

Hesham Rashid

ISM 6208 – Data Warehousing

University of South Florida

May 1st, 2026

Professor Don Berndt

Final Project

The main analytical question that drives this design is: What platform, genre, publisher, and regional trends have driven video game sales from 1980 to 2024, and does critical acclaim correlate with commercial success?

Executive Summary

In this report I will go over the findings of historical data on video game sales throughout history. The business problem that I am going to be investigating is what are some of the determinates of sales in the video game industry? Video games, especially retro ones, are very near and dear to my heart, so I wanted to create a report for it using SQL. The data was found from a Kaggle competition, (link can be found [here](#)) where there was a CSV file that had sales of over 60 thousand games, their manufacture, their company, their release year, etc. Some key findings that I found were:

- Action games were the most dominant in sales especially in the 7th generation (2006-2012), which corresponds to my childhood
- In terms of sales, Sony is the most dominant of the 3 major companies, although Microsoft produces the highest quality per title
- What critics think of games do matter, the better they think they are the more sales the games get

This report will go through explaining my findings, and the design of the data. To see the ETL script, the visualizations script, the schema, etc. you can go to the GitHub repository I made for this project [here](#).

Problem Statement

What platform, genre, publisher, and regional trends have driven video game sales from 1980 to 2024, and does critical acclaim correlate with commercial success?

Literature Review:

Data Warehousing: Snowflake. (2026).

Snowflake's star schema guide covers dimensional modeling fundamentals, fact and dimension table design, and query performance advantages. This connects with this project as I used a fact table in a star schema. As shown later in the report there are 7 dimension tables surrounding the fact_sales table that sits in the center.

Data Quality: IBM. (2025)

IBM's data quality guide covers the six pillars, accuracy, completeness, consistency, timeliness, validity, and uniqueness, and how each dimension affects organizational decision-making. My ETL pipeline that I created made data quality decisions to back some of these principles, I preserved null sales values rather than dropping rows (completeness), mapped inconsistent console abbreviations to full platform names (consistency), and used an Unknown surrogate key for missing dates (validity).

Data Visualization: Harvard Business School Online. (2021)

Harvard Business School's data storytelling guide explains how combining data, narrative, and visualization together creates more actionable and memorable insights than data alone. This will be shown in the reporting and story telling section where the data is visualized using matplotlib and pandas to make easier to understand charts.

Video Game Industry: Visual Capitalist. (2025)

Visual Capitalist's industry revenue chart shows U.S. video game revenue growing from roughly \$11 billion in the early 2000s to a peak of \$61.2 billion in 2021, driven by platform launches, mobile growth, and pandemic-era demand. This source provides real world context for the sales patterns that are shown in this report.

Data Collection:

Where the data came from:

The data came from a CSV file from Kaggle, again can be seen [here](#), the data set is titled "Video Game Sales 1980–2024" and contains 64,016 records of video game titles spanning over four decades.

What the data contains:

The data contains 14 columns covering game title, console, genre, publisher, developer, critic score, total sales, and regional sales broken down across North America, Japan, PAL, and Rest of World. The release dates span from 1971 to 2024 across 81 platforms.

Data limitations and quality observations:

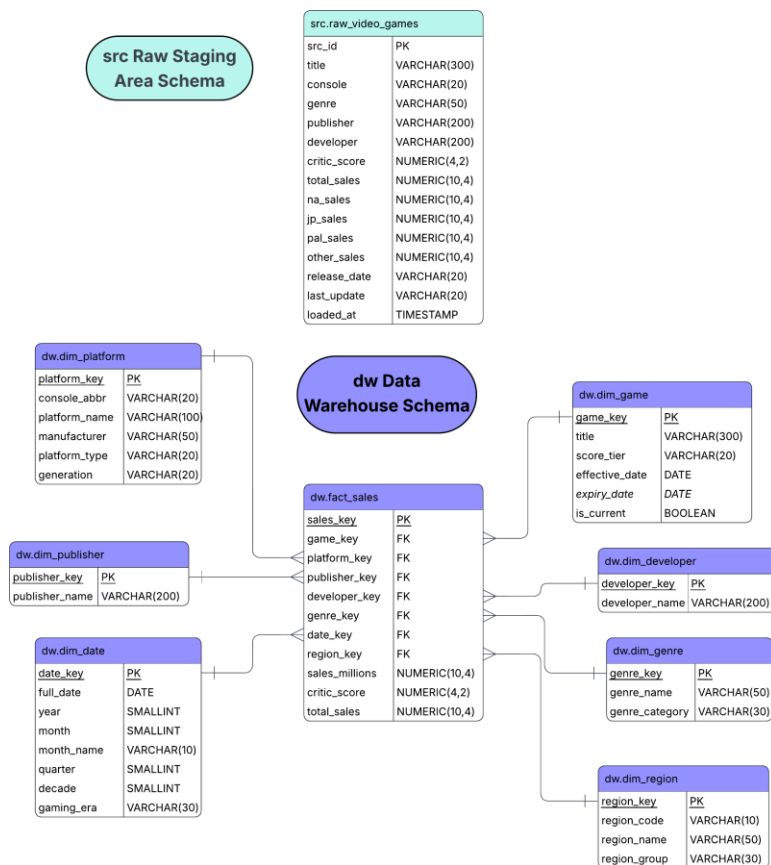
- critic_score had 57,338 null values out of 64,016 rows, meaning roughly 90% of games in the dataset were unscored
- total_sales had 45,094 nulls, interpreted as missing data rather than zero sales
- release_date had 7,051 nulls
- developer had 17 nulls
- Console abbreviations were abbreviated and required mapping to full platform names

