

Lecture Note #20: Data Analysis Tools Part #3

BUSI 201: Business Data Analysis

Spring 2024

Topic 1. Power Query: Pivot

Utilizing the Power Query tool, we can seamlessly transform data between the ‘wide’ and ‘long’ formats. The table on the left represents the ‘long’ format, while the one on the right represents the ‘wide’ format. The ability to switch between these formats proves valuable for tasks such as data cleaning and analysis. The choice of format may depend on the specific task at hand or the statistical software available to the user. Moving from the left to right is known as “Pivot,” and moving from the right to the left is “Unpivot.”

Variable 1	Variable 2	Value
Name1	Var1	20
Name1	Var2	50
Name1	Var3	30
Name2	Var1	40
Name2	Var2	70
Name2	Var3	10

Variable 1	ValueVar1	ValueVar2	ValueVar3
Name1	20	50	30
Name2	40	70	10

Let’s retrieve some data from the internet to practice. We’ll use the 2023 Worldwide Box Office statistics available at [Box Office Mojo](#). Navigate to the Data tab, choose From Web, enter the address, and click OK. The web URL is <https://www.boxofficemojo.com/year/world/2023/>.



Figure 1: Importing 2023 Box Office Data

Select Table 0 and eliminate unnecessary columns, specifically Rank, %, and %2. It's important to note that the current data is in the 'wide' format, with separate columns for domestic, foreign, and worldwide box office numbers. Our goal is to convert this into the 'long' format, where each movie occupies three rows, and each row contains information on worldwide, domestic, and foreign box office statistics. Go to **Transform**, and then choose the columns with box office statistics in the **blue box**. Next, select **Unpivot**.

The screenshot shows the Power Query Editor interface. The ribbon at the top has 'Transform' selected. In the main area, there is a table named 'Table 0 (2)' with data from 'Release Group'. The columns are 'Movie', 'Release Group', 'Worldwide', 'Domestic', and 'Foreign'. A blue box highlights the 'Worldwide', 'Domestic', and 'Foreign' columns. To the right, the 'Query Settings' pane shows the 'APPLIED STEPS' section with 'Removed Columns' listed.

Movie	Release Group	Worldwide	Domestic	Foreign
Barbie		\$141,769,400	\$656,169,400	\$805,900,000
2 The Super Mario Bros. Movie		\$136,888,017,00	\$574,934,330	\$786,953,687
Oppenheimer		\$90,205,530,00	\$325,326,530	\$624,879,000
Guardians of the Galaxy Vol. 3		\$845,555,777,00	\$558,995,815	\$486,559,962
Fast X		\$704,709,600,00	\$145,960,660	\$558,749,000
Spider-Man: Across the Spider-Verse		\$690,516,673,00	\$582,311,319	\$509,205,354
The Little Mermaid		\$569,626,289,00	\$298,172,056	\$271,454,233
Mission: Impossible - Dead Reckoning Part One		\$567,535,383,00	\$172,135,383	\$395,400,000
Elemental		\$495,851,987,00	\$154,426,697	\$341,425,290
Ant-Man and the Wasp: Quantumania		\$476,071,180,00	\$214,504,909	\$261,566,271
John Wick: Chapter 4		\$440,146,694,00	\$187,131,806	\$553,014,888
Transformers: Rise of the Beasts		\$438,966,592,00	\$157,066,392	\$281,900,000
Meg 2: The Trench		\$395,000,317,00	\$82,600,317	\$312,400,000
Indiana Jones and the Dial of Destiny		\$383,963,057,00	\$374,480,468	\$209,482,589
Five Nights at Freddy's		\$275,924,182,00	\$133,709,340	\$142,214,842
Creed III		\$275,248,615,00	\$156,248,615	\$119,000,000
The Fish		\$270,633,215,00	\$108,133,313	\$162,500,000
The Nun II		\$268,057,073,00	\$86,267,073	\$181,800,000
Sound of Freedom		\$247,801,879,00	\$184,174,541	\$63,627,338
20 Taylor Swift: The Eras Tour		\$246,626,030,00	\$515,835,894	\$70,690,136
Dungeons & Dragons: Honor Among Thieves		\$208,777,626,00	\$98,277,626	\$114,900,000
PAW Patrol: The Mighty Movie		\$192,495,899,00	\$645,195,899	\$127,300,000
The Equalizer 3		\$190,423,846,00	\$92,367,957	\$98,055,889
Indigo: The Red Door		\$189,086,877,00	\$82,156,962	\$106,929,815
Teenage Mutant Ninja Turtles: Mutant Mayhem		\$180,215,586,00	\$118,613,586	\$61,900,000
Scream VI		\$169,951,199,99	\$108,161,399	\$60,899,000

