Exploratory Data Analysis Example

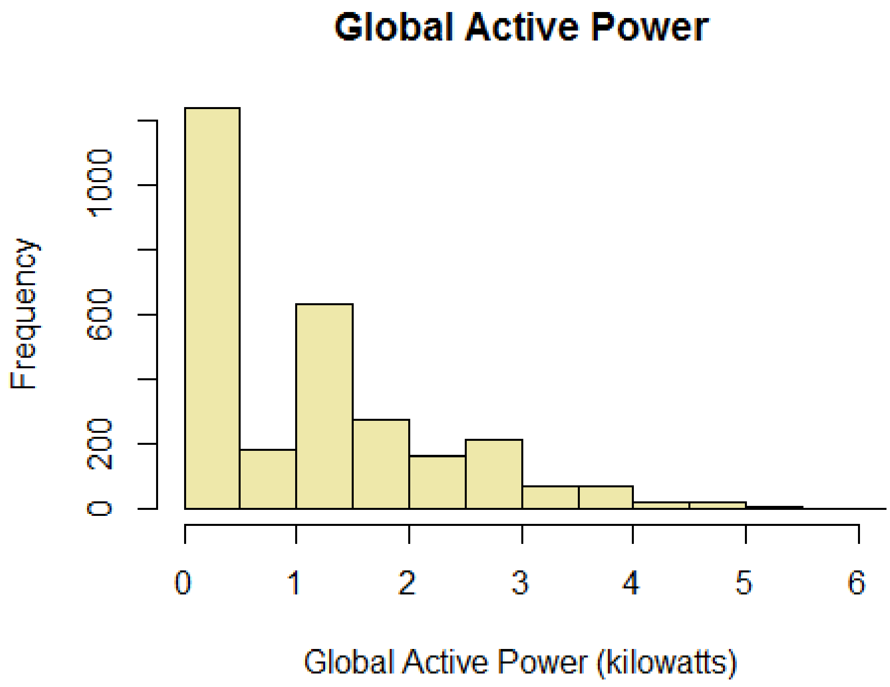
Linda Wang

January 17, 2016

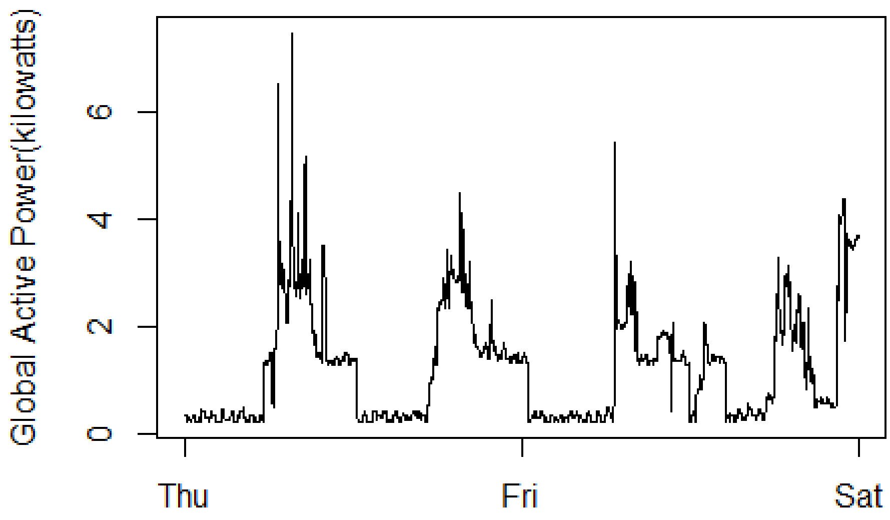
See detailed information after the code about what these scripts do.

The original project was done by writing 4 separate, independent script files. Expect to see some redundant code in this markdown file as the 4 scripts are laid out as they are by themselves.

```
data_row1 <-
read.table("household_power_consumption.txt",header=TRUE,nrow=1,na.strings="?
",sep=";")
nCols <- ncol(data_row1)
data_date <-
read.table("household_power_consumption.txt",colClasses=c("character",rep("NU
LL",nCols-1)),header=TRUE,as.is=TRUE,na.strings="?",sep=";")
data_date <- as.Date(as.vector(data_date[,1]),"%d/%m/%Y")
data_date <- data.frame(data_date)
colnames(data_date) <- c("Date")
start_date_idx <- which.max(data_date$Date>="2007-02-01")
end_date_idx <- which.min(data_date$Date<="2007-02-02")-1
data <-
read.table("household_power_consumption.txt",skip=start_date_idx,nrows=(end_d
ate_idx-
start_date_idx+1),col.names=names(data_row1),as.is=TRUE,na.strings="?",sep=";
")

hist(data$Global_active_power,breaks=12,freq=TRUE,col="palegoldenrod",border=
"Black",xlim=c(0,6),ylim=c(0,1200),xlab="Global Active Power
(kilowatts)",ylab="Frequency",main="Global Active Power")
```

