Assignment 4

Jingwen GAO

2025-10-24

Use the lung (survival) data:

1. Generate tabular and graphical summaries of missing data by variable.

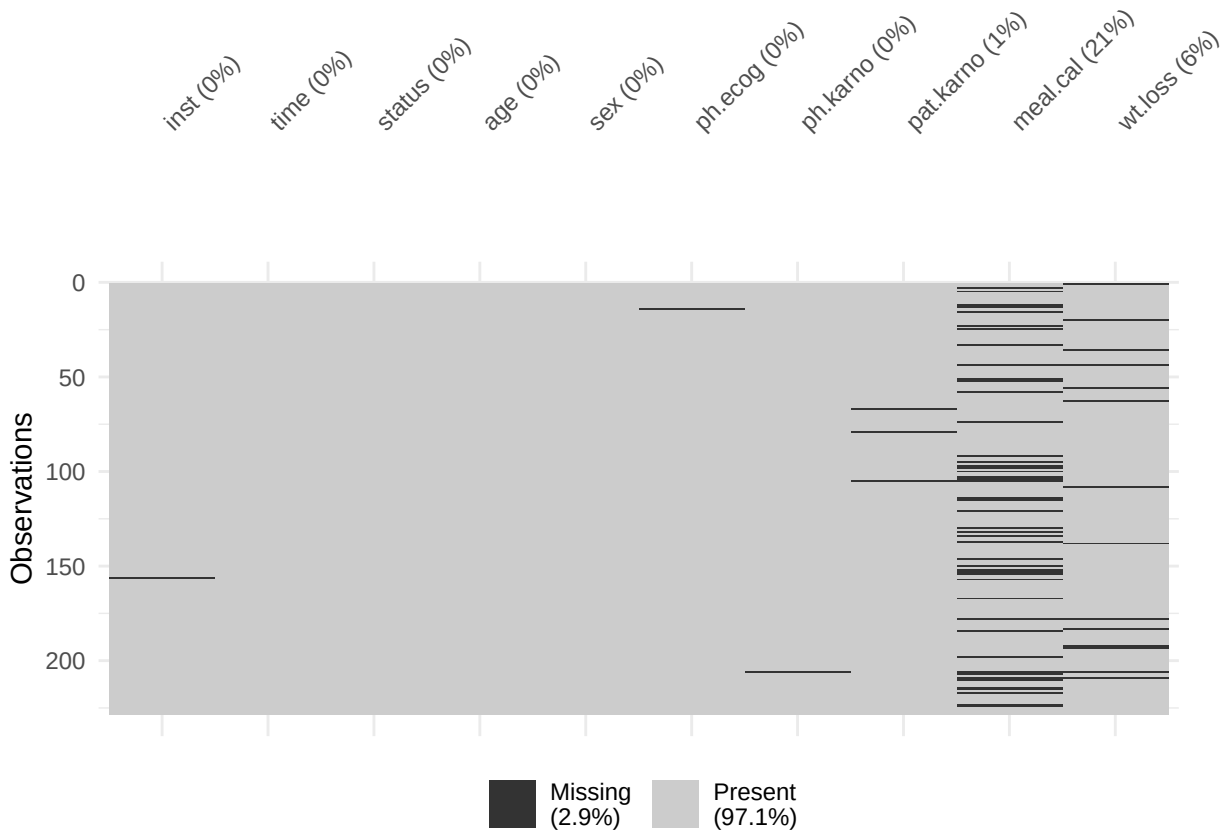
```
library(naniar)
library(finalfit)
library(survival)
miss_var_summary(lung)
```

```
## # A tibble: 10 x 3
##    variable  n_miss pct_miss
##    <chr>      <int>    <num>
##  1 meal.cal      47   20.6
##  2 wt.loss       14    6.14
##  3 pat.karno      3    1.32
##  4 inst           1    0.439
##  5 ph.ecog        1    0.439
##  6 ph.karno       1    0.439
##  7 time           0    0
##  8 status         0    0
##  9 age            0    0
## 10 sex            0    0
```

From the variable-wise tabular summary of missingness, `meal.cal` has the largest amount of missing data (47; 20.6%), followed by `wt.loss` (14; 6.14%). `pat.karno` has 3 missing values (1.32%), and `inst`, `ph.ecog`, and `ph.karno` each have 1 missing value (0.439%). `time`, `status`, `age`, and `sex` have no missing values.

