# Superviced machine learning

* In this section we will use regression and machine learning to analysis the data set and create a model on the variables in order to determine its prediction abilities.
* In this part we will give a short introduction to machine learning. Machine learning is teaching the machine to predict the outcome, based on a model and parameters. Machine learning can be both supervised and unsupervised. The difference is whether or not we provide the results to the training or not. In this analysis, we will try to get the machine to predict the prices of apartments in Copenhagen, based on the data set. Since the prices are already in the data set, we are able to test the prediction against real data. This is therefor a supervised machine learning.
* We will use both multiple and simple regression to make our prediction in order to examine the data set. We will be looking at two factors. First we will look at which parameters are significant, and secondly we will try to determine how significant the each parameter is. This will be done through regressive machine learning.
* In order to teach the machine, we need to give the machine data to train on. But since we also want to test the training, we also need a data set to test it on. If we gave the machine the same data set to train and test on, we would create a model, that predicted 100 %, since the machine would already know the outcome. We therefor need to split the data set and make sure that the data in the training set was not the data in the test. Otherwise we would be cheating.
* How to split the data set depends on the size of the data set. We want the machine to have as much data as possible to train on, so that it learns as much as possible, but we also want some data to test on. And the test data needs to be big enough so that the result becomes valid, remember Law of Large Numbers.   
  The bigger the data set, the smaller percentage of the data set is necessary for testing. It is therefor a trade of between how precise the model becomes and how sure we can be of the result.

# Regression analysis

* On the cleaned data we get a bell-like shape of the distribution of prices across the four areas in Copenhagen. This means that we can assume normal-distribution in order to make our regression analysis.
* In this data analysis we use simple linear regression on the data set, with prices being our dependent variable, and all other variables being independent. This will be the setting of the variables throughout the regression section.
* Furthermore since location is a string, this variable has been left out. Also price per square meter is per construction not a valid parameter to run as an independent variable on price. Therefor these three parameters have been removed from the dataframe before we do any regression. We continue the regression analysis of the data set using the six variables, 'm2', 'rooms', 'toilets', 'floor', 'age', 'locat\_dummy'.

# Part 1, Visual inspection

* Using six variables in reg.plot
* In this section we make a visual analysis in of the six selected variables. We make a simple scatterplot of each parameter on the first axis and price/price per square meters on the secondary axis. Since price per square meter includes square meters in construction, this visualization is not made.   
  The plots made with prices are given a red color, and plots with prices per square meters on the Y-axis is blue.
* If we look at the scatterplot of square meters, we se a clear trend. More square meters, leads to higher prices. Since we removed all data above 4.000.000 kr. It is unclear whether or not the plots on the fare right are outliers or indicators of decreasing marginal effects on square meters.
* For number of rooms in the first plot, we see a positive correlation. This means that more rooms in general means higher price. Often more rooms also indicate more space, which therefor is similar to the plot for square meters.  
  If we then look at the blue plot for rooms, which shows the correlation between number of rooms and the price per square meter, we see that apartments with two and three has the highest variance. Furthermore we see that the apartments with the highest prices per square meter is the 1 to 3 rooms apartments. The downward trend that we observe on the most expensive square meters underlines what we observed in the heatmap above, but as we also see, the cheapest square meter is located in the three-room category. Furthermore the trend of the least expensive square meters is horizontal or slightly upward going. Therefor any definitive correlation is difficult to conclude here.
* With regard to number of toilets, the outcome of the plots are very similar to the one with rooms. More toilets indicate bigger apartments, which again indicates higher prices. And the most expensive square meters are in the apartments with only one. Additionally only two observations are made with three toilets, which makes us question if we can do any analysis on this variable.
* Looking at floors, we see a different picture that when we looked at toilets and rooms. Here we see no real trend in which floor an apartment is located on and the total price. Especially if we do not include the two observations located on the seventh floor.   
  But if we look at the blue plot with square meter prices, we see an upward going trend for the highest priced square meters. This indicates that people in Copenhagen wish to live as high above the ground as possible. This makes sense if we think that apartments located closest to the ground get more noise from the street, gets less light from the sun and experience more break-ins.  
  Also for the apartments with the lowest price per square meter, do we see an upward going trend for the apartments from the groundfloor and up the next two floors. Then the trend goes down for the next two floors, only to again go up the last three floors. A qualified guess to why this is, is that the lest expensive square meters are located in buildings that do not have elevators. If we assume that the overall interest is to live as high as possible, then we get to the third floor before the challenges with walking up the stairs overcome the interest to live as high as possible. In building with six, seven and eight floors, elevators are often implemented, why we again here see an upward going trend.
* Age of the building does not seem to have any visual effect on the price or the price per square meter.
* Looking at location, we get the indication that people is willing to pay more for an apartment in Dummy 1 and 3, which are Copenhagen central (KBH K) and Copenhagen West (KBH V). However Copenhagen North (KBH N) looks slightly cheaper than the others and Copenhagen East (KBH OE) has the highest spread.

# Part 2, OLS on M2 and Age

* In this section, we will be running multiple and simple regressions on the data set to determine if any of the parameters are significant for the price of apartments in Copenhagen. We do this with a basic OLS regression.
* First of we remove location dummy and floor from the data set. Even though both variables seemed to have influence on prices in part 1, we cannot run regression on category variables, why the simplest solution is to remove them all together.
* Running the multiple regression on the data set of now four variables, we see that all four variables are significant and positive correlated with prices.   
  In section 1 we discussed how rooms and toilets are highly dependent on square meters, we can take out rooms and toilets and run the regression again. This lowers the variance and increase the significance of the model.
* We then run simple regression on each of the two parameters and see that both also are significant.

# Part 3, test/train model prediction

* In this section, we will use our experience from part 2 to see how well our model can predict prices based on the two selected parameters, square meters and age.
* Since we already have the prices, we can use the theory of supervised machine learning to estimate a regression model that can predict prices of apartments in Copenhagen.
* First of we split the data set into a training set and a test set. We chose 75/25 split due to the size of the data set. We have chosen to evaluate the success of the prediction, by using an R-squared score test and a Root mean squared error test.
* We find that running this regression, with square meters and age as variables, that we get a model where R-squared reaches almost 0,65. Even though we cannot translate that directly into a percentage score of how much it predicts correctly, we can say that our model has some prediction strength.   
  We then run the model on both parameters separately and find that square meters have almost all the prediction power, 0,64, and age have very little, 0,01.
* If we look at this result, it makes sense that people will pay more for more space, and the age of the building has very little influence. But is a little surprising, that older buildings cost more, than younger building. This however could be due to other reasons, such as location or size.