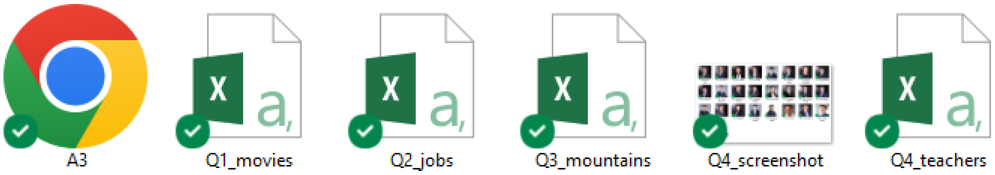
MSBA7001 Assignment 3

Module 1, 2023-24
HKU Business School

Contents

Instructions

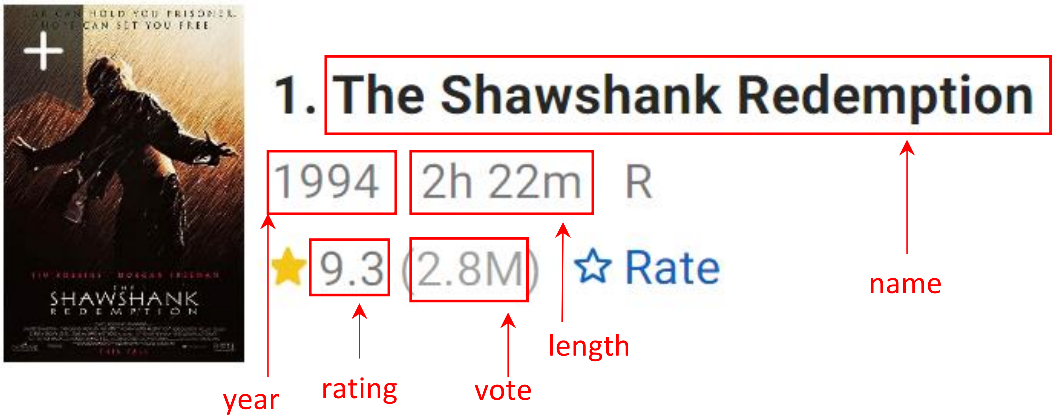
1. 4 questions, 6pts each. 24pts in total.
2. For every question, create an output heading, execute the required codes, and show your outputs.
3. Keep codes and data files in the same fold. Use relative file path.



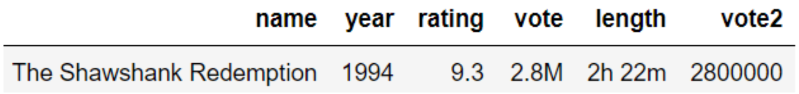
4. Partial points even if outputs are incorrect.
5. Save your codes in a Jupyter Notebook file named “A3.ipynb”
6. On Moodle, submit all files as shown in #3.
7. Due 11:30pm, Sept 30 (Saturday)

Q1 – top movies

Read this page: https://www.imdb.com/chart/top. Use only Regex (DO NOT use Beautiful Soup 4) to extract the following information from each movie.

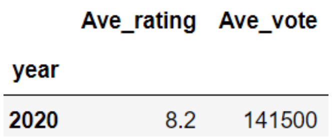


Store all 250 movies' info in a DataFrame called q1df. In addition, add a new column called vote2 which transforms votes to numerical values, for example, 2.8M to 2800000, 832K to 832000 (*hint*: pd.Series.mask and pd.Series.where). In q1df, make sure that the year and vote2 columns are int type, the rating column is float type, and all other columns are str type. The first movie is presented below for your reference.



name	year	rating	vote	length	vote2
The Shawshank Redemption	1994	9.3	2.8M	2h 22m	2800000

Furthermore, create a sample DataFrame called q1sample which shows the average ratings and average votes for movies released after the year 2020 (inclusive). The columns labels are "Ave_rating" and "Ave_vote". The first row is presented below for your reference.



