School of Computing and Information Systems
The University of Melbourne
COMP20008 - Elements of Data Processing, Semester 2, 2024

Assignment 2 – Victorian Vibes

Release:	Friday 6 September 2024
Due:	- *Group contract*: Friday 13 September at 5 PM - *Code and Report submission*: Monday, 7 October at 5 PM - *Slides submission*: Friday, 11 October at 5 PM - *Oral presentation*: Week 12 (14 to 18 Oct) - *Peer-Review:* Friday 18 October at 5 PM
Marks:	The Project will be marked out of 35 and will contribute 35% of your total mark.
Groups:	You should work in groups of 3 or 4
Main Contact:	Hasti Samadi (hasti.samadi@unimelb.edu.au)

1. Overview

In this project, you will build upon the dataset from Assignment 1, incorporating additional datasets that provide valuable insights into various suburbs and regions across Victoria. Your primary objective is to analyse these datasets to uncover trends, patterns, and key insights that can inform local government authorities in their planning efforts. By identifying critical areas for support, investment opportunities, and the unique challenges faced by each region, suburb, or Local Government Area (LGA), your analysis will contribute to shaping effective decision-making.

This assessment allows you to develop hands-on experience in data wrangling, processing, and analysis within an open-ended context.

You will produce both a written technical report and a presentation, tailored to a local government audience, summarizing your findings. The report should be accessible to readers with a moderate understanding of data analysis while clearly conveying your insights into the landscape of local government planning in Victoria.

2. Assignment Structure

Group Contract (Due: Friday 13 October at 5 PM)

You must submit a group contract outlining your team's goals, expectations, and policies for working on the project. A *group contract template* is provided. You are welcome to work with the provided template or customize it according to your preference. *Submit as a single PDF file via Canvas* *(Assignment 2: Group Contract).*

You may change your group contract throughout the semester, but proposed changes should be agreed to by all members. There are no marks directly allocated to the content of the Group Contract, but we may refer to it when assessing the relative contribution of each group member to resolve any dispute.

Failure to submit your Group Contract by the mentioned deadline will result in a deduction of 2 marks from the final assignment mark for all members of the group.

Code and Report Submission – 20 marks (Monday, 7 October at 5 PM)

1. **Report**: Your report should consist of ten to twelve single-column A4 pages. Maintain a line spacing of exactly 1 with normal margins and ensure that the font size is 11pt or above. Please note that if your report exceeds twelve pages (excluding the cover page), only the content within the first 12 pages will be reviewed and assessed. Any additional pages will be disregarded. The page limit includes all the text including references, captions, and any table or image. Tables and image content should be readable and sensible in size.
   The group name W[XX]G[X] and all group members' names should appear on the first page after the title of the report. *Submit as a single PDF file through Canvas/Turnitin (Assignment 2: Written Report)*
2. **Code**: One or more programs, written in Python, including all the code necessary to reproduce the results in your report (model implementation, data processing, visualization, and evaluation). Your code should be executable and have enough comments to make it understandable. You should also include a README file that briefly details your implementation and describes how to run your code to reproduce the results in the report. *Submit as a single zip file through Canvas/Turnitin (Assignment 2: Code and Comments).*

Slides Submission (Due: Friday, 11 October at 5 PM)

Submit the slides you plan to use for your oral presentation. These slides should effectively illustrate the key insights from your data analysis task. The slides must be submitted as a single PowerPoint (.pptx) or PDF file via Canvas/Turnitin (Assignment 2: Oral Presentation Slides). No other formats will be accepted.
Remember, you are required to use the exact slides you submit during your presentation.

Oral Presentation and Assessment – 13 marks (Due: from Monday 14 to Friday 18 October)

During week 12 of the semester (14-18 Oct) all teams should deliver an Oral Presentation of their work and findings for assignment 2 and answer some Oral Assessment questions. Some of the presentations will be conducted in the students' usual workshop room and some in other venues which will be announced shortly. Two markers will assess the oral presentations. See section 6 for more details.

Teamwork Evaluation – 2 marks (Due: Friday 24 May at 5 PM)

For this part of the assessment, every team member needs to evaluate both their own contributions to the assignment and the contributions of their teammates. This evaluation should align with the expectations you set in your submitted "group contract".

The evaluation should be delivered via Feedback Fruits available on Canvas (Assignment 2: Teamwork Evaluation).

Your group members' evaluations will determine individual group member evaluation scores worth 2 marks. If any member is identified as a non-contributor, these scores may be used to adjust those individual's marks for the report and oral presentation (worth 30 marks).

