Early-life socioeconomic inequality, adverse childhood experiences and biological aging in Colombia 1\_imputation data

Juan Rivillas

28/02/2022

## Contents Code 1: Data processing and preparation for analysis.

Code 1 is divided into three main steps:

Step 1. Preparing work space **(data manipulation steps)**. Step 2. **Imputation missing values** for data cleaning and exploration. Step 3. **Tidying data**: organize data in a consistent data structure.

Step 1 and 2 is described in this markdown.

Step 1. Preparing work space **(data manipulation steps)**. This step includes: 1.1. Working directory and loading of packages required for the tidying data. 1.2. Importing of original data base. 1.3. Verifying upload and exploring original data base. 1.4. Subset individuals with biomarkers -> filter BIOMARKER ==1 (Selection of 4,092 observations). 1.5. Verifying upload and exploring individual socioeconomic and biomarkers data (selection of 54 columns). 1.6. Overview of the entire dataframe (missing values, complete rates, histograms).

Step 2. **Imputation missing values** for data cleaning and exploration. This step includes: 2.1. Loading packages to work with missing values. 2.2. Descriptive statistics with missing values. 2.3. Understanding and exploring the missing value patterns. 2.4. Imputating missing values using multiple imputations with mice package. 2.5. Inspecting the quality of the imputed values.

## Step 1. Preparing work space (data manipulation steps).

**1.1. Working directory and loading of packages required for the tidying data**.

Assign file pathway to save automatically plots and outcomes in pdf format.

setwd("/Users/macbookpro/Documents/PhD Project/Data/SABE/Markdowns")

Install CRAN packages and install the relevant libraries. This code chunk shows the loading of packages required for the analyses.

**1.2. Importing of original SABE data base (Loading data EXCEL files)**. Note: All columns containind individual information were deleted to avoid any potential identification of individuals.

library(readxl)  
SABE\_Colombia <- read\_excel("~/Documents/PhD Project/Data/SABE/Dataframes/SABE\_Colombia.xlsx")

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP1917 / R1917C770: got 'CLORFENIRAMINA JARABE'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACY1917 / R1917C779: got 'DIFENHIDRAMINA'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP3240 / R3240C770: got 'HIDROXIDO DE ALUMINIO'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in APU3400 / R3400C1113: got 'INDICA QUE LE DUELE MUCHO AL  
## LEVANTARSE'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP3812 / R3812C770: got 'ASPROMIO'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP3918 / R3918C770: got 'PIROCICAN GEL'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACY3918 / R3918C779: got 'AACETATO DE ALUMINIO'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP4509 / R4509C770: got 'INSULINA DETEMIR 100ML'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in AOH7513 / R7513C1074: got 'SE SINTIO CANSADO'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in APU9431 / R9431C1113: got 'MIEDO A LASTIMARSE'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP9946 / R9946C770: got 'DIHIDROCODEINA BITARTRATO'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACY9946 / R9946C779: got 'NEOMICINA HIDROCORTISONA'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in APU10703 / R10703C1113: got 'ARTROSIS EN LAS RODILLAS'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP11348 / R11348C770: got 'AZARGA SOLUCION FAMILIA'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in APU13142 / R13142C1113: got '-la pam tiene problemas  
## respiratorios y se agita hablando, porlo cual los familiares no desean correr  
## riesgos con la salud del pam'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in AOH13832 / R13832C1074: got 'ME SENTÍ INSEGURO'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP14220 / R14220C770: got 'METFORMINA 850 MG'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACY14220 / R14220C779: got 'BECLOMETASONA'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in APU15888 / R15888C1113: got 'PAM NO PUEDE HACER EJERCICIO  
## POR DOLORES FUERTES EN LA RODILLA POR ARTRITIS'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP16309 / R16309C770: got 'ACETAMINOFEN'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACY16309 / R16309C779: got 'LOSARTAN'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in APU17636 / R17636C1113: got 'NO LO HACE SE MAREA'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in APU17639 / R17639C1113: got 'CHICUNGUNA'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in APU19207 / R19207C1113: got 'no finalizo quedo de pie y  
## gasto 16 segundos'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP19489 / R19489C770: got 'ALPRAZOLAN 0.5'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP19825 / R19825C770: got 'TRAZODONA 50 MG'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP21323 / R21323C770: got 'THIN LANCETS'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP21727 / R21727C770: got 'IPRATROPIO BRUMURO AEROSOL 0.02  
## MG'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP21837 / R21837C770: got 'ARAVA'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACY21837 / R21837C779: got 'CLORURO DE MAGNESIO'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP22463 / R22463C770: got 'TRIMETROPRINA'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACY22463 / R22463C779: got 'HIDROCLOROTIAZIDA'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP22703 / R22703C770: got 'insulina nph'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACY22703 / R22703C779: got 'insulina cristalina'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP22981 / R22981C770: got 'NAPROXENO'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACY22981 / R22981C779: got 'SINALGEN'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in ACP23262 / R23262C770: got 'ACIDO TRIOPICO'

## Warning in read\_fun(path = enc2native(normalizePath(path)), sheet\_i = sheet, :  
## Expecting logical in AOH23521 / R23521C1074: got 'La Pam Estaba cansada para  
## caminar por segunda vez.'

