CS3353
FOUNDATIONS OF DATA SCIENCE

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.

2. 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.

3. 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.

4. 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.

5.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.

6. 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.

7. 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.

8. 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.

9. Explain the main types/ categories of data?

* Structured

* Unstructured

* Natural language

* Machine-generated

* Graph-based

* Audio, video, and images

* Streaming

10. 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.

11. 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.

12. 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.

13. 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.

14. 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.

15. What are the different ways of combining data?

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

i. Joining- enriching an observation from one table with information from another table.
ii. Appending or stacking- adding the observations of one table to those of another table.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. Define data exploration process?

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

22. 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.

23. 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.

24. What are the types of database?

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

26. 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.

27. 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.

28. 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.

29. 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.

30. Explain SQL on Hadoop?

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

31. 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.

32. 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
3. Briefly explain the architecture of Data Mining
4. Briefly explain the architecture of Data warehouse
5. Explain briefly the need for Data Science with real time application
6.How do you set the research goal, retrieving data and Data preparation process in Data Science process
7. How do you set the research goal, retrieving data and Data preparation process in Data Science process
8. . Explain the Data exploration ,data modeling and presentation process in Data Science
9. Explain briefly above Normal curve.

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.

2. 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.

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

9. 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.

10. 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.

11. 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

12. 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.

13. 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.

14. 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.

15. 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.

$$standard\ deviation = \sqrt{\text{variance}}$$

16. 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.

17. 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.

18. 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)

19. 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.

PART B

1. Explain the different types of frequency distribution with example.

2. Describe Mean, Median, Mode and Averages with example.

8.a The IQ scores for a group of 35 high school dropouts are as follows:

91	85	84	79	80
87	96	75	86	104
95	71	105	90	77
123	80	100	93	108
98	69	99	95	90
110	109	94	100	103
112	90	90	98	89

(a) Construct a frequency distribution for grouped data.

3. Specify the real limits for the lowest class interval in this frequency distribution

4. Analyze how graph can be used to represent qualitative and quantitative data?

5. Generate the ungrouped and grouped frequency table for the following data
90,92,87,88,87,92,98,90,90,87,87,88,88,89,90,87,89,92,92,92,98,90,95,87,87

(i) How many people scored 98?
(ii) How many people scored 90 or less?
(iii) What proportion scored 87?

6. (i) Calculate the sum of square population standard deviation for the given x data value
13,10,11,7,9,11,9
(ii) Calculate the sample standard deviation for the given data 7,3,1,0,4

7. Suppose the IQ score have a bell shaped distribution with a mean of 100 and standard deviation of 15 then calculate the following,

(i) What percentage of people should have an IQ score between 85 and 115 .
(ii) What percentage of people should have an IQ score between 70 and 130
(iii) What percentage of people should have an IQ score more than 130
(iv) A person with an IQ score greater than 145 is considered as genius. Does the empirical rule support this statement

UNIT III

PART A

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.

2. 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.

3. 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.

4. 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.

5. 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

6. 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.

$$y = b0 + b1x1$$

7. 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.

8. 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. How correlation coefficient can be calculated for the quantitative data?

2. Explain the different type of regression analysis in detail.

3. Briefly explain in detail standard error of estimate.

4. Calculate the value of r using computation formula for the following data.

FRIENDS	SENT	RECEIVED
Dories	13	14
Steve	9	18
Mike	7	12
Andrea	5	10
John	1	6

5. Explain about regression towards mean in detail.

UNIT IV

PART A

1. What is NumPy in Python used for?

NumPy is a Python library used for working with arrays.

2. 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])

3. Write the output of the following numpy code:

i. np.array([3.14, 4, 2, 3])
ii. np.array([1, 2, 3, 4], dtype='float32')
iii. np.array([range(i, i + 3) for i in [2, 4, 6]])
iv. np.zeros(10, dtype=int)
v. np.ones((3, 5), dtype=float)
vi. np.full((3, 5), 3.14)
vii. np.arange(0, 20, 2)
viii. np.linspace(0, 1, 5)
ix. np.random.random((3, 3))
x. np.random.normal(0, 1, (3, 3))

OUTPUT:

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

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

iii. 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]])

4. 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
```

5. 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

6. How a pandas dataframe can be constructed?

i) From a single Series object

ii) From a list of dicts

iii) From a dictionary of Series objects

iv) From a two-dimensional NumPy array

v) 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.

8. 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')

9. 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

10. 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.

11. 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

3.Explain structured data in numpy array.

4.What is universal function? Explain clearly each function with example.

5. Explain aggregate functions with example

6.What is broadcasting and explain the rules with examples

7.Explain data objects in pandas.

8. Briefly explain the hierarchical indexing with examples

9. What is pivot table? Explain it clearly

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.

2. 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.

3. 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))
```

4. What functions can be used to draw the scatterplot?

plt.plot are the functions used to draw the scatter plots.

