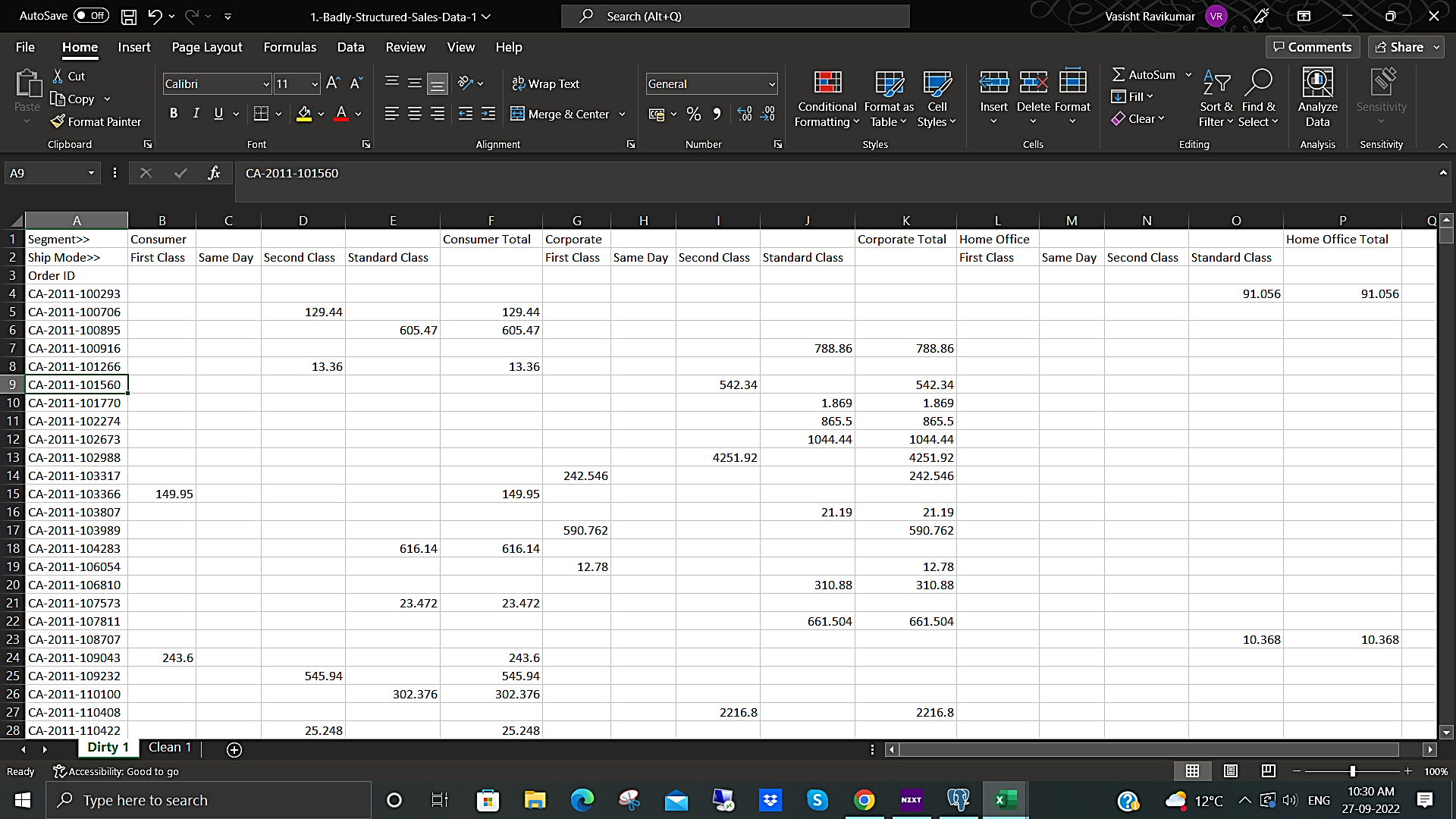
**Cleaning a poorly structured business data from an external source**

