Lecture Note #21: Data Analysis Tools Part #3

BUSI 201: Business Data Analysis

Fall 2023

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`.

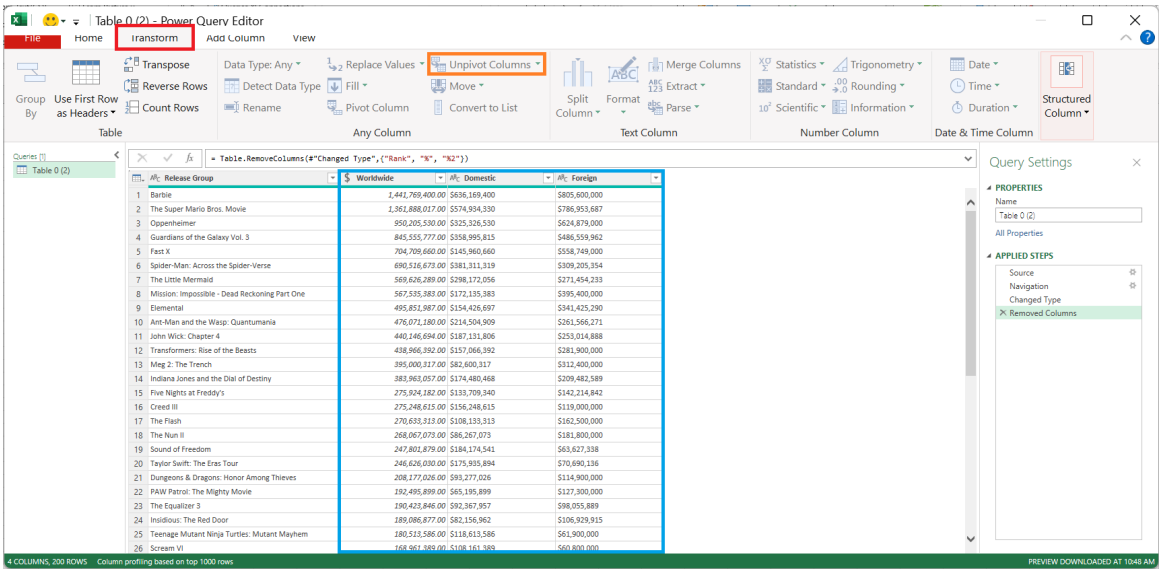


Figure 2: Unpivoting Box Office Statistics

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.

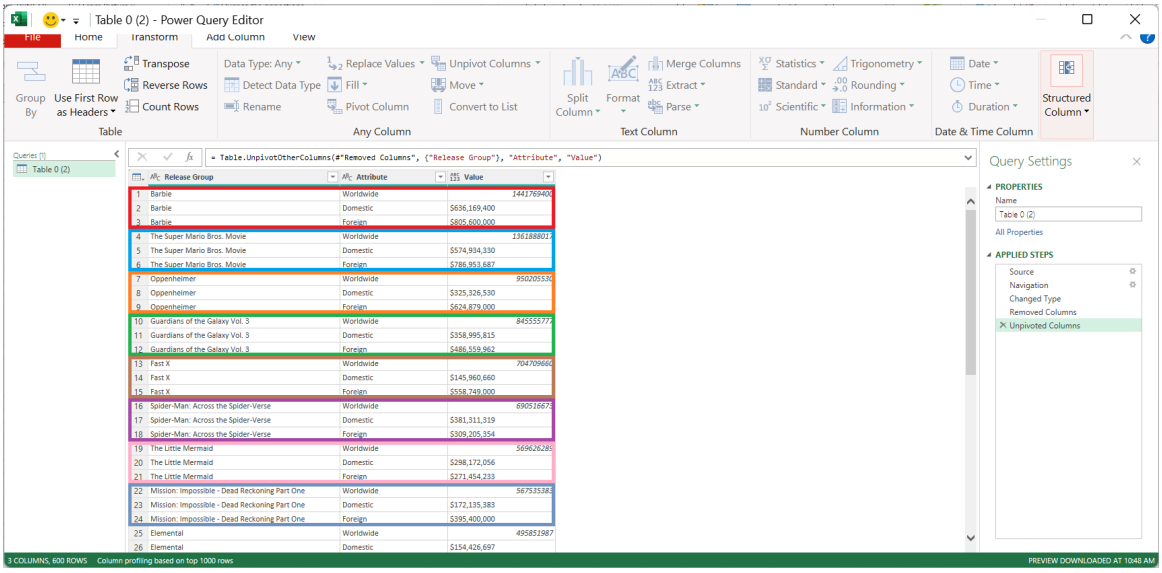


Figure 3: Box Office Statistics in 'Long'

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.

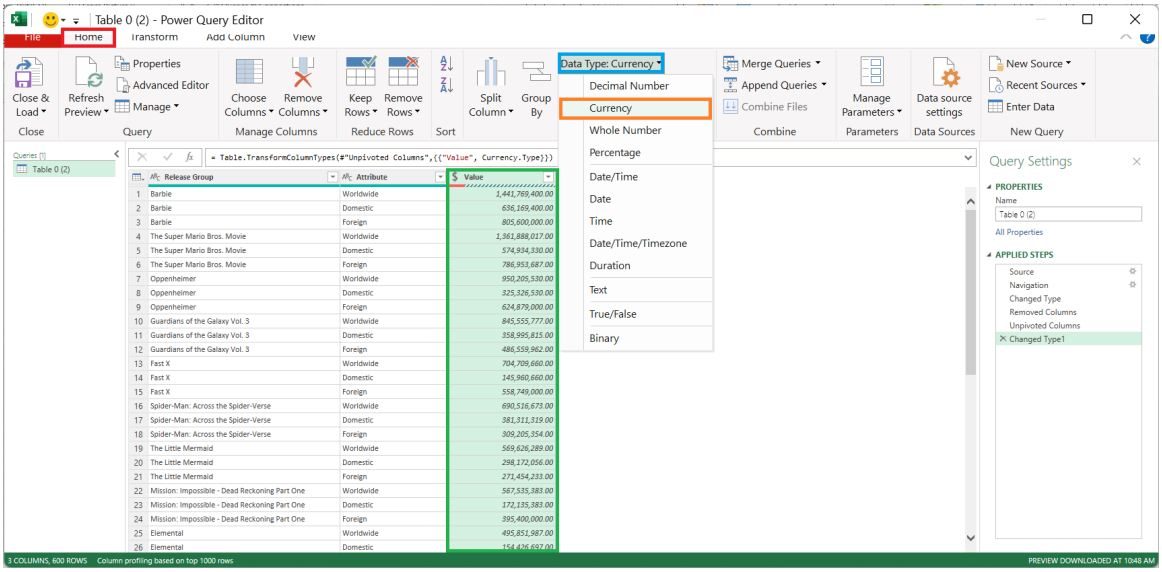


Figure 4: Changing Data Types

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.

