# lecture 3

## Table of contents

## 1 LUXEMBURG DATA PROJECT

```
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(purrr)
library(readxl)
library(stringr)
library(janitor)
```

```
Attaching package: 'janitor'

The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

## 1.1 Getting Data

```r
#the link for the data

url      <- "https://is.gd/1vvBAc"
raw_data <- tempfile(fileext = ".xslx")
download.file(url , raw_data , method = "auto" , mode = "wb")

sheets <- excel_sheets(raw_data)
```

```r
read_clean <- function(..., sheet){

  read_excel(..., sheet = sheet) |>

    mutate(year = sheet)

}

raw_data <- map(
  sheets,
  ~read_clean(raw_data,
              skip = 10,
              sheet = .)
) |>
  bind_rows() |>
  clean_names()
```

```
New names:
* `*` -> `*...3`
* `*` -> `*...4`
```

```r
raw_data
```

```
# A tibble: 1,343 x 9
   commune    nombre_doffres prix_moyen_annonce_e~1 prix_moyen_annonce_a~2 year
   <chr>               <dbl> <chr>                  <chr>                  <chr>
 1 Bascharage            192 593698.31000000006     3603.57                2010
 2 Beaufort              266 461160.29              2902.76                2010
 3 Bech                   65 621760.22              3280.51                2010
 4 Beckerich             176 444498.68              2867.88                2010
 5 Berdorf               111 504040.85              3055.99                2010
 6 Bertrange             264 795338.87              4266.46                2010
 7 Bettembou~            304 555628.29              3343.22                2010
 8 Bettendorf             94 495074.38              3235.26                2010
 9 Betzdorf              119 625914.47              3343.05                2010
10 Bissen                 70 516465.57              3321.65                2010
# i 1,333 more rows
# i abbreviated names: 1: prix_moyen_annonce_en_courant,
#   2: prix_moyen_annonce_au_m2_en_courant
# i 4 more variables: bech <chr>, x12 <dbl>, x3 <chr>, x4 <chr>
```

Some variables has their original names and we will change them to English

```
raw_data <- raw_data |>

  rename(

    locality = commune,

    n_offers = nombre_doffres,

    average_price_nominal_euros = prix_moyen_annonce_en_courant,

    average_price_m2_nominal_euros = prix_moyen_annonce_au_m2_en_courant,

    average_price_m2_nominal_euros = prix_moyen_annonce_au_m2_en_courant

  ) |>

  mutate(locality = str_trim(locality)) |>

  select(year, locality, n_offers, starts_with("average"))

raw_data
```

```
# A tibble: 1,343 x 5
   year  locality    n_offers average_price_nominal_euros average_price_m2_nom~1
   <chr> <chr>          <dbl> <chr>                       <chr>
 1 2010  Bascharage       192 593698.31000000006          3603.57
 2 2010  Beaufort         266 461160.29                   2902.76
 3 2010  Bech              65 621760.22                   3280.51
 4 2010  Beckerich        176 444498.68                   2867.88
 5 2010  Berdorf          111 504040.85                   3055.99
 6 2010  Bertrange        264 795338.87                   4266.46
 7 2010  Bettembourg      304 555628.29                   3343.22
 8 2010  Bettendorf        94 495074.38                   3235.26
 9 2010  Betzdorf         119 625914.47                   3343.05
10 2010  Bissen            70 516465.57                   3321.65
# i 1,333 more rows
# i abbreviated name: 1: average_price_m2_nominal_euros
```

let's find some typos

```
raw_data |>
  filter(grepl("Luxembourg" , locality)) |>
  count(locality)
```

```
# A tibble: 2 x 2
  locality             n
  <chr>            <int>
1 Luxembourg           9
2 Luxembourg-Ville     2
```

```
raw_data |> filter(grepl("P.tange" , locality)) |>
  count(locality)
```

```
# A tibble: 2 x 2
  locality     n
  <chr>    <int>
1 Petange      9
2 Pétange      2
```