This is a step-by-step run down of the practice done for cleaning a dirty data obtained from the website foresightbi.com (<https://foresightbi.com.ng/microsoft-power-bi/dirty-data-samples-to-practice-on/>)

Figure 1. Dirty Data as collected from the website



This data was probably taken directly from a pivot table which required a switching of all the **Segment** types from rows to columns and the **Ship Modes** from rows to columns. If this can be achieved, the **Sales data** can all be consolidated under a single column for a better look, also pivot tables and charts can be later drawn out of it. Unfortunately, the transpose option wouldn’t work in this case, as the rows and columns aren’t exactly interchanged.

To tackle this problem, the **Order ID,** the 4 **Ship Modes** columns (First Class, Same Day, Second Class and Standard Class) under each **Segment** (Consumer, Corporate, Home Office) were converted into a Power Query Editor.

To open any range in a Power Query Editor, selected the required range and click the **From Table/Range** from the **Get & Transform Data** Tab. This tab can be accessed from the **Data** Ribbon.

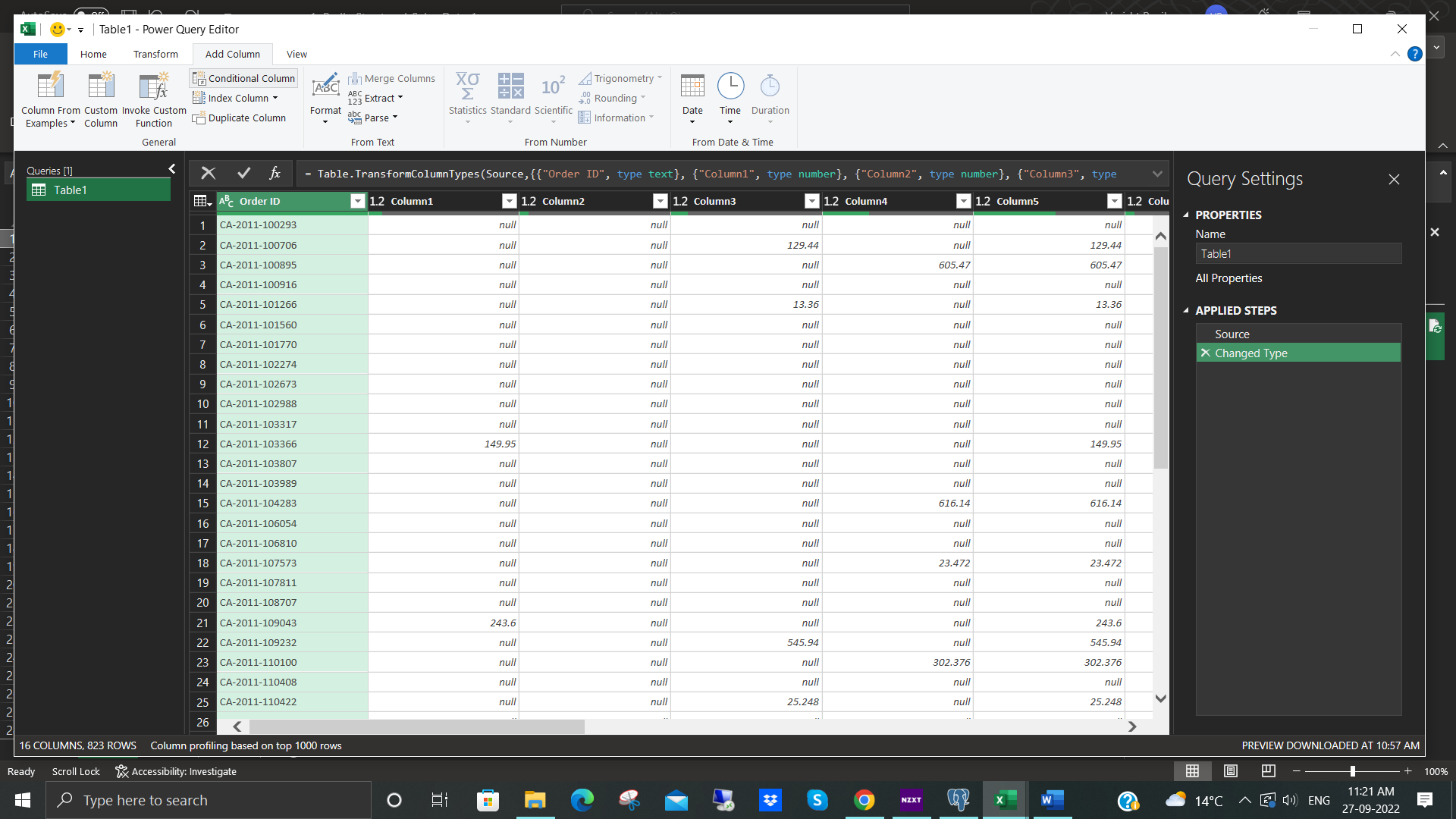


Figure 2. Power Query Editor as opened from the existing range from Figure 1

The Power Query Editor window opens with the selected range in the form of a table and the same range is also opened as a table in a new sheet. The changes done in the power query editor gets applied automatically to the new table created from the range.

The headings are changed from generic Column1, Column2, etc., to something meaningful, in this case, the **Ship Modes** for each column. The Grand Total columns copied from the original data is deleted inside the Power Query Editor as those columns aren’t necessary for this table formation. The separate Grand Totals can be obtained by simply adding a Pivot Table to the final cleaned table.

Using the **Conditional Columns** from the **Add Columns** ribbon in the Power Query Editor, new columns can be created fulfilling the necessary conditions. To achieve this, the null data from each column are kept as null and wherever there is a sales figure, those columns are renamed into their respective **Ship Modes** and **Segments,** in two separate columns**.** The dialog box contains drop down list of all the columns created with an option of IF ELSE option. This works like an OR operator where even if one of the values are not null, then the respective Ship Mode or Segment is printed in its desired column. The same method is carried out twice, once for the Ship Modes and once for the Segments. Please refer figure 3.

Once these steps are done, the Power Editor changes will be applied to the table after clicking the option **Close & Load**. This brings all the changes to the table created in two separate columns.

For Ease, the Order ID from the unaltered data is copied into a New Sheet. The column headings which were converted into values in the rows using power editor, can now be copied into the new table. Once this step is done, there is only one more data to be added to the newly created table to make this a completely cleaned data.

