Vidya Pratishthan's

Kamalnayan Bajaj Institute of Engineering and Technology

Vidyanagari, Baramati, Pune 413 133.

INDEX

Department of :- Artificial Intelligence And Data Science CLASS T.E

SR. NO.	NAME OF EXPERIMENT	Expt. Conducted on	Expt. Checked on	PAGE NO.	SIGN	REMARK
1.	Data-Wrangling-1	2/2/23	5/2/23	1		
2.	Data-Wrangling-2	9/2/23	12/2/23	3		
3.	Descriptive Statistics	16/2/23	17/2/23	6		
4.	Data Analytics-1	23/2/23	25/2/23	12		
5.	Data Analytics-2	2/3/23	3/3/23	15		
6.	Data Analytics-3	9/3/23	11/3/23	19		
7.	Text Analytics	16/3/23	27/3/23	22		
8.	Data Visualization-1	23/3/23	27/3/23	28		
9.	Data Visualization-2	6/4/23	15/4/23	31		

Rahul

CERTIFICATE

This is to certify that Mr./Miss Devrat Yashraj Deepak of

Class T.E(A.I.D.S) Roll No. 223031 has Satisfactorily Completed the term work of the subject.

Software Laboratory-3 for VI Semester of 20 22 - 20 23

Rahul

R.328

Date : / /

Staff Member / In-charge

H.O.D.

PRINCIPAL

Vidya Pratishthan's

Kamalnayan Bajaj Institute of Engineering and Technology

Vidyanagari, Baramati, Pune 413 133.

INDEX

Department of :- Artificial Intelligence And Data Science CLASS T.E

SR. NO.	NAME OF EXPERIMENT	Expt. Conducted on	Expt. Checked on	PAGE NO.	SIGN	REMARK
10.	Data Visualization-3	20/4/23	22/4/23	36	Rahul	
11.	Simple Program Scala	27/4/23	30/4/23	40		
12.	GINA.	30/4/23	1/5/23	42		
13.	Mini Projects	1/5/23	8/5/23	46		

CERTIFICATE

This is to certify that Mr./Miss Devrat Yashraj Deepak of Class T.E (A.I.D.S) Roll No. 223031 has Satisfactorily Completed the term work of the subject Software Laboratory 3 for VI Semester of 2022-2023

Date : / /

Rahul
Staff Member
In-charge

H.O.D.

PRINCIPAL

VPKBIET

Assignment No. 1 6/2/23

Title: Data Wrangling - 1

Objective: Perform the following operations using Python on any open source dataset (e.g. data.csv)

1. Import all the required Python libraries.
2. Locate open source data from web (e.g. https://www.kaggle.com) Provide clear description of the data and its source.
3. Load the Dataset into pandas dataframe.
4. Data Preprocessing: Check the missing values in the data using pandas isnull(), describe() function to get some statistics. Check dimension of data.
5. Data Formatting and Data Normalization: Summarize the types of variables by checking the data type (i.e. character, numeric, integer, factor and logical)
6. Turn categorical variables into quantitative variables using python.

Theory:

Data Wrangling: It is the process of gathering and collecting and transforming raw data into another format for better understanding, decision making, accessing and analysis in less time. Data Wrangling is also known as Data Munging.

Pandas: Pandas is a Python library used for working with datasets. It has function for analyzing, cleaning, exploring and manipulating data.

Data Preprocessing:

1. Check for missing values in the data using pandas isnull(): The isnull() method returns a DataFrame object where all the values are replaced with a Boolean value True or NULL/NA values, and otherwise False.

```
dataframe.isnull()
```

The notnull() method returns a DataFrame object where all the values are replaced with boolean value True for NOT NULL values, otherwise False.

```
dataframe.notnull()
```

2. describe(): The describe() method returns description of the data in the DataFrame. If the DataFrame contain numerical data, the description contain these information for each column:

Count - The number of non-empty values.
mean - average.
Std - Standard deviation

min - the minimum value
25 % - The 25 % percentile.
50 % - The 50 % percentile.
75 % - The 75 % percentile.
max - maximum value.

⑧. Pandas .Size, .Shape & n.dim are used to return Size, Shape and dimension of data frames & Series.

Syntax : dataframe.Size

Return : Return Size of dataframe/Series which is equivalent to total number of elements This is rows X columns.

Syntax : dataframe.Shape

Return : Return Shape (Rows, Columns) of dataframe

Syntax : dataframe.ndim

Return : Return dimension of dataframe/series. 1 for one dimension (Series), 2 for two dimension (dataframe)

VPKBIET

Data formatting & Normalization:

type() function is used to determine the type of data type.

Dtype() function returns the data type of the array.

Turn Categorical variables into quantative variables in Python:

Using replace() method

Replacing is one of the methods to convert categorical terms into numeric. For example. We will take a dataset of peoples salaries based on their level of education. This is an ordinal type of categorical variable. We will convert their education level into numeric terms.

replace(to_replace = None, value = None, inplace = False, limit = None, regex = False, method = 'pad')

Method 2: using get_dummies()
Replacing the values is not the most efficient way to convert them. Pandas provide a method called get_dummies which will return the dummy variable columns.

DATE

PAGE NO

Assignment No. 2 16/2/23

Title : Data Wrangling - 2

Objective : Create an "Academic Performance" dataset of students and perform the following operations using Python.

1. Scan all variables for missing value and inconsistencies. If there are missing values and/or inconsistencies, use any of the suitable techniques to deal with them.

2. Scan all numeric variables for outliers. If there are outliers, use any of the suitable techniques to deal with them.

3. Apply data transformation on at least one of the variables. The purpose of this transformation should be one of the following reasons:
to change the scale for better understanding of the variables, to convert non-linear relation into linear one, or to decrease the skewness and convert the distribution into a normal distribution.

Theory:

Missing Values (Data Cleaning):

Data Cleaning can be applied to remove data noise and Correct inconsistencies in data. Data cleaning routine work to "clean" the data by filling the missing values, smoothing noisy data correcting inconsistencies in data. If the data is dirty then it could lead to inaccurate results.

In pandas missing values are represented by

None : Python singleton object used for missing data in Python.

NaN : Not a Number Special floating point value recognized by all systems that use the standard IEEE floating point representation. Functions for detecting, removing, replacing until values in Pandas Dataframe.

isnull() : returns True for NaN values
notnull() : returns False for NaN values
dropna() : delete all Null values
fillna() : replaces NaN values with some value of their own.
replace() : Same as fillna().

Outliers : It is a data object that deviates significantly from the rest from the data objects & behaves in a different manner an outliers is an object that deviates significantly from the rest of the objects.

VPKBIET

Handling of Outliers:

① Trimming / removing the outliers:
We remove the outliers from the dataset.

② Quartile based flooring & capping:
Outliers is capped at certain value above the 90th percentile value / floored at a factor below the 10th percentile value.

③ Mean / Median imputation:
As the mean value is highly influenced by the outlier, it is advised to replace the outliers with the median value.

④ Data Transformation:
In data transformation, the data is transformed or consolidated into appropriate for mining.
Data transformation, the data involve the following

a. Smoothing:
It removes noise from the data such as techniques including regression & clustering. It helps in the predicting the patterns and collecting data.

b. Aggregation:
Data collection or aggregation is the method storing and presenting data in summary format. The data may be obtained from multiple data sources

to integrate data sources into a data anyltics description.

c. Discretization:

It is a process of transforming continous data into set of small variables, most data mining activities in the real world require continuous attributes.

d. Attribute Construction:

Where new attributes are created and applied to assist mining process from the given set of attributes. This simpligy the orginal data & makes mining more efficient.

e. Generalization:

It converts two level of converting all data variable to high level data attributes using concept hirarchy.

f. Normalization:

Data normalization involves converting all data variable into a given range. Techniques of normalization.

1. Min-Max

$$v' = \frac{V - min_A}{max_A - min_A} (new\_max_A - new\_min_A) + new\_min$$

2. Z-Score:

This value /score helps to understand how far the data point from mean.

Z-Score = (data point - mean) / Std. deviation

3. Decimal Scaling:

In this technique, we move the decimal point of values of the attribute.

A value v of attribute A is normalized by the following formula.

Normalized value of attribute: $\left(v^i / 10^i\right)$

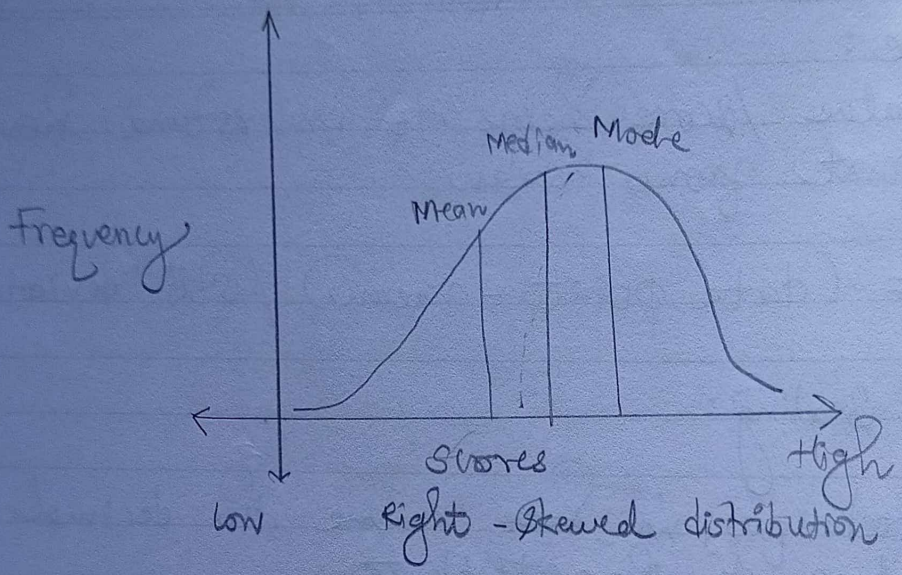
4. Scaling:

In scaling we are changing the range of the distribution of the data.

