DATA ANALYSIS & PREPROCESSING

Machine Learning – Yousef Elbaroudy

GUIDELINES

- Try to **SHUTDOWN YOUR PHONE** and focus on the important information mentioned through the session
- Understand the concepts that will be discussed, not to memorize it
- Apply what you understood in the practical session
- Don’t mind to ask about anything you want to know

Enjoy the Session ☺

SUMMER TRAINING PROJECT

Choose whether to be in groups (Teams) or individual

Try to implement the project in **INCREMENTAL WAY**

Starting with the next session (Exploratory Data Analysis, Data Cleaning and Preprocessing)

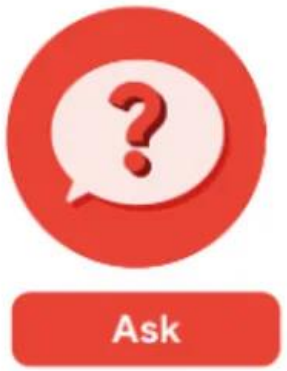
Whenever the Machine Learning Algorithms will be discussed; implement your required algorithm on your workflow and so on.

The criteria for the project evaluation will be explained next session

Prepare your ideas …

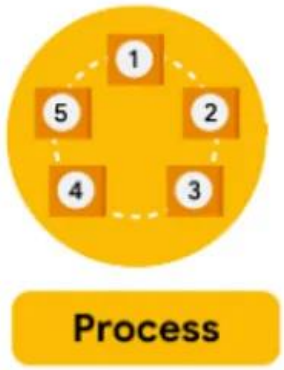
1. **Ask**: Business Challenge/Objective/Question
2. **Prepare**: Data generation, collection, storage, and data management
3. **Process**: Data cleaning/data integrity
4. **Analyze**: Data exploration, visualization, and analysis
5. **Share**: Communicating and interpreting results
6. **Act**: Putting your insights to work to solve the problem

WHAT IS DATA ANALYSIS ? (GOOGLE DATA ANALYTICS)

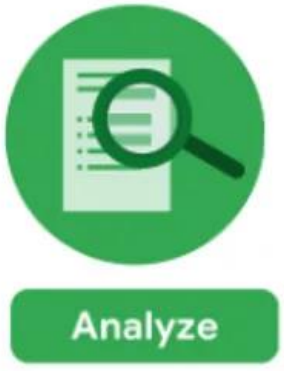


Ask questions and define the problem.

Prepare data by collecting and storing the information.

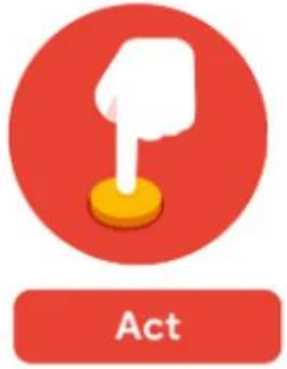


Process data by cleaning and checking the information.



Analyze data to find patterns, relationships, and trends.

Share data with your audience.



Act on the data and use the analysis results.

IN OTHER WORDS (UDACITY)

1. Question
2. Wrangle
3. Explore
4. Draw Conclusions
5. Communicate

Step 1: Ask Questions

- Given data then ask questions, or
- Ask questions then **gather** data

Step 2: Wrangle Data

a. **Gather** data to answer question

b. **Assess** data to identify any problems in your data's quality or structure

c. **Clean** data by modifying, replacing, or removing data

Step 3: Perform Exploratory Data Analysis (EDA)

- **Explore then augment** data to maximize the potential of
  - analyses & visualizations & models
- **Exploring** involves:
  - finding **patterns** in data
  - **visualizing** relationships in data
  - building **intuition** about what you're working with
- **After Exploring (optional)**
  - **Remove Outliers:**
  - **Feature Engineering:** create better features from data

Step 4: Draw Conclusions (or even make predictions)

- typically approached with **ML** or **inferential statistics**

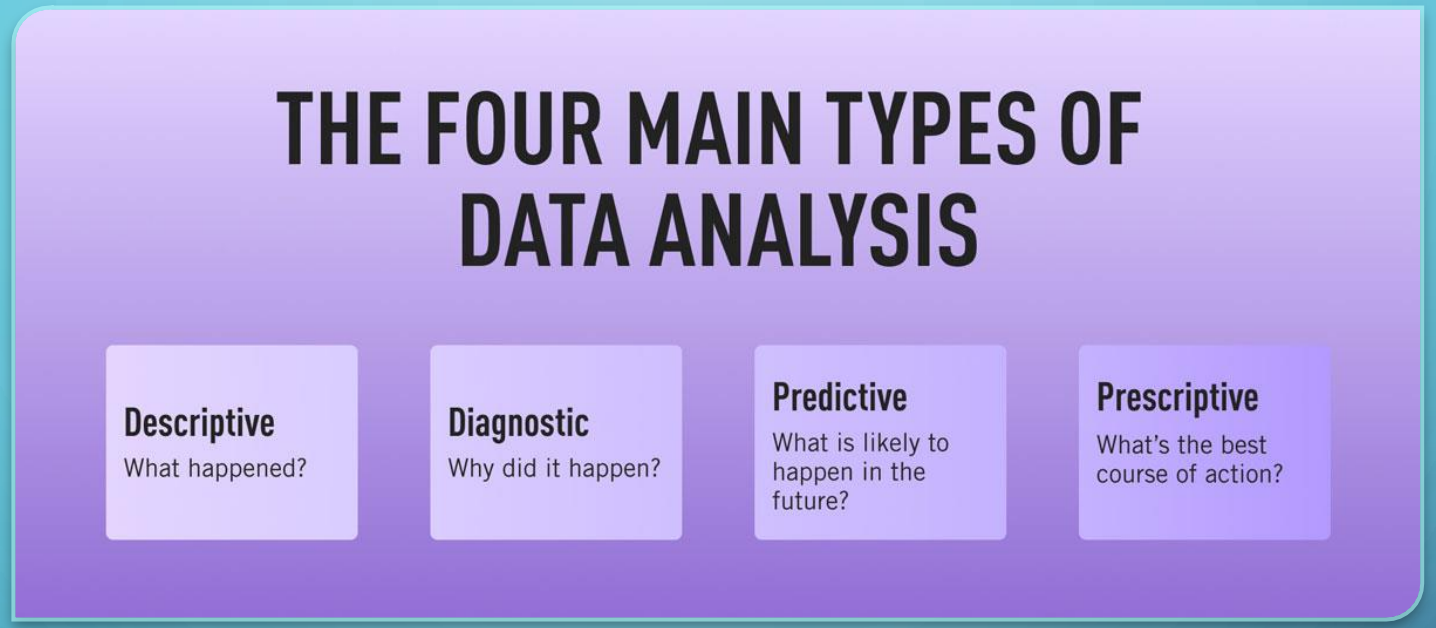
Step 5: Communicate Results

- often need to **justify** and **convey** meaning in the insights
- if your end goal is to build a system, you usually need to:
  - **share** what you've built
  - **explain** how you reached design decisions
  - **report** how well it performs
- communicate results by: report | slides | presentation | post | email | conversation
- **Data Visualization** will always be very valuable

Data + business knowledge = mystery solved

Blending data with business knowledge, plus maybe a touch of gut instinct, will be a common part of your process as a junior data analyst. The key is figuring out the exact mix for each particular project. A lot of times, it will depend on the goals of your analysis. That is why analysts often ask, “How do I define success for this project?”

TYPES OF ANALYSIS



THAT BRINGS US INTO "DATA QUALITY"

WHAT IS THE DIMENSIONS OF “DATA QUALITY” ?

- Data Quality is a measurement of the degree to which data is fit for purpose. Good data quality generates trust in data. Data Quality Dimensions are a measurement of a specific attribute of a data's quality.

Completeness

Completeness measures the degree to which all expected records in a dataset are present. At a data element level, completeness is the degree to which all records have data populated when expected.

COMPLETENESS

CustomerID	CustomerName	CustomerBirthDate	CustomerAccountType	CustomerAccountBalance	LatestAccountOpenDate
100000192	Robert Brown	4/12/2000	Loan	40390.00	12/20/2026
100000198	Maria Irving	12/1/2025	Deposit	-13280.00	10/21/2018
100000120	Ava Shiffer	10/31/1990	Credit Card	320	3/1/2020
100000192	Robert Brown	4/12/2000	Deposit	40390.00	12/20/2026
100000124	Matthew Martin	5/9/1965	Deposit	70102.00	5/4/2022
100000149		2/4/1988	Loan	0.00	9/20/1990

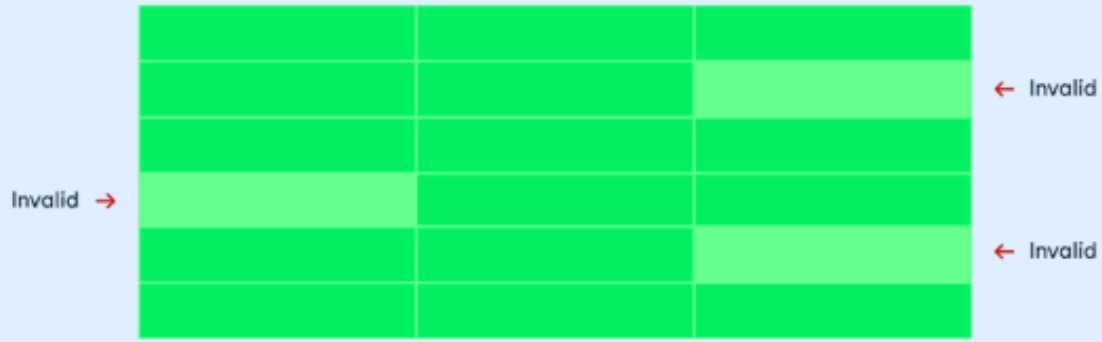
All records must have a value populated in the CustomerName field.

EXAMPLE

VALIDITY

- **Validity**

Validity measures the degree to which the values in a data element are valid.



CustomerID	CustomerName	CustomerBirthDate	CustomerAccountType	CustomerAccountBalance	LatestAccountOpenDate
100000192	Robert Brown	4/12/2000	Loan	40390.00	12/20/2026
100000198	Maria Irving	12/1/2025	Deposit	-13280.00	10/21/2018
100000120	Ava Shiffer	10/31/1990	Credit Card	320	3/1/2020
100000192	Robert Brown	4/12/2000	Deposit	40390.00	12/20/2026
100000124	Matthew Martin	5/9/1965	Deposit	70102.00	5/4/2022
100000149		2/4/1988	Loan	0.00	9/20/1990

- **Validity Example**
  - CustomerBirthDate value must be a date in the past.
  - CustomerAccountType value must be either Loan or Deposit.
  - LatestAccountOpenDate value must be a date in the past.

EXAMPLE

UNIQUENESS

- **Uniqueness**

Uniqueness measures the degree to which the records in a dataset are not duplicated.

All records must have a unique CustomerID and CustomerName.

