Missing Data Tutorial

Yichi Zhang

2023-03-30

Problems with Missingness

Assess Missingness

```
# import the linelist
linelist <- import("linelist_cleaned.rds")
# view the first 6 observations of the dataset
head(linelist)
```

```
##   case_id generation date_infection date_onset date_hospitalisation
## 1  5fe599          4     2014-05-08 2014-05-13           2014-05-15
## 2  8689b7          4           <NA> 2014-05-13           2014-05-14
## 3  11f8ea          2           <NA> 2014-05-16           2014-05-18
## 4  b8812a          3     2014-05-04 2014-05-18           2014-05-20
## 5  893f25          3     2014-05-18 2014-05-21           2014-05-22
## 6  be99c8          3     2014-05-03 2014-05-22           2014-05-23
##   date_outcome outcome gender age age_unit age_years age_cat age_cat5
## 1         <NA>    <NA>      m   2    years         2     0-4      0-4
## 2   2014-05-18 Recover      f   3    years         3     0-4      0-4
## 3   2014-05-30 Recover      m  56    years        56   50-69    55-59
## 4         <NA>    <NA>      f  18    years        18   15-19    15-19
## 5   2014-05-29 Recover      m   3    years         3     0-4      0-4
## 6   2014-05-24 Recover      f  16    years        16   15-19    15-19
##                              hospital       lon      lat infector source wt_kg
## 1                               Other -13.21574 8.468973   f547d6  other    27
## 2                             Missing -13.21523 8.451719     <NA>   <NA>    25
## 3 St. Mark's Maternity Hospital (SMMH) -13.21291 8.464817     <NA>   <NA>    91
## 4                        Port Hospital -13.23637 8.475476   f90f5f  other    41
## 5                    Military Hospital -13.22286 8.460824   11f8ea  other    36
## 6                        Port Hospital -13.22263 8.461831   aec8ec  other    56
##   ht_cm ct_blood fever chills cough aches vomit temp time_admission       bmi
## 1    48       22    no     no   yes    no   yes 36.8           <NA> 117.18750
## 2    59       22  <NA>   <NA>  <NA>  <NA>  <NA> 36.9          09:36  71.81844
## 3   238       21  <NA>   <NA>  <NA>  <NA>  <NA> 36.9          16:48  16.06525
## 4   135       23    no     no    no    no    no 36.8          11:22  22.49657
## 5    71       23    no     no   yes    no   yes 36.9          12:60  71.41440
## 6   116       21    no     no   yes    no   yes 37.6          14:13  41.61712
##   days_onset_hosp
## 1               2
## 2               1
## 3               2
## 4               2
## 5               1
```

```
## 6               1
```

```
# check the number of NAs in the dataset by columns
colSums(is.na(linelist))
```

```
##              case_id           generation       date_infection
##                    0                    0                 2087
##           date_onset date_hospitalisation         date_outcome
##                  256                    0                  936
##              outcome               gender                  age
##                 1323                  278                   86
##             age_unit            age_years              age_cat
##                    0                   86                   86
##             age_cat5             hospital                  lon
##                   86                    0                    0
##                  lat             infector               source
##                    0                 2088                 2088
##                wt_kg                ht_cm            ct_blood
##                    0                    0                    0
##                fever               chills                cough
##                  249                  249                  249
##                aches               vomit                 temp
##                  249                  249                  149
##       time_admission                  bmi      days_onset_hosp
##                  765                    0                  256
```

```
# the dimension of the original dataset
dim(linelist)
```

```
## [1] 5888   30
```

Functions to Remove Missing Values

```
# percent of ALL data frame values that are missing
pct_miss(linelist)
```

```
## [1] 6.688745
```

```
# percent of rows with any value missing
pct_miss_case(linelist)
```

```
## [1] 69.12364
```

