

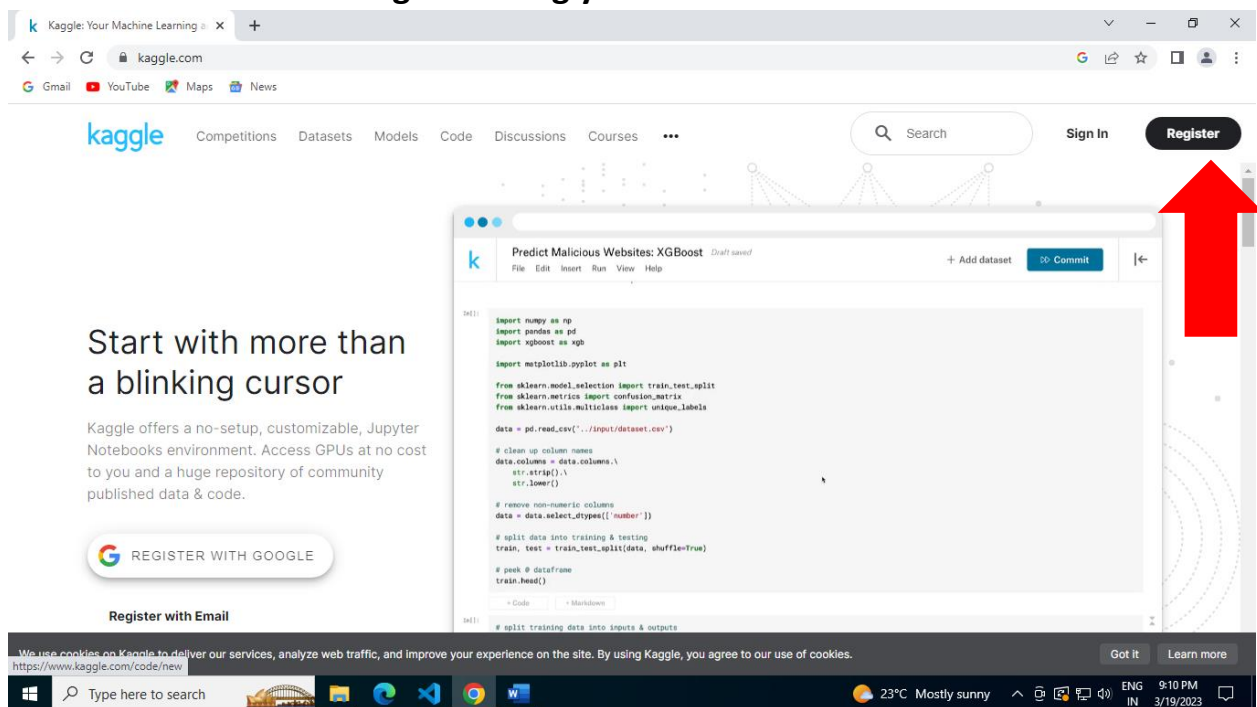
Practical 1: DSBDA

Tutorial

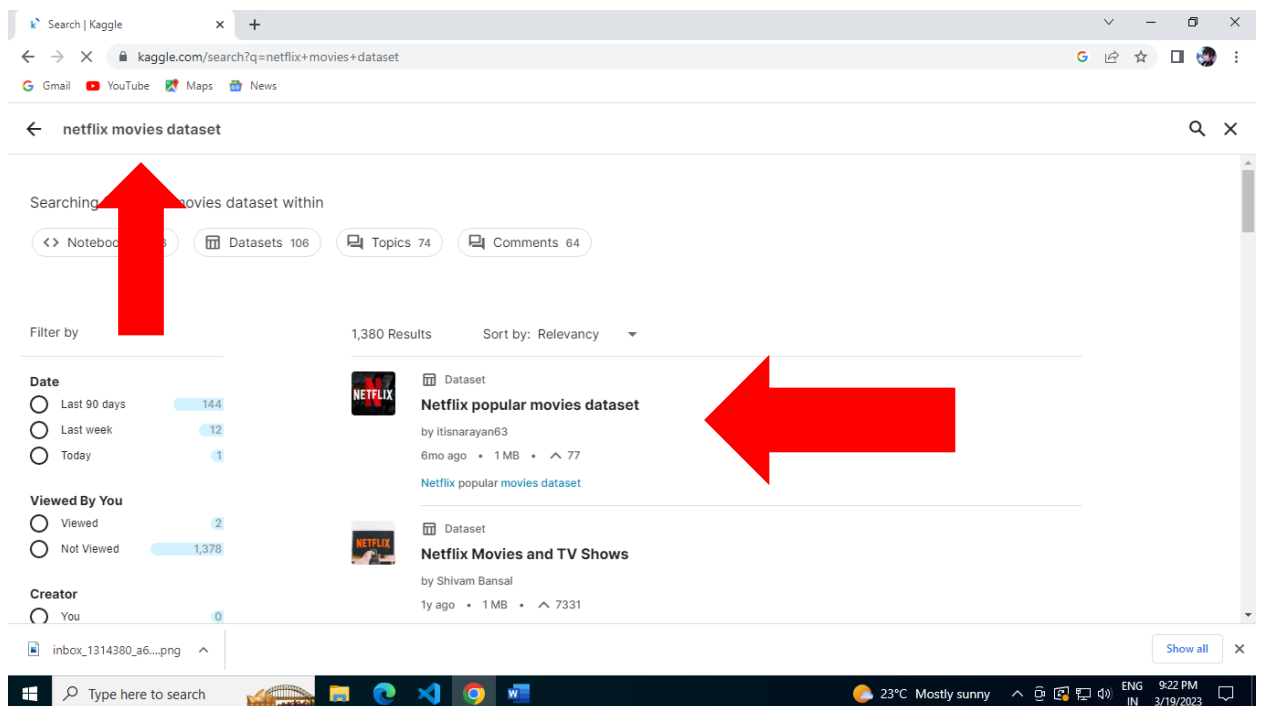
