UGANDA MARTYRS UNIVERSITY

FACULTY OF SCIENCE

DEPARTMENT OF NATURAL SCINCES

UNIVERSITY EXAMINATION

FIRST YEAR EXAMINATION FOR FIRST & SECOND YEAR EXAMINATION FOR BSC

GENERAL/ ECON & STAT

FINAL EXAMINATION

STATISTICAL ORGANIZATION

DATE OF EXAINATON: 20/07/2022

TIME: 3 HOURS

INSTRUCTIONS:

i. SECTION A IS COMPULSORY AND ONE NUMBER IN SECTION B

ii. Do not write anything on the question paper

iii. Show all workings and they have to be clear and tidy

iv. ENSURE YOUR KNIT YOUR DOCUMENT AND SEND TO https://drive.google.com/drive/folders/1bHB1tHYj7JpVVNgm9NQZWPyXxVuHowu6?usp=sharing

v. If you successfully complete a task, you will receive full points for that task and may move onto the next task.

vi. If you don't complete a task, or you have errors in your script or output, or R shows warning messages, you will receive 0 points to the current and all following tasks (unless stated otherwise). Hence, make sure to successfully complete each task before moving to the next one.

SECTION A

QUESTION ONE [30 MARKS]

I. ACCESS THE UCI MACHINE LEARNING REPOSITORY AND USE THE TURKISH HEADLINES DATASET.

II. Using dataset provided, you're required to conduct the following analysis.

a. Import the dataset in R using appropriate functions.
b. Describe the dataset dimensions and nature of attributes.
c. Return summary statistics and explain what these mean in relation to the data.
d. Correlation
e. Visualization
f. Joining datasets
g. Sub setting data sets.
h. Reshaping (pivoting long to wide and vise vasa)
i. Aggregate functions through grouping
j. Functional programing
k. Data search (missing values, outliers)
l. Imputation &/ deletion.

QUESTION TWO [50 MARKS]

You are a great scientist (congrats!). Sir Richard Branson heard about your great success and wants to pay you a ton of money to run a study to investigate the effect of automation on spacecraft pilots' performance. In short, Richard is interested in understanding whether or not using automation will improve pilots' performance.

The data from the experiment that you just finished running is saved in a csv file and is available here.

The dataset is organized as following:

- **Participant**: participant # 1 to 12
- **Mode**: Each pilot underwent two within-subject conditions: *Manual* and *Automated*. In the manual condition, the system was operated manually. In the automated condition, the automated system was engaged.
- **Events**: to understand pilots' reactions to unexpected events, pilots were instructed to press a button every time they were presented with an event (a light in this case). They were presented with 6 events in total: event #1 to 6.
- **RT**: Response Times (RT) to the event detection task were recorded in milliseconds.

Note that for this particular task you will need to set a directory which is unique for your machine and local folder. That's totally fine. When grading, I will change the directory in your script back to a different directory on my machine. However, make sure to assign a name to your dataset and always reference that dataset (or derivate datasets) throughout the script, and never reference the original csv file again in your script.

Task 2

Create the factor **Gender** and add it to the dataset.

- Participants 1-6 are males
- Participants 7-12 are females

Task 3

Create the factor **Age** and add it to the dataset. This factor has three levels: young, mid, old.

- Participants 1, 3, 4, 7 are **young**
- Participants 2, 5, 6, 8 are **mid**
- Participants 8 through 12 re **old**

Task 4

Using **mutate ()** or its combinations, create a **Response** variable from **RT** and add it to the dataset.

- If **RT** < 500, then **Response** is short
- If 501 < **RT** <1000 then **Response** is medium
- If **RT** > 1001 then **Response** is long

Task 5

Summarize the dataset to create **data_sum** that has mean, standard deviation and standard error of RT broken down by mode and gender. All nonnumeric values may need to be removed.

Task 6

Utilize **data_sum** to create a histogram with:

- **RT** on the y axis
- **Mode** on the x axis
- **gender** as grouping variable
- *Response Times (in ms)* for the y axis' title
- *System Mode* for the x axis' title

Assign the name **myPlot** to it.

Task 7

Complete 7.1 and 7.2

7.1 Run an ANOVA with RT as dependent measure and mode as within-subject factor.

. A few important things to keep in mind:

- You may need to omit empty cells or NAs from your dataset first.
- When running the ANOVA, R may show you the following message.

> Warning: Collapsing data to cell means. *IF* the requested effects are a subset of the full design, you must use the "within_full" argument, else results may be inaccurate.

If you see this or a similar message **together with the results of the ANOVA**, that's fine and you can move on to the following task. Make sure the results of the ANOVA are shown though.

7.2 Run an ANOVA with **gender** as between subject variable.

You might see a similar error message as in 7.1. If so (and even if you don't) move onto the next task and consider this task successfully completed, **provided the results of the ANOVA are shown.**

Task 8

Following what you have done for task 7.2, run an independent t-test to investigate the effect of gender on RT. In addition, calculate Cohen's d for this comparison.

Task 9

Turn the **data** dataset from a long format to a wide format so to have RT for Manual and Automated Modes into 2 separate columns.

Task 10

You run a survey study where you ask 30 participants, with different ages and genders, five questions each: Q1, Q2, Q3, Q4, Q5. Participants answer the five questions on a scale from 1 to 7. When importing the data into an Excel file, you do so by using a wide format. Unbeknownst to you, however, your adviser is fervently

- against using wide formats. So, before you present your work to them, you decide to quickly change the dataset from wide to long.

Access the dataset here, and turn it into a long format before your adviser finds out.

SECTION B

QUESTION THREE [20 Marks]

I. Define the following terms within R environment
   a. Package.
   b. Functions.

II. Explain the different atomic data structures in R and use appropriate examples to show how these are created.

III. Explain with examples the difference between user defined fictions and system functions in R.

IV. R programing language is considered to be an open source language, explain that this means.

QUESTION FOUR [20 Marks]

I. Let V = (1,2,3,4,5) be a vector within the R environment. Write code that can be used to carry out the following operations.
   a. Count the number of elements in V.
   b. Delete the 3rd element from V.
   c. Add a new element into V.
   d. Delete V from the R environment.

II. Given any matrix A in the R environment, describe with syntax the descriptive statistics that you would do.

QUESTION FIVE [20 Marks]

I. Describe the different components of a user defined function.

- II. Write a user defined R function to compute the standard deviation of any input numbers.

III. Evaluate the performance of the function by assigning a vector to the function name.