Data Processing Assignment 2

Riya Minesh Amin

9/19/2019

##Setup

Install and load the necessary packages to reproduce the report here:

```
library(readr)
library(readxl)
library(tidyr)
library(dplyr)
library(kableExtra)
library(knitr)
library(Hmisc)
library(outliers)
```

Read WHO Data

Read the WHO data using an appropriate function.

```
WHO<- read_excel("D:/WHO.xlsx")
```

Tidy Task 1:

```
WHO1 <- WHO %>% gather(code, value, 5:60)
WHO1
```

```
## # A tibble: 405,440 x 6
##    country     iso2  iso3   year code        value
##    <chr>       <chr> <chr> <dbl> <chr>       <chr>
##  1 Afghanistan AF    AFG    1980 new_sp_m014 NA
##  2 Afghanistan AF    AFG    1981 new_sp_m014 NA
##  3 Afghanistan AF    AFG    1982 new_sp_m014 NA
##  4 Afghanistan AF    AFG    1983 new_sp_m014 NA
##  5 Afghanistan AF    AFG    1984 new_sp_m014 NA
##  6 Afghanistan AF    AFG    1985 new_sp_m014 NA
##  7 Afghanistan AF    AFG    1986 new_sp_m014 NA
##  8 Afghanistan AF    AFG    1987 new_sp_m014 NA
##  9 Afghanistan AF    AFG    1988 new_sp_m014 NA
## 10 Afghanistan AF    AFG    1989 new_sp_m014 NA
## # ... with 405,430 more rows
```

Tidy Task 2:

```
WHO2 <- WHO1 %>% separate(code, into = c("new", "var", "sex"),sep = "_")
WHO2
```

```
## # A tibble: 405,440 x 8
##    country     iso2  iso3   year new   var   sex   value
##    <chr>       <chr> <chr> <dbl> <chr> <chr> <chr> <chr>
##  1 Afghanistan AF    AFG    1980 new   sp    m014  NA
##  2 Afghanistan AF    AFG    1981 new   sp    m014  NA
##  3 Afghanistan AF    AFG    1982 new   sp    m014  NA
##  4 Afghanistan AF    AFG    1983 new   sp    m014  NA
##  5 Afghanistan AF    AFG    1984 new   sp    m014  NA
##  6 Afghanistan AF    AFG    1985 new   sp    m014  NA
##  7 Afghanistan AF    AFG    1986 new   sp    m014  NA
##  8 Afghanistan AF    AFG    1987 new   sp    m014  NA
##  9 Afghanistan AF    AFG    1988 new   sp    m014  NA
## 10 Afghanistan AF    AFG    1989 new   sp    m014  NA
## # ... with 405,430 more rows
```

```
WHO3 <- WHO2 %>% separate(sex, into = c("sex","age"), sep = 1)
WHO3
```

```
## # A tibble: 405,440 x 9
##    country     iso2  iso3   year new   var   sex   age   value
##    <chr>       <chr> <chr> <dbl> <chr> <chr> <chr> <chr> <chr>
##  1 Afghanistan AF    AFG    1980 new   sp    m     014   NA
##  2 Afghanistan AF    AFG    1981 new   sp    m     014   NA
##  3 Afghanistan AF    AFG    1982 new   sp    m     014   NA
##  4 Afghanistan AF    AFG    1983 new   sp    m     014   NA
##  5 Afghanistan AF    AFG    1984 new   sp    m     014   NA
##  6 Afghanistan AF    AFG    1985 new   sp    m     014   NA
##  7 Afghanistan AF    AFG    1986 new   sp    m     014   NA
##  8 Afghanistan AF    AFG    1987 new   sp    m     014   NA
##  9 Afghanistan AF    AFG    1988 new   sp    m     014   NA
## 10 Afghanistan AF    AFG    1989 new   sp    m     014   NA
## # ... with 405,430 more rows
```

