greatlearning
Power Ahead

# Python for Visualization & EDA

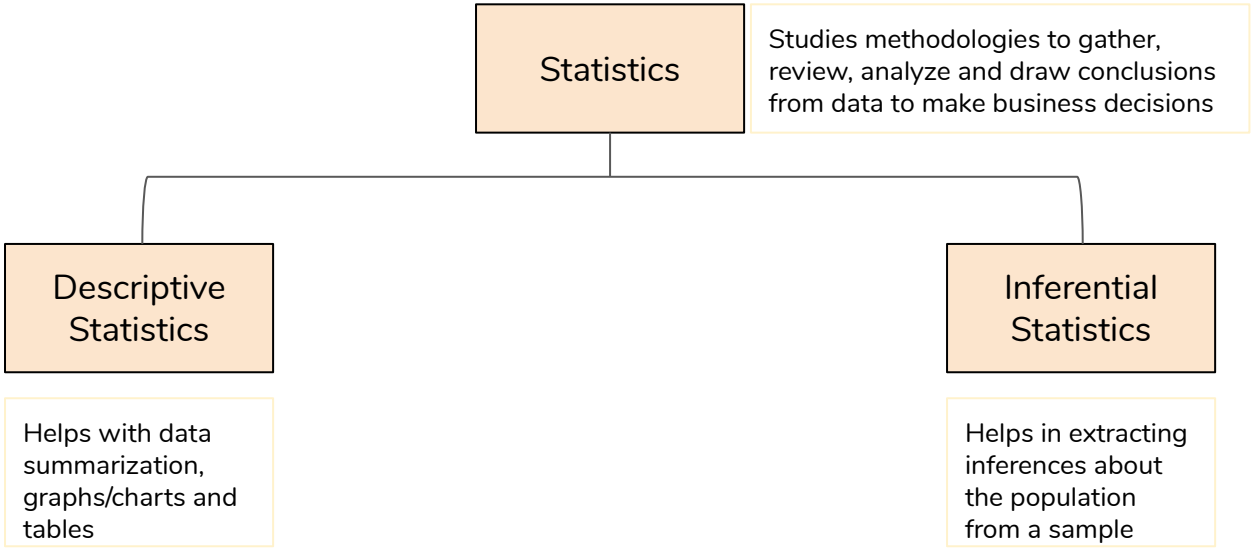
# Agenda

1. Basic Statistics
2. Types of data
3. Central Tendency and 3Ms
4. Measure of dispersion, Range, IQR, Variance, Standard Deviation
5. Covariance & Correlation
6. Introduction to Visualization
7. Common libraries for Visualization

# Pop Quiz

1. How does statistics help us with data analysis?
2. What do you understand by quantitative and qualitative data?
3. What is the difference between mean, median and mode?
4. What is the importance of visualisation for data analysis?
5. What are some of the popular charts/graphs available in Python?
6. Which plot would be the best to visualise average yearly rainfall, over the last decade - scatter plot or bar plot?
7. What is the difference between Covariance and Correlation?

# Statistics




**Importance of statistics for EDA:**

- Statistics provides the means and tools to find structure in the data
- It also give a deeper insight into what truths the data is showing.
- It is mainly used for quantitative data analysis and helps in analytical decision making.

# Terminologies in Statistics

**Population, Parameter, Sample, Statistic**

- A **population** is the universe of possible data for a specified object.
- A **parameter** is a numerical value associated with a population.
- A **sample** is a selection of observations from a population
- A **Statistic** is a numerical value associated with an observed sample.

**Example:** A marketing manager of an enterprise is facing a decision whether to introduce a new type of chair into the market or not. Consumer acceptance measured in a blind test is agreed upon as an appropriate basis for evaluation. Marketing of the new chair will be pursued only if the acceptance rate exceeds 30%. Otherwise, the new product will not be introduced in the market. A random sample of 200 consumers is collected in the blind test.

