

PEC 3: Desing & Implementation - Data load, clean-up, transformation & model selection

UOC - Alumno: Álvaro Rodríguez Sans

May 2020 - Delivery 23/05/2021

Índex

1	Data load	4
2	Inintial descriptive statistics and visualization	4
2.1	Data types and modifications	4
2.1.1	EM3 review	4
2.1.2	Transpose EM3 & review missing values	6
2.1.3	Google review	18
2.1.4	CNE review	63
2.2	Datasets combinations	74
2.2.1	CNE_tec_cas	74
2.2.2	GOG_CNE	75
2.2.3	Total	77
2.3	Dataframe plots	85
2.4	Time-series plots	88
2.5	STL - Seasonal and Trend decomposition using Loess	91
2.6	Till here 06-Apr-2021	96
	Bibliography	96

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

## Loading required package: lubridate
##
## Attaching package: 'lubridate'
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
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 ----- tidyverse 1.3.0 --
## 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 ----- tidyverse_conflicts() --
## x ggplot2::%+%()      masks psych::%+%()
## x ggplot2::alpha()    masks psych::alpha()
## x lubridate::as.difftime() masks base::as.difftime()
## x lubridate::date()   masks base::date()
## x dplyr::filter()     masks stats::filter()
## x lubridate::intersect() masks base::intersect()
## x dplyr::lag()        masks stats::lag()

```

```

## x lubridate::setdiff()      masks base::setdiff()
## x lubridate::union()       masks base::union()

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(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 ----- fpp3 0.4.0
## v tsibble      1.0.0      v feasts      0.2.1
## v tsibbledata  0.3.0

## -- Conflicts ----- fpp3_conflicts
## x ggplot2::%+%()      masks psych::%+%()
## x ggplot2::alpha()    masks psych::alpha()
## x lubridate::date()   masks base::date()
## x dplyr::filter()     masks stats::filter()
## x tsibble::intersect() masks base::intersect()
## x tsibble::interval() masks lubridate::interval()
## x dplyr::lag()         masks stats::lag()
## x fabletools::MAE()    masks DescTools::MAE()
## x fabletools::MAPE()   masks DescTools::MAPE()
## x fabletools::MSE()    masks DescTools::MSE()
## x fabletools::RMSE()   masks DescTools::RMSE()

```

```
## x tsibble::setdiff()   masks base::setdiff()
## x tsibble::union()    masks base::union()

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 Inintial descriptive statistics and visualization

2.1 Data types and modifications

We are going 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
```

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
```

2.1.2 Transpose EM3 & review missing values

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

```
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: According 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
```

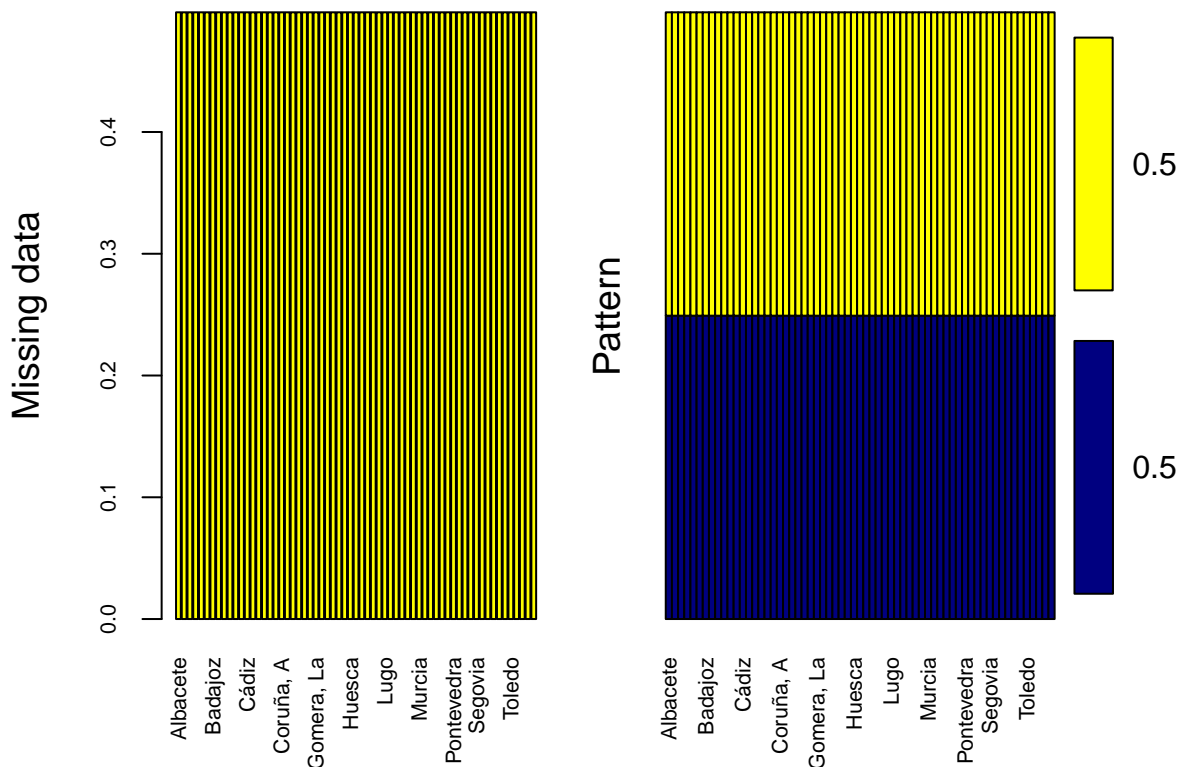
```
## # A tibble: 291 x 64
##   Periodo   Albacete `Alicante/Alaca~ Almería `Araba/Álava` Asturias Ávila
##   <date>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl> <dbl>
## 1 2019-11-18  25.2        28.1      24.4        31.9      29.9  26.6
## 2 2020-03-16   9.9        14.4      11.0        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>
```

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

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

  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

```



```
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 principles 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-).

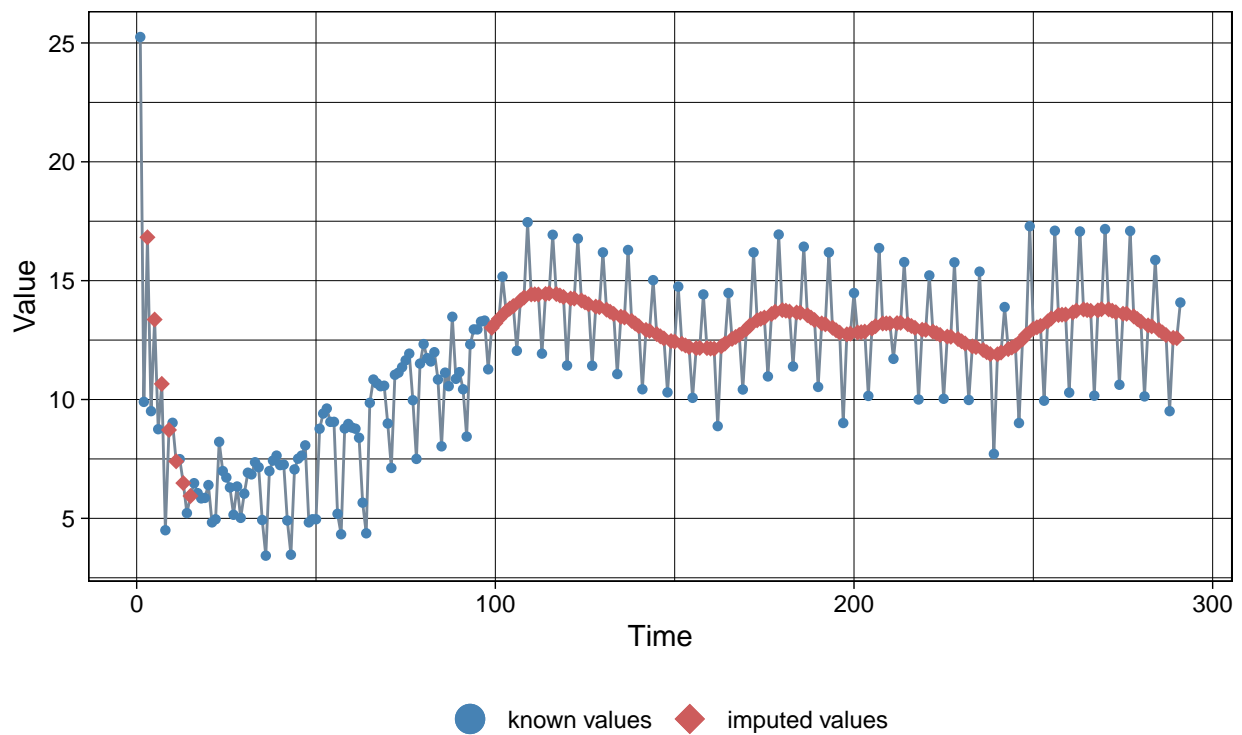
It is needed to transform the dataframe to a time series object.

```
# 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)
```

Imputed Values

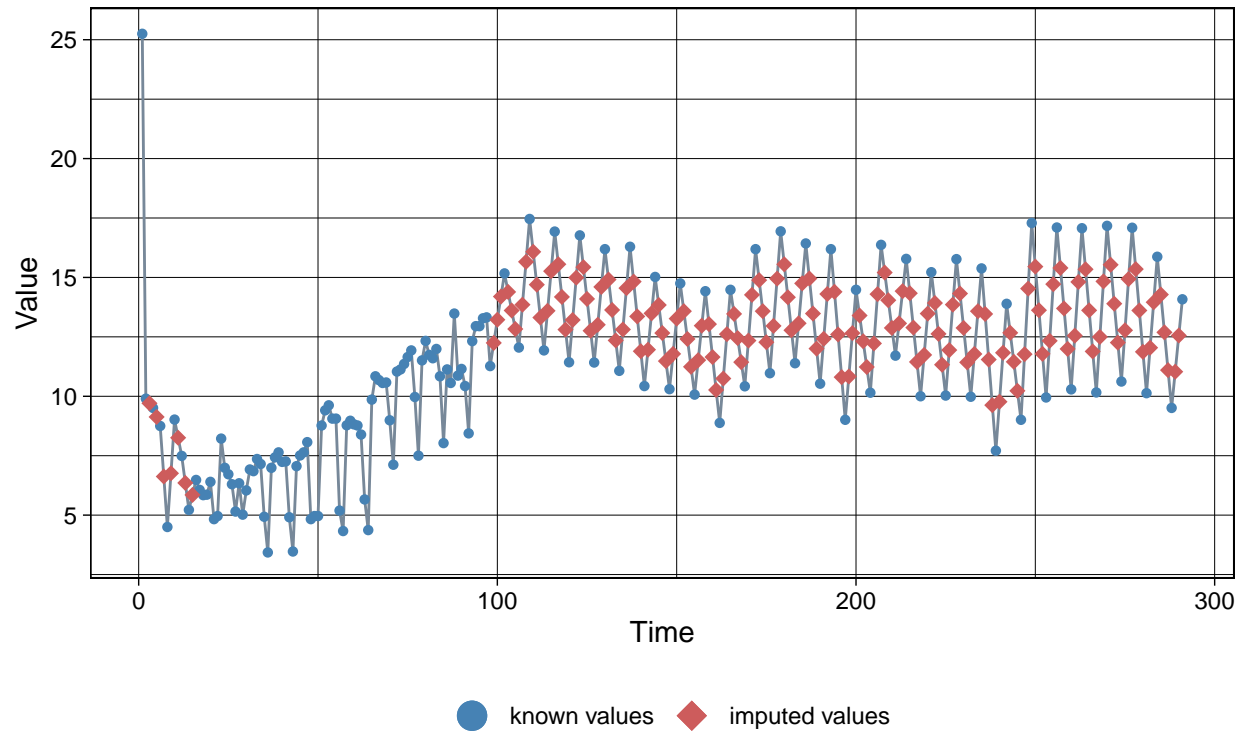
Visualization of missing value replacements



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

Imputed Values

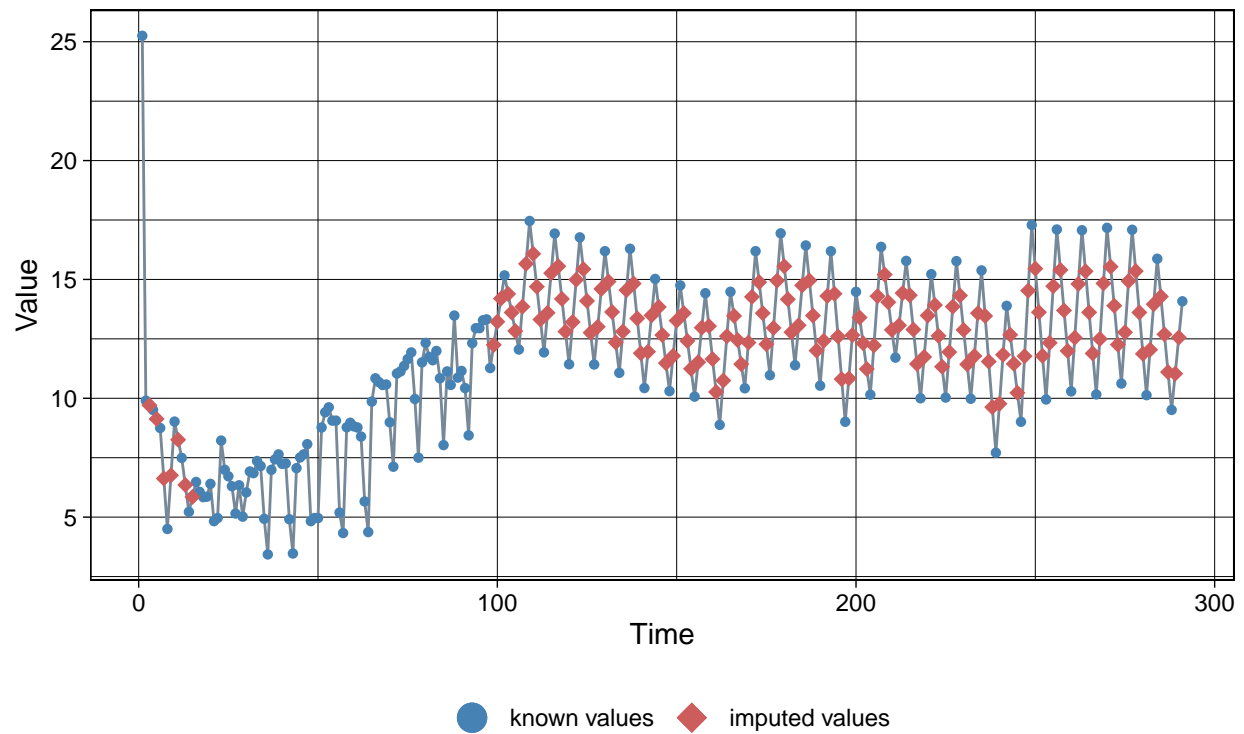
Visualization of missing value replacements



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

Imputed Values

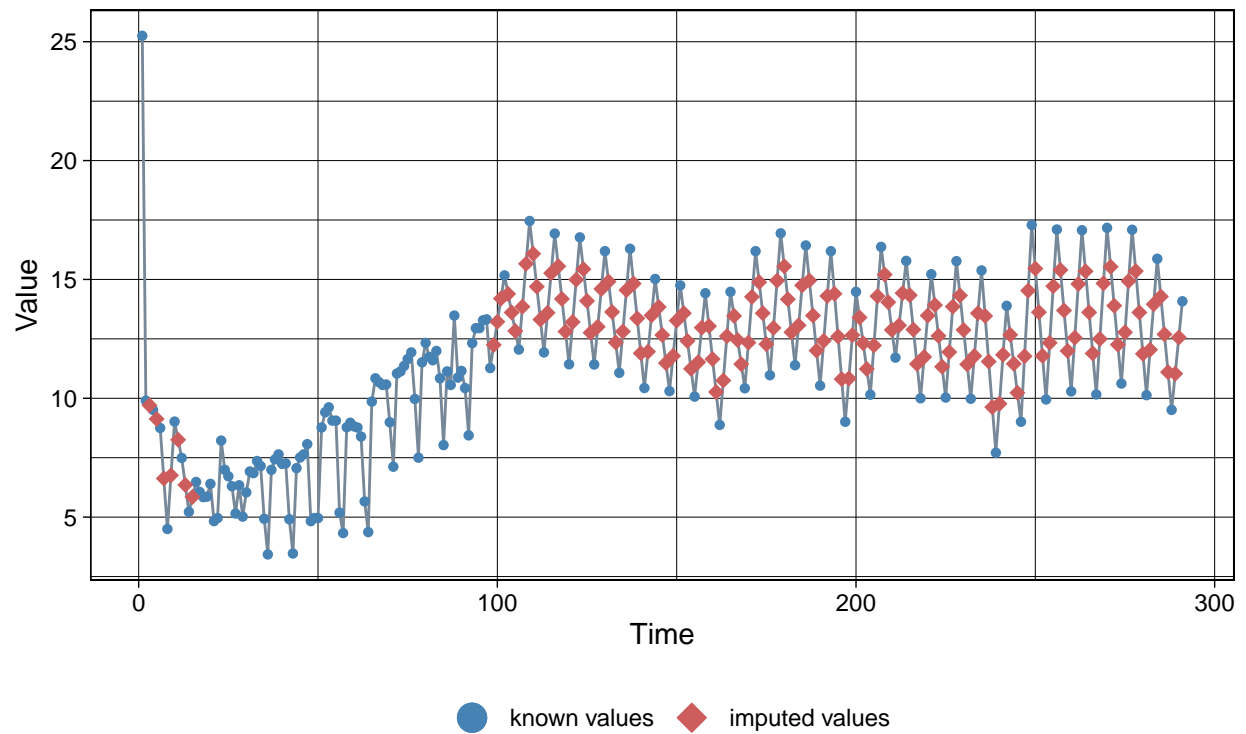
Visualization of missing value replacements



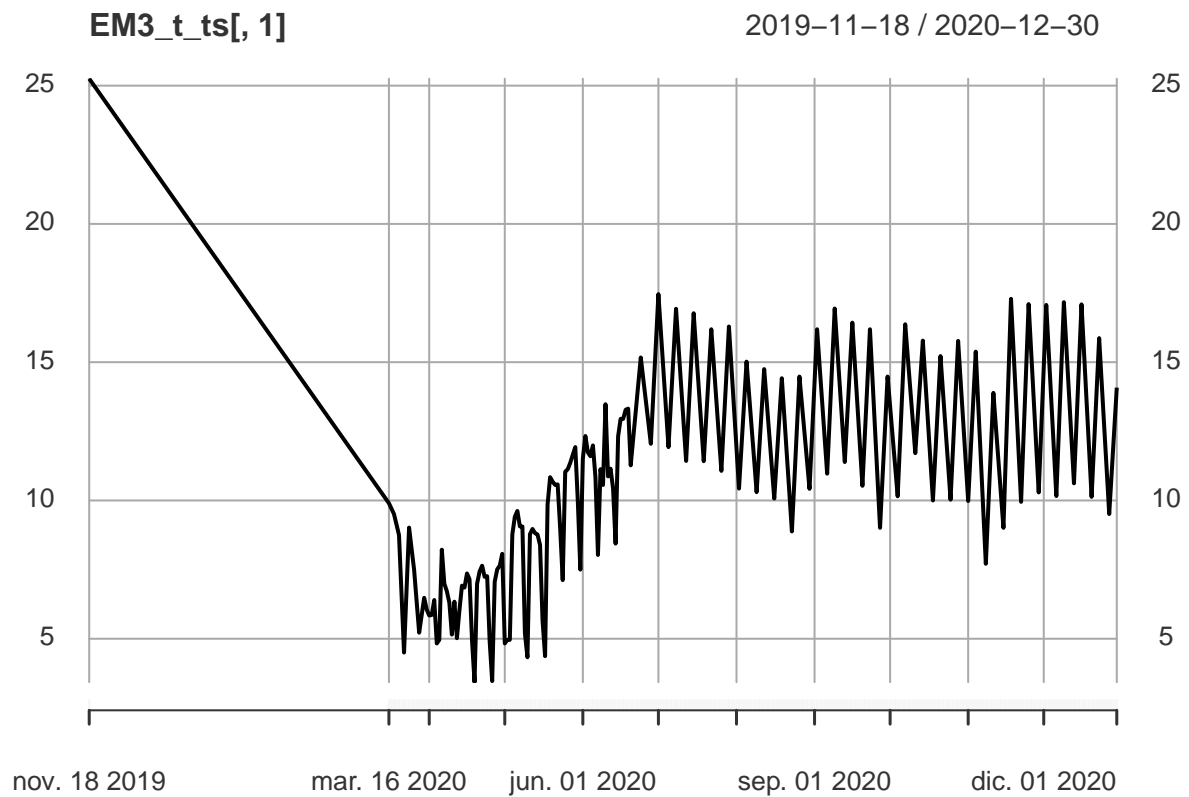
```
imp4 <- na_interpolation(EM3_t_ts[,1])  
ggplot_na_imputations(EM3_t_ts[,1], imp4)
```

