# Brief

I am investigating Video Game data to take a deep dive into video game playtime statistics, as well as review data. My goal from this investigation is to understand how playtime statistics have changed over time, as well as public opinion and reception to video games. Furthermore, I am curious in sales trends, and wish to briefly touch on them. The information I hope to learn will help me make educated assumptions about the video game market, player retention and statistics, and the ways these have changed over time.

# Steps For Cleaning

Step 1: Import data

Step 2: Run the following code to create new columns based off of the Variable column, using each distinct value from Variable to make a new column, and populate it with the corresponding data from Values:

let

Source = YourPreviousQueryName, // Replace YourPreviousQueryName with the name of your previous query where you loaded the data

Pivot = Table.Pivot(Source, List.Distinct(Source[Variable]), "Variable", "Value")

in

Pivot

Step 3: Use the GroupBy function to collapse all the duplicated entries for games into a single record, as currently there are many duplicates for a single game, each corresponding to a value in the newly formed columns as seen below:

A screenshot of a computer

Description automatically generated

To start with, the String columns need to be corrected, so that is Genre, Publisher, and Developer. This corrects any potential errors cropping up with the group by function when it aggregates the data later on. Once the data is in the correct format, the group by function can be run, a snippet of the correct data types can be seen below:

A screenshot of a computer

Description automatically generated

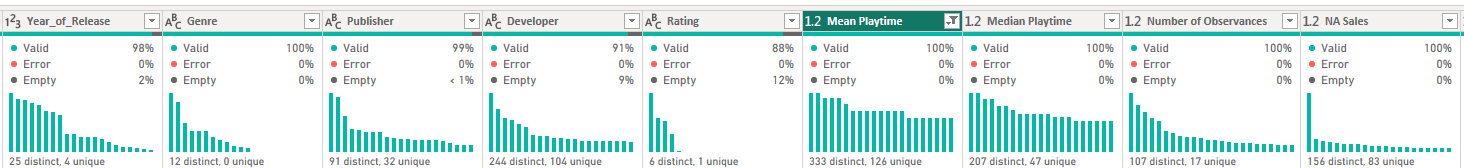
Now, group by can be run, which condenses all the rows under game name, platform, and year. The result of this is like so:

A screenshot of a computer

Description automatically generated

Step 4: Next, errors can be removed that were caused by the type conversion, specifically errored rows that contain a value of TBD from the User Score. By filtering these out, we have a mostly clean dataset now.

Step 5: Finally, the data can be inspected for missing rows and null values. As I am mostly interested in playtime statistics, I will filter out empty rows that lack data for playtime. With that done, there are still a few empty values here and there in the dataset, but overall, the data is now cleaned, as can be seen from the column quality and distribution headers:



A screenshot of a computer

Description automatically generated

Additional step: Convert the Year of Release column to String, and then to Date, so that it can be used as a Date value for figures later on:

