Dear students,

please read this carefully.

The exam is designed so that you do not get stuck in any particular part.

Your goal is to show us that you can think like a programmer, handle real-world issues with

processing data, and break the task into logical pieces.

You will most likely run into trouble. Do not give up! Coding is hard, and you will need to

gain a lot of experience, so we suggest approaching this with a cool head. Remember to use

Google and stackoverflow.com.

It is prohibited to share codes with each other! It is prohibited to use generative AI tools

such as Chat GPT.

INSTRUCTIONS

● Make sure you know your CUNI student number (like in your email

123456@fsv.cuni.cz).

● DEADLINE: 19:50, 8

th of April

● Present your work in a jupyter notebook in a GitHub repository, which you had in

through this form: https://forms.gle/wMjSWZQtuWKy3nv47

◦ Only pushed commits that fall within the time allowed will be taken into account.

● There should be no unreasonable delay between the deadline and the submission to

the Google form.

● Your code needs to run from scratch without errors.

● In total, you can receive 25 points.

● Generally, the topics tested are in lectures and seminars until Pandas II + Matplotlib.

If you show a decent attempt, you can receive points for partial work!

Data analysis – 25 points

We suggest creating a pandas DataFrame with the data. You might want to create a DF with

dates on the index, and a column per company.

PART 1

Get the data (0 points):

● Download your custom dataset from

https://ies-python-midterm.s3.eu-central-1.amazonaws.com/studentsSets/XXXXXXX

.zip

● where XXXXX stands for your CUNI student number.

o Make sure you copy the link correctly (no spaces, other characters (open it in a

text editor first))

o If you cannot find the file, let us know!

o It is a simple zip file, you should be able to open on Windows, Mac, unix

easily. There is no need to use python for this, you are welcome to do it

manually and unzip using your favorite tools.

● Make sure you can access the csv files within from your Jupyter notebook

● The name of the file is the ticker of a company, do not lose/rename it, you will need it.

● Make sure that Dates are correctly represented as Datetime variables

In the following analysis, use the Adjusted Close time series, unless specified

otherwise.

7x 2pt tasks + 1x 1pt task (15pts):

1. Is there a company that has no difference between the Close and Adj Close columns?

What does it mean from the financial point of view for the stock (you can get bonus

partial points)?

2. What is the highest and lowest price (Adj Close) each company recorded?

3. (1pt task) Calculate logarithmic returns from Adj Close. For each company report on

its, min, man, mean, median of the return distribution.

4. When did each company record the highest gain and highest loss for the day?

(logarithmic loss). Hint: idxmax

5. What is the average calendar weekly volume for each company? Hint: check how to

resample pandas DF

6. Which company recorded the highest total return over the whole period?

7. Plot the log-returns of the companies (ideally in the same plot)

8. Show the log-return distribution of the companies (ideally in the same plot)

SEE PART 2 on another page

Part 2

- Download the dataset about all S&P 500 companies

https://ies-python-midterm.s3.eu-central-1.amazonaws.com/companies/companies\_no

\_subindustry.csv . Note that pd.read\_csv can directly read public URLs.

2pt tasks (10 pts total):

1. Find out how many companies do not filled-in the date of inclusion (column

“included”) to S&P 500.

2. Delete the companies with no inclusion date and calculate which company is the

oldest/youngest constituent and tell us the average age of a constituent in the sample.

If you need to fix anything or make any assumptions, comment on them in the code.

Hint: pd.to\_datetime (some date column, dayfirst=True,errors='coerce')

3. Describe the distribution of companies across sectors and create a plot that

demonstrates the proportionality of the sectors (i.e. pie plot, or something like this)

4. Parse the “hq” column, extract the state of the hq and describe the distribution of the

states

5. Join the dataset with this one:

https://ies-python-midterm.s3.eu-central-1.amazonaws.com/companies/companies\_su

bindustry.csv

And join the two datasets based on an appropriate key. Report on distribution of

subindustries for the “Consumer Discretionary” GICS sector.

SUBMISSION CHECKPOINTS

o !! Save your work !!!

o Make sure your notebook runs from scratch – restart the kernel, import everything

and check it works

o Commit and push (commits not pushed to GitHub do not count!)