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

Practical No: 8

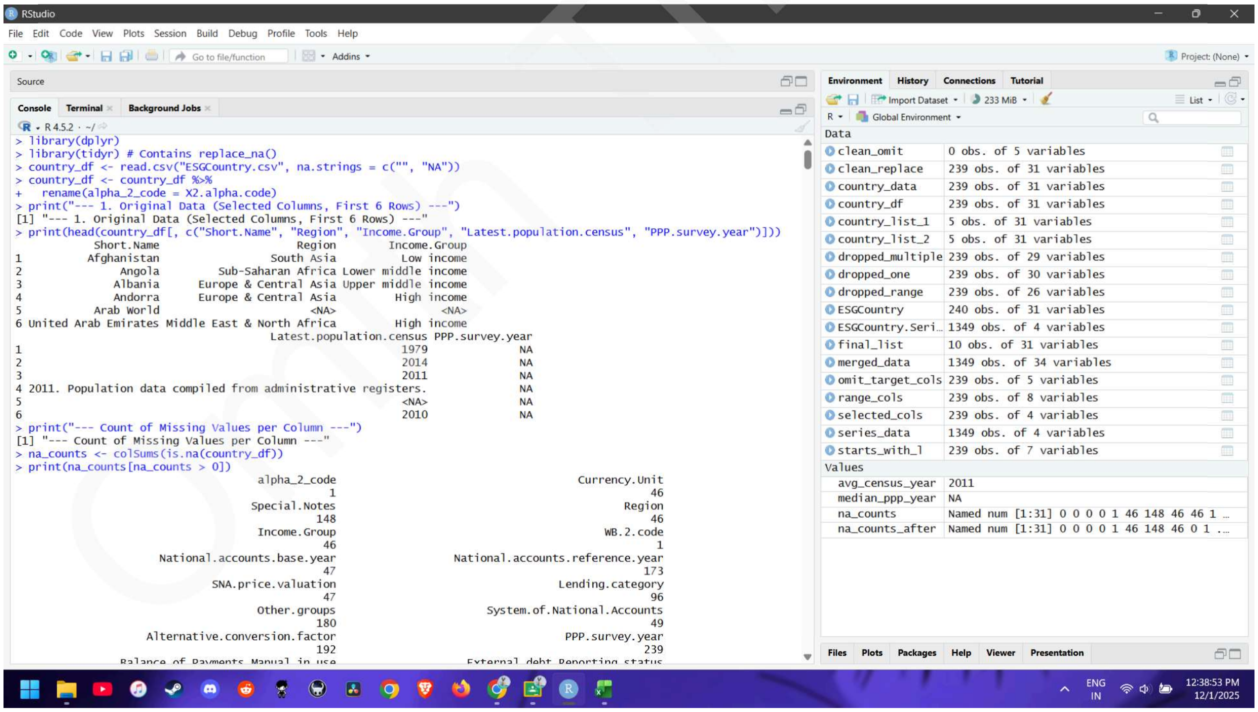
Aim: Applying basic data cleaning functions: handling missing values using na.omit()/replace_na() in R. import dataset.

