Data Visualization Project: Corruption Perceptions Index

Group E: Cheong Bae, Yumna Ikram, and Tamiko Isaacs

Department of Psychology, York University

PSYC 3031 A: Intermediate Statistics Laboratory

Dr. Monique Herbert

November 4, 2020

Data Visualization Project: Corruption Perceptions Index

Transparency International is an organization with a primary goal of ending public sector corruption in over 100 countries by promoting transparency, accountability, and integrity at all levels and across all sectors of society (Transparency International, n.d.). To understand the varying levels of corruption worldwide, Transparency International launched the Corruption Perceptions Index (CPI) in 1995 (Transparency International, n.d.). The CPI is a composite index that scores and ranks countries and territories based on how corrupt their public sector(s) are perceived to be by experts and business executives (Transparency International, n.d.). Additionally, the CPI provides an annual overview of the levels of corruption by ranking countries and territories from all over the world on a scale of 0 (highly corrupt) to 100 (very clean) (Transparency International, n.d.). In 2012, Transparency International revised the criteria used to create the CPI to allow for annual score comparisons (Transparency International, n.d.). The original dataset for this project comprised CPI annual scores from 2012 to 2017 for 180 countries. The data includes two categorical variables (country and year) and one continuous variable (CPI score). For this project, a revised dataset was created with a primary focus on visualizing the CPI scores over the course of 5 years leading up to 2017 for the final year's top 10 and bottom 10 scoring countries. The 20 countries alongside their 2017 rankings and CPI scores are presented in *Table 1*.

Table 1

Top 10 and bottom 10 scoring countries in 2017

Top countries	Rank	CPI Score	Bottom countries	Rank	CPI Score
New Zealand	1	89	Equatorial Guinea	171	17
Denmark	2	88	Korea, North	171	17

Norway	3	85	Libya	171	17
Switzerland	3	85	Guinea Bissau	171	17
Finland	3	85	Yemen	175	16
Sweden	6	84	Sudan	175	16
Singapore	6	84	Afghanistan	177	15
United Kingdom	8	82	Syria	178	14
Canada	8	82	South Sudan	179	12
Luxembourg	8	82	Somalia	180	9

Note. CPI = Corruption Perceptions Index.

Visualization Question

For the data visualization project, we compare the change of CPI scores between the top-scoring countries versus the bottom-scoring countries over a 5-year period. More specifically, through visualization, our project aims to determine whether there is a difference of potential significance in the pattern of change in the perceived levels of public sector corruption between the top 10 and bottom 10 scoring countries over a 5-year period.

Goals and Outcomes of the Visualization

Our visualization showcases the CPI scores of the top and bottom 10 scoring countries in the year 2017 over a 5-year period (2013-2017) using two slopegraphs: one for the bottom 10 and one for the top 10. Our goal with this visualization was to discern any possible pattern differences between the top- and bottom-scoring countries' CPI scores over this 5-year period. The initial expectation we had before visualizing our data was that the top 10 countries would have less fluctuation and more consistency in their scores year-to-year in comparison to the

bottom 10 countries. For the bottom 10 countries, we expected to see more fluctuation in their scores year-to-year (*Figure 1*).

Upon graphing our data, we found a discernible difference between the two slopegraphs, with the differences falling in line with our expectation of the top-scoring countries (*Figure 2*) maintaining approximate consistency in their scores. For instance, Switzerland's score fluctuated between 85 and 86. In contrast, the slopegraph of the bottom-scoring countries (*Figure 3*) showed more fluctuation in scores, with Yemen, Libya, Syria, and South Sudan all seeing larger drops in their scores in the year 2016, specifically. Visualization of the discernible difference between the change in CPI scores of the top- and bottom-scoring countries suggest there are differences in changing corruption activity potentially worth investigating with further research.

