Data Management Module

Hannah Chan-Henry

2026-02-18

```
library(tidyverse)
library(readxl)
```

The data set I am using is simulated data based on a real data set of tree forest data. The range of the values are accurate to the real data but the values are randomly generated within that range. This code should be mostly transferable when I work with the real data set.

```
tree_data<-read_xlsx("data/Generated tree data.xlsx", sheet= "Tree data")
head(tree_data)
```

```
## # A tibble: 6 x 11
##    Plot Species  Year  DBH1  DBH2  DBH3  DBH4 COND1 COND2 COND3 COND4
##   <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     1      12  1985  42.8  53.0  75.2  63.4    77    86    48    56
## 2     2      26  1985  35.7  35.1  84.5  74.7    13    66     4    87
## 3     3      10  1985  63.2  94.0  40.0  58.8     8    86    75     1
## 4     4      15  1985  94.4  65.4  95.2  43.9     9    89    61    60
## 5     5      28  1985  64.2  57.2  73.9  43.4    50    88    10    82
## 6     6      10  1985  57.4  19.6  73.8  30.1    33     0    80    22
```

```
species_code_data<-data.frame(read_xlsx("data/Generated tree data.xlsx", sheet= "Species code"))
head(species_code_data)
```

```
##   Code.Number         Scientific.Name
## 1           1         Cyathea capensis
## 2           2     Podocarpus falcatus
## 3           3   Podocarpus latifolius
## 4           4 Widdringtonia nodiflora
## 5           5          Strelitzia alba
## 6           6          Myrica serrata
```

```
description_data<-data.frame(read_xlsx("data/Generated tree data.xlsx", sheet= "Data description"))
description_data
```

```
##    Variable                                      Description
## 1      <NA>                                             <NA>
## 2      YEAR                        Plot establishment year
## 3   SPECIES                 Species number code: see Tree list
## 4      DBH1              DBH at first measurement (1985) in cm
## 5      DBH2            DBH at second measurement (1991) in cm
## 6      DBH3             DBH at third measurement (2001) in cm
## 7      DBH4            DBH at fourth measurement (2021) in cm
## 8     COND1  Tree condition code first measurement - see list
## 9     COND2 Tree condition code second measurement - see list
## 10    COND3  Tree condition code third measurement - see list
## 11    COND4 Tree condition code fourth measurement - see list
```

```
# some of the data is tidy but I need to further tidy the data.
tidy_tree_data1<-tree_data %>% select(-COND1, -COND2, -COND3, -COND4) %>%
  pivot_longer(cols = c("DBH1", "DBH2", "DBH3", "DBH4"),
               names_to = "Survey year", values_to = "DBH")

tidy_tree_data2<-tree_data %>% select(COND1, COND2, COND3, COND4) %>%
  pivot_longer(cols = c("COND1", "COND2", "COND3", "COND4"),
               names_to = "COND year", values_to = "COND")

tidy_tree_data<-cbind(tidy_tree_data1,tidy_tree_data2)

# change the codes like DBH1 and COND1 to the years which they represent

tidy_tree_data$`Survey year`[tidy_tree_data$`Survey year` == "DBH1"] <- "1985"
tidy_tree_data$`Survey year`[tidy_tree_data$`Survey year` == "DBH2"] <- "1991"
tidy_tree_data$`Survey year`[tidy_tree_data$`Survey year` == "DBH3"] <- "2001"
tidy_tree_data$`Survey year`[tidy_tree_data$`Survey year` == "DBH4"] <- "2021"

tidy_tree_data<-tidy_tree_data%>%select(-`COND year`)

#remove NA from data
tidy_tree_data <- na.omit(tidy_tree_data)
head(tidy_tree_data)
```

```
##   Plot Species Year Survey year      DBH COND
## 1    1      12 1985        1985 42.77341   77
## 2    1      12 1985        1991 52.96614   86
## 3    1      12 1985        2001 75.20837   48
## 4    1      12 1985        2021 63.37060   56
## 5    2      26 1985        1985 35.72565   13
## 6    2      26 1985        1991 35.12405   66
```

