KAMPUS MERDEKA
X
MY EDUSOLVE

Python Week #9 Data Analysis

Data Science - Team 3

Never Forget to import warnings

Petunjuk: Sorot teks, klik simbol tautan pada bilah alat, lalu pilih
More information can be found on the website Seaborn.com

Outline

Seaborn
- import
- load dataset
- create visualization
- plot(hist,bar,scatter,etc)

Statistic Descriptive
- Measure Central Tendency
- Measure of Dispersion

1 Seaborn

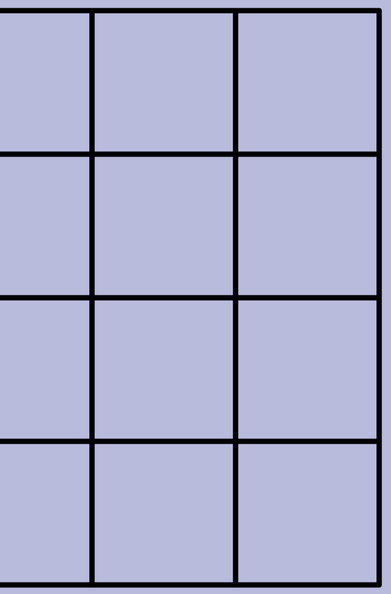
2 Statistic Descriptive

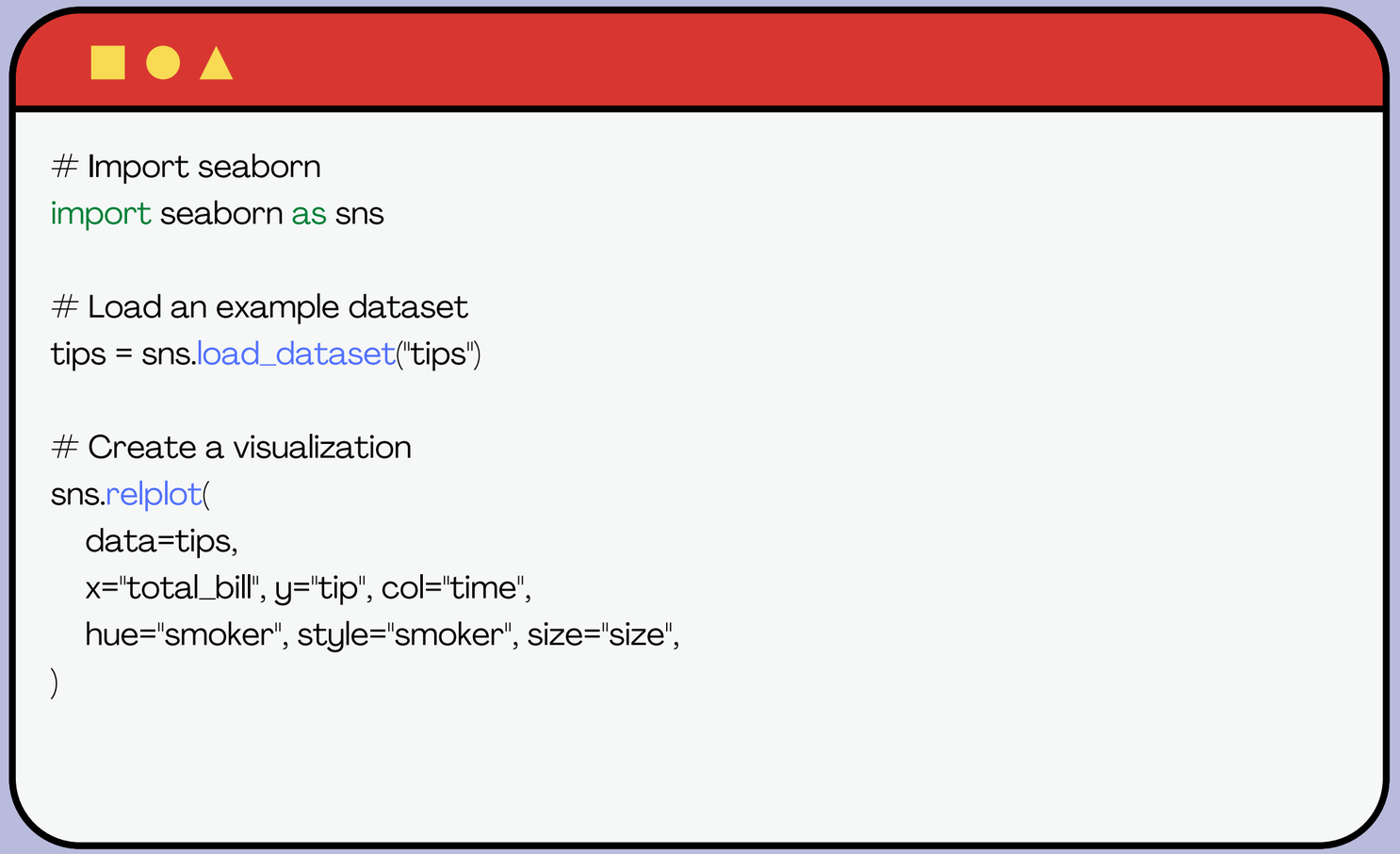
Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

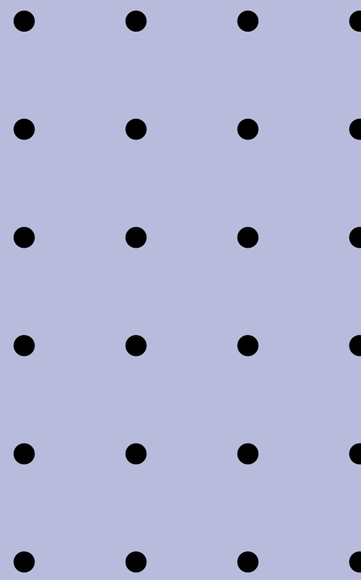
IMPORT,LOAD DATASET, CREATE VISUALIZATION

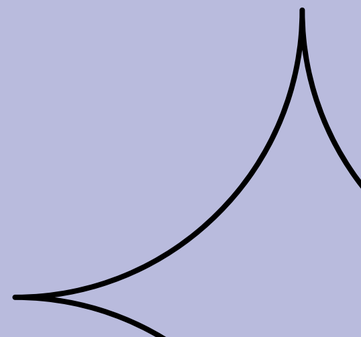
SEABORN HELPS YOU EXPLORE AND UNDERSTAND YOUR DATA. ITS PLOTTING FUNCTIONS OPERATE ON DATAFRAMES AND ARRAYS CONTAINING WHOLE DATASETS AND INTERNALLY PERFORM THE NECESSARY SEMANTIC MAPPING AND STATISTICAL AGGREGATION TO PRODUCE INFORMATIVE PLOTS. ITS DATASET-ORIENTED, DECLARATIVE API LETS YOU FOCUS ON WHAT THE DIFFERENT ELEMENTS OF YOUR PLOTS MEAN, RATHER THAN ON THE DETAILS OF HOW TO DRAW THEM.





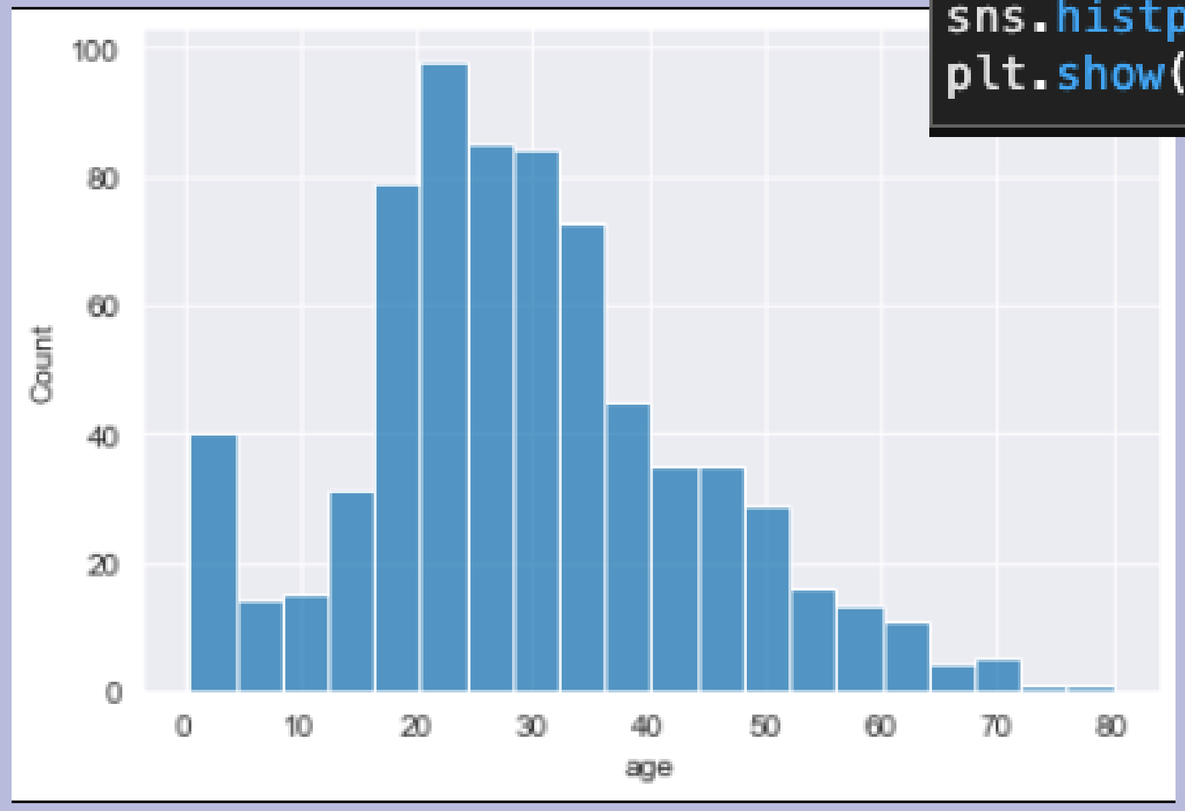
```
# Import seaborn
import seaborn as sns

# Load an example dataset
tips = sns.load_dataset("tips")

# Create a visualization
sns.relplot(
    data=tips,
    x="total_bill", y="tip", col="time",
    hue="smoker", style="smoker", size="size",
)
```





THERE IS NO UNIVERSALLY BEST WAY TO VISUALIZE DATA. DIFFERENT QUESTIONS ARE BEST ANSWERED BY DIFFERENT PLOTS. SEABORN MAKES IT EASY TO SWITCH BETWEEN DIFFERENT VISUAL REPRESENTATIONS BY USING A CONSISTENT DATASET-ORIENTED API.

