Lecture Note #20: Data Analysis Tools Part #2

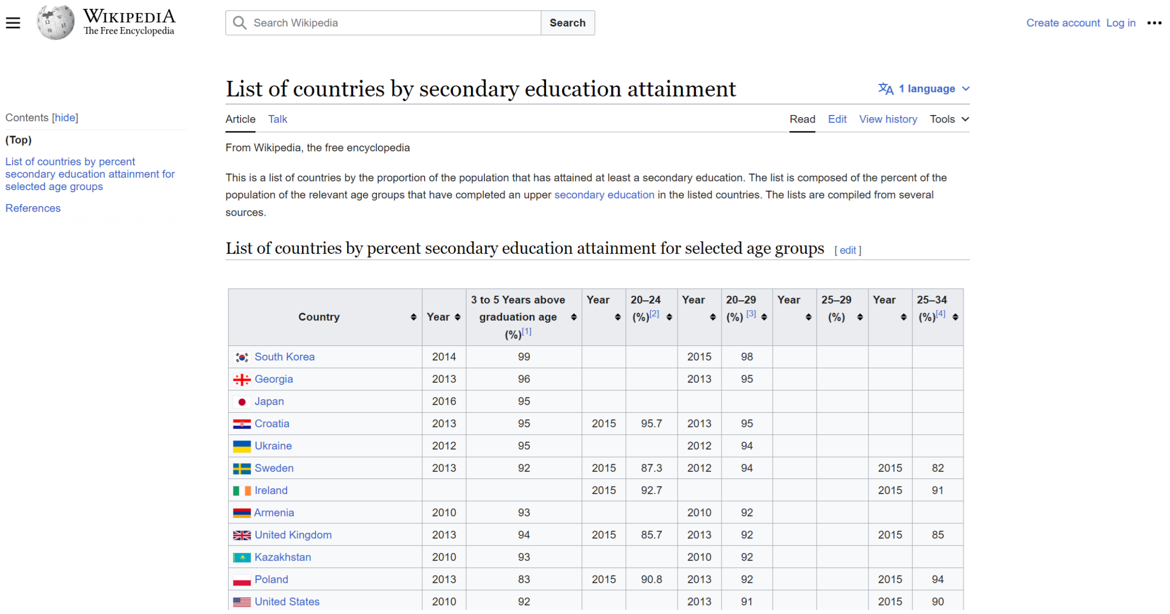
BUSI 201: Business Data Analysis

Fall 2023

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.



List of countries by secondary education attainment

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.

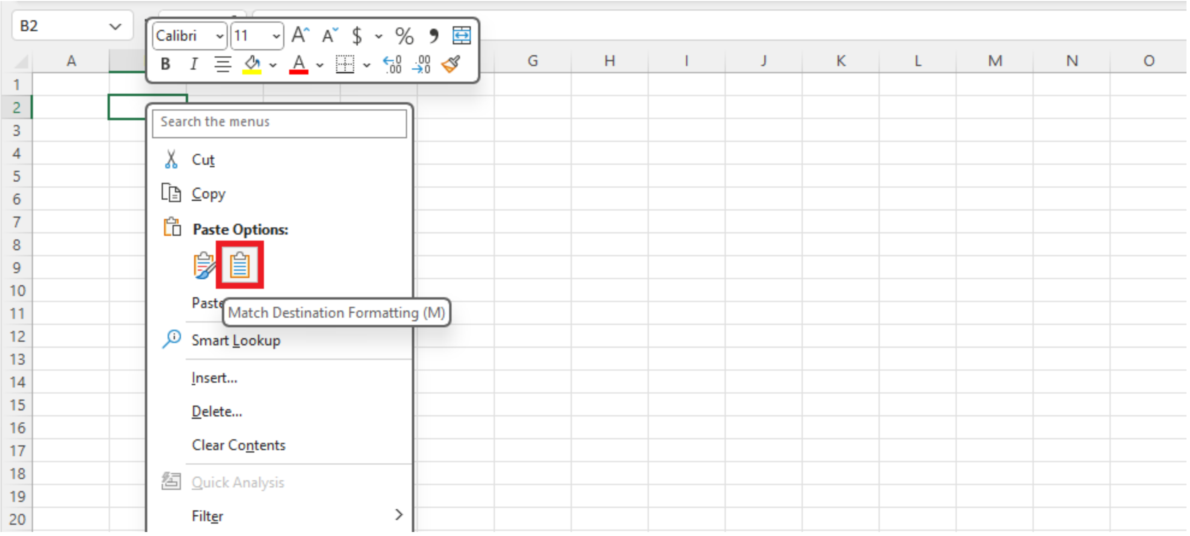
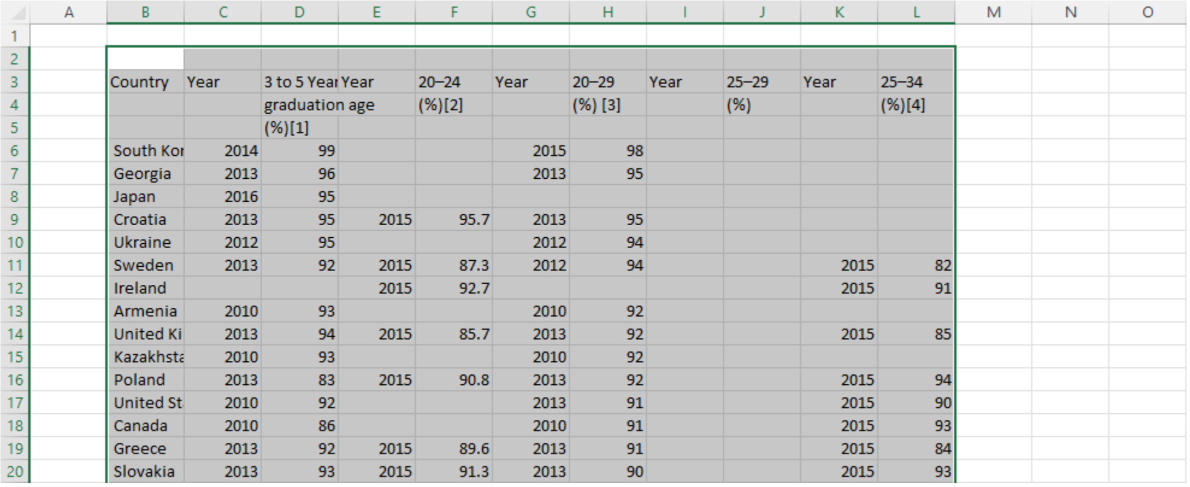


Figure 2: Pasting Options

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.



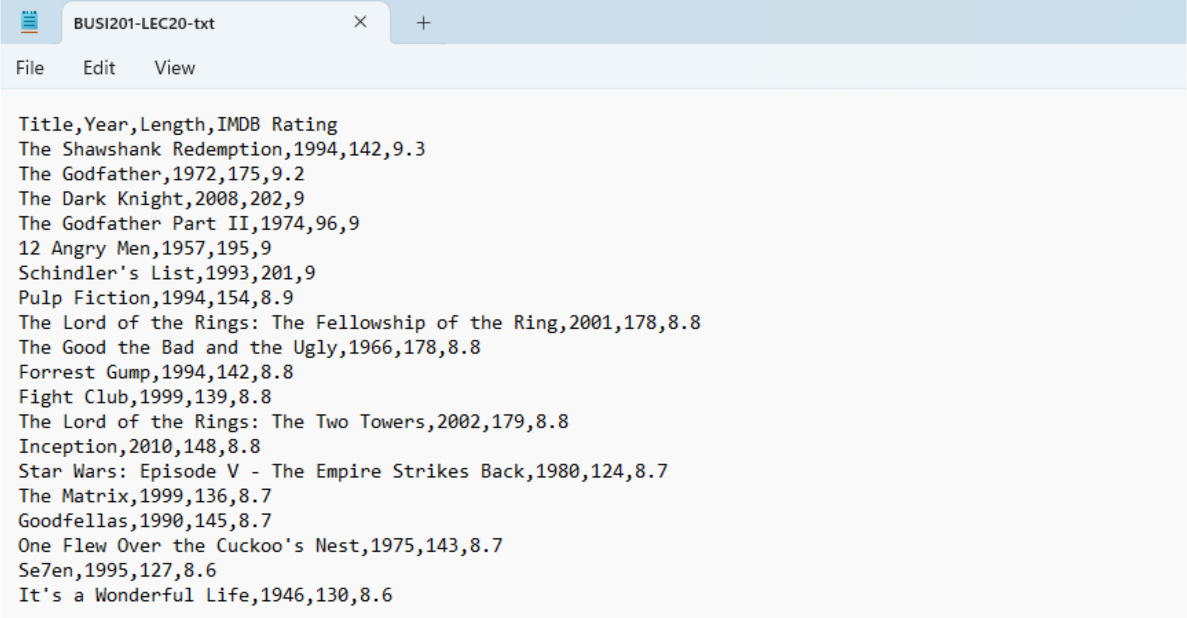
	A	B	C	D	E	F	G	H	I	J	K	L
1												
2												
3		Country	Year	3 to 5 Year	Year	20–24	Year	20–29	Year	25–29	Year	25–34
4				graduation age		(%)[2]		(%) [3]		(%)		(%)[4]
5				(%)[1]								
6		South Kor	2014	99			2015	98				
7		Georgia	2013	96			2013	95				
8		Japan	2016	95								
9		Croatia	2013	95	2015	95.7	2013	95				
10		Ukraine	2012	95			2012	94				
11		Sweden	2013	92	2015	87.3	2012	94			2015	82
12		Ireland			2015	92.7					2015	91
13		Armenia	2010	93			2010	92				
14		United Ki	2013	94	2015	85.7	2013	92			2015	85
15		Kazakhsta	2010	93			2010	92				
16		Poland	2013	83	2015	90.8	2013	92			2015	94
17		United St	2010	92			2013	91			2015	90
18		Canada	2010	86			2010	91			2015	93
19		Greece	2013	92	2015	89.6	2013	91			2015	84
20		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

