# Data Analysis Workshop - Activity 1

## Introduction

This activity is designed to be completed using MATLAB. You will be provided with a code snippet which will contain a working example. Your job is to modify this example code in order to complete the same task for a different data set.

To get started:

1. You will need to log on to a virtual machine which has MATLAB on it - click on the green VM icon on the desktop of your computer
2. Double click on the virtual machine called *desktop*
3. Once the virtual machine has loaded, open MATLAB from the windows *Start* menu

The following is a table which lists the dataset file names for each activity and part.

| **Part** | **Type** | **Filename** |
|---|---|---|
| Part 1 | Example | `slump_test.csv` |
| Part 1 | Task to Complete | `powergen.csv` |
| Part 2 | Example | `water_treatment.csv` |
| Part 2 | Task to Complete | `airquality_formatted.csv` |

## Part 1

### The Task

This part should take you about 10 or so minutes, and is designed to get you familiar with loading datasets into MATLAB, manipulating them, and creating some graphs.

An example code snippet has been provided for a different data set. The code snippet can be found in the learning materials on Learnline, it is called `activity1_part1_example.m`. This example code relies on a data set called `slump_test.csv`. The first thing you should do is to load `activity1_part1_example.m` into MATLAB and run the script. Your job will be to create a similar script which modifies the example code to accept and run with a different data set.

## Overview of the Data Set

The dataset you will use for this task is called `powergen.csv`. It contains 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the power plant was set to work with full load. Features consist of hourly average ambient variables Temperature (T), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V) to predict the net hourly electrical energy output (EP) of the plant.

A combined cycle power plant (CCPP) is composed of gas turbines, steam turbines, and heat recovery steam generators. In a CCPP, the electricity is generated by gas and steam turbines, which are combined in one cycle, and is transferred from one turbine to another. While the Vacuum is collected from and has effect on the Steam Turbine, the other three of the ambient variables effect the GT performance.

If you would like to know more about this data set, please visit the source:

`http://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant`

The data is provided in .csv formats. **There are example code snippets for a similar exercise that you can use to help you with the MATLAB syntax.**

## Instructions to Complete Part 1

1. Start by loading the `powergen.csv` data file into MATLAB. This file is located in the course materials on Learnline. To do this you will need to modify the following code from the example.

   ```
   % Load the .csv file into MATLAB and store it in a variable called data.
   % The ,1,0 offsets the top row because this includes non-numerical data
   data = csvread('slump_test.csv',1,0);
   ```

   How many observations are there? How many variables are there?

2. You now need to create 2 vectors. The first is `T` to hold ambient temperature data, and the second is `EP` to hold the data for energy generated. This requires you to modify the following code from the example:

   ```
   % Store each of the variables into a variable for easier use
   cement = data(:,1); % the 1 relates to column 1
   strength = data(:,11); % the 11 relates to column 11
   ```

   Note that you will have to determine which column your variables on interest are in (**hint: check the data over view**)

3. Create a histogram of the temperature of the variable to get an understanding of the underlying distribution. to do this you will have to modify the following code snippet:

```
% Create a histogram of the of the US Fuel Efficieny variable
figure(1) % Sets up a figure window
histogram(strength)
xlabel('Strength (MPa)')
ylabel('Frequency')
title('Distribution of Strength')
```

   Make sure you are including appropriate labels, including units. How would you describe the distribution of this variable?

4. Create a box and whisker plot of the electrical energy produced. You need to modify the following code snippet:

```
% Create a boxplot of the temperature
figure(2) % Sets up a figure window
boxplot(strength)
ylabel('Strength (MPa)')
title('Distribution of Strength')
```

   How would you describe the distribution of this variable?

5. The net hourly electrical energy generated should be correlated with the ambient temperature - can you think why? Create a scatter plot to help you visualise the relationship. You will need to modify the following code snippet:

```
% Create a scatter plot which shows the relationship between the amount of
% cement and the compressive strength
figure(3) % Sets up a figure window
scatter(cement,strength,'.')
xlabel('Cement (kg per cubic meter)')
ylabel('Strength (MPa)')
title('Cement vs Strength')
```

# Part 2

## The Task

This part of the activity will take about 10 minutes or so, and is designed to help further your understanding of manipulating data in the MATLAB working environment, and to plot some more complicated graphs

An example code snippet has been provided for a different data set. The code snippet can be found in the learning materials on Learnline, it is called `activity1_part2_example.m`. This example code relies on a data set called `water_treatment.csv`. The first thing you should do is to load this into MATLAB and run the script. Your job will be to create a similar script which modifies the example code to accept and run with a different data set.

## Overview of the Data Set

The dataset that you will use for this task is called `airquality_formatted.csv`. It contains **7674** instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level,within an Italian city.

Data were recorded from March 2004 to February 2005. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer.

