IPUMS USA Exercise 2 - R

OBJECTIVE: Gain an understanding of how an IPUMS USA dataset is structured and how it can be leveraged to explore your research interests. This exercise will use IPUMS USA to explore associations in household ownership, and trends in language spoken in the home.
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
- Analyze the data using sample code
- Validate data analysis work using answer key

IPUMS Variables
- MORTGAGE: Mortgage Status
- VALUEH: House Value
- LANGUAGE: Language spoken at home
- SEX: Sex
- AGE: Age
- METRO: Metropolitan status
- OWNERSHP: Ownership of dwelling
- COSTELEC: Annual electricity cost
- COSTGAS: Annual gas cost
- 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, if you prefer another programming style, please feel free to use it.

| %>% | The pipe operator eases coding nested functions – its read as "and then"
| as_factor | Converts the value labels provide 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 a variable

Answer Key (page 6)

Common Mistakes to Avoid
1. Not changing the working directory to the folder where your data is stored
2. Mixing up = and == ; To assign values to a variable, use "<" (or "="). Use "==" 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.
**Step 1**  
Make an Extract

**Step 2**  
Request the Data

### Registering with IPUMS

Go to [http://usa.ipums.org](http://usa.ipums.org), click LOG IN and create an IPUMS USA account. On login screen, enter email address and password and submit it!

- Similar to Exercise 1, create a data extract containing the following:
  - **Samples:** 2010 ACS (1-Yr)
  - **Variables:**
    - MORTGAGE: Mortgage Status
    - SEX: Sex
    - VALUEH: House value
    - AGE: Age
    - LANGUAGE: Language spoken at home
  - 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:
  - On the EXTRACT REQUEST page, click “ATTACH CHARACTERISTICS”. Check the box at the intersection of LANGUAGE and Mother, and SUBMIT

<table>
<thead>
<tr>
<th>Variable</th>
<th>Head</th>
<th>Father</th>
<th>Mother</th>
<th>Spouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERNUM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PERWT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEX</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LANGUAGE</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>LANGUAGE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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

Create a second extract without attaching any additional characteristics

- **Samples:** 2005 through 2010 ACS (1-Yr) Samples
- **Variables:**
  - METRO: Metropolitan status
  - OWNERSHIP: Ownership of dwelling
  - COSTWATR: Annual water cost
  - ROOMS: Number of rooms
  - COSTELEC: Annual electricity cost
  - UNITSSTR: Units in structure
  - COSTGAS: Annual gas cost
  - CPI99: 1999 Consumer Price Index
Part 1: Analyze the Sample

Analyze the Sample – Part I Frequencies

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?

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

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 from each household to represent that household’s characteristics.

D) 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)

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?

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?

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

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.
Part 2: Analyze the Sample

Analyze the Sample - Part II Frequencies

*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.*

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?

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

Part 3: Analyze the Sample

Analyze the Sample - Part III Advanced Exercises

*Now use 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?

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?
Part 3: Analyze the Sample

Analyze the Sample - Part III Advanced Exercises

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

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 note that electricity costs are expressed in contemporary dollars. To adjust costs for inflation a price index, CPI99, must be used. Go to the CPI99 variable description page. What year is the index year and how do you apply the inflation adjustment?

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?

H) 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.
ANSWERS: Analyze the Sample - Part I Frequencies
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 lived in homes that were mortgaged or had a deed of trust in 2010? What proportion of people in the sample lived in mortgaged homes? 1,523,041 people; 49.75%

C) Using weights, what proportion of the population lived in mortgaged homes in 2010? 47.46%

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: don’t use the weight quite yet) 42.20% of households mortgaged; 23.98% of households owned

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.

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.

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

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

**ANSWERS: Analyze the Sample - Part II Frequencies**

Investigate `LANGUAGE` variable frequencies.

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

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

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? 3.96%
   ii. Has a mother who speaks Chinese at home? 0.23%

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

**ANSWERS: Analyze the Sample - Part III Advanced Exercises**

*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%

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? ~$4500
   ii. A household not in the metro area? ~$2000

D) What is the approximate annual cost of electricity for:
i. A household in the metro area in 2010? ~$3800
   ii. A household not in the metro area? ~$1700

ANSWERS: Analyze the Sample - Part III Advanced Exercises

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)

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 (HINT: CPI99).
F) On the website, find the variable description for COSTELEC and note that electricity costs are expressed in contemporary dollars. To adjust costs for inflation a price index, CPI99, must be used. 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 * COSTELEC.

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? In nominal terms, the cost of gas is rising, but the cost of water seems not to change.

H) Has the annual cost of gas for a single family, 5 room home increased since 2005 in real terms? In real terms, the price is falling or staying about the same between 2005 and 2009.

Note: The variable CPI99 assigns an inflation index value according to the year of the observation.
EXERCISE 2

# Section 1: Analyze the Variables

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

# Section 2: Using weights

# D) 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)

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

# 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)
) +
  stat_bin(bins = 30)

# Analyze the Sample - Part II Frequencies
# Section 1: Analyze the Variables

# A) What were the three most commonly spoken languages in the US in 2010? ____________________________
# Note: The sort option automatically organizes the table into descending

data %>%
  group_by(LANGUAGE = haven::as_factor(LANGUAGE, level = "both")) %>%
Code

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? ________________

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?

# 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? ______

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 = haven::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"))