Intro datos missing

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. ¿Qué número absoluto de datos missing hay en esta base de datos? y ¿porcentaje? y ¿observaciones completas? ¿Te parecen muchos o pocos?

```
n_miss(data) #number of missing values
```

```
[1] 11109
```

```
pct_miss(data) #percentage of missing values
```

```
[1] 8.203125
```

```
## tot=dim(data)[1]*dim(data)[2] #todos los posibles valores
## n_miss(data)/tot
pct_complete_case(data)
```

```
[1] 23.47147
```

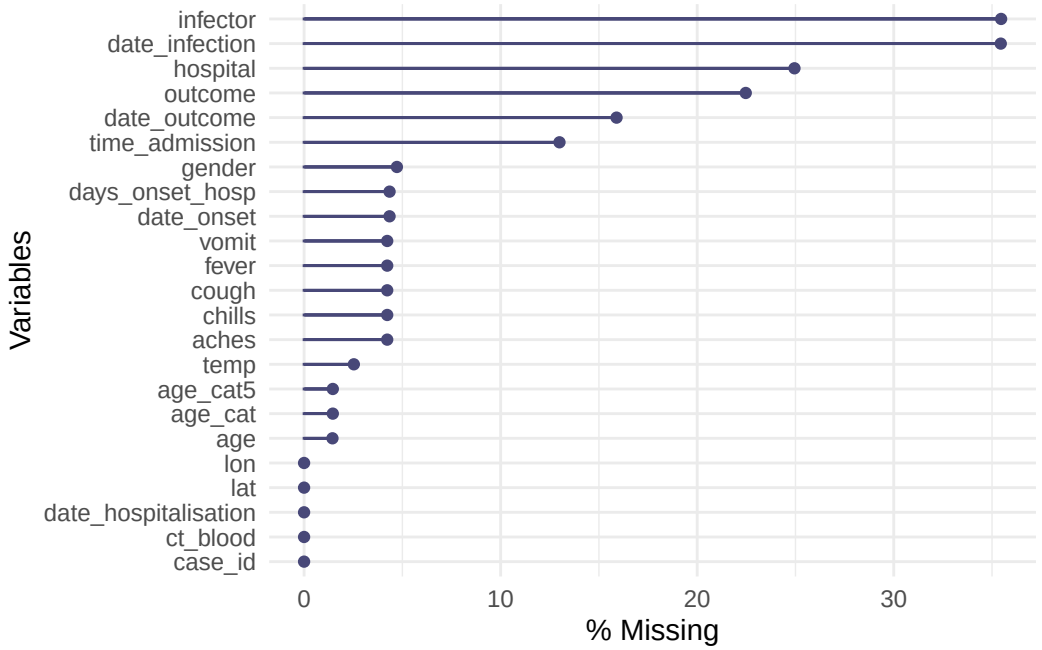
```
n_complete(data) #tot-n_miss(data)
```

```
[1] 124315
```

En términos absolutos parecen muchos datos, pero el porcentaje total de datos missing es un 8.6%. El problema real de esta base de datos es que sólo hay un 23.5% de observaciones completas y por tanto si quisieramos hacer un estudio multivariante con todas las variables o imputamos o si borramos perdemos mas de la mitad de las observaciones. Habrá que trabajar los datos missing para ver que hacemos con la base de datos.

2. Usando un gráfico ¿Qué variables son las que más porcentaje de datos missing tienen? ¿Hay variables que no tengan datos missing o que tengan menos de un 5%?

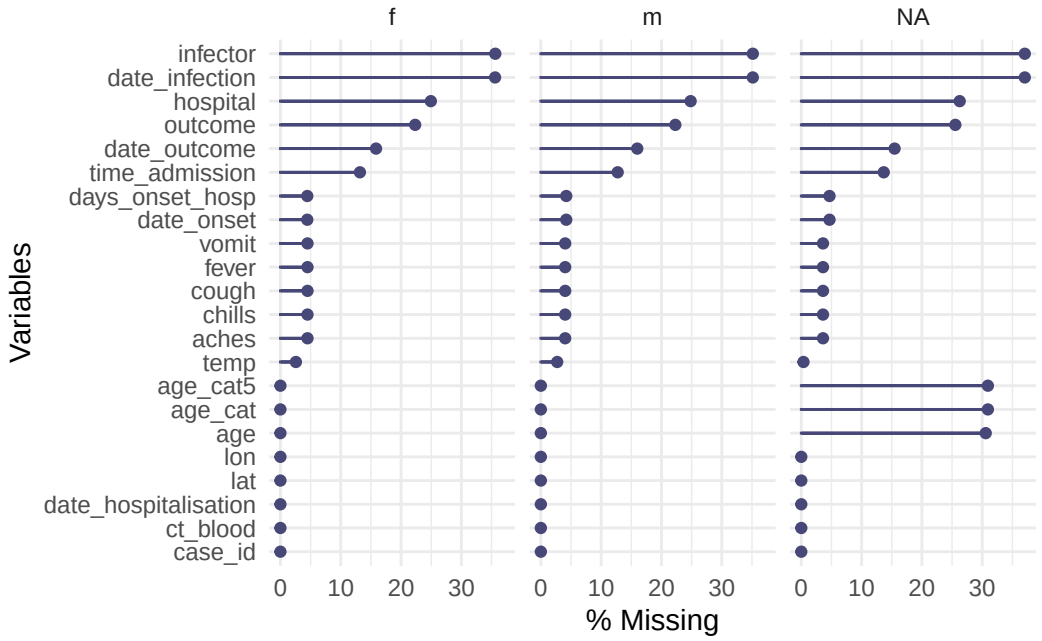
```
gg_miss_var(data,show_pct = TRUE)
```



