IPUMS Data Training Exercise:

An introduction to IPUMS USA

(Exercise 2 for R)

Learning goals

- Understand how IPUMS USA dataset is structured
- Create and download an IPUMS data extract
- Read the data into R

Summary

In this exercise, you will gain basic familiarity with the IPUMS USA data exploration and extract system to answer the following 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? You will create a data extract that includes the variables MORTGAGE, VALUEH, LANGUAGE, SEX, AGE, METRO, OWNERSHP, COSTELEC, COSTGAS, ROOMS, UNTSSTR; then you will use the sample code to analyze these data. After completing this exercise, you will have experience navigating the IPUMS USA website and should be able to leverage these data to explore your own research interests.

Register for an IPUMS Account

Go to https://usa.ipums.org/usa/ click on Login at the top, and apply for access. On login screen, enter email address and password and submit it!

Make a data extract

- Navigate to the IPUMS USA homepage and click on "Browse Data."

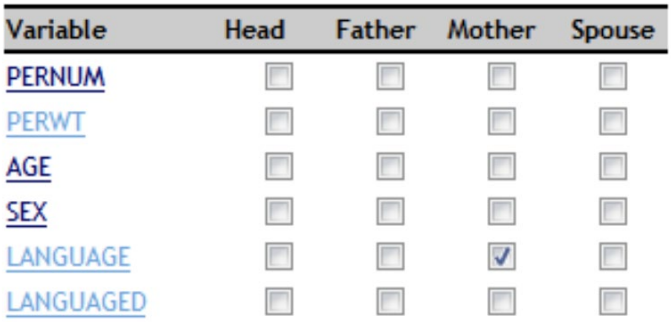
Select Samples – Extract #1: Associations in Household Ownership

- Go to the homepage and click SELECT DATA located at the top of the page.
- On the following webpage, click SELECT SAMPLES.
- Choose the 2010 ACS (1-year) sample by "check marking" the radio box to the left of the sample name.
- Once checked, click SUBMIT SAMPLE SELECTIONS.

Select Variables – Extract #1: Associations in Household Ownership

- Return to the SELECT DATA page. Using the variable table or search feature, find the variables:
    - MORTGAGE: Mortgage status
    - VALUEH: House value
    - LANGUAGE: Language spoken at home
    - SEX: Sex
    - AGE: Age
- Once you have located the variables, click the radio button `Add to cart' on the left side of the page. This selects them to be included in the data extract.

- Once the sample and variables are selected, click VIEW CART -> CREATE DATA EXTRACT.
- For this example, we will attach to each person case the language spoken by their mother if she resides in the household.
- To accomplish this, click "ATTACH CHARACTERISTICS" on the EXTRACT REQUEST page. Check the box at the intersection of LANGUAGE and Mother, and SUBMIT.



Variable	Head	Father	Mother	Spouse
PERNUM	☐	☐	☐	☐
PERWT	☐	☐	☐	☐
AGE	☐	☐	☐	☐
SEX	☐	☐	☐	☐
LANGUAGE	☐	☐	☑	☐
LANGUAGED	☐	☐	☐	☐

- Review and provide a short description for the extract and click SUBMIT EXTRACT.
- You will receive an e-mail when the data is available for download.

Now create a second extract

Select Samples – Extract #2: Housing Costs

- Go to the homepage and click SELECT DATA located at the top of the page.
- On the following webpage, click SELECT SAMPLES.
- Choose the 2005 through 2010 ACS (1-year) samples by "check marking" the radio box to the left of the sample names.
- Once checked, click SUBMIT SAMPLE SELECTIONS.

Select Variables - Extract #2: Housing Costs

- Return to the SELECT DATA page. Using the variable table or search feature, find the variables:
    - METRO: Metropolitan status
    - OWNERSHP: Ownership of dwelling
    - COSTELEC: Annual electricity cost
    - COSTGAS: Annual gas cost
    - COSTWATR: Annual water cost
    - ROOMS: Number of rooms
    - UNITSSTR: Units in structure
    - CPI99: CPI-U adjustment factor to 1999 dollars
- Once you have located the variables, click the radio button `Add to cart' on the left side of the page. This selects them to be included in the data extract. The radio button should then change from a `+' to a checkmark to confirm selection.
- Review and provide a short description for the extract and click SUBMIT EXTRACT.
- You will receive an e-mail when the data is available for download.

