Data Science

data visualization

Dr. Jamal Al Qundus | MEU | WS22

Outline

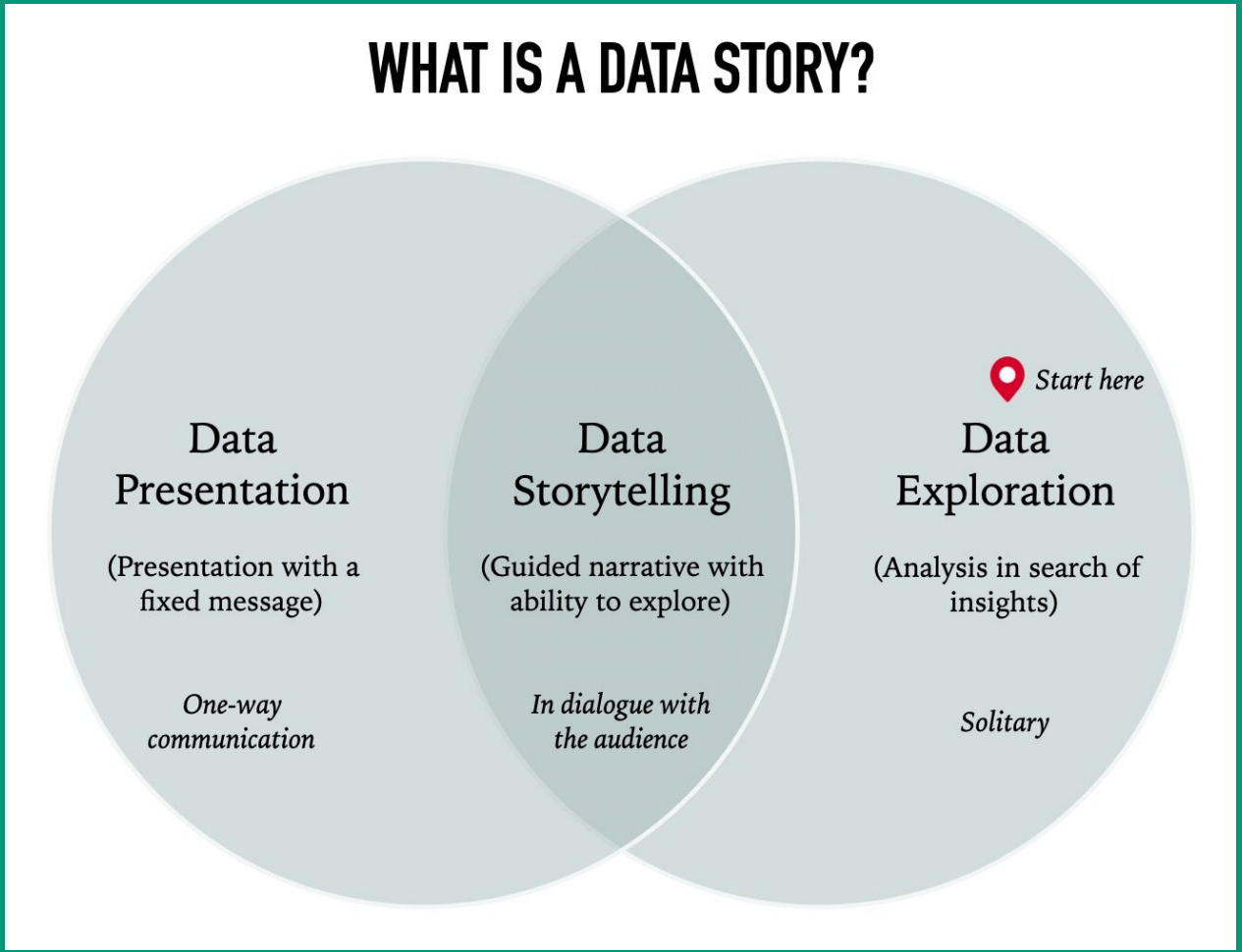
- Basics of visualization
- Data types and visualization types
- Software plotting libraries

Two types of visualization

Data exploration visualization: figuring out what is true

Data presentation visualization: convincing other people it is true

“Data exploration” is much broader than just visualization …



https://www.juiceanalytics.com/writing/5-differences-between-data-exploration-and-data-presentation

MEU جامعة الشرق الأوسط MIDDLE EAST UNIVERSITY Amman - Jordan

Importance of visualization

Before you run any analysis, build any machine learning system, etc, always visualize your data

If you can’t identify a trend or make a prediction for your dataset, neither will an automated algorithm

This is especially important to keep in mind as you hear stories of “superhuman” performance of AI methods (it is possible, but takes a long time, and is not the norm)

Visualization vs. statistics vs. analytics

Visualization almost always presents a more informative (though less quantitative) view of your data than statistics (the noun, not the field)

Statistics e.g. Mean, Median …

Data Visualization	Data Analytics
A graphical representation of information and data.	A process of analyzing data to make decision.
• Identify areas that need attention or improvement • Clarity which factors influence customer behaviour • Helps understand which products to place where	• Identify the underlying models and patterns • Acts as an input source for data visualization • Helps in improving the business by predicting the needs and conclusion.
The goal is to communicate information	To make more-informed business decisions
Static or interactive by a data engineer	Prescriptive and predictive analytics by a data analyst

https://www.geeksforgeeks.org/difference-between-data-visualization-and-data-analytics/

Outline

- Basics of visualization
- Data types and visualization types
- Software plotting libraries

Data types (Levels of Measurement)

Nominal: categorical data, no ordering Example – color: {r,g,b,…},Pet: {dog, cat, rabbit, ...}

Percentage as 20% … no statistics values as mean = ?

Ordinal: categorical data, with ordering Example – Satisfaction or rank: {1,2,3,4,5}

Percentage/frequency as 20% … (no?) statistics values as mean = ?

Interval: numerical data, a value has no fixed meaning Example – Age, size, weight

Ratio: numerical data, a value has special meaning Example – Age

Interval/Ratio can **discrete** as 5 customers, 17 points or **continuous** 4.2 miles 25 degrees

Statistical values Mean, Median …

Visualization Types

Most discussion of visualization types emphasizes what elements the chart is trying to convey

Instead, the focus is on the type and dimensionality of the underlying data

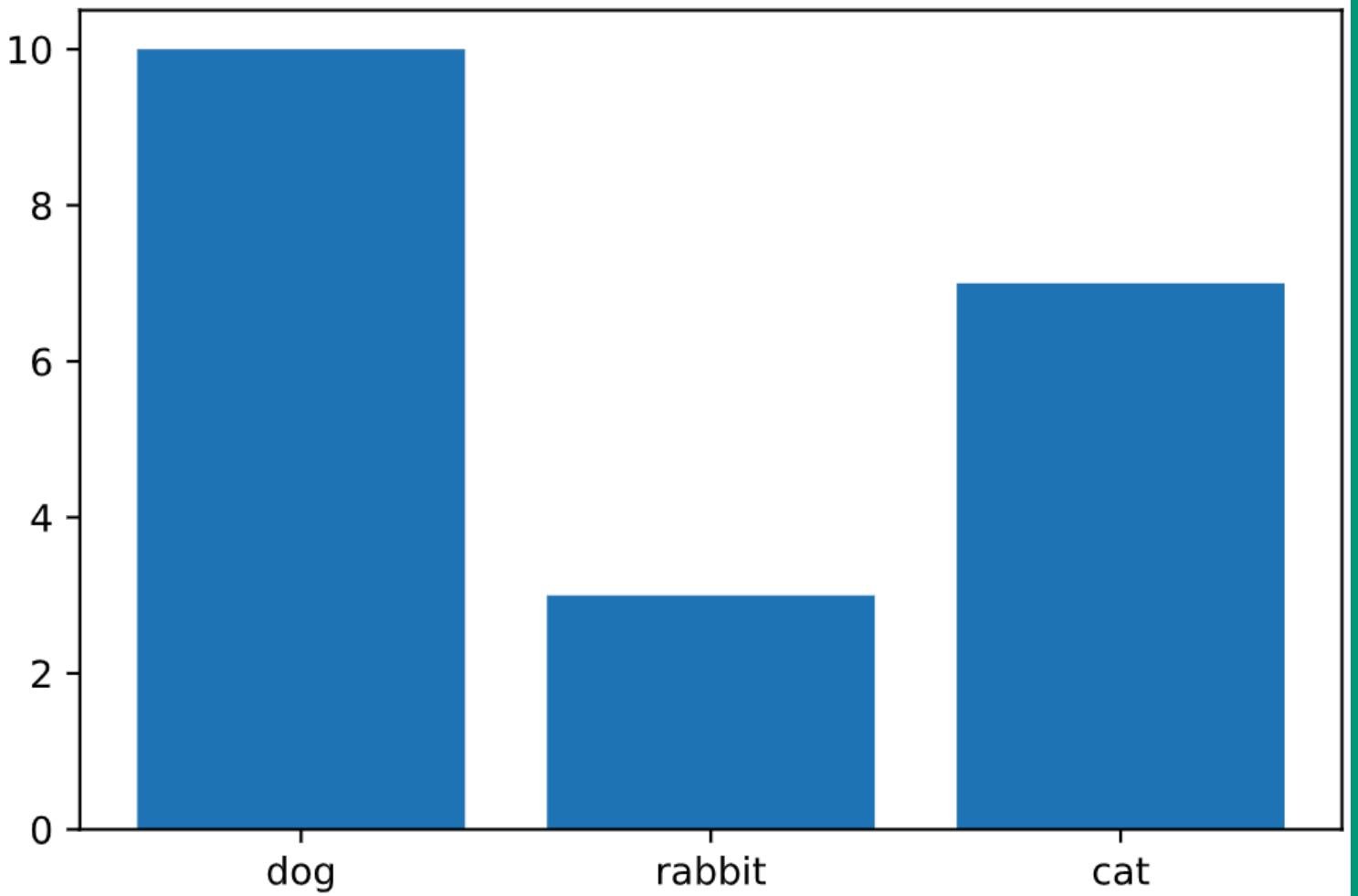
Visualization types (not an exhaustive list):

- 1D: bar chart, pie chart, histogram
- 2D: scatter plot, line plot, box and whisker plot, heatmap
- 3D+: scatter matrix, bubble chart

