# Chi Square in SAS® Enterprise Guide™

In this lesson we are going to utilize some of the SAS Enterprise Guide™ tools to conduct Chi-squared test for categorical variables with two or more categories in each. It is used mostly to compare percentages and we can use this test under the assumptions that we have independent observations.

## Dataset Description

**For this example, you will be revisiting the PEW2020 data.** About the PEW2020 data that you will be using:

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, living in 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:

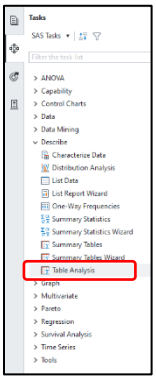
- How many of these are adults, 18 or older?
- What is your age?
- Case ID
- 'Country'
- Date
- What is the highest level of school you have completed or degree you have received?
- Are you of Hispanic or Latino origin or descent?
- Is your total annual household income from all sources, and before taxes ...?
- '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

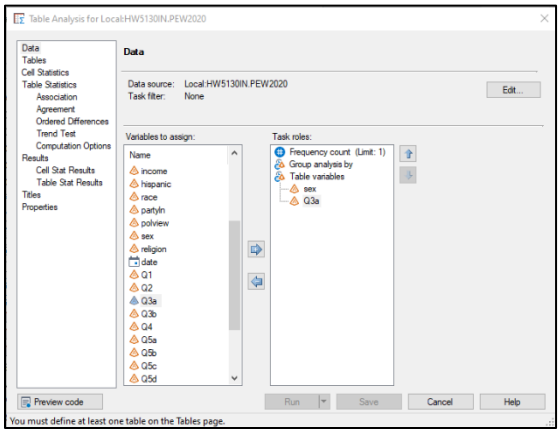
For this portion of the exercise, the question being tested is "Are men and women equally likely to prioritize US relationship to Germany over Russia?". The results from question 3 in the Pew survey and the sex variable will be used to investigate this question.

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

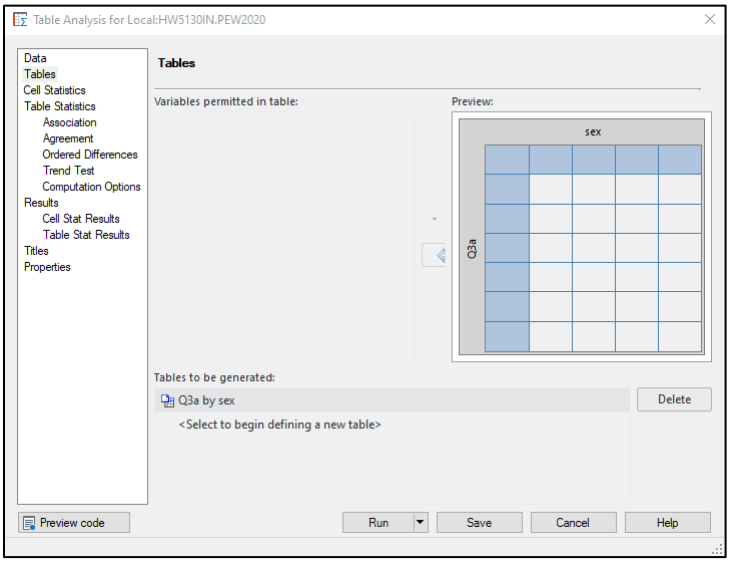
1. Bring the Pew2020 data set onto your Process Flow from your homework library.
2. Go to Tasks, expand Describe, and select Table Analysis.



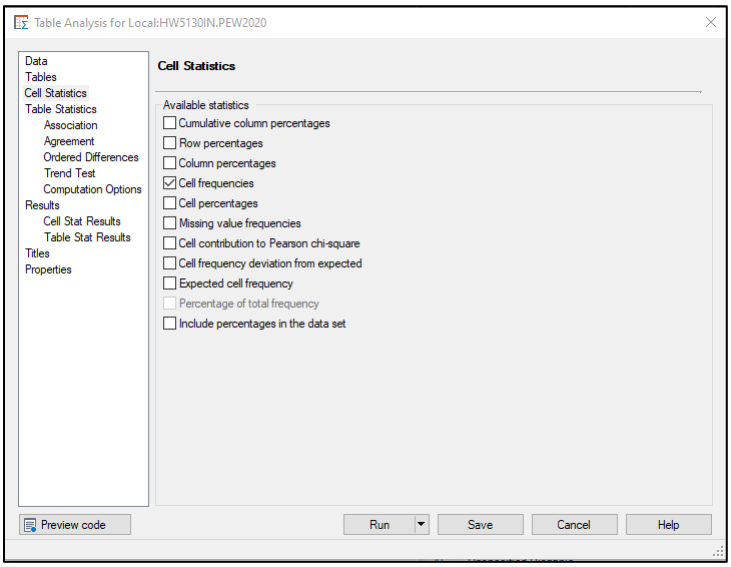

3. In the Table Analysis window:
   - Under Data, check that the Pew2020 data set is selected.
   - Select Q3a and sex as analysis variables.



