



Intermediate Pandas 1

☰ Tags	Pandas Python
↗ Class	
☑ Finished Yet?	☑
↗ Knowledge	🐼 <u>The Ninth Sprint: Essential Python for Data Science</u>

-Palmer Penguins Dataset:

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/df95dd0f-2134-4936-81cd-68c04cd60a57/penguins.csv>

-ใน Datalore สามารถปรับให้เป็น Dark Mode เพื่อความสบายตาได้ที่ View → Dark Mode

*Intermediate Pandas 1 และ 2 จะสลับไปใช้ DataLore แทน Google Colab

Lesson 21: Preview Penguins Data Frame

-Dataset penguins.csv ประกอบด้วยข้อมูลเกี่ยวกับเพนกวิน 3 สายพันธุ์

-เราสามารถใช่ DataFrame.head() ในการดู 5 แถวแรกของ Data Frame

-ค่า nan = Missing Value

```
[1] import pandas as pd

[2] penguins = pd.read_csv("penguins.csv")

[3] ▶ 0.1s
#Preview first 5 rows
penguins.head()
```

	species	island	bill_length...	bill_depth_mm	flipper_lengt...	body_mass
0	Adelie	Torgersen	39.1	18.7	181.0	
1	Adelie	Torgersen	39.5	17.4	186.0	
2	Adelie	Torgersen	40.3	18.0	195.0	
3	Adelie	Torgersen	nan	nan	nan	
4	Adelie	Torgersen	36.7	19.3	193.0	

5 rows x 8 columns

-เราสามารถใช่ DataFrame.tail() ในการดู 5 แถวสุดท้ายของ Data Frame

```
[4] ▶ 0.1s
#Preview last 5 rows
penguins.tail()
```

	species	island	bill_length...	bill_depth_mm	flipper_lengt...	body_mass_g
339	Gentoo	Biscoe	nan	nan	nan	nan
340	Gentoo	Biscoe	46.8	14.3	215.0	4850.0
341	Gentoo	Biscoe	50.4	15.7	222.0	5750.0
342	Gentoo	Biscoe	45.2	14.8	212.0	5200.0
343	Gentoo	Biscoe	49.9	16.1	213.0	5400.0

5 rows x 8 columns

-เราสามารถเช็คจำนวน row และ column ของ Data Frame ได้ด้วย DataFrame.shape

```
[5] ▶ 0.1s
#rows, columns
penguins.shape

(344, 7)
```

-เราสามารถเช็คข้อมูลของ Data Frame ได้ด้วย DataFrame.info()

```
[6] > 0.1s
#Penguins info
penguins.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 344 entries, 0 to 343
Data columns (total 7 columns):
#   Column                Non-Null Count  Dtype
---  -
0   species               344 non-null   object
1   island                344 non-null   object
2   bill_length_mm        342 non-null   float64
3   bill_depth_mm         342 non-null   float64
4   flipper_length_mm     342 non-null   float64
5   body_mass_g           342 non-null   float64
6   sex                   333 non-null   object
dtypes: float64(4), object(3)
memory usage: 18.9+ KB
```

Lesson 22: Select columns

-เราสามารถดึง column ได้ด้วยชื่อของ column หรือจะใช้ Dot Notation ก็ได้เช่นกัน ได้คำตอบเหมือนกัน (DataFrame.ColumnName) และอาจบวก .head() เพื่อดูเฉพาะ 5 แถวแรกสุดก็ได้เช่นกัน

[8] ▶ 0.1s

```
#Select columns
penguins["species"]
```

Table Raw Visualize Statistics

	Ab species ▼
0	Adelie
1	Adelie
2	Adelie
3	Adelie
4	Adelie
5	Adelie
6	Adelie
7	Adelie
8	Adelie

[9] ▶ 0.1s

penguins.species

Table Raw Visualize Statistics

	Ab species ▼
0	Adelie
1	Adelie
2	Adelie
3	Adelie
4	Adelie
5	Adelie
6	Adelie
7	Adelie
8	Adelie

-เราสามารถดึง column ใน Data Frame หลาย ๆ column พร้อมกันได้ด้วยการพิมพ์ชื่อ column ลงไปใน List

0.1s

```
penguins[["species", "island", "sex"]].head()
```

Table Raw Visualize Statistics

	Ab	species	Ab	island	Ab	sex
0		Adelie		Torgersen		MALE
1		Adelie		Torgersen		FEMALE
2		Adelie		Torgersen		FEMALE
3		Adelie		Torgersen		nan
4		Adelie		Torgersen		FEMALE

