Lab Program 4

R V Abhishek

2025-09-09

Data Import, Cleaning, and Export with Advanced Data Wrangling

Objective: Real World Data Cleaning Processes and emphasis on usage of advanced data wrangling techniques in R.

```
# Load necessary libraries
library(tidyverse)
```

```
-- 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   3.5.2     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- 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(titanic)
library(dplyr)
library(caret)
```

```
Loading required package: lattice

Attaching package: 'caret'

The following object is masked from 'package:purrr':

    lift
```

```
library(ggcorrplot)

# Load the Titanic dataset
data <- titanic::titanic_train

# Handle the missing data
# Replace missing values in the 'Age' column with the median age
data$Age[is.na(data$Age)] <- median(data$Age, na.rm = TRUE)
```

```r
# Replace missing values in the 'Embarked' column with the mode
mode_embarked <- as.character(names(sort(table(data$Embarked), decreasing = TRUE)[1]))
data$Embarked[is.na(data$Embarked)] <- mode_embarked

# Define the numeric columns for z-score and correlation calculation
numeric_columns <- c("Age", "SibSp", "Parch", "Fare")

# Remove outliers using z-score
z_scores <- as.data.frame(scale(data[, numeric_columns]))

# Identify the rows that have z_scores greater than 3 or less than -3 (outliers)
outlier_rows <- apply(z_scores, 1, function(row) any(abs(row) > 3))

# Filter out Outliers
data_cleaned <- data[!outlier_rows, ]

# Summarize the dataset before and after cleaning
summary_before <- summary(data)
summary_after <- summary(data_cleaned)

# Calculate Correlation Matrix (fixed)
correlation_matrix <- cor(data_cleaned[, numeric_columns], use = "complete.obs")

# Export cleaned data onto a new CSV file
write.csv(data_cleaned, "titanic_cleaned.csv", row.names = FALSE)

# Display Summaries
print("Summary Before Cleaning:")
```

```
[1] "Summary Before Cleaning:"
```

```r
print(summary_before)
```

```
  PassengerId       Survived          Pclass          Name
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
 Median :446.0   Median :0.0000   Median :3.000   Mode  :character
 Mean   :446.0   Mean   :0.3838   Mean   :2.309
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
 Max.   :891.0   Max.   :1.0000   Max.   :3.000
     Sex                 Age            SibSp           Parch
 Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
 Class :character   1st Qu.:22.00   1st Qu.:0.000   1st Qu.:0.0000
 Mode  :character   Median :28.00   Median :0.000   Median :0.0000
                    Mean   :29.36   Mean   :0.523   Mean   :0.3816
                    3rd Qu.:35.00   3rd Qu.:1.000   3rd Qu.:0.0000
                    Max.   :80.00   Max.   :8.000   Max.   :6.0000
    Ticket               Fare           Cabin             Embarked
 Length:891         Min.   :  0.00   Length:891         Length:891
 Class :character   1st Qu.:  7.91   Class :character   Class :character
 Mode  :character   Median : 14.45   Mode  :character   Mode  :character
                    Mean   : 32.20
```

```
                    3rd Qu.: 31.00
                    Max.   :512.33
```

```
print("Summary After Cleaning:")
```

```
[1] "Summary After Cleaning:"
```

```
print(summary_after)
```

```
  PassengerId       Survived          Pclass          Name
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:820
 1st Qu.:226.8   1st Qu.:0.0000   1st Qu.:2.000   Class :character
 Median :446.5   Median :0.0000   Median :3.000   Mode  :character
 Mean   :445.7   Mean   :0.3902   Mean   :2.311
 3rd Qu.:661.2   3rd Qu.:1.0000   3rd Qu.:3.000
 Max.   :891.0   Max.   :1.0000   Max.   :3.000
     Sex                 Age            SibSp            Parch
 Length:820         Min.   : 0.42   Min.   :0.0000   Min.   :0.0000
 Class :character   1st Qu.:23.00   1st Qu.:0.0000   1st Qu.:0.0000
 Mode  :character   Median :28.00   Median :0.0000   Median :0.0000
                    Mean   :29.44   Mean   :0.3488   Mean   :0.2549
                    3rd Qu.:35.00   3rd Qu.:1.0000   3rd Qu.:0.0000
                    Max.   :66.00   Max.   :3.0000   Max.   :2.0000
    Ticket               Fare            Cabin             Embarked
 Length:820         Min.   :  0.000   Length:820         Length:820
 Class :character   1st Qu.:  7.896   Class :character   Class :character
 Mode  :character   Median : 13.000   Mode  :character   Mode  :character
                    Mean   : 25.836
                    3rd Qu.: 27.000
                    Max.   :164.867
```