Another option for CSV files: SABE\_Colombia <- read\_csv(“SABE\_Colombia.csv”)

**1.3. Verifying upload and exploring original data base**. Here, Glimpse makes it possible to see every column within data frame. It’s a little like str applied to a data frame but it tries to show as much data as possible.

str(SABE\_Colombia)

## tibble [23,694 × 1,139] (S3: tbl\_df/tbl/data.frame)  
## $ ID : num [1:23694] 1 2 3 4 5 6 7 8 9 10 ...  
## $ PARTICIPA : num [1:23694] 1 1 1 1 1 1 1 1 1 1 ...  
## $ P102 : chr [1:23694] "11" "11" "11" "11" ...  
## $ P103 : chr [1:23694] "11001" "11001" "11001" "11001" ...  
## $ area\_residence : chr [1:23694] "1" "1" "1" "1" ...  
## $ P105 : chr [1:23694] "BOSA" "BOSA" "BOSA" "BOSA" ...  
## $ P114\_1 : num [1:23694] 1 1 1 1 2 1 2 1 1 1 ...  
## $ P114\_2 : num [1:23694] 1 1 1 2 2 2 2 1 1 2 ...  
## $ current\_sep : num [1:23694] 1 2 2 2 2 2 2 2 3 3 ...  
## $ P116 : num [1:23694] 1 1 1 1 1 2 2 1 2 2 ...  
## $ P117 : num [1:23694] 1 1 9 1 1 2 1 1 1 1 ...  
## $ P118 : num [1:23694] 1 1 9 1 1 1 2 1 2 2 ...  
## $ P119 : num [1:23694] 1 2 2 1 2 2 2 2 2 2 ...  
## $ sex : num [1:23694] 1 2 2 2 1 1 2 2 2 2 ...  
## $ age : num [1:23694] 60 68 79 70 66 62 79 62 62 65 ...  
## $ P122MES : num [1:23694] 7 5 7 4 12 12 12 3 7 12 ...  
## $ yearborn : num [1:23694] 1954 1946 1935 1944 1948 ...  
## $ P123 : num [1:23694] 19218834 26435628 28234267 41483660 19098477 ...  
## $ P124 : num [1:23694] 1 3 3 1 1 5 5 1 3 2 ...  
## $ P125\_P125T2\_1 : num [1:23694] 1 1 1 1 1 1 1 1 1 1 ...  
## $ P125\_P125T2\_2 : num [1:23694] 1 1 0 1 1 1 1 1 1 1 ...  
## $ P125\_P125T2\_3 : num [1:23694] 1 1 0 1 1 1 0 1 1 1 ...  
## $ P125\_P125T2\_4 : num [1:23694] 1 1 1 1 1 1 1 1 1 1 ...  
## $ P125 : num [1:23694] 4 4 2 4 4 4 3 4 4 4 ...  
## $ P126AR : num [1:23694] 1 1 1 1 1 1 1 1 1 1 ...  
## $ P126MESA : num [1:23694] 1 1 1 1 1 1 1 1 1 1 ...  
## $ P126PER : num [1:23694] 1 1 1 1 1 1 1 1 1 1 ...  
## $ P126REP : num [1:23694] 1 1 1 1 1 1 2 1 1 1 ...  
## $ P126 : num [1:23694] 3 3 3 3 3 3 3 3 3 3 ...  
## $ P127R9 : num [1:23694] 1 1 1 1 1 0 1 1 1 1 ...  
## $ P127R7 : num [1:23694] 1 1 0 1 1 0 0 0 1 1 ...  
## $ P127R5 : num [1:23694] 1 0 0 1 1 1 0 0 1 1 ...  
## $ P127R3 : num [1:23694] 1 0 0 0 1 1 0 0 1 1 ...  
## $ P127R1 : num [1:23694] 1 0 0 1 1 0 0 0 1 1 ...  
## $ P127 : num [1:23694] 5 2 1 4 5 2 1 1 5 5 ...  
## $ P128TPD : num [1:23694] 1 1 1 1 1 1 1 1 1 1 ...  
## $ P128DM : num [1:23694] 1 1 1 1 1 1 1 1 1 1 ...  
## $ P128PS : num [1:23694] 1 1 1 1 1 1 1 1 1 1 ...  
## $ P128 : num [1:23694] 3 3 3 3 3 3 3 3 3 3 ...  
## $ P129AR : num [1:23694] 0 0 1 1 1 0 0 1 1 1 ...  
## $ P129MESA : num [1:23694] 0 0 1 1 0 0 1 0 1 0 ...  
## $ P129PER : num [1:23694] 0 0 1 1 1 1 1 1 0 1 ...  
## $ P129 : num [1:23694] 0 0 3 3 2 1 2 2 2 2 ...  
## $ P130 : num [1:23694] 1 1 1 1 0 1 1 1 1 1 ...  
## $ MINMENPAM : num [1:23694] 16 13 13 18 17 14 13 14 18 18 ...  
## $ P131 : num [1:23694] 2 2 2 2 2 2 2 2 2 2 ...  
## $ P132 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P132A : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P133\_P133T2\_1 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P133\_P133T2\_2 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P133\_P133T2\_3 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P133\_P133T2\_4 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P133 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P134AR : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P134MESA : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P134PER : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P134REP : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P134 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P135R9 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P135R7 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P135R5 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P135R3 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P135R1 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P135 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P136TPD : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P136DM : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P136PS : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P136 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P137AR : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P137MESA : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P137PER : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P137 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P138 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ MINMENPRO : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P139\_1 : num [1:23694] 0 0 0 0 0 0 0 0 0 0 ...  
## $ P139\_2 : num [1:23694] 1 0 0 1 1 0 0 1 0 1 ...  
## $ P139\_3 : num [1:23694] 1 0 1 1 1 0 1 1 1 1 ...  
## $ P139\_4 : num [1:23694] 0 0 0 1 1 0 0 1 1 1 ...  
## $ P139\_5 : num [1:23694] 0 1 0 0 0 1 0 0 1 0 ...  
## $ P139\_6 : num [1:23694] 0 0 0 0 0 0 0 0 0 0 ...  
## $ P139\_7 : num [1:23694] 0 0 0 0 0 0 0 0 0 0 ...  
## $ P139CUAL : chr [1:23694] NA NA NA NA ...  
## $ P139A : num [1:23694] 5 1 5 7 7 1 1 7 5 7 ...  
## $ P140 : num [1:23694] 2 2 3 3 3 3 3 3 3 3 ...  
## $ health\_insurance : num [1:23694] 1 1 2 2 2 2 2 1 1 2 ...  
## $ P141A\_1 : num [1:23694] 0 0 NA NA NA NA NA 0 0 NA ...  
## $ P141A\_2 : num [1:23694] 0 0 NA NA NA NA NA 0 0 NA ...  
## $ P141A\_3 : num [1:23694] 0 0 NA NA NA NA NA 0 0 NA ...  
## $ P141A\_4 : num [1:23694] 0 0 NA NA NA NA NA 0 0 NA ...  
## $ P141A\_5 : num [1:23694] 1 1 NA NA NA NA NA 1 1 NA ...  
## $ P141A\_8 : num [1:23694] 0 0 NA NA NA NA NA 0 0 NA ...  
## $ P141A\_9 : num [1:23694] 0 0 NA NA NA NA NA 0 0 NA ...  
## $ P141B\_1 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P141B\_2 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P141B\_3 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P141B\_4 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P141B\_5 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P141B\_6 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## $ P141B\_8 : num [1:23694] NA NA NA NA NA NA NA NA NA NA ...  
## [list output truncated]

