M4T1.R

christiancobollogomez

2022-02-01

```
## Module 4 Deep Analytics and Visualization Task 1: Domain research and
# Exploratory data analysis.

#install.packages("RMySQL")
#install.packages("dplyr")
#install.packages("lubridate")

library(RMySQL)
```

```
## Loading required package: DBI
```

```
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
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
## Create a database connection
con = dbConnect(MySQL(),
                user='deepAnalytics',
                password='Sqltask1234!',
                dbname='dataanalytics2018',
                host='data-analytics-2018.cbrosir2cswx.us-east-1.rds.amazonaws.com')

dbListTables(con) # List of the 6 table inside the database.
```

```
## [1] "iris"    "yr_2006" "yr_2007" "yr_2008" "yr_2009" "yr_2010"
```

```
# We will use iris as an example.
```

```r
## Lists attributes contained in a table:
dbListFields(con,'iris') # Show 6 attributes
```

```
## [1] "id"            "SepalLengthCm" "SepalWidthCm"  "PetalLengthCm"
## [5] "PetalWidthCm"  "Species"
```

```r
# Still focusing on "iris", we can query the database. We can download all of
# the data or choose the specific attributes we're interested in.

## Use asterisk to specify all attributes for download
irisALL <- dbGetQuery(con, "SELECT * FROM iris")
```

```
## Warning in .local(conn, statement, ...): Unsigned INTEGER in col 0 imported as
## numeric
```

```r
## Use attribute names to specify specific attributes for download
irisSELECT <- dbGetQuery(con, "SELECT SepalLengthCm, SepalWidthCm FROM iris")


#########################

# We start taking all tables from yr_2006 to yr_2010

yr_2006ALL <- dbGetQuery(con, "SELECT * FROM yr_2006")
```

```
## Warning in .local(conn, statement, ...): Unsigned INTEGER in col 0 imported as
## numeric
```

```r
yr_2007ALL <- dbGetQuery(con, "SELECT * FROM yr_2007")
```

```
## Warning in .local(conn, statement, ...): Unsigned INTEGER in col 0 imported as
## numeric
```

```r
yr_2008ALL <- dbGetQuery(con, "SELECT * FROM yr_2008")
```

```
## Warning in .local(conn, statement, ...): Unsigned INTEGER in col 0 imported as
## numeric
```

```r
yr_2009ALL <- dbGetQuery(con, "SELECT * FROM yr_2009")
```

```
## Warning in .local(conn, statement, ...): Unsigned INTEGER in col 0 imported as
## numeric
```

```r
yr_2010ALL <- dbGetQuery(con, "SELECT * FROM yr_2010")
```

```
## Warning in .local(conn, statement, ...): Unsigned INTEGER in col 0 imported as
## numeric
```

```r
# Checking 2006
str(yr_2006ALL)
```

```
## 'data.frame':    21992 obs. of  10 variables:
##  $ id                   : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Date                 : chr  "2006-12-16" "2006-12-16" "2006-12-16" "2006-12-16" ...
##  $ Time                 : chr  "17:24:00" "17:25:00" "17:26:00" "17:27:00" ...
##  $ Global_active_power  : num  4.22 5.36 5.37 5.39 3.67 ...
##  $ Global_reactive_power: num  0.418 0.436 0.498 0.502 0.528 0.522 0.52 0.52 0.51 0.51 ...
##  $ Global_intensity     : num  18.4 23 23 23 15.8 15 15.8 15.8 15.8 15.8 ...
##  $ Voltage              : num  235 234 233 234 236 ...
```

```
##  $ Sub_metering_1       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Sub_metering_2       : num  1 1 2 1 1 2 1 1 1 2 ...
##  $ Sub_metering_3       : num  17 16 17 17 17 17 17 17 17 16 ...
```

```
summary(yr_2006ALL)
```

```
##        id            Date               Time           Global_active_power
##  Min.   :    1   Length:21992       Length:21992       Min.   :0.194
##  1st Qu.: 5499   Class :character   Class :character   1st Qu.:0.496
##  Median :10998   Mode  :character   Mode  :character   Median :1.708
##  Mean   :10998                                         Mean   :1.901
##  3rd Qu.:16496                                         3rd Qu.:2.692
##  Max.   :21996                                         Max.   :9.132
##  Global_reactive_power Global_intensity    Voltage      Sub_metering_1
##  Min.   :0.0000        Min.   : 0.80    Min.   :228.2   Min.   : 0.000
##  1st Qu.:0.0000        1st Qu.: 2.20    1st Qu.:238.8   1st Qu.: 0.000
##  Median :0.1140        Median : 7.20    Median :241.7   Median : 0.000
##  Mean   :0.1314        Mean   : 8.03    Mean   :241.4   Mean   : 1.249
##  3rd Qu.:0.1980        3rd Qu.:11.40    3rd Qu.:244.4   3rd Qu.: 0.000
##  Max.   :0.8000        Max.   :39.40    Max.   :251.7   Max.   :77.000
##  Sub_metering_2   Sub_metering_3
##  Min.   : 0.000   Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.: 0.00
##  Median : 0.000   Median : 0.00
##  Mean   : 2.215   Mean   : 7.41
##  3rd Qu.: 1.000   3rd Qu.:17.00
##  Max.   :74.000   Max.   :20.00
```

```
head(yr_2006ALL)
```

```
##   id       Date     Time Global_active_power Global_reactive_power
## 1  1 2006-12-16 17:24:00              4.216                 0.418
## 2  2 2006-12-16 17:25:00              5.360                 0.436
## 3  3 2006-12-16 17:26:00              5.374                 0.498
## 4  4 2006-12-16 17:27:00              5.388                 0.502
## 5  5 2006-12-16 17:28:00              3.666                 0.528
## 6  6 2006-12-16 17:29:00              3.520                 0.522
##   Global_intensity Voltage Sub_metering_1 Sub_metering_2 Sub_metering_3
## 1             18.4  234.84              0              1             17
## 2             23.0  233.63              0              1             16
## 3             23.0  233.29              0              2             17
## 4             23.0  233.74              0              1             17
## 5             15.8  235.68              0              1             17
## 6             15.0  235.02              0              2             17
```

```
tail(yr_2006ALL)
```

```
##          id       Date     Time Global_active_power Global_reactive_power
## 21987 21991 2006-12-31 23:54:00              2.576                 0.132
## 21988 21992 2006-12-31 23:55:00              2.574                 0.132
## 21989 21993 2006-12-31 23:56:00              2.576                 0.132
## 21990 21994 2006-12-31 23:57:00              2.586                 0.134
## 21991 21995 2006-12-31 23:58:00              2.648                 0.212
## 21992 21996 2006-12-31 23:59:00              2.646                 0.236
##       Global_intensity Voltage Sub_metering_1 Sub_metering_2 Sub_metering_3
## 21987             10.6  241.90              0              0              0
```

```
## 21988             10.6  241.89             0              0              0
## 21989             10.6  242.06             0              0              0
## 21990             10.6  242.61             0              0              0
## 21991             11.0  241.93             0              0              0
## 21992             11.0  241.89             0              0              0
```

