Data Quality Assignment

Gilbert Habaasa

2025-04-04

IPUMS Exercise

IPUMS International Data Extract and Analysis

Data Quality, European Doctoral School of Demography 2024-2025

INED – Paris (France)

Instructor: Mariona Lozano, Centre d'Estudis Demogràfics

Student: Gilbert Habaasa

Getting IPUMS Data into R

```
options(repos = c(CRAN = "https://cloud.r-project.org"))
install.packages("ipumsr")
```

```
## Installing package into 'C:/Users/admin/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)
```

```
## package 'ipumsr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\admin\AppData\Local\Temp\RtmpeeSZlD\downloaded_packages
```

```
setwd("C:\\Users\\admin\\OneDrive - London School of Hygiene and Tropical Medicine\\INED 2024\\Data Qual
# file.exists("C:\\Users\\admin\\OneDrive - London School of Hygiene and Tropical Medicine\\INED 2024\\

library(ipumsr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(readr)
ddi <- read_ipums_ddi("ipumsi_00001.xml")
data <- read_ipums_micro(ddi)
```

```
## Use of data from IPUMS International is subject to conditions including that users should cite the da
```

```r
data
```

```
## # A tibble: 12,596,631 x 15
##    COUNTRY    YEAR SAMPLE       SERIAL  HHWT URBAN   PERNUM PERWT RESIDENT AGE
##    <int+lbl> <int> <int+lbl>     <dbl> <dbl> <int+l>  <dbl> <dbl> <int+lb> <int>
##  1 484 [Mex~  2000 4.84e8 [Mex~   1001     3 2 [Urb~      1     3 NA       41
##  2 484 [Mex~  2000 4.84e8 [Mex~   1001     3 2 [Urb~      2     3 NA       37
##  3 484 [Mex~  2000 4.84e8 [Mex~   1001     3 2 [Urb~      3     3 NA       14
##  4 484 [Mex~  2000 4.84e8 [Mex~   1001     3 2 [Urb~      4     3 NA       19
##  5 484 [Mex~  2000 4.84e8 [Mex~   1002     3 2 [Urb~      1     3 NA       19
##  6 484 [Mex~  2000 4.84e8 [Mex~   1002     3 2 [Urb~      2     3 NA       17
##  7 484 [Mex~  2000 4.84e8 [Mex~   2001     3 2 [Urb~      1     3 NA       77
##  8 484 [Mex~  2000 4.84e8 [Mex~   2001     3 2 [Urb~      2     3 NA       30
##  9 484 [Mex~  2000 4.84e8 [Mex~   2001     3 2 [Urb~      3     3 NA       34
## 10 484 [Mex~  2000 4.84e8 [Mex~   3001     3 2 [Urb~      1     3 NA       39
## # i 12,596,621 more rows
## # i 5 more variables: SEX <int+lbl>, LIT <int+lbl>, EMPSTAT <int+lbl>,
## #   EMPSTATD <int+lbl>, OCCISCO <int+lbl>
```

Question and Variables

In this exercise you will gain basic familiarity with the IPUMS International data exploration and extract system to answer the following question: "What are the differences in urbanization, literacy, and occupational participation in Uganda and Mexico?" You will create a data extract that includes the following variables: **URBAN, SEX, EMPSTAT, OCCISCO, LIT, AGE**.

Variables and Code

URBAN: household location 1 = Rural 2 = Urban 9 = Unknown

SEX 1 = Male 2 = Female 9 = Missing/blank

URBAN: EMPSTAT: Employment status 1 = Employed 2 = Unemployed 3 = Not in labor force 9 = Unknown/Illegible

OCCISCO: Employment category 01 = Legislators, senior officials and managers 02 = Professionals 03 = Technicians and associate professionals 04 = Clerks 05 = Service workers and shop and market sales 06 = Skilled agricultural and fishery workers 07 = Crafts and related trades workers 08 = Plant and machine operators and assemblers 09 = Elementary occupations 10 = Armed forces 11 = Other occupations, unspecified or n.e.c. 97 = Response suppressed 98 = Unknown 99 = NIU (not in universe)

LIT: Literacy 0 = NIU (not in universe) 1 = No, illiterate 2 = Yes, literate 9 = Unknown, illegible or blank

AGE 000 = Less than 1 year old 001 = 1 … = … 140 = 140 999 = Missing

COUNTRY 484 = Mexico 800 = Uganda

Analyse the Data

Part 1: Variable documentation

For each variable below, search through the tabbed sections of the variable description to answer each question.