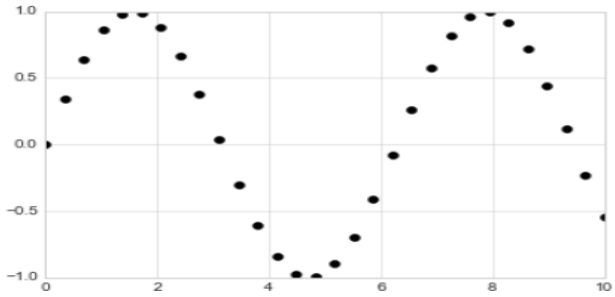
Example:

```
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')

5. 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.

6. 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.

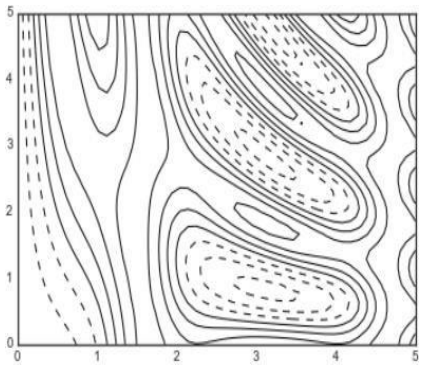
7. What are the functions can be used to draw the contour plots?

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:



8. 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.

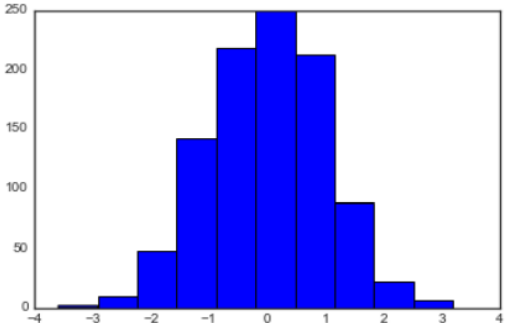
Example:

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

9. 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:



10. 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:

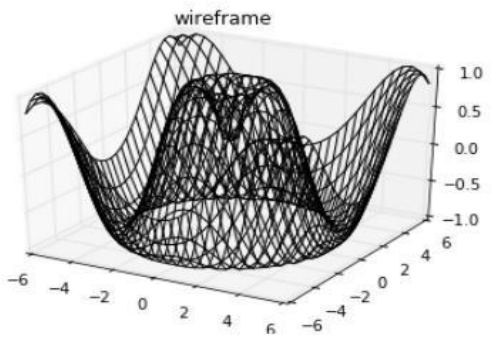
```
fig = plt.figure()

ax = plt.axes(projection='3d')

ax.plot_wireframe(X, Y, Z, color='black')

ax.set_title('wireframe')
```

Output:



11. 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:

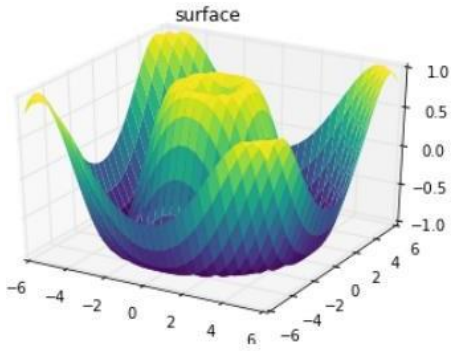
```
ax = plt.axes(projection='3d')

ax.plot_surface(X, Y, Z, rstride=1, cstride=1,

cmap='viridis', edgecolor='none')

ax.set_title('surface')
```

Output:



12. 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.

3. Explain contour plot and histogram.

4. What is 3D plotting? Explain it with example.

5. How graphical data can be projected using matplotlib? Explain with example.