The variables across the columns are as follows:

| **Column** | **Variable Name** |
|---|---|
| 1 | Year (YYYY) |
| 2 | Month (MM) |
| 3 | True hourly averaged concentration CO in mg/m$^3$(reference analyzer) |
| 4 | PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted) |
| 5 | True hourly averaged overall Non Metanic HydroCarbons concentration in µg/m$^3$ (reference analyzer) |
| 6 | True hourly averaged Benzene concentration in µg/m$^3$ (reference analyzer) |
| 7 | PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted) |
| 8 | True hourly averaged NOx concentration in ppb (reference analyzer) |
| 9 | PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted) |
| 10 | True hourly averaged NO2 concentration in µg/m$^3$ (reference analyzer) |
| 11 | PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted) |
| 12 | PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted) |
| 13 | Temperature in degrees Celsius |
| 14 | Relative Humidity (%) |
| 15 | AH Absolute Humidity |

If you would like to know more about this data set, please visit the source:

`http://archive.ics.uci.edu/ml/datasets/Air+Quality`

The data is provided in .csv format - note that sometimes you will have to format your own data prior to beginning your analysis. **There are code snippets for a similar exercise that you can use to help you with the MATLAB syntax that you need to use and to help you understand what to do.**

## Instructions

1. Start by loading the data file into MATLAB using a variable name `data`. The data set that you are working with is `airquality_formatted.csv`. You need to modify the following code snippet:

   ```
   % Load the .csv file into MATLAB and store it in a variable called data
   data = csvread('water_treatment.csv');
   ```

   Note that the `airquality_formatted.csv` data set has variable names in the first row - you will have to offset the `csvread()` (**hint: you did exactly this in the first part**).

2. We want to create a 2 histograms which compare the CO distributions according to year. To do this you first need to create a logical vector for the two different years present, 2004 and 2005. You will need to modify the following code snippet to do this:

   ```
   filter_1990 = (data(:,1) == 90);
   filter_1991 = (data(:,1) == 91);
   ```

   **ANSWER**

   This is step is tricky - the code needed to do this is:

   ```
   filter_2004 = (data(:,1) == 2004);
   filter_2005 = (data(:,1) == 2005);
   ```

3. Create two subsets of data (one for 2004 and one for 2005) using the logic vectors from above. The code snippet that you need to modify is:

   ```
   data_1990 = data(filter_1990,:);
   data_1991 = data(filter_1991,:);
   ```

   **ANSWER**

   This is step is tricky - the code needed to do this is:

   ```
   data_2004 = data(filter_2004,:);
   data_2005 = data(filter_2005,:);
   ```

4. Create two vectors containing the CO data for 2004 and 2005, called `CO_2004` and `CO_2005`. The column in the data set that you are interested in is the CO concentration. You will need to modify the following code:

```
vol_susolids_1990 = data_1990(:,27);
vol_susolids_1991 = data_1991(:,27);
```

   You will need to determine which column the CO concentration is in before you can do this - check the data overview section.

5. Create a new figure window. In this figure window, you will make 2 histogram plots side by side using the `subplot` command in MATLAB. You need to modify the following code snippet:

```
% A histogram of the Volatile Supended Solids in the water
% for 1990 compared to 1991
figure(1)

subplot(1,2,1) % Start the plot for the 1990
histogram(vol_susolids_1990)
xlabel('Volatile Supended Solids (ppm)')
ylabel('Frequency')
title('Volatile Supended Solids 2004')

subplot(1,2,2) % Start the plot for the 1991
histogram(vol_susolids_1991)
xlabel('Volatile Supended Solids (ppm)')
ylabel('Frequency')
title('Volatile Supended Solids 2005')
```

   You will need to choose appropriate labels and a title for your plots. Even though there are more samples in 2004 than there are in 2005, what might you say (approximately) about the two distributions?

6. Next we would like to visualise how the distribution changes each month using box and whisker plots. The first thing that you need to do is create a categorical array which contains a string label for each observation in the original data set (i.e. `2004/03` to represent an observation in 2004 in March). You will need to modify the following code snippet:

```
% Create a time series box and whisker diagram to show how the
% distribution changes over time

year_month = {}; % Initialises a catagorical array

% Adds all of the string labels to each observation
for i =1:length(data(:,27))
    year_month{i} = strcat(num2str(data(i,1)),'/', num2str(data(i,2)));
end
```

**ANSWER**

This is tricky - one way to implement this is as:

```
% Create a time series box and whisker diagram to show how the
% distribution changes over time

year_month = {}; % Initialises a catagorical array

% Adds all of the string labels to each observation
for i =1:length(data(:,3))
    year_month{i} = strcat(num2str(data(i,1)),'/', ...
        num2str(data(i,2)));
end
```

7. Finally plot your time series box and whisker plot - this is easy to implement - you will need to modify the following code snippet:

```
% Create the boxplot
figure(2);
boxplot(data(:,27),year_month);
set(gca,'FontSize',10,'XTickLabelRotation',90)
xlabel('Time (months)')
ylabel('Volatile Supended Solids (ppm)')
title('Volatile Supended Solids Distribution Change Over Time')
```

8. What can you say about the distribution of CO over time?