Datacleaning

Arvid Mikkers

Packages

We need the libabry tidyverse for data manipulation.

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------------------- tidy
```

```
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v dplyr   0.7.4
## v tidyr   0.8.0     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0
```

```
## -- Conflicts -------------------------------------------------------------------------- tidyverse_
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Data

RIVM

We have downloaded on July 19th, 2018 the data about life expectancy at birth from RIVM

We are interested in the LE for the total population.

```
dataLE1 <- read.csv2("../sourcedata/RIVM/LE.csv") %>%
  filter(Geslacht == "Totaal")%>%
  select(Gemeente, Bij.geboorte)
```

```
  colnames(dataLE1)[colnames(dataLE1)=="Bij.geboorte"] <- "LEtotalpop"
```

```
dataLE2 <- read.csv2("../sourcedata/RIVM/LE.csv") %>%
  filter(Geslacht == "Mannen")%>%
  select(Gemeente, Bij.geboorte)
```

```
  colnames(dataLE2)[colnames(dataLE2)=="Bij.geboorte"] <- "LEmen"
```

```
dataLE3 <- read.csv2("../sourcedata/RIVM/LE.csv") %>%
  filter(Geslacht == "Vrouwen")%>%
  select(Gemeente, Bij.geboorte)
```

```
  colnames(dataLE3)[colnames(dataLE3)=="Bij.geboorte"] <- "LEwomen"
```

```
dataLE4 <- full_join(dataLE1, dataLE2, by = "Gemeente")
```

```
dataLE <- full_join(dataLE4, dataLE3, by = "Gemeente")
```

CBS

Kerncijfers

We have downloaded on July 19th, 2018 the data about "kerncijfers" plus metadata from CBS

```
kerncijfers <- read.csv2("../sourcedata/CBS/kerncijfers.csv")
```

The data contains information about different aggregation levels. We are interested in the data of the municipalities. The id's for municipalities contain the string "GM". We will filter the municipalities. There are many municipalities without population, we will select municipalities with population.

```
kerncijfers2 <- kerncijfers %>%
  filter(grepl("GM", RegioS)) %>%
  filter(TotaleBevolking_1 != "NA")
```

The file `kerncijfers.csv` contains data per municipality about demographic characteristics. The data do not contain the names of the municipalities, but id's. To be able to merge these data with the RIVM data on life expecentancy, we will first produce a table with id's and names of municipalities from the metadata of the "kerncijfers". We need to rename TableInfos to RegioS and X to "Gemeente" to match the columns.

```
MD_kerncijfers <- read.csv2("../sourcedata/CBS/metadata_kerncijfers.csv")
```

```
Merge_table <- MD_kerncijfers %>%
  select(RegioS = TableInfos, Gemeente = X) %>%
  filter(grepl("GM", RegioS))

Merge_table$Gemeente <- as.character(Merge_table$Gemeente)

str(Merge_table)
```

```
## 'data.frame':    713 obs. of  2 variables:
##  $ RegioS  : Factor w/ 1226 levels "0","1","10","100",..: 1136 937 744 624 807 938 708 518 745 808 .
##  $ Gemeente: chr  "Aa en Hunze" "Aalburg" "Aalsmeer" "Aalten" ...
```

Now we can merge the `kerncijfers` with the `Merge_table`. First we will change the variable `RegioS` in the dataframes to be merged in to a character

```
Merge_table$RegioS <- as.character(Merge_table$RegioS)
kerncijfers2$RegioS <- as.character(kerncijfers2$RegioS)


kerncijfers3 <- right_join(Merge_table, kerncijfers2, by="RegioS")
```

And finally we can merge `dataLE` with `kerncijfers3` First we willl change the variable `Gemeente` in to a character.

```
dataLE$Gemeente <- as.character(dataLE$Gemeente)
kerncijfers3$Gemeente <- as.character(kerncijfers3$Gemeente)
data1 <- inner_join(dataLE, kerncijfers3, by = "Gemeente")
```

