R Notebook

Mike Kozlowski, HDS

Ebola outbreak data set

Data from a predicted data set

https://epirhandbook.com/en/missing-data.html (https://epirhandbook.com/en/missing-data.html)

Making use of the nanair package

```
require("VIM")
```

```
## Loading required package: VIM
```

```
## Warning: package 'VIM' was built under R version 4.2.3
```

```
## 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:datasets':
##
##     sleep
```

```
require("validate")
```

```
## Loading required package: validate
```

```
## Warning: package 'validate' was built under R version 4.2.3
```

```
require('robustbase')
```

```
## Loading required package: robustbase
```

```
## Warning: package 'robustbase' was built under R version 4.2.3
```

```
require("lmtest")
```

```
## Loading required package: lmtest
```

```
## Warning: package 'lmtest' was built under R version 4.2.3
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.2.3
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
##
## Attaching package: 'lmtest'
```

```
## The following object is masked from 'package:validate':
##
##     reset
```

```
require("rio")
```

```
## Loading required package: rio
```

```
## Warning: package 'rio' was built under R version 4.2.3
```

```
infile="C:\\Users\\Mike\\Documents\\DAT511\\5-3 class\\linelist_cleaned.rds"

linelist=import(infile)
```

This is a datase from a simulated ebola infection event, it is in a slightly odd data format, an rds or r data set file

We will look at it using the tools in VIM, but also another package called naniar, a tool to analyze Na values

What do we have in the data:

```
head(linelist)
```

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

6 rows | 1-7 of 31 columns

```
summary(linelist)
```

```
##    case_id            generation    date_infection         date_onset
##  Length:5888        Min.   : 0.00   Min.   :2014-03-19   Min.   :2014-04-07
##  Class :character   1st Qu.:13.00   1st Qu.:2014-09-06   1st Qu.:2014-09-16
##  Mode  :character   Median :16.00   Median :2014-10-11   Median :2014-10-23
##                     Mean   :16.56   Mean   :2014-10-22   Mean   :2014-11-03
##                     3rd Qu.:20.00   3rd Qu.:2014-12-05   3rd Qu.:2014-12-19
##                     Max.   :37.00   Max.   :2015-04-27   Max.   :2015-04-30
##                                     NA's   :2087         NA's   :256
##  date_hospitalisation  date_outcome          outcome
##  Min.   :2014-04-17   Min.   :2014-04-19   Length:5888
##  1st Qu.:2014-09-19   1st Qu.:2014-09-26   Class :character
##  Median :2014-10-23   Median :2014-11-01   Mode  :character
##  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
##     gender               age          age_unit           age_years
##  Length:5888        Min.   : 0.00   Length:5888        Min.   : 0.00
##  Class :character   1st Qu.: 6.00   Class :character   1st Qu.: 6.00
##  Mode  :character   Median :13.00   Mode  :character   Median :13.00
##                     Mean   :16.07                      Mean   :16.02
##                     3rd Qu.:23.00                      3rd Qu.:23.00
##                     Max.   :84.00                      Max.   :84.00
##                     NA's   :86                         NA's   :86
##     age_cat        age_cat5      hospital             lon
##  0-4    :1095   0-4    :1095   Length:5888        Min.   :-13.27
##  5-9    :1095   5-9    :1095   Class :character   1st Qu.:-13.25
##  20-29  :1073   10-14  : 941   Mode  :character   Median :-13.23
##  10-14  : 941   15-19  : 743                      Mean   :-13.23
##  30-49  : 754   20-24  : 638                      3rd Qu.:-13.22
##  (Other): 844   (Other):1290                      Max.   :-13.21
##  NA's   :  86   NA's   :  86
##       lat          infector            source              wt_kg
##  Min.   :8.446   Length:5888        Length:5888        Min.   :-11.00
##  1st Qu.:8.461   Class :character   Class :character   1st Qu.: 41.00
##  Median :8.469   Mode  :character   Mode  :character   Median : 54.00
##  Mean   :8.470                                         Mean   : 52.64
##  3rd Qu.:8.480                                         3rd Qu.: 66.00
##  Max.   :8.492                                         Max.   :111.00
##
##      ht_cm        ct_blood        fever              chills
##  Min.   :  4   Min.   :16.00   Length:5888        Length:5888
##  1st Qu.: 91   1st Qu.:20.00   Class :character   Class :character
##  Median :129   Median :22.00   Mode  :character   Mode  :character
##  Mean   :125   Mean   :21.21
##  3rd Qu.:159   3rd Qu.:22.00
##  Max.   :295   Max.   :26.00
##
##     cough              aches              vomit               temp
##  Length:5888        Length:5888        Length:5888        Min.   :35.20
##  Class :character   Class :character   Class :character   1st Qu.:38.20
##  Mode  :character   Mode  :character   Mode  :character   Median :38.80
##                                                           Mean   :38.56
##                                                           3rd Qu.:39.20
##                                                           Max.   :40.80
##                                                           NA's   :149
##  time_admission          bmi           days_onset_hosp
##  Length:5888        Min.   :-1200.00   Min.   : 0.000
##  Class :character   1st Qu.:   24.56   1st Qu.: 1.000
##  Mode  :character   Median :   32.12   Median : 1.000
##                     Mean   :   46.89   Mean   : 2.059
##                     3rd Qu.:   50.01   3rd Qu.: 3.000
```

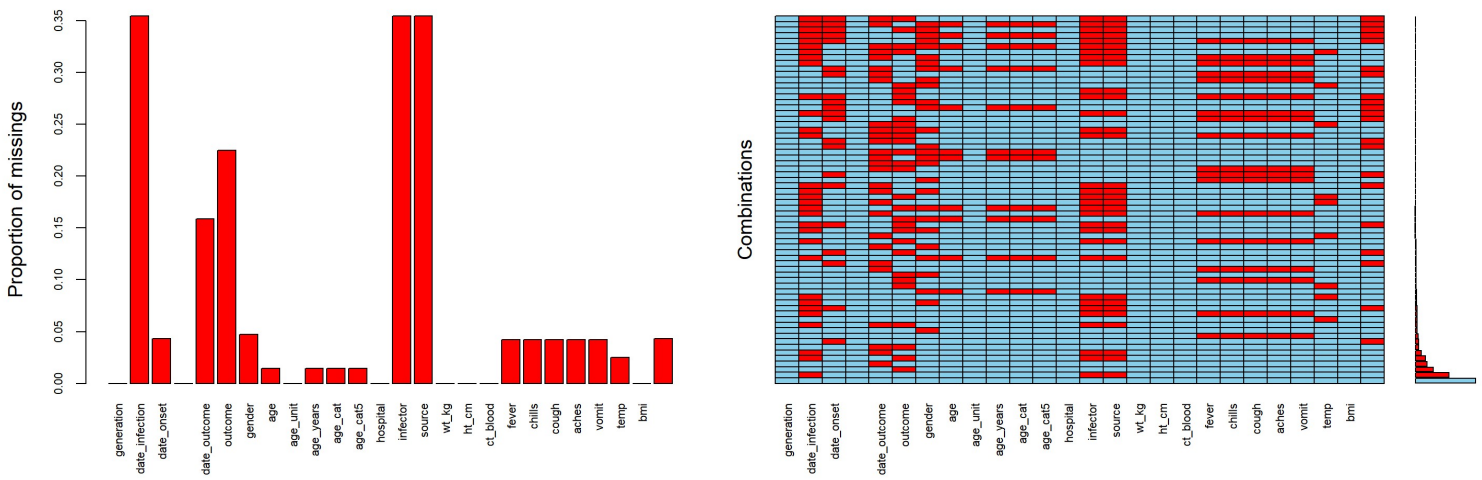
```
##                   Max.   : 1250.00   Max.   :22.000
##                                      NA's   :256
```

Just to simplify things a bit, let's drop some data

time_admission, lon, lat, case_id -these don't mean much to us

```
linelist=linelist[, !(names(linelist) %in% c("time_admission","lon","lat","case_id"))]
```

```
aggr(linelist,cex.axis=0.7)
```



#Question/Action

What are the top five most often missing items in this data set?

The top five most often missing items are date_infeection, date_outcome, outcome, infector, and source.

What are the top five most common combinations of missing items?

The top five most common combinations are date_infection, date_outcome, outcome, infector, and source.

The nanair package has an alternative way to show missing data information

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:validate':
##
##     expr
```

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## Warning: package 'tidyverse' was built under R version 4.2.3
```

```
## Warning: package 'readr' was built under R version 4.2.3
```

```
## Warning: package 'forcats' was built under R version 4.2.3
```

```
## Warning: package 'lubridate' was built under R version 4.2.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.0     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ lubridate 1.9.2     ✓ tibble    3.1.8
## ✓ purrr     1.0.1     ✓ tidyr     1.3.0
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::expr()   masks ggplot2::expr(), validate::expr()
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors
```

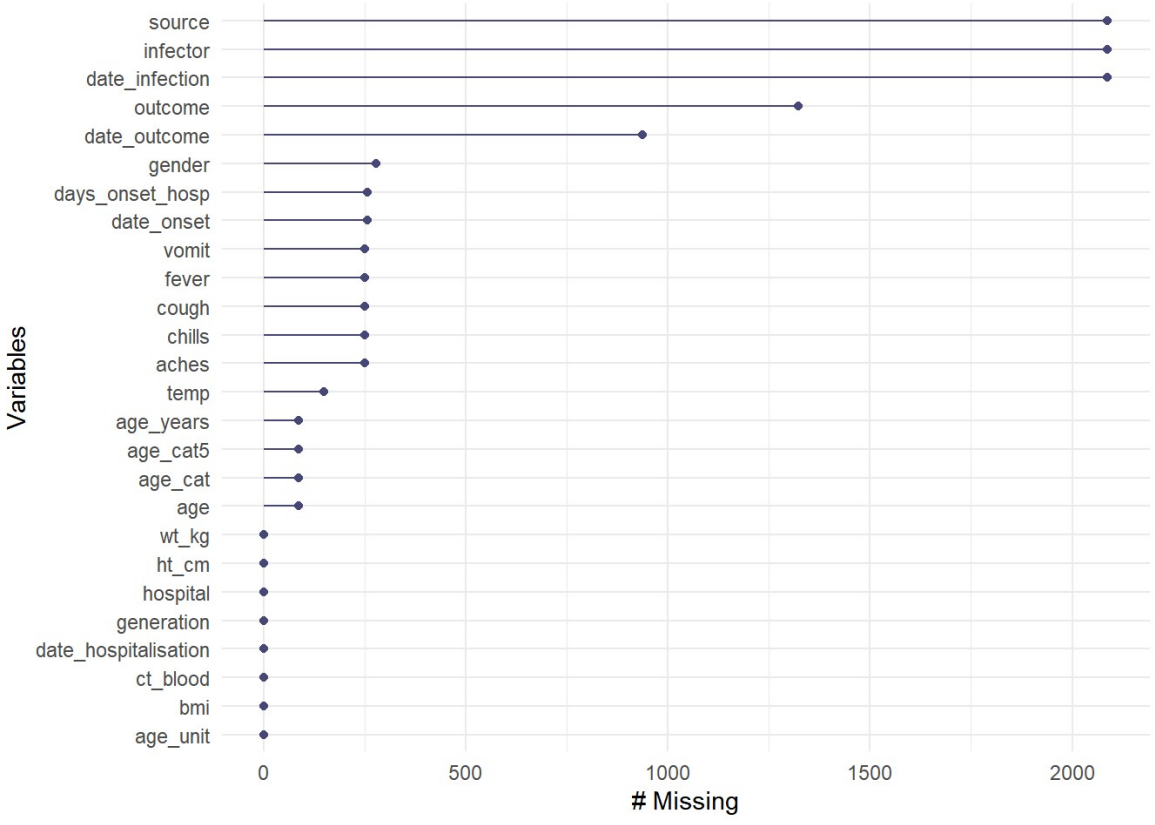
```
require(naniar)
```

```
## Loading required package: naniar
```

```
## Warning: package 'naniar' was built under R version 4.2.3
```

```
##
## Attaching package: 'naniar'
##
## The following object is masked from 'package:validate':
##
##     all_complete
```

