Chi-square in SAS® Studio

We are going to utilize SAS Studio to conduct Chi-square test for categorical variables with two or more categories in each.

Dataset Description

We will be using the **PEW2020 data.** The PEW2020 data comes from the September 2020 Pew survey in which telephone interviews were "conducted Sept. 22-28, 2020, among a national sample of 1,007 adults, 18 years of age or older, livingin the United States (301 respondents were interviewed on a landline telephone, and 706 were interviewed on a mobile phone, including 487 who had no landline telephone). A combination of landline and mobile phone random-digit-dial samples were used. Interviews were conducted in English(972) and Spanish (35). The combined landline and mobile phone sample is weighted to provide nationally representative estimates of the adult population 18 years of age and older."

The questions on this survey are as follows:

1. Which country currently is the most important partner for American foreign policy?
2. In general, how would you describe relations today between the United States and Germany? Would you say they are very good, somewhat good, somewhat bad or very bad?
3. Which is more important for the United States?
    a. Having a close relationship to Germany or having a close relationship to Russia?
    b. Having a close relationship to Germany or having a close relationship to China?
4. How would you rate the likelihood of the current rivalry between China and the United States escalating into a confrontation resembling the Cold War?
5. For each of the following issues, do you see Germany as a partner or not?
    a. Protecting the environment
    b. Dealing with China
    c. Dealing with Iran
    d. Promoting free trade
    e. Protecting European security
    f. Protecting democracy and human rights around the world
6. Which of these statements comes closer to your view, even if neither is exactly right? Once the coronavirus crisis is over, do you think ...?

The demographic data collected includes:

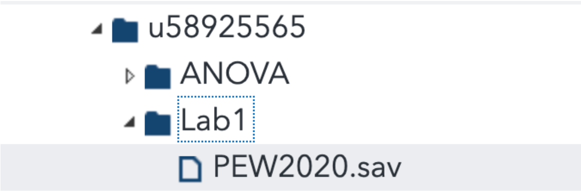
- How many of these are adults, 18 or older?
- What is your age?
- Case ID
- ‘Country’
- Date
- What is the highest level of school you have completed or degree you have received?
- Are you of Hispanic or Latino origin or descent?
- Is your total annual household income from all sources, and before taxes ...?
- ‘Marital status’
- Are you the parent or guardian of anyone under 18 in your household?
- As of today, do you lean more to the Republican Party or more to the Democratic Party?
- Generally speaking, would you describe your political views as ...?
- Race of Respondent
- What is your present religion, if any? Are you Protestant, Roman Catholic, Mormon, Orthodox such as Greek or Russian Orthodox, Jewish, Muslim, Buddhist, Hindu, atheist, agnostic, something else, or nothing in particular?
- ‘Sex of respondent’
- State
- ‘Survey’
- Including yourself, how many people are there living in your household?
- Weight

I. Select the Data Source and Performing a Chi-Square Analysis

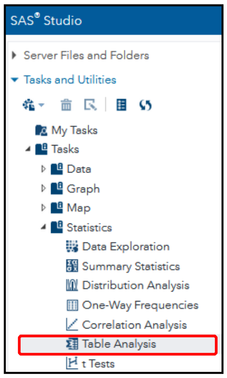
For this portion of the exercise, the question being tested is “Are men and women equally likely to prioritize US relationship to Germany over Russia?”. The results from question 3 in the Pew survey and the sex variable will be used to investigate this question.

First, let’s look at the counts for these two variables.

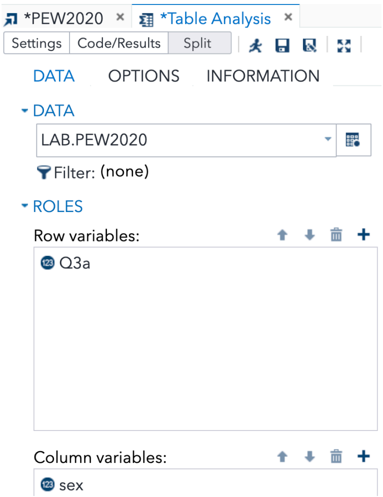
1. Bring the PEW2020 data to the software.



