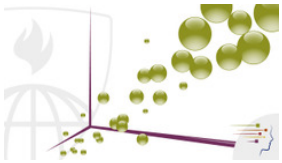
JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Exploratory Data Analysis

by Roger D. Peng, PhD, Jeff Leek, PhD, Brian Caffo, PhD



due in 4day 7h

Submission Phase	Evaluation Phase	Results Phase
1. Do assignment ☐	2. Evaluate peers ☐	3. See results ☐

☐ In accordance with the Honor Code, I certify that my answers here are my own work, and that I have appropriately acknowledged all external sources (if any) that were used in this work.

Save draft Submit for grading

Introduction

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. **Date**: Date in format dd/mm/yyyy
2. **Time**: time in format hh:mm:ss
3. **Global_active_power**: household global minute-averaged active power (in kilowatt)
4. **Global_reactive_power**: household global minute-averaged reactive power (in kilowatt)
5. **Voltage**: minute-averaged voltage (in volt)
6. **Global_intensity**: household global minute-averaged current intensity (in ampere)
7. **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).
8. **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.
9. **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.

Loading the data

When loading the dataset into R, 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 R. Make sure your computer has enough memory (most modern computers should be fine).
- 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 R using the `strptime()` and `as.Date()` functions.
- Note that in this dataset missing values are coded as `?`.

Making Plots

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

First you will need to fork and clone the following GitHub repository: https://github.com/rdpeng/ExData_Plotting1

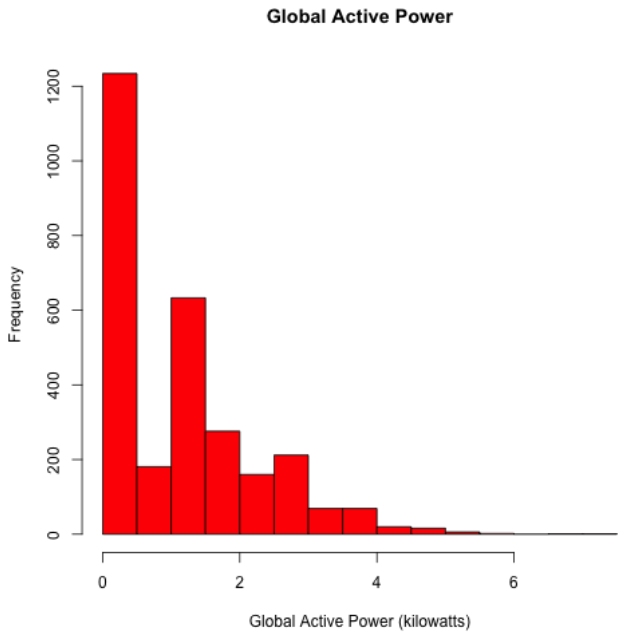
For each plot you should

- Construct the plot and save it to a PNG file with a width of 480 pixels and a height of 480 pixels.
- Name each of the plot files as `plot1.png`, `plot2.png`, etc.
- Create a separate R code file (`plot1.R`, `plot2.R`, etc.) that constructs the corresponding plot, i.e. code in `plot1.R` 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.
- Add the PNG file and R code file to your git repository

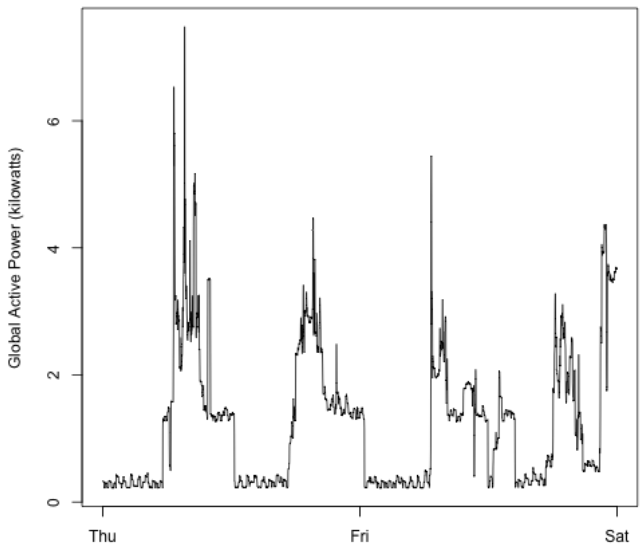
When you are finished with the assignment, push your git repository to GitHub so that the GitHub version of your repository is up to date. There should be four PNG files and four R code files.

The four plots that you will need to construct are shown below.