Figure 2: Unpivoting Box Office Statistics

By following the instructions above, you will achieve the results depicted in Figure 3. Now, you can observe that the data is arranged in the 'long' format, with each movie spanning three rows. Each row represents the film's worldwide, domestic, and foreign performance.

The screenshot shows the Power Query Editor interface. The ribbon at the top has 'Transform' selected. In the main area, there is a table named 'Table 0 (2)' with data from 'Release Group'. The columns are 'Movie', 'Release Group', 'Attribute', and 'Value'. A blue box highlights the 'Attribute' and 'Value' columns. To the right, the 'Query Settings' pane shows the 'APPLIED STEPS' section with 'Unpivoted Columns' listed.

Movie	Release Group	Attribute	Value
Barbie		Worldwide	\$141,769,400
Barbie		Domestic	\$656,169,400
Barbie		Foreign	\$805,900,000
2 The Super Mario Bros. Movie		Worldwide	\$136,888,017,00
2 The Super Mario Bros. Movie		Domestic	\$574,934,330
2 The Super Mario Bros. Movie		Foreign	\$786,953,687
Oppenheimer		Worldwide	\$90,205,530,00
Oppenheimer		Domestic	\$325,326,530
Oppenheimer		Foreign	\$624,879,000
Guardians of the Galaxy Vol. 3		Worldwide	\$845,555,777,00
Guardians of the Galaxy Vol. 3		Domestic	\$558,995,815
Guardians of the Galaxy Vol. 3		Foreign	\$486,559,962
Fast X		Worldwide	\$704,709,600,00
Fast X		Domestic	\$145,960,660
Fast X		Foreign	\$558,749,000
Spider-Man: Across the Spider-Verse		Worldwide	\$690,516,673,00
Spider-Man: Across the Spider-Verse		Domestic	\$582,311,319
Spider-Man: Across the Spider-Verse		Foreign	\$509,205,354
The Little Mermaid		Worldwide	\$569,626,289,00
The Little Mermaid		Domestic	\$298,172,056
The Little Mermaid		Foreign	\$271,454,233
Mission: Impossible - Dead Reckoning Part One		Worldwide	\$567,535,383,00
Mission: Impossible - Dead Reckoning Part One		Domestic	\$172,135,383
Mission: Impossible - Dead Reckoning Part One		Foreign	\$395,400,000
Elemental		Worldwide	\$495,851,987,00
Elemental		Domestic	\$154,426,697

Figure 3: Box Office Statistics in 'Long'

Changing Data Types

Upon observing Figure 3, it becomes apparent that the data type is not consistent across all entries. Specifically, the values for worldwide box office performance do not seem to match the other two. To address this disparity, navigate to **Home**, select the **Value** column, click **Data Type**, and then choose **Currency**. After following these steps, you will notice that the entire **Value** column now adheres to a uniform style.

The screenshot shows the Power Query Editor interface with the 'Table 0 (2)' query selected. In the top ribbon, the 'Home' tab is highlighted. On the right side, the 'Data Type' dropdown menu is open, showing various options like Decimal Number, Currency, Whole Number, Percentage, Date/Time, Date, Time, Duration, Text, True/False, and Binary. The 'Currency' option is highlighted with a red box. The main area displays a table with columns for 'Release Group', 'Attribute', and 'Value'. The 'Value' column contains numerical values representing box office earnings. The 'Query Settings' pane on the right shows the 'Name' as 'Table 0 (2)' and the 'Applied Steps' section, which includes a step named 'Changed Type1'.

