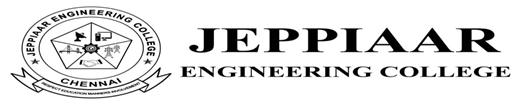
****

## DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

## CS3352- FOUNDATIONS OF DATA SCIENCE

## QUESTION BANK

## II YEAR A&B |BATCH:2022-2026

**Vision of Institution**

To build Jeppiaar Engineering College as an Institution of Academic Excellence in Technical education and Management education and to become a World Class University.

**Mission of Institution**

|  |  |
| --- | --- |
| **M1** | To excel in teaching and **learning, research and innovation** by promoting the principles of scientific analysis and creative thinking |
| **M2** | To participate in the production, **development and dissemination of knowledge** and interact with **national and international communities** |
| **M3** | To equip students with **values, ethics and life skills** needed to enrich their lives and enable them to meaningfully contribute to the **progress of society** |
| **M4** | To prepare students **for higher studies and lifelong learning**, enrich them with the **practical and entrepreneurial skills** necessary to excel as future professionals and contribute to **Nation’s economy** |

***Program Outcomes (POs)***

|  |  |
| --- | --- |
| **PO1** | **Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems. |
| **PO2** | **Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences. |
| **PO3** | **Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations |
| **PO4** | **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions. |
| **PO5** | **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations. |
| **PO6** | **The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice. |
| **PO7** | **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development. |
| **PO8** | **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice. |
| **PO9** | **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings. |
| **PO10** | **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions. |
| **PO11** | **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one’s own work, as a member and leader in a team, to manage projects and in multidisciplinary environments. |
| **PO12** | **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change. |

**Vision of Department**

To emerge as a globally prominent department, developing ethical computer professionals, innovators and entrepreneurs with academic excellence through quality education and research.

**Mission of Department**

|  |  |
| --- | --- |
| **M1** | To create **computer professionals** with an ability to identify and **formulate the engineering problems** and also to provide **innovative solutions** through **effective teaching learning process.** |
| **M2** | To **strengthen the core-competence** in computer science and engineering and to create an ability to **interact** effectively with industries. |
| **M3** | To produce engineers with good professional skills, **ethical values** and life skills for the **betterment of the society.** |
| **M4** | To encourage students towards **continuous and higher level learning** on technological advancements and provide a platform for **employment and self-employment.** |

#### Program Educational Objectives (PEOs)

|  |  |
| --- | --- |
| **PEO1** | **To address the real time complex engineering problems using innovative approach with strong core computing skills.** |
| **PEO2** | **To apply core-analytical knowledge and appropriate techniques and provide solutions to real time challenges of national and global society** |
| **PEO3** | **Apply ethical knowledge for professional excellence and leadership for the betterment of the society.** |
| **PEO4** | **Develop life-long learning skills needed for better employment and entrepreneurship** |

#### Program Specific Outcomes (PSOs)

Students will be able to

|  |  |
| --- | --- |
| **PSO1** | An ability to understand the core concepts of computer science and engineering and to enrich problem solving skills to analyze, design and implement software and hardware based systems of varying complexity. |
| **PSO2** | To interpret real-time problems with analytical skills and to arrive at cost effective and optimal solution using advanced tools and techniques. |
| **PSO3** | An understanding of social awareness and professional ethics with practical proficiency in the broad area of programming concepts by lifelong learning to inculcate employment and entrepreneurship skills. |

COURSE OBJECTIVES:

• To understand the data science fundamentals and process.

• To learn to describe the data for the data science process.

• To learn to describe the relationship between data.

• To utilize the Python libraries for Data Wrangling.

• To present and interpret data using visualization libraries in Python

UNIT I INTRODUCTION

Data Science: Benefits and uses – facets of data - Data Science Process: Overview – Defining research goals – Retrieving data – Data preparation - Exploratory Data analysis – build the model– presenting findings and building applications - Data Mining - Data Warehousing – Basic Statistical descriptions of Data

UNIT II DESCRIBING DATA

Types of Data - Types of Variables -Describing Data with Tables and Graphs –Describing Data with Averages - Describing Variability - Normal Distributions and Standard (z) Scores

UNIT III DESCRIBING RELATIONSHIPS

Correlation –Scatter plots –correlation coefficient for quantitative data –computational formula for correlation coefficient – Regression –regression line –least squares regression line – Standard error of estimate – interpretation of r2 –multiple regression equations –regression towards the mean

