Data Visualization:

is the representation of data through use of common graphics, such as Pie chart, bar chart, Histogram, scatter plot, and even animations. These visual information communicate complex data relationships and data-driven insights in a way that is easy to understand.

Importance of Data Visualization:

i) Improved Understanding: Data visualization helps in comprehending complex data by allowing individuals to identify patterns, trends, and correlations that may not be apparent in raw data.

ii) Enhanced Decision making: Data visualization enables decision-makers to gain insights quickly and make informed decisions. By presenting data in a visually appealing way, it becoms easier to identify key areas of focus, spot outliers, and understanding the impact of different factors.

iii) Effective Communication: Visual representations of data are often more accessible and engaging than raw number or text. Data visualization facilitates effective communication and story telling by Presenting information in a format that can be easily understood and remembered by the audience.

Purpose of Data Visualization: The main purpose of data visualization is to facilitate:

i) Data exploration: Visualization helps in understanding the structure and characteristics of data sets, enabling analysts to gain insight and formulate hypotheses

ii) Analyze Data: Visualization techniques allow analysts to identify trends, correlations, and patterns in data, aiding in data-driven decision-making and Problem solving.

iii) Communicate Findings: Visualization make it easy to communicate complex data and analysis results to a broader audience. They enable effective story telling, supporting presentations, reports and dashboards.

Benefits of Data visualization Tools:-

- Solves data inefficienties and absorb vast amount of data presented in visual formats.
- Increase the speed of decision making.
- Identify errors and inaccuracy in data quickly
- Access real time information and assist in management function.
- It promotes story telling and conveys the right message to the audience.
- Explore bussiness insights and achieve business

Basic principles of data visualizations:-

- Truffle's law: This law states that the human eye is drawn to the brightest and most contrasting elements in a visual. This means that it's important to use colours and fonts that will stand out and grab the viewer's attention.

- Gestalt's law: This theory of visual perception states that the human mind groups toghter similar objects or features. This can be used to create visul patterens and make data easier to understand.

Preattentive Processing:

is a type of visual processing that occurs before we consciously pay attention to something. This type of processing allows us to quickly identify objects and patterns in our environment. There are mainly four preattentive processing:

- Form: includes shap, size and orientation of an object.
- Color: includes hue, satusation, brightness of an object.
- Spatal positioning: includes location of an object in relation to other objects
- Motion: includes movement of an object.

Data visualization is a powerful tool that can be used to communicate information in a clear and concise way. By following these one can create effective & engaging data visualizations.

Exploratory Data Analysis (EDA) :-

is an approach used in data science and statistics to analyze and summarize the main characteristics of a dataset. It involves a variety of techniques and visualizations to understand the data, identify patterns, detect anomalies, and generate hypotheses for further investigation.

The primary goal of EDA is to gain insights into the data and develop the understanding of its structure, distribution, relationships b/w variables. It helps researchers and data scientists to formulate appropriate questions, determine the quality and reliability of the data, and select appropriate statistical methods for further analysis.

Steps involved in EDA :-

- Data collection (collecting required data from various sources)
- Finding all variables and understanding them (identify imp variables and understand their impact).
- cleaning the Dataset (Remove null values and irrelevant information)
- Identify correlated variables (Find relationship b/w variables)
- Choosing the right statistical methods (use appropriate statical tools for analysis)
- Visualizing and Analyzing Results: (Observe and interpret the findings)

Type of EDA :-

i) Univariate Analysis : focus on analyzing individual variables independently and uses visualizations such as histograms, box plots, bar charts, etc...

ii) Bivariate Analysis : focus on analyzing relationship b/w two variables. and uses visualizations such as scatter plots, line plots, correlation analysis

iii) Multivariable Analysis : analyzing relationship b/w multiple variables simultaneously and uses visualizations : heatmaps, Pie charts and clustered analysis.