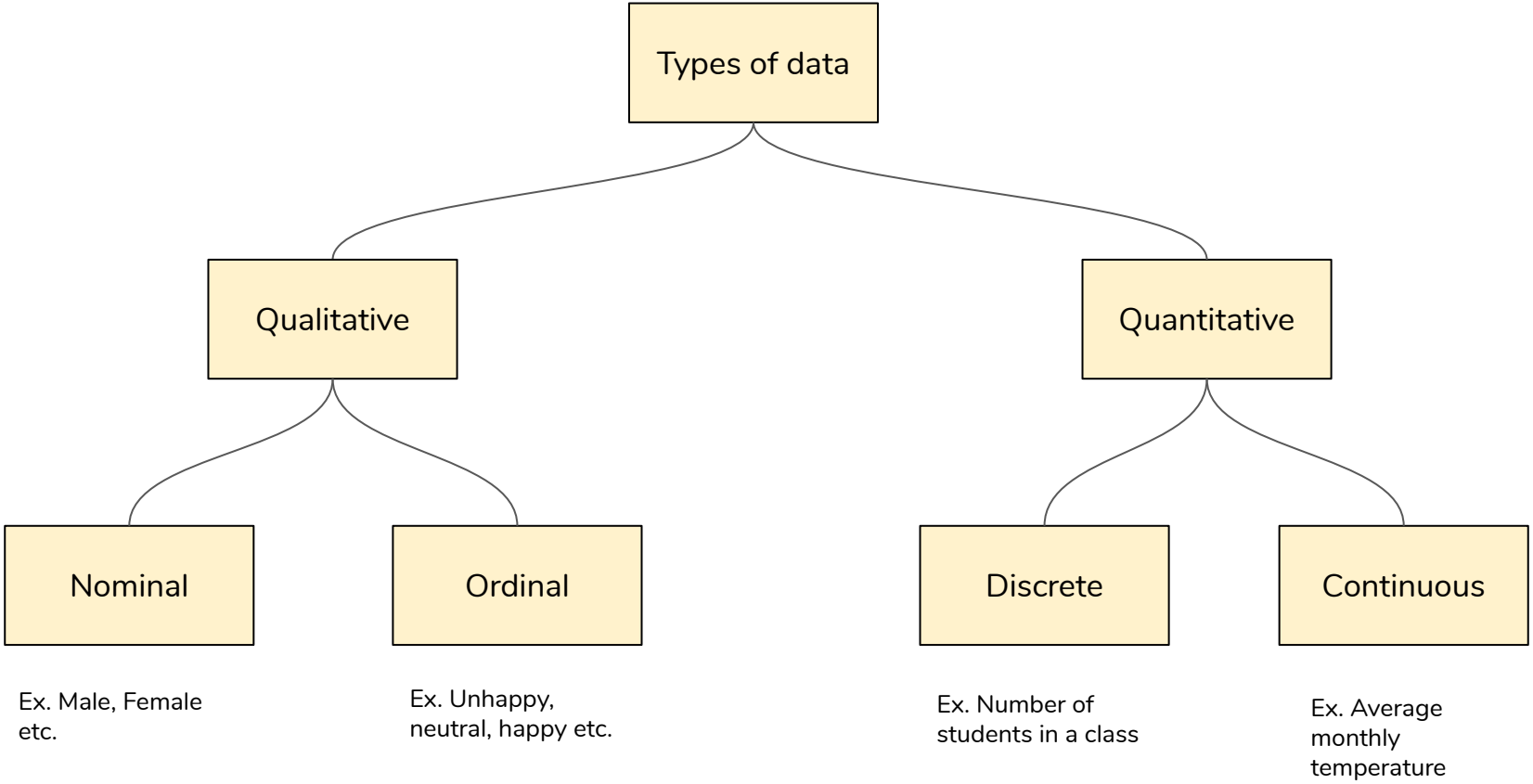
**Population -** all the target consumers

**Sample -** a random group of 200 people who have visited the store during the duration of the blind test

**Parameter -** the acceptance rate of the product

**Statistic -** the acceptance rate of the product from the blind test of 200 people

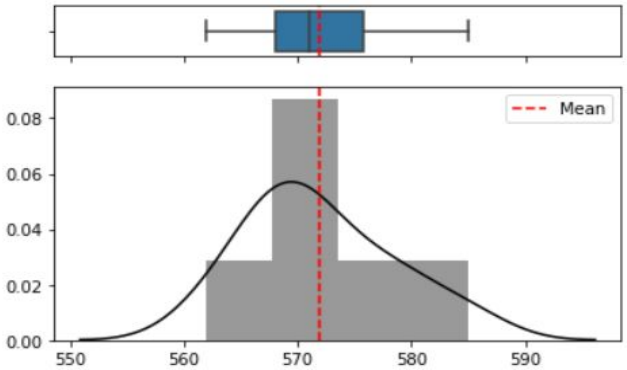
# Types of Data



# Mean

Mean is equal to the sum of all the values in the data set divided by the number of values in the data set.

