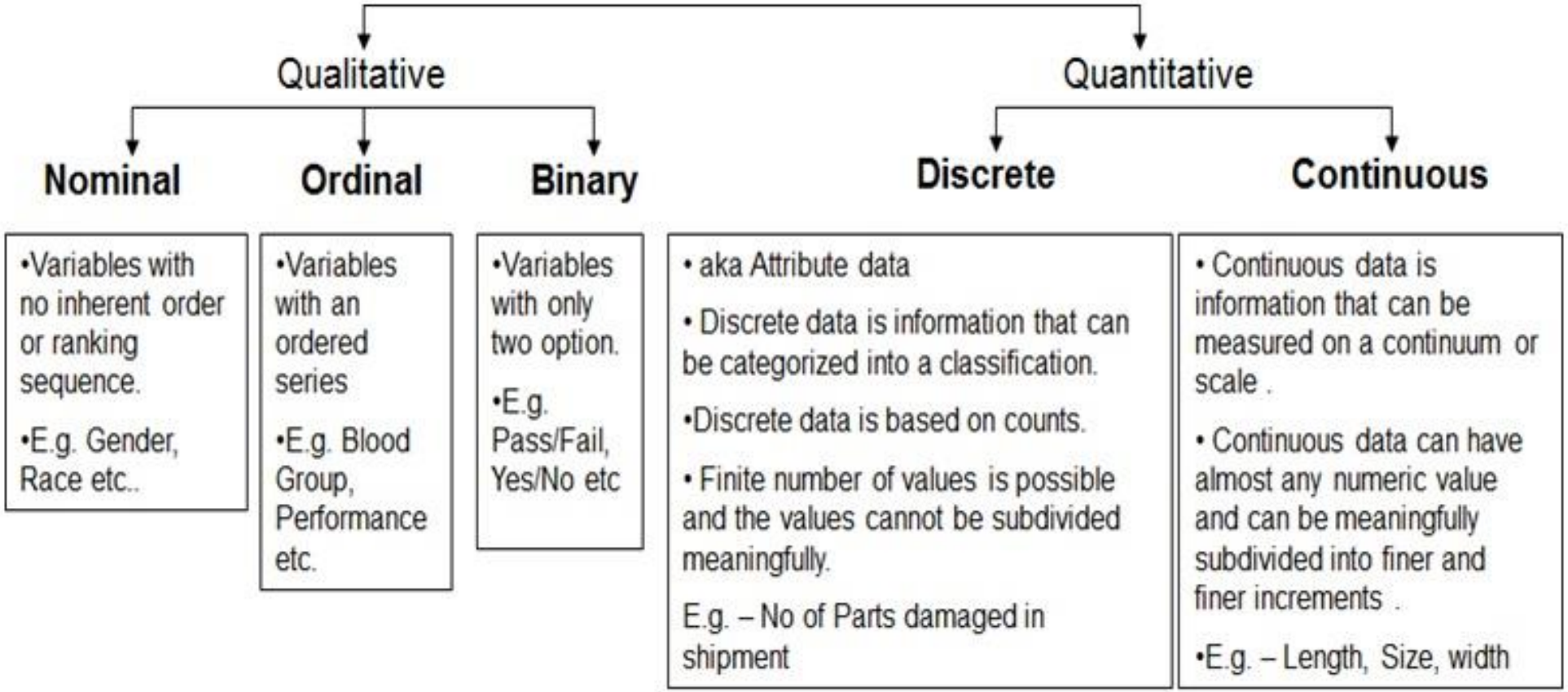
continuous data

quantitative data that can be measured



In this graph the days of the week are discrete data but the temperature is continuous data.

continuous data — infinite values
discrete data — finite values


Qualitative
Quantitative
Nominal
Ordinal
Binary
Discrete
Continuous
•Variables with no inherent order or ranking sequence.
•E.g. Gender, Race etc..
•Variables with an ordered series
•E.g. Blood Group, Performance etc.
•Variables with only two option.
•E.g. Pass/Fail, Yes/No etc
• aka Attribute data
• Discrete data is information that can be categorized into a classification.
•Discrete data is based on counts.
• Finite number of values is possible and the values cannot be subdivided meaningfully.
E.g. – No of Parts damaged in shipment
• Continuous data is information that can be measured on a continuum or scale .
• Continuous data can have almost any numeric value and can be meaningfully subdivided into finer and finer increments .
•E.g. – Length, Size, width

|Recap

Data Cleaning Basics

- The process of fixing or removing incorrect, corrupted, or irrelevant data.

