

# Introduction to Statistics

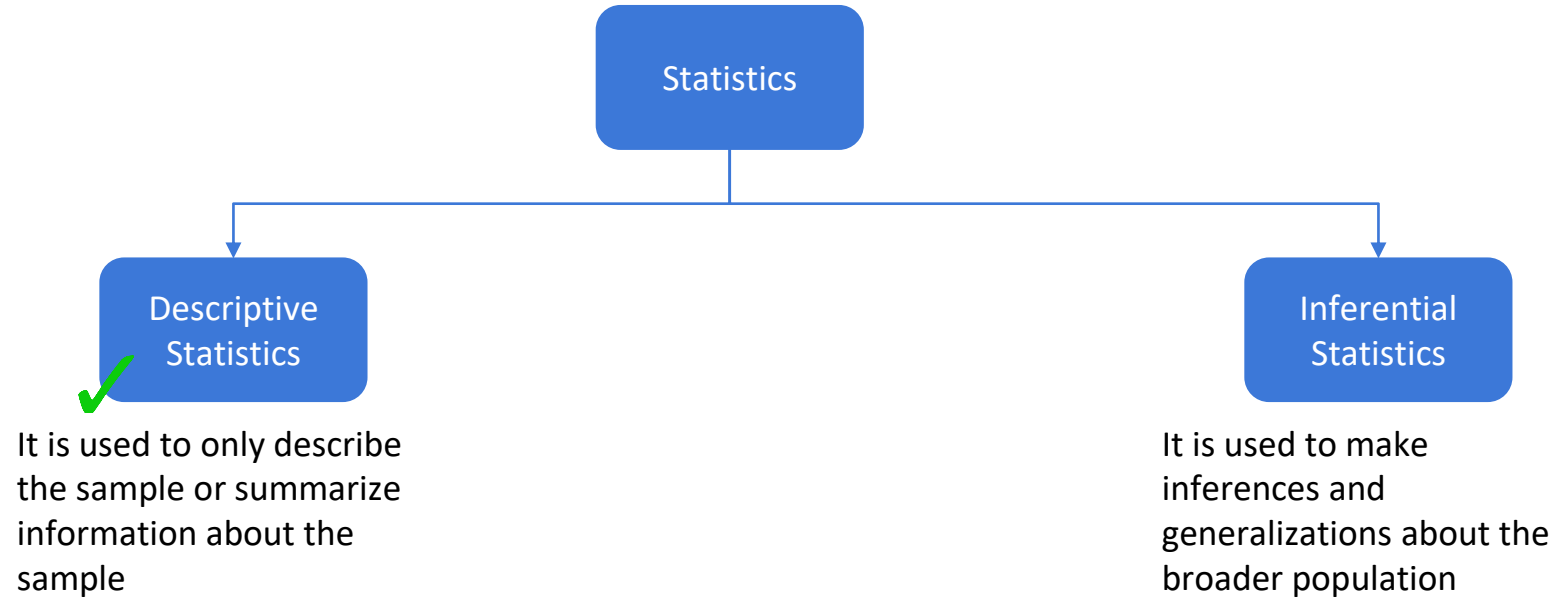
# Agenda

- Classification of Data
  - Time Series Data
  - Cross Sectional Data
  - Pooled Data
  - Panel Data
- Read Data From Various Sources
- Understanding the variables
  - Numerical Variable
  - Categorical Variable

# Agenda

- Descriptive Statistics
  - Measure of Central Tendency
  - Distribution of the data
  - Measure of Dispersion
  - Skewness and Kurtosis
  - Covariance and Correlation

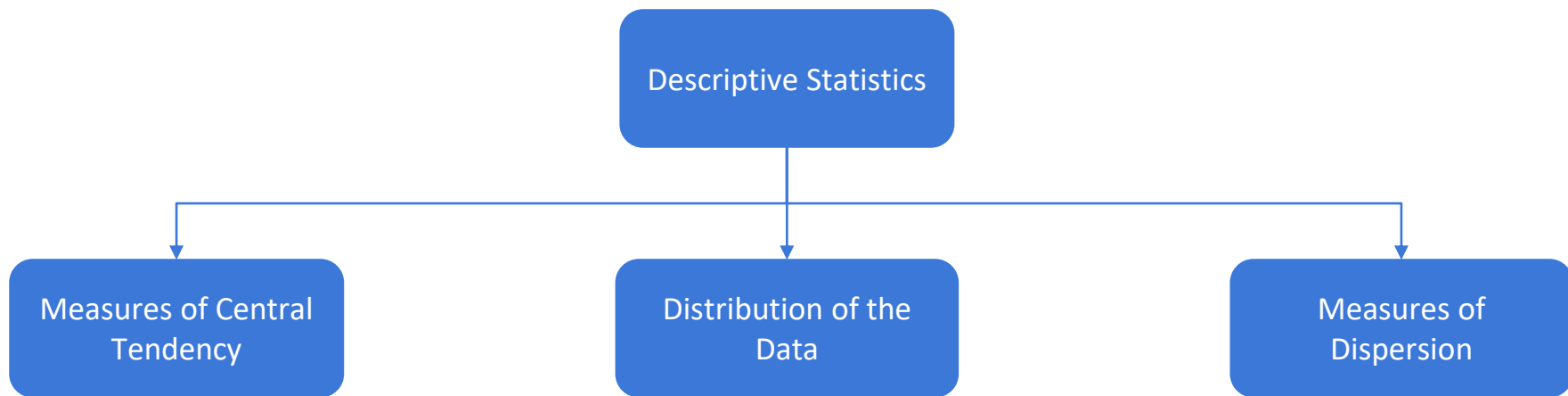
# Statistics



# Descriptive Statistics

# Descriptive Statistics

Descriptive statistics is the term given to the analysis of the data that helps describe, visualize or summarize the data



# Exploratory Data Analysis (EDA)

- EDA is a process of analyzing the datasets to summarize their main features using numerical and visual methods
- In this session, we shall cover the numerical methods to analyze the data
- The numerical methods includes descriptive statistics

# The Data



# What is Data?

Data is units of information in structured or unstructured format.

