Topic	DATA PREPROCESSING	
Class Description	**Students will review the data collected in the previous classes, download more data and merge the datasets into one.**	
Class	**PRO C129**	
Class time	**45 mins**	
Goal	• Understanding and reviewing data • Merging multiple datasets and pre-processing the data	

Resources Required	• Teacher Resources:
	  ○ Laptop with internet connectivity
	  ○ Earphones with mic
	  ○ Notebook and pen
	  ○ Smartphone
	• Student Resources:
	  ○ Laptop with internet connectivity
	  ○ Earphones with mic
	  ○ Notebook and pen	

Class structure	**Warm-Up** **Teacher-Led Activity 1** **Student-Led Activity 1** **Wrap-Up**	**5 mins** **15 mins** **20 mins** **05 mins**
Credit & Permissions:	**Exoplanet Exploration by NASA**	
WARM-UP SESSION - 10 mins		
Teacher Starts Slideshow  **Slide # to #** <**Note**: Only Applicable for Classes with VA> Refer to speaker notes and follow the instructions on each slide.		

Teacher Action	Student Action
Hey <student's name>. How are you? It's great to see you! Are you excited to learn something new today? **Following are the WARM-UP session deliverables:** • Greet the student. • Revision of previous class activities. • Quizzes.	**ESR**: Hi, thanks! Yes, I am excited about it! Click on the slide show tab and present the slides.
WARM-UP QUIZ Click on In-Class Quiz	
Continue WARM-UP Session **Slide # to #** <**Note**: Only Applicable for Classes with VA>	
Activity Details **Following are the session deliverables:** • Appreciate the student. • Narrate the story by using hand gestures and voice modulation methods to bring in more interest in students.	
Teacher Action	**Student Action**
In the last class, we completed scraping data from NASA's exoplanet catalog web page. We scraped data from exoplanets. Teacher Activity 2: Exoplanet Exploration Do you remember what exoplanets are? ***Note**: NASA's exoplanet catalog web page keeps updating as per the new planet discoveries. At the time of writing this document, the web page had **201 Pages** with **25 Planets per page** showing a total of **5009 planets data**.*	**ESR:** Exoplanets are those planets that we have found outside of our own solar system.

After scraping data from each page, the CSV below has **5009**(**id column 0-5008**) **planets data** with planet_type

4993	WTS-1 b	7653	4.01 Jupite	16.644	2012	https://exoplanets.nasa.gov/exoplanet-catalog/5162/wts-1-b/						
4994	WTS-2 b	3362	1.12 Jupite	15.954	2014	https://exoplanets.nasa.gov/exoplanet-catalog/5222/wts-2-b/						
4995	Xi Aquilae	183	2.8 Jupiter	4.70964	2007	https://exoplanets.nasa.gov/exoplanet-catalog/7060/xi-aquilae-b/						
4996	XO-1 b	534	0.83 Jupite	11.251	2006	https://exoplanets.nasa.gov/exoplanet-catalog/5164/xo-1-b/						
4997	XO-2 N b	503	0.566 Jupit	11.246	2007	https://exoplanets.nasa.gov/exoplanet-catalog/7153/xo-2-n-b/						
4998	XO-2 S b	494	0.26 Jupite	11.196	2014	https://exoplanets.nasa.gov/exoplanet-catalog/7154/xo-2-s-b/						
4999	XO-2 S c	494	1.37 Jupite	11.196	2014	https://exoplanets.nasa.gov/exoplanet-catalog/7155/xo-2-s-c/						
5000	XO-3 b	695	7.29 Jupite	9.854	2007	https://exoplanets.nasa.gov/exoplanet-catalog/5460/xo-3-b/						
5001	XO-4 b	889	1.42 Jupite	10.814	2008	https://exoplanets.nasa.gov/exoplanet-catalog/5223/xo-4-b/						
5002	XO-5 b	901	1.19 Jupite	12.198	2008	https://exoplanets.nasa.gov/exoplanet-catalog/5461/xo-5-b/						
5003	XO-6 b	768	4.4 Jupiter	10.247	2016	https://exoplanets.nasa.gov/exoplanet-catalog/3441/xo-6-b/						
5004	XO-7 b	764	0.709 Jupit	10.521	2019	https://exoplanets.nasa.gov/exoplanet-catalog/7546/xo-7-b/						
5005	YSES 2 b	357	6.3 Jupiter	10.885	2021	https://exoplanets.nasa.gov/exoplanet-catalog/7867/yses-2-b/						
5006	YZ Ceti b	12	0.7 Earths	12.074	2017	https://exoplanets.nasa.gov/exoplanet-catalog/7181/yz-ceti-b/						
5007	YZ Ceti c	12	1.14 Earth	12.074	2017	https://exoplanets.nasa.gov/exoplanet-catalog/7182/yz-ceti-c/						
5008	YZ Ceti d	12	1.09 Earth	12.074	2017	https://exoplanets.nasa.gov/exoplanet-catalog/7183/yz-ceti-d/						

updated_scraped_data

***Note**: NASA's exoplanet catalog web page keeps updating as per the new planet discoveries. At the time of writing this document, the web page had **201 Pages** with **25 Planets per page** showing a total of **5009 planets data**. Each planet's more data can be found by clicking on the planet name link.* Teacher Activity 3:11 Comae Berenices b

PLANET TYPE Gas Giant	DISCOVERY DATE 2007
MASS 19.4 Jupiters	PLANET RADIUS 1.08 x Jupiter (estimate)
ORBITAL RADIUS 1.29 AU	ORBITAL PERIOD 326 days
ECCENTRICITY 0.23	DETECTION METHOD Radial Velocity

Console output after scraping data from each planet **hyperlink.**

***Note**: Scraping data from each hyperlink for 5009 planets takes around 2-3 hours. Hence data has been scraped beforehand for teachers and students convenience.*

```
Data Scraping at hyperlink 5006 completed
https://exoplanets.nasa.gov/exoplanet-catalog/7181/yz-ceti-b/
Data Scraping at hyperlink 5007 completed
https://exoplanets.nasa.gov/exoplanet-catalog/7182/yz-ceti-c/
Data Scraping at hyperlink 5008 completed
https://exoplanets.nasa.gov/exoplanet-catalog/7183/yz-ceti-d/
Data Scraping at hyperlink 5009 completed
```

