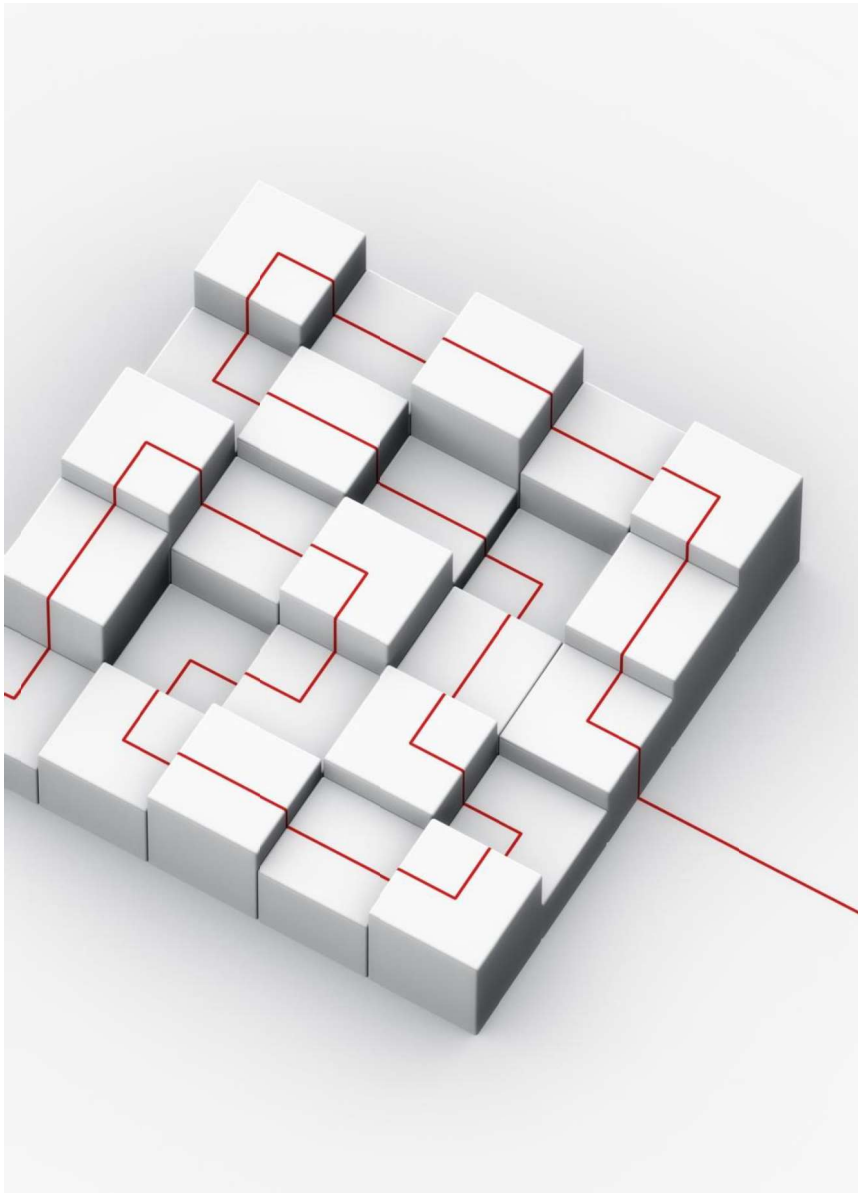
Chapter 01

Introduction to EDA and Descriptive Statistics

Dr. PHAUK Sokkhey
Dr. HAS Sothea

Department of Applied Mathematics and Statistics
Institute of Technology of Cambodia (ITC)

REVIEW ON DESCRIPTIVE STATISTICS



WHAT IS STATISTICS?

- **Statistics** is the branch of mathematics that deals with the ***collection, analysis, interpretation, presentation, and organization of data***.
- It involves the use of various methods and techniques to collect and analyze data, and to draw conclusions from that data.
- Statistics is used in a wide range of fields, including business, economics, social sciences, healthcare, and engineering.

WHAT IS STATISTICS?

- The **main goal** of statistics is to make sense of data by providing a way to describe and summarize the information contained within it.
- This can involve calculating various measures such as averages, standard deviations, and correlations, as well as visualizing the data using graphs and charts.
- Statistics also involves using probability theory to **model** and **analyze** random events, and to make predictions based on data.

STATISTICS

Statistics:

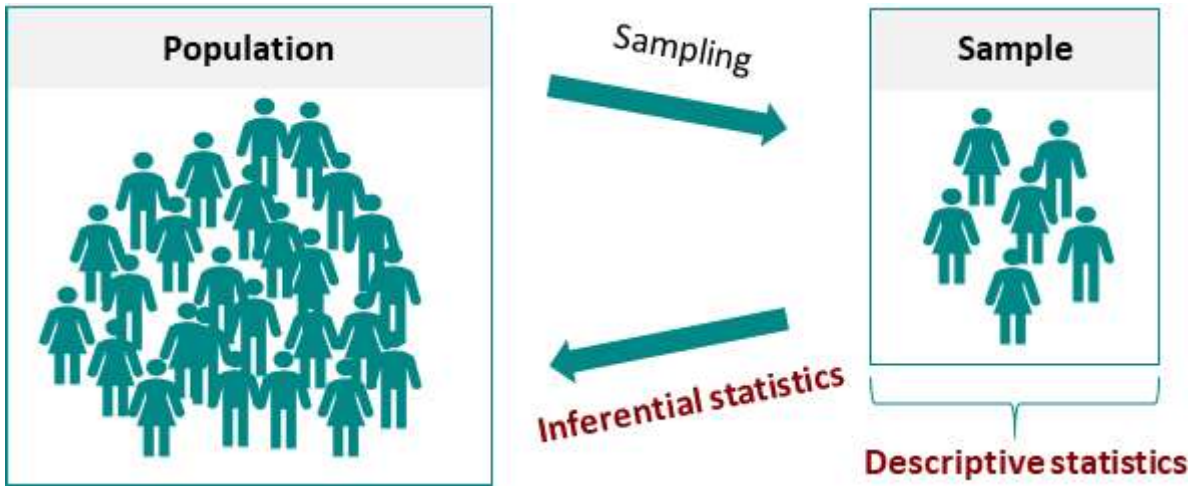
Descriptive Statistics

concerned with summarizing and describing data

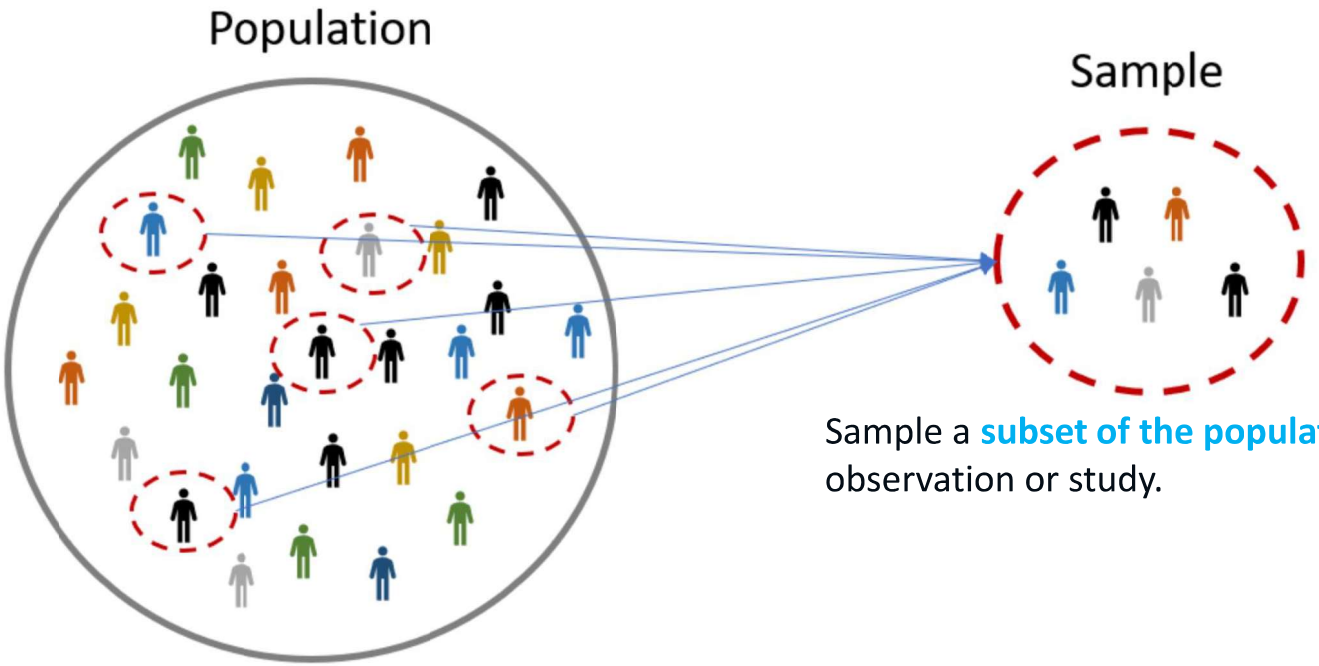
Inferential Statistics

concerned with making predictions and drawing conclusions about a larger population based on a sample of data

DESCRIPTIVE STATISTICS AND INFERENTIAL STATISTICS



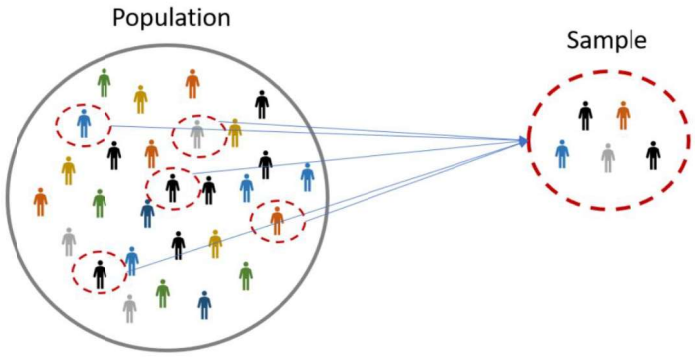
POPULATION VS SAMPLE



Sample a **subset of the population** that is selected for observation or study.

Population is the **entire group** of individuals, objects, or events that we are interested in studying

POPULATION VS SAMPLE



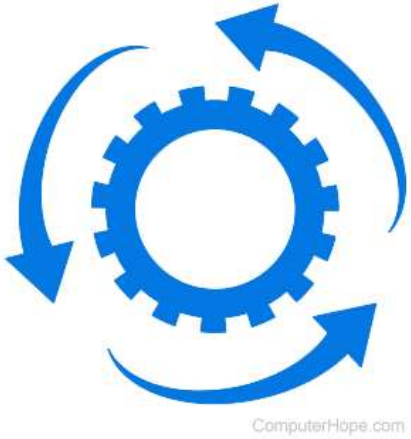
Why do we care of Sample and Population?

References: Population vs Sample

STATISTIC VS PARAMETER

Parameter

is a numerical characteristic of a the population.
For example, mean of the population is a parameter.

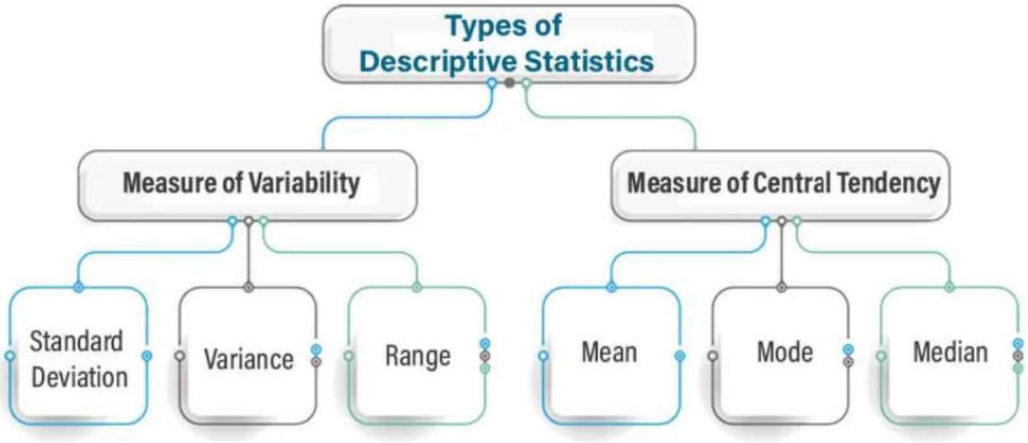


	Parameter	Statistic
Mean	μ mu	$\bar{x}$
Proportion	p	$\hat{p}$
Std. Dev.	σ sigma	s
Correlation	ρ rho	r
Slope	β beta	b

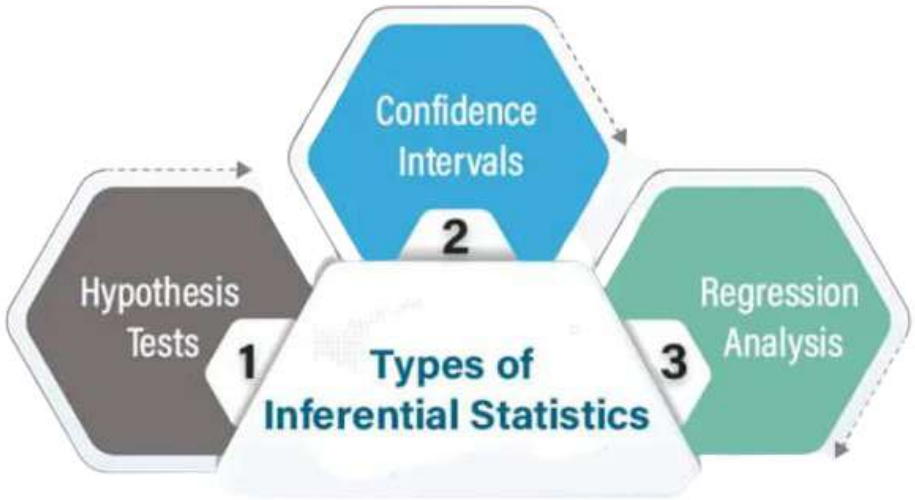
Statistic

is a numerical characteristic of a sample, rather than the population. Examples of statistics include the mean, median, standard deviation, and correlation coefficient.

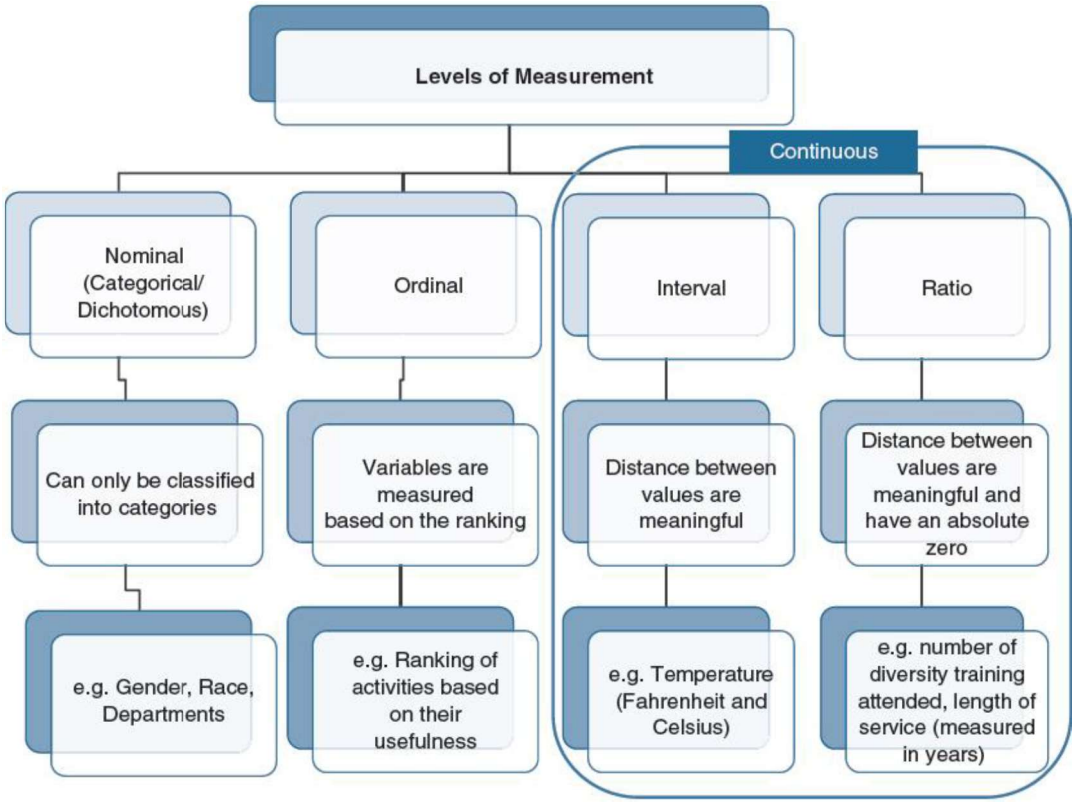
DESCRIPTIVE STATISTICS AND INFERENTIAL STATISTICS



DESCRIPTIVE STATISTICS AND INFERENTIAL STATISTICS

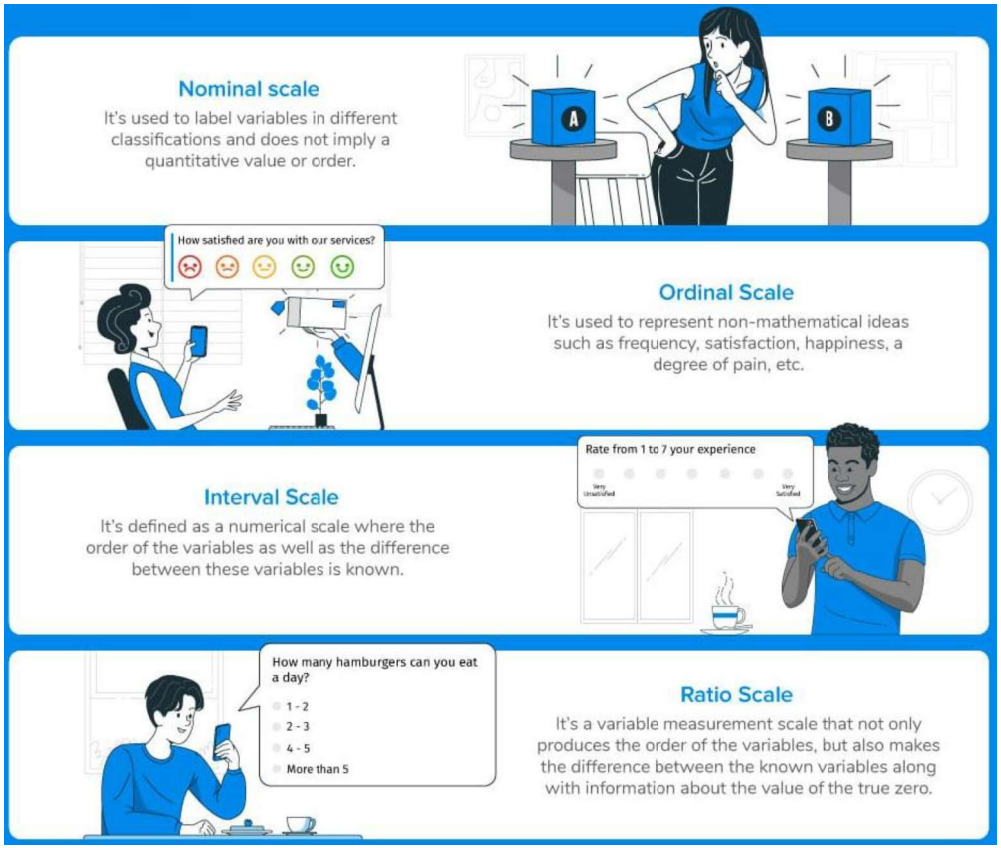


LEVEL OF MEASUREMENT



Watch video: Level of Measurement

EXAMPLE: LEVEL OF MEASUREMENT



TYPES OF DATA: QUANTITATIVE VS QUALITATIVE

Quantitative

Data that can be measured with numbers, such as duration or speed

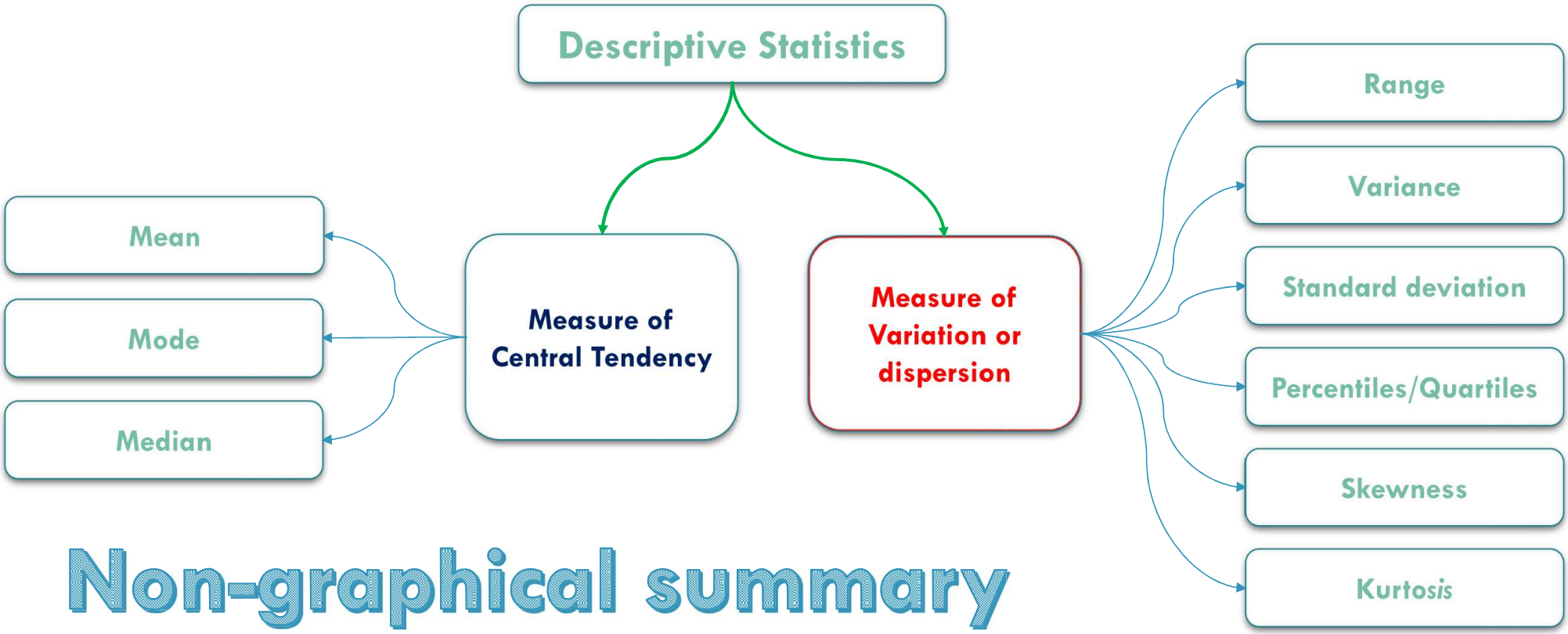
- **Discrete**: Whole numbers that can't be broken down, such as a number of items
- **Continuous**: Numbers that can be broken down, such as height or weight
  - **Interval**: Numbers with known differences between variables, such as time
  - **Ratio**: Numbers that have measurable intervals where difference can be determined, such as height or weight

Qualitative

Non-numerical data that is categorical, such as yes/no responses or eye colour

- **Nominal**: Data used for naming variables, such as hair colour
- **Ordinal**: Data used to describe the order of values, such as 1 = happy, 2 = neutral, 3 = unhappy

DESCRIPTIVE STATISTICS: QUANTITATIVE OR NUMERICAL DATA



Non-graphical summary

DESCRIPTIVE STATISTICS: QUANTITATIVE OR NUMERICAL DATA

- Large amounts of data can be overwhelming.
- A single number can summarize information about a dataset, such as the central tendency or the dispersion of the dataset.
- A common data summary is the arithmetic mean or mean , which is the sum of the data values in a dataset divided by the number of values in the dataset.

MEASURE OF CENTRAL TENDENCY: MEAN (AVERAGE)

- The "**Mean**" is the average of the data.
- Average can be identified by summing up all the numbers and then dividing them by the number of observation.

$$\text{Mean } (x) = \frac{\Sigma x}{n}$$

Example:

Data – 10,20,30,40,50 and Number of observations = 5
Mean = [10+20+30+40+50] / 5
Mean = 30

Outliers influence the central tendency of the data.

MEASURE OF CENTRAL TENDENCY: MEDIAN

- Median is the 50%th percentile of the data. It is exactly the center point of the data.
- Median can be identified by ordering the data and splits the data into two equal parts and find the number.
- It is the best way to find the center of the data.

