Exploring Real Data

1. The General Social Survey (GSS) data:

a. Background:
GSS is conducted by the National Opinion Research Centre (NORC), a Social Science Research Centre at University of Chicago, since 1972.

b. Purpose:
Monitor social change within the U.S. and compare U.S. with other countries. (The Canadian version of the GSS is available through Statistics Canada).

c. Topics:
Sociological and attitudinal surveys regarding education, employment, political opinion, and crime and violence.

d. Data:
Open access from University of Berkley: http://sda.berkeley.edu/

e. Data Analysis:
Survey Documentation and Analysis (SDA) 4.0 (Newest version) Archive: http://sda.berkeley.edu/archive.htm

i. Free web-based data analysis
http://sda.berkeley.edu/sdaweb/analysis/?dataset=gss14
ii. Export data into a CSV file and then import the data into a preferred statistical software for data analysis: PSPP (Free Statistical Software)

2. PSPP: https://www.gnu.org/software/pspp/

Download PSPP (e.g., PC or Mac OS): https://www.gnu.org/software/pspp/get.html

Part 1. Exploring the General Social Survey (GSS) data with SDA

In 2014 a survey on attitudes towards science and technology asked participants to indicate their opinion regarding the statement: "Science and technology are making our lives, healthier, easier, and more comfortable" as "strongly agree", "agree", "disagree", or "strongly disagree".

Objective: Univariate Data Analysis

Type of Data: An Ordinal Categorical Data
Levels: "strongly agree", "agree", "disagree", or "strongly disagree"

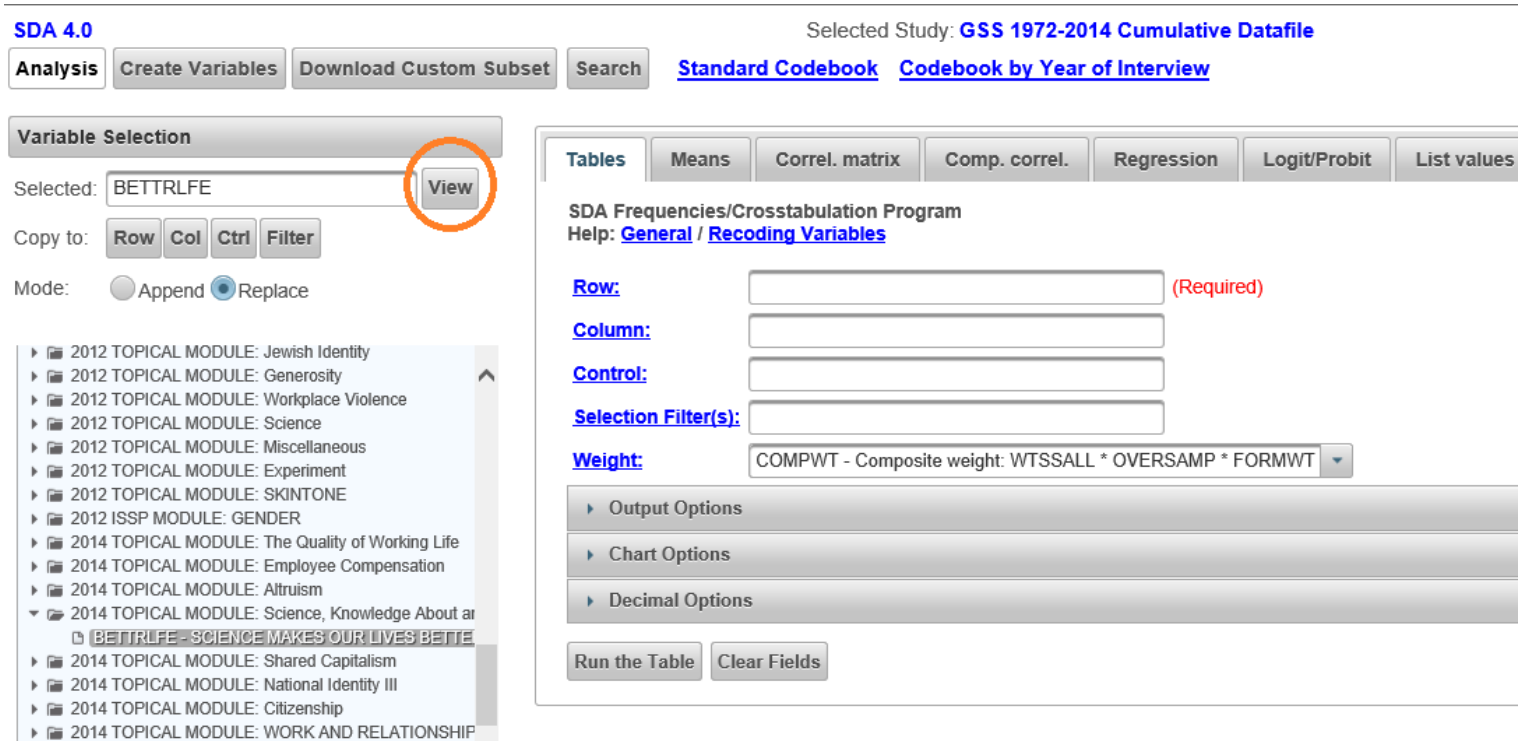
Steps:

1. Visit: http://sda.berkeley.edu/sdaweb/analysis/?dataset=gss14

2. Scroll down the variable tree, and click on the folder:
   2014 TOPICAL MODULE: Science, Knowledge About and Attitudes Towards.

3. Select the sub-folder, the item on this survey:
   BETTRLFE - SCIENCE MAKES OUR LIVES BETTER
   Note: BETTRLFE is the variable name.

4. Click view: Information regarding this variable, BETTRLFE.



5. Quick overview, frequency distribution, of the variable, BETTRLFE.

BETTRLFE **SCIENCE MAKES OUR LIVES BETTER**

Description of the Variable

I'm going to read to you some statements like those you might find in a newspaper or magazine article. For each statement, please tell me if you strongly agree, agree, disagree, or strongly disagree. C. Science and technology are making our lives, healthier, easier, and more comfortable.

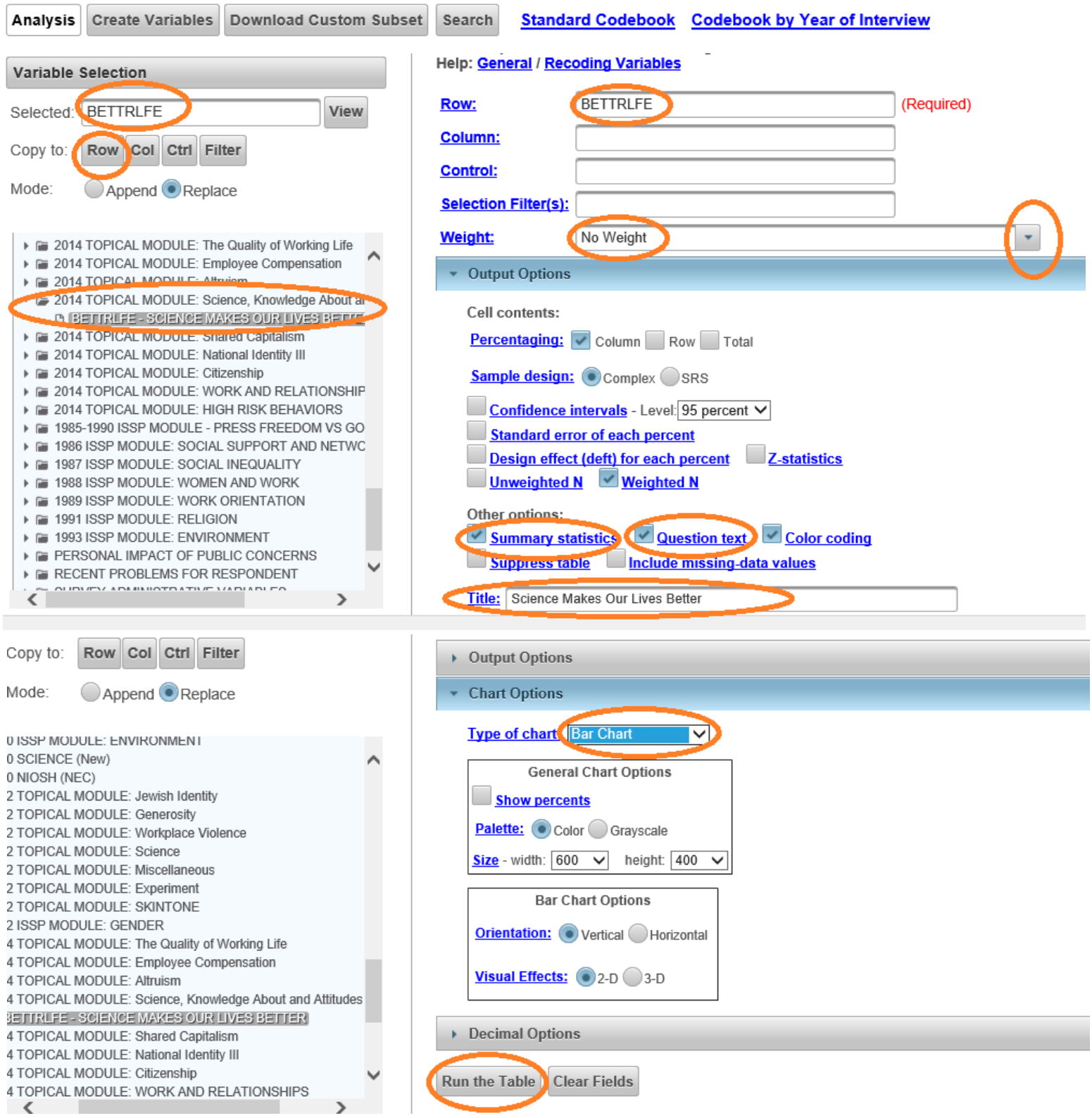
Percent	N	Value	Label
18.5	225	**1**	Strongly agree
63.4	772	**2**	Agree
16.5	201	**3**	Disagree
1.6	19	**4**	Strongly disagree
	58,360	**0**	IAP
	20	**8**	Don't know
	2	**9**	No answer
100.0	**59,599**		**Total**

Properties

Data type:	numeric
Missing-data codes:	0,8,9
Mean:	2.01
Std Dev:	.64
Record/column:	1/4772