CustomerID	CustomerName	CustomerBirthDate	CustomerAccountType	CustomerAccountBalance	LatestAccountOpenDate
100000192	Robert Brown	4/12/2000	Loan	40590.00	12/20/2026
100000198	Maria Irving	12/1/2025	Deposit	-13280.00	10/21/2018
100000120	Ava Shiffer	10/31/1990	Credit Card	320	3/1/2020
100000192	Robert Brown	4/12/2000	Deposit	40590.00	12/20/2026
100000124	Matthew Martin	5/9/1965	Deposit	70102.00	5/4/2022
100000149		2/4/1988	Loan	0.00	9/20/1990

EXAMPLE

TIMELINESS

SLA	Table Load Time
08:00 am	07:59 am
10:00 am	09:59 am
11:00 am	11:01 am

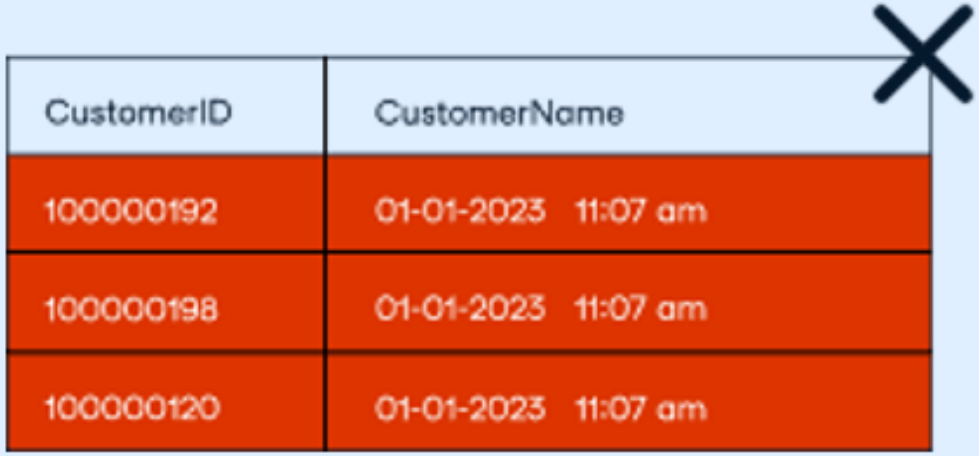
← Missed the SLA

Timeliness

Timeliness is the degree to which a dataset is available when expected and depends on service level agreements being set up between technical and business resources.

EXAMPLE

All records in the customer dataset must be loaded by the 9:00 am.

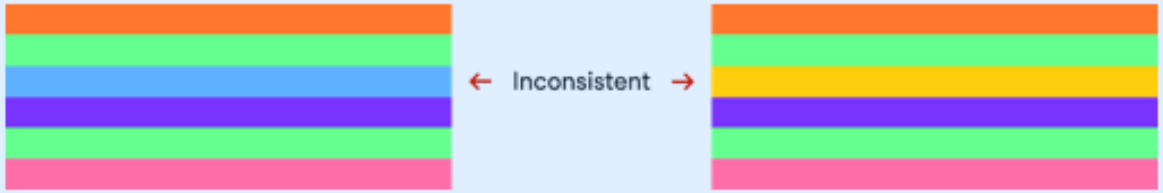


CustomerID	CustomerName
100000192	01-01-2023 11:07 am
100000198	01-01-2023 11:07 am
100000120	01-01-2023 11:07 am

CONSISTENCY

- **Consistency**

Consistency is a data quality dimension that measures the degree to which data is the same across all instances of the data. Consistency can be measured by setting a threshold for how much difference there can be between two datasets.



EXAMPLE

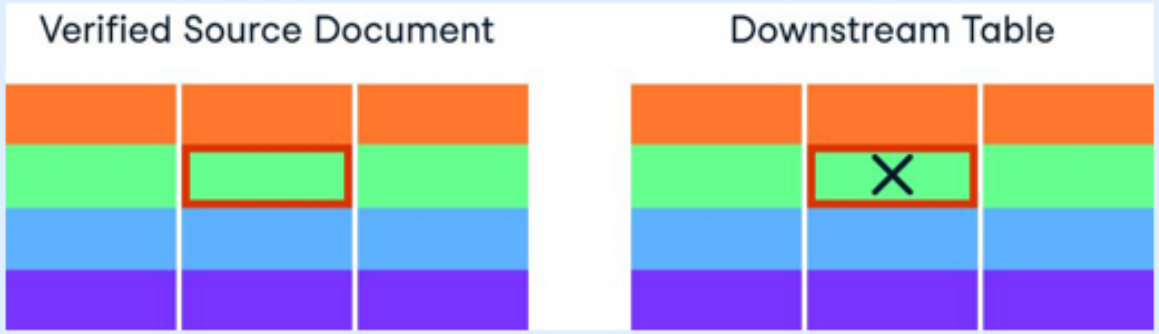
The count of records loaded today must be within +/- 5% of the count of records loaded yesterday.

Count of records in TargetCustomerTable	Record count difference from previous day
10,000,000	4,909,797 ✗
5,090,203	75 ✓
5,090,128	1 ✓

ACCURACY

- **Accuracy**

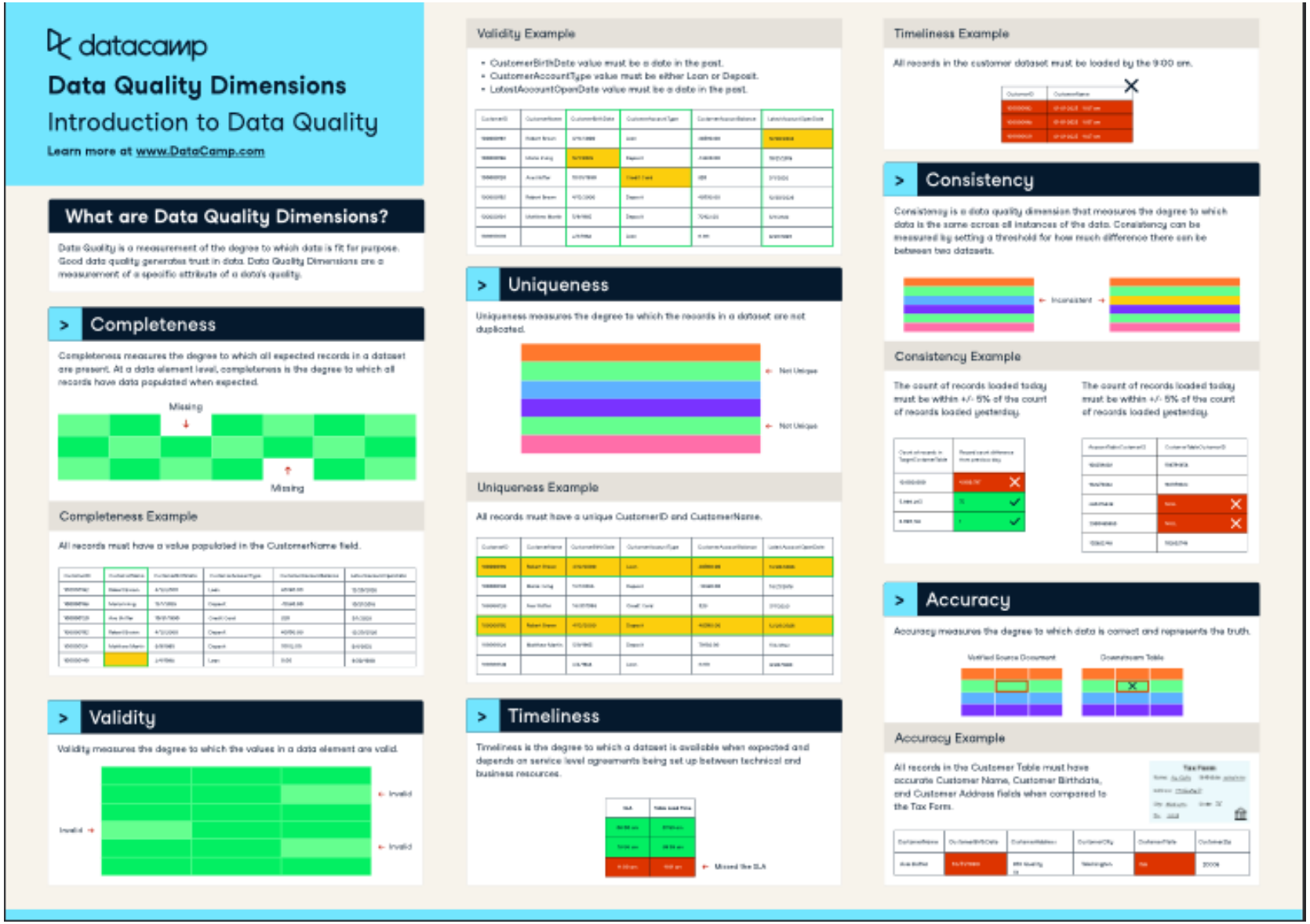
All records in the Customer Table must have accurate Customer Name, Customer Birthdate, and Customer Address fields when compared to the Tax Form.



EXAMPLE

All records in the Customer Table must have accurate Customer Name, Customer Birthdate, and Customer Address fields when compared to the Tax Form.

CustomerName	CustomerBirthDate	CustomerAddress	CustomerCity	CustomerState	CustomerZip
Ava Shiffer	10/31/1990	910 Quality St	Washington	WA	20008



https://images.datacamp.com/image/upload/v1678808802/Data_Quality_Dimensions_yebbuu.pdf

WELCOME TO THE WORLD OF
DATA CLEANING & DATA PREPROCESSING

- The dirty data includes things like **incomplete**, **inaccurate**, **irrelevant**, **corrupt** or **incorrectly formatted data**. The process also involves **deduplicating**, or '**deduping**'. This effectively means merging or removing identical data points.

- Note: Data cleaning is time-consuming with great importance comes great time investment. Data analysts spend anywhere from 60-80% of their time cleaning data.

BREAK (10 MINUTES)

WARNING !

The following concepts and operations will be applied are *NOT SEQUENTIAL*

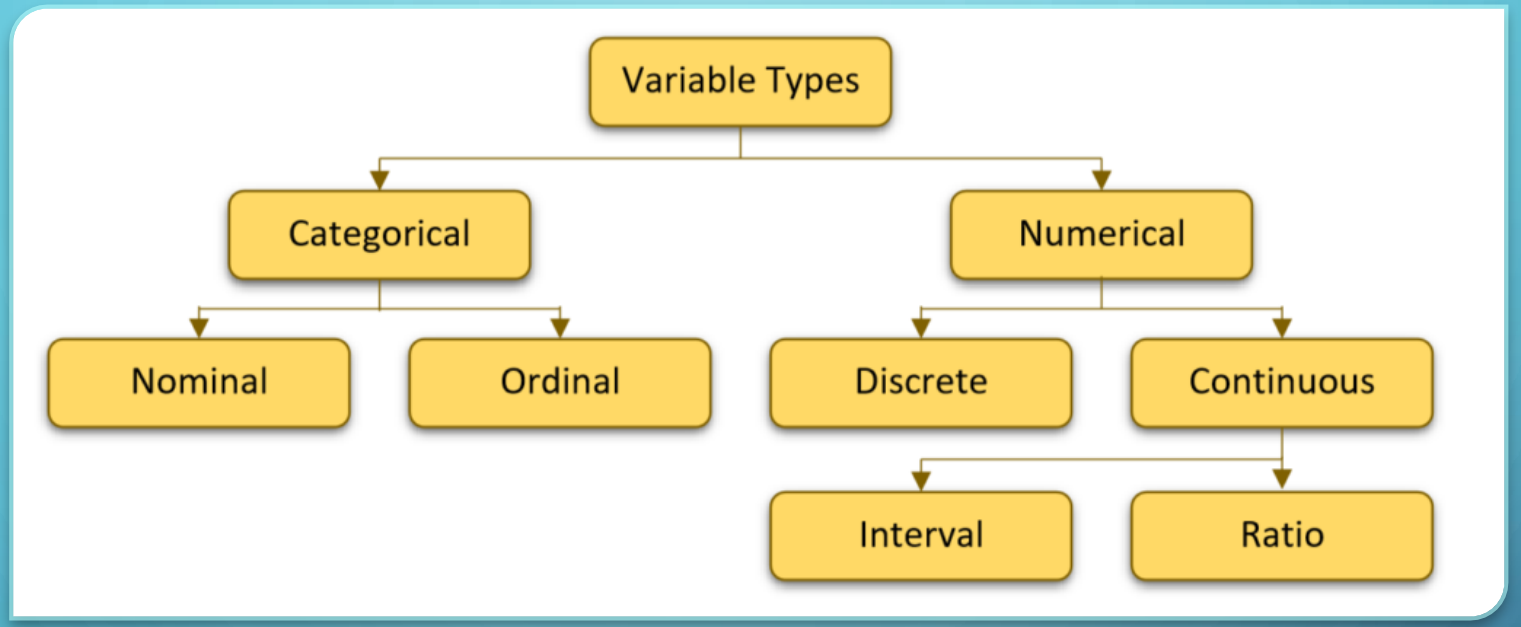
age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
39	0	2	138	220	0	1	152	0	0	1	0	2	1
58	0	0	130	197	0	1	131	0	0.6	1	0	2	1
47	1	2	130	253	0	1	179	0	0	2	0	2	1
35	1	1	122	192	0	1	174	0	0	2	0	2	1
58	1	1	125	220	0	1	144	0	0.4	1	4	3	1
56	1	1	130	221	0	0	163	0	0	2	0	3	1
56	1	1	120	240	0	1	169	0	0	0	0	2	1
55	0	1	132	342	0	1	166	0	1.2	2	0	2	1
41	1	1	120	157	0	1	182	0	0	2	0	2	1
38	1	2	138	175	0	1	173	0	0	2	4	2	1
38	1	2	138	175	0	1	173	0	0	2	4	2	1
67	1	0	160	286	0	0	108	1	1.5	1	3	2	0

