03_Data_Preparation

Aleksandra Kudaeva

28 Februar 2019

Preparation of Air-Pollution Data

The folloving packages were installed for data preparation of air pollution data: rvest readxl

```
#install and upload packages
packages = c("rvest",
             "readxl")

for (package in packages) {
  if(!require(package, character.only = TRUE)){
      install.packages(package, character.only = TRUE)
    }
  library(package, character.only = TRUE)
}
```

Uploading the Data

The data itself was downloaded in xls format from: "https://fbinter.stadt-berlin.de/fb/index.jsp"

```
# Upload the data
ap15 = read_excel("./Input/Air_Pollution_2015.xls",
                  sheet = 1)

# Original names are too long and contain special symbols and spaces
# Upload matching table for short names
mtch = read.csv2("./Input/matching.csv",
                 sep = ";")

# Rename variables
names(ap15) = mtch$new[match(names(ap15), mtch$old)]
```

Data Description

Initial Data Characteristics are the following:

```
#data summary
str(ap15)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    12374 obs. of  10 variables:
##  $ Nr          : chr  "6147" "86" "87" "2045" ...
##  $ Street      : chr  "Adalbertstr." "Adamstr." "Adamstr." "Adamstr." ...
##  $ Length      : num  40 124 103 34 13 73 129 39 197 288 ...
##  $ KFZ_daily   : num  5040 6930 6930 9810 9810 ...
##  $ NO2_yearly  : num  24.7 27 26.9 22.5 19 ...
```

```
##  $ PM10_yearly : num  21.6 18.9 18.9 18.3 17.7 ...
##  $ PM25_yearly : num  15.1 13.4 13.4 13 12.6 ...
##  $ NO2_index   : num  0.62 0.67 0.67 0.56 0.48 0.72 0.73 0.43 0.46 0.45 ...
##  $ PM10_index  : num  0.54 0.47 0.47 0.46 0.44 0.48 0.48 0.48 0.48 0.49 ...
##  $ Index_overal: num  1.16 1.14 1.14 1.02 0.92 1.2 1.21 0.91 0.94 0.94 ...
```

Several data columns needed to be reformated. For example, there were no unique way to write names of streets ()

```
# Reformat street section number
ap15$Nr = as.numeric(ap15$Nr)

# Format street names in the table, so that it is possible to merge 2 tables
ap15$str = sub("str.$|Straße|str$|-Straße|Strasse|straße|strasse|Str.",
               "",
               ap15$Street)
```

The data does not have an indicator od district, just the name of the street. To solve this problem, I use the table which matches street names and indixes/districts.

Uploading additional tables

Unfortunately, it was not possible to find a table of correspondence of street names and post indixes available for direct download. That is why I had to download it from the following web page separately for each district with help of "rvest" package: "https://berlin.kauperts.de"

```
#IMPORTANT: internet connection is needed

#create table with data for the whole Berlin
Berlin_streets = data.frame()

for (i in 1:dim(districts)[1]) {
  #generate a link to data for all the districts and sub-districts
  link=paste0("https://berlin.kauperts.de/Bezirke/",
              districts$District[i],"/Ortsteile/",
              districts$Sub.district[i],"/Strassen")

  #upload the data from the web-page with generated link
  webpage = read_html(link)
  tbls = html_nodes(webpage, "table")
  tab=html_table(tbls)[[1]]

  #add columns for district and sub-district
  tab$district=districts$District[i]
  tab$subdistrict=districts$Sub.district[i]

  #add created table to the table for the whole Berlin
  Berlin_streets=rbind(Berlin_streets, tab)
}

#save resulting table to csv
write.csv2(Berlin_streets, "./Output/Berlin_Streets.csv", row.names = FALSE)
```