



**DIPLOMA IN ARTIFICIAL INTELLIGENCE
AND MACHINE LEARNING**

CENTRALIZED QUESTION BANK

**1056234440 - DATA EXPLORATION AND
VISUALIZATION**

**DIRECTORATE OF TECHNICAL
EDUCATION GOVERNMENT OF
TAMILNADU**

DIPLOMA END SEMESTER /YEAR EXAMINATION-2025

Course: Artificial Intelligence and Machine Learning

Subject: Data exploration and Visualization

QP Code: 1056234440

Time:3 Hours

Date:

Session:

Max Marks:100

Answer the Following Questions

1. A) Working with R Variables and basic data structures - Vector, matrix, data frame
 - (a) Assume two vectors (2.1, 3.4, 2.5, 2.7, 2.9) and (0.3, 0.5, 0.6, 0.9, 1.1) are length and diameters of cylinders. Calculate the correlation between lengths and diameters.
 - (b) Assume the measurements are in centimeters. Recalculate the volumes so that their units are in cubic millimeters. Calculate the mean, standard deviation, and coefficient of variation of these new volumes.
 - (c) Construct a matrix with 10 columns and 10 rows, all filled with random numbers between 0 and 1.
 - (d) Calculate the row means of the above matrix. Also calculate the standard deviation across the row mean.
- B) OnCerealsDatasetperform
 - (a) How many cereal sin the data frame are 'hot' cereals?
 - (b) Take a subset of the data frame with only the Manufacturer 'K'
 - (c) Take a subset of the data frame of all cereals that have less than 80 calories, AND have more than 20 units of vitamins.
 - (d) Take a subset of the data frame containing cereals that contain atleast1 unit of sugar, and keep only the variables 'Cereal.name', 'calories' and 'vitamins'.
 - (e) For one of the above subsets ,write a new CSV file to disk
2. A) Working with data structure sin Pandas–data frame ,series
 - (a) Import the data set chit pole.
 - (b) Find the number of observations in the data set.
 - (c) Print the name of all the columns.
 - (d) How many items were ordered in total?
 - (e) Convert item price into a float.
 - (f) What was the total revenue generated during the period of the data set?
 - (g) How many orders were made in the period?

B) Use Titanic Data set and find

- (a) How many observations of 'Age' are missing from the data frame?
 - (b) For what proportion of the passengers is the age unknown? Was this proportion higher for 3rd class than 1st and 2nd?
 - (c) Count the number of passengers in each class (1st, 2nd, 3rd).
- 3. A) Working with data structures in Numpy - arrays
 - (a) Create a 1D array with a set of integer values.
 - (b) Create a 1D, 2D and 3D Boolean array.
 - (c) Extract all odd numbers from the created 1D array and replace them with -1.
 - (d) Reshape a 1D array to a 2D array with 2 rows and convert an array of arrays into a flat 1D array.
 - (e) Count the unique values in a numpy array.
 - (f) Create a new column from existing columns of a numpy array.
 - (g) Create a 2D array with a set of values and compute the row wise counts of all possible values in an array.
 - (h) Get the positions of top n values from a numpy array.
- B) Use the Hydro data set
 - (a) Make a line plot of storage versus Date
 - (b) Make the line thicker, and a dot-dashed style
 - (c) Make the same plot with points and change the color of the points in the following way: green if storage is over 500, orange if storage is between 235 and 500, and red if storage is below 235.
- 4. A) Import Titanic data set into python/R environment
 - (a) Take the last 15 names of passengers and sort them alphabetically
 - (b) What was the name of the oldest surviving male?
 - (c) Make a new variable called 'Status', based on the 'Survived' variable already in the data set. For passengers that did not survive, Status should be 'dead', for those who did, Status should be 'alive'.
- B) Use coweet a data set
 - (a) Make a scatter plot of biomass versus height, with the symbol color varying by species.
 - (b) Log-transform biomass, and redraw the plot.
- 5. A) Import Titanic dataset into python/R environment and perform descriptive analysis.
- B) Using data from titanic dataset