-function head() และ tail() สามารถใส่จำนวนในวงเล็บเพื่อระบุจำนวนแถวที่ต้องการดึงได้

Lesson 23: ILOC (Integer-location Based Indexing)

-iloc ใช้ในการหาข้อมูลอิงจากตำแหน่ง index เวลาดึงข้อมูล จะดึงเป็น Series ออกมา

[11] 0.1s

```
#iloc (integer-location based indexing for selection by position)
penguins.iloc[0]
```

Table Raw Visualize Statistics

	Ab 0 ▾
spe...	Adelie
isl...	Torgersen
bil...	39.1
bil...	18.7
fli...	181.0
bod...	3750.0
sex	MALE

-เราสามารถดึงออกมาได้หลายแถวในเวลาเดียวกัน

[12] 0.1s

```
penguins.iloc[ [0, 1, 2] ]
```

Table Raw Visualize Statistics

	Ab species ▾	Ab island ▾	bill_length... ▾	bill_depth_mm ▾	flipper_lengt... ▾	body_mass_g ▾
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0

-ถ้าจะดึงแถวตั้งแต่ 0 ถึง n ให้พิมพ์ DataFrame.iloc[0:n+1] แทน เพราะ Python จะถึงถึง n แต่ไม่นับแถว n เช่น อยากได้แถว 0 ถึง 2 จาก penguins ก็พิมพ์ penguins.iloc[0:3] เป็นต้น

[15] ▶ 0.1s

```
penguins.iloc[ 0:3 ]
```

	species	island	bill_length...	bill_depth_mm	flipper_lengt...	body_mass_g
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0

3 rows x 8 columns

-iloc ใช้ดึง column ได้ด้วย หลักการเหมือนกับการดึง row

[15] ▶ 0.1s

```
penguins.iloc[ 0:5 , [0, 1, 2] ]
```

	species	island	bill_length...
0	Adelie	Torgersen	39.1
1	Adelie	Torgersen	39.5
2	Adelie	Torgersen	40.3
3	Adelie	Torgersen	nan
4	Adelie	Torgersen	36.7

5 rows x 4 columns

-เราสามารถเก็บตารางที่ถูก Subset ด้วย iloc ไว้ในตัวแปรได้

[18] ▶ 0.1s

```
mini_penguins = penguins.iloc[ 0:5, 0:3]
mini_penguins
```

Table Raw Visualize Statistics

	Ab species ▼	Ab island ▼	.3 bill_length_... ▼
0	Adelie	Torgersen	39.1
1	Adelie	Torgersen	39.5
2	Adelie	Torgersen	40.3
3	Adelie	Torgersen	nan
4	Adelie	Torgersen	36.7

5 rows x 4 columns

[Jump to top](#)
[Jump to bottom](#)

Lesson 24: Filter Data Frame with One Condition

-เราสามารถกรอง Data Frame ของเราได้ด้วยการกำหนดเงื่อนไข สมมติว่า เราต้องการข้อมูลจาก penguins ที่อยู่ในเกาะ Torgersen ถ้าค่าความจริงเป็น True เท่ากับ island = Torgersen แต่ถ้าเป็น False ก็คืออยู่ในเกาะอื่น

[20] ▶ 0.1s

#Filter DataFrame
penguins["island"] == "Torgersen"

Table Raw Visualize Statistics

	Ab	island	▼
0		True	
1		True	
2		True	
3		True	
4		True	
5		True	
6		True	
7		True	
8		True	

-เราสามารถหาค่าความจริงเหล่านี้ในการกรอง Data Frame ของเราได้

[21] 0.1s

```
#Filter DataFrame
penguins[ penguins["island"] == "Torgersen" ]
```

Table Raw Visualize Statistics

	Ab	species	Ab	island	.3	bill_length...
0		Adelie		Torgersen		39.1
1		Adelie		Torgersen		39.5
2		Adelie		Torgersen		40.3
3		Adelie		Torgersen		nan
4		Adelie		Torgersen		36.7
5		Adelie		Torgersen		39.3
6		Adelie		Torgersen		38.9
7		Adelie		Torgersen		39.2
8		Adelie		Torgersen		34.1

-เราสามารถตั้งเงื่อนไขเป็นตัวเลขได้เช่นกัน เช่น กรองเฉพาะเพนกวินที่มีความยาวจะงอยไม่เกิน 34 มม.

