PEC 3: Desing and Implementation

UOC - Alumno: Álvaro Rodríguez Sans

May 2020 - Delivery 23/05/2021

Índex

1 Data load **5**

2 Initial descriptive statistics and visualization **6**
2.1 Data types and modifications . . . 6
2.1.1 EM3 review . . . 6
2.1.2 EM3 data transformation . . . 7
2.1.3 EM3 transpose . . . 7
2.1.4 EM3 review missing values & impute . . . 8
2.1.5 Google review . . . 20
2.1.6 Google autonomous-communities & provinces . . . 23
2.1.7 Google data transformation . . . 27
2.1.8 Google review missing values & impute . . . 29
2.1.9 CNE review . . . 65
2.1.10 CNE review missing values & impute . . . 66
2.1.11 CNE data transformation . . . 74
2.2 Datasets combinations . . . 76
2.2.1 CNE_tec_cas . . . 76
2.2.2 GOG_CNE . . . 77
2.2.3 Total . . . 79
2.3 Visual analysis . . . 89
2.3.1 Dataframe plots . . . 89
2.3.2 Time-series plots . . . 92
2.3.3 Correlation plots . . . 95
2.3.4 PCA . . . 98

3 Seasonal and trend decomposition **111**
3.0.1 ACF . . . 111
3.1 STL (Seasonal and Trend decomposition using Loess) . . . 112
3.2 Till here 06-Apr-2021 . . . 117

Bibliography **117**

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code. Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*. When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file). The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

The bibliographic references used for this practice have been: (Baayen 2008; Hothorn and Everitt 2014; Hyndman and Athanasopoulos 2021; Liviano Solas and Pujol Jover, n.d.; Teetor 2011; Vegas Lozano, n.d.).

```
if(!require(knitr)){
    install.packages('knitr', repos='http://cran.us.r-project.org')
    library(knitr)}
```

```
## Loading required package: knitr
```

```
if(!require(latexpdf)){
    install.packages('latexpdf', repos='http://cran.us.r-project.org')
    library(latexpdf)}
```

```
## Loading required package: latexpdf
```

```
if(!require(latex2exp)){
    install.packages('latex2exp', repos='http://cran.us.r-project.org')
    library(latex2exp)}
```

```
## Loading required package: latex2exp
```

```
if(!require(psych)){
    install.packages("psych", repos='http://cran.us.r-project.org')
    library(psych)}
```

```
## Loading required package: psych
```

```
if(!require(DescTools)){
    install.packages("DescTools", repos='http://cran.us.r-project.org')
    library(DescTools)}
```

```
## Loading required package: DescTools
```

```
##
## Attaching package: 'DescTools'
```

```
## The following objects are masked from 'package:psych':
##
##     AUC, ICC, SD
```

```
if(!require(tidyverse)){
    install.packages("tidyverse", repos='http://cran.us.r-project.org')
    library(tidyverse)}
```

```
## Loading required package: tidyverse
```

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

```
## v ggplot2 3.3.3     v purrr   0.3.3
## v tibble  3.0.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
if(!require(imputeTS)){
    install.packages("imputeTS", repos='http://cran.us.r-project.org')
```

```
    library(imputeTS)}
```

```
## Loading required package: imputeTS
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
if(!require(VIM)){
    install.packages('VIM', repos='http://cran.us.r-project.org')
    library(VIM)}
```

```
## Loading required package: VIM
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## Loading required package: data.table
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
## The following object is masked from 'package:DescTools':
##
##     %like%
```

```
## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues
```

```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
if(!require(stats)){
    install.packages("stats", repos='http://cran.us.r-project.org')
    library(stats)}
if(!require(tsbox)){
    install.packages("tsbox", repos='http://cran.us.r-project.org')
    library(tsbox)}
```

```
## Loading required package: tsbox
```

```
if(!require(fable)){
    install.packages("fable", repos='http://cran.us.r-project.org')
    library(fable)}
```

```
## Loading required package: fable
```

```
## Loading required package: fabletools
```

```
##
## Attaching package: 'fabletools'
```

```
## The following objects are masked from 'package:DescTools':
##
##     MAE, MAPE, MSE, RMSE
```

```
if(!require(fpp3)){
    install.packages("fpp3", repos='http://cran.us.r-project.org')
    library(fpp3)}
```

```
## Loading required package: fpp3
```

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

```
## v lubridate   1.7.8     v tsibbledata 0.3.0
## v tsibble     1.0.0     v feasts      0.2.1
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x ggplot2::%+%()          masks psych::%+%()
## x ggplot2::alpha()        masks psych::alpha()
## x data.table::between()   masks dplyr::between()
## x lubridate::date()       masks base::date()
## x dplyr::filter()         masks stats::filter()
## x data.table::first()     masks dplyr::first()
## x lubridate::hour()       masks data.table::hour()
## x tsibble::intersect()    masks base::intersect()
## x tsibble::interval()     masks lubridate::interval()
## x lubridate::isoweek()    masks data.table::isoweek()
## x tsibble::key()          masks data.table::key()
## x dplyr::lag()            masks stats::lag()
## x data.table::last()      masks dplyr::last()
## x fabletools::MAE()       masks DescTools::MAE()
## x fabletools::MAPE()      masks DescTools::MAPE()
## x lubridate::mday()       masks data.table::mday()
## x lubridate::minute()     masks data.table::minute()
## x lubridate::month()      masks data.table::month()
## x fabletools::MSE()       masks DescTools::MSE()
## x lubridate::quarter()    masks data.table::quarter()
## x fabletools::RMSE()      masks DescTools::RMSE()
## x lubridate::second()     masks data.table::second()
## x tsibble::setdiff()      masks base::setdiff()
## x data.table::transpose() masks purrr::transpose()
## x tsibble::union()        masks base::union()
## x lubridate::wday()       masks data.table::wday()
## x lubridate::week()       masks data.table::week()
## x lubridate::yday()       masks data.table::yday()
## x lubridate::year()       masks data.table::year()
```

```
if(!require(corrplot)){
    install.packages('corrplot', repos='http://cran.us.r-project.org')
    library(corrplot)}
```

```
## Loading required package: corrplot
```

```
## corrplot 0.84 loaded
```

```
if(!require(corrplot)){
    install.packages('corrplot', repos='http://cran.us.r-project.org')
    library(corrplot)}
if(!require(forecast)){
    install.packages('forecast', repos='http://cran.us.r-project.org')
    library(forecast)}
```

```
## Loading required package: forecast
```

```
##
## Attaching package: 'forecast'
```

```
## The following objects are masked from 'package:fabletools':
##
##     accuracy, forecast
```

```
## The following object is masked from 'package:DescTools':
##
##     BoxCox
```

```
if(!require(keras)){
    install.packages('keras', repos='http://cran.us.r-project.org')
    library(keras)}
```

```
## Loading required package: keras
```

```
if(!require(tensorflow)){
    install.packages('tensorflow', repos='http://cran.us.r-project.org')
    library(tensorflow)}
```

```
## Loading required package: tensorflow
```

```
knitr::opts_chunk$set(echo = TRUE)
```

1 Data load

Data is loaded from the sources stated at PEC1 and PEC2 (CNE, INE and Google).

- CNE-Covid-19
- INE-Covid-19
- Google-Covid-19

```
#library(dplyr)
# Source INE
EM3 <- read.csv('EM3-Movimiento de personas por provincias.csv',
                header=TRUE,
                sep = ";",
                stringsAsFactors = FALSE)

# Source Google
Google <- read.csv('Google-2020_ES_Region_Mobility_Report.csv',
                   header=TRUE,
                   sep = ";",
                   stringsAsFactors = FALSE)

# Source CNE
```

```
CNE_tecnica <- read.csv('CNE-casos_tecnica_provincia.csv',
                        header=TRUE,
                        sep = ",",
                        stringsAsFactors = FALSE)
CNE_casos <- read.csv('CNE-casos_hosp_uci_def_sexo_edad_provres.csv',
                      header=TRUE,
                      sep = ",",
                      stringsAsFactors = FALSE)
```

2 Initial descriptive statistics and visualization

2.1 Data types and modifications

We are gonig to check the **type of variable** that corresponds to each of the variables (numerical, factor, etc.) and **missing data / values or other anomalies** in each dataset.

2.1.1 EM3 review

We have the movement of people by provinces (we can see 146 rows by province, that correspond to days). In order to facilitate the comparison and have a valid reference on to what extent the mobility of the population should be considered to have varied, the data of a day of a week that can be considered "normal" are taken as a reference. For this study, the "normal" day that has been considered is the one that results from the average of the days 18 (Monday) to 21 (Thursday) of November 2019. It is indicated in the tables as the reference date 18/11/2019.

```
# Source INE
head(str(EM3,vec.len=2))
```

```
## 'data.frame':    9198 obs. of  3 variables:
##  $ Zonas.de.movilidad: chr  "Almería" "Almería" ...
##  $ Periodo           : chr  "30/12/2020" "27/12/2020" ...
##  $ Total             : chr  "17,17" "11,53" ...
```

```
## NULL
```

```
summary(EM3)
```

```
##  Zonas.de.movilidad   Periodo             Total
##  Length:9198        Length:9198        Length:9198
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
```

```
table(EM3$Zonas.de.movilidad)
```

```
##
##               Albacete       Alicante/Alacant                Almería
##                    146                    146                    146
##           Araba/Álava               Asturias                  Ávila
##                    146                    146                    146
##               Badajoz         Balears, Illes              Barcelona
##                    146                    146                    146
##               Bizkaia                 Burgos               Cáceres
##                    146                    146                    146
##                 Cádiz              Cantabria     Castellón/Castelló
##                    146                    146                    146
##                 Ceuta           Ciudad Real                Córdoba
```

```
##                   146                 146                 146
##             Coruña, A               Cuenca           Formentera
##                   146                 146                 146
##          Fuerteventura             Gipuzkoa              Girona
##                   146                 146                 146
##            Gomera, La       Gran Canaria             Granada
##                   146                 146                 146
##           Guadalajara          Hierro, El              Huelva
##                   146                 146                 146
##               Huesca               Ibiza               Jaén
##                   146                 146                 146
##             Lanzarote               León              Lleida
##                   146                 146                 146
##                 Lugo              Madrid              Málaga
##                   146                 146                 146
##              Mallorca             Melilla             Menorca
##                   146                 146                 146
##                Murcia             Navarra             Ourense
##                   146                 146                 146
##              Palencia          Palma, La        Palmas, Las
##                   146                 146                 146
##            Pontevedra          Rioja, La           Salamanca
##                   146                 146                 146
## Santa Cruz de Tenerife            Segovia             Sevilla
##                   146                 146                 146
##                 Soria           Tarragona            Tenerife
##                   146                 146                 146
##                Teruel              Toledo  Valencia/València
##                   146                 146                 146
##            Valladolid             Zamora            Zaragoza
##                   146                 146                 146
```

2.1.2 EM3 data transformation

We are going to **transform**:

- "Total" from "character" to "numerical"
- "Periodo" from "character" to "date"

```
EM3$Total <- sub(",", ".", EM3$Total)
EM3$Total <- as.numeric(EM3$Total)
EM3$Periodo <- as.Date(EM3$Periodo,format="%d/%m/%Y")
head(EM3)
```

```
##   Zonas.de.movilidad    Periodo Total
## 1            Almería 2020-12-30 17.17
## 2            Almería 2020-12-27 11.53
## 3            Almería 2020-12-23 17.81
## 4            Almería 2020-12-20 12.13
## 5            Almería 2020-12-16 18.28
## 6            Almería 2020-12-13 11.97
```

2.1.3 EM3 transpose

Due to the nature of this dataset we have to transpose it in order to analyse the missing values and impute them.

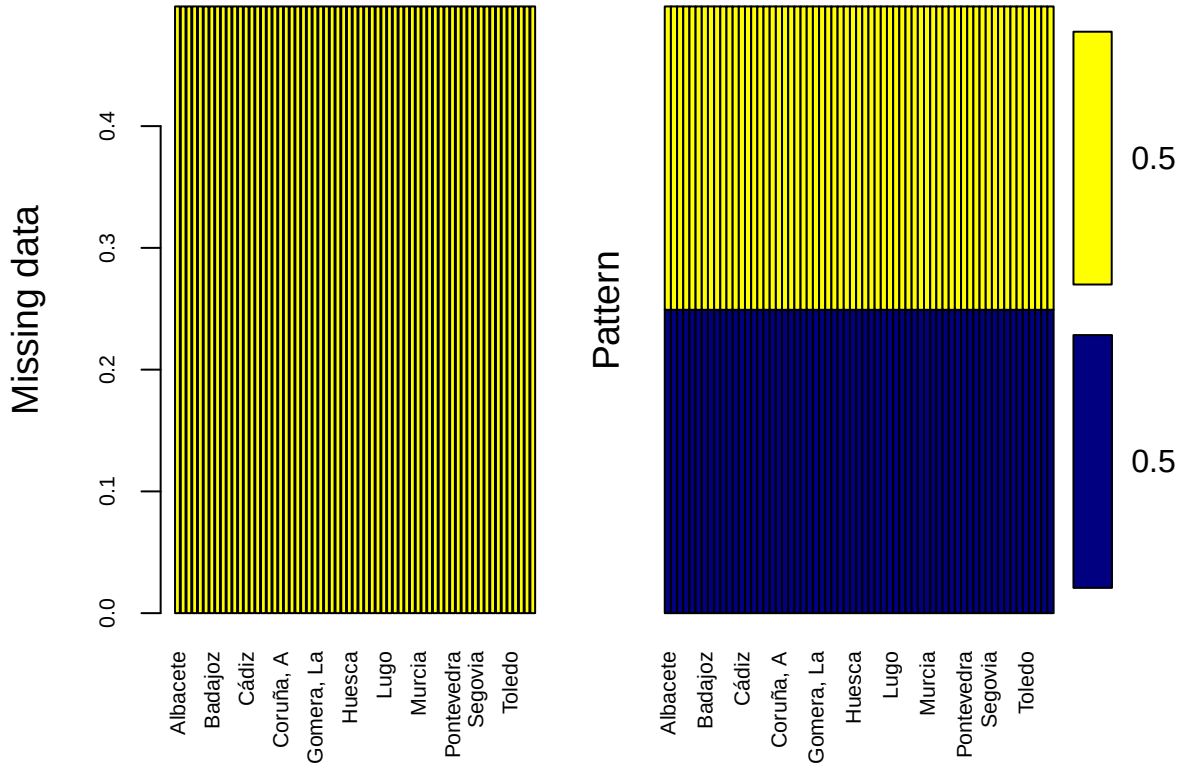
```r
if(!require(data.table)){
    install.packages('data.table', repos='http://cran.us.r-project.org')
    library(data.table)}

# Transpose dataframe
EM3_t<-dcast(EM3, Periodo~Zonas.de.movilidad, fill=NA)

# Create dates missing (for time series).
# Note: Acccording INE some dates are not provided.
EM3_t<-EM3_t %>%
  complete(Periodo = seq.Date(min(Periodo), max(Periodo), by="day"))

# Filter the interest period according INE EM3 study
EM3_t<- EM3_t %>%
  filter(Periodo <= "2019-11-18" | Periodo >= "2020-03-16")

EM3_t
```

```
## # A tibble: 291 x 64
##    Periodo    Albacete `Alicante/Alaca~ Almería `Araba/Álava` Asturias Ávila
##    <date>        <dbl>            <dbl>   <dbl>         <dbl>    <dbl> <dbl>
##  1 2019-11-18    25.2             28.1    24.4          31.9     29.9  26.6
##  2 2020-03-16     9.9             14.4    11.0          15.9     13.1   9.44
##  3 2020-03-17    NA               NA      NA            NA       NA    NA
##  4 2020-03-18     9.51            13.4     7.28         14.5     12.0   9.17
##  5 2020-03-19    NA               NA      NA            NA       NA    NA
##  6 2020-03-20     8.75            12.0     6.87         11.9     11.3   8.69
##  7 2020-03-21    NA               NA      NA            NA       NA    NA
##  8 2020-03-22     4.5              6.14    4.19          6.46     5.64  4.53
##  9 2020-03-23    NA               NA      NA            NA       NA    NA
## 10 2020-03-24     9.02            10.9     8.98         13.3     11.2   8.26
## # ... with 281 more rows, and 57 more variables: Badajoz <dbl>, `Balears,
## #   Illes` <dbl>, Barcelona <dbl>, Bizkaia <dbl>, Burgos <dbl>, Cáceres <dbl>,
## #   Cádiz <dbl>, Cantabria <dbl>, `Castellón/Castelló` <dbl>, Ceuta <dbl>,
## #   `Ciudad Real` <dbl>, Córdoba <dbl>, `Coruña, A` <dbl>, Cuenca <dbl>,
## #   Formentera <dbl>, Fuerteventura <dbl>, Gipuzkoa <dbl>, Girona <dbl>,
## #   `Gomera, La` <dbl>, `Gran Canaria` <dbl>, Granada <dbl>, Guadalajara <dbl>,
## #   `Hierro, El` <dbl>, Huelva <dbl>, Huesca <dbl>, Ibiza <dbl>, Jaén <dbl>,
## #   Lanzarote <dbl>, León <dbl>, Lleida <dbl>, Lugo <dbl>, Madrid <dbl>,
## #   Málaga <dbl>, Mallorca <dbl>, Melilla <dbl>, Menorca <dbl>, Murcia <dbl>,
## #   Navarra <dbl>, Ourense <dbl>, Palencia <dbl>, `Palma, La` <dbl>, `Palmas,
## #   Las` <dbl>, Pontevedra <dbl>, `Rioja, La` <dbl>, Salamanca <dbl>, `Santa
## #   Cruz de Tenerife` <dbl>, Segovia <dbl>, Sevilla <dbl>, Soria <dbl>,
## #   Tarragona <dbl>, Tenerife <dbl>, Teruel <dbl>, Toledo <dbl>,
## #   `Valencia/València` <dbl>, Valladolid <dbl>, Zamora <dbl>, Zaragoza <dbl>
```

2.1.4 EM3 review missing values & impute

We check the missing values by province (we are close to 150 by province).

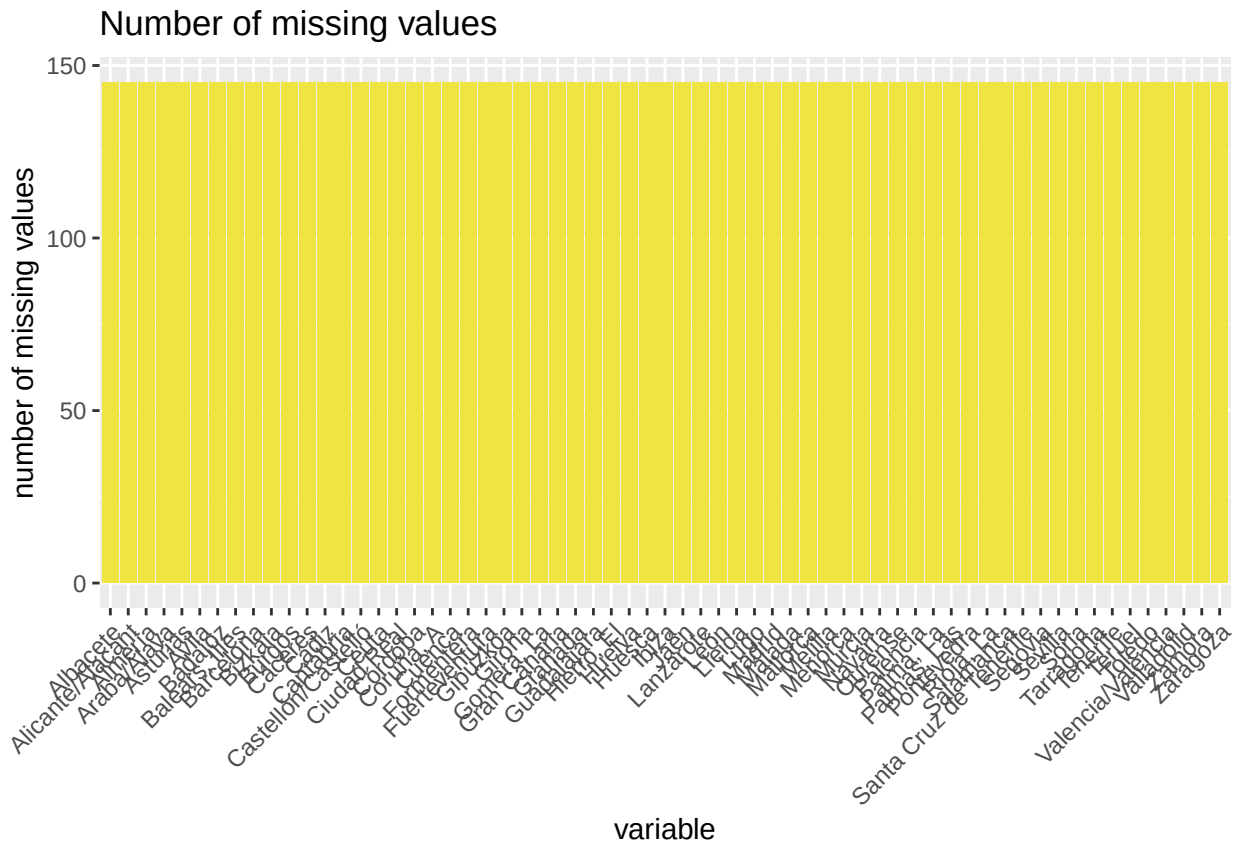
```r
aggr(EM3_t[,-1], col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(EM3_t[,-1]), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##                 Variable     Count
##                Albacete 0.4982818
##       Alicante/Alacant 0.4982818
##                Almería 0.4982818
##            Araba/Álava 0.4982818
##               Asturias 0.4982818
##                  Ávila 0.4982818
##                Badajoz 0.4982818
##         Balears, Illes 0.4982818
##              Barcelona 0.4982818
##                Bizkaia 0.4982818
##                 Burgos 0.4982818
##                Cáceres 0.4982818
##                  Cádiz 0.4982818
##              Cantabria 0.4982818
##     Castellón/Castelló 0.4982818
##                  Ceuta 0.4982818
##            Ciudad Real 0.4982818
##                Córdoba 0.4982818
##              Coruña, A 0.4982818
##                 Cuenca 0.4982818
##             Formentera 0.4982818
##          Fuerteventura 0.4982818
##                Gipuzkoa 0.4982818
```

```
##                  Girona 0.4982818
##              Gomera, La 0.4982818
##            Gran Canaria 0.4982818
##                 Granada 0.4982818
##             Guadalajara 0.4982818
##              Hierro, El 0.4982818
##                  Huelva 0.4982818
##                  Huesca 0.4982818
##                   Ibiza 0.4982818
##                    Jaén 0.4982818
##                Lanzarote 0.4982818
##                    León 0.4982818
##                  Lleida 0.4982818
##                    Lugo 0.4982818
##                  Madrid 0.4982818
##                  Málaga 0.4982818
##                 Mallorca 0.4982818
##                  Melilla 0.4982818
##                  Menorca 0.4982818
##                   Murcia 0.4982818
##                  Navarra 0.4982818
##                  Ourense 0.4982818
##                 Palencia 0.4982818
##               Palma, La 0.4982818
##             Palmas, Las 0.4982818
##              Pontevedra 0.4982818
##               Rioja, La 0.4982818
##              Salamanca 0.4982818
##  Santa Cruz de Tenerife 0.4982818
##                 Segovia 0.4982818
##                 Sevilla 0.4982818
##                   Soria 0.4982818
##               Tarragona 0.4982818
##                Tenerife 0.4982818
##                  Teruel 0.4982818
##                  Toledo 0.4982818
##       Valencia/València 0.4982818
##              Valladolid 0.4982818
##                  Zamora 0.4982818
##                Zaragoza 0.4982818
```

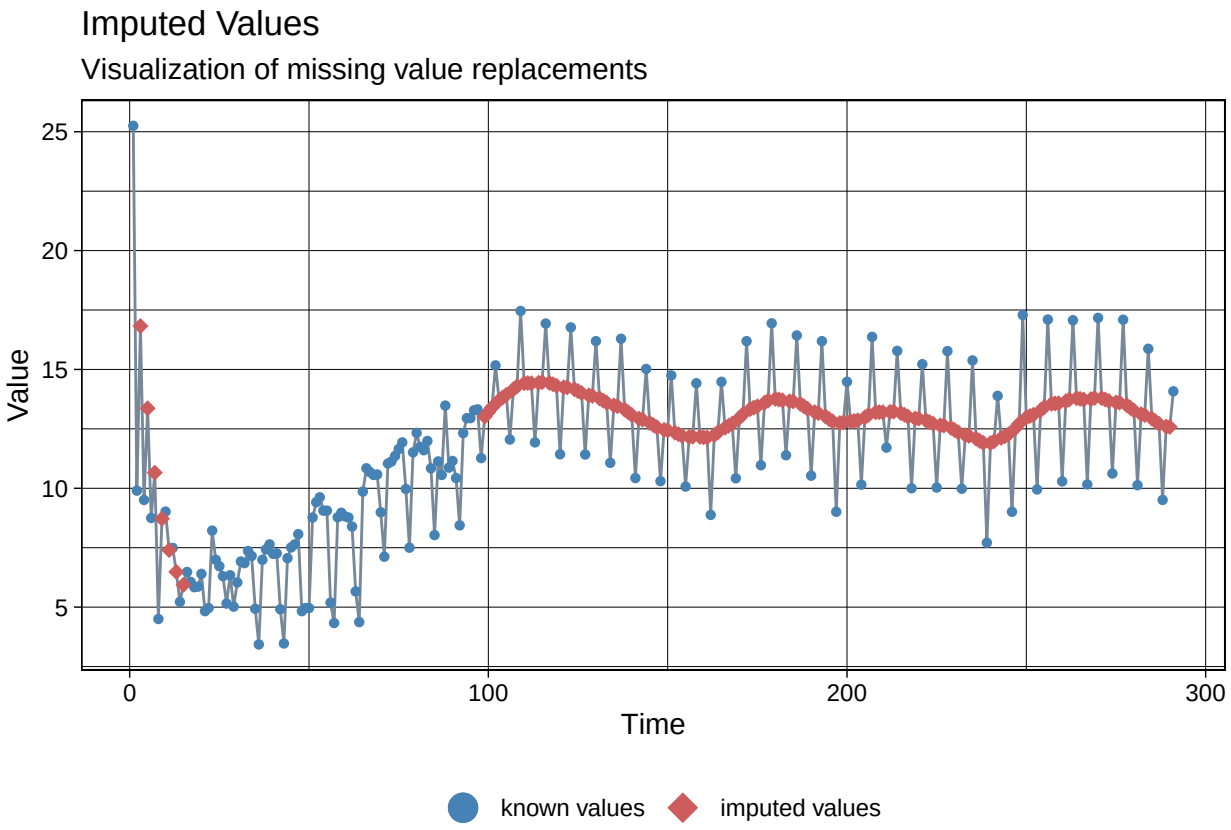
```r
EM3_t %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



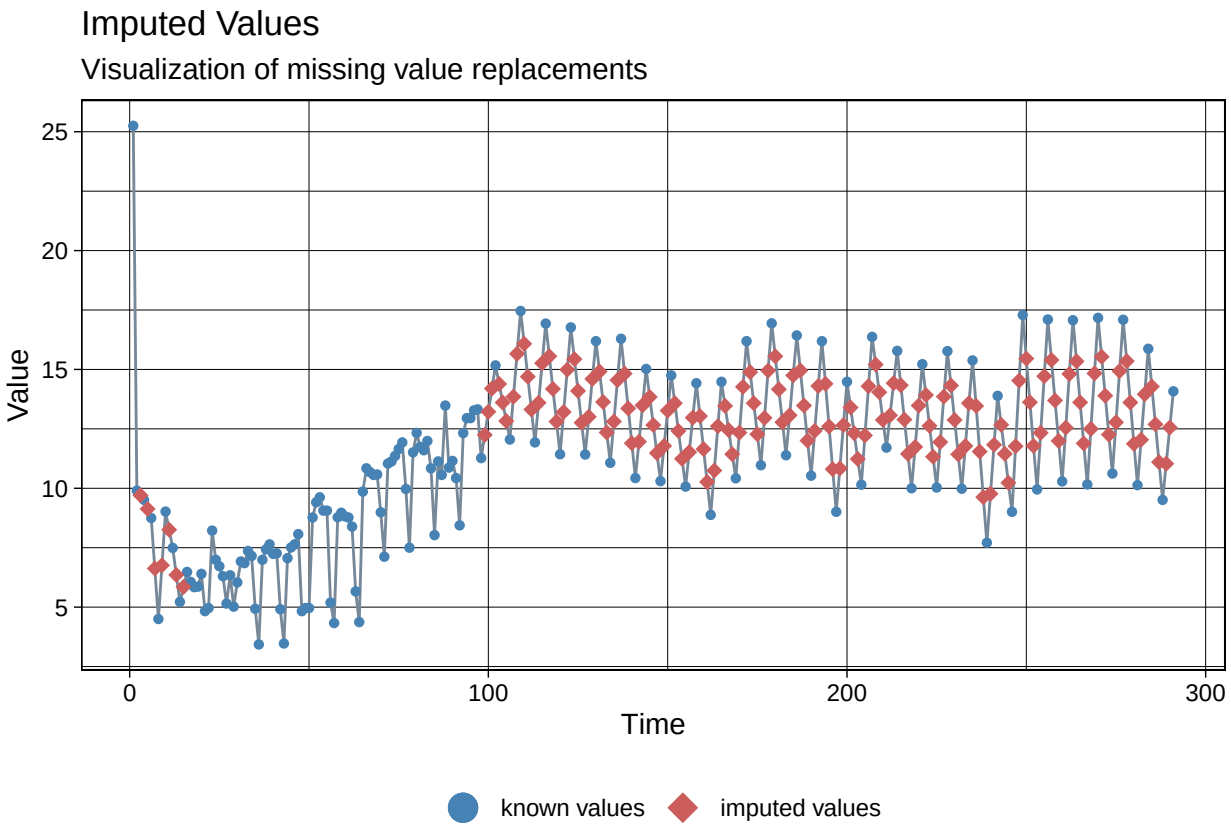
We impute the missing values following the principales stated for imputeTS. Thanks to this approach we almost double the amount of data for analysis by province (It was selected "na_seadec" due to it covers seasonality aspects -weekdays/weekends in our case-).

It is needed to transform the dataframe to a time series object.

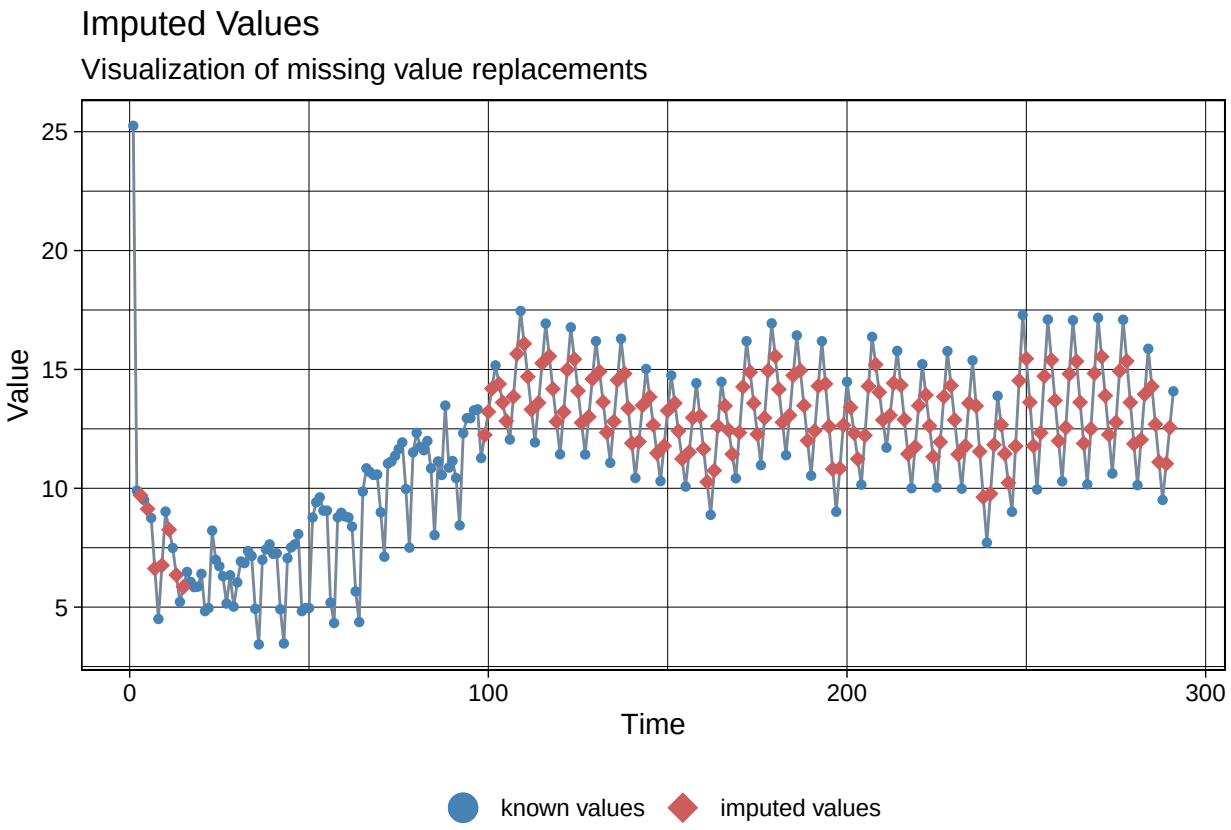
```
# Used to convert dataframe to ts object
library(xts)
EM3_t_ts<-xts(EM3_t[-1],EM3_t$Periodo)

# Impute the missing values with na_kalman, na_seadec, na_interpolation & na_seasplit
imp <- na_kalman(EM3_t_ts[,1])
ggplot_na_imputations(EM3_t_ts[,1], imp)
```



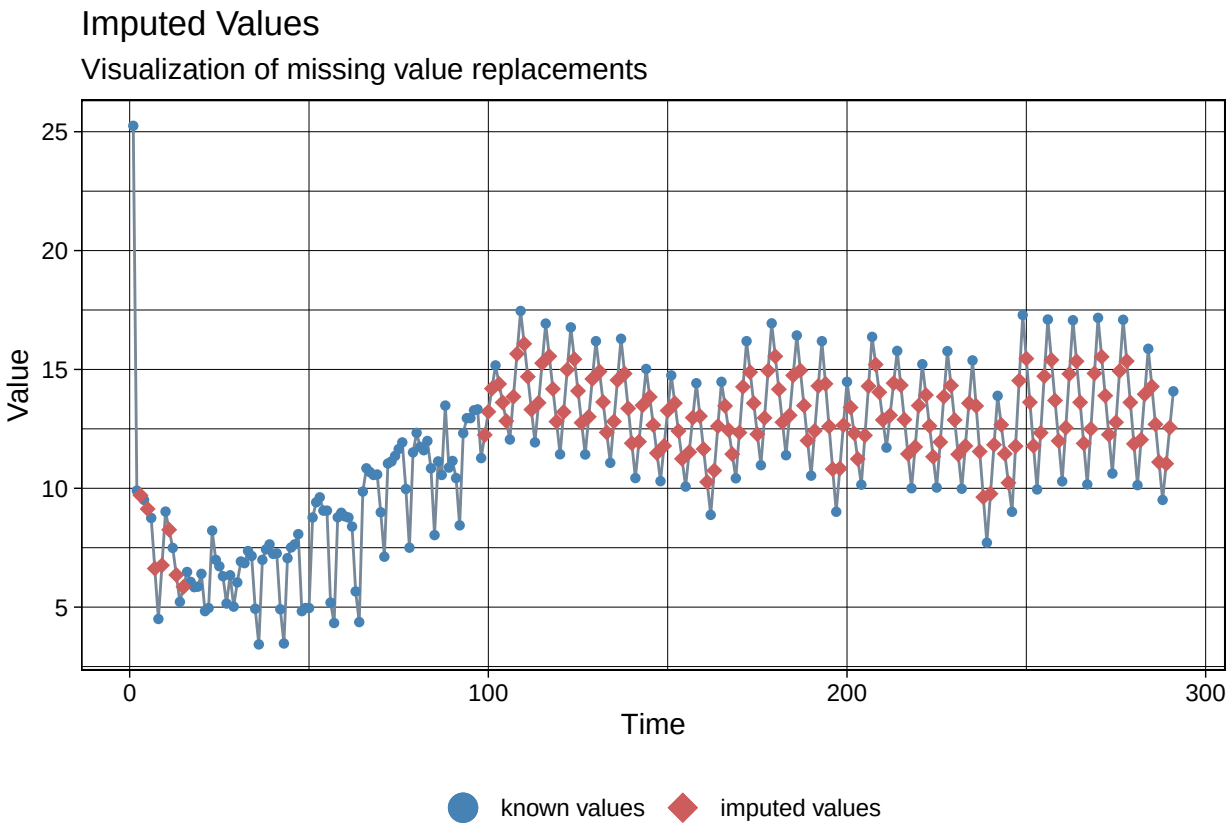

```
imp2 <- na_seadec(EM3_t_ts[,1])
ggplot_na_imputations(EM3_t_ts[,1], imp2)
```



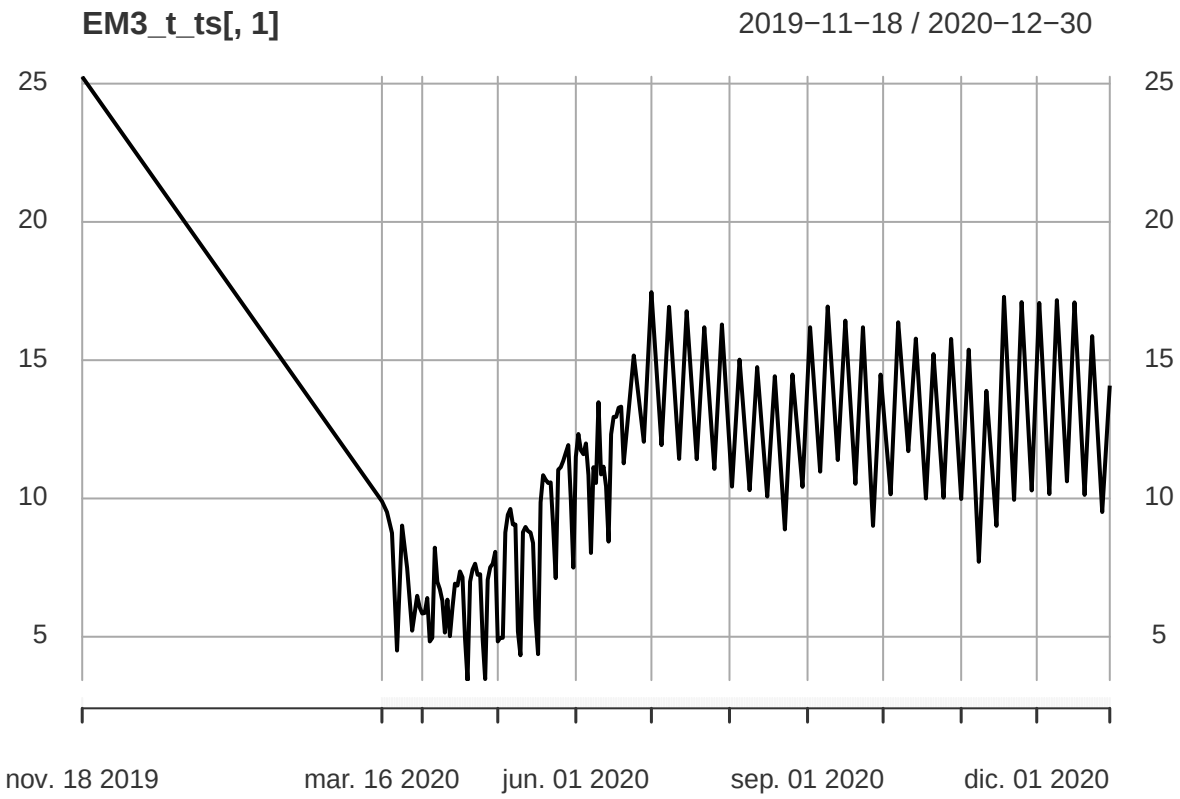

```
imp3 <- na_seasplit(EM3_t_ts[,1])
ggplot_na_imputations(EM3_t_ts[,1], imp3)
```



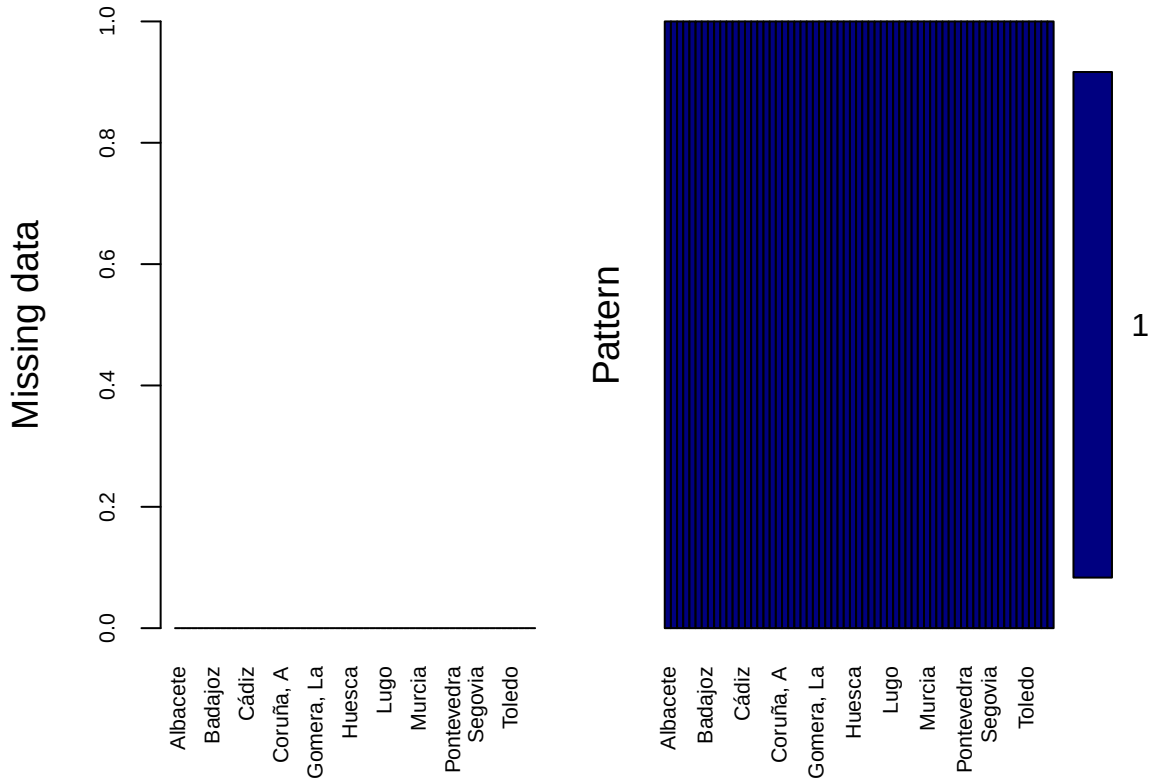

