Practical Assessment 2 Data Wrangling

	Assessment type:	Written report (PDF document) using R Markdown
	Due date:	26th May 2022, 5 pm Melbourne time
	Weighting:	30%
	Word limit:	Maximum 25 pages
	Feedback mode:	Feedback will be provided using Canvas marking tool and general text comments.

Group assessment

You will work on this assessment in a team of up to three students. You will select your own team member. Although you will work on the assessment together as a team, one of the team members will submit the report (PDF document) using R Markdown (otherwise there will be mixed up). Please write the details of your team members at the beginning of the report. If you prefer to work individually, that is also fine.

Purpose

The purpose of this assessment 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 a great way to show potential employers what you are capable of.

Overview

This assessment requires you to find some open data, and use your knowledge, skills gained during the course to preprocess the data. You will create a report using R Markdown to explain the steps taken by you in order to perform the data preprocessing tasks.

Assessment criteria and weighting

Please see the marking rubric to know the assessment criteria and weightage. Course learning outcomes This assessment is linked to the following course learning outcomes:

Course Learning outcomes

This assessment is linked to the following course learning outcomes:

- Accurately, logically and ethically combine data from multiple sources to make suitable for statistical analysis and draw valid interpretations.
- Articulate how data meets the best practice standards (e.g. tidy data principles).
- Select, perform and justify data validation processes for raw datasets.
- Use leading open source software (e.g. R) for reproducible, automated data processing.

Assignment data sources

Assessment 2 is open-ended however you are required to find suitable datasets that fulfil the minimum requirements given below. All the datasets that you use in this Assessment 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 the discussion board 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 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 assessment data (for example you can take crime statistics for the cities/states in Australia and merge this data set with cities/states' per capita income data).

2. Your data set should include **multiple data types** (numeric, character, factor, 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** (e.g., 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. **At least one of the data sets that you use should be Untidy**. You need to explain why the data set or data sets you used is/are 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 (e.g., the data may contain income and expense variables and you may create a savings variable out of the income and expense variables).

7. You are expected to **scan all variables for missing values, special values, and obvious errors (i.e., inconsistencies)**. If there are missing values, use any of the suitable techniques outlined in Module 5 to deal with them, reason and document your approach properly. **If there are no missing values in the data, then scan all variables for any special values and obvious errors**, use any of the suitable techniques outlined in Module 5 to deal with them, reason and document your approach properly.

8. You are expected **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. You are expected 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. The packages/functions 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 will be useful. 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. Please don't forget to put an explanation in your report if you do so.

● 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 as required.
● Depending on your level of knowledge gained in other courses (i.e., Applied Analytics 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.

Important Note:

Note that sometimes the order of the tasks may be different than the order given here. For example, you may need to tidy the data sets first to be able to create the common key to merge. Therefore, for such cases you may have a different ordering of the sections. 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).

Create the report using R Markdown

The assessment 2 report must be completed using the R Markdown template. Note that this is an R Markdown notebook template. Information for using the R Markdown package can be found Here. The R Markdown template must be updated with your name(s) and student number(s). You must use the headings and chunks provided in the template. You can add more chunks if required. Your report will be composed of the following sections. In the report, all R chunks and outputs need to be visible. Failure to do so will result in a loss of marks.

Sections of the report:

1. **Students' details [YAML input]:** Add students' full names, numbers, and the percentage of contributions in table "Group information". Add the leader's information (the one that submits assessment report) in "author" entries in the YAML header (located at the top of the R Markdown Template). If you work individually, then add only your own details in the table and write 100 % in percentage of contribution.

2. **Required packages [R code]:** Provide the packages required to reproduce the report.

3. **Executive Summary [Plain text]:** In your own words, provide a 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 (e.g., 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, briefly explain 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]:** Explain why your data (or one of the data sets) doesn't conform the tidy data principles (minimum requirement #5). Apply the required steps to reshape the data into a tidy format. 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, special values, and obvious errors (i.e., inconsistencies). 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.

Submission Format

● Upload the report as one single file (PDF) via the assessment 2 page in CANVAS.

● The easiest way to produce a PDF file from the RMarkdown is to Run all R chunks, then 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.

● After creating your PDF file make sure and check that your codes and outputs are visible.

