##Dealing with Missing Data

options(tidyverse.quiet = TRUE)  
library(titanic)  
library(tidyverse)  
library(tidymodels)

## ── Attaching packages ────────────────────────────────────── tidymodels 1.1.0 ──

## ✔ broom 1.0.4 ✔ rsample 1.1.1  
## ✔ dials 1.2.0 ✔ tune 1.1.1  
## ✔ infer 1.0.4 ✔ workflows 1.1.3  
## ✔ modeldata 1.1.0 ✔ workflowsets 1.0.1  
## ✔ parsnip 1.1.0 ✔ yardstick 1.2.0  
## ✔ recipes 1.0.6

## ── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
## ✖ scales::discard() masks purrr::discard()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ recipes::fixed() masks stringr::fixed()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ yardstick::spec() masks readr::spec()  
## ✖ recipes::step() masks stats::step()  
## • Dig deeper into tidy modeling with R at https://www.tmwr.org

library(mice) #package for imputation

##   
## Attaching package: 'mice'

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

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

library(VIM) #visualizing missingness

## Loading required package: colorspace

## Loading required package: grid

## The legacy packages maptools, rgdal, and rgeos, underpinning this package  
## will retire shortly. Please refer to R-spatial evolution reports on  
## https://r-spatial.org/r/2023/05/15/evolution4.html for details.  
## This package is now running under evolution status 0

## 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:recipes':  
##   
## prepare

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

library(naniar) #visualizing missingness  
library(skimr) #alternative way to view dataset summaries

##   
## Attaching package: 'skimr'

## The following object is masked from 'package:naniar':  
##   
## n\_complete

library(UpSetR) #visualizing missingness

Read in dataset

titanic = titanic::titanic\_train

Structure and summary

str(titanic)

## 'data.frame': 891 obs. of 12 variables:  
## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Survived : int 0 1 1 1 0 0 0 0 1 1 ...  
## $ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...  
## $ Name : chr "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...  
## $ Sex : chr "male" "female" "female" "female" ...  
## $ Age : num 22 38 26 35 35 NA 54 2 27 14 ...  
## $ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...  
## $ Parch : int 0 0 0 0 0 0 0 1 2 0 ...  
## $ Ticket : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Cabin : chr "" "C85" "" "C123" ...  
## $ Embarked : chr "S" "C" "S" "S" ...

summary(titanic)

## PassengerId Survived Pclass Name   
## Min. : 1.0 Min. :0.0000 Min. :1.000 Length:891   
## 1st Qu.:223.5 1st Qu.:0.0000 1st Qu.:2.000 Class :character   
## Median :446.0 Median :0.0000 Median :3.000 Mode :character   
## Mean :446.0 Mean :0.3838 Mean :2.309   
## 3rd Qu.:668.5 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :891.0 Max. :1.0000 Max. :3.000   
##   
## Sex Age SibSp Parch   
## Length:891 Min. : 0.42 Min. :0.000 Min. :0.0000   
## Class :character 1st Qu.:20.12 1st Qu.:0.000 1st Qu.:0.0000   
## Mode :character Median :28.00 Median :0.000 Median :0.0000   
## Mean :29.70 Mean :0.523 Mean :0.3816   
## 3rd Qu.:38.00 3rd Qu.:1.000 3rd Qu.:0.0000   
## Max. :80.00 Max. :8.000 Max. :6.0000   
## NA's :177   
## Ticket Fare Cabin Embarked   
## Length:891 Min. : 0.00 Length:891 Length:891   
## Class :character 1st Qu.: 7.91 Class :character Class :character   
## Mode :character Median : 14.45 Mode :character Mode :character   
## Mean : 32.20   
## 3rd Qu.: 31.00   
## Max. :512.33   
##

skim(titanic)