1D bar chart

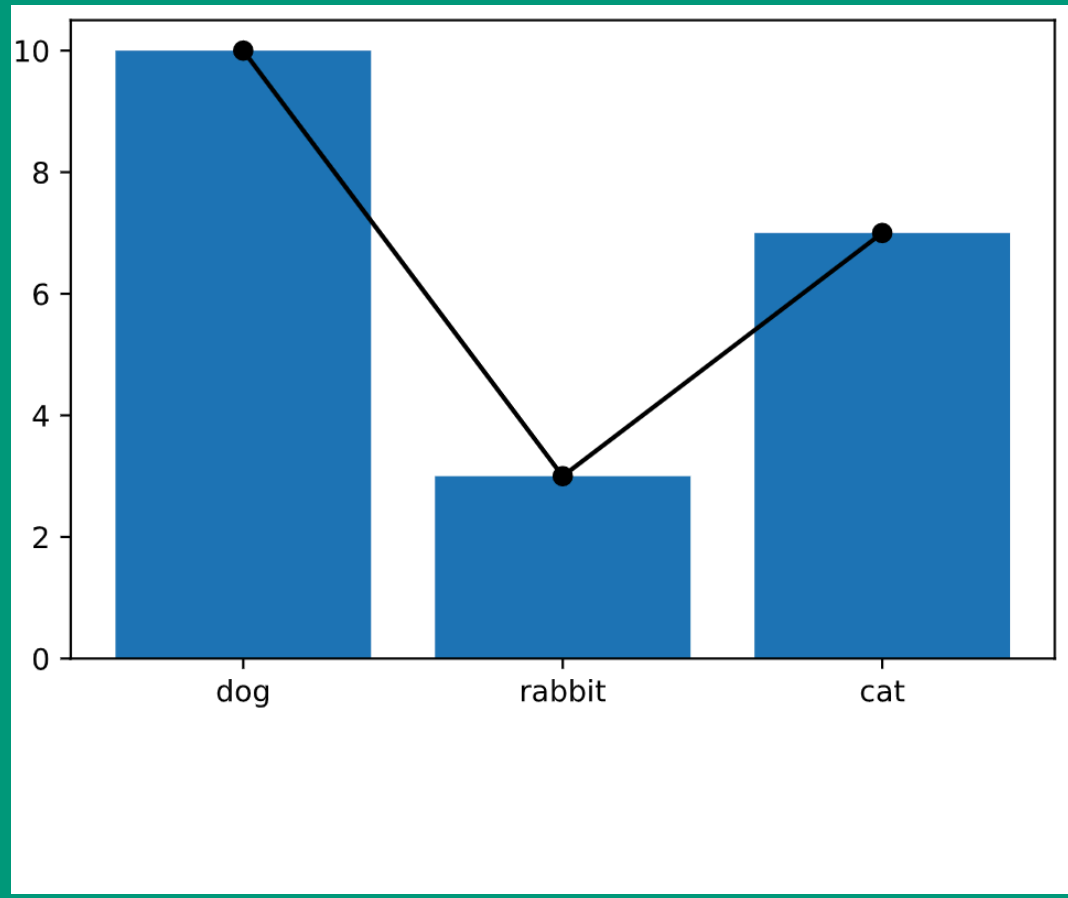
	Data
Nominal	✓
Ordinal	✓
Interval	✗
Ratio	✗

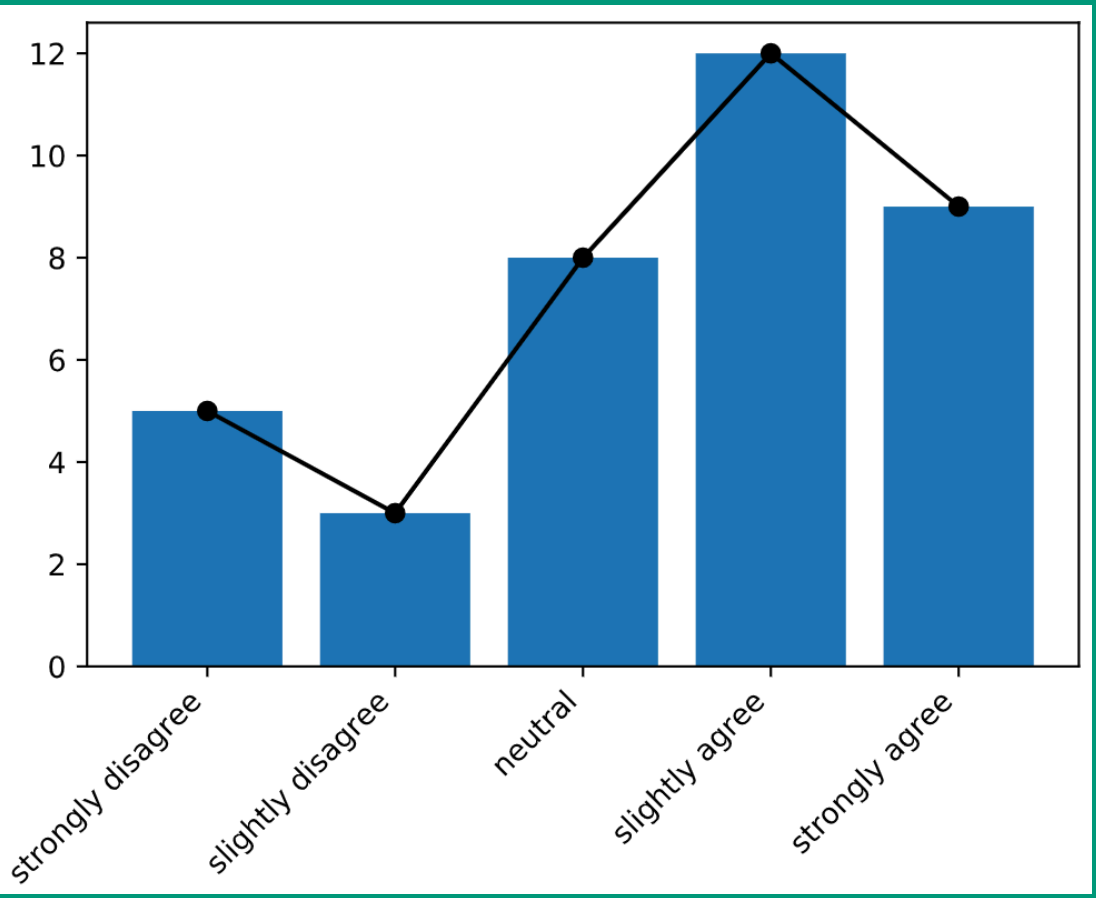
Suggestions, not rules



1D bar chart

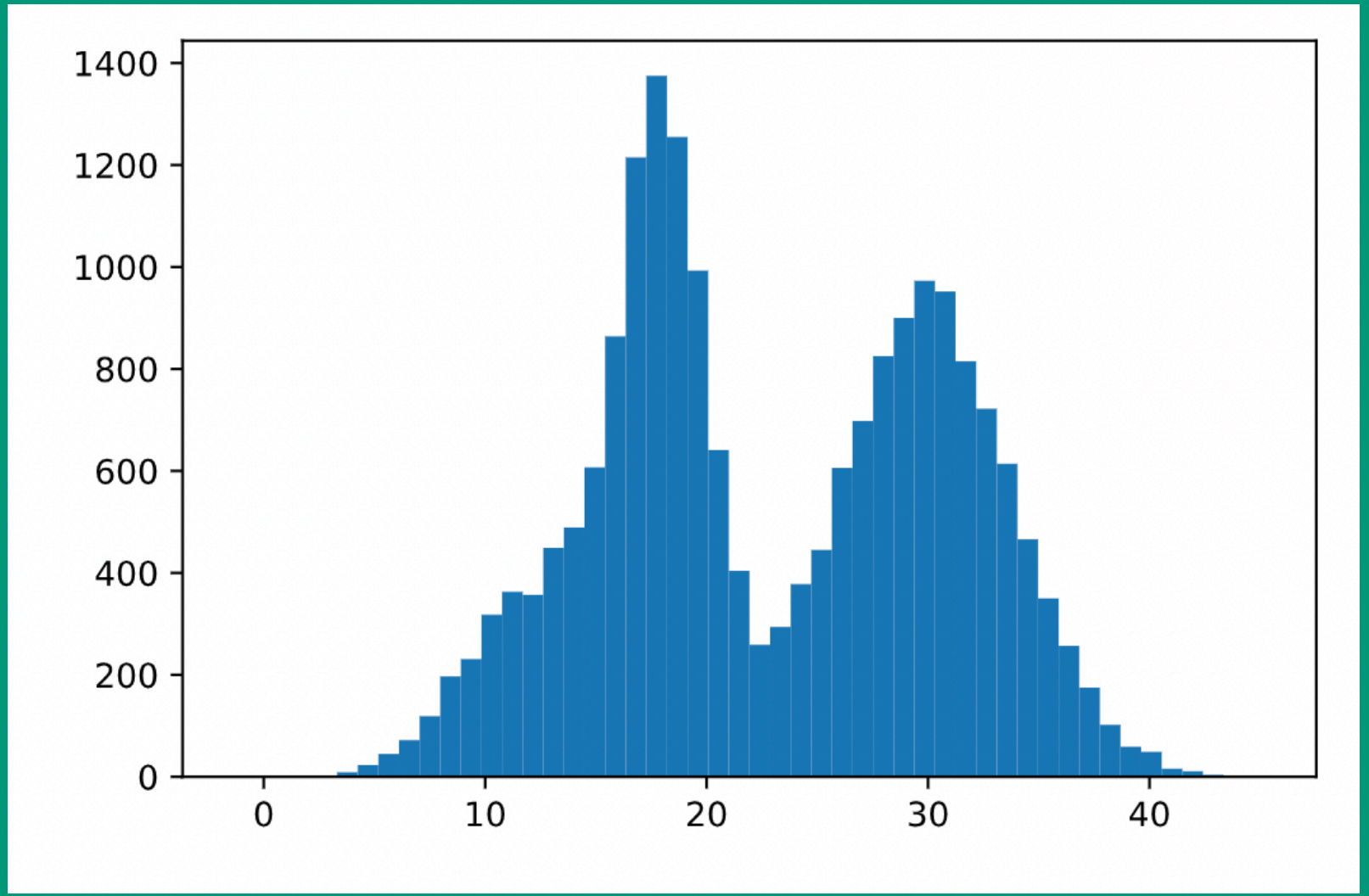
Don't use lines within a bar chart for categorial or ordinal features!