	Ave_rating	Ave_vote
year		
2020	8.2	141500

Finally, write q1df to a csv file named "Q1_movies.csv". In **four separate** cells, execute the following codes and show your outputs.

```
q1df.info()
```

```
q1df.head(3)
```

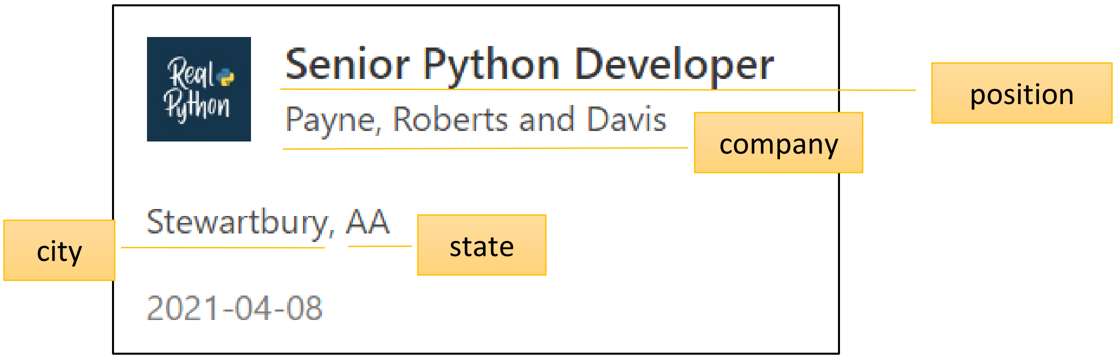
```
q1df.tail(3)
```

```
q1sample
```

Q2 – fake jobs

Note: this question is modified from a 2018 exam question.

Read this page: https://realpython.github.io/fake-jobs/. Extract the following information regarding job openings.



Senior Python Developer

Payne, Roberts and Davis

Stewartbury, AA

2021-04-08

Store all 100 jobs' info in a DataFrame called q2df. The first row is presented below for your reference.

position	company	city	state
Senior Python Developer	Payne, Roberts and Davis	Stewartbury	AA

Furthermore, create a sample DataFrame called q2sample that shows all "AA" state-based jobs that have "engineer" or "Engineer" in the name. The first row is presented below for your reference.

position	company	city	state
Energy engineer	Vasquez-Davidson	Christopherville	AA

Finally, write q2df to a csv file named "Q2_jobs.csv". In **four separate** cells, execute the following codes and show your outputs.

```
q2df.info()
```

```
q2df.head(3)
```

```
q2df.tail(3)
```