Data summary

|  |  |
| --- | --- |
| Name | titanic |
| Number of rows | 891 |
| Number of columns | 12 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 5 |
| numeric | 7 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Name | 0 | 1 | 12 | 82 | 0 | 891 | 0 |
| Sex | 0 | 1 | 4 | 6 | 0 | 2 | 0 |
| Ticket | 0 | 1 | 3 | 18 | 0 | 681 | 0 |
| Cabin | 0 | 1 | 0 | 15 | 687 | 148 | 0 |
| Embarked | 0 | 1 | 0 | 1 | 2 | 4 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PassengerId | 0 | 1.0 | 446.00 | 257.35 | 1.00 | 223.50 | 446.00 | 668.5 | 891.00 | ▇▇▇▇▇ |
| Survived | 0 | 1.0 | 0.38 | 0.49 | 0.00 | 0.00 | 0.00 | 1.0 | 1.00 | ▇▁▁▁▅ |
| Pclass | 0 | 1.0 | 2.31 | 0.84 | 1.00 | 2.00 | 3.00 | 3.0 | 3.00 | ▃▁▃▁▇ |
| Age | 177 | 0.8 | 29.70 | 14.53 | 0.42 | 20.12 | 28.00 | 38.0 | 80.00 | ▂▇▅▂▁ |
| SibSp | 0 | 1.0 | 0.52 | 1.10 | 0.00 | 0.00 | 0.00 | 1.0 | 8.00 | ▇▁▁▁▁ |
| Parch | 0 | 1.0 | 0.38 | 0.81 | 0.00 | 0.00 | 0.00 | 0.0 | 6.00 | ▇▁▁▁▁ |
| Fare | 0 | 1.0 | 32.20 | 49.69 | 0.00 | 7.91 | 14.45 | 31.0 | 512.33 | ▇▁▁▁▁ |

Factor conversion, recoding, and variable selection.

titanic = titanic %>% mutate(Survived = as\_factor(Survived)) %>%   
 mutate(Survived = fct\_recode(Survived, "No" = "0", "Yes" = "1" )) %>%  
 mutate(Pclass = as\_factor(Pclass)) %>% mutate(Sex = as.factor(Sex)) %>%  
 mutate(Embarked = as\_factor(Embarked)) %>%   
 mutate(Embarked = fct\_recode(Embarked,"Unknown"="","Cherbourg"="C","Southampton"="S","Queenstown"="Q")) %>%   
 select(Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked)  
  
str(titanic)

## 'data.frame': 891 obs. of 8 variables:  
## $ Survived: Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 1 1 2 2 ...  
## $ Pclass : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...  
## $ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...  
## $ Age : num 22 38 26 35 35 NA 54 2 27 14 ...  
## $ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...  
## $ Parch : int 0 0 0 0 0 0 0 1 2 0 ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Embarked: Factor w/ 4 levels "Southampton",..: 1 2 1 1 1 3 1 1 1 2 ...

skim(titanic)

Data summary

|  |  |
| --- | --- |
| Name | titanic |
| Number of rows | 891 |
| Number of columns | 8 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 4 |
| numeric | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| Survived | 0 | 1 | FALSE | 2 | No: 549, Yes: 342 |
| Pclass | 0 | 1 | FALSE | 3 | 3: 491, 1: 216, 2: 184 |
| Sex | 0 | 1 | FALSE | 2 | mal: 577, fem: 314 |
| Embarked | 0 | 1 | FALSE | 4 | Sou: 644, Che: 168, Que: 77, Unk: 2 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | 177 | 0.8 | 29.70 | 14.53 | 0.42 | 20.12 | 28.00 | 38 | 80.00 | ▂▇▅▂▁ |
| SibSp | 0 | 1.0 | 0.52 | 1.10 | 0.00 | 0.00 | 0.00 | 1 | 8.00 | ▇▁▁▁▁ |
| Parch | 0 | 1.0 | 0.38 | 0.81 | 0.00 | 0.00 | 0.00 | 0 | 6.00 | ▇▁▁▁▁ |
| Fare | 0 | 1.0 | 32.20 | 49.69 | 0.00 | 7.91 | 14.45 | 31 | 512.33 | ▇▁▁▁▁ |

What about Fare? There are some passengers in the dataset with a Fare of 0. How many and what does this mean?

titanic %>% filter(Fare == 0)

## Survived Pclass Sex Age SibSp Parch Fare Embarked  
## 1 No 3 male 36 0 0 0 Southampton  
## 2 No 1 male 40 0 0 0 Southampton  
## 3 Yes 3 male 25 0 0 0 Southampton  
## 4 No 2 male NA 0 0 0 Southampton  
## 5 No 3 male 19 0 0 0 Southampton  
## 6 No 2 male NA 0 0 0 Southampton  
## 7 No 2 male NA 0 0 0 Southampton  
## 8 No 2 male NA 0 0 0 Southampton  
## 9 No 3 male 49 0 0 0 Southampton  
## 10 No 1 male NA 0 0 0 Southampton  
## 11 No 2 male NA 0 0 0 Southampton  
## 12 No 2 male NA 0 0 0 Southampton  
## 13 No 1 male 39 0 0 0 Southampton  
## 14 No 1 male NA 0 0 0 Southampton  
## 15 No 1 male 38 0 0 0 Southampton

Looks like 15 passengers with a Fare of zero. Do we need to address this?

I’m tempted to replace the zeroes with “NA”. Let’s see how we’d do that. Not actually going to run the code.