To inspect which observations are missing in which variables, we use function `vis_miss()`.

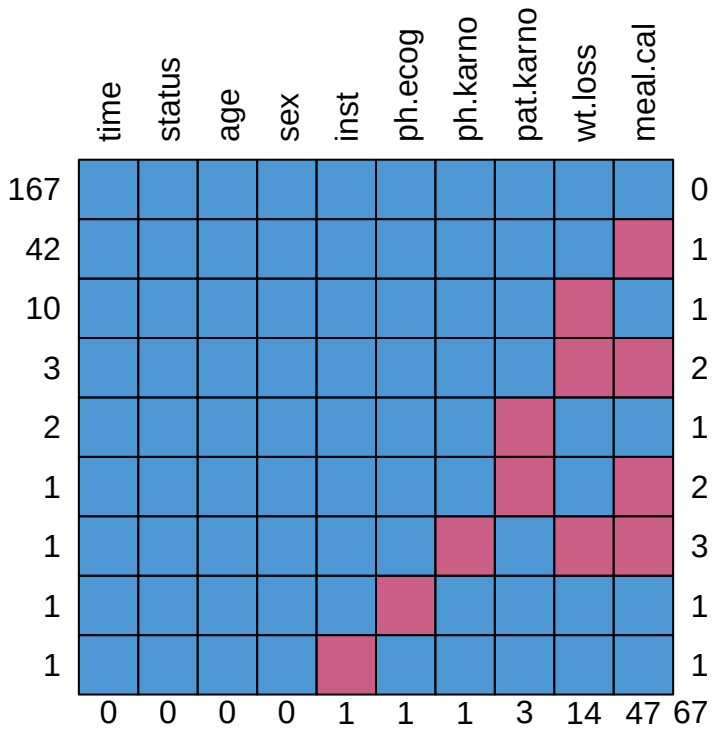
```
library(naniar)
library(ggplot2)

vis_miss(lung) + coord_cartesian(clip = "off") + theme(plot.margin = margin(t = 30,
    r = 20, b = 10, l = 10))
```



From the `vis_miss()` plot, missing values are mainly concentrated in the variables `meal.cal` and `wt.loss`, with only a few isolated missing points in `pat.karno`, `ph.karno`, `ph.ecog`, and `inst`. The overall missingness rate is low (2.9%), and the missing observations appear scattered rather than clustered by cases, suggesting that missingness may occur at random rather than systematically.

To inspect the joint missingness between variables, we use `mice::md.pattern()`.

```
library(mice)
md.pattern(lung, rotate.names = T)
```




```
##     time status age sex inst ph.ecog ph.karno pat.karno wt.loss meal.cal
## 167    1      1   1   1    1       1        1         1       1        1  0
## 42     1      1   1   1    1       1        1         1       1        0  1
## 10     1      1   1   1    1       1        1         1       0        1  1
## 3      1      1   1   1    1       1        1         1       0        0  2
## 2      1      1   1   1    1       1        1         0       1        1  1
## 1      1      1   1   1    1       1        1         0       1        0  2
## 1      1      1   1   1    1       1        0         1       0        0  3
## 1      1      1   1   1    1       0        1         1       1        1  1
## 1      1      1   1   1    0       1        1         1       1        1  1
##        0      0   0   0    1       1        1         3      14       47 67
```

There are 9 patterns of missing data. The plot shows that most missingness occurs independently in single variables. Out of 61 observations with missing values, 56 of them only have one missing value. Only a few records have simultaneous missingness across variables. The majority of the dataset (167 observations) have no missing values. This indicates a relatively simple missing data structure, with no strong multivariate missingness pattern, suggesting that the missingness is likely random rather than systematic.

