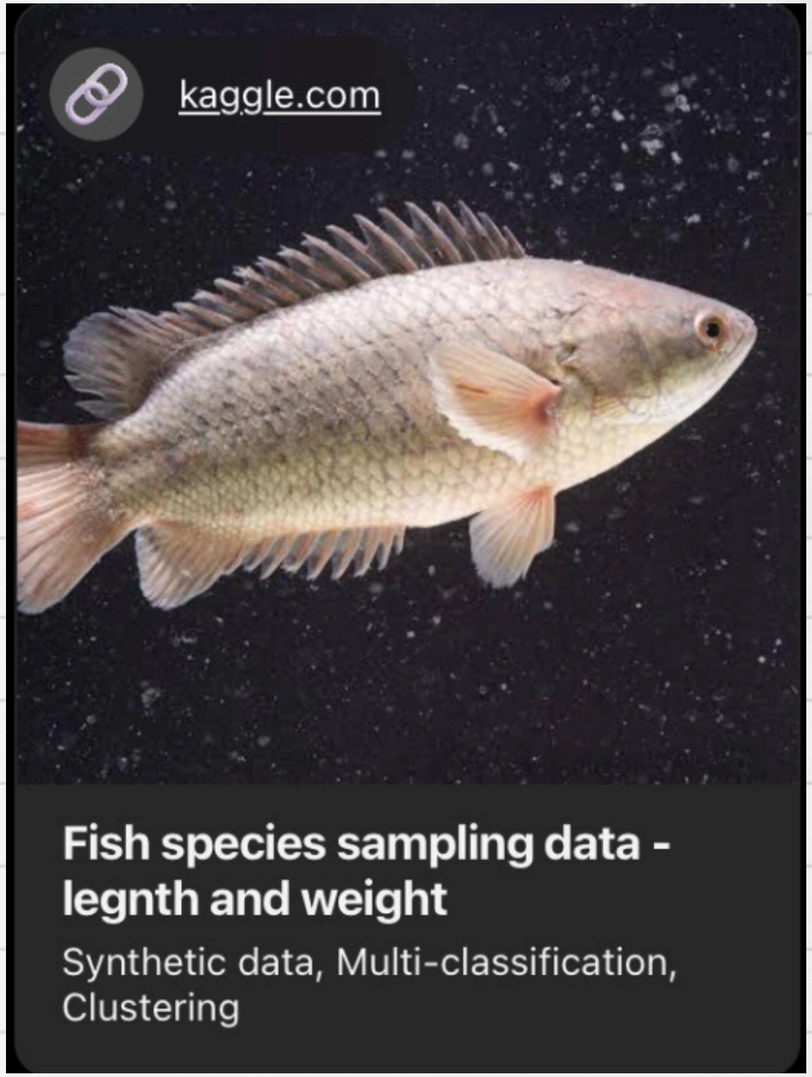
PROJECT 1
GROUP 3

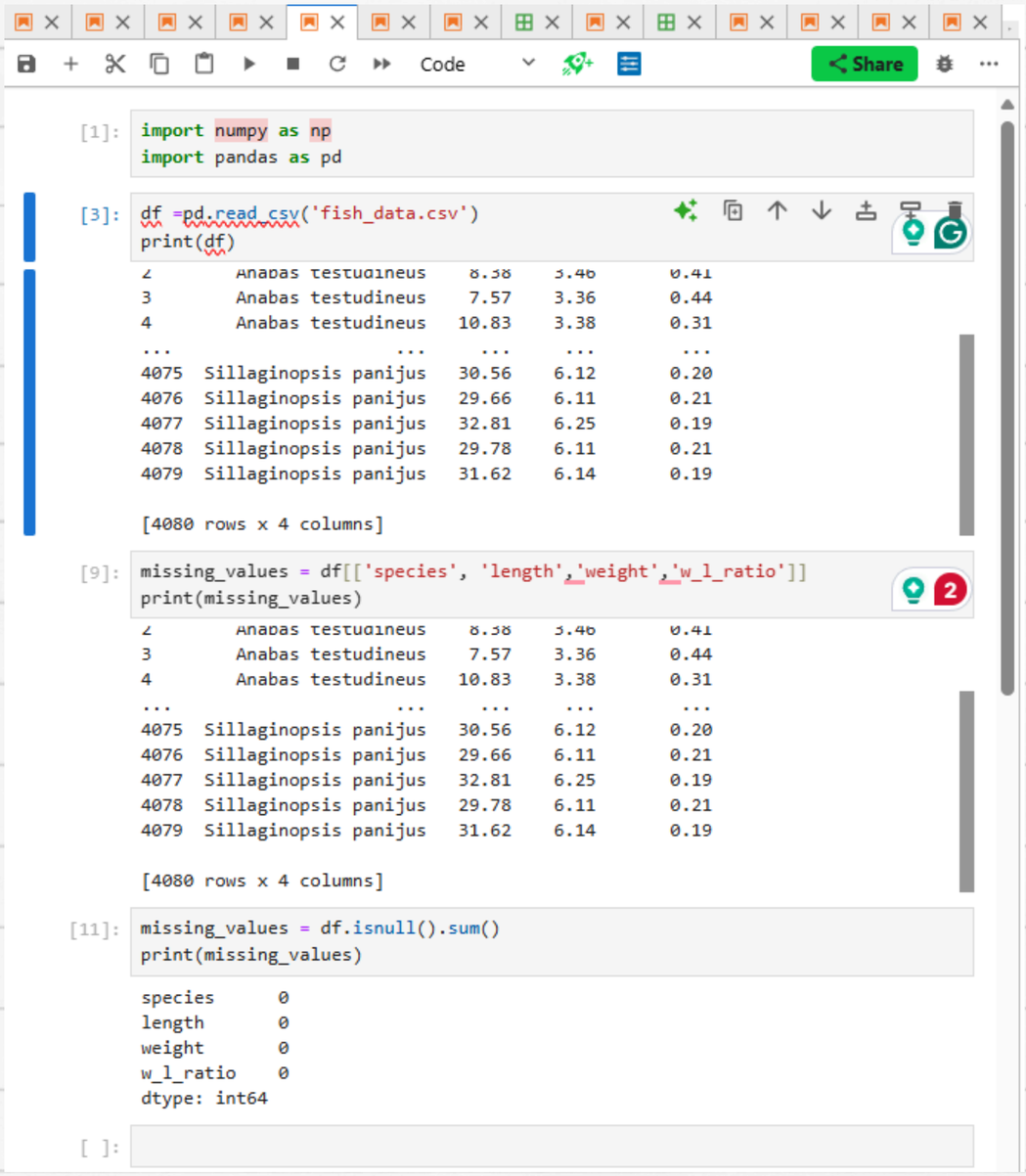
HAILEY BURNETTE, LEILA DIALLO, ELISE WEBB, & CRISTIAN SIMMS

INTRODUCTIONS: GROUP MEMBERS AND OUR DATASET

4080 samples of 9 different fish species found in the Tetulia River of Bangladesh
Quantitative data collected: height, weight, and height to weight ratio
Qualitative data: species of the sample



TASK #4- CRISTIAN



```
[1]: import numpy as np
     import pandas as pd

[3]: df =pd.read_csv('fish_data.csv')
     print(df)
```

```
2         Anabas testudineus    8.38    3.46       0.41
3         Anabas testudineus    7.57    3.36       0.44
4         Anabas testudineus   10.83    3.38       0.31
...                      ...     ...     ...        ...
4075  Sillaginopsis panijus   30.56    6.12       0.20
4076  Sillaginopsis panijus   29.66    6.11       0.21
4077  Sillaginopsis panijus   32.81    6.25       0.19
4078  Sillaginopsis panijus   29.78    6.11       0.21
4079  Sillaginopsis panijus   31.62    6.14       0.19

[4080 rows x 4 columns]
```

```
[9]: missing_values = df[['species', 'length','weight','w_l_ratio']]
     print(missing_values)
```

```
2         Anabas testudineus    8.38    3.46       0.41
3         Anabas testudineus    7.57    3.36       0.44
4         Anabas testudineus   10.83    3.38       0.31
...                      ...     ...     ...        ...
4075  Sillaginopsis panijus   30.56    6.12       0.20
4076  Sillaginopsis panijus   29.66    6.11       0.21
4077  Sillaginopsis panijus   32.81    6.25       0.19
4078  Sillaginopsis panijus   29.78    6.11       0.21
4079  Sillaginopsis panijus   31.62    6.14       0.19

[4080 rows x 4 columns]
```

```
[11]: missing_values = df.isnull().sum()
      print(missing_values)
```

```
species      0
length       0
weight       0
w_l_ratio    0
dtype: int64
```

```
[ ]:
```

DATA CLEANING WITH PANDAS

TASK # 5 – LEILA

My Task: Data Filtering and Selection (Number 5)

5. Data Filtering and Selection

o Filter the dataset based on conditions. For example: ■ Filter rows where a certain column's value is greater than a threshold (df[df['column'] > value]). o Use .loc[] and .iloc[] to select specific rows and columns.