After scraping data from each planet **hyperlink**, the CSV below has **5009**(**id column 0-5008**) **planets data** with planet_type, discovery_rate, mass, planet_radius, orbital_radius, orbital_period, eccentricity, detection_method

	A	B	C	D	E	F	G	H	I
996	4994	Gas Giant	2014	1.12 Jupite	1.363 x Jup	0.01855 AI	1 days	0	
997	4995	Gas Giant	2007	2.8 Jupiter	1.18 x Jupi	0.68 AU	136.8 days	0	
998	4996	Gas Giant	2006	0.83 Jupite	1.14 x Jupi	0.0488 AU	3.9 days	0	
999	4997	Gas Giant	2007	0.566 Jupit	0.993 x Jup	0.0368 AU	2.6 days	0.03	
000	4998	Gas Giant	2014	0.26 Jupite	0.971 x Jup	0.13 AU	18.2 days	0.18	
001	4999	Gas Giant	2014	1.37 Jupite	1.21 x Jupi	0.4756 AU	120.8 days	0.15	
002	5000	Gas Giant	2007	7.29 Jupite	1.41 x Jupi	0.0476 AU	3.2 days	0.29	
003	5001	Gas Giant	2008	1.42 Jupite	1.25 x Jupi	0.05524 AI	4.1 days	0	
004	5002	Gas Giant	2008	1.19 Jupite	1.14 x Jupi	0.0515 AU	4.2 days	0	
005	5003	Gas Giant	2016	4.4 Jupiter	2.07 x Jupi	0.0815 AU	3.8 days	0	
006	5004	Gas Giant	2019	0.709 Jupit	1.373 x Jup	0.04421 AI	2.9 days	0.04	
007	5005	Gas Giant	2021	6.3 Jupiter	1.14 x Jupi	115.0 AU	1176.5 yea	0	
008	5006	Terrestrial	2017	0.7 Earths	0.913 x Ea	0.01634 AI	2 days	0.06	
009	5007	Super Eart	2017	1.14 Earth	1.05 x Eart	0.02156 AI	3.1 days	0	
010	5008	Super Eart	2017	1.09 Earth	1.03 x Eart	0.02851 AI	4.7 days	0.07	

new_scraped_data

Now in today's class, we will combine the data we just scraped in the last class. We will also download some more data from an existing website and finally, we will merge the data as we pre-process it.

Isn't it interesting?

ESR: Yes!

Ok so let's start coding.	
Teacher Ends Slideshow 	
TEACHER-LED ACTIVITY - 10 mins	
Teacher Initiates Screen Share	
ACTIVITY • **Looking at the previous data** • **Downloading more data from the internet and then merging the data**	
Teacher Action	**Student Action**
We have two CSV files. 1. **updated_scraped_data.csv** has the following **headers.** ["name", "light_years_from_earth", "planet_mass", "stellar_magnitude", "discovery_date", "hyperlink"] 2. **New_scraped_data.csv** has following **headers:** ["planet_type","discovery_date", "mass", "planet_radius", "orbital_radius", "orbital_period", "eccentricity", "detection_method"] We'll be merging the data with the new data that we have downloaded.	
Let's start by loading the scraped data from the link. Clone the data using the **!git clone** command.	

```
!git clone https://github.com/procodingclass/PRO-NASA-Exoplanet-Scraped-Data
```

```
Cloning into 'PRO-NASA-Exoplanet-Scraped-Data'...
remote: Enumerating objects: 23, done.
remote: Counting objects: 100% (23/23), done.
remote: Compressing objects: 100% (18/18), done.
remote: Total 23 (delta 8), reused 15 (delta 3), pack-reused 0
Unpacking objects: 100% (23/23), done.
```

To analyze the data, read both the CSV as pandas DataFrames. Store the updated scraped data in **planet_df_1** and new scraped data into **new_plane_df_1**.

```
import pandas as pd
```

```
planet_df_1 = pd.read_csv('/content/PRO-NASA-Exoplanet-Scraped-Data/updated_scraped_data.csv')
new_planet_df_1 = pd.read_csv('/content/PRO-NASA-Exoplanet-Scraped-Data/new_scraped_data.csv')
```

Let's check both the DataFrames by using the **head() method.** All the headers are present along with the id.

```
planet_df_1.head()
```

	id	name	light_years_from_earth	planet_mass	stellar_magnitude	discovery_date	hyperlink
0	0	11 Comae Berenices b	304.0	19.4 Jupiters	4.72307	2007	https://exoplanets.nasa.gov/exoplanet-catalog/...
1	1	11 Ursae Minoris b	409.0	14.74 Jupiters	5.01300	2009	https://exoplanets.nasa.gov/exoplanet-catalog/...
2	2	14 Andromedae b	246.0	4.8 Jupiters	5.23133	2008	https://exoplanets.nasa.gov/exoplanet-catalog/...
3	3	14 Herculis b	58.0	4.66 Jupiters	6.61935	2002	https://exoplanets.nasa.gov/exoplanet-catalog/...
4	4	16 Cygni B b	69.0	1.78 Jupiters	6.21500	1996	https://exoplanets.nasa.gov/exoplanet-catalog/...

