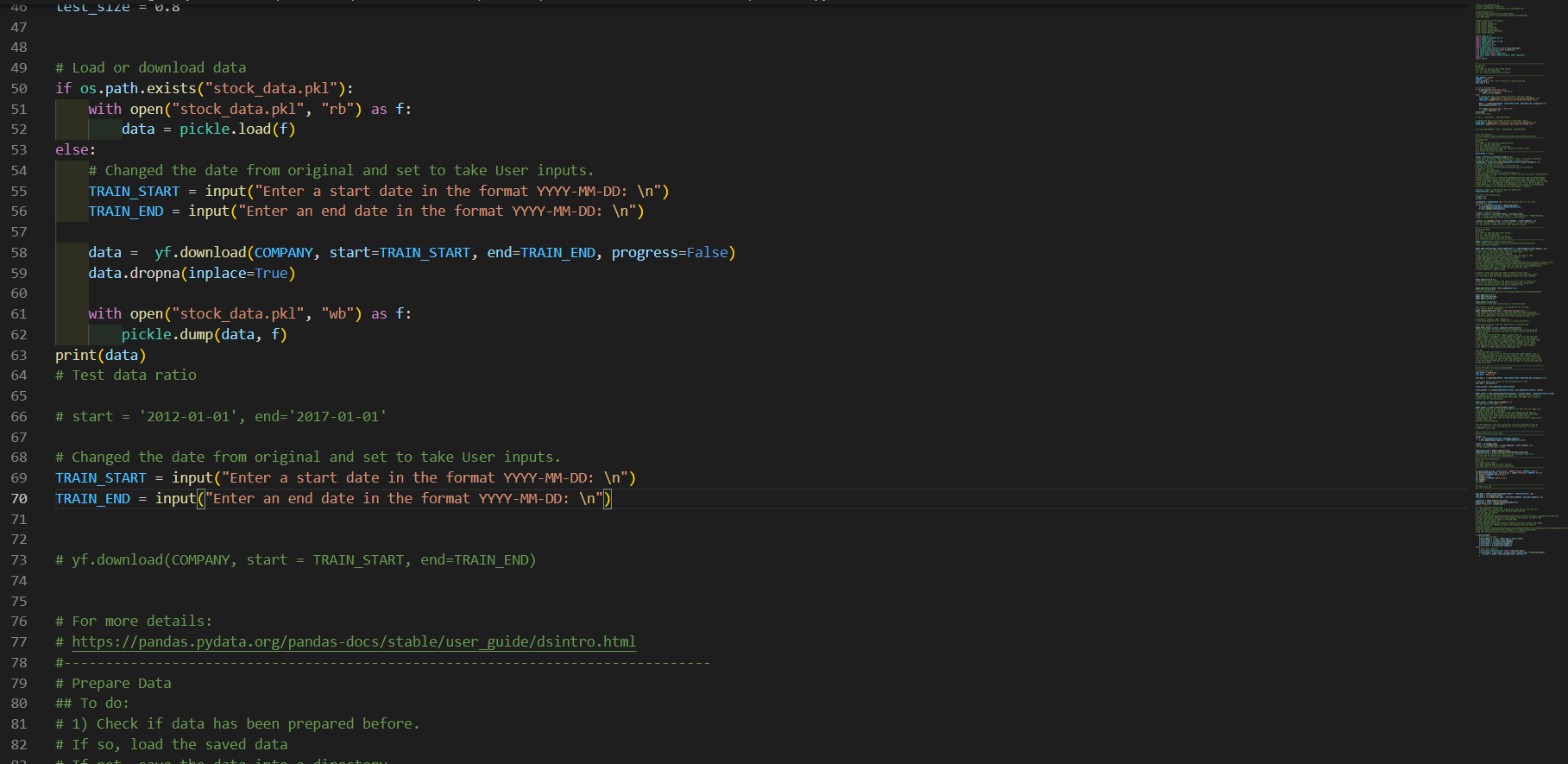
**COS30018 - Option B - Task 2: Data processing 1**

This tasks for this week are as follows :

a. This function will allow you to specify the start date and the end date for the whole

dataset as inputs



In this function I have changed the values of the “Train\_Start” and “Train\_End” variables to take user inputs instead of hard coding the dates and specified the format as YYYY-MM-DD. This function checks the date entered and then downloads the data based on the dates specidied and then the training and testing is done.