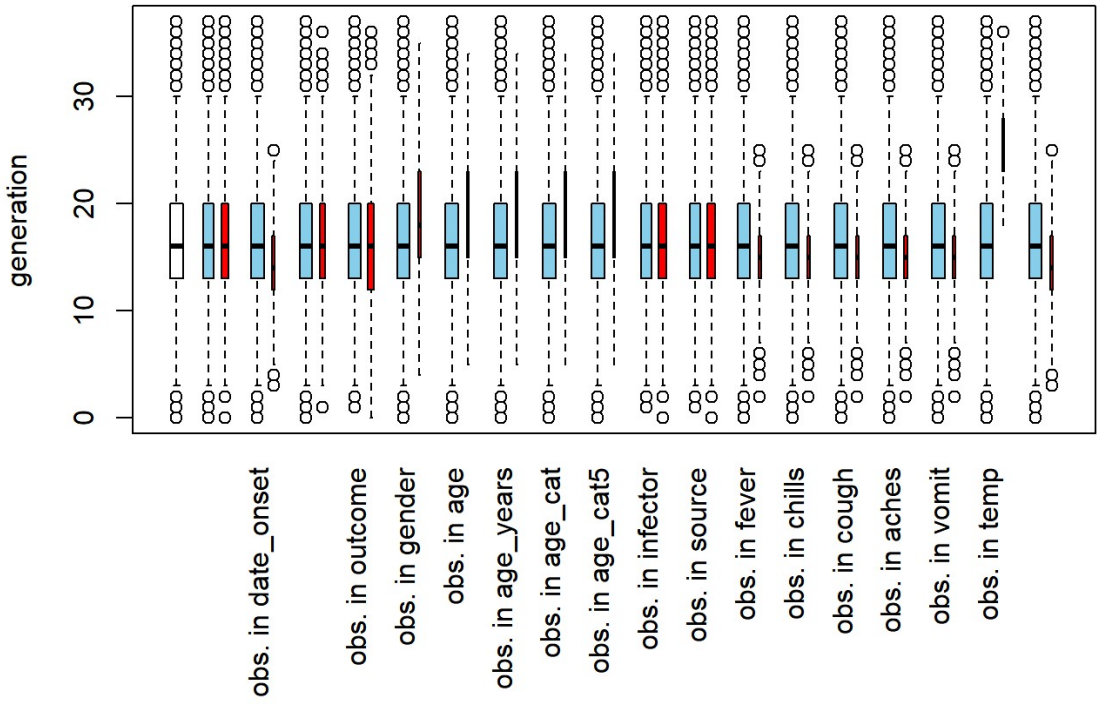
```
gg_miss_var(linelist)
```



Using VIM to look at patterns or shifts in data due to missing values

```
#windows()
VIM::pbox(linelist)
```

```
## Warning in createPlot(main, sub, xlab, ylab, labels, ca$at): not enough space
## to display frequencies
```



Question/Action

What variable in this plot shows a large shift associated with missing data values?

Infector has such a large shift that the scale is in scientific notation.

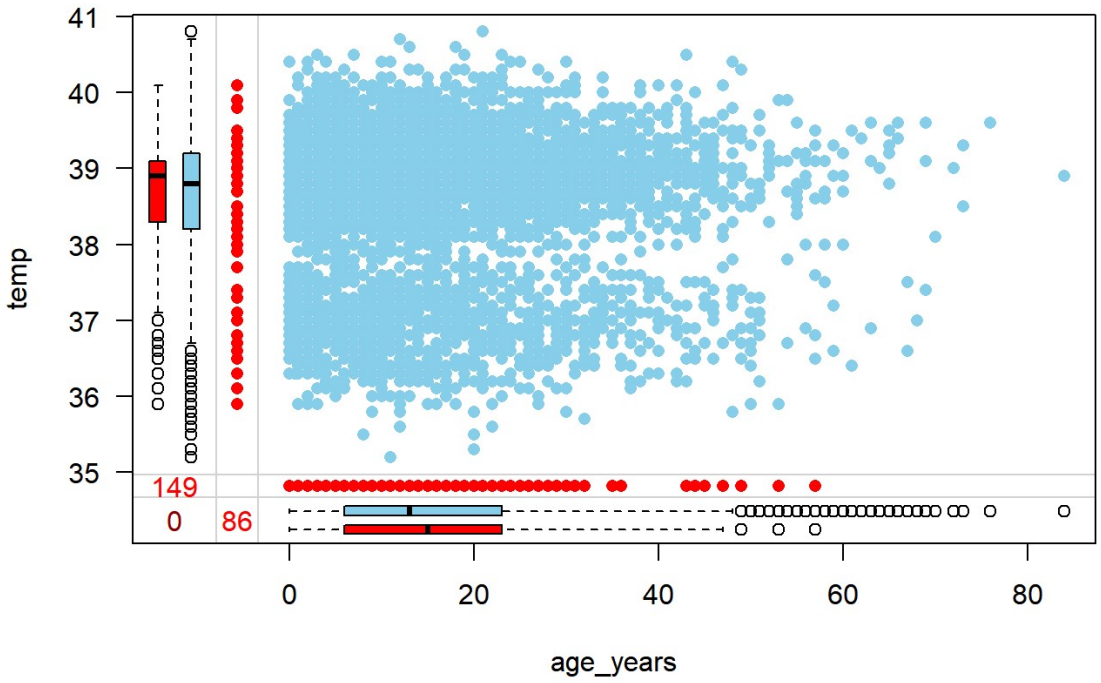
Describe what appears to be happening, and whether or not this makes sense.

It is happening in columns involving age, which tells me that missing age values are causing a huge shift in the data. This makes sense as age has a small range of ~0-100, so if a value were to be an error integer in the hundreds of thousands, this could occur.

We can look at a margin plot of age in years and temp

Here is the VIM version

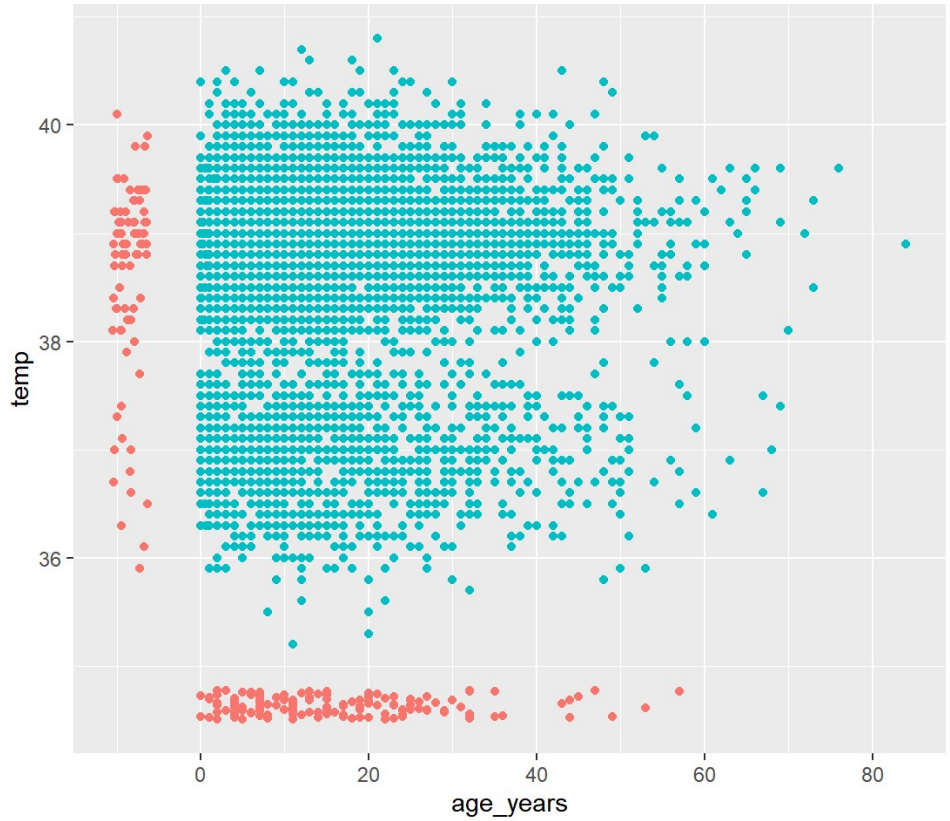
```
mydat=linelist[,c("age_years","temp")]
VIM::marginplot(mydat,las=1,pch=16)
```



Here is the nanair version of

the same plot

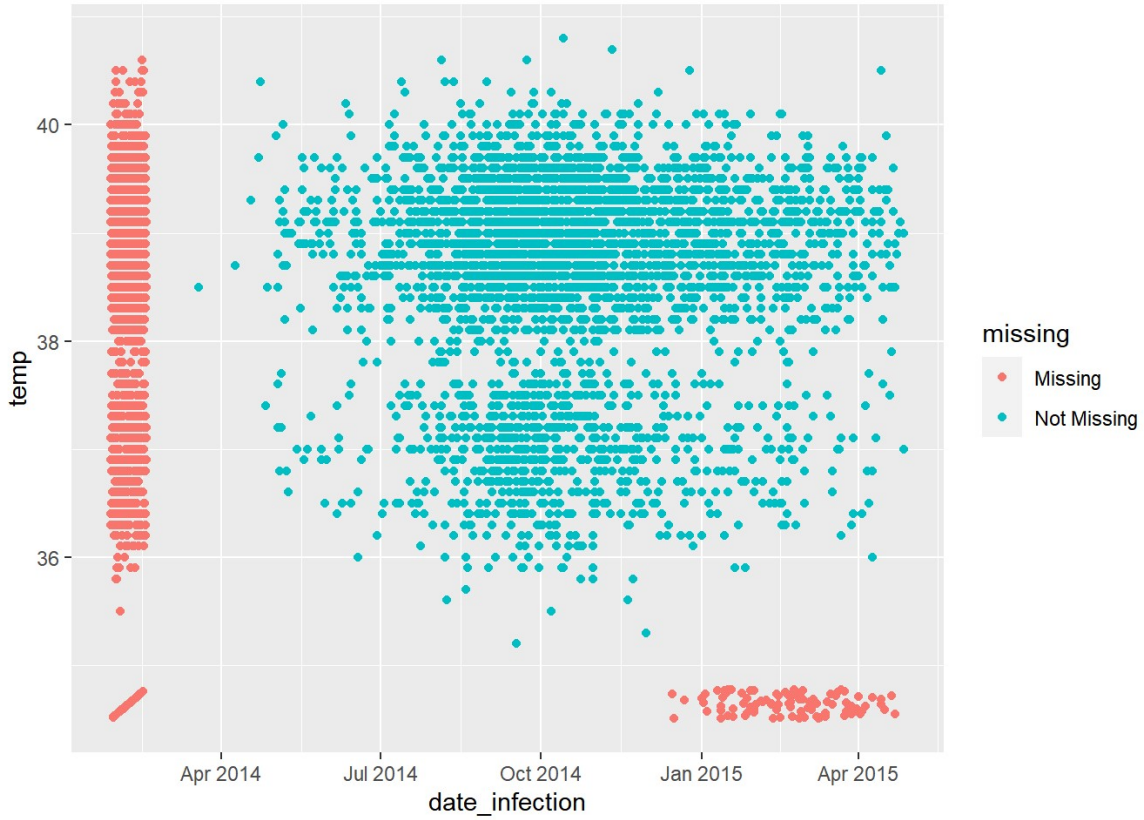
```
ggplot(
  data = linelist,
  mapping = aes(x = age_years, y = temp)) +
  geom_miss_point()
```