- Bad data could be:
- Empty cells
- Data in wrong format
- Wrong data
- Duplicates

Steps in cleaning data:

- Identify missing data (e.g., `df.isnull()` in Python).
- Handle duplicates (e.g., `df.drop_duplicates()`).
- Standardize formats (e.g., date formats).

- **Example**:
- Lab 2 covers handling missing values and removing duplicates in a dataset.

```python
# Importing pandas and numpy
import pandas as pd
import numpy as np

# Sample DataFrame with missing values
data = {'First Score': [100, 90, np.nan, 95],
        'Second Score': [30, 45, 56, np.nan],
        'Third Score': [np.nan, 40, 80, 98]}

df = pd.DataFrame(data)

# Checking for missing values using isnull()
missing_values = df.isnull()

print(missing_values)
```

Example 2: Filtering Data based on missing values

```
import pandas as pd

data = pd.read_csv("employees.csv")
bool_series = pd.isnull(data["Gender"])
missing_gender_data = data[bool_series]
print(missing_gender_data)
```

Output

	First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team
20	Lois	NaN	4/22/1995	7:18 PM	64714	4.934	True	Legal
22	Joshua	NaN	3/8/2012	1:58 AM	90816	18.816	True	Client Services
27	Scott	NaN	7/11/1991	6:58 PM	122367	5.218	False	Legal
31	Joyce	NaN	2/20/2005	2:40 PM	88657	12.752	False	Product
41	Christine	NaN	6/28/2015	1:08 AM	66582	11.308	True	Business Development
49	Chris	NaN	1/24/1980	12:13 PM	113590	3.055	False	Sales
51	NaN	NaN	12/17/2011	8:29 AM	41126	14.009	NaN	Sales
53	Alan	NaN	3/3/2014	1:28 PM	40341	17.578	True	Finance
60	Paula	NaN	11/23/2005	2:01 PM	48866	4.271	False	Distribution
64	Kathleen	NaN	4/11/1990	6:46 PM	77834	18.771	False	Business Development
69	Irene	NaN	7/14/2015	4:31 PM	100863	4.382	True	Finance
70	Todd	NaN	6/10/2003	2:26 PM	84692	6.617	False	Client Services
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
939	Ralph	NaN	7/28/1995	6:53 PM	70635	2.147	False	Client Services
945	Gerald	NaN	4/15/1989	12:44 PM	93712	17.426	True	Distribution
961	Antonio	NaN	6/18/1989	9:37 PM	103050	3.050	False	Legal
972	Victor	NaN	7/28/2006	2:49 PM	76381	11.159	True	Sales
985	Stephen	NaN	7/10/1983	8:10 PM	85668	1.909	False	Legal
989	Justin	NaN	2/10/1991	4:58 PM	38344	3.794	False	Legal
995	Henry	NaN	11/23/2014	6:09 AM	132483	16.655	False	Distribution

145 rows × 8 columns

Checking for Missing Values Using notnull()

```python
# Importing pandas and numpy
import pandas as pd
import numpy as np

# Sample DataFrame with missing values
data = {'First Score': [100, 90, np.nan, 95],
        'Second Score': [30, 45, 56, np.nan],
        'Third Score': [np.nan, 40, 80, 98]}

df = pd.DataFrame(data)

# Checking for non-missing values using notnull()
non_missing_values = df.notnull()

print(non_missing_values)
```

Output:

	First Score	Second Score	Third Score
0	True	True	False
1	True	True	True
2	False	True	True
3	True	False	True