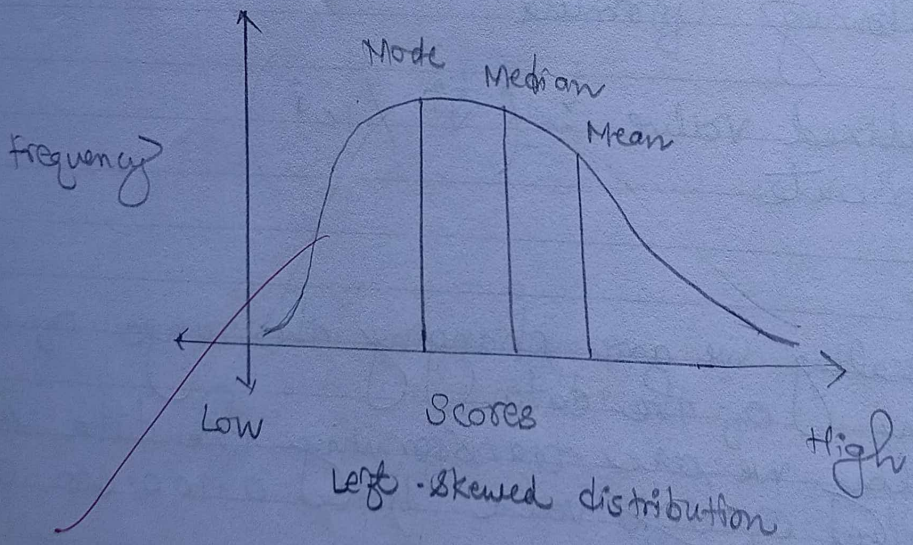
In scaling we are transforming the data so that it fits value specific value like 0-100 or 0-1.

Common types of scaling:

$$X_{new} = \frac{X_{old}}{X_{max}}$$



Right-Skewed distribution



Left-Skewed distribution

b. Min-Max Scaling:

$$X_{new} = \frac{X_{old} - X_{min}}{X_{max} - X_{min}}$$

This is the most popular & simple. Feature scaling, it takes each value & subtracts the minimum & then divides by the range (max - min)

Skewness:

If the value of a specific independent variable are skewed depending on the model, skewness may violate model assumption or may reduce the interpretation of feature importance.

"Skewness is a degree of asymmetry observed in a probability distribution that deviates from the symmetrical normal distribution in given set of data".

Mean = Median = Mode = 0

Types of Skewness:

1. Positive Skewed or right Skewed (Positive Skewness)

A positively skewed or right skewed distribution

has a long right tail, it is a sort of distribution where unlike symmetrically distributed data where all measures of the central tendency (mean, mode, median) equal each other, with positively skewed data, the measures of the central tendency

Positively skewed distribution is a type of distribution where mean, mode, median of distribution are positive rather than negative or zero.

Mean > Median > Mode.

2. Negative skewed or left skewed (Negative skewness)

A negatively skewed or left skewed distribution has long left tail. It is in the straight across reverse of positively skewed distribution. In statistics, negatively skewed distribution. Plotting right side of graph and the tail of the distribution is spreading on the left side.

Mode > Meadin > Mean

Rahul 16/2/23

Conclusion: Thus we implemented data wrangling 2. by applying data transformation on dataset also checking for missing values and handled it. Identifying outliers and handled with it using mentioned techniques.

VPKBIET

Assignment No. 5 19/2/2023

Title : Descriptive Statistics - Measures of Central Tendency and Variability.

Aim : Study the various measures of central tendency and variability with respect to the dataset.

Objectives : Perform the following operations on any dataset (e.g. data.csv).

1. Provide summary statistics (mean, median, mode, minimum, maximum, standard deviation) for a dataset (age, income etc) numeric variables grouped by one of the qualitative (categorical) variable.

   eg. Categorical Variable → age.

   quantative Variable → income.

   Provide summary statistics for income grouped by age. Create list of categorical variable.

2. Write a python program to display some basic statistical details like percentile, mean, standard deviation etc. of species of 'Iris-setosa', 'Iris-versicolor' & of 'iris.csv dataset.

VPKBIET

Theory: Central Tendency and variation are two measures of descriptive summary used in Statistics to summarize data.

Measures of Central Tendency shows the centre or middle of the dataset is located, whereas measures of variation shows the dispersion among data values.

Measures of Central Tendency:

"The central tendency is used to measure & represent whole dataset using single value.

1. Mean: The mean is average of given observations is defined as the sum of all the observations divided by the total number of observations.

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

function to calculate mean: numpy.mean()
e.g. df.mean()

2. Median: The median is central or middlemost value of dataset.

Odd number of observations = $\left(\frac{n+1}{2}\right)^{th}$ observation.

Even number of observations = average of $\frac{n}{2}^{th}$

$\left(\frac{n}{2}+1\right)^{th}$ observation.

$$\text{Median} = L + \left(\frac{\frac{n}{2} - C.F}{f}\right) \times h$$

L - Lower limit of median class, h = class size
C.F - Cummulative frequency, f - frequency

e.g. df.median()

⑥ Mode: The mode is the most repeated value of among the given set of data: among the given set of observation, mode is the value which has the maximum frequency.

$$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

f_1 is frequency of modal class, h is class size.
f_0 is frequency of the class preceding modal class.
f_1 is frequency of the class succeding modal class.

e.g. df.mode()

VPKBIET

Measures of Variability / Dispersion:

A measure of dispersion is a summary statistic that represents the amount of dispersion in a dataset. How spread out are the value? High dispersion signifies that they tend to fall further away.

① Standard Deviation:

The Standard deviation is the average amount of dispersion in your dataset.

① $$\sigma = \sqrt{\frac{\Sigma (X-\mu)^2}{N}}$$

σ = population Standard deviation.

Σ = Sum of

μ = population mean, N = number of values in population.

Formula Explanation

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

s = Sample Standard deviation.

Σ = Sum

X = each value

$\bar{x}$ = Sample mean.

VPKBIET

eg. df.Std()

2. Variance : The variance is the average of squared deviations from the mean. Variance is the square of standard deviation.

Formula

Explanation

① $\sigma^2 = \frac{\sum (x-\mu)^2}{N}$

σ^2 = Variance

μ = mean

n = number of values in sample.

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

eg df.Var()

Percentile : In Statistics, a percentile is a term that describes how a score compares to other scores from the same set.

$$P_x = \frac{x(n+1)}{100}$$

Function to calculate percentile: percentile()

e.g. df.percentile()

Groupby() Function in Python:

Pandas dataframe.groupby() function is used to split the data into groups, based on some criteria. Pandas objects can be split on any of their axes.

e.g. df.groupby(["car"]).mean()

Conclusion: Central Tendency and Variation are two measures used in Statistics to summarize data. Measures of Central tendency shows where the centre or middle of the data is located, whereas measure of variations shows the dispersion among data values.

Refer
21/2/23

VPKBIET

Assignment No. 4 2/3/2023

Title: Data Analytics-1

Objective: To predict the value of prices of the houses using given features. Create a linear regression model using Python to predict home prices using Boston Housing Dataset. (https://www.kaggle.com/c/boston-housing). The Boston Housing dataset contains information about various houses in Boston through different parameters. There are 506 samples & 14 feature variables in this dataset.

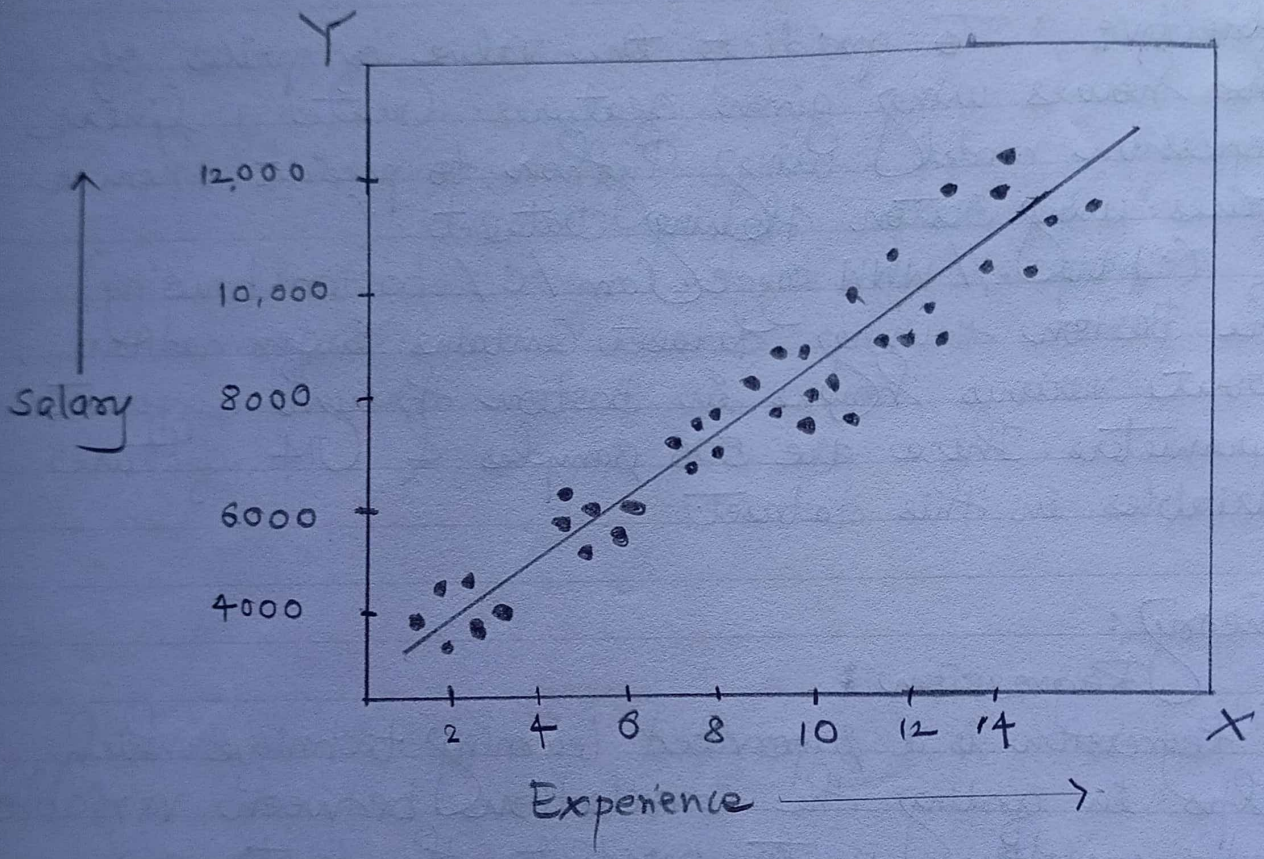
Theory:

Regression:

Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.