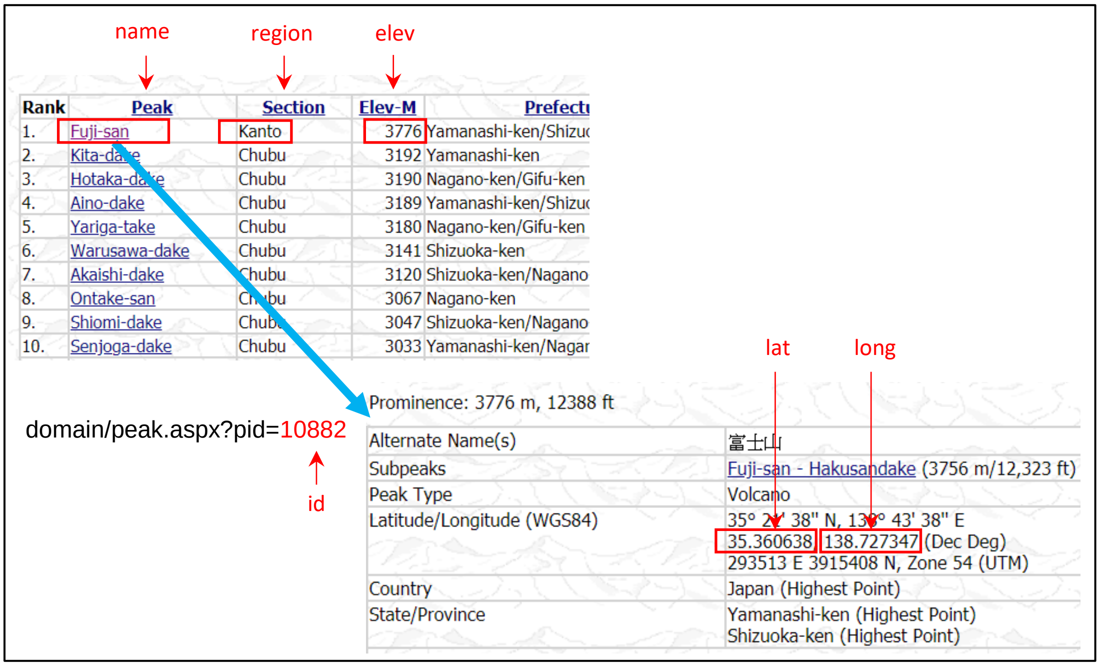
```
q2sample
```

Q3 – JP mountains

Note: this question is modified from a 2021 exam question.

There are “100 Famous Japanese Mountains” in Japan. Detailed mountain information can be found on the following webpage: https://www.peakbagger.com/list.aspx?lid=5651

Your task is to extract the highlighted values of all the 100 mountains.



Rank	Peak	Section	Elev-M	Prefectu
1.	Fuji-san	Kanto	3776	Yamanashi-ken/Shizuc
2.	Kita-dake	Chubu	3192	Yamanashi-ken
3.	Hotaka-dake	Chubu	3190	Nagano-ken/Gifu-ken
4.	Aino-dake	Chubu	3189	Yamanashi-ken/Shizuc
5.	Yariga-take	Chubu	3180	Nagano-ken/Gifu-ken
6.	Warusawa-dake	Chubu	3141	Shizuoka-ken
7.	Akaishi-dake	Chubu	3120	Shizuoka-ken/Nagano
8.	Ontake-san	Chubu	3067	Nagano-ken
9.	Shiomi-dake	Chubu	3047	Shizuoka-ken/Nagano
10.	Senjoga-dake	Chubu	3033	Yamanashi-ken/Nagar

Prominence: 3776 m, 12388 ft

Alternate Name(s)	富士山
Subpeaks	Fuji-san - Hakusandake (3756 m/12,323 ft)
Peak Type	Volcano
Latitude/Longitude (WGS84)	35° 21' 38" N, 138° 43' 38" E 35.360638 138.727347 (Dec Deg) 293513 E 3915408 N, Zone 54 (UTM)
Country	Japan (Highest Point)
State/Province	Yamanashi-ken (Highest Point) Shizuoka-ken (Highest Point)

When extracting the data, you are encouraged to check the first few mountains as a test before applying your code to the entire 100 mountains as it takes a little bit of time. You are also encouraged to print out a check point for every 10 mountains extracted.

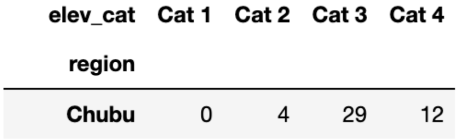
Store all 100 mountains’ info in a DataFrame called q3df. In addition, add a new column called elev_cat which transforms elevations to categorical values as follows:

- If elevation is greater than or equal to 0 and less than 1000, “Cat 1”
- If elevation is greater than or equal to 1000 and less than 2000, “Cat 2”
- If elevation is greater than or equal to 2000 and less than 3000, “Cat 3”
- If elevation is greater than or equal to 3000, “Cat 4”

In q3df, make sure that the elev column is int type, the lat and long columns are float type, and all other columns are str type. The first mountain is presented below for your reference.

name	region	elev	id	lat	long	elev_cat
Fuji-san	Kanto	3776	10882	35.360638	138.727347	Cat 4

Furthermore, create a sample DataFrame called q3sample which shows the total number of mountains in each region (row dimension) and under each category (column dimension). The first row is presented below for your reference.



elev_cat	Cat 1	Cat 2	Cat 3	Cat 4
region				
Chubu	0	4	29	12

Finally, write q3df to a csv file named "Q3_mountains.csv". In **four separate** cells, execute the following codes and show your outputs.

```
q3df.info()
```