Tidy Task 3:

```
WHO4 <- WHO2 %>% spread(var, value)
WHO4
```

```
## # A tibble: 101,360 x 10
##    country     iso2  iso3   year new   sex   ep    rel   sn    sp
##    <chr>       <chr> <chr> <dbl> <chr> <chr> <chr> <chr> <chr> <chr>
##  1 Afghanistan AF    AFG    1980 new   m014  NA    NA    NA    NA
##  2 Afghanistan AF    AFG    1981 new   m014  NA    NA    NA    NA
##  3 Afghanistan AF    AFG    1982 new   m014  NA    NA    NA    NA
##  4 Afghanistan AF    AFG    1983 new   m014  NA    NA    NA    NA
##  5 Afghanistan AF    AFG    1984 new   m014  NA    NA    NA    NA
##  6 Afghanistan AF    AFG    1985 new   m014  NA    NA    NA    NA
##  7 Afghanistan AF    AFG    1986 new   m014  NA    NA    NA    NA
##  8 Afghanistan AF    AFG    1987 new   m014  NA    NA    NA    NA
##  9 Afghanistan AF    AFG    1988 new   m014  NA    NA    NA    NA
## 10 Afghanistan AF    AFG    1989 new   m014  NA    NA    NA    NA
## # ... with 101,350 more rows
```

Tidy Task 4:

```
WHO5 <- WHO3 %>% mutate(age = factor(age, levels=c("014","1524","2534","3544","4554","5564","65"
), labels=c("<15","15-24","25-34","35-44","45-54","55-64","65>="), ordered=TRUE))
WHO6 <- WHO5 %>% mutate(sex = factor(sex))
```

Task 5: Filter & Select

```
WHO_subset<- WHO2 %>% filter(country %in% c("Afghanistan", "Albania", "Algeria")) %>% select(-(i
so2),-(new))
WHO_subset
```

```
## # A tibble: 5,712 x 6
##    country     iso3   year var   sex   value
##    <chr>       <chr> <dbl> <chr> <chr> <chr>
##  1 Afghanistan AFG    1980 sp    m014  NA
##  2 Afghanistan AFG    1981 sp    m014  NA
##  3 Afghanistan AFG    1982 sp    m014  NA
##  4 Afghanistan AFG    1983 sp    m014  NA
##  5 Afghanistan AFG    1984 sp    m014  NA
##  6 Afghanistan AFG    1985 sp    m014  NA
##  7 Afghanistan AFG    1986 sp    m014  NA
##  8 Afghanistan AFG    1987 sp    m014  NA
##  9 Afghanistan AFG    1988 sp    m014  NA
## 10 Afghanistan AFG    1989 sp    m014  NA
## # ... with 5,702 more rows
```

Read Species and Surveys data sets

```
species <- read_csv("D:/species.csv")
surveys <- read_csv("D:/surveys.csv")
```