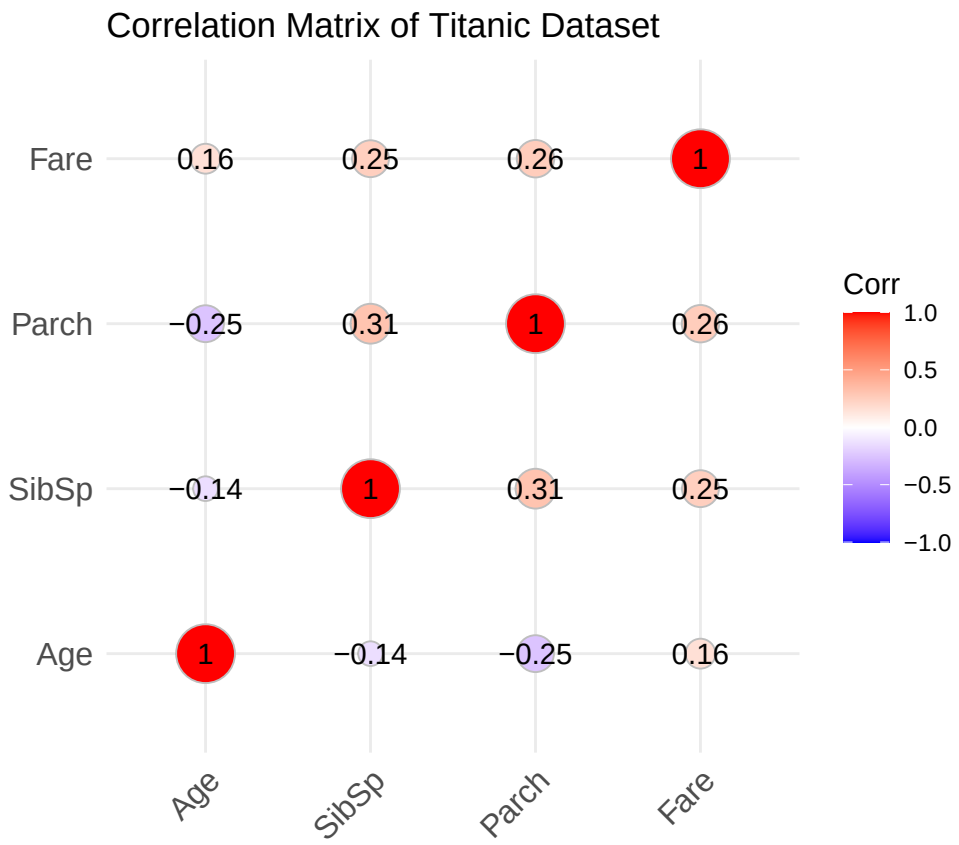
```
print("Correlation Matrix:")
```

```
[1] "Correlation Matrix:"
```

```
print(correlation_matrix)
```

```
             Age      SibSp      Parch      Fare
Age    1.0000000 -0.1439118 -0.2517719 0.1598100
SibSp -0.1439118  1.0000000  0.3072105 0.2472157
Parch -0.2517719  0.3072105  1.0000000 0.2599031
Fare   0.1598100  0.2472157  0.2599031 1.0000000
```

```
# Visualize Correlation Matrix (fixed)
ggcorrplot(correlation_matrix,
           method = "circle",
           lab = TRUE) +
  ggtitle("Correlation Matrix of Titanic Dataset")
```