1.Under the "Household" dropdown menu, find the "Geography" subcategory and click on the variable URBAN. What constitutes an urban area in each country? a.Mexico 2000:

Urban places are defined consistently across Mexican samples as localities with 2,500 or more persons.

b.Uganda 2002:

Urban areas in 2002 and 2014 are gazetted cities, municipalities and towns with more than 2,000 inhabitants.

2.What are the codes for URBAN?

1.Rural; 2.Urban; 9.Unknown.

3.Find the variable EMPSTAT. Is the reference period of work the same for Mexico and Uganda?

Mexico-Last week

Uganda-Last seven days

4.What is the universe for EMPSTAT in:

a.Mexico 2000?

Mexico-Persons age 12+

b.Uganda 2002?

Uganda-Persons age 5+

Part 2. Frequencies

5.Find codes page for the SAMPLE variable. What are the codes for:

a.Mexico 2000?

Value Code for Mexico-484

b.Uganda 2002?

Value Code for Uganda-800

```r
unique(data$COUNTRY)
```

```
## <labelled<integer>[2]>: Country
## [1] 484 800
##
## Labels:
##  value              label
##     32          Argentina
##     40            Austria
##     50         Bangladesh
##     51            Armenia
##     68            Bolivia
##     72            Botswana
##     76              Brazil
##    104             Myanmar
##    112             Belarus
##    116            Cambodia
##    120            Cameroon
##    124              Canada
##    152               Chile
##    156               China
##    170            Colombia
##    188          Costa Rica
##    192                Cuba
##    204               Benin
##    208             Denmark
##    214  Dominican Republic
##    218             Ecuador
##    222         El Salvador
##    231            Ethiopia
##    242                Fiji
##    246             Finland
##    250              France
##    275           Palestine
##    276             Germany
##    288               Ghana
##    300              Greece
##    320           Guatemala
##    324              Guinea
##    332               Haiti
##    340            Honduras
##    348             Hungary
##    352             Iceland
##    356               India
##    360           Indonesia
##    364                Iran
##    368                Iraq
##    372             Ireland
##    376              Israel
##    380               Italy
##    384       Côte d’Ivoire
##    388             Jamaica
##    400              Jordan
##    404               Kenya
##    417     Kyrgyz Republic
##    418                Laos
##    426             Lesotho
##    430             Liberia
##    454              Malawi
##    458            Malaysia
##    466                Mali
##    480           Mauritius
##    484              Mexico
##    496            Mongolia
##    504             Morocco
##    508          Mozambique
##    524              Nepal
##    528        Netherlands
##    558          Nicaragua
##    566            Nigeria
##    578             Norway
##    586           Pakistan
##    591             Panama
##    598   Papua New Guinea
##    600           Paraguay
##    604               Peru
##    608        Philippines
##    616             Poland
##    620           Portugal
##    630        Puerto Rico
##    642            Romania
##    643             Russia
##    646             Rwanda
##    662        Saint Lucia
##    686            Senegal
##    694       Sierra Leone
##    703    Slovak Republic
##    704            Vietnam
##    705           Slovenia
##    710       South Africa
##    716           Zimbabwe
##    724              Spain
##    728        South Sudan
##    729              Sudan
##    740           Suriname
##    752             Sweden
##    756        Switzerland
##    764           Thailand
##    768               Togo
##    780 Trinidad and Tobago
##    792             Turkey
##    800             Uganda
##    804            Ukraine
##    818              Egypt
##    826     United Kingdom
##    834           Tanzania
##    840      United States
##    854       Burkina Faso
##    858            Uruguay
##    862          Venezuela
##    894             Zambia
```

6.How many individuals are in the Mexico 2000 sample extract?

