W203 : Homework 5 : R Code and Answers to Questions

Natarajan Shankar

June 8, 2016

Natarajan Shankar

Homework for W203

June 8, 2016

```
#PART 2A
# include the Excel capability, without it this code will not work
library(xlsx)
```

```
## Loading required package: rJava
## Loading required package: xlsxjars
```

```
# Problem 10
# Load the data found in the file GDP_World_Bank.csv into the local directory
# read the data into a new data frame, gdb_data
gdp_data <- read.csv("GDP_World_Bank.csv", header = TRUE)

# Create a new variable, gdp_growth, that equals the nominal increase in GDP from 2011 to 2012
gdp_data$gdp_growth <- gdp_data$gdp2012 - gdp_data$gdp2011


#PART 2A Problem 10 Question: What is the mean of your new variable?
# Compute the mean of the new variable, filter out all the NA
growth_mean <- mean(gdp_data$gdp_growth, na.rm = TRUE)
growth_mean
```

```
## [1] 7172376796
```

Problem 10 Question: What is the mean of your new variable?

Mean GDP growth in $

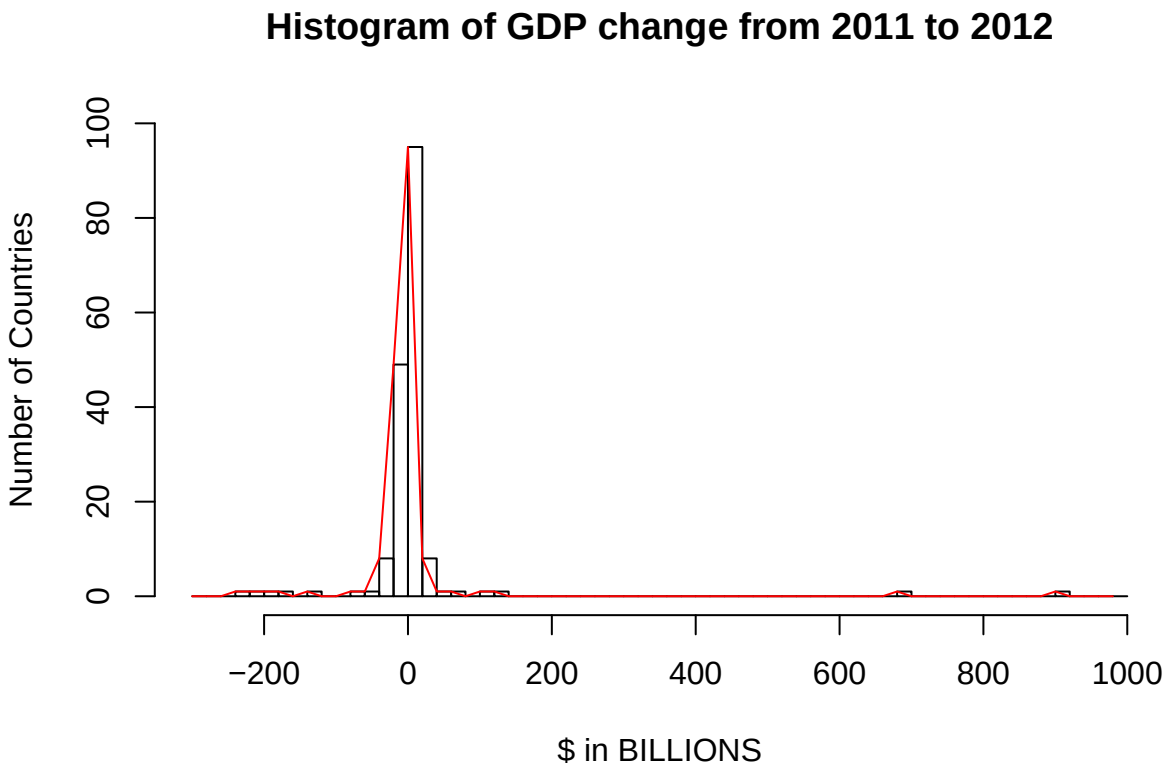
$7,172,376,796 ($7.17B)

```
# Convenience work
# Scale the gdp growth value by Billions so that it can be plotted
ONE_BILLION <- 1.0E+9
```

```r
gdp_data$gdp_growth_scaled  <- gdp_data$gdp_growth/ONE_BILLION

# for convenience, create a local vector with all the NAs removed
gdp_growth_no_NA <- na.omit(gdp_data$gdp_growth_scaled)

# Determine the lowest and the highest values of gdp_growth and round them off to the nearest 100
# The digits = -2 does the rounding
gdp_growth_lowend <- round(min(gdp_growth_no_NA), digits = -2)
gdp_growth_highend <- round(max(gdp_growth_no_NA), digits = -2)

# The rounding to the nearest 100 cuts off outliers at both ends.
# Ensure that outliers are included when plotting by ensuring that the axes are extended to cover the o
# and break the gdp_growth data into $20 Billion increments
breaks = seq(gdp_growth_lowend-100, gdp_growth_highend + 100, by=20)

# Part 2A
# Problem 11
# Create a histogram of of your new variable, gdp_grwoth
hist_data = hist(gdp_growth_no_NA, breaks= breaks, main = "Histogram of GDP change from 2011 to 2012",
                 ylab ="Number of Countries", xlab = "$ in BILLIONS", ylim=c(0,100), xlim=c(-300, 1000))

summary(gdp_growth_no_NA)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -230.0000   -0.3482    0.2017    7.1720    3.2830  910.0000
```

```r
#Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
#-230.0000   -0.3482    0.2017    7.1720    3.2830  910.0000

# Overlay a continuous plot on the histogram so be able to get a view of
# what distribution the data shows
length(hist_data$breaks) <- length(hist_data$breaks) -1
lines(hist_data$breaks, hist_data$counts, col="red")
```