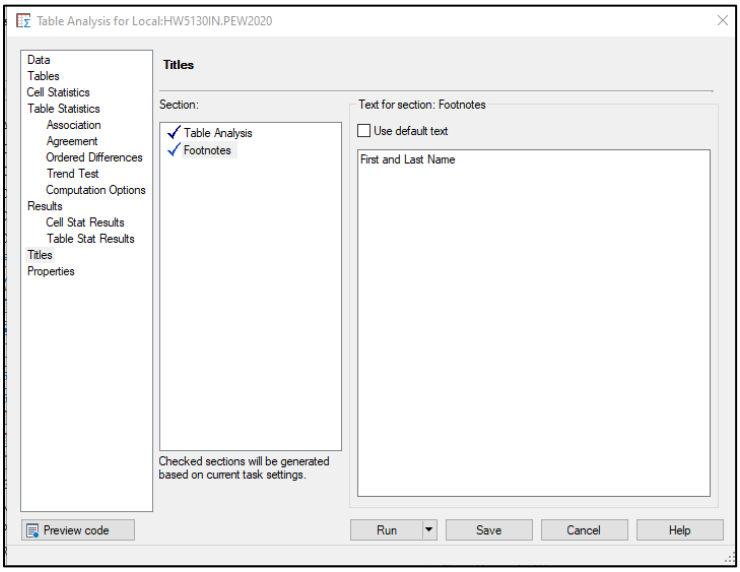
4. Under Tables, set sex as the columns and Q3a as the rows.




5. Under Cell Statistics, ensure that only Cell frequencies is selected.



6. Go to Titles to set the footnote as your First and Last Name.



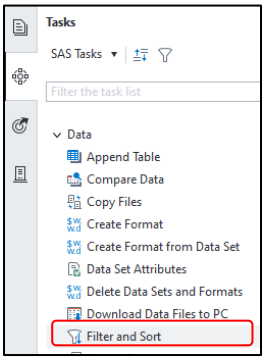

7. Click Run. You should obtain the following output:

**Table of Q3a by sex**

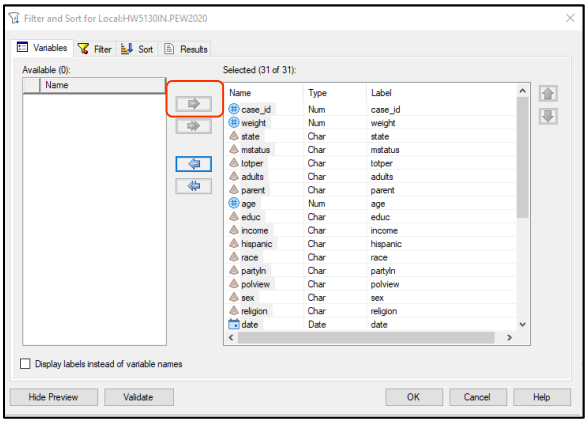
| Q3a | sex | | |
|---|---|---|---|
| | Female | Male | Total |
| Both relationships are equally important | 33 | 44 | 77 |
| DK/Refused | 11 | 6 | 17 |
| Having a close relationship to Germany | 324 | 327 | 651 |
| Having a close relationship to Russia | 121 | 128 | 249 |
| VOL: Neither | 9 | 4 | 13 |
| Total | 498 | 509 | 1007 |

First and Last Name

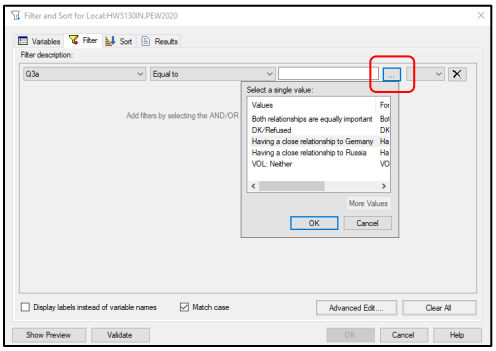
8. For this example, filter the data to only include 1. Having a close relationship to Germany and 2. Having a close relationship to Russia and exclude 3. Both relationships are equally important, 4. VOL: Neither, and 9. Don't Know / Refused. To do this, go to Tasks, expand Data, and select Filter and Sort.



