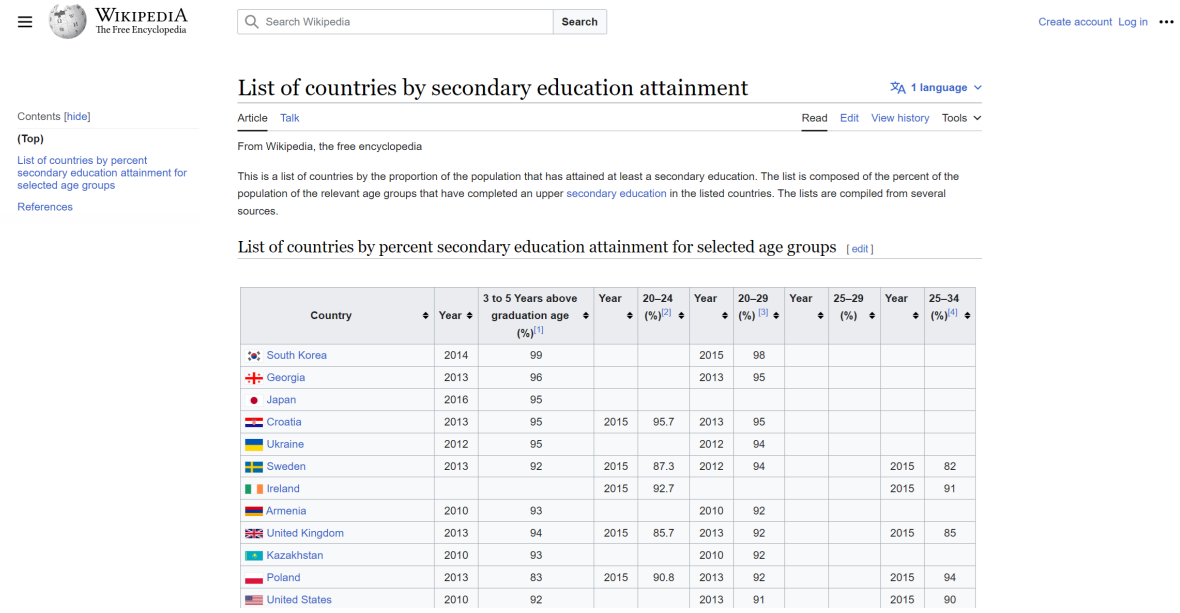
Lecture Note #19: Data Analysis Tools Part #2

BUSI 201: Business Data Analysis

Topic 1. Manually Importing Data

Sometimes, you will be the one recording data in an Excel spreadsheet. But sometimes, you will be importing data from outside sources into Excel to perform data analysis. We can either manually import data ourselves, or rely on built-in tools that Excel has to offer. First, we will examine some basic manual data importing from outside sources.

Suppose you are interested secondary education attainment around the world. A quick search may lead you to a `Wikipedia article` titled "List of countries by secondary education attainment."[1] Figure 1 below is a screenshot of said webpage captured as of November 2023.



Search Wikipedia | Search | Create account | Log in

List of countries by secondary education attainment

1 language

Article | Talk | Read | Edit | View history | Tools

Contents [hide]

(Top)

List of countries by percent secondary education attainment for selected age groups

References

From Wikipedia, the free encyclopedia

This is a list of countries by the proportion of the population that has attained at least a secondary education. The list is composed of the percent of the population of the relevant age groups that have completed an upper secondary education in the listed countries. The lists are compiled from several sources.

List of countries by percent secondary education attainment for selected age groups [edit]

Country	Year	3 to 5 Years above graduation age (%)[1]	Year	20–24 (%)[2]	Year	20–29 (%)[3]	Year	25–29 (%)	Year	25–34 (%)[4]
South Korea	2014	99			2015	98				
Georgia	2013	96			2013	95				
Japan	2016	95								
Croatia	2013	95	2015	95.7	2013	95				
Ukraine	2012	95			2012	94				
Sweden	2013	92	2015	87.3	2012	94			2015	82
Ireland			2015	92.7					2015	91
Armenia	2010	93			2010	92				
United Kingdom	2013	94	2015	85.7	2013	92			2015	85
Kazakhstan	2010	93			2010	92				
Poland	2013	83	2015	90.8	2013	92			2015	94
United States	2010	92			2013	91			2015	90

Figure 1: Wikipedia Article

One way to import this data into Excel is to simply copy and paste the entire table. You can copy the data in the table by left clicking and dragging to select the table, and then right clicking the selected table, and selecting copy.

[1] https://en.wikipedia.org/wiki/List_of_countries_by_secondary_education_attainment

You can choose two options when pasting the table data into Excel. You can choose to `Keep Source Formatting`, or `Match Destination Formatting` as shown in Figure 2. We will primarily be using the latter, as the source formatting is not necessarily well translated over to Excel.

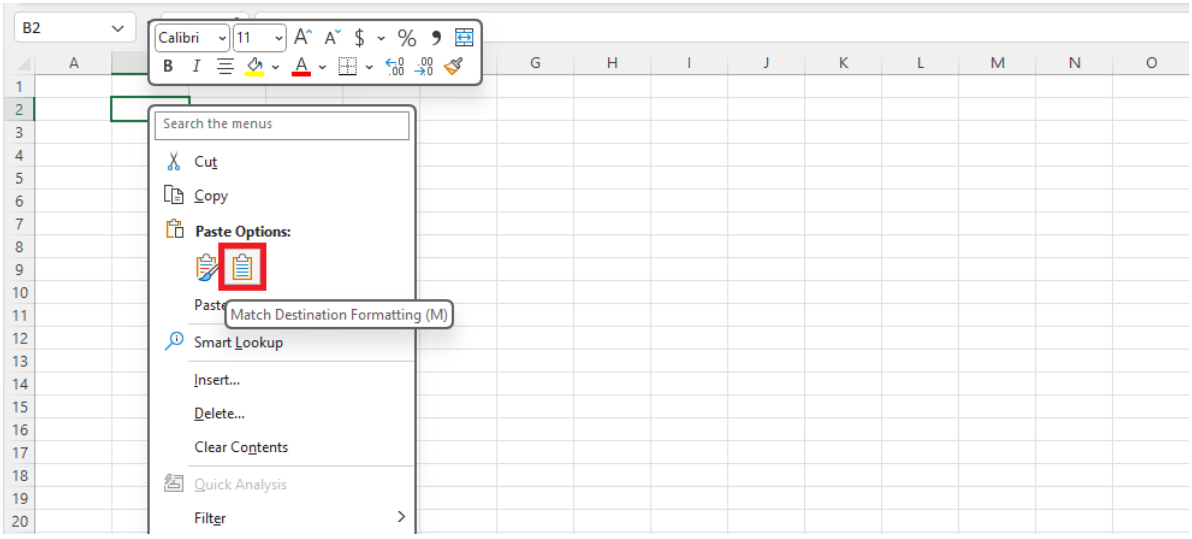


Figure 2: Pasting Options

Pasting the table we copied eariler while matching destination formatting, we can import the table as shown in Figure 3. Now that we have the table in Excel, we can use the tools that we have at our disposal to "clean" the data. Remove redundant rows and columns, sorting data by educational attainment, color-coding the table using conditional formatting, creating charts to visualize data, etc.

Country	Year	3 to 5 Year graduation (%)[1]	Year	20–24 age (%)[2]	Year	20–29 (%) [3]	Year	25–29 (%)	Year	25–34 (%)[4]
South Kor	2014	99			2015	98				
Georgia	2013	96			2013	95				
Japan	2016	95								
Croatia	2013	95	2015	95.7	2013	95				
Ukraine	2012	95			2012	94				
Sweden	2013	92	2015	87.3	2012	94			2015	82
Ireland			2015	92.7					2015	91
Armenia	2010	93			2010	92				
United Ki	2013	94	2015	85.7	2013	92			2015	85
Kazakhsta	2010	93			2010	92				
Poland	2013	83	2015	90.8	2013	92			2015	94
United St	2010	92			2013	91			2015	90
Canada	2010	86			2010	91			2015	93
Greece	2013	92	2015	89.6	2013	91			2015	84
Slovakia	2013	93	2015	91.3	2013	90			2015	93

Figure 3: Imported Table

This is a rather straightforward example of importing data. The original source material was already formatted as a table, and the importing process required little customization. Now, let us examine a case where the data requires a bit more work

CSV: Comma Separated Values

In some cases, you will encounter files in the form of pdfs or txt files. One such example can be found by downloading the `BUSI201-LEC19-txt` file. This file lists the top 20 movies of all time based on IMDB review scores as of November 2023.