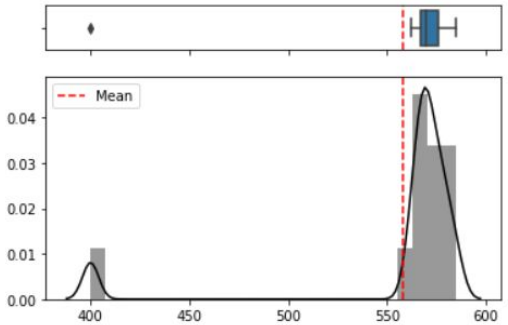
**Example:** Suppose over the last 12 days, a store sold 570, 568, 565, 572, 568, 585, 568, 578, 580, 575, 562, 572 litres of milk.




**Mean = 571.92**

But if store closed early on 1 day and sold only 400 litres of milk, the mean will be

**Mean = 557.75**

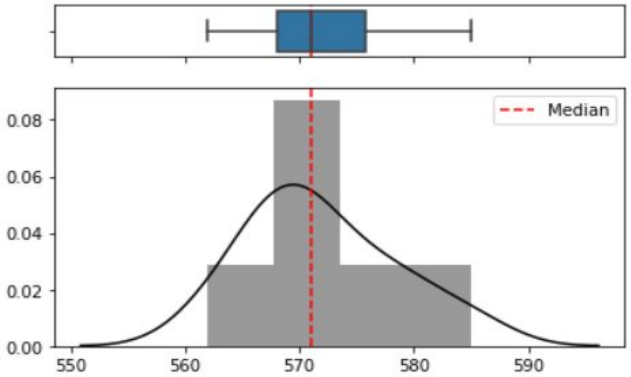



The mean has changed a lot. **Mean is affected by outliers.**

# Median

The median is the middle score for a set of data that has been arranged in order of magnitude.

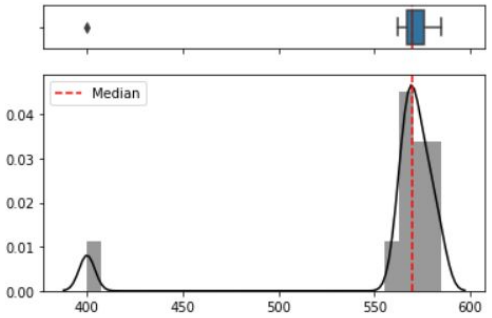
**Example:** Suppose over the last 12 days, a store sold 570, 568, 565, 572, 568, 585, 568, 578, 580, 575, 562, 572 litres of milk.




**Median = 571**

But if store closed early on 1 day and sold only 400 litres of milk, the median will be

**Median = 570**

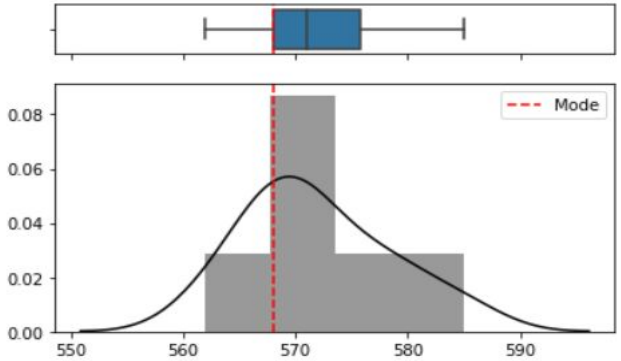



**Hence median is less affected by outliers.**

# Mode

The mode is the most frequent score in our data set. This is the only central tendency measure that can be used with nominal data, which have purely qualitative category assignments.