2. Do a Little MCAR test and interpret with regard to the assumption of MCAR.

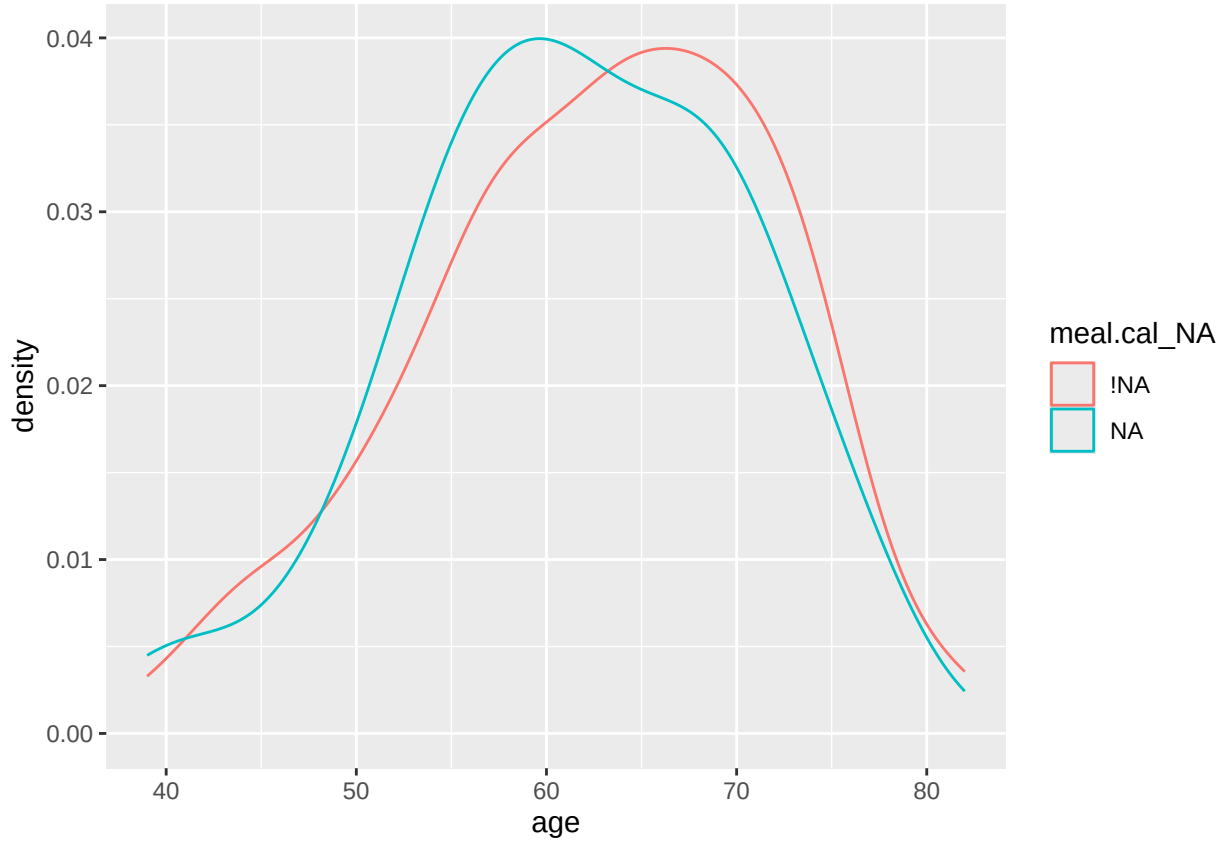
```
mcar_test(lung)
```

```
## # A tibble: 1 x 4
##   statistic    df p.value missing.patterns
##       <dbl> <dbl>   <dbl>            <int>
## 1      62.0    68   0.683                9
```

The null hypothesis of Little's MCAR test states that the missing data are Missing Completely at Random. The test result shows a chi-square statistic of 61.98 with 68 degrees of freedom and a p-value of 0.683. Since the p-value is much greater than common significance levels, we fail to reject the null hypothesis. Therefore, there is no evidence against the MCAR assumption, but to certify this missingness pattern, more investigation is needed.

