Cluster

DATA CLEANING

DATASET1

```
##
# Data Cleaning
##

rm(list=ls())

### Paths ####
pd1 = "../Data/Dataset1.- DatosConsumoAlimentarioMAPAporCCAA.txt" #Consumo
pd2 = "../Data/Dataset2.- Precios Semanales Observatorio de Precios Junta de Andalucia.txt" #Precio

pd4 = "../Data/Dataset4.- Comercio Exterior de EspaÃ±a.txt"
pd5 = "../Data/Dataset5_Coronavirus_cases.txt" #Covid

### Libraries ####
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

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

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract
```

```r
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
### Datasets ####

#### 1.Consumo ####
data1 = read.csv(pd1, sep = "|", dec = ",")
summary(data1) # 120 NA's en penetración
```

```
##     ï..AÃ.o         Mes                CCAA             Producto
##  Min.   :2018   Length:26634       Length:26634       Length:26634
##  1st Qu.:2018   Class :character   Class :character   Class :character
##  Median :2019   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2019
##  3rd Qu.:2019
##  Max.   :2020
##
##  Volumen..miles.de.kg. Valor..miles.de.â... Precio.medio.kg  PenetraciÃ³n....
##  Min.   :     0.0      Min.   :     0.0     Min.   : 0.000   Min.   : 0.00
##  1st Qu.:   103.8      1st Qu.:   227.2     1st Qu.: 1.390   1st Qu.:14.83
##  Median :   404.6      Median :   824.6     Median : 1.790   Median :34.58
##  Mean   :  3435.7      Mean   :  5607.8     Mean   : 2.166   Mean   :38.00
##  3rd Qu.:  1536.2      3rd Qu.:  2731.4     3rd Qu.: 2.680   3rd Qu.:58.66
##  Max.   :451789.9      Max.   :824313.8     Max.   :38.750   Max.   :99.93
##                                                              NA's   :120
##  Consumo.per.capita Gasto.per.capita    X             X.1
##  Min.   : 0.0000    Min.   : 0.000   Mode:logical   Mode:logical
##  1st Qu.: 0.0700    1st Qu.: 0.150   NA's:26634     NA's:26634
##  Median : 0.1900    Median : 0.420
##  Mean   : 0.6391    Mean   : 1.045
##  3rd Qu.: 0.5400    3rd Qu.: 1.010
##  Max.   :14.2900    Max.   :27.630
##
```

```r
# MUCHOS VALORES 0

data1 %<>% select(c(Ano = ï..AÃ.o, Mes, CCAA, Producto,
                    Volumen = Volumen..miles.de.kg., Valor = Valor..miles.de.â... ,
                    Precio_Medio = Precio.medio.kg, Penetracion = `PenetraciÃ³n....`,
                    Cons_cpt = Consumo.per.capita, Gasto_cpt = Gasto.per.capita))

summary(data1)
```

```
##       Ano            Mes                CCAA             Producto
##  Min.   :2018   Length:26634       Length:26634       Length:26634
```

```
##  1st Qu.:2018   Class :character   Class :character   Class :character
##  Median :2019   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2019
##  3rd Qu.:2019
##  Max.   :2020
##
##     Volumen             Valor           Precio_Medio     Penetracion
##  Min.   :     0.0   Min.   :     0.0   Min.   : 0.000   Min.   : 0.00
##  1st Qu.:   103.8   1st Qu.:   227.2   1st Qu.: 1.390   1st Qu.:14.83
##  Median :   404.6   Median :   824.6   Median : 1.790   Median :34.58
##  Mean   :  3435.7   Mean   :  5607.8   Mean   : 2.166   Mean   :38.00
##  3rd Qu.:  1536.2   3rd Qu.:  2731.4   3rd Qu.: 2.680   3rd Qu.:58.66
##  Max.   :451789.9   Max.   :824313.8   Max.   :38.750   Max.   :99.93
##                                                         NA's   :120
##     Cons_cpt         Gasto_cpt
##  Min.   : 0.0000   Min.   : 0.000
##  1st Qu.: 0.0700   1st Qu.: 0.150
##  Median : 0.1900   Median : 0.420
##  Mean   : 0.6391   Mean   : 1.045
##  3rd Qu.: 0.5400   3rd Qu.: 1.010
##  Max.   :14.2900   Max.   :27.630
##
```

```
data1 %>%
  filter(is.na(Penetracion))
```

```
##      Ano        Mes           CCAA             Producto   Volumen     Valor
## 1   2020      Julio Total Nacional              TOMATES  74111.99 120124.20
## 2   2020      Julio Total Nacional             CEBOLLAS  28290.45  35155.54
## 3   2020      Julio Total Nacional LECHUGA/ESC./ENDIVIA  17223.53  47468.82
## 4   2020      Julio Total Nacional            PIMIENTOS  22940.50  44005.62
## 5   2020      Julio Total Nacional        JUDIAS VERDES   9266.35  31458.49
## 6   2020      Julio Total Nacional                COLES   3315.25   3749.59
## 7   2020      Julio Total Nacional             NARANJAS  33012.21  46053.61
## 8   2020      Julio Total Nacional           MANDARINAS    718.23   1521.53
## 9   2020      Julio Total Nacional              LIMONES   9894.60  18370.43
## 10  2020      Julio Total Nacional             PLATANOS  42572.49  67823.14
## 11  2020      Julio Total Nacional             MANZANAS  25733.46  41543.94
## 12  2020      Julio Total Nacional                PERAS  10791.59  20791.84
## 13  2020      Julio Total Nacional          MELOCOTONES  34279.57  65720.55
## 14  2020      Julio Total Nacional         ALBARICOQUES   6350.30  14606.30
## 15  2020      Julio Total Nacional        FRESAS/FRESON   1885.53   7824.46
## 16  2020      Julio Total Nacional                MELON  75112.25  84902.47
## 17  2020      Julio Total Nacional               SANDIA 121701.30  90727.77
## 18  2020      Julio Total Nacional             CIRUELAS  10601.36  21042.86
## 19  2020      Julio Total Nacional              CEREZAS   9249.03  47511.59
## 20  2020      Julio Total Nacional                 UVAS   3613.12  11736.25
## 21  2020      Julio Total Nacional                 KIWI   9410.67  34961.04
## 22  2020      Julio Total Nacional      PATATAS FRESCAS  90175.98  79025.45
## 23  2020      Julio Total Nacional T.HORTALIZAS FRESCAS 242399.04 448389.61
## 24  2020      Julio Total Nacional     T.FRUTAS FRESCAS 451789.94 702647.20
## 25  2020     Agosto Total Nacional              TOMATES  64452.71 105649.30
## 26  2020     Agosto Total Nacional             CEBOLLAS  26491.05  33939.18
## 27  2020     Agosto Total Nacional LECHUGA/ESC./ENDIVIA  15760.23  41983.69
```

```
## 28  2020     Agosto Total Nacional            PIMIENTOS  23015.39  44520.41
## 29  2020     Agosto Total Nacional        JUDIAS VERDES   6575.16  26406.77
## 30  2020     Agosto Total Nacional               COLES   3460.95   3770.14
## 31  2020     Agosto Total Nacional             NARANJAS  22936.44  39001.81
## 32  2020     Agosto Total Nacional           MANDARINAS   1499.96   3048.43
## 33  2020     Agosto Total Nacional              LIMONES   9354.49  18078.57
## 34  2020     Agosto Total Nacional             PLATANOS  36262.03  58853.52
## 35  2020     Agosto Total Nacional             MANZANAS  25307.80  40563.22
## 36  2020     Agosto Total Nacional                PERAS  13387.91  24549.24
## 37  2020     Agosto Total Nacional          MELOCOTONES  31865.50  59874.19
## 38  2020     Agosto Total Nacional         ALBARICOQUES   1597.27   3676.89
## 39  2020     Agosto Total Nacional        FRESAS/FRESON   1312.50   6074.11
## 40  2020     Agosto Total Nacional               MELON  81329.97  80985.69
## 41  2020     Agosto Total Nacional              SANDIA  99670.37  66535.96
## 42  2020     Agosto Total Nacional             CIRUELAS  11904.57  24050.41
## 43  2020     Agosto Total Nacional             CEREZAS    931.26   5250.70
## 44  2020     Agosto Total Nacional                UVAS   7799.39  20798.26
## 45  2020     Agosto Total Nacional                KIWI   7943.37  30698.20
## 46  2020     Agosto Total Nacional      PATATAS FRESCAS  74085.89  66028.82
## 47  2020     Agosto Total Nacional T.HORTALIZAS FRESCAS 217270.54 409019.27
## 48  2020     Agosto Total Nacional     T.FRUTAS FRESCAS 401113.71 594976.88
## 49  2020 Septiembre Total Nacional              TOMATES  64894.57 115161.00
## 50  2020 Septiembre Total Nacional             CEBOLLAS  29660.96  36335.52
## 51  2020 Septiembre Total Nacional LECHUGA/ESC./ENDIVIA  15879.52  43313.25
## 52  2020 Septiembre Total Nacional            PIMIENTOS  27633.09  59697.46
## 53  2020 Septiembre Total Nacional        JUDIAS VERDES   6688.37  25314.37
## 54  2020 Septiembre Total Nacional               COLES   5498.32   6273.87
## 55  2020 Septiembre Total Nacional             NARANJAS  26007.96  47716.38
## 56  2020 Septiembre Total Nacional           MANDARINAS   5970.46  12196.50
## 57  2020 Septiembre Total Nacional              LIMONES   9375.09  18656.68
## 58  2020 Septiembre Total Nacional             PLATANOS  44651.93  73126.22
## 59  2020 Septiembre Total Nacional             MANZANAS  36432.66  58578.08
## 60  2020 Septiembre Total Nacional                PERAS  20333.21  34857.81
## 61  2020 Septiembre Total Nacional          MELOCOTONES  25601.73  55878.90
## 62  2020 Septiembre Total Nacional         ALBARICOQUES    290.57    528.15
## 63  2020 Septiembre Total Nacional        FRESAS/FRESON   1381.92   5657.11
## 64  2020 Septiembre Total Nacional               MELON  77570.51  72800.62
## 65  2020 Septiembre Total Nacional              SANDIA  51830.72  45308.76
## 66  2020 Septiembre Total Nacional             CIRUELAS  12767.94  24875.10
## 67  2020 Septiembre Total Nacional             CEREZAS    402.31   1152.25
## 68  2020 Septiembre Total Nacional                UVAS  21201.95  52061.69
## 69  2020 Septiembre Total Nacional                KIWI   8909.03  35924.77
## 70  2020 Septiembre Total Nacional      PATATAS FRESCAS  84453.33  75146.30
## 71  2020 Septiembre Total Nacional T.HORTALIZAS FRESCAS 244130.14 477824.67
## 72  2020 Septiembre Total Nacional     T.FRUTAS FRESCAS 385879.27 652149.54
## 73  2020    Octubre Total Nacional              TOMATES  52238.00  96244.33
## 74  2020    Octubre Total Nacional             CEBOLLAS  32327.00  40006.25
## 75  2020    Octubre Total Nacional LECHUGA/ESC./ENDIVIA  16010.76  45169.75
## 76  2020    Octubre Total Nacional            PIMIENTOS  26486.91  50703.32
## 77  2020    Octubre Total Nacional        JUDIAS VERDES   8567.84  29662.35
## 78  2020    Octubre Total Nacional               COLES   8010.26   8841.25
## 79  2020    Octubre Total Nacional             NARANJAS  39240.26  64232.30
## 80  2020    Octubre Total Nacional           MANDARINAS  32852.15  56137.96
## 81  2020    Octubre Total Nacional              LIMONES  11153.38  20670.52
```

```