Code:

```
# R Script: Handling Missing Values (Data Cleaning)

library(dplyr)
library(tidyr) # Contains replace_na()

# 1. IMPORT DATASET

country_df <- read.csv("ESGCountry.csv", na.strings = c("", "NA"))

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

print("--- 1. Original Data (Selected Columns, First 6 Rows) ---")
print(head(country_df[, c("Short.Name", "Region", "Income.Group", "Latest.population.census",
"PPP.survey.year")]))

print("--- Count of Missing Values per Column ---")
na_counts <- colSums(is.na(country_df))
print(na_counts[na_counts > 0])

# 2. METHOD A: REMOVE MISSING VALUES (na.omit)

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() ---")
print(paste("Original rows:", nrow(country_df)))
print(paste("Rows remaining:", nrow(clean_omit)))
print(head(clean_omit))


# 3. METHOD B: REPLACE MISSING VALUES (replace_na)

avg_census_year <-
round(mean(as.numeric(as.character(country_df$Latest.population.census)), na.rm = TRUE))
```

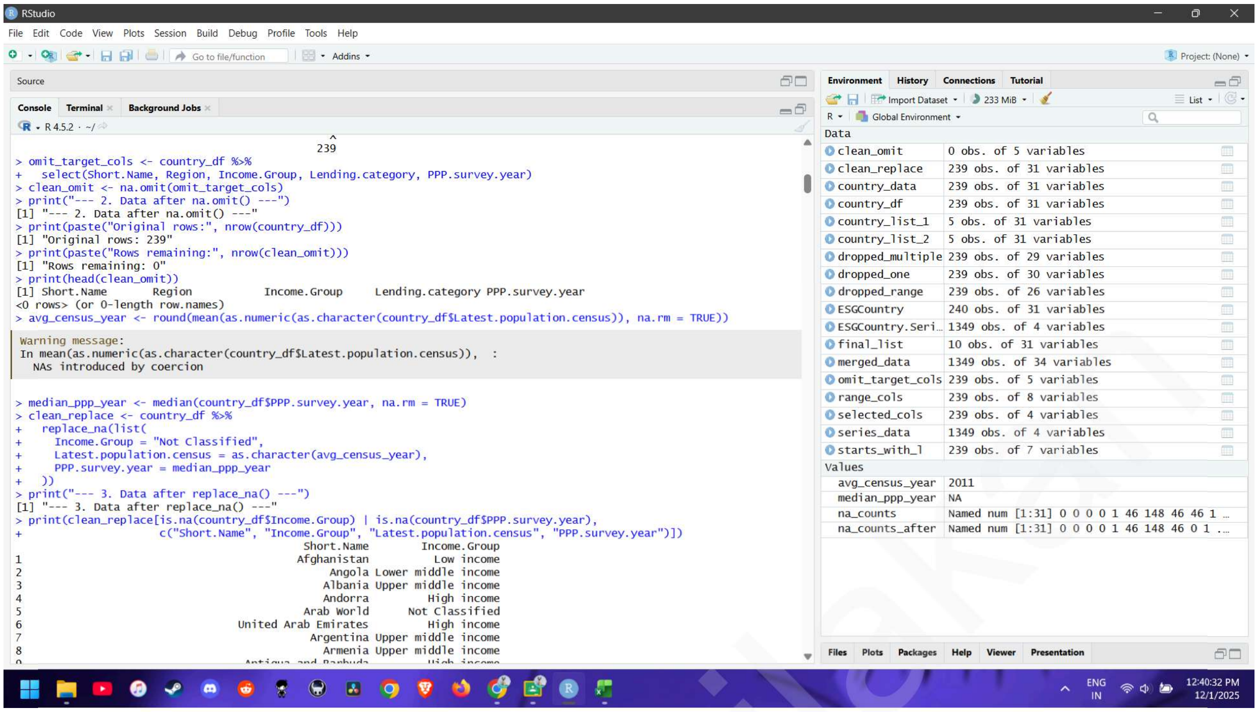
```
median_ppp_year <- median(country_df$PPP.survey.year, na.rm = TRUE)

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() ---")
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")])
print(head(clean_replace[, c("Short.Name", "Income.Group", "Latest.population.census",
"PPP.survey.year")]))

print("--- Remaining NAs after replacement ---")
na_counts_after <- colSums(is.na(clean_replace))
print(na_counts_after[c("Income.Group", "Latest.population.census", "PPP.survey.year")])
```

Output:



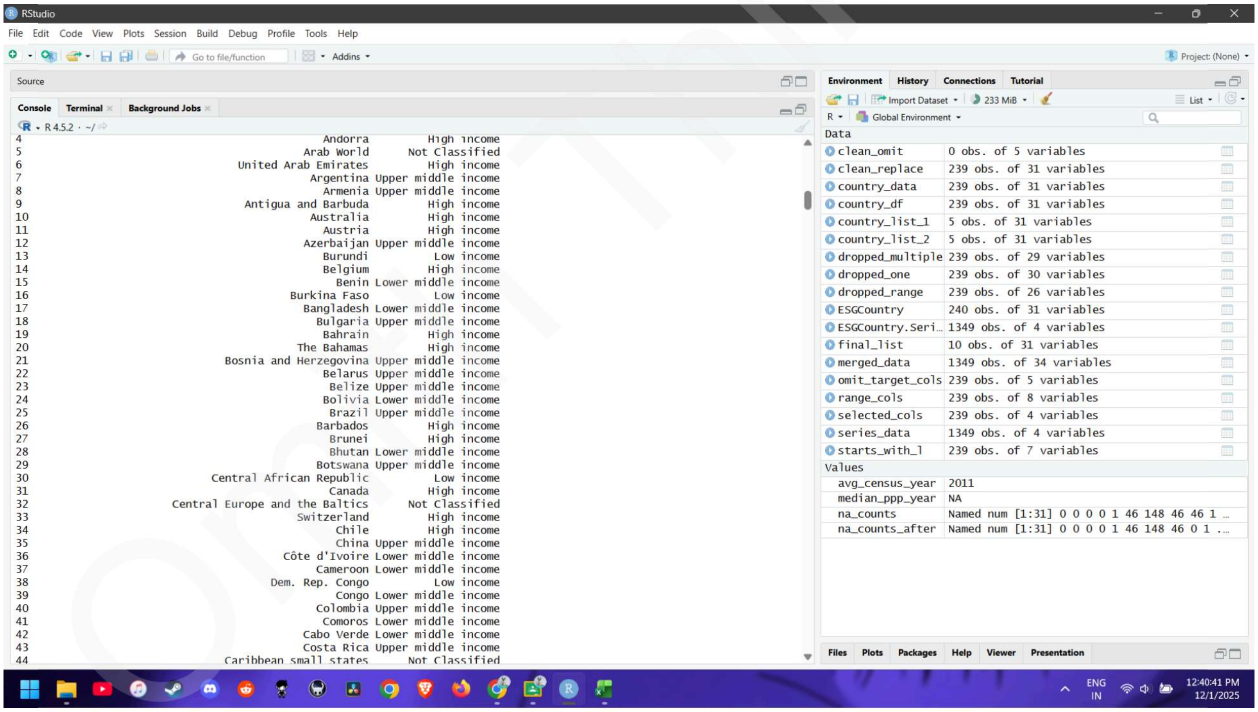
```
> library(dplyr)
> library(tidyr) # Contains replace_na()
> country_df <- read.csv("ESGCountry.csv", na.strings = c("", "NA"))
> country_df <- country_df %>%
+   rename(alpha_2_code = X2.alpha.code)
> print("--- 1. Original Data (Selected Columns, First 6 Rows) ---")
[1] "--- 1. Original Data (Selected Columns, First 6 Rows) ---"
> print(head(country_df[, c("Short.Name", "Region", "Income.Group", "Latest.population.census", "PPP.survey.year")]))
            Short.Name                     Region         Income.Group
1          Afghanistan                 South Asia           Low income
2               Angola         Sub-Saharan Africa  Lower middle income
3              Albania      Europe & Central Asia  Upper middle income
4              Andorra      Europe & Central Asia          High income
5           Arab World                       <NA>                 <NA>
6 United Arab Emirates Middle East & North Africa          High income
                         Latest.population.census PPP.survey.year
1                                            1979              NA
