MATH2349 Data Preprocessing Assignment 3

(Last Updated 15.10.2019)

Weight: 25%
Due date: 27 October 2019, 23:59 AEST.
Length: Maximum 20 pages
Feedback mode: Feedback will be provided using Turnitin's inline marking tool and general comments.

Purpose

The purpose of this final assignment is to put to work the tools and knowledge that you gain throughout this course. This provides you with multiple benefits.

- It will provide you with more experience using data preprocessing tools on real life data sets.
- It helps you to self-direct your learning and interests to find unique and creative ways to wrangle your data.
- It starts to build your data analytics portfolio. Portfolios (or e-portfolios) are great way to show potential employers what you are capable of.

Overview

This assignment requires you to find some open data, and use your knowledge gained during the course to preprocess the data. This is your opportunity to demonstrate all that you have learnt during this course. You will be awarded (with marks) the clearer you demonstrate your skills. This assignment is worth 25% and is due **27/10/2019**.

Groups

Students are permitted to work individually or in groups of up to 3 people for Assignment 3. **Each group must fill out the following form before 20/10/2019 to register their group details. After the deadline, group registrations won't be accepted.** Submit the details of your group here:

Group Registration Form

All group members must submit a copy of the report! Group members that are not registered and do not submit a report will not be acknowledged.

Assignment Data

Assignment 3 is open-ended, but there will be one key requirement. The data to be used must be open and ideally have a Creative Commons Licence. This will ensure you can share your work with anyone provided you make proper attribution. If you're not sure if data is

Open, contact the provider, read the documentation or post on Slack and I will investigate. Some open data sources are provided below, but I encourage you to find others:

- https://www.kaggle.com
- UCI Machine Learning Repository
- data.gov
- world bank
- amazon web services
- google data sets
- youtube video data sets
- analytics vidhya
- quandl
- driven data
- http://www.abs.gov.au/
- https://www.data.vic.gov.au/
- http://www.bom.gov.au/
- https://relational.fit.cvut.cz

Minimum Requirements for the Data sets

Considering this is a data preprocessing class, I do expect your data set to have certain requirements in terms of data attributes so that you can demonstrate your knowledge of data preprocessing.

The following are the minimum requirements for the data sets that I will look for:

1. **At least two data sets should be merged** to create your assignment data (for example you can take crime statistics for the cities/states in Australia and merge this data set with per capita income data)
2. Your data set should include multiple data types (numerics, characters, factors, dates, etc)
3. Your data set should include variables suitable for data type conversions so that you should be able to apply the **required data type conversions** (i.e., character -> factor, character -> date, numeric -> factor, etc. conversions)
4. Your data set should include **at least one factor variable** that needs to be labelled and/or ordered.
5. Your data set can be in a tidy/untidy format. **As a minimum, I expect you to determine whether your data is tidy or untidy.** If it is in a tidy format, you do not need to do anything. But if it is untidy, then you need to apply the required steps to reshape your data into a tidy format.
6. **At least one variable needs to be created/mutated** from the existing ones (i.e. the data may contain income and expense variables and you may create a savings variable out of the income and expense variables)
7. I expect you to **scan all variables for missing values/ inconsistencies.** If there are missing values/inconsistencies, use any of the suitable techniques outlined in Module 5 to deal with them, reason and document your approach properly.
8. I expect you to **scan all numeric variables for outliers.** If there are outliers, use any of the suitable techniques outlined in Module 6 to deal with them, reason and document your approach properly.
9. I expect you to apply data transformations on **at least one of the variables.** The purpose of this transformation should be one of the following reasons: i) to change

the scale for better understanding of the variable, ii) to convert a non-linear relation into linear one, or iii) to decrease the skewness and convert the distribution into a normal distribution.

10. I expect you to use **only** readr, xlsx, readxl, foreign, gdata, rvest, dplyr, tidyr, deductive, deducorrect, editrules, validate, Hmisc, forecast, stringr, lubridate, car, outliers, MVN, infotheo, MASS, caret, MLR , ggplot2, knitr and base R functions for this section. You can also use your own functions. This will show your accumulated knowledge that you gained throughout the semester in this course.

