MITx: 15.071x The Analytics Edge

Bookmarks

Unit 1: An Introduction to Analytics

Welcome to Unit 1

Initial Evaluation
Evaluations due Apr 26, 2016 at 00:00 UTC

The Analytics Edge: Intelligence, Happiness, and Health (Lecture Sequence)

Working with Data: An Introduction to R
Lecture Sequence Quick Questions

Understanding Food: Nutritional Education with Data (Recitation)

Assignment 1
Homework due Apr 26, 2016 at 00:00 UTC

Entrance Survey

Unit 1: An Introduction to Analytics > Assignment 1 > Demographics and Employment in the United States

Bookmark

DEMOGRAPHICS AND EMPLOYMENT IN THE UNITED STATES

In the wake of the Great Recession of 2009, there has been a good deal of focus on employment statistics, one of the most important metrics policymakers use to gauge the overall strength of the economy. In the United States, the government measures unemployment using the Current Population Survey (CPS), which collects demographic and employment information from a wide range of Americans each month. In this exercise, we will employ the topics reviewed in the lectures as well as a few new techniques using the September 2013 version of this rich, nationally representative dataset (available online).

The observations in the dataset represent people surveyed in the September 2013 CPS who actually completed a survey. While the full dataset has 385 variables, in this exercise we will use a more compact version of the dataset, CPSData.csv, which has the following variables:

PeopleInHousehold: The number of people in the interviewee's household.

Region: The census region where the interviewee lives.

State: The state where the interviewee lives.

MetroAreaCode: A code that identifies the metropolitan area in which the interviewee lives (missing if the interviewee does not live in a metropolitan area). The mapping from codes to names of metropolitan areas is provided in the file MetroAreaCodes.csv.

Age: The age, in years, of the interviewee. 80 represents people aged 80-84, and 85 represents people aged 85 and higher.

Married: The marriage status of the interviewee.

Sex: The sex of the interviewee.

Education: The maximum level of education obtained by the interviewee.

Race: The race of the interviewee.

Hispanic: Whether the interviewee is of Hispanic ethnicity.

CountryOfBirthCode: A code identifying the country of birth of the interviewee. The mapping from codes to names of countries is provided in the file CountryCodes.csv.

Citizenship: The United States citizenship status of the interviewee.

EmploymentStatus: The status of employment of the interviewee.

Industry: The industry of employment of the interviewee (only available if they are employed).

Problem 1.1 - Loading and Summarizing the Dataset

(1/1 point)

Load the dataset from CPSData.csv into a data frame called CPS, and view the dataset with the summary() and str() commands.

How many interviewees are in the dataset?

131302 ✓

You have used 1 of 3 submissions

Problem 1.2 - Loading and Summarizing the Dataset

(1/1 point)

Among the interviewees with a value reported for the Industry variable, what is the most common industry of employment? Please enter the name exactly how you see it.

Educational and health ✓

You have used 1 of 2 submissions

Problem 1.3 - Loading and Summarizing the Dataset

(2/2 points)

Recall from the homework assignment "The Analytical Detective" that you can call the sort() function on the output of the table() function to obtain a sorted breakdown of a variable. For instance, sort(table(CPS$Region)) sorts the regions by the number of interviewees from that region.

Which state has the fewest interviewees?

New Mexico ✓

Which state has the largest number of interviewees?

California ✓

You have used 1 of 3 submissions

Problem 1.4 - Loading and Summarizing the Dataset

(1/1 point)

What proportion of interviewees are citizens of the United States?

0.9421943 ✓

You have used 1 of 3 submissions

Problem 1.5 - Loading and Summarizing the Dataset

(1/1 point)
The CPS differentiates between race (with possible values American Indian, Asian, Black, Pacific Islander, White, or Multiracial) and ethnicity. A number of interviewees are of Hispanic ethnicity, as captured by the Hispanic variable. For which races are there at least 250 interviewees in the CPS dataset of Hispanic ethnicity? (Select all that apply.)

☑ American Indian

□ Asian

☑ Black

☑ Multiracial

□ Pacific Islander

☑ White

✓

You have used 1 of 2 submissions

Problem 2.1 - Evaluating Missing Values

(1/1 point)
Which variables have at least one interviewee with a missing (NA) value? (Select all that apply.)

□ PeopleInHousehold

□ Region

□ State

☑ MetroAreaCode

□ Age

☑ Married

□ Sex

☑ Education

□ Race

□ Hispanic

□ CountryOfBirthCode

□ Citizenship

☑ EmploymentStatus

☑ Industry

✓

You have used 1 of 2 submissions

Problem 2.2 - Evaluating Missing Values

(1/1 point)

Often when evaluating a new dataset, we try to identify if there is a pattern in the missing values in the dataset. We will try to determine if there is a pattern in the missing values of the Married variable. The function is.na(CPS$Married) returns a vector of TRUE/FALSE values for whether the Married variable is missing. We can see the breakdown of whether

Married is missing based on the reported value of the Region variable with the function table(CPS$Region, is.na(CPS$Married)). Which is the most accurate:

○ The Married variable being missing is related to the Region value for the interviewee.