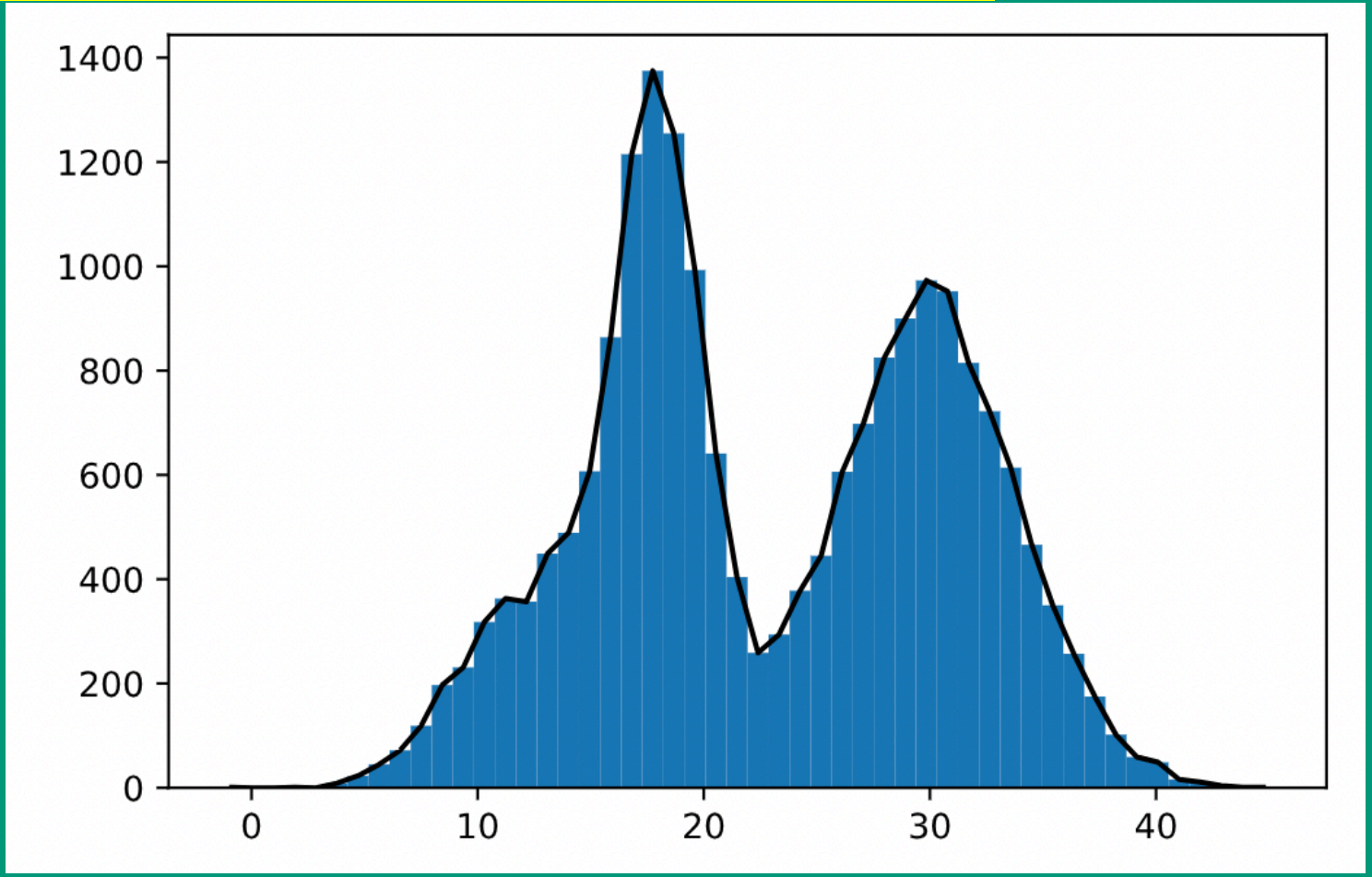
1D histogram

	Data
Nominal	✗
Ordinal	✗
Interval	✓
Ratio	✓



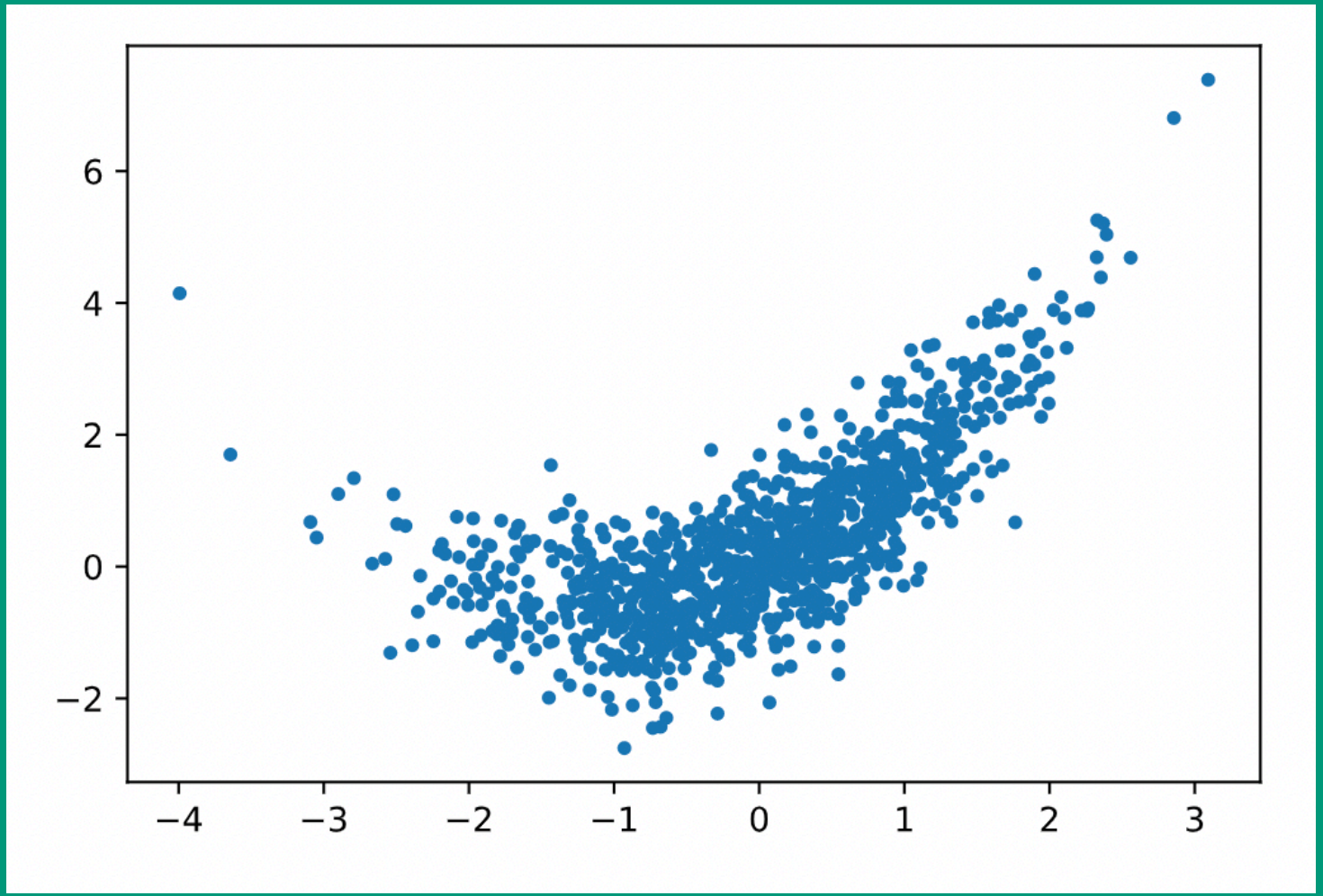
2D histogram

OK to use lines within a histogram (but not very informative)