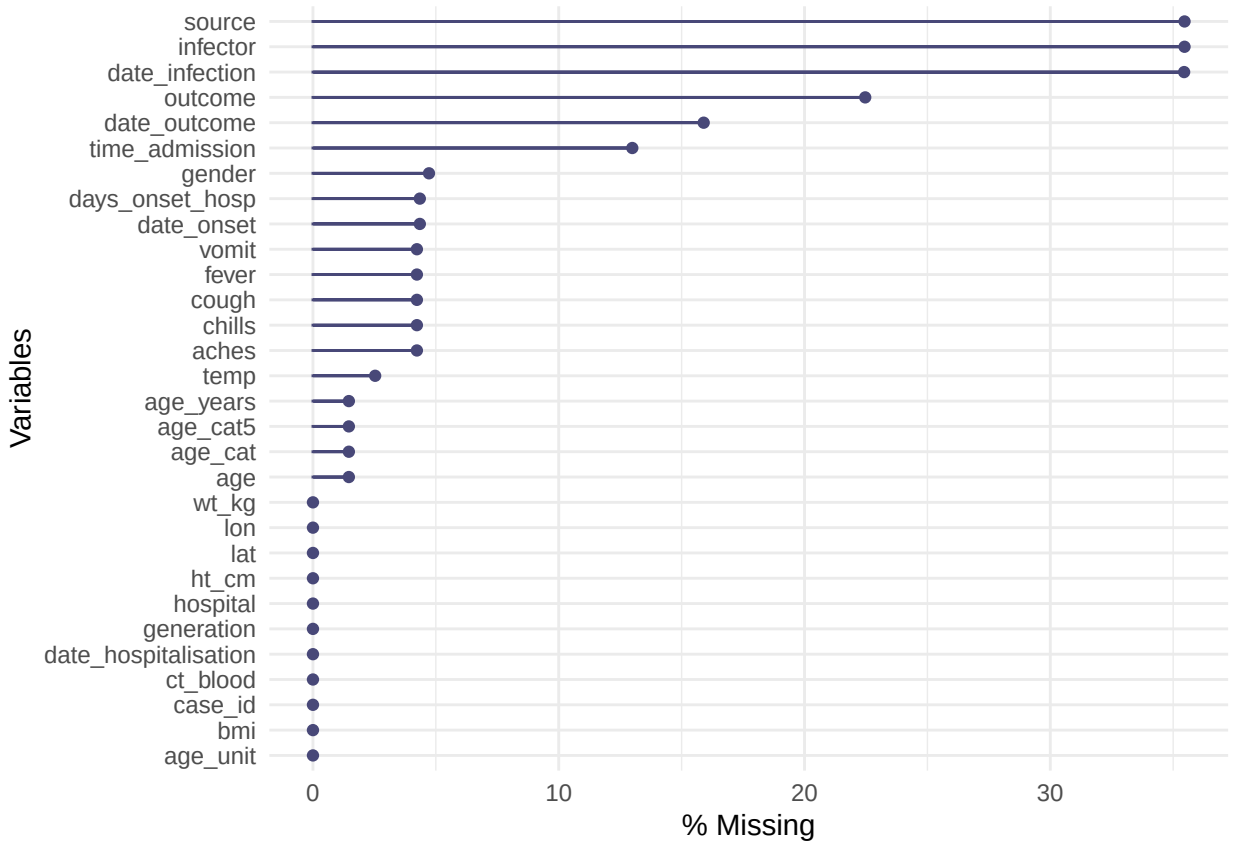
```
# percent of rows that are complete
pct_complete_case(linelist)
```

```
## [1] 30.87636
```

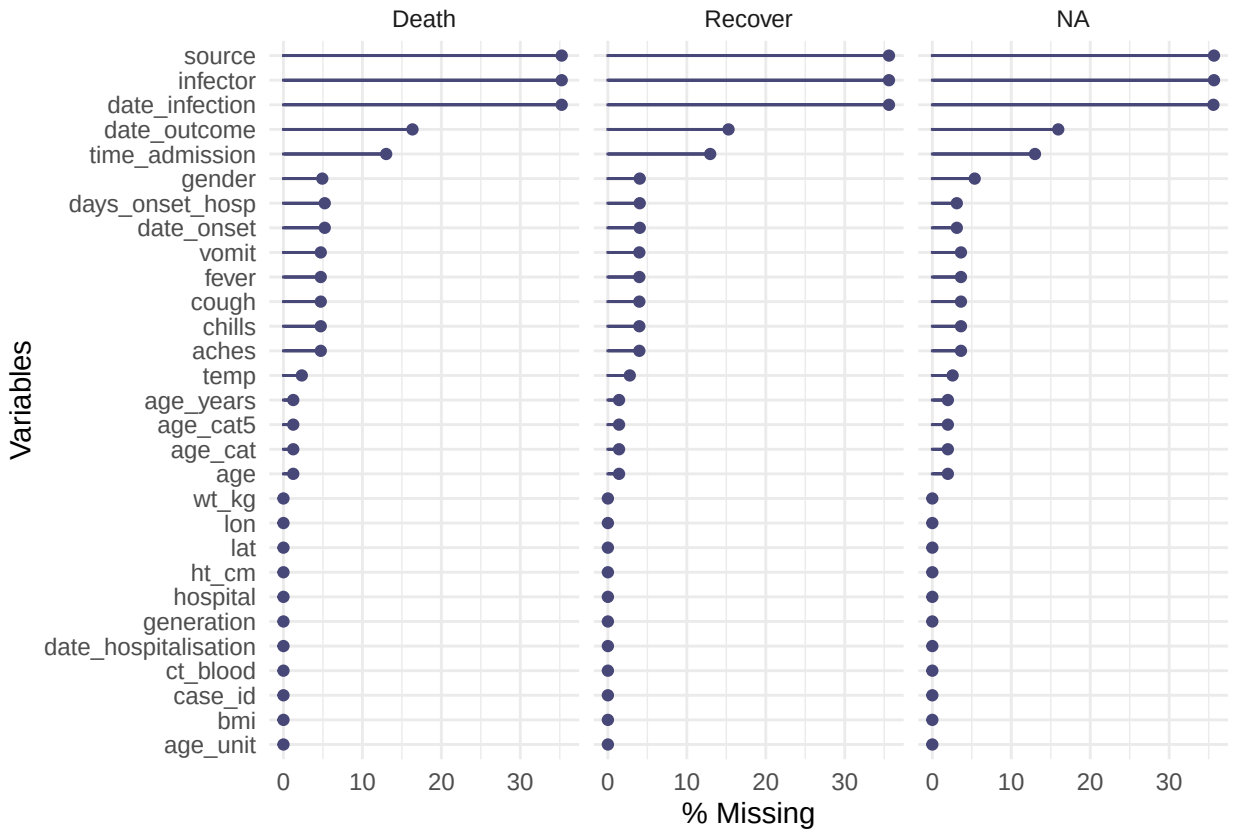
```
# complete.cases(linelist)
```

Visualize Missingness

```
## show the number of missing in each column
gg_miss_var(linelist, show_pct = TRUE)
```

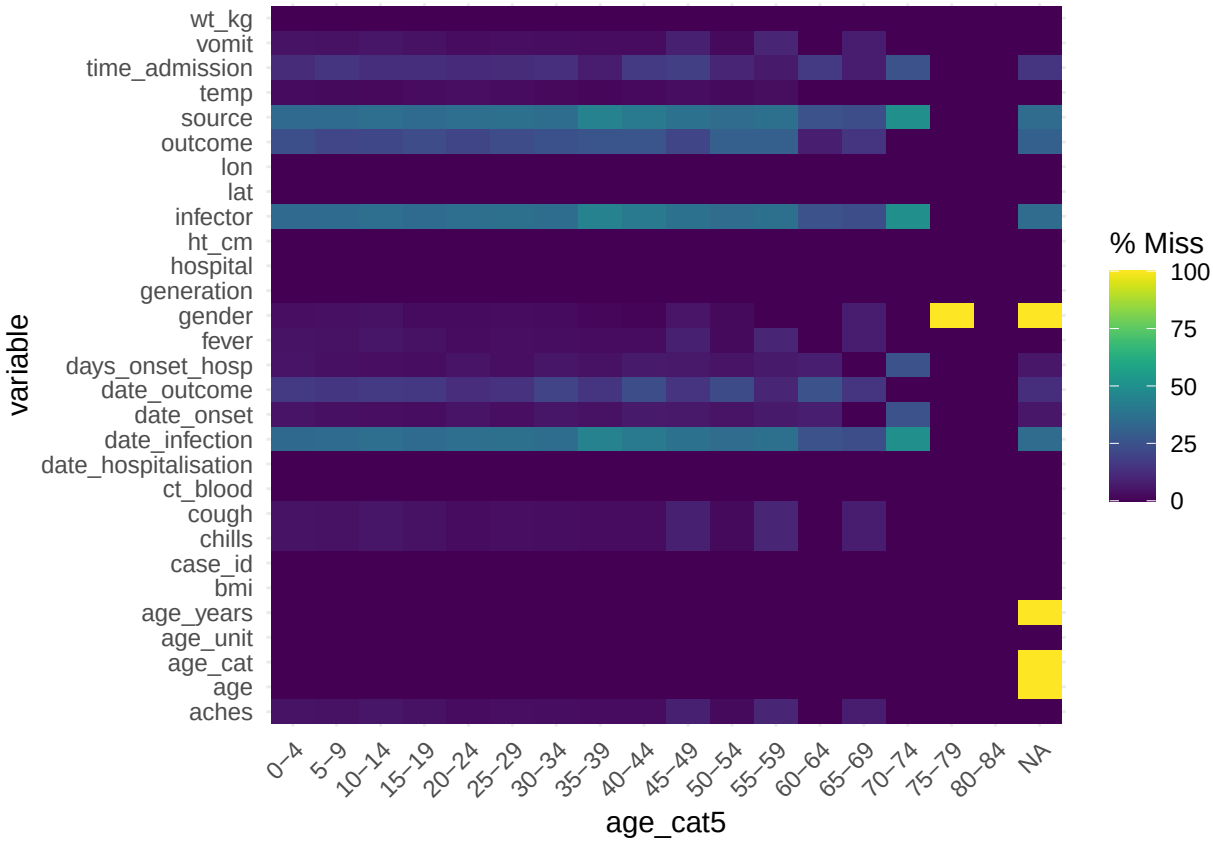



```
## split the data by a variable
linelist %>%
  gg_miss_var(show_pct = TRUE, facet = outcome)
```