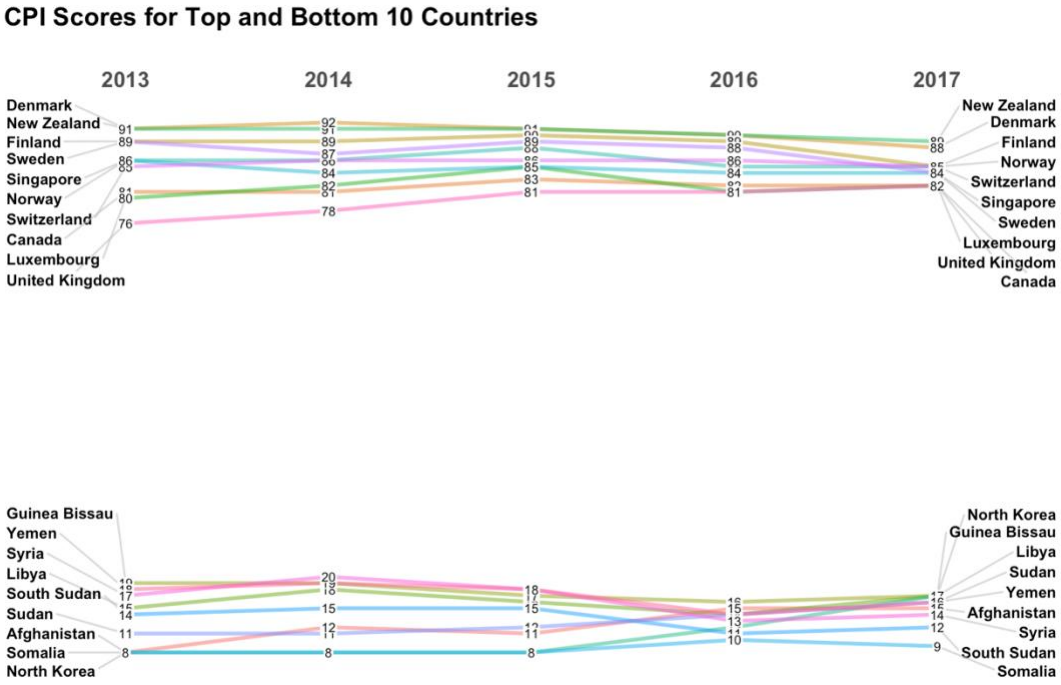

Figure 1. CPI scores for top- and bottom-scoring countries from 2013 to 2017.

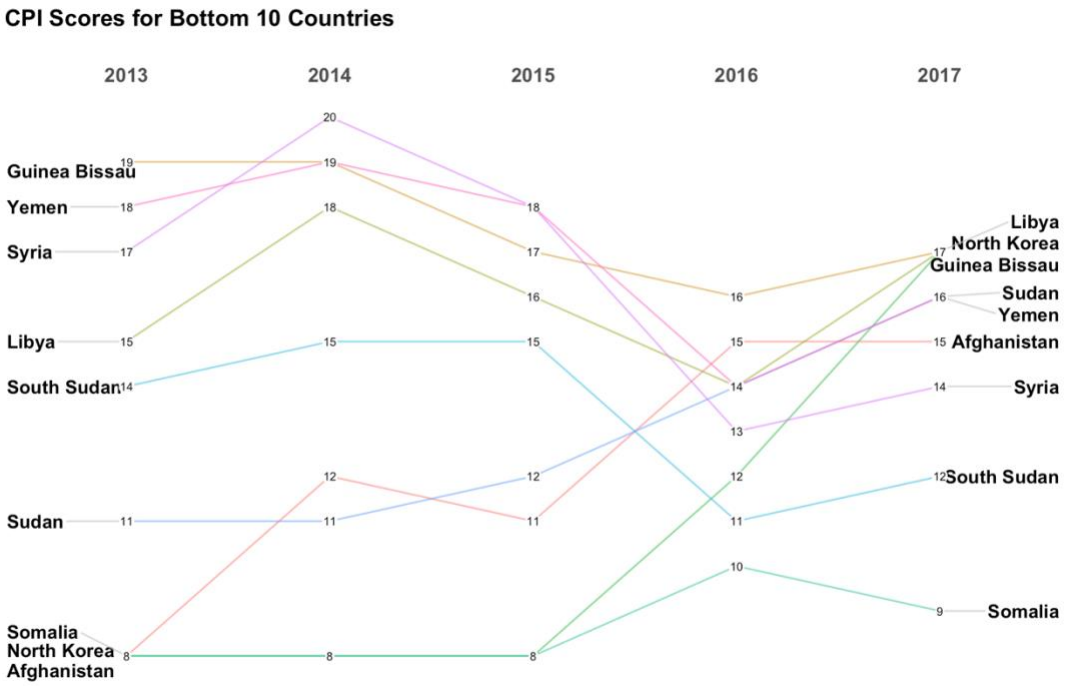

Figure 2. CPI scores for bottom-scoring countries from 2013 to 2017.

