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

UOC - Alumno: Álvaro Rodríguez Sans

Mayo 2020 - Delivery 23/05/2020

Índex

1 Data load **3**

2 Descriptive statistics and visualization **3**
2.1 Type of data and modifications 3
2.1.1 EM3 4
2.1.2 Google 5
2.1.3 CNE 12
2.2 Datasets combinations 14
2.2.1 CNE_tec_cas 14
2.2.2 GOG_CNE 15
2.2.3 Total 16

Bibliography **19**

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

```
## Loading required package: stringr
```

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

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

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

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

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

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

```
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
```

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

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

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

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

```
if(!require(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
```

```
## The following object is masked from 'package:data.table':
##
##     %like%
```

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

1 Data load

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

```
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 Descriptive statistics and visualization

2.1 Type of data and modifications

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

2.1.1 EM3

```
# 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 Google

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

```
## 'data.frame':    24242 obs. of  15 variables:
##  $ country_region_code                                : chr  "ES" ...
##  $ country_region                                     : chr  "Spain" ...
##  $ sub_region_1                                       : chr  "" ...
##  $ sub_region_2                                       : chr  "" ...
##  $ metro_area                                         : logi  NA ...
##  $ iso_3166_2_code                                    : chr  "" ...
##  $ census_fips_code                                   : logi  NA ...
##  $ place_id                                           : chr  "ChIJi7xhMnjjQgwR7KNoB5Qs7KY" ...
##  $ date                                               : chr  "15/02/2020" ...
##  $ retail_and_recreation_percent_change_from_baseline: int  2 2 ...
##  $ grocery_and_pharmacy_percent_change_from_baseline : int  -1 3 ...
##  $ parks_percent_change_from_baseline                 : int  26 13 ...
##  $ transit_stations_percent_change_from_baseline      : int  8 5 ...
##  $ workplaces_percent_change_from_baseline            : int  0 -1 ...
##  $ residential_percent_change_from_baseline           : int  -2 -2 ...
```

```
## NULL
```

```
summary(Google)
```

```
##  country_region_code country_region     sub_region_1       sub_region_2
##  Length:24242        Length:24242       Length:24242       Length:24242
##  Class :character    Class :character   Class :character   Class :character
##  Mode  :character    Mode  :character   Mode  :character   Mode  :character
```

```
##
##
##
##
##  metro_area     iso_3166_2_code    census_fips_code   place_id
##  Mode:logical   Length:24242       Mode:logical     Length:24242
##  NA's:24242     Class :character   NA's:24242       Class :character
##                 Mode  :character                    Mode  :character
##
##
##
##
##      date           retail_and_recreation_percent_change_from_baseline
##  Length:24242       Min.   :-97.00
##  Class :character   1st Qu.:-53.00
##  Mode  :character   Median :-32.00
##                     Mean   :-36.42
##                     3rd Qu.:-17.00
##                     Max.   : 71.00
##                     NA's   :56
##  grocery_and_pharmacy_percent_change_from_baseline
##  Min.   :-96.000
##  1st Qu.:-18.000
##  Median : -4.000
##  Mean   : -9.973
##  3rd Qu.:  3.000
##  Max.   :194.000
##  NA's   :396
##  parks_percent_change_from_baseline
##  Min.   :-94.0000
##  1st Qu.:-30.0000
##  Median : -5.0000
##  Mean   : -0.0038
##  3rd Qu.: 22.0000
##  Max.   :543.0000
##  NA's   :305
##  transit_stations_percent_change_from_baseline
##  Min.   :-100.00
##  1st Qu.: -46.00
##  Median : -30.00
##  Mean   : -32.63
##  3rd Qu.: -16.00
##  Max.   : 177.00
##  NA's   :832
##  workplaces_percent_change_from_baseline
##  Min.   :-92.0
##  1st Qu.:-37.0
##  Median :-24.0
##  Mean   :-26.7
##  3rd Qu.:-13.0
##  Max.   : 55.0
##  NA's   :42
##  residential_percent_change_from_baseline
##  Min.   :-10.000
```

```
##  1st Qu.:  4.000
##  Median :  7.000
##  Mean   :  9.419
##  3rd Qu.: 13.000
##  Max.   : 48.000
##  NA's   :267
```