Similarly, we'll print **new_planet_df_1** using the **head()** method.

```
new_planet_df_1.head()
```

	id	planet_type	discovery_date	mass	planet_radius	orbital_radius	orbital_period	eccentricity	detection_method
0	0	Gas Giant	2007	19.4 Jupiters	1.08 x Jupiter	1.29 AU	326 days	0.23	NaN
1	1	Gas Giant	2009	14.74 Jupiters	1.09 x Jupiter	1.53 AU	1.4 years	0.08	NaN
2	2	Gas Giant	2008	4.8 Jupiters	1.15 x Jupiter	0.83 AU	185.8 days	0.0	NaN
3	3	Gas Giant	2002	4.66 Jupiters	1.15 x Jupiter	2.93 AU	4.9 years	0.37	NaN
4	4	Gas Giant	1996	1.78 Jupiters	1.2 x Jupiter	1.66 AU	2.2 years	0.68	NaN

Now, we have to merge the data. If we look closely we have two repetitive columns.
let's remove a column from DataFrame. The discovery date and mass are repeated in both the data frames.

```
planet_df_1.head()
```

	id	name	light_years_from_earth	planet_mass	stellar_magnitude	discovery_date
0	0	11 Comae Berenices b	304.0	19.4 Jupiters	4.72307	2007
1	1	11 Ursae Minoris b	409.0	14.74 Jupiters	5.01300	2009
2	2	14 Andromedae b	246.0	4.8 Jupiters	5.23133	2008
3	3	14 Herculis b	58.0	4.66 Jupiters	6.61935	2002
4	4	16 Cygni B b	69.0	1.78 Jupiters	6.21500	1996

```
new_planet_df_1.head()
```

	id	planet_type	discovery_date	mass	planet_radius	orbital_radius	orbital_period	ecc
0	0	Gas Giant	2007	19.4 Jupiters	1.08 x Jupiter	1.29 AU	326 days	
1	1	Gas Giant	2009	14.74 Jupiters	1.09 x Jupiter	1.53 AU	1.4 years	
2	2	Gas Giant	2008	4.8 Jupiters	1.15 x Jupiter	0.83 AU	185.8 days	
3	3	Gas Giant	2002	4.66 Jupiters	1.15 x Jupiter	2.93 AU	4.9 years	
4	4	Gas Giant	1996	1.78 Jupiters	1.2 x Jupiter	1.66 AU	2.2 years	

Also, the detection method column was empty, so NAN is written. This column has to be removed.

```
new_planet_df_1.head()
```

	id	planet_type	discovery_date	mass	planet_radius	orbital_radius	orbital_period	eccentricity	detection_method
0	0	Gas Giant	2007	19.4 Jupiters	1.08 x Jupiter	1.29 AU	326 days	0.23	NaN
1	1	Gas Giant	2009	14.74 Jupiters	1.09 x Jupiter	1.53 AU	1.4 years	0.08	NaN
2	2	Gas Giant	2008	4.8 Jupiters	1.15 x Jupiter	0.83 AU	185.8 days	0.0	NaN
3	3	Gas Giant	2002	4.66 Jupiters	1.15 x Jupiter	2.93 AU	4.9 years	0.37	NaN
4	4	Gas Giant	1996	1.78 Jupiters	1.2 x Jupiter	1.66 AU	2.2 years	0.68	NaN

Use the **drop()** method to remove the columns. **inplace= True** updates the same DataFrame.

```
new_planet_df_1.drop(columns=['discovery_date', 'mass', 'detection_method'], inplace=True)
new_planet_df_1.head()
```

	id	planet_type	planet_radius	orbital_radius	orbital_period	eccentricity
0	0	Gas Giant	1.08 x Jupiter	1.29 AU	326 days	0.23
1	1	Gas Giant	1.09 x Jupiter	1.53 AU	1.4 years	0.08
2	2	Gas Giant	1.15 x Jupiter	0.83 AU	185.8 days	0.0
3	3	Gas Giant	1.15 x Jupiter	2.93 AU	4.9 years	0.37
4	4	Gas Giant	1.2 x Jupiter	1.66 AU	2.2 years	0.68

To merge the DataFrames and store them, define headers and create a new DataFrame **final_planet_df** with these headers. The headers are in the sequence of the DataFrames to be merged.

```
headers = ["name", "light_years_from_earth", "planet_mass", "stellar_magnitude", "discovery_date",
           "hyperlink", "planet_type","discovery_date", "mass", "planet_radius", "orbital_radius",
           "orbital_period", "eccentricity", "detection_method"]

final_planet_df = pd.DataFrame(columns=headers)
```

Use the **merge()** method to combine the two DataFrames.

```
final_planet_df = pd.merge(planet_df_1,new_planet_df_1)

final_planet_df.head()
```

	id	name	light_years_from_earth	planet_mass	stellar_magnitude	discovery_date
0	0	11 Comae Berenices b	304.0	19.4 Jupiters	4.72307	2007
1	1	11 Ursae Minoris b	409.0	14.74 Jupiters	5.01300	2009
2	2	14 Andromedae b	246.0	4.8 Jupiters	5.23133	2008
3	3	14 Herculis b	58.0	4.66 Jupiters	6.61935	2002
4	4	16 Cygni B b	69.0	1.78 Jupiters	6.21500	1996

Convert the final DataFrame to csv.