3. Data Sets

The provided files contain detailed community-level data for Victoria, covering demographic, health, and geographic information. Additionally, the datasets include data on the region's housing, expenditure on Electronic Gambling Machines (EGM), and offence record rates. This information is distributed across Four distinct files, offering a comprehensive view of various community aspects.

- 'Communities.csv' dataset comprises information on 1,080 communities and 226 columns with various details offering a comprehensive overview of community characteristics.
    - Geographic Information: region, location, distance from CBD and Distance to nearest public hospital
    - Demographic Information: density, population, cultural background, language diversity, education level, economic information, Housing and more
    - Health Information: Public Hospitals, Private Hospitals, Community Health Centres, Allied Health, emergency response information and more
- 'Houses_by_suburb.csv' dataset comprises the percentage shift in median prices per annum from 2013 to 2023. It includes the details of 790 regions across Victoria and 18 columns. You can add more columns if it helps your analysis.
    - The year columns contain data about house prices across different suburbs over the years.
    - Changes Percentage: difference in house prices between 2013 and 2023
- 'LGA_EGM.csv' dataset includes the total gaming expenditure on 'Electronic Gambling Machines (EGMs) for each 'Local Government Area (LGA)' in Victoria. It includes the details of 57 LGA across Victoria and 12 columns. You can add more columns if it helps your analysis.
- 'LGA_Offences.xlsx' dataset includes 6 separate tables:
    - Table1: Offences by Region and LGA
    - Table2: Offences by Type, LGA and Police Area

- Table3: Offences by Type, LGA and Suburb
- Table4: Offences by Location Type and LGA
- Table5: Offences by Investigation Status and LGA
- Table6: Drug Offences by Type and LGA

You do not need to use all datasets. Datasets you use will depend on your research question and the analysis approach your group agrees on. Details about using these files are provided in the README file.

4. Data Analysis Tasks

4.1. Research Question

The research question clarifies the purpose of your analysis. It identifies the problem or question being addressed, sets the context, and explains why the analysis is being conducted.

In your report, you must introduce (at least) ONE research question clearly and explicitly. We have presented a few examples of possible research questions in the accompanying video to provide you with some inspiration. However, each team needs to independently formulate their own research question based on the provided dataset.

While the possibility exists to explore more than one research question, it's important to note that the pursuit of several questions is not necessarily desirable or likely to lead to greater marks (i.e. full marks are obtainable for one well-studied research question). We will primarily evaluate the quality of your work by assessing the depth of your analysis, and the insights it yields, rather than simply covering a larger quantity of content or material.

4.2. Data Pre-processing

So far in the subject, you've learned various ways to prepare and organize data. These include techniques like filling in missing data (data imputation), reshaping data (data manipulation), adjusting data ranges (scaling) and grouping data into categories (discretizing). You've also explored methods to simplify complex data (dimensionality reduction) and handle text data (text processing) using tools like text vectorization and regular expressions.

For this project, you're encouraged to consider applying any of these methods to the provided datasets. Your objective is to implement a minimum of ONE data pre-processing technique, though you're welcome to utilize as many data pre-processing techniques as you see fit. The methods you select should logically support the research question(s) you have picked, and in your report and presentation, you should explain the reason for your selection of each method.

In your report and presentation, ensure you provide justifications and explanations for all methods you employ (for both pre-processing and supervised/unsupervised models). Present the results and highlight any interesting discoveries.

Remember, there's no single expected solution here. The more deeply you engage with and understand your data, the better set-up you will be for subsequent stages of your project.

4.3. Data Analysis

The data analysis and visualization phase are critical for converting the pre-processed data into valuable insights that can guide local government authorities in decision-making. This phase involves a thorough exploration of the datasets, revealing trends, patterns, and relationships that address the research question and provide actionable recommendations. In this section, we will expand on various analytical methods and visualization techniques, detailing how they can be applied to support planning, resource allocation, and problem-solving across regions, suburbs, and LGAs.

Visualization is an essential part of the data analysis process, as it allows complex information to be presented in an easily understandable form. Various tools and techniques can be used to create compelling visual narratives. You are encouraged to use a variety of analytical and visualization tools such as exploratory data analysis, correlation analysis and trend analysis.