Task 6: Join

Checking the imported data

```
str(species)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 54 obs. of  4 variables:
##  $ species_id: chr  "AB" "AH" "AS" "BA" ...
##  $ genus     : chr  "Amphispiza" "Ammospermophilus" "Ammodramus" "Baiomys" ...
##  $ species   : chr  "bilineata" "harrisi" "savannarum" "taylori" ...
##  $ taxa      : chr  "Bird" "Rodent" "Bird" "Rodent" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   species_id = col_character(),
##   ..   genus = col_character(),
##   ..   species = col_character(),
##   ..   taxa = col_character()
##   .. )
```

```
summary(species)
```

```
##   species_id           genus              species         
##  Length:54          Length:54          Length:54         
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##      taxa          
##  Length:54         
##  Class :character  
##  Mode  :character
```

```
str(surveys)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 35549 obs. of  8 variables:
##  $ record_id      : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ month          : num  7 7 7 7 7 7 7 7 7 7 ...
##  $ day            : num  16 16 16 16 16 16 16 16 16 16 ...
##  $ year           : num  1977 1977 1977 1977 1977 ...
##  $ species_id     : chr  "NL" "NL" "DM" "DM" ...
##  $ sex            : chr  "M" "M" "F" "M" ...
##  $ hindfoot_length: num  32 33 37 36 35 14 NA 37 34 20 ...
##  $ weight         : num  NA NA NA NA NA NA NA NA NA NA ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   record_id = col_double(),
##   ..   month = col_double(),
##   ..   day = col_double(),
##   ..   year = col_double(),
##   ..   species_id = col_character(),
##   ..   sex = col_character(),
##   ..   hindfoot_length = col_double(),
##   ..   weight = col_double()
##   .. )
```

```
summary(surveys)
```

```
##    record_id         month             day             year
##  Min.   :    1   Min.   : 1.000   Min.   : 1.00   Min.   :1977
##  1st Qu.: 8888   1st Qu.: 4.000   1st Qu.: 9.00   1st Qu.:1984
##  Median :17775   Median : 6.000   Median :16.00   Median :1990
##  Mean   :17775   Mean   : 6.478   Mean   :15.99   Mean   :1990
##  3rd Qu.:26662   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1997
##  Max.   :35549   Max.   :12.000   Max.   :31.00   Max.   :2002
##
##   species_id            sex            hindfoot_length     weight
##  Length:35549       Length:35549       Min.   : 2.00   Min.   :  4.00
##  Class :character   Class :character   1st Qu.:21.00   1st Qu.: 20.00
##  Mode  :character   Mode  :character   Median :32.00   Median : 37.00
##                                        Mean   :29.29   Mean   : 42.67
##                                        3rd Qu.:36.00   3rd Qu.: 48.00
##                                        Max.   :70.00   Max.   :280.00
##                                        NA's   :4111    NA's   :3266
```

Combine `surveys` and `species` data frames using the key variable `species_id`. For this task, you need to add the species information (`genus`, `species`, `taxa`) to the `surveys` data. Rename the combined data frame as `surveys_combined`.

```
# combining the surveys and species data frame into species_id
surveys_combined <- full_join(species, surveys, by = "species_id")
surveys_combined
```

```
## # A tibble: 35,555 x 11
##    species_id genus species taxa  record_id month   day  year sex
##    <chr>      <chr> <chr>   <chr>     <dbl> <dbl> <dbl> <dbl> <chr>
##  1 AB         Amph~ biline~ Bird       3126     7    21  1980 <NA>
##  2 AB         Amph~ biline~ Bird       3146     7    21  1980 <NA>
##  3 AB         Amph~ biline~ Bird       3152     7    21  1980 <NA>
##  4 AB         Amph~ biline~ Bird       3153     7    21  1980 <NA>
##  5 AB         Amph~ biline~ Bird       3586    12    15  1980 <NA>
##  6 AB         Amph~ biline~ Bird       3702     1    11  1981 <NA>
##  7 AB         Amph~ biline~ Bird       3705     1    11  1981 <NA>
##  8 AB         Amph~ biline~ Bird       3706     1    11  1981 <NA>
##  9 AB         Amph~ biline~ Bird       3775     1    12  1981 <NA>
## 10 AB         Amph~ biline~ Bird       4499     6     4  1981 <NA>
## # ... with 35,545 more rows, and 2 more variables: hindfoot_length <dbl>,
## #   weight <dbl>
```

Task 7: Calculate

```
# species ID
unique(surveys_combined$species_id)
```

```
##  [1] "AB" "AH" "AS" "BA" "CB" "CM" "CQ" "CS" "CT" "CU" "CV" "DM" "DO" "DS"
## [15] "DX" "EO" "GS" "NL" "NX" "OL" "OT" "OX" "PB" "PC" "PE" "PF" "PG" "PH"
## [29] "PI" "PL" "PM" "PP" "PU" "PX" "RF" "RM" "RO" "RX" "SA" "SB" "SC" "SF"
## [43] "SH" "SO" "SS" "ST" "SU" "SX" "UL" "UP" "UR" "US" "ZL" "ZM" NA
```