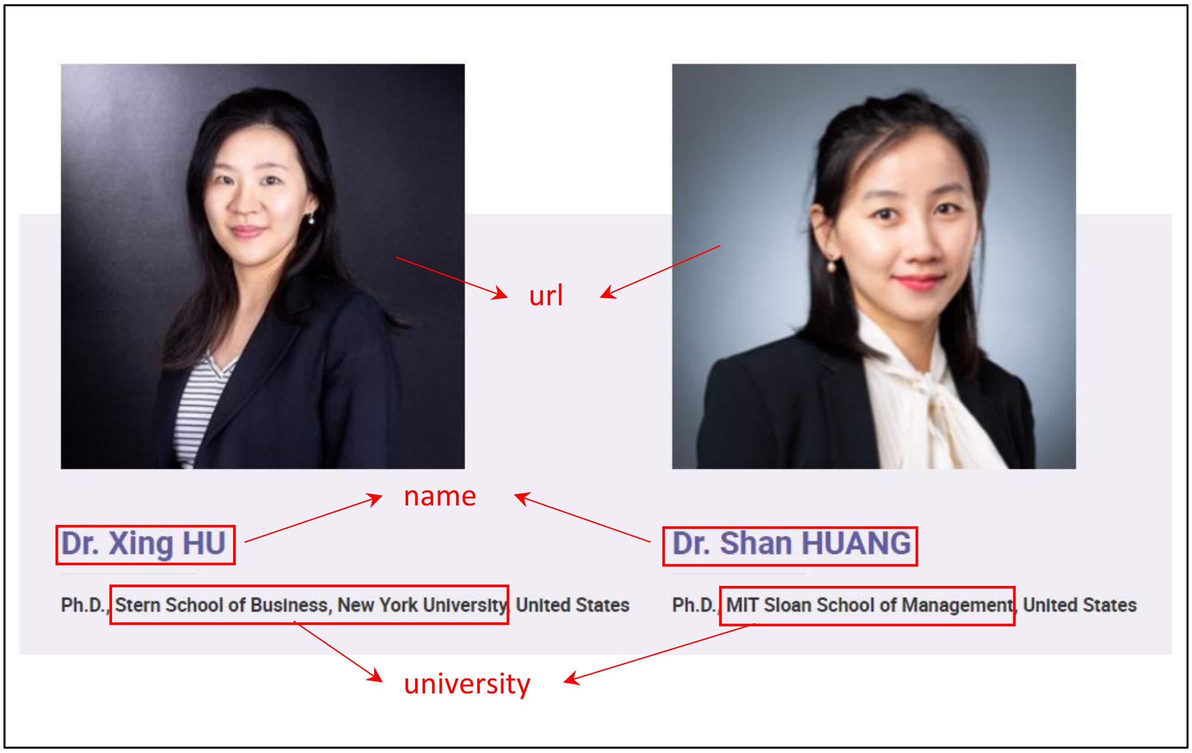
```
q3df.head(3)
```

```
q3df.tail(3)
```

```
q3sample
```

Q4 – MSc(BA) teachers

Build a crawler to extract profiles of all MSc(BA) teachers on the following page: https://msc.hkubs.hku.hk/articles/13?op=10&cd=99



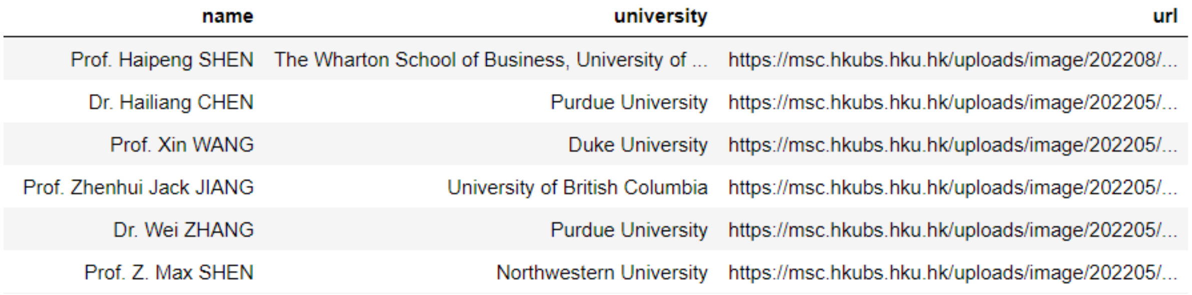
Dr. Xing HU

Ph.D., Stern School of Business, New York University, United States

Dr. Shan HUANG

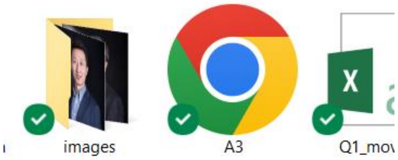
Ph.D., MIT Sloan School of Management, United States