- (a) Plot the age distribution
 - (b) Break the above distribution based on survival and plot this new distribution
 - (c) Plot the distribution of age based on survival and sex features
 - (d) Using bar plot visualize the percentage of survivors against sex grouped by class
- 6. A) Load Cereals data set into python /R environment
 - (a) Print first 10 observations from the data set and inspect the data types of the features.
 - (b) Add a new variable to the dataset called 'total carb', which is the sum of 'carbo' and 'sugars'.
 - (c) How many unique manufacturers are included in the data set?
 - (d) Rename the column 'Manufacturer' to 'Producer'.
- B) On Cereals Data set perform
 - (a) How many cereals in the data frame are 'hot' cereals?
 - (b) Take a subset of the data frame with only the Manufacturer 'K'
 - (c) Take a subset of the data frame of all cereals that have less than 80 calories, AND have more than 20 units of vitamins.
 - (d) Take a subset of the data frame containing cereas that contain atleast1 unit of sugar, and keep only the variables 'Cereal.name', 'calories' and 'vitamins'.
 - (e) For one of the above sub sets, write a new CSV file to disk
- 7 A) Read Hydro dataset into python/R environment and perform the following analysis.
 - (a) Change the first variable to a Date class. Are the successive measurements in the dataset always exactly one week apart?
 - (b) How many weeks was the dam level equal to or lower than the value of 235 Gwh?
- B) UseTitanicDatasetandfind
 - (a) How many observations of 'Age' are missing from the data frame?
 - (b) For what proportion of the passengers is the age known? Was this proportion higher for 3rd class than 1st and 2nd?
 - (c) Count the number of passenger sine ach class(1st,2nd,3rd).
- 8 A) Working with R Variables and basic data structures - Vector, matrix, data frame
 - (a) Assume two vectors (2.1, 3.4, 2.5, 2.7, 2.9) and (0.3, 0.5, 0.6, 0.9, 1.1) are length and diameters of cylinders. Calculate the correlation between lengths and diameters.

- (b) Assume the measurements are in centimeters. Recalculate the volumes so that their units are in cubic millimeters. Calculate the mean, standard deviation, and coefficient of variation of these new volumes.
- (c) Construct a matrix with 10 columns and 10 rows, all filled with random numbers between 0 and 1.
- (d) Calculate the row means of the above matrix. Also calculate the standard deviation across the row mean.

B) Using data from titanic data set

- (a) Plot the age distribution
- (b) Break the above distribution based on survival and plot this new distribution
- (c) Plot the distribution of age based on survival and sex features
- (d) Using bar plot visualize the percentage of survivors against sex grouped by class

9 A) Working with data structures in Pandas—data frame, series

- (a) Import the data set chit pole.
- (b) Find the number of observation sin the data set.
- (c) Print the name of all the columns.
- (d) How many items were ordered in total?
- (e) Convert item price into a float.
- (f) What was the total revenue generated during the period of the data set?
- (g) How many orders were made in the period?

B) Use the Hydro data set

- (a) Make a line plot of storage versus Date
- (b) Make the line thicker, and a dot-dashed style
- (c) Make the same plot with points and change the color of the point sin the following way: green if storage is over 500, orange if storage is between 235 and 500, and red if storage is below 235.

10 A) Working with data structures in Numpy-arrays

- (a) Create an 1D array with a set of integer values.
- (b) Create an 1D, 2D and 3D Boolean array.
- (c) Extract all odd numbers from the created 1D array and replace them with -1.
- (d) Reshape a 1D array to a 2D array with 2 rows and convert an array of arrays into a flat 1d array.

- (e) Count the unique values in an numpy array.
 - (f) Create a new column from existing columns of numpy array.
 - (g) Create a 2D array with a set of values and compute the row wise counts of all possible values in an array.
 - (h) Get the positions of top n values from numpy array.
- B) Use coweeet a data set
- (a) Make a scatter plot of biomass versus height, with the symbol color varying by species.
 - (b) Log-transform biomass ,and red raw the plot.
- 11 A) Import Titanic data set in to python / R environment
- (a) Take 15 random names of passengers sort them alphabetically
 - (b) What was the name of the oldest surviving male?
 - (c) Make a new variable called 'Status', based on the 'Survived' variable already in the data set. For passengers that did not survive, Status should be 'dead', for those who did, Status should be 'alive'.
- B) Use Titanic Data set and find
- (a) How many observations of 'Age' are missing from the data frame?
 - (b) For what proportion of the passengers is the age unknown? Was this proportion higher for 3rd class than 1st and 2nd?
 - (c) Count the number of passengers in each class (1st, 2nd, 3rd).
- 12 A) Import Titanic dataset into python/R environment and perform descriptive analysis.
- B) On Cereals Data set perform
- (a) How many cereals in the data frame are 'hot' cereals?
 - (b) Take a sub set of the data frame with only the Manufacturer 'K'
 - (c) Take a subset of the data frame of all cereals that have less than 80 calories, AND have more than 20 units of vitamins.
 - (d) Take a subset of the data frame containing cereals that contain atleast 1 unit of sugar, and keep only the variables 'Cereal.name', 'calories' and 'vitamins'.
 - (e) For one of the above sub sets, write a new CSV file to disk
- 13 A) Load Cereals data set in to python/R environment
- (a) Print first 10 observations from the dataset and inspect the data types of the features.
 - (b) Add a new variable to the dataset called 'total carb', which is the sum of 'carbo' and 'sugars'.
 - (c) How many unique manufacturers are included in the dataset?
 - (d) Rename the column 'Manufacturer' to 'Producer'.