Imputed Values

Visualization of missing value replacements



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



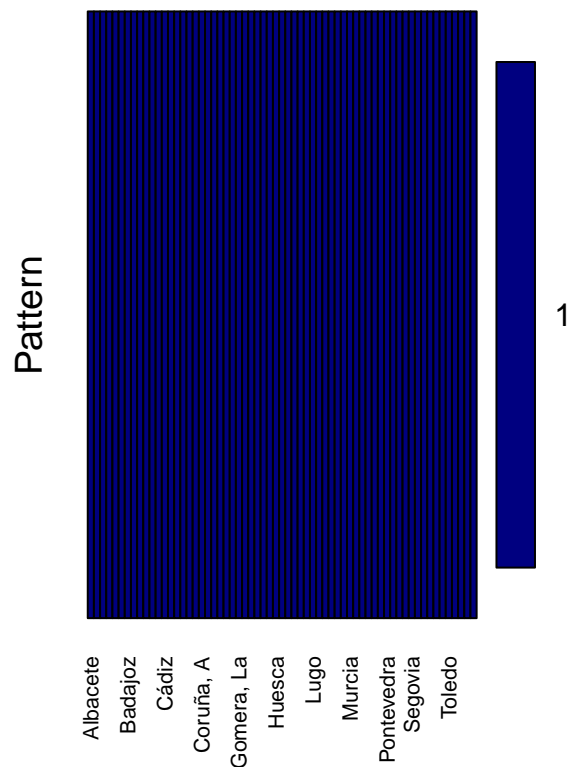
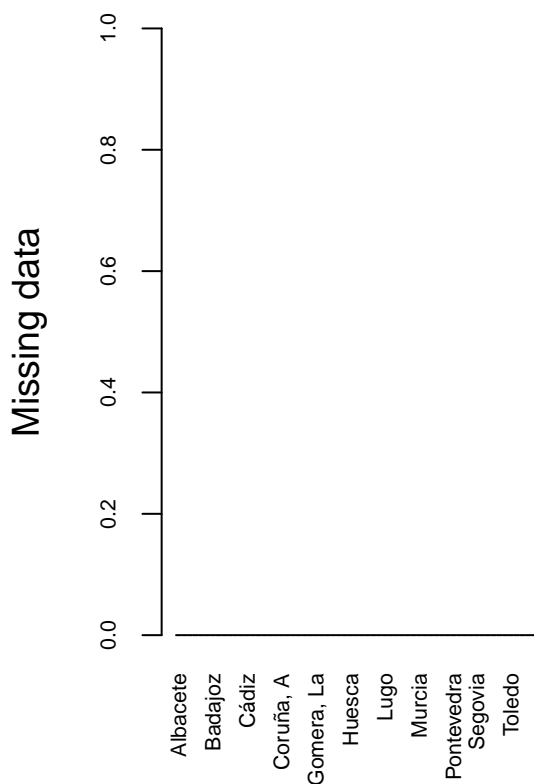
```
# We convert the time series object to a dataframe
EM3 <- ts_df(EM3_t_ts)

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))
```


Number of missing values

number of missing values

variable

```
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

2.1.3 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
##           AlmerÃ           Badajoz           Barcelona
##           385           385           385
##           Biscay           Burgos           CÃ¡ceres
##           385           385           385
##           CÃ¡diz           CÃ¡rdoba           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           M  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
```

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 Le  n" 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" "Almería"    385
## 4 "Andalusia" "Cádiz"      385
## 5 "Andalusia" "Córdoba"    385
## 6 "Andalusia" "Granada"    385
## 7 "Andalusia" "Huelva"     385
## 8 "Andalusia" "Jaén"       385
## 9 "Andalusia" "Málaga"     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 mentioned, 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 reference.

```
# Modification 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"] <- "Q"

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 Coruña"]<-"Coruña, A"
Google$sub_region_2[Google$sub_region_2=="Á\u0081lava"]<-"Araba/Álava"
Google$sub_region_2[Google$sub_region_2=="Á\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=="Almería"]<-"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=="C\u00c1ceres"]<-"C\u00c1ceres"
Google$sub_region_2[Google$sub_region_2=="C\u00c1diz"]<-"C\u00c1diz"
Google$sub_region_2[Google$sub_region_2=="C\u00c1rdoba"]<-"C\u00c1rdoba"
#Google$sub_region_2[Google$sub_region_2=="Cantabria"]<-"Cantabria"
Google$sub_region_2[Google$sub_region_2=="Castell\u00c1n"]<-"Castell\u00f3n/Castell\u00f3"
#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=="Ja\u00e9n"]<-"Ja\u00e9n"
Google$sub_region_2[Google$sub_region_2=="Las Palmas"]<-"Palmas, Las"
Google$sub_region_2[Google$sub_region_2=="Le\u00f3n"]<-"Le\u00f3n"
#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=="M\u00e1laga"]<-"M\u00e1laga"
#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)

```

```

##
##
##           4235           Albacete           Alicante/Alacant
##           385           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
## 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
## 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
## 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
## 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
```

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, place_id).
- D - “ES-” is eliminated 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$parcs_percent_change_from_baseline <- as.numeric(Google$parcs_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 Almeria AL 2020-02-15
## 2 Andalusia Almeria AL 2020-02-16
## 3 Andalusia Almeria AL 2020-02-17
## 4 Andalusia Almeria AL 2020-02-18
## 5 Andalusia Almeria 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
```

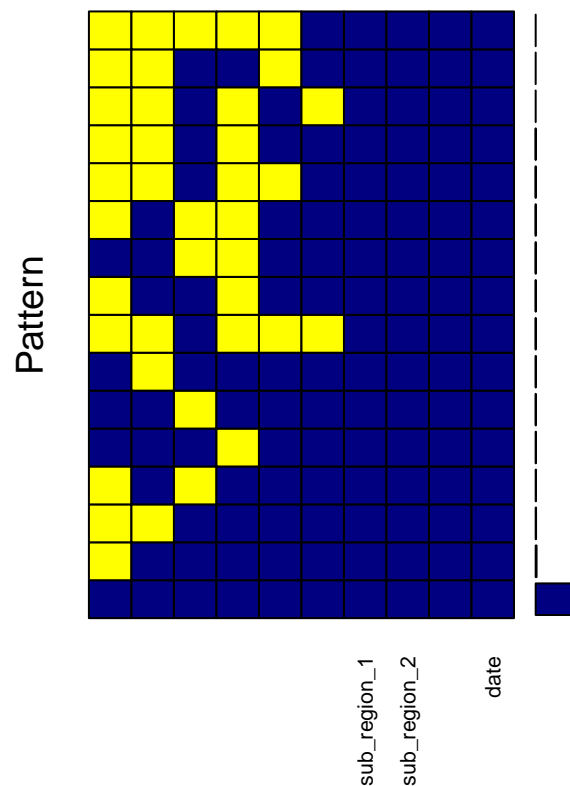
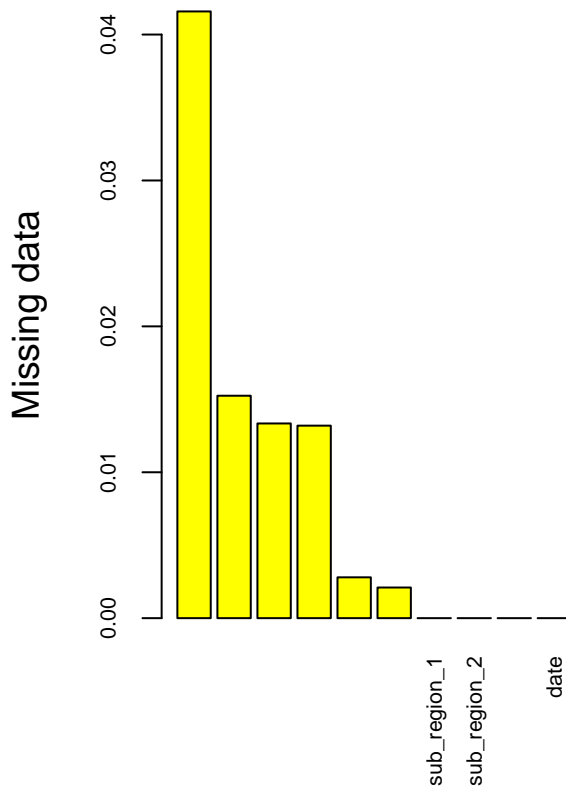
```
table(Google$iso_3166_2_code)
```

```
##
##  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
##  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
##  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)
```

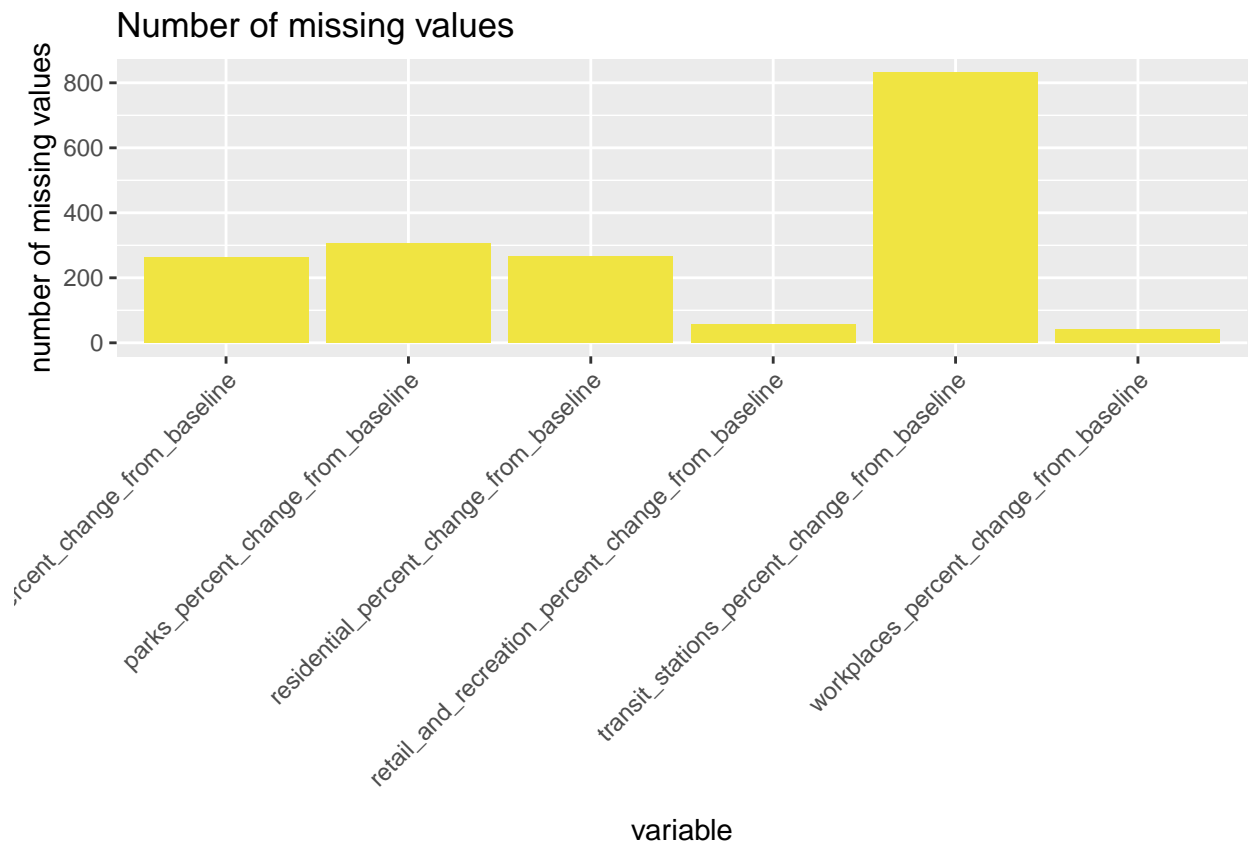
We check missing values.

```
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
```

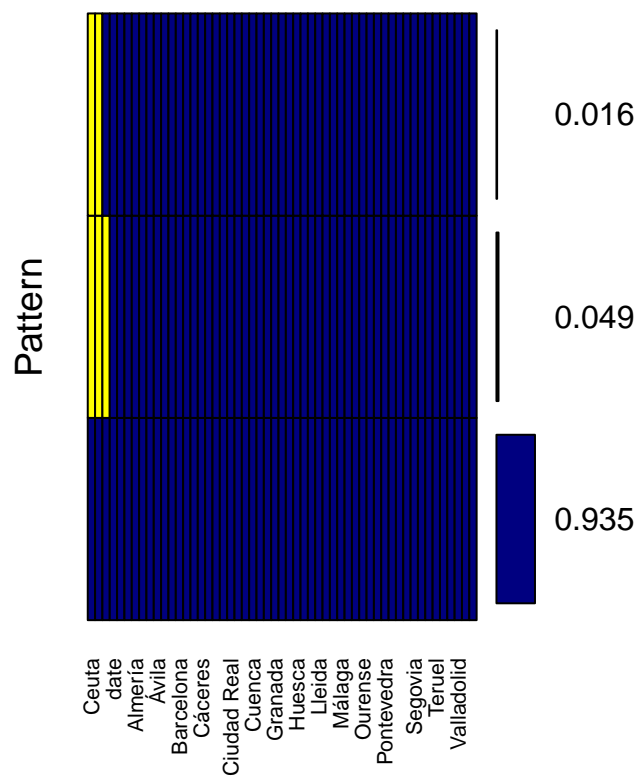
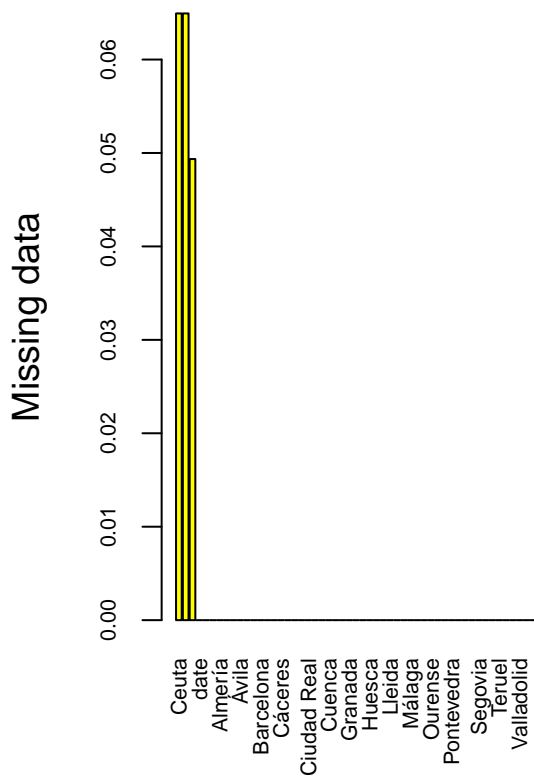
```
Google %>%
gather(key = "key", value = "val") %>%
mutate(is.missing = is.na(val)) %>%
group_by(key, is.missing) %>%
summarise(num.missing = n()) %>%
filter(is.missing==T) %>%
select(-is.missing) %>%
arrange(desc(num.missing)) %>%
ggplot() +
geom_bar(aes(x=key, y=num.missing), stat = 'identity', fill="#F0E442") +
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 input 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)

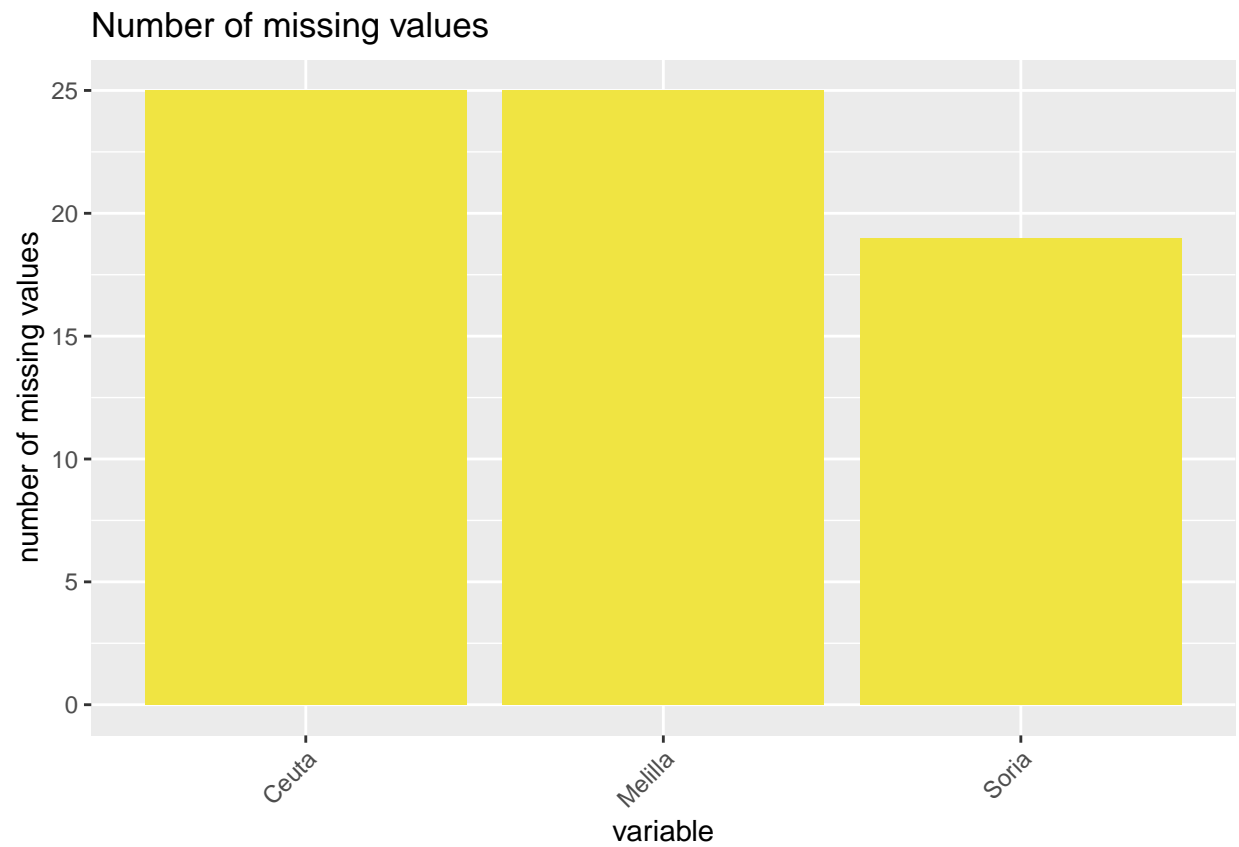
# Visualize missing values
aggr(Google_t_retail, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google_t_retail), cex.axis=.7,
     gap=3, ylab=c("Missing data", "Pattern"))
```



```
##
## 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))
```

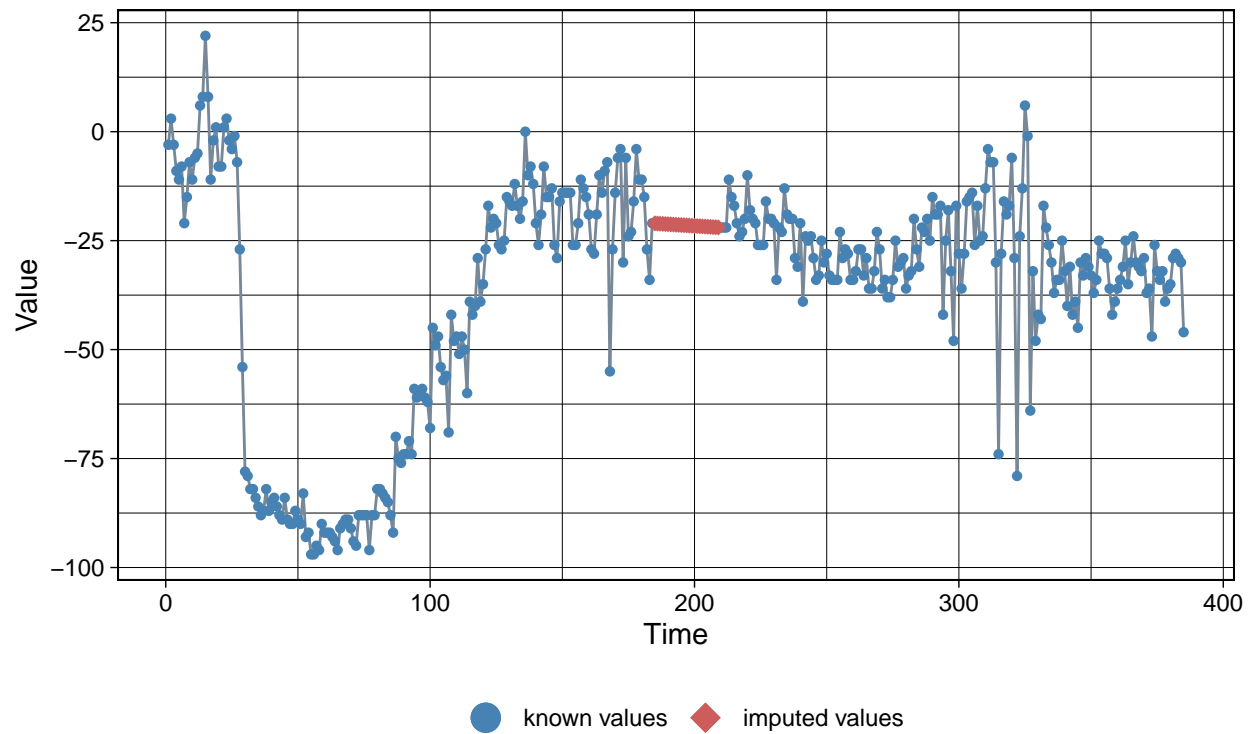


```
# Convert dataframe to ts object
Google_t_retail_ts<-xts(Google_t_retail[,1],Google_t_retail$date)

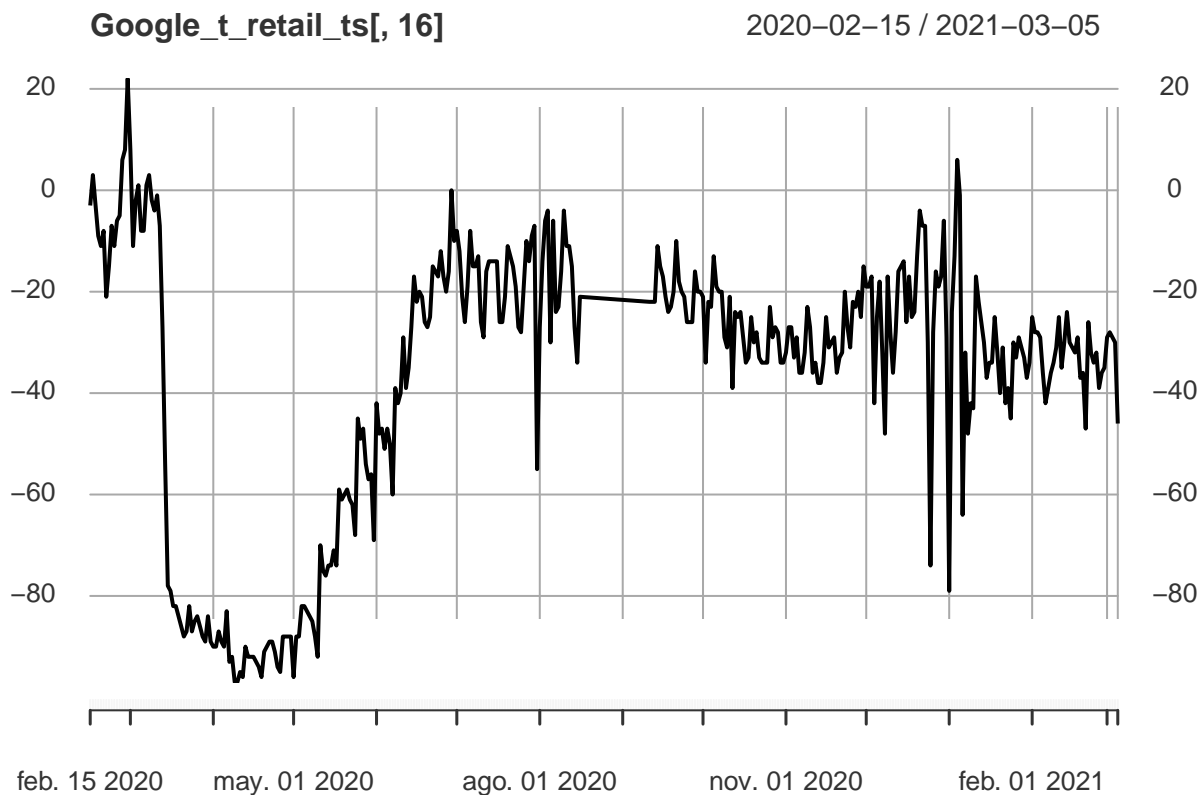
# Impute the missing values with na_seadec (i.e Ceuta)
imp5 <- na_seadec(Google_t_retail_ts[,16])
ggplot_na_imputations(Google_t_retail_ts[,16], imp5)
```


Imputed Values

Visualization of missing value replacements



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

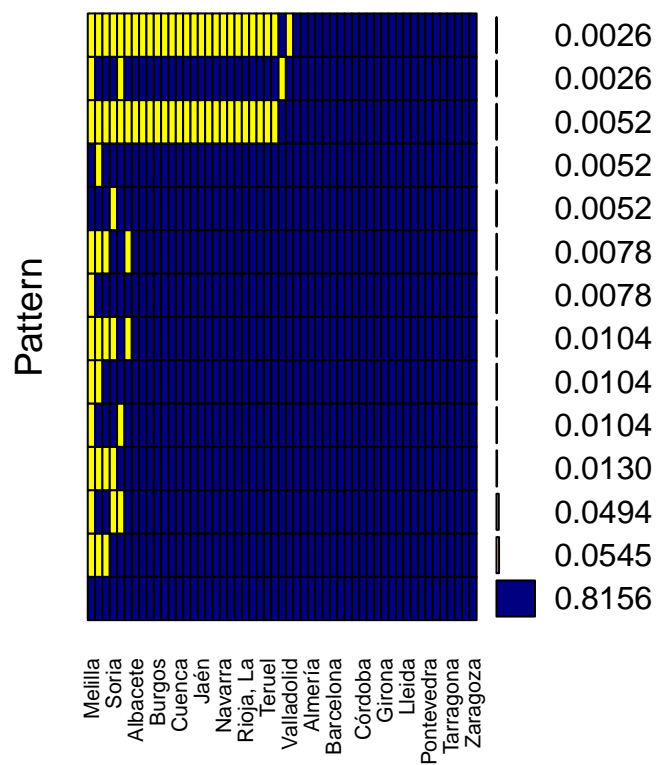
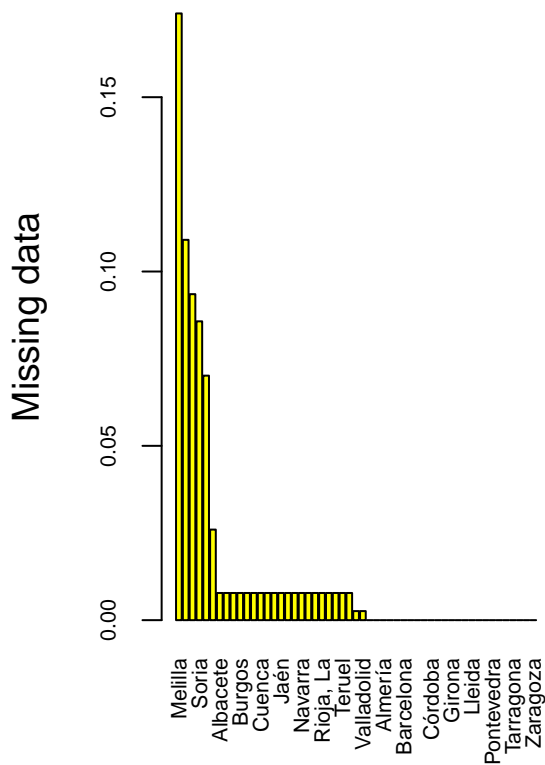


```
# We convert the time series object to a dataframe
Google_retail <- ts_df(Google_t_retail_ts)

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_base"

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

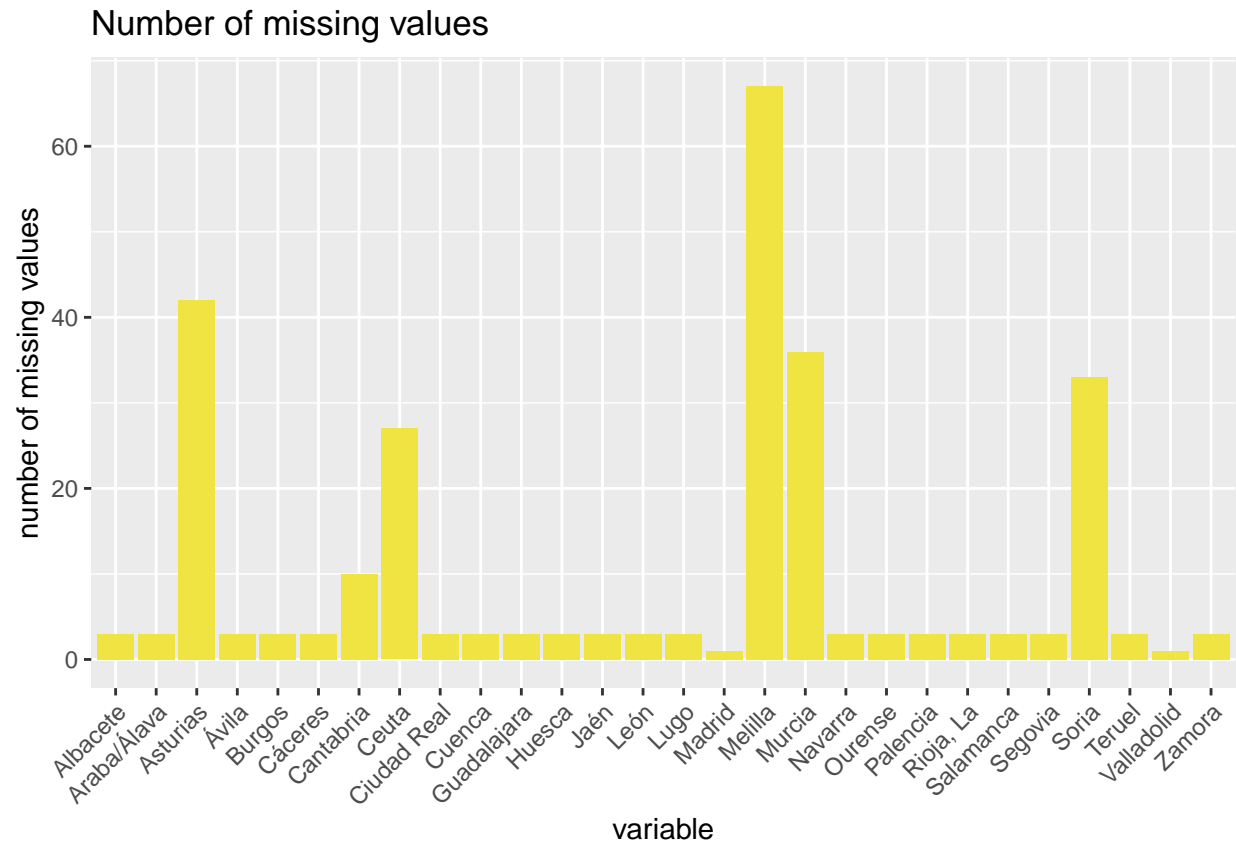
# Visualize missing values
aggr(Google_t_grocery, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google_t_grocery), cex.axis=.7,
     gap=3, ylab=c("Missing data", "Pattern"))
```



```
##
## 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))
```

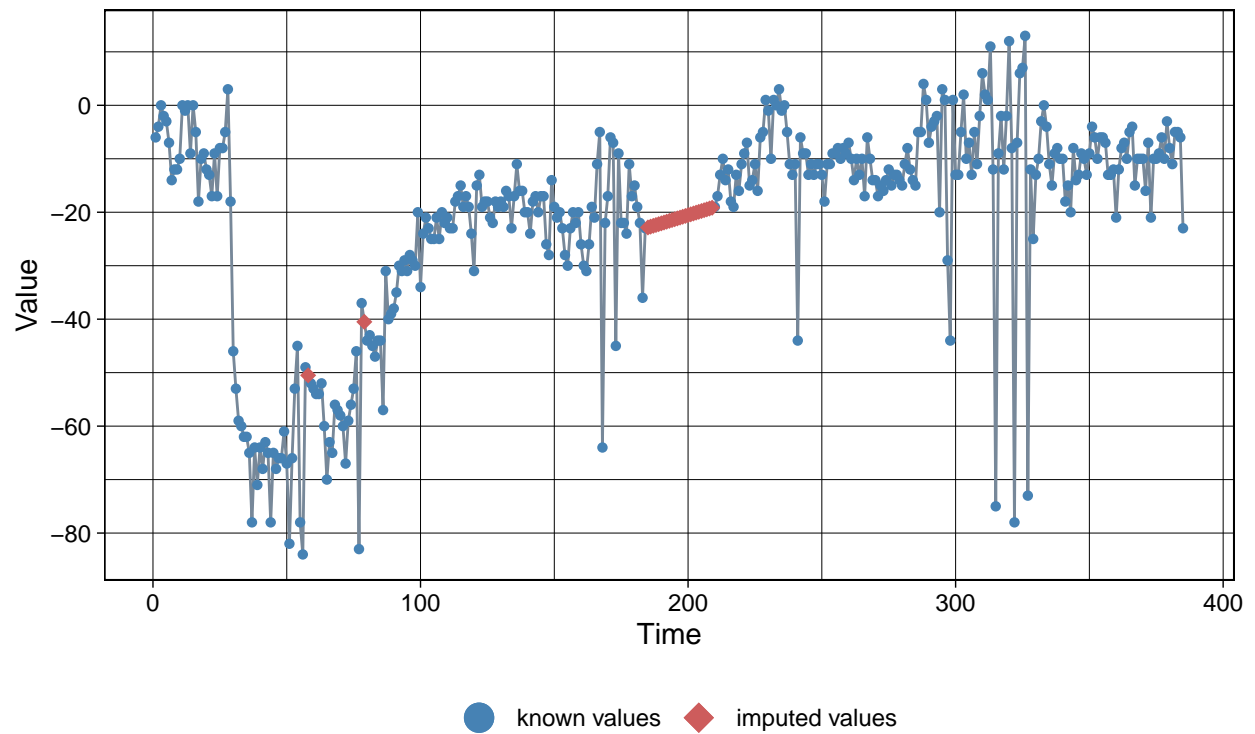


```
# Convert dataframe to ts object
Google_t_grocery_ts<-xts(Google_t_grocery[-1],Google_t_grocery$date)

# 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)
```

Imputed Values

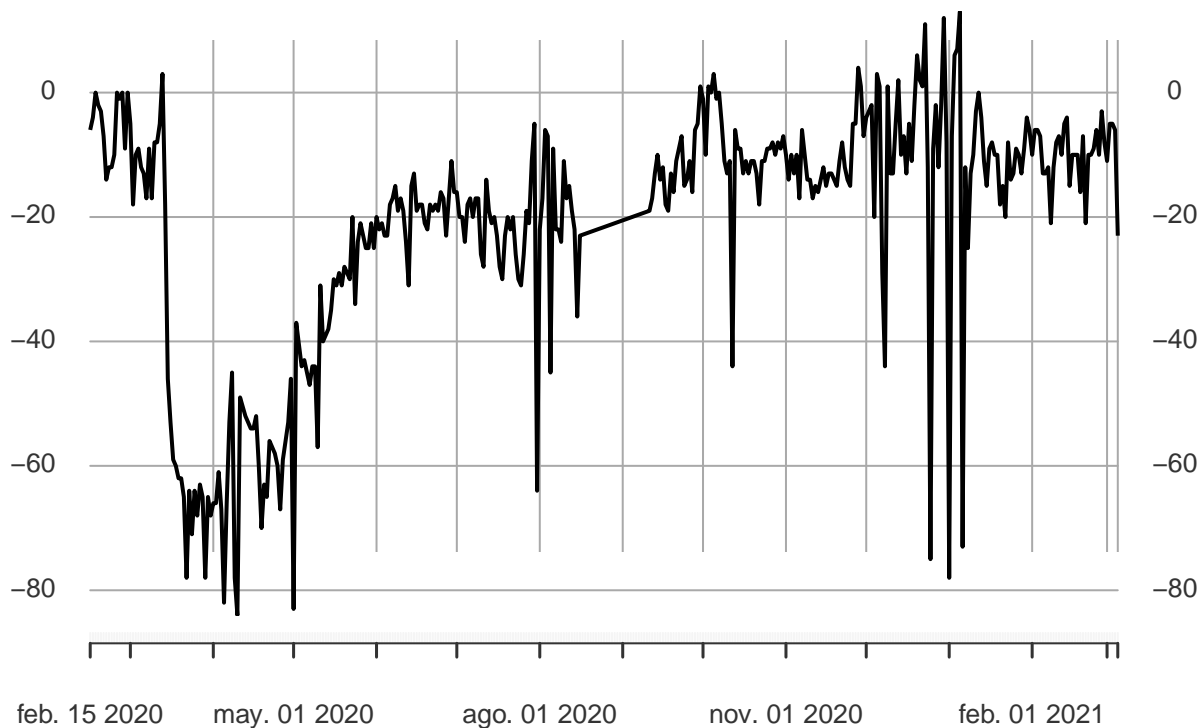
Visualization of missing value replacements



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

Google_t_grocery_ts[, 16]

2020-02-15 / 2021-03-05

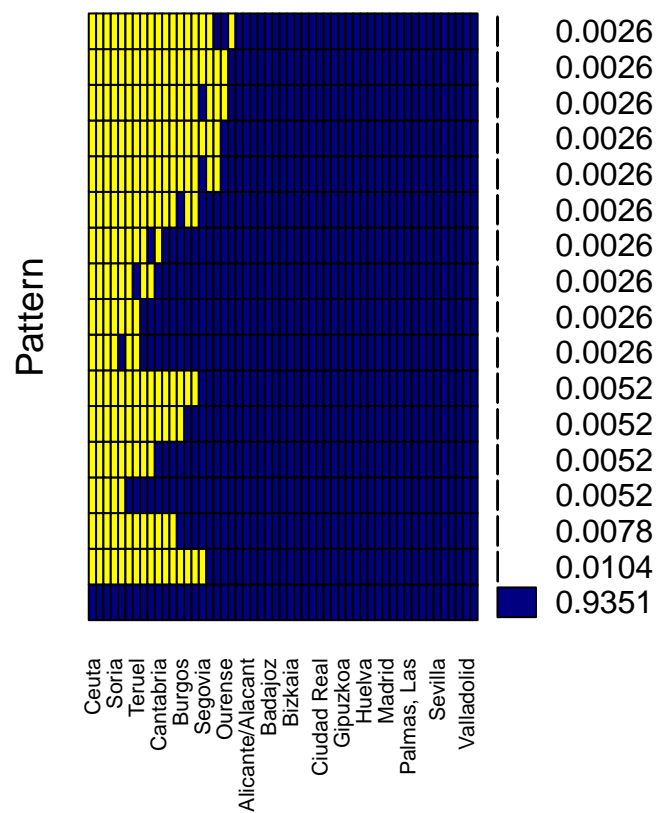
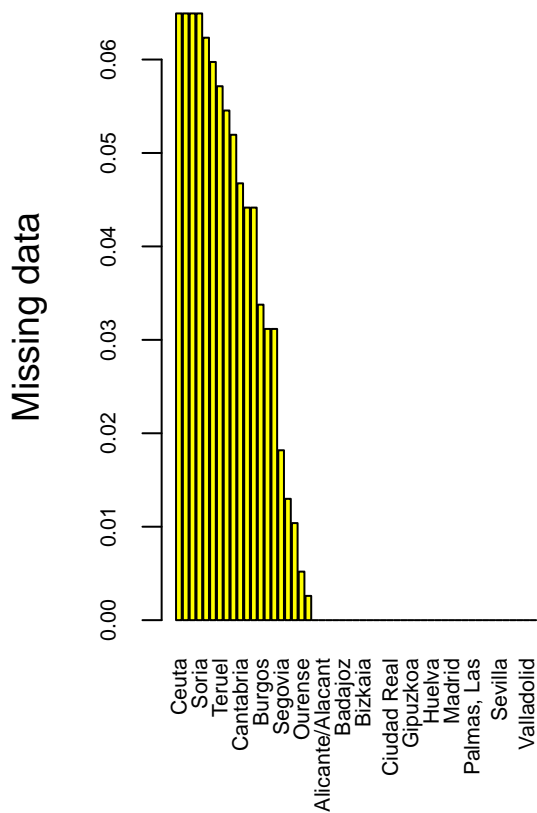


```
# We convert the time series object to a dataframe
Google_grocery <- ts_df(Google_t_grocery_ts)

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_base"