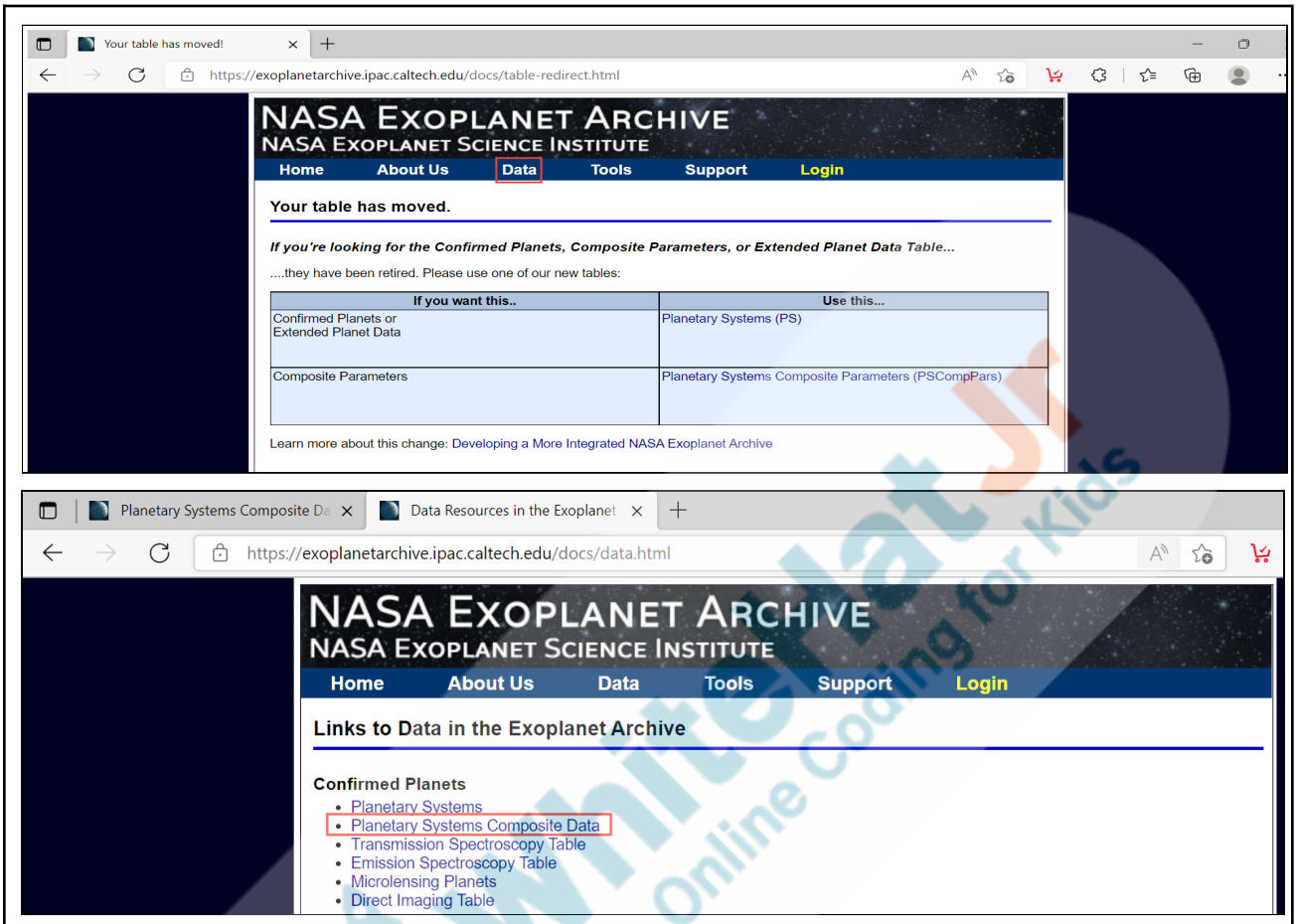
```
final_planet_df.to_csv('final_scraped_data.csv')
```

Great!
Now our dataset is combined.

Next, we have another website that has even more data. We will download the data in CSV format from there and then we will merge the two datasets.

For this, we will go to the following link:
The teacher opens the link Teacher Activity 4 to check the dataset.
Go to:

Data >> Planetary Systems Composite Data

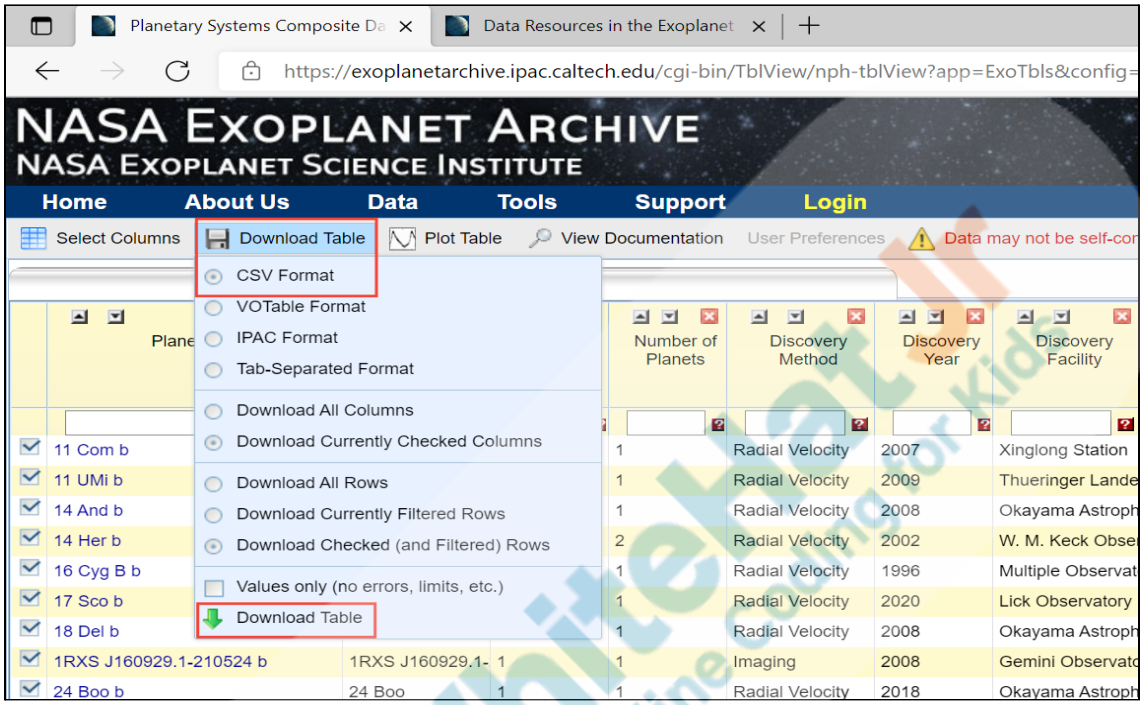


Now you can see that the data of exoplanets is present here.
This table has a lot of data that will help us analyze a lot of data about our universe, which planets are habitable and which ones are not.

***Note:** The teacher can use the link directly to go to the table. The link may change due to the timely updation of data.*

Let's download this data!

For downloading it, we will click on the **Download Table** on the Menu bar button at the top of the table and select

CSV Format and then click on the **Download Table** option with a green arrow.	
	