## 82  2020    Octubre Total Nacional             PLATANOS  55264.94  93993.44
## 83  2020    Octubre Total Nacional             MANZANAS  51560.58  80253.44
## 84  2020    Octubre Total Nacional                PERAS  26084.68  44372.81
## 85  2020    Octubre Total Nacional          MELOCOTONES  13165.82  32898.48
## 86  2020    Octubre Total Nacional         ALBARICOQUES    155.39    205.96
## 87  2020    Octubre Total Nacional        FRESAS/FRESON   1218.29   4844.54
## 88  2020    Octubre Total Nacional                MELON  39737.51  44238.72
## 89  2020    Octubre Total Nacional              SANDIA   9421.45   9490.50
## 90  2020    Octubre Total Nacional             CIRUELAS   7142.77  14346.21
## 91  2020    Octubre Total Nacional              CEREZAS    220.04    632.40
## 92  2020    Octubre Total Nacional                UVAS  25329.51  60093.95
## 93  2020    Octubre Total Nacional                KIWI  12491.20  45432.94
## 94  2020    Octubre Total Nacional      PATATAS FRESCAS  95916.00  83644.14
## 95  2020    Octubre Total Nacional T.HORTALIZAS FRESCAS 251090.91 496897.93
## 96  2020    Octubre Total Nacional     T.FRUTAS FRESCAS 379017.20 712633.15
## 97  2020  Noviembre Total Nacional             TOMATES  48644.85  87551.48
## 98  2020  Noviembre Total Nacional             CEBOLLAS  29522.24  36553.32
## 99  2020  Noviembre Total Nacional LECHUGA/ESC./ENDIVIA  15583.43  41557.28
## 100 2020  Noviembre Total Nacional            PIMIENTOS  19530.12  37862.19
## 101 2020  Noviembre Total Nacional        JUDIAS VERDES   6589.62  21713.50
## 102 2020  Noviembre Total Nacional                COLES   8192.86   9311.09
## 103 2020  Noviembre Total Nacional             NARANJAS  60360.94  70234.89
## 104 2020  Noviembre Total Nacional           MANDARINAS  49106.58  72536.43
## 105 2020  Noviembre Total Nacional              LIMONES   9798.40  16479.61
## 106 2020  Noviembre Total Nacional             PLATANOS  54888.95  88693.86
## 107 2020  Noviembre Total Nacional             MANZANAS  43081.70  67602.74
## 108 2020  Noviembre Total Nacional                PERAS  23814.58  40425.63
## 109 2020  Noviembre Total Nacional          MELOCOTONES   1740.22   4549.97
## 110 2020  Noviembre Total Nacional         ALBARICOQUES     85.10    145.17
## 111 2020  Noviembre Total Nacional        FRESAS/FRESON    938.76   4223.57
## 112 2020  Noviembre Total Nacional                MELON  12929.44  19804.29
## 113 2020  Noviembre Total Nacional              SANDIA    637.46    860.28
## 114 2020  Noviembre Total Nacional             CIRUELAS   2198.18   4372.42
## 115 2020  Noviembre Total Nacional              CEREZAS     95.29    538.92
## 116 2020  Noviembre Total Nacional                UVAS  15918.49  42755.62
## 117 2020  Noviembre Total Nacional                KIWI  12253.15  41949.06
## 118 2020  Noviembre Total Nacional      PATATAS FRESCAS  89679.98  76899.34
## 119 2020  Noviembre Total Nacional T.HORTALIZAS FRESCAS 238554.47 472592.86
## 120 2020  Noviembre Total Nacional     T.FRUTAS FRESCAS 341351.65 597092.66
##     Precio_Medio Penetracion Cons_cpt Gasto_cpt
## 1           1.62          NA     1.67      2.71
## 2           1.24          NA     0.64      0.79
## 3           2.76          NA     0.39      1.07
## 4           1.92          NA     0.52      0.99
## 5           3.39          NA     0.21      0.71
## 6           1.13          NA     0.07      0.08
## 7           1.40          NA     0.74      1.04
## 8           2.12          NA     0.02      0.03
## 9           1.86          NA     0.22      0.41
## 10          1.59          NA     0.96      1.53
## 11          1.61          NA     0.58      0.94
## 12          1.93          NA     0.24      0.47
## 13          1.92          NA     0.77      1.48
## 14          2.30          NA     0.14      0.33
```

```
## 15          4.15          NA     0.04      0.18
## 16          1.13          NA     1.69      1.92
## 17          0.75          NA     2.74      2.05
## 18          1.98          NA     0.24      0.47
## 19          5.14          NA     0.21      1.07
## 20          3.25          NA     0.08      0.26
## 21          3.72          NA     0.21      0.79
## 22          0.88          NA     2.03      1.78
## 23          1.85          NA     5.46     10.11
## 24          1.56          NA    10.18     15.85
## 25          1.64          NA     1.45      2.38
## 26          1.28          NA     0.60      0.77
## 27          2.66          NA     0.36      0.95
## 28          1.93          NA     0.52      1.00
## 29          4.02          NA     0.15      0.60
## 30          1.09          NA     0.08      0.09
## 31          1.70          NA     0.52      0.88
## 32          2.03          NA     0.03      0.07
## 33          1.93          NA     0.21      0.41
## 34          1.62          NA     0.82      1.33
## 35          1.60          NA     0.57      0.92
## 36          1.83          NA     0.30      0.55
## 37          1.88          NA     0.72      1.35
## 38          2.30          NA     0.04      0.08
## 39          4.63          NA     0.03      0.14
## 40          1.00          NA     1.83      1.83
## 41          0.67          NA     2.25      1.50
## 42          2.02          NA     0.27      0.54
## 43          5.64          NA     0.02      0.12
## 44          2.67          NA     0.18      0.47
## 45          3.86          NA     0.18      0.69
## 46          0.89          NA     1.67      1.49
## 47          1.88          NA     4.90      9.23
## 48          1.48          NA     9.04     13.42
## 49          1.77          NA     1.46      2.60
## 50          1.23          NA     0.67      0.82
## 51          2.73          NA     0.36      0.98
## 52          2.16          NA     0.62      1.35
## 53          3.78          NA     0.15      0.57
## 54          1.14          NA     0.12      0.14
## 55          1.83          NA     0.59      1.08
## 56          2.04          NA     0.13      0.28
## 57          1.99          NA     0.21      0.42
## 58          1.64          NA     1.01      1.65
## 59          1.61          NA     0.82      1.32
## 60          1.71          NA     0.46      0.79
## 61          2.18          NA     0.58      1.26
## 62          1.82          NA     0.01      0.01
## 63          4.09          NA     0.03      0.13
## 64          0.94          NA     1.75      1.64
## 65          0.87          NA     1.17      1.02
## 66          1.95          NA     0.29      0.56
## 67          2.86          NA     0.01      0.03
## 68          2.46          NA     0.48      1.17
```

```
## 69          4.03          NA     0.20      0.81
## 70          0.89          NA     1.90      1.70
## 71          1.96          NA     5.50     10.78
## 72          1.69          NA     8.70     14.71
## 73          1.84          NA     1.18      2.17
## 74          1.24          NA     0.73      0.90
## 75          2.82          NA     0.36      1.02
## 76          1.91          NA     0.60      1.14
## 77          3.46          NA     0.19      0.67
## 78          1.10          NA     0.18      0.20
## 79          1.64          NA     0.88      1.45
## 80          1.71          NA     0.74      1.27
## 81          1.85          NA     0.25      0.47
## 82          1.70          NA     1.25      2.12
## 83          1.56          NA     1.16      1.81
## 84          1.70          NA     0.59      1.00
## 85          2.50          NA     0.30      0.74
## 86          1.33          NA     0.00      0.00
## 87          3.98          NA     0.03      0.11
## 88          1.11          NA     0.90      1.00
## 89          1.01          NA     0.21      0.21
## 90          2.01          NA     0.16      0.32
## 91          2.87          NA     0.00      0.01
## 92          2.37          NA     0.57      1.36
## 93          3.64          NA     0.28      1.02
## 94          0.87          NA     2.16      1.89
## 95          1.98          NA     5.66     11.21
## 96          1.88          NA     8.54     16.08
## 97          1.80          NA     1.10      1.97
## 98          1.24          NA     0.67      0.82
## 99          2.67          NA     0.35      0.94
## 100         1.94          NA     0.44      0.85
## 101         3.30          NA     0.15      0.49
## 102         1.14          NA     0.18      0.21
## 103         1.16          NA     1.36      1.58
## 104         1.48          NA     1.11      1.64
## 105         1.68          NA     0.22      0.37
## 106         1.62          NA     1.24      2.00
## 107         1.57          NA     0.97      1.52
## 108         1.70          NA     0.54      0.91
## 109         2.61          NA     0.04      0.10
## 110         1.71          NA     0.00      0.00
## 111         4.50          NA     0.02      0.10
## 112         1.53          NA     0.29      0.45
## 113         1.35          NA     0.01      0.02
## 114         1.99          NA     0.05      0.10
## 115         5.66          NA     0.00      0.01
## 116         2.69          NA     0.36      0.96
## 117         3.42          NA     0.28      0.95
## 118         0.86          NA     2.02      1.73
## 119         1.98          NA     5.38     10.66
## 120         1.75          NA     7.70     13.47
```

```r
# FALTA INFORMACIÓN DE PENETRACIÓN A PARTIR DE JULIO/2020 (TODAS VARIABLES)
# TAMBIÉN QUITAMOS VALOR, QUE ES PRECIO*VOLUMEN

data1 %<>%
  select(-Penetracion, -Valor)

summary(data1)
```

```
##       Ano            Mes                CCAA              Producto
##  Min.   :2018   Length:26634       Length:26634       Length:26634
##  1st Qu.:2018   Class :character   Class :character   Class :character
##  Median :2019   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2019
##  3rd Qu.:2019
##  Max.   :2020
##     Volumen           Precio_Medio       Cons_cpt         Gasto_cpt
##  Min.   :     0.0   Min.   : 0.000   Min.   : 0.0000   Min.   : 0.000
##  1st Qu.:   103.8   1st Qu.: 1.390   1st Qu.: 0.0700   1st Qu.: 0.150
##  Median :   404.6   Median : 1.790   Median : 0.1900   Median : 0.420
##  Mean   :  3435.7   Mean   : 2.166   Mean   : 0.6391   Mean   : 1.045
##  3rd Qu.:  1536.2   3rd Qu.: 2.680   3rd Qu.: 0.5400   3rd Qu.: 1.010
##  Max.   :451789.9   Max.   :38.750   Max.   :14.2900   Max.   :27.630
```

```r
# POR EL MOMENTO, HACEMOS ANÁLISIS GENERAL DE ESPAÑA
data1 %<>%
  filter(CCAA == "Total Nacional")

# ARREGLAMOS FECHA
data1 %<>%
  mutate(Fecha = parse_date(paste(Mes, Ano), locale = locale("es"), format = "%B %Y"))


summary(data1)
```

