hpc3

Jonathan Talbot

9/30/2021

#install.packages("tidyr")  
#library(tidyverse)  
#install.packages("readr")  
#library(readr)  
#install.packages("rio")  
#library(rio)  
#install.packages("sqldf")  
#library(sqldf)  
#install.packages("lubridate")  
#library(lubridate)  
#install.packages("ggplot2")  
#library(ggplot2)  
#install.packages("magrittr")  
#library(magrittr)

df <-data.frame()  
df <- read.delim("household\_power\_consumption.txt", sep = ";") #load file from local drive  
  
df <- within(df, rm(Voltage, Global\_intensity))  
df<-tibble::as\_tibble(df)  
print(head(df,10))

## # A tibble: 10 x 7  
## Date Time Global\_active\_p~ Global\_reactive~ Sub\_metering\_1 Sub\_metering\_2  
## <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 16/12/2006 17:24:00 4.216 0.418 0.000 1.000   
## 2 16/12/2006 17:25:00 5.360 0.436 0.000 1.000   
## 3 16/12/2006 17:26:00 5.374 0.498 0.000 2.000   
## 4 16/12/2006 17:27:00 5.388 0.502 0.000 1.000   
## 5 16/12/2006 17:28:00 3.666 0.528 0.000 1.000   
## 6 16/12/2006 17:29:00 3.520 0.522 0.000 2.000   
## 7 16/12/2006 17:30:00 3.702 0.520 0.000 1.000   
## 8 16/12/2006 17:31:00 3.700 0.520 0.000 1.000   
## 9 16/12/2006 17:32:00 3.668 0.510 0.000 1.000   
## 10 16/12/2006 17:33:00 3.662 0.510 0.000 2.000   
## # ... with 1 more variable: Sub\_metering\_3 <dbl>

#install.packages("lubridate")  
library(lubridate)

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

df$Date <- lubridate::dmy(df$Date)  
head(df)

## # A tibble: 6 x 7  
## Date Time Global\_active\_power Global\_reactive\_power Sub\_metering\_1  
## <date> <chr> <chr> <chr> <chr>   
## 1 2006-12-16 17:24:00 4.216 0.418 0.000   
## 2 2006-12-16 17:25:00 5.360 0.436 0.000   
## 3 2006-12-16 17:26:00 5.374 0.498 0.000   
## 4 2006-12-16 17:27:00 5.388 0.502 0.000   
## 5 2006-12-16 17:28:00 3.666 0.528 0.000   
## 6 2006-12-16 17:29:00 3.520 0.522 0.000   
## # ... with 2 more variables: Sub\_metering\_2 <chr>, Sub\_metering\_3 <dbl>

df$DT <- lubridate::as\_datetime(paste(df$Date, df$Time, sep=" "))  
head(df)

## # A tibble: 6 x 8  
## Date Time Global\_active\_power Global\_reactive\_power Sub\_metering\_1  
## <date> <chr> <chr> <chr> <chr>   
## 1 2006-12-16 17:24:00 4.216 0.418 0.000   
## 2 2006-12-16 17:25:00 5.360 0.436 0.000   
## 3 2006-12-16 17:26:00 5.374 0.498 0.000   
## 4 2006-12-16 17:27:00 5.388 0.502 0.000   
## 5 2006-12-16 17:28:00 3.666 0.528 0.000   
## 6 2006-12-16 17:29:00 3.520 0.522 0.000   
## # ... with 3 more variables: Sub\_metering\_2 <chr>, Sub\_metering\_3 <dbl>,  
## # DT <dttm>

#install.packages("dplyr")  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

#limit to relevant dates  
print("number of rows of df:")

## [1] "number of rows of df:"

print(nrow(df))

## [1] 2075259

df\_limited <- dplyr::filter(df, Date == "2007-02-01" | Date == "2007-02-02")  
print("number of rows of df\_limited:")

## [1] "number of rows of df\_limited:"

print(nrow(df\_limited))

## [1] 2880

print(head(df\_limited))