#####
# 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"))
```

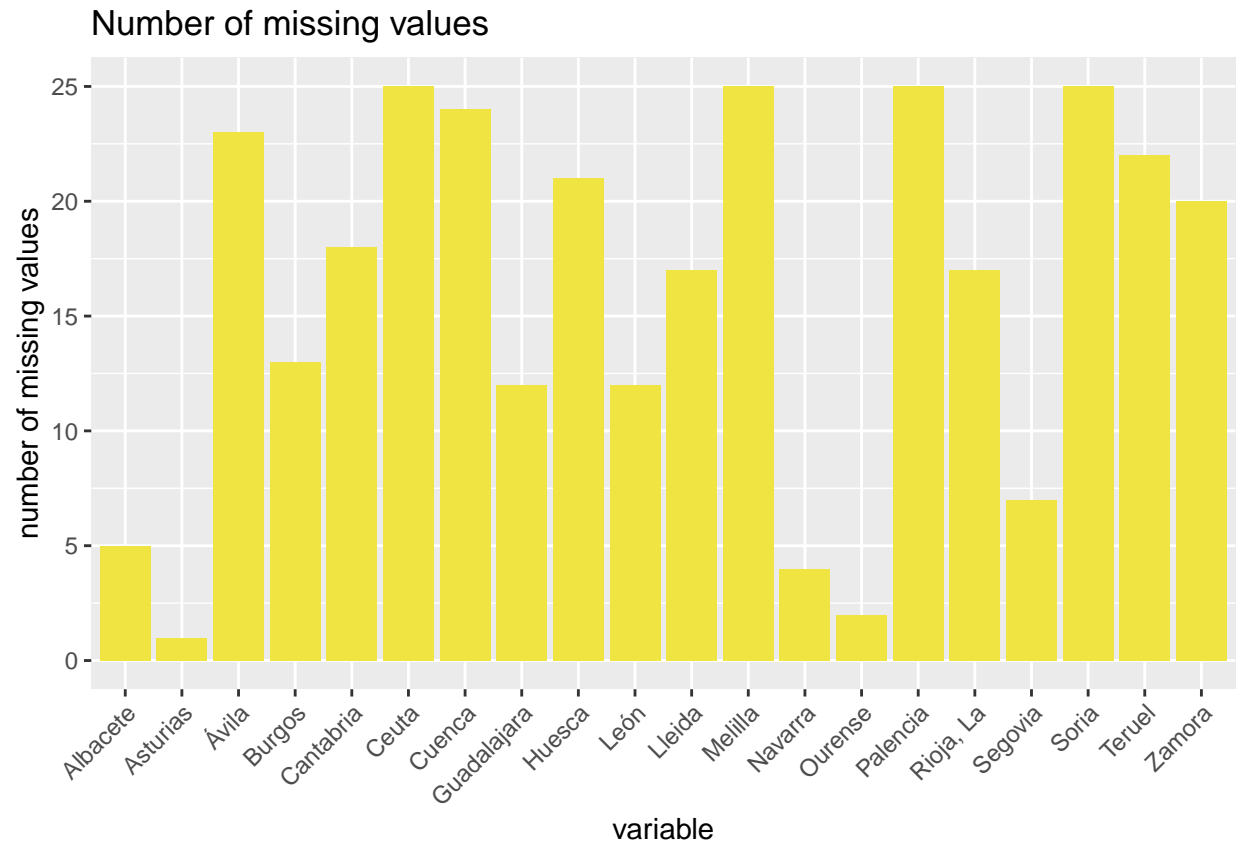


```
##
## Variables sorted by number of missings:
##      Variable      Count
##      Ceuta 0.064935065
##      Melilla 0.064935065
##      Palencia 0.064935065
##      Soria 0.064935065
##      Cuenca 0.062337662
##      Ávila 0.059740260
##      Teruel 0.057142857
##      Huesca 0.054545455
##      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))
```

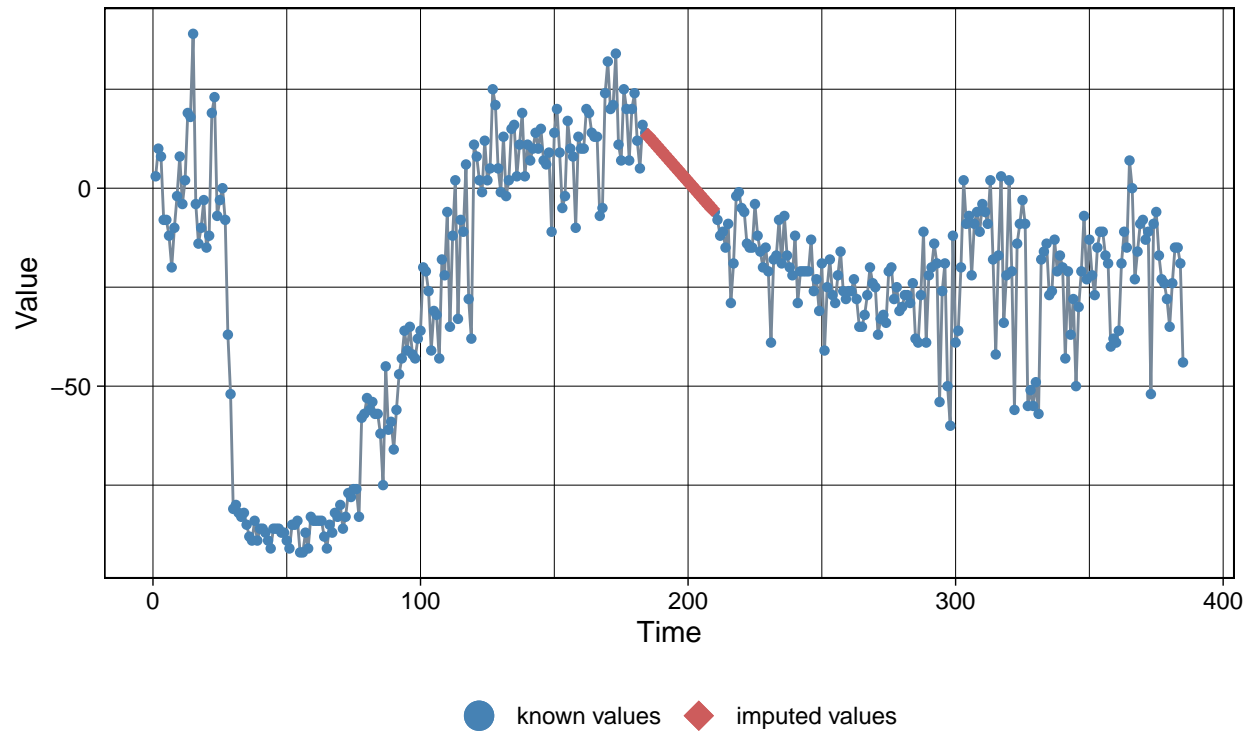


```
# Convert dataframe to ts object
Google_t_parks_ts<-xts(Google_t_parks[-1],Google_t_parks$date)

# 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)
```

Imputed Values

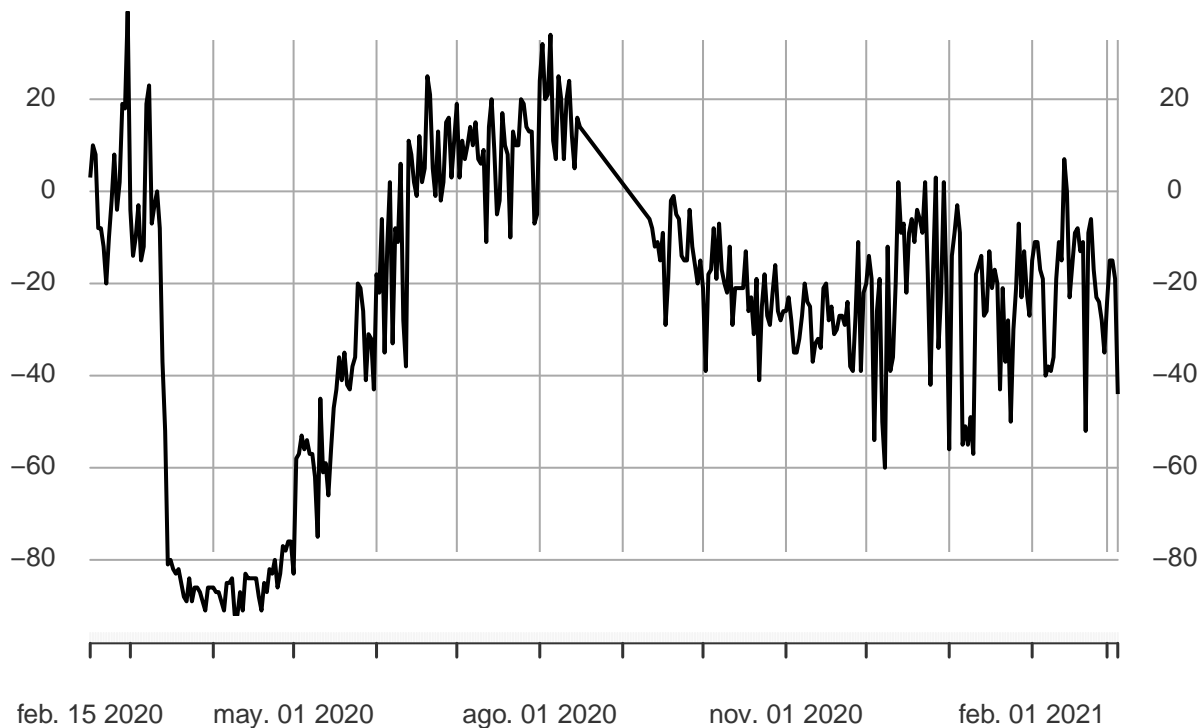
Visualization of missing value replacements



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

Google_t_parks_ts[, 16]

2020-02-15 / 2021-03-05

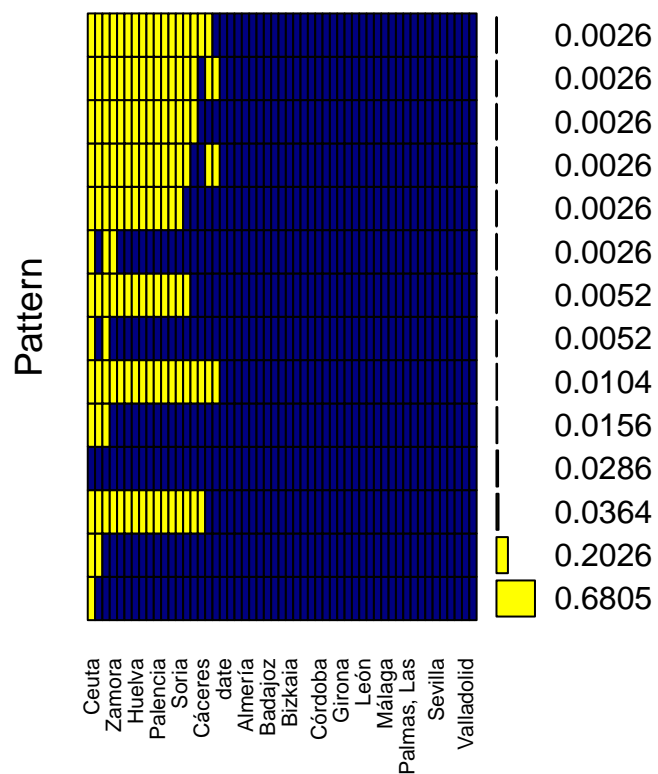
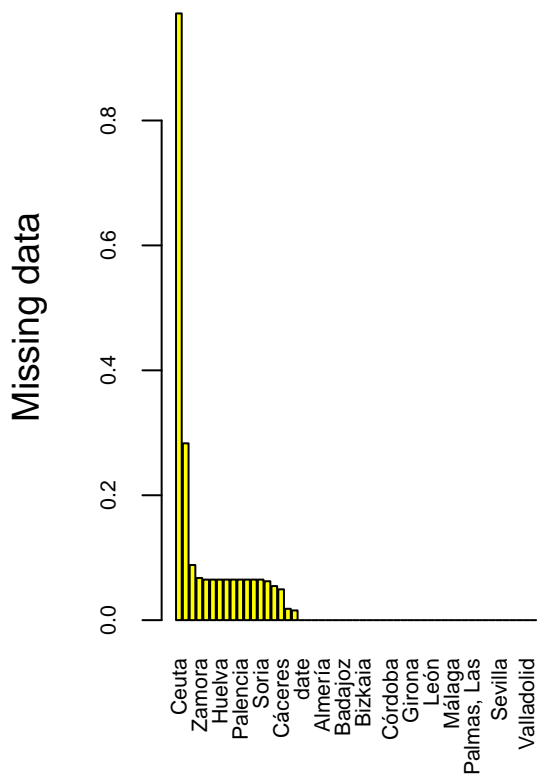


```
# We convert the time series object to a dataframe
Google_parks <- ts_df(Google_t_parks_ts)

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"))
```

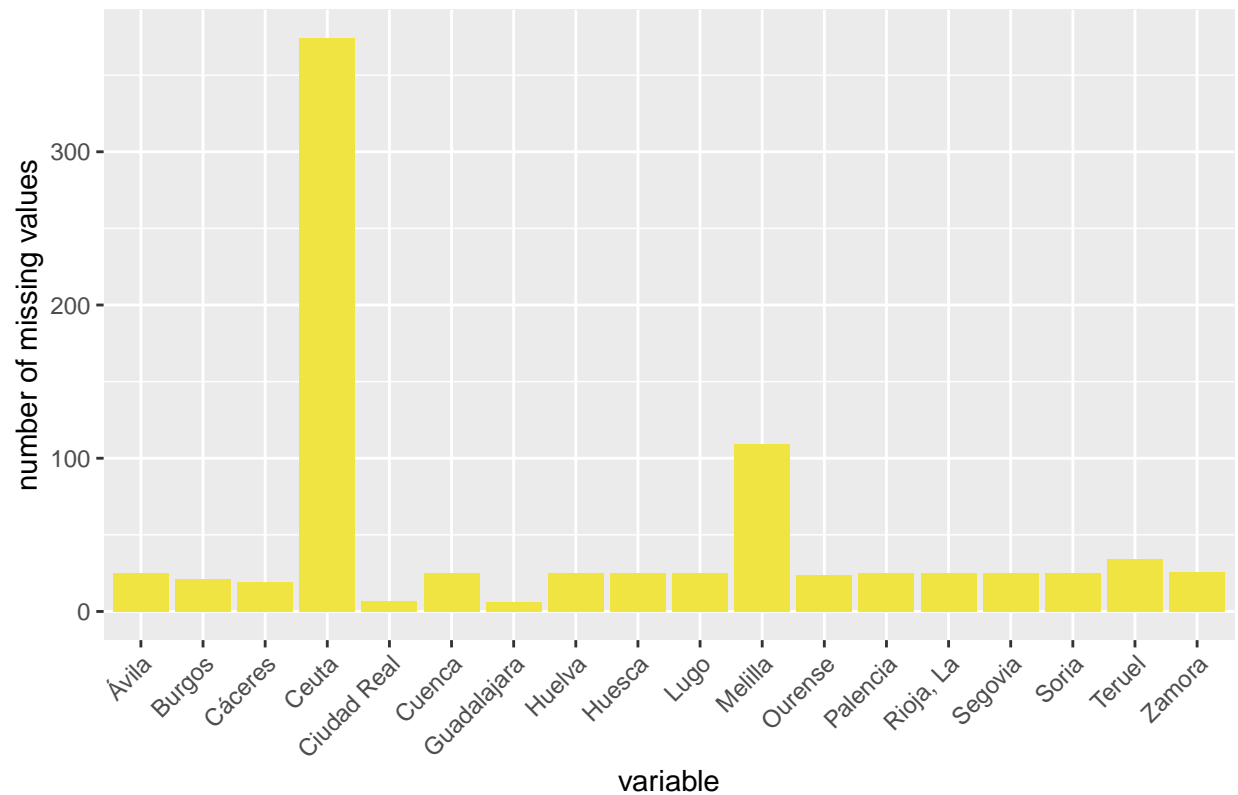


```
##
## Variables sorted by number of missings:
##      Variable      Count
##      Ceuta 0.97142857
##      Melilla 0.28311688
##      Teruel 0.08831169
##      Zamora 0.06753247
##      Ávila 0.06493506
##      Cuenca 0.06493506
##      Huelva 0.06493506
##      Huesca 0.06493506
##      Lugo 0.06493506
##      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))
```

Number of missing values

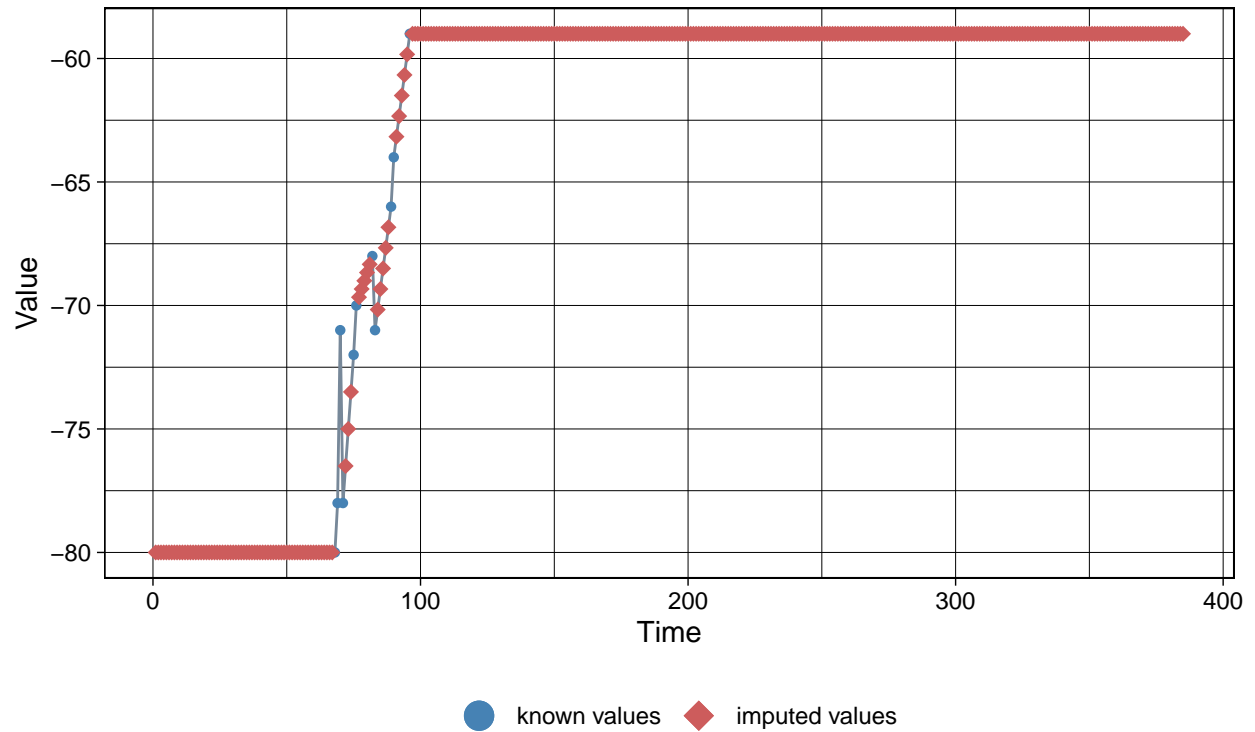


```
# Convert dataframe to ts object
Google_t_transit_ts<-xts(Google_t_transit[-1],Google_t_transit$date)

