Data Pre-Processing

What is our GOAL for this MODULE?

The goal of this module is to learn about observing and preprocessing data.

What did we ACHIEVE in the class TODAY?

- We took two different datasets, merged them together and processed the data.

Which CONCEPTS/CODING BLOCKS did we cover today?

- Logic Building
- Critical Thinking
- Python

How did we DO the activities?

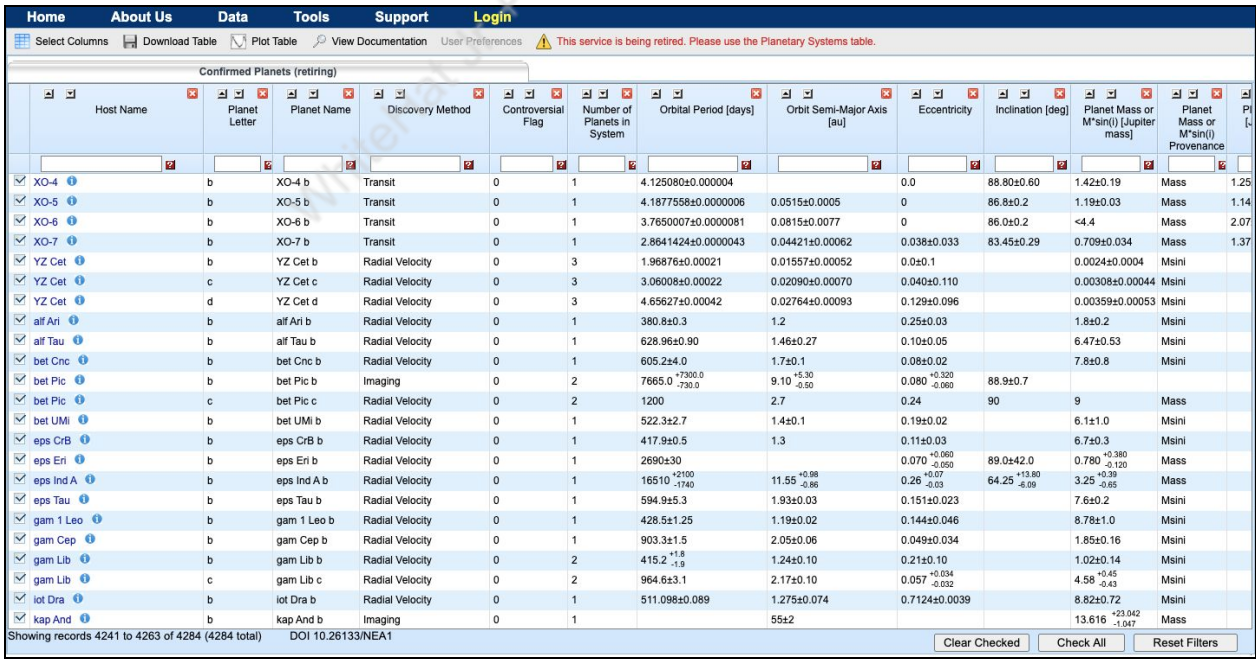
1. Take a look at the previous scraped data from last class.

   https://raw.githubusercontent.com/whitehatjr/web-scrapping-2/master/final.csv

2. Observe that the headers have 11 items while the data has 13 items.
3. Find the missing headers and add them.
4. Go to a new website and download the csv from there.

   https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=planets

5. Observe the two datasets and try to find a pattern in order to merge them.
6. Conclude that both the datasets are arranged in alphabetical order.
7. Observe that the second dataset has all the lowercase planet names at the bottom of the list separately.



8. Create a script that can arrange the second dataset alphabetically irrespective of

them being lowercase or uppercase.

9. Create a new directory and create a virtual environment.
10. Move the downloaded CSV to this new folder and rename it to dataset_1.csv.
11. Think about the logic.
    - ***Fun Fact:* All the lowercase letters have a higher ASCII value (lowercase a has an ASCII value of 97) and all the uppercase letters have a lower ASCII value (uppercase A has an ASCII value of 65).**
12. Using this fun fact, we conclude that we will convert all planet names (3rd column of the CSV) into either uppercase or lowercase and then sort the data.
13. Write the code on the logic we thought

```
import csv

data = []

with open("dataset_2.csv", "r") as f:
  csvreader = csv.reader(f)
  for row in csvreader:
     data.append(row)

headers = data[0]
planet_data = data[1:]

#Converting all planet names to lowercase
for data_point in planet_data:
  data_point[2] = data_point[2].lower()

#Sorting planet names in alphabetical order
planet_data.sort(key=lambda planet_data: planet_data[2])


with open("dataset_2_sorted.csv", "a+") as f:
  csvwriter = csv.writer(f)
  csvwriter.writerow(headers)
  csvwriter.writerows(planet_data)
```


14. Move the CSV of the data we scraped and rename it to dataset_1.csv.
15. Write a script to merge the two datasets and create a final.csv.

```
import csv
```

```
dataset_1 = []
dataset_2 = []

with open("dataset_1.csv", "r") as f:
  csvreader = csv.reader(f)
  for row in csvreader:
     dataset_1.append(row)

with open("dataset_2_sorted.csv", "r") as f:
  csvreader = csv.reader(f)
  for row in csvreader:
     dataset_2.append(row)

headers_1 = dataset_1[0]
planet_data_1 = dataset_1[1:]

headers_2 = dataset_2[0]
planet_data_2 = dataset_2[1:]

headers = headers_1 + headers_2
planet_data = []
for index, data_row in enumerate(planet_data_1):
  planet_data.append(planet_data_1[index] + planet_data_2[index])

with open("final.csv", "a+") as f:
  csvwriter = csv.writer(f)
  csvwriter.writerow(headers)
  csvwriter.writerows(planet_data)
```


What's NEXT?

In the next class, we will learn about data-cleaning. It is the final step before we start performing statistical analysis on our data.

EXTEND YOUR KNOWLEDGE:

You can read the following blog on data processing to understand more:
https://bigdataanalyticsnews.com/data-preparation-why-is-it-important/