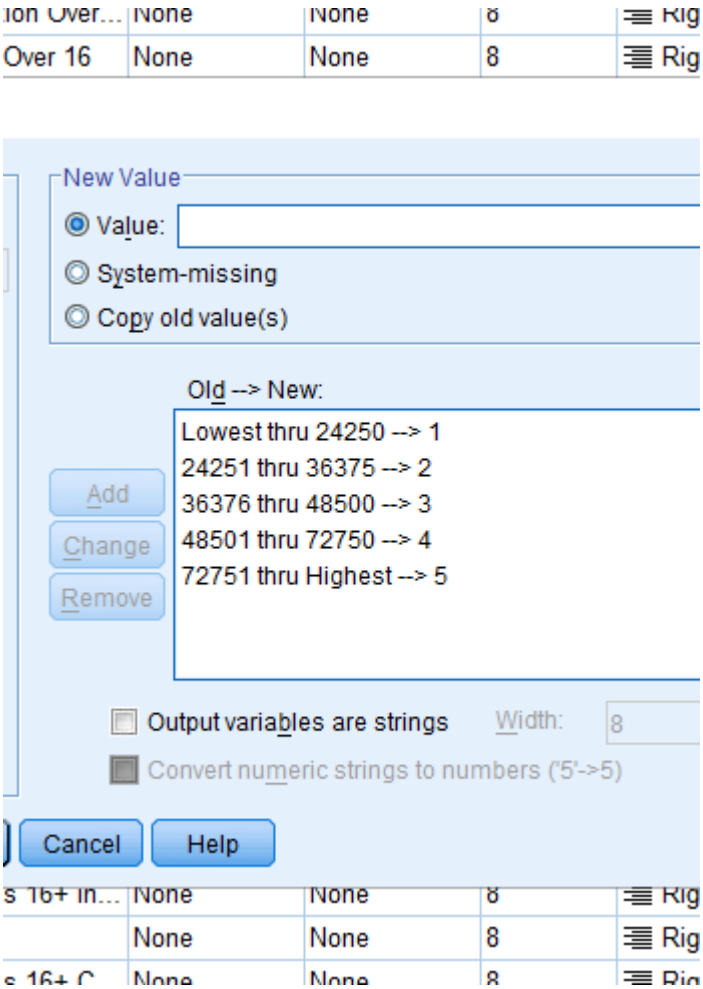


INTRODUCTION TO SPSS

Data management and analysis

ABSTRACT

In this tutorial, you will learn how to import datasets, recode data, run basic data analyses, among many other tasks. The data for this project is from the US Census Bureau.

Ayanda Masilela

CSSCR Workshops (2018)

TO USE THIS DOCUMENT

The document is arranged in sections, which can be overviewed in the Table of Contents. It is subsequently broken down as an outline.

Descriptions and purposes are listed in black

Instructions for execution are highlighted in purple

Questions are highlighted in burnt orange

System dialogue will often be highlighted in blue

YOU WILL NEED

- SPSS 19 – this tutorial has not been tested on later generations of SPSS

TABLE OF CONTENTS

GETTING STARTED

A. Download files
    a. <tinyurl.com/uwayatut>
        i. Navigate to "SPSS Tutorial Files"
        ii. Download all links