2. Go to Tasks and Utilities, expand Tasks, then expand Statistics, and select Table Analysis.



3. In the Table Analysis window:
   - under Data, check that the Pew2020 data set is selected
   - select Q3a as the row variable
   - select sex as the column variable



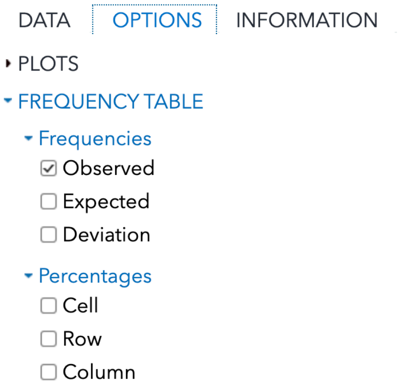
LAB.PEW2020

Filter: (none)

Row variables: Q3a

Column variables: sex

4. Make sure the Frequencies Observed is selected under options.



▸ PLOTS

▾ FREQUENCY TABLE

▾ Frequencies

☑ Observed

☐ Expected

☐ Deviation

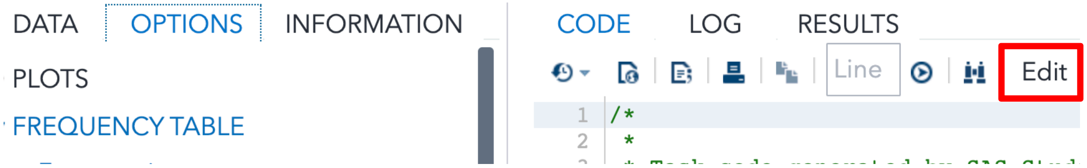
▾ Percentages

☐ Cell

☐ Row

☐ Column

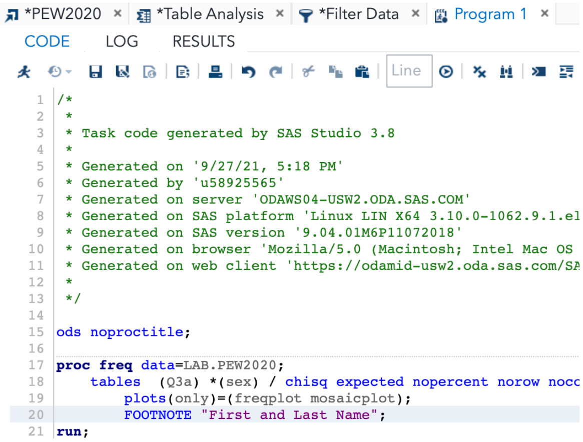
5. To add your First and Last Name in the footer of output, we will need to edit the SAS code. When the Code tab is highlighted, click Edit. A new program tab will be opened for you to edit the SAS code.



```
/*
 *
```

6. Go to the line before the Run statement, hit enter and then type the following code before the Run statement to add you own name to the output.

```
FOOTNOTE "First and Last Name";
```

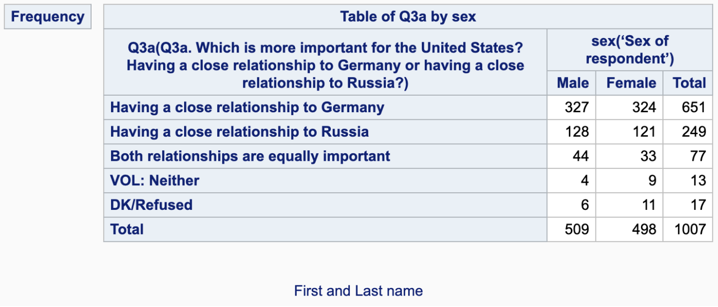


```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '9/27/21, 5:18 PM'
 * Generated by 'u58925565'
 * Generated on server 'ODAWS04-USW2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Macintosh; Intel Mac OS
 * Generated on web client 'https://odamid-usw2.oda.sas.com/SA
 *
 */

ods noproctitle;

proc freq data=LAB.PEW2020;
    tables  (Q3a) *(sex) / chisq expected nopercent norow nocc
        plots(only)=(freqplot mosaicplot);
        FOOTNOTE "First and Last Name";
run;
```

7. Click Run. You should obtain the following table:



Frequency

Table of Q3a by sex

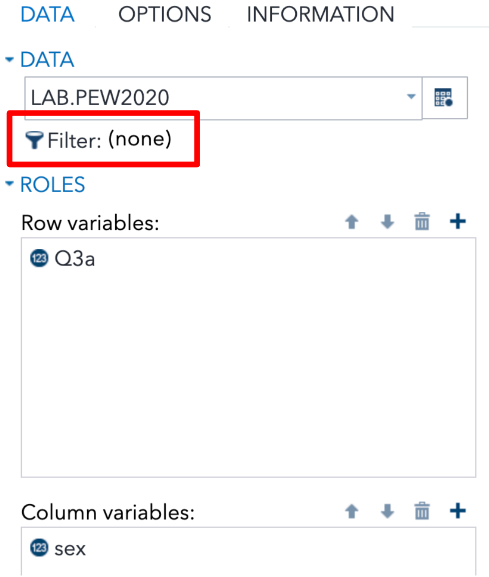
Q3a(Q3a. Which is more important for the United States? Having a close relationship to Germany or having a close relationship to Russia?)	sex('Sex of respondent')		
	Male	Female	Total
Having a close relationship to Germany	327	324	651
Having a close relationship to Russia	128	121	249
Both relationships are equally important	44	33	77
VOL: Neither	4	9	13
DK/Refused	6	11	17
Total	509	498	1007