```
imp4 <- na_interpolation(EM3_t_ts[,1])
ggplot_na_imputations(EM3_t_ts[,1], imp4)
```




```
# We select na_seadec for the dataset
EM3_t_ts <- na_seadec(EM3_t_ts)
plot(EM3_t_ts[,1])
```




```
# We convert the time series object to a dataframe
EM3 <- ts_df(EM3_t_ts)

names(EM3)[names(EM3) == "id"] <- "Zonas.de.movilidad"
names(EM3)[names(EM3) == "time"] <- "Periodo"
names(EM3)[names(EM3) == "value"] <- "Total"

# Transpose dataframe
EM3_t<-dcast(EM3, Periodo~Zonas.de.movilidad, fill=NA)

# We check again missing values (result should be zero)
aggr(EM3_t[,-1], col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(EM3_t[,-1]), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##                Variable Count
##                Albacete     0
##       Alicante/Alacant     0
##                Almería     0
##            Araba/Álava     0
##                Asturias     0
##                   Ávila     0
##                Badajoz     0
##         Balears, Illes     0
##              Barcelona     0
##                Bizkaia     0
##                  Burgos     0
##                Cáceres     0
##                  Cádiz     0
##               Cantabria     0
##     Castellón/Castelló     0
##                   Ceuta     0
##             Ciudad Real     0
##                 Córdoba     0
##               Coruña, A     0
##                  Cuenca     0
##              Formentera     0
##           Fuerteventura     0
##                Gipuzkoa     0
```

```
##                   Girona     0
##               Gomera, La     0
##             Gran Canaria     0
##                  Granada     0
##              Guadalajara     0
##               Hierro, El     0
##                   Huelva     0
##                   Huesca     0
##                    Ibiza     0
##                     Jaén     0
##                Lanzarote     0
##                     León     0
##                   Lleida     0
##                     Lugo     0
##                   Madrid     0
##                   Málaga     0
##                  Mallorca     0
##                   Melilla     0
##                   Menorca     0
##                    Murcia     0
##                   Navarra     0
##                   Ourense     0
##                  Palencia     0
##                 Palma, La     0
##               Palmas, Las     0
##                Pontevedra     0
##                 Rioja, La     0
##                Salamanca     0
##  Santa Cruz de Tenerife     0
##                   Segovia     0
##                   Sevilla     0
##                     Soria     0
##                 Tarragona     0
##                  Tenerife     0
##                    Teruel     0
##                    Toledo     0
##         Valencia/València     0
##                Valladolid     0
##                    Zamora     0
##                  Zaragoza     0
```

```
EM3_t %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


```
head(str(EM3,vec.len=2))
```

```
## 'data.frame':    18333 obs. of  3 variables:
##  $ Zonas.de.movilidad: chr  "Albacete" "Albacete" ...
##  $ Periodo           : Date, format: "2019-11-18" "2020-03-16" ...
##  $ Total             : num  25.2 9.9 ...
```

```
## NULL
```

```
summary(EM3)
```

```
##  Zonas.de.movilidad    Periodo               Total        
##  Length:18333       Min.   :2019-11-18   Min.   : 0.83  
##  Class :character   1st Qu.:2020-05-26   1st Qu.:10.57  
##  Mode  :character   Median :2020-08-07   Median :14.13  
##                     Mean   :2020-08-06   Mean   :13.79  
##                     3rd Qu.:2020-10-19   3rd Qu.:17.06  
##                     Max.   :2020-12-30   Max.   :36.70
```

```
table(EM3$Zonas.de.movilidad)
```

```
##
##               Albacete       Alicante/Alacant                Almería 
##                    291                    291                    291 
##           Araba/Álava               Asturias                  Ávila 
##                    291                    291                    291 
##                Badajoz         Balears, Illes              Barcelona 
##                    291                    291                    291
```

```
##                Bizkaia                Burgos               Cáceres
##                    291                   291                   291
##                Cádiz              Cantabria    Castellón/Castelló
##                    291                   291                   291
##                Ceuta           Ciudad Real               Córdoba
##                    291                   291                   291
##            Coruña, A                Cuenca            Formentera
##                    291                   291                   291
##         Fuerteventura              Gipuzkoa                Girona
##                    291                   291                   291
##            Gomera, La          Gran Canaria               Granada
##                    291                   291                   291
##           Guadalajara           Hierro, El                Huelva
##                    291                   291                   291
##                Huesca                Ibiza                  Jaén
##                    291                   291                   291
##             Lanzarote                 León                Lleida
##                    291                   291                   291
##                 Lugo               Madrid                Málaga
##                    291                   291                   291
##              Mallorca              Melilla               Menorca
##                    291                   291                   291
##                Murcia              Navarra               Ourense
##                    291                   291                   291
##              Palencia           Palma, La          Palmas, Las
##                    291                   291                   291
##            Pontevedra           Rioja, La             Salamanca
##                    291                   291                   291
## Santa Cruz de Tenerife             Segovia               Sevilla
##                    291                   291                   291
##                 Soria           Tarragona              Tenerife
##                    291                   291                   291
##                Teruel              Toledo    Valencia/València
##                    291                   291                   291
##            Valladolid              Zamora              Zaragoza
##                    291                   291                   291
```

2.1.5 Google review

Here we have data from autonomus communities and provinces.

```
#Source Google
head(str(Google,vec.len=1))
```

```
## 'data.frame':    24242 obs. of  15 variables:
##  $ country_region_code                              : chr  "ES" ...
##  $ country_region                                   : chr  "Spain" ...
##  $ sub_region_1                                     : chr  "" ...
##  $ sub_region_2                                     : chr  "" ...
##  $ metro_area                                       : logi  NA ...
##  $ iso_3166_2_code                                  : chr  "" ...
##  $ census_fips_code                                 : logi  NA ...
##  $ place_id                                         : chr  "ChIJi7xhMnjjQgwR7KNoB5Qs7KY" ...
##  $ date                                             : chr  "15/02/2020" ...
##  $ retail_and_recreation_percent_change_from_baseline: int  2 2 ...
```

```
##  $ grocery_and_pharmacy_percent_change_from_baseline : int  -1 3 ...
##  $ parks_percent_change_from_baseline                : int  26 13 ...
##  $ transit_stations_percent_change_from_baseline     : int  8 5 ...
##  $ workplaces_percent_change_from_baseline           : int  0 -1 ...
##  $ residential_percent_change_from_baseline          : int  -2 -2 ...
```

```
## NULL
```

```
summary(Google)
```

```
##  country_region_code country_region     sub_region_1       sub_region_2
##  Length:24242        Length:24242       Length:24242       Length:24242
##  Class :character    Class :character   Class :character   Class :character
##  Mode  :character    Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  metro_area     iso_3166_2_code    census_fips_code   place_id
##  Mode:logical   Length:24242       Mode:logical     Length:24242
##  NA's:24242     Class :character   NA's:24242       Class :character
##                 Mode  :character                    Mode  :character
##
##
##
##
##      date           retail_and_recreation_percent_change_from_baseline
##  Length:24242       Min.   :-97.00
##  Class :character   1st Qu.:-53.00
##  Mode  :character   Median :-32.00
##                     Mean   :-36.42
##                     3rd Qu.:-17.00
##                     Max.   : 71.00
##                     NA's   :56
##  grocery_and_pharmacy_percent_change_from_baseline
##  Min.   :-96.000
##  1st Qu.:-18.000
##  Median : -4.000
##  Mean   : -9.973
##  3rd Qu.:  3.000
##  Max.   :194.000
##  NA's   :396
##  parks_percent_change_from_baseline
##  Min.   :-94.0000
##  1st Qu.:-30.0000
##  Median : -5.0000
##  Mean   : -0.0038
##  3rd Qu.: 22.0000
##  Max.   :543.0000
##  NA's   :305
##  transit_stations_percent_change_from_baseline
##  Min.   :-100.00
##  1st Qu.: -46.00
##  Median : -30.00
##  Mean   : -32.63
```

```
##  3rd Qu.: -16.00
##  Max.   : 177.00
##  NA's   :832
##  workplaces_percent_change_from_baseline
##  Min.   :-92.0
##  1st Qu.:-37.0
##  Median :-24.0
##  Mean   :-26.7
##  3rd Qu.:-13.0
##  Max.   : 55.0
##  NA's   :42
##  residential_percent_change_from_baseline
##  Min.   :-10.000
##  1st Qu.:  4.000
##  Median :  7.000
##  Mean   :  9.419
##  3rd Qu.: 13.000
##  Max.   : 48.000
##  NA's   :267
```

```
table(Google$sub_region_1)
```

```
##
##                              Andalusia              Aragon             Asturias
##                385               3465               1540                385
##    Balearic Islands      Basque Country      Canary Islands            Cantabria
##                385               1540               1155                385
##   Castile-La Mancha   Castile and LeÃ³n           Catalonia              Ceuta
##               2310               3850               1925                378
## Community of Madrid         Extremadura            Galicia            La Rioja
##                385               1155               1925                385
##             Melilla            Navarre    Region of Murcia Valencian Community
##                379                385                385               1540
```

```
table(Google$sub_region_2)
```

```
##
##                                  A CoruÃ±a                Ã\201lava
##                  7687                385                385
##                Ã\201vila             Albacete             Alicante
##                   385                385                385
##              AlmerÃa              Badajoz            Barcelona
##                   385                385                385
##                Biscay              Burgos            CÃ¡ceres
##                   385                385                385
##                CÃ¡diz             CÃ³rdoba           CastellÃ³n
##                   385                385                385
##           Ciudad Real              Cuenca            Gipuzkoa
##                   385                385                385
##               Girona             Granada          Guadalajara
##                   385                385                385
##               Huelva             Huesca               JaÃ©n
##                   385                385                385
##            Las Palmas              LeÃ³n              Lleida
##                   385                385                385
```

```
##                   Lugo                MÃ¡laga               Palencia
##                    385                   385                   385
##             Pontevedra    Province of Ourense              Salamanca
##                    385                   385                   385
## Santa Cruz de Tenerife               Segovia                Seville
##                    385                   385                   385
##                  Soria             Tarragona                 Teruel
##                    385                   385                   385
##                 Toledo              Valencia             Valladolid
##                    385                   385                   385
##                 Zamora             Zaragoza
##                    385                   385
```

```
table(Google$iso_3166_2_code)
```

```
##
##        ES-A ES-AB ES-AL ES-AN ES-AR ES-AS ES-AV  ES-B ES-BA ES-BI ES-BU  ES-C
##   385   385   385   385   385   385   385   385   385   385   385   385   385
## ES-CA ES-CB ES-CC ES-CE ES-CL ES-CM ES-CN ES-CO ES-CR ES-CS ES-CT ES-CU ES-EX
##   385   385   385   378   385   385   385   385   385   385   385   385   385
## ES-GA ES-GC ES-GI ES-GR ES-GU  ES-H ES-HU ES-IB  ES-J  ES-L ES-LE ES-LU ES-MA
##   385   385   385   385   385   385   385   385   385   385   385   385   385
## ES-MC ES-MD ES-ML ES-NC ES-OR  ES-P ES-PO ES-PV ES-RI ES-SA ES-SE ES-SG ES-SO
##   385   385   379   385   385   385   385   385   385   385   385   385   385
## ES-SS  ES-T ES-TE ES-TF ES-TO  ES-V ES-VA ES-VC ES-VI  ES-Z ES-ZA
##   385   385   385   385   385   385   385   385   385   385   385
```

2.1.6 Google autonomous-communities & provinces

We check data grouped by autonomous communities and provinces.

```
Google %>% group_by(sub_region_1) %>% tally()
```

```
## # A tibble: 20 x 2
##    sub_region_1              n
##    <chr>                 <int>
##  1 ""                      385
##  2 "Andalusia"            3465
##  3 "Aragon"               1540
##  4 "Asturias"              385
##  5 "Balearic Islands"      385
##  6 "Basque Country"       1540
##  7 "Canary Islands"       1155
##  8 "Cantabria"             385
##  9 "Castile-La Mancha"    2310
## 10 "Castile and LeÃ³n"    3850
## 11 "Catalonia"           1925
## 12 "Ceuta"                378
## 13 "Community of Madrid"  385
## 14 "Extremadura"         1155
## 15 "Galicia"             1925
## 16 "La Rioja"             385
## 17 "Melilla"              379
## 18 "Navarre"              385
## 19 "Region of Murcia"     385
## 20 "Valencian Community"  1540
```

```
Google %>% group_by(sub_region_1) %>% count(sub_region_2)
```

```
## # A tibble: 63 x 3
## # Groups:   sub_region_1 [20]
##    sub_region_1 sub_region_2     n
##    <chr>        <chr>        <int>
##  1 ""           ""             385
##  2 "Andalusia"  ""             385
##  3 "Andalusia"  "AlmerÃa"      385
##  4 "Andalusia"  "CÃ¡diz"       385
##  5 "Andalusia"  "CÃ³rdoba"     385
##  6 "Andalusia"  "Granada"      385
##  7 "Andalusia"  "Huelva"       385
##  8 "Andalusia"  "JaÃ©n"        385
##  9 "Andalusia"  "MÃ¡laga"      385
## 10 "Andalusia"  "Seville"      385
## # ... with 53 more rows
```

In Spain there are **autonomous communities (AC)** and **autonomous cities (C)** that are considered as **provinces (Pr)**. This is the case for:

- AC - Asturias, Principality - Pr - Asturias
- AC - Balears, Illes - Pr - Balears, Illes
- AC - Cantabria - Pr - Cantabria
- AC - Madrid, Community - Pr - Madrid
- AC - Murcia, Region - Pr- Murcia
- AC - Navarra, Foral Community - Pr - Navarra
- AC - Rioja, La - Pr - Rioja, La
- C - Ceuta - C/Pr - Ceuta
- C - Melilla - C/Pr - Melilla

In this data set, the empty values in the "sub_region_2" column, for the autonomous communities mentinoed, will be replaced by the value contained in the "sub_region_1" column (A). Also we are going to modify the names of the provinces that have special characters in order to adopt the INE standards (B). See note.

Note The following links states the provinces in Spain INE CCAA and its ISO codes are going to be used as tables of referencence.

```
# Modidication provinces - A
Google$sub_region_2[Google$sub_region_1=="Balearic Islands"] <- "Balears, Illes"
Google$iso_3166_2_code[Google$sub_region_2=="Balears, Illes"] <- "PM"

Google$sub_region_2[Google$sub_region_1=="Asturias"] <- "Asturias"
Google$iso_3166_2_code[Google$sub_region_2=="Asturias"] <- "O"

Google$sub_region_2[Google$sub_region_1=="Cantabria"] <- "Cantabria"
Google$iso_3166_2_code[Google$sub_region_2=="Cantabria"] <- "S"

Google$sub_region_2[Google$sub_region_1=="Community of Madrid"] <- "Madrid"
Google$iso_3166_2_code[Google$sub_region_2=="Madrid"] <- "M"

Google$sub_region_2[Google$sub_region_1=="Region of Murcia"] <- "Murcia"
Google$iso_3166_2_code[Google$sub_region_2=="Murcia"] <- "MU"

Google$sub_region_2[Google$sub_region_1=="Navarre"] <- "Navarra"
Google$iso_3166_2_code[Google$sub_region_2=="Navarra"] <- "NA"
```

```
Google$sub_region_2[Google$sub_region_1=="La Rioja"] <- "Rioja, La"
Google$iso_3166_2_code[Google$sub_region_2=="Rioja, La"] <- "LO"

Google$sub_region_2[Google$sub_region_1=="Ceuta"] <- "Ceuta"
Google$iso_3166_2_code[Google$sub_region_2=="Ceuta"] <- "CE"

Google$sub_region_2[Google$sub_region_1=="Melilla"] <- "Melilla"
Google$iso_3166_2_code[Google$sub_region_2=="Melilla"] <- "ML"

# Modidication provinces - B
Google$sub_region_2[Google$sub_region_2=="A CoruÃ±a"]<-"Coruña, A"
Google$sub_region_2[Google$sub_region_2=="Ã\u0081lava"]<-"Araba/Álava"
Google$sub_region_2[Google$sub_region_2=="Ã\u0081vila"]<-"Ávila"
#Google$sub_region_2[Google$sub_region_2=="Albacete"]<-"Albacete"
Google$sub_region_2[Google$sub_region_2=="Alicante"]<-"Alicante/Alacant"
#Google$sub_region_2[Google$sub_region_2=="AlmerÃa"]<-"Almería"
#Google$sub_region_2[Google$sub_region_2=="Asturias"]<-"Asturias"
#Google$sub_region_2[Google$sub_region_2=="Badajoz"]<-"Badajoz"
#Google$sub_region_2[Google$sub_region_2=="Balears, Illes"]<-"Balears, Illes"
#Google$sub_region_2[Google$sub_region_2=="Barcelona"]<-"Barcelona"
Google$sub_region_2[Google$sub_region_2=="Biscay"]<-"Bizkaia"
#Google$sub_region_2[Google$sub_region_2=="Burgos"]<-"Burgos"
Google$sub_region_2[Google$sub_region_2=="CÃ¡ceres"]<-"Cáceres"
Google$sub_region_2[Google$sub_region_2=="CÃ¡diz"]<-"Cádiz"
Google$sub_region_2[Google$sub_region_2=="CÃ³rdoba"]<-"Córdoba"
#Google$sub_region_2[Google$sub_region_2=="Cantabria"]<-"Cantabria"
Google$sub_region_2[Google$sub_region_2=="CastellÃ³n"]<-"Castellón/Castelló"
#Google$sub_region_2[Google$sub_region_2=="Ceuta"]<-"Ceuta"
#Google$sub_region_2[Google$sub_region_2=="Ciudad Real"]<-"Ciudad Real"
#Google$sub_region_2[Google$sub_region_2=="Cuenca"]<-"Cuenca"
#Google$sub_region_2[Google$sub_region_2=="Gipuzkoa"]<-"Gipuzkoa"
#Google$sub_region_2[Google$sub_region_2=="Girona"]<-"Girona"
#Google$sub_region_2[Google$sub_region_2=="Granada"]<-"Granada"
#Google$sub_region_2[Google$sub_region_2=="Guadalajara"]<-"Guadalajara"
#Google$sub_region_2[Google$sub_region_2=="Huelva"]<-"Huelva"
#Google$sub_region_2[Google$sub_region_2=="Huesca"]<-"Huesca"
Google$sub_region_2[Google$sub_region_2=="JaÃ©n"]<-"Jaén"
Google$sub_region_2[Google$sub_region_2=="Las Palmas"]<-"Palmas, Las"
Google$sub_region_2[Google$sub_region_2=="LeÃ³n"]<-"León"
#Google$sub_region_2[Google$sub_region_2=="Lleida"]<-"Lleida"
#Google$sub_region_2[Google$sub_region_2=="Lugo"]<-"Lugo"
Google$sub_region_2[Google$sub_region_2=="MÃ¡laga"]<-"Málaga"
#Google$sub_region_2[Google$sub_region_2=="Madrid"]<-"Madrid"
#Google$sub_region_2[Google$sub_region_2=="Melilla"]<-"Melilla"
#Google$sub_region_2[Google$sub_region_2=="Murcia"]<-"Murcia"
#Google$sub_region_2[Google$sub_region_2=="Navarra"]<-"Navarra"
#Google$sub_region_2[Google$sub_region_2=="Palencia"]<-"Palencia"
#Google$sub_region_2[Google$sub_region_2=="Pontevedra"]<-"Pontevedra"
Google$sub_region_2[Google$sub_region_2=="Province of Ourense"]<-"Ourense"
#Google$sub_region_2[Google$sub_region_2=="Rioja, La"]<-"Rioja, La"
#Google$sub_region_2[Google$sub_region_2=="Salamanca"]<-"Salamanca"
#Google$sub_region_2[Google$sub_region_2=="Santa Cruz de Tenerife"]<-"Santa Cruz de Tenerife"
#Google$sub_region_2[Google$sub_region_2=="Segovia"]<-"Segovia"
```

```
Google$sub_region_2[Google$sub_region_2=="Seville"]<-"Sevilla"
#Google$sub_region_2[Google$sub_region_2=="Soria"]<-"Soria"
#Google$sub_region_2[Google$sub_region_2=="Tarragona"]<-"Tarragona"
#Google$sub_region_2[Google$sub_region_2=="Teruel"]<-"Teruel"
#Google$sub_region_2[Google$sub_region_2=="Toledo"]<-"Toledo"
Google$sub_region_2[Google$sub_region_2=="Valencia"]<-"Valencia/València"
#Google$sub_region_2[Google$sub_region_2=="Valladolid"]<-"Valladolid"
#Google$sub_region_2[Google$sub_region_2=="Zamora"]<-"Zamora"
#Google$sub_region_2[Google$sub_region_2=="Zaragoza"]<-"Zaragoza"
Google$sub_region_2 <- with(Google, ifelse(grepl("^Almer", sub_region_2),
                                           "Almería", sub_region_2))

table(Google$sub_region_2)
```

```
##
##                                Albacete       Alicante/Alacant
##                    4235             385                    385
##                 Almería       Araba/Álava               Asturias
##                     385             385                    385
##                   Ávila          Badajoz         Balears, Illes
##                     385             385                    385
##               Barcelona         Bizkaia                 Burgos
##                     385             385                    385
##                 Cáceres          Cádiz              Cantabria
##                     385             385                    385
##     Castellón/Castelló           Ceuta            Ciudad Real
##                     385             378                    385
##                 Córdoba       Coruña, A                 Cuenca
##                     385             385                    385
##                Gipuzkoa         Girona                Granada
##                     385             385                    385
##             Guadalajara         Huelva                 Huesca
##                     385             385                    385
##                   Jaén           León                 Lleida
##                     385             385                    385
##                   Lugo         Madrid                 Málaga
##                     385             385                    385
##                 Melilla         Murcia                Navarra
##                     379             385                    385
##                 Ourense       Palencia           Palmas, Las
##                     385             385                    385
##              Pontevedra       Rioja, La             Salamanca
##                     385             385                    385
## Santa Cruz de Tenerife        Segovia                Sevilla
##                     385             385                    385
##                   Soria       Tarragona                Teruel
##                     385             385                    385
##                  Toledo Valencia/València             Valladolid
##                     385             385                    385
##                  Zamora        Zaragoza
##                     385             385
```

```
table(Google$iso_3166_2_code)
```

```
##
##          CE  ES-A ES-AB ES-AL ES-AN ES-AR ES-AV  ES-B ES-BA ES-BI ES-BU  ES-C
##   385   378   385   385   385   385   385   385   385   385   385   385   385
## ES-CA ES-CC ES-CL ES-CM ES-CN ES-CO ES-CR ES-CS ES-CT ES-CU ES-EX ES-GA ES-GC
##   385   385   385   385   385   385   385   385   385   385   385   385   385
## ES-GI ES-GR ES-GU  ES-H ES-HU  ES-J  ES-L ES-LE ES-LU ES-MA ES-OR  ES-P ES-PO
##   385   385   385   385   385   385   385   385   385   385   385   385   385
## ES-PV ES-SA ES-SE ES-SG ES-SO ES-SS  ES-T ES-TE ES-TF ES-TO  ES-V ES-VA ES-VC
##   385   385   385   385   385   385   385   385   385   385   385   385   385
## ES-VI  ES-Z ES-ZA    LO     M    ML    MU    NA     O    PM     S
##   385   385   385   385   385   379   385   385   385   385   385
```

2.1.7 Google data transformation

We are going to **transform / eliminate**:

- A - Rows with "na" / "" in "sub_region_1" and "sub_region_2" columns are eliminated.
- B - Date column is transformed from "character" to "date".
- C - Some columns are eliminated due to they are not adding value or they contain blanks (country_region_code, country_region, metro_area, census_fips_code, pace_id).
- D - "ES-" is elimianted from "iso_3166_2_code" column.

```
# Transform / eliminate A
Google <- filter(Google, sub_region_1 != "", sub_region_2 != "" )

# Transform / eliminate B
Google$date <- as.Date(Google$date ,format="%d/%m/%Y")

# Transform / eliminate C
Google<-within(Google, rm(country_region_code,
                  country_region,
                  metro_area,
                  census_fips_code,
                  place_id))

# Transform / eliminate D
Google$iso_3166_2_code <- gsub("ES-", "", Google$iso_3166_2_code)

#Google$retail_and_recreation_percent_change_from_baseline <- as.numeric(Google$retail_and_recreation_p
#Google$grocery_and_pharmacy_percent_change_from_baseline <- as.numeric(Google$grocery_and_pharmacy_per
#Google$parks_percent_change_from_baseline <- as.numeric(Google$parks_percent_change_from_baseline)
#Google$transit_stations_percent_change_from_baseline <- as.numeric(Google$transit_stations_percent_cha
#Google$workplaces_percent_change_from_baseline <- as.numeric(Google$workplaces_percent_change_from_bas
#Google$residential_percent_change_from_baseline <- as.numeric(Google$residential_percent_change_from_b

head(Google,5)
```

```
##   sub_region_1 sub_region_2 iso_3166_2_code       date
## 1    Andalusia      Almería              AL 2020-02-15
## 2    Andalusia      Almería              AL 2020-02-16
## 3    Andalusia      Almería              AL 2020-02-17
## 4    Andalusia      Almería              AL 2020-02-18
## 5    Andalusia      Almería              AL 2020-02-19
```

```
##   retail_and_recreation_percent_change_from_baseline
## 1                                                  5
## 2                                                 -2
## 3                                                  0
## 4                                                 -3
## 5                                                 -1
##   grocery_and_pharmacy_percent_change_from_baseline
## 1                                                -3
## 2                                                 0
## 3                                                -2
## 4                                                -3
## 5                                                -3
##   parks_percent_change_from_baseline
## 1                                 40
## 2                                 -2
## 3                                  3
## 4                                 -2
## 5                                  3
##   transit_stations_percent_change_from_baseline
## 1                                            10
## 2                                             1
## 3                                             5
## 4                                             5
## 5                                             4
##   workplaces_percent_change_from_baseline
## 1                                       1
## 2                                       1
## 3                                       3
## 4                                       3
## 5                                       3
##   residential_percent_change_from_baseline
## 1                                       -2
## 2                                       -1
## 3                                       -1
## 4                                        0
## 5                                        0
```

```
table(Google$sub_region_2)
```

```
##
##               Albacete       Alicante/Alacant                Almería
##                    385                    385                    385
##            Araba/Álava               Asturias                  Ávila
##                    385                    385                    385
##                Badajoz         Balears, Illes              Barcelona
##                    385                    385                    385
##                Bizkaia                 Burgos                Cáceres
##                    385                    385                    385
##                  Cádiz              Cantabria     Castellón/Castelló
##                    385                    385                    385
##                  Ceuta           Ciudad Real                Córdoba
##                    378                    385                    385
##              Coruña, A                 Cuenca              Gipuzkoa
##                    385                    385                    385
##                 Girona                Granada           Guadalajara
```

```
##                  385                  385                  385
##               Huelva               Huesca                Jaén
##                  385                  385                  385
##                León               Lleida                Lugo
##                  385                  385                  385
##               Madrid               Málaga              Melilla
##                  385                  385                  379
##               Murcia              Navarra              Ourense
##                  385                  385                  385
##             Palencia         Palmas, Las           Pontevedra
##                  385                  385                  385
##           Rioja, La           Salamanca Santa Cruz de Tenerife
##                  385                  385                  385
##             Segovia              Sevilla               Soria
##                  385                  385                  385
##           Tarragona              Teruel               Toledo
##                  385                  385                  385
##    Valencia/València           Valladolid               Zamora
##                  385                  385                  385
##             Zaragoza
##                  385
```

```
table(Google$iso_3166_2_code)
```

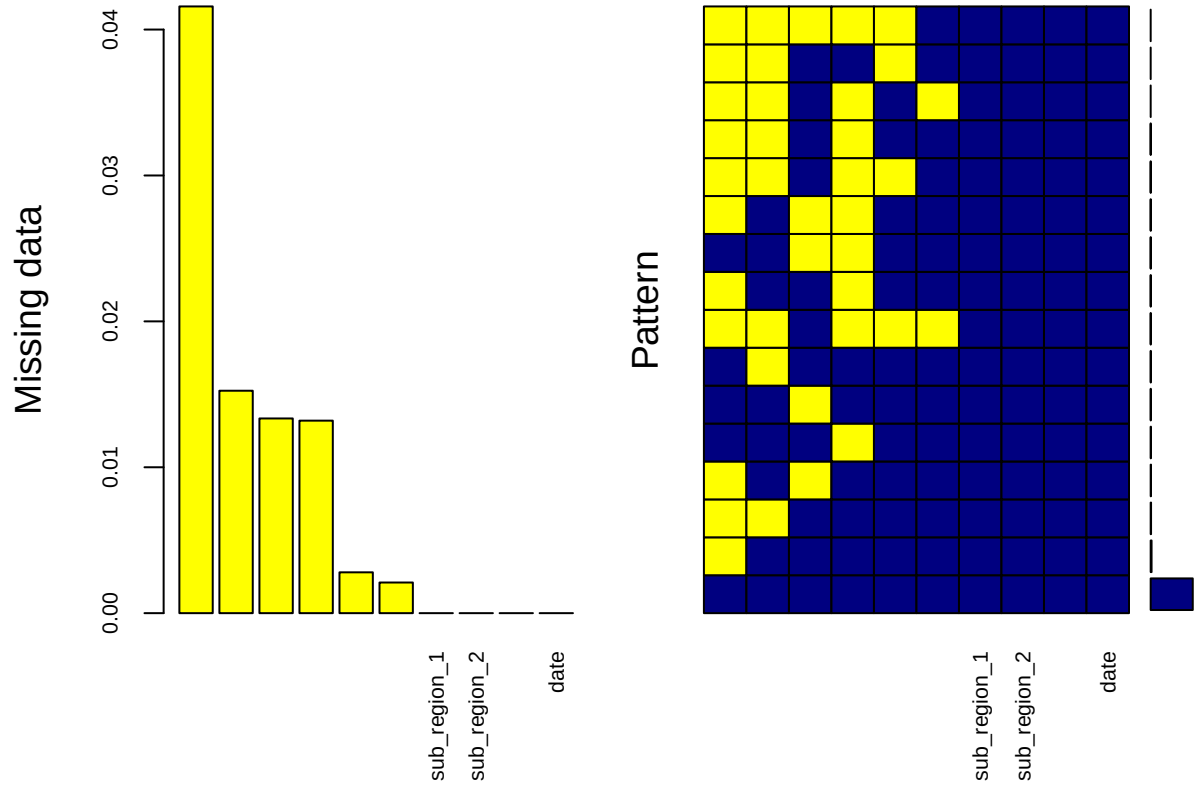
```
##
##   A  AB  AL  AV   B  BA  BI  BU   C  CA  CC  CE  CO  CR  CS  CU  GC  GI  GR  GU
## 385 385 385 385 385 385 385 385 385 385 385 378 385 385 385 385 385 385 385 385
##   H  HU   J   L  LE  LO  LU   M  MA  ML  MU  NA   O  OR   P  PM  PO   S  SA  SE
## 385 385 385 385 385 385 385 385 385 379 385 385 385 385 385 385 385 385 385 385
##  SG  SO  SS   T  TE  TF  TO   V  VA  VI   Z  ZA
## 385 385 385 385 385 385 385 385 385 385 385 385
```

```
#unique(Google$sub_region_2)
#unique(EM3$Zonas.de.movilidad)
```

2.1.8 Google review missing values & impute

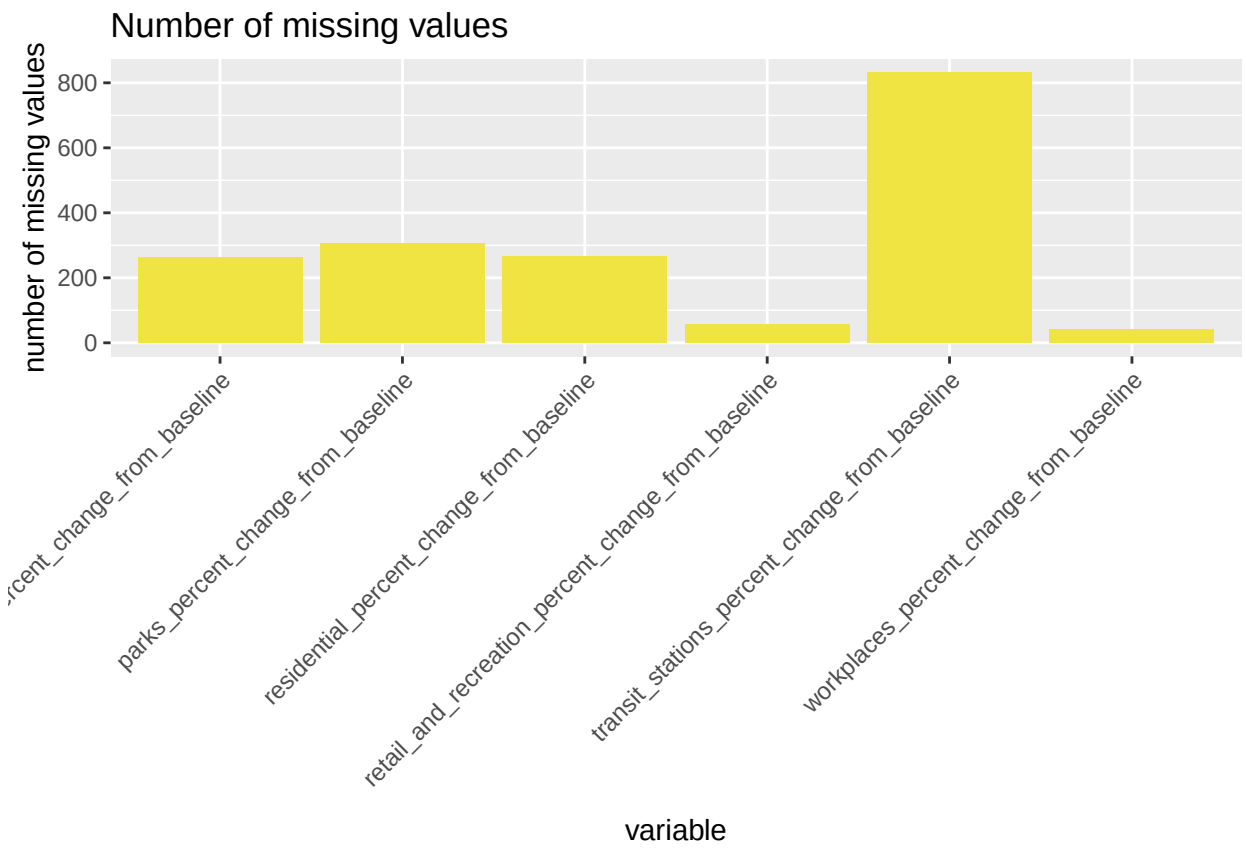
We check missing values.

```
aggr(Google, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```



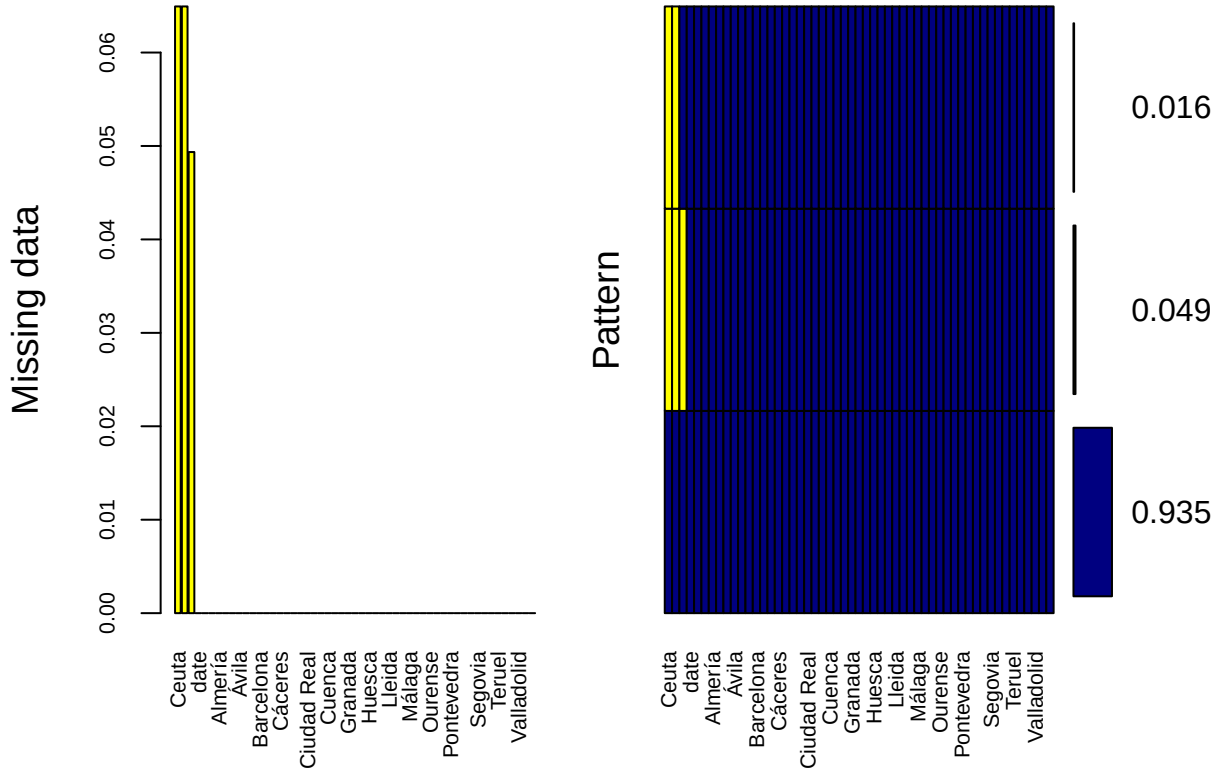

```
##
##  Variables sorted by number of missings:
##                                              Variable       Count
##       transit_stations_percent_change_from_baseline 0.041585445
##                    parks_percent_change_from_baseline 0.015244664
##             residential_percent_change_from_baseline 0.013345329
##    grocery_and_pharmacy_percent_change_from_baseline 0.013195382
##  retail_and_recreation_percent_change_from_baseline 0.002799020
##              workplaces_percent_change_from_baseline 0.002099265
##                                           sub_region_1 0.000000000
##                                           sub_region_2 0.000000000
##                                        iso_3166_2_code 0.000000000
##                                                   date 0.000000000
```

```
Google %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



We generate 6 new dataframes from the 6 features stated in order to imput missing values using the approach "imputeTS".

