COMPUTATION II – ALGORITHMS & DATA STRUCTURES

2023 PROJECT GUIDELINES

Contact:

Berfin Sakallioglu, NOVA IMS

bsakallioglu@novaims.unl.pt

1 Project Description

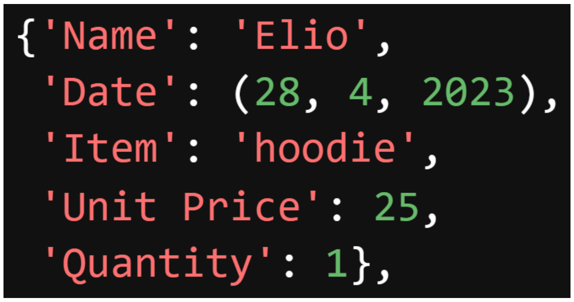
You are asked to apply your knowledge about algorithms and data structures to analyze a dataset.

- All the communication will be done via email to: bsakallioglu@novaims.unl.pt.
- The group members will be chosen by the students filling the following sheet. The oral defense schedule will be published there as well: Sheet
- This is a group project. The groups will have 4-5 members. If you are having trouble finding a group, please send an email.
- You are required to use the provided Jupyter Notebook on Moodle for implementing your code. It is essential to add clear and concise **comments** to your code to enhance readability and understanding.
- You are allowed to use `NumPy and Matplotlib` but you are not allowed to use `pandas`.
- The deadline is 11th of June. More information is given in Chapter 3 Evaluation Guidelines.

2 Project Objectives

2.1 Dataset

You have been provided with a dataset containing the purchase history of a hypothetical Students' Union [1]. The dataset is a list of dictionaries, with each dictionary representing a single purchase containing the following information:



```
{'Name': 'Elio',
 'Date': (28, 4, 2023),
 'Item': 'hoodie',
 'Unit Price': 25,
 'Quantity': 1},
```

- **Name:** The name of the student.

  The values associated with these keys are strings containing the names of the students. Assuming that each person has a unique name (Although this may not always be the case in practice, especially in Portugal!), there will inevitably be cases where multiple purchases are associated with the same name in the dataset. This is to be expected, as it is logical that a single student may make multiple purchases on different days, or even on the same day.

[1] The union, all the names, and data are fictitious. No identification with actual persons, institutions and products is intended or should be inferred.

- **Date:** Date of the purchase.

  The value associated with the 'Date' key is a tuple consisting of three elements, which represents a date in the format of **(DD, MM, YYYY)** (since we're not lunatics). The first element corresponds to the day of the purchase, the second to the month, and the third to the year. Please note that the dataset provided to you only includes purchases made between January 1st, 2023 and May 1st, 2023.

- **Item:** The name of the item that was purchased.

  The key represents the name of the item that was purchased from the Students' Union. This refers to the specific product that was bought from the range of items offered by the SU.

- **Unit Price:** The unit price of the item.

  The 'Unit Price' key represents the price of the item purchased, expressed in Euros as a floating-point number. Please note that this value represents the price per unit of the item, rather than the total value of the purchase.

- **Quantity:** Quantity of the item.

  The 'Quantity' key represents the number of items that were purchased, expressed as an integer value.

2.2 Extraction of Information

In this section, you will extract information from the dataset for each person and save this information. As you read through the tasks, it may be useful to consider what **data structure** you will use to store the extracted information. One option is to create a new dataset that includes customer names and the extracted information, while another option is to extend the existing dataset with additional fields for the extracted information.

2.2.1 Recency

Using the provided dataset, calculate and save the number of days since each student's/customer's latest purchase. You can use May 1st as the current date for this calculation. This will provide us with information about the **recency** of each student's last purchase, indicating how recent they made a purchase from the SU.

2.2.2 Frequency

Employing the dataset that has been given, calculate and save the total number of purchases made by each student throughout the given date interval. This information will tell us how many times each student bought something from the SU.

2.2.3 Monetary Value

Using the provided dataset, calculate and save the **total** amount spent by each student on purchases throughout the year. This information will tell us the total value of all the items purchased by each student from the SU at the given date interval.

Plot **histograms** for recency, frequency and monetary value of the dataset. (You can make use of `Matplotlib`.)

2.3 Sorting

In this section, you will sort the student names based on their calculated **recency**, **frequency**, and **monetary** value. It is important to use an efficient and suitable sorting algorithm for this task.

First RFM Score (RFM)

The three independent steps for assigning recency, frequency, and monetary value scores to each student are as follows:

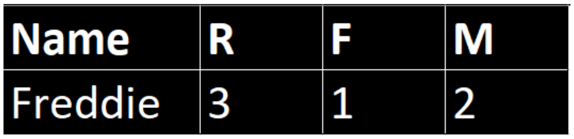
- Sort the students based on their calculated recency, with the student who made the most recent purchase appearing at the top/beginning of the structure.

