Topic	Data Cleaning	
Class Description	**Students will clean the data that was retrieved in the previous classes.**	
Class	**C130**	
Class time	**45 mins**	
Goal	● Understanding and reviewing data ● Cleaning the data as per our need	
Resources Required	● Teacher Resources ○ Laptop with internet connectivity ○ Earphones with mic ○ Notebook and pen ● Student Resources ○ Laptop with internet connectivity ○ Earphones with mic ○ Notebook and pen	
Class structure	**Warm Up** **Teacher-led Activity** **Student-led Activity** **Wrap up**	**5 mins** **15 min** **15 min** **5 min**

CONTEXT

- **Review the concepts learned in the earlier classes**

Class Steps	Teacher Action	Student Action
Step 1: Warm Up (5 mins)	Hi <Student Name>! In the last class, we added more data and merged the two databases that we had. Can you recall the logic we used to merge the two CSVs?	**ESR:** - We first sorted the data from the second database in alphabetical order irrespective of whether it is uppercase or lowercase and then we merged the two!

	Great! Now in today's class, we will understand the meaning of all the columns that we have, in our CSV and then we will see how we can clean our data. This means how we can remove all the unwanted data to make it easier for us to use. Are you excited?	**ESR:** "Yes!"
	Before we get started I have an exciting quiz question for you! Are you ready to answer this question? Teacher click on the  button on the bottom right corner of your screen to start the In-Class Quiz. A quiz will be visible to both you and the student. Encourage the student to answer the quiz question. The student may choose the wrong option, help the student to think correctly about the question and then answer again. After the student selects the correct option, the  button will start appearing on your screen. Click the End quiz to close the quiz pop-up and continue the class.	**ESR:** yes

	alright ,Let's get started!	
Teacher Initiates Screen Share		
CHALLENGE • **Making the student understand the meaning of all the columns** • **Making the student understand how to clean the data**		
Step 2: Teacher-led Activity (15 min)	*(Before beginning the class, please make the student download the CSV from the link below. This CSV is the latest version of the data on which we have to perform data cleaning.)* *<Teacher can download from* ***Teacher Activity 1****>* https://github.com/whitehatjr/Data-cleaning/blob/master/final.csv	*<Student can download from* ***Student Activity 1****>*
	Let's just start by understanding the meaning of all the columns one by one.	

name - This is the name of the exo-planet.

light_years_from_earth - This is the distance of this planet from Earth in light years. 1 light year is the distance light can travel in one year, and light it super fast. It can travel 9.461 Trillion km in 1 year.

planet_mass - This is the mass of the planet with respect to Earth or Jupiter (Jupiter is the metric for Gas Giants while Earth is the metric for all other types of planets).

stellar_magnitude - This is the brightness of the host star of the planet when observed from Earth (just as the sun is our host star).

discovery_date - This is the year of discovery for the exo-planet.

hyperlink - This is just the hyperlink that we scraped.

planet_type - This is the type of the planet (Gas Giant, Super Earth, etc.).

temp_planet_date - This is a duplicate.

temp_planet_mass - This is another duplicate.

planet_radius - This is the radius of the exo-planet with respect to Earth or Jupiter.

orbital_radius - This is the average distance of this exo-planet from its sun. Just like our solar system has 1 sun, there are multiple solar systems that contain many planets and sun(s).

orbital_period - This is the time it takes to complete one orbit of it's sun.

eccentricity - This denotes how circular the orbit is. It might be oval in shape too. The lower the eccentricity, the more circular is the orbit.

pl_hostname - The name of the host solar system.

pl_letter - The letter given to this planet.

pl_name - The name of this planet (short version).

pl_discmethod - This is the discovery method which was used to find this exo-planet.

pl_controvflag - This is a boolean (0, 1) which says if the existence of this planet is questioned or not.

pl_pnum - This is the number of planets that are there in its solar system.

pl_orbper - This is again, the orbital period in days.

