# 12-752: Data-Driven Energy Management of Buildings Assignment#1

Mario Bergés
Assistant Professor
Department of Civil and Environmental Engineering
Carnegie Mellon University
Pittsburgh, PA 15213
marioberges@cmu.edu

November 6, 2015

Some notes before you begin:

- Make sure you document everything you do, and not just write down the answer to the question. This will both help during grading as well as improving your learning process.
- Do not write down any solution or process that you do not understand. If you feel that you do not understand how to do something, seek some help: e-mail the TAs or the instructor
- To submit your assignment, please do so using Blackboard. Two files should be uploaded via Blackboard: (a) the IPython Notebook (i.e. a .ipynb file) documenting all the tasks found in the assignment and all of your answers (including the output of your code); and (b) a PDF copy of this notebook
- Please upload a single compressed ZIP file containing the above, and name it as follows: *andrewID_assignment_#.zip* (where *andrewID* is your AndrewID and # is the assignment number

## 1 Loading the Dataset

For this assignment, we will be exploring a specific dataset that we will be discussing in class: the EIA's Residential Energy Consumption Survey (RECS) from 2009. Specifically, we will explore the "microdata" which contains the survey responses for the 12,000+ participants in the study.

You can obtain a copy of the dataset here:

- recs2009_public.csv
- public_layout.csv

The first file is the survey data itself, while the second one contains descriptions for each of the survey questions in the same order as they appear in the first file. Download both files into the folder where you will be running the Python code from.

The first part of your Python program should basically load the CSV file into memory.

**Task #1 [5%]:** Load the CSV file into memory.

Then, you should make sure that you have done this correctly:

**Task #2 [5%]:** Print out the first row of the file on the screen.

If all is working well, your data should be have exactly 12,083 rows. Check that this is the case:

**Task #3 [5%]:** Print out the size of the variable you are using to store the contents of the CSV file.

Do you know what is the data type for the variable you are using to store this data?

**Task #4 [5%]:** Print out the data type for the variable (e.g., list, array, etc.

## 2 Exploring the Dataset Graphically

In this section we will simply explore the dataset using common Exploratory Data Analysis techniques.

**Task #5 [10%]:** Using scatter plots, find a pair of variables that shows a strong linear relationship (a positive linear correlation). Using what you know about these variables (i.e. the layout file) explain why this relationship makes (or doesn't) make sense.

**Task #6 [10%]:** Create a histogram of the total energy use by buildings with more than 1 window in the heated areas. Compare that with the same histogram but for homes with no windows in heated areas. What does this information tell you, if anything?

**Task #7 [10%]:** Create a box plot for the kWh usage of refrigerators. Using information contained in the survey for this home, can you explain why the highest outlier (more than 10,000 kWh) exists?

**Task #8 [10%]:** Ensure that all of your plots have the axes labeled approrpiately.

## 3 Exploring the Dataset Quantitatively

Now we move on to use more quantitative tools for exploring the dataset.

**Task #9 [20%]:** Calculate the mean ($\mu$), median ($M$) and the standard deviation ($\sigma$) for each variable (i.e. column). Sort this list by the standard deviation, in descending order, and print out the tuples ($\mu$, $M$, $\sigma$).

In other words, you will print a list that looks something like this (don't pay attention to the exact values, though):

(3.5, 4.5, 0.2)
(2.1, 4, 0.15)
(100.5, 254.3, 23.4)
...
Where each line corresponds to a variable and the number in parentheses are the mean, median and standard deviation for that variable.

**Task #10 [20%]:** Using a similar process to the one you used to load the "microdata" CSV file into memory, load the "layout" file into memory. Using this infromation, re-print the list from Task #9 but preface each tuple with the full description of the variable that corresponds to it.

In other words, your list will look something like this (again, don't pay attention to the exact numbers):

Electricity usage for water heating, in kilowatt-hours, 2009: (3.5, 4.5, 0.2)
Number of sliding glass doors in heated areas: (2.1, 4, 0.15)
Total Site Electricity usage, in kilowatt-hours: (100.5, 254.3, 23.4)
...