Review and submit your extract

- Click on the "View Cart" button underneath your data cart.
- Review your variable and sample selection to ensure your extract is complete.
    - You may notice a number of additional variables you did not select are in your cart; IPUMS preselects a number of key technical variables, which are automatically included in your data extract.

- Add additional variables or samples if they are missing from your extract, or click the "Create Data Extract" button.
- Review the Extract Request screen that summarizes your extract; add a description of your extract (e.g., "USA Exercise 2: Household Ownership" or "USA Exercise 2: Housing Costs") and click "Submit Extract".
- You will receive an email when your data extract is available to download.

Getting the data into your statistics software

The following instructions are for R. If you would like to use a different stats package, see: https://ipums.org/support/exercises

Download the Data

- Go to https://usa.ipums.org/usa/ and click on Download or Revise Extracts.
- Right-click on the Data link next to the extract you created.
- Choose "Save Target As..." (or "Save Link As...").
- Save into "Documents" (Documents should pop up as the default location).
- Do the same 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.

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")
```

Read 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 have 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, tidyr, and ggplot2 packages, so if you want to run the same code, run the following commands (but if you know other ways better, feel free to use them):

```
library(dplyr)
library(tidyr)
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-labels vignette in the R package. From R, run command:

```
vignette("value-labels", package = "ipumsr")
```

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:

Code	Purpose
%>%	The pipe operator 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 provided for IPUMS data into a factor variable for R
summarize	Summarize a dataset's 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
weighted.mean	Get the weighted mean of the variable

Common Mistakes to Avoid

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

A note on IPUMS USA and sample weighting

Many of the data samples provided by IPUMS USA are based on statistical survey techniques to obtain a nationally representative sample of the population. This means that persons with some characteristics are over-represented in the samples, while others are underrepresented.

To obtain representative statistics, users should always apply IPUMS USA sample weights for the population of interest (persons/households). IPUMS USA provides both person (PERWT) and household—level (HHWT) sampling weights to assist users with applying a consistent sampling weight procedure across data samples. While appropriate use of sampling weights will produce correct point estimates (e.g., means, proportions), many researchers believe that it is also necessary to use additional statistical techniques that account for the complex sample design to produce correct standard errors and statistical tests.

IPUMS USA has provided the variables STRATA and CLUSTER for this purpose. While unnecessary for the following analytic exercises focused on mean and proportional estimates, a further discussion can be found on the IPUMS USA website: ANALYSIS AND VARIANCE ESTIMATION WITH IPUMS USA
https://usa.ipums.org/usa/complex_survey_vars/userNotes_variance.shtml

Analyze the Data

Part 1: Frequencies

This part of the exercise uses Extract #1: Associations in Household Ownership.

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

   _____

   _____

   _____

2. 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 = haven::as_factor(MORTGAGE)) %>%
   summarize(n = n()) %>%
   mutate(pct = n / sum(n))
```

3. Using weights, what proportion of the population had a mortgage in 2010?

   _____

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

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 (HHWT) and select only one person (filter using PERNUM = 1) from each household to represent that household's characteristics.

4. What proportion of households **in the sample** had a mortgage? What proportion **of the sample** owned their home?

   _____

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

5. What proportion of households **across the country** in 2010 had a mortgage?

   _____

6. 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 = haven::as_factor(MORTGAGE)) %>%
   summarize(n = sum(HHWT)) %>%
   mutate(pct = n / sum(n))
```


7. What is the average value of:
    a. A home that is mortgaged? _____
    b. A home that is owned? _____

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

8. What could explain this difference? *Note: Exclude cases where house value is missing.*

    _____

    _____

    _____

9. Under the description tab on the website for VALUEH, read the first user note. Follow the relevant link on the codes tab to find the top codes for VALUEH (2010 ACS/PRCS topcodes by state). How could this complicate your data analysis? Create and 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) ) +
stat_bin(bins = 30)
```


Part 2: Frequencies

10. What were the three most commonly spoken languages in the US in 2010?

    _____

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

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

    _____

    _____

12. What percent of individuals who speak English at home:
    a. Has a mother who speaks Spanish at home? _____
    b. Has a mother who speaks Chinese at home? _____

