**Video games sales analysis**

**Introduction of the data set:**

The data set contains over 16,500 games data from the year 1980 to 2020.

**Let’s start the analysis:**

We will conduct this analysis by follow the steps of the google analytics course, which are:

* ASK
* Prepare
* Process
* Analyze
* Share
* Act

**Phase 1, Ask:**

**Questions for Data analysis:**

|  |  |  |
| --- | --- | --- |
| **Data analysis Questions** | | |
| **Q.NO** | **Overall** | **2019** |
| 1 | Genre with the most sales | Which published has the more sales in 2019 |
| 2 | Maximum number of gameon each platform | Top sold game in 2019 |
| 3 | Platform having most sales | Most sales in North America in 2019 |
| 4 | Number of games released by publishers | Most sales in Europe in 2019 |
| 5 | Are games being made increasing per year | Most sales in Japan in 2019 |
| 6 | What are the most popular platform per region | Top selling games of each publisher in 2019 |
| 7 | What is the most popular publishers per region | Most successful platform in 2019 |
| 8 | What is the most popular genre per region | Which genre has the most sales in 2019 |
| 9 | Bestselling games per year |  |
| 10 | Explore top publishers |  |
| 11 | Trend in the preferred genres from 1985 - 2020 |  |
| 12 |  |  |
| 13 |  |  |

**Phase 2, Prepare:**

* This data set was located and download from kaggle. Here is the link

<https://www.kaggle.com/datasets/gregorut/videogamesales>

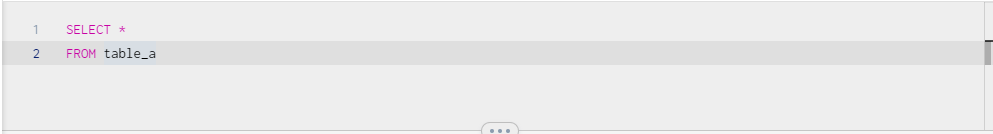
* This data set has information about, the games, publisher, year of release, North America’s sales, Europe sales, other region sales and Global sales, which will help bring insights to the questions in the ask phase.
* This data has no licence but the common licence of [Attribution 4.0 International (CC BY 4.0)](https://creativecommons.org/licenses/by/4.0/) states that data set can be adjusted, tampered or modified according to the need. Only citation of the data set source is required.
* The data provided in the link is clean, but for the sake of showing skills in this portfolio, it was made unclean. Following components were added to make the data unclean:
  + Extra rows (done)
  + Duplicates (done)
  + Extra spaces in cells contains strings (done)
  + Missing Values (done)
  + Alignment wrong. (done)
  + Splitting of columns to another sheet for SQL joining. (done)
* For the sake of showing skills, I have split the data set into two tables and uploaded them in [dataworld](https://data.world) data storage, so that I can demonstrate some SQL skills in retrieving the data. The name of the two tables are table\_a and table\_b.

**Using SQL to retrieve data from the dataworld database:**

Let us start the process by locating the video game analysis dataset in the dataworld database. After locating the dataset. Here is the process:

* The Schema of the dataset shows that data sets shows that there are two tables, **table\_a** and **table\_b.**
* **table\_a** has **7 fields** and **16,804 records**.
* **table\_b** has **5 fields** and **16,804 records**.
* Using the following command to inspect table\_a:
* Select \*
* From table\_a

**SQL Query console:**

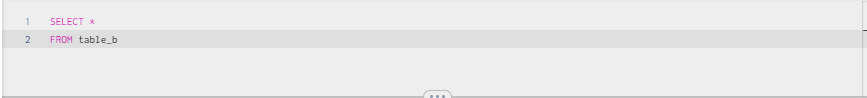


**Result:**

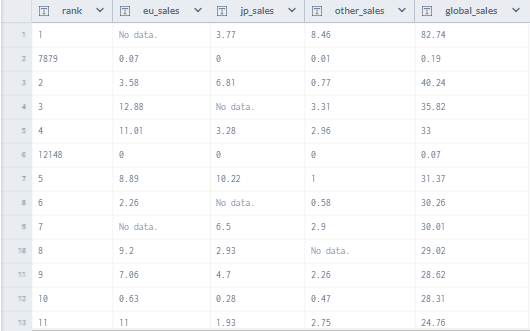
****

* **Using the following command to inspect table\_b:**
  + SELECT \*
  + FROM table\_b

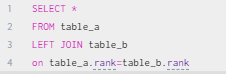
**SQL Query console:**



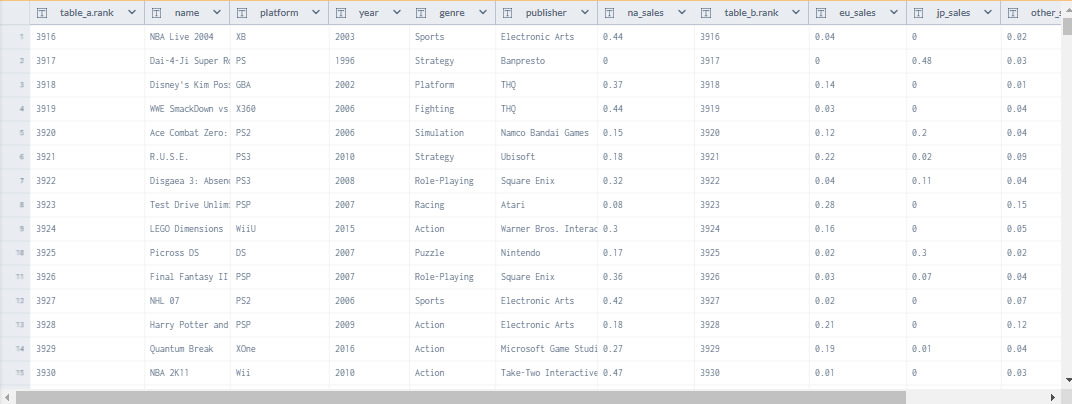
**Result:**



* By observing the fields name of both the tables, I have noted that there is a field called rank which is present in both the tables. With the help of SQL Join query, I will combine the data set of both the table with this common rank field.
* Using join query to combine the data of both the tables in a single table. Now there are multiple join methods to join data but for this analysis, I want all the data of the left table which is table\_a and the data of table\_b, which matches with table\_a. So for this purpose, I will use the left join query.
* **Using Left Join query to combine data of 2 tables in a single table**:
  + Select \*
  + From table\_a
  + Left Join table\_b
  + On table\_a.rank=table\_brank



