



ITC 6460 Cloud Analytics

Final Project Report – Video Games Market Analysis

Group: Husky 2

By

- Katherine LaConte

- Likhitha Varakala

- Manmitha Pantangi

- Shivam Sinha

Introduction

The dataset for this analysis encompasses two comprehensive files: `game_sales_data.csv` and `vgsales.csv`. These datasets provide an extensive overview of the video game industry, cataloging various aspects such as game titles, platforms, release years, genre classifications, and sales figures across different regions including North America, Europe, Japan, and other territories. This rich compilation of data offers a unique lens through which to view the dynamics of video game sales globally, capturing the trends, preferences, and shifts in the market over time. It serves as a critical foundation for the analytical tasks ahead, enabling a multifaceted exploration of what drives success in the video game industry.

Out[4]:

	Rank	Name	Platform	Publisher	Developer	Critic_Score	User_Score	Total_Shipped	Year
0	1	Wii Sports	Wii	Nintendo	Nintendo EAD	7.7	8.0	82.90	2006
1	2	Super Mario Bros.	NES	Nintendo	Nintendo EAD	10.0	8.2	40.24	1985
2	3	Counter-Strike: Global Offensive	PC	Valve	Valve Corporation	8.0	7.5	40.00	2012
3	4	Mario Kart Wii	Wii	Nintendo	Nintendo EAD	8.2	9.1	37.32	2008
4	5	PLAYERUNKNOWN'S BATTLEGROUNDS	PC	PUBG Corporation	PUBG Corporation	8.6	4.7	36.60	2017
...
19595	19594	FirePower for Microsoft Combat Flight Simulator 3	PC	GMX Media	Shockwave Productions	NaN	NaN	0.01	2004
19596	19595	Tom Clancy's Splinter Cell	PC	Ubisoft	Ubisoft	9.4	NaN	0.01	2003
19597	19596	Ashita no Joe 2: The Anime Super Remix	PS2	Capcom	Capcom	NaN	NaN	0.01	2002
19598	19597	Tokyo Yamanote Boys for V: Main Disc	PSV	Rejet	Rejet	NaN	NaN	0.01	2017
19599	19598	NadePro!! Kisama no Seiyuu Yatte Miro!	PS2	GungHo	GungHo Works	NaN	NaN	0.01	2009

19600 rows × 9 columns

Fig 1: Dataset 1

In [5]:

df2

Out[5]:

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37
...
16593	16596	Woody Woodpecker in Crazy Castle 5	GBA	2002.0	Platform	Kemco	0.01	0.00	0.00	0.00	0.01
16594	16597	Men in Black II: Alien Escape	GC	2003.0	Shooter	Infogrames	0.01	0.00	0.00	0.00	0.01
16595	16598	SCORE International Baja 1000: The Official Game	PS2	2008.0	Racing	Activision	0.00	0.00	0.00	0.00	0.01
16596	16599	Know How 2	DS	2010.0	Puzzle	7G//AMES	0.00	0.01	0.00	0.00	0.01
16597	16600	Spirits & Spells	GBA	2003.0	Platform	Wanadoo	0.01	0.00	0.00	0.00	0.01

16598 rows × 11 columns

Fig 2: Dataset 2

Data Cleaning

The data cleaning process for the video game sales datasets, `game_sales_data.csv` and `vg-sales.csv`, was a meticulous task aimed at ensuring the data's accuracy, completeness, and consistency before analysis. This foundational step involved several critical actions to refine the datasets, making them suitable for in-depth examination and analysis. Initially, **duplicate entries** were identified and removed to prevent skewed results, ensuring that each record uniquely represented a video game. This step was crucial for maintaining the integrity of the dataset and avoiding inflation of sales figures or misrepresentation of market trends.

Handling **missing values** was another significant aspect of the data cleaning process, requiring careful consideration to decide whether to impute missing data based on available information or remove entries lacking critical details. This decision-making process was vital for preserving the dataset's quality without introducing bias or inaccuracies.

