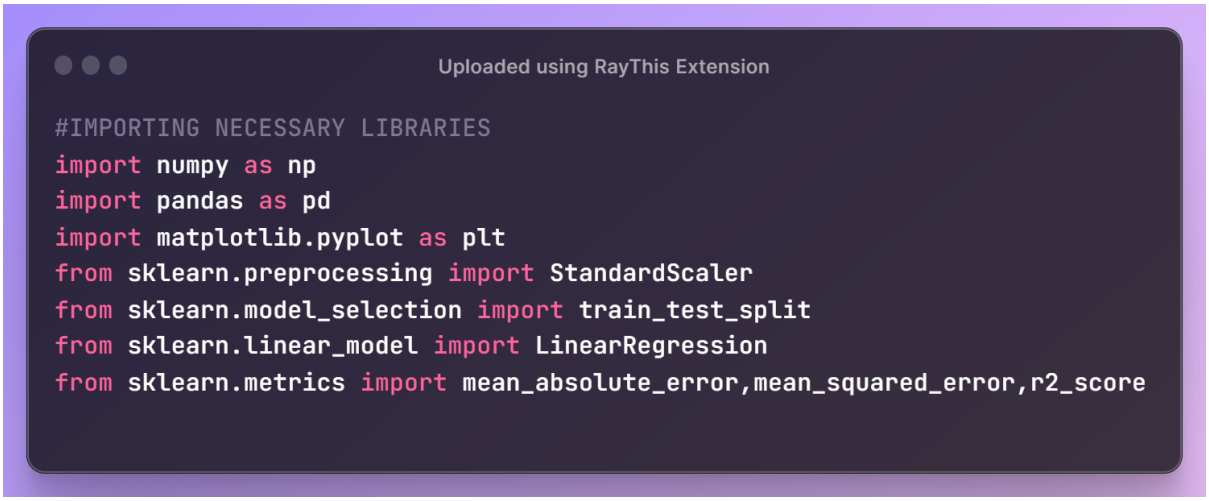


Over here we are importing the libraries,

1.Numpy is for sorting the data into an array.

2.Pandas is for reading the data.

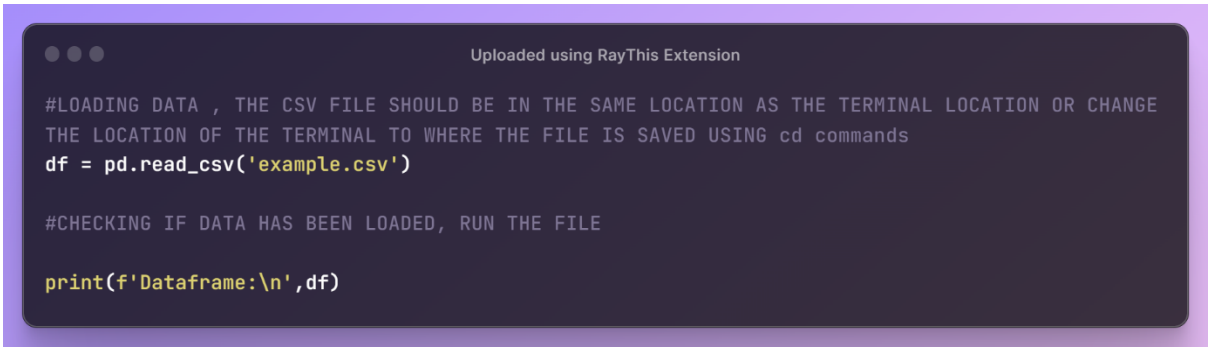
3.Matplotlib is for visualization but we don't have to import it, so you can skip this line.

4.Also the sklearn.preprocessing import Standard Scaler is also supposed to be skipped.