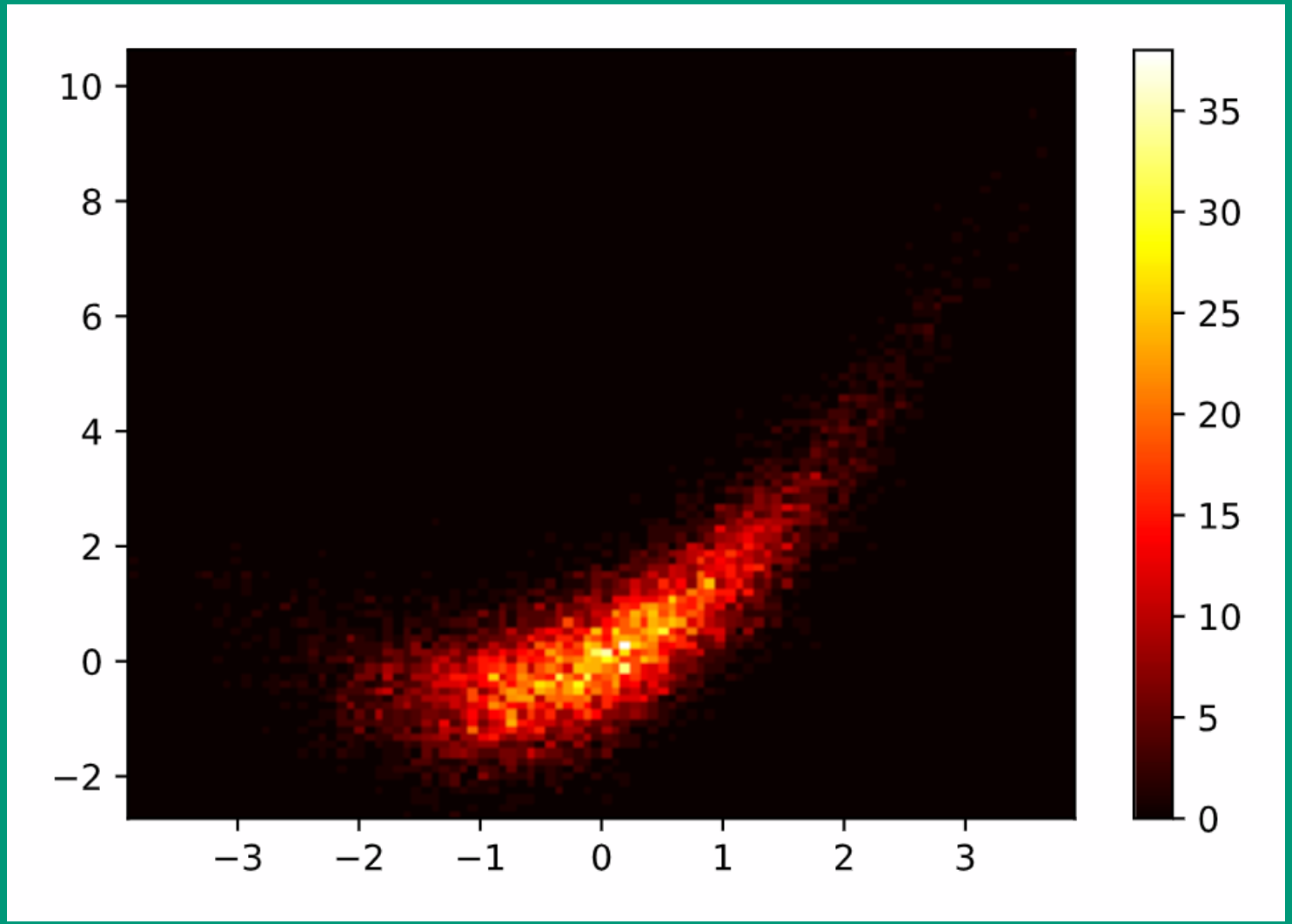
2D scatter plot

	Dim 1	Dim 2
Nominal	✗	✗
Ordinal	✗	✗
Interval	✓	✓
Ratio	✓	✓



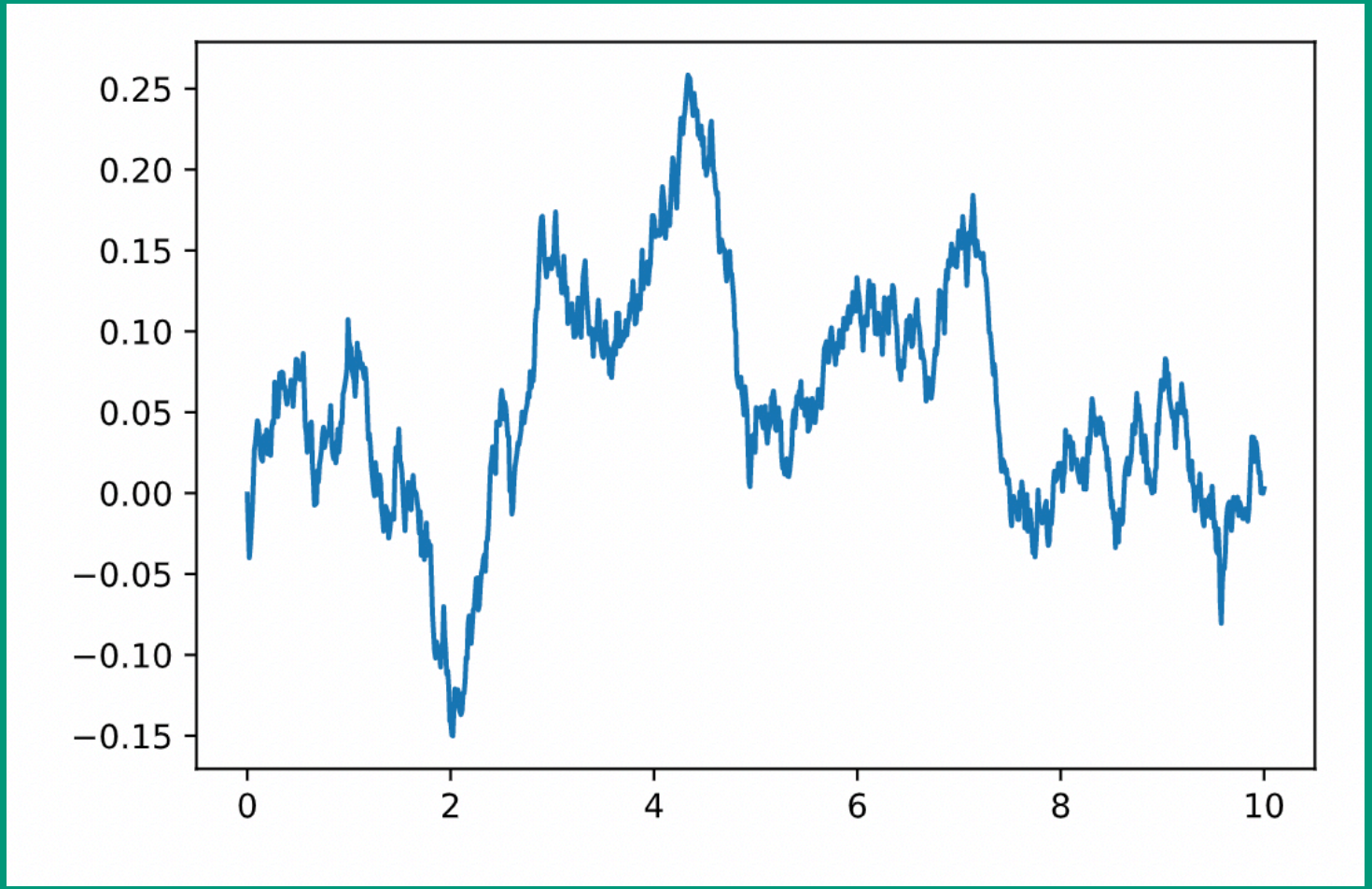
2D heatmap (density, or 2D historam)

	Dim 1	Dim 2
Nominal	✗	✗
Ordinal	✗	✗
Interval	✓	✓
Ratio	✓	✓



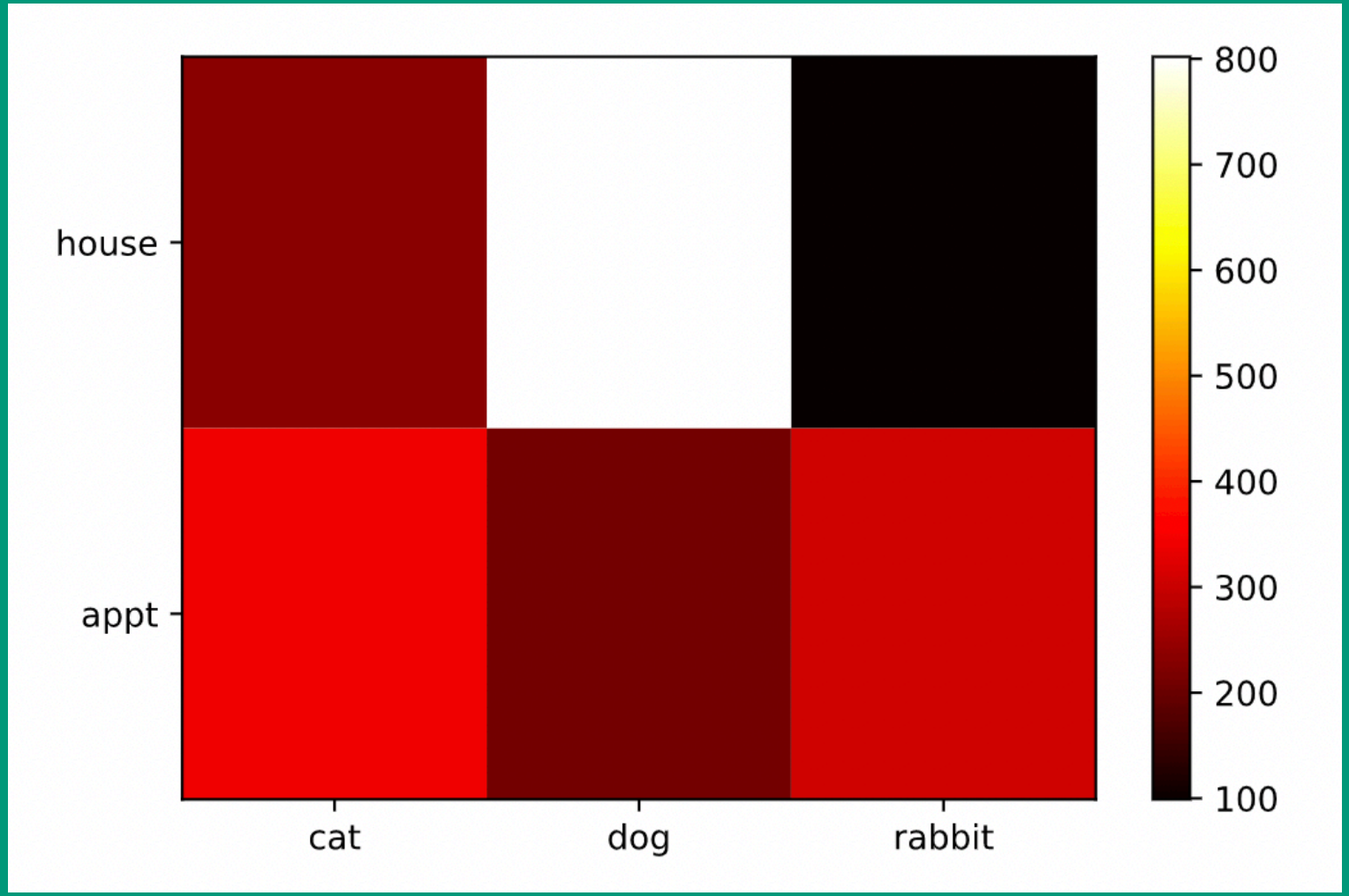
2D line plot

	Dim 1	Dim 2
Nominal	✗	✗
Ordinal	✗	✗
Interval	✓	✓
Ratio	✓	✓



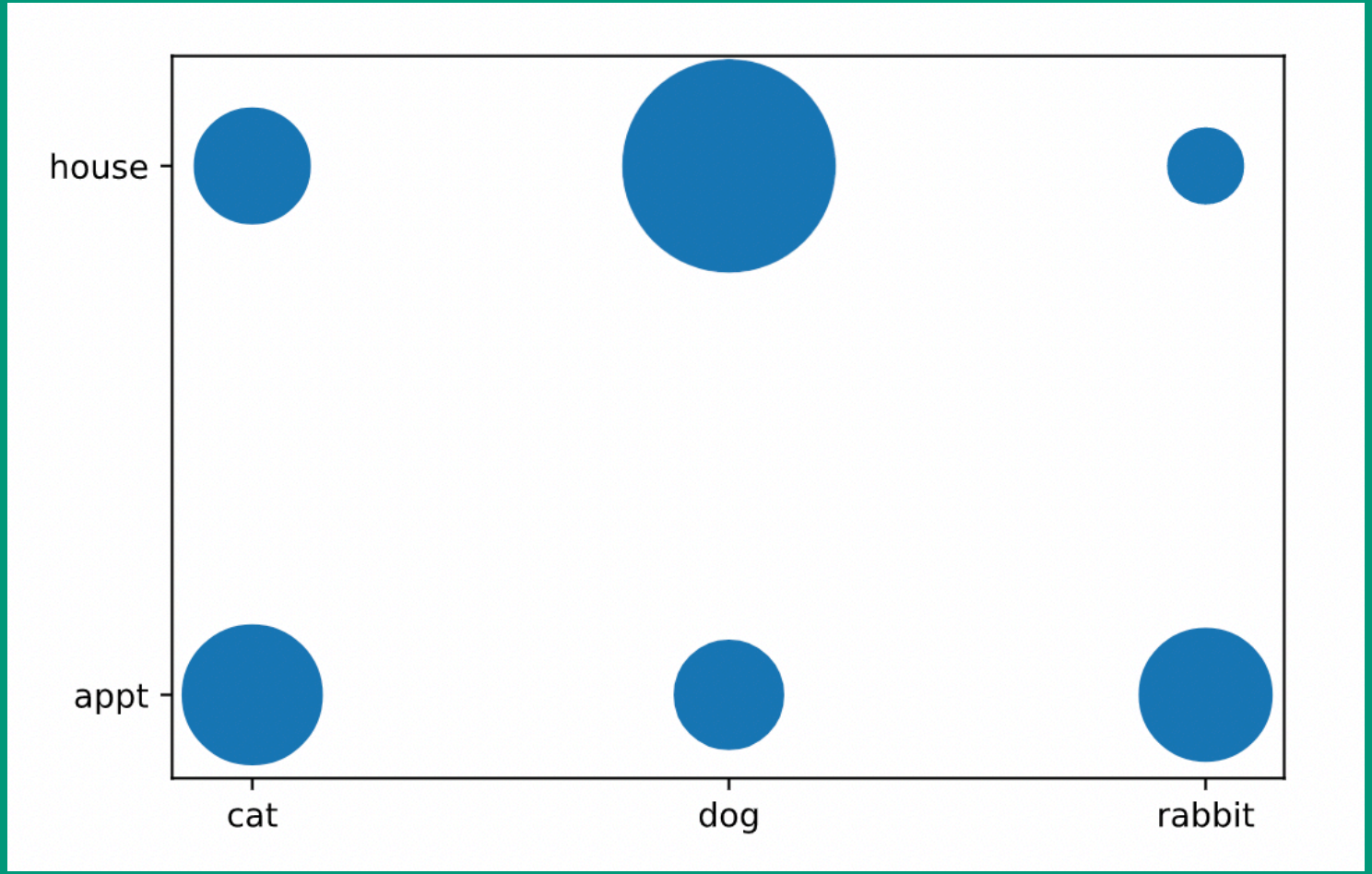
2D heatmap (matrix)