Selected Study: General Social Survey Cumulative Datafile 1972-2014

6. Making some changes to SDA options and then running the table:



7. After we run the table, for valid cases (n = 1217), this is (part of) the SDA output:

Science Makes Our Lives Better

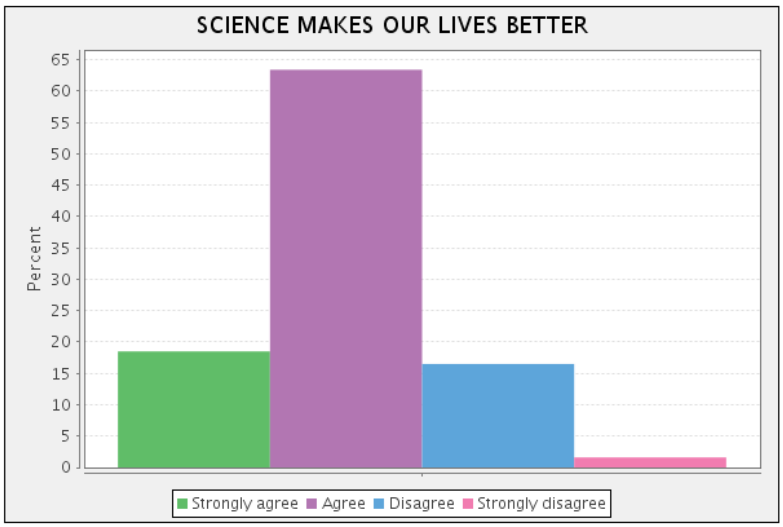
SDA 4.0: Tables

General Social Survey Cumulative Datafile 1972-2014

Aug 20, 2016 (Sat 01:36 PM PDT)

Variables					
Role	**Name**	**Label**	**Range**	**MD**	**Dataset**
Row	**BETTRLFE**	SCIENCE MAKES OUR LIVES BETTER	1-4	0,8,9	1

Frequency Distribution		
Cells contain: -**Column percent** -N of cases		**Distribution**
BETTRLFE	1: Strongly agree	**18.5** 225
	2: Agree	**63.4** 772
	3: Disagree	**16.5** 201
	4: Strongly disagree	**1.6** 19
	COL TOTAL	***100.0*** *1,217*



- We can prompt students with questions such as:
  - What was the most response to this GSS statement? Answer: Agree
- We can develop students' numeracy skills:
  - Leave the total count (e.g., 1217) and the percentages for each category (e.g., 18.5%, 63.4%, 16.5%, 1.6%).
  - Hide the frequencies (e.g., 225, 772, 201, 19) in the frequency distribution.
  - Ask students, "how many participants agreed to this GSS question?"

    Take the percentage in the "agree" category and multiply by the total count:

    63.4% (or 0.634) x 1217 ≅ 772

8. Treat an ordinal categorical variable as a quantitative variable:

Objective: Obtain summary statistics such as the Mean, Median, Mode, and Standard Deviation.

Science Makes Our Lives Better

SDA 4.0: Tables

General Social Survey Cumulative Datafile 1972-2014

Aug 20, 2016 (Sat 01:36 PM PDT)

Variables					
Role	**Name**	**Label**	**Range**	**MD**	**Dataset**
Row	**BETTRLFE**	SCIENCE MAKES OUR LIVES BETTER	1-4	0,8,9	1

Frequency Distribution		
Cells contain: -**Column percent** -N of cases		**Distribution**
BETTRLFE	1: Strongly agree	**18.5** 225
	2: Agree	**63.4** 772
	3: Disagree	**16.5** 201
	4: Strongly disagree	**1.6** 19
	COL TOTAL	***100.0*** *1,217*

- We can prompt students with questions such as:
  - What is the mean response to this GSS statement?
  - What is the median response to this GSS statement?
  - What is the mode response to this GSS statement?
  - What is the shape of the distribution to this GSS statement?

  - Find the mean in two different ways (some rounding are involved):

1. Mean = $\frac{sum(level\ x\ frequncy\ of\ that\ level)}{total\ count} = \frac{1(225)+2(772)+3(201)+4(19)}{1217} = \frac{2448}{1217} = 2.01$

2. (Weighted) Mean = sum (level x proportion for each level)
   = (0.185 x 1) + (2 x 0.634) + (3 x 0.165) + (4 x 0.016) = 2.03

- Find the median, 50th percentile:

  Add percentages until we reach 50%. In this example:

  18.5% (less than 50%; so add the next %) + 63.40% > 50% (median is here).

  Median is 2 (refers to "agree" category).

- Mean is very close to the median (slightly bigger):

  (Mean = 2.01) ≅ (Median = 2)
  ∴ We have an approximately symmetrical distribution.

- Mode refers to the most frequent category. In this example, it is level 2 (refers to "agree" category).

  Mean, Median, and Mode are about the same.
  ∴ We have an approximately, bell-shaped symmetrical distribution.

Below is what SDA provides as summary statistics:

Summary Statistics					
Mean =	2.01	Std Dev =	.64	Coef var =	.32
Median =	2.00	Variance =	.41	Min =	1.00
Mode =	2.00	Skewness =	.34	Max =	4.00
Sum =	2,448.00	Kurtosis =	.50	Range =	3.00

9. SDA provides information regarding how the data were collected:

Sample design: stratified cluster sample

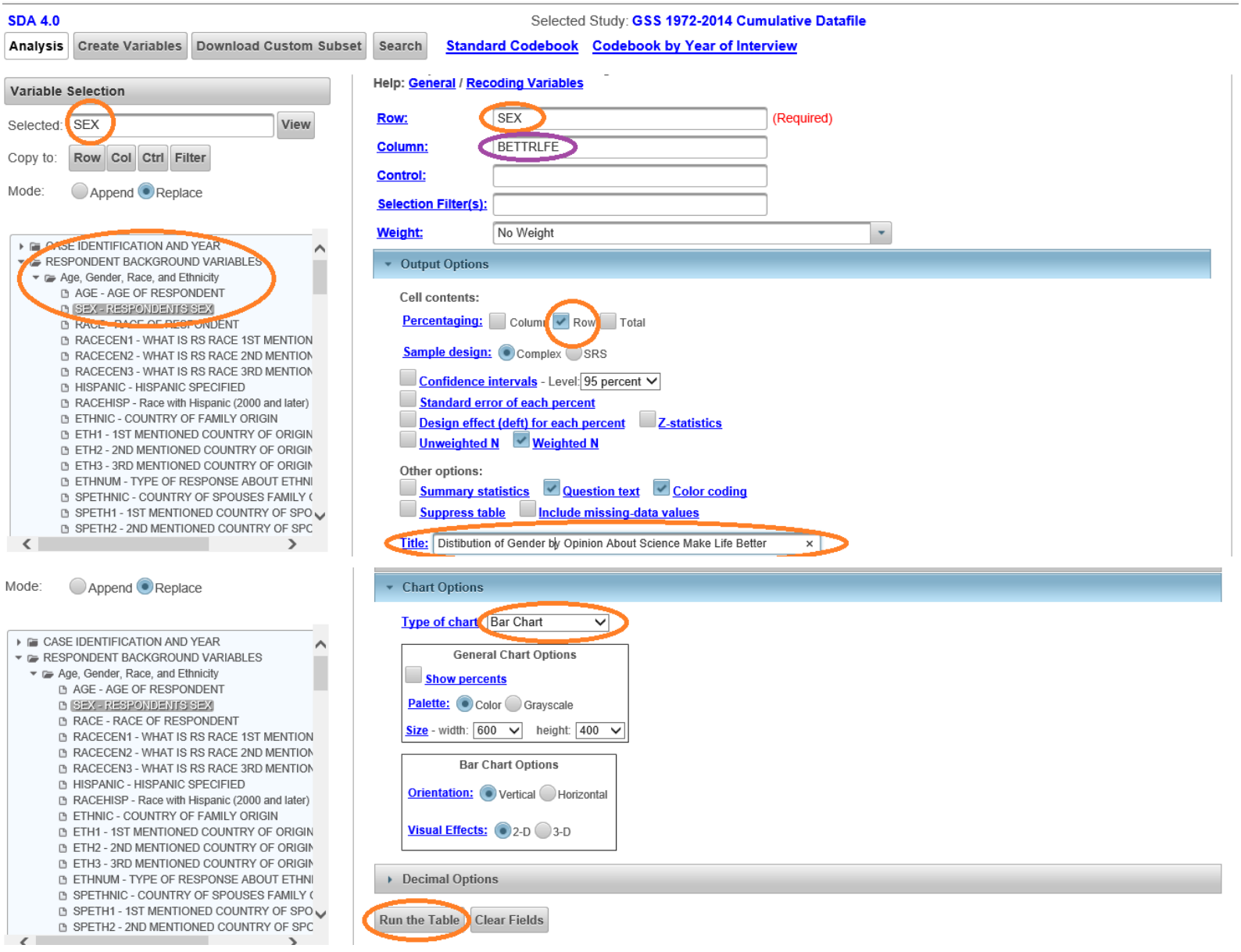
Stratum variable = SAMPLERC
Cluster variable = RSAMPCODE

Consider, adding a layer to the data analysis. In this example, we will add the variable sex of respondents into the data analysis.

Objective: Bivariate Data Analysis (Association between Two Categorical Variables)
Variables and their Type:

- **Opinion About Science Make Life Better:**
  An Ordinal Categorical Variable
  Levels: "strongly agree", "agree", "disagree", or "strongly disagree"
- **Sex of respondents:**
  A Nominal Categorical
  Levels: "Male", or "Female"

10. Select "SEX" as the "Row" Variable; And, "BETTRLFE" as the "Column" variable. Follow the screen shots below to makes the changes to SDA and then click on "Run the Table".



11. The contingency table for the distribution of:

"Opinion About Science Makes our Lives Better by Sex of Respondents".

Distibution of Gender by Opinion About Science Make Life Better

SDA 4.0: Tables

General Social Survey Cumulative Datafile 1972-2014

Aug 21, 2016 (Sun 03:35 PM PDT)

Variables

Role	Name	Label	Range	MD	Dataset
Row	**SEX**	RESPONDENTS SEX	1-2	0	1
Column	**BETTRLFE**	SCIENCE MAKES OUR LIVES BETTER	1-4	0,8,9	1

Frequency Distribution

Cells contain: -**Row percent** -N of cases		BETTRLFE				
		1 Strongly agree	2 Agree	3 Disagree	4 Strongly disagree	***ROW TOTAL***
SEX	1: MALE	**22.1** 120	**61.0** 331	**15.1** 82	**1.8** 10	***100.0*** *543*
	2: FEMALE	**15.6** 105	**65.4** 441	**17.7** 119	**1.3** 9	***100.0*** *674*
	COL TOTAL	***18.5*** *225*	***63.4*** *772*	***16.5*** *201*	***1.6*** *19*	***100.0*** *1,217*

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	Z
N in each cell:	Smaller than expected			Larger than expected			

We can develop students' statistical literacy (read and interpret the table):

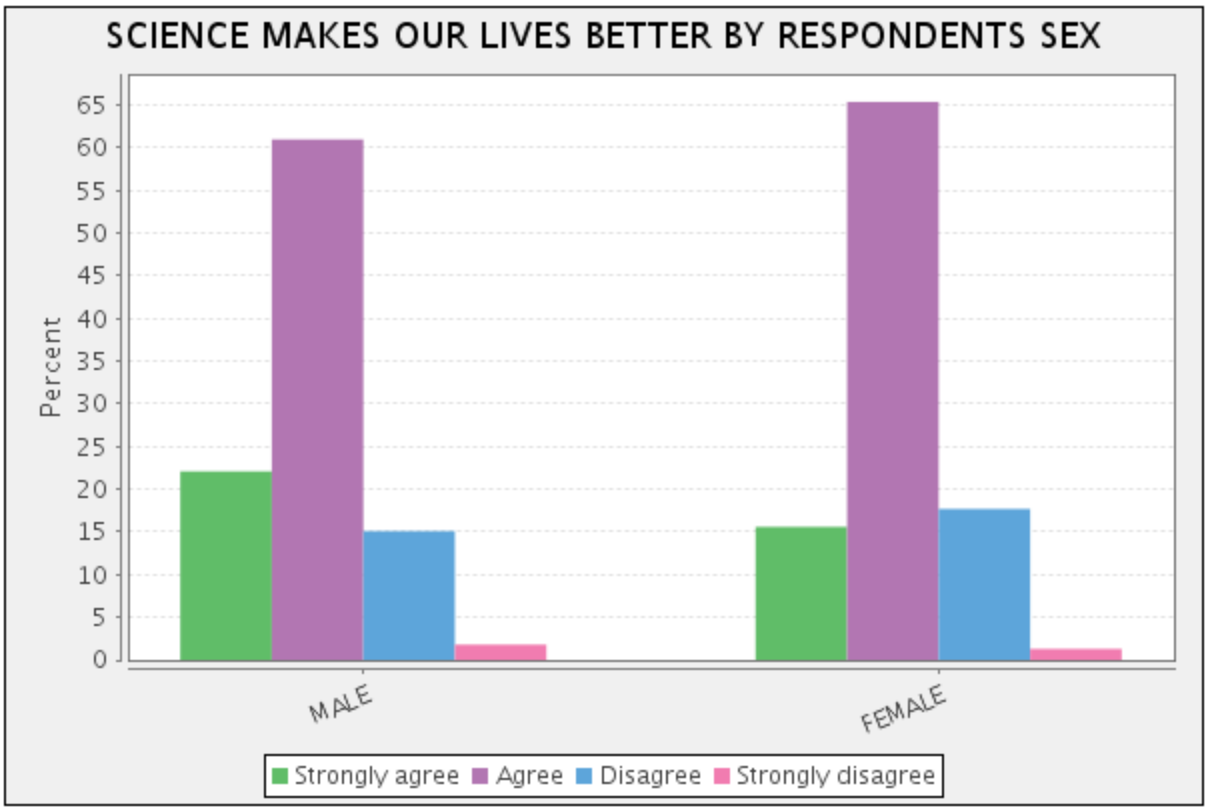
- Marginal Distributions:
  - E.g., percentage of male respondents: 543/1217 x 100 = 44.61%

- Conditional Distributions:
  - Percentage of **male** participants who agreed: 331/543 x 100 = 60.96%;
  - Percentage of **female** participants who agreed: 441/674 x 100 = 65.43%
  - **Compare the percentages for males and females who agreed: Female participants were more likely to agree than the males.**

- Joint Distributions:
  - E.g., percentage of the participants who were male and agreed to the GSS statement: 331/1217 x 100 = 27.20%

12. The nested bar chart for the distribution of:

"Opinion About Science Makes our Lives Better by Sex of Respondents".



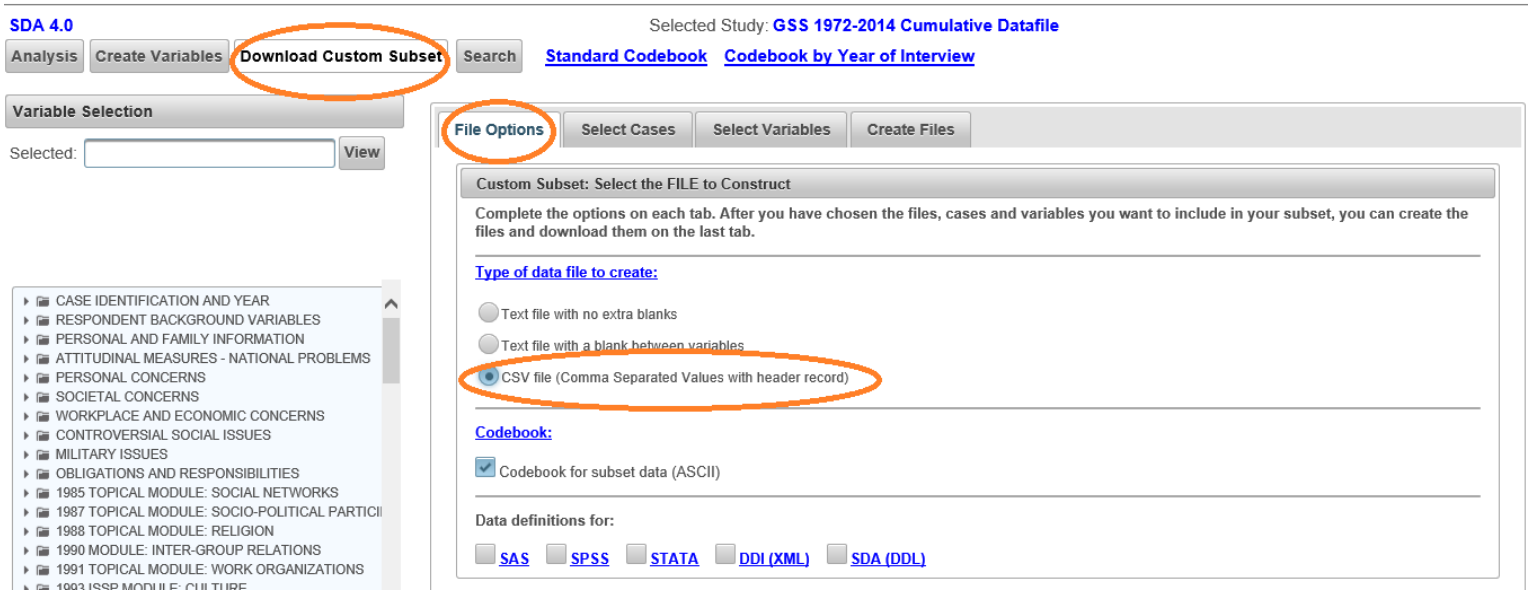
We can ask students to describe any obvious pattern they see in the graph.