Examples:

- Collection of relevant tweets
- Records of yield in a farm over a period of time
- Records of stock price every minute
- Records of performance of a sports person

# Classification of Data

In general, there are four types of data. They are:

- Time Series Data
- Cross sectional data
- Pooled Data
- Panel Data

# Time series data

- Time series data is the set of observations on a variable at different time points
- The data may be collected daily, weekly, monthly, or annually

Date	Prev.Close	Open.Price	High.Price	Low.Price
01-Jan-1996	408	407	407.9	405
02-Jan-1996	407.9	407	409	406.25
03-Jan-1996	406.25	409	409	409
04-Jan-1996	409	405	407	405
05-Jan-1996	406.3	401.5	401.5	401.5
08-Jan-1996	401.5	401.5	405	402
09-Jan-1996	404	404	400	395
10-Jan-1996	399.5	399.5	397	395
11-Jan-1996	396	399	405	396
12-Jan-1996	405	403	403	400

Sample data of  
daily stock  
price of a  
company



This file is meant for personal use by nbilagi@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

# Cross-sectional data

Cross-sectional data are set of observations on two or more variables at the same time point

Produce in the 2019 in (quintals)			
Region	Rice	Wheat	Maize
Region 1	28.76	86.66	8.23
Region 2	84.03	65.45	41.94
Region 3	71.75	28.99	38.43
Region 4	14	68.87	7.04
Region 5	43.97	54.27	86.51

Sample data of a farm produce in five regions in the year 2019



# Pooled data

Pooled data is combination of both time series data and cross sectional data.

	Produce in the 2011		Produce in the 2012	
Appliance	Production	Profit (in '000 Rs)	Production	Profit (in '000 Rs)
Radio	280	13	121	8
Television	840	645	1456	5192
Washing Machine	775	2899	152	5706
Computer	876	6887	1005	6349
Air Conditioner	439	5427	328	8095

Sample data of the number of appliances produced in an industry for two years

# Panel data

- Panel data is type of pooled data
- The same cross-sectional unit is surveyed over time
- Also known as longitudinal or micro-panel data

Appliance	Year	Production	Profit (in '000 Rs)
Radio	2010	280	13
Television	2010	840	645
Washing Machine	2010	775	2899
Computer	2010	876	6887
Air Conditioner	2010	439	5427
Radio	2011	234	123
Television	2011	835	645
Washing Machine	2011	564	2899
Computer	2011	874	6887
Air Conditioner	2011	435	5427

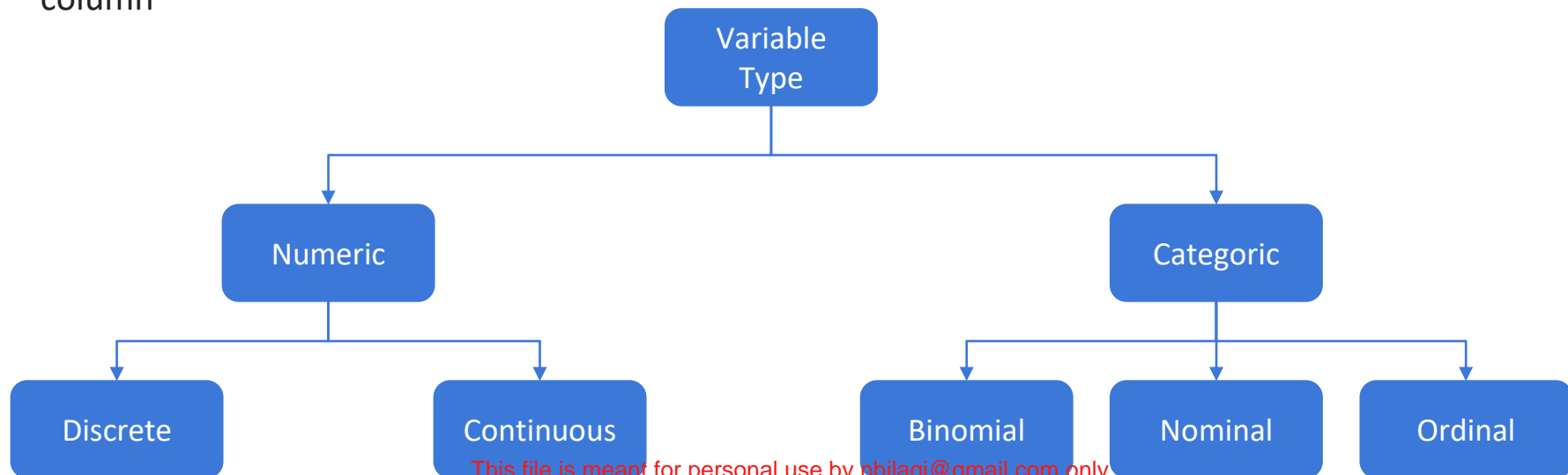
Sample data of the number of appliances produced and profits earned by a company over a period of two years



# Understanding the variable

# Understand the variable

To understand the descriptive statistics, we need to first understand the type of each variable or column





# Numerical Variable

- A variable which takes a numeric value is called a **numeric variable**
- Also known as **quantitative variable**
- A **discrete numeric variable** is a random variable which takes discrete values, i.e. values from the set of whole numbers only. It can take countably finite values

## Examples:

1. Number of cars passing by a toll-gate every one minute
2. Number of defective items in a box

# Numerical Variable

A **continuous numeric variable** is a variable which can have infinite number of values within a range

## Examples:

1. The amount of rainfall in millimeters
2. Price of a stock

# Categorical Variable

- Categorical variable has two or more levels
- Also known as **qualitative variable**

## Examples:

1. Colors: red, orange, yellow
2. Gender: male, female
3. Economic status: low, medium, high

# Categorical Variable

## Nominal Data:

- Nominal data has **no order** and has **two or more than two categories**
- It represents discrete units and are used to label variables
- Example: Housing type - Apartment, Bungalow, Penthouse

## Binary Data:

- Binary data has **no order** and has **strictly two categories**
- Also known as dichotomous variable
- Example: Presence or absence of a disease

# Categorical Variable

## Ordinal Data:

- Ordinal data is **ordered** nominal data
- It represents discrete and ordered units
- Example: Education - Primary, Secondary, High School, College and University which are ordered



- At times data may appear to be numeric but is actually categorical
- Consider the adjoining table. The 'Gender' column has value 1 and 0
- The numeric values used to represent a categorical variable cannot be used for numerical computation

Name	Gender	Age
Ria	1	45
Arya	1	23
Sam	0	34
Joe	0	54
Harry	0	12
Sarah	1	43

# Data Analysis



# Descriptive Statistics

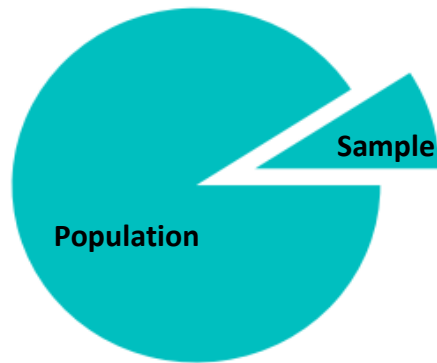


# What is descriptive statistics?

- Descriptive statistics is the term given to the analysis of the data that helps describe, visualize or summarize the data
- Helps exhibit the patterns in the data
- However, descriptive statistics only help in understanding the data and not make conclusions regarding any hypothesis
- We can not generalise the facts, they are confined only to the data at hand (that may be the [sample](#))

# Population & Sample

- **Population** is a collection of all the individuals or objects
- **Sample** is a subset of the population which is a representative of the population



## Example:

Suppose a company producing electric bulbs wants to know the average life of a bulb. If the details on average life of all bulbs are available, then this information is regarded as the population.

