Chapter 2 - Data Exercise 2

Attila Serfozo

2020.10.25

Process of Data Cleaning

```
## For data cleaning I only used the tidyverse package
rm(list=ls())
library(tidyverse)

## Import raw data from ford_focus_scraping.csv
data_in <-
"D:/Egyetem/CEU/Fall_Term/Coding_1-Data_Management_and_Analysis_with_R/Coding/Data_Exercises/Data/"
focusdb <- read_csv(paste0(data_in,"Raw/ford_focus_scraping.csv"))

## Remove unnecessary variables left from scraping
# These are mainly links and page references
focusdb <- select ( focusdb, -c(Pagination ,
                                ‘web-scraper-start-url‘,
                                ‘Pagination-href‘,
                                Open_element) )

## Create an ID from web-scraper-order variable, it can be used as unique variable
focusdb <- separate ( focusdb , ‘web-scraper-order‘, "-" ,
                      into = c("garbage" ,"ID"))

#Dropping the garbage variable, which is the leftover from the ID creation
focusdb <- select ( focusdb , -garbage)

# Convert ID to numeric values
focusdb$ID <- as.numeric( focusdb$ID)

# Order table by ID number
focusdb <- arrange(focusdb , ID)

## Separate the Performance variables kW and horsepower into two seperate variable
focusdb <- separate(focusdb, Performance, "," ,
                    into = c("Performance_kW", "Performance_HP"))

## Creatinge numbers from all numeric variables, therefore dropping all the
## unit of measure characters from the observations like "Ft", km", "kW", etc.
focusdb <- mutate(focusdb,
            Price = as.numeric(gsub("[^0-9]","",Price ) ),
            Kilometers = as.numeric(gsub("[^0-9]","",Kilometers ) ),
```

```r
            Cylinder_capacity = as.numeric(gsub("[^0-9]","",Cylinder_capacity )),
            Performance_kW = as.numeric(gsub("[^0-9]","",Performance_kW ) ),
            Performance_HP = as.numeric(gsub("[^0-9]","",Performance_HP ) ))

# Rename variables to show unit of measure
focusdb <- rename(focusdb,
                  Price_HUF = Price,
                  Cylinder_capacity_cm3 = Cylinder_capacity,
                  Link = `Open_element-href`)

## Filter out observations with missing prices
focusdb <- filter(focusdb,
                  Price_HUF != "NA")

## Converting Transmission observations to Manual/Automatic
focusdb <-
  mutate(focusdb, Transmission = ifelse(focusdb$Transmission ==
                                          "Manuális (6 fokozatú)", "Manuális",
                                  ifelse(focusdb$Transmission ==
                                          "Manuális (5 fokozatú)", "Manuális",
                                          "Automata") ) )

## Converting Registration date to Registration year and creating Age variable
# Separating year and month
focusdb <- separate(focusdb, Registration_date, "/" ,
                     into = c("Registration_date","Month") )

# Changing NA to january in month variable
focusdb$Month[is.na(focusdb$Month)] <- 1

# Formatting year and month as number
focusdb <- mutate(focusdb,
                  Registration_date = as.numeric(gsub("[^0-9]","",Registration_date ) ),
                   Month = round(as.numeric(gsub("[^0-9]","",Month ))/12,3))

# Creating the date of registration with decimals
focusdb <- mutate(focusdb, Registration = Registration_date + Month)

# Calculating age of cars by deducting their reg.date from 2020/10 (2020.833)
focusdb <- mutate(focusdb, Age = 2020.833-focusdb$Registration)

# Removing month and registration in decimals variables
focusdb <- select ( focusdb, -c(Month, Registration) )

## Reordering the columns for the final table
focusdb <- focusdb [, c(1,3,4,5,15,7,9,12,11,10,13,6,8,14,2)]

## Writing out the cleaned database to a csv
write_excel_csv(focusdb , paste0(data_in,"Clean/ford_focus_scraping.csv"))
```

The Clean dataset

I started the data cleaning process with 244 observations from the raw data. During the cleaning process I needed to decide how to handle missing values. As price and age are the main variables of the analysis it was an obvious choice to exclude the 2 observation which did not have values for price. Missing month values in registration date were replaced with January to calculate ages of cars. Missing values were also common in cylinder capacity variable, but this can be acceptable as we can use kW and HP for measurement of performance if we need for further analysis. Missing values in color were the most common, they can also be neglected this time, but if required can be added manually from the website based on pictures of the cars.

As a result I finished with 242 observations in the dataset. The cleaned dataset consists of the following variables, which can be find detailed in the variables.xlsx as well.

- ID - Unique identificator
- Name - Name of the ad
- Price_HUF - Price of the cars in HUF
- Registration_date - Year of registration
- Age - Age of cars in years
- Kilometers - Kilometers run
- Fuel_type - Petrol/Diesel/Electric
- Performance_HP - Performance in horsepower
- Performance_kW - Performance in kilowatt
- Cylinder_capacity_cm3 - Cylinder capacity
- Transmission - Transmission type
- Condition - Condition of the car
- Color - Color of the vehicle
- Nr_pictures - Number of pictures from the car in the ad
- Link - Link to the ad