Course 2 Section 1.15 - Dealing with missing data

Jiaying Wu

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
#BiocManager::install(version = "3.11")
BiocManager::install("impute", version = "3.10")
```

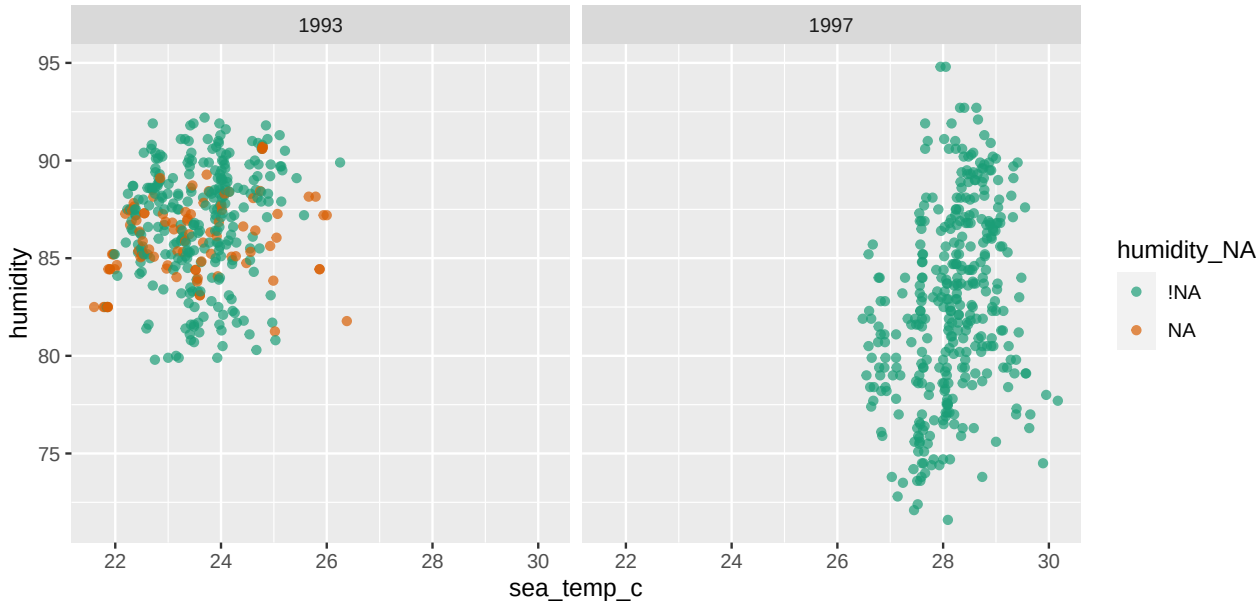
```
library(tidyverse)
library(impute)
library(visdat)
library(naniar)
```

In RStudio, change the code to plot sea temperature against humidity, with colour representing missing humidity values. What do you learn about the imputations?

```
tao_impute <- bind_shadow(oceanbuoys) %>%
  arrange(year, sea_temp_c, humidity) %>%
  select(sea_temp_c, humidity)
tao_impute <- impute.knn(as.matrix(tao_impute), 5)
```

