ShilpiSharma_Aug_SVAP_Asmt_R2

Shilpi Sharma

10/8/2017

Domain - Employment (People)

Topic - Unemployment Analysis and Comparison at the country and gender level

```
# Loading required libraries
library(rvest)
```

```
## Loading required package: xml2
```

```
library(tidyr)
library(dplyr)
```

```
##
## 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)
library(pander)
library(tidyverse)
```

```
## Loading tidyverse: tibble
## Loading tidyverse: readr
## Loading tidyverse: purrr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(readxl)
library(stringr)
library(RColorBrewer)
library(lattice)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

Frame the questions

- Which are the top 10 countries with the highest unemployment ratio of males to females for both the timeframes 1991 and 2016?
- Which country has the highest takers of intermediate and advanced education in 2016?

Acquire the Data

Getting the unemployment data for different countries from the World Bank database

```
setwd("/Users/shilpisharma/SVAPData")
getwd()
```

```
## [1] "/Users/shilpisharma/SVAPData"
```

```
UData=read_excel("Unemployment.xls")
summary(UData)
```

```
##    Country          Male Unemployment 1991 Male Unemployment 2016
##  Length:226         Length:226             Length:226
##  Class :character   Class :character       Class :character
##  Mode  :character   Mode  :character       Mode  :character
##  Female Unemployment 1991 Female Unemployment 2016
##  Length:226               Length:226
##  Class :character         Class :character
##  Mode  :character         Mode  :character
##  Male Youth Unemployment 1991 Male Youth Unemployment 2016
##  Length:226                   Length:226
##  Class :character             Class :character
##  Mode  :character             Mode  :character
##  Female Youth Unemployment 1991 Female Youth Unemployment 2016
##  Length:226                     Length:226
##  Class :character               Class :character
##  Mode  :character               Mode  :character
##  Percent of total force with basic education
##  Length:226
##  Class :character
##  Mode  :character
##  Percent of total force with intermediate education
##  Length:226
##  Class :character
##  Mode  :character
##  Percent of total force with advanced education
##  Length:226
##  Class :character
##  Mode  :character
```

```
attach(UData)
str(UData)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    226 obs. of  12 variables:
##  $ Country                                          : chr  "Afghanistan" "Albania" "Algeria" "Americ
##  $ Male Unemployment 1991                           : chr  "1.1" "15.2" "17.6" ".." ...
##  $ Male Unemployment 2016                           : chr  "7.7" "16.5" "9.2" ".." ...
##  $ Female Unemployment 1991                         : chr  "1.8" "10.5" "42.9" ".." ...
##  $ Female Unemployment 2016                         : chr  "12.4" "16.1" "19.7" ".." ...
##  $ Male Youth Unemployment 1991                     : chr  "2.5" "37.4" "34.3" ".." ...
##  $ Male Youth Unemployment 2016                     : chr  "17" "36.8" "22.6" ".." ...
##  $ Female Youth Unemployment 1991                   : chr  "3.4" "26.3" "66.4" ".." ...
##  $ Female Youth Unemployment 2016                   : chr  "23.6" "35.7" "44.3" ".." ...
##  $ Percent of total force with basic education      : chr  ".." "13.8" ".." ".." ...
##  $ Percent of total force with intermediate education: chr  ".." "20.4" ".." ".." ...
##  $ Percent of total force with advanced education   : chr  ".." "19.1" ".." ".." ...
```

Refine the Data

- Check for Quality and Consistency
- Missing values
- Outlier treatment
- Remove unneeded
- Format Data Types

Changing Column Names to shorter names

```
dim(UData)
```

```
## [1] 226  12
```

```
column_name <- c('country', 'maleUnempt91', 'maleUnempt16', 'femaleUnempt91', 'femaleUnempt16', 'maleYou
colnames(UData) <- column_name
str(UData)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    226 obs. of  12 variables:
##  $ country            : chr  "Afghanistan" "Albania" "Algeria" "American Samoa" ...
##  $ maleUnempt91       : chr  "1.1" "15.2" "17.6" ".." ...
##  $ maleUnempt16       : chr  "7.7" "16.5" "9.2" ".." ...
##  $ femaleUnempt91     : chr  "1.8" "10.5" "42.9" ".." ...
##  $ femaleUnempt16     : chr  "12.4" "16.1" "19.7" ".." ...
##  $ maleYouthUnempt91  : chr  "2.5" "37.4" "34.3" ".." ...
##  $ maleYouthUnempt16  : chr  "17" "36.8" "22.6" ".." ...
##  $ femaleYouthUnempt91: chr  "3.4" "26.3" "66.4" ".." ...
##  $ femaleYouthUnempt16: chr  "23.6" "35.7" "44.3" ".." ...
##  $ basicEduPct        : chr  ".." "13.8" ".." ".." ...
##  $ intermediateEduPct : chr  ".." "20.4" ".." ".." ...
##  $ advancedEduPct     : chr  ".." "19.1" ".." ".." ...
```

Change the data types from chr to numeric and date

```
head(UData)
```

```
## # A tibble: 6 x 12
##          country maleUnempt91 maleUnempt16 femaleUnempt91 femaleUnempt16
##            <chr>        <chr>        <chr>          <chr>          <chr>
## 1    Afghanistan          1.1          7.7            1.8           12.4
## 2        Albania         15.2         16.5           10.5           16.1
## 3        Algeria         17.6          9.2           42.9           19.7
## 4 American Samoa           ..           ..             ..             ..
## 5        Andorra           ..           ..             ..             ..
## 6         Angola          6.3          6.2            7.2            7.1
## # ... with 7 more variables: maleYouthUnempt91 <chr>,
## #   maleYouthUnempt16 <chr>, femaleYouthUnempt91 <chr>,
## #   femaleYouthUnempt16 <chr>, basicEduPct <chr>,
## #   intermediateEduPct <chr>, advancedEduPct <chr>
```

