Top of Form

**Exercise 2- Data Pre-processing in R**

**Exercise Overview**

In this exercise, you will use the RStudio interface to load data from a CSV file, to explore the statistics that this software provides for different attribute types, and to apply data pre-processing filters.  The data pre-processing filters include selection of a subset of variables, discretization methods, handling missing values, and more.

**Exercise Instructions**

This exercise consists of two parts. The first part will help you master and apply the key concepts from course readings to the data analysis. **The first part is not graded and you do not need to submit a report for part 1 before starting part 2. However, you should do it** to get maximum benefit from the exercise. **Part 2 must be submitted**.

|  |  |
| --- | --- |
| **Part 1**  Complete the exercise in the Word document on data provided in the cars.csv file.  You do not need to submit anything for this part.  Keep in mind that your output might be slightly different depending on your operating system and RStudio version. | Car Risk Level Bar Chart |

**Part 2**

Use the provided CreditApproval.csv dataset for this part. Follow the instructions below to write an R script with comments and to write a report on your findings and results interpretation.

**The report should be in your own words and should cover the following key points in the same order as below.**

1. **Introduction**
2. Read the dataset description at <http://archive.ics.uci.edu/ml/datasets/Credit+Approval>   Explain what data is in the dataset, and what each row represents.
3. Use the read.csv command to load the CreditApproval.csv data into RStudio.  Then, run the command to preview the first 10 data rows.  Include the command, the output screenshot, and an explanation of what the output shows.
4. Run the str command.  Add a screenshot of the output to the report, and explain what the output tells you about the Credit Approval data structure.
5. **Descriptive statistics -**Run the summary command to display the descriptive statistics for all variables. Include the command, output, and answers to the following questions in the report:
6. What descriptive statistics does the output show for numeric variables in the credit approval data?  Provide some examples.
7. What descriptive statistics does the output show for factor variables in the credit approval data?  Provide some examples.
8. What are the differences between the str command and summary command?
9. **Variable filters**
10. Discretization
11. In your own words, explain what discretization is.
12. Run 3 different discretization methods discussed in the tutorial.  For each method, include the command you ran, and output screenshot.  Discuss the input parameters you used, including the number of bins, and interpret the command output.
13. Compare the discretization methods you ran, and discuss when you would use each method.
14. Run the command to remove one of the dataset variables, and run a second command to validate that the variable no longer exists.  Include both commands and the output in the report.  Explain when and why we remove variable(s).
15. **Row filters**
16. Missing values
    1. Run the command(s) to check if the data has missing values. Include the command(s) and the output in the report.
    2. What variables have missing values, and how many values is each variable missing?  Support the answer with an output screenshot.
    3. Choose one of the numeric variables with missing values, and run the command to replace the missing values with the variable mean.  Verify that the variable no longer has missing values.  Include and discuss the commands you ran and the commands’ output.
    4. Why do we need to handle the missing values?
    5. Besides replacing the value with the mean, what alternative approaches for handling missing values could we use?
17. Run the command to sort the data by one of the variables.  Then run the command to validate the sorting.  (See example in the tutorial) Include the commands and the screenshot in the report.  What are some reasons for sorting data?
18. **Data visualization -**Run the plot function for one of the variables.  Include the command, the output, and an explanation of what the output shows.
19. **Summary**
20. Why is data pre-processing important?
21. What are the differences between row and variable filters?
22. Which part of this exercise did you find the most challenging? What approach did you take to resolve the challenge?

**Exercise Deliverables (Part 2 only)**

Submit the following files into your Exercise 2 Assignment Folder:

* Word document with answers to all questions in order.
* An R script with all commands for part 2. The order of commands must follow the order of corresponding questions.

**Exercise Grading**

* ***This exercise is worth 2% of the course grade.***
* *This exercise* ***will not*** *be graded if an R script is not submitted.*
* *To get full credit, the R code must have comments that explain what the code does*
* Cite references on a references list and in text as needed.
* *The report must be your original work and should address all key points above in order.*
* *Keep in mind that no points will be returned for any misplaced answers that did not get credit.*
* *Grammatical and spelling errors may affect the exercise grade.*

***Start working on the exercise early, and post any questions in class. This exercise may take 3-5 hours to complete.***

Bottom of Form