香港都會大學
科技學院
Hong Kong Metropolitan University
School of Science and Technology

STAT S261F

2022 Spring Mini-Project 1 (UG)

Data Analytics with Applications

Mini-Project 1: Introduction to Data Science

Instruction for submission:

1. This is an individual project and is a component of **personalized assessment**.
2. You should enter your **HKMU student ID** in answering **Question 1**. Failing to do so will result in **ZERO** mark for question 1.
3. Your submission should follow the format of a typical project.
4. It should contain a cover page, table of content, chapters and page numbers.
5. You may use the templates provided by MS word to create the cover page and table of content.
6. Adhere to font size 12, single-line spacing.
7. The structure of your mini-project report

<Cover page>

Include HKMU logo, course name and a project title:

STATS 261F Data Analytics and Application

Mini-project 1: Introduction to Data Science

Remember to put down **your name** and **student ID**

<Table of Content>

Include chapters and page numbers

<Chapter 1 – Basic R operations and Linear Algebra>

Put down the question, your R code answer and the output from R in the report.

<Chapter 2 – Data Collection and Feature Extraction for Text Mining>

Put down the question, your R code answer and the output from R in the report.

<Chapter 3 – Working Principle of my codes>

Explanation of your codes answered in 2j)

Mark Distribution	
Correct Formatting	5 %
Question 1	55%
Question 2	40%

Question 1. Basic R operations and Linear Algebra (55%)

Show clearly all the R codes and their corresponding output for each question.

Table 1 shows the English and Chinese scores of students A – H.

	A	B	C	D	E	F	G	H
English	<1st and 2nd digit >	<5th and 6th digit >	54	75	12	80	90	59
Chinese	<3rd and 4th digit >	<7th and 8th digit >	60	53	49	70	15	30

a) Fill in the scores for students A and B with **your HKMU student ID**. Students who do not follow this instruction will receive **ZERO** mark for the whole question 1. If the 3rd digit, 5th digit or 7th digit of your student ID is zero, just assign a 1-digit score (i.e. 4th digit, 6th digit, 8th digit). If a pair of digits are zero (e.g. both 3rd and 4th digit are zero), just assign zero as the score.

b) Create a matrix called **student_matrix** with 2 rows and 8 columns, where
   i) Row names are "English" and "Chinese", respectively.
   ii) Column names are A,B,C,D,E,F,G,H.

(10 marks)

c) Transpose the matrix such that "English" and "Chinese" become columns.

(5 marks)

d) Find the student getting lowest mark in English.

(5 marks)

e) Find the student getting highest mark in Chinese.

(5 marks)

f) Count the number of students who score higher than 50 in English.

(5 marks)

g) Count the number of students who score lower than average in Chinese.

(5 marks)

h) Create a vector called **total** and calculate the total scores for **each student** as
   Total[i] = English[i] + Chinese[i] for each i-th student using **vector/matrix operations**

(5 marks)

i) Combine the vector **total** to the matrix **student_matrix**.

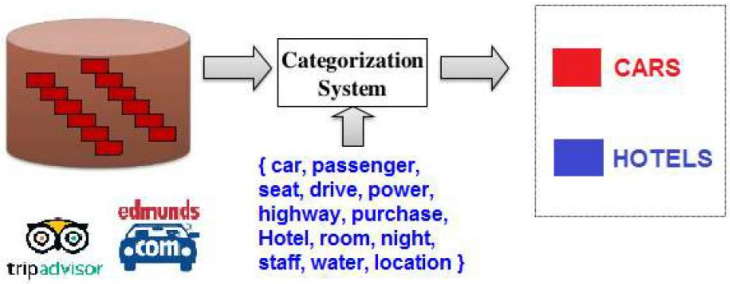
(5 marks)

j) Create a **matrix** called **grades** and assign grades to the marks of each subject as follows:

(10 marks)

Marks	Grade
>= 80	A
>= 60 and < 80	B
>= 50 and <60	C
>=40 and < 50	D
Below 40	FL

Question 2. Data Collection and Feature Extraction for Text Mining (40%)



In this question, you shall perform data collection and feature extraction for datasets obtained from **trip advisor** and **Edmunds** for a text classifier.

Show clearly all the R codes and their corresponding output for each question.

a) Unzip **mini-project-dataset.zip** obtained from OLE.

b) Create a vector called **dictionary**, which contains the following words {"car", "passenger", "seat", "drive", "power", "highway", "purchase", "hotel", "room", "night", "staff", "water", "location"}

(1 mark)

c) Read the file "01.txt" and save the text to a variable **txt01**.

(1 mark)

d) Convert all text in txt01 to lower case. (e.g. from "ABCDefg" to "abcdefg")

(1 mark)

e) Create a new variable **keyword1** and assign keyword1 as the 1st element of dictionary, i.e. "car"

(1 mark)

f) Create a new matrix **allfeatures** with 15 rows and 13 columns. Fill all elements with zero.

(1 mark)

g) Count how many times the word "car" appearing in **txt01** and save the count to allfeatures[1][1]. You may refer to homework 1 for reference.

(5 marks)

h) Repeat steps g) and h) to count the remaining keywords of **dictionary** and save the counts to **allfeatures**[1][i] for i=2,3,4,5,…to 13 using a **FOR LOOP**.

(5 marks)

i) Repeat steps c) to i) to read and count the remaining files, 02.txt to 10.txt, 31.txt to 35.txt and save the counts to the remaining rows of the matrix **allfeatures**. Name the rows and columns of the matrix as follows:

	car	passenger	seat	….	staff	water	location
01txt							
02txt							
….							
35txt							

(12 marks)

j) Explain your codes in j) **line by line** how it works.

(13 marks)

-END-