Cleaning-Bad-Data-in-R

2025-03-24

1. Missing Data

1.2. Missing Fields

After loading the heating data I need to transform and format some values. When I apply mutate to transform homes `heating %>% mutate(homes = as.numeric(homes))` from Class: character, R show me a Caused by warning: ! NAs introducidos por coerción. Then I ask for the data using a filter and found characters values (".", "Z"). Finally mutate the values to zero after check with other sources.

```
# Load the tidyverse
library(tidyverse)
```

```
# Load the data file
heating <- read_csv("./exercise_files/1_2/heating.csv")

# Tidy the data
heating <- heating %>% gather(key="age", value="homes", -Source)

knitr::kable(
  summary(heating),
  digits=1, align=rep('c', 5))
```

Source	age	homes
Length:112	Length:112	Length:112
Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character

```
knitr::kable(
  head(heating %>% mutate(homes = as.numeric(homes)), 7),
   digits=1, align=rep('c', 5))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `homes = as.numeric(homes)`.
## Caused by warning:
## ! NAs introducidos por coerción
```

Source	age	homes
Warm-air furnace	Under 25 years old	2546
Steam or hot water system	Under 25 years old	326
Electric heat pump	Under 25 years old	529
Built-in electric units	Under 25 years old	280
Floor, wall, or other built-in hot-air units without ducts	Under 25 years old	267
Room heaters with flue	Under 25 years old	15

Source	age	homes
Room heaters without flue	Under 25 years old	18

```
knitr::kable(
  heating %>% filter(is.na(as.numeric(homes))),
  digits=1, align=rep('c', 5))
```

```
## Warning: There was 1 warning in `filter()`.
## i In argument: `is.na(as.numeric(homes))`.
## Caused by warning:
## ! NAs introducidos por coerción
```

Source	age	homes
Cooking stove	Under 25 years old	.
Cooking stove	25 to 29 years old	Z
Cooking stove	30 to 34 years old	.

```
heating <-
  heating %>%
  mutate(homes=ifelse(homes=='.', 0, homes)) %>%
  mutate(homes=ifelse(homes=='Z', 0, homes)) %>%
  mutate(homes = as.numeric(homes))

  summary(heating)
```

```
##     Source              age                homes
##  Length:112         Length:112         Min.   :    0.00
##  Class :character   Class :character   1st Qu.:   37.25
##  Mode  :character   Mode  :character   Median :  153.50
##                                        Mean   : 1056.12
##                                        3rd Qu.:  674.00
##                                        Max.   :15348.00
```

1.3. Missing Rows

In this exercise I have lands acres from USA, but there are just 42 obs, however the count of states are 50. So is necessary create a data table in tiddy format with the other ones.

```
# Load the data file
land <- read_csv("./exercise_files/1_3/publiclands.csv")

summary(land)
```

```
##     State           PublicLandAcres
##  Length:42          Min.   :   16000
##  Class :character   1st Qu.:  606250
##  Mode  :character   Median : 1156000
##                     Mean   : 4577905
##                     3rd Qu.: 7592500
##                     Max.   :22083000
```

```
nrow(land)
```

```
## [1] 42
```

```
unique(land$State)
```

```
##  [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"
##  [5] "California"     "Colorado"       "Florida"        "Georgia"
##  [9] "Idaho"          "Illinois"       "Indiana"        "Kansas"
## [13] "Kentucky"       "Louisiana"      "Maine"          "Michigan"
## [17] "Minnesota"      "Mississippi"    "Missouri"       "Montana"
## [21] "Nebraska"       "Nevada"         "New Hampshire"  "New Mexico"
## [25] "New York"       "North Carolina" "North Dakota"   "Ohio"
## [29] "Oklahoma"       "Oregon"         "Pennsylvania"   "South Carolina"
## [33] "South Dakota"   "Tennessee"      "Texas"          "Utah"
## [37] "Vermont"        "Virginia"       "Washington"     "West Virginia"
## [41] "Wisconsin"      "Wyoming"
```

```
missing_states <- tibble(State= c('Connecticut', 'Delawere', 'Hawai', 'Iowa', 'Maryland',
                                  'Massachusetts', 'New Jersey', 'Rhode Island'),
                         PublicLandAcres=c(0,0,0,0,0,0,0,0))

land <- rbind(land, missing_states)

knitr::kable(
  tail(land, 10),
    digits=1, align=rep('c', 5))
```

State	PublicLandAcres
Wisconsin	1523000
Wyoming	9238000
Connecticut	0
Delawere	0
Hawai	0
Iowa	0
Maryland	0
Massachusetts	0

State	PublicLandAcres
New Jersey	0
Rhode Island	0

1.4. Agregation and Missing Values

Sometimes when we calculate some statistics like sum or count, the result is NA because R do not have all the values. In this case we need to add na.rm = TRUE.

```
# Load the data file
employees <- read_csv("./exercise_files/1_4/employees.csv")

knitr::kable(employees, digits=1, align=rep('c', 5))
```

FirstName	LastName	Salary	NumDependents
Alexander	Hamilton	40000	3
Aaron	Burr	50000	2
George	Washington	60000	1
Maria	Reynolds	NA	4
Angelica	Schuyler	10000	NA
Hercules	Mulligan	20000	0

```
sum(employees$Salary)
```

```
## [1] NA
```

```
mean(employees$Salary)
```

```
## [1] NA
```

```
max(employees$Salary)
```

```
## [1] NA
```

```
sum(employees$Salary, na.rm = TRUE)
```

```
## [1] 180000
```

```
mean(employees$Salary, na.rm = TRUE)
```

```
## [1] 36000
```

```
max(employees$Salary, na.rm = TRUE)
```

```
## [1] 60000
```