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

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

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

```
##
##                                  A CoruÃ±a                Ã\201lava
##                  7687                385                   385
##                Ã\201vila             Albacete              Alicante
##                   385                385                   385
##               AlmerÃa              Badajoz             Barcelona
##                   385                385                   385
##                Biscay              Burgos              CÃ¡ceres
##                   385                385                   385
##                CÃ¡diz              CÃ³rdoba            CastellÃ³n
##                   385                385                   385
##           Ciudad Real              Cuenca              Gipuzkoa
##                   385                385                   385
##               Girona              Granada            Guadalajara
##                   385                385                   385
##               Huelva              Huesca                JaÃ©n
##                   385                385                   385
##            Las Palmas              LeÃ³n               Lleida
##                   385                385                   385
##                 Lugo              MÃ¡laga              Palencia
##                   385                385                   385
##            Pontevedra  Province of Ourense            Salamanca
##                   385                385                   385
## Santa Cruz de Tenerife              Segovia              Seville
##                   385                385                   385
##                Soria             Tarragona               Teruel
##                   385                385                   385
##               Toledo              Valencia            Valladolid
##                   385                385                   385
##               Zamora             Zaragoza
##                   385                385
```

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

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

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

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

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

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

- AC - Melilla - C/Pr - iMelilla

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

Note The following link states the provinces in Spain INE CCAA. It is going to be used as table referencence.

```
# Modidication provinces - A
Google$sub_region_2[Google$sub_region_1=="Balearic Islands"] <- "Balears, Illes"
Google$sub_region_2[Google$sub_region_1=="Asturias"] <- "Asturias"
Google$sub_region_2[Google$sub_region_1=="Cantabria"] <- "Cantabria"
Google$sub_region_2[Google$sub_region_1=="Community of Madrid"] <- "Madrid"
Google$sub_region_2[Google$sub_region_1=="Region of Murcia"] <- "Murcia"
Google$sub_region_2[Google$sub_region_1=="Navarre"] <- "Navarra"
Google$sub_region_2[Google$sub_region_1=="La Rioja"] <- "Rioja, La"
Google$sub_region_2[Google$sub_region_1=="Ceuta"] <- "Ceuta"
Google$sub_region_2[Google$sub_region_1=="Melilla"] <- "Melilla"

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

```
Google$sub_region_2[Google$sub_region_2=="Navarra"]<-"Navarra"
Google$sub_region_2[Google$sub_region_2=="Palencia"]<-"Palencia"
Google$sub_region_2[Google$sub_region_2=="Pontevedra"]<-"Pontevedra"
Google$sub_region_2[Google$sub_region_2=="Province of Ourense"]<-"Ourense"
Google$sub_region_2[Google$sub_region_2=="Rioja, La"]<-"Rioja, La"
Google$sub_region_2[Google$sub_region_2=="Salamanca"]<-"Salamanca"
Google$sub_region_2[Google$sub_region_2=="Santa Cruz de Tenerife"]<-"Santa Cruz de Tenerife"
Google$sub_region_2[Google$sub_region_2=="Segovia"]<-"Segovia"
Google$sub_region_2[Google$sub_region_2=="Seville"]<-"Sevilla"
Google$sub_region_2[Google$sub_region_2=="Soria"]<-"Soria"
Google$sub_region_2[Google$sub_region_2=="Tarragona"]<-"Tarragona"
Google$sub_region_2[Google$sub_region_2=="Teruel"]<-"Teruel"
Google$sub_region_2[Google$sub_region_2=="Toledo"]<-"Toledo"
Google$sub_region_2[Google$sub_region_2=="Valencia"]<-"Valencia/València"
Google$sub_region_2[Google$sub_region_2=="Valladolid"]<-"Valladolid"
Google$sub_region_2[Google$sub_region_2=="Zamora"]<-"Zamora"
Google$sub_region_2[Google$sub_region_2=="Zaragoza"]<-"Zaragoza"
Google$sub_region_2 <- with(Google, ifelse(grepl("^Almer", sub_region_2),
                                           "Almería", sub_region_2))

table(Google$sub_region_2)
```

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

```
##                   385                   385                   385
##                Toledo     Valencia/València            Valladolid
##                   385                   385                   385
##                Zamora              Zaragoza
##                   385                   385
```

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

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

```
# 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)
head(Google,5)
```

```
##   sub_region_1 sub_region_2 iso_3166_2_code       date
## 1    Andalusia      Almería              AL 2020-02-15
## 2    Andalusia      Almería              AL 2020-02-16
## 3    Andalusia      Almería              AL 2020-02-17
## 4    Andalusia      Almería              AL 2020-02-18
## 5    Andalusia      Almería              AL 2020-02-19
##   retail_and_recreation_percent_change_from_baseline
## 1                                                  5
## 2                                                 -2
## 3                                                  0
## 4                                                 -3
## 5                                                 -1
##   grocery_and_pharmacy_percent_change_from_baseline
## 1                                                -3
## 2                                                 0
## 3                                                -2
## 4                                                -3
## 5                                                -3
##   parks_percent_change_from_baseline
## 1                                 40
## 2                                 -2
## 3                                  3
## 4                                 -2
## 5                                  3
##   transit_stations_percent_change_from_baseline
```

```
## 1                                          10
## 2                                           1
## 3                                           5
## 4                                           5
## 5                                           4
##   workplaces_percent_change_from_baseline
## 1                                       1
## 2                                       1
## 3                                       3
## 4                                       3
## 5                                       3
##   residential_percent_change_from_baseline
## 1                                       -2
## 2                                       -1
## 3                                       -1
## 4                                        0
## 5                                        0
```

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

2.1.3 CNE

```
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
```

```
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
```

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).

