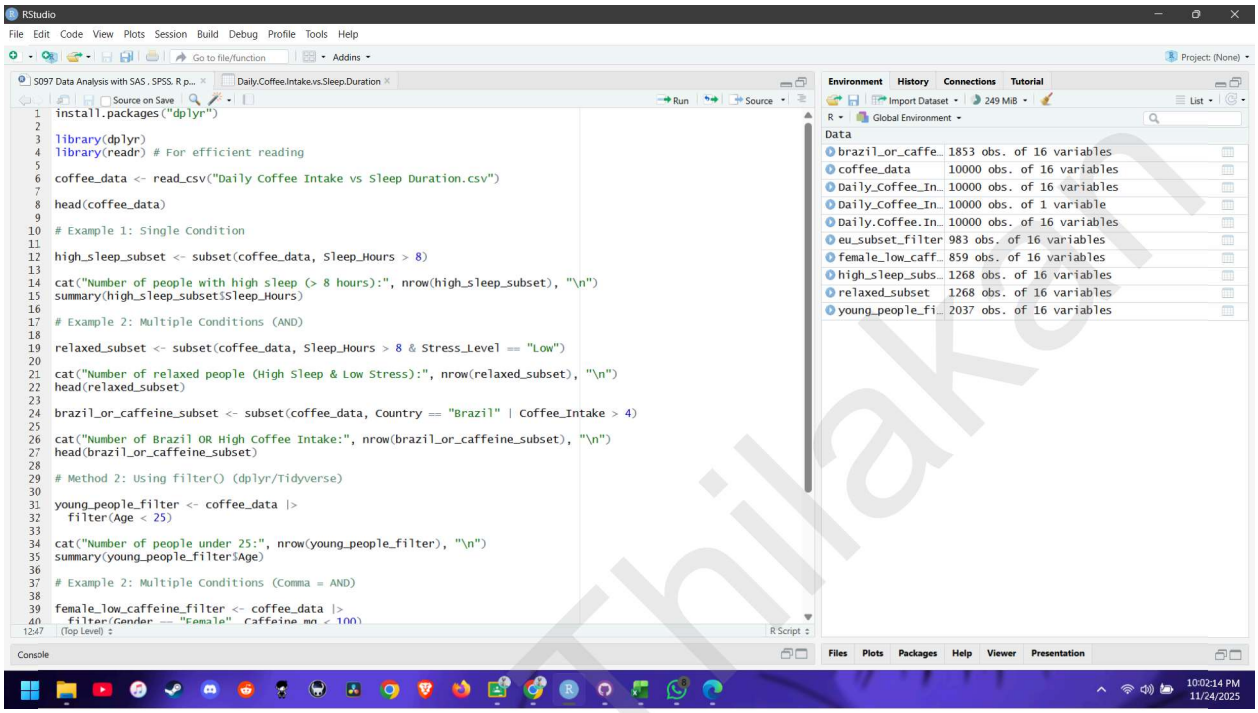
SHETH LUJ AND SIR MV COLLEGE
Subject: Data Analysis with SAS / SPSS /R

