Infant Mortality Rate Around the World

Jinsol Kim 10/31/2017

Introduction

This post intends to perform and learn more about data analysis with R. One of the best way to utilize data analytics skills seemed to be finding a way to contribute to the society. One of the social problems that this post wishes to cover is the infant mortality. Many worldly organizations have been making efforts to find a way to allocate their resources to effectively lower the infant mortality; however, the problem still remains as a major issue. To find out the best allocation of aid, infant mortality rate from the United Nations Population Division's *World Population Prospects* is examined in different perspectives.

Preparation

The packages to be used to import, display, and analyze the data are installed.

```
library(readr)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2

##
## 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(ggplot2)
```

Data

Importing Data

The raw data file for this assignment is downloaded from the World Bank website and are included in the `data/` folder along with data dictionary file.

```
# import the data API_SP.DYN.IMRT.IN_DS2_en_csv_v2.csv
index <- read_csv('data/API_SP.DYN.IMRT.IN_DS2_en_csv_v2.csv',
                  skip = 4,
                  col_types = cols(
                    CountryName = col_character(),
                    CountryCode = col_character(),
                    IndicatorName = col_character(),
                    IndicatorCode = col_character(),
                    Year = col_integer(),
                    Indicator = col_double()
                   )
)

# import the data Metadata_Country_API_SP.DYN.IMRT.IN_DS2_en_csv_v2.csv
metadata <- read_csv('data/Metadata_Country_API_SP.DYN.IMRT.IN_DS2_en_csv_v2.csv',
                  col_types = cols(
                    CountryCode = col_character(),
                    Region = col_character(),
                    IncomeGroup = col_character(),
                    SpecialNotes = col_character(),
                    TableName = col_character()
                   )
)
```

Manipulating Data

The two data frames that include indices and metadata are merged using `inner_join()` by `CountryCode`.

```
data_full <- inner_join(index, metadata, by = "CountryCode")
```

Then, the data frame is manipulated using `select()` to limit columns to those will be used.

```
data_full <- select(data_full, CountryName, CountryCode, Region, IncomeGroup, Year, Indicator)
```

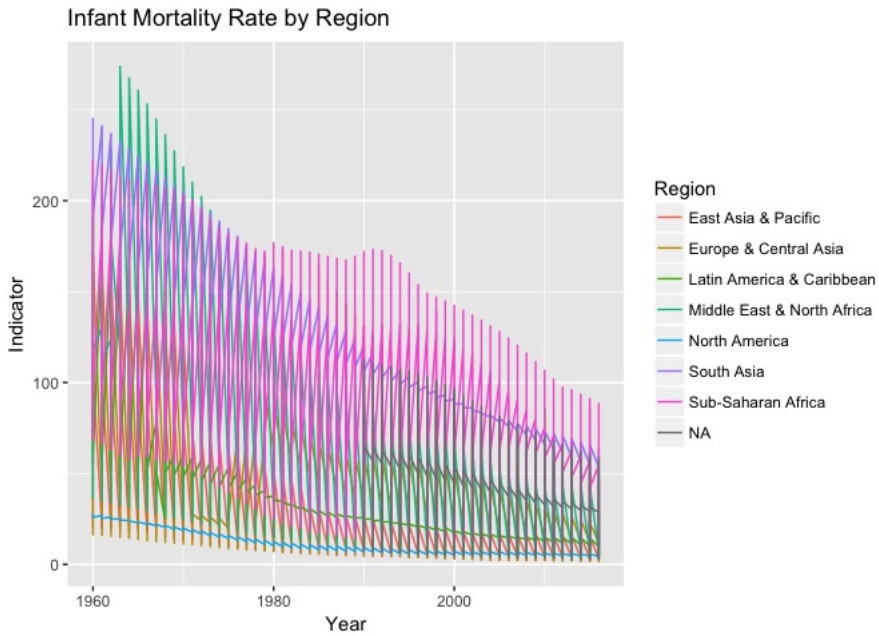
Then, the countries that does not have any data captured in the dataset is excluded using `filter()` to enable better data analysis.

```
data_full <- filter(data_full, is.na(Indicator) == FALSE)
```

Analysis by Region

To carry out data analysis be region, the graph was created using the package `ggplot2`.

```
ggplot(data = data_full, aes(x = Year, y = Indicator, group = Region)) +
  geom_line(aes(color = Region)) +
  ggtitle("Infant Mortality Rate by Region")
```

