MSBA7001 Exercises II

Module 1, 2024-25
HKU Business School

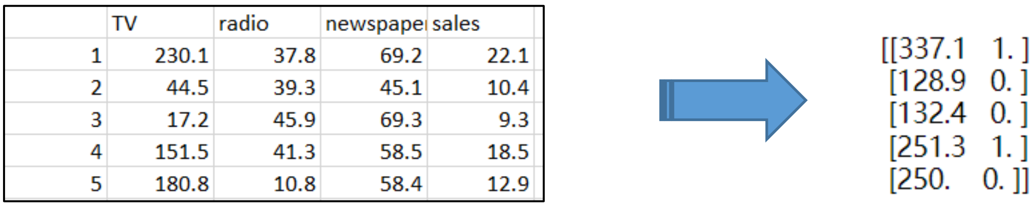
Contents

NumPy and pandas

Exercise – create a new array

Use “advertising.csv”. Create a new array called new where:

- the 1st column is the sum of TV, radio, and newspaper, and
- the 2nd column is either 1 (if sales > mean of sales) or 0 (otherwise).



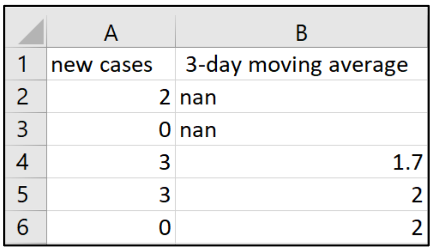
	TV	radio	newspape	sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

```
[[337.1   1. ]
 [128.9   0. ]
 [132.4   0. ]
 [251.3   1. ]
 [250.    0. ]]
```

Print out the first five rows of the array in the answer.

Exercise – moving average

Use “hk_covid_stats.csv”. Calculate the moving average of daily new cases at an interval of three days. Write the result to “moving_ave.csv”. Note that there may be missing values in the data.

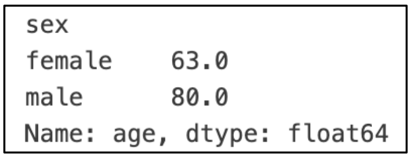


	A	B
1	new cases	3-day moving average
2	2	nan
3	0	nan
4	3	1.7
5	3	2
6	0	2

Exercise – titanic passengers

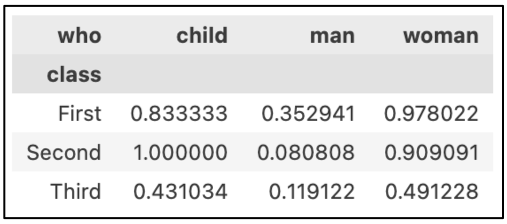
Use “titanic_passengers.csv”. Process the data as follows:

1. Show the oldest man and woman who survived.



```
sex
female    63.0
male      80.0
Name: age, dtype: float64
```

2. Show survival rate by "class" and "who".



who	child	man	woman
class			
First	0.833333	0.352941	0.978022
Second	1.000000	0.080808	0.909091
Third	0.431034	0.119122	0.491228

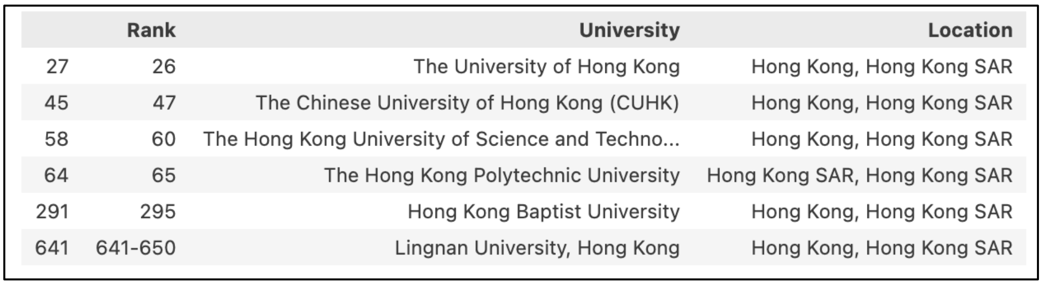
Exercise – university ranking

Extract university ranking from the following page:

https://www.litzusa.com/en-US/StudyusaRecords/detail/Times-Higher-Education-World-University-Ranking-THE