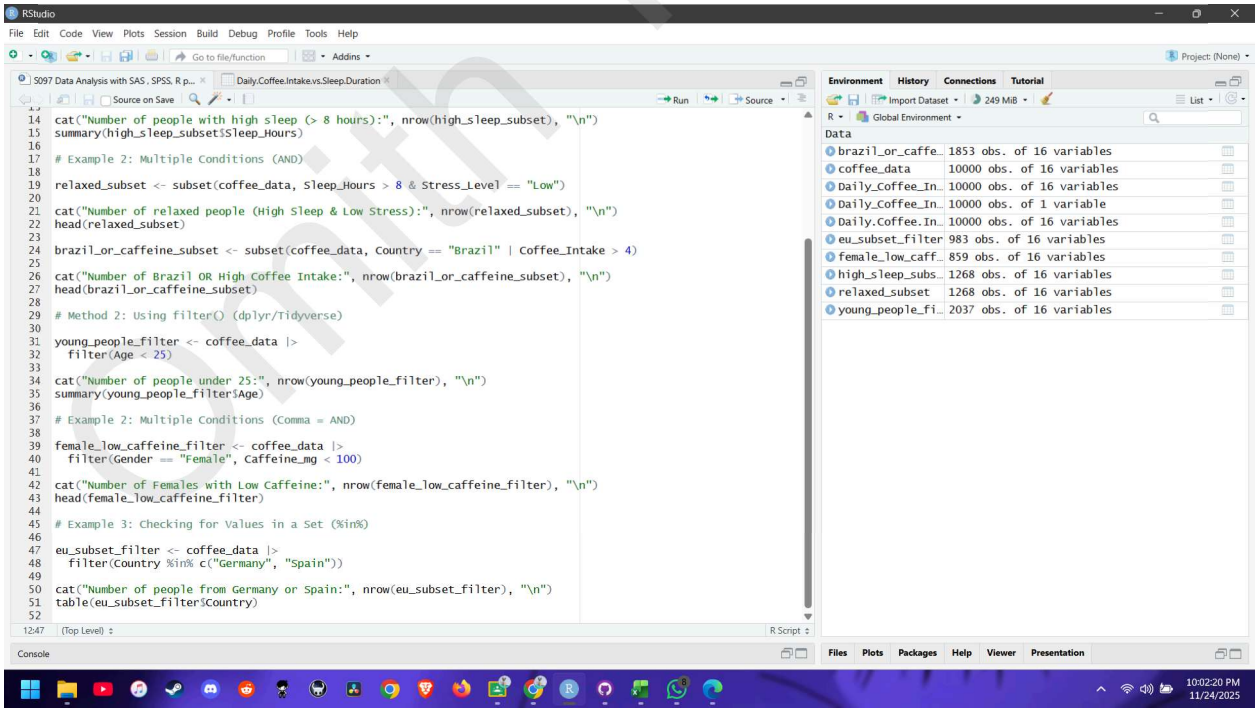
Practical No: 4

Aim: Applying conditional filters subset() or filter() in R.

Code:



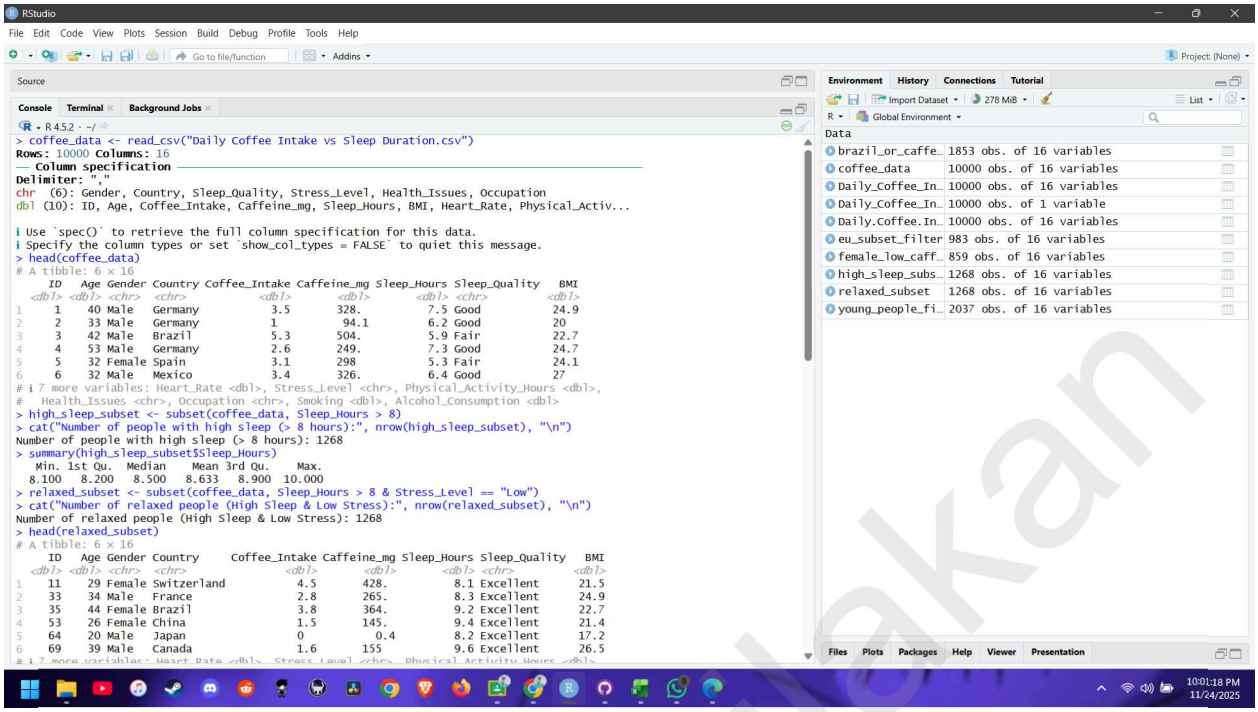
```
install.packages("dplyr")

library(dplyr)
library(readr) # For efficient reading

coffee_data <- read_csv("Daily Coffee Intake vs Sleep Duration.csv")

head(coffee_data)

# Example 1: Single Condition

high_sleep_subset <- subset(coffee_data, Sleep_Hours > 8)

cat("Number of people with high sleep (> 8 hours):", nrow(high_sleep_subset), "\n")
summary(high_sleep_subset$Sleep_Hours)

# Example 2: Multiple Conditions (AND)

relaxed_subset <- subset(coffee_data, Sleep_Hours > 8 & Stress_Level == "Low")

cat("Number of relaxed people (High Sleep & Low Stress):", nrow(relaxed_subset), "\n")
head(relaxed_subset)

brazil_or_caffeine_subset <- subset(coffee_data, Country == "Brazil" | Coffee_Intake > 4)

cat("Number of Brazil OR High Coffee Intake:", nrow(brazil_or_caffeine_subset), "\n")
head(brazil_or_caffeine_subset)

# Method 2: Using filter() (dplyr/Tidyverse)

young_people_filter <- coffee_data |>
  filter(Age < 25)

cat("Number of people under 25:", nrow(young_people_filter), "\n")
summary(young_people_filter$Age)

# Example 2: Multiple Conditions (Comma = AND)

female_low_caffeine_filter <- coffee_data |>
  filter(Gender == "Female", Caffeine_mg < 100)
```



```
cat("Number of people with high sleep (> 8 hours):", nrow(high_sleep_subset), "\n")
summary(high_sleep_subset$Sleep_Hours)

# Example 2: Multiple Conditions (AND)

relaxed_subset <- subset(coffee_data, Sleep_Hours > 8 & Stress_Level == "Low")

cat("Number of relaxed people (High Sleep & Low Stress):", nrow(relaxed_subset), "\n")
head(relaxed_subset)

brazil_or_caffeine_subset <- subset(coffee_data, Country == "Brazil" | Coffee_Intake > 4)

cat("Number of Brazil OR High Coffee Intake:", nrow(brazil_or_caffeine_subset), "\n")
head(brazil_or_caffeine_subset)

# Method 2: Using filter() (dplyr/Tidyverse)

young_people_filter <- coffee_data |>
  filter(Age < 25)

cat("Number of people under 25:", nrow(young_people_filter), "\n")
summary(young_people_filter$Age)

# Example 2: Multiple Conditions (Comma = AND)

female_low_caffeine_filter <- coffee_data |>
  filter(Gender == "Female", Caffeine_mg < 100)

cat("Number of Females with Low Caffeine:", nrow(female_low_caffeine_filter), "\n")
head(female_low_caffeine_filter)

# Example 3: Checking for Values in a Set (%in%)

eu_subset_filter <- coffee_data |>
  filter(Country %in% c("Germany", "Spain"))

cat("Number of people from Germany or Spain:", nrow(eu_subset_filter), "\n")
table(eu_subset_filter$Country)
```