Graph interpretation:

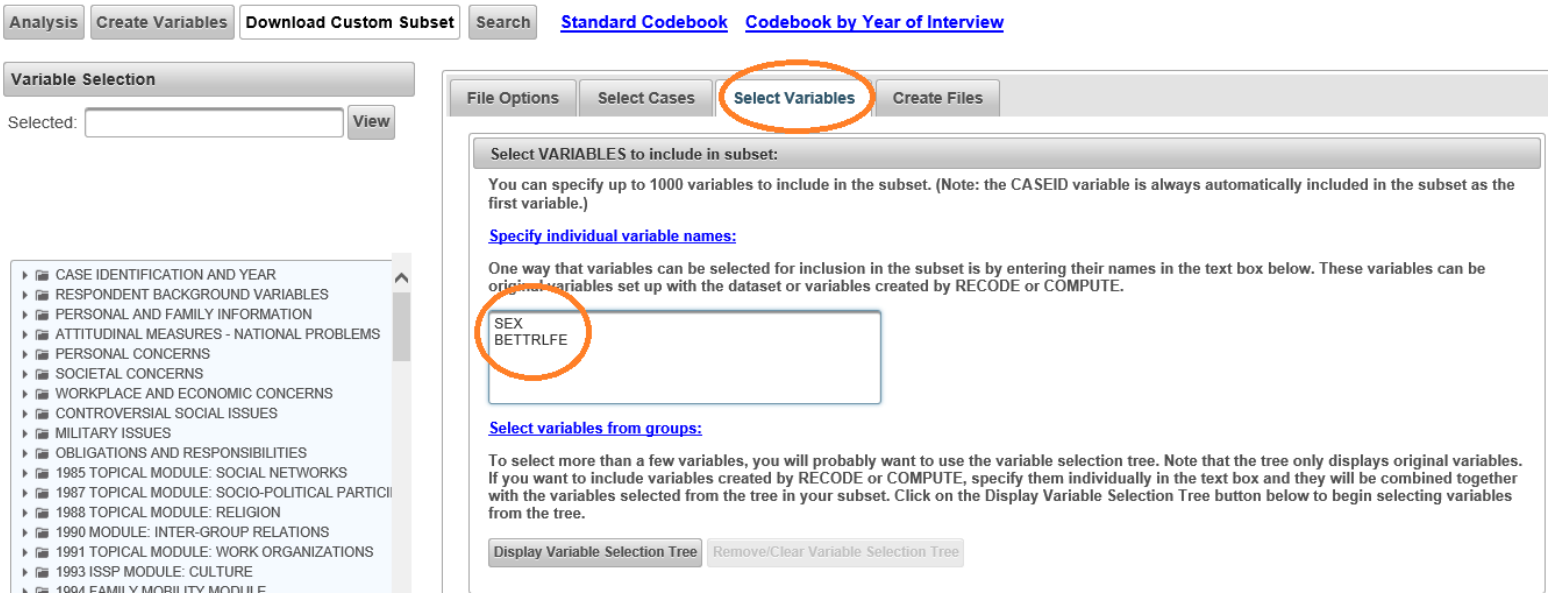
Both male and female participants mostly agreed to the GSS statement, "Science and technology make our lives better". More of the female respondents agreed to the GSS statement than the males.

Export Data into a CSV file

1. In SDA, from the tool bar menu, go to “Download Custom Subset”.
2. In the “File Options” tab, select “CSV file from the “Type of data file to create”.