UNIT IV PYTHON LIBRARIES FOR DATA WRANGLING

Basics of Numpy arrays –aggregations –computations on arrays –comparisons, masks, boolean logic – fancy indexing – structured arrays – Data manipulation with Pandas – data indexing and selection – operating on data – missing data – Hierarchical indexing – combining datasets – aggregation and grouping – pivot tables

UNIT V DATA VISUALIZATION

Importing Matplotlib – Line plots – Scatter plots – visualizing errors – density and contour plots – Histograms – legends – colors – subplots – text and annotation – customization – three dimensional plotting - Geographic Data with Basemap - Visualization with Seaborn.

COURSE OUTCOMES:

At the end of this course, the students will be able to:

CO1: Define the data science process

CO2: Understand different types of data description for data science process

CO3: Gain knowledge on relationships between data

CO4: Use the Python Libraries for Data Wrangling

CO5: Apply visualization Libraries in Python to interpret and explore data

TOTAL:45 PERIODS

TEXT BOOKS

1. David Cielen, Arno D. B. Meysman, and Mohamed Ali, “Introducing Data Science”, Manning Publications, 2016. (Unit I)

2. Robert S. Witte and John S. Witte, “Statistics”, Eleventh Edition, Wiley Publications, 2017. (Units II and III)

3. Jake VanderPlas, “Python Data Science Handbook”, O’Reilly, 2016. (Units IV and V)

REFERENCES: 1. Allen B. Downey, “Think Stats: Exploratory Data Analysis in Python”, Green Tea Press,2014

# UNIT I PART A

1. What is data science?

Data science is the study of data. It involves developing methods of recording, storing, and analyzing data to effectively extract useful information. The goal of data science is to gain insights and knowledge from any type of data — both structured and unstructured.

1. Explain the role of data science in business?

Data science is the requirement of every business to make business forecasts and predictions based on facts and figures, which are collected in the form of data and processed through data science. Data science is also important for marketing promotions and campaigns.

1. Explain the role of data science in Medical Research?

Data science is necessary for research and analysis in health care, which makes it easier for practitioners in both fields to understand the challenges and extract results through analysis and insights proposed based on data.

The advanced technology of machine learning and artificial intelligence has been infused in the field of medical science to conduct biomedical and genetic researches more easily.

1. Explain the role of data science in health care?

The clinical history of a patient and the medicinal treatment given is computerized and kept as a digitalized record, which makes it easier for the practitioner and doctors to detect the complex diseases at an early stage and understand its complexity.

1. Explain the role of data science in education?

Online personality tests and career counseling suggest students and individuals select a career path based on their choices. This is done with the help of data science, which has designed algorithms and predictive rationalities that conclude with the provided set of choices.

1. Explain the role of data science in social media?

Most of the data of consumer behavior, choices, and preferences are being collected through online platforms, which help the business grow. The systems are smart enough to predict and analyze a

user’s mood, behavior, and opinions towards a specific product, incident, or event with the help of texts, feedbacks, thoughts, views, reactions, and suggestions.

1. Explain the role of data science in Technology?

The new technology emerging around us to ease life and alter our lifestyle is all due to data science. Some applications collect sensory data to track our movements and activities and also provide knowledge and suggestions concerning our health, such as blood pressure and heart rate. Data science is also being implemented to advance security and safety technology, which is the most important issue nowadays.

1. Explain the role of data science in financial institution?

Financial institutions use data science to predict stock markets, determine the risk of lending money, and learn how to attract new clients for their services.

1. Explain the main types/ categories of data?

* Structured
* Unstructured
* Natural language
* Machine-generated
* Graph-based
* Audio, video, and images
* Streaming

1. Explain natural language? Is natural language is instructed data

Yes, Natural language is a special type of unstructured data; it’s challenging to process because it requires knowledge of specific data science techniques and linguistics. The natural language processing is used in entity recognition, topic recognition, summarization, text completion, and sentiment analysis.

Example: Predictive text, Search results.

1. What is Machine-generated data with example?

Machine-generated data is information that’s automatically created by a computer, process, application, or other machine without human intervention. Machine-generated data is becoming a major data resource and will continue to do so.The analysis of machine data relies on highly scalable tools, due to its high volume and speed.