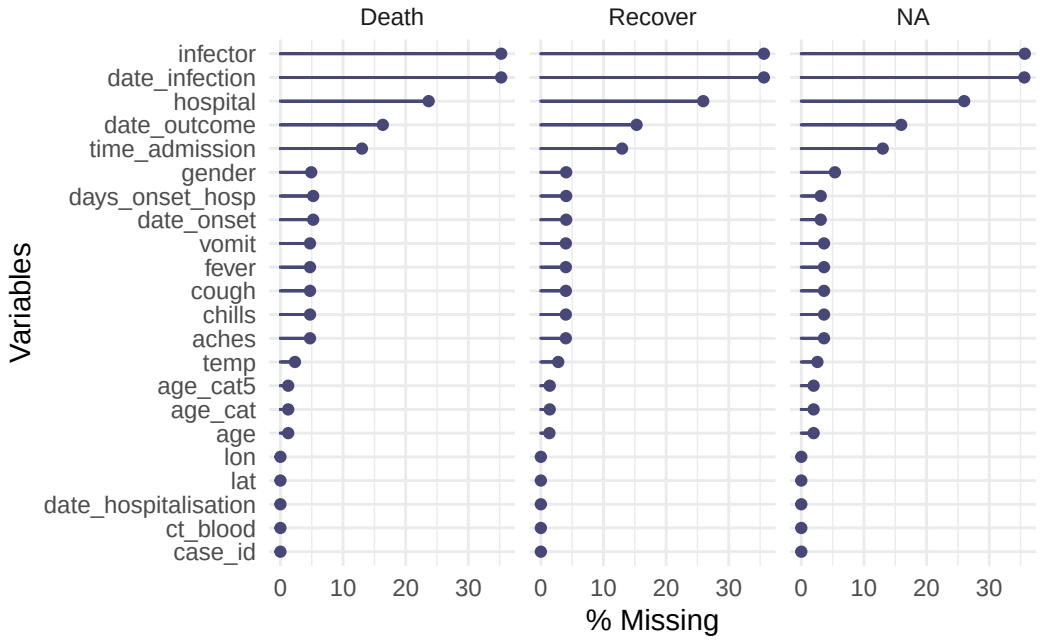
Infector y date_infection tiene más de un 30% de valores missing, hospital, outcome, date_outcome y time_admission tiene entre 10 y 25% habría que pensar si imputar o borrar las variables enteras, seguramente dependa de los análisis posteriores. La gran mayoría de variables tienen menos del 5% y esta es una buena cifra para imputar datos.

3. Usando un gráfico, representa los datos missing de toda la base de datos por los valores de la variable gender, outcome, age_cat y hospital. ¿Podemos ver algún patrón?

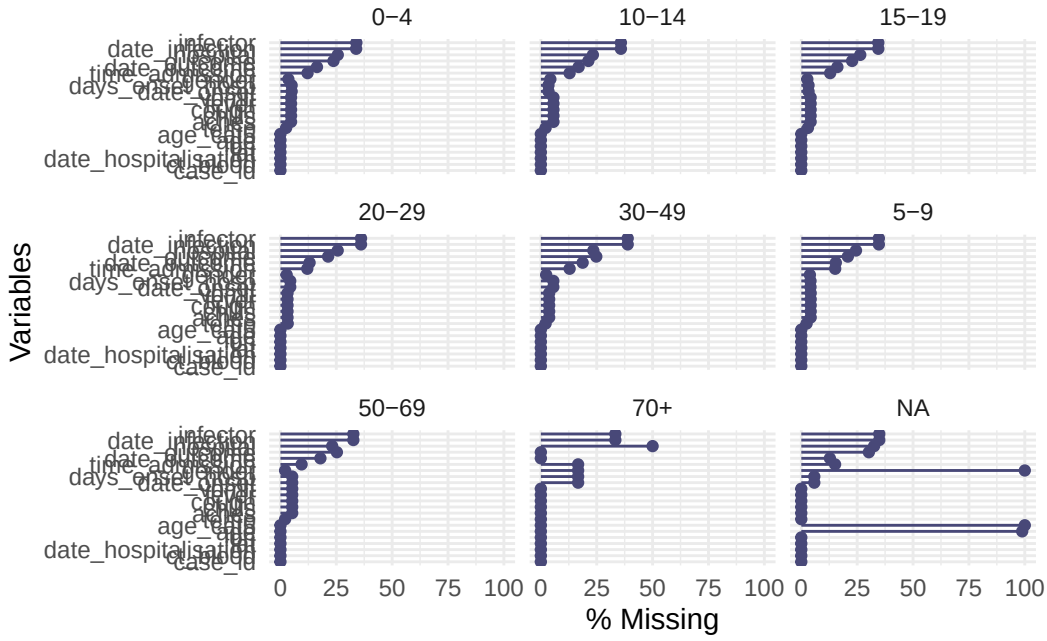
```
gg_miss_var(data, show_pct = TRUE,facet = gender)
```



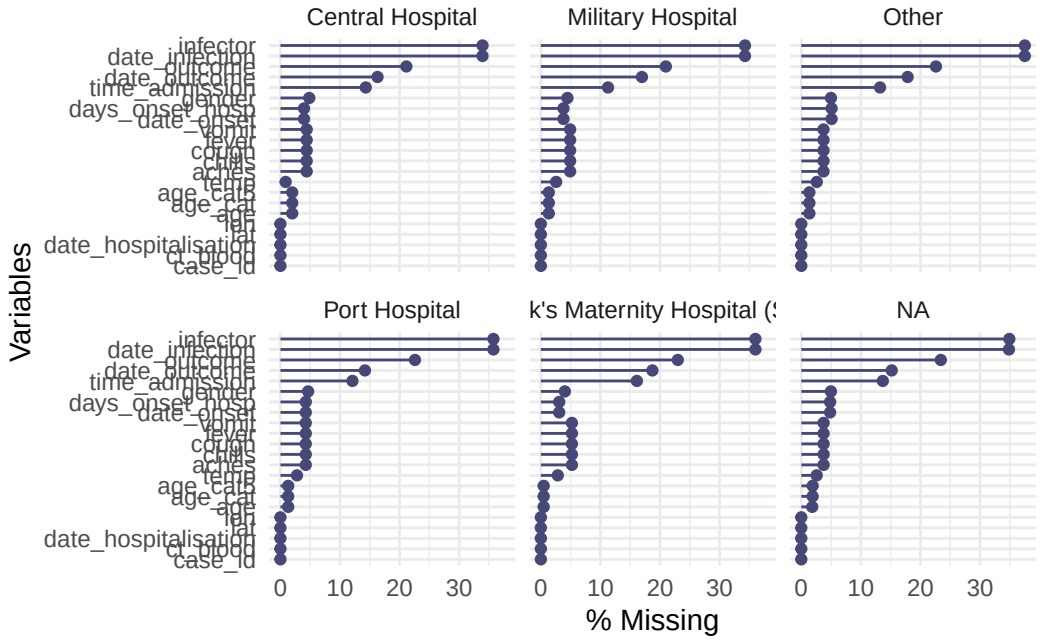

```
gg_miss_var(data, show_pct = TRUE,facet = outcome)
```




```
gg_miss_var(data, show_pct = TRUE,facet = age_cat)
```

