Preprocessing

```r
library(readr)
library(magrittr)
library(lubridate)
library(pander)
library(data.table)
training <- read_csv("../data/training.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   date = col_datetime(format = ""),
##   WeekStatus = col_character(),
##   Day_of_week = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```r
test <- read_csv("../data/testing.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   date = col_datetime(format = ""),
##   WeekStatus = col_character(),
##   Day_of_week = col_character()
## )
## See spec(...) for full column specifications.
```

Data description

```
date time year-month-day hour:minute:second
Appliances, energy use in Wh
lights, energy use of light fixtures in the house in Wh
T1, Temperature in kitchen area, in Celsius
RH_1, Humidity in kitchen area, in %
T2, Temperature in living room area, in Celsius
RH_2, Humidity in living room area, in %
T3, Temperature in laundry room area
RH_3, Humidity in laundry room area, in %
T4, Temperature in office room, in Celsius
RH_4, Humidity in office room, in %
T5, Temperature in bathroom, in Celsius
RH_5, Humidity in bathroom, in %
T6, Temperature outside the building (north side), in Celsius
RH_6, Humidity outside the building (north side), in %
T7, Temperature in ironing room , in Celsius
RH_7, Humidity in ironing room, in %
T8, Temperature in teenager room 2, in Celsius
RH_8, Humidity in teenager room 2, in %
T9, Temperature in parents room, in Celsius
RH_9, Humidity in parents room, in %
```

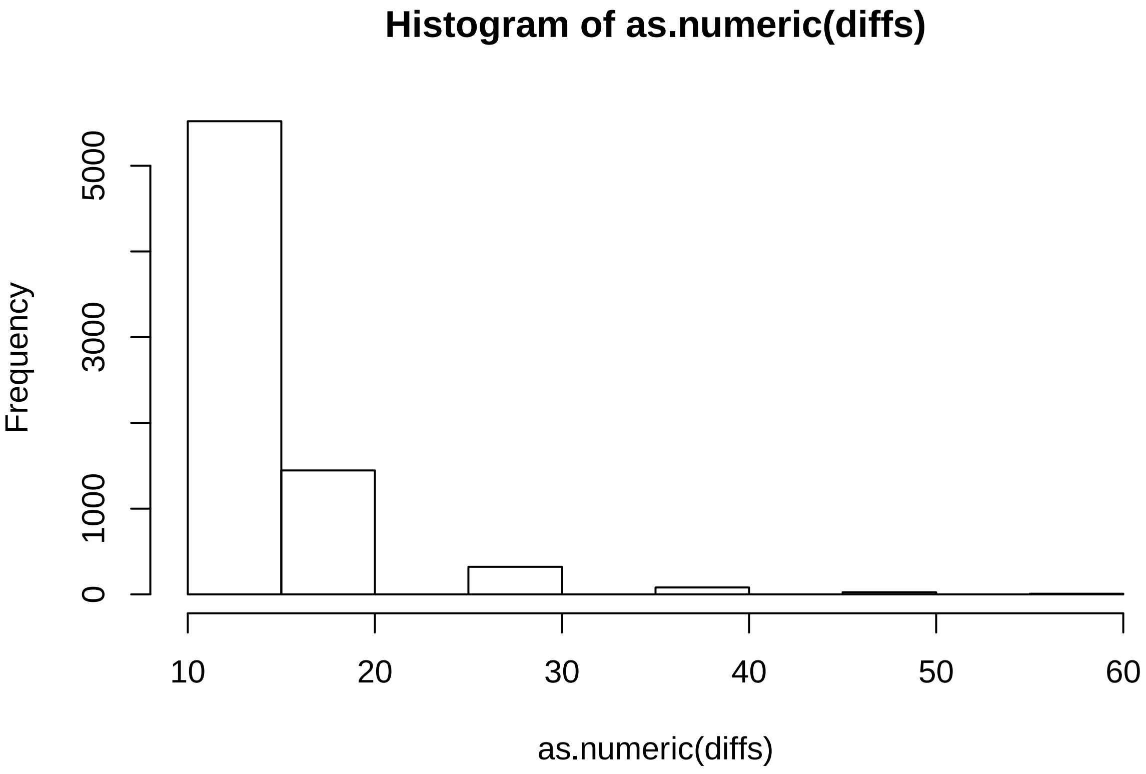
```
To, Temperature outside (from Chièvres weather station), in Celsius
Pressure (from Chièvres weather station), in mm Hg
RH_out, Humidity outside (from Chièvres weather station), in %
Windspeed (from Chièvres weather station), in m/s
Visibility (from Chièvres weather station), in km
Tdewpoint (from Chièvres weather station), °C
rv1, Random variable 1, nondimensional
rv2, Rnadom variable 2, nondimensional
NSM, time in seconds
WeekStatus, weekday vs weekend
Day_of_week, obvious
```

Exploring

```
timediff <- function(xs){
    odds <- xs[c(TRUE,FALSE)]
    evens <- xs[!c(TRUE,FALSE)]
    out <- difftime(evens, odds)
    rev(rev(out)[-1])
}
diffs <- timediff(training[[1]])
```

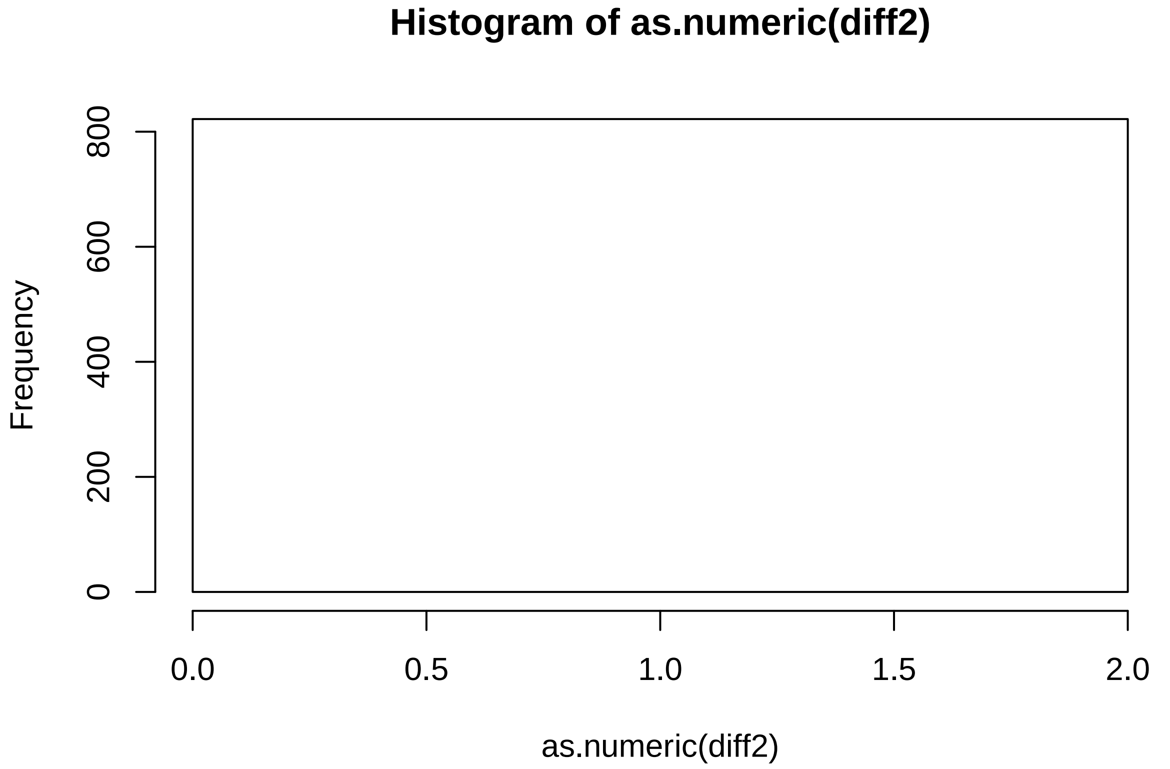
```
## Warning in unclass(time1) - unclass(time2): longer object length is not a
## multiple of shorter object length
```

```
ht <- function(xs){
    c(head(xs),
      tail(xs))
}
hist(as.numeric(diffs))
```