Data type conversion was carried out to align the data with the requirements of the analysis tools and techniques to be applied later. Sales figures were converted to numeric types to facilitate mathematical operations and comparisons, while release years were adjusted to date formats to enable temporal analyses. This step ensured that the datasets were in a state that accurately reflected the reality of the market, allowing for meaningful insights to be drawn.

Finally, **merging the datasets** was a complex but essential task, requiring alignment of similar columns, reconciliation of genre classifications, and ensuring consistent data granularity. This integration process created a unified dataset that offered a comprehensive view of the video game market, laying a solid foundation for the subsequent analytical phases.

Through these meticulous data cleaning and preparation steps, the datasets were transformed into a reliable source for analysis, setting the stage for uncovering valuable insights into video game sales trends and market dynamics.

AWS Services

In this analysis project, the utilization of Amazon Web Services (AWS) played a pivotal role in handling, analyzing, and visualizing large-scale video game sales data. Through the combined use of AWS S3, IAM, AWS Glue, Amazon SageMaker, and AWS QuickSight, a comprehensive and

secure data analysis pipeline was established, enabling the extraction of valuable insights from the datasets `game_sales_data.csv` and `vgsales.csv`.

AWS S3 served as the cornerstone for data storage, offering a highly durable, scalable, and secure solution for managing the extensive datasets required for this analysis. By leveraging S3, we ensured that our data was stored in a centralized location, accessible from anywhere, yet protected against unauthorized access and loss. This facilitated seamless integration with other AWS services for further processing and analysis.

AWS Identity and Access Management (IAM) was instrumental in safeguarding the analysis process. By meticulously managing permissions and roles, IAM enabled us to define who could access the datasets and analytical tools, thus ensuring that only authorized personnel had the ability to interact with the sensitive data. This level of security is critical in maintaining the integrity of the data analysis process and protecting against potential data breaches or unauthorized access.

AWS Glue played a critical role in the data preparation phase. As a fully managed extract, transform, and load (ETL) service, AWS Glue automated the time-consuming tasks of data cataloging, cleaning, enrichment, and transformation. It turned raw data from S3 into a structured format that was ready for analysis. This automation not only streamlined the data preparation process but also ensured consistency and reliability in the data being analyzed.

Amazon SageMaker was leveraged for its powerful machine learning capabilities. It enabled the building, training, and deployment of a regression model aimed at predicting global video game sales. SageMaker's comprehensive and user-friendly environment accelerated the development of the model by providing access to high-performance computing resources, pre-built algorithms, and the flexibility to experiment with different model configurations. This facilitated the creation of a predictive model that could accurately forecast sales trends based on historical data and other influencing factors.

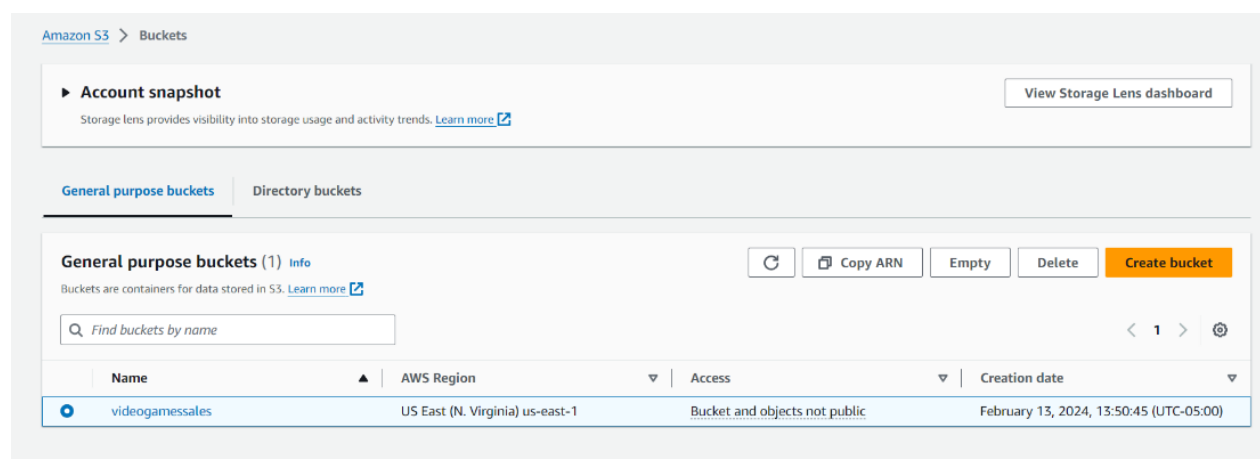
AWS QuickSight was the final piece in the data analysis pipeline, bringing the insights and findings to life through interactive dashboards and visualizations. QuickSight's ability to connect directly to data stored in AWS services, such as S3 and AWS Glue, allowed for real-time analysis and visualization. The dashboards created provided a dynamic and intuitive interface for exploring the data, enabling stakeholders to quickly understand the analysis outcomes and make informed