```
tail(UData)
```

```
## # A tibble: 6 x 12
##               country maleUnempt91 maleUnempt16 femaleUnempt91
##                 <chr>        <chr>        <chr>          <chr>
## 1          South Asia          3.7          3.6            4.9
## 2  Sub-Saharan Africa          7.1          6.4            9.1
## 3          Low income          4.7          4.8            6.4
## 4 Lower middle income          4.6          4.5            5.9
## 5 Upper middle income          6.7          6.3            6.4
## 6         High income          6.3          6.1            7.5
## # ... with 8 more variables: femaleUnempt16 <chr>,
## #   maleYouthUnempt91 <chr>, maleYouthUnempt16 <chr>,
## #   femaleYouthUnempt91 <chr>, femaleYouthUnempt16 <chr>,
## #   basicEduPct <chr>, intermediateEduPct <chr>, advancedEduPct <chr>
```

```
# Coercing all the columns datatype except Country from chr to numeric -> NAs introduced
UData$maleUnempt91 <- as.numeric(UData$maleUnempt91)
```

```
## Warning: NAs introduced by coercion
```

```
UData$maleUnempt16 <- as.numeric(UData$maleUnempt16)
```

```
## Warning: NAs introduced by coercion
```

```
UData$femaleUnempt91 <- as.numeric(UData$femaleUnempt91)
```

```
## Warning: NAs introduced by coercion
```

```
UData$femaleUnempt16 <- as.numeric(UData$femaleUnempt16)
```

```
## Warning: NAs introduced by coercion
```

```
UData$maleYouthUnempt91 <- as.numeric(UData$maleYouthUnempt91)
```

```
## Warning: NAs introduced by coercion
```

```
UData$maleYouthUnempt16 <- as.numeric(UData$maleYouthUnempt16)
```

```
## Warning: NAs introduced by coercion
```

```
UData$femaleYouthUnempt91 <- as.numeric(UData$femaleYouthUnempt91)
```

```
## Warning: NAs introduced by coercion
```

```
UData$femaleYouthUnempt16 <- as.numeric(UData$femaleYouthUnempt16)
```

```
## Warning: NAs introduced by coercion
```

```
UData$basicEduPct <- as.numeric(UData$basicEduPct)
```

```
## Warning: NAs introduced by coercion
```

```
UData$intermediateEduPct <- as.numeric(UData$intermediateEduPct)
```

```
## Warning: NAs introduced by coercion
```

```
UData$advancedEduPct <- as.numeric(UData$advancedEduPct)
```

```
## Warning: NAs introduced by coercion
```

```
tail(UData)
```

```
## # A tibble: 6 x 12
##               country maleUnempt91 maleUnempt16 femaleUnempt91
##                 <chr>        <dbl>        <dbl>          <dbl>
## 1          South Asia          3.7          3.6            4.9
## 2  Sub-Saharan Africa          7.1          6.4            9.1
## 3          Low income          4.7          4.8            6.4
## 4 Lower middle income          4.6          4.5            5.9
## 5 Upper middle income          6.7          6.3            6.4
## 6         High income          6.3          6.1            7.5
## # ... with 8 more variables: femaleUnempt16 <dbl>,
## #   maleYouthUnempt91 <dbl>, maleYouthUnempt16 <dbl>,
## #   femaleYouthUnempt91 <dbl>, femaleYouthUnempt16 <dbl>,
## #   basicEduPct <dbl>, intermediateEduPct <dbl>, advancedEduPct <dbl>
```

Filter all the rows except the last four rows

```
df <- UData %>% filter(row_number() < 223)
tail(df)
```

```
## # A tibble: 6 x 12
##                       country maleUnempt91 maleUnempt16 femaleUnempt91
##                         <chr>        <dbl>        <dbl>          <dbl>
## 1      Europe & Central Asia          8.7          8.2            9.9
## 2  Latin America & Caribbean          6.6          6.7           10.4
## 3 Middle East & North Africa         10.3          8.9           21.4
## 4              North America          7.5          5.3            6.7
## 5                 South Asia          3.7          3.6            4.9
## 6         Sub-Saharan Africa          7.1          6.4            9.1
## # ... with 8 more variables: femaleUnempt16 <dbl>,
## #   maleYouthUnempt91 <dbl>, maleYouthUnempt16 <dbl>,
## #   femaleYouthUnempt91 <dbl>, femaleYouthUnempt16 <dbl>,
## #   basicEduPct <dbl>, intermediateEduPct <dbl>, advancedEduPct <dbl>
```

```
# Removing the special characters in Korea country field
df <- df %>% within(country[str_detect(country, 'Korea, Dem+')] <- 'Korea Dem.')
```

```
tail(df)
```

```
## # A tibble: 6 x 12
##                      country maleUnempt91 maleUnempt16 femaleUnempt91
##                        <chr>        <dbl>        <dbl>          <dbl>
## 1      Europe & Central Asia          8.7          8.2            9.9
## 2  Latin America & Caribbean          6.6          6.7           10.4
## 3 Middle East & North Africa         10.3          8.9           21.4
## 4              North America          7.5          5.3            6.7
## 5                 South Asia          3.7          3.6            4.9
## 6         Sub-Saharan Africa          7.1          6.4            9.1
## # ... with 8 more variables: femaleUnempt16 <dbl>,
## #   maleYouthUnempt91 <dbl>, maleYouthUnempt16 <dbl>,
## #   femaleYouthUnempt91 <dbl>, femaleYouthUnempt16 <dbl>,
## #   basicEduPct <dbl>, intermediateEduPct <dbl>, advancedEduPct <dbl>
```

Transform the Data

First Question - Finding the top 10 countries with the highest unemployment ratio of males to females