## # A tibble: 6 x 8  
## Date Time Global\_active\_power Global\_reactive\_power Sub\_metering\_1  
## <date> <chr> <chr> <chr> <chr>   
## 1 2007-02-01 00:00:00 0.326 0.128 0.000   
## 2 2007-02-01 00:01:00 0.326 0.130 0.000   
## 3 2007-02-01 00:02:00 0.324 0.132 0.000   
## 4 2007-02-01 00:03:00 0.324 0.134 0.000   
## 5 2007-02-01 00:04:00 0.322 0.130 0.000   
## 6 2007-02-01 00:05:00 0.320 0.126 0.000   
## # ... with 3 more variables: Sub\_metering\_2 <chr>, Sub\_metering\_3 <dbl>,  
## # DT <dttm>

df[is.na(df)] <- 0 #replace NAs with 0s  
any(is.na(df)) # check for any NAs remaining - should return False

## [1] FALSE

df$Global\_active\_power<-as.numeric(df$Global\_active\_power) #"Numeric" #change class of Global\_active\_power column from character to numeric

## Warning: NAs introduced by coercion

df$Global\_reactive\_power<-as.numeric(df$Global\_reactive\_power)

## Warning: NAs introduced by coercion

df$Sub\_metering\_1<-as.numeric(df$Sub\_metering\_1)

## Warning: NAs introduced by coercion

df$Sub\_metering\_2<-as.numeric(df$Sub\_metering\_2)

## Warning: NAs introduced by coercion

df$Sub\_metering\_3<-as.numeric(df$Sub\_metering\_3)  
print(head(df,10))

## # A tibble: 10 x 8  
## Date Time Global\_active\_power Global\_reactive\_power Sub\_metering\_1  
## <date> <chr> <dbl> <dbl> <dbl>  
## 1 2006-12-16 17:24:00 4.22 0.418 0  
## 2 2006-12-16 17:25:00 5.36 0.436 0  
## 3 2006-12-16 17:26:00 5.37 0.498 0  
## 4 2006-12-16 17:27:00 5.39 0.502 0  
## 5 2006-12-16 17:28:00 3.67 0.528 0  
## 6 2006-12-16 17:29:00 3.52 0.522 0  
## 7 2006-12-16 17:30:00 3.70 0.52 0  
## 8 2006-12-16 17:31:00 3.7 0.52 0  
## 9 2006-12-16 17:32:00 3.67 0.51 0  
## 10 2006-12-16 17:33:00 3.66 0.51 0  
## # ... with 3 more variables: Sub\_metering\_2 <dbl>, Sub\_metering\_3 <dbl>,  
## # DT <dttm>

#convert .txt to .csv  
library("sqldf")

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

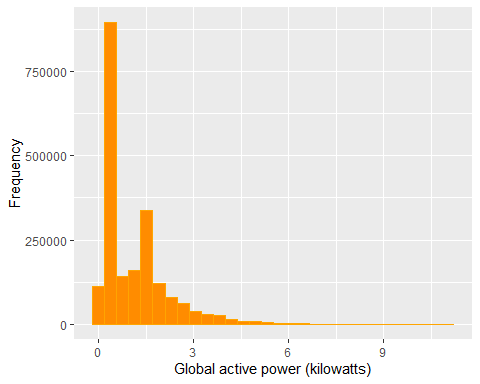
write.csv(df,"household\_power\_consumption.csv", row.names = TRUE)  
print(summary(df))

## Date Time Global\_active\_power  
## Min. :2006-12-16 Length:2075259 Min. : 0.076   
## 1st Qu.:2007-12-12 Class :character 1st Qu.: 0.308   
## Median :2008-12-06 Mode :character Median : 0.602   
## Mean :2008-12-05 Mean : 1.092   
## 3rd Qu.:2009-12-01 3rd Qu.: 1.528   
## Max. :2010-11-26 Max. :11.122   
## NA's :25979   
## Global\_reactive\_power Sub\_metering\_1 Sub\_metering\_2 Sub\_metering\_3   
## Min. :0.000 Min. : 0.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.:0.048 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.000   
## Median :0.100 Median : 0.000 Median : 0.000 Median : 1.000   
## Mean :0.124 Mean : 1.122 Mean : 1.299 Mean : 6.378   
## 3rd Qu.:0.194 3rd Qu.: 0.000 3rd Qu.: 1.000 3rd Qu.:17.000   
## Max. :1.390 Max. :88.000 Max. :80.000 Max. :31.000   
## NA's :25979 NA's :25979 NA's :25979   
## DT   
## Min. :2006-12-16 17:24:00   
## 1st Qu.:2007-12-12 00:18:30   
## Median :2008-12-06 07:13:00   
## Mean :2008-12-06 07:13:00   
## 3rd Qu.:2009-12-01 14:07:30   
## Max. :2010-11-26 21:02:00   
##

#install.packages("ggplot2")  
library(ggplot2)  
  
#create plot  
#png("Rplot1.png",width = 480, height = 480, units = "px", pointsize = 12)  
  
#myplot<-  
 ggplot(data=df, mapping=aes(x=Global\_active\_power ))+  
geom\_histogram(color="orange", fill="dark orange")+  
 ylab("Frequency")+  
 xlab("Global active power (kilowatts)")

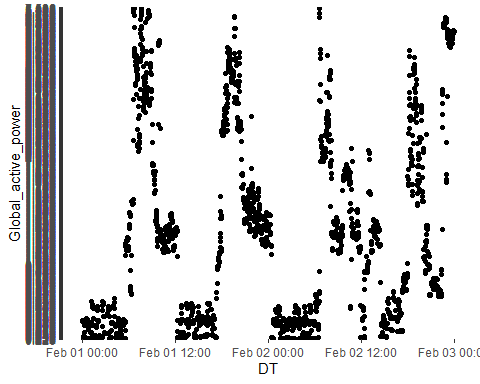
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 25979 rows containing non-finite values (stat\_bin).



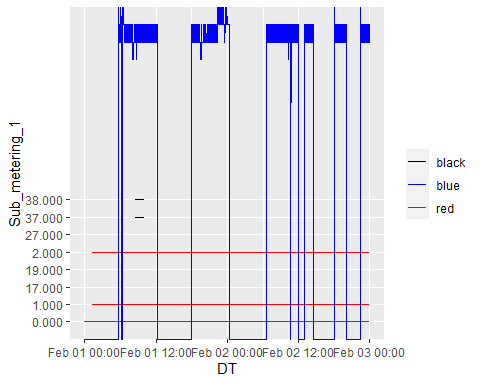
#save as .png 480 x 480  
  
#print(myplot)  
#dev.off()

#print(head(df\_limited))  
#Plot number 2  
#png("Rplot2.png",width = 480, height = 480, units = "px", pointsize = 12)  
  
#library(ggplot2)  
  
#myplot2<-  
ggplot2::ggplot(data=df\_limited, mapping=aes(x=DT, y=Global\_active\_power))+  
 geom\_point()



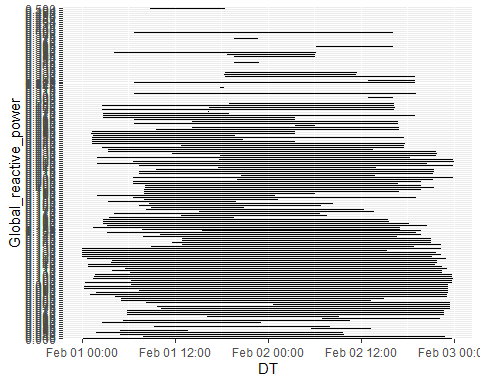
#print(myplot2)  
#dev.off()

#png("Rplot3.png", width = 480, height = 480, units = "px", pointsize = 12)   
  
#myplot3 <-   
 ggplot(data=df\_limited)+  
 geom\_line(mapping=aes(x=DT, y = Sub\_metering\_1, color = "black"))+  
 geom\_line(mapping=aes(x=DT, y = Sub\_metering\_2, color = "red"))+   
 geom\_line(mapping=aes(x=DT, y = Sub\_metering\_3, color = "blue"))+  
 labs(color="")+  
 scale\_color\_manual(values=c("black","blue","red"))



#print(myplot3)  
#dev.off()

#myplot4   
   
  
  
#Plot number 4  
#myplot4 <-png(filename = "Rplot4.png", width = 480, height = 480, units = "px", pointsize = 12)   
  
ggplot(data=df\_limited,mapping = aes(x=DT,y=Global\_reactive\_power))+  
 geom\_line()



#print(myplot4)  
#dev.off()

#png(filename = "Rplot4.png", width = 480, height = 480, units = "px", pointsize = 12)   
  
  
#myplot5<-par(mfrow=c(2,2))  
#myplot  
#myplot2  
#myplot3  
#myplot4  
  
#print(myplot5)  
#dev.off()