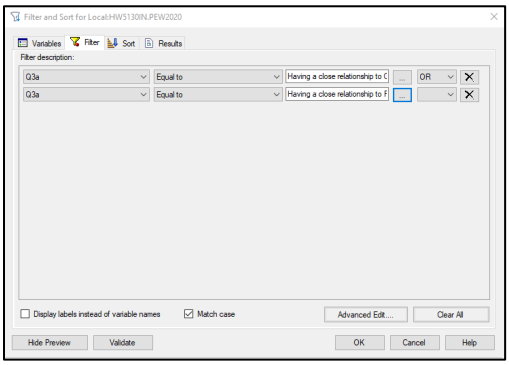

9. In the filter and sort window, highlight all of the variables and click the arrow pointing to the right to move them into the "Selected"column.



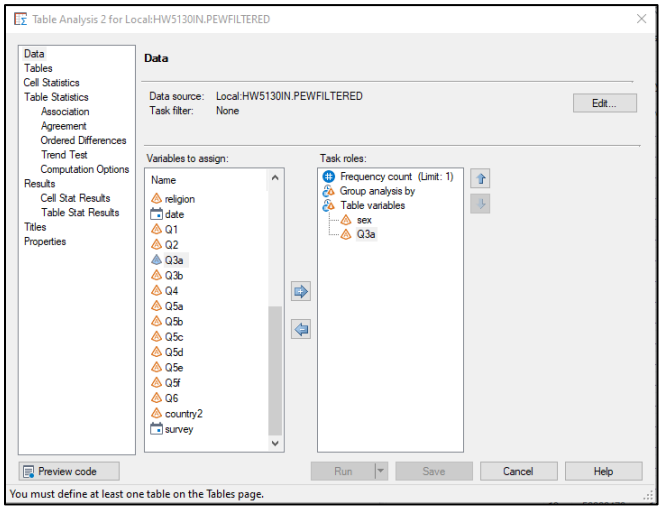

10. On the Filter tab. In the far left drop down menu, select Q3a. For the next drop down box, select "Equal to". For the next box, click on the ellipse and select "Having a close relationship to Germany".



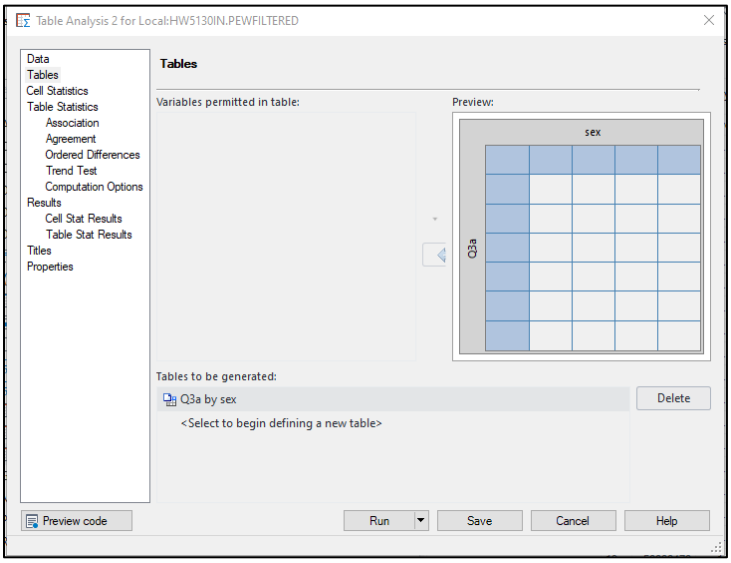
11. In the last box, select “or”. Then, follow the same procedure for “Having a close Relationship to Russia”.