**Result:**



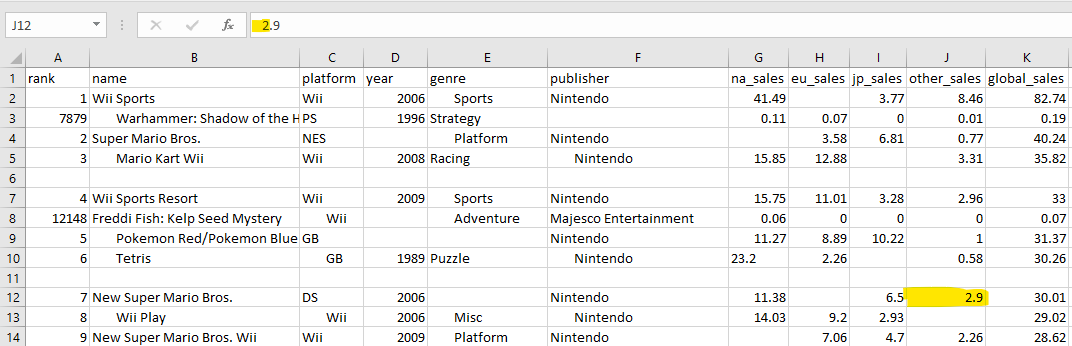
**Phase 3, Process:**

Now that the data is selected and retrieved, it’s time to clean the data by making it ready for analysis. To make the data clean, a data analyst must look for the following things in the data sets:

* Proper structure formatting.
* Extra spaces in the cell.
* Removing empty rows within the data
* Filling missing values.
* Duplications
* Exclude any irrelevant data.

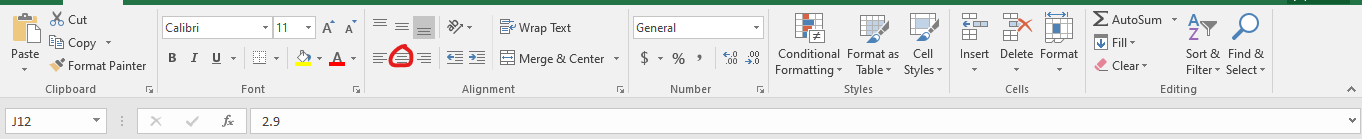
**Proper structure formatting.**

By inspecting the data sheet, I noticed that the alignment of some data in the rows is not correct. There are some cells that have extra spaces within the cell but there are some that are not properly aligned. The image below shows that the cell value is not properly align with the rest of the data set and no extra space is present because the formula bar shows.



Let’s align it this for it to look clean.

* Select the whole data set with CTRL + A and click on the centre align in the button in the home tab of the ribbon.



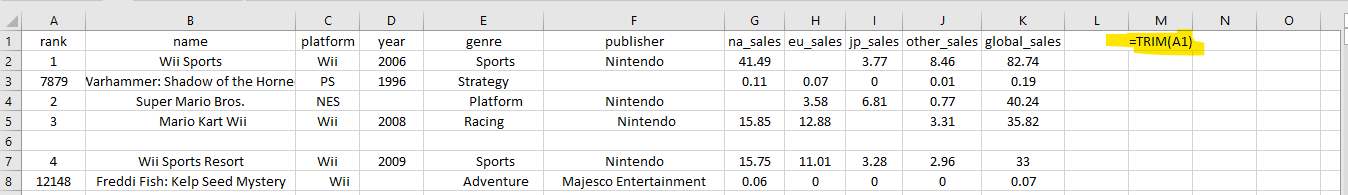
* This method will present the data in centre alignment and as a result will look clean.



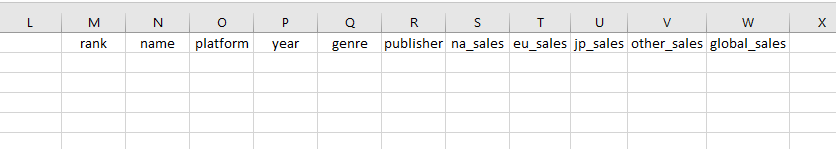
* Now if some cells are still not properly aligned, this means that there are some extra spaces in their cell. These spaces can be removed in the coming method.

**Remove Extra spaces in the cells:**

* Extra spaces in the cells of the data can be remove with the trim function.
* At right side of the Headers in row 1, in the cell M1, insert the trim function with the cell A1 as a relative reference

****

* This will make the text “rank” appear without any extra spaces, now click on the fill handle of cell M1 and drag it 10 columns, right side in row 1 because the total columns of this data set excluding the rank column is 10.

****

* Now I will select all the column headers in row 1 and then I will select the fill handle which can be seen at the bottom right of cell w1.After selecting the fill handle of Cell W1, I will drag it to where the data set ends. This method will fill the new rows with the data without any extra spaces.
* Now I need to replace the data having extra spaces with the new extra space-less data. This can be done as follow:
  + - Select the new corrected data.
    - Copy it with CTRL + C
    - Highlight all the data set which have extra spaces
    - Press right click and paste the data as values on the erroneous dataset.
    - This will replace the unclean extra space data with the relatively cleaner version of it.
* Now our data set is properly aligned and have no extra spaces.

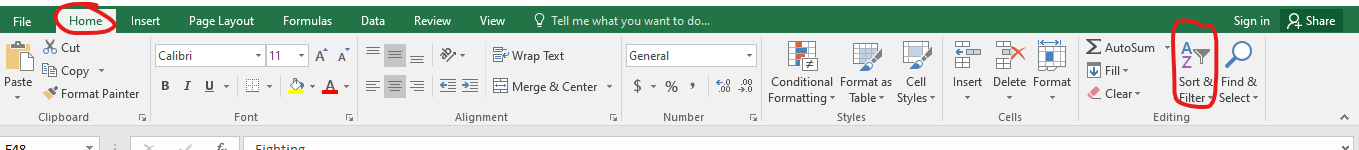
**Result:**

****

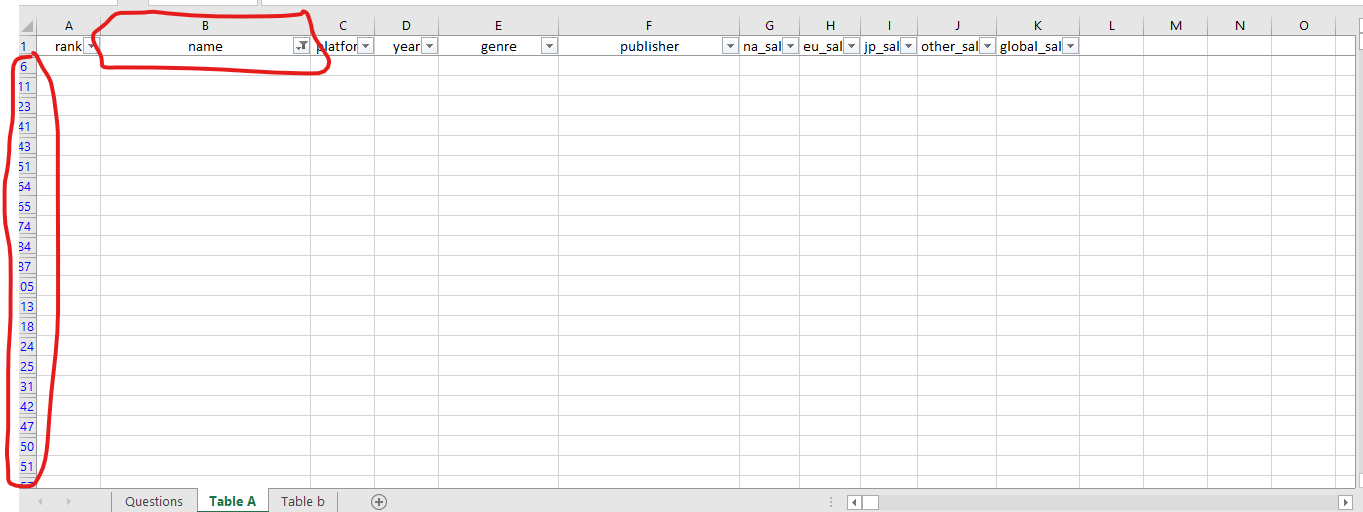
**Removing extra empty rows within the data set.**

Our next objective is to look for extra empty rows with in the dataset, and by skimming through the data I have found that there are some present. These extra empty rows can be removed with two methods.