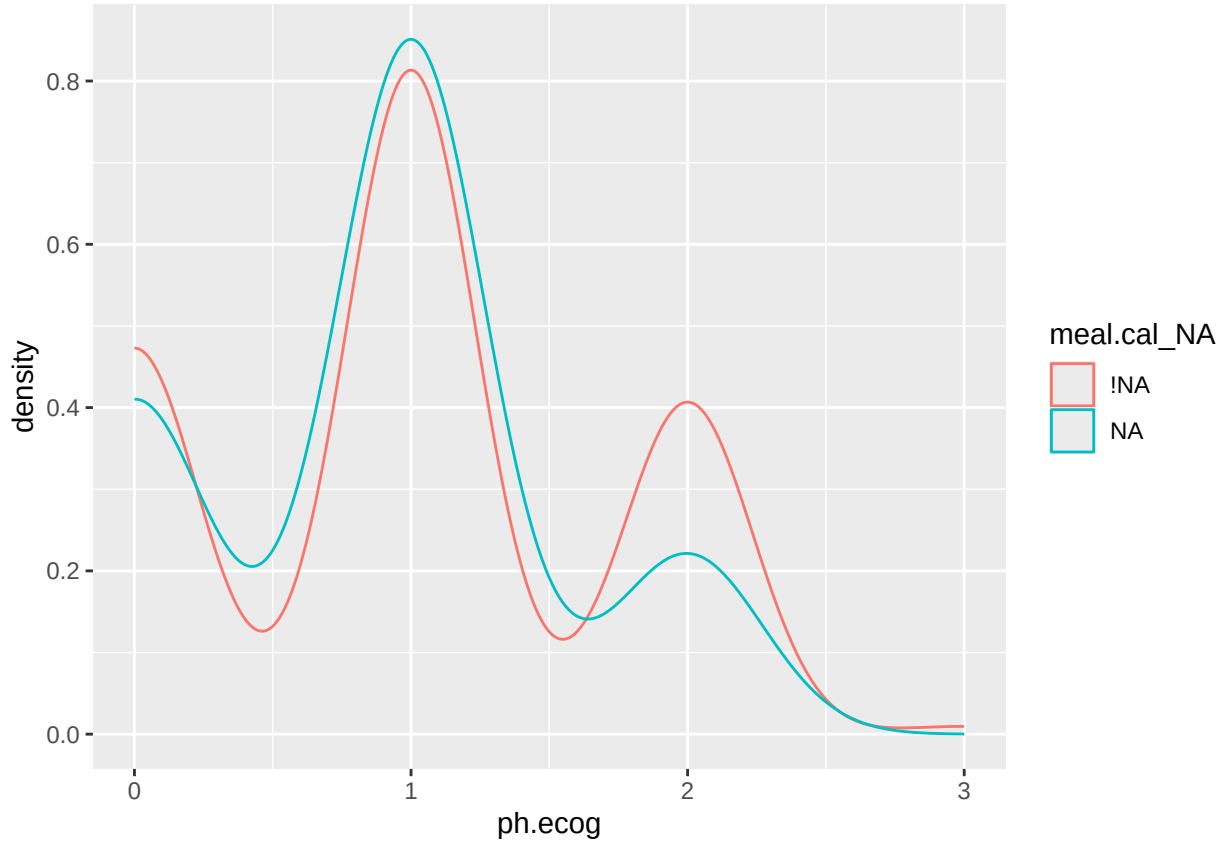
3. We'll focus on meal calories, because of the magnitude of missing data on that variable, and try to distinguish between MAR and MNAR models: Are the distributions of age, ph.ecog, ph.karno (pat.karno is highly correlated with ph.karno so we can omit that), and wt.loss dependent on missing data in meal.cal? Demonstrate with a series of density plots of each by a nabular meal.cal variable. Summarize in a sentence or 2. Do the findings argue for MAR or MNAR?

```
library(ggplot2)
lung %>%
    nabular() %>%
    ggplot(aes(x = age, color = meal.cal_NA)) + geom_density()
```