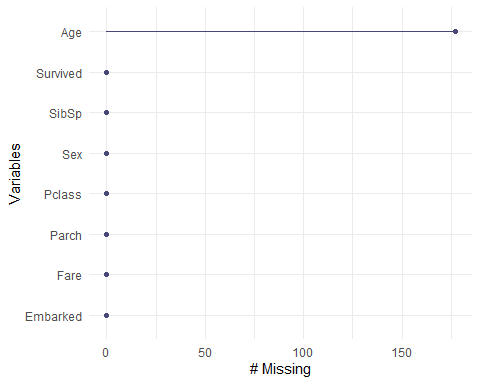
#titanic = titanic %>% mutate(Fare = na\_if(Fare, "0"))

Visualizing missingness. There are MANY ways to look at missingness in R. Typically we are interested in proportion of missingness by variable, by case (row), and by factor.

Here’s a helpful link: <https://cran.r-project.org/web/packages/naniar/vignettes/naniar-visualisation.html>. NOTE: If you have lots of variables and/or rows, these plots can quickly become cluttered.

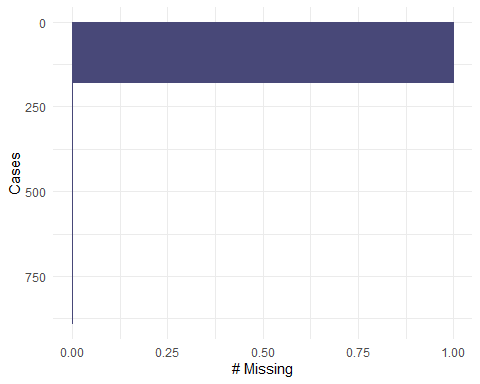
Simple view of missingess

gg\_miss\_var(titanic)



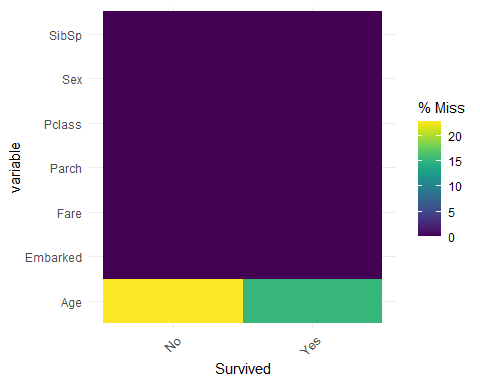
By case

gg\_miss\_case(titanic) #x axis is number of missing values in each row (case)



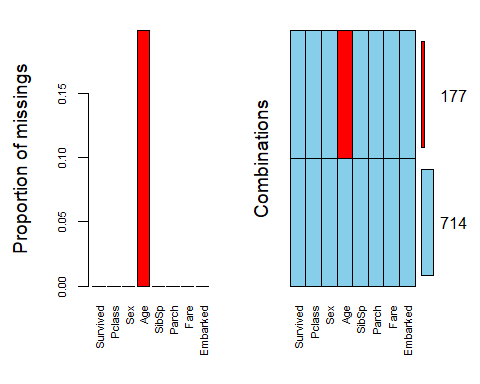
By a factor (here we choose the variable Survived)

gg\_miss\_fct(x = titanic, fct = Survived)



Looking at missingness by variable and combinations of missingness using “aggr” from VIM package.

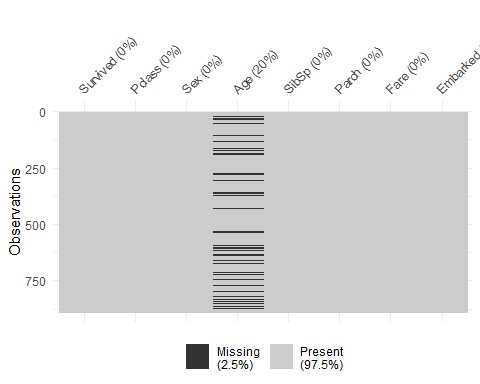
vim\_plot = aggr(titanic, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



#the cex.axis reduces size of text on x-axis so labels fit better

A view of missingness by variable and row.

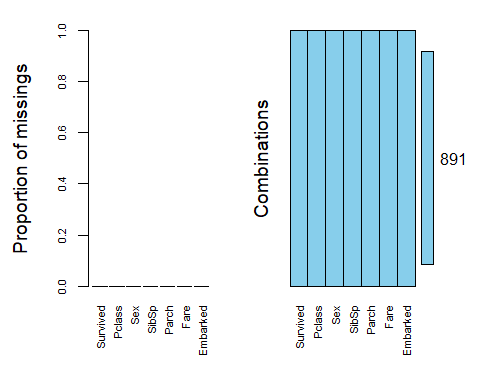
vis\_miss(titanic) #from the naniar package



So how do we do handle our missing data in this dataset?