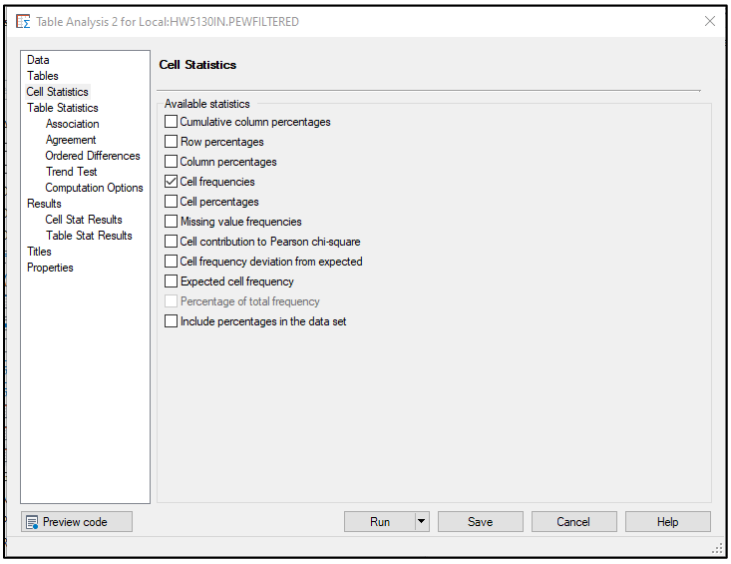

12. On the Results Tab, title the Output Name, label the dataset “PewFiltered” in your homework library. Click OK.
13. Now, let’s perform the analysis. Go back In the Table Analysis window:
    - under Data, check that the PewFiltered data set is selected
    - select Q3a and sex as table variables



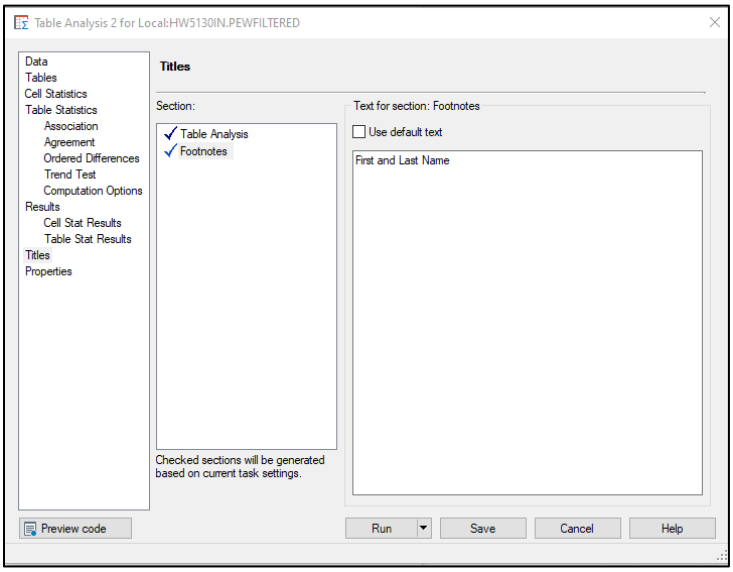
14. Under Tables, set sex as the columns and Q3a as the rows.




15. Under Cell Statistics, ensure that only Cell frequencies is selected.



16. Go to Titles to set the footnote as your First and Last Name.



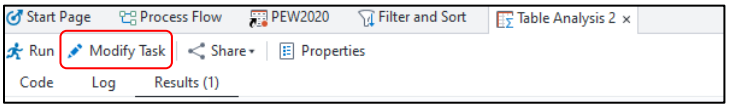

17. Click Run. You should obtain the following output:

Table of Q3a by sex

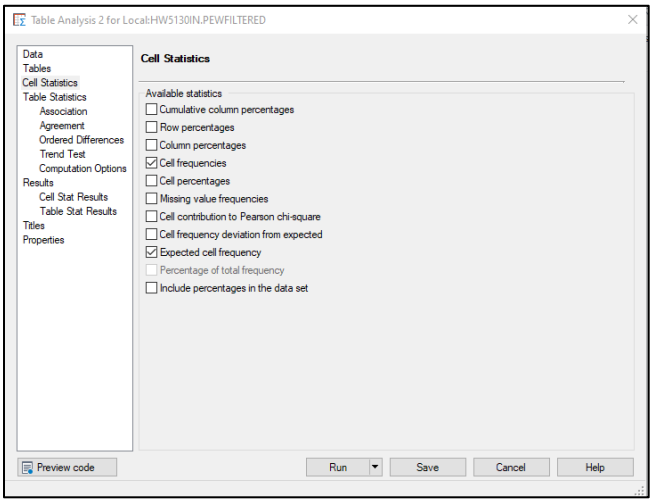
| Q3a | sex | | |
|---|---|---|---|
| | Female | Male | Total |
| Having a close relationship to Germany | 324 | 327 | 651 |
| Having a close relationship to Russia | 121 | 128 | 249 |
| Total | 445 | 455 | 900 |