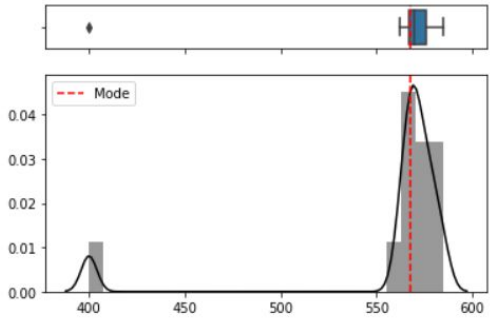
**Example:** Suppose over the last 12 days, a store sold 570, 568, 565, 572, 568, 585, 568, 578, 580, 575, 562, 572 litres of milk.




**Mode = 568**

But if store closed early on 1 day and sold only 400 litres of milk, the mode will be

**Mode = 568**



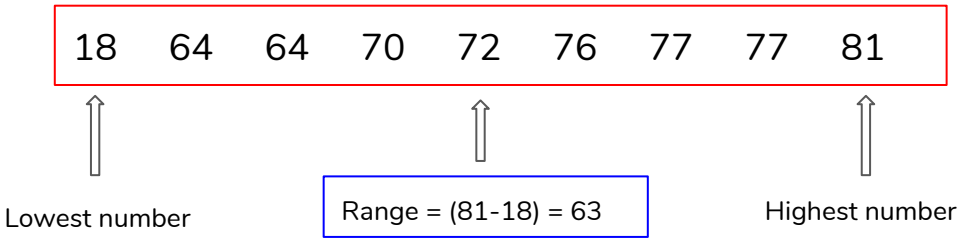

**Hence mode is not affected by outliers.**

# Measure of dispersion, Range and IQR

**Measures of dispersion**: It indicates how large the spread of distribution in around the central tendency.

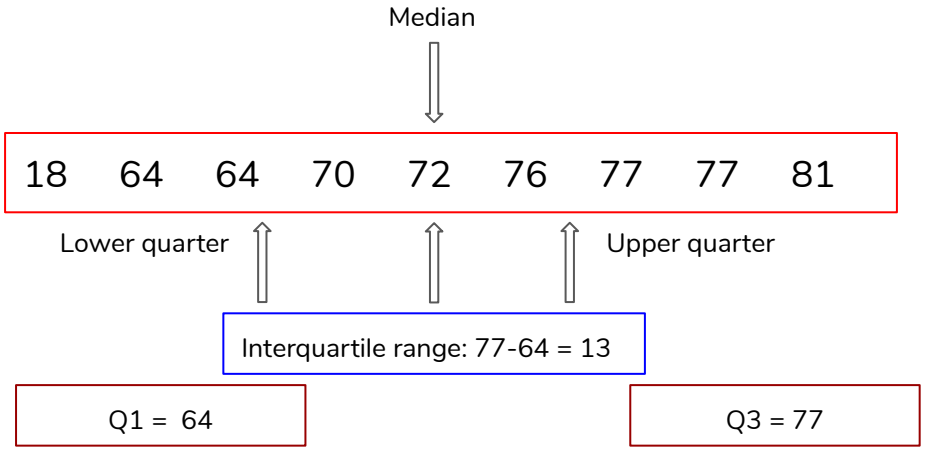
**Range:** Range is the simplest of all measures of dispersion. It is calculated as the difference between maximum and minimum value in dataset.

range = X(maximum) - X(minimum)




**Interquartile range (IQR):** It is a measure of variability, based on dividing a data set into quartiles i.e. into four parts represented by Q1,Q2,Q3 and Q4.

IQR = Q3 - Q1



# Variance and Standard Deviation

**Variance:** It describes how far each value in the data set is from the mean.

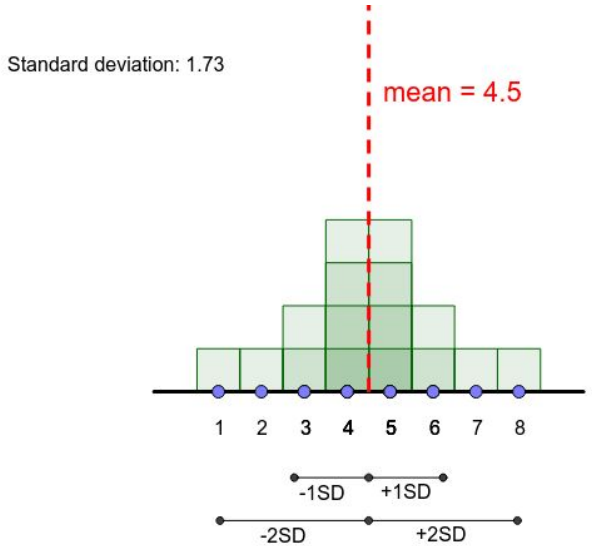
**Standard Deviation:** It is a measure of how spread out the numbers in a distribution are