In [21]:
```
filtered_df = df[df['length'] > 5.0] #Filtering the data to only include lengths over a ce

#Selecting specific rows and columns using .loc[]
selected_data_loc = df.loc[0:4, ['length', 'weight']]
```

In [30]:
```
#Selecting rows based on index and columns based on index position using .iloc[]
selected_data_iloc = df.iloc[0:5, [1, 2]]
```

In [26]:
```
# Results displayed:
filtered_df.head()
```

Out[26]:

	length	weight
0	10.66	3.45
1	6.91	3.27
2	8.38	3.46
3	7.57	3.36
4	10.83	3.38

In [32]:
```
print(selected_data_loc)
```

```
   length  weight
0   10.66    3.45
1    6.91    3.27
2    8.38    3.46
3    7.57    3.36
4   10.83    3.38
```

In [34]:
```
print(selected_data_iloc)
```

```
   length  weight
0   10.66    3.45
1    6.91    3.27
2    8.38    3.46
3    7.57    3.36
4   10.83    3.38
```

Filtering and Selection of specific variables.

TASK #6-ELISE

SORTING

```
#sorting the length column from largest to smallest
length_sorted_df = df.sort_values(by='length',ascending=False )
print(length_sorted_df)
```

	species	length	weight	w_l_ratio
3722	Sillaginopsis panijus	33.86	6.23	0.18
3634	Sillaginopsis panijus	33.85	6.27	0.19
3809	Sillaginopsis panijus	33.84	6.19	0.18
3773	Sillaginopsis panijus	33.82	6.21	0.18
3678	Sillaginopsis panijus	33.66	6.22	0.18
...	...	...	...	...
469	Anabas testudineus	6.39	3.37	0.53
72	Anabas testudineus	6.39	3.06	0.48
286	Anabas testudineus	6.39	3.21	0.50
353	Anabas testudineus	6.37	3.25	0.51
190	Anabas testudineus	6.36	3.25	0.51

```
# sorting the w_l_ratio column from smallest to largest
wl_sorted_df = df.sort_values(by=['w_l_ratio'],ascending=True)
print(wl_sorted_df)
```

	species	length	weight	w_l_ratio	Rank	length_rank
550	Coilia dussumieri	24.74	2.09	0.08	570.0	570.0
828	Coilia dussumieri	25.00	2.05	0.08	522.5	522.5
720	Coilia dussumieri	25.53	2.15	0.08	465.0	465.0
671	Coilia dussumieri	24.69	2.27	0.09	583.0	583.0
834	Coilia dussumieri	25.13	2.34	0.09	503.0	503.0
...	...	...	...	...	...	...
1912	Pethia conchonius	7.39	4.55	0.62	3932.0	3932.0
1904	Pethia conchonius	7.21	4.54	0.63	3968.5	3968.5
2265	Pethia conchonius	7.29	4.61	0.63	3952.5	3952.5
1900	Pethia conchonius	7.25	4.54	0.63	3960.0	3960.0
1985	Pethia conchonius	7.10	4.52	0.64	3988.5	3988.5

```
# sorting the species column descending and weight column ascending from largest to smallest
sw_sorted_df = df.sort_values(by=['species','weight'], ascending = [False, True])
print(sw_sorted_df)
```

	species	length	weight	w_l_ratio	Rank	length_rank	\
3797	Sillaginopsis panijus	27.75	6.01	0.22	455.0	455.0	
3828	Sillaginopsis panijus	27.76	6.01	0.22	454.0	454.0	
3732	Sillaginopsis panijus	28.18	6.02	0.21	447.0	447.0	
3635	Sillaginopsis panijus	28.27	6.03	0.21	443.0	443.0	
3667	Sillaginopsis panijus	28.91	6.03	0.21	419.5	419.5	
...	...	...	...	...	...	...	
396	Anabas testudineus	9.66	3.51	0.36	3384.5	3384.5	
406	Anabas testudineus	10.04	3.51	0.35	3310.5	3310.5	
275	Anabas testudineus	9.47	3.53	0.37	3432.5	3432.5	
156	Anabas testudineus	9.97	3.58	0.36	3321.5	3321.5	
447	Anabas testudineus	8.57	3.58	0.42	3655.0	3655.0	