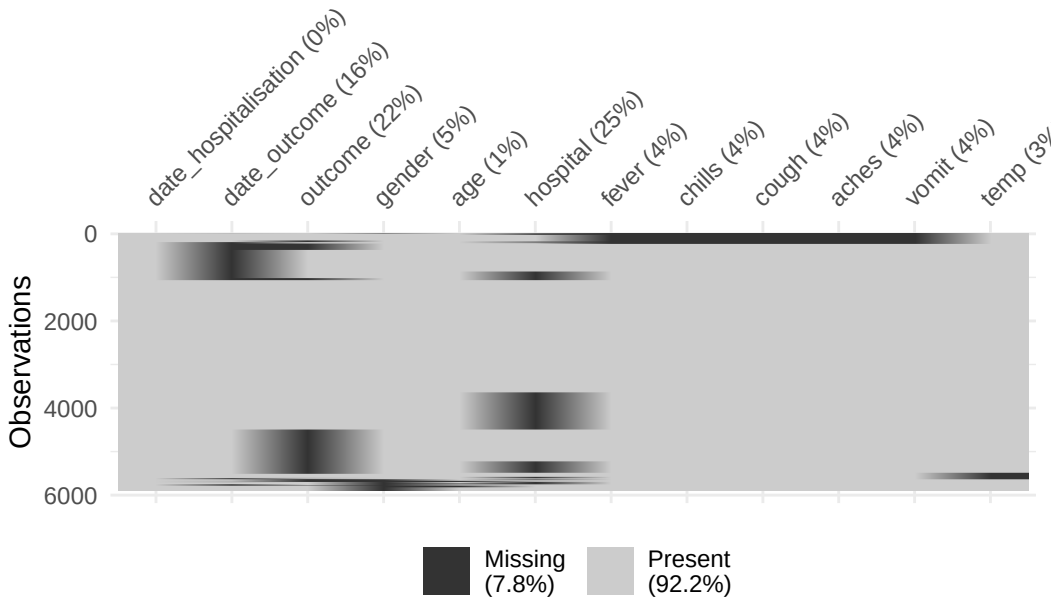


```
gg_miss_var(data, show_pct = TRUE,facet = hospital)
```

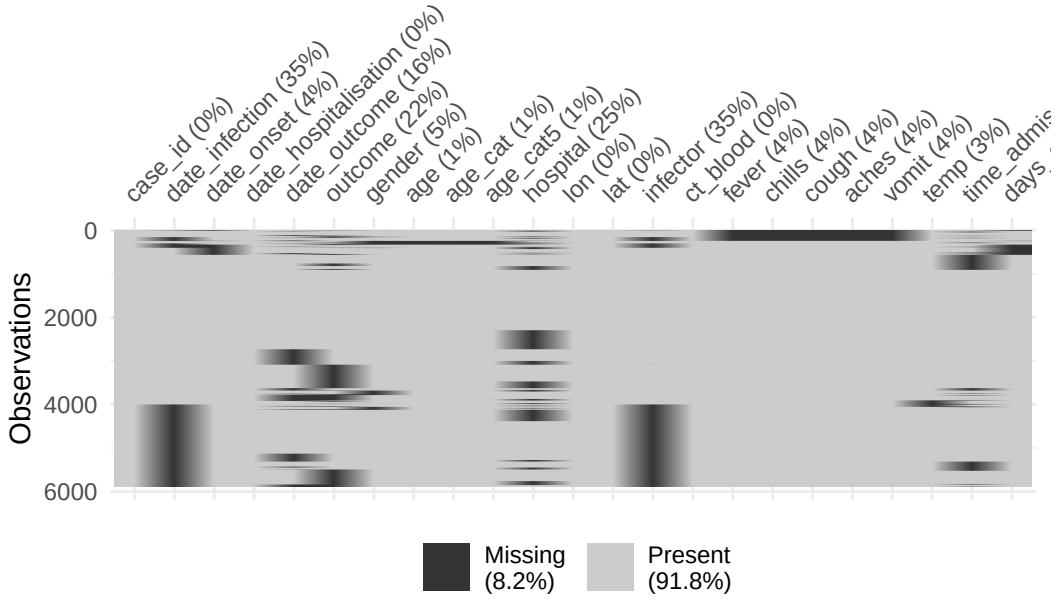


4. Imagímate que quieres estudiar los días que han pasado desde la fecha de hospitalización hasta el outcome según las características de los pacientes (age, gender, hospital) y sus síntomas (fever, chills, cough, aches vomit, temp). ¿Qué heatmap representarías para ver los datos missing? ¿Qué problemas vas a encontrar en este análisis?

```
vis_miss(select(data,date_hospitalisation,date_outcome,outcome,gender,age,hospital,fever,chi
```




```
vis_miss(data,cluster = T)
```



El principal problema que vemos aquí es que las variables correspondientes a los síntomas corresponden a las mismas observaciones, por tanto, por un lado, no sabemos por qué puede

