



DIPLOMA IN COMPUTER ENGINEERING

CENTRALIZED QUESTION BANK

4052653 - Data Science and Big Data Practical

**DIRECTORATE OF TECHNICAL
EDUCATION GOVERNMENT OF
TAMILNADU**

DIPLOMA END SEMESTER / YEAR EXAMINATION – 2023

Course : Computer Engineering

Subject : Data Science and Big Data Practical

QP Code : 4052653

Time : 3 Hours **Date** :

Session:

Max Marks:100

ANSWER THE FOLLOWING QUESTIONS

1. Load the data about the exam fee paid by the students of all branches of your college. Perform the following operations on it using Excel.
 - a) Arrange the data branch wise within 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.
Reg no, 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 vlookup 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 dataset from the Auto-MPG 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 dataset from the Auto-MPG repository and
 - a) Identify the relationship between the variables using correlation.
 - b) Identify the independent and the dependent variables.
 - c) Perform the linear regression on the related variables and find the regression equation.
 - d) Estimate the performance of the regression model.

5. Load any external csv data file and store it in a Pandas Data Frame.
 - a) Check the shape and column types of the Data Frame (rows and columns).
[Note: Use df.info () and df shape()]
 - b) Subset the data column by names, by index, by range.
 - c) Subset data based on index label, row index, multiple rows.
 - d) Subset based on rows and columns

6. DESCRIPTIVE STATISTICS 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. DATA VISUALIZATION
 - a) Load the Auto-MPG dataset from csv file into pandas.
 - b) Analyze the Behavior of the Number of Cylinders and Horsepower Using a Box plot
 - c) Find the relationship between horsepower and weight using the scatter plot using the data from Auto-MPG:
 - d) Find the outliers 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 dataset and visualize giving an overview of relationships among data.
- C. Fit a simple linear regression model using libraries such as Numpy or Scikit-learn.
(import Linear Regression from sklearn. linear_model)
 - Import the packages and classes you need.
 - Provide data for independent and dependent variables.
 - Create a regression model and fit it with existing data.
 - Check the results of model fitting to know whether the model is satisfactory.

10. OUTLIER Detection

When analysing 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 a 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.