It would be challenging for the company to test each and every bulb produced. In such a scenario, the company would draw a sample from the produced bulbs to test.

# Descriptive statistics

The major characteristics we observe in each variable of the data set are:

1. Measures of central tendency
2. Distribution of the data
3. Measures of dispersion

The above characteristics help us in understanding the data and are part of the exploratory data analysis

# Measures of Central Tendency

# Measures of central tendency

- A measure of central tendency is a single value that identifies the central position of the data
- It includes mean, median, mode and partition values

# Mean

- Mean is a central tendency of the data
- It is defined as the sum of all observations divided by the number of observations:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{N}$$

**where N is the total number of observations**

- The whole data is spread out around the mean

# Mean

## Merits:

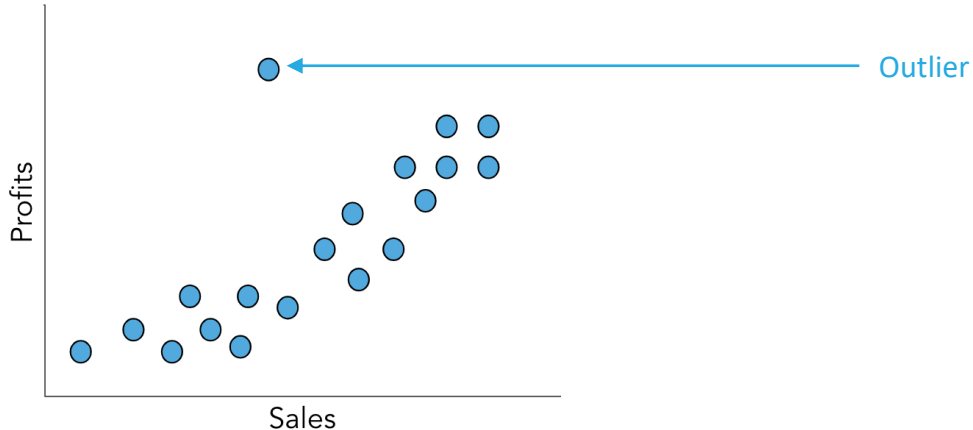
- Based on all observation

## Demerits:

- Affected by extreme observations (**outliers**)
- Arithmetically it is not possible to obtain the mean of a data with missing values
- Mean can not be calculated for categorical data

# Outlier

- An outlier is a value that behaves differently than other observations
- It does not follow the usual pattern





# Obtain the mean

Each column of a DataFrame is a pandas Series.

Let us create a series and find its mean.

```
prices = pd.Series([0,0,35,40,10,54,87,12,95,64,56,4,45,76,34,56,
                    87,34,56,32,56,48,89,42,65,100,99,98,100,96,99])
```

```
# obtain the mean of 'prices'
prices.mean()
```

```
57.064516129032256
```

# Obtain the mean

Arithmetically, it is not possible to compute mean of missing data. But, python ignores the missing values and calculates the mean of the available data in the series

Introduce missing values in  
the previous series



```
# let us introduce some missing values in the series
# the remaining values are same as the series 'prices'
prices_missing = pd.Series([0,0,35,40,10,54,87,np.NaN,12,95,64,56,4,45,76,34,np.NaN,56,
                        87,34,56,32,56,48,np.NaN,89,42,65,100,99,98,100,96,99])
```

```
# obtain the mean of prices_missing
prices_missing.mean()
```

```
57.064516129032256
```



## Mean of series with missing values

We notice that the means obtained are the same

This is because the function drops the missing values before calculating the mean value

```
# obtain the mean of 'prices'  
prices.mean()
```

```
57.064516129032256
```

```
# obtain the mean of price_missing  
prices_missing.mean()
```

```
57.064516129032256
```

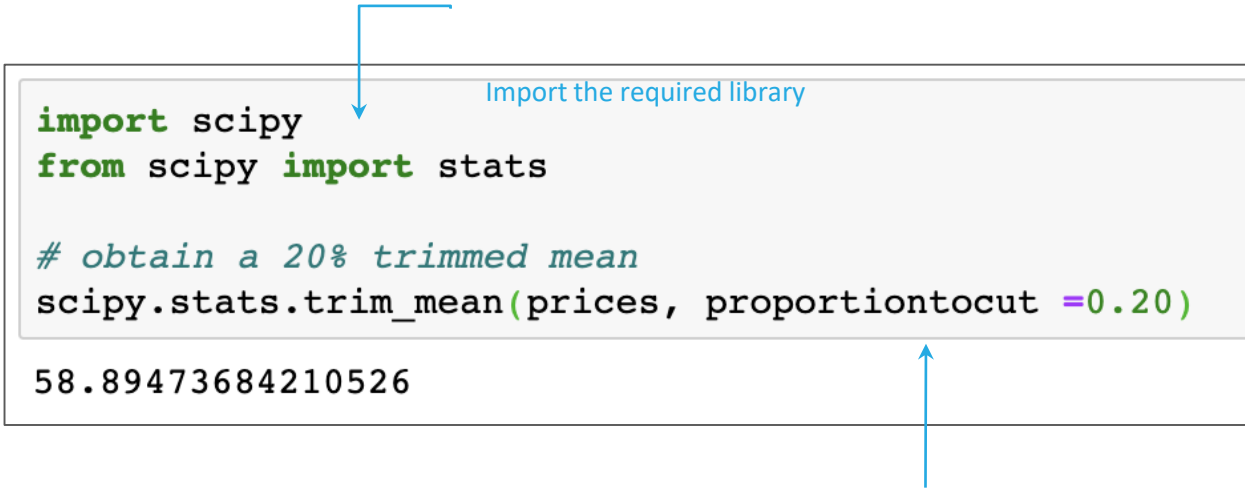
# Trimmed mean

- Arithmetic mean obtained by ignoring lowest and highest  $\alpha\%$  observations is called as the  $\alpha\%$  - trimmed mean
- Using a trimmed mean, it is possible to eliminate the influence of outliers
- Note the observations are arranged either in increasing or decreasing order
- Also called as **truncated mean**

Example: The trimmed means are used in scoring the performance of an athlete during olympic games, to minimize the extreme scoring from possible biased judgements.