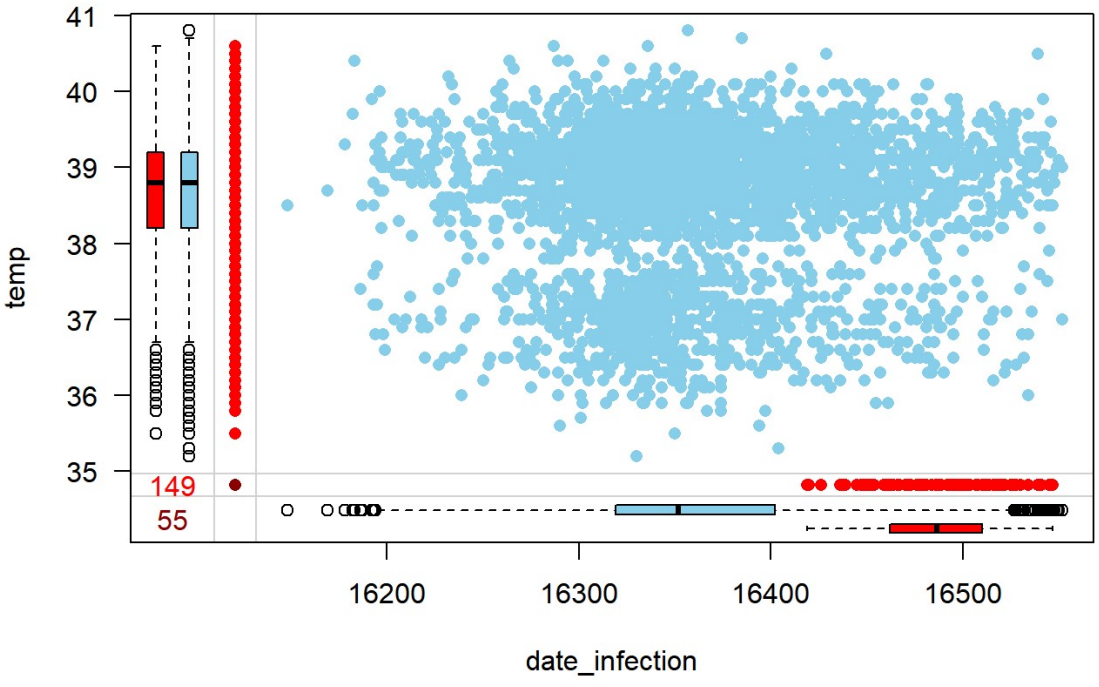
#Question/Action

Based on your answer to the last question, about which variable seems to show the largest shift associated with missing data, produce both styles of margin plots as a way of investigating the relationship you saw earlier.

```
ggplot(
  data = linelist,
  mapping = aes(x = date_infection, y = temp)) +
  geom_miss_point()
```

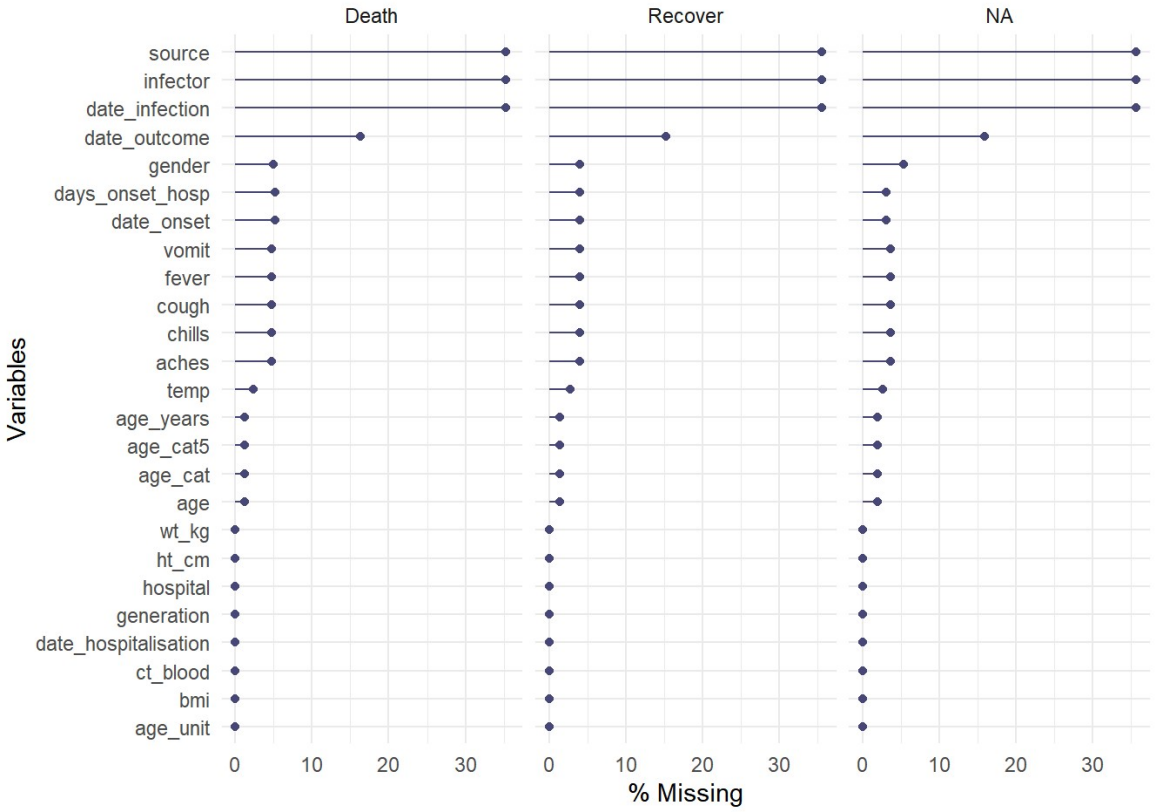



```
mydat=linelist[,c("date_infection","temp")]
VIM::marginplot(mydat,las=1,pch=16)
```



We use a nanair plot of missing rates as a function of a categorical, such as outcome

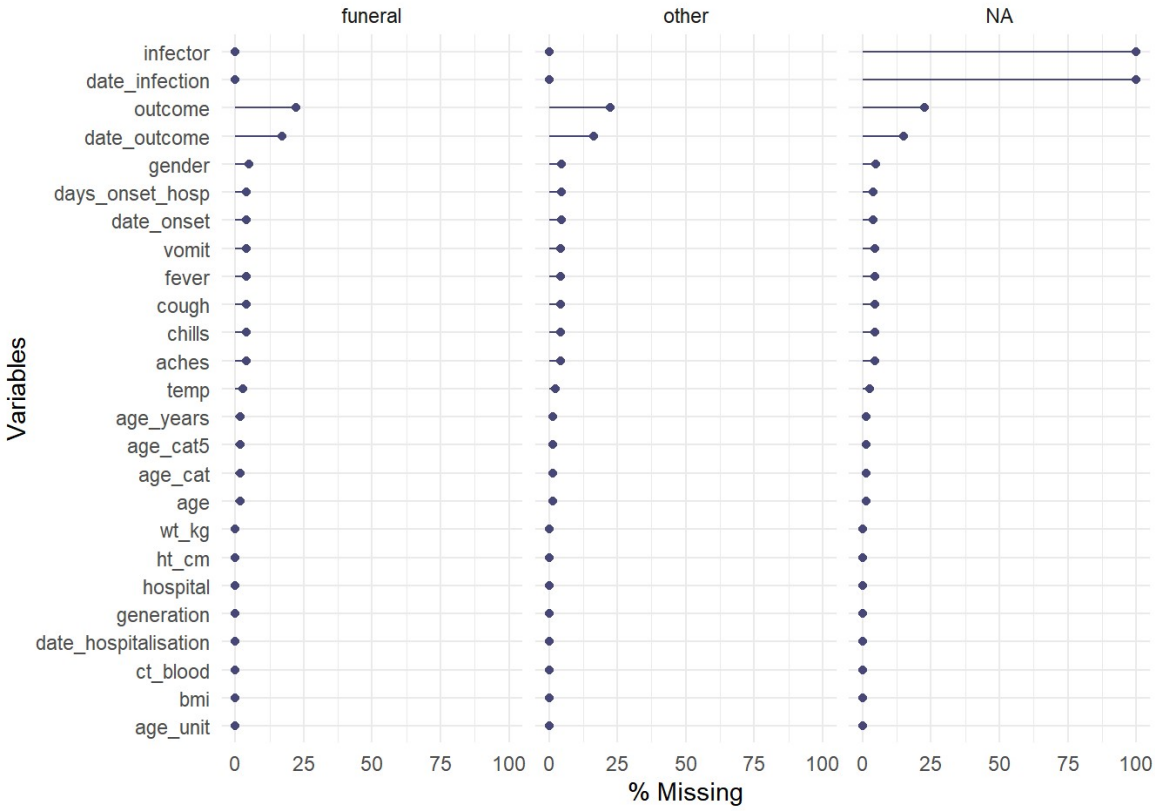
```
linelist %>%
  gg_miss_var(show_pct = TRUE, facet = outcome)
```



Question/Action

Use this type of faceting by a factor to look at another factor that might related to missing values

```
linelist %>%
  gg_miss_var(show_pct = TRUE, facet = source)
```



#visualizing where in the data set there are missing values

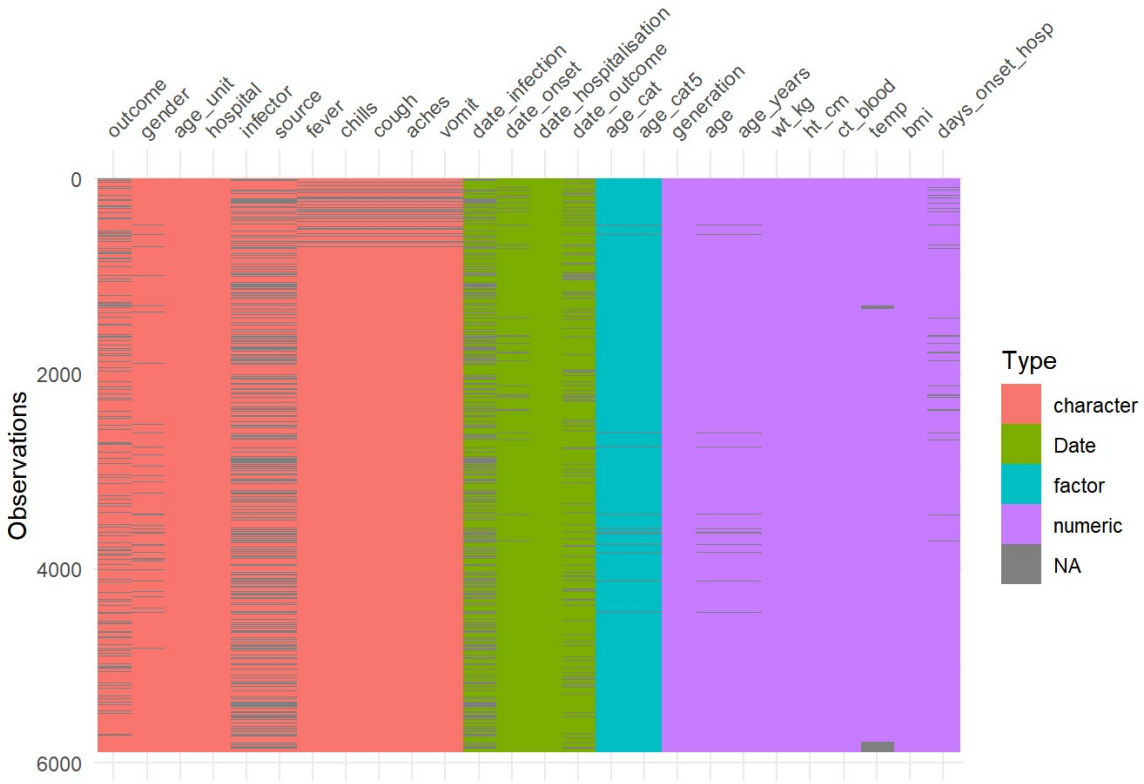
The visdat function can show us the patterns of missing data and the data types

```
require(visdat)
```

