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

Practical No: 9

Aim: Performing text manipulation using str_sub(), str_split() (R). import dataset.

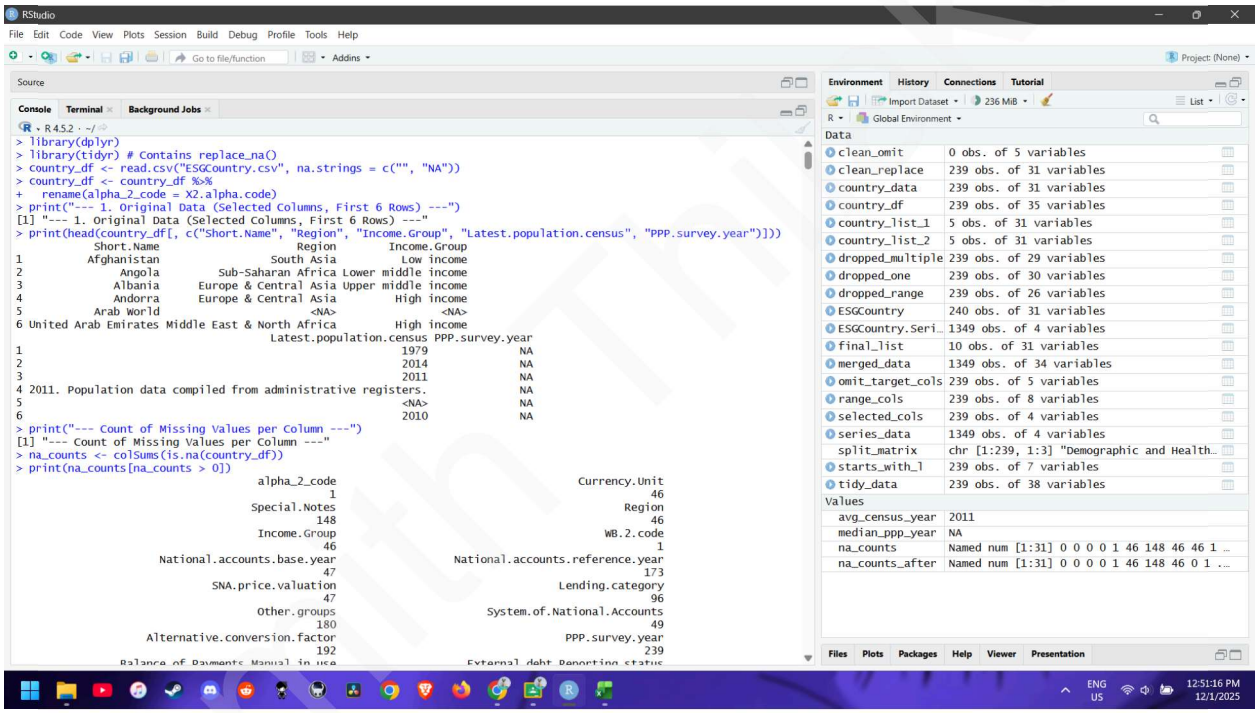
Code:

```
# R Script: Text Manipulation with stringr (Adapted for ESGCountry.csv)
library(stringr)
library(tidyr)
library(dplyr)

# 1. IMPORT DATASET

country_df <- read.csv("ESGCountry.csv")

country_df <- country_df %>%
  rename(alpha_2_code = X2.alpha.code)

print("--- Original Dataset (Key Columns) ---")
print(head(country_df[, c("Country.Code", "Long.Name", "Latest.household.survey")], 5))


# 2. USING str_sub() (Substring)

country_df$Prefix_5 <- str_sub(country_df$Long.Name, 1, 5)

country_df$Suffix_5 <- str_sub(country_df$Long.Name, -5, -1)

print("--- Data after str_sub() ---")
print(country_df %>% select(Long.Name, Prefix_5, Suffix_5) %>% head(5))


# 3. USING str_split() (Split String)

# Method B: Split Fixed (Returns a matrix, easier to assign to columns)
split_matrix <- str_split(country_df$Latest.household.survey, ", ", simplify = TRUE)

country_df$Survey_Type <- split_matrix[, 1] # Text before the comma
country_df$Survey_Detail <- split_matrix[, 2] # Text after the comma

print("--- Data after str_split() (Manual Assignment) ---")
print(country_df %>% select(Latest.household.survey, Survey_Type, Survey_Detail) %>%
head(5))
```