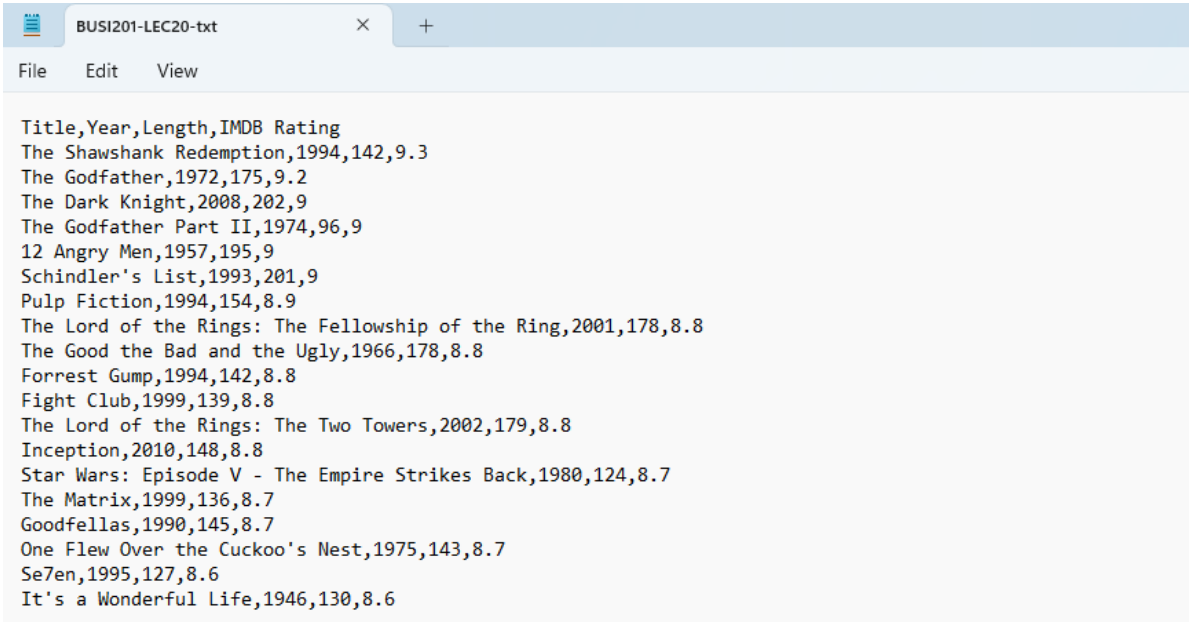


Figure 4: Data in TXT Format

Copy the text file, and paste in into an Excel spreadsheet. The initial result will not be ideal, since each line in the text file will populate a single cell. We must call up the text import wizard by clicking `Paste Options`, and then `Use Text Import Wizard`.

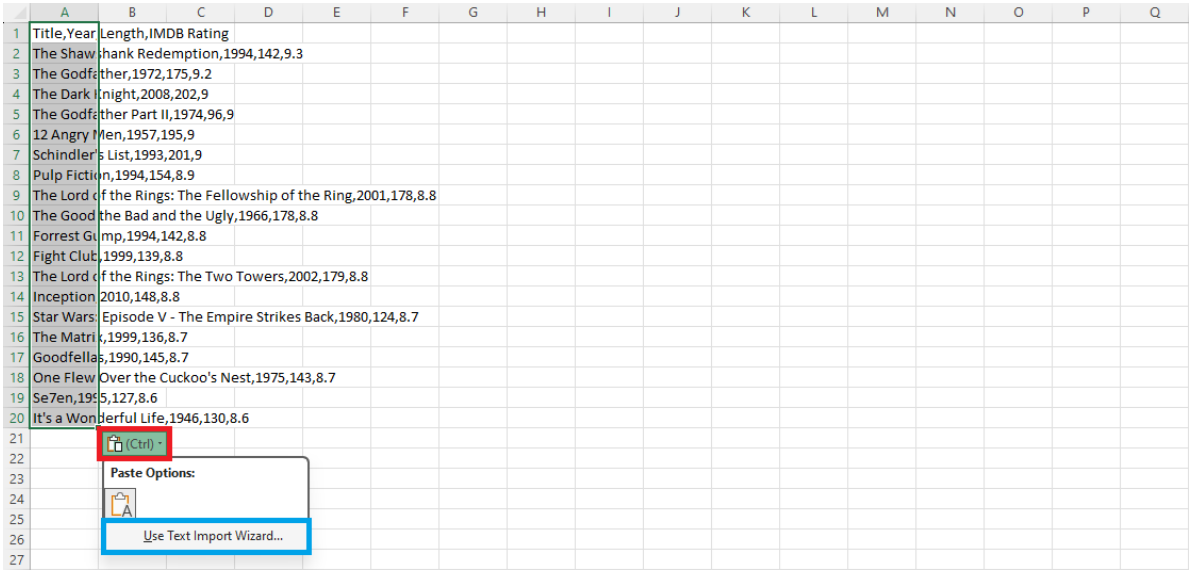


Figure 5: Text File Pasted to Excel

The text import wizard pop-up is shown in Figure 6. Note that in the blue box that the source data is set to `Delimited`, since the entries are separated by commas. You should choose this same format when each field is separated by tabs as well.[2] This will be the default for most cases when you download a text file with data. Click `Next` to move along.

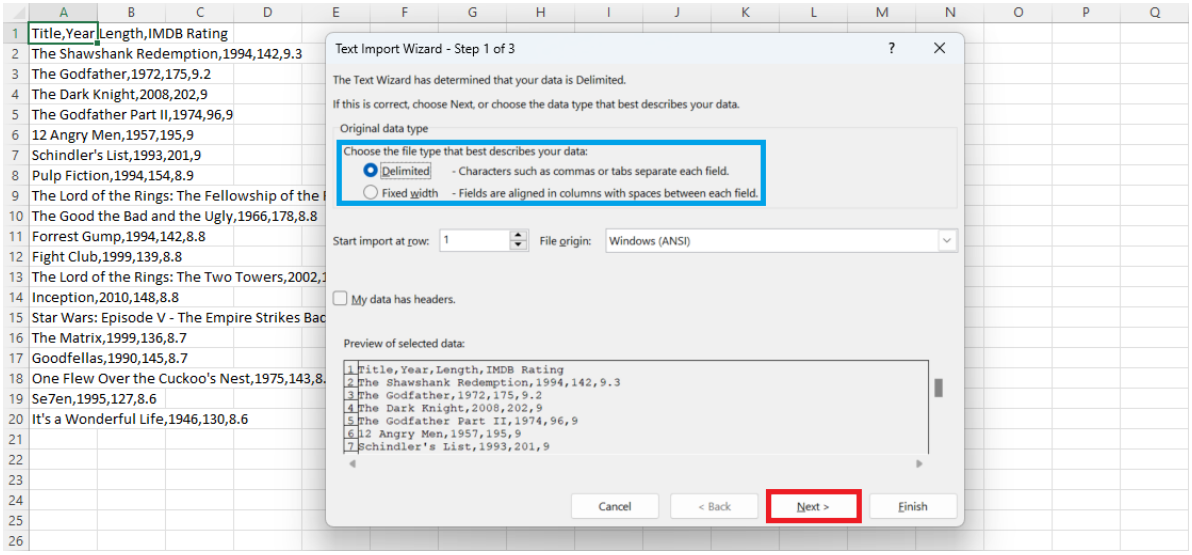


Figure 6: Text Import Wizard Step #1

In this next stage, we can tell Excel that the fields are separated by commas. So, in the red box of Figure 7, deselect `Space`, and select `Comma`. Observe how the preview in the blue box changes depending on the selected delimiters.