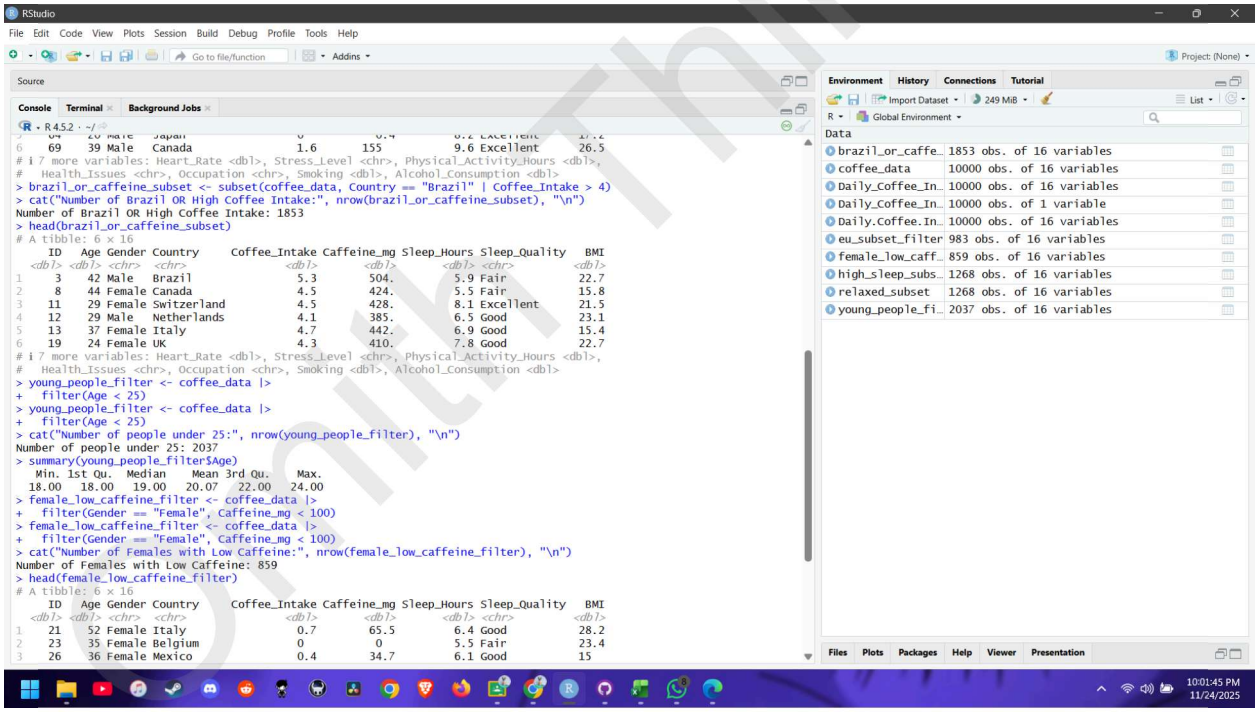
Output:

Omith Thilakan
S097



```
> coffee_data <- read_csv("Daily Coffee Intake vs Sleep Duration.csv")
Rows: 10000 Columns: 16
── Column specification ────────────────────────────────────────
Delimiter: ","
chr  (6): Gender, Country, Sleep_Quality, Stress_Level, Health_Issues, Occupation
dbl (10): ID, Age, Coffee_Intake, Caffeine_mg, Sleep_Hours, BMI, Heart_Rate, Physical_Activ...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
> head(coffee_data)
# A tibble: 6 × 16
     ID   Age Gender Country Coffee_Intake Caffeine_mg Sleep_Hours Sleep_Quality   BMI
  <dbl> <dbl> <chr>  <chr>           <dbl>       <dbl>       <dbl> <chr>         <dbl>
1     1    40 Male   Germany           3.5         328.         7.5 Good           24.9
2     2    33 Male   Germany           1            94.1        6.2 Good           20
3     3    42 Male   Brazil            5.3         504.         5.9 Fair           22.7
4     4    53 Male   Germany           2.6         249.         7.3 Good           24.7
5     5    32 Female Spain             3.1         298          5.3 Fair           24.1
6     6    32 Male   Mexico            3.4         326.         6.4 Good           27
# ℹ 7 more variables: Heart_Rate <dbl>, Stress_Level <chr>, Physical_Activity_Hours <dbl>,
#   Health_Issues <chr>, Occupation <chr>, Smoking <dbl>, Alcohol_Consumption <dbl>
> high_sleep_subset <- subset(coffee_data, Sleep_Hours > 8)
> cat("Number of people with high sleep (> 8 hours):", nrow(high_sleep_subset), "\n")
Number of people with high sleep (> 8 hours): 1268
> summary(high_sleep_subset$Sleep_Hours)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.100   8.200   8.500   8.633   8.900  10.000
> relaxed_subset <- subset(coffee_data, Sleep_Hours > 8 & Stress_Level == "Low")
> cat("Number of relaxed people (High Sleep & Low Stress):", nrow(relaxed_subset), "\n")
Number of relaxed people (High Sleep & Low Stress): 1268
> head(relaxed_subset)
# A tibble: 6 × 16
     ID   Age Gender Country     Coffee_Intake Caffeine_mg Sleep_Hours Sleep_Quality   BMI
  <dbl> <dbl> <chr>  <chr>               <dbl>       <dbl>       <dbl> <chr>         <dbl>
1    11    29 Female Switzerland           4.5         428.         8.1 Excellent      21.5
2    33    34 Male   France                2.8         265.         8.3 Excellent      24.9
3    35    44 Female Brazil                3.8         364.         9.2 Excellent      22.7
4    53    26 Female China                 1.5         145.         9.4 Excellent      21.4
5    64    20 Male   Japan                 0             0.4        8.2 Excellent      17.2
6    69    39 Male   Canada                1.6         155          9.6 Excellent      26.5
```