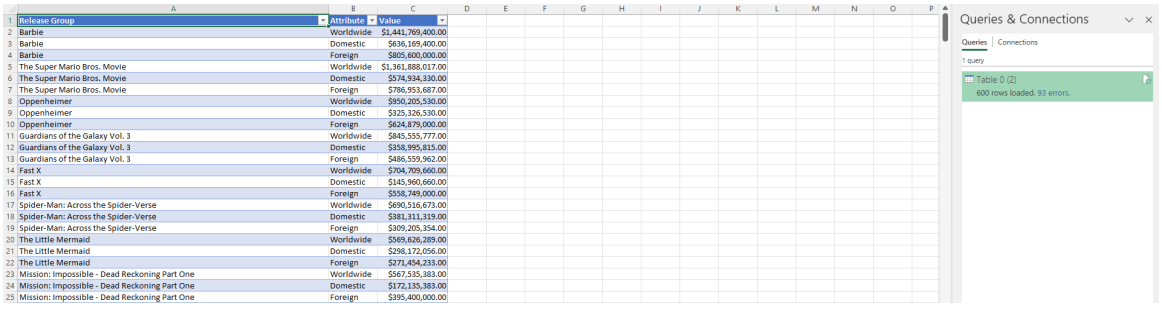


Figure 5: Imported Data

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-LEC21-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.

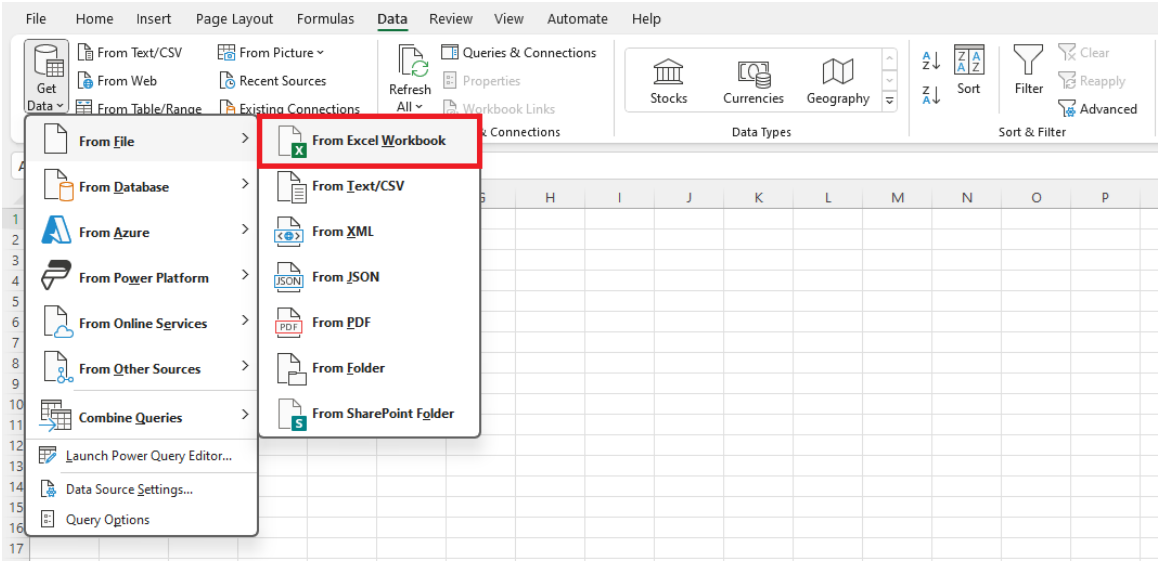


Figure 6: Importing from Workbooks

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`.

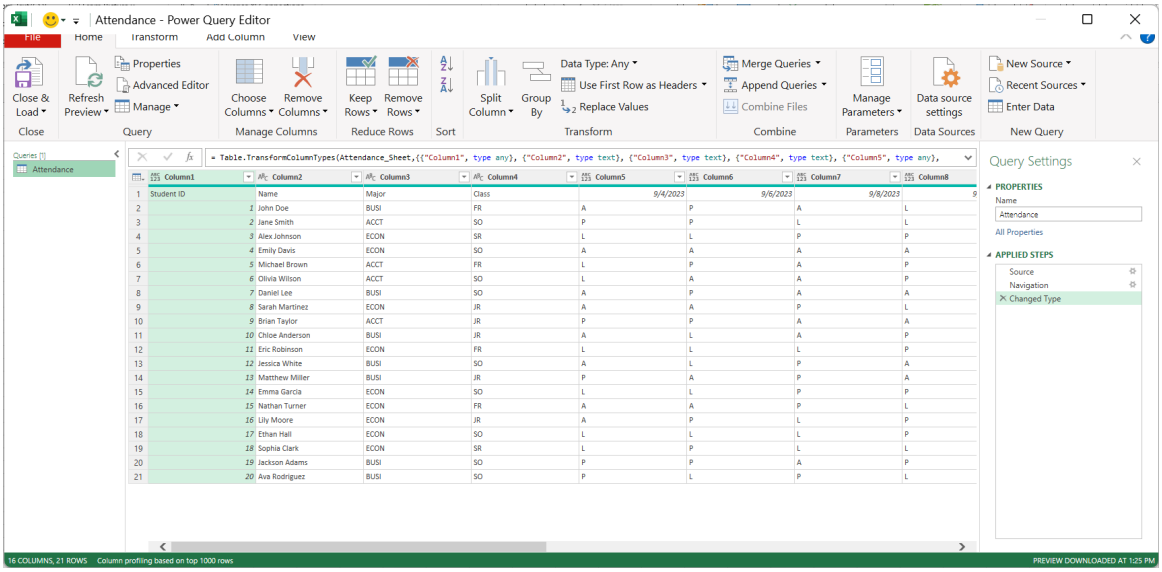


Figure 7: Power Query Editor for `Attendance`

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`.