  You need to divide the sorted students into three equal parts based on when they made their last purchase. Then, assign a **recency score** to each part, with a score of 3 for the most recent third of customers, a score of 2 for the middle third, and a score of 1 for the least recent third.

- Resort the students based on their calculated frequency values, highest frequency being on top/beginning. Assign a **frequency score** of 3 to the top one-third of the list, 2 to the second one-third, and 1 to the remaining one-third.

- Resort the customer list based on their calculated monetary values, largest value being on top/beginning. Assign a **monetary score** of 3 to the top one-third of the list, 2 to the second one-third, and 1 to the remaining one-third.

After these steps, we will have assigned an **RFM score** (recency, frequency, monetary value) to each student.



Name	R	F	M
Freddie	3	1	2

Using the name of the students and their RFM score, perform another sorting of the data:

- Sort the names using their RFM score in descending order. In the end, we are expecting to see the students with RFM score 333 at the top/beginning of the list and 111 at the bottom/end of the list.

Can you determine which students are most valuable to our SU, considering that our organization is evil and values students solely based on their purchase history? (see page 2 footnote)

Second RFM Score (RFM')

Recalculate the recency, frequency and monetary value scores for each student as following:

- Follow the first step of the First RFM Score calculation. Rank the students by their recency and create 3 buckets with **scores** from 3 to 1.
- **Within each bucket of recency**, rank students by frequency and create 3 buckets. Assign a **frequency score** of 3 to the top one-third of the list, 2 to the second one-third, and 1 to the remaining one-third.
- **Within each bucket of frequency**, rank customers by monetary and create 3 buckets. Assign a **monetary score** of 3 to the top one-third of the list, 2 to the second one-third, and 1 to the remaining one-third.

Use `scipy.stats.spearmanr` to correlate RFM and RFM'. This function calculates the Spearman Rank Correlation which is used to measure the correlation between two ranked variables.

2.4 Searching

- Using linear search, find the RFM scores of your group members. (Assuming the dataset contains names of everyone, otherwise apologies, you can search another name.)

Who do you think is the best student/customer among your group?

- Sort the students in alphabetical order. (Hint: You may assign a numerical value to indicate the position of a letter in the alphabet using a dictionary.)
- Search for your group members using binary search and find their RFM scores.

How much time, in terms of seconds, did it take to find a name using linear search and binary search? Compare the results. (You can use the `timeit` module.)

3 Evaluation Guidelines

3.1 Deliverables

The project must be delivered via Moodle as a `.ipynb` notebook. The notebook provided to you on Moodle under the Project section must be used as a template, and the order and structure of the chapters must be maintained. Ensure that all questions are answered in the notebook. Before submitting, every cell in the notebook must be executed to display the corresponding clean output, if any. The `.ipynb` notebook must also include the names, surnames, and student numbers of all group members.

3.2 Evaluation criteria

For both the first and second examination epoch, the final grade of the curricular unit is obtained by the following formula:

$$final\_grade = 0.5 * group\_project + 0.5 * written\_exam - behaviour\_penalty$$

A minimum grade of 9.5 is required for both evaluation components. The grade of the project will be the same in the first and second examination epochs. Grade of this group project will be given by the following formula:

$$group\_project = 0.8 * notebook\_grade \ + \ 0.2 * defense\_grade$$

The following (unsorted) list provides details about each major criterion:

- The code must be *clean* and run without errors, the code cells must be very well commented. Clarity, synthesis, and objectiveness are highly appreciated.
- Ability to explore, prepare and visualize the data.
- Ability to correctly perform sorting.
- Ability to correctly perform searching.
- Ability to use suitable data structures.
- Ability to justify your decisions with *a priori* reasoning and comment upon the obtained results.

3.3 Deadline

The last day to deliver the project is the 11th of June, Sunday night at 23:59, 35 days after the date of the release of the project, therefore, there will be no deadline extension. You can resubmit your project on Moodle as many times as you need until the deadline.

3.4 Defense

There will be a mandatory presential defense where all the group members must participate. Different members of the same group can receive different grades. There is no need for any kind of slides or presentations for the defense, it will be in Q&A format. The absence of a group element implies 0 on their grade of defense. The schedule of the defenses will be published closer to the date by Berfin Sakallioglu (It is planned to be done on the following week days after the deadline.).

3.5 Grade Discussion (i.e., aftermath)

Students will be provided with an opportunity to optionally discuss their grades; for that, all the group members must manifest their respective intention and appear (all) in the scheduled time slot. The discussion will involve an explanation of how the evaluation criteria were applied. The outcome of the discussion can be threefold: (1) the grade is maintained, (2) the grade is increased or (3) the grade is decreased. The grade can be either increased or decreased if and only if it will be evident that the evaluation criteria were applied incorrectly. Additionally, the grade can be decreased (individually) if some of the group members exhibit a clear absence of knowledge regarding the project and the course.