Plot1 for EDA PS 1

Harold Pollack

9/14/2018

Read the UCI data. Then save the dates 2/1/2007 and 2/2/2007. Note that the date variable is dd/mm/y. So some care must be taken. ALso cache=True given size of the dataset.

mydata <- read.csv("/Users/haroldpollack/Documents/coursera\_EDA/pS1/household\_power\_consumption.txt",header=TRUE,sep=";")  
mydata$date\_formatted <-as.Date(mydata$Date,format="%d/%m/%Y")  
str(mydata)

## 'data.frame': 2075259 obs. of 10 variables:  
## $ Date : Factor w/ 1442 levels "1/1/2007","1/1/2008",..: 342 342 342 342 342 342 342 342 342 342 ...  
## $ Time : Factor w/ 1440 levels "00:00:00","00:01:00",..: 1045 1046 1047 1048 1049 1050 1051 1052 1053 1054 ...  
## $ Global\_active\_power : Factor w/ 4187 levels "?","0.076","0.078",..: 2082 2654 2661 2668 1807 1734 1825 1824 1808 1805 ...  
## $ Global\_reactive\_power: Factor w/ 533 levels "?","0.000","0.046",..: 189 198 229 231 244 241 240 240 235 235 ...  
## $ Voltage : Factor w/ 2838 levels "?","223.200",..: 992 871 837 882 1076 1010 1017 1030 907 894 ...  
## $ Global\_intensity : Factor w/ 222 levels "?","0.200","0.400",..: 53 81 81 81 40 36 40 40 40 40 ...  
## $ Sub\_metering\_1 : Factor w/ 89 levels "?","0.000","1.000",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ Sub\_metering\_2 : Factor w/ 82 levels "?","0.000","1.000",..: 3 3 14 3 3 14 3 3 3 14 ...  
## $ Sub\_metering\_3 : num 17 16 17 17 17 17 17 17 17 16 ...  
## $ date\_formatted : Date, format: "2006-12-16" "2006-12-16" ...

summary(mydata)

## Date Time Global\_active\_power  
## 1/1/2007 : 1440 17:24:00: 1442 ? : 25979   
## 1/1/2008 : 1440 17:25:00: 1442 0.218 : 9565   
## 1/1/2009 : 1440 17:26:00: 1442 0.216 : 9363   
## 1/1/2010 : 1440 17:27:00: 1442 0.322 : 9350   
## 1/10/2007: 1440 17:28:00: 1442 0.324 : 9304   
## 1/10/2008: 1440 17:29:00: 1442 0.326 : 9187   
## (Other) :2066619 (Other) :2066607 (Other):2002511   
## Global\_reactive\_power Voltage Global\_intensity   
## 0.000 : 481561 ? : 25979 1.000 : 172785   
## ? : 25979 241.180: 3559 1.400 : 164720   
## 0.100 : 21902 241.120: 3289 1.200 : 140425   
## 0.092 : 21841 241.350: 3285 1.600 : 118212   
## 0.098 : 21812 241.020: 3284 1.800 : 95395   
## 0.096 : 21792 241.270: 3247 0.800 : 82272   
## (Other):1480372 (Other):2032616 (Other):1301450   
## Sub\_metering\_1 Sub\_metering\_2 Sub\_metering\_3 date\_formatted   
## 0.000 :1880175 0.000 :1436830 Min. : 0.000 Min. :2006-12-16   
## 1.000 : 84936 1.000 : 378224 1st Qu.: 0.000 1st Qu.:2007-12-12   
## ? : 25979 2.000 : 157075 Median : 1.000 Median :2008-12-06   
## 2.000 : 19017 ? : 25979 Mean : 6.458 Mean :2008-12-05   
## 38.000 : 16119 3.000 : 7189 3rd Qu.:17.000 3rd Qu.:2009-12-01   
## 37.000 : 14892 37.000 : 6638 Max. :31.000 Max. :2010-11-26   
## (Other): 34141 (Other): 63324 NA's :25979

# table(mydata$Date)  
#  
# now extract the two dates called for in the assignment  
#  
data\_for\_graphing <-subset(mydata,(Date== "1/2/2007" | Date == "2/2/2007"))  
str(data\_for\_graphing)