Example 4: Filtering Data with Non-Missing Values

```python
# Importing pandas
import pandas as pd

# Reading data from a CSV file
data = pd.read_csv("employees.csv")

# Identifying non-missing values in the 'Gender' column
non_missing_gender = pd.notnull(data["Gender"])

# Filtering rows where 'Gender' is not missing
non_missing_gender_data = data[non_missing_gender]

display(non_missing_gender_data)
```

Output:

	First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team
0	Douglas	Male	8/6/1993	12:42 PM	97308	6.945	True	Marketing
1	Thomas	Male	3/31/1996	6:53 AM	61933	4.170	True	NaN
2	Maria	Female	4/23/1993	11:17 AM	130590	11.858	False	Finance
3	Jerry	Male	3/4/2005	1:00 PM	138705	9.340	True	Finance
4	Larry	Male	1/24/1998	4:47 PM	101004	1.389	True	Client Services
5	Dennis	Male	4/18/1987	1:35 AM	115163	10.125	False	Legal
6	Ruby	Female	8/17/1987	4:20 PM	65476	10.012	True	Product
7	NaN	Female	7/20/2015	10:43 AM	45906	11.598	NaN	Finance
8	Angela	Female	11/22/2005	6:29 AM	95570	18.523	True	Engineering
9	Frances	Female	8/8/2002	6:51 AM	139852	7.524	True	Business Development
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
994	George	Male	6/21/2013	5:47 PM	98874	4.479	True	Marketing
996	Phillip	Male	1/31/1984	6:30 AM	42392	19.675	False	Finance
997	Russell	Male	5/20/2013	12:39 PM	96914	1.421	False	Product
998	Larry	Male	4/20/2013	4:45 PM	60500	11.985	False	Business Development
999	Albert	Male	5/15/2012	6:24 PM	129949	10.169	True	Sales

855 rows × 8 columns

Data Analysis

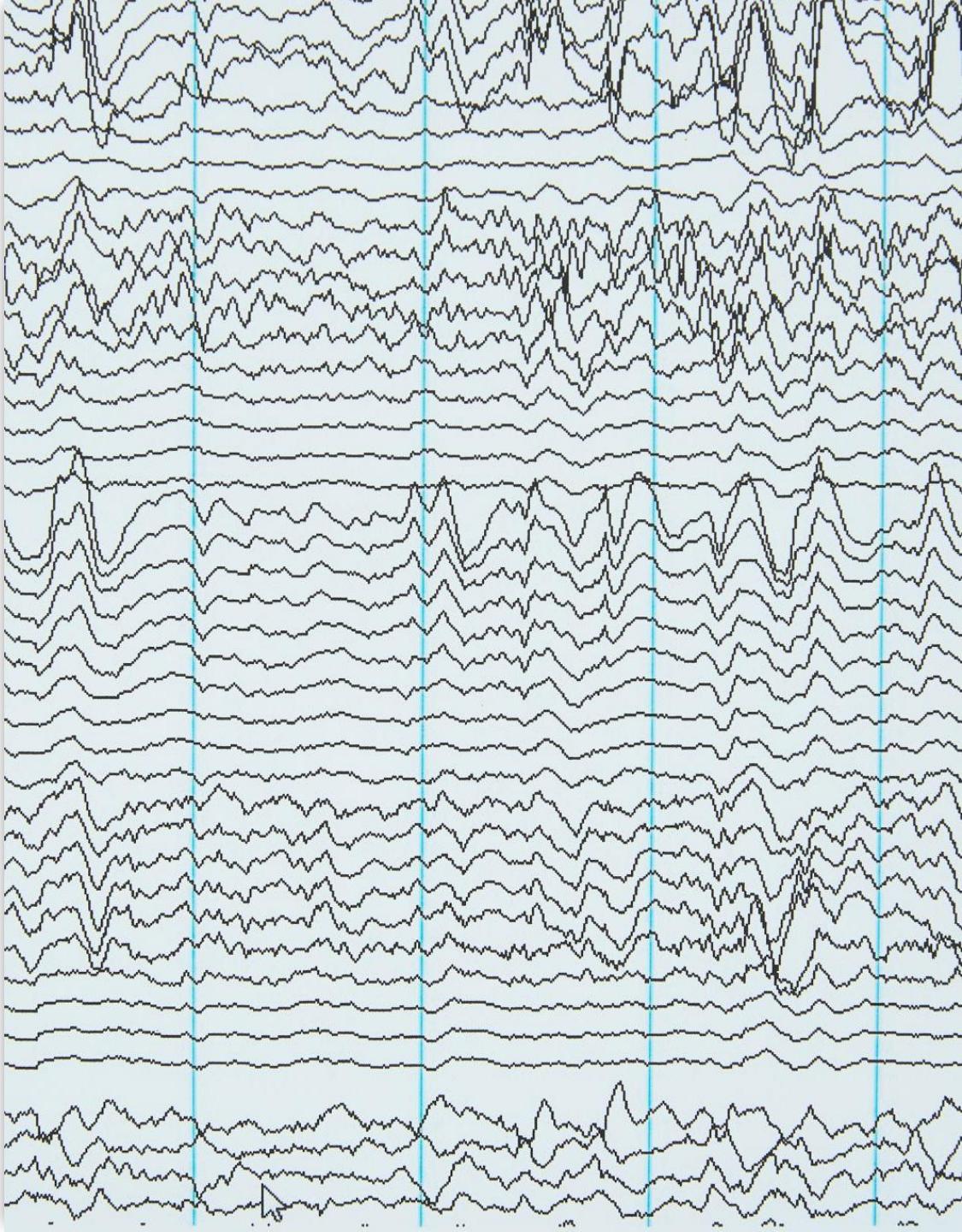
The objective is to understand the dataset's structure and summary statistics.