decisions. By leveraging QuickSight's advanced visualization features, we were able to highlight key trends, patterns, and anomalies in the video game sales data, making the insights accessible to a broad audience.

In summary, the synergy between AWS S3, IAM, AWS Glue, Amazon SageMaker, and AWS QuickSight formed the backbone of this comprehensive analysis project. This integration not only facilitated a secure and efficient data analysis workflow but also enabled the extraction of actionable insights from complex datasets, demonstrating the power of AWS in supporting data-driven decision-making in the video game industry.

Steps

Create an S3 Bucket: Log into the AWS Management Console, navigate to S3, and create a new bucket to store your datasets (game_sales_data.csv and vgsales.csv). Ensure to configure the bucket settings according to your privacy and access requirements.



Use AWS Glue for Data Cataloging:

Create a Crawler: In AWS Glue, set up a new crawler pointing to your S3 bucket. The crawler scans your data and infers schemas.

One crawler successfully created
The following crawler is now created: "videogames"

[AWS Glue](#) > [Crawlers](#) > videogames

videogames

Last updated (UTC)
February 13, 2024 at 19:00:20

Run crawlerEditDelete


Crawler properties



Name videogames	IAM role gluecrawler	Database games_data	State READY
Description -	Security configuration -	Lake Formation configuration -	Table prefix -
Maximum table threshold -			

▶ Advanced settings

Run the Crawler: Execute the crawler to populate the AWS Glue Data Catalog with tables representing your datasets.

Crawler run details

Run ID
 6b471c31-12de-4fa7-a33c-6b56d1e233d9

Tables added (1) csv	Partitions added -
Tables updated -	Partitions updated -
Tables deleted -	Partitions deleted -
Start time (UTC) February 13, 2024 at 19:07:48	Status  Completed
End time (UTC) February 13, 2024 at 19:08:55	Duration 67244
DPU hours -	Log View log 

Close

Table Schema:

Schema (15)

View and manage the table schema.

Q Filter schemas

< 1 > ⚙

Edit schema as JSON

Edit schema

#	Column name	Data type	Partition key	Comment
1	name	string	-	-
2	platform	string	-	-
3	publisher	string	-	-
4	developer	string	-	-
5	critic_score	double	-	-
6	user_score	double	-	-
7	total_shipped	double	-	-
8	year	bigint	-	-
9	genre	string	-	-
10	na_sales	double	-	-
11	eu_sales	double	-	-
12	jp_sales	double	-	-
13	other_sales	double	-	-
14	global_sales	double	-	-
15	year_as_date	string	-	-

Querying Data with AWS Athena:

Navigate to AWS Athena in the AWS Management Console.

Ensure Athena is set to use the Data Catalog populated by Glue.

Write and execute SQL queries against your datasets to analyze video game sales trends, genre performance, and other insights.