	Dim 1	Dim 2
Nominal	✓	✓
Ordinal	✓	✓
Interval	✗	✗
Ratio	✗	✗



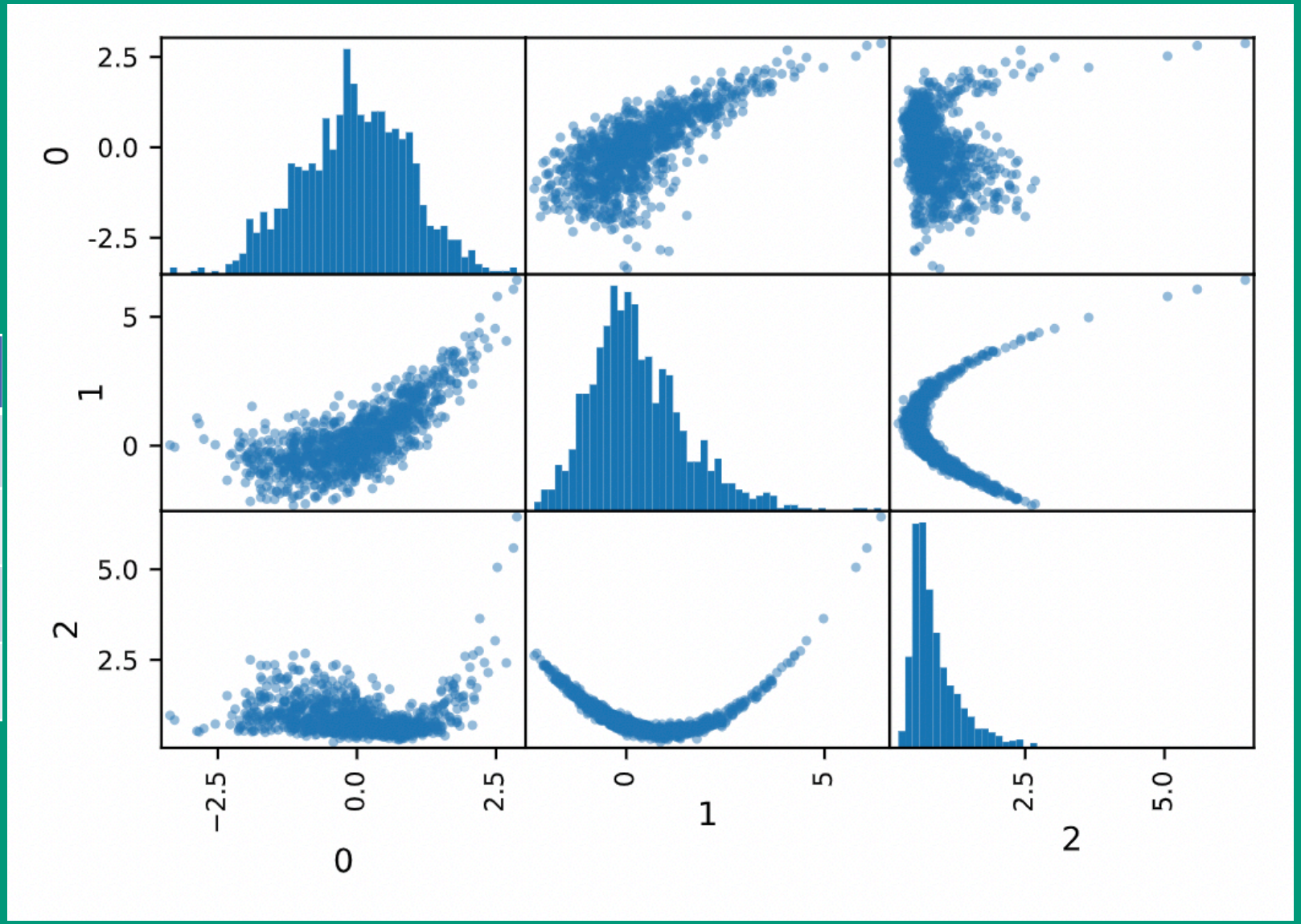
2D bubble plot

	Dim 1	Dim 2
Nominal	✓	✓
Ordinal	✓	✓
Interval	✗	✗
Ratio	✗	✗



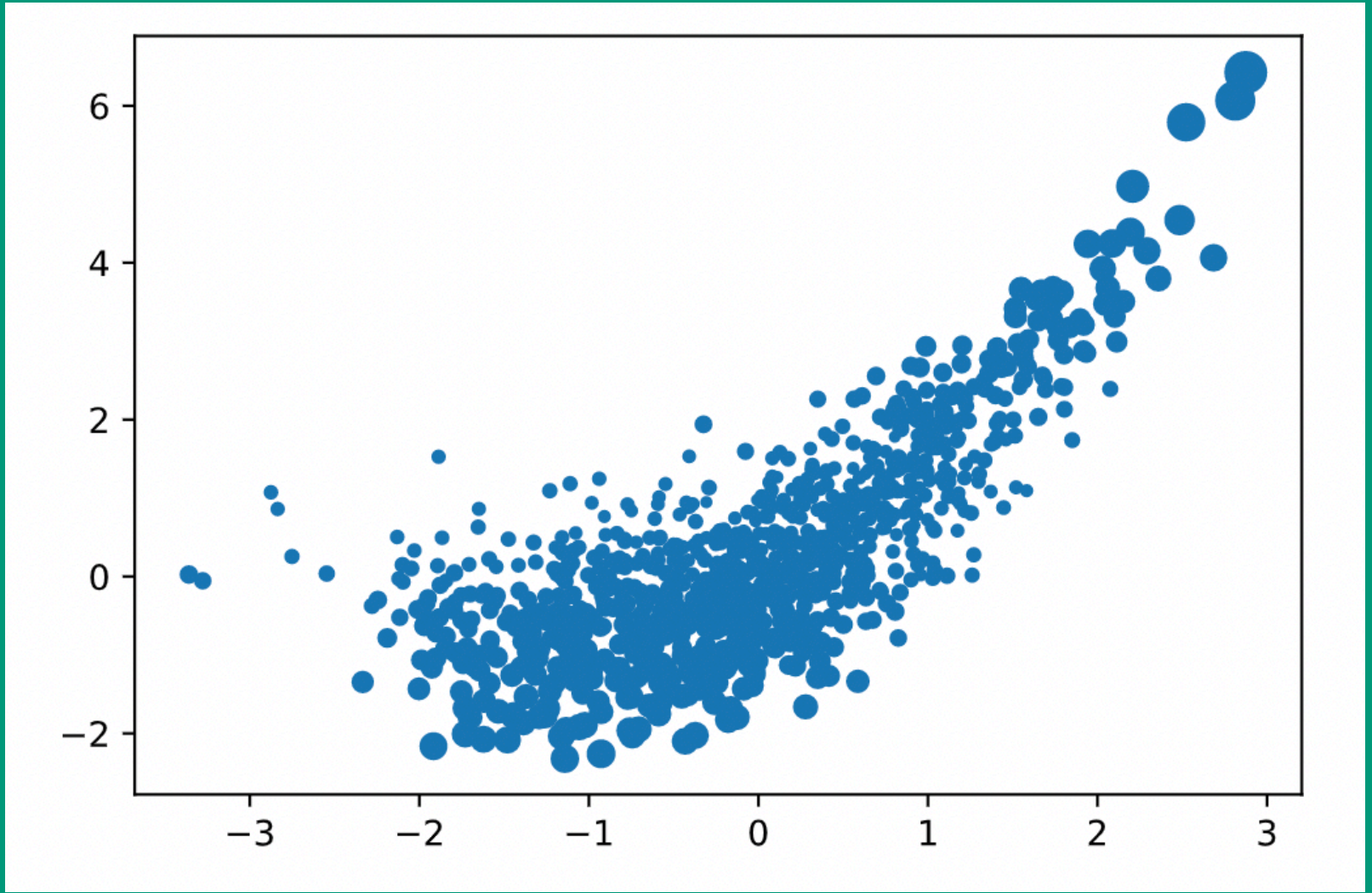
3D scatter matrix

	Dim 1	Dim 2	Dim 3
Nominal	✗	✗	✗
Ordinal	✗	✗	✗
Interval	✓	✓	✓
Ratio	✓	✓	✓



3D bubble plot

	Dim 1	Dim 2	Dim 3
Nominal	✗	✗	✗
Ordinal	✗	✗	✗
Interval	✓	✓	✓
Ratio	✓	✓	✓



Outline

- Basics of visualization
- Data types and visualization types
- Software plotting libraries