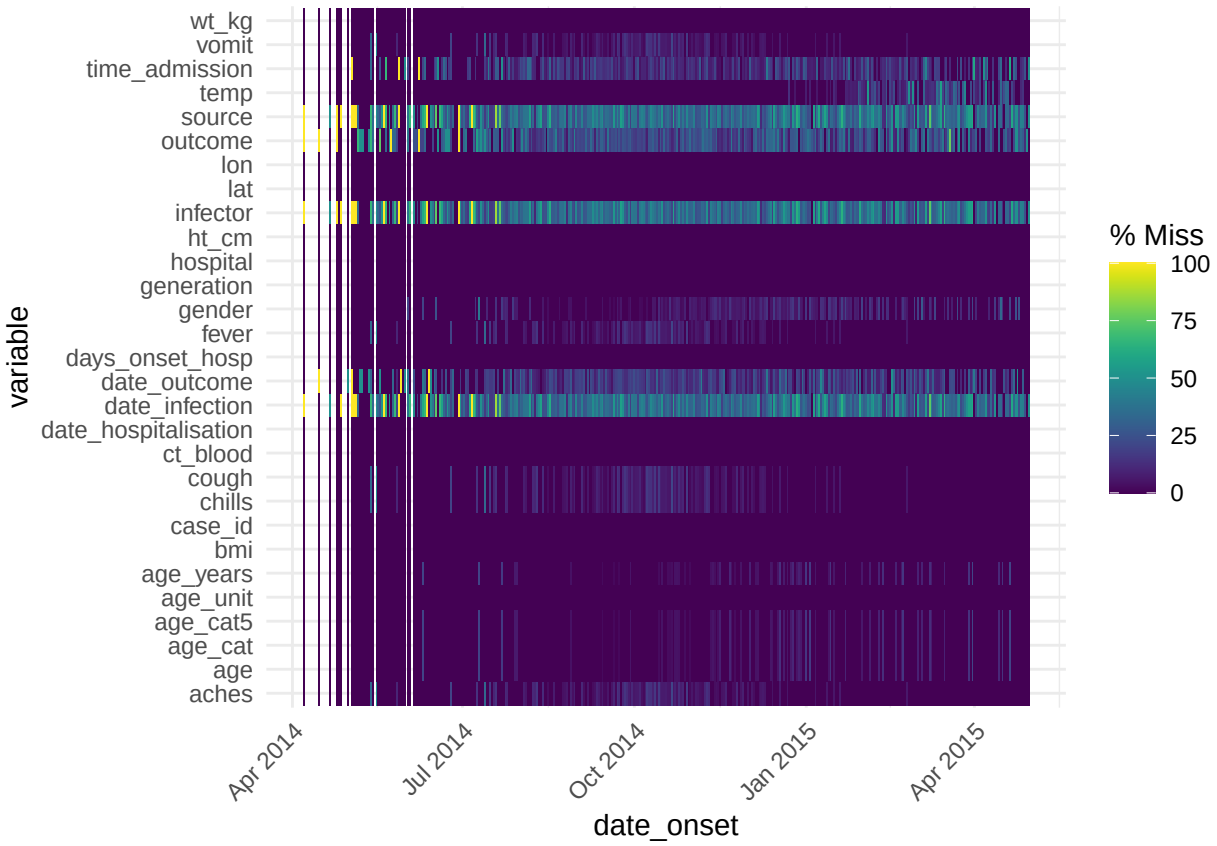

```
ggplot(
  data = linelist,
  mapping = aes(x = age_years, y = temp)) +
  geom_miss_point()
```