- B) Use the Hydro data set
- (a) Make a line plot of storage versus Date
 - (b) Make the line thicker, and a dot-dashed style
 - (c) Make the same plot with points and change the color of the points in the following way: green if storage is over 500, orange if storage is between 235 and 500, and red if storage is below 235.
- 14 A) Read Hydro data set into python / R environment and perform the following analysis.
- (a) Change the first variable to a Date class .Are the successive measurements in the dataset always exactly one week apart?
 - (b) How many weeks was the dam level equal to or lower than the value of 235Gwh?
- B) Using data from titanic data set
- (a) Plot the age distribution
 - (b) Break the above distribution based on survival and plot this new distribution
 - (c) Plot the distribution of age based on survival and sex features
- Using bar plot visualize the percentage of survivors against sex grouped by class
- 15 A) Working with R Variables and basic data structures - Vector, matrix, data frame
- (a) Assume two vectors (2.1, 3.4, 2.5, 2.7, 2.9) and (0.3, 0.5, 0.6, 0.9, 1.1) are length and diameters of cylinders. Calculate the correlation between lengths and diameters.
 - (b) Assume the measurements are in centimeters. Recalculate the volumes so thattheirunits are incubicmillimeters.Calculate the mean,standard deviation, and coefficient of variation of these new volumes.
 - (c) Construct a matrix with 10 columns and 10 rows, all filled with random numbers between 0 and 1.
 - (d) Calculate the row means of the above matrix .Also calculate the standard deviation across the row mean.
- B) Use coweet a dataset
- (a) Make a scatter plot of biomass versus height, with the symbol color varying by species.
 - (b) Log-transform bio mass ,and red raw the plot.
- 16 A) Working with data structures in Pandas–data frame, series
- (a) Import the data set chit pole.
 - (b) Find the number of observations in the data set.
 - (c) Print the name of all the columns.

- (d) How many items were ordered in total?
 - (e) Convert item price into a float.
 - (f) What was the total revenue generated during the period of the data set?
 - (g) How many orders were made in the period?
- B) Using data from titanic data set
- (a) Plot the age distribution
 - (b) Break the above distribution based on survival and plot this new distribution
 - (c) Plot the distribution of age based on survival and sex features
 - (d) Using bar plot visualize the percentage of survivors against sex grouped by class
- 17 A) Working with data structures in Numpy-arrays
- (a) Create an 1D array with a set of integer values.
 - (b) Create an 1D, 2D and 3D Boolean array.
 - (c) Extract all odd numbers from the created 1D array and replace them with -1.
 - (d) Reshape a 1D array to a 2D array with 2 rows and convert an array of arrays into a flat 1d array.
 - (e) Count the unique values in anumpy array.
 - (f) Create a new column from existing columns of anumpy array.
 - (g) create a 2D array with a set of values and compute the row wise count of all possible values in an array.
 - (h) Get the positions of top n values from anumpy array.
- B) On Cereals Data set perform
- (a) How many cereals in the data frame are 'hot' cereals?
 - (b) Take a sub set of the data frame with only the Manufacturer 'K'
 - (c) Take a subset of the data frame of all cereals that have less than 80 calories, AND have more than 20 units of vitamins.
 - (d) Take a subset of the data frame containing cereals that contain at least 1 unit of sugar, and keep only the variables 'Cereal.name', 'calories' and 'vitamins'.
 - (e) For one of the above subsets, write a new CSV file to disk
- 18 A) Import Titanic data set into python/ R environment
- (a) Take 15 random names of passengers sort them alphabetically
 - (b) What was the name of the oldest surviving male?
 - (c) Make a new variable called 'Status', based on the 'Survived' variable already in the data set .For passengers that did not survive, Status should be

'dead', for those who did, Status should be 'alive'.

- B) Use the Hydro data set
- (a) Make a line plot of storage versus Date
 - (b) Make the line thicker, and a dot-dashed style
 - (c) Make the same plot with points and change the color of the points in the following way: green if storage is over 500, orange if storage is between 235 and 500, and red if storage is below 235.
- 19 A) Import Titanic dataset into python/R environment and perform descriptive analysis.
- B) Use coweeet a data set
- (a) Make a scatter plot of biomass versus height, with the symbol color varying by species.
 - (b) Log-transform biomass, and redraw the plot.
- 20 A) Read Hydro data set into python / R environment and perform the following analysis.
- (a) Change the first variable to a Date class. Are the successive measurements in the dataset always exactly one week apart?
 - (b) How many weeks was the dam level equal to or lower than the value of 235Gwh?
- B) Using data from titanic data set
- (a) Plot the age distribution
 - (b) Break the above distribution based on survival and plot this new distribution
 - (c) Plot the distribution of age based on survival and sex features
 - (d) Using bar plot visualize the percentage of survivors against sex grouped by class

Allocation Of Marks

S. No	Description	Marks
1	Aim(05) Program from Part A (30)	35
2	Aim(05) Program from Part B (30)	35
3	Executing any one program (Part A or Part B)	15
4	Output	10
5	VivaVoce	5
TOTAL		100