First and Last Name

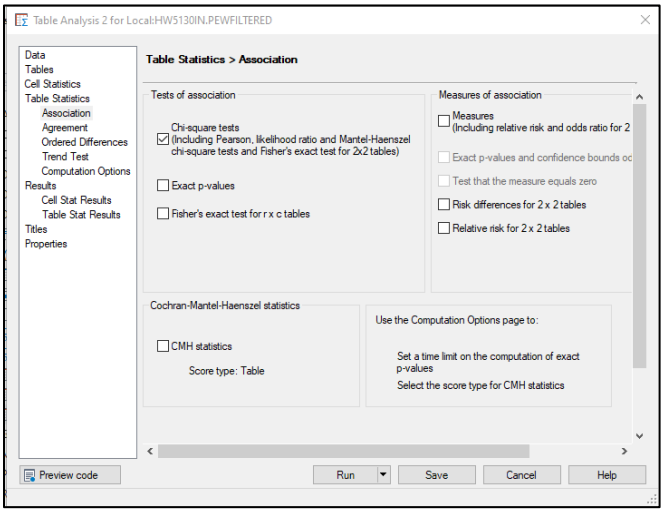
18. Now, select Modify Task.



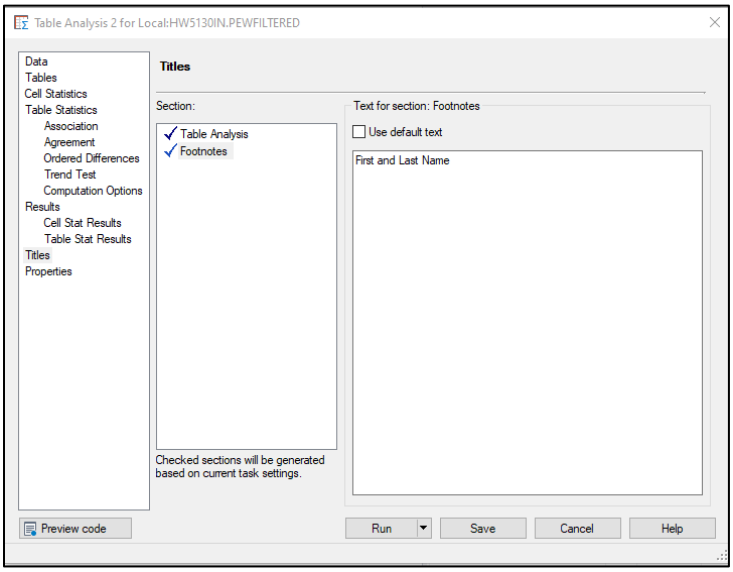

19. Under Cell Statistics, select Cell frequencies and Expected cell frequencies.




20. Under Table Statistics: Association, select Chi-square tests.



21. Go to Titles to set the footnote as your First and Last Name.




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

| Table of Q3a by sex | | | |
|---|---|---|---|
| | sex | | |
| Q3a | Female | Male | Total |
| Having a close relationship to Germany | 324<br>321.88 | 327<br>329.12 | 651 |
| Having a close relationship to Russia | 121<br>123.12 | 128<br>125.88 | 249 |
| Total | 445 | 455 | 900 |

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

| 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 is:
$H_0$: There is not a relationship between gender and the opinion for the preferred country that they believe 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.

Next, we need to look up the critical value for 1 degree of freedom at the 0.05 significance level. This critical value is 3.841. Therefore, the decision rule for this scenario is to reject $H_0$ if $\chi^2 > 3.841$. Since our calculated $\chi^2$ 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.

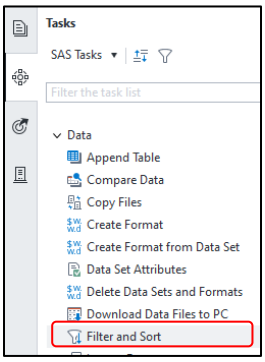
In layman's term, with an insignificant result it implies that the probabilities between gender and their opinion is therefore an independent event. This would then allow you to conclude that gender does not impact the relationship tied to the opinion. In the event that instead the Chi Square test was positive, we would infer that there is some at least one of the cells with the Chi Square frequency table that is outside the expected distribution of 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* testing is required to isolate which combinations have a significant relationship.

Farther down in the output, you will see the Phi Coefficient, $\varphi$, which is a Chi-Square based measure of association. The Chi-Square coefficient depends on both the sample size and the strength of the relationship. The Phi Coefficient is independent of the sample size. When you have a 2×2 analysis, such as the one in this example, $\varphi$ can be interpreted as the symmetric percent difference which measures the percent of concentration of cases on the diagonal. Additionally, in a 2×2 analysis it is identical to the correlation coefficient. The Phi Coefficient is used in the calculation of Cramers V, is also used to describe the association between two nominal variables, but 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 $\varphi$ to retain the direction of the association. In this example, the Cramer's V value is -0.0105, which indicates that while the groups may be going in different directions (i.e., making different selections), there is little to no association between the behavior of these two groups.

