Minnesota Population Center

Training and Development

IPUMS â€“ International Extraction and Analysis

Exercise 1

OBJECTIVE: Gain an understanding of how the IPUMS dataset is structured and how it can be leveraged to explore your research interests. This exercise will use the IPUMS to explore demographic and population characteristics of Mexico and Uganda.

Research Questions

What are the differences in urbanization, literacy, and occupational participation between Mexico and Uganda?

Objectives

- Create and download an IPUMS data extract
- Decompress data file and read data into R
- Analyze the data using sample code
- Validate data analysis work using answer key

IPUMS Variables

- URBAN: Household location
- SEX: Sex
- EMPSTAT: Employment status
- OCCISCO: Employment category
- FLOOR: Flooring material
- LIT: Literacy
- AGE: Age

R Code to Review

This tutorial's sample code and answers use the so-called "tidyverse" style, but R has the blessing (and curse) that there are many different ways to do almost everything. If you prefer another programming style, please feel free to use it. But, for your reference, these are some quick explanations for commands that this tutorial will use:

- **%>%** - The pipe operator which helps make code with nested function calls easier to read. When reading code, it can be read as "and then". The pipe makes it so that code like `ingredients %>% stir() %>% cook()` is equivalent to `cook(stir(ingredients))` (read as "take *ingredients* and then *stir* and then *cook*").
- **as_factor** - Converts the value labels provide for IPUMS data into a factor variable for R

- **summarize** - Summarize a datasets observations to one or more groups
- **group_by** - Set the groups for the summarize function to group by
- **filter** - Filter the dataset so that it only contains these values
- **mutate** - Add on a new variable to a dataset
- **ggplot** - Make graphs using ggplot2
- **weighted.mean** - Get the weighted mean of the a variable

Review Answer Key (At end of document)

Common Mistakes to Avoid

1) Not changing the working directory to the folder where your data is stored
2) Mixing up = and == ; To assign a value in generating a variable, use "<-" (or "="). Use "==" to test for equality.

Note: In this exercise, for simplicity we will use "weighted.mean". For analysis where variance estimates are needed, use the survey or srvyr package instead.

Registering with IPUMS

- Go to http://international.ipums.org, click on User Registration and Login and Apply for access. On login screen, enter email address and password and submit it!

Step 1: Make an Extract

- Go back to homepage and go to Select Data
- Click the Select Samples box and check the box for the 2000 sample for Mexico and 2002 for Uganda
- Click the Submit sample selections box
- Using the drop down menu or search feature, select the following variables:
  - URBAN: Household location
  - SEX: Sex
  - EMPSTAT: Employment status
  - OCCISCO: Employment category
  - FLOOR: Flooring material
  - LIT: Literacy
  - AGE: Age

Step 2: Request the Data

- Click the blue VIEW CART button under your data cart

- Review variable selection

- Click the blue Create Data Extract button
- Review the â€˜Extract Request Summaryâ€™ screen, describe your extract and click Submit Extract
- You will get an email when the data is available to download
- To get to page to download the data, follow the link in the email, or follow the Download and Revise Extracts link on the homepage

Getting the data into your statistics software

The following instructions are for R. If you would like to use a different stats package, see: http://international.ipums.org/international/extract_instructions.shtml

Step 1: Download the Data

- Go to http://international.ipums.org and click on Download or Revise Extracts
- Right-click on the data link next to extract you created
- Choose "Save Target As..." (or "Save Link As...")
- Save into "Documents" (that should pop up as the default location)
- Do the same thing for the DDI link next to the extract
- (Optional) Do the same thing for the R script
- You do not need to decompress the data to use it in R

Step 2: Install the ipumsr package

- Open R from the Start menu
- If you haven't already installed the ipumsr package, in the command prompt, type the following command:

```
install.packages("ipumsr")
```

Step 3: Read in the data

- Set your working directory to where you saved the data above by adapting the following command (Rstudio users can also use the "Project" feature to set the working directory. In the menubar, select File -> New Project -> Existing Directory and then navigate to the folder):

```
setwd("~/") # "~/" goes to your Documents directory on most computers
```

