Data Mining- Assignment 1	
Due	---
Grading	This assignment will be graded from 0 to 10.
Grading Measures	• The quality of your data mining strategy and results • The argumentation, validity, and clarity of your report
How to deliver:	• Group of 1 student are allowed. • Write down your report to this assignment in a .pdf file with the following name "<your student number><your name>.pdf", e.g., "012345JanJansen.pdf". ⊙ upload your report and excel file

HOW TO DO IT:

0. First, start becoming familiar with the basic statistics tools in Excel, by playing with the example provided in "DataAnalysisSession1.xls", in which the data set "DataSession1.txt" has been analyzed.
1. Work out Assignment 1 remembering what you have learnt in the class and practiced at point 0.

Open the Excel file "Seismic data" and save it as "<your student number><your name>.xls". It contains 920 readings from one seismic recording station over a 7 day period. Answer to the following questions:

1.1: Compute:

- the number of elements in the data set
- the median
- the minimum and maximum values
- the range of the data set
- the first and third quartile, and the IQR
- the 95% confidence interval

1.2: Following the instructions provided in "How to make a scatter plot with Excel.pdf", create a scatter plot of the data set, and attach a copy here below:

SCATTER PLOT!

1.3 Considering the statistics computed at the previous point, together with the scatter plot of the data set, can you tell anything about the distribution of the data set? Does the data set seem to be organized following a Normal distribution? Do outliers are present in the data set? Please, provide your answer in the blank space below, without overstepping it!

1.4: Bearing in mind that outliers in a data set are identified as the values which lie outside of the whiskers in a box-whiskers plot, and that the whiskers extend to 1.5 times the IQR below the first or above the third quartile: try to identify possible outliers in the data set. To do that, adapt the following Excel command to your file:

```
=IF(OR(A1>$B$1,A1<$B$2),"Outlier","Not Outlier")
```

1.5: Once the outliers have been identified, copy the original data set and delete the outliers from it. Compute the mean, median, and standard deviation of the new data set. How the median has changed?
Was this data set affected considerably by the presence of the outliers or not? Why? Please, provide an answer in the blank space below, without overstepping it!

1.6: Create a box-whiskers plot of the new data set, computing the whiskers by means of the standard rule: 1.5 times the IQR below the first or above the third quartile. Provide a copy of it here below:

BOX-WHISKERS PLOT!

1.7: Create a histogram of the new data set. Provide a copy of it here below:

HISTOGRAM!

2. In 22 patients with an unusual liver disease the plasma alkaline phosphatase was found by a certain laboratory to have a mean value of 39 King-Armstrong units, standard deviation 3.4 units. What is the 95% confidence interval within which the mean of the population of such cases whose specimens come to the same laboratory may be expected to lie?