```
# We just have 21.992 observations. It starts from 2006-12-16. Just half month.
```

```
# Checking 2007
str(yr_2007ALL)
```

```
## 'data.frame':    521669 obs. of  10 variables:
##  $ id                   : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Date                 : chr  "2007-01-01" "2007-01-01" "2007-01-01" "2007-01-01" ...
##  $ Time                 : chr  "00:00:00" "00:01:00" "00:02:00" "00:03:00" ...
##  $ Global_active_power  : num  2.58 2.55 2.55 2.55 2.55 ...
##  $ Global_reactive_power: num  0.136 0.1 0.1 0.1 0.1 0.1 0.096 0 0 0 ...
##  $ Global_intensity     : num  10.6 10.4 10.4 10.4 10.4 10.4 10.4 10.2 10.2 10.2 ...
##  $ Voltage              : num  242 242 242 242 242 ...
##  $ Sub_metering_1       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Sub_metering_2       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Sub_metering_3       : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
summary(yr_2007ALL)
```

```
##        id              Date               Time           Global_active_power
##  Min.   :     1   Length:521669      Length:521669      Min.   : 0.082
##  1st Qu.:130423   Class :character   Class :character   1st Qu.: 0.278
##  Median :264606   Mode  :character   Mode  :character   Median : 0.504
##  Mean   :263456                                        Mean   : 1.117
##  3rd Qu.:395178                                        3rd Qu.: 1.548
##  Max.   :525600                                        Max.   :10.670
##  Global_reactive_power Global_intensity    Voltage      Sub_metering_1
##  Min.   :0.0000        Min.   : 0.400   Min.   :223.5   Min.   : 0.000
##  1st Qu.:0.0000        1st Qu.: 1.200   1st Qu.:236.9   1st Qu.: 0.000
##  Median :0.1000        Median : 2.400   Median :239.7   Median : 0.000
##  Mean   :0.1174        Mean   : 4.764   Mean   :239.4   Mean   : 1.232
##  3rd Qu.:0.1860        3rd Qu.: 6.400   3rd Qu.:241.8   3rd Qu.: 0.000
##  Max.   :1.1480        Max.   :46.400   Max.   :252.1   Max.   :78.000
##  Sub_metering_2   Sub_metering_3
##  Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.000   Median : 0.000
##  Mean   : 1.638   Mean   : 5.795
##  3rd Qu.: 1.000   3rd Qu.:17.000
##  Max.   :78.000   Max.   :20.000
```

```
head(yr_2007ALL)
```

```
##   id       Date     Time Global_active_power Global_reactive_power
## 1  1 2007-01-01 00:00:00              2.580                 0.136
## 2  2 2007-01-01 00:01:00              2.552                 0.100
## 3  3 2007-01-01 00:02:00              2.550                 0.100
```

```
## 4  4 2007-01-01 00:03:00              2.550                 0.100
## 5  5 2007-01-01 00:04:00              2.554                 0.100
## 6  6 2007-01-01 00:05:00              2.550                 0.100
##   Global_intensity Voltage Sub_metering_1 Sub_metering_2 Sub_metering_3
## 1             10.6  241.97              0              0              0
## 2             10.4  241.75              0              0              0
## 3             10.4  241.64              0              0              0
## 4             10.4  241.71              0              0              0
## 5             10.4  241.98              0              0              0
## 6             10.4  241.83              0              0              0
```

```
tail(yr_2007ALL)
```

```
##            id       Date     Time Global_active_power Global_reactive_power
## 521664 525595 2007-12-31 23:54:00               1.648                 0.102
## 521665 525596 2007-12-31 23:55:00               1.746                 0.204
## 521666 525597 2007-12-31 23:56:00               1.732                 0.210
## 521667 525598 2007-12-31 23:57:00               1.732                 0.210
## 521668 525599 2007-12-31 23:58:00               1.684                 0.144
## 521669 525600 2007-12-31 23:59:00               1.628                 0.072
##        Global_intensity Voltage Sub_metering_1 Sub_metering_2 Sub_metering_3
## 521664              6.8  241.56              0              0             18
## 521665              7.2  242.41              0              0             18
## 521666              7.2  242.42              0              0             18
## 521667              7.2  242.50              0              0             18
## 521668              7.0  242.18              0              0             18
## 521669              6.6  241.79              0              0             18
```

```
# Full year.



# Checking 2008
str(yr_2008ALL)
```

```
## 'data.frame':    526905 obs. of  10 variables:
##  $ id                   : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Date                 : chr  "2008-01-01" "2008-01-01" "2008-01-01" "2008-01-01" ...