Customer	Genre	Age	Annual In	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79
17	Female	35	21	35
18	Male	20	21	66

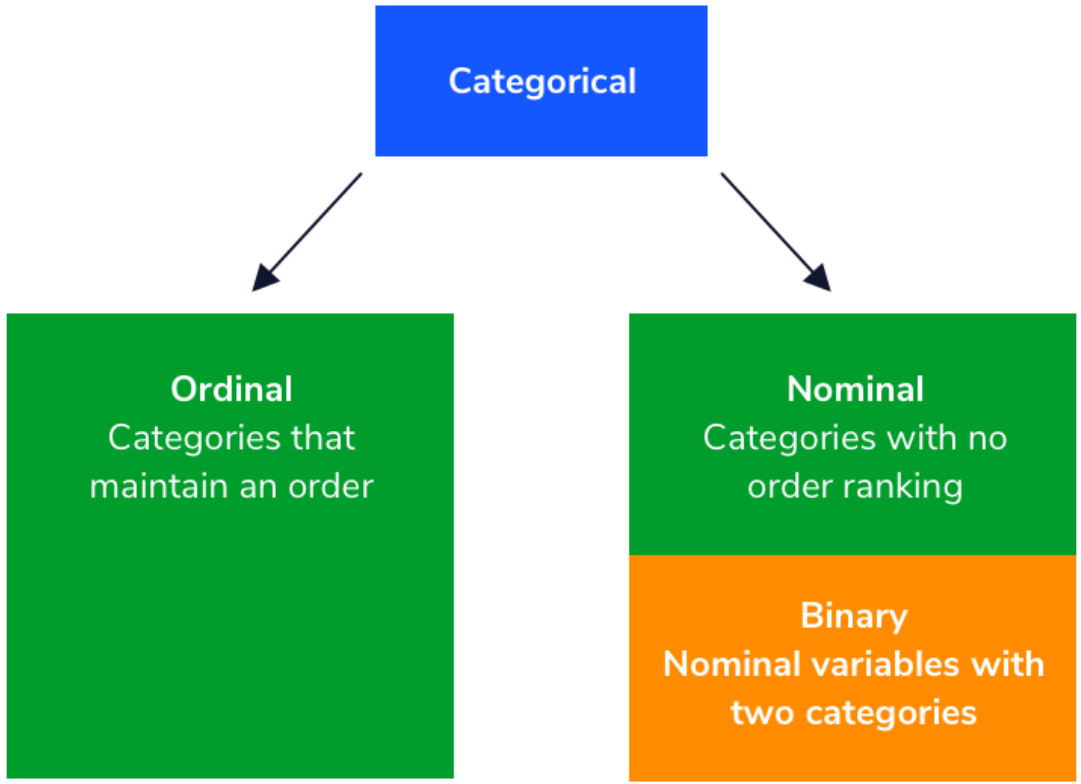
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1

6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0

TYPES OF ATTRIBUTES



CATEGORICAL: NOMINAL VS. ORDINAL



Categorical

Nominal


Pen


Pencil


Eraser


Cow


Dog


Cat

Ordinal


Excellent


Good

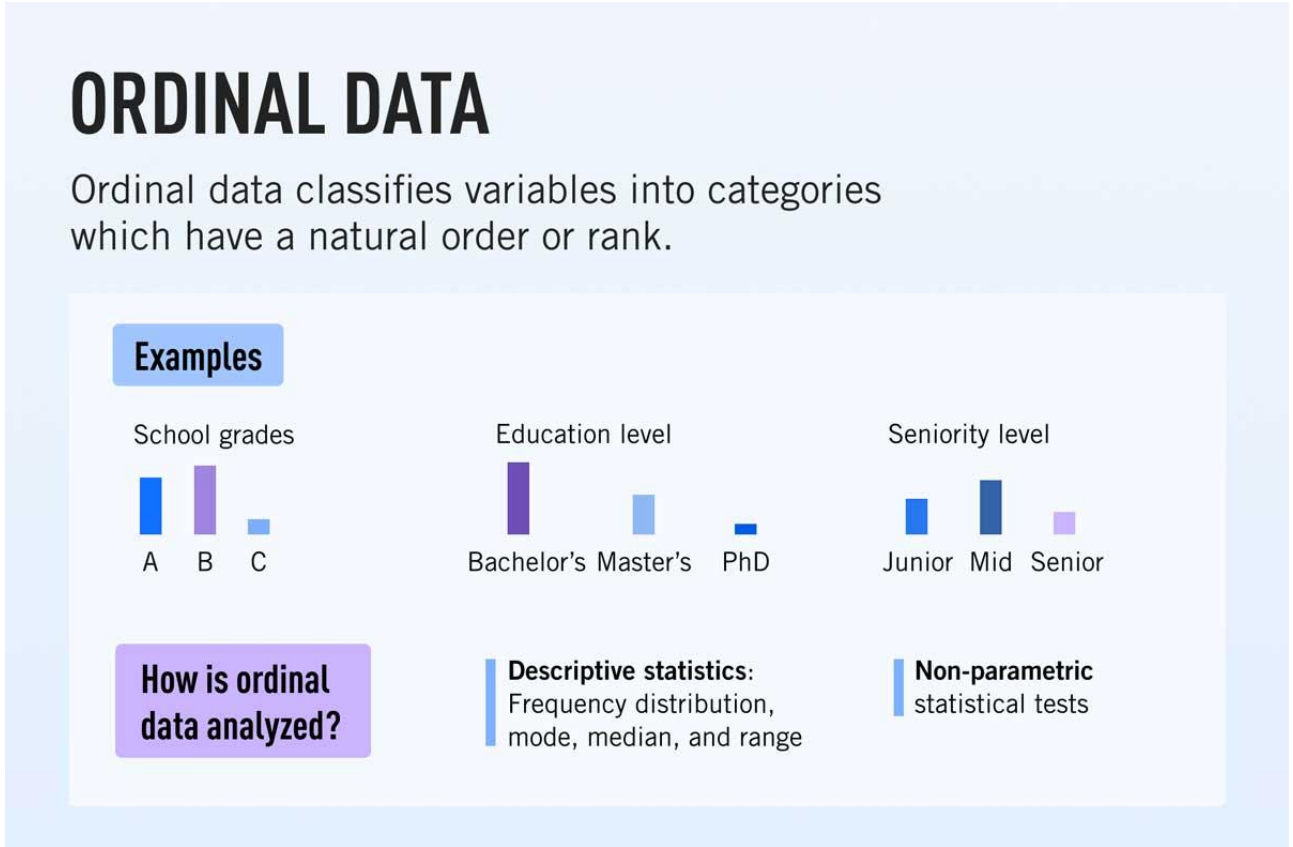

Bad


Fantastic


Okay


Don’t Like

ORDINAL ATTRIBUTE



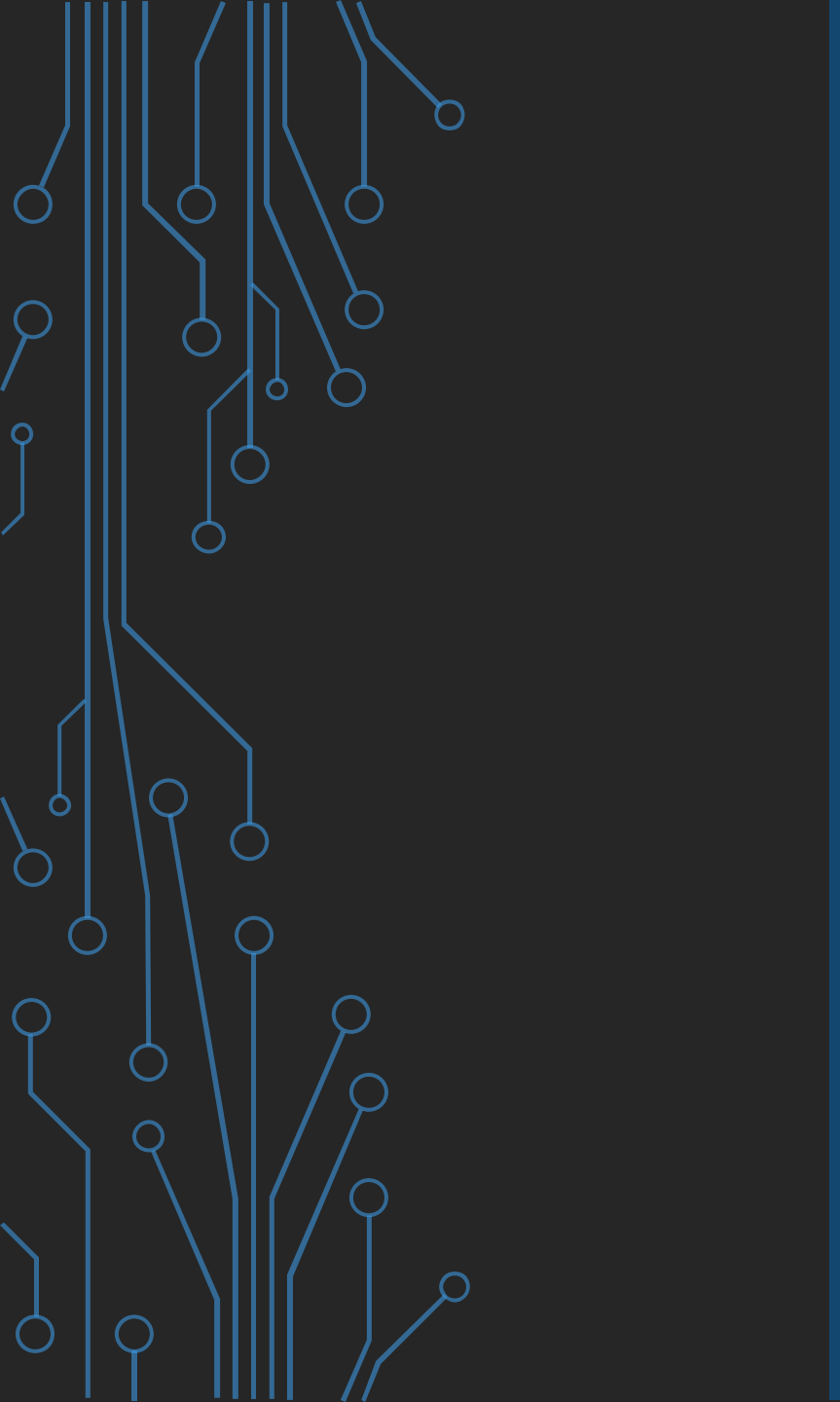
TEST YOUR UNDERSTANDING

Attribute	Values
Cancer detected	Yes, No
result	Pass , Fail

Attribute	Value
Grade	A,B,C,D,E,F
Basic pay scale	16,17,18

Attribute	Value
Profession	Teacher, Business man, Peon
ZIP Code	301701, 110040

Attribute	Values
Colours	Black, Brown, White
Categorical Data	Lecturer, Professor, Assistant Professor



DOES COMPUTER UNDERSTAND TEXT ?