Species ID ‘DM’ was randomely selected for the next part.

```
# new dataset by filtering for DM
DM <- subset(surveys_combined, surveys_combined$species_id == "DM")
```

```
# structure of DM
str(DM)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    10596 obs. of  11 variables:
##  $ species_id     : chr  "DM" "DM" "DM" "DM" ...
##  $ genus          : chr  "Dipodomys" "Dipodomys" "Dipodomys" "Dipodomys" ...
##  $ species        : chr  "merriami" "merriami" "merriami" "merriami" ...
##  $ taxa           : chr  "Rodent" "Rodent" "Rodent" "Rodent" ...
##  $ record_id      : num  3 4 5 8 9 12 13 14 15 16 ...
##  $ month          : num  7 7 7 7 7 7 7 7 7 7 ...
##  $ day            : num  16 16 16 16 16 16 16 16 16 16 ...
##  $ year           : num  1977 1977 1977 1977 1977 ...
##  $ sex            : chr  "F" "M" "M" "M" ...
##  $ hindfoot_length: num  37 36 35 37 34 38 35 NA 36 36 ...
##  $ weight         : num  NA NA NA NA NA NA NA NA NA NA ...
```

Convert month into a factor.

```
# Converting into a factor
DM$month <- factor(DM$month)
```

```
# conversion
str(DM$month)
```

```
##  Factor w/ 12 levels "1","2","3","4",..: 7 7 7 7 7 7 7 7 7 7 ...
```

```
# Calculate the averge weight of DM excluding NA in each month
DM_WeightByMonth <- aggregate(DM$weight ~ month, DM, mean, na.action = na.omit)
```

```
# Output
kable(DM_WeightByMonth, col.names = c('Month', 'Average Weight of DM'), align = rep('c')) %>%
  kable_styling(bootstrap_options = "striped", full_width = F, "condensed") %>%
  column_spec(1, bold = TRUE, border_right = TRUE, width = "5em") %>%
  column_spec(2, width = "10em")
```

Month	Average Weight of DM
1	42.93697
2	43.95270
3	45.19864
4	44.75049
5	43.18730
6	41.52889
7	41.93692
8	41.84119
9	43.35076
10	42.50429
11	42.35932
12	42.98561

```
# Calculate the averge weight of DM excluding NA in each month
DM_HFLengthByMonth <- aggregate(DM$hindfoot_length ~ month, DM, mean, na.action = na.omit)
```

```
# Output
kable(DM_HFLengthByMonth, col.names = c('Month', 'Average Hindfoot Length of DM'),
      align = rep('c')) %>%
  kable_styling(bootstrap_options = "striped", full_width = F, "condensed") %>%
  column_spec(1, bold = TRUE, border_right = TRUE, width = "5em") %>%
  column_spec(2, width = "10em")
```

Month	Average Hindfoot Length of DM
1	36.09476
2	36.18777
3	36.11765
4	36.20646
5	35.81556
6	35.97699
7	35.71283
8	35.79850
9	35.84908
10	35.94261
11	35.94831
12	36.04545

