Data entry and codebook creation

Overview

Although you will be working with previously collected data, it is important to understand what data looks like as well as how it is coded and entered into a spreadsheet or dataset for analysis. This can help you identify and avoid problems later when reading data into an analysis software program. For example if you mix letters and numbers in the same cell, the variable will be treated as character not numeric.

There are three pieces to this assignment.

1. Entering raw data into a spreadsheet.
2. Creating a codebook
3. Importing the data into your software program of choice (SPC)

Using the PDF copies of medical records for 5 patients seeking treatment in a hospital emergency room you are going to do data entry and create a codebook

Submission Instructions

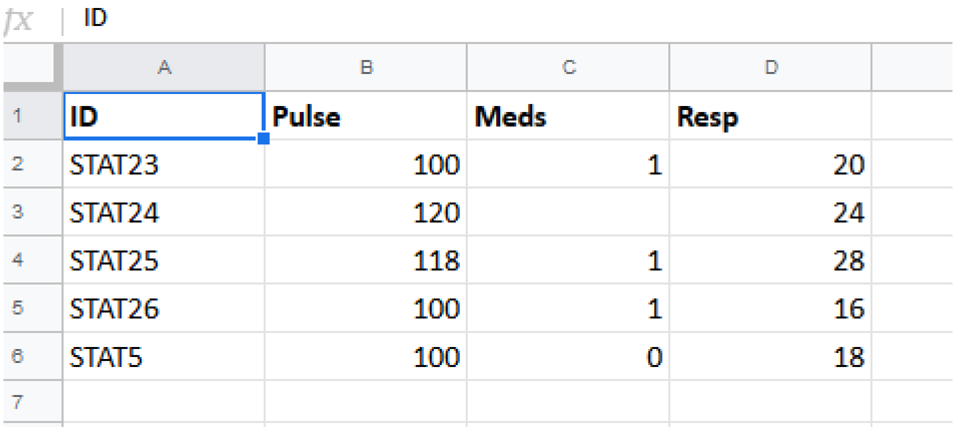
You will enter data and create your codebook directly in Google Sheets.

- Start a new spreadsheet in the **01 Data Entry** folder in our shared Google Drive.
- Name this file `medrecords_userid` where *userid* is your chico state user id.
- Name the following worksheets: `data`, `codebook`, `import`.

Assignment

1. Data Entry

1. Select 4 variables recorded on the medical forms
    - one should be a unique identifier, at least one should be a quantitative variable and at least one should be a categorical variable. (Read PMA5 Ch2 for this information)
2. Select a brief name for each variable - write this in the first row
    - Use good variable naming conventions:
        - short
        - no special characters
        - no spaces
        - doesn't start with a number
3. Determine what range of values is needed for recording each variable
4. Enter the data for each patient, one patient per row.
5. If data is missing for a particular value, leave the cell blank.



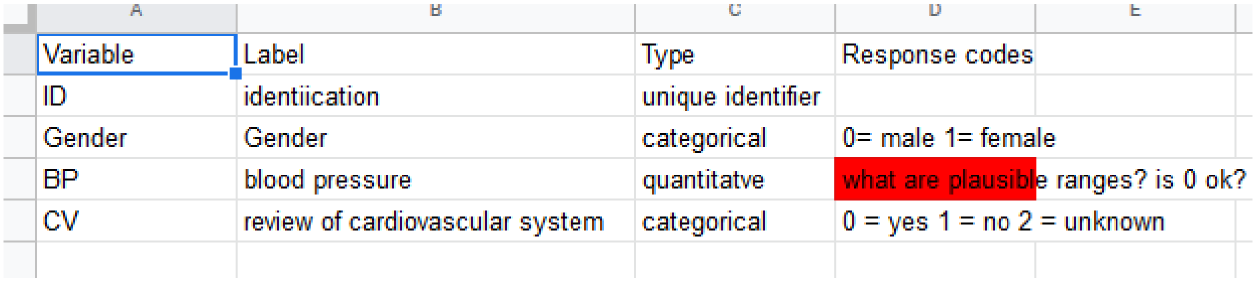
ID	Pulse	Meds	Resp
STAT23	100	1	20
STAT24	120		24
STAT25	118	1	28
STAT26	100	1	16
STAT5	100	0	18

Recall the *tidy data principles* state to put one observation per row, and one variable (characteristic) per column.

2. Codebook Creation

In a separate worksheet list the variable names, labels, data types, and response code or ranges in separate columns (4 columns total).

An example of what this should look like is below. (With the exception of the red error)

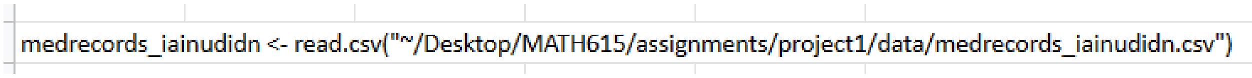


Variable	Label	Type	Response codes
ID	identiication	unique identifier	
Gender	Gender	categorical	0= male 1= female
BP	blood pressure	quantitatve	what are plausible ranges? is 0 ok?
CV	review of cardiovascular system	categorical	0 = yes 1 = no 2 = unknown