What is linear Regression?

The objective of a linear regression model to find a relationship between one or more features (independent variables) and a continuous target variables (dependent variables). When there is only feature it is called Uni-variate / Simple



Equation: $Y = aX + b$

Here, Y = dependent variables (target variables)
X = Independent variables (predictor variables)
a, b are the linear coefficients.

Linear Regression and if there are multiple features it is called Multiple Linear Regression.

Linear Regression is a Statistical regression method which is used for predictive analysis.

It is one of the very simple and easy algorithms which is used for predictive analysis and shows a relationship between the continuous variables.

- The relationship between variables in the linear regression model can be explained using the below image. Here we are predicting the salary of an employee on the basis of the year of experience.

Where can be Linear Regression be used ?

It is very powerful technique and can be used on to understand the factors the influence profitability. It can be used to forecasts sales in the coming months by analyzing the sales data for previous months. It can also be used to gain various insights about customer behaviour. By the end of the day we will build a model which looks like the below picture i.e determine a line which best fits the data.

Some popular applications of linear regression are:

- Analysing trends and sales estimates.
- Sales forecasting

- Real estate prediction.
- Arriving at ETA's in traffic.

Evaluation:

1. RMSE : It measures the average difference between values predicted by a model and actual values. It provides an estimation of how well the model is able to predict the target value (accuracy). The lower the value of the Root Mean Squared Error, the better the model is

2. MSE : The Mean Squared Error measures how close regression line is to a set of data points. It is risk function corresponding to the expecting value of the squared error loss. Mean Square Error is calculated by taking the average, specifically the mean of errors squared from data as it relates to a function. Lesser the MSE ⇒ Smaller is the error ⇒ Better the estimator.

3. AE = Absolute error refers to the magnitude of difference the prediction of an observation and true value that of observation.

Conclusion: Thus, we perqormed, data analytics, by creating linear regression model to predict house prices. We have also evaluated perqormance of the model.

6/3/23

VPKBIET

Assignment No. 5

Title : Data Analytics 2.

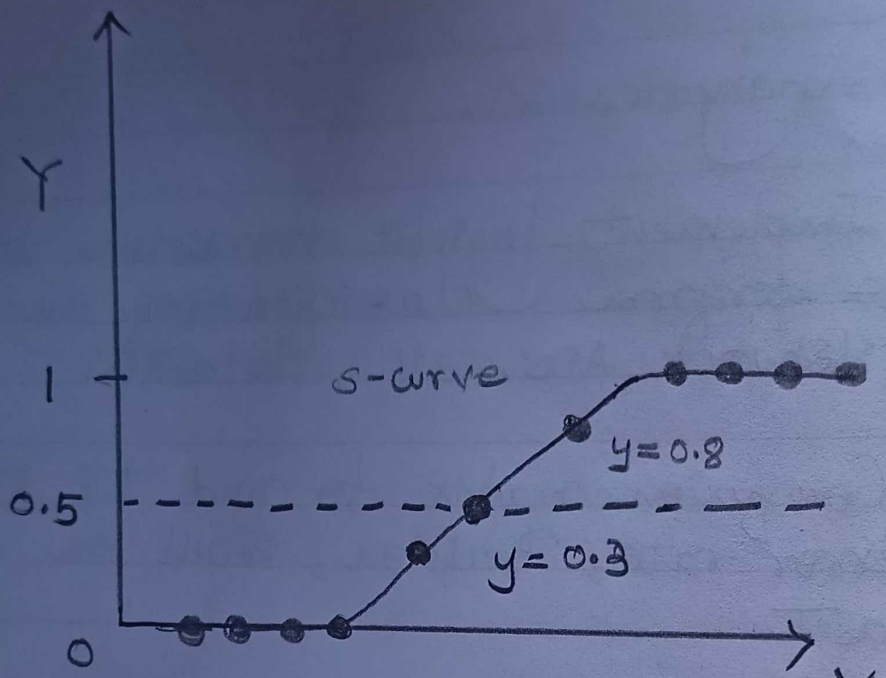
Objective : 1. Implement logistic regression using Python to perform classification on Social_Network_ADS.csv dataset.

2. Compute Confusion matrix to find TP, FP, TN, FN Accuracy, Error rate, Precision, Recall on the given dataset.

Theory:

What is Logistic Regression?

- It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic Regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or false. But instead of giving the exact value as 0 or 1. it gives the probabilistic values which lies between 0 or 1.
- In Logistic Regression, instead of fitting regression line, we fit an "S" shaped logistic function which predicts two maximum values (0 or 1).
- In Logistic Regression, indicates curves curve the



Logistic Function

likelihood of something such as whether the cells are cancerous or not a mouse is obese or not based on its weight, etc.

- Logistic Regression is significant machine learning algorithm because it has the ability to provide probabilities and classify the observation new data using continuous and discrete datasets.

- Logistic Regression can be used to classify the observation using different types and can easily determine the most effective variables used for the classification. The below image is showing the logistic function.

The logistic Regression equation can be obtained from the linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

We know the equation of the straight line can be written as:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \dots + b_n x_n$$

In logistic Regression y can be between 0 & 1 only so for this lets divide the above equation by (1-y):

$\frac{y}{1-y}$; 0 for $y=0$; and Infinity for $y=1$

① Accuracy : $\frac{TN + TP}{TN + TP + FN + FP}$

② Error rate : $1 - \text{Accuracy} = \frac{FN + FP}{TN + TP + FN + FP}$

③ Precision : $\frac{TP}{FP + TP}$

④ Recall : $\frac{TP}{TP + FN}$

⑤ F1 Score : $\frac{2 \times (\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{precision})}$

But we need range between $(-\infty$ to $+\infty)$, then take algorithm of the equation. It will become,

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots\ldots + b_n x_n$$

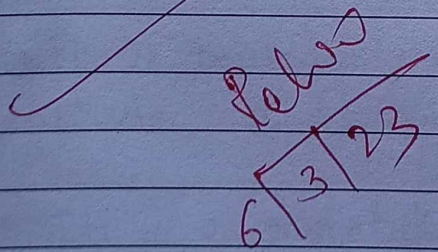
The above equation is the final equation Logistic Regression.

- Accuracy: It defines the ratio of correctly predicted with all parameters.
- TN : It stands for True Negative The case is negative & predicted positive.
- FN : (False negative) the case was predicted negative and positive.
- FP : (False positive) the case was negative but predicted positive.
- TP : (True Positive) Case was positive & predicted positive.
- Precision : Accuracy of positive predictions.

- Recall : It is the ability of a classifier to find all positive instances. It is defined as ratio of True positives to the sum of true positives & false negatives.

- F1-Score : It is harmonic mean of precision & recall such that best score is 1 & worst is 0.0. F1 scores are less than accuracy measure as they embed precision & recall into their computation.

★ Conclusion : Thus we performed data analytics by creating logistic regression model and computed Confusion matrix to find TP, FP, TN, FN, accuracy, error rate, Precision, Recall.



Data Wra - 1
Dat Wray - 2

VPKBIET

Assignment No. 6:

Title: Data Analytics 6.

Objective: 1. Implement simple Naive Bayes classification algorithm using Python on iris dataset

2. Compete Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall on the given dataset.

Theory:

Naive Bayes Classifier Algorithm:

- Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that include a high-dimensional training set.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of object.
- Some popular examples of Naive Bayes Algorithm are spam filtration, sentimental analysis, and classifying articles.

Naive Bayes
Data Vis - 1 (Hidden
Data Vis - 2 Hidden
Descriptive Stat -

VPKBIET

Bayes Theorem:

Bayes theorem is also known as Bayes law; which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability probability.

The formula for Bayes theorem:

$$P(A/B) = \frac{P(B/A)\ P(A)}{P(B)}$$

Where,

$P(A/B)$ is posterior probability: Probability of hypothesis A on the observed event B.

$P(B/A)$ is likelihood probability: Probability of the evidence given that the probability of a hypotheses is true.

$P(A)$ is Prior Probability: Probability of hypothesis before observating the evidence.

$P(B)$ is Marginal Probability: Probability of the Evidence.

VPKBIET

Types of Naive Bayes Model:

These are types of Naive Bayes model, which is given below:

- Gaussian: The Gaussian model assumes that the features follow a normal distribution. This means if predictions take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.

- Multinomial: The multinomial Naive Bayes classifier is used when data is multinomial distributed. It is primarily used for the document classification problems, it means a particular document belongs to which category such as sports, politics, education.

The classifier uses for frequency of words the predictors

- Bernoulli: The Bernoulli classifier works similar to the multimodal classifier, but the predictor variables are the independent Boolean variables. Such as if a particular word is present or not in a document. This model is also famous for document classification tasks.

Steps to implement:

- Data Pre-processing step.
- Fitting Naive Bayes to the Training set:

- predicting the test result.
- Test accuracy of the result (Creation of confusion matrix)
- Visualizing the test set result.

Conclusion: Thus, we performed, data analytics by creating Naive Bayes Classification on iris dataset and computed confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall.

Rahul
29/3/23

VPKBIET

Assignment No. 7

Title : Text Analytics

Objective : 1. Extract, Sample, document and apply following document pre-processing methods:

Tokenization, POS Tagging, Stop words removal, Stemming Lemmatization.

2. Create representation of documents by Calculating Term Frequency and Inverse Document Frequency.

Theory : Tokenization : In this step, the text is split into smaller units.
Output : Sentences are tokenized into words.

POS Tagging : It is also called grammatical tagging is the process of marking up a word in text as corresponding to a particular part of speech, based on both its definition and its context.

Stop words removal : Stopwords are the commonly used words and are removed from the text as they do not add any value to the analysis.
eg. [i, me, my, myself, we, our, etc.]

Stemming: It is also known as the text standard step where the words are stemmed or diminished to their root / base form. For example, words like 'programmer', 'programming', 'program' will be stemmed to 'program'.

Lemmatization: It stems the word but makes sure that it does not lose its meaning. It has a predefined dictionary that stores the context of words in the dictionary while diminishing.