2                                            2014              NA
3                                            2011              NA
4 2011. Population data compiled from administrative registers.  NA
5                                            <NA>              NA
6                                            2010              NA
> print("--- Count of Missing Values per Column ---")
[1] "--- Count of Missing Values per Column ---"
> na_counts <- colSums(is.na(country_df))
> print(na_counts[na_counts > 0])
                     alpha_2_code                          Currency.Unit
                                1                                     46
                    Special.Notes                                 Region
                              148                                     46
                     Income.Group                              WB.2.code
                               46                                      1
          National.accounts.base.year     National.accounts.reference.year
                               47                                    173
                SNA.price.valuation                       Lending.category
                               47                                     96
                     Other.groups               System.of.National.Accounts
                              180                                     49
      Alternative.conversion.factor                        PPP.survey.year
                              192                                    239
    Balance.of.Payments.Manual.in.use             External.debt.Reporting.status
```

Data	
clean_omit	0 obs. of 5 variables
clean_replace	239 obs. of 31 variables
country_data	239 obs. of 31 variables
country_df	239 obs. of 31 variables
country_list_1	5 obs. of 31 variables
country_list_2	5 obs. of 31 variables
dropped_multiple	239 obs. of 29 variables
dropped_one	239 obs. of 30 variables
dropped_range	239 obs. of 26 variables
ESGCountry	240 obs. of 31 variables
ESGCountry.Seri…	1349 obs. of 4 variables
final_list	10 obs. of 31 variables
merged_data	1349 obs. of 34 variables
omit_target_cols	239 obs. of 5 variables
range_cols	239 obs. of 8 variables
selected_cols	239 obs. of 4 variables
series_data	1349 obs. of 4 variables
starts_with_l	239 obs. of 7 variables
Values	
avg_census_year	2011
median_ppp_year	NA
na_counts	Named num [1:31] 0 0 0 0 1 46 148 46 46 1 …
na_counts_after	Named num [1:31] 0 0 0 0 1 46 148 46 0 1 …



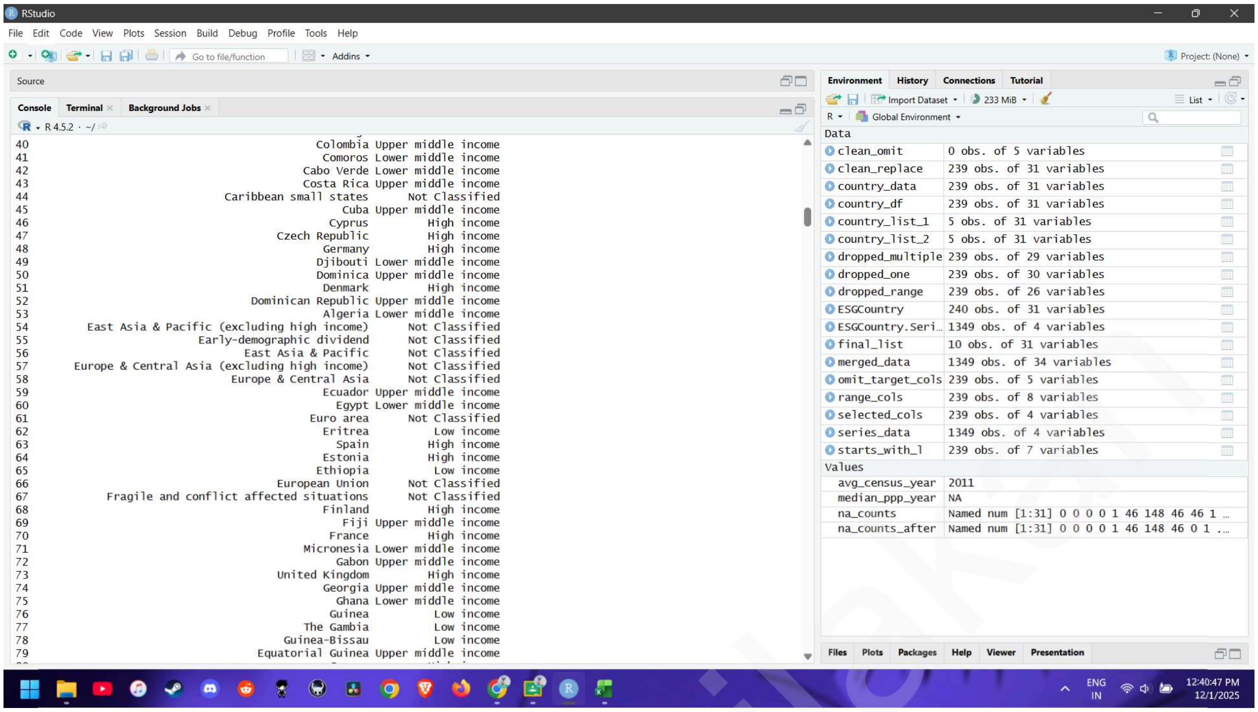
```
                                                   ^
                                                  239
> omit_target_cols <- country_df %>%