State (Nominal Scale)
Maharashtra
Tamil Nadu
Delhi
Karnataka
Gujarat
Uttar Pradesh

State (Label Encoding)
3
4
0
2
1
5

id	color
1	red
2	blue
3	green
4	blue

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

FIRST:
LABEL ENCODING & ONE-HOT ENCODING

```
Original Categorical Data: ["Red", "Green", "Blue", "Green", "Red"]

Label Encoding:
---------------
Encoded Data: [0, 1, 2, 1, 0]
Explanation: Each unique category is assigned a unique integer label.
```

LABEL ENCODING (CHATGPT)

```
One-Hot Encoding:
-----------------
Encoded Data:
[1, 0, 0]   # Red
[0, 1, 0]   # Green
[0, 0, 1]   # Blue
[0, 1, 0]   # Green
[1, 0, 0]   # Red
Explanation: A binary vector is used for each category.
```

ONE-HOT ENCODING (CHATGPT)

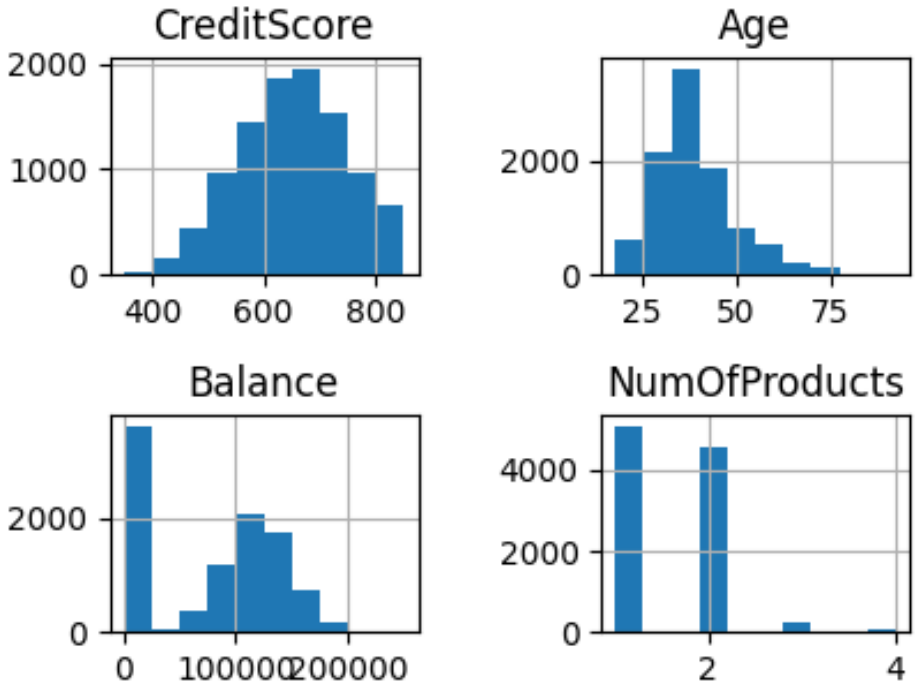
LABEL ENCODING VS. ONE-HOT ENCODING

- **Label Encoding** assigns a unique integer label to each category. It results in a single column of integers, which may imply ordinal relationships between the labels that may not exist.

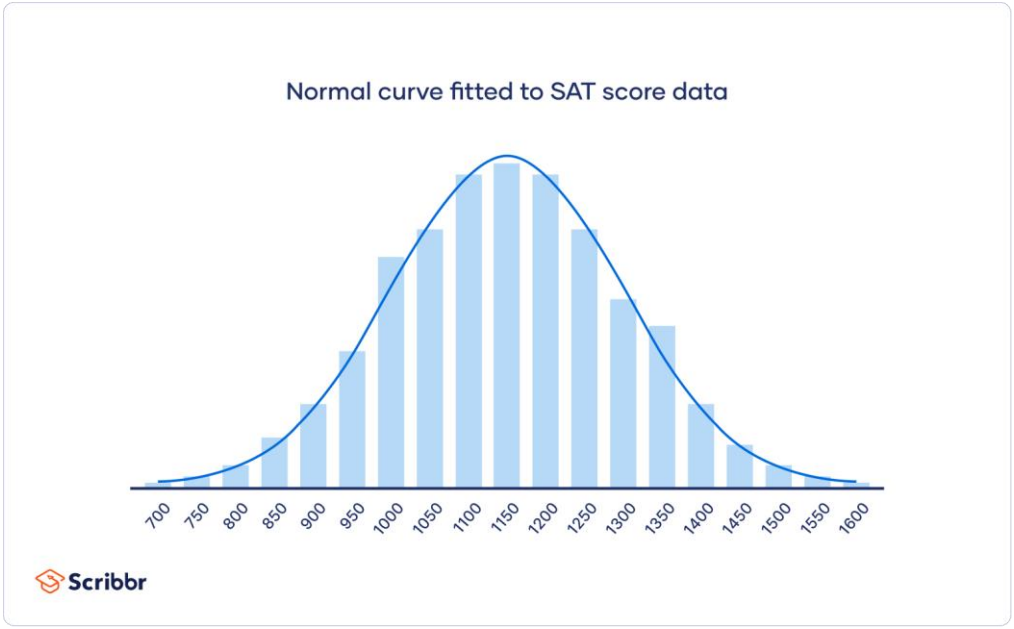
- **One-Hot Encoding** represents each category as a binary vector. It creates multiple columns, one for each category, with binary values (0 or 1) indicating the presence or absence of a category.

WHAT KIND OF STATISTICS ?

Frequency Distribution	Mode, Median	Bar Charts, Pie Charts	Relative Frequency and Proportions
This involves counting the occurrences of each category in a categorical attribute.	The mode is the category with the highest frequency in a categorical attribute.	These graphical representations are commonly used to visualize the distribution of categorical data.	These metrics provide insights into the proportions of different categories within the categorical variable.



SECOND: DISTRIBUTIONS (HISTOGRAMS)



NORMAL, GAUSSIAN OR BELL DISTRIBUTION

WHY NORMAL DISTRIBUTION ?

Central Limit Theorem

- **Central Limit Theorem**: One of the key reasons for the prominence of the normal distribution is the Central Limit Theorem.
- It states that the distribution of the sum (or average) of a large number of independent, identically distributed random variables approaches a normal distribution, regardless of the underlying distribution of the individual variables.

WHY NORMAL DISTRIBUTION ?

Simplicity and Familiarity

- The normal distribution has a simple and well-defined shape, characterized by its mean and standard deviation.
- This simplicity makes it easy to work with mathematically and visually. Additionally, many statistical techniques and tests assume or perform better when the data approximates a normal distribution.

WHY NORMAL DISTRIBUTION ?

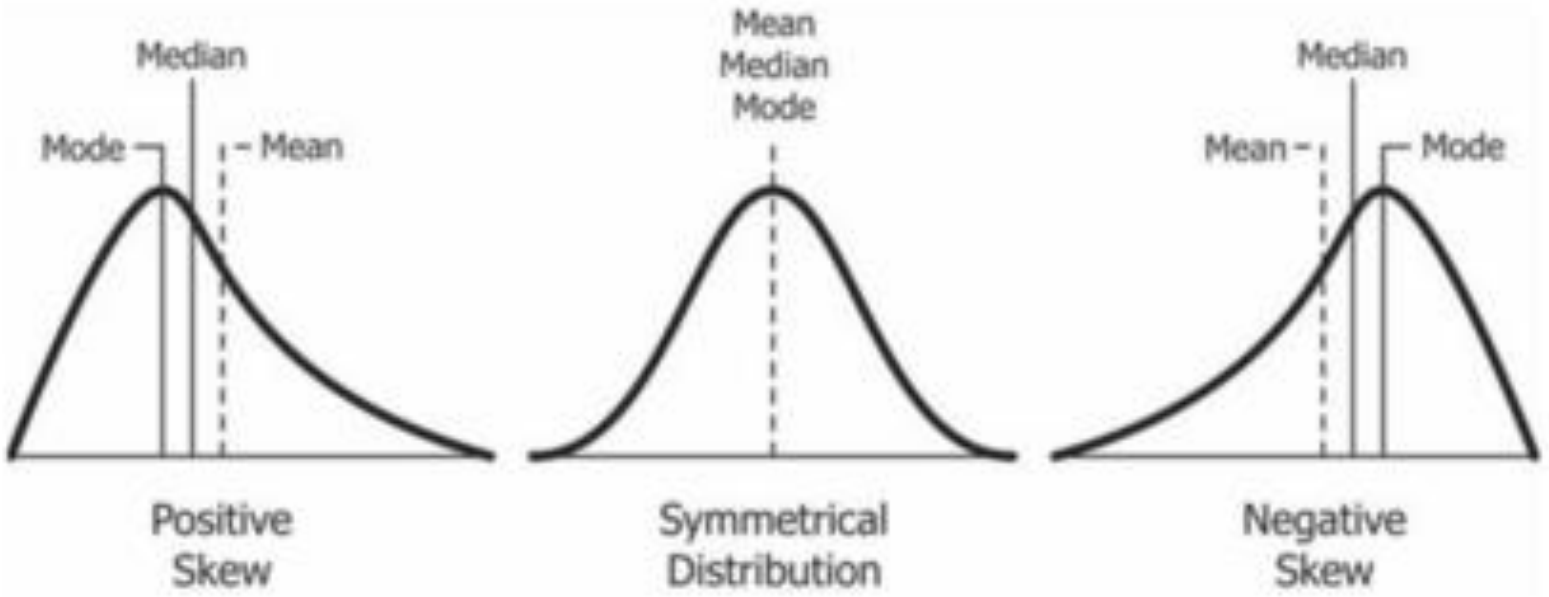
Parameter Estimation

- In many statistical analyses, such as linear regression, maximum likelihood estimation, and hypothesis testing, assumptions of normality simplify the estimation of parameters and the calculation of confidence intervals and p-values.

WHY NORMAL DISTRIBUTION ?