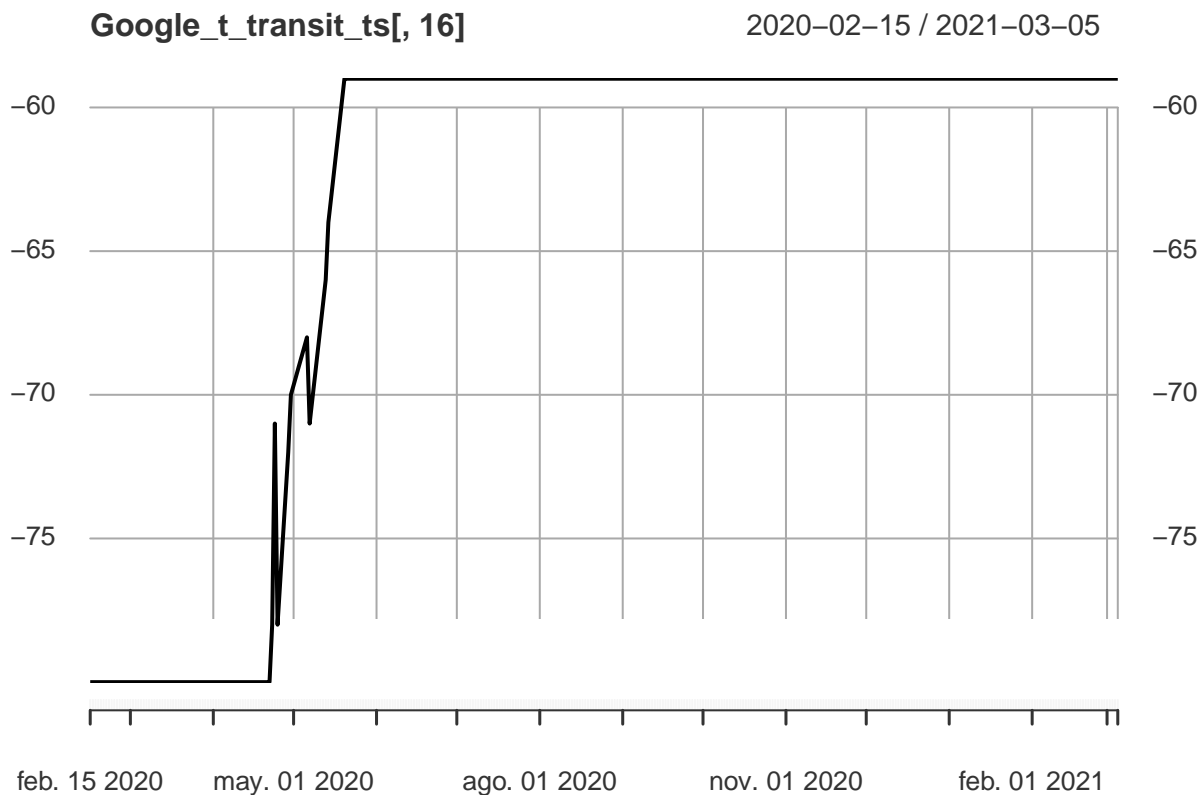
# Impute the missing values with na_seadec (i.e Ceuta)
imp8 <- na_seadec(Google_t_transit_ts[,16])
ggplot_na_imputations(Google_t_transit_ts[,16], imp8)
```

Imputed Values

Visualization of missing value replacements



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

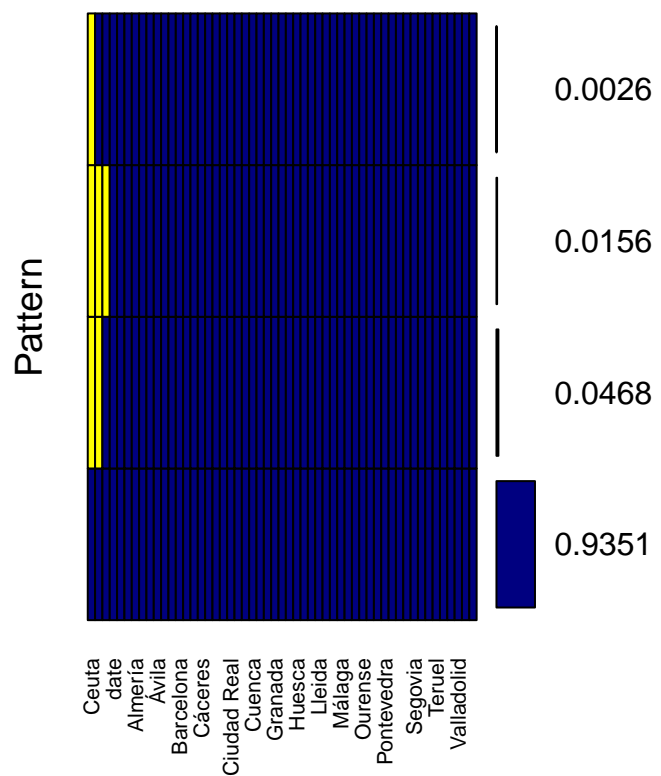
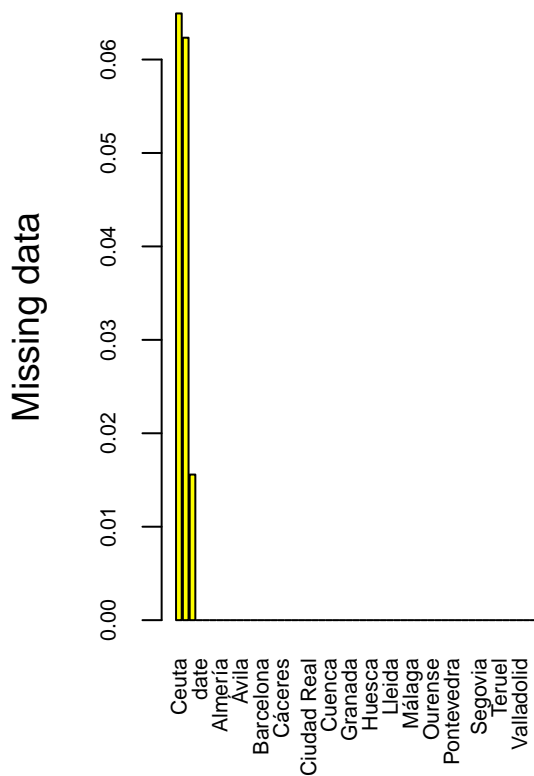



```
# We convert the time series object to a dataframe
Google_transit <- ts_df(Google_t_transit_ts)

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)

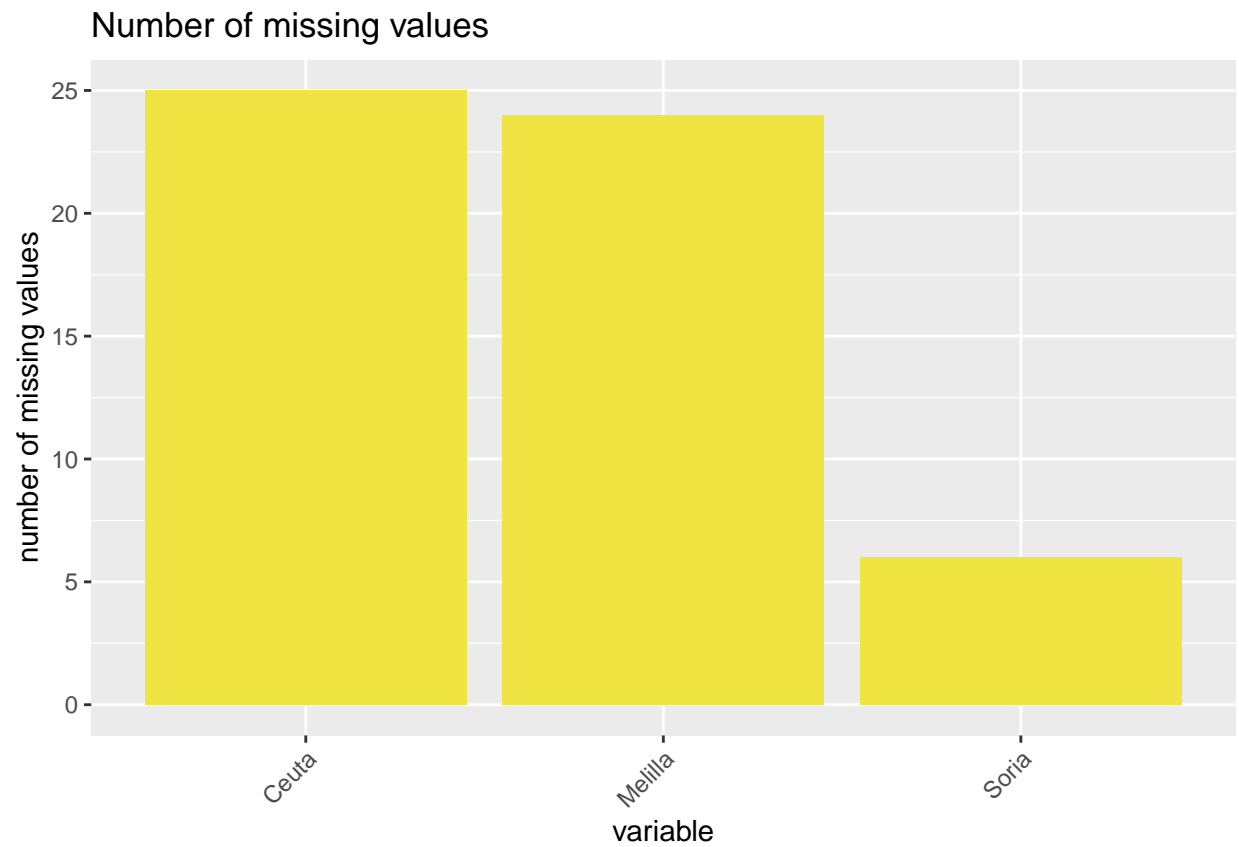
# Visualize missing values
aggr(Google_t_workplaces, col=c('navyblue','yellow'),
     numbers=TRUE, sortVars=TRUE,
     labels=names(Google_t_workplaces), cex.axis=.7,
     gap=3, ylab=c("Missing data", "Pattern"))
```



```
##
## 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
##          Guadaluja 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
```

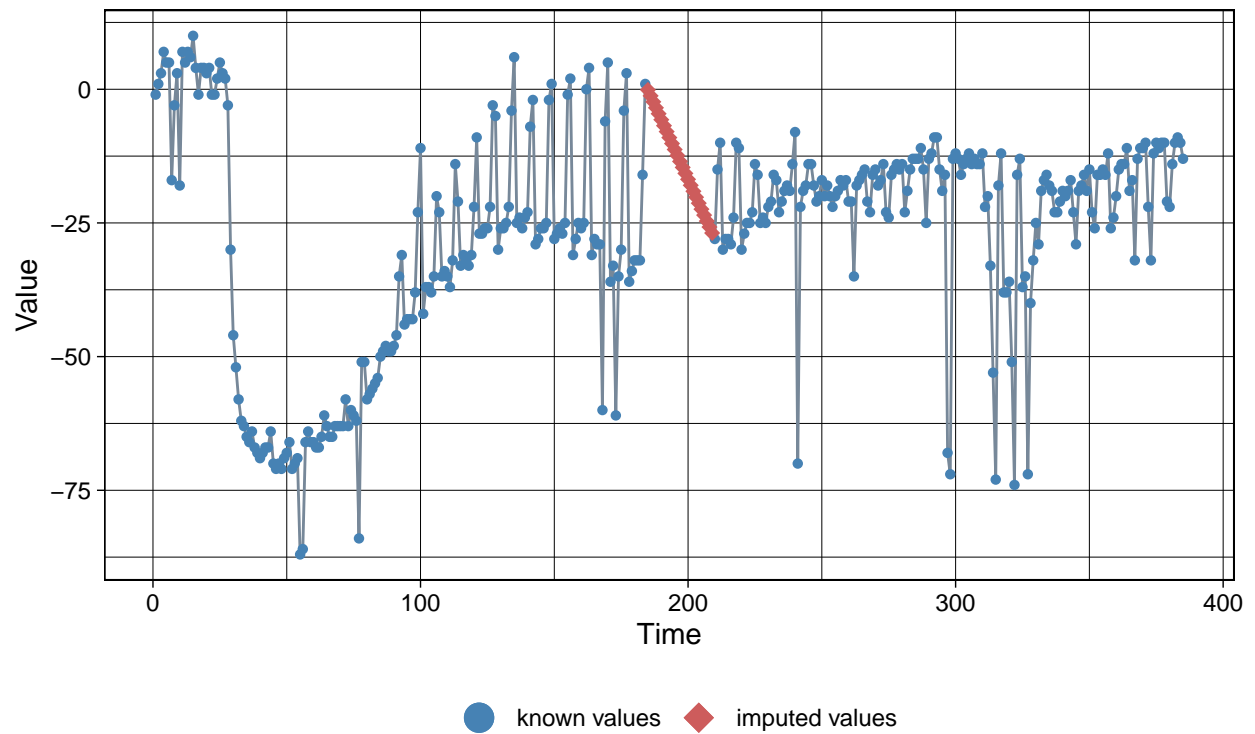
```
Google_t_workplaces %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), 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)
```

Imputed Values

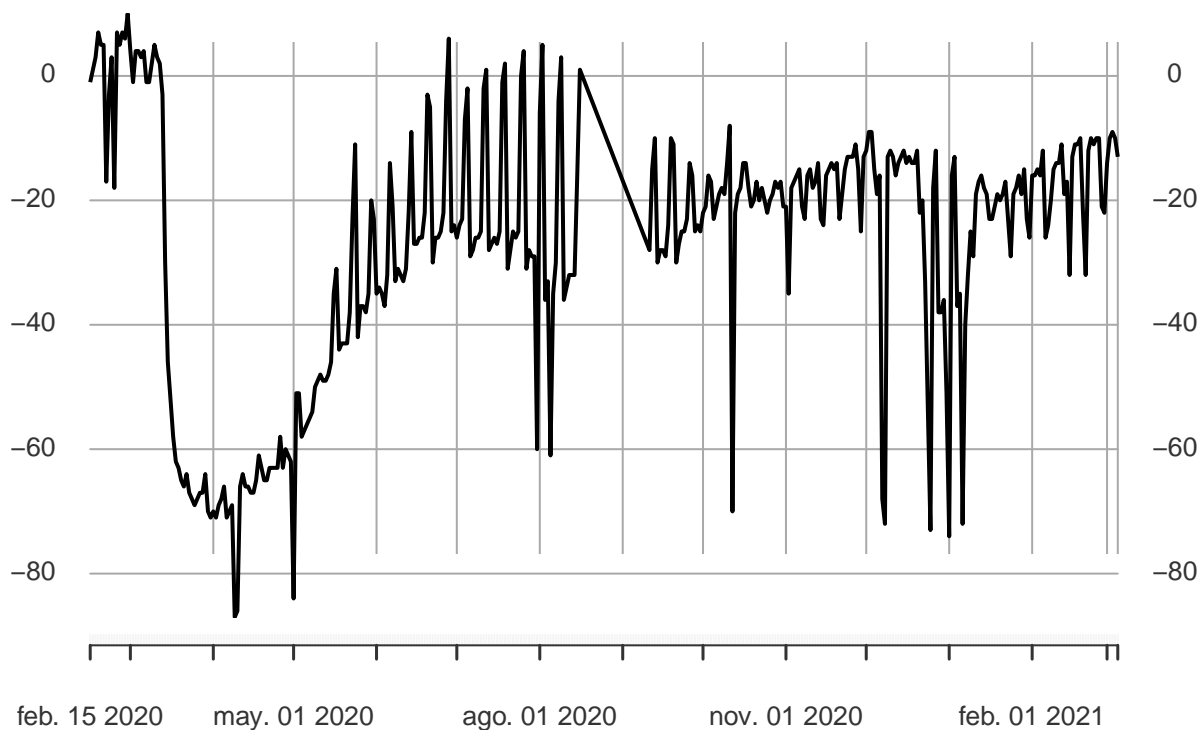
Visualization of missing value replacements



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

Google_t_workplaces_ts[, 16]

2020-02-15 / 2021-03-05

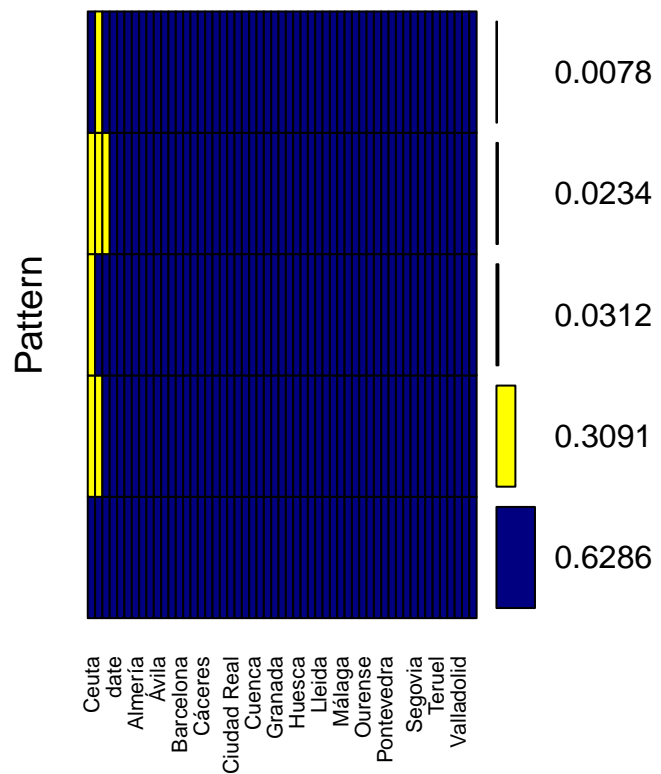
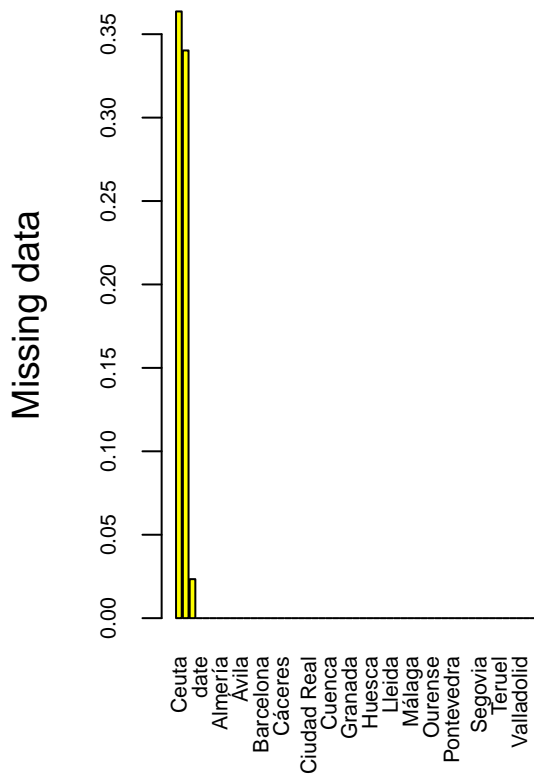


```
# We convert the time series object to a dataframe
Google_workplaces <- ts_df(Google_t_workplaces_ts)

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
##          Guadaluja 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))
```

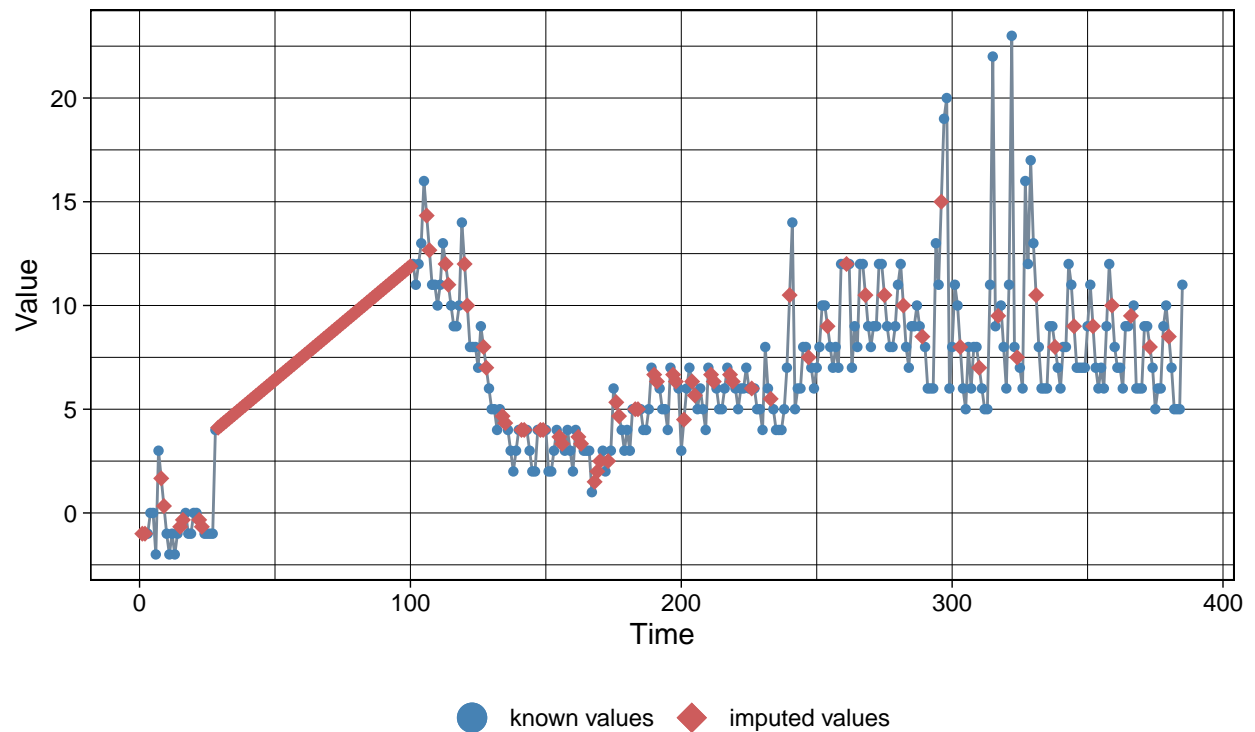



```
# Convert dataframe to ts object
Google_t_residential_ts<-xts(Google_t_residential[-1],Google_t_residential$date)