Note: *Download the dataset.* Next we'll read this dataset using Pandas DataFrame and print it.	

```
archive_planet_df["pl_name"] = archive_planet_df["pl_name"].str.lower()

archive_planet_df.head()
```

	id	pl_name	hostname	sy_snum	sy_pnum	discoverymethod	disc_year	disc_facility
0	0	11 com b	11 Com	2	1	Radial Velocity	2007	Xinglong Station
1	1	11 umi b	11 UMi	1	1	Radial Velocity	2009	Thueringer Landessternwarte Tautenburg
2	2	14 and b	14 And	1	1	Radial Velocity	2008	Okayama Astrophysical Observatory
3	3	14 her b	14 Her	1	2	Radial Velocity	2002	W. M. Keck Observatory

It has header names in short form and also 85 columns are present.

```
archive_planet_df.head()
```

	id	pl_name	hostname	sy_snum	sy_pnum	discoverymethod	disc_year	disc_facility	pl_controv_flag	pl_orbper	...
0	0	11 Com b	11 Com	2	1	Radial Velocity	2007	Xinglong Station	0	326.03000	...
1	1	11 UMi b	11 UMi	1	1	Radial Velocity	2009	Thueringer Landessternwarte Tautenburg	0	516.21997	...
2	2	14 And b	14 And	1	1	Radial Velocity	2008	Okayama Astrophysical Observatory	0	185.84000	...
3	3	14 Her b	14 Her	1	2	Radial Velocity	2002	W. M. Keck Observatory	0	1773.40002	...
4	4	16 Cyg B b	16 Cyg B	3	1	Radial Velocity	1996	Multiple Observatories	0	798.50000	...

5 rows × 85 columns

We can check the number of rows and columns using the **shape()** method.

```
final_planet_df.shape
(5009, 12)

archive_planet_df.shape
(5009, 85)
```

Great! Now, we need to merge the data. For this, you are provided with both datasets. Let's write the code to merge them.
Teacher Stops Screen Share
So now it's your turn. Please share your screen with me.
Teacher Starts Slideshow **Slide # to #** <**Note**: Only Applicable for Classes with VA> Refer to speaker notes and follow the instructions on each slide.
We have one more class challenge for you. Can you solve it? Let's try. I will guide you through it.
Teacher Ends Slideshow
STUDENT-LED ACTIVITY - 20 mins
• **Ask the student to press the ESC key to come back to the panel.** • **Guide the student to start Screen Share.** • **The teacher gets into Full Screen.**

Student Initiates Screen Share	
ACTIVITY • **The student tries to find patterns between the data** • **The student writes some code to merge 2 datasets**	
Teacher Action	**Student Action**
Open Student Activity 2 to download the datasets. This contains the dataset we merged (**final_scraped_data**) and the dataset that I have just downloaded from Planetary Systems Composite data. (**PSCompPars.csv**). ***Note:** Guide the student to open Student Activity 2 and download the dataset.*	
NASA's EXOPLANET EXPLORATION Note: At the scrapping data from this website, this web page had 5009 planets across 201 page(with 25 planets per page) 1. **updated_scraped_data.csv** : ["name", "light_years_from_earth", "planet_mass", "stellar_magnitude", "discovery_date", "hyperlink"] 2. **new_scraped_data.csv** : ["planet_type","discovery_date", "mass", "planet_radius", "orbital_radius", "orbital_period","eccentricity", "detection_method"] NASA's EXOPLANET ARCHIVE **PSCompPars.csv**	
Okay, now the first thing that we have to do is that we need to look at the data and try to identify a pattern within the data. If we look at the names of these planets in both the datasets, we can see that the data we scraped earlier i.e. **final_planet_df** has a full version of the names while the planet data we downloaded has a shorter version of the names.	

```
final_planet_df.head()
```

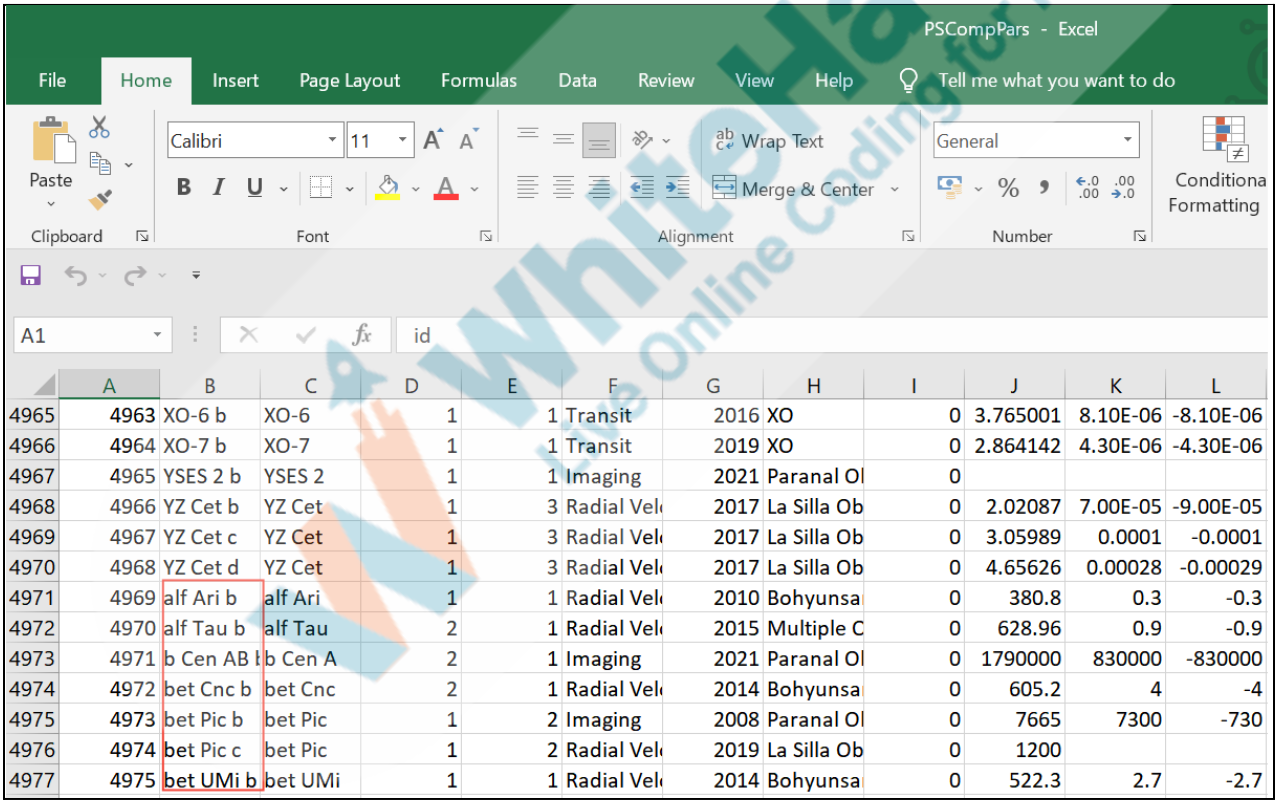
	id	name	light_years_from_earth	planet_mass	stellar_magnitude	discovery_date
0	0	11 Comae Berenices b	304.0	19.4 Jupiters	4.72307	2007
1	1	11 Ursae Minoris b	409.0	14.74 Jupiters	5.01300	2009
2	2	14 Andromedae b	246.0	4.8 Jupiters	5.23133	2008
3	3	14 Herculis b	58.0	4.66 Jupiters	6.61935	2002
4	4	16 Cygni B b	69.0	1.78 Jupiters	6.21500	1996

```
archive_planet_df.head()
```