Now since we are collecting data for planets that exist so far away from us, there is no way for us to know the actual values of a planet, such as their orbital period, radius, etc. so we do calculations for it. Each calculation is based on observation such as here, can have a margin of error in the actual value. Thus, all the columns with ***err1*** *and* ***err2*** are the scope of errors, and we will ignore them.

pl_orbperlim - This is again the radius of the orbit of the planet.

pl_orbeccen - This is again the eccentricity of the planet.

pl_orbincl - This is the orbital inclination, which means that it is the tilt of the exo-planet's orbit when it revolves around its sun.

pl_bmassj - This is again the mass of the planet.

pl_bmassprov - This is the unit to calculate the mass.

pl_radj - This is again, the radius of the planet.

pl_dens - This is the density of the planet.

pl_ttvflag - This is a flag that indicates if this planet's orbit exhibits any timing variations from other planets in the system.

pl_kepflag - This is a flag that tells if the solar system exhibits a planetary system (multiple planets) based on ***Kepler Field Mission.***

pl_k2flag - This is a flag that tells if the solar system exhibits a planetary system based on the ***K2 Mission.***

pl_nnotes - This is just the number of notes associated with the planet.

ra_str - This is the right ascension of the planetary system, which is the east-west coordinate by which the position of this planet is measured.

dec_str - This is the north-south coordinate by which the position of the planet is measured.

st_dist - This is again the distance of the planet from Earth.

gaia_dist - This is again the distance of the planet from Earth in Gaia Parallax. Gaia Parallax is the coordinate that is calculated with Trigonometry.

st_optmag - This is the Optical magnitude (discussed earlier).

st_optband - There are different bands in light. This is the band of the optical magnitude.

gaia_gmag - This is the magnitude of the host star of the planet measured in G-Band.

st_teff - This is the temperature of the host star in Kelvin.

st_mass - This is the amount of mass contained in the host star.

st_rad - This is the radius of the host star.

rowupdate - This is the date of last update for this exo-planet. ***pl_facility -*** Facility at which the planet was discovered (There are many facilities that are observing and looking for new planets/stars in our galaxy).		
	Great! Now we understand the meaning of all the rows. Let's dive into the data-cleaning part now!	
	Let's start by creating a virtual environment in a new directory: **python3.8 -m venv venv** Let's source the virtual environment: MACOS/UBUNTU:- **source venv/bin/activate** WINDOWS:- **venv\Scripts\activate.bat**	*The student creates a virtual environment.*
	Okay! Now look at what I'm doing closely.	*The student observes and listens.*

	I will first import pandas as pd and then I will import csv to create the final output csv after cleaning this data.
	
	`import pandas as pd`
	`import csv`
	
	Now, also we need to make sure to do so.	

pip install pandas

Now, I will move my CSV (downloaded from the link above) from the previous class and read it in a dataframe. Do you remember what Dataframes are?

```
df = pd.read_csv("final.csv")
```

Now let's just print the shape of this dataframe.

```
print(df.shape)
```

If we run this script, we can see that the shape is printed to be:

(4284, 85)

This means that we have 4,284 rows (the same number as our exo-planets) and we have 85 columns. We will first begin with removing all the unwanted columns, as we have many of them.

```
import pandas as pd
import csv

df = pd.read_csv("final.csv")
print(df.shape)
```

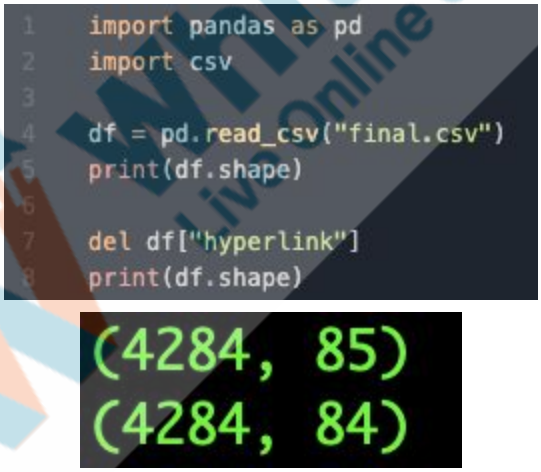
Teacher Stops Screen Share

Now it's your turn. Please share your screen with me.

