IPUMS Data Training Exercise

An Introduction to IPUMS PMA

(Exercise 2 for R)

Learning goals

- Create and download an IPUMS PMA data extract
- Decompress data file and read data in R
- Analyze the data using sample code

Summary

In this exercise, you will gain an understanding of how IPUMS PMA service delivery point datasets are structured and how it can be leveraged to explore your research interests. This exercise will use the PMA dataset to explore basic frequencies of births, facilities and rural or urban locations. You will create data extracts that include the variables: EAID, FACILITYTYPEGEN, FACILITYADV, PILLOSBS, PILLOUTDAY, and URBAN; then, you will use a sample code to analyze these data.

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
ggplot	Initializes a graphic object (histogram, box, plot, etc.)

Common Mistakes to Avoid

- Not changing the working directory to the folder where your data is stored.

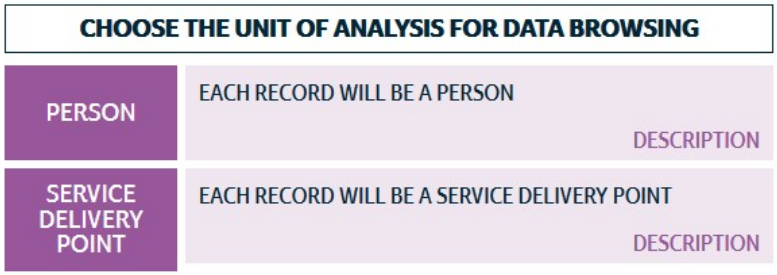
- Mixing up = and ==; to assign a value in generating a variable, use "<-" (or "="). Use "==" to test for equality.

Registering with IPUMS

Go to http://pma.ipums.org, click on Register to Use IPUMS PMA on the left hand side of the screen. Click the Register for IPUMS PMA button and fill out the form to apply for access. You will have to wait for your account to be approved to access the data. Once you receive the approval email, click "Log In" at the top of the page and use your email and password.

Select Samples

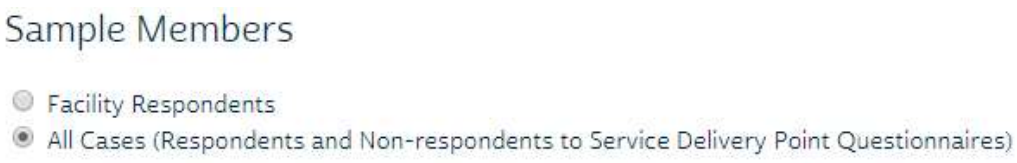
- Choose the Service Delivery Point unit of analysis



- Click the Select Samples box, check the box for the Kenya 2016 R5

- Scroll to the bottom of the page and click the radio button option for All Cases. The default is Facility Respondents

Click the Submit Sample Selections box



Select Variables

- The search tool allows you to search for variables. Observe the options for limiting your search results by variable characteristics or variable type.
- You may add a variable to your cart by clicking on the plus sign in the "Add to Cart" column of the topical variable list, or list of search results.
- You may view information about the variable by clicking on the variable name, and navigating through the tabs that include a description of the variable, codes and value labels, the universe of persons asked the question, and information on the comparability of the variable among other pieces of information. If you are reviewing variable-specific information, you may click on the "Add to Cart" button near the top of the screen to add this variable to your data cart.
- Using the drop down menu or search feature, select the following variables:

  EAID: Enumeration area (primary sampling unit)

  FACILITYTYPEGEN: Type of facility

  FACILITYADV: Advanced facility

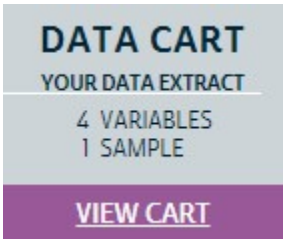
  PILLOBS: Observed and in or out of stock of birth control pills

  PILLOUTDAY: Number of days birth control pills have been out of stock

URBAN: Urban or rural status

Review and submit your extract

Click the purple VIEW CART button under your data cart



Review variable selection. Note that certain variables appear in your data cart even if you did not select them, and they are not included in the constantly updated count of variables in your data cart. The preselected variables are needed for weighting, for variance estimation, or to identify the year, country, and round of a sample.