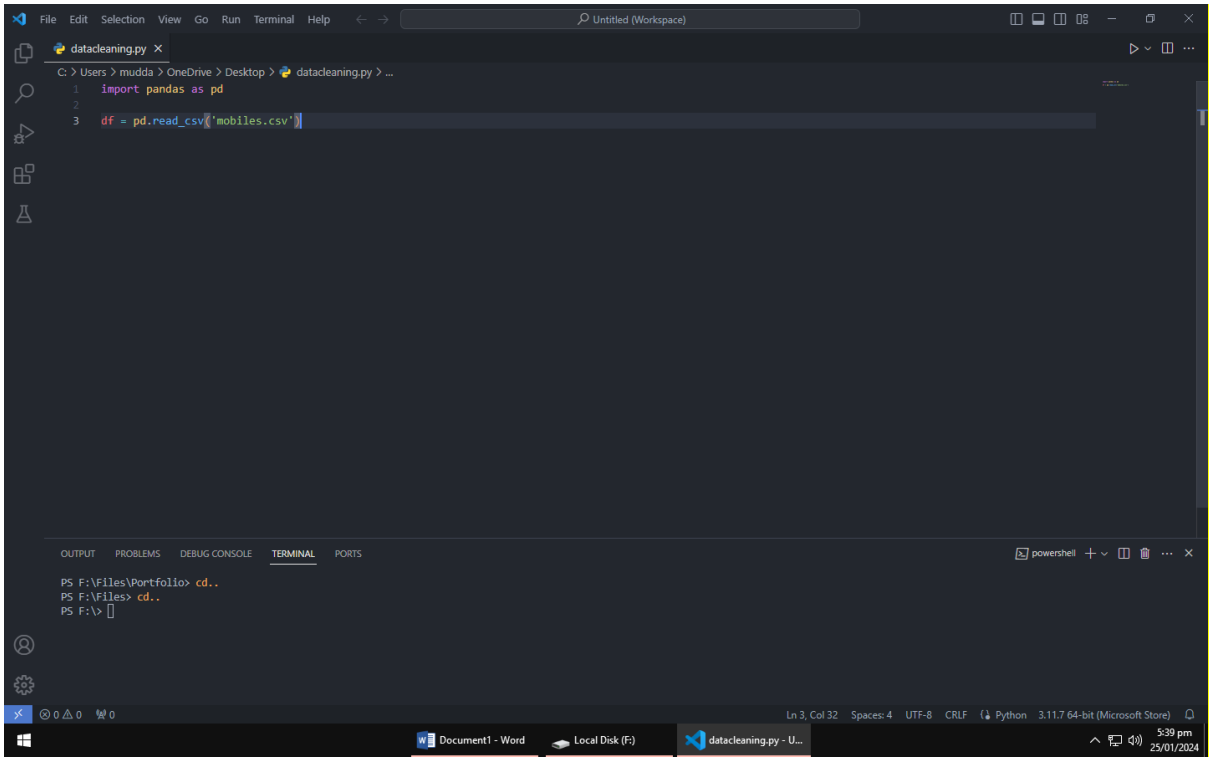
5.Sklearn model_selection line is necessary in this line we split the data into training and test set this is necessary for checking if the model performs well on data on which it has not been trained on.