```
# Calculating for 1991 data
uemptRatio91 = df$maleUnempt91/df$femaleUnempt91
df <- cbind(df, uemptRatio91)
str(df)
```

```
## 'data.frame':    222 obs. of  13 variables:
##  $ country            : chr  "Afghanistan" "Albania" "Algeria" "American Samoa" ...
##  $ maleUnempt91       : num  1.1 15.2 17.6 NA NA 6.3 NA 5.4 17.5 NA ...
##  $ maleUnempt16       : num  7.7 16.5 9.2 NA NA 6.2 NA 5.8 15 NA ...
##  $ femaleUnempt91     : num  1.8 10.5 42.9 NA NA 7.2 NA 6.5 20.9 NA ...
##  $ femaleUnempt16     : num  12.4 16.1 19.7 NA NA 7.1 NA 7.7 18.7 NA ...
##  $ maleYouthUnempt91  : num  2.5 37.4 34.3 NA NA 10.8 NA 10.6 35.4 NA ...
##  $ maleYouthUnempt16  : num  17 36.8 22.6 NA NA 10.5 NA 14.3 31.5 NA ...
##  $ femaleYouthUnempt91: num  3.4 26.3 66.4 NA NA 12.4 NA 12.2 46.9 NA ...
##  $ femaleYouthUnempt16: num  23.6 35.7 44.3 NA NA 11.9 NA 19 44.9 NA ...
##  $ basicEduPct        : num  NA 13.8 NA NA NA NA NA 8.3 15.1 NA ...
##  $ intermediateEduPct : num  NA 20.4 NA NA NA NA NA 6.8 19.1 NA ...
##  $ advancedEduPct     : num  NA 19.1 NA NA NA NA NA 5.5 17.9 NA ...
##  $ uemptRatio91       : num  0.611 1.448 0.41 NA NA ...
```

```
dfUemptRatio91 <- df %>% arrange(desc(uemptRatio91)) %>% head(10)

# Calculating for 2016 data
uemptRatio16 = df$maleUnempt16/df$femaleUnempt16
df <- cbind(df, uemptRatio16)
str(df)
```

```
## 'data.frame':    222 obs. of  14 variables:
##  $ country            : chr  "Afghanistan" "Albania" "Algeria" "American Samoa" ...
##  $ maleUnempt91       : num  1.1 15.2 17.6 NA NA 6.3 NA 5.4 17.5 NA ...
##  $ maleUnempt16       : num  7.7 16.5 9.2 NA NA 6.2 NA 5.8 15 NA ...
##  $ femaleUnempt91     : num  1.8 10.5 42.9 NA NA 7.2 NA 6.5 20.9 NA ...
##  $ femaleUnempt16     : num  12.4 16.1 19.7 NA NA 7.1 NA 7.7 18.7 NA ...
```