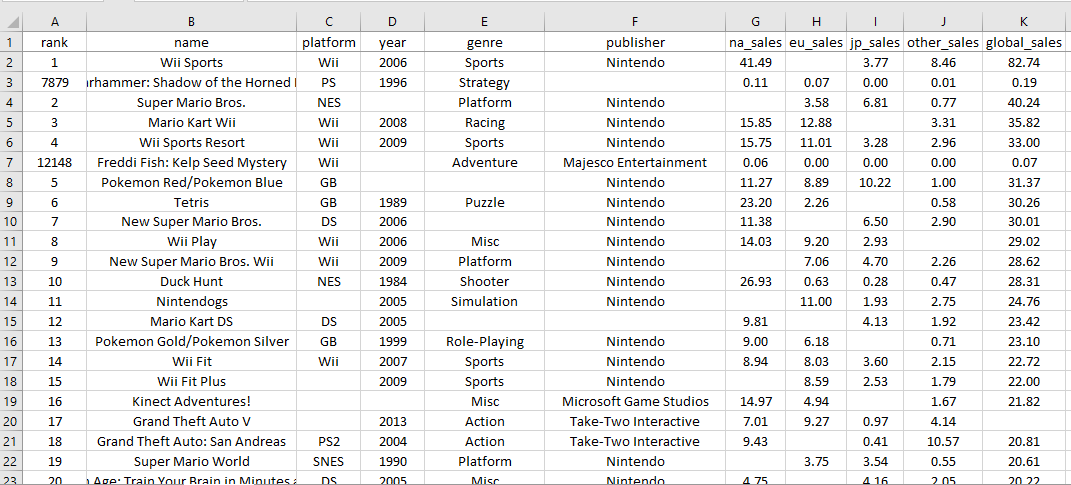
* **Using filters**:
  + First select the whole data set.
  + In the Home tab at the top, find the Sort & Filter button and click on.



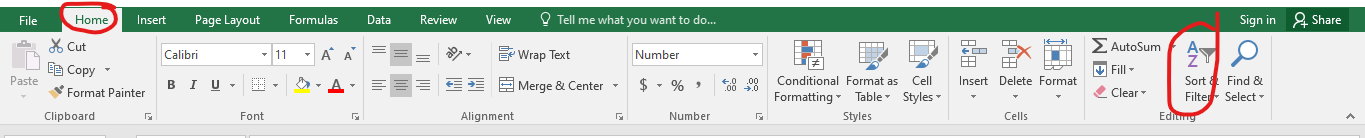
* + Small arrows will appear beside in the header’s cell. That is the filter option.
  + You can apply filter on any column, I chose to apply a filter on the name column and in the filter option I selected only blank cells. This will hide all the filled rows and only the rows that are empty will appear.



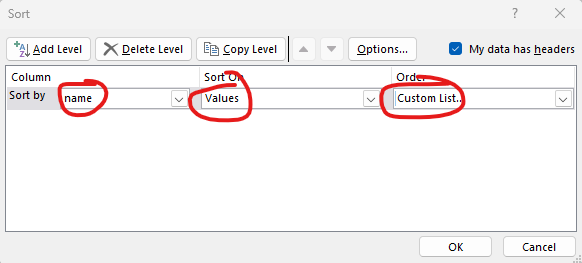
* + The numbers on the left hand side shows the specific row which is empty and is present in the middle of the data set.
  + Select all these rows and right click on the row section and press delete.
  + This will delete all the empty rows that are present in the middle of the data set.
  + Now remove the filter by clicking the Sort & Filter button again.
  + After this the rows that have data in it will appear and the extra empty rows will not be visible.

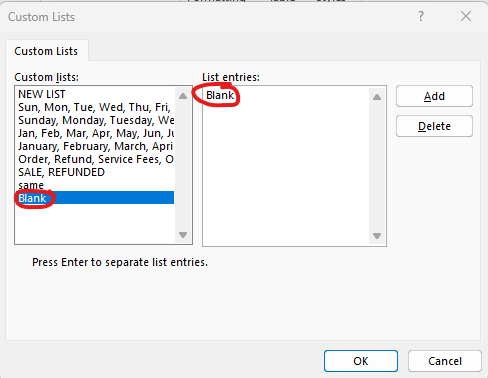


* **Using the sort method**:
* First select the whole data.
* Then Click on the Sort & filter option in the home tab and click on Custom Sort.