- Run the following command from the console, adapting it so it refers to the extract you just created (note the number may not be the same depending on how many extracts you've already made):

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

# Or, if you downloaded the R script, the following is equivalent:
#    source("ipumsi_00001.R")
```

- This tutorial will also rely on the dplyr and ggplot2 packages, so if you want to run the same code, run the following command (but if you know other ways better, feel free to use them):

```
library(dplyr)
library(ggplot2)
```

- To stay consistent with the exercises for other statistical packages, this exercise does not spend much time on the helpers to allow for translation of the way IPUMS uses labelled values to the way base R does. You can learn more about these in the value-labes vignette in the R package. From R run command: `vignette("value-labels", package = "ipumsr")`

Analyze the Sample â€“ Part I Variable Documentation

Section 1: Analyze the Variables

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

A) Under â€œHouseholdâ€⊚ and subcategory â€œGeographyâ€⊚, select the URBAN variable. What constitutes an urban area:

i. In Mexico in 2000? _________________________

ii. In Uganda in 2002? ________________________

B) What are the codes for URBAN?

____

C) Find the variable EMPSTAT (employment status). Is the reference period of work the same for these two samples?

____

D) What is the universe for EMPSTAT:

i. In Mexico 2000? ____________________________________

ii. In Uganda 2002? ___________________________________

Analyze the Sample â€“ Part II Frequencies

Section 1: Analyze the Data

A) Website: Find the codes page for the SAMPLE variable and write down the code values for Mexico 2000 and Uganda 2002.

____

B) How many individuals are in the Mexico 2000 sample extract?

____

C) How many individuals are in the Uganda 2002 sample extract?

____

```r
data %>%
  group_by(SAMPLE = as_factor(lbl_clean(SAMPLE), levels = "both")) %>%
  summarize(n = n())
```

D) How many individuals in the sample lived in urban areas? Mexico 2000 _______________ Uganda 2002 ____________

E) What proportion of individuals in the sample lived in urban areas? Mexico 2000 __________ Uganda 2002 __________

```r
data %>%
  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    URBAN = as_factor(URBAN)
  ) %>%
  summarize(n = n()) %>%
  mutate(pct = n / sum(n))
```

Section 2: Weighted Frequencies (PERWT)

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

F) Using weights, what is the total population of each country? Mexico 2000 ________________ Uganda 2002 _______________

G) Using weights, how many individuals lived in urban areas? Mexico 2000 ________________ Uganda 2002 _______________

H) Using weights, what proportion of individuals lived in urban areas? Mexico 2000 ______________ Uganda 2002 ___________

```r
data %>%
  group_by(SAMPLE = as_factor(lbl_clean(SAMPLE), levels = "both")) %>%
  summarize(n = sum(PERWT))

data %>%
  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    URBAN = as_factor(URBAN)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))
```

Section 3: When to use the household weights (HHWT)

Suppose you were interested not in the number of people living in urban areas, but in the number of households. To get this statistic you would need to use the household weight.

In order to use household weight, you should be careful to select only one person from each household to represent that household's characteristics. You will need to apply the household weight (HHWT). To identify only one person from each household, use the â€œfilterâ€• statement to select only cases where the PERNUM equals 1.

Analyze the Sample â€“ Part III Trends in the Data

Section 1: Analyze the Data

A) Using weights, which occupational category has the highest percentage of workers from each country? Mexico 2000 ________________ Uganda 2002 _______________

```
data %>%
  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    OCCISCO = as_factor(OCCISCO)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n)) %>%
  arrange(SAMPLE, desc(pct)) %>%
  top_n(3, pct)
```

B) Which occupational category has the highest percentage of female workers in each country? Mexico 2000 ________________ Uganda 2002 _______________

