COMPUTING SUBJECT:	Machine Learning
TYPE:	WORK ASSIGNMENT
IDENTIFICATION:	Importing and working with datasets
COPYRIGHT:	*Michael Claudius*
DEGREE OF DIFFICULTY:	Medium
TIME CONSUMPTION:	**< 2 hours**
EXTENT:	< 50 cells
OBJECTIVE:	ML Project analysis of data
COMMANDS:	

IDENTIFICATION: Housing 1

The Mission

Work and understanding datasets from external sources and then analysed by the developer.
Remark
Working with data and understanding their contentsis an essential factor in machine learning.

The problem

When analyzing large datasets, you will need to use various in-built libraries for plotting and calculation. You need to have a dataset and to create a Notebook-project.
You have already in a previous exercise downloaded the dataset, "*housing*.csv" (house values) from the following GitHub - https://github.com/ageron/handson-ml2, to your PC.
Now it is time to fetch a complete program.

Assignment 1: Reading a program into the Jupyter Notebook

The program is in the repository you made when you downloaded from GitHub.

1. Navigate to the folder "*handson-ml2-master"* holding the downloaded GitHub repository on your PC. Notice the subfolder "*02_end_to_end_machine_learning_project".*
This project we need to access:

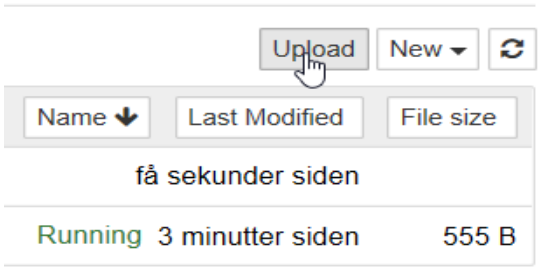
You now have two options:

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

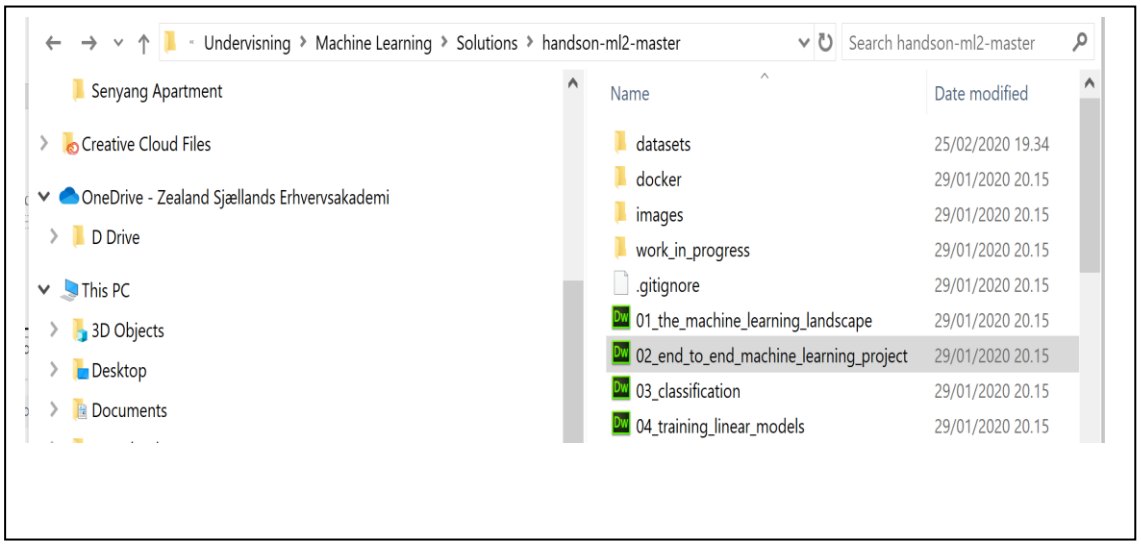
As you previously have done the same for the dataset, I will only describe option A.

***2. Option A*:** Upload the "*02_end_to_end_machine_learning_project* to your solutions-folder:

Start Jupyter, navigate to *Machine Learning/ Solutions* and click "Upload" on the right side of the screen.



Your file-explorer will now appear. Navigate to your repository. Here you will find the folder named "*handson-ml2-master*. The folder contains different projects stored as ipynb-files.



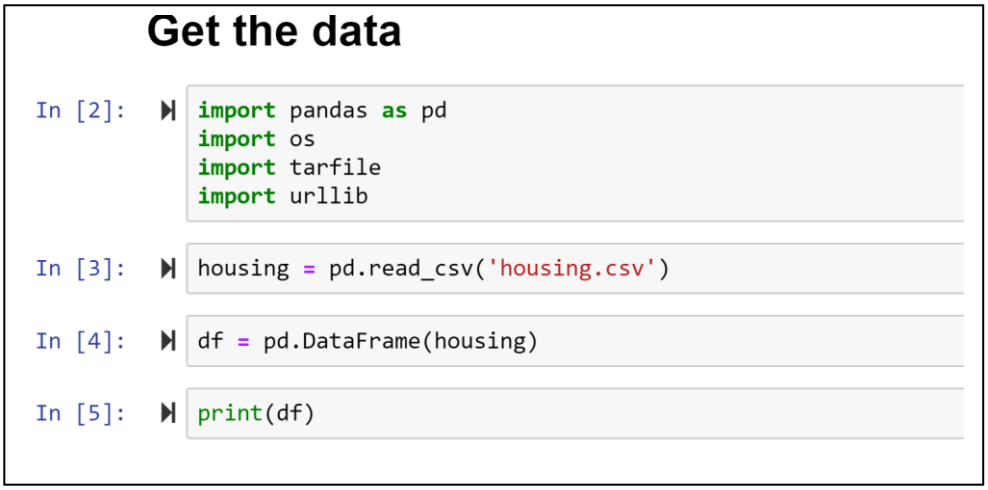
Choose the "*02_end_to_end_machine_learning_project* file, it will now appear in your Jupyter folder. Then you must click *Upload*.

3. In Jupyter, first create a copy of the project and save the copy as Housing2A.ipynb.

Now we can adjust the program code without destroying the original file

Assignment 2: Application program: Adjusting Jupyter Notebook program

1. First cut away all cells below the headline "Prepare data" (from cell number 48).
Secondly change the start cells to use your local housing data, *"housing.csv"*, like you did before in a previous exercise.
Thus the original cell [2] to cell [5] should be changed to resemble the following cells:



2. The DataFrame you just created contains many 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 (try e.g. 10) as argument.

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

Now we can execute the code without destroying the original file.

Assignment 3: Exploring the data and linear regression

1. Run the cells one by one, and make sure all group members understand the principles not necessary all details. Make notes on the fly in Google docs.
On the way discuss the topics and write down the answers to the following questions:

a. What is a feature?
b. State some of the features in the data set.
c. What is the difference between a label and a feature?
d. What is the outcome of the functions *head*, *info* and *describe*.
e. Why is it smart to use the function *train_split_test* and **not** *split_train_test* ?
f. Stratified vs. random. What is the advantage of a stratified test set?
g. Why is correlation matrix interesting?
h. State the 3 features with highest correlation to median-house-value.
i. Combination of features can lead to higher correlation.
j. How is it done in the housing-project?
k. What can we do with features with low correlations (less 5%)?