One idea: Column-wise deletion of the “Age” variable (creating a new data frame with this variable removed).

titanic\_coldel = titanic %>% select(-Age)   
vim\_plot = aggr(titanic\_coldel, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



skim(titanic\_coldel)

Data summary

|  |  |
| --- | --- |
| Name | titanic\_coldel |
| Number of rows | 891 |
| Number of columns | 7 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 4 |
| numeric | 3 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| Survived | 0 | 1 | FALSE | 2 | No: 549, Yes: 342 |
| Pclass | 0 | 1 | FALSE | 3 | 3: 491, 1: 216, 2: 184 |
| Sex | 0 | 1 | FALSE | 2 | mal: 577, fem: 314 |
| Embarked | 0 | 1 | FALSE | 4 | Sou: 644, Che: 168, Que: 77, Unk: 2 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SibSp | 0 | 1 | 0.52 | 1.10 | 0 | 0.00 | 0.00 | 1 | 8.00 | ▇▁▁▁▁ |
| Parch | 0 | 1 | 0.38 | 0.81 | 0 | 0.00 | 0.00 | 0 | 6.00 | ▇▁▁▁▁ |
| Fare | 0 | 1 | 32.20 | 49.69 | 0 | 7.91 | 14.45 | 31 | 512.33 | ▇▁▁▁▁ |

Second idea: Row-wise deletion of any row with at least one NA:

titanic\_rowdel = titanic %>% drop\_na()   
#alternatively can specify which variable(s) on which to do the drop\_na  
#drop\_na(Age)  
skim(titanic\_rowdel)

Data summary

|  |  |
| --- | --- |
| Name | titanic\_rowdel |
| Number of rows | 714 |
| Number of columns | 8 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 4 |
| numeric | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| Survived | 0 | 1 | FALSE | 2 | No: 424, Yes: 290 |
| Pclass | 0 | 1 | FALSE | 3 | 3: 355, 1: 186, 2: 173 |
| Sex | 0 | 1 | FALSE | 2 | mal: 453, fem: 261 |
| Embarked | 0 | 1 | FALSE | 4 | Sou: 554, Che: 130, Que: 28, Unk: 2 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | 0 | 1 | 29.70 | 14.53 | 0.42 | 20.12 | 28.00 | 38.00 | 80.00 | ▂▇▅▂▁ |
| SibSp | 0 | 1 | 0.51 | 0.93 | 0.00 | 0.00 | 0.00 | 1.00 | 5.00 | ▇▁▁▁▁ |
| Parch | 0 | 1 | 0.43 | 0.85 | 0.00 | 0.00 | 0.00 | 1.00 | 6.00 | ▇▁▁▁▁ |
| Fare | 0 | 1 | 34.69 | 52.92 | 0.00 | 8.05 | 15.74 | 33.38 | 512.33 | ▇▁▁▁▁ |

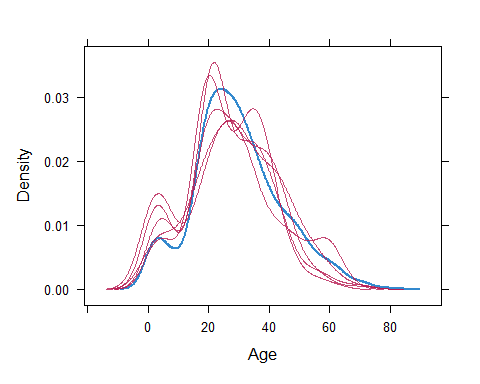
Doing this discards 177 rows of data.

Finally, imputation. Let’s use the “mice” package to do imputation.

set.seed(1234) #sets seed for random number generator  
imp\_age = mice(titanic, m=5, method='pmm', printFlag=FALSE)  
#m is the number of imputations, 5 is a reasonable value as a default  
#pmm is "predictive mean matching" = imputation method for numeric data  
#printFlag reduces amount of output  
summary(imp\_age)

## Class: mids  
## Number of multiple imputations: 5   
## Imputation methods:  
## Survived Pclass Sex Age SibSp Parch Fare Embarked   
## "" "" "" "pmm" "" "" "" ""   
## PredictorMatrix:  
## Survived Pclass Sex Age SibSp Parch Fare Embarked  
## Survived 0 1 1 1 1 1 1 1  
## Pclass 1 0 1 1 1 1 1 1  
## Sex 1 1 0 1 1 1 1 1  
## Age 1 1 1 0 1 1 1 1  
## SibSp 1 1 1 1 0 1 1 1  
## Parch 1 1 1 1 1 0 1 1

densityplot(imp\_age, ~Age) #red imputed, blue original

 Merge the imputed values into our titanic data frame

titanic\_complete = complete(imp\_age)   
summary(titanic\_complete)