+   select(Short.Name, Region, Income.Group, Lending.category, PPP.survey.year)
> clean_omit <- na.omit(omit_target_cols)
> print("--- 2. Data after na.omit() ---")
[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
8                                            Armenia Upper middle income
9                                Antigua and Barbuda         High income
```

```
Data
clean_omit          0 obs. of 5 variables
clean_replace       239 obs. of 31 variables
country_data        239 obs. of 31 variables
country_df          239 obs. of 31 variables
country_list_1      5 obs. of 31 variables
country_list_2      5 obs. of 31 variables
dropped_multiple    239 obs. of 29 variables
dropped_one         239 obs. of 30 variables
dropped_range       239 obs. of 26 variables
ESGCountry          240 obs. of 31 variables
ESGCountry.Seri…    1349 obs. of 4 variables
final_list          10 obs. of 31 variables
merged_data         1349 obs. of 34 variables
omit_target_cols    239 obs. of 5 variables
range_cols          239 obs. of 8 variables
selected_cols       239 obs. of 4 variables
series_data         1349 obs. of 4 variables
starts_with_l       239 obs. of 7 variables
Values
  avg_census_year   2011
  median_ppp_year   NA
  na_counts         Named num [1:31] 0 0 0 0 1 46 148 46 46 1 …
  na_counts_after   Named num [1:31] 0 0 0 0 1 46 148 46 0 1 .…
```



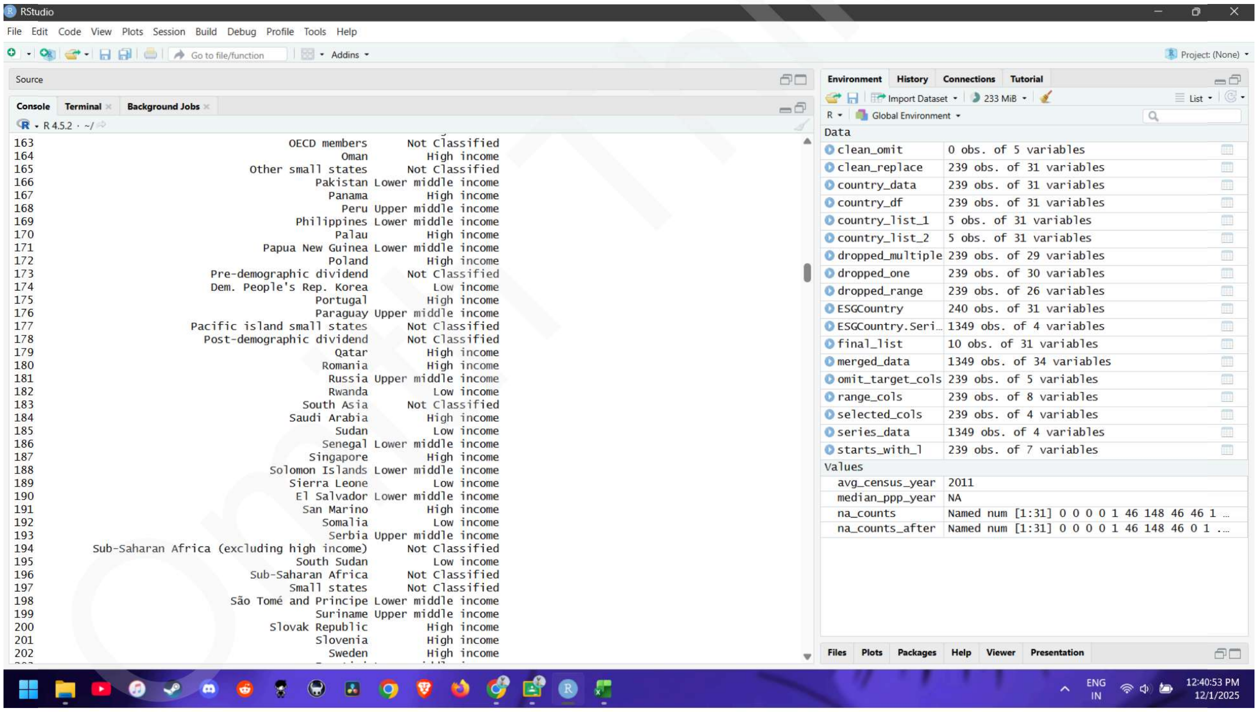
```
4                                            Andorra         High income
5                                         Arab World      Not Classified
6                               United Arab Emirates         High income
7                                          Argentina Upper middle income
8                                            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
39                                             Congo Lower middle income
40                                          Colombia Upper middle income
41                                           Comoros Lower middle income
42                                        Cabo Verde Lower middle income
43                                        Costa Rica Upper middle income
44                           Caribbean small states      Not Classified
```

```
Data
clean_omit          0 obs. of 5 variables
clean_replace       239 obs. of 31 variables
country_data        239 obs. of 31 variables
country_df          239 obs. of 31 variables
country_list_1      5 obs. of 31 variables
country_list_2      5 obs. of 31 variables
dropped_multiple    239 obs. of 29 variables
dropped_one         239 obs. of 30 variables
dropped_range       239 obs. of 26 variables
