CITS2402 Introduction to Data Science
Semester 2, 2024

Assignment

Assessed, worth 20%. Due: 11:59pm, Friday 4th October 2024

1 Aim

This assignment aims to investigate the similarities and differences between Australia and New Zealand regarding a demographic feature of your choice by comparing data from the latest available Australian Census (2021) and New Zealand Census (2023) [1].

You may choose what census topic you are most interested in from the data, provided it is not one we addressed in the lecture or lab case studies (such as age, number of children, or travel to work). For best results, you should also choose a topic with multiple categories (not, for example, binary categories). You may focus on particular categories of interest. You should use appropriate ways of visualising the data that best demonstrate the similarities and differences in your conclusions.

To focus the assignment, it may be helpful to frame it as a question you seek to answer. You should clearly state in your opening paragraphs what it is that you are seeking to answer.

When selecting your topic, choose data that lends well to in-depth analysis and graphical comparison. Simple comparisons, such as population size, are unlikely to achieve high marks. Review several alternatives before settling on your topic to ensure it offers rich data for analysis and meaningful insights. Choose a topic that interests you and allows for a comprehensive and engaging exploration using the available census data.

Here are some context **examples** to inspire your choice:

- Comparing urbanisation trends in both countries.
- Examining the impact of migration on cultural diversity in both countries.
- Analysing the relationship between education levels and life satisfaction in both countries.
- Comparing housing or health indicators and their impact on overall well-being in both countries.

It is important to understand from the beginning that, while programming is important, this assignment is not merely a "coding exercise". Equally significant are the context and rationale behind your work, the sound and replicable use of data, and the clear and compelling presentation of your results.

[1] You may choose a country other than NZ if there is one you are particularly interested in, providing the census data is publicly available and you link the source.

2 Learning outcomes

This assignment demonstrates competencies in:

- sourcing information from (authoritative) public data repositories.
- extracting and cleaning information needed to answer a question about the data.
- analysing and interpreting the data.
- visualising data to aid understanding and communicate results.
- writing comprehensive and informative scientific reports to communicate your findings.

3 Authorship

The assignment may be done individually or in groups of up to three students.

Each student's name and student number (whether completed by one or more students) must be provided in the declaration at the top of the assignment template `CITS2402-Assignment-template.ipynb`.

Where the assignment has been completed by more than one student, **only one copy should be provided** that it is clear which version to mark.

The suffix "-template" should be replaced with the corresponding student numbers. For instance, if you are doing your assignment with another person, you should rename your file as `CITS2402-Assignment-STDNO1-STDNO2.ipynb`, where 'STDNO1' and STDNO2' are the corresponding student numbers involved in the submission.

The submission must be the student's (s) own work. Any material used in the assignment from other sources must be clearly stated and referenced.

Your report, including all explanations and code, must be provided in a single notebook. The notebook should contain headings and explanations in markdown cells and executable code in Python code cells (as is done in the lab sheets).

It is recommended that you download a backup of your final completed submission directory for your own records.

4 Data

Finding the appropriate data is part of the exercise. You cannot expect that the two countries will provide the data in the same way. It is recommended that you begin searching for your data early on, starting with the Australian Bureau of Statistics (ABS) and Stats NZ - Census data. The metadata in the spreadsheets should be used to identify the relevant tables for your investigation. Only the individual tables in CSV format (not all the data) should be used in your submission.

Your report should clearly explain how you located the relevant data. This explanation should be detailed enough to allow the reader to replicate your steps and obtain their own raw data to test your code.

Your code should only need to access the file system for reading. You should avoid saving images, writing files, etc.

5 Submission

Your submission consists of:

1. The Python notebook `CITS2402-Assignment-STDNO1-STDNO2.ipynb`, including all explanations and code. It should contain headings and explanations in markdown cells and executable code in Python cells.
2. The PDF version of the Python notebook. Both files should be named following the instructions above.
3. Any data files that you use to run the code. Data files should not be more than 1MB. If you wish to include any images (not required), they should be no larger than 200KB.

Submit your files to LMS **as a ZIP file** before the due date and time. You can submit them multiple times. Only the latest version will be marked. Pay attention if you are submitting all requested files when making a new submission. Your submission will follow the rules provided in LMS.

Before submitting your assignment, you must ensure it runs without errors in the Google Colab environment. This is to avoid markers having local problems with libraries you may choose to use. Therefore, your code should run seamlessly without having to install any package on Google Colab. The Colab environment has embedded essential data science libraries (numpy, pandas, matplotlib, etc). Your mark will be zero if they cannot run your code in Google Colab.

Important:

- You must submit your assignment as .IPYNB ***and*** as an electronic file in PDF format (do not send DOCX, or any other file format). Only PDF format and .IPYNB is accepted, and any other file formats will receive a zero mark.
- Failing to submit any of the required files will result in a zero mark. You should include all data files you are using.
- The data files you are using should be clearly specified. The marker should be able to download the same data files from the website(s) and run your analysis.
- You should provide comments on your code.
- By submitting your assignment, you acknowledge you have read all instructions provided in this document and LMS.