SQL Queries

Popular Genres by Global Sales: Identifying the most successful genres worldwide, highlighting consumer preferences.

```

1  -- Most Popular Genres by Global Sales
2  SELECT genre, SUM(global_sales) AS Total_Global_Sales
3  FROM csv
4  GROUP BY genre
5  ORDER BY Total_Global_Sales DESC
6  limit 10;

```

# ▾	genre ▾	Total_Global_Sales
1	Action	1005.7399999999909
2	Shooter	735.8199999999971
3	Platform	616.7299999999999
4	Sports	546.5299999999996
5	Racing	478.20999999999896
6	Misc	465.91999999999814
7	Role-Playing	295.9099999999984
8	Simulation	238.75000000000054
9	Fighting	219.85000000000034
10	Adventure	149.61000000000098

Highest Critic Scores: Filtering games with top reviews to understand the impact of critical acclaim on sales.

```

1  -- Highest Critic Scores
2  SELECT name, genre, developer, platform, critic_score
3  FROM csv
4  ORDER BY critic_score DESC
5  LIMIT 10;

```


#	name	genre	developer	platform	critic_score
1	Final Fantasy II	Role-Playing	Square	SNES	10.000000000000002
2	Super Mario Bros.	Platform	Nintendo EAD	NES	10.000000000000002
3	Super Mario Kart	Racing	Nintendo EAD	SNES	10.000000000000002
4	The Legend Of Zelda: Ocarina Of Time	Action	Nintendo EAD	N64	9.900000000000002
5	Goldeneye 007	Shooter	Rare Ltd.	N64	9.800000000000002
6	Super Mario 64	Platform	Nintendo EAD	N64	9.7
7	NFL 2K	Sports	Visual Concepts	DC	9.7
8	Super Mario Galaxy 2	Platform	Nintendo EAD Tokyo	WII	9.7
9	Super Mario Galaxy	Platform	Nintendo EAD Tokyo	WII	9.7
10	The Orange Box	Shooter	Valve Software	X360	9.7

Genre Performance Across Platforms and Over Time: Examining how different genres fare on various gaming platforms and their popularity trends.

```

1  -- Performance of Genres Across Different Platforms and over time
2  SELECT year, platform, genre, SUM(global_sales) AS Total_Sales
3  FROM csv
4  GROUP BY year, genre, platform
5  ORDER BY Total_Sales desc
6  LIMIT 15;

```

#	year	platform	genre	Total_Sales
1	2006	WII	Sports	83.3
2	2009	WII	Sports	70.98999999999997
3	1985	NES	Platform	42.25
4	2008	WII	Racing	39.220000000000006
5	2005	DS	Simulation	37.99
6	2010	X360	Shooter	37.570000000000014
7	2011	PS3	Shooter	37.569999999999999
8	2009	PS3	Action	36.09
9	2006	DS	Platform	34.480000000000001
10	2001	PS2	Racing	33.829999999999984
11	2011	PS3	Action	33.360000000000001
12	2005	PS2	Action	32.620000000000005
13	2011	X360	Shooter	32.440000000000005
14	1989	GB	Puzzle	32.2
15	2009	WII	Platform	31.38

Regional Genre Performance: Comparing genre success in different geographic regions, offering insights into regional market tastes.

```

1  -- Name, Genre Performance by Region
2  SELECT name, genre, SUM(na_sales) AS NA_Sales, SUM(eu_sales) AS EU_Sales, SUM(jp_sales) AS JP_Sales, SUM(other_sales) AS Other_Sales
3  FROM csv
4  GROUP BY genre, name
5  ORDER BY SUM(global_sales) DESC
6  LIMIT 10;

```

#	name	genre	NA_Sales	EU_Sales	JP_Sales	Other_Sales
1	Wii Sports	Sports	41.49	29.02	3.77	8.46
2	Super Mario Bros.	Platform	29.08	3.58	6.81	0.77
3	Mario Kart Wii	Racing	15.85	12.88	3.79	3.31
4	Wii Sports Resort	Sports	15.75	11.01	3.28	2.96
5	Call Of Duty: Modern Warfare 3	Shooter	15.58	11.290000000000003	0.62	3.3500000000000005
6	Tetris	Puzzle	23.2	2.26	4.22	0.58
7	New Super Mario Bros.	Platform	11.38	9.23	6.5	2.9
8	Call Of Duty: Black Ops II	Shooter	14.080000000000002	11.049999999999999	0.72	3.88
9	Call Of Duty: Black Ops	Shooter	17.009999999999998	8.690000000000001	0.59	3.12
10	New Super Mario Bros. Wii	Platform	14.59	7.06	4.7	2.26