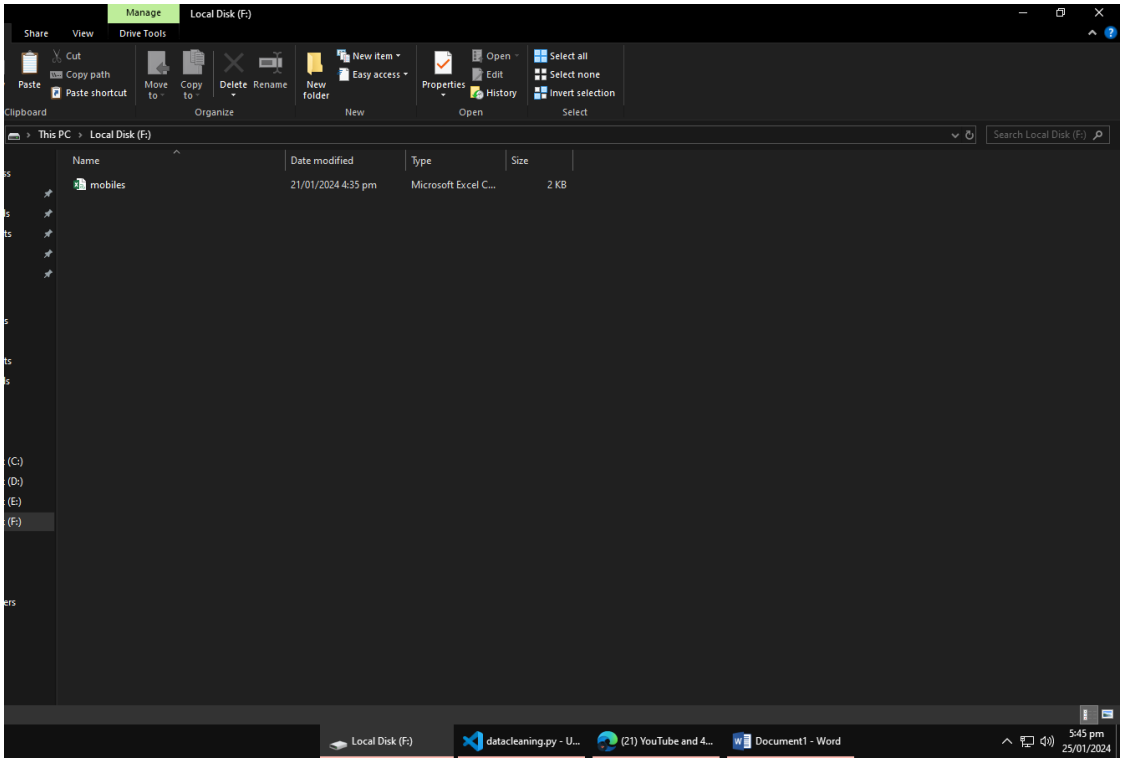
6.Sklearn.metrics is used for calculating various scores like mean squared error, mean absolute error and r2 score, these scores are used to improve the model's performance

I will also send some videos for train test split and metrics.



This is an important step here we import the data into python file. The condition is that instead of example.csv you should name the file you are work on for example if you have mobile dataset it should be named mobiles.csv or mobile.csv, meaning the name of the file should be the same in code with extension as it is in the folder.