First and Last name

8. Q3a is a numerical variable with user defined format.

Q3a Variable value	User defined format
1	Having a close relationship to Germany
2	Having a close relationship to Russia and exclude
3	Both relationships are equally important
4	VOL: Neither
9	DK/Refused

For this example, filter the data to only include 1: Having a close relationship to Germany and 2: Having a close relationship to Russia. To do this, click filter.



LAB.PEW2020

Filter: (none)

Row variables: Q3a

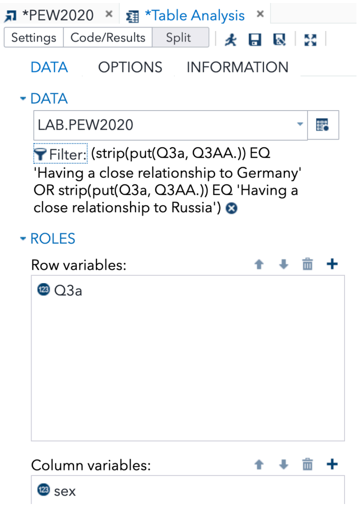
Column variables: sex

In filter box, input either of below codes.

```
(strip(put(Q3a, Q3AA.)) EQ 'Having a close relationship to Germany'
OR strip(put(Q3a, Q3AA.)) EQ 'Having a close relationship to
Russia')
```

or

```
Q3a EQ 1 OR Q3a EQ 2
```



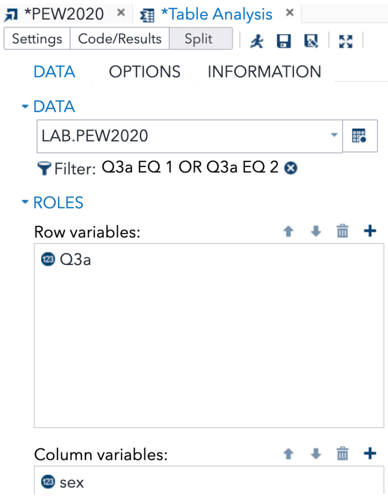
LAB.PEW2020

Filter: (strip(put(Q3a, Q3AA.)) EQ 'Having a close relationship to Germany' OR strip(put(Q3a, Q3AA.)) EQ 'Having a close relationship to Russia')

Row variables: Q3a

Column variables: sex

or



LAB.PEW2020

Filter: Q3a EQ 1 OR Q3a EQ 2

Row variables: Q3a

Column variables: sex

9. With Frequencies Observed selected under Options, click Run, you should have the following table.

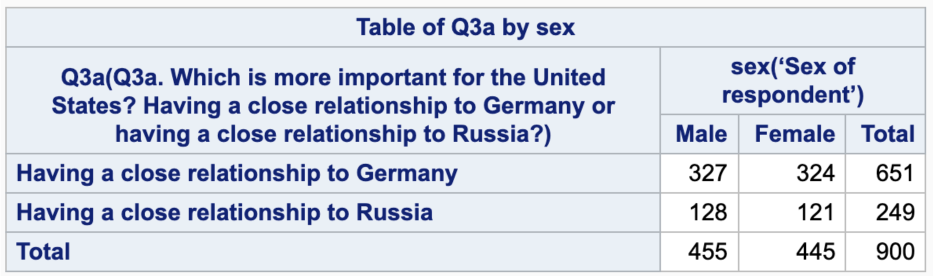


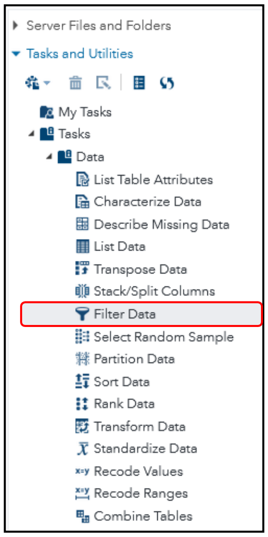
Table of Q3a by sex

Q3a(Q3a. Which is more important for the United States? Having a close relationship to Germany or having a close relationship to Russia?)	sex(‘Sex of respondent’)		
	Male	Female	Total
Having a close relationship to Germany	327	324	651
Having a close relationship to Russia	128	121	249
Total	455	445	900

