R Notebook

Code

Install and Load Libraries

Hide

```
#install packages
install.packages("RMySQL")
```

```
Error in install.packages : Updating loaded packages
```

Hide

```
#If there's trouble with installing RMySQL, try this:
#install.packages('RMySQL', dependencies=TRUE, repos='http://cran.rstudio.com/')
#chooseCRANmirror() I tried chooseCRANmirror() with selection 65. It somehow works.
```

Hide

```
install.packages("dplyr")
```

```
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.6/dplyr_0.8.4.tgz'
Content type 'application/x-gzip' length 6846395 bytes (6.5 MB)
==================================================
downloaded 6.5 MB


The downloaded binary packages are in
    /var/folders/hm/2md7sccd0479bw81zsh0yyq80000gn/T//RtmpuAP3mN/downloaded_packages
```

Hide

```
install.packages("ggplot2")
```

```
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.6/ggplot2_3.2.1.tg
z'
Content type 'application/x-gzip' length 3973186 bytes (3.8 MB)
==================================================
downloaded 3.8 MB


The downloaded binary packages are in
    /var/folders/hm/2md7sccd0479bw81zsh0yyq80000gn/T//RtmpuAP3mN/downloaded_packages
```

Hide

```
install.packages("tidyr")
```

```
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.6/tidyr_1.0.2.tgz'
Content type 'application/x-gzip' length 1020461 bytes (996 KB)
==================================================
downloaded 996 KB
```

```

The downloaded binary packages are in
    /var/folders/hm/2md7sccd0479bw81zsh0yyq80000gn/T//RtmpuAP3mN/downloaded_packages
```

Hide

```
install.packages("lubridate")
```

```
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.6/lubridate_1.7.4.t
gz'
Content type 'application/x-gzip' length 1512972 bytes (1.4 MB)
==================================================
downloaded 1.4 MB
```

```

The downloaded binary packages are in
    /var/folders/hm/2md7sccd0479bw81zsh0yyq80000gn/T//RtmpuAP3mN/downloaded_packages
```

Hide

```
# load libraries
library(RMySQL)
library(dplyr)
```

```
Registered S3 method overwritten by 'dplyr':
  method           from
  print.rowwise_df

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
```

Hide

```
library(ggplot2)
library(tidyr)
library(lubridate)
```

```

Attaching package: ‘lubridate’

The following object is masked from ‘package:base’:

    date
```

Hide

```
library(scales)
```

Hide

```
# Set specific options of for libraries
## Only use scientific notation for values greather than set amount
options(scipen=10000000)
```

Hide

```
#confirm libraries
(.packages())
```

```
 [1] "lubridate" "tidyr"     "dplyr"     "RMySQL"    "DBI"       "stats"     "graphics"
"grDevices" "utils"
[10] "datasets"  "methods"   "base"
```

Connect to Database and Obtain Data

Hide

```
# Create a database connection
con = dbConnect(MySQL(), user='deepAnalytics', password='Sqltask1234!', dbname='dataanal
ytics2018', host='data-analytics-2018.cbrosir2cswx.us-east-1.rds.amazonaws.com')
```

Hide

```
#summary of connection
summary(con)
```

```
<MySQLConnection:0,1>
  User:   deepAnalytics
  Host:   data-analytics-2018.cbrosir2cswx.us-east-1.rds.amazonaws.com
  Dbname: dataanalytics2018
  Connection type: data-analytics-2018.cbrosir2cswx.us-east-1.rds.amazonaws.com via TCP/
IP

Results:
```

Hide

```
dbGetInfo(con)
```

```
$host
[1] "data-analytics-2018.cbrosir2cswx.us-east-1.rds.amazonaws.com"

$user
[1] "deepAnalytics"

$dbname
[1] "dataanalytics2018"

$conType
[1] "data-analytics-2018.cbrosir2cswx.us-east-1.rds.amazonaws.com via TCP/IP"

$serverVersion
[1] "5.6.10"

$protocolVersion
[1] 10

$threadId
[1] 100314

$rsId
list()
```

Hide

```
# List the tables contained in the database.
my_tables <- dbListTables(con)
my_tables
```

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

Hide

```
# there are 6 tables: "iris"    "yr_2006" "yr_2007" "yr_2008" "yr_2009" "yr_2010"
```