	id	pl_name	hostname	sy_snum	sy_pnum	discoverymethod	disc_year	disc_facility
0	0	11 Com b	11 Com	2	1	Radial Velocity	2007	Xinglong Station
1	1	11 UMi b	11 UMi	1	1	Radial Velocity	2009	Thueringer Landessternwarte Tautenburg
2	2	14 And b	14 And	1	1	Radial Velocity	2008	Okayama Astrophysical Observatory
3	3	14 Her b	14 Her	1	2	Radial Velocity	2002	W. M. Keck Observatory
4	4	16 Cyg B b	16 Cyg B	3	1	Radial Velocity	1996	Multiple Observatories

The name **11 Comae Berenices b** from the data we scraped is written as **11 Com b** in the data we downloaded. Similarly, all the names are different.	*The student tries to find a pattern for a couple of minutes.*

Therefore we cannot use the names of these data points as a metric to merge the two datasets. Can you think of any patterns that both of the datasets follow?	**ESR:** The names are in alphabetical order.
That's great! But there is one thing we need to be careful about. Both the datasets are in alphabetical order, however, the second dataset (with tables) has the planet names that start with a lower case alphabet in the bottom, segregated separately in alphabetical order while the data we scraped is in perfect alphabetical order.	



	A	B	C	D	E	F	G	H	I	J	K	L
4965	4963	XO-6 b	XO-6	1	1	Transit	2016	XO	0	3.765001	8.10E-06	-8.10E-06
4966	4964	XO-7 b	XO-7	1	1	Transit	2019	XO	0	2.864142	4.30E-06	-4.30E-06
4967	4965	YSES 2 b	YSES 2	1	1	Imaging	2021	Paranal Ol	0			
4968	4966	YZ Cet b	YZ Cet	1	3	Radial Vel	2017	La Silla Ob	0	2.02087	7.00E-05	-9.00E-05
4969	4967	YZ Cet c	YZ Cet	1	3	Radial Vel	2017	La Silla Ob	0	3.05989	0.0001	-0.0001
4970	4968	YZ Cet d	YZ Cet	1	3	Radial Vel	2017	La Silla Ob	0	4.65626	0.00028	-0.00029
4971	4969	alf Ari b	alf Ari	1	1	Radial Vel	2010	Bohyunsa	0	380.8	0.3	-0.3
4972	4970	alf Tau b	alf Tau	2	1	Radial Vel	2015	Multiple C	0	628.96	0.9	-0.9
4973	4971	b Cen AB b	b Cen A	2	1	Imaging	2021	Paranal Ol	0	1790000	830000	-830000
4974	4972	bet Cnc b	bet Cnc	2	1	Radial Vel	2014	Bohyunsa	0	605.2	4	-4
4975	4973	bet Pic b	bet Pic	1	2	Imaging	2008	Paranal Ol	0	7665	7300	-730
4976	4974	bet Pic c	bet Pic	1	2	Radial Vel	2019	La Silla Ob	0	1200		
4977	4975	bet UMi b	bet UMi	1	1	Radial Vel	2014	Bohyunsa	0	522.3	2.7	-2.7

Here, we can see the dataset has all the planet names until alphabet Y in alphabetical order. After it, the planet names with lower case alphabet are below it. These are also present in alphabetical order.	**ESR:** We need to arrange the second dataset in alphabetical order irrespective of if the name is

What do you think we need to do first to merge the two datasets? That's right! Let's write a code that can quickly do that.	uppercase or lower case.

```
archive_planet_df["pl_name"] = archive_planet_df["pl_name"].str.lower()
```

```
archive_planet_df.head()
```

	id	pl_name	hostname	sy_snum	sy_pnum	discoverymethod	disc_year	disc_facility
0	0	11 com b	11 Com	2	1	Radial Velocity	2007	Xinglong Station
1	1	11 umi b	11 UMi	1	1	Radial Velocity	2009	Thueringer Landessternwarte Tautenburg
2	2	14 and b	14 And	1	1	Radial Velocity	2008	Okayama Astrophysical Observatory
3	3	14 her b	14 Her	1	2	Radial Velocity	2002	W. M. Keck Observatory
4	4	16 cyg b b	16 Cyg B	3	1	Radial Velocity	1996	Multiple Observatories

To sort the names by alphabetical order we'll use the **sort_values()** method of pandas to arrange them in alphabetical order.