Summary statistics: Mean, median, mode, range, standard deviation.

Data visualisation: Histograms, box plots.

Correlation snalysis: Relationships between variables.

- Lab 1 provides examples of running summary statistics and correlations?.



Introduction to Data Visualisation

The graphical representation of data to communicate insights effectively.

Why Visualise?

Makes complex data easier to understand.

Highlights patterns and outliers.

Choosing good colors for your charts is hard

There is a whole lot of material on this side of visualising data.

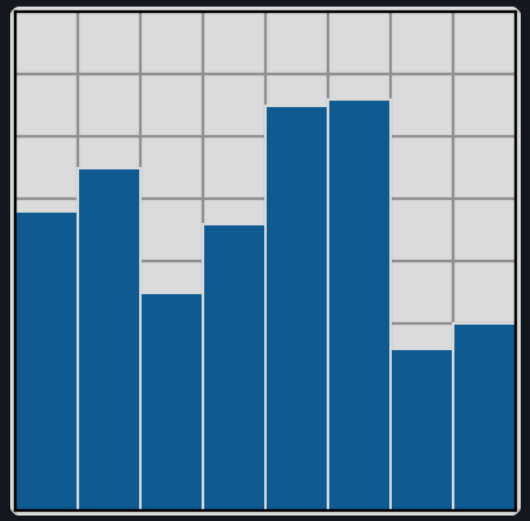
Lab 3 on DataVisualisation has an exercise on this.

https://blog.datawrapper.de/beautifulcolors/

Types of charts_1

bar(x, height)

See `bar`.

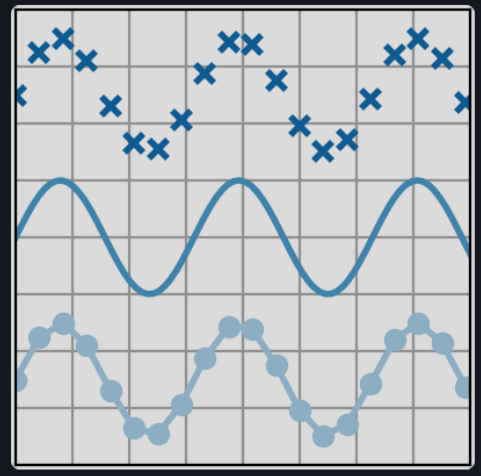


Types of charts_2

plot(x, y)

Plot y versus x as lines and/or markers.