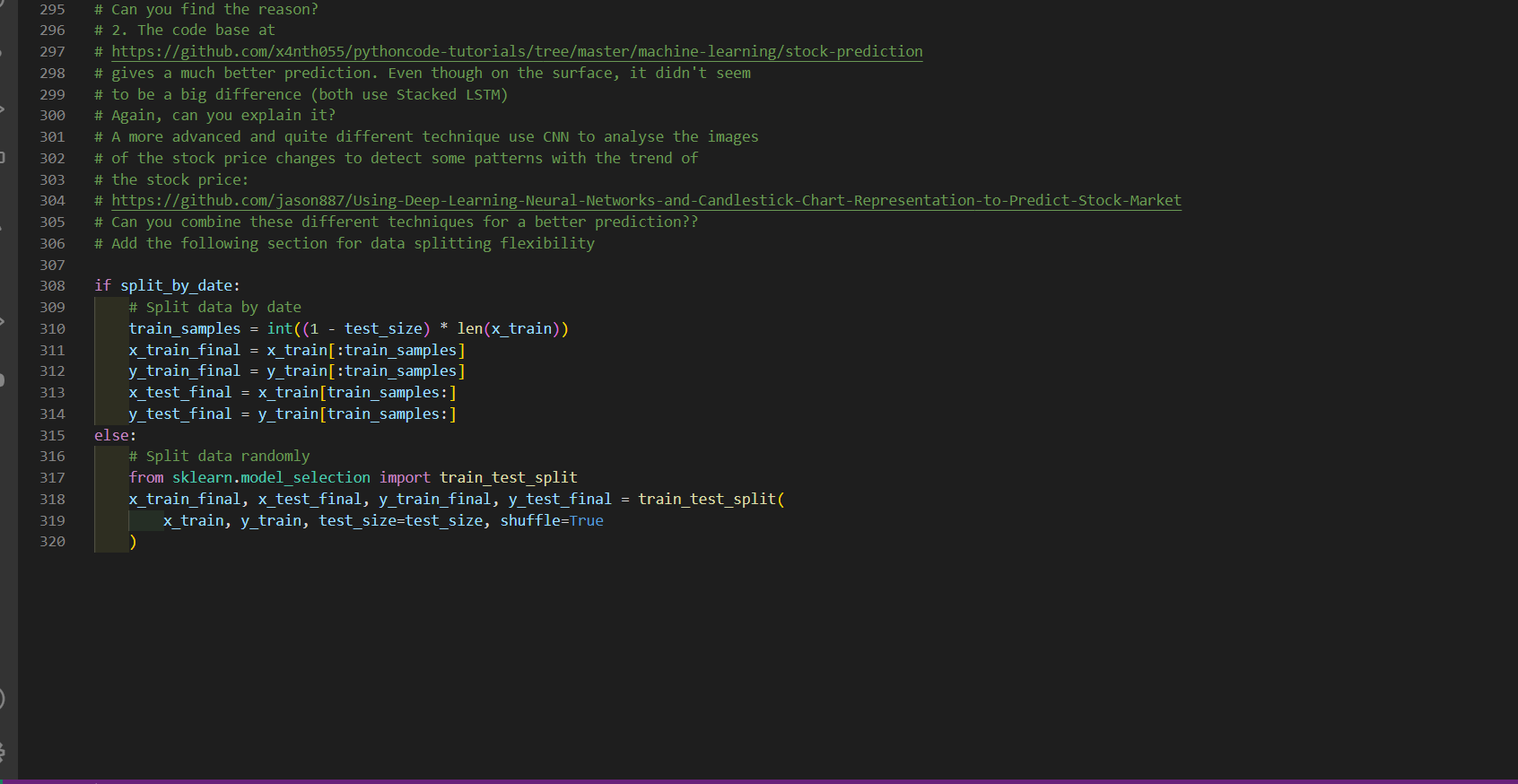
b. This function will allow you to deal with the NaN issue in the data