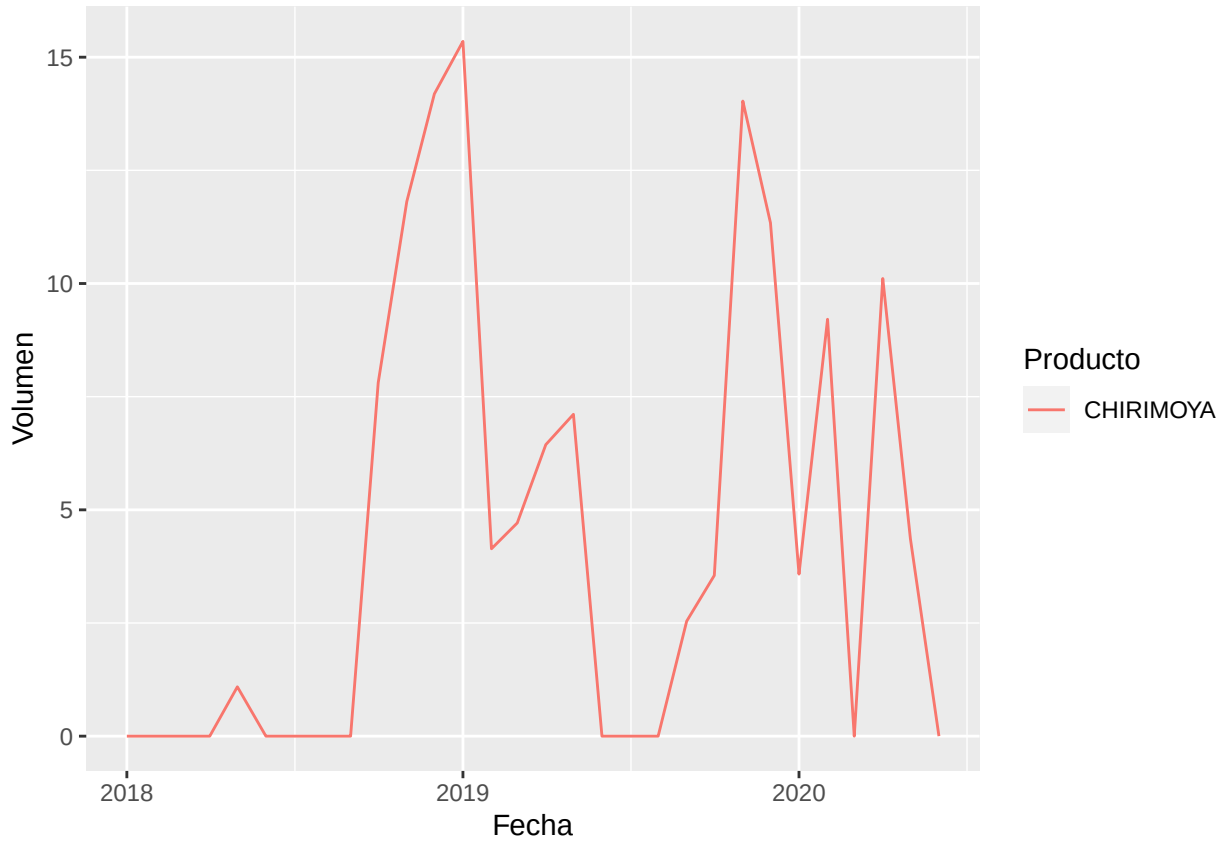
```
##       Ano            Mes                CCAA              Producto
##  Min.   :2018   Length:1593        Length:1593        Length:1593
##  1st Qu.:2018   Class :character   Class :character   Class :character
##  Median :2019   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2019
##  3rd Qu.:2020
##  Max.   :2020
##     Volumen         Precio_Medio      Cons_cpt         Gasto_cpt
##  Min.   :     0   Min.   :0.000   Min.   : 0.0000   Min.   : 0.000
##  1st Qu.:  3906   1st Qu.:1.460   1st Qu.: 0.0900   1st Qu.: 0.180
##  Median :  8994   Median :1.820   Median : 0.2000   Median : 0.470
##  Mean   : 30607   Mean   :2.181   Mean   : 0.6746   Mean   : 1.103
##  3rd Qu.: 26421   3rd Qu.:2.690   3rd Qu.: 0.5800   3rd Qu.: 1.080
##  Max.   :451790   Max.   :7.370   Max.   :10.1800   Max.   :18.600
##      Fecha
##  Min.   :2018-01-01
##  1st Qu.:2018-09-01
##  Median :2019-05-01
```

```
##  Mean   :2019-05-06
##  3rd Qu.:2020-01-01
##  Max.   :2020-11-01
```

```
# QUÉ HACEMOS CON VALORES 0? SON COMO NA'S?
data1 %>%
  filter(Volumen == 0) %>%
  select(Producto) %>%
  unique()
```

```
##    Producto
## 1 CHIRIMOYA
```

```
data1 %>%
  filter(Producto == "CHIRIMOYA") %>%
  ggplot() +
  geom_line(aes(x = Fecha, y = Volumen, color = Producto))
```