| Data | |
|---|---|
| brazil_or_caffe… | 1853 obs. of 16 variables |
| coffee_data | 10000 obs. of 16 variables |
| Daily_Coffee_In… | 10000 obs. of 16 variables |
| Daily_Coffee_In… | 10000 obs. of 1 variable |
| Daily.Coffee.In… | 10000 obs. of 16 variables |
| eu_subset_filter | 983 obs. of 16 variables |
| female_low_caff… | 859 obs. of 16 variables |
| high_sleep_subs… | 1268 obs. of 16 variables |
| relaxed_subset | 1268 obs. of 16 variables |
| young_people_fi… | 2037 obs. of 16 variables |



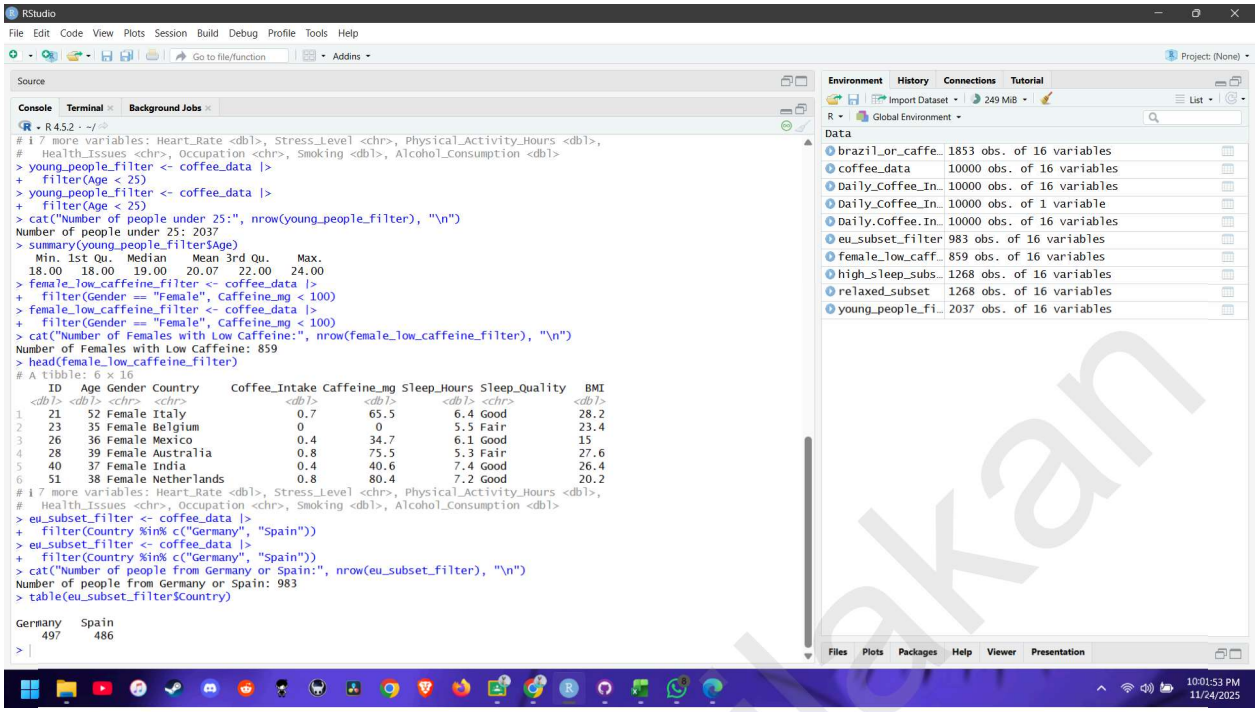
```
# ℹ 7 more variables: Heart_Rate <dbl>, Stress_Level <chr>, Physical_Activity_Hours <dbl>,
#   Health_Issues <chr>, Occupation <chr>, Smoking <dbl>, Alcohol_Consumption <dbl>
> brazil_or_caffeine_subset <- subset(coffee_data, Country == "Brazil" | Coffee_Intake > 4)
> cat("Number of Brazil OR High Coffee Intake:", nrow(brazil_or_caffeine_subset), "\n")
Number of Brazil OR High Coffee Intake: 1853
> head(brazil_or_caffeine_subset)
# A tibble: 6 × 16
     ID   Age Gender Country     Coffee_Intake Caffeine_mg Sleep_Hours Sleep_Quality   BMI
  <dbl> <dbl> <chr>  <chr>               <dbl>       <dbl>       <dbl> <chr>         <dbl>
1     3    42 Male   Brazil              5.3         504.         5.9 Fair           22.7
2     8    44 Female Canada              4.5         424.         5.5 Fair           15.8
3    11    29 Female Switzerland         4.5         428.         8.1 Excellent      21.5
4    12    29 Male   Netherlands         4.1         385.         6.5 Good           23.1
5    13    37 Female Italy               4.7         442.         6.9 Good           15.4
6    19    24 Female UK                  4.3         410.         7.8 Good           22.7
# ℹ 7 more variables: Heart_Rate <dbl>, Stress_Level <chr>, Physical_Activity_Hours <dbl>,
#   Health_Issues <chr>, Occupation <chr>, Smoking <dbl>, Alcohol_Consumption <dbl>
> young_people_filter <- coffee_data |>