Problem 11 Question: Is the data of the new variable Normally distributed? Describe its shape.

1. The data is not Normally distributed
2. Plot is positive skewed
3. Plot does not show a Normal distribution (does not look identical on both sides of the Mean)
4. Displays a positive kurtosis (leptokurtic)

```
# Part 2A
# Problem 12
# Create a new Boolean variable that equals TRUE if a contry's GDP growth is higher than the Mean
gdp_data$high_growth <- ifelse(gdp_data$gdp_growth > growth_mean, TRUE, FALSE)

# Part 2A
# Problem 12: Question: How many Countries have above average growth?
nrow(na.omit(gdp_data[gdp_data$high_growth == TRUE,]))
```

```
## [1] 31
```

Problem 12: Question: How many Countries have above average growth?

[1] 31 Countries have above average growth

```
# Problem 12: Question: How many Countries have below average growth?
nrow(na.omit(gdp_data[gdp_data$high_growth == FALSE,]))
```

```
## [1] 142
```

Problem 12: Question: How many Countries have below average growth?

[1] 142 Countries have below average growth

Problem 12: Question: Explain the result in terms of the shape of the gdp_growth distribution?

1. There is a cluster of frequent scores at the left side of the distribution and the frequency tails off on the right side.
2. Hence there is a positive skew. Mathematically, the frequency number above the Mean is much less than the frequency number below the Mean.
3. Also, the outliers on the right add to the long tail

Problem 13

Find one new data set

Source : https://www.worlddata.info/downloads/

Countries.csv file downloadable at

https://www.worlddata.info/downloads/

```
country_data <- read.csv("countries.csv", header = TRUE, sep = ";")

# Ensure that gdp_data abd country_data have the same name of the sorting column
colnames(country_data)[1] <- "Country"

# Keep track of how many rows each file has. We'll merge the smaller file into the larger one.
# Contry Data has more rows than gdp_data
nrow(country_data)
```

```
## [1] 247
```

```
# [1] 247
nrow(gdp_data)
```

```
## [1] 212
```

```
# [1] 212

# Create a new file by merging both files
enhanced_country_data <- merge(country_data, gdp_data, by = "Country", all = TRUE)
nrow(enhanced_country_data)
```