Seaborn.histplot




```
# contoh hishplot
sns.histplot(titanic['age'])
plt.show()
```

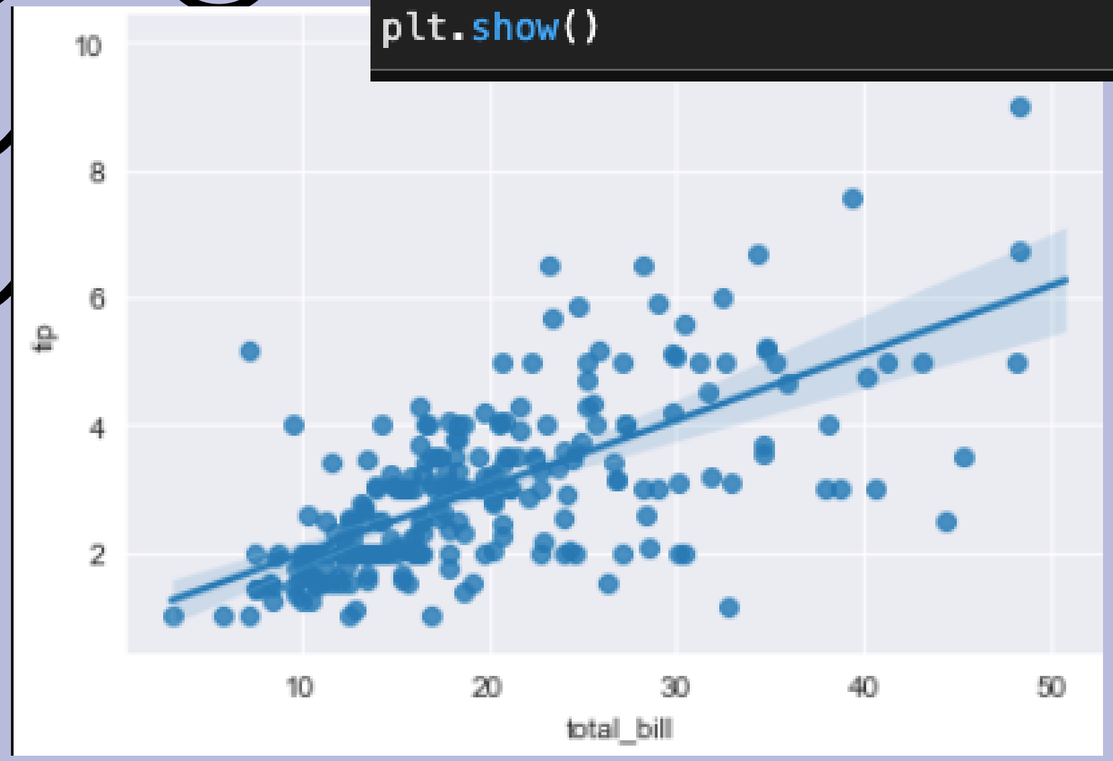
Plot univariate or bivariate histograms to show distributions of datasets.

A histogram is a classic visualization tool that represents the distribution of one or more variables by counting the number of observations that fall within discrete bins.

seaborn.scatterplot

SEABORN HAS SEVERAL PLOTS THAT CAN BE USED TO VISUALIZE THE DISTRIBUTION OF DATA, NAMELY SCATTERPLOT, REGPLOT AND IMPLOT. THIS PLOT HAS A PALETTE, SOME OF WHICH ARE;

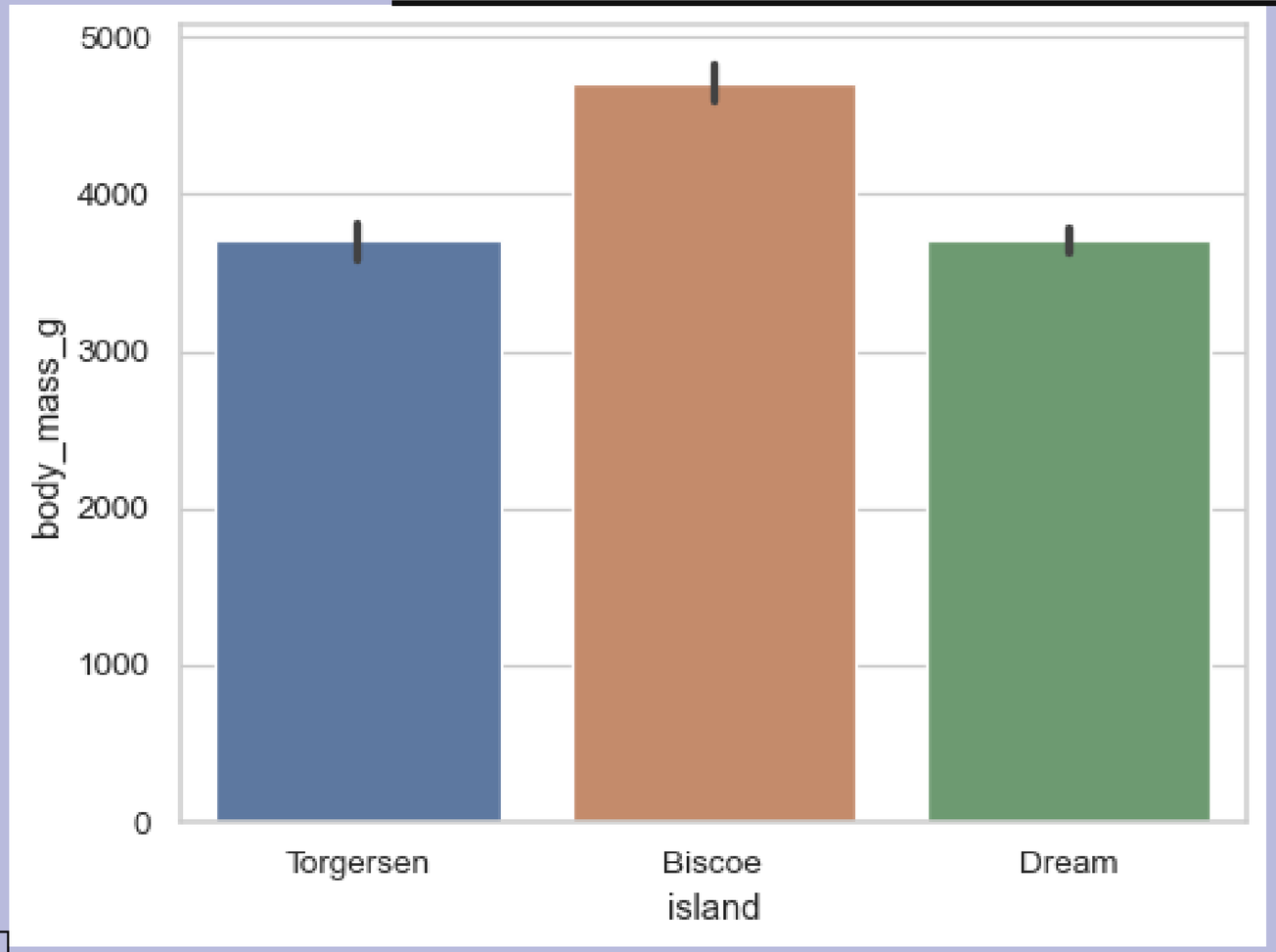
```
# contoh scatterplot
plt.figure(figsize=(8,5))
sns.scatterplot(data=tips, x ='total_bill', y='tip',
                hue = "day", palette = "Set1", style='smoker')
plt.title("judul")
plt.show()
```



The relationship between x and y can be shown for different subsets of the data using the hue, size, and style parameters. These parameters control what visual semantics are used to identify the different subsets. It is possible to show up to three dimensions independently by using all three semantic types, but this style of plot can be hard to interpret and is often ineffective. Using redundant semantics (i.e. both hue and style for the same variable) can be helpful for making graphics more accessible.

seaborn.barplot

```
df = sns.load_dataset("penguins")
sns.barplot(data=df, x="island", y="body_mass_g")
plt.show()
```