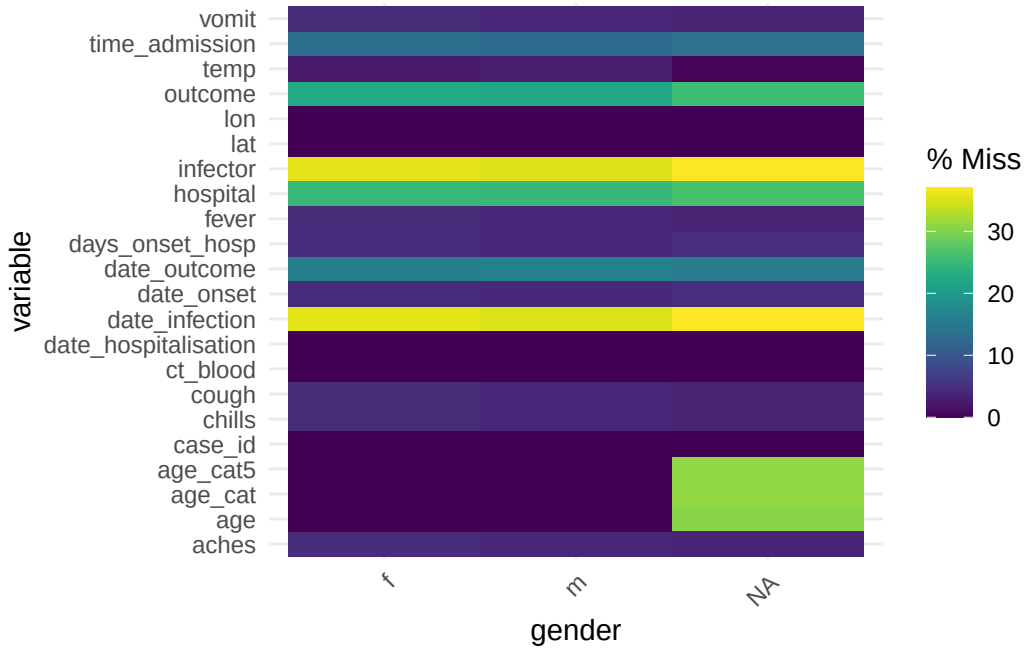
ser debido esto y puede que haya alguna razón detrás que no sepamos y por otro lado, no vamos a poder usar el resto de síntomas para imputar estos datos missing.

También vemos que la fecha date_outcome tiene muchos datos missing, esto hace que si construimos una nueva variable con los días que han pasado desde la fecha de hospitalización hasta el outcome, tendremos todos los missing de esta variable.

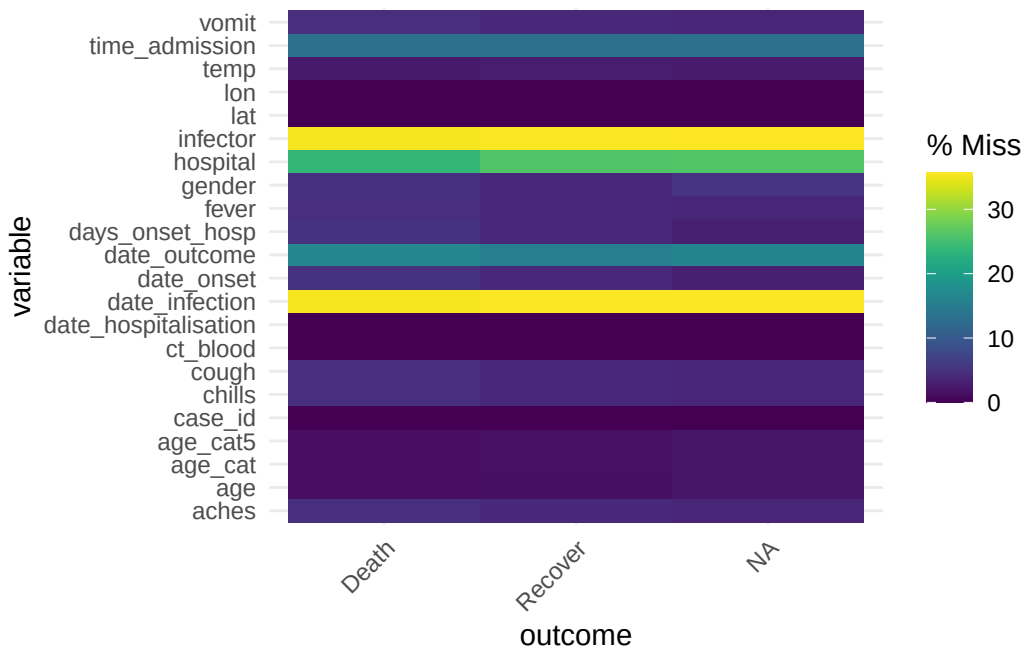
5. Representa un heatmap de todos los datos por las variables categóricas (gender, outcome y hospital y por las fechas date_hospitalisation y date_outcome). ¿Qué observas en estos gráficos?

```
gg_miss_fct(data, gender)
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `gender = (function (x) ...`.
Caused by warning:
! `fct_explicit_na()` was deprecated in forcats 1.0.0.
i Please use `fct_na_value_to_level()` instead.
i The deprecated feature was likely used in the naniar package.
  Please report the issue at <https://github.com/njtierney/naniar/issues>.
```



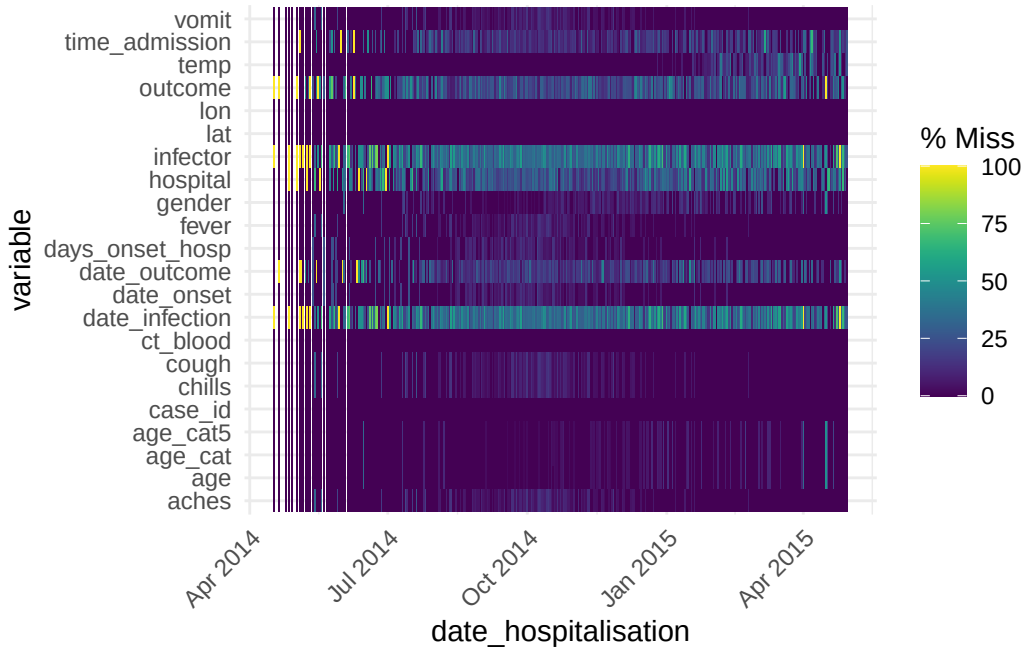
```
gg_miss_fct(data, outcome)
```




```
gg_miss_fct(data, date_infection)
```