##  $ Time                 : chr  "00:00:00" "00:01:00" "00:02:00" "00:03:00" ...
##  $ Global_active_power  : num  1.62 1.63 1.62 1.61 1.61 ...
##  $ Global_reactive_power: num  0.07 0.072 0.072 0.07 0.07 0 0 0 0 0 ...
##  $ Global_intensity     : num  6.6 6.6 6.6 6.6 6.6 6.4 6.4 6.4 6.4 6.4 ...
##  $ Voltage              : num  241 242 242 241 241 ...
##  $ Sub_metering_1       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Sub_metering_2       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Sub_metering_3       : num  18 18 18 18 18 17 18 18 18 18 ...
```

```
summary(yr_2008ALL)
```

```
##        id              Date               Time           Global_active_power
##  Min.   :     1   Length:526905      Length:526905      Min.   : 0.076
##  1st Qu.:131732   Class :character   Class :character   1st Qu.: 0.300
##  Median :263461   Mode  :character   Mode  :character   Median : 0.566
##  Mean   :263474                                         Mean   : 1.072
##  3rd Qu.:395191                                         3rd Qu.: 1.518
##  Max.   :527040                                         Max.   :10.348
```

```
##  Global_reactive_power Global_intensity    Voltage      Sub_metering_1
##  Min.   :0.0000        Min.   : 0.200   Min.   :224.6   Min.   : 0.00
##  1st Qu.:0.0460        1st Qu.: 1.400   1st Qu.:238.9   1st Qu.: 0.00
##  Median :0.0940        Median : 2.400   Median :240.7   Median : 0.00
##  Mean   :0.1171        Mean   : 4.552   Mean   :240.6   Mean   : 1.11
##  3rd Qu.:0.1840        3rd Qu.: 6.400   3rd Qu.:242.5   3rd Qu.: 0.00
##  Max.   :1.3900        Max.   :44.600   Max.   :250.9   Max.   :80.00
##  Sub_metering_2   Sub_metering_3
##  Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.000   Median : 1.000
##  Mean   : 1.256   Mean   : 6.034
##  3rd Qu.: 1.000   3rd Qu.:17.000
##  Max.   :76.000   Max.   :31.000
```

```
head(yr_2008ALL)
```

```
##   id       Date     Time Global_active_power Global_reactive_power
## 1  1 2008-01-01 00:00:00               1.620                 0.070
## 2  2 2008-01-01 00:01:00               1.626                 0.072
## 3  3 2008-01-01 00:02:00               1.622                 0.072
## 4  4 2008-01-01 00:03:00               1.612                 0.070
## 5  5 2008-01-01 00:04:00               1.612                 0.070
## 6  6 2008-01-01 00:05:00               1.546                 0.000
##   Global_intensity Voltage Sub_metering_1 Sub_metering_2 Sub_metering_3
## 1              6.6  241.25              0              0             18
## 2              6.6  241.74              0              0             18
## 3              6.6  241.52              0              0             18
## 4              6.6  240.82              0              0             18
## 5              6.6  240.80              0              0             18
## 6              6.4  240.66              0              0             17
```

```
tail(yr_2008ALL)
```

```
##            id       Date     Time Global_active_power Global_reactive_power
## 526900 527035 2008-12-31 23:54:00               0.484                 0.064
## 526901 527036 2008-12-31 23:55:00               0.484                 0.064
## 526902 527037 2008-12-31 23:56:00               0.482                 0.064
## 526903 527038 2008-12-31 23:57:00               0.482                 0.064
## 526904 527039 2008-12-31 23:58:00               0.480                 0.064
## 526905 527040 2008-12-31 23:59:00               0.482                 0.062
##        Global_intensity Voltage Sub_metering_1 Sub_metering_2 Sub_metering_3
## 526900              2.2  248.15              0              0              0
## 526901              2.2  247.69              0              0              0
## 526902              2.2  247.35              0              0              0
## 526903              2.2  246.99              0              0              0
## 526904              2.2  246.52              0              0              0
## 526905              2.2  246.97              0              0              0
```

```
# Full year.



# Checking 2009
str(yr_2009ALL)
```

```
## 'data.frame':    521320 obs. of  10 variables:
##  $ id                   : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Date                 : chr  "2009-01-01" "2009-01-01" "2009-01-01" "2009-01-01" ...