Hide

```
# Lists attributes contained in a table
list_db_fields_custom_function<- function (x) {dbListFields(con,x)}
lapply(my_tables,list_db_fields_custom_function)
```

```
[[1]]
[1] "id"            "SepalLengthCm" "SepalWidthCm"  "PetalLengthCm" "PetalWidthCm"  "Spe
cies"

[[2]]
 [1] "id"                   "Date"                 "Time"                 "Global_act
ive_power"
 [5] "Global_reactive_power" "Global_intensity"     "Voltage"              "Sub_meteri
ng_1"
 [9] "Sub_metering_2"        "Sub_metering_3"

[[3]]
 [1] "id"                   "Date"                 "Time"                 "Global_act
ive_power"
 [5] "Global_reactive_power" "Global_intensity"     "Voltage"              "Sub_meteri
ng_1"
 [9] "Sub_metering_2"        "Sub_metering_3"

[[4]]
 [1] "id"                   "Date"                 "Time"                 "Global_act
ive_power"
 [5] "Global_reactive_power" "Global_intensity"     "Voltage"              "Sub_meteri
ng_1"
 [9] "Sub_metering_2"        "Sub_metering_3"

[[5]]
 [1] "id"                   "Date"                 "Time"                 "Global_act
ive_power"
 [5] "Global_reactive_power" "Global_intensity"     "Voltage"              "Sub_meteri
ng_1"
 [9] "Sub_metering_2"        "Sub_metering_3"

[[6]]
 [1] "id"                   "Date"                 "Time"                 "Global_act
ive_power"
 [5] "Global_reactive_power" "Global_intensity"     "Voltage"              "Sub_meteri
ng_1"
 [9] "Sub_metering_2"        "Sub_metering_3"
```

Hide

```
# tables for the years 2006 -2010 have the same attributes. Column names are the same.
```

Hide

```
# We are only using Date, Time and Submeters for our analysis.

yr_2006SELECT <- dbGetQuery(con, "SELECT Date, Time, Sub_metering_1, Sub_metering_2, Sub
_metering_3 FROM yr_2006")
yr_2007SELECT <- dbGetQuery(con, "SELECT Date, Time, Sub_metering_1, Sub_metering_2, Sub
_metering_3 FROM yr_2007")
yr_2008SELECT <- dbGetQuery(con, "SELECT Date, Time, Sub_metering_1, Sub_metering_2, Sub
_metering_3 FROM yr_2008")
install.packages("RMySQL")
```

```
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.6/RMySQL_0.10.19.tg
z'
Content type 'application/x-gzip' length 1760084 bytes (1.7 MB)
==================================================
downloaded 1.7 MB
```

```


The downloaded binary packages are in
    /var/folders/hm/2md7sccd0479bw81zsh0yyq80000gn/T//RtmpnXV7tH/downloaded_packages
```

Hide

```
yr_2009SELECT <- dbGetQuery(con, "SELECT Date, Time, Sub_metering_1, Sub_metering_2, Sub
_metering_3 FROM yr_2009")
install.packages("dplyr")
```

```
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.6/dplyr_0.8.4.tgz'
Content type 'application/x-gzip' length 6846395 bytes (6.5 MB)
==================================================
downloaded 6.5 MB
```

```


The downloaded binary packages are in
    /var/folders/hm/2md7sccd0479bw81zsh0yyq80000gn/T//RtmpnXV7tH/downloaded_packages
```

Hide

```
install.packages("tidyr")
```

```
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.6/tidyr_1.0.2.tgz'
Content type 'application/x-gzip' length 1020461 bytes (996 KB)
==================================================
downloaded 996 KB
```

```


The downloaded binary packages are in
    /var/folders/hm/2md7sccd0479bw81zsh0yyq80000gn/T//RtmpnXV7tH/downloaded_packages
```

Hide

```
install.packages("lubridate")
```

```
Error in install.packages : Updating loaded packages
```

Hide

```
yr_2010SELECT <- dbGetQuery(con, "SELECT Date, Time, Sub_metering_1, Sub_metering_2, Sub
_metering_3 FROM yr_2010")
```

Explore and prepare data

Note: MySQL tables are read into R as data.frames, but without coercing character or logical data into factors. Similarly while exporting data.frames, factors are exported as character vectors. Integer columns are usually imported as R integer vectors, except for cases such as BIGINT or UNSIGNED INTEGER which are coerced to R's double precision vectors to avoid truncation (currently R's integers are signed 32-bit quantities). Time variables are imported/exported as character data, so you need to convert these to your favorite date/time representation.

Investigate Data

Hide

```
# Function to explore tables. Prints out structure, summary, head and tail of data for e
very table.
investigateDF <- function(df) {list(str(df), summary(df),head(df),tail(df))}
```

Hide

```
# Investigates tables from 2006 to 2010
investigateDF(yr_2006SELECT)
```

```
'data.frame':   21992 obs. of  5 variables:
 $ 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" ...
 $ 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 ...
[[1]]
NULL

[[2]]
     Date               Time           Sub_metering_1   Sub_metering_2   Sub_metering_3
 Length:21992       Length:21992       Min.   : 0.000   Min.   : 0.000   Min.   : 0.00
 Class :character   Class :character   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.00
 Mode  :character   Mode  :character   Median : 0.000   Median : 0.000   Median : 0.00
                                       Mean   : 1.249   Mean   : 2.215   Mean   : 7.41
                                       3rd Qu.: 0.000   3rd Qu.: 1.000   3rd Qu.:17.00
                                       Max.   :77.000   Max.   :74.000   Max.   :20.00

[[3]]
```