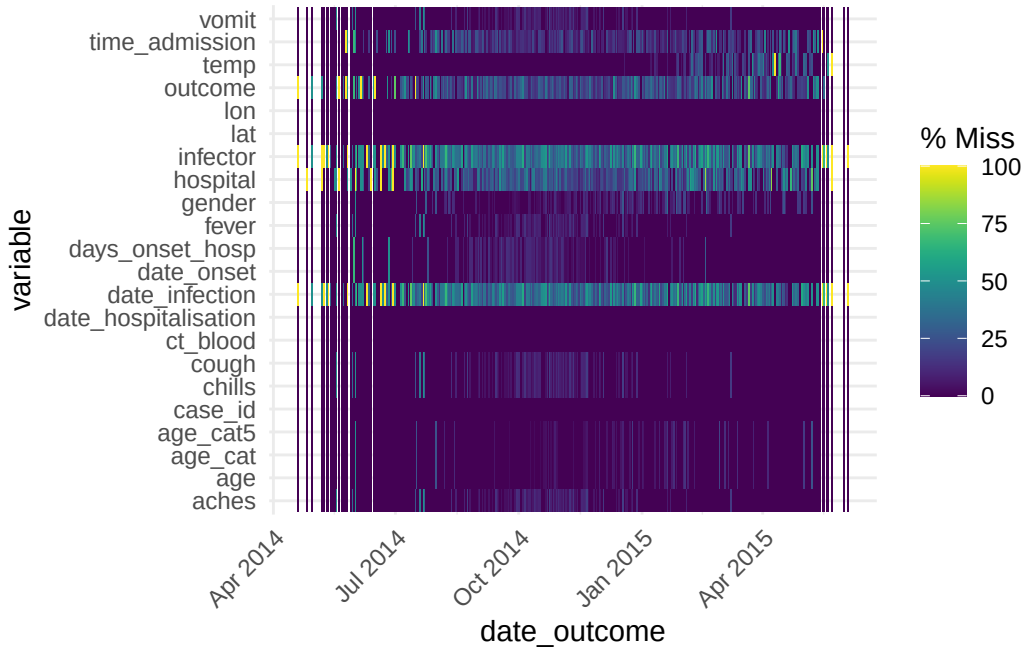
```
Warning: Removed 22 rows containing missing values or values outside the scale range
(`geom_tile()`).
```


```
gg_miss_fct(data, date_hospitalisation)
```



```
gg_miss_fct(data, date_outcome)
```

```
Warning: Removed 22 rows containing missing values or values outside the scale range
(`geom_tile()`).
```



Para gender y outcome no se observa ningún patrón destacable más que el que ya sabíamos de los missing en la variable gender y age para las mismas observaciones. Sobre las fechas, no hay nignún patrón en los missing, pero si parece que al inicio de la recogida (inicio de la epidemia) y al final había menos datos que al final, que los datos recogidos son más continuos en el tiempo.

6. Usando las variables "shadow", realiza un gráfico que muestren la distribución de las distintas fechas considerando los datos missing (NA) y no missing (!NA) de la variable edad y gender. ¿Qué conclusión sacarías?

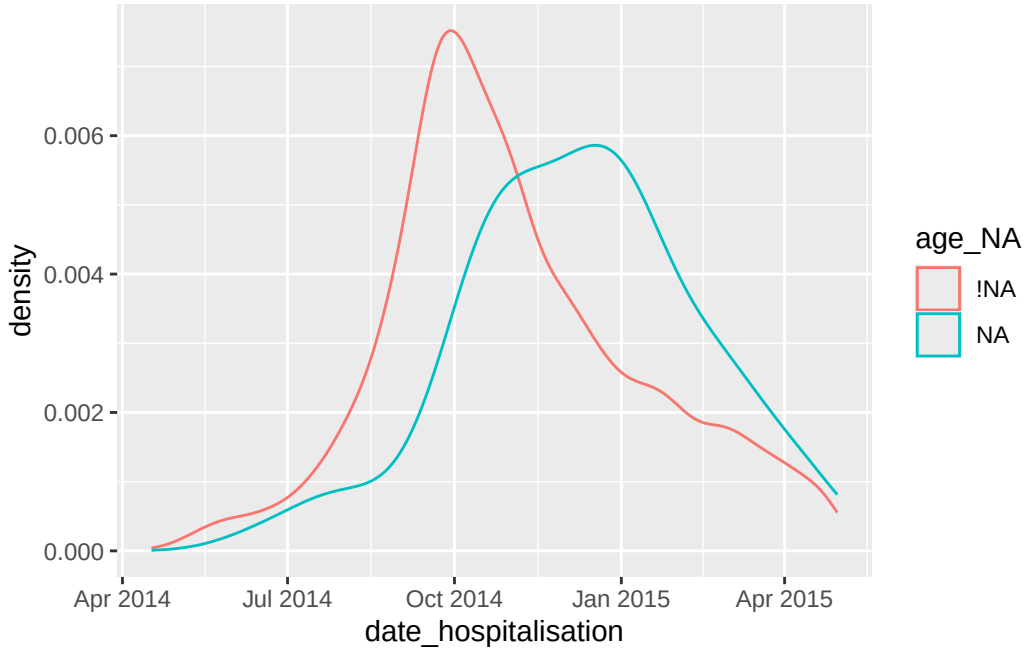
```
shadowed_data <- data %>%
  bind_shadow()

shadowed_data[,c("gender","gender_NA")]
```

```
# A tibble: 5,888 x 2
   gender gender_NA
   <fct>  <fct>
 1 m      !NA
 2 f      !NA
 3 m      !NA
 4 f      !NA
 5 m      !NA
 6 f      !NA
 7 f      !NA
 8 f      !NA
 9 m      !NA
10 f      !NA
# i 5,878 more rows
```

```
table(shadowed_data$gender,shadowed_data$gender_NA,useNA = "always")
```