# 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)
```

Imputed Values

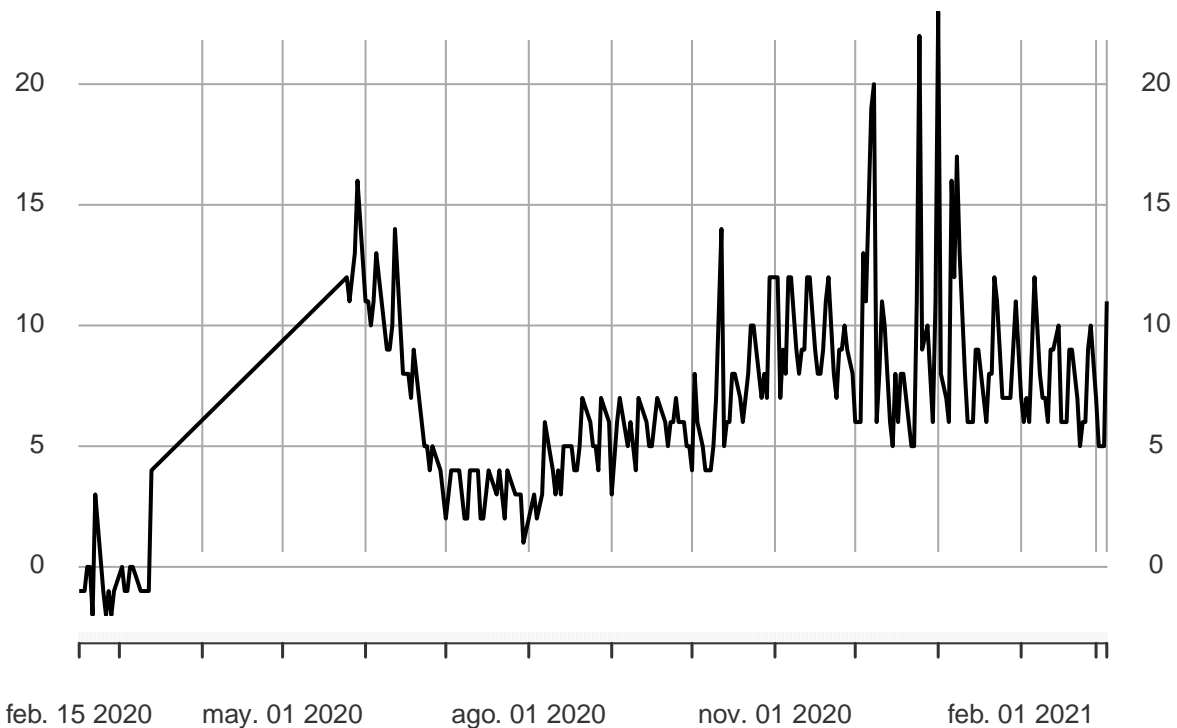
Visualization of missing value replacements



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

Google_t_residential_ts[, 16]

2020-02-15 / 2021-03-05



```
# 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_baseline"
```

Now we merge the previous dataframes into new one with the imputed values 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)

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
```

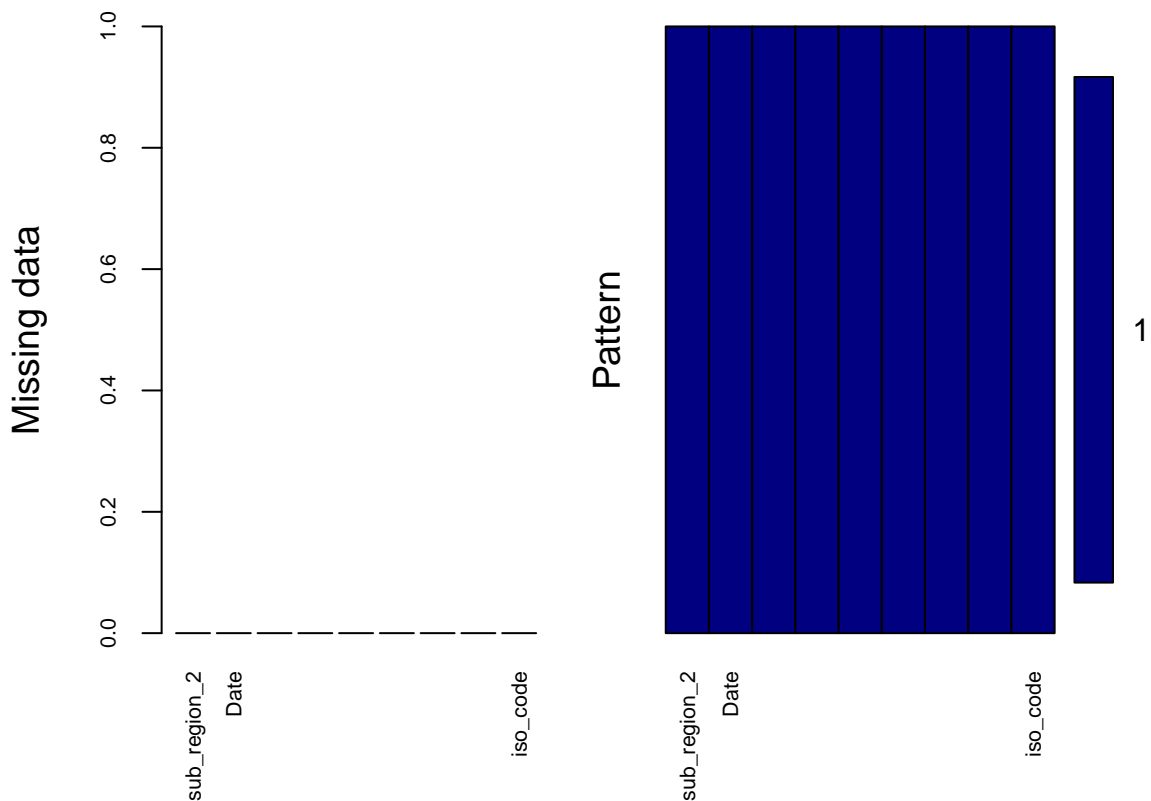
```
##          León          Lleida          Lugo
##          385          385          385
##          Madrid        Málaga        Melilla
##          385          385          385
##          Murcia        Navarra        Ourense
##          385          385          385
##          Palencia      Palmas, 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
##  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
##  SG  SO  SS  T  TE  TF  TO  V  VA  VI  Z  ZA
## 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))
```

Number of missing values

number of missing values

variable

2.1.4 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
```

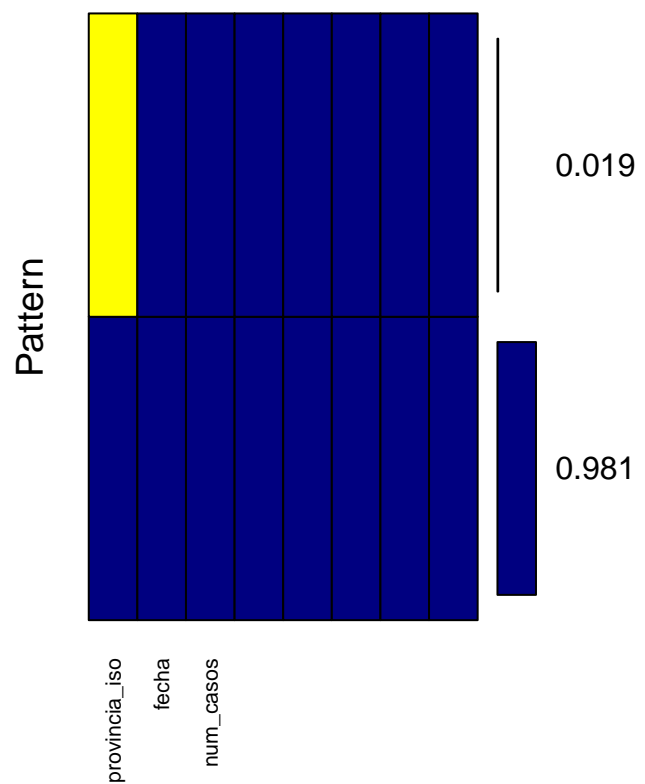
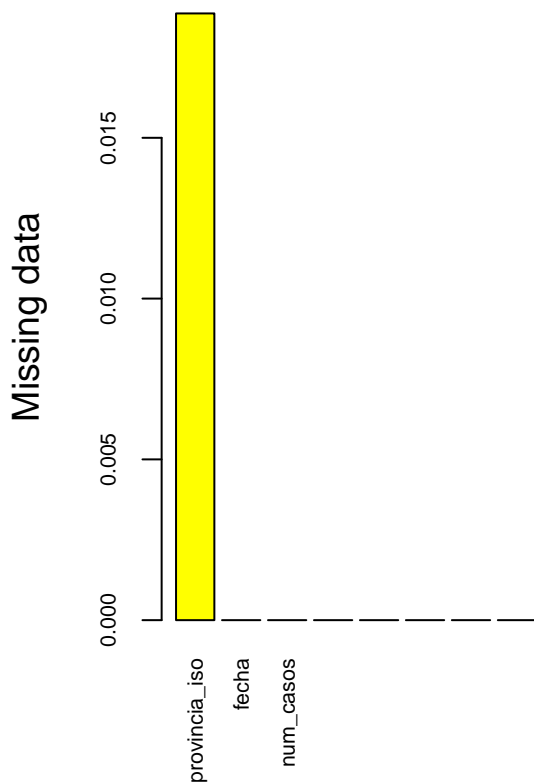
```
##                               Mean   : 136.9   Mean   : 109.6
##                               3rd Qu.: 120.0   3rd Qu.: 100.0
##                               Max.    :6972.0   Max.    :6546.0
## num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## Min.    : 0.0000          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
##  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
## SG  SO  SS  T  TE  TF  TO  V  VA  VI  Z  ZA
## 442 442 442 442 442 442 442 442 442 442 442 442
```

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"))
```

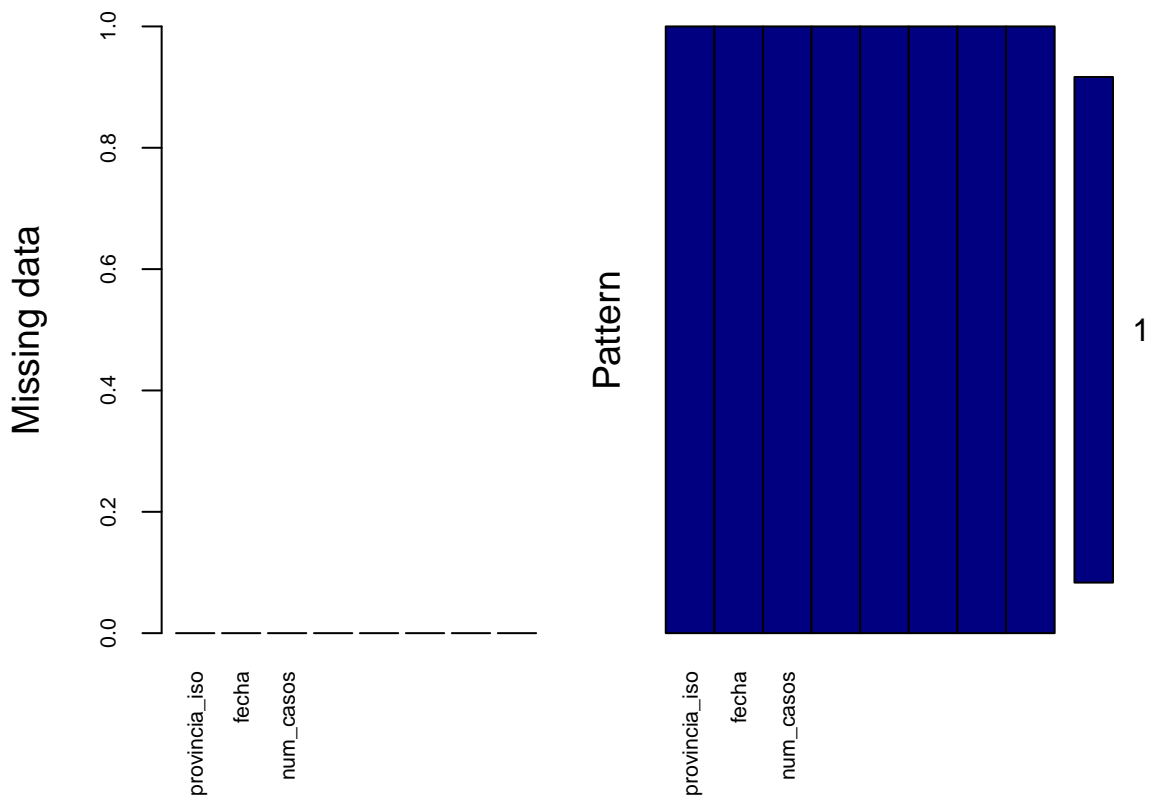



```
##
## Variables sorted by number of missings:
##           Variable      Count
##      provincia_iso 0.01886792
##           fecha 0.00000000
##      num_casos 0.00000000
## num_casos_prueba_pcr 0.00000000
## num_casos_prueba_test_ac 0.00000000
## num_casos_prueba_ag 0.00000000
## num_casos_prueba_elisa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

```
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))
```

Number of missing values

number of missing values

variable

```
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
```

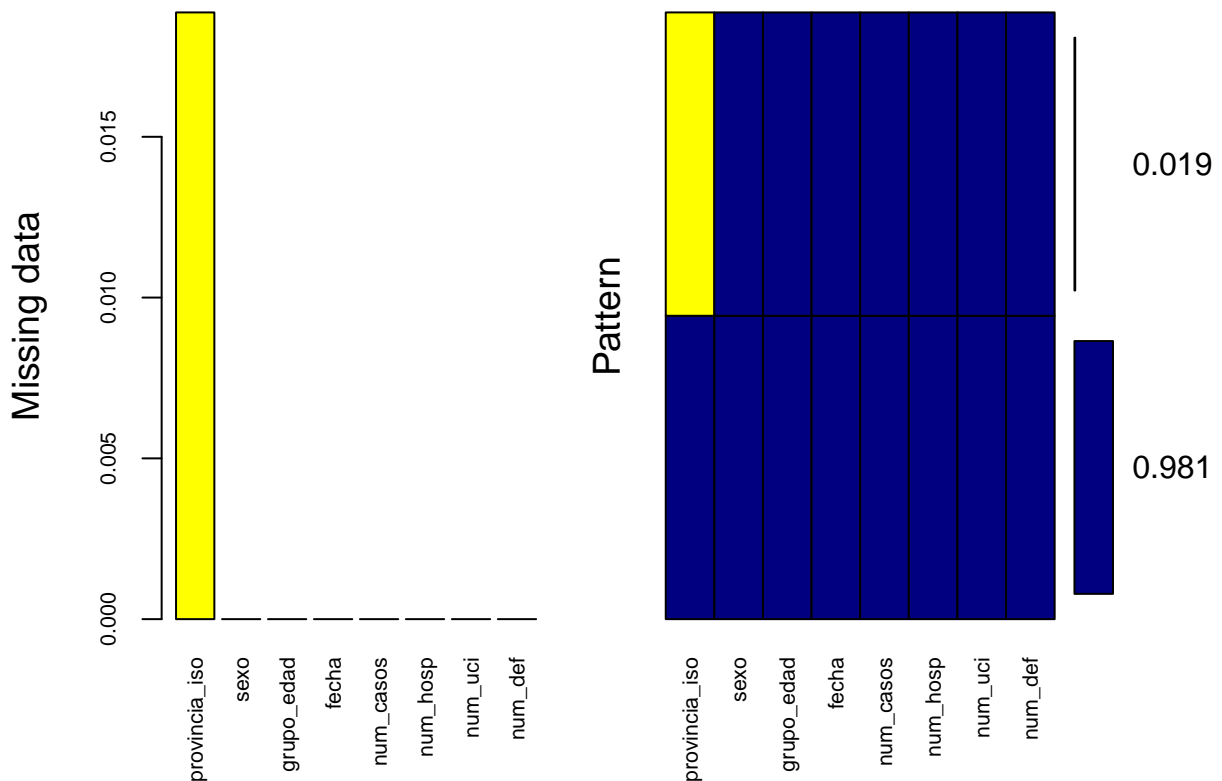
```
## 3rd Qu.: 2.000 3rd Qu.: 0.0000 3rd Qu.: 0.00000 3rd Qu.: 0.0000
## Max. :771.000 Max. :269.0000 Max. :35.00000 Max. :100.0000
```

```
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
```

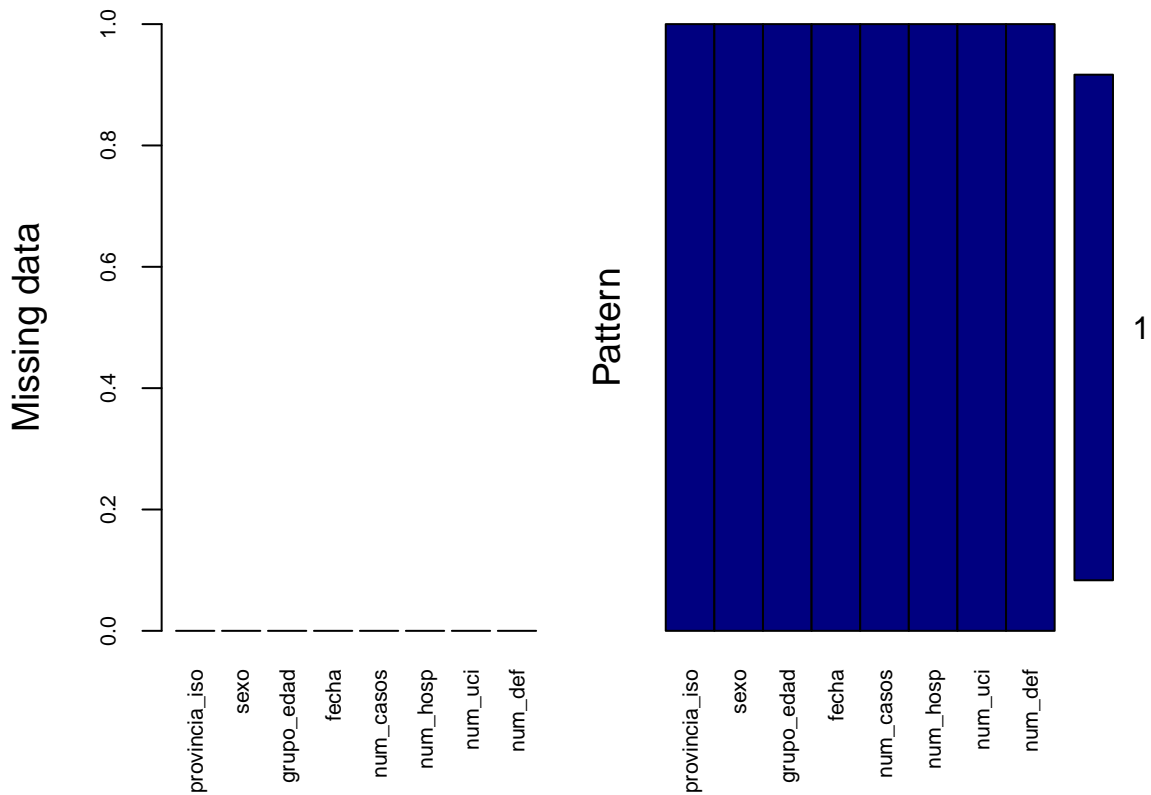
```
CNE_casos %>%
  gather(key = "key", value = "val") %>%
  mutate(is.missing = is.na(val)) %>%
  group_by(key, is.missing) %>%
  summarise(num.missing = n()) %>%
  filter(is.missing==T) %>%
  select(-is.missing) %>%
  arrange(desc(num.missing)) %>%
  ggplot() +
  geom_bar(aes(x=key, y=num.missing), 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))
```

Number of missing values

number of missing values

variable

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 grouped 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
##   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
##  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
```



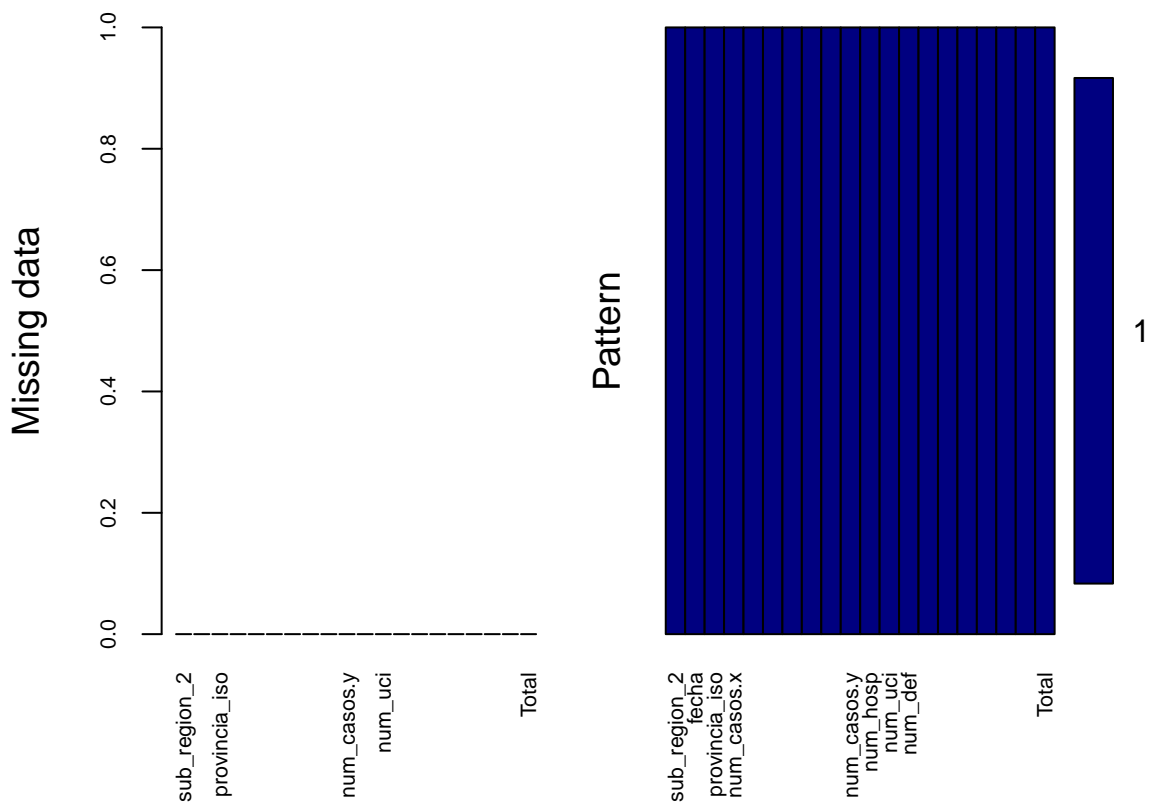
```
##      Valencia/València      Valladolid      Zamora
##              290              290              290
##      Zaragoza
##              290
```

```
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))
```

Number of missing values

number of missing values

variable

```
# 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

```

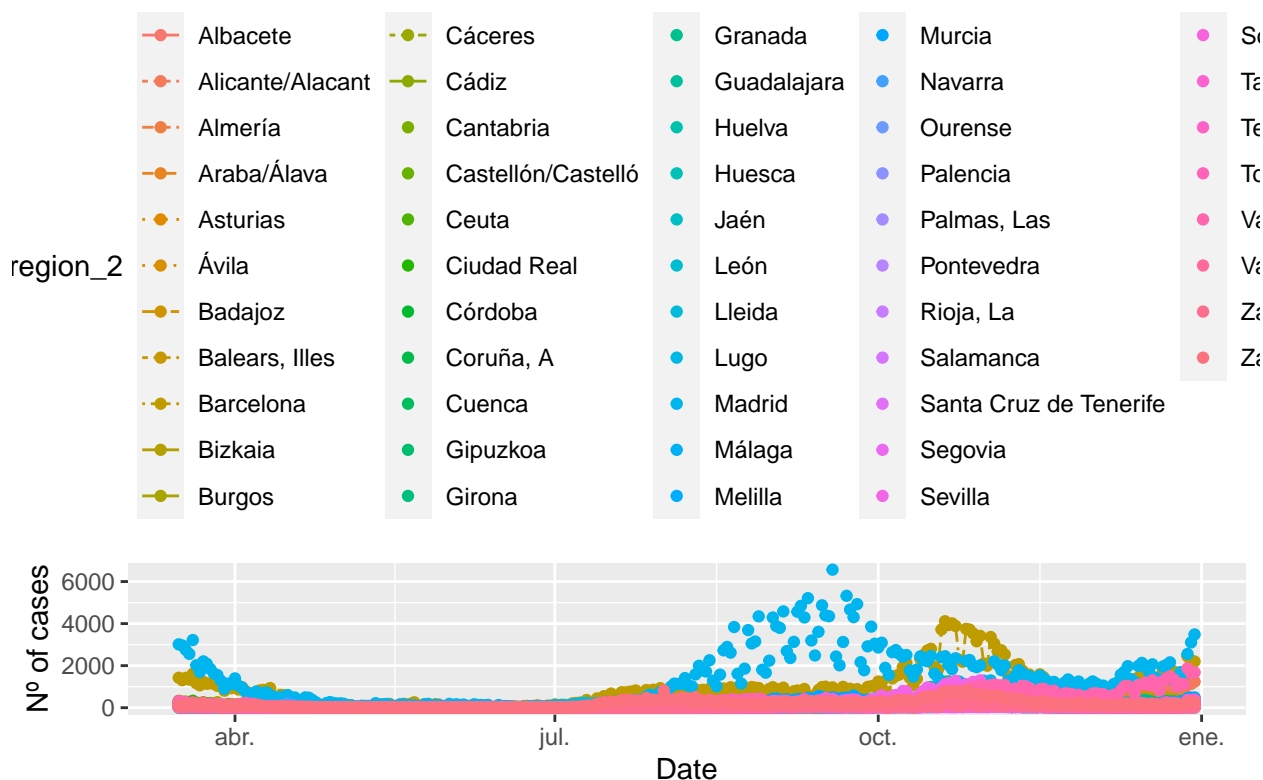
```
write.csv2(Total,"D:\\UOC Master Data Science\\_ M2.882 - TFM - Área 5\\UOC - Guia - PECS\\Pec3\\Total.
```

2.3 Dataframe plots

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

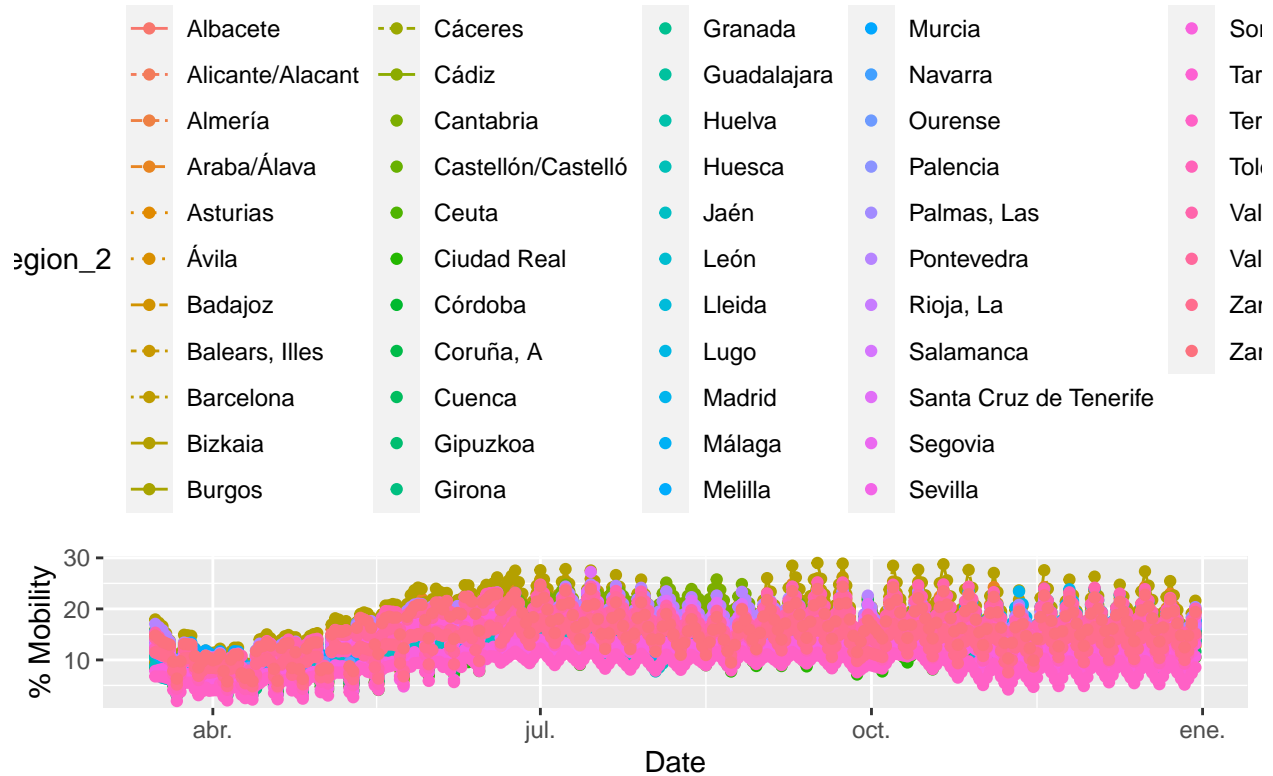
```
# Line plots
# All num_casos.x
ggplot(Total, aes(x=fecha, y=num_casos.x, group=sub_region_2)) +
  geom_line(aes(linetype=sub_region_2, color=sub_region_2))+
  geom_point(aes(color=sub_region_2))+
  theme(legend.position="top") +
  labs(title="Cases by Province",
        x = "Date", y = "Nº of cases")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

Cases by Province



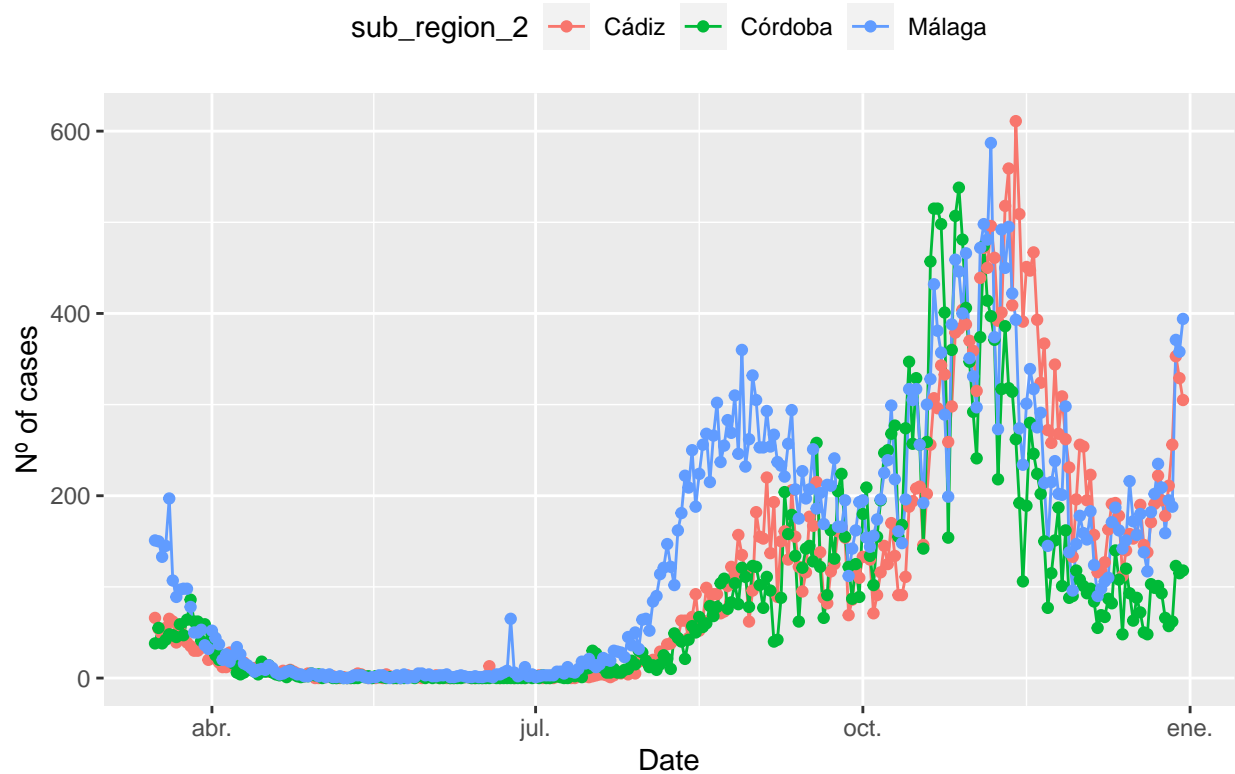
```
# All Total (mobility)
ggplot(Total, aes(x=fecha, y=Total, group=sub_region_2)) +
  geom_line(aes(linetype=sub_region_2, color=sub_region_2))+
  geom_point(aes(color=sub_region_2))+
  theme(legend.position="top") +
  labs(title="Mobility Change by Province",
        x = "Date", y = "% Mobility")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

Mobility Change by Province



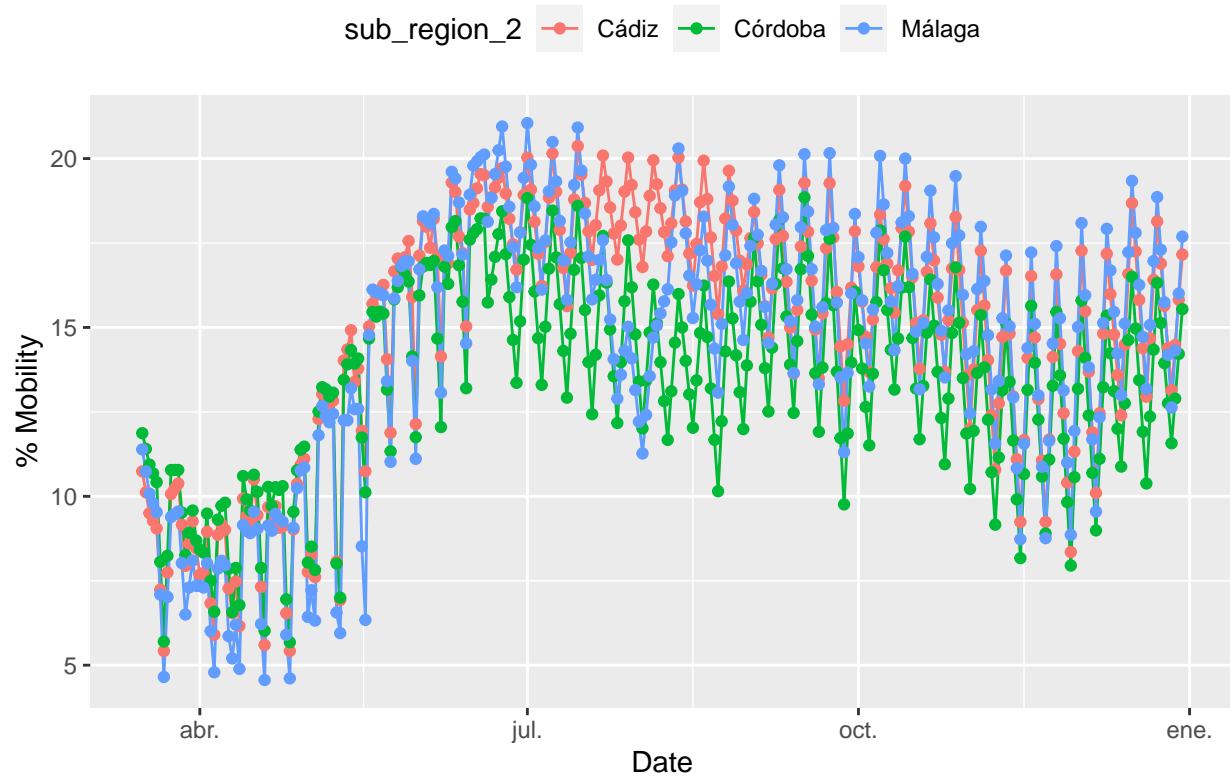
```
# Mal, Cor and Cad - num_casos.x
Total %>%
  filter(sub_region_2 == "Málaga" | sub_region_2 == "Cádiz" |
         sub_region_2 == "Córdoba") %>%
  ggplot(aes(x=fecha, y=num_casos.x))+
  geom_line(aes(color=sub_region_2))+
  geom_point(aes(color=sub_region_2))+
  theme(legend.position="top") +
  labs(title="Cases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

Cases by Province (Málaga, Córdoba and Cádiz)



```
# 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")
```

Mobility Change by Province (Málaga, Córdoba and Cádiz)



2.4 Time-series plots

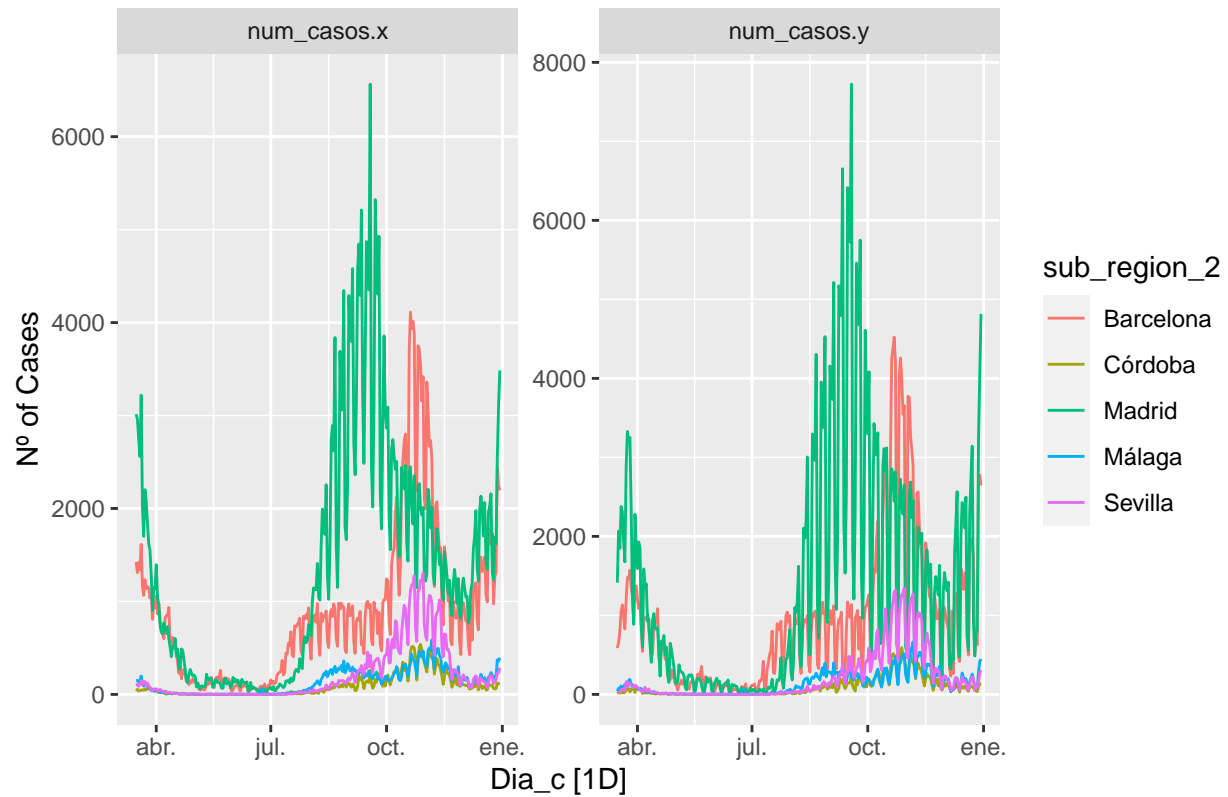
We have generated some plots from the **time-series** object generated.