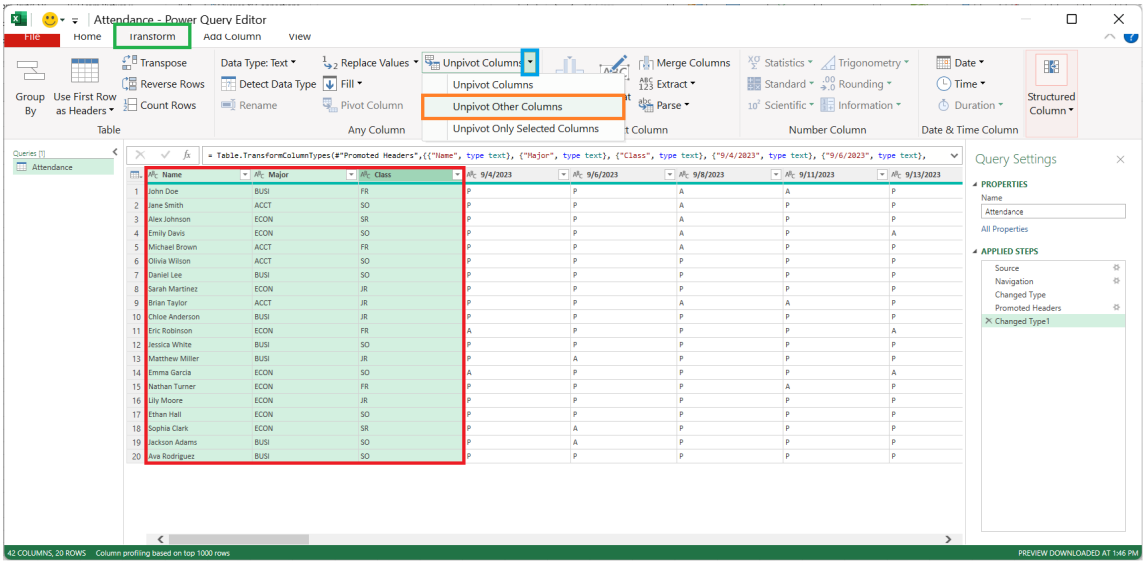


Figure 8: Unpivoting the `Attendance` Sheet

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.

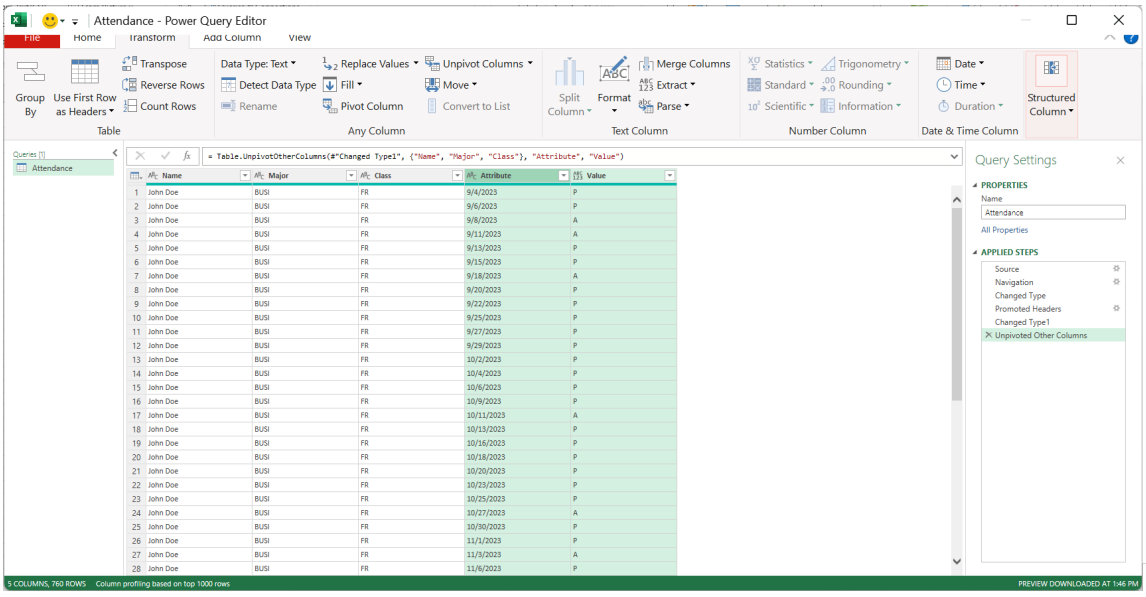


Figure 9: Unpivoted `Attendance` Data

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.

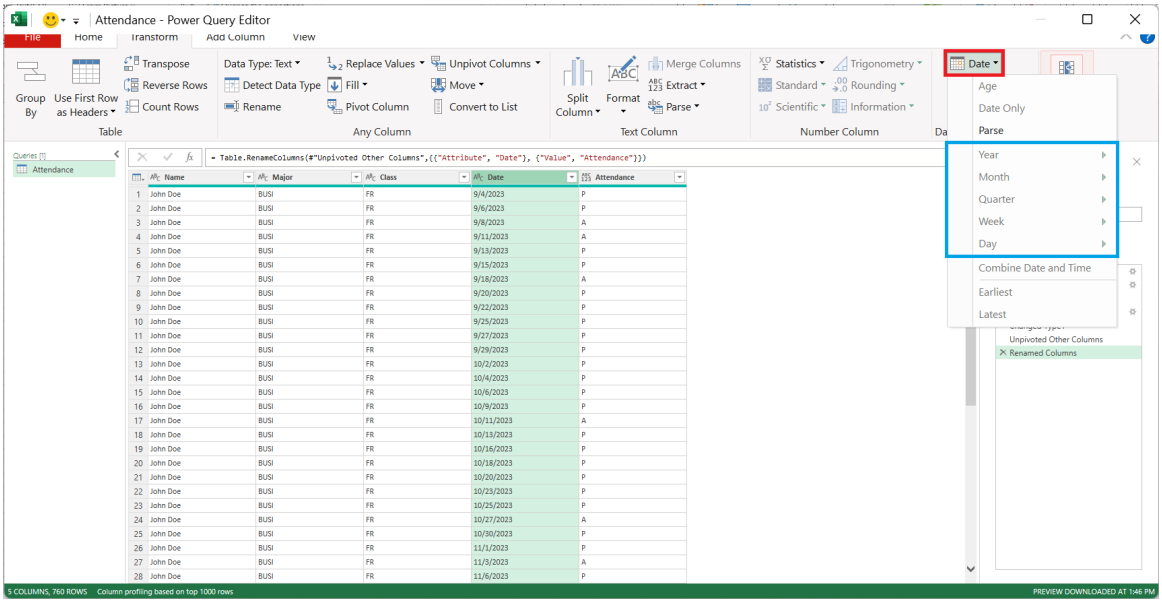


Figure 10: Dates Not Recognized

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.