Now I need to add to my meta data as I have changed the column names. I also need to remove the variables that are no longer being used.

```
new_row <- data.frame(Variable = c("DBH","COND","Survey year"), Description = c("Diameter at Breast Hei
description_new<-rbind(description_data,new_row)
description_new<-description_new[-c(1,4,5,6,7,8,9,10,11),]
description_new
```

```
##       Variable                        Description
## 2         YEAR            Plot establishment year
## 3      SPECIES Species number code: see Tree list
## 12         DBH          Diameter at Breast Height
## 13        COND            Condition code measured
## 14 Survey year Year which measurements were taken
```

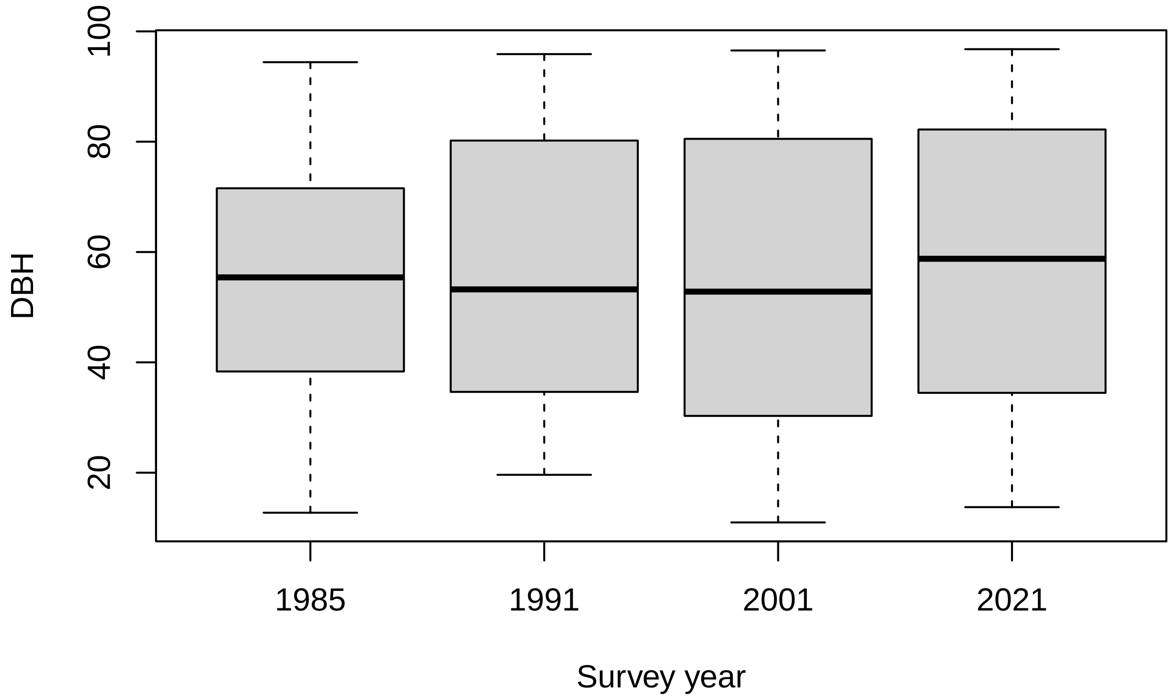
I need to remove any species that are in the metadata but are not in the species data.

```
allow<-unique(tidy_tree_data$Species)
filtered_species_names <- species_code_data[species_code_data$Code.Number %in% allow, ]
# the number of species that were removed from the code name list
#nrow(species_code_data)-nrow(filtered_species_names)
```

3 species were removed from the species code list.

Now I need to create a summary the data. We should see very little pattern in the data as the values are random. This is just an example of what could be done with the larger, real data set.

```
boxplot(tidy_tree_data$DBH~tidy_tree_data$`Survey year`, xlab = "Survey year", ylab="DBH")
```


DBH
Survey year