# Obtain the trimmed mean

While working in python, we can directly use the built in function, `trim_mean()`, available in the library `scipy`.



```
import scipy
from scipy import stats

# obtain a 20% trimmed mean
scipy.stats.trim_mean(prices, proportiontocut = 0.20)

58.89473684210526
```

Parameter 'proportiontocut' denotes the  $\alpha\%$  values to exclude. It takes value between 0 to 0.51

# Median

- Median is the middle most observation in the data when it is arranged either in ascending or descending order of their values
- It divides the data into two equal parts. Thus, it is a **positional average**
- Median will be the middle term, if the number of observations is odd
- Median will be average of middle two terms, if number of observations is even

# Median

## Merits:

- It is not affected by extreme values

## Demerits:

- It is not based on all observations

## Obtain the median

```
prices.median()
```

```
56.0
```

**Interpretation:** The median value is 56. It implies that the value 56 divides the data in two equal parts. There are 50% of the observations above and below this price.



# Using mean and median

- For a numeric variable, the missing value is replaced by the mean if there are no outliers present
- In case if outliers are present the missing values are replaced by the median
- For symmetric data median = mean, so either of the value can be used

# Mode

- Mode of the data is the value which has the highest frequency
- In other words, it is most repeated observation
- A set of observations having two modes is said to be **bimodal**
- A set of observations may have more than two modes. Such data is said to be **multimodal**

# Mode

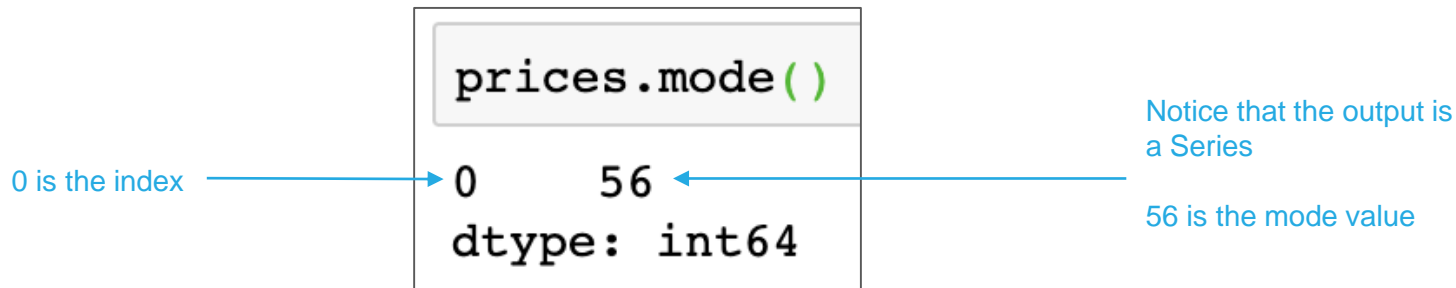
## Merits:

- It is applicable to both numeric and categorical data
- It is not affected by extreme observations

## Demerits:

- It is not based on all observations

# Obtain the mode



**Interpretation:** The mode value is 56. It implies that the value 56 occurs the most number of times. Also note that there is only one mode. The data is **unimodal**.

# Obtain the mode

```
# create a series of pen prices
prices_pens = pd.Series([0,35,40,10,54,35,100,100])

# display the value counts of the
prices_pens.value_counts()
```

```
100    2
35     2
54     1
40     1
10     1
0      1
dtype: int64
```

Number of times each value is  
appearing in the data

Notice values 35 and 100  
each occur 2 times

```
# print the mode
prices_pens.mode()
```

```
0    35
1   100
dtype: int64
```

The mode values

**Interpretation:** The mode values are 35 and 100. It implies that these values occur equal number of times and have the highest frequency.

Since there are two modes, the data is **bimodal**.

This file is meant for personal use by nbilagi@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

# Obtain the mode

```
# create a series of pen prices
prices_pens = pd.Series([0,0,35,40,10,54,76,34,56,87,34,56,100,99,100,99])

# display the value counts of the
prices_pens.value_counts()
```

```
56      2
99      2
100     2
34      2
0       2
76      1
10      1
40      1
87      1
54      1
35      1
dtype: int64
```

Number of times each value is  
appearing in the data

Notice values 0, 34, 56, 99, 100  
each occur 2 times

```
# print the mode
prices_pens.mode()
```

```
0      0
1     34
2     56
3     87
4     99
5    100
dtype: int64
```

The mode values

**Interpretation:** The mode values are 0, 34, 56, 87, 99, 100. It implies that these values occur equal number of times and have the highest frequency.

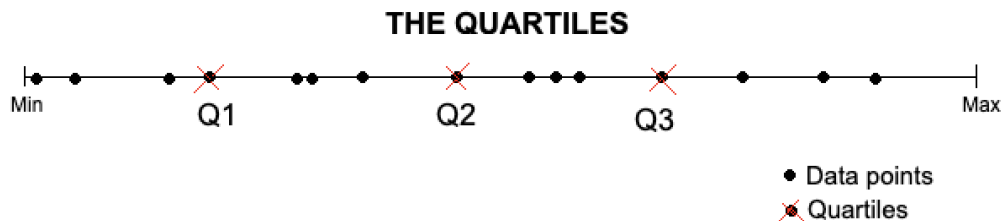
Also note that there are 6 mode values. Data is **multimodal**.

# Partition values

- Partition values are the values which divide the dataset into equal parts
- Note that partition values may not be equispaced
- Also known as **quantiles**
- We have seen that the **median** divides the data into two equal halves. Similarly in order to make four parts we use **quartiles**, for ten parts we use **deciles**

# Quartiles

- The values that divide the dataset into four equal parts are called quartiles



- 25% of the data lies below the first quartile (Q1) and 75% above it
- The second quartile divides the data points in two equal halves. Q2 is the median
- The third quartile (Q3) divides the observation in 75%-25% meaning that 75% of the data lies below Q3 and 25% above it



# Obtaining the quartiles

The partition values are obtained using the quantile function specifying the percentage of data below the required partition value.

```
# to get the first quartile  
prices.quantile(.25)
```

34.5

```
# to get the second quartile  
prices.quantile(.50)
```

56.0

```
# to get the third quartile  
prices.quantile(.75)
```

88.0

# Deciles

- The values that divide the dataset into ten equal parts are called deciles
- 10% of the data lies below the first decile (D1) and 90% above it
- Similarly, 20% of the data lies below the second decile (D2) and 80% above it and so on
- The fifth decile is the same as the second quartile i.e.  $D5 = Q2 = \text{Median}$
- There are in all 9 deciles

Obtain the nine deciles (D9) for the variable 'prices' defined before.

Similarly we define percentiles as values which divides the data into 100 equal parts.

Obtain the 25th (P25), 50th (P50) and 75th (P75) percentile.