dim(SABE\_Colombia)

## [1] 23694 1139

#glimpse(SABE\_Colombia)  
head(SABE\_Colombia, 50)

## # A tibble: 50 × 1,139  
## ID PARTICIPA P102 P103 area\_residence P105 P114\_1 P114\_2 current\_sep  
## <dbl> <dbl> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 1 1 11 11001 1 BOSA 1 1 1  
## 2 2 1 11 11001 1 BOSA 1 1 2  
## 3 3 1 11 11001 1 BOSA 1 1 2  
## 4 4 1 11 11001 1 BOSA 1 2 2  
## 5 5 1 11 11001 1 BOSA 2 2 2  
## 6 6 1 11 11001 1 BOSA 1 2 2  
## 7 7 1 11 11001 1 BOSA 2 2 2  
## 8 8 1 11 11001 1 BOSA 1 1 2  
## 9 9 1 11 11001 1 ENGATIVA 1 1 3  
## 10 10 1 11 11001 1 ENGATIVA 1 2 3  
## # … with 40 more rows, and 1,130 more variables: P116 <dbl>, P117 <dbl>,  
## # P118 <dbl>, P119 <dbl>, sex <dbl>, age <dbl>, P122MES <dbl>,  
## # yearborn <dbl>, P123 <dbl>, P124 <dbl>, P125\_P125T2\_1 <dbl>,  
## # P125\_P125T2\_2 <dbl>, P125\_P125T2\_3 <dbl>, P125\_P125T2\_4 <dbl>, P125 <dbl>,  
## # P126AR <dbl>, P126MESA <dbl>, P126PER <dbl>, P126REP <dbl>, P126 <dbl>,  
## # P127R9 <dbl>, P127R7 <dbl>, P127R5 <dbl>, P127R3 <dbl>, P127R1 <dbl>,  
## # P127 <dbl>, P128TPD <dbl>, P128DM <dbl>, P128PS <dbl>, P128 <dbl>, …

**1.4. subset individuals with biomarkers** -> filter BIOMARKER ==1 (Selection of 4,092 observations). There are 22,639 individuals and 1,139 columns in the original database. The column “Biomarkers” was filtered to select individuals with collected biomarkers data (4,092 observations).

newdata <- SABE\_Colombia[ which(SABE\_Colombia$biomarkers=='1'), ]

Selection of 52 sociodemographic, socioeconomic, and health outcomes variables present in the data.

