Lab Assignment 3: Data Visualization

Due: 02:00 PM Feb 04, Wednesday

The purpose of this lab is to do exploratory data analysis by data visualization. You will learn how to draw histogram, bar plot, scatter plot, line plot etc., with Python `matplotlib` package. Interactive visualization with Tableau software, D3, plotly will come next week.

1 Data visualization as a science

- Watch $0:00-6:00, 9:27-14:00$ of the very insightful and inspiring TED talk from Hans Rosling.
- A great GIF explains the basic principle of data visualization.
- Three popular galleries for data visualization, Tableau, Matplotlib, D3

2 Python visualization packages

2.1 Introduction to related packages

Numpy: Enables exactly a same environment in Python as that in Matlab, e.g, data named as matrix or arrays, fast matrix operations, numerical computation, etc. Two instant notices are, 1). use brackets [] instead of parenthesis () for matrix or array indexing, 2). indexes start from 0 instead of 1.

See what happened with the following code in Spyder:

```
import numpy as np
a = np.array([1,2,3,4])
b = np.zeros((3,5))
c = np.random.rand(4,6)
c[[2,3], 3] = 0
colmean = np.mean(c, axis = 0)
rowmean = np.mean(c, axis=1)
nonzeromean = np.mean(c[nonzero(c[:,3]), 3])
```

Pandas: Allows analyst to read, preprocessing complex table data. It's compatible with R, so the functions provided by this package are similar to those in R, such as the example from line 11 to line 47 in the demo_visualization.py.

2.2 Visualization with Matplotlib package

A mind thinking for visualization from my own experience.

1). What attributes do you have, eg. categorical, continuous, time series, geographical?

2). What story do you want to tell? What kind of plots can help you to tell the story, eg. histogram, word cloud, bar plot, box plot, scatter plot, time series plot?

3). Do you have the data prepared for the plot you want, eg. word frequency of twitter data?

4). Present different information using dot size, color, text, shape, etc, e.g., in the talk, Hans Rosling used colors to present continents.

5). Choose the best way to present, e.g., one picture per figure, or multiple pictures in one figure with subplots, interactive plots or static plots?

6). Add more to your plots, e.g., comfortable color palette, concise and professional lines and dots, understandable legend, labels, annotations.

Histogram, Scatter, Line Plot

- Download demo_visualization.py.
- In Spyder, `import demo_visualization.py`.
- Call demo_visualization.hist_plot().
- Call demo_visualization.scatter_plot1().
- Call demo_visualization.scatter_plot2().
- Call demo_visualization.line_plot().

Detailed comments are given in the script, try to understand them.

3 Lab Assignment [1]

This assignment uses data from the UC Irvine Machine Learning Repository, a popular repository for machine learning datasets. In particular, we will be using the "Individual household electric power consumption Data Set" which I have made available on the course web site:

- **Dataset**: Electric power consumption [20Mb]
- **Description**: Measurements of electric power consumption in one household with a one-minute sampling rate over a period of almost 4 years. Different electrical quantities and some sub-metering values are available.

The following descriptions of the 9 variables in the dataset are taken from the UCI web site:

- Date: Date in format dd/mm/yyyy
- Time: time in format hh:mm:ss

[1] @Copyright Cousera Exploratory Data Analysis - Lab 1

- Global_active_power: household global minute-averaged active power (in kilowatt)
- Global_reactive_power: household global minute-averaged reactive power (in kilowatt)
- Voltage: minute-averaged voltage (in volt)
- Global_intensity: household global minute-averaged current intensity (in ampere)
- Sub_metering_1: energy sub-metering No. 1 (in watt-hour of active energy). It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered).
- Sub_metering_2: energy sub-metering No. 2 (in watt-hour of active energy). It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.
- Sub_metering_3: energy sub-metering No. 3 (in watt-hour of active energy). It corresponds to an electric water-heater and an air-conditioner.