[23] 0.1s

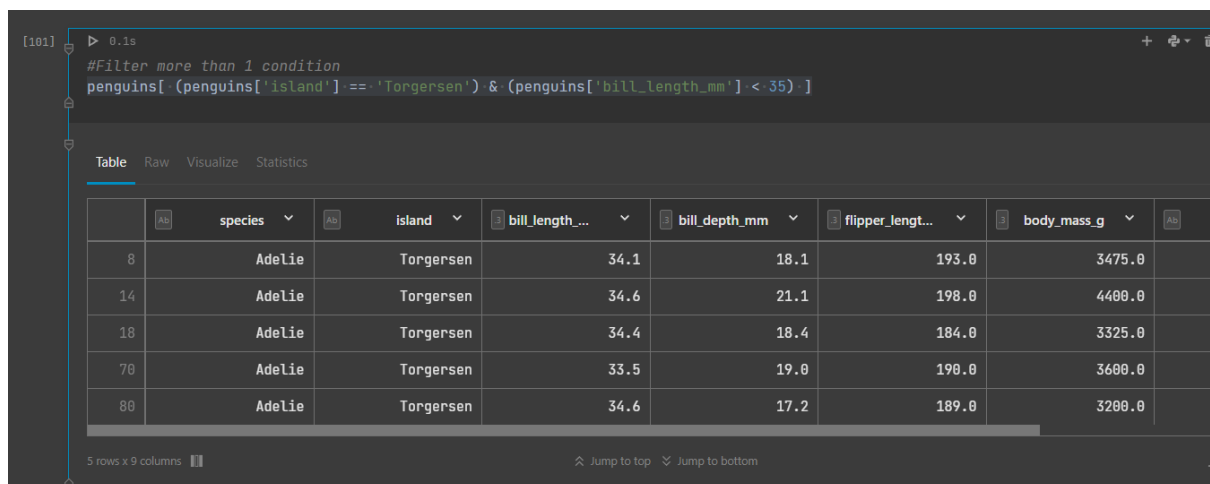
```
penguins[ penguins['bill_length_mm'] < 34 ]
```

Table Raw Visualize Statistics

	Ab	species	Ab	island	.3	bill_length...	.3	bill_depth_mm
70		Adelie		Torgersen		33.5		19.0
98		Adelie		Dream		33.1		16.1
142		Adelie		Dream		32.1		15.5

Lesson 25: Filter Data Frame more than One Condition

-เราสามารถกรอง Data Frame ได้มากกว่า 1 Condition ด้วยสัญลักษณ์ตัวเชื่อม (ใช้ได้ทั้ง & [AND] และ | [OR])

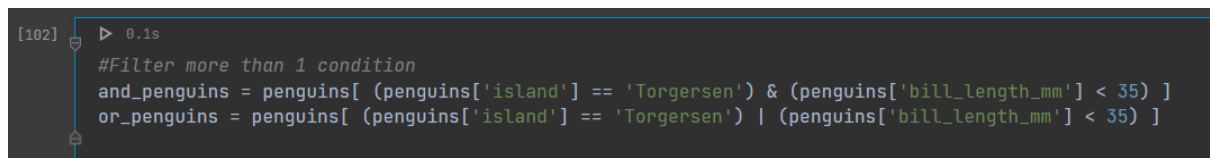


```
[101] ▶ 0.1s
#Filter more than 1 condition
penguins[ (penguins['island'] == 'Torgersen') & (penguins['bill_length_mm'] < 35) ]
```

	species	island	bill_length...	bill_depth_mm	flipper_lengt...	body_mass_g
8	Adelie	Torgersen	34.1	18.1	193.0	3475.0
14	Adelie	Torgersen	34.6	21.1	198.0	4400.0
18	Adelie	Torgersen	34.4	18.4	184.0	3325.0
70	Adelie	Torgersen	33.5	19.0	190.0	3600.0
80	Adelie	Torgersen	34.6	17.2	189.0	3200.0

5 rows x 9 columns

-แต่ละเงื่อนไขต้องอยู่ในวงเล็บ



```
[102] ▶ 0.1s
#Filter more than 1 condition
and_penguins = penguins[ (penguins['island'] == 'Torgersen') & (penguins['bill_length_mm'] < 35) ]
or_penguins = penguins[ (penguins['island'] == 'Torgersen') | (penguins['bill_length_mm'] < 35) ]
```

Lesson 26: Query

-เราสามารถกรองข้อมูลด้วยเงื่อนไขหลายเงื่อนไขได้ด้วย .query()

[25] 0.1s

```
#Filter with .query()
penguins.query('island == "Torgersen"') #island == 'Torgersen'
```