myvars <- c("ID", "current\_sep", "sep\_occupation", "family\_sep","area\_residence","educational\_level","health\_insurance","neglected\_food", "household\_violence", "migration", "migration\_yo","emotional\_abuse","poor\_health2",  
"early\_infection1","early\_infection2","early\_infection3","early\_infection4","early\_infection5","early\_infection6","early\_infection7","hdl", "ldl", "total\_cholesterol", "triglycerides", "glucose", "glycated\_haemoglobin", "weight", "height", "wc","systolic\_bprh","dyastolic\_bprh","systolic\_bplh", "dyastolic\_bplh", "hta", "diabetes", "cvd", "medication","P835","P835\_1","P835\_2","P835\_3","P835\_4","P835\_5","P835\_6","P835\_7","age", "sex", "ethnic\_group", "alcohol", "smoking","cash\_transfer1", "cash\_transfer2")

**1.5. Verifying upload and exploring individual socioeconomic and biomarker data**.

biomarkers <- newdata[myvars]

**1.6. Overview of the entire dataframe** (missing values, complete rates, histograms). A first glance there are 9 character variables and 43 numeric variables. There are 15 variables with not missing value, there are 34 variables with 20% missing values, and there are five columns with 50% missing values (medication and diastolic and systolic blood pressure in the dataframe).

skimr::skim(biomarkers)