Task 8: Missing Values

```
# Converting into a factor
surveys_combined$year <- factor(surveys_combined$year)
```

```
# conversion
str(surveys_combined$year)
```

```
##  Factor w/ 26 levels "1977","1978",..: 4 4 4 4 4 5 5 5 5 5 ...
```

```
# YEAR 1998
surveys_combined_year <- subset(surveys_combined, surveys_combined$year == "1998")
```

```
str(surveys_combined_year)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1610 obs. of  11 variables:
##  $ species_id     : chr  "AB" "AB" "AH" "AH" ...
##  $ genus          : chr  "Amphispiza" "Amphispiza" "Ammospermophilus" "Ammospermophilus" ...
##  $ species        : chr  "bilineata" "bilineata" "harrisi" "harrisi" ...
##  $ taxa           : chr  "Bird" "Bird" "Rodent" "Rodent" ...
##  $ record_id      : num  28842 28959 27462 27547 27571 ...
##  $ month          : num  11 12 1 3 3 3 3 5 5 5 ...
##  $ day            : num  21 22 31 1 1 2 2 29 29 29 ...
##  $ year           : Factor w/ 26 levels "1977","1978",..: 22 22 22 22 22 22 22 22 22 22 ...
##  $ sex            : chr  NA NA NA NA ...
##  $ hindfoot_length: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ weight         : num  NA NA NA NA NA NA NA NA NA NA ...
```

```
#  species ID (those which were surveyed in 1998)
unique(surveys_combined_year$species_id)
```

```
##  [1] "AB" "AH" "CB" "CT" "DM" "DO" "DS" "DX" "NL" "OT" "PB" "PC" "PE" "PF"
## [15] "PL" "PM" "PP" "PX" "RM" "SA" "SS" NA
```

```
# Creating a new list value for a count of NA value (from weight)
NA_count <- sapply(surveys_combined_year$weight, function(y) sum(length(which(is.na(y)))))
NA_count
```

```
##    [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##   [35] 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##   [69] 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1
##  [103] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [137] 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [171] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
##  [205] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [239] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [273] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [307] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
##  [341] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [375] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [409] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [443] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [477] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [511] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [545] 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0
##  [579] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [613] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [647] 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 1 0 1 1 0 0 0 0 0 0 0 0 0
##  [681] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [715] 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0
##  [749] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [783] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [817] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [851] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 1 1 1 1
##  [885] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [919] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [953] 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [987] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1021] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
## [1055] 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1089] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1123] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1157] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1191] 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1225] 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1259] 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
## [1293] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
## [1327] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1361] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1395] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1429] 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1463] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1497] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [1531] 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1
## [1565] 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [1599] 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
# Creating a new column in surveys_combined_year
surveys_combined_year$WeightNA <- NA_count
```

```
# sum of NA
weight_by_species <- aggregate(WeightNA ~ species_id, surveys_combined_year, FUN = length)
weight_by_species
```

```
##    species_id WeightNA
## 1          AB        2
## 2          AH       33
## 3          CB        4
## 4          CT        1
## 5          DM      503
## 6          DO      111
## 7          DS        9
## 8          DX        2
## 9          NL       32
## 10         OT      164
## 11         PB      329
## 12         PC        1
## 13         PE       24
## 14         PF       26
## 15         PL        7
## 16         PM      103
## 17         PP      208
## 18         PX        1
## 19         RM       13
## 20         SA        2
## 21         SS       15
```

The total missing values in “weight” column grouped by species (in the year 1998)

```
# Output the results
kable(weight_by_species, col.names = c('Species ID (1998)', 'Number of NA Weight Values'),
      align = rep('c')) %>%
  kable_styling(bootstrap_options = "striped", full_width = F, "condensed") %>%
  column_spec(1, bold = TRUE, border_right = TRUE, width = "8em") %>%
  column_spec(2, width = "8em")
```