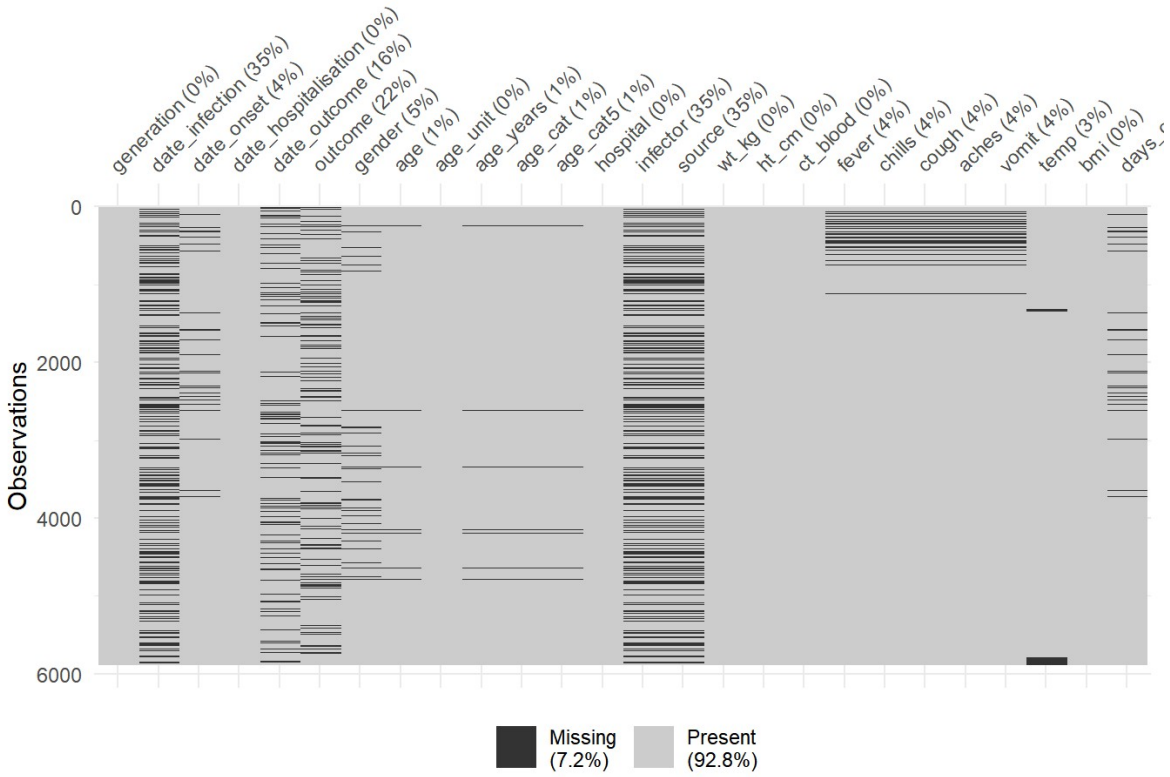
```
## Loading required package: visdat
```

```
## Warning: package 'visdat' was built under R version 4.2.3
```

```
vis_dat(linelist)
```




```
vis_miss(linelist)
```

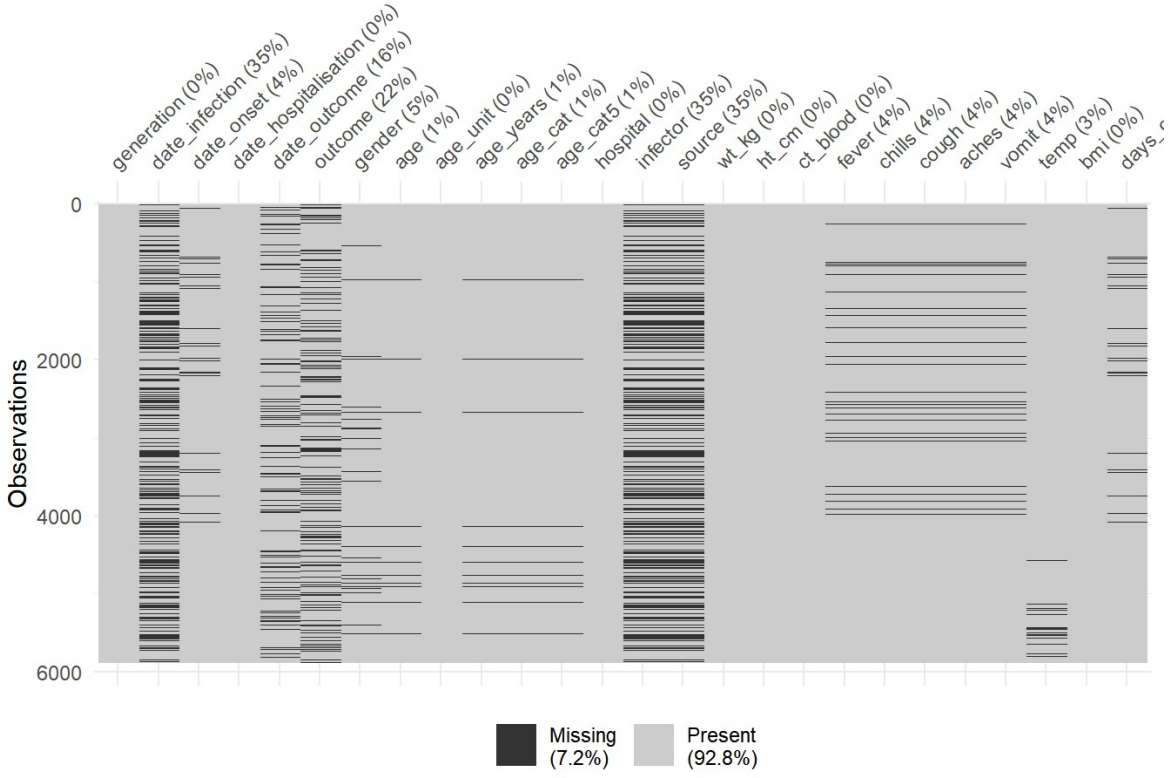


Question/Action

Sort the data set by some value, maybe age, date of admission, or some other variable.

Then send the ordered data set results into vis_miss so the missing information is ordered- do you see any patterns?

```
sortedData <- linelist[order(linelist$date_hospitalisation),]
vis_miss(sortedData)
```

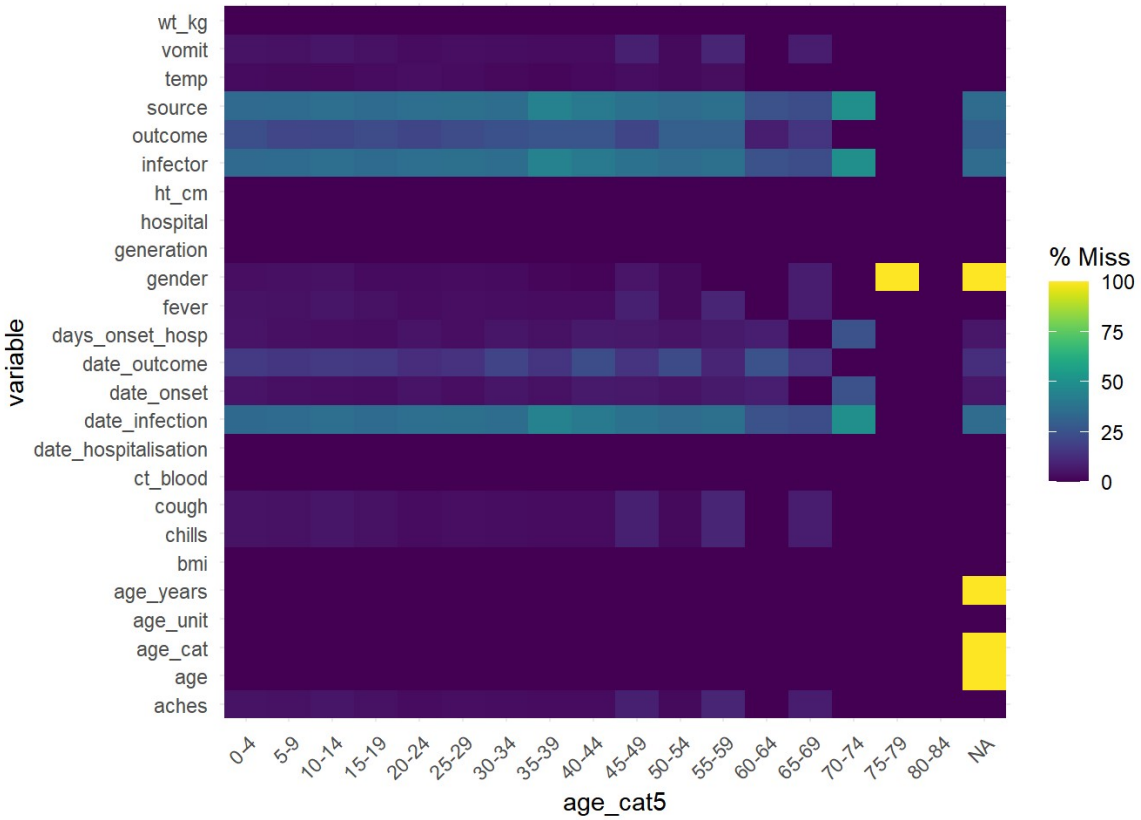


It appears that infector, source, and date_infection seem to match when data is missing, as well as fever, cough, aches, and vomit.

We can heatmpa the rates of missing information by some factor as well

```
gg_miss_fct(linelist, age_cat5)
```

```
## Warning: There was 1 warning in `mutate()`.
## ℹ In argument: `age_cat5 = (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 <]8;;https://github.com/njtierney/naniar/issueshttps://github.com/njtierney/naniar/issues]8;;>.
```



We can create a “shadowed” data set that has columns that indicate if there are missing values in other columns

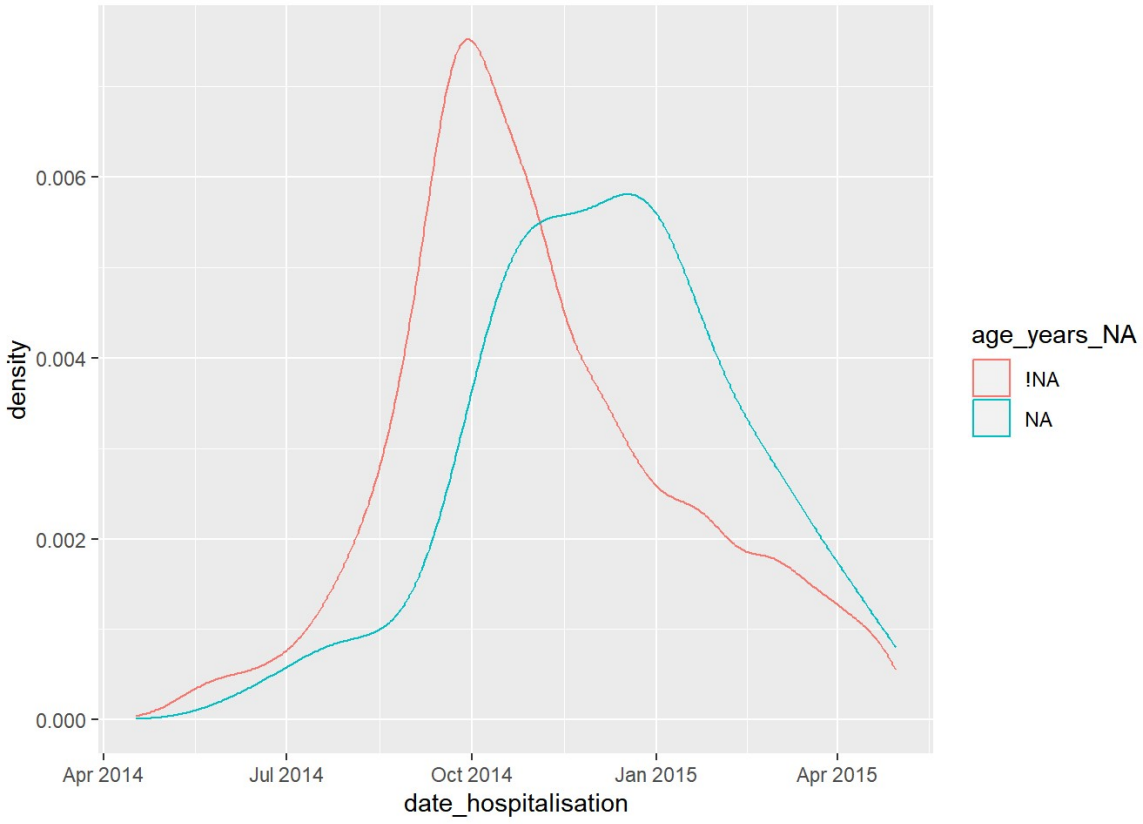
```
shadowed_linelist <- linelist %>%
  bind_shadow()

names(shadowed_linelist)
```

```
##  [1] "generation"             "date_infection"
##  [3] "date_onset"             "date_hospitalisation"
##  [5] "date_outcome"           "outcome"
##  [7] "gender"                 "age"
##  [9] "age_unit"               "age_years"
## [11] "age_cat"                "age_cat5"
## [13] "hospital"               "infector"
## [15] "source"                 "wt_kg"
## [17] "ht_cm"                  "ct_blood"
## [19] "fever"                  "chills"
## [21] "cough"                  "aches"
## [23] "vomit"                  "temp"
## [25] "bmi"                    "days_onset_hosp"
## [27] "generation_NA"          "date_infection_NA"
## [29] "date_onset_NA"          "date_hospitalisation_NA"
## [31] "date_outcome_NA"        "outcome_NA"
## [33] "gender_NA"              "age_NA"
## [35] "age_unit_NA"            "age_years_NA"
## [37] "age_cat_NA"             "age_cat5_NA"
## [39] "hospital_NA"            "infector_NA"
## [41] "source_NA"              "wt_kg_NA"
## [43] "ht_cm_NA"               "ct_blood_NA"
## [45] "fever_NA"               "chills_NA"
## [47] "cough_NA"               "aches_NA"
## [49] "vomit_NA"               "temp_NA"
## [51] "bmi_NA"                 "days_onset_hosp_NA"
```