Introduction TF_IDF

TF_IDF stands for 'Term Frequency - Inverse Document Frequency'. This is technique to quantify words in a set of documents. we generally compute a score for each word to signify its importance in the document and corpus.

Terminology t - term (word) d - document (set of words) N - Count of Corpus - the total document set

Term Frequency: This measures the importance of the documents in a whole set of corpus. This is very similar to TF but the only difference is that TF is the frequency counter for a term t in document d, whereas DF is the count of term t in the document set N.

DF(t) = Occurence of T in N documents.

Inverse Document Frequency: IDF is the inverse of the document frequency which measures the informativeness of term t. When we calculate IDF, it will be very low for the most occurring words such as stop words.

$$idf(t) = N / df$$

Finally, by taking a multiplicative value of TF and IDF we get the TF-IDF score. There are many different variations of TF-IDF but for now, let us concentrate on this basic version.

$$tf\text{-}idf(t, d) = tf(t, d) * \log(N/(df+1))$$

Conclusion: Thus, we performed, text analytics by initially performing pre-processing on the texts and then applying TF-IDF on it.

Review
27/4/23.

VPKBIET

Assignment No. 8

Title: Data Visualization I.

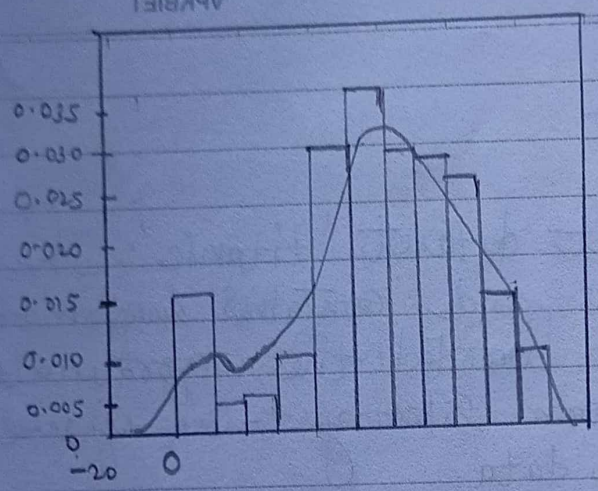
Objective: 1. Use the inbuilt dataset 'titanic'. The dataset contains 891 rows and contains information about the passengers who boarded the unfortunate Titanic ship. Use the seaborn library to see if we can find any patterns in data.

2. Write a code to check how the price of the ticket (column name: 'fare') for each passenger is distributed by plotting a histogram.

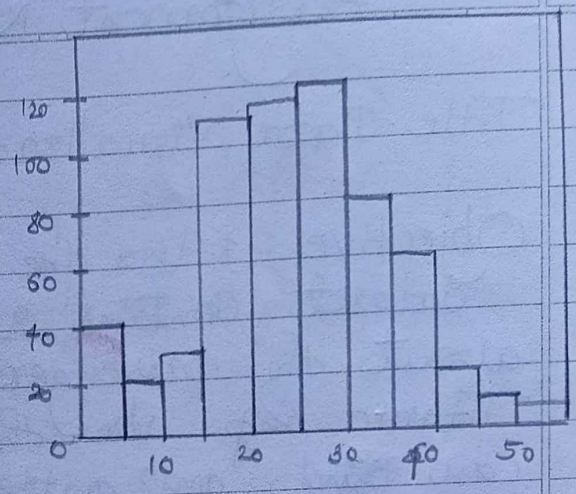
Theory: Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs and maps, data visualization tools provide an accessible way to see and understand trends, outliers and patterns in data. In the world of Big data, and data visualization tools and technologies are essential to analyze massive amounts of information and make data driven decisions

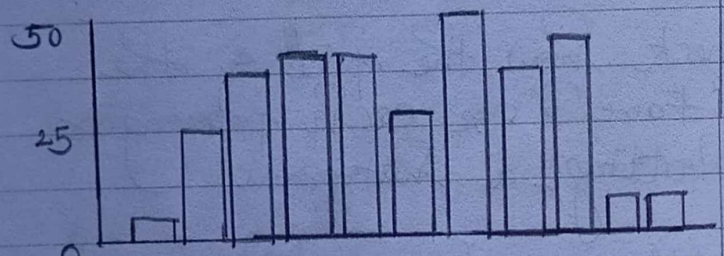
The seaborn library is built on top of Matplotlib and offers many advanced data visualization capabilities. Seaborn, the seaborn library can be used to draw a variety of charts and matrix plots, grid plots, regression plots etc. it can be used to draw distribution and categorical plots as well.

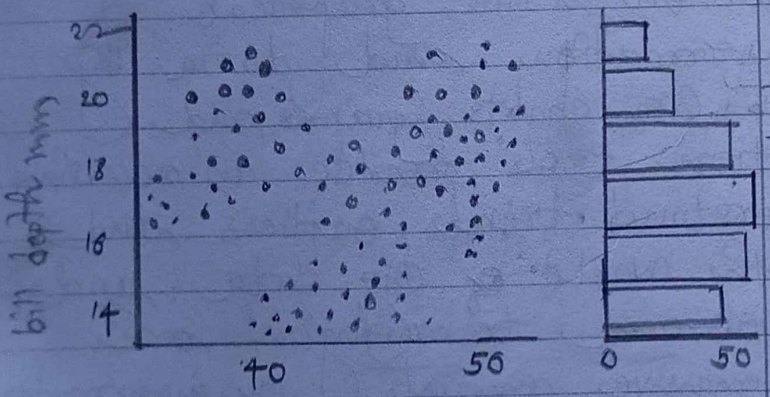
1. Distribution plot

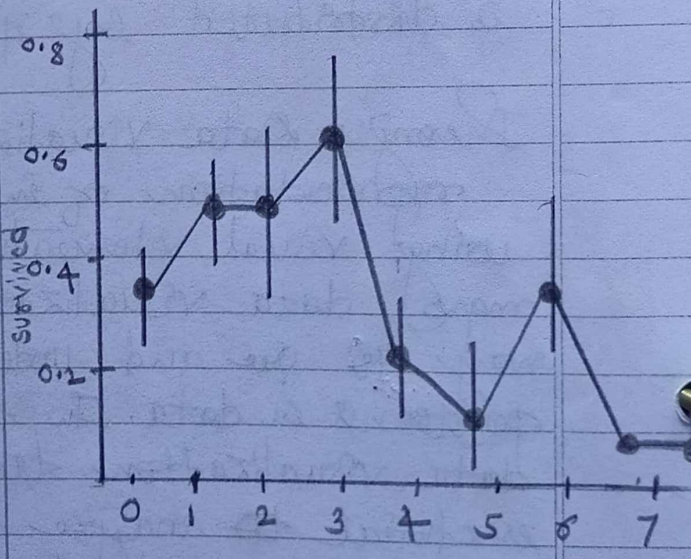


2. Histogram plot









3. Joint plot

4. Pair plot

- Distribution plot

Distribution plots visually asseses the distribution of sample data by comparing the empirical distribution of the data with the theoritical values expected from a specified distribution.

sns.distplot(dataset['fare'])

- Histogram plot

In statistics, a histogram is a graphical representation of the distribution of data. The histogram is represented by a set of rectangles, adjacent to each other, where each bar represent a kind of data.

sns.distplot(dataset['fare'], kde = False, bins = 10)

- Joint plot

Joints plot is the way of understanding the relationship between two variables and the distribution of the individual each variables.

sns.jointplot(x='age', y='fare', data=dataset)

sns.jointplot(x='age', y='fare', data=dataset, kind='hex')

VPKBIET

- pair plot:

Pair plot is used to understand the best set of the features to explain a relationship between two variables or to form the most separated clusters.

Conclusion: Thus we successfully implemented Simple Data Visualization techniques using Python on Titanic Data set.

Reborn
24/3/23

VPKBIET

Assignment No.

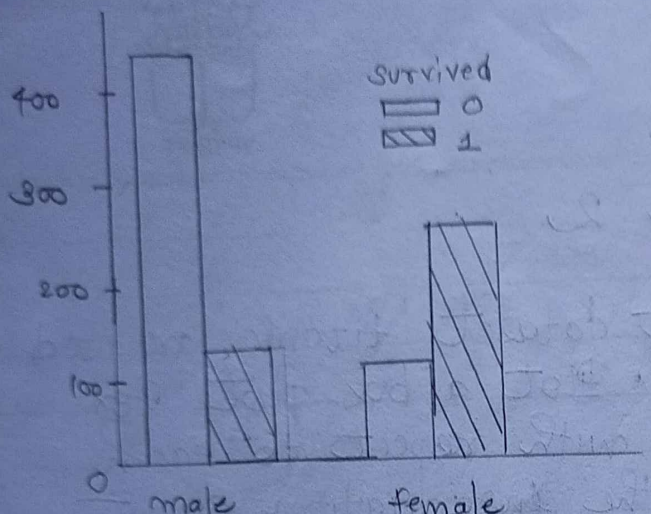
Title: Data Visualization 2

Objective: 1. Use the inbuilt dataset 'titanic' are used in above problem. Plot a box plot for distribution of age with respect to each gender along with the information about whether they survived or not. (Column: 'sex' and 'age')

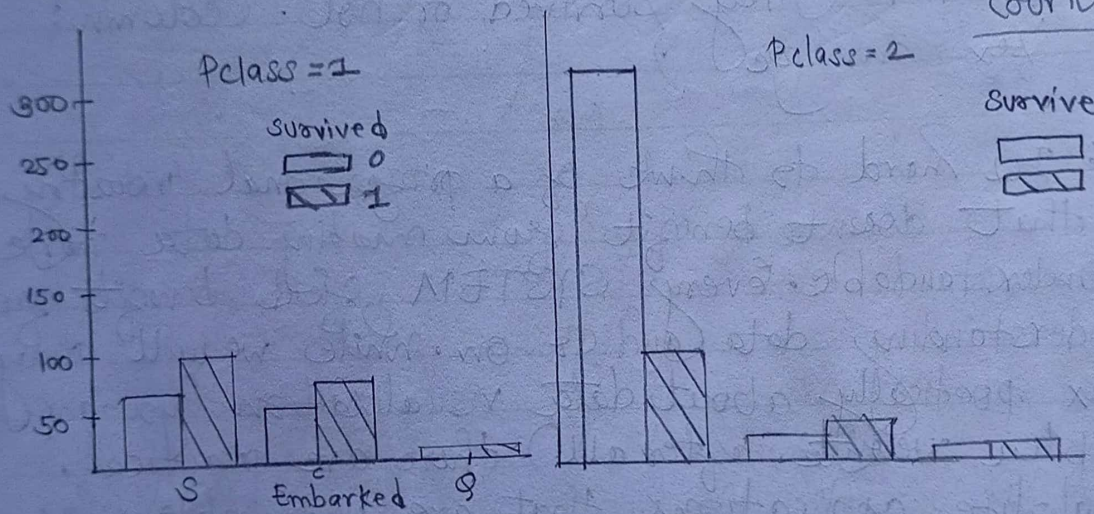
Theory: It is hard to think of a professional industry that doesn't benefit from making data more understandable. Every SYSTEM field benefits from understanding data (and so on. While we will always wax poetically about data visualization (you are in Tableu website, after all) there are practical, real-life applications that are undeniable. Since visualization is so prolific, it's also one of the most powerful useful professional skills to develop. Skill sets are changing to accomodate a data driven world.