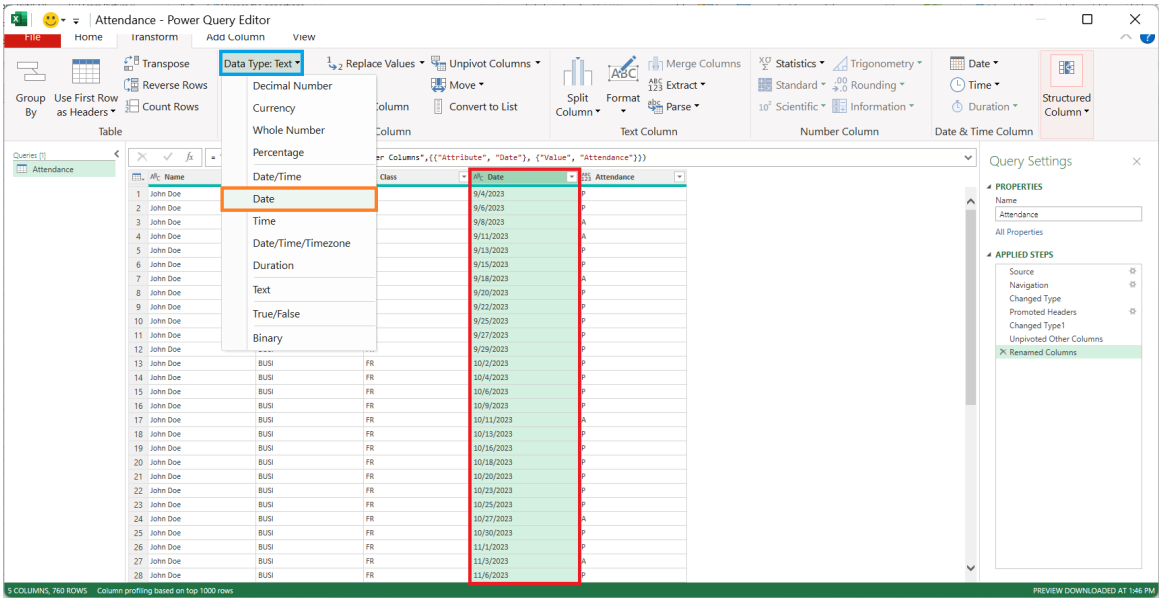


Figure 11: Dates Not Recognized

[1] Double-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.

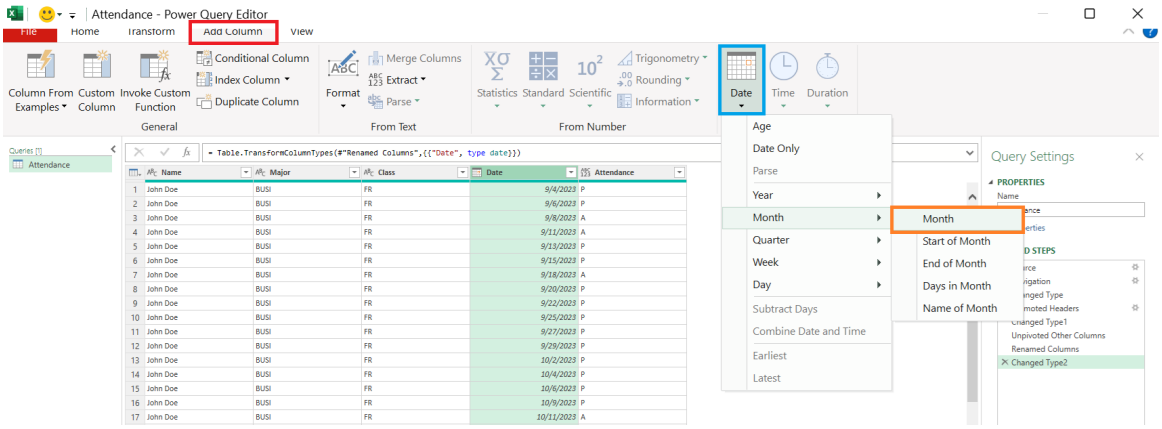


Figure 12: Adding Column: Month

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.

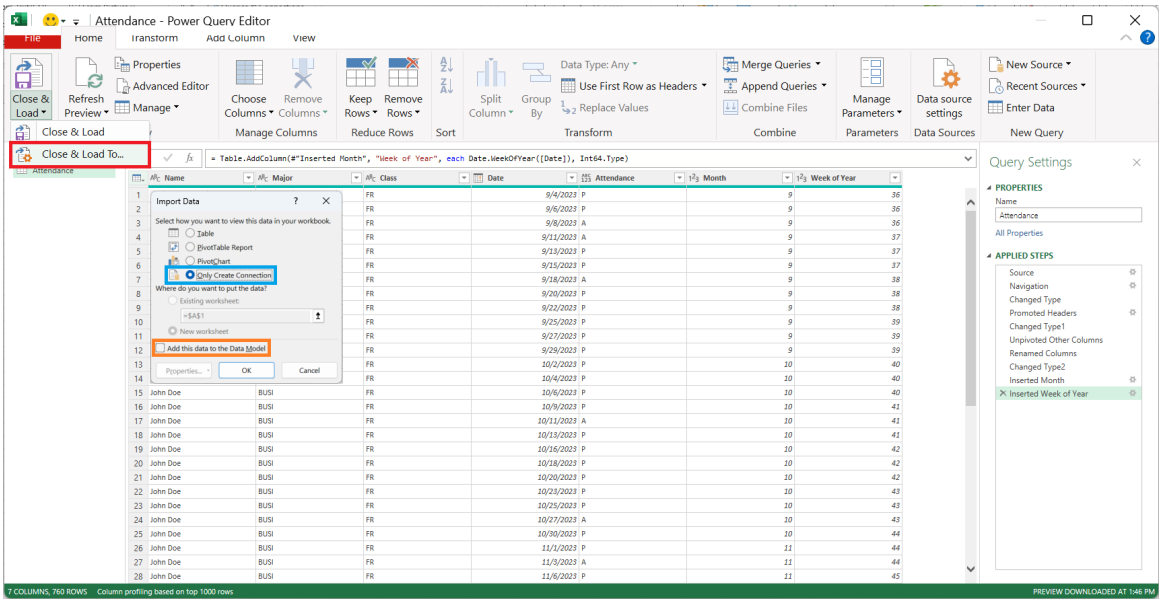


Figure 13: Loading to Excel

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.