- Click the Create Data Extract button
- Review the 'Extract Request Summary' screen, describe your extract and click Submit Extract
- You will get an email when the data is available to download.
- To get to the page to download the data, follow the link in the email, or follow the My Data Extracts link on the homepage.

Getting the data into your statistics software

Download the data

Go to http://pma.ipums.org/ and click on My Data Extracts

Extract Number	Date	Formatted Data	Fixed-width Text Files						
			Data	Command Files				Codebook	
51	2018-10-26	--	Download .DAT	SPS	SAS	STATA	R	Basic	DDI

- 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

Install and load packages for R

Open R from the Start menu

If you haven't already installed any of the following packages, type:

```
install.packages("ipumsr")
install.packages("dplyr")
install.packages("ggplot2")
```

Next (or if you have already installed the packages on your computer), type:

```
library(ipumsr)
library(dplyr)
library(ggplot2)
options(tibble.print_max = Inf)
```

Read data into R

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 (replace the #s below with the number of your extract):

```
ddi <- read_ipums_ddi("pma_000##.xml")
SDP <- read_ipums_micro(ddi)
```

This exercise demonstrates how to merge SDP data to the HHF dataset used in Exercise 1. Please see instructions in Exercise 1 to ensure that your HHF extract contains the variables used below. Load the HHF data using the following commands (your HHF extract will have different #s than the SDP extract):

```
ddi <- read_ipums_ddi("pma_000##.xml")
HHF <- read_ipums_micro(ddi)
```

NOTE: 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")`

Analyze the Sample

Part 1: Exploring Facility Types

1. Create a frequency table for FACILITYTYPEGEN showing the proportion of each type of facility surveyed in Kenya 2016 Round 5. _____

```
SDP %>%
 count(type <- as_factor(FACILITYTYPEGEN)) %>%
 mutate(prop = prop.table(n))
```

2. According to the Universe tab, what facilities are included in the surveyed universe for FACILITYTYPEGEN _____
   _____

3. Users should note that many variables in the service delivery point (SDP) survey have a universe defined by FACILITYADV, a country-specific designation of "advanced facility" types. Create a crosstab to see which types of facilities from the previous question were designated as "advanced facilities" in Kenya for 2016.

```
SDP %>%
  mutate(ADVANCED = FACILITYADV==1)%>%
  group_by(as_factor(FACILITYTYPEGEN), ADVANCED)%>%
  summarize()
```

4. Consult the Comparability tab for FACILITYADV, taking care to note that advanced facility designations vary by country, and sometimes vary by survey round within a country. Locate the entry for Kenya, and determine whether its advanced facility designation matches what you found in Question 3. Is the designation consistent for

all Kenyan survey rounds that included this variable? _____

_____

_____

Part 2: Descriptive Statistics

5. Consider the variable PILLOBS, which describes whether the SDP had an observable stock of birth control pills on the day of the interview. According to the Codes tab, what are the possible responses for SDPs surveyed in Kenya 2016?

_____

_____

6. According to the Comparability tab, possible responses to PILLOBS may vary from sample to sample. How so? _____

_____

7. According to the Universe tab, what facilities are included in the surveyed universe for PILLOBS? _____

_____

8. Among facilities that usually provide birth control pills shown in PILLOBS, what type of facility was least likely to have supplies of birth control pills in-stock on the day of the interview? What proportion of facilities of this type were out of stock? (Restrict analysis only to completed interviews and in-universe cases).

_____

_____

```
SDP%>%
  filter(PILLOBS < 90)%>%
  count(FACILITYTYPEGEN, PILLOBS)%>%
  group_by(FACILITYTYPEGEN)%>%
  mutate(type = as_factor(FACILITYTYPEGEN))%>%
  mutate(obs = as_factor(PILLOBS))%>%
  mutate(prop_type = prop.table(n))%>%
  select(type, obs, n, prop_type)
```

Part 3: Data Visualization

For facilities that were out of birth control pills, PILLOUTDAY shows the number of days that supplies had been unavailable. Because some SDPs had been out of stock for more than 90 days, NIU and missing value codes for PILLOUTDAY are coded as values 9994, 9997, and 9999 in order to exceed the range of valid responses.

