CSE5DEV: Data Exploration and Analysis

Assessment 2: Assignment on data exploration and analysis (detailed task instructions)	
Assignment type	Written Report
Weighting	25%
Word count / length	1000 words (indicative)
SILOs	1,2,3,4,5
Due date	Sunday, 10 September 2022 23.59 (Melbourne time*)

Topic overview

This assignment is to test your knowledge and ability in data cleaning, data exploration and data analysis techniques for application to dimensionality problems and evaluate the performance of these techniques. For this assignment, you will be given datasets to do data exploration and analysis. You need to create a report that expresses the key insights about the data with appropriate visualisations.

Assessment criteria

This assessment will measure your ability to:

- Accurately apply data analysis concepts and implement steps as instructed to produce correct output - 60%
- Synthesise and evaluate research on relevant topics and evaluate and apply findings to produce an appropriate outcome. - 40%

Guidelines

- Use one R Markdown file for this assignment.
- Implement each step in one Chunk (block) of code. Do not put your code in one Chunk (block) of code.
- You should write a piece of code for each process. Do not manually add, delete, or make any sort of calculation.
- You need to submit the R Markdown file and the generated HTML file.

Submission format

Upload 2 files in the LMS submission space.

- an R Markdown file and
- the generated HTML file

Detailed task instructions

Task 1. 40 marks (10 marks each)

The Dataset for task 1: Studentmarks.csv.

Dataset Description: The student marks data set contains student id, student name, date of birth, and the marks they got in 2020, 2021, and 2022.

For the dataset, complete the following tasks:

a. Calculate the age of the students and add that as a new column age1.

b. Split the dob column into date month and year and then calculate the age based on the year column only and add that as a new column age2

c. Create a scatter plot for Studentname versus Marks of all three years.

d. Calculate the total marks of all students and filter the who got at least 200 marks in total and create a bar chart for students vs totalmarks in the descending order of their marks.

Task 2. 60 marks (15 marks each)

The Dataset for task 2: Movies.csv

The film industry or motion picture industry is one of the largest sources of entertainment in the world. The industry produces thousands of films annually and rakes in billions of dollars in revenue. As a consultant, your tasks are to analyze the Dataset obtained from IMDB. An online database of information related to films, television programs, and video games, including cast, production crew, fictional characters, biographies, plot summaries, trivia, and reviews.

Dataset Description:

The movie dataset gathers data on movies and a few attributes about its structure, like movie duration in minutes, release year, director, and actors, and so on. Besides these variables, there are a few columns of interest regarding the movie evaluation, such as IMDB score, reviews, audience's votes, and Facebook's likes.

The goal is to bring a perspective in one of these variables (e.g. profit) and its relationship with the other variables in the Dataset. Exclude variables of non-interest (if applicable).

dependent variable (y)

For the Dataset, complete the following tasks:

a. Handle missing values: Remove or impute all missing values? Support the method you choose (remove or impute) with appropriate arguments. You need to show the step by step process.

b. Based on the Dataset, calculate "Profit" and determine the relationship between "Profit" and other variables (e.g. IMDB score). Hint: Profit = Gross – Budget. Use line plot or scatter plot to find the relationship. Present a summary of your findings

right formula? negative values?

c. Calculate the correlation between the relevant variable(s) used in the Dataset. Present a summary of your findings

d. Based on the correlation matrix, list and plot Strong and Weak Correlations. Provide appropriate reasoning about the findings.

Options: Inpute the mean,median,mode or Do not impute (get rid of rows or columns that has NA values), else continue (to observe how it impacts the entire dataset)

This is an individual Assignment. You are not permitted to work as a group when writing this assignment.

Copying, Plagiarism: Plagiarism is the submission of somebody else's work in a manner that gives the impression that the work is your own. The Department of Computer Science and Information Technology treats plagiarism very seriously. When it is detected, penalties are strictly imposed. Students are referred to the Department of Computer Science and Information Technology's Handbook and policy documents with regard to plagiarism and assignment return, and also to the section of 'Academic Integrity' on the subject learning guide.

No extensions will be given: Penalties are applied to late assignments (5% of total assignment mark given is deducted per day, accepted up to 5 days after the due date only). If there are circumstances that prevent the assignment being submitted on time, an application for special consideration may be made. See Student Handbook for details. Note that delays caused by computer downtime cannot be accepted as a valid reason for a late submission without penalty. Students must plan their work to allow for both scheduled and unscheduled downtime.

La Trobe Learning Hub:

The Learning Hub has been designed to develop and extend your academic skills.

You can access the Library services remotely through the library website.

Referencing guidelines

Please use APA 7th edition as your referencing style. For more information, see the Academic Referencing Tool of the Library.

The reference list is not included in the word count. In-text citations are included in the word count.

