Country Analysis

Arin Parsa

7/21/2020

Contents

Purpose

The purpose of this assignment is to learn how to clean and merge datasets, match by column using the GDP and Educational Data of countries. Additionally, the assignment helps to learn applying a function on column grouped by another column. Furthermore, the assignment helps to learn about quantiles and how to use the cut function and breaks.

Load and clean datasets

Gross Domestic Product data for the 190 ranked countries in this data set:

[https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv]

Educational data from this data set:

[https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv]

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(data.table)
```

```
##
```

```
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
gdp <- read.csv("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv")
edu <- read.csv("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv")

## gdp dataset needs to be cleaned. Skip first five rows, load 190 rows
## change column names to match edu table, change Rank to numeric

gdp <- gdp[5:194,]

gdp <- gdp %>% rename(Rank = Gross.domestic.product.2012, CountryCode = X) %>% mutate(Rank = as.numeric
```

Assignment 1

Match the data based on the country shortcode. How many of the IDs match? Sort the data frame in descending order by GDP rank (so United States is last). What is the 13th country in the resulting data frame?

```
merge_df <- merge(gdp, edu, by = "CountryCode")
merge_df <- merge_df %>% arrange(desc(Rank))
names(merge_df)[4] <- "CountryName"
merge_df[13,c("CountryCode", "CountryName")]
```

```
##    CountryCode         CountryName
## 13         KNA St. Kitts and Nevis
```

Assignment 2

What is the average GDP ranking for the "High income: OECD" and "High income: nonOECD" group?

```
merge_df <- merge_df %>% group_by(Income.Group) %>% filter(Income.Group %in% c("High income: OECD", "Hig

tapply(merge_df$Rank, merge_df$Income.Group, mean, na.rm = TRUE)
```

```
## High income: nonOECD    High income: OECD
##             91.91304             32.96667
```

Assignment 3

Cut the GDP ranking into 5 separate quantile groups. Make a table versus Income.Group. How many countries are Lower middle income but among the 38 nations with highest GDP?

```
merge_df <- merge(gdp, edu, by = "CountryCode")
merge_df <- merge_df %>% arrange(desc(Rank))

breaks <- quantile(merge_df[,"Rank"], probs = seq(0, 1, 0.2), na.rm = TRUE)
breaks
```

```
##    0%   20%   40%   60%   80%  100%
##   1.0  38.6  76.2 113.8 152.4 190.0
```

```r
merge_df$quantileGDP <- cut(merge_df$Rank, breaks = breaks)
# Alternative approach:
# mergeDT <- setDT(merge_df)
# mergeDT$quantileGDP <- cut(mergeDT[,Rank], breaks = breaks)

## Finding number of countries that are among on the 38 nations with the highest GDP,
## but are lower middle income
merge_df <- filter(merge_df, Income.Group == "Lower middle income")
nrow(merge_df[merge_df$quantileGDP == "(1,38.6]",])
```

```
## [1] 5
```