See `plot`.

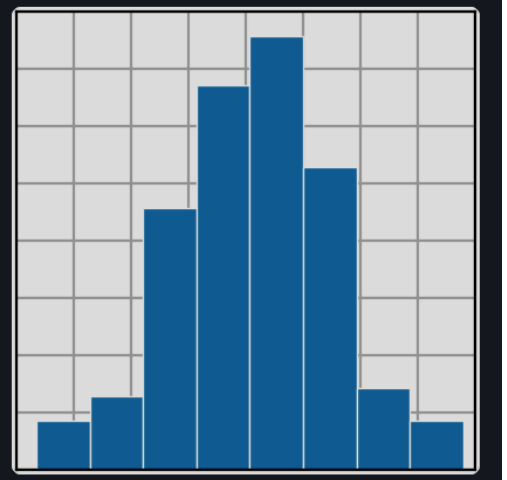


Types of charts_3

hist(x)

Compute and plot a histogram.

See `hist`.



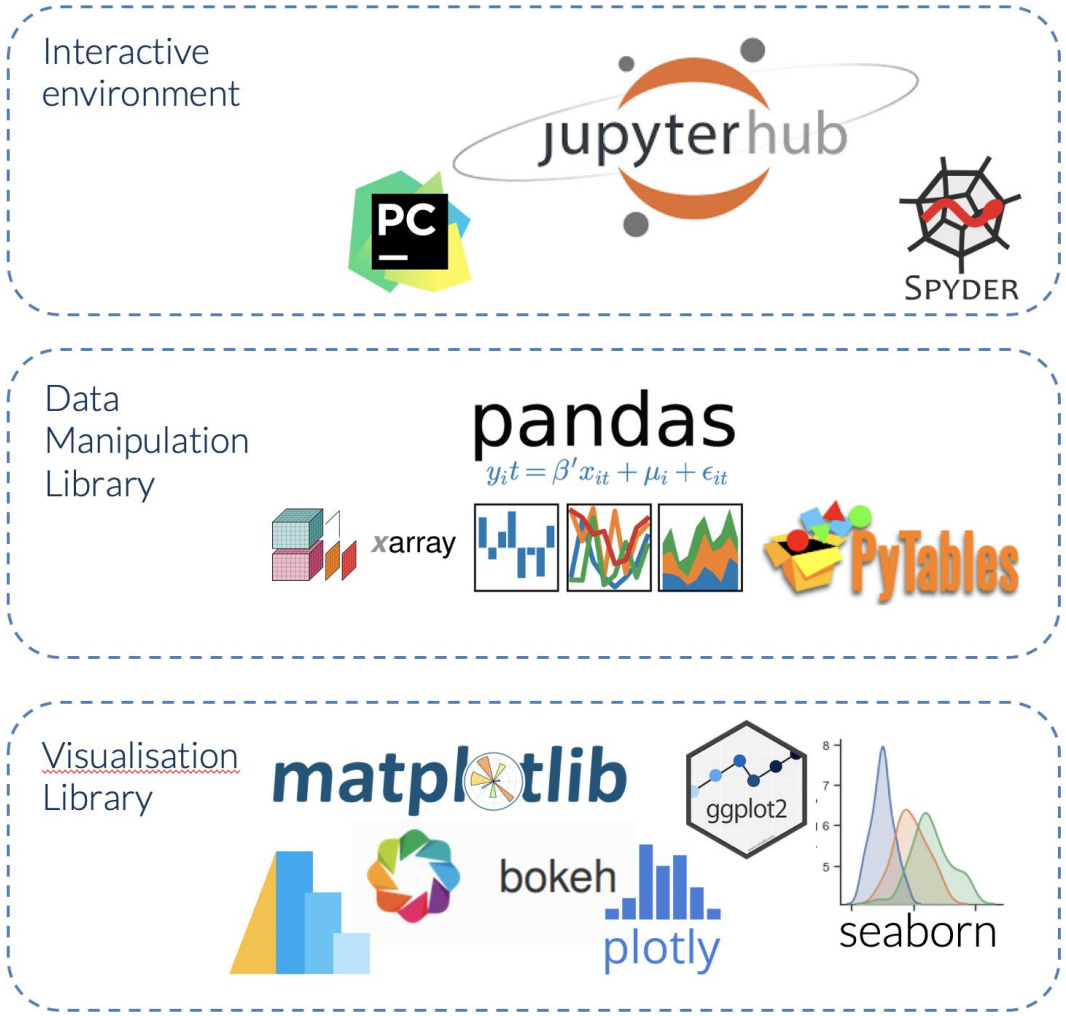
Visualisation in Python

- Some b**asic libraries**:

- **Matplotlib**: Versatile but requires detailed coding.
