library(titanic)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.8 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.1  
## ✔ readr 2.1.2 ✔ forcats 0.5.2  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(tidymodels)

## ── Attaching packages ────────────────────────────────────── tidymodels 1.0.0 ──  
## ✔ broom 1.0.1 ✔ rsample 1.1.0  
## ✔ dials 1.0.0 ✔ tune 1.0.0  
## ✔ infer 1.0.3 ✔ workflows 1.0.0  
## ✔ modeldata 1.0.1 ✔ workflowsets 1.0.0  
## ✔ parsnip 1.0.1 ✔ yardstick 1.0.0  
## ✔ recipes 1.0.1   
## ── 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) #for imputations

##   
## 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  
## 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) #view dataset summaries

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

library(UpSetR) #visualizing missingness

titanic <-titanic::titanic\_train

summarise(titanic)

## data frame with 0 columns and 1 row

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" ...

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 | ▇▁▁▁▁ |

titanic <- select(titanic, c(Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked))

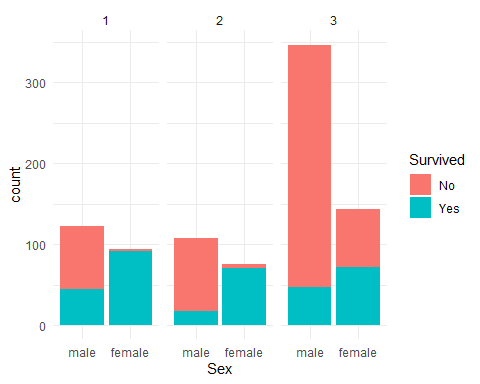
titanic <- titanic %>% mutate\_if(is.character, as\_factor)  
#str(titanic)  
  
titanic <- titanic %>% mutate\_if(is.integer, as\_factor)  
str(titanic)

## 'data.frame': 891 obs. of 8 variables:  
## $ Survived: Factor w/ 2 levels "0","1": 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 "male","female": 1 2 2 2 1 1 1 1 2 2 ...  
## $ Age : num 22 38 26 35 35 NA 54 2 27 14 ...  
## $ SibSp : Factor w/ 7 levels "0","1","2","3",..: 2 2 1 2 1 1 1 4 1 2 ...  
## $ Parch : Factor w/ 7 levels "0","1","2","3",..: 1 1 1 1 1 1 1 2 3 1 ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Embarked: Factor w/ 4 levels "S","C","Q","": 1 2 1 1 1 3 1 1 1 2 ...

titanic <- titanic %>%  
 mutate(Survived = fct\_recode(Survived, "No"= "0", "Yes" = "1")) %>%  
 mutate(Embarked = fct\_recode(Embarked, "Unknown"= "", "Cherbourg" = "C", "Southampton" ="S", "Queenstown"= "Q"))

library(esquisse)

#esquisser()  
library(ggplot2)  
  
ggplot(titanic) +  
 aes(x = Sex, fill = Survived) +  
 geom\_bar() +  
 scale\_fill\_hue(direction = 1) +  
 theme\_minimal() +  
 facet\_wrap(vars(Pclass))



summary(titanic)

## Survived Pclass Sex Age SibSp Parch   
## No :549 1:216 male :577 Min. : 0.42 0:608 0:678   
## Yes:342 2:184 female:314 1st Qu.:20.12 1:209 1:118   
## 3:491 Median :28.00 2: 28 2: 80   
## Mean :29.70 3: 16 3: 5   
## 3rd Qu.:38.00 4: 18 4: 4   
## Max. :80.00 5: 5 5: 5   
## NA's :177 8: 7 6: 1   
## Fare Embarked   
## Min. : 0.00 Southampton:644   
## 1st Qu.: 7.91 Cherbourg :168   
## Median : 14.45 Queenstown : 77   
## Mean : 32.20 Unknown : 2   
## 3rd Qu.: 31.00   
## Max. :512.33   
##

#How many have 0.00 Fare?