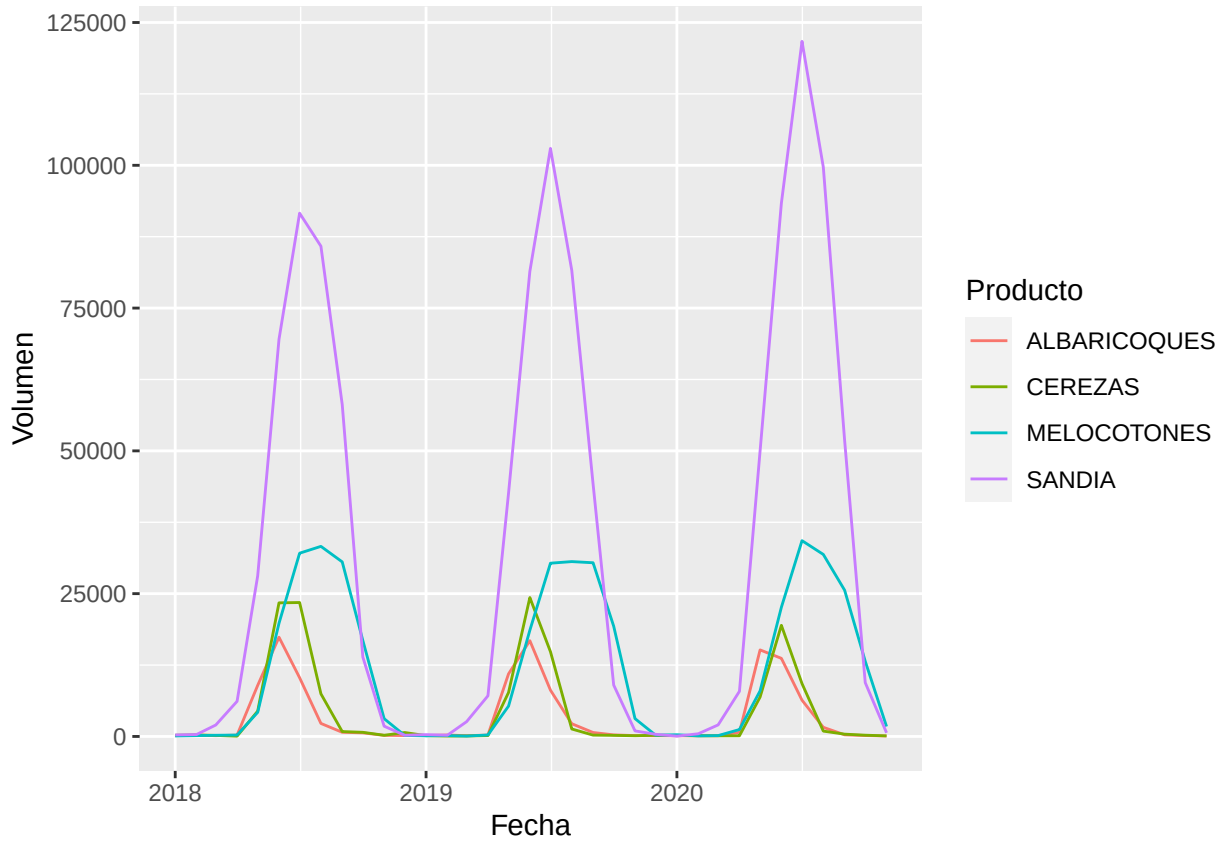

```
# LA CHIRIMOYA ES EL ÚNICO PRODUCTO CON ALGÚN "VOLUMEN == 0", LA SACAMOS DEL DATASET
data1 %<>%
  filter(Producto != "CHIRIMOYA")


summary(data1)
```

```
##       Ano            Mes                CCAA             Producto
##  Min.   :2018   Length:1563        Length:1563        Length:1563
##  1st Qu.:2018   Class :character   Class :character   Class :character
##  Median :2019   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2019
##  3rd Qu.:2020
##  Max.   :2020
##     Volumen          Precio_Medio      Cons_cpt         Gasto_cpt
##  Min.   :    49.6   Min.   :0.630   Min.   : 0.0000   Min.   : 0.000
##  1st Qu.:  4097.1   1st Qu.:1.465   1st Qu.: 0.0900   1st Qu.: 0.195
##  Median :  9375.1   Median :1.820   Median : 0.2100   Median : 0.480
##  Mean   : 31194.2   Mean   :2.169   Mean   : 0.6875   Mean   : 1.124
##  3rd Qu.: 26727.6   3rd Qu.:2.670   3rd Qu.: 0.5900   3rd Qu.: 1.090
##  Max.   :451789.9   Max.   :6.960   Max.   :10.1800   Max.   :18.600
##      Fecha
##  Min.   :2018-01-01
##  1st Qu.:2018-09-01
##  Median :2019-05-01
##  Mean   :2019-05-07
##  3rd Qu.:2020-01-01
##  Max.   :2020-11-01
```

