UNIVERSITY OF AUCKLAND
Waipapa Taumata Rau
NEW ZEALAND

BUSINESS SCHOOL
DEPARTMENT OF INFORMATION SYSTEMS AND OPERATIONS MANAGEMENT

BUSAN 300: Data Wrangling

2022 Semester One

Wrangling Project Specifications

Contents

1. Purpose

The purpose of this project is for you to achieve all learning outcomes of the course. The process of working through it is more valuable than the mark received.

In attempting and completing this project correctly, you will:

- experience sourcing real, raw data
- refine your data wrangling skills
- encounter real data problems and employ the tools and techniques learned in this class and beyond to solve them
- improve your data-based problem-solving skills
- be patient when completing your work

2. Guidelines

2.1 Submission

Project Proposal: submit a single document in PDF format containing your project proposal to Canvas and have it approved by the stated deadline. Attach links to data files (e.g. Google Drive) when needed.

Project Report: submit a single document in PDF format containing all the contents of your project report to Canvas by the stated deadline. Attach links to data files (e.g. Google Drive) when needed.

2.2 Weighting

The project is worth **15%** of your final grade.

2.3 Academic honesty and integrity

This project is an individual assessment. You must complete all the work yourself. Do not submit work that you did not produce. Do not work in a way which could result in parties producing the same or very similar work.

In attempting this assignment you agree to adhere to all the principles and practices of academic honesty and integrity for the University of Auckland outlined here: https://www.auckland.ac.nz/en/about/learning-and-teaching/policies-guidelines-and-procedures/academic-integrity-info-for-students.html. Any form of cheating, plagiarism, assistance in cheating, unfair collaboration, or other behaviour deemed to be academic misconduct will not be tolerated. Academic misconduct will be dealt with according to the University's Student Academic Conduct Statute.

3. Tasks

This wrangling project is an exploration of public data with an intention to discover insights of interest to New Zealand or a global audience. The project is split into two parts: project proposal and project report. You should follow the steps described in the following sections in the order that they are listed.

3.1 Project Proposal

Have your project proposal approved by the stated deadline.

Each student's project should be sufficiently interesting (i.e. worth doing), doable, and be of similar complexity to other projects (to ensure fairness in marking and learning experience). To ensure your project meets this standard, please spend sufficient time to read the project requirements, explore potential datasets, and plan your project.

Summarise your plan in a project proposal document which briefly states:

1. The two datasets you will use in the project
    - **At least one dataset** must be from the list provided here. Use these datasets to form the basis for your project. We ask you to use at least one provided dataset to simulate a real working environment, where you may be tasked with analysing data on a particular initial topic and must then find additional resources to support your findings
    - Clearly state the sources: what are the data, their file types and where do they come from?
    - Provide links to the source of the files
2. How you plan to combine the two datasets
    - Consider what attribute(s) they have in common
    - Consider what technique(s) you plan to use
3. The intended final format of your combined dataset (e.g. Excel workbook, SQL database, MongoDB store, etc.)
4. A backup plan for completing this project
    - Consider what you could do to avoid forfeiting all 15% of your final grade due to poor forethought or underestimation of the requirements of this project etc.
    - You should enact this plan if you are unable to successfully combine your two datasets by Week 11

Copy and use the template provided **here** within a text editor and write no more than 300 words. Submit your proposal as a PDF file to the "Project Proposal" Canvas assignment and ensure it is approved by the stated deadline. You will receive feedback on your proposal. If your proposal is approved (marked as "complete" on Canvas) you should proceed with your proposed datasets. If your proposal is rejected (marked as "incomplete" on Canvas) you should meet with the teaching staff to discuss the reason for rejection, then revise and resubmit. You are encouraged to repeat this process until your proposal is approved, and hence it is recommended that you submit your initial proposal as early as possible.

There are no marks for completing this task. Your proposal must be approved for your project report to be accepted and marked.

3.2 Project Report: Source and audit two datasets

Source and audit two disparate datasets from separate sources. You will be required to combine the datasets into a single dataset, so ensure the chosen datasets are suitably related.

Remember that the datasets will be used to attempt to discover insights of interest to the New Zealand or the global public.

For learning purposes, your datasets must be:

- *sufficiently different* in format/file type from one another, and
- *sufficiently complex* - complexity can arise from
    - large data size (e.g. tens of thousands of rows/instances, or tens of columns)

- non-uniform data structures (e.g. the data is an amalgamation from multiple sources, or different attributes exist for different instances)
- dirty data

Summarise your sourcing and auditing activities in the report. See the requirements in the "Wrangling Details" section below.

3.3 Project Report: Pose questions

Pose **three** meaningful questions that could only be answered if the two datasets were combined. **These questions should be impossible to answer if the datasets were not combined**. You can assume any problem or situation/scenario under which these questions are posed. You will be required to *attempt* to answer the questions in your report, whether you obtain correct answers is not important as long as your attempt is sensible.