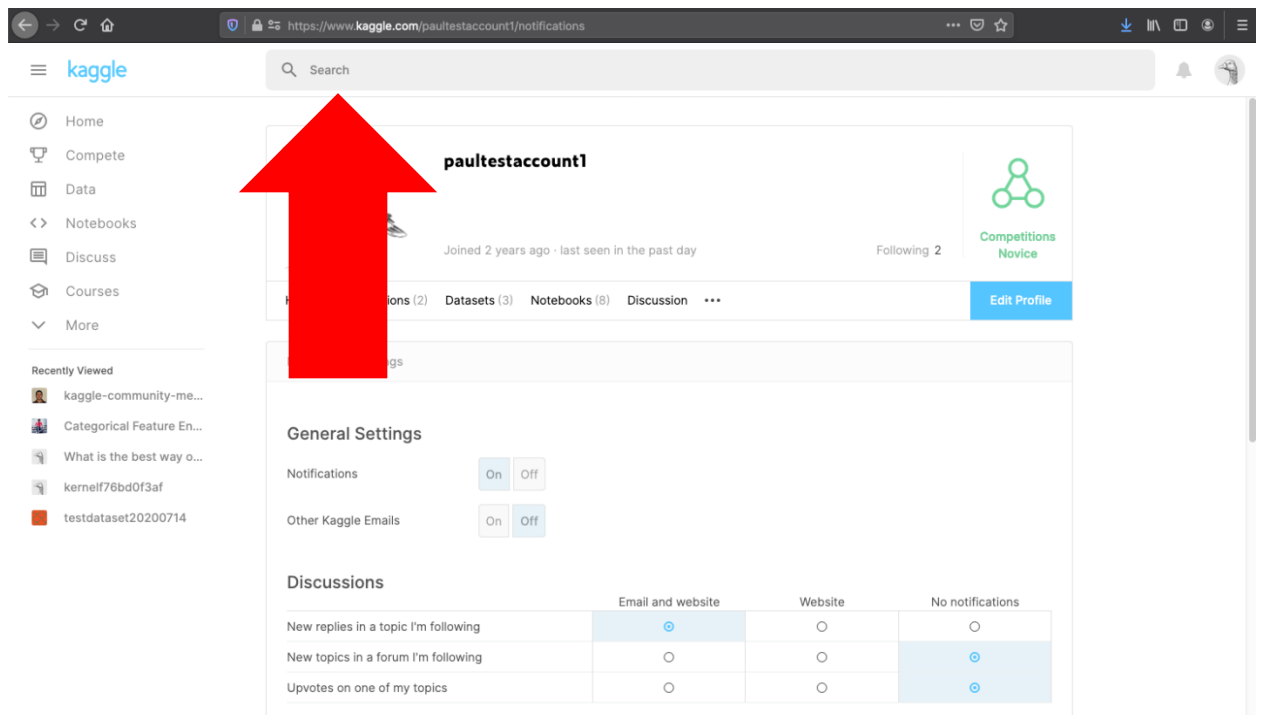
The first practical is of data wrangling. We need to obtain data sets first.

Steps:

1. First, open the web browser and go to <https://www.kaggle.com>
2. On the website register using your email

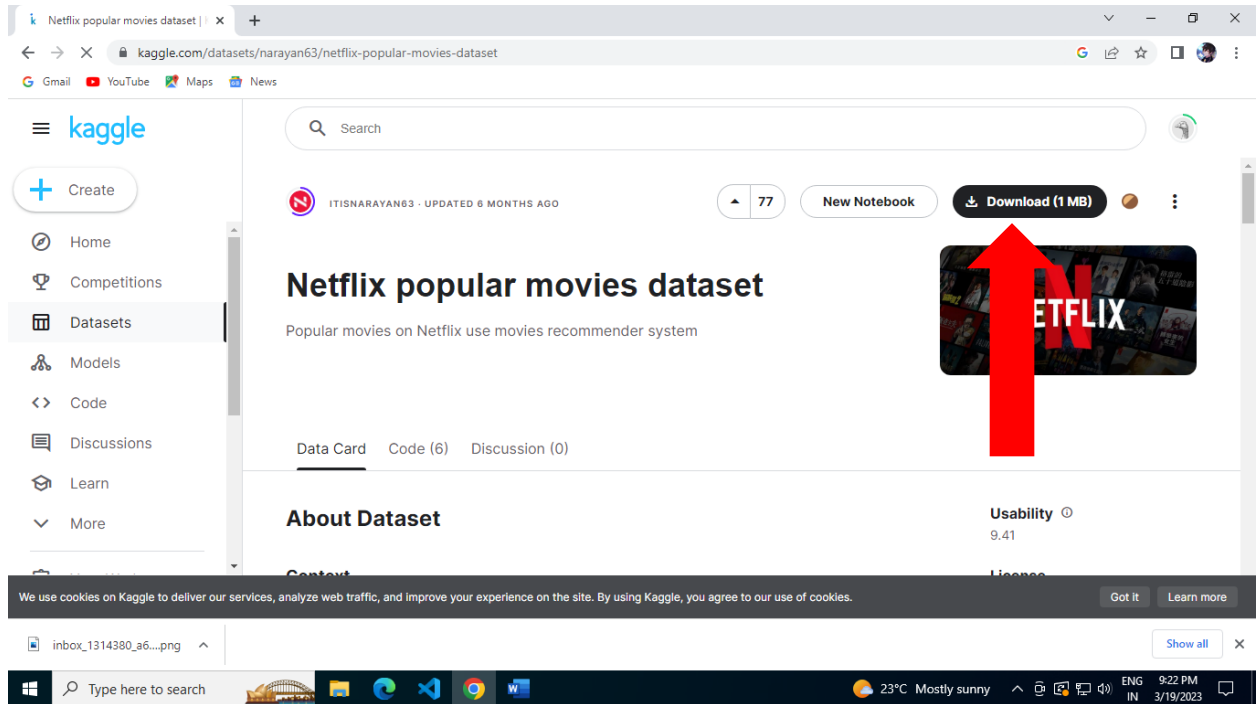


3. After registering you'll be automatically signed in, for those who have already registered can simply sign in using their account. Now go to the search bar on the top of the page and search for any datasets of your choice in the csv format.

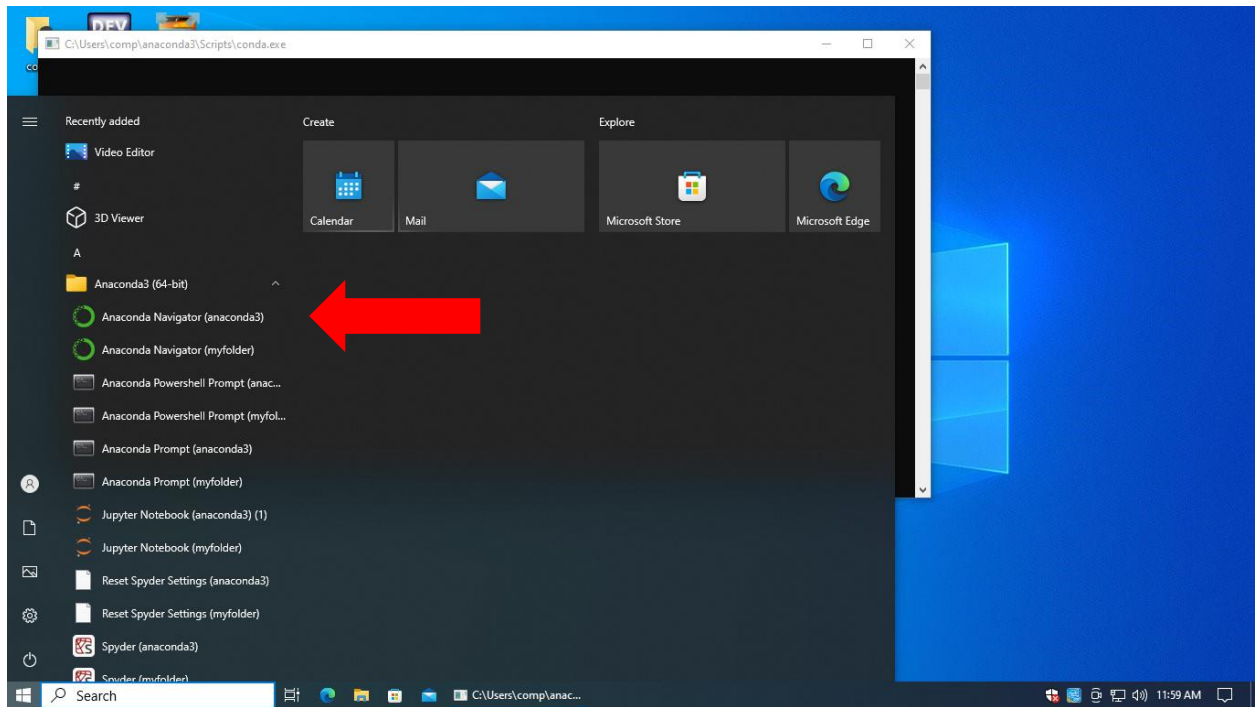


- After you've entered your desired dataset in the search bar click on the dataset you want from the list. You can also filter and sort by using the left-hand side of the website.