```
## Warning in knnimp(x, k, maxmiss = rowmax, maxp = maxp): 2 rows with more than 50 % entries missing;
##  mean imputation used for these rows
```

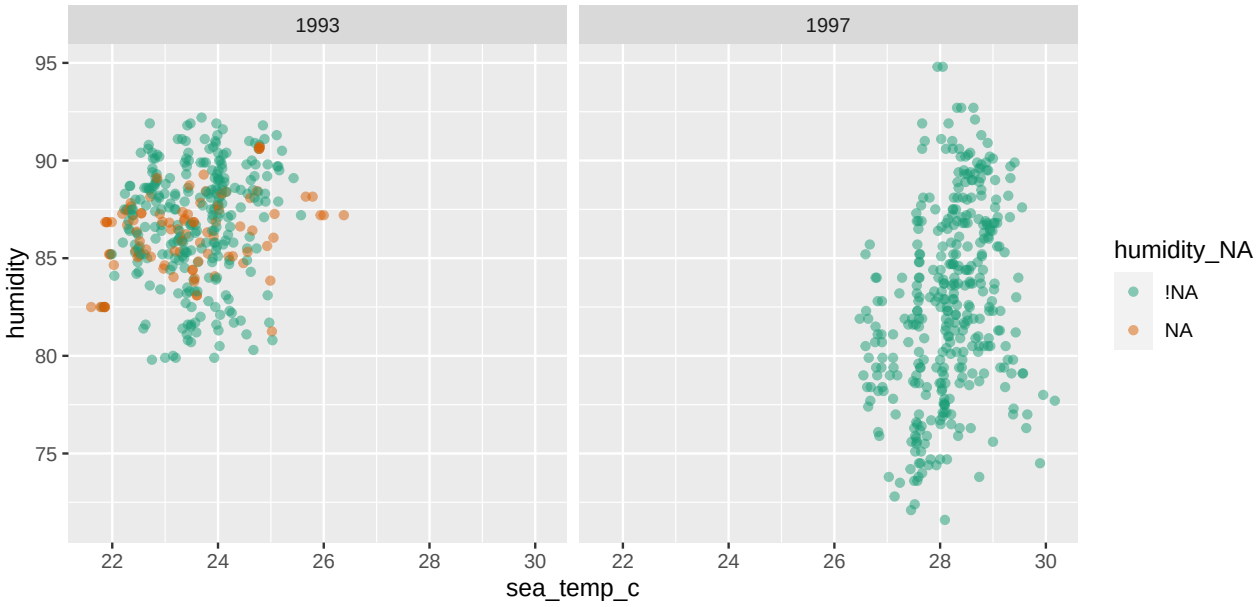
```
tao_shadow <- bind_shadow(oceanbuoys) %>%
  arrange(year, sea_temp_c, humidity) %>%
  mutate(sea_temp_c = tao_impute$data[,1],
         humidity = tao_impute$data[,2])
ggplot(tao_shadow,
       aes(x = sea_temp_c,
           y = humidity,
           colour=humidity_NA)) +
  geom_point(alpha=0.7) +
  facet_wrap(~year) +
  scale_colour_brewer(palette="Dark2") +
  theme(aspect.ratio=1)
```




```
tao_impute_93 <- bind_shadow(oceanbuoys) %>%
  arrange(year, sea_temp_c, humidity) %>%
  filter(year=="1993") %>%
  select(sea_temp_c, humidity)
tao_impute_93 <- impute.knn(as.matrix(tao_impute_93), 5)
```

```
## Warning in knnimp(x, k, maxmiss = rowmax, maxp = maxp): 2 rows with more than 50 % entries missing;
##  mean imputation used for these rows
```

```
tao_impute_97 <- bind_shadow(oceanbuoys) %>%
  arrange(year, sea_temp_c, humidity) %>%
  filter(year=="1997") %>%
  select(sea_temp_c, humidity)
tao_impute_97 <- impute.knn(as.matrix(tao_impute_97), 5)
tao_impute <- rbind(tao_impute_93$data, tao_impute_97$data)
tao_shadow <- bind_shadow(oceanbuoys) %>%
  arrange(year, sea_temp_c, humidity) %>%
  mutate(sea_temp_c = tao_impute[,1],
         humidity = tao_impute[,2])
ggplot(tao_shadow,
       aes(x = sea_temp_c,
           y = humidity,
           colour= humidity_NA)) +
  geom_point(alpha=0.5) +
  facet_wrap(~year) +
  scale_colour_brewer(palette="Dark2") +
  theme(aspect.ratio=1)
```


1993
1997
humidity
95
90
85
80
75
22
24
26
28
30
sea_temp_c
humidity_NA
!NA
NA