```
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)
   )
```

13. What percent of men under the age of 30 speak Spanish at home?

_____

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

Part 3: Advanced Exercises

This part of the exercise uses Extract #2: Housing Costs.

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

_____

_____

_____

15. What is the proportion of households in the central city who owned their home:
    a. in 2008?

    _____

    b. in 2010?

    _____

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


Create a graph for annual utility costs by metropolitan status

16. What is the approximate annual cost of *water* for:

    a. A household in the metro area in 2010?

    _____

    b. A household not in the metro area?

    _____

```
data <- data %>%

   mutate(in_metro = case_when(METRO == 1 ~ "nonmetro", METRO
     %in% 2:4 ~ "metro", METRO == 0 ~ NA_character_))

data %>%

  filter(PERNUM == 1 & YEAR == 2010 & COSTWATR != 0 & COSTWATR <
     9990 & !is.na(in_metro)) %>%

  group_by(in_metro) %>%

  summarize(COSTWATR = weighted.mean(COSTWATR, HHWT))

  ggplot(aes(x = in_metro, y = COSTWATR)) +

    geom_col(fill = "royalblue3")
```

17. What is the approximate annual cost of *electricity* for:

    a. A household in the metro area in 2010?

    _____

    b. A household not in the metro area?

    _____

```
data %>%
  filter(PERNUM == 1 & YEAR == 2010 & COSTELEC != 0 & COSTELEC <
     9990 & !is.na(in_metro)) %>%
  group_by(in_metro) %>%
  summarize(COSTELEC = weighted.mean(COSTELEC, HHWT)) %>%
  ggplot(aes(x = in_metro, y = COSTELEC)) +
    geom_col(fill = "royalblue3")
```

18. In this sample, is there a simple correlation between the number of rooms and the annual cost of electricity?

_____

_____

```
data_subset <- data %>%
  filter(PERNUM == 1 & COSTELEC > 0 & COSTELEC < 9990 & ROOMS >
     0) %>%
  select(COSTELEC, ROOMS, HHWT)


wtd_covariance <- cov.wt(data_subset %>% select(COSTELEC,
     ROOMS), wt = data_subset$HHWT, cor = TRUE)
wtd_covariance$cor
```

Next, create a graph that will display the average cost of gas and water 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 detached, single-family house with 5 rooms.

19. On the website, find the variable description for COSTGAS and note that gas costs are expressed in contemporary dollars. To adjust costs for inflation a price index, use CPI99. Go to the CPI99 variable description page. What year is the index year and how do you apply the inflation adjustment?

    _____

    _____

20. Did the annual cost of gas for a single family, 5-room home increase between 2005 and 2010 in nominal terms? What about the annual cost of water?

    _____

    _____

```
data_summary <- data %>%

filter(PERNUM == 1 & COSTGAS != 0 & COSTGAS < 9990 & COSTWATR !=
     0 & COSTWATR < 9990) %>%

filter(UNITSSTR == 3 & ROOMS == 5) %>%

group_by(YEAR = YEAR) %>%

summarize(COSTGAS = weighted.mean(COSTGAS, HHWT), COSTWATR =
     weighted.mean(COSTWATR, HHWT) ) %>%

gather(key, value, COSTGAS, 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"))
```


21. Has the annual cost of gas for a single family, 5 room home increased since 2005 in real terms? *Note: The variable CPI99 assigns an inflation index value according to the year of the observation.*

_____

_____

```
data_summary <- data %>%
     filter(PERNUM == 1 & COSTGAS != 0 & COSTGAS < 9990) %>%
filter(UNITSSTR == 3 & ROOMS == 5) %>%
     mutate(COSTGAS = COSTGAS * CPI99) %>%
     group_by(YEAR) %>%
     summarize( COSTGAS = weighted.mean(COSTGAS, HHWT) )
```