# Summary

- Mean, median and mode can be used to impute the missing values in the data
- For a symmetric distribution, mean = median = mode
- For any data,

$$Q2 = D5 = \text{Median} = P50$$

- Mean is affected by the presence of outliers
- Trimmed mean is useful when the outliers are present in the data
- Median and mode are not affected by the presence of outliers

# Distribution of the Data

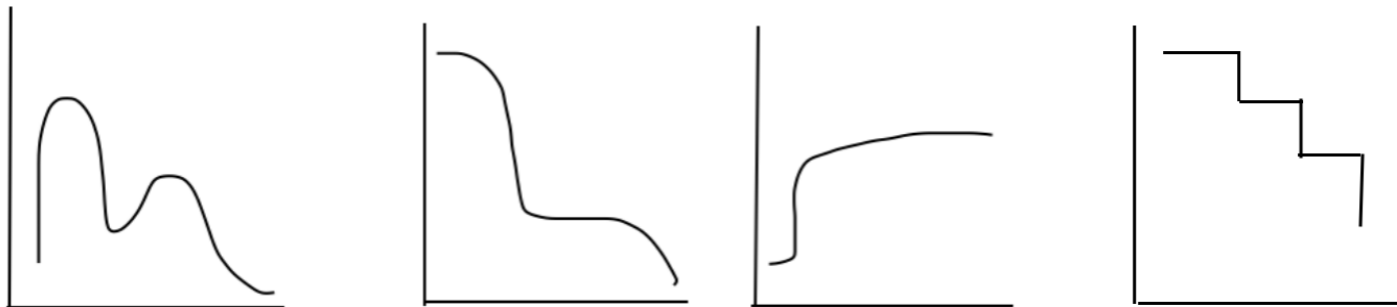
# Distribution of the data

- The distribution is a summary of the frequency of values taken by a variable
- The distribution of the data gives information on the shape and spread of the data
- On plotting the histogram or a frequency curve for a variable, we actually look at how the data is distributed over its range

# Distribution of the data

- The data may be distributed in a random matter
- However there are standard distributions in statistics like the normal distribution, binomial distribution, student's-t distribution and so on

Some distributions



# Measures of Dispersion



# Measures of Dispersion

- The measure of dispersion refers to the variability within the data
- Variability is the measure of how close or far the data lie from the central value
- Reliability of a measure of central tendency is more if dispersion in the data is less
- It includes range, variance, standard deviation, coefficient of variation and interquartile range

# Range

- Range is the difference between the largest and the smallest observation
- It is defined as

$$\text{Range} = X_n - X_1$$

where  $X_n$  is the largest value and  $X_1$  is the smallest value

# Range

## Merits:

- It is a basic way to measure the variability

## Demerits:

- If either the smallest or the largest value turns out to be an outlier then the range value is unreliable
- The other in-between observations do not affect the range value
- It does not give a clear picture of variability in the data

This file is meant for personal use by nbilagi@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

# Obtain the range

```
# Series 1
year_establishment = pd.Series((1955, 1900, 1980, 1990, 2000))
Range = year_establishment.max() - year_establishment.min()
Range
```

100

```
# Series 2
year_establishment = pd.Series((1920, 1990, 1925, 1930, 2020))
Range = year_establishment.max() - year_establishment.min()
Range
```

100

Use the formula for range

**Interpretation:** The range of both the variables is 100. However, it seen that the first series has observations towards the maximum and the second series has observations towards the minimum. Thus, **range is not a good measure of dispersion**.

# Variance

Variance is the arithmetic mean of squares of deviations taken from the mean. It is given as

$$\text{Variance } (\sigma^2) = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}$$

**Interpretation:** The variance looks at how spread out the observations are from the mean. The higher the variance more the data is spread out.

## Obtain the variance

```
# obtain the variance for the variable 'prices'  
prices.var()
```

```
1054.3290322580644
```

More the data is spread out away from the mean, higher the variance.



## The unit of measurement of variance

- The unit of measurement of the variance is not the same as that of the original data
- Consider an industry producing screws. The diameter of each screw is noted in millimeters
- The variance in the diameter is measured in squared millimeters
- Thus a usual practice is to take square root of the variance which would have the same unit as the original data
- The value thus obtained is called the standard deviation of the data

# Standard deviation

The standard deviation of the variable is the square root of the variance.

$$\text{Standard deviation } (\sigma) = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}}$$



# Obtain the standard deviation

```
prices.std()
```

```
32.470433200960905
```

**Interpretation:** The standard deviation looks at how spread out the observations are from the mean. The higher the standard deviation more the data is spread out.

# Coefficient of variation

- The coefficient of variation is a statistical measure of dispersion of data points around the mean

$$\text{Coefficient of Variation (C.V.)} = \frac{\text{Standard Deviation}}{|\text{Mean}|} \times 100 = \frac{\sigma}{|\bar{x}|} \times 100$$

- It is always expressed in percentage and is unit free
- Generally, used to compare dispersion of two or more groups
- According to Karl Pearson, C.V. is the percentage variation in the mean and S.D. is the total variation in mean

# Obtain the coefficient of variation

```
# importing variation from scipy.stats
from scipy.stats import variation

scipy.stats.variation(prices)

0.5597598237090757
```

The variation() from the scipy library computes the coefficient of variation.

Note: The name of the function can be quite misleading; do not confuse with 'variance'.

# Obtain the coefficient of variation

Consider two financial portfolios, A and B. Obtain the the better portfolio in which an investor should invest.

Portfolio	Expected Return (%)	Volatility (%)
A	25	30
B	10	15

# Obtain the coefficient of variation

Obtain the coefficient of variation for the two portfolios

```
# the expected return of portfolio A (in %)
expected_return_A = 25

# the volatility of portfolio A (in %)
volatility_A = 10

# obtain the coefficient of variation
cv = volatility_A/expected_return_A

# print the coefficient of variation
cv
```

0.4

```
# the expected return of portfolio B (in %)
expected_return_B = 30

# the volatility of portfolio B (in %)
volatility_B = 15

# obtain the coefficient of variation
cv = volatility_B/expected_return_B

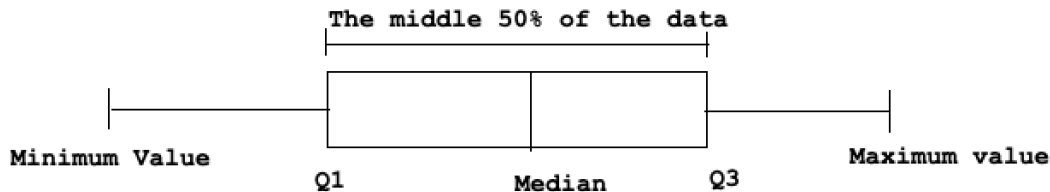
# print the coefficient of variation
cv
```

0.5

**Interpretation:** The coefficient of variation for portfolio A is lower than that of portfolio B. So, the investor should invest in portfolio A because it is comparatively more consistent in its returns.