EDA uses :-

- understand the data better
- Identify patterns and trends
- spot anomalies
- Test hypothesis
- Make better decisions

Machine learning Algorithms:

ML algorithms are computational models that enable computers to learn Patterns and make predictions or decisions without being explicitly programmed These algorithms learns and generalize patterns from input data to perform specific tasks, such as classification, regression, clustering or recommendation

Supervised learning: labled data is given and trained in presence of both input & output
eg:- Classification: image, text, voice recognition; Regression: Fraud, weather detection

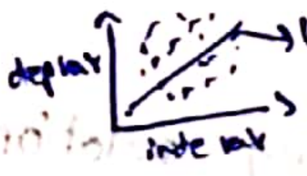
Unsupervised:- unlabled data, find hidden patterns by itself. eg: Clustering: Segmentation anomaly detection, association mining. Dimensionality reduction. (Prediction & recommendation models)

Semi supervised: mixture of both sup & unsup learning. Data scientists may feed an algorithm mostly labled training data, but the model is free to explore the data on its own and develop its own understanding. eg (Machine translation) Fraud detection, labelling the

Reinforcement:- feedback based ML recieves feedback from env by agent through actions.
eg (Robotics, Videogames, Resource management)

different ML algorithms:-

1.) **Linear Regression**: Supervised learning alg used for regression tasks. It models the relationship blw a dependent variable and one (or) more independent variable by fitting a linear eq to the observed data. The goal is to find best fitting line that minimizes the diff blw the predicted and actual values.




```
From sklearn.model_selection import train_test_split
From sklearn.linear_model import LinearRegression, LogisticRegression.
From sklearn.tree import DecisionTreeRegression, DecisionTreeClassifier.
From sklearn.ensemble import RandomForestRegression, RandomForestClassifier.
From sklearn.SVM import SVR, SVC
From sklearn.neighbors import KNeighborsRegression, Classifier
From sklearn.naive_bayes import GaussianNB
From sklearn.metrics import accuracy_score, mean_square_error
import numpy as np.
```

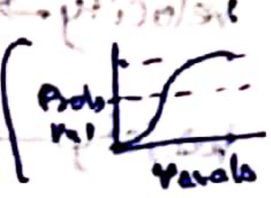
```
# sample input data
X = np.array([[1,2],[2,3],[3,4],[4,5],[5,6]])
y = np.array([10,20,30,40,50]).

# load and split the dataset:
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

linear_regression = Linear_Regression()
linear_regression.fit(X_train, y_train)
linear_Prediction = linear_regression.predict(X_test)

#model evaluation
mse = mean_squared_error(y_test, linear_Predictions)
accuracy = accuracy_score(y_test, linear_Prediction)

Print("Linear Regression MSE :", mse)
Print("Linear Regression Accuracy :", accuracy)
```

ii) Logistic Regression : supervised learning alg used for classification tasks. models relationship by fitting logistic function to data. Predicts probability of an instance belonging to a particular class. []

iii) Decision Trees : sup algo handles both reg & classi tasks, Tree like structure where internal nodes represents features, branch → decisions : leafnode → outcomes. []

iv) Random Forest : combins multiple decision trees to improve predicted accuracy. creat forest of Decision trees. handles high-dim data and avoid overfitting.

v) Support vector Machine : sup algo handles both reg & classi tasks. finds optimal hyperplane that maximizes the margin b/w diff classes, enabling effective separation. handle complex decision boundaries, work well on both linear, non-linear data. []

vi) K-Nearest Neighbors : sup for reg & classi tasks, Classify new instances based on the majority vote of their NN. "K" determines no/of neighbors considered for decision making []

vii) K-mean clustering : unsup algo for classi, regg. divides dataset into K clusters based on the similarity of instances. Adjust centeroids until convergence. []

ML uses : predictive analysis, NLP, CV, Robotics, etc...

a) Underfitting and Overfitting in Machine Learning:-

* Underfitting:-

Underfitting occurs when a machine learning model does not learn the training data well enough. This can happen for a number of reasons, including:

- The model is too simple and cannot capture the underlying patters in data
- The model is not trained on enough data.
- The data is noisy and corrupted.

Underfitting can be identified by looking at the model's performance on the training data and the test data. If the model performs poorly on both training and testing data, then it is likely that the model is underfitting.

* Overfitting:-

Overfitting occurs when a machine learning model learns the training data too well. This can happen for a number of reasons, including:

- The model is too complex and learns the noise in the data as well as the underlying patterns.
- The model is trained on too much data.
- The data is not representative of the real world.