```
# 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))

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
```

2.2 Datasets combinations

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

2.2.1 CNE_tec_cas

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

```
# CNE_casos_g
CNE_casos_g = CNE_casos %>% group_by(provincia_iso, fecha) %>% summarise_at(vars(num_casos, num_hosp, nu
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=c("provincia_iso","fecha"))
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
```

2.2.2 GOG_CNE

- GOG_CNE -> CNE_tec_cas + Google

```
# New dataframe GOG_CNE
GOG_CNE<-merge(CNE_tec_cas, Google, by.x=c("provincia_iso","fecha"), by.y=c("iso_3166_2_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_1     sub_region_2
## 1 Valencian Community Alicante/Alacant
## 2 Valencian Community Alicante/Alacant
## 3 Valencian Community Alicante/Alacant
## 4 Valencian Community Alicante/Alacant
## 5 Valencian Community Alicante/Alacant
##   retail_and_recreation_percent_change_from_baseline
## 1                                                  3
## 2                                                 -2
## 3                                                  0
## 4                                                 -5
## 5                                                  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
```

2.2.3 Total

- Total -> GOG_CNE + EM3

```
# 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 2020-03-18            AB         114                  107
## 3     Albacete 2020-03-20            AB         131                  121
## 4     Albacete 2020-03-22            AB         125                  112
## 5     Albacete 2020-03-24            AB         107                   91
##   num_casos_prueba_test_ac num_casos_prueba_ag num_casos_prueba_elisa
## 1                        5                   0                     0
## 2                        7                   0                     0
## 3                       10                   0                     0
## 4                       13                   0                     0
## 5                       16                   0                     0
##   num_casos_prueba_desconocida num_casos.y num_hosp num_uci num_def
## 1                            0          65       43       3       7
## 2                            0          26       24       7       7
## 3                            0          85       63       4       6
## 4                            0          60       61       8      13
## 5                            0          53       76       7      14
##        sub_region_1 retail_and_recreation_percent_change_from_baseline
## 1 Castile-La Mancha                                               -81
## 2 Castile-La Mancha                                               -83
```

```
## 3 Castile-La Mancha                                                    -87
## 4 Castile-La Mancha                                                    -94
## 5 Castile-La Mancha                                                    -87
##   grocery_and_pharmacy_percent_change_from_baseline
## 1                                               -32
## 2                                               -32
## 3                                               -34
## 4                                               -80
## 5                                               -55
##   parks_percent_change_from_baseline
## 1                                -73
## 2                                -70
## 3                                -74
## 4                                -88
## 5                                -81
##   transit_stations_percent_change_from_baseline
## 1                                           -66
## 2                                           -70
## 3                                           -76
## 4                                           -89
## 5                                           -79
##   workplaces_percent_change_from_baseline
## 1                                     -51
## 2                                     -58
## 3                                     -68
## 4                                     -66
## 5                                     -64
##   residential_percent_change_from_baseline Total
## 1                                       22  9.90
## 2                                       23  9.51
## 3                                       32  8.75
## 4                                       23  4.50
## 5                                       28  9.02
```

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

```
## 'data.frame':    6663 obs. of  21 variables:
##  $ sub_region_2                                       : chr  "Albacete" ...
##  $ fecha                                              : Date, format: "2020-03-16" ...
##  $ provincia_iso                                      : chr  "AB" ...
##  $ num_casos.x                                        : int  137 114 ...
##  $ num_casos_prueba_pcr                               : int  132 107 ...
##  $ num_casos_prueba_test_ac                           : int  5 7 ...
##  $ 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 26 ...
##  $ num_hosp                                           : int  43 24 ...
##  $ num_uci                                            : int  3 7 ...
##  $ num_def                                            : int  7 7 ...
##  $ sub_region_1                                       : chr  "Castile-La Mancha" ...
##  $ retail_and_recreation_percent_change_from_baseline: int  -81 -83 ...
##  $ grocery_and_pharmacy_percent_change_from_baseline : int  -32 -32 ...
##  $ parks_percent_change_from_baseline                 : int  -73 -70 ...
##  $ transit_stations_percent_change_from_baseline      : int  -66 -70 ...
```

