PySpark - SQL

Student: Joaquín Bembhy

Professor: Danilo Ardagna

Course: Cloud Technologies & Big Data Frameworks

Dataset:

This dataset contains a list of video games with sales greater than 100,000 copies. The following data is provided for each game:

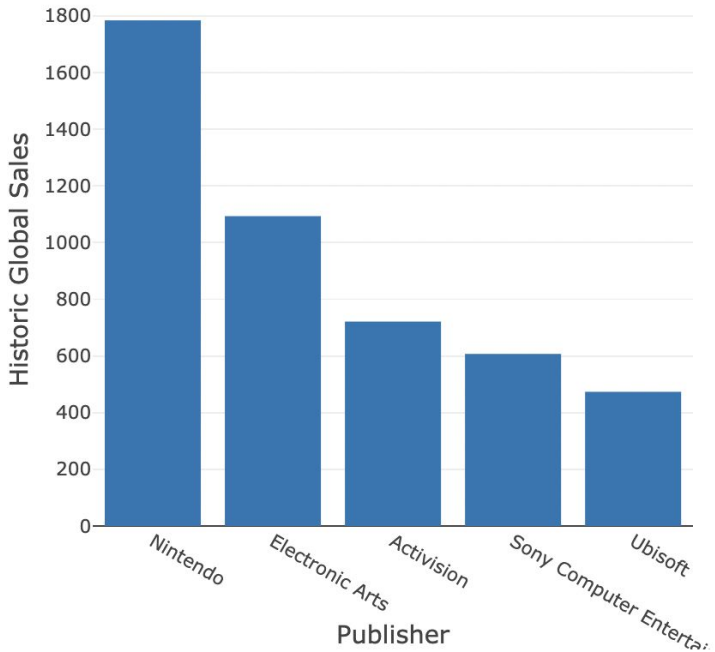
- **Rank** - Ranking of overall sales
- **Name** - The games name
- **Platform** - Platform of the games release (i.e. PC,PS4, etc.)
- **Year** - Year of the game's release
- **Genre** - Genre of the game
- **Publisher** - Publisher of the game
- **NA_Sales** - Sales in North America (in millions)
- **EU_Sales** - Sales in Europe (in millions)
- **JP_Sales** - Sales in Japan (in millions)
- **Other_Sales** - Sales in the rest of the world (in millions)
- **Global_Sales** - Total worldwide sales.

Pre-processing:

- Some of the column types were imported with the incorrect format, therefore, the type of this columns was modified.
- We also checked for NULL values, but the dataset didn’t have any.
- Now we have a clean dataset to work with.

Query 1:

- Calculate the 5 publishers with the highest historic ‘Global_Sales’
  - We want to view what are the publishers which have sold the most ($) in terms of global sales.

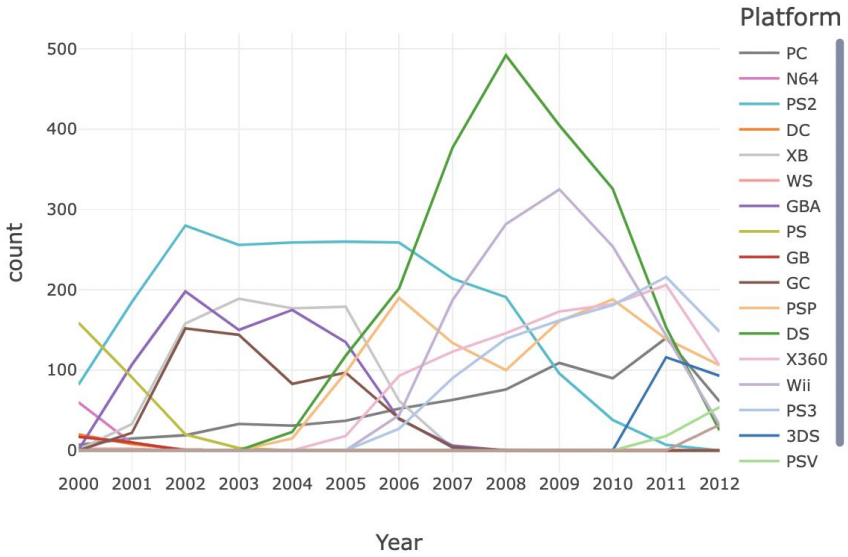


Insights:

- Nintendo leads historic global sales with almost 1800 million dollars, followed by EA, Activision, Sony and finally Ubisoft.

