Working with data - Basics

by Ahmet Sacan.

```
url='http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-ca

% Make sure the folder containing bmes files has been added to your path.
file=bmes.downloadurl(url);
```

```
--- NOTICE: Attempting to download & save url [ http://archive.ics.uci.edu/ml/machine-learning-databases/breast-canc
```

```
t = readtable(file,'filetype','text');
```

Setting up the attribute names

if the datafile contained the attribute names (typically in the first row), readtable() would have detected that and set up the attribute names for us. But for this dataset, the file does not contain the attribute names so we have to give readtable a little more help.

```
% I manually extracted the attribute names from: http://archive.ics.uci.edu/ml/machine-learning
attributes={'id','thickness','cellsizeuniform','cellshapeuniform','adhesion','cellsize','barenu

t.Properties.VariableNames=attributes;
```

Using the table object

```
t(1:5,1:4)
```

```
ans = 5×4 table
```

	id	thickness	cellsizeuniform	cellshapeuniform
1	1000025	5	1	1
2	1002945	5	4	4
3	1015425	3	1	1
4	1016277	6	8	8
5	1017023	4	1	1

```
t{1:5,1:4}
```

```
ans = 5×4
     1000025           5           1           1
     1002945           5           4           4
     1015425           3           1           1
     1016277           6           8           8
     1017023           4           1           1
```

```
t(1:5,'thickness')
```

```
ans = 5×1 table
```

	thickness
1	5
2	5
3	3
4	6
5	4

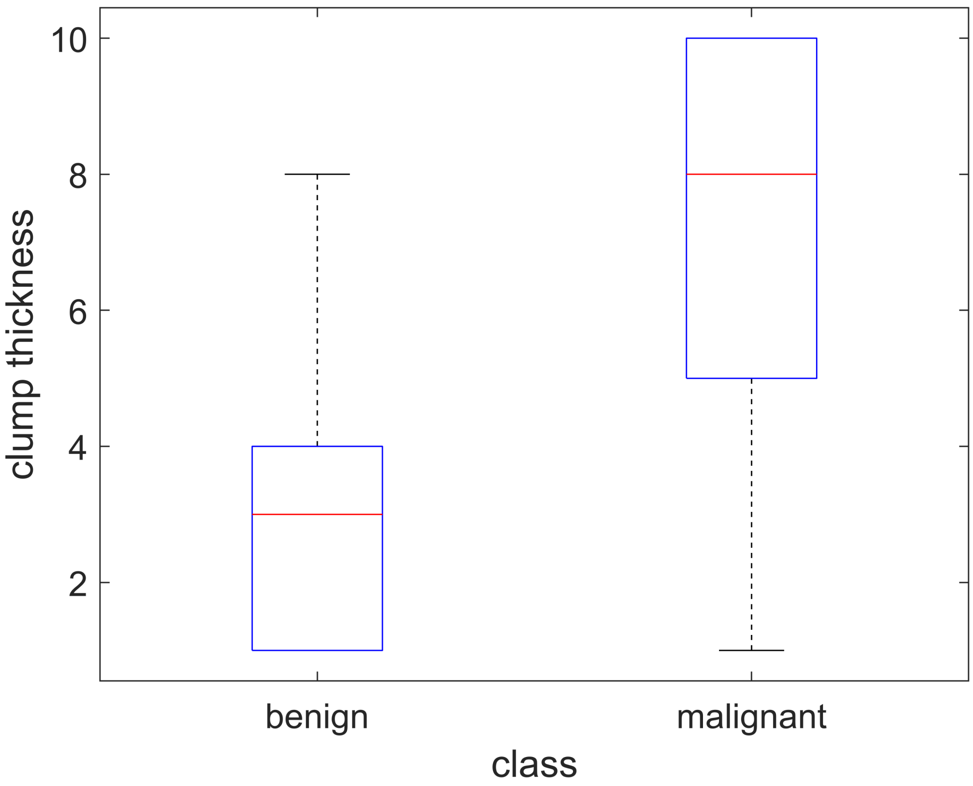
```
t(1:5,{'thickness','class'})
```

```
ans = 5×2 table
```

	thickness	class
1	5	2
2	5	2
3	3	2
4	6	2
5	4	2

boxplot visualization for groups

```
% we have manually determined that class=2 is benign and class=4 is malignant.
boxplot(t{:, 'thickness'}, t{:,'class'} , 'Labels',{'benign','malignant'});
xlabel('class'); ylabel('clump thickness');
```



Calculating correlation

```
[correlation,pvalcorr] = corr(t{:,'thickness'},t{:,'cellsize'})
```

```
correlation = 0.5218
pvalcorr = 4.4982e-50
```

Calculating significance of difference

Is there significant difference in the clump thickness, between benign & malignant samples?

```
Ibenign = t{:,'class'} == 2;
Imalignant  = t{:,'class'} == 4;

[~,pvalue]=ttest2(t{Ibenign, 'thickness'}, t{Imalignant, 'thickness'})
```

```
pvalue = 6.8356e-111
```