```
##  $ maleYouthUnempt91  : num  2.5 37.4 34.3 NA NA 10.8 NA 10.6 35.4 NA ...
##  $ maleYouthUnempt16  : num  17 36.8 22.6 NA NA 10.5 NA 14.3 31.5 NA ...
##  $ femaleYouthUnempt91: num  3.4 26.3 66.4 NA NA 12.4 NA 12.2 46.9 NA ...
##  $ femaleYouthUnempt16: num  23.6 35.7 44.3 NA NA 11.9 NA 19 44.9 NA ...
##  $ basicEduPct        : num  NA 13.8 NA NA NA NA NA 8.3 15.1 NA ...
##  $ intermediateEduPct : num  NA 20.4 NA NA NA NA NA 6.8 19.1 NA ...
##  $ advancedEduPct     : num  NA 19.1 NA NA NA NA NA 5.5 17.9 NA ...
##  $ uemptRatio91       : num  0.611 1.448 0.41 NA NA ...
##  $ uemptRatio16       : num  0.621 1.025 0.467 NA NA ...
```

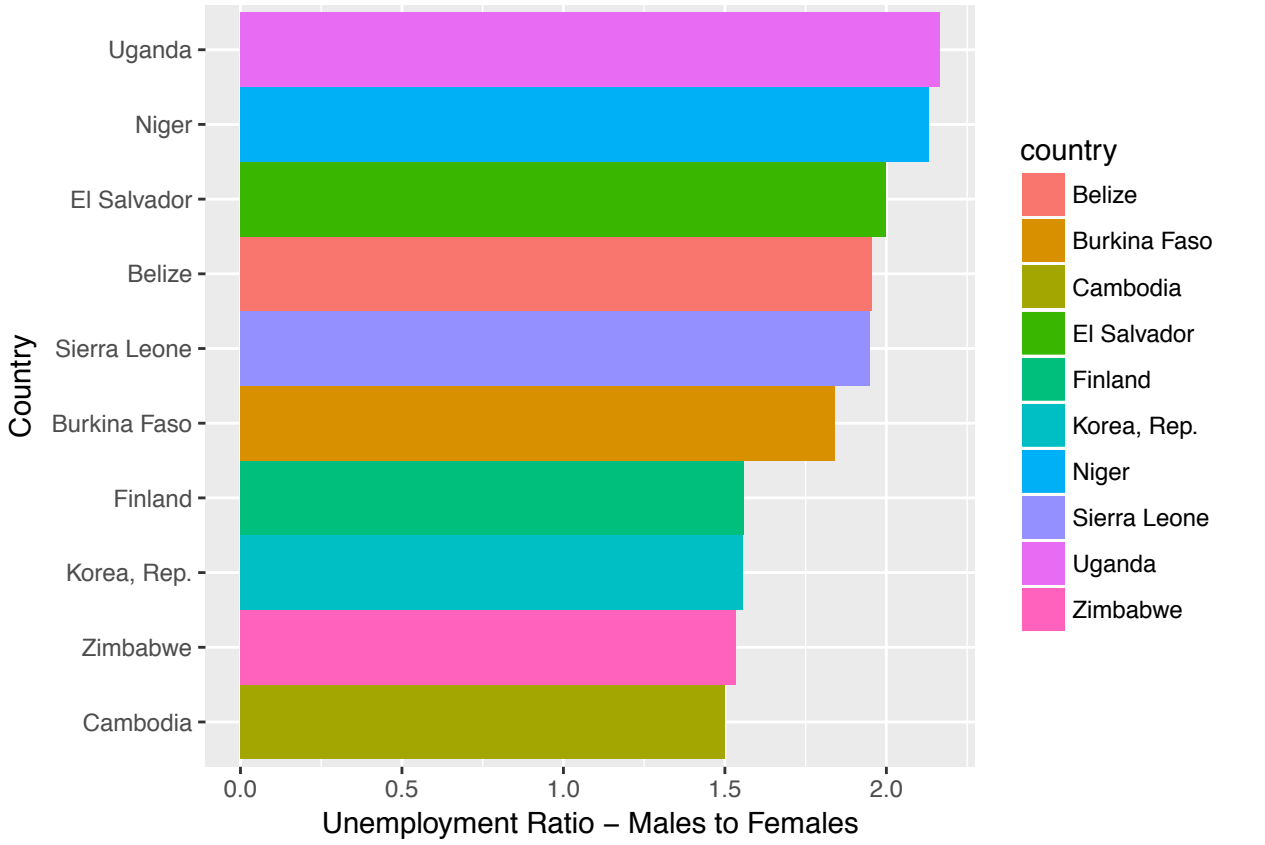
```r
dfUemptRatio91 <- df %>% arrange(desc(uemptRatio91)) %>% head(10)
dfUemptRatio16 <- df %>% arrange(desc(uemptRatio16)) %>% head(10)
```

Solution : Uganda had the highest unemployment ratio of males to females in 1991 but it moved down to spot 10 in 2016, whereas Sierra has the highest unemployement ratio in 2016.

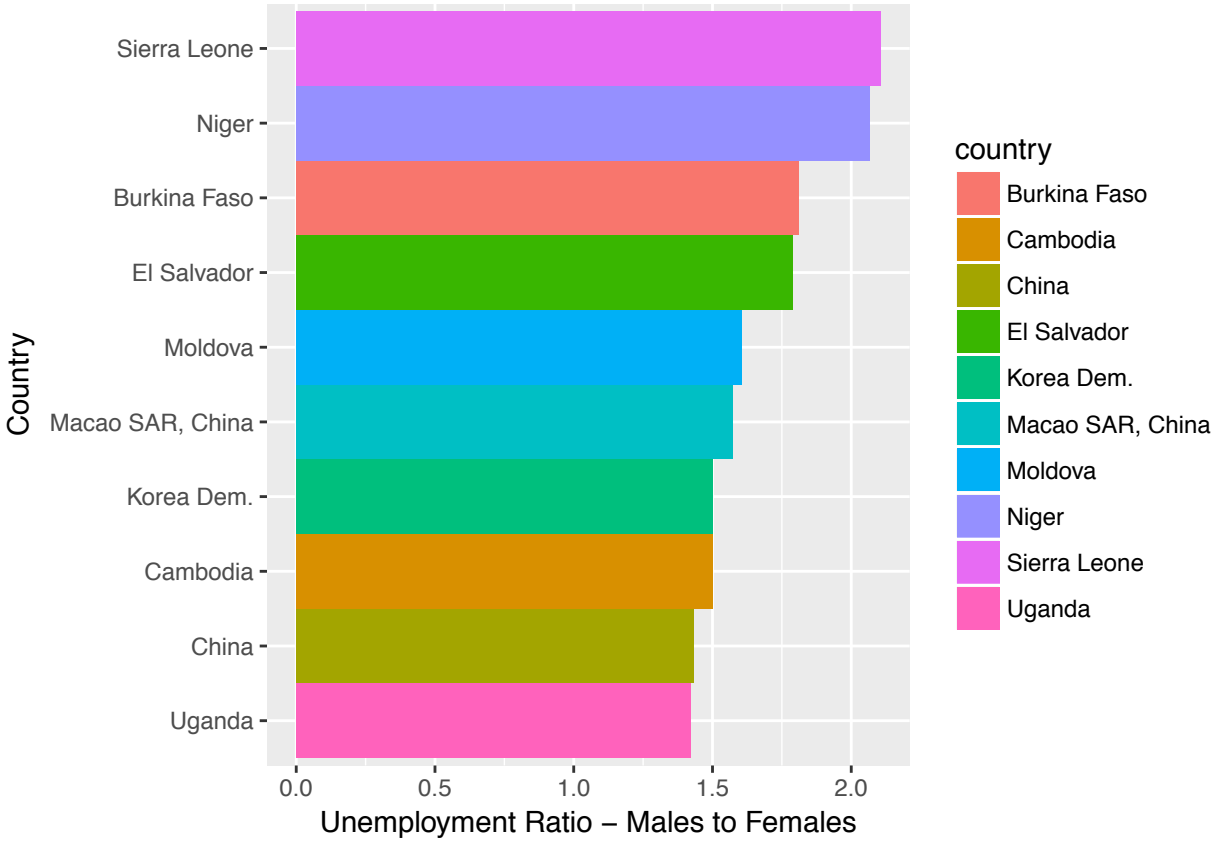
Explore - Visualize

Depicting through Bar Graph

