COMPUTING SUBJECT:	Machine Learning
TYPE:	WORK ASSIGNMENT
IDENTIFICATION:	Importing and working with datasets
COPYRIGHT:	*Michael Claudius*
DEGREE OF DIFFICULTY:	Medium
TIME CONSUMPTION:	**< 1 hour**
EXTENT:	< 50 cells
OBJECTIVE:	ML Project analysis of data
COMMANDS:	

IDENTIFICATION: Housing 2

The Mission

Work and understanding datasets from external sources and then analysed by the developer.
Remark
Working with data and understanding their contents is 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 have already in previous exercises downloaded the dataset, "*housing*.csv" (house values) from the following GitHub - https://github.com/ageron/handson-ml2, to your PC.
Also you have downloaded the complete program "*02_end_to_end_machine_learning_project"* and created your own project.

Now it is time to run the last part of the program with preparing data and training models.

Assignment 1: Application program: Adjusting Jupyter Notebook program

From the original program, copy all cells below the headline "Prepare data" (from cell number 48) to the end of your own project.

Assignment 2: Preparing the data and training models

Run the new cells one by one, and make sure all group members understand the basic 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 to do with missing feature values?

b. How is this done in housing project?
Give a code example.

c. What is an Imputer?
Show code how to use *SimpleImputer*

d. How to handle text and categorical attributes?

e. Explain the idea behind *OneHotEncoder.*

f. What is the idea behind min-max scaling and standard scaling?

g. Why is scaling needed needed?

h. If you use *DecisionTreeRegression* you get an error rmse = 0.0.
Happy? Is there a better evaluation?

i. What is fine tuning?

   What is the meaning of the hyper parameters:
   n_estimators:
   bootstrapping:
   in RandomForest model.

j. Features have different importance in the model.
   Which features have low importance in housing project?
   What to do with them?

k. What is the idea behind k-fold cross validation?

l. What is a grid search ?