DIPLOMA IN INFORMATION TECHNOLOGY

CENTRALIZED QUESTION BANK

4052653 - Data Science and Big Data Practical

DIRECTORATE OF TECHNICAL

EDUCATIONGOVERNMENT OF

TAMILNADU

DIPLOMA END SEMESTER / YEAR EXAMINATION – 2023

Course : Information Technology

Subject : Data Science and Big Data Practical **QP Code :** 4052653

Time : **3 Hours **Date : **Session:** **Max Marks:**100

ANSWER ALL THE QUESTION

1. Load the data about the exam fee paid by the students of all branches of your college. Perform the following operation sonit using Excel.

   a. Arrange the data branch wise with in the branch and arrange register numbers. Replace all names with CAPITAL.

   b. Count the number of students in each branch and semester

   c. Calculate the total fee paid by students of each branch.

   d. Find the minimum and the maximum fee paid by the student.

   e. Find the sum, average, max, min of fee paid in each branch

2. Load the data collected from all students during online answer paper submission with the following details for each exam. Regno, name, course_ code, subject _ code, semester, number_ of_ pages (nop), mode _of _ dispatch, email_ id, mobile_ number.
   Perform the following operations using Excel.

   a. Check the file for any missing data in the columns.

   b. Count the number of students appeared for the exam .

   c. Count the number of papers (subjects) submitted by each student (Using register number).

   d. Create a new column by concatenating register number and the subject code. Using this column, perform the v lookup function to find the number of pages (nop) written by the students in that subject, and the mode of dispatch.

   e. Count the number of students appeared (submitted) for each subject.

   f. Count the number of different (unique) subject _codes that have been submitted.

3. Read the data set from the Auto-MP G repository and perform the descriptive Statistics on the data using Excel-Data Analysis. Verify the same using the statistical functions of Excel.

4. Read the data set from the Auto-MP G repository and

   a. Identify the relationship between the variables using correlation..
   b. Identify the in dependent and the dependent variables.
   c. Perform the linear regression on the related variables and find the.
   d. regression equation.
   e. Estimate the performance of the regression model.

5. Load any external csv data file and store it in a P and as Data Frame.

   a. Check the shape and column types of the Data Frame(rows and columns).[Note: Usedf.info()and df. shape()].
   b. Subset the data column by names, by index, by range.
   c. Subset data base don index label, row index, multiple rows.
   d. Subset base don rows and columns.

6. DESCRIPTIVESTATISTICS using Python-Pandas

   a. Write a Python script to find basic descriptive statistics on AUTO-MPG dataset.
   b. Find the values of the descriptive statistics.
   c. Determine the measures of a central location, such as mean, markers such as Quartiles or percentiles, and measures of variability or spread, such as the standard deviation.

7. READING AND WRITING DIFFERENT TYPES OF DATASETS

   a. Reading different types of data sets (.txt, .csv) from Web and disk and writing in file in specific disk location.
   b. Reading Excel data sheet using Pandas.
   c. Export the values from the Data Frame to several other formats.

8. DATAVISUALIZATION

   a. Load the Auto-MPG dataset from csv file into pandas.
   b. Analyze the Behavior of the Number of Cylinder sand Horse power Using a Box plot
   c. Find the relationship between horse power and weight using the scatter plot using the data from Auto _MPG:
   d. Find the out liers using plot.
   e. Plot the histogram, bar chart and pie chart on sample data

9. COVARIANCE and CORRELATION

   a. Find the correlation and covariance between two variables.
   b. Plot the correlation plot on the data set and visualize giving an over view of relationships a mong data.
   c. Fit a simple linear regression model using libraries such as NumpyorScikit-learn. (Import Linear Regression from sk learn. linear _ model)

- Import the packages and classes you need.
- Provide data for independent and dependent variables.

- Create a regression model and fitit with existing data.
  Check the results of model fitting to know whether the model is satisfactory.

10. OUTLIER Detection

When analyzing data collected as part of a science experiment it may be desirable to remove the most extreme values before performing other calculations. Write a function that takes a list of values and an non-negative integer, n, as its parameters.

The function should create a new copy of the list with the n largest elements and the n smallest elements removed. Then it should return the new copy of the list as the function‘s only result. The order of the elements in the returned list does not have to match the order of the elements in the original list.

11. Text Processing

b. Open a text file and read all the lines of the file.
c. Token is e(separate the words) the text.
d. Count the total number of lines, total number of word sand unique words
e. Sort the words alphabetically.
f. Find the most frequent and least frequent words.
g. List the words having certain suffixes.
   Note: You can open aTamiltextfileusing'UTF-16' encoding.

12. Text Processing-II

Load atextfilecontaining a list of words into a Data Frame. Apply the following functions and verify the results.
Replace(), repeat(), count(pattern), starts with(pattern), ends with(pattern),find(pattern),find all(pattern).

DETAILLEDALLOCATIONOFMARKS

Writing answer for any one program from the list	45Marks
Executing the program	35Marks
Result with printout of the Program	10Marks
Demonstration of Mini Project	5 Marks
VIVA–VOCE	5 Marks
TOTAL	100Marks