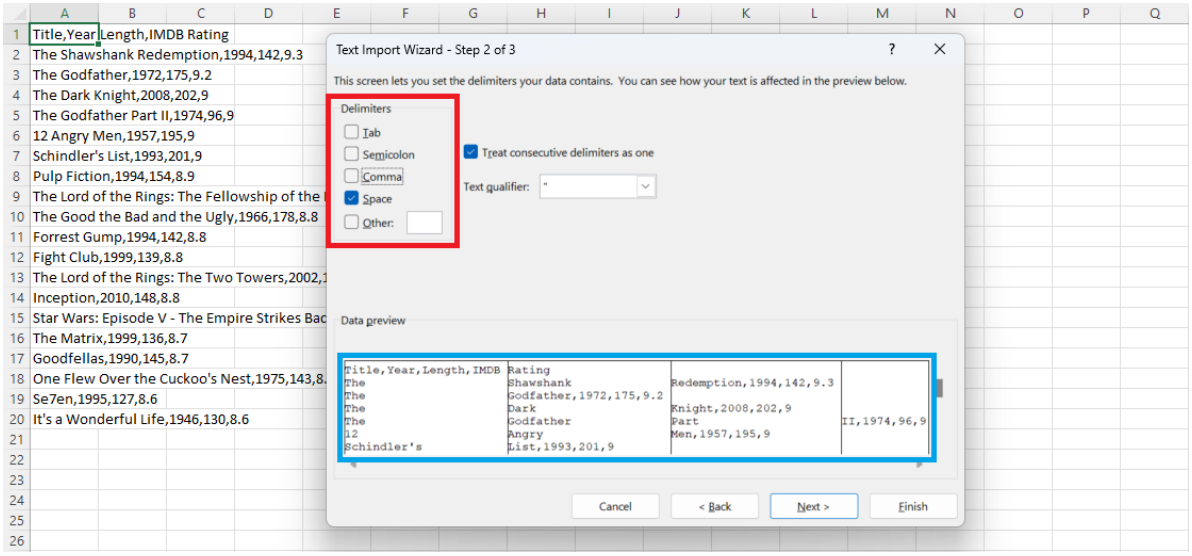


Figure 7: Text Import Wizard Step #2

[2] The `tab` key.

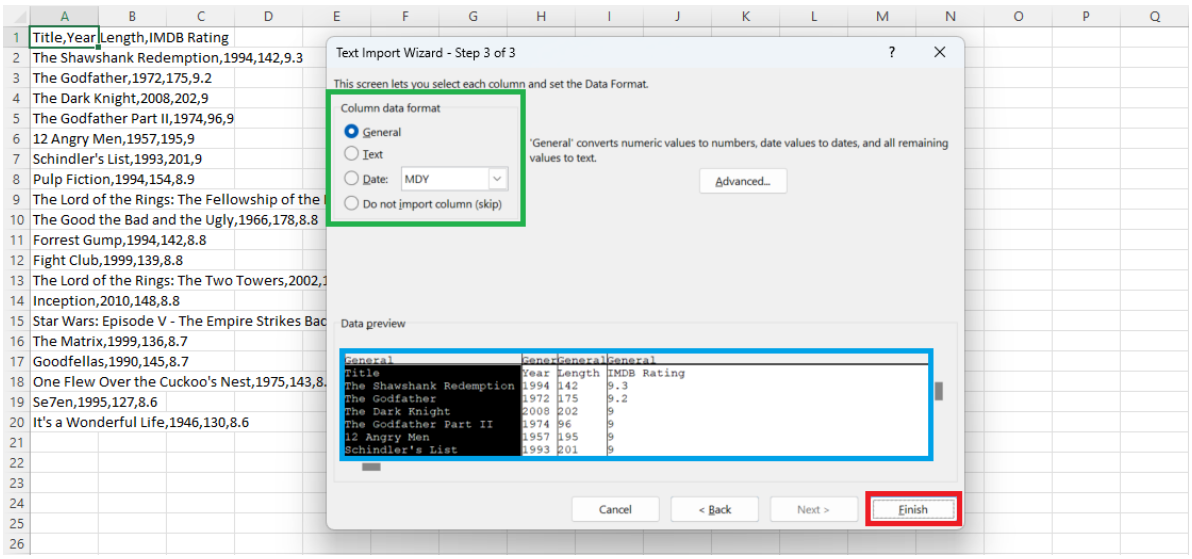


Figure 8: Text Import Wizard Step #3

In the third and last step of the text import wizard, users can set the data format for each column using the options in the green box. Then, check the preview in blue box, and select `Finish`. Figure 9 is the resulting table that is generated using the text import wizard.

	Title	Year	Length	IMDB Rating
2	The Shaws	1994	142	9.3
3	The Godfa	1972	175	9.2
4	The Dark K	2008	202	9
5	The Godfa	1974	96	9
6	12 Angry M	1957	195	9
7	Schindler's	1993	201	9
8	Pulp Fictic	1994	154	8.9
9	The Lord c	2001	178	8.8
10	The Good	1966	178	8.8
11	Forrest Gu	1994	142	8.8
12	Fight Club	1999	139	8.8
13	The Lord c	2002	179	8.8
14	Inception	2010	148	8.8
15	Star Wars:	1980	124	8.7
16	The Matri:	1999	136	8.7
17	Goodfellas	1990	145	8.7
18	One Flew	1975	143	8.7
19	Se7en	1995	127	8.6
20	It's a Won	1946	130	8.6

Figure 9: Table Generated Using Text Import Wizard

Topic 2. Power Query: From Web

Instead of manually importing data, we can use Power Query to import, transform, and clean data. We will first examine how to directly import data from websites. Let us return to the Wikipedia article on secondary education attainment. Navigate to the `Data` tab, and select `From Web`. Type in the URL in the blue box, and then click `OK`.

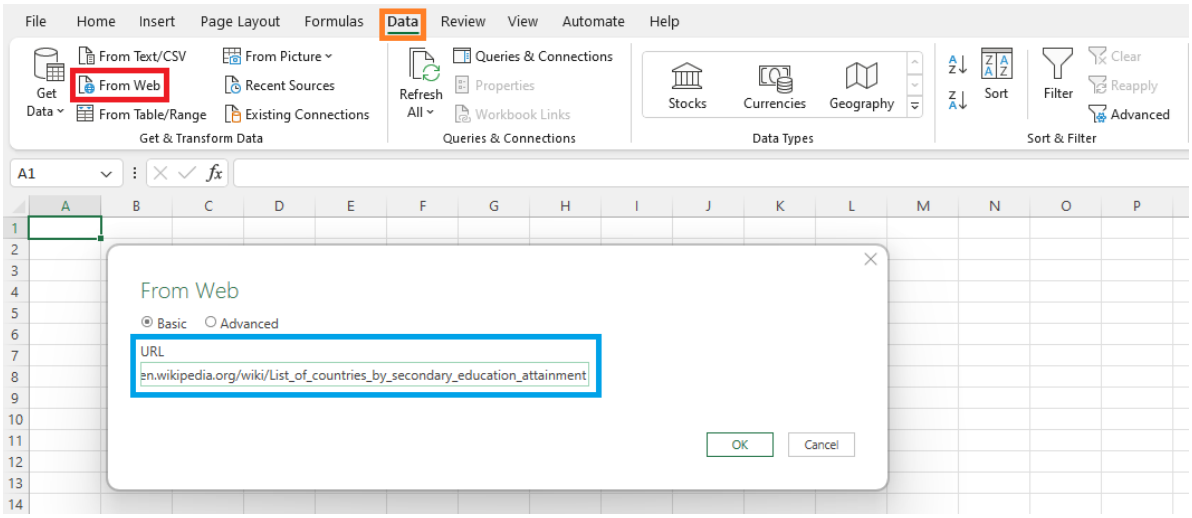


Figure 10: Importing from the Web

Once Excel establishes a link to the webpage, it will open up the `Navigator` page as shown in Figure 11. To the left hand side, you will find the objects included in the webpage. For this purpose, we should select `Table0`, and check out a preview of the table in the blue box. If the preview in the blue box indeed matches the table you wish to import, select `Transform Data`.

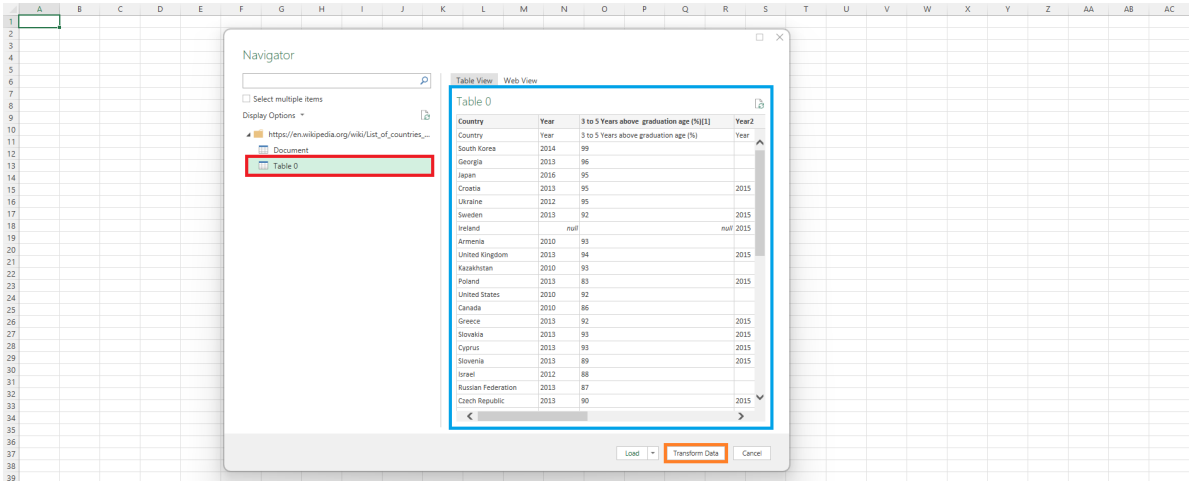


Figure 11: Selecting the Data

A new window named `Power Query Editor` will pop up, which allows the users to edit the data before we import it to Excel. The most basic operations here will be operating on rows and columns.