```
data %>%
  filter(SEX == 2) %>%
  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    OCCISCO = as_factor(OCCISCO)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n)) %>%
  arrange(SAMPLE, desc(pct)) %>%
  top_n(3, pct)
```

Section 2: Compare the distribution of occupational activity among people in the labor force

Note that in order to do your analysis, you must decide whether you are analyzing the total population or the people participating in the labor force. The previous commands yielded totals and percentages of people within an occupation among all people in the population. If you want to know how women's work is distributed among women in the labor force, you have to limit your analysis to people who are employed. To find out who is working, look at employment status category 1, "employed."

A) What is the labor force participation distribution by gender in each country?
Mexico 2000 %:_______
Uganda 2002 %:_______

```
data %>%
  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    SEX = as_factor(SEX)
  ) %>%
  summarize(pct = weighted.mean(EMPSTAT == 1, PERWT))
```

B) What percentage of women within the labor force is working:

i. In Agriculture; Mexico 2000:______ In Uganda 2002:________

ii. In Service; Mexico 2000:_______ Uganda 2002:________

```
agg_and_service <- c(
  "Skilled agricultural and fishery workers",
  "Service workers and shop and market sales"
)

data %>%
  filter(EMPSTAT == 1 & SEX == 2) %>%
  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    OCCISCO = as_factor(OCCISCO)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n)) %>%
  arrange(SAMPLE, OCCISCO) %>%
  filter(OCCISCO %in% agg_and_service)
```

Analyze the Sample â€“ Part IV Graphical Analysis

Section 1: Graph the Data

A) What percent of the population is literate in each sample?

____

B) How are universe differences seen on the graph?

____

```
data_summary <- data %>%
  group_by(
    SAMPLE = as_factor(SAMPLE),
    LIT = as_factor(LIT)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))

ggplot(data_summary, aes(x = LIT, y = pct)) +
  geom_col() +
  facet_wrap(~as_factor(SAMPLE)) +
  theme(axis.text.x = element_text(angle = 20, hjust = 1))
```

Section 2: Recode literacy to look at literacy rates across age

```
data <- data %>%
 mutate(
    LIT_BIN = LIT %>%
      lbl_na_if(~.lbl %in% c("NIU (not in universe)", "Unknown/missing")) %>%
      {. == 2}
  )

data_summary <- data %>%
  filter(AGE < 999 & !is.na(LIT_BIN)) %>%
  group_by(SAMPLE = as_factor(SAMPLE), AGE = as.numeric(AGE)) %>%
  summarize(MEAN_LIT = weighted.mean(LIT_BIN, PERWT))

ggplot(data_summary, aes(x = AGE, y = MEAN_LIT, color = SAMPLE, group =
SAMPLE)) +
  geom_line()
```

Section 3: Analyze Recoded Data

A) Which country has higher overall literacy?

____

B) At (approximately) which ages are literacy rates highest? Mexico 2000 ____________ Uganda 2002 ______________

C) How are universe differences seen on the graph?

____

D) In which country are literacy rates nearly equal for men and women?

____

```
data_summary <- data %>%
  group_by(SAMPLE = as_factor(SAMPLE), SEX = as_factor(SEX)) %>%
  summarize(MEAN_LIT = weighted.mean(LIT_BIN, PERWT, na.rm = TRUE))
```

```
ggplot(data_summary, aes(x = SAMPLE, y = MEAN_LIT, fill = SEX)) +
  scale_fill_manual(values = c("#7570b3", "#e6ab02")) +
  geom_col(position = "dodge")
```

E) What type of floor material is most common in Uganda 2002?

____