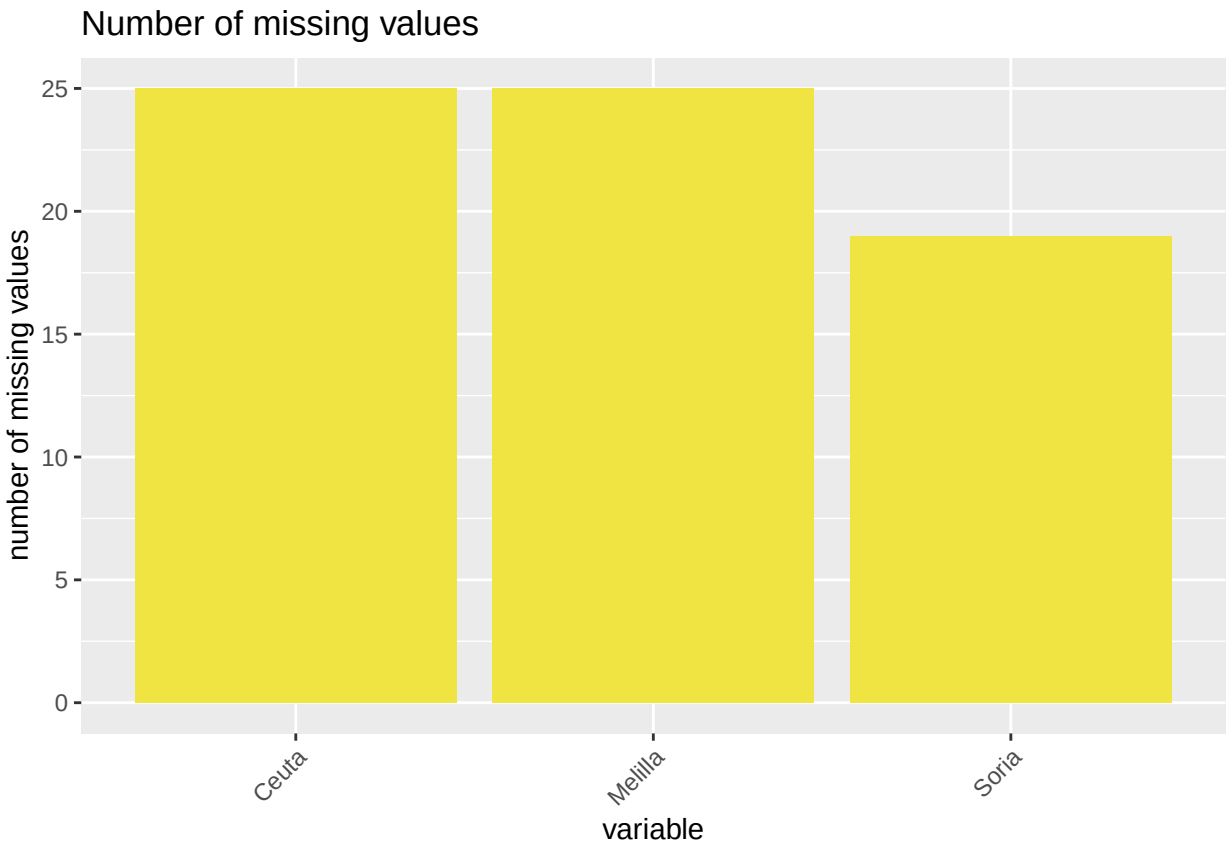
```
# Transpose dataframe
Google_retail<-Google[c(2,4,5)]
Google_t_retail<-dcast(Google_retail, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_retail, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google_t_retail), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```



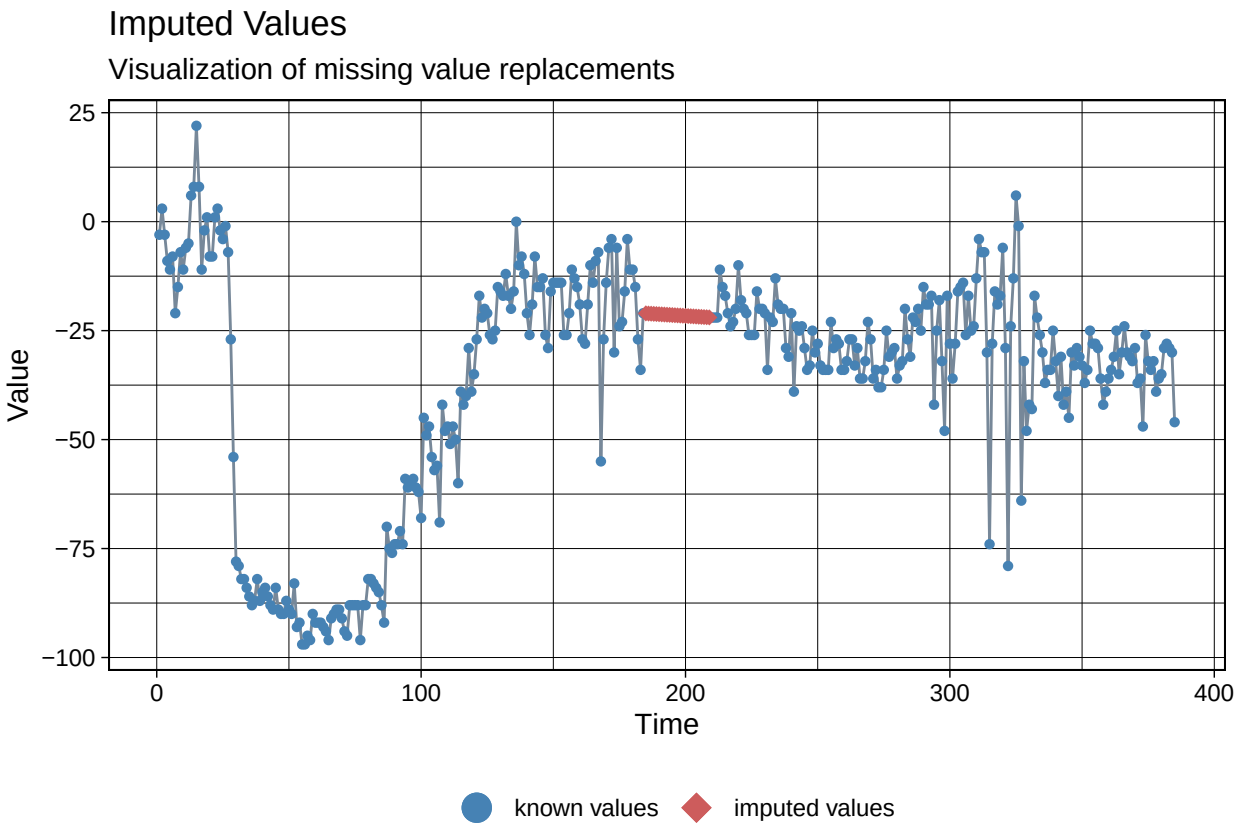

```
##
##  Variables sorted by number of missings:
##                 Variable      Count
##                    Ceuta 0.06493506
##                  Melilla 0.06493506
##                    Soria 0.04935065
##                     date 0.00000000
##                 Albacete 0.00000000
##        Alicante/Alacant 0.00000000
##                  Almería 0.00000000
##              Araba/Álava 0.00000000
##                 Asturias 0.00000000
##                    Ávila 0.00000000
##                  Badajoz 0.00000000
##           Balears, Illes 0.00000000
##                Barcelona 0.00000000
##                  Bizkaia 0.00000000
##                   Burgos 0.00000000
##                  Cáceres 0.00000000
##                    Cádiz 0.00000000
##                Cantabria 0.00000000
##      Castellón/Castelló 0.00000000
##              Ciudad Real 0.00000000
##                  Córdoba 0.00000000
##                Coruña, A 0.00000000
##                  Cuenca 0.00000000
```

```
##                Gipuzkoa 0.00000000
##                  Girona 0.00000000
##                 Granada 0.00000000
##              Guadalajara 0.00000000
##                  Huelva 0.00000000
##                  Huesca 0.00000000
##                    Jaén 0.00000000
##                    León 0.00000000
##                  Lleida 0.00000000
##                    Lugo 0.00000000
##                  Madrid 0.00000000
##                  Málaga 0.00000000
##                  Murcia 0.00000000
##                 Navarra 0.00000000
##                 Ourense 0.00000000
##                Palencia 0.00000000
##             Palmas, Las 0.00000000
##              Pontevedra 0.00000000
##                Rioja, La 0.00000000
##                Salamanca 0.00000000
##  Santa Cruz de Tenerife 0.00000000
##                 Segovia 0.00000000
##                 Sevilla 0.00000000
##               Tarragona 0.00000000
##                  Teruel 0.00000000
##                  Toledo 0.00000000
##       Valencia/València 0.00000000
##              Valladolid 0.00000000
##                  Zamora 0.00000000
##                Zaragoza 0.00000000
```

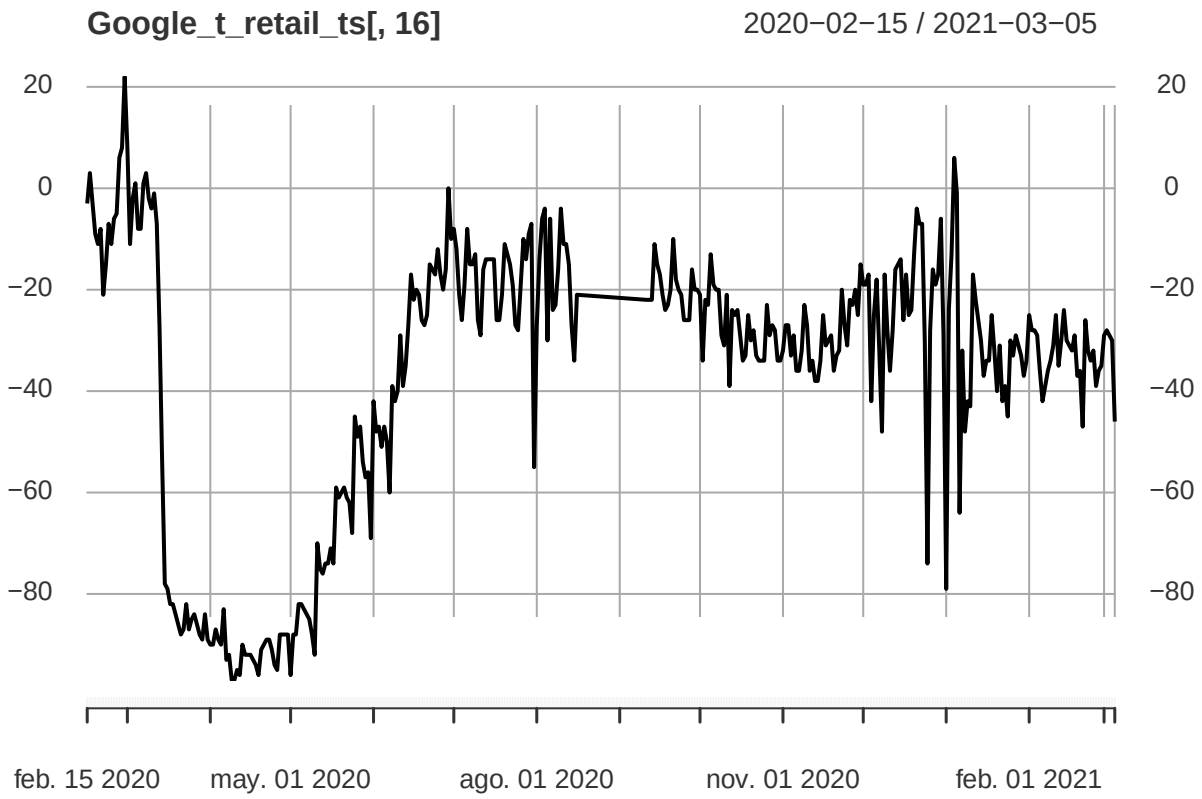
```
Google_t_retail %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



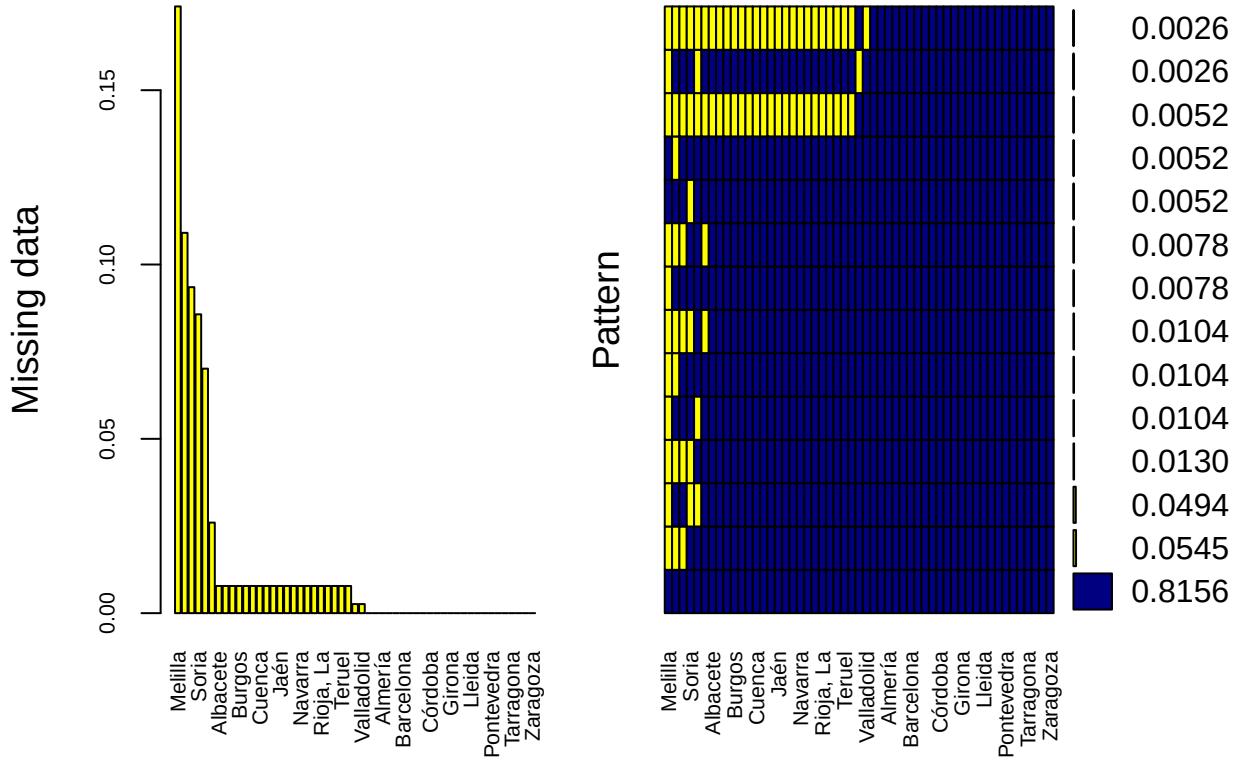

```
# Convert dataframe to ts object
Google_t_retail_ts<-xts(Google_t_retail[-1],Google_t_retail$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp5 <- na_seadec(Google_t_retail_ts[,16])
ggplot_na_imputations(Google_t_retail_ts[,16], imp5)
```




```
# We select na_seadec for the dataset
Google_t_retail_ts <- na_seadec(Google_t_retail_ts)
plot(Google_t_retail_ts[,16])
```



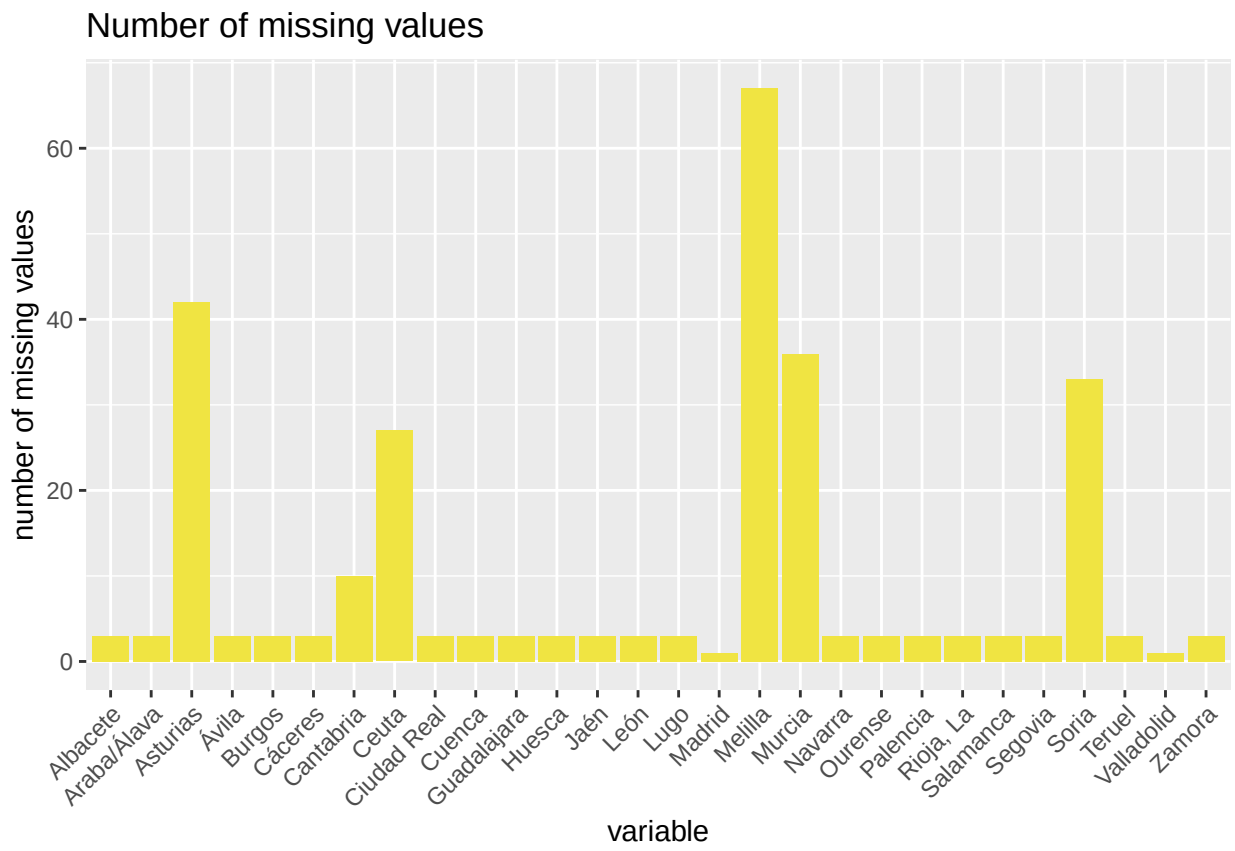

```
# We convert the time series object to a dataframe
Google_retail <- ts_df(Google_t_retail_ts)

names(Google_retail)[names(Google_retail) == "id"] <- "sub_region_2"
names(Google_retail)[names(Google_retail) == "time"] <- "Date"
names(Google_retail)[names(Google_retail) == "value"] <- "retail_and_recreation_percent_change_from_bas

######################################################################
# Transpose dataframe
Google_grocery<-Google[c(2,4,6)]
Google_t_grocery<-dcast(Google_grocery, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_grocery, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google_t_grocery), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```



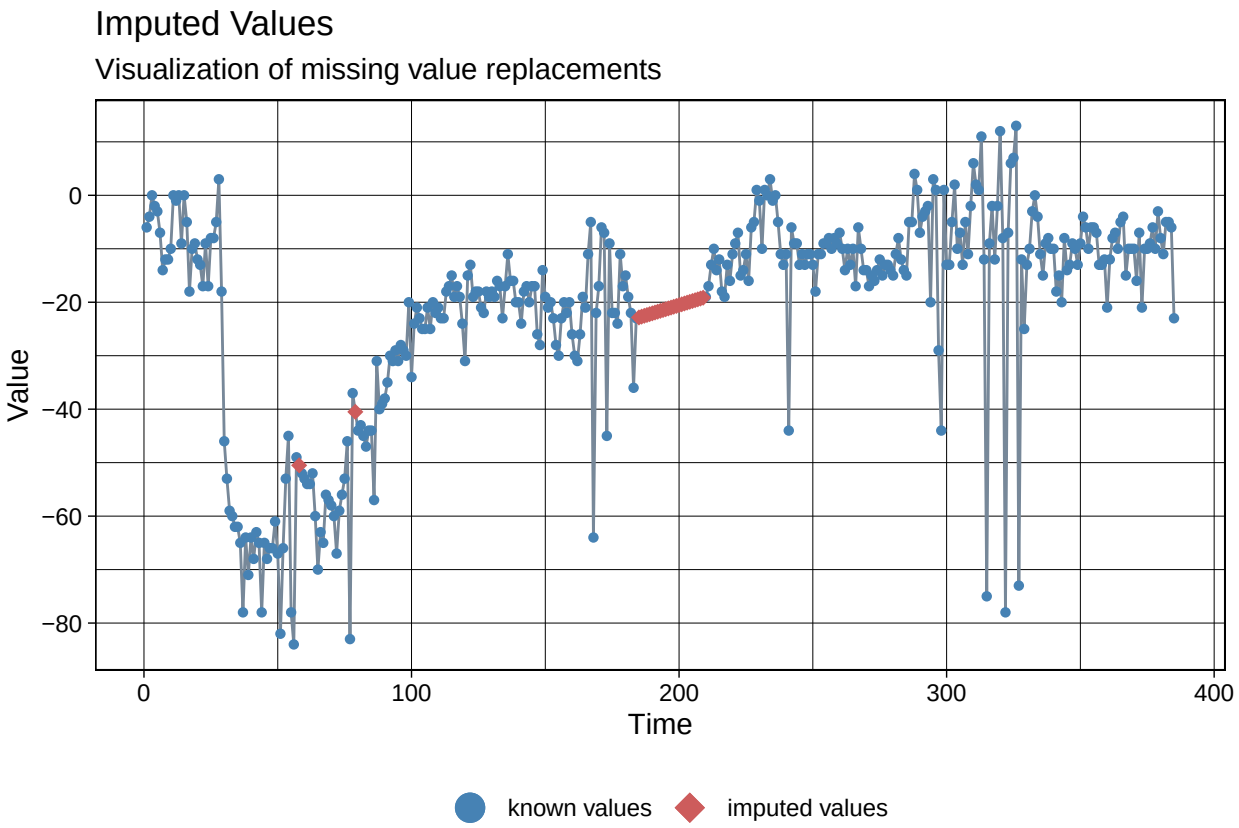

```
##
##  Variables sorted by number of missings:
##                 Variable       Count
##                  Melilla 0.174025974
##                 Asturias 0.109090909
##                   Murcia 0.093506494
##                    Soria 0.085714286
##                    Ceuta 0.070129870
##                Cantabria 0.025974026
##                 Albacete 0.007792208
##              Araba/Álava 0.007792208
##                    Ávila 0.007792208
##                   Burgos 0.007792208
##                  Cáceres 0.007792208
##              Ciudad Real 0.007792208
##                   Cuenca 0.007792208
##              Guadalajara 0.007792208
##                   Huesca 0.007792208
##                     Jaén 0.007792208
##                     León 0.007792208
##                     Lugo 0.007792208
##                  Navarra 0.007792208
##                  Ourense 0.007792208
##                 Palencia 0.007792208
##                Rioja, La 0.007792208
##                Salamanca 0.007792208
```

```
##                 Segovia 0.007792208
##                  Teruel 0.007792208
##                  Zamora 0.007792208
##                  Madrid 0.002597403
##              Valladolid 0.002597403
##                    date 0.000000000
##       Alicante/Alacant 0.000000000
##                 Almería 0.000000000
##                 Badajoz 0.000000000
##         Balears, Illes 0.000000000
##               Barcelona 0.000000000
##                 Bizkaia 0.000000000
##                   Cádiz 0.000000000
##     Castellón/Castelló 0.000000000
##                 Córdoba 0.000000000
##               Coruña, A 0.000000000
##                Gipuzkoa 0.000000000
##                  Girona 0.000000000
##                 Granada 0.000000000
##                  Huelva 0.000000000
##                  Lleida 0.000000000
##                  Málaga 0.000000000
##             Palmas, Las 0.000000000
##              Pontevedra 0.000000000
##  Santa Cruz de Tenerife 0.000000000
##                 Sevilla 0.000000000
##               Tarragona 0.000000000
##                  Toledo 0.000000000
##       Valencia/València 0.000000000
##                Zaragoza 0.000000000
```

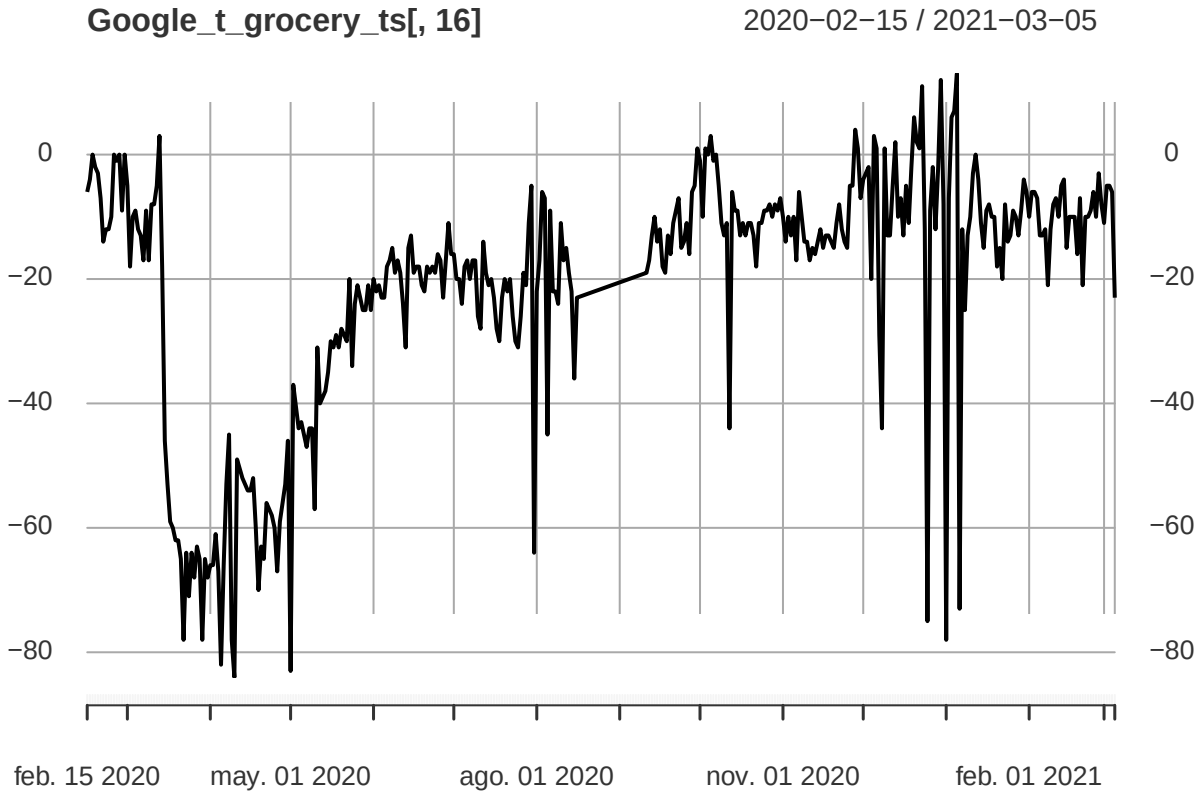
```
Google_t_grocery %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



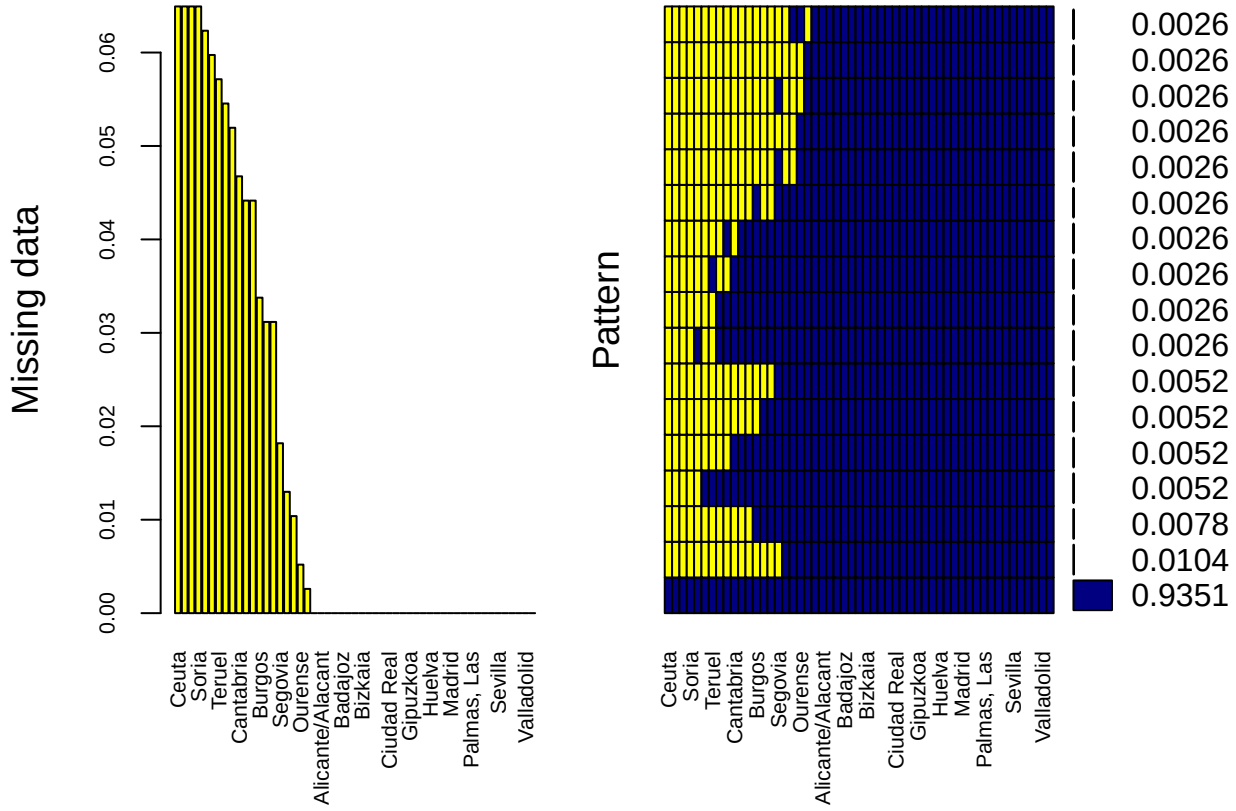

```
# Convert dataframe to ts object
Google_t_grocery_ts<-xts(Google_t_grocery[-1],Google_t_grocery$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp6 <- na_seadec(Google_t_grocery_ts[,16])
ggplot_na_imputations(Google_t_grocery_ts[,16], imp6)
```




```
# We select na_seadec for the dataset
Google_t_grocery_ts <- na_seadec(Google_t_grocery_ts)
plot(Google_t_grocery_ts[,16])
```



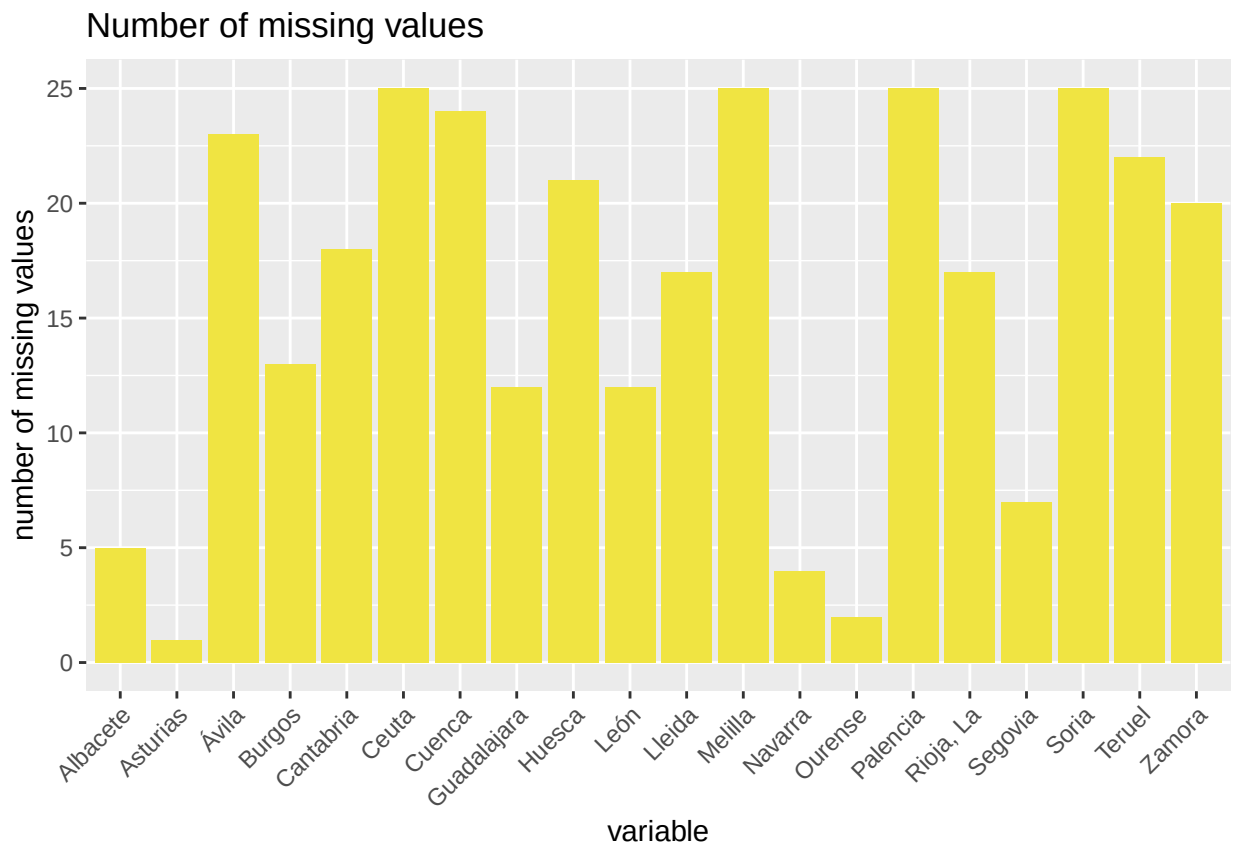

```
# We convert the time series object to a dataframe
Google_grocery <- ts_df(Google_t_grocery_ts)

names(Google_grocery)[names(Google_grocery) == "id"] <- "sub_region_2"
names(Google_grocery)[names(Google_grocery) == "time"] <- "Date"
names(Google_grocery)[names(Google_grocery) == "value"] <- "grocery_and_pharmacy_percent_change_from_bas

#####################################################################
# Transpose dataframe
Google_parks<-Google[c(2,4,7)]
Google_t_parks<-dcast(Google_parks, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_parks, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google_t_parks), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```



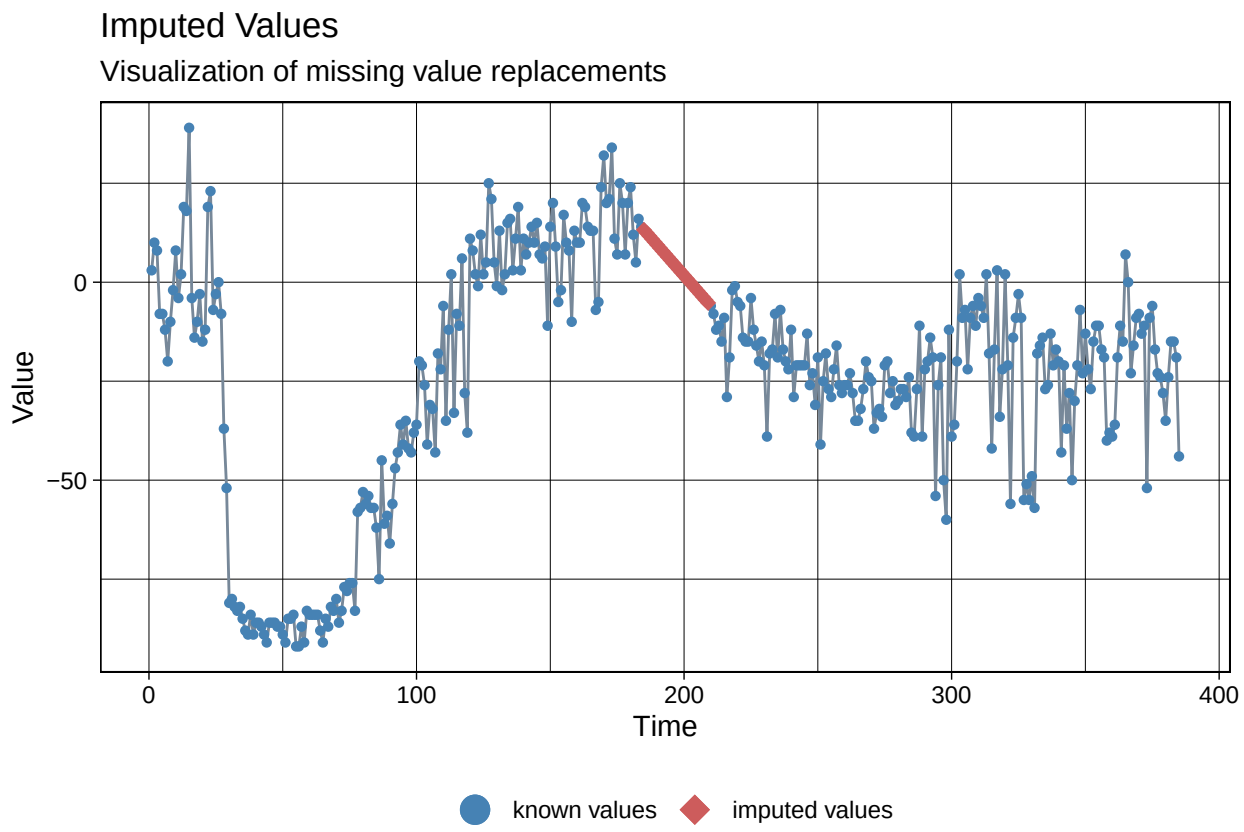

```
##
##  Variables sorted by number of missings:
##                 Variable       Count
##                    Ceuta 0.064935065
##                  Melilla 0.064935065
##                 Palencia 0.064935065
##                    Soria 0.064935065
##                   Cuenca 0.062337662
##                    Ávila 0.059740260
##                   Teruel 0.057142857
##                   Huesca 0.054545455
##                   Zamora 0.051948052
##                Cantabria 0.046753247
##                   Lleida 0.044155844
##                Rioja, La 0.044155844
##                   Burgos 0.033766234
##              Guadalajara 0.031168831
##                    León 0.031168831
##                  Segovia 0.018181818
##                 Albacete 0.012987013
##                  Navarra 0.010389610
##                  Ourense 0.005194805
##                 Asturias 0.002597403
##                     date 0.000000000
##         Alicante/Alacant 0.000000000
##                  Almería 0.000000000
```

```
##             Araba/Álava 0.000000000
##                 Badajoz 0.000000000
##          Balears, Illes 0.000000000
##               Barcelona 0.000000000
##                 Bizkaia 0.000000000
##                 Cáceres 0.000000000
##                   Cádiz 0.000000000
##      Castellón/Castelló 0.000000000
##             Ciudad Real 0.000000000
##                 Córdoba 0.000000000
##               Coruña, A 0.000000000
##                Gipuzkoa 0.000000000
##                  Girona 0.000000000
##                 Granada 0.000000000
##                  Huelva 0.000000000
##                    Jaén 0.000000000
##                    Lugo 0.000000000
##                  Madrid 0.000000000
##                  Málaga 0.000000000
##                  Murcia 0.000000000
##             Palmas, Las 0.000000000
##              Pontevedra 0.000000000
##               Salamanca 0.000000000
##  Santa Cruz de Tenerife 0.000000000
##                 Sevilla 0.000000000
##               Tarragona 0.000000000
##                  Toledo 0.000000000
##       Valencia/València 0.000000000
##              Valladolid 0.000000000
##                Zaragoza 0.000000000
```

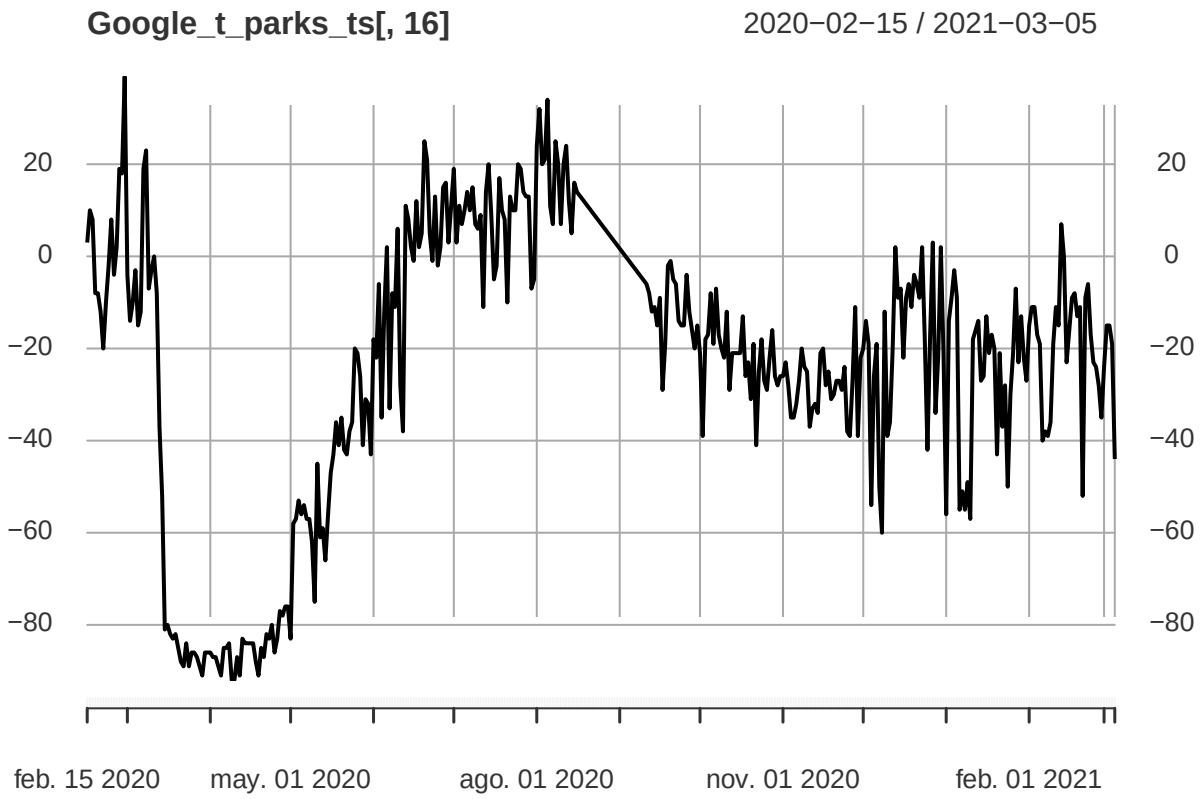
```
Google_t_parks %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