Median of data doesn't effect by outlier

Example:

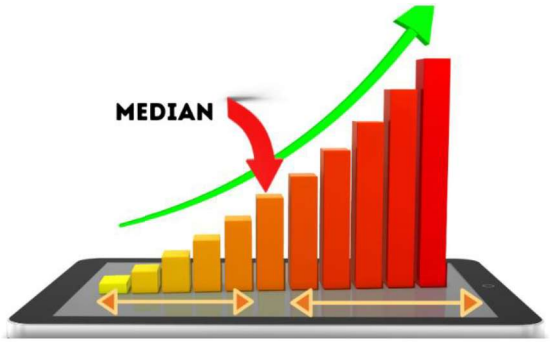
[3, 4, 5, 6, 7]

Median
(Odd number of data)

[3, 4, 5, 6, 7, 8]

(5 + 6) / 2 = 5.5

Median
(Even number of data)



Median Formula

if n is odd,

$$median = \left(\frac{n+1}{2}\right)^{th}$$

if n is even,

$$median = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$$

n = number of terms
th = n(th) number

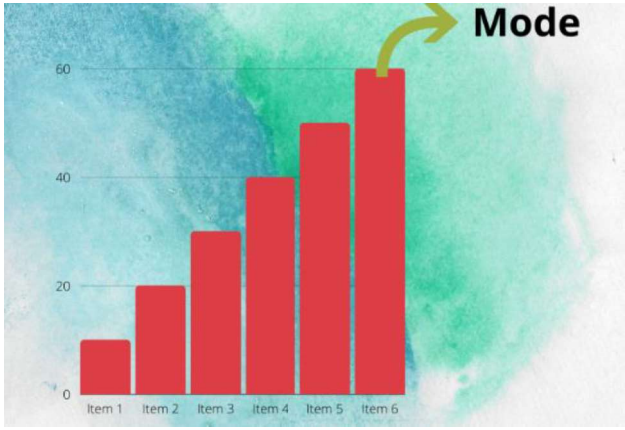
MEASURE OF CENTRAL TENDENCY: MODE

- **Mode** is frequently occurring data or elements.
- If an element occurs the highest number of times, it is the mode of that data. If no number in the data is repeated, then there is no mode for that data.
- There can be more than one mode in a dataset if two values have the same frequency and also the highest frequency.

Outliers don't influence the Mode.
The mode can be calculated for both quantitative and qualitative data.

Example:

Data – 1,3,4,6,7,3,3,5,10,3
Mode is 3
because 3 has the highest frequency (4 times)



MEASURE OF VARIATION OR DISPERSION

Why measure of dispersion is concerned?

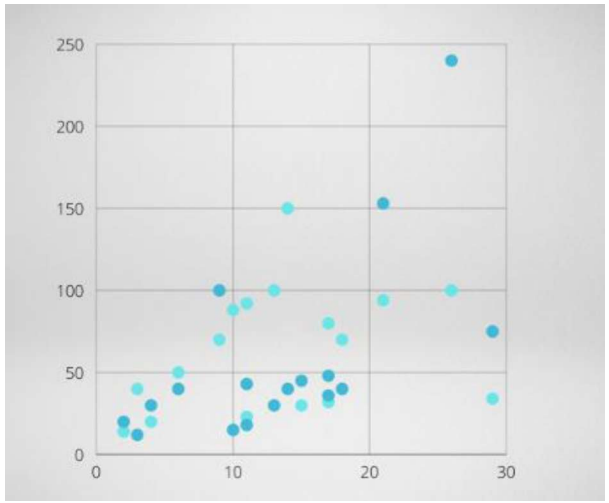
- The dispersion is the "**Spread of the data**". It measures how far the data is spread.
- In most of the dataset, the data values are closely located near the mean.
- On some other dataset, the values are widely spread out of the mean. These dispersions of data can be measured by

Standard deviation

Range

Inter quartile range

Variance



EXAMPLE: WHY MEASURE OF VARIATION OR DISPERSION IS CONCERN?

Money Matters

Two companies, each with 1,000 employees

Both same mean salary: $100K But

C1: Every employee makes $100K — 100M total

C2: Every employee makes $1, CEO $99,999,001 — 100M-999

Which will you join?

Same mean Very different distributions

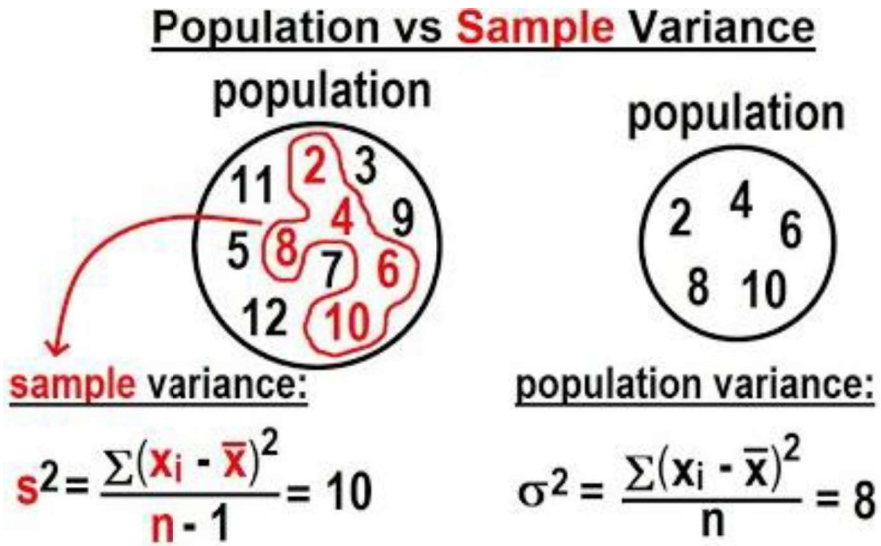
Mean ain't all Variation matters!

VARIANCE

- The variance is a measure of variability. It is the **average squared deviation from the mean**.
- How far a value is from the mean is known as its *deviation*; the variance is the *average* of the squared deviations
- The symbol σ^2 represents the population variance and the symbol for s^2 represents

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$ Population Variance

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$ Sample Variance

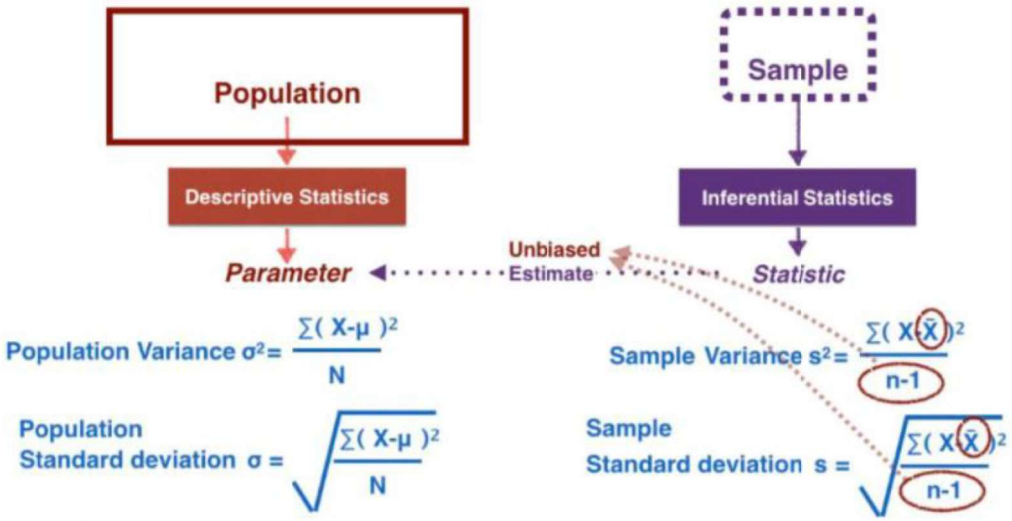


STANDARD DEVIATION

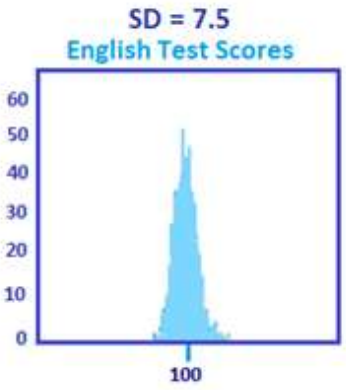
- The variance is a measure of variability. It is the **average squared deviation from the mean**.
- The symbol σ^2 represents the population variance and the symbol for s^2 represents sample variance.

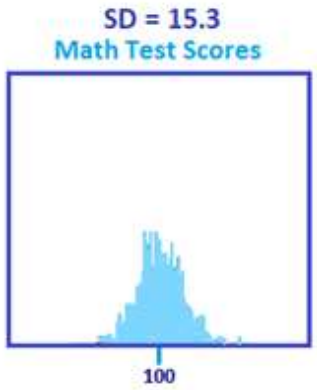
$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$ Population Variance

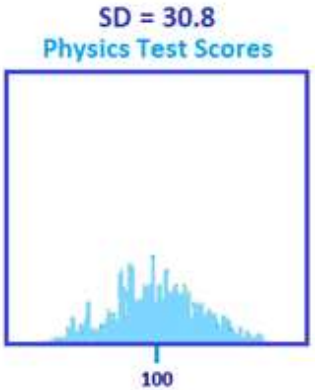
$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$ Sample Variance



STANDARD DEVIATION







What does standard deviation tell you about students' performance in these tests?

EXERCISE: CALCULUS THE SUMMARY STATISTICS FOR SONG SIZE

Song	Artist	Genre	Size (MB)	Length (sec)
My Friends	D. Williams	Alternative	3.83	247
Up the Road	E. Clapton	Rock	5.62	378
Jericho	k.d. lang	Folk	3.48	225
Dirty Blvd.	L. Reed	Rock	3.22	209
Nothingman	Pearl Jam	Rock	4.25	275

What is the Average size of song?

What genre are most popular (mode)?

RANGE

Range = Maximum – Minimum

- Maximum Value = 100th Percentile
- Minimum Value = 0th Percentile
- Another measure of variation; not preferred because based on extreme values. However, it is a good measurement to observe the different and investigate the spread of dataset.

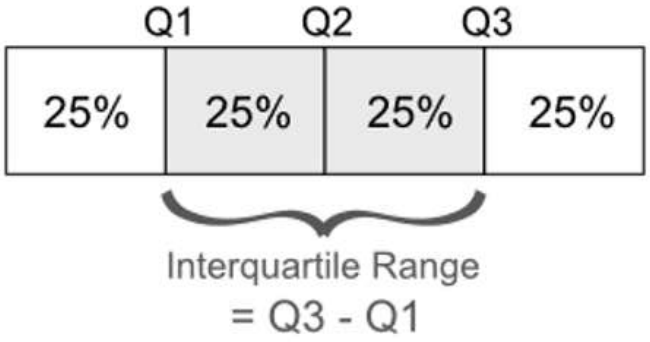
PERCENTILE/QUARTILES

Common Percentiles

- Q3: Upper Quartile = 75^{th} Percentile
- Q1: Lower Quartile = 25^{th} Percentile
- One quarter of the values fall below the lower quartile and one quarter fall above the upper quartile

The Interquartile Range (IQR)

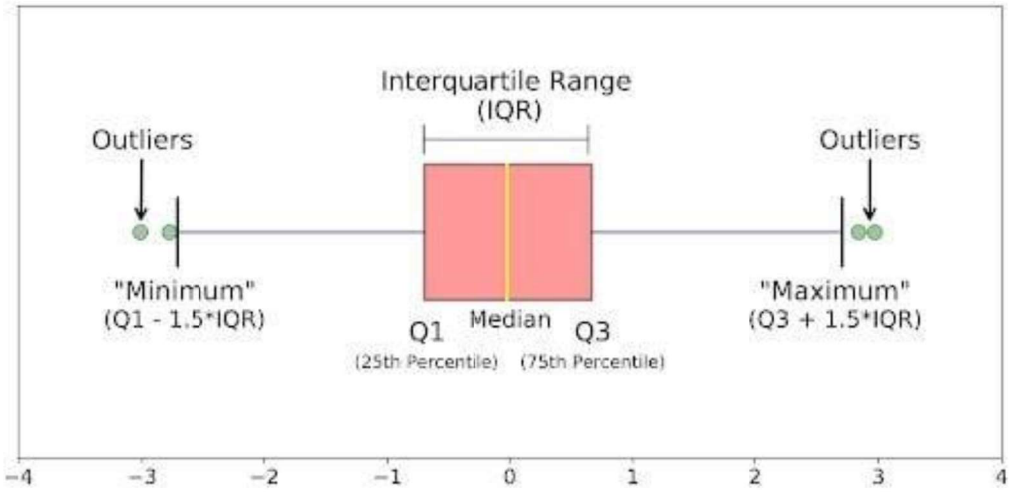
- **IQR** = 75^{th} Percentile – 25^{th} Percentile
- A measure of variation based on quartiles
- Used to accompany the median (Q2)



FIVE NUMBER SUMMARY IN BOXPLOT

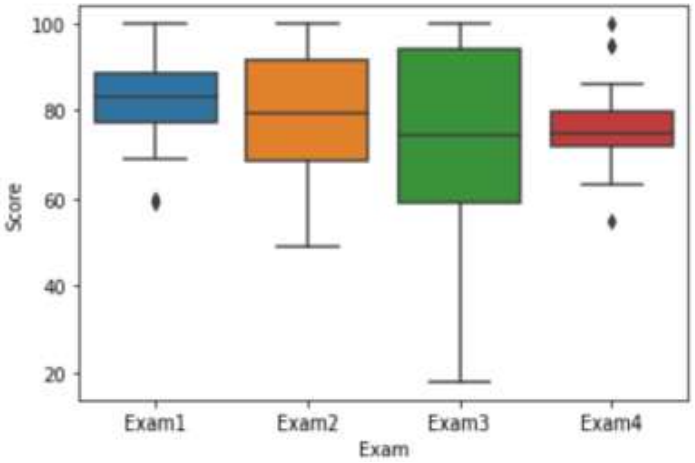
The Five Number Summary

- Maximum
- Upper Quartile (Q3)
- Median
- Lower Quartile (Q1)
- Minimum



Don't confused Max and Min here with Max and Min of the data

FIVE NUMBER SUMMARY IN BOXPLOT



The Five Number Summary

- Maximum
- Upper Quartile (Q3)
- Median
- Lower Quartile (Q1)
- Minimum

Why these values are concerned here?