# Quartiles

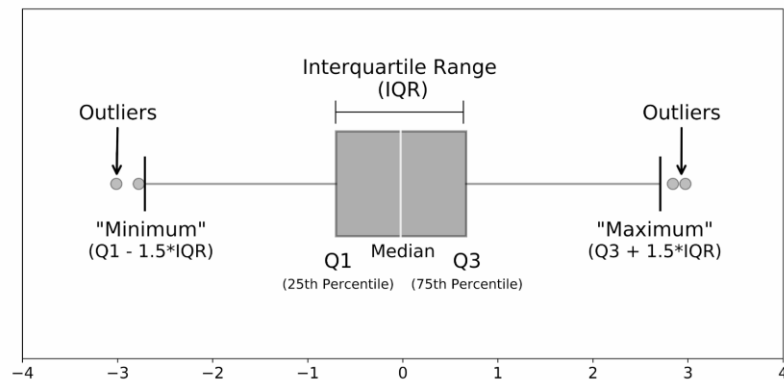
- We have seen quartiles are kind of partition values
- With the help of the quartiles we understand the spread of the data
- Quartiles when plotted with box plot help in identifying the outliers



This file is meant for personal use by nbilagi@gmail.com only.  
Sharing or publishing the contents in part or full is liable for legal action.

# Understanding box plot

- A box plot is a way of displaying the distribution of data based on a five number summary (minimum, first quartile (Q1), median, third quartile (Q3), and maximum)
- The whiskers give an idea of the spread of the data
- The dots outside of the whiskers are the outliers



# Interquartile range

- The interquartile range is defined as the difference between the third and the first quartile
- It gives the range in which the middle 50% of the data lies
- The value of IQR is used to remove outliers in the data

$$\text{IQR} = Q_3 - Q_1$$



# Obtain interquartile range

```
# obtain the first quartile  
Q1 = prices.quantile(0.25)  
  
# obtain the quartile  
Q3 = prices.quantile(0.75)  
  
# obtain the IQR  
IQR = Q3 - Q1  
  
# print IOR  
IQR
```

53.5

**Interpretation:** The interquartile range of the prices is 53.5

# Z-score

- Every value in the dataset has the z-score
- To find the z-score of a value, subtract the mean from it and divide it by the standard deviation
- The z-scores are extensively used in statistics especially in hypothesis testing
- It is also used to **normalize** the data, ie is it scales the data to get the mean of the scaled data as 0 and variance as 1

$$\text{Z- score} = \frac{x_i - \bar{x}}{\sigma}$$

# Z-score

The z-score tells how much is the value away from the mean

Z-score Value	What it means?
0	The corresponding value is same as the mean
< 0	The corresponding value is less than the mean
> 0	The corresponding value is greater than the mean
< -3 or > 3	The corresponding value is potential outlier

This file is meant for personal use by nbilagi@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

# Summary

- Range is not a reliable measure as it do not consider the in-between observations
- The variables with a near-zero standard deviation have less significance in the analysis
- Coefficient of variation (CV) is the unitless quantity
- Box-plot is used to visualize five-number summary (minimum, Q1, Q2, Q3, and maximum)
- IQR can be used to detect and remove the outliers in the data

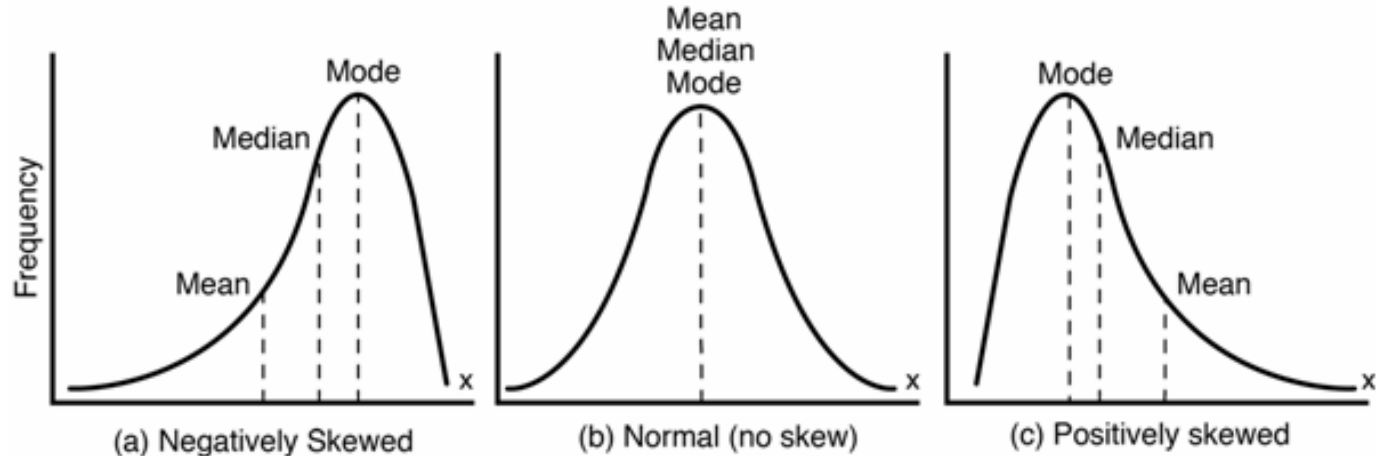
# Skewness

# Skewness

- Skewness is a lack of symmetry or departure from symmetry
- If the distribution of the data is elongated on either side then the data is said to be skewed
- If the distribution of the data is elongated on the left side then the data is said to be left skewed
- If the distribution of the data is elongated on the right side then the data is said to be right skewed

# Skewness

The graph of skewed distribution



mean < median < mode

mean = median = mode

mean > median > mode

This file is meant for personal use by nbilagi@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

# Calculate the skewness

The Bowley's Coefficient of Skewness is defined as

$$S_b = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$

where  $Q_1$ ,  $Q_2$  and  $Q_3$  are first, second and third quartiles respectively

Value of skewness	Interpretation
$S_b < 0$	The distribution is negatively skewed
$S_b = 0$	The distribution is not skewed (symmetric distribution)
$S_b > 0$	The distribution is positively skewed

This file is meant for personal use by nbilagi@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.



# Calculate the skewness

The Bowley's coefficient of skewness:

```
# calculate Bowley's coefficient of skewness
# calculate first quartile
Q1 = prices.quantile(0.25)

# calculate second quartile
Q2 = prices.quantile(0.5)

# calculate third quartile
Q3 = prices.quantile(0.75)

s_b = (Q3 - 2*Q2 + Q1) / (Q3 - Q1)

# Bowley's coefficient of skewness
print(s_b)

0.19626168224299065
```

**Interpretation:** The coefficient is positive but near zero. Thus, the variable 'prices' is said to be near symmetric.

# Calculate the skewness

The Karl Pearson's Coefficient of Skewness is defined as

$$S_k = \frac{3(\text{mean} - \text{median})}{\sigma}$$

Where  $\sigma$  is the standard deviation

Value of skewness	Interpretation
$S_k < 0$	The distribution is negatively skewed
$S_k = 0$	The distribution is not skewed (symmetric distribution)
$S_k > 0$	The distribution is positively skewed

This file is meant for personal use by nbilagi@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.

