Project #1 **04.08.2018**

Project objective:

In our theory classes in Week 1 and 2, we have learned the following.

1) Data cube model for multidimensional data
2) Measurement of central tendency
3) Basics on Programming with R

The above-mentioned concepts are very much linked to the tasks in data analytics. Covering the aforementioned topics the following five projects have been planned. You are advised to implement them **using R programming only**. Implementation in any other programming environment will not be accepted.

Note: You have to complete one project in the list of FIVE topics. To see the topic you have to work, please see the link http://cse.iitkgp.ac.in/~dsamanta/courses/da/students.html

Topic 1

Solve the following using R programming only.

a) Write R program to create a database with the following columns.
   i. Name (Categorical)
   ii. Roll Number (Nominal)
   iii. DoB (Nominal)
   iv. Aadhar Number (Nominal)
   v. Gender (Categorical)
   vi. Mobile Number (Numeric)
   vii. Email-Id (Categorical)
b) Enter at least 10 records into the table.
c) You should save all the entries in the table in secondary storage.
d) Read the table into R's workspace memory.
e) Delete all the records whose "age" is greater than 30 years as on today.

Topic 2

Use the IRIS database. Read the IRIS database into R working memory space and do the following.

a) Find mean, median, mode and midrange (with 10% as the outliers) to the attributes which are applicable to each.
b) Find the standard deviation and variance of those observations having highest mode with respect to "species".
c) Arrange the table in descending order of the sum of the values in the first four columns.

Topic 3

Create three databases with two columns "distance" and "time" contacting the data travelled by a car in some interval of time(s).

Case 1: Different distances covered by the car in the same interval of time.

Case 2: Same distances covered by the car in the different intervals of time.

Case 3: Different distances covered by the car in the different intervals of time.

Enter at least 50 observations for each of the above-mentioned case and then save the database permanently.

Compute the following mean calculation, whichever is applicable.

a) Arithmetic mean (AM)
b) Geometric mean (GM).
c) Harmonic mean (HM).
d) Check the validity AM $\geq$GM $\geq$HM
e) Compute midrange with p = 40%

Topic 4

Use the database GLASS. With reference to this dataset, write programs in R for the following operations.

a) Draw the box plot with the five number summary.
b) Remove the outliers using IQR.
c) Draw the box plot with the dataset after the removal of outliers.
d) Compare the results (plot two box plots on the same graph).

Topic 5

Consider the dataset ADULT. Here, metadata is as follows.

Age: Continuous.

Workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

Fnlwgt: Continuous.

Education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

Education-num: Continuous.
Marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
Occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
Relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
Sex: Female, Male.
Capital-gain: Continuous.
Capital-loss: Continuous.
Hours-per-week: Continuous.
Native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad &Tobago, Peru, Hong, Holand-Netherlands.
Class: >50K, <=50K.

a) Pre-process the dataset. Remove those observations containing missing values.
b) Transform the non-numeric columns into numeric using label encoding and one-hot encoding.
c) Sort the data frame in the descending order of "Hours-per-week".
d) Show all the information whose native country is "US".
e) Subset the dataset based on male and female belonging to the column "sex".

Submission procedure:

1. Prepare a report which should include tool used, methodology followed, reasonable assumptions, if any, etc. Please write the name of the topic, roll number, name and your mobile number on the top of the report.
2. Submit the program files (all are executable).
3. You may create a tar file including the above data using any zip program and submit the same to Moodle system at https://10.5.18.110/moodle/login/index.php .
4. Plagiarism, if found should be taken seriously. It is -100 marks for the common submissions.
5. **Last date of submission : 19.08.2018, 24:00 hours (hard deadline).**