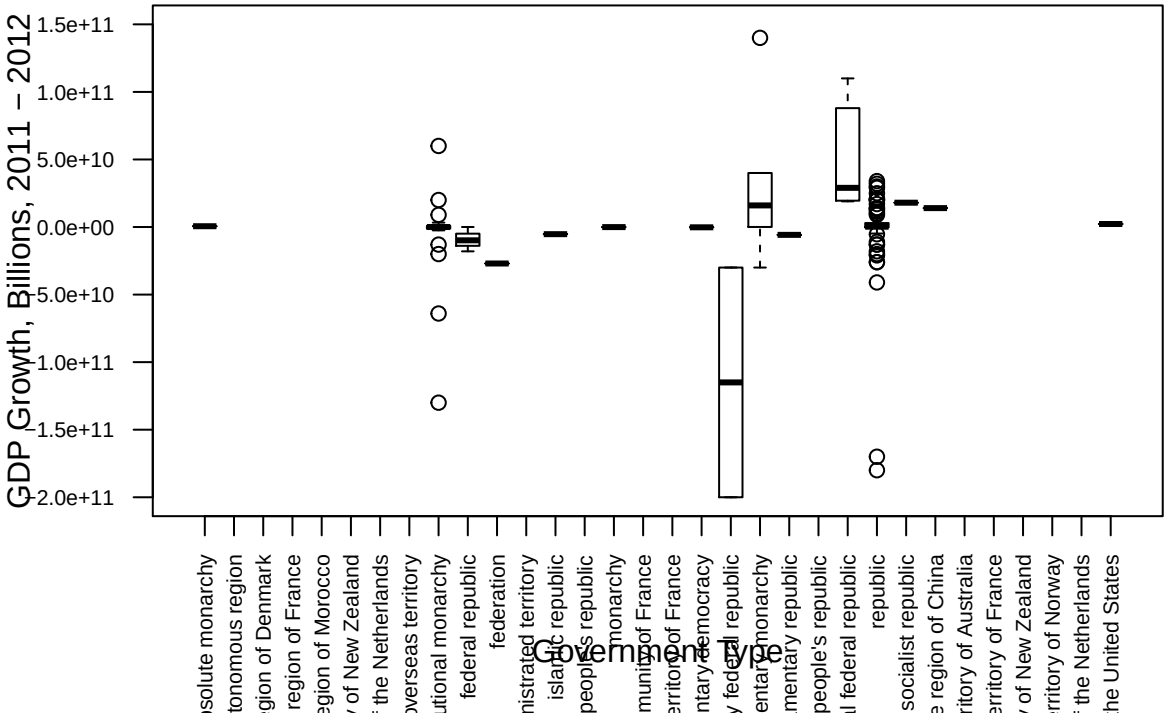
```
## [1] 269
```

```
#[1] 269 - Why does the new file have 269 rows? Should only have 247 with perfect merge

# look at differences between the files
setdiff(enhanced_country_data$Country, country_data$Country)
```

```
##  [1] "Brunei Darussalam"              "Channel Islands"
##  [3] "Congo Dem. Rep."                "Congo Rep."
##  [5] "Cote d Ivoire"                  "Faeroe Islands"
##  [7] "Korea Dem. Rep."                "Korea Rep."
##  [9] "Kyrgyz Republic"                "Lao PDR"
## [11] "Macao"                          "Micronesia"
## [13] "Myanmar"                        "Russian Federation"
## [15] "Sint Maarten (Dutch part)"      "Slovak Republic"
## [17] "St. Kitts and Nevis"            "St. Lucia"
## [19] "St. Martin (French part)"       "St. Vincent and the Grenadines"
## [21] "Syrian Arab Republic"           "Yemen, Rep."
```

```
# Move the data into an excel file so that it can be edited and cleaned up
write.xlsx(x = enhanced_country_data, file = "enhanced_country_data.xlsx", sheetName = "TestSheet", row

# File has been cleaned up, read it back in
merged_country_data <- read.xlsx("enhanced_data_merged_fixed.xlsx", sheetIndex=1)

# MERGE SUCCESSFULLY COMPLETED
# Total known countries = 247
# 3 new countries added at the bottom of file (Channel Islands, Faeroe Islands and Macao)


# Check whether Goovenment type has a bearing on GDP Growth
plot(merged_country_data$Government, merged_country_data$gdp_growth,  las=2, ylim=c(-2e+11, 1.5e+11), c
```

Linkage between Govenment Type and GDP Growth, 2011 − 2012



The graph shows that:

1. Economies in democracies and federal republics show wide bands of GDP growth
2. The GDP range with Constitutional monarchies mimics the wide range shown by Democratic and Federal Republic givernments
3. The coomunist and the Socialist govetrnemnts show stability in economy with neither a large growth nor a large fall