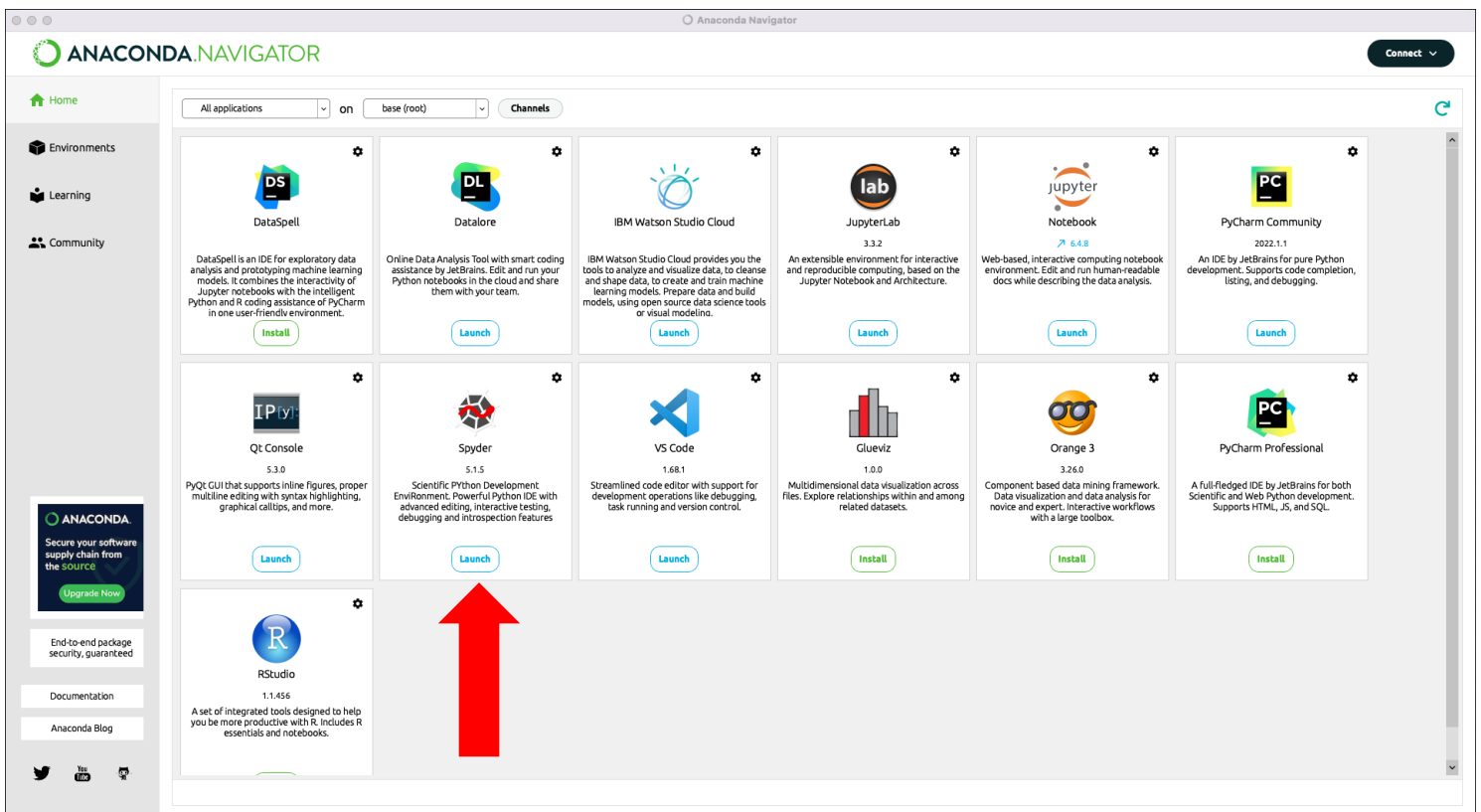
5. On the selected page click on download in the top right-hand side.