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

$$Standard\ Deviation, \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Where $x_i$ = *data set values*

$\bar{x}$ = *mean of the data set*



# Covariance & Correlation

## Covariance

- Covariance is a measure of association between two variables.
- It represents association in units of the two variables.

## Correlation

- Correlation is also a measure of association between two variables.
- Moreover, it is a dimensionless quantity and thus enables comparison beyond units.

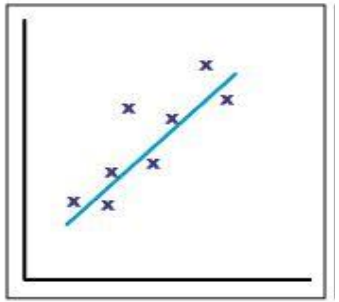
$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y}$$

Correlation between X and Y

Standard deviation of X

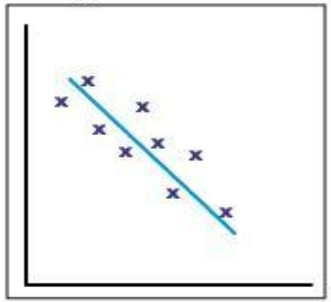
Standard deviation of Y

**Positive correlation**



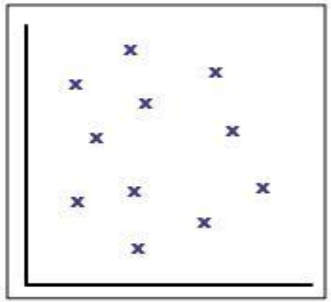
The points lie close to a straight line, which has a positive slope.
This shows that as one variable increases, the other increases.

**Negative correlation**



The points lie close to a straight line, which has a negative slope.
This shows that as one variable increases, the other decreases.

**No correlation**

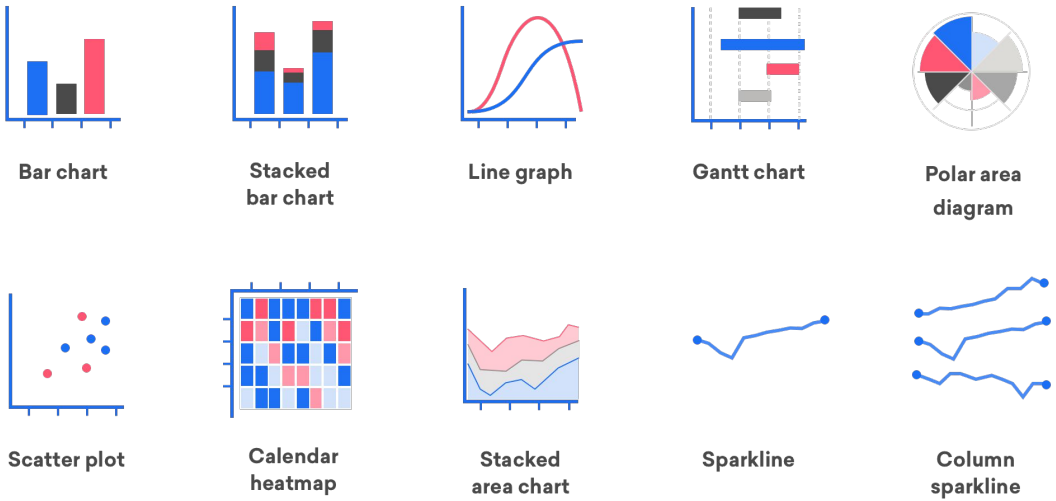


There is no pattern to the points.

This shows that there is no connection between the two variables.

# Introduction to Visualization

- Data visualization is the representation of data or information in a graph, chart, or other visual format to communicate relationships of the data with images.
- We need data visualization because a visual summary of information makes it easier to identify patterns and trends than looking through thousands of rows on a spreadsheet.