Omith Thilakan
S097

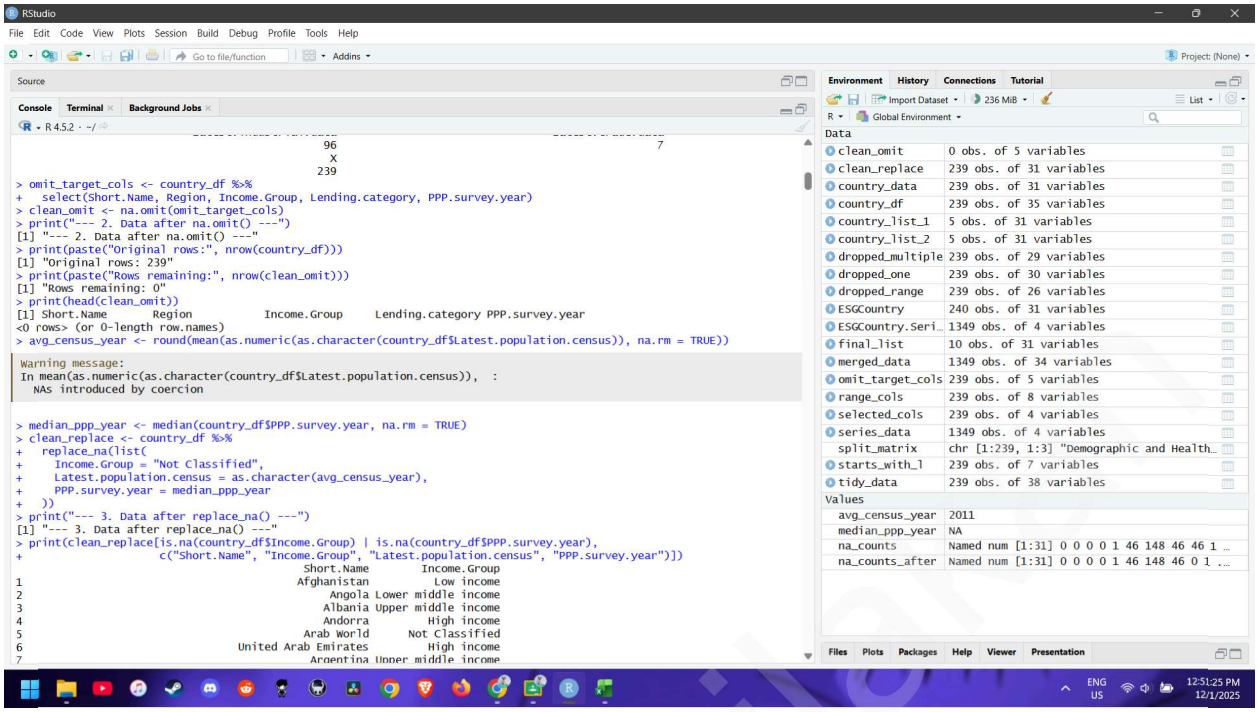
```
# 4. BONUS: The "Tidy" Way (separate)

tidy_data <- country_df %>%
  separate(Long.Name,
           into = c("Status_1", "Status_2", "Rest_of_Name"),
           sep = " ",
           extra = "merge", # Merges all remaining parts into the last column
           remove = FALSE) # Keep the original Long.Name column

print("--- Bonus: The 'separate' function (easier splitting) ---")
print(tidy_data %>% select(Long.Name, Status_1, Status_2, Rest_of_Name) %>% head(5))
```

Output:

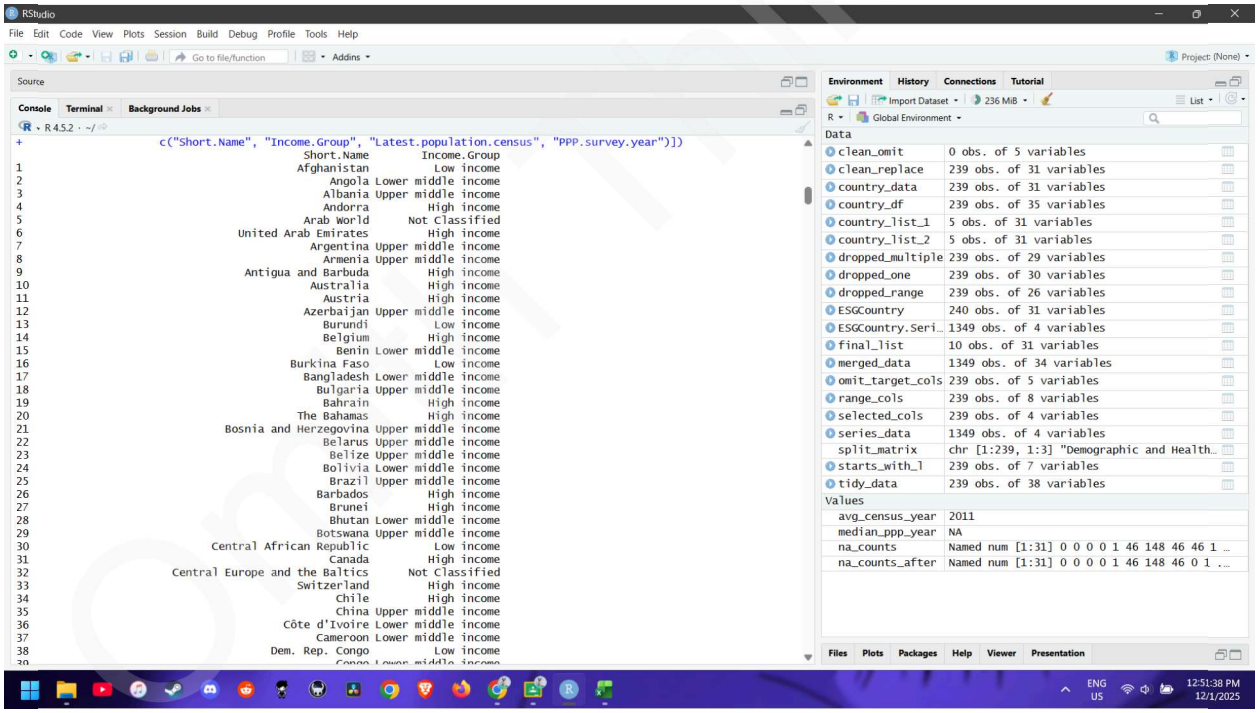


Omith Thilakan
S097

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



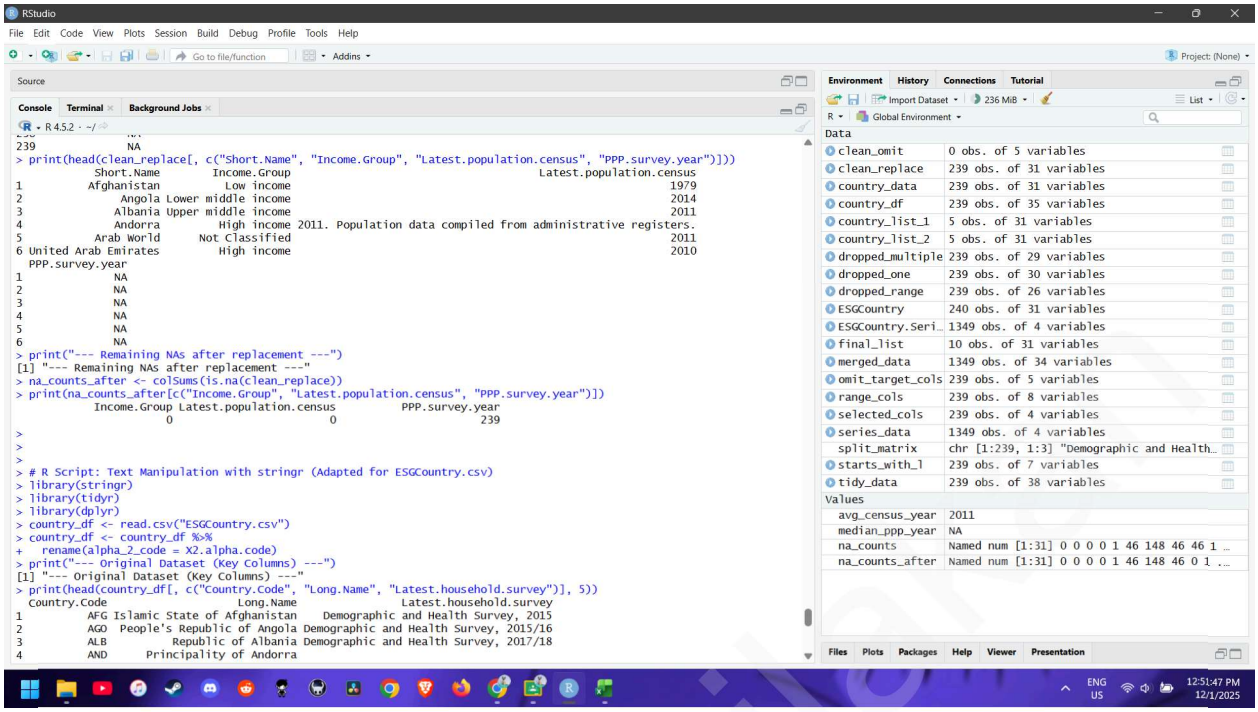
```
                                        96                                        7
                                         X
                                       239
> omit_target_cols <- country_df %>%
+   select(Short.Name, Region, Income.Group, Lending.category, PPP.survey.year)
> clean_omit <- na.omit(omit_target_cols)
> print("--- 2. Data after na.omit() ---")
[1] "--- 2. Data after na.omit() ---"
> print(paste("Original rows:", nrow(country_df)))
[1] "Original rows: 239"
> print(paste("Rows remaining:", nrow(clean_omit)))
[1] "Rows remaining: 0"
> print(head(clean_omit))
[1] Short.Name       Region           Income.Group     Lending.category PPP.survey.year
<0 rows> (or 0-length row.names)
> avg_census_year <- round(mean(as.numeric(as.character(country_df$Latest.population.census)), na.rm = TRUE))
Warning message:
In mean(as.numeric(as.character(country_df$Latest.population.census)),  :
  NAs introduced by coercion

> median_ppp_year <- median(country_df$PPP.survey.year, na.rm = TRUE)
> clean_replace <- country_df %>%
+   replace_na(list(
+     Income.Group = "Not Classified",
+     Latest.population.census = as.character(avg_census_year),
+     PPP.survey.year = median_ppp_year
+   ))
> print("--- 3. Data after replace_na() ---")
[1] "--- 3. Data after replace_na() ---"
> print(clean_replace[is.na(country_df$Income.Group) | is.na(country_df$PPP.survey.year),
+                     c("Short.Name", "Income.Group", "Latest.population.census", "PPP.survey.year")])
                                  Short.Name        Income.Group
1                                Afghanistan          Low income
2                                     Angola Lower middle income
3                                    Albania Upper middle income
4                                    Andorra         High income
5                                 Arab World      Not Classified
6                      United Arab Emirates         High income
7                                  Argentina Upper middle income
```