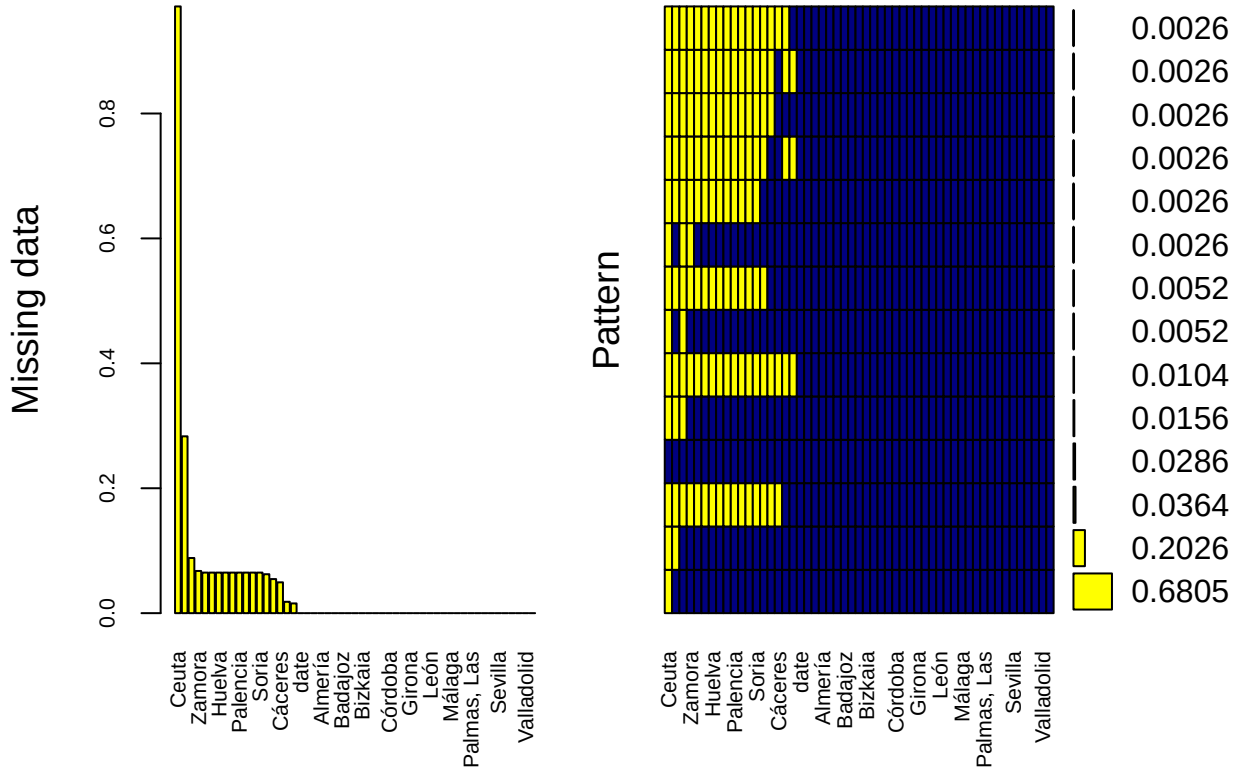

```
# Convert dataframe to ts object
Google_t_parks_ts<-xts(Google_t_parks[-1],Google_t_parks$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp7 <- na_seadec(Google_t_parks_ts[,16])
ggplot_na_imputations(Google_t_parks_ts[,16], imp7)
```




```
# We select na_seadec for the dataset
Google_t_parks_ts <- na_seadec(Google_t_parks_ts)
plot(Google_t_parks_ts[,16])
```



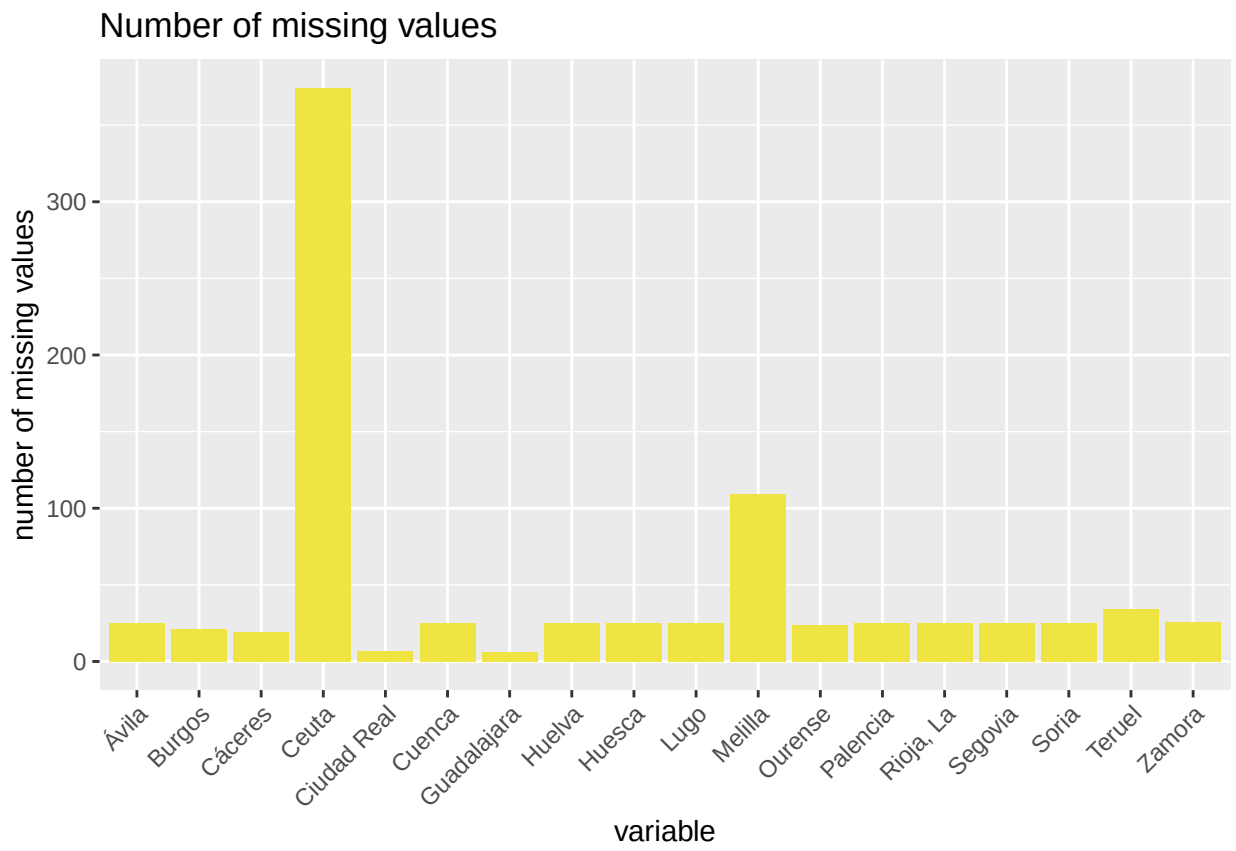

```
# We convert the time series object to a dataframe
Google_parks <- ts_df(Google_t_parks_ts)

names(Google_parks)[names(Google_parks) == "id"] <- "sub_region_2"
names(Google_parks)[names(Google_parks) == "time"] <- "Date"
names(Google_parks)[names(Google_parks) == "value"] <- "parks_percent_change_from_baseline"

#####################################################################
# Transpose dataframe
Google_transit<-Google[c(2,4,8)]
Google_t_transit<-dcast(Google_transit, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_transit, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google_t_transit), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```



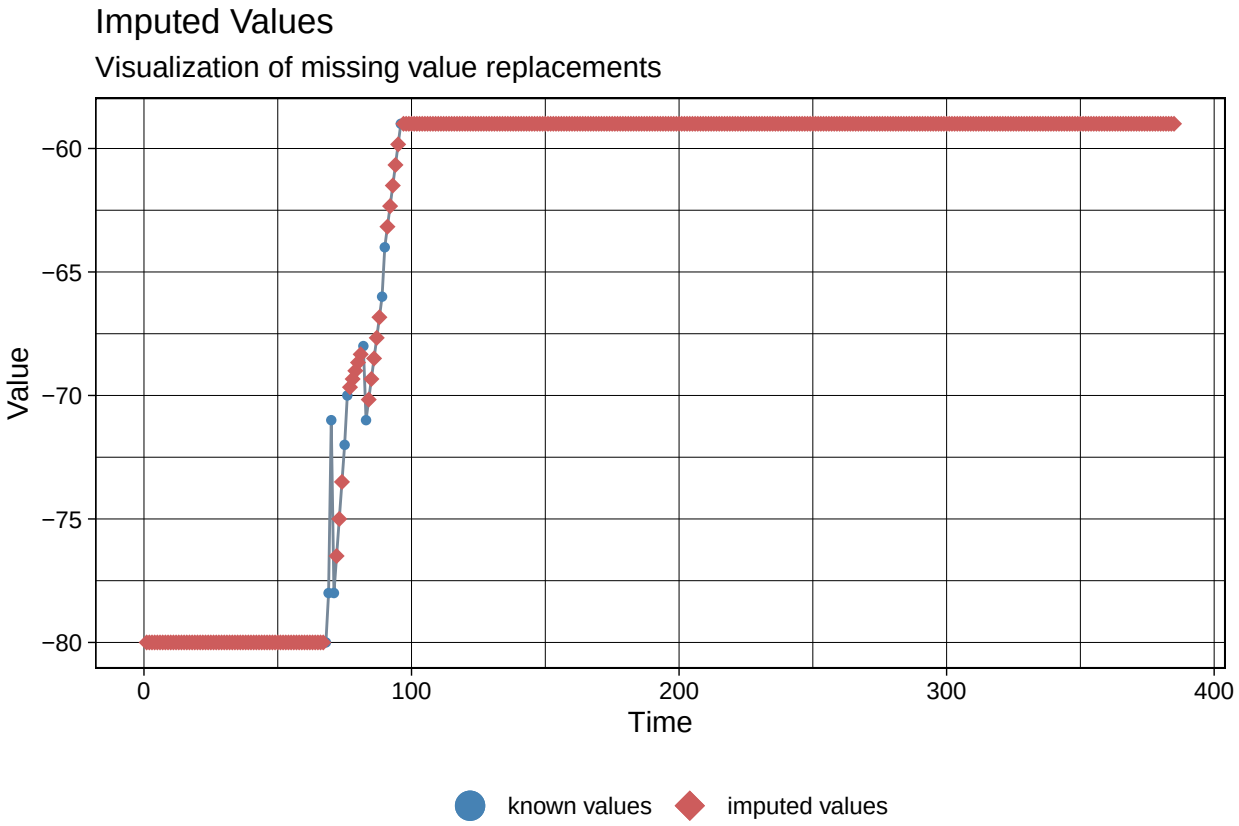

```
##
##  Variables sorted by number of missings:
##                 Variable      Count
##                    Ceuta 0.97142857
##                  Melilla 0.28311688
##                   Teruel 0.08831169
##                   Zamora 0.06753247
##                    Ávila 0.06493506
##                   Cuenca 0.06493506
##                   Huelva 0.06493506
##                   Huesca 0.06493506
##                     Lugo 0.06493506
##                  Palencia 0.06493506
##                 Rioja, La 0.06493506
##                  Segovia 0.06493506
##                    Soria 0.06493506
##                  Ourense 0.06233766
##                   Burgos 0.05454545
##                  Cáceres 0.04935065
##              Ciudad Real 0.01818182
##              Guadalajara 0.01558442
##                     date 0.00000000
##                 Albacete 0.00000000
##         Alicante/Alacant 0.00000000
##                  Almería 0.00000000
##              Araba/Álava 0.00000000
```

```
##                Asturias 0.00000000
##                 Badajoz 0.00000000
##          Balears, Illes 0.00000000
##               Barcelona 0.00000000
##                 Bizkaia 0.00000000
##                   Cádiz 0.00000000
##               Cantabria 0.00000000
##      Castellón/Castelló 0.00000000
##                 Córdoba 0.00000000
##               Coruña, A 0.00000000
##                 Gipuzkoa 0.00000000
##                   Girona 0.00000000
##                 Granada 0.00000000
##                    Jaén 0.00000000
##                    León 0.00000000
##                  Lleida 0.00000000
##                  Madrid 0.00000000
##                  Málaga 0.00000000
##                  Murcia 0.00000000
##                 Navarra 0.00000000
##             Palmas, Las 0.00000000
##               Pontevedra 0.00000000
##                Salamanca 0.00000000
##  Santa Cruz de Tenerife 0.00000000
##                  Sevilla 0.00000000
##                Tarragona 0.00000000
##                   Toledo 0.00000000
##       Valencia/València 0.00000000
##              Valladolid 0.00000000
##                Zaragoza 0.00000000
```

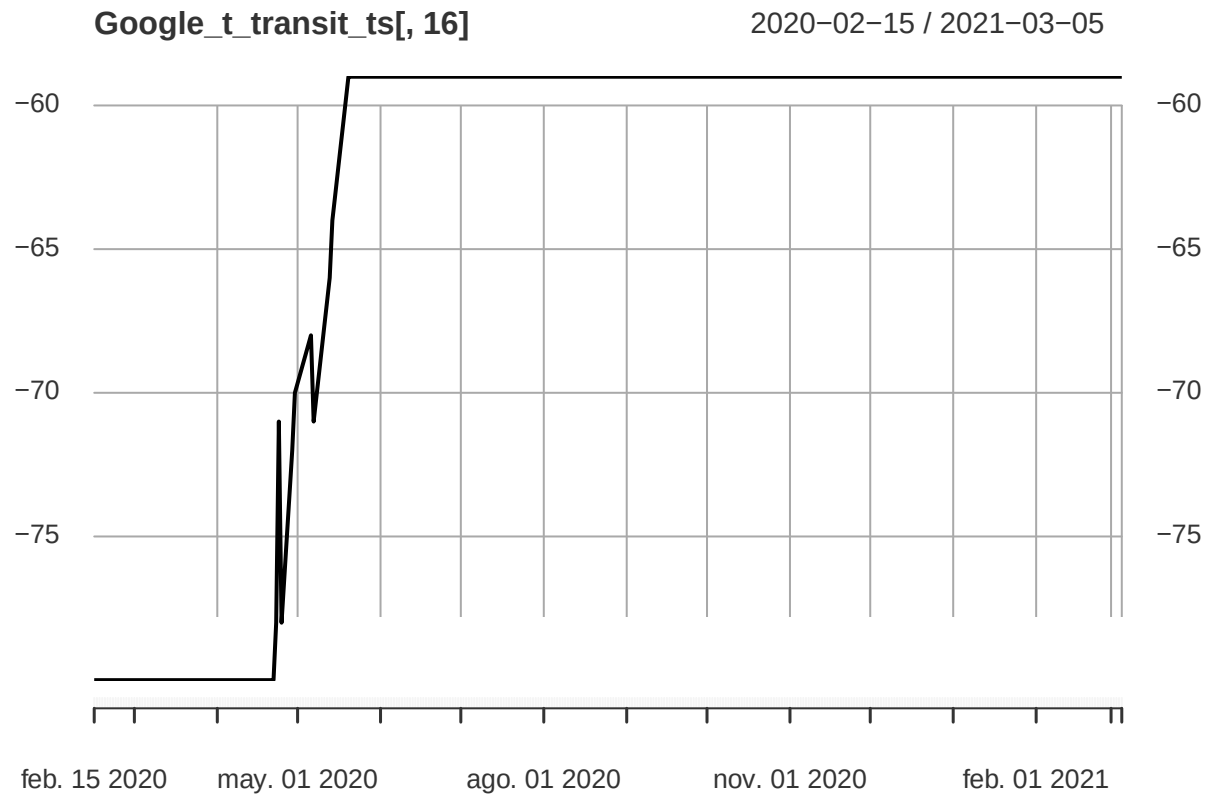
```
Google_t_transit %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



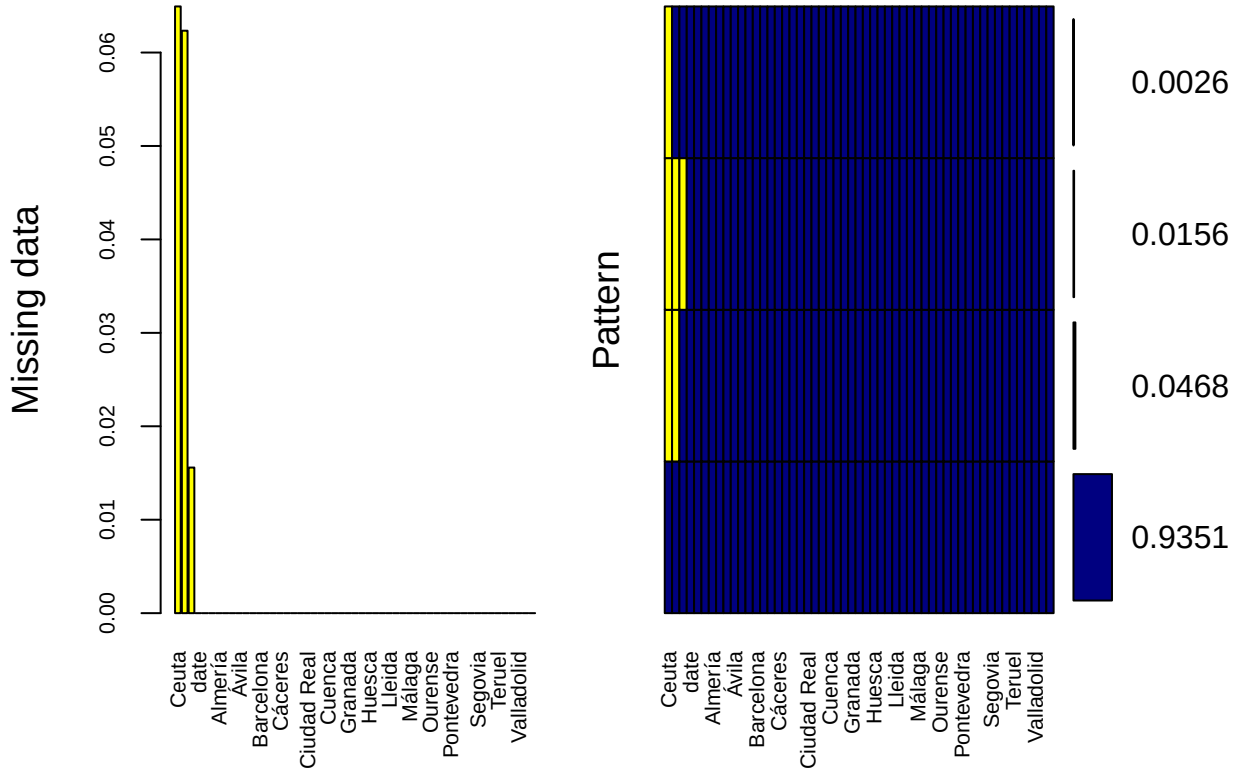

```
# Convert dataframe to ts object
Google_t_transit_ts<-xts(Google_t_transit[-1],Google_t_transit$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp8 <- na_seadec(Google_t_transit_ts[,16])
ggplot_na_imputations(Google_t_transit_ts[,16], imp8)
```




```
# We select na_seadec for the dataset
Google_t_transit_ts <- na_seadec(Google_t_transit_ts)
plot(Google_t_transit_ts[,16])
```



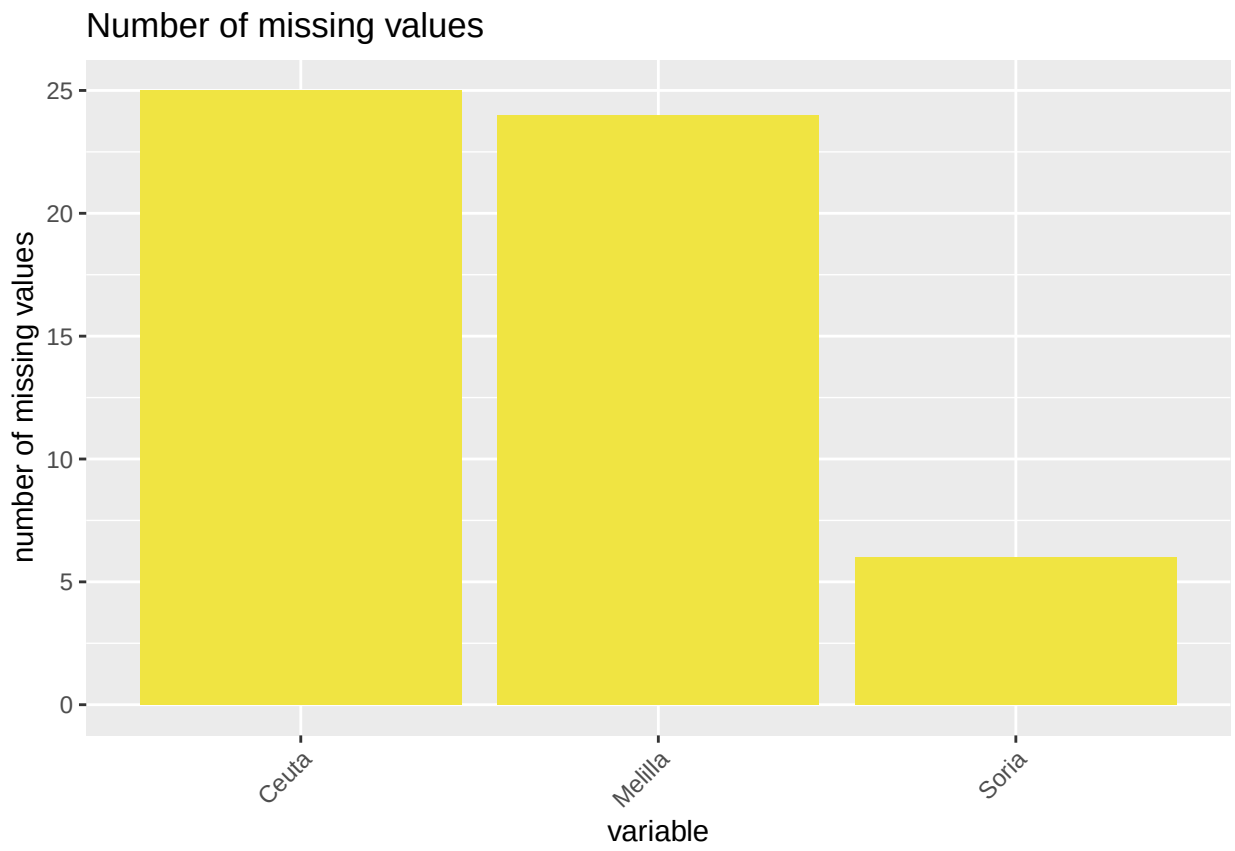

```
# We convert the time series object to a dataframe
Google_transit <- ts_df(Google_t_transit_ts)

names(Google_transit)[names(Google_transit) == "id"] <- "sub_region_2"
names(Google_transit)[names(Google_transit) == "time"] <- "Date"
names(Google_transit)[names(Google_transit) == "value"] <- "transit_stations_percent_change_from_baseli

######################################################################
# Transpose dataframe
Google_workplaces<-Google[c(2,4,9)]
Google_t_workplaces<-dcast(Google_workplaces, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_workplaces, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google_t_workplaces), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```



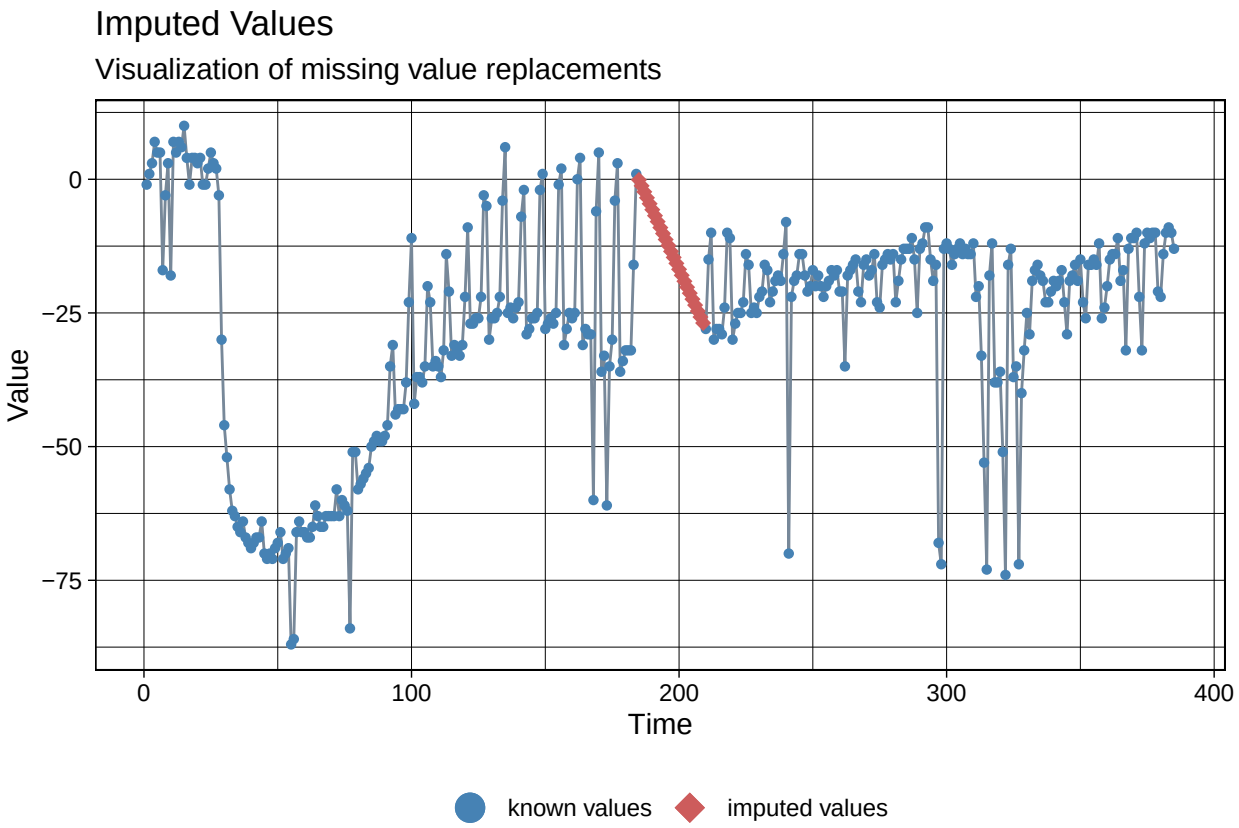

```
##
##  Variables sorted by number of missings:
##                 Variable      Count
##                    Ceuta 0.06493506
##                  Melilla 0.06233766
##                    Soria 0.01558442
##                     date 0.00000000
##                 Albacete 0.00000000
##        Alicante/Alacant 0.00000000
##                  Almería 0.00000000
##              Araba/Álava 0.00000000
##                 Asturias 0.00000000
##                    Ávila 0.00000000
##                  Badajoz 0.00000000
##           Balears, Illes 0.00000000
##                Barcelona 0.00000000
##                  Bizkaia 0.00000000
##                   Burgos 0.00000000
##                  Cáceres 0.00000000
##                    Cádiz 0.00000000
##                Cantabria 0.00000000
##      Castellón/Castelló 0.00000000
##              Ciudad Real 0.00000000
##                  Córdoba 0.00000000
##                Coruña, A 0.00000000
##                   Cuenca 0.00000000
```

```
##                Gipuzkoa 0.00000000
##                  Girona 0.00000000
##                 Granada 0.00000000
##              Guadalajara 0.00000000
##                  Huelva 0.00000000
##                  Huesca 0.00000000
##                   Jaén 0.00000000
##                   León 0.00000000
##                  Lleida 0.00000000
##                    Lugo 0.00000000
##                  Madrid 0.00000000
##                  Málaga 0.00000000
##                  Murcia 0.00000000
##                 Navarra 0.00000000
##                 Ourense 0.00000000
##                Palencia 0.00000000
##              Palmas, Las 0.00000000
##               Pontevedra 0.00000000
##                Rioja, La 0.00000000
##                Salamanca 0.00000000
##  Santa Cruz de Tenerife 0.00000000
##                  Segovia 0.00000000
##                  Sevilla 0.00000000
##                Tarragona 0.00000000
##                   Teruel 0.00000000
##                   Toledo 0.00000000
##       Valencia/València 0.00000000
##              Valladolid 0.00000000
##                  Zamora 0.00000000
##                 Zaragoza 0.00000000
```

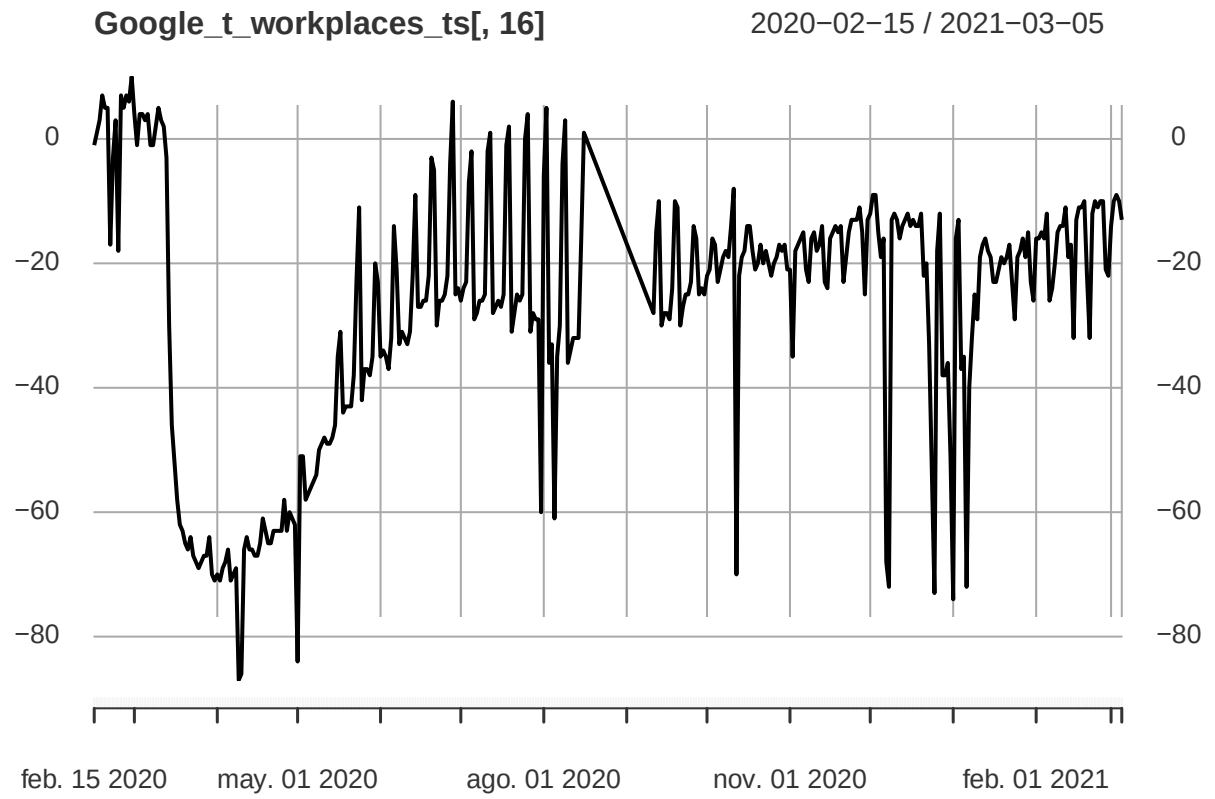
```
Google_t_workplaces %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



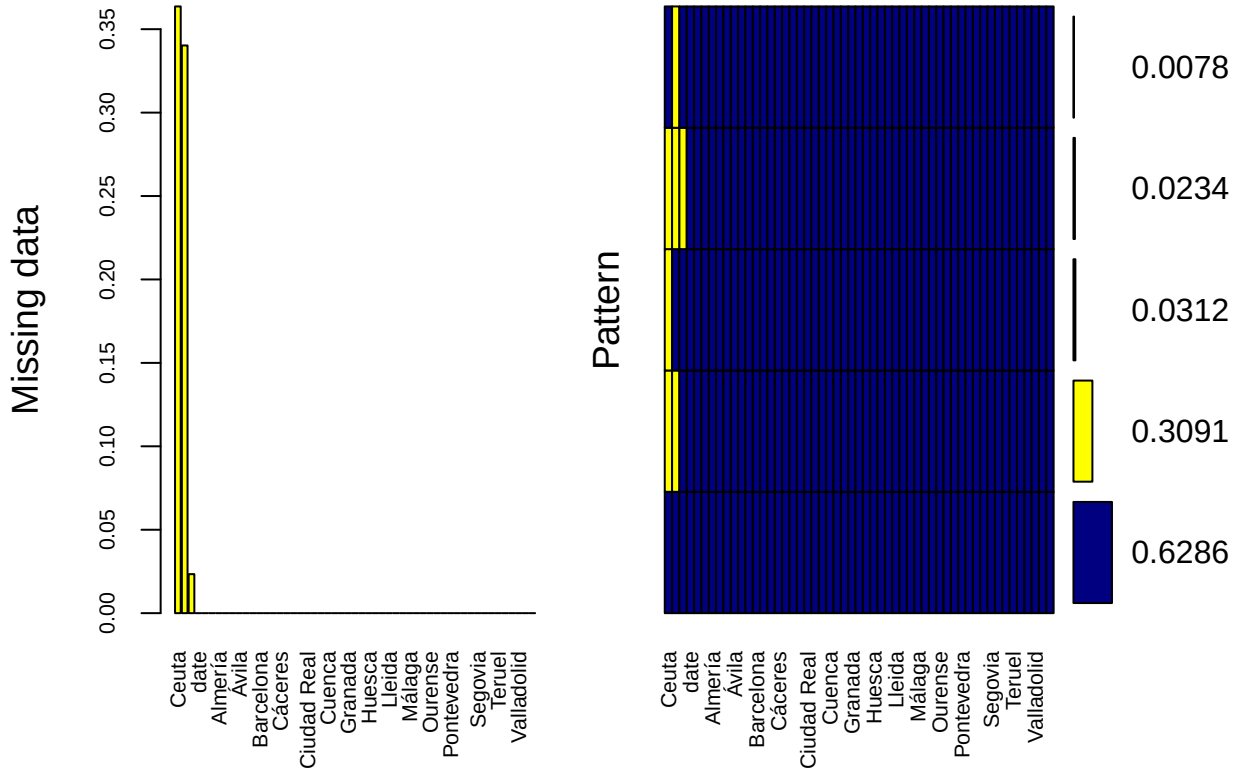

```
# Convert dataframe to ts object
Google_t_workplaces_ts<-xts(Google_t_workplaces[-1],Google_t_workplaces$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp9 <- na_seadec(Google_t_workplaces_ts[,16])
ggplot_na_imputations(Google_t_workplaces_ts[,16], imp9)
```




```
# We select na_seadec for the dataset
Google_t_workplaces_ts <- na_seadec(Google_t_workplaces_ts)
plot(Google_t_workplaces_ts[,16])
```



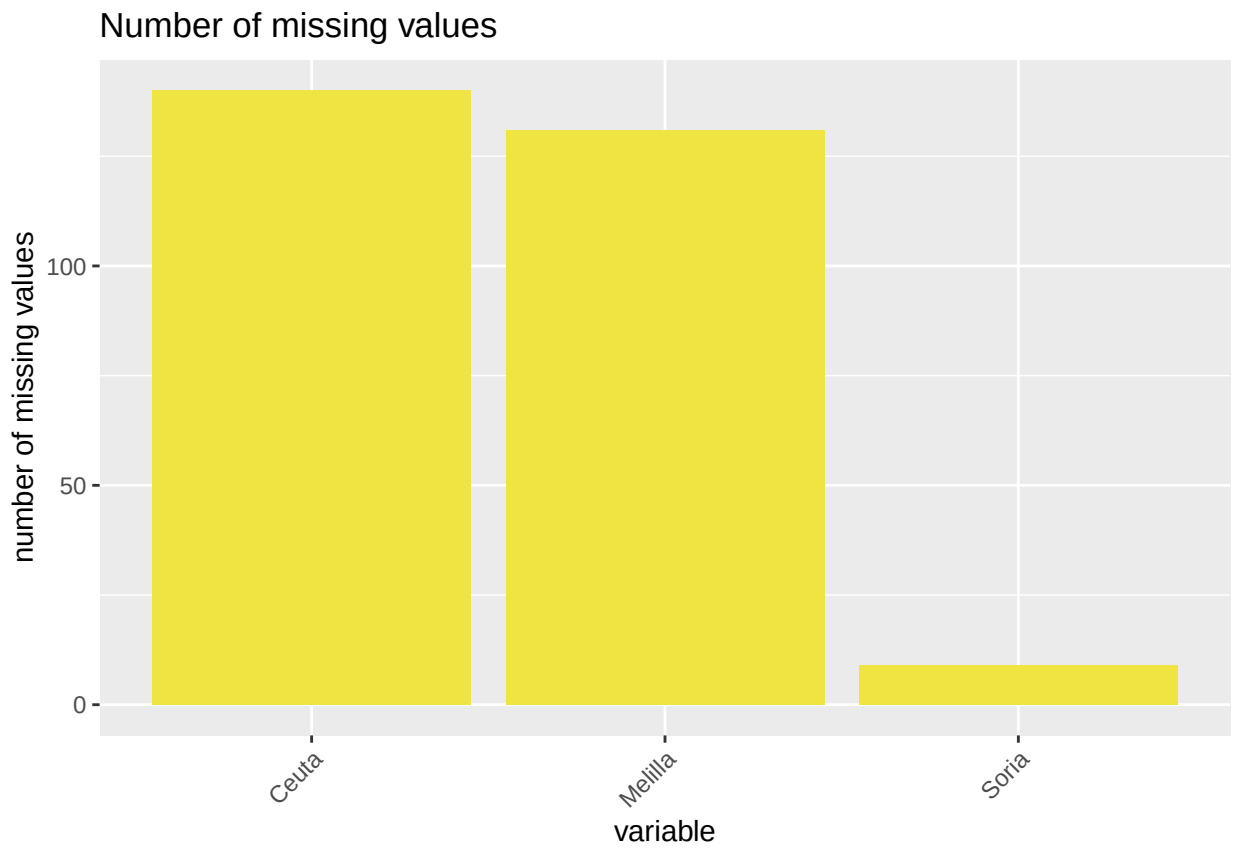

```
# We convert the time series object to a dataframe
Google_workplaces <- ts_df(Google_t_workplaces_ts)

names(Google_workplaces)[names(Google_workplaces) == "id"] <- "sub_region_2"
names(Google_workplaces)[names(Google_workplaces) == "time"] <- "Date"
names(Google_workplaces)[names(Google_workplaces) == "value"] <- "workplaces_percent_change_from_baseli

#####################################################################
# Transpose dataframe
Google_residential<-Google[c(2,4,10)]
Google_t_residential<-dcast(Google_residential, date~sub_region_2, fill=NA)

# Visualize missing values
aggr(Google_t_residential, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google_t_residential), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```



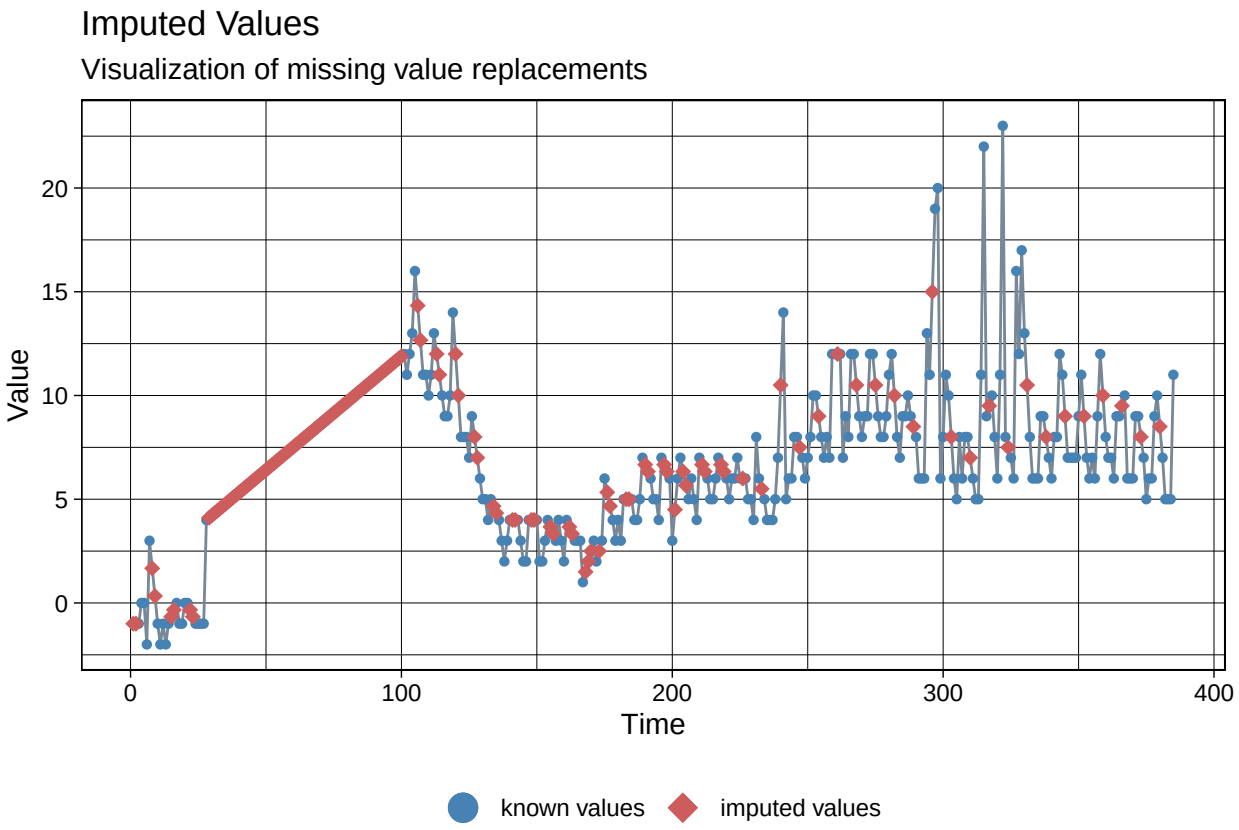

```
##
##  Variables sorted by number of missings:
##                 Variable      Count
##                    Ceuta 0.36363636
##                  Melilla 0.34025974
##                    Soria 0.02337662
##                     date 0.00000000
##                 Albacete 0.00000000
##         Alicante/Alacant 0.00000000
##                  Almería 0.00000000
##              Araba/Álava 0.00000000
##                 Asturias 0.00000000
##                    Ávila 0.00000000
##                  Badajoz 0.00000000
##            Balears, Illes 0.00000000
##                 Barcelona 0.00000000
##                  Bizkaia 0.00000000
##                   Burgos 0.00000000
##                  Cáceres 0.00000000
##                    Cádiz 0.00000000
##                Cantabria 0.00000000
##       Castellón/Castelló 0.00000000
##              Ciudad Real 0.00000000
##                  Córdoba 0.00000000
##                Coruña, A 0.00000000
##                   Cuenca 0.00000000
```

```
##                Gipuzkoa 0.00000000
##                  Girona 0.00000000
##                 Granada 0.00000000
##              Guadalajara 0.00000000
##                  Huelva 0.00000000
##                  Huesca 0.00000000
##                   Jaén 0.00000000
##                   León 0.00000000
##                  Lleida 0.00000000
##                    Lugo 0.00000000
##                  Madrid 0.00000000
##                  Málaga 0.00000000
##                  Murcia 0.00000000
##                 Navarra 0.00000000
##                 Ourense 0.00000000
##                Palencia 0.00000000
##              Palmas, Las 0.00000000
##               Pontevedra 0.00000000
##                Rioja, La 0.00000000
##                Salamanca 0.00000000
##  Santa Cruz de Tenerife 0.00000000
##                  Segovia 0.00000000
##                  Sevilla 0.00000000
##                Tarragona 0.00000000
##                   Teruel 0.00000000
##                   Toledo 0.00000000
##       Valencia/València 0.00000000
##              Valladolid 0.00000000
##                  Zamora 0.00000000
##                 Zaragoza 0.00000000
```

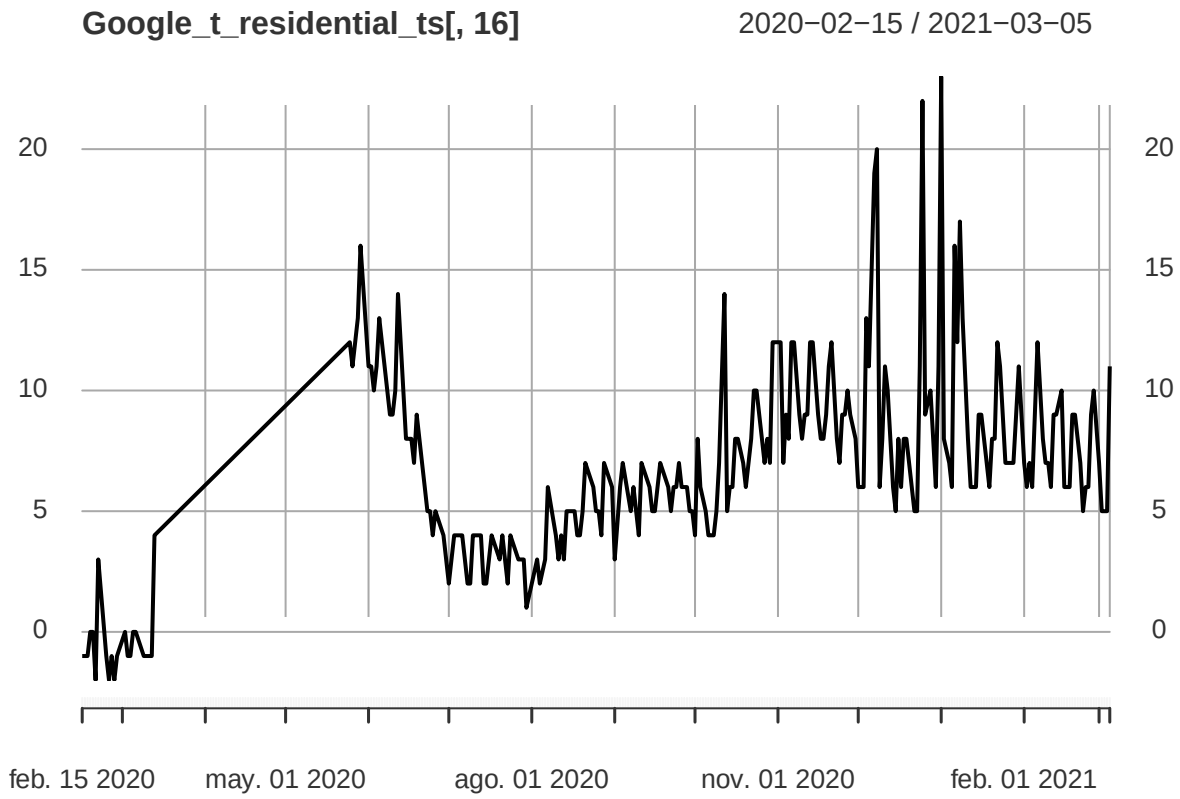
```
Google_t_residential %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Convert dataframe to ts object
Google_t_residential_ts<-xts(Google_t_residential[-1],Google_t_residential$date)

# Impute the missing values with na_seadec (i.e Ceuta)
imp10 <- na_seadec(Google_t_residential_ts[,16])
ggplot_na_imputations(Google_t_residential_ts[,16], imp10)
```




```
# We select na_seadec for the dataset
Google_t_residential_ts <- na_seadec(Google_t_residential_ts)
plot(Google_t_residential_ts[,16])
```




```
# We convert the time series object to a dataframe
Google_residential <- ts_df(Google_t_residential_ts)

names(Google_residential)[names(Google_residential) == "id"] <- "sub_region_2"
names(Google_residential)[names(Google_residential) == "time"] <- "Date"
names(Google_residential)[names(Google_residential) == "value"] <- "residential_percent_change_from_bas
```

Now we merge the previous dataframes into new one with the imputed vaules and we add the ISO code for the province.