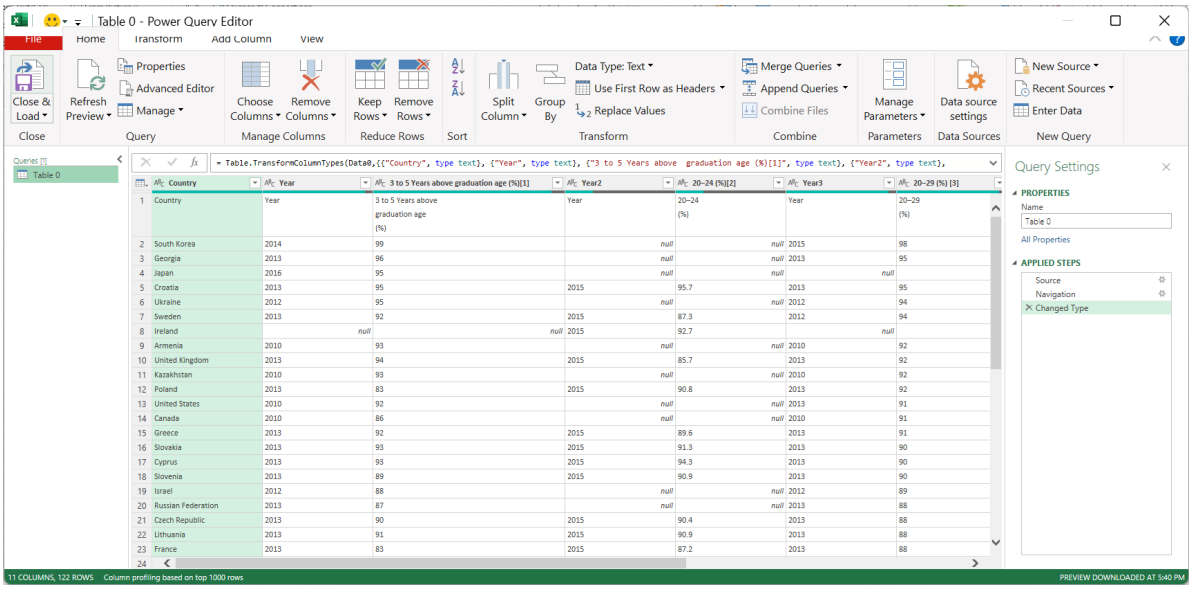


Figure 12: Power Query Editor

Editing Columns

First, to select the columns that are relevant, we can click `Choose Columns` button shown in the red box in Figure 13. Then, you can choose the columns that you would like to have included in the table that will be imported into Excel. You can "uncheck" the items in the blue box that you would not like to have imported.

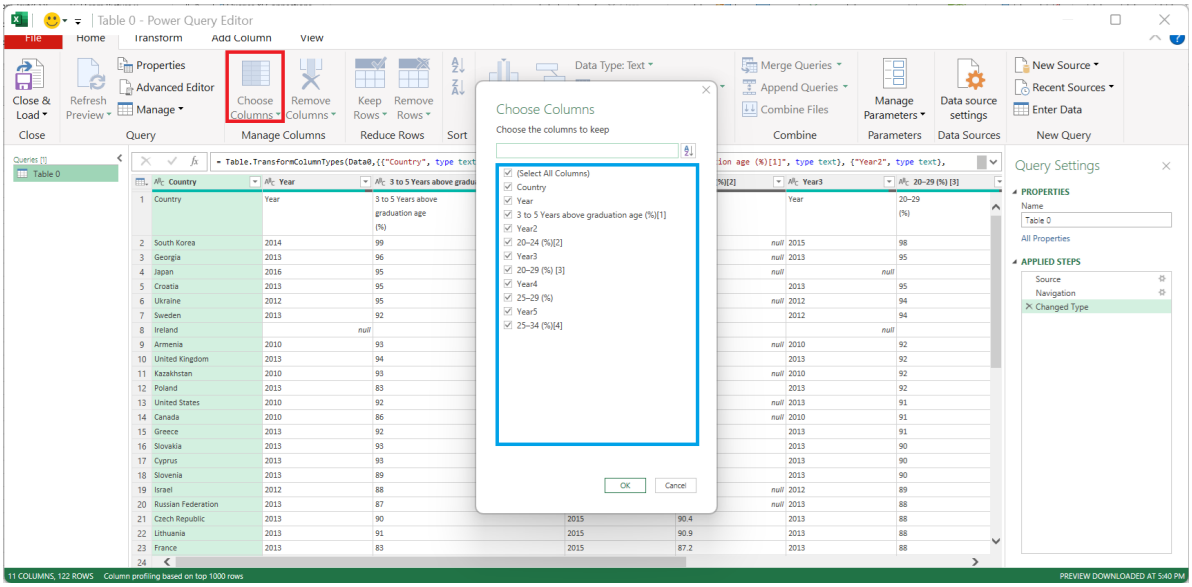


Figure 13: Power Query Editor: Choosing Columns

Editing Rows

See the orange box in Figure 14. We removed the `Year` columns in the previous step, and we can see that the Power Query editor records this change. Choosing the gear icon to the right of each item, you can see the specific changes you made to the imported data. This is a massive improvement over manually editing data.

Next, we can remove rows that are irrelevant for our purposes. For this table, we can see that the variable names are repeated in the first row of the table. We can remove this row by clicking `Remove Rows`, then selecting `Remove Top Rows` in the red box. Remove the first row of this table by typing in `1` in the blue box, and click `OK`.

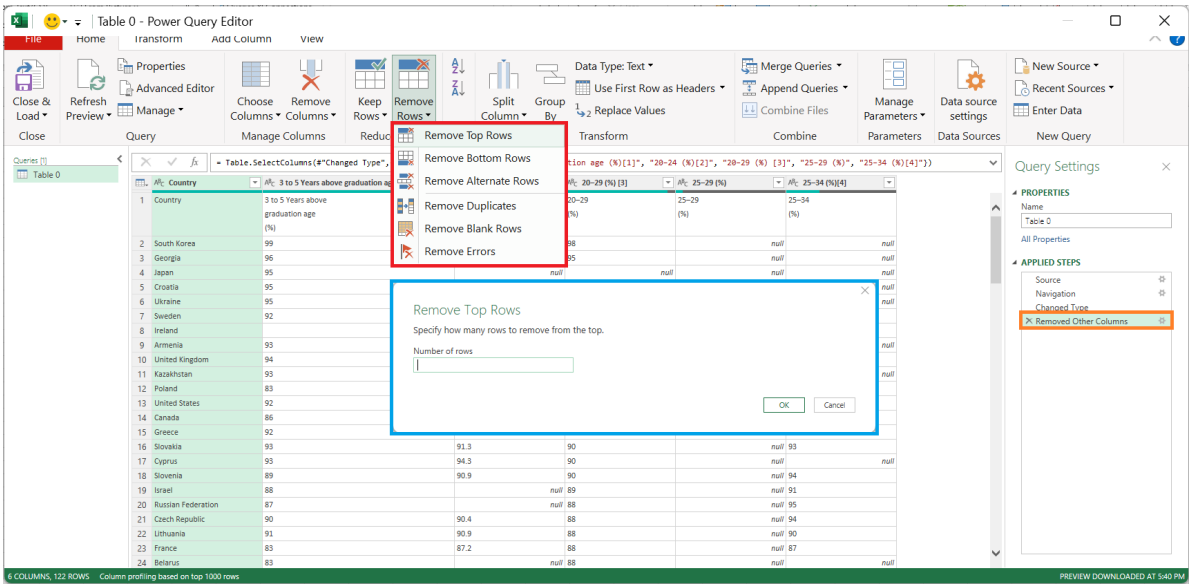


Figure 14: Power Query Editor

The other options included in either `Keep Rows` or `Remove Rows` may prove quite useful, and we encourage that readers try out these options:

- Keep / Remove Top Rows: Keep /Remove only the top N rows from
- Keep / Remove Bottom Rows: Keep / Remove only the bottom N rows from this table.
- Keep / Remove Range of Rows: Specify the number of rows to keep / remove starting at a specific row.
- Keep / Remove Duplicates: Keep / Remove rows containing duplicated values in the currently selected columns.
- Keep / Remove Errors: Keep / Remove only rows containing errors in the currently selected columns.