	Date	Time	Sub_metering_1	Sub_metering_2	Sub_metering_3
	<chr>	<chr>	<dbl>	<dbl>	<dbl>
1	2006-12-16	17:24:00	0	1	17
2	2006-12-16	17:25:00	0	1	16
3	2006-12-16	17:26:00	0	2	17
4	2006-12-16	17:27:00	0	1	17
5	2006-12-16	17:28:00	0	1	17
6	2006-12-16	17:29:00	0	2	17

6 rows

```
[[4]]
```

	Date	Time	Sub_metering_1	Sub_metering_2	Sub_metering_3
	<chr>	<chr>	<dbl>	<dbl>	<dbl>
21987	2006-12-31	23:54:00	0	0	0
21988	2006-12-31	23:55:00	0	0	0
21989	2006-12-31	23:56:00	0	0	0
21990	2006-12-31	23:57:00	0	0	0
21991	2006-12-31	23:58:00	0	0	0
21992	2006-12-31	23:59:00	0	0	0

6 rows

Hide

```
investigateDF(yr_2007SELECT)
```

```
'data.frame':   521669 obs. of  5 variables:
 $ 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" ...
 $ 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 ...
[[1]]
NULL


[[2]]
     Date               Time           Sub_metering_1   Sub_metering_2   Sub_metering_3
 Length:521669      Length:521669      Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 Class :character   Class :character   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
 Mode  :character   Mode  :character   Median : 0.000   Median : 0.000   Median : 0.000
                                       Mean   : 1.232   Mean   : 1.638   Mean   : 5.795
                                       3rd Qu.: 0.000   3rd Qu.: 1.000   3rd Qu.:17.000
                                       Max.   :78.000   Max.   :78.000   Max.   :20.000

[[3]]
```

Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_metering_3 <dbl>
1 2007-01-01	00:00:00	0	0	0
2 2007-01-01	00:01:00	0	0	0
3 2007-01-01	00:02:00	0	0	0
4 2007-01-01	00:03:00	0	0	0
5 2007-01-01	00:04:00	0	0	0
6 2007-01-01	00:05:00	0	0	0

6 rows

```
[[4]]
```

	Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_metering_3 <dbl>
521664	2007-12-31	23:54:00	0	0	18
521665	2007-12-31	23:55:00	0	0	18

	Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_metering_3 <dbl>
521666	2007-12-31	23:56:00	0	0	18
521667	2007-12-31	23:57:00	0	0	18
521668	2007-12-31	23:58:00	0	0	18
521669	2007-12-31	23:59:00	0	0	18

6 rows

Hide

```
investigateDF(yr_2008SELECT)
```

```
'data.frame':   526905 obs. of  5 variables:
 $ 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" ...
 $ 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 ...
[[1]]
NULL

[[2]]
     Date               Time           Sub_metering_1  Sub_metering_2   Sub_metering_3
 Length:526905      Length:526905      Min.   : 0.00   Min.   : 0.000   Min.   : 0.000
 Class :character   Class :character   1st Qu.: 0.00   1st Qu.: 0.000   1st Qu.: 0.000
 Mode  :character   Mode  :character   Median : 0.00   Median : 0.000   Median : 1.000
                                       Mean   : 1.11   Mean   : 1.256   Mean   : 6.034
                                       3rd Qu.: 0.00   3rd Qu.: 1.000   3rd Qu.:17.000
                                       Max.   :80.00   Max.   :76.000   Max.   :31.000

[[3]]
```

	Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_metering_3 <dbl>
1	2008-01-01	00:00:00	0	0	18
2	2008-01-01	00:01:00	0	0	18
3	2008-01-01	00:02:00	0	0	18
4	2008-01-01	00:03:00	0	0	18
5	2008-01-01	00:04:00	0	0	18
6	2008-01-01	00:05:00	0	0	17

6 rows

```
[[4]]
```

	Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_metering_3 <dbl>
526900	2008-12-31	23:54:00	0	0	0
526901	2008-12-31	23:55:00	0	0	0
526902	2008-12-31	23:56:00	0	0	0
526903	2008-12-31	23:57:00	0	0	0
526904	2008-12-31	23:58:00	0	0	0
526905	2008-12-31	23:59:00	0	0	0