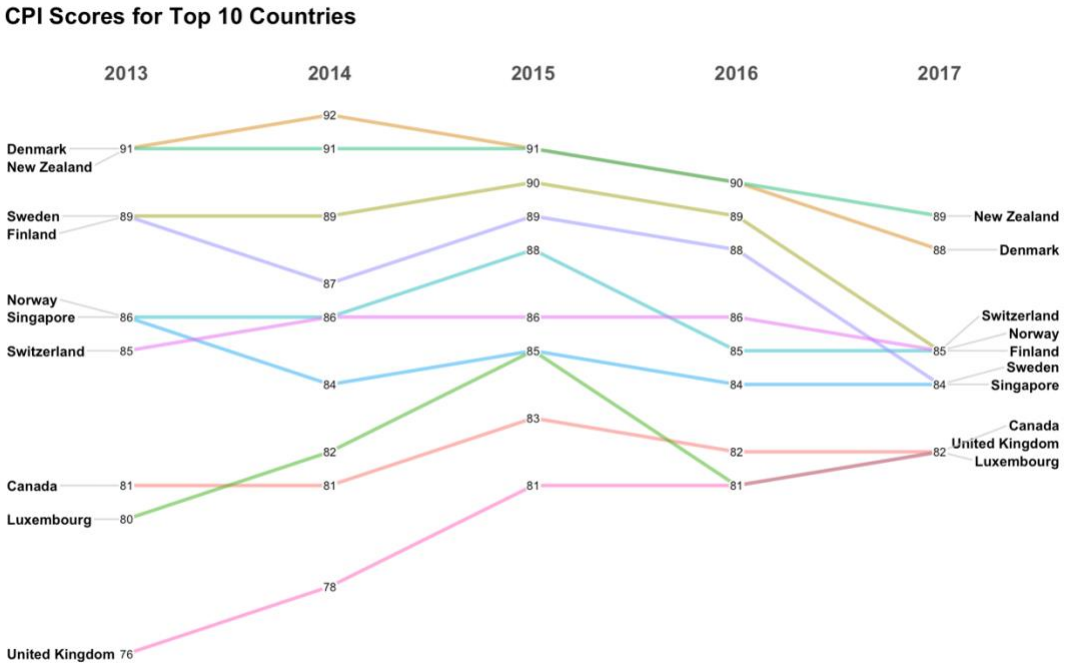

Figure 3. CPI scores for top-scoring countries from 2013 to 2017.