```r
# Depicting 1991 data through Bar Graph
ggplot(dfUemptRatio91) +
  aes(reorder(country, uemptRatio91), uemptRatio91, fill=country) +
  geom_col(width = 1) + xlab("Country") + ylab("Unemployment Ratio - Males to Females") +
  coord_flip()
```

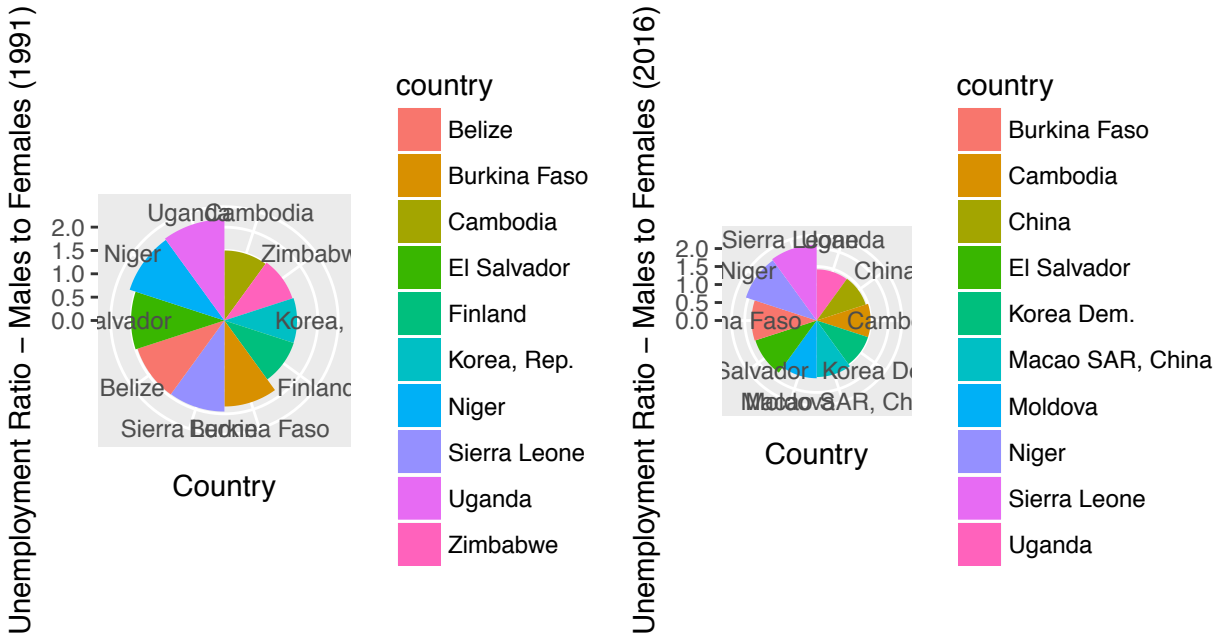


```
# Depicting 2016 data through Bar Graph
ggplot(dfUemptRatio16) +
  aes(reorder(country, uemptRatio16), uemptRatio16, fill=country) +
  geom_col(width = 1) + xlab("Country") + ylab("Unemployment Ratio - Males to Females") +
  coord_flip()
```



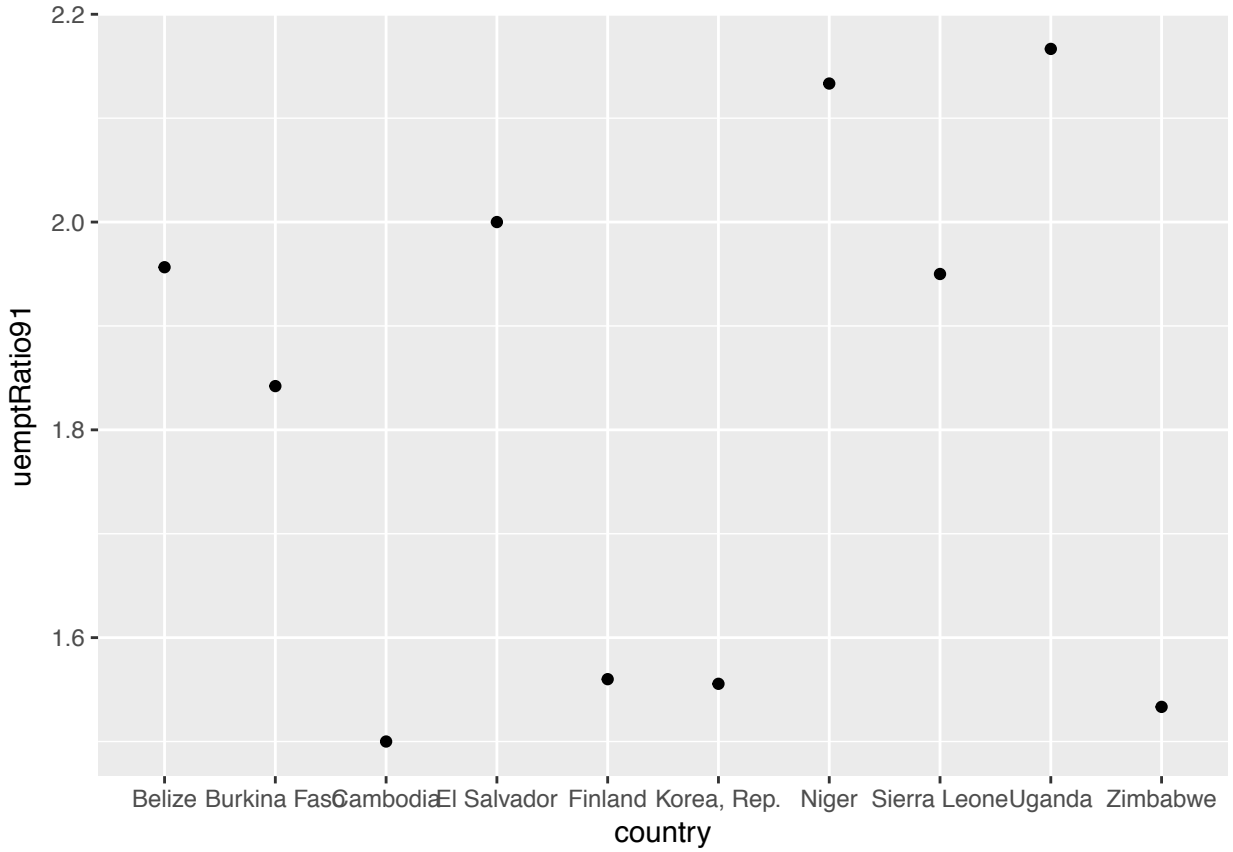
Depicting through pie charts