Top Performing Games in Each Genre: Showcasing standout titles within each genre for a focused look at market leaders

```

1  -- Top Performing Games in Each Genre
2  SELECT genre, name, critic_score, global_sales
3  FROM csv
4  WHERE (genre, global_sales) IN (
5      SELECT genre, MAX(global_sales)
6      FROM csv
7      GROUP BY genre
8  )
9  ORDER BY global_sales DESC
10 LIMIT 10;

```

Predictive Modeling

Data Preprocessing:

The dataset used for training and testing the model consists of video game sales information, including details on platform, genre, release year, and publisher. Missing values in the 'Year'

column were imputed with the mean value. Additionally, the 'Publisher' column was transformed, categorizing publishers with fewer than 50 occurrences as 'Small Publisher.'

Categorical variables such as 'Platform,' 'Genre,' and 'Publisher' were one-hot encoded to represent them numerically. Numerical features were standardized using the StandardScaler from scikit-learn.

Model Architecture:

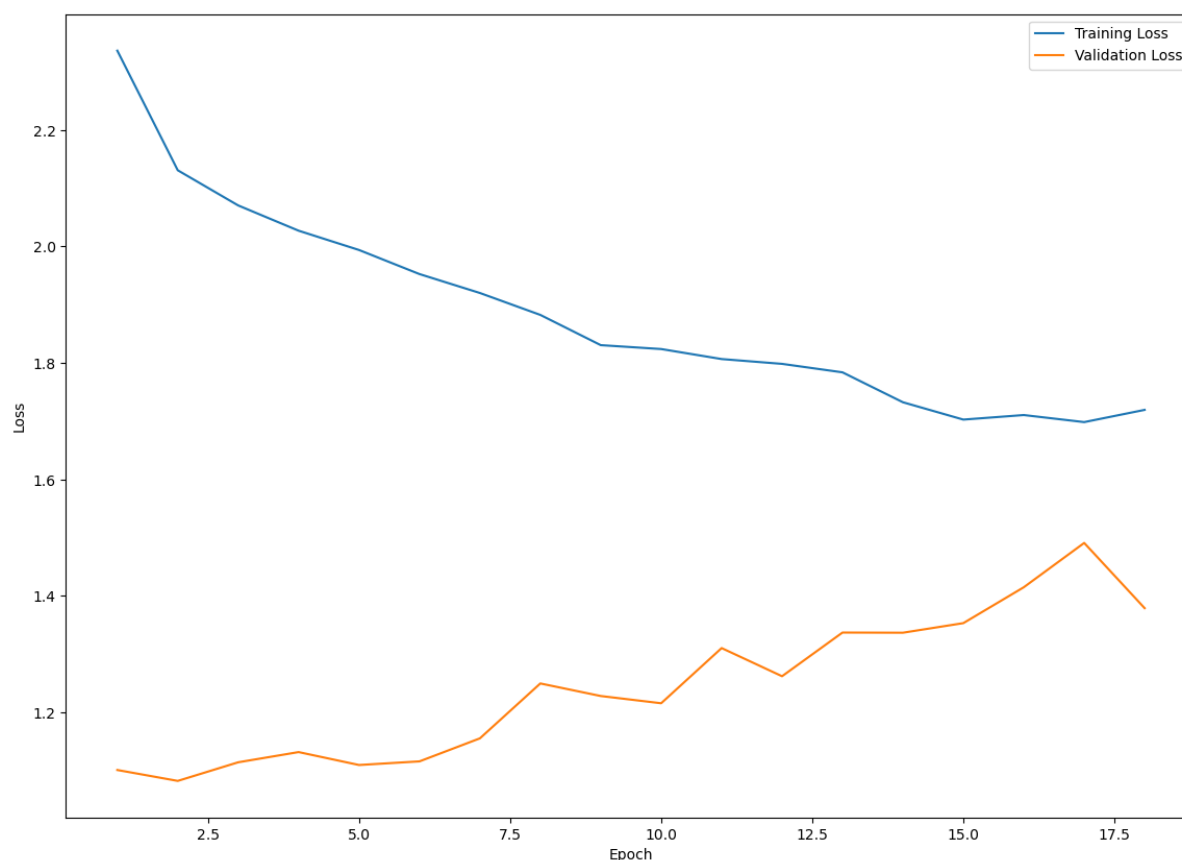
The neural network model comprises an input layer with 91 neurons (after one-hot encoding), two hidden layers with 128 neurons each, and an output layer with 1 neuron for regression purposes. The model is compiled with the mean squared error (MSE) loss function and RMSprop optimizer.

