Lab Assignment 3: Data Visualization Basics

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 packages useful for data analysis

2.1 Data analysis with Numpy and Pandas

Numpy: Numpy is a very popular mathematical package that emulates Matlab in how it treats matrices (such as data tables). Two major differences compared to Matlab are 1) using brackets [] instead of parenthesis () for matrix or array indexing, 2) indexing start from 0 instead of 1. To get going with Numpy, study what happens after every line of the following code.

```
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: Pandas is another package that might be useful for reading and preprocessing tabular data. It emulates R and the functions provided by this package are similar to those in R language. If you do not know R, do not worry. To get an idea how Pandas work, study lines 11 to 47 in the demo_visualization.py. Take a look at Pandas documentation to get an idea how to use it and what functionality it offers.

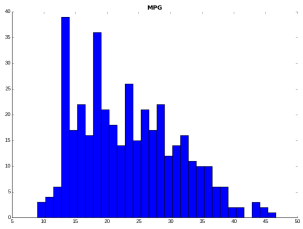

Figure 1: Histogram

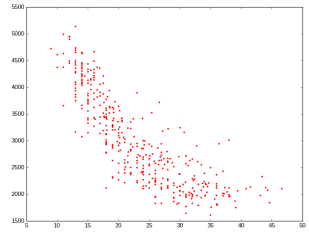

Figure 2: Scatter 1

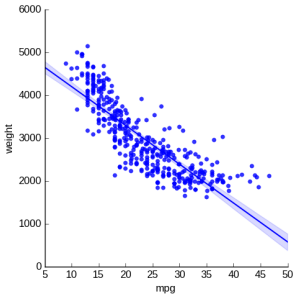

Figure 3: Scatter 2

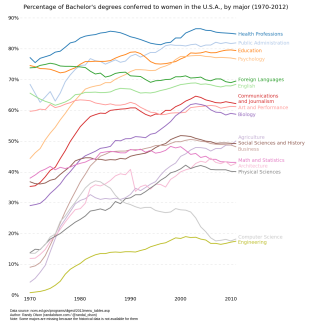

Figure 4: Line Plot

2.2 Data visualizaiton with Matplotlib

Matplotlib is the most popular package for data visualization. Take a moment to look at the gallery Matplotlib. Take a look at Matplotlib documentation to get an idea how to use it and what functionality it offers.

Data visualization is half science and half art. Please think about it when creating displays of your data. 1) What kinds of attributes do you have in the data set, e.g. categorical, continuous, time series, geographical? 2) What story do you want to tell? What kind of plots can help you to tell the story, e.g. histogram, word cloud, bar plot, box plot, scatter plot, time series plot, heatmap? 3) Do you have the data prepared for the plot you want, e.g. word frequency of twitter data? 4) What is a good way to present the data, e.g., one picture per figure, or multiple pictures in one figure with subplots, interactive plots or static plots? 5) How to improve visual appeal of the visualizations, e.g., choice of color palette, concise and professional lines and dots, understandable legend, labels, annotations (e.g., in his talk, Hans Rosling used colors to present continents).

Histogram, Scatter, Line Plot

Let us make some plots.

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

You will be able to see the four figures: Figure 1, Figure 2, Figure 3, Figure 4:

Detailed comments are given in the script.

3 Exploratory analysis of Car MPG data

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 " Auto MPG Data Set" available from https://archive.ics.uci.edu/ml/datasets/Auto+MPG. We provide a code in Lab_3_carmpg.py that allows you to load the data into python. Do the following:

1. How many cars and how many attributes are in the data set.