```
# Plotting 1991 data on a pie chart
plot1 <- ggplot(dfUemptRatio91, facets = ~mygroup) +
  aes(reorder(country, uemptRatio91), uemptRatio91, fill=country) +
  geom_col(width = 1) + xlab("Country") + ylab("Unemployment Ratio - Males to Females (1991)") +
  coord_flip() +
  coord_polar()

# Plotting 2016 data on a pie chart
plot2 <- ggplot(dfUemptRatio16, facets = ~mygroup) +
  aes(reorder(country, uemptRatio16), uemptRatio16, fill=country) +
  geom_col(width = 1) + xlab("Country") + ylab("Unemployment Ratio - Males to Females (2016)") +
  coord_flip() +
  coord_polar()

grid.arrange(plot1, plot2, ncol=2)
```

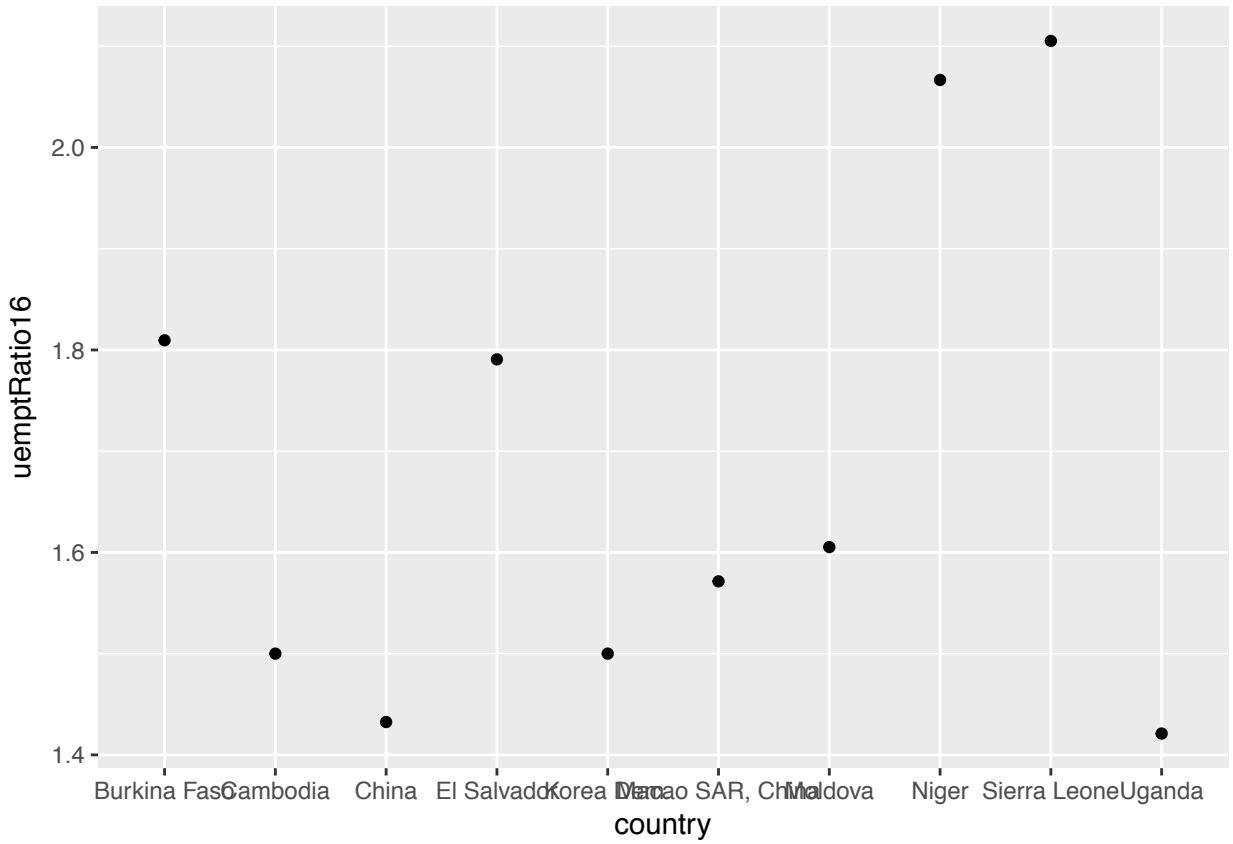


Depicting both the 1991 and 2016 unemployment ratio data through same Scatter Plot