```
# New dataframe Google_b
Google_b <- merge(Google_retail, Google_grocery) %>%
              merge(Google_parks) %>%
              merge(Google_transit) %>%
              merge(Google_workplaces) %>%
              merge(Google_residential)

# We add the iso code for the province
Google_b$iso_code <- NA
Google_b$iso_code<-Google[match(Google_b$sub_region_2, Google$sub_region_2),3]
rm("Google")
Google<-Google_b
rm("Google_b")
head(Google,5)
```

```
##   sub_region_2       Date retail_and_recreation_percent_change_from_baseline
## 1     Albacete 2020-02-15                                                3
```

```
## 2     Albacete 2020-02-16                                                 5
## 3     Albacete 2020-02-17                                                -2
## 4     Albacete 2020-02-18                                                -3
## 5     Albacete 2020-02-19                                                 0
##   grocery_and_pharmacy_percent_change_from_baseline
## 1                                                -5
## 2                                                 1
## 3                                                 3
## 4                                                -1
## 5                                                 1
##   parks_percent_change_from_baseline
## 1                                 35
## 2                                 40
## 3                                  7
## 4                                 -4
## 5                                  7
##   transit_stations_percent_change_from_baseline
## 1                                            13
## 2                                            18
## 3                                            20
## 4                                             6
## 5                                             9
##   workplaces_percent_change_from_baseline
## 1                                       1
## 2                                       0
## 3                                       5
## 4                                       4
## 5                                       4
##   residential_percent_change_from_baseline iso_code
## 1                                       -3       AB
## 2                                       -4       AB
## 3                                       -1       AB
## 4                                       -1       AB
## 5                                       -1       AB
```

```
table(Google$sub_region_2)
```

```
##
##               Albacete       Alicante/Alacant                Almería
##                    385                    385                    385
##           Araba/Álava               Asturias                  Ávila
##                    385                    385                    385
##                Badajoz         Balears, Illes              Barcelona
##                    385                    385                    385
##                Bizkaia                 Burgos                Cáceres
##                    385                    385                    385
##                  Cádiz              Cantabria     Castellón/Castelló
##                    385                    385                    385
##                  Ceuta            Ciudad Real                Córdoba
##                    385                    385                    385
##              Coruña, A                 Cuenca               Gipuzkoa
##                    385                    385                    385
##                 Girona                Granada           Guadalajara
##                    385                    385                    385
##                 Huelva                 Huesca                  Jaén
```

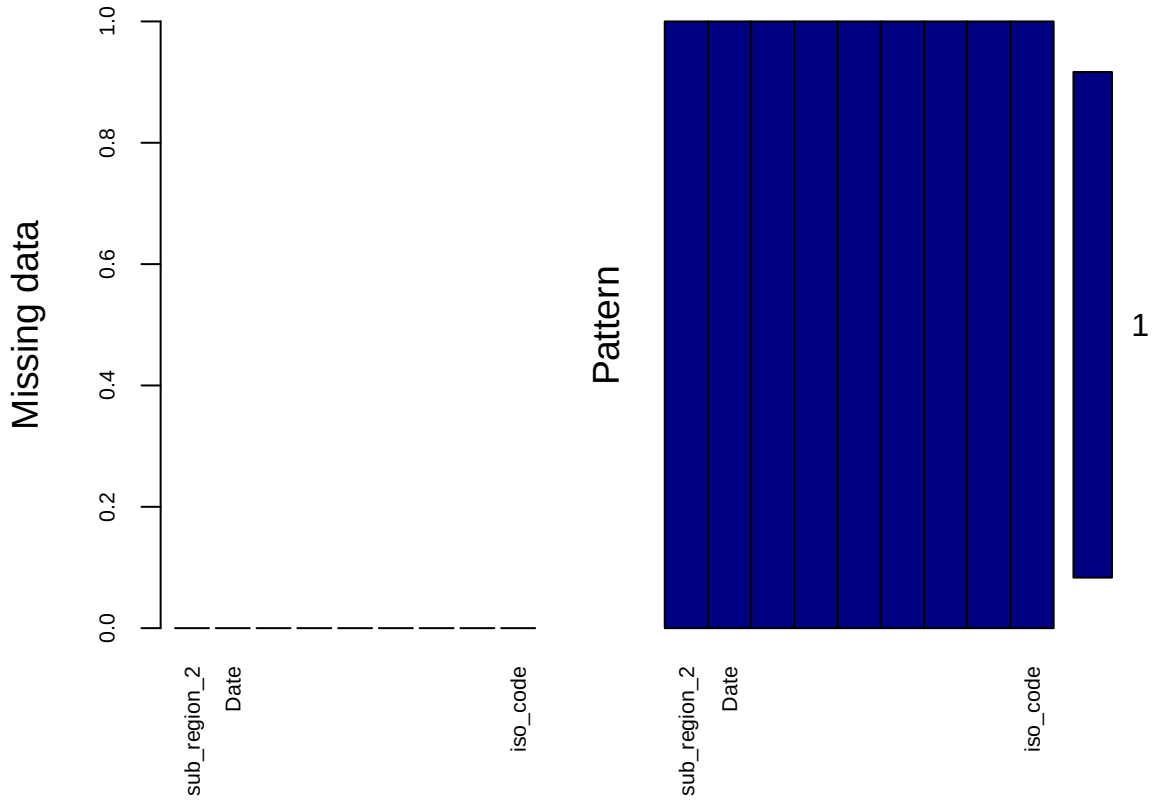
```
##                   385                   385                   385
##                  León                 Lleida                  Lugo
##                   385                   385                   385
##                Madrid                Málaga               Melilla
##                   385                   385                   385
##                Murcia               Navarra               Ourense
##                   385                   385                   385
##              Palencia           Palmas, Las             Pontevedra
##                   385                   385                   385
##            Rioja, La             Salamanca Santa Cruz de Tenerife
##                   385                   385                   385
##               Segovia               Sevilla                 Soria
##                   385                   385                   385
##             Tarragona                Teruel                Toledo
##                   385                   385                   385
##     Valencia/València             Valladolid                 Zamora
##                   385                   385                   385
##              Zaragoza
##                   385
```

```
table(Google$iso_code)
```

```
##
##   A  AB  AL  AV   B  BA  BI  BU   C  CA  CC  CE  CO  CR  CS  CU  GC  GI  GR  GU
## 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385
##   H  HU   J   L  LE  LO  LU   M  MA  ML  MU  NA   O  OR   P  PM  PO   S  SA  SE
## 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385
##  SG  SO  SS   T  TE  TF  TO   V  VA  VI   Z  ZA
## 385 385 385 385 385 385 385 385 385 385 385 385
```

We check missing values. We should obtain zero missing values.

```
aggr(Google, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##                                                Variable Count
##                                             sub_region_2     0
##                                                    Date     0
##  retail_and_recreation_percent_change_from_baseline     0
##   grocery_and_pharmacy_percent_change_from_baseline     0
##                  parks_percent_change_from_baseline     0
##       transit_stations_percent_change_from_baseline     0
##             workplaces_percent_change_from_baseline     0
##           residential_percent_change_from_baseline     0
##                                             iso_code     0
```

```
Google %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

2.1.9 CNE review

The CSV files are provided per "imputed date" (fecha)":

- **cases_technic_province.csv** - Number of cases by diagnostic technique and province (of residence)
- **cases_hosp_uci_def_sexo_edad_provres.csv** - Number of hospitalizations, number of ICU admissions and number of deaths by sex, age and province of residence.

```
head(str(CNE_tecnica, vec.len=3))
```

```
## 'data.frame':    23426 obs. of  8 variables:
##  $ provincia_iso               : chr  "A" "AB" "AL" ...
##  $ fecha                       : chr  "2020-01-01" "2020-01-01" "2020-01-01" ...
##  $ num_casos                   : int  0 0 0 0 0 0 0 0 ...
##  $ num_casos_prueba_pcr        : int  0 0 0 0 0 0 0 0 ...
##  $ num_casos_prueba_test_ac    : int  0 0 0 0 0 0 0 0 ...
##  $ num_casos_prueba_ag         : int  0 0 0 0 0 0 0 0 ...
##  $ num_casos_prueba_elisa      : int  0 0 0 0 0 0 0 0 ...
##  $ num_casos_prueba_desconocida: int  0 0 0 0 0 0 0 0 ...
```

```
## NULL
```

```
summary(CNE_tecnica)
```

```
##  provincia_iso         fecha             num_casos      num_casos_prueba_pcr
##  Length:23426       Length:23426       Min.   :   0.0   Min.   :   0.0
##  Class :character   Class :character   1st Qu.:   2.0   1st Qu.:   2.0
##  Mode  :character   Mode  :character   Median :  32.0   Median :  26.0
```

```
##                                     Mean   : 136.9   Mean   : 109.6
##                                     3rd Qu.: 120.0   3rd Qu.: 100.0
##                                     Max.   :6972.0   Max.   :6546.0
##  num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
##  Min.   : 0.0000          Min.   :   0.00     Min.   : 0.0000
##  1st Qu.: 0.0000          1st Qu.:   0.00     1st Qu.: 0.0000
##  Median : 0.0000          Median :   0.00     Median : 0.0000
##  Mean   : 0.2037          Mean   :  26.21     Mean   : 0.1602
##  3rd Qu.: 0.0000          3rd Qu.:   9.00     3rd Qu.: 0.0000
##  Max.   :32.0000          Max.   :3267.00     Max.   :71.0000
##  num_casos_prueba_desconocida
##  Min.   :  0.0000
##  1st Qu.:  0.0000
##  Median :  0.0000
##  Mean   :  0.7122
##  3rd Qu.:  0.0000
##  Max.   :505.0000
```

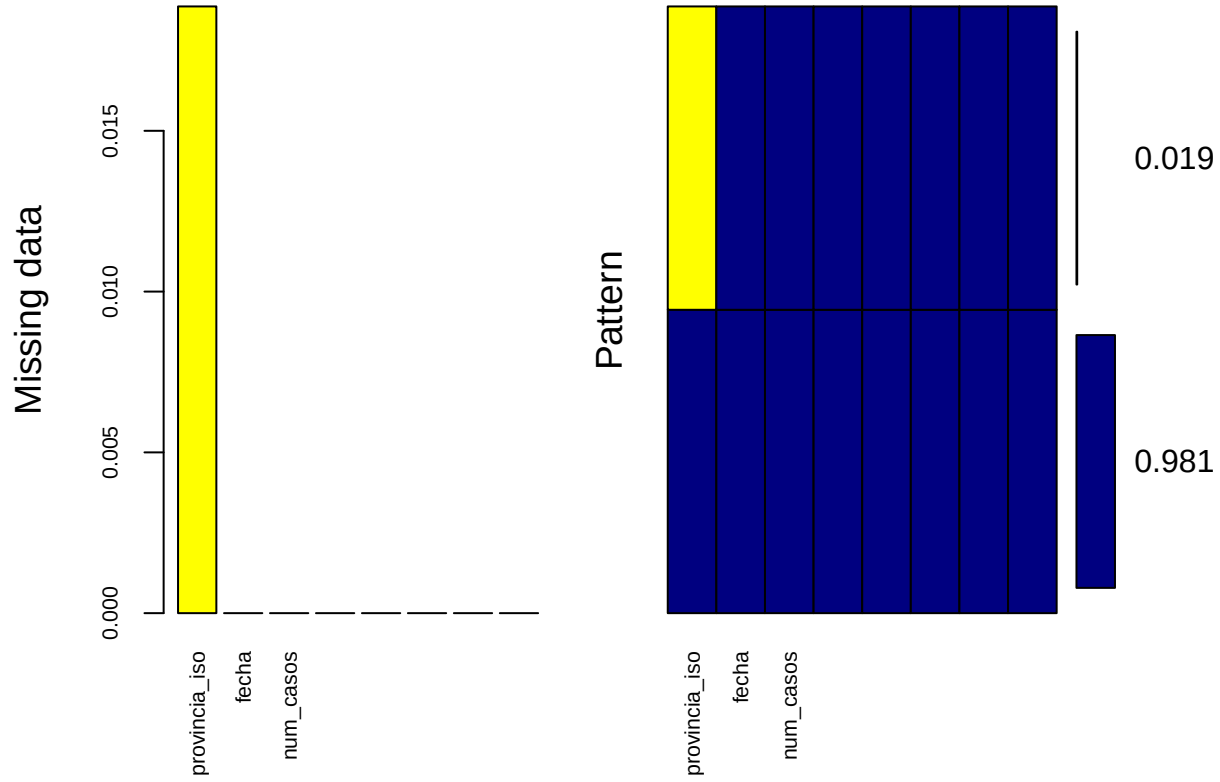
```
table(CNE_tecnica$provincia_iso)
```

```
##
##   A  AB  AL  AV   B  BA  BI  BU   C  CA  CC  CE  CO  CR  CS  CU  GC  GI  GR  GU
## 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442
##   H  HU   J   L  LE  LO  LU   M  MA  ML  MU  NC   O  OR   P  PM  PO   S  SA  SE
## 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442
##  SG  SO  SS   T  TE  TF  TO   V  VA  VI   Z  ZA
## 442 442 442 442 442 442 442 442 442 442 442 442
```

2.1.10 CNE review missing values & impute

We check missing values for CNE_tecnica. In this case we omit the NA values.

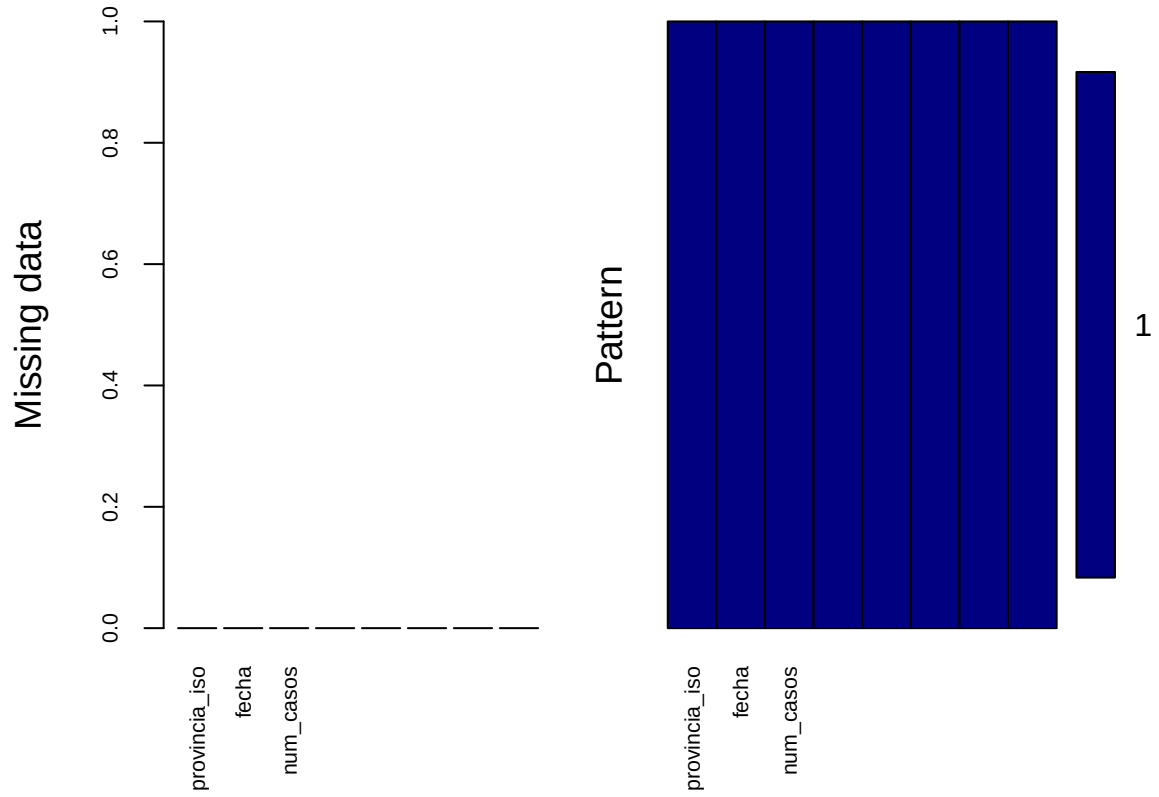
```
aggr(CNE_tecnica, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(CNE_tecnica), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##                       Variable      Count
##                 provincia_iso 0.01886792
##                         fecha 0.00000000
##                     num_casos 0.00000000
##           num_casos_prueba_pcr 0.00000000
##       num_casos_prueba_test_ac 0.00000000
##            num_casos_prueba_ag 0.00000000
##          num_casos_prueba_elisa 0.00000000
##  num_casos_prueba_desconocida 0.00000000
```

```r
CNE_tecnica %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


```
####################################
CNE_tecnica <- na.omit(CNE_tecnica)
####################################

aggr(CNE_tecnica, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(CNE_tecnica), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##                       Variable Count
##                  provincia_iso     0
##                          fecha     0
##                      num_casos     0
##           num_casos_prueba_pcr     0
##       num_casos_prueba_test_ac     0
##            num_casos_prueba_ag     0
##          num_casos_prueba_elisa     0
##  num_casos_prueba_desconocida     0
```

```
CNE_tecnica %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


```
head(str(CNE_casos,vec.len=3))
```

```
## 'data.frame':    702780 obs. of  8 variables:
##  $ provincia_iso: chr  "A" "A" "A" ...
##  $ sexo         : chr  "H" "H" "H" ...
##  $ grupo_edad   : chr  "0-9" "10-19" "20-29" ...
##  $ fecha        : chr  "2020-01-01" "2020-01-01" "2020-01-01" ...
##  $ num_casos    : int  0 0 0 0 0 0 0 0 ...
##  $ num_hosp     : int  0 0 0 0 0 0 0 0 ...
##  $ num_uci      : int  0 0 0 0 0 0 0 0 ...
##  $ num_def      : int  0 0 0 0 0 0 0 0 ...
```

```
## NULL
```

```
summary(CNE_casos)
```

```
##  provincia_iso          sexo            grupo_edad           fecha          
##  Length:702780      Length:702780      Length:702780      Length:702780     
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##    num_casos          num_hosp           num_uci            num_def       
##  Min.   :  0.000   Min.   :  0.0000   Min.   : 0.00000   Min.   :  0.0000  
##  1st Qu.:  0.000   1st Qu.:  0.0000   1st Qu.: 0.00000   1st Qu.:  0.0000  
##  Median :  0.000   Median :  0.0000   Median : 0.00000   Median :  0.0000  
##  Mean   :  4.562   Mean   :  0.4611   Mean   : 0.04117   Mean   :  0.1036  
```

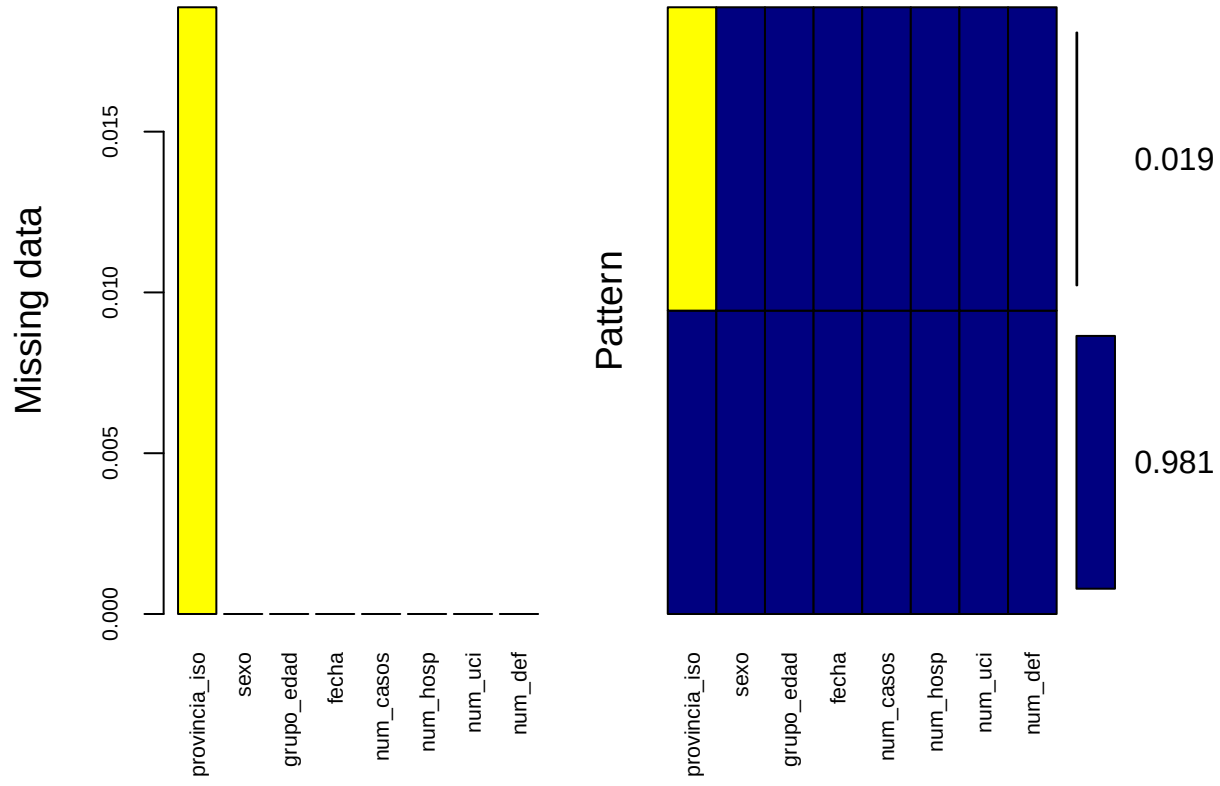
```
##  3rd Qu.:  2.000   3rd Qu.:  0.0000   3rd Qu.: 0.00000   3rd Qu.:  0.0000
##  Max.   :771.000   Max.   :269.0000   Max.   :35.00000   Max.   :100.0000
```

```
table(CNE_casos$provincia_iso)
```

```
##
##     A    AB    AL    AV     B    BA    BI    BU     C    CA    CC    CE    CO
## 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260
##    CR    CS    CU    GC    GI    GR    GU     H    HU     J     L    LE    LO
## 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260
##    LU     M    MA    ML    MU    NC     O    OR     P    PM    PO     S    SA
## 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260
##    SE    SG    SO    SS     T    TE    TF    TO     V    VA    VI     Z    ZA
## 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260 13260
```

We check missing values for CNE_casos. In this case also we omit the NA values.

```
aggr(CNE_casos, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(CNE_casos), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```



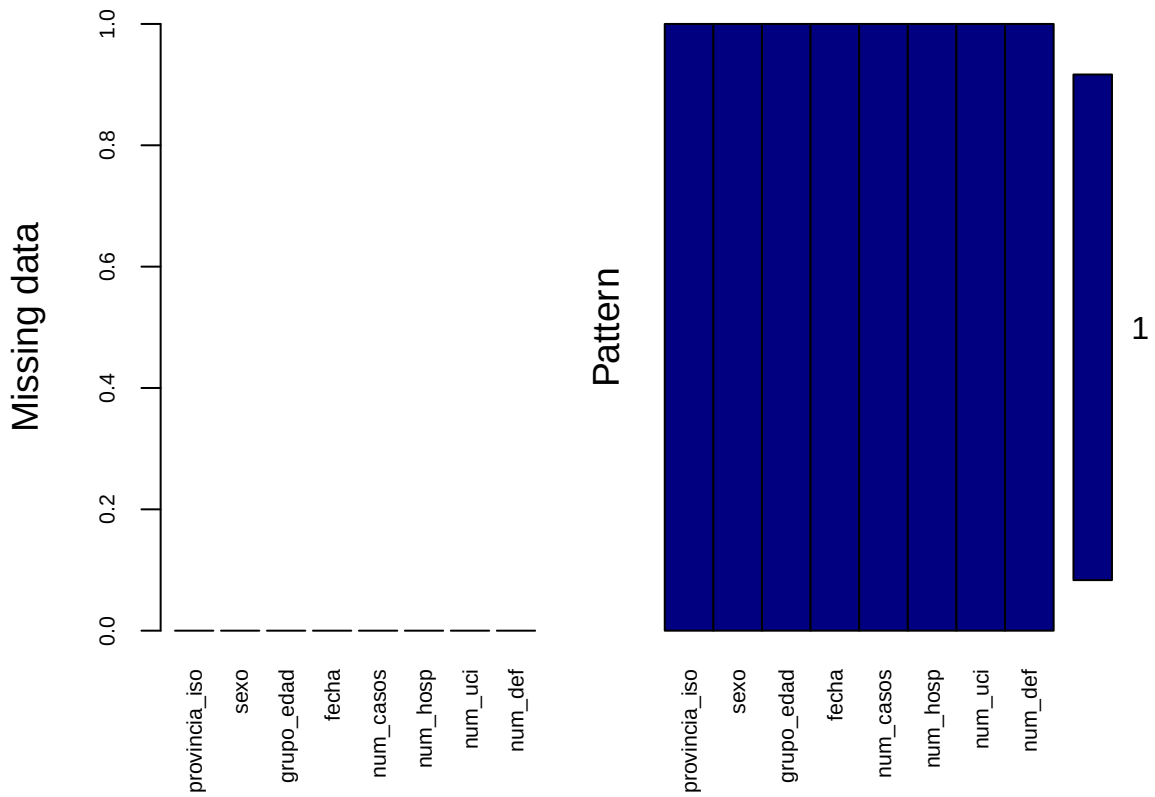

```
##
##  Variables sorted by number of missings:
##       Variable       Count
##  provincia_iso 0.01886792
##           sexo 0.00000000
##     grupo_edad 0.00000000
```

```
##          fecha 0.00000000
##      num_casos 0.00000000
##       num_hosp 0.00000000
##        num_uci 0.00000000
##        num_def 0.00000000
```

```
CNE_casos %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


```
################################
CNE_casos <- na.omit(CNE_casos)
################################

aggr(CNE_casos, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(CNE_casos), cex.axis=.7,
```

```
     gap=3, ylab=c("Missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##       Variable Count
##  provincia_iso     0
##           sexo     0
##     grupo_edad     0
##          fecha     0
##      num_casos     0
##       num_hosp     0
##        num_uci     0
##        num_def     0
```

```
CNE_casos %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
```

```
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

2.1.11 CNE data transformation

We are going to **transform / eliminate**:

- A - "Fecha" column is transformed (in both datasets) from "character" to "date".
- B - "Grupo_edad" and "Sexo" columns are eliminated from dataset "CNE_casos" due to they are not adding value (mobility does not include this variable).
- C - We change NC iso code to NA (Navarra) in both dataframes.

```
# Transform / eliminate A
CNE_tecnica$fecha <- as.Date(CNE_tecnica$fecha ,format="%Y-%m-%d")
CNE_casos$fecha <- as.Date(CNE_casos$fecha ,format="%Y-%m-%d")

# Transform / eliminate B
CNE_casos<-within(CNE_casos, rm(grupo_edad, sexo))

# Iso code update for Navarra C
CNE_tecnica$provincia_iso[CNE_tecnica$provincia_iso=="NC"] <- "NA"
CNE_casos$provincia_iso[CNE_casos$provincia_iso=="NC"] <- "NA"

head(CNE_tecnica,5)
```

```
##   provincia_iso      fecha num_casos num_casos_prueba_pcr
## 1             A 2020-01-01         0                    0
## 2            AB 2020-01-01         0                    0
```

```
## 3            AL 2020-01-01         0                   0
## 4            AV 2020-01-01         0                   0
## 5             B 2020-01-01         0                   0
##   num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## 1                        0                   0                      0
## 2                        0                   0                      0
## 3                        0                   0                      0
## 4                        0                   0                      0
## 5                        0                   0                      0
##   num_casos_prueba_desconocida
## 1                            0
## 2                            0
## 3                            0
## 4                            0
## 5                            0
```

```
head(CNE_casos,5)
```

```
##   provincia_iso      fecha num_casos num_hosp num_uci num_def
## 1             A 2020-01-01         0        0       0       0
## 2             A 2020-01-01         0        0       0       0
## 3             A 2020-01-01         0        0       0       0
## 4             A 2020-01-01         0        0       0       0
## 5             A 2020-01-01         0        0       0       0
```

We check both dataframes offers the same total results.

```
CNE_tecnica %>%
  group_by(provincia_iso) %>%
  summarise_at(vars(num_casos), sum)
```

```
## # A tibble: 52 x 2
##    provincia_iso num_casos
##    <chr>             <int>
##  1 A                143555
##  2 AB                26916
##  3 AL                47032
##  4 AV                11084
##  5 B                382992
##  6 BA                45886
##  7 BI                80588
##  8 BU                29808
##  9 C                 51272
## 10 CA                70428
## # ... with 42 more rows
```

```
CNE_casos %>%
  group_by(provincia_iso) %>%
  summarise_at(vars(num_casos), sum)
```

```
## # A tibble: 52 x 2
##    provincia_iso num_casos
##    <chr>             <int>
##  1 A                143555
##  2 AB                26916
##  3 AL                47032
##  4 AV                11084
```

```
##  5 B                382992
##  6 BA                45886
##  7 BI                80588
##  8 BU                29808
##  9 C                 51272
## 10 CA                70428
## # ... with 42 more rows
```

2.2 Datasets combinations

We proceed to **combine** the different data sets into one.

2.2.1 CNE_tec_cas

- CNE_casos_g, a groupped dataframe due to the columns eliminated in previous step (grupo_edad, sexo)
- CNE_tec_cas -> CNE_tecnica + CNE_casos_g

Here we merge by columns "provincia_iso","fecha".

```
# CNE_casos_g
CNE_casos_g = CNE_casos %>%
  group_by(provincia_iso, fecha) %>%
  summarise_at(vars(num_casos, num_hosp, num_uci, num_def), sum)
head(CNE_casos_g,5)
```

```
## # A tibble: 5 x 6
## # Groups:   provincia_iso [1]
##   provincia_iso fecha      num_casos num_hosp num_uci num_def
##   <chr>         <date>         <int>    <int>   <int>   <int>
## 1 A             2020-01-01         0        1       0       0
## 2 A             2020-01-02         0        0       0       0
## 3 A             2020-01-03         0        0       0       0
## 4 A             2020-01-04         0        0       0       0
## 5 A             2020-01-05         0        1       0       0
```

```
# New dataframe CNE_tec_cas
CNE_tec_cas<-merge(CNE_tecnica,
                   CNE_casos_g, by.x=c("provincia_iso","fecha"),
                   by.y=c("provincia_iso","fecha"))

# We check both dataframes offers the same total results
CNE_tecnica %>%
  group_by(provincia_iso) %>%
  summarise_at(vars(num_casos), sum)
```

```
## # A tibble: 52 x 2
##    provincia_iso num_casos
##    <chr>            <int>
##  1 A               143555
##  2 AB               26916
##  3 AL               47032
##  4 AV               11084
##  5 B               382992
##  6 BA               45886
##  7 BI               80588
##  8 BU               29808
```

```
##  9 C                51272
## 10 CA               70428
## # ... with 42 more rows
```

```
CNE_casos_g %>%
  group_by(provincia_iso) %>%
  summarise_at(vars(num_casos), sum)
```

```
## # A tibble: 52 x 2
##    provincia_iso num_casos
##    <chr>             <int>
##  1 A                143555
##  2 AB                26916
##  3 AL                47032
##  4 AV                11084
##  5 B                382992
##  6 BA                45886
##  7 BI                80588
##  8 BU                29808
##  9 C                 51272
## 10 CA                70428
## # ... with 42 more rows
```

```
head(CNE_tec_cas,5)
```

```
##   provincia_iso      fecha num_casos.x num_casos_prueba_pcr
## 1             A 2020-01-01           0                    0
## 2             A 2020-01-02           0                    0
## 3             A 2020-01-03           0                    0
## 4             A 2020-01-04           0                    0
## 5             A 2020-01-05           0                    0
##   num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## 1                        0                   0                      0
## 2                        0                   0                      0
## 3                        0                   0                      0
## 4                        0                   0                      0
## 5                        0                   0                      0
##   num_casos_prueba_desconocida num_casos.y num_hosp num_uci num_def
## 1                            0           0        1       0       0
## 2                            0           0        0       0       0
## 3                            0           0        0       0       0
## 4                            0           0        0       0       0
## 5                            0           0        1       0       0
```

```
table(CNE_tec_cas$provincia_iso)
```

```
##
##   A  AB  AL  AV   B  BA  BI  BU   C  CA  CC  CE  CO  CR  CS  CU  GC  GI  GR  GU
## 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442
##   H  HU   J   L  LE  LO  LU   M  MA  ML  MU  NA   O  OR   P  PM  PO   S  SA  SE
## 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442 442
##  SG  SO  SS   T  TE  TF  TO   V  VA  VI   Z  ZA
## 442 442 442 442 442 442 442 442 442 442 442 442
```

2.2.2 GOG_CNE

- GOG_CNE -> CNE_tec_cas + Google

Here we merge by columns “provincia_iso” / “fecha” and “iso_3166_2_code” / “date”.

```
# New dataframe GOG_CNE
GOG_CNE<-merge(CNE_tec_cas,
               Google,
               by.x=c("provincia_iso","fecha"),
               by.y=c("iso_code","Date"))
head(GOG_CNE,5)
```

```
##   provincia_iso      fecha num_casos.x num_casos_prueba_pcr
## 1             A 2020-02-15           1                    1
## 2             A 2020-02-16           1                    1
## 3             A 2020-02-17           1                    1
## 4             A 2020-02-18           1                    1
## 5             A 2020-02-19           1                    1
##   num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## 1                        0                   0                     0
## 2                        0                   0                     0
## 3                        0                   0                     0
## 4                        0                   0                     0
## 5                        0                   0                     0
##   num_casos_prueba_desconocida num_casos.y num_hosp num_uci num_def
## 1                            0           0        1       0       0
## 2                            0           0        0       0       0
## 3                            0           0        1       0       0
## 4                            0           0        1       0       0
## 5                            0           0        2       1       0
##       sub_region_2 retail_and_recreation_percent_change_from_baseline
## 1 Alicante/Alacant                                                 3
## 2 Alicante/Alacant                                                -2
## 3 Alicante/Alacant                                                 0
## 4 Alicante/Alacant                                                -5
## 5 Alicante/Alacant                                                 1
##   grocery_and_pharmacy_percent_change_from_baseline
## 1                                                -1
## 2                                                 1
## 3                                                 2
## 4                                                -2
## 5                                                 1
##   parks_percent_change_from_baseline
## 1                                 34
## 2                                  8
## 3                                  9
## 4                                -14
## 5                                 10
##   transit_stations_percent_change_from_baseline
## 1                                             7
## 2                                             5
## 3                                             7
## 4                                            -2
## 5                                             3
##   workplaces_percent_change_from_baseline
## 1                                       0
## 2                                      -2
## 3                                       3
```

```
## 4                                       2
## 5                                       3
##   residential_percent_change_from_baseline
## 1                                       -1
## 2                                       -1
## 3                                        0
## 4                                        1
## 5                                        0
```

```
table(GOG_CNE$provincia_iso)
```

```
##
##   A  AB  AL  AV   B  BA  BI  BU   C  CA  CC  CE  CO  CR  CS  CU  GC  GI  GR  GU
## 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385
##   H  HU   J   L  LE  LO  LU   M  MA  ML  MU  NA   O  OR   P  PM  PO   S  SA  SE
## 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385 385
##  SG  SO  SS   T  TE  TF  TO   V  VA  VI   Z  ZA
## 385 385 385 385 385 385 385 385 385 385 385 385
```

2.2.3 Total

- Total -> GOG_CNE + EM3

Here we merge by columns "sub_region_2" / "fecha" and "Zonas.de.movilidad" / "Periodo". With this dataset we have 21 features for study.