ESGCountry          240 obs. of 31 variables
ESGCountry.Seri…    1349 obs. of 4 variables
final_list          10 obs. of 31 variables
merged_data         1349 obs. of 34 variables
omit_target_cols    239 obs. of 5 variables
range_cols          239 obs. of 8 variables
selected_cols       239 obs. of 4 variables
series_data         1349 obs. of 4 variables
starts_with_l       239 obs. of 7 variables
Values
  avg_census_year   2011
  median_ppp_year   NA
  na_counts         Named num [1:31] 0 0 0 0 1 46 148 46 46 1 …
  na_counts_after   Named num [1:31] 0 0 0 0 1 46 148 46 0 1 .…
```



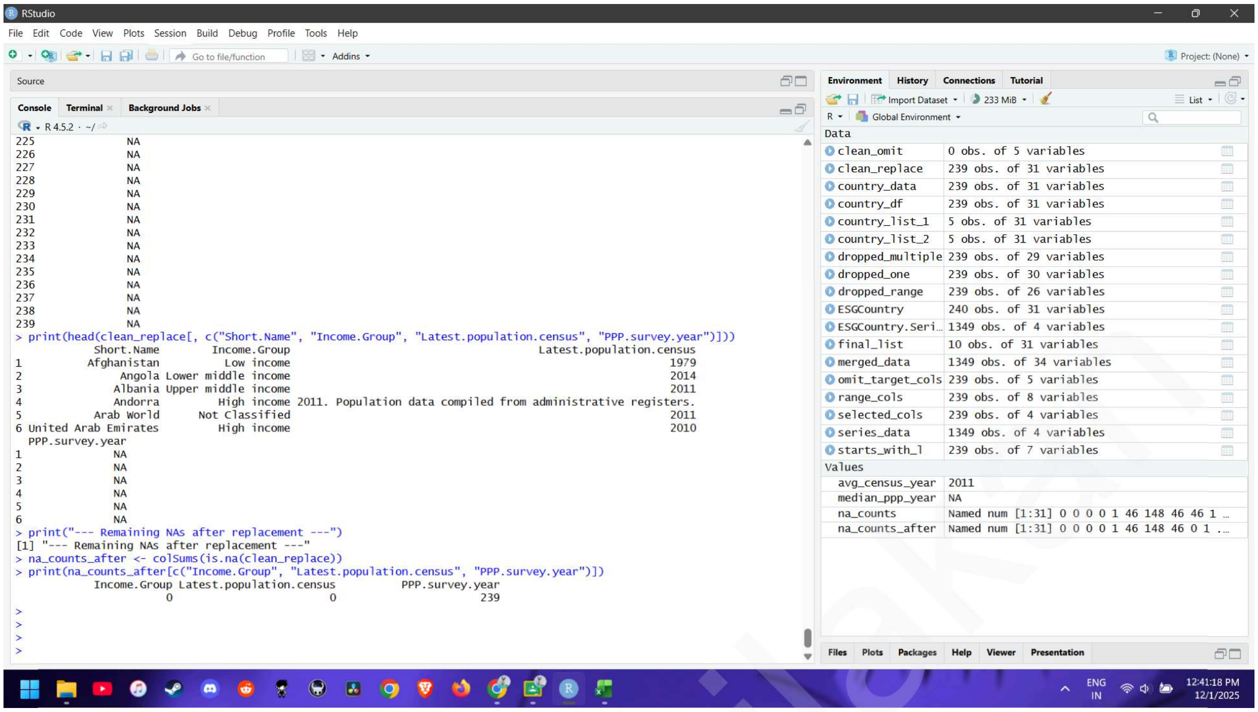
```
40                                       Colombia Upper middle income
41                                        Comoros Lower middle income
42                                     Cabo Verde Lower middle income
43                                     Costa Rica Upper middle income
44                         Caribbean small states      Not Classified
45                                           Cuba Upper middle income
46                                         Cyprus         High income
47                                 Czech Republic         High income
48                                        Germany         High income
49                                       Djibouti Lower middle income
50                                       Dominica Upper middle income
51                                        Denmark         High income
52                             Dominican Republic Upper middle income
53                                        Algeria Lower middle income
54    East Asia & Pacific (excluding high income)      Not Classified
55                     Early-demographic dividend      Not Classified
56                            East Asia & Pacific      Not Classified
57  Europe & Central Asia (excluding high income)      Not Classified
58                          Europe & Central Asia      Not Classified
59                                        Ecuador Upper middle income
60                                          Egypt Lower middle income
61                                      Euro area      Not Classified
62                                        Eritrea          Low income
63                                          Spain         High income
64                                        Estonia         High income
65                                       Ethiopia          Low income
66                                 European Union      Not Classified
67       Fragile and conflict affected situations      Not Classified
68                                        Finland         High income