Query 2:

- Find how many games each platform released each year between 2000 and 2012

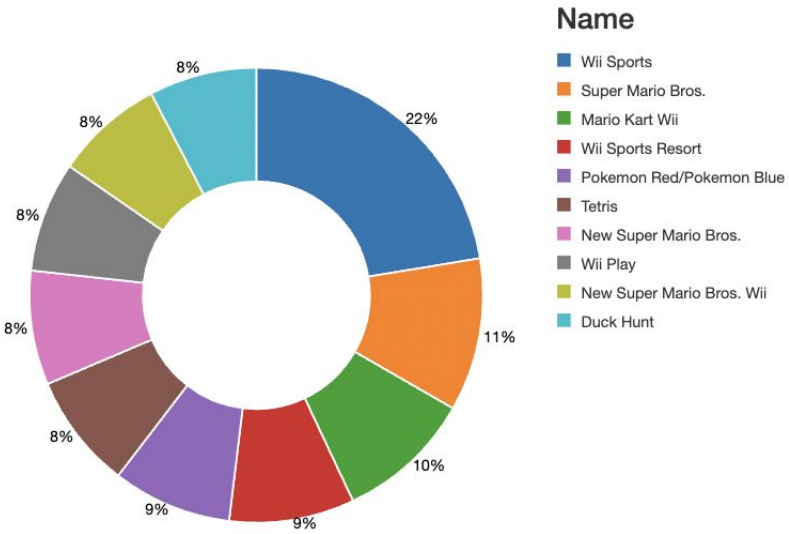


Insights:

- We can see how PS2 had a constant release of games (around 250 per year) during the 2000 first decade, but it started to decline by ends of it.
- DS became extremely popular after it's release, and many games started coming out quickly. But after the year 2008 it started crashing and by 2012 almost no more games were released. This is probably due to the introduction of other platforms as Wii, PS3, XBOX 360, and PSP (which is a direct substitute for it's portability).
- The Wii suffered the same destiny as the Nintendo DS, as many games were released quickly, but then it suddenly declined.
- There seems to be a tendency for a life cycle of platforms, but this should be confirmed with further exploration

Query 3:

- Find the contribution of the top 10 games (in global sales) to the global revenue
    - We first have to calculate the historic global revenue (sum of all the global sales)
    - Then divide the global sales of each game for that number

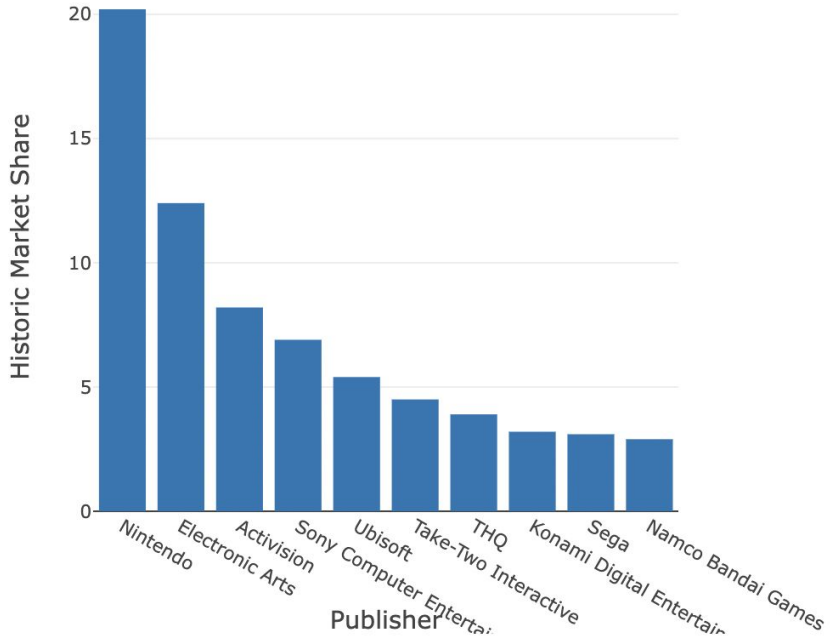


Insights:

- From the sum of the top 10 total global sales, Wii Sports represented the 22%
- Super Mario Bros represented the 11%, so, we know that Wii Sports represented at lest the double than all the other top 10 games.
- Mario Kart for Wii represented 10%
- Wii Sports Resorts and Pokemon Red/Pokemon Blue 9%
- All the other games around 8%

Query 4:

- In terms of historic global revenue, which platform has the highest market share?
    - We want to know the market share of each platform in terms of historic sales

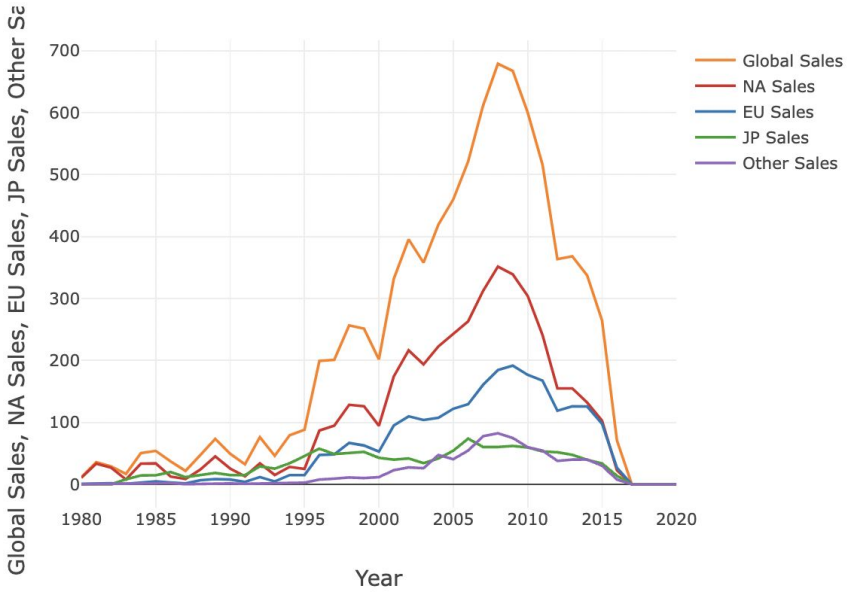


Insights:

- We can clearly see how Nintendo has 20% of the historic total market share
- Electronic Arts (EA) ranks second with 12.4%
- Activision runs 3rd with 8.2%
- And so on...

Query 5:

- Find the distribution of sales per year of release for each region
    - We must group by Year and sum the sales for each column.



Insights:

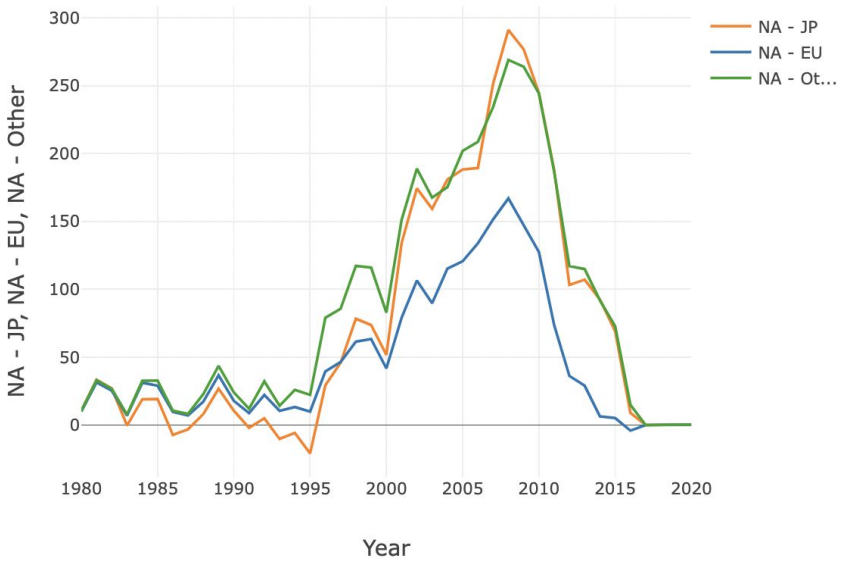
- NA accounts for a huge amount of total sales, specially with their games released after the year 2000.
- JP sales have stayed relatively constant throughout the last decade, compared to other regions
- Sales released around the year 2008 are the ones which have given the highest revenue, but then it declined for all of the regions

Query 6:

- Compute the sales difference between NA_Sales and the other regions
    - We defined a UDF for (x-y) and applied it to the data frame.

Insights:

- Before the year 1995, Japan had games with higher sales than North America, but then NA games sales boomed.
- Difference in sales between NA and other regions before 1995 weren't that high, but later the gap increased hugely.
- After the year 2000, NA difference with JP and Other parts of the world was much higher than the difference between EU.
- The gap started decreasing after 2008 (but looking at previous graphs, general sales have to, so this doesn't mean much)



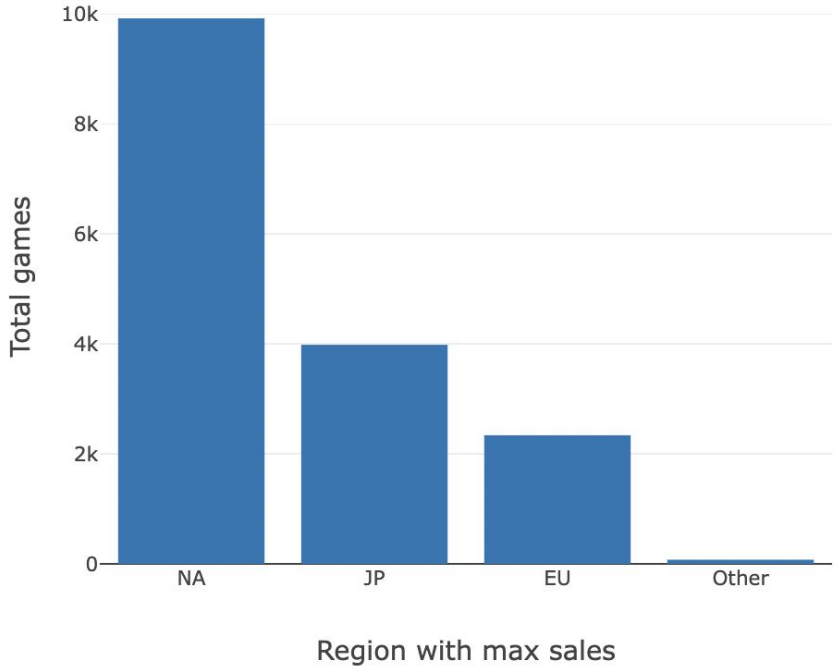
Query 7:

- In how many regions did each games sales led?
  - We want to know the number of games which sales was the maximum for each region.
  - We first retrieve the max number for each row (game)
  - We then create a UDF to compare the sales of each region with the max and run it in the data frame

---

Insights:

- **North America**: almost 10k games which sold the most in this region
- **Japan**: almost 4k games that sold the most in this region
- **Europe**: above 2k games that sold the most in this region
- **Other**: only 76 games that sold the most



Query 8:

- What is the name of the best selling game for each region?
    - We must create a Window with a partition regarding the ‘Region with max sales’ column
    - Now we compute the maximum of each group and then filter out the rows so that the value in the max column in df_maxReg is equal to the max of the window.

	Game	Region	Sales
1	Wii Sports	NA	41.49
2	Nintendogs	EU	11
3	Grand Theft Auto: San Andreas	Other	10.57
4	Pokemon Black/Pokemon White	JP	5.65

Most sold game per region:

1. **NA**: Wii Sports
2. **EU**: Nintendogs
3. **Other**: Grand Theft Auto: San Andreas
4. **JP**: Pokemon Black/Pokemon White

Query 9:

- Perform a **Join** operation to visualize how much did the top performing games sold in each region.
    - We performed this to be able to know what were the sales in each region for the games in Query 8.

Insights

- We can see that Grand Theft Auto has been developed for different platforms, but the region in which it sold the most was 'Other' (highest rank)
- Worldwide, Wii Sports is the highest sold game, while Nintendogs is Ranked 11th, Grand Theft Auto 18th (for PS2) and Pokemon 27th.

	Rank	Name	Platform	Year	Region with max sales	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
1	1	Wii Sports	Wii	2006	NA	41.49	29.02	3.77	8.46	82.74
2	11	Nintendogs	DS	2005	EU	9.07	11	1.93	2.75	24.76
3	18	Grand Theft Auto: San Andreas	PS2	2004	Other	9.43	0.4	0.41	10.57	20.81
4	27	Pokemon Black/Pokemon White	DS	2010	JP	5.57	3.28	5.65	0.82	15.32
5	875	Grand Theft Auto: San Andreas	XB	2005	Other	1.26	0.61	0	0.09	1.95
6	2122	Grand Theft Auto: San Andreas	PC	2005	Other	0	0.92	0	0.05	0.98