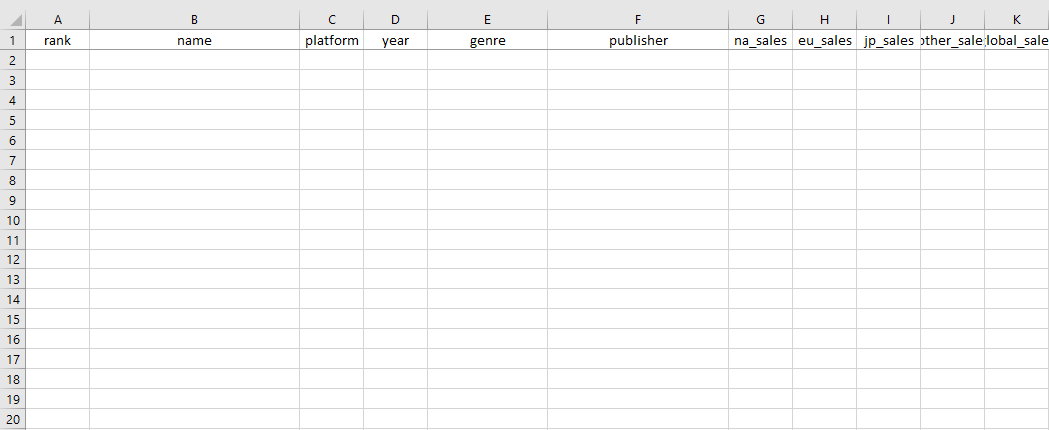


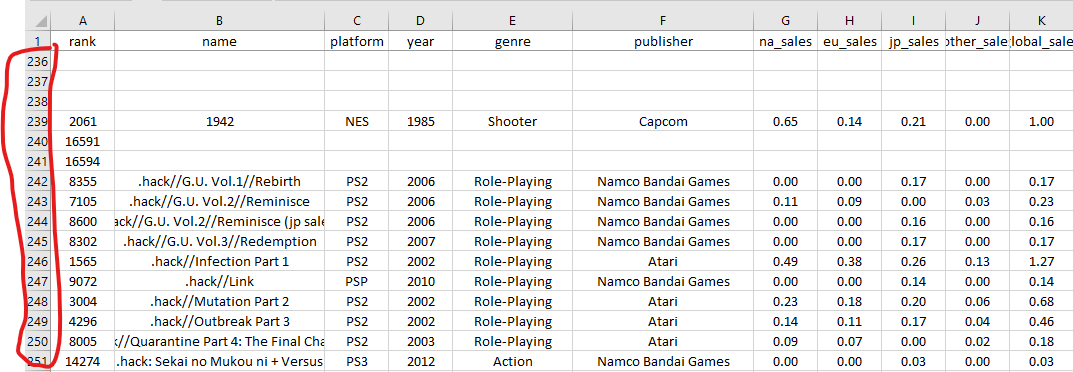
* In the custom sort window, I chose column name to sort by, then sort on values and in orders field, I chose custom list.



* In the custom list window, I added a new entry called Blank. This entry will make all the blank rows appear in the ascending order.
* 

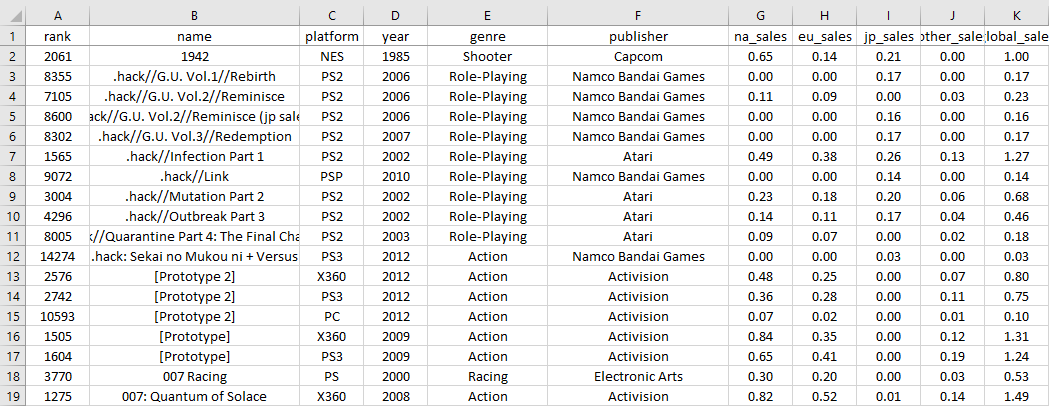
Result:





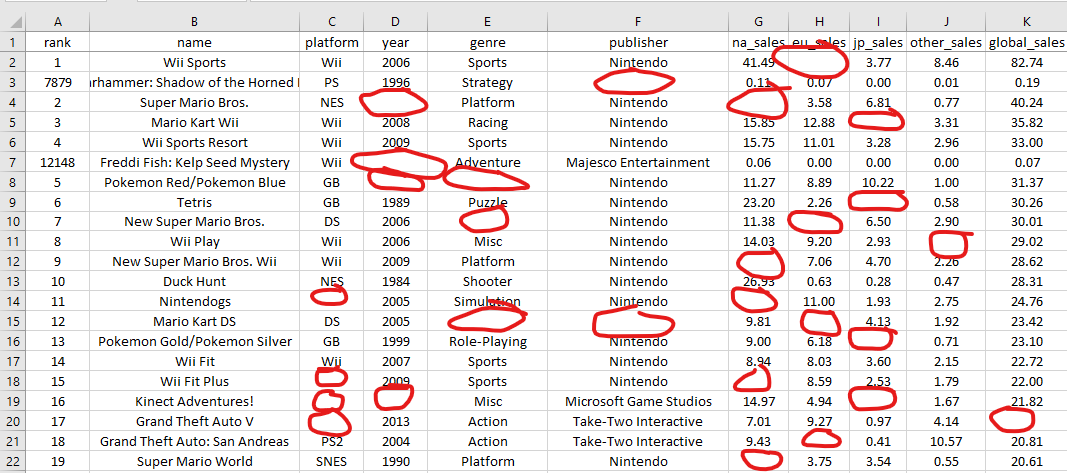
* The cell which have data present in it are arranged in the bottom.
* Now remove the empty rows that are arrange in the ascending order. This will remove all the empty rows.

**Final result:**



**Filling missing values**:

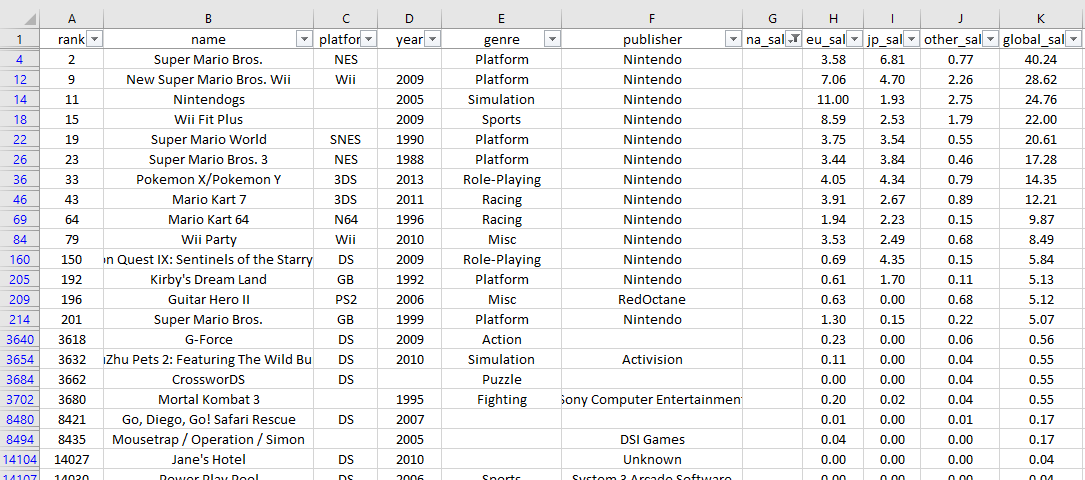
By observing the data, I have noticed that there are some empty cells in rows.



Judging by the data these empty cells can be fixed through excels arithmetic formulas and also some internet research. To make this process easy, I will fill these cells with the help of filter option.

**Using filter to show only empty cells:**

* First, I will fill the sales empty cells.
* Apply a blank filter on na\_sales column. All the empty cells in the na\_sales column will appear.