69                                           Fiji Upper middle income
70                                         France         High income
71                                     Micronesia Lower middle income
72                                          Gabon Upper middle income
73                                 United Kingdom         High income
74                                        Georgia Upper middle income
75                                          Ghana Lower middle income
76                                         Guinea          Low income
77                                     The Gambia          Low income
78                                  Guinea-Bissau          Low income
79                              Equatorial Guinea Upper middle income
```

```
Data
clean_omit         0 obs. of 5 variables
clean_replace      239 obs. of 31 variables
country_data       239 obs. of 31 variables
country_df         239 obs. of 31 variables
country_list_1     5 obs. of 31 variables
country_list_2     5 obs. of 31 variables
dropped_multiple   239 obs. of 29 variables
dropped_one        239 obs. of 30 variables
dropped_range      239 obs. of 26 variables
ESGCountry         240 obs. of 31 variables
ESGCountry.Seri…   1349 obs. of 4 variables
final_list         10 obs. of 31 variables
merged_data        1349 obs. of 34 variables
omit_target_cols   239 obs. of 5 variables
range_cols         239 obs. of 8 variables
selected_cols      239 obs. of 4 variables
series_data        1349 obs. of 4 variables
starts_with_l      239 obs. of 7 variables
Values
avg_census_year    2011
median_ppp_year    NA
na_counts          Named num [1:31] 0 0 0 0 1 46 148 46 46 1 …
na_counts_after    Named num [1:31] 0 0 0 0 1 46 148 46 0 1 .…
```



```
163                               OECD members      Not Classified
164                                       Oman         High income
165                         Other small states      Not Classified
166                                   Pakistan Lower middle income
167                                     Panama         High income
168                                       Peru Upper middle income
169                                Philippines Lower middle income
170                                      Palau         High income
171                           Papua New Guinea Lower middle income
172                                     Poland         High income
173                   Pre-demographic dividend      Not Classified
174                   Dem. People's Rep. Korea          Low income
175                                   Portugal         High income