```
data_summary <- data %>%
  filter(as_factor(SAMPLE) == "Uganda 2002") %>%
  group_by(FLOOR = as_factor(FLOOR)) %>%
  summarize(n = n()) %>%
  mutate(pct = n / sum(n))

data_summary
```

ANSWERS Analyze the Sample â€“ Part I Variable Documentation

Section 1: Analyze the Variables

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

A) Under â€œHouseholdâ€◎ and subcategory â€œGeographyâ€◎, select the URBAN variable. What constitutes an urban area:

i. In Mexico in 2000?

2,500+ people

ii. In Uganda in 2002?

2,000+ people

B) What are the codes for URBAN?

1 Rural 2 Urban

C) Find the variable EMPSTAT (employment status). Is the reference period of work the same for these two samples?

Both samples use a reference week.

D) What is the universe for EMPSTAT:

i. In Mexico 2000?

Persons age 12+

ii. In Uganda 2002?

Persons age 5+

ANSWERS Analyze the Sample â€“ Part II Frequencies

Section 1: Analyze the Data

A) Website: Find the codes page for the SAMPLE variable and write down the code values for Mexico 2000 and Uganda 2002.
Mexico 2000: 484200001; Uganda 2002: 800200201

B) How many individuals are in the Mexico 2000 sample extract?
10,099,182 persons

C) How many individuals are in the Uganda 2002 sample extract?
2,497,449 persons

```
data %>%
  group_by(SAMPLE = as_factor(lbl_clean(SAMPLE), levels = "both")) %>%
  summarize(n = n())
#> # A tibble: 2 x 2
#>                       SAMPLE        n
#>                       <fctr>    <int>
#> 1 [484200001] Mexico 2000 10099182
#> 2 [800200201] Uganda 2002  2497449
```

D) How many individuals in the sample lived in urban areas?
Mexico 2000 *5,976,764* Uganda 2002 *306,054*

E) What proportion of individuals in the sample lived in urban areas?
Mexico 2000 *59.2%* Uganda 2002 *12.3%*

```
data %>%
  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    URBAN = as_factor(URBAN)
  ) %>%
  summarize(n = n()) %>%
  mutate(pct = n / sum(n))
#> # A tibble: 4 x 4
#> # Groups:   SAMPLE [2]
#>         SAMPLE  URBAN       n       pct
#>         <fctr> <fctr>   <int>     <dbl>
#> 1 Mexico 2000  Rural 4122418 0.4081933
#> 2 Mexico 2000  Urban 5976764 0.5918067
#> 3 Uganda 2002  Rural 2191395 0.8774534
#> 4 Uganda 2002  Urban  306054 0.1225466
```

Section 2: Weighted Frequencies (PERWT)

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

F) Using weights, what is the total population of each country?
Mexico 2000 *97,014,867* Uganda 2002 *24,974,490*

G) Using weights, how many individuals lived in urban areas?
Mexico 2000 *72,409,464* Uganda 2002 *3,060,540*

H) Using weights, what proportion of individuals lived in urban areas?
Mexico 2000 *74.6%* Uganda 2002 *12.3%*

Comparing frequencies and proportions, you can see that unweighted sample data from Mexico grossly misrepresent the population. The Mexico data was designed specifically to oversample rural areas. Weighting corrects the proportional representation of individuals or households.

```
data %>%
  group_by(SAMPLE = as_factor(lbl_clean(SAMPLE), levels = "both")) %>%
  summarize(n = sum(PERWT))
#> # A tibble: 2 x 2
#>                     SAMPLE        n
#>                     <fctr>    <dbl>
#> 1 [484200001] Mexico 2000 97014867
#> 2 [800200201] Uganda 2002 24974490

data %>%
  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    URBAN = as_factor(URBAN)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))
#> # A tibble: 4 x 4
#> # Groups:   SAMPLE [2]
#>        SAMPLE  URBAN        n       pct
#>        <fctr> <fctr>    <dbl>     <dbl>
#> 1 Mexico 2000  Rural 24605403 0.2536251
#> 2 Mexico 2000  Urban 72409464 0.7463749
#> 3 Uganda 2002  Rural 21913950 0.8774534
#> 4 Uganda 2002  Urban  3060540 0.1225466
```

Section 3: When to use the household weights (HHWT)

Suppose you were interested not in the number of people living in urban areas, but in the number of households. To get this statistic you would need to use the household weight.

In order to use household weight, you should be careful to select only one person from each household to represent that household's characteristics. You will need to apply the household weight (HHWT). To identify only one person from each household, use the â€œfilterâ€• statement to select only cases where the PERNUM equals 1.

ANSWERS Analyze the Sample â€“ Part III Trends in the Data

Section 1: Analyze the Data

A) Using weights, which occupational category has the highest percentage of workers from each country?

Mexico 2000 *6.5% Crafts and Related Trades* Uganda 2002 *21.5% of people work in Agriculture*

