STATS 7022 - Data Science PG Assignment 1

Dang Thinh Nguyen

2024-06-13

Question 4: Technical Communication

Data cleaning is an important step in data analysis. Raw data could exist multiple issues that need to be solved such as irrelevant value, wrong structure, missing value, duplicate data, erroneous data, or other irregularities (Hall, 2024). Cleaning data improves the quality, accurate, consistent and reliable (Hall, 2024).

To handle the raw data, there are several tasks corresponding to its issues:

1. Irrelevant or duplicated data should be removed to reduce the noise of the data (Tableau, 2024).
2. Correcting the wrong structure (categorical nominal, categorical ordinal, quantitative discrete or quantitative continuous) and modifying the values that have same meaning but different describe (i.e. “No recording”, “No record”, “None”) are essentially to synchronize data structure; hence, dataset is being consistent (Tableau, 2024).
3. Outliers are extremely large or extremely small and cause skewness in the dataset. Removing the outliers is a crucial step to improve performance; however, this step should be considered if the outliers play an important role to explain a specific meaning in the dataset (Tableau, 2024).
4. About missing values, there are three methods to proceed: dropping data, modifying data or finding an alternative method to address the missing values (Tableau, 2024). Deleting missing values is a simple and direct method; however, it might cause losing information which is a remarkable point in a small dataset. Missing values could be placed by alternative values, such as median, same value with relative observations, etc. An issue of this method is that it might cause bias since the new values are generated. The third method, finding an alternative method to address the missing values, is to utilize a method that reduces the effects of missing values to the integrity of whole dataset.

In R, “tidyverse” and “tidymodels” provide a comprehensive set of tools for data cleaning:

1. Loading libraries and creating a small sample dataset:

```r
pacman::p_load(tidyverse, tidymodels, dplyr)
data <- tibble(
  Name = c("John", "Alice", "Bob", "Alice", "Eve",
           "John", "Jane", "Mike", "Sarah"),
  Age = c(45, 30, 30, 30, 40, 45, 35, 40, 1300),
  Gender = c("Male", "Female", "Male", "Female", "Female",
             "Male", "Female", "Male", "Female"),
  Income = c('50000', '60000', 'No record', '60000', 'No record',
             '90000', '70000', '80000', NA)
)
```

2. Taming data to correct the structure:

```r
data <- data %>%
  mutate(Income = ifelse(data$Income == 'No record', NA, data$Income))
data <- data %>% mutate(Age = as.integer(Age),
                        Gender = factor(Gender),
                        Income = as.double(Income))
```

3. Remove duplicates

```r
data <- distinct(data)
```

4. Filtering outliers

```r
data <- data %>% filter(!(Age > 100) | is.na(Age))
```

5. Handling missing values

```r
# Removing
data <- data %>% na.omit()

# Filtering with median
median_income <- median(data$Income, na.rm = TRUE)
data$Income[is.na(data$Income)] <- median_income
```

Reference

1. Hall, D 2024, *Data cleaning: definition, techniques & best practices for 2024*, *Technology Advice*, viewed 12 June 2024, https://technologyadvice.com/blog/information-technology/data-cleaning/.

2. Tableau 2024, *Guide to data cleaning: definition, benefits, components, and how to clean your data*, Tableau, viewed 12 June 2024, https://www.tableau.com/learn/articles/what-is-data-cleaning.