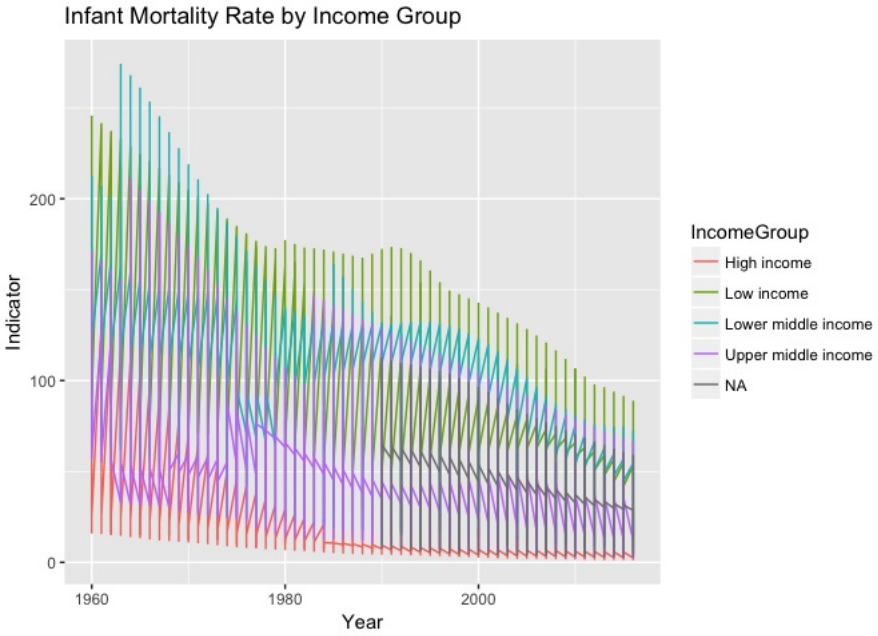


The graph shows that the Middle East and North African region had the highest infant mortality rate during the 1960; however, as the region experienced steep decline in infant mortality, Sub-Saharan Africa became the region with the highest infant mortality after the late 1970s.

Analysis by Nation's Income Group

Then, another graph was created using the package `ggplot2` to examine the data in different perspective.

```
ggplot(data = data_full, aes(x = Year, y = Indicator, group = IncomeGroup)) +
  geom_line(aes(color = IncomeGroup)) +
  ggtitle("Infant Mortality Rate by Income Group")
```



The graph of infant mortality rate by region also shows similar trend. Although the infant mortality rate was the highest in nations with lower middle income, the low income nations have the highest infant mortality rate after the late 1970s.

Income Distribution in Different Regions

To further examine whether the trends in two infant mortality graphs in previous parts were relevant, the analysis on income distribution of nations by region was carried out.

In order to take a look at the income distribution in different regions, a new data frame with the list of countries was created using the functions `select()` and `unique()` based on the full data frame.

```
country_list <- select(data_full, CountryName, CountryCode, Region, IncomeGroup)
country_list <- unique(country_list)
```

Then, another data frame with the proportion of each income group in different regions was created.