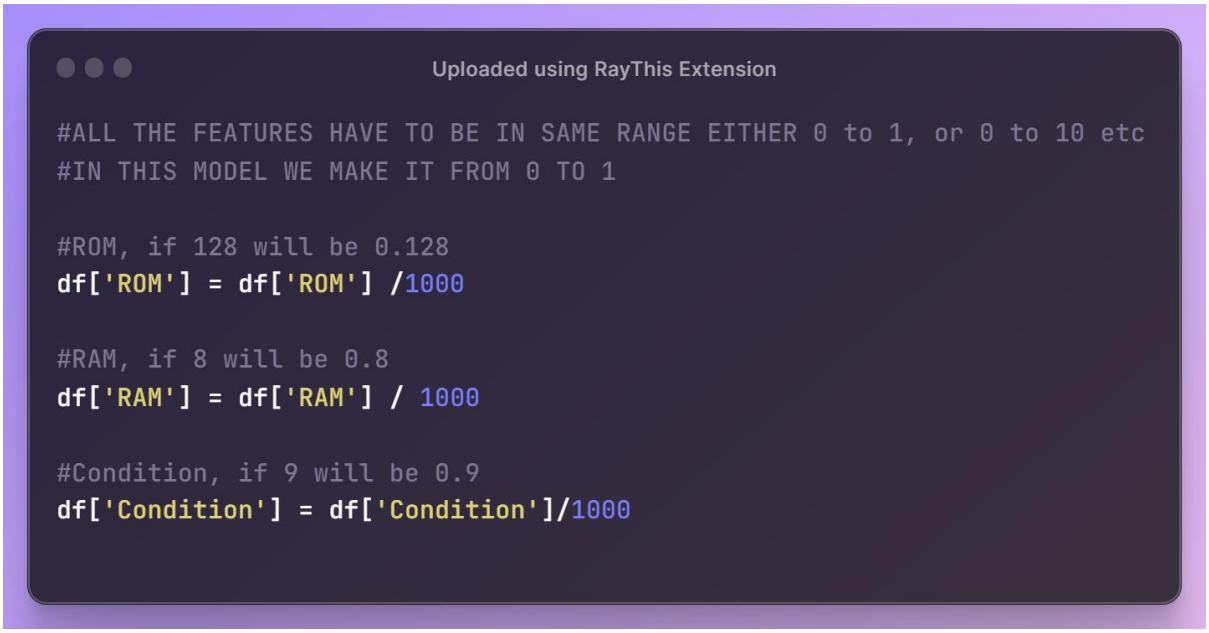

File Edit Selection View Go Run Terminal Help
Untitled (Workspace)
datacleaning.py
C: > Users > mudda > OneDrive > Desktop > datacleaning.py > ...
1 import pandas as pd
3 df = pd.read_csv('mobiles.csv')
OUTPUT PROBLEMS DEBUG CONSOLE TERMINAL PORTS
powershell
PS F:\Files\Portfolio> cd..
PS F:\Files> cd..
PS F:\>
Ln 3, Col 32 Spaces: 4 UTF-8 CRLF Python 3.11.7 64-bit (Microsoft Store)
Document1 - Word
Local Disk (F:)
datacleaning.py - U...
5:39 pm
25/01/2024


Manage
Local Disk (F:)
Share View Drive Tools
Paste Cut Copy path Paste shortcut
Move to Copy to Delete Rename
New folder New item Easy access
Properties Open Edit History
Select all Select none Invert selection
Clipboard Organize New Open Select
This PC > Local Disk (F:)
Search Local Disk (F:)
Name Date modified Type Size
mobiles 21/01/2024 4:35 pm Microsoft Excel C... 2 KB
(C:)
(D:)
(E:)
(F:)
Local Disk (F:)
datacleaning.py - U...
(21) YouTube and 4...
Document1 - Word
5:45 pm
25/01/2024

Another Important thing for loading the files is that the files should be in the same location as the terminal here you see that I have placed the files in Local Disk F, and the location where my terminal in VSCODE is at is also Local Disk F.

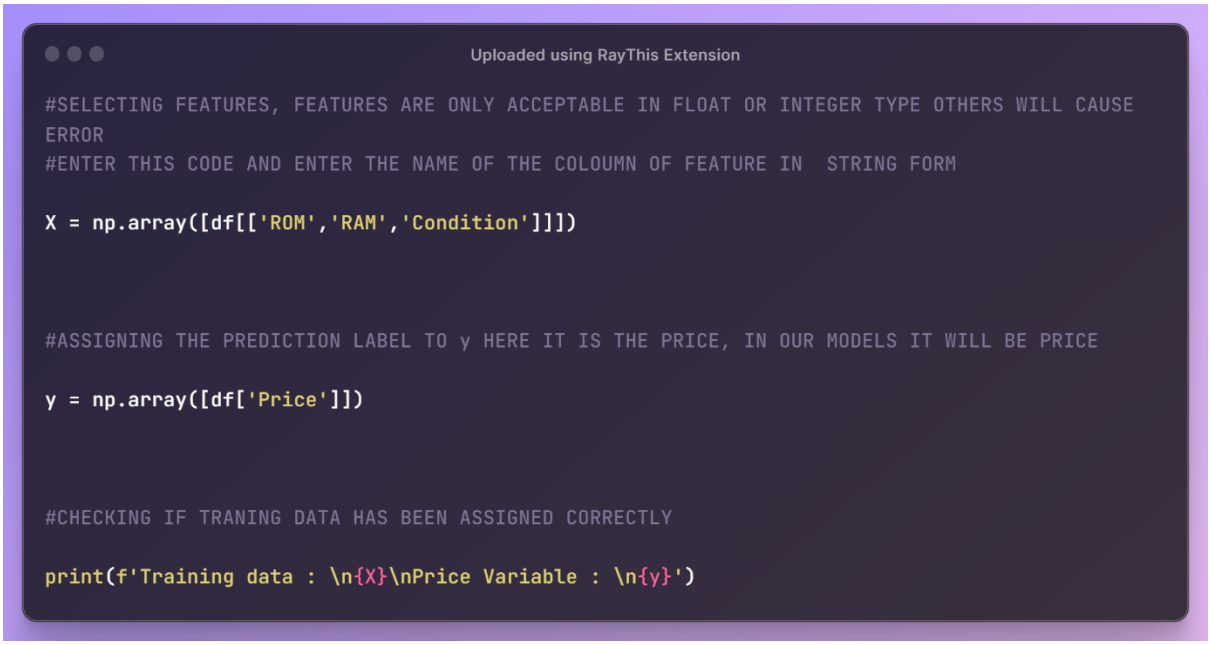
Or you can use cd commands to navigate to the folder where your file is at or you can also try adding the path where the file is at before file name.

