My Data

Ilkka Kiistala

7 Apr 2015

Baana Biker data

I'm going to combine biker data collected at Baana and Helsinki weather data.

Load biker data

Load biker data from Helsinki Region Infoshare:

```
url <- 'http://www.hel.fi/hel2/tietokeskus/data/helsinki/ksv/Baanan_pyorailijamaarat.xlsx'
download.file(url, destfile="bikers.xlsx")
```

```
library(xlsx)
```

```
## Loading required package: rJava
## Loading required package: methods
## Loading required package: xlsxjars
```

```
bikers <- read.xlsx("bikers.xlsx", 1)
```

Check biker data

```
head(bikers)
```

```
##   viikko    päiväys päivä NA. viikkosumma NA..1 NA..2 NA..3 NA..4 NA..5
## 1      1 2012-12-31    ma  78          NA    NA    NA    NA    NA    NA
## 2     NA 2013-01-01    ti  44          NA    NA    NA    NA    NA    NA
## 3     NA 2013-01-02    ke 192          NA    NA    NA    NA    NA    NA
## 4     NA 2013-01-03    to 303          NA    NA    NA    NA    NA    NA
## 5     NA 2013-01-04    pe 312          NA    NA    NA    NA    NA    NA
## 6     NA 2013-01-05    la 118          NA    NA    NA    NA    NA    NA
##   NA..6 NA..7 NA..8 NA..9 NA..10
## 1    NA    NA    NA    NA     NA
## 2    NA    NA    NA    NA     NA
## 3    NA    NA    NA    NA     NA
## 4    NA    NA    NA    NA     NA
## 5    NA    NA    NA    NA     NA
## 6    NA    NA    NA    NA     NA
```

```
nrow(bikers)
```

```
## [1] 518
```

```
str(bikers)
```

```
## 'data.frame':    518 obs. of  15 variables:
##  $ viikko     : num  1 NA NA NA NA NA NA 2 NA NA ...
##  $ päiväys    : Date, format: "2012-12-31" "2013-01-01" ...
##  $ päivä      : Factor w/ 7 levels "ke","la","ma",..: 3 6 1 7 4 2 5 3 6 1 ...
##  $ NA.        : num  78 44 192 303 312 118 105 362 382 348 ...
##  $ viikkosumma: num  NA NA NA NA NA ...
##  $ NA..1      : logi  NA NA NA NA NA NA ...
##  $ NA..2      : logi  NA NA NA NA NA NA ...
##  $ NA..3      : logi  NA NA NA NA NA NA ...
##  $ NA..4      : logi  NA NA NA NA NA NA ...
##  $ NA..5      : logi  NA NA NA NA NA NA ...
##  $ NA..6      : logi  NA NA NA NA NA NA ...
##  $ NA..7      : logi  NA NA NA NA NA NA ...
##  $ NA..8      : logi  NA NA NA NA NA NA ...
##  $ NA..9      : logi  NA NA NA NA NA NA ...
##  $ NA..10     : logi  NA NA NA NA NA NA ...
```

Cleanup

We only need date and number of bikers.

```
bikers <- bikers[, c(2,4)]
```

Let's rename the columns.

```
head(bikers)
```

```
##      päiväys NA.
## 1 2012-12-31  78
## 2 2013-01-01  44
## 3 2013-01-02 192
## 4 2013-01-03 303
## 5 2013-01-04 312
## 6 2013-01-05 118
```

```
names(bikers) <- c("date", "bikers")
head(bikers)
```

```
##         date bikers
## 1 2012-12-31     78
## 2 2013-01-01     44
## 3 2013-01-02    192
## 4 2013-01-03    303
## 5 2013-01-04    312
## 6 2013-01-05    118
```

Load weather data

Helsinki daily temperature observations

The Finnish Meteorological Institute (FMI) is a research and service agency under the Ministry of Transport and Communications.

Observations are accessible via FMI Open Data WFS service. Quering the service requires registration, which provides user with an API key.

```
# Example API call:
http://data.fmi.fi/fmi-apikey/insert-your-apikey-here/wfs?request=getFeature&storedquery_id=
```

See their Open Data Manual: http://en.ilmatieteenlaitos.fi/open-data-manual-fmi-wfs-services

Pre-editing the data

After fetching the data as XML, it was parsed into following form:

```
2013-01-01 tday 2.8
2013-01-02 tday 2.3
2013-01-03 tday 1.0
2013-01-04 tday 1.6
2013-01-05 tday -2.5
2013-01-06 tday -4.6
2013-01-07 tday -4.7
2013-01-08 tday -0.1
```

Importing the temperature data into R

```
tem <- read.csv("helsinki-temperatures.tsv", sep=" ", header=FALSE, stringsAsFactors=FALSE)
summary(tem)
```

```
##       V1                 V2                  V3
##  Length:810         Length:810         Min.   :-15.700
##  Class :character   Class :character   1st Qu.:  1.300
##  Mode  :character   Mode  :character   Median :  5.650
##                                        Mean   :  7.228
##                                        3rd Qu.: 15.175
##                                        Max.   : 25.700
```