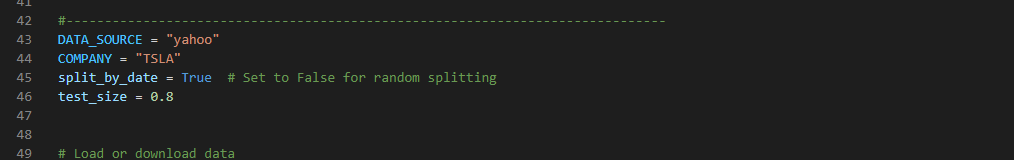
This was present in the example code that was given (P1) I just had to add a single line that is data.dropna( inplace = true), which deals wit the issues for any null values or NaN.

c. This function will also allow you to use different methods to split the data into

train/test data; e.g. you can split it according to some specified ratio of train/test and

you can specify to split it by date or randomly.

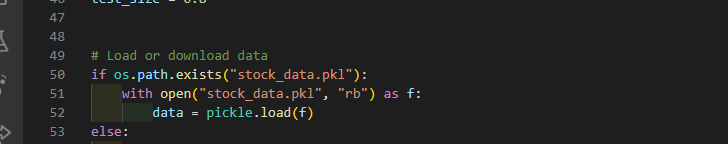


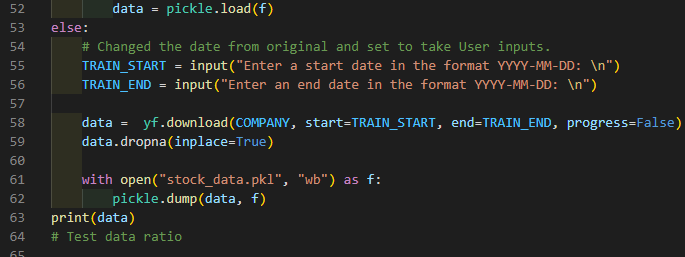
This was a tricky function to do at first, however in this function , I imported a library train\_test\_split from sklearn. This function by default is set to true which will split the data by Time. If it is set to false then it will split the data randomly as the if else condition is made by me. The “ratio” or the “test\_size” is also set to 0.8 by default so that it takes only 80 percent of data for training and the rest is used for testing purposes.

This was also present in the example code that was given to us, with some simple modifications it runs as required .

d. This function will have the option to allow you to store the downloaded data on your

local machine for future uses and to load the data locally to save time.



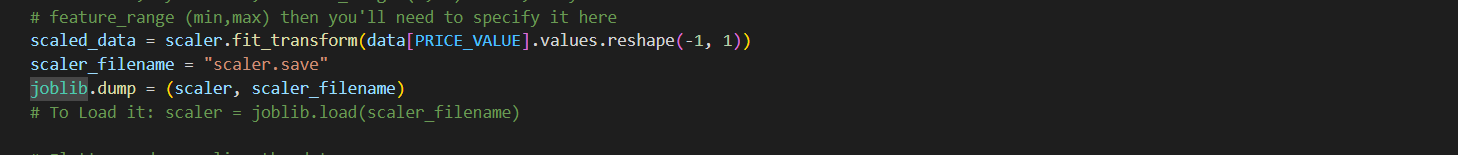
This function was made by me importing another library called pickle and import os. I have made this function so that it saves the last configuration and then loads it automatically the next time we want to use it. In the above example we can see that the data is saved in a file name “stock\_data.pkl” and there is an if condition applied with it . If the path with tha file name exists it will load the last saved file and start from there to save time.

If there is no such file in the directory it will simply ask for the user input for dates and run the program from the beginning.

This idea was taken from the internet and the explanation is provided in: <https://www.askpython.com/python/examples/save-data-in-python> .