Examples of machine data are web server logs, call detail records, network event logs, and telemetry.

1. What is Graph-based or network data?

Graph or network data is, in short, data that focuses on the relationship or adjacency of objects. The graph structures use nodes, edges, and properties to represent and store graphical data. Graph-based data is a natural way to represent social networks.

1. List the name of the steps involved in data science processing?
   1. SETTING THE RESEARCH GOAL
   2. RETRIEVING DATA
   3. DATA PREPARATION
   4. DATA EXPLORATION
   5. DATA MODELLING
   6. PRESENTATION AND AUTOMATION.
2. What are OUTLIERS?

An outlier is an observation that seems to be distant from other observations or, more specifically, one observation that follows a different logic or generative process than the other observations.

1. What are the different ways of combining data?

The two operations that are performed in various data types to combine them are as follows:

1. Joining- enriching an observation from one table with information from another table.
2. Appending or stacking- adding the observations of one table to those of another table.
3. What is big data?

Big data is a blanket term for any collection of data sets so large or complex that it becomes difficult to process using traditional data management techniques. They are characterized by the four Vs: velocity, variety, volume, and veracity.

1. What is data Cleansing?

Data cleansing is a sub process of the data science process that focuses on removing errors in your data so your data becomes a true and consistent representation of the processes.

1. Define the term setting the research goal (in data science processing).

The term setting the research goal mean Defining what, why, and how of your projection in a project charter.

1. Explain the term Retrieving data.

The term retrieving data means Finding and getting access to data needed in your project. This data is either found within the company or retrieved from a third party.

1. What is data preparation and process?

It is the process of Checking and remediating data errors, enriching the data with data from other data sources, and transforming it into a suitable format for your models.

1. Define data exploration process?

Data exploration process of Diving deeper into your data using descriptive statistics and visual techniques.

1. What is data modelling?

It is the process of Building a model is an iterative process that involves selecting the variables for the model, executing the model, and model diagnostics. It is generally Using machine learning and statistical techniques to achieve our project goal.

1. Explain Presentation and automation process?

Finally, present the results to the business. These results can take many forms, ranging from presentations to research reports. Presenting the results to the stakeholders and industrializing the analysis process for repetitive reuse and integration with other tools.

1. What are the types of database?

* Column databases
* Document stores
* Streaming data
* Key-value stores
* SQL on Hadoop
* New SQL
* Graph databases.

1. What is a column database?

In this method Data is stored in columns, which allows algorithms to perform much faster queries. Newer technologies use cell-wise storage. Table-like structures are still important.

1. Explain document stores?

Document stores no longer use tables, but store every observation in a document. This allows for a much more flexible data scheme.

1. What is Streaming data?

Data is collected, transformed, and aggregated not in batches but in real time. Although we’ve categorized it here as a database to help you in tool selection, it’s more a particular type of problem that drove creation of technologies such as Storm.

1. Define Key-value stores?

When Data isn’t stored in a table; rather you assign a key for every value, such as org.marketing.sales.2015: 20000. This scales well but places almost all the implementation on the developer.

1. Explain SQL on Hadoop?

Batch queries on Hadoop are in a SQL-like language that uses the map-reduce framework in the background

1. What is New SQL?

This class combines the scalability of NoSQL databases with the advantages of relational databases. They all have a SQL interface and a relational data model.

1. Explain Graph databases?

Not every problem is best stored in a table. Particular problems are more naturally translated into graph theory and stored in graph databases. A classic example of this is a social network.

## PART B

1 i)Explain the benefits of data science ii)List the facets of data with example

2. Briefly explain the steps in data science process with diagram

# UNIT II PART A

1. What is Frequency distribution?

Frequency distribution is a collection of observations produced by sorting observations into classes and showing their frequency (f) of occurrence in each class.

1. What are the types and uses of Frequency distributions?

Types of Frequency distributions:

* + Grouped frequency distribution.
  + Ungrouped frequency distribution.
  + Cumulative frequency distribution.
  + Relative frequency distribution.
  + Relative cumulative frequency distribution.

Uses:

A frequency distribution helps us to detect any pattern in the data (assuming a pattern exists) by superimposing some order on the inevitable variability among observations.

1. What is Grouped frequency distribution:

When observations are sorted into classes of more than one value, the result is referred to as a frequency distribution for grouped data.