##  $ Time                 : chr  "00:00:00" "00:01:00" "00:02:00" "00:03:00" ...
##  $ Global_active_power  : num  0.484 0.484 0.482 0.482 0.482 0.57 0.59 0.588 0.586 0.586 ...
##  $ Global_reactive_power: num  0.062 0.062 0.062 0.06 0.062 0 0.078 0.078 0.078 0.078 ...
##  $ Global_intensity     : num  2.2 2.2 2.2 2.2 2.2 2.6 2.6 2.6 2.6 2.6 ...
##  $ Voltage              : num  248 248 248 248 247 ...
##  $ Sub_metering_1       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Sub_metering_2       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Sub_metering_3       : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
summary(yr_2009ALL)
```

```
##        id              Date               Time           Global_active_power
##  Min.   :     1   Length:521320      Length:521320      Min.   : 0.122
##  1st Qu.:130398   Class :character   Class :character   1st Qu.: 0.318
##  Median :264038   Mode  :character   Mode  :character   Median : 0.622
##  Mean   :262890                                        Mean   : 1.079
##  3rd Qu.:395266                                        3rd Qu.: 1.514
##  Max.   :525600                                        Max.   :11.122
##  Global_reactive_power Global_intensity    Voltage      Sub_metering_1
##  Min.   :0.0000        Min.   : 0.400   Min.   :223.2   Min.   : 0.000
##  1st Qu.:0.0520        1st Qu.: 1.400   1st Qu.:240.1   1st Qu.: 0.000
##  Median :0.1060        Median : 2.800   Median :241.9   Median : 0.000
##  Mean   :0.1314        Mean   : 4.555   Mean   :241.9   Mean   : 1.137
##  3rd Qu.:0.2060        3rd Qu.: 6.200   3rd Qu.:243.6   3rd Qu.: 0.000
##  Max.   :1.2400        Max.   :48.400   Max.   :254.2   Max.   :82.000
##  Sub_metering_2   Sub_metering_3
##  Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.000   Median : 1.000
##  Mean   : 1.136   Mean   : 6.823
##  3rd Qu.: 1.000   3rd Qu.:18.000
##  Max.   :77.000   Max.   :31.000
```

```
head(yr_2009ALL)
```

```
##   id       Date     Time Global_active_power Global_reactive_power
## 1  1 2009-01-01 00:00:00              0.484                 0.062
## 2  2 2009-01-01 00:01:00              0.484                 0.062
## 3  3 2009-01-01 00:02:00              0.482                 0.062
## 4  4 2009-01-01 00:03:00              0.482                 0.060
## 5  5 2009-01-01 00:04:00              0.482                 0.062
## 6  6 2009-01-01 00:05:00              0.570                 0.000
##   Global_intensity Voltage Sub_metering_1 Sub_metering_2 Sub_metering_3
## 1              2.2  247.86              0              0              0
## 2              2.2  247.72              0              0              0
## 3              2.2  247.75              0              0              0
## 4              2.2  247.52              0              0              0
## 5              2.2  246.94              0              0              0
## 6              2.6  246.94              0              0              0
```

```
tail(yr_2009ALL)
```

```
##            id       Date     Time Global_active_power Global_reactive_power
```

```
## 521315 525595 2009-12-31 23:54:00                1.704                 0.128
## 521316 525596 2009-12-31 23:55:00                1.746                 0.158
## 521317 525597 2009-12-31 23:56:00                1.786                 0.234
## 521318 525598 2009-12-31 23:57:00                1.784                 0.232
## 521319 525599 2009-12-31 23:58:00                1.792                 0.236
## 521320 525600 2009-12-31 23:59:00                1.792                 0.238
##        Global_intensity Voltage Sub_metering_1 Sub_metering_2 Sub_metering_3
## 521315              7.0  239.43              0              0             18
## 521316              7.2  239.95              0              0             18
## 521317              7.4  240.09              0              0             19
## 521318              7.4  239.99              0              0             18
## 521319              7.4  240.62              0              0             18
## 521320              7.4  240.82              0              0             19
```

```
# Full year.



# Checking 2010
str(yr_2010ALL)
```

```
## 'data.frame':    457394 obs. of  10 variables:
##  $ id                   : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Date                 : chr  "2010-01-01" "2010-01-01" "2010-01-01" "2010-01-01" ...
##  $ Time                 : chr  "00:00:00" "00:01:00" "00:02:00" "00:03:00" ...
##  $ Global_active_power  : num  1.79 1.78 1.78 1.75 1.69 ...
##  $ Global_reactive_power: num  0.236 0.234 0.234 0.186 0.102 0.1 0.1 0.102 0.072 0 ...
##  $ Global_intensity     : num  7.4 7.4 7.4 7.2 7 7 7 7 6.8 6.6 ...
##  $ Voltage              : num  241 240 240 240 240 ...
##  $ Sub_metering_1       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Sub_metering_2       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Sub_metering_3       : num  18 18 19 18 18 19 18 18 19 18 ...
```

```
summary(yr_2010ALL)
```

```
##        id              Date               Time           Global_active_power
##  Min.   :     1   Length:457394      Length:457394      Min.   :0.138
##  1st Qu.:119509   Class :character   Class :character   1st Qu.:0.336
##  Median :233860   Mode  :character   Mode  :character   Median :0.700
##  Mean   :236335                                        Mean   :1.061
##  3rd Qu.:355437                                        3rd Qu.:1.512
##  Max.   :475023                                        Max.   :9.724
##  Global_reactive_power Global_intensity    Voltage      Sub_metering_1
##  Min.   :0.0000        Min.   : 0.600   Min.   :225.3   Min.   : 0.0000
##  1st Qu.:0.0540        1st Qu.: 1.400   1st Qu.:239.8   1st Qu.: 0.0000
##  Median :0.1000        Median : 3.000   Median :241.5   Median : 0.0000
##  Mean   :0.1294        Mean   : 4.478   Mean   :241.5   Mean   : 0.9875
##  3rd Qu.:0.2000        3rd Qu.: 6.200   3rd Qu.:243.2   3rd Qu.: 0.0000
##  Max.   :1.1240        Max.   :43.000   Max.   :253.5   Max.   :88.0000
##  Sub_metering_2   Sub_metering_3
##  Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 1.000
##  Median : 0.000   Median : 1.000
##  Mean   : 1.102   Mean   : 7.244
##  3rd Qu.: 1.000   3rd Qu.:18.000
```

```
##  Max.   :80.000   Max.   :31.000
```

```
head(yr_2010ALL)
```

```
##   id       Date     Time Global_active_power Global_reactive_power
## 1  1 2010-01-01 00:00:00               1.790                 0.236
## 2  2 2010-01-01 00:01:00               1.780                 0.234
## 3  3 2010-01-01 00:02:00               1.780                 0.234
## 4  4 2010-01-01 00:03:00               1.746                 0.186
## 5  5 2010-01-01 00:04:00               1.686                 0.102
## 6  6 2010-01-01 00:05:00               1.686                 0.100
##   Global_intensity Voltage Sub_metering_1 Sub_metering_2 Sub_metering_3
## 1              7.4  240.65              0              0             18
## 2              7.4  240.07              0              0             18
## 3              7.4  240.15              0              0             19
## 4              7.2  240.26              0              0             18
## 5              7.0  240.12              0              0             18
## 6              7.0  240.10              0              0             19
```

```
tail(yr_2010ALL)
```

```
##            id       Date     Time Global_active_power Global_reactive_power
## 457389 475018 2010-11-26 20:57:00               0.946                     0
## 457390 475019 2010-11-26 20:58:00               0.946                     0
## 457391 475020 2010-11-26 20:59:00               0.944                     0
## 457392 475021 2010-11-26 21:00:00               0.938                     0
## 457393 475022 2010-11-26 21:01:00               0.934                     0
## 457394 475023 2010-11-26 21:02:00               0.932                     0
##        Global_intensity Voltage Sub_metering_1 Sub_metering_2 Sub_metering_3
## 457389              4.0  240.33              0              0              0
## 457390              4.0  240.43              0              0              0
## 457391              4.0  240.00              0              0              0
## 457392              3.8  239.82              0              0              0
## 457393              3.8  239.70              0              0              0
## 457394              3.8  239.55              0              0              0
```