## 'data.frame': 2880 obs. of 10 variables:  
## $ Date : Factor w/ 1442 levels "1/1/2007","1/1/2008",..: 16 16 16 16 16 16 16 16 16 16 ...  
## $ Time : Factor w/ 1440 levels "00:00:00","00:01:00",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ Global\_active\_power : Factor w/ 4187 levels "?","0.076","0.078",..: 127 127 126 126 125 124 124 124 124 82 ...  
## $ Global\_reactive\_power: Factor w/ 533 levels "?","0.000","0.046",..: 44 45 46 47 45 43 43 43 44 2 ...  
## $ Voltage : Factor w/ 2838 levels "?","223.200",..: 1823 1840 1859 1898 1824 1737 1754 1771 1778 1797 ...  
## $ Global\_intensity : Factor w/ 222 levels "?","0.200","0.400",..: 8 8 8 8 8 8 8 8 8 6 ...  
## $ Sub\_metering\_1 : Factor w/ 89 levels "?","0.000","1.000",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ Sub\_metering\_2 : Factor w/ 82 levels "?","0.000","1.000",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ Sub\_metering\_3 : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ date\_formatted : Date, format: "2007-02-01" "2007-02-01" ...

summary(data\_for\_graphing)

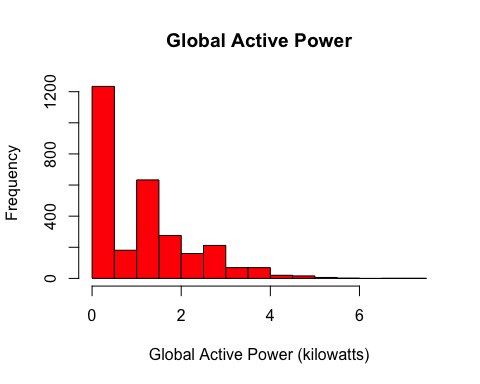
## Date Time Global\_active\_power Global\_reactive\_power  
## 1/2/2007:1440 00:00:00: 2 0.226 : 60 0.000 : 991   
## 2/2/2007:1440 00:01:00: 2 0.224 : 51 0.116 : 71   
## 1/1/2007: 0 00:02:00: 2 0.228 : 51 0.122 : 67   
## 1/1/2008: 0 00:03:00: 2 0.232 : 50 0.094 : 52   
## 1/1/2009: 0 00:04:00: 2 0.230 : 39 0.118 : 51   
## 1/1/2010: 0 00:05:00: 2 0.300 : 33 0.124 : 49   
## (Other) : 0 (Other) :2868 (Other):2596 (Other):1599   
## Voltage Global\_intensity Sub\_metering\_1 Sub\_metering\_2  
## 238.130: 11 1.000 : 391 0.000 :2764 0.000 :2305   
## 241.570: 10 1.400 : 356 1.000 : 70 1.000 : 408   
## 242.180: 10 5.600 : 192 37.000 : 21 2.000 : 167   
## 239.360: 9 1.200 : 163 2.000 : 16 ? : 0   
## 239.820: 9 5.800 : 122 38.000 : 6 10.000 : 0   
## 241.410: 9 2.000 : 112 17.000 : 1 11.000 : 0   
## (Other):2822 (Other):1544 (Other): 2 (Other): 0   
## Sub\_metering\_3 date\_formatted   
## Min. : 0.000 Min. :2007-02-01   
## 1st Qu.: 0.000 1st Qu.:2007-02-01   
## Median : 0.000 Median :2007-02-01   
## Mean : 8.501 Mean :2007-02-01   
## 3rd Qu.:17.000 3rd Qu.:2007-02-02   
## Max. :19.000 Max. :2007-02-02   
##

# table(data\_for\_graphing$Date)  
table(data\_for\_graphing$date\_formatted)

##   
## 2007-02-01 2007-02-02   
## 1440 1440

make the initial histogram and save as plot1.png

#  
# note that data\_for\_graphing$Global\_active\_power is a factor. It has to be changed to numeric for graphing  
#  
# Global\_active\_power is a factor. So one must be careful converting it to numeric  
#  
Global\_active\_power\_numeric <- as.numeric(as.character(data\_for\_graphing$Global\_active\_power))   
hist(Global\_active\_power\_numeric,col="red",xlab="Global Active Power (kilowatts)",main="Global Active Power",ylab="Frequency")



#  
# now save the png file  
#  
dev.copy(png,'/Users/haroldpollack/Documents/coursera\_datascience3/Plot1/plot1.png')

## quartz\_off\_screen   
## 3

dev.off()

## quartz\_off\_screen   
## 2