B. Open test_scores.sav and DP03CensusData.sav
    a. Data view
        i. Open the test_scores.sav window
        ii. Looks like a spreadsheet/database
        iii. All of the information you would like to analyze
        iv. Look at the top right: How many variables are visible? _____
    b. Variable View
        i. Click the "Variable" bubble at the bottom left of your test_sores.sav window
        ii. First column tells you the names of your columns
        iii. Review *Type* and *Measure*
            1. Type: Generally Numeric or String
                a. Click the […] to observe all types
                b. String – Generally nominal
                c. Numerical – Can be nominal, ordinal, or scalar
            2. Measure
                a. Nominal
                b. Ordinal
                c. Scale – Includes **interval** and **ratio**
            3. Double-click the column head for any column, and it will transport you to Variable View (Variable View tab also on the bottom of the window)
        iv. How does this relate to data types in statistical analysis?
            1. Nominal – Unique identifiers (student ID, driver's license #)
            2. Ordinal – Intentional ordering in a series (Like – Neutral – Dislike)
            3. Interval – The distance between measures has meaning (30-40°F is the same distance as 70-80°F)
            4. Ratio – Always an absolute zero that is meaningful (distance, quantity)
        v. Label and Value
            1. Label: adds more detail about the nature of your column
                a. Column headers are limited in length and format
                b. Labels can help to clarify what is within
                c. Compare labels between the two data rows

i. How are *school_setting* and *classroom* differently represented?_____

2. Value: sometimes it is useful to code values symbolically
    a. 1 = yes, 0 = no, etc.
    b. The value column is where we can store this information

vi. Other Variable Properties
1. Width
    a. Defines the overall length of cell contents
    b. When dealing with decimals, everything to the left *and* right of the decimal
2. Decimal
    a. Number of decimals viewable
    b. In a continuous dataset with many significant figures, there may be three or more decimal places
    c. If this is a coded column (1 = Urban, 2 = Suburban, 3 = Rural), you can reduce the number of decimals to "0" to keep things looking clean
        i. NOTE: often when you recode variables, even as categories, SPSS will automatically assign 2 decimal places
        ii. Adjust accordingly
3. Role
    a. Input. The variable will be used as an input (e.g., predictor, independent variable).
    b. Target. The variable will be used as an output or target (e.g., dependent variable).
    c. Both. The variable will be used as both input and output.
    d. None. The variable has no role assignment.
4. Missing
    a. Datasets with missing variables often code "99" or "9999"
    b. With this column, you can tell SPSS that these values mean "NO DATA"

MANAGING YOUR DATA

The data for section A comes from the US Census Bureau

A. Working with less clean datasets
    a. Open DP03Metadata.xlsx and DP03CensusData.csv
        i. Metadata is an additional file or set of files that provide detailed information about the structure of your dataset

ii. Observe Row 1 of DP03CensusData.csv – there are many strange codes that don't make much sense
iii. Open DP03Metadata.xlsx and look in Column 1 – these codes correspond with descriptors
iv. This is essentially the same data structure as the test_scores.sav file, but in a different format
v. Scroll so that Rows 248 and 252 are visible
    1. What are the codes for each row?

       _____
    2. What, in brief, do they represent?

       _____

       _____

B. Importing Datasets
    a. DP03Censusdata.sav is a successfully imported version of DP03CensusData.xlsx, but we will practice importing, as you will surely encounter this challenge
        i. Click File -> Open Data
        ii. At the bottom of the new window, click in the **Files of type** drop menu
        iii. Click **All files (*.*)**, and then navigate to the folder where DP03CensusData.csv is stored, select, and click **OK**
        iv. Click "Next >" in the new window
        v. This will open the Import Wizard. You can track your progress in the text at the top of the little window
        vi. *How are your variables arranged?*
            1. DP03Censusdata.csv is a comma delimited file
                a. It is essentially a text document, but rows and columns are separated by a comma
                b. "Fixed Width" would apply with a file type like DBF or MS Access
        vii. *Are variable names included at the top of your file?*
            1. Yes! Remember how the codes in the DP03Censusdata.csv Excel spreadsheets were housed in Row 1? Selecting **Yes** tells SPSS to look for those codes in Row 1, and will turn them into column headers (variables)
            2. Click "Next >"
        viii. *The first case of data begins on which line number?*
            1. Make sure that "2" is selected
        ix. *How are your cases represented?*
            1. Select *Each line represents a case*
        x. *How many cases do you want to import?*