```
# Interrupted at 2010-11-26.

## So, Full years: 2007, 2008, 2009. We will combine all this information
## in a single database. Just the full years.

DF<-bind_rows(yr_2007ALL, yr_2008ALL, yr_2009ALL) # Combine dataframes


# Checking DF
str(DF)
```

```
## 'data.frame':    1569894 obs. of  10 variables:
##  $ id                   : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Date                 : chr  "2007-01-01" "2007-01-01" "2007-01-01" "2007-01-01" ...
##  $ Time                 : chr  "00:00:00" "00:01:00" "00:02:00" "00:03:00" ...
##  $ Global_active_power  : num  2.58 2.55 2.55 2.55 2.55 ...
##  $ Global_reactive_power: num  0.136 0.1 0.1 0.1 0.1 0.1 0.096 0 0 0 ...
##  $ Global_intensity     : num  10.6 10.4 10.4 10.4 10.4 10.4 10.4 10.2 10.2 10.2 ...
##  $ Voltage              : num  242 242 242 242 242 ...
##  $ Sub_metering_1       : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ Sub_metering_2       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Sub_metering_3       : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
summary(DF)
```

```
##        id              Date               Time           Global_active_power
##  Min.   :     1   Length:1569894     Length:1569894     Min.   : 0.076
##  1st Qu.:130851   Class :character   Class :character   1st Qu.: 0.300
##  Median :264035   Mode  :character   Mode  :character   Median : 0.566
##  Mean   :263274                                         Mean   : 1.089
##  3rd Qu.:395212                                         3rd Qu.: 1.524
##  Max.   :527040                                         Max.   :11.122
##  Global_reactive_power Global_intensity    Voltage      Sub_metering_1
##  Min.   :0.0000        Min.   : 0.200   Min.   :223.2   Min.   : 0.000
##  1st Qu.:0.0460        1st Qu.: 1.400   1st Qu.:238.7   1st Qu.: 0.000
##  Median :0.1000        Median : 2.600   Median :240.8   Median : 0.000
##  Mean   :0.1219        Mean   : 4.624   Mean   :240.6   Mean   : 1.159
##  3rd Qu.:0.1920        3rd Qu.: 6.400   3rd Qu.:242.8   3rd Qu.: 0.000
##  Max.   :1.3900        Max.   :48.400   Max.   :254.2   Max.   :82.000
##  Sub_metering_2   Sub_metering_3
##  Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.000   Median : 1.000
##  Mean   : 1.343   Mean   : 6.216
##  3rd Qu.: 1.000   3rd Qu.:17.000
##  Max.   :78.000   Max.   :31.000
```

```
head(DF)
```

```
##   id       Date     Time Global_active_power Global_reactive_power
## 1  1 2007-01-01 00:00:00              2.580                 0.136
## 2  2 2007-01-01 00:01:00              2.552                 0.100
## 3  3 2007-01-01 00:02:00              2.550                 0.100
## 4  4 2007-01-01 00:03:00              2.550                 0.100
## 5  5 2007-01-01 00:04:00              2.554                 0.100
## 6  6 2007-01-01 00:05:00              2.550                 0.100
##   Global_intensity Voltage Sub_metering_1 Sub_metering_2 Sub_metering_3
## 1             10.6  241.97              0              0              0
## 2             10.4  241.75              0              0              0
## 3             10.4  241.64              0              0              0
## 4             10.4  241.71              0              0              0
## 5             10.4  241.98              0              0              0
## 6             10.4  241.83              0              0              0
```

```
tail(DF)
```

```
##             id       Date     Time Global_active_power Global_reactive_power
## 1569889 525595 2009-12-31 23:54:00               1.704                 0.128
## 1569890 525596 2009-12-31 23:55:00               1.746                 0.158
## 1569891 525597 2009-12-31 23:56:00               1.786                 0.234
## 1569892 525598 2009-12-31 23:57:00               1.784                 0.232
## 1569893 525599 2009-12-31 23:58:00               1.792                 0.236
## 1569894 525600 2009-12-31 23:59:00               1.792                 0.238
##         Global_intensity Voltage Sub_metering_1 Sub_metering_2 Sub_metering_3
## 1569889              7.0  239.43              0              0             18
## 1569890              7.2  239.95              0              0             18
```

```
## 1569891              7.4  240.09              0              0             19
## 1569892              7.4  239.99              0              0             18
## 1569893              7.4  240.62              0              0             18
## 1569894              7.4  240.82              0              0             19
```

```
# They are in correct order.

## Now we will combine Date and Time to convert them to the correct format

## Combine Date and Time attribute values in a new attribute column
DF<-cbind(DF,paste(DF$Date,DF$Time), stringsAsFactors=FALSE)

## Give the new attribute in the 11th column a header name

colnames(DF)[11] <-"DateTime"

## Move the DateTime attribute within the dataset.
DF <- DF[,c(ncol(DF), 1:(ncol(DF)-1))]
head(DF) # Now the last column DateTime is the first one.