Data summary

|  |  |
| --- | --- |
| Name | biomarkers |
| Number of rows | 4092 |
| Number of columns | 52 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 9 |
| numeric | 43 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| area\_residence | 0 | 1.00 | 1 | 1 | 0 | 3 | 0 |
| P835 | 1080 | 0.74 | 3 | 61 | 0 | 1218 | 0 |
| P835\_1 | 1614 | 0.61 | 3 | 128 | 0 | 1087 | 0 |
| P835\_2 | 2181 | 0.47 | 2 | 65 | 0 | 899 | 0 |
| P835\_3 | 2746 | 0.33 | 3 | 43 | 0 | 650 | 0 |
| P835\_4 | 3217 | 0.21 | 3 | 44 | 0 | 508 | 0 |
| P835\_5 | 3527 | 0.14 | 3 | 38 | 0 | 347 | 0 |
| P835\_6 | 3752 | 0.08 | 3 | 34 | 0 | 242 | 0 |
| P835\_7 | 3882 | 0.05 | 5 | 42 | 0 | 167 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | 0 | 1.00 | 11844.32 | 7213.21 | 2.0 | 5063.25 | 11773.0 | 18281.5 | 23646.0 | ▇▅▇▆▇ |
| current\_sep | 0 | 1.00 | 1.98 | 0.84 | 1.0 | 1.00 | 2.0 | 3.0 | 6.0 | ▇▂▁▁▁ |
| sep\_occupation | 497 | 0.88 | 5.38 | 11.22 | 1.0 | 1.00 | 3.0 | 5.0 | 99.0 | ▇▁▁▁▁ |
| family\_sep | 686 | 0.83 | 1.83 | 0.90 | 1.0 | 1.00 | 2.0 | 2.0 | 9.0 | ▇▂▁▁▁ |
| educational\_level | 0 | 1.00 | 3.20 | 6.44 | 1.0 | 2.00 | 2.0 | 3.0 | 99.0 | ▇▁▁▁▁ |
| health\_insurance | 0 | 1.00 | 1.69 | 0.77 | 1.0 | 1.00 | 2.0 | 2.0 | 9.0 | ▇▁▁▁▁ |
| neglected\_food | 686 | 0.83 | 1.78 | 0.72 | 1.0 | 1.00 | 2.0 | 2.0 | 9.0 | ▇▁▁▁▁ |
| household\_violence | 686 | 0.83 | 1.85 | 0.53 | 1.0 | 2.00 | 2.0 | 2.0 | 9.0 | ▇▁▁▁▁ |
| migration | 0 | 1.00 | 1.84 | 0.37 | 1.0 | 2.00 | 2.0 | 2.0 | 2.0 | ▂▁▁▁▇ |
| migration\_yo | 3428 | 0.16 | 43.41 | 21.93 | 1.0 | 25.00 | 48.5 | 59.0 | 99.0 | ▅▃▇▅▁ |
| emotional\_abuse | 686 | 0.83 | 3.98 | 0.74 | 1.0 | 4.00 | 4.0 | 4.0 | 9.0 | ▁▇▁▁▁ |
| poor\_health2 | 686 | 0.83 | 1.90 | 0.56 | 1.0 | 2.00 | 2.0 | 2.0 | 9.0 | ▇▁▁▁▁ |
| early\_infection1 | 687 | 0.83 | 2.01 | 0.43 | 1.0 | 2.00 | 2.0 | 2.0 | 9.0 | ▇▁▁▁▁ |
| early\_infection2 | 686 | 0.83 | 2.00 | 0.46 | 1.0 | 2.00 | 2.0 | 2.0 | 9.0 | ▇▁▁▁▁ |
| early\_infection3 | 686 | 0.83 | 1.64 | 1.06 | 1.0 | 1.00 | 2.0 | 2.0 | 9.0 | ▇▁▁▁▁ |
| early\_infection4 | 686 | 0.83 | 2.01 | 0.31 | 1.0 | 2.00 | 2.0 | 2.0 | 9.0 | ▇▁▁▁▁ |
| early\_infection5 | 686 | 0.83 | 2.00 | 0.52 | 1.0 | 2.00 | 2.0 | 2.0 | 9.0 | ▇▁▁▁▁ |
| early\_infection6 | 686 | 0.83 | 1.96 | 0.36 | 1.0 | 2.00 | 2.0 | 2.0 | 9.0 | ▇▁▁▁▁ |
| early\_infection7 | 686 | 0.83 | 1.98 | 0.61 | 1.0 | 2.00 | 2.0 | 2.0 | 9.0 | ▇▁▁▁▁ |
| hdl | 5 | 1.00 | 45.55 | 13.54 | 5.0 | 36.00 | 44.0 | 53.0 | 120.0 | ▁▇▃▁▁ |
| ldl | 4 | 1.00 | 126.21 | 35.95 | 30.0 | 101.00 | 125.0 | 149.0 | 338.0 | ▂▇▂▁▁ |
| total\_cholesterol | 4 | 1.00 | 194.58 | 41.78 | 60.0 | 166.00 | 193.0 | 221.0 | 439.0 | ▁▇▃▁▁ |
| triglycerides | 5 | 1.00 | 163.00 | 96.90 | 42.0 | 105.00 | 141.0 | 192.5 | 2062.0 | ▇▁▁▁▁ |
| glucose | 14 | 1.00 | 100.78 | 36.84 | 31.0 | 86.00 | 93.0 | 103.0 | 543.0 | ▇▁▁▁▁ |
| glycated\_haemoglobin | 9 | 1.00 | 13.80 | 1.73 | 0.2 | 12.70 | 13.8 | 14.9 | 24.1 | ▁▁▇▃▁ |
| weight | 86 | 0.98 | 73.07 | 75.34 | 30.0 | 56.00 | 64.5 | 73.0 | 777.0 | ▇▁▁▁▁ |
| height | 230 | 0.94 | 146.39 | 124.73 | 1.0 | 144.00 | 153.0 | 162.0 | 777.0 | ▇▅▁▁▁ |
| wc | 194 | 0.95 | 105.00 | 91.56 | 66.0 | 85.00 | 93.0 | 101.0 | 777.0 | ▇▁▁▁▁ |
| systolic\_bprh | 2104 | 0.49 | 188.54 | 463.96 | 84.0 | 122.00 | 136.0 | 154.0 | 8888.0 | ▇▁▁▁▁ |
| dyastolic\_bprh | 2121 | 0.48 | 113.46 | 411.06 | 43.0 | 68.00 | 76.0 | 89.0 | 8888.0 | ▇▁▁▁▁ |
| systolic\_bplh | 2104 | 0.49 | 232.05 | 818.13 | 60.0 | 117.00 | 131.0 | 150.0 | 8888.0 | ▇▁▁▁▁ |
| dyastolic\_bplh | 2118 | 0.48 | 89.32 | 219.01 | 42.0 | 68.00 | 75.0 | 83.0 | 8888.0 | ▇▁▁▁▁ |
| hta | 0 | 1.00 | 1.45 | 0.58 | 1.0 | 1.00 | 1.0 | 2.0 | 9.0 | ▇▁▁▁▁ |
| diabetes | 0 | 1.00 | 1.85 | 0.52 | 1.0 | 2.00 | 2.0 | 2.0 | 9.0 | ▇▁▁▁▁ |
| cvd | 0 | 1.00 | 1.86 | 0.40 | 1.0 | 2.00 | 2.0 | 2.0 | 9.0 | ▇▁▁▁▁ |
| medication | 0 | 1.00 | 1.26 | 0.44 | 1.0 | 1.00 | 1.0 | 2.0 | 2.0 | ▇▁▁▁▃ |
| age | 0 | 1.00 | 70.49 | 8.03 | 60.0 | 64.00 | 69.0 | 76.0 | 105.0 | ▇▅▂▁▁ |
| sex | 0 | 1.00 | 1.60 | 0.49 | 1.0 | 1.00 | 2.0 | 2.0 | 2.0 | ▅▁▁▁▇ |
| ethnic\_group | 686 | 0.83 | 4.43 | 1.61 | 1.0 | 4.00 | 5.0 | 5.0 | 9.0 | ▂▆▇▁▁ |
| alcohol | 0 | 1.00 | 4.58 | 1.21 | 1.0 | 5.00 | 5.0 | 5.0 | 9.0 | ▁▁▇▁▁ |
| smoking | 0 | 1.00 | 2.93 | 1.14 | 1.0 | 2.00 | 4.0 | 4.0 | 9.0 | ▇▇▁▁▁ |
| cash\_transfer1 | 0 | 1.00 | 1.98 | 0.39 | 1.0 | 2.00 | 2.0 | 2.0 | 9.0 | ▇▁▁▁▁ |
| cash\_transfer2 | 0 | 1.00 | 1.79 | 0.59 | 1.0 | 2.00 | 2.0 | 2.0 | 9.0 | ▇▁▁▁▁ |