To check that the file is loaded we print the variable we assigned the file to, if it dosent cause any error and prints the whole file it means you have loaded the file correctly

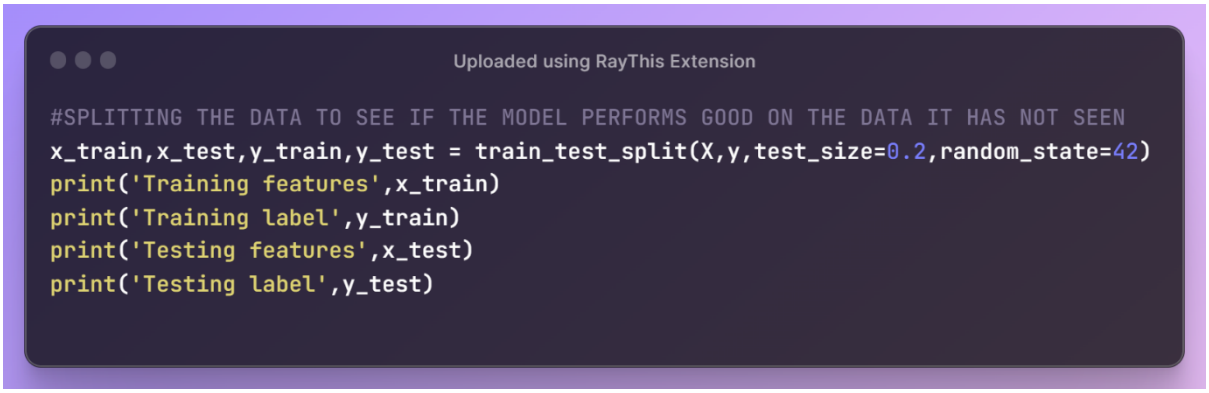


Here we have used a technique called feature engineering/feature scaling. It is used to scale the features in the same range for example if RAM is 8gb and ROM is 128gb, we divide the ROM by 100 to make it 1.28 to make it in the same range as 0 to 10.

But in the code I have divided it by 1000 to make it in range 0 to 1 it is not necessary to divide all the features(RAM,ROM etc) by a same number we just have to keep them in the same range, we could divide the ‘Condition’ by 10 but to make it simple I divided all the features by 1000.



Now we use numpy to load the data into arrays, all the training features(On which the model is trained on) will go into a single variable but for output value(price) there should be a separate variable.