6 rows

Hide

```
investigateDF(yr_2009SELECT)
```

```
'data.frame':   521320 obs. of  5 variables:
 $ 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" ...
 $ 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 ...
[[1]]
NULL

[[2]]
     Date               Time           Sub_metering_1   Sub_metering_2   Sub_metering_3
 Length:521320      Length:521320      Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 Class :character   Class :character   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
 Mode  :character   Mode  :character   Median : 0.000   Median : 0.000   Median : 1.000
                                       Mean   : 1.137   Mean   : 1.136   Mean   : 6.823
                                       3rd Qu.: 0.000   3rd Qu.: 1.000   3rd Qu.:18.000
                                       Max.   :82.000   Max.   :77.000   Max.   :31.000

[[3]]
```

	Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_metering_3 <dbl>
1	2009-01-01	00:00:00	0	0	0
2	2009-01-01	00:01:00	0	0	0
3	2009-01-01	00:02:00	0	0	0
4	2009-01-01	00:03:00	0	0	0

	Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_metering_3 <dbl>
5	2009-01-01	00:04:00	0	0	0
6	2009-01-01	00:05:00	0	0	0

6 rows

```
[[4]]
```

	Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_metering_3 <dbl>
521315	2009-12-31	23:54:00	0	0	18
521316	2009-12-31	23:55:00	0	0	18
521317	2009-12-31	23:56:00	0	0	19
521318	2009-12-31	23:57:00	0	0	18
521319	2009-12-31	23:58:00	0	0	18
521320	2009-12-31	23:59:00	0	0	19

6 rows

Hide

```
investigateDF(yr_2010SELECT)
```

```
'data.frame':   457394 obs. of  5 variables:
 $ 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" ...
 $ 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 ...
[[1]]
NULL

[[2]]
     Date               Time           Sub_metering_1    Sub_metering_2   Sub_metering_3
 Length:457394      Length:457394      Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000

 Class :character   Class :character   1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 1.000
 Mode  :character   Mode  :character   Median : 0.0000   Median : 0.000   Median : 1.000
                                       Mean   : 0.9875   Mean   : 1.102   Mean   : 7.244
                                       3rd Qu.: 0.0000   3rd Qu.: 1.000   3rd Qu.:18.000
                                       Max.   :88.0000   Max.   :80.000   Max.   :31.000

[[3]]
```

Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_metering_3 <dbl>
1 2010-01-01	00:00:00	0	0	18
2 2010-01-01	00:01:00	0	0	18
3 2010-01-01	00:02:00	0	0	19
4 2010-01-01	00:03:00	0	0	18
5 2010-01-01	00:04:00	0	0	18
6 2010-01-01	00:05:00	0	0	19

6 rows

```
[[4]]
```

	Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_metering_3 <dbl>
457389	2010-11-26	20:57:00	0	0	0
457390	2010-11-26	20:58:00	0	0	0
457391	2010-11-26	20:59:00	0	0	0
457392	2010-11-26	21:00:00	0	0	0
457393	2010-11-26	21:01:00	0	0	0
457394	2010-11-26	21:02:00	0	0	0

6 rows

```
NA
```

2006 data starts at December 16, 2006. Therefore, we will not include it in the dataset because we only want tables with a whole year's worth of data in it. 2010 starts at January 1 and ends at November 26. It is missing December's data. We will also not include it.We will only use 2007 to 2009.

Hide

```
#Combine tables into one dataframe (using dplyr)
df2007_2009 <- bind_rows(yr_2007SELECT,yr_2008SELECT,yr_2009SELECT)
install.packages("lubridate")
```

```
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.6/lubridate_1.7.4.t
gz'
Content type 'application/x-gzip' length 1512972 bytes (1.4 MB)
==================================================
downloaded 1.4 MB
```

```
The downloaded binary packages are in
    /var/folders/hm/2md7sccd0479bw81zsh0yyq80000gn/T//RtmpnXV7tH/downloaded_packages
```

Hide

```
investigateDF(df2007_2009)
```

```
'data.frame':   1569894 obs. of  5 variables:
 $ 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" ...
 $ 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 ...
[[1]]
NULL

[[2]]
     Date               Time           Sub_metering_1   Sub_metering_2   Sub_metering_3
 Length:1569894     Length:1569894     Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 Class :character   Class :character   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
 Mode  :character   Mode  :character   Median : 0.000   Median : 0.000   Median : 1.000
                                       Mean   : 1.159   Mean   : 1.343   Mean   : 6.216
                                       3rd Qu.: 0.000   3rd Qu.: 1.000   3rd Qu.:17.000
                                       Max.   :82.000   Max.   :78.000   Max.   :31.000

[[3]]
```

	Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_metering_3 <dbl>
1	2007-01-01	00:00:00	0	0	0
2	2007-01-01	00:01:00	0	0	0
3	2007-01-01	00:02:00	0	0	0
4	2007-01-01	00:03:00	0	0	0
5	2007-01-01	00:04:00	0	0	0
6	2007-01-01	00:05:00	0	0	0

6 rows

```
[[4]]
```

	Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_metering_3 <dbl>
1569889	2009-12-31	23:54:00	0	0	18

	Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_metering_3 <dbl>
1569890	2009-12-31	23:55:00	0	0	18
1569891	2009-12-31	23:56:00	0	0	19
1569892	2009-12-31	23:57:00	0	0	18
1569893	2009-12-31	23:58:00	0	0	18
1569894	2009-12-31	23:59:00	0	0	19

6 rows

```
NA
```

Create DateTime Objects

Hide

```
# Combine Date and Time attribute values in a new attribute column
df2007_2009 <-cbind(df2007_2009,paste(df2007_2009$Date,df2007_2009$Time), stringsAsFacto
rs=FALSE)
```

Hide

```
# Give the new attribute in the 6th column a header name
colnames(df2007_2009)[6] <-"DateTime"
head(df2007_2009)
```

	Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_metering_3 <dbl>	DateTime <chr>
1	2007-01-01	00:00:00	0	0	0	2007-01-01 00:00:00
2	2007-01-01	00:01:00	0	0	0	2007-01-01 00:01:00
3	2007-01-01	00:02:00	0	0	0	2007-01-01 00:02:00
4	2007-01-01	00:03:00	0	0	0	2007-01-01 00:03:00
5	2007-01-01	00:04:00	0	0	0	2007-01-01 00:04:00
6	2007-01-01	00:05:00	0	0	0	2007-01-01 00:05:00

6 rows

Hide

```
# Move the DateTime attribute within the dataset
df2007_2009 <- df2007_2009[,c(ncol(df2007_2009), 1:(ncol(df2007_2009)-1))]
head(df2007_2009)
```

	DateTime <chr>	Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_metering <db
1	2007-01-01 00:00:00	2007-01-01	00:00:00	0	0	
2	2007-01-01 00:01:00	2007-01-01	00:01:00	0	0	
3	2007-01-01 00:02:00	2007-01-01	00:02:00	0	0	
4	2007-01-01 00:03:00	2007-01-01	00:03:00	0	0	
5	2007-01-01 00:04:00	2007-01-01	00:04:00	0	0	
6	2007-01-01 00:05:00	2007-01-01	00:05:00	0	0	

6 rows

Hide

```
# Convert DateTime from character to POSIXct
df2007_2009$DateTime <- as.POSIXct(df2007_2009$DateTime, "%Y/%m/%d %H:%M:%S")
```

```
unknown timezone '%Y/%m/%d %H:%M:%S'unknown timezone '%Y/%m/%d %H:%M:%S'unknown timezone
'%Y/%m/%d %H:%M:%S'unknown timezone '%Y/%m/%d %H:%M:%S'
```

Hide

```
attr(df2007_2009$DateTime, "tzone") <- "Europe/Paris"

#Verify
str(df2007_2009)
```

```
'data.frame':   1569894 obs. of  6 variables:
 $ DateTime      : POSIXct, format: "2007-01-01 01:00:00" "2007-01-01 01:01:00" "2007-01
-01 01:02:00" "2007-01-01 01:03:00" ...
 $ 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" ...
 $ 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 ...
```

Hide

```
head(df2007_2009)
```

	DateTime <S3: POSIXct>	Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_metering <db
1	2007-01-01 01:00:00	2007-01-01	00:00:00	0	0	
2	2007-01-01 01:01:00	2007-01-01	00:01:00	0	0	
3	2007-01-01 01:02:00	2007-01-01	00:02:00	0	0	

	DateTime	Date	Time	Sub_metering_1	Sub_metering_2	Sub_metering
	<S3: POSIXct>	<chr>	<chr>	<dbl>	<dbl>	<db
4	2007-01-01 01:03:00	2007-01-01	00:03:00	0	0	
5	2007-01-01 01:04:00	2007-01-01	00:04:00	0	0	
6	2007-01-01 01:05:00	2007-01-01	00:05:00	0	0	

6 rows

Hide

```
# Create "year, quarter, month, week, weekday, day, dateTZ(different than original date
 [chr string] with time zone applied], hour, and minute attributes
df2007_2009$year <- year(df2007_2009$DateTime)
df2007_2009$quarter <- quarter(df2007_2009$DateTime)
df2007_2009$month <- month(df2007_2009$DateTime)
df2007_2009$week <- week(df2007_2009$DateTime)
df2007_2009$weekday <- weekdays(df2007_2009$DateTime)
df2007_2009$day <- day(df2007_2009$DateTime)
df2007_2009$dateTZ <- date(df2007_2009$DateTime)
df2007_2009$hour <- hour(df2007_2009$DateTime)
df2007_2009$minute <- minute(df2007_2009$DateTime)
```