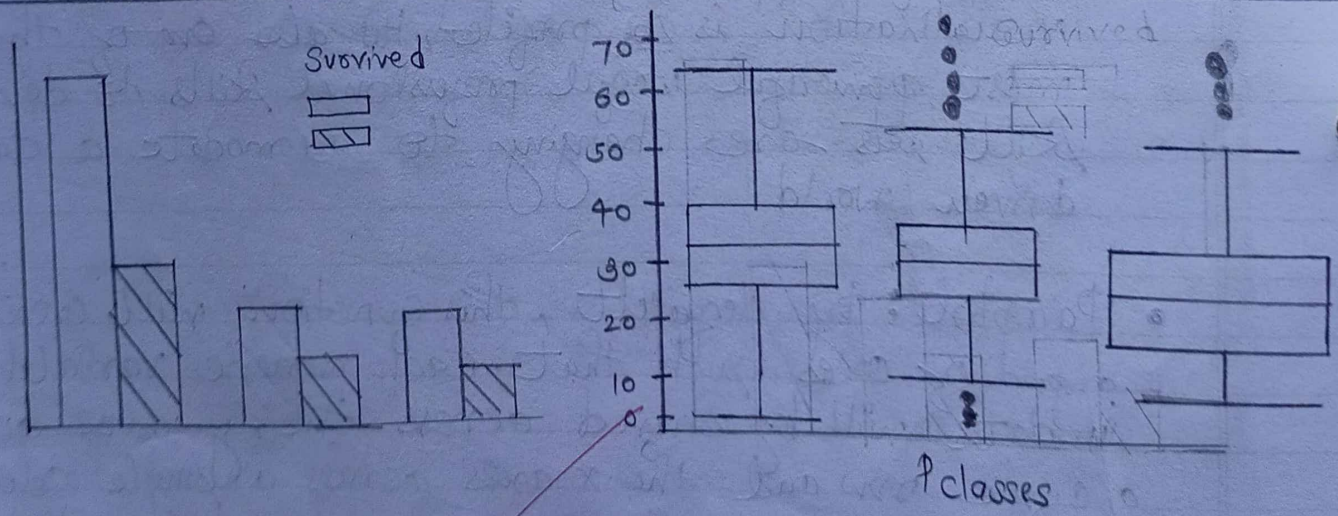
- Pairplot: By defaults, this function will create a grid of axes such that each numeric variable in data will by shared across the y-axes across a single row and the x-axes across a single column.
- Barplot: A barplot (or barchart) is one of the most common type of graphic. It shows the relationship numeric, categoric variable.

Bar plot



Count plot





Box plot

- Count plot: A count plot be thought of as 'its a histogram across a categorical, instead of quantative variable. The basic API and options are identical to those of barplot(), so you can compare counts across nested variables.

```
sns.countplot(x=df["Class"])
```

- Box plot: A box and whisker plot also called box plot - displays the five-number of set of data. The five-number summary is the minimum, first quartile median, third quartile and maximum. In a box plot we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median.

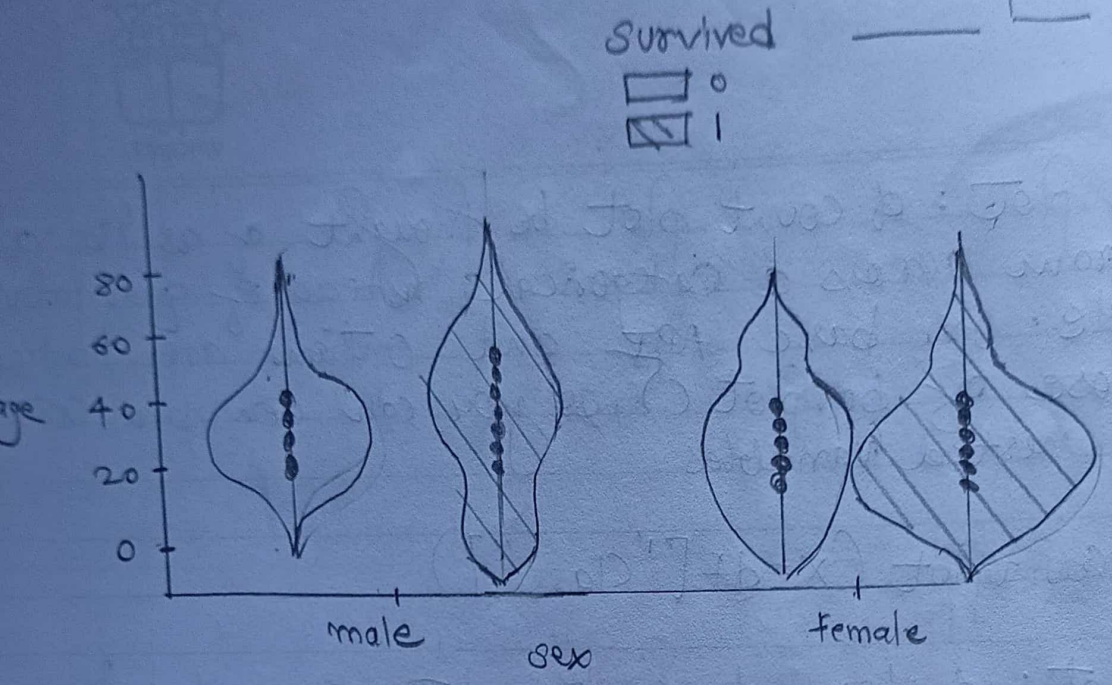
```
sns.boxplot(x=df["age"])
```

- Violin plot: A violin plot is hybrid of box plot and a kernal density plot, which shows peaks in the data. It is used to visualize the distribution of the numerical data.

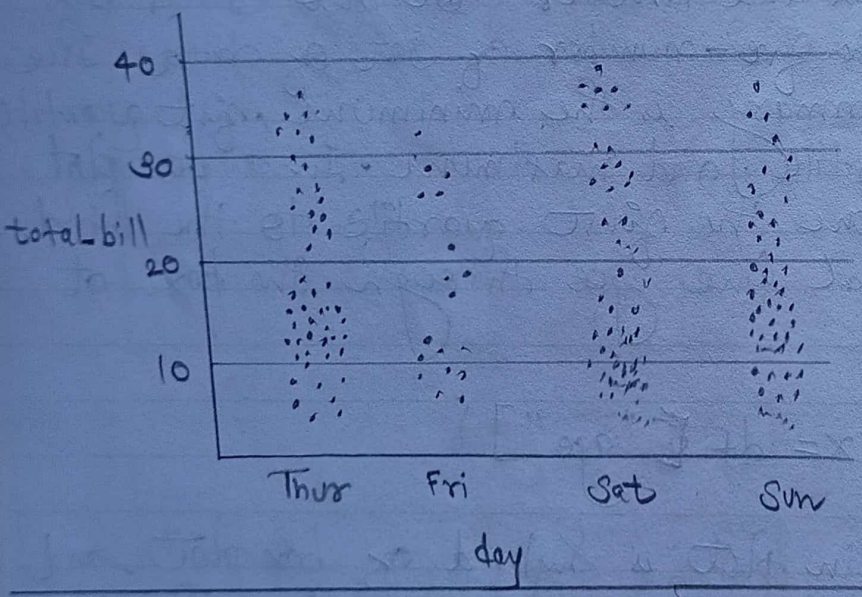
```
sns.violinplot(x=df["age"])
```

- Strip plot: Strip plot can be drawn on its own, but it is the also good complement to a box or violin plot in cases where you want to show all observations along with some representation of the underlying distribution.

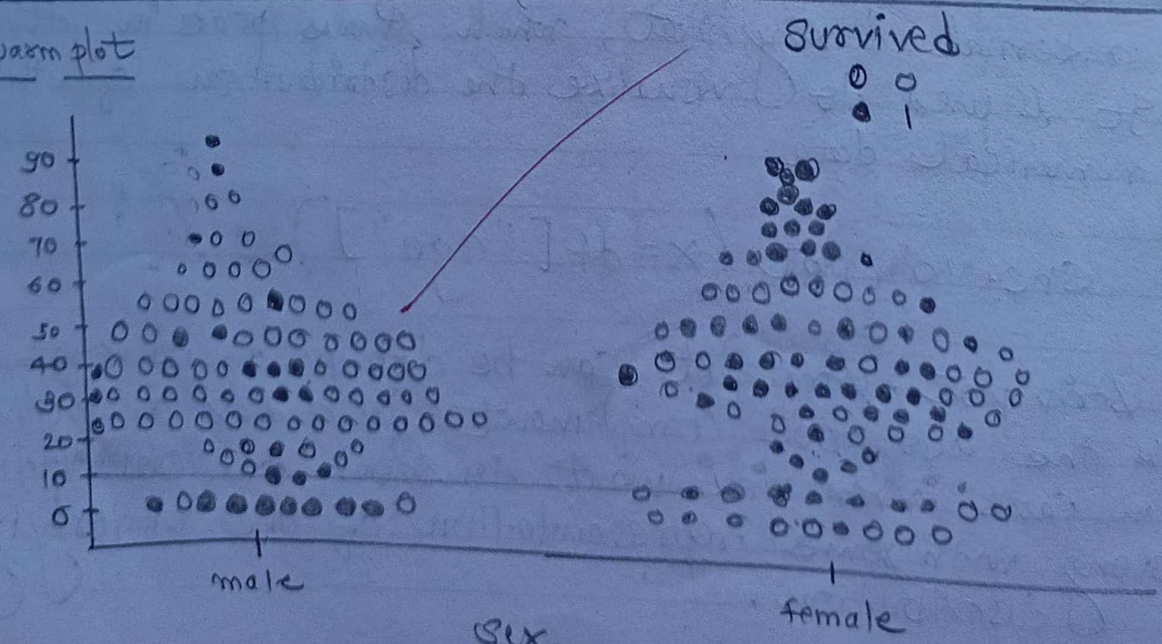
violin plot



strip plot



Swarm plot



VPKBIET

```
sns.stripplot(data = tips, x = "total_bill")
```

- Swarmplot: This function is similar to stripplot() but the points are adjusted (only along the categorical axis) so that they don't overlap. This gives a better representation of the distribution of values, but it doesn't not scale well to large numbers of the distribution of values, but it does not scale well to large numbers of observations. This style of plot is sometimes called a "beeswarm".

