Gender Statistics: Data Wrangling And Cleaning

Chad Landreth

12/2/2019

```
library(stringr)
library(dplyr)
library(tidyr)
library(magrittr)
library(readr)
library(kableExtra)
```

The purpose of this document is to properly source the respective dataset for this study and provide detailed steps and code for wrangling and cleaning the data. Ultimately, this document will serve as the resource of full traceability from the sourced, raw dataset to the "cleaned", tidy dataset for uploading to the Tableau application and for further visual analysis.

Source:

The dataset was curated and downloaded, along with its metadata, from the Gender Statistics database under the Databases section of The World Bank Group website at the following link: Gender Statistics

Fields:

Series Name - Name (description) identifying the series
Series Code - Unique variable-length code identifying the series
Country Name - Name (description) identifying the country (economy)
Country Code - Unique 3-letter code identifying the country (economy)
1999 [YR1999] - The measures or indicators (values) of the series within the 1999 year
2000 [YR2000] - The measures or indicators (values) of the series within the 2000 year
2001 [YR2001] - The measures or indicators (values) of the series within the 2001 year
2002 [YR2002] - The measures or indicators (values) of the series within the 2002 year
2003 [YR2003] - The measures or indicators (values) of the series within the 2003 year
2004 [YR2004] - The measures or indicators (values) of the series within the 2004 year
2005 [YR2005] - The measures or indicators (values) of the series within the 2005 year
2006 [YR2006] - The measures or indicators (values) of the series within the 2006 year
2007 [YR2007] - The measures or indicators (values) of the series within the 2007 year
2008 [YR2008] - The measures or indicators (values) of the series within the 2008 year
2009 [YR2009] - The measures or indicators (values) of the series within the 2009 year
2010 [YR2010] - The measures or indicators (values) of the series within the 2010 year
2011 [YR2011] - The measures or indicators (values) of the series within the 2011 year
2012 [YR2012] - The measures or indicators (values) of the series within the 2012 year
2013 [YR2013] - The measures or indicators (values) of the series within the 2013 year
2014 [YR2014] - The measures or indicators (values) of the series within the 2014 year
2015 [YR2015] - The measures or indicators (values) of the series within the 2015 year
2016 [YR2016] - The measures or indicators (values) of the series within the 2016 year
2017 [YR2017] - The measures or indicators (values) of the series within the 2017 year
2018 [YR2018] - The measures or indicators (values) of the series within the 2018 year

Series:

FX.OWN.TOTL.FE.ZS

- Percent | Annual | Weighted average
  - Account ownership at a financial institution or with a mobile-money-service provider, female
  - (% of population ages 15+)

FX.OWN.TOTL.MA.ZS

- Percent | Annual | Weighted average
  - Account ownership at a financial institution or with a mobile-money-service provider, male
  - (% of population ages 15+)

SL.FAM.WORK.FE.ZS

- Percent | Annual | Weighted average
  - Contributing family workers, female
  - (% of female employment) (modeled ILO estimate)

SL.FAM.WORK.MA.ZS

- Percent | Annual | Weighted average
  - Contributing family workers, male
  - (% of male employment) (modeled ILO estimate)

IC.REG.COST.PC.FE.ZS

- Percent | Annual | Unweighted average
  - Cost of business start-up procedures, female
  - (% of GNI per capita)

IC.REG.COST.PC.MA.ZS

- Percent | Annual | Unweighted average
  - Cost of business start-up procedures, male
  - (% of GNI per capita)

SL.EMP.MPYR.FE.ZS

- Percent | Annual | Weighted average

  - Employers, female
  - (% of female employment) (modeled ILO estimate)

SL.EMP.MPYR.MA.ZS

- Percent | Annual | Weighted average
  - Employers, male
  - (% of male employment) (modeled ILO estimate)

SL.AGR.EMPL.FE.ZS

- Percent | Annual | Weighted average
  - Employment in agriculture, female
  - (% of female employment) (modeled ILO estimate)

SL.AGR.EMPL.MA.ZS

- Percent | Annual | Weighted average
  - Employment in agriculture, male
  - (% of male employment) (modeled ILO estimate)

SL.IND.EMPL.FE.ZS

- Percent | Annual | Weighted average
  - Employment in industry, female
  - (% of female employment) (modeled ILO estimate)