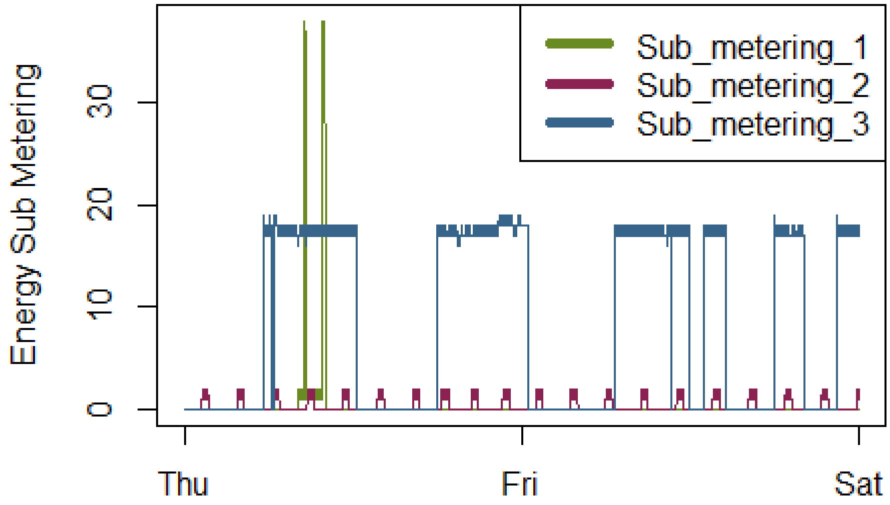


```
data_row1 <-
read.table("household_power_consumption.txt",header=TRUE,nrow=1,na.strings="?
",sep=";")
nCols <- ncol(data_row1)
data_date <-
read.table("household_power_consumption.txt",colClasses=c("character",rep("NU
LL",nCols-1)),header=TRUE,as.is=TRUE,na.strings="?",sep=";")
data_date <- as.Date(as.vector(data_date[,1]),"%d/%m/%Y")
data_date <- data.frame(data_date)
colnames(data_date) <- c("Date")
start_date_idx <- which.max(data_date$Date>="2007-02-01")
end_date_idx <- which.min(data_date$Date<="2007-02-02")-1
data <-
read.table("household_power_consumption.txt",skip=start_date_idx,nrows=(end_d
ate_idx-
start_date_idx+1),col.names=names(data_row1),as.is=TRUE,na.strings="?",sep=";
")

plot(strptime(paste(data$Date,data$Time),"%d/%m/%Y
%H:%M:%S"),data$Global_active_power,type="l",xlab="",ylab="Global Active
Power(kilowatts)")
```