Figure 4: Changing Data Types

Importing to Excel

By selecting **Close & Load** after importing the results mentioned above, we will successfully import the target data from the Box Office Mojo webpage into an Excel worksheet, as illustrated in Figure 5. Choose **Attendance** in the navigator window and proceed to select **Transform Data** as usual.

The screenshot shows an Excel spreadsheet with data from the 'Table 0 (2)' query. The columns are labeled A through P. The data includes movie titles, release groups, attributes, and values. The 'Value' column contains currency formatted numbers. To the right of the spreadsheet, the 'Queries & Connections' ribbon is visible, showing the status '600 rows loaded, 53 errors.'

Figure 5: Imported Data

Another Use of Pivot in Power Query

This time, let's explore the Pivot function in Power Query and learn how to import data from another spreadsheet. Refer to Figure 6 for guidance on importing data from other workbooks. Choose BUSI201-LEC20-Workbook.xlsx, which includes a hypothetical attendance sheet for a college-level course. We will use the Power Query Pivot function to transform this data.

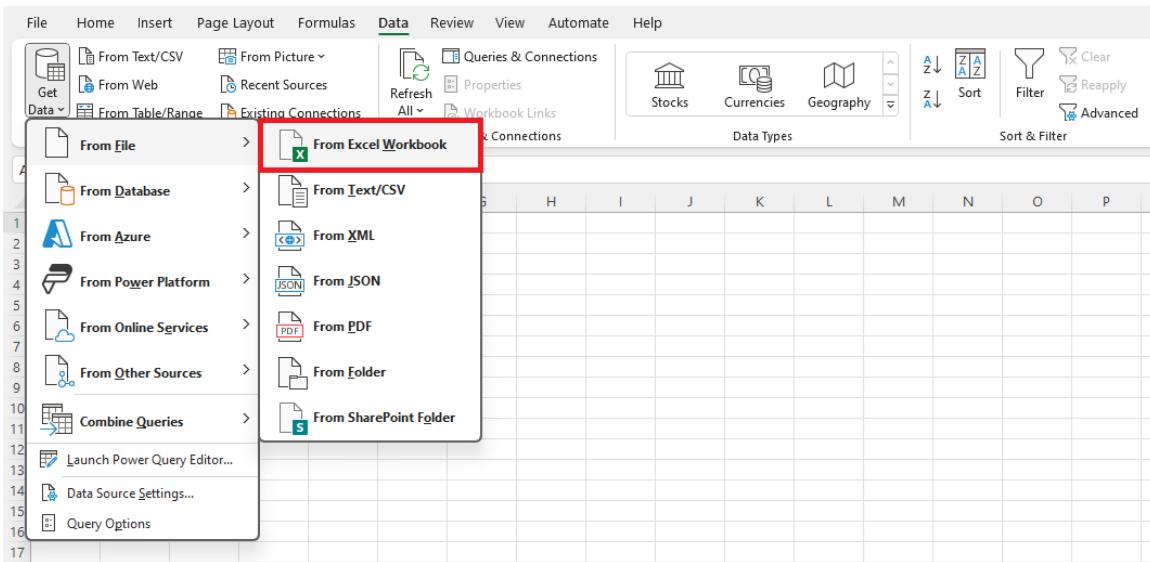


Figure 6: Importing from Workbooks

The attendance sheet will be visible in Power Query, as illustrated in Figure 7. Before proceeding with pivots, there are some changes we need to make. As the first row of the data contains variable names, select [Use First Row as Headers](#).

A screenshot of the Power Query Editor window. The title bar says 'Attendance - Power Query Editor'. The ribbon has tabs for File, Home, Transform, Add Column, and View. The 'Transform' tab is selected. The main area shows a table with 16 columns and 21 rows of student attendance data. The first row contains column headers like 'Student ID', 'Name', 'Major', etc. The 'Query Settings' pane on the right shows the 'Source' step with 'Changed Type' applied. The status bar at the bottom says '16 COLUMNS, 21 ROWS' and 'Column profiling based on top 1000 rows'.