Overfitting can be identified by looking at the model's performance on training and testing data. If the model performs well on the training data but poorly on the test data, then it is likely that the model is overfitting.

* Preventing Underfitting and Overfitting:

There are a no/of things that can be done to prevent underfitting and overfitting, including:

- Selecting the right model: The complexity of the model should be chosen based on the amount of training data and the complexity of the problem

Datafications:-

- Process of turning aspects of the physical world into data that can be analyzed. done through using sensors, cameras, etc..
- key part of development of IOT
- ndlof benifits, such as helping us to understand the world around us better and improve decision making. eg:- Traffic Traking, monitoring our healths,
- having risks, such as invading our privacy and being used to discriminate ppl
- eg:- smart phones : fitness trackers ; self driving cars.

Datafication rapidly growing → major impact on our lifes → in comming years.

Big data:

term used to describe data sets that are so large and complex that they are difficult (or) impossible to process using traditional data processing applications.

- Primary characteristics : Volume → Big data sets are extremely large (petabytes in size)
  - velocity → Big data sets are generated & Processed at high velocity
  - variety → structured (organized in regular way eg : databases)
    - unstructured (not organized in any way eg : text, img, video)
    - semi structured (not organized in regular way but have some structure, eg : XML, Json data),

eg:- social media data, sensor data, log data, etc...

Hypothesis testing :-

is an act in statistics whereby an analyst tests an assumption regarding a population parameter. statistical method to determine the whether there is significant diff b/w two sets of data. Two types : Null hypothesis (no diff b/w two sets of data) Ho
alternative hypothesis (There is diff b/w two data sets) H1

The goal of hypothesis testing is to determine whether there is enough evidence to reject the null hypothesis. if null hypo is rejected then alternative hyp. is accepted.

- Type I error : Rejecting the null hypothesis when it is actually true.

significance level is (α) Probability of making type I error, often sets at 0.05 or 0.01,

Type II error: Failing to reject the null hypothesis when it is actually false

Power of a test (β) the probability of Type II error calculated by $1-\beta$.

-) best way to reduce type I & II error is to increase the sample size which helps to detect diff blw two sets of data.

Clustering :-

unsupervied ML algorithm that grps data points into clusters based on similarity

the goal of clustering is to find grp of data pts similar to each other, and diff from data points in other clusters.

Types of clustering:

i) K-mean clustering:

is a simple and popular clustering alg that grps data pts into fixed no\|of clusters. The alg work by iteratively assigning data pts to clusters with the nearest mean, until the convergance.

steps :- i) Choose no\|of clusters → ii) Initialize cluster centroids → iii) Assign data pts to [redacted] cluster with nearest centroid → iv) update the mean of centroids in a cluster → v) repeate until convergance.

(of centroids, Produce suboptimal results)

-) k-mean is good choice for data sets with fixed no\|of clusters, sensitive in initial choice

ii) Hierarchical clustering:

is a type of clustering alg that builds a tree-like structure of clusters where each cluster can be a parent (or) child of other cluster. The algorithm starts by treating each data points as its own cluster. Then it repeatedly merge the similar clusters until there is only one cluster left.

-> It can be agglomerative (bottom-up) or divisive (top-down)

-> does not requries to specifies the no\|of clusters in advance.

iii) DBSCAN: (Density based spatial clustering of Applications with Noise)

identifies and grps data points that are densely packed, while considering pts that are not densely packed as noise. doesn't requries to specify the no\|of clusters in advance.

-) Capable of identifying clusters of diff shaps & size, computationall expensive, tough to choose esp, min pts.

esp : min distance of neighbourhood.

min pts : min pts within neighbourhood to make it as core pts

-) core points, boader pts, noise

Eg of clustering:- Customer segmentation, Data exploration, NLP, image segmentation, anomaly detection,

Python code for data cleaning :-

```
import pandas as pd

data = pd.read_csv ('data.csv')  #load the data set
missing_values = data.isnull().sum()  # Check for missing values.
Print ("Missing values:", missing_values)

data = data.dropmissing_values().  #remove missing values & duplicates.
data = data.drop_duplicates();

data ['Data'] = pd.to_datetime (data['Date']).  #convert data type.
data [' value'] = pd.to_numeric (data['Date value'])

data ['Name'] = data ['Name'].str.strip()  #remove trailing (or) leading whitespaces
data ['Name'] = data['Name'].str.lower()  # Convert to lower case.

data = data[(data['value'] > lower_threshold) & (data['value'] upper_threshold)]
data.to_csv ('cleaned_data.csv', index = False)  # export cleaned data into new CSV file
```