Academic integrity and plagiarism

'Academic integrity means being honest in academic work and taking responsibility for learning the conventions of scholarship . . . Academic integrity education is integral to the learning experience at La Trobe University . . . The University requires its academic staff and students to observe the highest ethical standards in all aspects of academic work, and it demonstrates its commitment to these values by awarding due credit for honestly conducted scholarly work, and by penalising academic misconduct and all forms of cheating' (La Trobe University, n.d.).

The penalty for submitting an assignment as your own that is the work of a third-party may be as severe as 'exclusion from the University without readmission' (La Trobe University, 2020, p. 4).

Refer to the *Academic integrity – schedule of responses and penalties for academic misconduct*.

You should familiarise yourself with the Academic integrity website and complete the academic integrity module (AIM) in your LMS.

If you have any questions regarding academic integrity, your Subject or Course Coordinator will be able to assist.

Students are also referred to the Department of Computer Science and Computer Engineering's Handbook and policy documents about plagiarism and assignment return, and to the document on 'Academic Misconduct' in the subject learning guide.

References

La Trobe University. (2020). *Academic integrity – schedule of responses and penalties for academic misconduct*. Retrieved February 2, 2021, from https://www.latrobe.edu.au/__data/assets/pdf_file/0006/847923/academic-integrity-schedule-of-responses.pdf

La Trobe University. (n.d.). *Academic integrity policy*. Retrieved February 2, 2021, from https://policies.latrobe.edu.au/document/view.php?id=221

Assessment criteria / grading rubric

CRITERIA	A: Excellent (>80%)	B: Very good (70–79%)	C: Good (60–69%)	D: Acceptable (50–59%)	N: Unacceptable (<50%)
Accurately apply data analysis concepts and implement steps as instructed to produce correct output. (The marking criteria applies to each individual questions from 1a to 1d - **10 marks each**) **(Total: 40 marks)**	The output is correct with all the required information. The output doesn't have any errors. The answers demonstrated sophisticated knowledge of the subject content by integrating major and minor data analysis concepts. Implemented each step in one block of code with appropriate comments. **(8-10 marks)**	The output is correct with most of the required information. The output has minor insignificant errors. The answers demonstrated a well-developed knowledge of the subject content by integrating major and minor data analysis concepts. Implemented steps in one block of code with mostly appropriate comments. **(7 marks)**	Meet all requirements. Minor errors which may lead to incorrect answer. The answers demonstrated a developed knowledge of the subject content by integrating major data analysis concepts. Implemented each step in one block of code but some comments are missing or incorrect. **(6 marks)**	Meet all requirements. Includes errors which result in an incorrect answer. The answers demonstrated a developing knowledge of the subject content by integrating major data analysis concepts. Implemented most steps in one block of code but comments are missing or incorrect. **(5 marks)**	Inadequately developed program that does not reflect the necessary requirements. Major errors and do not run. The answers did not demonstrate enough knowledge of the major data analysis concepts. Poorly written without any structure **(0-4 marks)**
Accurately apply data analysis concepts and implement steps as instructed to produce correct output. (The marking criteria applies to each individual questions from 2a to 2d - **15 marks each**) **(Total: 60 marks)**	The output is correct with all the required information, and it is displayed in the correct format There are no errors. The packages, functions, algorithms and standards are implemented correctly. The answers demonstrate excellent knowledge of the subject content by integrating major and minor data exploration and data analysis concepts. Implemented each step in one block of code with highly appropriate comments. **(12 - 15 marks)**	The output is correct with most of the required information, and it is displayed mostly in the correct format There are few errors. The packages, functions, algorithms and standards are implemented mostly correctly. The answers demonstrate well-developed knowledge of the subject content by integrating major and minor data exploration and data analysis concepts. Implemented each step in one block of code with appropriate comments. **(11 marks)**	The output is not correct or is displayed in incorrect format The packages, functions, algorithms and standards are implemented but with some minor errors The answers demonstrate a developed knowledge of the subject content by integrating major data exploration and data analysis concepts Implemented each step in one block of code but some comments are missing or incorrect. **(9 marks)**	The output is not correct or is displayed in incorrect format The packages, functions, algorithms and standards are implemented but with many minor errors The answers demonstrate developing knowledge of the subject content by integrating major data exploration and data analysis concepts Implemented each step in one block of code but many comments are missing or incorrect. **(8 marks)**	The output is not present or has major errors and do not run The packages, functions, algorithms and standards are not implemented correctly. The answers do not demonstrate enough knowledge of the major data exploration and data analysis concepts. Poorly written without any structure **(0-7 marks)**

TASK 1	MARKS	TASK 2	MARKS	Total marks:
1a		2a		
1b		2b		
1c		2c		
1d		2d		