We can then create histograms or time series of the distributions of missing data

```
ggplot(data = shadowed_linelist,          # data frame with shadow columns
  mapping = aes(x = date_hospitalisation, # numeric or date column
                colour = age_years_NA)) + # shadow column of interest
  geom_density()                          # plots the density curves
```

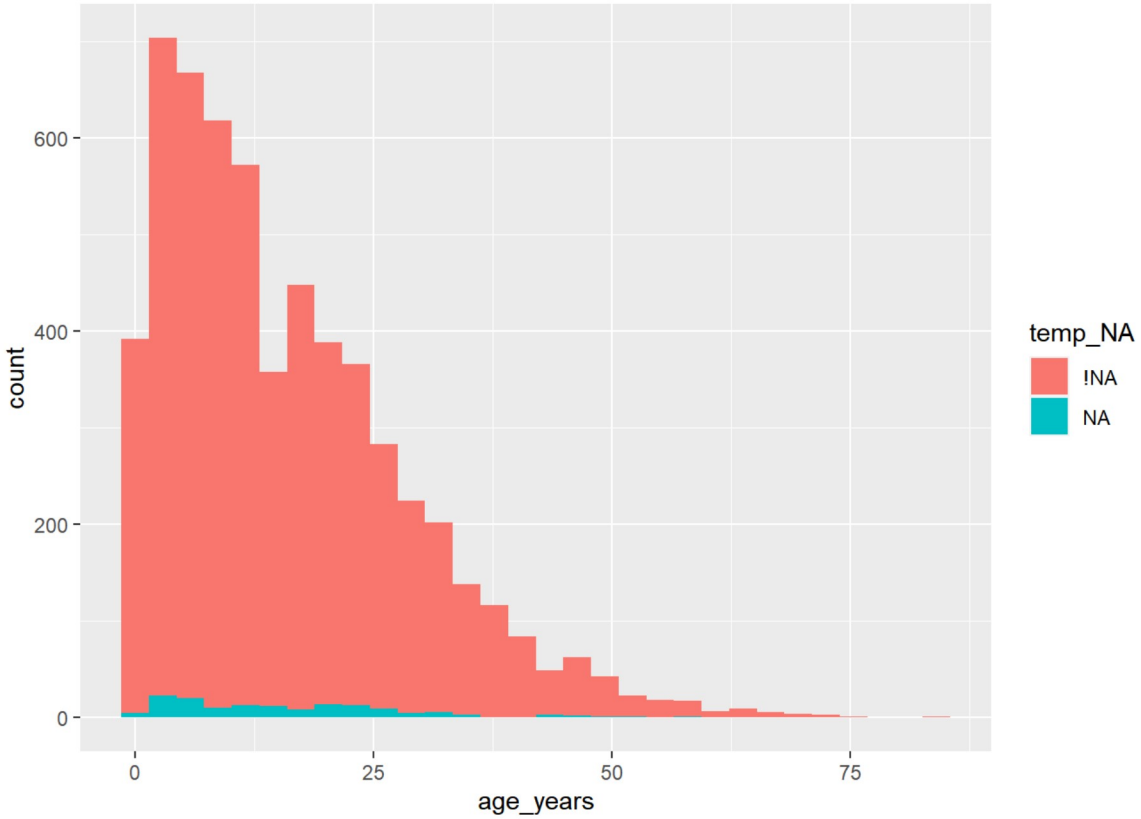


Looking at where in a distribution the data is missing

```
ggplot(data = shadowed_linelist, aes(x=age_years,fill=temp_NA))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 86 rows containing non-finite values (`stat_bin()`).
```



#Modelling missingness

We can build a model to predict when data will be missing

The function add_prop_miss adds a column to the data that computes the proportion of missing data on the line and adds that to the last column

```
linelist2=add_prop_miss(linelist)
head(linelist2)
```

	generation <dbl>	date_infection <date>	date_onset <date>	date_hospitalisation <date>	date_outcome <date>	outcome <chr>
1	4	2014-05-08	2014-05-13	2014-05-15	<NA>	*NA*
2	4	<NA>	2014-05-13	2014-05-14	2014-05-18	Recover
3	2	<NA>	2014-05-16	2014-05-18	2014-05-30	Recover
4	3	2014-05-04	2014-05-18	2014-05-20	<NA>	*NA*
5	3	2014-05-18	2014-05-21	2014-05-22	2014-05-29	Recover
6	3	2014-05-03	2014-05-22	2014-05-23	2014-05-24	Recover

6 rows | 1-7 of 28 columns

It is then possible to build a predictive model (a regression tree) that will predict the value of prop_miss_all, in other words the proportion of missing data based on the other variables

In the discussion of imputing, we saw how predictive models could be used to impute missing values.

In this application, we are using the predictive model to predict when a particular row in the data frame will have missing information.

Models such as regression trees can tell us which variables are being used to make predictions, in a regression tree these are called the “importance” value of each predictor variable. We fit a regression tree to the model and then use the importance values to determine which of the predictors is informative about when data is likely to be missing. Then we can go back to the visual methods to understand what the relationship is.

Below, we fit a regression tree to the data to predict the proportion missing variable, using all the other variables, and then look at the summary() of the model to find the importance values. The summary of this model is quite long, we have to dig through it a bit to find the importances.

```
require("rpart")
```

```
## Loading required package: rpart
```

```
rmodel=rpart(prop_miss_all~., linelist2)
```

The model importance values tell us which variables are effective at predicting missing data

```
summary(rmodel)
```

```
## Call:
## rpart(formula = prop_miss_all ~ ., data = linelist2)
##   n= 5888 
## 
##           CP nsplit rel error   xerror       xstd
## 1 0.03849279      0 1.0000000 1.000281 0.02338594
## 2 0.03351233      2 0.9230144 1.086485 0.02568671
## 3 0.02759585      3 0.8895021 1.051708 0.02463689
## 4 0.01602697      5 0.8343104 1.011147 0.02217159
## 5 0.01289415      6 0.8182834 1.007242 0.02221775
## 6 0.01000000      7 0.8053893 1.009549 0.02238935
## 
## Variable importance
##             infector                 temp                fever 
##                   79                   12                    5 
##               source       date_infection             hospital 
##                    1                    1                    1 
## date_hospitalisation 
##                    1 
## 
## Node number 1: 5888 observations,    complexity param=0.03849279
##   mean=0.07218071, MSE=0.00578312 
##   left son=2 (5373 obs) right son=3 (515 obs)
##   Primary splits:
##       infector             splits as  LLLLLLLRLLLLLLLRLLLLLLLRRLLLLLLRLLLLRLLLLLLLLLLLLRLLLLLRRLLLLLLRLLLLLLLLLLLL
LLLLLLLRLLLLLLLLLLLLLLLLLRLLLRLLLLRLRLLLLLLLLLLLLLLLLLRRLRRLLLLLLLLLRLLLLRLLLLLLRLLLLLLLLLLLLLLRLLLLLLRLRLRLLLRLLLLLLLL
LLLLLLLLLLLLLLLLLLLLLLLLRLLLLRLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLRLLLRRLLLLLLLLLLLRLLLLLLLLLRRLRLLRLLLRLLLLRLLLLLLRLLLLRL
LLLLRLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLRLLRLLLLRRLLLLLRLLLLRLLLLRLLLRLLLLLLRRRLLLLLLLRLLRLLLRLLLLLLLLLLLLLRLRLLLL
LLLLRLLLLLLLLLRLLRLLLLLLLLLLLRLRLLRLLRLLLLLLLRLLLLLLLLLLLRLLLLLLRRLLLLLLLLLLLLLLLLLLLLLLLLRLRLLLRLLLLLLLLLLLLLLLLLLRLLLL
LLLRRLLLLLLLLLRRLLLLLLLLLLLRLLRLLRRLLRLLLLLLLLLLLLLLLLLLRLLLLRLLLLLLLLRLLLRLLLLLLRRLLLLLLLLLLLLLLLLLLLLLLLLLLLRLLRLLLRLL
LLLLLLLRRLLLRRLRLRLLLLLLLLLLLLLLLLLLLRLLLLLLLRLLLLLLLLLLRLLLLLLRLLLLLLLLRLLLLLLLRLRLLRLLRLLLLLLLLLLLLLLLLLLLLLLLLLL
LLLRLRLLLLLLLLLLLLLRLLLLLRLLLLLLLLLLLLLLLLLLLLLLLLLLLRRLLLLLRLLLLLLLLLLLLLLLLRLLLLLLLLLLRLLLRLRLRLLLLLRLLLLLLRLLLLLLLL
LRLLLLLRLLRLLLLLLLLLLLLLRLLLLLLLLLLRLRRLLLLLRLLLLLLLRRLLLLLRLLLLLLLLLLRLLLRLRRLRLLRRLLRLLLLLLLLLLLLLRLLLLLLLLLLLLLLLLL
LLLRLLLLLLLRLLLLLLLLLLLLLLRLLLLLLLLLLLLLLRLRLLLLLLLLLRLLRRLLLLLLLLLRLRLLRLLLRLLLLLLLLLLLLLLLLLLLLLLLLLLLLRLLLLLLLL
LLLLLLLRLLLLLLLLRRLLLLLLLLLRRLLLRLLLLLLLLLLLLLLRLRLLLLLLLLLRLLLRLLLLLLRLRLLLLLLLLLRLLLLLRLLLLLLLLLLLLLLLLLRLLLRLRRL
LLLLLLLRLLLLLLRLLLLLLLLLLRLLLLLRLLLLLLLLLRLLRLRLLRLLLLLLLLLRLLLLLLLRLLLRLLLLRLLRLLLLLLLLLLLLLLLLRLLLLLLRLLLLLLLLLRLR
LLLLRLLLLLRLLLLLRRLLLLRLLLRLLLLLLLRLLLRLLLLRLLRLLLLLLRLLLRLLLLLLLLLLLLLLLLLRLLLLLLLLLLLLLLRLLLLLLLLRLLLRLLLLLLLRLLLLL
LLLRRLLRLLLLLLLLLLLRRLLRLLLLLLLRLLLRLRRLLLLLLLLLLLLLLRLLLRLLLLLLLLRLLLLLLLLLLLLRLLLRLLLRRLLLLLLLRLLLLLLLLLLLRLLLLLLLRLL
LLLLRLRLLLLLLRLLLLLLLLLLLLLLLLLLLLLLLLLRLLRLLLLLLRLLRRLLLLLLLLLLLLLLLLLLLRLLLRLRLLLLLLLLLLLLLLLLLLLLLLLLLLLRLLLLRLLLLLLLL
LLLLLRLRLLLLLLRLLLLLLLRLLLLLLLRLLLRLLLLLLLLRLLLLLRLLLLLLRLLLLLLLLLLLLLLLLRLLLLLLLLLRLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLR
LLLRRLLLLLLLLLLLLLLLLLRLLLRLLLLLLLLLLLLLLLLLLLLLLLLLLLRLRLLLLLRLLLRLLLLLRLLLLLLLLLLLLLLLLLLLLLLLRLLLLLLLLLLLLLLRRR
LLLLLLLLLLLLLLLLLRLRLLLLLLRLLLLRLLLRLLLRLLLRRLRLLLLLLRLLLLLLRLLLLLRLLLLLRLLRRLLLRLRLLLRRLLLLLLLLLLLLRLLLRRLLLLLLLLLRLL
LLLLLLLRLLLRLRLLLLLLLLLLLLLLRLLLLLLRLLLLLLLLLLRLLLLLLLLRLRLLLLRLLLLLLLLLLLLLLLLLLLLLLRRRRLLLLLLLLLLLLLLLRLLLLLLLLLLLL
LLLLLLRLLLLLLLLLLRLLLLLRLLLRLLLRLLLLLLLLLLLLLLLLLRLLLLRLLLLLLLLLLLLLLLLLRLLRLLLLRLLLLLRLLLLLLLLRLLLLLLRRLLLLRLRLLLLLLL
LLLLLLLLLLLLLLLRLRRLLLLLLRRLLLLLLLLLRLLLLLLLLLLLLLLRLLRLRRLRLLLLLLLLLRLLLLLLLLLLLLLLLRLLLLLLLLLLLRLLRLLRLLRLLLLLLLLLRRLLL
LRLLLLLRRLLLLLLLLLLLLRLLLLRLLLLLRLLLRLLLLRLLLLLLRLRRLLLLLLRLLLLRLLRLLLLRLLRLLRRLLLLLLLLLRRLLLRLLRLRLLLLLLLLLLLLLLLL
LLLLLLLLLRLLRLRLLLLRLLLLLLLRLLLLLLLRLLLLLLLLRLLLLLLLRLLLLLLLLLLRLLLLLRLLLLLLLLLLLLLLRLLLLLLLLLLLRLLRLLLRLLLLLLLRRLLLLRL
RRLLLLLLLLLLLLLLLLRRLLLLLRLLLLRLLLLRLLRLLLLLLLL, improve=0.135474500, (2088 missing)
##       temp                 < 37.95     to the right, improve=0.042394820, (149 missing)
##       ct_blood             < 20.5      to the left,  improve=0.006225116, (0 missing)
##       days_onset_hosp      < 2.5       to the right, improve=0.003747335, (256 missing)
##       date_hospitalisation < 16373.5   to the right, improve=0.002134539, (0 missing)
## 
## Node number 2: 5373 observations,    complexity param=0.03351233
##   mean=0.06766024, MSE=0.00545252 
##   left son=4 (4191 obs) right son=5 (1182 obs)
##   Primary splits:
##       infector     splits as  LLRLLLL-LLLLRLL-LRLRLRL--RRLRLL-LRLR-RLLRLLLRLLLR-RLRLL--LRLRLL-RRLLLLLLRLLLLLLLLLL-
LLLRLRLLLLLLRRLLL-RRR-LRRL-L-LLRLLLLRRRLLLLRR--L--RLLRLLLLL-RLLL-LRRLLL-LLLRLLLLRLLRR-LLRLLL-R-L-LRL-LRLRLLLLLLLRLRRR
RLRLLLLLLRLLLLLLRL-LRRL-LLLLLRLRRLLRLLRLLLRLLLLRLRLLLRLR-LLL--LRLLRLLRRL-RRLRRLLLR--L-LL-LLL-RLLL-LLRRRL-LRRR-RLRLL-LLL
RLLLRLLRLLLRLLRLLLLLLLLRLLLRLLLLRRRR-LL-LRLL--RRLRL-LLRL-LRLL-RLR-LLLRLR---LLLRLLR-RL-LLL-LLLLRLLLLRLL-L-RLRLRLLL-RLL
LRLLLR-RL-LLLLLRLLLRR-L-LL-LL-LLRLLRL-LLLLLLLLRL-LLLLLL--RLRLLRLRLRLLRLLLLLLRLL-R-RRL-RLRLRLLLLLRLRLLLLL-LLLRLLR--LLR
```