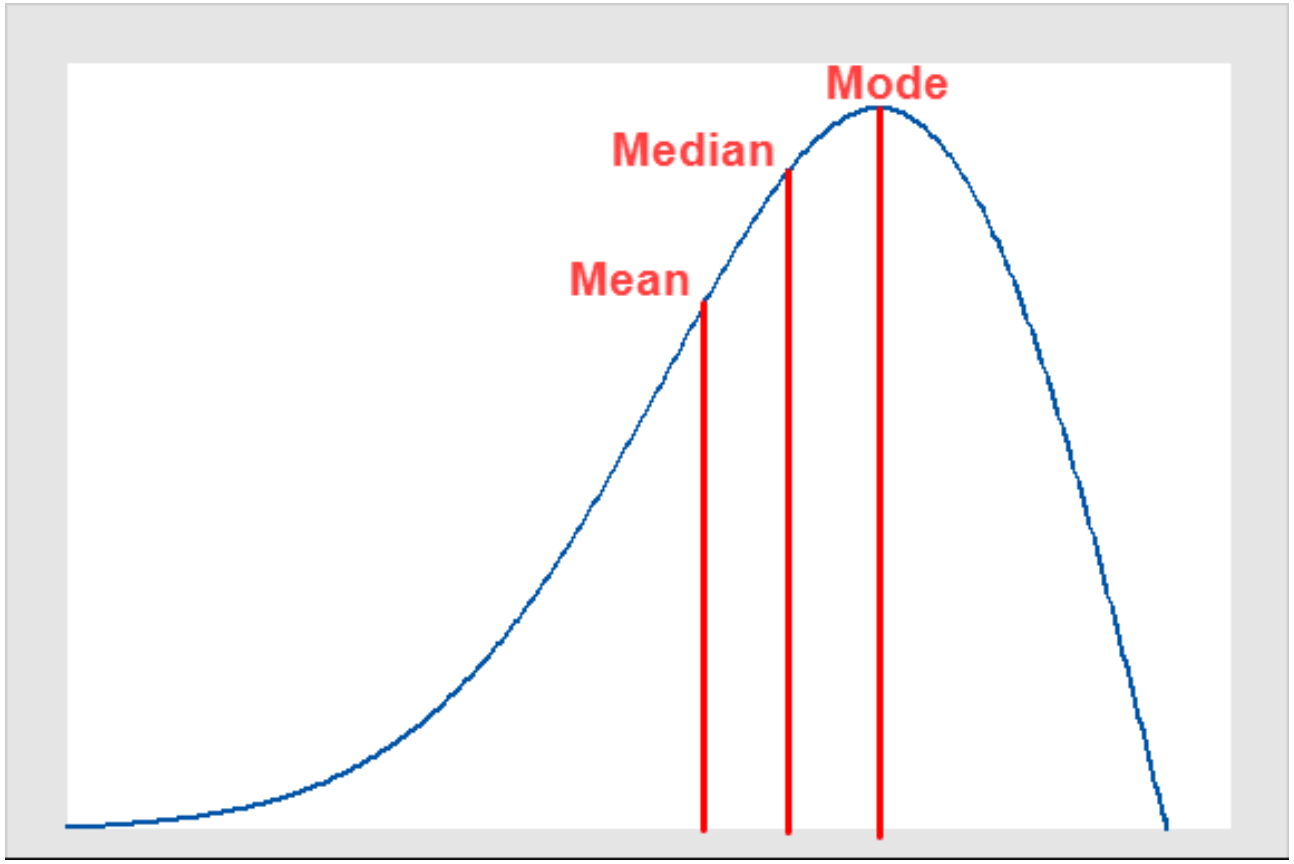
Data Transformation

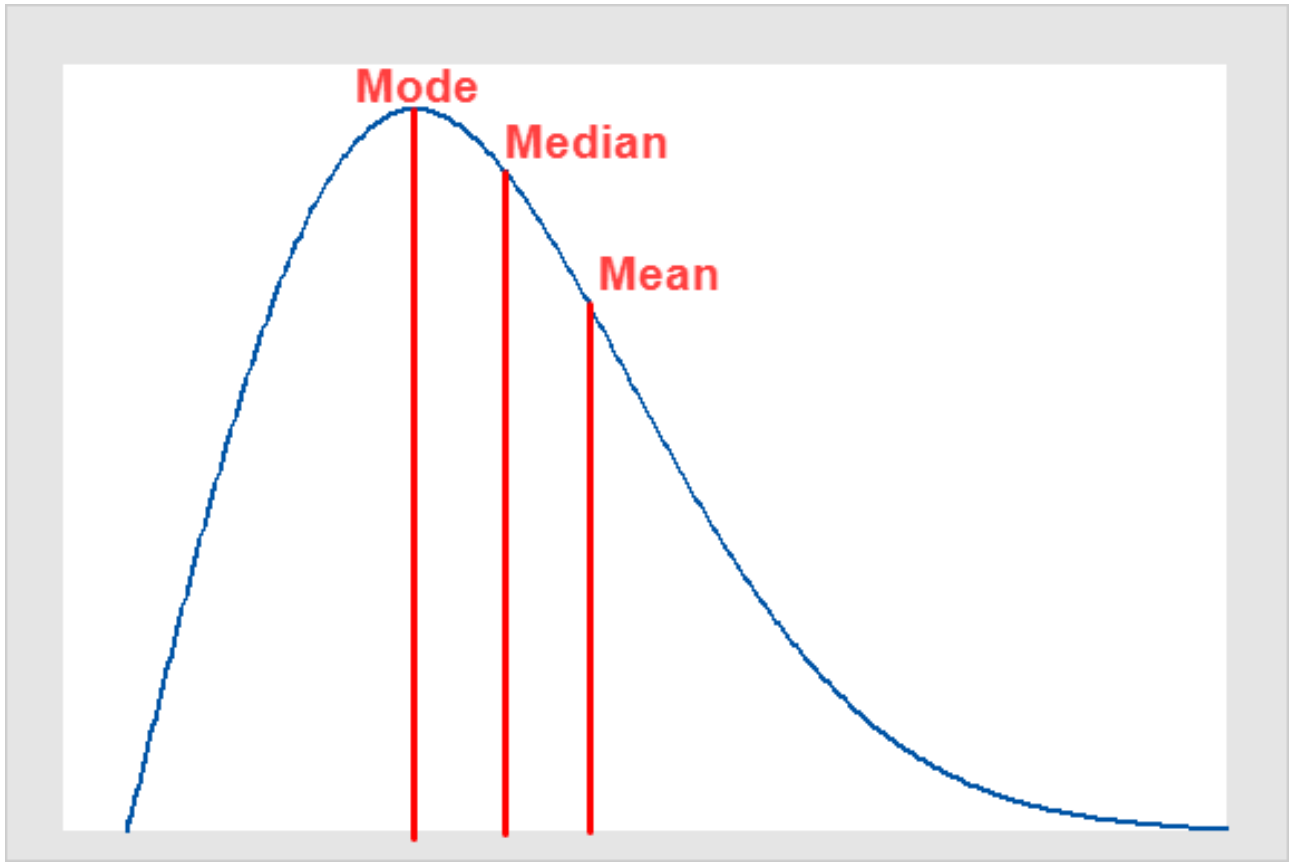
- ***WHEN DATA IS NOT NORMALLY DISTRIBUTED***, it is sometimes transformed (e.g., logarithmically or with Box-Cox transformation) to achieve normality. Normalizing the data can make it more suitable for certain analyses.



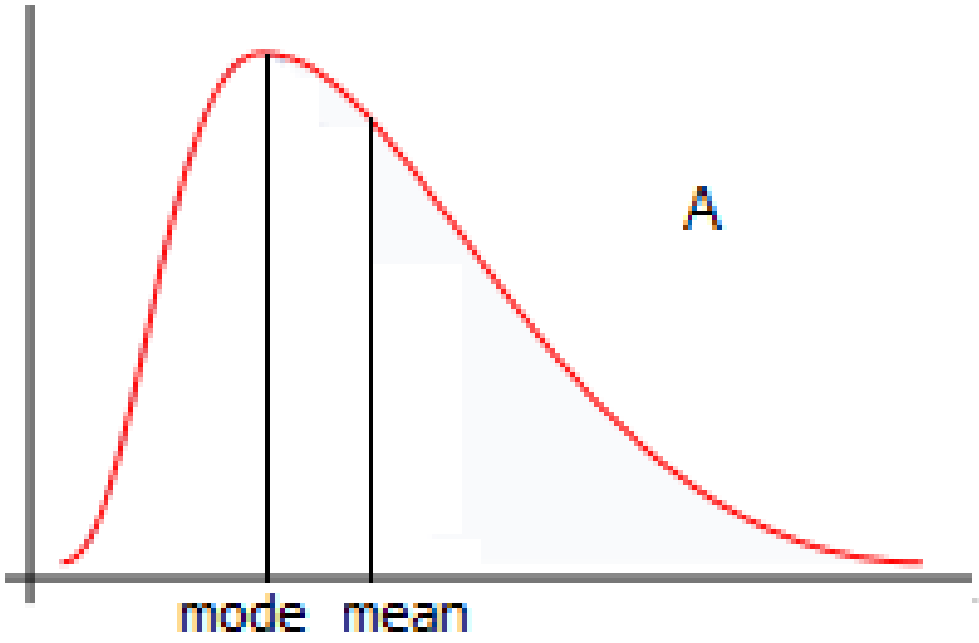
LEFT-SKEWED & RIGHT-SKEWED DISTRIBUTIONS

LEFT-SKEWED DISTRIBUTION

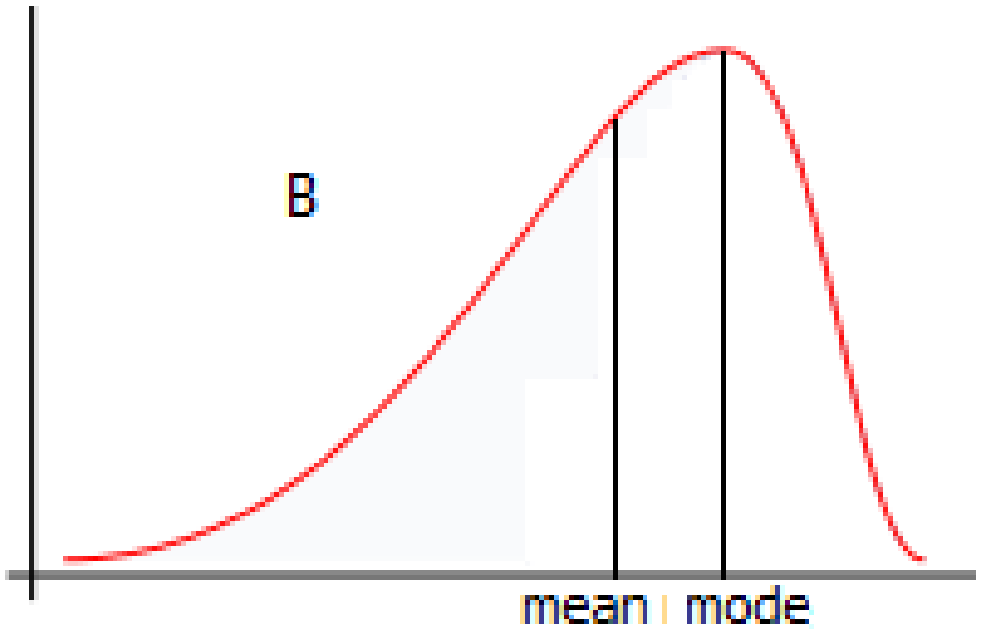




RIGHT-SKEWED DISTRIBUTION

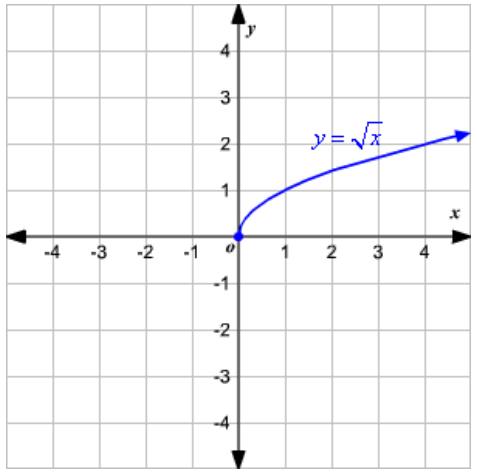


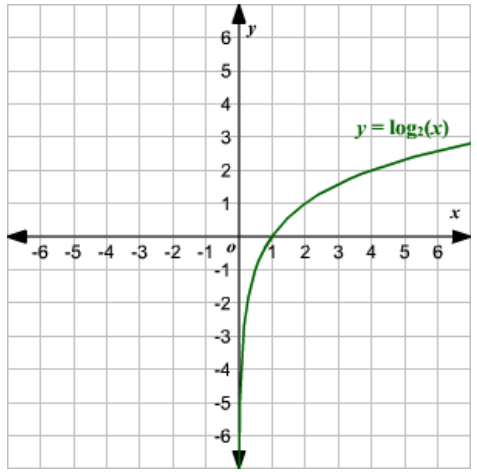
Gumbel distribution skew to right
mean>mode