Figure 7: Power Query Editor for Attendance

Our goal is to transform this data into a form that will work with PivotTables. In its current form, building a PivotTable based on this data is challenging, as each date's attendance record takes up an entire column. We will transform this data by “unpivoting” some columns. Select the variables that should not be “unpivoted,” which are Name, Major, and Class in this example. Then, navigate to **Transform**, select the drop-down menu next to **Unpivot Columns**, and choose **Unpivot Other Columns**.

The screenshot shows the Power Query Editor interface with the 'Attendance' query selected. The ribbon at the top has 'Transform' highlighted. In the 'Transform' tab, the 'Unpivot Columns' dropdown is open, and the option 'Unpivot Other Columns' is highlighted with a red box. The main table view shows student data with columns for Name, Major, Class, and dates from 9/4/2023 to 9/13/2023. The rows represent individual attendance records. To the right, the 'Query Settings' pane shows the 'Name' field set to 'Attendance' and the 'Applied Steps' list containing 'Promoted Headers' and 'Changed Type1'. The bottom status bar indicates '42 COLUMNS, 20 ROWS'.

Figure 8: Unpivoting the Attendance Sheet

The “unpivoted” data will appear as shown in Figure 9. We will now demonstrate how to efficiently handle dates in Excel using Power Query and construct tables showing monthly/weekly attendance of the students in the dataset.

The screenshot shows the Power Query Editor interface with the 'Attendance' query selected. The ribbon at the top has 'Transform' highlighted. The main table view now displays the unpivoted data, where each row represents a student and their attendance across dates. The columns are Name, Major, Class, Attribute (representing the date), and Value (representing P or A). The 'Applied Steps' list now includes 'Promoted Headers', 'Changed Type1', and 'Unpivoted Other Columns'. The bottom status bar indicates '5 COLUMNS, 769 ROWS'.

Figure 9: Unpivoted Attendance Data

Dates in Power Query

Let's start by renaming the columns from Attribute to Date and from Value to Attendance.¹ At this point, Excel may not recognize the unpivoted Date column as dates. Refer to Figure 11, and notice that all options in the blue box are unavailable because the variables are not recognized as dates.

Figure 10: Dates Not Recognized

Select the Date column, go to Data Type, and set the data type to Date. This will enable the Power Query editor to properly recognize that the Date column consists of dates.

Figure 11: Dates Not Recognized

¹Double-click the variable names at the top of the table to rename the columns.

Now, let's add additional columns to this dataset. Navigate to **Add Column**, select **Date**, and from the dropdown menu, choose **Month**. Refer to Figure 12 for the process displayed in the Power Query Editor. Following a similar process, we will add Weeks to this dataset.

The screenshot shows the Power Query Editor interface. The 'Add Column' button is highlighted in red. A dropdown menu is open under the 'Date' button, with 'Month' selected. The 'Month' step is listed in the 'Applied Steps' pane on the right. The main area displays a table with columns: Name, Major, Class, Date, and Attendance.

Figure 12: Adding Column: Month

Load to Excel: Connections Only

Once you have added the months and weeks, return to the Home tab in the Editor and select **Close & Load To**. In the pop-up window, click **Only Create Connection** and select **OK**. If you wish to use the Power Pivot function, you may also check the **Add this data to the Data Model** option.

The screenshot shows the 'Close & Load To...' dialog in the Power Query Editor. The 'Only Create Connection' option is selected. The 'Add this data to the Data Model' checkbox is checked. The 'Query Settings' pane on the right shows the connection name 'Attendance'. The main area displays a table with columns: Name, Major, Class, Date, Attendance, Month, and Week of Year.

Figure 13: Loading to Excel

In contrast to our previous cases, we will not be directly importing the data into this new workbook but merely creating a connection. Notice in Figure 14 that no actual data has been imported, but there is a connection shown in the red box.