RANGE

Range = Maximum – Minimum

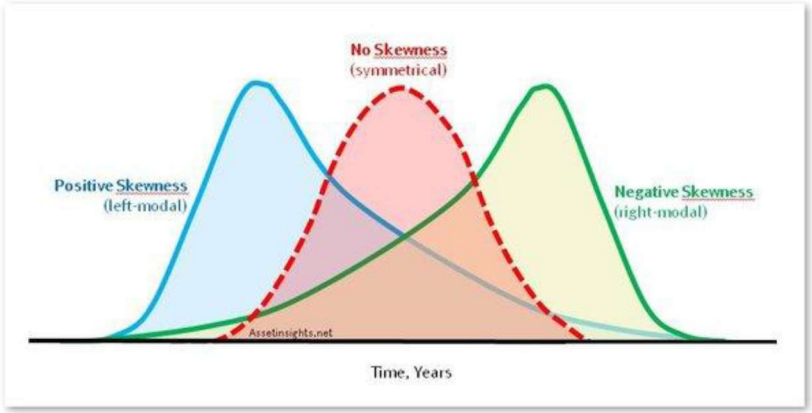
- Maximum Value = 100th Percentile
- Minimum Value = 0th Percentile
- Another measure of variation; not preferred because based on extreme values. However, it is a good measurement to observe the different and investigate the spread of dataset.

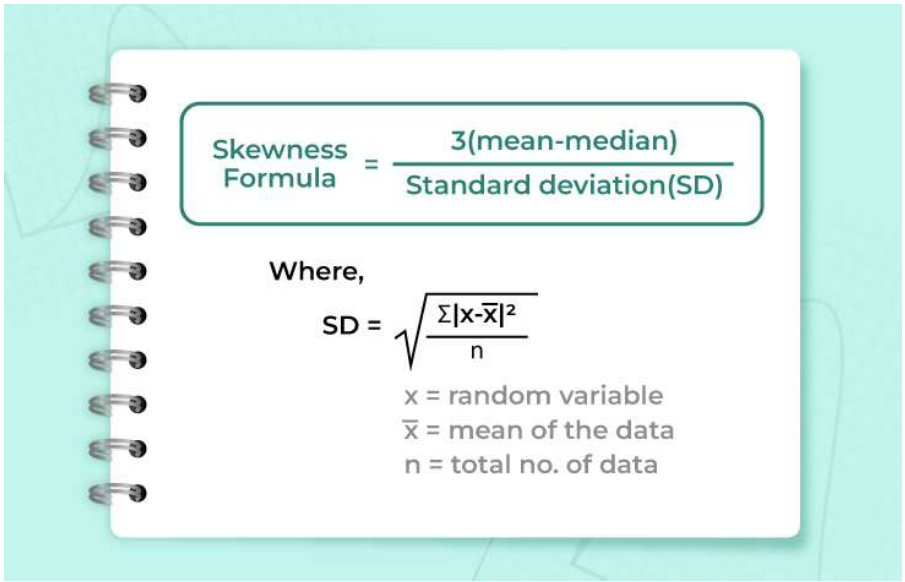
The Five Number Summary

- Maximum
- Upper Quartile (Q3)
- Median
- Lower Quartile (Q1)
- Minimum

SHAPE OF DATA: SKEWNESS

- **Skewness** is a measurement of the distortion of symmetrical distribution or asymmetry in a data set.
- Skewness is demonstrated on a bell curve when data points *are not distributed symmetrically to the left and right sides of the median on a bell curve.*
- If the bell curve is shifted to the left or the right, it is said to be skewed.





As a general rule of thumb: If skewness is less than -1 or greater than 1, the distribution is highly skewed. If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed. If skewness is between -0.5 and 0.5, the distribution is approximately symmetric.

SHAPE OF DATA: KURTOSIS

- Skewness and kurtosis are both measures of the shape of a probability distribution.
- Skewness measures the degree of asymmetry of a distribution, while kurtosis measures the degree of peakedness or flatness of a distribution relative to the normal distribution.

Kurtosis Formula

$$\text{Kurtosis} = n * \frac{\sum_i^n (Y_i - \bar{Y})^4}{\sum_i^n (Y_i - \bar{Y}^2)^2}$$

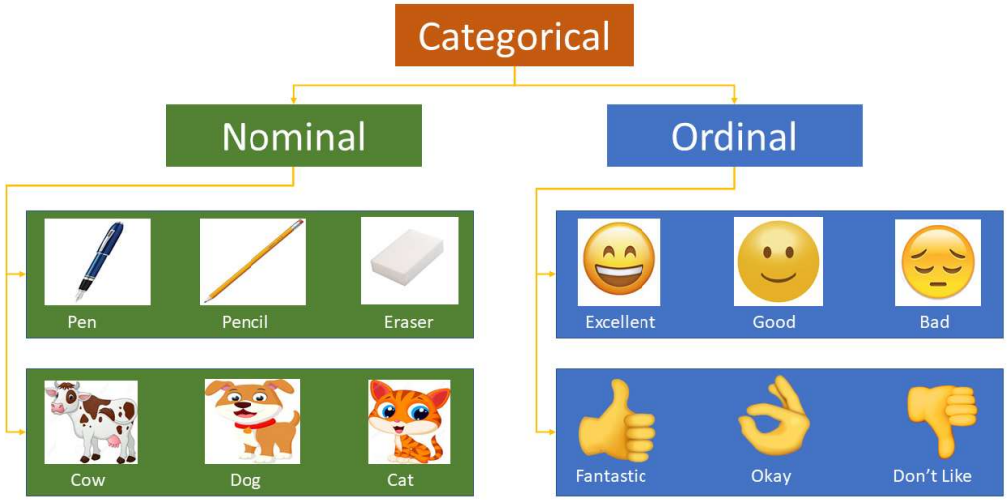
In general, a kurtosis value greater than 3 indicates a distribution that is more peaked than the normal distribution (i.e., leptokurtic), while a kurtosis value less than 3 indicates a distribution that is less peaked than the normal distribution (i.e., platykurtic

GROUP DISCUSSION: SKEWNESS AND KURTOSIS

What is the important of skewness and Kurtosis for Data Analysis?

DESCRIPTIVE STATISTICS: QUALITATIVE OR CATEGORICAL VARIABLES

- One way of displaying data for a single categorical variable is by using a **table** , or in particular a **frequency distribution table**.
- This is a table which displays the various categories for a variable, along with the corresponding frequencies (i.e. how often each category occurs in the data) and usually associated percentages (sometimes including **cumulative percentages** , which give the sum of all the percentages up to and including that row of the table).



Non-graphical summary

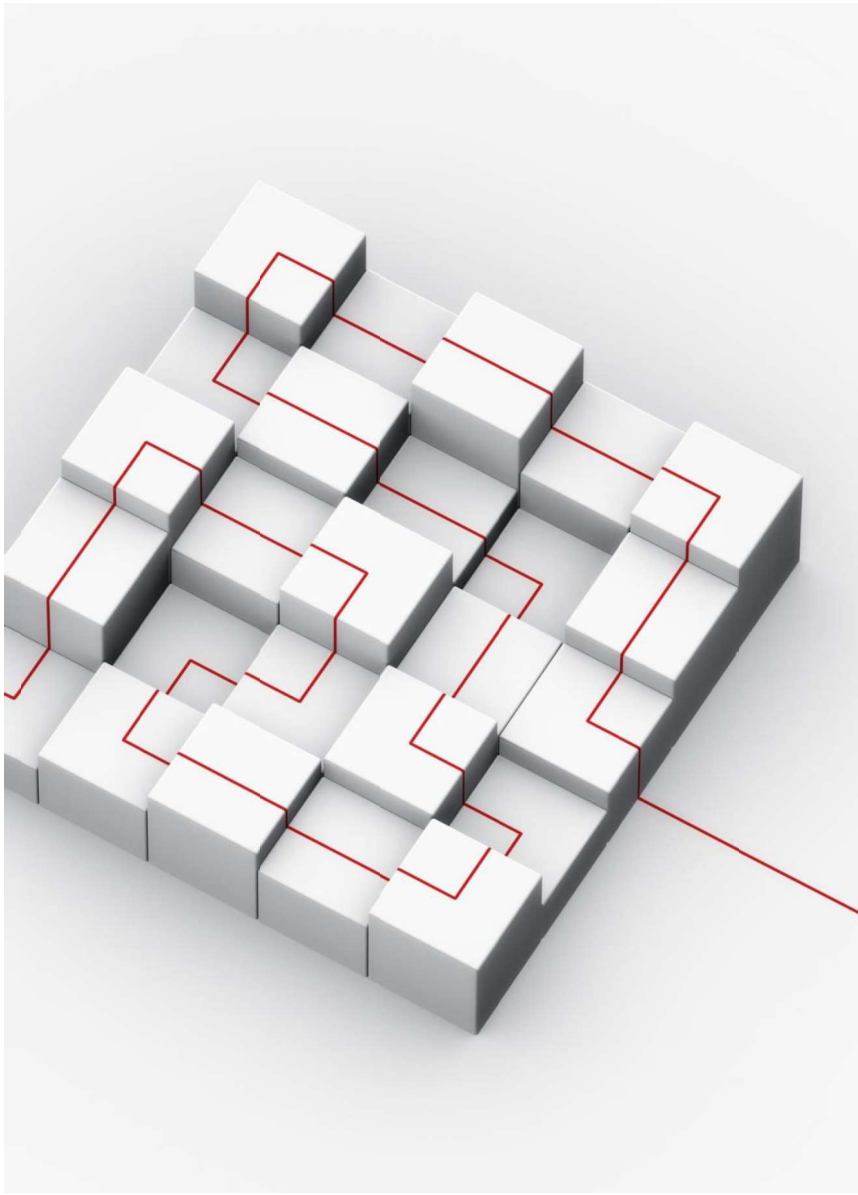
DESCRIPTIVE STATISTICS: QUALITATIVE OR CATEGORICAL VARIABLES

For example, the following is a frequency table showing the frequency of each category of marital status in a sample of 80 people, along with the corresponding percentages:

Marital status

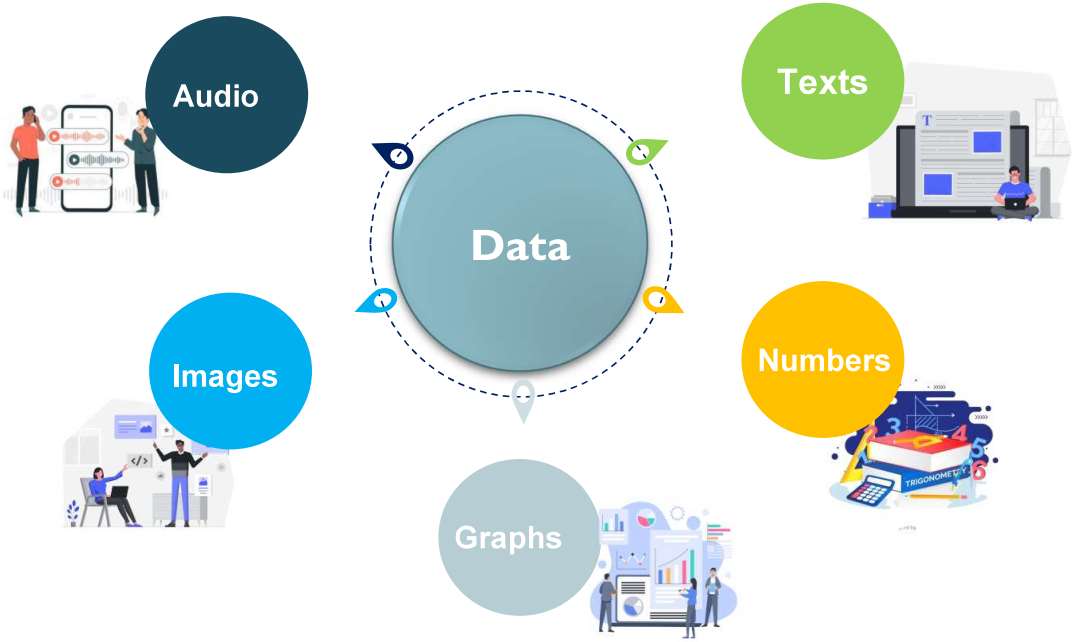
		Frequency	Percent	Cumulative Percent
Valid	Single	44	55.0	55.0
	Married	29	36.3	91.3
	Other	7	8.8	100.0
	Total	80	100.0	

DESCRIPTIVE STATISTICS WITH DATA VISUALIZATION



DATA

- **Data** is information, especially facts, picture, text, audio or numbers, usually collected or computed for purposes of analysis.

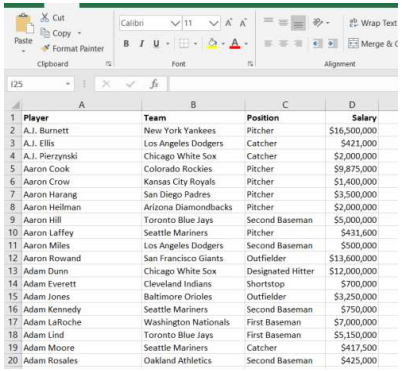


WHAT IS DATA VISUALIZATION?

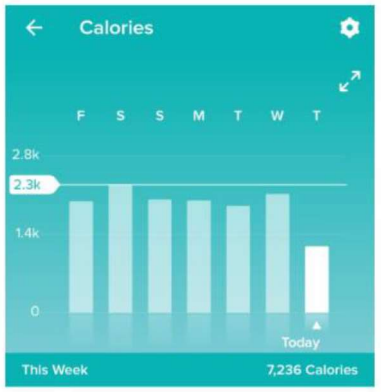
- Data visualization is the graphical or visual representation of data.
- It helps to highlight the most useful insights from a dataset, making it easier to spot trends, patterns, outliers, and correlations.

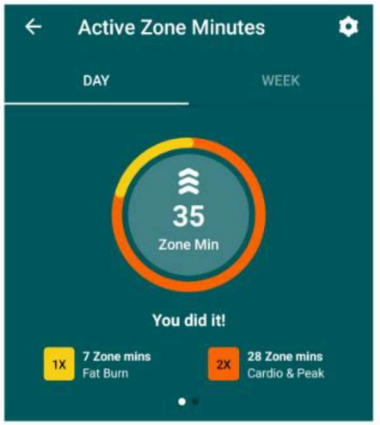
INTRODUCTION

Imagine you're presented with a spreadsheet containing columns and rows of data. You probably won't be able to decipher the data without delving into it, and it's unlikely that you'll be able to spot trends and patterns at first glance. Now imagine seeing the same data presented as a bar chart, or on a color-coded map. It's much easier to see what the data is telling you, right?



	A	B	C	D
1	Player	Team	Position	Salary
2	A.J. Burnett	New York Yankees	Pitcher	$16,500,000
3	A.J. Ellis	Los Angeles Dodgers	Catcher	$421,000
4	A.J. Pierzynski	Chicago White Sox	Catcher	$2,000,000
5	Aaron Cook	Colorado Rockies	Pitcher	$9,875,000
6	Aaron Crow	Kansas City Royals	Pitcher	$1,400,000
7	Aaron Harang	San Diego Padres	Pitcher	$3,500,000
8	Aaron Heilman	Arizona Diamondbacks	Pitcher	$2,000,000
9	Aaron Hill	Toronto Blue Jays	Second Baseman	$5,000,000
10	Aaron Laffey	Seattle Mariners	Pitcher	$431,600
11	Aaron Miles	Los Angeles Dodgers	Second Baseman	$500,000
12	Aaron Rowand	San Francisco Giants	Outfielder	$13,600,000
13	Adam Dunn	Chicago White Sox	Designated Hitter	$12,000,000
14	Adam Everett	Cleveland Indians	Shortstop	$700,000
15	Adam Jones	Baltimore Orioles	Outfielder	$3,250,000
16	Adam Kennedy	Seattle Mariners	Second Baseman	$750,000
17	Adam LaRoche	Washington Nationals	First Baseman	$7,000,000
18	Adam Lind	Toronto Blue Jays	First Baseman	$5,150,000
19	Adam Moore	Seattle Mariners	Catcher	$417,500
20	Adam Rosales	Oakland Athletics	Second Baseman	$425,000





TWO MAIN PURPOSES OF DATA VISUALIZATION

Exploration

When faced with a new dataset, one of the first things you'll do is carry out an **exploratory data analysis**. At this stage, visualizations can make it easier to get a sense of what's in your dataset and to spot any noteworthy trends or anomalies. Ultimately, you're getting an initial lay of the land and finding clues as to what the data might be trying to tell you.

Explanation

Once you've conducted your analysis and have figured out what the data is telling you, you'll want to share these insights with others—key business stakeholders who can take action based on the data, for example, or public audiences who have an interest in your topic area. Explanatory data visualizations help you tell this story, and it's up to you to determine which visualizations will help you to do so most effectively.

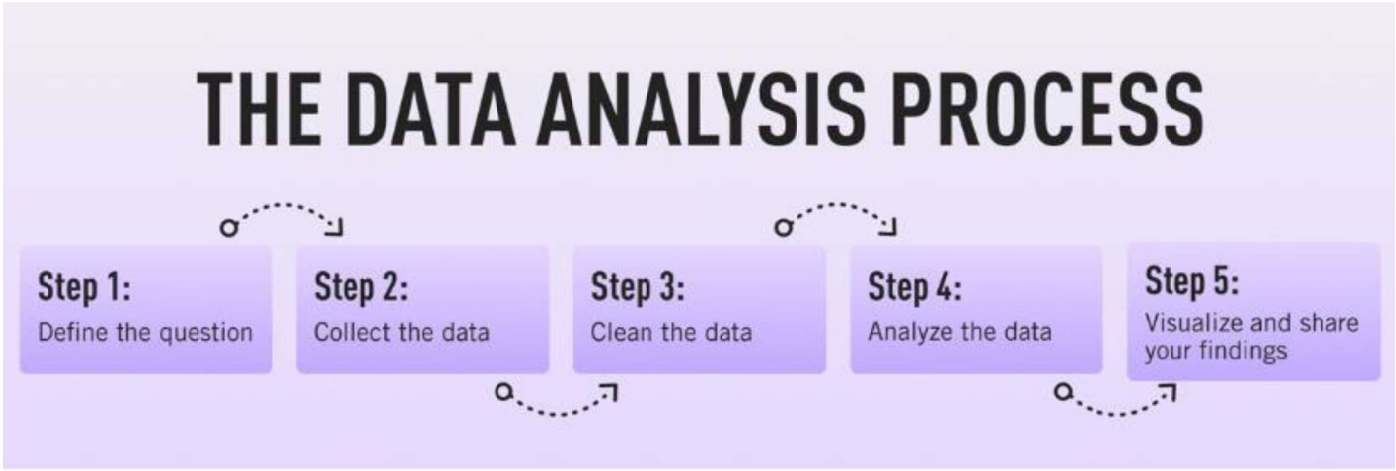
EFFECTIVE DATA VISUALIZATION AT A GLANCE

Data visualization allows you to:

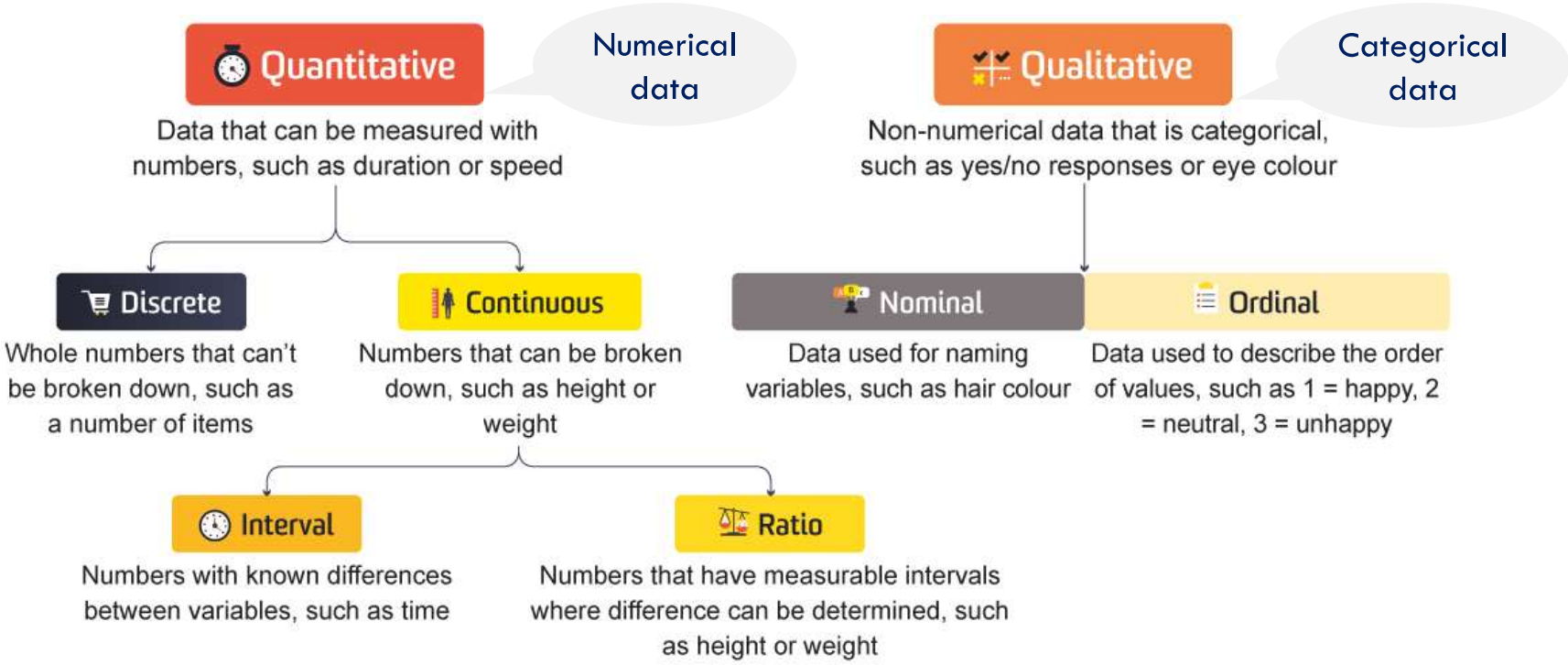
- **Get an initial understanding** of your data by making trends, patterns, and outliers easily visible to the naked eye
- **Comprehend large volumes** of data quickly and efficiently
- **Communicate insights** and findings to non-data experts, making your data accessible and actionable
- **Tell a meaningful** and impactful story, highlighting the most relevant information for a given context

WHEN DO DATA VISUALIZATION

Aside from exploratory data visualization which takes place in the early stages, data visualization usually comprises the final step in **the data analysis process**. To recap, the data analysis process can be set out as follows:

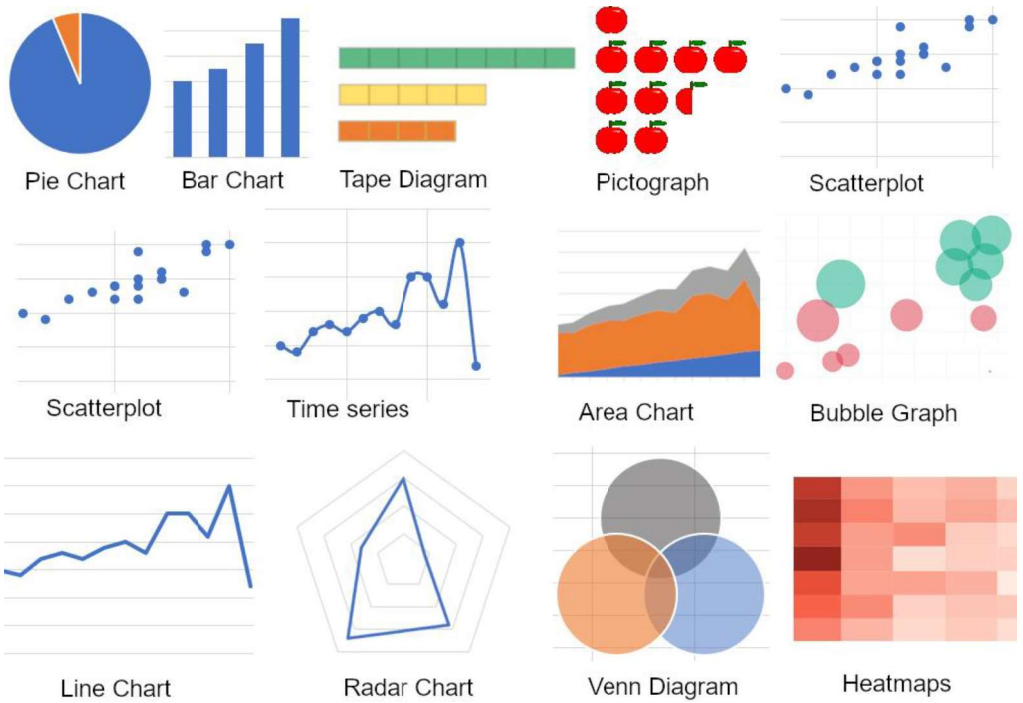


RECALL: DATA VISUALIZATION LINK WITH TYPES OF DATA



COMMON TYPES OF DATA VISUALIZATION

While **descriptive statistics' summaries** are useful for condensing information, **visual summaries** can provide even more context and detail in a small amount of space. Here are examples of graphs that are commonly used:

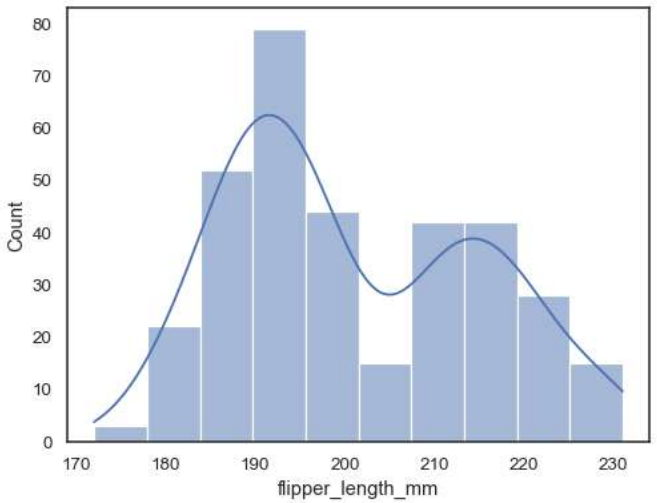


HISTOGRAM

A **histogram** is a graphical representation of the distribution of **numerical data**. It consists of a series of bars, where the height of each bar represents the frequency or relative frequency of data values falling within a particular interval or "bin" on the horizontal axis.