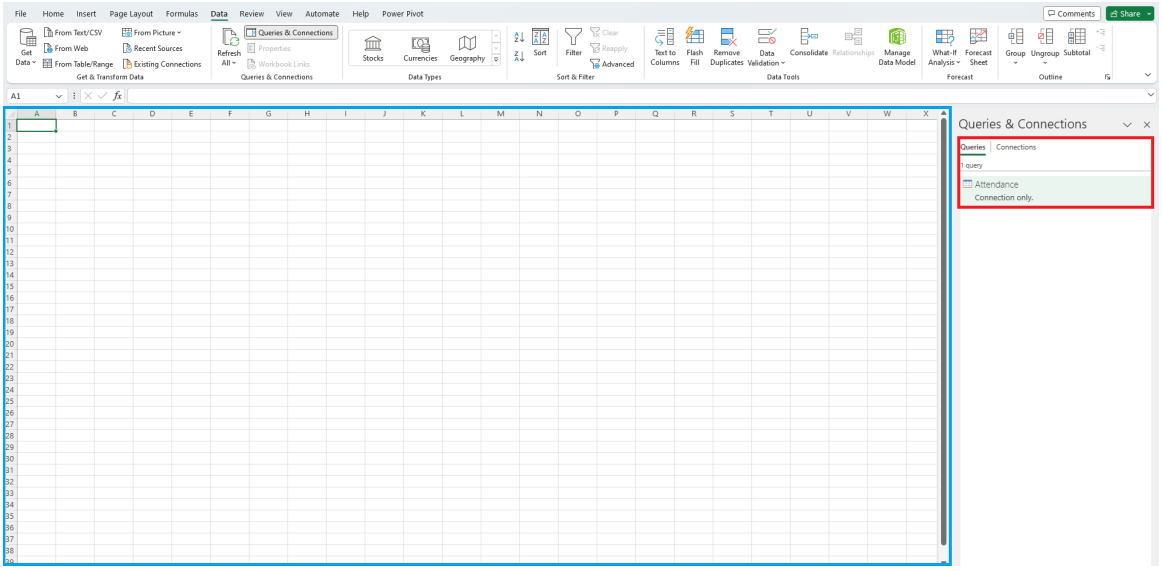


Figure 14: Connection Created

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.

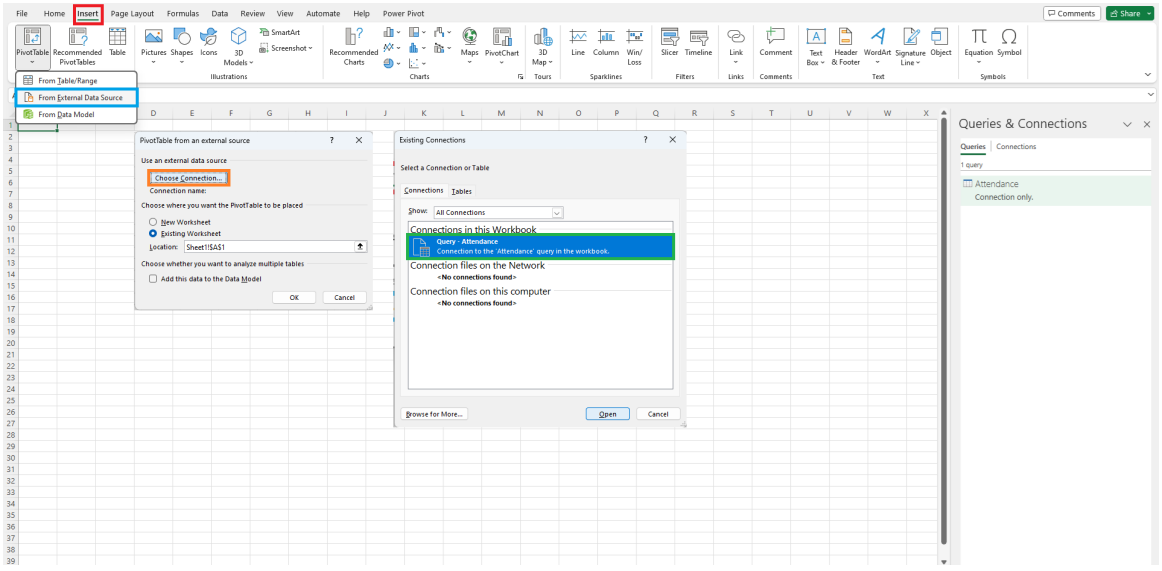
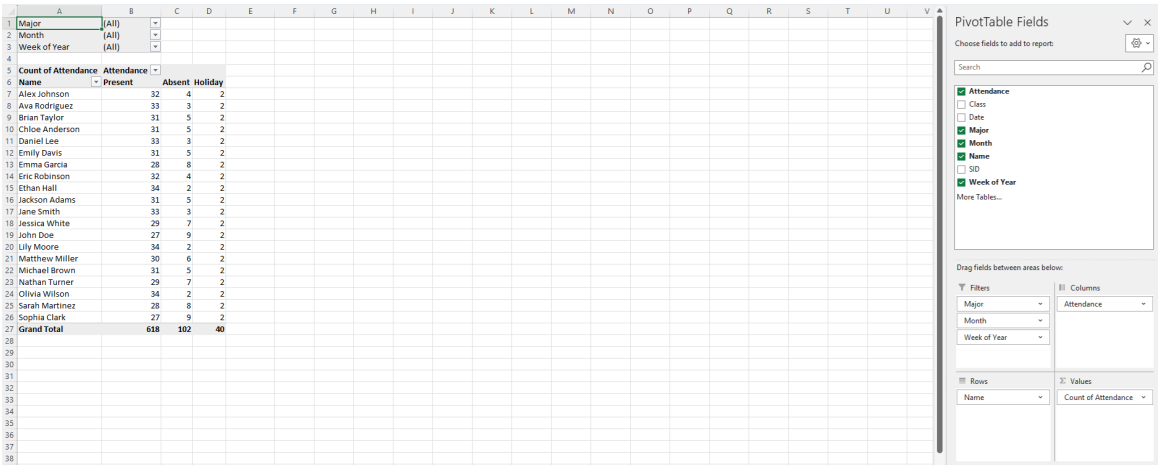


Figure 15: PivotTable with External Connections

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.