Hide

```
# verify new attributes
head(df2007_2009)
```

	DateTime	Date	Time	Sub_metering_1	Sub_metering_2	Sub_metering
	<S3: POSIXct>	<chr>	<chr>	<dbl>	<dbl>	<db
1	2007-01-01 01:00:00	2007-01-01	00:00:00	0	0	
2	2007-01-01 01:01:00	2007-01-01	00:01:00	0	0	
3	2007-01-01 01:02:00	2007-01-01	00:02:00	0	0	
4	2007-01-01 01:03:00	2007-01-01	00:03:00	0	0	
5	2007-01-01 01:04:00	2007-01-01	00:04:00	0	0	
6	2007-01-01 01:05:00	2007-01-01	00:05:00	0	0	

6 rows | 1-8 of 15 columns

Hide

```
tail(df2007_2009)
```

	DateTime	Date	Time	Sub_metering_1	Sub_metering_2	Sub_n
	<S3: POSIXct>	<chr>	<chr>	<dbl>	<dbl>	
1569889	2010-01-01 00:54:00	2009-12-31	23:54:00	0	0	

	DateTime <S3: POSIXct>	Date <chr>	Time <chr>	Sub_metering_1 <dbl>	Sub_metering_2 <dbl>	Sub_n
1569890	2010-01-01 00:55:00	2009-12-31	23:55:00	0	0	
1569891	2010-01-01 00:56:00	2009-12-31	23:56:00	0	0	
1569892	2010-01-01 00:57:00	2009-12-31	23:57:00	0	0	
1569893	2010-01-01 00:58:00	2009-12-31	23:58:00	0	0	
1569894	2010-01-01 00:59:00	2009-12-31	23:59:00	0	0	

6 rows | 1-8 of 15 columns

Missing Values

Hide

```
sum(is.na(df2007_2009))
# no missing values
```

Data Documentation

Source: http://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption# (http://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption#)

Abstract: 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.

Attribute Information:

sub_metering_1: energy sub-metering No. 1 (in). 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.

Summary Statistics

Hide

```
# view mean, mode, standard deviation, quartiles, and characterization of distribution.
summary(df2007_2009)
```

```
    DateTime                       Date               Time           Sub_metering_1   Su
b_metering_2
 Min.   :2007-01-01 01:00:00   Length:1569894     Length:1569894     Min.   : 0.000   Mi
n.   : 0.000
 1st Qu.:2007-10-03 08:39:15   Class :character   Class :character   1st Qu.: 0.000   1s
t Qu.: 0.000
 Median :2008-07-01 22:05:30   Mode  :character   Mode  :character   Median : 0.000   Me
dian : 0.000
 Mean   :2008-07-02 03:54:14                                        Mean   : 1.159   Me
an   : 1.343
 3rd Qu.:2009-03-31 14:32:45                                        3rd Qu.: 0.000   3r
d Qu.: 1.000
 Max.   :2010-01-01 00:59:00                                        Max.   :82.000   Ma
x.   :78.000
 Sub_metering_3        year         quarter         month             week         weekd
ay
 Min.   : 0.000   Min.   :2007   Min.   :1.00   Min.   : 1.000   Min.   : 1.00   Length:
1569894
 1st Qu.: 0.000   1st Qu.:2007   1st Qu.:2.00   1st Qu.: 4.000   1st Qu.:13.00   Class :
character
 Median : 1.000   Median :2008   Median :3.00   Median : 7.000   Median :27.00   Mode  :
character
 Mean   : 6.216   Mean   :2008   Mean   :2.51   Mean   : 6.529   Mean   :26.62
 3rd Qu.:17.000   3rd Qu.:2009   3rd Qu.:4.00   3rd Qu.:10.000   3rd Qu.:40.00
 Max.   :31.000   Max.   :2010   Max.   :4.00   Max.   :12.000   Max.   :53.00
      day            dateTZ                hour          minute
 Min.   : 1.00   Min.   :2007-01-01   Min.   : 0.0   Min.   : 0.00
 1st Qu.: 8.00   1st Qu.:2007-10-03   1st Qu.: 5.0   1st Qu.:14.25
 Median :16.00   Median :2008-07-01   Median :12.0   Median :30.00
 Mean   :15.71   Mean   :2008-07-01   Mean   :11.5   Mean   :29.50
 3rd Qu.:23.00   3rd Qu.:2009-03-31   3rd Qu.:18.0   3rd Qu.:44.00
 Max.   :31.00   Max.   :2010-01-01   Max.   :23.0   Max.   :59.00
```