Store all 25 teachers’ info in a DataFrame called q4df. The first few rows are presented below for your reference.



name	university	url
Prof. Haipeng SHEN	The Wharton School of Business, University of ...	https://msc.hkubs.hku.hk/uploads/image/202208/...
Dr. Hailiang CHEN	Purdue University	https://msc.hkubs.hku.hk/uploads/image/202205/...
Prof. Xin WANG	Duke University	https://msc.hkubs.hku.hk/uploads/image/202205/...
Prof. Zhenhui Jack JIANG	University of British Columbia	https://msc.hkubs.hku.hk/uploads/image/202205/...
Dr. Wei ZHANG	Purdue University	https://msc.hkubs.hku.hk/uploads/image/202205/...
Prof. Z. Max SHEN	Northwestern University	https://msc.hkubs.hku.hk/uploads/image/202205/...

In addition, manually create another folder called “images” in the same directory of the codes file. Save all 25 teachers’ profile images in the “images” folder.



How to download and save images

Python treats images as the type of bytes. A byte consists of 8 binary numbers. Therefore, when writing an image, use the mode 'wb' (write binary) and return content from the response (see below).

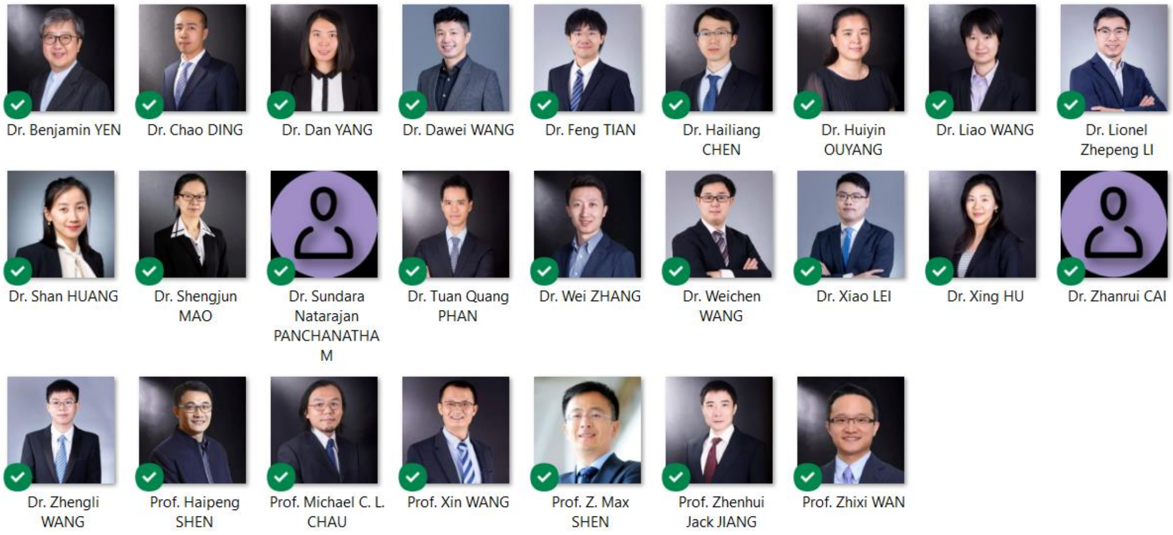
```
import requests

r = requests.get(url)
with open(filepath, 'wb') as handle:
    handle.write(r.content)
```

The image title should have the following structure: name.jpg, where name comes from the name column. For instance:

- Prof. Michael C.L. CHAU.jpg
- Dr. Shan HUANG.jpg

If a teacher does not have a profile image, still download the place holder (see below). Take a screenshot of the images and name it "Q4_screenshot.png" (or .jpg, .jpeg). Save this screenshot in the same folder as your codes file (see Instructions #3).



Finally, write q4df to a csv file named "Q4_teachers.csv". In **two separate** cells, execute the following codes and show your outputs.

```
q4df
```

```
q4df.info()
```