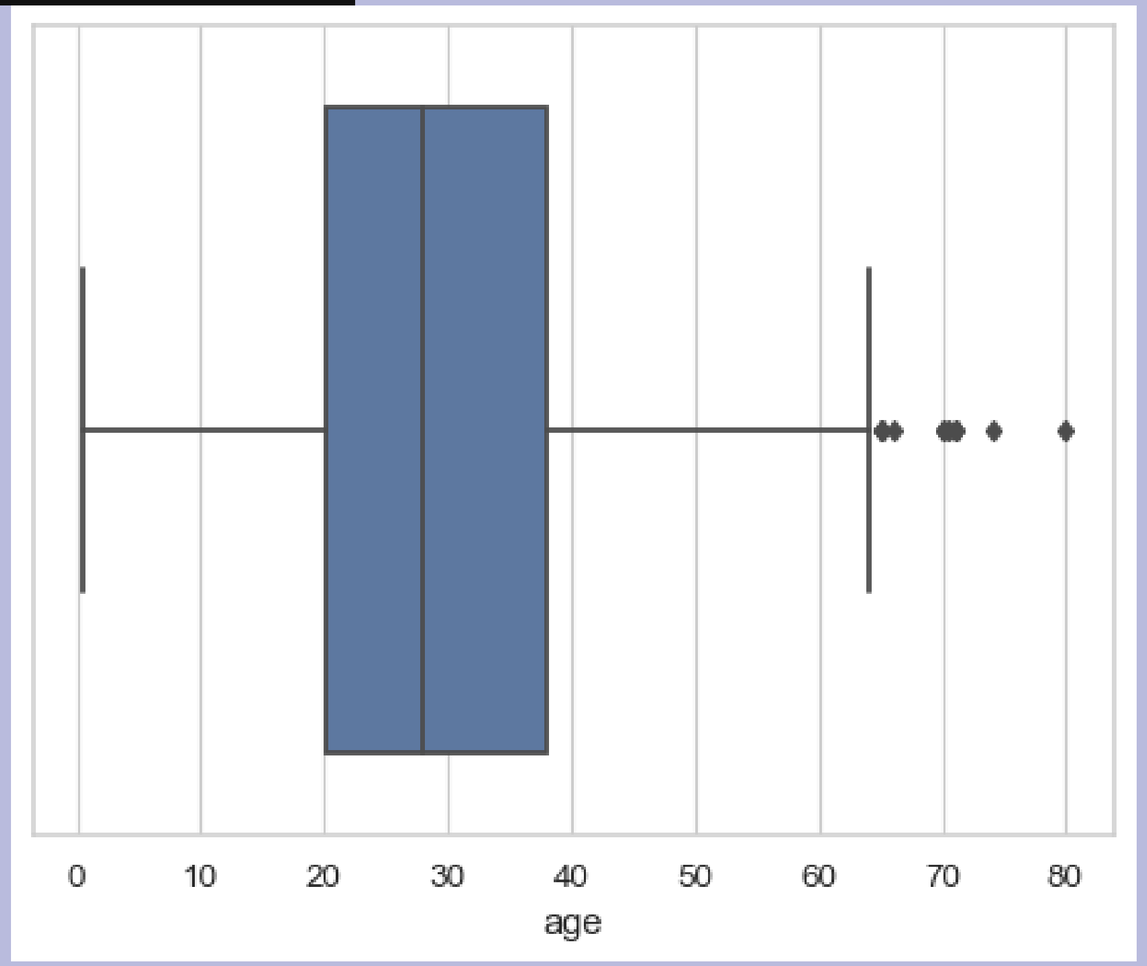


A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle and provides some indication of the uncertainty around that estimate using error bars. Bar plots include 0 in the quantitative axis range, and they are a good choice when 0 is a meaningful value for the quantitative variable, and you want to make comparisons against it.

seaborn.boxplot

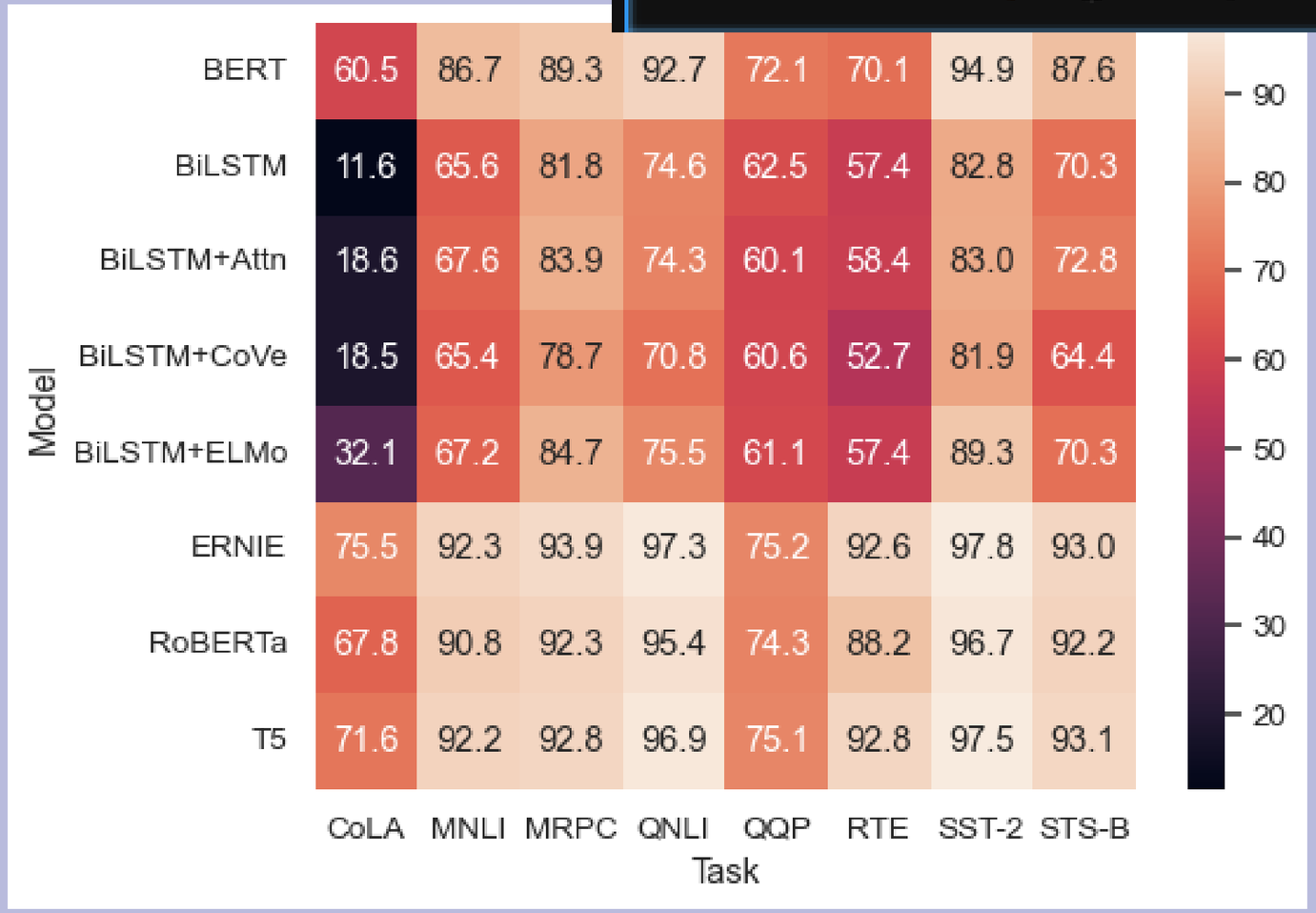
```
df = sns.load_dataset("titanic")
sns.boxplot(x=df["age"])
```

A box plot (or box-and-whisker plot) shows the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable. The box shows the quartiles of the dataset while the whiskers extend to show the rest of the distribution, except for points that are determined to be “outliers” using a method that is a function of the inter-quartile range.



seaborn.heatmap

```
sns.heatmap(glue, annot=True, fmt=".1f")
```



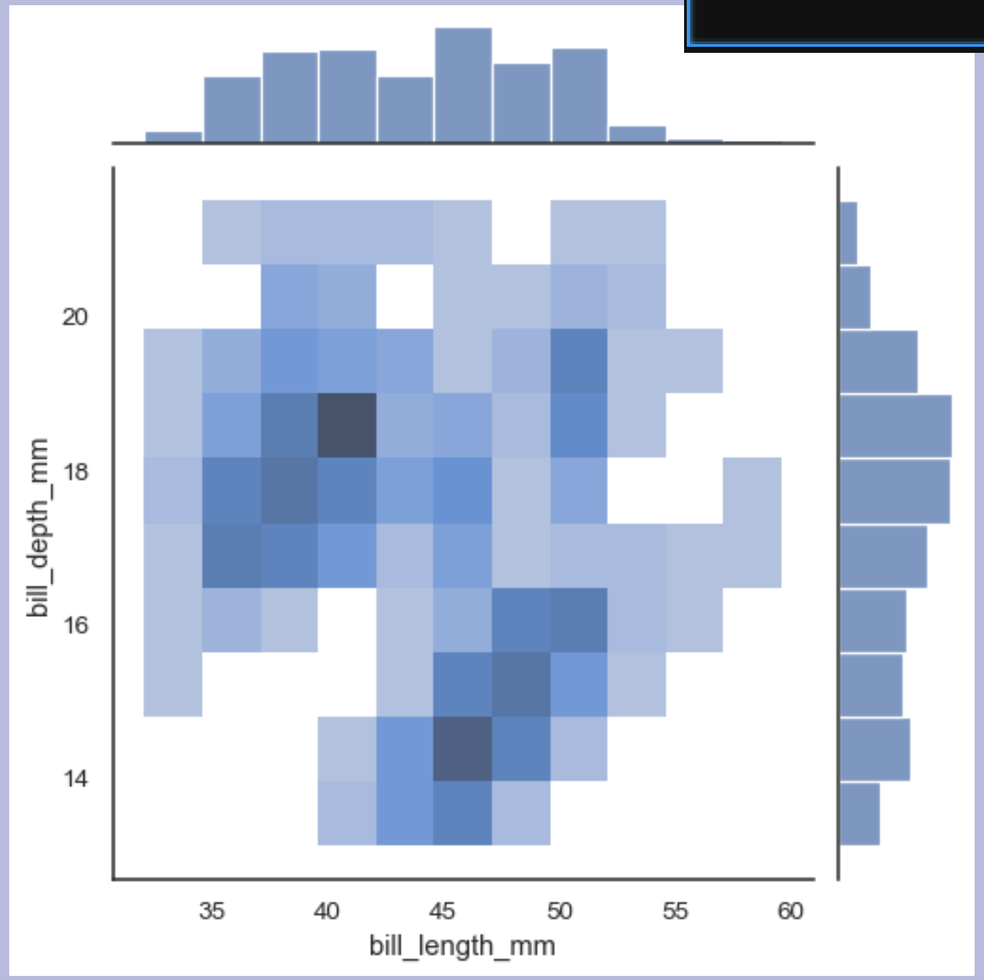
This is an Axes-level function and will draw the heatmap into the currently-active Axes if none is provided to the ax argument. Part of this Axes space will be taken and used to plot a colormap, unless cbar is False or a separate Axes is provided to cbar_ax.