Species ID (1998)	Number of NA Weight Values
AB	2
AH	33
CB	4
CT	1
DM	503
DO	111

Species ID (1998)	Number of NA Weight Values
DS	9
DX	2
NL	32
OT	164
PB	329
PC	1
PE	24
PF	26
PL	7
PM	103
PP	208
PX	1
RM	13
SA	2
SS	15

```
# Drop the WeightNA column
surveys_combined_year <- surveys_combined_year[-c(12)]
surveys_combined_year
```

```
## # A tibble: 1,610 x 11
##    species_id genus species taxa  record_id month   day year  sex
##    <chr>      <chr> <chr>   <chr>     <dbl> <dbl> <dbl> <fct> <chr>
##  1 AB         Amph~ biline~ Bird      28842    11    21 1998  <NA>
##  2 AB         Amph~ biline~ Bird      28959    12    22 1998  <NA>
##  3 AH         Ammo~ harrisi Rode~     27462     1    31 1998  <NA>
##  4 AH         Ammo~ harrisi Rode~     27547     3     1 1998  <NA>
##  5 AH         Ammo~ harrisi Rode~     27571     3     1 1998  <NA>
##  6 AH         Ammo~ harrisi Rode~     27628     3     2 1998  <NA>
##  7 AH         Ammo~ harrisi Rode~     27646     3     2 1998  <NA>
##  8 AH         Ammo~ harrisi Rode~     27956     5    29 1998  <NA>
##  9 AH         Ammo~ harrisi Rode~     27959     5    29 1998  <NA>
## 10 AH         Ammo~ harrisi Rode~     27971     5    29 1998  <NA>
## # ... with 1,600 more rows, and 2 more variables: hindfoot_length <dbl>,
## #   weight <dbl>
```

Determine the mean values

```
# Determination of the mean values
average_species_year <- aggregate(surveys_combined_year$weight ~species_id, surveys_combined_yea
r, mean, na.action = na.omit)
```

```
average_species_year
```

```
##    species_id surveys_combined_year$weight
## 1          DM                     43.13140
## 2          DO                     49.73118
## 3          DS                    116.00000
## 4          NL                    159.46667
## 5          OT                     24.67568
## 6          PB                     30.08224
## 7          PE                     20.30435
## 8          PF                      8.72000
## 9          PL                     16.71429
## 10         PM                     20.59140
## 11         PP                     16.26699
## 12         RM                     13.10000
```

```
# renaming the column (data cleaning) to perform next function
names(average_species_year) <- c("species_id", "weight")
```

```
# surveys_weight_imputed created
surveys_weight_imputed <- left_join(surveys_combined_year, average_species_year, by = "species_i
d") %>%
  mutate(weight = ifelse(is.na(weight.x), weight.y, weight.x)) %>%
  select(-weight.y, weight.x)
```

Task 9: Inconsistencies or Special Values

```
# surveys_combined_year
sum(is.na(surveys_combined_year$weight))
```

```
## [1] 215
```

```
# surveys_weight_imputed
sum(is.na(surveys_weight_imputed$weight))
```

```
## [1] 81
```

```
# Select species ID 'DS' from list above
nacheck <- subset(surveys_combined_year, surveys_combined_year$species_id == "DS")
```

```
dim(nacheck)
```

```
## [1]  9 11
```

```
# surveys_weight_imputed
sum(is.na(nacheck$weight))
```

```
## [1] 2
```

```
# Not a Number Count
sum(is.nan(surveys_weight_imputed$weight))
```

```
## [1] 0
```

```
# Infinite Count
sum(is.infinite(surveys_weight_imputed$weight))
```

```
## [1] 0
```

```
# Checking for finite values
sum(is.finite(surveys_weight_imputed$weight))
```

```
## [1] 1529
```

```
# Checking structure
str(surveys_weight_imputed$weight)
```

```
##  num [1:1610] NA NA NA NA NA NA NA NA NA NA ...
```