1. All of the cases!
    a. The other options may seem odd, but sometimes researchers only desire a limited number or random subset of cases
    b. This frequently happens with super-massive datasets numbering in the hundreds, thousands, or even millions
    c. Click “Next >”

xi. *Which delimiters appear between variables?*
1. Make sure that “Comma” is selected
    a. Since this is a .CSV file, it is comma delimited, even if SPSS does not display the commas in the preview box

xii. Text Wizard Window: *Invalid variable names for this application have been found and changed.*
1. Not to worry, this happens often
2. SPSS can be picky about the characters that can be used as column headers, as well as the lengths of column headers
    a. It dislikes spaces, *, ?, / and many other characters
3. This does not change anything in your source file
4. Click **OK**

xiii. *Specifications for variable(s) selected in the data preview*
1. This window gives you a preview of what your dataset will look like when it’s ready
2. You can edit variable names and set the data format in this window
3. You can also modify the character length
4. To change a variable name, simply scroll to the desired column header and edit the “Variable name:” dialogue
5. Scroll all the way to HC01_VC85 and change the name to “MEDHH_INC”
6. Note how the data format now says **NUMERIC**, whereas in Column 1 it said **STRING**
7. Click “Next >”

xiv. Would you like to save this file format of your text file?
1. You may do so, especially if you will be working with other census datasets
2. You can import this particular arrangement that we made and skip all of the formatting steps the next time you import a census dataset
3. Click **Finish**

4. Your dataset should pop up. Save the file as "MyDP03.sav" if you were successful!

[NOTE TO INSTRUCTOR: If a student's data doesn't show up, don't spend time diagnosing the problem. Instead have them open and work with DP03CensusData.sav. The necessary steps to continue with the lesson have already been performed on this dataset]

5. Go to your Variable view and scroll so that rows 128 and 130 are visible
    a. You can retroactively change variable names
    b. Double-click on HV01_VC86, and change the text to "MNHH_INC"
    c. Click and hold "MEDHH_INC" in the far left, blue column where the list number is located, and drag it somewhere near the top of your list – this may take a moment
    d. Do the same for "MNHH_INC"

C. Recoding Your Data
    a. For ease of use or analytical purposes, it may be worthwhile to recode your data
        i. Remember in the *school_settings* row that Urban, Suburban, and Rural areas were coded as 1, 2, and 3
            1. This is a **1-to-1 categorical recode**
        ii. Recoding can also be done with ranges of numbers to generate categories
            1. This is a **many-to-1 categorical recode**
    b. 1-to-1 categorical recode
        i. Open the MYDP03.sav window (or DP03CensusData.sav) window
        ii. Click **Data View**, and locate the Washington entry for the column GEO_id2: **it's 53**
            1. GEOID is a code used for the Federal Information Processing Standard (FIPS) organizing system
            2. Any geographic data created in the US will be arranged based on these standards
            3. All states are assigned a unique two digit number
            4. Counties are assigned the state two di
        iii. Click **Variable View**
        iv. Scroll to the right, and change the **Measure** for GEO_id2 to "Nominal"
        v. At the top of your window, click **Transform -> Recode into Different Variables**
            1. **Recode into Different vs Same Variables**
                a. Same Variables will overwrite the current variable column

b. This is generally unadvisable

vi. Click GEO_id2
vii. Change the *Output Variable Name* to "Class"
viii. Change the *Label* to "State or County?"
ix. Double-click GEO_id2
x. Click the **Old and New Values** button
xi. This process only recodes to numbers, so we will class STATES as "0" and COUNTIES as "1"
xii. For the *Old Value* on the left side of the window, type "53"
    1. For the *New Value* type 0
xiii. To assign a code to counties, click the *All other values* radio button
    1. For the *New Value* type 1
    2. Click **Continue**
    3. Click **Change** if you haven't already
    4. Click **OK**
xiv. ADJUSTING VALUES
    1. Click **Variable View** and scroll to the bottom intil you see the Class row
    2. In the **Decimals** column, click the corresponding cell, and reduce decimals to "0"
    3. In the **Values** column, click the cell corresponding to Class and the […] button
xv. Enter the following labels for each value and click "Add"
    1. 0 = State
    2. 1 = County