Model Training:

The dataset was split into training and validation sets, with 80% of the data used for training. The model was trained for 18 epochs, and the training process was visualized using a plot displaying the training and validation loss over epochs.

Model Evaluation:

The trained model can now be used for predicting global video game sales. Predictions are made on a test set, and performance is evaluated using the mean squared error (MSE) metric. This metric provides insights into how well the model generalizes to unseen data.



X-axis (Epoch): This axis represents the number of training epochs, which are iterations over the entire dataset. Each point on the x-axis corresponds to one complete pass through your training data.

Y-axis (Loss): The y-axis represents the loss, specifically the mean squared error (MSE) in your case. The loss is a measure of how well the model is performing on the training and validation sets. It quantifies the difference between the predicted values and the actual target values.

Training Loss (Blue Line): The blue line shows the training loss at each epoch. This is the value of the loss function computed on the training set. The goal during training is to minimize this loss, indicating that the model is learning and improving its predictive capability.

Validation Loss (Orange Line): The orange line shows the validation loss at each epoch. This is the value of the loss function computed on a separate validation set, which the model has not seen during training. The validation loss helps to assess how well the model generalizes to new, unseen data.

Legend: The legend in the upper left corner helps identify which line corresponds to training loss and which corresponds to validation loss.

The purpose of monitoring these losses over epochs is to identify patterns that might indicate issues like overfitting (if training loss keeps decreasing but validation loss starts increasing) or underfitting (if both losses remain high). It helps you understand the training progress and the model's ability to generalize to new data.