1. What is ungrouped frequency distribution?

When observations are sorted into classes of single values the result is referred to as a frequency distribution for ungrouped data.

1. What is cumulative frequency distribution?

Cumulative frequency distributions show the total number of observations in each class and in all lower-ranked classes. This type of distribution can be used effectively with sets of scores, such as test scores for intellectual or academic aptitude, when relative standing within the distribution assumes primary importance.

1. What is relative frequency distribution?

Relative frequency distributions show the frequency of each class as a part or fraction of the total frequency for the entire distribution.

1. Define Percentile Ranks?

The percentile rank of a score indicates the percentage of scores in the entire distribution with similar or smaller values than that score.

1. What is Histograms?

A histogram is the most commonly used graph to show frequency distributions. It is a bar-type graph for quantitative data. The common boundaries between adjacent bars emphasize the continuity of the data, as with continuous variables.

1. Explain any three Features of histograms?

Equal units along the horizontal axis (the X axis, or abscissa) reflect the various class intervals of the frequency distribution. Equal units along the vertical axis (the Y axis, or ordinate) reflect increases in frequency. (The units along the vertical axis do not have to be the same width as those along the horizontal axis.). The body of the histogram consists of a series of bars whose heights reflect the frequencies for the various classes.

1. What is Frequency Polygon?

Frequency Polygon is a line graph for quantitative data that also emphasizes the continuity of continuous variables. Frequency polygons may be constructed directly from frequency distributions.

1. What is mean?

The mean is found by adding all scores and then dividing by the number of scores. MEAN = SUM OF ALL SCORES /NUMBER OF SCORES

1. What is median?

The median reflects the middle value when observations are ordered from least to most. The median splits a set of ordered observations into two equal parts, the upper and lower halves. In other words, the median has a percentile rank of 50, since observations with equal or smaller values constitute 50 percent of the entire distribution.

1. What is mode?

Mode reflects the value of the most frequently occurring score.It is easy to assign a value to the mode if the data are organized.

1. What if a distribution have More than one mode or no mode at all?

Distributions can have more than one mode (or no mode at all). Distributions with two obvious peaks, even though they are not exactly the same height, are referred to as bimodal. Distributions with more than two peaks are referred to as multimodal. The presence of more than one mode might reflect important differences among subsets of data.

1. Explain Range, variance and standard deviation?

**Range**: The range is the difference between the largest and smallest scores

**Variance**: The variance is a measure of variability

**Standard deviation**: Standard deviation is a measure of the amount of variations or dispersion of a set of value. A low standard deviation indicates that the values tend to be close to the mean of the set, while high standard deviation indicates that the values are spread out over a wider range.

𝑠𝑡𝑎𝑛𝑑𝑎𝑟𝑑 𝑑𝑒𝑣𝑖𝑎𝑡𝑖𝑜𝑛 = √variance

1. What is DEGREES OF FREEDOM (df)?

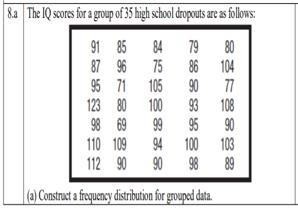
Degrees of freedom (df) refers to the number of values that are free to vary, given one or more mathematical restrictions, in a sample being used to estimate a population characteristic.

1. What is INTERQUARTILE RANGE (IQR)?

Interquartile range (IQR), is simply the range for the middle 50 percent of the scores. More specifically, the IQR equals the distance between the third quartile (or 75th percentile) and the first quartile.

## PART B

1. Explain the different types of frequency distribution with example.
2. Describe Mean, Median, Mode and Averages with example.



1. Specify the real limits for the lowest class interval in this frequency distribution
2. Analyze how graph can be used to represent qualitative and quantitative data?

# UNIT III

**PART A**

1. What is NumPy in Python used for?

NumPy is a Python library used for working with arrays.

