Imputacion Datos Missing Ébola

Silvia Pineda

Con la misma base de datos de la epidemia de ébola haz los siguientes ejercicios

Lectura de datos

```
library(naniar)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
load("linelist.Rdata")
head(data)
```

```
  case_id date_infection date_onset date_hospitalisation date_outcome outcome
1  5fe599     2014-05-08 2014-05-13           2014-05-15         <NA>    <NA>
2  8689b7           <NA> 2014-05-13           2014-05-14   2014-05-18 Recover
3  11f8ea           <NA> 2014-05-16           2014-05-18   2014-05-30 Recover
4  b8812a     2014-05-04 2014-05-18           2014-05-20         <NA>    <NA>
5  893f25     2014-05-18 2014-05-21           2014-05-22   2014-05-29 Recover
6  be99c8     2014-05-03 2014-05-22           2014-05-23   2014-05-24 Recover
```

```
  gender age age_cat age_cat5                              hospital       lon
1      m   2     0-4      0-4                                 Other -13.21574
2      f   3     0-4      0-4                                  <NA> -13.21523
3      m  56   50-69    55-59 St. Mark's Maternity Hospital (SMMH) -13.21291
4      f  18   15-19    15-19                          Port Hospital -13.23637
5      m   3     0-4      0-4                      Military Hospital -13.22286
6      f  16   15-19    15-19                          Port Hospital -13.22263
       lat infector ct_blood fever chills cough aches vomit temp time_admission
1 8.468973   f547d6       22    no     no   yes    no   yes 36.8           <NA>
2 8.451719     <NA>       22  <NA>   <NA>  <NA>  <NA>  <NA> 36.9          09:36
3 8.464817     <NA>       21  <NA>   <NA>  <NA>  <NA>  <NA> 36.9          16:48
4 8.475476   f90f5f       23    no     no    no    no    no 36.8          11:22
5 8.460824   11f8ea       23    no     no   yes    no   yes 36.9          12:60
6 8.461831   aec8ec       21    no     no   yes    no   yes 37.6          14:13
  days_onset_hosp
1               2
2               1
3               2
4               2
5               1
6               1
```

```
str(data)
```

```
'data.frame':   5888 obs. of  23 variables:
 $ case_id             : chr  "5fe599" "8689b7" "11f8ea" "b8812a" ...
 $ date_infection      : Date, format: "2014-05-08" NA ...
 $ date_onset          : Date, format: "2014-05-13" "2014-05-13" ...
 $ date_hospitalisation: Date, format: "2014-05-15" "2014-05-14" ...
 $ date_outcome        : Date, format: NA "2014-05-18" ...
 $ outcome             : Factor w/ 2 levels "Death","Recover": NA 2 2 NA 2 2 2 1 2 1 ...
 $ gender              : Factor w/ 2 levels "f","m": 2 1 2 1 2 1 1 1 2 1 ...
 $ age                 : num  2 3 56 18 3 16 16 0 61 27 ...
 $ age_cat             : Factor w/ 8 levels "0-4","10-14",..: 1 1 7 3 1 3 3 1 7 4 ...
 $ age_cat5            : Factor w/ 17 levels "0-4","10-14",..: 1 1 12 3 1 3 3 1 13 5 ...
 $ hospital            : Factor w/ 5 levels "Central Hospital",..: 3 NA 5 4 2 4 NA NA NA NA
 $ lon                 : num  -13.2 -13.2 -13.2 -13.2 -13.2 ...
 $ lat                 : num  8.47 8.45 8.46 8.48 8.46 ...
 $ infector            : chr  "f547d6" NA NA "f90f5f" ...
 $ ct_blood            : int  22 22 21 23 23 21 21 22 22 22 ...
 $ fever               : Factor w/ 2 levels "no","yes": 1 NA NA 1 1 1 NA 1 1 1 ...
 $ chills              : Factor w/ 2 levels "no","yes": 1 NA NA 1 1 1 NA 1 1 1 ...
```

```
$ cough               : Factor w/ 2 levels "no","yes": 2 NA NA 1 2 2 NA 2 2 2 ...
$ aches               : Factor w/ 2 levels "no","yes": 1 NA NA 1 1 1 NA 1 1 1 ...
$ vomit               : Factor w/ 2 levels "no","yes": 2 NA NA 1 2 2 NA 2 2 1 ...
$ temp                : num  36.8 36.9 36.9 36.8 36.9 37.6 37.3 37 36.4 35.9 ...
$ time_admission      : chr  NA "09:36" "16:48" "11:22" ...
$ days_onset_hosp     : int  2 1 2 2 1 1 2 1 1 2 ...
```

```
summary(data)
```

```
  case_id          date_infection         date_onset
Length:5888        Min.   :2014-03-19   Min.   :2014-04-07
Class :character   1st Qu.:2014-09-06   1st Qu.:2014-09-16
Mode  :character   Median :2014-10-11   Median :2014-10-23
                   Mean   :2014-10-22   Mean   :2014-11-03
                   3rd Qu.:2014-12-05   3rd Qu.:2014-12-19
                   Max.   :2015-04-27   Max.   :2015-04-30
                   NA's   :2087         NA's   :256
date_hospitalisation  date_outcome           outcome      gender
Min.   :2014-04-17   Min.   :2014-04-19   Death  :2582   f   :2807
1st Qu.:2014-09-19   1st Qu.:2014-09-26   Recover:1983   m   :2803
Median :2014-10-23   Median :2014-11-01   NA's   :1323   NA's: 278
Mean   :2014-11-03   Mean   :2014-11-12
3rd Qu.:2014-12-17   3rd Qu.:2014-12-28
Max.   :2015-04-30   Max.   :2015-06-04
                     NA's   :936
     age           age_cat        age_cat5
Min.   : 0.00   0-4    :1095   0-4    :1095
1st Qu.: 6.00   5-9    :1095   5-9    :1095
Median :13.00   20-29  :1073   10-14  : 941
Mean   :16.01   10-14  : 941   15-19  : 743
3rd Qu.:23.00   30-49  : 754   20-24  : 638
Max.   :84.00   (Other): 844   (Other):1290
NA's   :85      NA's   :  86   NA's   :  86
                                  hospital         lon              lat
Central Hospital                     : 454   Min.   :-13.27   Min.   :8.446
Military Hospital                    : 896   1st Qu.:-13.25   1st Qu.:8.461
Other                                : 885   Median :-13.23   Median :8.469
Port Hospital                        :1762   Mean   :-13.23   Mean   :8.470
St. Mark's Maternity Hospital (SMMH): 422   3rd Qu.:-13.22   3rd Qu.:8.480
NA's                                 :1469   Max.   :-13.21   Max.   :8.492

  infector            ct_blood      fever       chills      cough
```

```
 Length:5888        Min.   :16.00   no  :1090   no  :4540   no  : 773
 Class :character   1st Qu.:20.00   yes :4549   yes :1099   yes :4866
 Mode  :character   Median :22.00   NA's: 249   NA's: 249   NA's: 249
                    Mean   :21.21
                    3rd Qu.:22.00
                    Max.   :26.00

  aches       vomit           temp       time_admission     days_onset_hosp
 no  :5095   no  :2836   Min.   :35.20   Length:5888        Min.   : 0.000
 yes : 544   yes :2803   1st Qu.:38.20   Class :character   1st Qu.: 1.000
 NA's: 249   NA's: 249   Median :38.80   Mode  :character   Median : 1.000
                         Mean   :38.56                      Mean   : 2.059
                         3rd Qu.:39.20                      3rd Qu.: 3.000
                         Max.   :40.80                      Max.   :22.000
                         NA's   :149                        NA's   :256
```

