CATÓLICA PORTO BUSINESS SCHOOL

Coursework

Overview and Submission

The coursework covers all topics that we have discussed during the course. The grouping by tasks indicates the overarching focus of the questions. However, you may combine different concepts to solve the questions effectively.

Please use the file `submission_YOURSTUDENTID.py` to solve the questions. I already prepared the structure and added additional hints to make work easier. **Please replace "YOURSTUDENTID" in the file name with your actual student ID. Also make sure to add your student ID in line 20.**

Only submit the one Python file to Blackboard; do not upload the data sets or anything else. While you can output variables for debugging purposes, please make sure to remove all print statements prior to your submission.

Grading

Most of the time, there is no single correct answer. While you should aim to solve the problems with as little code as possible, you get full marks if your code produces the expected outputs (almost regardless of how you made it work). Nevertheless, your code must be flexible enough to work with slightly different inputs. To assure this is the case, please use variables rather than hard-coded values whenever possible.

One of the most crucial aspects in practice is proper documentation of your code. Hence, you may add as many comments and documentation strings as necessary to make it easy for someone else to understand your workings.

Make sure to work clean. PyCharm will indicate if you are not following best practice. For instance, there should be a space before and after an equal sign when defining variables. However, there is no space when using keyword arguments.

```
my_variable = 10  # a space before and after the equal sign
df = pd.read_csv('./data/file.csv', nrows=100)  # no spaces
```

The points for the assignment are awarded as follows:

- 80% Quality of the code (would your code work even if the actual inputs are slightly messier/different to the sample data sets?)
- 20% Quality of comments and readability of the code

CATÓLICA PORTO BUSINESS SCHOOL

Task 1: Python fundamentals

Q1: The three essential built-in data types are lists, tuples, and dictionaries. While two of them might seem similar in structure, they are commonly used for different purposes. How would you use them in your research projects (give one example for each of the data types and describe in one sentence why you would use the data type in this context)? Make sure to use different examples for each data type.

Q2: Use a list comprehension to calculate the square root ($\sqrt{n}$) of all numbers between 11 and 23 (inclusive). Hence, your list may contain 13 items. Make sure to use the given variables for the lower and upper limit (your code should work with other ranges too).

Q3: You decide to create an empty DataFrame in preparation for one of your analyses. The DataFrame should consist of `num_rows` rows and `num_cols` columns that are labelled as in the screenshot (beginning from one). Both the number of rows and columns can vary from anywhere between 2 and 100. Save the DataFrame as a pickle file to the following director: `./data-task1/q3_df.pkl`

	Portfolio_1	Portfolio_2	Portfolio_3	Portfolio_4	Portfolio_5	Portfolio_6	Portfolio_7	Portfolio_8	Portfolio_9	Portfolio_10	Portfolio_11	Portfolio_12	Portfolio_13	Portfolio_14
Company_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Company_2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Company_3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Company_4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Company_5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Company_6	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Q4: Rewrite the following for loop using a list comprehension. Explain the functioning of the for loop and if conditions in one or two bullet points in documentation string. (The code below is not a screenshot and can be copied.)

```
companies = ('Apple', 'Amazon', 'Alphabet', 'Microsoft', 'Visa')
companies_new = list()

for company in companies:
    if 'p' in company:
        companies_new.append(company.upper())
    else:
        companies_new.append(company.lower())

print(companies_new)
```

Q5: Create an empty text file (i.e., a file without content) in the subfolder `./data-task1/` for each of the companies given in the tuple `companies`. Use the company name as its file name, e.g., `Microsoft.txt`. Make sure that your code would work with other inputs too.

Q6: Read all text files in the subfolders of `./data-task1/q6/` and add the file content (a four-digit number) to a list. Make sure to only read files with a file extension `.txt`. Hint: you may want to automate the process. There should not be any code changes necessary if the number of files or subdirectories changes.

Task 2: pandas

We focus on executive compensation data that is provided by Capital IQ. More specifically, we work with the annual compensation table "ANNCOMP" of ExecuComp.

ANNCOMP lists all named executives, titles, and their compensation data. Compensation data includes items such as: salary, stock options, bonuses, and shares owned.

You can find a description of all the variables in the attached Word document: `./data-task2/_Description-Execucomp.docx`

We are mainly interested in the following fields:

- GENDER: Identifies the gender of the named executive officer
- AGE: Age of the executive as reported in the annual proxy statement
- JOINED_CO: The date that the named executive officer joined the company
- LEFTCO: The date that the named executive officer left the company
- CEOANN: "CEO" in this field indicates that this person was the CEO for all or most of the indicated fiscal year
- BECAMECEO: The date that the individual became CEO
- LEFTOFC: The date that the named executive officer left the position of CEO
- SALARY: The dollar value of the base salary (cash and non-cash) earned by the named executive officer during the fiscal year in thousands of dollars
- TOTAL_CURR: Total current compensation comprised of salary and bonus in thousands of dollars

You find the data set in the subfolder `./data-task2/`

Q10: Load the two CSV files `company-info.csv` and `compensation-data.csv`, and assign them to the corresponding variables (see Python file for more information).

Q11: Create a new column called `female`. This dummy variable should be equal to 1 if the executive is female and 0 otherwise. Make sure to not add additional columns; only `female`.

Q12: In 2017, what is the name of the company with the highest percentage/share(!) of female executives? For example, a firm with a total of 10 executives of which two are female, has a 0.2 or 20% share of female executives. You are required to return the company name or names. Hence, you must merge the two data sets. There are three companies with the same maximum value of female executives. You get full marks if you return either all of them or only one of them. Hint: you need pandas' split-apply-combine approach for this question.

Q13: In 2016, what is the average age of an executive per company? Ignore NaN values.

Q14: Calculate the tenure of CEOs. If there is no end date, you can assume that the executive is still appointed as CEO. If there is no start date, you are unable to calculate a tenure for that CEO; use an NaN value for the duration instead. Make sure to remove duplicate values (we have yearly observations and end up with a lot of duplicates). Save the resulting DataFrame as a pickled file `./data-task2/ceo_data.pkl`. Include only the columns listed in the Python file.

Task 3: Matplotlib

This task builds on what we already did in class. Hence, the data sets may seem familiar. You find the data set in the subfolder `./data-task3/`

Q21: Load the daily returns for Microsoft and BP.

Q22: You want to look into the montly performance of the two companies. Compute the cumulative returns per month for both companies. You end up with a new DataFrame containing the cumulative return in the rows and the two companies in the columns (similar to what we did in class).

Q23: Visualize the the twelve months of data using Matplotlib. Do not use the complicated Matplotlib syntax; you can solve the question using the simple pandas syntax as described in the pandas User Guide.