Plot 1

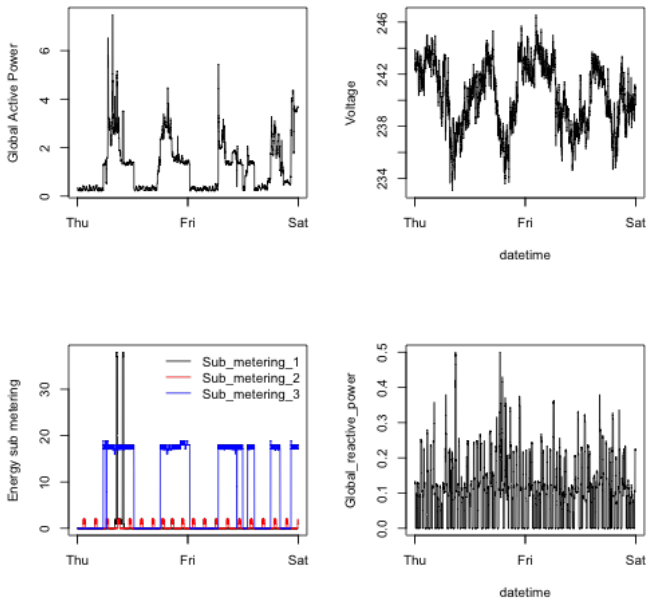


Plot 2



Plot 3

Plot 4



Please submit the URL pointing to your GitHub repository containing the completed R code for this assignment.

B I ☐ ☐ ☐ Link <code> Math Edit: Rich ☐ Preview

Evaluation/feedback on the above work

Note: this section can only be filled out during the evaluation phase.

Was a valid GitHub URL containing a git repository submitted?

0 points: A valid GitHub URL was NOT submitted (or URL is broken)

1 point: The submitted URL points to a GitHub repository

Does the GitHub repository contain at least one commit beyond the original fork?

0 points: No, there are no commits beyond the original fork
1 point: Yes, there is at least one commit beyond the original fork

Overall evaluation/feedback

Note: this section can only be filled out during the evaluation phase.

Please examine the plot files in the GitHub repository. Do the plot files appear to be of the correct graphics file format?

0 points: No, at least one of the files appears to be in the wrong format
1 point: Yes, all of the files appear to be in the correct format

Please view the image file for Plot 1 from the GitHub repository. Does the plot appear correct?

0 points: No, the plot appears incorrect in at least 1 major discrepancy from the reference plot (e.g. wrong data), or at least 2 minor discrepancies (e.g. x-label is incorrect, title is incorrect), or the plot was not viewable
1 point: The plot is mostly correct with at most one minor discrepancy from the reference plot

Please evaluate the code for Plot 1, but **do not run the code** on your computer. Does the code appear to create the plot reference plot given in the assignment?

0 points: The code does not create the reference plot, or is not viewable/present in repository
1 point: The code is mostly correct, but does not reproduce the reference plot exactly
2 points: The code reproduces the reference plot exactly

Please view the image file for Plot 2 from the GitHub repository. Does the plot appear correct?

0 points: No, the plot appears incorrect in at least 1 major discrepancy from the reference plot (e.g. wrong data), or at least 2 minor discrepancies (e.g. x-label is incorrect, title is incorrect), or the plot was not viewable
1 point: The plot is mostly correct with at most one minor discrepancy from the reference plot

Please evaluate the code for Plot 2, but **do not run the code** on your computer. Does the code appear to create the plot reference plot given in the assignment?

0 points: The code does not create the reference plot, or is not viewable/present in repository
1 point: The code is mostly correct, but does not reproduce the reference plot exactly
2 points: The code reproduces the reference plot exactly

Please view the image file for Plot 3 from the GitHub repository. Does the plot appear correct?

0 points: No, the plot appears incorrect in at least 1 major discrepancy from the reference plot (e.g. wrong data), or at least 2 minor discrepancies (e.g. x-label is incorrect, title is incorrect), or the plot was not viewable
1 point: The plot is mostly correct with at most one minor discrepancy from the reference plot

Please evaluate the code for Plot 3, but **do not run the code** on your computer. Does the code appear to create the plot reference plot given in the assignment?

0 points: The code does not create the reference plot, or is not viewable/present in repository
1 point: The code is mostly correct, but does not reproduce the reference plot exactly
2 points: The code reproduces the reference plot exactly

Please view the image file for Plot 4 from the GitHub repository. Does the plot appear correct?

0 points: No, the plot appears incorrect in at least 1 major discrepancy from the reference plot (e.g. wrong data), or at least 2 minor discrepancies (e.g. x-label is incorrect, title is incorrect), or the plot was not viewable
1 point: The plot is mostly correct with at most one minor discrepancy from the reference plot

Please evaluate the code for Plot 4, but **do not run the code** on your computer. Does the code appear to create the plot reference plot given in the assignment?

0 points: The code does not create the reference plot, or is not viewable/present in repository
1 point: The code is mostly correct, but does not reproduce the reference plot exactly
2 points: The code reproduces the reference plot exactly

Please use the space below to provide constructive feedback to the student who submitted the work. Point out the submission's strengths as well as areas in need of improvement. You may also use this space to explain your grading decisions.

You've written 0 words

☐ In accordance with the Honor Code, I certify that my answers here are my own work, and that I have appropriately acknowledged all external sources (if any) that were used in this work.

Save draft Submit for grading