- **Ask Student to press ESC key to come back to panel**

	• **Guide Student to start Screen Share** • **Teacher gets into Fullscreen**	
	ACTIVITY • **Student codes to remove unwanted columns** • **Student makes necessary changes to the data**	
Step 3: Student-Led Activity (15 min)	Okay! Now since we already have a dataframe, let's list down all the columns that we want to remove: ***hyperlink*** ***temp_planet_date*** ***temp_planet_mass*** ***pl_letter*** ***pl_name*** ***pl_controvflag*** ***pl_pnum*** ***pl_orbper*** ***pl_orbpererr1*** ***pl_orbpererr2*** ***pl_orbperlim*** ***pl_orbsmax*** ***pl_orbsmaxerr1*** ***pl_orbsmaxerr2*** ***pl_orbsmaxlim*** ***pl_orbeccen*** ***pl_orbeccenerr1*** ***pl_orbeccenerr2*** ***pl_orbeccenlim*** ***pl_orbinclerr1*** ***pl_orbinclerr2*** ***pl_orbincllim*** ***pl_bmassj*** ***pl_bmassjerr1*** ***pl_bmassjerr2*** ***pl_bmassjlim***	

	pl_bmassprov ***pl_radj*** ***pl_radjerr1*** ***pl_radjerr2*** ***pl_radjlim*** ***pl_denserr1*** ***pl_denserr2*** ***pl_denslim*** ***pl_ttvflag*** ***pl_kepflag*** ***pl_k2flag*** ***pl_nnotes*** ***ra*** ***dec*** ***st_dist*** ***st_disterr1*** ***st_disterr2*** ***st_distlim*** ***gaia_dist*** ***gaia_disterr1*** ***gaia_disterr2*** ***gaia_distlim*** ***st_optmag*** ***st_optmagerr*** ***st_optmaglim*** ***st_optband*** ***gaia_gmag*** ***gaia_gmagerr*** ***gaia_gmaglim*** ***st_tefferr1*** ***st_tefferr2*** ***st_tefflim*** ***st_masserr1*** ***st_masserr2*** ***st_masslim*** ***st_raderr1*** ***st_raderr2***	

	st_radlim ***rowupdate*** ***pl_facility***	
	That is a lot of columns to remove! All of it is the data that we do not require, hence we can delete it. To remove a column from a dataframe, we do: `del df["hyperlink"]` This will remove the **hyperlink** column from the dataframe! Let's print the shape of our DF after this line to cross check!	*Student observes.*
		

	Now, Let us delete all the columns that we do not want in a similar way!
	
	`del df["hyperlink"]`
	`del df["temp_planet_date"]`
	`del df["temp_planet_mass"]`
	`del df["pl_letter"]`
		Student deletes all the columns.

```
del df["pl_name"]
del df["pl_controvflag"]
del df["pl_pnum"]
del df["pl_orbper"]
del df["pl_orbpererr1"]
del df["pl_orbpererr2"]
del df["pl_orbperlim"]
del df["pl_orbsmax"]
del df["pl_orbsmaxerr1"]
del df["pl_orbsmaxerr2"]
del df["pl_orbsmaxlim"]
del df["pl_orbeccen"]
del df["pl_orbeccenerr1"]
del df["pl_orbeccenerr2"]
del df["pl_orbeccenlim"]
del df["pl_orbinclerr1"]
del df["pl_orbinclerr2"]
del df["pl_orbincllim"]
del df["pl_bmassj"]
del df["pl_bmassjerr1"]
del df["pl_bmassjerr2"]
del df["pl_bmassjlim"]
del df["pl_bmassprov"]
del df["pl_radj"]
del df["pl_radjerr1"]
del df["pl_radjerr2"]
del df["pl_radjlim"]
del df["pl_denserr1"]
del df["pl_denserr2"]
del df["pl_denslim"]
del df["pl_ttvflag"]
del df["pl_kepflag"]
del df["pl_k2flag"]
del df["pl_nnotes"]
del df["ra"]
del df["dec"]
del df["st_dist"]
```

```
del df["st_disterr1"]
del df["st_disterr2"]
del df["st_distlim"]
del df["gaia_dist"]
del df["gaia_disterr1"]
del df["gaia_disterr2"]
del df["gaia_distlim"]
del df["st_optmag"]
del df["st_optmagerr"]
del df["st_optmaglim"]
del df["st_optband"]
del df["gaia_gmag"]
del df["gaia_gmagerr"]
del df["gaia_gmaglim"]
del df["st_tefferr1"]
del df["st_tefferr2"]
del df["st_tefflim"]
del df["st_masserr1"]
del df["st_masserr2"]
del df["st_masslim"]
del df["st_raderr1"]
del df["st_raderr2"]
del df["st_radlim"]
del df["rowupdate"]
del df["pl_facility"]
```