LRRLRL--LRLRLLLLLLL-RR-RL--LL-RLLLLRLRLLLLLRLL-RLRL-LLLLRLL-RLR-RRLLLL--LLLLRLLLRRLLLRRLRLLLRLLRLRL-RL-RLL-LRRRRLRLL
--LRL--L-L-RLRRRLLLRLRRLRRLL-LLRLRRL-LLLLRLLLRLR-LRRLLR-RRLRLRLRL-LRLLLLR-R-LR-RR-RLLLLLRRRRLLLLRLLLRLLRRRLLLLLRR-L-R
LLRLLLLLRLRR-LLLLL-LLLRLLLLLLLLLLRRLLRLLLRLLRRRR--LRRRR-LLLLRLLLRRLLLRL-RRLLLLLLL-LLL-L-L-LLRLR-RLLRRR-LRLLRRLLLR-LLLLR
-LL-LLRLLRRRLLLLR-LLLLLLLLR-R--LLRLLR-LLLRLRL--LLRLL-RLRLLLRRR-RRR-R--L-RR--LL-RRLLLLRLLLRLL-RRLLLLLLRRRLLRLLLRLL-LLL
RRRR-RLRRLLRRLLRLRL-LLLLRLLLLRLL-L-RRLRRRRLLL-LL--RLRRLLRLLL-L-RL-RLR-LRLRLLLLLRLLLLRLLRLLLRRRLLRRLRL-RRLLRLLLLLLLRRR
-LLRRLLRR--RLLLLRLR--LRL-RLRRLLLRLLLRLRLL-L-RRLRRLLL-LRL-RLLRLR-R-RRRLRLLRL-RRLLLL-LLRLLLRLRLRLRLLLRL-LRL-L--RLLRLLLL
-LRLRLL-LLRLLLLLLL-RLLRLL-RRLLRRLLRL-LL-R-LR-LRLRRLLL-RLLRRLR-RLR-LRLL-RR-LLRLLRLLLRLLLLR-LLRLLRR-LRLRLLLRL-R-RLLR-RR
RLL-RRLLL--RLLL-LLL-RRRLLLR-LRL-LLRL-LL-LLRLL-LLR-LLLLRLRRLLLLLLLLRL-LLLLLLLLLLRRLLR-LLLRLLLR-LLL-LLRRLL-LLRRLLRL--RL
-RLLLLLRRLLL--LL-LLLLRLR-LRL-L--RRRLLLRLRRLR-LLL-RLRLRRLL-LRLLRRLLLLLL-RLL-RRR--LLLLLRLR-LLLLRLRRRLR-LRRLLL-LLLLLR-L-
LLLLRL-LLRLLLLLLLLLLLRLRRLRRLL-RR-RLLRRR-RL--LLRLRRLLLLRRRLRRLR-RRL-R-RLLRRRLRLLRRLLRLRLRLLLLLRL-RLLL-LRLLLLLLLLLLR-L
-LLLRR-LLRRLLL-RRRLLLR-LRL-LRLLLLRRLL-LLRLL-LRLLRL-LRRLLRLLLRLLLRL-LRLLLRRLR-LLRLLLLLLLLLLRRLLRLLRRLLLLRLLLLR-RLL--LL
RLLLLRLLLLLLL-RLR-LLLLLLLLRRRLLRLRLLLLLLLLLRRLRLLL-L-LRLLL-LRL-RRLRLL-LLLLLLRLLRLLRLLRRLLLLLL-LRLRLRRRLRLLLL---LLLLLLR
LLRLRLRLRL-L-LRRLLL-RLLR-RLL-LRL-LRL--L-LRLRLL-RLLRL-LLLLR-RLLLR-LL--LLL-L-LRR--LLLRLLLLLLLLR-LRL--RLRRLRLL-RLLLLRRLL
-RLR-L-LLRLLLLRRRLLL-LLLLLL-RRRRLLRLLR-LLLLLLLR-L-LLRL-LLLRLRLLRRRRLLLLRLRLLRLL----LLLLRRRLRLLRRR-LLLLLRRRRLLLRLLRRL-
LLLLLLRLLRL-LLRLR-LLL-RRL-LLLLLLLRLLLLRRR-LLLL-LRLLRLLRLRLLRRLL-LL-RRLL-LLLRL-RLLLRRLL-LLRLRR--RLRL-L-LLLLLLLLRLLLRLL
LRRLRLL-R--LLLLRL--LRLLLRLRL-RRLLLLLRLRRLR-LR-L--L-LLLLLLRLR-LLLLRLLLLLRRLL-RRRRRRRLLLL-LL-RL-RR-RLLLRLLR--LRLL-LLLLR
--LLRRLLRLLLLL-LLLR-LLLLL-LLR-LLRLL-LRLLLL-L--RLLLRL-LLRL-LL-RRRR-LL-RL--LRLLLRLRL--LLL-RR-L-LLRRLLLRLLRLLRLLRRLLLRRL
RL-RL-R-LRRL-LRRRRLL-LRLLLLRL-RLRRRRRL-LLLLLRL-LLLLRRRR-RLRRL-RLLLRLRRLRLLLR-LRLRRRLLLL-RR-RRL-LLLLRLL--RLLR-R--LLLLR
LLLRRLLLLRL--RLRRL-RRLL-RRRR-RL-LRRLLRRL, improve=0.030339970, (2088 missing)
temp   < 37.95   to the right, improve=0.017539360, (133 missing)
ct_blood   < 20.5   to the left, improve=0.002614700, (0 missing)
age_cat5   splits as LLLLLLLRRRRRLLRLL-, improve=0.001979148, (34 missing)
date_outcome < 16382.5   to the right, improve=0.001581080, (774 missing)
Surrogate splits:
date_infection   < 16198.5   to the right, agree=0.644, adj=0.003, (1 split)
date_hospitalisation < 16203.5   to the right, agree=0.644, adj=0.003, (2087 split)
wt_kg   < 94.5   to the left, agree=0.643, adj=0.002, (0 split)
date_onset   < 16203.5   to the right, agree=0.643, adj=0.001, (0 split)
age_cat5   splits as LLLLLLLLLLLLLLLRL-, agree=0.643, adj=0.001, (0 split)
##
Node number 3: 515 observations,   complexity param=0.03849279
mean=0.1193428, MSE=0.006794811
left son=6 (404 obs) right son=7 (111 obs)
Primary splits:
infector   splits as -------L-------L-------LR------L----L------------R-----LL------R
------------------L-----------------R---L----R-L----------------LL-LL---------L----L------L-------------R------L-L-L
---R---------------------------------L----L---------------------------------L---LL----------L---------LL-L--L---L----L
------L----R-----L--R--L----RL-----L----R----L---L------LLL-------R--R---L
------------R-R--------R---------L--L-----------L-L--L--L-------L----------L------RL---------------------L-R---L
------------------R-------RR---------LL-----------L--L--RR--L----------------L----R-------L---L------LL
---------------------------L--R---L---------RL---LR-L-L-----------------L-------R-----------L------L---------L-------
L-L--L--R-------------------------------L-L-----------L-----L---------------------------LL-----L---------------L
---------L---L-L-L-----R------L----------R-----L--R-------------R---------L-LL------L-------RR-----L---------L---L-LL
-L--RL--L-------------L--------------------L-------R--------------R------------L-L----------L--LR----------L-L--R---L
-------------------------------R---------------L--------LL--------RL---R----------------L-R--------L---R------L-L
---------L------L------------------R---L-RL--------L------L----------R------L----------L--L-L--L--------L-------L---R
----R--R---------------R-------L---------L-L----L-----L-----LL----L---R-------R---R----R--L-----R---R
------------------L---------------L--------L---R------L--------LL--L-----------RL--R-------L---L-LR-----------L---L
--------R------------R---L---LL--------L-----------R------L------R-R------R----------------------L--L------L--RL
------------------L---R-L---------------------------L----R-------------L-L-----L-------R-------L---L----------L-----L
------L---------------L---------L---------------------------------L---LR---------------R---L
-------------------------------L-L-----R---L------L----------------------R--------------LLL-----------------L-L
------L----L---L---L---LL-R------R-----L-----L-----L--LL---L-L---RL-------------L---LL--------R---------L---L-R
-------------L------L----------L--------R-L----L-----------------------LLLL--------------L------------------L
-----------L-----L---L---L---------------L----L----------------R--R----L-----L--------L------LL----R-R
----------------------L-LL------LR---------L-------------L--L-RL-L---------R--------------L-----------L--R--L--R
--------RR----R-----RR------------L----R-----L---L-----L------R-LR------L----R--L----L--L--LL---------LL---R--L-R
--------------------------R--L-L----L-------L--------L--------L-------R--------L-----L--------------R----------L--L
---L-------LL----L-RR----------------LL-----L----L----L--L--------, improve=0.39040250, (0 missing)