SL.IND.EMPL.MA.ZS

- Percent | Annual | Weighted average
  - Employment in industry, male
  - (% of male employment) (modeled ILO estimate)

SL.SRV.EMPL.FE.ZS

- Percent | Annual | Weighted average
  - Employment in services, female
  - (% of female employment) (modeled ILO estimate)

SL.SRV.EMPL.MA.ZS

- Percent | Annual | Weighted average

  - Employment in services, male
  - (% of male employment) (modeled ILO estimate)

SL.EMP.TOTL.SP.FE.ZS

- Percent | Annual | Weighted average
  - Employment to population ratio, 15+, female
  - (%) (modeled ILO estimate)

SL.EMP.TOTL.SP.MA.ZS

- Percent | Annual | Weighted average
  - Employment to population ratio, 15+, male
  - (%) (modeled ILO estimate)

SL.EMP.1524.SP.FE.ZS

- Percent | Annual | Weighted average
  - Employment to population ratio, ages 15-24, female
  - (%) (modeled ILO estimate)

SL.EMP.1524.SP.MA.ZS

- Percent | Annual | Weighted average
  - Employment to population ratio, ages 15-24, male
  - (%) (modeled ILO estimate)

fin1.t.a.2

- Percent | Triennial | Weighted average
  - Financial institution account, female
  - (% age 15+)

fin1.t.a.1

- Percent | Triennial | Weighted average
  - Financial institution account, male
  - (% age 15+)

NY.GDP.MKTP.CD

- United States Dollars | Annual | Gap-filled total
  - Gross Domestic Product
  - (current US$)

SL.TLF.ACTI.1524.FE.ZS

- Percent | Annual | Weighted average
  - Labor force participation rate for ages 15-24, female
  - (%) (modeled ILO estimate)

SL.TLF.ACTI.1524.MA.ZS

- Percent | Annual | Weighted average
  - Labor force participation rate for ages 15-24, male
  - (%) (modeled ILO estimate)

SL.TLF.CACT.FE.ZS

- Percent | Annual | Weighted average
  - Labor force participation rate, female
  - (% of female population ages 15+) (modeled ILO estimate)

SL.TLF.ACTI.FE.ZS

- Percent | Annual | Weighted average
  - Labor force participation rate, female
  - (% of female population ages 15-64) (modeled ILO estimate)

SL.TLF.CACT.MA.ZS

- Percent | Annual | Weighted average
  - Labor force participation rate, male
  - (% of male population ages 15+) (modeled ILO estimate)

SL.TLF.ACTI.MA.ZS

- Percent | Annual | Weighted average
  - Labor force participation rate, male
  - (% of male population ages 15-64) (modeled ILO estimate)

SL.TLF.TOTL.FE.IN

- Whole Number | Annual | Sum
  - Labor force, female

SL.TLF.TOTL.FE.ZS

- Percent | Annual | Weighted average
  - Labor force, female
  - (% of total labor force)

SL.TLF.TOTL.IN

- Whole Number | Annual | Sum
  - Labor force, total

SG.LAW.EQRM.WK

- 1 = Yes, 0 = No | Annual
  - Law mandates equal remuneration for females and males for work of equal value
  - (1=yes; 0=no)

SG.LAW.NODC.HR

- 1 = Yes, 0 = No | Annual
  - Law mandates nondiscrimination based on gender in hiring
  - (1=yes; 0=no)

SG.LAW.LEVE.PU

- 1 = Yes, 0 = No | Annual
  - Law mandates paid or unpaid maternity leave
  - (1=yes; 0=no)

SG.LAW.CHMR

- 1 = Yes, 0 = No | Annual
  - Law prohibits or invalidates child or early marriage
  - (1=yes; 0=no)

SG.LEG.DVAW

- 1 = Yes, 0 = No | Annual
  - Legislation exists on domestic violence
  - (1=yes; 0=no)

SG.LEG.SXHR.EM

- 1 = Yes, 0 = No | Annual
  - Legislation exists on sexual harassment in employment
  - (1=yes; 0=no)

SG.LEG.MRRP

- 1 = Yes, 0 = No | Annual
  - Legislation explicitly criminalizes marital rape
  - (1=yes; 0=no)

