

Student Name:

Submission Date:

DBST 667 – Data Mining

Dr. Irene Tsapara

Week 2 Individual Exercise

Deliverables: Two Files: (1) Submit this lab report with answers to all questions including output screenshots into the ‘Individual Exercises Week 2’ assignment folder. (2) Submit an R script that contains all commands with comments that briefly describe each commands purpose.

Grading: This exercise is worth 2% of the course grade. All questions must be answered in your own words with any paraphrased references properly cited using in-text citations and a reference list as needed. In addition, grammatical and spelling errors may affect the grade.

Part 2 – Run an exercise on the CreditApproval.csv data set, completing this report and providing the commands, output screenshots, and discussion/interpretation as requested. Ensure that all commands are saved in this report and in an R script.

a. Introduction:

- i. Read the dataset description at [UCI Machine Learning: Credit Approval](#). In your own words, describe your understanding of the dataset, what the attributes (columns) mean, and what each observation (row) represents.

- ii. Run the `read.csv()` command to load the data into a variable named ‘credit’. Then, run the command to preview the first 10 data rows in ‘credit’. Include the command and output screenshot. *Note: Ensure that you use the `utils::read.csv()` command and not any other similar commands from other packages.*

Command: >

Output:

- iii. **Run the `str()` command on the ‘credit’ dataset. Include the command, output screenshot, and a brief description of how the structure is presented.**

Command: >

Output:

Description:

b. Descriptive Statistics:

- i. **Run the `summary()` command on the ‘credit’ dataset to display the descriptive statistics for all variables. Include the command and output screenshot.**

Command: >

Output:

- ii. Choose two numeric attributes from 'credit', run the `summary()` command on both, and provide your interpretation of each of the six descriptive statistics.

Command: >

Output:

Command: >

Output:

Interpretation:

- iii. Choose two factor attributes from 'credit', run the `summary()` command on both, and provide your interpretation of each of the six descriptive statistics.

Command: >

Output:

Command: >

Output:

Interpretation:

- iv. What differences did you observe between the output of the `str()` and `summary()` commands?

c. Variable Filters – Discretization and Removing Variables:

- i. What is discretization? Provide a one-paragraph, masters-level response in your own words.
- ii. Run the three different discretization methods discussed in the tutorial (equal interval, equal frequency, k-means clustering). For each method, include the command and output screenshot. For all commands, provide a one-paragraph discussion of the input parameters used, the number of bins, and your interpretation of the output.

Command: >

Output:

Command: >

Output:

Command: >

Output:

Discussion:

- iii. **Compare and contrast the discretization methods above providing at least one example of when you would use each one.**

- iv. **Run a command to remove one of the attributes from the ‘credit’ dataset. Run another command to demonstrate that the attribute was successfully removed. Include both commands and output screenshots as well as a discussion of when and why variables should be removed from a dataset.**

Command: >

Output:

Command: >

Output:

Discussion:

d. Row Filters – Handling Missing Values and Sorting:

- i. **Run a command to check if the ‘credit’ dataset has any missing values. Your command and output should show all attributes along with how many observations total have missing values. Include the command and output screenshot.**

Command: >

Output:

- ii. **Choose one of the numeric attributes with missing values and run the command to replace the missing values with the attribute mean. Then run the command to verify that the variable no longer has missing values. Include both commands and output screenshots.**

Command: >

Output:

Command: >

Output:

- iii. **Why is it important to handle missing values in your dataset prior to beginning your primary data analysis? Provide a one-paragraph, masters-level response in your own words.**

- iv. **Describe at least one alternative approach for handling missing values other than replacing the values with the attribute mean. Provide a one-paragraph, masters-level response in your own words.**
- v. **Run the command to sort the ‘credit’ dataset by one of the attributes. Then run the command to validate the sorting. Include both commands and output screenshots as well as a discussion where you provide at least two reasons why data should be sorted.**

Command: >

Output:

Command: >

Output:

Discussion:

e. Data Visualization:

- i. Run the `plot()` function for one of the variables in the ‘credit’ dataset. Include the command, output screenshot, and a one-paragraph, masters-level interpretation of what the plot shows.**

Command: >

Output:

Discussion:

f. Summary:

- i. Why is data pre-processing important? Describe at least two advantages that pre-processing results in as well as two disadvantages of not pre-processing. Provide a one-paragraph, masters-level response in your own words.**
- ii. What differences did you observe between variable filters and row filters? Provide at least one scenario for each filter type where implementing the filter would benefit your data analysis. Provide a one-paragraph, masters-level response in your own words.**
- iii. (Not Graded) Which section of this exercise did you find the most challenging? What approach did you take to resolve this challenge? Were there any sections from this report or the tutorial which could use additional focus?**

References