temp   < 38.05   to the right, improve=0.30936080, (16 missing)
days_onset_hosp < 2.5   to the right, improve=0.09334726, (109 missing)
ct_blood   < 20.5   to the left, improve=0.06162580, (0 missing)

```
##       wt_kg           < 14.5      to the left,  improve=0.01281189, (0 missing)
##   Surrogate splits:
##       ht_cm < 234       to the left,  agree=0.788, adj=0.018, (0 split)
##
## Node number 4: 4191 observations,    complexity param=0.02759585
##   mean=0.07539967, MSE=0.006481
##   left son=8 (3323 obs) right son=9 (868 obs)
##   Primary splits:
##       temp                < 37.95     to the right, improve=0.021427170, (62 missing)
##       infector            splits as  LL-LLLR-RLLL-LL-R-L-L-R----L-LR-R-R---RL-LRL-LLL---L-LL--L-L-RL---LLLLLL-LLR
LLLLRLR-RRL-L-RLLLLL--LLL-----L--L-L-LL-RLLL---LLLL----L---LL-LLLLL--LLL-L--LLL-LLL-LLRL-LL---LL-LLL---R-R-R-L-L-RRLL
LRL-L----L-LLLLL-RLLLL-L-L--R-LLLLL-L--LR-LL-LLL-RLLL-L-LLL-L--LLR--L-LL-LL--R---L--RLL---L-LR-LLL--LLL-LR---L-L-----
L-LL-LLL-LLL-LR-LLL-LR-RLLRLLLL-LLL-LLRL-----LL-R-LL----L-L-RL-L-L-LL--R--LLL-R----LLL-LL---L-LLL-LLLL-LLLL-RL-L--R-L
-RLL--LLL-RLL---L-LLLLL-RLR---R-LL-LL-LL-LL-L-LLLLLLLR-L-RLLLRL---L-LL-L-L-LL-LLLLLL-LL-----L--L-L-LLRLR-L-LLRRR-LLL-
LL---RL-L--L-L--L-L-RLLLLRL-----L--LL--LLLR-L-LLLLL-LL--L-L-LRLL-LL--L----LRLL--LRLL-LLL--RLR--L-LLL-LL-R-L--R--LL-L
----L-LL--L-L--L-L--L---LLL-L--L--LL-RL-L--R-LLLR-LLL-L--R--LR----L-L-L-L-L-LLLL----L------LLLLL----RLLR-LRL-LL---RLL
RL---L--RL-LLLL-L---LLLRL-RLL-LLLRLLLLL--LL-LLL-LL------L-----LLLL-RLR--LLL-R---LRLLLLL-LLL-L-L-LL-L---LL----R-LL--LL
L--LLRL--LR-LL-LL---LRLL--LLLLLLLL-----LL-RL--RLL-L-L--LR-LL--L-RLL-----------L-----RL---RLLL-LLL-LL---LRLLLL---LL-LL
R-LL-LLL------L--LL--LL-L-R-LRLL-LLLL-LL-L---R----LRR-LL---L--LL-LLL-L--L--L--L-L-RLLLL-RLLR-RL-LLL---RR--L-L---LL-RL
LLRLL----LR--LL-----LRLL-L---L-L--L--RLL-LLL-R-LL-L---R--LLL-L-R--LL-L-------L-LR-R---LLLL-LR-LLL-L-L-L-LLL-R-R-L-L
---LL-LLLL-L-L-LR-LL-LRLRLLL--LL-LR---RL--LL-L-LL---L--L-L--LLL--RL--R---L--L-LL----LR-LL-LRL-LLRL--LR-LR---L-L-LLL-R
----LL-----LL---LLL---LLL-LLL----LLR--L-L-RL-L-LL-LR-RR-LL--LLLL-R--LLLLLLRL-L-LLLLLLLLLL--LL--LLL-LLR--LLR-RL--LL-LL
--LR-R---L--LLRLL--RLL--LL-LRLR-L--L-R-L-----LRR-L--L--RLL--L-L--LL-L-LL--LLLLLL--RR------LLLLL-L--LLLL-L---R--L--LRL
-LLLLL--L-LLLL-L-RL-LLLLRLLLLLL-L--L--LL-----LR-----L--RL-L--LRLL---L--L----L----RL---L-LL--LL-L-L-RLRLL-L--LLR-L-LLL
RLRLLLL--L-LLL---LL--LLL----LRL--L-R-L-LLLL--RL-LR-LR-L-LL-L-L--LR-LRL-LLL-R-L-RRL--L--LL-LRLRLRLRRL--LL-LL--LLRL-LLL
L---RL--LR-LLLL-LLLLLLL--R--LLLLLLLL---LL-L-LLLLRRLL--L-LLL-L-L-LLL-L-L---L-LL-LLLLLL-RL-RL-LL--LRLLR-L-R-L---L-LLLL
---LLLLLL-LR-R-R-L-L-L-R--LLL--RL---LR-L-L-L-L--R-L-L-LL--RL-L-RLLL---LRL--LL--LLL-L-L----LLR-LLLLRLLL--L-L---L--L-LL
--LRLL--LL--L--L-RL-LLLR---LLR-LLLLLL-----LL-RR--LLLLLLL--L-LR-L-LRL-L-LL----LLLR-L-LR-LR----LRLL---L-LR----LLRLL----
LLL-RR--L-LLLLLL-RL-L-LL-R--LLL---R-LLLLLLL-RLLL----LLLL-L-LL-RL-L-LL--LL-LL---LL-LLL-L--LLL--LL-LL-L-----L-L-L-LLLRL
LLR-LLR-LRL--L-RL----LLRR-L--L-LRL-L-L---LLLRL-L--R--L--R--R-LLRLLL-L--RLLL-LLLLL--LL--------LRLL-RL--L-----LLL-LL---
L-LR-RLLL---LL--LL-LRLLL-LLR--LLLLR-LL--RL-RL-L-LLLL-L---RRL-L-LL-L-LL------LL--L--L-LLL-L-L--LLL----L-LL--LLL-RL-LL-
LL--LRR--L-L--R---L--L-L----LL-R-LRLL-L--L-----L-LLLLR-L-LLLL------L--L--LRL-R--L-LRR--L-L---LLLL------L-RLRL-LR---LL
-----LLLL-RLR--LLLR-L---L--R---LR-------R-L--LL--L, improve=0.005043291, (2080 missing)
##       ct_blood            < 20.5      to the left,  improve=0.002788106, (0 missing)
##       date_hospitalisation < 16325.5   to the left,  improve=0.002515674, (0 missing)
##       date_onset          < 16320.5   to the left,  improve=0.002441433, (83 missing)
##   Surrogate splits:
##       fever splits as  RL, agree=0.977, adj=0.886, (62 split)
##
## Node number 5: 1182 observations
##   mean=0.04021866, MSE=0.0008404332
##
## Node number 6: 404 observations,    complexity param=0.01602697
##   mean=0.09234577, MSE=0.004968948
##   left son=12 (272 obs) right son=13 (132 obs)
##   Primary splits:
##       temp            < 38.05     to the right, improve=0.27282140, (16 missing)
##       days_onset_hosp < 2.5       to the right, improve=0.06385318, (97 missing)
##       infector        splits as  -------L-------L-------L-------L----R------------------RR
-------------------------L---------------------L------L----------------RR-LL---------L----L------L
--------------------L-L-L-------------------------------------L----L--------------------------------L---LL----------L
---------LL-L--R---L----R------L----------L-----------------------------------------L-----L-----R---------R---R
------RLL--------------R---------------------------------L--L-----------L-L--L--L-------L----------L-------R
----------------------R-----L-------------------------------------RL-----------L--L------L----------------L
------------L---L------LL--------------------------L------L----------L---L--L-L-----------------L------------------
R------R---------L-------R-L--L---------------------------------L-L------------L-----L--------------------------RL
-----L---------------L---------L---R-L-L------------R----------------L-------------------------L-LR------L
--------------R---------L---R-LL-L---L--R-------------L--------------------L---------------------------------L-L
----------L--R-----------L-R------L-------------------------------------------R--------LR---------L
--------------------L----------L----------R-R---------L------L----------------------L--L--------L------L
-----------------R----------L--L-L--L--------L-------L-----------------------------------L---------R-L----L-----L
-----LL----L----------------------L---------------------------L---------------L--------L----------L--------RL--L
```