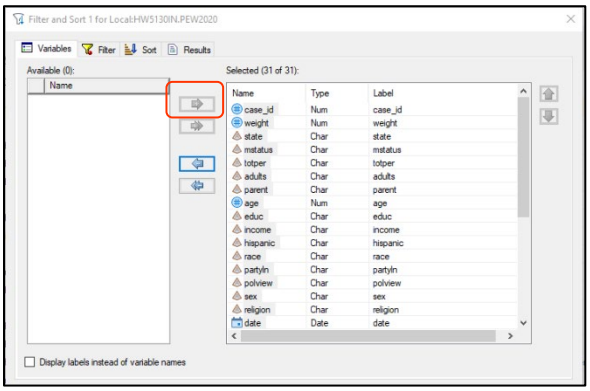
## II. **Prepare the Data for the Homework**

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.

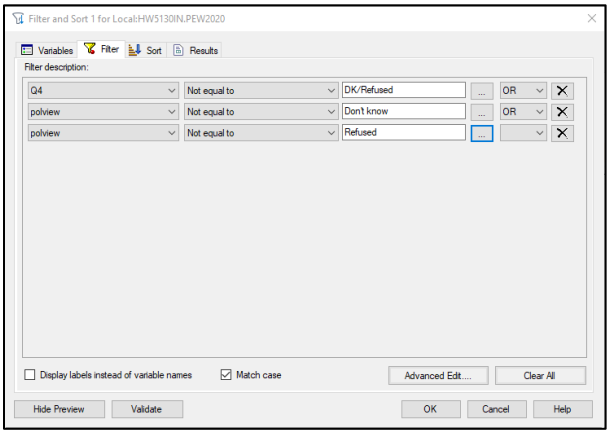
1. Select the Pew2020 Data that is in your homework library.
2. We will begin by filtering out the cases where Q4 takes the value of “DK/Refused” and where polyview variable takes the value of “Don’t Know” or “Refused”. To do this, go to Tasks, expand Data, and select Filter and Sort.



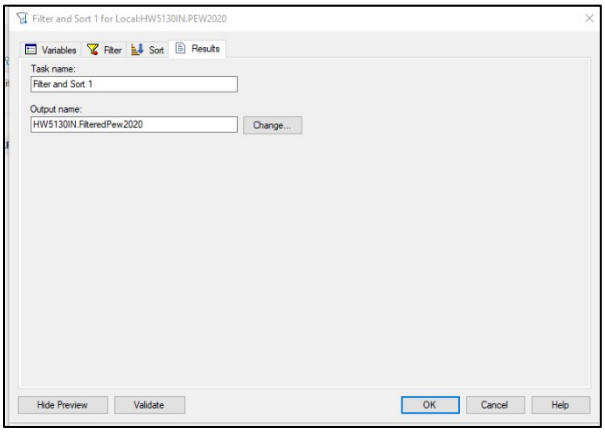

3. In the filter and sort window, highlight all of the variables and click the arrow pointing to the right to move them into the “Selected” column.



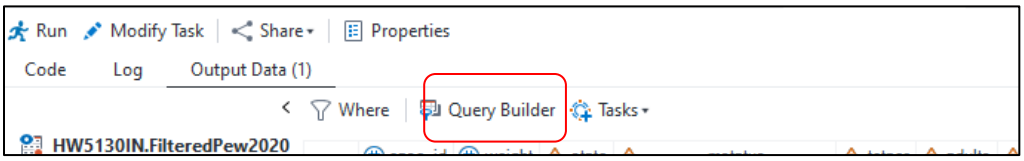

4. On the Filter tab. In the filter data window, set Q4 not equal to “DK/Refused”, or polview not equal to “Don’t Know”, or polview not equal to “Refused”.



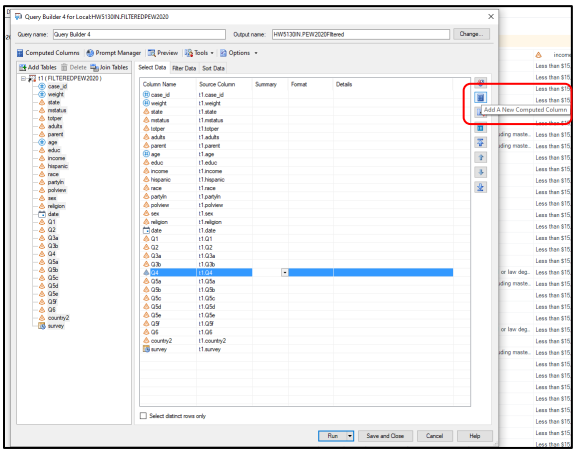
5. On the results tab, name the output FilteredPew2020.