```
data %>%
  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    OCCISCO = as_factor(OCCISCO)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n)) %>%
  arrange(SAMPLE, desc(pct)) %>%
  top_n(3, pct)
#> # A tibble: 6 x 4
#> # Groups:   SAMPLE [2]
#>        SAMPLE                                   OCCISCO        n
#>        <fctr>                                    <fctr>    <dbl>
#> 1 Mexico 2000                     NIU (not in universe) 61870249
#> 2 Mexico 2000          Crafts and related trades workers  6293986
#> 3 Mexico 2000 Service workers and shop and market sales  6166733
#> 4 Uganda 2002                     NIU (not in universe) 17465580
#> 5 Uganda 2002  Skilled agricultural and fishery workers  5360990
#> 6 Uganda 2002 Service workers and shop and market sales   644650
#> # ... with 1 more variables: pct <dbl>
```

B) Which occupational category has the highest percentage of female workers in each country?

Mexico 2000 *Service, shop and market sales 5.5%* Uganda 2002 *Agricultural work 21.1%*

```
data %>%
  filter(SEX == 2) %>%
  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    OCCISCO = as_factor(OCCISCO)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n)) %>%
  arrange(SAMPLE, desc(pct)) %>%
  top_n(3, pct)
#> # A tibble: 6 x 4
#> # Groups:   SAMPLE [2]
#>        SAMPLE                                   OCCISCO        n
#>        <fctr>                                    <fctr>    <dbl>
#> 1 Mexico 2000                     NIU (not in universe) 38380825
#> 2 Mexico 2000 Service workers and shop and market sales  2719745
#> 3 Mexico 2000                    Elementary occupations  2303786
#> 4 Uganda 2002                     NIU (not in universe)  9228410
#> 5 Uganda 2002  Skilled agricultural and fishery workers  2649150
#> 6 Uganda 2002 Service workers and shop and market sales   300550
#> # ... with 1 more variables: pct <dbl>
```

Section 2: Compare the distribution of occupational activity among people in the labor force

Note that in order to do your analysis, you must decide whether you are analyzing the total population or the people participating in the labor force. The previous commands yielded totals and percentages of people within an occupation among all people in the population. If you want to know how women's work is distributed among women in the labor force, you have to limit your analysis to people who are employed. To find out who is working, look at employment status category 1, "employed."

A) What is the labor force participation distribution by gender in each country?
Mexico 2000 %: *50.3% of males and 22.9% of females are employed*
Uganda 2002 %: *33.7% of males and 26.5% of females are employed*

```
data %>%
  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    SEX = as_factor(SEX)
  ) %>%
  summarize(pct = weighted.mean(EMPSTAT == 1, PERWT))
#> # A tibble: 4 x 3
#> # Groups:   SAMPLE [?]
#>        SAMPLE    SEX       pct
#>        <fctr> <fctr>     <dbl>
#> 1 Mexico 2000   Male 0.5029563
#> 2 Mexico 2000 Female 0.2286285
#> 3 Uganda 2002   Male 0.3369137
#> 4 Uganda 2002 Female 0.2647869
```

B) What percentage of women within the labor force is working:

i. In Agriculture; Mexico 2000: *4.7%* In Uganda 2002: *79.7%*

ii. In Service; Mexico 2000: *23.9%* Uganda 2002: *9.0%*