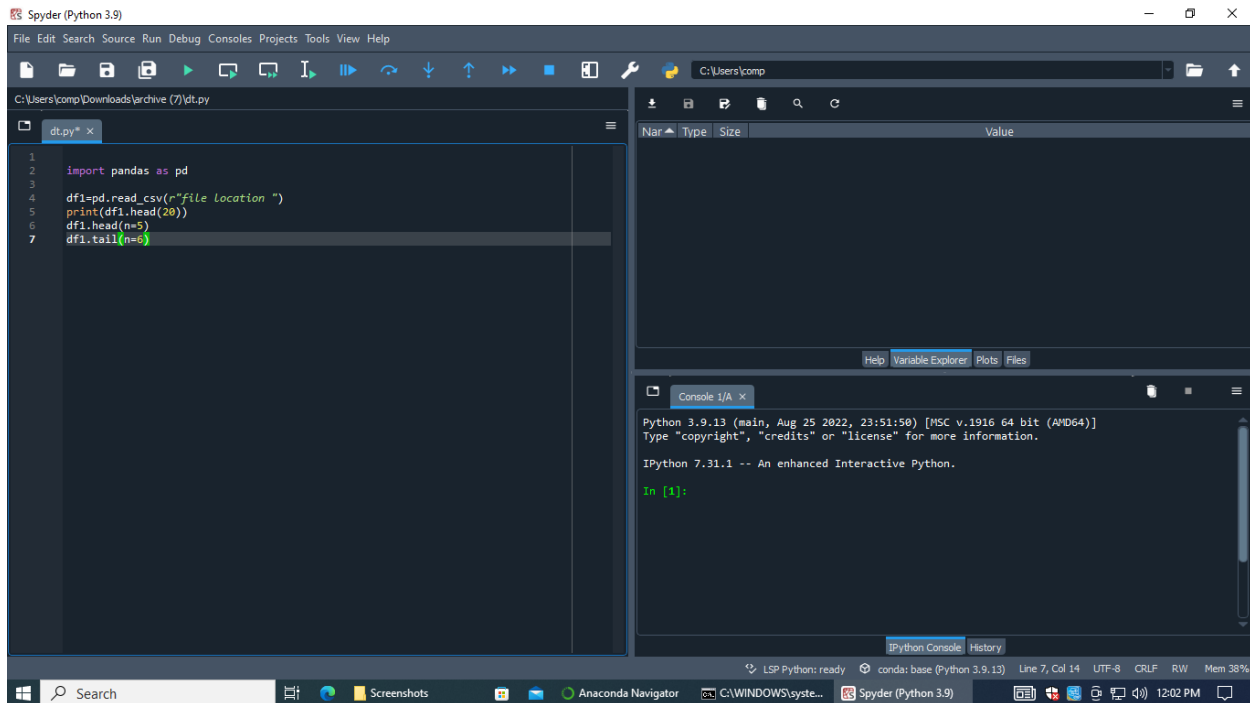
6. If the download is in a zipped folder then unzip it using any of the softwares like winzip, winrar, 7zip etc.(download if not installed, link for 7 zip : <https://www.7-zip.org/>)
7. After successfully downloading the dataset , head over to the browser once again and download anaconda.(link : <https://www.anaconda.com/>)
8. Click on download for windows, and install it.
9. After installing anaconda head to the start menu and type 'Anaconda Navigator'. Click on the first option with the green circle anaconda logo.



10. After the navigator opens, find spyder and launch it.



11. After opening spider, a command prompt and a Editing GUI will appear. You can code in python in this GUI. As we have installed anaconda it'll automatically install python before opening Spyder.



12. After installing python, we can start to code. Create a new file and import the panda's library using the lines:

"Import pandas as pd"

13. On the next line create a data frame using the syntax:

Name of data frame = `pd.read_csv(r' file location in the pc ')`

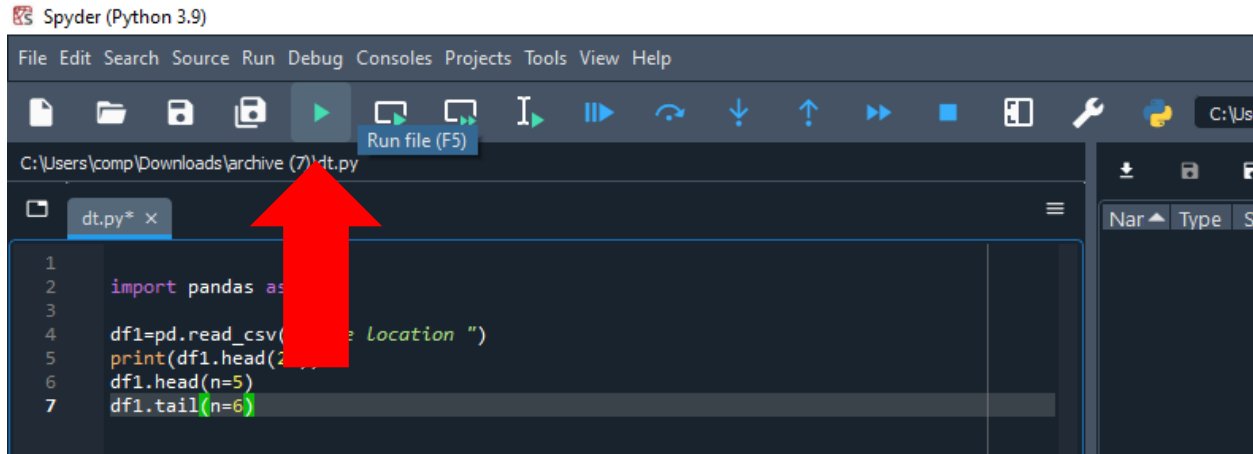
for eg:

`df1= pd.read_csv(r'C:\Users\comp\Documents\sushant\n_movies.csv')`

here *"n_movies.csv"* is the name of the dataset.

14. On the next line print the dataset by typing: *"print(df1) "*

15. Now click on the run button (green triangle) on the navbar.



16.The console will now display the name and the contents of the dataset.

17. Now after the dataset is successfully displayed, we need to perform operations on the database. Here is the list of all operations to be performed:

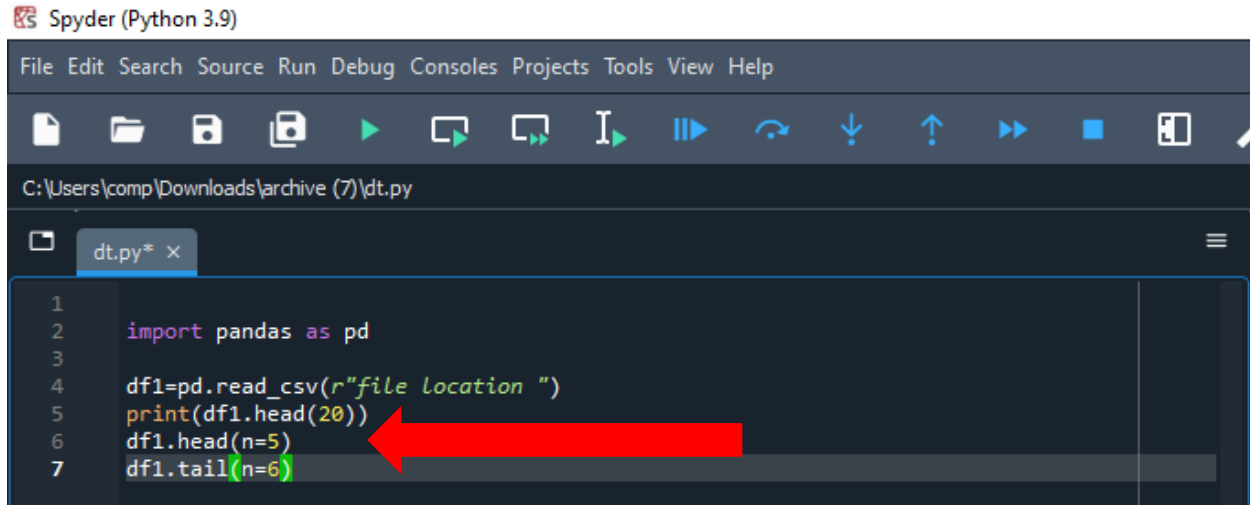
5. Panda Dataframe functions for Data Preprocessing :

Dataframe Operations:

Sr. No	Data Frame Function	Description
1	dataset.head(n=5)	Return the first n rows.
2	dataset.tail(n=5)	Return the last n rows.
3	dataset.index	The index (row labels) of the Dataset.
4	dataset.columns	The column labels of the Dataset.
5	dataset.shape	Return a tuple representing the dimensionality of the Dataset.
6	dataset.dtypes	Return the dtypes in the Dataset. This returns a Series with the data type of each column. The result's index is the original Dataset's columns.