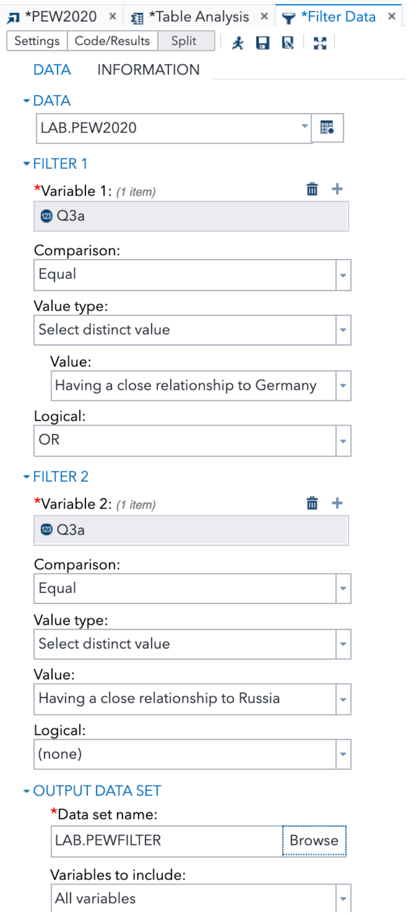
Note: If a footnote with your name is needed for this output, you must edit code as earlier rather than directly click Run.

Alternatively, you could filter data first and then perform the Table Analysis on the filtered data.

10. Under Tasks and Utilities, expand Tasks, expand Data, then select Filter Data.



11. In the filter window, select Q3a as Variable 1, set comparison to Equal, select value type to distinct value, select "Having a close relationship to Germany", then for logical select OR. Repeatthese selections for variable 2 to be "Having a close Relationship to Russia". Under Output, set dataset name as Lab5.PEWFILTER. Click Run.



DATA

LAB.PEW2020

FILTER 1

*Variable 1: *(1 item)*

Q3a

Comparison:

Equal

Value type:

Select distinct value

Value:

Having a close relationship to Germany

Logical:

OR

FILTER 2

*Variable 2: *(1 item)*

Q3a

Comparison:

Equal

Value type:

Select distinct value

Value:

Having a close relationship to Russia

Logical:

(none)

OUTPUT DATA SET

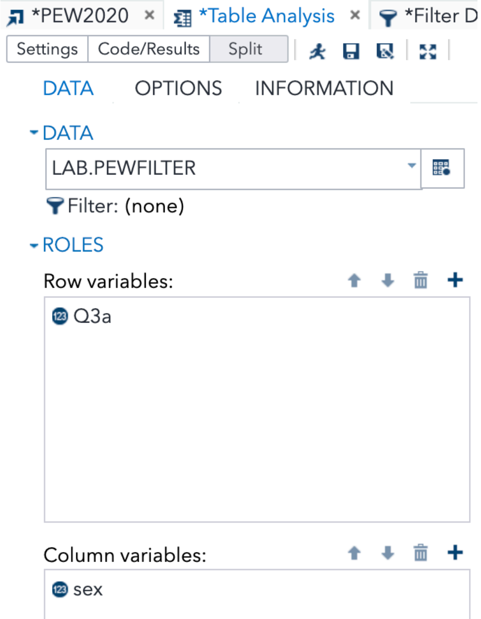
*Data set name:

LAB.PEWFILTER

Variables to include:

All variables

12. Go back In the Table Analysis window:
    - under Data, check that filtered data set is selected
    - select Q3a as the row variable
    - select sex as the column variable



DATA

LAB.PEWFILTER

Filter: (none)

ROLES

Row variables:

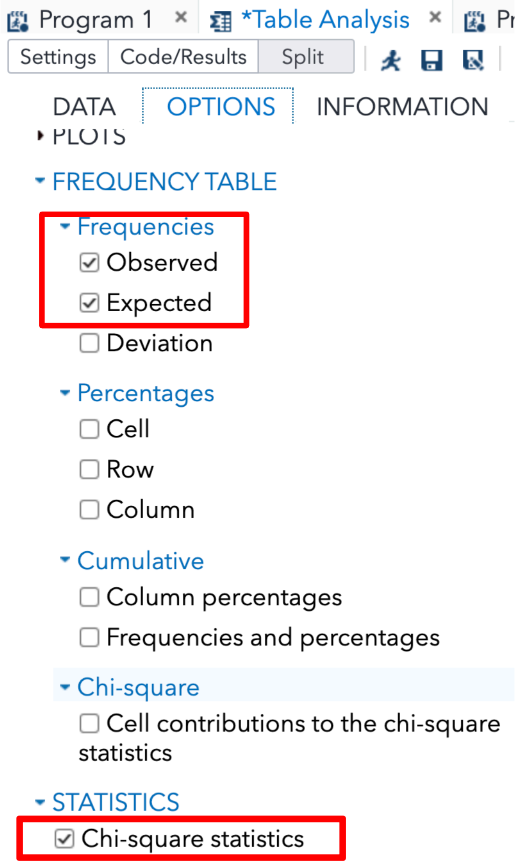
Q3a

Column variables:

sex

13. Click Run. You should generate the same table.

14. If we want to generate expected frequencies and Chi-square statistic too, in the Table Analysis task, select Frequencies Observed and Expected and Chi-square statistics.



