**DATA 301 Lab Instructions**

**Marking and Evaluation (15 marks)**[**¶**](https://firas.moosvi.com/courses/data301/notes/week03/lab1.html#marking-and-evaluation-15-marks)

This lab is to be done on an individual basis. Marks are awarded by ***precisely*** following these requirements:

* **Loading (1 mark) -** Create an Excel file called **lab1\_(yourstudentnum).xlsx** (e.g. **lab1\_11111111.xlsx**) where the first sheet is called *rawdata* and contains the loaded data set with no changes.
* **Cleaning and Formatting (5 marks) -** create a second sheet called *data* that contains the data set after all incorrect and out-of-range values have been removed.
  + Any data that is not a number, is less than 0, or above 100 should be replaced by an **empty string (“”)** (1 mark). Hint: May need to use [IF](https://support.office.com/en-us/article/IF-function-69aed7c9-4e8a-4755-a9bc-aa8bbff73be2) function.
  + Header fields must be in bold font (0.5 marks).
  + Timestamp field must be formatted as a date field in this format yyyy/mm/dd hh:mm:ss (2 marks). Hint: May need to use [DATEVALUE](https://support.office.com/en-us/article/DATEVALUE-function-df8b07d4-7761-4a93-bc33-b7471bbff252) function and function like [MID](https://support.office.com/en-us/article/MID-MIDB-functions-d5f9e25c-d7d6-472e-b568-4ecb12433028) to convert date. [More info](http://chandoo.org/wp/2010/03/23/text-to-date-convertion/). Note that both the DATE and TIME need to be converted and added together to get a DATETIME.
  + Data must be sorted by siteid (ascending), sensorid (ascending), then timestamp (ascending) (0.5 marks).
  + Add a column called global\_sensor\_id that is generated by siteid&”\_”&sensorid (1 mark).
* **Summary (3 marks)** - create a third sheet called *summary* that contains the following items:
  + Count of readings in B3, maximum reading value in B4, minimum reading value in B5, average reading in B6 (1 mark).
  + Maximum reading by any sensor at site 2 in B8 (1 mark). Hint: Ctrl+Shift+Enter will be useful. See: [Array Output using Ctrl+Shift+Enter](http://superuser.com/questions/674566/when-to-use-ctrlshiftenter-and-when-to-use-enter-in-excel)
  + Number of readings for sensor 2 at site 1 in B10 (1 mark). Note: Include **all** readings including readings that were out of range/errors/blank.
* **Transformation (3 marks) -** create a fourth sheet called *transform* that contains a table where each row is a time and each column contains values for the sensor.
  + Creating data table (1 mark).
  + Replacing missing values with NA() (0.5 marks). [More info](http://www.criticaltosuccess.com/when-data-points-are-bad-or-missing-how-do-you-create-a-usable-and-decent-looking-excel-chart/) and [Using ISNA() function](http://www.extendoffice.com/documents/excel/2487-excel-conditional-formatting-if-na.html#a1)
  + Adding sparklines in last column for each row (0.5 marks).
  + Add conditional formatting so that cells with #N/A are filled in red and cells with values >= 90 are filled in green (1 mark).
* **Visualization (2 marks) -** create a **XY Scatter chart (with smooth lines and markers)** on a sheet called *chart* that shows the readings by time for all three sensors at site 1. Add an appropriate trendline for sensor 1 at site 1. Also, using all the data from site 3, make a histogram to show the overall distribution of the data(you can use bin size of 10) and use a box plot to show the distribution of data from site 2 for each sensor separately.
* **Analysis (1 marks) -** in the *summary* sheet put in cell B12 if sensor 1 at site 2 has any data trend (none, exponential, linear, etc.).