V2_removing_data

August 17, 2025

1 Activity: Removing Data

1.1 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`.

[4]:
```python
import pandas as pd


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


df_no_type = df.drop(columns='Type')
print(df.head())
print(df_no_type.head())
```

```
         Plant Name          Type    Origin  Height (cm)
0            Orchid    Ornamental  Tropical           30
1              Fern  Ground Cover  Tropical           40
2            Bamboo         Grass      Asia          600
3            Cactus     Succulent   America           60
4  Bird of Paradise    Ornamental    Africa          150
         Plant Name    Origin  Height (cm)
0            Orchid  Tropical           30
1              Fern  Tropical           40
2            Bamboo      Asia          600
3            Cactus   America           60
4  Bird of Paradise    Africa          150
```

[ ]:
```python
# 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`.

[5]:
```
import pandas as pd
df_duplicates = df[df.duplicated()]
print(df_duplicates)
```

```
          Plant Name         Type           Origin  Height (cm)
6             Cactus    Succulent          America           60
30         Rafflesia       Flower   Southeast Asia           20
47      Kangaroo Paw       Flower        Australia           60
48      Bougainvillea       Shrub    South America          400
49  Bird of Paradise   Ornamental           Africa          150
50      Venus Flytrap  Carnivorous   North America           15
51              Rose       Flower             Asia           60
```

[ ]:
```
# Question 2 Grading Checks

assert isinstance(df_duplicates, pd.DataFrame), 'Have you created a DataFrame␣
 ↪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`.

[7]:
```
df_plant_type_duplicates = df[df.duplicated(subset=["Plant Name", "Type"])]
print(df_plant_type_duplicates)
```

```
          Plant Name         Type           Origin  Height (cm)
6             Cactus    Succulent          America           60
22            Bamboo        Grass             Asia          500
30         Rafflesia       Flower   Southeast Asia           20
47      Kangaroo Paw       Flower        Australia           60
48      Bougainvillea       Shrub    South America          400
49  Bird of Paradise   Ornamental           Africa          150
50      Venus Flytrap  Carnivorous   North America           15
51              Rose       Flower             Asia           60
53             Tulip       Flower           Europe           30
55         Sunflower       Flower    North America          180
60            Cactus    Succulent         Americas           30
62            Bamboo        Grass             Asia          900
67         Aloe Vera    Succulent           Africa           30
73           Jasmine        Shrub             Asia           90
```

[ ]:
```
# Question 3 Grading Checks

assert isinstance(df_plant_type_duplicates, pd.DataFrame), 'Have you created 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.

[8]:
```
clean_mask = ~df.duplicated(subset=["Plant Name", "Origin"], keep="last")

df_cleaned = df[clean_mask]
print(df_cleaned)
```

```
    Plant Name          Type                     Origin  Height (cm)
0        Orchid    Ornamental                   Tropical           30
1          Fern  Ground Cover                   Tropical           40
5   Banana Tree          Tree                   Tropical          300
6        Cactus     Succulent                    America           60
7      Monstera    Ornamental                   Tropical           70
..          …            …                        …          …
71        Ficus          Tree                       Asia          200
72    Columbine        Flower              North America           30
73      Jasmine         Shrub                       Asia           90
74      Fuchsia        Flower  Central and South America           40
75     Amaranth        Flower                    Various           80

[66 rows x 4 columns]
```

[ ]:
```
# Question 4 Grading Checks

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