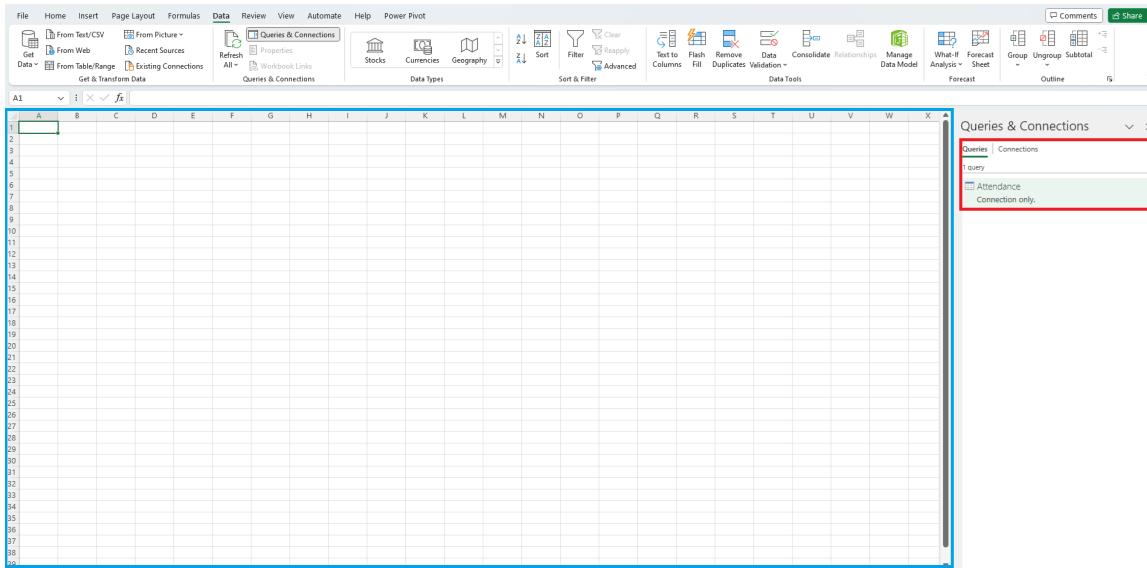


Figure 14: Connection Created

Even though the data is not included in this specific workbook, we can still create a PivotTable due to the connection we just established. Navigate to the **Insert** tab and select **PivotTable**, but be sure to choose **From External Data Source**. Then, in the new pop-up, select **Choose Connection** and confirm that the external data source is the **Query-Attendance** that we just created.

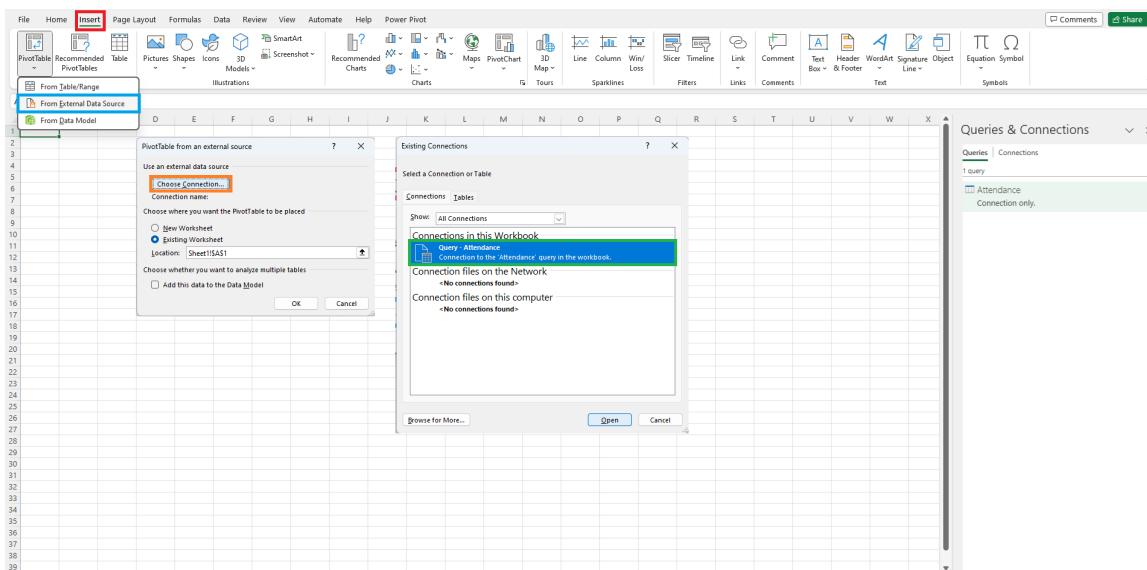


Figure 15: PivotTable with External Connections

Then, we can use the standard PivotTable framework to generate tables that show us the attendance status of all students over three months as shown in Figure 16.