Loading Data to Excel

Once the table is edited to satisfaction, we can load it to Excel by clicking `Close & Load` in the red box. It is recommended that users check the `Applied Steps` in the blue box before loading the table to Excel.

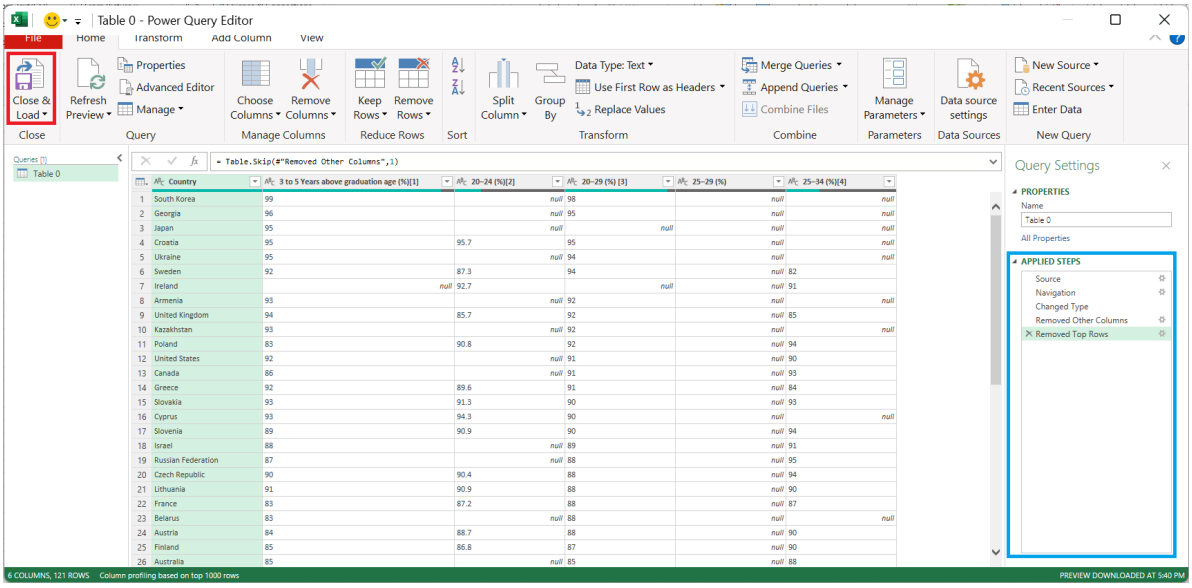


Figure 15: Loading Data to Excel

Figure 16 shows the data imported to Excel. The data will automatically be organized as a table as shown in the red box, and the default name will follow the object name we found in Figure 11.

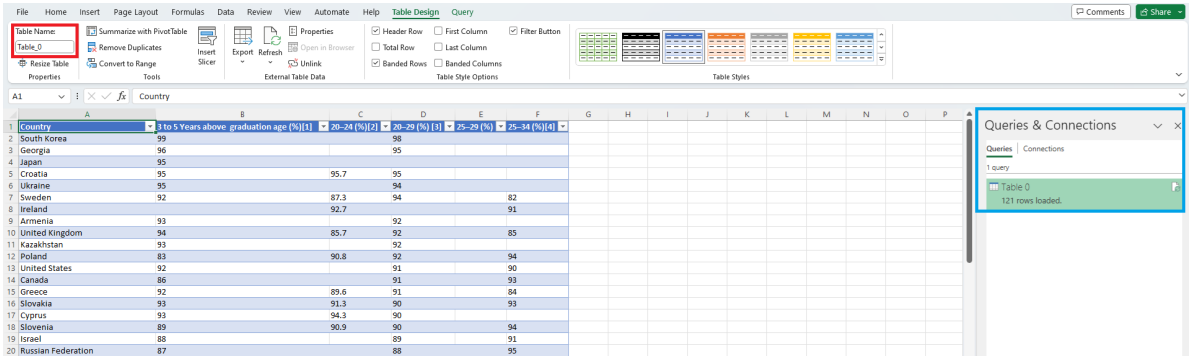


Figure 16: Imported to Excel

Topic 3. Power Query: TXT, Splitting, Duplicating, & Grouping

We can also import text files via Power Query. Let us return to the text file we used previously, `BUSI201-LEC20-txt`. We may import a text file into Excel using the same Power Query framework by selecting `From Text/CSV` under the `Data` tab.

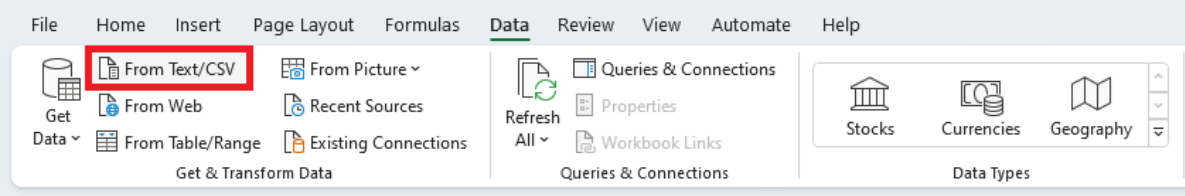
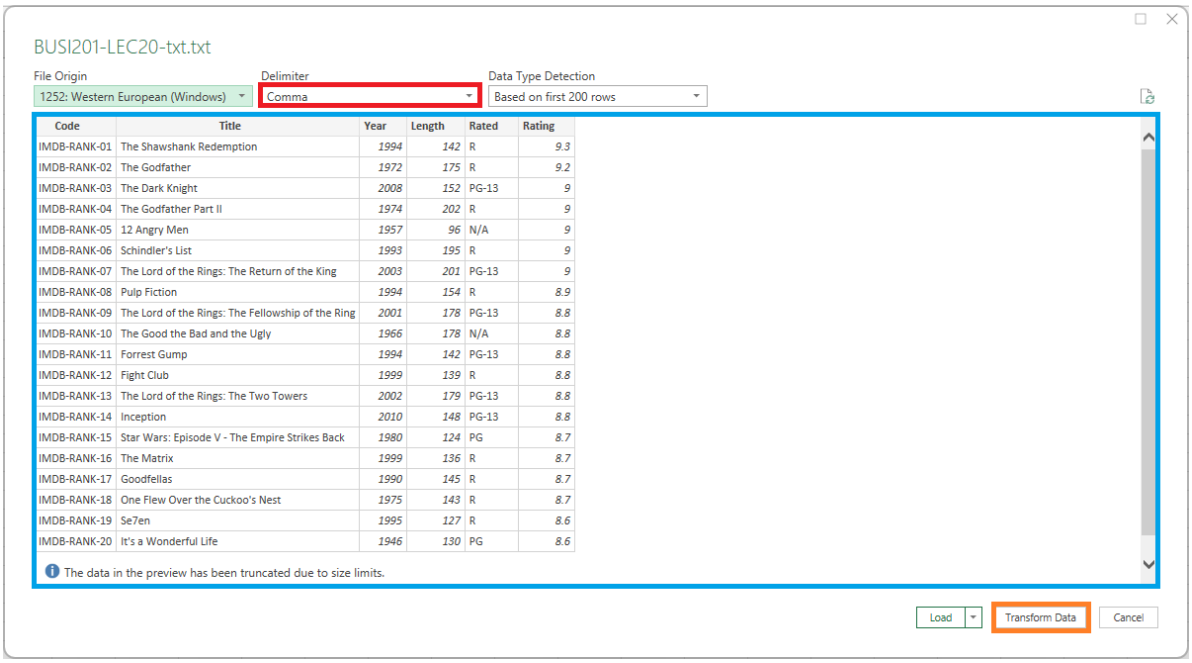


Figure 17: Loading Text / CSV to Excel

The window in Figure 18 should pop up when the text file is correctly selected. Since our text file is separated using commas, the delimiter is correctly set to `Commas`, and the preview in the blue box shows the correct layout for our table. Select `Transform Data`.