## Survived Pclass Sex Age SibSp   
## No :549 1:216 female:314 Min. : 0.42 Min. :0.000   
## Yes:342 2:184 male :577 1st Qu.:20.00 1st Qu.:0.000   
## 3:491 Median :28.00 Median :0.000   
## Mean :29.02 Mean :0.523   
## 3rd Qu.:38.00 3rd Qu.:1.000   
## Max. :80.00 Max. :8.000   
## Parch Fare Embarked   
## Min. :0.0000 Min. : 0.00 Southampton:644   
## 1st Qu.:0.0000 1st Qu.: 7.91 Cherbourg :168   
## Median :0.0000 Median : 14.45 Queenstown : 77   
## Mean :0.3816 Mean : 32.20 Unknown : 2   
## 3rd Qu.:0.0000 3rd Qu.: 31.00   
## Max. :6.0000 Max. :512.33

Now we can continue with our modeling.

Looking at a different dataset.

heart = read\_csv("heart.csv")

## Rows: 4240 Columns: 16  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (16): male, age, education, currentSmoker, cigsPerDay, BPMeds, prevalent...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

str(heart)

## spc\_tbl\_ [4,240 × 16] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ male : num [1:4240] 1 0 1 0 0 0 0 0 1 1 ...  
## $ age : num [1:4240] 39 46 48 61 46 43 63 45 52 43 ...  
## $ education : num [1:4240] 4 2 1 3 3 2 1 2 1 1 ...  
## $ currentSmoker : num [1:4240] 0 0 1 1 1 0 0 1 0 1 ...  
## $ cigsPerDay : num [1:4240] 0 0 20 30 23 0 0 20 0 30 ...  
## $ BPMeds : num [1:4240] 0 0 0 0 0 0 0 0 0 0 ...  
## $ prevalentStroke: num [1:4240] 0 0 0 0 0 0 0 0 0 0 ...  
## $ prevalentHyp : num [1:4240] 0 0 0 1 0 1 0 0 1 1 ...  
## $ diabetes : num [1:4240] 0 0 0 0 0 0 0 0 0 0 ...  
## $ totChol : num [1:4240] 195 250 245 225 285 228 205 313 260 225 ...  
## $ sysBP : num [1:4240] 106 121 128 150 130 ...  
## $ diaBP : num [1:4240] 70 81 80 95 84 110 71 71 89 107 ...  
## $ BMI : num [1:4240] 27 28.7 25.3 28.6 23.1 ...  
## $ heartRate : num [1:4240] 80 95 75 65 85 77 60 79 76 93 ...  
## $ glucose : num [1:4240] 77 76 70 103 85 99 85 78 79 88 ...  
## $ TenYearCHD : num [1:4240] 0 0 0 1 0 0 1 0 0 0 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. male = col\_double(),  
## .. age = col\_double(),  
## .. education = col\_double(),  
## .. currentSmoker = col\_double(),  
## .. cigsPerDay = col\_double(),  
## .. BPMeds = col\_double(),  
## .. prevalentStroke = col\_double(),  
## .. prevalentHyp = col\_double(),  
## .. diabetes = col\_double(),  
## .. totChol = col\_double(),  
## .. sysBP = col\_double(),  
## .. diaBP = col\_double(),  
## .. BMI = col\_double(),  
## .. heartRate = col\_double(),  
## .. glucose = col\_double(),  
## .. TenYearCHD = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

summary(heart)