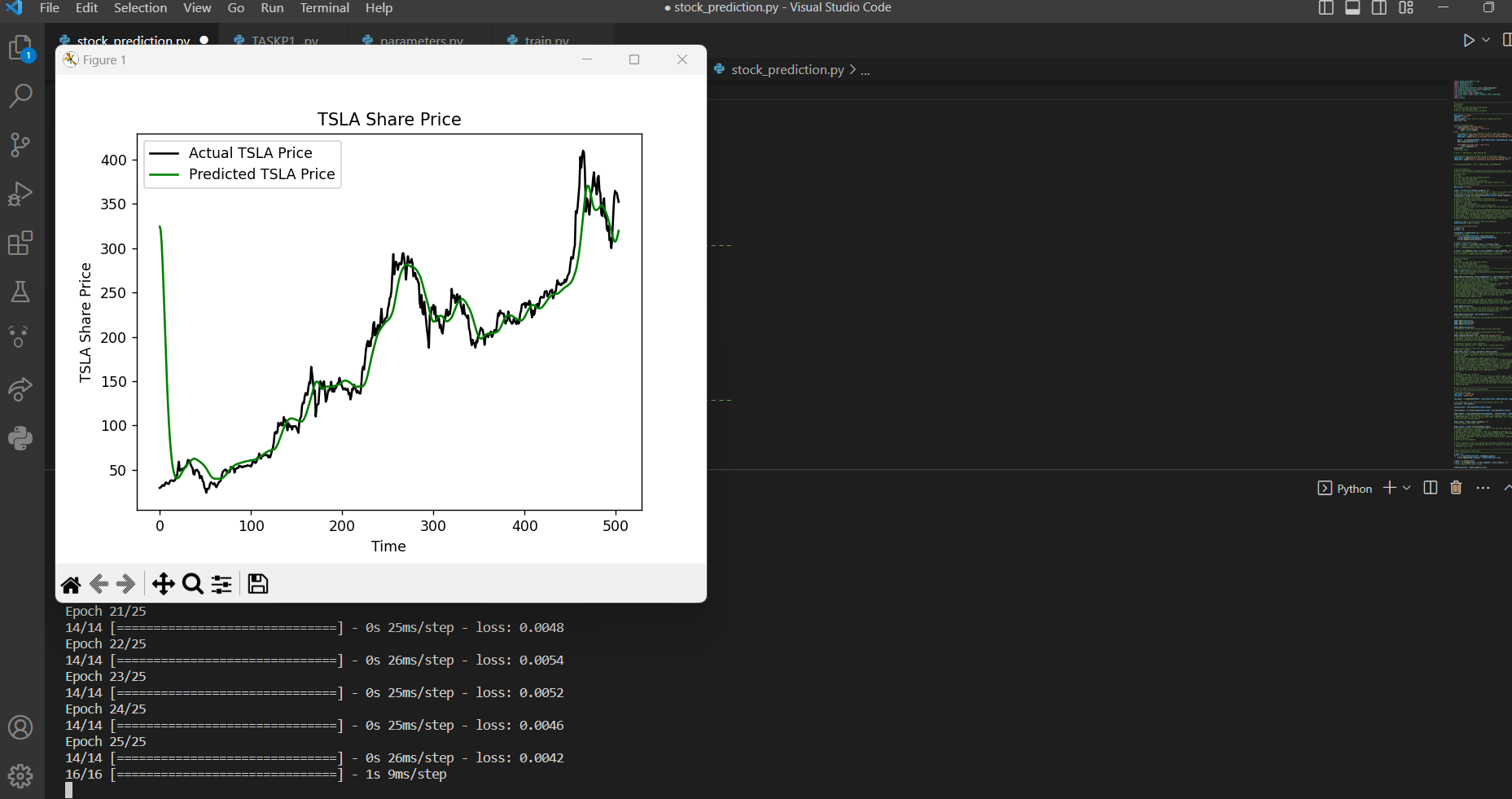
e. This function will also allow you to have an option to scale your feature columns and

store the scalers in a data structure to allow future access to these scalers

At last , in this function , I have imported another library called as joblib .

This provides just two lines of code that saves our scaler function into a file named “scaler.save”

To load this, I have also added the comment which loads the saved data.

Additional Observations: 

This is the final output with the current version of code. I noticed that the curve is not as accurate as previous version because I set the user input for only two years of data, where as in the previous task it was about five years. Therefore, using the functionality of the “ratio”, only 80% data was used for the training and 20% was used for testing. Hope this assignment demonstrates my clear understanding of the projects and it also fulfills the requirements that were given in this task. :)