```
# 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

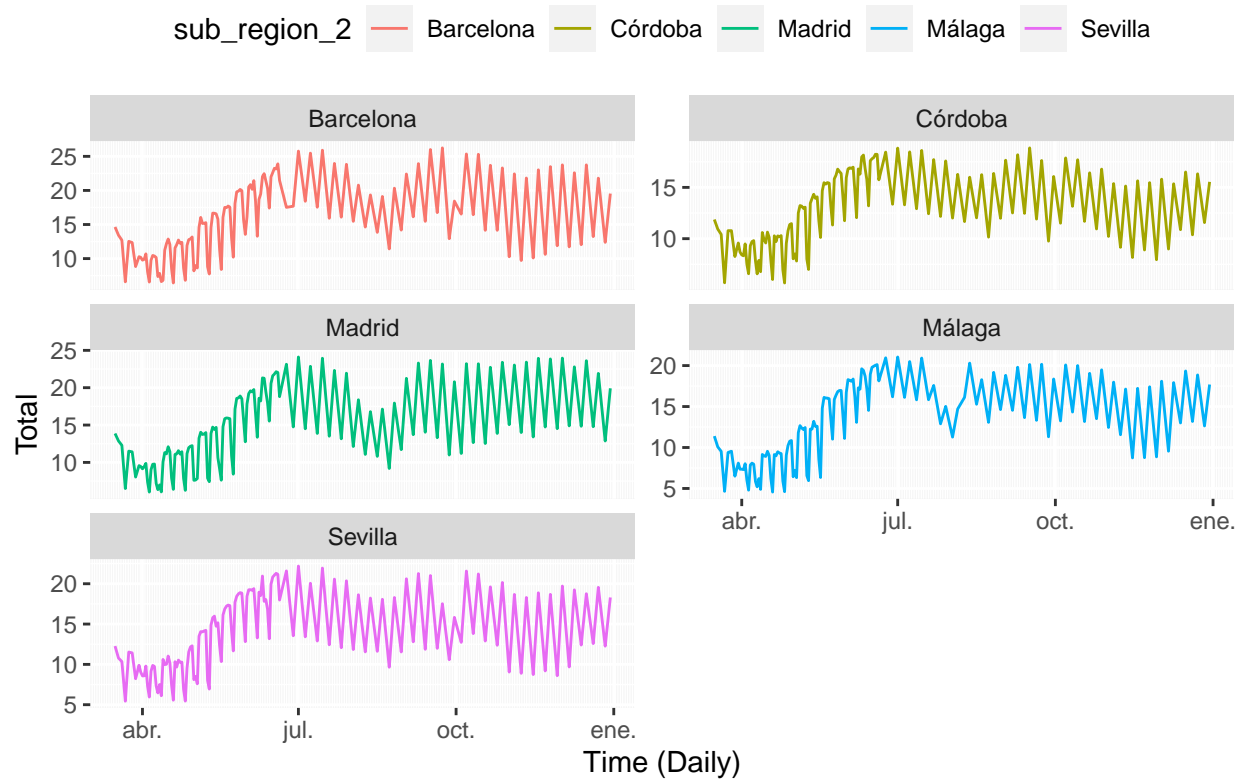
# Plots
# A num_casos.x,num_casos.y
autoplot(Total_ts_b, vars(num_casos.x,num_casos.y)) +
  labs(y = "Nº of Cases",
       title = "Reported Cases (CNE A vs B)")
```


Reported Cases (CNE A vs B)



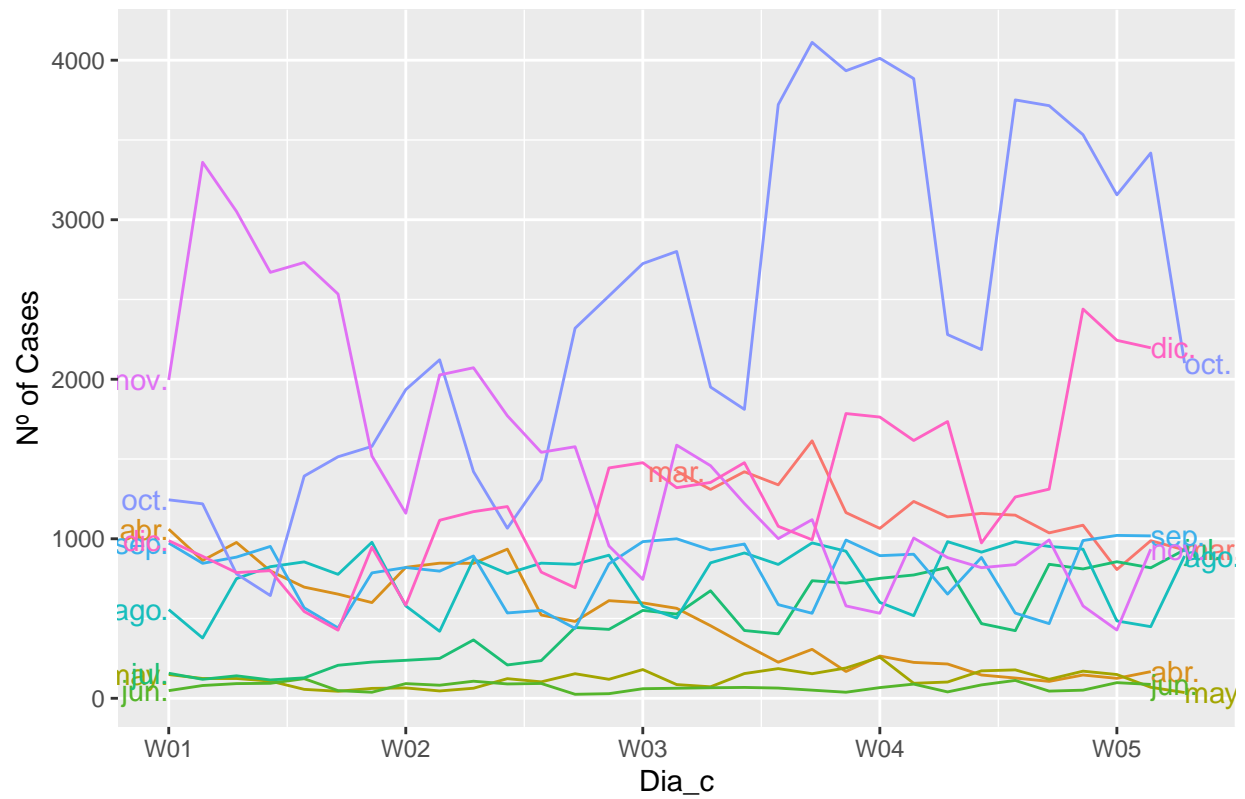
```
# B Total (mobility)
autoplot(Total_ts_b, Total) +
  facet_wrap(~sub_region_2, scales = "free_y", ncol=2) +
  theme(legend.position = "top") +
  scale_x_date(date_minor_breaks = "1 day", name = "Time (Daily)") +
  ggtitle(label = "Mobility Change by Province (Málaga, Córdoba and Cádiz)")
```

Mobility Change by Province (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")
```

Barcelona – Infections by Month



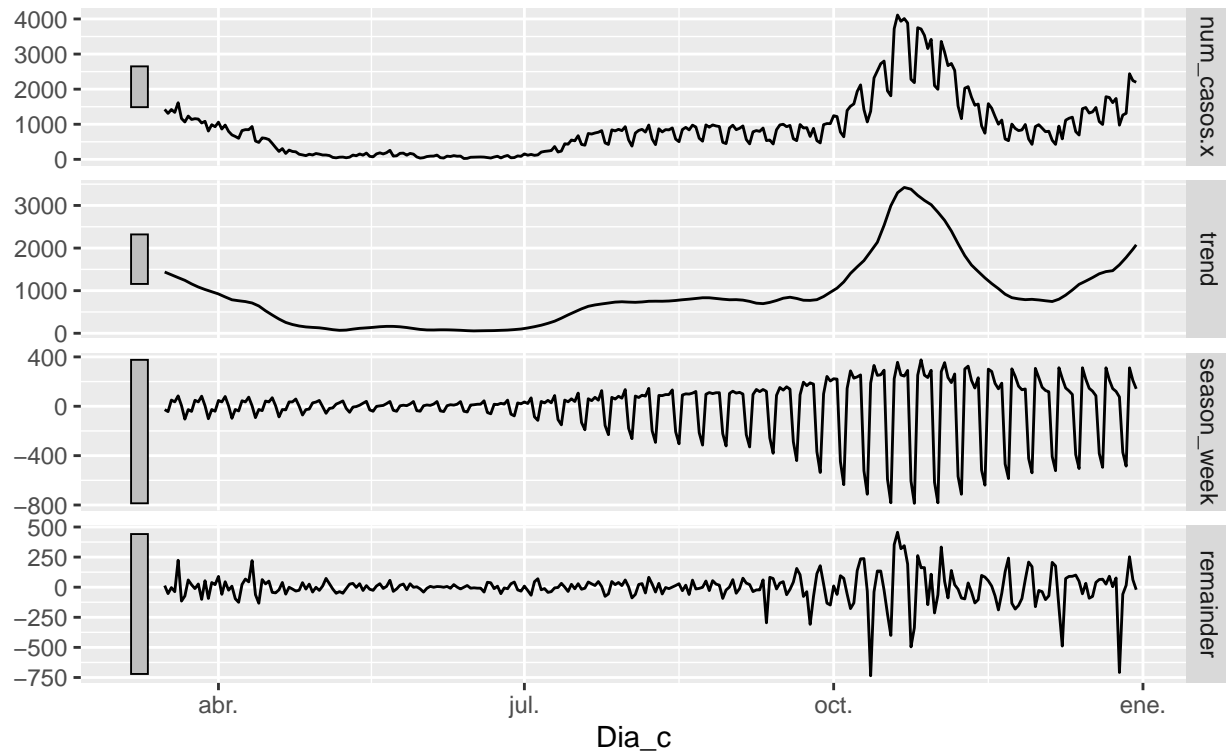
2.5 STL - Seasonal and Trend decomposition using Loess

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

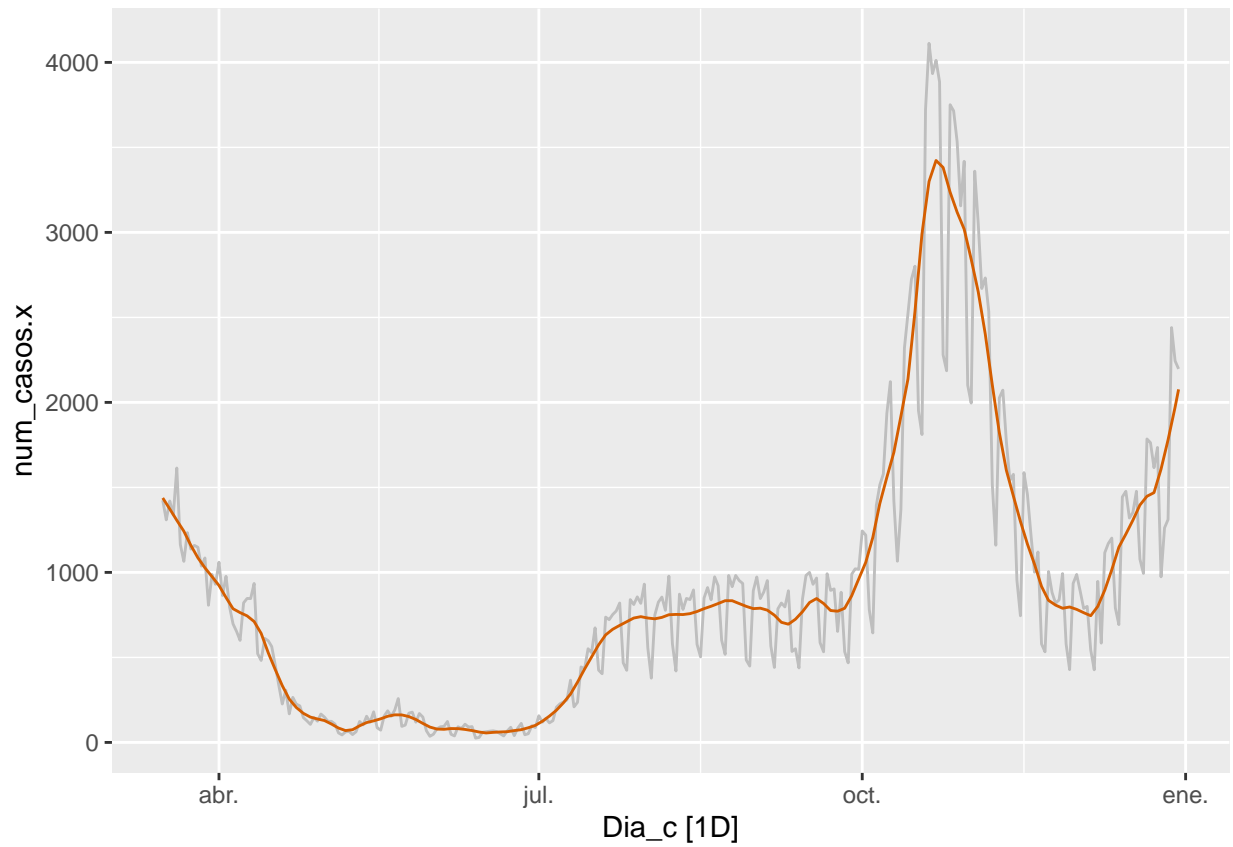
components(dcmp) %>% autoplot()
```

STL decomposition

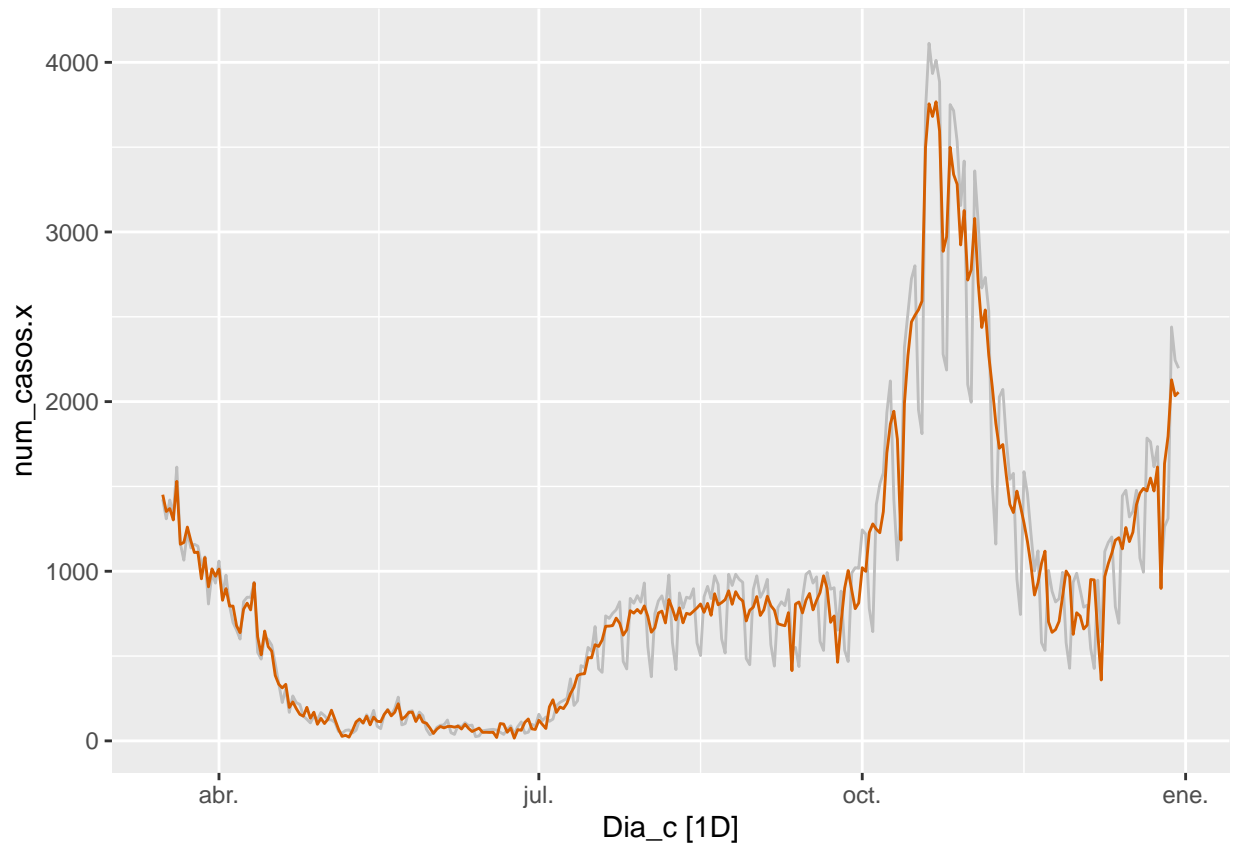
$\text{num_casos.x} = \text{trend} + \text{season_week} + \text{remainder}$



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



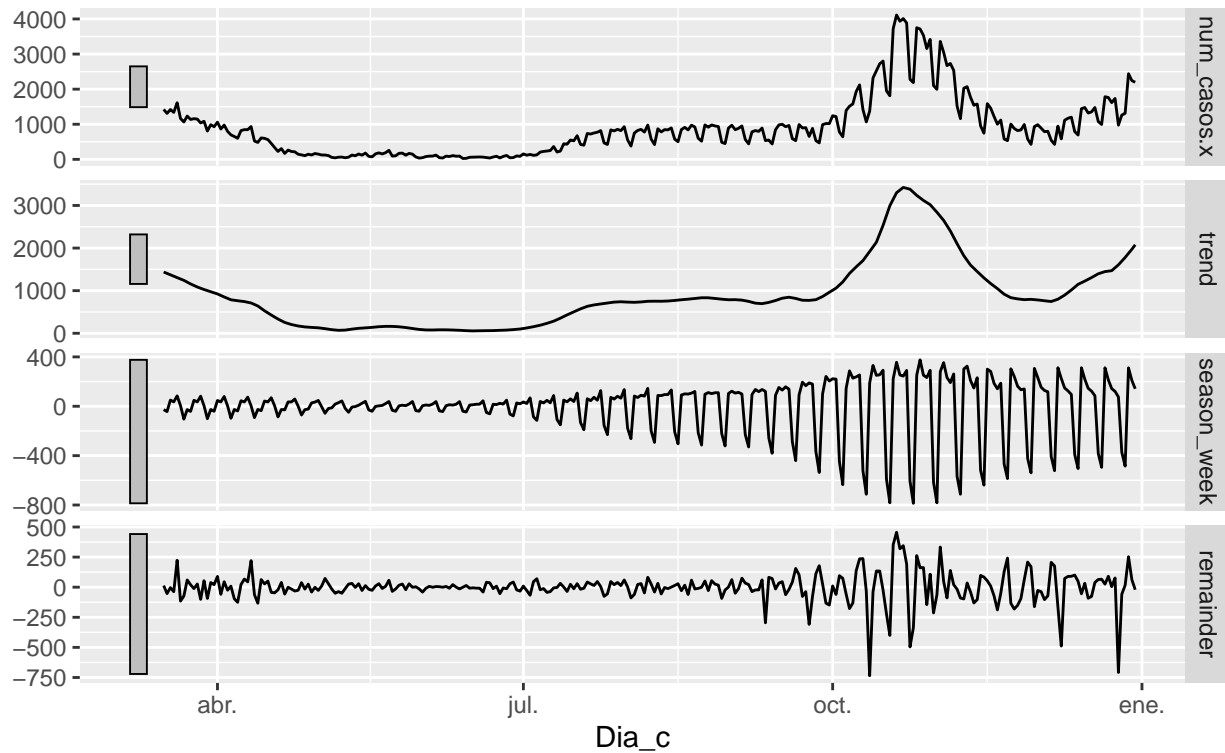
```
components(dcmp) %>%  
  as_tsibble() %>%  
  autoplot(num_casos.x, color="gray") +  
  geom_line(aes(y=season_adjust), color = "#D55E00")
```



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

STL decomposition

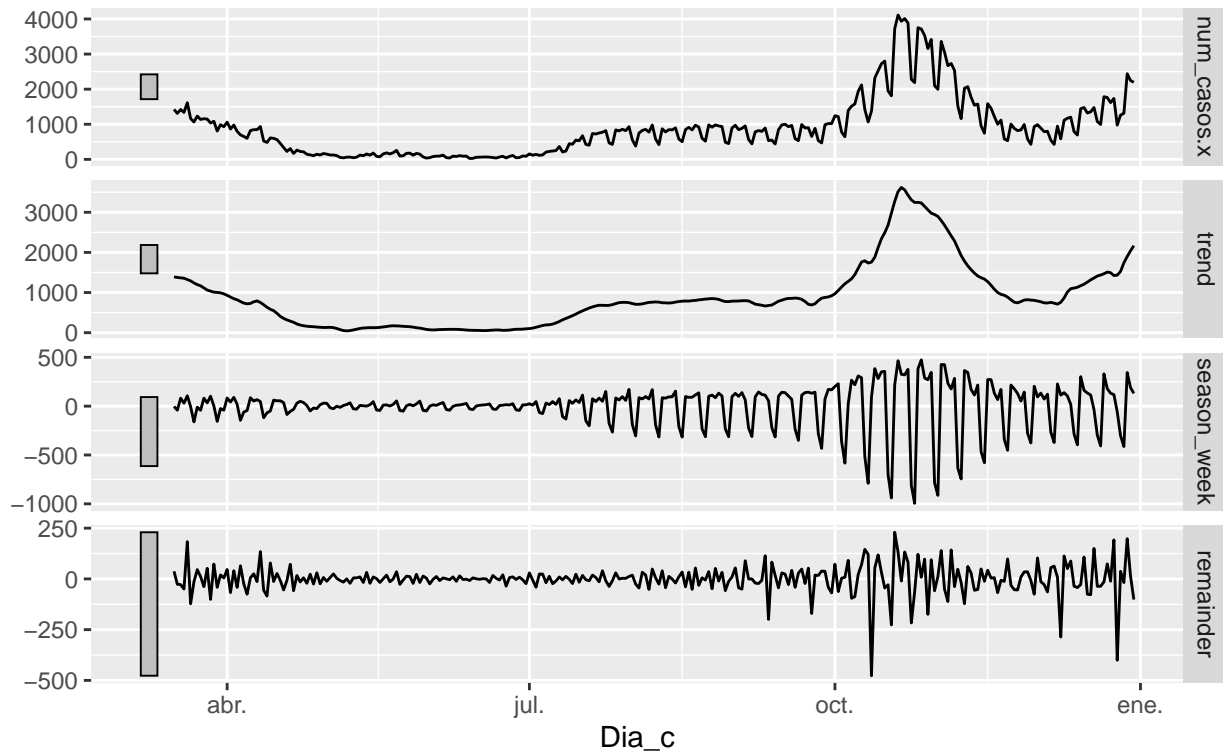
$\text{num_casos.x} = \text{trend} + \text{season_week} + \text{remainder}$



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

STL decomposition

$\text{num_casos.x} = \text{trend} + \text{season_week} + \text{remainder}$



2.6 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. *Análisis de Datos Y Estadística 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.