Optional things that you can do to preprocess data

- You can subset your data by selecting variables and/or filtering in (or out) cases.
- Your data set can include date or string information or both. If this is the case, I expect you to apply required date conversions for dates and string manipulations for strings.
- Your data set can include variables that need to be separated or joined (i.e. the data may contain year, month and day information separately and you may join them to create a new variable).
- Depending on your level of knowledge gained in other courses (i.e. Introduction to Statistics and/or Machine Learning, etc) you may apply data normalisation, feature selection and feature extraction. Note that, this is an optional task and you don't have to apply any of these techniques if you don't know the theory and the fundamentals.
- **For those who are more advanced or daring, I am happy to listen your ideas! Provided that you get your data set and processing plan approved by me you can go creative and challenge yourself.**

Submission Instructions

The Assignment 3 report must be completed using the R Markdown template provided here:

R Markdown Template - Assignment 3

You must use the headings and chunks provided in the template, you may add additional sections and R chunks if you require. **In the report, all R chunks and outputs need to be visible. Failure to do so will result in a loss of marks.**

You must also publish your report to RPubs (see here) and **submit this RPubs link to the google form given below:**

RPubs link submission form

If you are working as a group, **ONLY ONE GROUP MEMBER** needs to complete this step. This online version of the report will be used for marking. Failure to submit your link will delay your feedback and risk late penalties.

Report Section Details

1. **Report title and group/individual details [Plain text]**: You can add the title of your report and student(s) details by updating the “title” and “author” entries at the top of the R Markdown Template.
2. **Required packages [R code]:** Provide the packages required to reproduce the report. Make sure you fulfilled the minimum requirement #10.
3. **Executive Summary [Plain text]**: In your own words, provide a brief summary of the preprocessing. Explain the steps that you have taken to preprocess your data. Write this section last after you have performed all data preprocessing. (Word count Max: 300 words)
4. **Data [Plain text & R code & Output]:** A clear description of data sets, their sources, and variable descriptions should be provided. In this section, you must also provide the R codes with outputs (head of data sets) that you used to import/read/scrape the data set. You need to fulfil the minimum requirement #1 and merge at least two data sets to create the one you are going to work on. In addition to the R codes and outputs, you need to explain the steps that you have taken.
5. **Understand [Plain text & R code & Output]:** Summarise the types of variables and data structures, check the attributes in the data and apply proper data type conversions. In addition to the R codes and outputs, explain briefly the steps that you have taken. In this section, show that you have fulfilled minimum requirements 2-4.
6. **Tidy & Manipulate Data I [Plain text & R code & Output]:** Check if the data conforms the tidy data principles. If your data is untidy, reshape your data into a tidy format (minimum requirement #5). In addition to the R codes and outputs, explain everything that you do in this step.
7. **Tidy & Manipulate Data II [Plain text & R code & Output]:** Create/mutate at least one variable from the existing variables (minimum requirement #6). In addition to the R codes and outputs, explain everything that you do in this step.
8. **Scan** I **[Plain text & R code & Output]:** Scan the data for missing values, inconsistencies and obvious errors. In this step, you should fulfil the minimum requirement #7. In addition to the R codes and outputs, explain your methodology (i.e. explain why you have chosen that methodology and the actions that you have taken to handle these values) and communicate your results clearly.
9. **Scan** II **[Plain text & R code & Output]:** Scan the numeric data for outliers. In this step, you should fulfil the minimum requirement #8. In addition to the R codes and outputs, explain your methodology (i.e. explain why you have chosen that methodology and the actions that you have taken to handle these values) and communicate your results clearly.
10. **Transform [Plain text & R code & Output]:** Apply an appropriate transformation for at least one of the variables. In addition to the R codes and outputs, explain everything that you do in this step. In this step, you should fulfil the minimum requirement #9.

NOTE: Follow the order outlined above in the report as possible as you can. Note that sometimes the order of the tasks may be different than the order given here. Any further or optional pre-processing tasks can be added to the template using an additional section in the R Markdown file. Make sure your code is visible (within the margin of the page). Do not use View() to show your data, instead give headers (using head())

This assignment is worth 25% and must be uploaded to the **Assignment 3 link** submission by **27/10/2019**. The report must be uploaded to Turnitin as a PDF with your code chunks and outputs showing. **YOU SHOULD ALSO PROVIDE THE RPUBS LINK OF YOUR REPORT. You can use this form to submit your Rpubs link.)**

Extensions will only be granted in accordance with the RMIT University Extension and Special Consideration Policy. No exceptions. Assignments submitted late will be penalised (see Course Information for further details).