Student T-test :-

Student T-test is a statistical hypothesis test that is used to compare the mean of two independent group of data. The t-test calculates a t-value, which is a ratio of the difference between the two means to the standard error of the difference.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_\Delta} \quad \text{where} \quad s_\Delta = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

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

where $\bar{x}_1$ = mean of first set of values,

$\bar{x}_2$ = mean of second set of values,

s_1 = standard deviation of first set of values,

s_2 = " " second " " "

n_1 = Total no. of values in first set,

n_2 = " " second "

Data preprocessing :-

Is the process of transforming raw data into a format that can be easily analyzed by ML algorithms. It involves several steps, including data cleaning, integration, transformation and reduction.

i) Data cleaning :-

- process of identifing and removing (or) correcting errors, inconsistencies and the missing values from data.
- Techniques used for data cleaning includes imputation, deduplication, outlier removal and handling missing values.
- Example : filling in missing values with estimated values when there is missing data in the 'age' coloum of dataset, this is called imputation.

ii) Data integration :-

- Process of combining data from multiple sources and resolving inconsistencies or conflicts.
- Techniques used includes data fusion, data alignment and data merging,
- Example: merging of customer data based on common identifier, such as customer ID and then align the information.

iii) Data transformation :-

- Process of converting data into a format suitable for machine learning algorithms
- Techniques used includes normalization, feature scaling, encoding & data discretization
- Example : Converting categorical variable like colour into numerical value by encoding enables the ML algorithm to use in it directly.

iv) Data Reduction :-

- Process of reducing the amount of data while retaining as much relevant info as possible
- Techniques used includes feature selection, dimensionality reduction and data sampling.
- Example : Selecting a subset of features that are most relevant to the analysis by dimensionality reduction to increase computation speed of the model.

Statistics : is the study of the collection, analysis, interpretation, presentation and organization of data.

Eg :- mean of marks obtained by 50 students in a class.

→ Basics of statistics include the measure of central tendency (mean, median, mode) and the measure of dispersion (variance and standard deviation)

Types of Statistics :-

i) **Descriptive statistics :-**

These statistics are used to summarize and describe a data set. Histograms, pie charts, bar and scatter plots are common ways to summarize data and present it in table and graphs.

Here, the summarization is done using :-

- Measures of central tendency : mean, median, mode
- Measures of dispersion : range, variance, standard deviation.
- Measures of shape : skewness, kurtosis.

For instance, if we have a data set containing the weights of 20 people, we could use descriptive statistics to summarize and describe the data, such as calculating the mean weight, the range of weights, or the skewness of the weight distribution.

ii) **Inferential statistics :-**

These statistics are used to draw conclusions or make predictions about a population based on a sample of data.

Here, the prediction is done using :- (t-tests, ANOVA, chi-square test & linear regression)

- Hypothesis testing : testing whether a sample mean is significantly different from population mean
- Confidence intervals : estimating the range of values within which a population parameter (i.e. mean) likely to fall.
- Regression analysis : analyzing the relationship between variables and predicting values based on that relationship.

For example, we might use inferential statistics to test whether a new drug reduces cholesterol levels in a population : we would take a sample of People and randomly assign them to a treatment group or a control group, and then use statistical tests to determine wheather there is a significant difference in cholesterol levels between the two groups.

Probability :-

Probability is a mathematical concept that predicts how likely events are to occur. The probability values are expressed between 0 & 1. The definition of probability is the degree to which something is likely to occur.

These are mainly two types of probability distributions :

i) Discrete probability :-

These distributions are associated with random variables that can only take on a finite or countable number of values, such as the outcome of a coin toss or the number of defective items in a batch of products.

ii) Continuous probability :-

These distributions on other hand are associated with random variables that can take on any value within a range, such as the height or weight of individuals in a population.

-> Examples of common probability distribution include the normal distribution (also known as Gaussian distribution), the binomial distribution, the poisson distribution and the exponential distribution.