Objective - Data Import, Cleaning, and Export with Adult Income Dataset

```
# Load necessary libraries
library(tidyverse)
library(dplyr)
library(caret)
library(ggcorrplot)

# Load the Adult Income dataset
data <- read.csv("D:/Coding/Coding/Time Series Analysis/Lab 4/adult.data", header = FALSE)

# Assign column names based on the dataset documentation
colnames(data) <- c('age', 'workclass', 'fnlwgt', 'education', 'education_num', 'marital_status', 'occu

# Handle missing values represented by '?'
data[data == '?'] <- NA

# Replace categorical missing values with mode
replace_mode <- function(x){
  mode_val <- as.character(names(sort(table(x), decreasing = TRUE)[1]))
  x[is.na(x)] <- mode_val
  return(x)
}

data <- data %>%
  mutate_if(is.character, replace_mode)
```

```r
# Replace numeric missing values with median
data <- data %>%
  mutate_if(is.numeric, ~ifelse(is.na(.), median(., na.rm = TRUE), .))

# Define the remove_outliers function
remove_outliers <- function(x){
  z_scores <- scale(x)
  x[abs(z_scores) <= 3]
}

# Remove outliers using z-score
numeric_columns <- sapply(data, is.numeric)

# Apply z-score outlier removal to numeric columns
data_cleaned <- data %>%
  filter(!apply(as.data.frame(scale(data[, numeric_columns])), 1, function(row) any(abs(row) > 3)))

# Summarize before and after cleaning
summary_before <- summary(data)
summary_after <- summary(data_cleaned)

# Calculate correlation Matrix
correlation_matrix <- cor(data_cleaned[, numeric_columns], use = "complete.obs")

# Export as CSV
write.csv(data_cleaned, "cleaned_adult_income_data.csv", row.names = FALSE)

# Display Summaries
print("Summary Before Cleaning:")
```

```
[1] "Summary Before Cleaning:"
```

```r
print(summary_before)
```

```
      age          workclass             fnlwgt          education
 Min.   :17.00   Length:32561       Min.   :  12285   Length:32561
 1st Qu.:28.00   Class :character   1st Qu.: 117827   Class :character
 Median :37.00   Mode  :character   Median : 178356   Mode  :character
 Mean   :38.58                      Mean   : 189778
 3rd Qu.:48.00                      3rd Qu.: 237051
 Max.   :90.00                      Max.   :1484705
 education_num   marital_status      occupation        relationship
 Min.   : 1.00   Length:32561       Length:32561       Length:32561
 1st Qu.: 9.00   Class :character   Class :character   Class :character
 Median :10.00   Mode  :character   Mode  :character   Mode  :character
 Mean   :10.08
 3rd Qu.:12.00
 Max.   :16.00
     race               sex             capital_gain    capital_loss
 Length:32561       Length:32561       Min.   :    0   Min.   :   0.0
 Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.0
 Mode  :character   Mode  :character   Median :    0   Median :   0.0
```

```
                                   Mean   : 1078   Mean   :  87.3
                                   3rd Qu.:    0   3rd Qu.:   0.0
                                   Max.   :99999   Max.   :4356.0
 hours_per_week  native_country        income
 Min.   : 1.00   Length:32561       Length:32561
 1st Qu.:40.00   Class :character   Class :character
 Median :40.00   Mode  :character   Mode  :character
 Mean   :40.44
 3rd Qu.:45.00
 Max.   :99.00
```

```
print("Summary After Cleaning:")
```

```
[1] "Summary After Cleaning:"
```

```
print(summary_after)
```

```
      age          workclass             fnlwgt         education
 Min.   :17.00   Length:29828       Min.   : 12285   Length:29828
 1st Qu.:27.00   Class :character   1st Qu.:117509   Class :character
 Median :37.00   Mode  :character   Median :177667   Mode  :character
 Mean   :38.14                      Mean   :185193
 3rd Qu.:47.00                      3rd Qu.:234279
 Max.   :79.00                      Max.   :506329
 education_num   marital_status      occupation        relationship
 Min.   : 3.00   Length:29828       Length:29828       Length:29828
 1st Qu.: 9.00   Class :character   Class :character   Class :character
 Median :10.00   Mode  :character   Mode  :character   Mode  :character
 Mean   :10.08
 3rd Qu.:12.00
 Max.   :16.00
    race               sex             capital_gain      capital_loss
 Length:29828       Length:29828       Min.   :    0.0   Min.   :   0.000
 Class :character   Class :character   1st Qu.:    0.0   1st Qu.:   0.000
 Mode  :character   Mode  :character   Median :    0.0   Median :   0.000
                                       Mean   :  570.2   Mean   :   1.209
                                       3rd Qu.:    0.0   3rd Qu.:   0.000
                                       Max.   :22040.0   Max.   :1258.000
 hours_per_week native_country        income
 Min.   : 4.0   Length:29828       Length:29828
 1st Qu.:40.0   Class :character   Class :character
 Median :40.0   Mode  :character   Mode  :character
 Mean   :39.9
 3rd Qu.:45.0
 Max.   :77.0
```

```
print("Correlation Matrix:")
```

```
[1] "Correlation Matrix:"
```

```r
print(correlation_matrix)
```

```
                     age       fnlwgt education_num capital_gain capital_loss
age           1.00000000 -0.074427786   0.041427102  0.131043981  0.020824647
fnlwgt       -0.07442779  1.000000000  -0.037482414 -0.002378925  0.002583047
education_num 0.04142710 -0.037482414   1.000000000  0.154844283  0.009481359
capital_gain  0.13104398 -0.002378925   0.154844283  1.000000000 -0.009038231
capital_loss  0.02082465  0.002583047   0.009481359 -0.009038231  1.000000000
hours_per_week 0.09219535 -0.015375555   0.150513483  0.097209049 -0.003089539
              hours_per_week
age               0.092195352
fnlwgt           -0.015375555
education_num     0.150513483
capital_gain      0.097209049
capital_loss     -0.003089539
hours_per_week    1.000000000
```

```r
# Visualize Correlation Matrix (fixed)
ggcorrplot(correlation_matrix,
           method = "circle",
           lab = TRUE) +
  ggtitle("Correlation Matrix of Adult Income Dataset")
```