Source: https://morphocode.com/location-time-urban-data-visualization/

# Common libraries for Visualization

1. **Matplotlib:**
   - Matplotlib is a popular graphical subroutine and is used widely for data visualization applications.
   - It provides a context, one in which one or more plots can be drawn before the image is shown or saved to file. The context can be accessed via functions on *pyplot*.

2. **Seaborn:**
   - Seaborn is complementary to Matplotlib and it specifically targets statistical data visualization.
   - A saying around matplotlib and seaborn is, "*matplotlib tries to make easy things easy and hard things possible, seaborn tries to make a well-defined set of hard things easy too.*"

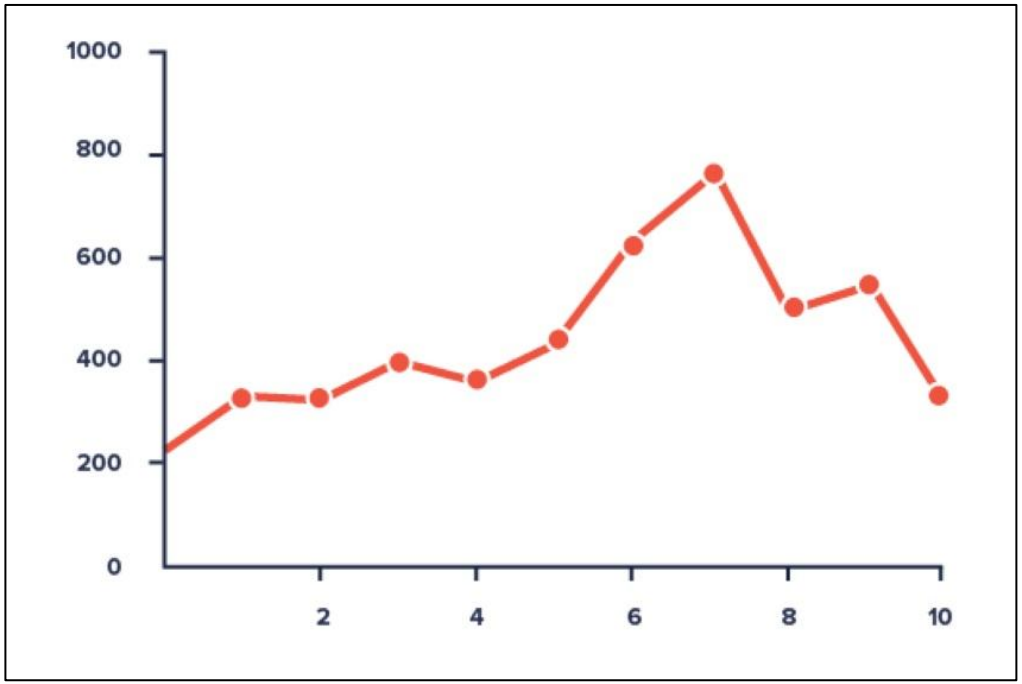
3. **Plotly:**
   - Plotly provides a web-service for hosting graphs.
   - It is mainly used for interactive visualization, dashboards etc.

greatlearning
Power Ahead

# Appendix

# Line Chart

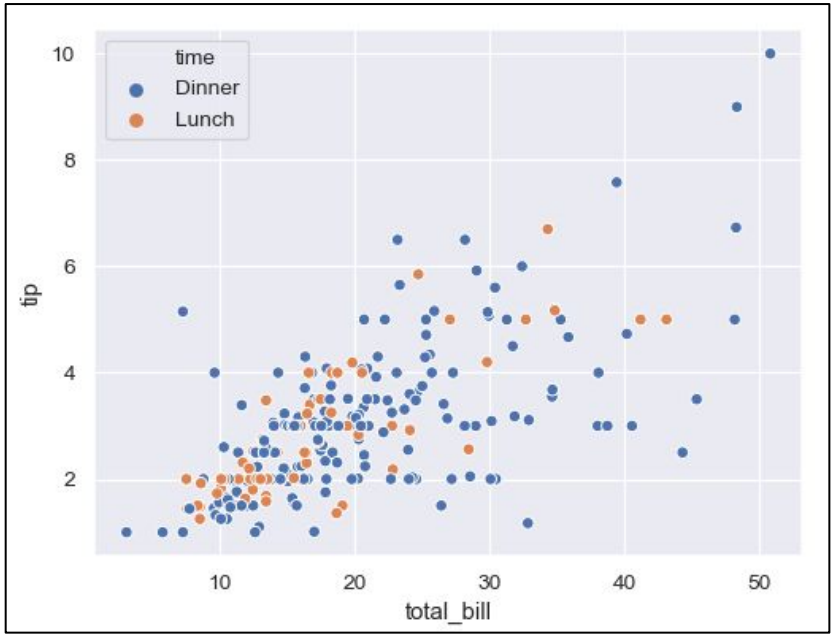
A **line graph** is a graphical display of information that changes continuously over time.