## Step 2. Imputation missing values for data cleaning and exploration.

**2.1. Loading packages to work with missing values**.

library(mice) #Loading the mice package

##   
## Attaching package: 'mice'

## The following object is masked from 'package:stats':  
##   
## filter

## The following objects are masked from 'package:base':  
##   
## cbind, rbind

library(VIM) #Loading the following package for looking at the missing values

## Loading required package: colorspace

## Warning: package 'colorspace' was built under R version 4.1.2

## Loading required package: grid

## VIM is ready to use.

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

##   
## Attaching package: 'VIM'

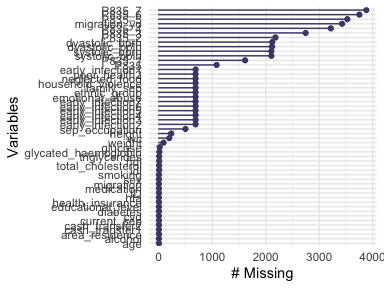
## The following object is masked from 'package:datasets':  
##   
## sleep

library(lattice)  
library (naniar)#tidy ways to summarise, visualize, and manipulate misisng data.

**2.2. Descriptive statistics with missing values**. To visualise the missings for each variable.

gg\_miss\_var(biomarkers)

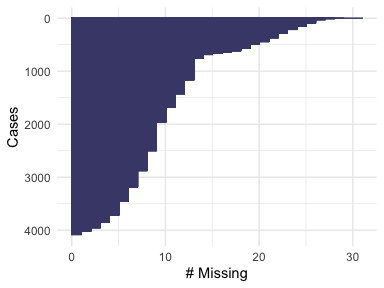
## Warning: It is deprecated to specify `guide = FALSE` to remove a guide. Please  
## use `guide = "none"` instead.



str(biomarkers) #to see every column within data frame Look at the percent missing by setting and set the y limits to be between 0 and 100.

look at the missings across cases.

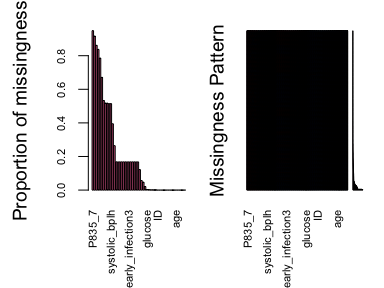
gg\_miss\_case(biomarkers)



Plot variables sorted by number of missing and individuals with missings values to identify and visualize patterns of missing entries.

biomarkers\_miss = aggr(biomarkers, col=mdc(1:2), numbers=TRUE, sortVars=TRUE, labels=names(biomarkers), cex.axis=.7, gap=3, ylab=c("Proportion of missingness","Missingness Pattern"))

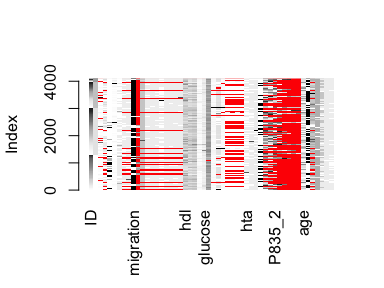
## Warning in plot.aggr(res, ...): not enough vertical space to display frequencies  
## (too many combinations)



##   
## Variables sorted by number of missings:   
## Variable Count  
## P835\_7 0.9486803519  
## P835\_6 0.9169110459  
## P835\_5 0.8619257087  
## migration\_yo 0.8377321603  
## P835\_4 0.7861681329  
## P835\_3 0.6710654936  
## P835\_2 0.5329912023  
## dyastolic\_bprh 0.5183284457  
## dyastolic\_bplh 0.5175953079  
## systolic\_bprh 0.5141739980  
## systolic\_bplh 0.5141739980  
## P835\_1 0.3944281525  
## P835 0.2639296188  
## early\_infection1 0.1678885630  
## family\_sep 0.1676441838  
## neglected\_food 0.1676441838  
## household\_violence 0.1676441838  
## emotional\_abuse 0.1676441838  
## poor\_health2 0.1676441838  
## early\_infection2 0.1676441838  
## early\_infection3 0.1676441838  
## early\_infection4 0.1676441838  
## early\_infection5 0.1676441838  
## early\_infection6 0.1676441838  
## early\_infection7 0.1676441838  
## ethnic\_group 0.1676441838  
## sep\_occupation 0.1214565005  
## height 0.0562072336  
## wc 0.0474095797  
## weight 0.0210166178  
## glucose 0.0034213099  
## glycated\_haemoglobin 0.0021994135  
## hdl 0.0012218964  
## triglycerides 0.0012218964  
## ldl 0.0009775171  
## total\_cholesterol 0.0009775171  
## ID 0.0000000000  
## current\_sep 0.0000000000  
## area\_residence 0.0000000000  
## educational\_level 0.0000000000  
## health\_insurance 0.0000000000  
## migration 0.0000000000  
## hta 0.0000000000  
## diabetes 0.0000000000  
## cvd 0.0000000000  
## medication 0.0000000000  
## age 0.0000000000  
## sex 0.0000000000  
## alcohol 0.0000000000  
## smoking 0.0000000000  
## cash\_transfer1 0.0000000000  
## cash\_transfer2 0.0000000000

**2.3. Understanding and exploring the missingness patterns**. This graph check association between the value or a variable and the missingness of another one using plots.

