Minnesota Population Center

Training and Development

IPUMS â€“ USA Extraction and Analysis

Exercise 2

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 associations in household ownership, and trends in language spoken in the home.

11/13/2017

Research Questions

What proportion of households in the US has a mortgage? Is the mother's spoken language a consistent determinant of a child's preferred language? How are utility costs changing over time, and are changes in cost different by urban status?

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

- MORTGAGE: Mortgage Status
- METRO: Metropolitan status
- VALUEH: House value
- OWNERSHP: Ownership of dwelling
- LANGUAGE: Language spoken at home
- COSTELEC: Annual electricity cost
- SEX: Age
- COSTGAS: Annual gas cost
- AGE: Sex
- ROOMS: Number of rooms
- UNITSSTR: Units in structure

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
- **gather** - Use tidyr's gather to help reshape data when making graphs
- **weighted.mean** - Get the weighted mean of the a variable

Review Answer Key (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.
3) Not including missing values when needed. The attached characteristics have missing values when the person doesn't have the relationship, but sometimes you want to treat that as a "No", not as a missing value.

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

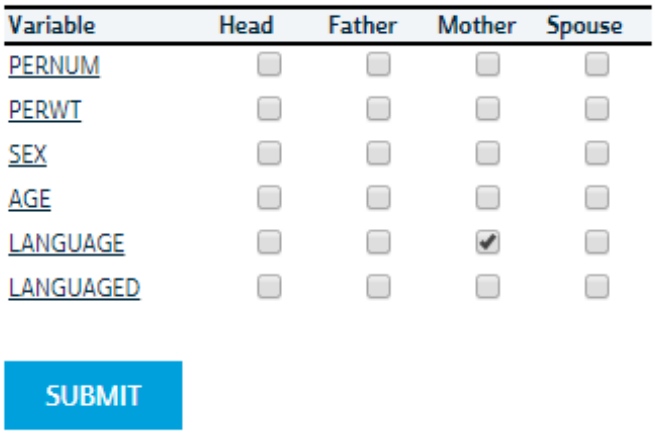
Go to http://usa.ipums.org, click on IPUMS Registration and Login and Apply for access. On login screen, enter email address and password and submit it!

Step 1: Make an Extract

- Go back to homepage and go to Select Data
- Click the Select Samples box, check the box for the 2010 ACS sample, then select â€œSubmit sample selectionsâ€•
- Using the drop down menu or search feature, select the following variables:
  - MORTGAGE: Mortgage Status
  - SEX: Age
  - VALUEH: House value
  - AGE: Sex
  - LANGUAGE: Language spoken at home

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
- Click â€œAttach Characteristicsâ€•. Check the box at the intersection of LANGUAGE and Mother, and Submit



- Review the â€˜Extract Request Summaryâ€™, describe your extract and click Submit Extract. You will get an email when the data is available to download
- Follow the Download and Revise Extracts link on the homepage, or the link in the email
- Do the same for a second extract. Choose the ACS samples 2005 through 2010, and the following variables:
  - METRO: Metropolitan status
  - COSTWATR: Annual water cost
  - OWNERSHP: Ownership of dwelling
  - ROOMS: Number of rooms
  - COSTELEC: Annual electricity cost
  - UNITSSTR: Units in structure
  - COSTGAS: Annual gas cost

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://usa.ipums.org/usa/extract_instructions.shtml

Step 1: Download the Data

- Go to http://usa.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("usa_00001.xml")
data <- read_ipums_micro(ddi)

# Or, if you downloaded the R script, the following is equivalent:
#    source("usa_00001.R")
```

- This tutorial will also rely on the dplyr package, 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)
library(tidyr)
```

- 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 Frequencies

Section 1: Analyze the Variables

Get a basic frequency of the MORTGAGE variable.

A) Find the codes page on the website for the MORTGAGE variable and write down the code value, and what category each code represents.

_____

B) How many people in the sample had a mortgage or deed of trust on their home in 2010? What proportion of the sample had a mortgage?

_____

```
data %>%
  group_by(MORTGAGE = as_factor(MORTGAGE)) %>%
  summarize(n = n()) %>%
  mutate(pct = n / sum(n))
```

C) Using weights, what proportion of the population had a mortgage in 2010?

_____

```
data %>%
  group_by(MORTGAGE = as_factor(MORTGAGE)) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))
```

Section 2: Using weights

Using household weights (HHWT)

Suppose you were interested not in the number of people with mortgages, but in the number of households that had mortgages. 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. And you will need to apply the household weight (HHWT).

D) What proportion of households in the sample had a mortgage? What proportion of the sample owned their home? (Hint: donâ€™t use the weight quite yet)

_____

```
data %>%
  filter(PERNUM == 1) %>%
  group_by(MORTGAGE = as_factor(MORTGAGE)) %>%
  summarize(n = n()) %>%
  mutate(pct = n / sum(n))
```

E) What proportion of households had a mortgage across the country in 2010?

_____

F) What proportion of households owned their home? Does the sample over or under-represent households who own their home?

_____

```
data %>%
  filter(PERNUM == 1) %>%
  group_by(MORTGAGE = as_factor(MORTGAGE)) %>%
  summarize(n = sum(HHWT)) %>%
  mutate(pct = n / sum(n))
```

G) What is the average value of:

i. A home that is mortgaged? _____

ii. A home that is owned? _____

H) What could explain this difference? _____

```
data %>%
  filter(VALUEH != 0 & VALUEH != 9999999 & PERNUM == 1) %>%
  group_by(MORTGAGE = as_factor(MORTGAGE)) %>%
  summarize(VALUEH = weighted.mean(VALUEH, HHWT))
```

Note: The missing value code for house value is excluded.

Section 3: Graph the Data

I) Under the description tab on the website for VALUEH, reader the first user note. On the codes page, find the top codes by state for VALUEH, under 2010 ACS/PRCS topcodes by state. How could this complicate your data analysis? Check a histogram of your data to rule out any bias.

_____

_____

```
data_summary <- data %>%
  filter(VALUEH != 0 & VALUEH != 9999999 & PERNUM == 1)

ggplot(data_summary, aes(x = as.numeric(VALUEH), weight = HHWT)) +
  geom_histogram()
```

Analyze the Sample - Part II Frequencies

Section 1: Analyze the Variables

Investigate LANGUAGE variable frequencies.

A) What were the three most commonly spoken languages in the US in 2010?

_____

Note: The sort option automatically organizes the table into descending frequency.

```r
data %>%
  group_by(LANGUAGE = as_factor(LANGUAGE, level = "both")) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n)) %>%
  arrange(desc(n))
```

B) Using the code page on the website for LANGUAGE, find the codes for the three most commonly spoken languages. _____

C) What percent of individuals who speak English at home:

i. Has a mother who speaks Spanish at home? _____

ii. Has a mother who speaks Chinese at home? _____

```r
data %>%
  filter(LANGUAGE == 1) %>%
  summarize(
    mom_spanish = weighted.mean(LANGUAGE_MOM == 12, PERWT, na.rm = TRUE),
    mom_chinese = weighted.mean(LANGUAGE_MOM == 43, PERWT, na.rm = TRUE)
  )
```

D) What percent of men under the age of 30 speak Spanish at home?

_____

```r
data %>%
  filter(as_factor(SEX) == "Male" & AGE < 30) %>%
  group_by(LANGUAGE = as_factor(LANGUAGE)) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n)) %>%
  arrange(desc(pct))
```

Analyze the Sample - Part III Advanced Exercises

Section 1: Analyze the Data

Revisit Step 3 to import the second extract into R.

A) On the website, what are the codes for METRO? What is the code for a single family house, detached in the variable UNITSSTR?