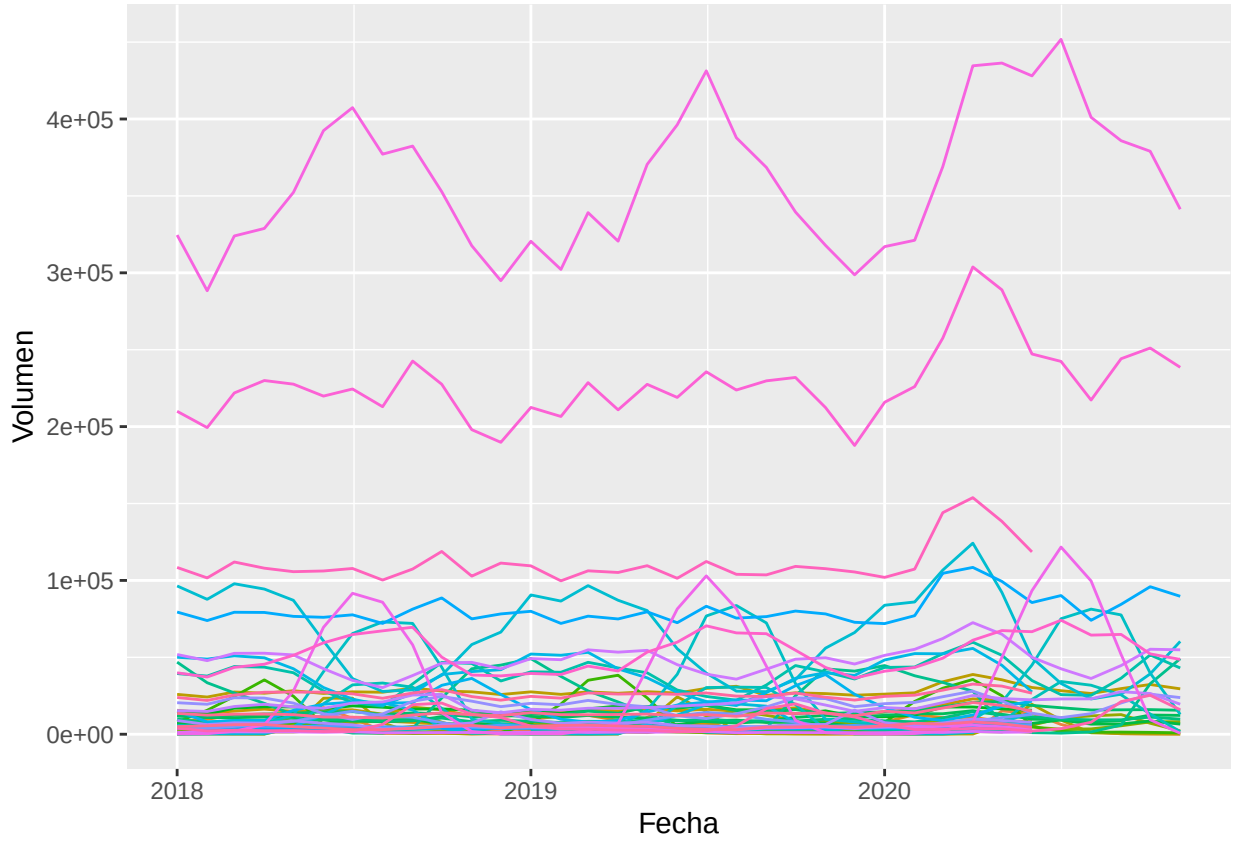
```
# QUEDAN PRODUCTOS CON OTRAS VARIABLES == 0
data1 %>%
  filter(Gasto_cpt == 0 | Cons_cpt == 0) %>%
  select(Producto) %>%
  unique()
```

```
##        Producto
## 1   MELOCOTONES
## 2  ALBARICOQUES
## 3       CEREZAS
## 13       SANDIA
```

```
data1 %>%
  filter(Producto == "MELOCOTONES" | Producto == "ALBARICOQUES" | Producto == "CEREZAS" | Producto == "S
  ggplot() +
  geom_line(aes(x = Fecha, y = Volumen, color = Producto))
```




```
# MELOCOTONES   ALBARICOQUES, CEREZAS,SANDIA tienen algún valor 0, pero porque son de temporada

data1 %>%
  ggplot() +
  geom_line(aes(x = Fecha, y = Volumen, color = Producto), show.legend = FALSE)
```




```
# Estandarizamos las variables
data1 %<>%
  group_by(Producto) %>%
  mutate(scVolumen = scale(Volumen),
         scPrecio_Medio = scale(Precio_Medio),
         scCons_cpt = scale(Cons_cpt),
         scGasto_cpt = scale(Gasto_cpt)) %>%
  ungroup()


data1 %>%
  ggplot() +
  geom_line(aes(x = Fecha, y = scVolumen, color = Producto), show.legend = FALSE)
```


```
# OJO! NO TODAS LAS VERDURAS LLEGAN HASTA NOVIEMBRE 2020

## Unidades
# Volumen: en miles de kg, litros o unidades en caso de huevos
# Valor: en miles de €
# Penetración: % de hogares que lo compran
```

