School of Computing and Information Systems
The University of Melbourne
COMP20008 - Elements of Data Processing, Semester 2, 2023

Assignment 2 - Netflix Unleashed: Exploring the Magic in Data

Release:	Monday 3 Sep 2023
Due:	• *Group Registration:* Friday 8 Sep at 5 PM • *Group contract*: Friday 15 Sep at 5 PM • *Code and Report submission*: Friday, 6 Oct at 5 PM • *Slides submission*: Monday, 9 Oct at 9 AM • *Oral presentation*: Week 11 • *Peer-Review:* Friday 13 Oct at 5 PM
Marks:	The Project will be marked out of 40 and will contribute 40% of your total mark.
Groups:	You may choose to form a group of 4
Main Contact:	Hasti Samadi (hasti.samadi@unimelb.edu.au)

1. Overview

In this project you will aim to undertake a comprehensive analysis of a collection of datasets containing detailed information about movies and TV shows available on the Netflix platform. Your primary objective is to meticulously examine the provided data to extract intriguing insights. The outcomes of your analysis will be communicated through both a presentation and a well-structured written technical report targeted towards a managerial audience.

This assessment presents a valuable opportunity for you to immerse yourself in the realm of data analysis within the framework of an open-ended research challenge. It serves as a platform for improving your data analysis skills and cultivating your adeptness in creative problem-solving. From a technical standpoint, this project entails employing suitable data processing and analytical tools, as well as delving into the application of various machine learning algorithms to unearth hidden patterns in the data.

The focal point of this endeavour lies in crafting a brief technical report. This report serves as a testament to the knowledge you have acquired and should be comprehensible to a reader with a reasonable level of understanding of the data analysis. Through this report, you will adeptly convey your insights and discoveries, offering a valuable perspective on the landscape of movies and TV shows on Netflix.

2. Assignment Structure

Group registration (Due: Friday 8 Sep at 5 PM)

You will need to form a project group of 3-4 members (maximum 7 groups per workshop). Members MUST be from the same workshop group. You will need to register your group via

your workshop tutor and elect the group leader who will be responsible for submitting all team deliverables.

Students who have not formed a group will be randomly assigned to a group. For students who propose groups of less than 4 members, we reserve the right to add members to the group or to reassign the members to other groups, if this proves necessary for overall load balancing.

Group Contract – 1 mark (Due: Friday 15 Sep 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 via Canvas/Assignment 2: Task Explanation page. You are welcome to work with the provided template or customize it according to your preference. *Submit as a single PDF file through Canvas/Turnitin.*

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 – 27 marks (Due: Friday 6 Oct at 5 PM)

1. **Report**: A written report, of 2000 to 3000 words. All group members' names and student IDs should appear on the first page of the report. The word count includes all the text including references, captions, and the tables' content. *Submit as a single PDF file through Canvas/Turnitin.*
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. *Submit as a single zip file through Canvas/Turnitin.*

Slides Submission (Due: Monday 9 Oct at 9 AM)

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

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

Oral Presentations – 10 marks (Due: from Monday 9 Oct to Friday 13 Oct)

During week 11 all teams should deliver an oral presentation of their work and findings for assignment 2. The presentations will be conducted in the students' usual workshop room and be assessed by two individual markers.

Teamwork Assessment – 2 marks (Due: Friday 13 Oct at 5 PM)

For this assessment, each team member will complete a form to evaluate both their own contributions and those of their teammates in the assignment. This form should align with the expectations you set in your “group contract” submitted in week 9.

Details about this team assessment form will be published via the Canvas/Assignment2: Teamwork assessment page closer to the assignment due date.

The evaluations provided by your fellow group members will serve as the basis for calculating individual teamwork scores. In addition, these scores, hold the potential to influence the adjustment of individual marks assigned for the report and/or oral presentation for each team member.

3. Data Sets

You are provided with two separate CSV files. The main dataset (Titles.csv) contains information about Netflix movies and TV shows, while the second file (Credits.csv) contains information about the 'actors' and 'directors' for the titles in the main dataset.

You can choose to use only the main dataset or use both provided files. What you use will depend on your research question and the analysis approach your group agreed on.

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 one research question clearly and explicitly. We have presented a set of research questions in the accompanying video to provide you with 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 crucial to note that the pursuit of several inquiries does not inherently lead to a greater number of marks being awarded. We will primarily evaluate the quality of your work through a focus on the depth of your analysis, the valuable insights it yields, and the quality of your outcomes, rather than simply covering a larger quantity of content or material.

4.2. Data Pre-processing

So far, 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 your project, you're encouraged to apply any of these methods to the provided datasets that you find suitable. Applying these techniques will enhance data management and render the data ready for extracting intriguing insights. The objective is to implement a minimum of three techniques, though you're welcome to utilize as many data preparation techniques as you see fit. A well-prepared dataset amplifies the value of your analysis. Utilize visualizations to discern how you can optimize your data preparation.

In your report and presentation, ensure you provide justifications and explanations for the methods you employ, present the results, and highlight any noteworthy discoveries.

Remember, there's no universal solution here. The more you engage with and understand your data, the more proficient you'll become in advancing to the subsequent stage of your project.

4.3. Use of supervised and unsupervised models

In this subject, we explore a range of machine-learning 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 available too.

Feel free to choose any method(s) that you'd like to further investigate and 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 techniques, though you're welcome to utilise more if you so choose. You might opt to employ two supervised models, two unsupervised methods, or even one of each.

In your report and presentation, it's vital to articulate your rationale behind the 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; elucidate the implications of these outcomes.

This presents an opportunity to apply your acquired knowledge and showcase your grasp of these techniques. Your ingenuity in method selection and explanation will greatly influence the success of your assignment.

NOTE: You are strongly encouraged to make use of any existing Python libraries (such as *sklearn* or *scipy*) in your attempts at this assignment.

5. Report

Your primary submission of 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 introduce 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.
- **Findings and Interpretation:** Provide a list of 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.
- **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 Canvas/Assignment 2: Task Explanation page. You are welcome to work with the provided template or customize it according to your preference.

6. 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.
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 comprehensive list of the findings you unearthed during your analysis. Subsequently, delve into an in-depth 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 **8 minutes**. If your presentation doesn't finish on time the markers will interrupt and stop you and it will also negatively impact your results. There may be a further 4-5 minutes of questions and answers from the markers.
- **Presenters**: All the members of the team should attend the presentation (unless they have an exemption granted by the teaching team). It is required for every member of the group to deliver some content and be actively involved in the following Q&A.
- **Slides**: To ensure fairness for all groups and prevent last-minute modification based on other teams' work, you will be asked to use the exact version of the slides that you have submitted to Canvas when presenting.

6. Assessment Criteria

The report will be marked according to the rubric published via the Canvas/Assignment 2: Task Explanation page.

The oral presentations will also be marked according to the rubric published via the Canvas/Assignment 2: Task Explanation page.

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 an executable code that supports your findings, we reserve the right to give your team **zero** marks for the report section.

7. Terms and Conditions

8.1 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.

8.2 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.

8.3 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 inappropriate levels 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 appears 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.