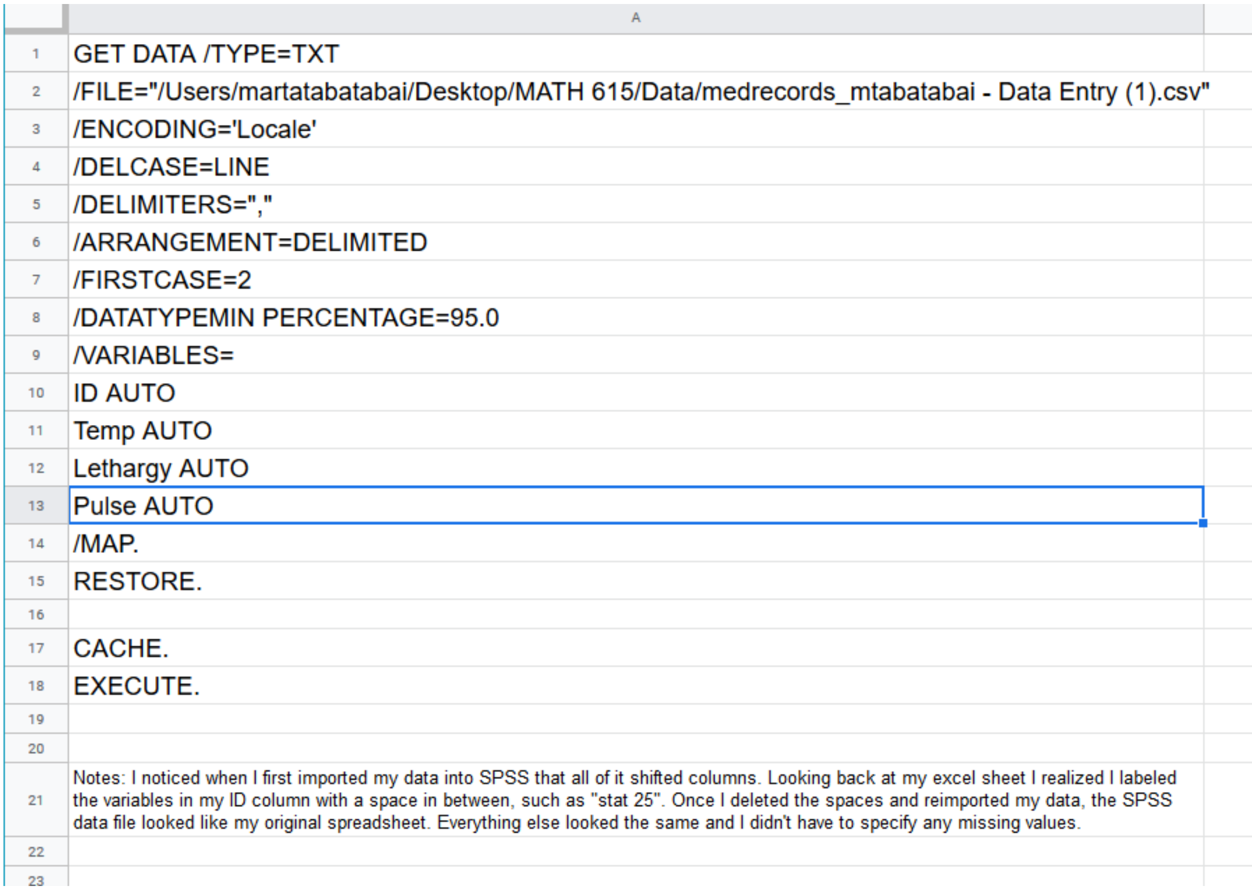
3. Data Import

1. Export your file to your hard drive as a Comma Separated Value (`*.csv`).
   - If it asks you, only save the`data` worksheet.
2. Using your software program of choice, import this data into the program using point and click GUI methods.
   - Code is fine if you already know how. Point and click is also fine for now.
   - Read the collaborative notes for your SPC to learn others have done this.
3. Note and record any problems that you noticed and/or had to fix at the bottom of your `codebook` worksheet.
   - Does your data file look like your spreadsheet?
   - Did you have to specify missing values in any specific way?
4. Show that it worked by taking a screenshot of the code, and the view of the data. Paste both in the `codebook` worksheet.

Examples of import code



```
medrecords_iainudidn <- read.csv("~/Desktop/MATH615/assignments/project1/data/medrecords_iainudidn.csv")
```



```
GET DATA /TYPE=TXT
/FILE="/Users/martatabatabai/Desktop/MATH 615/Data/medrecords_mtabatabai - Data Entry (1).csv"
/ENCODING='Locale'
/DELCASE=LINE
/DELIMITERS=","
/ARRANGEMENT=DELIMITED
/FIRSTCASE=2
/DATATYPEMIN PERCENTAGE=95.0
/VARIABLES=
ID AUTO
Temp AUTO
Lethargy AUTO
Pulse AUTO
/MAP.
RESTORE.

CACHE.
EXECUTE.
```

Notes: I noticed when I first imported my data into SPSS that all of it shifted columns. Looking back at my excel sheet I realized I labeled the variables in my ID column with a space in between, such as "stat 25". Once I deleted the spaces and reimported my data, the SPSS data file looked like my original spreadsheet. Everything else looked the same and I didn't have to specify any missing values.

Grading Rubric

- Data entry
  - Data for 4 distinct variables are entered
  - The three specifed data types are included
  - Used proper variable names
  - Missing data properly treated
- Codebook
  - Each variable in the data is present
  - Ranges of plausible data are defined
  - Levels of categorical variables are defined
- Data import code
  - Code looks correct
  - Path to data leads to math 615 folder
  - Data was read in correctly
    - first row was read in as variable names
    - missing data properly accounted for
  - Screenshot present