c. Many-to-1 categorical recode
    i. Open the MYDP03.sav window (or DP03CensusData.sav) window
    ii. We will be analyzing qualifications for health care coverage subsidies based on poverty levels
    iii. Based on income data from 2015 and intended for 2016 data analyses
    iv. Open the "2016plchart.pdf" file to observe the full extent of the analysis
    v. We will work with the assumed household size of "4", and create classes for 100%, 200%, 300%, and 400%
        1. Thus, that will set our ranges to the following

100%	$24,250
200%	$36,375
300%	$48,500
400%	$72,750

vi. At the top of your window, click **Transform -> Recode into Different Variables**
    1. Scroll to MEDHH_INC should be near the top of the list – Click it and press the arrow
    2. Set the Output Variable Name to "PercPovLev"
        a. SPSS won't accept % as a column header character
    3. Set the Label to "% Poverty Level"

vii. Click the **Old and New Values** button

viii. In the Old Value half (left) click the "Range, LOWEST through value:" radio button
    1. Type $24,250

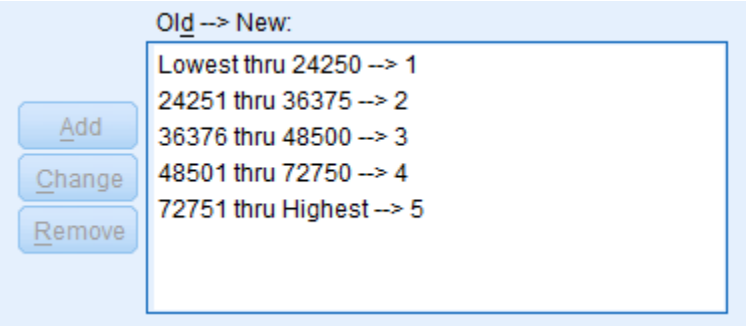
ix. In the New Value Half (right) click the **Value** radio button, and enter "1"
    1. Click the **Add** button
    2. You should now see syntax that says "Lowest thru 24250 --> 1"
        a. This expresses a range of 0 - 24250

x. To the left, click the **Range** radio button
    1. Type 24251 for the top entry, and 36375 in the through: entry
    2. Change the **Value** on the right to "2", and click **Add**
    3. Why is it important to set the range for the next tier up 1 greater than the upper extent of the previous range?

        _____

xi. Repeat ***steps x. 1-3*** for the 200% - 300% and 300% - 400% range using the appropriate numbers

xii. The last step is similar to the first, but with "Range, value through HIGHEST:"
    1. When your are finished, it should look like this…



    2. Click **Continue**
    3. Click **Change**
    4. Click **OK**

xiii. A new window has popped up!
    1. This is your output window. It tracks all of the calculations, recodes, etc .that you carry out

2. You can save it any time you generate new outputs, and would prefer to not repeat those steps
3. Scroll to the bottom, and it will detail your latest process of recoding data

xiv. Go back to the **Data View** of your DP03 census spreadsheet, and scroll all the way to the right

1. Your new recoded values should now be present, but with decimals
2. Click the **Variabel View**
    a. In the **Decimals** column, change the "2" to a "0"
    b. Just like you did the Median and Mean household income variables, slide "PercPovLev" towards the top of the variable list so that it is easily accessible
3. ADJUSTING VALUES
    a. In the **Values** column, click the cell corresponding to PrecPovLev and the […] button
    b. Enter the following labels for each value and click "Add":

Value	Label
1	< = 100% Poverty Level
2	< = 200% Poverty Level
3	< = 300% Poverty Level
4	< = 400% Poverty Level
5	> 400% Poverty Level

    c. If you make a mistake, simply highlight the value you would like to change, type the new text, and click the **Change** button
    d. Click **OK**
    e. Note how the **Measure** has changed from *Scale* to *Nominal.* Why did this happen?