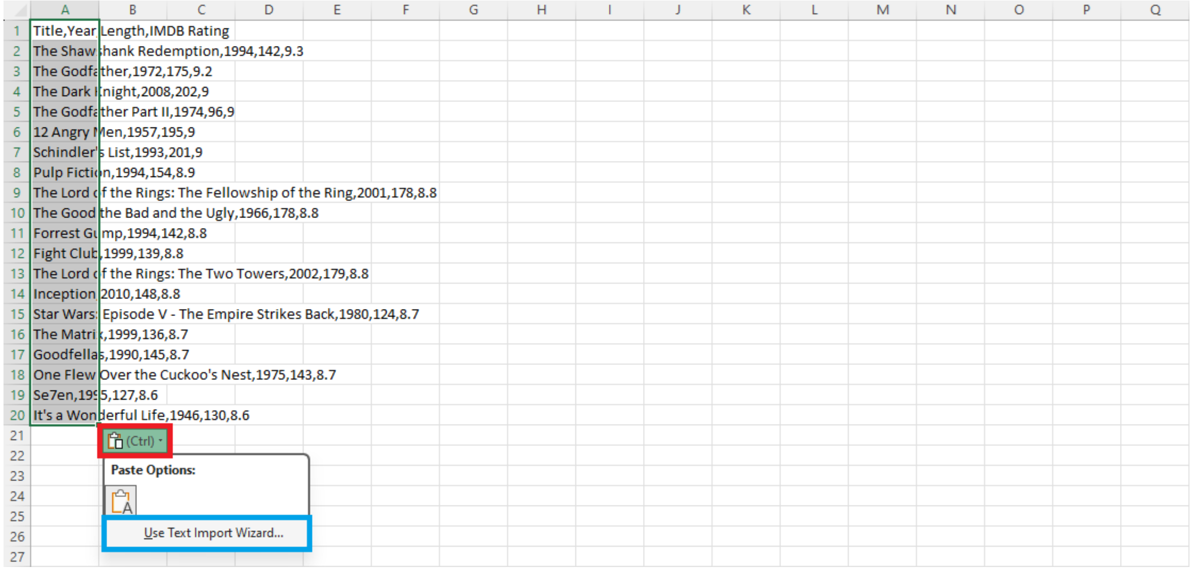
In some cases, you will encounter files in the form of pdfs or txt files. One such example can be found by downloading the BUSI201-LEC20-txt file. This file lists the top 20 movies of all time based on IMDB review scores as of November 2023.



```
Title,Year,Length,IMDB Rating
The Shawshank Redemption,1994,142,9.3
The Godfather,1972,175,9.2
The Dark Knight,2008,202,9
The Godfather Part II,1974,96,9
12 Angry Men,1957,195,9
Schindler's List,1993,201,9
Pulp Fiction,1994,154,8.9
The Lord of the Rings: The Fellowship of the Ring,2001,178,8.8
The Good the Bad and the Ugly,1966,178,8.8
Forrest Gump,1994,142,8.8
Fight Club,1999,139,8.8
The Lord of the Rings: The Two Towers,2002,179,8.8
Inception,2010,148,8.8
Star Wars: Episode V - The Empire Strikes Back,1980,124,8.7
The Matrix,1999,136,8.7
Goodfellas,1990,145,8.7
One Flew Over the Cuckoo's Nest,1975,143,8.7
Se7en,1995,127,8.6
It's a Wonderful Life,1946,130,8.6
```

Figure 4: Data in TXT Format

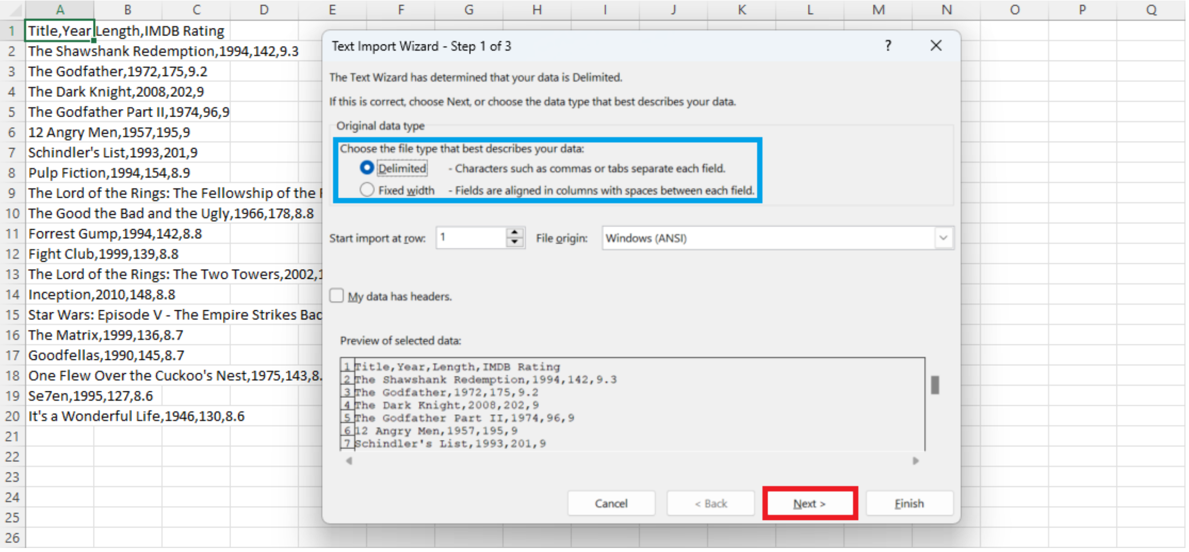
Copy the text file, and paste in into an Excel spreadsheet. The initial result will not be ideal, since each line in the text file will populate a single cell. We must call up the text import wizard by clicking Paste Options, and then Use Text Import Wizard.



```
1   Title,Year,Length,IMDB Rating
2   The Shawshank Redemption,1994,142,9.3
3   The Godfather,1972,175,9.2
4   The Dark Knight,2008,202,9
5   The Godfather Part II,1974,96,9
6   12 Angry Men,1957,195,9
7   Schindler's List,1993,201,9
8   Pulp Fiction,1994,154,8.9
9   The Lord of the Rings: The Fellowship of the Ring,2001,178,8.8
10  The Good the Bad and the Ugly,1966,178,8.8
11  Forrest Gump,1994,142,8.8
12  Fight Club,1999,139,8.8
13  The Lord of the Rings: The Two Towers,2002,179,8.8
14  Inception,2010,148,8.8
15  Star Wars: Episode V - The Empire Strikes Back,1980,124,8.7
16  The Matrix,1999,136,8.7
17  Goodfellas,1990,145,8.7
18  One Flew Over the Cuckoo's Nest,1975,143,8.7
19  Se7en,1995,127,8.6
20  It's a Wonderful Life,1946,130,8.6
21
22
23
24
25
26
27
```

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.



```
1  Title,Year,Length,IMDB Rating
2  The Shawshank Redemption,1994,142,9.3
3  The Godfather,1972,175,9.2
4  The Dark Knight,2008,202,9
5  The Godfather Part II,1974,96,9
6  12 Angry Men,1957,195,9
7  Schindler's List,1993,201,9
8  Pulp Fiction,1994,154,8.9
9  The Lord of the Rings: The Fellowship of the I
10 The Good the Bad and the Ugly,1966,178,8.8
11 Forrest Gump,1994,142,8.8
12 Fight Club,1999,139,8.8
13 The Lord of the Rings: The Two Towers,2002,1
14 Inception,2010,148,8.8
15 Star Wars: Episode V - The Empire Strikes Bac
16 The Matrix,1999,136,8.7
17 Goodfellas,1990,145,8.7
18 One Flew Over the Cuckoo's Nest,1975,143,8.
19 Se7en,1995,127,8.6
20 It's a Wonderful Life,1946,130,8.6
```

Text Import Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

● Delimited - Characters such as commas or tabs separate each field.

○ Fixed width - Fields are aligned in columns with spaces between each field.