```
agg_and_service <- c(
  "Skilled agricultural and fishery workers",
  "Service workers and shop and market sales"
)

data %>%
  filter(EMPSTAT == 1 & SEX == 2) %>%
  group_by(
    SAMPLE = as_factor(lbl_clean(SAMPLE)),
    OCCISCO = as_factor(OCCISCO)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n)) %>%
  arrange(SAMPLE, OCCISCO) %>%
  filter(OCCISCO %in% agg_and_service)
#> # A tibble: 4 x 4
#> # Groups:   SAMPLE [2]
#>        SAMPLE                                OCCISCO       n        pct
#>        <fctr>                                 <fctr>   <dbl>      <dbl>
#> 1 Mexico 2000 Service workers and shop and market sales 2719745 0.23908213
#> 2 Mexico 2000  Skilled agricultural and fishery workers  538457 0.04733365
#> 3 Uganda 2002 Service workers and shop and market sales  300550 0.09042878
#> 4 Uganda 2002  Skilled agricultural and fishery workers 2649150 0.79707005
```

ANSWERS Analyze the Sample â€“ Part IV Graphical Analysis

Section 1: Graph the Data

A) What percent of the population is literate in each sample?

Mexico 2000 ~78%; Uganda 2002 ~45%

B) How are universe differences seen on the graph?

NIU is included as a separate category; within universe % would be higher.

```
data_summary <- data %>%
  group_by(
    SAMPLE = as_factor(SAMPLE),
    LIT = as_factor(LIT)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))

ggplot(data_summary, aes(x = LIT, y = pct)) +
  geom_col() +
  facet_wrap(~as_factor(SAMPLE)) +
  theme(axis.text.x = element_text(angle = 20, hjust = 1))
```

Section 2: Recode literacy to look at literacy rates across age

```
data <- data %>%
 mutate(
    LIT_BIN = LIT %>%
      lbl_na_if(~.lbl %in% c("NIU (not in universe)", "Unknown/missing")) %>%
      {. == 2}
  )

data_summary <- data %>%
  filter(AGE < 999 & !is.na(LIT_BIN)) %>%
  group_by(SAMPLE = as_factor(SAMPLE), AGE = as.numeric(AGE)) %>%
  summarize(MEAN_LIT = weighted.mean(LIT_BIN, PERWT))

ggplot(data_summary, aes(x = AGE, y = MEAN_LIT, color = SAMPLE, group =
SAMPLE)) +
  geom_line()
```

Section 3: Analyze Recoded Data

A) Which country has higher overall literacy?

Mexico 2000

B) At (approximately) which ages are literacy rates highest?

Mexico 2000 ~*13-25* Uganda 2002 ~*14-18*

C) How are universe differences seen on the graph?

Lines begin at different ages (5 in Mexico, 10 in Uganda). Apart from universe, Mexico records higher ages which are included with corresponding literacy rates in the graph.

D) In which country are literacy rates nearly equal for men and women?

Mexico 2000

```
data_summary <- data %>%
  group_by(SAMPLE = as_factor(SAMPLE), SEX = as_factor(SEX)) %>%
  summarize(MEAN_LIT = weighted.mean(LIT_BIN, PERWT, na.rm = TRUE))

ggplot(data_summary, aes(x = SAMPLE, y = MEAN_LIT, fill = SEX)) +
  scale_fill_manual(values = c("#7570b3", "#e6ab02")) +
  geom_col(position = "dodge")
```

E) What type of floor material is most common in Uganda 2002?

None (earth floor)

```
data_summary <- data %>%
  filter(as_factor(SAMPLE) == "Uganda 2002") %>%
  group_by(FLOOR = as_factor(FLOOR)) %>%
  summarize(n = n()) %>%
  mutate(pct = n / sum(n))

data_summary
#> # A tibble: 8 x 3
#>                     FLOOR       n         pct
#>                    <fctr>   <int>       <dbl>
#> 1   NIU (not in universe)   26548 0.010630047
#> 2 None/unfinished (earth) 1931982 0.773582163
#> 3                Concrete   84225 0.033724412
#> 4            Cement screed  399943 0.160140607
#> 5                    Stone    9954 0.003985667
#> 6                    Brick   15790 0.006322451
#> 7                     Wood   11469 0.004592286
#> 8  Other finished, n.e.c.   17538 0.007022366
```