Matplotlib

Matplotlib is the standard for plotting in Python / Jupyter Notebook

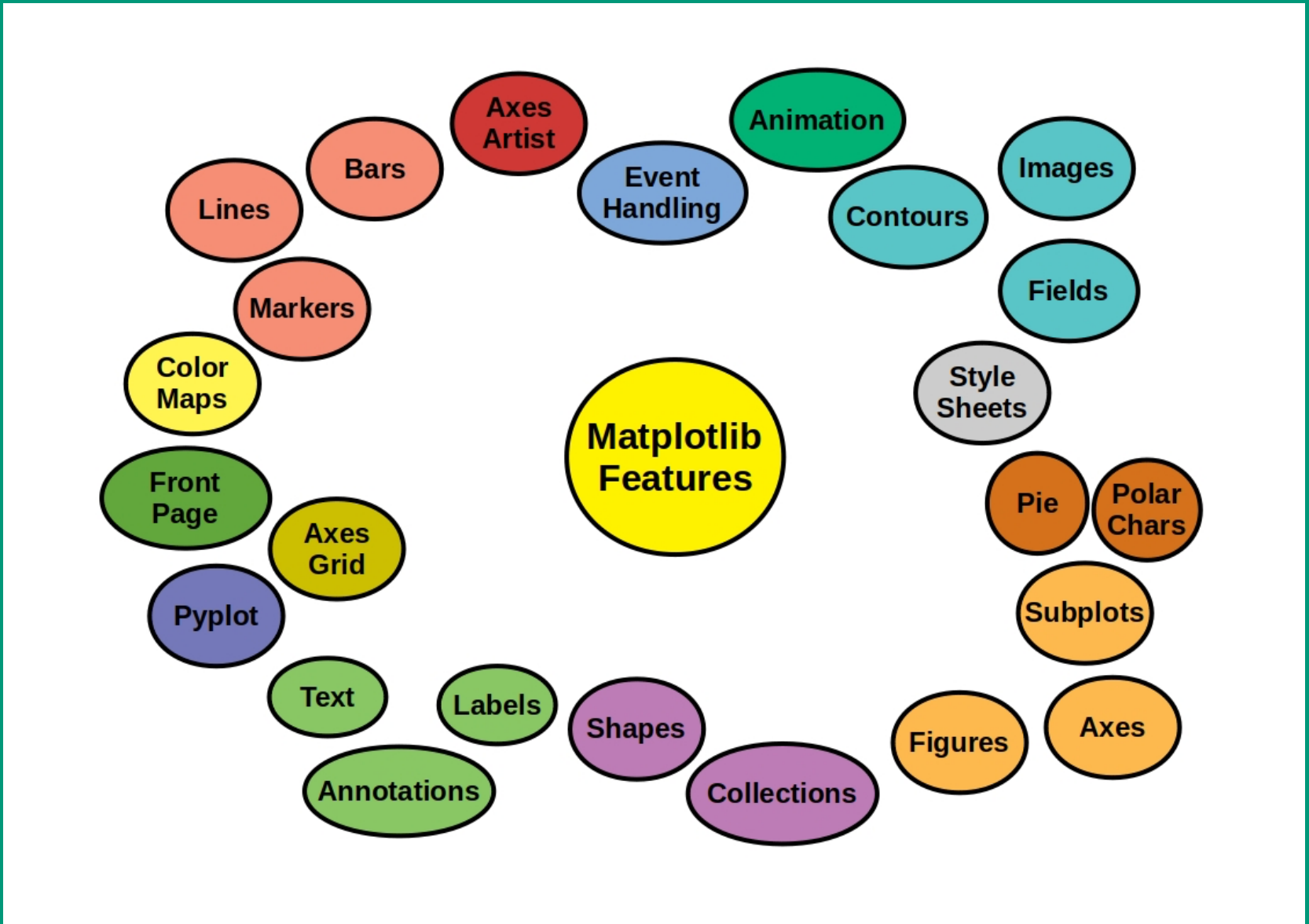
Matplotlib used to generate fairly ugly plots by default, but in recent versions this is no longer the case, so minimal need for additional libraries

It is aimed at generating static plots, not very good for interacting with data (with a few exceptions)

A number of additional libraries provide some level of interactive plot (and static plots), but matplotlib is enough of a standard that we'll use it here

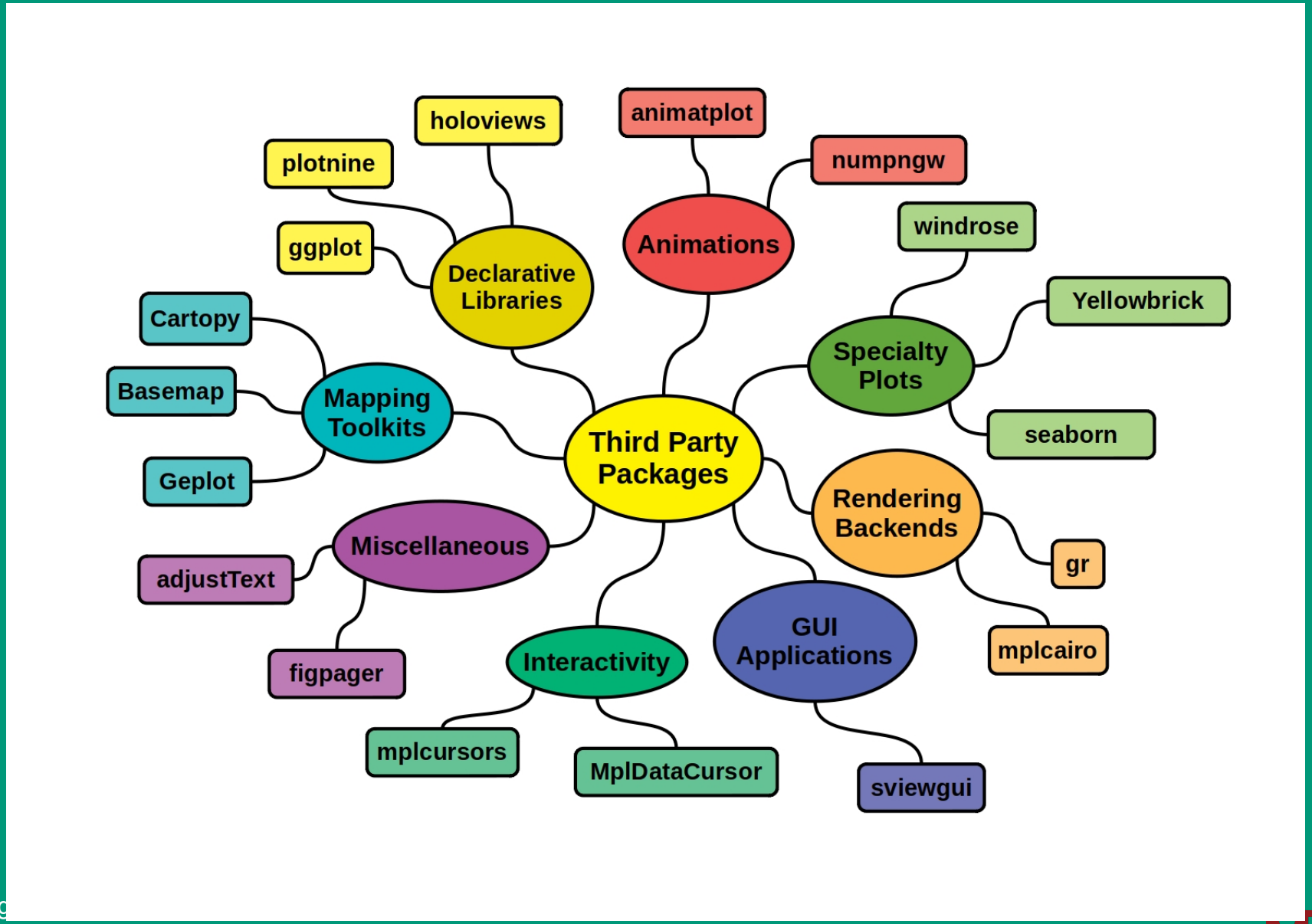
seeborn https://seaborn.pydata.org

Matplotlib



https://starship-knowledge.com/matplotlib-vs-seaborn

Matplotlib



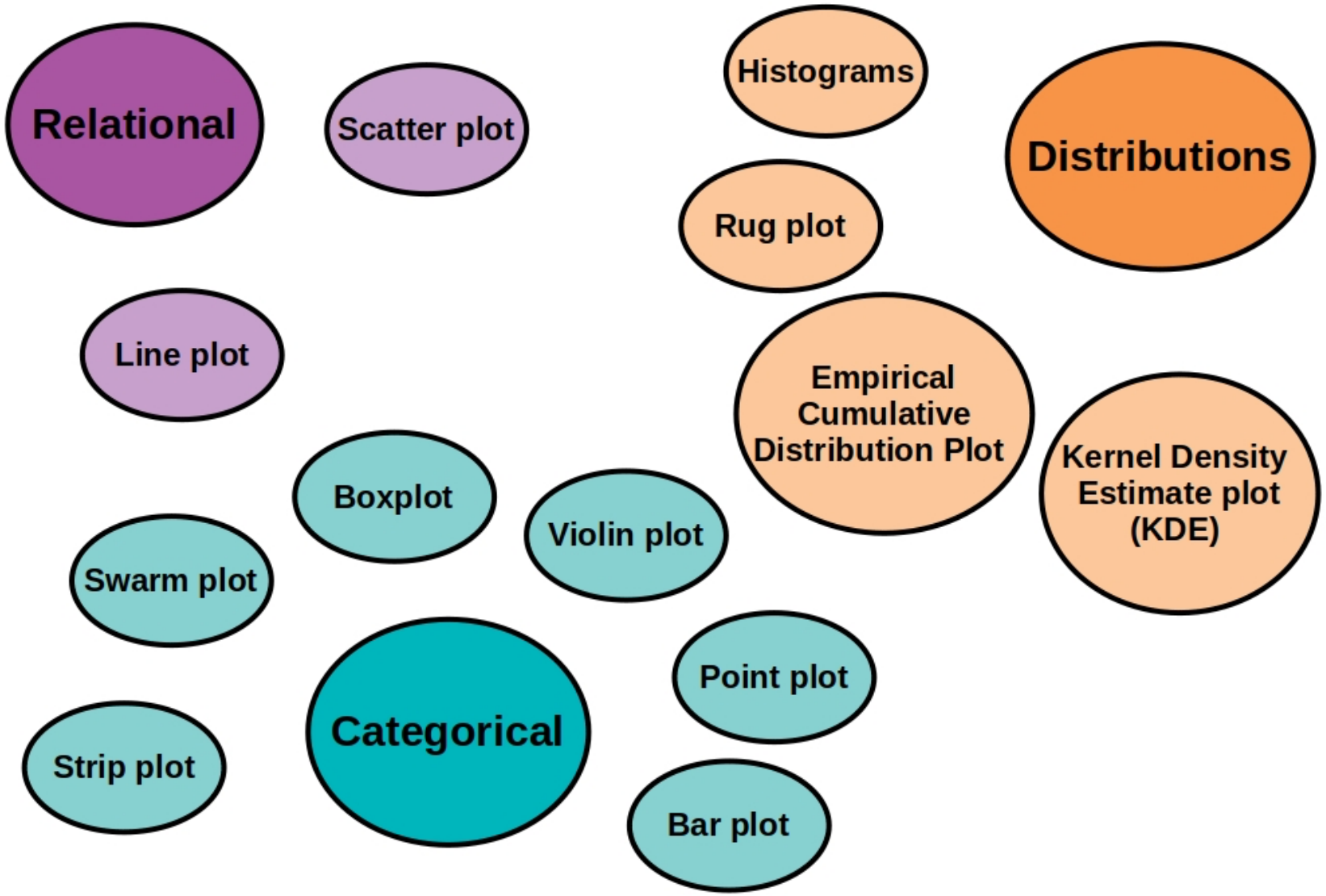
https://starship-knowledg

Seaborn

provides built-in themes for designing matplotlib graphs

linear regression models

optimization when processing NumPy and Pandas data structures.