Hide

```
str(df2007_2009)
```

```
'data.frame':   1569894 obs. of  15 variables:
 $ DateTime      : POSIXct, format: "2007-01-01 01:00:00" "2007-01-01 01:01:00" "2007-01
-01 01:02:00" "2007-01-01 01:03:00" ...
 $ 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" ...
 $ 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 ...
 $ year          : num  2007 2007 2007 2007 2007 ...
 $ quarter       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ month         : num  1 1 1 1 1 1 1 1 1 1 ...
 $ week          : num  1 1 1 1 1 1 1 1 1 1 ...
 $ weekday       : chr  "Monday" "Monday" "Monday" "Monday" ...
 $ day           : int  1 1 1 1 1 1 1 1 1 1 ...
 $ dateTZ        : Date, format: "2007-01-01" "2007-01-01" "2007-01-01" "2007-01-01" ...
 $ hour          : int  1 1 1 1 1 1 1 1 1 1 ...
 $ minute        : int  0 1 2 3 4 5 6 7 8 9 ...
```

Notes:

DateTime Minimum: 2007-01-01 01:00:00
Median:2008-07-01 22:05:30 Maximum:2010-01-01 00:59:00

We didn't include 2010 in our dataset but, due to applying differing time zones, we have January 1, 2010 in our dataset. We may beed to remove this.

Total Energy Consumption for Submeters

Submeter_3 has the highest mean of 6.216 followed by submeter_2 with 1.343, and submeter_2 with 1.159.

Hide

```
# Sum of Energy for each Submeter
sum(df2007_2009$Sub_metering_1)
sum(df2007_2009$Sub_metering_2)
sum(df2007_2009$Sub_metering_3)

#[1] 1819989
#[2] 2108410
#[3] 9758843
```

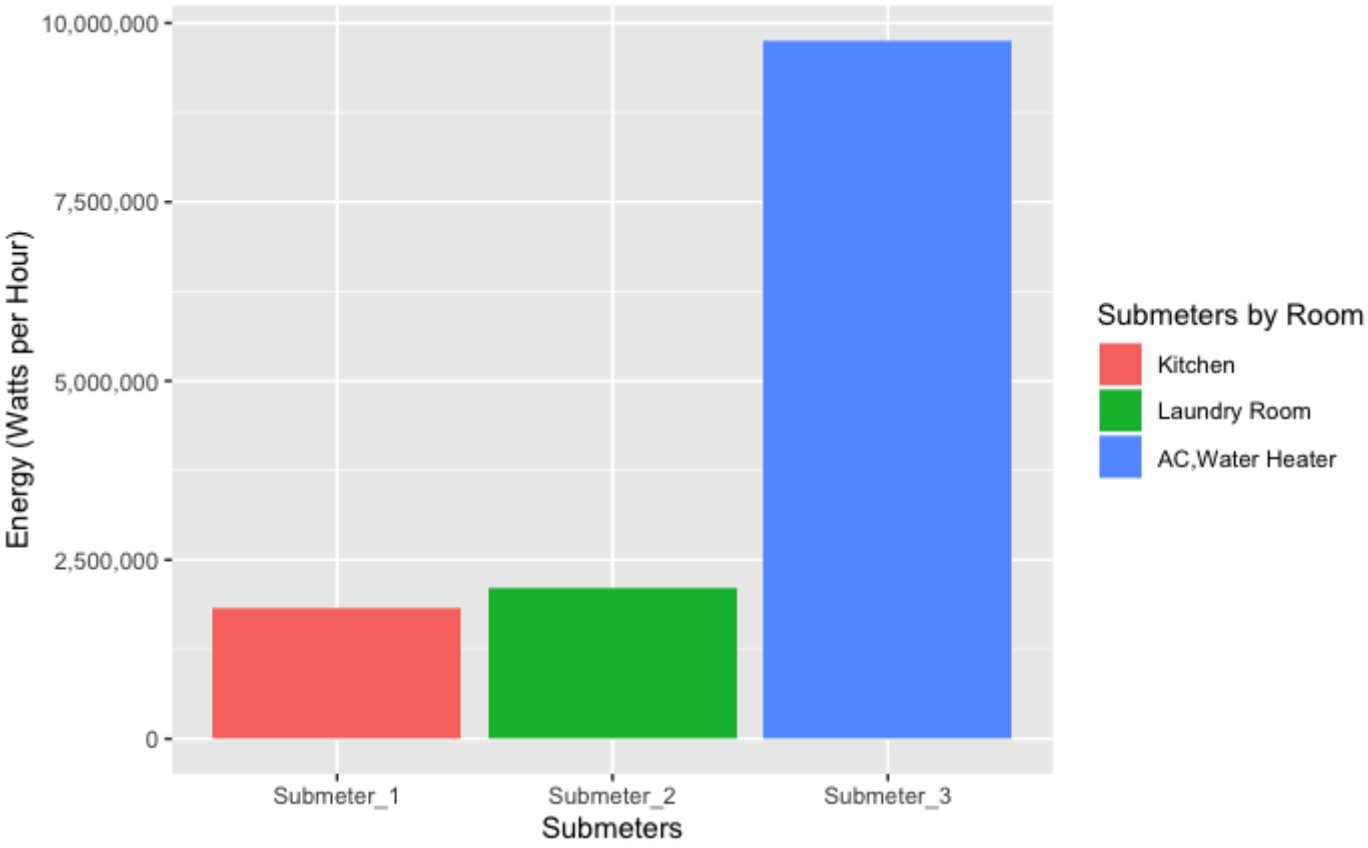
Submeter_3, which is for the water heater and A/C, uses the most power as compared to all the other submeters. Submeter_2 and submeter_1 have similar means and similar total usage. ## Visualize Total Energy Consumption for each Submeter

