MATH2349 Data Preprocessing
Assignment 1

(Last Updated 18.02.2018)

Weight: 10%
Due date: 25 March, 2018, 23:59 AEST.
Length: Maximum 8 pages
Feedback mode: Feedback will be provided using Turnitin's inline marking tool and general comments.

This assignment requires you to locate an open data from the web, import it into R, and reflect upon the data types, formats and structures in your data set.

This assignment is worth 10% and is due **25/03/2018**. The instructions are as follows:

Assignment Instructions

1- Locate an open source of data from the web. This can be a tabular, spreadsheet data (i.e., .txt, .csv, .xls, .xlsx files), data sets from other statistical software (i.e., SPSS, SAS, Stata etc. data files), or you can scrape HTML table data.

Some sources for open data are provided below, but I encourage you to find others:

- http://www.abs.gov.au/
- https://www.data.vic.gov.au/
- http://www.bom.gov.au/
- https://www.kaggle.com

As a minimum, the data set should include:

- one numeric variable.
- one qualitative (categorical) variable.

There is no limit on the number of observations and number of variables. But keep in mind that when you have a very large data set, it will increase your reading time.

2- Read/Import the data into R, then save it as a data frame. You can use Base R functions or *readr*, *xlsx*, *readxl*, *foreign*, *rvest* packages for this purpose.

3- Inspect the data frame and variables using R functions. As a minimum, you should:

- check the dimensions of the data frame.

- check the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set.
- check the levels of factor variables, rename/rearrange them if required.
- check the column names in the data frame, rename them if required.

4- Subset the data frame using first 10 observations (include all variables). Then convert it to a matrix. If you get an error, explain why you got this error.

5- Subset the data frame including only first and the last variable in the data set, save it as an R object file (.RData).

Submission Instructions

The assignment 1 report must be completed using the R Markdown template provided here:

R Markdown Template - Assignment 1

Note that this is an R Markdown notebook template. Information for using the R Markdown package can be found here and here. The R Markdown template must be updated with your name and student number. You must use the headings and chunks provided in the template. You can add more chunks if required. You report will be composed of the following sections.

Report sections:

1. **Report title and student details [YAML input]**: You can add the title of your report (i.e. Assignment 1) and student number by updating the "title" and "author" entries in the YAML header (located at the top of the R Markdown Template).
2. **Data Description [Plain text]:** A clear description of data and its source (i.e. URL of the web site) should be provided.
3. **Read/Import Data [Plain text & R code & Output]:** Read/Import the data into R, then save it as a data frame. You can use Base R functions or *readr*, *xlsx*, *readxl*, *foreign*, *rvest* packages for this purpose. In this section, you must provide the R codes with outputs (i.e. head of data set) and explain everything that you do in order to import/read/scrape the data set.
4. **Inspect and Understand [Plain text & R code & Output]:** Summarise the types of variables and data structures, check the attributes in the data. Provide the R codes with outputs and explain everything that you do in this step.
5. **Subsetting I [Plain text & R code & Output]:** Subset the data frame using first 10 observations (include all variables). Then convert it to a matrix. Provide the R codes with outputs and explain everything that you do in this step. If you get an error, explain why you got this error.

6. **Subsetting II [Plain text & R code & Output]:** Subset the data frame including only first and the last variable in the data set, save it as an R object file (.RData). Provide the R codes with outputs and explain everything that you do in this step.

The report must be uploaded to Turnitin as a **PDF** with your code chunks showing. The easiest way to achieve this is to **Preview** your notebook in HTML (by clicking Preview) → **Open in Browser** (Chrome) → Right click on the report in Chrome → Click **Print** and Select the **Destination** Option to **Save as PDF**.

This assignment is worth 10% and must be uploaded to the Assignment 1 Turnitin link by **25/03/2018**.

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 effort. Assignments will be submitted through Turnitin, so if you've copied from a classmate, it will be detected. It is your responsibility to ensure you do not copy or do not allow another classmate to copy your work. If plagiarism is detected, both the copier and the student copied from will be responsible. It is good practice to never share assignment files with other students. 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.

Assignment 1 Marking Rubric

Criteria	Not acceptable (0)	Needs Improvement (1)	Excellent (2)
Locate data (10%)	No data source was given or the data didn't meet the minimum requirements.	The data source was given but it was described poorly.	A complete data source was provided and data met the minimum requirements.
Read/Import and save data (20%)	The attempt to read/import data set was unsuccessful.	The attempt to read/import data set was successful but unable to save the data in the correct format.	Able to read/import and save the data correctly.
Inspect data (30%)	There was no attempt to inspect the data and the variables in the data set.	There was an attempt to inspect the data and variables but it didn't meet the minimum requirements	A complete inspection of data and variables.
Subset and convert to a matrix (20%)	Unable to subset the data frame correctly.	Subsetting data frame was successful, but attempt to convert it to a matrix was missing or needed improvement.	A complete subsetting and data type conversion were provided.
Subset and save as an R object (20%)	Unable to subset the data frame and save it as an R object.	Able to subset the data frame correctly but failed to save it as an R object.	A complete subsetting and data conversion were provided.