annot : bool or rectangular dataset, optional

If True, write the data value in each cell. If an array-like with the same shape as data, then use this to annotate the heatmap instead of the data. Note that DataFrames will match on position, not index.

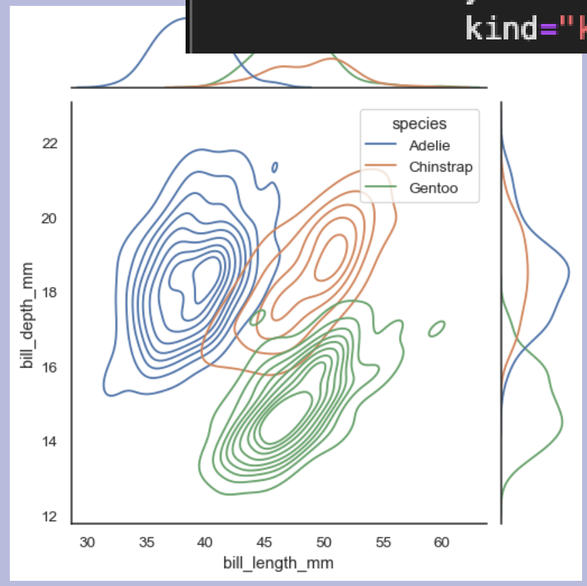
seaborn.jointplot

```
sns.jointplot(data=penguins, x="bill_length_mm",
              y="bill_depth_mm",
              kind="hist")
```



This function provides a convenient interface to the JointGrid class, with several canned plot kinds. This is intended to be a fairly lightweight wrapper; if you need more flexibility, you should use JointGrid directly.

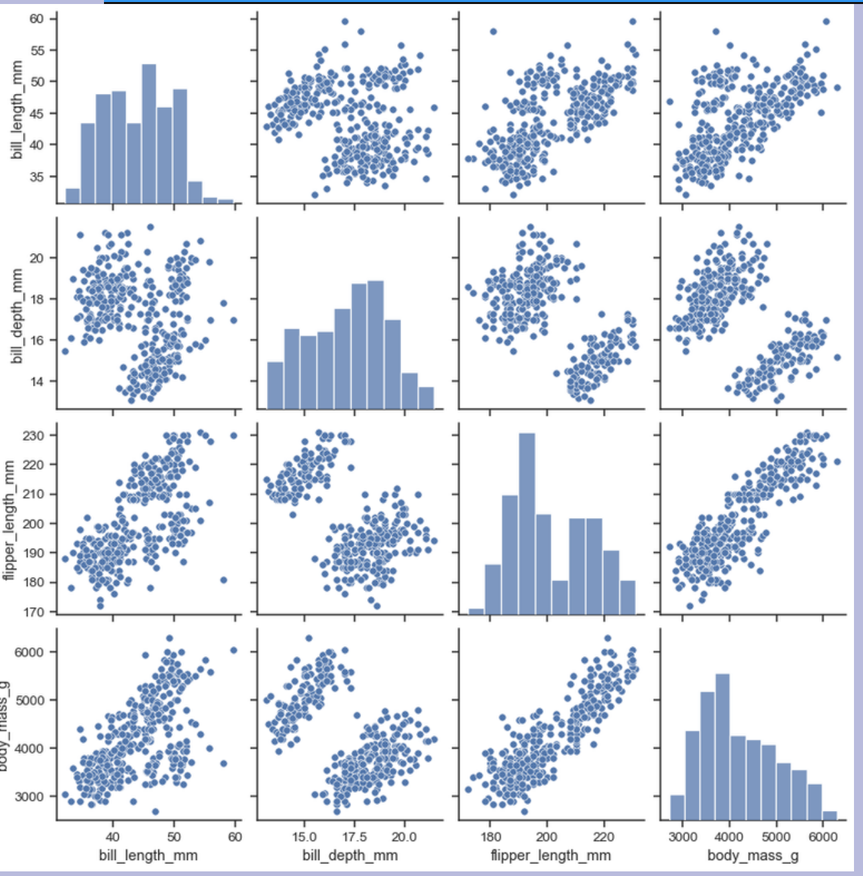
```
sns.jointplot(data=penguins, x="bill_length_mm",
              y="bill_depth_mm", hue="species",
              kind="kde")
```



Several different approaches to plotting are available through the kind parameter. Setting kind="kde" will draw both bivariate and univariate KDEs:

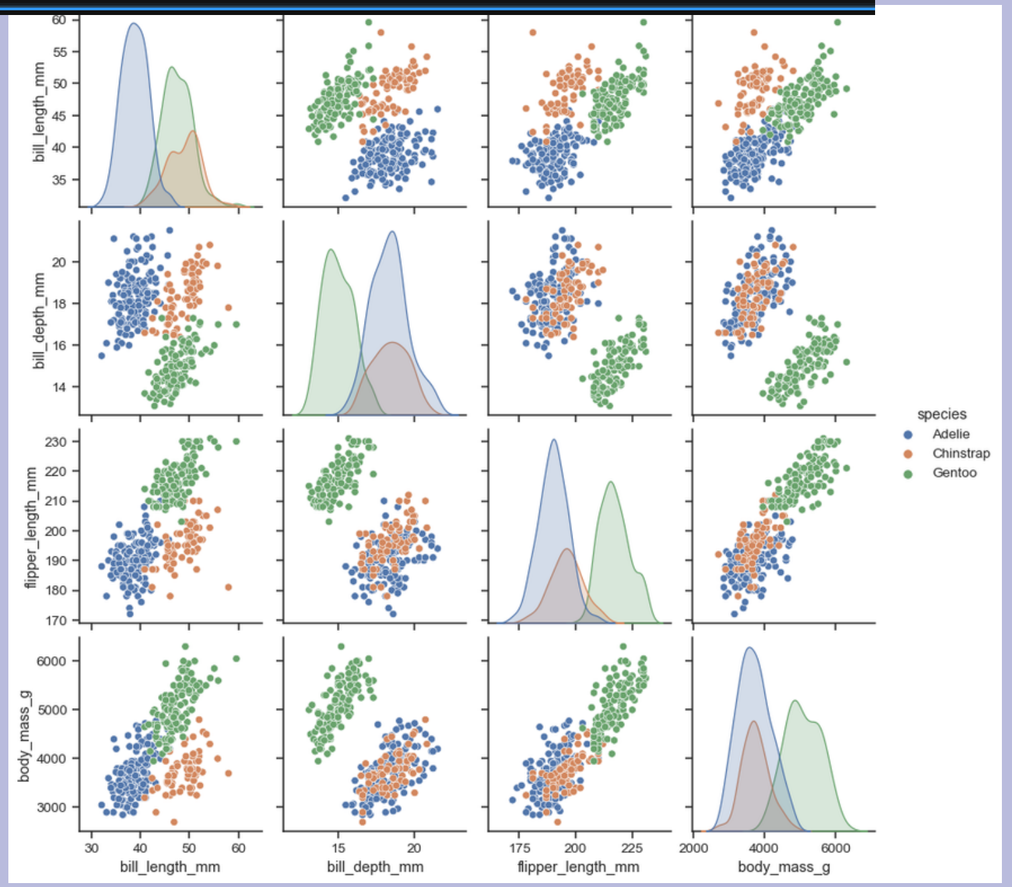
seaborn.pairplot

```
penguins = sns.load_dataset("penguins")
sns.pairplot(penguins)
```



By default, this function will create a grid of Axes such that each numeric variable in data will by shared across the y-axes across a single row and the x-axes across a single column. The diagonal plots are treated differently: a univariate distribution plot is drawn to show the marginal distribution of the data in each column.