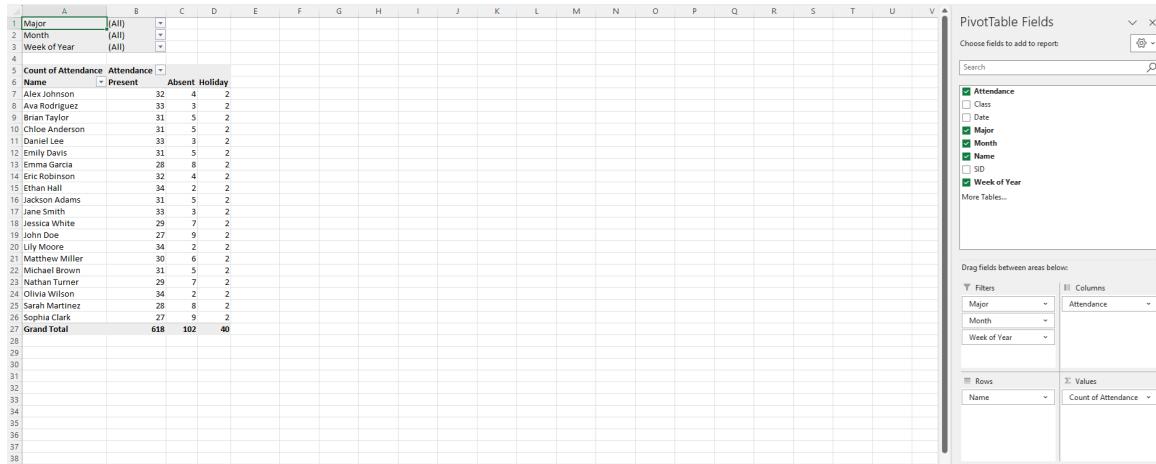


Figure 16: Completed PivotTable

Topic 2. Merging Data: Manual Approach

Sometimes, we may encounter situations where we need to merge data from different sources. For example, open the workbook BUSI201-LEC20-Workbook.xlsx and go to the worksheet **Shell**. As shown in the accompanying Figure 17, you will find a worksheet with student IDs, names, classes, and majors, where the quiz, midterm, and final scores are currently empty.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	C
1	SID	Name	Major	Class	Quiz	Midterm	Final							
2	39630	John Doe	BUSI	FR										
3	88741	Jane Smith	ACCT	SO										
4	58518	Alex Johnson	ECON	SR										
5	69909	Emily Davis	ECON	SO										
6	43307	Michael Brown	ACCT	FR										
7	49372	Olivia Wilson	ACCT	SO										
8	95461	Daniel Lee	BUSI	SO										
9	55398	Sarah Martinez	ECON	JR										
10	62040	Brian Taylor	ACCT	JR										
11	32403	Chloe Anderson	BUSI	JR										
12	34358	Eric Robinson	ECON	FR										
13	10536	Jessica White	BUSI	SO										
14	30357	Matthew Miller	BUSI	JR										
15	49615	Emma Garcia	ECON	SO										
16	28778	Nathan Turner	ECON	FR										
17	33565	Lily Moore	ECON	JR										
18	69282	Ethan Hall	ECON	SO										
19	92750	Sophia Clark	ECON	SR										
20	47053	Jackson Adams	BUSI	SO										
21	97585	Ava Rodriguez	BUSI	SO										

Figure 17: Shell Worksheet

To manually merge information, we must first identify a unique identifier. If there is no unique identifier, we must create one. Fortunately, in this specific case, the student ID number (SID) serves as a unique identifier. With a unique identifier available, we can use the VLOOKUP function to retrieve students' performance data from the other three worksheets.

Specifically, the formula we must use will be:

- E2: =VLOOKUP(\$A2,Quiz!\$A\$2:\$E\$21,5,0)
- F2: =VLOOKUP(\$A2,Midterm!\$A\$2:\$E\$21,5,0)
- G2: =VLOOKUP(\$A2,Final!\$A\$2:\$E\$21,5,0)

See Figure 18 for the manually merged results.

A	B	C	D	E	F	G	H	I	J	K
1	SID	Name	Major	Class	Quiz	Midterm	Final			
2	39630	John Doe	BUSI	FR	65.883544	63.848031	89.667084			
3	88741	Jane Smith	ACCT	SO	70.622386	76.023977	94.590742			
4	58518	Alex Johnson	ECON	SR	44.240421	59.19928	94.398706			
5	69909	Emily Davis	ECON	SO	32.011265	73.41517	95.950031			
6	43307	Michael Brown	ACCT	FR	22.790446	90.155576	91.733427			
7	49372	Olivia Wilson	ACCT	SO	72.342314	95.111828	90.850339			
8	95461	Daniel Lee	BUSI	SO	100	78.101099	98.221627			
9	55398	Sarah Martinez	ECON	JR	77.773608	88.937366	87.426706			
10	62040	Brian Taylor	ACCT	JR	74.419605	100	90.713715			

Figure 18: Manually Matched Values

Topic 3. Power Query: Merging Data

We can also rely on Power Query to merge the data instead of manually merging it. Open another workbook and import the workbook BUSI201-LEC21-Workbook.xlsx into the Power Query Editor. Select the entire workbook, as shown in Figure 19, and choose [Transform Data](#).

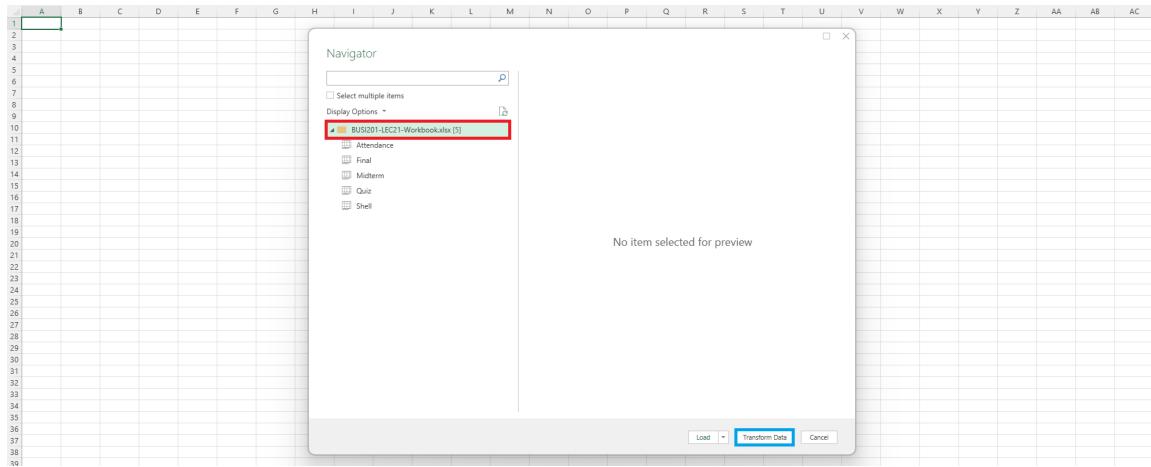


Figure 19: Importing the Entire Workbook

Select the tables in the **red box** in Figure 20. Right-click and select [Add as New Query](#). Each table will be added as a new query in the **orange box**. After adding each element as a new query, you may need to navigate to each individual query to set [Use First Row as Headers](#).