1. Write a python program create an array?

Import numpy as np np.array([1,4,2,5,3])

**OUTPUT**: array([1,4,2,5,3])

1. Write the output of the following numpy code:

i. np.array([3.14, 4, 2, 3])

* 1. np.array([1, 2, 3, 4], dtype='float32')
  2. np.array([range(i, i + 3) for i in [2, 4, 6]])
  3. np.zeros(10, dtype=int)
  4. np.ones((3, 5), dtype=float) vi. np.full((3, 5), 3.14)

1. np.arange(0, 20, 2)
2. np.linspace(0, 1, 5)
3. np.random.random((3, 3))
4. np.random.normal(0, 1, (3, 3)) OUTPUT:

i.array([3.14, 4. , 2. , 3. ])

1. array([1., 2., 3., 4.], dtype=float32)
2. array([[2, 3, 4],

[4, 5, 6],

[6, 7, 8]])

iv.array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0])

v. array([[1., 1., 1., 1., 1.],

[1., 1., 1., 1., 1.],

[1., 1., 1., 1., 1.]])

vi. array([[3.14, 3.14, 3.14, 3.14, 3.14],

[3.14, 3.14, 3.14, 3.14, 3.14],

[3.14, 3.14, 3.14, 3.14, 3.14]])

vii. array([ 0, 2, 4, 6, 8, 10, 12, 14, 16, 18])

viii. array([0. , 0.25, 0.5 , 0.75, 1. ])

ix. array([[0.15175932, 0.58606546, 0.68749167],

[0.75655189, 0.23198831, 0.94250833],

[0.95147981, 0.50405388, 0.22004745]])

x. array([[-0.98220551, -0.27991827, -1.58428463],

[-0.99791504, 0.10710667, -1.15115236],

[ 0.76783606, -0.83683471, -0.07508393]])

1. Define Series Object.

A Pandas Series is a one-dimensional array of indexed data. It can be created from a list or array. Example:

data = pd.Series([0.25, 0.5, 0.75, 1.0]) print(data )

The output will be displayed as

|  |  |
| --- | --- |
| 0 | 0.25 |
| 1 | 0.50 |
| 2 | 0.75 |
| 3 | 1.00 |

dtype: float64

1. What is Data frame?

A DataFrame is an analog of a two-dimensional array with both flexible row indices and flexible column names. DataFrame as a sequence of aligned Series objects.

Example:

states = pd.DataFrame({'population': population, 'area': area}) print(states)

Output:

|  |  |  |
| --- | --- | --- |
|  | area | population |
| California | 423967 | 38332521 |
| Florida | 170312 | 19552860 |
| Illinois | 149995 | 12882135 |
| New York | 141297 | 19651127 |
| Texas | 695662 | 26448193 |

1. How a pandas dataframe can be constructed?
2. From a single Series object
3. From a list of dicts
4. From a dictionary of Series objects
5. From a two-dimensional NumPy array
6. From a NumPy structured array
7. What are indexers?

Pandas provides some special indexer attributes that explicitly expose indexing schemes. They are loc, iloc, and ix.

loc attribute - allows indexing and slicing that always references the explicit index.

iloc attribute - allows indexing and slicing that always references the implicit Python-style index. Ix - is a hybrid of the two, and for Series objects is equivalent to standard []-based indexing.

1. How missing data can be handled in python?

**None**, a Python singleton object that is often used for missing data in Python code. Example:

import numpy as np import pandas as pd

val = np.array([1, None, 3, 4]) (val)

Output: array([1, None, 3, 4], dtype=object)

The other missing data representation, NaN ( Not a Number), is a special floating-point value recognized by all systems that use the standard IEEE floating-point representation:

Example

val = np.array([1, np.nan, 3, 4]) print( val.dtype)

Output: dtype('float64')

1. How the operations can be performed on null values in pandas data structure?

There are several useful methods for detecting, removing, and replacing null values in Pandas data structures.

They are:

isnull() - Generate a Boolean mask indicating missing values notnull() - Opposite of isnull()

dropna() - Return a filtered version of the data

fillna() - Return a copy of the data with missing values filled or imputed

1. Define Hierarchical Indexing.

Hierarchical indexing also known as multi-indexing is used to incorporate multiple index levels within a single index. In this way, higher-dimensional data can be compactly represented within the familiar one-dimensional Series and two-dimensional DataFrame objects.

1. What is pivot table?

The pivot table takes simple column-wise data as input, and groups the entries into a two- dimensional table that provides a multidimensional summarization of the data.

## PART B

1.Briefly explain the basics of numpy arrays with example 2.Describe about fancy indexing with example

1. Explain structured data in numpy array.
2. What is universal function? Explain clearly each function with example.
3. Explain aggregate functions with example
4. What is broadcasting and explain the rules with examples 7.Explain data objects in pandas.
5. Briefly explain the hierarchical indexing with examples
6. What is pivot table? Explain it clearly

## UNIT IV PART A

1. Define Normal curve and its properties.

The normal curve is a theoretical curve defined for a continuous variable, and noted for its symmetrical bell-shaped form.