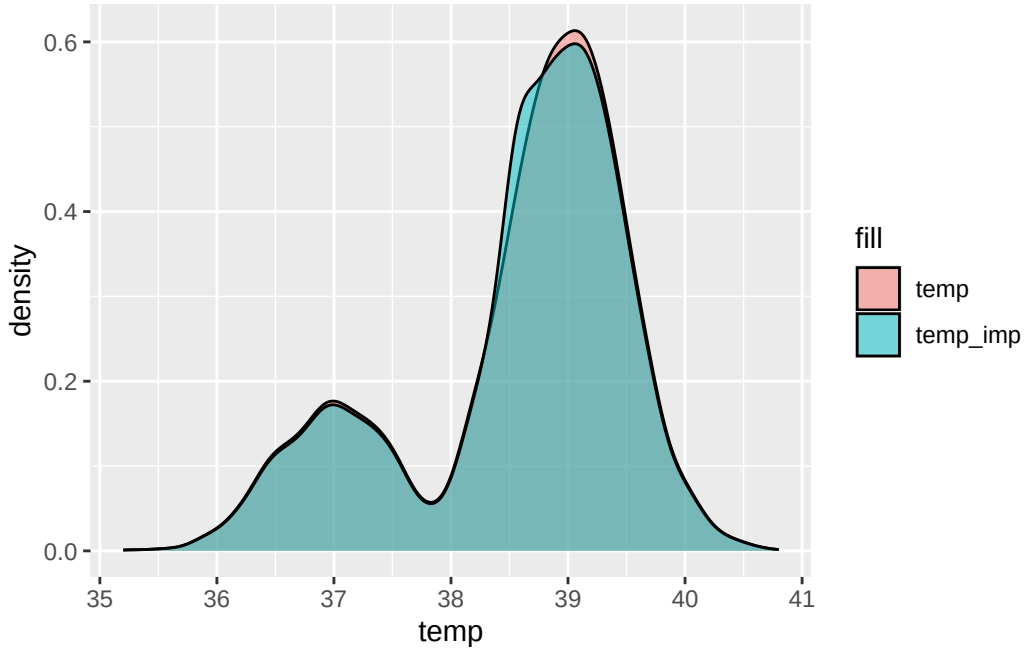
1. Imputa la variable temp con las 3 formas que hemos visto en la teoría y comparalas.

```
###MEDIA
data$temp_imp<-data$temp
mean(data$temp_imp, na.rm = TRUE)
```

```
[1] 38.55829
```

```
data$temp_imp[is.na(data$temp_imp)] <- mean(data$temp_imp, na.rm = TRUE)

ggplot(data, aes(x = temp, fill = "temp")) +
  geom_density(alpha = 0.5) +
  geom_density(aes(x = temp_imp, fill = "temp_imp"), alpha = 0.5)
```

```
Warning: Removed 149 rows containing non-finite outside the scale range
(`stat_density()`).
```



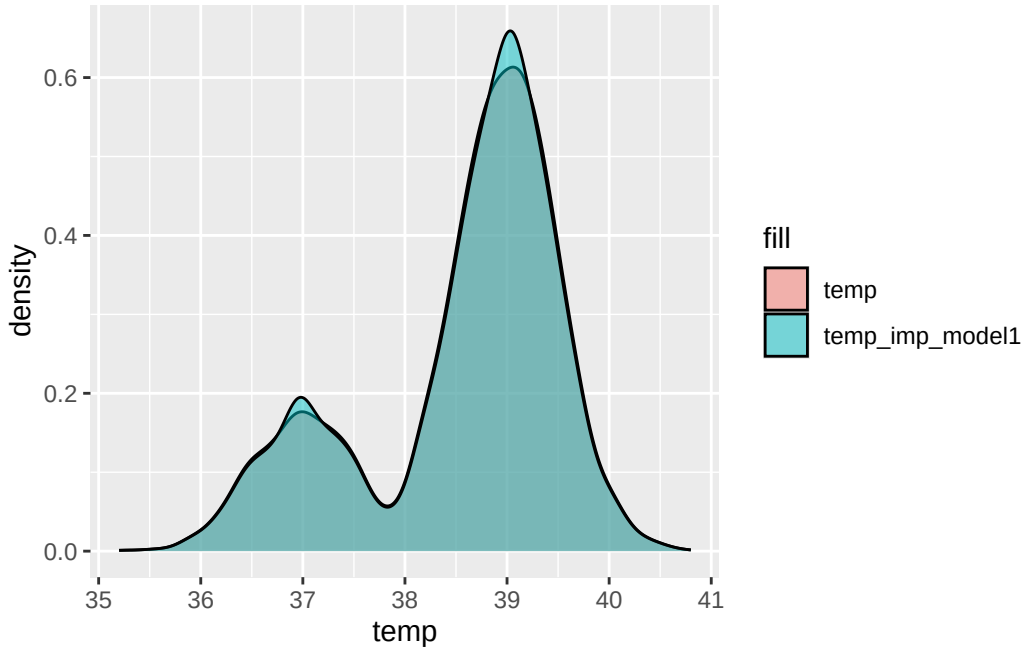

```r
## MODELO REGRESIóN

# ajustar un modelo de regresión lineal de temperatura ~ fiebre
model1 <- lm(temp ~ fever, data = data)
summary(model1)
```

```
Call:
lm(formula = temp ~ fever, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.77314 -0.32113 -0.02113  0.32686  1.77887

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.97314    0.01451  2548.3   <2e-16 ***
feveryes     2.04798    0.01613   126.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4701 on 5488 degrees of freedom
```

```
  (398 observations deleted due to missingness)
Multiple R-squared:  0.7459,    Adjusted R-squared:  0.7459
F-statistic: 1.611e+04 on 1 and 5488 DF,  p-value: < 2.2e-16
```

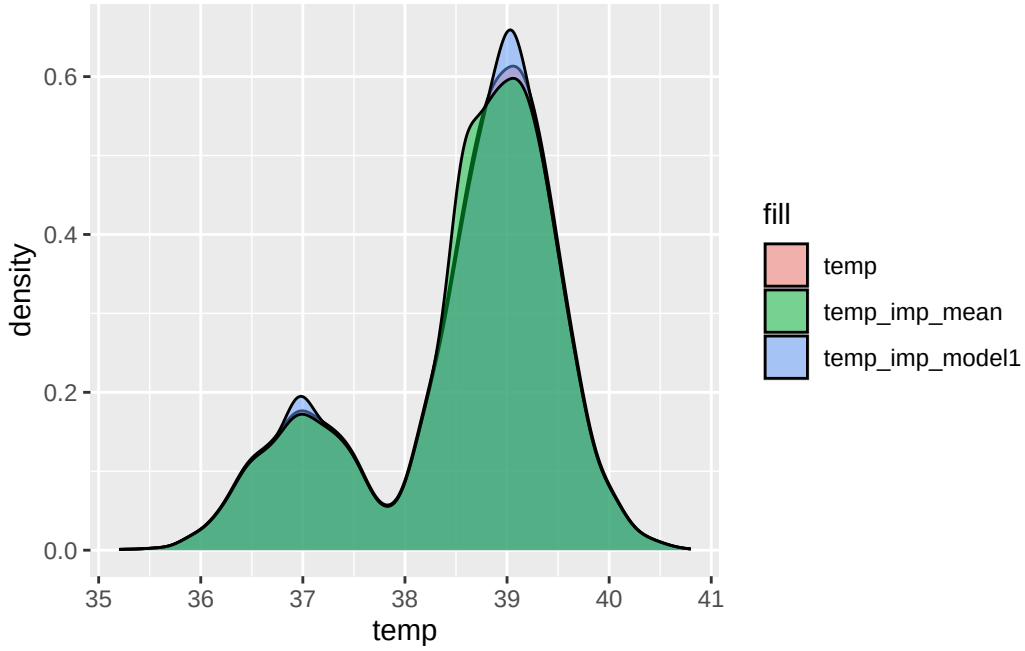
```r
#predecir los valores solo para las observaciones faltantes
predictions <- predict(model1,newdata = data [is.na(data$temp),])


## Crear una nueva variable de linelist con la temperatura imputada
data$temp_imp_model1 <- data$temp
data$temp_imp_model1[is.na(data$temp)]<- predictions



# Hacer un gráfico para comparar las observaciones
ggplot(data, aes(x = temp, fill = "temp")) +
  geom_density(alpha = 0.5) +
  geom_density(aes(x = temp_imp_model1, fill = "temp_imp_model1"), alpha = 0.5)
```

```
Warning: Removed 149 rows containing non-finite outside the scale range
(`stat_density()`).
```



```r
# Hacer un gráfico para comparar las observaciones con la media y la regresión
ggplot(data, aes(x = temp, fill = "temp")) +
  geom_density(alpha = 0.5) +
  geom_density(aes(x = temp_imp_model1, fill = "temp_imp_model1"), alpha = 0.5) +
  geom_density(aes(x = temp_imp, fill = "temp_imp_mean"), alpha = 0.5)
```

```
Warning: Removed 149 rows containing non-finite outside the scale range
(`stat_density()`).
```




```r
##Regresión con incertidumbre##.

#*Podemos regresar la incertidumbre a las imputaciones sumando el error de predicción.
# La idea es simular observaciones bajo el modelo:
summary(model1) ##Cogemos el residual standard error
```

```
Call:
lm(formula = temp ~ fever, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-1.77314 -0.32113 -0.02113  0.32686  1.77887

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.97314    0.01451  2548.3   <2e-16 ***
feveryes     2.04798    0.01613   126.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4701 on 5488 degrees of freedom
  (398 observations deleted due to missingness)
Multiple R-squared:  0.7459,    Adjusted R-squared:  0.7459
F-statistic: 1.611e+04 on 1 and 5488 DF,  p-value: < 2.2e-16
```

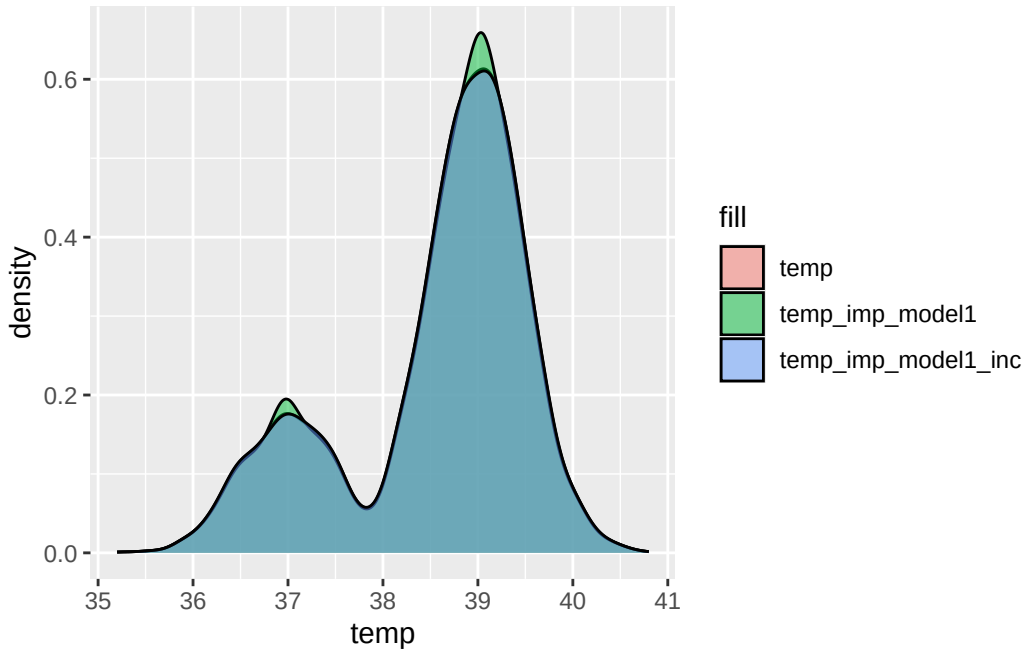
```r
set.seed(3)
rnorm(sum(is.na(data$temp)), 0, sd = 0.47)
```

```
  [1] -0.452108705 -0.137487090  0.121630462 -0.541501986  0.092017928
  [6]  0.014158254  0.040146334  0.524806800 -0.572862985  0.595663299
 [11] -0.350047350 -0.531672728 -0.336688490  0.118746614  0.071461482
 [16] -0.144598522 -0.447918146 -0.304674121  0.575427403  0.093911456
 [21] -0.271887349 -0.442881345 -0.095752244 -0.783243175 -0.227693901
 [26] -0.348304151  0.545489416  0.475671549 -0.033876883 -0.534287680
 [31]  0.423293623  0.400332110  0.342026132  0.346156008 -0.165500920
 [36]  0.331592291  0.611168255  0.017978447 -0.460263372  0.373067779
 [41]  0.369658230 -0.145917672  0.798475877 -0.373459043  0.163765727
 [46] -1.064738505 -0.076236481  0.531506546 -0.214106609 -0.422608168
 [51]  0.341614284 -0.380437224  0.125530004 -0.816513944 -0.663369814
 [56] -0.213169077 -0.486680899  0.640207160  0.431204666 -0.369016816
 [61]  0.269553541  0.431552218  0.120455018  0.165424281  0.551938558
 [66] -0.225997796 -0.196849969  0.448903018 -0.605833107  0.087512794
 [71] -0.014722986  0.219535736  0.481372907  0.125658473  0.108958268
 [76]  0.351368458  0.572022200  0.180178422 -0.464384826 -0.073720868
 [81]  0.815701552 -0.165580204  0.323660821  0.575470865  0.373319263
 [86] -0.003009127  0.103000799 -0.416637963  0.206687337 -0.416603183
 [91] -0.401294674 -0.465297335 -0.305912536  0.495354930 -0.183712676
 [96] -0.033175605 -0.217163880  0.254226885  0.437868436 -0.098358942
[101]  0.290154523 -0.190386431  0.494958769  0.283073596  0.478206753
[106]  0.285838639  0.097165918 -0.891931827 -0.320813931  0.226229055
[111] -0.217624588 -0.131478597 -0.194434368  0.760820265 -0.338896184
[116] -0.212953784  0.006700866  0.101409372  0.088768992 -0.023569792
[121] -0.702847227  0.172883744  0.243057690 -0.227637669  0.317182139
```

```
[126] -0.358350844  0.181454682 -0.312081572 -0.810441772  0.543469967
[131]  0.325948097  0.067283487  0.701622373 -0.767112135  0.060087631
[136] -1.129721952  0.678646280 -0.413079734 -0.614026014 -0.412283527
[141] -0.547258816 -0.931703410 -0.465273790 -0.071291763  0.428878192
[146]  0.191604810 -0.583826656 -0.302066374  0.907214534
```

```r
data$temp_imp_model1_inc <- data$temp
data$temp_imp_model1_inc[is.na(data$temp)]<- predictions + rnorm(sum(is.na(data$temp)), 0, sd

# Hacer un gráfico para comparar las observaciones con la media y la regresión
ggplot(data, aes(x = temp, fill = "temp")) +
  geom_density(alpha = 0.5) +
  geom_density(aes(x = temp_imp_model1, fill = "temp_imp_model1"), alpha = 0.5) +
  geom_density(aes(x = temp_imp_model1_inc, fill = "temp_imp_model1_inc"), alpha = 0.5)
```

```
Warning: Removed 149 rows containing non-finite outside the scale range
(`stat_density()`).
```

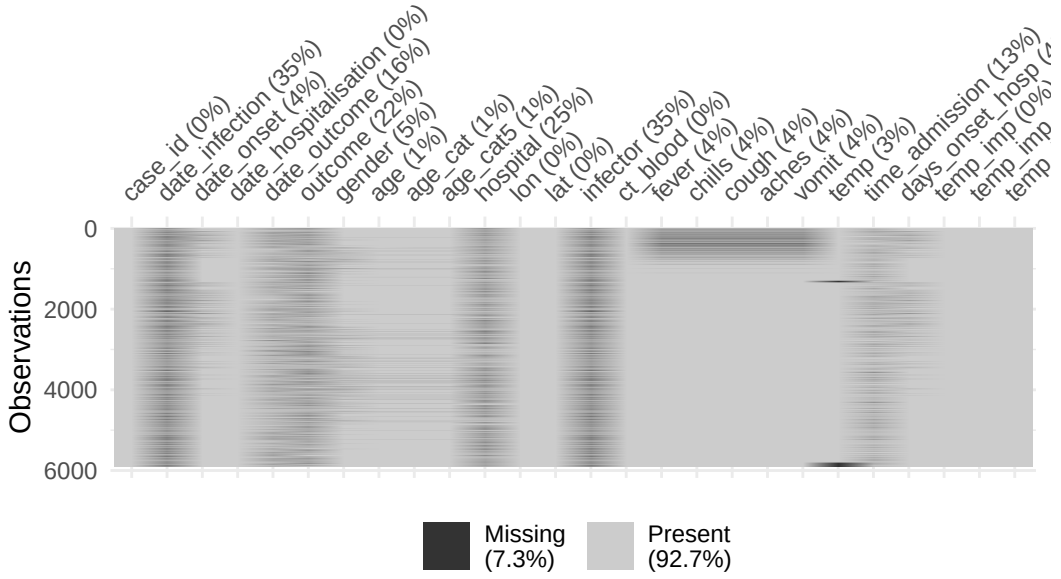


En este caso vemos como la imputación por la media nos imputa todos los individuos por el valor 38.5 asumiento que todos los faltantes tienen fiebre, pero esto no tiene por qué ser así, aprovechamos que todos los individuos a los que les falta la variable temp, tienen dato en fever, así que podemos hacer un modelo de regresión y usar ese modelo para predecir los

datos faltantes. Como la variable fiebre es dicotómica, cuando esta sea 0 (no fiebre), imputará por el alpha y cuando esta sea 1 (si fiebre) imputará por el alpha + la estimación de la beta. Para terminar de hacer una buena imputación, a las predicciones de este modelo le podemos añadir una incertidumbre mediante un valor aleatorio que decimos genera con una normal (0, sigma2), la desviación típica nos ayudará a que los valores sean plausibles y para ello, usaremos el error estándar residual que nos genera el propio modelo. Con esta aproximación conseguimos generar una variable imputada con una distribución similar a la original.

2. ¿Podemos imputar `age`, `gender` y `outcome` usando la media/moda? ¿Qué opinas de esta imputación?

```
vis_miss(data)
```

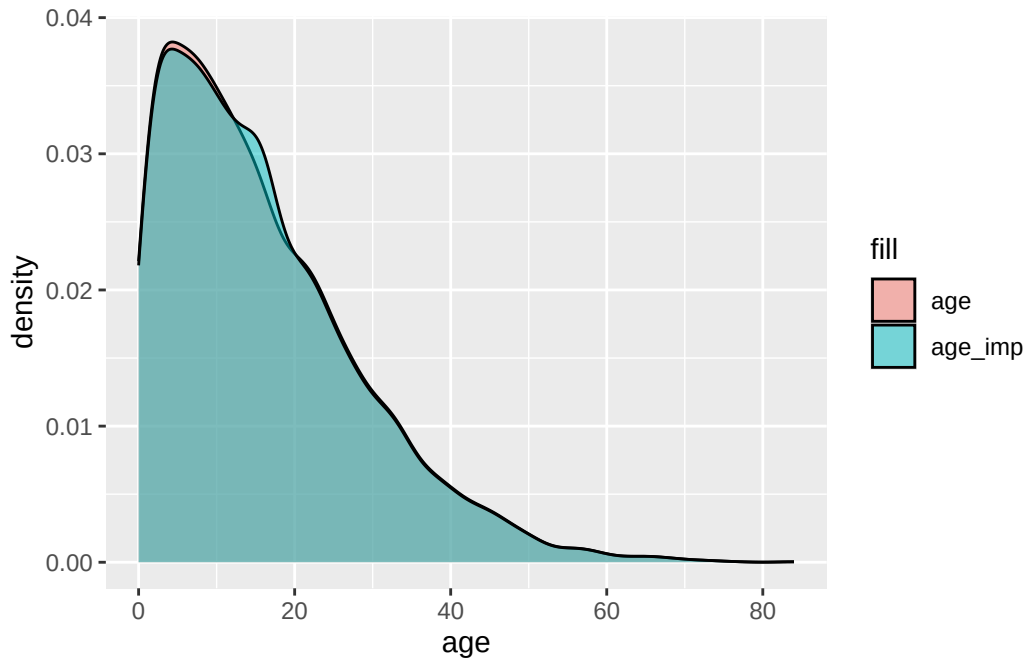



```
##AGE
data$age_imp<-data$age
mean(data$age_imp, na.rm = TRUE)
```

```
[1] 16.01068
```

```r
# 16.01068
data$age_imp[is.na(data$age_imp)] <- mean(data$age_imp, na.rm = TRUE)

##Gráfico para representar las diferencias
ggplot(data, aes(x = age, fill = "age")) +
geom_density(alpha = 0.5) +
geom_density(aes(x = age_imp, fill = "age_imp"), alpha = 0.5)
```

```
Warning: Removed 85 rows containing non-finite outside the scale range
(`stat_density()`).
```



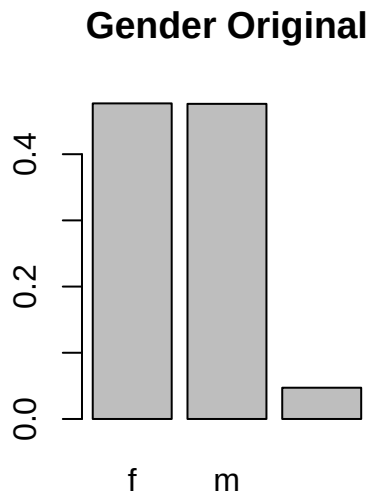

```r
##GENDER
prop.table(table(data$gender,useNA = "always"))
```

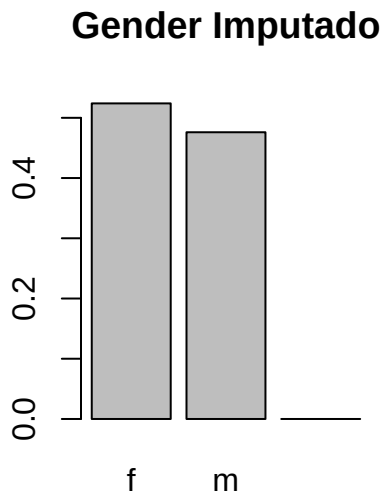
```
         f          m       <NA>
0.47673234 0.47605299 0.04721467
```

```r
data$gender_imp <-data$gender

data$gender_imp[is.na(data$gender_imp)] <- "f"
prop.table(table(data$gender_imp,useNA = "always"))
```

```
       f        m     <NA>
0.523947 0.476053 0.000000
```

```r
par(mfrow = c(1, 2))  # Divide la ventana gráfica en 1 fila y 2 columnas
barplot(prop.table(table(data$gender, useNA = "always")), main = "Gender Original")
barplot(prop.table(table(data$gender_imp, useNA = "always")), main = "Gender Imputado")
```






```r
##OUTCOME
prop.table(table(data$outcome,useNA = "always"))
```

```
    Death   Recover      <NA>
0.4385190 0.3367867 0.2246943
```

```r
data$outcome_imp <-data$outcome

data$outcome_imp[is.na(data$outcome_imp)] <- "Death"
prop.table(table(data$outcome_imp,useNA = "always"))
```

```
    Death   Recover      <NA>
0.6632133 0.3367867 0.0000000
```

```r
par(mfrow = c(1, 2))  # Divide la ventana gráfica en 1 fila y 2 columnas
barplot(prop.table(table(data$outcome, useNA = "always")), main = "Outcome Original")
barplot(prop.table(table(data$outcome_imp, useNA = "always")), main = "Outcome Imputado")
```


```r
##Modelo de regresión logística para la imputación de outcome
summary(glm(outcome~days_onset_hosp,data = data,family = "binomial"))
```

```
Call:
glm(formula = outcome ~ days_onset_hosp, family = "binomial",
    data = data)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.45750    0.04220  -10.84  < 2e-16 ***
days_onset_hosp  0.09985    0.01408    7.09 1.34e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5962.2  on 4349  degrees of freedom
Residual deviance: 5909.5  on 4348  degrees of freedom
  (1538 observations deleted due to missingness)
AIC: 5913.5

Number of Fisher Scoring iterations: 4
```

La variable edad la podemos imputar con la media porque tiene un % pequeño de datos missing y la distrbución lo permite.

La variable gender de la misma forma tiene un % pequeño y aunque imputemos por la categoría más frecuente, no se va de la frecuencia.

La variable outcome no se puede imputar por la moda ya que el % de missing es muy elevado y generamos una variable muy desigual a la original y completamente desbalanceado. Habría que buscar otras formas. En este caso como la variable outcomes es una variable cualitativa dicotómica, necesitamos hacer uso de una regresión logística en vez de una regresión lineal. En R se haría usando glm() con la opción binomial, pero esto no lo habéis estudiado, así que en este caso solo con saber que no se puede usar la moda es suficiente.

3. ¿Qué podríamos hacer con los missing de hospital para poder usar esa variable sin tener que borrar el 25% de los datos que son los que corresponden a esta variable?

```r
data$hospital_imp<-data$hospital
levels(data$hospital_imp) <- c(levels(data$hospital_imp), "Missing")
data$hospital_imp[which(is.na(data$hospital_imp))] <- "Missing"

table(data$hospital_imp)
```

```
                    Central Hospital                     Military Hospital
                                 454                                   896
                               Other                         Port Hospital
                                 885                                  1762
St. Mark's Maternity Hospital (SMMH)                              Missing
                                 422                                  1469
```

Al ser un % tan elevado si borramos todas las observaciones estaríamos borrando mucho. Por como es esta variable, vemos que los missing podría tener un significado, como por ejemplo ser un mismo hospital o pertenecer a una región socioeconómica más desfavorecida. Por tanto una opción sería generar una categoría para poder agrupar esos datos y tratar los NAs como una categoría más.

4. ¿Podemos imputar la variable `fever` a partir de `temp`? ¿Cómo procederías? ¿Encuentras alguna discrepancia a la hora de ejecutar esta acción?

```
library(psych)
```

```
Attaching package: 'psych'

The following objects are masked from 'package:ggplot2':

    %+%, alpha
```

```
describeBy(data$temp,data$fever)
```

```
 Descriptive statistics by group
group: no
   vars    n  mean   sd median trimmed  mad  min max range  skew kurtosis   se
X1    1 1050 36.97 0.48     37   36.98 0.59 35.2  38   2.8 -0.22     -0.3 0.01
------------------------------------------------------------
group: yes
   vars    n  mean   sd median trimmed  mad min  max range skew kurtosis   se
X1    1 4440 39.02 0.47     39   39.01 0.44  38 40.8   2.8 0.21    -0.16 0.01
```

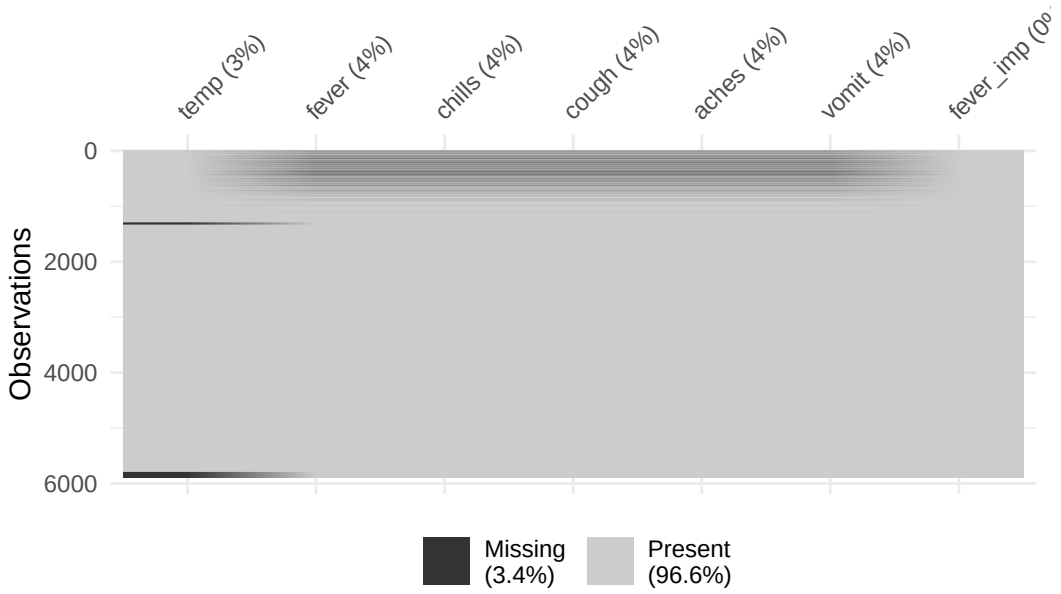
```
##Max sin fiebre es 38 y min con fiebre es 38

##No hay una clara definición, habría que tomar la decisión para corregir la definición

library(dplyr)
data$fever_imp<-if_else(is.na(data$temp),data$fever, if_else(data$temp<38,"no","yes"))
table(data$fever,data$fever_imp)
```

```
       no  yes
no   1085    5
yes     0 4549
```

Para imputar la variable fever es tan fácil como aplicar la definición de fiebre $>$ 38 SI, fiebre $<$ 38 NO. Lo único es que al mirar esta definición, nos hemos dado cuenta que a veces 38 está incluído en el si y a veces en el no, aprovechamos para corregir la definición.

5. ¿Podríamos imputar el resto de síntomas?

```
vis_miss(select(data,temp,fever,chills,cough,aches,vomit,fever_imp))
```




```
table(data$fever_imp,data$chills,useNA = "always")
```

```
       no  yes <NA>
no    873  212  248
yes  3667  887    1
<NA>    0    0    0
```

```r
table(data$fever_imp,data$cough,useNA = "always")
```

```
        no  yes <NA>
  no   149  936  248
  yes  624 3930    1
  <NA>   0    0    0
```

```r
table(data$fever_imp,data$aches,useNA = "always")
```

```
        no  yes <NA>
  no   989   96  248
  yes 4106  448    1
  <NA>   0    0    0
```

```r
table(data$fever_imp,data$vomit,useNA = "always")
```

```
        no  yes <NA>
  no   521  564  248
  yes 2315 2239    1
  <NA>   0    0    0
```

Todos los que se han imputado en fiebre se han imputado como fiebre = no, pero cuano vemos una tabla cruzada con e resto de síntomas, vemos que no todos los que no tienen fiebre tampoco tienen chills, o vomit etc, entonces podría ser un poco arriesgado, asumir que todos los síntomas son no. Podríamos intentar ver si con un modelo de regresión lo solucionamos, pero no hemos encontrado ninguna asociación, por tanto una posible opción es imputar por la moda y asumir el error.