SG.LEG.SXHR

- 1 = Yes, 0 = No | Annual
  - Legislation specifically addresses sexual harassment
  - (1=yes; 0=no)

SG.OWN.PRRT.MR

- 1 = Yes, 0 = No | Annual
  - Married men and married women have equal ownership rights to property
  - (1=yes; 0=no)

SG.LAW.OBHB.MR

- 1 = Yes, 0 = No | Annual
  - Married women are required by law to obey their husbands
  - (1=yes; 0=no)

SH.MMR.WAGE.ZS

- Percent | Annual
  - Maternal leave benefits
  - (% of wages paid)

SH.MMR.LEVE

- Whole Number | Annual
  - Maternity leave
  - (days paid)

SG.MMR.LEVE.EP

- 1 = Yes, 0 = No | Annual
  - Mothers are guaranteed an equivalent position after maternity leave
  - (1=yes; 0=no)

SG.JOB.NOPN.EQ

- 1 = Yes, 0 = No | Annual
  - Nonpregnant and nonnursing women can do the same jobs as men
  - (1=yes; 0=no)

SL.EMP.OWAC.FE.ZS

- Percent | Annual | Weighted average
  - Own-account workers, female
  - (% of female employment) (modeled ILO estimate)

SL.EMP.OWAC.MA.ZS

- Percent | Annual | Weighted average
  - Own-account workers, male
  - (% of male employment) (modeled ILO estimate)

SP.POP.TOTL.FE.IN

- Whole Number | Annual | Sum
  - Population, female

SP.POP.TOTL

- Whole Number | Annual | Sum
  - Population, total

SG.GEN.PARL.ZS

- Percent | Annual | Weighted average
  - Proportion of seats held by women in national parliaments
  - (%)

SG.GEN.MNST.ZS

- Percent | Annual | Weighted average
  - Proportion of women in ministerial level positions
  - (%)

SL.TLF.CACT.FM.ZS

- Percent | Annual | Weighted average
  - Ratio of female to male labor force participation rate
  - (%) (modeled ILO estimate)

SL.UEM.1524.FM.ZS

- Percent | Annual | Weighted average
  - Ratio of female to male youth unemployment rate
  - (% ages 15-24) (modeled ILO estimate)

fin18.t.d.2

- Percent | Triennial | Weighted average
  - Saved any money in the past year, female
  - (% age 15+)

fin18.t.d.1

- Percent | Triennial | Weighted average
  - Saved any money in the past year, male
  - (% age 15+)

SL.EMP.SELF.FE.ZS

- Percent | Annual | Weighted average
  - Self-employed, female
  - (% of female employment) (modeled ILO estimate)

SL.EMP.SELF.MA.ZS

- Percent | Annual | Weighted average
  - Self-employed, male
  - (% of male employment) (modeled ILO estimate)

IC.REG.PROC.FE

- Decimal | Annual | Unweighted average
  - Start-up procedures to register a business, female
  - (number)

IC.REG.PROC.MA

- Decimal | Annual | Unweighted average
  - Start-up procedures to register a business, male
  - (number)

IC.REG.DURS.FE

- Decimal | Annual | Unweighted average
  - Time required to start a business, female
  - (days)

IC.REG.DURS.MA

- Decimal | Annual | Unweighted average
  - Time required to start a business, male
  - (days)

SL.UEM.TOTL.FE.ZS

- Percent | Annual | Weighted average
  - Unemployment, female
  - (% of female labor force) (modeled ILO estimate)

SL.UEM.TOTL.MA.ZS

- Percent | Annual | Weighted average
  - Unemployment, male

  - (% of male labor force) (modeled ILO estimate)

SL.UEM.1524.FE.ZS

- Percent | Annual | Weighted average
  - Unemployment, youth female
  - (% of female labor force ages 15-24) (modeled ILO estimate)

SL.UEM.1524.MA.ZS

- Percent | Annual | Weighted average
  - Unemployment, youth male
  - (% of male labor force ages 15-24) (modeled ILO estimate)

SL.EMP.VULN.FE.ZS

- Percent | Annual | Weighted average
  - Vulnerable employment, female
  - (% of female employment) (modeled ILO estimate)

SL.EMP.VULN.MA.ZS

- Percent | Annual | Weighted average
  - Vulnerable employment, male
  - (% of male employment) (modeled ILO estimate)