https://starship-knowledge.com/matplotlib-v

Pandas

Pandas offer tools for cleaning and process your data. It is the most popular Python library that is used for data analysis. In pandas, a data table is called a dataframe.

```
# Python code demonstrate creating
import pandas as pd

# initialise data of lists.
data = {'Name':[ 'Mohe' , 'Karnal' , 'Yrik' , 'jack' ],
                                'Age':[ 30 , 21 , 29 , 28 ]}
# Create DataFrame
df = pd.DataFrame( data )

# Print the output.
df
```

	Name	Age
0	Mohe	30
1	Karnal	21
2	Yrik	29
3	jack	28

Example (1)

```
# import module
import pandas
# load the csv
data = pandas.read_csv("nba.csv")
# show first 5 column
data.head()
```

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0.0	PG	25.0	6-2	180.0	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99.0	SF	25.0	6-6	235.0	Marquette	6796117.0
2	John Holland	Boston Celtics	30.0	SG	27.0	6-5	205.0	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28.0	SG	22.0	6-5	185.0	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8.0	PF	29.0	6-10	231.0	NaN	5000000.0

Seaborn

Seaborn is an amazing visualization library for statistical graphics plotting in Python. It is built on the top of matplotlib library and also closely integrated into the data structures from pandas.

pip install seaborn

 # for more examples check out https://seaborn.pydata.org/tutorial/introduction

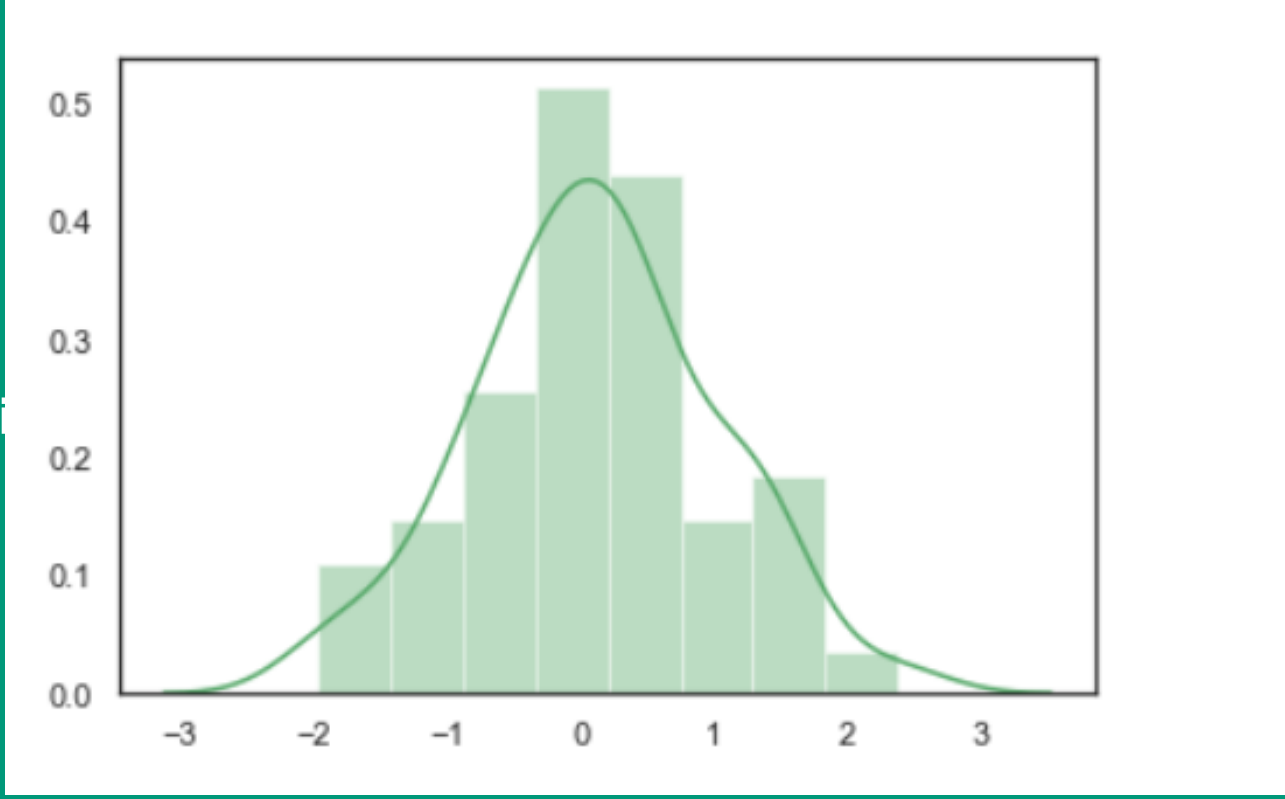
Seaborn

```
# Importing libraries
import numpy as np
import seaborn as sns

# Selecting style as white, dark, whitegrid, darkgri
sns.set( style = "white" )

# Generate a random univariate dataset
rs = np.random.RandomState( 10 )
d = rs.normal( size = 50 )

# Plot a simple histogram and kde with binsize determined automatically
sns.distplot(d, kde = True, color = "g")
```



Seaborn: statistical data visualization

Line Plot

Scatter Plot

Box plot

Bar plot

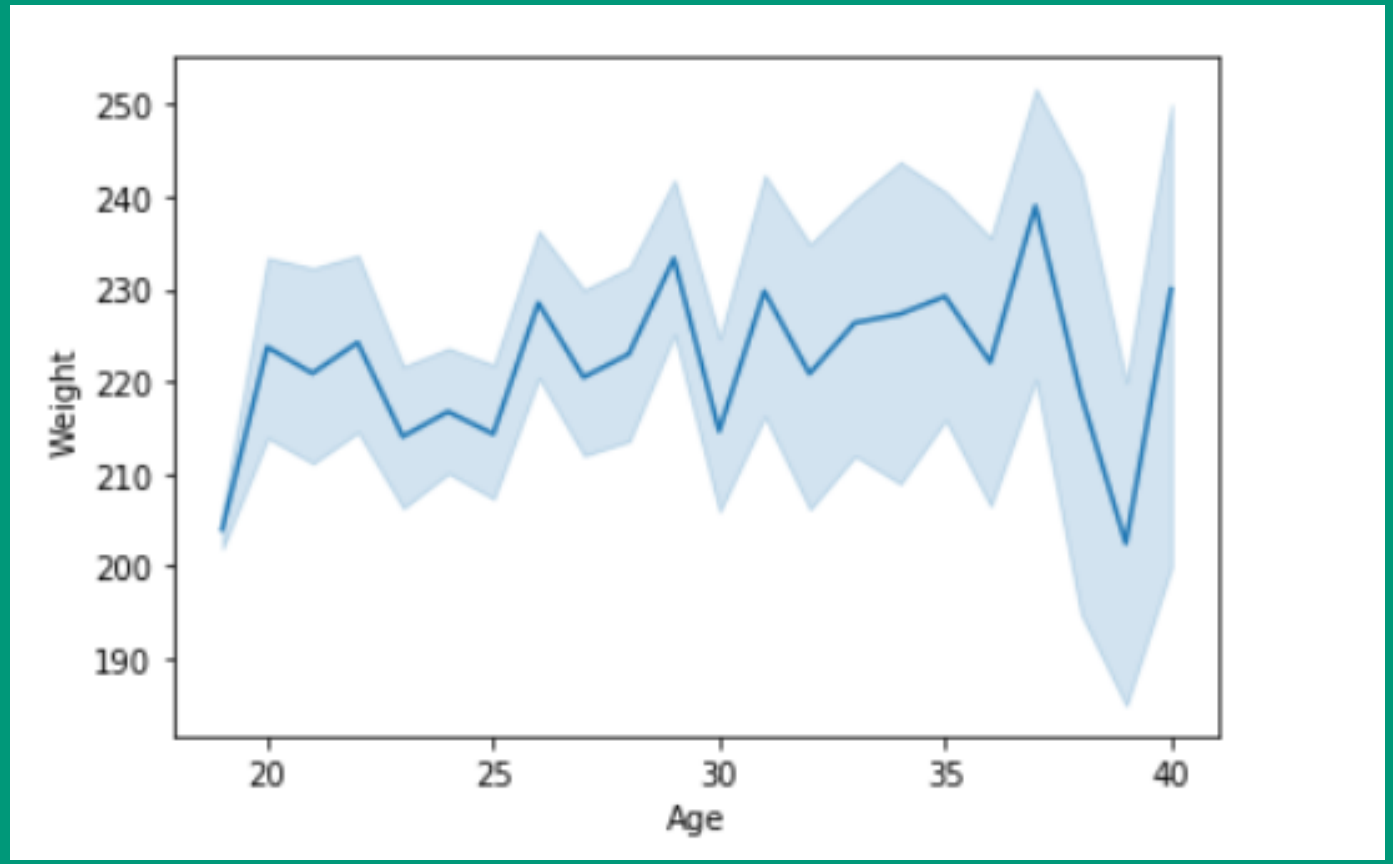
…

Seaborn Line plot

```
# import module
import seaborn as sns
import pandas

# loading csv
data = pandas.read_csv("nba.csv")

# plotting lineplot
sns.lineplot( data['Age'], data['Weight'])
```




```
# check out the parameter hue =data["Position"]
```