		Columns with mixed types are stored with the object dtype.
7	<code>dataset.columns.values</code>	Return the columns values in the Dataset in array format
8	<code>dataset.describe(include='all')</code>	<p>Generate descriptive statistics. to view some basic statistical details like percentile, mean, std etc. of a data frame or a series of numeric values.</p> <p>Analyzes both numeric and object series, as well as Dataset column sets of mixed data types.</p>
9	<code>dataset['Column name']</code>	Read the Data Column wise.
10	<code>dataset.sort_index(axis=1, ascending=False)</code>	Sort object by labels (along an axis).
11	<code>dataset.sort_values(by="Column name")</code>	Sort values by column name.
12	<code>dataset.iloc[5]</code>	Purely integer-location based indexing for selection by position.
13	<code>dataset[0:3]</code>	Selecting via [], which slices the rows.
14	<code>dataset.loc[:, ["Col_name1", "col_name2"]]</code>	Selection by label
15	<code>dataset.iloc[:n, :]</code>	a subset of the first n rows of the original data
16	<code>dataset.iloc[:, :n]</code>	a subset of the first n columns of the original data
17	<code>dataset.iloc[:m, :n]</code>	a subset of the first m rows and the first n columns

18. Here dataset is the name of the data base, after writing the code mentioned in step 12,13, and 14 start writing the first function :
`df1.head(n=5)` below the print statement.

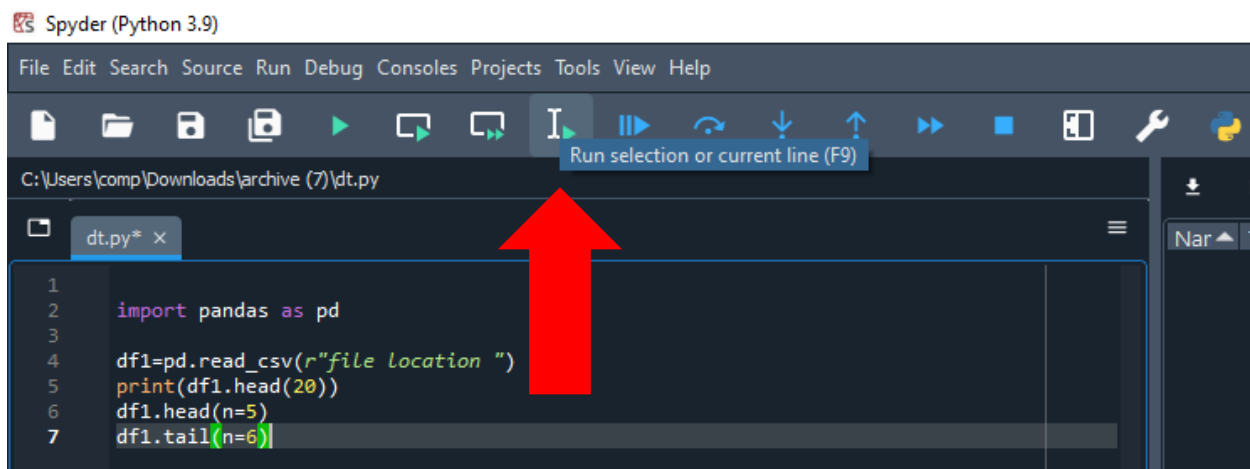


The screenshot shows the Spyder Python IDE interface. The file explorer on the left shows a file named `dt.py`. The editor window displays the following code:

```
1
2 import pandas as pd
3
4 df1=pd.read_csv(r"file location ")
5 print(df1.head(20))
6 df1.head(n=5)
7 df1.tail(n=6)
```

A red arrow points to the `df1.head(n=5)` line, indicating it should be executed.

19. Now execute this line by clicking on the “ execute current line button ” on the nav bar.



The screenshot shows the Spyder Python IDE interface. The file explorer on the left shows a file named `dt.py`. The editor window displays the same code as in the previous screenshot. A red arrow points to the 'Run selection or current line (F9)' button in the toolbar, indicating it should be clicked to execute the current line.

20. By doing so your function will be executed and output will be displayed in the console.
21. Similarly write the next function below the previous one and execute current line.
22. Repeat step 21 till all the functions have been executed successfully.
23. At last, copy the entire code and the output and save it in a text file.
24. Save your file and exit anaconda.