# School of Computing and Information Systems
# The University of Melbourne
# COMP20008 - Elements of Data Processing, Semester 1, 2024

### Assignment 2 – Who else likes this book?

| | |
|---|---|
| Release: | Friday 19 Apr 2024 at 9 AM |
| Due: | • *Group contract*: Friday 26 Apr at 5 PM<br>• *Code and Report submission*: Friday, 10 May at 5 PM<br>• *Slides submission*: Monday, 13 May at 9 AM<br>• *Oral presentation*: Week 11<br>• *Peer-Review:* Friday 24 May 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 2 or 3 |
| Main Contact: | Hasti Samadi (hasti.samadi@unimelb.edu.au) |

## 1. Overview

In this project, you will undertake an analysis of a collection of datasets containing detailed information about books and their reviews by users of an online bookstore. Your overall objective is to analyse the data and extract insights. These insights are intended to help managers of the bookstore decide which kinds of books they should buy (and not buy) in the future for best sales, and which books they should recommend to buyers as possible additional purchases. The outcomes of your analysis will be communicated through both a presentation and a written technical report targeted towards a managerial audience.

This assessment presents an opportunity for you to gain experience in data wrangling, processing and analysis for an open-ended task.

You will deliver a brief technical report summarising your findings which should be comprehensible to a reader with a reasonable level of understanding of data analysis. Through this report, you will communicate your insights and discoveries on the landscape of book reviews.

## 2. Assignment Structure

- **Group Contract – 2 marks (Due: Friday 26 Apr 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 vary 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.

**Code and Report Submission – 22 marks (Due: Friday, 10 May 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, only the content within the first 12 pages will be reviewed and assessed. Any additional pages will be disregarded. Conversely, submitting a report shorter than eight pages will result in a penalty. 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: Group 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: Monday, 13 May at 9 AM)**

You will need to submit the slides you are going to use for delivering your oral presentation. These slides should illustrate your insights derived from the data analysis task you've undertaken. *Submit as a single PowerPoint (.pptx) or PDF file through Canvas/Turnitin. (Assignment 2: Oral Presentation Slides) No other format is acceptable.*

You will be required to use the exact slides that you have submitted for your presentation.

**Oral Presentation and assessment – 8 marks (Due: from Monday 13 May to Friday 17 May)**

During week 11 all teams should deliver an oral presentation of their work and findings for assignment 2. 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 Fruit 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

### 3.1 Main Data sets

The provided files contain data regarding various books, users, and their corresponding book ratings. You will find this information distributed across three distinct CSV files.

- ‘BX-Books.csv’ dataset comprises information on 18,185 books, including their International Standard Book Number (ISBN), Title, Author, Year of publication and Publisher.
- ‘BX-Users.csv’ dataset comprises anonymised information on 48,299 users of the online bookstore including their ID, City, State, Country and Age.
- ‘Bx-Ratings.csv’ dataset includes the reviews of the provided users on the given books. The columns include the user ID, book ISBN and the rating associated with that review.

What datasets you use will depend on your research question and the analysis approach your group agrees on. Details about using these text features are provided in the README file.

### 3.2 Recommendation Data Sets

Considering the nature of these files, there is an opportunity to develop a recommendation system capable of predicting the ratings that users might assign to new books. While incorporating this into your research question is an optional challenge, groups opting to implement the recommendation system can substitute it with the two supervised or unsupervised models outlined in section 4.3.

To assist you with implementing a recommendation system we have provided three separate CSV files:

- ‘BX-NewBooks.csv’ dataset information on 8,924 new books, including their ISBN, Title, Author, Year of publication and Publisher.
- ‘BX-NewBooks-Users.csv’ dataset comprises information on 8,520 users of the online bookstore including their ID, City, State, Country and Age. These users are not new and they have a history of rating books in the system. Your goal can be to predict the ratings that these users can provide for the books in the ‘BX-New-Books.csv’ dataset.
- ‘BX-NewBooks-Ratings.csv’ dataset contains the real ratings provided by users for the new books listed in the 'BX-NewBooks.csv' dataset, which are associated with the users' information in the 'BX-NewBooks-Users.csv' dataset. You can utilize this information to compare the predicted ratings generated by your recommendation systems against the actual ratings provided by users, allowing for comprehensive evaluation and validation of your models.

Please keep in mind that if you are not implementing a recommendation system you are not allowed to use these datasets.

## 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, it is essential to 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), converting data (encoding), 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 TF-IDF.

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 three data pre-processing techniques, 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. It would be good if you also describe the importance (effect) of these discoveries in terms of sales.

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. Use of supervised and unsupervised models

In this subject, we explore certain Machine Learning related techniques. These include identifying relationships between variables (correlation), predicting outcomes based on known data (supervised models like Decision trees and linear regression), and finding 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, or two unsupervised methods, or one of each. As highlighted earlier, you have the flexibility to incorporate a recommendation system as your machine learning model implementation. Implementing a recommendation system will satisfy the minimum expectations of section 4.3.

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, 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 what you have done.
- **Introduction**: This section introduces the purpose of the report, the problem or question being addressed, and introduces the data sources used. It sets the context and explains why the analysis was conducted.
- **Methodology**: Detailed explanation of the methods, techniques, and tools employed for data preparation, analysis, and interpretation. When writing this section, you can assume that the reader is familiar with the technical terms.
- **Data Exploration and Analysis:** Present the results of your data analysis. This section may include descriptive statistics, visualizations, and insights gained from exploring the data. Use charts, graphs, and tables to illustrate patterns, trends, and relationships.
- **Results**: Summarize the most important insights obtained from the supervised and/or unsupervised learning models you have used. Focus on answering the main questions or

addressing the problem you have introduced in the introduction. Present the results, in terms of evaluation metric(s) and, ideally, illustrative examples and diagrams.

- **Discussion and Interpretation:** Provide a list of interesting findings and an in-depth interpretation of them. 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. Explain what needs to be done to improve your analysis.
- **Conclusion:** Summarize the main points of the report and reiterate the key findings and recommendations. Emphasise the value and potential impact of the analysis.
- **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

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 **9 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 may be a further **10 minutes** of questions and answers 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.

### 6.1. 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. Tutors 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 have to 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 this[1] Kaggle page under the Creative Commons CC0 license.

[1] https://www.kaggle.com/datasets/ruchi798/bookcrossing-dataset