**Data Science - Case Study1 R Markdown File**

Nagesh Padiyar

Contents

Introduction 1

Folder Description 1

Directions to run the code 2

Further Work 8

References 8

Introduction

The following documents describes the analysis performed Gross Domestic Product data on 190 countries and the educational data from each of those countries.

The data is downloaded from the following sources:

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

Original data sources:

http://data.worldbank.org/data-catalog/GDP-ranking-table http://data.worldbank.org/data-catalog/ed-stats

• File Names created in the local directories:

getdata\_Fdata\_FEDSTATS\_Country.csv

getdata\_Fdata\_FGDP.csv

• The data is delimited by a comma (,)

• Data description (FEDSTATS\_Country.csv)

1. CountryCode
2. Long Name
3. Income Group
4. Region
5. Lending Category
6. Other Groups
7. Currency Unit
8. Latest Population Census
9. Latest Household Survey
10. Special Notes
11. National Accounts Special Year
12. System Of National Accounts
13. SNA Price Valuation
14. Alternative Conversion Factor
15. PPP Survey Year
16. Balance Of Payments Manual In Use
17. External Debt Reporting Status
18. System Of Trade
19. Government Accounting Concept
20. IMF data dissemination standard
21. Source of most recent and expenditure data
22. Vital Registration Complete
23. Latest Agricultural Census
24. Latest Industrial Data
25. Latest Water Withdrawal Data
26. 2-alpha code
27. WB-2 Code
28. Table Name
29. Short Name

* GDP Data description (FGDP.csv)

1. Ranking
2. Economy
3. Millions of Dollars (GDP)

• Problems with the data –

1. The data had Null values for a lot of fields.

**Folder Description**

The folder Case Study 1 GDP Data analysis contains the following

1. Case\_Study\_1\_Q1.R – This R program merges the data in the two files based on Country ShortCode. It also determines the number of IDs which match
2. Case\_Study\_1\_Q2.R – This R program sorts the data frame in ascending order by GDP (so United States is last). It also determines the 13th country in the resulting data frame.
3. Case\_Study\_1\_Q3.R – This R program calculates the average GDP rankings for the “High income: OECD” and “High income: nonOECD” groups.
4. Case\_Study\_1\_Q4.R – This R program plots the GDP for all the countries. It uses ggplot2 to color the plot by Income Group
5. Case\_Study2\_Q5.R – This R program cuts the GDP ranking into 5 separate quantile.

**Directions to run the code**

Call library (dplyr)

library(dplyr)

Set the Working Directory where you downloaded the .CSV files

setwd("C:/Users/Nagesh Padiyar/SMU Datascience/Homework Assignments/Case Study 1");

Read .csv data file for Country Data

country\_data <- read.csv("getdata\_Fdata\_FEDSTATS\_Country.csv", header = FALSE, sep = ",", na.strings = c("","NA"));

Set the first column name to CountryCode

colnames(country\_data)[1] <- 'CountryCode';

Get sub set of country data where country code is not NULL

country\_data\_sub <- subset(x=country\_data, !is.na(CountryCode));

Get the Head and Summary Information for Country Data Sub

head(country\_data\_sub);

summary(country\_data\_sub);

str(country\_data\_sub);

names(country\_data\_sub);

Read .csv data file for GDP Data

gdp\_data <- read.csv("getdata\_Fdata\_FGDP.csv", header = FALSE, sep = ",", na.strings = c("","NA"));

nrow(gdp\_data);

Set the first column name to CountryCode

colnames(gdp\_data)[1] <- 'CountryCode';

Get sub set of country data where country code is not NULL

gdp\_data\_sub <- subset(x=gdp\_data, !is.na(CountryCode));

Get the Head and Summary Information for GDP Data Sub

names(gdp\_data\_sub);

head(gdp\_data\_sub);

summary(gdp\_data\_sub);

Merged the Country and GDP data files using CountryCode which is common field between the two data sets

merged\_data = merge(country\_data\_sub, gdp\_data\_sub, by.x="CountryCode", by.y="CountryCode", all=TRUE);

Get the Head and Summary information for the Merged Data file

head(merged\_data);

summary(merged\_data);

names(merged\_data);

str(merged\_data);

Edit the merged data file to see it in table format

edit(merged\_data);

Total number of records in the merged data is 239.

Question #2

Sort the data frame in ascending order by GDP (so United States is last). What is the 13th country in the resulting data frame?

Set the Column Name for the Merged Data as GDPRanking

colnames(merged\_data)[32] <- 'GDPRanking';

This is to make sure that we are removing NULL values before sorting

gdpranking\_data\_sub <- subset(x=merged\_data, !is.na(GDPRanking));

Check the number of records after removing NULL values. It is 190.

nrow(gdpranking\_data\_sub)

Use the attach command before using the Order function on GDPRanking

attach(gdpranking\_data\_sub)

This command sorts the data using the column GDPRanking by ascending order

sort\_merged\_data <- merged\_data[order(as.numeric(as.character(GDPRanking))),];

This command sorts the data using the column GDPRanking in descending order

sort\_merged\_data <- merged\_data[order(-as.numeric(as.character(GDPRanking))),];

You can look at the data using the edit command

edit(sort\_merged\_data);

The 13th Country from the top after sorting the data in St. Kitts and Nevis.

Actually Since St. Kitts and Newis and Grenada are both tied at 178 in the GDP Ranking.

Question #3

What are the average GDP rankings for the "High income: OECD" and "High income: nonOECD" groups?

unique(V3.x)

[1] High income: nonOECD Low income Lower middle income Upper middle income High income: OECD Income Group <NA>

Levels: High income: nonOECD High income: OECD Income Group Low income Lower middle income Upper middle income

sort\_merged\_data\_sub <- subset(x=sort\_merged\_data, !is.na(GDPRanking));

sort\_merged\_V3\_X\_NA\_sub <- subset(x=sort\_merged\_data\_sub, !is.na(V3.x));

Number of records with NOT NULL values for V3.x. The answer is 189

nrow(sort\_merged\_V3\_X\_NA\_sub)

with(sort\_merged\_V3\_X\_NA\_sub, mean(as.numeric(as.character(GDPRanking[V3.x == "High income: OECD"]))))

[1] 32.96667

sort\_merged\_V3\_X\_NA\_sub <- subset(x=sort\_merged\_V3\_X\_NA\_sub, !is.na(GDPRanking));

sort\_merged\_V3\_X\_NA\_sub <- subset(x=sort\_merged\_V3\_X\_NA\_sub, !is.na(V3.x));

with(sort\_merged\_V3\_X\_NA\_sub, mean(as.numeric(as.character(GDPRanking[V3.x == "High income: nonOECD"]))))

[1] 91.91304

# Question #4

# Plot the GDP for all of the countries. Use ggplot2 to color your plot by Income Group.

library(dplyr);

library(ggplot2);

install.packages(dplyr);

g <- ggplot(data=sort\_merged\_V3\_X\_NA\_sub,aes(x=V3.x,y=GDPRanking)) + geom\_point(color="firebrick");

g;

The output of the ggplot2 is as follows:



The PDF file is as follows:

