Module 2 Exercise

February 5, 2018

Instructions

For the remaining modules, you will be working directly in RStudio.

1 Create an R-Script File

a. In RStduio, create a new R Script file. Save it as "Exercise1" in a new project folder.
b. Install the packages you will be using for analysis (if not already installed):
   - tidyverse
   - knitr

```
library(rio)
library(tidyverse)
```

c. Set the working directory to your project folder.

2 Import the data

For this exercise, you will be working with data from the "World Development Indicators" data set produced by the World Bank.

a. Download the dataset from https://data.worldbank.org/data-catalog/world-development-indicators.
   - You will need to unzip the download and move the "WDIData.csv" file into a suitable project folder. This will be your data set.
b. Import the "WDIData.csv" data set (with a suitable name for the data)

```
world.dev.data <- import("WDIData.csv")
```

```
##
Read 12.0% of 415800 rows
Read 19.2% of 415800 rows
Read 24.1% of 415800 rows
Read 31.3% of 415800 rows
Read 43.3% of 415800 rows
Read 55.3% of 415800 rows
Read 72.2% of 415800 rows
Read 91.4% of 415800 rows
Read 415800 rows and 63 (of 63) columns from 0.191 GB file in 00:00:19
```

c. Using pipes, convert the data frame into a tibble.

```
world.dev.data <- world.dev.data %>% as.tibble()
```

3 Filter the data set for the desired indicators

"WDIData.csv" is a pretty large data set. For the exercise, you will be working with indicators related to poverty and inequality. Reduce the size of the data set in memory by filtering for only the indicator codes listed below:

Indicator	Short Description
NY.GDP.PCAP.KD	GDP per capita (constant 2010 US$)
SP.POP.TOTL	Population, total
SL.TLF.CACT.FM.ZS	Ratio of female to male labor force participation rate (%) (modeled ILO estimate)
SE.SEC.CUAT.UP.ZS	Educational attainment, at least completed upper secondary, population 25+, total (cum. %)
SL.UEM.NEET.ZS	Share of youth not in education, employment or training, total (% of youth population)
SL.UEM.BASC.ZS	Unemployment with basic education (% of total labor force with basic education)
SI.POV.UMIC	Poverty headcount ratio at $5.50 a day (2011 PPP) (% of population)
SI.DST.FRST.20	Income share held by lowest 20%
SI.DST.02ND.20	Income share held by second 20%
SI.DST.03RD.20	Income share held by third 20%
SI.DST.10TH.10	Income share held by highest 10%

a. First create a vector containing the indicator codes you want to keep.
b. Rename the indicator code column to just "indicator" to make it easier to work with.
c. Use **filter()** with the `%in%` operator to selected the desired observations.

```
myindics <- c("NY.GDP.PCAP.KD", "SP.POP.TOTL", "SL.TLF.CACT.FM.ZS",
              "SE.SEC.CUAT.UP.ZS", "SL.UEM.NEET.ZS", "SL.UEM.BASC.ZS",
              "SI.POV.UMIC","SI.DST.FRST.20","SI.DST.02ND.20",
              "SI.DST.03RD.20","SI.DST.10TH.10")

world.dev.data <- world.dev.data %>% rename(indicator = "Indicator Code")
world.dev.data <- world.dev.data %>% filter(indicator %in% myindics)
```

4 Tidy the data set

a. First, get a sense of the data structure.
b. Drop the following columns: "v63", "Indicator Name"

```
world.dev.data$V63 <- NULL
world.dev.data[,"Indicator Name"] <- NULL
```

c. In order to easily work with the data, you will first need to tidy it.

- Use the **gather** and **spread** functions so that the data adheres to tidy data principles.
- You may need to think about how to specify the columns to be gathered since they appear as numbers.

```
world.dev.data <- world.dev.data %>% gather(key="year", value="value", "1960":"2017") %>%
  spread(key="indicator", value="value")
```

5 Rename and change class of columns

- To make the data truly presentable, rename the indicator columns to a short descriptive variable names (be sure not to include any spaces).
- Change the class of indicators to numeric.

```
world.dev.data$NY.GDP.PCAP.KD <-as.numeric(world.dev.data$NY.GDP.PCAP.KD)
world.dev.data$year <-as.numeric(world.dev.data$year)
world.dev.data$SP.POP.TOTL <-as.numeric(world.dev.data$SP.POP.TOTL)
world.dev.data$SL.TLF.CACT.FM.ZS <-as.numeric(world.dev.data$SL.TLF.CACT.FM.ZS)
world.dev.data$SE.SEC.CUAT.UP.ZS <-as.numeric(world.dev.data$SE.SEC.CUAT.UP.ZS)
```

```
world.dev.data$SL.UEM.NEET.ZS <-as.numeric(world.dev.data$SL.UEM.NEET.ZS)
world.dev.data$SL.UEM.BASC.ZS <-as.numeric(world.dev.data$SL.UEM.BASC.ZS)
world.dev.data$SI.POV.UMIC <-as.numeric(world.dev.data$SI.POV.UMIC)
world.dev.data$SI.DST.FRST.20 <-as.numeric(world.dev.data$SI.DST.FRST.20)
world.dev.data$SI.DST.02ND.20 <-as.numeric(world.dev.data$SI.DST.02ND.20)
world.dev.data$SI.DST.03RD.20 <-as.numeric(world.dev.data$SI.DST.03RD.20)
world.dev.data$SI.DST.10TH.10 <-as.numeric(world.dev.data$SI.DST.10TH.10)
```