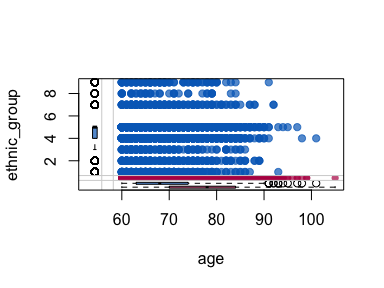
matrixplot(biomarkers, sortby = 2)



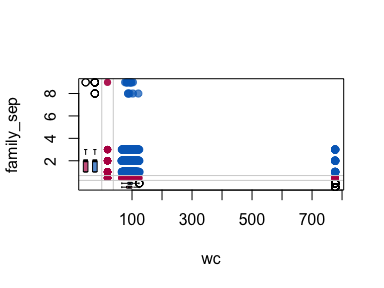
##   
## Click in a column to sort by the corresponding variable.  
## To regain use of the VIM GUI and the R console, click outside the plot region.

Here, we apply additional helpful visual approach (box plot) to draw margin plot of missing data patterns. For instance, the red box plot on the left shows the distribution of age with systolic blood pressure missing at the bottom of the graph. If the red and blue box plots are very similar, then the assumption of the MCAR data is correct. This means the probability of being missing is the same for all cases, then the data are said to be missing completely at random (MCAR). This approach was applied to six combination of variables with missing data. Overall, missing data is distributed similarly observations in the data.

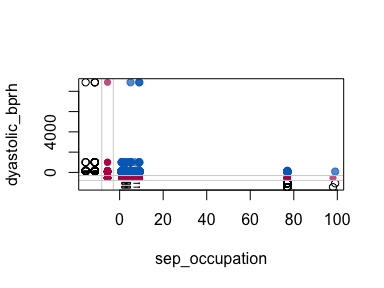
marginplot(biomarkers[, c("age", "ethnic\_group")], col = mdc(1:2), cex.numbers = 1.2, pch = 19)



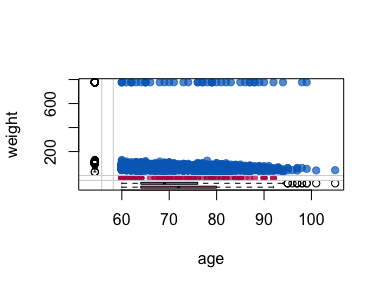
marginplot(biomarkers[, c("wc", "family\_sep")], col = mdc(1:2), cex.numbers = 1.2, pch = 19)



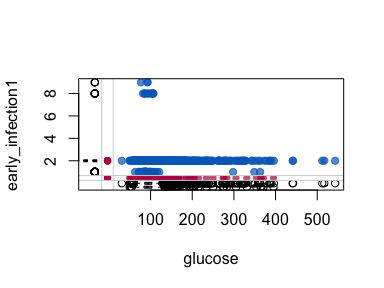
marginplot(biomarkers[, c("sep\_occupation", "dyastolic\_bprh")], col = mdc(1:2), cex.numbers = 1.2, pch = 19)



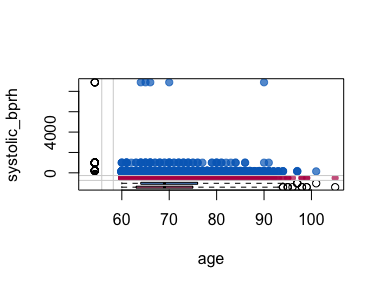
marginplot(biomarkers[, c("age", "weight")], col = mdc(1:2), cex.numbers = 1.2, pch = 19)



marginplot(biomarkers[, c("glucose", "early\_infection1")], col = mdc(1:2), cex.numbers = 1.2, pch = 19)



marginplot(biomarkers[, c("age","systolic\_bprh")], col = mdc(1:2), cex.numbers = 1.2, pch = 19)



#summary with missing values  
pct\_miss(biomarkers) #percentage of missing value in the data

## [1] 20.60352

n\_miss(biomarkers) #number of missing values

## [1] 43841

n\_complete(biomarkers) #without missing value

## [1] 168943

n\_miss(biomarkers$migration\_yo) #number of missing value for var

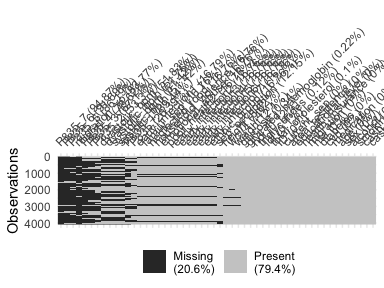
## [1] 3428

as\_shadow(biomarkers) #a matrix with missing and non missing bind\_shadow(biomarkers) #Initial matrix concatenated with the matrix with missing and non missing

Check percentage of missing data patterns in each column. There are 79% observations with not missing values. There are 21% observations with missing values in medication variables, dyastolic and systolic blood pressure, early-life infection and childhood experiences variables (which corresponds 42,841 missing values).