# Calculate the skewness

The Karl Pearson's coefficient of skewness:

```
# calculate Karl Pearson's coefficient of skewness
# calculate the mean
price_mean = prices.mean()

# calculate median
price_med = prices.median()

# calculate standard deviation
price_sd = prices.std()

s_k = (3*(price_mean - price_med)) / price_sd

# Karl Pearson's coefficient of skewness
print(s_k)

0.09835250325524642
```

**Interpretation:** The coefficient is positive but near zero. Thus, the variable 'prices' is said to be near symmetric.

This file is meant for personal use by nbilagi@gmail.com only.  
Sharing or publishing the contents in part or full is liable for legal action.

# Calculate the skewness

- The adjusted Fisher-Pearson Coefficient of Skewness is defined as

$$G_1 = \sqrt{\frac{N(N-1)}{N-2}} \frac{m_3}{m_2^{3/2}}$$

Where,  $m_2$  and  $m_3$  are the 2<sup>nd</sup> and 3<sup>rd</sup> central moments respectively.

$$m_2 = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^2 \quad \text{and} \quad m_3 = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^3$$

- This coefficient of skewness is based on the central moments of the distribution

# Calculate the skewness

Value of skewness	Interpretation
$G_1 < 0$	The distribution is negatively skewed
$G_1 = 0$	The distribution is not skewed (symmetric distribution)
$G_1 > 0$	The distribution is positively skewed

# Obtain the skewness

In python, the 'skew()' returns the value of adjusted Fisher-Pearson coefficient of skewness.

```
# use the 'skew()' to calculate the skewness
prices.skew()

-0.20497188990720197
```

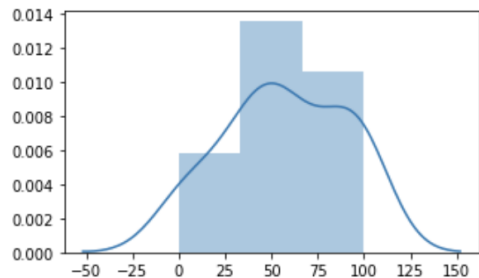
**Interpretation:** The variable prices is slightly negative skewed. It can be said it is close to symmetric.

```
# import the required libraries
import seaborn as sns
import matplotlib.pyplot as plt

# set the plot size
plt.figure(figsize=(5,3))

# plot a distribution plot
sns.distplot(prices)

# display the plot
plt.show()
```





## Coefficient of skewness

- The magnitude of the coefficient of skewness explains the extent of skewness in the distribution
- The +/- sign of a coefficient provides the direction of skewness
- The coefficient compares the distribution of the sample to the normal distribution. Thus, the zero value represents the symmetric distribution
- A higher magnitude of the coefficient indicates that the distribution highly differs from the normal distribution

# Summary

## Bowley's Coefficient of Skewness

- Based on quartiles
- Value lies between -1 to +1
- Can be used to calculate skewness for open-ended distribution
- Preferred if the data has extreme values (outliers)

## Karl Pearson's Coefficient of Skewness

- Based on mean, median and standard deviation
- Usually value lies between -3 to +3
- Can not be used for open-ended distribution, as we can not calculate the mean in such case

## Fisher - Pearson Coefficient of Skewness

- Based on central moments
- Mostly used measure of skewness by different softwares
- Can not be used for open-ended distribution, as we can not calculate the mean in such case

This file is meant for personal use by nbilagi@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.



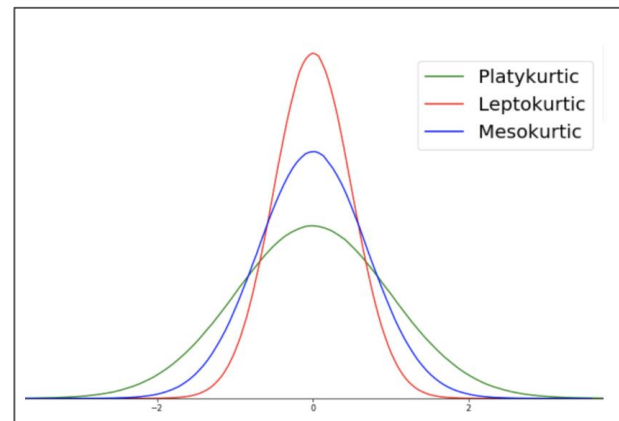
# Kurtosis

# Kurtosis

- Kurtosis measures the peakedness of the distribution
- In other words, it is a statistical measure that defines how the tails of the distribution differ from the normal distribution
- Kurtosis identifies whether the tails of a given distribution contain extreme values
- A histogram is an effective way to show both the skewness and kurtosis of a data set

# Kurtosis

Value	Thickness of Tails	Interpretation
Kurtosis < 0	Thin	The distribution is <b>platykurtic</b>
Kurtosis = 0	Normal	The distribution is <b>mesokurtic</b>
Kurtosis > 0	Thick	The distribution is <b>leptokurtic</b>



# Obtain the kurtosis

We use `kurt()` to find the kurtosis

**Interpretation:** The variable 'prices' is platykurtic since  $-1.036 < 0$ . This means there are more dispersed prices with lighter tails. i.e. less extreme values.

```
prices.kurt()
```

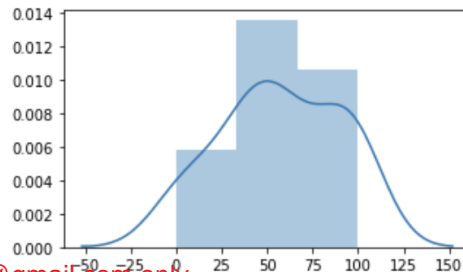
```
-1.0366224980536156
```

```
# import the required libraries
import seaborn as sns
import matplotlib.pyplot as plt
```

```
# set the plot size
plt.figure(figsize=(5,3))
```

```
# plot a distribution plot
sns.distplot(prices)
```

```
# display the plot
plt.show()
```



# Summary

- A mesokurtic distribution is basically a normal distribution
- A leptokurtic distribution has heavy tails on both the sides that indicates the presence of large outliers. Also the distribution is concentrated towards the mean than the normal distribution
- A platykurtic distribution has flat/lighter tails on both the sides and it indicates that there are small amount of outliers in the distributions. Also, the values are more dispersed and fewer values are close to the mean