#hue : *(optional) This parameter take column name for colour encoding.*

Seaborn scatter plot

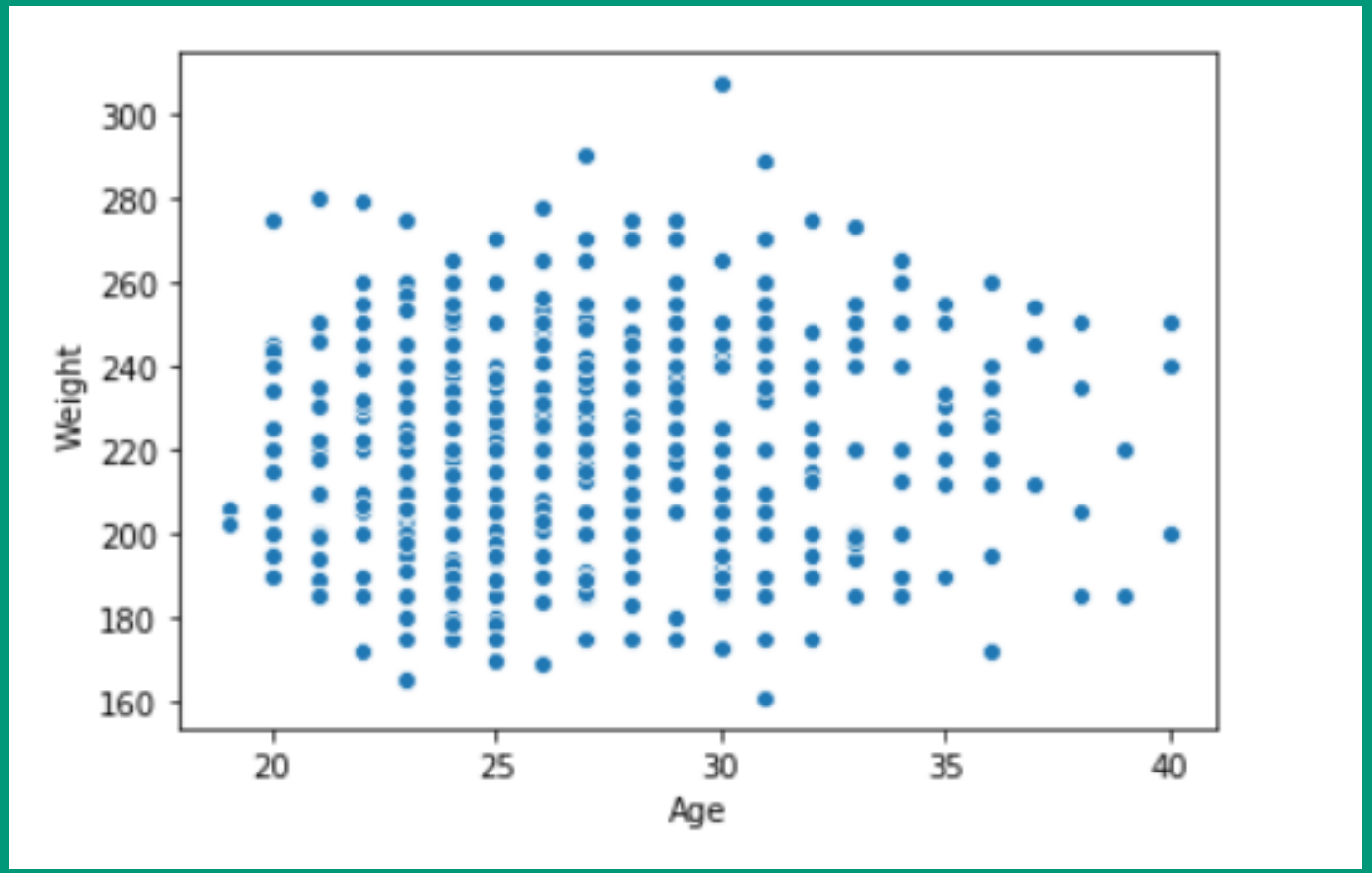
```
# import module
import seaborn
import pandas

# load csv
data = pandas.read_csv("nba.csv")

# plotting
seaborn.scatterplot(data['Age'],data['Weight’])

# check out the parameter hue =data["Position"]
```

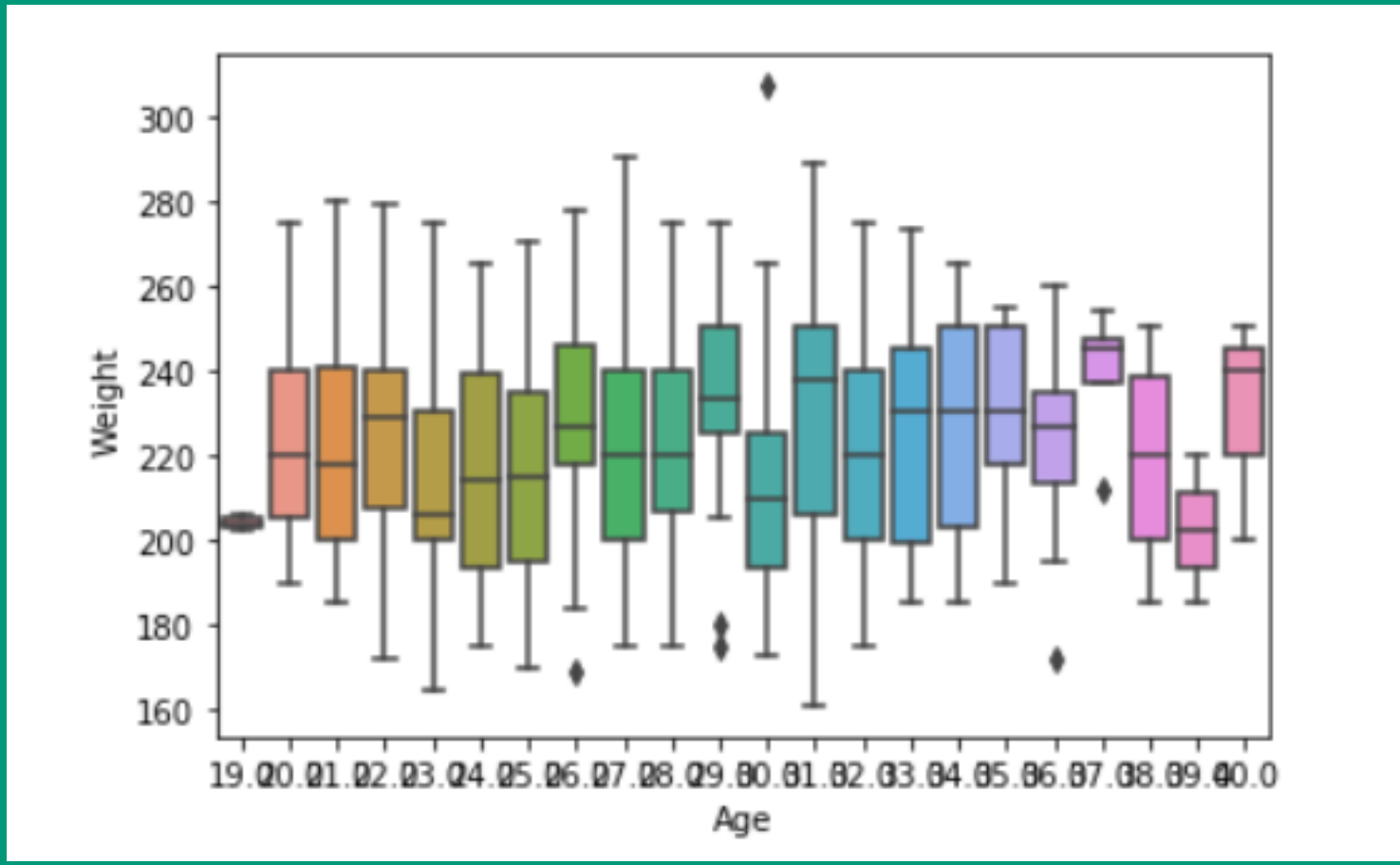
#hue : *(optional) This parameter take column name for colour encoding.*



Seaborn Box plot

```
# import module
import seaborn as sns
import pandas

# read csv and plotting
data = pandas.read_csv( "nba.csv" )
sns.boxplot( data['Age'], data['Weight'])
```

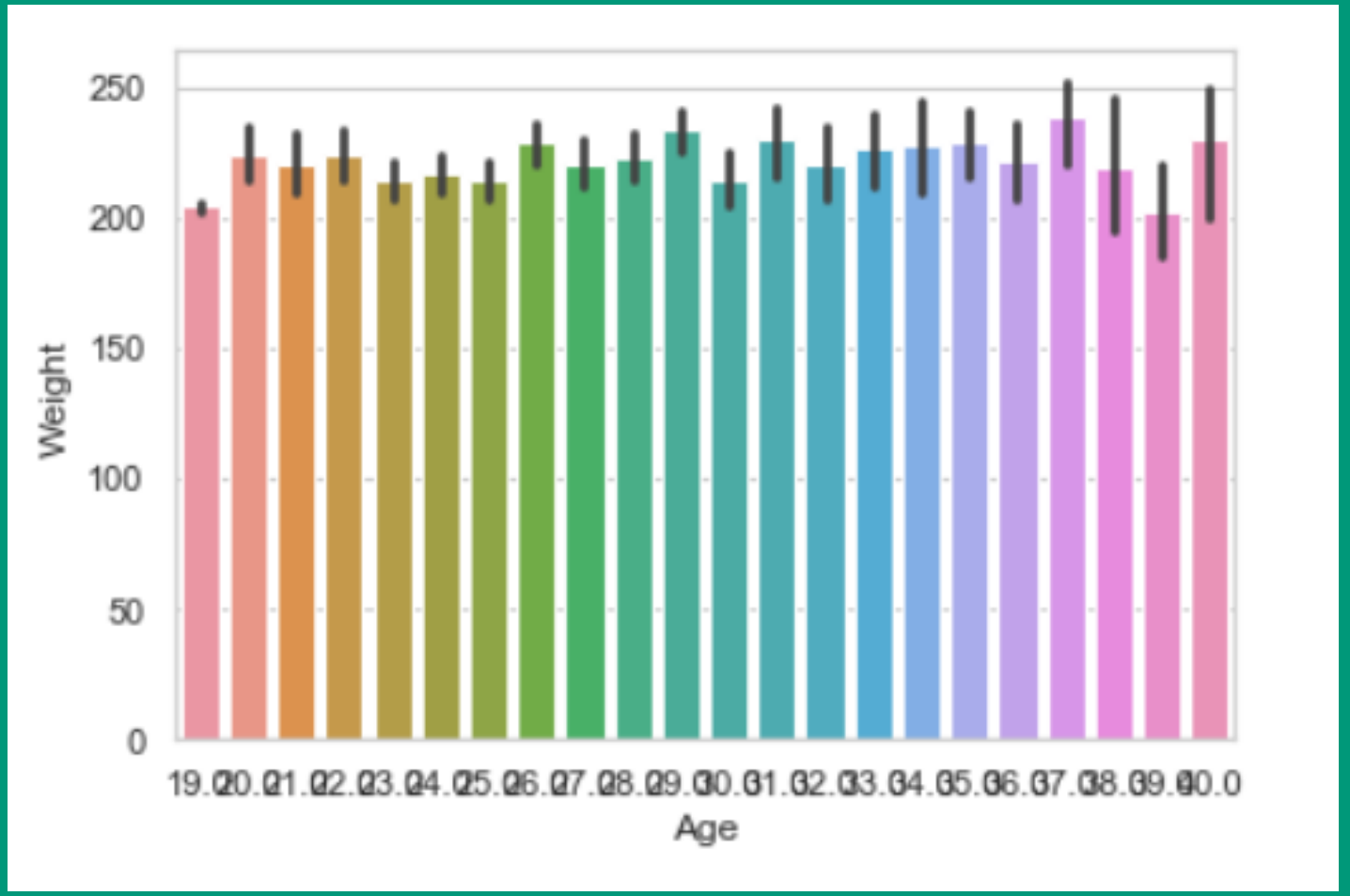


Seaborn Bar plot

```
# import module
import seaborn


seaborn.set(style = 'whitegrid')


# read csv and plot
data = pandas.read_csv("nba.csv")
seaborn.barplot(x ="Age", y ="Weight", data = data)
```



Thank you!

Questions?

Contact

Middle East University

Amman-Jordan- Zip-Code
(Postal Address): (11831)
Phone : +962 6 4790222
Fax : +962 6 4129613
Mobile: +962 79 7122000
Mobile: +962 79 9969933
https://www.meu.edu.jo

Dr. Jamal Al Qundus
Phone +962 (-)
email: jalqundus@meu.edu.jo
Office: B 218-I