Code	Title	Year	Length	Rated	Rating
IMDB-RANK-01	The Shawshank Redemption	1994	142	R	9.3
IMDB-RANK-02	The Godfather	1972	175	R	9.2
IMDB-RANK-03	The Dark Knight	2008	152	PG-13	9
IMDB-RANK-04	The Godfather Part II	1974	202	R	9
IMDB-RANK-05	12 Angry Men	1957	96	N/A	9
IMDB-RANK-06	Schindler's List	1993	195	R	9
IMDB-RANK-07	The Lord of the Rings: The Return of the King	2003	201	PG-13	9
IMDB-RANK-08	Pulp Fiction	1994	154	R	8.9
IMDB-RANK-09	The Lord of the Rings: The Fellowship of the Ring	2001	178	PG-13	8.8
IMDB-RANK-10	The Good the Bad and the Ugly	1966	178	N/A	8.8
IMDB-RANK-11	Forrest Gump	1994	142	PG-13	8.8
IMDB-RANK-12	Fight Club	1999	139	R	8.8
IMDB-RANK-13	The Lord of the Rings: The Two Towers	2002	179	PG-13	8.8
IMDB-RANK-14	Inception	2010	148	PG-13	8.8
IMDB-RANK-15	Star Wars: Episode V - The Empire Strikes Back	1980	124	PG	8.7
IMDB-RANK-16	The Matrix	1999	136	R	8.7
IMDB-RANK-17	Goodfellas	1990	145	R	8.7
IMDB-RANK-18	One Flew Over the Cuckoo's Nest	1975	143	R	8.7
IMDB-RANK-19	Se7en	1995	127	R	8.6
IMDB-RANK-20	It's a Wonderful Life	1946	130	PG	8.6

Figure 18: Loading Text / CSV to Excel

Suppose that you want to create a column that splits the first column into many columns that has information on the ranking, which platform the rankings are based on, and the year when the rankings were taken. This can be acheived by splitting the `Code` column into many parts.

Splitting Columns

Select `Split Column`, and then choose `By Delimiter`, since the `Code` data is linked via short dashes. There are many different methods to split columns, and those methods may be useful depending on the type of data.

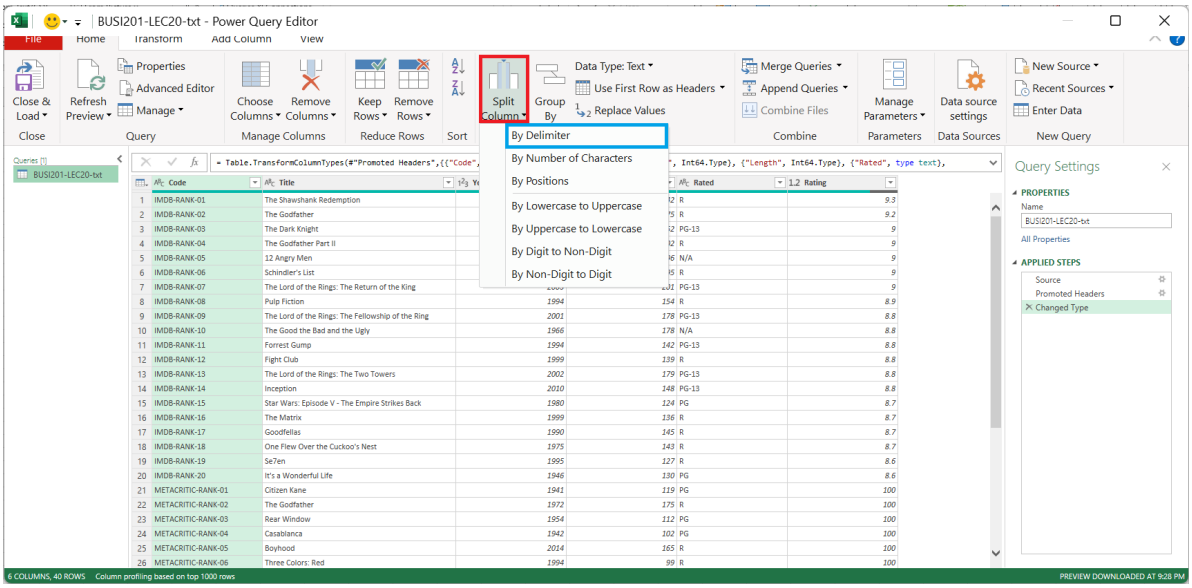


Figure 19: Splitting Columns in Power Query

We can tell Excel which delimiter will be used to split the column in the red box in Figure 20. Set up the options as shown in the red box and blue box to split the column `Code`. Choosing any of the other options in the blue box will allow the user to split the column in various ways.

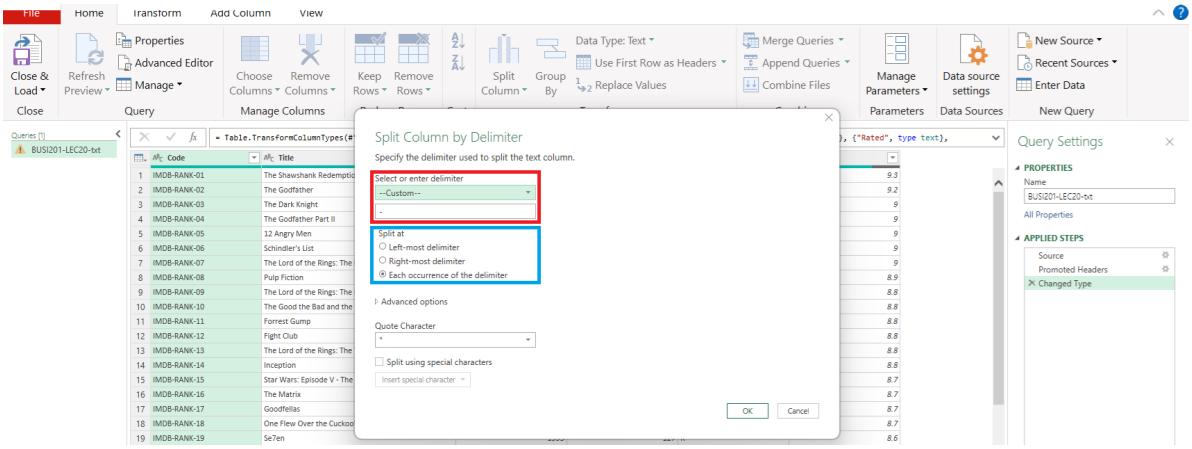


Figure 20: By Delimiter Options

Clicking `OK`, the column `Code` will be split into three parts as shown in the red box in Figure 21. The original column has been split by each occurrence of `-`, creating columns named `Code.1`, `Code.2`, and `Code.3`. You may double click the header containing the names of the columns to rename the columns.

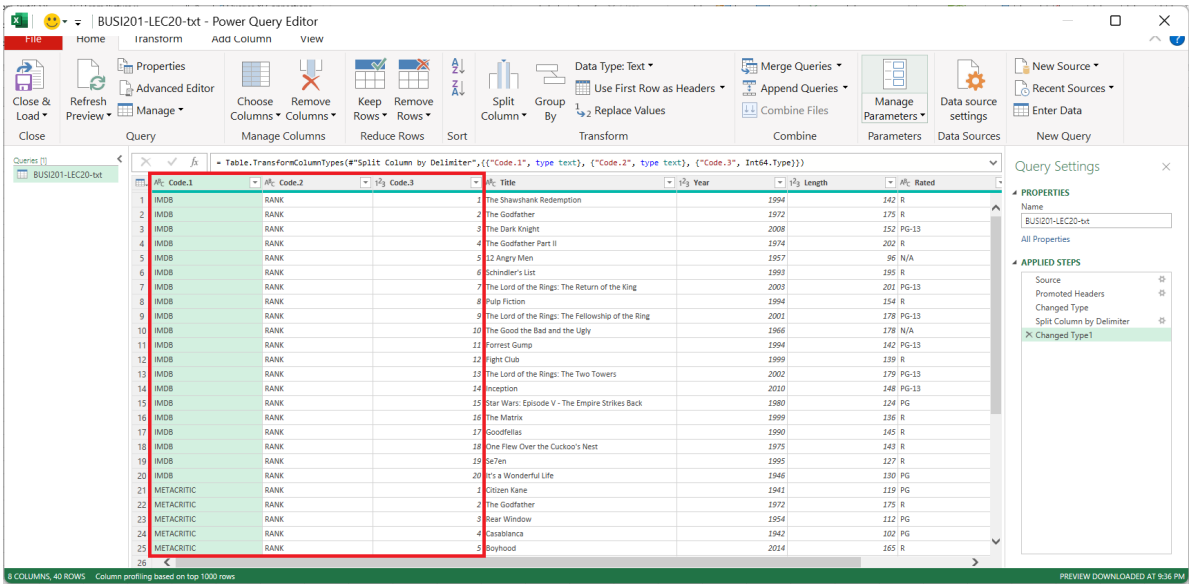


Figure 21: Splitting `Code`

Creating Duplicate Queries

We can create duplicate queries by right clicking the original query in the red box in Figure 22, and clicking `Duplicate`. We will later be using these duplicates to generate new variables, and merge data.