Collaboration

You are permitted to discuss and collaborate on the assignment with your classmates. However, the write-up of the report must be an individual/group effort. Assignments will be submitted through Turnitin, so if you've copied from a fellow classmate/group, it will be detected. It is your responsibility to ensure you do not copy or do not allow another classmate/groups to copy your work. If plagiarism is detected, both the copier and the student/group copied from will be responsible. It is good practice to never share assignment files with other students/groups. You should ensure you understand your responsibilities by reading the RMIT University website on academic integrity. Ignorance is no excuse.

Learning Objectives Assessed

This assignment assesses the following Course Learning Objectives:

1. Critically reflect upon different data sources, types, formats and structures.
2. Apply data integration techniques to import and combine different sources of data.
3. Apply different data manipulation techniques to recode, filter, select, split, aggregate, and reshape the data into a format suitable for statistical analysis.
4. Justify data by detecting and handling missing values, outliers, inconsistencies and errors.
5. Demonstrate practical experience by having been exposed to real data problems.
6. Effectively use leading open source software for reproducible, automated data preprocessing.

Assignment 3 Marking Rubric

Criteria	Not acceptable (0)	Needs Improvement (3)	Excellent (5)
Executive Summary (5%)	No executive summary was provided.	The executive summary was provided but there was room for improvement.	A complete summary of the data preprocessing tasks was provided.
Data (10%)	No data source was given or the data didn't meet the minimum requirement #1, or the attempt to read/import/merge data sets were unsuccessful.	The data source was given but it was described poorly, or variable descriptions were missing or, the attempt to read/import/merge data sets were successful but there was room for improvement.	A complete and clear description of data sets, their sources, and variable descriptions were provided and data met the minimum requirement #1.
Understand (15%)	There was no attempt to inspect the data and the variables in the data set and unable to meet the minimum requirements #2-4.	There was an attempt to inspect the data and variables but it didn't meet the minimum requirements #2-4, or there was room for improvement.	A complete inspection of data and variables, inspection met the minimum requirements #2-4.
Tidy & Manipulate I (10%)	Unable to reflect on tidy data principles (minimum requirement #5)	The data set was untidy and there was an attempt to tidy/manipulate the data but it wasn't aligned with the tidy data principles or it was poorly described.	Able to reflect on the tidy data principles or a complete set of tasks were provided to tidy and manipulate the data properly.
Tidy & Manipulate II (5%)	Unable to create/mutate at least one variable from the existing variables (minimum requirement #6)	Able to create/mutate at least one variable from the existing variables but there was room for improvement or it was poorly described.	Able to create/mutate at least one variable from the existing variables and fulfil the (minimum requirement #6).
Scan I (20%)	Unable to scan for and deal with missing values, inconsistencies and obvious errors (minimum requirement #7). Some scripts were provided in an attempt to scan the data, however no methodology/actions were taken to handle those values.	Able to scan the data for missing values, inconsistencies and obvious errors, but the task needed improvements in the methodology. For example: - A methodology was applied to scan missing values, inconsistencies and obvious errors, however there was no attempt to check whether this approach can be safely applied, OR - A methodology was applied to scan missing values, inconsistencies and obvious errors, however the approach taken was not suitable/safe to apply, OR - The methodology was not explained enough, the results and outputs weren't presented in a clearer way.	A complete set of tasks were provided to scan the data for missing values, inconsistencies, and errors. -A safe and suitable methodology was followed to scan and deal with missing values, inconsistencies and obvious errors - The methodology taken was explained thoroughly. - The results and outputs were presented clearly.
Scan II (20%)	Unable to scan the data for outliers (minimum requirement #8). Some scripts were provided in an attempt to scan the data, however no methodology/actions were taken to handle those values.	Able to scan the data for outliers, but the task needed improvements in the methodology. For example: - A methodology was applied to scan and deal with outliers however there was no attempt to check whether this approach can be safely applied, OR	A complete set of tasks were provided to scan the data for outliers. - A safe and suitable methodology was followed to scan and deal with outliers

		- A methodology was applied to scan and deal with outliers however the approach taken was not suitable/safe to apply. - The methodology was not explained enough, the results and outputs weren't presented in a clearer way.	- The methodology taken was explained thoroughly. - The results and outputs were presented clearly.
Transform (10%)	Unable to apply an appropriate transformation for at least one of the variables (minimum requirement #9).	There was an attempt to apply a transformation to the data but it was poorly described OR there was room for improvement.	A complete set of tasks were provided to apply the transformation properly.
Succinct (5%)	The report was too long and/or lacked clarity.	The report could be written more succinctly. There was unnecessary detail that distracted from the main findings (like outputs were too long or there were unnecessary displays)	The report was written succinctly and clearly.