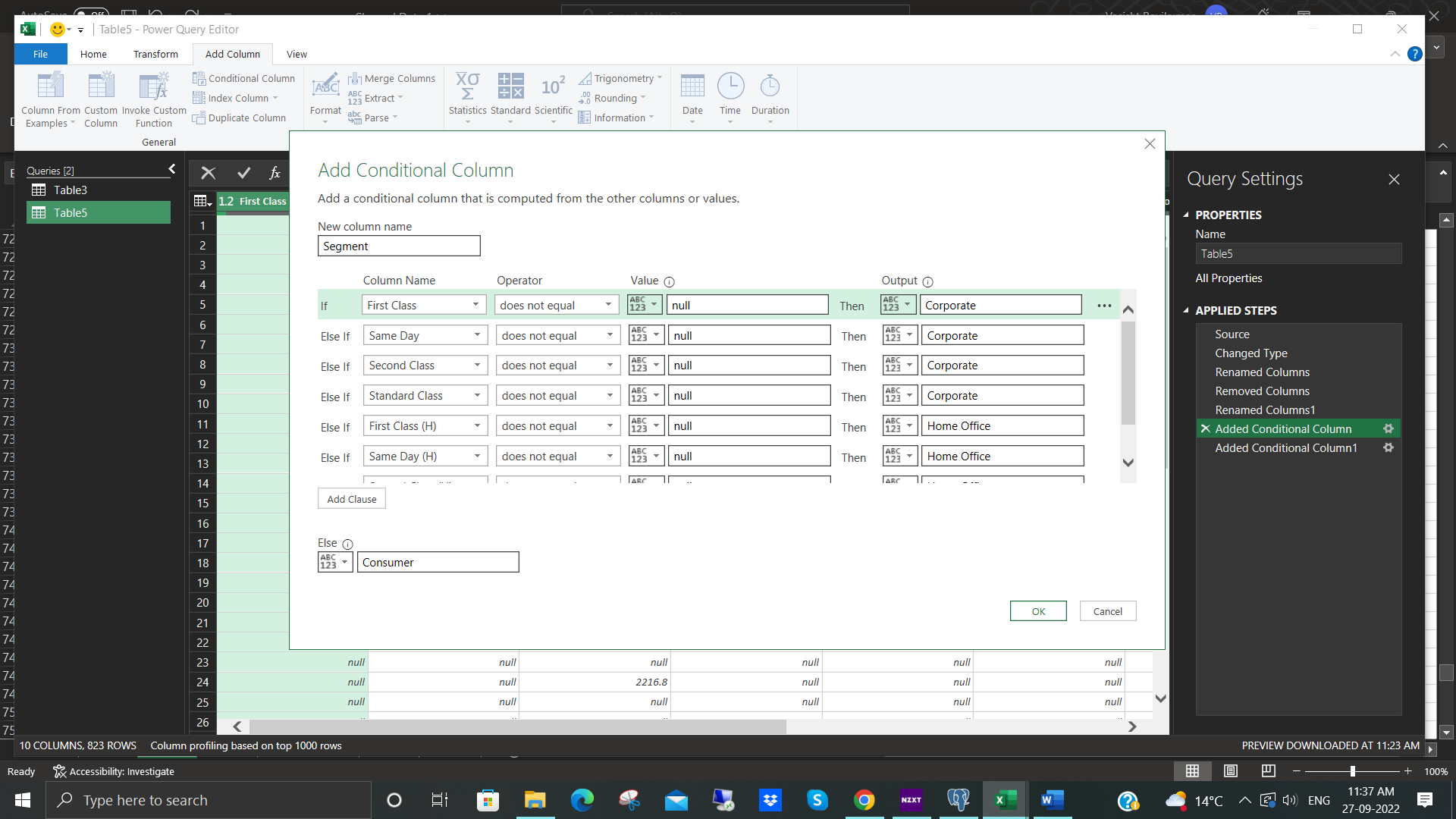


Figure 3. Adding a conditional column for each **Segment.** If a value in a column is not equal to null, then the output would be the respective **Segment** value of thatcolumn**.**

To add the sales figures, copy the data from the old sheet to the new sheet. Now, it’ll be divided among 15 different columns. We can erase the **Grand Total** Columns as it is unnecessary for this tidying up. Once those are deleted, we’re left with 12 columns, each column having the **Sales** data for each **Ship Mode** per **Segment**. These can all be combined into one single column using concatenation.

The formula that was used to arrive at this was =CONCAT(D2:O2). This joins all the data together. There wouldn’t be any null value in this case as the type of **Segment** cannot be two of Home, Corporate or Consumer at the same time. They must be mutually exclusive. Also, one **Order ID** cannot have orders from two different types of Orders nor can have different **Ship Mode**. Hence, it is safe to Concatenate all these values without joining 2 or more numbers together. In all these cases, if there Is a value in one of the categories, it is null in all the other categories.

Once Concatenated, the values were copied and pasted special as Values to get a number inside each cell instead of the Concatenation formula. Then number formatting is done as currency to get the Sales figures in dollars.