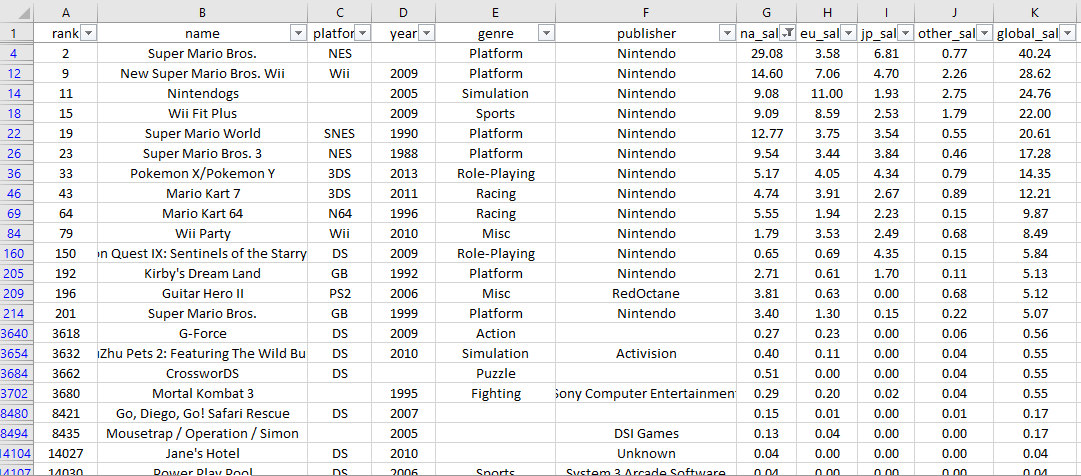


* Now use excels arithmetic formulas to fill these cell. I will use start will Cell G4 and apply the formula:

=K4-(H4+I4+J4)

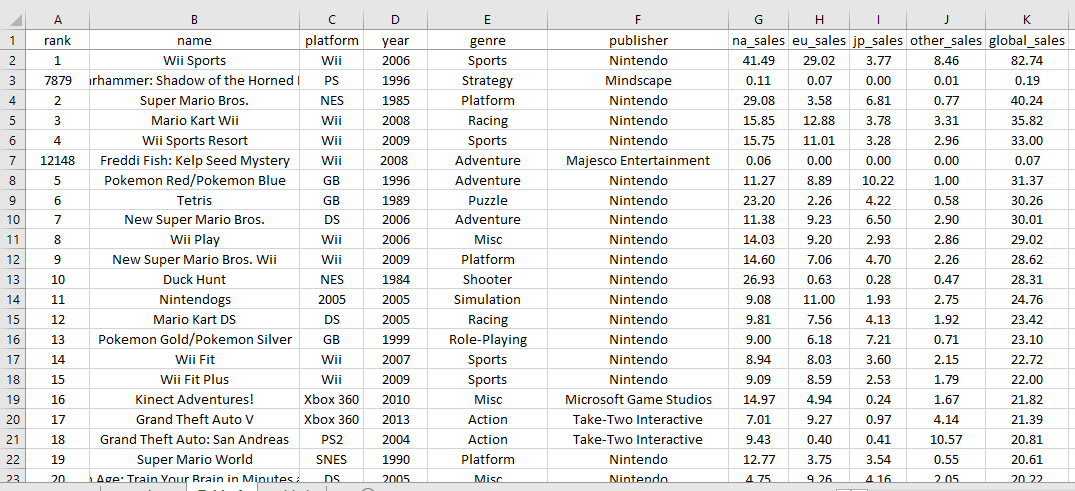
* Now that the first value of na\_sales have appeared I will drag the fill handle and this will fill the all the empty cell in the na\_sales column.

Result:



* Now use this method with the rest of the sales column.
* After filling the sales column, I searched on the internet to fill the missing empty cells in the platform, year and genre column.

**Result:**



**Duplication:**

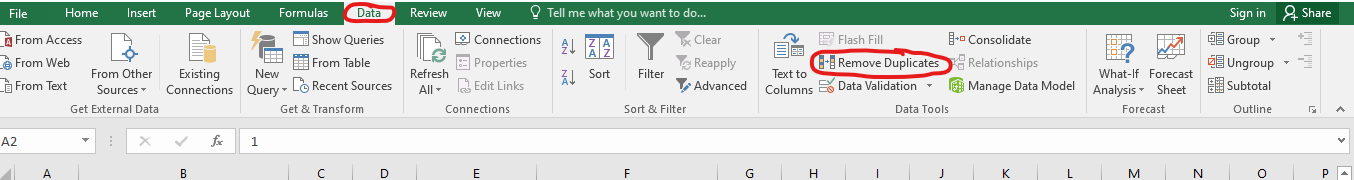
Duplications can be removed from a number of methods.

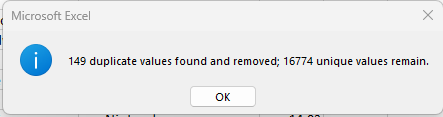
**Microsoft Excel:**

In Microsoft Excel, you there also many ways to remove duplicates.

**Using the remove duplicate button**:

This is the easiest method of removing duplicates in your dataset. You select the entire data with Ctrl + A and then go to the data tab, there find and click on the remove duplicate button. This will remove duplicates from the data sets.





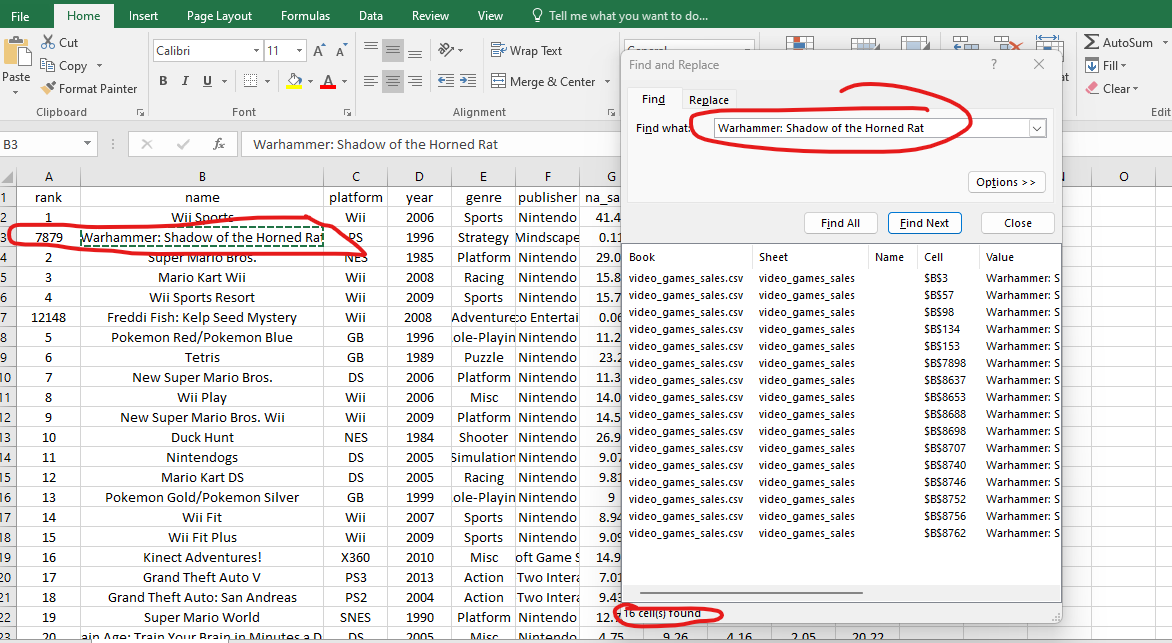
**Using SQL to remove duplication:**

SQL make very easy to query for a data set that have no duplication. This can be done with the SQL’s Distinct statement. Here is the procedure:

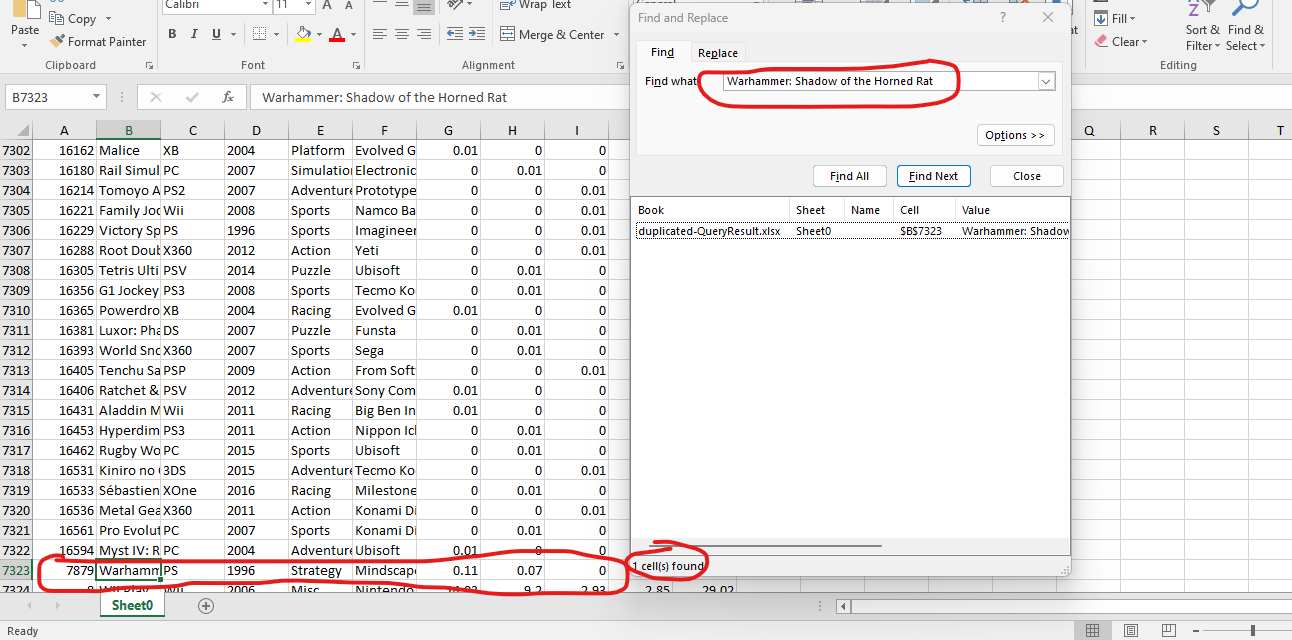
* I will upload the data set sheet to [data.world](https://data.world), which contain duplicates.
* Then I will use the following SQL statement to query for a data set that have unique values.



Before:



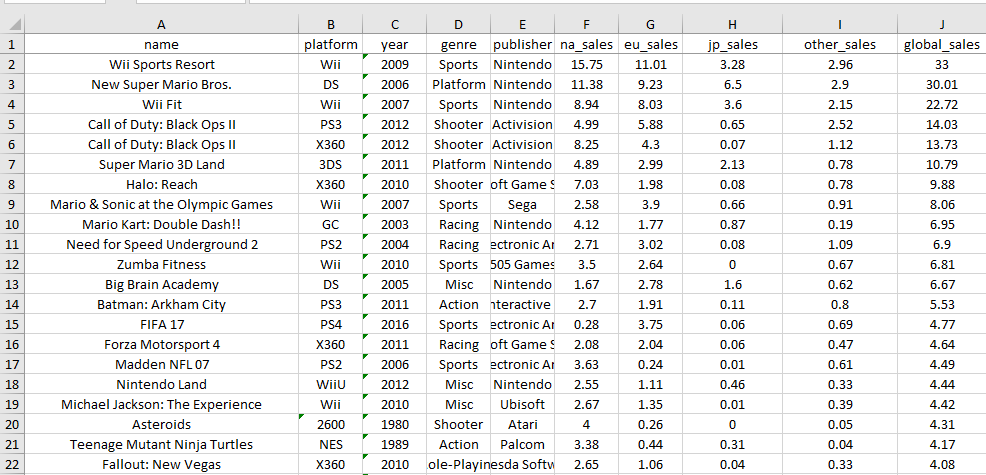
After:



**Excluding irrelevant data:**

It is a good practice to exclude irrelevant data from the data set. In this data set, I have noticed that the column rank is relevant to my analysis. So I will exclude that column and after this our data set will become clean and ready for analysis.

**Final look of the data set:**



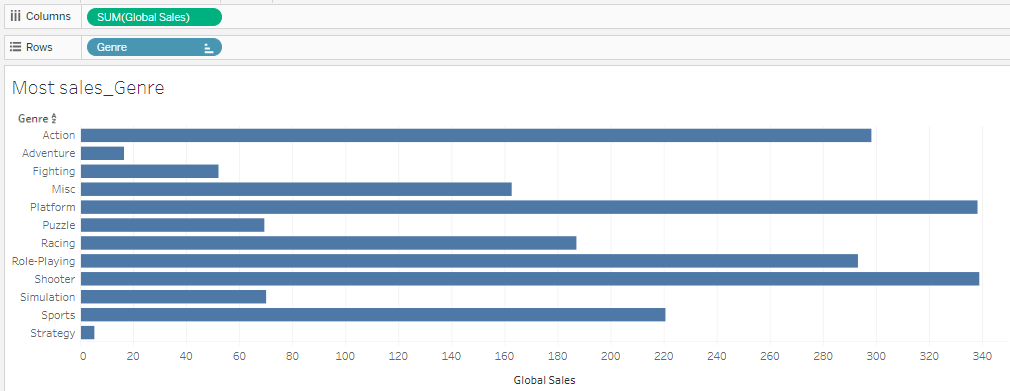
**Phase 4 Analyse:**

For Analysis of the data set to solve questions which were formulated in the Ask phase, I will be using Tableau public to develop insight.