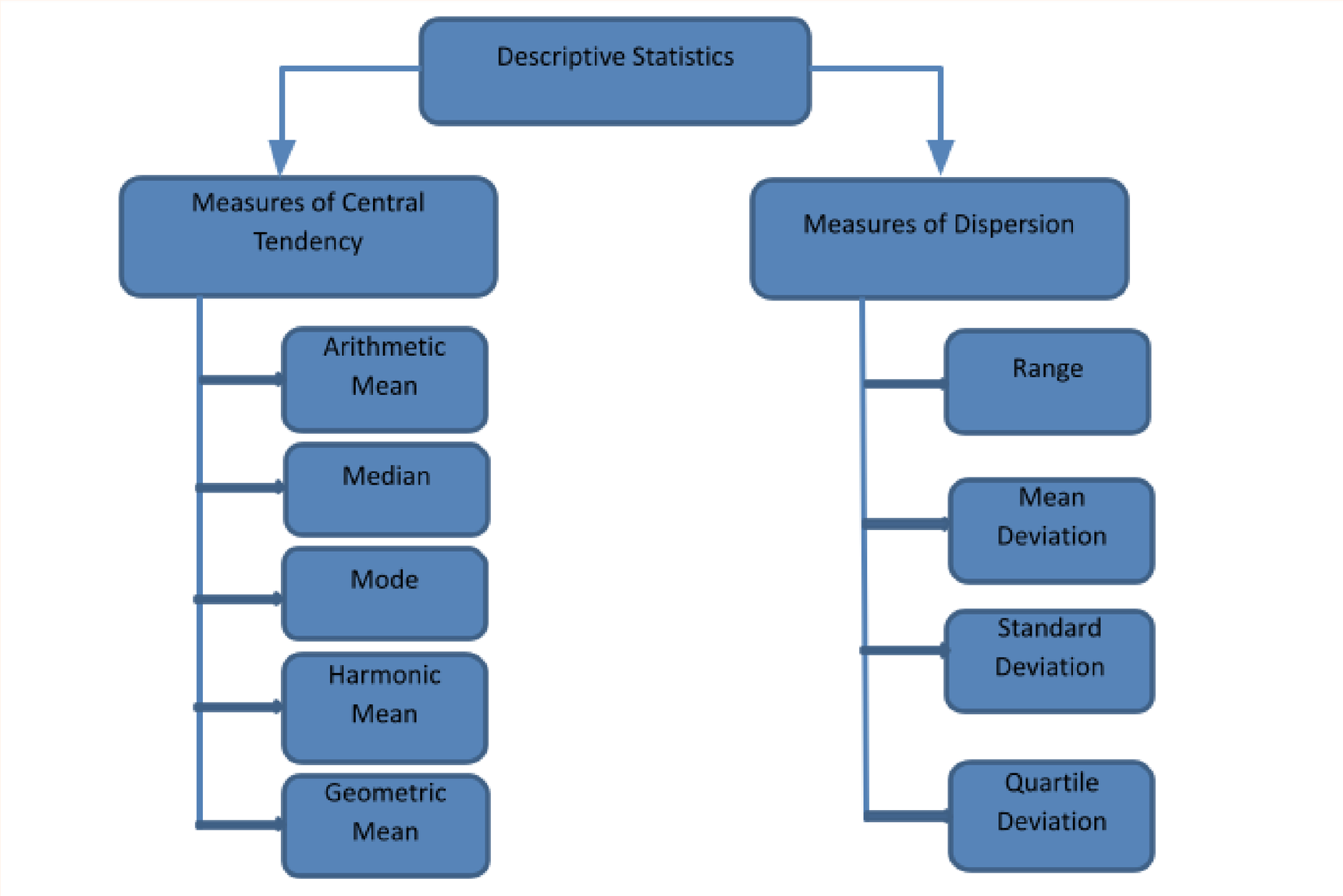
```
sns.pairplot(penguins, hue="species")
```



Descriptive Statistic

Descriptive Statistics is a statistical analysis process that focuses on the management, presentation and classification of data, with this process the power presented is more attractive, easier to understand, and able to give more meaning to data users, it can be said that this is a basic analysis that must be mastered by everyone who working with data

Examples that we usually do in this type of statistics usually include data analysis, graphing, and calculating various types of data measurements



Measure Of Central Tendency

- Measure of central tendency is the most common method used in descriptive power analysis. This method focuses on describing the condition of the data at the center point.

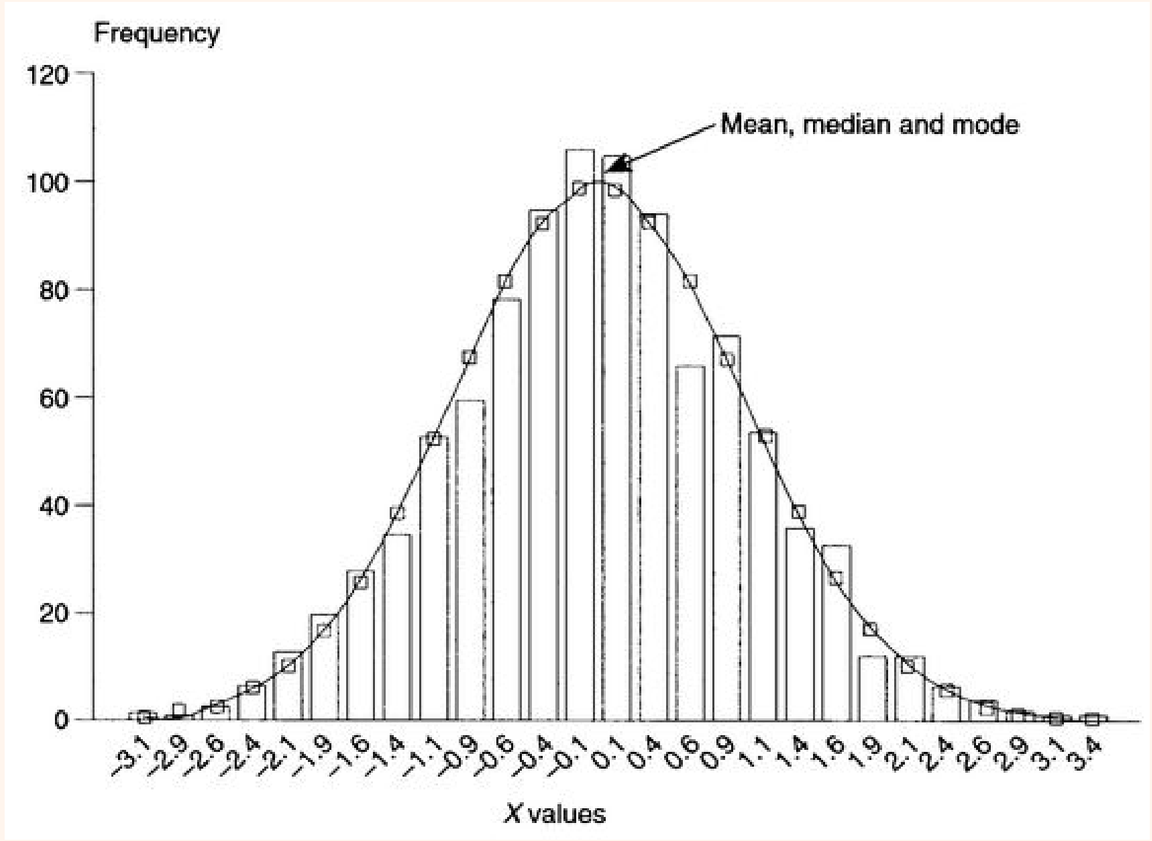
- In general, we can see how the condition of the data by seeing where the data center is located. Usually the data center itself will be at the middle value. although it's not always the case

- To prove this mathematically, the measurements that are often used are the mean, median, and mode

- The **mean** is the average of a set of data that we have. The formula is very simple. we only need to add up the values of all the data we have and divide by the number of data.

- The **median** is the middle value of a data. when we have a set of data. we can sort the data from smallest to largest value. if we have an odd number of data. then the median value of the data will immediately become the median. but if we have even data. million need to find the mean of the mean of the data.

- **Mode** is the value that occurs most frequently in a data set. we just need to see which value occurs most frequently in the group. when the number of frequencies. every data is the same, then the mode value does not exist