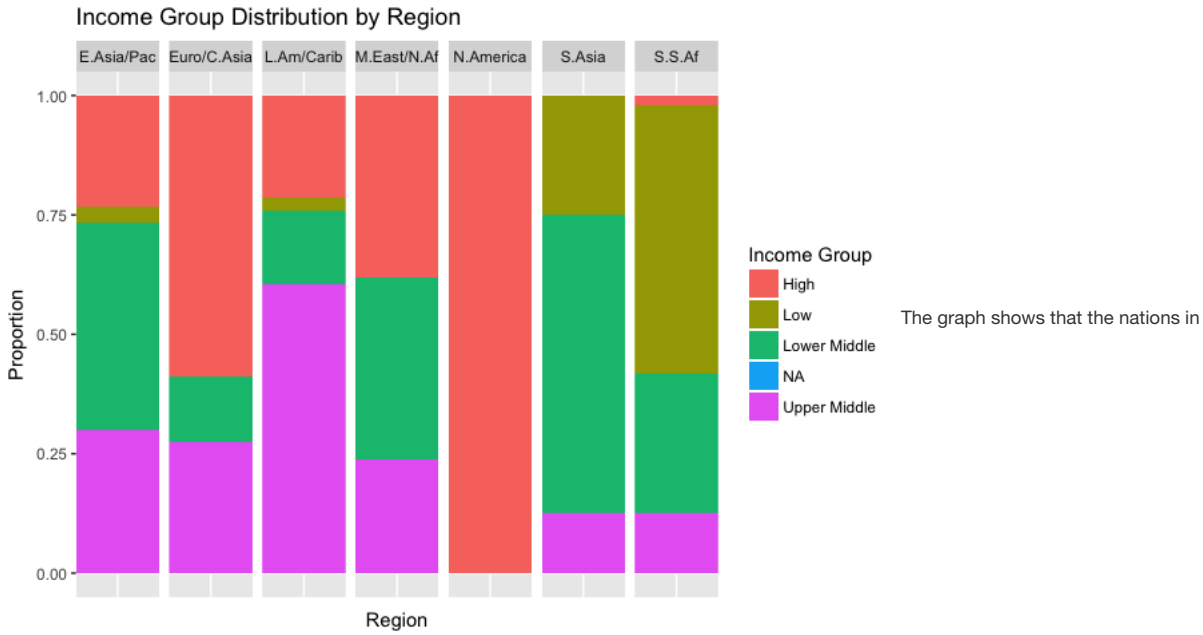
```
data_prop <- data.frame(
      IncomeGroup = rep(c('Low income', 'Lower middle income', 'Upper middle income', 'High income', 'NA'), 7),
      Region = factor(rep(c('East Asia & Pacific', 'South Asia', 'Europe & Central Asia', 'Middle East & North Afr
ica', 'Sub-Saharan Africa', 'North America', 'Latin America & Caribbean'), each = 5)),
      TotalCountry = as.numeric(t(c(rep(count(filter(country_list, Region == "East Asia & Pacific")), 5),
                      rep(count(filter(country_list, Region == "South Asia")), 5),
                      rep(count(filter(country_list, Region == "Europe & Central Asia")), 5),
                      rep(count(filter(country_list, Region == "Middle East & North Africa")), 5),
                      rep(count(filter(country_list, Region == "Sub-Saharan Africa")), 5),
                      rep(count(filter(country_list, Region == "North America")), 5),
                      rep(count(filter(country_list, Region == "Latin America & Caribbean")), 5)
                      ))),
      CountIncGroup = as.numeric(t(c(count(filter(filter(country_list, Region == "East Asia & Pacific"), IncomeGro
up == "Low income")),
                       count(filter(filter(country_list, Region == "East Asia & Pacific"), IncomeGroup == "Lower
middle income")),
                       count(filter(filter(country_list, Region == "East Asia & Pacific"), IncomeGroup == "Upper
middle income")),
                       count(filter(filter(country_list, Region == "East Asia & Pacific"), IncomeGroup == "High i
ncome")),
                       count(filter(filter(country_list, Region == "East Asia & Pacific"), IncomeGroup == "NA")),
                       count(filter(filter(country_list, Region == "South Asia"), IncomeGroup == "Low income")),
                       count(filter(filter(country_list, Region == "South Asia"), IncomeGroup == "Lower middle in
come")),
                       count(filter(filter(country_list, Region == "South Asia"), IncomeGroup == "Upper middle in
come")),
                       count(filter(filter(country_list, Region == "South Asia"), IncomeGroup == "High income")),
                       count(filter(filter(country_list, Region == "South Asia"), IncomeGroup == "NA")),
                       count(filter(filter(country_list, Region == "Europe & Central Asia"), IncomeGroup == "Low
income")),
                       count(filter(filter(country_list, Region == "Europe & Central Asia"), IncomeGroup == "Lowe
r middle income")),
                       count(filter(filter(country_list, Region == "Europe & Central Asia"), IncomeGroup == "Uppe
r middle income")),
                       count(filter(filter(country_list, Region == "Europe & Central Asia"), IncomeGroup == "High
income")),
                       count(filter(filter(country_list, Region == "Europe & Central Asia"), IncomeGroup == "NA")
),
                       count(filter(filter(country_list, Region == "Middle East & North Africa"), IncomeGroup ==
"Low income")),
                       count(filter(filter(country_list, Region == "Middle East & North Africa"), IncomeGroup ==
"Lower middle income")),
                       count(filter(filter(country_list, Region == "Middle East & North Africa"), IncomeGroup ==
"Upper middle income")),
                       count(filter(filter(country_list, Region == "Middle East & North Africa"), IncomeGroup ==
"High income")),
                       count(filter(filter(country_list, Region == "Middle East & North Africa"), IncomeGroup ==
"NA")),
                       count(filter(filter(country_list, Region == "Sub-Saharan Africa"), IncomeGroup == "Low inc
ome")),
                       count(filter(filter(country_list, Region == "Sub-Saharan Africa"), IncomeGroup == "Lower m
iddle income")),
                       count(filter(filter(country_list, Region == "Sub-Saharan Africa"), IncomeGroup == "Upper m
iddle income")),
                       count(filter(filter(country_list, Region == "Sub-Saharan Africa"), IncomeGroup == "High in
come")),
                       count(filter(filter(country_list, Region == "Sub-Saharan Africa"), IncomeGroup == "NA")),
                       count(filter(filter(country_list, Region == "North America"), IncomeGroup == "Low income")
),
                       count(filter(filter(country_list, Region == "North America"), IncomeGroup == "Lower middle
income")),
                       count(filter(filter(country_list, Region == "North America"), IncomeGroup == "Upper middle
income")),
                       count(filter(filter(country_list, Region == "North America"), IncomeGroup == "High income"
)),
                       count(filter(filter(country_list, Region == "North America"), IncomeGroup == "NA")),
                       count(filter(filter(country_list, Region == "Latin America & Caribbean"), IncomeGroup == "
Low income")),
                       count(filter(filter(country_list, Region == "Latin America & Caribbean"), IncomeGroup == "
Lower middle income")),
                       count(filter(filter(country_list, Region == "Latin America & Caribbean"), IncomeGroup == "
Upper middle income")),
                       count(filter(filter(country_list, Region == "Latin America & Caribbean"), IncomeGroup == "
High income")),
                       count(filter(filter(country_list, Region == "Latin America & Caribbean"), IncomeGroup == "
NA"))
                       )))
)

data_prop <- mutate(data_prop,
                    Prop = data_prop$CountIncGroup / data_prop$TotalCountry)
```