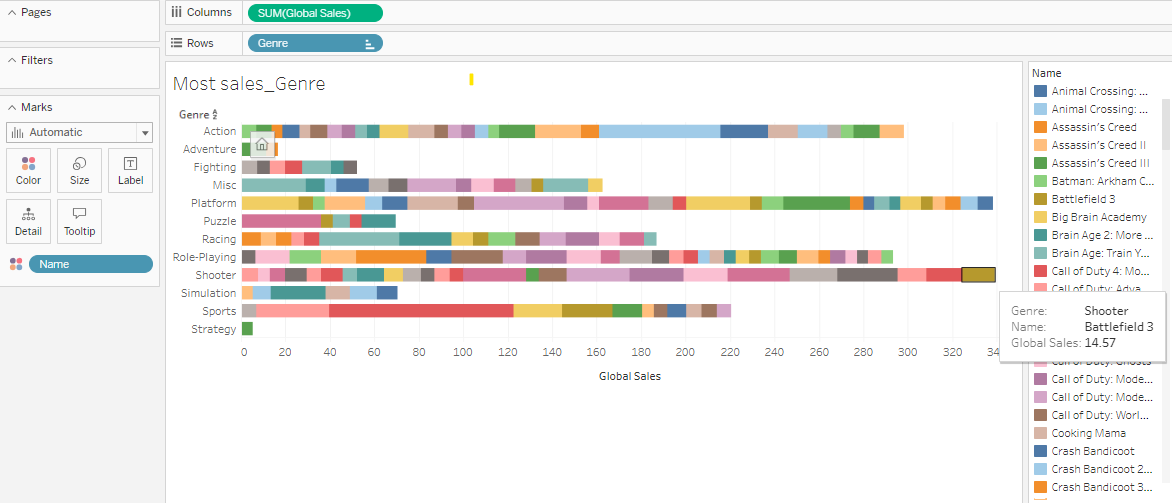
|  |  |
| --- | --- |
| **Data Analysis Questions** | |
| **Q.NO** | **Overall** |
| 1 | Genre with the most sales |
| 2 | Maximum number of gameon each platform |
| 3 | Platform having most sales |
| 4 | Are games being made increasing per year |
| 5 | Bestselling games per year |
| 6 | Trend in the preferred genres from 1985 – 2020 |

Question 1:

**Genre with the most sales:**



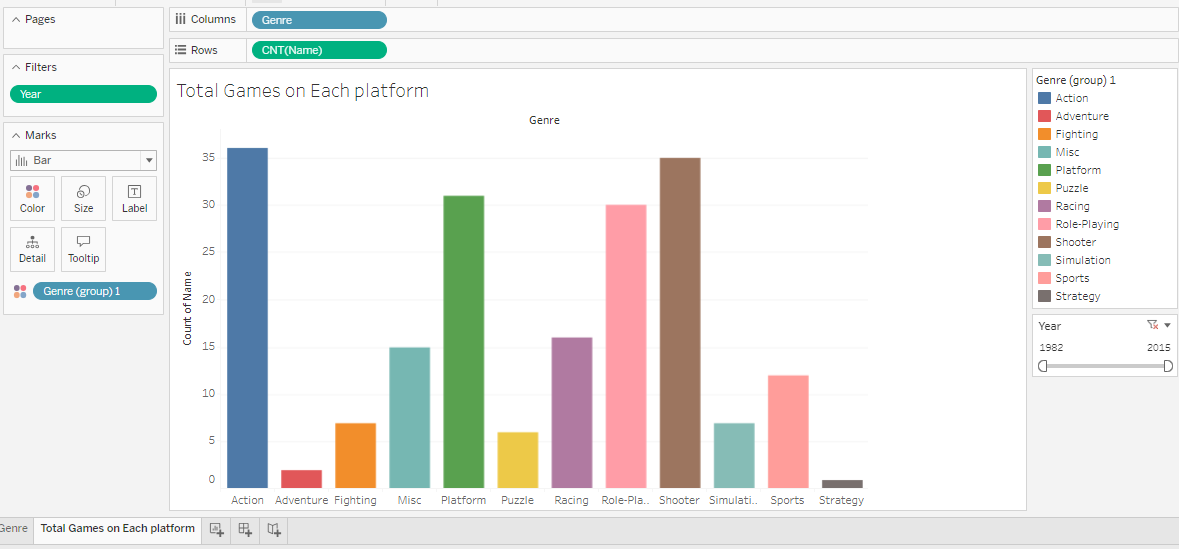
* Judging by the column graph, it appears that the genre with the most sales is the shooter genre.
* We can also find the games with the most global sales with respect to their genre in this graph too.



* Now we can see the game with the most global sale with respect to their genre. If I hovers my mouse to the tip of the shooter genre bar. I can see that Battlefield 3 has the most global sale in this genre.

**Question 2:**

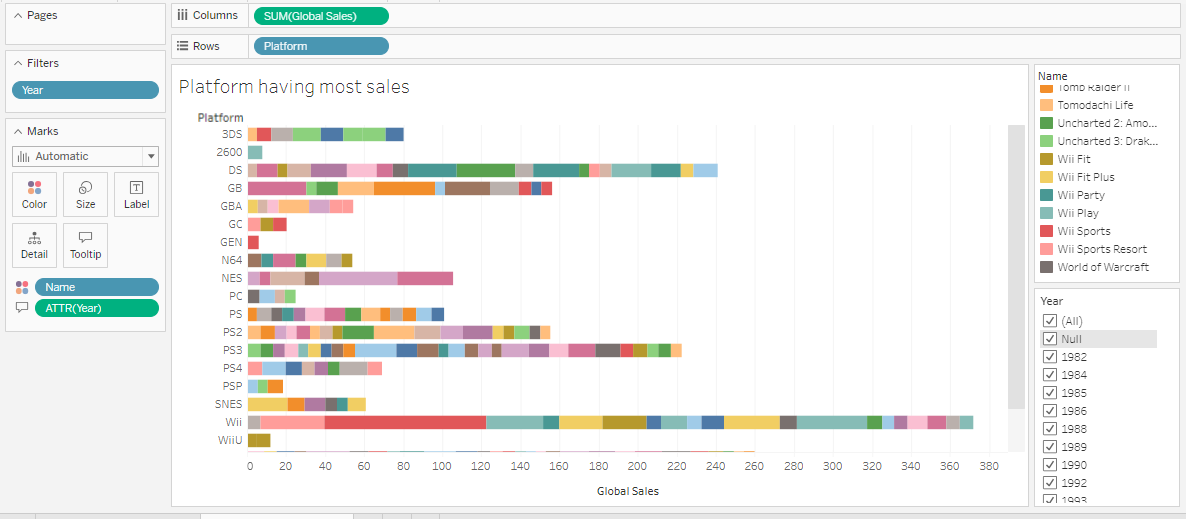
**Maximum number of games on each Genre:**

****

* According to the bar chart created, the Maximum number of games on each genre is the action genre, having 36 games from the period 1982 – 2015.
* This chart also has a year filter, which can filter the data according to the year.