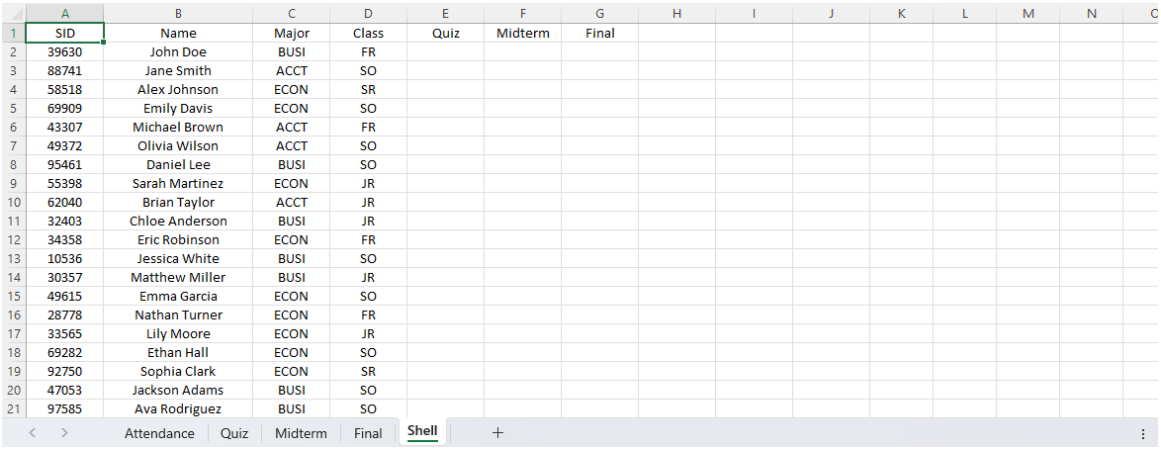
Major	(All)	
Month	(All)	
Week of Year	(All)	

Count of Attendance	Attendance		
Name	Present	Absent	Holiday
Alex Johnson	32	4	2
Ava Rodriguez	33	3	2
Brian Taylor	31	5	2
Chloe Anderson	31	5	2
Daniel Lee	33	3	2
Emily Davis	31	5	2
Emma Garcia	28	8	2
Eric Robinson	32	4	2
Ethan Hall	34	2	2
Jackson Adams	31	5	2
Jane Smith	33	3	2
Jessica White	29	7	2
John Doe	27	9	2
Lily Moore	34	2	2
Matthew Miller	30	6	2
Michael Brown	31	5	2
Nathan Turner	29	7	2
Olivia Wilson	34	2	2
Sarah Martinez	28	8	2
Sophia Clark	27	9	2
Grand Total	618	102	40

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-LEC21-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
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			

Attendance | Quiz | Midterm | Final | Shell

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
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`.

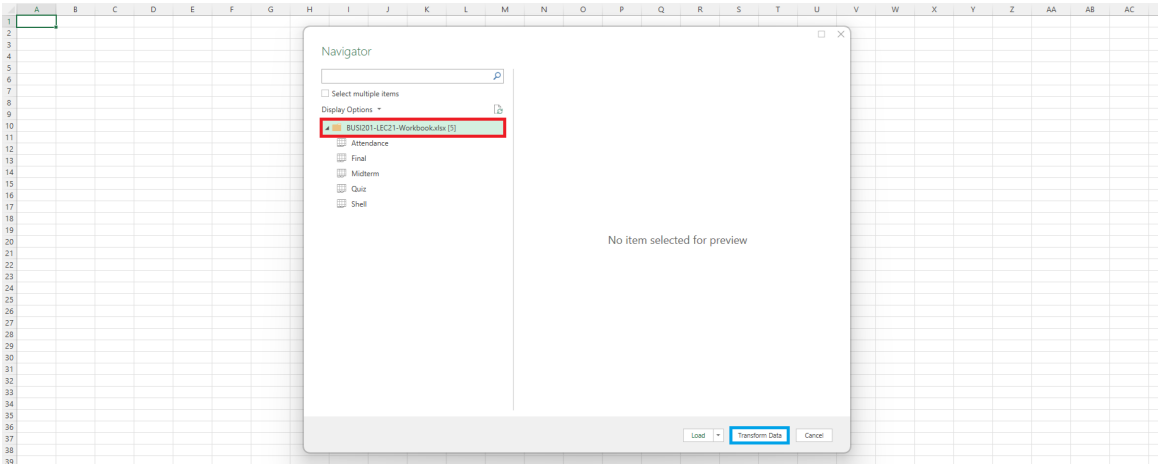


Figure 19: Importing the Entire Workbook

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`.

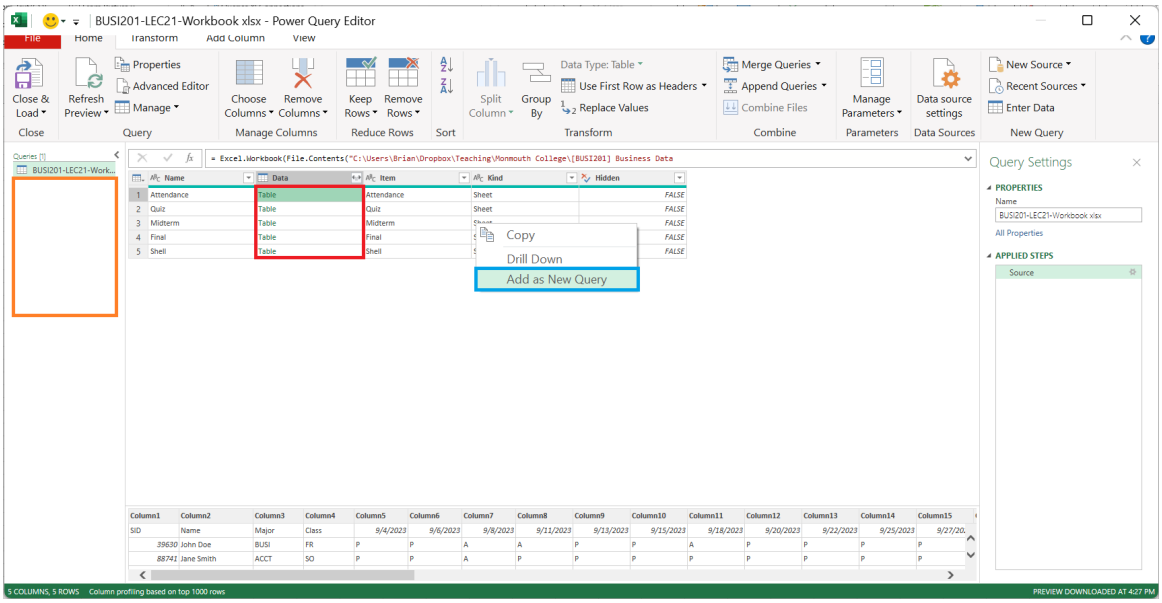


Figure 20: Adding Tables to Query

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.

