Video Game Sales Analysis

Title Page

- Dataset Name: Video Game Sales Analysis (Video Game Sales from Kaggle Dataset)
- Team Members: Zack Lee, Esther Law, Max Shuford

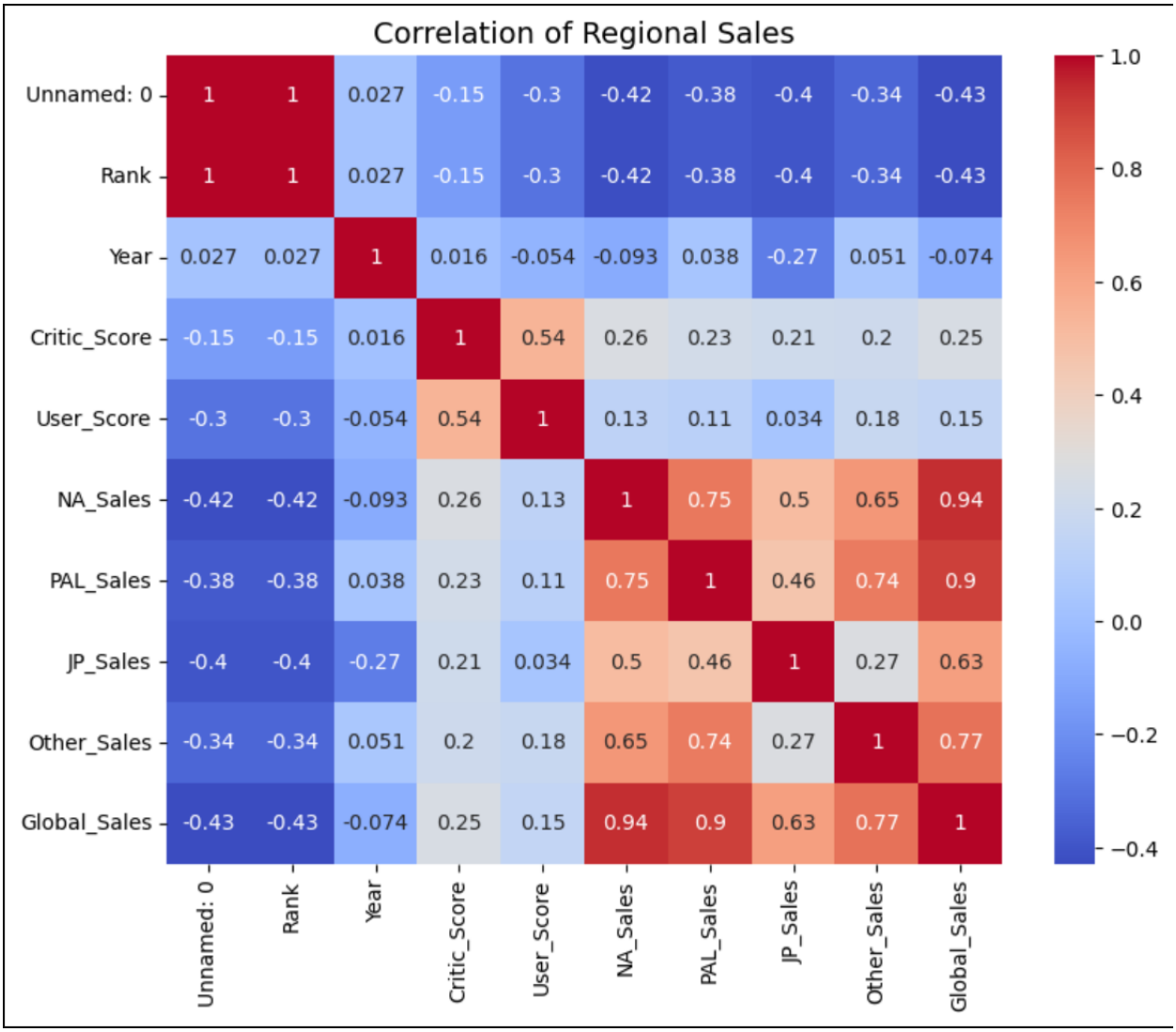
Introduction

- Dataset description: This dataset includes the year of the games released, the genre of the game, the name of the publisher, the name of the developer, as well as the sales of each game in North America, PAL region, and Japan as well as other sales and global sales.