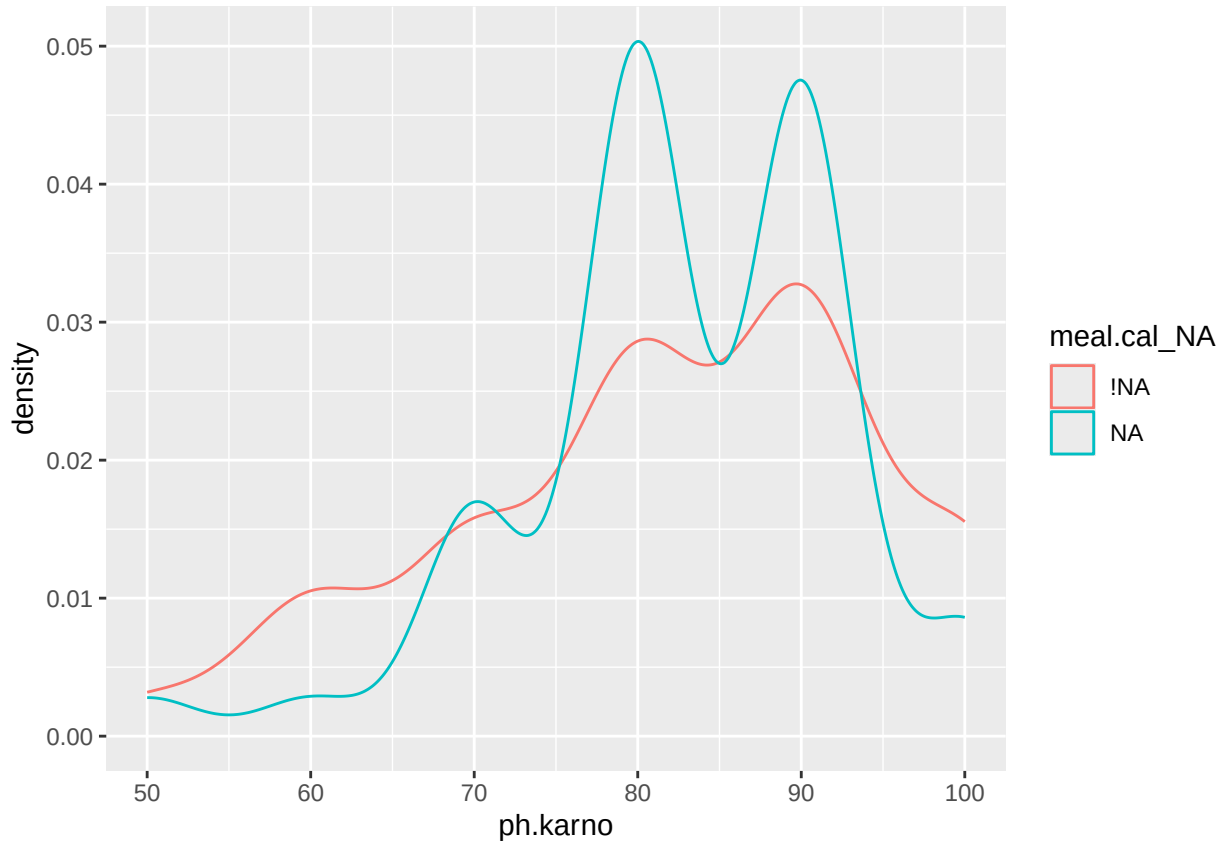
The two curves are largely overlapping, with only minor differences in shape and location. This suggests that `age` is not strongly associated with whether `meal.cal` is missing.

```
lung %>%
    nabular() %>%
    ggplot(aes(x = ph.ecog, color = meal.cal_NA)) + geom_density()
```