Histogram Uses:

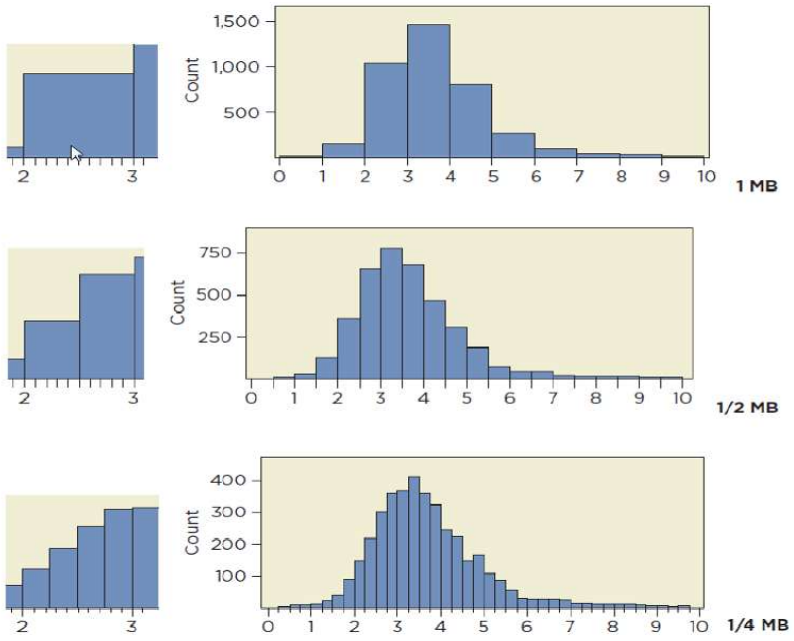
- When **the data is continuous**.
- When you want to represent the shape of the **data's distribution (e.g. normal?)**.
- When you want to see whether the outputs of two or more processes are different.
- To summarize **large data sets** graphically.
- To communicate the data distribution quickly to others.



HISTOGRAM

Histogram of Song Sizes

- Using intervals of different lengths yield different histograms
- Narrow intervals expose details smoothed over by wider intervals
- Most software packages determine the right length to use automatically



Histograms of Song Sizes – Different Intervals

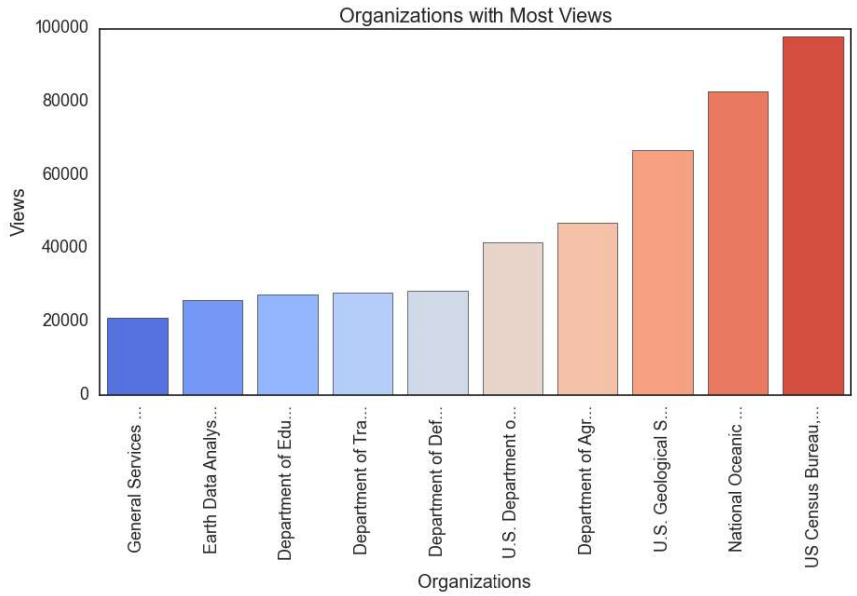
	Method formula bins	Method formula width
Square-root	$\sqrt{n}$	$\frac{\max(values) - \min(values)}{\sqrt{n}}$
Sturges 1926	$\text{ceil}(\log_2 n) + 1$	$\frac{\max(values) - \min(values)}{\text{ceil}(\log_2 n) + 1}$
Rice 1944	$2 * \sqrt[3]{n}$	$\frac{\max(values) - \min(values)}{2 * \sqrt[3]{n}}$
Scott 1979	$\frac{\max(values) - \min(values)}{3.5 * \frac{stdev(values)}{\sqrt[3]{n}}}$	$3.5 * \frac{stdev(values)}{\sqrt[3]{n}}$
Freedman-Diaconis 1981	$\frac{\max(values) - \min(values)}{2 * \frac{IQR(values)}{\sqrt[3]{n}}}$	$2 * \frac{IQR(values)}{\sqrt[3]{n}}$

BAR CHART

Bar charts represent **categorical data** with rectangular bars (to understand what is categorical data see categorical data examples). Bar graphs are among the most popular types of graphs and charts in economics, statistics, marketing, and visualization in digital customer experience.

Bar chart Uses:

- **Bar charts** represent **categorical data** with rectangular bars (to understand what is categorical data see categorical data examples).
- Bar graphs are among the most popular types of graphs and charts in economics, statistics, marketing, and visualization in digital customer experience.



EXAMPLE: LOOKING AT DATA

Which hosts send the most visitors to Amazon's Web site?

- The Data set consists of 188996 visits
- ***Host*** is a categorical variable
- To answer this equation we must describe the variable in ***Host***

For example:

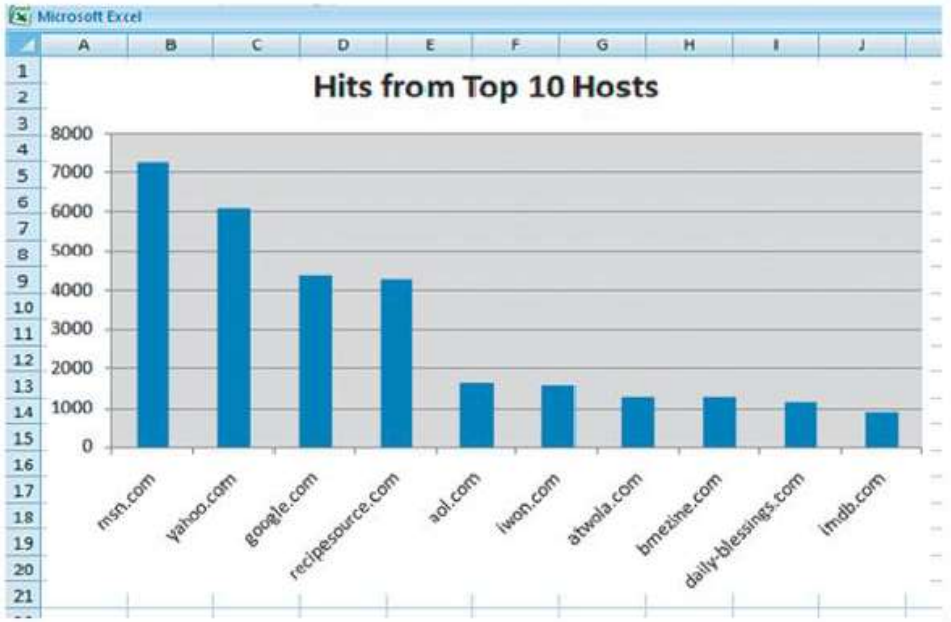
5 observations of visits in Amazon

Purchase	Host	Householdsize	Region	Income Range
No	msn.com	5	North central	35-50 k
No		5	West	50-75 k
No	google.com	4	West	50-75 k
Yes	yahoo.com	3	North Central	35-50 k
No	dealtime.com	4	South	<15 k

EXAMPLE: LOOKING AT DATA

Host	Frequency	Proportion
Typed "amazon.com"	89,919	0.47577
msn.com	7,258	0.03840
yahoo.com	6,078	0.03216
google.com	4,381	0.02318
recipesource.com	4,283	0.02266
aol.com	1,639	0.00867
iwon.com	1,573	0.00832
atwola.com	1,289	0.00682
bmezine.com	1,285	0.00680
daily-blessings.com	1,166	0.00617
imdb.com	886	0.00469
couponmountain.com	813	0.00430
earthlink.net	790	0.00418
popupad.net	589	0.00312
overture.com	586	0.00310
dotcomscoop.com	577	0.00305
netscape.com	544	0.00288
dealtime.com	543	0.00287
att.net	533	0.00282
postcards.org	532	0.00281
24hour-mall.com	503	0.00266
Other	63,229	0.33455
Total	**188,996**	**1.00**

Top 10 Hosts



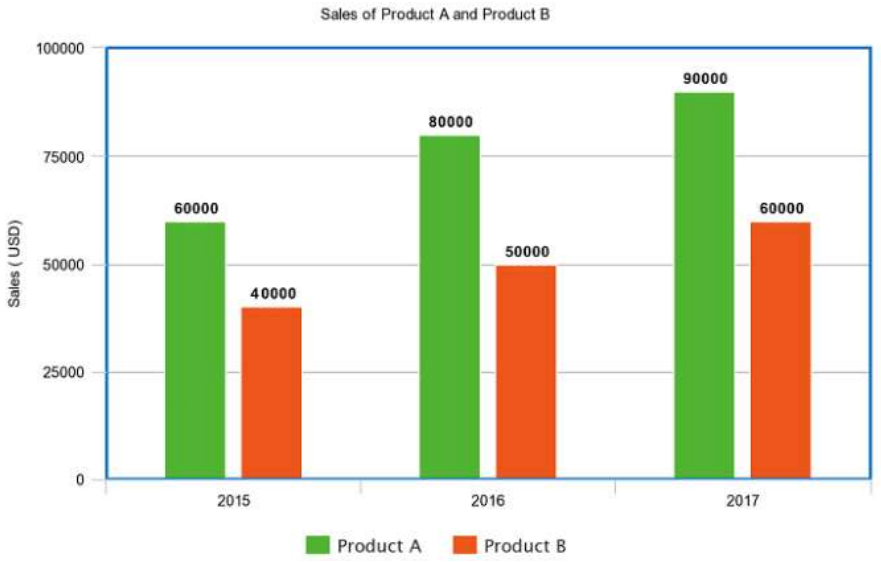
Bar chart for Top 10 Hosts

GROUPED BAR CHART

A **grouped bar chart** is a type of bar chart that displays **multiple sets** of data side-by-side, with each set of data represented by a group of bars. The bars within each group are typically clustered together and are **visually connected,** while the groups themselves are separated by some amount of space.

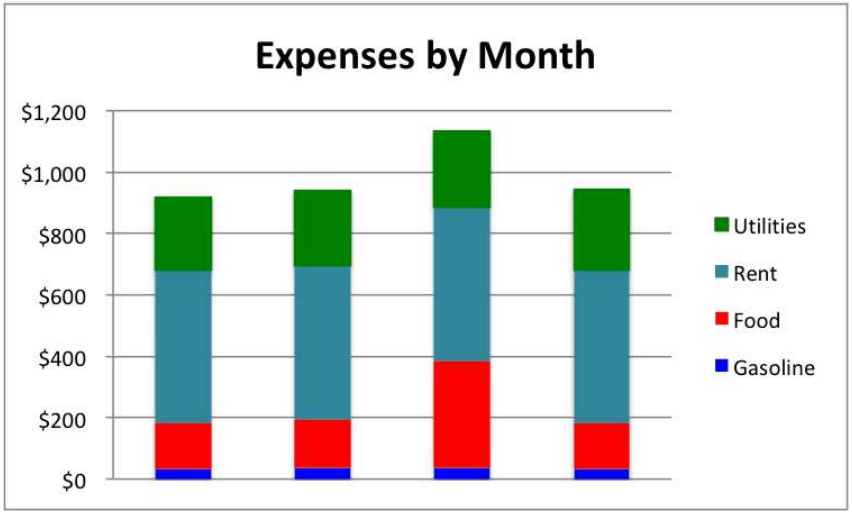
Bar chart Uses:

- Comparing multiple subcategories
- Comparing across multiple groups
- Highlighting differences
- Visualizing trends



STACKED BAR CHART

A **stacked bar chart** is a type of bar chart that displays multiple sets of data stacked on top of each other, with each set of data represented by a segment of the bar. The bars in a stacked bar chart are divided into segments, with each segment representing a different category or subcategory. Each segment is colored differently to represent a different variable or data series.

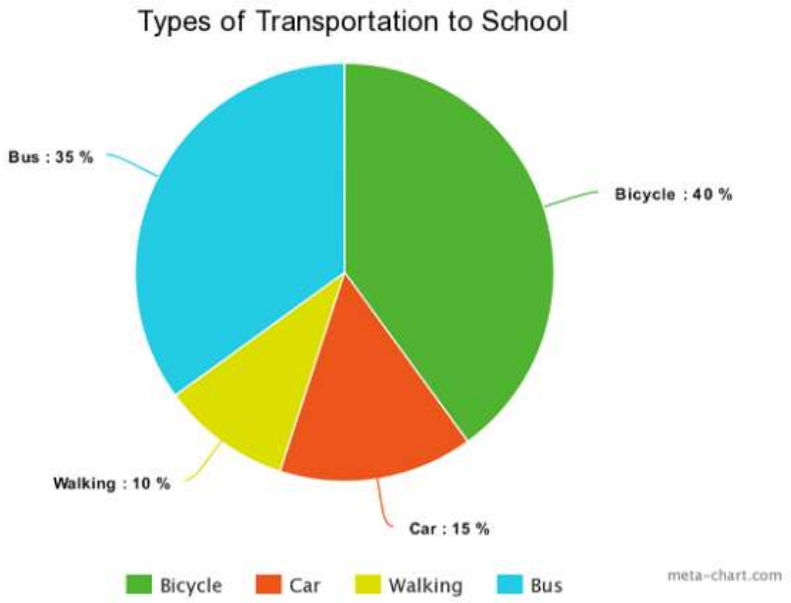


PIE CHART

When it comes to statistical types of graphs and charts, the **pie chart** (or the circle chart) has a crucial place and meaning. It displays data and statistics in an easy-to-understand 'pie-slice' format and illustrates **numerical proportion**.

Pie Chart Uses:

- When you want to create and **represent the composition** of something.
- It is very useful for displaying **nominal or ordinal** categories of data.
- To **show percentage** or proportional data.
- When **comparing areas of growth** within a business such as profit.
- Pie charts work best for displaying data for **3 to *7*** categories.

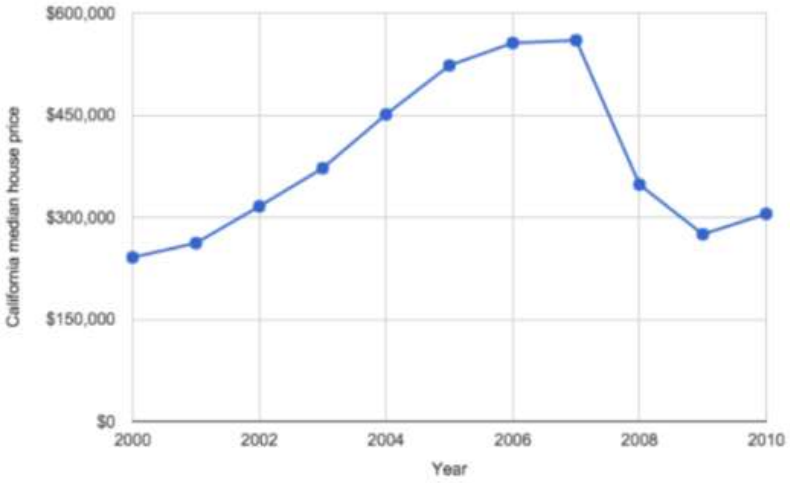


LINE GRAPH

A line graph, also known as a line chart, is a type of graph that displays data as a series of points connected by straight lines. Line graphs have an x-axis and a y-axis. In the most cases, time is distributed on the horizontal axis. The graph typically has two axes, with the horizontal axis representing the independent variable (such as time, distance, or category) and the vertical axis representing the dependent variable (such as temperature, sales, or quantity)

Line Graphs Uses:

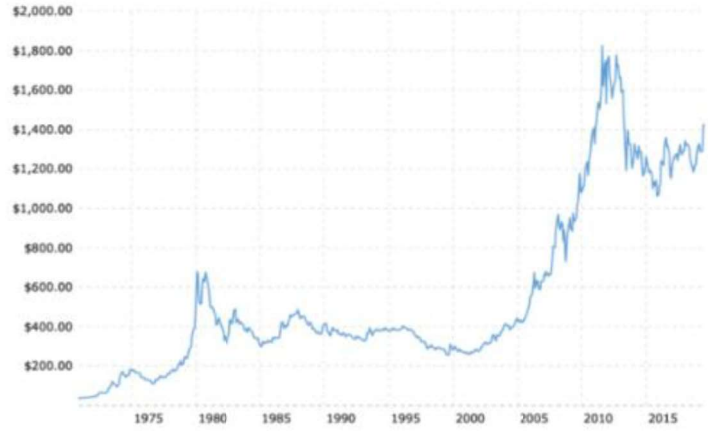
- When you want **to show trends**. For example, how house prices have increased over time or categories.
- When you want **to make predictions** based on a data history over time.
- When **comparing** two or more different variables, situations, and information over a given period of time.



TIME SERIES GRAPH

Time series graphs and line graphs are both types of graphs used to visualize how a variable changes over time. Time series graphs are specifically designed to display time-related data, while line graphs can be used to display any type of data.

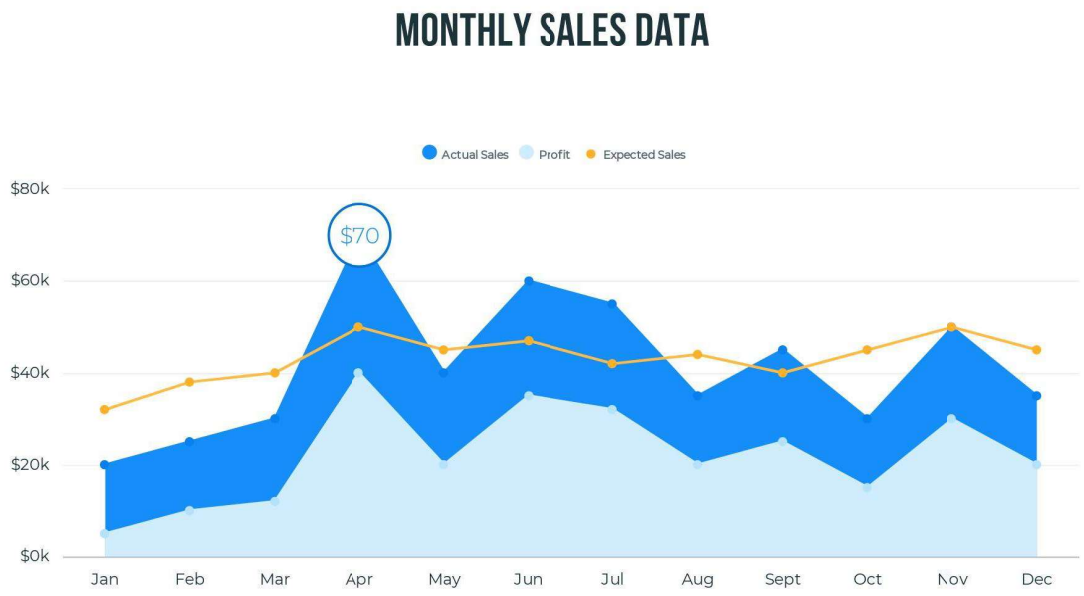
1.Trend analysis: Time series graphs are particularly effective for identifying trends and patterns in data over time.
2.Seasonal analysis: Time series graphs can also be useful for identifying seasonal patterns in data.
3.Forecasting: Time series graphs can be used to make predictions about future trends or patterns in data.
4.Comparing multiple variables: Time series graphs can be effective for comparing how multiple variables change over time.



Credit: Macrotrends

AREA CHART

Area charts show the change in one or **several quantities over time**. They are very similar to the line chart. However, the area between axis and line are usually filled with colors.

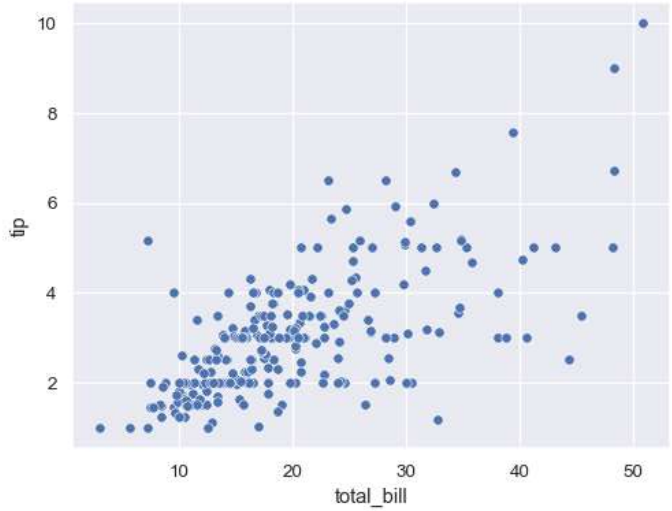


SCATTER PLOT

The **scatter plot** is an X-Y diagram that shows a **relationship between two variables**. It is used to plot data points on a vertical and a horizontal axis. The purpose is to show **how much one variable affects another**.

Scatter plot uses:

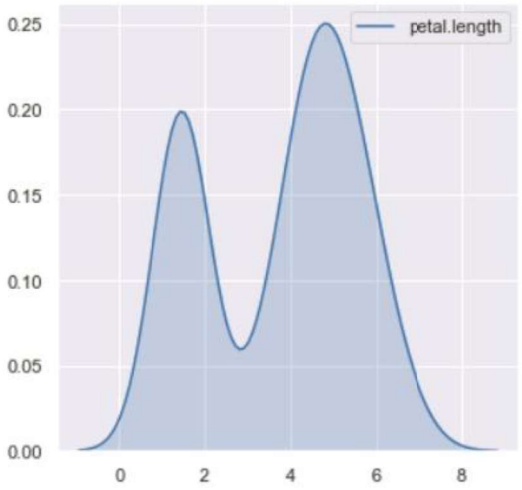
- When trying to find out whether there is a **relationship between 2 variables**.
- **To predict** the behavior of dependent variable based on the measure of the independent variable.
- When having **paired numerical data**.
- When working with root cause analysis tools to identify the potential for problems.
- When you just want to visualize the correlation between 2 large datasets.



DENSITY PLOT

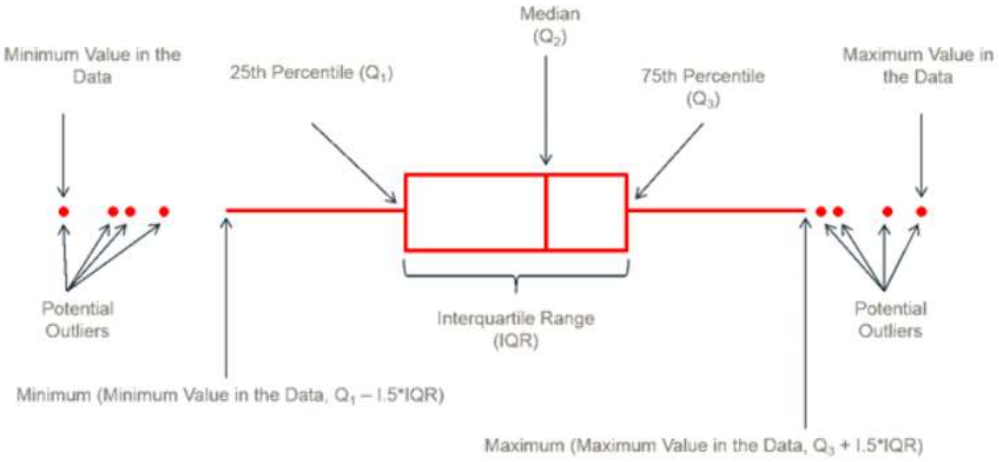
A **density plot** is like a smoother version of a histogram. Generally, the kernel density estimate is used in density plots to show the probability density function of the variable. A continuous curve, which is the kernel is drawn to generate a smooth density estimation for the whole data.

- **Identify the shape of the distribution:** Density plots can help you to identify the shape of the distribution of a continuous variable. For example, a normal distribution will have a symmetrical bell-shaped curve, while a bimodal distribution will have two peaks.
- **Compare distributions:** Density plots can be used to compare the distributions of different variables or subsets of the same variable. By overlaying multiple density plots on the same graph, you can easily see how the distributions differ or overlap.
- **Identify outliers:** Density plots can be used to identify outliers or extreme values in the data. Outliers will be visible as points outside the main body of the distribution.