Conclusion: Thus we have successfully implemented simple data visualization techniques using Python on Titanic dataset.

Roll no
24/3/23

Assignment 10

Data Visualization III

Download the Iris flower dataset or any other dataset into a DataFrame. (e.g., https://archive.ics.uci.edu/ml/datasets/Iris (https://archive.ics.uci.edu/ml/datasets/Iris)). Scan the dataset and give the inference as:

1. List down the features and their types (e.g., numeric, nominal) available in the dataset.
2. Create a histogram for each feature in the dataset to illustrate the feature distributions.
3. Create a boxplot for each feature in the dataset.
4. Compare distributions and identify outliers

In [1]:

```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
```

In [2]:

```
df=pd.read_csv('IRIS.csv')
```

In [3]:

```
df
```

Out[3]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

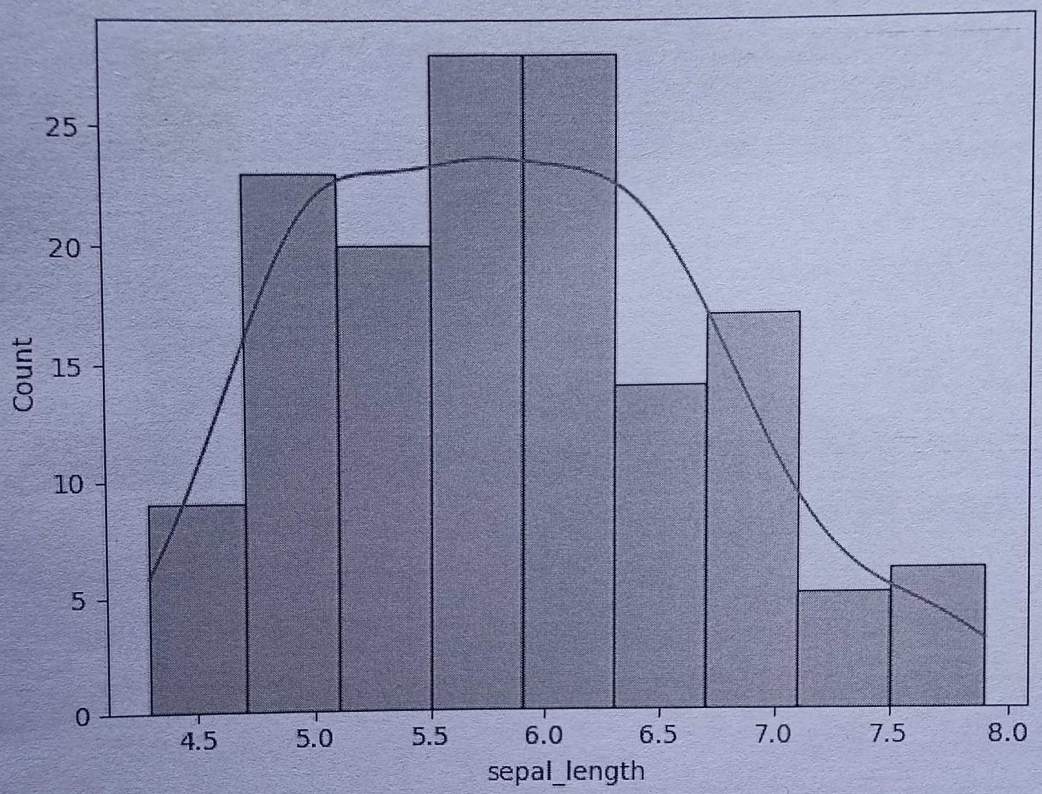