FREQUENCY TABLE

Frequencies

☑ Observed

☑ Expected

☐ Deviation

Percentages

☐ Cell

☐ Row

☐ Column

Cumulative

☐ Column percentages

☐ Frequencies and percentages

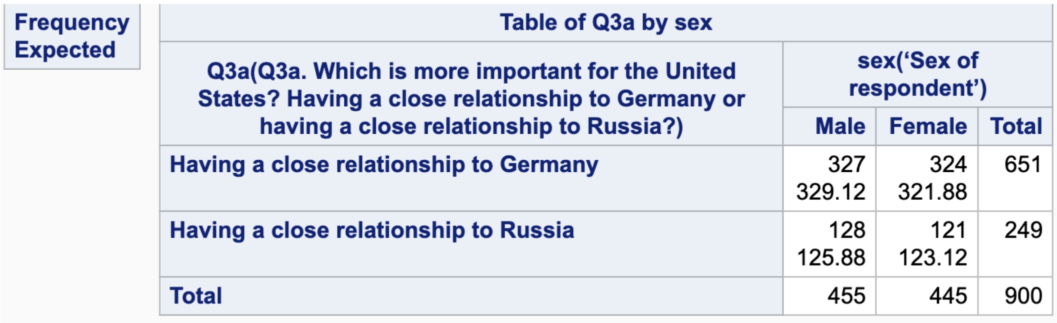
Chi-square

☐ Cell contributions to the chi-square statistics

STATISTICS

☑ Chi-square statistics

15. Click Run to obtain the following table from the output.

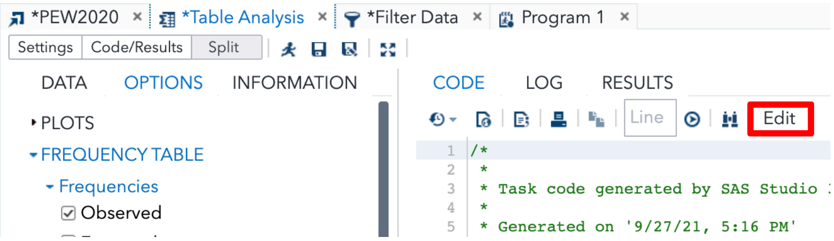


Frequency
Expected

Table of Q3a by sex

Q3a(Q3a. Which is more important for the United States? Having a close relationship to Germany or having a close relationship to Russia?)	sex('Sex of respondent')		
	Male	Female	Total
Having a close relationship to Germany	327 329.12	324 321.88	651
Having a close relationship to Russia	128 125.88	121 123.12	249
Total	455	445	900

16. Rather than directly click the Run, we can add footnote by editing the code in a newly opened program.

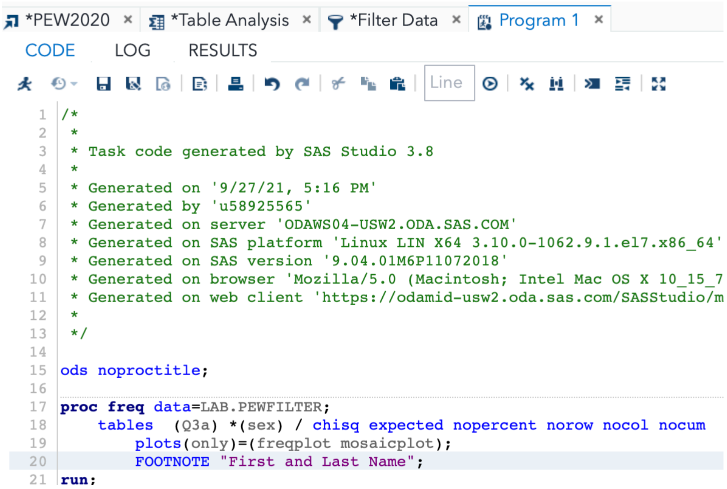


FREQUENCY TABLE

Frequencies

☑ Observed

```
/*
 *
 * Task code generated by SAS Studio :
 *
 * Generated on '9/27/21, 5:16 PM'
```