176                                   Paraguay Upper middle income
177                Pacific island small states      Not Classified
178                  Post-demographic dividend      Not Classified
179                                      Qatar         High income
180                                    Romania         High income
181                                     Russia Upper middle income
182                                     Rwanda          Low income
183                                 South Asia      Not Classified
184                               Saudi Arabia         High income
185                                      Sudan          Low income
186                                    Senegal Lower middle income
187                                  Singapore         High income
188                            Solomon Islands Lower middle income
189                               Sierra Leone          Low income
190                                El Salvador Lower middle income
191                                 San Marino         High income
192                                    Somalia          Low income
193                                     Serbia Upper middle income
194 Sub-Saharan Africa (excluding high income)      Not Classified
195                                South Sudan          Low income
196                         Sub-Saharan Africa      Not Classified
197                               Small states      Not Classified
198                      São Tomé and Principe Lower middle income
199                                   Suriname Upper middle income
200                            Slovak Republic         High income
201                                   Slovenia         High income
202                                     Sweden         High income
```

```
Data
clean_omit         0 obs. of 5 variables
clean_replace      239 obs. of 31 variables
country_data       239 obs. of 31 variables
country_df         239 obs. of 31 variables
country_list_1     5 obs. of 31 variables
country_list_2     5 obs. of 31 variables
dropped_multiple   239 obs. of 29 variables
dropped_one        239 obs. of 30 variables
dropped_range      239 obs. of 26 variables
ESGCountry         240 obs. of 31 variables
ESGCountry.Seri…   1349 obs. of 4 variables
final_list         10 obs. of 31 variables
merged_data        1349 obs. of 34 variables
omit_target_cols   239 obs. of 5 variables
range_cols         239 obs. of 8 variables
selected_cols      239 obs. of 4 variables
series_data        1349 obs. of 4 variables
starts_with_l      239 obs. of 7 variables
Values
avg_census_year    2011
median_ppp_year    NA
