Bonus Activity: Data Preparation Using Gephi, Excel, and OpenRefine

This bonus section shows you how this Movie Actor dataset was created. It was partially created using one of the automation plugins mentioned in the workshop, "Convert Excel and csv files to networks"

1. The original IMDB movie dataset was found on data.world. (Note: You will need to complete a free registration with data.world to access the data.)
2. Download the dataset from the link above and open up the csv file in Excel and save it as an .xlsx file so that it works better with the plugin.
3. Install the plugin you will need in Gephi by going to the *Tools* menu and then selecting *Plugins.* From the *Available Plugins* tab, search for "Convert Excel and csv files to networks"
4. Select it from the list, and then click on the *Install* button below the list and follow the prompts to install the plugin. Note: Installing plugins can only be done on your own computer. This plugin should already be installed in the Map & Data Library computer lab.
5. Next, select *File* and then *New Project.*
6. Select *File* and then *Import...*
7. Under *Category*, select *Data importer (co-occurrences)*. You should see *Convert Excel and csv files to networks* listed in under *Wizard Type,* and then click *Next*.
8. Click on the *select file* button and browse to the *IMDB-Movie-Data.xls* file in the *Movie Actors* folder, leave the defaults when prompted to select the appropriate sheet, and then click *Next*.
9. For options 1 and 2, select *Actors* from the list. We are going to create a connection between two actors if they starred in the same movie together (i.e., they are listed in the same row together in the spreadsheet). Then click *Next.*
10. Select *comma* for the delimiter and click on *Next*.
11. We are not creating a dynamic network here, so just click on *Next.*
12. Select all three options and then click on *Next.*
13. This screen will confirm how the network will be constructed and what attributes will be captured. Click on *Finish.*
14. Make sure to select that it is an *Undirected* graph, select *Append to existing workspace*, and click on *OK.*
15. Examine the nodes and edges tables to see the resulting dataset.

For this activity's dataset (which was modified from the original dataset), I continued to modify this dataset. I used *Export table* to export the edges tables to a csv file. From there I sorted the edges by weight, only keeping edges with a weight of 2 or more. Using the edge table and Excel, I worked backwards to create a unique list of nodes (this tutorial explains how).

I then used a data cleaning tool called OpenRefine to augment my actor data to also include date of birth and country of citizenship. I also used the tool to do a quick calculation for age subtracting their birth year from 2021. To learn more about OpenRefine, you can take our Working with Messy Data in OpenRefine self-paced online course OR follow our online OpenRefine tutorials (the augmenting activity 2 is most directly related to this example).