Data selection

```
data1a <- data1 %>%

  select(LEtotalpop, LEwomen, LEmen, Gemeente, Gescheiden_32, TotaalMetMigratieachtergrond_44, Bevolkin



data1a$Gescheiden_32 <- as.numeric(as.character(data1a$Gescheiden_32))
data1a$TotaalMetMigratieachtergrond_44 <- as.numeric(as.character(data1a$TotaalMetMigratieachtergrond_4
data1a$GemiddeldeHuishoudensgrootte_89 <- as.numeric(as.character(data1a$GemiddeldeHuishoudensgrootte_8
data1a$Koopwoningen_94 <- as.numeric(as.character(data1a$Koopwoningen_94))
data1a$AfstandTotZiekenhuis_216 <- as.numeric(as.character(data1a$AfstandTotZiekenhuis_216))
data1a$k_80JaarOfOuder_21 <- as.numeric(as.character(data1a$k_80JaarOfOuder_21))
str(data1a)
```

```
## 'data.frame':    374 obs. of  15 variables:
##  $ LEtotalpop                     : num  82.5 80.9 83.2 82 81 82.8 82.7 81.7 79.7 81.8 ...
##  $ LEwomen                        : num  85 81.9 85.7 84.1 83.1 84.6 84.4 83.6 81.6 83.3 ...
##  $ LEmen                          : num  80 79.8 80.8 79.9 78.9 81.1 80.9 79.7 77.7 80.3 ...
##  $ Gemeente                       : chr  "Aa en Hunze" "Aalburg" "Aalsmeer" "Aalten" ...
##  $ Gescheiden_32                  : num  8.2 5.3 7.7 6.1 7.5 7.2 8.8 11 9.8 11.2 ...
##  $ TotaalMetMigratieachtergrond_44: num  5.9 5.5 18.1 11 4 13.6 19.3 21.6 24.9 40 ...
##  $ Bevolkingsdichtheid_57         : int  91 259 1555 279 274 2270 1150 974 1076 1533 ...
##  $ GemiddeldeHuishoudensgrootte_89: num  2.26 2.65 2.41 2.35 2.39 2.46 2.43 2.1 2.24 2.38 ...
##  $ Woningdichtheid_93             : int  40 100 640 122 117 935 470 456 482 617 ...
##  $ Koopwoningen_94                : num  71.1 71.7 63.1 69.4 63.3 60.1 67.3 57.7 54.6 63.8 ...
##  $ GemiddeldeWoningwaarde_99      : int  218 239 269 185 159 194 248 183 151 181 ...
##  $ AfstandTotZiekenhuis_216       : num  11.5 11.8 9.7 14.5 15.7 8.5 6.3 4 3.7 4.4 ...
##  $ TotaleBevolking_1              : int  25243 13038 31299 26912 28007 19955 24985 107615 72425 1981
##  $ k_80JaarOfOuder_21             : num  5.6 3.8 4.5 5.2 4 4.9 3.8 4.1 4.4 2 ...
##  $ k_80JaarOfOuder_12             : int  1408 490 1398 1397 1125 985 944 4457 3192 4009 ...
```

Gemeentefonds

We need to add some information to the dataset about socio-economic factors (such as income, education and benefits).

We have downloaded on July 19th, 2018 the data about "Gemeentefonds" plus metadata from CBS

We need the variable `RegioS` to be able to merge. We will `filter` this variable for all observations containing ("GM")

For Income we will select the average income and the percentage of low income people. For education we will take the percentage of lower educated people We also want to select the percentage of one person households. This variable is not available, but we will construct this variable from the number of households and the number of one person households.

```
GF <- read.csv2("../sourcedata/CBS/gemeentefonds.csv")

GF2 <- GF %>%
  filter(grepl("GM", RegioS)) %>%
  select(RegioS, GemiddeldGestandaardiseerdInkomen_41, InkomenTot120SociaalMinimum_13, LagerOpgeleidenP
  mutate(Percentage_eenpersoonshuishoudens = (Eenpersoonshuishoudens_44/Huishoudens_32)*100)


GF3 <- GF2 %>%
  select(-Huishoudens_32, -Eenpersoonshuishoudens_44)

GF3$GemiddeldGestandaardiseerdInkomen_41 <- as.numeric(as.character(GF3$GemiddeldGestandaardiseerdInkom
```

```
## Warning: NAs introduced by coercion
```

```
GF3$InkomenTot120SociaalMinimum_13 <- as.numeric(as.character(GF3$InkomenTot120SociaalMinimum_13))
GF3$LagerOpgeleidenPercentage_5 <- as.numeric(GF3$LagerOpgeleidenPercentage_5)

str(GF3)
```

```
## 'data.frame':    390 obs. of  5 variables:
##  $ RegioS                            : Factor w/ 391 levels "GM0003","GM0005",..: 330 234 128 65 2
##  $ GemiddeldGestandaardiseerdInkomen_41: num  23000 23700 25000 23500 20900 22500 24800 21500 20100 1
##  $ InkomenTot120SociaalMinimum_13      : num  8.2 6.2 6.6 7.6 13 8.7 6.7 11.9 17.6 16 ...
##  $ LagerOpgeleidenPercentage_5         : num  12 31 23 20 30 18 14 16 21 17 ...
##  $ Percentage_eenpersoonshuishoudens   : num  26.4 23.3 27.9 27.9 28.7 ...
```