Nulls were preserved in the fact table rather than dropped, representing data collection gaps rather than absence of sales activity, this connects back to the IBM data quality source about completeness (IBM. 2025)

Database Design

The data is stored in a Postgres Database (retro_games_db). There are two schemas in the database. The first schema is a raw staging area (src schema) this is where the data gets extracted from the CSV files to the Postgres Database. Then the data gets transformed to fit the second schema, which is the dw, data warehouse, schema that has a central fact table and 7 dimension tables surrounding it. There is an ETL script that is written in python that handles the extraction of the data from the CSV file to the src schema, then transform the data in the src schema to fit the dw schema, then load the data into the dw schema for us to do queries on.

This is the ERD to visualize the schemas and how the fact table connects with the dimension tables:



Dimension Table Descriptions Here is a brief description of each dimension table:

- **dim_game** This dimension stores each game title along with a derived score_tier column that buckets the raw critic score into four categories: Acclaimed (8+), Good (7-8), Mixed (5-7), and Poor (<5). This dimension implements SCD Type 2 to track changes in game records over time. The source dataset included a last_update column indicating when a record was revised, which was used as the expiry_date. This allows historical analysis of how sales figures were revised after initial reporting.
- **dim_platform** The source data stored platform information as abbreviated codes such as GC, PS2, and X360 which are not meaningful on their own. This dimension enriches those abbreviations with full platform names, manufacturer, platform type and console generation. This was added manually through a lookup table in the ETL pipeline.
- **dim_date** Dates were parsed from the raw release_date column and broken into year, month, quarter, and decade. A derived gaming_era column was added to group years into historically meaningful console generations. An Unknown surrogate record was inserted as the first row to handle the 7,051 games with missing release dates, allowing those records to still participate in the fact table without being dropped.
- **dim_genre** Genre values in the source data formed a fixed controlled vocabulary of 20 genres. Because both the genre names and their category mappings are fully known and static, this dimension was seeded directly in the DDL rather than loaded dynamically from the CSV.
- **dim_region** Regional sales in the source data were stored as separate columns, na_sales, jp_sales, pal_sales, and other_sales, rather than as rows. This dimension normalizes those columns into four region records.

- **dim_publisher / dim_developer** These are simple lookup dimensions loaded dynamically from the unique publisher and developer values in the source CSV. Both include an Unknown fallback record to handle the 17 null developer values and any publisher records labeled Unknown in the source data.

Fact Table Description: Each grain (row) of the fact table represents one row per game, platform, and region.

OLTP Companion Design: For the OLTP, the tables could look something like:

OLTP Tables:

- games: game_id, title, genre_id, developer_id, release_date
- platforms: platform_id, platform_name, manufacturer_id
- manufacturers: manufacturer_id, manufacturer_name, country
- publishers: publisher_id, publisher_name
- developers: developer_id, developer_name
- reviews: review_id, game_id, platform_id, critic_score, review_date, source
- sales_transactions: transaction_id, game_id, platform_id, region_id, sale_date, units_sold, sale_price
- regions: region_id, region_name

This OLTP schema is fully normalized. Manufacturer is its own table separate from platform, and individual sale transactions are recorded row by row with a price and unit count. The dw schema collapses all of this into wide dimension tables and a single fact table aggregated at the

game level. The OLTP system could handle individual transactions per day and can handle fast writes, while the dw schema would be loaded periodically to handle analytical queries. Queries on the OLTP system would be simple lookups like "how many units of GTA V sold today" while queries on the dw schema answer broader questions like "which genre dominated sales in the 7th console generation"

Exploratory Data Analysis

Here I will go through the queries I split up the queries into 3 topics; game trends over time, differences between platforms, and critic scores vs sales.

Topic 1, Game Trends Over Time:

Query 1, Top genre per era:

```
-- Top genre per era (ranked)
-- Uses RANK() to find the #1 genre in each gaming era
WITH genre_era_sales AS (
    SELECT
        d.gaming_era,
        g.genre_name,
        ROUND(SUM(f.total_sales)::NUMERIC, 2) AS total_sales_millions,
        COUNT(DISTINCT f.game_key) AS game_count
    FROM dw.fact_sales f
    JOIN dw.dim_date d ON f.date_key = d.date_key
    JOIN dw.dim_genre g ON f.genre_key = g.genre_key
    WHERE f.region_key = 1
        AND d.gaming_era != 'Unknown'
        AND f.total_sales IS NOT NULL
    GROUP BY d.gaming_era, g.genre_name
),
ranked AS (
    SELECT *,
        RANK() OVER (PARTITION BY gaming_era ORDER BY total_sales_millions DESC) AS rnk
    FROM genre_era_sales
)
SELECT
    gaming_era,
    genre_name,
    total_sales_millions,
    game_count,
    rnk AS rank_in_era
FROM ranked
WHERE rnk <= 3
ORDER BY gaming_era, rnk;
```