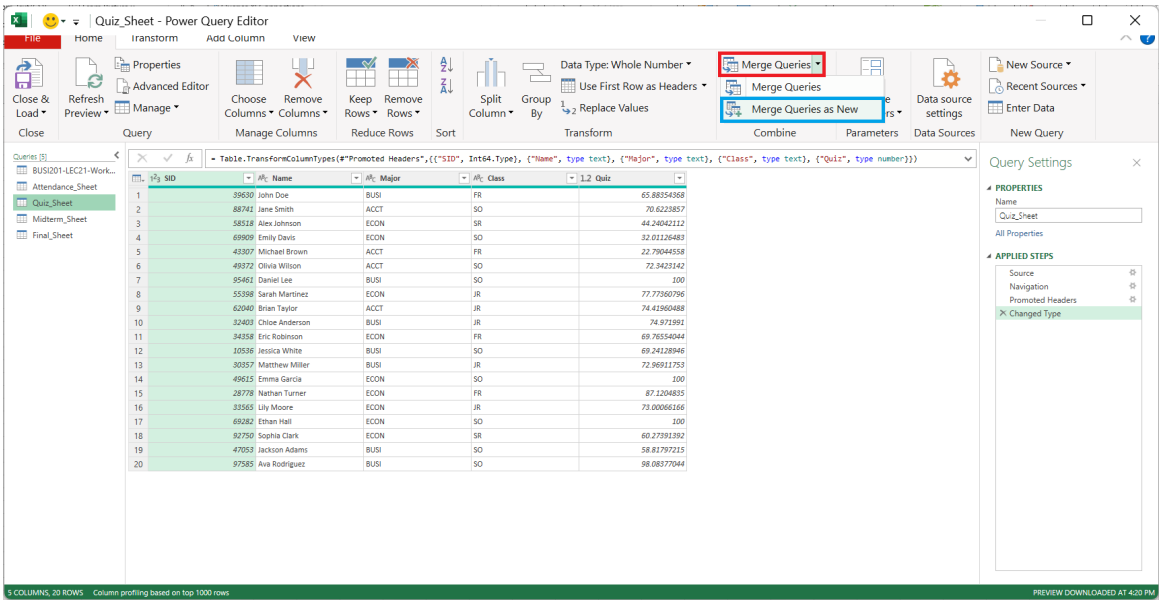


Figure 21: Initiate Merge

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`.

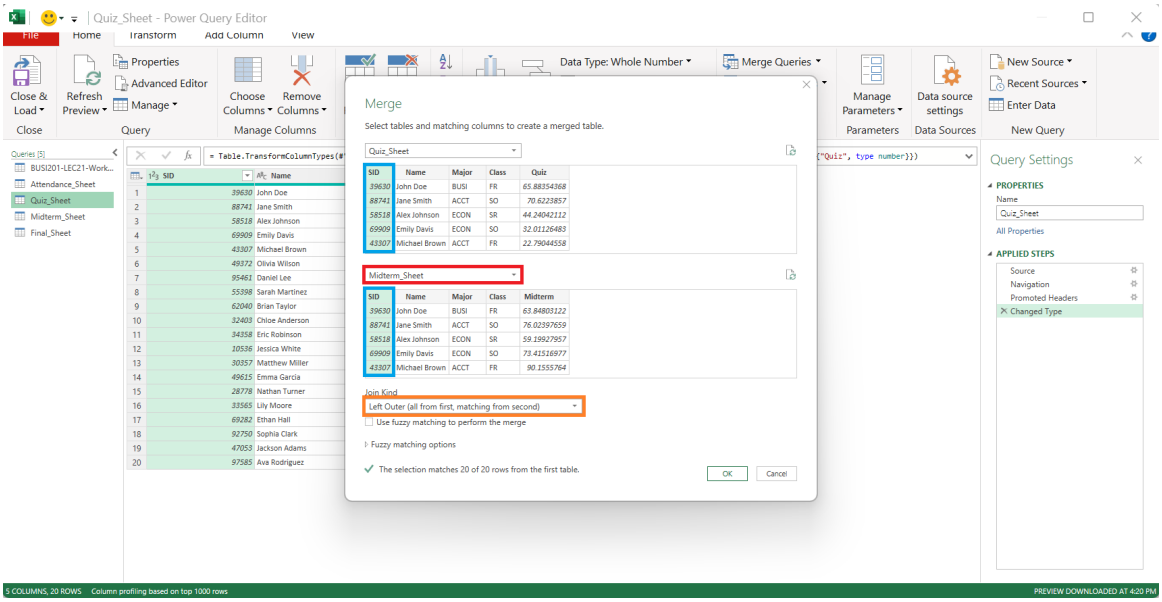


Figure 22: First Round of Merging

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.

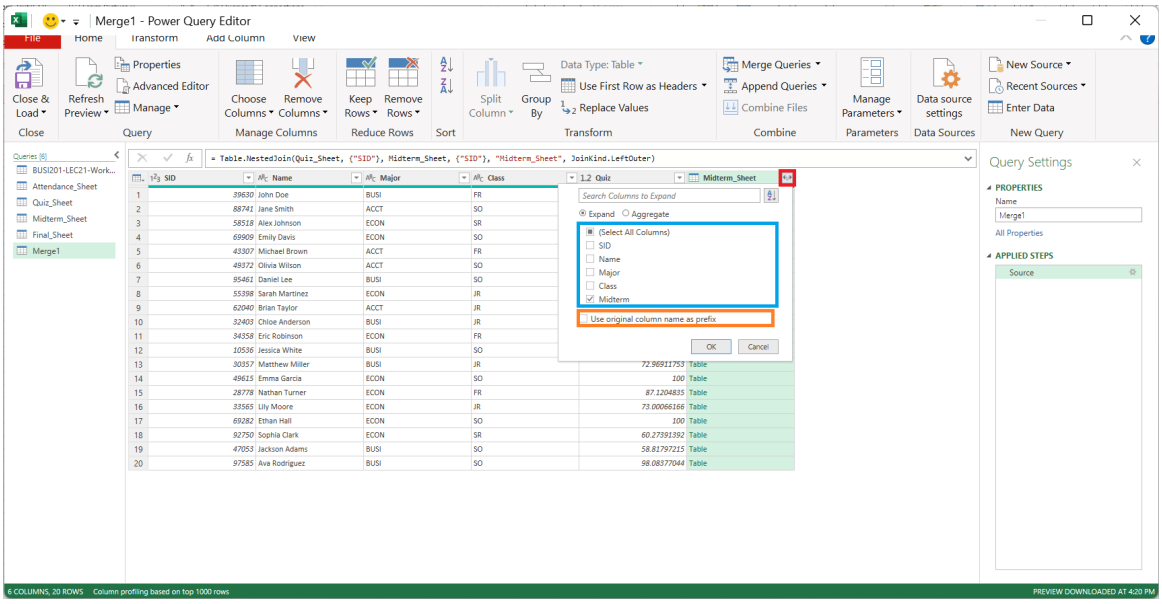


Figure 23: First Round of Merging (cont'd)

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.