○ The Married variable being missing is related to the Sex value for the interviewee.

● The Married variable being missing is related to the Age value for the interviewee. ✓

○ The Married variable being missing is related to the Citizenship value for the interviewee.

○ The Married variable being missing is not related to the Region, Sex, Age, or Citizenship value for the interviewee.

You have used 1 of 2 submissions

Problem 2.3 - Evaluating Missing Values

(2/2 points)

As mentioned in the variable descriptions, MetroAreaCode is missing if an interviewee does not live in a metropolitan area. Using the same technique as in the previous question, answer the following questions about people who live in non-metropolitan areas.

How many states had all interviewees living in a non-metropolitan area (aka they have a missing MetroAreaCode value)? For this question, treat the District of Columbia as a state (even though it is not technically a state).

2 ✓

How many states had all interviewees living in a metropolitan area? Again, treat the District of Columbia as a state.

3 ✓

You have used 2 of 3 submissions

Problem 2.4 - Evaluating Missing Values

(1/1 point)

Which region of the United States has the largest proportion of interviewees living in a non-metropolitan area?

● Midwest ✓

○ Northeast

○ South

○ West

You have used 1 of 1 submissions

Problem 2.5 - Evaluating Missing Values

(4/4 points)

While we were able to use the table() command to compute the proportion of interviewees from each region not living in a metropolitan area, it was somewhat tedious (it involved manually computing the proportion for each region) and isn't something you would want to do if there were a larger number of options. It turns out there is a less tedious way to compute the proportion of values that are TRUE. The mean() function, which takes the average of the values passed to it, will treat TRUE as 1 and FALSE as 0, meaning it returns the proportion of values that are true. For instance, mean(c(TRUE, FALSE, TRUE, TRUE)) returns 0.75. Knowing this, use tapply() with the mean function to answer the following questions:

Which state has a proportion of interviewees living in a non-metropolitan area closest to 30%?

Wisconsin ✓

Which state has the largest proportion of non-metropolitan interviewees, ignoring states where all interviewees were non-metropolitan?

Montana ✓

You have used 1 of 4 submissions

Problem 3.1 - Integrating Metropolitan Area Data

(2/2 points)

Codes like MetroAreaCode and CountryOfBirthCode are a compact way to encode factor variables with text as their possible values, and they are therefore quite common in survey datasets. In fact, all but one of the variables in this dataset were actually stored by a numeric code in the original CPS datafile.

When analyzing a variable stored by a numeric code, we will often want to convert it into the values the codes represent. To do this, we will use a dictionary, which maps the the code to the actual value of the variable. We have provided dictionaries MetroAreaCodes.csv and CountryCodes.csv, which respectively map MetroAreaCode and CountryOfBirthCode into their true values. Read these two dictionaries into data frames MetroAreaMap and CountryMap.

How many observations (codes for metropolitan areas) are there in MetroAreaMap?

271 ✓

How many observations (codes for countries) are there in CountryMap?

149 ✓

You have used 1 of 3 submissions

Problem 3.2 - Integrating Metropolitan Area Data

(2/2 points)

To merge in the metropolitan areas, we want to connect the field MetroAreaCode from the CPS data frame with the field Code in MetroAreaMap. The following command merges the two data frames on these columns, overwriting the CPS data frame with the result:

```
CPS = merge(CPS, MetroAreaMap, by.x="MetroAreaCode", by.y="Code", all.x=TRUE)
```

The first two arguments determine the data frames to be merged (they are called "x" and "y", respectively, in the subsequent parameters to the merge function). by.x="MetroAreaCode" means we're matching on the MetroAreaCode variable from the "x" data frame (CPS), while by.y="Code" means we're matching on the Code variable from the "y" data frame (MetroAreaMap). Finally, all.x=TRUE means we want to keep all rows from the "x" data frame (CPS), even if some of the rows' MetroAreaCode doesn't match any codes in MetroAreaMap (for those familiar with database terminology, this parameter makes the operation a left outer join instead of an inner join).

Review the new version of the CPS data frame with the summary() and str() functions. What is the name of the variable that was added to the data frame by the merge() operation?

MetroArea ✓

How many interviewees have a missing value for the new metropolitan area variable? Note that all of these interviewees would have been removed from the merged data frame if we did not include the all.x=TRUE parameter.

34238 ✓

You have used 1 of 3 submissions

Problem 3.3 - Integrating Metropolitan Area Data

(1/1 point)

Which of the following metropolitan areas has the largest number of interviewees?

○ Atlanta-Sandy Springs-Marietta, GA

○ Baltimore-Towson, MD

● Boston-Cambridge-Quincy, MA-NH ✓

○ San Francisco-Oakland-Fremont, CA

You have used 1 of 1 submissions

Problem 3.4 - Integrating Metropolitan Area Data

(2/2 points)
Which metropolitan area has the highest proportion of interviewees of Hispanic ethnicity? Hint: Use tapply() with mean, as in the previous subproblem. Calling sort() on the output of tapply() could also be helpful here.

Laredo, TX ✓

You have used 1 of 5 submissions

Problem 3.5 - Integrating Metropolitan Area Data

(2/2 points)
Remembering that CPS$Race == "Asian" returns a TRUE/FALSE vector of whether an interviewee is Asian, determine the number of metropolitan areas in the United States from which at least 20% of interviewees are Asian.

4 ✓

You have used 1 of 5 submissions

Problem 3.6 - Integrating Metropolitan Area Data

(1/1 point)

Normally, we would look at the sorted proportion of interviewees from each metropolitan area who have not received a high school diploma with the command:

sort(tapply(CPS$Education == "No high school diploma", CPS$MetroArea, mean))

However, none of the interviewees aged 14 and younger have an education value reported, so the mean value is reported as NA for each metropolitan area. To get mean (and related functions, like sum) to ignore missing values, you can pass the parameter na.rm=TRUE. Passing na.rm=TRUE to the tapply function, determine which metropolitan area has the smallest proportion of interviewees who have received no high school diploma.

Iowa City, IA ✓

You have used 1 of 3 submissions

Problem 4.1 - Integrating Country of Birth Data

(2/2 points)

Just as we did with the metropolitan area information, merge in the country of birth information from the CountryMap data frame, replacing the CPS data frame with the result. If you accidentally overwrite CPS with the wrong values, remember that you can restore it by re-loading the data frame from CPSData.csv and then merging in the metropolitan area information using the command provided in the previous subproblem.

What is the name of the variable added to the CPS data frame by this merge operation?

Country ✓

How many interviewees have a missing value for the new country of birth variable?

176 ✓

You have used 1 of 3 submissions

Problem 4.2 - Integrating Country of Birth Data

(2/2 points)

Among all interviewees born outside of North America, which country was the most common place of birth?

Philippines ✓

You have used 1 of 5 submissions

Problem 4.3 - Integrating Country of Birth Data

(2/2 points)

What proportion of the interviewees from the "New York-Northern New Jersey-Long Island, NY-NJ-PA" metropolitan area have a country of birth that is not the United States? For this computation, don't include people from this metropolitan area who have a missing country of birth.

0.3092993 ✓ **Answer:** 0.309

EXPLANATION

From table(CPS$MetroArea == "New York-Northern New Jersey-Long Island, NY-NJ-PA", CPS$Country != "United States"), we can see that 1668 of interviewees from this metropolitan area were born outside the United States and 3736 were born in the United States (it turns out an additional 5 have a missing country of origin). Therefore, the proportion is 1668/(1668+3736)=0.309.

You have used 2 of 5 submissions

Problem 4.4 - Integrating Country of Birth Data

(3/3 points)

Which metropolitan area has the largest number (note -- not proportion) of interviewees with a country of birth in India? Hint -- remember to include na.rm=TRUE if you are using tapply() to answer this question.

- ○ Boston-Cambridge-Quincy, MA-NH
- ○ Minneapolis-St Paul-Bloomington, MN-WI
- ● New York-Northern New Jersey-Long Island, NY-NJ-PA ✓
- ○ Washington-Arlington-Alexandria, DC-VA-MD-WV

In Brazil?

- ● Boston-Cambridge-Quincy, MA-NH ✓
- ○ Minneapolis-St Paul-Bloomington, MN-WI
- ○ New York-Northern New Jersey-Long Island, NY-NJ-PA
- ○ Washington-Arlington-Alexandria, DC-VA-MD-WV

In Somalia?

- ○ Boston-Cambridge-Quincy, MA-NH
- ● Minneapolis-St Paul-Bloomington, MN-WI ✓
- ○ New York-Northern New Jersey-Long Island, NY-NJ-PA
- ○ Washington-Arlington-Alexandria, DC-VA-MD-WV

EXPLANATION

To obtain the number of TRUE values in a vector of TRUE/FALSE values, you can use the sum() function. For instance, sum(c(TRUE, FALSE, TRUE, TRUE)) is 3. Therefore, we can obtain counts of people born in a particular country living in a particular metropolitan area with:

```
sort(tapply(CPS$Country == "India", CPS$MetroArea, sum, na.rm=TRUE))
```

```
sort(tapply(CPS$Country == "Brazil", CPS$MetroArea, sum, na.rm=TRUE))
```

```
sort(tapply(CPS$Country == "Somalia", CPS$MetroArea, sum, na.rm=TRUE))
```

We see that New York has the most interviewees born in India (96), Boston has the most born in Brazil (18), and Minneapolis has the most born in Somalia (17).

You have used 1 of 1 submissions

Please remember not to ask for or post complete answers to homework questions in this discussion forum.