Project Assignment #2 20.09.2017

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) Correlation analysis
   (Non-Parametric based statistical inference)
   i. Karl Pearson's Correlation Analysis
   ii. Charles Spearman's Correlation Analysis
   iii. Chi-Square Correlation Analysis
3) Regression analysis
   Simple linear regression
   Multiple linear regression
   Non-linear regression analysis
4) Auto Regression Analysis

The projects under this assignment are to practice the concepts on the above topics with real life data. You are advised to implement all the projects as stated below using (preferably) R programming or any other programming environment like Python or Mat Lab.

Topic 1

Reference: MOVIE data with 5043 observations

a) Calculate population mean from all the movies up to 2015 on imdb_score.
b) Collect a sample of all the movies in the year 2016.
c) Test the hypothesis that "popularity of films (as imdb score) increases".
   To test the hypothesis consider following:
   i. Population standard deviation is known.
   ii. Population standard deviation is unknown

Topic 2

Reference: NUTRITION data with 80 observations

a) Decide whether rating is correlated with sugar content in the product.
b) If correlation exist then what type of correlation (i.e. positive, negative, linear, non-linear)
   Calculate r^2 to support your answer.
   For non-linearity test you should try with up to 3 degree models.

Topic 3

Reference: SALARY data with 1,48,654 observations

Database contains salary information of different employees in different organisations. It is required to test whether Overtime Pay, Other Pay and benefits altogether increases with Basic Pay for the year 2014.

Topic 4

Reference: SNACKS data with 100 observations

a) Find the Spearman correlation matrix of all the ordinal attributes
b) Determine the coefficient of determination (Spearman).
c) Interpret the result from the two tables.

Topic 5

Reference: GAMES data with 16,719 observations

Draw the relevance contingency table to test the hypothesis "action video game is highly rated among teens".
T=teens.(rating column in game data)

Topic 6

Reference: STOCK data for the year 2016-2017

For the given data from stock exchange predict the stock value in the month 1/10/2017.

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 the 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: 22.10.2017, 11:55 hours (hard deadline).**