🌲 Intermediate Pandas Part1 🐼

```
import pandas as pd
```

🔒 load data

```
penguins = pd.read_csv('penguins.csv')
```

✅ preview first 5 rows

```
# preview first 5 rows
penguins.head()
```

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	FEMALE

Next steps: Generate code with penguins | View recommended plots | New interactive sheet

✅ preview last 5 rows

```
# preview last 5 rows
penguins.tail()
```

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
339	Gentoo	Biscoe	NaN	NaN	NaN	NaN	NaN
340	Gentoo	Biscoe	46.8	14.3	215.0	4850.0	FEMALE
341	Gentoo	Biscoe	50.4	15.7	222.0	5750.0	MALE
342	Gentoo	Biscoe	45.2	14.8	212.0	5200.0	FEMALE
343	Gentoo	Biscoe	49.9	16.1	213.0	5400.0	MALE

```
# shape of dataframe
penguins.shape
```

```
(344, 7)
```

```
# information of dataframe
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
```

✅ Select column

```
# select column
penguins['species']
```

Show hidden output

```
penguins.species
```

Show hidden output

```
penguins.species.head()
```

	species
0	Adelie
1	Adelie
2	Adelie
3	Adelie
4	Adelie

dtype: object

```
# select multiple column
penguins[['species', 'island', 'sex']].head()
```

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

```
penguins[['species', 'island', 'sex']].tail(8)
```

Show hidden output

✅integer location based indexing (iloc)

```
# integer location based indexing (iloc)
# เลือก row ตามตัวเลข index
penguins.iloc[0]
```

	0
species	Adelie
island	Torgersen
bill_length_mm	39.1
bill_depth_mm	18.7
flipper_length_mm	181.0
body_mass_g	3750.0
sex	MALE

dtype: object

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

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE

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

Show hidden output

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

	species	island	body_mass_g
0	Adelie	Torgersen	3750.0
1	Adelie	Torgersen	3800.0
2	Adelie	Torgersen	3250.0
3	Adelie	Torgersen	NaN
4	Adelie	Torgersen	3450.0

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

Show hidden output

Next steps: Generate code with mini_penguins | View recommended plots | New interactive sheet

✅ Filter dataframe with one condition

[ ] ↳ 3 cells hidden

Filter dataframe more one condition

```
# filter more than one condition
# `and` opertor
penguins[ (penguins['island'] == 'Torgersen') & (penguins['bill_length_mm'] < 35) ]
```

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
8	Adelie	Torgersen	34.1	18.1	193.0	3475.0	NaN
14	Adelie	Torgersen	34.6	21.1	198.0	4400.0	MALE
18	Adelie	Torgersen	34.4	18.4	184.0	3325.0	FEMALE
70	Adelie	Torgersen	33.5	19.0	190.0	3600.0	FEMALE
80	Adelie	Torgersen	34.6	17.2	189.0	3200.0	FEMALE

```
# `or` opertor
filtered_penguins = penguins[ (penguins['island'] == 'Torgersen') | (penguins['bill_length_mm'] < 35) ]
```

✅ Query > filter with .query()

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

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
8	Adelie	Torgersen	34.1	18.1	193.0	3475.0	NaN
14	Adelie	Torgersen	34.6	21.1	198.0	4400.0	MALE
18	Adelie	Torgersen	34.4	18.4	184.0	3325.0	FEMALE
70	Adelie	Torgersen	33.5	19.0	190.0	3600.0	FEMALE
80	Adelie	Torgersen	34.6	17.2	189.0	3200.0	FEMALE

⛔ missing values

```
# check missing in each column
penguins.isna().sum()
```

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

dtype: int64

```
# filter missing values in column `sex`
penguins[penguins['sex'].isna()]
```

Show hidden output

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

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
339	Gentoo	Biscoe	NaN	NaN	NaN	NaN	NaN

⛔ Drop na

```
# drop na
clean_penguins = penguins.dropna()

clean_penguins.head(3)
```

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE

Next steps: Generate code with clean_penguins | View recommended plots | New interactive sheet

✅ fill missing values > mean imputation

```
# fill missing values
penguins.head()
```

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FEMALE
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	FEMALE

Next steps: Generate code with penguins | View recommended plots | New interactive sheet

```
# fill na with mean
avg_value = penguins['bill_length_mm'].mean()
avg_value
```

```
np.float64(43.9219298245614)
```

```
penguins['bill_length_mm'] = penguins['bill_length_mm'].fillna(avg_value)
penguins.head()
```

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.10000	18.7	181.0	3750.0	MALE
1	Adelie	Torgersen	39.50000	17.4	186.0	3800.0	FEMALE
2	Adelie	Torgersen	40.30000	18.0	195.0	3250.0	FEMALE
3	Adelie	Torgersen	43.92193	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.70000	19.3	193.0	3450.0	FEMALE

Next steps: Generate code with penguins | View recommended plots | New interactive sheet

✅ Sort Dataframe

```
## sort bill_length_mm low to high, high to low
penguins.dropna().sort_values('bill_length_mm', ascending=False).head(3)
```

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
253	Gentoo	Biscoe	59.6	17.0	230.0	6050.0	MALE
169	Chinstrap	Dream	58.0	17.8	181.0	3700.0	FEMALE
321	Gentoo	Biscoe	55.9	17.0	228.0	5600.0	MALE

```
# sort multiple column
penguins.dropna().sort_values(['island','bill_length_mm'], ascending=[True,False])
```

Show hidden output

✅ Unique and Count

```
# unique values
unique_islands = penguins['island'].unique()
unique_islands_df = pd.DataFrame(unique_islands, columns=['island'])
unique_islands_df
```

	island
0	Torgersen
1	Biscoe
2	Dream

Next steps: Generate code with unique_islands_df | View recommended plots | New interactive sheet

```
# count values
penguins['species'].value_counts()
```

species	count
Adelie	152
Gentoo	124
Chinstrap	68

dtype: int64

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

# rename column
result.columns = ['island', 'species', 'count']

result
```

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

Next steps: Generate code with `result` View recommended plots New interactive sheet