	A-Z gaming_era	A-Z genre_name	123 total_sales_millions	123 game_count	123 rank_in_era
1	2nd Gen (1976-1982)	Shooter	18.11	16	1
2	2nd Gen (1976-1982)	Action	16.87	31	2
3	2nd Gen (1976-1982)	Puzzle	3.27	5	3
4	3rd/4th Gen (1983-1992)	Platform	25.17	26	1
5	3rd/4th Gen (1983-1992)	Action	20.48	38	2
6	3rd/4th Gen (1983-1992)	Shooter	11.37	17	3
7	5th Gen (1993-1998)	Sports	89.91	207	1
8	5th Gen (1993-1998)	Fighting	71.36	141	2
9	5th Gen (1993-1998)	Racing	48.55	102	3
10	6th Gen (1999-2005)	Sports	372.74	712	1
11	6th Gen (1999-2005)	Action	301.33	602	2
12	6th Gen (1999-2005)	Racing	225.99	508	3
13	7th Gen (2006-2012)	Action	502.58	929	1
14	7th Gen (2006-2012)	Sports	473.35	851	2
15	7th Gen (2006-2012)	Shooter	426.35	460	3
16	8th Gen (2013-2019)	Shooter	322.11	161	1
17	8th Gen (2013-2019)	Action	244.46	494	2
18	8th Gen (2013-2019)	Sports	237.55	162	3
19	9th Gen (2020+)	Role-Playing	1.89	6	1
20	9th Gen (2020+)	Action	0.4	5	2
21	9th Gen (2020+)	Misc	0.34	5	3

These are the top 3 genres in terms of sales for each generation. This is an interesting query as it shows changes in genre interest over time and what genres stay popular. For example, we can see that in the early days of video games, puzzle games were a high seller, but as games became more sophisticated that fell out. Shooter and Action games are constantly in the top 3 of most generations showing how popular they are as genres.

Query 2, fastest growing genres between the 2000s and the 2010s:

```
-- Fastest growing genres (2000s vs 2010s)
SELECT
  g.genre_name,
  ROUND(SUM(CASE WHEN d.decade = 2000 THEN f.total_sales ELSE 0 END)::NUMERIC, 2) AS sales_2000s,
  ROUND(SUM(CASE WHEN d.decade = 2010 THEN f.total_sales ELSE 0 END)::NUMERIC, 2) AS sales_2010s,
  ROUND((
    SUM(CASE WHEN d.decade = 2010 THEN f.total_sales ELSE 0 END) -
    SUM(CASE WHEN d.decade = 2000 THEN f.total_sales ELSE 0 END)
  )::NUMERIC, 2) AS sales_change,
  ROUND((
    (SUM(CASE WHEN d.decade = 2010 THEN f.total_sales ELSE 0 END) -
    SUM(CASE WHEN d.decade = 2000 THEN f.total_sales ELSE 0 END)) /
    NULLIF(SUM(CASE WHEN d.decade = 2000 THEN f.total_sales ELSE 0 END), 0) * 100
  )::NUMERIC, 2) AS pct_change
FROM dw.fact_sales f
JOIN dw.dim_date d ON f.date_key = d.date_key
JOIN dw.dim_genre g ON f.genre_key = g.genre_key
WHERE f.region_key = 1
  AND d.decade IN (2000, 2010)
  AND f.total_sales IS NOT NULL
GROUP BY g.genre_name
ORDER BY pct_change DESC NULLS LAST;
```

	A2 genre_name	123 sales_2000s	123 sales_2010s	123 sales_change	123 pct_change
1	Action-Adventure	0.99	147.48	146.49	14,796.97
2	Visual Novel	0.07	5.68	5.61	8,014.29
3	Music	0.85	50.9	50.05	5,888.24
4	MMO	0.32	8.99	8.67	2,709.38
5	Party	0.64	5.57	4.93	770.31
6	Education	0.36	0.61	0.25	69.44
7	Shooter	361.39	560.41	199.02	55.07
8	Role-Playing	170.05	215.43	45.38	26.69
9	Action	573.81	451.53	-122.28	-21.31
10	Sports	619.59	426.83	-192.76	-31.11
11	Misc	342.87	174.32	-168.55	-49.16
12	Fighting	170.67	83.13	-87.54	-51.29
13	Strategy	61.4	27.08	-34.32	-55.9
14	Adventure	208.93	83.74	-125.19	-59.92
15	Puzzle	66.85	23.66	-43.19	-64.61
16	Platform	210.14	72.03	-138.11	-65.72
17	Racing	333.18	111.44	-221.74	-66.55
18	Simulation	203.59	67.88	-135.71	-66.66
19	Sandbox	0	1.89	1.89	[NULL]
20	Board Game	0	0.33	0.33	[NULL]

In this query we can see the boom in sales for the sales of action-adventure games from the 2000s to the 2010s. This could be because of the rise of the genre, as in the 2000s there were more only action games, while the action-adventure genre as a whole boomed. We can also see the introduction of sandbox and board game genre games in the 2010s in the data set, as well as the fall off of platform, racing and simulation games.