## male age education currentSmoker   
## Min. :0.0000 Min. :32.00 Min. :1.000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:42.00 1st Qu.:1.000 1st Qu.:0.0000   
## Median :0.0000 Median :49.00 Median :2.000 Median :0.0000   
## Mean :0.4292 Mean :49.58 Mean :1.979 Mean :0.4941   
## 3rd Qu.:1.0000 3rd Qu.:56.00 3rd Qu.:3.000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :70.00 Max. :4.000 Max. :1.0000   
## NA's :105   
## cigsPerDay BPMeds prevalentStroke prevalentHyp   
## Min. : 0.000 Min. :0.00000 Min. :0.000000 Min. :0.0000   
## 1st Qu.: 0.000 1st Qu.:0.00000 1st Qu.:0.000000 1st Qu.:0.0000   
## Median : 0.000 Median :0.00000 Median :0.000000 Median :0.0000   
## Mean : 9.006 Mean :0.02962 Mean :0.005896 Mean :0.3106   
## 3rd Qu.:20.000 3rd Qu.:0.00000 3rd Qu.:0.000000 3rd Qu.:1.0000   
## Max. :70.000 Max. :1.00000 Max. :1.000000 Max. :1.0000   
## NA's :29 NA's :53   
## diabetes totChol sysBP diaBP   
## Min. :0.00000 Min. :107.0 Min. : 83.5 Min. : 48.0   
## 1st Qu.:0.00000 1st Qu.:206.0 1st Qu.:117.0 1st Qu.: 75.0   
## Median :0.00000 Median :234.0 Median :128.0 Median : 82.0   
## Mean :0.02571 Mean :236.7 Mean :132.4 Mean : 82.9   
## 3rd Qu.:0.00000 3rd Qu.:263.0 3rd Qu.:144.0 3rd Qu.: 90.0   
## Max. :1.00000 Max. :696.0 Max. :295.0 Max. :142.5   
## NA's :50   
## BMI heartRate glucose TenYearCHD   
## Min. :15.54 Min. : 44.00 Min. : 40.00 Min. :0.0000   
## 1st Qu.:23.07 1st Qu.: 68.00 1st Qu.: 71.00 1st Qu.:0.0000   
## Median :25.40 Median : 75.00 Median : 78.00 Median :0.0000   
## Mean :25.80 Mean : 75.88 Mean : 81.96 Mean :0.1519   
## 3rd Qu.:28.04 3rd Qu.: 83.00 3rd Qu.: 87.00 3rd Qu.:0.0000   
## Max. :56.80 Max. :143.00 Max. :394.00 Max. :1.0000   
## NA's :19 NA's :1 NA's :388

skim(heart)

Data summary

|  |  |
| --- | --- |
| Name | heart |
| Number of rows | 4240 |
| Number of columns | 16 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 16 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| male | 0 | 1.00 | 0.43 | 0.50 | 0.00 | 0.00 | 0.0 | 1.00 | 1.0 | ▇▁▁▁▆ |
| age | 0 | 1.00 | 49.58 | 8.57 | 32.00 | 42.00 | 49.0 | 56.00 | 70.0 | ▃▇▆▆▂ |
| education | 105 | 0.98 | 1.98 | 1.02 | 1.00 | 1.00 | 2.0 | 3.00 | 4.0 | ▇▆▁▃▂ |
| currentSmoker | 0 | 1.00 | 0.49 | 0.50 | 0.00 | 0.00 | 0.0 | 1.00 | 1.0 | ▇▁▁▁▇ |
| cigsPerDay | 29 | 0.99 | 9.01 | 11.92 | 0.00 | 0.00 | 0.0 | 20.00 | 70.0 | ▇▃▁▁▁ |
| BPMeds | 53 | 0.99 | 0.03 | 0.17 | 0.00 | 0.00 | 0.0 | 0.00 | 1.0 | ▇▁▁▁▁ |
| prevalentStroke | 0 | 1.00 | 0.01 | 0.08 | 0.00 | 0.00 | 0.0 | 0.00 | 1.0 | ▇▁▁▁▁ |
| prevalentHyp | 0 | 1.00 | 0.31 | 0.46 | 0.00 | 0.00 | 0.0 | 1.00 | 1.0 | ▇▁▁▁▃ |
| diabetes | 0 | 1.00 | 0.03 | 0.16 | 0.00 | 0.00 | 0.0 | 0.00 | 1.0 | ▇▁▁▁▁ |
| totChol | 50 | 0.99 | 236.70 | 44.59 | 107.00 | 206.00 | 234.0 | 263.00 | 696.0 | ▆▇▁▁▁ |
| sysBP | 0 | 1.00 | 132.35 | 22.03 | 83.50 | 117.00 | 128.0 | 144.00 | 295.0 | ▇▇▁▁▁ |
| diaBP | 0 | 1.00 | 82.90 | 11.91 | 48.00 | 75.00 | 82.0 | 90.00 | 142.5 | ▁▇▅▁▁ |
| BMI | 19 | 1.00 | 25.80 | 4.08 | 15.54 | 23.07 | 25.4 | 28.04 | 56.8 | ▅▇▁▁▁ |
| heartRate | 1 | 1.00 | 75.88 | 12.03 | 44.00 | 68.00 | 75.0 | 83.00 | 143.0 | ▂▇▃▁▁ |
| glucose | 388 | 0.91 | 81.96 | 23.95 | 40.00 | 71.00 | 78.0 | 87.00 | 394.0 | ▇▁▁▁▁ |
| TenYearCHD | 0 | 1.00 | 0.15 | 0.36 | 0.00 | 0.00 | 0.0 | 0.00 | 1.0 | ▇▁▁▁▂ |