9. Calculate the mean shortage of pills for *all* in-universe facilities in PILLOUTDAY, taking care to exclude any value above 9000. Then find the mean for *each facility type* in FACILITYTYPEGEN, and display the result as a bar chart. (Restrict analysis only to valid responses from SDPs in universe for PILLOUTDAY).

```
SDP%>%
  filter(PILLOUTDAY < 9000)%>%
  summarise(mean(PILLOUTDAY))
SDP%>%
  filter(PILLOUTDAY < 9000)%>%
  group_by(as_factor(FACILITYTYPEGEN))%>%
  summarise(mean(PILLOUTDAY))
```


```
SDP%>%
  filter(PILLOUTDAY < 9000)%>%
  group_by(facility_type = as_factor(FACILITYTYPEGEN))%>%
  summarise(mean_days = mean(PILLOUTDAY))%>%
  ggplot() + geom_col(aes(x =  facility_type, y = mean_days)) +
coord_flip()
```

10. Suppose you suspect that the apparent difference between the facilities in 9 is really a disparity between types of facilities that are most likely to be found in urban vs. rural areas. Create a pair of bar charts groups by URBAN to test if this is true. Are there differences between urban and rural facilities of each type?

```
SDP%>%
  filter(PILLOUTDAY < 9000)%>%
  group_by(facility_type = as_factor(FACILITYTYPEGEN), urban =
as_factor(URBAN))%>%
  summarize(mean_days = mean(PILLOUTDAY))%>%
  ggplot(aes(x = facility_type, y=mean_days)) +
geom_col(aes(fill = urban), position = position_dodge()) +
coord_flip()
```

Part 4: Combining SDP and HHF Data

Users should note that PMA2020 surveyed facilities in the same sampling areas as households and females in the same survey round. These SDP data are *not meant to be nationally representative.* Instead, they are meant to portray the health provision environment of the surveyed households and women. Thus, there are no sampling weights for SDP variables.

The files do contain a weight for the sampling units EAWEIGHT, which is a probability weight representing the likelihood of the enumeration area (EA) being selected for sampling. The collectors of the original data do not recommend using EAWEIGHT to weight SDP variables. Rather, the best use of SDP variables is to calculate summary statistics at the EA level and attach them to the Household and Female (HHF) dataset using the EAID variable as a source of contextual information for each woman's service delivery environment.

For example, one could use the variables PILLOBS and PILLOUTDAY to calculate whether any facility in each EAID was out of stock of birth control pills and the mean number of days the facility or facilities in each EAID were out of stock of pills, respectively. These summary statistics may be merged with the HHF dataset in order to show whether each female respondent had reliable local access to birth control pills.

11. Create a table showing the number of women aged 15-49 (ELIGIBLE == 1) sampled in the Kenya 2016 Round 5 Household and Female dataset (HHF) who resided in each enumeration area where birth control pills were not available at all local facilities in the SDP survey. How many enumeration areas in Kenya 2016 meet these criteria?_____

```
HHF%>%
  mutate(pillobs = HHF$EAID %in% subset(SDP$EAID, SDP$PILLOBS ==
3))%>%
```


```
group_by(EAID)%>%
filter(pillobs==TRUE & ELIGIBLE==1)%>%
count(pillobs)%>%
select(EAID, n)
```

12. Looking at the table created in 11, what is *the total number* of sampled women aged 15-49 (ELIGIBLE == 1) in the Kenya 2016 Round 5 Household and Female dataset (HHF) who resided in an enumeration area where birth control pills were not available at all local facilities in the SDP survey. _____

13. Run a logistic regression model to predict the association between women currently using the pill (FPNOWUSPILL) and the mean shortage duration (PILLOUTDAY) for each enumeration area that was out of pills on the day of the SDP interview. Adjust your model to be representative of all Kenyan women using FQWEIGHT. Recode values for FPNOWUSPILL that are not in universe or missing to zero.