DATASET5

```
#### 5.Covid ####
data5 = read.csv(pd5, sep = "|", dec = ",")
summary(data5) # NA's en pop y cumulative
```

```
##    dateRep               day              month             year
##  Length:58690       Min.   : 1.00   Min.   : 1.000   Min.   :2019
##  Class :character   1st Qu.: 8.00   1st Qu.: 5.000   1st Qu.:2020
##  Mode  :character   Median :16.00   Median : 7.000   Median :2020
##                     Mean   :15.99   Mean   : 6.801   Mean   :2020
##                     3rd Qu.:24.00   3rd Qu.: 9.000   3rd Qu.:2020
##                     Max.   :31.00   Max.   :12.000   Max.   :2020
##
##      cases             deaths         countriesAndTerritories    geoId
##  Min.   : -8261   Min.   :-1918.00   Length:58690            Length:58690
```

```
##  1st Qu.:     0   1st Qu.:    0.00   Class :character        Class :character
##  Median :    14   Median :    0.00   Mode  :character        Mode  :character
##  Mean   :  1061   Mean   :   24.77
##  3rd Qu.:   245   3rd Qu.:    4.00
##  Max.   :207913   Max.   : 4928.00
##
##  countryterritoryCode  popData2019        continentExp
##  Length:58690         Min.   :8.150e+02   Length:58690
##  Class :character     1st Qu.:1.325e+06   Class :character
##  Mode  :character     Median :7.813e+06   Mode  :character
##                       Mean   :4.125e+07
##                       3rd Qu.:2.861e+07
##                       Max.   :1.434e+09
##                       NA's   :108
##  Cumulative_number_for_14_days_of_COVID.19_cases_per_100000
##  Min.   :-147.4196
##  1st Qu.:   0.6745
##  Median :   6.3393
##  Mean   :  59.5616
##  3rd Qu.:  47.0096
##  Max.   :1900.8362
##  NA's   :2864
```

```
data5 %<>% mutate(Date = dmy(dateRep)) %>%
  select(c(Territory = countriesAndTerritories, Code = countryterritoryCode,
           Continent = continentExp, Date, Cases = cases, Death = deaths,
           Cumulative = Cumulative_number_for_14_days_of_COVID.19_cases_per_100000,
           Pop = popData2019)) %>%
  drop_na(Pop)
###
# No creo que necesitemos todos los territorios.
# Los que no tienen población registrada eliminados sin miedo
###
str(data5)
```

```
## 'data.frame':    58582 obs. of  8 variables:
##  $ Territory : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Code      : chr  "AFG" "AFG" "AFG" "AFG" ...
##  $ Continent : chr  "Asia" "Asia" "Asia" "Asia" ...
##  $ Date      : Date, format: "2020-11-29" "2020-11-28" ...
##  $ Cases     : int  228 214 0 200 185 246 252 154 232 282 ...
##  $ Death     : int  11 15 0 12 13 17 8 12 25 5 ...
##  $ Cumulative: num  6.85 6.78 6.4 7.34 7.2 ...
##  $ Pop       : int  38041757 38041757 38041757 38041757 38041757 38041757 38041757 38041757 38041757
```

```
summary(data5)
```

```
##    Territory             Code             Continent              Date
##  Length:58582       Length:58582       Length:58582       Min.   :2019-12-31
##  Class :character   Class :character   Class :character   1st Qu.:2020-05-04
##  Mode  :character   Mode  :character   Mode  :character   Median :2020-07-13
##                                                           Mean   :2020-07-10
##                                                           3rd Qu.:2020-09-21
```

```
##                                                        Max.   :2020-11-29
##
##      Cases              Death             Cumulative              Pop
##  Min.   : -8261   Min.   :-1918.00   Min.   :-147.4196   Min.   :8.150e+02
##  1st Qu.:     0   1st Qu.:    0.00   1st Qu.:   0.6745   1st Qu.:1.325e+06
##  Median :    14   Median :    0.00   Median :   6.3393   Median :7.813e+06
##  Mean   :  1063   Mean   :   24.81   Mean   :  59.5616   Mean   :4.125e+07
##  3rd Qu.:   246   3rd Qu.:    4.00   3rd Qu.:  47.0096   3rd Qu.:2.861e+07
##  Max.   :207913   Max.   : 4928.00   Max.   :1900.8362   Max.   :1.434e+09
##                                      NA's   :2756
```

```
# MIRAMOS CUANDO EMPIEZA COVID EN ESPAÑA (CASOS DE COVID)

data5 %>%
  filter(Territory == "Spain") %>%
  filter(Cases != 0) %>%
  filter(Date == min(Date))
```

```
##   Territory Code Continent       Date Cases Death Cumulative      Pop
## 1     Spain  ESP    Europe 2020-02-01     1     0 0.00213051 46937060
```

```
# EN FEBRERO EMPIEZAN A HABER CASOS EN ESPAÑA

data5 %>%
  filter(Territory == "Spain") %>%
  filter(Cases != 0) %>%
  filter(Date == max(Date))
```

```
##   Territory Code Continent       Date Cases Death Cumulative      Pop
## 1     Spain  ESP    Europe 2020-11-27 10853   294   361.3712 46937060
```

```
# DATOS HASTA NOVIEMBRE

# ANÁLISIS DESDE FEBRERO 2020 HASTA NOVIEMBRE 2020?
```

INDICE EFECTO COVID

COVID empieza en febrero:

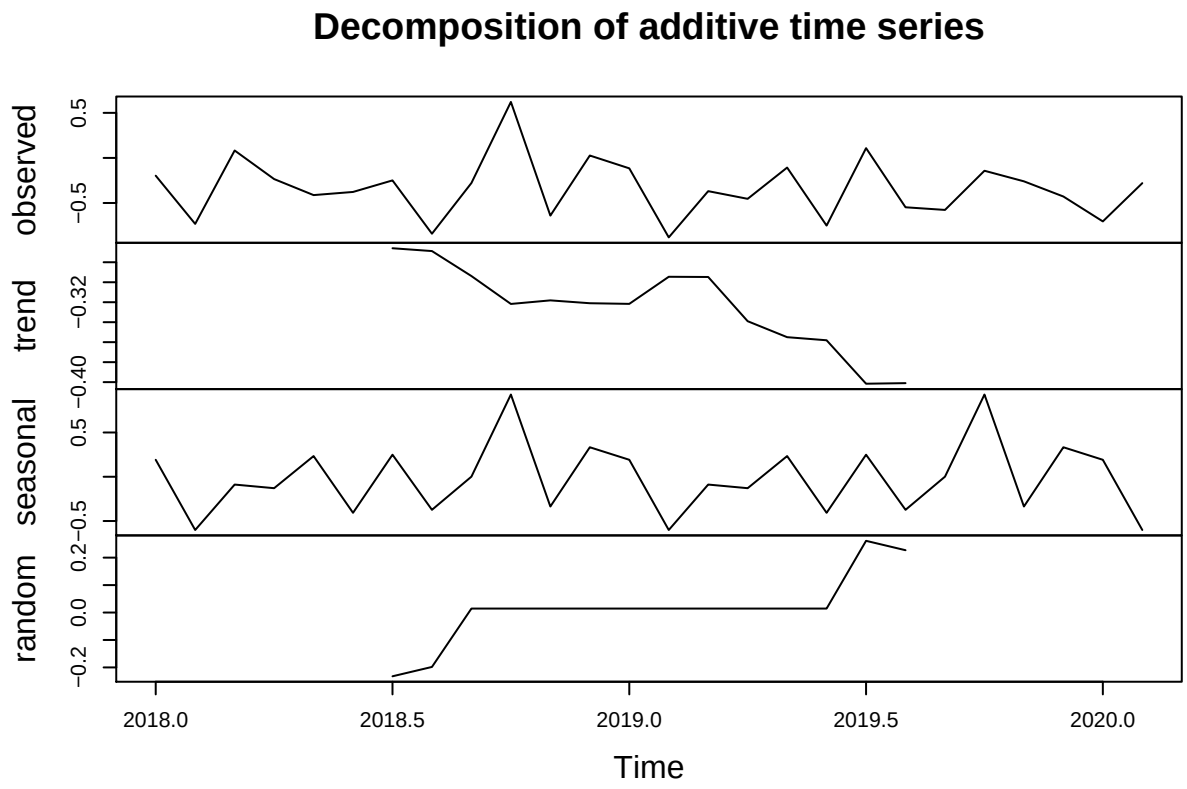
(1) estimar que hubiera pasado (para cada fruta/verdura) a partir de febrero si no hubiera ocurrido el COVID (predicción hasta noviembre)

(2) comparar con lo que realmente ha ocurrido

(3) estimar el efecto con la diferencia

```
# PREDICCIÓN TEMPORAL CON PATATAS(POR EJEMPLO) <- hacerlo dinámico para todas las frutas/verduras (y va
patatasPreCovid <- data1 %>%
  filter(Fecha <= "2020-02-01", Producto == "TOTAL PATATAS") %>%
  ungroup()
```

```r
patatasPostCovid <- data1 %>%
  filter(Fecha >= "2020-02-01", Producto == "TOTAL PATATAS") %>%
  ungroup()

serieTemporal <- ts(patatasPreCovid$scVolumen, start = 2018, frequency = 12)

plot(decompose(serieTemporal))
```