Results

[Part 1: Zack Lee] for Student A

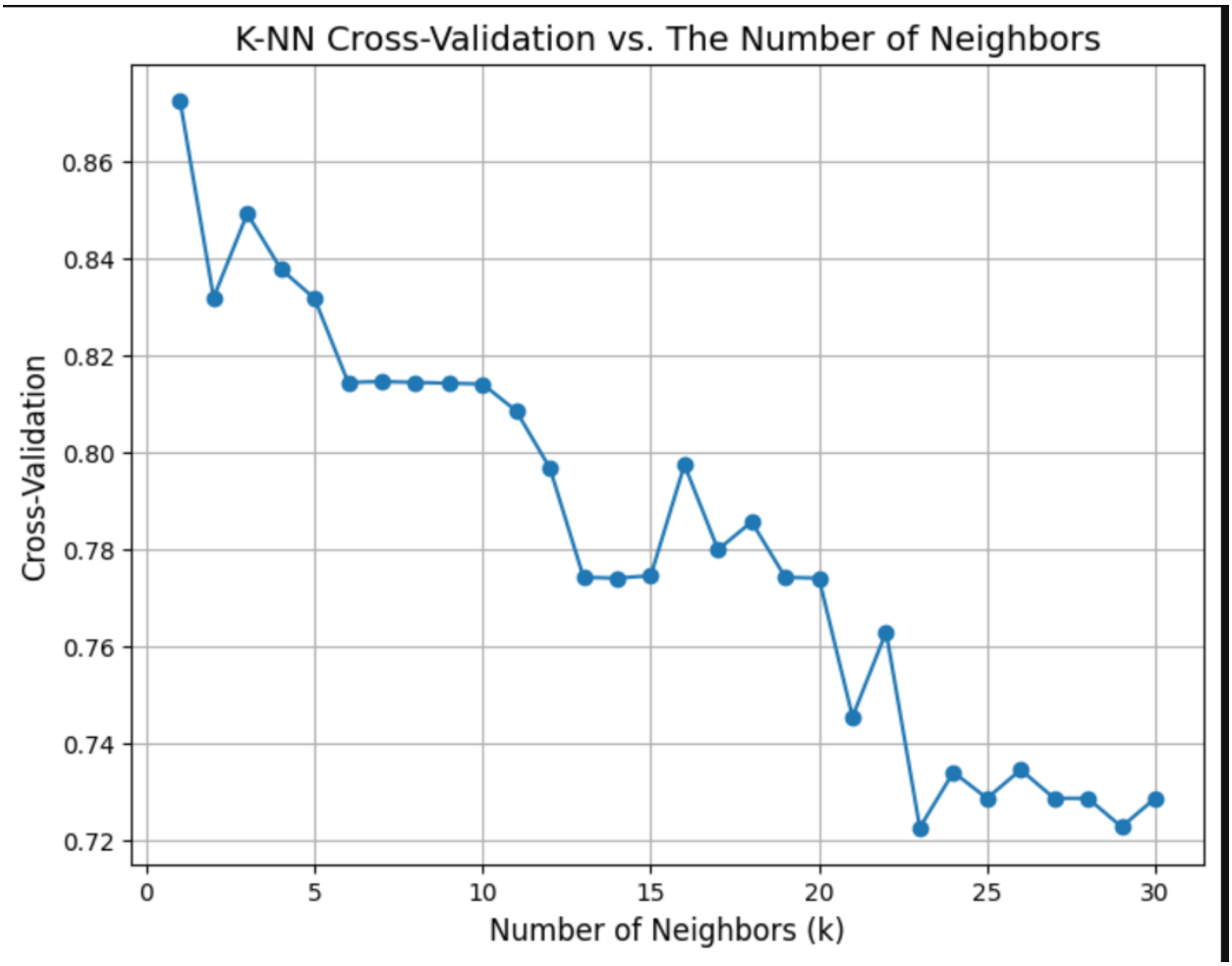
[Question 1]: Which regions' sales correlate most strongly with each other, and which ones correlate to global sales the most?



1. Purpose:
    a. The purpose of this correlation Matrix is to find which regions' sales correlate with each other the most.
2. Methodology:
    a. By using a correlation matrix, I was able to find the correlation between the different regional sales.
3. Explain the graph (e.g. what is x axis, what is y axis),
    a. In the correlation matrix, I have all the numeric values on the x and y axes, and then using that, it shows the correlation value between the different attributes, with the diagonal line being a correlation between itself so the value will always be 1. Then on the right is shows what the different values mean and the more red the value is or closer to 1 the better correlation and the bluer it is or closer to -1 the less correlation there is.
4. Harvest Highlights:

a. The highlights in this are that most of the regional sales depend on the NA region while the other sales rely on the PAL regional sales for their correlation. Then with the global sales the NA region has the highest correlatio