#factor conversions  
heart = heart %>% mutate(male = as\_factor(male)) %>%  
 mutate(education = as\_factor(education)) %>%  
 mutate(currentSmoker = as\_factor(currentSmoker)) %>%  
 mutate(BPMeds = as\_factor(BPMeds)) %>%  
 mutate(prevalentStroke = as\_factor(prevalentStroke)) %>%  
 mutate(prevalentHyp = as\_factor(prevalentHyp)) %>%  
 mutate(diabetes = as\_factor(diabetes)) %>%  
 mutate(TenYearCHD = as\_factor(TenYearCHD))  
  
#recode variables  
heart = heart %>% mutate(male = fct\_recode(male, "Yes" = "1","No"="0")) %>%   
 mutate(currentSmoker = fct\_recode(currentSmoker, "YesSmokes"="1","NoSmokes"="0")) %>%  
 mutate(BPMeds = fct\_recode(BPMeds, "YesBPMeds"="1","NoBPMeds"="0")) %>%  
 mutate(prevalentStroke = fct\_recode(prevalentStroke, "YesStroke"="1","NoStroke"="0")) %>%  
 mutate(prevalentHyp = fct\_recode(prevalentHyp, "YesHyp"="1","NoHyp"="0")) %>%  
 mutate(diabetes = fct\_recode(diabetes, "YesDiabetes"="1","NoDiabetes"="0")) %>%  
 mutate(TenYearCHD = fct\_recode(TenYearCHD, "YesTenYearCHD"="1","NoTenYearCHD"="0")) %>%  
 mutate(education = fct\_recode(education, "Some HS"="1","HS"="2","Some College"="3","College or More"="4"))

skim(heart)

Data summary

|  |  |
| --- | --- |
| Name | heart |
| Number of rows | 4240 |
| Number of columns | 16 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 8 |
| numeric | 8 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

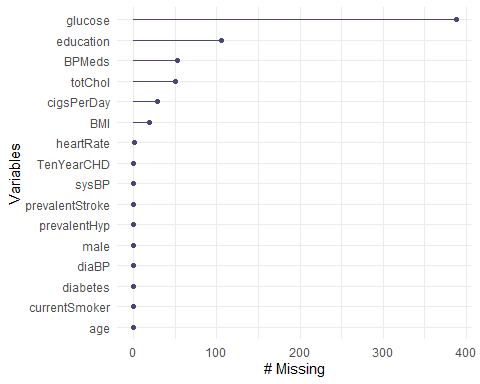
| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| male | 0 | 1.00 | FALSE | 2 | No: 2420, Yes: 1820 |
| education | 105 | 0.98 | FALSE | 4 | Som: 1720, HS: 1253, Som: 689, Col: 473 |
| currentSmoker | 0 | 1.00 | FALSE | 2 | NoS: 2145, Yes: 2095 |
| BPMeds | 53 | 0.99 | FALSE | 2 | NoB: 4063, Yes: 124 |
| prevalentStroke | 0 | 1.00 | FALSE | 2 | NoS: 4215, Yes: 25 |
| prevalentHyp | 0 | 1.00 | FALSE | 2 | NoH: 2923, Yes: 1317 |
| diabetes | 0 | 1.00 | FALSE | 2 | NoD: 4131, Yes: 109 |
| TenYearCHD | 0 | 1.00 | FALSE | 2 | NoT: 3596, Yes: 644 |

**Variable type: numeric**

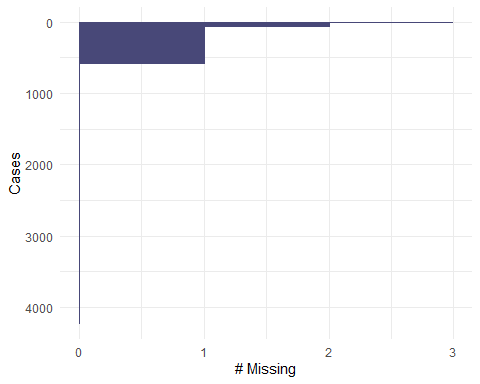
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | 0 | 1.00 | 49.58 | 8.57 | 32.00 | 42.00 | 49.0 | 56.00 | 70.0 | ▃▇▆▆▂ |
| cigsPerDay | 29 | 0.99 | 9.01 | 11.92 | 0.00 | 0.00 | 0.0 | 20.00 | 70.0 | ▇▃▁▁▁ |
| totChol | 50 | 0.99 | 236.70 | 44.59 | 107.00 | 206.00 | 234.0 | 263.00 | 696.0 | ▆▇▁▁▁ |
| sysBP | 0 | 1.00 | 132.35 | 22.03 | 83.50 | 117.00 | 128.0 | 144.00 | 295.0 | ▇▇▁▁▁ |
| diaBP | 0 | 1.00 | 82.90 | 11.91 | 48.00 | 75.00 | 82.0 | 90.00 | 142.5 | ▁▇▅▁▁ |
| BMI | 19 | 1.00 | 25.80 | 4.08 | 15.54 | 23.07 | 25.4 | 28.04 | 56.8 | ▅▇▁▁▁ |
| heartRate | 1 | 1.00 | 75.88 | 12.03 | 44.00 | 68.00 | 75.0 | 83.00 | 143.0 | ▂▇▃▁▁ |
| glucose | 388 | 0.91 | 81.96 | 23.95 | 40.00 | 71.00 | 78.0 | 87.00 | 394.0 | ▇▁▁▁▁ |