Topic 2, Differences Between Platforms:

Query 3, the big 3 companies ranked in terms of sales:

```
-- Total sales by manufacturer overall
SELECT
    p.manufacturer,
    ROUND(SUM(f.total_sales)::NUMERIC, 2) AS total_sales_millions,
    COUNT(DISTINCT f.game_key)           AS total_titles,
    COUNT(DISTINCT p.platform_key)       AS platform_count,
    ROUND(AVG(f.total_sales)::NUMERIC, 4) AS avg_sales_per_title
FROM dw.fact_sales f
JOIN dw.dim_platform p ON f.platform_key = p.platform_key
WHERE f.region_key = 1
    AND f.total_sales IS NOT NULL
    AND p.manufacturer IN ('Nintendo', 'Sony', 'Microsoft')
GROUP BY p.manufacturer
ORDER BY total_sales_millions DESC;
```

	A-Z manufacturer	123 total_sales_millions	123 total_titles	123 platform_count	123 avg_sales_per_title
1	Sony	3,265.75	7,074	7	0.4327
2	Nintendo	1,664.73	6,362	13	0.2473
3	Microsoft	1,361	2,587	4	0.5097

Really straightforward query that shows Sony has the most sales compared to the other 2 companies with over \$3 Billion. It is also important to note that Microsoft has the least sales, however also the least number of titles, so they actually have higher average sales per title, possibly showing a higher quality of game be title.

Query 4, Ranking each company's platform into quadrants:

```
-- Ranks every Nintendo/Sony/Microsoft platform into sales quartiles 1 being lowest, 4 being highest
SELECT
    p.manufacturer,
    p.platform_name,
    p.generation,
    p.platform_type,
    ROUND(SUM(f.total_sales)::NUMERIC, 2) AS total_sales_millions,
    COUNT(DISTINCT f.game_key) AS title_count,
    NTILE(4) OVER (
        PARTITION BY p.manufacturer
        ORDER BY SUM(f.total_sales)
    ) AS sales_quartile
FROM dw.fact_sales f
JOIN dw.dim_platform p ON f.platform_key = p.platform_key
WHERE f.region_key = 1
    AND f.total_sales IS NOT NULL
    AND p.manufacturer IN ('Nintendo', 'Sony', 'Microsoft')
GROUP BY p.manufacturer, p.platform_name, p.generation, p.platform_type
ORDER BY sales_quartile;
```

	A-Z manufacturer	A-Z platform_name	A-Z generation	A-Z platform_type	123 total_sales_millions	123 title_count	123 sales_quartile
1	Microsoft	Xbox Live Arcade	7th Gen	Digital	0.2	8	1
2	Sony	PlayStation Vita	8th Gen	Handheld	63.02	675	1
3	Sony	PlayStation Network	7th Gen	Digital	3.81	15	1
4	Nintendo	Virtual Console	7th Gen	Digital	0.35	6	1
5	Nintendo	Game Boy Color	5th Gen	Handheld	4.34	3	1
6	Nintendo	Game Boy	4th Gen	Handheld	19.84	46	1
7	Nintendo	Nintendo Wii U	8th Gen	Home Console	35.42	148	1
8	Sony	PlayStation Portable	7th Gen	Handheld	245.29	1,337	2
9	Sony	PlayStation 4	8th Gen	Home Console	539.92	906	2
10	Nintendo	Nintendo Switch	9th Gen	Hybrid	36.46	263	2
11	Nintendo	Nintendo Entertainment System	3rd Gen	Home Console	47.93	48	2
12	Nintendo	Super Nintendo	4th Gen	Home Console	65.71	197	2
13	Microsoft	Xbox	6th Gen	Home Console	232.05	836	2
14	Nintendo	Nintendo 3DS	8th Gen	Handheld	99.27	560	3
15	Microsoft	Xbox One	8th Gen	Home Console	268.96	524	3
16	Nintendo	Nintendo 64	5th Gen	Home Console	93.79	278	3
17	Nintendo	GameCube	6th Gen	Home Console	119.53	528	3
18	Sony	PlayStation	5th Gen	Home Console	546.25	1,126	3
19	Sony	PlayStation 3	7th Gen	Home Console	839.7	1,346	3
20	Nintendo	Game Boy Advance	6th Gen	Handheld	224.48	896	4
21	Microsoft	Xbox 360	7th Gen	Home Console	859.79	1,301	4
22	Sony	PlayStation 2	6th Gen	Home Console	1,027.76	2,123	4
23	Nintendo	Nintendo Wii	7th Gen	Home Console	459.44	1,355	4
24	Nintendo	Nintendo DS	7th Gen	Handheld	458.17	2,388	4

This query takes each platform for each of the companies and ranks them based on the sales associated with their games. In the 4th quadrant those are the platforms that sold the most games, we can see some really popular systems like the PlayStation 2, Xbox 360 and the Nintendo Wii. Quadrant 1 has systems whose games sold the least, which also corresponds to less popular systems like the PlayStation Vita, and Nintendo Virtual Console. The 4 quadrants are made using NTILE window function, which divides each manufacturer's platforms into four equal buckets ranked by total sales

Topic 3, Critic Score vs Sales:

Query 5, Do critic scores translate to sales:

```
-- Top-level view of whether critical acclaim translates to sales
SELECT
    gm.score_tier,
    COUNT(DISTINCT f.game_key) AS game_count,
    ROUND(AVG(f.total_sales)::NUMERIC, 4) AS avg_sales_millions,
    ROUND(MAX(f.total_sales)::NUMERIC, 4) AS max_sales_millions,
    ROUND(SUM(f.total_sales)::NUMERIC, 2) AS total_sales_millions
FROM dw.fact_sales f
JOIN dw.dim_game gm ON f.game_key = gm.game_key
WHERE f.region_key = 1
    AND f.total_sales IS NOT NULL
    AND gm.score_tier != 'Unscored'
GROUP BY gm.score_tier
ORDER BY avg_sales_millions DESC;
```

	AZ score_tier	123 game_count	123 avg_sales_millions	123 max_sales_millions	123 total_sales_millions
1	Acclaimed (8+)	1,054	1.3071	20.32	1,742.33
2	Good (7-8)	1,010	0.5757	10.13	685.67
3	Mixed (5-7)	1,040	0.4137	10.41	517.95
4	Poor (<5)	315	0.2739	4.06	95.86

This query does show that when games are rated higher by critics they tend to do better in sales.