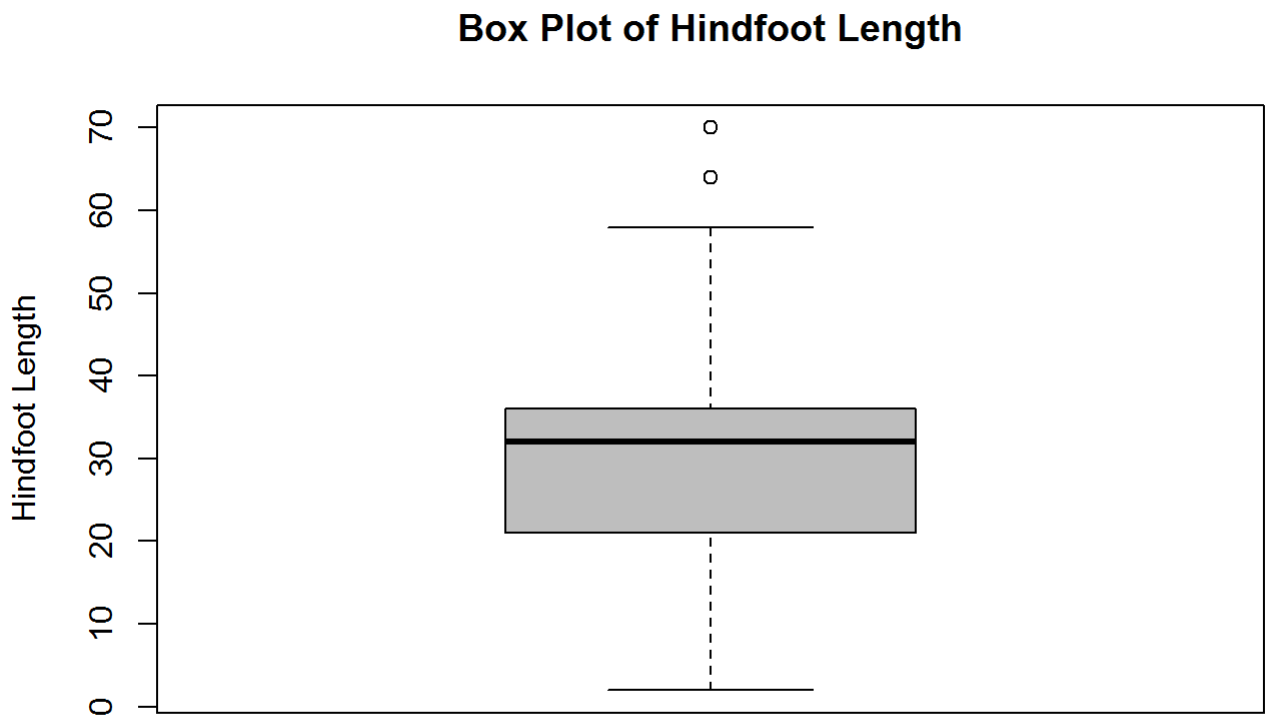
The `surveys_weight_imputed` still includes NA values because each 'weight' valuation from the chosen year (1998) was to start with NA for these species. This implies that they could not be packed with a mean as it was impossible to generate a mean. All the other contorls above showed theoutcomes that were anticipated.

Task 10: Outliers

```
unique(surveys_combined$species_id)
```

```
##  [1] "AB" "AH" "AS" "BA" "CB" "CM" "CQ" "CS" "CT" "CU" "CV" "DM" "DO" "DS"
## [15] "DX" "EO" "GS" "NL" "NX" "OL" "OT" "OX" "PB" "PC" "PE" "PF" "PG" "PH"
## [29] "PI" "PL" "PM" "PP" "PU" "PX" "RF" "RM" "RO" "RX" "SA" "SB" "SC" "SF"
## [43] "SH" "SO" "SS" "ST" "SU" "SX" "UL" "UP" "UR" "US" "ZL" "ZM" NA
```

```
surveys_combined$hindfoot_length %>%  boxplot(main="Box Plot of Hindfoot Length",
                                              ylab="Hindfoot Length", col = "grey")
```




```
# Checking summary statistics
summary(surveys_combined$hindfoot_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2.00   21.00   32.00   29.29   36.00   70.00    4117
```

```
# Droping all the Na
surveys_combined <- dplyr::filter(surveys_combined,  !is.na(hindfoot_length))
```

```
# summary statistics
summary(surveys_combined$hindfoot_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   21.00   32.00   29.29   36.00   70.00
```

```
# z score summary statistics
zscores <- surveys_combined$hindfoot_length %>%  scores(type = "z")
zscores %>% summary()
```

