PANDAS FOR DATA SCIENCE

MILS BURASAKORN

CR. CHRIS PHILIPS ON BEHANCE

SOFTWARE DEVELOPER

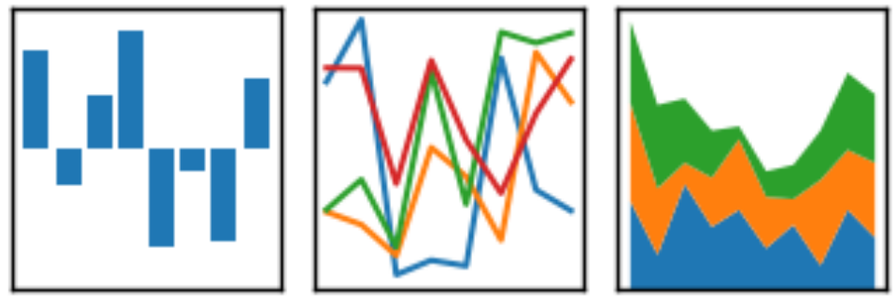
@ PRONTO TOOLS

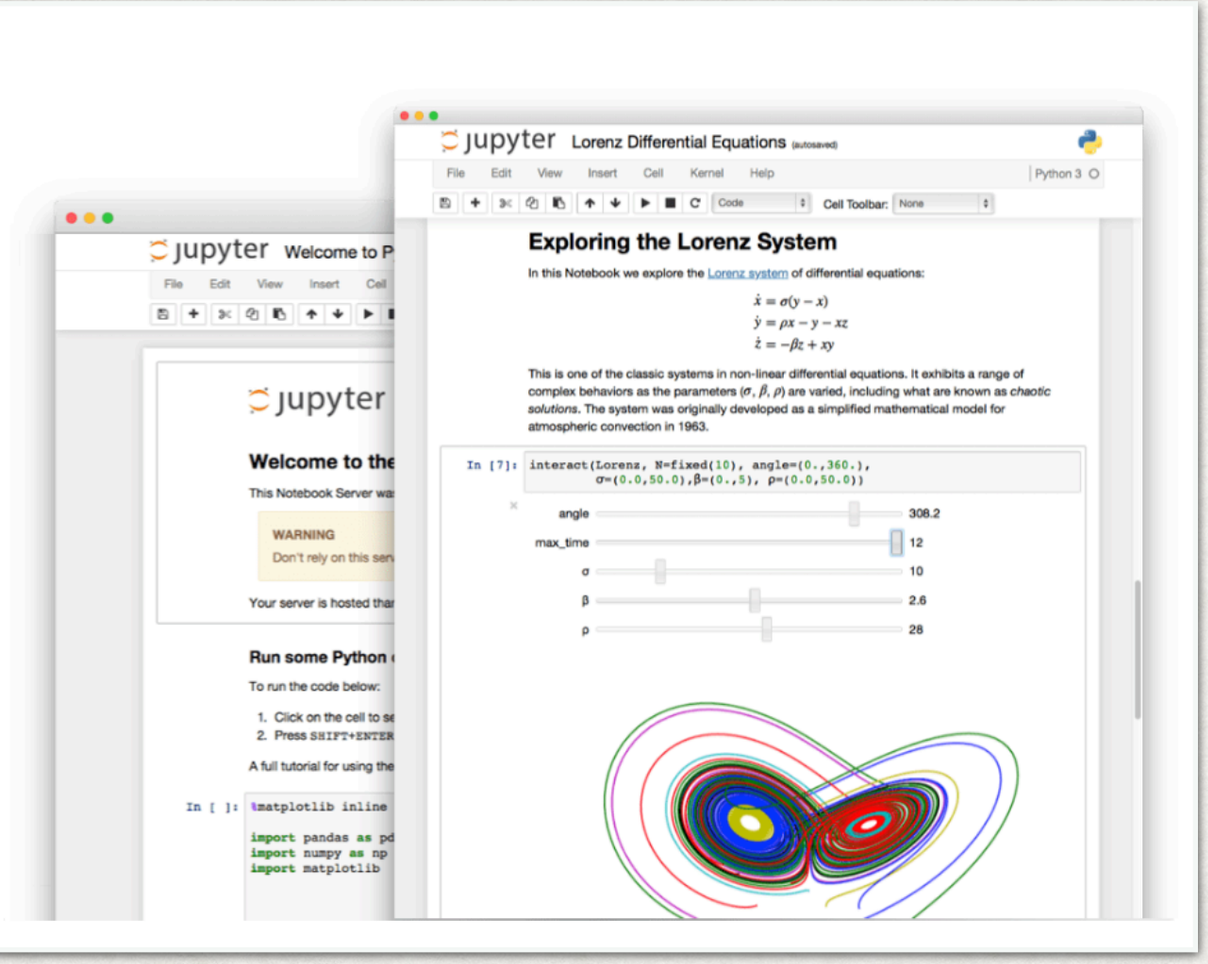