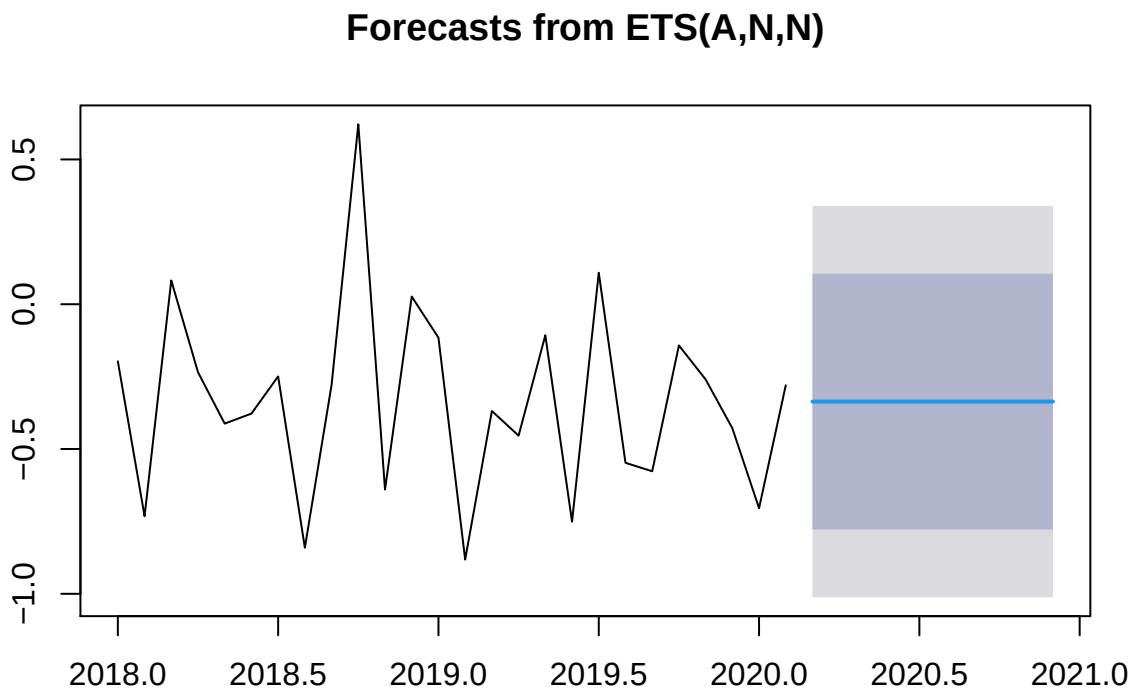



```r
library(forecast)
```

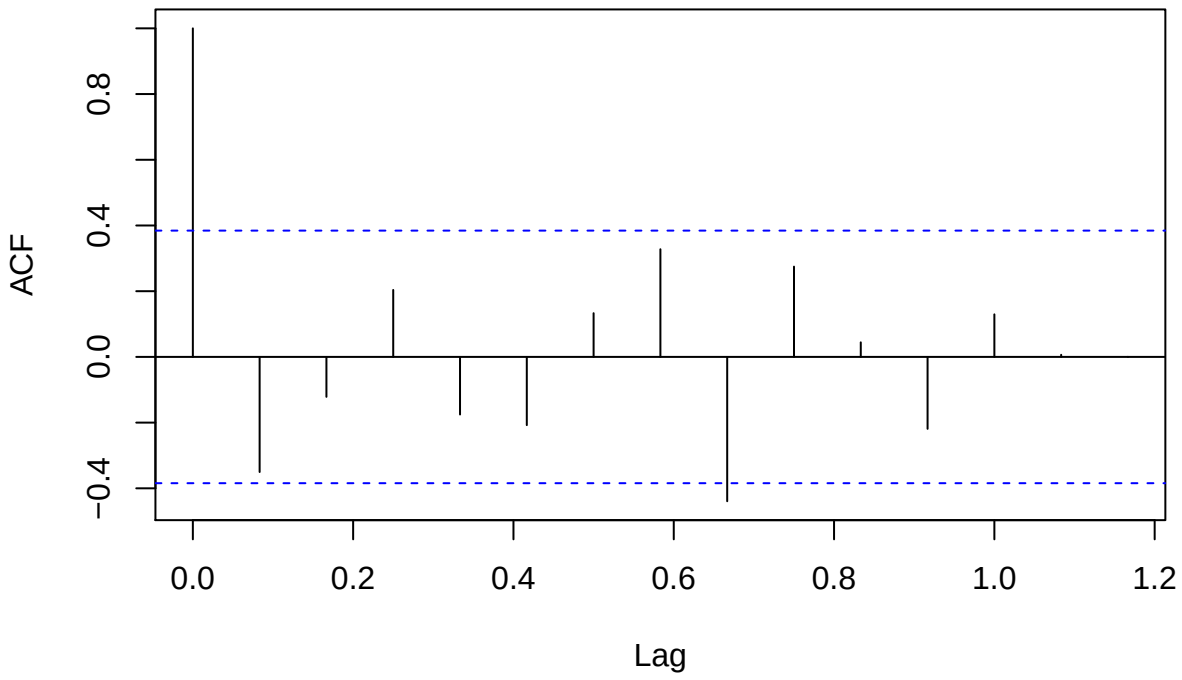
```
## Warning: package 'forecast' was built under R version 4.0.4
```

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

```r
arimaFit <- auto.arima(serieTemporal, stepwise = FALSE, approximation = FALSE) # CAMBIAR STEPWISE SI TA
forec <- forecast(serieTemporal, h = 10)
plot(forec)
```




```
acf(serieTemporal)
```