Hide

```
# Create dataframe of sums
sum_of_submeters <- data.frame(Submeter_1 = sum(df2007_2009$Sub_metering_1), Submeter_2
 = sum(df2007_2009$Sub_metering_2), Submeter_3 = sum(df2007_2009$Sub_metering_3))
```

Hide

```
# plot data
ggplot(data = sum_of_submeters_long, aes(x = Submeters, y = Total_Energy_Usage, fill = S
ubmeters)) +
    geom_col()+
   scale_y_continuous(label=comma)+
   ylab("Energy (Watts per Hour)")+
   scale_fill_discrete(name = "Submeters by Room", labels = c("Kitchen", "Laundry Room",
"AC,Water Heater"))
```



Visualize Energy Consumption Over Time

Hide

```
plot(df2007_2009$Sub_metering_1)
```

Most observations are between 0 and 40, with occassional high usage. There are two primary breaks in the data. Where there is minimum usage. Perhaps, this is vacation tim

Hide

```
plot(df2007_2009$DateTime,df2007_2009$Sub_metering_1, main = "Submeter 1")
plot(df2007_2009$DateTime,df2007_2009$Sub_metering_2, main = "Submeter 2")
plot(df2007_2009$DateTime,df2007_2009$Sub_metering_3, main = "Submeter 3")
```

Hide

```
ggplot(df2007_2009, aes(x = DateTime, y = Sub_metering_1)) +
  geom_line(aes(group = quarter))
ggplot(df2007_2009, aes(x = DateTime, y = Sub_metering_2)) +
  geom_line(aes(group = quarter))
ggplot(df2007_2009, aes(x = DateTime, y = Sub_metering_3)) +
  geom_line(aes(group = quarter))
```

#submeter_1

Hide

```
summary(df2007_2009$Sub_metering_1)
```

Hide

```
ggplot(data=df2007_2009, aes(x=DateTime, y=Sub_metering_1)) + geom_line()+ylab("Energy")
+ xlab("Time")
```

Hide

```
ggplot(data=df2007_2009, aes(x=DateTime, y=Sub_metering_1)) + geom_point()+ylab("Energy"
)+ xlab("Time")
ggplot(data=df2007_2009, aes(x=DateTime, y=Sub_metering_2)) + geom_point()+ylab("Energy"
)+ xlab("Time")

ggplot(data=df2007_2009, aes(x=DateTime, y=Sub_metering_3)) + geom_point()+ylab("Energy"
)+ xlab("Time")
```

Hide

```
plot(df2007_2009$DateTime, df2007_2009$Sub_metering_3, ylab="Energy", xlab="Time")
```

Hide

```
ggplot(data=df2007_2009, aes(x=DateTime, y=Sub_metering_3)) + geom_line()+ylab("Energy")
+ xlab("Time")
```

Hide

```
ggplot(data=df2007_2009, aes(x=Sub_metering_1)) + geom_freqpoly()+ylab("Frequency")+ xla
b("Energy")
ggplot(data=df2007_2009, aes(x=Sub_metering_2)) + geom_freqpoly()+ylab("Frequency")+ xla
b("Energy")
ggplot(data=df2007_2009, aes(x=Sub_metering_3)) + geom_freqpoly()+ylab("Frequency")+ xla
b("Energy")
```

Hide

```
frequency_submeter1 <- data.frame(table(df2007_2009$Sub_metering_1))
names(frequency_submeter1)[names(frequency_submeter1) == "Var1"] <- "Energy"
names(frequency_submeter1)[names(frequency_submeter1) == "Freq"] <- "Frequency"

arrange(frequency_submeter1, -frequency_submeter1$Frequency)
```

Hide

```
df_1_submeter = gather(sum_of_submeters, key = "Submeters") %>%
    group_by(Submeters) %>%
    summarize(Total_Energy_Usage = sum(value, na.rm = TRUE))

ggplot(data = sum_of_submeters_long, aes(x = Submeters, y = Total_Energy_Usage, fill = S
ubmeters)) +
    geom_col()+
   scale_y_continuous(label=comma)+
   #ggtitle("Total Energy Usage over 3 Years")+
   ylab("Energy (Watts per Hour)")+
   scale_fill_discrete(name = "Submeters by Room", labels = c("Kitchen", "Laundry Room",
"AC,Water Heater"))
```