Vital Events Pre-Processing

Katie Schilling - 501130072

07/02/2022

Required tools to be loaded

```
library(dlookr)
```

```
##
## Attaching package: 'dlookr'

## The following object is masked from 'package:base':
##
##     transform
```

```
library(lubridate)
```

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

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

```
library(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
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag

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

```
library(moments)
```

```
##
## Attaching package: 'moments'

## The following objects are masked from 'package:dlookr':
##
##     kurtosis, skewness
```

```
library(ggpubr)
```

```
## Loading required package: ggplot2
```

```
library(smooth)
```

```
## Loading required package: greybox

## Package "greybox", v1.0.4 loaded.

##
## Attaching package: 'greybox'

## The following object is masked from 'package:lubridate':
##
##     hm

## This is package "smooth", v3.1.5
## By the way, have you already tried adam() function from smooth?
```

```
library(greybox)
library(forecast)
```

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

##
## Attaching package: 'forecast'

## The following object is masked from 'package:greybox':
##
##     forecast
```

```
## The following object is masked from 'package:ggpubr':
##
##     gghistogram
```

```
library(funModeling)
```

```
## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following object is masked from 'package:dlookr':
##
##     describe

## The following objects are masked from 'package:base':
##
##     format.pval, units

## funModeling v.1.9.4 :)
## Examples and tutorials at livebook.datascienceheroes.com
##  / Now in Spanish: librovivodecienciadedatos.ai
```

Import the first data set

```
Vital_events<- read.csv("C:/Users/Katie Schilling/Downloads/vital_events_data_by_month_1994-2021_q2 (1)
## Change the column names to cleaner versions
colnames(Vital_events)<- c("Month","Year","Births", "Marriages","Deaths","Stillbirths")
```

View the data to see what information is present

```
head(Vital_events)
```

```
##               Month Year Births Marriages Deaths Stillbirths
## 1   January/janvier 1994  11631      2078   8094          75
## 2 February/fÃ©vrier 1994  11254      2650   6428          62
## 3        March/mars 1994  13003      2557   6503          73
## 4       April/avril 1994  12576      3967   6224          74
## 5           May/mai 1994  13240      6493   6483          67
## 6         June/juin 1994  13072      7754   6187          66
```

Clean up of the vitals Data event. Removal of the french version of the month, as many of them did not import properly. Makes the data easier to read, view and work with.

```
Vital_events[Vital_events == "January/janvier"] <- "January"
Vital_events[Vital_events == "February/fÃ©vrier"] <- "February"
Vital_events[Vital_events == "March/mars"] <- "March"
Vital_events[Vital_events == "April/avril"] <- "April"
Vital_events[Vital_events == "May/mai"] <- "May"
Vital_events[Vital_events == "June/juin"] <- "June"
Vital_events[Vital_events == "July/juillet"] <- "July"
Vital_events[Vital_events == "August/aoÃ»t"] <- "August"
Vital_events[Vital_events == "September/septembre"] <- "September"
Vital_events[Vital_events == "October/octobre"] <- "October"
Vital_events[Vital_events == "November/novembre"] <- "November"
Vital_events[Vital_events == "December/dÃ©cembre"] <- "December"
```

Check data now to see if the changes are sufficent

```
head(Vital_events)
```

```
##      Month Year Births Marriages Deaths Stillbirths
## 1  January 1994  11631      2078   8094          75
## 2 February 1994  11254      2650   6428          62
## 3    March 1994  13003      2557   6503          73
## 4    April 1994  12576      3967   6224          74
## 5      May 1994  13240      6493   6483          67
## 6     June 1994  13072      7754   6187          66
```