Is there an association between the number of days that the facilities in the woman's enumeration area are out of stock of pills and the woman's current use of the pill for family planning? _____

```
model_data <- HHF%>%

  left_join(SDP%>%

              group_by(EAID)%>%

              summarize(pilloutday = mean(subset(PILLOUTDAY,
PILLOUTDAY < 9000))))%>%

  mutate(fpnowuspill = case_when(as.numeric(FPNOWUSPILL) > 90 ~
0, TRUE ~ as.numeric(FPNOWUSPILL)))%>%

  mutate(pilloutday = case_when(is.na(pilloutday) ~ 0, TRUE ~
pilloutday))



model <- glm(FPNOWUSPILL ~ pilloutday, data = model_data, family
= binomial, weights = round(FQWEIGHT))

summary(model)

exp(coef(model))
```


ANSWERS

Part 1: Exploring Facility Types

1. Create a frequency table for FACILITYTYPEGEN showing the proportion of each type of facility surveyed in Kenya 2016 Round 5.

```
> SDP %>%
+   count(type <- as_factor(FACILITYTYPEGEN)) %>%
+   mutate(prop = prop.table(n))
# A tibble: 7 x 3
  `type <- as_factor(FACILITYTYPEGEN)`     n    prop
  <fct>                                <int>   <dbl>
1 Hospital                                79 0.185
2 Health center                           90 0.210
3 Health clinic                           16 0.0374
4 Other health facility                    1 0.00234
5 Dispensary                             190 0.444
6 Pharmacy/chemist/drug shop              48 0.112
7 Other                                    4 0.00935
```

2. According to the Universe tab, what facilities are included in the surveyed universe for FACILITYTYPEGEN? All service delivery points

3. Users should note that many variables in the service delivery point (SDP) survey have a universe defined by FACILITYADV, a country-specific designation of "advanced facility" types. Create a crosstab to see which types of facilities from the previous question were designated as "advanced facilities" in Kenya for 2016. All are advanced, except for Pharmacy / Chemist / Drug Shop

```
> SDP %>%
+   mutate(ADVANCED = FACILITYADV==1)%>%
+   group_by(as_factor(FACILITYTYPEGEN), ADVANCED)%>%
+   summarize()
# A tibble: 7 x 2
# Groups:   as_factor(FACILITYTYPEGEN) [?]
  `as_factor(FACILITYTYPEGEN)` ADVANCED
  <fct>                        <lgl>
1 Hospital                     TRUE
2 Health center                TRUE
3 Health clinic                TRUE
4 Other health facility        TRUE
5 Dispensary                   TRUE
6 Pharmacy/chemist/drug shop   FALSE
7 Other                        TRUE
```


4. Consult the Comparability tab for FACILITYADV, taking care to note that advanced facility designations vary by country, and sometimes vary by survey round within a country. Locate the entry for Kenya, and determine whether its advanced facility designation matches what you found in Question C. Is the designation consistent for all Kenyan survey rounds that included this variable? It does match, and all Kenyan rounds interviewed have the same designation._

Part 2: Descriptive Statistics

5. Consider the variable PILLOBS, which describes whether the SDP had an observable stock of birth control pills on the day of the interview. According to the Codes tab, what are the possible responses for SDPs surveyed in Kenya 2016?

1 - In-stock and observed	94 - Not interviewed (SDP questionnaire)
2 - In-stock but not observed	98 - No response or missing
3 - Out of stock	99 - NIU (not in universe)

6. According to the Comparability tab, possible responses to PILLOBS may vary from sample to sample. How so? Some early samples include less detail, providing dichotomous responses based on whether the interviewer observed contraceptive pills in-stock. In these early samples, if contraceptive pills were not observed, they were assumed to be "out of stock". In later surveys, interviewers had the option of reporting that contraceptive pills were "in-stock but not observed".

7. According to the Universe tab, what facilities are included in the surveyed universe for PILLOBS? Service delivery points that provide contraceptive pills.