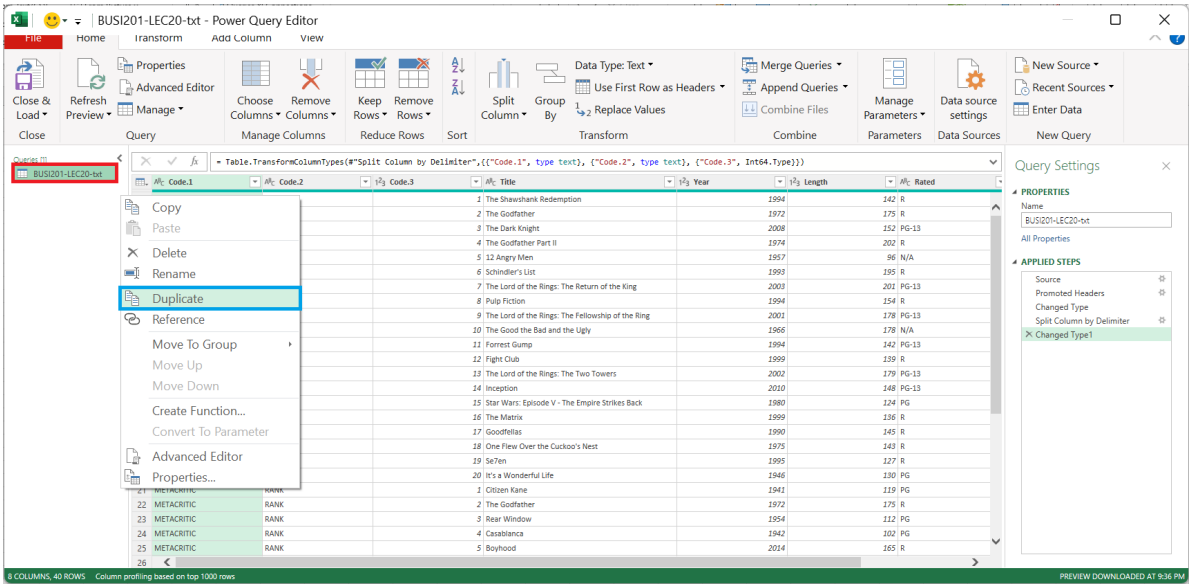


Figure 22: Creating Duplicates

Grouping

We can use the grouping tool to generate new variables based on this data. Suppose we wanted to know how many movies are in the top movies by its ratings; R, PG, PG-13, etc. Click `Group By`, and setting up the options as shown in the blue box in Figure 23.[3]

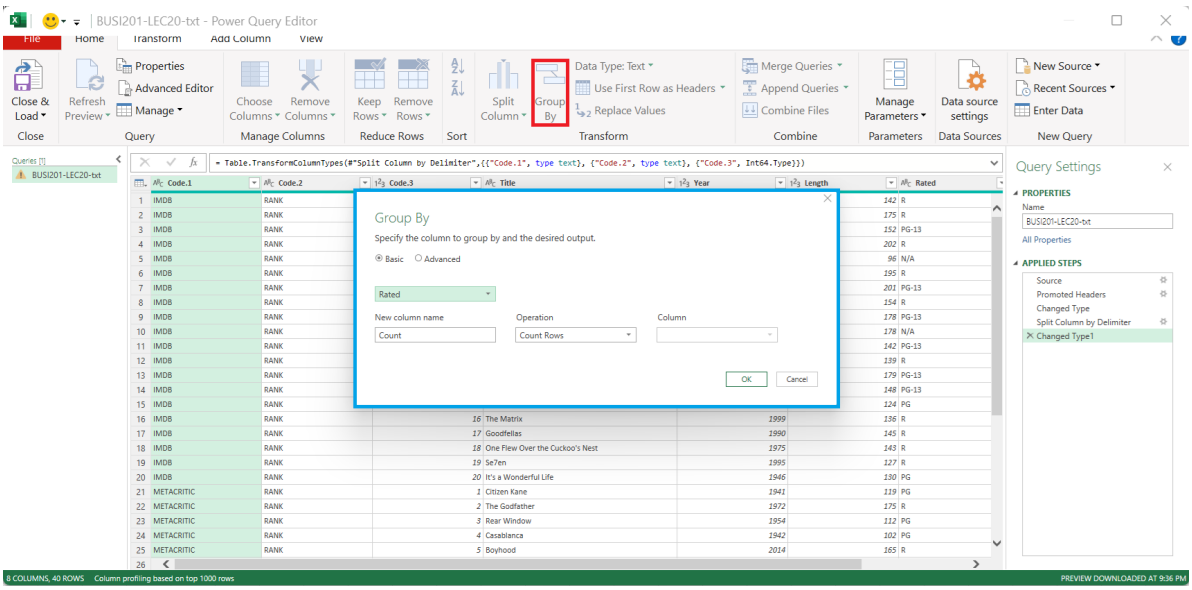


Figure 23: Grouping by `Rated`

Figure 24 shows us how the data will be transformed following the grouping described above. We will later see how we can merge query tables to consolidate multiple data sources.

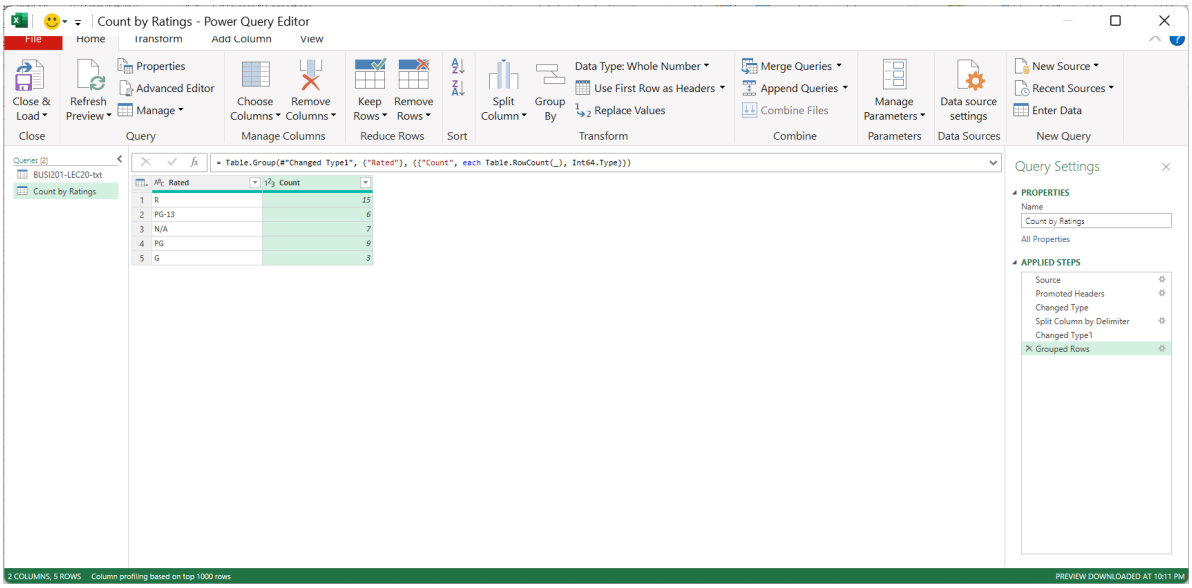
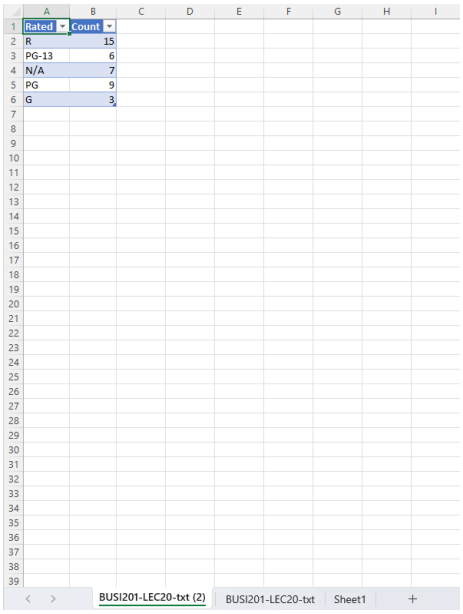


Figure 24: Grouped by `Rated`

[3] At this point, we ignore the issue of duplicates. For instance, `The Godfather` is included in all three lists.