MEASURE OF DISPERSION / SPREAD

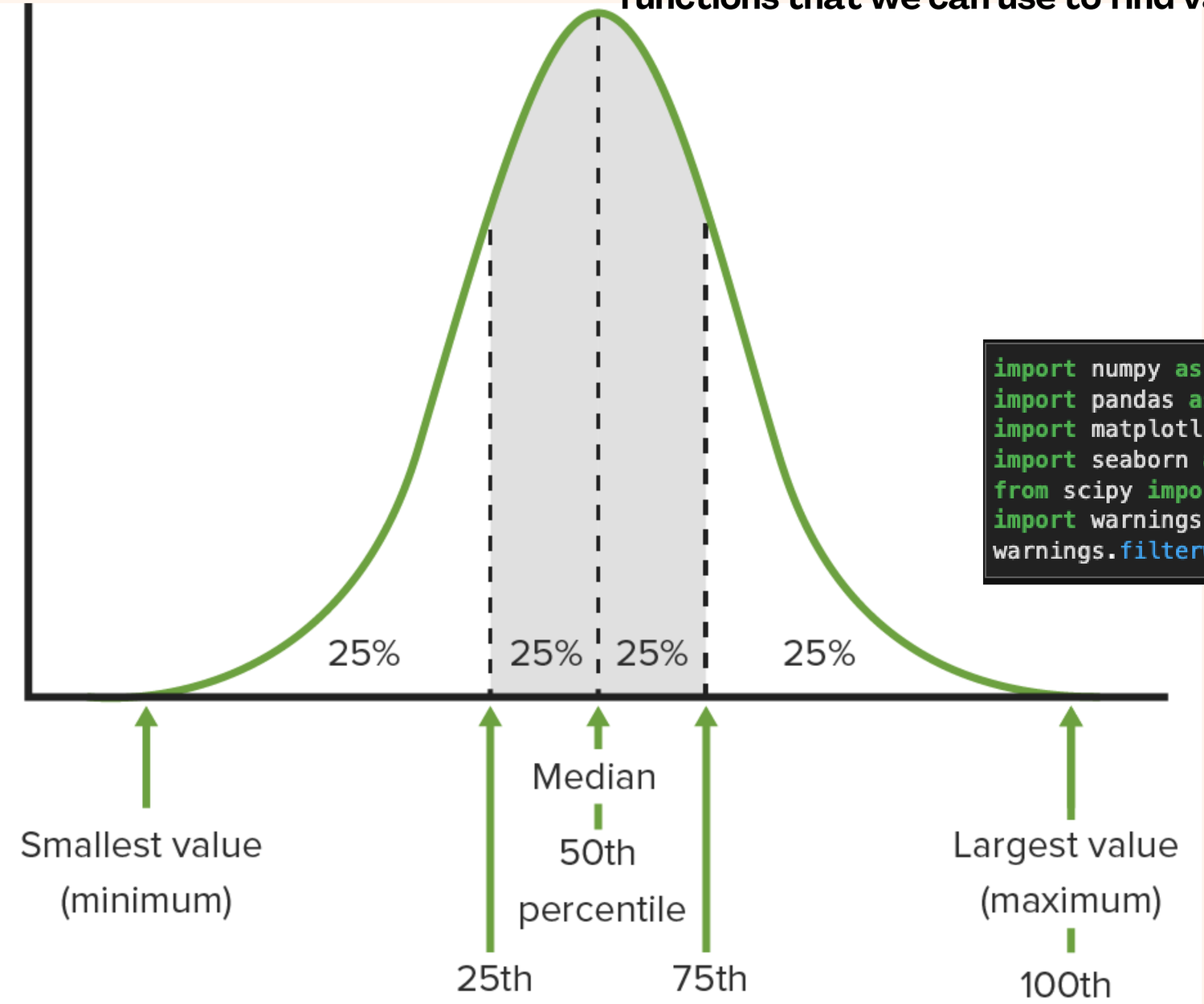
measure of dispersion (size of dispersion _ is a measure to present how the distribution of each data. Measrue of dispersion shows how the condition of a data spreads in the group of data that we have, this allows us to analyze how far the data is spread from the size of the playback

If the distribution of the data is low, this indicates that the data is spread not far from the center, on the contrary, if the distribution is far, it indicates that the data is far from the center.

There are several that are often used to measure the distribution of data, including:

- Range shows how far the distribution is regardless of the shape of the distribution.
- Quartile is a measure of the spread that divides the data into 4 parts, as the name implies, the quartiles divide the data into 25% in each part
- Variance is a measurement of the variability of the data that knows how far the data held above.
- Standard Deviation is the most commonly used measure of spread because it provides clear and intuitive information. To get the standard deviation value we only need to do the square root of the variance. The standard deviation describes how different the values in our data are from the mean.
- Skewness is a statistic used to describe the distribution of data whether it is skewed to the left, right or symmetrical
- Kurtosis is a statistic used to describe whether the distribution of data tends to be flat or tapered

Python and some of its libraries such as Numpy, Pandas, Scipy, which have built-in functions that we can use to find values in addition to




```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
import warnings
warnings.filterwarnings("ignore")
```

CONSIDERATIONS IN FILLING NULL VALUES

Pros :

mean
1. in calculations always consider all values
2. not sensitive to the increase in the amount of data.
3. suitable for homogeneous data

median
1. not affected by outliers.
2. can be used for both qualitative and quantitative data
3. suitable for heterogeneous data

Mode
1. Not affected by outliers
2. suitable for both quantitative and qualitative data

Cons :

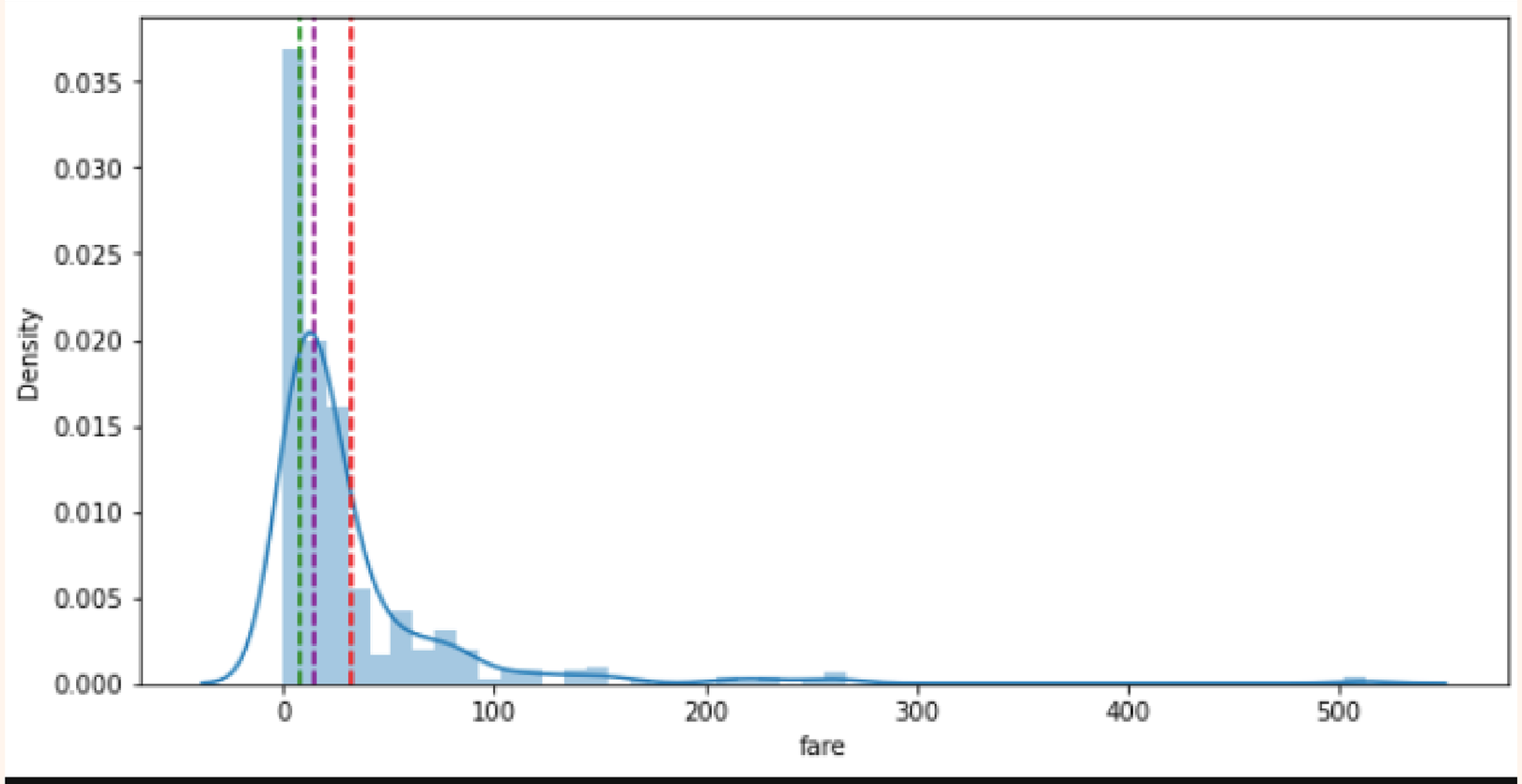
mean
1. Very sensitive to data outliers. If there are many outliers, the average becomes less representative.
2. Cannot be used for qualitative data
3. Not suitable for heterogeneous data

Median
1. Does not consider all data values.
2. does not reflect the population average,
3. Sensitive to increasing the amount of data.

Mode
1. Not as much as it is in one dataset
2. Sometimes in one dataset there are two or more modes, if that happens then the mode becomes difficult to use
3. not considering all values
4. Sensitive to increasing the amount of data

VISUALIZATION CENTRAL TENDENCY

```
# Visualisasi Central Tendency
plt.figure(figsize=(10,5))
sns.distplot(titanic['fare'])
plt.axvline(titanic['fare'].mean(), color='red', ls='--')
plt.axvline(titanic['fare'].median(), color='purple', ls="--")
plt.axvline(titanic['fare'].mode()[0], color='green', ls='--')
plt.show()
```



MEASURE OF DISPERESE WITH NUMPY AND SCIPY

```python
#contoh measure of disperese dengan numpy dan scipy
print("nilai range =", np.max(titanic['fare'])-np.min(titanic['fare']))
print("nilai stdev =", np.std(titanic['fare']))
print("nilai variance =", np.std(titanic['fare']**2))
print("nilai quantile1 =", np.quantile(titanic['fare'], .25))
print("nilai quantile2 =", np.percentile(titanic['fare'], 50))
print("nilai quantile3 =", np.percentile(titanic['fare'], 75))
print("nilai skewness =", stats.skew(titanic['fare']))
print("nilai kurtorsis =", stats.kurtosis(titanic['fare']))
```

```
nilai range = 512.3292
nilai stdev = 49.66553444477411
nilai variance = 17316.267565034144
nilai quantile1 = 7.9104
nilai quantile2 = 14.4542
nilai quantile3 = 31.0
nilai skewness = 4.7792532923723545
nilai kurtorsis = 33.20428925264474
```

```python
# Contoh measure of disperse dengan pandas
print("nilai range =", titanic['fare'].max()-titanic['fare'].min())
print("nilai std =", titanic['fare'].std())
print("nilai variance =", titanic['fare'].var())
print("nilai quantile1 =", titanic['fare'].quantile(.25))
print("nilai quantile2 =", titanic['fare'].quantile(.50))
print("nilai quantil3 =", titanic['fare'].quantile(.75))
print("nilai skewnes =", titanic['fare'].skew())
print("nilai kurtosis =", titanic['fare'].kurtosis())
```

```
nilai range = 512.3292
nilai std = 49.693428597180905
nilai variance = 2469.436845743117
nilai quantile1 = 7.9104
nilai quantile2 = 14.4542
nilai quantil3 = 31.0
nilai skewnes = 4.787316519674893
nilai kurtosis = 33.39814088089868
```

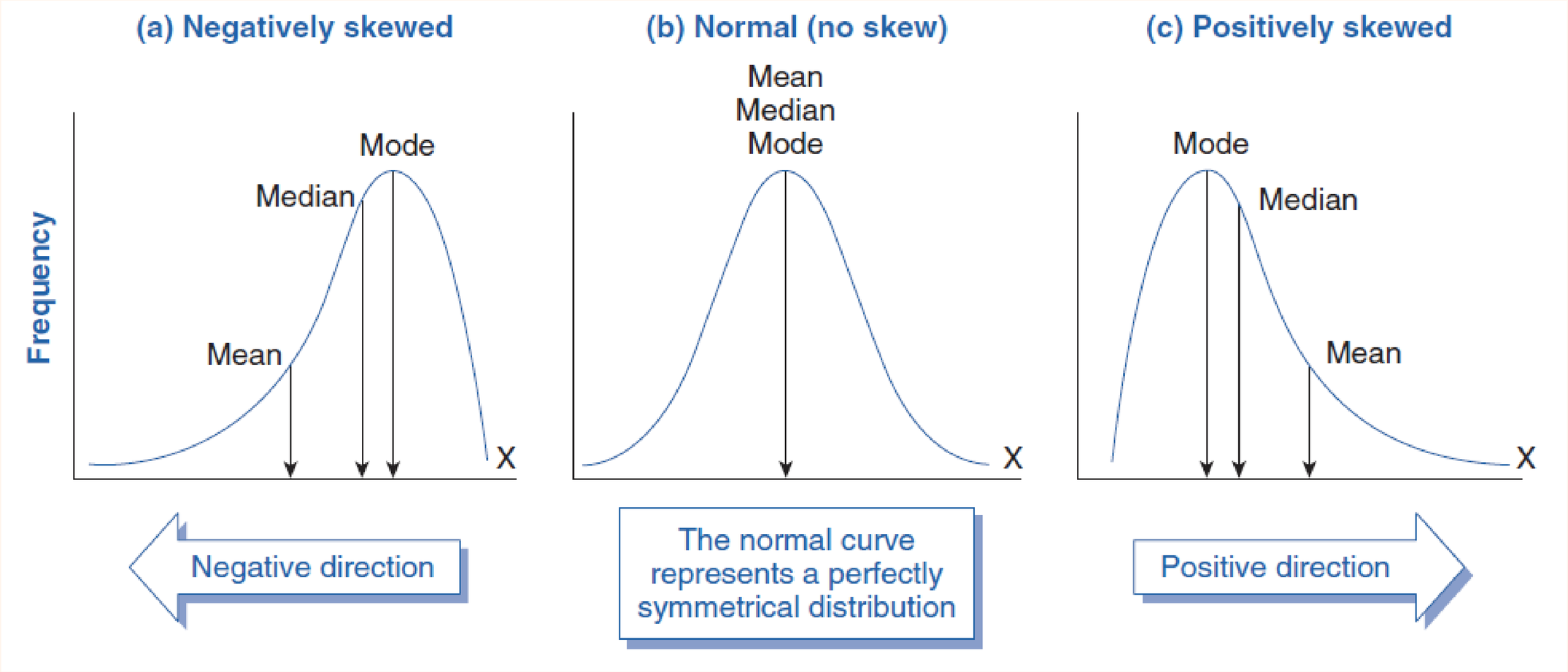
MEASURE OF DISPERESE WITH PANDAS

SKEWNESS

Skewness is a measure of the asymmetry in the distribution of values. Skewness can be positive, negative and zero.

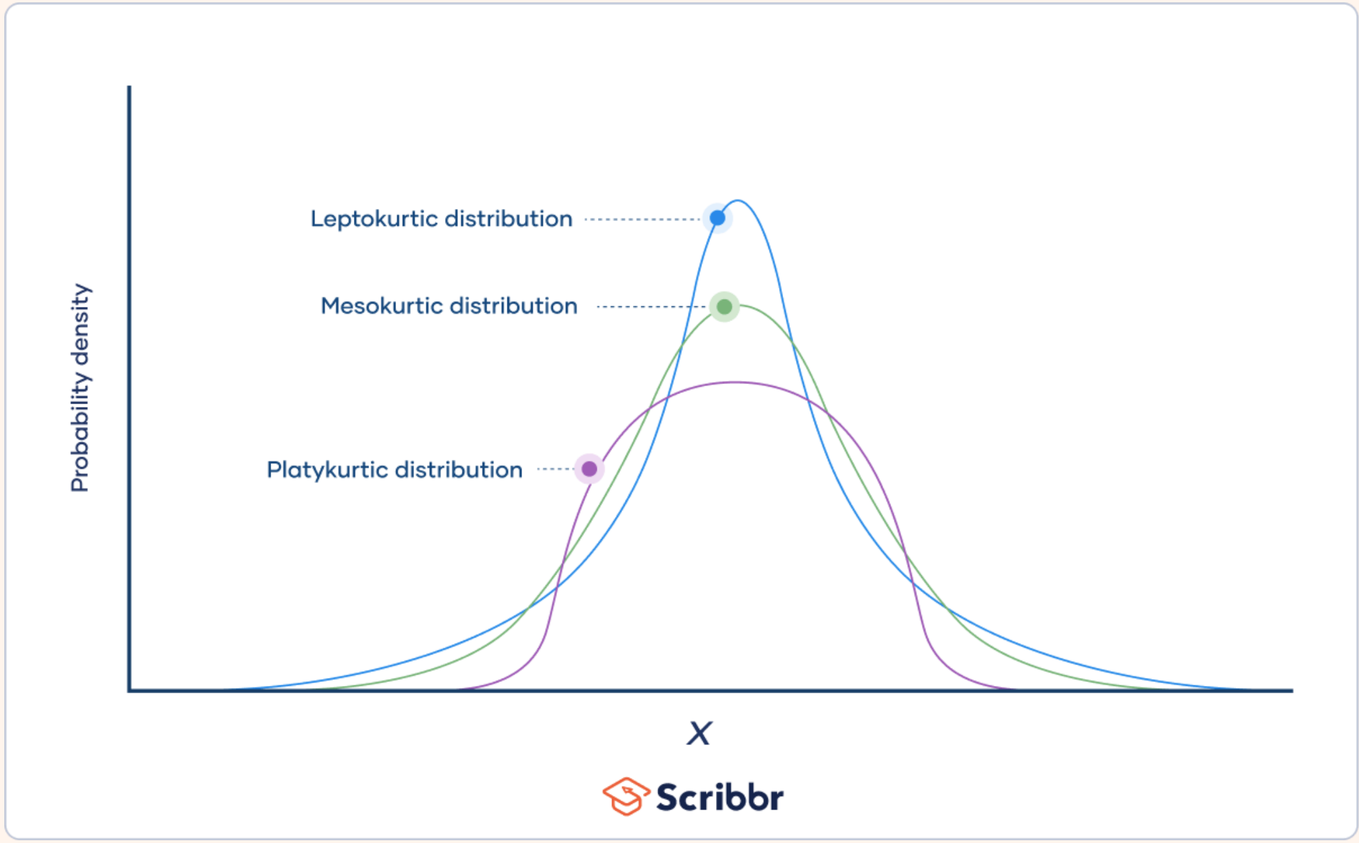
- Skewness with a positive value (> 0) means that the tail of the distribution is on the right, meaning that most of the distribution is in a low value (left of the curve).
- Skewness which is negative (<0) means that the tail of the distribution is on the left, meaning that most of the distribution is in the high value (right of the curve)
- skewness which is zero = (0) means that the value is symmetrically distributed, with the same distance between the right and left tails.

For example, if the value of skewness = -0.817, it means that it is a negative value. but not far from the value 0, meaning that the data distribution tends to be symmetrical, or almost normal



Kurtois

Kurtosis is an indicator to show the degree of tailedness, the greater the value of kurtosis, the sharper the curve.