```
# Depicting 1991 data through Scatter Plot
ggplot(dfUemptRatio91) +
  aes(country, uemptRatio91) +
  geom_point()
```




```
# Depicting 2016 data through Scatter Plot
ggplot(dfUemptRatio16) +
  aes(country, uemptRatio16) +
  geom_point()
```



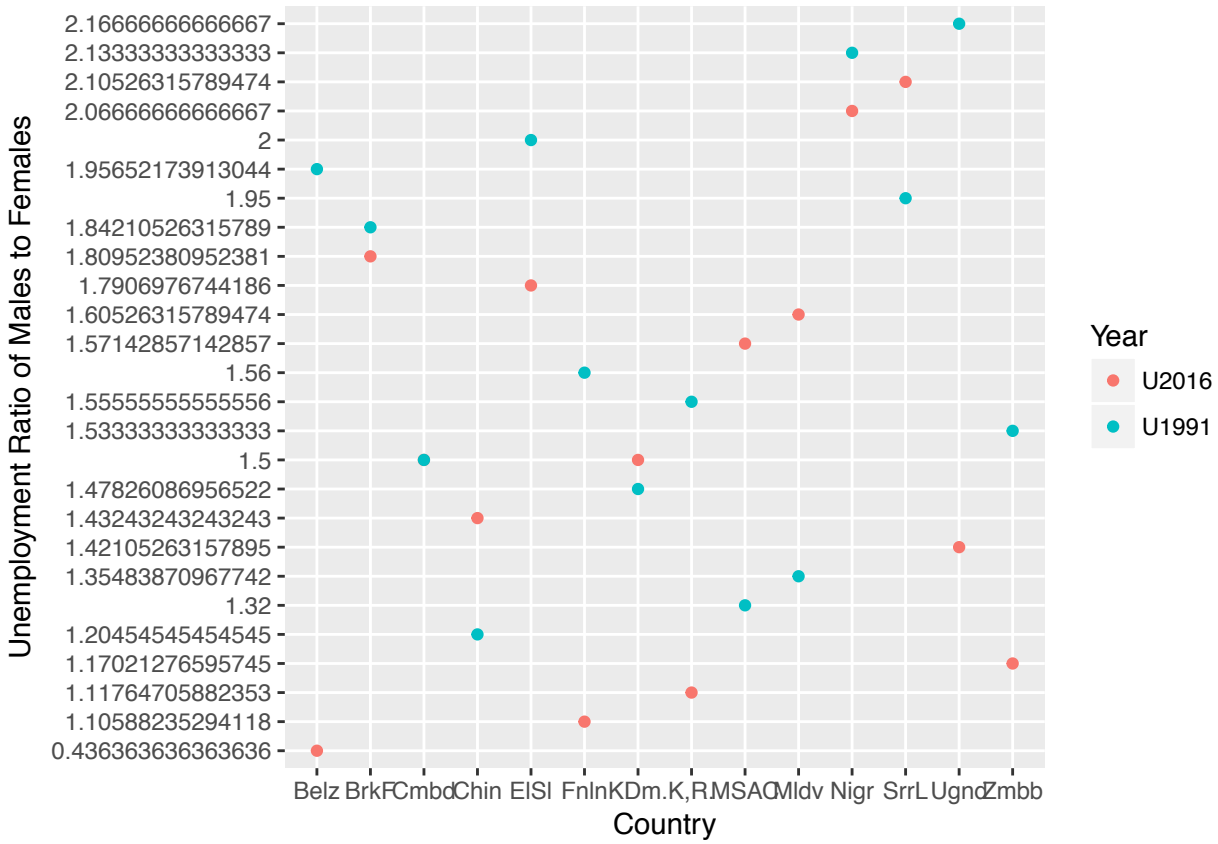

```
# Depicting both timeframes in the same scatter plot
# Combining the 1991 and 2016 ratio data frames
d <- rbind(dfUemptRatio91, dfUemptRatio16)
# Removing duplicate rows
d <- unique(d)
# Extracting the columns of interest
d1 <- as.data.frame(cbind(country=d$country, U2016=d$uemptRatio16, U1991=d$uemptRatio91))
str(d1)
```

```
## 'data.frame':    14 obs. of  3 variables:
##  $ country: Factor w/ 14 levels "Belize","Burkina Faso",..: 13 11 5 1 12 2 6 8 14 3 ...
##  $ U2016  : Factor w/ 13 levels "0.436363636363636",..: 5 12 10 1 13 11 2 3 4 7 ...
##  $ U1991  : Factor w/ 14 levels "1.20454545454545",..: 14 13 12 11 10 9 8 7 6 5 ...
```

```
# Transposing the data from wide to long format
df2 <- melt(data = d1, id = "country")
```

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```

```
# Renaming the variable column name to Year
colnames(df2)[colnames(df2) == 'variable'] <- 'Year'
# Plotting on a scatter plot
ggplot(data = df2, aes(x = country, y = value, colour = Year, group = 1)) + geom_point() + xlab("Country
```



Second Question - Which country has the highest takers of intermediate and advanced education in 2016?

Stacked Bar Plot with Colors and Legend

```
# Extracting the required education indicator columns from the cleansed data frame
percents <- as.data.frame(cbind(country = df$country, intermediateEducation = df$intermediateEduPct, ad
# Removing the NAs
percents <- subset(percents, !is.na(intermediateEducation) & !is.na(advancedEducation))
# Getting the top 10 intermediate education takers
topIntermediate <- percents %>% arrange(desc(intermediateEducation)) %>% head(10)
# Getting the top 10 advanced education takers
topAdvanced <- percents %>% arrange(desc(advancedEducation)) %>% head(10)
# Combining the Intermediate top ten and Advanced top ten data
combinedEduData <- unique(rbind(topIntermediate, topAdvanced))
str(combinedEduData)
```