_____

_____

B) What is the proportion of households in the central city who owned their home in 2008? _____ In 2010? _____

```r
data %>%
  filter(PERNUM == 1 & METRO == 2) %>%
  group_by(YEAR) %>%
  summarize(own = weighted.mean(OWNERSHP == 1, HHWT))
```

Section 2: Graph the Data

Create a graph for annual utility costs by metropolitan status

C) What is the approximate annual cost of water for:

i. A household in the metro area in 2010? _____

ii. A household not in the metro area? _____

D) What is the approximate annual cost of electricity for:

i. A household in the metro area in 2010? _____

ii. A household not in the metro area? _____

```
data_summary <- data %>%
  filter(PERNUM == 1 & YEAR == 2010 & COSTELEC != 0 & COSTELEC < 9990 &
COSTWATR != 0 & COSTWATR < 9990) %>%
  group_by(METRO = as_factor(METRO)) %>%
  summarize(
    COSTELEC = weighted.mean(COSTELEC, HHWT),
    COSTWATR = weighted.mean(COSTWATR, HHWT)
    ) %>%
  gather(key, value, COSTELEC, COSTWATR)

ggplot(data_summary, aes(x = METRO, y = value, fill = key)) +
  geom_col(position = "dodge") +
  theme(axis.text.x = element_text(angle = 20, hjust = 1)) +
  scale_fill_manual(values = c("#7570b3", "#e6ab02"))
```

E) Is there a simple correlation between the number of rooms and the annual cost of electricity? _____

```
cor(data$COSTELEC, data$ROOMS)
```

Next, create a graph that will display the average cost of electricity and gas over time, controlling for the number of rooms and the units in structure. To control for these variables, look at the specific case of a single family house, detached with 5 rooms. Because the graph will also observe prices over time, inflation must be controlled for.

F) On the website, find the variable description for COSTELEC and follow the link that discusses adjusting for inflation. What year is the index year?

_____

G) Has the annual cost of gas for a single family, 5-room home increased since 2005 in nominal terms? What about the annual cost of water?

_____

```
data_summary <- data %>%
  filter(PERNUM == 1 & COSTELEC != 0 & COSTELEC < 9990 & COSTWATR != 0 &
COSTWATR < 9990) %>%
  filter(UNITSSTR == 3 & ROOMS == 5) %>%
```

```
  group_by(YEAR = YEAR) %>%
  summarize(
    COSTELEC = weighted.mean(COSTELEC, HHWT),
    COSTWATR = weighted.mean(COSTWATR, HHWT)
    ) %>%
  gather(key, value, COSTELEC, COSTWATR)

ggplot(data_summary, aes(x = YEAR, y = value, fill = key)) +
  geom_col(position = "dodge") +
  theme(axis.text.x = element_text(angle = 20, hjust = 1)) +
  scale_fill_manual(values = c("#7570b3", "#e6ab02"))
```

H) Has the annual cost of gas for a single family, 5 room home increased since 2005 in real terms? _____

Note: The variable ADJUST assigns an inflation index value according to the year of the observation. There is not yet an index for 2010, so exclude 2010.

```
inc_adj <- data.frame(YEAR = 2005:2010, ADJUST = c(0.853, 0.826, 0.804,
0.774, 0.777, 0.764))
data <- left_join(data %>% mutate(YEAR = zap_ipums_attributes(YEAR)),
inc_adj, by = "YEAR")

data_summary <- data %>%
  filter(PERNUM == 1 & COSTGAS != 0 & COSTGAS < 9990) %>%
  filter(UNITSSTR == 3 & ROOMS == 5) %>%
  mutate_at(vars(COSTGAS), function(x, adj) x * adj, adj = .$ADJUST) %>%
  group_by(YEAR) %>%
  summarize(
    COSTGAS = weighted.mean(COSTGAS, HHWT)
  )

ggplot(data_summary, aes(x = YEAR, y = COSTGAS)) +
  geom_col(position = "dodge") +
  theme(axis.text.x = element_text(angle = 20, hjust = 1)) +
  scale_fill_manual(values = c("#7570b3", "#e6ab02"))
```

ANSWERS Analyze the Sample â€“ Part I Frequencies

Section 1: Analyze the Variables

Get a basic frequency of the MORTGAGE variable.