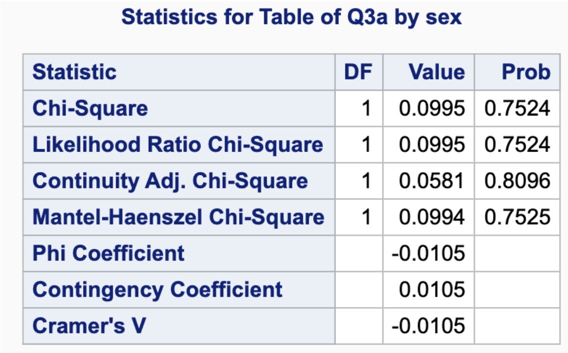
```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '9/27/21, 5:16 PM'
 * Generated by 'u58925565'
 * Generated on server 'ODAWS04-USW2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7
 * Generated on web client 'https://odamid-usw2.oda.sas.com/SASStudio/m
 *
 */

ods noproctitle;

proc freq data=LAB.PEWFILTER;
    tables  (Q3a) *(sex) / chisq expected nopercent norow nocol nocum
        plots(only)=(freqplot mosaicplot);
        FOOTNOTE "First and Last Name";
run;
```

17. Click Run, we can generate the same output with the footnote.

18. Scroll down in the output to find the following table.



Statistics for Table of Q3a by sex

Statistic	DF	Value	Prob
Chi-Square	1	0.0995	0.7524
Likelihood Ratio Chi-Square	1	0.0995	0.7524
Continuity Adj. Chi-Square	1	0.0581	0.8096
Mantel-Haenszel Chi-Square	1	0.0994	0.7525
Phi Coefficient		-0.0105	
Contingency Coefficient		0.0105	
Cramer's V		-0.0105	

The hypothesis statements for this problem are:
H_0: There is not a relationship between gender and the opinion for the preferred country that theybelieve the US should prioritize between Germany and Russia.
H_1: There is a relationship between gender and the opinion for the preferred country that they believe the US should prioritize between Germany and Russia.

We can look up the critical value for 1 degree of freedom at the 0.05 significance level (using a Chi-square table or the Chi-square Distribution in https://gallery.shinyapps.io/dist_calc/). This critical value is 3.841. Therefore, the decision rule for this scenario is to reject H_0 if $\chi^2 > 3.841$. Since ourcalculated χ^2 value is less than our critical value, we fail to reject the null hypothesis and conclude that there is insufficient evidence to support that there is a relationship between gender and opinion for prioritizing Russia or Germany. It is easier to read the Prob from table which shows the p-value of 0.7524 which also indicates the not significant test result.

In layman's term, an insignificant result implies that the gender and their opinion are independent events. If the Chi-square test was significant, we would infer that there is some observed cells within the frequency table outside the expected values. It is important to realize that as the number of categories (2×3, 2×4, 3×3, etc.) increase the Chi-square test will not identify which relationships are considered different enough from their expected probabilities on their own and so additional *post hoc* analysis is needed to isolate which combinations have a significant relationship.

Farther down in the output, you will see the Phi Coefficient, $\varphi = \sqrt{\frac{\chi^2}{n}}$, a Chi-square based measure of association. When you have a 2×2 analysis, such as the one in this example, φ can be interpreted as the symmetric percent difference which measures the percent of concentration of cases on the diagonal. The Phi Coefficient is used in the calculation of Cramer's V and describes the association between two nominal variables, so the values are constrained to being between 0 (no association) to 1 (perfect association). The formula used by SAS to calculate Cramer's V also allows for the retention of the sign for φ to retain the direction of the association, but we should ignore the sign for nominal variables. In this example, the Cramer's V value is -0.0105, which indicates that there is little to no association between the behavior of these two variables.

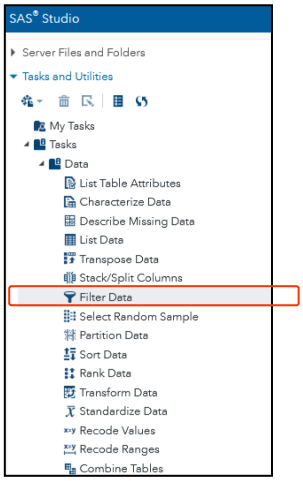
II. Prepare the Data for the Homework

For this example, we are interested in the research question “Are males more liberal than females?”. However, there are several values that need to be recoded first.