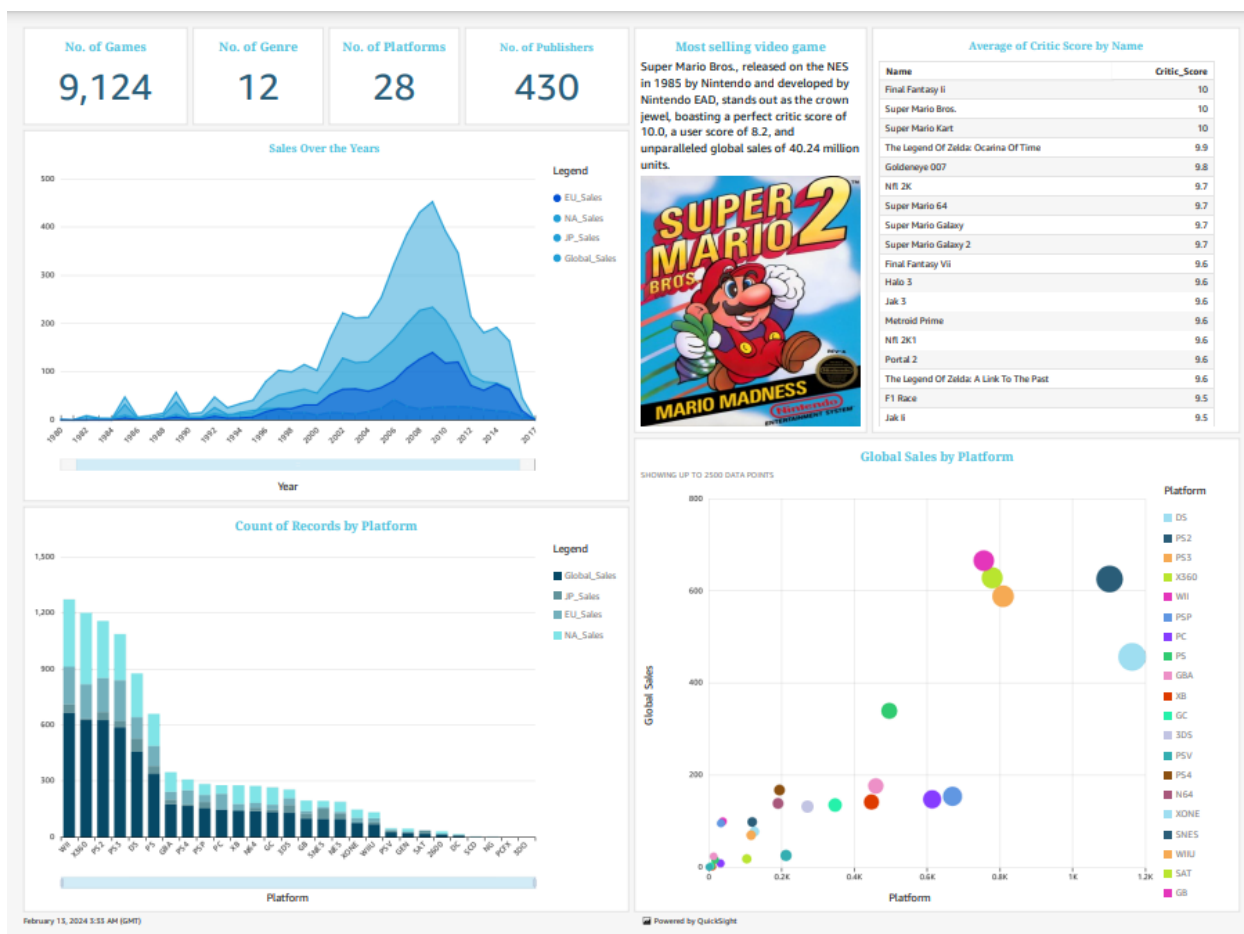
Results:

```
In [39]: y_test
Out[39]: Rank
          3341    0.60
          16176   0.01
          12303   0.06
          4426   0.44
          1243   1.51
          ...
          5805   0.31
          5074   0.38
          14130   0.03
          16126   0.01
          9768   0.12
          Name: Global_Sales, Length: 3308, dtype: float64
```

```
In [37]: y_train
Out[37]: Rank
          15075   0.02
          3397   0.59
           811   2.07
           757   2.16
          12726   0.06
          ...
          14525   0.03
          12617   0.06
           7150   0.22
           622   2.48
          13756   0.04
          Name: Global_Sales, Length: 13232, dtype: float64
```

Dashboard and Visualization

The dashboard presents a comprehensive analysis of video game sales, highlighting key metrics such as the number of games, genres, platforms, publishers, and sales over the years. It features detailed charts and visualizations, including sales by platform, critic scores for top games, and global sales distribution. Notably, "**Super Mario Bros.**" emerges as the top-selling game, showcasing its significant impact. This dashboard, powered by QuickSight, effectively synthesizes vast data into accessible insights, aiding stakeholders in understanding market trends and making informed decisions.



Conclusion

The comprehensive analysis, leveraging AWS services and advanced data processing techniques, provided deep insights into the video game industry's sales trends, consumer preferences, and market dynamics. By examining datasets, applying predictive modeling, and visualizing findings through QuickSight dashboards, we've uncovered valuable patterns that can inform strategic decisions. This project not only demonstrates the power of cloud computing in handling complex data analyses but also highlights the potential for data-driven approaches to anticipate market shifts and guide industry stakeholders towards informed, strategic decisions.