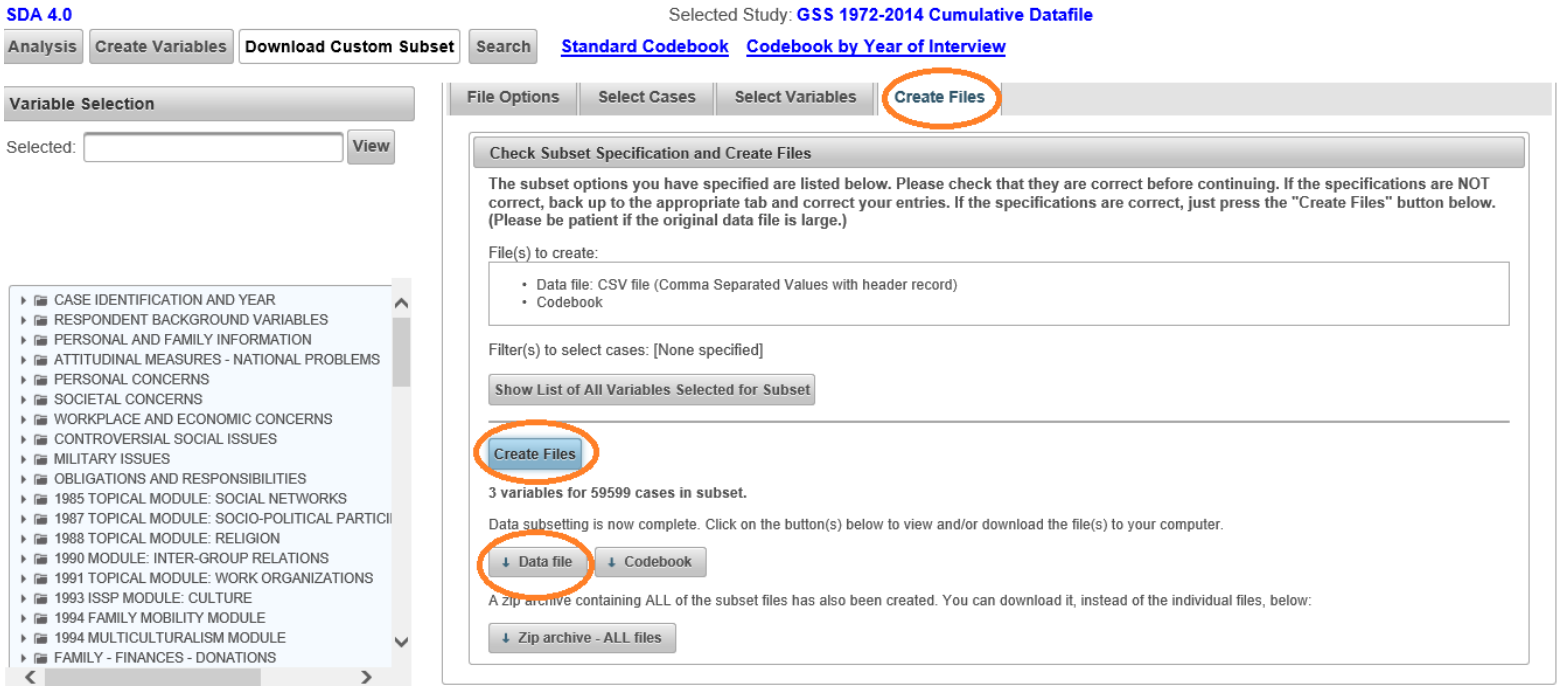
3. In “Select Variables” tab, under “Specify individual variables names”, type in the box the names of the two variables: SEX, BETTRLFE.



4. In “Create Files” tab, click on the icon “Create Files”.
5. Next, click on the icon “Data files’ and save the data in your “My Document” folder.
6. Click on the icon “Codebook”. Note the numeric references for the levels of the variables in the data file:

SEX: “1 = Male”, “2 = Female”
BETTRLFE: “1 = Strongly Agree”, “2 = Agree”, “3 = Disagree”, “4 = Strongly Disagree”

The above levels for each variable are the only levels we want to include into our data analysis. We will omit the missing values coded as, e.g., 8, 9, from our data analysis (later on in PSPP).



Part 2. Import Data in PSPP

❖ Open PSPP.

1. Select Files to Import:

   ▪ In menu bar, go to File > Import Data > My Document
   - Select the saved data file
   - Click Next (Bottom of the page)

2. Select the Lines to Import:

   ▪ Click Next

3. Select the First Line:

   ▪ Select Line "1"
   ▪ Check off the box: Line Above Selected Line Contains Variable Names
   ▪ Click Next

4. Choose Separators:

   ▪ Click Next

5. Adjust Variable Formats:

   ▪ Click Apply.

❖ Go to Variable View (Bottom of the page).

   ▪ For variable SEX, make value labels:
     - Value: 1, Value Label: Male; Click on Add
     - Value: 2, Value Label: Female; Click on Add
     - Click Ok.

   ▪ For variable BETTRLFE, make value labels:
     - Value: 1, Value Label: Strongly Agree; Click on Add
     - Value: 2, Value Label: Agree; Click on Add
     - Value: 3, Value Label: Disagree; Click on Add
     - Value: 4, Value Label: Strongly Disagree; Click on Add
     - Click Ok.

- In menu bar, go to File > New > Syntax
  - Copy the syntax below and paste into the syntax editor:

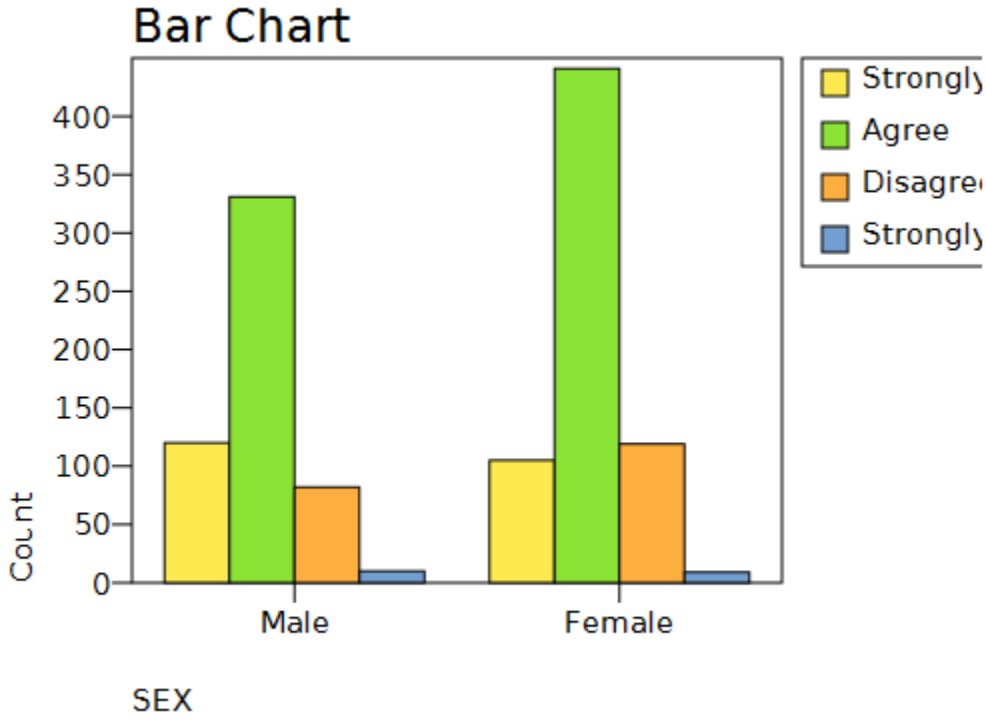
```
SELECT IF (SEX = 1 OR SEX = 2).
EXECUTE.

SELECT IF (BETTRLFE = 1 OR BETTRLFE = 2 OR BETTRLFE = 3 OR
BETTRLFE = 4).
EXECUTE.
```

  - In the syntax menu bar, click on RUN > ALL.

- In menu bar, go to Bar Chart.
  - In the Bar Chart box that appears:
    - Variable: BETTRLFE
    - Variable Axis: SEX
    - Variable Cluster: BETTRLFE
    - Click OK.

- PSPP output:



❖ In menu bar, go to Analyze > Descriptive Statistics > Crosstabs

- In the Crosstabs box that appears:
  - Rows: Select variable “SEX”
  - Columns: Select variable “BETTRLFE”
  - Click OK.

❖ PSPP output:

▸ CROSSTABS

CROSSTABS

```
CROSSTABS
      /TABLES= SEX       BY    BETTRLFE
      /FORMAT=AVALUE TABLES PIVOT
      /STATISTICS=CHISQ
      /CELLS=COUNT ROW COLUMN TOTAL.
```

Summary.

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
SEX * BETTRLFE	1217	100.0%	0	0.0%	1217	100.0%

SEX * BETTRLFE [count, row %, column %, total %].

SEX	*BETTRLFE*				
	Strongly Agree	Agree	Disagree	Strongly Disagree	Total
Male	120.00	331.00	82.00	10.00	543.00
	22.10%	60.96%	15.10%	1.84%	100.00%
	53.33%	42.88%	40.80%	52.63%	44.62%
	9.86%	27.20%	6.74%	.82%	44.62%
Female	105.00	441.00	119.00	9.00	674.00
	15.58%	65.43%	17.66%	1.34%	100.00%
	46.67%	57.12%	59.20%	47.37%	55.38%
	8.63%	36.24%	9.78%	.74%	55.38%
Total	225.00	772.00	201.00	19.00	1217.00
	18.49%	63.43%	16.52%	1.56%	100.00%
	100.00%	100.00%	100.00%	100.00%	100.00%
	18.49%	63.43%	16.52%	1.56%	100.00%

Chi-square tests.

Statistic	*Value*	*df*	*Asymp. Sig. (2-tailed)*
Pearson Chi-Square	9.55	3	.023
Likelihood Ratio	9.50	3	.023
Linear-by-Linear Association	4.74	1	.030
N of Valid Cases	1217		

❖ Options to save (Export) PSPP outputs:

- In the output window, go to File > Export > "Give a name to your output"
- At the bottom of the box that appears, from the drop down menu, select the format that you want to save your PSPP output: e.g., HTML(*.html)
- Click "Save"

Here is the PSPP output in html format:

SEX * BETTRLFE [count, row %, column %, total %].

	BETTRLFE				
SEX	Strongly Agree	Agree	Disagree	Strongly Disagree	Total
Male	120.00	331.00	82.00	10.00	543.00
	22.10%	60.96%	15.10%	1.84%	100.00%
	53.33%	42.88%	40.80%	52.63%	44.62%
	9.86%	27.20%	6.74%	.82%	44.62%
Female	105.00	441.00	119.00	9.00	674.00
	15.58%	65.43%	17.66%	1.34%	100.00%
	46.67%	57.12%	59.20%	47.37%	55.38%
	8.63%	36.24%	9.78%	.74%	55.38%
Total	225.00	772.00	201.00	19.00	1217.00
	18.49%	63.43%	16.52%	1.56%	100.00%
	100.00%	100.00%	100.00%	100.00%	100.00%
	18.49%	63.43%	16.52%	1.56%	100.00%

❖ PSPP Codes:

```
ELECT IF (SEX = 1 OR SEX = 2).
EXECUTE.

SELECT IF (BETTRLFE = 1 OR BETTRLFE = 2 OR BETTRLFE = 3 OR
BETTRLFE = 4).
EXECUTE.

GRAPH /BAR = COUNT BY SEX BY BETTRLFE.

CROSSTABS
    /TABLES= SEX        BY      BETTRLFE
    /FORMAT=AVALUE TABLES PIVOT
    /STATISTICS=CHISQ
    /CELLS=COUNT ROW COLUMN TOTAL.
```

~ The End ~

Thanks for your participation ☺

Asal Aslemand