8. Among facilities that usually provide birth control pills shown in PILLOBS, what type of facility was least likely to have supplies of birth control pills in-stock on the day of the interview? What proportion of facilities of this type were out of stock? (Restrict analysis only to completed interviews and in-universe cases). Health clinics were most likely to be out of pills with 25% out of stock.

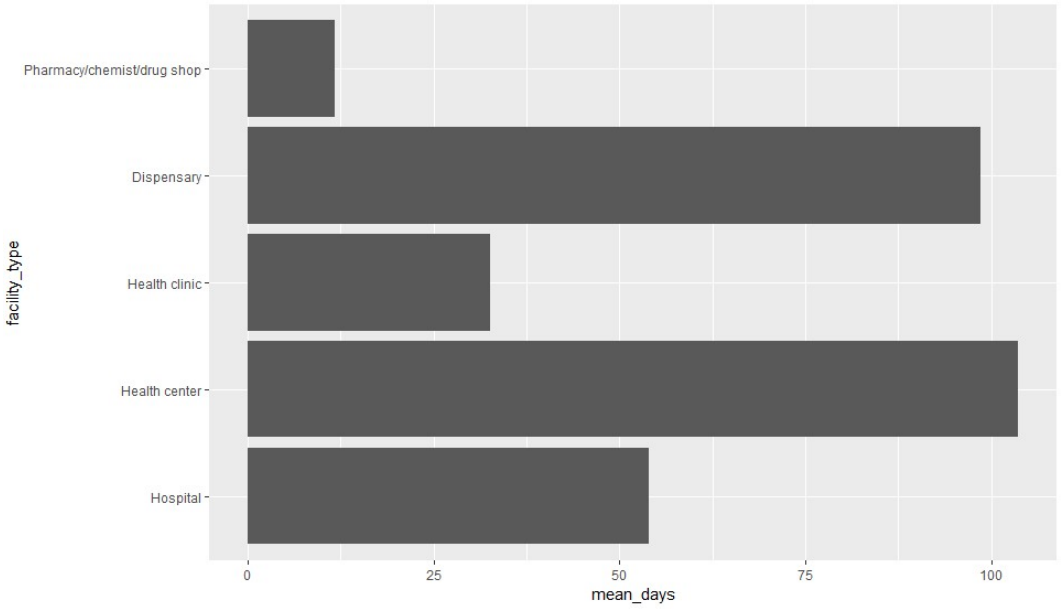
```
> SDP%>%
+   filter(PILLOBS < 90)%>%
+   count(FACILITYTYPEGEN, PILLOBS)%>%
+   group_by(FACILITYTYPEGEN)%>%
+   mutate(type = as_factor(FACILITYTYPEGEN))%>%
+   mutate(obs = as_factor(PILLOBS))%>%
+   mutate(prop_type = prop.table(n))%>%
+   select(type, obs, n, prop_type)
Adding missing grouping variables: `FACILITYTYPEGEN`
# A tibble: 17 x 5
# Groups:   FACILITYTYPEGEN [7]
   FACILITYTYPEGEN type                  obs                        n prop_type
   <int+lbl>       <fct>                 <fct>                  <int>     <dbl>
 1 1               Hospital              In-stock and observed     69    0.896
 2 1               Hospital              In-stock but not observed  1    0.0130
 3 1               Hospital              Out of stock               7    0.0909
 4 2               Health center         In-stock and observed     69    0.812
 5 2               Health center         In-stock but not observed  2    0.0235
 6 2               Health center         Out of stock              14    0.165
 7 3               Health clinic         In-stock and observed      8    0.667
 8 3               Health clinic         In-stock but not observed  1    0.0833
 9 3               Health clinic         Out of stock               3    0.25
10 4               Other health facility In-stock and observed      1    1
11 6               Dispensary            In-stock and observed    144    0.783
12 6               Dispensary            In-stock but not observed  3    0.0163
13 6               Dispensary            Out of stock              37    0.201
14 7               Pharmacy/chemist/drug shop In-stock and observed 36    0.878
15 7               Pharmacy/chemist/drug shop In-stock but not observed 2 0.0488
16 7               Pharmacy/chemist/drug shop Out of stock           3    0.0732
17 9               Other                 In-stock and observed      1    1
```

Part 3: Data Visualization

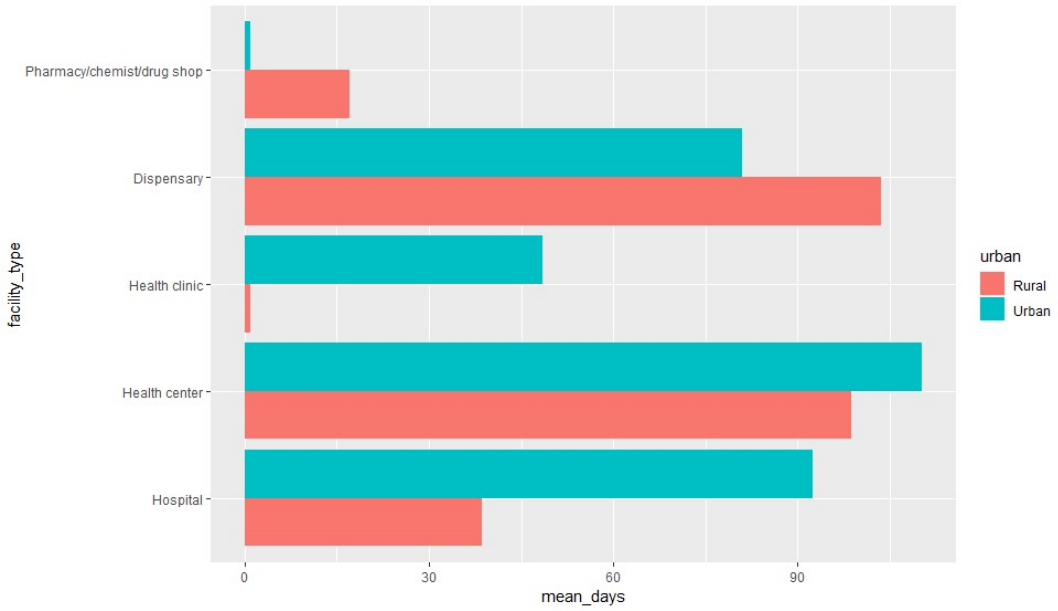
9. Calculate the mean shortage of pills for *all* in-universe facilities in PILLOUTDAY, taking care to exclude any value above 9000. Then find the mean for *each facility type* in FACILITYTYPEGEN, and display the result as a bar chart. (Restrict analysis only to valid responses from SDPs in universe for PILLOUTDAY).

```
> SDP%>%
+   filter(PILLOUTDAY < 9000)%>%
+   summarise(mean(PILLOUTDAY))
# A tibble: 1 x 1
  `mean(PILLOUTDAY)`
               <dbl>
1               87.5
```

```
> SDP%>%
+   filter(PILLOUTDAY < 9000)%>%
+   group_by(facility_type = as_factor(FACILITYTYPEGEN), urban = as_factor(URBAN))%>%
+   summarize(mean(PILLOUTDAY))
# A tibble: 10 x 3
# Groups:   facility_type [?]
   facility_type              urban `mean(PILLOUTDAY)`
   <fct>                      <fct>              <dbl>
 1 Hospital                   Rural               38.6
 2 Hospital                   Urban               92.5
 3 Health center              Rural               98.8
 4 Health center              Urban              110.
 5 Health clinic              Rural                1
 6 Health clinic              Urban               48.5
 7 Dispensary                 Rural              104.
 8 Dispensary                 Urban               80.9
 9 Pharmacy/chemist/drug shop Rural               17
10 Pharmacy/chemist/drug shop Urban                1
```



10. Suppose you suspect that the apparent difference between the facilities in 9 is really a disparity between types of facilities that are most likely to be found in urban vs. rural areas. Create a pair of bar charts groups by URBAN to test if this is true. Are there differences between urban and rural facilities of each type? Yes, there are differences in each facility type. Rural Pharmacy/Chemist/Drug shops and Dispensaries have higher mean days than the urban of the same facilities, and the other facilities (Health Clinic, Health Center, and Hospital) have a higher mean in Urban areas than rural



Part 4: Combining SDP and HHF Data

11. Create a table showing the number of women aged 15-49 (ELIGIBLE == 1) sampled in the Kenya 2016 Round 5 Household and Female dataset (HHF) who resided in each enumeration area where birth control pills were not available at all local facilities in the SDP survey. How many enumeration areas in Kenya 2016 meet these criteria?