- This plot shows the relationship between the sales and the no. of days
- We can say that sales has been the highest on day 7

# Scatter Plot

- A **scatter plot** uses dots to represent values for two different numeric variables.
- The position of each dot on the horizontal and vertical axis indicates values for an individual data point.
- Scatter plots are used to observe relationships between variables.



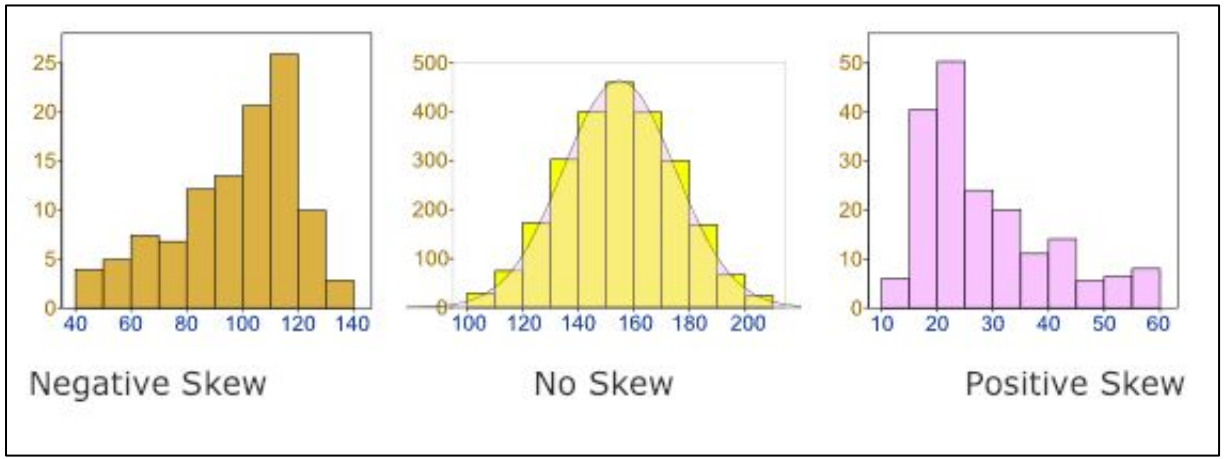

- This plot shows the relationship between the tip and the total bill at the time of lunch and dinner.
- We can say if the total bill is large, the tip can also be large

# Histogram and skewness in data

- A **histogram** is a graphical display of data using bars of different heights.
- In a histogram, each bar groups numbers into ranges

