Durham University

Computer Science

Summative Assignment

Module code and title	COMP2271 Data Science
Academic year	2023-24
Coursework title	Data cleaning and analysis coursework
Coursework credits	5 credits
% of module's final mark	25%
Lecturer	Jingyun Wang
Submission date*	Tuesday, December 12, 2023 14:00
Estimated hours of work	10 hours
Submission method	Gradescope (code)

Additional coursework files	*amazon_laptop_2023.xlsx*
Required submission items and formats	*A zip file includes 2 files:* • *"userID.jpynb" or "userID.py"* • *userID_report.pdf*

* This is the deadline for all submissions except where an approved extension is in place.

Late submissions received within 5 working days of the deadline will be capped at 40%.
Late submissions received later than 5 days after the deadline will receive a mark of 0.
It is your responsibility to check that your submission has uploaded successfully and obtain a submission receipt.

Your work must be done by yourself (or your group, if there is an assigned groupwork component) and comply with the university rules about plagiarism and collusion. Students suspected of plagiarism, either of published or unpublished sources, including the work of other students, or of collusion will be dealt with according to University guidelines (https://www.dur.ac.uk/learningandteaching.handbook/6/2/4/).

Data Cleaning and analysis

The submission for coursework should be in the form of a report (for Problem1 and 2) named "**userID_report.pdf**" (the userID is your 6-digit userID.) The whole report should be **no more than 750 words**. Put your word count at the beginning of the report. (No abstract is needed. This word count includes headings, captions of figures/graphs/charts and tables, footnotes/endnotes, and contents of tables but **does not include references**. Use **APA style** of referencing.)
You also need submit a code named "userID.jpynb" or "userID.py". Put "Problem 1" and "Problem 2" in the comments to identify the solution for each question. (The mark proportion for the report is 70% and for the code is 30%).

Background

The dataset **amazon_laptop_2023. xlsx** consists of comprehensive collection of latest available laptops scraped from Amazon.com. The data includes product details such as the rating, price, operating system, title, review count, and display size. Its metadata is attached in the second spreadsheet.

Problem 1 (40 Marks):
Use the **data cleaning** methods you learned in the lectures to clean and process the **amazon_laptop_2023.xlsx** and then save to another file called "**amazon_laptop_2023_cleaned.xlsx**". For the report of this question, please use the following table as the template:

Column name	Situation of the column	Cleaning actions/steps	Justification/Explanation

Feel free to use figures and citations to support your Justification. Figures should come after the above table with captions.

Problem 2 (60 Marks):
Assume a customer want to buy a laptop in amazon with a budget less than 1500 dollars. Imagine any 2 types of customers with different requirements and recommend 5 laptops based on their different needs. For example, if the customer needs to travel a lot, what are her/his best options.

Create a Python program to analyse the data in **amazon_laptop_2023.xlsx**. Use **seaborn** (other visualization libraries are not allowed, except for using Matplotlib to specify figures' size and labels) and **pandas** to support the data processing. In addition to the code, you also need to submit a report to explain the reason of your recommendation with the support of visual graphs.

- Comment your code properly;
- Provide reader-friendly visualization as taught in the lectures;
- Do not show you code in the report!

PLAGIARISM and COLLUSION

Students suspected of plagiarism, either of published work or work from unpublished sources, including the work of other students, or of collusion will be dealt with according to Computer Science and University guidelines.

Please see https://www.dur.ac.uk/learningandteaching.handbook/6/2/4/ **and** https://www.dur.ac.uk/learningandteaching.handbook/6/2/4/1/ **for further information**