- **Seaborn**: Simplifies statistical visualizations.

But actually…..there are many more ➽

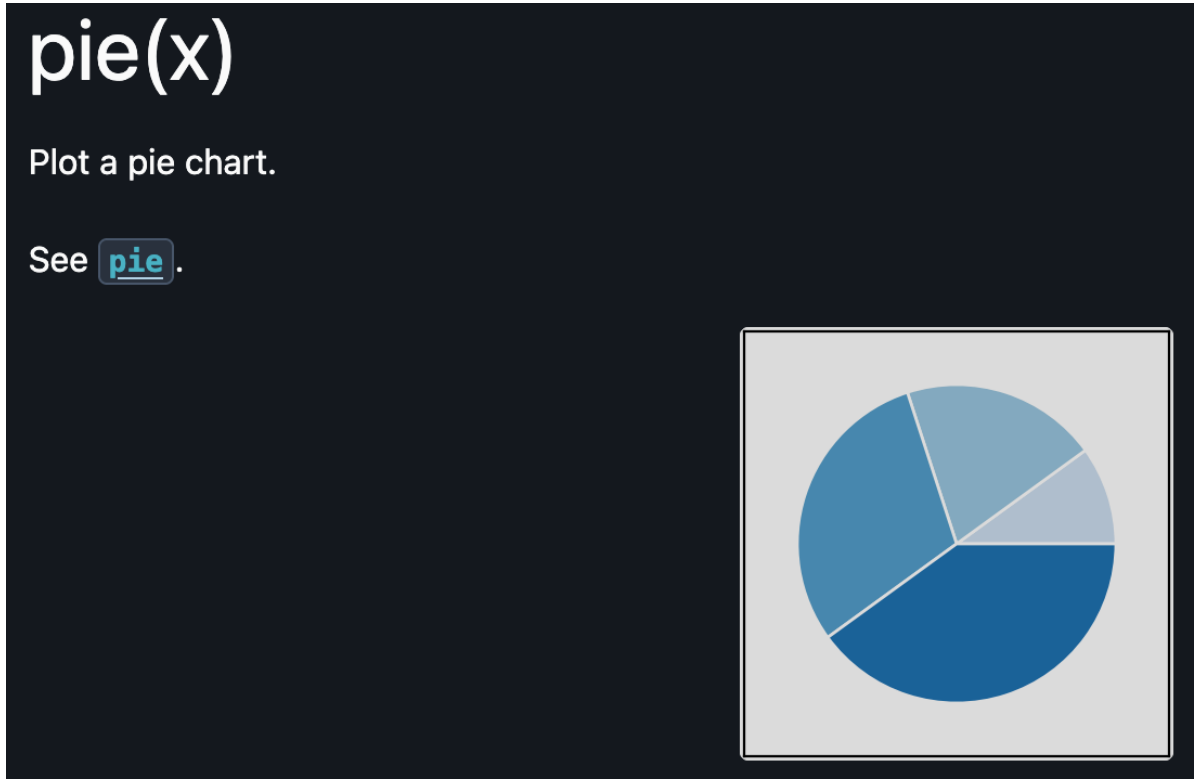
Data Analysis Stacks



Choosing the right visualisation

Pie Chart:

- Best for showing proportions.
- Example: Market share of companies in an industry.