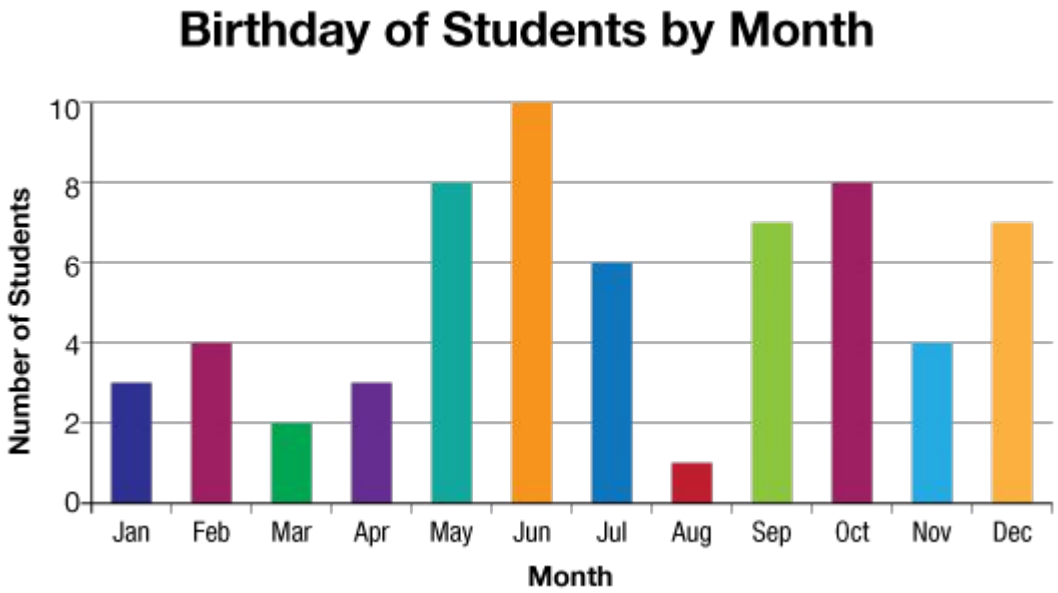
Skewness refers to distortion or asymmetry in a symmetrical bell curve in a set of data

- If the curve is shifted to the left, it is called left skewed. (leftmost curve in the below fig.)
- If the curve is shifted to the right, it is called right skewed. (rightmost curve in the below fig.)



# Bar Plot

- A bar chart is a chart that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.
- The bars can be plotted vertically or horizontally.



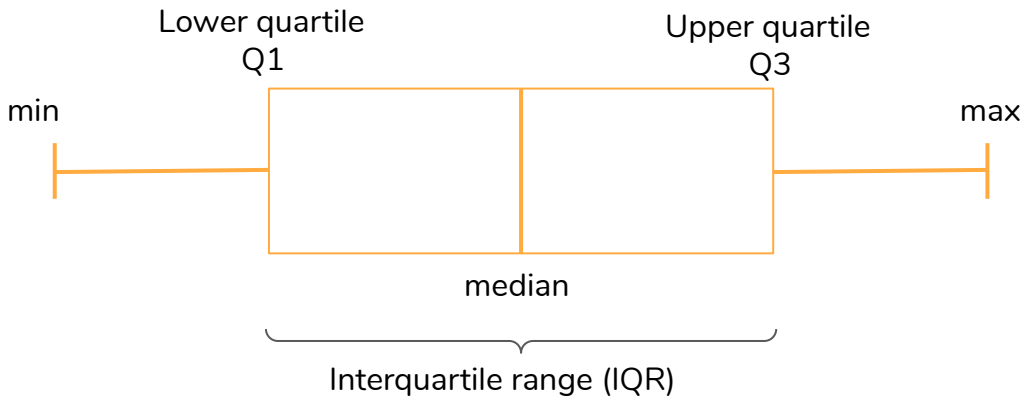

- Most of the students celebrated their birthday in June.
- In August, very less students celebrated their birthdays.

# Five number summary and Box Plot

The five number summary gives you a rough idea about what your data set looks like. It includes five items:

- The minimum.
- Q1 (the first quartile, or the 25% mark)
- The median.
- Q3 (the third quartile, or the 75% mark)
- The maximum.

A box plot is a type of chart often used in exploratory data analysis to visually show the distribution of numerical data and skewness through displaying the data quartiles.



# Chart Selection

| X Variable | Y Variable | Purpose of analysis | Type of chart | Example |
|---|---|---|---|---|
| Continuous (numerical) | Continuous (numerical) | How Y changes with X | Scatter plot | How cholesterol varies with Age? |
| Continuous (numerical) | Categorical | How range of X varies for various category levels | Box plot | Cholesterol variation with Men and Women |
| Categorical | Categorical | What is the number or % of records of X which falls under each category | Stacked bar | How many men have heart disease compared to women? |
| Continuous | - | Look at the distribution of the values of the X variable | Histogram, boxplot | Distribution of cholesterol ranges |
| Impact of 2 X variables on a Y variable | | | Facet_grid() | Distribution of chol across mean and women – compared for people who have and don't have heart disease |

greatlearning
*Power Ahead*

# Happy Learning !