When loading the dataset into Python, please consider the following:

- The dataset has 2,075,259 rows and 9 columns. First calculate a rough estimate of how much memory the dataset will require in memory before reading into Python. Make sure your computer has enough memory (most modern computers should be fine). To increase memory for the VM, by click on 'Edit the Virtual Machine' to increase both the RAM size and disk size.
- We will only be using data from the dates 2007-02-01 and 2007-02-02. One alternative is to read the data from just those dates rather than reading in the entire dataset and subsetting to those dates.
- You may find it useful to convert the Date and Time variables to Date/Time classes in Python using the Pandas package data.Date = pd.to_datetime(data.Date). The benefit of converting the column to datetime type is it will be easy to select a range of datetime, for example, `data[data.Date['2013-1-15':'2013-1-28']]` selects all records from 2013-1-15 to 2013-1-28.
- Note that in this dataset missing values are coded as ?.

Assignment: Our overall goal here is simply to examine how household energy usage varies over a 2-day period (2007-02-01 and 2007-02-02) in February, 2007. Your task is to reconstruct the following plots below, all of which were constructed using the base plotting system.

For each plot you should

- Construct the plot and save it to a PNG file.
- Name each of the plot files as plot1.png, plot2.png, etc.
- Complete four functions in the skeleton Python script Lab_3.py(`plot1()`, `plot2()`, etc.) that construct the corresponding plot, i.e. code in `plot1()` constructs the plot1.png plot. Your code file should include code for reading the data so that the plot can be fully reproduced. You should also include the code that creates the PNG file.

Submissions: Lab_3.py

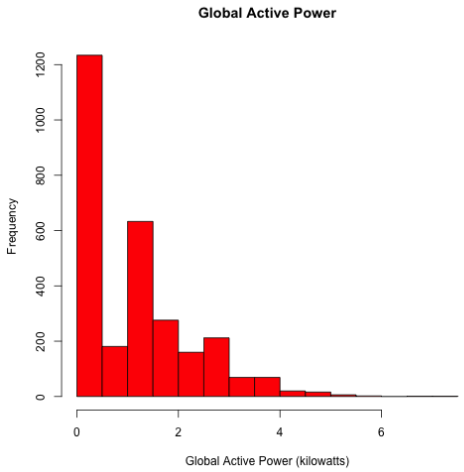

Figure 1: Plot 1

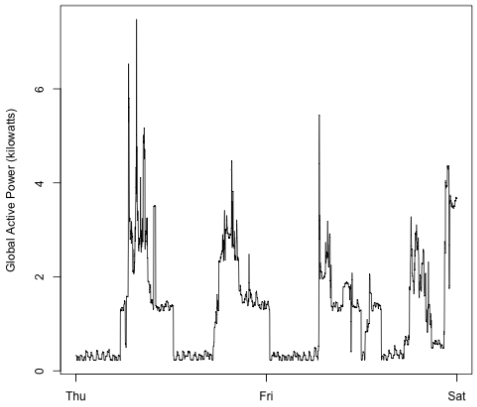

Figure 2: Plot 2

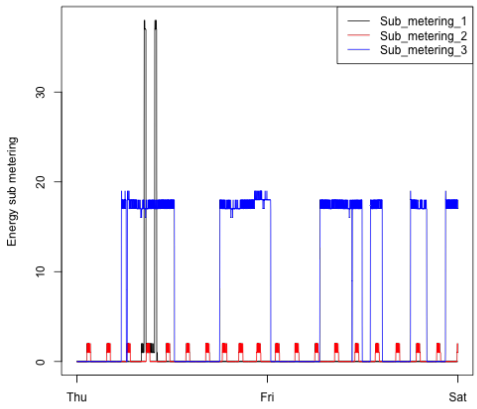

Figure 3: Plot 3

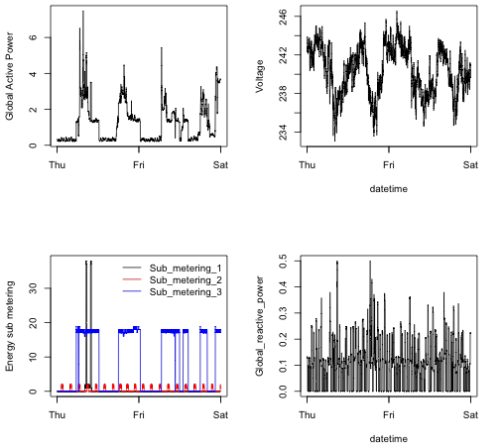

Figure 4: Plot 4

The four plots that you will need to construct.