Properties of Normal curve:

The normal curve is symmetrical, its lower half is the mirror image of its upper half. Being bell shaped, the normal curve peaks above a point midway along the horizontal spread and then tapers off gradually in either direction from the peak(without actually touching the horizontal axis, since, in theory, the tails of a normal curve extend infinitely far)

1. What is Z-score?

A z score is a unit-free, standardized score that, regardless of the original units of measurement, indicates how many standard deviations a score is above or below the mean of its distribution.

where X is the original score and μ and σ are the mean and the standard deviation, respectively.

1. What is correlation?

Correlation is a statistical measure that expresses the extent to which two variables are linearly related. It’s a common tool for describing simple relationships without making a statement about cause and effect.

The sample correlation coefficient, r, quantifies the strength of the relationship. Correlations are also tested for statistical significance.

1. Define Scatterplots?

A scatterplot is a graph containing a cluster of dots that represents all pairs of scores.With a little training, you can use any dot cluster as a preview of a fully measured relationship.

1. What is correlation coefficient?

A correlation coefficient is a number between –1 and 1 that describes the relationship between pairs of variables. The type of correlation coefficient, designated as r, that describes the linear relationship between pairs of variables form quantitative data.

1. Define Regression.

A predictive modeling technique that evaluates the relation between dependent (i.e. the target variable) and independent variables is known as regression analysis. Regression analysis can be used for forecasting, time series modeling, or finding the relation between the variables and predict continuous values.

1. Write the types of Regression analysis. Types of Regression Analysis:
   1. Simple Linear Regression
   2. Multiple Linear Regression
   3. Polynomial Regression
   4. Logistic Regression
   5. Ridge Regression
   6. Lasso Regression
   7. Bayesian Linear Regression
   8. Decision Tree Regression
   9. Random Forest Regression
2. Diffrentiate single and multiple linear regression? Simple Linear Regression:

Linear regression is the most basic form of regression algorithms in machine learning. The model consists of a single parameter and a dependent variable has a linear relationship. We denote simple linear regression by the following equation given below

y = mx + c + e

where m is the slope of the line, c is an intercept, and e represents the error in the model Multiple Linear Regression:

When the number of independent variables increases, it is called the multiple linear regression models.

y = b0 + b1x1

1. What is ridge regression?

Ridge Regression is another type of regression in machine learning and is usually used when there is a high correlation between the parameters. This is because as the correlation increases the least square estimates give unbiased values.

1. What is decision tree?

The decision tree as the name suggests works on the principle of conditions. It is efficient and has strong algorithms used for predictive analysis. It has mainly attributed that include internal nodes, branches, and a terminal node. Every internal node holds a “test” on an attribute, branches hold the conclusion of the test and every leaf node means the class label. It is used for both classifications as well as regression which are both supervised learning algorithms.

## PART B

1. Explain briefly above Normal curve.
2. How correlation coefficient can be calculated for the quantitative data? 3.Explain the different type of regression analysis in detail.

4.Briefly explain in detail standard error of estimate.

## UNIT V PART A

1. What is the purpose of matplotlib?

Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy.

One of Matplotlib’s most important features is its ability to play well with many operating systems and graphics backends.

1. Write the dual interface of matplotlib?

The dual interfaces of matplotlib are: a convenient MATLAB-style state-based interface, and a more powerful object-oriented interface.

1. How to draw a simple line plot using matplotlib? import matplotlib.pyplot as plt plt.style.use('seaborn-whitegrid')

import numpy as np fig = plt.figure()

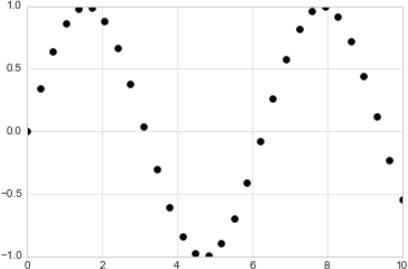
ax = plt.axes()

x = np.linspace(0, 10, 1000) ax.plot(x, np.sin(x))

1. What functions can be used to draw the scatterplot? plt.plot are the functions used to draw the scatter plots. Example:

x = np.linspace(0, 10, 30) y = np.sin(x)

plt.plot(x, y, 'o', color='black'); Output:



A second, more powerful method of creating scatter plots is the plt.scatter function, which can be used very similarly to the plt.plot function.