The screenshot shows the Power Query Editor with the following details:

- File ribbon:** Home, Transform, Add Column, View.
- Toolbar:** Close & Load, Refresh, Advanced Editor, Properties, Choose Columns, Remove Columns, Manage Columns, Keep Rows, Remove Rows, Sort, Data Type: Table, Use First Row as Headers, Merge Queries, Append Queries, Combine Files, Manage Parameters, Data source settings, New Source, Recent Sources, Enter Data, New Query.
- Queries pane:** Lists five queries: Attendance, Quiz, Midterm, Final, and Shell. The 'Attendance' entry is highlighted with a red box.
- Preview pane:** Shows the data for the 'Attendance' table. A red box highlights the table in the preview. An orange box highlights the 'Attendance' table in the preview pane. A blue box highlights the 'Add as New Query' button in the context menu.
- Query Settings pane:** Shows the properties for the current query, which is 'Source'. It includes sections for **PROPERTIES** (Name: BUSI201-LEC21-Workbook.xlsx) and **APPLIED STEPS** (Source).

Figure 20: Adding Tables to Query

To merge the data and achieve a result similar to the final product in **Topic 2**, start by selecting the Quiz_Sheet query. Then, choose **Merge Queries as New** to create a new query that can be exported later. Alternatively, you can select **Merge Queries** to merge the data directly into the Quiz_Sheet query.

Figure 21: Initiate Merge

In the pop-up window, select the query that holds the data you want to merge with Quiz_Sheet to create a new query. You can do this by choosing the appropriate query in the **red box** in Figure 22. Then, left-click on the unique identifiers in both queries, as shown in the **blue boxes**, and click OK.

Figure 22: First Round of Merging

We need to make some changes to the newly generated query. Click the dropdown menu in the red box in Figure 23. Uncheck all elements other than the variable you want to include in the query, as seen in the blue box. You may also uncheck the **Use original column name as prefix** for brevity.

Figure 23: First Round of Merging (cont'd)

The newly generated query resulting from the steps detailed above is visible in Figure 24. We can complete this merging process by following similar steps to merge final exam data into this new Merge1 query.

Figure 24: First Round of Merging Completed

Figure 25: All Data Merged in Query

To merge the final exam information into this query, select **Merge Queries** this time instead of **Merge Queries as New**, and follow the exact same steps. Choose the student ID numbers, and go through the checkboxes. The **Merge1** query following this process is shown in Figure 25. Importing this query into Excel, the result is displayed in Figure 26.

	A	B	C	D	E	F	G	H	I	J	K	L
1	SID	Name	Major	Class	Quiz	Midterm	Final					
2	39630	John Doe	BUSI	FR	65.88354368	63.84803122	89.66708395					
3	88741	Jane Smith	ACCT	SO	70.6223857	76.02397659	94.59074178					
4	58518	Alex Johnson	ECON	SR	44.24042112	59.19927957	94.39870638					
5	69909	Emily Davis	ECON	SO	32.01126483	73.41516977	95.95003058					
6	43307	Michael Brown	ACCT	FR	22.79044558	90.1555764	91.73342698					
7	49372	Olivia Wilson	ACCT	SO	72.3423142	95.11182753	90.85033899					
8	95461	Daniel Lee	BUSI	SO	100	78.10109853	98.22162731					
9	55398	Sarah Martinez	ECON	JR	77.77360796	88.93736639	87.42670642					
10	62040	Brian Taylor	ACCT	JR	74.41960488	100	90.71371502					
11	32403	Chloe Anderson	BUSI	JR	74.971991	62.26719079	84.57406171					
12	34358	Eric Robinson	ECON	FR	69.76554044	67.57332321	86.69297478					
13	10536	Jessica White	BUSI	SO	69.24128946	93.20209547	97.07938781					
14	30357	Matthew Miller	BUSI	JR	72.96911753	84.40236014	93.17842013					
15	49615	Emma Garcia	ECON	SO	100	69.94307312	88.48818443					
16	28778	Nathan Turner	ECON	FR	87.1204835	75.86425166	95.65823593					
17	33565	Lily Moore	ECON	JR	73.00066166	88.40781056	92.30112903					
18	69282	Ethan Hall	ECON	SO	100	83.40013971	94.87143174					
19	92750	Sophia Clark	ECON	SR	60.27391392	100	87.99410653					
20	47053	Jackson Adams	BUSI	SO	58.81797215	70.92857179	88.74993291					
21	97585	Ava Rodriguez	BUSI	SO	98.08377044	72.35184801	86.68562663					
22												

Figure 26: Merged Query Imported