43 enumeration areas.

```
> HHF%>%
+   mutate(pillobs = HHF$EAID %in% subset(SDP$EAID, SDP$PILLOBS == 3))%>%
+   group_by(EAID)%>%
+   filter(pillobs==TRUE & ELIGIBLE==1)%>%
+   count(pillobs)%>%
+   select(EAID, n)
# A tibble: 43 x 2
# Groups:   EAID [43]
    EAID     n
   <dbl> <int>
 1  4013    33
 2  4047    44
 3  4163    47
 4  4207    38
 5  4212    23
 6  4214    38
 7  4234    50
 8  4245    45
 9  4318    36
10  4336    41
11  4356    47
12  4361    45
13  4373    39
14  4431    40
15  4456    50
16  4485    51
17  4531    42
18  4626    23
19  4628    27
20  4639    36
21  4655    39
22  4662    34
23  4676    56
24  4677    44
25  4707    56
26  4711    44
27  4712    22
28  4719    46
29  4760    52
30  4768    29
31  4784    50
32  4795    63
33  4871    51
34  4885    25
35  4887    45
36  4895    55
37  4899    34
38  4905    38
39  4938    36
40  4952    54
41  4963    39
42  4965    36
43  4974    34
```


12. Looking at the table created in A), what is *the total number* of sampled women aged 15-49 (ELIGIBLE == 1) in the Kenya 2016 Round 5 Household and Female dataset (HHF) who resided in an enumeration area where birth control pills were not available at all local facilities in the SDP survey?

These are the first three EAIDs in the list:

EAID 4013 = 33 women EAID 4163 = 47 women

EAID 4047 = 44 women

13. Run a logistic regression model to predict the association between women currently using the pill (FPNOWUSPILL) and the mean shortage duration (PILLOUTDAY) for each enumeration area that was out of pills on the day of the SDP interview…. Is there such an association?

The likelihood that a sampled woman uses birth control pills remains the same regardless of the average number of days that her local SDP had none available (odds ratio = 1.000). However, this finding is not statistically significant (p = 0.728).

FPNOWUSPILL	Odds Ratio	P value	95% CI
PILLOUTDAY	1.000	0.728	(0.998, 1.002)

```
> model_data <- HHF%>%
+   left_join(SDP%>%
+               group_by(EAID)%>%
+               summarize(pilloutday = mean(subset(PILLOUTDAY, PILLOUTDAY < 9000))))%>%
+   mutate(fpnowuspill = case_when(as.numeric(FPNOWUSPILL) > 90 ~ 0, TRUE ~ as.numeric(FPNOWUSPILL)))%>%
+   mutate(pilloutday = case_when(is.na(pilloutday) ~ 0, TRUE ~ pilloutday))
Joining, by = "EAID"
>
> model <- glm(FPNOWUSPILL ~ pilloutday, data = model_data, family = binomial, weights = round(FQWEIGHT))
> summary(model)

Call:
glm(formula = FPNOWUSPILL ~ pilloutday, family = binomial, data = model_data,
    weights = round(FQWEIGHT))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5405   0.0000   0.0000   0.0000   4.2755

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.9979074  0.0649217 -46.177   <2e-16 ***
pilloutday  -0.0003712  0.0010675  -0.348    0.728
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2378.5  on 5521  degrees of freedom
Residual deviance: 2378.4  on 5520  degrees of freedom
AIC: 2382.4

Number of Fisher Scoring iterations: 5

> exp(coef(model))
(Intercept)  pilloutday
 0.04989136  0.99962886
```