```
# New dataframe Total
Total<-merge(GOG_CNE,
             EM3,
             by.x=c("sub_region_2","fecha"),
             by.y=c("Zonas.de.movilidad","Periodo"))
head(Total,5)
```

```
##   sub_region_2      fecha provincia_iso num_casos.x num_casos_prueba_pcr
## 1     Albacete 2020-03-16            AB         137                  132
## 2     Albacete 2020-03-17            AB         128                  123
## 3     Albacete 2020-03-18            AB         114                  107
## 4     Albacete 2020-03-19            AB         149                  133
## 5     Albacete 2020-03-20            AB         131                  121
##   num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## 1                        5                   0                      0
## 2                        5                   0                      0
## 3                        7                   0                      0
## 4                       16                   0                      0
## 5                       10                   0                      0
##   num_casos_prueba_desconocida num_casos.y num_hosp num_uci num_def
## 1                            0          65       43       3       7
## 2                            0          29       40       4       2
## 3                            0          26       24       7       7
## 4                            0          22       40       5       7
## 5                            0          85       63       4       6
##   retail_and_recreation_percent_change_from_baseline
## 1                                                -81
## 2                                                -84
## 3                                                -83
## 4                                                -93
```

```
## 5                                                -87
##   grocery_and_pharmacy_percent_change_from_baseline
## 1                                               -32
## 2                                               -41
## 3                                               -32
## 4                                               -92
## 5                                               -34
##   parks_percent_change_from_baseline
## 1                                -73
## 2                                -74
## 3                                -70
## 4                                -80
## 5                                -74
##   transit_stations_percent_change_from_baseline
## 1                                           -66
## 2                                           -72
## 3                                           -70
## 4                                           -86
## 5                                           -76
##   workplaces_percent_change_from_baseline
## 1                                     -51
## 2                                     -56
## 3                                     -58
## 4                                     -85
## 5                                     -68
##   residential_percent_change_from_baseline Total
## 1                                       22 9.900
## 2                                       23 9.705
## 3                                       23 9.510
## 4                                       35 9.130
## 5                                       32 8.750
```

```
head(str(Total,vec.len=1))
```

```
## 'data.frame':    15080 obs. of  20 variables:
##  $ sub_region_2                                       : chr  "Albacete" ...
##  $ fecha                                              : Date, format: "2020-03-16" ...
##  $ provincia_iso                                      : chr  "AB" ...
##  $ num_casos.x                                        : int  137 128 ...
##  $ num_casos_prueba_pcr                               : int  132 123 ...
##  $ num_casos_prueba_test_ac                           : int  5 5 ...
##  $ num_casos_prueba_ag                                : int  0 0 ...
##  $ num_casos_prueba_elisa                             : int  0 0 ...
##  $ num_casos_prueba_desconocida                       : int  0 0 ...
##  $ num_casos.y                                        : int  65 29 ...
##  $ num_hosp                                           : int  43 40 ...
##  $ num_uci                                            : int  3 4 ...
##  $ num_def                                            : int  7 2 ...
##  $ retail_and_recreation_percent_change_from_baseline: num  -81 -84 ...
##  $ grocery_and_pharmacy_percent_change_from_baseline : num  -32 -41 ...
##  $ parks_percent_change_from_baseline                 : num  -73 -74 ...
##  $ transit_stations_percent_change_from_baseline      : num  -66 -72 ...
##  $ workplaces_percent_change_from_baseline            : num  -51 -56 ...
##  $ residential_percent_change_from_baseline           : num  22 23 ...
##  $ Total                                              : num  9.9 ...
```

```
## NULL
```

```
summary(Total)
```

```
##  sub_region_2           fecha            provincia_iso       num_casos.x
##  Length:15080       Min.   :2020-03-16   Length:15080       Min.   :   0
##  Class :character   1st Qu.:2020-05-27   Class :character   1st Qu.:   5
##  Mode  :character   Median :2020-08-07   Mode  :character   Median :  39
##                     Mean   :2020-08-07                      Mean   : 126
##                     3rd Qu.:2020-10-19                      3rd Qu.: 120
##                     Max.   :2020-12-30                      Max.   :6565
##  num_casos_prueba_pcr num_casos_prueba_test_ac num_casos_prueba_ag
##  Min.   :   0.0       Min.   : 0.0000          Min.   :   0.00
##  1st Qu.:   5.0       1st Qu.: 0.0000          1st Qu.:   0.00
##  Median :  35.0       Median : 0.0000          Median :   0.00
##  Mean   : 110.2       Mean   : 0.2832          Mean   :  15.19
##  3rd Qu.: 105.0       3rd Qu.: 0.0000          3rd Qu.:   4.00
##  Max.   :6546.0       Max.   :32.0000          Max.   :1465.00
##  num_casos_prueba_elisa num_casos_prueba_desconocida  num_casos.y
##  Min.   : 0.0000        Min.   : 0.0000              Min.   :   0
##  1st Qu.: 0.0000        1st Qu.: 0.0000              1st Qu.:   6
##  Median : 0.0000        Median : 0.0000              Median :  37
##  Mean   : 0.1989        Mean   : 0.1317              Mean   : 127
##  3rd Qu.: 0.0000        3rd Qu.: 0.0000              3rd Qu.: 117
##  Max.   :71.0000        Max.   :65.0000              Max.   :7724
##     num_hosp          num_uci           num_def
##  Min.   :   0.00   Min.   :  0.000   Min.   :  0.000
##  1st Qu.:   1.00   1st Qu.:  0.000   1st Qu.:  0.000
##  Median :   4.00   Median :  0.000   Median :  1.000
##  Mean   :  14.86   Mean   :  1.281   Mean   :  3.437
##  3rd Qu.:  12.00   3rd Qu.:  1.000   3rd Qu.:  3.000
##  Max.   :1930.00   Max.   :135.000   Max.   :334.000
##  retail_and_recreation_percent_change_from_baseline
##  Min.   :-97.00
##  1st Qu.:-57.00
##  Median :-30.00
##  Mean   :-37.29
##  3rd Qu.:-17.00
##  Max.   : 71.00
##  grocery_and_pharmacy_percent_change_from_baseline
##  Min.   :-96.00
##  1st Qu.:-24.00
##  Median : -6.00
##  Mean   :-11.75
##  3rd Qu.:  4.00
##  Max.   :194.00
##  parks_percent_change_from_baseline
##  Min.   :-94.000
##  1st Qu.:-30.000
##  Median : -2.000
##  Mean   :  5.809
##  3rd Qu.: 30.000
##  Max.   :543.000
##  transit_stations_percent_change_from_baseline
##  Min.   :-100.00
```

```
##  1st Qu.: -53.00
##  Median : -31.00
##  Mean   : -35.19
##  3rd Qu.: -17.00
##  Max.   :  74.00
##  workplaces_percent_change_from_baseline
##  Min.   :-92.00
##  1st Qu.:-43.00
##  Median :-26.00
##  Mean   :-29.08
##  3rd Qu.:-13.00
##  Max.   : 55.00
##  residential_percent_change_from_baseline     Total
##  Min.   :-10.00                           Min.   : 1.95
##  1st Qu.:  4.00                           1st Qu.:11.36
##  Median :  7.00                           Median :14.39
##  Mean   : 10.14                           Mean   :14.20
##  3rd Qu.: 14.00                           3rd Qu.:17.11
##  Max.   : 48.00                           Max.   :29.00
```

```
table(Total$sub_region_2)
```

```
##
##               Albacete       Alicante/Alacant                Almería
##                    290                    290                    290
##            Araba/Álava               Asturias                  Ávila
##                    290                    290                    290
##                Badajoz        Balears, Illes              Barcelona
##                    290                    290                    290
##                Bizkaia                 Burgos                Cáceres
##                    290                    290                    290
##                 Cádiz              Cantabria     Castellón/Castelló
##                    290                    290                    290
##                 Ceuta            Ciudad Real                Córdoba
##                    290                    290                    290
##              Coruña, A                 Cuenca               Gipuzkoa
##                    290                    290                    290
##                Girona                Granada            Guadalajara
##                    290                    290                    290
##                Huelva                 Huesca                  Jaén
##                    290                    290                    290
##                 León                  Lleida                  Lugo
##                    290                    290                    290
##                Madrid                 Málaga                Melilla
##                    290                    290                    290
##                Murcia                Navarra                Ourense
##                    290                    290                    290
##              Palencia            Palmas, Las             Pontevedra
##                    290                    290                    290
##             Rioja, La              Salamanca Santa Cruz de Tenerife
##                    290                    290                    290
##               Segovia                Sevilla                  Soria
##                    290                    290                    290
##             Tarragona                 Teruel                 Toledo
##                    290                    290                    290
```

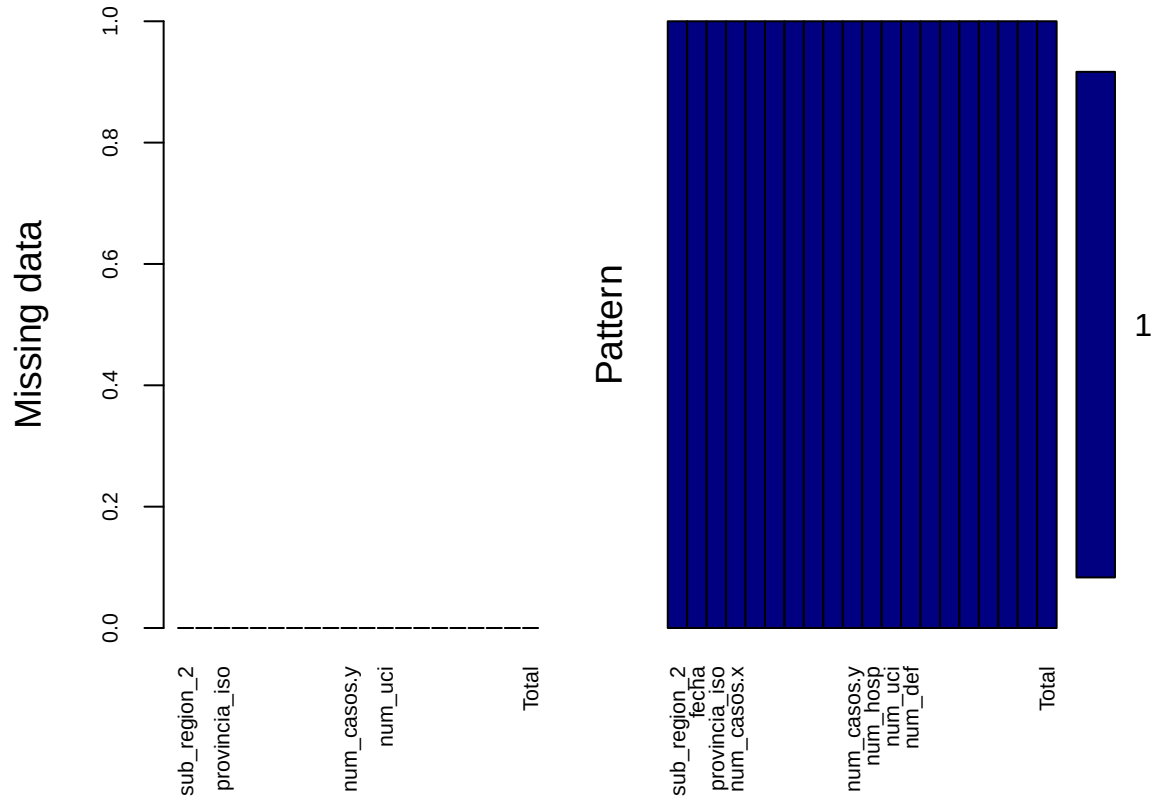
```
##      Valencia/València             Valladolid                 Zamora
##                    290                     290                    290
##               Zaragoza
##                    290
```

```
table(Total$provincia_iso)
```

```
##
##   A  AB  AL  AV   B  BA  BI  BU   C  CA  CC  CE  CO  CR  CS  CU  GC  GI  GR  GU
## 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290
##   H  HU   J   L  LE  LO  LU   M  MA  ML  MU  NA   O  OR   P  PM  PO   S  SA  SE
## 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290
##  SG  SO  SS   T  TE  TF  TO   V  VA  VI   Z  ZA
## 290 290 290 290 290 290 290 290 290 290 290 290
```

We check the missing values. We should have zero missing values

```
aggr(Total, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Total), cex.axis=.7,
     gap=3, ylab=c("Missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##                                       Variable Count
##                                   sub_region_2     0
##                                          fecha     0
##                                   provincia_iso     0
```

```
##                                       num_casos.x     0
##                                  num_casos_prueba_pcr     0
##                               num_casos_prueba_test_ac     0
##                                    num_casos_prueba_ag     0
##                                 num_casos_prueba_elisa     0
##                            num_casos_prueba_desconocida     0
##                                       num_casos.y     0
##                                          num_hosp     0
##                                           num_uci     0
##                                           num_def     0
##  retail_and_recreation_percent_change_from_baseline     0
##   grocery_and_pharmacy_percent_change_from_baseline     0
##                   parks_percent_change_from_baseline     0
##       transit_stations_percent_change_from_baseline     0
##             workplaces_percent_change_from_baseline     0
##           residential_percent_change_from_baseline     0
##                                               Total     0
```

```
Total %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing)) %>%
    ggplot() +
    geom_bar(aes(x=key, y=num.missing), stat = 'identity',fill="#F0E442") +
    labs(x='variable', y="number of missing values",
         title='Number of missing values') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


```
# Review results
# Discrepancies due to different time-frames when merge CNE dataframes (see previous checks)
Total %>%
  group_by(provincia_iso) %>%
  summarise_at(vars(num_casos.x,num_casos.y), sum)
```

```
## # A tibble: 52 x 3
##    provincia_iso num_casos.x num_casos.y
##    <chr>               <int>       <int>
##  1 A                   56493       55068
##  2 AB                  16459       16626
##  3 AL                  21488       21372
##  4 AV                   6525        6681
##  5 B                  257034      261208
##  6 BA                  23165       22612
##  7 BI                  56867       57728
##  8 BU                  22742       22978
##  9 C                   25641       25604
## 10 CA                  31593       31225
## # ... with 42 more rows
```

```
# CSV file generation
head(Total,5)
```

```
##   sub_region_2      fecha provincia_iso num_casos.x num_casos_prueba_pcr
## 1     Albacete 2020-03-16            AB         137                  132
## 2     Albacete 2020-03-17            AB         128                  123
```

```
## 3     Albacete 2020-03-18            AB         114                  107
## 4     Albacete 2020-03-19            AB         149                  133
## 5     Albacete 2020-03-20            AB         131                  121
##   num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## 1                        5                   0                      0
## 2                        5                   0                      0
## 3                        7                   0                      0
## 4                       16                   0                      0
## 5                       10                   0                      0
##   num_casos_prueba_desconocida num_casos.y num_hosp num_uci num_def
## 1                            0          65       43       3       7
## 2                            0          29       40       4       2
## 3                            0          26       24       7       7
## 4                            0          22       40       5       7
## 5                            0          85       63       4       6
##   retail_and_recreation_percent_change_from_baseline
## 1                                                -81
## 2                                                -84
## 3                                                -83
## 4                                                -93
## 5                                                -87
##   grocery_and_pharmacy_percent_change_from_baseline
## 1                                               -32
## 2                                               -41
## 3                                               -32
## 4                                               -92
## 5                                               -34
##   parks_percent_change_from_baseline
## 1                                -73
## 2                                -74
## 3                                -70
## 4                                -80
## 5                                -74
##   transit_stations_percent_change_from_baseline
## 1                                           -66
## 2                                           -72
## 3                                           -70
## 4                                           -86
## 5                                           -76
##   workplaces_percent_change_from_baseline
## 1                                     -51
## 2                                     -56
## 3                                     -58
## 4                                     -85
## 5                                     -68
##   residential_percent_change_from_baseline Total
## 1                                       22 9.900
## 2                                       23 9.705
## 3                                       23 9.510
## 4                                       35 9.130
## 5                                       32 8.750
```

```
str(Total)
```

```
## 'data.frame':    15080 obs. of  20 variables:
```

```
##  $ sub_region_2                                        : chr  "Albacete" "Albacete" "Albacete" "Albace
##  $ fecha                                               : Date, format: "2020-03-16" "2020-03-17" ...
##  $ provincia_iso                                       : chr  "AB" "AB" "AB" "AB" ...
##  $ num_casos.x                                         : int  137 128 114 149 131 129 125 112 107 78 .
##  $ num_casos_prueba_pcr                                : int  132 123 107 133 121 120 112 103 91 64 ..
##  $ num_casos_prueba_test_ac                            : int  5 5 7 16 10 9 13 9 16 14 ...
##  $ num_casos_prueba_ag                                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ num_casos_prueba_elisa                              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ num_casos_prueba_desconocida                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ num_casos.y                                         : int  65 29 26 22 85 24 60 194 53 22 ...
##  $ num_hosp                                            : int  43 40 24 40 63 60 61 108 76 66 ...
##  $ num_uci                                             : int  3 4 7 5 4 2 8 2 7 9 ...
##  $ num_def                                             : int  7 2 7 7 6 10 13 16 14 16 ...
##  $ retail_and_recreation_percent_change_from_baseline: num  -81 -84 -83 -93 -87 -92 -94 -85 -87 -87
##  $ grocery_and_pharmacy_percent_change_from_baseline : num  -32 -41 -32 -92 -34 -60 -80 -46 -55 -55
##  $ parks_percent_change_from_baseline                : num  -73 -74 -70 -80 -74 -85 -88 -77 -81 -76
##  $ transit_stations_percent_change_from_baseline     : num  -66 -72 -70 -86 -76 -80 -89 -76 -79 -78
##  $ workplaces_percent_change_from_baseline           : num  -51 -56 -58 -85 -68 -64 -66 -61 -64 -65
##  $ residential_percent_change_from_baseline          : num  22 23 23 35 32 27 23 26 28 28 ...
##  $ Total                                               : num  9.9 9.71 9.51 9.13 8.75 ...
```

```
summary(Total)
```

```
##  sub_region_2           fecha            provincia_iso       num_casos.x
##  Length:15080       Min.   :2020-03-16   Length:15080       Min.   :   0
##  Class :character   1st Qu.:2020-05-27   Class :character   1st Qu.:   5
##  Mode  :character   Median :2020-08-07   Mode  :character   Median :  39
##                     Mean   :2020-08-07                      Mean   : 126
##                     3rd Qu.:2020-10-19                      3rd Qu.: 120
##                     Max.   :2020-12-30                      Max.   :6565
##  num_casos_prueba_pcr num_casos_prueba_test_ac num_casos_prueba_ag
##  Min.   :   0.0       Min.   : 0.0000          Min.   :   0.00
##  1st Qu.:   5.0       1st Qu.: 0.0000          1st Qu.:   0.00
##  Median :  35.0       Median : 0.0000          Median :   0.00
##  Mean   : 110.2       Mean   : 0.2832          Mean   :  15.19
##  3rd Qu.: 105.0       3rd Qu.: 0.0000          3rd Qu.:   4.00
##  Max.   :6546.0       Max.   :32.0000          Max.   :1465.00
##  num_casos_prueba_elisa num_casos_prueba_desconocida  num_casos.y
##  Min.   : 0.0000        Min.   : 0.0000              Min.   :   0
##  1st Qu.: 0.0000        1st Qu.: 0.0000              1st Qu.:   6
##  Median : 0.0000        Median : 0.0000              Median :  37
##  Mean   : 0.1989        Mean   : 0.1317              Mean   : 127
##  3rd Qu.: 0.0000        3rd Qu.: 0.0000              3rd Qu.: 117
##  Max.   :71.0000        Max.   :65.0000              Max.   :7724
##     num_hosp          num_uci           num_def
##  Min.   :   0.00   Min.   :  0.000   Min.   :  0.000
##  1st Qu.:   1.00   1st Qu.:  0.000   1st Qu.:  0.000
##  Median :   4.00   Median :  0.000   Median :  1.000
##  Mean   :  14.86   Mean   :  1.281   Mean   :  3.437
##  3rd Qu.:  12.00   3rd Qu.:  1.000   3rd Qu.:  3.000
##  Max.   :1930.00   Max.   :135.000   Max.   :334.000
##  retail_and_recreation_percent_change_from_baseline
##  Min.   :-97.00
##  1st Qu.:-57.00
##  Median :-30.00
```

```
##  Mean   :-37.29
##  3rd Qu.:-17.00
##  Max.   : 71.00
##  grocery_and_pharmacy_percent_change_from_baseline
##  Min.   :-96.00
##  1st Qu.:-24.00
##  Median : -6.00
##  Mean   :-11.75
##  3rd Qu.:  4.00
##  Max.   :194.00
##  parks_percent_change_from_baseline
##  Min.   :-94.000
##  1st Qu.:-30.000
##  Median : -2.000
##  Mean   :  5.809
##  3rd Qu.: 30.000
##  Max.   :543.000
##  transit_stations_percent_change_from_baseline
##  Min.   :-100.00
##  1st Qu.: -53.00
##  Median : -31.00
##  Mean   : -35.19
##  3rd Qu.: -17.00
##  Max.   :  74.00
##  workplaces_percent_change_from_baseline
##  Min.   :-92.00
##  1st Qu.:-43.00
##  Median :-26.00
##  Mean   :-29.08
##  3rd Qu.:-13.00
##  Max.   : 55.00
##  residential_percent_change_from_baseline     Total
##  Min.   :-10.00                           Min.   : 1.95
##  1st Qu.:  4.00                           1st Qu.:11.36
##  Median :  7.00                           Median :14.39
##  Mean   : 10.14                           Mean   :14.20
##  3rd Qu.: 14.00                           3rd Qu.:17.11
##  Max.   : 48.00                           Max.   :29.00
```

```
table(Total$provincia_iso)
```

```
##
##   A  AB  AL  AV   B  BA  BI  BU   C  CA  CC  CE  CO  CR  CS  CU  GC  GI  GR  GU
## 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290
##   H  HU   J   L  LE  LO  LU   M  MA  ML  MU  NA   O  OR   P  PM  PO   S  SA  SE
## 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290 290
##  SG  SO  SS   T  TE  TF  TO   V  VA  VI   Z  ZA
## 290 290 290 290 290 290 290 290 290 290 290 290
```

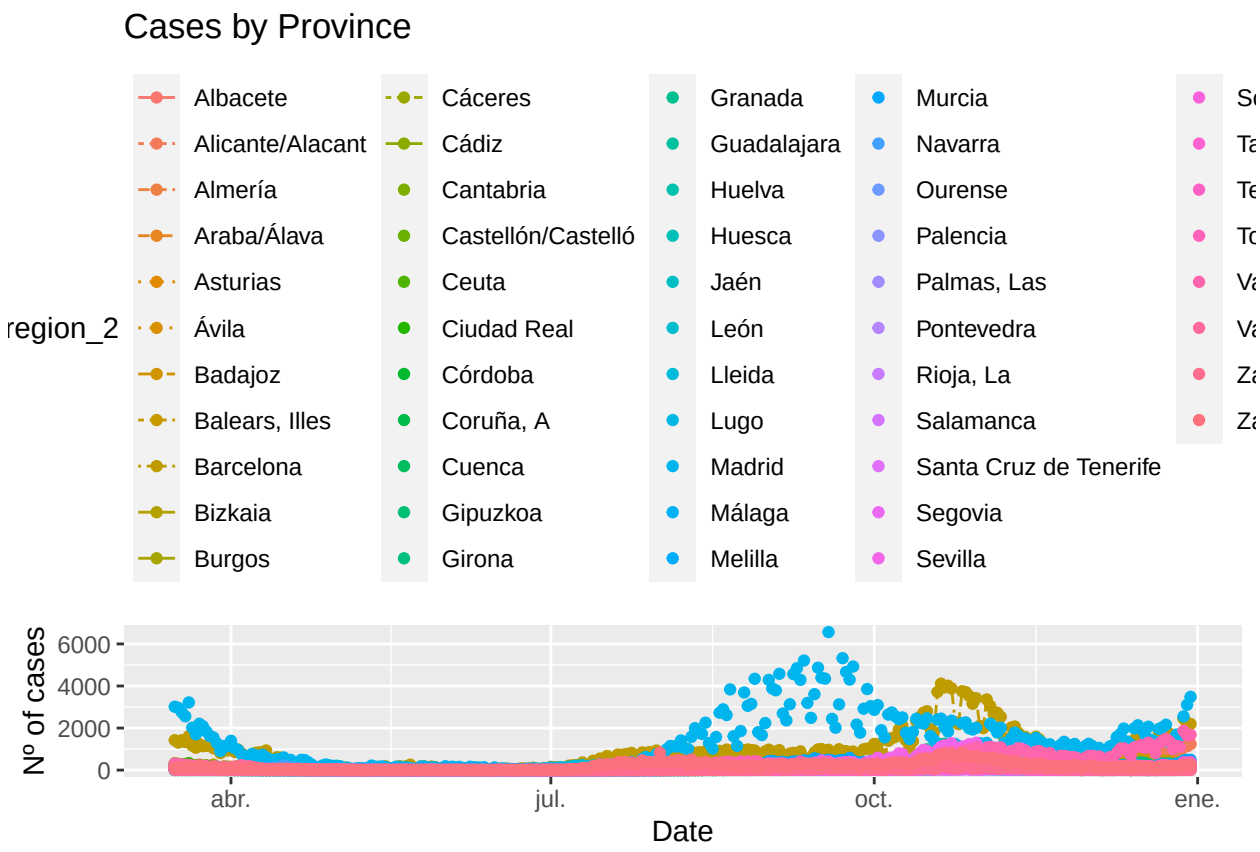
```
write.csv2(Total,"D:\\UOC Master Data Science\\_ M2.882 - TFM - Área 5\\UOC - Guia - PECS\\Pec3\\Total.
```

2.3 Visual analysis

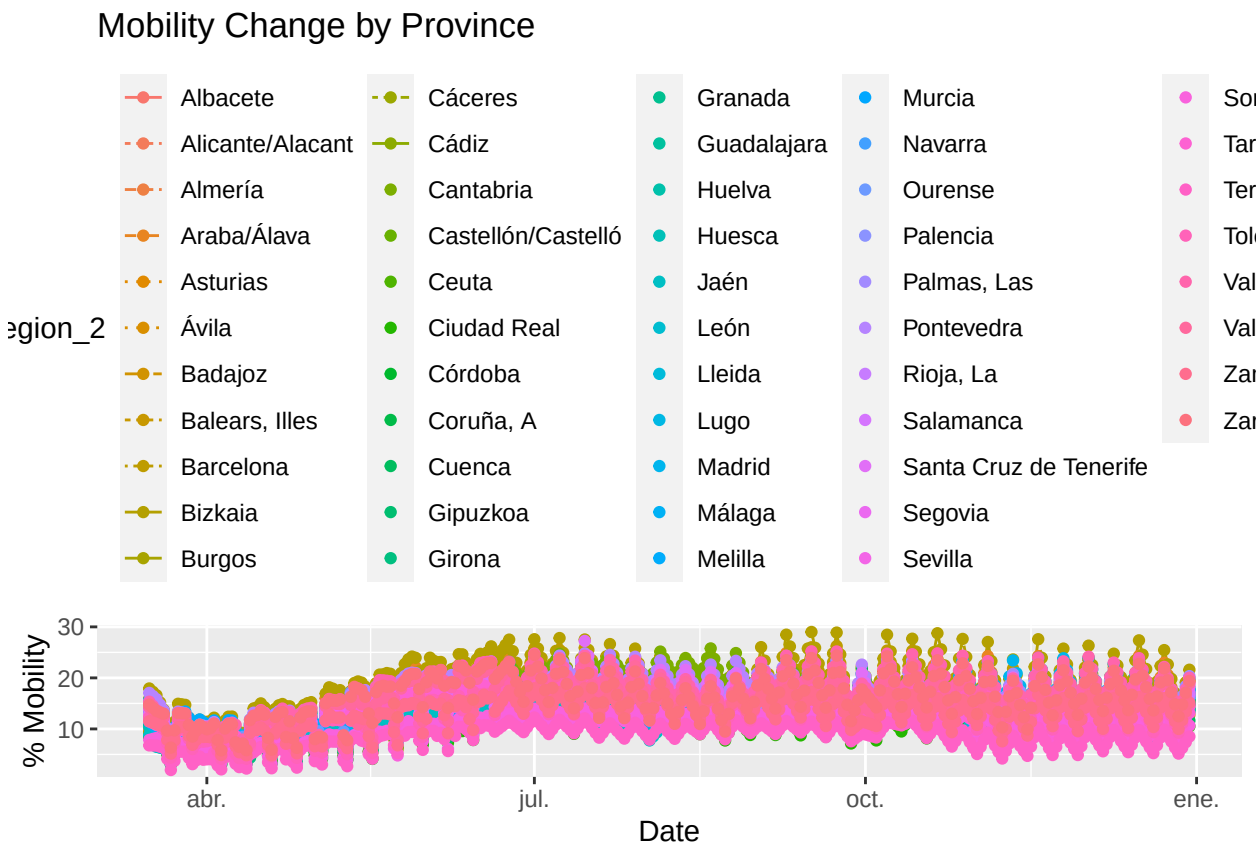
2.3.1 Dataframe plots

We have generated some plots from the **dataframe** object generated.

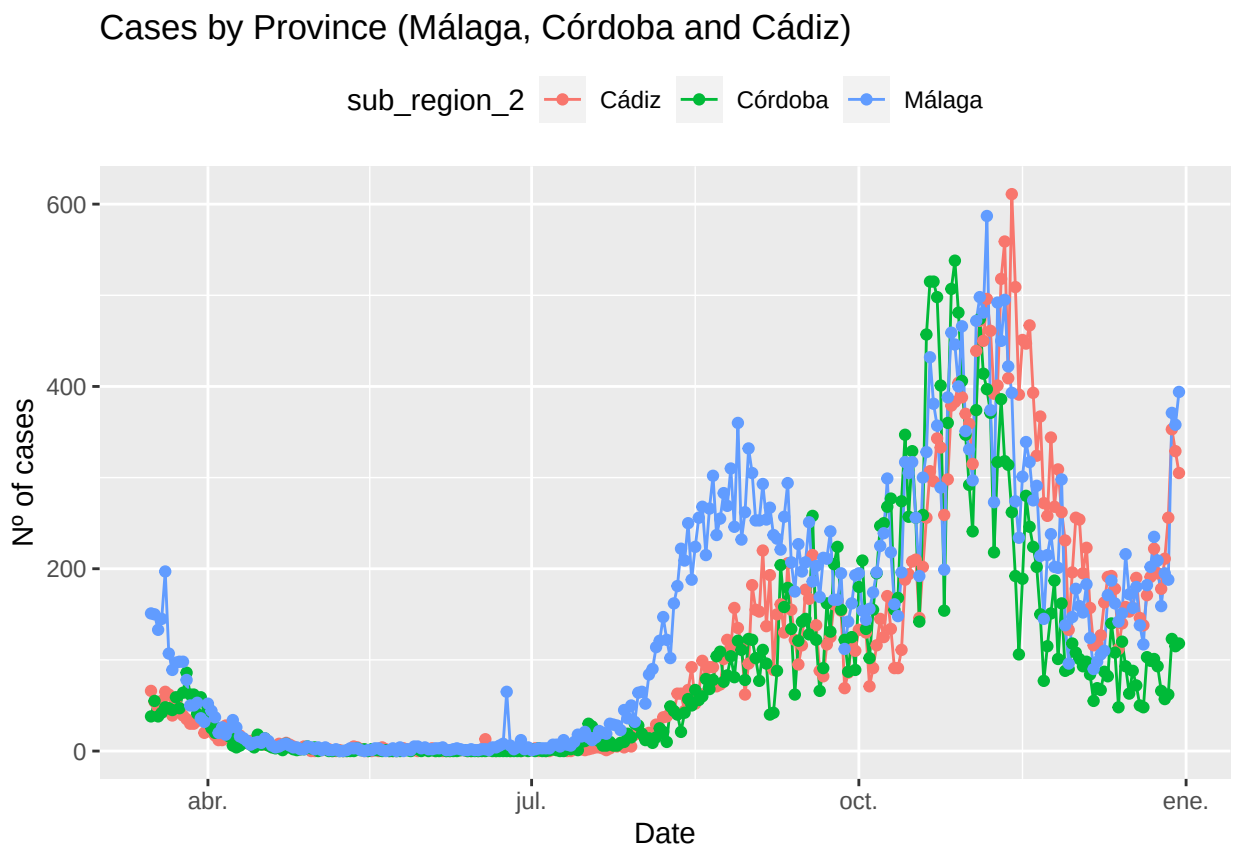
```
# Line plots
# All num_casos.x
ggplot(Total, aes(x=fecha, y=num_casos.x, group=sub_region_2)) +
  geom_line(aes(linetype=sub_region_2, color=sub_region_2))+
  geom_point(aes(color=sub_region_2))+
  theme(legend.position="top") +
  labs(title="Cases by Province",
       x ="Date", y = "Nº of cases")
```



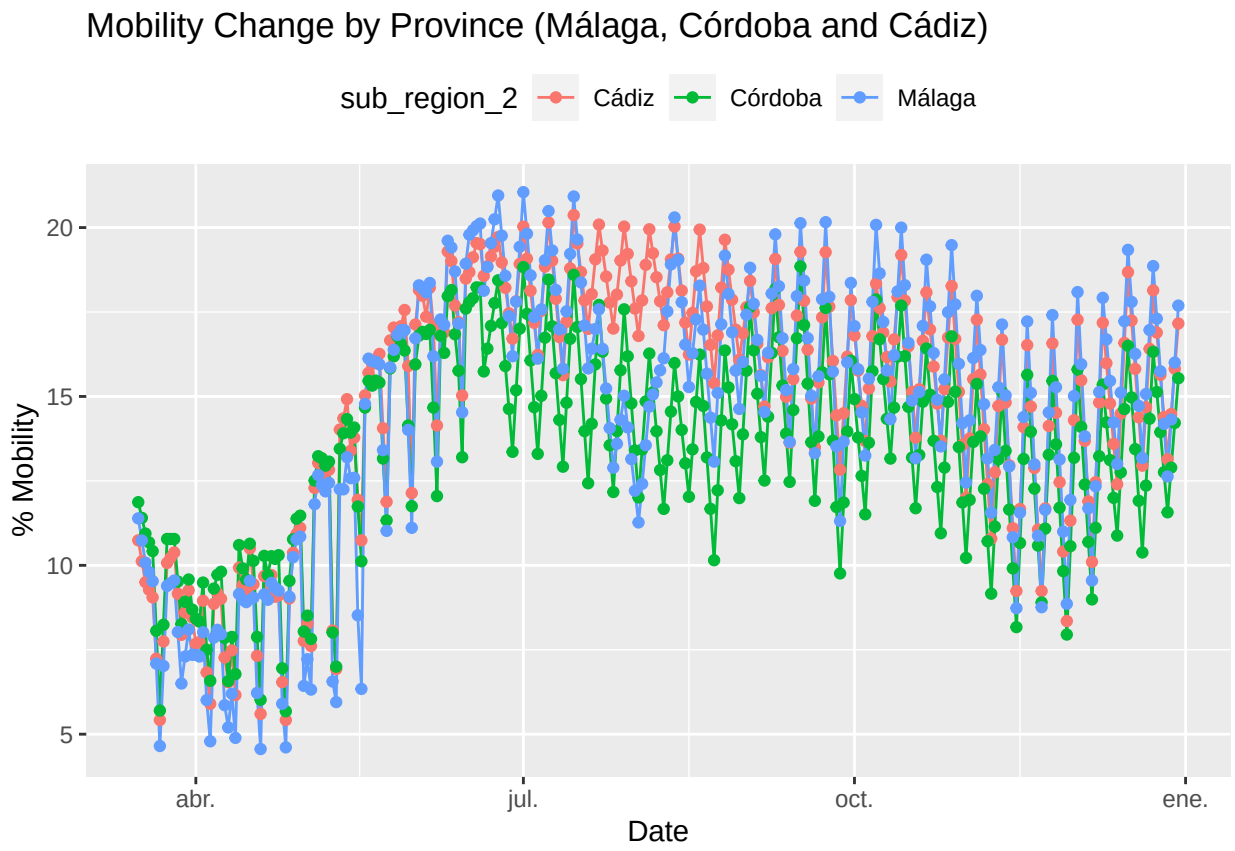

```
# All Total (mobility)
ggplot(Total, aes(x=fecha, y=Total, group=sub_region_2)) +
  geom_line(aes(linetype=sub_region_2, color=sub_region_2))+
  geom_point(aes(color=sub_region_2))+
  theme(legend.position="top") +
  labs(title="Mobility Change by Province",
       x ="Date", y = "% Mobility")
```




```
# Mal, Cor and Cad - num_casos.x
Total %>%
  filter(sub_region_2 == "Málaga" | sub_region_2 == "Cádiz" |
         sub_region_2 == "Córdoba") %>%
  ggplot(aes(x=fecha, y=num_casos.x))+
    geom_line(aes(color=sub_region_2))+
    geom_point(aes(color=sub_region_2))+
    theme(legend.position="top") +
    labs(title="Cases by Province (Málaga, Córdoba and Cádiz)",
        x ="Date", y = "Nº of cases")
```



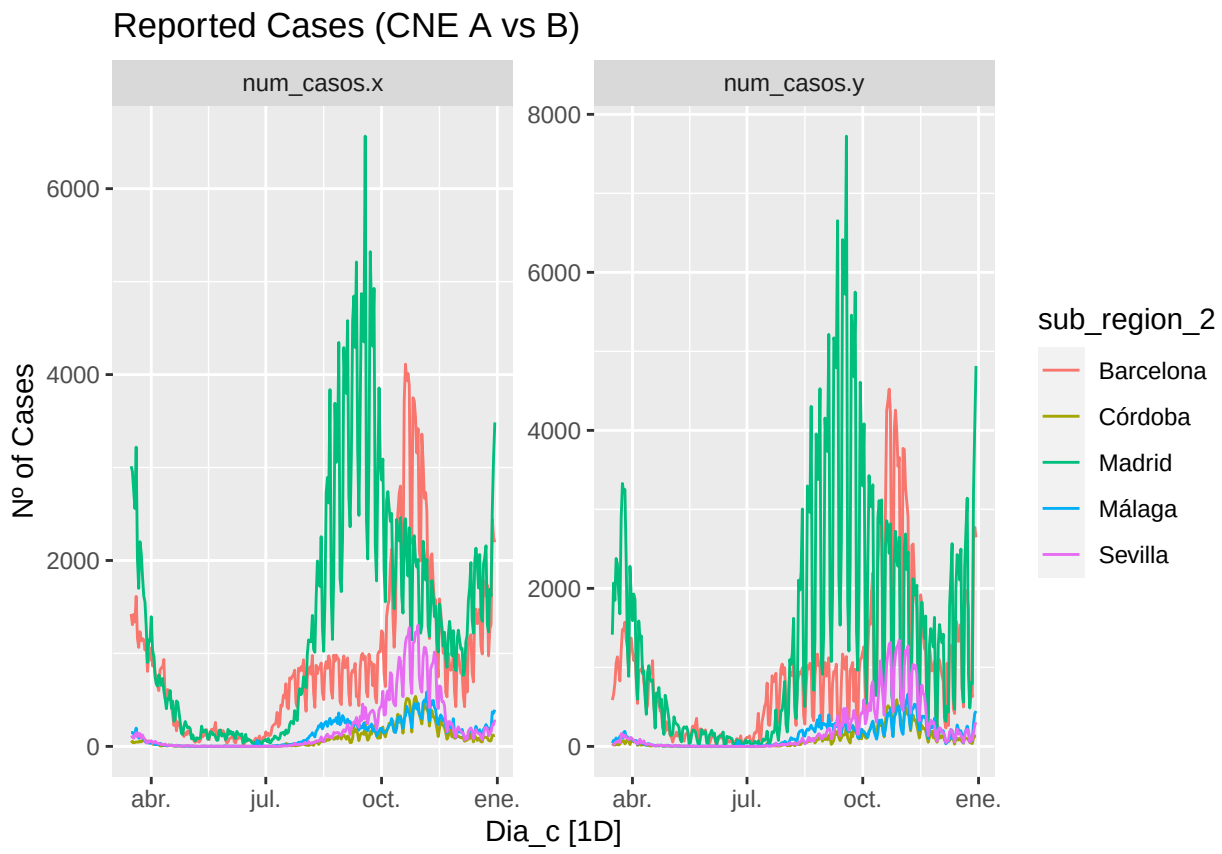

```
# Mal, Cor and Cad - Total (mobility)
Total %>%
  filter(sub_region_2 == "Málaga" | sub_region_2 == "Cádiz" |
         sub_region_2 == "Córdoba") %>%
  ggplot(aes(x=fecha, y=Total))+
    geom_line(aes(color=sub_region_2))+
    geom_point(aes(color=sub_region_2))+
    theme(legend.position="top") +
  labs(title="Mobility Change by Province (Málaga, Córdoba and Cádiz)",
        x ="Date", y = "% Mobility")
```



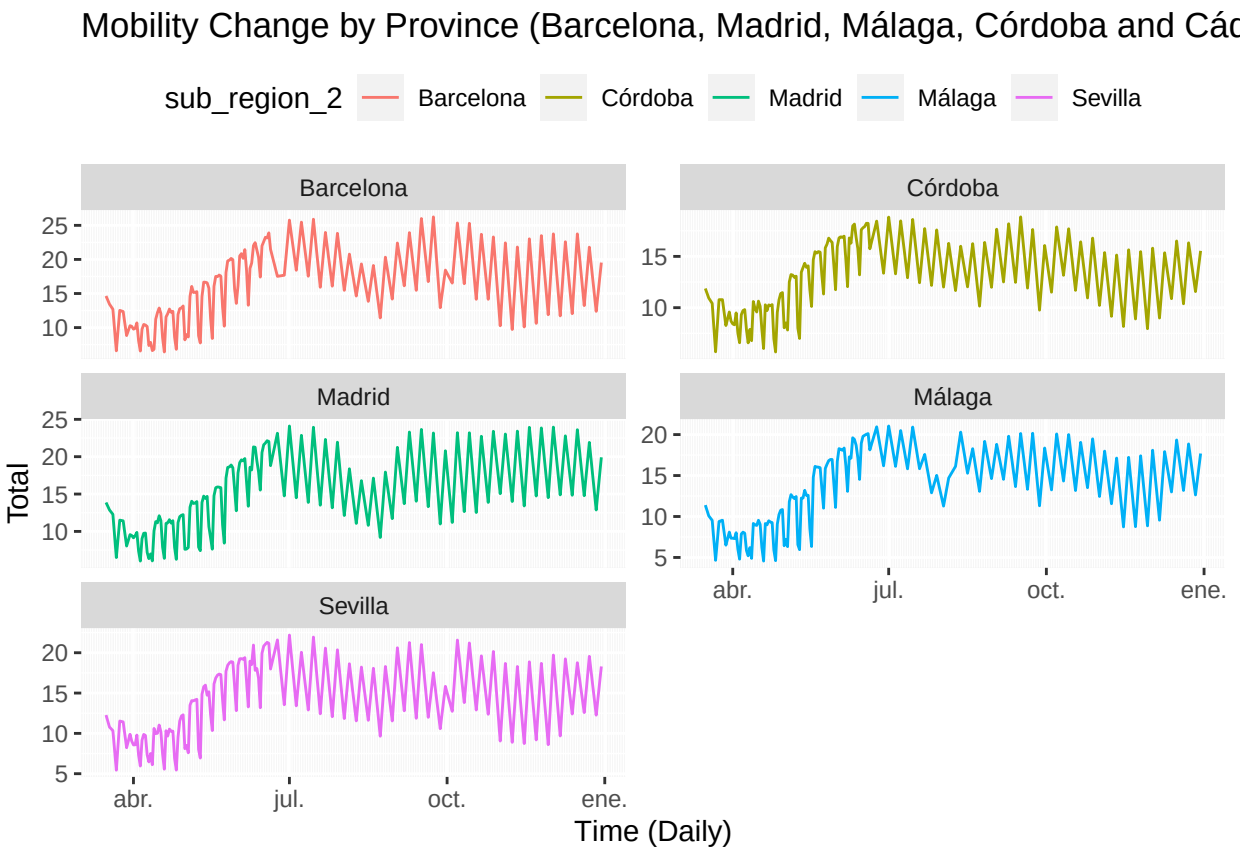
2.3.2 Time-series plots

We have generated some plots from the **time-series** object generated.