**Question 3:**

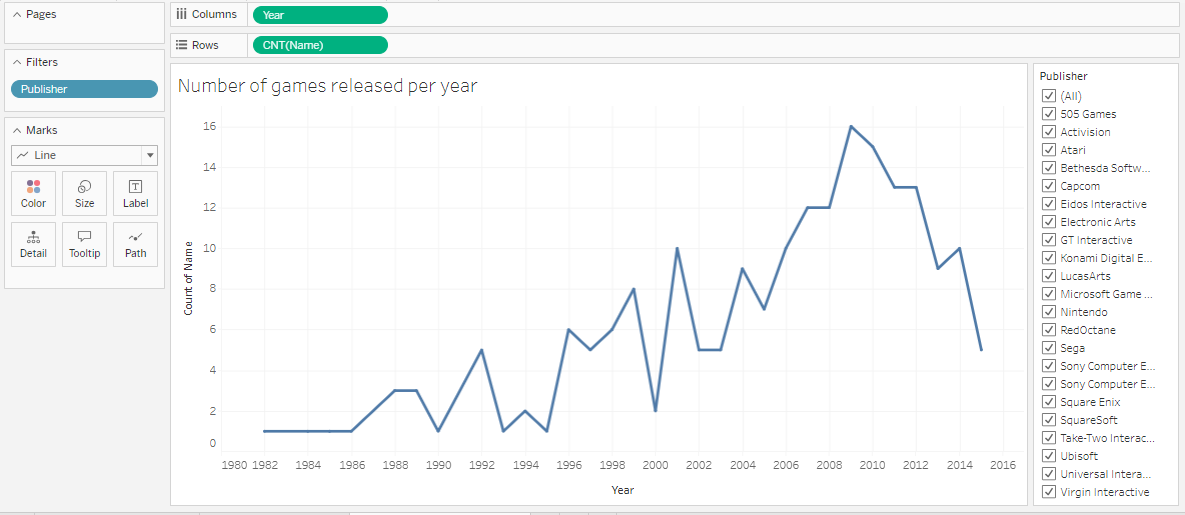
**Platform having most sales.**

****

* According to this column chart, the platform that have the most sales is the WII platform.

**Question 4:**

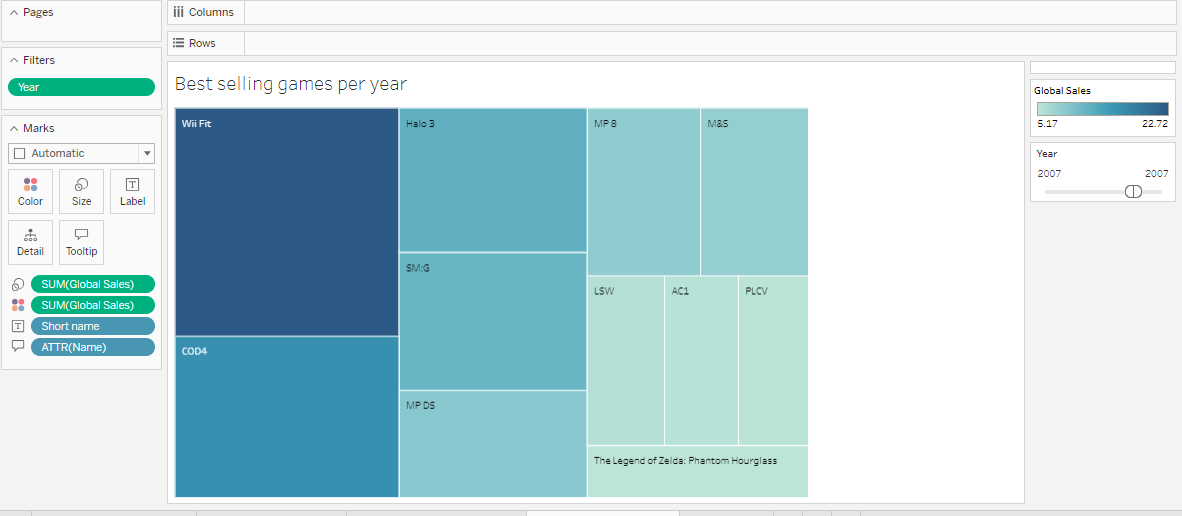
**Are games being made increasing per year?**

****

* By analysing this line chart, I came to know that the amount of games that have been released increased from 1982 to 2010, after that it’s a downward trend, meaning that the amount of games being made are decreasing.

**Question 5:**

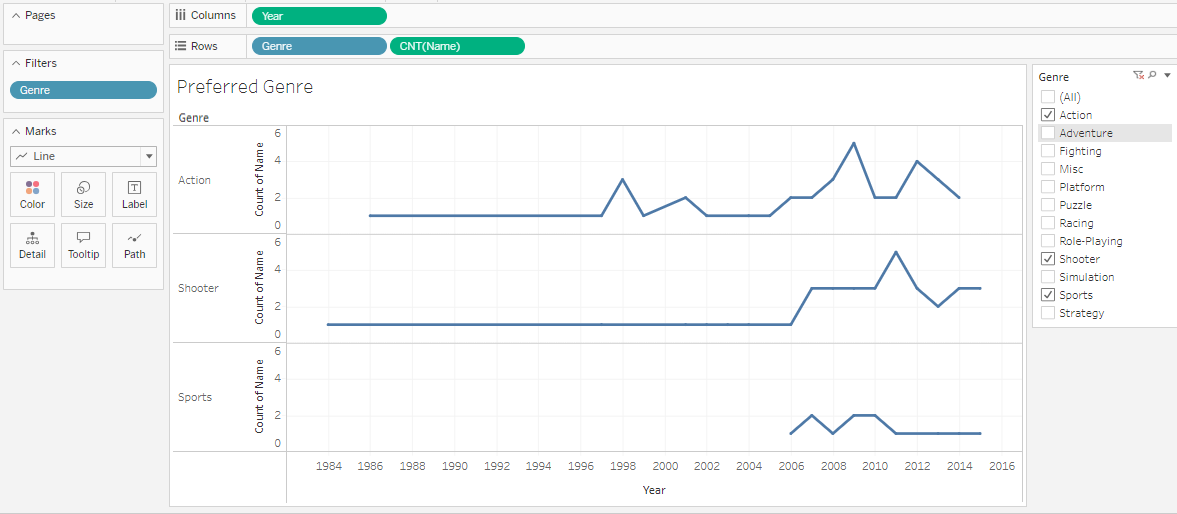
**Best-selling games per year?**

****

* By this heat map. It appear that for the year 2007, the best-selling game is WII fit, followed by Call of duty 4 and Halo 3.

**Question 6:**

**Genre preference from 1984-2015:**

****

* I have compared the action genre, shooter and sports genre with each other and the line graph show a trend in their preference throught the years.