Visualization

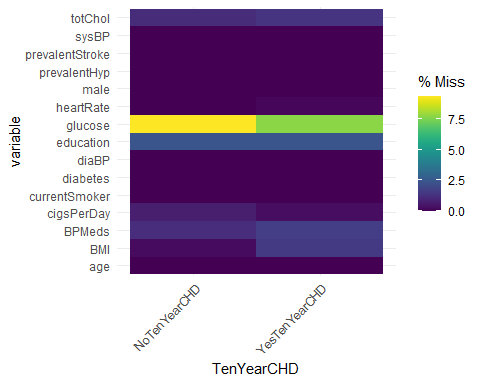
gg\_miss\_var(heart)

 By case

gg\_miss\_case(heart) #x axis is number of missing values in each row (case)

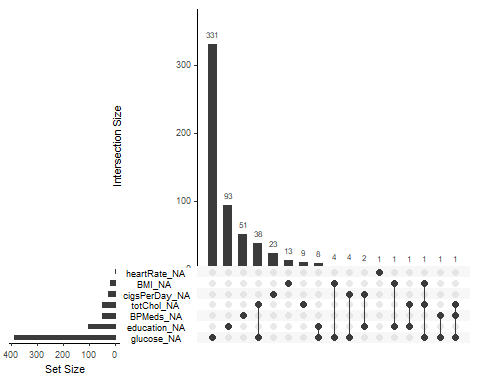
 By a factor

gg\_miss\_fct(x = heart, fct = TenYearCHD)



A view of patterns of missingness

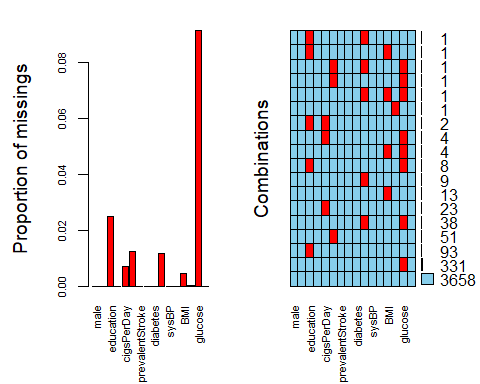
gg\_miss\_upset(heart, nsets = 7) #from the UpSetR package, must have at least two variables with missingness to use this plot



#note nsets = 7 refers to then number of variables to show in the plot. I chose 7 as there are 7 variables with missingness

Not a lot (barely any) patterned missigness.

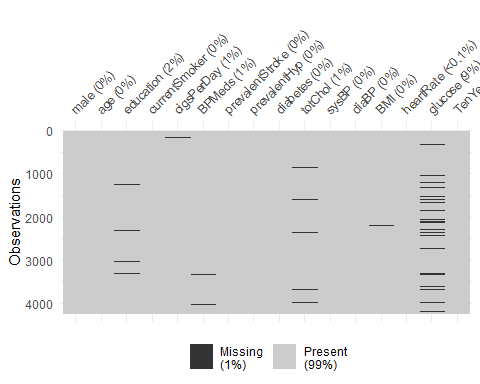
vim\_plot = aggr(heart, numbers = TRUE, prop = c(TRUE, FALSE),cex.axis=.7)



#the cex.axis reduces size of text on x-axis so labels fit better

A view of missingness by variable and row.

vis\_miss(heart) #from the naniar package

 So what’s our strategy here? It’s a bit of “artistry” as there is no definitive “right” way. With the exception of “education” and “BPMeds” the missingness is confined to numeric variables. Numeric variables tend to make good candidates for imputation (although we can impute categorical variables too).

A question: Should we impute health data at all? My temptation is to do row-wise deletion and call it a day. That would leave us with 3,658 rows.