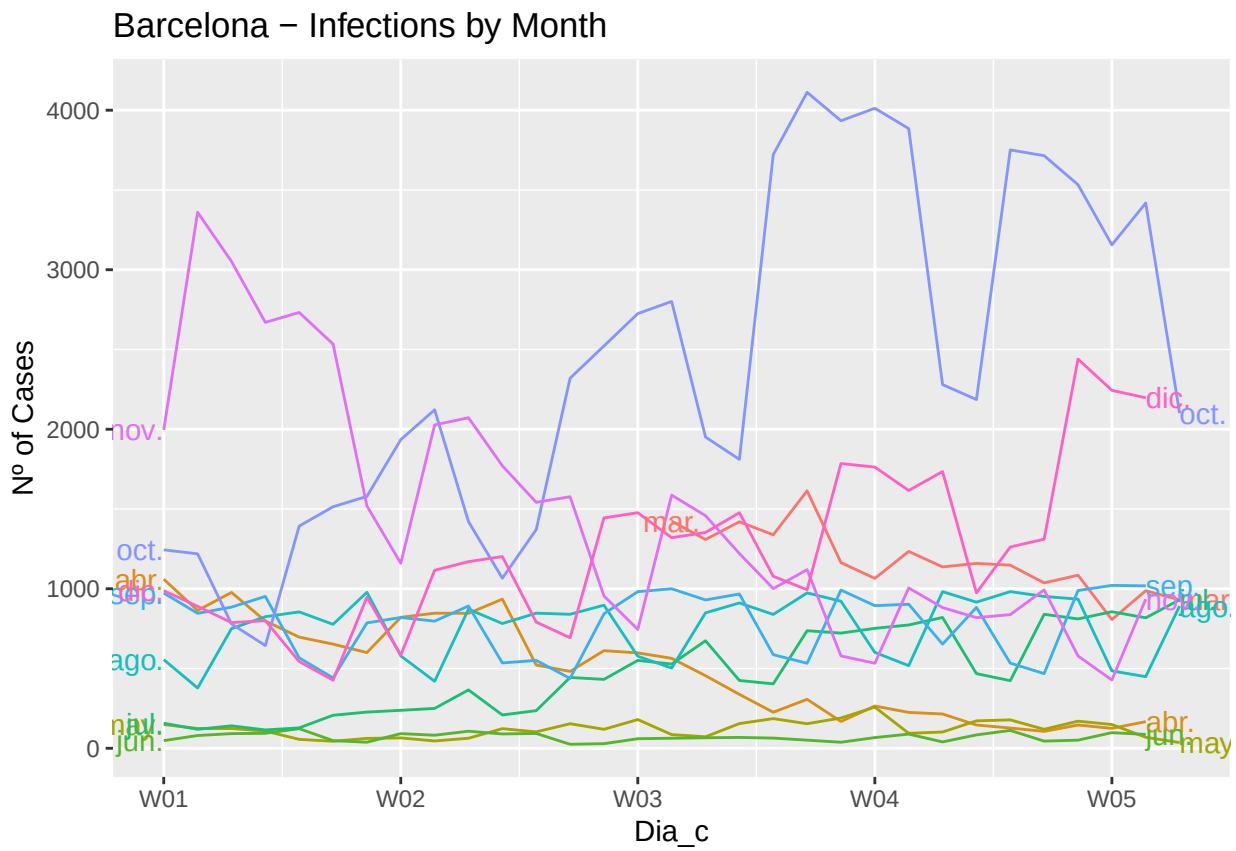
```
# Convert dataframe to ts object
Total_ts <- Total[-3] %>%
  mutate(Dia_c = as_date(fecha)) %>%
  select(-fecha) %>%
  as_tsibble(key = c(sub_region_2),
             index = Dia_c)

# Filter for Bar, Mad, Mal, Cor and, Cad
Total_ts %>% filter(sub_region_2 == "Barcelona" | sub_region_2 == "Madrid" |
                    sub_region_2 == "Málaga" | sub_region_2 == "Sevilla" |
                    sub_region_2 == "Córdoba") -> Total_ts_b

# Plots
# A num_casos.x,num_casos.y
autoplot(Total_ts_b, vars(num_casos.x,num_casos.y)) +
  labs(y = "Nº of Cases",
       title = "Reported Cases (CNE A vs B)")
```




```
# B Total (mobility)
autoplot(Total_ts_b, Total) +
  facet_wrap(~sub_region_2, scales = "free_y", ncol=2) +
  theme(legend.position = "top") +
  scale_x_date(date_minor_breaks = "1 day", name = "Time (Daily)") +
  ggtitle(label = "Mobility Change by Province (Barcelona, Madrid, Málaga, Córdoba and Cádiz)")
```

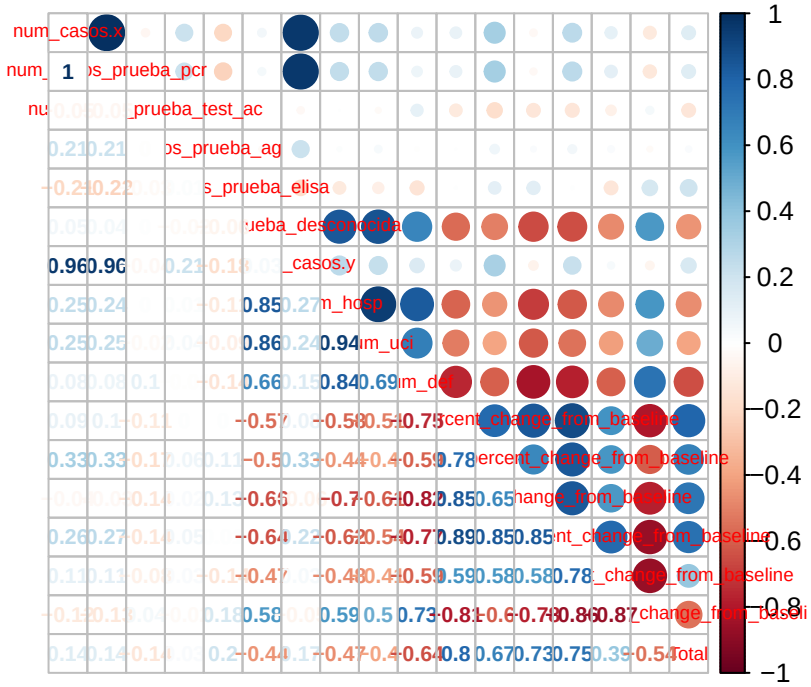



```
# C sub_region_2 == "Barcelona" by month
Total_ts %>% filter(sub_region_2 == "Barcelona") %>%
  gg_season(num_casos.x, period = "month", labels = "both") +
  theme(legend.position = "top") +
  labs(y="Nº of Cases", title="Barcelona - Infections by Month")
```

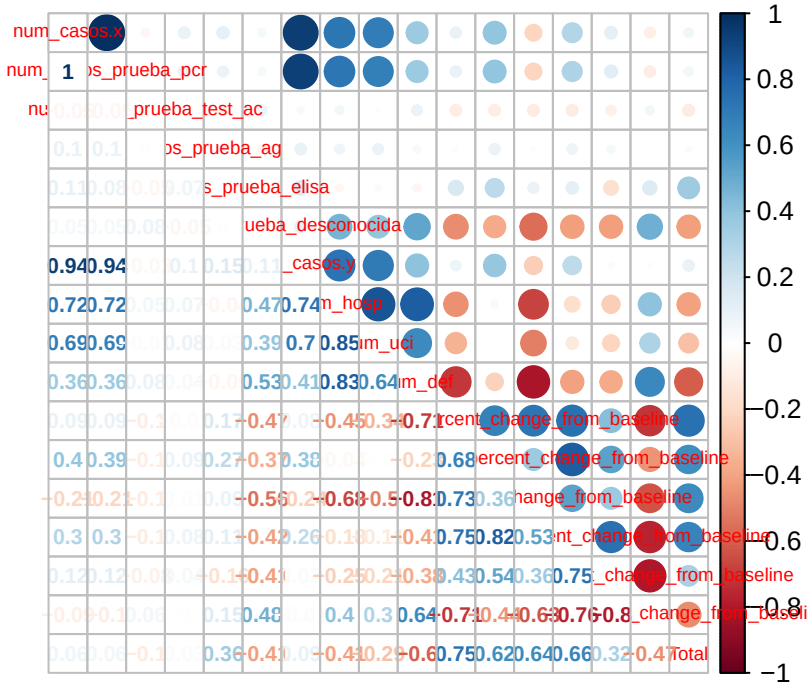


2.3.3 Correlation plots

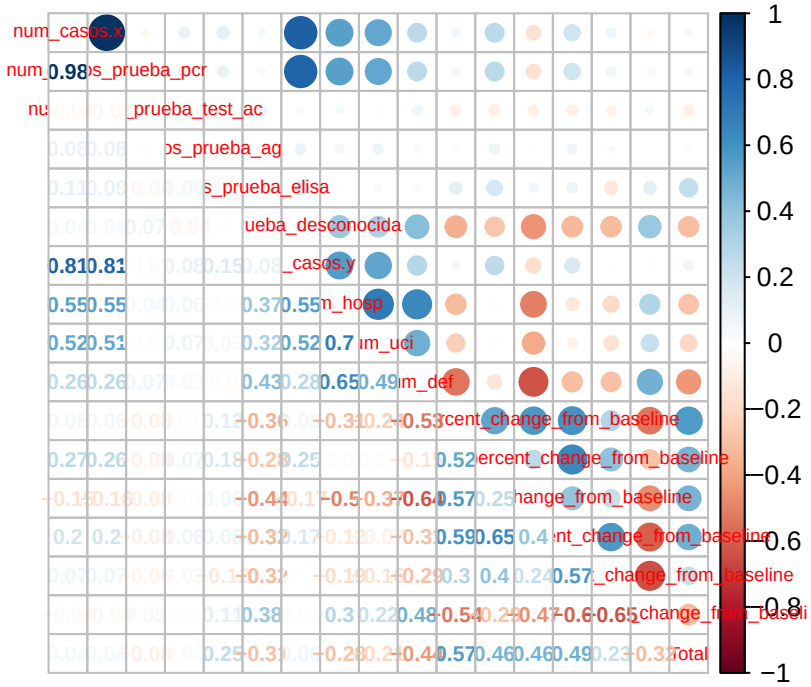
```
# Filter to "sub_region_2" == "Barcelona"
# Character / date columns are eliminated

# pearson
Total.res<-Total %>%
  filter(sub_region_2 == "Barcelona")
Total.res<-cor(Total.res[,c(-1,-2,-3)],method="pearson")
corrplot.mixed(Total.res,upper="circle",number.cex=.65,tl.cex=.6)
```




```
# spearman
Total.res<-Total %>%
  filter(sub_region_2 == "Barcelona")
Total.res<-cor(Total.res[,c(-1,-2,-3)],method="spearman")
corrplot.mixed(Total.res,upper="circle",number.cex=.65,tl.cex=.6)
```




```
# kendall
Total.res<-Total %>%
  filter(sub_region_2 == "Barcelona")
Total.res<-cor(Total.res[,c(-1,-2,-3)],method="kendall")
corrplot.mixed(Total.res,upper="circle",number.cex=.65,tl.cex=.6)
```



2.3.4 PCA

```
pca <- prcomp(Total.res, scale = T)
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.0590 1.8914 1.12142 1.02409 0.83504 0.57825 0.46266
## Proportion of Variance 0.5504 0.2104 0.07398 0.06169 0.04102 0.01967 0.01259
## Cumulative Proportion  0.5504 0.7609 0.83485 0.89655 0.93756 0.95723 0.96982
##                           PC8     PC9     PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.41006 0.33620 0.28445 0.21986 0.17734 0.16911 0.15372
## Proportion of Variance 0.00989 0.00665 0.00476 0.00284 0.00185 0.00168 0.00139
## Cumulative Proportion  0.97971 0.98636 0.99112 0.99397 0.99582 0.99750 0.99889
##                           PC15    PC16      PC17
## Standard deviation     0.1369 0.01274 1.826e-17
## Proportion of Variance 0.0011 0.00001 0.000e+00
## Cumulative Proportion  1.0000 1.00000 1.000e+00
```

```
pca$rotation
```

```
##                                                  PC1          PC2
## num_casos.x                              -0.1354826901 -0.469400471
## num_casos_prueba_pcr                     -0.1356873829 -0.469237915
## num_casos_prueba_test_ac                 -0.0782958432  0.237502756
## num_casos_prueba_ag                       0.0005323686  0.008790618
```

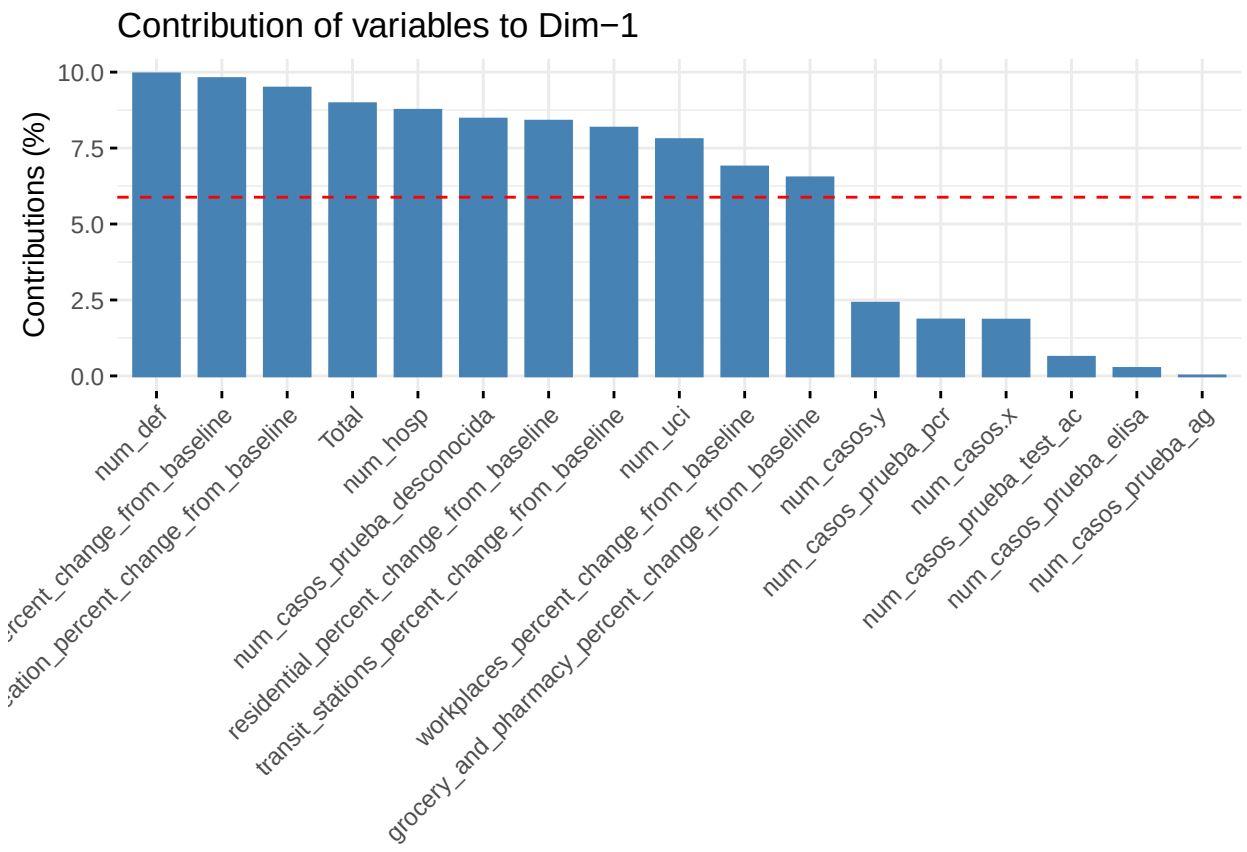
```
## num_casos_prueba_elisa                                0.0495912937  0.004385553
## num_casos_prueba_desconocida                         -0.2907430443  0.113686282
## num_casos.y                                          -0.1546876637 -0.450080493
## num_hosp                                             -0.2956764064 -0.199498016
## num_uci                                              -0.2788863801 -0.230508957
## num_def                                              -0.3153112111 -0.030904089
## retail_and_recreation_percent_change_from_baseline    0.3077977291 -0.116093786
## grocery_and_pharmacy_percent_change_from_baseline     0.2553316609 -0.261732823
## parks_percent_change_from_baseline                    0.3128404870  0.012776948
## transit_stations_percent_change_from_baseline         0.2855716515 -0.218931952
## workplaces_percent_change_from_baseline               0.2622358364 -0.170838668
## residential_percent_change_from_baseline             -0.2895648649  0.166332144
## Total                                                 0.2993175108 -0.081611685
##                                                              PC3          PC4
## num_casos.x                                          -0.002875288 -0.028445983
## num_casos_prueba_pcr                                 -0.012292151 -0.026795983
## num_casos_prueba_test_ac                             -0.392712472 -0.268237834
## num_casos_prueba_ag                                  -0.034460138  0.950253303
## num_casos_prueba_elisa                                0.811470962 -0.060011278
## num_casos_prueba_desconocida                          0.019674417 -0.078105637
## num_casos.y                                           0.041668217 -0.036946432
## num_hosp                                             -0.051419434 -0.010413059
## num_uci                                              -0.018126448  0.006323566
## num_def                                              -0.031133681  0.009747098
## retail_and_recreation_percent_change_from_baseline    0.033200491 -0.075398568
## grocery_and_pharmacy_percent_change_from_baseline     0.027494995 -0.030995413
## parks_percent_change_from_baseline                    0.029299658 -0.011018427
## transit_stations_percent_change_from_baseline        -0.099387689 -0.018021579
## workplaces_percent_change_from_baseline              -0.303553292  0.028981871
## residential_percent_change_from_baseline              0.211975965  0.003369575
## Total                                                 0.173800838 -0.065471155
##                                                              PC5          PC6
## num_casos.x                                          -0.15653475  0.16549669
## num_casos_prueba_pcr                                 -0.15256567  0.17028866
## num_casos_prueba_test_ac                             -0.81534671 -0.10365823
## num_casos_prueba_ag                                  -0.25969423 -0.01360833
## num_casos_prueba_elisa                               -0.34805371 -0.37185910
## num_casos_prueba_desconocida                          0.20580692 -0.11429059
## num_casos.y                                          -0.15827091  0.15297764
## num_hosp                                              0.03855580 -0.11688342
## num_uci                                               0.06172358 -0.06173530
## num_def                                               0.08791553 -0.26888200
## retail_and_recreation_percent_change_from_baseline -0.02731712  0.18700422
## grocery_and_pharmacy_percent_change_from_baseline     0.01445198 -0.34534708
## parks_percent_change_from_baseline                   -0.01705528  0.39173313
## transit_stations_percent_change_from_baseline         0.03520761 -0.26003250
## workplaces_percent_change_from_baseline               0.10943614 -0.48832683
## residential_percent_change_from_baseline              0.03104587  0.13040337
## Total                                                -0.05468582  0.19337305
##                                                              PC7          PC8
## num_casos.x                                           0.11538168  0.07515692
## num_casos_prueba_pcr                                  0.11048811  0.07713501
## num_casos_prueba_test_ac                             -0.08510797 -0.12117026
## num_casos_prueba_ag                                  -0.03926649 -0.14832473
```

```
## num_casos_prueba_elisa                                 0.24859208  0.01374512
## num_casos_prueba_desconocida                           0.19845835 -0.75630827
## num_casos.y                                            0.04117476 -0.02860659
## num_hosp                                              -0.11367668 -0.07611704
## num_uci                                               -0.02723992 -0.20583241
## num_def                                               -0.20890547  0.15436921
## retail_and_recreation_percent_change_from_baseline -0.10130686 -0.31700087
## grocery_and_pharmacy_percent_change_from_baseline  -0.63305922 -0.04093499
## parks_percent_change_from_baseline                     0.15076652  0.04293438
## transit_stations_percent_change_from_baseline      -0.06852096 -0.19335324
## workplaces_percent_change_from_baseline                0.37588850  0.19906444
## residential_percent_change_from_baseline           -0.39494991  0.23079899
## Total                                                 -0.26415699 -0.26912855
##                                                              PC9         PC10
## num_casos.x                                           -0.204798939 -0.059620970
## num_casos_prueba_pcr                                  -0.206074966 -0.057649643
## num_casos_prueba_test_ac                               0.031408747  0.010328003
## num_casos_prueba_ag                                   -0.027397924 -0.004813823
## num_casos_prueba_elisa                                 0.085159377 -0.047189805
## num_casos_prueba_desconocida                          -0.349092565 -0.125363275
## num_casos.y                                           -0.130984317  0.031574368
## num_hosp                                               0.283065654  0.067094187
## num_uci                                                0.717067401  0.169083317
## num_def                                               -0.009119052  0.013475201
## retail_and_recreation_percent_change_from_baseline  0.099531350 -0.317145228
## grocery_and_pharmacy_percent_change_from_baseline  -0.110976209 -0.296856620
## parks_percent_change_from_baseline                     0.283874181 -0.240873233
## transit_stations_percent_change_from_baseline          0.094114251 -0.074265913
## workplaces_percent_change_from_baseline               -0.125028631  0.240735868
## residential_percent_change_from_baseline              -0.178700468  0.013386663
## Total                                                 -0.118124812  0.793981499
##                                                             PC11          PC12
## num_casos.x                                           -0.049617465 -0.0988357734
## num_casos_prueba_pcr                                  -0.044726847 -0.1058640000
## num_casos_prueba_test_ac                              -0.042562355 -0.0389190655
## num_casos_prueba_ag                                    0.003758095 -0.0008550931
## num_casos_prueba_elisa                                 0.023430627 -0.0013867741
## num_casos_prueba_desconocida                          -0.154872738 -0.1615965585
## num_casos.y                                            0.033429358  0.1359821524
## num_hosp                                               0.302911164  0.0671255962
## num_uci                                               -0.400459562  0.0477114671
## num_def                                                0.553043315 -0.1950704924
## retail_and_recreation_percent_change_from_baseline  0.286708424  0.6275693247
## grocery_and_pharmacy_percent_change_from_baseline  -0.348092535 -0.0594956464
## parks_percent_change_from_baseline                    -0.082227368 -0.4958947989
## transit_stations_percent_change_from_baseline          0.264287632 -0.4433548371
## workplaces_percent_change_from_baseline               -0.216784094  0.2008296543
## residential_percent_change_from_baseline              -0.278596900  0.0313367951
## Total                                                  0.074087201 -0.0830357636
##                                                             PC13         PC14
## num_casos.x                                            0.3503737964 -0.03447444
## num_casos_prueba_pcr                                   0.3691954013 -0.03383423
## num_casos_prueba_test_ac                              -0.0023721487 -0.01431289
## num_casos_prueba_ag                                    0.0009495141 -0.01895981
```

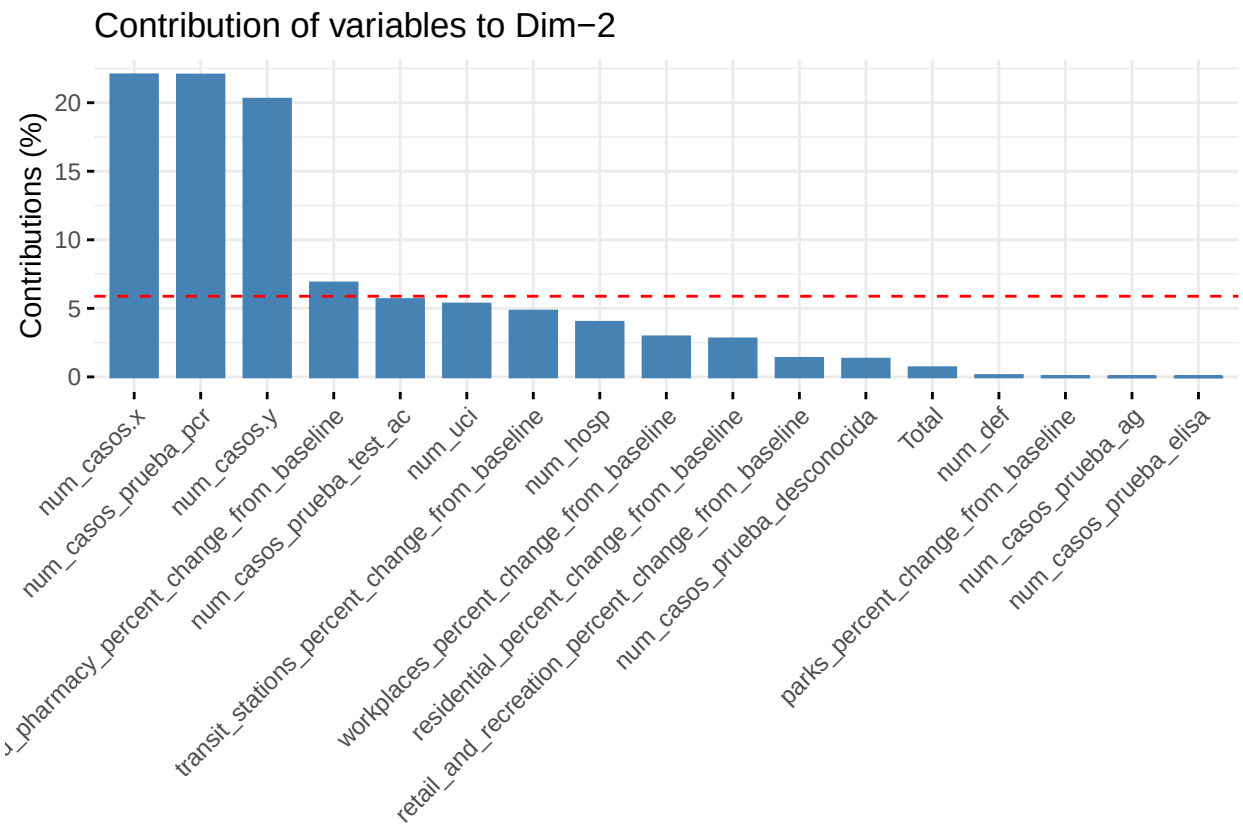
```
## num_casos_prueba_elisa                                0.0189278939 -0.03470038
## num_casos_prueba_desconocida                         -0.0341231415 -0.11380732
## num_casos.y                                          -0.8087585769  0.15136854
## num_hosp                                             -0.0093409355 -0.32920173
## num_uci                                               0.0905105752  0.12476243
## num_def                                              -0.0179131471 -0.20425404
## retail_and_recreation_percent_change_from_baseline    0.1591640063  0.05248551
## grocery_and_pharmacy_percent_change_from_baseline    -0.0689717137 -0.30356466
## parks_percent_change_from_baseline                   -0.1967768342 -0.33947866
## transit_stations_percent_change_from_baseline         0.0373896719  0.66046288
## workplaces_percent_change_from_baseline              -0.0493969622 -0.09490414
## residential_percent_change_from_baseline              0.0240600499  0.34670086
## Total                                                 0.0590314137 -0.13712447
##                                                             PC15          PC16
## num_casos.x                                           0.021886709 -0.710827710
## num_casos_prueba_pcr                                  0.020989875  0.696614910
## num_casos_prueba_test_ac                              0.002128322 -0.009045443
## num_casos_prueba_ag                                  -0.001031653 -0.006169593
## num_casos_prueba_elisa                               -0.019875868  0.004055133
## num_casos_prueba_desconocida                         -0.000227830 -0.014131490
## num_casos.y                                           0.065935929  0.012850260
## num_hosp                                             -0.738022542 -0.005532761
## num_uci                                               0.291185664 -0.005062923
## num_def                                               0.517898894 -0.032682612
## retail_and_recreation_percent_change_from_baseline    0.103782290 -0.031882076
## grocery_and_pharmacy_percent_change_from_baseline     0.033876519  0.014999252
## parks_percent_change_from_baseline                   -0.032945065 -0.039812604
## transit_stations_percent_change_from_baseline        -0.169052448 -0.008281975
## workplaces_percent_change_from_baseline              -0.059237981 -0.040905347
## residential_percent_change_from_baseline             -0.223342981 -0.056238440
## Total                                                 0.040900018 -0.009746369
##                                                             PC17
## num_casos.x                                          -0.04694046
## num_casos_prueba_pcr                                  0.08912178
## num_casos_prueba_test_ac                              0.09544863
## num_casos_prueba_ag                                   0.05828430
## num_casos_prueba_elisa                                0.04498527
## num_casos_prueba_desconocida                          0.16508413
## num_casos.y                                          -0.01242631
## num_hosp                                              0.04584112
## num_uci                                               0.06322805
## num_def                                               0.30912666
## retail_and_recreation_percent_change_from_baseline    0.33484091
## grocery_and_pharmacy_percent_change_from_baseline    -0.13871900
## parks_percent_change_from_baseline                    0.41803506
## transit_stations_percent_change_from_baseline         0.04834211
## workplaces_percent_change_from_baseline               0.44739326
## residential_percent_change_from_baseline              0.57600490
## Total                                                 0.07657033
```

```
if(!require(FactoMineR)){
    install.packages('FactoMineR', repos='http://cran.us.r-project.org')
    library(FactoMineR)}
if(!require(factoextra)){
```

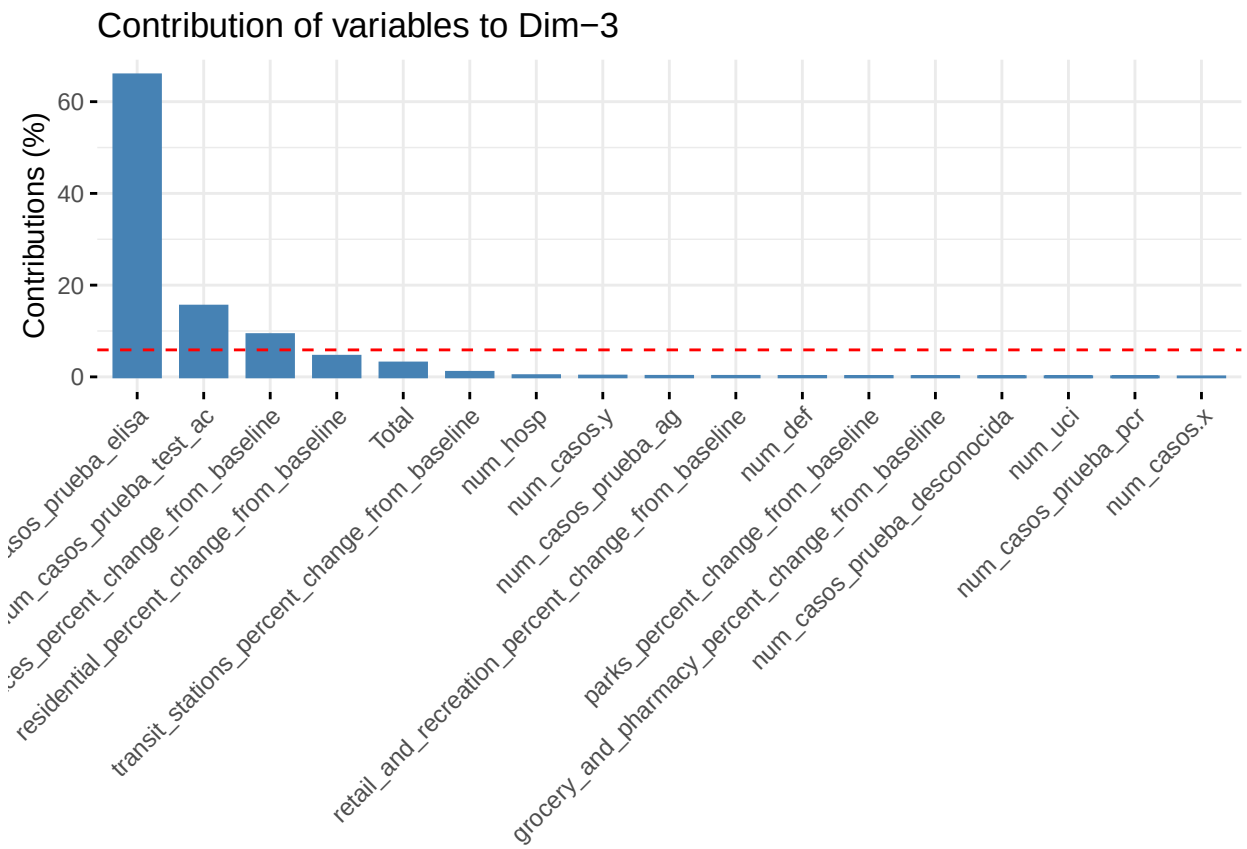
```
    install.packages('factoextra', repos='http://cran.us.r-project.org')
    library(factoextra)}

# Var contribution for PC1-PC5
fviz_contrib(pca, choice = "var", axes = 1)
```



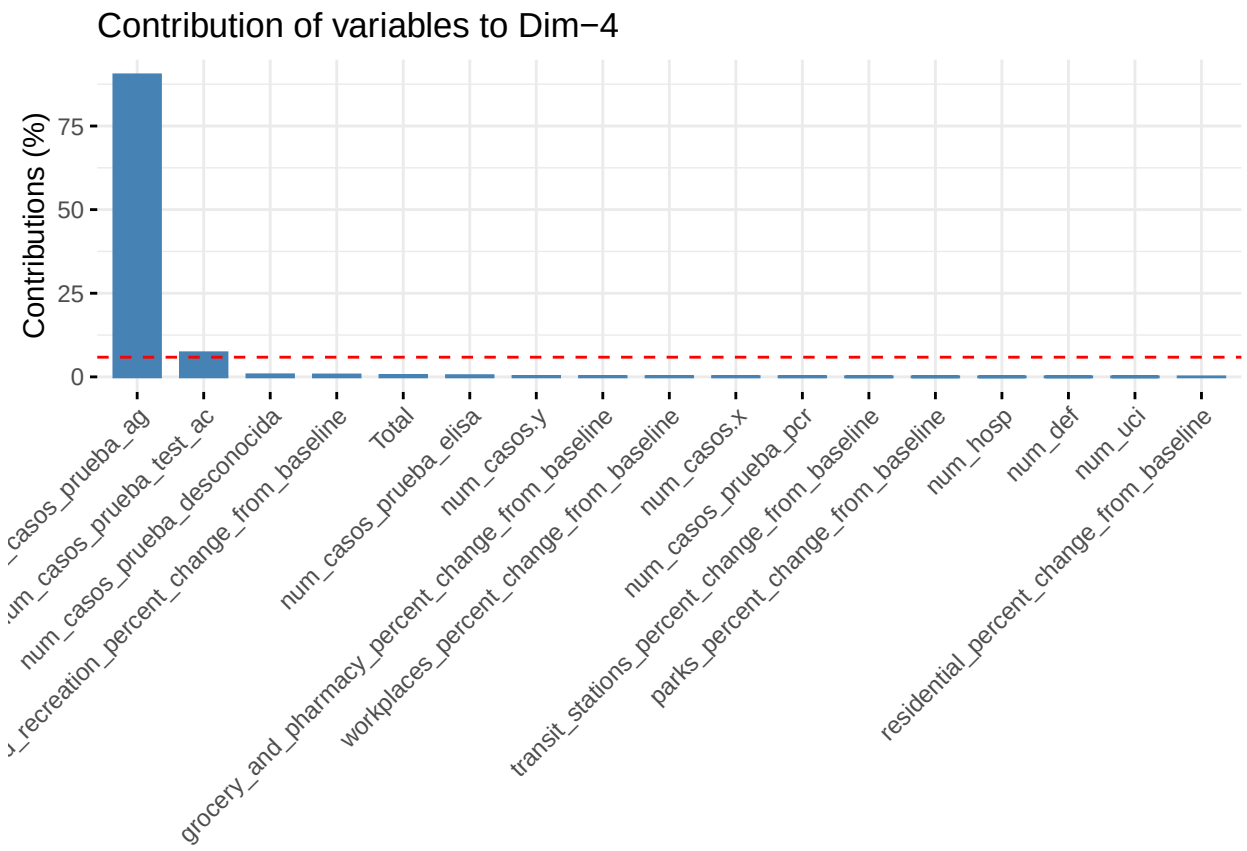

```
fviz_contrib(pca, choice = "var", axes = 2)
```



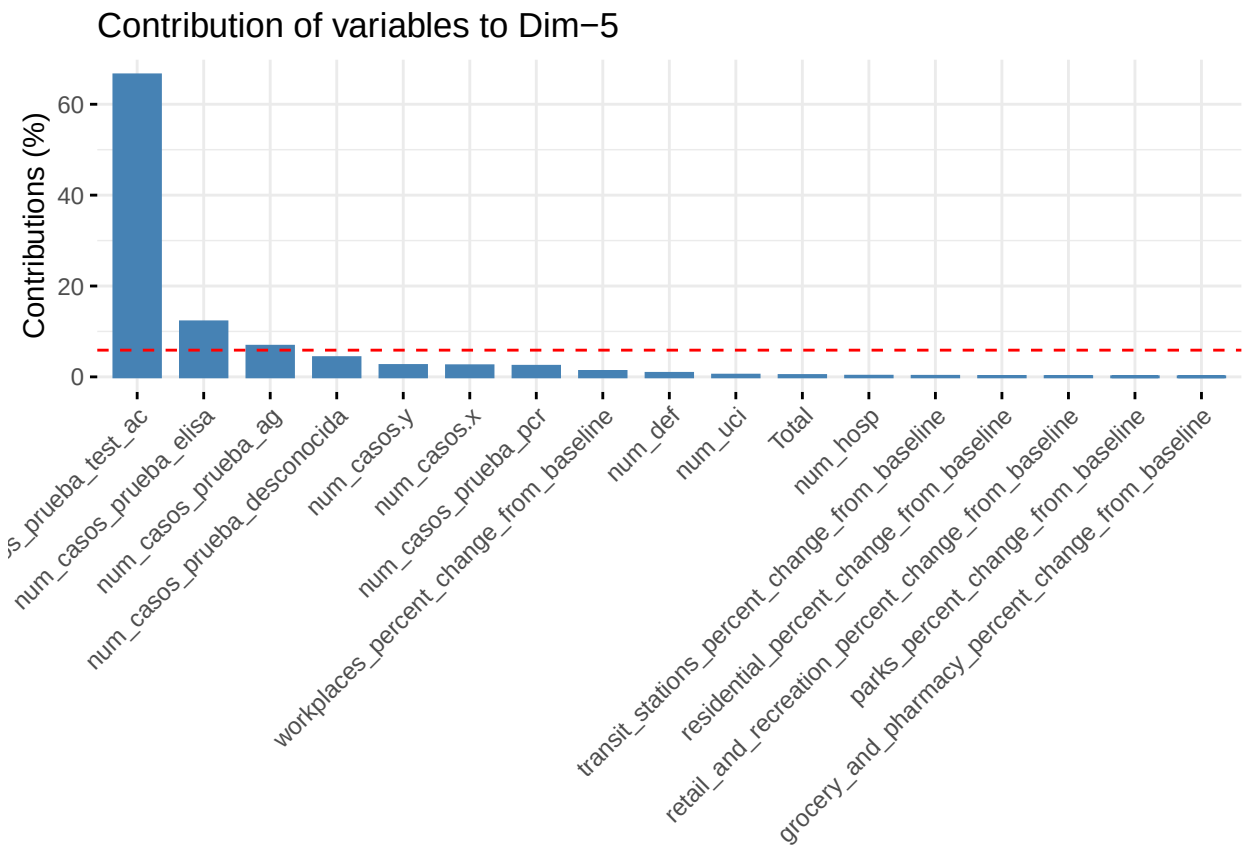

```
fviz_contrib(pca, choice = "var", axes = 3)
```



