COMPUTING SUBJECT:	Machine Learning
TYPE:	WORK ASSIGNMENT
IDENTIFICATION:	Importing and working with datasets
COPYRIGHT:	*Michael Claudius and Mahdi Petersen*
DEGREE OF DIFFICULTY:	Easy
TIME CONSUMPTION:	< 45 minutes
EXTENT:	< 50 lines
OBJECTIVE:	Get an understanding of data migration
COMMANDS:	

IDENTIFICATION: Importing Datasets

The Mission
Work with datasets from external sources so the data can be read, printed and/or plotted to screen.
Remark
Working with data and understanding its' contents is an essential factor in machine learning.

The problem
When working with large datasets, you will need to know how to read them to a Panda's DataFrame so you can prepare and model the data. The datasets have be downloaded and you have to ensure that your data can be located from its path.

Assignment 1: Reading csv-file to a Pandas DataFrame in Jupyter Notebook
Data comes from a source, and in this example, you are to work with data stored in a csv-file. Your assignment is to collect a set of data from the internet, save it in your project directory or in a preferred folder, and read it to a Pandas DataFrame.

1. Create a new ipynb-file and name it "housingtest".

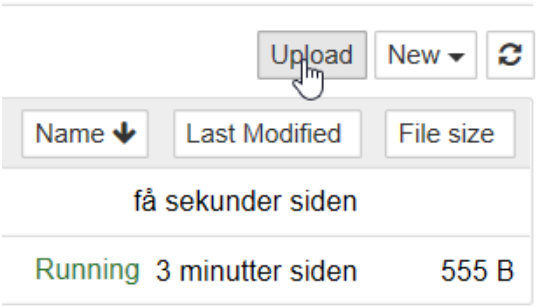
2. Navigate to following GitHub - https://github.com/ageron/handson-ml2 , clone or download the repository to your PC. When done, you will find a folder named "datasets". The folder contains different datasets stored as CSV-files. First, we will be working with *housing*.csv (house prices) and later on in other assignments *oecd_bli_2015.csv* (life satisfaction).
Note: I downloaded to local C:\ drive and then copied the whole Github-project to my *OneDrive Zealand*, as Anaconda has directly access to the *OneDrive Zealand* from the project repository.

You now have two options:

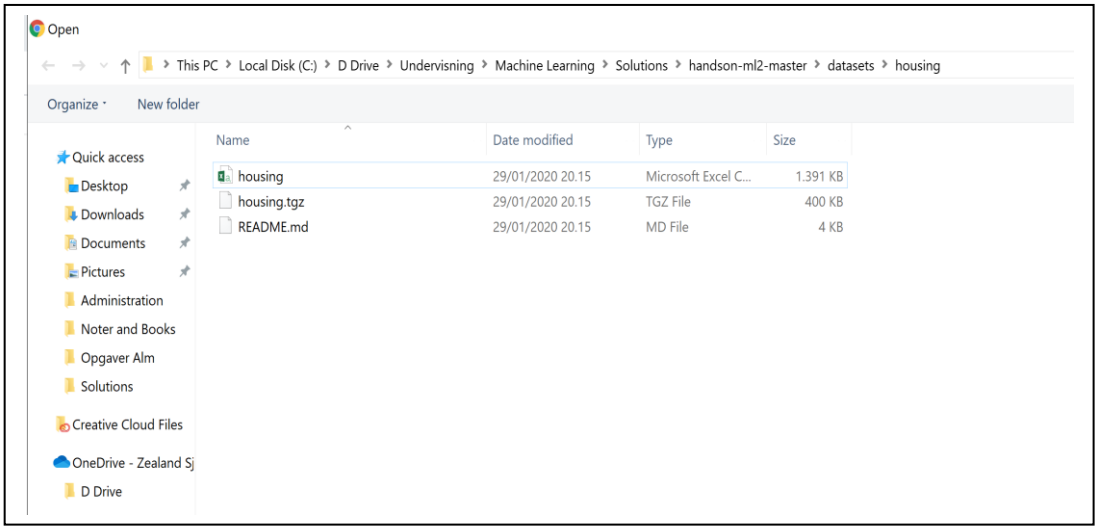
A. *Copy the housing.csv directly to your Jupyter project folder, or*
B. *Access it from the folder, holding the cloned/downloaded GitHub repository.*

***3. First Option A*:** Upload the housing.csv to your solutions-folder:

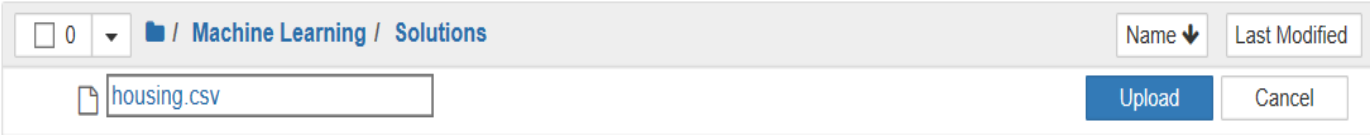
Click "Upload" on the right side of the screen.



Your file-explorer will now appear. Navigate to your repository. Here you will find a folder named "datasets". The folder contains different datasets stored as CSV-files.



Choose the housing file, it will now appear in your Jupyter folder.
Then you must click *Upload*.



4. In Jupyter, first create a new Notebook.
In first cell import pandas (pd) as done previously. Then declare a variable to hold the data by using the pandas' pd.read_csv function.

If you saved the CSV-file in your project folder:

```
housing = pd.read_csv('housing.csv')
```

The CSV-file is now ready to be stored in a data frame. Simply create a DataFrame object as you did in the previous assignment and let it take your data variable as argument.

```
df = pd.DataFrame(housing)
```

The DataFrame you just created contains a lot of data. In order to see just a small fraction call the head() function. It will show you the top five results by default, but you can give an optional number as argument.

```
print(df.head())
```

Your output should look like this.

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	\
0	-122.23	37.88	41.0	880.0	129.0	
1	-122.22	37.86	21.0	7099.0	1106.0	
2	-122.24	37.85	52.0	1467.0	190.0	
3	-122.25	37.85	52.0	1274.0	235.0	
4	-122.25	37.85	52.0	1627.0	280.0	

	population	households	median_income	median_house_value	ocean_proximity
0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	496.0	177.0	7.2574	352100.0	NEAR BAY
3	558.0	219.0	5.6431	341300.0	NEAR BAY
4	565.0	259.0	3.8462	342200.0	NEAR BAY

5. Second Option B: If you saved the CSV-file in another folder or directory then You have to specify the path to the .csv file in the folder datasets folder

```
housing1 = pd.read_csv(r'C:\Users\EASJ\OneDrive……\handson-ml2-master\datasets\housing\housing.csv')
```

Remember to add the file name and extension, as the last part of the path.
Then in the cell type:

```
df1 = pd.DataFrame(housing1)
print(df1)
```

Try to run, hopefully you get the same output as before.

UPS: Did you forget the r'C:\ ??