-> Probability distributions are widely used in fields such as statistics, physics, engineering, finance, and many other areas of science and industry to model and analyze a wide range of phenomena.

SVD (singular value decomposition) :-

$A_{mxn} = U_{mxm} \cdot \Sigma_{mxn} \cdot N^T_{nn}$

rotation stretching rotation

here U & V are orthogonal i.e $U \cdot U^T = I$, $V \cdot V^T = I$.

& Σ is a rectangular diagonal matrix with singular values

$A^T \cdot A = (V \cdot \Sigma^T \cdot U^T)(U \cdot \Sigma \cdot V^T)$

$= V \cdot \Sigma^T \cdot \Sigma \cdot V^T$

$A \cdot A^T = (U \cdot \Sigma \cdot V^T)(V \cdot \Sigma^T \cdot U^T)$

$= U \cdot \Sigma \cdot \Sigma^T \cdot U^T$

rotation : $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ → unitary transformation

stretching : $\begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}$

$A^T \cdot A = V$

$A \cdot A^T = U$

eg:- $A = \begin{vmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{vmatrix}_{2x3}$

$A^T = \begin{vmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{vmatrix}_{3x2}$

$A \cdot A^T = \begin{vmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{vmatrix} \begin{vmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{vmatrix} = \begin{vmatrix} 11 & 1 \\ 1 & 11 \end{vmatrix}$

$\lambda^2 - S_1\lambda + S_2$

$\Rightarrow |A - \lambda I| x = 0$

$\Rightarrow \begin{vmatrix} 11-\lambda & 1 \\ 1 & 11-\lambda \end{vmatrix} = 0$ $\Rightarrow (11-\lambda)^2 - 1^2 = 0$ $\Rightarrow (11-\lambda)^2 = 1$

$\Rightarrow \pm(11-\lambda) = 1$

$\Rightarrow \lambda = 10 ; \lambda = 12$

Put λ values in above equation.

$\begin{vmatrix} 11-10 & 1 \\ 1 & 11-10 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} = 0$ $\Rightarrow \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix} \begin{matrix} x_1 \\ x_2 \end{matrix}$ $\Rightarrow$ $x_1 + x_2 = 0$ $\therefore x_1 = -x_2$

$x_1 + x_2 = 0$

$\begin{vmatrix} 11-12 & 1 \\ 1 & 11-12 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} = 0$ $\Rightarrow \begin{vmatrix} -1 & 1 \\ 1 & -1 \end{vmatrix} \begin{matrix} x_1 \\ x_2 \end{matrix}$ $\Rightarrow$ $-x_1 + x_2 = 0$ $\therefore x_1 = x_2$

$x_1 - x_2 = 0$

we need to write in decending order so, $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$

now calculating orthogonal, for above $\Rightarrow \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{matrix} \rightarrow \sqrt{2} \\ \rightarrow \sqrt{2} \end{matrix}$ $\Rightarrow \begin{vmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{vmatrix}$

now for V : $A^T.A$.

$$\Rightarrow \begin{vmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{vmatrix} \begin{vmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{vmatrix} \Rightarrow \begin{vmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{vmatrix} \longrightarrow \lambda^3 - S_1\lambda^2 + S_2\lambda - S_3 = 0$$

$$S_1 = 22$$

$$S_2 = 100 + (+4) + 16 = 120$$

$$S_3 = 10(4) - 0 + 2(-20) = 0$$

$$\Rightarrow \lambda^3 - 22\lambda^2 + 120\lambda = 0$$

$$\Rightarrow \lambda(\lambda^2 - 22\lambda + 120) = 0 \qquad \therefore \lambda = 0 ; \lambda = 10; \lambda = 12$$

$$\Rightarrow \lambda^2 - 12\lambda - 10\lambda + 120 = 0 \Rightarrow \lambda(\lambda - 12) - 10(\lambda - 12) = 0 \Rightarrow (\lambda - 10)(\lambda - 12).$$

