Programming for Data Analytic
SOFT8032
Second Examination

October 24, 2025

1 Second Assessment. First Project

This project contributes 30% in your final mark. This is an individual project and has to be all done by yourself. You are not allowed to disclose your code to anyone else. All submissions are subjected to Generative AI detection tools.

You MUST present your submission during your lab time (details is provided in the bottom of the project spec).

IMPORTANT: **This project is designed so that you can use the functions, syntax, and topics covered in the lectures and labs. Therefore, using any functions, syntax, or topics outside the module content (lectures/labs) will result in a deduction of points.**

1.1 Project Specification

The objective of this project is to analyze the information and content in the dataset named **shopping.csv**. Please perform the following tasks:

1. The goal of this task is to analyze the relationship between products and shipping types. First, you need to clean the dataset (e.g., fix incorrect entries or remove noises). Then, for each unique product, identify which shipping type is the most popular, the second most popular, and so on, down to the least popular. Finally, print each product name along with a list of shipping types sorted from the most to the least popular, including their respective percentages of popularity.

2. The goal of this task is to perform an analytical study on distinct customer segments or groups based on their purchasing values, separated by gender. Each customer in the dataset is associated with a current purchase amount in USD and a number of previous purchases ('Previous Purchases'). To approximate the total purchase value for each customer, we define a new feature called 'Total Purchased USD', calculated as $Purchase\ Amount\ (USD) \times Previous\ Purchases$.

   (a) Use an appropriate visualization technique to visually display the extent of each customer segment separated by gender (Male and Female). The segments are defined in Table 1. Apply all the necessary visualization features and details to the plot.

Segment	Total Purchased USD
Segment1	0–500
Segment2	500–1000
Segment3	1000–1500
Segment4	1500–2000
Segment5	2000–2500
Segment6	2500–3000
Segment7	3000–3500
Segment8	3500–4000
Segment9	4000–4500
Segment10	4500–5000
Segment11	5000–5500
Segment12	5500–6000

Table 1: Segments and Their Corresponding Ranges in Total Purchased USD

(b) In the console, separately print the population of each gender for each of the segments above.

Note: If any non-numerical values exist in the required columns for this task, the corresponding rows should be ignored.

3. Product Analysis:

   In this task, the *Previous Purchases* for each product will be analyzed.

   For each product, calculate the following:

   (a) For a given product, identify all **unique age groups**. For each unique age group, calculate the *average of 'Previous Purchases'*. For example, if a product has 10 unique age groups, you will have 10 averages of 'Previous Purchases'. Then, calculate the average of these 10 values and denote it as A.

   (b) Calculate the **overall average of 'Previous Purchases'** for a given product: Compute the average 'Previous Purchases' across all entries for a given product and denote it as B.

   Thus, for each product, you will have two metrics: A and B. Use an appropriate visualization technique to display both metrics for all products. Additionally, print the names of the products where $A < B$ in the console.

4. Date analysis.

   - Use the information in the `Dates` column to visually depict the popularity of seasons. Note that the first season starts on March 21, and each season lasts three months. A popular season is one with the highest number of sales.
   - Use an appropriate visualization technique to show the popularity of individual months. Print the names of the top three most popular months across all entries. A popular month is one with the highest number of sales.

- Write a function that takes a product name as input and visually depicts the number of sales for that product year by year. Hint: for a given product, group the dataset by year, then apply an appropriate visualization technique to show the sales distribution across each year for that product.

5. Apply one analytical task of your choice. Make sure the chosen task is useful for people in this industry and also complex enough. Use comment section and explain the idea of your task.

Note: Printing unnecessary information will result in negative marks. Print only what is required.

Note that all visualization plots need to have proper labels and annotations. Lack of visualization features attracts penalty.

Efficiency is crucial for this project, as a more efficient code can expedite the process and reduce the likelihood of errors in the analysis. For example, avoid unnecessary loops and hard coded values, and so forth. Also please use only the syntax, functions and libraries that have been covered in the lecture and lab.

1.2 Submission and Deadline

Please only submit one python file and type your name and id as comment on the top of the file. Each task should be implemented as a separate function with an interpretation provided as a comment below the function if needed. You may define additional functions if needed. The deadline for this project is 16th Nov 2025 at 23:59

1.2.1 LATE SUBMISSION PENALTIES

Please refer to the late submission penalties below:

Up to 1 calendar week delay: 10 marks penalty is applied.

Up to 2 calendar weeks delay: 20 marks penalty is applied

Over 2 calendar weeks delay: no submission will be accepted.

Any content copied and pasted may be sent to the plagiarism board for disciplinary proceedings. All work must be your own and any information taken/inspired from other sources should be clearly referenced.

.

You are required to present your code to me during the labs in week 10, 11 or 12, depending on when you submit your work. No marks will be assigned if you do not present your code during the lab. Please note that you are required to present your project individually, not to the entire class. You should be prepared to answer questions about your code.

If you have any questions about this project, please contact Farshad Ghassemi Toosi at farshad.toosi@mtu.ie or via Canvas.

1.3 Rubric

This rubric is subject to change.

1. Correct task implementation and code efficiency (Data extraction, data cleaning, code efficiency and correct visualization if needed etc.). (100%)

2. Relatively correct task implementation and code efficiency (Data extraction, data cleaning, code efficiency and correct visualization if needed etc.). (70%)

3. Partly correct task implementation and code efficiency (Data extraction, data cleaning, code efficiency and correct visualization if needed etc.). (40%)

4. Wrong task implementation. (0%)