Table Raw Visualize Statistics

	Ab	species	Ab	island	bill_length...	bill_dept
0		Adelie		Torgersen	39.1	
1		Adelie		Torgersen	39.5	
2		Adelie		Torgersen	40.3	
3		Adelie		Torgersen	nan	
4		Adelie		Torgersen	36.7	
5		Adelie		Torgersen	39.3	
6		Adelie		Torgersen	38.9	
7		Adelie		Torgersen	39.2	
8		Adelie		Torgersen	34.1	

-กรองหลายเงื่อนไข (ถ้าเป็น & = AND, | = OR)

[27] 0.1s

```
penguins.query('island == "Torgersen" & bill_length_mm < 35')
```


Table Raw Visualize Statistics

	Ab	species	Ab	island	bill_length...	bill_depth_mm	t...
8		Adelie		Torgersen	34.1	18.1	
14		Adelie		Torgersen	34.6	21.1	
18		Adelie		Torgersen	34.4	18.4	
70		Adelie		Torgersen	33.5	19.0	
80		Adelie		Torgersen	34.6	17.2	

Lesson 27: Missing Values

-Missing value ใน Python จะเรียกว่าค่า nan เราสามารถใช้ .isna() ในการเช็คข้อมูลตรงไหนเป็น nan ได้ (True = nan, False = ไม่ใช่ nan)

```
[31] ▶ 0.1s
#Check missing values in each column
penguins.isna().sum()
```



	3
species	0
island	0
bill_length_mm	2
bill_depth_mm	2
flipper_length_mm	2
body_mass_g	2
sex	11

-เราสามารถกรอง row ที่ column ชื่อ sex มี nan ได้ดังนี้

```
penguins[ penguins['sex'].isna() ]
```

```
[33] > 0.1s
#Filter missing values in column 'sex'
penguins[ penguins['sex'].isna() ]
```

species	island	bill_length...	bill_depth_mm	flipper_lengt...	body_mass_g	sex
Adelie	Torgersen	nan	nan	nan	nan	nan
Adelie	Torgersen	34.1	18.1	193.0	3475.0	nan
Adelie	Torgersen	42.0	20.2	190.0	4250.0	nan
Adelie	Torgersen	37.8	17.1	186.0	3300.0	nan
Adelie	Torgersen	37.8	17.3	180.0	3700.0	nan
Adelie	Dream	37.5	18.9	179.0	2975.0	nan
Gentoo	Biscoe	44.5	14.3	216.0	4100.0	nan
Gentoo	Biscoe	46.2	14.4	214.0	4650.0	nan
Gentoo	Biscoe	47.3	13.8	216.0	4725.0	nan
Gentoo	Biscoe	44.5	15.7	217.0	4875.0	nan
Gentoo	Biscoe	nan	nan	nan	nan	nan

11 rows x 8 columns

-เราสามารถ Drop row ที่มี nan ก็ได้ด้วย .dropna()

```
[34] > 0.1s
#Drop missing values
clean_penguins = penguins.dropna()

clean_penguins
```

species	island	bill_length...	bill_depth_mm	flipper_lengt...	body_mass_g	sex
---------	--------	----------------	---------------	------------------	-------------	-----

Table Raw Visualize Statistics

Lesson 28: Fill Missing Values

-นอกจากการ Drop row ที่มี nan ที่ทิ้งไปแล้ว เราสามารถแทนที่ค่า nan ได้ด้วย Mean Imputation (แทนที่ nan ด้วย mean ของ column นั้น ๆ)

[36]

```
##Mean imputation
#Fill missing value
top5_penguins = penguins.head(5)
```

[37]

```
#Find mean
avg_value = top5_penguins['bill_length_mm'].mean()
print(avg_value)
```

38.9

[38]

```
▶ 0.1s
#Fill missing value with avg_value
top5_penguins = top5_penguins['bill_length_mm'].fillna(value = avg_value)
top5_penguins
```

Table Raw Visualize Statistics

	bill_length_...	
0	39.1	
1	39.5	
2	40.3	

Lesson 29: Sort Data Frame

-เราสามารถใช้ `.sort_values('column ที่ต้องการอ้างอิง')` ในการจัดเรียง row ในตารางอิงจากค่าใน column นั้น ๆ น้อยไปมาก หรือมากไปน้อยก็ได้ [Default = น้อยไปมาก]

▶ 0.1s

```
#Sort dataframe
penguins.sort_values('bill_length_mm')
```

Table Raw Visualize Statistics

	Ab species ▼	Ab island ▼	.3 bill_length_... ▼	.3 bil
142	Adelie	Dream	32.1	
98	Adelie	Dream	33.1	
70	Adelie	Torgersen	33.5	
92	Adelie	Dream	34.0	
8	Adelie	Torgersen	34.1	
18	Adelie	Torgersen	34.4	
54	Adelie	Biscoe	34.5	
80	Adelie	Torgersen	34.6	
14	Adelie	Torgersen	34.6	