```
##  $ workplaces_percent_change_from_baseline           : int  -51 -58 ...
##  $ residential_percent_change_from_baseline          : int  22 23 ...
##  $ Total                                             : num  9.9 9.51 ...
```

```
## NULL
```

```
summary(Total)
```

```
##  sub_region_2           fecha            provincia_iso       num_casos.x
##  Length:6663        Min.   :2020-03-16   Length:6663        Min.   :   0.00
##  Class :character   1st Qu.:2020-04-28   Class :character   1st Qu.:   2.00
##  Mode  :character   Median :2020-06-03   Mode  :character   Median :  13.00
##                     Mean   :2020-06-29                      Mean   :  64.23
##                     3rd Qu.:2020-08-26                      3rd Qu.:  62.00
##                     Max.   :2020-12-30                      Max.   :3934.00
##
##  num_casos_prueba_pcr num_casos_prueba_test_ac num_casos_prueba_ag
##  Min.   :   0.0       Min.   : 0.0000          Min.   :  0.000
##  1st Qu.:   2.0       1st Qu.: 0.0000          1st Qu.:  0.000
##  Median :  11.0       Median : 0.0000          Median :  0.000
##  Mean   :  57.3       Mean   : 0.4939          Mean   :  5.988
##  3rd Qu.:  54.0       3rd Qu.: 0.0000          3rd Qu.:  0.000
##  Max.   :3933.0       Max.   :29.0000          Max.   :452.000
##
##  num_casos_prueba_elisa num_casos_prueba_desconocida  num_casos.y
##  Min.   : 0.0000        Min.   : 0.0000              Min.   :   0.00
##  1st Qu.: 0.0000        1st Qu.: 0.0000              1st Qu.:   3.00
##  Median : 0.0000        Median : 0.0000              Median :  16.00
##  Mean   : 0.2611        Mean   : 0.1962              Mean   :  65.16
##  3rd Qu.: 0.0000        3rd Qu.: 0.0000              3rd Qu.:  62.00
##  Max.   :71.0000        Max.   :65.0000              Max.   :4370.00
##
##     num_hosp          num_uci           num_def        sub_region_1
##  Min.   :  0.000   Min.   : 0.0000   Min.   :  0.000   Length:6663
##  1st Qu.:  0.000   1st Qu.: 0.0000   1st Qu.:  0.000   Class :character
##  Median :  2.000   Median : 0.0000   Median :  1.000   Mode  :character
##  Mean   :  9.862   Mean   : 0.8553   Mean   :  3.226
##  3rd Qu.:  9.000   3rd Qu.: 1.0000   3rd Qu.:  3.000
##  Max.   :805.000   Max.   :78.0000   Max.   :189.000
##
##  retail_and_recreation_percent_change_from_baseline
##  Min.   :-97.0
##  1st Qu.:-84.0
##  Median :-47.0
##  Mean   :-51.2
##  3rd Qu.:-26.0
##  Max.   : 59.0
##  NA's   :14
##  grocery_and_pharmacy_percent_change_from_baseline
##  Min.   :-96.00
##  1st Qu.:-43.00
##  Median :-17.00
##  Mean   :-20.31
##  3rd Qu.: -2.00
##  Max.   :194.00
```

```
##  NA's   :114
##  parks_percent_change_from_baseline
##  Min.   :-94.00
##  1st Qu.:-62.00
##  Median :-17.00
##  Mean   :-15.63
##  3rd Qu.: 12.00
##  Max.   :486.00
##  NA's   :62
##  transit_stations_percent_change_from_baseline
##  Min.   :-100.00
##  1st Qu.: -74.00
##  Median : -45.00
##  Mean   : -47.72
##  3rd Qu.: -28.00
##  Max.   :  74.00
##  NA's   :258
##  workplaces_percent_change_from_baseline
##  Min.   :-92.00
##  1st Qu.:-57.00
##  Median :-34.00
##  Mean   :-34.78
##  3rd Qu.:-16.00
##  Max.   : 55.00
##  NA's   :10
##  residential_percent_change_from_baseline     Total
##  Min.   :-10.00                           Min.   : 1.95
##  1st Qu.:  6.00                           1st Qu.: 9.54
##  Median : 11.00                           Median :12.83
##  Mean   : 13.64                           Mean   :13.08
##  3rd Qu.: 23.00                           3rd Qu.:16.53
##  Max.   : 48.00                           Max.   :29.00
##  NA's   :187
```

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.

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

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

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