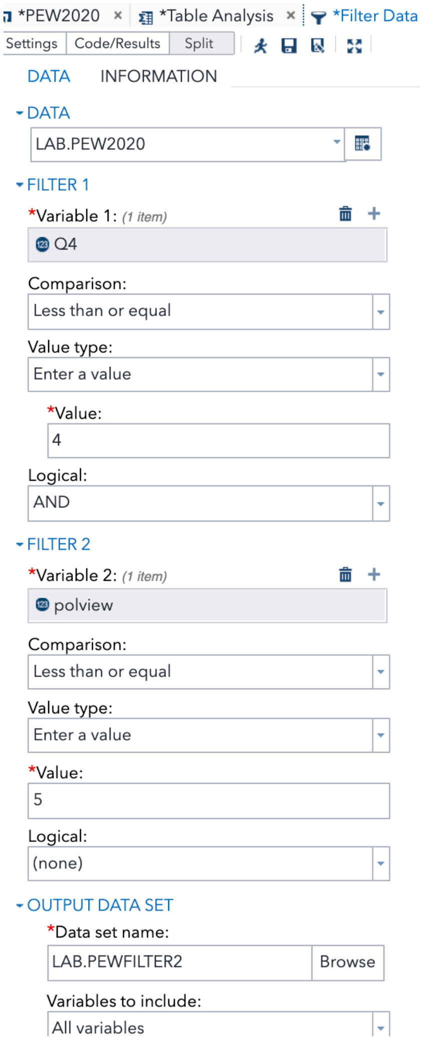
1. Select the PEW2020 Data that is in your homework library.
2. Q4 and polview variable are both numerical with user defined formats as below

Q4 Variable value	User defined format
1	Very likely
2	Somewhat likely
3	Somewhat unlikely
4	Very Unlikely
9	DK/Refused
polview Variable value	**User defined format**
1	Very conservative
2	Somewhat conservative
3	Moderate
4	Somewhat liberal
5	Very liberal
8	Don’t know
9	Refused

We will filter out the cases where Q4 takes the value of “DK/Refused” and where polview variable takes the value of “Don’t Know” or “Refused”. Go to Tasks and Utilities, expand tasks, then Expand Data, and select Filter Data.



3. In the filter data window, set Variable 1 to Q4 less than or equal to 4, Logical to AND, Variable 2 to polview less thanor equal to 5. Set the new data set name as Lab5.PEW2020FILTER2. This will create a data set that excludes cases where Q4 takes the value of “DK/Refused” and where polyview variable takes the value of “Don’t Know” or “Refused”. Click Run.



DATA

LAB.PEW2020

FILTER 1

*Variable 1: (1 item)
Q4

Comparison:
Less than or equal

Value type:
Enter a value

*Value:
4

Logical:
AND

FILTER 2

*Variable 2: (1 item)
polview

Comparison:
Less than or equal

Value type:
Enter a value

*Value:
5

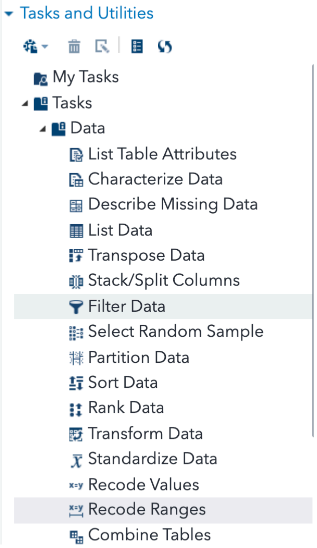
Logical:
(none)

OUTPUT DATA SET

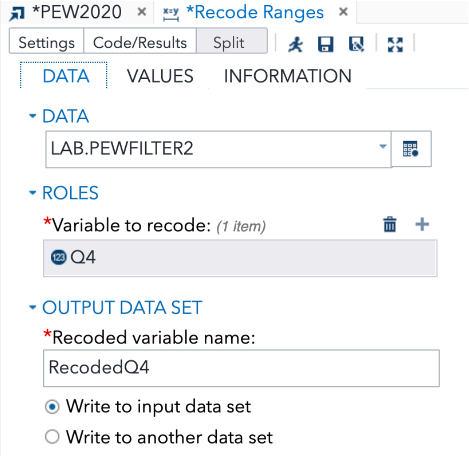
*Data set name:
LAB.PEWFILTER2 Browse

Variables to include:
All variables

4. We will recode some of your variables. Under Tasks and Utilities, Expand Tasks, expand data, and select Recode Ranges.



5. On the Data portion of the Recode Ranges window:
    - Make sure that you have the PEW2020FILTER2 file selected.
    - Under Roles, select “Numerical variable” and set Q4 as the variable to recode.
    - Under Output Data Set, set the recoded variable name to RecodedQ4 and select writeto input data set.