-ถ้าอยากให้เรียงมากไปน้อย ให้เพิ่ม parameter ดังนี้: ascending = False

[41] 0.1s

```
#Sort dataframe
penguins.sort_values('bill_length_mm', ascending = False)
```

Table Raw Visualize Statistics

	Ab species ▾	Ab island ▾	.3 bill_length_... ▾	.3 bill
253	Gentoo	Biscoe	59.6	
169	Chinstrap	Dream	58.0	
321	Gentoo	Biscoe	55.9	
215	Chinstrap	Dream	55.8	
335	Gentoo	Biscoe	55.1	
283	Gentoo	Biscoe	54.3	
183	Chinstrap	Dream	54.2	
191	Chinstrap	Dream	53.5	
327	Gentoo	Biscoe	53.4	

-เราสามารถ Chain Function ของเราได้ เช่น dropna ตามด้วย sort และแสดงผล head(10) เป็นต้น

[42] 0.1s

```
#Sort dataframe
penguins.dropna().sort_values('bill_length_mm', ascending = False).head(10)
```

Table Raw Visualize Statistics

	species	island	bill_length...	bill_depth_mm	fl
253	Gentoo	Biscoe	59.6	17.0	
169	Chinstrap	Dream	58.0	17.8	
321	Gentoo	Biscoe	55.9	17.0	
215	Chinstrap	Dream	55.8	19.8	
335	Gentoo	Biscoe	55.1	16.0	
283	Gentoo	Biscoe	54.3	15.7	
183	Chinstrap	Dream	54.2	20.8	
191	Chinstrap	Dream	53.5	19.9	
327	Gentoo	Biscoe	53.4	15.8	
181	Chinstrap	Dream	52.8	20.0	

-เราสามารถ Sort ได้มากกว่า 1 column

0.1s

sort multiple columns

```
penguins.dropna().sort_values(['island', 'bill_length_mm'])
```

Table Raw Visualize Statistics

	species	island	bill_length...	bill
54	Adelie	Biscoe	34.5	
52	Adelie	Biscoe	35.0	
100	Adelie	Biscoe	35.0	
25	Adelie	Biscoe	35.3	
66	Adelie	Biscoe	35.5	
60	Adelie	Biscoe	35.7	
22	Adelie	Biscoe	35.9	
64	Adelie	Biscoe	36.4	
58	Adelie	Biscoe	36.5	

*ถ้าเจอ Function ที่ไม่คุ้นเคย ลองเสิร์ช Google หรืออ่านใน Documentation เพื่อทำให้เข้าใจหน้าที่การทำงานของ Function ได้ดียิ่งขึ้น

Lesson 30: Unique and Count

-เราสามารถหาค่าที่ไม่ซ้ำกัน (Unique) ได้ด้วย .unique()

```
[45] ▶ 0.1s
#Unique values
penguins['species'].unique()
```

Table Raw Visualize Statistics

	Ab	▼
0	Adelie	
1	Chinstrap	
2	Gentoo	

-เราสามารถนับค่าที่ไม่ซ้ำกันได้ด้วย .value_counts()

```
[46] ▶ 0.1s
#Count unique values
penguins['species'].value_counts()
```

Table Raw Visualize Statistics

	.3	species	▼
Adelie	152		
Gentoo	124		
Chinstrap	68		

-เราสามารถ Count ได้มากกว่า 1 column (Multi-index)

[48] ▶ 0.1s

```
#Count more than one column
penguins[ ['island', 'species'] ].value_counts().reset_index()
```

Table Raw Visualize Statistics

	Ab	island	Ab	species	.3	0
0		Biscoe		Gentoo		124
1		Dream		Chinstrap		68
2		Dream		Adelie		56
3		Torgersen		Adelie		52
4		Biscoe		Adelie		44

-เราสามารถเปลี่ยนชื่อของ column ได้ด้วยการฝากค่าไว้ในตัวแปร แล้วพิมพ์ชื่อ column ใหม่กับชื่อเก่า

[52] ▶ 0.1s

```
#Count more than one column
result = penguins[ ['island', 'species'] ].value_counts().reset_index()

result.columns = ['island', 'species', 'count']

result
```

Table Raw Visualize Statistics

	Ab	island	Ab	species	.3	count
0		Biscoe		Gentoo		124
1		Dream		Chinstrap		68
2		Dream		Adelie		56
3		Torgersen		Adelie		52
4		Biscoe		Adelie		44