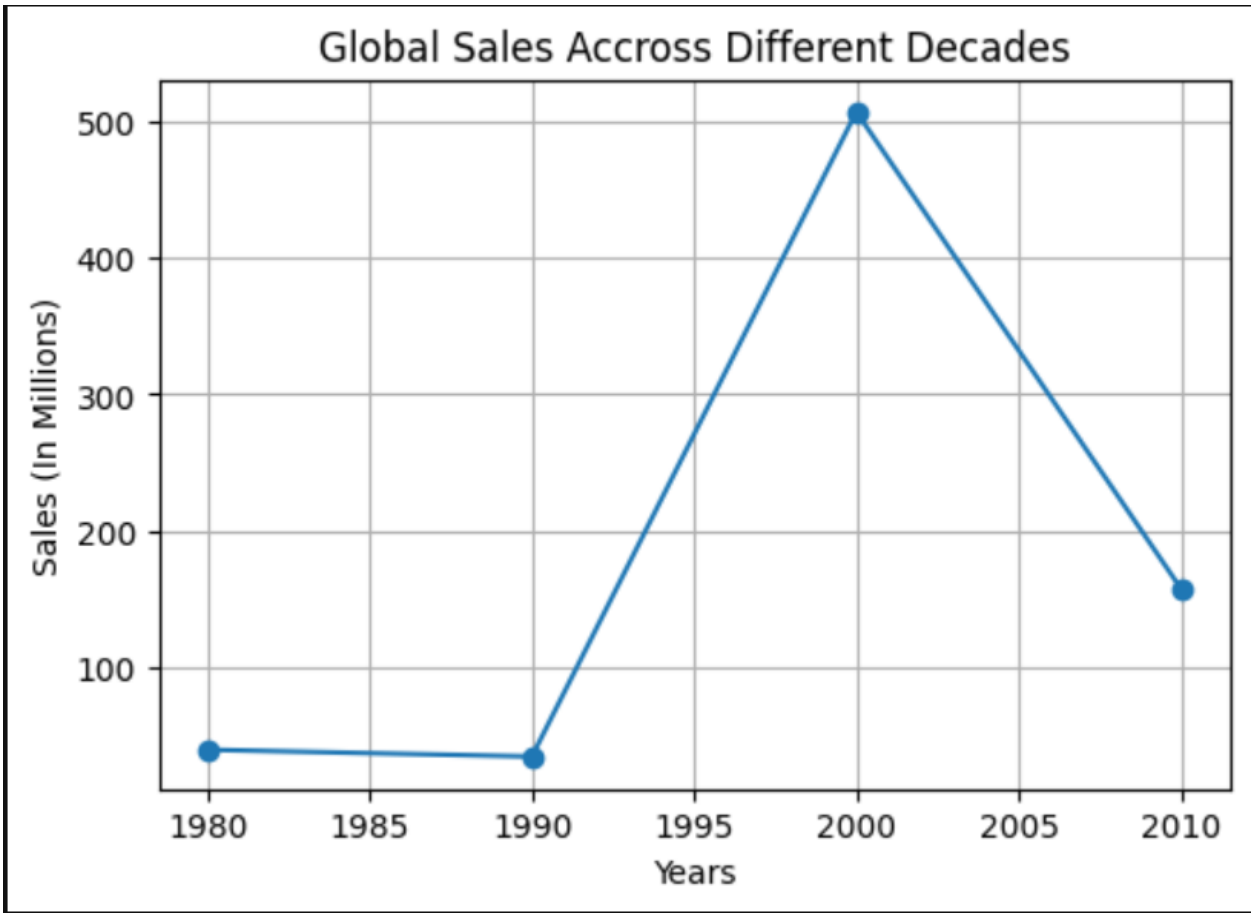
[Question 2]: Can we classify video games into low, medium, and high selling games based on their regional sales using KNN?



1. Purpose:
    a. The purpose of this is to correctly classify the games into different classes based on their global sales or how well they performed, using low, medium, and high metrics for the amount of global sales
2. Methodology:
    a. Using KNN classification to classify the data into different classes to help understand the data better
3. Explain the graph (e.g. what is x axis, what is y axis),
    a. The x-axis represents the tested values of k, and the y-axis represents the cross-validation score or the percentage of accurately classified games

4. Harvest Highlights: I was able to accurately classify the games into 3 different groups with an accuracy of 87% using a K value of 1. With the KNN classification and confusion matrix, for the low-selling games, I classified 24/25 correctly, for the medium games, 3/5 correctly, and for the high games, 16/23 games correctly.

[Question 3]: How have global sales changed over the decades?