DATA

LAB.PEWFILTER2

ROLES

*Variable to recode: *(1 item)*

Q4

OUTPUT DATA SET

*Recoded variable name:

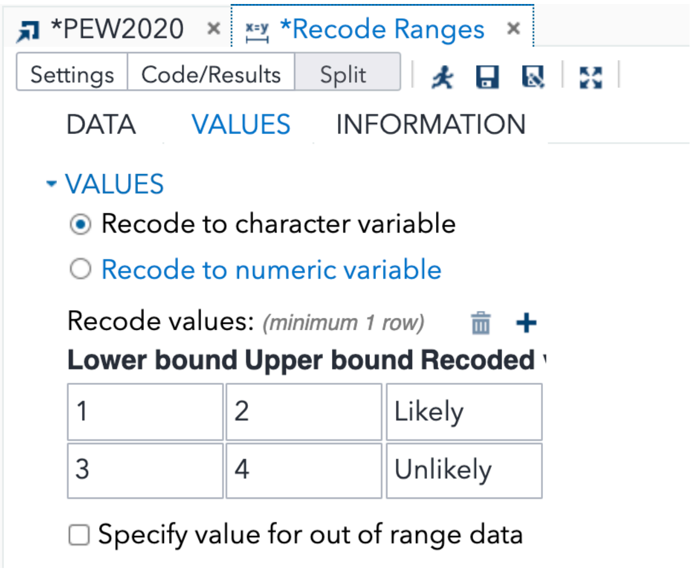
RecodedQ4

● Write to input data set

○ Write to another data set

6. On the values tab, use the following to prepare the old and new values.

Q4 Variable value	User defined format	RecodedQ4 Variable Values
1	Very likely	Likely
2	Somewhat likely	
3	Somewhat unlikely	Unlikely
4	Very Unlikely	



VALUES

● Recode to character variable

○ Recode to numeric variable

Recode values: *(minimum 1 row)*

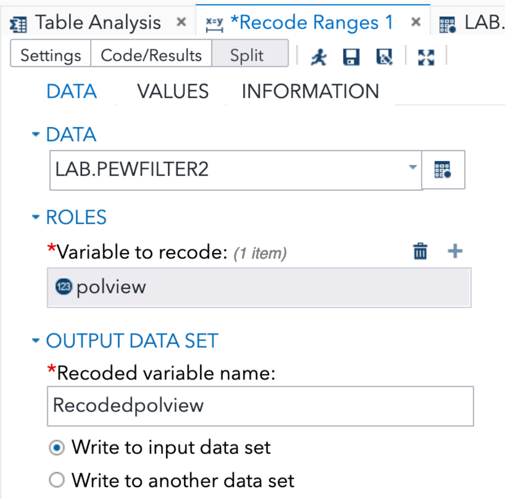
Lower bound	Upper bound	Recoded
1	2	Likely
3	4	Unlikely

☐ Specify value for out of range data

7. Click Run. Open the data set to confirm that the variable RecodedQ4 is in your list of variables.

8. Open another Recode Ranges window to recode the values for polview to make a new variable named Recodedpolview in the input data set as follows.

polview Variable value	User defined format	Recodedpolview Variable Values
1	Very conservative	Conservative
2	Somewhat conservative	
3	Moderate	Moderate
4	Somewhat liberal	Liberal
5	Very liberal	



DATA

LAB.PEWFILTER2

ROLES

*Variable to recode: (1 item)

polview

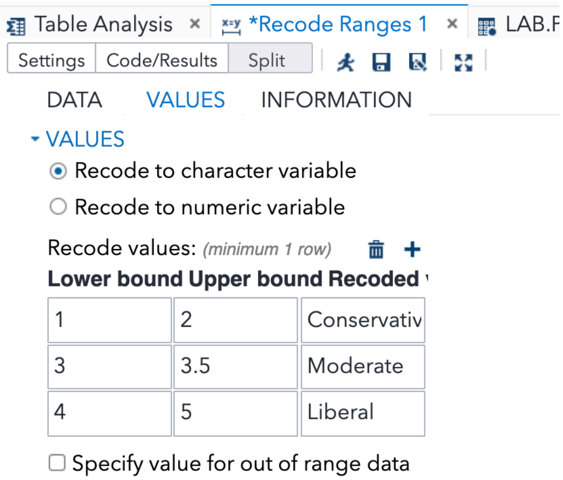
OUTPUT DATA SET

*Recoded variable name:

Recodedpolview

● Write to input data set

○ Write to another data set



VALUES

● Recode to character variable

○ Recode to numeric variable

Recode values: (minimum 1 row)

Lower bound	Upper bound	Recoded '
1	2	Conservativ
3	3.5	Moderate
4	5	Liberal

□ Specify value for out of range data

Note that the upper and lower bound of the range for Moderate are both 3 which will be considered wrong in SAS, so we can set 3.5 as a upper bound even if 3.5 is not an existing value of the polview variable that only takes integers.

9. Click Run. You should now have both the RecodedQ4 variable and the Recodedpolview variable in this data set.