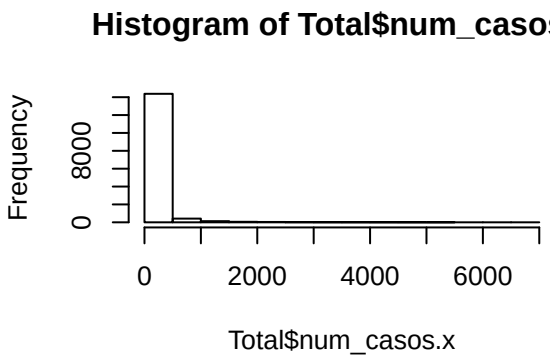

```
fviz_contrib(pca, choice = "var", axes = 4)
```

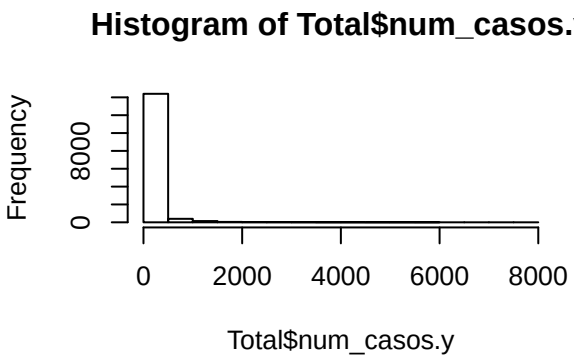


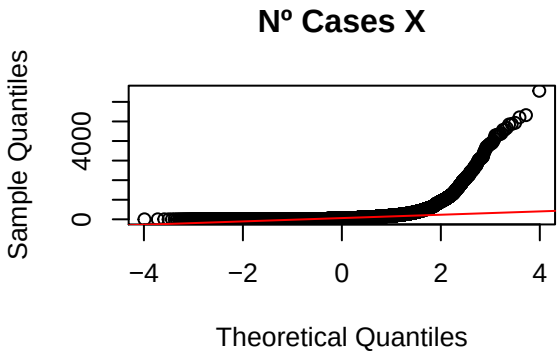

```
fviz_contrib(pca, choice = "var", axes = 5)
```

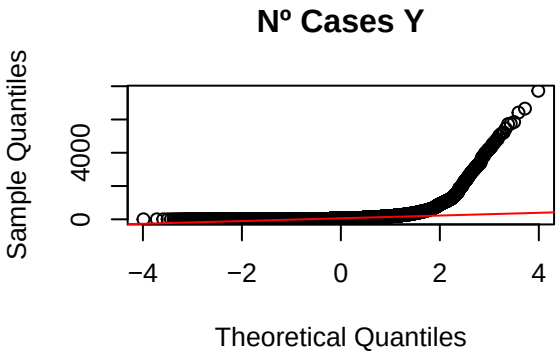



```r
par(mfrow=c(2,2))

hist(Total$num_casos.x)
hist(Total$num_casos.y)

qqnorm(Total$num_casos.x, main="Nº Cases X")
qqline(Total$num_casos.x,col=2)

qqnorm(Total$num_casos.y, main="Nº Cases Y")
qqline(Total$num_casos.y,col=2)
```

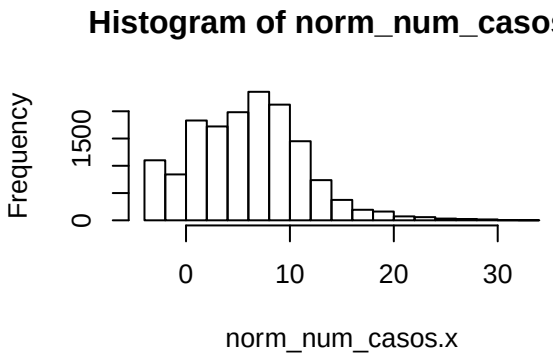


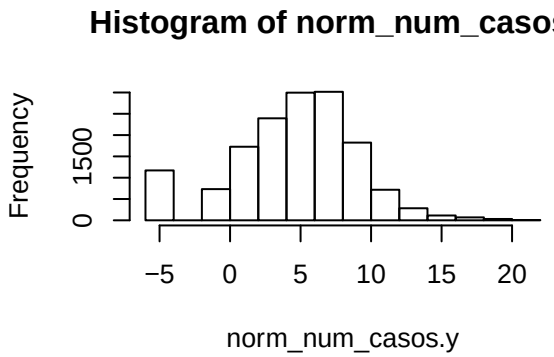


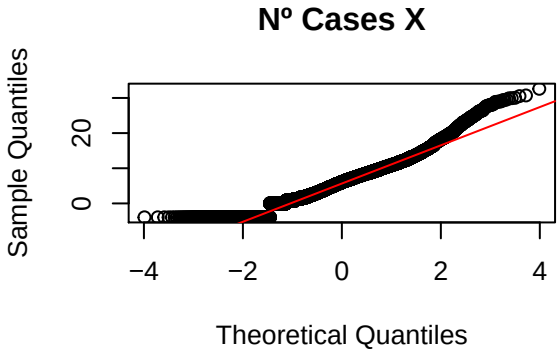


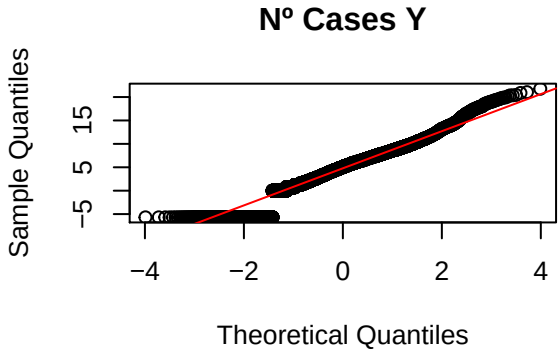



```
###############################################
norm_num_casos.x <- BoxCox(Total$num_casos.x, lambda = BoxCoxLambda(Total$num_casos.x))
norm_num_casos.y <- BoxCox(Total$num_casos.y, lambda = BoxCoxLambda(Total$num_casos.y))

hist(norm_num_casos.x)
hist(norm_num_casos.y)

qqnorm(norm_num_casos.x, main="Nº Cases X")
qqline(norm_num_casos.x,col=2)

qqnorm(norm_num_casos.y, main="Nº Cases Y")
qqline(norm_num_casos.y,col=2)
```










```
# Columns removal according PCA results and SME knowledge
Total_ts <- Total_ts[c(-4,-5,-6,-8)]
Total_ts_b <- Total_ts_b[c(-4,-5,-6,-8)]
table(Total_ts$sub_region_2)
```

```
##
##              Albacete      Alicante/Alacant               Almería
##                   290                   290                   290
##         Araba/Álava              Asturias                 Ávila
##                   290                   290                   290
##               Badajoz        Balears, Illes              Barcelona
##                   290                   290                   290
##               Bizkaia               Burgos               Cáceres
##                   290                   290                   290
##                 Cádiz             Cantabria    Castellón/Castelló
##                   290                   290                   290
##                 Ceuta           Ciudad Real               Córdoba
##                   290                   290                   290
##           Coruña, A               Cuenca              Gipuzkoa
##                   290                   290                   290
##                Girona               Granada           Guadalajara
##                   290                   290                   290
##                Huelva               Huesca                 Jaén
##                   290                   290                   290
##                 León               Lleida                 Lugo
##                   290                   290                   290
```

```
##                 Madrid                 Málaga                Melilla
##                    290                    290                    290
##                 Murcia                Navarra                Ourense
##                    290                    290                    290
##              Palencia           Palmas, Las             Pontevedra
##                    290                    290                    290
##             Rioja, La             Salamanca Santa Cruz de Tenerife
##                    290                    290                    290
##               Segovia                Sevilla                  Soria
##                    290                    290                    290
##             Tarragona                 Teruel                 Toledo
##                    290                    290                    290
##      Valencia/València             Valladolid                 Zamora
##                    290                    290                    290
##               Zaragoza
##                    290
```

```
str(Total_ts)
```

```
## tsibble [15,080 x 15] (S3: tbl_ts/tbl_df/tbl/data.frame)
##  $ sub_region_2                                      : chr [1:15080] "Albacete" "Albacete" "Albacete
##  $ num_casos.x                                       : int [1:15080] 137 128 114 149 131 129 125 112
##  $ num_casos_prueba_pcr                              : int [1:15080] 132 123 107 133 121 120 112 103
##  $ num_casos_prueba_desconocida                      : int [1:15080] 0 0 0 0 0 0 0 0 0 0 ...
##  $ num_hosp                                          : int [1:15080] 43 40 24 40 63 60 61 108 76 66
##  $ num_uci                                           : int [1:15080] 3 4 7 5 4 2 8 2 7 9 ...
##  $ num_def                                           : int [1:15080] 7 2 7 7 6 10 13 16 14 16 ...
##  $ retail_and_recreation_percent_change_from_baseline: num [1:15080] -81 -84 -83 -93 -87 -92 -94 -85
##  $ grocery_and_pharmacy_percent_change_from_baseline : num [1:15080] -32 -41 -32 -92 -34 -60 -80 -46
##  $ parks_percent_change_from_baseline                : num [1:15080] -73 -74 -70 -80 -74 -85 -88 -77
##  $ transit_stations_percent_change_from_baseline     : num [1:15080] -66 -72 -70 -86 -76 -80 -89 -76
##  $ workplaces_percent_change_from_baseline           : num [1:15080] -51 -56 -58 -85 -68 -64 -66 -61
##  $ residential_percent_change_from_baseline          : num [1:15080] 22 23 23 35 32 27 23 26 28 28 .
##  $ Total                                             : num [1:15080] 9.9 9.71 9.51 9.13 8.75 ...
##  $ Dia_c                                             : Date[1:15080], format: "2020-03-16" "2020-03-
##  - attr(*, "key")= tibble [52 x 2] (S3: tbl_df/tbl/data.frame)
##   ..$ sub_region_2: chr [1:52] "Albacete" "Alicante/Alacant" "Almería" "Araba/Álava" ...
##   ..$ .rows       : list<int> [1:52]
##   .. ..$ : int [1:290] 1 2 3 4 5 6 7 8 9 10 ...
##   .. ..$ : int [1:290] 291 292 293 294 295 296 297 298 299 300 ...
##   .. ..$ : int [1:290] 581 582 583 584 585 586 587 588 589 590 ...
##   .. ..$ : int [1:290] 871 872 873 874 875 876 877 878 879 880 ...
##   .. ..$ : int [1:290] 1161 1162 1163 1164 1165 1166 1167 1168 1169 1170 ...
##   .. ..$ : int [1:290] 1451 1452 1453 1454 1455 1456 1457 1458 1459 1460 ...
##   .. ..$ : int [1:290] 1741 1742 1743 1744 1745 1746 1747 1748 1749 1750 ...
##   .. ..$ : int [1:290] 2031 2032 2033 2034 2035 2036 2037 2038 2039 2040 ...
##   .. ..$ : int [1:290] 2321 2322 2323 2324 2325 2326 2327 2328 2329 2330 ...
##   .. ..$ : int [1:290] 2611 2612 2613 2614 2615 2616 2617 2618 2619 2620 ...
##   .. ..$ : int [1:290] 2901 2902 2903 2904 2905 2906 2907 2908 2909 2910 ...
##   .. ..$ : int [1:290] 3191 3192 3193 3194 3195 3196 3197 3198 3199 3200 ...
##   .. ..$ : int [1:290] 3481 3482 3483 3484 3485 3486 3487 3488 3489 3490 ...
##   .. ..$ : int [1:290] 3771 3772 3773 3774 3775 3776 3777 3778 3779 3780 ...
##   .. ..$ : int [1:290] 4061 4062 4063 4064 4065 4066 4067 4068 4069 4070 ...
##   .. ..$ : int [1:290] 4351 4352 4353 4354 4355 4356 4357 4358 4359 4360 ...
##   .. ..$ : int [1:290] 4641 4642 4643 4644 4645 4646 4647 4648 4649 4650 ...
```

```
##   .. ..$ : int [1:290] 4931 4932 4933 4934 4935 4936 4937 4938 4939 4940 ...
##   .. ..$ : int [1:290] 5221 5222 5223 5224 5225 5226 5227 5228 5229 5230 ...
##   .. ..$ : int [1:290] 5511 5512 5513 5514 5515 5516 5517 5518 5519 5520 ...
##   .. ..$ : int [1:290] 5801 5802 5803 5804 5805 5806 5807 5808 5809 5810 ...
##   .. ..$ : int [1:290] 6091 6092 6093 6094 6095 6096 6097 6098 6099 6100 ...
##   .. ..$ : int [1:290] 6381 6382 6383 6384 6385 6386 6387 6388 6389 6390 ...
##   .. ..$ : int [1:290] 6671 6672 6673 6674 6675 6676 6677 6678 6679 6680 ...
##   .. ..$ : int [1:290] 6961 6962 6963 6964 6965 6966 6967 6968 6969 6970 ...
##   .. ..$ : int [1:290] 7251 7252 7253 7254 7255 7256 7257 7258 7259 7260 ...
##   .. ..$ : int [1:290] 7541 7542 7543 7544 7545 7546 7547 7548 7549 7550 ...
##   .. ..$ : int [1:290] 7831 7832 7833 7834 7835 7836 7837 7838 7839 7840 ...
##   .. ..$ : int [1:290] 8121 8122 8123 8124 8125 8126 8127 8128 8129 8130 ...
##   .. ..$ : int [1:290] 8411 8412 8413 8414 8415 8416 8417 8418 8419 8420 ...
##   .. ..$ : int [1:290] 8701 8702 8703 8704 8705 8706 8707 8708 8709 8710 ...
##   .. ..$ : int [1:290] 8991 8992 8993 8994 8995 8996 8997 8998 8999 9000 ...
##   .. ..$ : int [1:290] 9281 9282 9283 9284 9285 9286 9287 9288 9289 9290 ...
##   .. ..$ : int [1:290] 9571 9572 9573 9574 9575 9576 9577 9578 9579 9580 ...
##   .. ..$ : int [1:290] 9861 9862 9863 9864 9865 9866 9867 9868 9869 9870 ...
##   .. ..$ : int [1:290] 10151 10152 10153 10154 10155 10156 10157 10158 10159 10160 ...
##   .. ..$ : int [1:290] 10441 10442 10443 10444 10445 10446 10447 10448 10449 10450 ...
##   .. ..$ : int [1:290] 10731 10732 10733 10734 10735 10736 10737 10738 10739 10740 ...
##   .. ..$ : int [1:290] 11021 11022 11023 11024 11025 11026 11027 11028 11029 11030 ...
##   .. ..$ : int [1:290] 11311 11312 11313 11314 11315 11316 11317 11318 11319 11320 ...
##   .. ..$ : int [1:290] 11601 11602 11603 11604 11605 11606 11607 11608 11609 11610 ...
##   .. ..$ : int [1:290] 11891 11892 11893 11894 11895 11896 11897 11898 11899 11900 ...
##   .. ..$ : int [1:290] 12181 12182 12183 12184 12185 12186 12187 12188 12189 12190 ...
##   .. ..$ : int [1:290] 12471 12472 12473 12474 12475 12476 12477 12478 12479 12480 ...
##   .. ..$ : int [1:290] 12761 12762 12763 12764 12765 12766 12767 12768 12769 12770 ...
##   .. ..$ : int [1:290] 13051 13052 13053 13054 13055 13056 13057 13058 13059 13060 ...
##   .. ..$ : int [1:290] 13341 13342 13343 13344 13345 13346 13347 13348 13349 13350 ...
##   .. ..$ : int [1:290] 13631 13632 13633 13634 13635 13636 13637 13638 13639 13640 ...
##   .. ..$ : int [1:290] 13921 13922 13923 13924 13925 13926 13927 13928 13929 13930 ...
##   .. ..$ : int [1:290] 14211 14212 14213 14214 14215 14216 14217 14218 14219 14220 ...
##   .. ..$ : int [1:290] 14501 14502 14503 14504 14505 14506 14507 14508 14509 14510 ...
##   .. ..$ : int [1:290] 14791 14792 14793 14794 14795 14796 14797 14798 14799 14800 ...
##   .. ..@ ptype: int(0)
##   ..- attr(*, ".drop")= logi TRUE
##  - attr(*, "index")= chr "Dia_c"
##   ..- attr(*, "ordered")= logi TRUE
##  - attr(*, "index2")= chr "Dia_c"
##  - attr(*, "interval")= interval [1:1] 1D
##   ..@ .regular: logi TRUE
```

```
summary(Total_ts)
```

```
##  sub_region_2        num_casos.x   num_casos_prueba_pcr
##  Length:15080       Min.   :   0   Min.   :   0.0
##  Class :character   1st Qu.:   5   1st Qu.:   5.0
##  Mode  :character   Median :  39   Median :  35.0
##                     Mean   : 126   Mean   : 110.2
##                     3rd Qu.: 120   3rd Qu.: 105.0
##                     Max.   :6565   Max.   :6546.0
##  num_casos_prueba_desconocida    num_hosp          num_uci
##  Min.   : 0.0000              Min.   :   0.00   Min.   :  0.000
##  1st Qu.: 0.0000              1st Qu.:   1.00   1st Qu.:  0.000
```

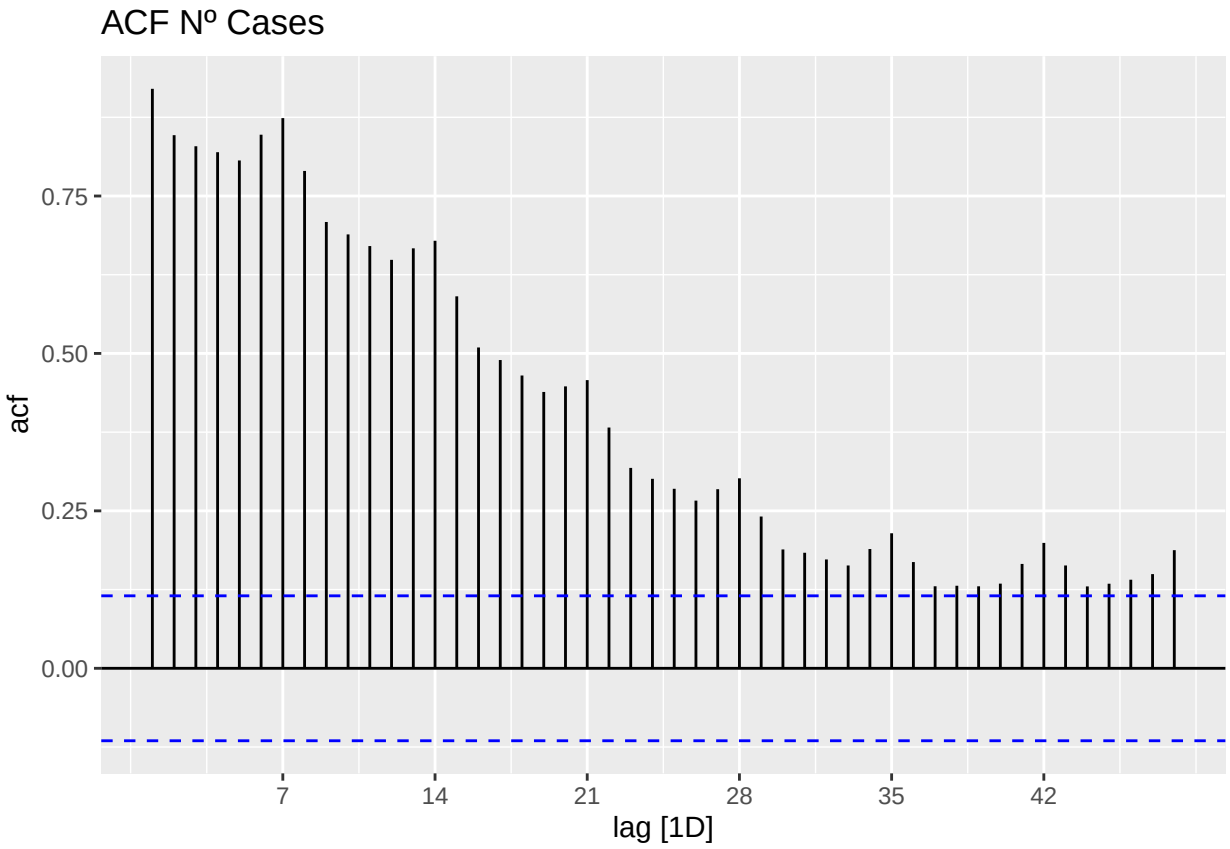
```
##  Median : 0.0000              Median :   4.00   Median :  0.000
##  Mean   : 0.1317              Mean   :  14.86   Mean   :  1.281
##  3rd Qu.: 0.0000              3rd Qu.:  12.00   3rd Qu.:  1.000
##  Max.   :65.0000              Max.   :1930.00   Max.   :135.000
##     num_def        retail_and_recreation_percent_change_from_baseline
##  Min.   :  0.000   Min.   :-97.00
##  1st Qu.:  0.000   1st Qu.:-57.00
##  Median :  1.000   Median :-30.00
##  Mean   :  3.437   Mean   :-37.29
##  3rd Qu.:  3.000   3rd Qu.:-17.00
##  Max.   :334.000   Max.   : 71.00
##  grocery_and_pharmacy_percent_change_from_baseline
##  Min.   :-96.00
##  1st Qu.:-24.00
##  Median : -6.00
##  Mean   :-11.75
##  3rd Qu.:  4.00
##  Max.   :194.00
##  parks_percent_change_from_baseline
##  Min.   :-94.000
##  1st Qu.:-30.000
##  Median : -2.000
##  Mean   :  5.809
##  3rd Qu.: 30.000
##  Max.   :543.000
##  transit_stations_percent_change_from_baseline
##  Min.   :-100.00
##  1st Qu.: -53.00
##  Median : -31.00
##  Mean   : -35.19
##  3rd Qu.: -17.00
##  Max.   :  74.00
##  workplaces_percent_change_from_baseline
##  Min.   :-92.00
##  1st Qu.:-43.00
##  Median :-26.00
##  Mean   :-29.08
##  3rd Qu.:-13.00
##  Max.   : 55.00
##  residential_percent_change_from_baseline     Total           Dia_c
##  Min.   :-10.00                          Min.   : 1.95   Min.   :2020-03-16
##  1st Qu.:  4.00                          1st Qu.:11.36   1st Qu.:2020-05-27
##  Median :  7.00                          Median :14.39   Median :2020-08-07
##  Mean   : 10.14                          Mean   :14.20   Mean   :2020-08-07
##  3rd Qu.: 14.00                          3rd Qu.:17.11   3rd Qu.:2020-10-19
##  Max.   : 48.00                          Max.   :29.00   Max.   :2020-12-30
```

3 Seasonal and trend decomposition

3.0.1 ACF

```
# Filter time-series for Bar, Mad, Mal, Cor and, Cad
Total_ts_b %>%
```

```
  filter(sub_region_2 == "Barcelona") %>%
  ACF(num_casos.x, lag_max = 48) %>%
  autoplot() +
  labs(title="ACF Nº Cases")
```

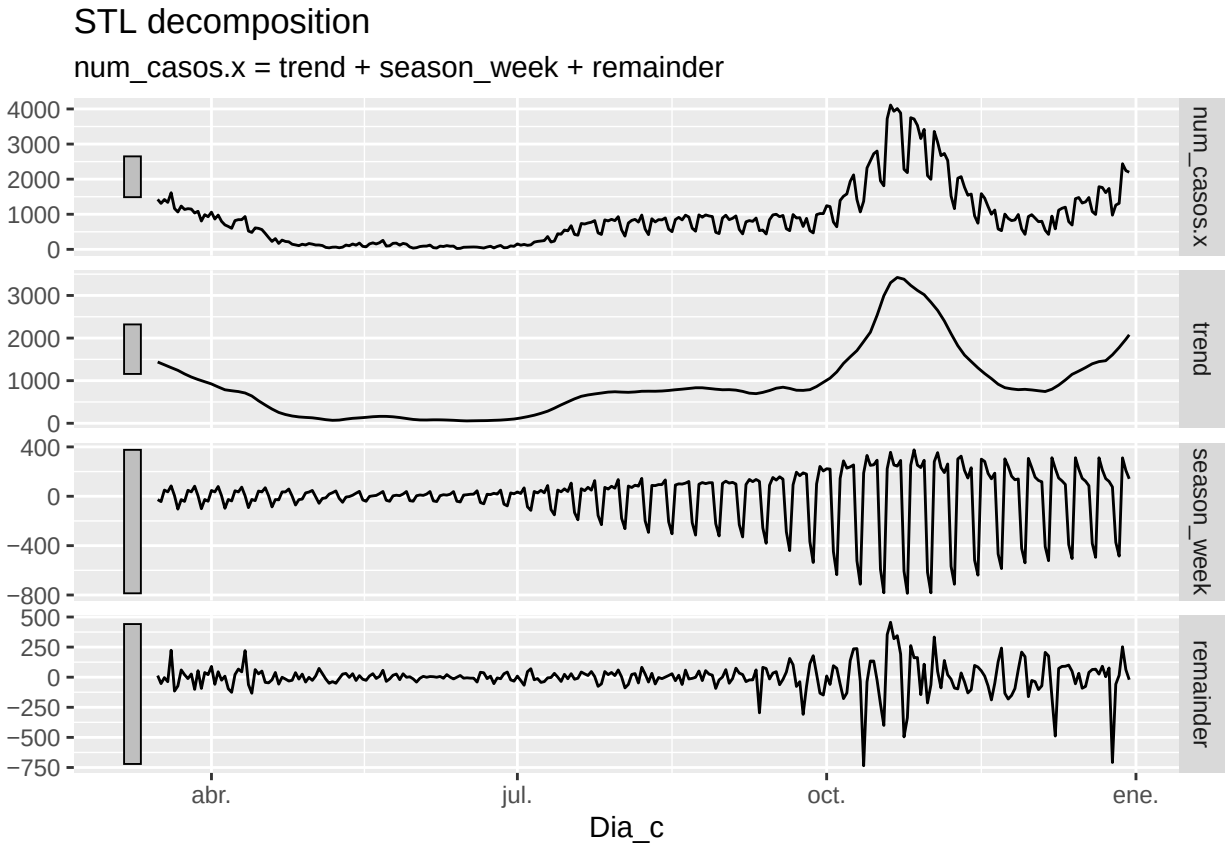


3.1 STL (Seasonal and Trend decomposition using Loess)

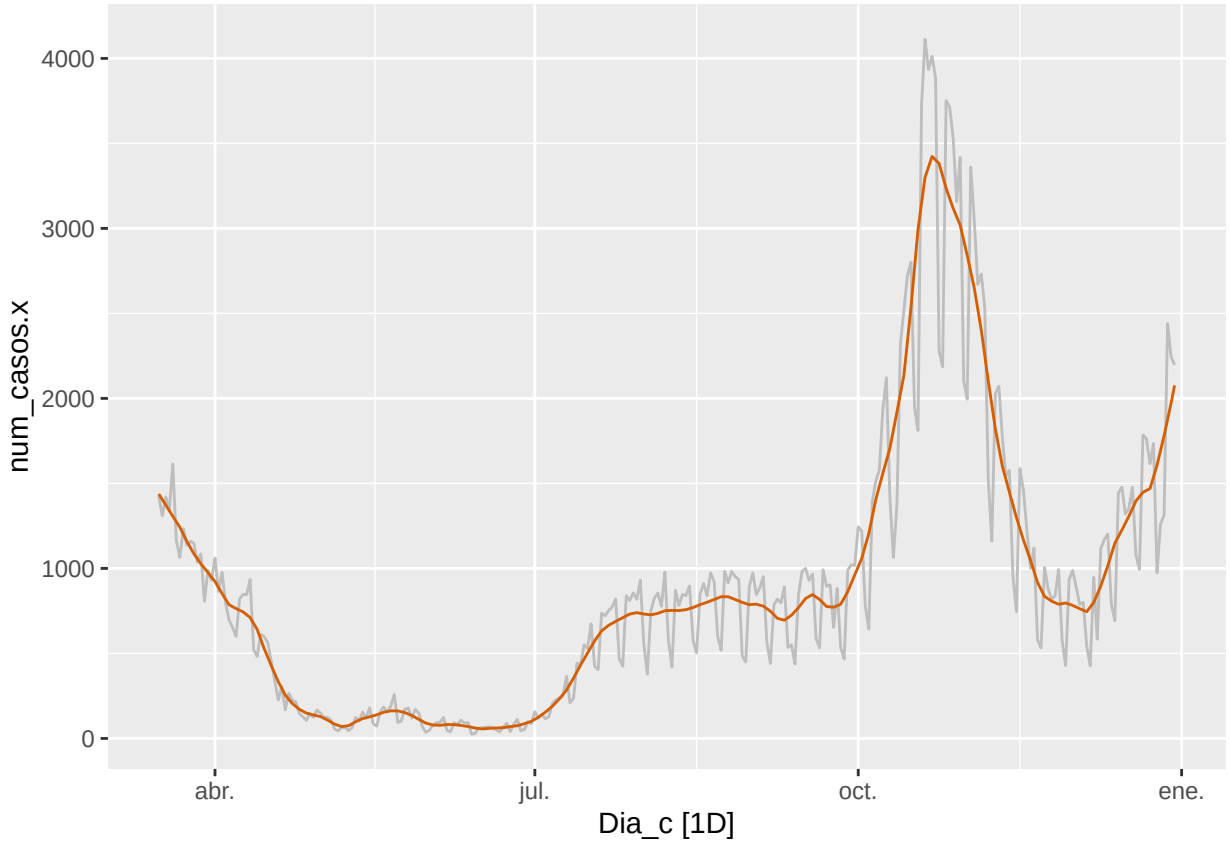
As stated by (Hyndman and Athanasopoulos 2021)... "STL has several advantages over classical decomposition, and the SEATS and X-11 methods:

- Unlike SEATS and X-11, STL will handle any type of seasonality, not only monthly and quarterly data.
- The seasonal component is allowed to change over time, and the rate of change can be controlled by the user.
- The smoothness of the trend-cycle can also be controlled by the user.
- It can be robust to outliers (i.e., the user can specify a robust decomposition), so that occasional unusual observations will not affect the estimates of the trend-cycle and seasonal components. They will, however, affect the remainder component"...

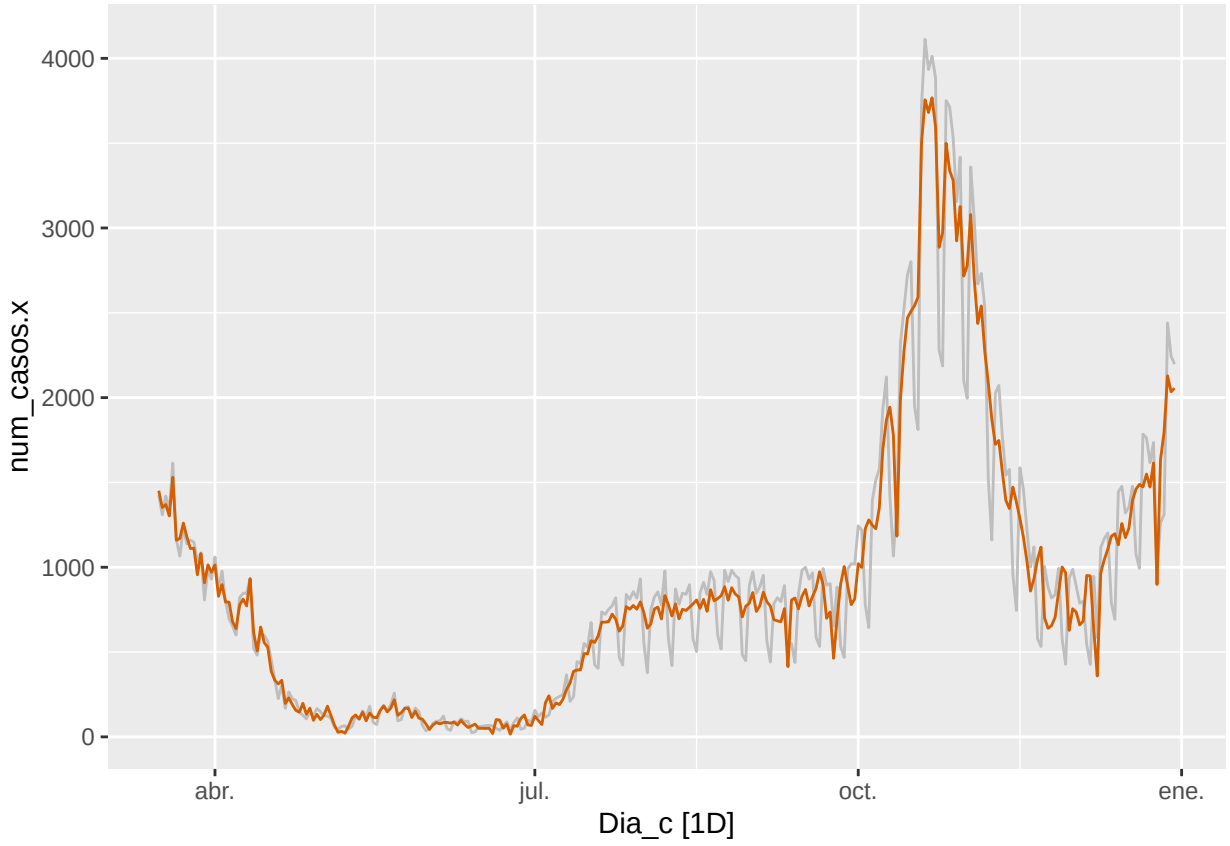
```
dcmp <- Total_ts %>%
  filter(sub_region_2 == "Barcelona") %>%
  model(STL(num_casos.x))

components(dcmp) %>% autoplot()
```



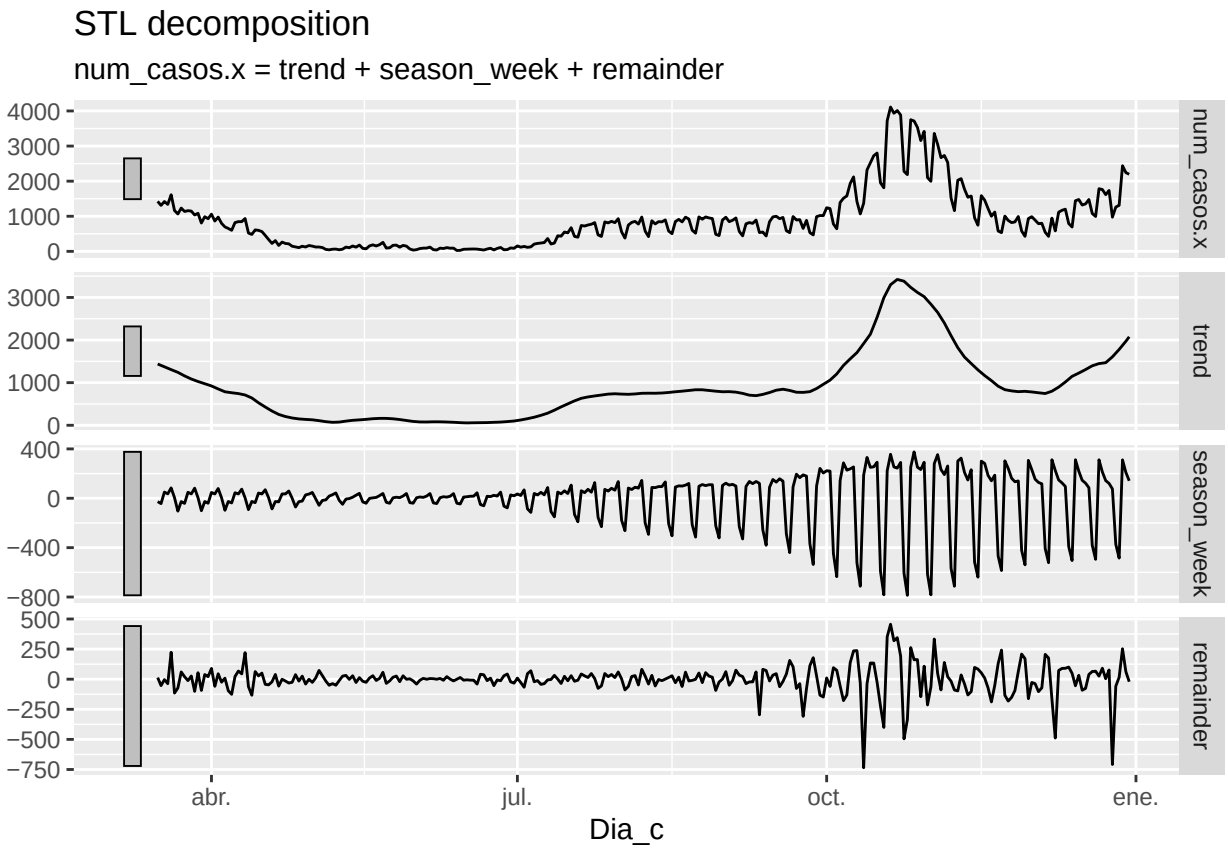

```
components(dcmp) %>%
  as_tsibble() %>%
  autoplot(num_casos.x, color="gray") +
  geom_line(aes(y=trend), color = "#D55E00")
```



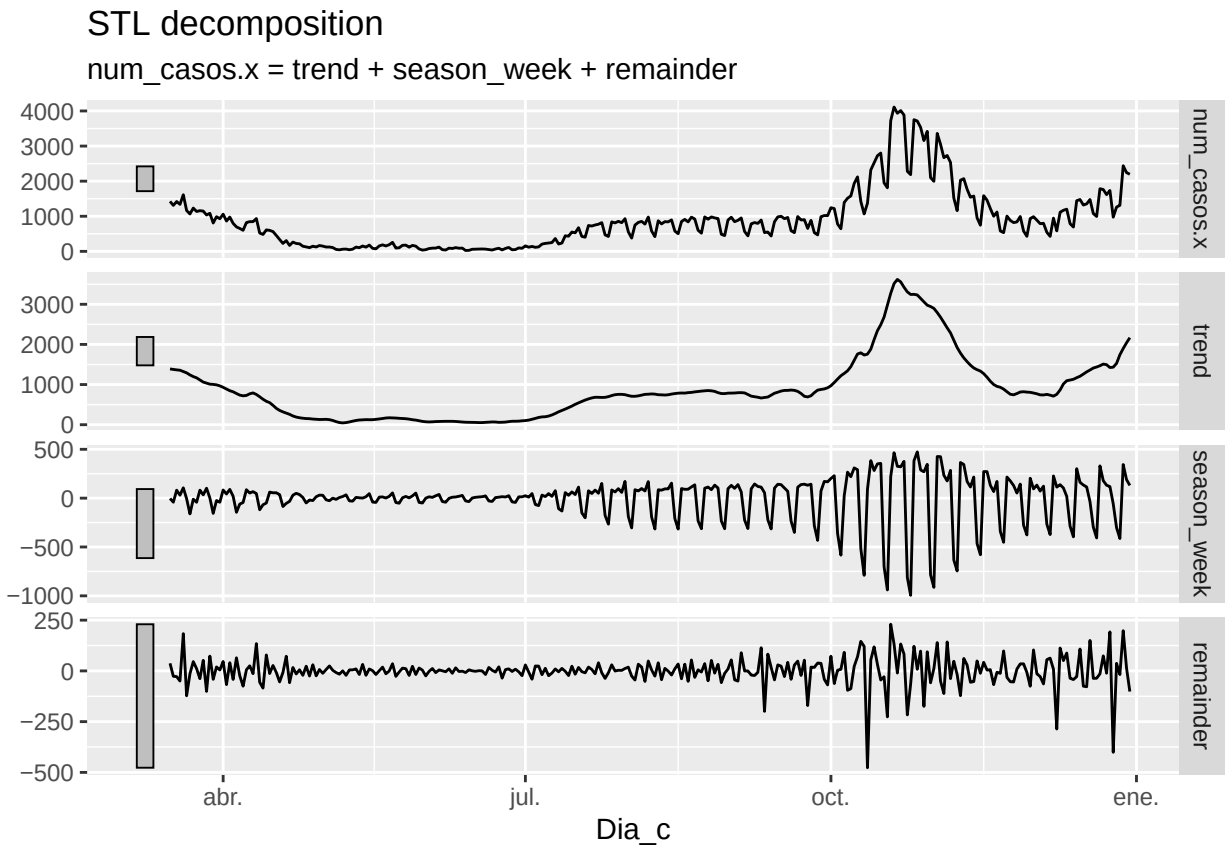

```
components(dcmp) %>%
  as_tsibble() %>%
  autoplot(num_casos.x, color="gray") +
  geom_line(aes(y=season_adjust), color = "#D55E00")
```




```
############################
Total_ts %>%
  filter(sub_region_2 == "Barcelona") %>%
  model(STL(num_casos.x)) %>%
  components() %>%
  autoplot()
```




```
Total_ts %>%
  filter(sub_region_2 == "Barcelona") %>%
  model(STL(num_casos.x ~ season(window = 7) +
              trend(window = 7))) %>%
  components() %>%
  autoplot()
```



3.2 Till here 06-Apr-2021

Bibliography

Baayen, R Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.

Hothorn, Torsten, and Brian S Everitt. 2014. *A Handbook of Statistical Analyses Using R*. CRC press.

Hyndman, Rob J., and George Athanasopoulos. 2021. "Forecasting: Principles and Practice, 3rd." OTexts: Melbourne, Australia. OTexts.com.

Liviano Solas, Daniel, and Maria Pujol Jover. n.d. *Analisis de Datos Y Estadastica Descriptiva Con R Y R-Commander*. UOC.

Teetor, Paul. 2011. *R Cookbook: Proven Recipes for Data Analysis, Statistics, and Graphics*. O'Reilly Media, Inc.

Vegas Lozano, Esteban. n.d. *Preprocesamiento de Los Datos*. UOC.