$\therefore \lambda_1 = 12 ; \lambda_2 = 10 ; \lambda_3 = 0$

by cramers rule :-

$$\therefore \text{for } \lambda = 12 :- \begin{vmatrix} (0-12) & 0 & 2 \\ 0 & (10-12) & 4 \\ 2 & 4 & (2-12) \end{vmatrix} \begin{matrix} x_1 \\ x_2 \\ x_3 \end{matrix} = 0 \Rightarrow \frac{x_1}{4} = \frac{-x_2}{-8} = \frac{x_3}{4}$$

$$\begin{matrix} x_1 \\ x_2 \\ x_3 \end{matrix} = \begin{vmatrix} 1 \\ 2 \\ 1 \end{vmatrix}$$

$$\text{for } \lambda = 10 :- \begin{vmatrix} 0 & 0 & 2 \\ 0 & 0 & 4 \\ 2 & 4 & -2 \end{vmatrix} \begin{matrix} x_1 \\ x_2 = 0 \\ x_3 \end{matrix} \Rightarrow \frac{x_1}{-16} = \frac{-x_2}{-8} = \frac{x_3}{0} \Rightarrow \begin{matrix} x_1 \\ x_2 \\ x_3 \end{matrix} = \begin{vmatrix} 2 \\ -1 \\ 0 \end{vmatrix}$$

$$\text{for } \lambda = 0 :- \begin{vmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{vmatrix} \begin{matrix} x_1 \\ x_2 = 0 \\ x_3 \end{matrix} \Rightarrow \frac{x_1}{4} = \frac{-x_2}{-8} = \frac{x_3}{-20} \Rightarrow \begin{matrix} x_1 \\ x_2 \\ x_3 \end{matrix} = \begin{vmatrix} 1 \\ 2 \\ -5 \end{vmatrix}$$

$$V = \begin{vmatrix} 1 & 2 & 1 \\ 2 & -1 & 2 \\ 1 & 0 & -5 \end{vmatrix} \Rightarrow \text{orthogonal } V = \begin{vmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{vmatrix}$$

$$V^T = \begin{vmatrix} 1 & 2 & 1 \\ 2 & -1 & 0 \\ 1 & 2 & -5 \end{vmatrix}$$

$$\text{orthogonal } V = \begin{vmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{vmatrix}$$

where λ_1 & λ_2 are 12 & 10 as they are common in both AA^T ; A^TA and should be in descending orde

Σ = diagnol matrix with same dimensionality as A $\Rightarrow \begin{bmatrix} \sqrt{\lambda_1} & 0 & 0 \\ 0 & \sqrt{\lambda_2} & 0 \end{bmatrix}$

$\begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{bmatrix}$ ← decending orde

$$\therefore A = \begin{vmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{vmatrix} \begin{vmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{vmatrix} \begin{vmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{vmatrix}$$

a) PCA for

x	2.5	0.5	2.2	1.9	3.1	2.3	2.0	1.0	1.5	1.1
y	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

$n = 10$

x	y	$x_i - \bar{x}$ (A)	$y_i - \bar{y}$ (B)	AB	A^2	B^2
2.5	2.4	0.69	0.49	0.3381	0.4761	0.2401
0.5	0.7	-1.31	-1.21	1.5851	1.7161	1.4641
2.2	2.9	0.39	0.99	0.3861	0.1521	0.9801
1.9	2.2	0.09	0.29	0.0261	0.0081	0.0841
3.1	3.0	1.29	1.09	1.4061	1.6641	1.1881
2.3	2.7	0.49	0.79	0.3871	0.2401	0.6241
2.0	1.6	0.19	-0.31	-0.4589	0.0361	0.0961
1.0	1.1	-0.81	-0.81	0.6561	0.6561	0.6561
1.5	1.6	-0.31	-0.31	0.0961	0.0961	0.0961
1.1	0.9	-0.71	-1.01	0.7171	0.5041	1.0201
				5.539	5.549	6.449

$\bar{x} = 1.81$ $\bar{y} = 1.91$

Covariance matrix $= \begin{bmatrix} Cov(x,x) & Cov(x,y) \\ Cov(y,x) & Cov(y,y) \end{bmatrix}$ ∴ here $Cov(x,y) = Cov(y,x)$

where; $Cov(x,x) = \sum_{n=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$; $Cov(x,y) = \sum_{n=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

⇒ Covariance matrix (A) $= \frac{\begin{bmatrix} 5.549 & 5.539 \\ 5.539 & 6.449 \end{bmatrix}}{9} = \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix}$

now, we need to find eigen values & eigen vectors for the above covariance matrix.