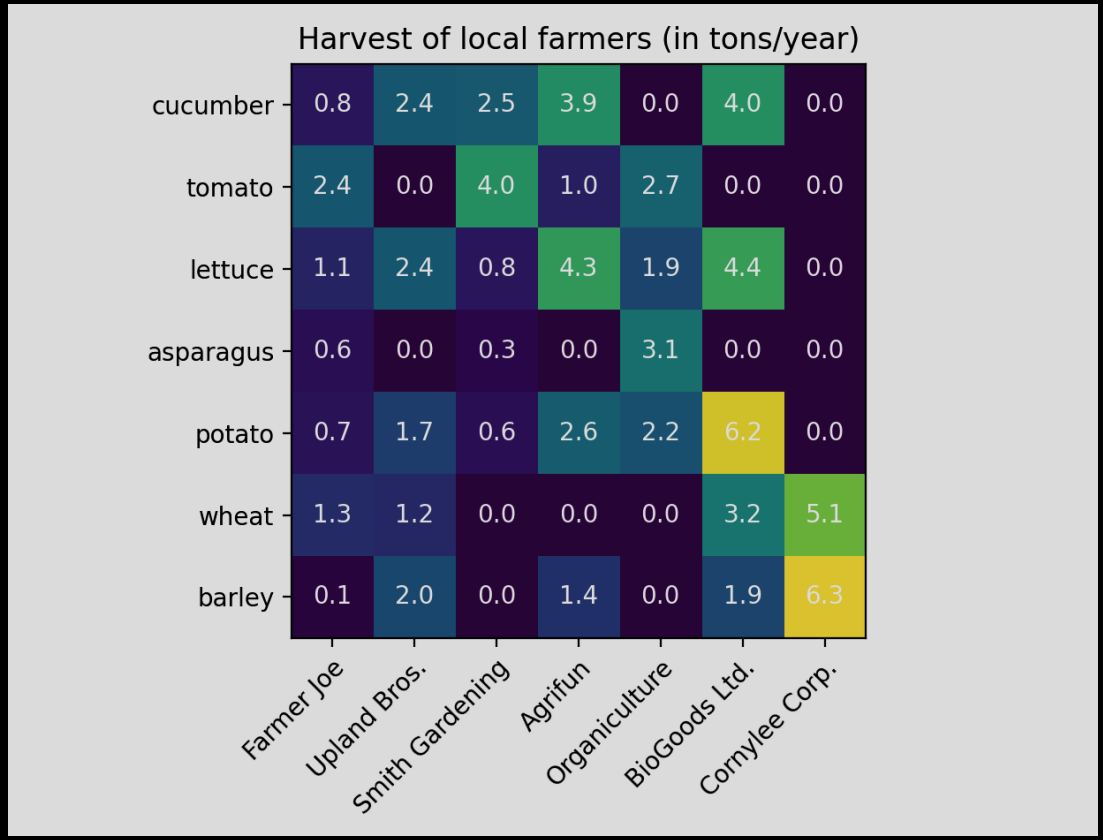


Choosing the right visualisation

Heatmap:

- Best for showing correlations or intensity.
- Example: Correlation matrix for harvest in tons/year.

https://matplotlib.org/stable/gallery/images_contours_and_fields/image_annotated_heatmap.html



Interactive Data Visualisation

Tools:

Plotly: For interactive plots.

Bokeh

Why Interactive?

Allows users to explore data.

Enhances engagement.

Lab 3 DataVisualisation showcases an example of interactive data visualisation

Data Storytelling

Combining data analysis, visuals, and narrative.

Key Components:

- **Data**: Ensure quality and relevance.
- **Visuals**: Choose the right charts.
- **Narrative**: Make the insights relatable.

Lab 3 DataVisualisation has elements of narrative

Labs

There are 3 Lab notebooks to work with – today is slightly different in that you have exercises and examples with running code and some explanation and solutions.

Explore them, make notes, ask questions!