Convert the month names to the corresponding month number

```
Vital_events <- Vital_events
Vital_events["Month"][Vital_events["Month"]== "January"]  <- 01
Vital_events["Month"][Vital_events["Month"]== "February"]  <- 02
Vital_events["Month"][Vital_events["Month"]== "March"]  <- 03
Vital_events["Month"][Vital_events["Month"]== "April"]  <- 04
Vital_events["Month"][Vital_events["Month"]== "May"]  <- 05
Vital_events["Month"][Vital_events["Month"]== "June"]  <- 06
Vital_events["Month"][Vital_events["Month"]== "July"]  <- 07
Vital_events["Month"][Vital_events["Month"]== "August"]  <-  08
Vital_events["Month"][Vital_events["Month"]== "September"]  <- 09
Vital_events["Month"][Vital_events["Month"]== "October"]  <- 10
Vital_events["Month"][Vital_events["Month"]== "November"]  <- 11
Vital_events["Month"][Vital_events["Month"]== "December"]  <- 12
```

Change the Year and Month columns to numeric values, and then create a column with the 2 values combined in the proper Month and Year format. Assigned day to the 1st day of each month so that I had a full date to work with

```
Vital_events$Month <- as.numeric(Vital_events$Month)
Vital_events$Year <- as.numeric(Vital_events$Year)
Vital_events$Date <- sprintf("%d-%02d-%s", Vital_events$Year, Vital_events$Month, "1")
```

Re-order Final dataset so that date is the first column and sort by date

```r
Vital_events <- Vital_events[c(7, 3:6)]
Vital_events <- Vital_events[order(Vital_events$Date),]
```

View Cleaned Data Set

```r
describe(Vital_events)
```

```
## Vital_events
##
##  5  Variables      330  Observations
## ---------------------------------------------------------------------------
## Date
##        n  missing distinct
##      330        0      330
##
## lowest : 1994-01-1 1994-02-1 1994-03-1 1994-04-1 1994-05-1
## highest: 2021-02-1 2021-03-1 2021-04-1 2021-05-1 2021-06-1
## ---------------------------------------------------------------------------
## Births
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      330        0      310        1    11763    832.6    10441    10694
##      .25      .50      .75      .90      .95
##    11260    11818    12288    12674    12886
##
## lowest : 10020 10059 10062 10103 10113, highest: 13104 13143 13195 13240 13398
## ---------------------------------------------------------------------------
## Marriages
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      330        0      320        1     5085     3171     1940     2101
##      .25      .50      .75      .90      .95
##     2596     3559     7627     9242     9804
##
## lowest :   597  1142  1314  1460  1722, highest: 10801 10830 11004 11083 11532
## ---------------------------------------------------------------------------
## Deaths
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      330        0      315        1     7500     1114     6202     6402
##      .25      .50      .75      .90      .95
##     6706     7326     8094     8897     9395
##
## lowest :  5926  6039  6060  6062  6064, highest: 10161 10712 10844 11121 11390
## ---------------------------------------------------------------------------
## Stillbirths
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      330        0      102        1    90.74    31.55     52.0     59.9
##      .25      .50      .75      .90      .95
##     73.0     90.5    114.0    126.0    132.0
##
## lowest :   0   2   8  19  24, highest: 146 147 148 150 156
## ---------------------------------------------------------------------------
```

```
summary(Vital_events)
```

```
##      Date               Births        Marriages         Deaths
##  Length:330         Min.   :10020   Min.   :  597   Min.   : 5926
##  Class :character   1st Qu.:11260   1st Qu.: 2596   1st Qu.: 6706
##  Mode  :character   Median :11818   Median : 3559   Median : 7326
##                     Mean   :11763   Mean   : 5085   Mean   : 7500
##                     3rd Qu.:12288   3rd Qu.: 7627   3rd Qu.: 8094
##                     Max.   :13398   Max.   :11532   Max.   :11390
##   Stillbirths
##  Min.   :  0.00
##  1st Qu.: 73.00
##  Median : 90.50
##  Mean   : 90.74
##  3rd Qu.:114.00
##  Max.   :156.00
```

Export CLeaned Dataset

```
write.csv(Vital_events,"C:/Users/Katie Schilling/Downloads/vital_events_clean.csv", row.names = FALSE)
```