A) Find the codes page on the website for the MORTGAGE variable and write down the code value, and what category each code represents.
0 N/A; 1 No, owned free and clear; 2 Check mark on manuscript (probably yes); 3 Yes, mortgaged/ deed of trust or similar debt; 4 Yes, contract to purchase

B) How many people in the sample had a mortgage or deed of trust on their home in 2010? What proportion of the sample had a mortgage?
1,523,041 people; 49.75%

```
data %>%
  group_by(MORTGAGE = as_factor(MORTGAGE)) %>%
  summarize(n = n()) %>%
  mutate(pct = n / sum(n))
#> # A tibble: 4 x 3
#>                                        MORTGAGE       n         pct
#>                                          <fctr>   <int>       <dbl>
#> 1                                           N/A  888238 0.290113441
#> 2                        No, owned free and clear  628501 0.205278976
#> 3 Yes, mortgaged/ deed of trust or similar debt 1523041 0.497450756
#> 4                      Yes, contract to purchase   21912 0.007156827
```

C) Using weights, what proportion of the population had a mortgage in 2010?
47.46%

```
data %>%
  group_by(MORTGAGE = as_factor(MORTGAGE)) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))
#> # A tibble: 4 x 3
#>                                        MORTGAGE         n         pct
#>                                          <fctr>     <dbl>       <dbl>
#> 1                                           N/A 107104033 0.346223180
#> 2                        No, owned free and clear  53211365 0.172010404
#> 3 Yes, mortgaged/ deed of trust or similar debt 146806345 0.474564385
#> 4                      Yes, contract to purchase   2227946 0.007202031
```

Section 2: Using weights

Using household weights (HHWT)

Suppose you were interested not in the number of people with mortgages, but in the number of households that had mortgages. 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. And you will need to apply the household weight (HHWT).

D) What proportion of households in the sample had a mortgage? What proportion of the sample owned their home? (Hint: donâ€™t use the weight quite yet)
42.20% of households mortgaged; 23.98% of household owned

```
data %>%
  filter(PERNUM == 1) %>%
```

```
  group_by(MORTGAGE = as_factor(MORTGAGE)) %>%
  summarize(n = n()) %>%
  mutate(pct = n / sum(n))
#> # A tibble: 4 x 3
#>                                          MORTGAGE      n         pct
#>                                            <fctr>  <int>       <dbl>
#> 1                                             N/A 426976 0.332619758
#> 2                         No, owned free and clear 307864 0.239829988
#> 3 Yes, mortgaged/ deed of trust or similar debt 541709 0.421998230
#> 4                       Yes, contract to purchase   7127 0.005552024
```

E) What proportion of households had a mortgage across the country in 2010?
40.53% of households

F) What proportion of households owned their home? Does the sample over or under-represent households who own their home?
20.07% of households, sample over-represents households that own their own home or have a mortgage.

```
data %>%
  filter(PERNUM == 1) %>%
  group_by(MORTGAGE = as_factor(MORTGAGE)) %>%
  summarize(n = sum(HHWT)) %>%
  mutate(pct = n / sum(n))
#> # A tibble: 4 x 3
#>                                          MORTGAGE        n         pct
#>                                            <fctr>    <dbl>       <dbl>
#> 1                                             N/A 47607067 0.388455433
#> 2                         No, owned free and clear 24600389 0.200729752
#> 3 Yes, mortgaged/ deed of trust or similar debt 49666435 0.405259087
#> 4                       Yes, contract to purchase   680881 0.005555728
```

G) What is the average value of:

i. A home that is mortgaged? *$267,941.30*

ii. A home that is owned? *$219,110.30*

H) What could explain this difference?
? Perhaps homes that have already been paid off are older and less expensive, or it takes less time to pay off a home that is worth less.

```
data %>%
  filter(VALUEH != 0 & VALUEH != 9999999 & PERNUM == 1) %>%
  group_by(MORTGAGE = as_factor(MORTGAGE)) %>%
  summarize(VALUEH = weighted.mean(VALUEH, HHWT))
```

Note: The missing value code for house value is excluded.

Section 3: Graph the Data

I) Under the description tab on the website for VALUEH, reader the first user note. On the codes page, find the top codes by state for VALUEH, under 2010 ACS/PRCS

topcodes by state. How could this complicate your data analysis? Check a histogram of your data to rule out any bias.
There doesnâ€™t seem to be a significant cluster around the topcodes, so the data sample may not be noticeably biased.

```
data_summary <- data %>%
  filter(VALUEH != 0 & VALUEH != 9999999 & PERNUM == 1)

ggplot(data_summary, aes(x = as.numeric(VALUEH), weight = HHWT)) +
  geom_histogram()
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

ANSWERS Analyze the Sample - Part II Frequencies

Section 1: Analyze the Variables

Investigate LANGUAGE variable frequencies.

A) What were the three most commonly spoken languages in the US in 2010?
English, Spanish, Chinese

```
data %>%
  group_by(LANGUAGE = as_factor(LANGUAGE, level = "both")) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n)) %>%
  arrange(desc(n))
#> # A tibble: 62 x 3
#>                   LANGUAGE         n         pct
#>                     <fctr>     <dbl>       <dbl>
#>  1            [1] English 229753346 0.742697841
#>  2           [12] Spanish  36981799 0.119546909
#>  3       [0] N/A or blank  20095332 0.064959923
#>  4           [43] Chinese   2717534 0.008784667
#>  5 [31] Hindi and related   2104371 0.006802564
#>  6            [11] French   2098698 0.006784225
#>  7 [54] Filipino, Tagalog   1697559 0.005487508
#>  8        [50] Vietnamese   1390472 0.004494823
#>  9             [2] German   1222054 0.003950397
#> 10            [49] Korean   1117307 0.003611793
#> # ... with 52 more rows
```

B) Using the code page on the website for LANGUAGE, find the codes for the three most commonly spoken languages.
01 English; 12 Spanish; 43 Chinese

C) What percent of individuals who speak English at home:

i. Has a mother who speaks Spanish at home? *3.89%*

ii. Has a mother who speaks Chinese at home? *2.22%*

```
data %>%
  filter(LANGUAGE == 1) %>%
  summarize(
    mom_spanish = weighted.mean(LANGUAGE_MOM == 12, PERWT, na.rm = TRUE),
    mom_chinese = weighted.mean(LANGUAGE_MOM == 43, PERWT, na.rm = TRUE)
  )
#> # A tibble: 1 x 2
#>   mom_spanish mom_chinese
#>         <dbl>       <dbl>
#> 1  0.03890631 0.002225844
```

D) What percent of men under the age of 30 speak Spanish at home? *13.4%*

```
data %>%
  filter(as_factor(SEX) == "Male" & AGE < 30) %>%
  group_by(LANGUAGE = as_factor(LANGUAGE)) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n)) %>%
  arrange(desc(pct))
#> # A tibble: 62 x 3
#>              LANGUAGE        n         pct
#>                <fctr>    <dbl>       <dbl>
#>  1           English 41701183 0.649009608
#>  2      N/A or blank 10264433 0.159748841
#>  3           Spanish  8615686 0.134088834
#>  4           Chinese   459082 0.007144848
#>  5 Hindi and related   381566 0.005938441
#>  6            French   329330 0.005125474
#>  7        Vietnamese   250195 0.003893869
#>  8            German   194703 0.003030229
#>  9            Arabic   187287 0.002914811
#> 10 Filipino, Tagalog   176588 0.002748299
#> # ... with 52 more rows
```

ANSWERS Analyze the Sample - Part III Advanced Exercises

Section 1: Analyze the Data

Revisit Step 3 to import the second extract into R.

A) On the website, what are the codes for METRO? What is the code for a single family house, detached in the variable UNITSSTR? *UNITSSTR: 03 1-family house, detached; METRO: 0 Not identifiable; 1 Not in metro area; 2 Central city; 3 Outside central city; 4 Central city status unknown*

B) What is the proportion of households in the central city who owned their home in 2008? *44.51%* In 2010? *42.92%*