2. How many distinct car companies are represented in the data set? What is the name of the car with the best MPG? What car company produced the most 8-cylinder cars? What are the names of 3-cylinder cars? Do some internet search that can tell you about the history and popularity of those 3-cylinder cars.

3. What is the range, mean, and standard deviation of each attribute? Pay attention to potential missing values.

4. Plot histograms for each attribute. Pay attention to the appropriate choice of number of bins. Write 2-3 sentences summarizing some interesting aspects of the data by looking at the histograms.

5. Plot a scatterplot of weight vs. MPG attributes. What do you conclude about the relationship between the attributes? What is the correlation coefficient between the 2 attributes?

6. Plot a scatterplot of year vs. cylinders attributes. Add a small random noise to the values to make the scatterplot look nicer. What can you conclude? Do some internet search about the history of car industry during 70's that might explain the results.(*Hint:* `data.mpg + np.random.random(len(data.mpg))` will add small random noise)

7. Show 2 more scatterplots that are interesting do you. Discuss what you see.

8. Plot a time series for all the companies that show how many new cars they introduces during each year. Do you see some interesting trends? (*Hint:* `data.car_name.str.split()[0]` returns a vector of the first word of `car_name` column.)

9. Calculate the pairwise correlation, and draw the heatmap with Matplotlib. Do you see some interesting correlation? (*Hint:* `data.iloc[:,0:8].corr()`, `plt.pcolor()` draws the heatmap.)

4 Electric power consumption data [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:

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

- Date: Date in format `dd/mm/yyyy`
- Time: time in format `hh:mm:ss`
- 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['15/01/2008':'28/01/2008']]` selects all records from 2013-1-15 to 2013-1-28.
- Note that in this dataset missing values are coded as `?`.

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. It's not necessarily to be exactly the same, e.g., for Figure 6, the x-axis can be based on hours.

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_electricpower.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.

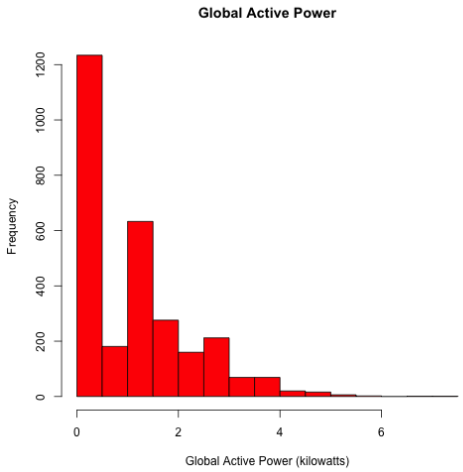

Figure 5: Plot 1

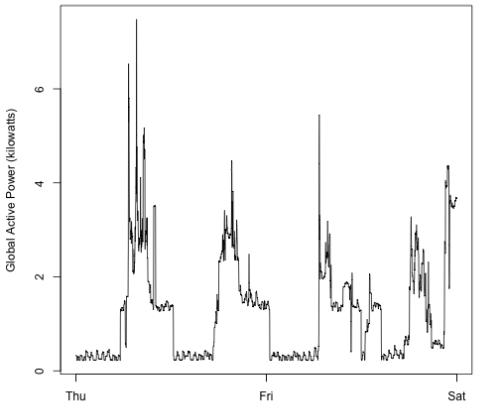

Figure 6: Plot 2

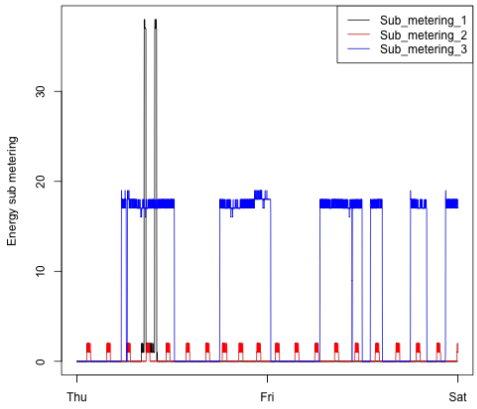

Figure 7: Plot 3

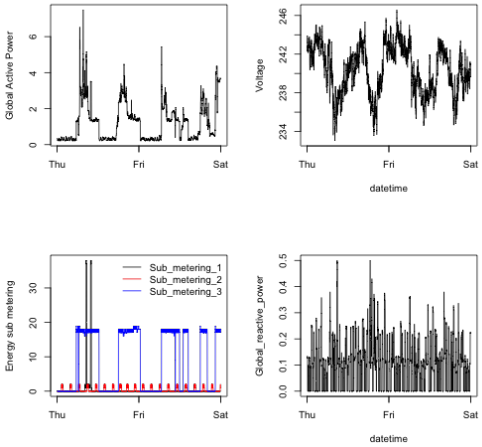

Figure 8: Plot 4

Write a story about the energy use of the household during the two days. Base the story on the plots you have made.

5 Submission

Only use Numpy, Pandas, Matplotlib for all the tasks of this lab. Please inform me if you really have a reason to use other packages. You are allowed to submit an assignment up to five (5) times in total. Each submission will replace the previous.You can discuss your solutions, search for materials online, but all your code should be written by youself.

Please submit the Python scripts Lab_3_carmpg.py, Lab_3_electricpower.py, and your writeup documents for Section 3-4 through blackboard.