This makes sense, better games get rated higher, and better games are more popular, so they sell more.

Query 6, What are some “critic-proof” games:

```
-- The "critic-proof" games – massive sales despite mixed/poor reviews
SELECT
    gm.title,
    p.platform_name,
    p.manufacturer,
    g.genre_name,
    d.year,
    f.total_sales,
    f.critic_score,
    gm.score_tier
FROM dw.fact_sales f
JOIN dw.dim_game gm ON f.game_key = gm.game_key
JOIN dw.dim_platform p ON f.platform_key = p.platform_key
JOIN dw.dim_genre g ON f.genre_key = g.genre_key
JOIN dw.dim_date d ON f.date_key = d.date_key
JOIN dw.dim_region r ON f.region_key = r.region_key
WHERE r.region_code = 'NA'
    AND gm.score_tier IN ('Mixed (5-7)', 'Poor (<5)')
    AND f.total_sales IS NOT NULL
ORDER BY f.total_sales DESC
LIMIT 20;
```

	AZ title	AZ platform_name	AZ manufacturer	AZ genre_name	123 year	123 total_sales	123 critic_score	AZ scor
1	Call of Duty: Ghosts	Xbox 360	Microsoft	Shooter	2,013	10.41	6.9	Mixed (
2	Cooking Mama	Nintendo DS	Nintendo	Simulation	2,006	5.66	6.6	Mixed (
3	Crash Bandicoot: The Wrath of Cortex	PlayStation 2	Sony	Platform	2,001	5.42	6.9	Mixed (
4	Medal of Honor: Rising Sun	PlayStation 2	Sony	Shooter	2,003	5.13	5.9	Mixed (
5	FIFA 15	PlayStation 3	Sony	Sports	2,014	4.56	6.9	Mixed (
6	Mario & Sonic at the Olympic Winter Games	Nintendo Wii	Nintendo	Sports	2,009	4.54	6.8	Mixed (
7	Michael Jackson: The Experience	Nintendo Wii	Nintendo	Misc	2,010	4.37	5.6	Mixed (
8	Teenage Mutant Ninja Turtles	Nintendo Entertainment System	Nintendo	Platform	1,989	4.17	5.9	Mixed (
9	Star Wars Battlefront (2015)	Xbox One	Microsoft	Shooter	2,015	4.15	6.9	Mixed (
10	Carnival Games	Nintendo Wii	Nintendo	Misc	2,007	4.06	4.2	Poor (<
11	MySims	Nintendo DS	Nintendo	Simulation	2,007	3.66	6.7	Mixed (
12	The Simpsons: Road Rage	PlayStation 2	Sony	Racing	2,001	3.61	6.1	Mixed (
13	Cooking Mama 2: Dinner With Friends	Nintendo DS	Nintendo	Simulation	2,007	3.58	6.9	Mixed (
14	Grand Theft Auto 2	PlayStation	Sony	Action	1,999	3.42	6.9	Mixed (
15	Crash Bash	PlayStation	Sony	Misc	2,000	3.39	6.8	Mixed (
16	007: Tomorrow Never Dies	PlayStation	Sony	Shooter	1,999	3.21	6.2	Mixed (
17	FIFA 15	Xbox 360	Microsoft	Sports	2,014	2.91	6.9	Mixed (
18	Sega Superstars Tennis	Xbox 360	Microsoft	Sports	2,008	2.89	6.9	Mixed (
19	Cooking Mama: Cook Off	Nintendo Wii	Nintendo	Simulation	2,007	2.89	5.9	Mixed (
20	Imagine: Babyz	Nintendo DS	Nintendo	Simulation	2,007	2.87	3.5	Poor (<

This query shows games that had low scores by critics, but still sold really well. This can be seen as the critics got it wrong, or games that have a cult following. An example of the cult following could be the Cooking Mama series, and an example of the critics got it wrong could be the Call of Duty Ghosts.

