Middle East Technical University **02.05.2022**
Department of Statistics
2021-2022 Spring

ASSIGNMENT 2 (STAT 412)

(Due to 15 May 2022, Sunday 22:00)

Instructions: (Read carefully)

1. Please do your assignment alone.
2. You can consult with me for help on the assignment, but please don't get help anyone else.
3. NO late submission will be accepted since the answers of this assignment will be shared with you soon after the deadline.
4. Please do not forget to write **your name in all uploaded files** (html / pdf and R script / R markdown)
5. Please show **all the details** of your solutions and interpret all results.
   (Uninterpreted output will cause you to lose points)
6. Please prefer to use R Markdown to do your assignment. Then, create a pdf or html file through it.

(Note: If you have no idea about Markdown, you can prepare your homework by copy and paste method. If you confront any question or problem, please inform me. I will help you.)

7. Please submit your Markdown file or R script (.rmd) in "Assignment 2" section on ODTUCLASS.
8. Please submit your assignment file (pdf) in "Assignment 2 Turnitin report" section on ODTUCLASS.

QUESTIONS

1) *(30 pts)*This data set gives the birth rates and economic development for 30 Nations .The data consist of 30 observations on 5 variables which are birth rates, per capita income, proportion (ratio?) of population in farming, and infant mortality during early 1950s for 30 nations.

Variable Number	Variable Name	Description
1	nat	Nation name 1-30
2	birth.r	Birth Rate in 1953-1954
3	inc	Per Capita Income
4	farm	Proportion of population on farm
5	mort	Infant Mortality Rate in 1953-1954

Data are given in the birthrate.txt file. You need to add variable names by using R.

Firstly, you are supposed to implement 5-fold Cross Validation (*please use set.seed(123)*). Afterwards , estimate a linear models (response variable is birth rate) on 5 folds. Secondly, you are supposed to calculate MSE between the predicted values and the actual values on the test set for each iteration. You interpret your results

Middle East Technical University **02.05.2022**
Department of Statistics
2021-2022 Spring

ASSIGNMENT 2 (STAT 412)

(Due to 15 May 2022, Sunday 22:00)

(**Hint:** the first fold is treated as a validation set, and remaining 4 folds (training set) is used to fit parameters. This process is repeated 5 time, each time , a different observations is treated as validation set).

2) *(70 pts)*The data set titled NY.txt is collected with an interest in observing the impact of new housing projects in New York State Municipalities. The objective is to understand the cost impact on municipal expenditures (spending) resulting from the proposed construction of new housing projects. Since many of the services provided by a municipality are funded largely through property taxes, it is clearly of interest to try to determine whether these projects will produce an increase in expenditures. The columns in the data set correspond to:

Column Description
1 - Row number (you can just ignore this)
2 - State Code
3 - Country Code
4- Expenditure per person (measured in $) (response)
5 - Wealth per person (measures richness only related to real estate property values)
6 – Total population
7 - Percent intergovernmental (percentage of revenue (government income) that comes from state and federal grants or subsidies (support))
8 - Density (=Population/Area)
9 - Mean Income per person
10 - id # (for matching)
11 – Population growth rate
Missing values are denoted with NA.

NOTE: Please set same seed to generate a reproducible random sampling for example set.seed(123)

a. You are expected to preprocess the dataset before going further and analyzing it. In other words, please check whether the data is messy or not. if data is untidy, please convert the data to the appropriate format for further analysis (For example, handling missing cases, performing data cleaning procedure, assigning column names , etc)

b. You are expected to explore the data to identify hidden patterns and insights about data

c. Do you need transformation on data? If your answer is yes, explain the reason why it is needed and apply appropriate transformation for variables. If your answer is “no”, explain the reason why it is not needed.

Middle East Technical University **02.05.2022**
Department of Statistics
2021-2022 Spring

ASSIGNMENT 2 (STAT 412)

(Due to 15 May 2022, Sunday 22:00)

d. You are expected to conduct Validation set approach (80% training, 20% testing)

e. Construct the appropriate model and report your most important findings

Bonus (10 pts): What can your future work be for this analysis?