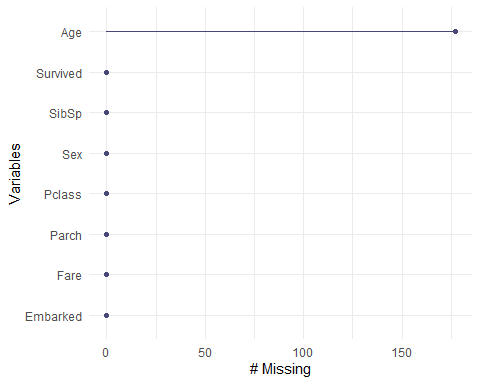
titanic %>% filter(Fare == 0)

## 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

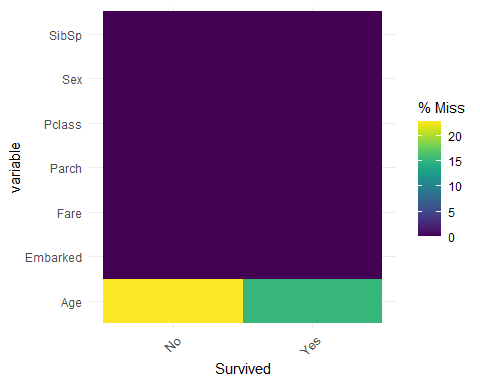
#Simple view of Missingness

gg\_miss\_var(titanic)

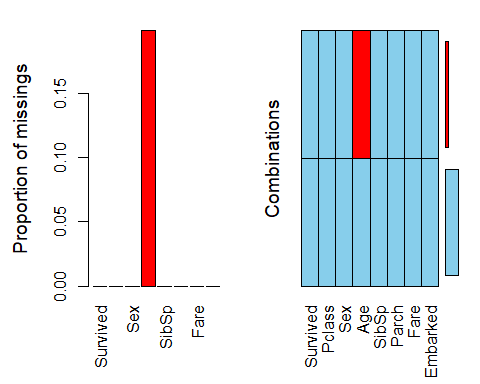
## Warning: It is deprecated to specify `guide = FALSE` to remove a guide. Please  
## use `guide = "none"` instead.



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

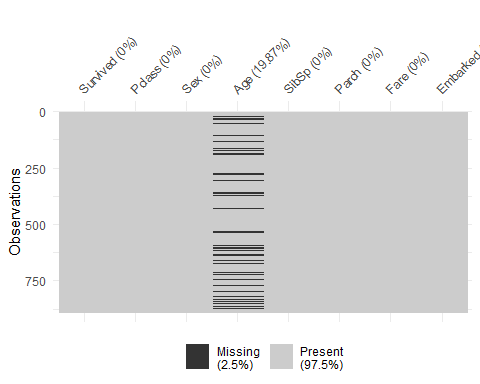


Vim\_plot = aggr(titanic)



vis\_miss(titanic)

## Warning: `gather\_()` was deprecated in tidyr 1.2.0.  
## Please use `gather()` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was generated.



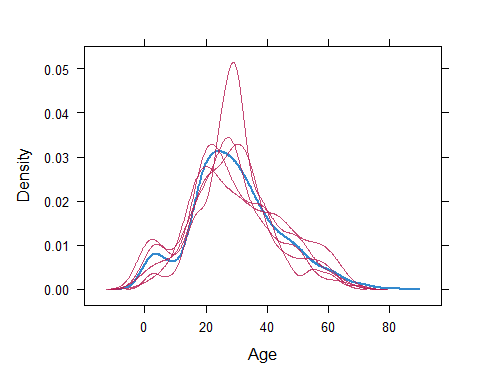
set.seed(1234)  
imp\_age <- mice(titanic, m=5, method = 'pmm', printFlag = FALSE)

## Warning: Number of logged events: 25

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  
## Number of logged events: 25   
## it im dep meth out  
## 1 1 1 Age pmm SibSp8  
## 2 1 2 Age pmm SibSp8  
## 3 1 3 Age pmm SibSp8  
## 4 1 4 Age pmm SibSp8  
## 5 1 5 Age pmm SibSp8  
## 6 2 1 Age pmm SibSp8

densityplot(imp\_age, ~Age)

 #Blue line are the original values and red lines are imputed values

titanic\_complete <- complete(imp\_age)  
summary(titanic\_complete)

## Survived Pclass Sex Age SibSp Parch   
## No :549 1:216 male :577 Min. : 0.42 0:608 0:678   
## Yes:342 2:184 female:314 1st Qu.:20.75 1:209 1:118   
## 3:491 Median :28.00 2: 28 2: 80   
## Mean :29.66 3: 16 3: 5   
## 3rd Qu.:39.00 4: 18 4: 4   
## Max. :80.00 5: 5 5: 5   
## 8: 7 6: 1   
## Fare Embarked   
## Min. : 0.00 Southampton:644   
## 1st Qu.: 7.91 Cherbourg :168   
## Median : 14.45 Queenstown : 77   
## Mean : 32.20 Unknown : 2   
## 3rd Qu.: 31.00   
## Max. :512.33   
##

Vim\_plot = aggr(titanic\_complete)