- Rename all other variables to one-word names as well.

```
world.dev.data <- world.dev.data %>%
  rename(countrycode = "Country Code", country = "Country Name",
         pcgdp = NY.GDP.PCAP.KD, pop = SP.POP.TOTL,
         emp_femaletomale = SL.TLF.CACT.FM.ZS,
         secondary.complet = SE.SEC.CUAT.UP.ZS,
         disengaged.youth = SL.UEM.NEET.ZS,
         unemp.basic.educ = SL.UEM.BASC.ZS,
         povertyrte = SI.POV.UMIC,
         incshare_low20 = SI.DST.FRST.20, incshare_2nd20 = SI.DST.02ND.20,
         incshare_3rd20 = SI.DST.03RD.20,incshare_top10 = SI.DST.10TH.10)
```

6 Further restrict the sample

We are specifically interested in inequality and poverty measures for "high-income" countries, with at least 4 million people, over the last ten years. If you look carefully at the dataset, you will notice that there are multiple issues that need to be addressed:

a. There are a number of non-country regions that are present which need to be removed.

- The list of county codes that need to be filtered out are supplied in regions.csv.
- Import this file and filter for country codes not in `regions`.
- **Tip**: Use the following code to get `regions` into a format where you can use it with an %in% filter:

```
regions <- import("regions.csv")
regions <- as.character(regions$CountryCode)
```

```
world.dev.data <- world.dev.data %>% filter(!(countrycode %in% regions))
```

- Drop the `countrycode` column once you have removed all regional observations.

```
world.dev.data$countrycode <- NULL
```

b. The sample should be restricted to countries with a population greater than 4 million, GDP per capita greater than $12,000, and year greater than or equal to 2007.

```
world.dev.data <- world.dev.data %>%
  filter(year >= 2000, pcgdp >= 12000,pop > 4000000)
```

7 Create new variables

Now, create some new variables that we might be interested in:

- log of GDP per capita
- The absolute number of people in poverty (at $5.50 a day)
- Income share held by the bottom 60% of the population

- Ratio of income held by top 10% to bottom 60%

```
world.dev.data <- world.dev.data %>%
    mutate(logGDP = log(pcgdp),
           numpoverty = povertyrte * pop,
           income60 = incshare_low20 + incshare_2nd20 + incshare_3rd20,
           income10to60 = incshare_top10 / income60)
```

8 Create and view summary statistics

a. By country, create average values (with option `na.rm=TRUE`) for:

- Ratio of income held by top 10% to bottom 60%
- GDP per capita
- Ratio of female to male labor force partipation
- Share of population having completely at least completed upper secondary schooling
- Share of youth not in education, employment or training, total (% of youth population)
- Unemployment with basic education (% of total labor force with basic education)

```
my_summary <- world.dev.data %>% group_by(country, year) %>%
  summarize(income10to60.avg = mean(income10to60, na.rm=TRUE),
            gdp.avg = mean(pcgdp, na.rm=TRUE),
            emp_femaletomale.avg = mean(emp_femaletomale, na.rm=TRUE),
            secondary.complet.avg = mean(secondary.complet, na.rm=TRUE),
            disengaged.yth.avg = mean(disengaged.youth, na.rm=TRUE),
            unemp.basic.educ.avg = mean(unemp.basic.educ, na.rm=TRUE)
            )

export(my_summary, "C:/Users/AN.4271/OneDrive - Handelshögskolan i Stockholm/Teaching/R Course/Modules/
```

b. Then use percent_rank to create a table of percentile ranks for each of the indicators in (a) by country.

```
my_summary_prank <- my_summary %>%
  mutate(income10to60.xtile=percent_rank(income10to60.avg),
         gdp.xtile = percent_rank(gdp.avg),
         emp_femaletomale.xtile = percent_rank(emp_femaletomale.avg),
         secondary.complet.xtile = percent_rank(secondary.complet.avg),
         disengaged.yth.xtile = percent_rank(disengaged.yth.avg),
         unemp.basic.xtile = percent_rank(unemp.basic.educ.avg))
```

c. Keep only the percentile rank columns and sort by per capita GDP percentile rank.

```
  my_summary_prank <- my_summary_prank %>%
    select(income10to60.xtile, gdp.xtile, emp_femaletomale.xtile,
           secondary.complet.xtile, disengaged.yth.xtile, unemp.basic.xtile) %>%
    arrange(desc(gdp.xtile))
```

```
## Adding missing grouping variables: `country`
```

```
  my_summary_prank[1:20, 1:5]
```

country	income10to60.xtile	gdp.xtile	emp_femaletomale.xtile	secondary.complet.xtile
Chile	NA	1	1.0000000	NA
Venezuela, RB	NA	1	0.3333333	0.2222222
Australia	NA	1	1.0000000	NA
Austria	0.5625	1	0.7500000	0.3750000

country	income10to60.xtile	gdp.xtile	emp_femaletomale.xtile	secondary.complet.xtile
Belgium	NA	1	0.9375000	NA
Canada	NA	1	0.8750000	NA
Czech Republic	NA	1	1.0000000	NA
Denmark	0.3125	1	0.4375000	0.2500000
Finland	0.4375	1	0.3750000	0.2500000
France	NA	1	1.0000000	NA
Germany	NA	1	1.0000000	NA
Greece	0.2500	1	0.4375000	0.1250000
Hong Kong SAR, China	NA	1	0.9375000	NA
Hungary	NA	1	0.5000000	NA
Ireland	NA	1	1.0000000	NA
Israel	NA	1	1.0000000	NA
Italy	0.0000	1	0.3750000	0.1875000
Japan	NA	1	1.0000000	NA
Korea, Rep.	NA	1	1.0000000	NA
Netherlands	NA	1	1.0000000	NA

9 Display basic descriptive statistics for all variables in the general data set.

```
summary(world.dev.data)
```

```
##    country               year          pcgdp       secondary.complet
##  Length:558         Min.   :2000   Min.   :12052   Min.   :15.68
##  Class :character   1st Qu.:2005   1st Qu.:22766   1st Qu.:56.94
##  Mode  :character   Median :2009   Median :40138   Median :70.27
##                     Mean   :2008   Mean   :38325   Mean   :66.09
##                     3rd Qu.:2013   3rd Qu.:48282   3rd Qu.:78.60
##                     Max.   :2016   Max.   :91617   Max.   :89.77
##                                                    NA's   :290
##  incshare_2nd20  incshare_3rd20  incshare_top10  incshare_low20
##  Min.   : 8.00   Min.   :12.70   Min.   :20.30   Min.   :1.500
##  1st Qu.:12.30   1st Qu.:16.60   1st Qu.:22.60   1st Qu.:6.900
##  Median :12.90   Median :17.20   Median :24.60   Median :7.900
##  Mean   :12.89   Mean   :17.02   Mean   :24.93   Mean   :7.792
##  3rd Qu.:13.90   3rd Qu.:17.60   3rd Qu.:26.10   3rd Qu.:8.800
##  Max.   :15.10   Max.   :19.00   Max.   :39.70   Max.   :9.900
##  NA's   :313     NA's   :313     NA's   :313     NA's   :313
##    povertyrte     emp_femaletomale unemp.basic.educ disengaged.youth
##  Min.   : 0.000   Min.   :21.14    Min.   : 0.80    Min.   : 3.40
##  1st Qu.: 0.400   1st Qu.:71.01    1st Qu.: 8.20    1st Qu.: 7.50
##  Median : 0.800   Median :77.50    Median :12.25    Median :11.10
##  Mean   : 2.123   Mean   :74.44    Mean   :13.55    Mean   :11.35
##  3rd Qu.: 2.100   3rd Qu.:82.19    3rd Qu.:16.60    3rd Qu.:13.80
##  Max.   :42.400   Max.   :90.12    Max.   :48.40    Max.   :31.60
##  NA's   :313                       NA's   :130      NA's   :189
##       pop               logGDP         numpoverty          income60
##  Min.   :  4027200   Min.   : 9.397   Min.   :0.000e+00   Min.   :23.30
##  1st Qu.:  6824175   1st Qu.:10.033   1st Qu.:4.240e+06   1st Qu.:36.00
##  Median : 10728356   Median :10.600   Median :1.445e+07   Median :37.60
##  Mean   : 43412752   Mean   :10.441   Mean   :6.271e+07   Mean   :37.69
##  3rd Qu.: 47279651   3rd Qu.:10.785   3rd Qu.:4.428e+07   3rd Qu.:40.50
```

```
##  Max.   :359479269   Max.   :11.425   Max.   :1.136e+09   Max.   :43.10
##                                       NA's   :313         NA's   :313
##   income10to60
##  Min.   :0.4822
##  1st Qu.:0.5569
##  Median :0.6532
##  Mean   :0.6762
##  3rd Qu.:0.7270
##  Max.   :1.6223
##  NA's   :313
```