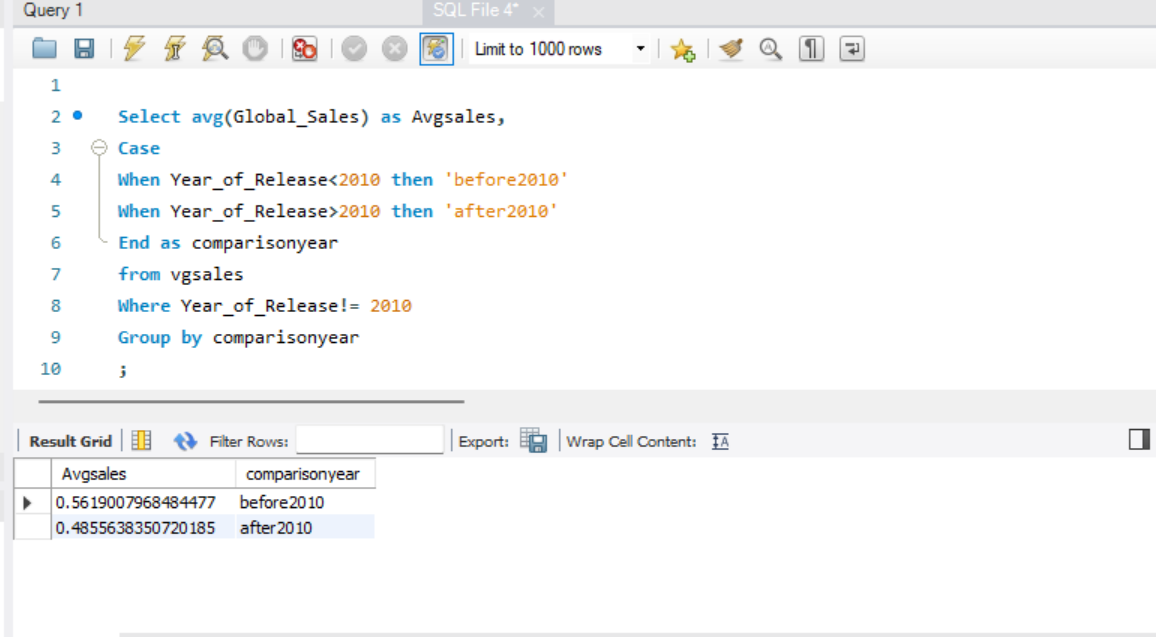
Data 1202 DATA ANALYSIS TOOLS

Data Transformation in SQL

We proceeded through a series of processes to arrive at the correct end solution. Importing the massive dataset of 16720 rows was the first stage.

**Question Number 1** : The goal was to determine whether the average sale was higher before or after 2010.



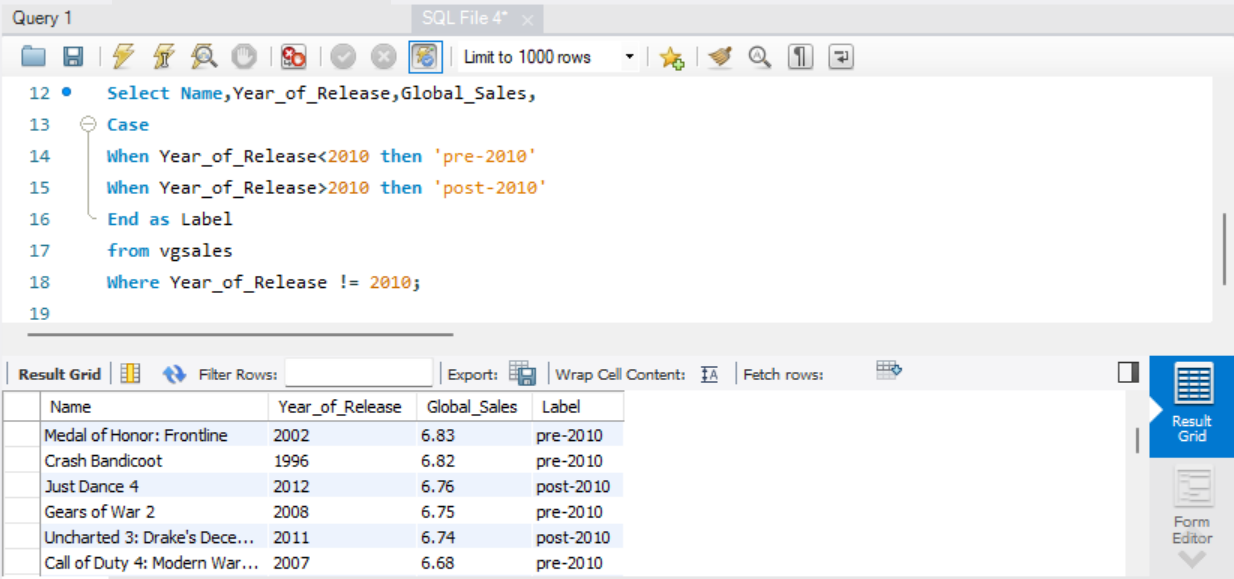
The following are the procedures we took to get at the solution, during which we encountered various technical issues due to a lack of coding knowledge :

1. **Dividing the dataset into two parts** – We distributed the data set into two parts, one for “Before 2010” and another for “After 2010”. This step was done using the Case-When clause. We added a new column that says Before 2010 if the year of release was before 2010, and After 2010 if it was after or in 2010.
2. **Calculating the average of Global Sales** – Specifically for the average, before and after the year 2010 which was done using the function AVG() for which we have to take some assistance from w3schools to ensure the codes worked properly.
3. **Calculating the average separately** - Figuring out how to calculate the average of the entire row individually before and after 2010 was one of the most difficult tasks. We eventually obtained the answer using the Groupby clause after attempting many techniques. This clause grouped the data rows depending on the new column we created. This column contains data for just two values: Before2010 and After2010.

With all these three steps, we were able to display the average sales separately before 2010 and after 2010 as a clear comparison.

To answer the first question, the data show that the average global sales before 2010 were **HIGHER** than average video game global sales after 2010.

**Question Number 2** : We are required to create a new column with two values representing data from before and after 2010.



After solving question 1, this section was much easy to complete because we had previously done so in step 1. The Case-When clause was used to bifurcate the whole data set based on the year of the release, and the column was renamed "Label."

The only issue we had was that we didn't know what to do with the rows with the year of release 2010 because the data necessary was before and after 2010. Because there was an obvious comparison between the data in 2010 and the data in other rows, the new column displayed the value "post-2010" by default for these rows.

After addressing our uncertainties with the professor, we used the "where clause" to eliminate all the rows with the year of release as 2010.

Finally, we came up with the data that shows the year before 2010 and the year after 2010. After running the script, it displays the name of the video game along with the year of release, with pre-2010 indicating a release before 2010, and post-2010 indicating a release after 2010.