Start import at row: 1 File origin: Windows (ANSI)

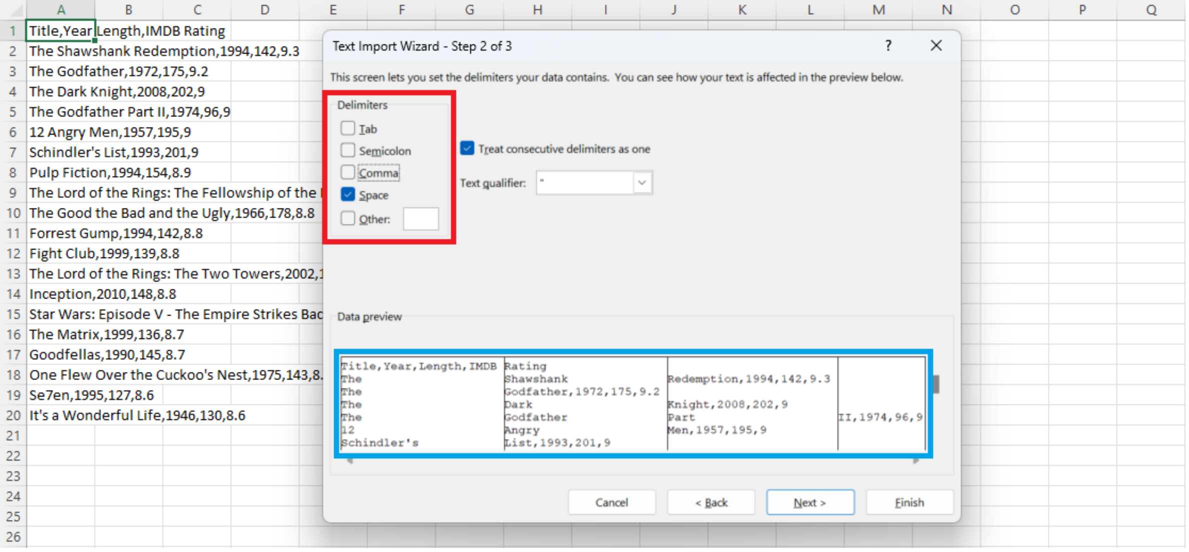
□ My data has headers.

Preview of selected data:

```
1 Title,Year,Length,IMDB Rating
2 The Shawshank Redemption,1994,142,9.3
3 The Godfather,1972,175,9.2
4 The Dark Knight,2008,202,9
5 The Godfather Part II,1974,96,9
6 12 Angry Men,1957,195,9
7 Schindler's List,1993,201,9
```

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.



```
1  Title,Year,Length,IMDB Rating
2  The Shawshank Redemption,1994,142,9.3
3  The Godfather,1972,175,9.2
4  The Dark Knight,2008,202,9
5  The Godfather Part II,1974,96,9
6  12 Angry Men,1957,195,9
7  Schindler's List,1993,201,9
8  Pulp Fiction,1994,154,8.9
9  The Lord of the Rings: The Fellowship of the I
10 The Good the Bad and the Ugly,1966,178,8.8
11 Forrest Gump,1994,142,8.8
12 Fight Club,1999,139,8.8
13 The Lord of the Rings: The Two Towers,2002,1
14 Inception,2010,148,8.8
15 Star Wars: Episode V - The Empire Strikes Bac
16 The Matrix,1999,136,8.7
17 Goodfellas,1990,145,8.7
18 One Flew Over the Cuckoo's Nest,1975,143,8.
19 Se7en,1995,127,8.6
20 It's a Wonderful Life,1946,130,8.6
```

Text Import Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

□ Tab

□ Semicolon

□ Comma

☑ Space

□ Other:

☑ Treat consecutive delimiters as one

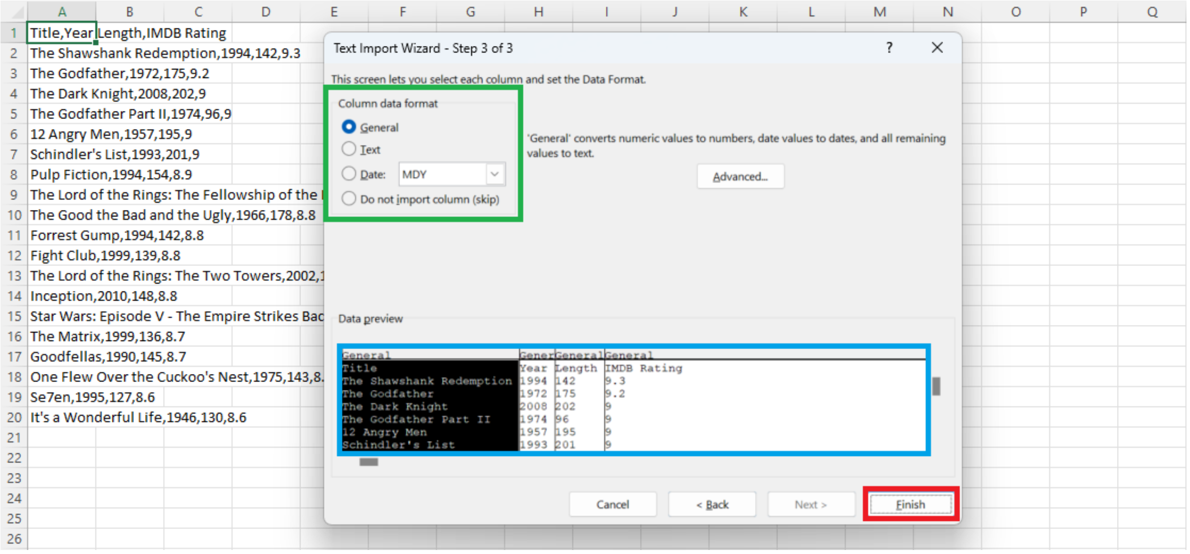
Text qualifier: "

Data preview

```
Title,Year,Length,IMDB  Rating
The                     Shawshank                Redemption,1994,142,9.3
The                     Godfather,1972,175,9.2
The                     Dark                     Knight,2008,202,9
The                     Godfather                Part                     II,1974,96,9
12                      Angry                    Men,1957,195,9
Schindler's             List,1993,201,9
```

Figure 7: Text Import Wizard Step #2

[2] The tab key.



	A
1	Title,Year,Length,IMDB Rating
2	The Shawshank Redemption,1994,142,9.3
3	The Godfather,1972,175,9.2
4	The Dark Knight,2008,202,9
5	The Godfather Part II,1974,96,9
6	12 Angry Men,1957,195,9
7	Schindler's List,1993,201,9
8	Pulp Fiction,1994,154,8.9
9	The Lord of the Rings: The Fellowship of the I
10	The Good the Bad and the Ugly,1966,178,8.8
11	Forrest Gump,1994,142,8.8
12	Fight Club,1999,139,8.8
13	The Lord of the Rings: The Two Towers,2002,1
14	Inception,2010,148,8.8
15	Star Wars: Episode V - The Empire Strikes Bac
16	The Matrix,1999,136,8.7
17	Goodfellas,1990,145,8.7
18	One Flew Over the Cuckoo's Nest,1975,143,8.
19	Se7en,1995,127,8.6
20	It's a Wonderful Life,1946,130,8.6

Text Import Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format

● General
○ Text
○ Date: MDY
○ Do not import column (skip)

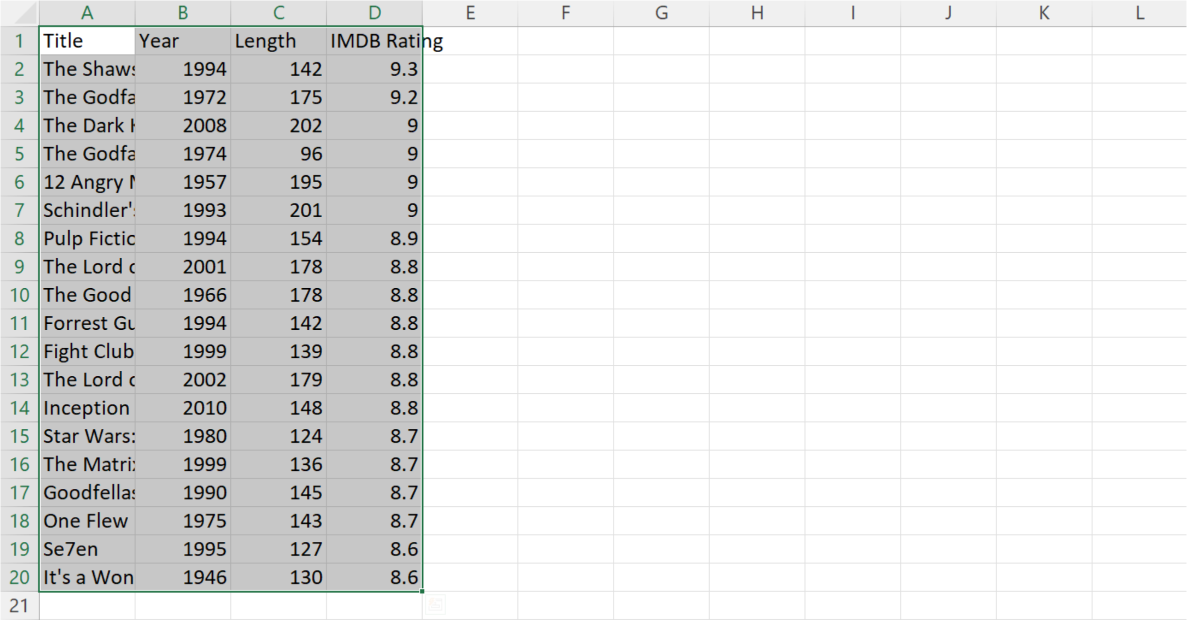
'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Data preview

```
General                 GenerGeneralGeneral
Title                   Year Length IMDB Rating
The Shawshank Redemption1994 142    9.3
The Godfather           1972 175    9.2
The Dark Knight         2008 202    9
The Godfather Part II   1974 96     9
12 Angry Men            1957 195    9
Schindler's List        1993 201    9
```

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.

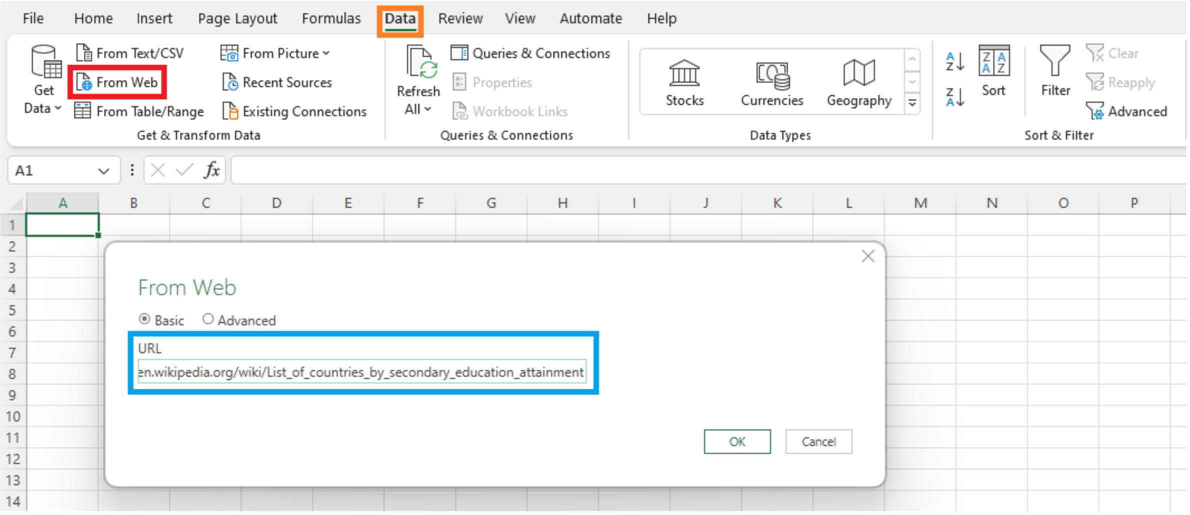


	A	B	C	D
1	Title	Year	Length	IMDB Rating
2	The Shaws	1994	142	9.3
3	The Godfa	1972	175	9.2
4	The Dark I	2008	202	9
5	The Godfa	1974	96	9
6	12 Angry I	1957	195	9
7	Schindler'	1993	201	9
8	Pulp Fictic	1994	154	8.9
9	The Lord c	2001	178	8.8
10	The Good	1966	178	8.8
11	Forrest Gu	1994	142	8.8
12	Fight Club	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	Goodfella:	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.



From Web

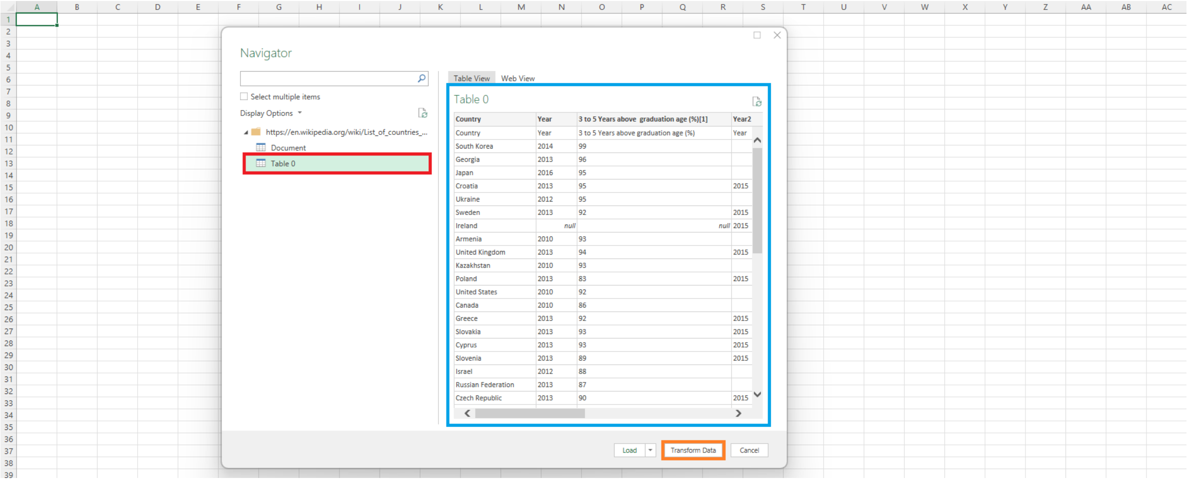
● Basic ○ Advanced

URL

en.wikipedia.org/wiki/List_of_countries_by_secondary_education_attainment

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.



Navigator

□ Select multiple items

Display Options

https://en.wikipedia.org/wiki/List_of_countries_...

Document

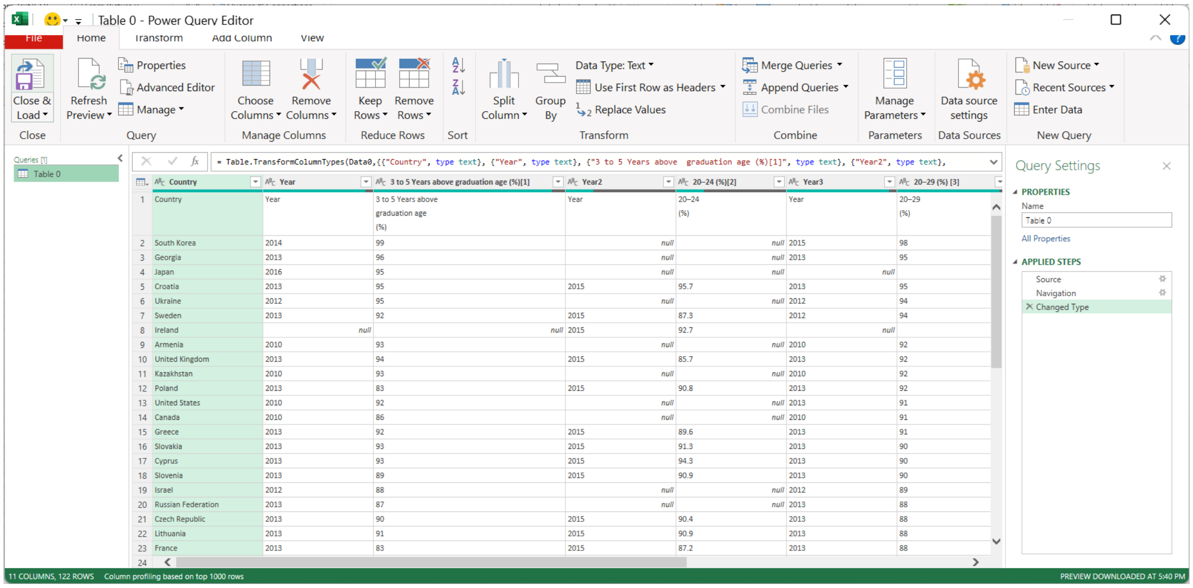
Table 0

Table 0

Country	Year	3 to 5 Years above graduation age (%)[1]	Year2
Country	Year	3 to 5 Years above graduation age (%)	Year
South Korea	2014	99	
Georgia	2013	96	
Japan	2016	95	
Croatia	2013	95	2015
Ukraine	2012	95	
Sweden	2013	92	2015
Ireland	null	null	2015
Armenia	2010	93	
United Kingdom	2013	94	2015
Kazakhstan	2010	93	
Poland	2013	83	2015
United States	2010	92	
Canada	2010	86	
Greece	2013	92	2015
Slovakia	2013	93	2015
Cyprus	2013	93	2015
Slovenia	2013	89	2015
Israel	2012	88	
Russian Federation	2013	87	
Czech Republic	2013	90	2015

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.



```
= Table.TransformColumnTypes(Data0,{{"Country", type text}, {"Year", type text}, {"3 to 5 Years above  graduation age (%)[1]", type text}, {"Year2", type text},
```

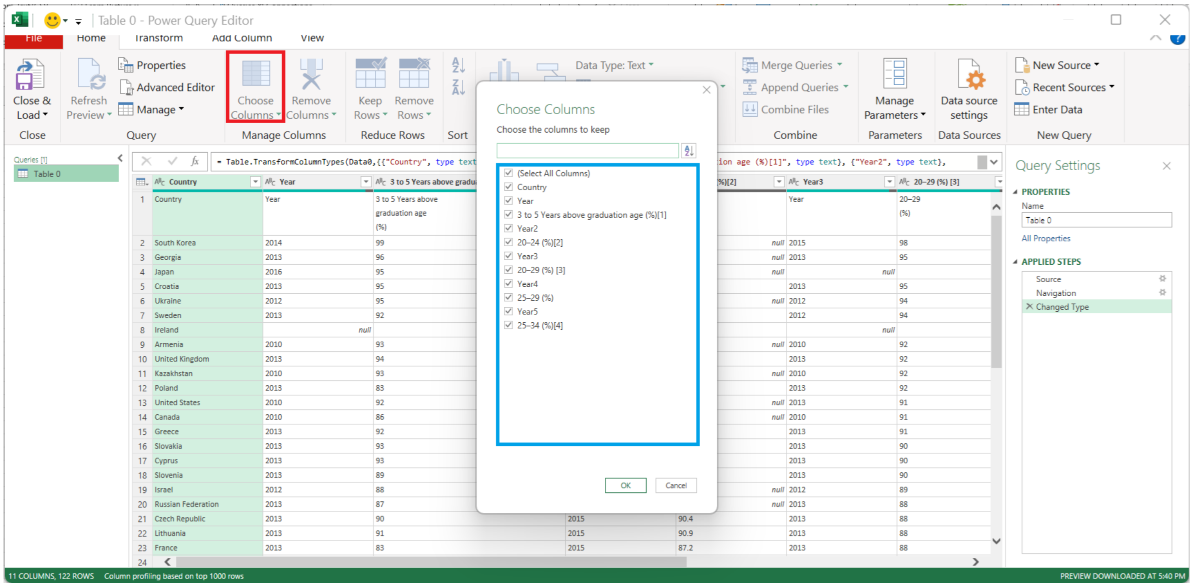
	Country	Year	3 to 5 Years above graduation age (%)[1]	Year2	20–24 (%)[2]	Year3	20–29 (%) [3]
1	Country	Year	3 to 5 Years above graduation age (%)	Year	20–24 (%)	Year	20–29 (%)
2	South Korea	2014	99	null	null	2015	98
3	Georgia	2013	96	null	null	2013	95
4	Japan	2016	95	null	null	null	
5	Croatia	2013	95	2015	95.7	2013	95
6	Ukraine	2012	95	null	null	2012	94
7	Sweden	2013	92	2015	87.3	2012	94
8	Ireland	null	null	2015	92.7	null	
9	Armenia	2010	93	null	null	2010	92
10	United Kingdom	2013	94	2015	85.7	2013	92
11	Kazakhstan	2010	93	null	null	2010	92
12	Poland	2013	83	2015	90.8	2013	92
13	United States	2010	92	null	null	2013	91
14	Canada	2010	86	null	null	2010	91
15	Greece	2013	92	2015	89.6	2013	91
16	Slovakia	2013	93	2015	91.3	2013	90
17	Cyprus	2013	93	2015	94.3	2013	90
18	Slovenia	2013	89	2015	90.9	2013	90
19	Israel	2012	88	null	null	2012	89
20	Russian Federation	2013	87	null	null	2013	88
21	Czech Republic	2013	90	2015	90.4	2013	88
22	Lithuania	2013	91	2015	90.9	2013	88
23	France	2013	83	2015	87.2	2013	88
24							

11 COLUMNS, 122 ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED AT 5:40 PM

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.



Choose Columns

Choose the columns to keep

- ☑ (Select All Columns)
- ☑ Country
- ☑ Year
- ☑ 3 to 5 Years above graduation age (%)[1]
- ☑ Year2
- ☑ 20–24 (%)[2]
- ☑ Year3
- ☑ 20–29 (%) [3]
- ☑ Year4
- ☑ 25–29 (%)
- ☑ Year5
- ☑ 25–34 (%)[4]

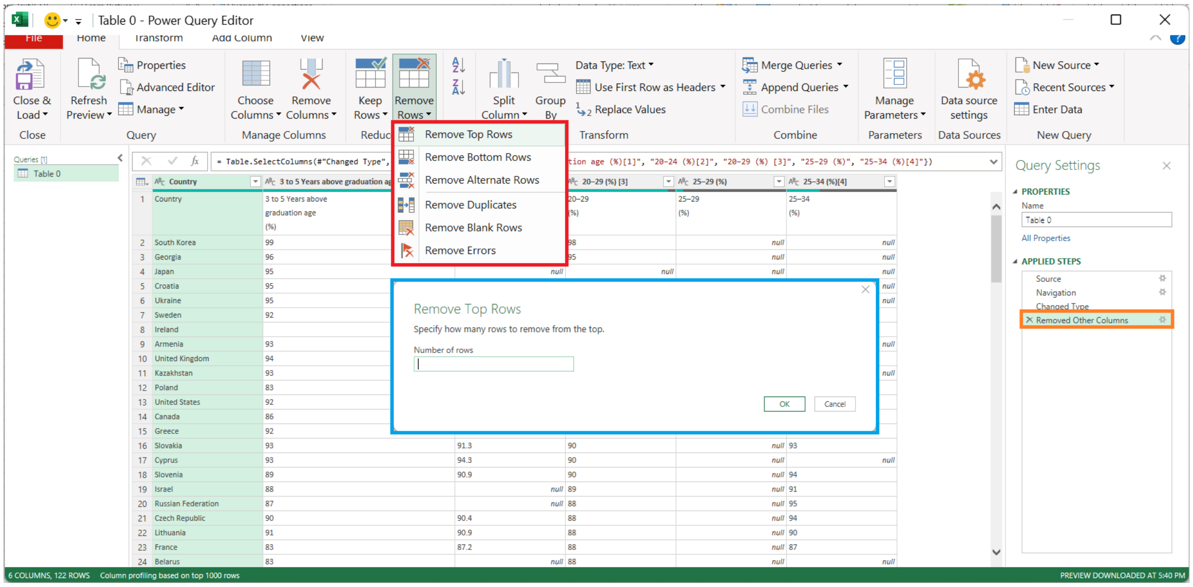
11 COLUMNS, 122 ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED AT 5:40 PM

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.



```
= Table.SelectColumns(#"Changed Type",
tion age (%)[1]", "20-24 (%)[2]", "20-29 (%) [3]", "25-29 (%)", "25-34 (%)[4]"})
```

Name: Table 0

Remove Top Rows

Specify how many rows to remove from the top.

Number of rows

	Country	3 to 5 Years above graduation ag		20–29 (%) [3]	25–29 (%)	25–34 (%)[4]
1	Country	3 to 5 Years above graduation age (%)		20–29 (%)	25–29 (%)	25–34 (%)
2	South Korea	99		98	null	null
3	Georgia	96		95	null	null
4	Japan	95	null	null	null	null
5	Croatia	95				null
6	Ukraine	95				null
7	Sweden	92				
8	Ireland					
9	Armenia	93				null
10	United Kingdom	94				
11	Kazakhstan	93				null
12	Poland	83				
13	United States	92				
14	Canada	86				
15	Greece	92				
16	Slovakia	93	91.3	90	null	93
17	Cyprus	93	94.3	90	null	null
18	Slovenia	89	90.9	90	null	94
19	Israel	88	null	89	null	91
20	Russian Federation	87	null	88	null	95
21	Czech Republic	90	90.4	88	null	94
22	Lithuania	91	90.9	88	null	90
23	France	83	87.2	88	null	87
24	Belarus	83	null	88	null	null

6 COLUMNS, 122 ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 5:40 PM

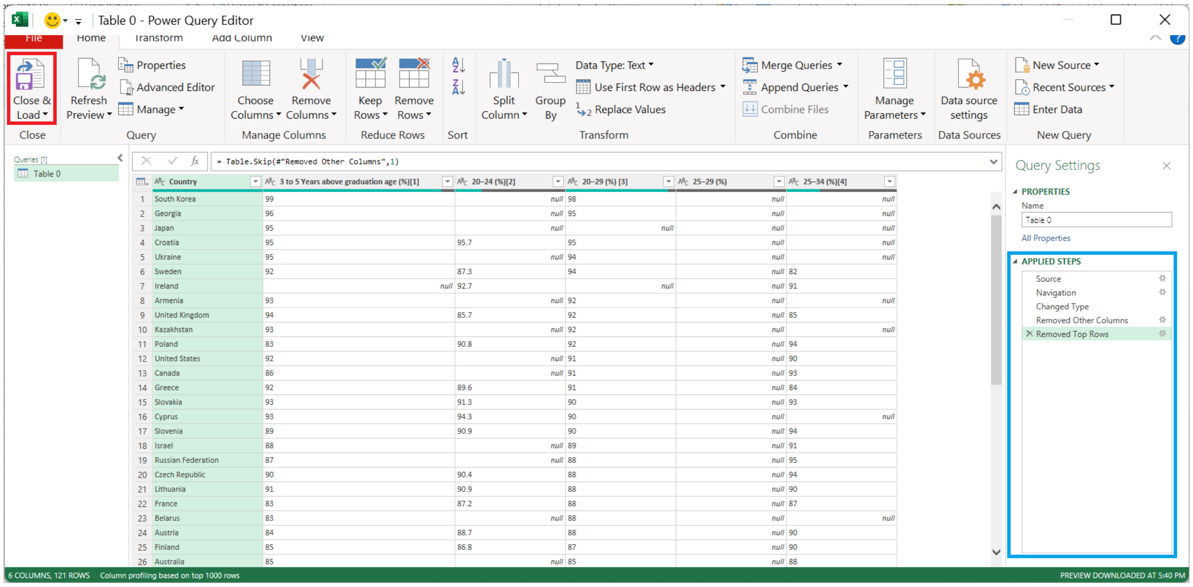
Figure 14: Power Query Editor

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.



```
= Table.Skip(#"Removed Other Columns",1)
```

	Country	3 to 5 Years above graduation age (%)[1]	20–24 (%)[2]	20–29 (%) [3]	25–29 (%)	25–34 (%)[4]
1	South Korea	99	null	98	null	null
2	Georgia	96	null	95	null	null
3	Japan	95	null	null	null	null
4	Croatia	95	95.7	95	null	null
5	Ukraine	95	null	94	null	null
6	Sweden	92	87.3	94	null	82
7	Ireland	null	92.7	null	null	91
8	Armenia	93	null	92	null	null
9	United Kingdom	94	85.7	92	null	85
10	Kazakhstan	93	null	92	null	null
11	Poland	83	90.8	92	null	94
12	United States	92	null	91	null	90
13	Canada	86	null	91	null	93
14	Greece	92	89.6	91	null	84
15	Slovakia	93	91.3	90	null	93
16	Cyprus	93	94.3	90	null	null
17	Slovenia	89	90.9	90	null	94
18	Israel	88	null	89	null	91
19	Russian Federation	87	null	88	null	95
20	Czech Republic	90	90.4	88	null	94
21	Lithuania	91	90.9	88	null	90
22	France	83	87.2	88	null	87
23	Belarus	83	null	88	null	null
24	Austria	84	88.7	88	null	90
25	Finland	85	86.8	87	null	90
26	Australia	85	null	85	null	88

Query Settings

PROPERTIES

Name: Table 0

All Properties

APPLIED STEPS

- Source
- Navigation
- Changed Type
- Removed Other Columns
- Removed Top Rows

6 COLUMNS, 121 ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED AT 5:40 PM

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.

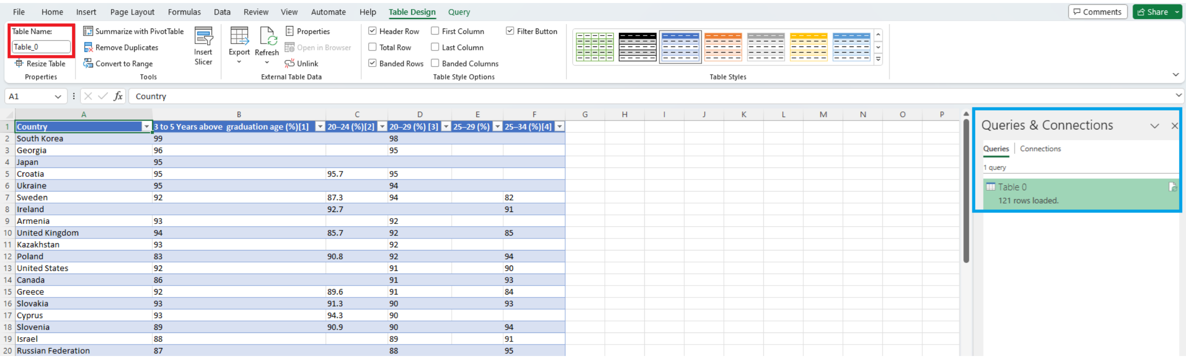


Table Name: Table_0

☑ Header Row ☐ First Column ☑ Filter Button
☐ Total Row ☐ Last Column
☑ Banded Rows ☐ Banded Columns

A1 Country

	A	B	C	D	E	F
1	Country	3 to 5 Years above graduation age (%)[1]	20–24 (%)[2]	20–29 (%) [3]	25–29 (%)	25–34 (%)[4]
2	South Korea	99		98		
3	Georgia	96		95		
4	Japan	95				
5	Croatia	95	95.7	95		
6	Ukraine	95		94		
7	Sweden	92	87.3	94		82
8	Ireland		92.7			91
9	Armenia	93		92		
10	United Kingdom	94	85.7	92		85
11	Kazakhstan	93		92		
12	Poland	83	90.8	92		94
13	United States	92		91		90
14	Canada	86		91		93
15	Greece	92	89.6	91		84
16	Slovakia	93	91.3	90		93
17	Cyprus	93	94.3	90		
18	Slovenia	89	90.9	90		94
19	Israel	88		89		91
20	Russian Federation	87		88		95

Queries & Connections

Queries | Connections

1 query

Table 0
121 rows loaded.

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.

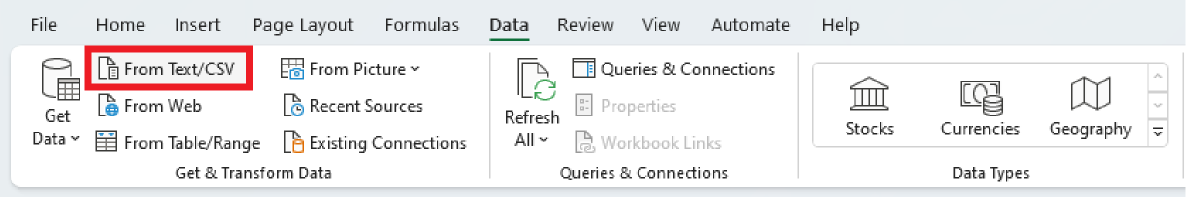
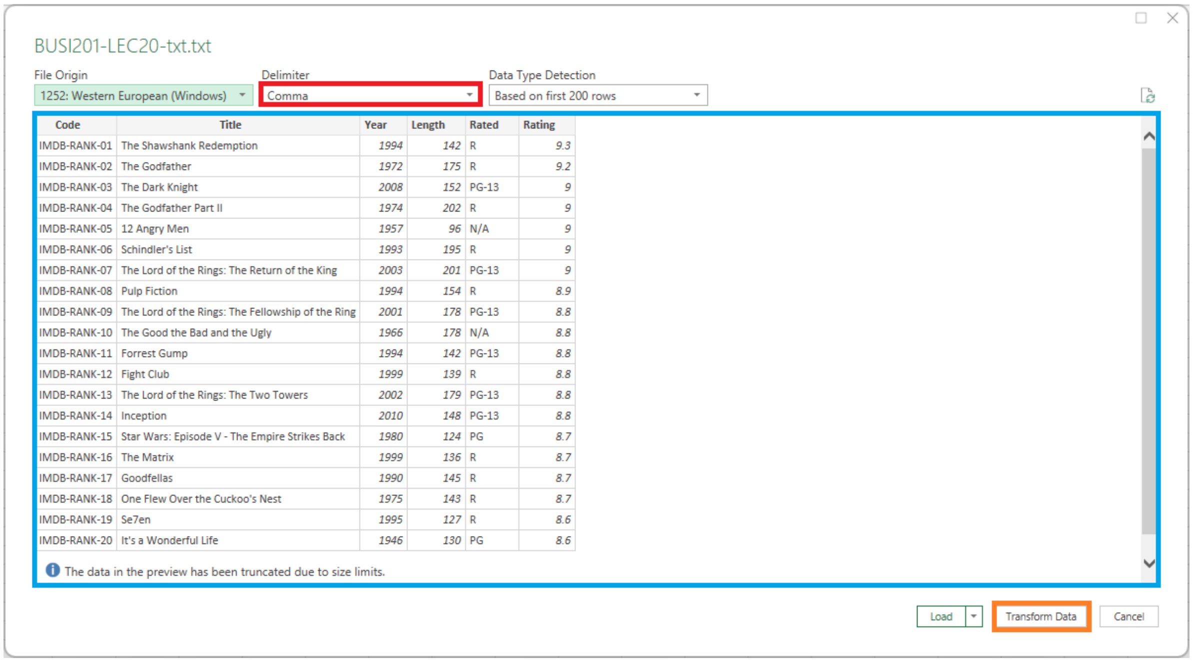


Figure 17: Loading Text / CSV to Excel

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`.



BUSI201-LEC20-txt.txt

File Origin: 1252: Western European (Windows)
Delimiter: Comma
Data Type Detection: Based on first 200 rows

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

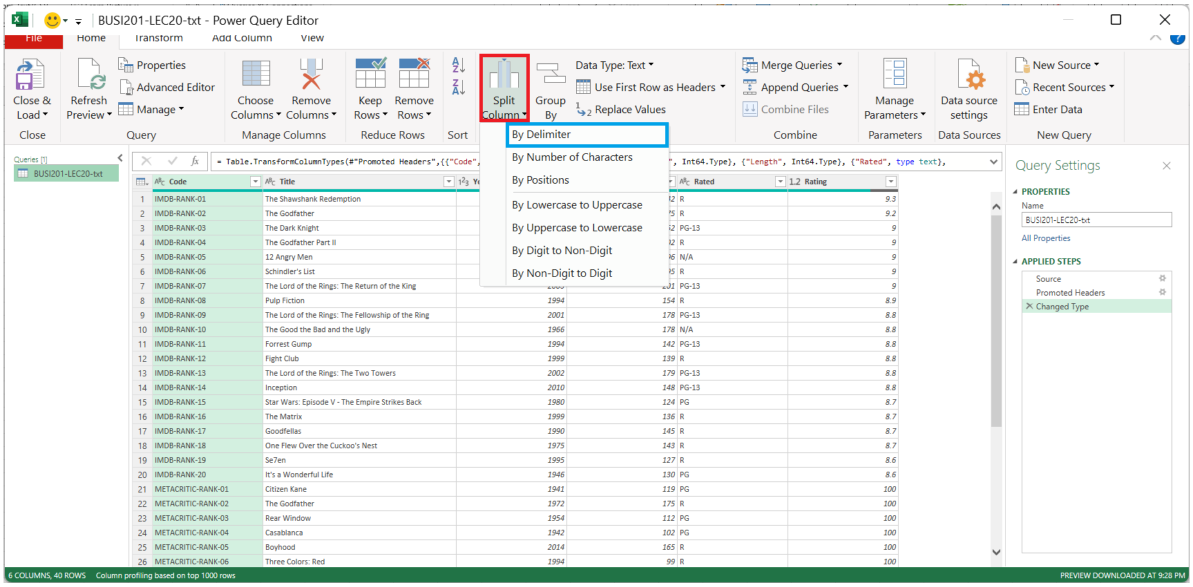
ⓘ The data in the preview has been truncated due to size limits.

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.

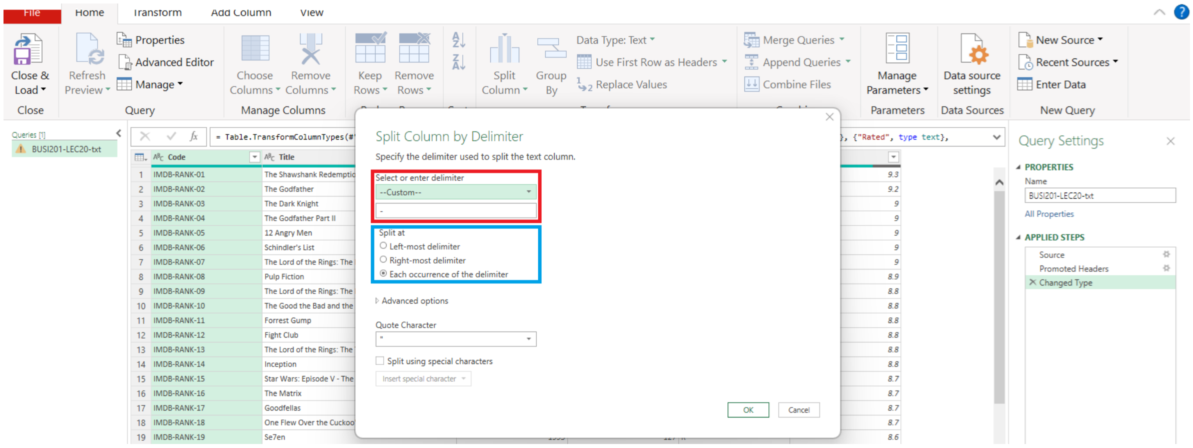


```
= Table.TransformColumnTypes(#"Promoted Headers",{{"Code", ... ", Int64.Type}, {"Length", Int64.Type}, {"Rated", type text},
```

	Code	Title	Ye		Rated	Rating
1	IMDB-RANK-01	The Shawshank Redemption		42	R	9.3
2	IMDB-RANK-02	The Godfather		75	R	9.2
3	IMDB-RANK-03	The Dark Knight		52	PG-13	9
4	IMDB-RANK-04	The Godfather Part II		02	R	9
5	IMDB-RANK-05	12 Angry Men		96	N/A	9
6	IMDB-RANK-06	Schindler's List		95	R	9
7	IMDB-RANK-07	The Lord of the Rings: The Return of the King		201	PG-13	9
8	IMDB-RANK-08	Pulp Fiction	1994	154	R	8.9
9	IMDB-RANK-09	The Lord of the Rings: The Fellowship of the Ring	2001	178	PG-13	8.8
10	IMDB-RANK-10	The Good the Bad and the Ugly	1966	178	N/A	8.8
11	IMDB-RANK-11	Forrest Gump	1994	142	PG-13	8.8
12	IMDB-RANK-12	Fight Club	1999	139	R	8.8
13	IMDB-RANK-13	The Lord of the Rings: The Two Towers	2002	179	PG-13	8.8
14	IMDB-RANK-14	Inception	2010	148	PG-13	8.8
15	IMDB-RANK-15	Star Wars: Episode V - The Empire Strikes Back	1980	124	PG	8.7
16	IMDB-RANK-16	The Matrix	1999	136	R	8.7
17	IMDB-RANK-17	Goodfellas	1990	145	R	8.7
18	IMDB-RANK-18	One Flew Over the Cuckoo's Nest	1975	143	R	8.7
19	IMDB-RANK-19	Se7en	1995	127	R	8.6
20	IMDB-RANK-20	It's a Wonderful Life	1946	130	PG	8.6
21	METACRITIC-RANK-01	Citizen Kane	1941	119	PG	100
22	METACRITIC-RANK-02	The Godfather	1972	175	R	100
23	METACRITIC-RANK-03	Rear Window	1954	112	PG	100
24	METACRITIC-RANK-04	Casablanca	1942	102	PG	100
25	METACRITIC-RANK-05	Boyhood	2014	165	R	100
26	METACRITIC-RANK-06	Three Colors: Red	1994	99	R	100

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.



Split Column by Delimiter

Specify the delimiter used to split the text column.

Select or enter delimiter: --Custom--

-

Split at
○ Left-most delimiter
○ Right-most delimiter
● Each occurrence of the delimiter

Advanced options

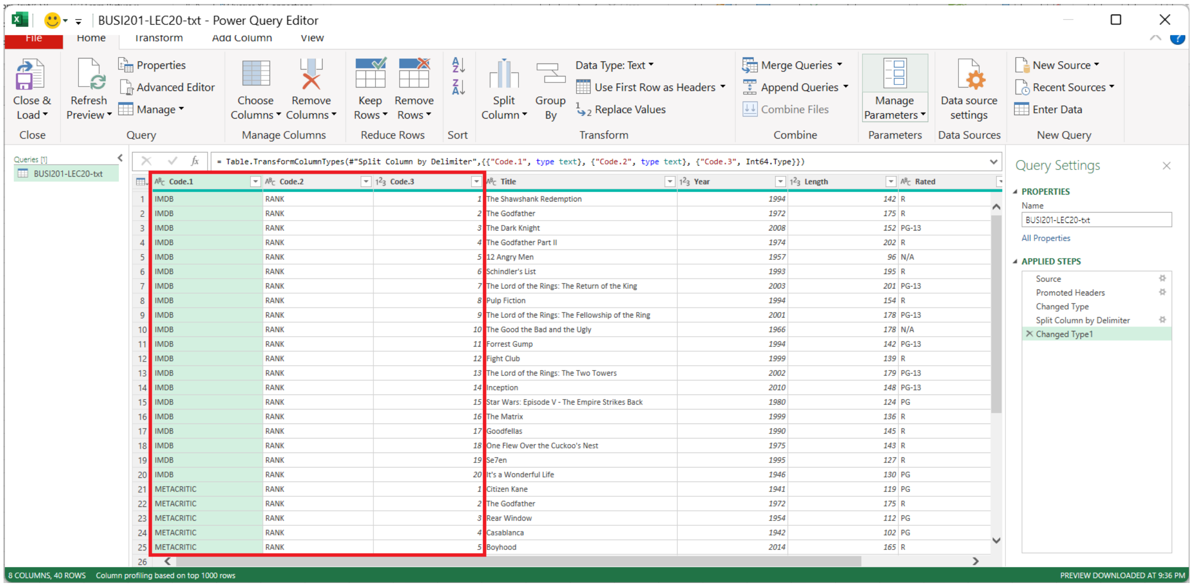
Quote Character: "

□ Split using special characters

Insert special character

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.



```
= Table.TransformColumnTypes(#"Split Column by Delimiter",{{"Code.1", type text}, {"Code.2", type text}, {"Code.3", Int64.Type}})
```

	Code.1	Code.2	Code.3	Title	Year	Length	Rated
1	IMDB	RANK	1	The Shawshank Redemption	1994	142	R
2	IMDB	RANK	2	The Godfather	1972	175	R
3	IMDB	RANK	3	The Dark Knight	2008	152	PG-13
4	IMDB	RANK	4	The Godfather Part II	1974	202	R
5	IMDB	RANK	5	12 Angry Men	1957	96	N/A
6	IMDB	RANK	6	Schindler's List	1993	195	R
7	IMDB	RANK	7	The Lord of the Rings: The Return of the King	2003	201	PG-13
8	IMDB	RANK	8	Pulp Fiction	1994	154	R
9	IMDB	RANK	9	The Lord of the Rings: The Fellowship of the Ring	2001	178	PG-13
10	IMDB	RANK	10	The Good the Bad and the Ugly	1966	178	N/A
11	IMDB	RANK	11	Forrest Gump	1994	142	PG-13
12	IMDB	RANK	12	Fight Club	1999	139	R
13	IMDB	RANK	13	The Lord of the Rings: The Two Towers	2002	179	PG-13
14	IMDB	RANK	14	Inception	2010	148	PG-13
15	IMDB	RANK	15	Star Wars: Episode V - The Empire Strikes Back	1980	124	PG
16	IMDB	RANK	16	The Matrix	1999	136	R
17	IMDB	RANK	17	Goodfellas	1990	145	R
18	IMDB	RANK	18	One Flew Over the Cuckoo's Nest	1975	143	R
19	IMDB	RANK	19	Se7en	1995	127	R
20	IMDB	RANK	20	It's a Wonderful Life	1946	130	PG
21	METACRITIC	RANK	1	Citizen Kane	1941	119	PG
22	METACRITIC	RANK	2	The Godfather	1972	175	R
23	METACRITIC	RANK	3	Rear Window	1954	112	PG
24	METACRITIC	RANK	4	Casablanca	1942	102	PG
25	METACRITIC	RANK	5	Boyhood	2014	165	R
26							

Name: BUSI201-LEC20-txt

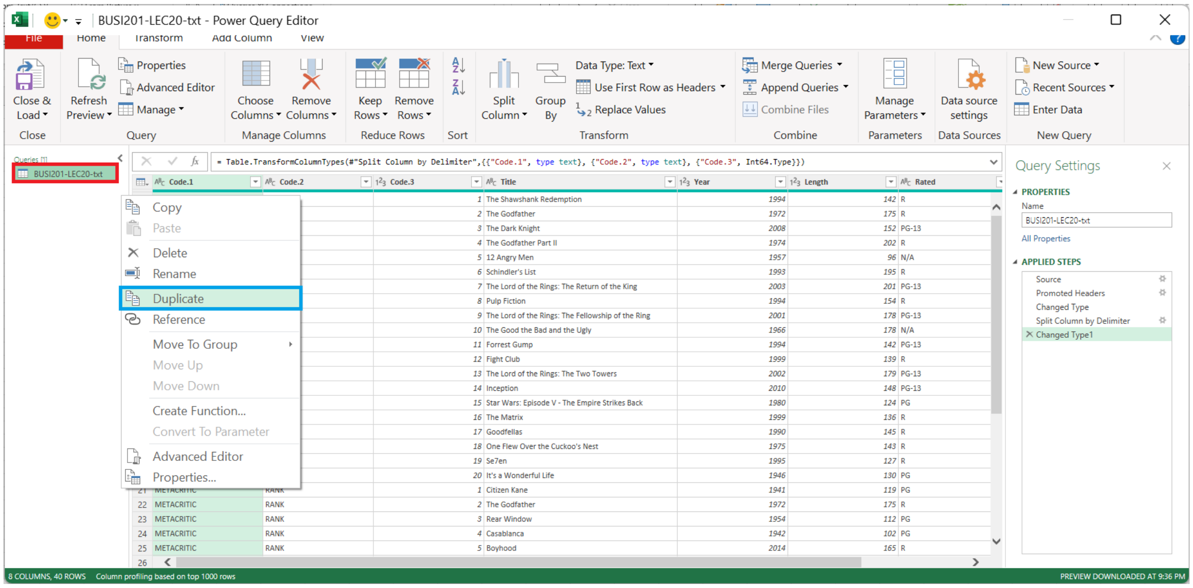
APPLIED STEPS:
- Source
- Promoted Headers
- Changed Type
- Split Column by Delimiter
- Changed Type1

8 COLUMNS, 40 ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED AT 9:36 PM

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.



```
= Table.TransformColumnTypes(#"Split Column by Delimiter",{{"Code.1", type text}, {"Code.2", type text}, {"Code.3", Int64.Type}})
```

	Code.1	Code.2	Code.3	Title	Year	Length	Rated
			1	The Shawshank Redemption	1994	142	R
			2	The Godfather	1972	175	R
			3	The Dark Knight	2008	152	PG-13
			4	The Godfather Part II	1974	202	R
			5	12 Angry Men	1957	96	N/A
			6	Schindler's List	1993	195	R
			7	The Lord of the Rings: The Return of the King	2003	201	PG-13
			8	Pulp Fiction	1994	154	R
			9	The Lord of the Rings: The Fellowship of the Ring	2001	178	PG-13
			10	The Good the Bad and the Ugly	1966	178	N/A
			11	Forrest Gump	1994	142	PG-13
			12	Fight Club	1999	139	R
			13	The Lord of the Rings: The Two Towers	2002	179	PG-13
			14	Inception	2010	148	PG-13
			15	Star Wars: Episode V - The Empire Strikes Back	1980	124	PG
			16	The Matrix	1999	136	R
			17	Goodfellas	1990	145	R
			18	One Flew Over the Cuckoo's Nest	1975	143	R
			19	Se7en	1995	127	R
			20	It's a Wonderful Life	1946	130	PG
21	METACRITIC	RANK	1	Citizen Kane	1941	119	PG
22	METACRITIC	RANK	2	The Godfather	1972	175	R
23	METACRITIC	RANK	3	Rear Window	1954	112	PG
24	METACRITIC	RANK	4	Casablanca	1942	102	PG
25	METACRITIC	RANK	5	Boyhood	2014	165	R
26							

Name: BUSI201-LEC20-txt

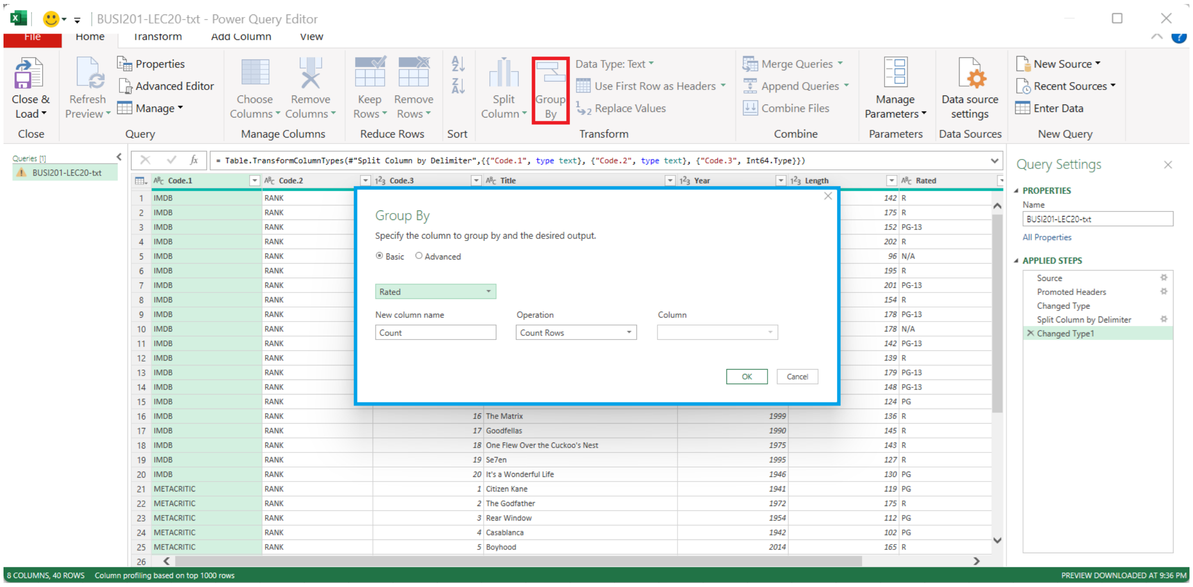
APPLIED STEPS:
- Source
- Promoted Headers
- Changed Type
- Split Column by Delimiter
- Changed Type1

8 COLUMNS, 40 ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED AT 9:36 PM

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]



```
= Table.TransformColumnTypes(#"Split Column by Delimiter",{{"Code.1", type text}, {"Code.2", type text}, {"Code.3", Int64.Type}})
```

Group By

Specify the column to group by and the desired output.

● Basic ○ Advanced

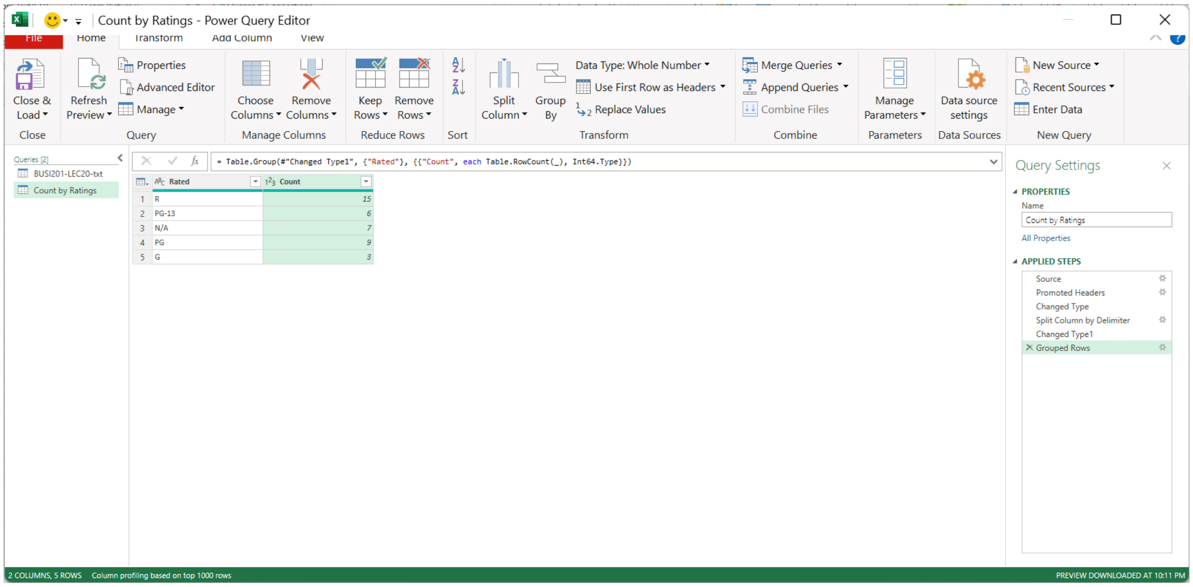
Rated

New column name	Operation	Column
Count	Count Rows	

OK Cancel

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.



```
= Table.Group(#"Changed Type1", {"Rated"}, {{"Count", each Table.RowCount(_), Int64.Type}})
```

	Rated	Count
1	R	15
2	PG-13	6
3	N/A	7
4	PG	9
5	G	3

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.