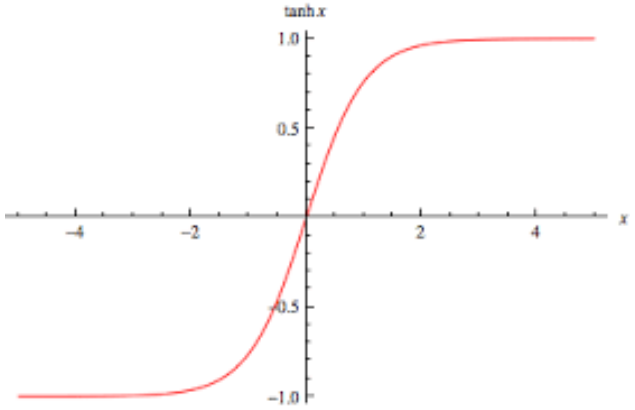


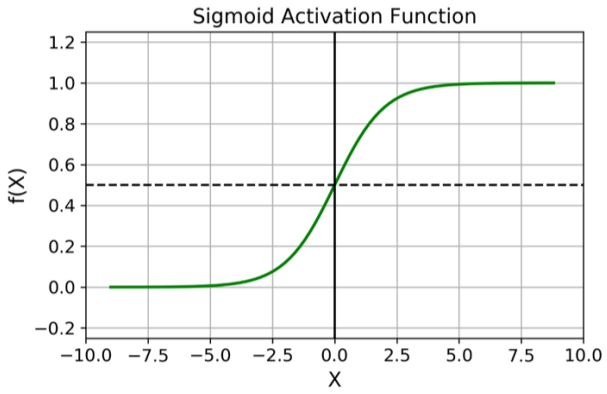
Mirrored Gumbel skew to left
mean<mode

HOW TO FIX SKEWNESS DISTRIBUTIONS ? (TRANSFORMATIONS)







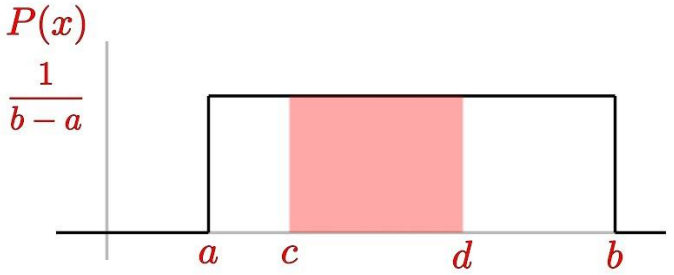


1. **Logarithmic Transformation**: Taking the logarithm of the data can compress the range of values and reduce the impact of extreme values. This is especially useful when the data spans several orders of magnitude.
2. **Square Root Transformation**: This transformation is milder than the logarithmic transformation but still reduces the impact of extreme values and brings the distribution closer to normal.
3. **Box-Cox Transformation**: The Box-Cox transformation is a family of power transformations that can be used to stabilize variance and make the data more closely resemble a normal distribution. The optimal transformation parameter is determined based on the data.
4. **Reciprocal Transformation**: Taking the reciprocal (1/x) of the data can help reduce the influence of large values and bring the distribution closer to symmetry.
5. **Exponential Transformation**: For data with a strong exponential growth pattern, applying the reciprocal of the exponential function (e^x) can help make the data more symmetric.
6. **Square Transformation**: Squaring the data can also help reduce the impact of extreme values, but it may have a stronger effect on data distribution compared to other methods.

1. **Power Transformation (Box-Cox or Yeo-Johnson):** Similar to addressing right-skewed data, you can apply power transformations like the Box-Cox or Yeo-Johnson transformations to stabilize variance and make the data distribution more symmetric.
2. **Square Root Transformation:** Just as with right-skewed data, taking the square root of the data can help mitigate the impact of extreme small values and bring the distribution closer to symmetry.
3. **Logarithmic Transformation:** While this is typically used for right-skewed data, in some cases, taking the logarithm of the inverse of the data (i.e., ln(1/x)) can help address left-skewness.
4. **Exponential Transformation:** Applying the exponential function (e^x) to the data can help emphasize larger values and reduce the skewness.

UNIFORM DISTRIBUTION

Uniform Distribution



$$Mean: \quad \mu = \frac{a+b}{2}$$

$$S.D.: \quad \sigma = \sqrt{\frac{(b-a)^2}{12}}$$

Probability

$$P(c \le X \le d) = \frac{d-c}{b-a}$$

HOW TO TRANSFORM UNIFORM INTO NORMAL DISTRIBUTION ?

Box-Cox Transformation

WHAT IS BOX-COX TRANSFORMATION ?

- **The Box Cox Transformation** is a popular method of transforming non-normal dependent variables into a normal shape.
- This technique helps to stabilize variance and can improve the accuracy of any subsequent statistical tests or models.
- It involves taking **the natural logarithm of a variable and then raising it to some power** (lambda) which is determined by maximum likelihood estimation.
- The lambda value will depend on how skewed the data is, meaning that a different lambda will be used for different data sets.

WHY BOX-COX TRANSFORMATION ?

There are three main reasons for using the Box Cox transformation:

1. To stabilise the variance
2. To improve normality
3. To make patterns in the data more easily recognisable

BREAK (10 MINUTES)

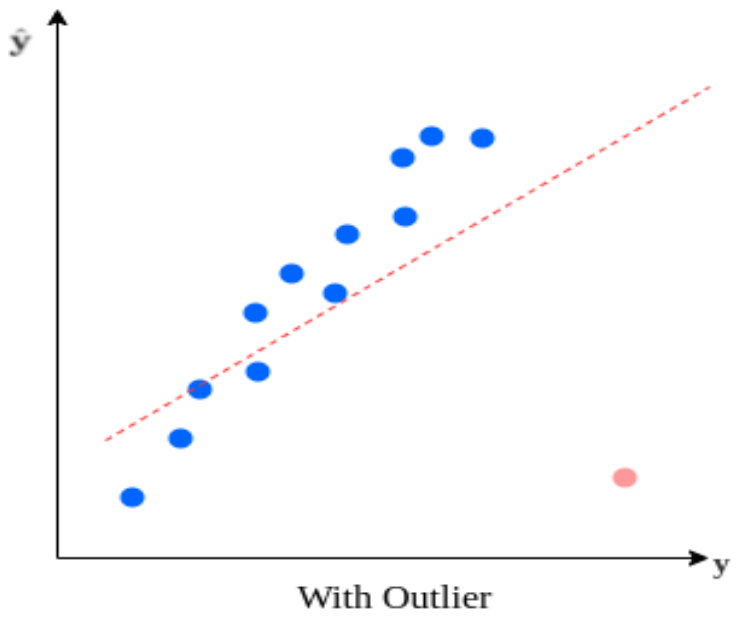
WHAT IS OUTLIERS ?

- **Outliers** are data points that significantly deviate from the rest of the data in a dataset.
- They can be unusually high or low values compared to the majority of the data points.
- Outliers can arise due to various reasons, including measurement errors, data entry mistakes, or genuine extreme values in the underlying phenomenon being studied.

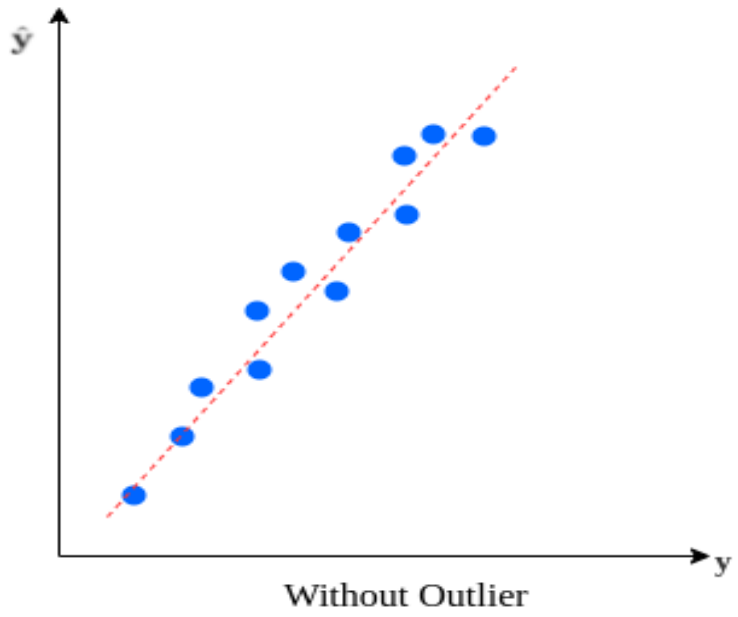
Outliers can have a significant impact on statistical analyses and machine learning models, potentially leading to biased or inaccurate results

1. **Visual Inspection**: Plotting data using histograms, box plots, scatter plots, or other graphical representations can help identify potential outliers.
2. **Z-Score**: The z-score measures how many standard deviations a data point is away from the mean. Data points with z-scores beyond a certain threshold (typically around ±2 to ±3) are considered outliers.
3. **Modified Z-Score**: Similar to the z-score, the modified z-score accounts for the median and uses the median absolute deviation (MAD) as a measure of variability.
4. **Interquartile Range (IQR)**: The IQR is the range between the first quartile (25th percentile) and the third quartile (75th percentile) of the data. Data points outside a certain range beyond the quartiles are considered outliers.

HOW TO DETECT OUTLIERS ?