Summary reveals some repeating days, so to be sure, we need to check that their data match. But before that, let's remove the columns with 'tday' values and name our columns.

```
names(tem)
```

```
## [1] "V1" "V2" "V3"
```

```
tem[,2] <- NULL
names(tem)
```

```
## [1] "V1" "V3"
```

```
names(tem) <- c("date", "tday")
head(tem)
```

```
##         date tday
## 1 2012-12-31  2.2
## 2 2013-01-01  2.8
## 3 2013-01-02  2.3
## 4 2013-01-03  1.0
## 5 2013-01-04  1.6
## 6 2013-01-05 -2.5
```

Now it's easier to refer to the columns. Let's check those repeating dates.

```
date_count <- as.data.frame(table(tem$date), stringsAsFactors=FALSE)
# subset(date_count, Freq > 1)
# multidate <- as.vector(subset(date_count, Freq > 1, select="Var1"))
multidate <- subset(date_count, Freq > 1, select="Var1")[,1]

# tem[tem$date %in% as.vector(multidate),]
head( tem[tem$date %in% multidate,] )
```

```
##          date tday
## 32 2013-01-31  1.8
## 33 2013-01-31  1.8
## 64 2013-03-03 -6.2
## 65 2013-03-03 -6.2
## 93 2013-03-31 -0.9
## 94 2013-03-31 -0.9
```

No anomalies in the pairs. It is safe to remove duplicates.

```
nrow(tem)
```

```
## [1] 810
```

```
tail(tem)
```

```
##           date tday
## 805 2015-02-11  3.8
## 806 2015-02-12  2.3
## 807 2015-02-13  1.8
## 808 2015-02-14  0.9
## 809 2015-02-15 -4.3
## 810 2015-02-16 -2.0
```

```
tem <- tem[!duplicated(tem),]
nrow(tem)
```

```
## [1] 771
```

```
tail(tem)
```

```
##           date tday
## 805 2015-02-11  3.8
## 806 2015-02-12  2.3
## 807 2015-02-13  1.8
## 808 2015-02-14  0.9
## 809 2015-02-15 -4.3
## 810 2015-02-16 -2.0
```

```
head( tem[tem$date %in% multidate,] )
```

```
##          date tday
## 32 2013-01-31  1.8
## 64 2013-03-03 -6.2
## 93 2013-03-31 -0.9
## 95 2013-04-01 -0.3
## 97 2013-04-02  1.7
## 99 2013-04-03  1.4
```

```
# This recalculates rownames
rownames(tem) <- NULL
tail(tem)
```

```
##           date tday
## 766 2015-02-11  3.8
## 767 2015-02-12  2.3
## 768 2015-02-13  1.8
## 769 2015-02-14  0.9
## 770 2015-02-15 -4.3
## 771 2015-02-16 -2.0
```

Merging the two data frames

```
nrow(bikers)
```

```
## [1] 518
```

```
nrow(tem)
```

```
## [1] 771
```

Let's find out the latest dates we have in the two data sets.

```r
max(bikers$date)
```

```
## [1] "2014-06-01"
```

```r
max(tem$date)
```

```
## [1] "2015-02-16"
```

```r
# the earlier date is usually in the biker data, but let's play it safe
mutually_latest_date <- min(max(bikers$date), max(tem$date))
mutually_latest_date
```

```
## [1] "2014-06-01"
```

```r
# check that row count matches
nrow(subset(bikers, date <= mutually_latest_date))
```

```
## [1] 518
```

```r
nrow(subset(tem, date <= mutually_latest_date))
```

```
## [1] 518
```

```r
bikers <- subset(bikers, date <= mutually_latest_date)
tem <- subset(tem, date <= mutually_latest_date)
```

Now their row counts match, so let's merge them. First, some values to check after merge:

```r
tail(bikers);tail(tem)
```

```
##           date bikers
## 513 2014-05-27   3084
## 514 2014-05-28   2488
## 515 2014-05-29    952
## 516 2014-05-30   2566
## 517 2014-05-31   1543
## 518 2014-06-01   2229
```

```
##           date tday
## 513 2014-05-27  8.4
## 514 2014-05-28  7.4
## 515 2014-05-29  9.4
## 516 2014-05-30 11.3
## 517 2014-05-31 11.3
## 518 2014-06-01 13.4
```

Merge.

```
# bw as in "Bikers and Weather"
bw <- merge(bikers, tem)
nrow(bw)
```

```
## [1] 0
```

The date column is of different type. We need to transform tem$date into Date type.

```
tem$date <- as.Date(tem$date)
bw <- merge(bikers, tem)
nrow(bw)
```

```
## [1] 518
```

```
tail(bw)
```

```
##           date bikers tday
## 513 2014-05-27   3084  8.4
## 514 2014-05-28   2488  7.4
## 515 2014-05-29    952  9.4
## 516 2014-05-30   2566 11.3
## 517 2014-05-31   1543 11.3
## 518 2014-06-01   2229 13.4
```

Add weekday data column

```
bw$weekday <- format( as.Date(bw$date), "%w" )
```

Description of columns

- date: the date of the biker count / weather observations
- bikers: number of bikers passing the electronic counter device at Baana
- tday: average temperature of the day

- weekday: weekday of the day, 0=Sunday, 1=Monday