+   filter(Age < 25)
> young_people_filter <- coffee_data |>
+   filter(Age < 25)
> cat("Number of people under 25:", nrow(young_people_filter), "\n")
Number of people under 25: 2037
> summary(young_people_filter$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   18.00   19.00   20.07   22.00   24.00
> female_low_caffeine_filter <- coffee_data |>
+   filter(Gender == "Female", Caffeine_mg < 100)
> female_low_caffeine_filter <- coffee_data |>
+   filter(Gender == "Female", Caffeine_mg < 100)
> cat("Number of Females with Low Caffeine:", nrow(female_low_caffeine_filter), "\n")
Number of Females with Low Caffeine: 859
> head(female_low_caffeine_filter)
# A tibble: 6 × 16
     ID   Age Gender Country     Coffee_Intake Caffeine_mg Sleep_Hours Sleep_Quality   BMI
  <dbl> <dbl> <chr>  <chr>               <dbl>       <dbl>       <dbl> <chr>         <dbl>
1    21    52 Female Italy               0.7          65.5        6.4 Good           28.2
2    23    35 Female Belgium             0             0          5.5 Fair           23.4
3    26    36 Female Mexico              0.4          34.7        6.1 Good           15
```

| Data | |
|---|---|
| brazil_or_caffe… | 1853 obs. of 16 variables |
| coffee_data | 10000 obs. of 16 variables |
| Daily_Coffee_In… | 10000 obs. of 16 variables |
| Daily_Coffee_In… | 10000 obs. of 1 variable |
| Daily.Coffee.In… | 10000 obs. of 16 variables |
| eu_subset_filter | 983 obs. of 16 variables |
| female_low_caff… | 859 obs. of 16 variables |
| high_sleep_subs… | 1268 obs. of 16 variables |
| relaxed_subset | 1268 obs. of 16 variables |
| young_people_fi… | 2037 obs. of 16 variables |



Omith Thilakan
S097

SHETH LUJ AND SIR MV COLLEGE
Subject: Data Analysis with SAS / SPSS /R






Omith Thilakan
S097