Programming Assignment #2

Download a dataset from Kaggle, and write an R program to do the designated computations.

Name: 王韋峰, 劉家均

Student ID : 411221424, 411221426

Date : 2025/10/28

1. Problem description

The purpose of this assignment is to practice data analysis using **R** by performing specific computations on a real dataset. The dataset chosen for this task is the *Airline Dataset Updated - v2.csv* from Kaggle. The program must accomplish the following:

A) Calculate the number of male and female passengers.
B) Find the average age of male passengers.
C) Find the average age of female passengers.
D) List the top 10 nationalities based on frequency.
The program uses the **readr** and **dplyr** packages to read, filter, and summarize data efficiently.

2. Highlight the way you write your program

1. Imported necessary R libraries: **readr** and **dplyr** for reading and manipulating CSV data.
2. Loaded the dataset using **read_csv()** and stored it in a data frame named **airline**.
3. Used **count()** to compute the number of passengers by gender.
4. Filtered the dataset to separate male and female passengers using **filter()**.
5. Used **summarise()** with **mean()** to calculate the average age for each gender, ignoring missing values with **na.rm = TRUE**.
6. Counted the frequency of each nationality and sorted it in descending order to list the top 10. This approach minimizes the use of loops and takes advantage of R's vectorized and declarative data operations.

3. The program listing

```r
library(readr)
library(dplyr)

# Load dataset
airline <- read_csv("../data_set/Airline Dataset/Airline Dataset Updated - v2.csv")

# (A) Count male and female passengers
gender_count <- count(airline, Gender)

# (B) Average age of male passengers
male_data <- filter(airline, Gender == "Male")
avg_male_age <- summarise(male_data, mean_age = mean(Age, na.rm = TRUE))

# (C) Average age of female passengers
female_data <- filter(airline, Gender == "Female")
avg_female_age <- summarise(female_data, mean_age = mean(Age, na.rm = TRUE))

# (D) Top 10 nationalities
nationality_count <- count(airline, Nationality, sort = TRUE)
top_nationalities <- head(nationality_count, 10)

# Output results
cat("(A) Male passengers:",
    gender_count$n[gender_count$Gender == "Male"],
    ", Female passengers:",
    gender_count$n[gender_count$Gender == "Female"], "\n")
cat("(B) Average Age of Male Passengers:", round(avg_male_age$mean_age, 2), "\n")
cat("(C) Average Age of Female Passengers:", round(avg_female_age$mean_age, 2), "\n")
cat("(D) Top 10 Nationalities:\n")
print(top_nationalities)
```

4. The run result

```
(A) Male passengers: 49598 , Female passengers: 49021
(B) Average Age of Male Passengers: 45.49
(C) Average Age of Female Passengers: 45.52
(D) Top 10 Nationalities:
# A tibble: 10 × 2
   Nationality       n
   <chr>         <int>
 1 China         18317
 2 Indonesia     10559
 3 Russia         5693
 4 Philippines    5239
 5 Brazil         3791
 6 Portugal       3299
 7 Poland         3245
 8 France         2907
 9 Sweden         2397
10 United States  2105
```

5. Discussion

This exercise demonstrated how powerful and convenient **R** can be for data analytics tasks. Compared with traditional programming, R's syntax allows complex computations to be expressed in very few lines.

The **dplyr** package simplifies data manipulation by providing clear and intuitive functions like **filter()**, **summarise()**, and **count().** Each operation can be chained with pipes (`|>` or `%>%`) to make the code clean and readable. This approach is not only efficient but also mirrors real-world data workflows in big data analysis.

Through this task, we realized that R excels in statistical data handling, particularly for quick computations and exploratory analysis. While Python is more flexible across domains, R provides an elegant and concise syntax for statistical modeling and data summarization.

Overall, this assignment strengthened our understanding of how R can be applied to real datasets. we gained practical skills in summarizing categorical and numerical data, identifying key demographic insights, and creating reproducible analytical workflows—all fundamental skills in the era of Big Data.