Document your questions in the report. See the requirements in the "Project Summary" section and "Questions and Answers" section below.

3.4 Project Report: Combine datasets to be stored in a single format

Wrangle your two datasets to clean and combine them into a single dataset. Store the combined dataset in a single data store (i.e. its "final storage format"). Data stores could be a file (e.g. Excel workbook) or a database (e.g. Microsoft SQL Server, Microsoft Access, SQLite, MongoDB). **You must not use any file converter or automation tool to transform data.**

If your datasets are too large for the tools used in this course to handle, you may use a sufficiently large subset of the data to build a proof-of-concept that the data can in fact be combined.

Document your wrangling processes in the report. See the requirements in the "Wrangling Details" section below.

3.5 Project Report: Answer questions

Using your cleaned and combined dataset in its "final storage format", **attempt** to answer the three questions you posed in any way you can. Answering is likely to involve the use of one or many of the following: PivotTables, XPath, MongoDB queries, visualisations, SQL queries etc.

You are advised to keep your exploration simple. Bear in mind that the learning outcomes of this assessment relate to data wrangling process and technique but not statistical analysis or data mining.

Document your answers in the report. See the requirements in the "Project Summary" section and "Questions and Answers" section below.

3.6 Project Report: Write report

Document the previous tasks in **one** report. Lay out your report in three sections in this order:

1) Project Summary
2) Wrangling Details
3) Questions and Answers

The requirements of each section are as the following.

3.6.1 Project Summary

Write this section for a business audience (e.g. a senior business decision maker).

Summarise your entire project. Include the three questions you posed, each with a short summary and discussion of your answer/conclusion for it. Provide some insight about your findings.

3.6.2 Wrangling Details

Write this section for a "data expert" audience (e.g. a classmate).

Detail the following wrangling processes in your report. Specifically:

For each dataset

- Its origin. Provide a direct link to the data, or if that is not possible, explain how you obtained the data
- Its general characteristics. What format is it in; what is the structure; how many columns/fields; how many records? etc.
- An initial audit. What observations did you make about the data? Are there any obvious or potential problems that may have to be dealt with before combining it with the other dataset?

Steps you performed to combine the datasets

- The level of detail should be so that any of your classmates can replicate the process
    - Stating what you did is more important than stating how you did it. For example, rather than explaining what menu items you clicked on, it is sufficient to state that you "removed duplicates on the user id and timestamp fields"
    - Include **relevant screenshots** of intermediate steps etc.
- State all transformations performed on each dataset individually before they were combined
- State all transformations performed to combine the datasets
- State the steps performed to store the combined dataset in its "final storage format"

Provide a link to your final combined data store so that it can be inspected. To do this, please upload your combined data store to Google Drive and set the viewing permissions to "Anyone with this link can view". Please provide the link to your data store clearly in your project report. Marks will be deducted if the final data store is not provided or is inaccessible.

3.6.3 Questions and Answers

Write this section for a "data expert" audience (e.g. a classmate).

Detail how you used your combined dataset, in its "final storage format", to reach the answers/conclusions to your questions. Specifically:

For each question

- State the question and your answer/conclusion
- State the steps performed to produce the answer. Include what tools/software you used and what queries you executed etc.
    - The goal here is not so that the process can be replicated, but to show your answer/conclusion is sound or otherwise reasonable
    - Even if no exact answer was found, the steps to reach that result still need to be documented. You should also state what would be needed to reach an answer

- Include relevant screenshots of intermediate steps etc.

Submit the report as a single PDF file to the "Project Report" Canvas assignment by the stated deadline.

4. Marking Guide

The experience of completing the project is intended to be more important than the results you produce.

No template for the report will be provided to not limit anyone's style and creativity. The content and structure you are required to provide are explicitly given in Section 3.

The table below summarises how your project will generally be marked. Where tasks are completed exceptionally well (demonstrated by its execution or documentation or both), marks for those tasks can compensate for tasks which are not completed as well.

Criterion and Marks (out of 15)
Project Summary: Appropriateness (2) • The summary is written for a general business audience • A business worker can read **only** this section and find out all they need to know about the project **Project Summary: Accuracy (2)** • The summary is **concise**, **descriptive**, **accurate**, and free from writing errors • The summary provides all the context necessary to understand and appreciate the project
Wrangling Details: Data Sourcing and Auditing (2) • The **origin**, **characteristics**, and initial **audit** is written for each dataset • The characteristics and audit of the datasets are accurate and meaningful **Wrangling Details: Data Combining and Storing (6)** • The documented process to transform and combine the data is **comprehensive**, **accurate**, **meaningful/replicable,** and free from writing errors
Questions and Answers: Answering (2) • The documented process to *attempt* to discover answers is **logical** and free from writing errors • The answers and resulting conclusions are sound and reasonable **Questions and Answers: Overall Considerations (1)** • The project has sufficient complexity and appeal • The work submitted is professional, including formatting, grammar etc.

5. Document Change History

v1.0 2022-01-11

- Initial release