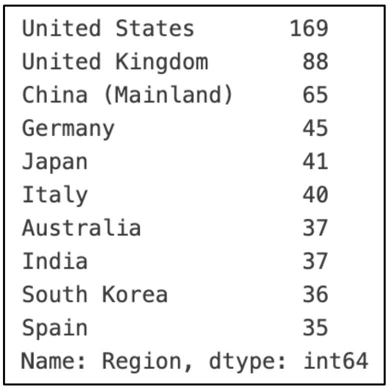
Create a DataFrame based on this ranking and process the data as follows:

1. Show the ranking of universities in Hong Kong.



	Rank	University	Location
27	26	The University of Hong Kong	Hong Kong, Hong Kong SAR
45	47	The Chinese University of Hong Kong (CUHK)	Hong Kong, Hong Kong SAR
58	60	The Hong Kong University of Science and Techno...	Hong Kong, Hong Kong SAR
64	65	The Hong Kong Polytechnic University	Hong Kong SAR, Hong Kong SAR
291	295	Hong Kong Baptist University	Hong Kong, Hong Kong SAR
641	641-650	Lingnan University, Hong Kong	Hong Kong, Hong Kong SAR

2. Show top 10 countries/territories with the highest number of universities.

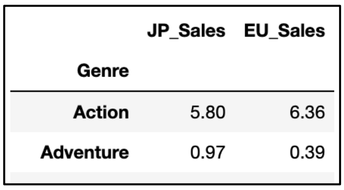


```
United States        169
United Kingdom        88
China (Mainland)      65
Germany               45
Japan                 41
Italy                 40
Australia             37
India                 37
South Korea           36
Spain                 35
Name: Region, dtype: int64
```

Exercise – game sales

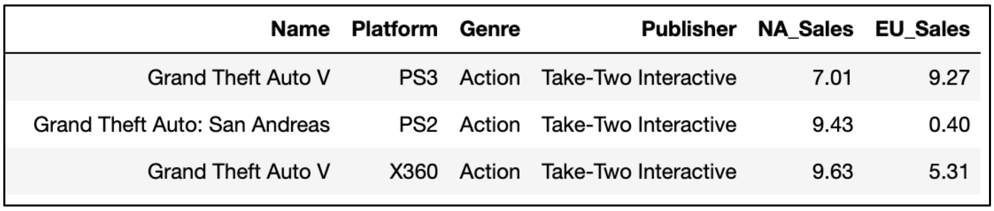
Use "game_sales.csv". Process the data as follows:

1. Show all PS4 games whose names include years, e.g., Just Dance 2016.
2. Show total sales in Japan and EU for each game genre after the year 2015.



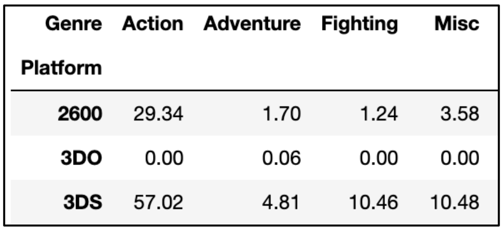
	JP_Sales	EU_Sales
Genre		
Action	5.80	6.36
Adventure	0.97	0.39

3. Show action games whose global sales exceed 10.
4. Create a sample based on #3. Keep NA_Sales and EU_Sales in the sample.



Name	Platform	Genre	Publisher	NA_Sales	EU_Sales
Grand Theft Auto V	PS3	Action	Take-Two Interactive	7.01	9.27
Grand Theft Auto: San Andreas	PS2	Action	Take-Two Interactive	9.43	0.40
Grand Theft Auto V	X360	Action	Take-Two Interactive	9.63	5.31

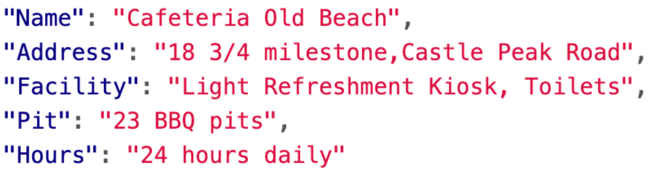
5. Create a pivot table to show the sum of global sales by platform (row dimension) and by genre (column dimension).



