# CS 40003 : Data Analytics

## Project Assignment #2 27.08.2018

## Project objective:

Towards the data analytics activities, statistical learning is one of the interesting task, which if carried out effectively discover many hidden information. In this course, we have studied the following topics as "statistical learning":

1) Hypothesis testing
   (Parametric-based statistical inference)
2) Sampling distributions
   i. Z distribution
   ii. t distribution
   iii. Chi-Square distribution

The projects under this assignment are to practice the concepts on the above-mentioned topics with real life data. You are advised to implement only ONE of the following projects as assigned you. You are advised to implement the project using R programming language. A sincere student should attempt to implement all the projects, if possible; however, only the assigned project need to be submitted ad will be evaluated.

## Topic 1

**Reference: HOURLY_WEATHER**

Data description is given below:
Number of attributes = 26, no. of instances = 1048576
Attributes:
wsid - Weather station id, elvt – Elevation, lat – Latitude, lon – Longitude, inme - Station number, prov - State (Province), date - Date of observation, hr - The hour (0-23), prcp - Amount of precipitation in millimetres (last hour), stp - Air pressure for the hour in hPa to tenths (instant), smax - Maximum air pressure for the last hour in hPa to tenths, smin- Minimum air pressure for the last hour in hPa to tenths, gbrd - Solar radiation KJ/m2, temp- Air temperature (instant) in celsius degrees, dewp- Dew point temperature (instant) in celsius degrees, tmax - Maximum temperature for the last hour in celsius degrees, dmax - Maximum dew point temperature for the last hour in celsius degrees, tmin - Minimum temperature for the last hour in celsius degrees, dmin- Minimum dew point temperature for the last hour in celsius degrees, hmdy - Relative humid in % (instant), hmax - Maximum

relative humid temperature for the last hour in %, hmin - Minimum relative humid temperature for the last hour in %, wdsp - Wind speed in metres per second, wdct - Wind direction in radius degrees (0-360), gust - Wind gust in metres per second

a. Have the population with reference to the attribute “temp”. Calculate the population variance.
b. Create a sample of size 30.
c. Infer the mean of population μ = 20 considering the confidence interval 5%.

## Topic 2

**Reference: RAINFALL with 680 instances as the population.**

a. Consider the attribute “max pressure” as an attribute under testing.
b. Create a sample of size 30.
c. Infer the mean of population μ = 1000 considering confidence level 5%.

## Topic 3

**Reference: MOVIE data with 1001 observations.**

a. Calculate population mean from all the movies up to 2015 on “Rating”.
b. Collect a sample of all the movies in the year 2016.
c. Test the hypothesis that “popularity of films (as Rating) increases”.

To test the hypothesis consider following:

i. Population standard deviation is unknown.
ii. Population standard deviation is known.

## Topic 4

**Reference : ENERGY data with 769 observations**.

Data description is given below.

X1 : Relative Compactness
X2 : Surface Area
X3 : Wall Area
X4 : Roof Area
X5 : Overall Height
X6 : Orientation
X7 : Glazing Area
X8 : Glazing Area Distribution
y1 : Heating Load
y2 : Cooling Load

a. Create 50 random samples each of size 30 from the given database.
b. Verify whether the population follows normal distribution or not.
c. How you can tell about the population mean from the sampling distribution.

## Topic 5

**Reference: HEART data with 336 observations.**

Attributes:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

For the given sample find the following.

(a) Calculate the sample variance.

(b) Test the hypothesis that the population variance is 2500 with 5% level of confidence.

(c) What should be the p-value so that the null hypothesis will be rejected?

**Submission procedure:**

1. Prepare a report which should include tool used, methodology followed, reasonable assumptions, if any, etc.
2. Submit the program file.
3. You may create a tar file including the above document 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: 09.09.2018, 24:00 hours (hard deadline).**