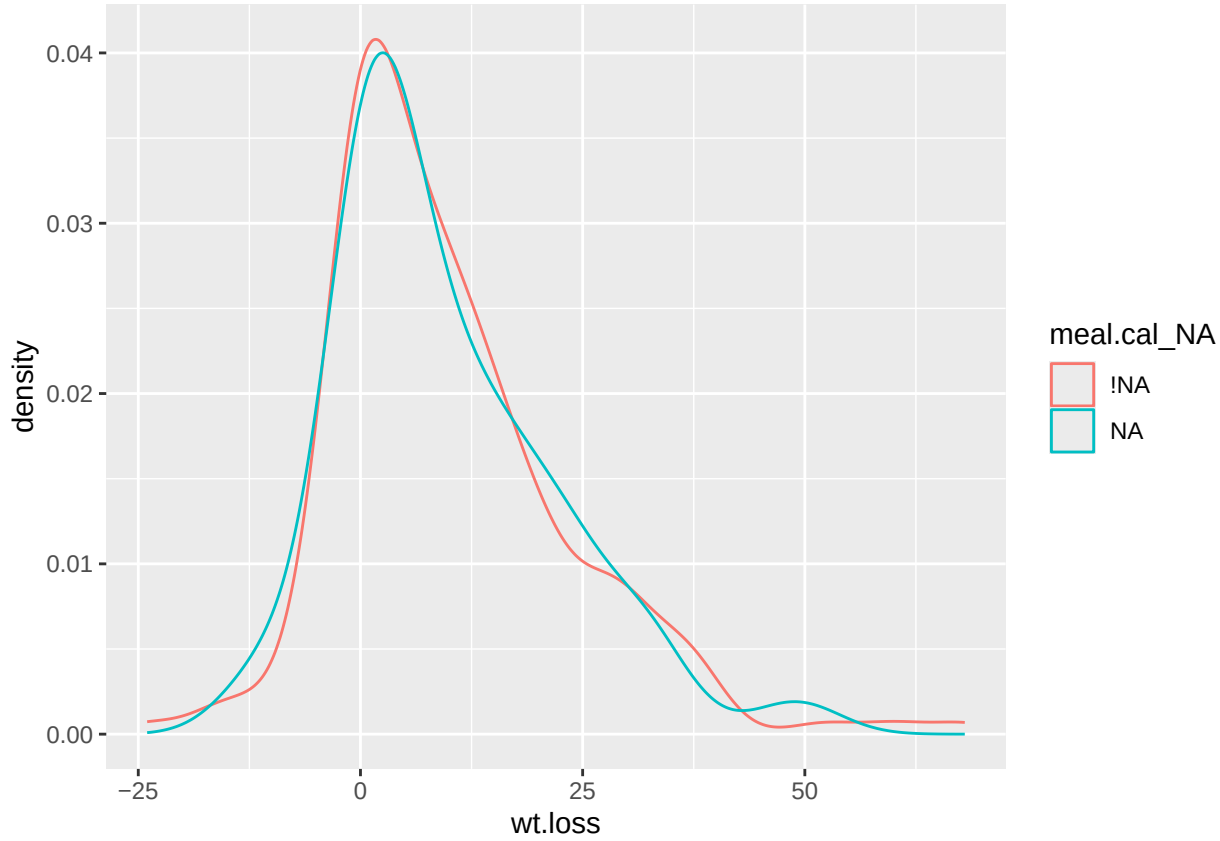
The two curves are similar overall, though the group with missing `meal.cal` appears to have slightly higher `ph.ecog` scores. This mild difference suggests that missingness in `meal.cal` may be related to `ph.ecog`, implying that the data are more consistent with a MAR mechanism rather than MNAR.

```
lung %>%
    nabular() %>%
    ggplot(aes(x = ph.karno, color = meal.cal_NA)) + geom_density()
```



The two curves are significantly different. Missing `meal.cal` is associated with a moderate high `ph.karno`(75-95). This suggests that missingness in `meal.cal` may be related to `ph.karno`, implying that the data are more consistent with a MAR mechanism rather than MNAR.

```
lung %>%
    nabular() %>%
    ggplot(aes(x = wt.loss, color = meal.cal_NA)) + geom_density()
```



The two curves are largely overlapping. This suggests that `wt.loss` is not strongly associated with whether `meal.cal` is missing.

These results suggest that missingness in `meal.cal` is modestly related to `ph.ecog` and `ph.karno`, but not to `age` or `wt.loss`. Therefore, the missing data pattern is more consistent with a Missing at Random (MAR) mechanism rather than Missing Not at Random (MNAR).