There are 10099182 individuals in the Mexico 2000 sample extract

```
data |> filter(COUNTRY==484) |> count()
```

```
## # A tibble: 1 x 1
##           n
##        <int>
## 1 10099182
```

7.How many individuals are in the Uganda 2002 sample extract?

There are 2497449 individuals in the Uganda 2002 sample extract

```
data |> filter(COUNTRY==800) |> count()
```

```
## # A tibble: 1 x 1
##         n
##     <int>
## 1 2497449
```

8.What proportion of individuals in the sample lived in urban areas in each country?

a.Mexico 2000:

In Mexico, 59.2% of individuals in the 2000 sample lived in urban areas.

b.Uganda 2002:

In Uganda, 12.3% of individuals in the 2002 sample lived in urban areas.

```
100*prop.table(table(data$URBAN,data$COUNTRY),2)
```

```
##
##           484      800
##   1 40.81933 87.74534
##   2 59.18067 12.25466
```

Part 3. Weighted frequencies

To get a more accurate estimate of the actual proportion of individuals living in urban areas, you will have to use the person weight.

9.Using weights, what is the total population of each country?

a.Mexico 2000:

97014867 people

b.Uganda 2002:

24974490 people

```
data |> filter(COUNTRY==484) |> summarise(total_population=sum(PERWT))
```

```
## # A tibble: 1 x 1
##   total_population
##              <dbl>
## 1         97014867
```

```
data |> filter(COUNTRY==800) |> summarise(total_population=sum(PERWT))
```

```
## # A tibble: 1 x 1
##   total_population
##              <dbl>
## 1         24974490
```

10.Using weights, how many individuals lived in urban areas in each country?

a.Mexico 2000:

72409464 people

b.Uganda 2002:

3060540 people

```
data |> group_by (COUNTRY, URBAN) %>% summarise(living_in_urban=sum(PERWT))
```

```
## ‘summarise()‘ has grouped output by ’COUNTRY’. You can override using the
## ‘.groups‘ argument.
```

```
## # A tibble: 4 x 3
## # Groups:   COUNTRY [2]
##   COUNTRY      URBAN     living_in_urban
##   <int+lbl>    <int+lbl>           <dbl>
## 1 484 [Mexico] 1 [Rural]        24605403
## 2 484 [Mexico] 2 [Urban]        72409464
## 3 800 [Uganda] 1 [Rural]        21913950
## 4 800 [Uganda] 2 [Urban]         3060540
```

11.Using weights, what proportion of individuals lived in urban areas in each country?

a.Mexico: 2000:

Using weights, 74.6% of individuals in the Mexico 2000 sample lived in urban areas.

b.Uganda 2002:

Using weights, 12.3% of individuals in the Uganda 2002 sample lived in urban areas.

```
data |>
  group_by(COUNTRY) |>
  summarise(prop_urban = sum(PERWT[URBAN == 2]) / sum(PERWT) * 100)
```

```
## # A tibble: 2 x 2
##   COUNTRY      prop_urban
##   <int+lbl>         <dbl>
## 1 484 [Mexico]       74.6
## 2 800 [Uganda]       12.3
```

Part 4. Trends

12.Using weights, which occupational category has the highest percentage of workers?

a.In Mexico 2000:

Crafts and related trades workers (17.9%)

b.In Uganda 2002:

Skilled agricultural and fishery workers (71.4%)

```
data |> filter(OCCISCO != 98 & OCCISCO != 99 ) |>
  group_by(COUNTRY, OCCISCO) |>
  summarise(weighted_n = sum(PERWT, na.rm = TRUE), .groups = "drop") |>
  group_by(COUNTRY) |>
  mutate(perc = weighted_n / sum(weighted_n) * 100)|>
  arrange(desc(perc))  |> print(n=50)
```

```
## # A tibble: 21 x 4
## # Groups:   COUNTRY [2]
##    COUNTRY      OCCISCO                                       weighted_n    perc
##    <int+lbl>    <int+lbl>                                          <dbl>   <dbl>
##  1 800 [Uganda]  6 [Skilled agricultural and fishery workers]    5360990 71.4
##  2 484 [Mexico]  7 [Crafts and related trades workers]           6293986 17.9
##  3 484 [Mexico]  5 [Service workers and shop and market sale~    6166733 17.5
##  4 484 [Mexico]  6 [Skilled agricultural and fishery workers]    5675532 16.1
##  5 484 [Mexico]  9 [Elementary occupations]                      4976531 14.2
##  6 484 [Mexico]  8 [Plant and machine operators and assemble~    3507871  9.98
##  7 484 [Mexico]  4 [Clerks]                                      3101337  8.82
##  8 800 [Uganda]  5 [Service workers and shop and market sale~     644650  8.59
##  9 484 [Mexico]  2 [Professionals]                               2888521  8.22
## 10 800 [Uganda]  7 [Crafts and related trades workers]            410990  5.47
## 11 800 [Uganda]  3 [Technicians and associate professionals]      402410  5.36
## 12 800 [Uganda]  9 [Elementary occupations]                       360750  4.80
## 13 484 [Mexico]  3 [Technicians and associate professionals]     1095576  3.12
## 14 800 [Uganda]  8 [Plant and machine operators and assemble~     190070  2.53
## 15 484 [Mexico]  1 [Legislators, senior officials and manage~     707943  2.01
## 16 484 [Mexico] 11 [Other occupations, unspecified or n.e.c.]     664732  1.89
## 17 800 [Uganda]  2 [Professionals]                                 64490  0.859
## 18 800 [Uganda]  4 [Clerks]                                        44290  0.590
## 19 800 [Uganda]  1 [Legislators, senior officials and manage~      27900  0.372
## 20 484 [Mexico] 10 [Armed forces]                                  64523  0.184
## 21 800 [Uganda] 97 [Response suppressed]                            2370  0.0316
```

13.Which occupation category has the highest percentage of female workers?

a.In Mexico 2000:

Code # 05=Service workers and shop and market sales (7.7%)

b.In Uganda 2002:

Code # 06 = Skilled agricultural and fishery workers (35.3%)

```
data |> filter(OCCISCO != 98 & OCCISCO != 99) |>
  group_by(COUNTRY, OCCISCO, SEX) |>
  summarise(weighted_n = sum(PERWT, na.rm = TRUE), .groups = "drop") |>
  group_by(COUNTRY) |>
  mutate(perc = weighted_n / sum(weighted_n) * 100)|>
  arrange(desc(perc))  |> print(n=50)
```

```
## # A tibble: 42 x 5
## # Groups:   COUNTRY [2]
##    COUNTRY      OCCISCO                               SEX       weighted_n    perc
##    <int+lbl>    <int+lbl>                             <int+l>        <dbl>   <dbl>
##  1 800 [Uganda]  6 [Skilled agricultural and fishery~ 1 [Mal~    2711840 3.61e+1
##  2 800 [Uganda]  6 [Skilled agricultural and fishery~ 2 [Fem~    2649150 3.53e+1
##  3 484 [Mexico]  7 [Crafts and related trades worker~ 1 [Mal~    5159621 1.47e+1
##  4 484 [Mexico]  6 [Skilled agricultural and fishery~ 1 [Mal~    5137075 1.46e+1
##  5 484 [Mexico]  5 [Service workers and shop and mar~ 1 [Mal~    3446988 9.81e+0
##  6 484 [Mexico]  8 [Plant and machine operators and ~ 1 [Mal~    2783577 7.92e+0
##  7 484 [Mexico]  5 [Service workers and shop and mar~ 2 [Fem~    2719745 7.74e+0
##  8 484 [Mexico]  9 [Elementary occupations]           1 [Mal~    2672745 7.61e+0
##  9 484 [Mexico]  9 [Elementary occupations]           2 [Fem~    2303786 6.56e+0
## 10 484 [Mexico]  4 [Clerks]                           2 [Fem~    1696101 4.83e+0
## 11 800 [Uganda]  5 [Service workers and shop and mar~ 1 [Mal~     344100 4.58e+0
## 12 484 [Mexico]  2 [Professionals]                    1 [Mal~    1533968 4.36e+0
## 13 800 [Uganda]  7 [Crafts and related trades worker~ 1 [Mal~     316660 4.22e+0
## 14 800 [Uganda]  5 [Service workers and shop and mar~ 2 [Fem~     300550 4.00e+0
## 15 484 [Mexico]  4 [Clerks]                           1 [Mal~    1405236 4.00e+0
## 16 484 [Mexico]  2 [Professionals]                    2 [Fem~    1354553 3.85e+0
## 17 800 [Uganda]  9 [Elementary occupations]           1 [Mal~     271350 3.61e+0
## 18 800 [Uganda]  3 [Technicians and associate profes~ 1 [Mal~     264520 3.52e+0
## 19 484 [Mexico]  7 [Crafts and related trades worker~ 2 [Fem~    1134365 3.23e+0
## 20 800 [Uganda]  8 [Plant and machine operators and ~ 1 [Mal~     185420 2.47e+0
## 21 484 [Mexico]  8 [Plant and machine operators and ~ 2 [Fem~     724294 2.06e+0
## 22 800 [Uganda]  3 [Technicians and associate profes~ 2 [Fem~     137890 1.84e+0
## 23 484 [Mexico]  3 [Technicians and associate profes~ 1 [Mal~     616499 1.75e+0
## 24 484 [Mexico]  6 [Skilled agricultural and fishery~ 2 [Fem~     538457 1.53e+0
## 25 484 [Mexico]  1 [Legislators, senior officials an~ 1 [Mal~     520216 1.48e+0
## 26 484 [Mexico]  3 [Technicians and associate profes~ 2 [Fem~     479077 1.36e+0
## 27 800 [Uganda]  7 [Crafts and related trades worker~ 2 [Fem~      94330 1.26e+0
## 28 484 [Mexico] 11 [Other occupations, unspecified o~ 1 [Mal~     428167 1.22e+0
## 29 800 [Uganda]  9 [Elementary occupations]           2 [Fem~      89400 1.19e+0
## 30 484 [Mexico] 11 [Other occupations, unspecified o~ 2 [Fem~     236565 6.73e-1
## 31 800 [Uganda]  2 [Professionals]                    1 [Mal~      46980 6.26e-1
## 32 484 [Mexico]  1 [Legislators, senior officials an~ 2 [Fem~     187727 5.34e-1
## 33 800 [Uganda]  4 [Clerks]                           2 [Fem~      22320 2.97e-1
## 34 800 [Uganda]  4 [Clerks]                           1 [Mal~      21970 2.93e-1
## 35 800 [Uganda]  1 [Legislators, senior officials an~ 1 [Mal~      20700 2.76e-1
## 36 800 [Uganda]  2 [Professionals]                    2 [Fem~      17510 2.33e-1
## 37 484 [Mexico] 10 [Armed forces]                     1 [Mal~      63690 1.81e-1
## 38 800 [Uganda]  1 [Legislators, senior officials an~ 2 [Fem~       7200 9.59e-2
## 39 800 [Uganda]  8 [Plant and machine operators and ~ 2 [Fem~       4650 6.19e-2
## 40 800 [Uganda] 97 [Response suppressed]              1 [Mal~       1760 2.34e-2
## 41 800 [Uganda] 97 [Response suppressed]              2 [Fem~        610 8.12e-3
## 42 484 [Mexico] 10 [Armed forces]                     2 [Fem~        833 2.37e-3
```

14.What is the labour force participation distribution by gender in each country?

a.Mexico 2000:

Men-71.7% ; Female-31.4%

b.Uganda 2002:

Men-43.5% ; Female-33.7%

```
unique(data$EMPSTAT)
```

```
## <labelled<integer>[5]>: Activity status (employment status) [general version]
## [1] 1 3 0 2 9
##
## Labels:
##  value                  label
##      0 NIU (not in universe)
##      1               Employed
##      2             Unemployed
##      3               Inactive
##      9        Unknown/missing
```

```
data |>
  filter(EMPSTAT %in% c(1, 2, 3)) |>  # esclude 0, 9
  mutate(
    labour_force = ifelse(EMPSTAT %in% c(1, 2), 1, 0)
  ) |>
  group_by(COUNTRY, SEX) |>
  summarise(
    total = sum(PERWT, na.rm = TRUE),
    lf = sum(PERWT * labour_force, na.rm = TRUE),
    lfpr = lf / total * 100,
    .groups = "drop"
  )
```

```
## # A tibble: 4 x 5
##   COUNTRY      SEX            total       lf  lfpr
##   <int+lbl>    <int+lbl>      <dbl>    <dbl> <dbl>
## 1 484 [Mexico] 1 [Male]   33653918 24119201  71.7
## 2 484 [Mexico] 2 [Female] 36571068 11479937  31.4
## 3 800 [Uganda] 1 [Male]   10150370  4410680  43.5
## 4 800 [Uganda] 2 [Female] 10278010  3461510  33.7
```

15.What percentage of women within the labour force is working:

a.In agriculture in Mexico 2000:

4.7% of women within labourforce in Mexico work in agriculture

b.In agriculture in Uganda 2002:

76.5% of women within labourforce in Uganda work in agriculture

c.In service in Mexico 2000:

23.7% of women within labourforce in Uganda work in services employment

d.In service in Uganda 2002:

8.7% of women within labourforce in Uganda work in services employment

```
unique(data$OCCISCO)
```

```
## <labelled<integer>[14]>: Occupation, ISCO general
##  [1]  5 99  7  6  9 11  8  4  3  2  1 10 98 97
##
## Labels:
##  value                                      label
##      1 Legislators, senior officials and managers
##      2                              Professionals
##      3    Technicians and associate professionals
##      4                                     Clerks
##      5  Service workers and shop and market sales
##      6   Skilled agricultural and fishery workers
##      7          Crafts and related trades workers
##      8 Plant and machine operators and assemblers
##      9                     Elementary occupations
##     10                               Armed forces
##     11   Other occupations, unspecified or n.e.c.
##     97                        Response suppressed
##     98                                    Unknown
##     99                      NIU (not in universe)
```

```
# AGRICULTURE IN MEXICO AND UGANDA

# MEXICO
# Women labour force in Mexico
subset_data <- data[data$COUNTRY == 484 &
                      data$SEX == 2 &
                      data$EMPSTAT %in% c(1, 2), ]

# Occupational categories
agriculture_codes <- c("6")  # Agricultural workers

# Numerator 1: Women in agriculture occupation (OCCISCO == 6)
numerator1 <- sum(subset_data$PERWT[subset_data$EMPSTAT == 1 & subset_data$OCCISCO == 6], na.rm = TRUE)

# Denominator: Total women in labour force (working+Non-working)
denominator1 <- sum(subset_data$PERWT, na.rm = TRUE)


# Mexico Percent of Women in labourforce working in agriculture
mexico_perc_agric <- (numerator1 / denominator1) * 100
mexico_perc_agric
```

```
## [1] 4.690418
```

```
#UGANDA
# Women labour force in Uganda
subset_data <- data[data$COUNTRY == 800 &
                      data$SEX == 2 &
                      data$EMPSTAT %in% c(1, 2), ]

# Numerator 2: Women in agriculture occupation (OCCISCO == 6)
numerator2 <- sum(subset_data$PERWT[subset_data$EMPSTAT == 1 & subset_data$OCCISCO == 6], na.rm = TRUE)

# Denominator 2: Total women in labour force (working+Non-working)
denominator2 <- sum(subset_data$PERWT, na.rm = TRUE)


# Uganda Percent of Women in labourforce working in agriculture
```

```
uganda_perc_agric <- (numerator2 / denominator2) * 100
uganda_perc_agric
```

```
## [1] 76.53163
```

```
# SERVICES IN MEXICO AND UGANDA

# MEXICO
# Women labour force in Mexico
subset_data <- data[data$COUNTRY == 484 &
                      data$SEX == 2 &
                      data$EMPSTAT %in% c(1, 2), ]

# Occupational categories
service_codes <- c("5")  # Service workers

# Numerator 3: Women in service occupation (OCCISCO == 5)
numerator3 <- sum(subset_data$PERWT[subset_data$EMPSTAT == 1 & subset_data$OCCISCO == 5], na.rm = TRUE)

# Denominator 3: Total women in labour force (working+Non-working)
denominator3 <- sum(subset_data$PERWT, na.rm = TRUE)


# Mexico Percent of Women in labourforce working in service Occupation
mexico_perc_services <- (numerator3 / denominator3) * 100
mexico_perc_services
```

```
## [1] 23.69129
```

```
#UGANDA
# Women labour force in Uganda
subset_data <- data[data$COUNTRY == 800 &
                      data$SEX == 2 &
                      data$EMPSTAT %in% c(1, 2), ]

# Numerator 4: Women in Service occupation (OCCISCO == 6)
numerator4 <- sum(subset_data$PERWT[subset_data$EMPSTAT == 1 & subset_data$OCCISCO == 5], na.rm = TRUE)

# Denominator 4: Total women in labour force (working+Non-working)
denominator4 <- sum(subset_data$PERWT, na.rm = TRUE)


# Uganda Percent of Women in labourforce working in Service Occupation
uganda_perc_services <- (numerator4 / denominator4) * 100
uganda_perc_services
```

```
## [1] 8.682627
```

Part 5: Graphical Analysis

16.What percentage of the population is literate in each country?

a.Mexico 2000:

77.7% of the Population in Mexico 2000 is literate

b.Uganda 2002:

45.1% of the Population in Uganda 2002 is literate

```
#Mexico 2000: Percentage literate, excluding response suppressed, unknown, and NIU
mexico_literacy <- data %>%
  filter(COUNTRY == 484, YEAR == 2000, LIT == 2, !LIT %in% c(97, 98, 99)) %>%
  summarise(literate_weight = sum(PERWT, na.rm = TRUE)) %>%
  mutate(
    mexico_literacy_percentage = literate_weight /
      sum(data$PERWT[data$COUNTRY == 484 & data$YEAR == 2000], na.rm = TRUE) * 100
  )
mexico_literacy$mexico_literacy_percentage
```

```
## [1] 77.6563
```

```
#Uganda 2002: Percentage literate, excluding response suppressed, unknown, and NIU
uganda_literacy <- data %>%
  filter(COUNTRY == 800, YEAR == 2002, LIT == 2, !LIT %in% c(97, 98, 99)) %>%
  summarise(literate_weight = sum(PERWT, na.rm = TRUE)) %>%
  mutate(
    uganda_literacy_percentage = literate_weight /
      sum(data$PERWT[data$COUNTRY == 800 & data$YEAR == 2002], na.rm = TRUE) * 100
  )

uganda_literacy$uganda_literacy_percentage
```