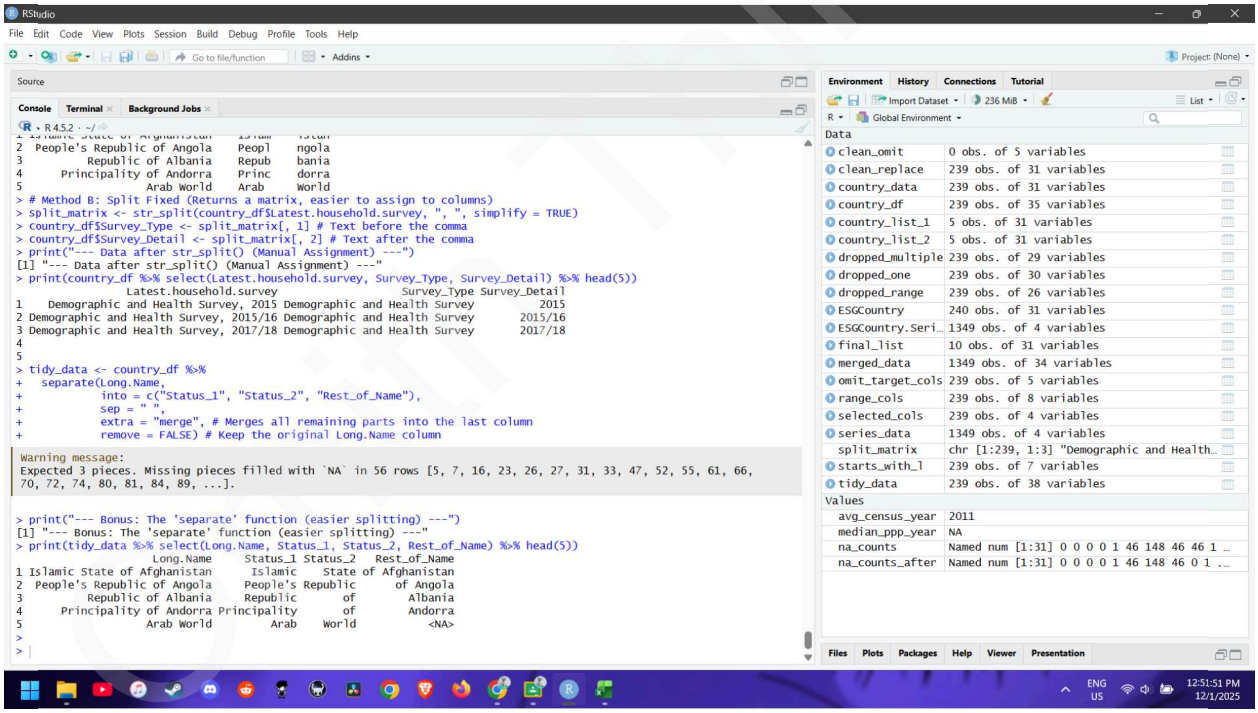


```
+                     c("Short.Name", "Income.Group", "Latest.population.census", "PPP.survey.year")])
                                  Short.Name        Income.Group
1                                Afghanistan          Low income
2                                     Angola Lower middle income
3                                    Albania Upper middle income
4                                    Andorra         High income
5                                 Arab World      Not Classified
6                      United Arab Emirates         High income
7                                  Argentina Upper middle income
8                                    Armenia Upper middle income
9                          Antigua and Barbuda         High income
10                                 Australia         High income
11                                   Austria         High income
12                                Azerbaijan Upper middle income
13                                   Burundi          Low income
14                                   Belgium         High income
15                                     Benin Lower middle income
16                              Burkina Faso          Low income
17                                Bangladesh Lower middle income
18                                  Bulgaria Upper middle income
19                                   Bahrain         High income
20                                The Bahamas         High income
21                    Bosnia and Herzegovina Upper middle income
22                                   Belarus Upper middle income
23                                    Belize Upper middle income
24                                   Bolivia Lower middle income
25                                    Brazil Upper middle income
26                                  Barbados         High income
27                                    Brunei         High income
28                                    Bhutan Lower middle income
29                                  Botswana Upper middle income
30                  Central African Republic          Low income
31                                    Canada         High income
32          Central Europe and the Baltics      Not Classified
33                               Switzerland         High income
34                                     Chile         High income
35                                     China Upper middle income
36                           Côte d'Ivoire Lower middle income
37                                  Cameroon Lower middle income
38                          Dem. Rep. Congo          Low income
```

Omith Thilakan
S097

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





Omith Thilakan
S097