Then, the package `sqldf` was used to tidy the data frame, and the barplot of income group proportion in different regions was created using the package `ggplot2`.

```
library(sqldf)
```

```
## Loading required package: gsubfn

## Loading required package: proto

## Warning in doTryCatch(return(expr), name, parentenv, handler): unable to load shared object '/Library/Framework
s/R.framework/Resources/modules//R_X11.so':
##   dlopen(/Library/Frameworks/R.framework/Resources/modules//R_X11.so, 6): Library not loaded: /opt/X11/lib/libS
M.6.dylib
##   Referenced from: /Library/Frameworks/R.framework/Resources/modules//R_X11.so
##   Reason: image not found

## Could not load tcltk.  Will use slower R code instead.

## Loading required package: RSQLite
```

```
data_prop = sqldf("select Prop,
                  CASE WHEN IncomeGroup == 'Low income' THEN 'Low'
                       WHEN IncomeGroup == 'Lower middle income' THEN 'Lower Middle'
                       WHEN IncomeGroup == 'Upper middle income' THEN 'Upper Middle'
                       WHEN IncomeGroup == 'High income' THEN 'High'
                       WHEN IncomeGroup == 'NA' THEN 'NA'
                  END income,
                  CASE WHEN Region == 'East Asia & Pacific' THEN 'E.Asia/Pac'
                       WHEN Region == 'South Asia' THEN 'S.Asia'
                       WHEN Region == 'Europe & Central Asia' THEN 'Euro/C.Asia'
                       WHEN Region == 'Middle East & North Africa' THEN 'M.East/N.Af'
                       WHEN Region == 'Sub-Saharan Africa' THEN 'S.S.Af'
                       WHEN Region == 'North America' THEN 'N.America'
                       WHEN Region == 'Latin America & Caribbean' THEN 'L.Am/Carib'
                  END region from data_prop")

ggplot(data = data_prop, aes(x = factor(1), y = Prop, fill = factor(income)), ) +
  geom_bar(width = 1, stat = "identity") +
  facet_grid(facets=. ~ region) +
  xlab('Region') +
  ylab('Proportion') +
  labs(fill='Income Group') +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank()) +
  ggtitle("Income Group Distribution by Region")
```



The graph shows that the nations in Sub-Saharan Africa were mostly in the low income group; surprisingly, the region with the highest proportion of lower middle income group was, however, the South Asia.

Nonetheless, notable proportion of nations in the Middle East and North Africa were in the high income group, and this seems to have contributed to lower infant mortality rate of the Middle East and North Africa after the late 1970s.

Take Home Message

The world is changing quite fast and require skills to digest larger and larger data. R seems to be the tool that is the most suitable for people in such modern world.

Problems like the infant mortality rate examined in this post require careful approach with precision and speed at the same time. Lower accuracy or slower reaction on such problem could result in millions of infants losing their lives.

The data examined in this post was very large; as mentioned in the data dictionary, the original csv file included 15053 rows. Accordingly, when I first tried to see how the Excel would handle the data, the data seemed too heavy for the Excel. However, the data analysis was much faster and efficient with R. Although it takes an effort to learn the language and structure to use R, it is certain that R is one of the best data analysis toolkit for “Data Analysis Cycle” composed of data preparation, actual analysis, and reporting.

References

- https://data.worldbank.org/indicator/SP.DYN.IMRT.IN
- https://www.r-chart.com/2010/07/pie-charts-in-ggplot2.html
- https://cran.rstudio.com/web/packages/sqldf/sqldf.pdf
- https://www.rdocumentation.org/packages/sqldf/versions/0.4-11
- https://strengejacke.wordpress.com/2013/03/05/easily-plotting-grouped-bars-with-ggplot/
- https://www.red-gate.com/simple-talk/dotnet/software-tools/data-manipulation-in-r--beyond-sql/
- http://www.cookbook-r.com/Graphs/
- https://blog.exploratory.io/filter-data-with-dplyr-76cf5f1a258e