_____

_____

D. Calculating New Variables

SPSS also allows you to calculate new fields based on existing fields. We'll continue working with the census data for this section

a. Calculating population density

i. Population density is simply the measure of people per square measure, in this case miles
ii. If D = population density, P = population, and A = Area

$$D = P/A$$

iii. Click **Transform -> Compute New Variable**
iv. Set the *Target Variable* to Pop_Dens
v. You can also edit the desired label at this time to "Population Density"
vi. Keep the *Type* as *Numeric*
vii. To fill the **Numeric Expression** click "Population" [TOTAL_POP]
viii. Next, click the "/" for division
ix. Finally, click "Area (miles)" [SQ-MI]
x. Your syntax should read "TOTAL_POP / SQ_MI"
xi. Click **OK**
xii. Scroll all the way to the right in your data view, and you should now see population densities
xiii. In your **Variable View** slide the new population density variable so it is next to the "TOTAL_POP" variable
xiv. Head back to the **Data View**, and scroll all the way to the left
xv. Right-click the column header for "Pop_Dens", and click **Sort Descending**
    1. Which county has the highest population density? _____
xvi. Right-click the column header for "Pop_Dens", and click **Sort Ascending**
    1. Which county has the lowest population density? _____
    2. How does the population density of the state of Washington compare to the county with the highest population density? _____

ANALYZING DATA

That's why we're here! In this section, we will cover minor preparation steps, creating histograms, creating charts (plots, lines, bars, it's up to you as the instructor!), and some basic tests. Feel free to repeat these steps with the test score data.

1. Since we are looking to analyze counties only, we'll need to remove the entry for Washington state
    a. At the top of the window, click **Data -> Select Cases**
        i. Select the radio button for "If condition is satisfied", then click the **If** button

ii. Scroll to the bottom of the list and locate "State or County?", and double-click it
iii. Click the "=" operator
iv. Type "1"
v. Click **Continue**, then **OK**
vi. Head to your output window to see if the process was successful
vii. Next, in your **Data/Variable View** window, look to the bottom right corner
    1. It should say "Filter On"
viii. Go to your **Data View**, and note how the entry for Washington State now has a black slash through its leftmost blue column
    1. It will not be included in further analyses
    2. It hasn't been permanently removed! Rather, it has been filtered out

2. Descriptive Statistics
    a. These will describe the basics of our dataset, such as mean and standard deviation, etc.
        i. These traits can help you identify which tests (T-Tests, etc) are appropriate to run
    b. At the top of the window, click **Analyze -> Descriptive Statistics -> Descriptives**
        i. Double-click *Population* and *Median Household Income*
        ii. Click the **Options** button
            1. Add *Range*
            2. Click **Continue**, then **OK**
        iii. Now go to your output window – there should now be a table with descriptive statistics
            1. What is the standard deviation of median household income? _____
            2. What is the range of populations? _____
    c. Next, we'll use the **Frequencies** option
        i. This tool allows us more flexibility to look at categorical data, whereas the Descriptives function was limited to continuous data
            1. It also offers us a few more descriptive statistical options, such as measures of central tendency (median, mode in addition to mean)
        ii. **Analyze -> Descriptive Statistics -> Frequencies**
            1. Scroll to the bottom of the list and locate "% Poverty Level"
            2. To the right, click **Charts**
                a. Click the radio button for **Pie**
                b. Click the radio button for **Percentagies**
                    i. Click **Continue,** then **OK**