Apply clustering to find similar missingness groups by setting.

vis\_miss(biomarkers, sort\_miss = TRUE)

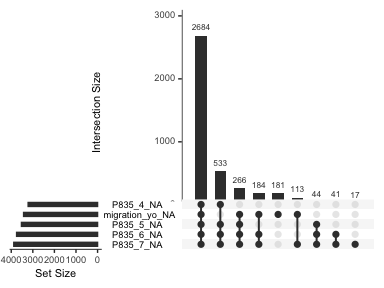


There are a lot of different clusters here, to see clear relationships here we apply another technique to try arranging by different variables using arrange. It is possibly to identify common missigness patterns: there are same 16% missing values in family sep, early-life infection, childhood experiences variables, and ethnic groups (n=686).

Four measurement variables of dyastolic and systolic blood pressure have the same number of individuals with missing values (51%).

Check the combination and patterns of missingness by looking at an upset plot of the missingness. Here, medication has the most missing values and there are 2684 cases where five medications have missing values together.

gg\_miss\_upset(biomarkers)

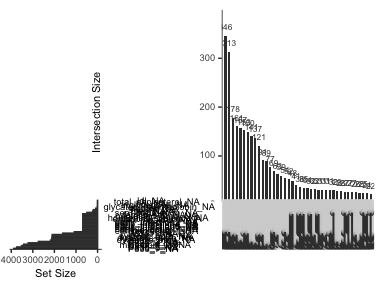


Look at histogram to clearly depicts the influence of missing values in the variables checking intersection sizes across variables.

n\_var\_miss(biomarkers)

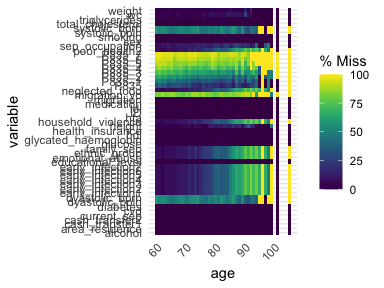
## [1] 36

gg\_miss\_upset(biomarkers, nsets = n\_var\_miss(biomarkers))



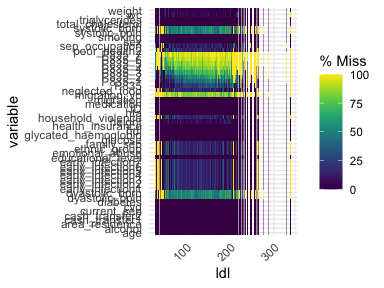
Looking at the missingness patterns by age missing values.

gg\_miss\_fct(biomarkers, age)



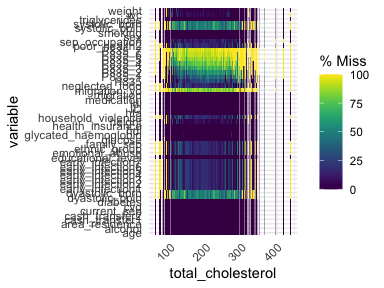
gg\_miss\_fct(biomarkers, ldl)

## Warning: Removed 51 rows containing missing values (geom\_tile).



gg\_miss\_fct(biomarkers, total\_cholesterol)

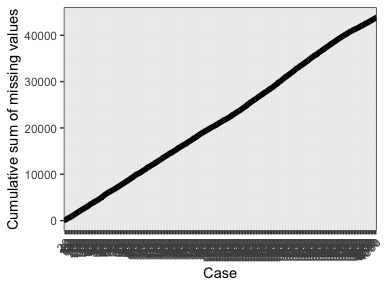
## Warning: Removed 51 rows containing missing values (geom\_tile).



Explore the missingness by some categorical variable using facetting to take a closer look at the distribution of missing values.

Reviewing the cumulative sum of missing values, reading the rows of the dataframe from the top to bottom using a plot.

gg\_miss\_case\_cumsum(biomarkers)



**2.4. imputing missing values using** using predictive mean matching as imputation method with mice package (Generates Multivariate Imputations by Chained Equations MICE).

Note:This step takes 60 minutes processing data.

Here is an explanation of the parameters used:

m – Refers to 5 imputed data sets. maxit – Refers to no. of iterations taken to impute missing values. method – Refers to method used in imputation. we used predictive mean matching.

Check the imputation method used for each variable (mice makes it very easy to do).

Check the imputed data for each observation (first column left) within each imputed dataframe (first row at the top). The missing values have been replaced with the imputed values in the first of the five dataframe.

tempDatafamily\_sep tempDataethnic\_group tempDatasystolic\_bprh tempDatasystolic\_bplh tempDatadyastolic\_bprh tempDatadyastolic\_bplh

Since there are 5 imputed data sets, we select any using complete() function to export imputed data. After imputation, there are 90% observations with missing any information and 10% observation with missing values (which corresponds 21,999 missing values).

**2.5. Inspecting the quality of imputed data**. Inspecting the distribution of original and imputed data and the quality of the imputations using the xyplot() and densityplot() functions to verify imputations. First, using a scatterplot and plot age against all the other variables. Second, check whether that the shape of the magenta points (imputed) matches the shape of the blue ones (observed) using density plots. The matching shape tells us that the imputed values are indeed “plausible values”.

densityplot(tempData) densityplot(tempDataearly\_infection2) densityplot(tempDatadyastolic\_bprh) densityplot(tempDataweight) densityplot(tempDataethnic\_group)

Save imputed data base for transformation data.