```
ggplot(data_summary, aes(x = YEAR, y = COSTGAS)) +
     geom_col(position = "dodge", fill = "darkblue") +
     theme(axis.text.x = element_text(angle = 20, hjust = 1))
```


Answers

Part 1: Frequencies

1. 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

2. 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%

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

Using household weights (HHWT)

4. What proportion of households in the sample had a mortgage? What proportion of the sample owned their home? *(Hint: don't use the weight quite yet)*
   42.20% of households mortgaged; 23.98% of household owned

5. What proportion of households across the country had a mortgage in 2010?
   40.53% of households

6. 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.

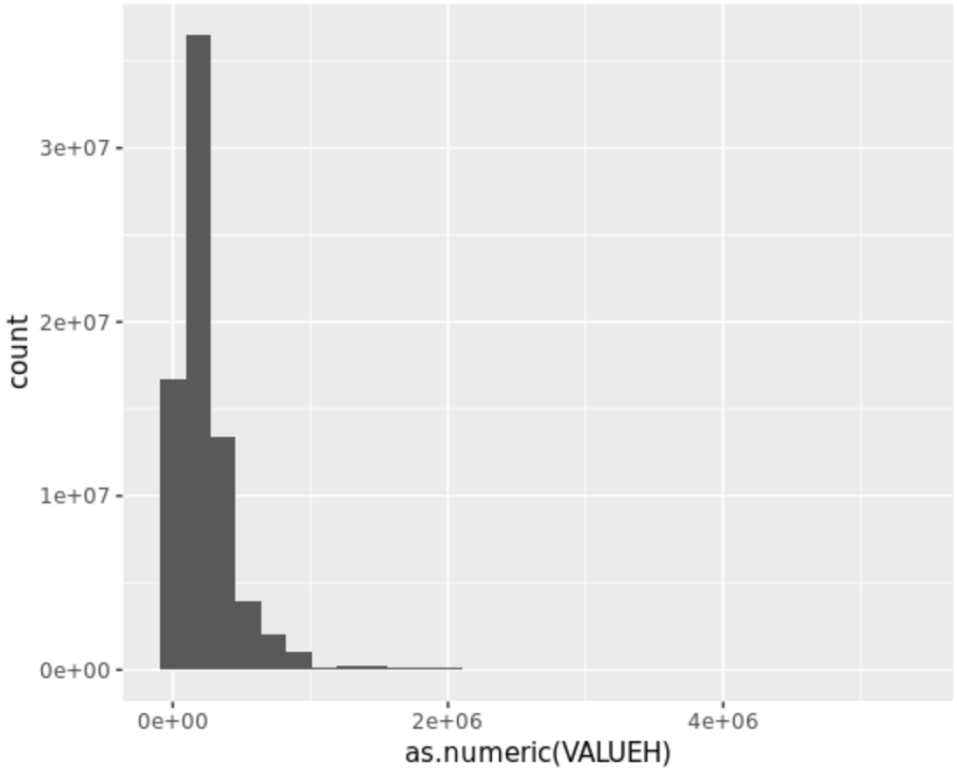
7. What is the average value of:
    a. A home that is mortgaged? $267,890
    b. A home that is owned? $219,015

8. 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.

9. Under the description tab on the website for VALUEH, read the first user note. Follow the relevant link on the codes tab to find the top codes for VALUEH (2010 ACS/PRCS topcodes by state). How could this complicate your data analysis? Create and check a histogram of your data to rule out any bias.

    There doesn't seem to be a significant cluster around the topcodes, so the data sample may not be noticeably biased.



Part 2: Frequencies

10. What were the three most commonly spoken languages in the US in 2010?

    English, Spanish, Chinese

11. 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

12. What percent of individuals who speak English at home:
    a. Has a mother who speaks Spanish at home? 3.89%
    b. Has a mother who speaks Chinese at home? 0.22%