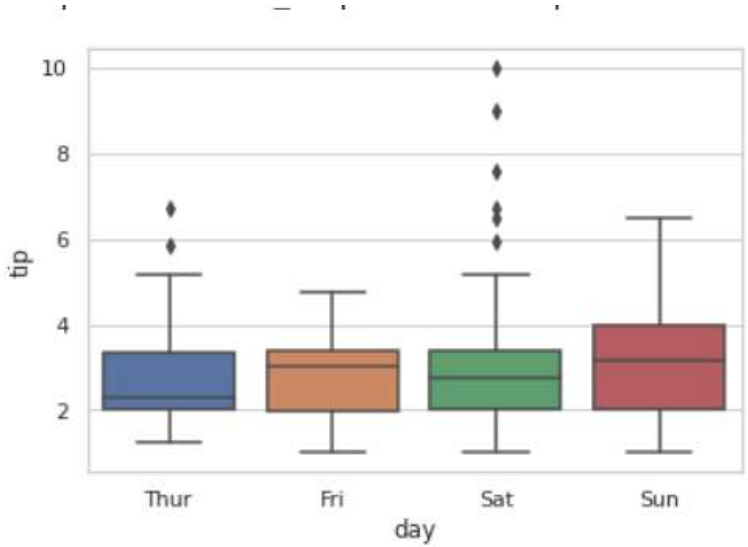


BOX PLOTS

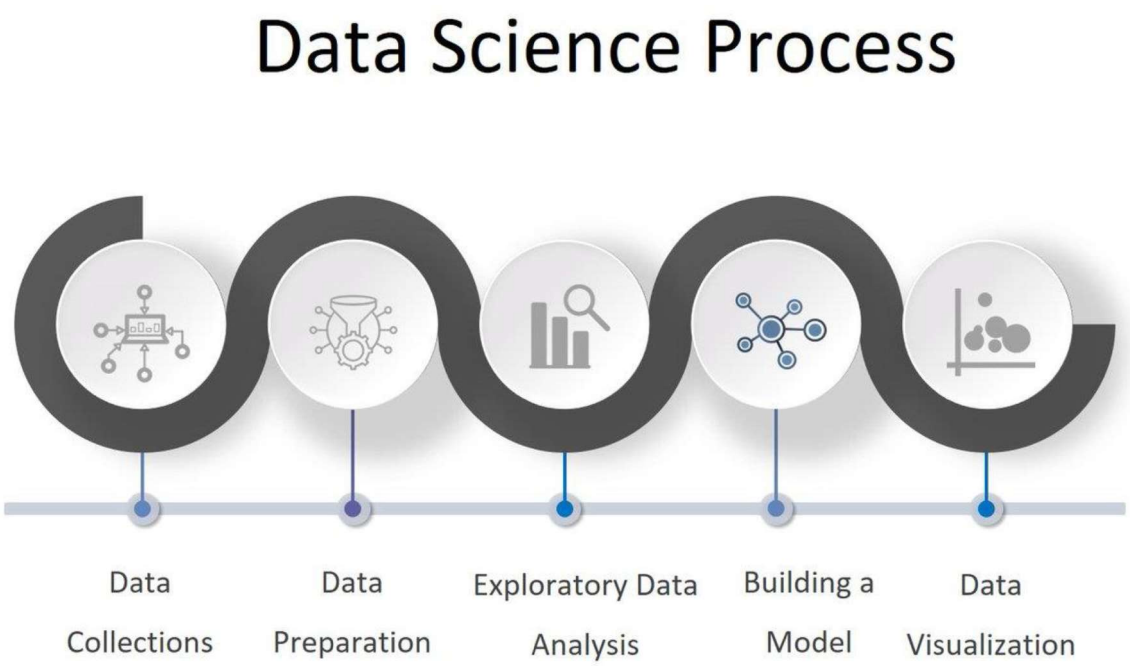
A **box-plot** is a very useful and standardized way of **displaying the distribution** of data based on a five-number summary (minimum, first quartile, second quartile(median), third quartile, maximum). It helps in understanding these parameters of the distribution of data and is extremely helpful in **detecting outliers**.



(Source:leansigmacorporation.com)



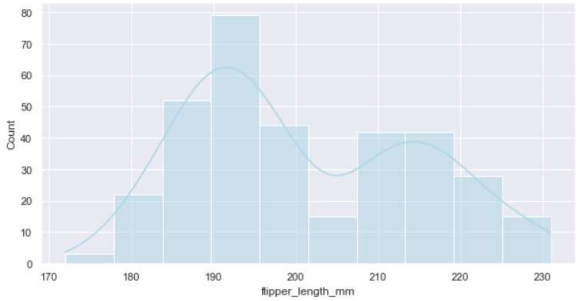
UNIVARIATE AND MULTIVARIATE ANALYSIS



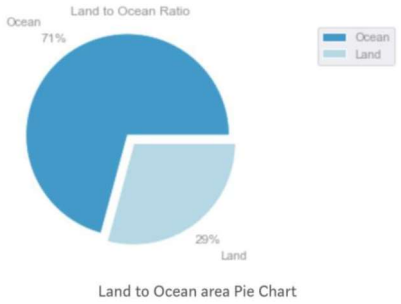
Exploratory Data Analysis (EDA) is a crucial method used in statistics and data science to analyze and understand datasets is majorly performed using the following methods:

- **Univariate analysis:** provides summary statistics for each field in the raw data set (or) summary only on one variable.
- **Bivariate analysis:** is performed to find the relationship between each variable in the dataset and the target variable of interest (or) using 2 variables and finding the relationship between them.
- **Multivariate analysis:** is performed to understand interactions between different fields in the dataset (or) finding interactions between variables more than 2.

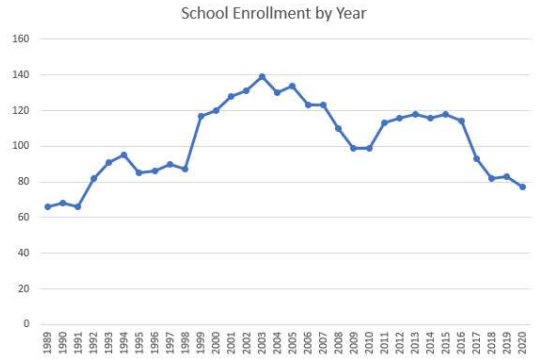
UNIVARIATE ANALYSIS



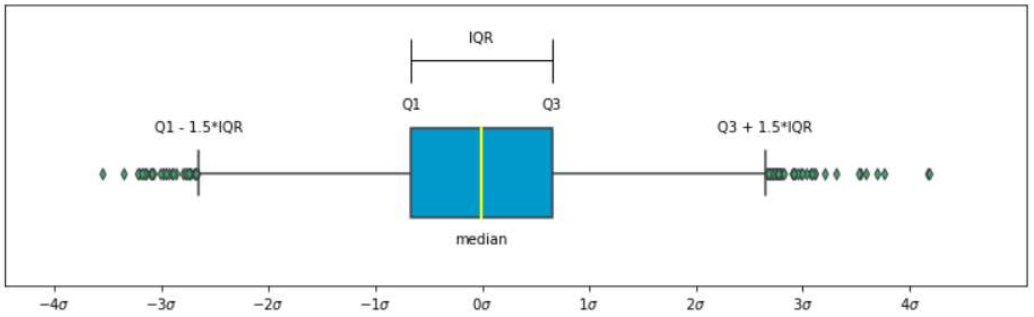
Bar chart



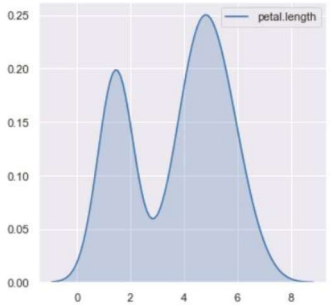
Pie chart



Line and time series graph



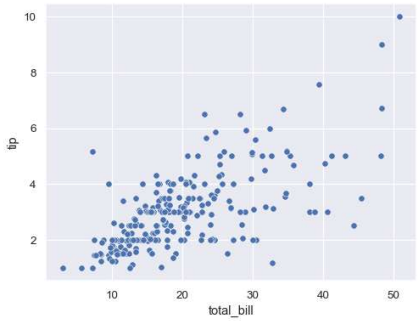
Boxplot



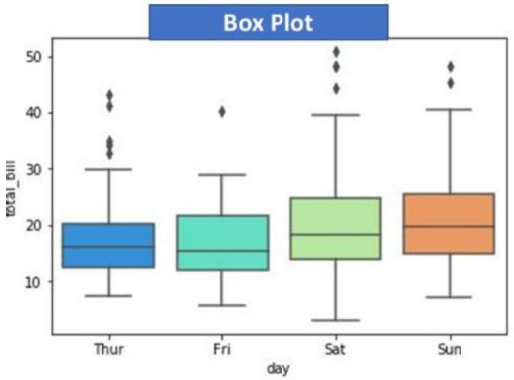
Density plot

Univariate analysis: Link

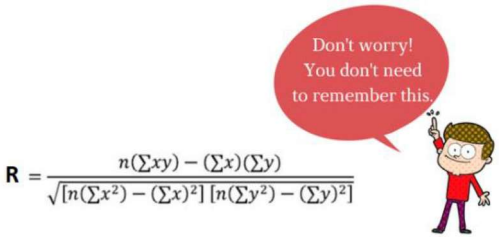
BIVARIATE ANALYSIS



Scatter plot



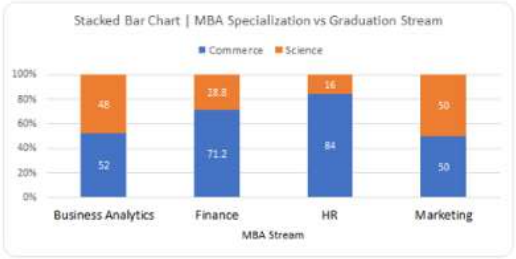
Multiple boxplot



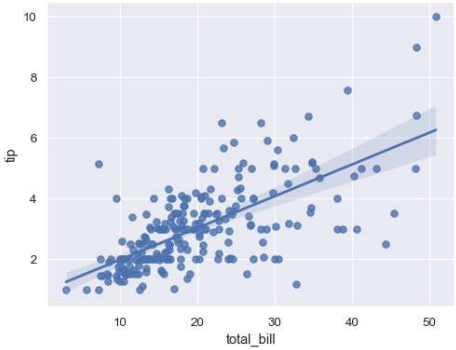
Correlation coefficient

CROSSTAB		MBA Specialization				
		Business Analytics	Finance	HR	Marketing	Row Total
Graduation Stream	Commerce	13	57	21	35	126
	Science	12	23	4	35	74
	Column Total	25	80	25	70	200

Crosstab



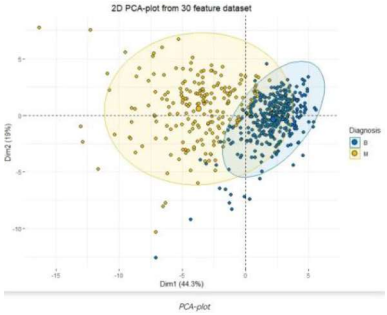
Stacked bar plot



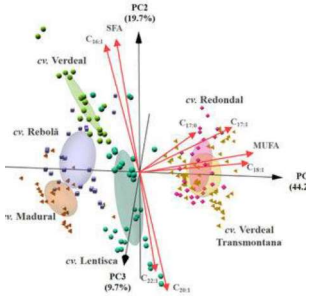
Linear regression

Bivariate analysis: Link

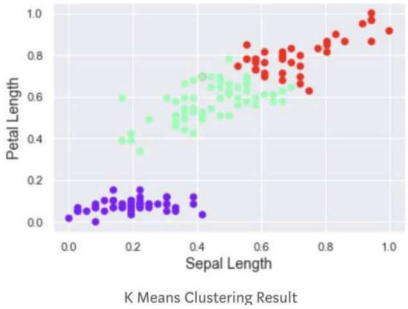
MULTIVARIATE ANALYSIS



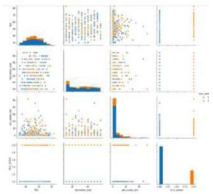
Principal Component Analysis (PCA)



Linear Discriminant Analysis (LDA)

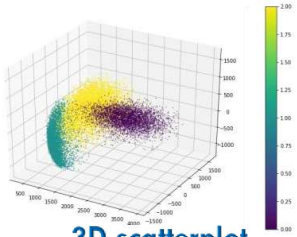


Clustering Analysis

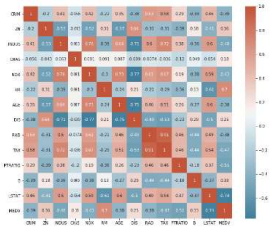


Pair plot

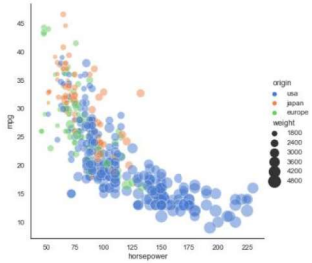
Line and time series graph



3D scatterplot



Heatmap



Bubble chart

Multivariate analysis: Link

Q & A