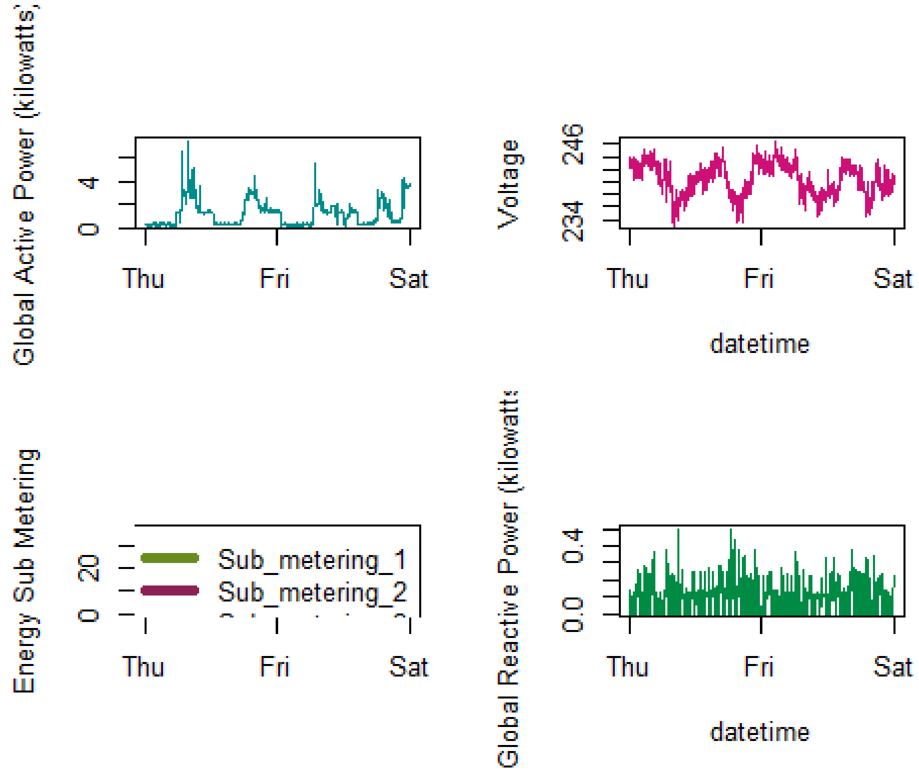


```
data_row1 <-
read.table("household_power_consumption.txt",header=TRUE,nrow=1,na.strings="?
",sep=";")
nCols <- ncol(data_row1)
data_date <-
read.table("household_power_consumption.txt",colClasses=c("character",rep("NU
LL",nCols-1)),header=TRUE,as.is=TRUE,na.strings="?",sep=";")
data_date <- as.Date(as.vector(data_date[,1]),"%d/%m/%Y")
data_date <- data.frame(data_date)
colnames(data_date) <- c("Date")
start_date_idx <- which.max(data_date$Date>="2007-02-01")
end_date_idx <- which.min(data_date$Date<="2007-02-02")-1
data <-
read.table("household_power_consumption.txt",skip=start_date_idx,nrows=(end_d
ate_idx-
start_date_idx+1),col.names=names(data_row1),as.is=TRUE,na.strings="?",sep=";
")

plot(strptime(paste(data$Date,data$Time),"%d/%m/%Y
%H:%M:%S"),data$Sub_metering_1,type="l",xlab="",ylab="Energy Sub
Metering",col="olivedrab4")
lines(strptime(paste(data$Date,data$Time),"%d/%m/%Y
%H:%M:%S"),data$Sub_metering_2,col="violetred4")
lines(strptime(paste(data$Date,data$Time),"%d/%m/%Y
%H:%M:%S"),data$Sub_metering_3,col="steelblue4")
```

```
legend("topright",c("Sub_metering_1","Sub_metering_2","Sub_metering_3"),lty=c
(1,1,1),lwd=c(5,5,5),col=c("olivedrab4","violetred4","steelblue4"))
```



```
data_row1 <-
read.table("household_power_consumption.txt",header=TRUE,nrow=1,na.strings="?
",sep=";")
nCols <- ncol(data_row1)
data_date <-
read.table("household_power_consumption.txt",colClasses=c("character",rep("NU
LL",nCols-1)),header=TRUE,as.is=TRUE,na.strings="?",sep=";")
data_date <- as.Date(as.vector(data_date[,1]),"%d/%m/%Y")
data_date <- data.frame(data_date)
colnames(data_date) <- c("Date")
start_date_idx <- which.max(data_date$Date>="2007-02-01")
end_date_idx <- which.min(data_date$Date<="2007-02-02")-1
data <-
read.table("household_power_consumption.txt",skip=start_date_idx,nrows=(end_d
ate_idx-
start_date_idx+1),col.names=names(data_row1),as.is=TRUE,na.strings="?",sep=";
")

par(mfrow=c(2,2))
plot(strptime(paste(data$Date,data$Time),"%d/%m/%Y
%H:%M:%S"),data$Global_active_power,type="l",xlab="",ylab="Global Active
Power (kilowatts)",col="cyan4")
plot(strptime(paste(data$Date,data$Time),"%d/%m/%Y
%H:%M:%S"),data$Voltage,type="l",xlab="datetime",ylab="Voltage",col="deeppink
3")
plot(strptime(paste(data$Date,data$Time),"%d/%m/%Y
%H:%M:%S"),data$Sub_metering_1,type="l",xlab="",ylab="Energy Sub
Metering",col="olivedrab4")
lines(strptime(paste(data$Date,data$Time),"%d/%m/%Y
%H:%M:%S"),data$Sub_metering_2,col="violetred4")
lines(strptime(paste(data$Date,data$Time),"%d/%m/%Y
%H:%M:%S"),data$Sub_metering_3,col="steelblue4")
legend("topright",c("Sub_metering_1","Sub_metering_2","Sub_metering_3"),lty=c
(1,1,1),lwd=c(5,5,5),col=c("olivedrab4","violetred4","steelblue4"))
plot(strptime(paste(data$Date,data$Time),"%d/%m/%Y
%H:%M:%S"),data$Global_reactive_power,type="l",xlab="datetime",ylab="Global
Reactive Power (kilowatts)",col="springgreen4")
```



Introduction

This assignment uses data from the UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml/), 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
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.

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.

For each plot you should 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.