Also, use the **tail()** method to check the data present at the bottom of the dataset.

```
archive_planet_df = archive_planet_df.sort_values('pl_name')

archive_planet_df.tail(10)
```

	id	pl_name	hostname	sy_snum	sy_pnum	discoverymethod	disc_year
4959	4959	xo-2 s c	XO-2 S	2	3	Radial Velocity	2014
4960	4960	xo-3 b	XO-3	1	1	Transit	2007
4961	4961	xo-4 b	XO-4	1	1	Transit	2008
4962	4962	xo-5 b	XO-5	1	1	Transit	2008
4963	4963	xo-6 b	XO-6	1	1	Transit	2016
4964	4964	xo-7 b	XO-7	1	1	Transit	2019
4965	4965	yses 2 b	YSES 2	1	1	Imaging	2021

Also, print the same for **final_planet_df**.

```
final_planet_df.tail(10)
```

	id	name	light_years_from_earth	planet_mass	stellar_magnitude	discovery_date
4999	4999	XO-2 S c	494.0	1.37 Jupiters	11.196	2014
5000	5000	XO-3 b	695.0	7.29 Jupiters	9.854	2007
5001	5001	XO-4 b	889.0	1.42 Jupiters	10.814	2008
5002	5002	XO-5 b	901.0	1.19 Jupiters	12.198	2008
5003	5003	XO-6 b	768.0	4.4 Jupiters	10.247	2016

Merge these two datasets using the **merge()** method of pandas. Also, check the dimensions of the resultant DataFrame using the **shape()** method.

```
merge_planets_df = pd.merge(final_planet_df, archive_planet_df, on="id" )

merge_planets_df.shape
```

```
(5009, 96)
```

We got 96 which is the addition of the number of columns for final and archive datasets. Since id is the common column it is considered once only.

```
merge_planets_df.columns
```

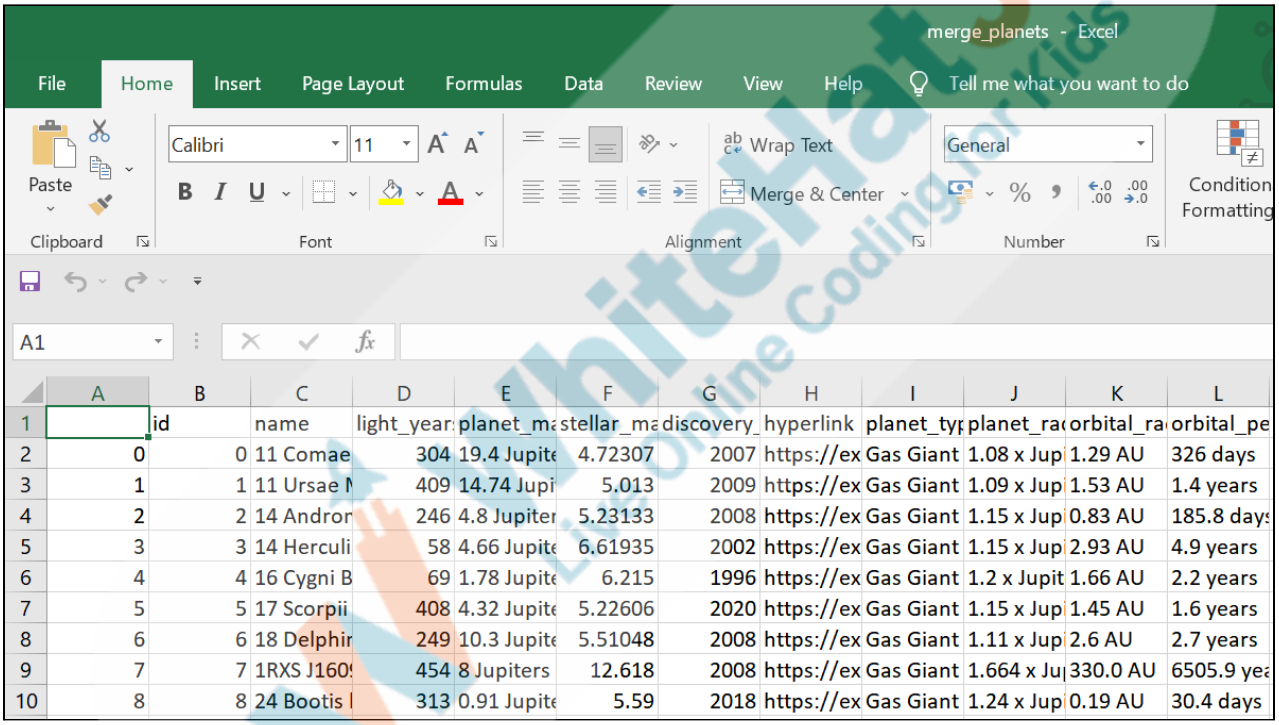
```
Index(['id', 'name', 'light_years_from_earth', 'planet_mass',
       'stellar_magnitude', 'discovery_date', 'hyperlink', 'planet_type',
       'planet_radius', 'orbital_radius', 'orbital_period', 'eccentricity',
       'pl_name', 'hostname', 'sy_snum', 'sy_pnum', 'discoverymethod',
       'disc_year', 'disc_facility', 'pl_controv_flag', 'pl_orbper',
       'pl_orbpererr1', 'pl_orbpererr2', 'pl_orbperlim', 'pl_orbsmax',
       'pl_orbsmaxerr1', 'pl_orbsmaxerr2', 'pl_orbsmaxlim', 'pl_rade',
       'pl_radeerr1', 'pl_radeerr2', 'pl_radelim', 'pl_radj', 'pl_radjerr1',
       'pl_radjerr2', 'pl_radjlim', 'pl_bmasse', 'pl_bmasseerr1',
       'pl_bmasseerr2', 'pl_bmasselim', 'pl_bmassj', 'pl_bmassjerr1',
       'pl_bmassjerr2', 'pl_bmassjlim', 'pl_bmassprov', 'pl_orbeccen',
       'pl_orbeccenerr1', 'pl_orbeccenerr2', 'pl_orbeccenlim', 'pl_insol',
       'pl_insolerr1', 'pl_insolerr2', 'pl_insollim', 'pl_eqt', 'pl_eqterr1',
       'pl_eqterr2', 'pl_eqtlim', 'ttv_flag', 'st_spectype', 'st_teff',
       'st_tefferr1', 'st_tefferr2', 'st_tefflim', 'st_rad', 'st_raderr1',
       'st_raderr2', 'st_radlim', 'st_mass', 'st_masserr1', 'st_masserr2',
       'st_masslim', 'st_met', 'st_meterr1', 'st_meterr2', 'st_metlim',
       'st_metratio', 'st_logg', 'st_loggerr1', 'st_loggerr2', 'st_logglim',
       'rastr', 'ra', 'decstr', 'dec', 'sy_dist', 'sy_disterr1', 'sy_disterr2',
       'sy_vmag', 'sy_vmagerr1', 'sy_vmagerr2', 'sy_kmag', 'sy_kmagerr1',
       'sy_kmagerr2', 'sy_gaiamag', 'sy_gaiamagerr1', 'sy_gaiamagerr2'],
      dtype='object')
```