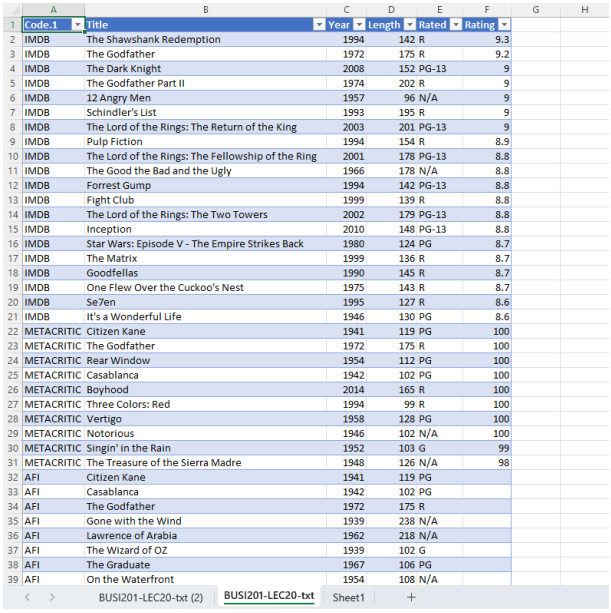
Topic 4. Power Query: Refreshing Data

You might still believe that manually importing data isn't too troublesome, considering the need to learn another tool. However, data imported using Power Query offers a crucial advantage over manual imports – it enables us to refresh the tables in Excel when the source data changes. Import the table using Power Query from **Topic 3** to obtain the two worksheets depicted below.



Rated	Count
R	15
PG-13	6
N/A	7
PG	9
G	3

BUSI201-LEC20-txt (2) | BUSI201-LEC20-txt | Sheet1



Code.1	Title	Year	Length	Rated	Rating
IMDB	The Shawshank Redemption	1994	142	R	9.3
IMDB	The Godfather	1972	175	R	9.2
IMDB	The Dark Knight	2008	152	PG-13	9
IMDB	The Godfather Part II	1974	202	R	9
IMDB	12 Angry Men	1957	96	N/A	9
IMDB	Schindler's List	1993	195	R	9
IMDB	The Lord of the Rings: The Return of the King	2003	201	PG-13	9
IMDB	Pulp Fiction	1994	154	R	8.9
IMDB	The Lord of the Rings: The Fellowship of the Ring	2001	178	PG-13	8.8
IMDB	The Good the Bad and the Ugly	1966	178	N/A	8.8
IMDB	Forrest Gump	1994	142	PG-13	8.8
IMDB	Fight Club	1999	139	R	8.8
IMDB	The Lord of the Rings: The Two Towers	2002	179	PG-13	8.8
IMDB	Inception	2010	148	PG-13	8.8
IMDB	Star Wars: Episode V - The Empire Strikes Back	1980	124	PG	8.7
IMDB	The Matrix	1999	136	R	8.7
IMDB	Goodfellas	1990	145	R	8.7
IMDB	One Flew Over the Cuckoo's Nest	1975	143	R	8.7
IMDB	Se7en	1995	127	R	8.6
IMDB	It's a Wonderful Life	1946	130	PG	8.6
METACRITIC	Citizen Kane	1941	119	PG	100
METACRITIC	The Godfather	1972	175	R	100
METACRITIC	Rear Window	1954	112	PG	100
METACRITIC	Casablanca	1942	102	PG	100
METACRITIC	Boyhood	2014	165	R	100
METACRITIC	Three Colors: Red	1994	99	R	100
METACRITIC	Vertigo	1958	128	PG	100
METACRITIC	Notorious	1946	102	N/A	100
METACRITIC	Singin' in the Rain	1952	103	G	99
METACRITIC	The Treasure of the Sierra Madre	1948	126	N/A	98
AFI	Citizen Kane	1941	119	PG	
AFI	Casablanca	1942	102	PG	
AFI	The Godfather	1972	175	R	
AFI	Gone with the Wind	1939	238	N/A	
AFI	Lawrence of Arabia	1962	218	N/A	
AFI	The Wizard of OZ	1939	102	G	
AFI	The Graduate	1967	106	PG	
AFI	On the Waterfront	1954	108	N/A	

BUSI201-LEC20-txt (2) | BUSI201-LEC20-txt | Sheet1

Suppose now that the rating for the top-rated movie on IMDB, "The Shawshank Redemption," has been updated to 9.5. Additionally, let's assume that "The Dark Knight" has had its PG-13 rating updated to an R rating. Open the file `BUSI201-LEC20-txt`, update the ratings in the source file, and then save the file.

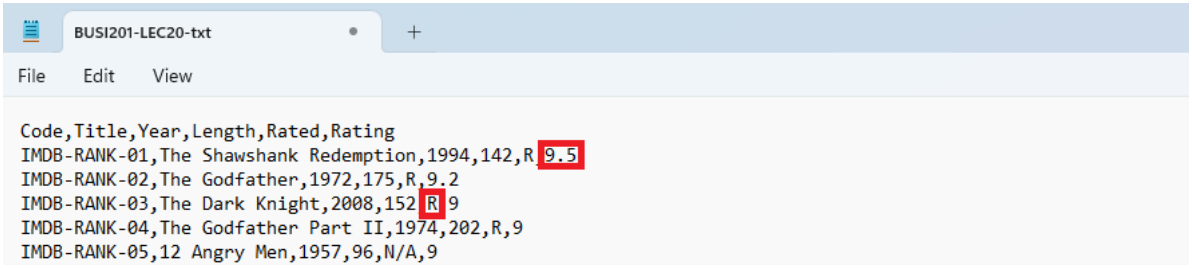


Figure 25: Updating Source Data

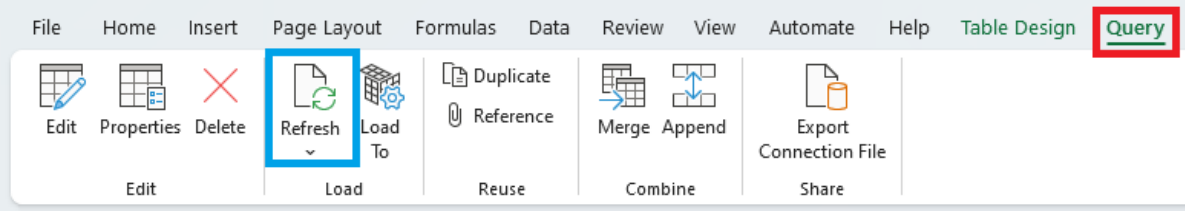


Figure 26: Refreshing Tables

Return to the Excel worksheet `BUSI201-LEC20-txt` containing the main source, click the tables generated using Power Query to make the `Query` tab available. Then, click `Refresh` to update the table.

	A	B	C	D	E	F	G	H
1	Code.1	Title	Year	Length	Rated	Rating		
2	IMDB	The Shawshank Redemption	1994	142	R	9.5		
3	IMDB	The Godfather	1972	175	R	9.2		
4	IMDB	The Dark Knight	2008	152	R	9		
5	IMDB	The Godfather Part II	1974	202	R	9		
6	IMDB	12 Angry Men	1957	96	N/A	9		
7	IMDB	Schindler's List	1993	195	R	9		

Figure 27: First Table Refreshed

See Figure 27 to observe the refreshed table. Follow the same workflow to update the second table, where we grouped the data. Once you refresh the second table in the worksheet `BUSI201-LEC20-txt(2)`, you can see that the grouped table displays the updated distribution as shown in Figure 28.

	A	B	C	D	E
1	Rated	Count			
2	R	16			
3	N/A	7			
4	PG-13	5			
5	PG	9			
6	G	3			
7					

Figure 28: Second Table Refreshed

While we cannot demonstrate here, if information on a webpage is updated, you may update the contents of your Excel spreadsheet by clicking refresh.