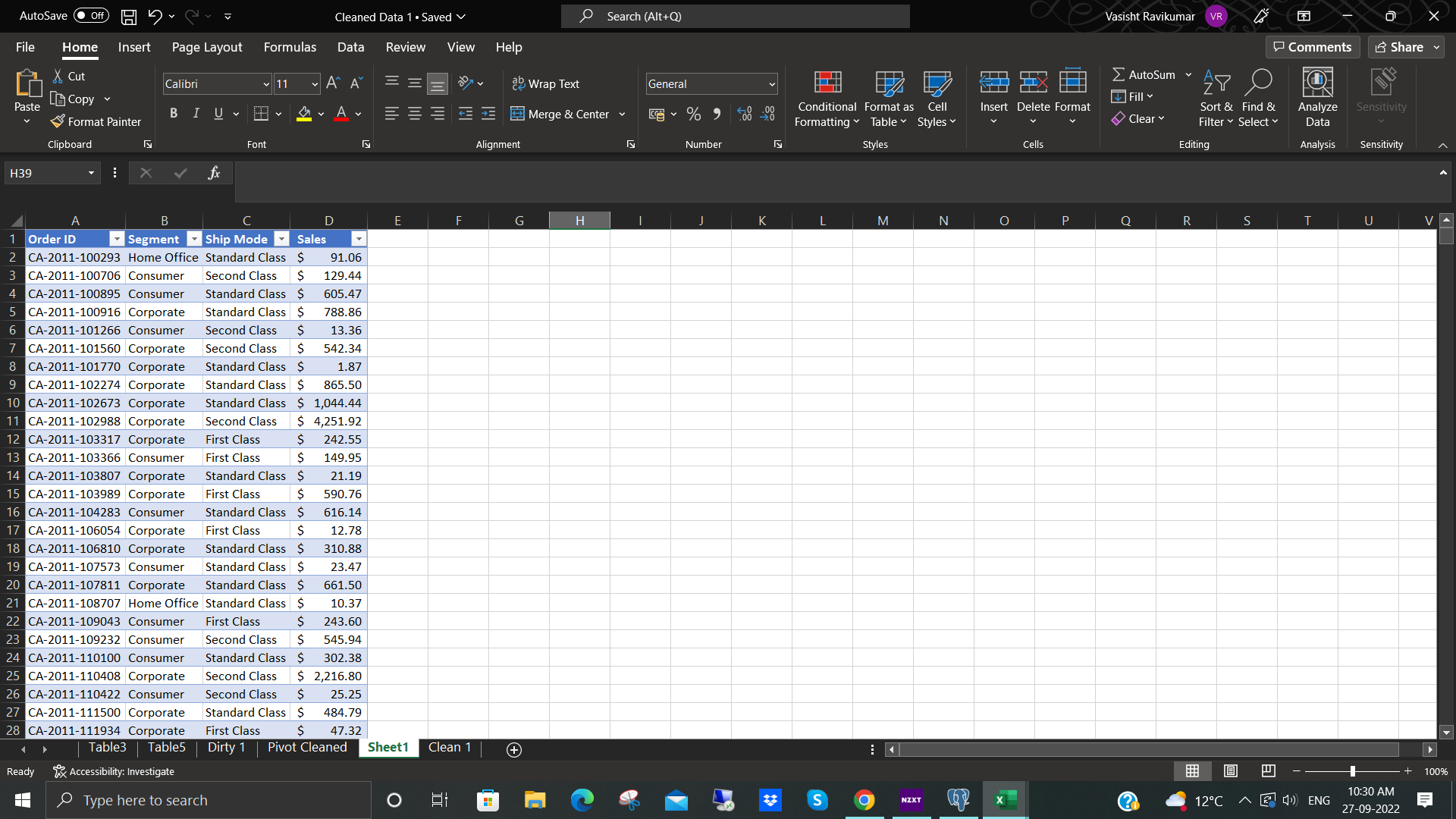


Figure 4 Final Cleaned Data in a table format

The final data looks clean, and the data can be easily read and converted into charts or graphs for more analysis. It can be seen in Figure 5.