Now we can merge `GF3` with `Merge_table`, we need to change `RegioS` in to a character. And GF4 with `data1`, we need to change `Gemeente` in to a character.

```
GF3$RegioS <- as.character(GF3$RegioS)


GF4 <-right_join(Merge_table, GF3, by="RegioS")

GF4$gemeente <- as.character(GF4$Gemeente)
data2 <- left_join(data1a, GF4, by= "Gemeente")
```

Health monitor

We have downloaded on July 19th, 2018 the data about the health plus metadata from CBS

```
HM <- read.csv2("../sourcedata/CBS/health_monitor.csv")
summary(HM$Perioden)
```

```
## 2016JJ00
##     3888
```

We add the names of the municipalities to our table instead of the ID's the CBS uses.

```
HM$RegioS <- as.character(HM$RegioS)


HM2 <- full_join(Merge_table,HM , by="RegioS") %>%
 filter(grepl("GM", RegioS)) %>%
  filter(Leeftijd == 10000, Marges== "MW00000")
```

We would like to add the following data to our dataset

```
HM3 <- HM2 %>%
  select(Gemeente, ErvarenGezondheidGoedZeerGoed_1, EenOfMeerLangdurigeAandoeningen_2, NormaalGewicht_9

str(HM3)
```

```
## 'data.frame':    390 obs. of  7 variables:
##  $ Gemeente                          : chr  "Aa en Hunze" "Aalburg" "Aalsmeer" "Aalten" ...
##  $ ErvarenGezondheidGoedZeerGoed_1   : Factor w/ 421 levels "       .","    43.4",..: 300 292 310 261
##  $ EenOfMeerLangdurigeAandoeningen_2: Factor w/ 446 levels "       .","    11.0",..: 172 86 150 187 
##  $ NormaalGewicht_9                  : Factor w/ 403 levels "       .","    19.4",..: 186 223 239 207
##  $ VoldoetAanFitnorm_14              : Factor w/ 514 levels "       .","     7.4",..: 177 103 160 208
##  $ UrenMantelzorgPerWeek_19          : Factor w/ 219 levels "       .","     2.6",..: 98 67 50 54 1 1
##  $ WekelijkseSporters_16             : Factor w/ 519 levels "       .","    14.0",..: 261 181 341 277
```

We will make the variables numeric

```
HM3$ErvarenGezondheidGoedZeerGoed_1 <- as.numeric(as.character(HM3$ErvarenGezondheidGoedZeerGoed_1))
```

```
## Warning: NAs introduced by coercion
```

```
HM3$EenOfMeerLangdurigeAandoeningen_2 <- as.numeric(as.character(HM3$EenOfMeerLangdurigeAandoeningen_2))
```

```
## Warning: NAs introduced by coercion
```

```
HM3$NormaalGewicht_9 <- as.numeric(as.character(HM3$NormaalGewicht_9))
```

```
## Warning: NAs introduced by coercion
```

```
HM3$VoldoetAanFitnorm_14 <- as.numeric(as.character(HM3$VoldoetAanFitnorm_14))
```

```
## Warning: NAs introduced by coercion
```

```
HM3$UrenMantelzorgPerWeek_19 <- as.numeric(as.character(HM3$UrenMantelzorgPerWeek_19))
```

```
## Warning: NAs introduced by coercion
```

```
HM3$WekelijkseSporters_16 <- as.numeric(as.character(HM3$WekelijkseSporters_16))
```

```
## Warning: NAs introduced by coercion
```