--------, improve=0.05409909, (0 missing)
ct_blood   < 20.5   to the left, improve=0.04544234, (0 missing)
age_cat5   splits as LLRLLLRRLRLLRR----, improve=0.01820498, (20 missing)
Surrogate splits:
infector splits as ----------R-------R----L------------------LL----------L
age_cat5   splits as LLRLLLRRLRLLRR----, improve=0.01820498, (20 missing)
Surrogate splits:
infector splits as ----------R-------R----L------------------LL----------L
ct_blood   < 20.5   to the left, improve=0.04544234, (0 missing)
Node number 7: 111 observations
mean=0.2176022, MSE=0.00113267
Node number 8: 3323 observations,   complexity param=0.02759585
mean=0.0694229, MSE=0.005113027
left son=16 (2862 obs) right son=17 (461 obs)
Primary splits:
infector   splits as LL---LR-LLLL--L-R-L-L-R----L-LR-R-R---RL--RL-LLL---L-LL--L-L-LL---LLLLLL--LR
wt_kg   < 94.5   to the left, agree=0.668, adj=0.023, (15 split)
ht_cm   < 229.5   to the left, agree=0.665, adj=0.015, (0 split)
bmi   < 235.034   to the left, agree=0.665, adj=0.015, (0 split)


---LL--L-L-L-L-LR-LL-LR-RLL----L--R---RL--LL-L-LL---L--L-L--LLL--R---R---L--L-LL----LL-LL-LRL-LLLL--L--LR---L-L-LLL-R
----LL-----LL---LLL----LL-LL-----L-L----L-RL-L-LL--R-RR-LL--L-LL-L--LLL-LLRL-L-LLLLLLLLLL--LL--LLL--LR--LLR-RL--LL-L
----R-L---L--LLRLL--RL---L--LLLR-L--L-R-L-----LRR-L--L--R-L--L-L---L----L--L-LLLL--RR------LLLL-----LL-L-L---L--L--LR
L-LL-LL--L--LLL-L-LL-LLLLRLLLLLL-L--L--LL-----LR-----L--RL----LRLL---L--L----L-----L------L--L--L-L-RLRLL----L-R---LL
LRLRLLLL--L-L-L----L--LLL----LRL--L-R-L-LLLL--R---R--R-L-LL-L-L--LR-LRL-LLL-R-L-RR---L--LL-LR-RLRLRRL--LL-LL---LRL-LL
LL---RL--LR-L-LL-LLLLL-L--L--LLLLLLLL----L-L-LLLLRRLL--L-LLL-L-L-LLL-L-L---L-LL-LLLLLL-RL-R--LL--LRLLLR---R-L---L-LL-
L---LL-LLL-LR-L-R-L-L-L-R---LL--RL---LR-L-L-L-L--R-L-L-LL--RL-L-RLLL----RL---L--LL----L----LLR---LLR--L--L-L------L-L
---LR----LL--L----RL-L-LR---LLR-LLLLLL--------LR--LL--L------L----LR--L-LL----L-LR-L-LR-L-----LR-L---L-LR------L-L
----LLL-RL--L-LL-LLL-RL-L-LL-R--LLL---R-LLLLL-L-R-LL----LLLL---LL-RL-L-L---L--LL---LL-LLL-L--LLL---L-LL-L-----L-L-L-L
LLRLLLR-LLR-LRL--L-RL----LLRR-L--L-LR--L-L---LLLRL-L--R--L--R--R-LLRLLL-L--RLLL-LL--L--LL--------LRLL-RL--------LLL-L
L---L-LR-R--L---LL---L-LRLLL-L-L--LLLLR-L---LL-R--L-LLLL-----L-L-L-L--L-LL------LL--L--L-LLL-L-L--LLL----L-LL--LLL-R
--L--LL--LRR--L-L--R---L--L-L----LL-R-L-LL-L--L-----L-LLLLR-L-LLLL---------L--LRL-R--L-LRL--L-L---L-LL------L-LLRL-LR
---LL-----LLL--R-R--LLLR-L------R---LL-------L-L--LL--L, improve=0.007752307, (1610 missing)
date_onset   < 16414.5   to the left, improve=0.003979285, (57 missing)
generation   < 20.5   to the left, improve=0.002407005, (0 missing)
date_hospitalisation < 16417.5   to the left, improve=0.002186176, (0 missing)
wt_kg   < 79.5   to the left, improve=0.001754285, (0 missing)
Surrogate splits:
bmi   < -502.9586 to the right, agree=0.733, adj=0.004, (1610 split)
wt_kg < -7.5   to the right, agree=0.732, adj=0.002, (0 split)
##
Node number 9: 868 observations
mean=0.09828075, MSE=0.01105777
##
Node number 12: 272 observations
mean=0.06674208, MSE=0.002354579
##
Node number 13: 132 observations,   complexity param=0.01289415
mean=0.1451049, MSE=0.006221779
left son=26 (73 obs) right son=27 (59 obs)
Primary splits:
infector   splits as ------------------------------L----------------------R
--L-----------------L-L----------L----R
--------------------------R-R-L--------------------------------------R-------------------------------------R----L
----------R---------R--------------R------L---R-----L
--------------------L----------------R---L-------L-------L----------R
-------R----------------------L-----L-----------------------------------R
---R---L
--L------R---------L-------L-R
-----------------------------------R-L------------L-----L----------------------------LR--------------------L
-------------R---L---L--R--------------------R---------L
------------L--R-------------R---L-L
------L---R
---R---------L---------------------------------------R
------------------------R-----------------------------------L---R
------------R------L---L
--------R---------------------------------R-------------R---R
---L---R------R---R
---L--------------------------R-------------------------------R-----L
------R

----------R---L--------------------------R------------R---R
---------------R---------------L---L-R----L-------------R--------------------------L---------------R------L
---L---------------R-----------------------------------R
---R
---R--------------------L------------------------R
-----------------R-------------------------------------L----------L---R-----------------------------L
---------------------L-------L---------------------------------R
---L------L-------R-----R
-------------------L------R----L----------------, improve=0.53460650, (0 missing)
days_onset_hosp < 2.5   to the right, improve=0.10922300, (22 missing)
ct_blood   < 19.5   to the left, improve=0.07316438, (0 missing)

```
##       date_onset      < 16348.5   to the right, improve=0.06042450, (22 missing)
##       source          splits as  RL, improve=0.04417949, (0 missing)
##   Surrogate splits:
##       source               splits as  RL, agree=0.621, adj=0.153, (0 split)
##       date_infection       < 16319.5   to the right, agree=0.614, adj=0.136, (0 split)
##       hospital             splits as  RRLLLR, agree=0.614, adj=0.136, (0 split)
##       date_hospitalisation < 16353     to the right, agree=0.606, adj=0.119, (0 split)
##       age_cat5             splits as  LRRLLLRLLL--LL----, agree=0.598, adj=0.102, (0 split)
##
## Node number 16: 2862 observations
##   mean=0.07737999, MSE=0.005414551
##
## Node number 17: 461 observations
##   mean=0.02002336, MSE=0.0004077007
##
## Node number 26: 73 observations
##   mean=0.09325606, MSE=0.004819559
##
## Node number 27: 59 observations
##   mean=0.2092568, MSE=0.000515053
```

#Looking at this summary,

Variable importance infector temp fever source 79 12 5 1 date_infection hospital date_hospitalisation 1 1 1

the term that predicts missing data most effectively are the infector and temp values

Looking at the correlation of missing entries

We have the dataframe shadowed_linelist which has columns added to indicate whether or not the corresponding variables are missing or not. We could look at this to see the extent to which missing values are correlated with other missing values

```
colnames(shadowed_linelist)
```

```
##  [1] "generation"             "date_infection"
##  [3] "date_onset"             "date_hospitalisation"
##  [5] "date_outcome"           "outcome"
##  [7] "gender"                 "age"
##  [9] "age_unit"               "age_years"
## [11] "age_cat"                "age_cat5"
## [13] "hospital"               "infector"
## [15] "source"                 "wt_kg"
## [17] "ht_cm"                  "ct_blood"
## [19] "fever"                  "chills"
## [21] "cough"                  "aches"
## [23] "vomit"                  "temp"
## [25] "bmi"                    "days_onset_hosp"
## [27] "generation_NA"          "date_infection_NA"
## [29] "date_onset_NA"          "date_hospitalisation_NA"
## [31] "date_outcome_NA"        "outcome_NA"
## [33] "gender_NA"              "age_NA"
## [35] "age_unit_NA"            "age_years_NA"
## [37] "age_cat_NA"             "age_cat5_NA"
## [39] "hospital_NA"            "infector_NA"
## [41] "source_NA"              "wt_kg_NA"
## [43] "ht_cm_NA"               "ct_blood_NA"
## [45] "fever_NA"               "chills_NA"
## [47] "cough_NA"               "aches_NA"
## [49] "vomit_NA"               "temp_NA"
## [51] "bmi_NA"                 "days_onset_hosp_NA"
```

In the shadowed_linelist, columns 27-52 have the counts of the missing data, lines 1-26 have the data itself

What we want to do is compute the correlation matrix of the NA count columns and show it as a heatmap

This allows us to see the extent to which missing entries are paired with other missing values

```
require("reshape2")
```

```
## Loading required package: reshape2
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
#force the na values to be 0 and the non-na to be 1

nacor=cor((shadowed_linelist[,27:52]=="!NA")*1.0)
```

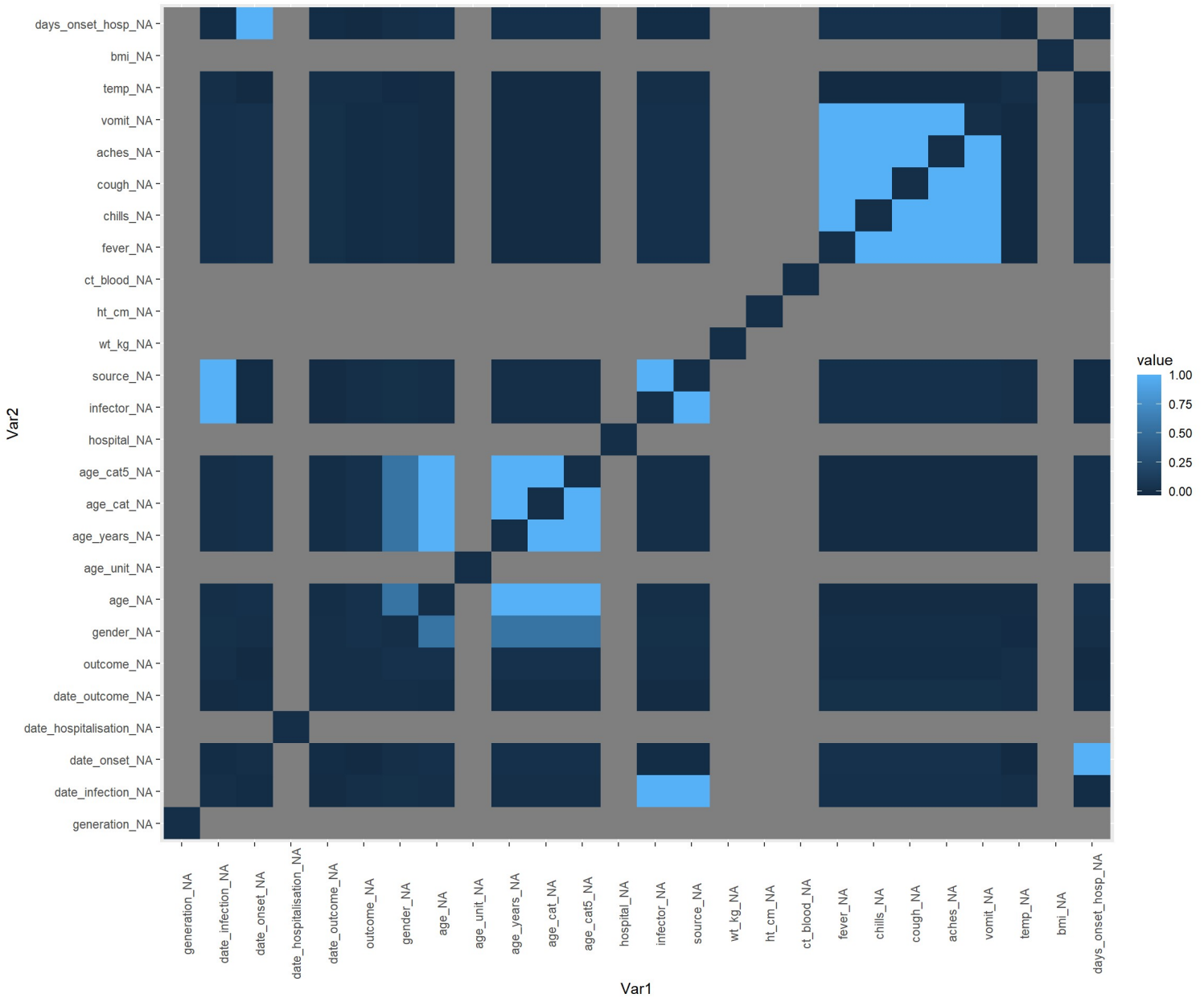
```
## Warning in cor((shadowed_linelist[, 27:52] == "!NA") * 1): the standard
## deviation is zero
```

```
# replace all the values on the diagonal with zeros
# the diagonals are all ones,  and this distorts the heatmap

diag(nacor)<-0

# melted version of the matrix of data, use this in ggplot to create a heatmap
melted_nacor=melt(nacor)

ggplot(data=melted_nacor,aes(x=Var1,y=Var2,fill=value))+geom_tile()+theme(axis.text.x=element_text(angle=90))
```



Question/Action

What are the high correlation pairs or groups in the above diagram?

date_infection with infector and source, age with it's related fields of age_unit, age_years, age_cat, and age_cat5, source with infector, and the symptoms as a group, being fever, chills, cough, comit, and temp.

Do these make sense? Are they likely to have a common cause? Why?

These definitely make sense - if you do not know the infector or source, you're unlikely to know when the person was infected, as they lack that information. Age being correlated to age related fields all make sense, source with infector being correlated is logical as well, and if the disease is likely to have a very common base of symptoms, it would make sense that patients are expected to show all symptoms, and it makes sense that patients that are missing some symptoms are missing others.