SPECIAL COURSE IN SOFTWARE ENGINEERING, AUTUMN 2024

EXERCISE 4 – MORE DATA ANALYSIS

In this exercise, you will write program to conduct more data analysis using ***regression, t-test, and chi-square***.

THIS EXERCISE WILL BE GRADED (MAX: 10 points)

SUBMISSION DEADLINE: OCTOBER 18, 2024, MIDNIGHT

Task 1. dataset_loading()

This function simply loads the dataset that's available in the root folder.

Task 2. list_column_types()

This function is similar to one you implemented in **GEX3**. The function sorts all the columns into interval, numeric ordinal, non-numeric ordinal, and nominal.

It stores the column names and the unique elements in that column in a dictionary called **column_types**

Task 3. select_variable()

The function is similar to the function you implemented in **GEX3.** The function lists all the available columns and their types. Asks the user the type of data that is needed to implement the test.

This function is called from the **main()** function, depending on the test the user has selected to perform.

If the user selects **t-test or Mann-Whitney U test**, then this function should have the logic to tell the user if a column has more than two categories. If that's the case, the user should be asked to enter another variable.

The function takes type of **data (e.g., nominal, interval), max_categories (e.g., 2, for t-test)** and **allow_skip (e.g., True or False)** as parameters.

The function returns the variable name that the user enters.

Task 4. check_normality()

The function is similar to the function you implemented in **GEX3.** The function checks if the column has more than 2,000 items. If it has, then it will check normality using Anderson-Darling Test. Otherwise, it will use Shapiro-Wilk test.

The function accepts the **column,** and **size_limit** as the parameter, and returns the **statistic** and **p_value** from the test.

Task 5. perform_regression()

The function takes two columns as the parameters, handles, NA entries, and performs simple linear regression.

Task 6. t_test_or_mannwhitney()

The function accepts a continuous variable and a categorical variable (nominal or ordinal) as its parameters. The test first checks for normality and based on the result from that **check_normality()** function, either **t-Test** is conducted or **Mann-Whitney U Test**.

The function returns the **test statistic** and **p_value**.

Task 7. chi_square_test()

The function accepts two **categorical variables (nominal**) as its parameters.

The function returns the **test statistic** and **p_value**.

Task 8. main()

After creating the object and loading the data, the function presents the following options to the user to pick from:

1. t-test or Mann-Whitney U Test
2. Chi-square Test
3. Linear Regression

Based on the user choice, the appropriate function is called, while making sure that first the right variable is selected from **select_variable** function. If the user chooses '**1**', then the program should inform the user if there are any nominal variable where there are more than 2 categories. If all the columns have more than 2 categories, the program should inform the user that **t-Test or Mann-Whitney U Test is not possible** with the available dataset.

Similarly, if there are no nominal variables suitable for **Chi-square Test**, the user should be informed. And in case of regression, if there are not enough continuous type variables, then **Linear Regression** should also not be possible.