- There is a section in your LMS, Assignment - Updates and Clarifications, where you will find updates or clarifications about the tasks when necessary. It is your responsibility to check this page regularly.
- You will be assessed on your thinking and process, not only on your results. You should demonstrate you understand the concepts involved.
- Your answer must be concise. You will be graded on thoughtfulness. If you are writing long answers, rethink what you are doing. Probably, it is the wrong path.
- You can ask in the lab or during consultation if you need clarification about the assignment.
- You should be aware that some algorithms can take a while to run. A good approach to improving the Python speed is using the vectorised forms discussed in class. In this case, it is strongly recommended that you start your assignment soon to accommodate the computational time.

6 Code

The code will be executed with a fresh kernel for marking, so (as usual) you should ensure it runs with a clean kernel before submission.

Any supporting data must be in the same directory. Data files should not be more than 1MB.

It is recommended that development is done in the notebooks - in the past students have had code fail due to pasting from other environments.

7 Rubric

The assignment questions involve analysing data. You should present a well-organised report and aim to write concisely.

The assignment will be marked for clarity and professionalism of both the exposition and the coding.

Please read the assignment instructions and rubric carefully when preparing your code and report.

It is recommended that you structure the report in a way that is consistent with the data science lifecycle. You should use headings to help structure your report.

Consider these aspects as examples of what is expected (see rubric for more details):

- Plots and figures: Your report should include appropriate and well-presented visualisations that are meaningful to the analysis. All your diagrams/plots should have proper titles, axis labels, values, etc., to help the reader understand what you are plotting.

- Your report should include the correct use of the data science lifecycle steps and the interpretation of the results obtained from these steps.

- Presentation of results: Describe the results of your analysis and their interpretations. **Software output is not a valid output**. You must format and present your answer appropriately (tables, graphs, etc.). You should not add irrelevant information when presenting the results.

- Discussion: Based on your results, describe the conclusions of your analysis.

- Overall: Do not report out a long list of numbers if you do not explain what they are and what they mean. Do not generate plots after plots without explaining what they were doing or their purposes. Instead, think about the best way to display/illustrate your approach. Pay attention to all details, including the graph labels, presentation quality, and clarity.

Markers will pay attention to the following components (roughly equal weighting).

Context and data:

- Adequate context has been provided to understand the question and why it is important.

- The question you seek to understand is clearly stated.

- It is clear what data is used and its provenance. Instructions allow the reader to easily source the data (to make the work replicable).

- Complete context and information about the data (e.g. the unit of measure, description of the categories in the topic, etc) are provided.

- Relevant differences between the data from different sources and assumptions you have to make for comparison are clearly described.

- If you are extracting only part of the data, your code should be accompanied by a brief description of what you are extracting and why.

Data lifecycle, structure, and presentation:

Your data cleaning steps should be accompanied by a brief description of any steps you took to transform the data from its raw form into usable form.

- The route from the data to the results is clearly set out, and steps are explained.

- It is clear what format the raw data took, what is extracted and why.

- Any data cleaning and conversion is clearly and concisely outlined.

- The processing or analysis necessary to extract and compile the results is clearly explained.

Results, visualisation and conclusion:

- The results are clearly stated and connected back to the original data and assumptions.
- Appropriate and informative choices are made for visualisation(s) (plots).
- The visualisations are clearly and professionally presented.
- Conclusions are connected to relevant features of the visualisations.

Coding:

There is no single “right” way to write the code. However, the following should be considered:

- Code is clear and easy to read and comprehend. Considerations should include the use of meaningful variable names, the use of comments and/or docstrings for key steps/blocks (you do not need to comment every line; this tends to obscure the key steps), and the use of functions.
- Code is appropriately *concise*. Code should not be pared down to a bare minimum at the expense of clarity and readability. However, you should try to avoid unnecessary extraneous code.
- Code is reasonably efficient. (It is not necessary to achieve ultimate efficiency at the expense of writing clear, logical code. However, you should avoid obvious unnecessary inefficiencies.)
- Code is well *structured*. Functional decomposition is used to separate tasks into meaningful components.
- Ensure you avoid repeating unnecessary blocks of code by using functions. Additionally, refrain from hard-coding values directly within functions; instead, use function arguments to pass these values.

Professionalism and Challenge:

- Overall, the report forms a compelling and illuminating narrative.
- The report is not unnecessarily long or repetitive and provides all the information completely but concisely.
- The report reveals aspects of the data that are not trivially obvious.
- The report is of a quality that an employer would be comfortable showing to a client.

8 Plagiarism and penalty on late submissions

See the URL below about late submission of assignments:

https://ipoint.uwa.edu.au/app/answers/detail/a_id/2711/~/consequences-for-late-assignment-submission

Plagiarism: In accordance with University Policy, you certify that all work submitted for this assignment is your own and that all material drawn from other sources has been fully acknowledged.