Genre	Action	Adventure	Fighting	Misc
Platform				
2600	29.34	1.70	1.24	3.58
3DO	0.00	0.06	0.00	0.00
3DS	57.02	4.81	10.46	10.48

Exercise – barbeque sites

Use "bbq.json". This file includes information about 41 barbeque sites across Hong Kong. See one site's information below:

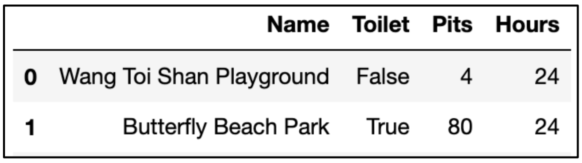


```
"Name": "Cafeteria Old Beach",
"Address": "18 3/4 milestone,Castle Peak Road",
"Facility": "Light Refreshment Kiosk, Toilets",
"Pit": "23 BBQ pits",
"Hours": "24 hours daily"
```

Your job is to extract data from each site as follows:

- Name: retrieve the name of the site, e.g., 'Cafeteria Old Beach'
- Toilet: from 'Facility', find out whether the site has toilet. As long as the word 'Toilet', 'Toilets', 'toilet', or 'toilets' can be found, the Toilet value should be True, otherwise, False. In the case of this site, the value is True
- Pits: retrieve the number of BBQ pits, e.g., 23
- Hours: retrieve the opening hours, e.g., 24

Store the result in a DataFrame called bbqfull. The column names should be Name, Toilet, Pits, Hours. Make sure that the Toilet column is bool type, the Pits and Hours columns are int64 or int type. The first two rows are presented below for your reference.



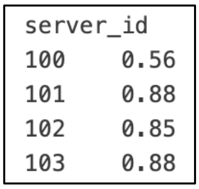
	Name	Toilet	Pits	Hours
0	Wang Toi Shan Playground	False	4	24
1	Butterfly Beach Park	True	80	24

In addition, create a DataFrame called bbqsample which includes only the sites that have toilet and the number of pits greater than or equal to 10. Note that the index of bbqsample needs to be reset. Finally, write bbqfull to "bbq_clean.csv".

Exercise – server utilization

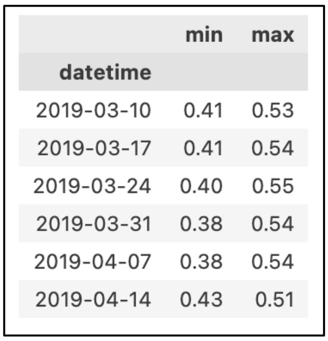
Use "server_utilization.csv". Process the data as follows:

1. Show every server's peak cpu utilization between 7pm and 9pm. For example:



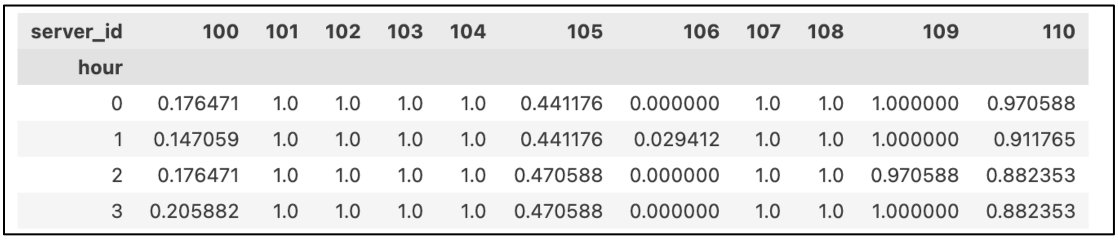
```
server_id
100      0.56
101      0.88
102      0.85
103      0.88
```

2. For server_id 100, show its weekly min and max cpu utilization between 12pm and 6pm.



	min	max
datetime		
2019-03-10	0.41	0.53
2019-03-17	0.41	0.54
2019-03-24	0.40	0.55
2019-03-31	0.38	0.54
2019-04-07	0.38	0.54
2019-04-14	0.43	0.51

3. Create a pivot table to show the proportion of "high" CPU utilization for server_id 100 – 110 and hour. Note that if the utilization is less than 0.5, it's "low", otherwise "high". For example:



server_id	100	101	102	103	104	105	106	107	108	109	110
hour											
0	0.176471	1.0	1.0	1.0	1.0	0.441176	0.000000	1.0	1.0	1.000000	0.970588
1	0.147059	1.0	1.0	1.0	1.0	0.441176	0.029412	1.0	1.0	1.000000	0.911765
2	0.176471	1.0	1.0	1.0	1.0	0.470588	0.000000	1.0	1.0	0.970588	0.882353
3	0.205882	1.0	1.0	1.0	1.0	0.470588	0.000000	1.0	1.0	1.000000	0.882353