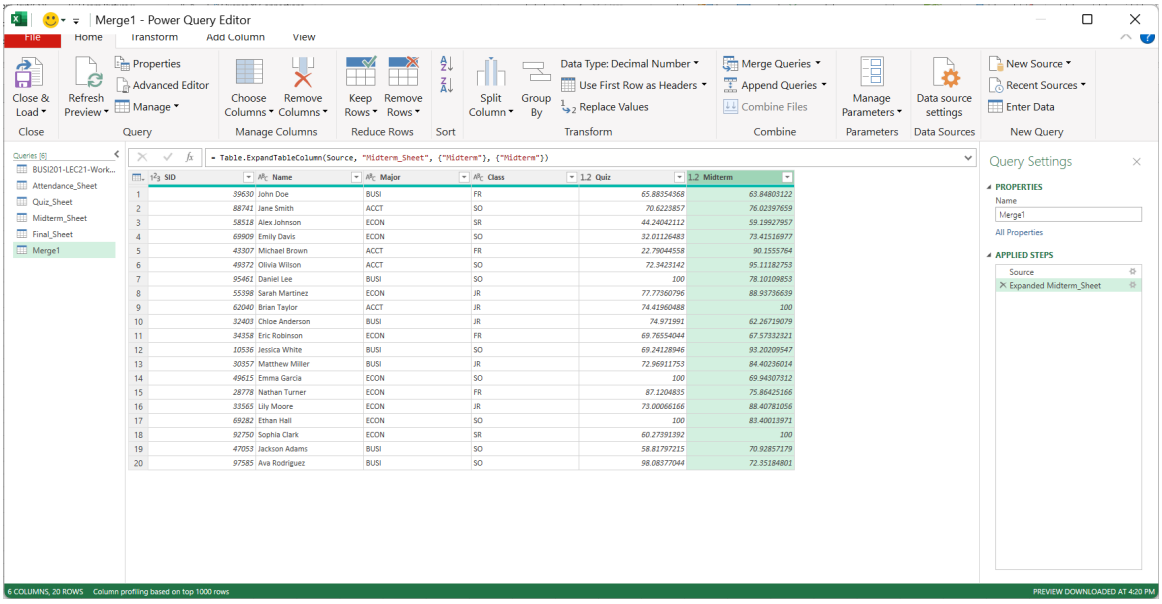


Figure 24: First Round of Merging Completed

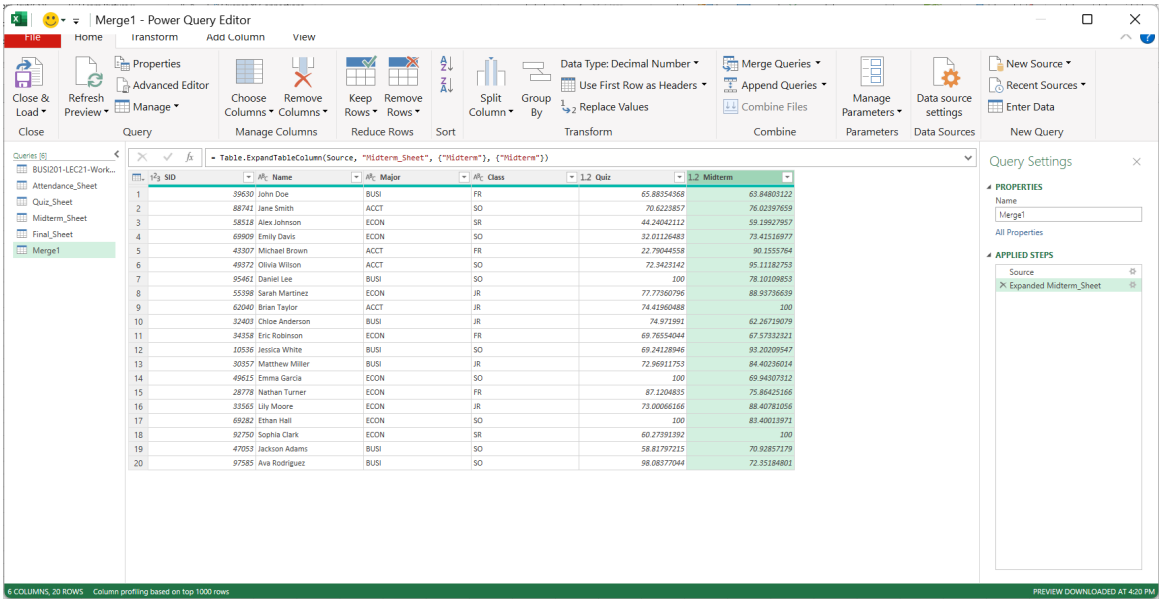


Figure 25: All Data Merged in Query

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.

SID	Name	Major	Class	Quiz	Midterm	Final
39630	John Doe	BUSI	FR	65.88354368	63.84803122	89.66708395
88741	Jane Smith	ACCT	SO	70.6223857	76.02397659	94.59074178
58518	Alex Johnson	ECON	SR	44.24042112	59.19927957	94.39870638
69909	Emily Davis	ECON	SO	32.01126483	73.41516977	95.95003058
43307	Michael Brown	ACCT	FR	22.79044558	90.1555764	91.73342698
49372	Olivia Wilson	ACCT	SO	72.3423142	95.11182753	90.85033899
95461	Daniel Lee	BUSI	SO	100	78.10109853	98.22162731
55398	Sarah Martinez	ECON	JR	77.77360796	88.93736639	87.42670642
62040	Brian Taylor	ACCT	JR	74.41960488	100	90.71371502
32403	Chloe Anderson	BUSI	JR	74.971991	62.26719079	84.57406171
34358	Eric Robinson	ECON	FR	69.76554044	67.57332321	86.69297478
10536	Jessica White	BUSI	SO	69.24128946	93.20209547	97.07938781
30357	Matthew Miller	BUSI	JR	72.96911753	84.40236014	93.17842013
49615	Emma Garcia	ECON	SO	100	69.94307312	88.48818443
28778	Nathan Turner	ECON	FR	87.1204835	75.86425166	95.65823593
33565	Lily Moore	ECON	JR	73.00066166	88.40781056	92.30112903
69282	Ethan Hall	ECON	SO	100	83.40013971	94.87143174
92750	Sophia Clark	ECON	SR	60.27391392	100	87.99410653
47053	Jackson Adams	BUSI	SO	58.81797215	70.92857179	88.74993291
97585	Ava Rodriguez	BUSI	SO	98.08377044	72.35184801	86.68562663

Figure 26: Merged Query Imported