$$\Rightarrow |A - \lambda I| = 0$$

$$\Rightarrow \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = 0$$

$$\Rightarrow \begin{bmatrix} (0.6166 - \lambda) & 0.6154 \\ 0.6154 & (0.7166 - \lambda) \end{bmatrix} = 0 \quad \text{---eq(1)}$$

By finding det for above matrix, we will get an equation, which can be solved to get the eigen values:-

$$\Rightarrow (0.6166 - \lambda)(0.7166 - \lambda) - (0.6154)^2 = 0$$

$$\Rightarrow 0.441 - 0.7166\lambda - 0.6166\lambda + \lambda^2 - 0.378 = 0$$

$$\Rightarrow \lambda^2 - 1.3332\lambda + 0.063 = 0$$

now, we will find roots for above equation :- $\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

$$\Rightarrow \underline{\lambda_1 = 1.283} \quad \& \quad \underline{\lambda_2 = 0.049}$$

now putting λ_1 in eq---(1) for eigen vectors

$$\begin{bmatrix} (0.6166 - 1.283) & 0.6154 \\ 0.6154 & (0.7166 - 1.283) \end{bmatrix}$$

(or)

eigen vectors can be found by using :- $CV = \lambda V$

$$\Rightarrow \begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 1.283 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

$$\Rightarrow 0.6166x_1 + 0.6154y_1 = 1.283x_1 \Rightarrow x_1 = 0.923y_1$$

$$\Rightarrow 0.6154x_1 + 0.7166y_1 = 1.283y_1$$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 0.923 \\ 1 \end{bmatrix}$$

$$\therefore \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} \frac{x_1}{\sqrt{x_1^2+y_1^2}} \\ \frac{y_1}{\sqrt{x_1^2+y_1^2}} \end{bmatrix} = \begin{bmatrix} \frac{0.923}{\sqrt{1.851}} \\ \frac{1}{\sqrt{1.851}} \end{bmatrix} = \begin{bmatrix} \frac{0.923}{1.36} \\ \frac{1}{1.36} \end{bmatrix} = \begin{bmatrix} 0.678 \\ 0.735 \end{bmatrix}$$

for $\lambda = 0.049$:- $CV = \lambda V$

$$\begin{bmatrix} 0.6166 & 0.6154 \\ 0.6154 & 0.7166 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0.049 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

$\Rightarrow 0.6166x_1 + 0.6154y_1 = 0.049x_1 \quad \Rightarrow \quad x_1 = \frac{-0.6154}{0.5676} y_1$

$0.6154x_1 + 0.7166y_1 = 0.049y_1$

$\Rightarrow x_1 = -1.084 y_1$

$$\Rightarrow \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} \frac{-1.084}{\sqrt{2.175}} \\ \frac{1}{\sqrt{2.175}} \end{bmatrix} = \begin{bmatrix} \frac{-1.084}{1.474} \\ \frac{1}{1.474} \end{bmatrix} = \begin{bmatrix} -0.735 \\ 0.678 \end{bmatrix}$$

SVD :- $A_{m \times n} \rightarrow$ rank (r)

$\Sigma_{m \times n}$; $U_{m \times m}$; $V_{n \times n}$

$\Rightarrow A = U \Sigma V^T$

- from given "A" find $A^T.A$ & find eigen value for that matrix and also find σ values [like $\sigma_1 = \sqrt{\lambda_1}$; $\sigma_2 = \sqrt{\lambda_2}$, etc...].
- from eigen values find eigen vectors and place beside each other by normalizing them to get "V".
- find Σ, as it is a diagnol matrix of values "σ" & make its rank equal to A i.e Rank(Σ) = Rank(A).
- now $U_1 = \frac{1}{\sigma_1} AV_1$ & $U_2 = \frac{1}{\sigma_2} AV_2$ $\Rightarrow$ $U = [U_1, U_2]$

$A = U \Sigma V^T$

$U \rightarrow AA^T$; $\Sigma \rightarrow$ diagnol matrix ; $V^T \rightarrow A^TA$

U = normalized Eigen vector matrix AA^T

V^T = normalized eigen vector matrix A^TA.