Activity: Removing Data

Introduction

In this activity you will practice using Pandas functionality to check for and remove any unwanted data from a dataset. This activity will cover the following topics:

- Removing columns from a DataFrame
- Removing rows from a DataFrame
- Removing rows based on a condition
- Checking for duplicate data

Question 1

Create a `DataFrame` called `df` from the given CSV file `exotic_plants_data.csv`, then drop the column `Type` and assign the result to a new `DataFrame` called `df_no_type`.

In [9]:
```
import pandas as pd


# Your code here

df = pd.read_csv('exotic_plants_data.csv')

df_no_type = df.drop(columns = ['Type'])


#df.head()
```

In [10]:
```
# Question 1 Grading Checks

assert isinstance(df, pd.DataFrame), 'Have you created a DataFrame named df?'
assert isinstance(df_no_type, pd.DataFrame), 'Have you created a DataFrame named df_no_type?'
```

Question 2

Check the `df` `DataFrame` for any duplicate rows and assign the result to a new `DataFrame` called `df_duplicates`.

In [30]:
```
# Your code here

df_duplicates = df[df.duplicated()]
```

In [31]:
```
# Question 2 Grading Checks

assert isinstance(df_duplicates, pd.DataFrame), 'Have you created a DataFra
me named df_duplicates?'
```

Question 3

Check the `df` `DataFrame` for any duplicate rows based on the `Plant Name` and `Type` columns and assign the result to a new `DataFrame` called `df_plant_type_duplicates`.

In [28]:
```
# Your code here

df_plant_type_duplicates = df[df.duplicated(subset = ['Plant Name', 'Typ
e'])]
```

In [29]:
```
# Question 3 Grading Checks

assert isinstance(df_plant_type_duplicates, pd.DataFrame), 'Have you create
d a DataFrame named df_duplicates?'
```

Question 4

Create a mask called `clean_mask` that will clean up any duplicates in the `df` DataFrame that have the same `Plant Name` and `Origin` and only keep the most up-to-date duplicate entry.

In [24]:
```
# Your code here

clean_mask = ~df.duplicated(subset=["Plant Name", "Origin"], keep = 'last')
```

In [25]:
```
# Question 4 Grading Checks

assert isinstance(clean_mask, pd.Series), 'Have you created a Series named
clean_mask?'
```