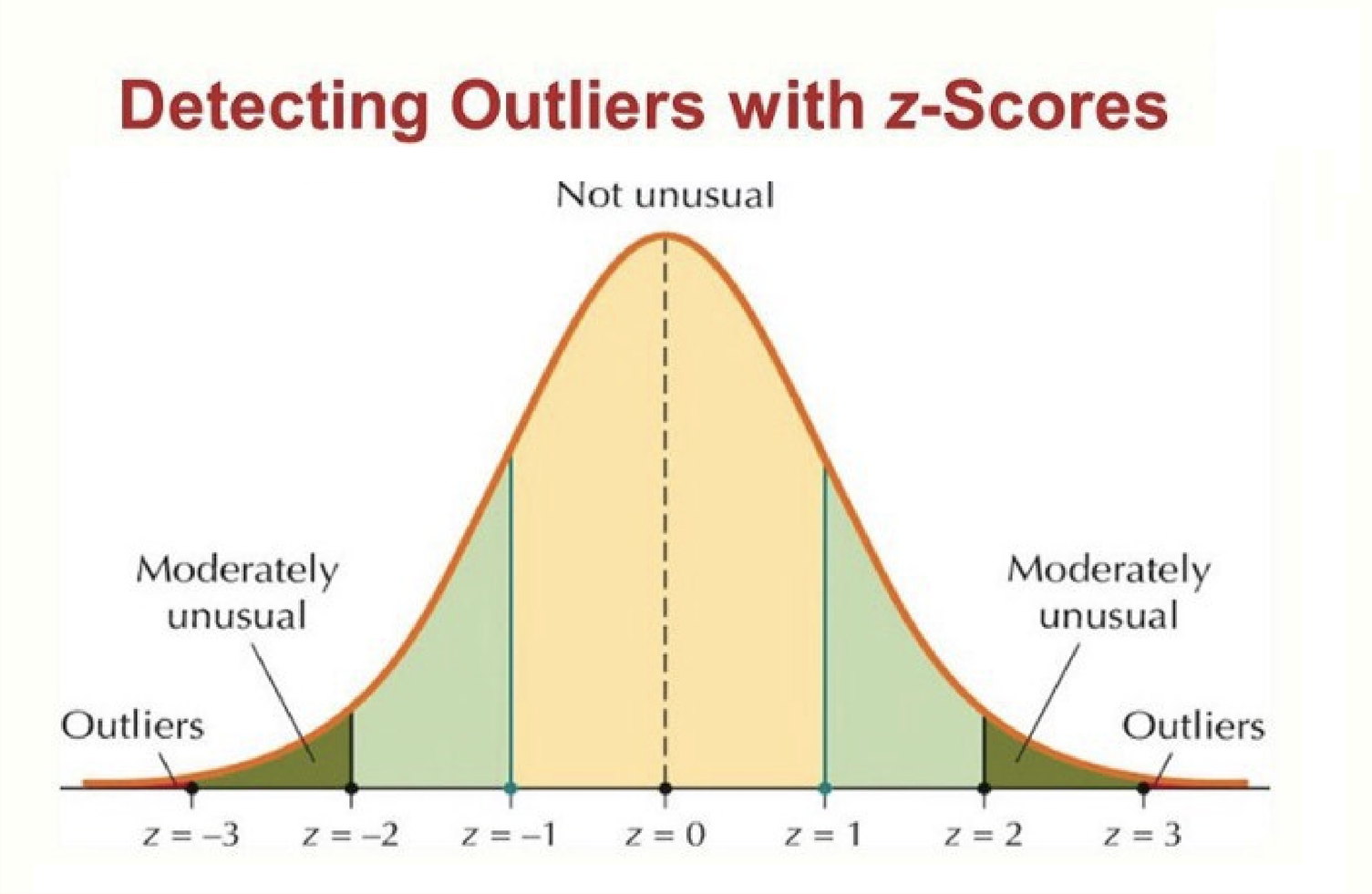


The reference value for kurtosis is 3. it means that on a normal kurtosis test:

- f the kurtosis value is > 3, then the distribution curve is called leptokurtic.
- If < 3 then it is called palticurtic.
- If the value of kurtosis = 3, then the curve includes a normal distribution, it can be called meskurtic or mesokurtotic

Standard Score(z-score)

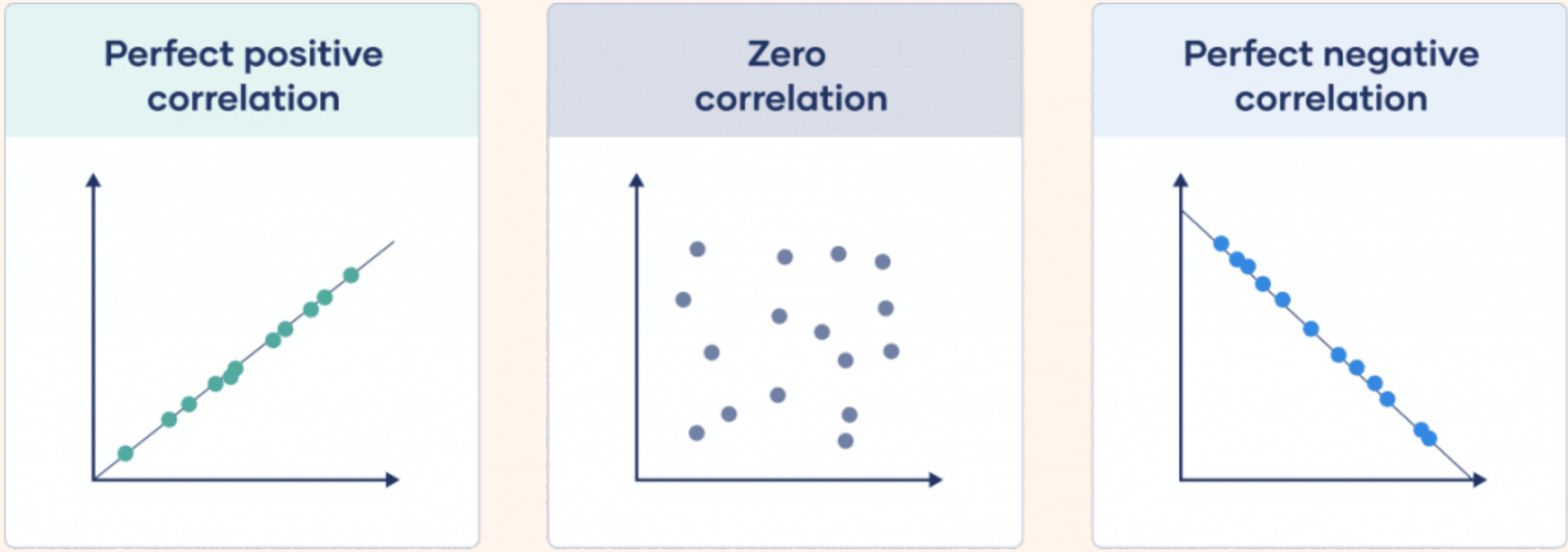
represents the value of the deviation of a data entry against the mean of the dataset measured based on the standard deviation



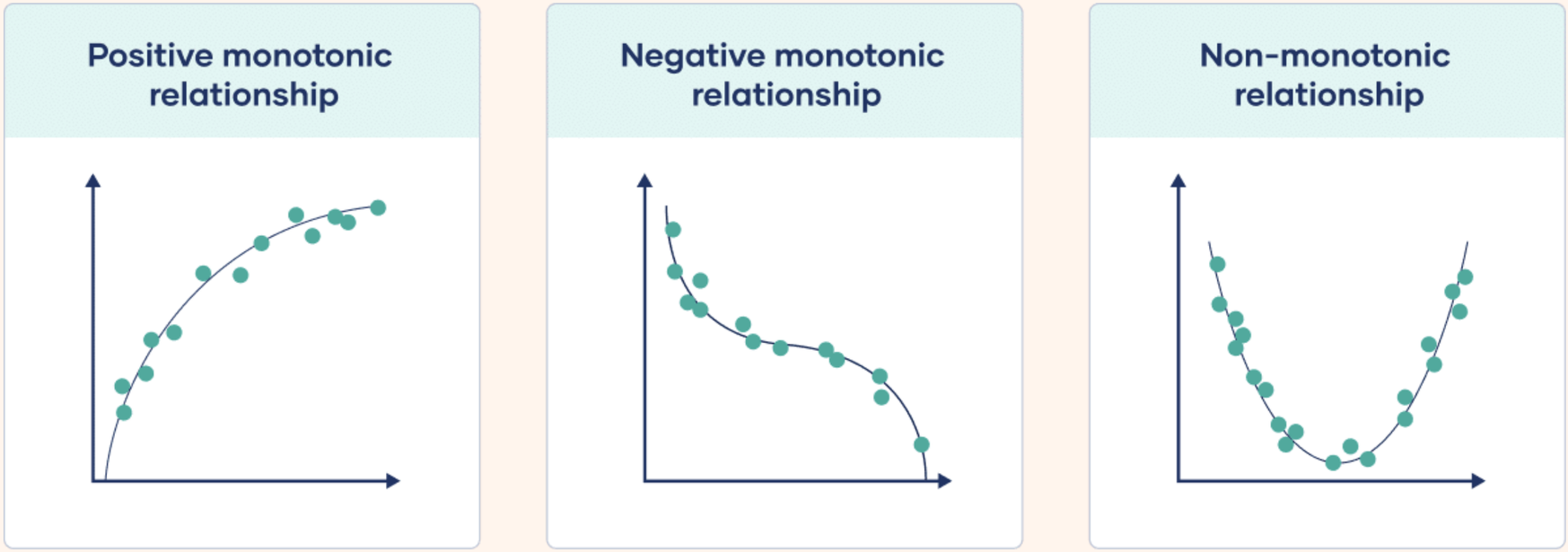
CORRELATION COEFFICIENT

In simple terms correlation means relationship, the word correlation has different meanings in various fields such as statistics and research methods.

Pearson correlation Evaluation of linear relationship between two continuous variables



Spearman Correlation Evaluation of monotonic relationship between two variables. Spearman correlation is based on the ranked values of each variable.



- The value of the correlation coefficient is positive, meaning that for every increase in the positive value of x, there is an increase in the value of y.
- The value of the correlation coefficient is negative, meaning that for every increase in the positive value of x, there is a decrease in the value of y.
- The correlation coefficient value is 0 , meaning that there is no relationship between x and y.

THANK YOU!