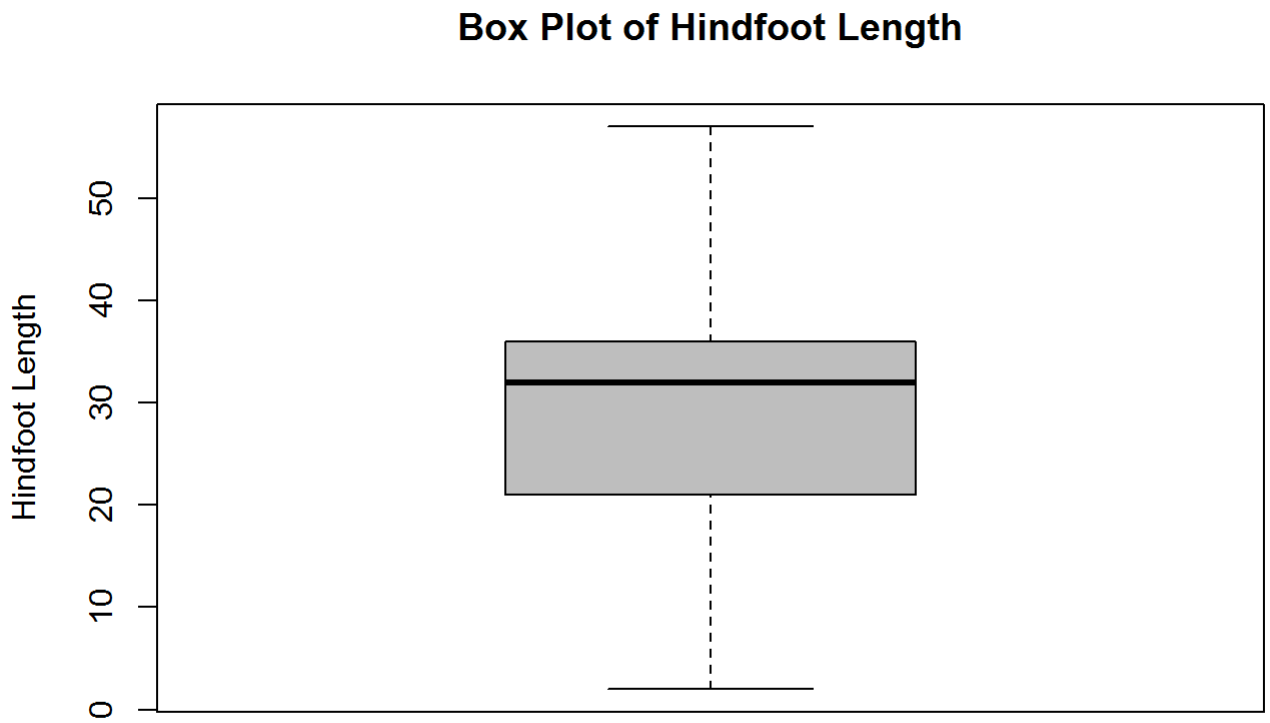
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.8530 -0.8665  0.2835  0.0000  0.7017  4.2565
```

```
# z score values
surveys_combined$hindfoot_length[ which( abs(zscores) >3 )]
```

```
## [1] 58 64 58 70
```

```
# Imputing outliers
surveys_combined$hindfoot_length[ which( abs(zscores) >3 )] <- mean(surveys_combined$hindfoot_le
ngth,
                                                                    na.rm = TRUE)
```

```
# Checking results
surveys_combined$hindfoot_length %>%  boxplot(main="Box Plot of Hindfoot Length",
                                              ylab="Hindfoot Length", col = "grey")
```



By imputation, the outliers were seperated. In the event of these information, the technique of removing outiliers doesnot involve a great deal of thought. This is because there is an enormous variety of distinct species taking over the hindfoot lenght, so this statistical assessment has minimal relevant significance. In addition, the outliers only represented four values out of more than 30,000 so any method can handle with these outliers.