Example:

plt.scatter(x, y, marker='o')

1. Write the difference between plot and scatter functions?

The primary difference of plt.scatter from plt.plot is that it can be used to create scatter plots where the properties of each individual point (size, face color, edge color, etc.) can be individually controlled or mapped to data.

1. Define contour plot?

Contour plot used to plot three dimensional data into two dimensional data. A contour plot can be created with the plt.contour function. It takes three arguments: a grid of x values, a grid of y values, and a grid of z values.

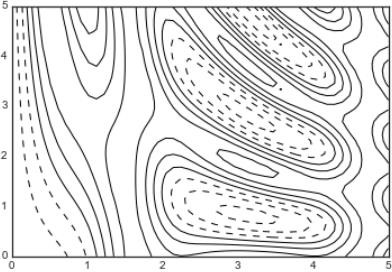
1. What are the functions can be used to draw the contour plots?

plt.contour, plt.contourf, and plt.imshow are the functions used to draw the contour plots. Example:

x = np.linspace(0, 5, 50)

y = np.linspace(0, 5, 40) X, Y = np.meshgrid(x, y) Z = f(X, Y)

plt.contour(X, Y, Z, colors='black') Output:



1. What is the purpose of using histogram?

A simple histogram is very useful in understanding a dataset. It is used to specify the frequency distributions between two varibales.

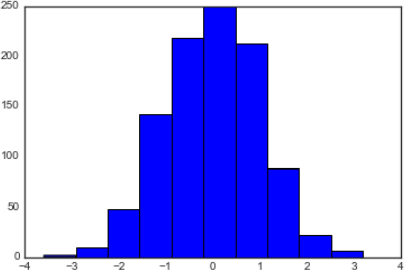
Example:

plt.style.use('seaborn-white') data = np.random.randn(1000) plt.hist(data)

1. Write the source code to draw a simple histogram? import numpy as np

import matplotlib.pyplot as plt plt.style.use('seaborn-white') data = np.random.randn(1000) plt.hist(data)

Output:



1. How to create a three-dimensional wireframe plot?

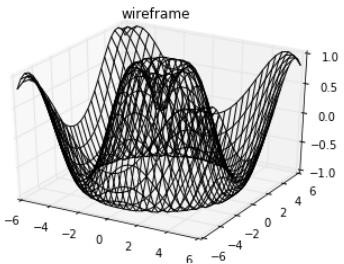
The three-dimensional plots that work on gridded data are wireframes. It take a grid of values and project it onto the specified three dimensional surface, and can make the resulting three-dimensional forms quite easy to visualize.

Example:

fig = plt.figure()

ax = plt.axes(projection='3d') ax.plot\_wireframe(X, Y, Z, color='black') ax.set\_title('wireframe')

Output:



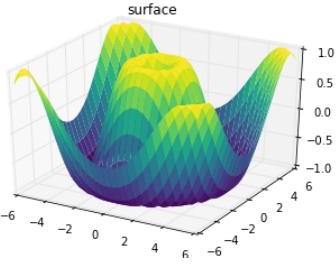
1. Define surface plot?

A surface plot is like a wireframe plot, but each face of the wireframe is a filled polygon. Adding a colormap to the filled polygons can aid perception of the topology of the surface being visualized.

Example:

ax = plt.axes(projection='3d') ax.plot\_surface(X, Y, Z, rstride=1, cstride=1, cmap='viridis', edgecolor='none') ax.set\_title('surface')

Output:



1. What is the use of seaborn?

Seaborn is a open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis. Seaborn works easily with dataframes and the Pandas library. The graphs created can also be customized easily.

Below are a few benefits of Data Visualization:

Graphs can help us find data trends that are useful in any machine learning or forecasting project. Graphs make it easier to explain your data to non-technical people.

Visually attractive graphs can make presentations and reports much more appealing to the reader.

## PART B

1.What is matplotlib? Specify the two interfaces used by it. 2.Briefly explain about the line plot and scatter plot.

1. Explain contour plot and histogram.
2. What is 3D plotting? Explain it with example.
3. How graphical data can be projected using matplotlib? Explain with example.