Top of Form

Tiania Chan

Data 630

Assignment 2

|  |  |
| --- | --- |
|  |  |

**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. A screenshot of a computer

   Description automatically generated with medium confidence
   1. Head is showing the first 10 rows of data. Since the window isn’t large enough to show all at once, the data looks like it is stacked.
5. 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.
6. Graphical user interface, text

   Description automatically generated
   1. This is printing out the dataframe as a string so that you can see all of the data structure
7. **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:
8. Graphical user interface, text

   Description automatically generated
9. What descriptive statistics does the output show for numeric variables in the credit approval data?  Provide some examples.
   1. The descriptive statistics from the summary command shows the basic shape and some data that can be helpful in getting a high level understanding of the data, For example, getting the min and max of each column for key, male, age, debt, etc gives us an idea of the range of data to look for. Looking at the mean and median gives the reader an idea of where the data might be if it is a normal distribution.
10. What descriptive statistics does the output show for factor variables in the credit approval data?  Provide some examples.
    1. An example of a factor variable shown in summary statistics would be education level. It shows c,w,q,I,aa,ff, and then a broad category labeled other. Another example would be ethnicity which is coded as v,h,bb,ff, j. and other.
11. What are the differences between the str command and summary command?
    1. The summary command focuses more on numerical values and focuses on displaying the information found in the columns so that a user gets an idea as to the shape of the data in the dataframe’s columns. On the other hand, the str command shows more of the data structure of the dataframe by listing the columns, the data type, identifying factor variables and giving a sneak peek as to what the first few rows of data. Combined together, this can really help a data analyst get a better idea what queries to use and how to clean the data as well.
12. **Variable filters**
13. Discretization
14. In your own words, explain what discretization is.
    1. Discretization is where a user may take noisy data that is spread out and categorize them into easily digestible categories or levels. In the reading, we use age and label the data as youth, adult, and senior.
15. 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.
16. Text

    Description automatically generated
    1. The first discretization method I used was interval on the numerical value column, Age. I wanted to see if we broke it down into 5 equal intervals, what the data would look like. It is displaying the interval beginning and end and age in each of the intervals. So we can see that 27 years to 40 years makes up majority of the data in the column Age.
17. Text

    Description automatically generated
18. 2. The second discretization method I used was the cluster to see if the data is grouped around certain values. A better way to go about this would have been to plot the data first to understand if there are any clusters first then run the cluster discretization method.
19. Text

    Description automatically generated
    1. The last discretization method I rant was frequency to see if there was an even distribution with debt in the dataset. In this case there is not as you can see majority are below 9.43 so the width of the category is largest between 9.43 and 28.
20. Compare the discretization methods you ran, and discuss when you would use each method.
21. 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).
    1. Graphical user interface, text

       Description automatically generated
    2. I chose to delete the 2nd variable which was income so that is why instead the variable Age is missing from the second data frame whereas the first data frame contains Age. Also you can see that creditshort has 16 variables where as creditapproval and creditapproval2 both have 17 variables. A reason why a user may want to delete a variable is if the dataset is too large so they want to speed up the processing time, to clean the data so it only contains relevant data so any future users aren’t having to look at irrelevant data, or if there are simply too many nulls, a user may choose to delete a variable instead of keep a variable since it is not used.
22. **Row filters**
23. 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.
       1. 
       2. A picture containing building, computer, window blind

          Description automatically generated
    2. What variables have missing values, and how many values is each variable missing?  Support the answer with an output screenshot.
    3. A screenshot of a computer

       Description automatically generated with medium confidence
       1. According to the screenshot, only Age and Zipcode have missing values with Age missing 12 and Zipcode missing 13.
    4. 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.
       1. A screenshot of a computer

          Description automatically generated with medium confidence
       2. The two commands I ran to replace the null values in Age with the mean value was creditapproval2$Age[is.na(creditapproval2$Age)]<- mean(creditapproval2$Age, na.rm=TRUE) which took a look at the null values in the Age variable and assigned the mean of the Age variable to each of the null. I was then able to check to make sure this was executed correctly by running apply(creditapproval2,2,function(creditapproval2)sum(is.na(creditapproval2))) again which shows that the Age variable now shows 0 null values.
    5. Why do we need to handle the missing values?
       1. The reason for handling missing values is because that can be considered unclean data and makes analysis of the data more difficult to complete. One method of handling missing data is by substituting with the mean of the column
    6. Besides replacing the value with the mean, what alternative approaches for handling missing values could we use?
       1. Besides replacing null values with the mean, we can also remove rows with null values if it is a small enough percentage of the data, replace with other measures of central tendencies, replace with a blank, or replace with a null value. The method that is taken really depends on the data set and the demands of the requirement as to which will work best.
24. 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?
25. Text

    Description automatically generated
    1. Sorting data is useful if you need to analyze the data by time, value, or if you need to group by the value of a variable. It can also help make machine learning algorithms run faster as they search through data, they can find similar values easier if they are closer together.
26. **Data visualization -**Run the plot function for one of the variables.  Include the command, the output, and an explanation of what the output shows.
27. A screenshot of a computer

    Description automatically generated with medium confidence
    1. The output shows a bar chart in the color dark green with the number of customers in each age range. The maximum appears to be 12 people between the ages of 30 and 34 with a higher concentration of customers in their 20s to 20s with a spike of 12 people between ages 30 to 34.
28. **Summary**
29. Why is data pre-processing important?
    1. Data preprocessing is extremely important mostly because real world data is rarely as clean as manicured data that we see in school projects. (Hans, Kamber, Pei, 2011) The reasons for low quality data are numerous and most of the time spent extracting meaning from data is spent cleaning. It is a well known trope that data scientists spend 80% of the time cleaning data and only 20% on building models. The reason why is because without clean data, extracting meaning and building models would be inefficient and meaningless with all the extra noise that unclean data has.
30. What are the differences between row and variable filters?
    1. Besides syntactical differences in R, the most obvious difference between row and variable filters is that rows have all of the variables whereas variables only contain the one type of value. The way I think about it is columns are variables and rows are the slices of the columns. With a variable filter, the number of variables changes so the row length also changes. On the other hand, a row filter can add, remove, and process rows but the variable length will remain the same.
31. Which part of this exercise did you find the most challenging? What approach did you take to resolve the challenge?
    1. The most difficult part of this exercise was using the discretization methods. I may have over thought them by wanting to figure out which method would be best for each variable before I ran the commands which meant that it took me more time before I actually ran that part of the assignment. This I overcame by deciding against that way of thinking and simply going to try the method on the variable and use each as a learning experiment to see what worked and what didn’t and why. I also had a typo which caused me to spend the most time on this section of the assignment.

Bottom of Form