- Exploratory Data Analysis (EDA) is the initial step in the data analysis process. It helps in understanding the structure of the data, identifying potential relationships, and uncovering hidden patterns. EDA provides a foundation for more advanced analyses by summarizing the data and offering a visual representation of its key characteristics. You can use descriptive statistics such as central tendency measures (mean, median), standard deviation, ranges and frequency distribution as well as visualisations such as histograms, boxplots, pie charts, bar charts and tables to conduct this analysis.
- Correlation analysis investigates the relationships between two or more variables, helping to identify which factors influence or are related to each other. You can use methods such as the Pearson Correlation Coefficient and visualisation tools such as heatmaps and scatter plots to conduct this analysis. You are also highly encouraged to use your own intuition and understanding the perform a deeper and more intuitive analysis in this section.
- Trend analysis is employed to study changes in variables over time. By understanding how certain factors evolve, you can anticipate changes and provide recommendations more effectively. You can use tools such as line charts and area charts to facilitate your analysis.

You are expected to develop a minimum of TWO analytic descriptions for your findings and highlight critical findings that point to a specific trend or understanding using the information from your visualisations.

4.4. Use of Supervised and Unsupervised Models

In this subject, we explore certain Machine Learning related techniques. These include predicting outcomes based on known data (supervised models like Decision trees and linear regression) and finding clusters and patterns in data without prior labels (unsupervised methods like k-means and agglomerative clustering). Many other techniques are possible too.

Feel free to choose any Machine Learning method(s) that are suitable for answering your research question. Your choices should be substantiated and clarified in both your report and presentation. The objective is to implement a minimum of TWO Machine Learning techniques, though you're welcome to utilise more if you so choose. You might opt to employ two supervised models, two unsupervised methods, or one of each.

In your report and presentation, it's important to articulate your rationale behind the machine learning methods you chose. Provide a concise overview of your approach and outline how you assessed the effectiveness of your chosen methods. Equally important is your interpretation of the results and their implications.

NOTE: You are welcome (and indeed strongly encouraged) to make use of any relevant existing Python libraries (such as *sklearn* or *scipy*) in your work on this assignment.

5. Report

Your primary submission for this assignment is your report. The report should follow the structure of a technical paper. It should describe your approach and observations, both in data preparation, data analysis and the machine learning algorithms you tried. Its main aim is to provide the reader with knowledge about the problem, in particular critical analysis of your results and discoveries.

The following is the expected structure of the report for this assignment.

- **Executive Summary:** A concise overview of the entire report, summarizing the objectives, methods used, key findings, and recommendations. This section provides a high-level snapshot of your work and its outcomes.
- **Introduction**: Introduce the purpose of the report, the problem or research question being addressed, and the data sources used. Set the context for your analysis and explain the significance of the study.
- **Methodology**: Offer a detailed explanation of the methods, techniques, and tools used for data preparation, analysis, and interpretation. Assume the reader has a basic understanding of technical terms, so focus on the specifics of your approach.
- **Data Exploration and Preprocessing**: Present the results from your data preprocessing and data analysis. Describe any preprocessing techniques applied to clean and prepare the data. This section should include descriptive statistics, visualizations, and insights gained from exploring the data and analysing the trends and correlations in the provided datasets. Use charts, graphs, and tables to illustrate patterns, trends, and relationships.
- **Data Modelling:** In this section, you need to explain your rationale for selecting the supervised (e.g., regression, classification) or unsupervised (e.g., clustering) models used. Describe how the models were implemented, the key features selected, and any preprocessing tailored for them. Specify the evaluation metrics (e.g., accuracy, precision,

recall) and justify why they were chosen. Discuss the results in relation to the research question.

- **Discussion and Interpretation:** Summarize the most important insights obtained from your data analysis and data segmentation and/or predictive modelling. Focus on answering the main questions or addressing the problem you have introduced in the introduction. Provide an in-depth interpretation of your findings. Bullet points or numbered lists can help highlight these findings. Explain the significance of the patterns observed. Explain why these findings are interesting and valuable. Discuss any unexpected or interesting insights that emerged. (This is the most important section of your report)

  Remember we are more interested in seeing evidence that you have thought about the task and can identify reasons behind your different results in different experiments. You should think beyond simple numbers to the reasons that underlie them and connect them back to your research question. You can also add complementary experiments and their results in this section.
- **Limitations and improvement opportunities:** Address the limitations of the analysis, such as data constraints, potential biases, or assumptions made. Suggest potential improvements or alternative approaches for future analyses.
- **Conclusion:** Recap the key findings and recommendations from the report. Emphasize the overall value and potential impact of your analysis, linking back to the objectives outlined in the introduction.
- **References:** List any sources, references, or citations used in the report, especially if you've drawn upon external research or literature to inform your analysis.

We've supplied a template for the report via the assignment page. You are welcome to work with the provided template or customize it according to your preference.