3. Head to your **Output** window, and you will now have a table featuring the frequencies of occurrences of each class of poverty level based on median household income
4. You should also see a pie chart, which provides a visual for the proportion of income brackets
5. You will find in SPSS that there are multiple ways to generate charts, histograms, etc. This is but one way.
6. How many counties have a median household income that is greater than 400% of the poverty level? _____

d. Click **Analyze -> Descriptive Statistics -> Frequencies**
   i. Double-click *Mean Household Income* and *Median Household Income*
      1. Click the **Statistics** button
      2. Add Quantiles, Mean, Median, Standard Deviation, and Range
         a. Click **Continue**
         b. Click **OK**
      3. In the **Output** window, you will now have a table featuring the statistics for median and mean household income
         a. For this particular dataset, frequencies are not as useful, but in a dataset such as test scores, frequencies of a particular grade may be more useful
         b. Nonetheless, this tool gives us a few more statistical description options than the basic descriptives tool

3. Histograms

Histograms can tell us about the overall distribution of our data. We'll head back to the test scores data, since it has a greater sample size and demonstrates more variation within its dataset.

a. Click on **Graphs -> Legacy Dialogues - > Histogram**
   i. Add Number of Students in the Classroom to the *Variable*
      1. Click **OK**
      2. What is the most frequently occurring classroom size? _____
      3. This gives us an overall view of classroom sizes across all settings, but what if we want to compare these results in different school settings (urban, suburban, rural)?
   ii. Go back to the **Histogram** dialogue, and add Number of Students in the Classroom to the *Variable*
   iii. Add School Setting to *Rows*
      1. Click **OK**

2. In your output, you will now see a histogram that lists the frequencies of each classroom size, and furthermore divided by the classroom settings or Urban, Suburban, and Rural
3. Which setting appears to have the largest classroom sizes generally? _____
4. Which setting appears to have the smallest classroom sizes generally?

   _____

4. Scatterplots

Scatterplots are a good way to see correlations and trends

a. **Graphs -> Legacy Dialogues -> Scatter/Dot**
   i. Select *Simple Scatter*
      1. Click **Define**
   ii. Select your dependent variable (Y Axis): Pre-test
   iii. Select your independent variable (X Axis): Number of students in classroom
      1. Click **OK**
      2. What can you infer from these results? _____

         _____

5. Splitting Datasets

"Splitting" enables us to carry out the same tasks, but subdivided based on a category. We will continue working with school settings

[NOTE TO INSTRUCTOR: Some may ask why when we created histograms we were able to input a classification variable and SPSS knew exactly how to generate that particular graph in three iterations.

- In SPSS, there are often a couple of ways to do the right thing
- You can indeed split the dataset as we will in the upcoming steps
- Splitting the dataset enables us to double down on subdivisions

a. **Data -> Split File**
   i. Select the *Compare Groups* radio button
   ii. Select School Setting
      1. Click **OK**
      2. Note that in the bottom, right corner, the window says "Split by School Setting"
   iii. Now repeat the steps from the Scatter Plots section

iv. Briefly describe the outcomes of classroom size in relation to test score for each classroom setting:

_____

_____

_____

b. Box (and Whisker) Plot
   i. Click the **Variable View** tab, and look at the variables School Setting and School Type
   ii. Click the Values […] button and observe the data represented
      1. What does 0 represent? What does 1 represent?

         _____
   iii. We've already split the dataset. We can take advantage of our graph builder's interface tools to further subdivide our analysis
      1. In this case, we will build a box plot that subdivides Public and Non-Public schools WITHIN the categories Urban, Suburban, and Rural

c. **Graphs -> Legacy Dialogues -> Boxplot**
   i. Keep the default choices selected, and click the **Define** button
      1. Add "Number of students in the classroom" to *Vriable*
      2. Add "School type" to *Category Axis*, then click **OK**
      3. You should now see a table 3 boxplots, one for each School setting
         a. Which school type tends to have the largest classroom sizes? _____
         b. Which combination of school type and school setting has the largest classroom sizes? _____