WHAT NORMAL PEOPLE THINK OF

WHAT DATA SCIENTISTS THINK OF

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

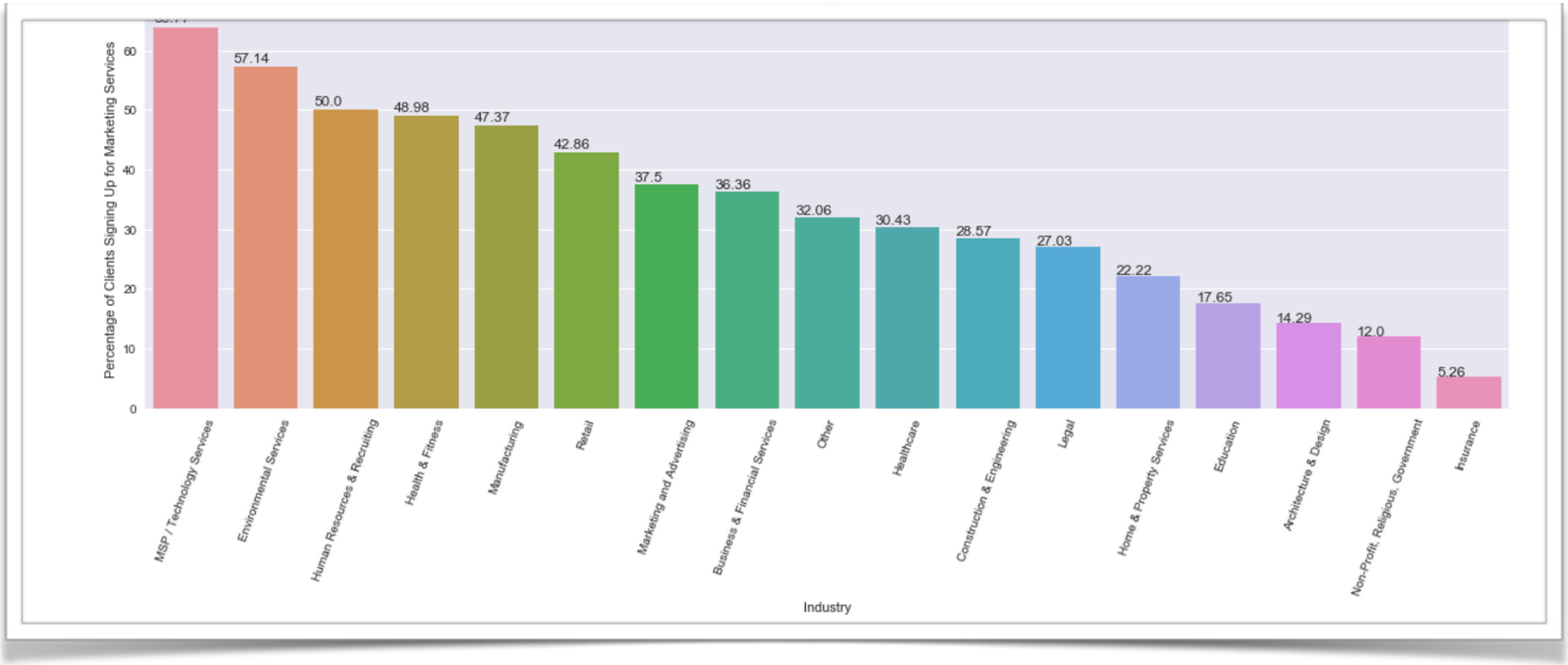


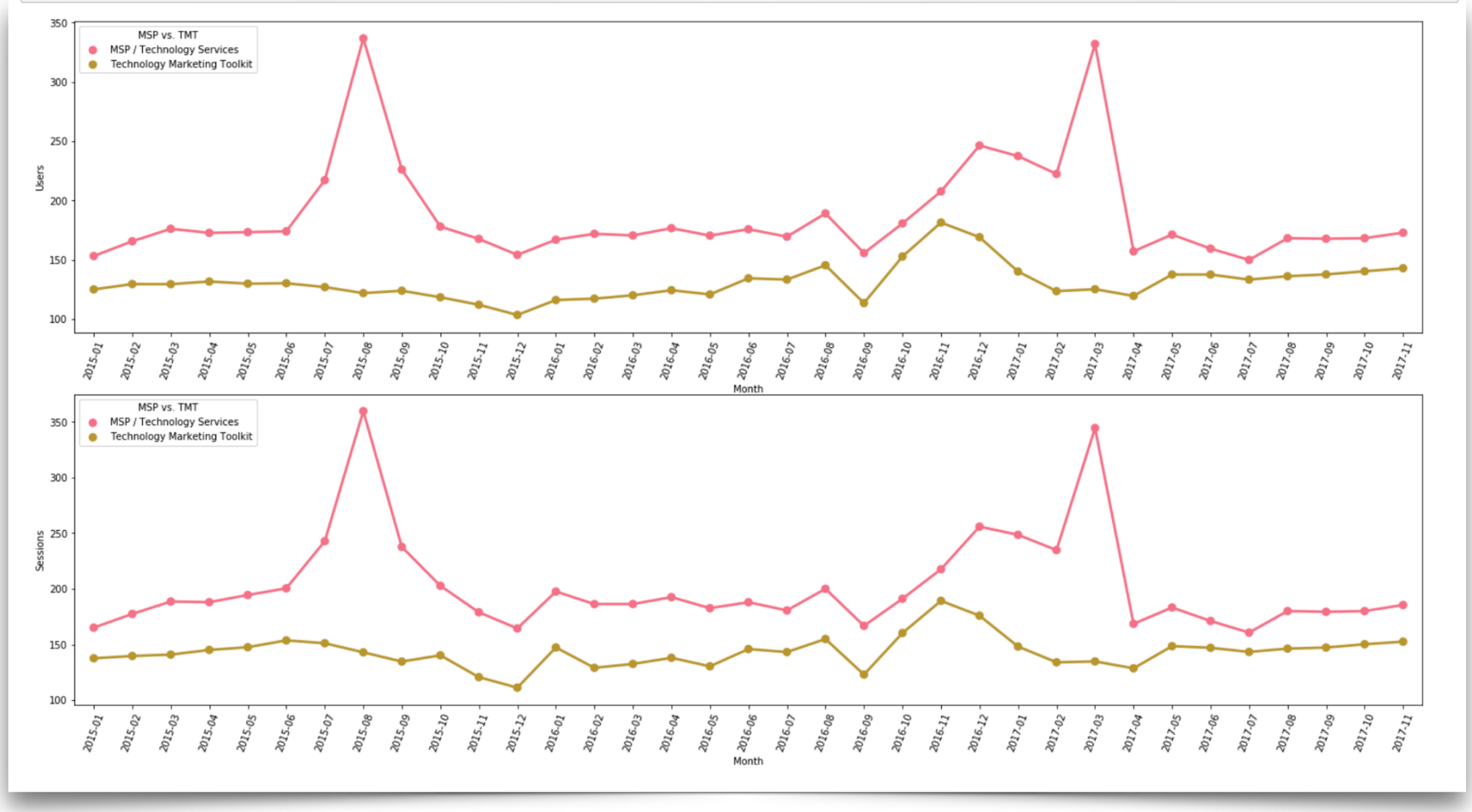


WHAT IS PANDAS?

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use **data structures** and **data analysis tools** for the Python programming language.


ONE DOES NOT SIMPLY
SAY NO TO PANDA
quickmeme.com


Percentage of Clients Signing Up for Marketing Services
0
10
20
30
40
50
60
57.14
50.0
48.98
47.37
42.86
37.5
36.36
32.06
30.43
28.57
27.03
22.22
17.65
14.29
12.0
5.26
MSP / Technology Services
Environmental Services
Human Resources & Recruiting
Health & Fitness
Manufacturing
Retail
Marketing and Advertising
Business & Financial Services
Other
Healthcare
Construction & Engineering
Legal
Home & Property Services
Education
Architecture & Design
Non-Profit, Religious, Government
Insurance
Industry


MSP vs. TMT
MSP / Technology Services
Technology Marketing Toolkit
Users
100
150
200
250
300
350
2015-01
2015-02
2015-03
2015-04
2015-05
2015-06
2015-07
2015-08
2015-09
2015-10
2015-11
2015-12
2016-01
2016-02
2016-03
2016-04
2016-05
2016-06
2016-07
2016-08
2016-09
2016-10
2016-11
2016-12
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
2017-07
2017-08
2017-09
2017-10
2017-11
Month
MSP vs. TMT
MSP / Technology Services
Technology Marketing Toolkit
Sessions
100
150
200
250
300
350
2015-01
2015-02
2015-03
2015-04
2015-05
2015-06
2015-07
2015-08
2015-09
2015-10
2015-11
2015-12
2016-01
2016-02
2016-03
2016-04
2016-05
2016-06
2016-07
2016-08
2016-09
2016-10
2016-11
2016-12
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
2017-07
2017-08
2017-09
2017-10
2017-11
Month

WHAT IS PANDAS?

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use **data structures** and **data analysis tools** for the Python programming language.

DATA STRUCTURE

Based on Python

1. Python List

data = [176, 308, 890, 63]

data[2]

2. Python Dict

data = { 'name': ‘Kan’, ‘age’: ’36' }

data[‘age’]

3. Series

```
>> import pandas as pd
>> pd.Series()
```

```
data = [176, 308, 890, 63]
```

Q: series_data[2] ?

pd.Series(data)

0	176
1	308
2	890
3	63

```
data = {'a': 176, 'b': 308, 'c': 890, 'd': 63}
```

Q: series_data['c'] ?

Q: series_data[2] ?

a	176
b	308
c	890
d	63

SERIES

```
data = {'a': 176, 'b': 308, 'c': 890, 'd': 63}

index = ['b', 'c', 'd', 'e', 'f']
```

```
pd.Series(data, index=index)
```

b	308
c	890
d	63
e	NaN
f	NaN

SERIES

```
pd.Series(data, index=index)
```

```
series_data.index
```

```
Index(['b', 'c', 'd', 'e', 'f'], dtype='object')
```

```
series_data.values
```

```
array([308.,   890.,   63.,   nan,   nan])
```

b	308
c	890
d	63
e	NaN
f	NaN

Q: series_data.append(pd.Series([19, 27, 54, 3])) ?

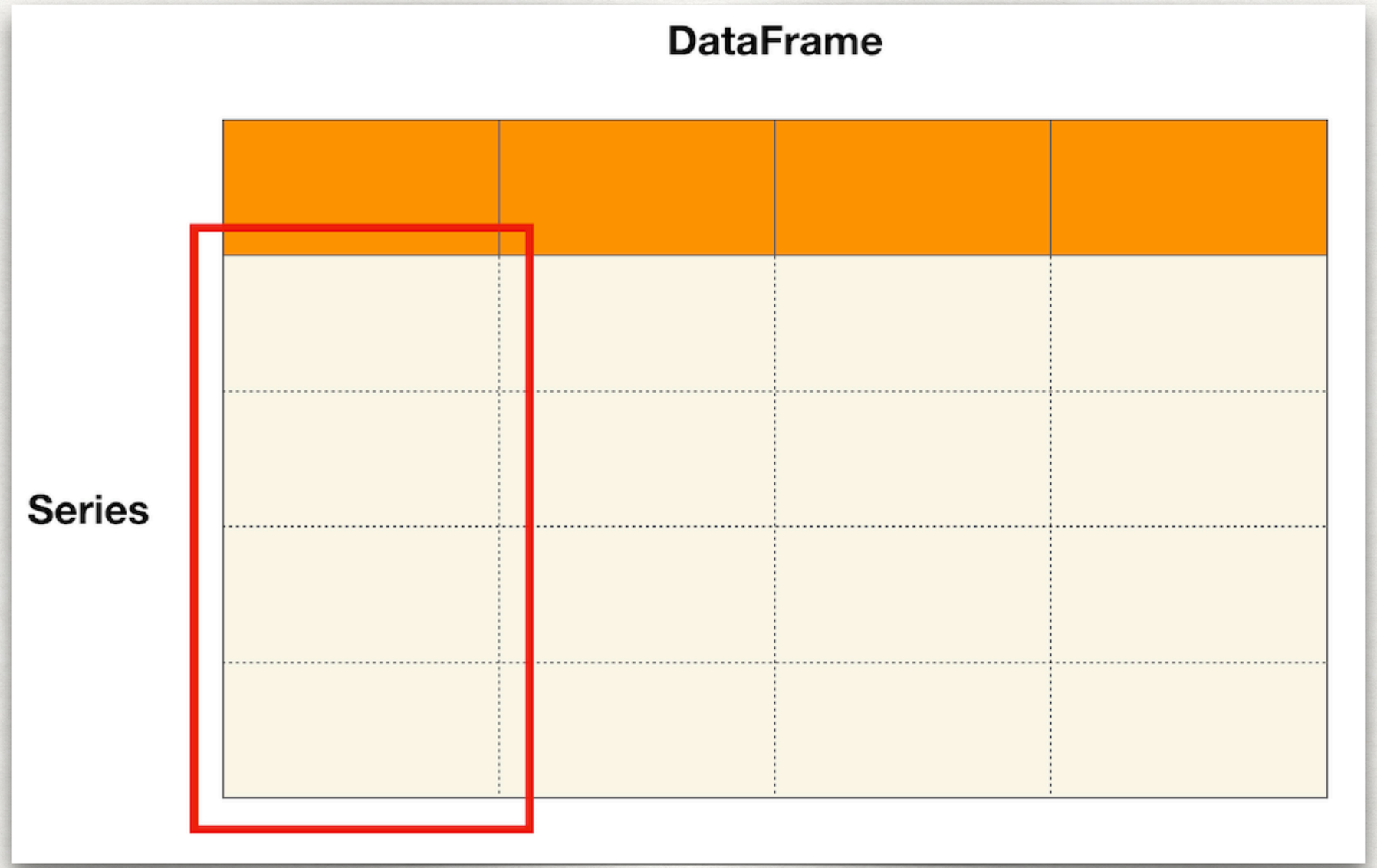
SERIES

Q: series_data.append(pd.Series([19, 27, 54, 3])) ?

QUIZ - SERIES

กดสร้าง New Notebook แล้วลองสร้าง Series จาก data ต่อไปนี้

```
food_data_dict = {
    'Germany': 'sauerkraut',
    'Spain': 'paella',
    'Italy': 'pizza',
    'USA': 'Hamburger'
}
```


DataFrame
Series

DATAFRAME

```
personal_data_dict = {
    'age': [39, 50, 38],
    'education': ['Bachelors', 'Bachelors', 'HS-grad'],
    'occupation': ['Adm-clerical', 'Tech-support', 'Sales'],
    'sex': ['Male', 'Female', 'Female'],
    'capital-gain': [2174, 111, 993]
}
```

```
df = pd.DataFrame(personal_data_dict)
```

	age	education	occupation	sex	capital-gain
0	39	Bachelors	Adm-clerical	Male	2174
1	50	Bachelors	Tech-support	Female	111
2	38	HS-grad	Sales	Female	993

DATAFRAME

```
df = pd.DataFrame(personal_data_dict)
```

	age	education	occupation	sex	capital-gain
0	39	Bachelors	Adm-clerical	Male	2174
1	50	Bachelors	Tech-support	Female	111
2	38	HS-grad	Sales	Female	993

```
df = pd.index
```

```
df = pd.values
```

```
df = pd.education
```

```
df = pd['education']
```

QUIZ - DATAFRAME

ลองสร้าง Dataframe **จาก** data **ต่อไปนี้**

```
zoo_data_dict = {
    'animal': ['elephant', 'lion', 'kangaroo'],
    'uniq_id': [1001, 1014, 1902],
    'water_need': [500, 120, 200],
}
```

2. GETTING THE DATA

1. CSV file
2. Excel file
3. JSON file
4. URL

etc.

2. WRITING THE DATA

1. CSV file
2. Excel file
3. JSON file

etc.