First Semester 2019
2104529 Computation Methods in Industrial Engineering
Homework 3

Homework

This homework contains 4 questions (100 points) and dues on 17 November 2019. (the non-public data for this homework are available in couseville (`Course Home > Assignments>Homework 03`)

Question 1 [20 points]
Obtain "Teaching Assistant Evaluation" data set from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Teaching+Assistant+Evaluation) and visualize the data using **two** different techniques. (**note:** your score depends on aesthetic, insight, and genuine)

Question 2 [20 points]
After marked 30 questions, an instructor notice a possible cheating of the following 10 students ('Stu-1' ... 'Stu-10'). The questions are multiple-choice question and available in sheet '`multiCheat`', and the solution of each question is listed in the 'Solution' row. Can you detect cheaters (source and copier)?

Question 3 [20 points]
In cryptography, one of the most ancient method to encode a message is called a simple substitution cipher, a method that substitute each character of these message with other unique charter, for example

H	E	L	L	O		W	O	R	L	D
					↓					
O	V	N	N	Z		T	Z	F	N	C

(a) [5 points] encode the message in sheet '`Encryption`', particularly "THE FACULTY OF ENGINEERING ... AND THE WORLD" (`ENCRY_MSG`) using the key generate by the following R code:

```
set.seed(13)
key <- as.vector(sample(LETTERS,size=26,replace=F))
```

(b) [5 points] The most popular way in which a decoder uses to attack the simple substitution cipher is called "frequency analysis". Elaborate this method along with the ENGLISH frequency table and suggest how you would implement it

(c) [5 points] To counter against "frequency analysis", an encoder changes the substitute table by rotating one letter backward or forward in every encoded character. With the original key and the forward rotating key scheme, "HELLO WORLD" becomes "OAHZY JWXQH". Develop R code to encode with the original key and the forward rotating key scheme and encode the message "THE FACULTY OF ENGINEERING ... AND THE WORLD" (**trivial:** this is the concept of ENIGMA machine used to encode in WWII)

(d) [5 points] Use the original key and the forward rotating key scheme, decode the (`DECRY_MSG`) message, particularly "EARWPSKIPTME WL ... HKUJSZVS"

(e) [5 points (bonus)] Implement "frequency analysis" using R code (**hint:** **table** (·), **unlist** (·), (**strsplit** (·, split="")), **sub** (·))

Question 4 [40 points]

Consider "Dresses Attribute Sales" data set from UCI Machine Learning Repository that (https://archive.ics.uci.edu/ml/datasets/Dresses_Attribute_Sales) consists of 'Attribute' and 'Dress Sale'

After separate the data into a training data set (450 instances) and a testing data set (51 instances) with seed `set.seed (13)` and answer the following questions

(a) [5 points] In both data set, list the clean up issue that you have done, e.g., missing value, typo, 'outlier')

(b) [5 points] In the training data set, analyze and visualize daily sale pattern of recommended data.

(c) [5 points] In the training data set, show **one or two important descriptive statistic** of sale grouped by 'style', 'price', 'season', 'NeckLine', 'SleeveLength' 'Material' 'FabricType' 'Decoration' and 'Pattern Type' using **dplyr::summerize** (**hint:** There should be less than 18 tables. If you cannot explain the feature, it is NOT important)

(d) [5 points] In the training data set, apply **stats::princomp()** and graph correlation using **corrplot::corrplot()**

(e) [5 points] From the training data set, develop a classification method to whether a dress would be recommended or not

(f) [5 points] From the training data set, develop a forecasting method to predict the sale.

(g) [10 points] Compare the result in Questions 4.e and 4.f with the testing data set. (**hint:** what is the suitable accuracy measurement)