Looks like we have some sort of issue with uneven sampling. Lets average this dataset every 30 minutes to try to fix

```
smooth_time <- function(df, by = "30 min"){
    df[[1]]  <- floor_date(df[[1]], unit = by)
    df
}
smoothed <- smooth_time(training)
mn_or_val <- function(xs) {
    if(is.character(xs)){
        tail(names(sort(table( xs ))),1)
    }
    else {
        mean(xs)
    }
}
# requires there to be a column named date
collapse_dates <- function(df, unit){
    smoothed <- smooth_time(df, by = unit)
     setDT(smoothed)[, lapply(.SD, mn_or_val), by = .(date)]
}
train_clean <- collapse_dates(training,"2 hours")
diff2 <- timediff(train_clean[[1]])
hist(as.numeric(diff2))
```



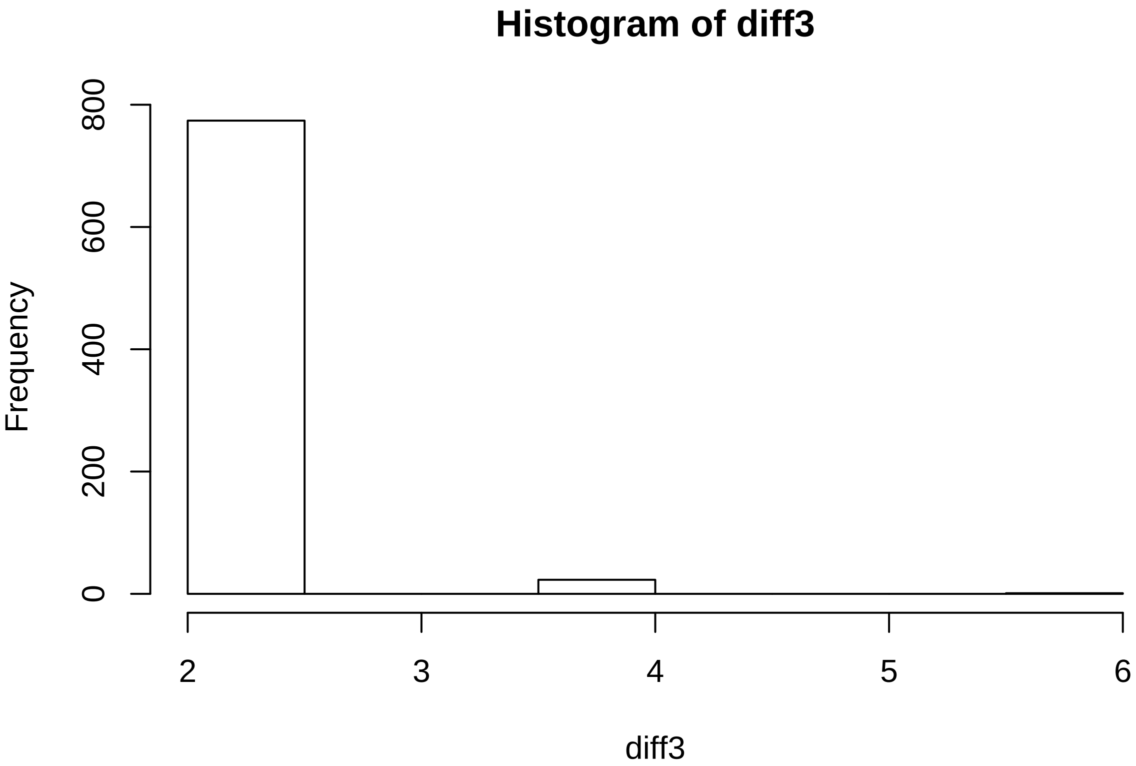
```
pander(table(as.numeric(diff2)))
```

2
822

```
test_clean <- collapse_dates(test, "2 hours")
diff3 <- as.numeric(timediff(test_clean[[1]]))
```

```
## Warning in unclass(time1) - unclass(time2): longer object length is not a
## multiple of shorter object length
```

```
hist(diff3)
```



```
pander(table(diff3))
```

2	4	6
774	23	1

I think we can live with this