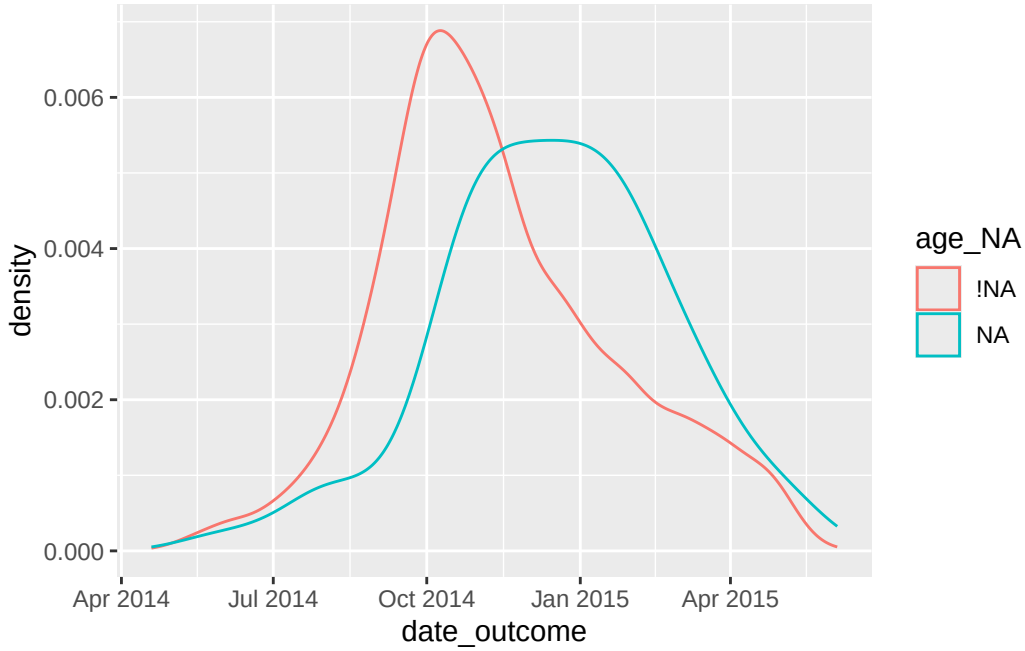
```
        !NA   NA <NA>
  f    2807    0    0
  m    2803    0    0
  <NA>    0  278    0
```

```
ggplot (data = shadowed_data,mapping = aes(x = date_hospitalisation,colour = age_NA)) +
        geom_density()
```



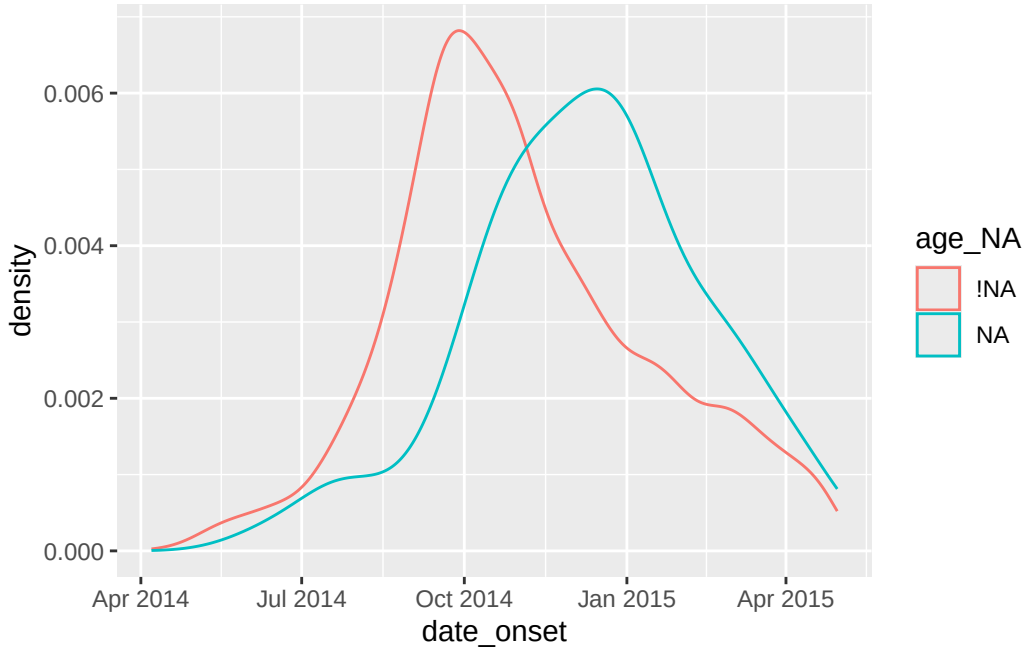

```
ggplot (data = shadowed_data,mapping = aes(x = date_outcome, colour = age_NA)) +
         geom_density()
```

```
Warning: Removed 936 rows containing non-finite outside the scale range
(`stat_density()`).
```