R Code with Comments

```
# In order to visualize this data using slopegraphs, install the following
packages:
# tidyverse, CGPfunctions, psych, gcookbook

# After completing installation, load these packages using the library()
function. This allows R to recognize and use the functions from these
packages
library(tidyverse)

## ── Attaching packages
─────────────────────────────────────────────────────────────────────────
tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.2     ✓ purrr   0.3.4
## ✓ tibble  3.0.3     ✓ dplyr   1.0.2
## ✓ tidyr   1.1.2     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.5.0

## ── Conflicts
─────────────────────────────────────────────────────────────────────────
── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(CGPfunctions)

## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car

library(psych)

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

library(gcookbook)


# Before doing anything with the data, R needs to recognize it. Therefore it
needs to be imported. Do this using the read_csv() function
# After importing the data set, turn it into a data frame using the
```

```
data_frame() function. This will be helpful when using the actual slopegraph code later, where it asks for data frame
# After importing and making the data set into a data frame, use the View() function - this will get R to create a new tab next to the R Script so the data can be seen in a table. This is helpful to have as a reference and to ensure the data was imported correctly.
CPI_Scores_Sheet1 <- read_csv("CPI Scores - Sheet1.csv")
```

```
## Parsed with column specification:
## cols(
##   Year = col_double(),
##   Country = col_character(),
##   `CPI Score` = col_double()
## )
```

```
data_frame(CPI_Scores_Sheet1)
```

```
## Warning: `data_frame()` is deprecated as of tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
## # A tibble: 100 x 3
##     Year Country     `CPI Score`
##    <dbl> <chr>             <dbl>
##  1  2013 New Zealand          91
##  2  2014 New Zealand          91
##  3  2015 New Zealand          91
##  4  2016 New Zealand          90
##  5  2017 New Zealand          89
##  6  2013 Denmark              91
##  7  2014 Denmark              92
##  8  2015 Denmark              91
##  9  2016 Denmark              90
## 10  2017 Denmark              88
## # … with 90 more rows
```

```
View(CPI_Scores_Sheet1)
```

```
# In order to visualize the data later, it is important that the variables are listed as the appropriate type. Find out what the variables are listed as in R using the sapply() function
sapply(CPI_Scores_Sheet1, class)
```

```
##        Year     Country   CPI Score
##   "numeric" "character"   "numeric"
```

```
# For the slopegraph function, the variable "year" must be a factor variable, currently it is numeric
# To change the variable "year" to a factor, use the c() function to make the variable "year" into a vector, and then use the lapply() function to change it from numeric to factor
```

```
year <- c(1:1)
CPI_Scores_Sheet1[,year] <- lapply(CPI_Scores_Sheet1[,year] , factor)


# In order to verify that this was done correctly, use the sapply() function
again to check the variable types
sapply(CPI_Scores_Sheet1, class)

##        Year     Country   CPI Score
##    "factor" "character"   "numeric"

# It was noticed at this point that the variable name for CPI scores needed
to be changed to either CPI_score or CPIscore to get the slopegraph function
to work as R does not work with variable names comprising separate words
# Use the colnames() function to change the 3rd column variable "CPI Score"
to "CPIscore"
colnames(CPI_Scores_Sheet1)[3] <- "CPIscore"


# After changing this, the newggslopegraph() function can now be used. This
function requires specifying the data frame being graphed, the variable used
for the time intervals (year), the variable for the unit of measure
(CPIscore), and the variable by which the data will be grouped (Country)
newggslopegraph(dataframe = CPI_Scores_Sheet1,
                Times = Year,
                Measurement = CPIscore,
                Grouping = Country,
                Title = "CPI Scores for Top and Bottom 10 Countries",
                SubTitle ="",
                Caption = NULL)

##
## Converting 'Year' to an ordered factor

# This visualization came out looking cluttered, and it was difficult to
clearly see patterns in countries' scores
# It was then decided that, for goals of this visualization, it would make
more sense to split the top and bottom 10 into 2 separate slopegraphs
# In order to create 2 graphs, the top and bottom countries needed to be made
into 2 separate sets of data
# This was done by subsetting the data using the slice() function - this
function subsets data based on rows selected by the coder
# After subsetting, the View() function was used to ensure that the data was
subsetted correctly
bottom_10 <- CPI_Scores_Sheet1 %>% slice(51:100)
view(bottom_10)
top_10 <- CPI_Scores_Sheet1 %>% slice(1:50)
view(top_10)
```

```
# After successfully subsetting the data, used the the newggslopegraph()
function again to visualize the Bottom 10 Countries' Scores
newggslopegraph(dataframe = bottom_10,
                Times = Year,
                Measurement = CPIscore,
                Grouping = Country,
                Title = "CPI Scores for Bottom 10 Countries",
                SubTitle ="",
                Caption = NULL)

##
## Converting 'Year' to an ordered factor

# After successfully creating a slopegraph for the Bottom 10, the same
newggslopegraph() function was used to visualize the Top 10 Countries' Scores
newggslopegraph(dataframe = top_10,
                Times = Year,
                Measurement = CPIscore,
                Grouping = Country,
                Title = "CPI Scores for Top 10 Countries",
                SubTitle ="",
                Caption = NULL)

##
## Converting 'Year' to an ordered factor
```

Limitations

There were limitations with the data, as well as with the visualization applied. With respect to the data, Transparency International calculated the CPI scores using data from 13 external sources comprising experts and business executives (Transparency International, n.d.). The data sources measured public sector corruption in the following forms: bribery, public funds diversion, corruption prosecution, and legal protections for whistleblowers, journalists, and investigators (Transparency International, n.d.). Corruption operates illegally, however, and manifests in activities often hidden and not included in the calculation of the CPI scores (Transparency International, n.d.). Thus, it is important to note that the CPI does not account for all corrupt activities, including tax fraud, money laundering, and illicit money flow (Transparency International, n.d.). The inclusion of these activities is difficult, with their

exposure typically dependent on scandals and the success of subsequent investigation (Transparency International, n.d.). With critical activities omitted from the production of the CPI scores, the validity of the data visualization is challenged and therefore should not be considered complete representations of corruption.

In the visualization of the data, the bottom-scoring country Equatorial Guinea did not have recorded CPI scores from 2014 to 2016 and had to be omitted from the data representation due to missing data. Another limitation was the significant disparity of score values between the top and bottom countries, minimizing the ability to discern meaningful patterns of change when both top- and bottom-scoring countries were represented on the same slopegraph. Separating the data into two slopegraphs, one visualizing the top-scoring countries and the other visualizing the bottom-scoring countries, resolved this issue. The y-axis of CPI scores were minimized to a smaller range, and the consequent visualization in two separate slopegraphs allowed the values to be represented with greater clarity.

References

Transparency International (n.d.). *Corruption Perceptions Index*. Retrieved October 24, 2020, from https://www.transparency.org/en/cpi.