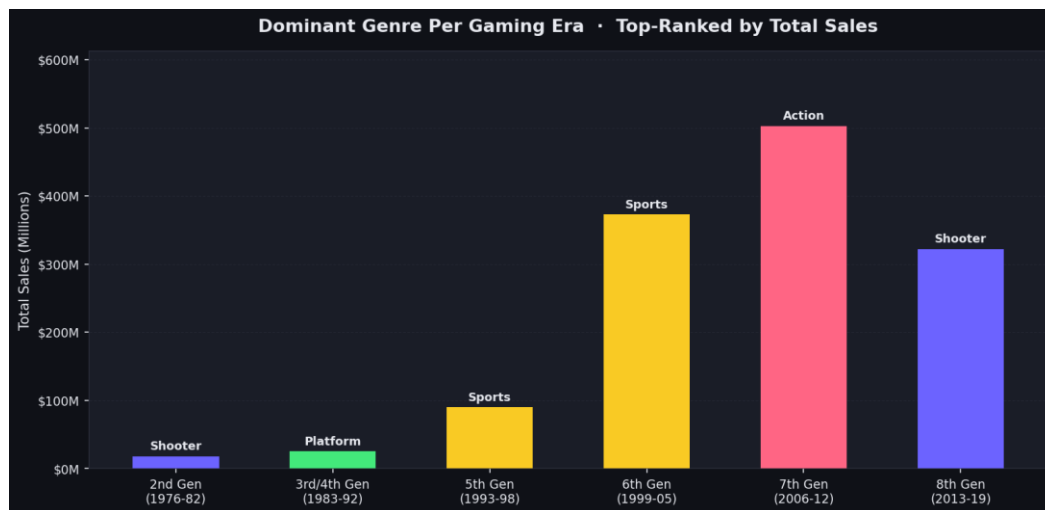
Reporting, Modeling, Storytelling

Here I will go through the visualizations, again these are split into the 3 topics mentioned earlier.

I created these visualizations using a python script that used matplotlib and pandas. The script can be seen in the project's [GitHub Repository](#). Note that each chart corresponds to the numbered query above

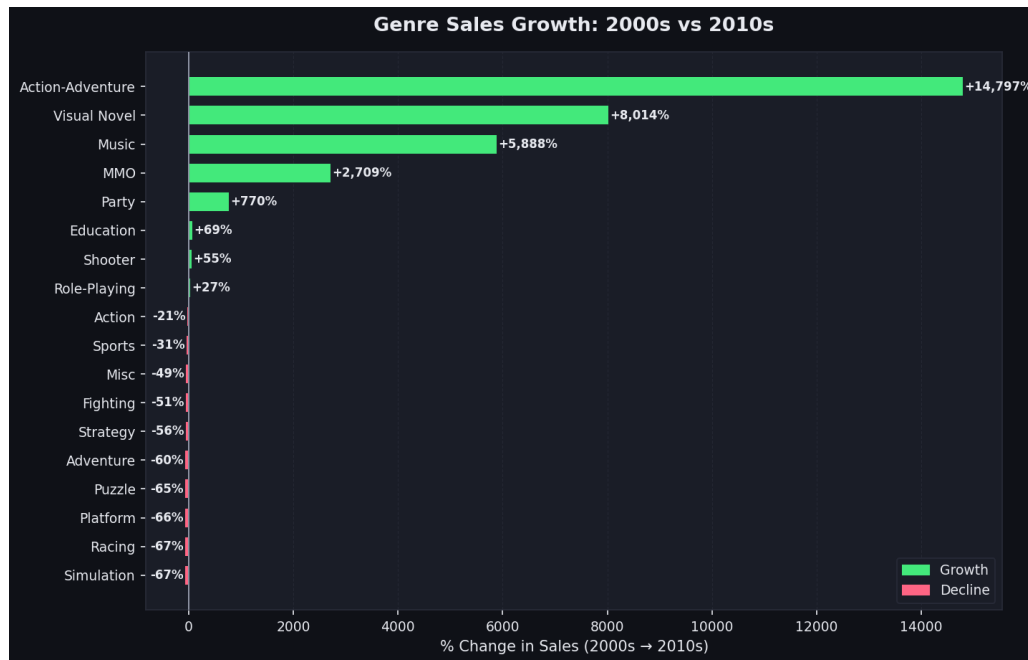
Topic 1, Game Trends Over Time:

Chart 1, Top genre per era:



In this chart for simplicity, I only did the top genre per era. This visually shows the boom of sports games in rise of 3D models in gaming, especially in the PlayStation 1 and 2 eras. It also shows the maturing of the industry with the rise of action games in the 7th generation and the revival of shooter games in the 8th generation.

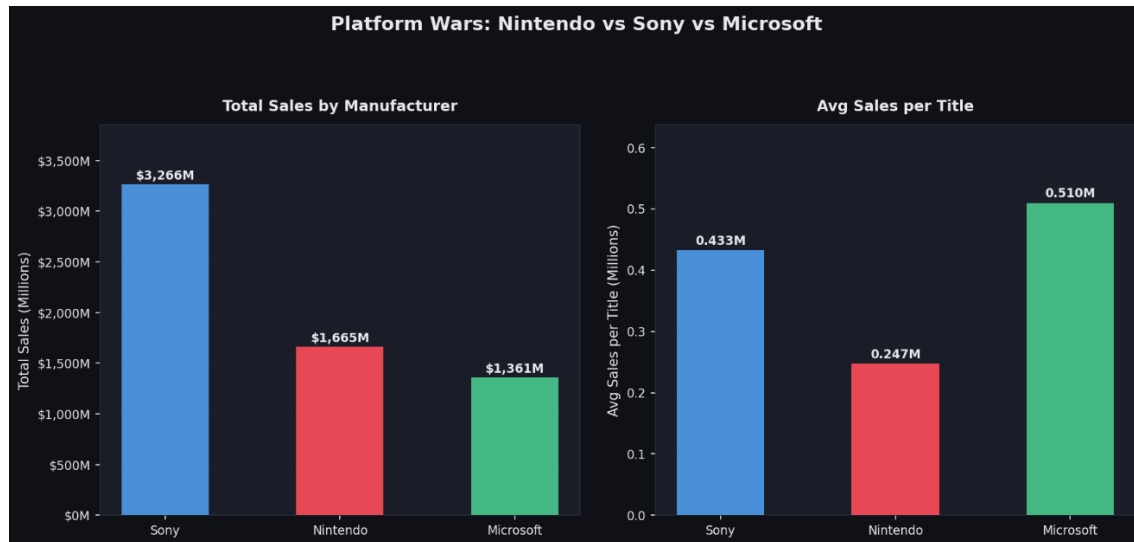
Chart 2, fastest growing genres between the 2000s and the 2010s:



This chart shows the massive boom of the Action-Adventure genre that we talked about earlier, but in a visual perspective shows how much larger the boom is compared to the other genres, that also grew significantly.

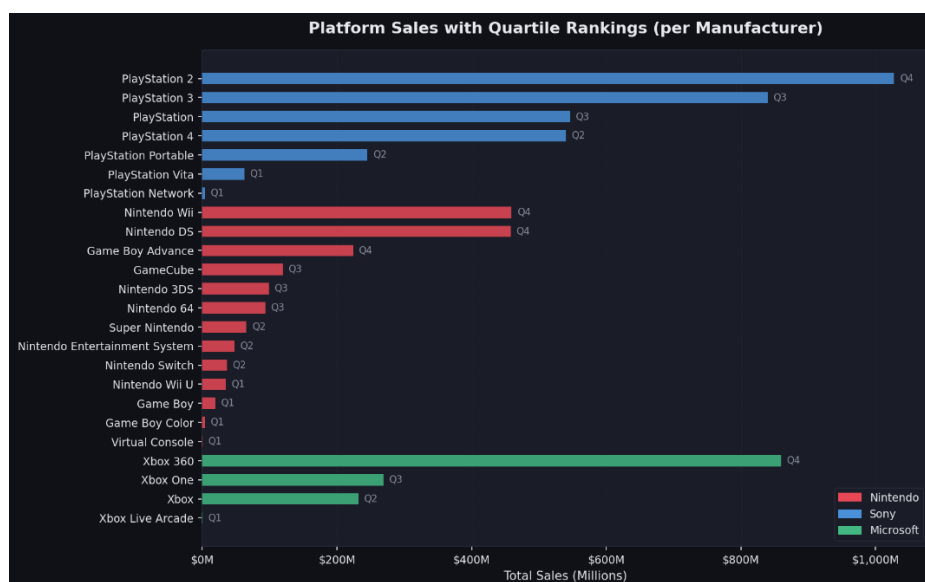
Topic 2, Differences Between Platforms:

Chart 3, the big 3 companies ranked in terms of sales:



This visually shows how much more Sony has in terms of sales compared to the other 2 companies, however it shows how Microsoft is not far off from Sony and actually leads in sales per title.

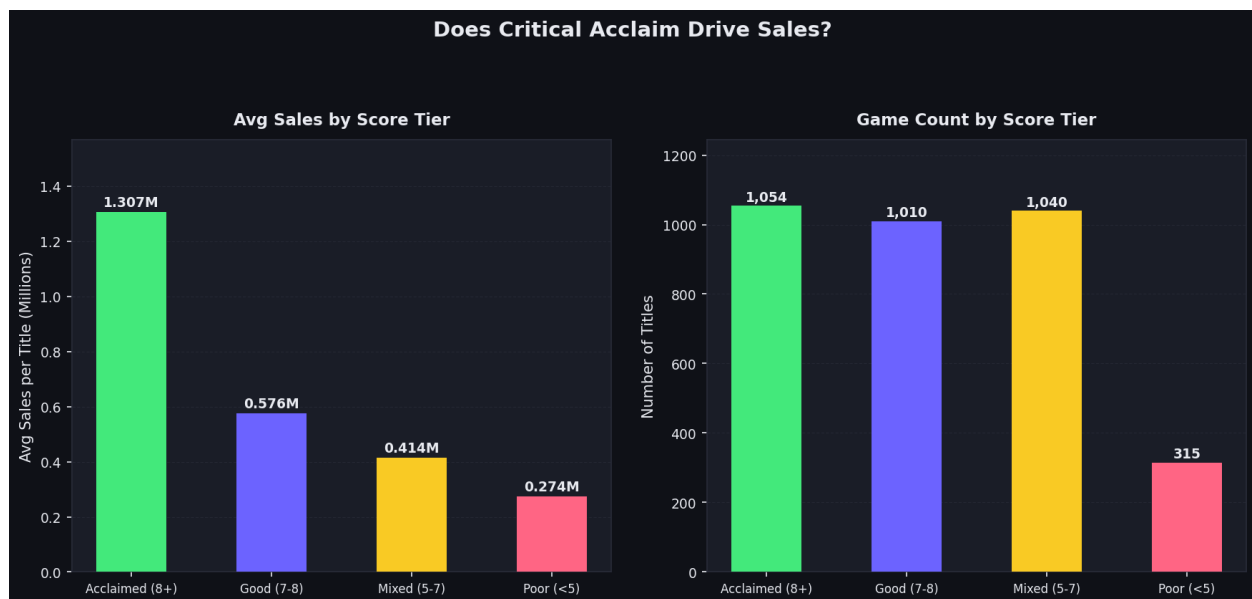
Chart 4, Ranking each company's platform into quadrants:



This visualization shows the sales of games per platform but also categorizes the quadrants. This is good to see how Nintendo's 2 quadrant 4 platforms actually performed worse than one of Sony's quadrant 2 platforms. It also shows how since Microsoft only has 4 platforms in this dataset, each quadrant has its own platform.

