Visualization and data processing

Total Marks - 60

Sourav Das

21 May 2021

## Assessment 2

The following exercises are designed to assess your understanding of concepts, implementation and interpretation of topics in Visualization and Data Processing. Some questions may require you to search and use R functions that we have not used so far. In all following questions submit codes and output.

Note: The questions in this assessment may have multiple correct solutions. Hence submission of R code is essential. Almost no statistical background is presumed knowledge for this assessment. All methods required for solution are available on the content pages of Weeks 2-5 of this subject. Some of them have been covered in detail during collaborate sessions.

**Answer sheet and script (code).**

All outputs must be followed by a short statement, for full marks. You could use on of the following alternatives in responding to the questions.

1. Please use a new word document to provide your responses sequentially using Section and Question numbers. You could paste your R code followed by output and discussion in the word document.
2. In a word document provide your output and discussion. Submit the –annotated (commented)- R code separately. Codes without annotation won’t receive full marks.
3. If you know Rmarkdown, you could use that to create an integrated report. However, Rmarkdown is NOT a requirement for this subject.

## A. Visualization: Section Marks 10

Import the data *oneworld.csv* into R. Show the class of all variables. Submit code and output. The objective in this question is explore the relationship between the variables GDP vs Industry by regions.

1. Create a new ordered categorical variable GDPcat - with three categories such that the proportion of countries in each categories is 40%, 40% and 20%, respectively, in increasing order of magnitude of the variable GDPName the three categories to be “Low” “Medium” and “High”. Show your code and the solution.

Hint: Remove any missing observations. There is a function in R that helps find such percentage values for variables. Marks (4)

1. Use a visualization tool to display the relationship between *GDP* and Infant Mortality, stratified by *Regions*, on a single plot. You must use ggplot2 for visualization. For plotting you could use either the variable GDP or the variable GDPcat. Comment on your observations. Hint: There are multiple possible correct responses, but you must justify your choice. Note that your choice of the GDP variable would dictate the choice of the plot function. Marks(6)

## B. Data Processing – I: Section Marks 15

1. Write an R function to identify the proportion of missing observations in a variable or column of tabular data. Marks (4)
2. Implement this function across all variables of the dataset *airquality*. This dataset is available with base R. Show your code and the result— that is, the proportion missing in each variable. Marks (2)
3. Specify a variable from this dataset that you would select for univariate missing value imputation. Hint: Justify your variable choice based on the count or proportion of missing observations, noting that univariate imputation reduces the natural variation of a variable. Marks (2)
4. Show your code -in base R - to replace all missing observations in the chosen variable with some value, based on the data. Justify the choice of replacement value. Hint: Read the appropriate section on your Weekly content page to perform this task. Marks (3)
5. Summarize the results after imputation and compare with the pre-imputed variable. Hint: In commenting you can use a use a descriptive statistics function or a visualization tool learned on this subject. Marks (4)

## C. Data Processing – II: Section Marks 20

Clinician scientists at Royal Melbourne hospital are investigating the relationship between fecal calprotectin (FC) as a non-invasive diagnostic alternative to Inflammatory Bowel Disease (IBD) and Acute Sever Ulcerative Colitis (ASUC). It would also contribute to standard of care. A subset of the dataset is provided as bowel.csv. The data has several missing values, causes of missing-ness are often unknown.

In the following show questions show your R working and output.

1. In the bowel.csv dataset,
   1. Count the number of missing observations on the variable FC and in the overall dataset. Marks (2)
   2. Perform a univariate imputation on the variable *FC*. Your solution should include Marks (3)
      1. code,
      2. result, and
      3. Justification for the choice of imputation value you used. Hint: please consult the relevant topic page.
2. Select two variables that may have association with *FC*.
3. Justify your choice using dplyr()based tool(s). Show your working in R. Marks(4)
4. Use your investigation in a.) to replace all missing values in the variable FC. Show your code. Marks (6)
5. Present a comparison - with discussion - between the two types of imputation in the context of variable *FC*. You may use tools from topics in Weeks 2 and 3, for comparison. Marks (5).

## D. Text Analytics: Section Marks 15

Mysterydocs.RData is a collection of unstructured text documents. The response to the questions below must include comments, wherever applicable. This question tests both implementation and conceptual understanding of data analytic tools that you have learned in this subject.

1. How many documents are there in the dataset? Justify your answer using an R function (show your code and result). Mark (1)

2. Using methods of Week 5 Topic 2,

a. clean this collection of texts and convert it into tabular data. You must use at least 5 cleaning steps, including stemming.

b. Show your code and the last six rows and first five columns (only) of the tabular data that you created. Marks (4)

3. Create a subset of the tabular data you constructed in QD2., retaining only those words that have occurred at least 200 times. Use a visualization tool to show the frequency distribution of words in this subset data. Hint: Select an appropriate visualization tool from your learnings of Week 3. Marks (3).

4. Using tools learned in this subject to quantify and depict any similarities between the documents.

Is there any obvious structure in the similarity matrix as depicted by the plot? If so what does this structure mean in terms of the original data? One or two sentences at the most.

Please use the original tabular data that you created in Q.D2. You have to use an appropriate (relevant) weighting metric described on your content page. Hint: You have to use (at least) one quantitative measure and (at least) one visualization tool to justify your answer. For visualization of the similarity matrix you may use R functions such as levelplot() or image()or any other suitable plotting function. You would have to research the implementation of these functions. Show your code, results and comment on your findings. Marks (7)