```
data %>%
  filter(PERNUM == 1 & METRO == 2) %>%
  group_by(YEAR) %>%
  summarize(own = weighted.mean(OWNERSHP == 1, HHWT))
#> # A tibble: 6 x 2
#>    YEAR       own
#>   <int>     <dbl>
#> 1  2005 0.4833379
#> 2  2006 0.4529056
#> 3  2007 0.4518961
#> 4  2008 0.4450600
#> 5  2009 0.4378750
#> 6  2010 0.4292035
```

Section 2: Graph the Data

Create a graph for annual utility costs by metropolitan status

C) What is the approximate annual cost of water for:

i. A household in the metro area in 2010? *~1700*

ii. A household not in the metro area? *~1750*

D) What is the approximate annual cost of electricity for:

i. A household in the metro area in 2010? *~600*

ii. A household not in the metro area? *~500*

```
data_summary <- data %>%
  filter(PERNUM == 1 & YEAR == 2010 & COSTELEC != 0 & COSTELEC < 9990 &
COSTWATR != 0 & COSTWATR < 9990) %>%
  group_by(METRO = as_factor(METRO)) %>%
  summarize(
    COSTELEC = weighted.mean(COSTELEC, HHWT),
    COSTWATR = weighted.mean(COSTWATR, HHWT)
    ) %>%
  gather(key, value, COSTELEC, COSTWATR)

ggplot(data_summary, aes(x = METRO, y = value, fill = key)) +
  geom_col(position = "dodge") +
  theme(axis.text.x = element_text(angle = 20, hjust = 1)) +
  scale_fill_manual(values = c("#7570b3", "#e6ab02"))
```

E) Is there a simple correlation between the number of rooms and the annual cost of electricity?

There seems to be a weak positive correlation between number of rooms and the cost of electricity. (0.11)

```
cor(data$COSTELEC, data$ROOMS)
#> [1] 0.113366
```

Next, create a graph that will display the average cost of electricity and gas over time, controlling for the number of rooms and the units in structure. To control for these variables, look at the specific case of a single family house, detached with 5 rooms. Because the graph will also observe prices over time, inflation must be controlled for.

F) On the website, find the variable description for COSTELEC and follow the link that discusses adjusting for inflation. What year is the index year?
1999

G) Has the annual cost of gas for a single family, 5-room home increased since 2005 in nominal terms? What about the annual cost of water?
Both appear to be rising

```
data_summary <- data %>%
  filter(PERNUM == 1 & COSTELEC != 0 & COSTELEC < 9990 & COSTWATR != 0 &
COSTWATR < 9990) %>%
  filter(UNITSSTR == 3 & ROOMS == 5) %>%
  group_by(YEAR = YEAR) %>%
  summarize(
    COSTELEC = weighted.mean(COSTELEC, HHWT),
    COSTWATR = weighted.mean(COSTWATR, HHWT)
    ) %>%
  gather(key, value, COSTELEC, COSTWATR)

ggplot(data_summary, aes(x = YEAR, y = value, fill = key)) +
  geom_col(position = "dodge") +
  theme(axis.text.x = element_text(angle = 20, hjust = 1)) +
  scale_fill_manual(values = c("#7570b3", "#e6ab02"))
```

H) Has the annual cost of gas for a single family, 5 room home increased since 2005 in real terms?
Decreasing Note: The variable ADJUST assigns an inflation index value according to the year of the observation.

```
inc_adj <- data.frame(YEAR = 2005:2010, ADJUST = c(0.853, 0.826, 0.804,
0.774, 0.777, 0.764))
data <- left_join(data %>% mutate(YEAR = zap_ipums_attributes(YEAR)),
inc_adj, by = "YEAR")

data_summary <- data %>%
  filter(PERNUM == 1 & COSTGAS != 0 & COSTGAS < 9990) %>%
```

```r
  filter(UNITSSTR == 3 & ROOMS == 5) %>%
  mutate_at(vars(COSTGAS), function(x, adj) x * adj, adj = .$ADJUST) %>%
  group_by(YEAR) %>%
  summarize(
    COSTGAS = weighted.mean(COSTGAS, HHWT)
  )

ggplot(data_summary, aes(x = YEAR, y = COSTGAS)) +
  geom_col(position = "dodge") +
  theme(axis.text.x = element_text(angle = 20, hjust = 1)) +
  scale_fill_manual(values = c("#7570b3", "#e6ab02"))
```