na_counts          Named num [1:31] 0 0 0 0 1 46 148 46 46 1 …
na_counts_after    Named num [1:31] 0 0 0 0 1 46 148 46 0 1 .…
```



```
225              NA
226              NA
227              NA
228              NA
229              NA
230              NA
231              NA
232              NA
233              NA
234              NA
235              NA
236              NA
237              NA
238              NA
239              NA
> print(head(clean_replace[, c("Short.Name", "Income.Group", "Latest.population.census", "PPP.survey.year")]))
            Short.Name        Income.Group                                         Latest.population.census
1          Afghanistan          Low income                                                             1979
2               Angola Lower middle income                                                             2014
3              Albania Upper middle income                                                             2011
4              Andorra         High income 2011. Population data compiled from administrative registers.
5           Arab World      Not Classified                                                             2011
6 United Arab Emirates         High income                                                             2010
  PPP.survey.year
1              NA
2              NA
3              NA
4              NA
5              NA
6              NA
> print("--- Remaining NAs after replacement ---")
[1] "--- Remaining NAs after replacement ---"
> na_counts_after <- colSums(is.na(clean_replace))
> print(na_counts_after[c("Income.Group", "Latest.population.census", "PPP.survey.year")])
            Income.Group Latest.population.census          PPP.survey.year
                       0                        0                      239
>
>
>
>
```

Data	
clean_omit	0 obs. of 5 variables
clean_replace	239 obs. of 31 variables
country_data	239 obs. of 31 variables
country_df	239 obs. of 31 variables
country_list_1	5 obs. of 31 variables
country_list_2	5 obs. of 31 variables
dropped_multiple	239 obs. of 29 variables
dropped_one	239 obs. of 30 variables
dropped_range	239 obs. of 26 variables
ESGCountry	240 obs. of 31 variables
ESGCountry.Seri…	1349 obs. of 4 variables
final_list	10 obs. of 31 variables
merged_data	1349 obs. of 34 variables
omit_target_cols	239 obs. of 5 variables
range_cols	239 obs. of 8 variables
selected_cols	239 obs. of 4 variables
series_data	1349 obs. of 4 variables
starts_with_l	239 obs. of 7 variables
Values	
avg_census_year	2011
median_ppp_year	NA
na_counts	Named num [1:31] 0 0 0 0 1 46 148 46 46 1 …
na_counts_after	Named num [1:31] 0 0 0 0 1 46 148 46 0 1 .…