Project 2: Data set 1: Untidy Hypothetical Data

Mohammad Zahid Chowdhury

2025-03-08

Introduction : This is a hypothetical untidy dataset and this dataset contains four variables including name, age, scores, and address. This data set has missing values, multiple information in one coulumn. I saved this data as csv and then clean it. After cleaning this data I will analyze the data.

Loading required packages:

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.2
```

```
## Warning: package 'readr' was built under R version 4.4.2
```

```
## Warning: package 'dplyr' was built under R version 4.4.2
```

```
## Warning: package 'stringr' was built under R version 4.4.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(ggplot2)
```

Read the dataset:

```r
students_score <- read.csv("https://raw.githubusercontent.com/zahid607/Project-2/refs/heads/main/Student

colnames(students_score)
```

```
## [1] "Name"    "Age"     "Scores"  "Address"
```

```r
head(students_score)
```

```
##          Name Age                                Scores       Address
## 1    John Doe  25               Math: 80, Science: 85  123 Main St.
## 2  Jane Smith  30               Math: 92, Science: 88   456 Oak St.
## 3 Sarah White  22               Math: 75, History: 80  789 Pine St.
## 4   Bob Brown  28               Math: 85, Science: 90
## 5 Carol Green  26 Math: 78, Science: 80, History: 85 101 Maple St.
## 6  Alice Blue  NA               Math: 90, History: 85   202 Elm St.
```

Transform the data:

```r
# Separate the Scores column into Math, Science, and History
students_score <- students_score %>%
  separate(Scores, into = c("Math", "Science", "History"), sep = ", ", remove = FALSE)
```

```
## Warning: Expected 3 pieces. Missing pieces filled with `NA` in 5 rows [1, 2, 3,
## 4, 6].
```

```r
# Check the cleaned data
head(students_score)
```

```
##          Name Age                                Scores     Math     Science
## 1    John Doe  25               Math: 80, Science: 85 Math: 80 Science: 85
## 2  Jane Smith  30               Math: 92, Science: 88 Math: 92 Science: 88
## 3 Sarah White  22               Math: 75, History: 80 Math: 75 History: 80
## 4   Bob Brown  28               Math: 85, Science: 90 Math: 85 Science: 90
## 5 Carol Green  26 Math: 78, Science: 80, History: 85 Math: 78 Science: 80
## 6  Alice Blue  NA               Math: 90, History: 85 Math: 90 History: 85
##       History       Address
## 1        <NA>  123 Main St.
## 2        <NA>   456 Oak St.
## 3        <NA>  789 Pine St.
## 4        <NA>
## 5 History: 85 101 Maple St.
## 6        <NA>   202 Elm St.
```

Reshape the column & the “Math:”, “Science:”, “History:” text, leaving only the numeric values:

```r
students_score <- students_score %>%
  mutate(
    Math = as.numeric(gsub("Math: ", "", Math)),
    Science = as.numeric(gsub("Science: ", "", Science)),
    History = as.numeric(gsub("History: ", "", History))
  )
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `Science = as.numeric(gsub("Science: ", "", Science))`.
## Caused by warning:
## ! NAs introduced by coercion
```

```r
head(students_score)
```

```
##          Name Age                              Scores Math Science History
## 1    John Doe  25              Math: 80, Science: 85   80      85      NA
## 2  Jane Smith  30              Math: 92, Science: 88   92      88      NA
## 3 Sarah White  22              Math: 75, History: 80   75      NA      NA
## 4   Bob Brown  28              Math: 85, Science: 90   85      90      NA
## 5 Carol Green  26 Math: 78, Science: 80, History: 85   78      80      85
## 6  Alice Blue  NA              Math: 90, History: 85   90      NA      NA
##         Address
## 1  123 Main St.
## 2   456 Oak St.
## 3  789 Pine St.
## 4
## 5 101 Maple St.
## 6   202 Elm St.
```

```r
View(students_score)
```

Number of Missing Data:

```r
# Check for missing data
missing_data <- colSums(is.na(students_score))
print(missing_data)
```

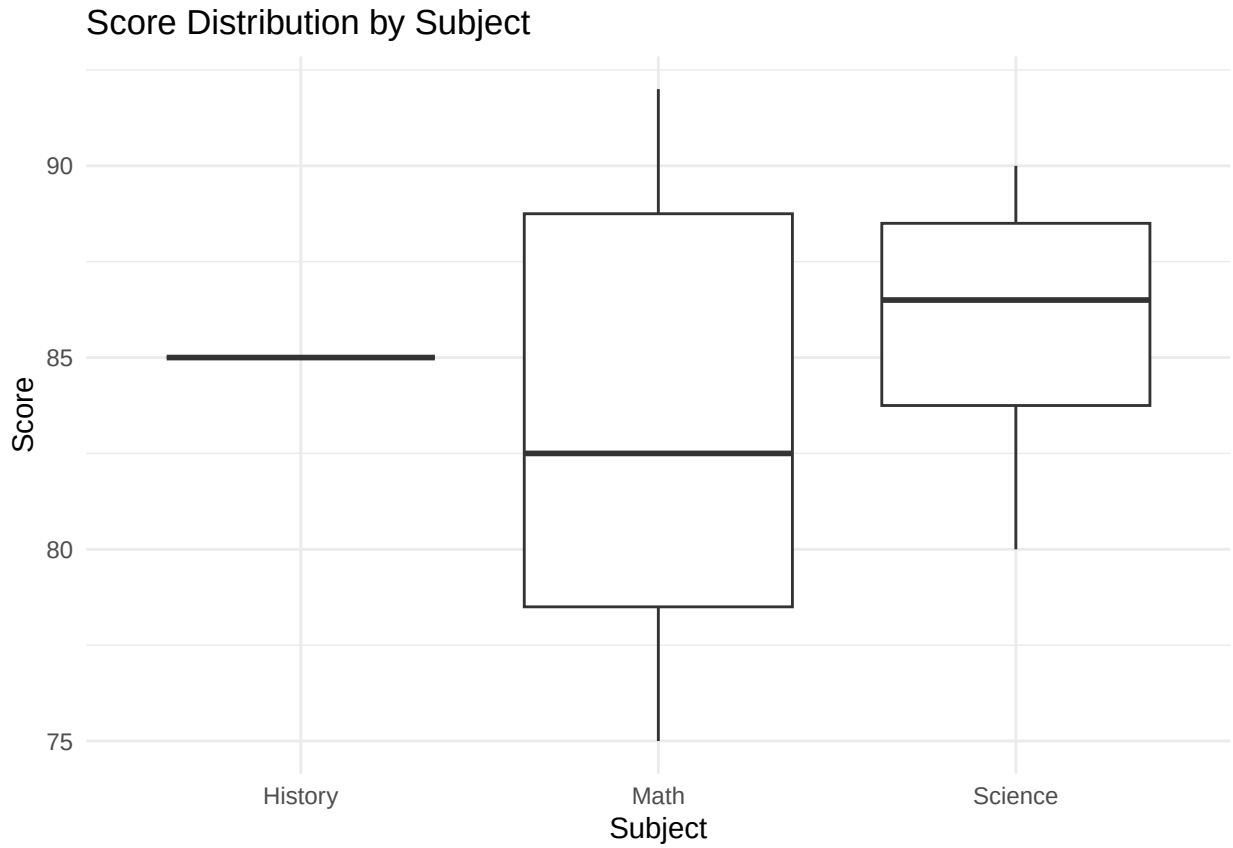
```
##    Name     Age  Scores    Math Science History Address
##       0       1       0       0       2       5       0
```

Boxplots to compare the distribution of scores in different subjects and any outlier detection.

```r
# Reshape data for plotting
students_score <- students_score %>%
  pivot_longer(cols = c("Math", "Science", "History"),
               names_to = "Subject", values_to = "Score")

ggplot(students_score, aes(x = Subject, y = Score)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Score Distribution by Subject")
```

```
## Warning: Removed 7 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



Average Score of Students.

```r
students_score_avg <- students_score %>%
  group_by(Name) %>%
  summarise(avg_score = mean(Score, na.rm = TRUE)) %>%
  arrange(desc(avg_score))

head(students_score_avg)
```

```
## # A tibble: 6 x 2
##   Name        avg_score
##   <chr>           <dbl>
## 1 Alice Blue       90
## 2 Jane Smith       90
## 3 Bob Brown        87.5
## 4 John Doe         82.5
## 5 Carol Green      81
## 6 Sarah White      75
```

Conclusion: In this analysis, I worked with a hypothetical dataset containing information about students, including their names, ages, scores in various subjects (Math, Science, History), and their addresses. The primary goal of this analysis was to clean, transform, and explore the data to gain insights into the students' academic performance across different subjects.