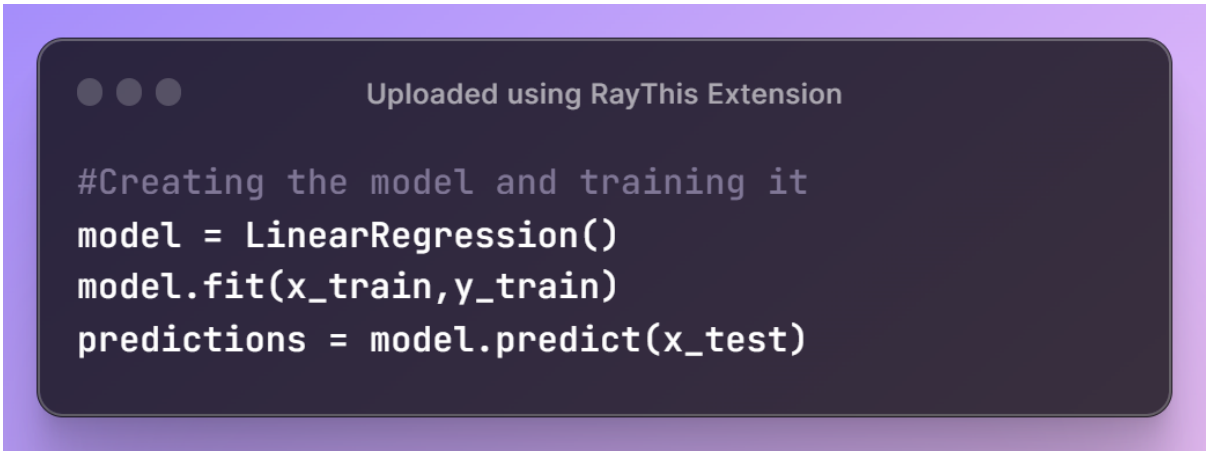
Now we use train test split function to split the dataset into training and test data,

X is the training features, y is the output variable,

test size means that how much of the data should be used as test set here 0.2 means 20% data is used as test set,

random state 42 means that the when we run the code multiple times the training and test data should be the same for example out of 5 total examples example 1 would be test set and example 2,3,4,5 would be training and when the second time the code is run it will select any one of the training set randomly(2,3,4,5) and except the test example others and the example 1 would be selected as training set.

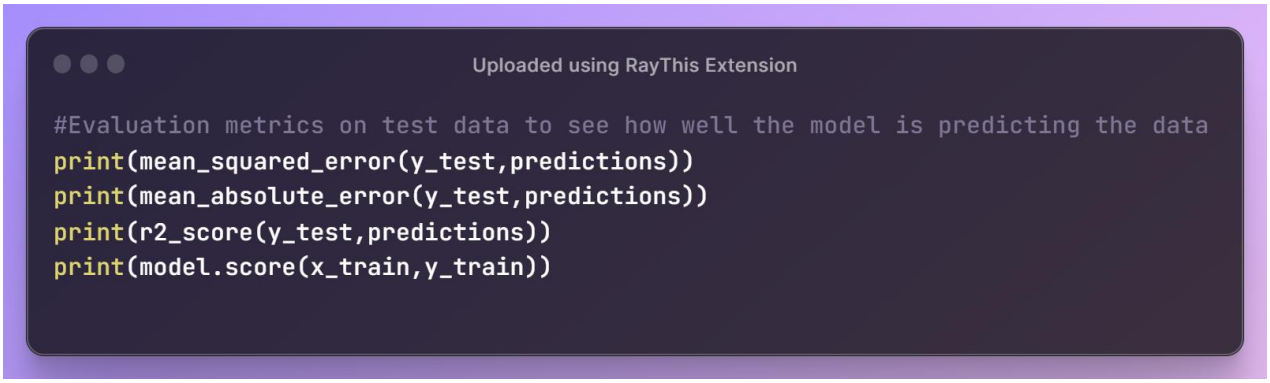
42 is a random number we can use any number we like except 0



Now here we create the main model named LinearRegression we assign it to a variable named model,

Then we fit the training data on the model, we don’t use the test data because it is only for checking if the model performs well.

Then we make predictions on the model using test data, you can print the prediction variable to check what predictions the model is making



Finally these are some error metrics, this will check the difference between the actual price and what the model predicted

I will send a video for each one of them.

The only important part in the code is reading the data if you have read the data correctly then there should be no problem.

The model scores will be very low, so I will improve the scores of the model using various techniques and I will explain them to you too.