



Chapter 4 - exercise 2: Hãy thực hiện những yêu cầu liên quan tới Data Frame như sau:

1. Tạo Data Frame từ dictionary như sau: dic_1 = {'X':[78,85,96,80,86], 'Y':[84,94,89,83,86], 'Z':[86,97,96,72,83]}. In data frame này
2. Tạo Data Frame df2 từ dictionary exam_data, với index của data frame là labels như sau:
exam_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'],

```
'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],  
'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],  
'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}
```

 và labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']

In data frame vừa tạo và in info() của data frame

3. Tạo data frame mới từ data frame df2, chỉ chứa 2 cột là name và score. In type, shape của data frame này. In head của data frame này
4. Tạo data frame mới từ data frame df2, chỉ chứa 2 cột là name và score, và chỉ có các dòng 1, 3, 5, 6. In data frame này
5. Từ data frame df2, In các dòng có dữ liệu null
6. Từ data frame df2, In các dòng có score > 15 và <=20
7. Cập nhật điểm ở dòng 'd' thành 18
8. In thống kê chung của df2
9. Cho biết trong df2 điểm mấy có tần suất xuất hiện nhiều nhất? In những dòng này. Cho biết có bao nhiêu dòng thỏa điều kiện?
10. Thêm dòng k có nội dung như sau: [1, 'Suresh', 'yes', 15.5] vào df2. In tail của df2.
11. Thêm dòng l có nội dung như sau: [2, 'Janny', 'yes', 12.5] vào df2. In tail của df2. Xóa bỏ dòng i của df2. In lại tail.
12. Tạo df3 = Dữ liệu df2 sắp xếp tăng dần theo điểm
13. Thêm cột result căn cứ dữ liệu của cột 'score', nếu >=10 thì giá trị = 1, ngược lại = 0. In df2.
14. Trong df2, thay tên 'Emily' thành 'Samantha'. In df2
15. Duyệt df2, in name, score, result: nếu = 1 thì in 'Pass', ngược lại thì in 'Fail'

```
In [1]: %config IPCompleter.greedy = True  
import numpy as np  
import pandas as pd
```

```
In [2]: # Câu 1: Tạo Data Frame từ dictionary như sau:
'''
dic_1 = {'X':[78,85,96,80,86], 'Y':[84,94,89,83,86], 'Z':[86,97,96,72,83]}
In data frame này
'''
dic_1 = {'X':[78,85,96,80,86], 'Y':[84,94,89,83,86], 'Z':[86,97,96,72,83]}
df1 = pd.DataFrame(dic_1)
df1
```

Out[2]:

	X	Y	Z
0	78	84	86
1	85	94	97
2	96	89	96
3	80	83	72
4	86	86	83

```
In [3]: # Câu 2: Tạo Data Frame df2 từ dictionary exam_data, với index của data frame là L
'''
exam_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'],
'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],
'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],
'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}
labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']
'''
# In data frame vừa tạo và in info() của data frame
exam_data = {'name': ['Anastasia', 'Dima', 'Katherine', 'James', 'Emily', 'Michael', 'Matthew', 'Laura', 'Kevin', 'Jonas'],
'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],
'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],
'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}
labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']

df2 = pd.DataFrame(exam_data, index = labels)
df2
```

Out[3]:

	name	score	attempts	qualify
a	Anastasia	12.5	1	yes
b	Dima	9.0	3	no
c	Katherine	16.5	2	yes
d	James	NaN	3	no
e	Emily	9.0	2	no
f	Michael	20.0	3	yes
g	Matthew	14.5	1	yes
h	Laura	NaN	1	no
i	Kevin	8.0	2	no
j	Jonas	19.0	1	yes



In [4]: df2.info()

```
<class 'pandas.core.frame.DataFrame'>
Index: 10 entries, a to j
Data columns (total 4 columns):
name          10 non-null object
score         8 non-null float64
attempts      10 non-null int64
qualify       10 non-null object
dtypes: float64(1), int64(1), object(2)
memory usage: 400.0+ bytes
```

In [5]: *# Câu 3: Tạo data frame mới từ data frame df2, chỉ chứa 2 cột là name và score.*
In type, shape của data frame này. In head của data frame này

```
df3 = df2[['name', 'score']]
print(type(df3))
print(df3.shape)
```

```
<class 'pandas.core.frame.DataFrame'>
(10, 2)
```

In [6]: df3.head()

Out[6]:

	name	score
a	Anastasia	12.5
b	Dima	9.0
c	Katherine	16.5
d	James	NaN
e	Emily	9.0

In [7]: *# Câu 4: Tạo data frame mới từ data frame df2, chỉ chứa 2 cột là name và score, và*
In data frame này

```
df4 = df2.iloc[[1, 3, 5, 6], [1, 3]]
df4
```

Out[7]:

	score	qualify
b	9.0	no
d	NaN	no
f	20.0	yes
g	14.5	yes



```
In [8]: df4i = df2.loc[["b", "d", "f", "g"], ["name", "score"]]
df4i
```

Out[8]:

	name	score
b	Dima	9.0
d	James	NaN
f	Michael	20.0
g	Matthew	14.5

```
In [9]: # Câu 5: Từ data frame df2, In các dòng có dữ liệu null
df2[df2['score'].isnull()]
```

Out[9]:

	name	score	attempts	qualify
d	James	NaN	3	no
h	Laura	NaN	1	no

```
In [10]: # Câu 6: Từ data frame df2, In các dòng có score > 15 và <=20
df2[(df2['score'] > 15) & (df2['score'] <= 20)]
```

Out[10]:

	name	score	attempts	qualify
c	Katherine	16.5	2	yes
f	Michael	20.0	3	yes
j	Jonas	19.0	1	yes

```
In [11]: # Câu 7: Cập nhật điểm ở dòng 'd' thành 18
df2.loc['d', 'score'] = 18
df2.head()
```

Out[11]:

	name	score	attempts	qualify
a	Anastasia	12.5	1	yes
b	Dima	9.0	3	no
c	Katherine	16.5	2	yes
d	James	18.0	3	no
e	Emily	9.0	2	no



```
In [12]: # Câu 8: In thống kê chung của df2
df2.describe(include = 'all')
```

Out[12]:

	name	score	attempts	qualify
count	10	9.000000	10.000000	10
unique	10	NaN	NaN	2
top	Michael	NaN	NaN	no
freq	1	NaN	NaN	5
mean	NaN	14.055556	1.900000	NaN
std	NaN	4.633064	0.875595	NaN
min	NaN	8.000000	1.000000	NaN
25%	NaN	9.000000	1.000000	NaN
50%	NaN	14.500000	2.000000	NaN
75%	NaN	18.000000	2.750000	NaN
max	NaN	20.000000	3.000000	NaN

```
In [13]: # Câu 9: Cho biết điểm mấy có tần suất xuất hiện nhiều nhất? In những dòng này. Ch
mark = df2['score'].mode()
print(mark[0])
sc1 = df2[(df2['score']==mark[0])]
sc1
```

9.0

Out[13]:

	name	score	attempts	qualify
b	Dima	9.0	3	no
e	Emily	9.0	2	no

```
In [14]: print('Số dòng thỏa điều kiện:')
sc1.shape[0]
```

Số dòng thỏa điều kiện:

Out[14]: 2

```
In [15]: # Câu 10: Thêm dòng k có nội dung như sau: [1, 'Suresh', 'yes', 15.5] vào df2
df2.loc['k'] = ['Suresh', 15.5, 1, 'yes']
df2
```