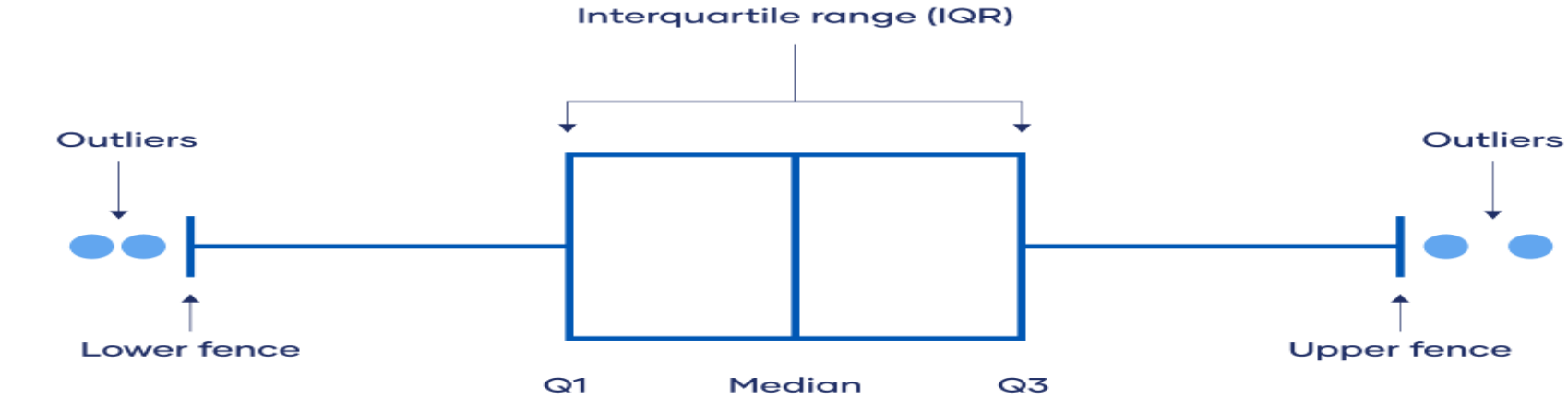


With Outlier



Without Outlier

OUTLIERS (SCATTER PLOT)

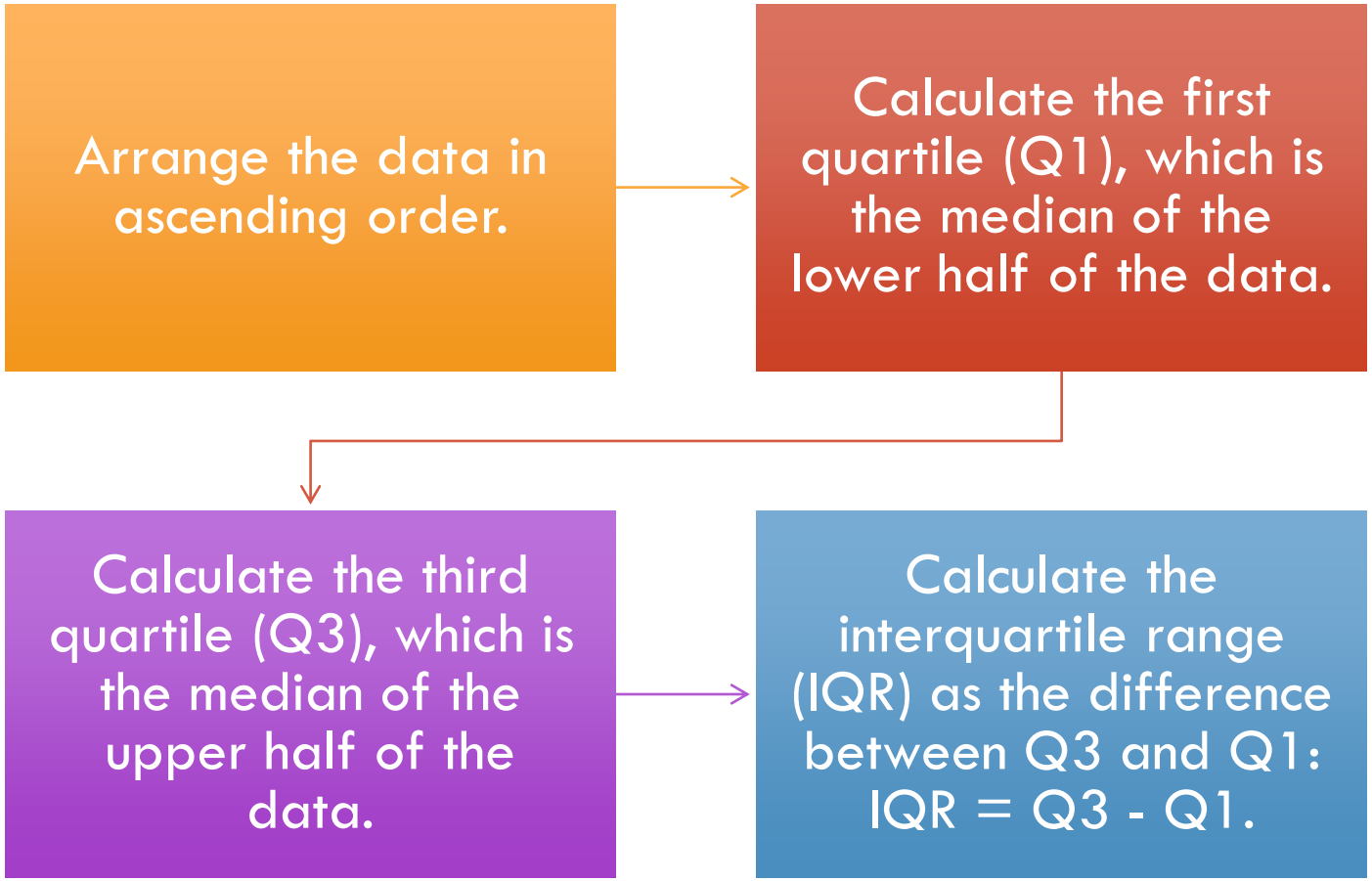


OUTLIERS (BOX PLOT)

INTERQUARTILE RANGE (IQR)

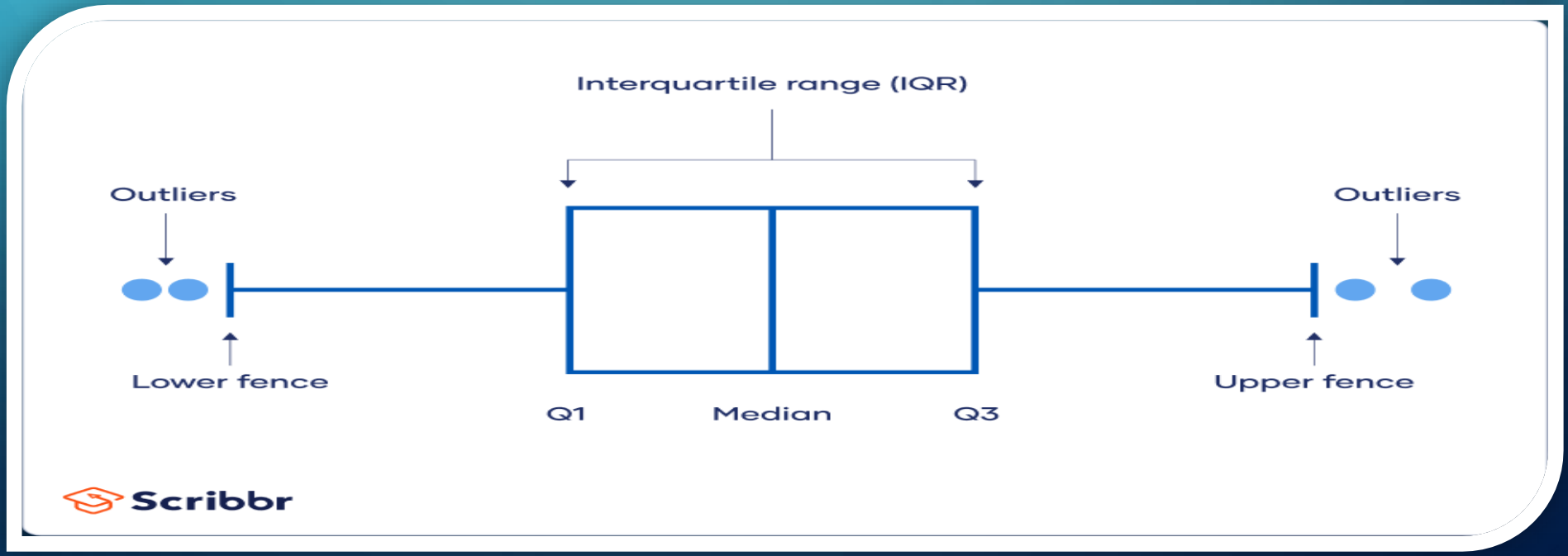
- **The interquartile range (IQR)** is a statistical measure used to describe the spread or dispersion of a dataset
- It is a measure of variability that focuses on the middle 50% of the data, specifically the range between the first quartile (Q1) and the third quartile (Q3).
- Quartiles are values that divide a dataset into four equal parts, each containing 25% of the data.
- **The IQR is a robust measure of dispersion because it is not influenced by extreme values (outliers) as much as the range or standard deviation.**

HOW TO CALCULATE IT ?



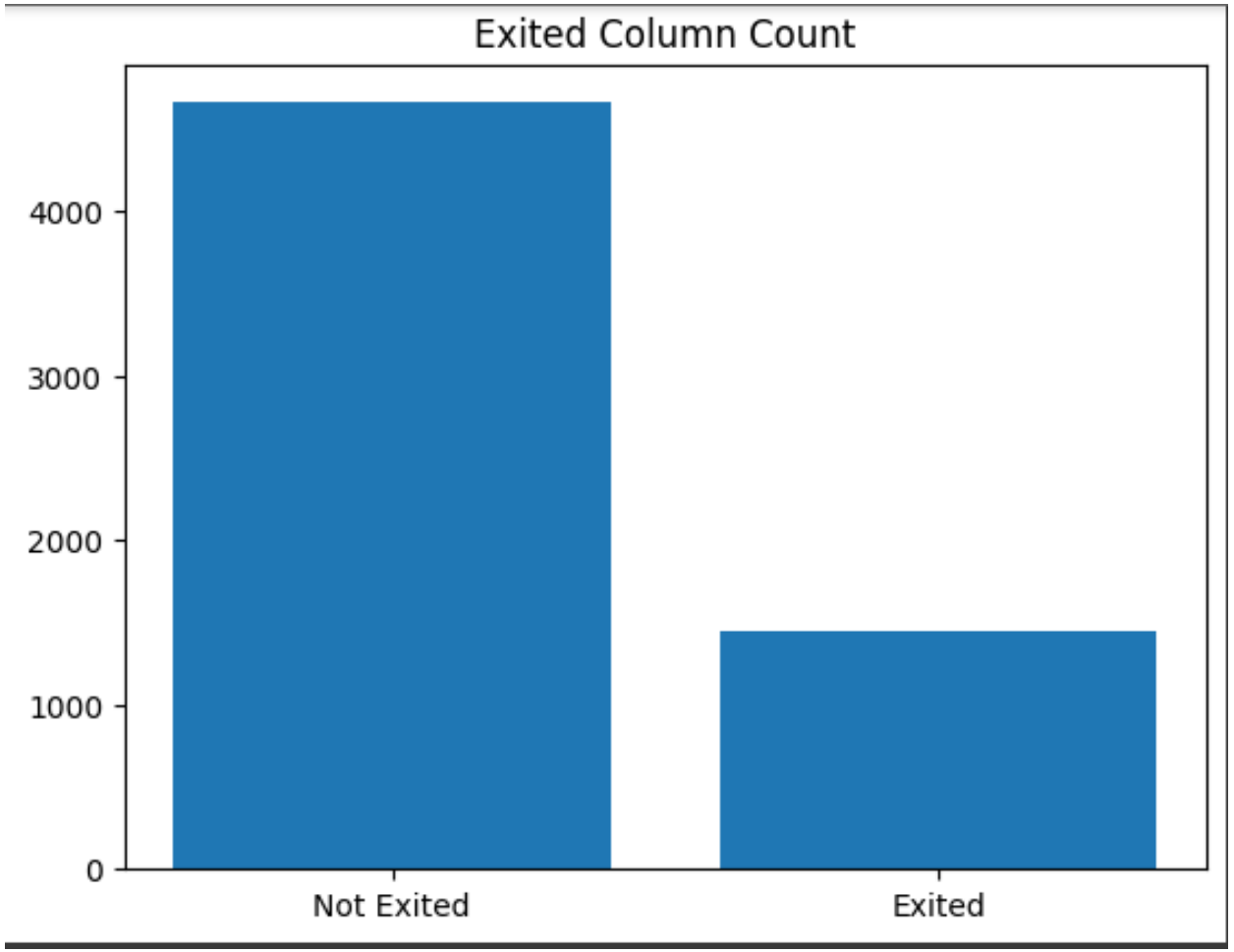
IQR & BOX PLOTS

- The IQR is commonly used in various statistical analyses and data visualization techniques, such as box plots. In a box plot
- the box represents the IQR, and the "whiskers" extend to the minimum and maximum values within a certain range (often 1.5 times the IQR) or to the actual data points within that range



HOW TO HANDLE OUTLIERS ?

1. **Removal**: Outliers can be removed from the dataset if they are likely to be data entry errors or have a negligible impact on the analysis.
2. **Transformation**: Applying data transformations (e.g., logarithmic, square root) can reduce the impact of outliers and make the data distribution more normal.
3. **Capping or Winsorization**: Capping or replacing extreme values with a predefined threshold can help mitigate the effect of outliers.
4. **Imputation**: Outliers can be replaced with more typical values using statistical imputation methods.
5. **Advanced Models**: Robust statistical models and machine learning algorithms that are less sensitive to outliers can be used.



IMBALANCED DATASETS (BAR CHARTS)

WHAT IS IMBALANCING ?

- An **imbalanced dataset** refers to a dataset in which the distribution of classes (categories or labels) is significantly skewed or uneven.
- In other words, one class has a much larger number of instances compared to one or more other classes.
- This imbalance can pose challenges for various machine learning algorithms and statistical analyses, particularly those that assume a roughly equal distribution of classes.

WHAT IS THE CHALLENGES AGAINST IMBALANCED DATASETS ?

1. **Bias in Model Performance**: Algorithms tend to be biased toward the majority class, resulting in poorer performance on the minority class.
2. **Limited Learning**: Algorithms may struggle to learn patterns from the minority class due to the limited number of instances.
3. **Misclassification Costs**: In some applications, misclassifying the minority class may be more costly than misclassifying the majority class.
4. **Evaluation Metrics**: Traditional accuracy may not be an appropriate metric for imbalanced datasets. Other metrics like precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) are more suitable.
5. **Overfitting**: Algorithms may overfit the majority class, resulting in poor generalization to new data.

SOLUTIONS

1. **Resampling**: This involves either oversampling the minority class, undersampling the majority class, or generating synthetic samples (e.g., using SMOTE - Synthetic Minority Over-sampling Technique).
2. **Cost-sensitive Learning**: Modify the algorithm's cost function to penalize misclassification of the minority class more heavily.
3. **Ensemble Methods**: Techniques like boosting or bagging can help improve the performance of algorithms on imbalanced datasets.

SOLUTIONS (CONT.)

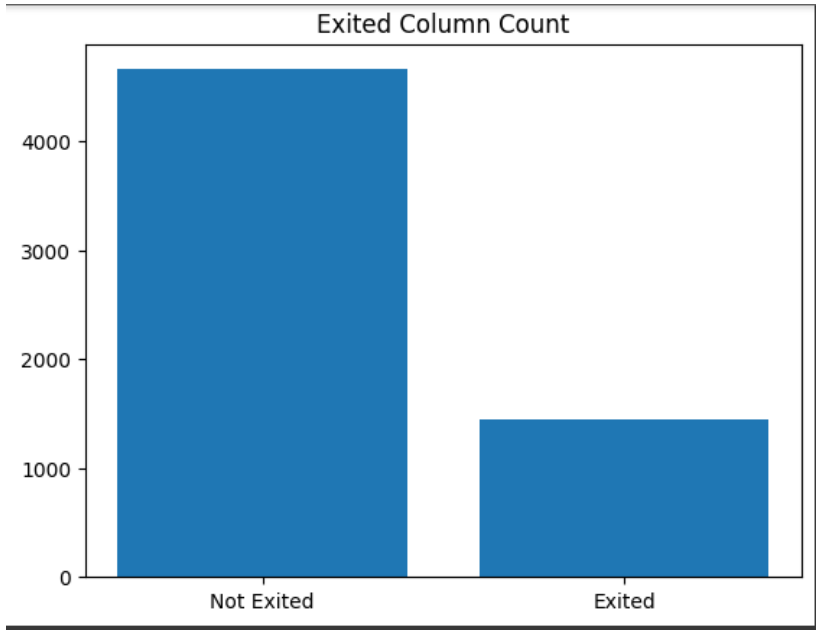
4. **Anomaly Detection**: Use specialized anomaly detection algorithms that are designed to handle imbalanced scenarios.
5. **Evaluation Metrics**: Focus on precision, recall, F1-score, and AUC-ROC to evaluate model performance.
6. **Data Augmentation**: Generate additional data instances for the minority class using techniques like text augmentation (in NLP) or image augmentation (in computer vision).
7. **Transfer Learning**: Utilize pre-trained models and fine-tune them for the imbalanced dataset.

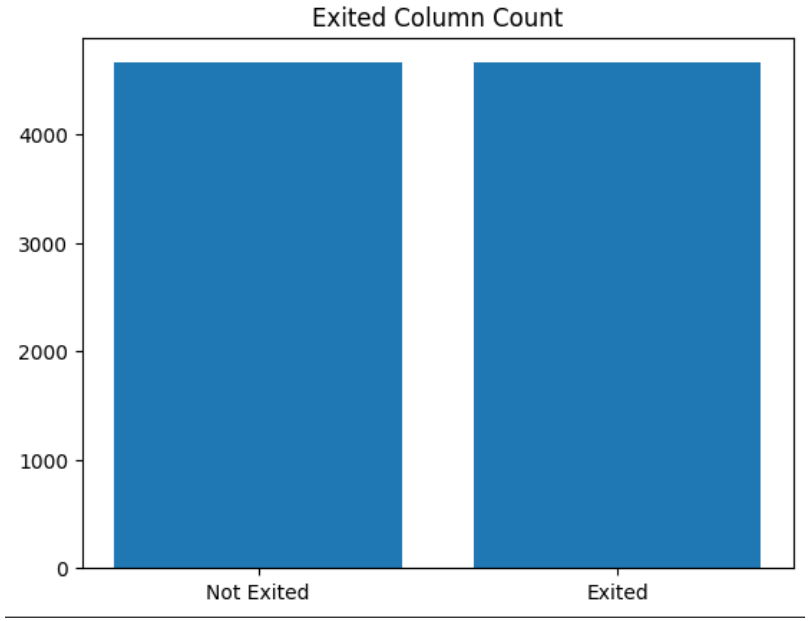
SYNTHETIC-MINORITY OVERSAMPLING TECHNIQUE (SMOTE OVERSAMPLING)

- It is a popular technique used to address the issue of class imbalance in machine learning datasets.
- **SMOTE** is specifically designed to increase the representation of the minority class by generating synthetic samples.
- In a class-imbalanced dataset, the minority class (i.e., the class with fewer instances) is often underrepresented, which can lead to biased model performance and poor generalization.

SMOTE ALGORITHM

1. For each instance in the minority class, SMOTE selects k nearest neighbors from the same class. The value of k is a parameter chosen by the user.
2. Synthetic samples are generated by interpolating between the selected instance and its k nearest neighbors. This is done by randomly selecting a neighbor and computing the difference between the feature values of the instance and the neighbor. A random fraction of this difference is added to the instance to create a new synthetic sample.
3. The process is repeated until the desired level of balance is achieved.





BEFORE AND AFTER BALANCING

NORMALIZATION & SCALING

- **Normalization and scaling** are techniques used in data preprocessing to transform the features (variables) of a dataset into a specific range or distribution.

- These techniques are important for preparing data before feeding it into machine learning algorithms, as they can improve the performance and convergence of various models.

NORMALIZATION

- **Normalization** involves transforming the entire dataset so that each feature has a similar scale.
- The goal is to bring all features to a similar range, typically between 0 and 1.
- This is particularly useful for algorithms that rely on distances or gradients, such as k-nearest neighbors (KNN) and gradient descent-based optimization algorithms.

TYPES OF NORMALIZATION

1. **Min-Max Scaling**: Scales the data to a specified range (e.g., [0, 1]) using the formula: `x_normalized = (x - min(x)) / (max(x) - min(x))`.
2. **Z-Score Normalization (Standardization)**: Transforms the data to have a mean of 0 and a standard deviation of 1 using the formula: `x_standardized = (x - mean(x)) / std(x)`.

SCALING

- Scaling, on the other hand, focuses on adjusting the range of features without necessarily changing their distribution or statistical properties.

- Scaling can be important for algorithms that use distance measures but don't necessarily require a specific range for input features.

TYPES OF SCALING

1. **Min-Max Scaling**: As mentioned above, this technique scales features to a specific range.
2. **Z-Score Scaling (Standardization)**: As mentioned above, this technique standardizes features to have a mean of 0 and a standard deviation of 1.
3. **Robust Scaling**: This technique is similar to Z-score scaling but uses the median and interquartile range, making it more robust to outliers.
4. **Max Absolute Scaling**: Scales features by dividing them by their maximum absolute value, resulting in values within the range [-1, 1].
5. **Unit Vector Scaling (Normalization)**: Scales features to have a Euclidean norm (magnitude) of 1, which can be useful for algorithms that rely on vector distances.

NORMALIZATION VS. SCALING

- Use **Normalization** when you want to ensure that all features are on a similar scale, which can be important for distance-based algorithms.
- Use **Scaling** when you want to adjust the range of features without necessarily changing their distribution, and the specific scale of features isn't critical.

PEARSON CORRELATION

- **Pearson's correlation coefficient** or Pearson's r, is a statistical measure that quantifies the linear relationship between two continuous variables.

- It assesses how closely the data points of two variables align on a straight line. The coefficient ranges from -1 to 1, where:

  - A value of +1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other also increases proportionally.

  - A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other decreases proportionally.

  - A value of 0 indicates no linear relationship between the variables; they are independent of each other.

PEARSON CORRELATION (CONT.)

- In essence, Pearson correlation helps to determine if there is a consistent pattern in the way two variables change in relation to each other.

- It's important to note that Pearson correlation specifically measures linear relationships, so if the relationship between the variables is not linear, the correlation coefficient may not accurately capture their association.

The formula for calculating Pearson correlation coefficient between two variables, X and Y, based on their sample data, is as follows:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Where:

- X_i and Y_i are individual data points of variables X and Y, respectively.
- $\bar{X}$ and $\bar{Y}$ are the means (averages) of variables X and Y, respectively.

PEARSON CORRELATION (CONT.)

	0	1	2	3	4	5	6	7	8	9
0	1	0.35	0.4	0.46	0.073	-0.23	-0.73	0.48	-0.44	0.015
1	0.35	1	-0.28	0.57	-0.29	0.38	-0.36	0.64	0.25	0.19
2	0.4	-0.28	1	-0.52	0.15	-0.14	-0.093	0.016	-0.43	-0.38
3	0.46	0.57	-0.52	1	-0.23	-0.23	-0.48	0.47	0.28	0.45
4	0.073	-0.29	0.15	-0.23	1	-0.1	-0.15	-0.52	-0.61	-0.19
5	-0.23	0.38	-0.14	-0.23	-0.1	1	-0.03	0.42	0.21	0.095
6	-0.73	-0.36	-0.093	-0.48	-0.15	-0.03	1	-0.49	0.38	-0.35
7	0.48	0.64	0.016	0.47	-0.52	0.42	-0.49	1	0.38	0.42
8	-0.44	0.25	-0.43	0.28	-0.61	0.21	0.38	0.38	1	0.15
9	0.015	0.19	-0.38	0.45	-0.19	0.095	-0.35	0.42	0.15	1

TIME FOR PRACTICALITY (20 MINUTES)

QUESTIONS


THANK YOU