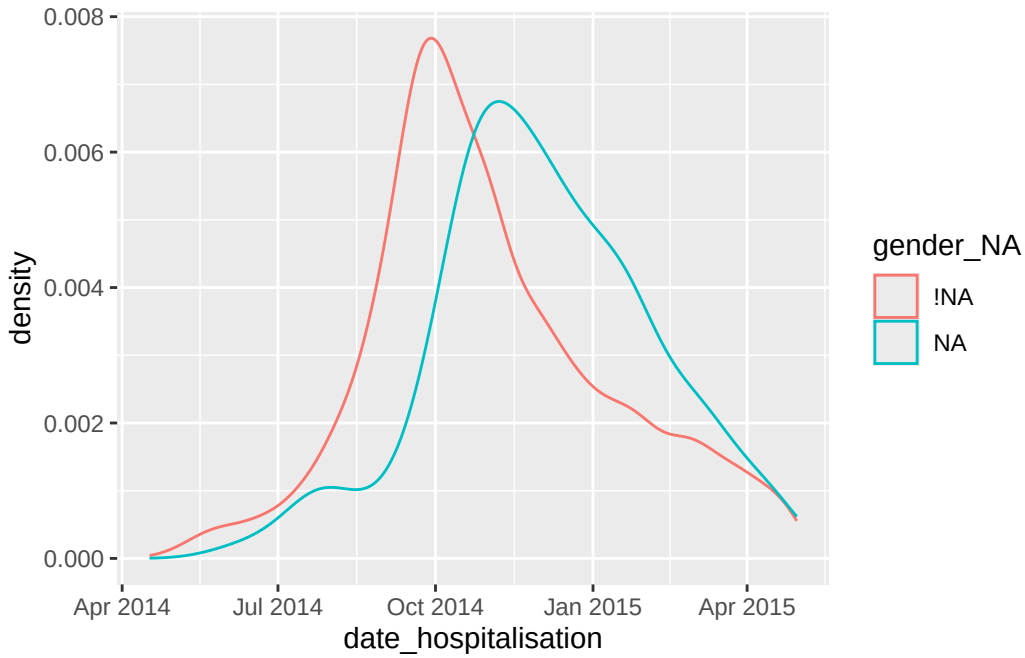



```
ggplot (data = shadowed_data, mapping = aes(x = date_onset, colour = age_NA)) +
          geom_density()
```

```
Warning: Removed 256 rows containing non-finite outside the scale range
(`stat_density()`).
```

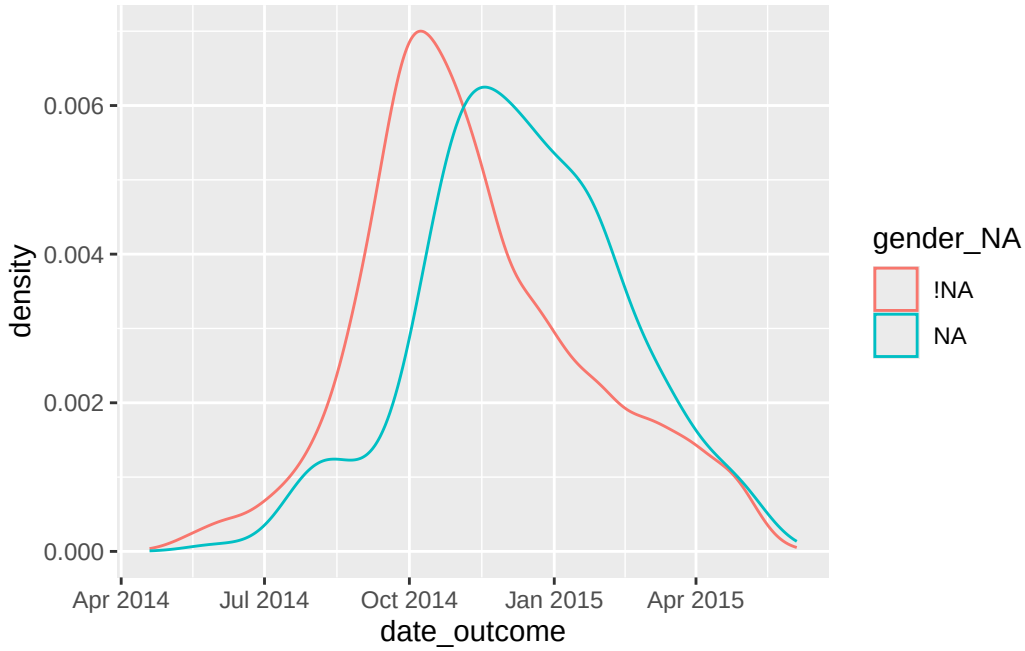



```
ggplot (data = shadowed_data, mapping = aes(x = date_hospitalisation,colour = gender_NA)) +
```



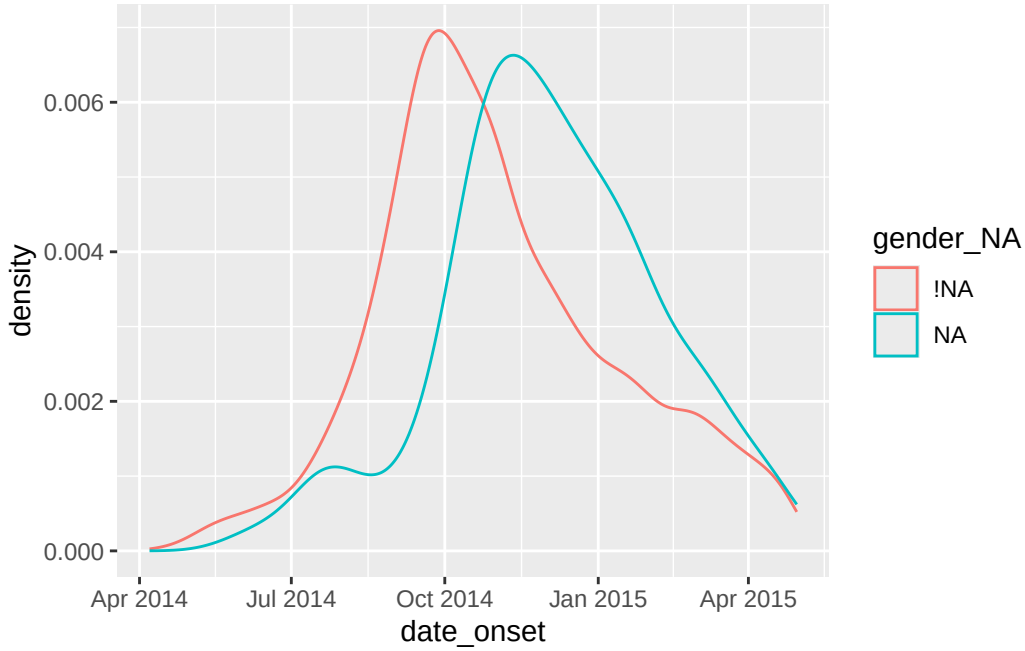
```
ggplot (data = shadowed_data, mapping = aes(x = date_outcome, colour = gender_NA)) +
```

```
Warning: Removed 936 rows containing non-finite outside the scale range
(`stat_density()`).
```




```
ggplot (data = shadowed_data, mapping = aes(x = date_onset, colour = gender_NA)) +
              geom_density()
```

```
Warning: Removed 256 rows containing non-finite outside the scale range
(`stat_density()`).
```