Out[15]:

	name	score	attempts	qualify
a	Anastasia	12.5	1	yes
b	Dima	9.0	3	no
c	Katherine	16.5	2	yes
d	James	18.0	3	no
e	Emily	9.0	2	no
f	Michael	20.0	3	yes
g	Matthew	14.5	1	yes
h	Laura	NaN	1	no
i	Kevin	8.0	2	no
j	Jonas	19.0	1	yes
k	Suresh	15.5	1	yes

```
In [16]: # Câu 11: Thêm dòng l có nội dung như sau: [2, 'Janny', 'yes', 12.5] vào df2.
# In tail của df2. Xóa bỏ dòng l của df2. In lại tail.
df_temp = pd.DataFrame({'name': ['Janny'],
                        'score': [12.5],
                        'attempts': [2],
                        'qualify': ['yes']}, columns=["name", "score", "attempts", "qualify"])
df_temp
```

Out[16]:

	name	score	attempts	qualify
0	Janny	12.5	2	yes

```
In [17]: df2.append(df_temp)
df2.tail()
```

Out[17]:

	name	score	attempts	qualify
g	Matthew	14.5	1	yes
h	Laura	NaN	1	no
i	Kevin	8.0	2	no
j	Jonas	19.0	1	yes
k	Suresh	15.5	1	yes



```
In [18]: df3 = df2.drop(['k'])
df3.tail()
```

Out[18]:

	name	score	attempts	qualify
f	Michael	20.0	3	yes
g	Matthew	14.5	1	yes
h	Laura	NaN	1	no
i	Kevin	8.0	2	no
j	Jonas	19.0	1	yes

```
In [19]: # Câu 12: Sắp xếp df2 tăng dần theo điểm
df4 = df3
df4['score'] = df4['score'].apply(pd.to_numeric)
df4 = df4.sort_values(by='score')
```

```
In [20]: # Câu 13: Thêm cột result căn cứ dữ liệu của cột 'score', nếu >=10 thì giá trị = 1
df2['result'] = df2['score'].map(lambda x: 1 if x >= 10 else 0)
df2
```

Out[20]:

	name	score	attempts	qualify	result
a	Anastasia	12.5	1	yes	1
b	Dima	9.0	3	no	0
c	Katherine	16.5	2	yes	1
d	James	18.0	3	no	1
e	Emily	9.0	2	no	0
f	Michael	20.0	3	yes	1
g	Matthew	14.5	1	yes	1
h	Laura	NaN	1	no	0
i	Kevin	8.0	2	no	0
j	Jonas	19.0	1	yes	1
k	Suresh	15.5	1	yes	1

In []:



```
In [21]: # Câu 14: Trong df2, thay tên 'Emily' thành 'Samantha'
df2['name'] = df2['name'].replace('Emily', 'Samantha')
df2
```

Out[21]:

	name	score	attempts	qualify	result
a	Anastasia	12.5	1	yes	1
b	Dima	9.0	3	no	0
c	Katherine	16.5	2	yes	1
d	James	18.0	3	no	1
e	Samantha	9.0	2	no	0
f	Michael	20.0	3	yes	1
g	Matthew	14.5	1	yes	1
h	Laura	NaN	1	no	0
i	Kevin	8.0	2	no	0
j	Jonas	19.0	1	yes	1
k	Suresh	15.5	1	yes	1

```
In [22]: # Câu 15: Duyệt df2, in name, score, result: nếu = 1 thì in 'Pass', ngược lại thì
for index, row in df2.iterrows():
    print(row['name'], ' - Score:', row['score'], ' - Result:', "Pass" if row['resu
```

```
Anastasia - Score: 12.5 - Result: Pass
Dima - Score: 9.0 - Result: Fail
Katherine - Score: 16.5 - Result: Pass
James - Score: 18.0 - Result: Pass
Samantha - Score: 9.0 - Result: Fail
Michael - Score: 20.0 - Result: Pass
Matthew - Score: 14.5 - Result: Pass
Laura - Score: nan - Result: Fail
Kevin - Score: 8.0 - Result: Fail
Jonas - Score: 19.0 - Result: Pass
Suresh - Score: 15.5 - Result: Pass
```

In []:

