Import Library

```
import pandas as pd
```

Load data

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

```
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

summarise dataframe `.describe()`

```
# summarise dataframe
penguins.describe()
```

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
count	342.000000	342.000000	342.000000	342.000000
mean	43.921930	17.151170	200.915205	4201.754386
std	5.459584	1.974793	14.061714	801.954536
min	32.100000	13.100000	172.000000	2700.000000
25%	39.225000	15.600000	190.000000	3550.000000
50%	44.450000	17.300000	197.000000	4050.000000
75%	48.500000	18.700000	213.000000	4750.000000
max	59.600000	21.500000	231.000000	6300.000000

```
# average, mean
penguins['bill_length_mm'].mean()
```

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

```
# std: standard deviation
penguins['bill_length_mm'].std()
```

```
5.459583713926532
```

```
# median
penguins['bill_length_mm'].median()
```

```
44.45
```

Group by + sum/mean

```
# group by + sum/mean
penguins.groupby('species')['bill_length_mm'].mean()
```

species	bill_length_mm
Adelie	38.791391
Chinstrap	48.833824
Gentoo	47.504878

dtype: float64

```
# group by aggregate function
penguins.groupby('species')['bill_length_mm'].agg(['min', 'mean', 'median', 'std', 'max'])
```

species	min	mean	median	std	max
Adelie	32.1	38.791391	38.80	2.663405	46.0
Chinstrap	40.9	48.833824	49.55	3.339256	58.0
Gentoo	40.9	47.504878	47.30	3.081857	59.6

```
# group by more than one column
result = penguins.groupby(['island', 'species'])['bill_length_mm'].agg(['min', 'mean', 'max']).reset_index

result.to_csv('penguins_result.csv')
```

```
# if youe code is long
penguins.groupby(['island', 'species'])['bill_length_mm'] \
    .agg(['min', 'mean', 'max']) \
    .reset_index()
```

	island	species	min	mean	max
0	Biscoe	Adelie	34.5	38.975000	45.6
1	Biscoe	Gentoo	40.9	47.504878	59.6
2	Dream	Adelie	32.1	38.501786	44.1
3	Dream	Chinstrap	40.9	48.833824	58.0
4	Torgersen	Adelie	33.5	38.950980	46.0

.map values MALE:m, FEMALE:f

```
penguins['sex_new'] = penguins['sex'].map({'MALE':'m', 'FEMALE':'f'}).fillna('other')

penguins.head()
```

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

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

numpy

```
# import numpy
import numpy as np
np.mean(penguins['bill_length_mm'])
```

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

```
# pandas style
penguins['bill_length_mm'].mean()
```

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

```
# other function of numpy
print (np.sum(penguins['bill_depth_mm']))
print (np.std(penguins['body_mass_g']))
```

```
5865.700000000001
800.7812292384519
```

🐼 numpy.where (condition like ifelse)

```
score = pd.Series([80, 50, 62, 95, 20])
```

```
grade = np.where(score >= 80, 'passed', 'failed')

grade = pd.DataFrame(grade)
grade.columns = ['result']
grade
```

	result
0	passed
1	failed
2	failed
3	passed
4	failed

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

```
df = penguins.query("species == 'Adelie' ")[['species', 'island', 'bill_length_mm']].dropna()
df.head()
```

	species	island	bill_length_mm
0	Adelie	Torgersen	39.1
1	Adelie	Torgersen	39.5
2	Adelie	Torgersen	40.3
4	Adelie	Torgersen	36.7
5	Adelie	Torgersen	39.3

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

```
df['new_column'] = np.where(df['bill_length_mm']> 40, True, False) #boolean
df.head(10)
```

	species	island	bill_length_mm	new_column
0	Adelie	Torgersen	39.1	False
1	Adelie	Torgersen	39.5	False
2	Adelie	Torgersen	40.3	True
4	Adelie	Torgersen	36.7	False

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

🐼 Merge Dataframe (join table)

```
# create data frame
left = {
    'key': [1, 2, 3, 4],
    'name': ['sun', 'joe', 'jane', 'anna'],
    'age':[25, 28, 30, 22]
}

right = {
    'key': [1, 2, 3, 4],
    'city': ['Bangkok', 'London', 'Seoul', 'Tokyo'],
    'zip': [1001, 2504, 2094, 9802]
}

df_left = pd.DataFrame(left)
df_right = pd.DataFrame(right)
```

```
df_left
```

	key	name	age
0	1	sun	25
1	2	joe	28
2	3	jane	30
3	4	anna	22

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

```
import pandas as pd
df_result = pd.merge(df_left, df_right, on='key')

df_result
```

	key	name	age	city	zip
0	1	sun	25	Bangkok	1001
1	2	joe	28	London	2504
2	3	jane	30	Seoul	2094
3	4	anna	22	Tokyo	9802

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

Start coding or generate with AI.