SL.EMP.WORK.FE.ZS

- Percent | Annual | Weighted average
  - Wage and salaried workers, female
  - (% of female employment) (modeled ILO estimate)

SL.EMP.WORK.MA.ZS

- Percent | Annual | Weighted average
  - Wage and salaried workers, male
  - (% of male employment) (modeled ILO estimate)

SG.CRT.TSTM.WT

- 1 = Yes, 0 = No | Annual

  - Woman's testimony carries the same evidentiary weight in court as a man's
  - (1=yes; 0=no)

SG.IND.WORK.EQ

- 1 = Yes, 0 = No | Annual
  - Women are able to work in the same industries as men
  - (1=yes; 0=no)

Data Wrangling and Cleaning:

Please Note:
In the course of choosing each of the Series for the dataset download from the WBG website, some Series are extraneous to the study to be performed in the Tableau application. In these cases, the extraneous Series will be dropped from the dataset before saving the tidied-up dataset to a CSV file. For example, four Series were chosen for Unemployment percentages, however the two total percentages (female, male) are necessary for this study while the two youth percentages (female, male, ages 15-24) will be dropped from the dataset. The specific code for this operation will be called out in the comments and/or markdown text, so users reading this can make his or her own choice of dropping these Series.

1. The raw dataset is read into a dataframe (df) from a CSV-format file.

- The first line of the file is designated as the dataframe column headers.
- All empty strings are replaced with NA.

```
df <- read.csv("Data_Extract_From_Gender_Statistics.csv",
               header = TRUE, na.strings = c("", "NA"))
```

2. All data points are typecast to character strings.

Since there are various data types represented in each row, casting all data points to character strings allows us to retain explicit control of the resulting data types when applying the **gather()** function from the **dplyr** package later on. Most likely, **gather()** will automatically convert all data points to character strings if they are not explicitly casted here, but there is no guarantee of that.

```
df[] <- lapply(df, as.character)
```

3. Replace all whitespace, double dots, and single dots with underscores in each column header.

```
names(df) <- gsub(" ", "_", names(df))
names(df) <- gsub("\\.\\.", "_", names(df))
names(df) <- gsub("\\.", "_", names(df))
```

4. Convert all columns representing the year into a simpler 4-digit year character string.

- For example, the column header for the year 2010 will be converted from "2010_[YR2010]" to "2010".

```
names(df)[5:length(names(df))] <-
  substr(names(df)[5:length(names(df))], 2, 5)
```

5. Drop the Series_Name and Country_Code columns from the dataframe.

- The Series_Name for many of the series are substantially too long as field names for the visual analysis in Tableau, so the Series_Code will be kept instead for brevity. When building the visualizations, however, the dataset's metadata will be referenced to properly label them.
- The Country_Code will not be clear to the audience in the visualizations, so the Country_Name will be kept instead for clarity.

```
df <- select(df, -c(1,4))
```

6. Shift the Country_Name column to the first column of the dataframe (**before** the Series_Code column).

```
df <- select(df, Country_Name, everything())

# <<< Piping Method >>>
#
# df <- df %>%
#     select(-c(1,4)) %>%
#     select(Country_Name, everything())
#
```

7. Replace all single dots with underscores in each Series_Code value.

Eventually, these Series_Code values will become column headers when applying the **spread()** function of the **dplyr** package. This replacement maintains consistency between all column headers. For example, the Series_Code value for "SP.POP.TOTL" will become "SP_POP_TOTL".

```
df$Series_Code <- gsub("\\.", "_", df$Series_Code)
```

8. Sort all rows in ascending order alphabetically by Country_Name first, then by Series_Code using the **arrange()** function of the **dplyr** package.

```
df <- arrange(df, Country_Name, Series_Code)
```

9. Move all column headers specifying the year under a single key column (Year) and all respective column values under a single value column (Series_Value) using the **gather()** function of the **tidyr** package.