```

```
##              DateTime id       Date     Time Global_active_power
## 1 2007-01-01 00:00:00  1 2007-01-01 00:00:00               2.580
## 2 2007-01-01 00:01:00  2 2007-01-01 00:01:00               2.552
## 3 2007-01-01 00:02:00  3 2007-01-01 00:02:00               2.550
## 4 2007-01-01 00:03:00  4 2007-01-01 00:03:00               2.550
## 5 2007-01-01 00:04:00  5 2007-01-01 00:04:00               2.554
## 6 2007-01-01 00:05:00  6 2007-01-01 00:05:00               2.550
##   Global_reactive_power Global_intensity Voltage Sub_metering_1 Sub_metering_2
## 1                 0.136             10.6  241.97              0              0
## 2                 0.100             10.4  241.75              0              0
## 3                 0.100             10.4  241.64              0              0
## 4                 0.100             10.4  241.71              0              0
## 5                 0.100             10.4  241.98              0              0
## 6                 0.100             10.4  241.83              0              0
##   Sub_metering_3
## 1              0
## 2              0
## 3              0
## 4              0
## 5              0
## 6              0
```

```
#You will now want to convert the new DateTime attribute to a DateTime data type
# called POSIXct. After converting to POSIXct we will add the time zone to
# prevent warning messages. The data description suggests that the data is from
# France.

## Convert DateTime from character to POSIXct
DF$DateTime <- as.POSIXct(DF$DateTime, "%Y/%m/%d %H:%M:%S")
```

```
## Warning in strptime(xx, f, tz = tz): unknown timezone '%Y/%m/%d %H:%M:%S'

## Warning in as.POSIXct.POSIXlt(x): unknown timezone '%Y/%m/%d %H:%M:%S'

## Warning in strptime(x, f, tz = tz): unknown timezone '%Y/%m/%d %H:%M:%S'

## Warning in as.POSIXct.POSIXlt(as.POSIXlt(x, tz, ...), tz, ...): unknown timezone
```

```
## '%Y/%m/%d %H:%M:%S'
```

```
## Add the time zone
attr(DF$DateTime, "tzone") <- "Europe/Paris"

## Inspect the data types
str(DF)
```

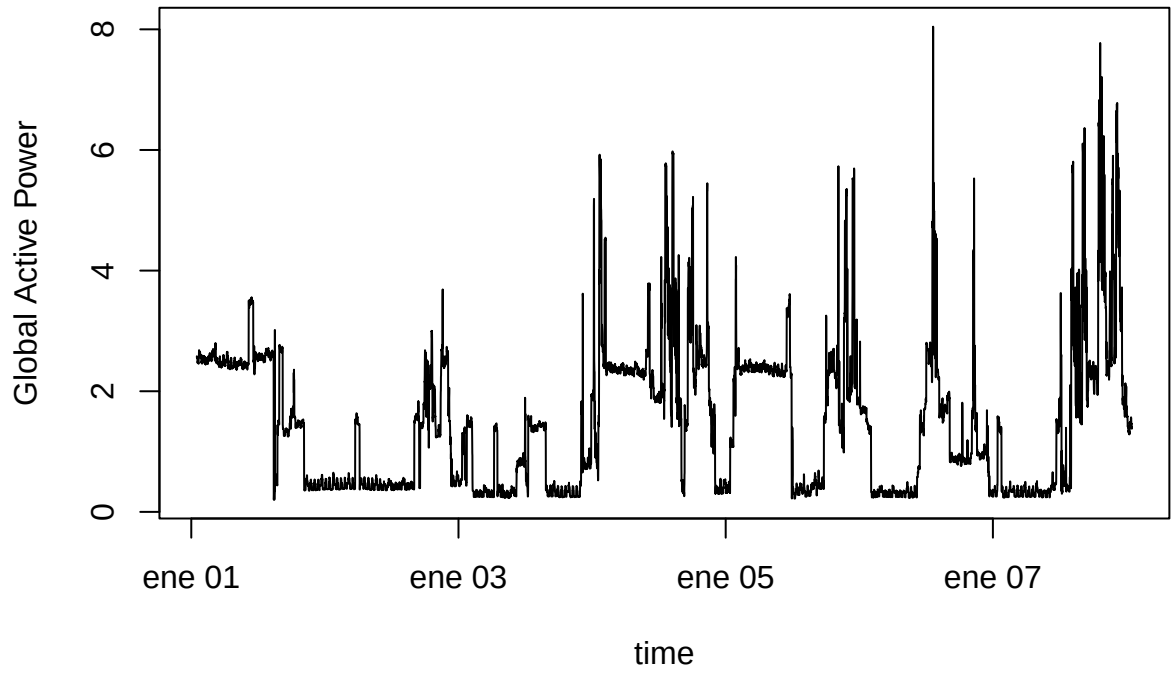
```
## 'data.frame':    1569894 obs. of  11 variables:
##  $ DateTime             : POSIXct, format: "2007-01-01 01:00:00" "2007-01-01 01:01:00" ...
##  $ id                   : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Date                 : chr  "2007-01-01" "2007-01-01" "2007-01-01" "2007-01-01" ...