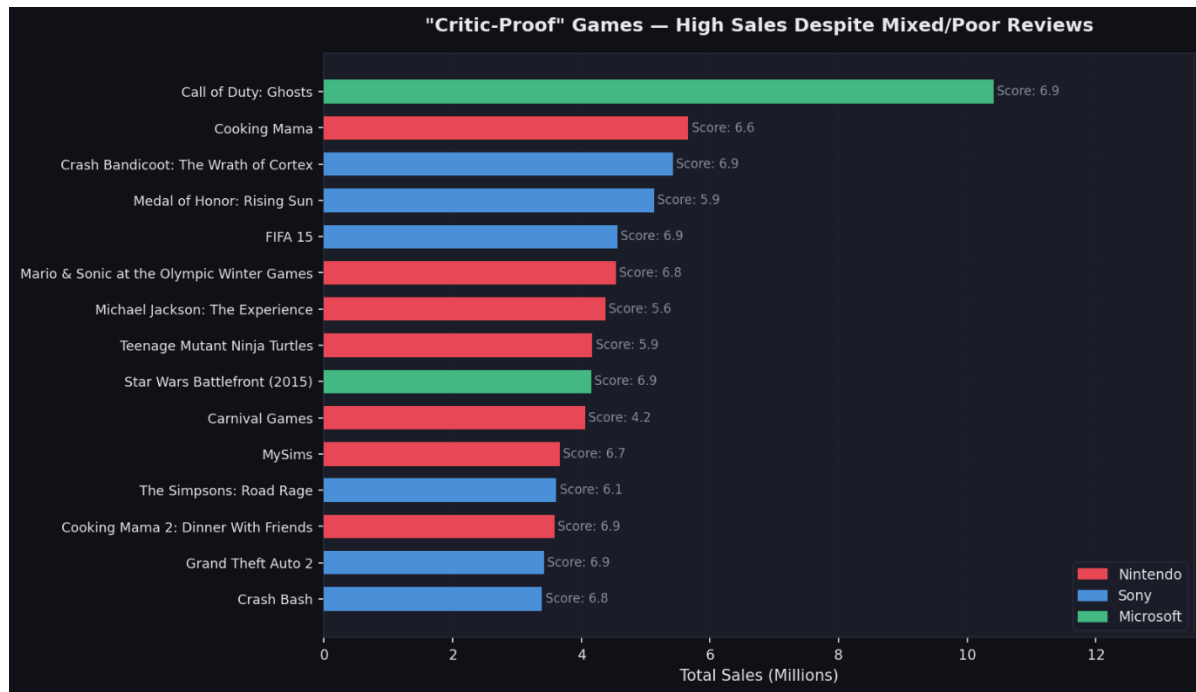
Topic 3, Critic Score vs Sales:

Chart 5, Do critic scores translate to sales:



Here we can see how lopsided the sales of games with acclaimed critic scores are compared to the games with lower scores. The sales of the games with an acclaimed score are double the games with just a good score, it shows that there is a strong correlation between critic scores and sales.

Chart 6, What are some “critic-proof” games:



This visualization shows some of the games that had poor reviews but still performed well. The games are also color coded by the companies to help differentiate.

Conclusion:

This project set out to answer the question of what platform, genre, publisher, and regional trends have driven video game sales from 1980 to 2024, and whether critical acclaim correlates with commercial success. Through the design of a star schema data warehouse, an ETL pipeline, and a series of analytical queries and visualizations, we can see several meaningful patterns emerged from the data.

Genre preferences have shifted significantly across console generations, moving from Platform and Puzzle games in the early era to the dominance of Action and Shooter titles in the modern era, reflecting both the evolution of game design and changing player demographics. In the platform wars, Sony leads in total sales volume, but Microsoft's higher average sales per title suggests a different publishing strategy focused on fewer, higher profile releases. Finally, the data confirms that critical acclaim does correlate with commercial success, acclaimed titles average significantly higher sales than mixed or poor reviewed games, however the critic-proof games analysis shows that brand loyalty and franchise recognition can override critical reception entirely.

From a data warehousing perspective, this project demonstrated the value of dimensional modeling in organizing complex, multi-dimensional data into a structure that supports efficient and meaningful analysis. The two schema architecture, SCD Type 2 implementation, and careful null handling decisions all contributed to a robust analytical foundation that could realistically support business decision making in a game publishing or retail context.

References:

Divekar, B. (2024). *Video game sales and industry data 1980–2024* [Data set]. Kaggle.

<https://www.kaggle.com/datasets/bhushandivekar/video-game-sales-and-industry-data-1980-2024/data>

Harvard Business School Online. (2021). *Data storytelling: How to effectively tell a story with data*. <https://online.hbs.edu/blog/post/data-storytelling>

IBM. (2025). *6 pillars of data quality and how to improve your data*.

<https://www.ibm.com/products/tutorials/6-pillars-of-data-quality-and-how-to-improve-your-data>

Rashid, H. (2026). *Video game sales data warehouse* [Source code]. GitHub.

<https://github.com/hrashid13/Retro-Game-DW>

Snowflake. (2026). *What is a star schema? A complete guide for data modeling*.

<https://www.snowflake.com/en/fundamentals/star-schema/>

Visual Capitalist. (2025). *Charted: Video game industry revenue in the U.S. (2002–2024)*.

<https://www.visualcapitalist.com/video-game-industry-revenue-over-time/>