150 rows × 5 columns

In [4]:

```
sns.histplot(df['sepal_length'],kde=True)
```

Out[4]:

```
<AxesSubplot: xlabel='sepal_length', ylabel='Count'>
```

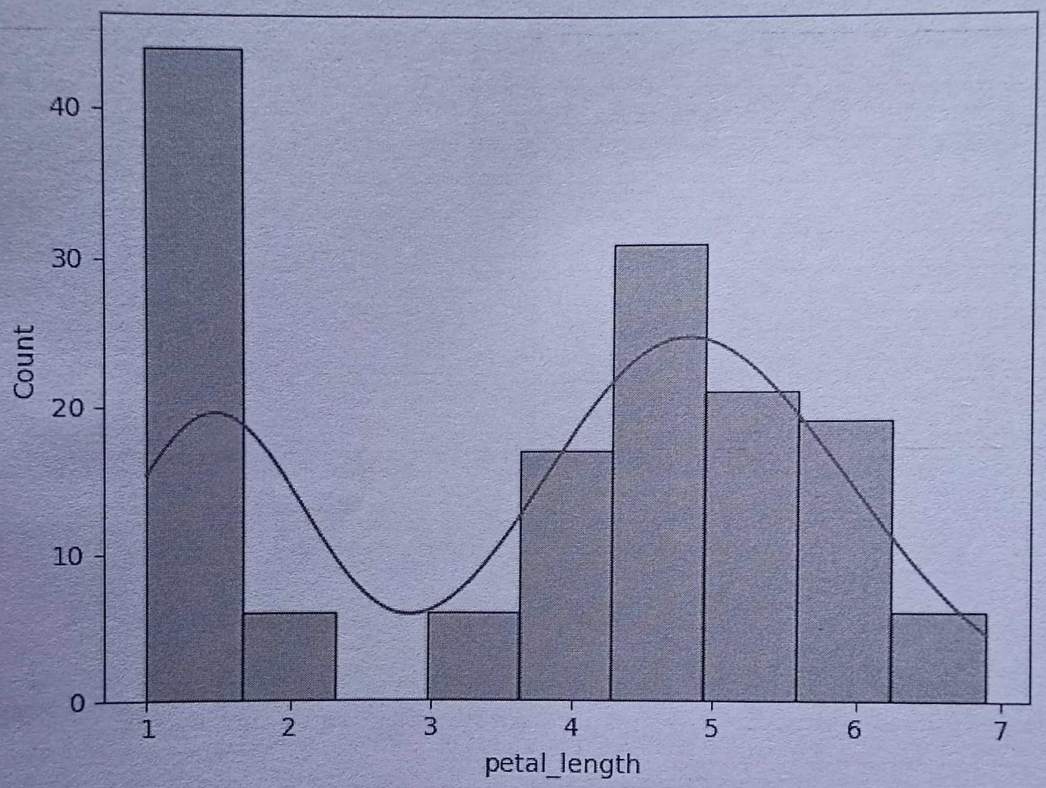


In [5]:

```
sns.histplot(df['petal_length'],kde=True)
```

Out[5]:

```
<AxesSubplot: xlabel='petal_length', ylabel='Count'>
```

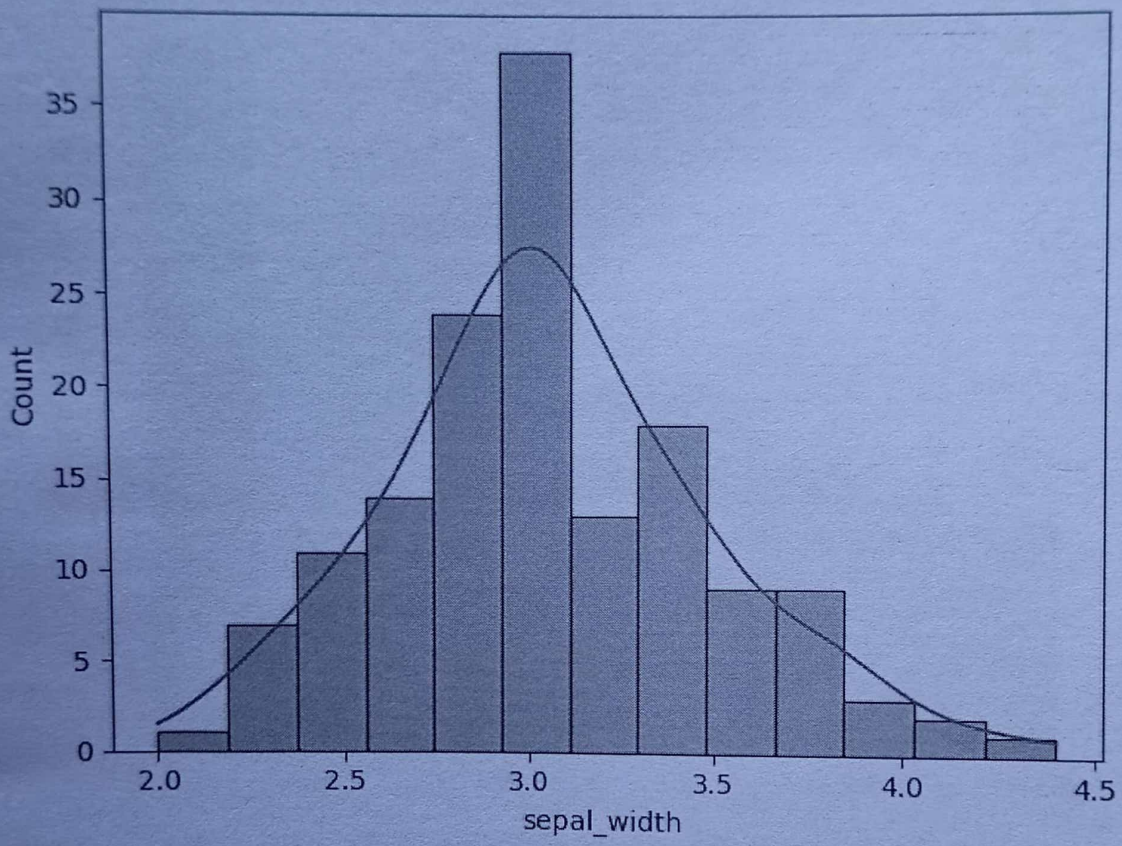


In [6]:

```
sns.histplot(df['sepal_width'],kde=True)
```

Out[6]:

```
<AxesSubplot: xlabel='sepal_width', ylabel='Count'>
```

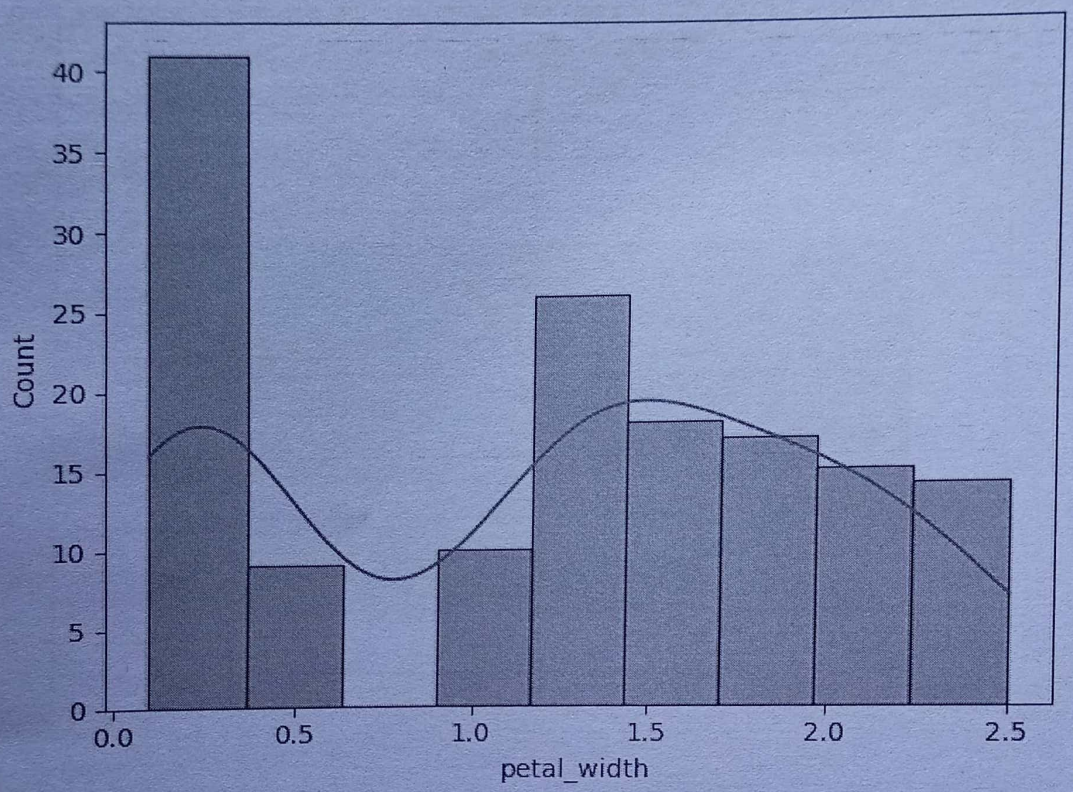


In [7]:

```
sns.histplot(df['petal_width'],kde=True)
```

Out[7]:

```
<AxesSubplot: xlabel='petal_width', ylabel='Count'>
```

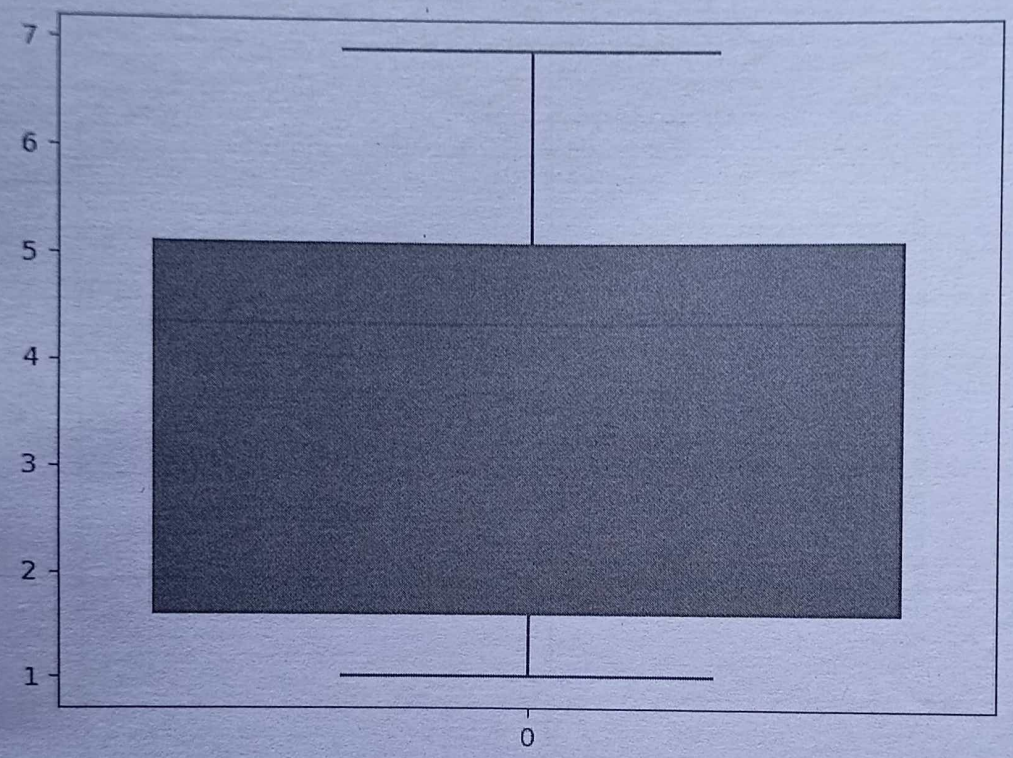


In [8]:

```
sns.boxplot(df['petal_length'],color='Red')
```

Out[8]:

<AxesSubplot: >

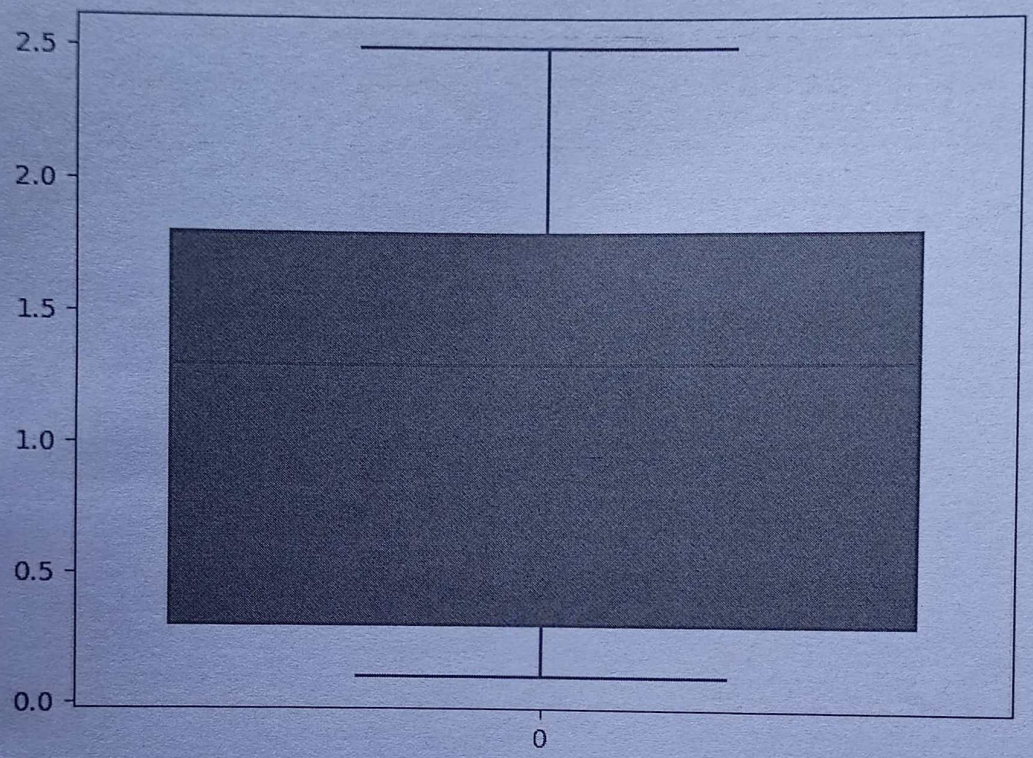


In [9]:

```
sns.boxplot(df['petal_width'],color='Red')
```

Out[9]:

<AxesSubplot: >

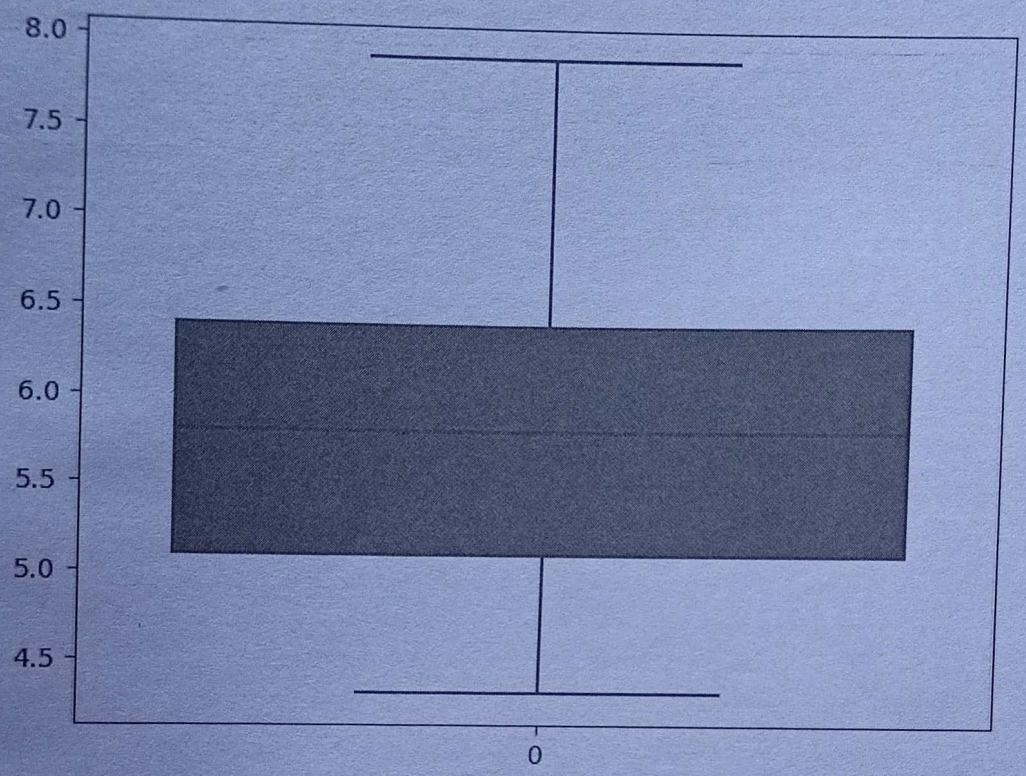


In [10]:

```
sns.boxplot(df['sepal_length'],color='Red')
```

Out[10]:

<AxesSubplot: >

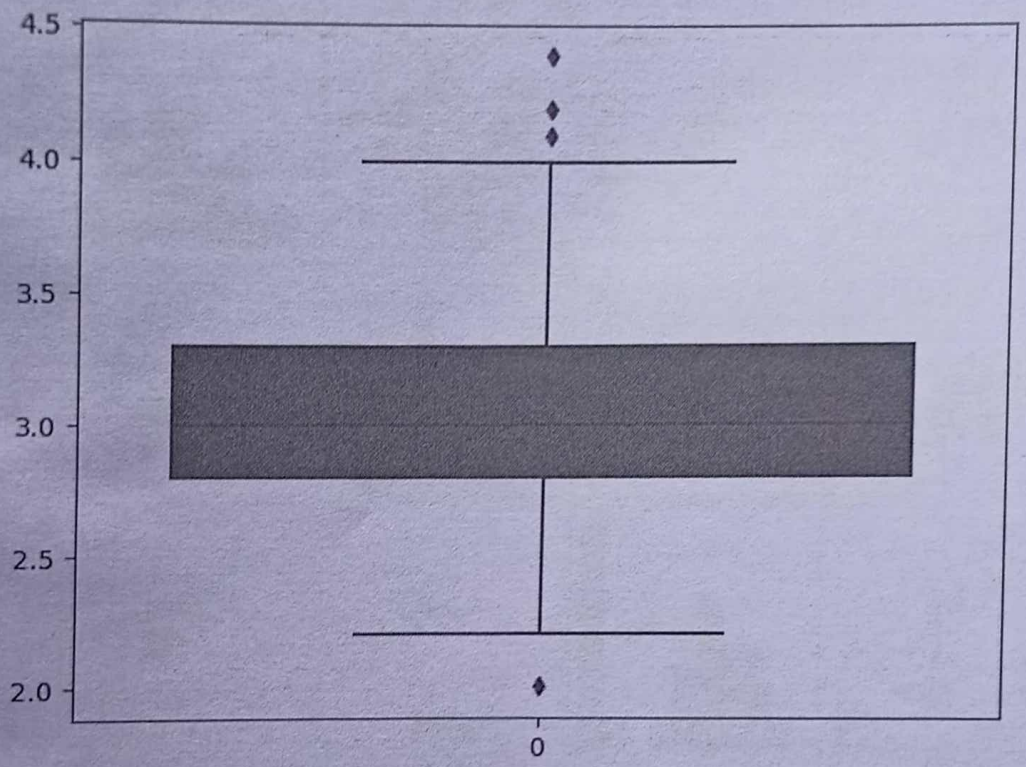


In [11]:

```
sns.boxplot(df['sepal_width'],color='Red')
```

Out[11]:

<AxesSubplot: >

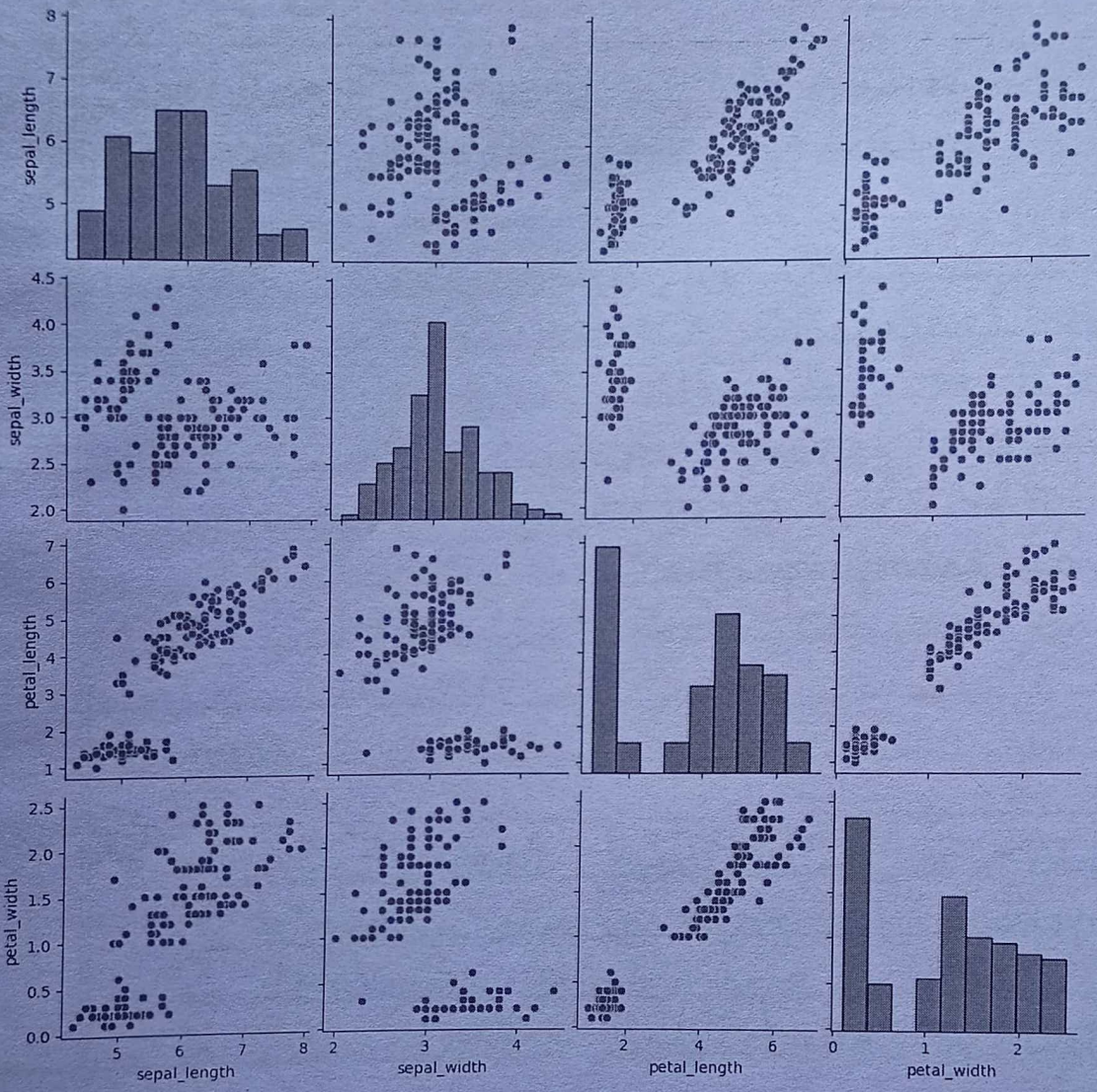


In [12]:

```
sns.pairplot(df)
```

Out[12]:

```
<seaborn.axisgrid.PairGrid at 0x20f245f47d0>
```



VPKBIET

Assignment No. 11

Title: Scala

Objective: To write a simple program is SCALA using Apache Spark framework.

Theory: Scala stands for Scalable language. Scala is general-purpose, high level, multi-paradigm programming language. It is a pure object-oriented programming language which also provides support to the functional programming approach. Scala programs can convert to bytecodes and can run on the JVM.

Features of SCALA:

- Object Oriented: Every value in SCALA is a object so it is purely object-oriented programming language. The behaviour and type of objects are depicted by the classes and traits in SCALA.
- Functional: It is also a functional programming language as every function is a value and every value is an object. It provides the support for the high-order functions, nested functions, anonyms functions etc.
- Statically Typed: The process of verifying and then enforcing the constraints of types is done at compile time in Scala. In most cases, the user has no need to specify a type.

- Extensible: New language constructs can be added to Scala in the form of libraries. Scala is designed to interpolate with the JRE (Java Runtime Environments).

- Concurrent & Synchronize Processing: Scala allows the user to write codes in an immutable manner that makes it easy to apply the parallelism (Synchronize) and concurrency.

Syntax of Scala:

① Objects in scala:

A typical scala program creates many objects, which as you know, interact by invoking methods.

```
var obj = new Dog();
```

② Functions in Scala:

One can divide up the code into separate function

```
def function_name ([parameter_list])
        [: return_type]
{
    // function body
}
```

VPKBIET

Variable in Scala:

Variable are nothing but reserved memory to store values

Var my_var : String : "Foo"

Conditional Statements:

: If statement - Nested if else

: If-else statement - If else-if

```
if (condition):
  {
      // Statement to to executed
  }
if (condition)
  {  // if block statements
  }
   else {
      // else block statements
  }
```

Conclusion: Thus we performed a simple string comparision program in Scale.

Rel 2
15/5