And now, our output should look something like this:

```
(4284, 85)
(4284, 19)
```

We deleted 66 Columns!

	Okay! Now let's check the names of all the columns with the following code: `print(list(df))` This will give us the list of all the headers.	*Student prints the headers.*

```
['name', 'light_years_from_earth', 'planet_mass', 'stellar_magnitude', 'discovery_date',
'planet_type', 'planet_radius', 'orbital_radius', 'orbital_period', 'eccentricity', 'pl_h
ostname', 'pl_discmethod', 'pl_orbincl', 'pl_dens', 'ra_str', 'dec_str', 'st_teff', 'st_m
ass', 'st_rad']
```

	Let's change the name of these headers and make them more readable. Our headers look fine up until eccentricity. We will change the others with the following code:
	
	`df = df.rename({`
	`'pl_hostname': "solar_system_name",`
	`'pl_discmethod': "planet_discovery_method",`
	`'pl_orbincl': "planet_orbital_inclination",`
	`'pl_dens': "planet_density",`
	`'ra_str': "right_ascension",`
	`'dec_str': "declination",`
	`"st_teff": "host_temperature",`
		Student changes the headers and checks the result.

```
                'st_mass':
"host_mass",
                'st_rad':
"host_radius"
            }, axis='columns')
```

Here, this will change all the headers. If we now print the list of headers, it will be something like this:

```
['name', 'light_years_from_earth', 'planet_mass', 'stellar_magnitude', 'discovery_date',
'planet_type', 'planet_radius', 'orbital_radius', 'orbital_period', 'eccentricity', 'sola
r_system_name', 'planet_discovery_method', 'planet_orbital_inclination', 'planet_density'
, 'right_ascension', 'declination', 'host_temperature', 'host_mass', 'host_radius']
```

Great! Our data looks much more clean and readable now. We have reduced the number of columns from 85 to 19 and we have made our headers much more readable. It looks like our data is ready to be used to perform statistics. One last thing that we need to do is that we have to create a main CSV which we are going to use for our statistics. Let's do that!

The student creates a csv from the given dataframe.

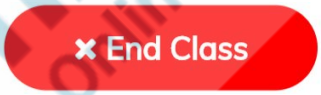
```
df.to_csv('main.csv')
```

We are done!

Teacher Guides Student to Stop Screen Share

FEEDBACK

- **Appreciate the student for their efforts**
- **Identify 2 strengths and 1 area of progress for the student**

Step 4: Wrap-Up (5 min)	So, in this class, we completed the pre-requisites of how we prepare our data before performing any statistics. One thing that we are yet to do is that we have to make our data uniform, but we will do it as we perform statistics and build models. How was your experience?	**ESR:** varied
	Congratulations! You are now prepared to find your own datasets and prepare it for your research/case studies! Next class, we will be applying some statistics to this data to see if we can find something interesting.	-
Teacher Clicks 		

Activity	Activity Name	Links
Teacher Activity 1	Data from last class	https://github.com/whitehatjr/Data-cleaning/blob/master/final.csv
Teacher Activity 2	Solution	https://github.com/whitehatjr/Data-cleaning
Student Activity 1	Data from last class	https://github.com/whitehatjr/Data-cleaning/blob/master/final.csv