TASK #6-ELISE

RANKING

```
# ranking the length column descending from largest to smallest
df_rank = df['length_rank'] = df['length'].rank(ascending=False)
df_rank
```

```
0       3185.0
1       4021.5
2       3711.5
3       3893.5
4       3149.0
         ...
4075     300.5
4076     381.0
4077      49.0
4078     370.5
4079     164.0
Name: length, Length: 4080, dtype: float64
```

```
df_rank = df
df['species_rank'] = df['species'].rank(ascending=False)
df_rank
```

	species	length	weight	w_l_ratio	Rank	length_rank	sub_group_rank	species_rank
0	Anabas testudineus	10.66	3.45	0.32	3185.0	3185.0	2.0	3842.5
1	Anabas testudineus	6.91	3.27	0.47	4021.5	4021.5	1.0	3842.5
2	Anabas testudineus	8.38	3.46	0.41	3711.5	3711.5	4.0	3842.5
3	Anabas testudineus	7.57	3.36	0.44	3893.5	3893.5	2.0	3842.5
4	Anabas testudineus	10.83	3.38	0.31	3149.0	3149.0	2.0	3842.5
...	...	...	...	...	...	...	...	...
4075	Sillaginopsis panijus	30.56	6.12	0.20	300.5	300.5	1.0	228.0
4076	Sillaginopsis panijus	29.66	6.11	0.21	381.0	381.0	1.0	228.0
4077	Sillaginopsis panijus	32.81	6.25	0.19	49.0	49.0	1.0	228.0
4078	Sillaginopsis panijus	29.78	6.11	0.21	370.5	370.5	1.5	228.0
4079	Sillaginopsis panijus	31.62	6.14	0.19	164.0	164.0	1.0	228.0

4080 rows × 8 columns

```
# ranking the species column in descending order
```

TASK # 7 – HAILEY

Basic Operations

mean
median
mode
range

```
mean_w_l_ratio=df['w_l_ratio'].mean()
print(f'The mean length is: {mean_w_l_ratio}.')
```

```
The mean length is: 0.252781862745098.
```

```
median_w_l_ratio=df['w_l_ratio'].median()
print(f'The median length is: {median_w_l_ratio}.')
```

```
The median length is: 0.19.
```

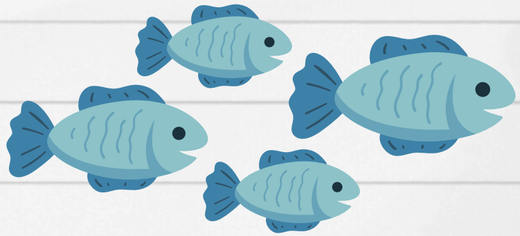
```
mode_w_l_ratio=df['w_l_ratio'].mode()
print(f'The median length is: {mode_w_l_ratio}.')
```

```
The median length is: 0    0.18
Name: w_l_ratio, dtype: float64.
```

```
range_w_l_ratio = df['w_l_ratio'].max() - df['w_l_ratio'].min()
print(f'The range of the weight length ratio is: {range_w_l_ratio}.')
```

```
The range of the weight length ratio is: 0.56.
```

TASK #7 RESULTS:



- Mean of length: 17.35 cm
- Mean of weight: 3.79 g
- Mean of l : w ratio: 0.25
- Largest fish by length: Sillaginopsis panijus at 31.07 cm
- Smallest fish by length: Anabas testudineus at 8.18 cm
- Largest fish by weight: Sillaginopsis panijus at 6.14 g
- Smallest fish by weight: Puntius lateristriga at 2.63 g
- Largest fish by l : w ratio: Pethia conchonius at 0.49
- Smallest fish by l : w ratio: Coilia dussumieri at 0.11

THANK YOU VERY MUCH!