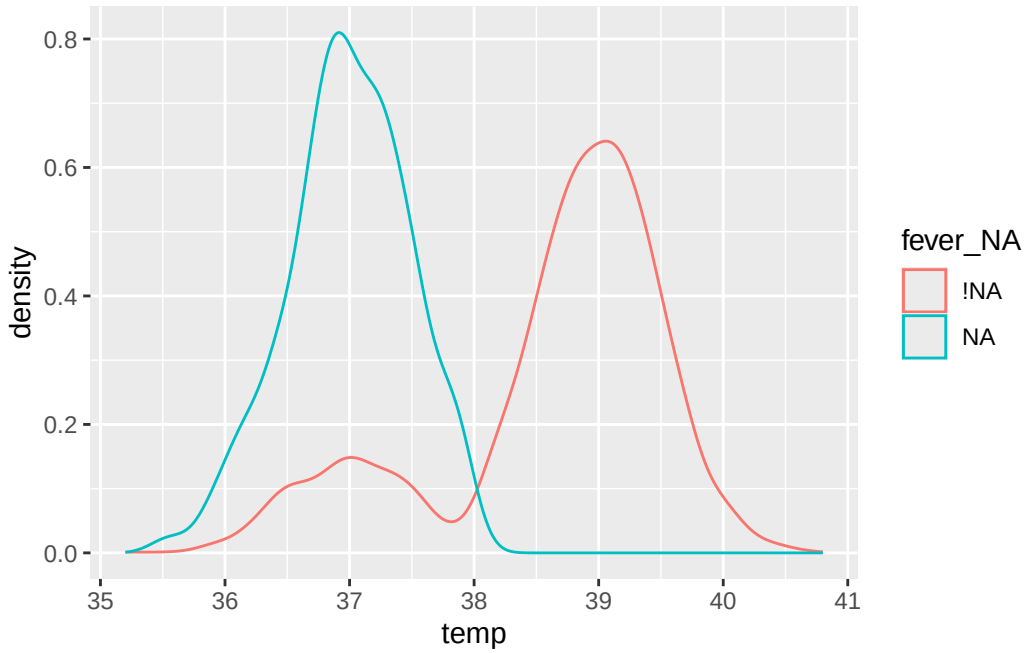


Para ambas variables (edad y gender) y todas las fechas lo que se observa es que los datos missing corresponden a fechas más tardías desde el inicio de la recogida de datos (epidemia). Podría corresponder a una situación en la que al inicio que hay menos gente acudiendo a los hospitales la recogida de datos se hace de forma completa y adecuada y según pasa el tiempo y los hospitales colapsan, la recogida de datos es muchas veces imposible.

7. Usando las variables "shadow", realiza un gráfico para ver si las personas que tienen missing en los síntomas, corresponden a personas más sanas (puedes usar la variable temp)

```
shadowed_data <- data %>%
  bind_shadow()


ggplot (data = shadowed_data, mapping = aes(x = temp, colour = fever_NA)) +
  geom_density()
```

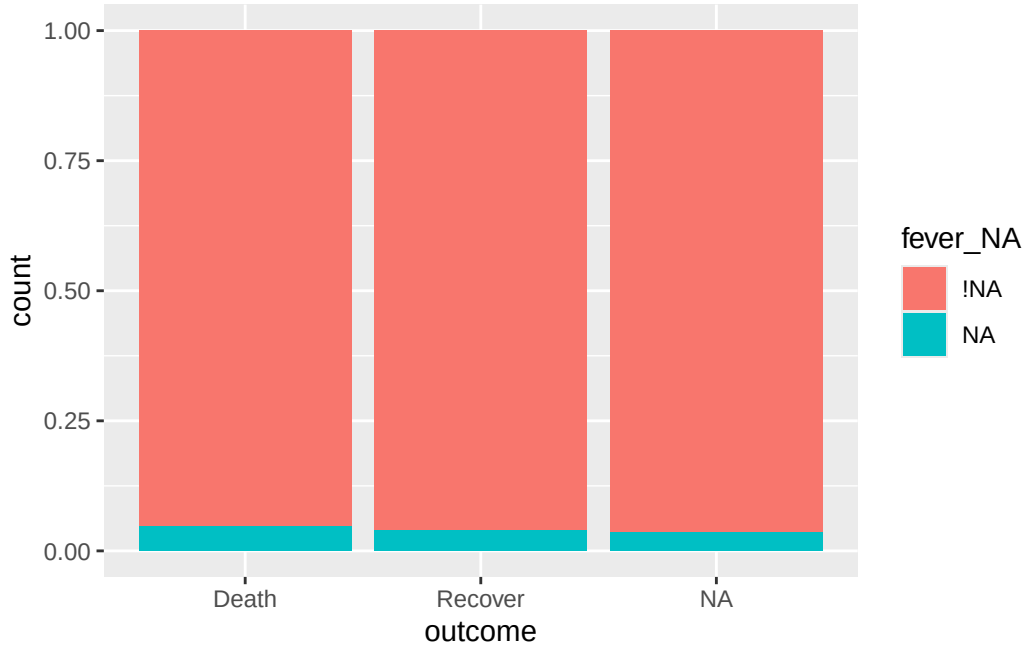
```
Warning: Removed 149 rows containing non-finite outside the scale range
(`stat_density()`).
```




```
table(shadowed_data$outcome,shadowed_data$fever_NA)
```

```
           !NA   NA
  Death    2460  122
  Recover 1904   79
```

```
ggplot(shadowed_data, aes(x = outcome, fill = fever_NA)) +
  geom_bar(position = "fill")
```



En el gráfico de temperatura se observa que los datos faltantes de la variable fever y de todos los síntomas son en aquellos que no tuvieron fiebre, en cambio cuando representamos el outcome con los missing de esta misma, no parece haber ningún patrón.