##  $ Time                 : chr  "00:00:00" "00:01:00" "00:02:00" "00:03:00" ...
##  $ Global_active_power  : num  2.58 2.55 2.55 2.55 2.55 ...
##  $ Global_reactive_power: num  0.136 0.1 0.1 0.1 0.1 0.1 0.096 0 0 0 ...
##  $ Global_intensity     : num  10.6 10.4 10.4 10.4 10.4 10.4 10.4 10.2 10.2 10.2 ...
##  $ Voltage              : num  242 242 242 242 242 ...
##  $ Sub_metering_1       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Sub_metering_2       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Sub_metering_3       : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
# Extract "Year" information from DateTime using the Lubridate "year" function
# and create an attribute for year

## Create "year" attribute with lubridate

DF$year <- year(DF$DateTime)

# Now there is a new column with the year.
# Like "year", Libridate also has functions to create attributes for quarter,
# month, week, weekday, day, hour and minute.

DF$month <- month(DF$DateTime)
DF$day <- day(DF$DateTime)
DF$minute <- minute(DF$DateTime)
DF$hour <- hour(DF$DateTime)

# Remark: apparently, the hour denotes the hour where it belongs. So, the first
# 59 minutes belong the "hour 1".

summary(DF)
```

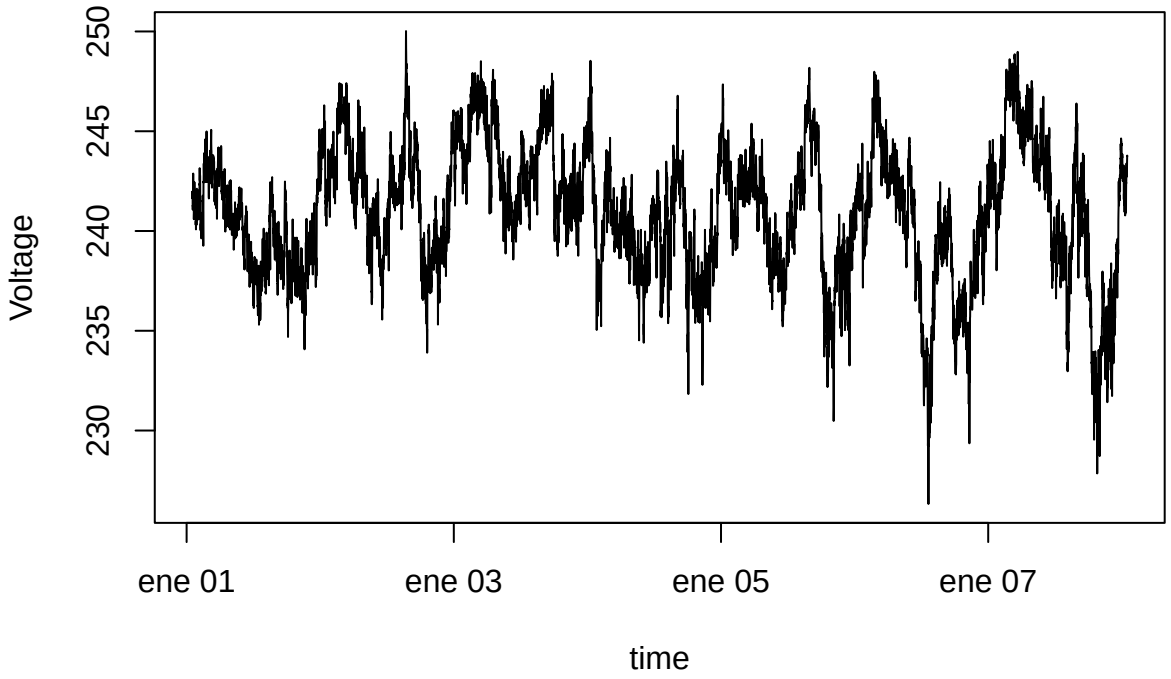
```
##     DateTime                          id             Date
##  Min.   :2007-01-01 01:00:00   Min.   :     1   Length:1569894
##  1st Qu.:2007-10-03 08:39:15   1st Qu.:130851   Class :character
##  Median :2008-07-01 22:05:30   Median :264035   Mode  :character
##  Mean   :2008-07-02 03:54:14   Mean   :263274
##  3rd Qu.:2009-03-31 14:32:45   3rd Qu.:395212
##  Max.   :2010-01-01 00:59:00   Max.   :527040
##      Time           Global_active_power Global_reactive_power Global_intensity
##  Length:1569894     Min.   : 0.076      Min.   :0.0000        Min.   : 0.200
##  Class :character   1st Qu.: 0.300      1st Qu.:0.0460        1st Qu.: 1.400
##  Mode  :character   Median : 0.566      Median :0.1000        Median : 2.600
##                     Mean   : 1.089      Mean   :0.1219        Mean   : 4.624
##                     3rd Qu.: 1.524      3rd Qu.:0.1920        3rd Qu.: 6.400
##                     Max.   :11.122      Max.   :1.3900        Max.   :48.400
```

```
##     Voltage      Sub_metering_1   Sub_metering_2   Sub_metering_3
##  Min.   :223.2   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:238.7   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
##  Median :240.8   Median : 0.000   Median : 0.000   Median : 1.000
##  Mean   :240.6   Mean   : 1.159   Mean   : 1.343   Mean   : 6.216
##  3rd Qu.:242.8   3rd Qu.: 0.000   3rd Qu.: 1.000   3rd Qu.:17.000
##  Max.   :254.2   Max.   :82.000   Max.   :78.000   Max.   :31.000
##       year          month             day            minute           hour
##  Min.   :2007   Min.   : 1.000   Min.   : 1.00   Min.   : 0.00   Min.   : 0.0
##  1st Qu.:2007   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:14.25   1st Qu.: 5.0
##  Median :2008   Median : 7.000   Median :16.00   Median :30.00   Median :12.0
##  Mean   :2008   Mean   : 6.529   Mean   :15.71   Mean   :29.50   Mean   :11.5
##  3rd Qu.:2009   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:44.00   3rd Qu.:18.0
##  Max.   :2010   Max.   :12.000   Max.   :31.00   Max.   :59.00   Max.   :23.0
```

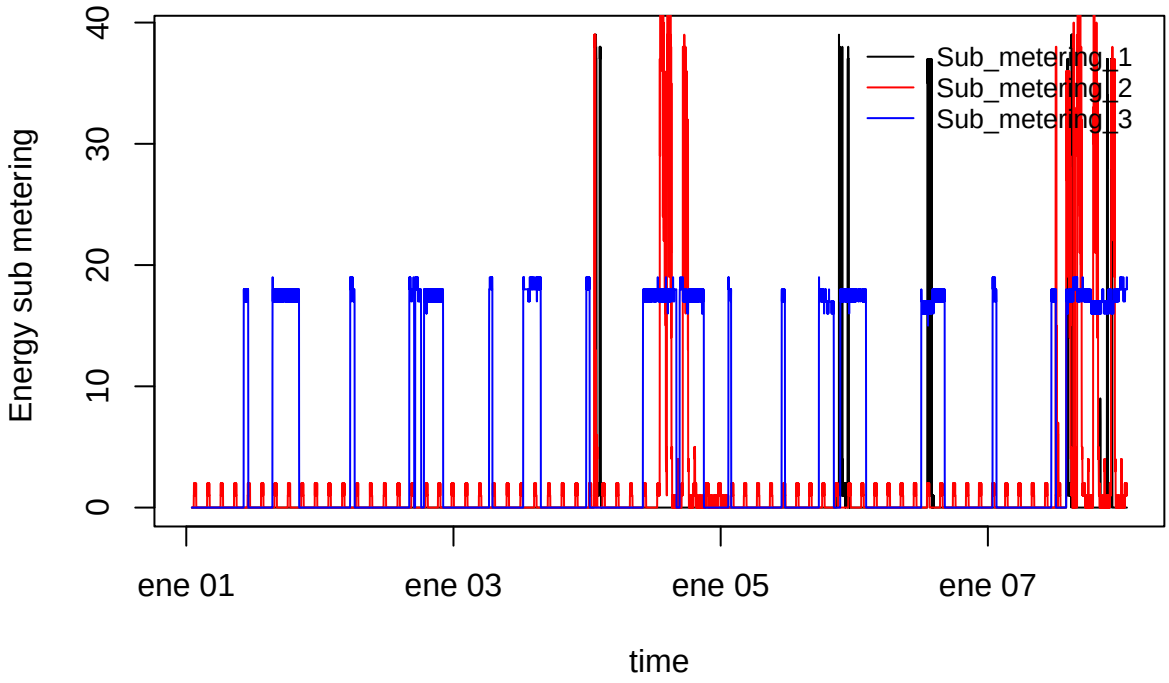
```
## A priori analysis of the descriptives:

# Global active mean: 1.089 kilowatt.
# Global reactive mean: 0.1219 kilowatt
# Global intensity mean: 4.624 amperes.
# Voltage mean: 240.6 volts.
# sub metering 1 (kitchen): 1.159 watt/hour.
# sub metering 2 (laundry room): 1.343 watt/hour.
# sub metering 3 (water-heater and air-conditioner): 6.216 watt/hour.

# in the sub meterings all 3 minimums are 0, but the maximums are 82, 78 and 31,
# respectively. It is interesting that sub metering 3 is clearly the one using
# more power in mean (more than a x4 wrt sub metering 2), but the maximum is 31
# is much lower than maximums in sub metering 1 and 2. By the descriptives, we
# may expect pretty "non-smooth" distributions in all three cases, but
# specifically in 1 and 2. In those cases we can expect high pikes,
# since in those cases we are dealing with  heavy on demand power-using tools
# like ovens , washing machines, etc.

# However, it is somehow suspicious that sub metering 2 has a median of 0.
# Supposedly, refrigerator should be on sub metering 2, and it is a pretty
# consuming appliance that should always be on.

# We may confirm some insights through some extra EDA. We'll do it on a random
# week.

df<- filter(DF, Date >= "2007-01-01" & Date <= "2007-01-07") # Take a week

# 1st plot: global active power/date
plot(df$DateTime, df$Global_active_power, type = "l", main = NULL, xlab = "time", ylab = "Global Active
```




```r
# 2nd plot voltage/date
plot(df$DateTime, df$Voltage, type = "l", main = NULL, xlab = "time", ylab = "Voltage")
```




```r
# 3rd plot Energy Sub meterings/date
plot(df$DateTime, df$Sub_metering_1, type = "l", main = NULL, xlab = "time", ylab = "Energy sub meterin
lines(df$DateTime, df$Sub_metering_2, type = "l", col = "red", main = NULL, xlab = "time", ylab = "")
lines(df$DateTime, df$Sub_metering_3, type = "l", col = "blue", main = NULL, xlab = "time", ylab = "")
legend("topright", legend = c("Sub_metering_1","Sub_metering_2","Sub_metering_3"),
       col=c("black", "red", "blue"), lty=1, cex=0.8, box.lty=0, inset = 0.02)
```


Energy sub metering
0
10
20
30
40
ene 01
ene 03
ene 05
ene 07
time
Sub_metering_1
Sub_metering_2
Sub_metering_3