Specifying the *names(df)[-(1:2)]* argument passes all but the first two column names to the **gather()** function. In this case, the two excluded column names are Country_Name and Series_Code.

```
df <- gather(df, names(df)[-(1:2)], key = "Year", value = "Series_Value")
```

10. Drop any rows containing any NA's.

Any of the series values in the dataset showing the double dots represent "Not Applicable / No Response", so by keeping these double dots, any rows containing any NA's are truly erroneous and should be removed from the dataframe.

```
df <- na.omit(df)
```

11. Move all unique values under the Series_Code column into individual column headers and all values under the Series_Value column underneath the column header representing the Series_Code for that value using the **spread()** function of the **tidyr** package.

```
df <- spread(df, key = "Series_Code", value = "Series_Value")

# <<< Piping Method >>>
#
# df <- df %>%
#     arrange(Country_Name, Series_Code) %>%
#     gather(names(df)[-(1:2)], key = "Year", value = "Series_Value") %>%
#     na.omit() %>%
#     spread(key = "Series_Code", value = "Series_Value")
#
```

12. Drop unnecessary columns with respect to the visual analysis in Tableau.

KEEP	Col_Idx_K	DROP	Col_Idx_D
SL_EMP_TOTL_SP_FE_ZS	44	SL_EMP_1524_SP_FE_ZS	36
SL_EMP_TOTL_SP_MA_ZS	45	SL_EMP_1524_SP_MA_ZS	37
SL_TLF_CACT_FE_ZS	60	SL_TLF_ACTI_1524_FE_ZS	56
		SL_TLF_ACTI_FE_ZS	58
SL_TLF_CACT_MA_ZS	62	SL_TLF_ACTI_1524_MA_ZS	57
		SL_TLF_ACTI_MA_ZS	59
SL_TLF_TOTL_FE_ZS	64	SL_TLF_TOTL_FE_IN	63
		SL_TLF_TOTL_IN	65
SL_UEM_TOTL_FE_ZS	69	SL_UEM_1524_FE_ZS	66
		SL_UEM_1524_FM_ZS	67
SL_UEM_TOTL_MA_ZS	70	SL_UEM_1524_MA_ZS	68
SP_POP_TOTL	71	SP_POP_TOTL_FE_IN	72

```
# Note to User:
# Comment out the below code if you wish
# to keep all columns up to this point

df <- select(df, -c(36,37,56,57,58,59,63,65,66,67,68,72))
```

14. Filter the dataframe down based on those countries with the highest top 30 total population and/or with the highest top 30 gross domestic product.

Note to User:
For the purpose of the visual analysis in Tableau, the focus is on the aforementioned criteria. The below block of code can be commented out if you wish to include all countries in the final dataset rather than based on this criteria.

```
# Cast the Total Population dataframe column to integer,
# then return the top 30 countries with the highest
# population in 2018

df_pop <- df %>% select(c(Country_Name, Year, SP_POP_TOTL))
df_pop$SP_POP_TOTL <- as.integer(df_pop$SP_POP_TOTL)
```

```r
top_countries_pop <- df_pop %>%
  filter(Year == "2018") %>%
  arrange(desc(SP_POP_TOTL)) %>%
  top_n(30, SP_POP_TOTL) %>%
  select(Country_Name)

top_countries_pop <- top_countries_pop[["Country_Name"]]
df_pop <- NULL

# Cast the Gross Domestic Product dataframe column
# to numeric, then return the top 30 countries with
# the highest GDP in 2018

df_gdp <- df %>% select(c(Country_Name, Year, NY_GDP_MKTP_CD))
df_gdp$NY_GDP_MKTP_CD <- as.numeric(df_gdp$NY_GDP_MKTP_CD)

top_countries_gdp <- df_gdp %>%
  filter(Year == "2018") %>%
  arrange(desc(NY_GDP_MKTP_CD)) %>%
  top_n(30, NY_GDP_MKTP_CD) %>%
  select(Country_Name)

top_countries_gdp <- top_countries_gdp[["Country_Name"]]
df_gdp <- NULL

# Filter the dataframe doum to the countries in the
# top 30 of the highest total population in 2018 or
# the top 30 of the highest GDP in 2018

df <- df %>%
  filter(Country_Name %in% top_countries_pop |
           Country_Name %in% top_countries_gdp)

top_countries_pop <- NULL
top_countries_gdp <- NULL
```

15. Write the dataframe to a CSV format file using the **write_csv()** function of the **readr** package.

```r
write_csv(df, "gender_dataset.csv")

df <- NULL
```