B.E / B.Tech. PRACTICAL END SEMESTER EXAMINATIONS, NOVEMBER/DECEMBER 2022

Third Semester

AD3301 - DATA EXPLORATION AND VISUALIZATION

(Regulations 2021)

Time : 3 Hours Answer any one Question Max. Marks 100

Aim/Principle/Apparatus required/Procedure	Tabulation/Circuit/ Program/Drawing	Calculation & Results	Viva-Voce	Record	Total
20	30	30	10	10	100

1. Data preprocessing methods on student and labor datasets Implement data cube for data ware house on 3-dimensional data.

2. For any dataset (eg: migration of the country) that contains immigration details and do the following i).Create an area plot for top 6 immigrant countries ii. Create and year-wise immigrant bar chart. iii .Create a boxplot. iv. Show the total no. of immigrant's countries using AreaChart and Pie chart. v. Create a scatter Histogram for the immigrants for a particular year.

3. Perform EDA with an Email dataset and perform the following

   i)Importing a pandas data frame

   ii)Use any one plots

4. For the given data set that contains the data of flights that were on time in January for the years 2019 and 2020. Using the two data sets visualize the data using matplot and plotly libraries to depict the following: Show the difference in statistics for distance for both the years using the appropriate plotting technique. Visualize the no. of flights whose destination airport id is 11778 and 11267using a bar plot or bar chart. Create a Sunburst Plot for both the years depicting the difference among them.

5. Import the panda data frames with any dataset.

6. a. Write an R script to find basic descriptive statistics using summary

   b. Write an R script to find subset of dataset by using subset ()

7. Perform data analysis and representation on a map using Map dataset with Mouse Rollover effect.

8. a. Find the data distributions using box and scatter plot.

   b. Find the outliers using plot.

   c. Plot the histogram, bar chart and pie chart on sample data

9. The following training examples map descriptions of individuals

   onto high, medium and low credit-worthiness.

   medium skiing design single twenties no ->highRisk

   high golf trading married forties yes ->lowRisk

   ow speedway transport married thirties yes ->medRisk

   medium football banking single thirties yes ->lowRisk

   high flying media married fifties yes ->highRisk

   ow football security single twenties no ->medRisk

   medium golf media single thirties yes ->medRisk

   medium golf transport married forties yes ->lowRisk

   high skiing banking single thirties yes ->highRisk

   ow golf unemployed married forties yes ->highRisk

   Input attributes are (from left to right) income, recreation, job, status, agegroup, home-owner. Find the unconditional probability of `golf' and the conditional probability of `single' given `medRisk' in the dataset?

10. Perform EDA and apply to investigate the data and summarize the key insights.

11. Perform EDA on Iris data set and perform any on visualization techniques and perform an analysis report.

12. Perform the following R-operation for the Air Quality dataset

    i)Mean

    ii)Head

    iii)Summary

13. With any data set modify Data of a Data Frame with an Expression in R Programming using with() Function

14. Plot a pair plot to visualize the relationship between the features in a pairwise manner. A pair plot enables us to visualize both distributions of single variables as well as the relationship between pairs of variables.(Use any kind of dataset)

15. Use sample() function in R Language which creates random sample based on the parameters provided in the function call. Either a vector or a positive integer as the object in the function parameter.

16. Perform the following operation with any numerical data set

    1. Unique values in the data
    2. Visualize the Unique counts
    3. Find the Null values

17. Perform EDA and find the Correlation Plot

18. Perform the exploratory analysis and implement through the use of the Pandas library using Python/R.

    The process consists of the following

    1. Importing a dataset.
    2. Understanding the big picture.
    3. Preparation.
    4. Understanding of variables.
    5. Study of the relationships between variables.

19. For numeric data, perform EDA approach is as follows:

    i. Plot univariate distribution of each numeric data

    ii. If categorical data are available, plot univariate distribution by each categorical value

    iii. Plot pairwise joint distribution of numeric data

20. Perform the Time series EDA and do the plotted time series charts to check for trends and seasonality.

    [For ease of analysis, the code automatically sums up the numeric data by daily, monthly and yearly frequency before plotting the charts. It achieves this through Pandas' resample method.]