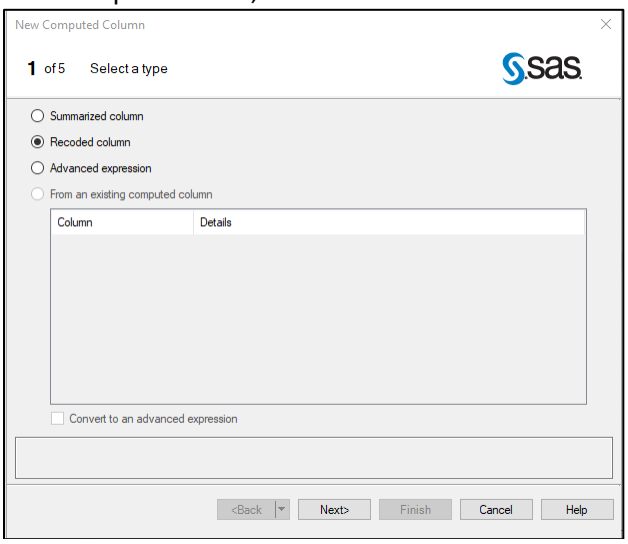

6. Click OK.
7. Now, you need to recode some of your variables. Select Query Builder.



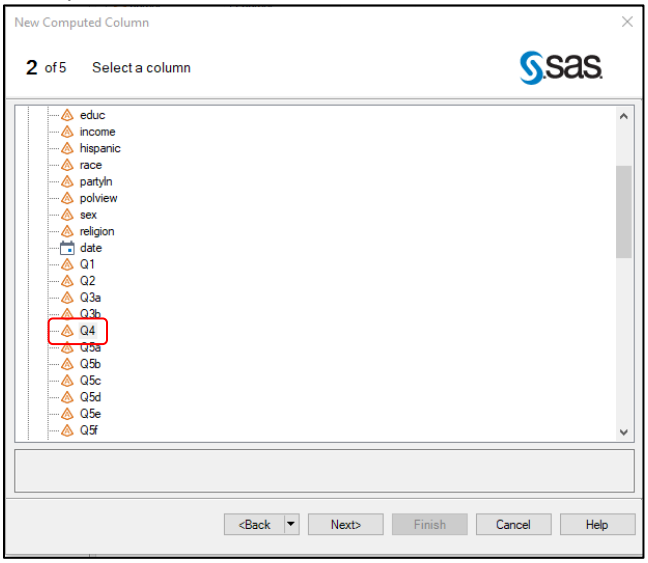

8. In the Query Builder window:
    - Change the output name to Pew2020Filtered in your homework library.
    - Select all of the variables and bring them into the selected data tab.
    - Click the “Add a New Computed Column” button.



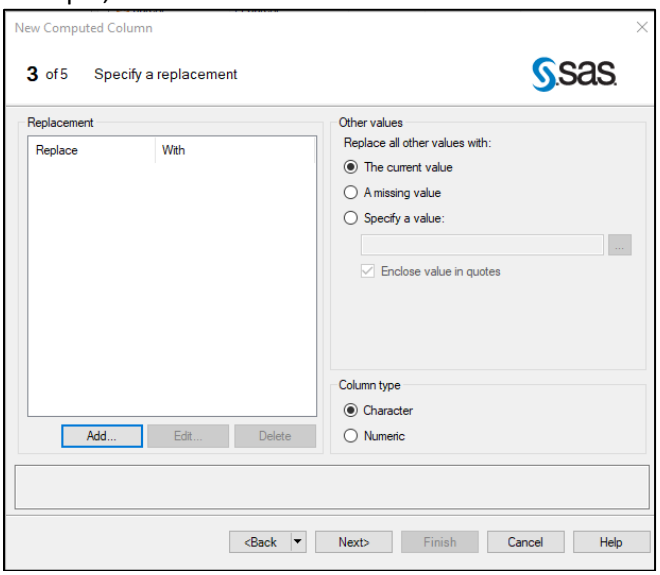
9. In the step 1 window, select “Recoded column”.



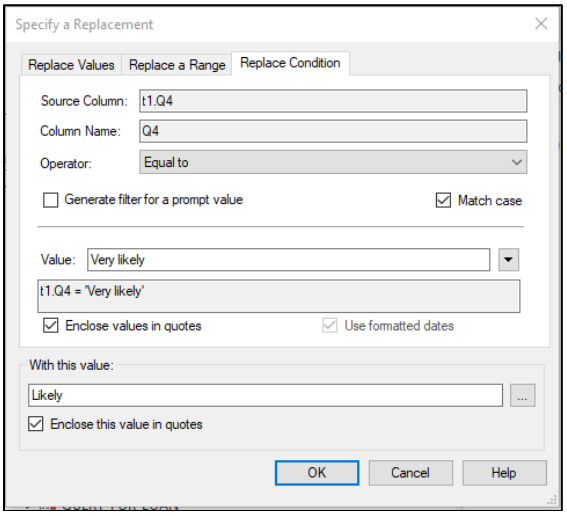

10. In step 2, select Q4. Click Next.



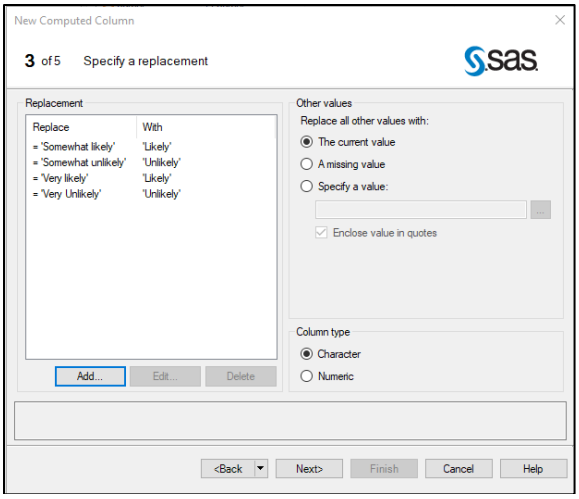
11. In step 5, select Add to create a new condition for recoding your data.



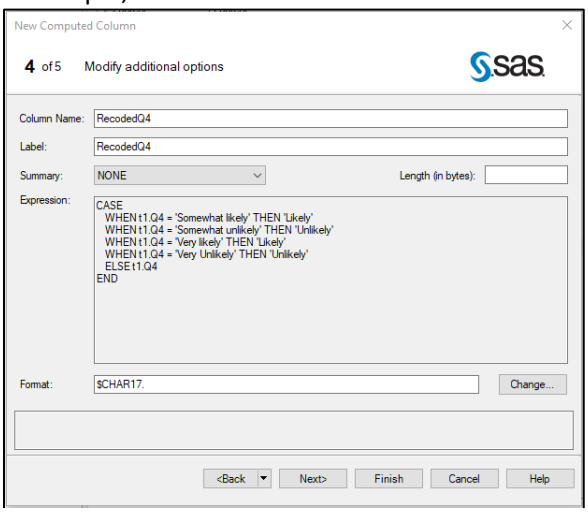

12. Select the Replace Condition tab. Set the Operator to “equal to”. Set the value Very likely (select “Enclose value in quotes”) to be replaced with Likely (select “Enclose value in quotes”). Click OK.



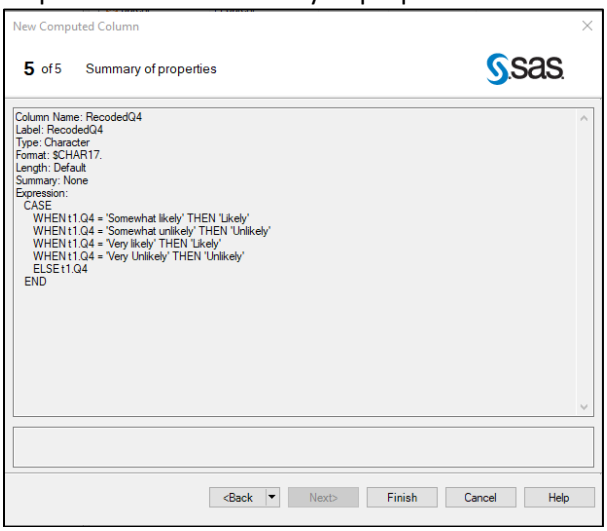

13. Repeat this process for coding “Somewhat likely” as “Likely” as well as “Somewhat Unlikely” and “Very Unlikely” as Unlikely. Keep “Replace all other values with” “the current value” selected. Click Next.




14. For Step 4, name and label the new column “RecodedQ4”. Click Next.




15. Step 5 shows the summary of properties. Click Finish.



16. Now, follow this same procedure to recode the following values for the variable “polview”, which should be renamed “RecodedPolview”
    - Change “Very conservative” and “Somewhat conservative” to conservative.
    - Change “Very liberal” and “Somewhat liberal” to libral.

17. Click Run. You should now see both the RecodedQ4 variable and the recodedpolview variable in this data set.