# Covariance

# Covariance

- Covariance is a measure of how much two random variables vary together
- Covariance is similar to variance, but variance explains how a single variable varies, and covariance explains how two variables vary together
- It cannot tell whether the value indicates a strong relationship or weak relationship, since the covariance can take any value

# Covariance

The covariance for two variables X and Y are given by

$$COV(X,Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$X_i$  = values taken by variable X ,  $\forall X \in [1, n]$

$Y_i$  = values taken by variable Y ,  $\forall Y \in [1, n]$

$\bar{X}$  = mean of  $X_i$

$\bar{Y}$  = mean of  $Y_i$



# Obtain the covariance

```
# create imports series
imports_raw_material = pd.Series([10,11,14,14,20,22,16,12,15,13])

# create exports series
exports_finished_products = pd.Series([12,14,15,16,21,26,21,15,16,14])

# calculate the covarinace
imports_raw_material.cov(exports_finished_products)

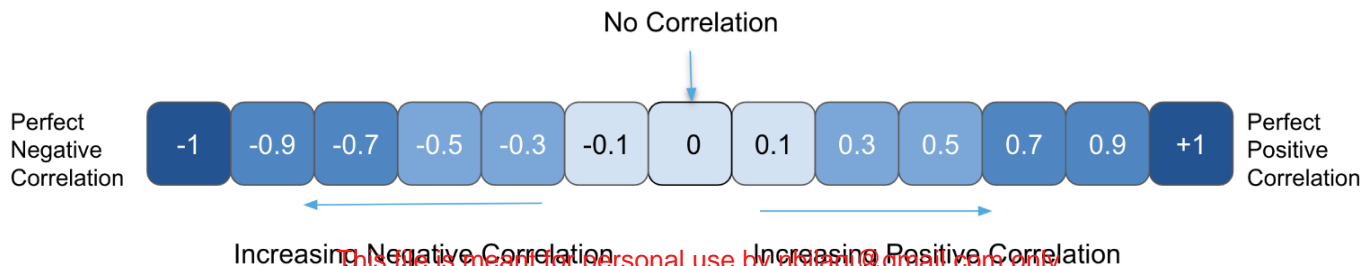
15.444444444444443
```

**Interpretation:** There is a positive covariance between imports of raw material and export of finished goods

# Correlation

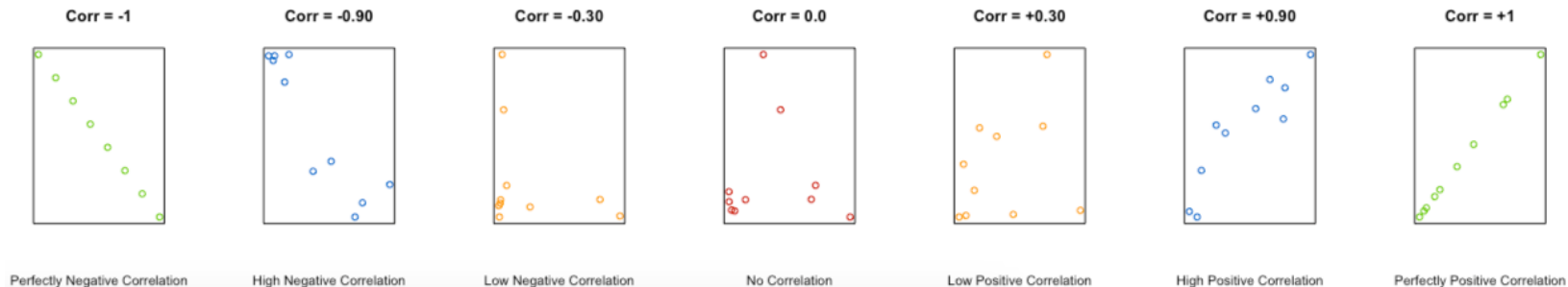
# Correlation

- Correlation is the extent of linear relationship among numeric variables
- It indicates the extent to which two variables increase or decrease in parallel
- The value of a correlation coefficient ranges between -1 and 1



# Correlation

- A positive correlation exists if variable increase/decrease simultaneously
- A negative correlation exists if one variable increases, while the other variable decreases



# Obtain the correlation

```
# create imports series
imports_raw_material = pd.Series([10,11,14,14,20,22,16,12,15,13])

# create exports series
exports_finished_products = pd.Series([12,14,15,16,21,26,21,15,16,14])

# calculate the correlation
imports_raw_material.corr(exports_finished_products)

0.9458496510416226
```

**Interpretation:** There is a high positive correlation between imports of raw material and export of finished goods

# Correlation

- Correlation is normalized form of covariance

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$$

where  $\sigma_x$  is the standard deviation of X and  $\sigma_y$  is the standard deviation of Y

- The two terms conceptually they are almost identical

# Covariance and correlation

## Covariance

- tells whether two variables are related by measuring how the variables change in relation to each other
- Gives the direction of the relationship

## Correlation

- Similar to covariance
- Tells how strong is the relationship

# Measures of central tendency

A measure of central tendency is a single value that identifies the central position of the data

Measure of Central Tendency	Description	Example
Mean	Ratio of sum of all observations to the number of observations	price = [1, 2, 3, 4, 5] mean = 3
Median	Middle most observation in the data when it is arranged either in ascending or descending order of their values	price = [6, 8, 9, 18, 20] median = 9
Mode	Value which has the highest frequency	price = [6, 18, 9, 18, 20] mode = 18

This file is meant for personal use by nbilagi@gmail.com only.

Sharing or publishing the contents in part or full is liable for legal action.



# Measures of central tendency

Measure of Central Tendency	Description
Partition values	The values that divide the data into desired number of equal parts
Quartiles	The values that divide the data into four equal parts

# Measures of dispersion

The measure of dispersion refers to the variability within the data where variability is the measure of how close or far the data lie from the central value

Measures of Dispersion	Description	Example
Range	The difference between the largest and smallest value in the data	price = [1, 2, 3, 4, 5] range = 4
Variance	The arithmetic mean of squares of deviations taken from the mean	price = [6, 18, 9, 18, 20] variance = 31.36
Standard deviation	The square root of variance <small>This file is meant for personal use by nbilagi@gmail.com only</small>	price = [6, 18, 9, 18, 20] Std deviation = 5.6

# Measures of dispersion

Measures of Dispersion	Description
Coefficient of variation	A relative measure of deviation, it is the ratio of standard deviation to the mean
Interquartile Range	The difference between the first and the third quartile

# Skewness and kurtosis

	Description
Skewness	Lack of symmetry or departure from symmetry
Kurtosis	Measures the peakedness of the distribution. In other words, defines how the tails of the distribution differ from the normal distribution

# Covariance and correlation

	Description
Covariance	Tells whether two variable are related or not and gives the direction of the relationship (positive/negative)
Correlation	Tells whether two variable are related and along with the direction of the relationship (positive/negative), it tells how strong the relationship is