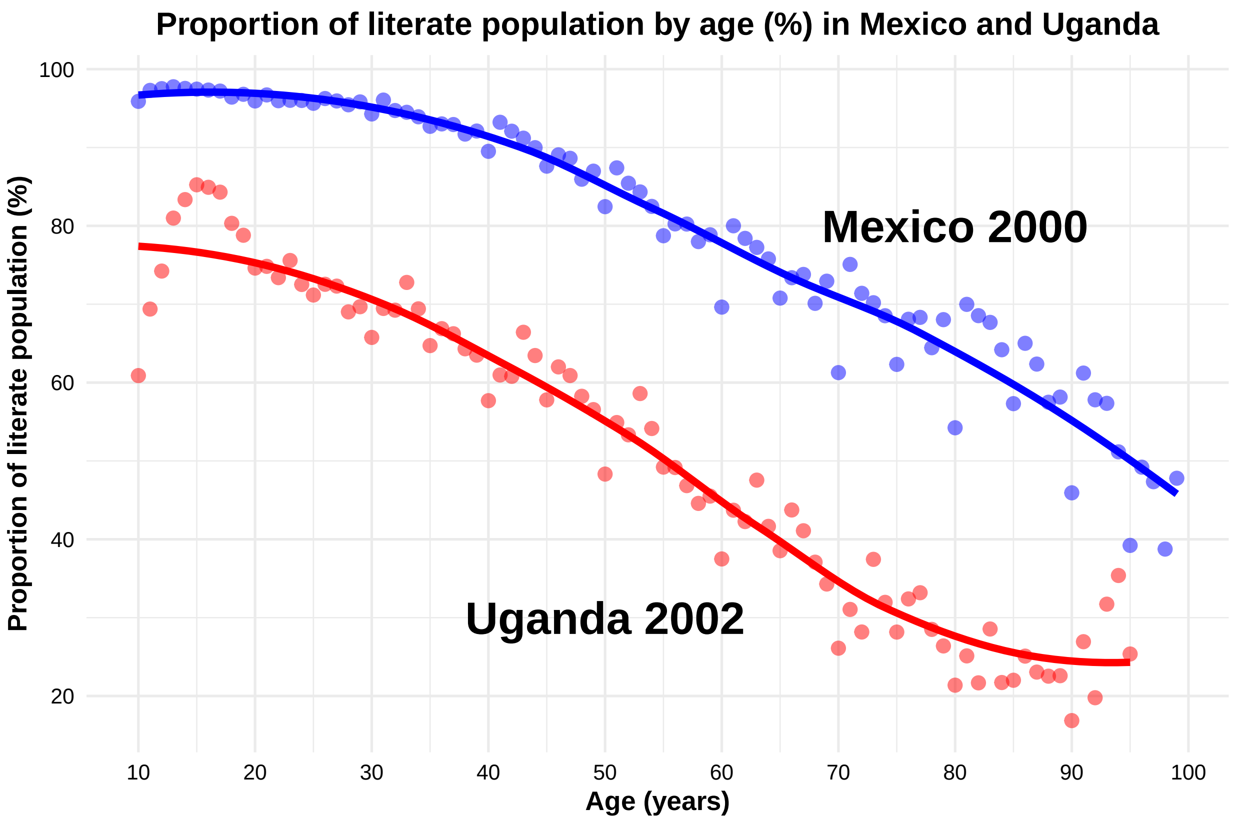
```
## [1] 45.08969
```

17.(OPTIONAL) Create a graph to visualize differences in the percentage of literacy by AGE and SEX in both countries.

```
data |>
  filter(LIT %in% 1:2, AGE < 100) |>
  summarise(n = sum(PERWT), .by = c(COUNTRY, LIT, AGE)) |>
  mutate(prop = round(100*n/sum(n),2), .by = c(COUNTRY,AGE)) |>
  mutate(
    COUNTRY = factor(COUNTRY, levels = c(484, 800), labels = c("Mexico","Uganda"))
  ) |>
  filter(LIT == 2, AGE >= 10) |>
  ggplot() +
  aes(x = AGE, y = prop, color = COUNTRY, group = COUNTRY) +
  geom_point(size = 2, alpha = .5, show.legend = FALSE) +
  geom_smooth(se = FALSE,linewidth = 1.3, show.legend = FALSE) +
  scale_color_manual(values = c("blue","red")) +
  scale_y_continuous(breaks = seq(0,100,20)) +
  scale_x_continuous(breaks = seq(0,100,10)) +
  labs(
    x = "Age (years)",
    y = "Proportion of literate population (%)",
```

```r
    caption = "Source: IPUMS International Census data",
    title = "Proportion of literate population by age (%) in Mexico and Uganda"
  ) +
  annotate(
    geom="text",
    x=80,
    y=80,
    label="Mexico 2000",
    color="black",
    fontface =2,
    size = 6
  ) +
  annotate(
    geom="text",
    x=50,
    y=30,
    label="Uganda 2002",
    color="black",
    fontface =2,
    size = 6
  ) +
  theme_minimal(base_size = 10) +
  theme(
    plot.title = element_text(color = "black", face = "bold", hjust = .5),
    plot.caption = element_text(color = "black", face = "italic", hjust = 1),
    axis.title.x = element_text(color = "black", face = "bold", hjust = .5),
    axis.title.y = element_text(color = "black", face = "bold", hjust = .5),
    axis.text = element_text(color = "black")
  )
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Source: IPUMS International Census data