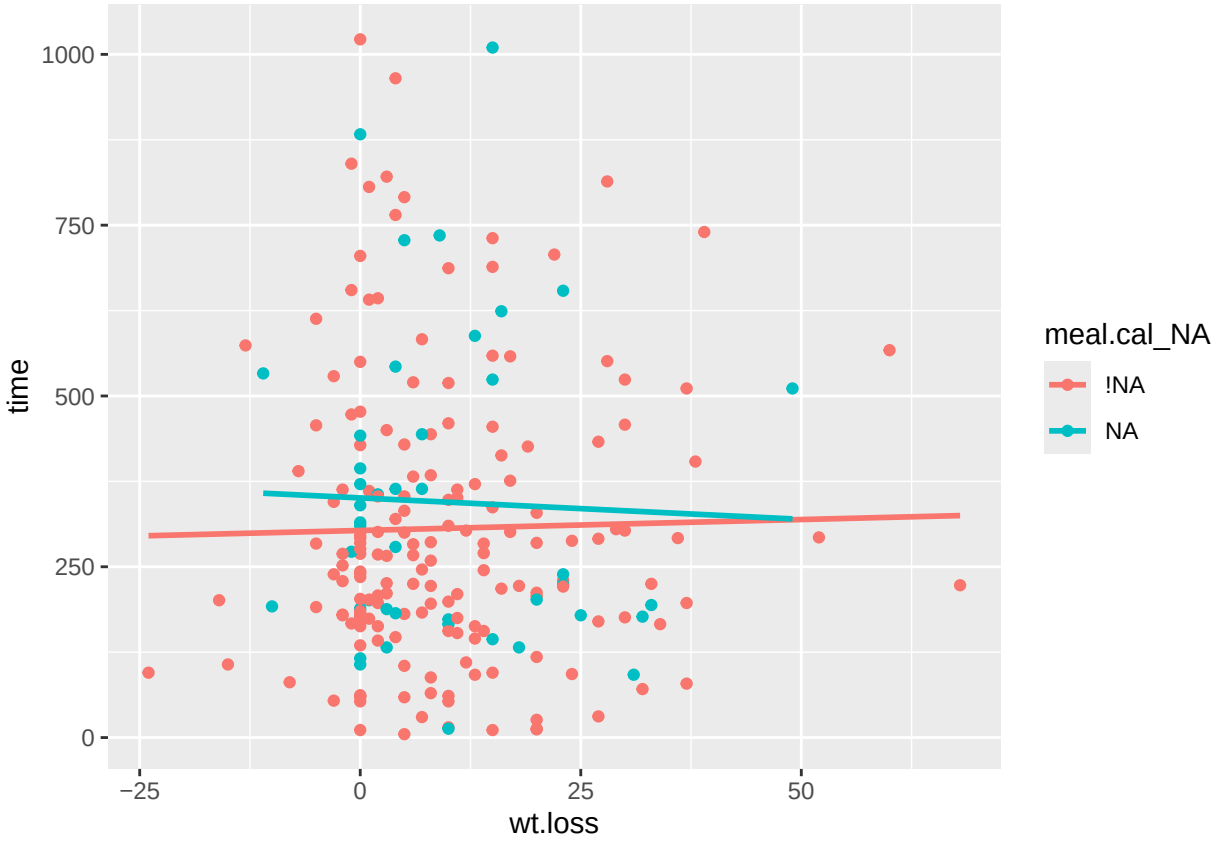
4. Explain a) how a measure of meal calories might suppress its own values, and b) which missing data pattern this would be consistent with.

If meals with very high (or very low) true calories make respondents uncomfortable to reveal, they may intentionally leave the item blank. Because the probability of missingness then depends on the unobserved true value of meal.cal, this mechanism is consistent with MNAR.

By contrast, if the item is missed simply because calorie calculation is complicated or respondents forget to record, independently of the true calorie amount, then the mechanism would be MCAR.

5. If researchers are interested in the relationship between survival time (y) and weight loss (x), show that the relationship is not dependent on missingness in meal calories.

```
lung %>%
    nabular() %>%
    ggplot(aes(x = wt.loss, y = time, color = meal.cal_NA)) + geom_point() + geom_smooth(method = "lm",
    se = F)
```



The scatter plot shows the relationship between survival time and weight loss, separated by whether meal.cal is missing (meal.cal_NA). The fitted regression lines for the two groups are nearly parallel and close to each other, indicating that the slope and general trend of time versus wt.loss are very similar regardless of meal.cal missingness.This visual pattern suggests that the association between survival time and weight loss does not depend on whether meal.cal is missing.

```
summary(lm(time ~ wt.loss * meal.cal_NA, data = nabular(lung)))
```

```
##
## Call:
## lm(formula = time ~ wt.loss * meal.cal_NA, data = nabular(lung))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -331.53 -147.09  -39.45  102.95  718.87
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           303.1346    20.3435  14.901   <2e-16 ***
## wt.loss                 0.3203     1.2253   0.261    0.794
## meal.cal_NANA          47.6602    46.1751   1.032    0.303
## wt.loss:meal.cal_NANA  -0.9470     2.9425  -0.322    0.748
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 213.6 on 210 degrees of freedom
```

```
##   (14 observations deleted due to missingness)
## Multiple R-squared:  0.005863,   Adjusted R-squared:  -0.008339
## F-statistic: 0.4128 on 3 and 210 DF,  p-value: 0.744
```

Fitting `time ~ wt.loss * meal.cal_NA` shows a non-significant interaction ($\beta = -0.95, p = 0.75$), indicating that the slope of survival time on weight loss is the same for records with and without missing meal.cal. Thus, the `time~wt.loss` relationship does not depend on missingness of `meal.cal_NA`.