```
## 'data.frame':    19 obs. of  3 variables:
##  $ country             : Factor w/ 222 levels "Afghanistan",..: 27 176 19 134 158 29 35 107 109 58
##  $ intermediateEducation: Factor w/ 64 levels "0.4","1.3","1.8",..: 64 64 63 62 61 60 60 59 58 57 ..
##  $ advancedEducation    : Factor w/ 61 levels "0.6","1.3","1.6",..: 41 48 41 51 35 35 45 54 40 45 ..
```

```
summary(combinedEduData)
```

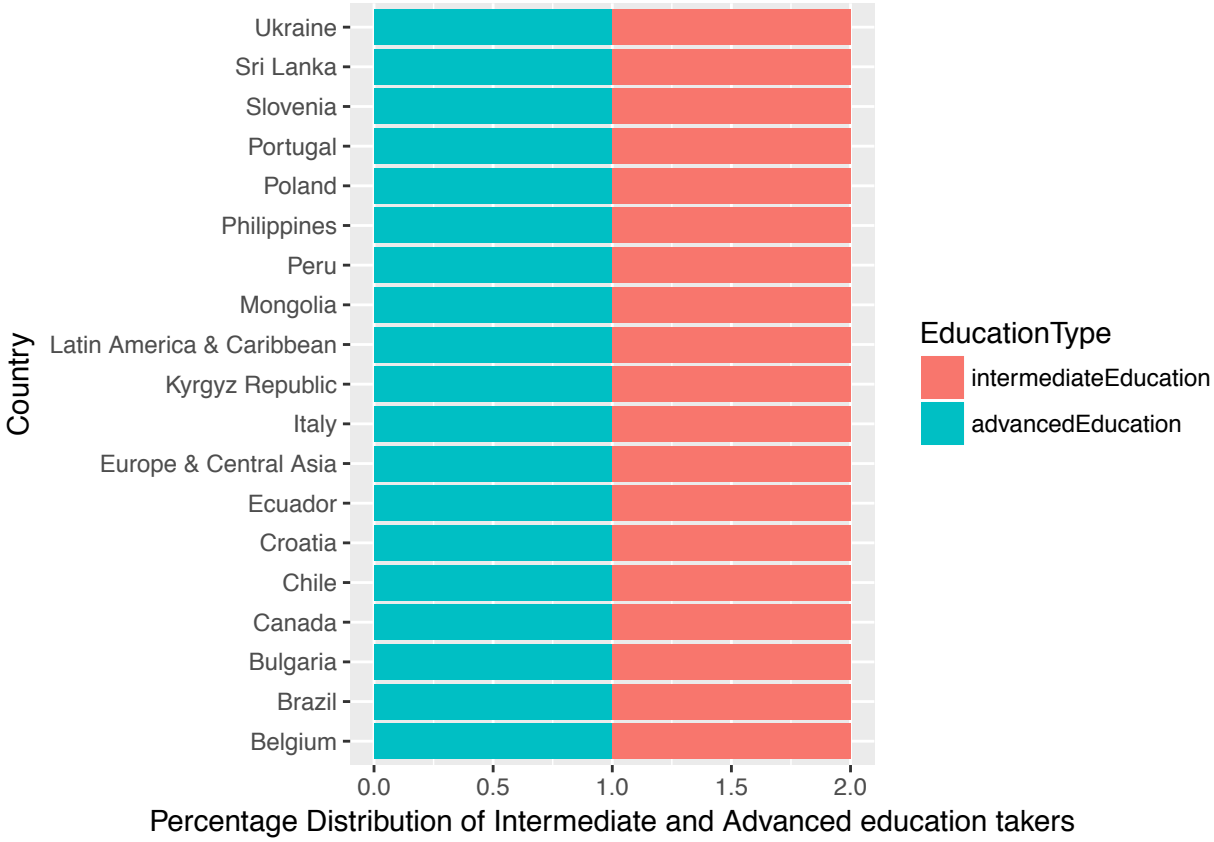
```
##      country   intermediateEducation advancedEducation
##  Belgium : 1   8.3    : 2            4      :2
```

```
##  Brazil   : 1   9.9    : 2           4.6    :2
##  Bulgaria: 1   10.1   : 1           5.5    :2
##  Canada   : 1   10.3   : 1           8.4    :2
##  Chile    : 1   11.4   : 1           4.5    :1
##  Croatia  : 1   12     : 1           5.8    :1
##  (Other) :13   (Other):11           (Other):9
```

```
# Depicting through stacked bar chart
# Transposing the data from wide to long format
df3 <- melt(data = combinedEduData, id = "country")
```

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```

```
# Renaming the variable column name to Education Type
colnames(df3)[colnames(df3) == 'variable'] <- 'EducationType'
ggplot(df3, aes(x=country)) + geom_bar(aes(fill = EducationType)) + xlab("Country") +
  ylab("Percentage Distribution of Intermediate and Advanced education takers") + coord_flip()
```



Conclusion - Insights gained

- Uganda has the highest disproportion between males and females with males being more unemployed in 1991 timeframe.
- Uganda reduced this ratio from 2.16 to 1.42 in 2016.
- Sierra Leone had the highest unemployment ratio of males to females in 2016.
- Brazil has highest number of people completing intermediate education.
- Croatia has the highest number of people completing the advanced education.