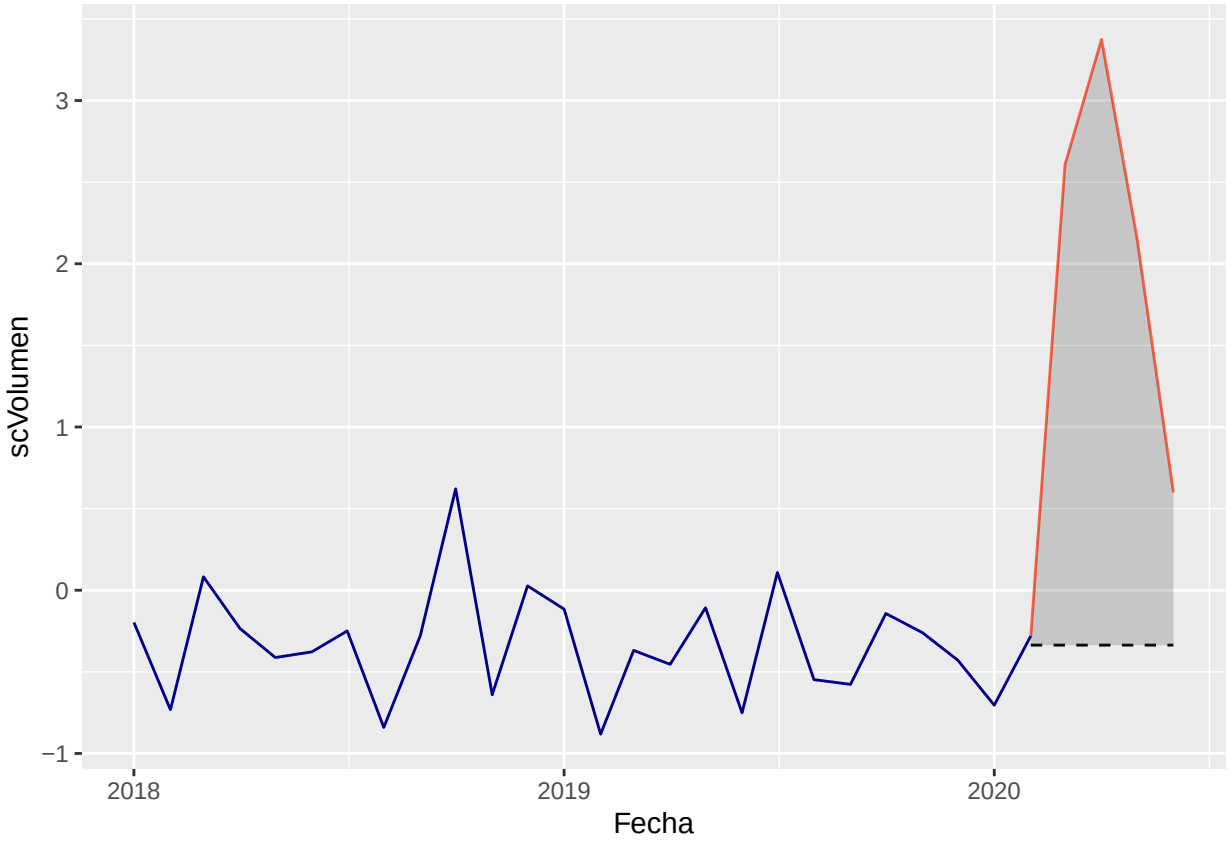

```
library(tseries)
```

```
## Warning: package 'tseries' was built under R version 4.0.4
```

```
adf.test(serieTemporal) # No se puede rechazar la h0 de que la serie no es estacionaria
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  serieTemporal
## Dickey-Fuller = -3.0272, Lag order = 2, p-value = 0.1811
## alternative hypothesis: stationary
```

```
patatasPostCovid$predVolumen <- as.vector(forec$mean)[1:nrow(patatasPostCovid)]

ggplot() +
  geom_line(data = patatasPreCovid, mapping = aes(x = Fecha, y = scVolumen), color = "darkblue") +
  geom_line(data = patatasPostCovid, mapping = aes(x = Fecha, y = scVolumen), color = "tomato") +
  geom_line(data = patatasPostCovid, mapping = aes(x = Fecha, y = predVolumen), color = "black", linety
  geom_ribbon(data = patatasPostCovid, mapping = aes(ymin = scVolumen, ymax = predVolumen, x = Fecha),
```




```
# EL AREA ES EL ÍNDICE EFECTO COVID. EL ÍNDICE DEBE DIVIDIRSE ENTRE EN NÚMERO DE MESES PREDICHOS, YA QU
patatasIndex <- sum(patatasPostCovid$scVolumen - patatasPostCovid$predVolumen) / nrow(patatasPostCovid)
# si es positivo, el covid ha hecho que el valor sea mayor, si es negativo, lo contrario
```

FUNCIÓN AUTOMATIZAR ÍNDICES

```
df = data1
prod = "TOTAL PATATAS"
var = "scVolumen"
covindex <- function(df, prod, var, plt = FALSE) {
  df %<>%
    filter(Producto == prod) %>%
    select(Fecha, var)
  preCovid <- filter(df, Fecha <= "2020-02-01")
  postCovid <- filter(df, Fecha >= "2020-02-01")
  st <- ts(preCovid[,2], start = 2018, frequency = 12)
  arimaPred <- forecast(st, h = 10)
  postCovid[,3] <- as.vector(arimaPred$mean)[1:nrow(postCovid)]
  index <- sum(postCovid[,2] - postCovid[,3]) / nrow(postCovid)
  if(plt) {
    plot <- ggplot(data = postCovid) +
      geom_line(data = preCovid, mapping = aes_string(x = "Fecha", y = var)) +
      geom_line(aes_string(x = "Fecha", y = var), color = "tomato") +
      geom_line(aes_string(x = "Fecha", y = "...3"), linetype = "dashed") +
```

```
      geom_ribbon(aes_string(ymin = var, ymax = "...3", x = "Fecha"), alpha = 0.2)
    return(list(index = index, plot = plot))
  } else {
    return(list(index = index))
  }
}

covindex(data1, "CEBOLLAS", "scPrecio_Medio")
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use ‘all_of(var)‘ instead of ‘var‘ to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
## $index
## [1] 0.5362106
```

tabla de indices

```
unique(data1$Producto) # hay 50 productos
```

```
##  [1] "TOTAL PATATAS"        "PATATAS FRESCAS"      "PATATAS CONGELADAS"
##  [4] "PATATAS PROCESADAS"   "T.HORTALIZAS FRESCAS" "TOMATES"
##  [7] "CEBOLLAS"             "AJOS"                 "COLES"
## [10] "PEPINOS"              "JUDIAS VERDES"        "PIMIENTOS"
## [13] "CHAMPIÃ‘ONES+O.SETAS" "LECHUGA/ESC./ENDIVIA" "ESPARRAGOS"
## [16] "VERDURAS DE HOJA"     "BERENJENAS"           "ZANAHORIAS"
## [19] "CALABACINES"          "OTR.HORTALIZAS/VERD." "BROCOLI"
## [22] "ALCACHOFAS"           "VERD./HORT. IV GAMA"  "T.FRUTAS FRESCAS"
## [25] "NARANJAS"             "MANDARINAS"           "LIMONES"
## [28] "PLATANOS"             "MANZANAS"             "PERAS"
## [31] "MELOCOTONES"          "NECTARINAS"           "ALBARICOQUES"
## [34] "FRESAS/FRESON"        "MELON"                "SANDIA"
## [37] "CIRUELAS"             "CEREZAS"              "UVAS"
## [40] "KIWI"                 "AGUACATE"             "PIÃ‘A"
## [43] "OTRAS FRUTAS FRESCAS" "POMELO"               "FRUTAS IV GAMA"
## [46] "PATATAS FRITAS"       "APIO"                 "COLIFLOR"
## [49] "PUERRO"               "MANGO"
```

```
colnames(data1)[10:13] # Hay 4 variables interesantes
```

```
## [1] "scVolumen"      "scPrecio_Medio" "scCons_cpt"     "scGasto_cpt"
```

```
tabla <- matrix(nrow = 50, ncol = 4, dimnames = list(unique(data1$Producto), colnames(data1)[10:13]))

for(prods in unique(data1$Producto)) {
  for(inds in colnames(data1)[10:13]) {
    tabla[prods, inds] <- covindex(data1, prods, inds)$index
  }
```

```
}

covindex(data1, "ESPARRAGOS", "scVolumen") # la predicción no va demasiado bien...
```

```
## $index
## [1] 0.9720303
```

ACP y Clustering

```
data1819 <- data1 %>%
  filter(CCAA == "Total Nacional") %>%
  filter(Ano %in% c(2018, 2019)) %>%
  group_by(Producto, Mes) %>%
  summarise(meanPrecio = mean(Precio_Medio),
            meanVolumen = mean(Volumen),
            meanCons = mean(Cons_cpt),
            meanGasto = mean(Gasto_cpt))

data20 <- data1 %>%
  filter(CCAA == "Total Nacional") %>%
  filter(Ano  == 2020)

datafin <- left_join(data20, data1819) %>%
  mutate(difPrecio = (Precio_Medio - meanPrecio) / meanPrecio,
         difVolumen = (Volumen - meanVolumen) / meanVolumen,
         difCons = (Cons_cpt - meanCons) / meanCons,
         difGasto = (Gasto_cpt - meanGasto) / meanGasto) %>%
  select(Producto, Mes, starts_with("dif")) %>%
  group_by(Producto) %>%
  summarise(difPrecio = mean(difPrecio),
            difVolumen = mean(difVolumen),
            difCons = mean(difCons),
            difGasto = mean(difGasto))

as.matrix(datafin)
rownames(datafin) = datafin$Producto
a <- na.omit(as.data.frame(as.matrix(datafin)[,-1]))
a %<>% mutate(across(c(difPrecio,difVolumen,difCons,difGasto), as.numeric))
rownames(a) = datafin$Producto
prcomp()
```