Referencing guidelines

You must acknowledge all the sources of information you have used in your assessments. Refer to the RMIT Easy Cite Referencing Tool to see examples and tips on how to reference in the appropriate style. You can also refer to the Library Referencing Page for more tools such as
EndNote, referencing tutorials and referencing guides for printing. Use the RMIT Harvard referencing method for this assessment.

Collaboration

You are permitted to discuss and collaborate on the assessment with other groups. However, the write-up of the report must be with your own allocated group effort. Assignments will be submitted through Turnitin, so if you've copied from other groups, it will be detected. It is your responsibility to ensure you do not copy or do not allow another group to copy your work. If plagiarism is detected, both groups will be responsible. It is good practice to never share assessment files with others. You should ensure you understand your responsibilities by reading the RMIT University website on Academic Integrity. Ignorance is no excuse.

Academic integrity and plagiarism

Academic integrity is about the honest presentation of your academic work. It means acknowledging the work of others while developing your own insights, knowledge, and ideas. You should take extreme care that you have:

● Acknowledged words, data, diagrams, models, frameworks, and/or ideas of others you have quoted (i.e., directly copied), summarised, paraphrased, discussed, or mentioned in your assessment through the appropriate referencing methods.

● Provided a reference list of the publication details so your reader can locate the source if necessary. This includes material taken from internet sites.

If you do not acknowledge the sources of your material, you may be accused of plagiarism because you have passed off the work and ideas of another person, without appropriate referencing, as if they were your own.

RMIT University treats plagiarism as a very serious offense constituting misconduct. Plagiarism covers a variety of inappropriate behaviours, including:

● Failure to properly document a source.

● Copyright material from the internet or databases.

● Collusion between students. For further information on our policies and procedures, please refer to the University Website.

Assessment declaration

When you submit work electronically, you agree to the Assessment Declaration.

Extensions and special consideration

This course follows the RMIT University Assessment policy for extensions and special consideration. Information is available Here. Ensure you understand these guidelines before applying.
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 below for further details).

Late submission of assessment

Late submissions, without an approved extension or special consideration, will incur a penalty of 10% of the total mark per day for up to 5 days late (so the maximum late penalty is 50%). Submissions more than 5 days late are not accepted.

Penalty for exceeding maximum number of 25 Pages

A penalty of 5% of the total mark will be applied per each extra page.

Assessment2 marking rubric

Criteria	To meet all requirements and get the full point, you must complete the following criteria for each part
Executive Summary (5)	A complete summary of the data preprocessing tasks was provided
Data (10)	1. Complete & clear description of • data sets, • their sources, • variable descriptions were provided. 2. Data met the minimum requirement #1. 3. Merging data was correct (you may want to merge your data sets after tidying data set(s)). 4. R codes with outputs (head of data) were provided 5. Brief explanations of steps were given.
Understand (15)	1. Complete inspection of • data structure, • variables types, were done. 2. Attributes were checked & proper data type conversions were applied. 3. Inspection met the minimum requirements #2-4. 4. R codes with outputs were provided. 5. Brief explanations of steps were given.
Tidy & Manipulate 1(15)	1. Able to reflect on the tidy data principles. 2. Clear explanation was provided. 3. Complete set of tasks were provided to tidy and manipulate the data properly (you may want to tidy your data set(s) before merging them). 4. R codes with outputs were provided. 5. Brief explanations of steps were given.

Tidy & Manipulate 2(5)	1. Able to create/mutate at least one variable from the existing variables fulfilling the (minimum requirement #6). 2. R codes with outputs were provided. 3. Brief explanations of steps were given.
Scan I (20)	1. Complete set of tasks were provided to scan the data for missing values, special values and obvious errors (minimum requirement #7). 2. Safe and suitable methodology was followed to scan and deal with missing values, special values and obvious errors. 3. Methodology taken was explained thoroughly. 4. R codes were provided. 5. Results and outputs were presented clearly.
Scan II (20)	1. Complete sets of tasks were provided to scan the data for outliers. 2. Safe and suitable methodology was followed to scan and deal with outliers. 3. Methodology taken was explained thoroughly. 4. R codes were provided. 5. Results and outputs were presented clearly.
Transform (5)	1. Complete set of tasks were provided to apply the transformation properly, fulfilling requirement #9. 2. R codes with outputs were provided. 3. Brief explanations of steps were given.
Succinct (5)	The report was written succinctly and clearly.