CS 40003 : Data Analytics

Project Assignment #1 04.08.2017

Project objective:

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

1) Data cube model for multidimensional data
2) Measurement of central tendency
3) Descriptive statistics
4) Probability distributions
5) Sampling distributions

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 (preferably) R programming or any other programming environment like Python or Mat Lab.

Topic 1

Reference: CAR data with 50 observations

a) Carefully observe the data. Apparently what it seems?
b) It is proposed to analyse the data with the calculations of AM (Arithmetic Mean), GM (Geometric Mean) and HM (Harmonic Mean). Calculate such mean measures. Which mean calculation has a real significance to the data? Justify your answers.
c) Do you suggest any other measurement(s) which might be useful implications?

Topic 2

Reference: EARTHQUAKE data with 8086 observations

a) The table includes the severity of earthquakes at different places in India during the year 2016. You are advised to browse the data carefully. Point out the discrepancy(ies), if any.
b) For the given data, calculate the "*Five point summary*" and hence draw the box plot.
c) Use the ITR calculation and then decide nay data as outlier(s). Remove the outlier(s), if found. Taking the cleaned data, obtain the box plot? Compare the two box plots.

Topic 3

Reference: AUTOMOBILE data with 205 observations

a) Categorize all the attributes listed in the table according to the NOIR topology?

b) Apply the applicable central tendency measures to any four attributes taking one attribute from each category.

c) Consider the attribute “peak-rpm” and “city-mpg”? Find which probability distribution(s) they are likely to follow?

Topic 4

Reference: IRISH data with 50 observations

a) Consider the 150 observations as very close to population data. Find the population mean.

b) Assume a sample of size 50 chosen at random, find the population variance.

c) Compare the sample variance with that of population variance?

Topic 5

Reference: WEATHER data during 1901-2002

a) The data pertaining to weather from National Data Centres across the major cities in India. The data are in PDF form, which can be easily converted to CSV (Comma Separate Value) format or XLS (Excel Worksheet) format according to your requirement.

b) Store the data using data cube model.

c) Apply the operation(s) (e.g., slice, dice, roll up drill down) to extract a particular data (e.g., the data about North-East region) from the data cube you have obtained.

Submission procedure:

1. Prepare a report which should include Tool used, methodology followed, reasonable assumptions, if any, etc. You may consider separate report for each topic.
2. Submit three program files (all are executable) separately for each topic.
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.
5. **Last date of submission is: 27.08.2017, 12:55 hours (hard deadline).**