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




```
## change over time
gg_miss_fct(linelist, date_onset)
```

```
## Warning: Removed 29 rows containing missing values (‘geom_tile()‘).
```




```
outcome_missing <- linelist %>%
    mutate(week = lubridate::floor_date(date_onset, "week")) %>%   # create new week column
    group_by(week) %>%                                             # group the rows by week
    summarise(                                                     # summarize each week
        n_obs = n(),                                                  # number of records

        outcome_missing = sum(is.na(outcome) | outcome == ""),        # number of records missing the v
        outcome_p_miss  = outcome_missing / n_obs,                    # proportion of records missing t

        outcome_dead    = sum(outcome == "Death", na.rm=T),           # number of records as dead
        outcome_p_dead  = outcome_dead / n_obs) %>%                   # proportion of records as dead

    tidyr::pivot_longer(-week, names_to = "statistic") %>%         # pivot all columns except week, to
    filter(stringr::str_detect(statistic, "_p_"))                  # keep only the proportion values

ggplot(data = outcome_missing)+
    geom_line(
      mapping = aes(x = week, y = value, group = statistic, color = statistic),
      size = 2,
      stat = "identity")+
    labs(title = "Weekly outcomes",
         x = "Week",
         y = "Proportion of weekly records") +
     scale_color_discrete(
       name = "",
       labels = c("Died", "Missing outcome"))+
    scale_y_continuous(breaks = c(seq(0,1,0.1)))+
```

```
  theme_minimal()+
  theme(legend.position = "bottom")
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
```

```
## Warning: Removed 2 rows containing missing values ('geom_line()').
```

Address Missingness

Mean Imputation

```
linelist <- linelist %>%
    mutate(temp_replace_na_with_mean = replace_na(temp, mean(temp, na.rm = T)))
```

Regression Imputation

```
simple_temperature_model_fit <- lm(temp ~ fever + age_years, data = linelist)

#using our simple temperature model to predict values just for the observations where temp is missing
predictions_for_missing_temps <- predict(simple_temperature_model_fit,
                                         newdata = linelist %>% filter(is.na(temp)))
```

```
model_dataset <- linelist %>%
    select(temp, fever, age_years)
temp_imputed <- mice(model_dataset,
                     method = "norm.predict",
                     seed = 1,
```

```
                      m = 1,
                      print = F)
```

```
## Warning: Number of logged events: 1
```

Multiple Imputation

```
# imputing missing values for all variables in our model_dataset, and creating 10 new imputed datasets
multiple_imputation = mice(
    model_dataset,
    seed = 1,
    m = 10,
    print = FALSE)
```

```
## Warning: Number of logged events: 1
```

```
model_fit <- with(multiple_imputation, lm(temp ~ age_years + fever))
base::summary(mice::pool(model_fit))
```

```
##          term     estimate    std.error    statistic        df       p.value
## 1 (Intercept) 3.703143e+01 0.0270863456 1.367162e+03  26.83673  1.583113e-66
## 2   age_years 3.867829e-05 0.0006090202 6.350905e-02 171.44363  9.494351e-01
## 3    feveryes 1.978044e+00 0.0193587115 1.021785e+02 176.51325 5.666771e-159
```

Exercise

```
fit <- lm(Ozone ~ Wind, data = airquality)
head(na.action(fit))
```

```
##  5 10 25 26 27 32
##  5 10 25 26 27 32
```

```
naprint(na.action(fit))
```

```
## [1] "37 observations deleted due to missingness"
```

```
colSums(is.na(airquality))
```

```
##   Ozone Solar.R    Wind    Temp   Month     Day
##      37       7       0       0       0       0
```