13. What percent of men under the age of 30 speak Spanish at home? 13.4%

Part 3: Advanced Exercises

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

    METRO: 0 Not identifiable; 1 Not in metro area; 2 Central city; 3 Outside central city; 4 Central city status unknown; UNITSSTR: 03 1-family house, detached

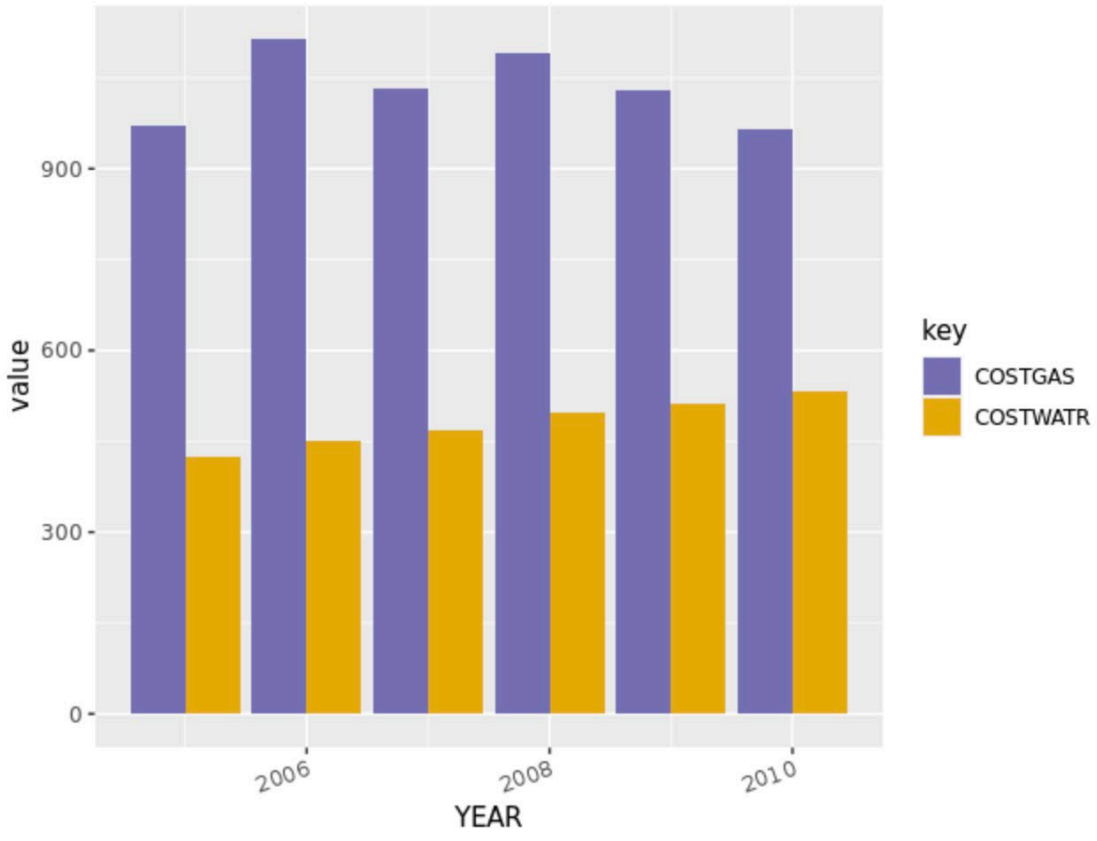
15. What is the proportion of households in the central city who owned their home:
    a. in 2008? 44.51%
    b. in 2010? 42.92%

16. What is the approximate annual cost of *water* for:
    a. A household in the metro area in 2010? ~$575
    b. A household not in the metro area? ~500

17. What is the approximate annual cost of *electricity* for:
    a. A household in the metro area in 2010? ~$1700
    b. A household not in the metro area? ~$1750

18. In this sample, 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. (0.30)

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 detached, single-family house with 5 rooms.



19. On the website, find the variable description for COSTGAS and note that gas costs are expressed in contemporary dollars. To adjust costs for inflation a price index, use CPI99. Go to the CPI99 variable description page. What year is the index year and how do you apply the inflation adjustment? 1999; real costs adjusted for inflation and indexed to the 1999 U.S. dollars are estimated by generating a new variable CPI99 * COSTGAS.
20. Did the annual cost of gas for a single family, 5-room home increased between 2005 and 2010 What about the annual cost of water?
    In nominal terms, the cost of gas has fluctuated over time, but the cost of water has steadily increased.
21. Did the annual cost of gas for a single family, 5 room home increased between 2005 and 2010 in real terms?
    In *real terms*, the gas prices fluctuated over time.