Convert the DataFrame into csv. Use the download method to download the merged CSV file.

```
# Convert to CSV
merge_planets_df.to_csv('merge_planets.csv')

## Download CSV
from google.colab import files
files.download('merge_planets.csv')
```

Let's check the **merge_planets.csv** file.



	A	B	C	D	E	F	G	H	I	J	K	L
1		id	name	light_year	planet_ma	stellar_ma	discovery_	hyperlink	planet_typ	planet_ra	orbital_ra	orbital_pe
2	0	0	11 Comae	304	19.4 Jupite	4.72307	2007	https://ex	Gas Giant	1.08 x Jup	1.29 AU	326 days
3	1	1	11 Ursae M	409	14.74 Jupi	5.013	2009	https://ex	Gas Giant	1.09 x Jup	1.53 AU	1.4 years
4	2	2	14 Andror	246	4.8 Jupiter	5.23133	2008	https://ex	Gas Giant	1.15 x Jup	0.83 AU	185.8 days
5	3	3	14 Herculi	58	4.66 Jupite	6.61935	2002	https://ex	Gas Giant	1.15 x Jup	2.93 AU	4.9 years
6	4	4	16 Cygni B	69	1.78 Jupite	6.215	1996	https://ex	Gas Giant	1.2 x Jupit	1.66 AU	2.2 years
7	5	5	17 Scorpii	408	4.32 Jupite	5.22606	2020	https://ex	Gas Giant	1.15 x Jup	1.45 AU	1.6 years
8	6	6	18 Delphir	249	10.3 Jupite	5.51048	2008	https://ex	Gas Giant	1.11 x Jup	2.6 AU	2.7 years
9	7	7	1RXS J160	454	8 Jupiters	12.618	2008	https://ex	Gas Giant	1.664 x Ju	330.0 AU	6505.9 yea
10	8	8	24 Bootis I	313	0.91 Jupite	5.59	2018	https://ex	Gas Giant	1.24 x Jup	0.19 AU	30.4 days

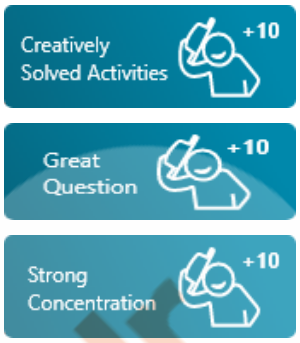
So, as you can see we have successfully performed the preprocessing of data on the given datasets. This is useful to analyze the data and get insights from data.

Teacher Guides Student to Stop Screen Share

WRAP-UP SESSION - 05 mins

Teacher Starts Slideshow  **Slide # to #** <**Note**: Only Applicable for Classes with VA>
Activity details **Following are the WRAP-UP session deliverables:** ● Appreciate the student. ● Revise the current class activities. ● Discuss the quizzes.
WRAP-UP QUIZ Click on In-Class Quiz
Continue WRAP-UP Session  **Slide # to #** <**Note**: Only Applicable for Classes with VA>
Activity Details **Following are the session deliverables:** ● Explain the facts and trivia ● Next class challenge ● Project for the day ● Additional Activity (Optional)
FEEDBACK ● **Appreciate and compliment the student for trying to learn a difficult concept.** ● **Get to know how they are feeling after the session.** ● **Review and check their understanding.**

Teacher Action	Student Action
You get “hats-off” for your excellent work!	*Make sure you have given at least 2 hats-off during the class for:*

In the next class, we'll be looking at more techniques of data cleaning i.e removing unwanted data.	
PROJECT OVERVIEW DISCUSSION Refer the document below in Activity Links Sections	
Teacher Clicks 	

ACTIVITY LINKS		
Activity Name	**Description**	**Links**
Teacher Activity 1	Boilerplate Code	https://colab.research.google.com/drive/1CCJOl8fn8WirjJ-DeqFwGewDLJeKGogU?usp=sharing
Teacher Activity 2	Exoplanet Exploration	https://exoplanets.nasa.gov/discovery/exoplanet-catalog/
Teacher Activity 3	11 Comae Berenices b	https://exoplanets.nasa.gov/exoplanet-catalog/6988/11-comae-berenices-b/
Teacher Activity 4	NASA Exoplanet Archive	https://exoplanetarchive.ipac.caltech.edu/docs/table-redirect.html
Teacher Activity 5	Dataset	https://github.com/procodingclass/PRO-NASA-Exoplanet-Scraped-Data
Teacher Activity 6	Reference Code	https://colab.research.google.com/drive/1ilGwmFoTz44f8x1oyHwFJBPmoE2-S99K?usp=sharing
Teacher Reference 1	Project	https://s3-whjr-curriculum-uploads.whjr.online/6479d832-9780-4e43-ba73-af1a226b078c.pdf
Teacher Reference 2	Project Solution	https://colab.research.google.com/drive/1mMOhmdKx-q40ZbbDoNG1jTBJ2RKMecMo?usp=sharing
Teacher Reference 3	Visual-Aid	Will be added after VA creation
Teacher Reference 4	In-Class Quiz	https://s3-whjr-curriculum-uploads.whjr.online/c704493f-5bfa-4707-9425-18c3d69ee258.pdf
Student Activity 1	Dataset	https://github.com/procodingclass/PRO-NASA-Exoplanet-Scraped-Data
Student Activity 2	Boilerplate code	https://colab.research.google.com/drive/10G98up218Tg8UG-NCetB2lHO9BvhpBo8?usp=sharing