6. Oral Presentation and Assessment

6.1. Oral Presentation

You need to conduct an oral presentation explaining what you have done for assignment 2. Your presentation should encompass the key components below:

1. *Introduction of Research Question*: Begin by introducing the research question that guided your assignment. Explain briefly why it is relevant to the managers of the bookstore.
2. *Methods, Techniques, and Tools*: Elaborate on the methods, techniques, and tools you employed for both data preparation and data analysis. Explain how you gathered, cleaned, and structured the data, as well as the analytical techniques and machine learning techniques you utilized.
3. *Presentation of Results*: Share the outcomes derived from your data analysis. Provide a concise overview of the insights you gained through your analytical process.

4. *List of Findings and In-Depth Interpretation*: Present a list of the findings from your analysis. Then provide an interpretation of these findings, shedding light on the significance and implications they hold in relation to your research question.
5. *Limitations and Improvement Opportunities*: Address the limitations encountered during your study, discussing any constraints or challenges that might have influenced the results. Furthermore, demonstrates suggested potential areas for improvement and development.

The presentation requirements are as follows:

- **Timing**: Your presentation should take exactly **7 minutes**. If your presentation doesn't finish on time the markers will interrupt and stop you and it will also negatively impact your mark. There will be a further **15 minutes** of questions from the markers.
- **Presenters**: Attendance at the presentation is mandatory for all team members unless they have been granted an exemption by the teaching staff. Each member of the group is expected to contribute to the presentation content.
- **Slides**: To ensure fairness for all groups and prevent last-minute modifications based on other teams' work, when presenting you will be asked to use the exact version of the slides that you submitted to Canvas.

6.2. Oral Assessment

After the presentation, there will be an oral assessment of all team members' knowledge of the assignment. During this Q&A session, each member will be evaluated individually. Markers will ask questions about the **entire** report, rather than focusing on your specific sections. All members are required to respond independently to oral questions regarding both the report and the presentation. Our findings from the oral assessment can impact your report marks.

7. Teamwork

As mentioned previously, 2 marks for this assignment are determined by the results of your teamwork evaluation task. However, based on these assessments and past records, we will identify any non-contributing members and adjust the overall assignment grade accordingly.

The group contract outlines the expectations and responsibilities of each group member. It's crucial that every member actively participates in this assignment. Remember, your comprehension of the entire project will be assessed during the oral evaluation.

If you encounter any challenges with inactive team members who aren't responsive to your inquiries, please reach out to Hasti for assistance in finding a solution.

8. Assessment Criteria

The report will be marked according to the rubric published via the assignment page. The oral presentations and oral assessments will also be marked according to their published rubric.

Although your code is not assessed directly, you must submit the code that produced the results presented in your report. If you do not submit executable code that supports your findings, we reserve the right to give your team **zero** marks for the report section.

9. Terms and Conditions

9.1 Changes/Updates to the Assignment Specifications

We will use Canvas to advertise any (hopefully small-scale) changes or clarifications in the assignment specifications. Any addendums made to the assignment specifications via Canvas will supersede the information contained in this version of the specifications.

It is your responsibility to ensure you are adhering to the latest iteration of these specifications should updates be announced.

9.2 Late Submissions

There will be no extensions granted, and no late submissions allowed to ensure a smooth run of the oral presentations.

For students who are demonstrably unable to submit in time, we may be able to offer alternative arrangements, but these could involve not being able to complete the oral presentation component, with the associated work being reweighted. The arrangement will be sought on a case-by-case basis. Please email Hasti (hasti.samadi@unimelb.edu.au) with documentation of the reasons for the delay.

9.3 Academic Honesty

While it is acceptable to discuss the assignment with others in general terms, excessive collaboration with students outside of your group is considered cheating. Your submissions will be examined for originality and will invoke the University’s Academic Misconduct Policy where either an inappropriate level of collaboration or plagiarism appears to have taken place.

We highly recommend (re)taking the academic honesty training module in this subject's Canvas. We will be checking submissions for originality and will invoke the University's Academic Misconduct policy where inappropriate levels of collusion or plagiarism appear to have taken place. Content produced by generative AI (including, but not limited to, ChatGPT) is not your own work, and submitting such content will be treated as a case of academic misconduct, in line with the University's academic integrity policy and specifically recent guidance on the use of ChatGPT and other Large Language Models in student work.

9.4 Data Acknowledgement

The data used in this assignment is extracted from the datasets provided on the DATA VIC [1] portal under the Creative Commons CC0 license.

[1] https://www.data.vic.gov.au/