Now we can merge the Health Monitor Data with the rest of our data.

```r
data3 <- left_join(data2, HM3, by= "Gemeente")

str(data3)
```

```
## 'data.frame':    374 obs. of  27 variables:
##  $ LEtotalpop                          : num  82.5 80.9 83.2 82 81 82.8 82.7 81.7 79.7 81.8 ...
##  $ LEwomen                             : num  85 81.9 85.7 84.1 83.1 84.6 84.4 83.6 81.6 83.3 ...
##  $ LEmen                               : num  80 79.8 80.8 79.9 78.9 81.1 80.9 79.7 77.7 80.3 ...
##  $ Gemeente                            : chr  "Aa en Hunze" "Aalburg" "Aalsmeer" "Aalten" ...
##  $ Gescheiden_32                       : num  8.2 5.3 7.7 6.1 7.5 7.2 8.8 11 9.8 11.2 ...
##  $ TotaalMetMigratieachtergrond_44     : num  5.9 5.5 18.1 11 4 13.6 19.3 21.6 24.9 40 ...
##  $ Bevolkingsdichtheid_57              : int  91 259 1555 279 274 2270 1150 974 1076 1533 ...
##  $ GemiddeldeHuishoudensgrootte_89     : num  2.26 2.65 2.41 2.35 2.39 2.46 2.43 2.1 2.24 2.38 ...
##  $ Woningdichtheid_93                  : int  40 100 640 122 117 935 470 456 482 617 ...
##  $ Koopwoningen_94                     : num  71.1 71.7 63.1 69.4 63.3 60.1 67.3 57.7 54.6 63.8 ...
##  $ GemiddeldeWoningwaarde_99           : int  218 239 269 185 159 194 248 183 151 181 ...
##  $ AfstandTotZiekenhuis_216            : num  11.5 11.8 9.7 14.5 15.7 8.5 6.3 4 3.7 4.4 ...
##  $ TotaleBevolking_1                   : int  25243 13038 31299 26912 28007 19955 24985 107615 72425
##  $ k_80JaarOfOuder_21                  : num  5.6 3.8 4.5 5.2 4 4.9 3.8 4.1 4.4 2 ...
##  $ k_80JaarOfOuder_12                  : int  1408 490 1398 1397 1125 985 944 4457 3192 4009 ...
##  $ RegioS                              : chr  "GM1680" "GM0738" "GM0358" "GM0197" ...
##  $ GemiddeldGestandaardiseerdInkomen_41: num  23000 23700 25000 23500 20900 22500 24800 21500 20100 2
##  $ InkomenTot120SociaalMinimum_13      : num  8.2 6.2 6.6 7.6 13 8.7 6.7 11.9 17.6 16 ...
##  $ LagerOpgeleidenPercentage_5         : num  12 31 23 20 30 18 14 16 21 17 ...
##  $ Percentage_eenpersoonshuishoudens   : num  26.4 23.3 27.9 27.9 28.7 ...
##  $ gemeente                            : chr  "Aa en Hunze" "Aalburg" "Aalsmeer" "Aalten" ...
##  $ ErvarenGezondheidGoedZeerGoed_1     : num  79.5 78.7 80.5 75.6 81.3 75.3 77.9 72.6 72.3 73.2 ...
##  $ EenOfMeerLangdurigeAandoeningen_2   : num  33.5 24.9 31.3 35 35.9 33.4 32.1 36.2 38.2 34.1 ...
##  $ NormaalGewicht_9                    : num  43.7 47.4 49 45.8 40.8 47.9 47.8 52.7 42.9 44 ...
##  $ VoldoetAanFitnorm_14                : num  28.6 21.2 26.9 31.7 26.1 19.9 22.3 28.2 32 18.7 ...
##  $ UrenMantelzorgPerWeek_19            : num  12.8 9.7 8 8.4 NA NA NA 8.7 10.2 16.4 ...
##  $ WekelijkseSporters_16               : num  46.3 38.3 54.3 47.9 41.7 42.3 50.1 53.3 47.1 49.7 ...
```

Rijkswaterstaat

We have downloaded on September 27th, 2018 the data about "CO2 emssions" from Rijkswaterstaat

```
RWS <- read.csv2("../sourcedata/Rijkswaterstaat/CO2.csv", sep =",")
str(RWS)
```

```
## 'data.frame':    381 obs. of  2 variables:
##  $ Gemeenten                       : Factor w/ 381 levels "'s-Hertogenbosch",..: 2 3 4 5 6 7 8 9 10 1
##  $ Totaal.bekende.CO2.uitstoot.2016: int  NA NA NA 134297 159597 201343 117913 579053 571012 831527
```

Since in our dataset the column containing the Municipalities is called "Gemeente" and in the datset from "Rijkswaterstaat" is called "Gemeenten" we have to change the name before we can merge it with our dataset.

```
colnames(RWS)[colnames(RWS)=="Gemeenten"] <- "Gemeente"

rijkswaterstaat1 <- RWS
str(RWS)
```

```
## 'data.frame':    381 obs. of  2 variables:
##  $ Gemeente                        : Factor w/ 381 levels "'s-Hertogenbosch",..: 2 3 4 5 6 7 8 9 10 1
##  $ Totaal.bekende.CO2.uitstoot.2016: int  NA NA NA 134297 159597 201343 117913 579053 571012 831527
```

Now we can merge the data from "Rijkswaterstaat" with our previous made dataset so that we have one finaldataset to work with.

```
data4 <- left_join(data3, rijkswaterstaat1, by= "Gemeente")
```

```
## Warning: Column `Gemeente` joining character vector and factor, coercing
## into character vector
```

Finally we will save data4 as cvs

```
write.csv2(data4, "../Sourcedata/Analysis/Datafile.csv")
```