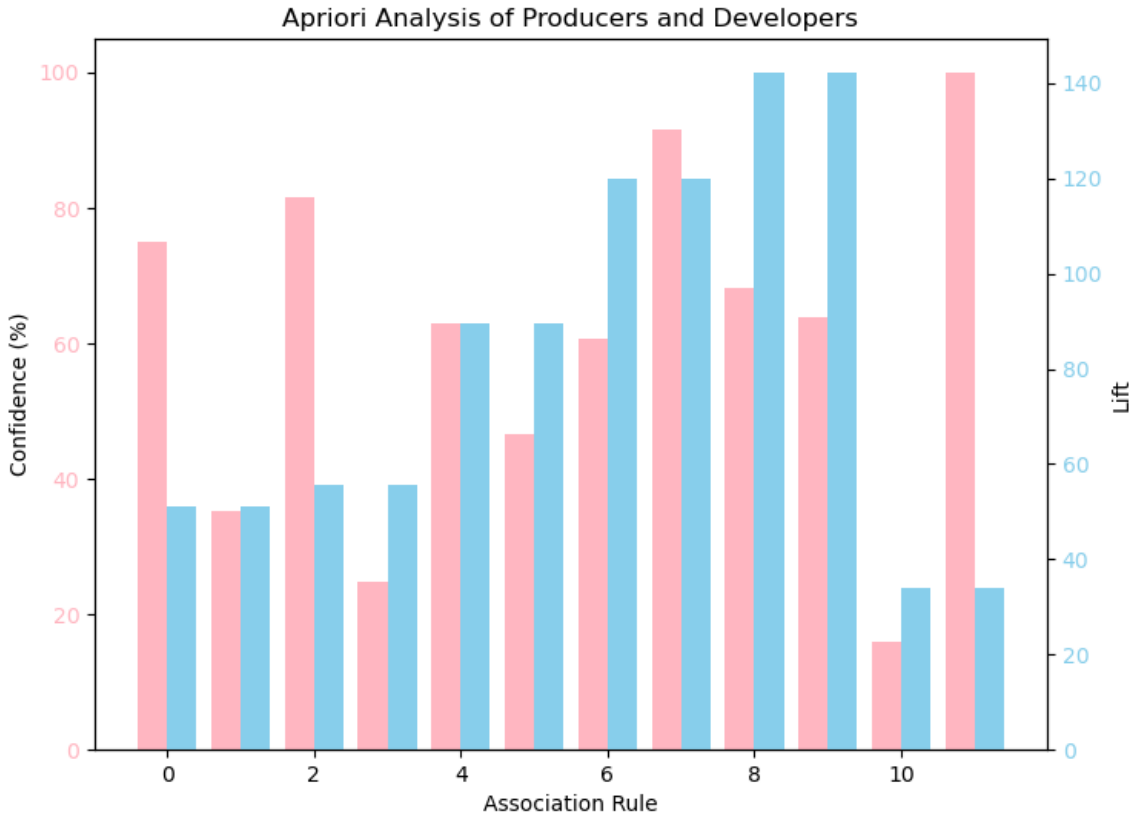


1. Purpose:
    a. The purpose of this graph and question (there are other images in the final project file) is to show the temporal change in global sales over the different decades
2. Methodology:
    a. By using the pandas, NumPy libraries, and matplotlib, I was able to dive deeper into the dataset, pull out the four different decades, the global sales per decade, and the average games sold per decade, the number of games released per decade, and plot my findings. I was also able to get specific publishers and their sales for each decade.
3. Explain the graph (e.g. what is x axis, what is y axis)
    a. The graph above shows the global sales per decade with the x-axis being the years or decade, and the y-axis being the global sales for that decade.
4. Harvest Highlights:

a. I think the biggest takeaway is the temporal trends over the years and how the global sales either skyrocketed or fell off, and based on the findings in the graph above, I was able to look for specific decades and find what games were released that contributed to those temporal trends. I was also able to see how many games were made by each publisher, each decade and how that also contributed to the global sales trends.

[Part 2: Esther Law] for Student B

[Question 4]: Are certain publishers strongly associated with specific developers?



1. Purpose

The purpose of this graph is to visualize the confidence and lift values of association rules resulting from the Apriori analysis of video game Producers and Developers.

2. Methodology

To conduct this analysis, we first read the CSV file into a DataFrame and filter out the publisher and developer columns into a list. We then use TransactionEncoder to convert the list of publishers and developers to one-hot format. We then run the Apriori Algorithm to retrieve the

total frequent itemsets. We then use these itemsets to generate the association rules. We then graph the information, converting confidence to their percentage values.

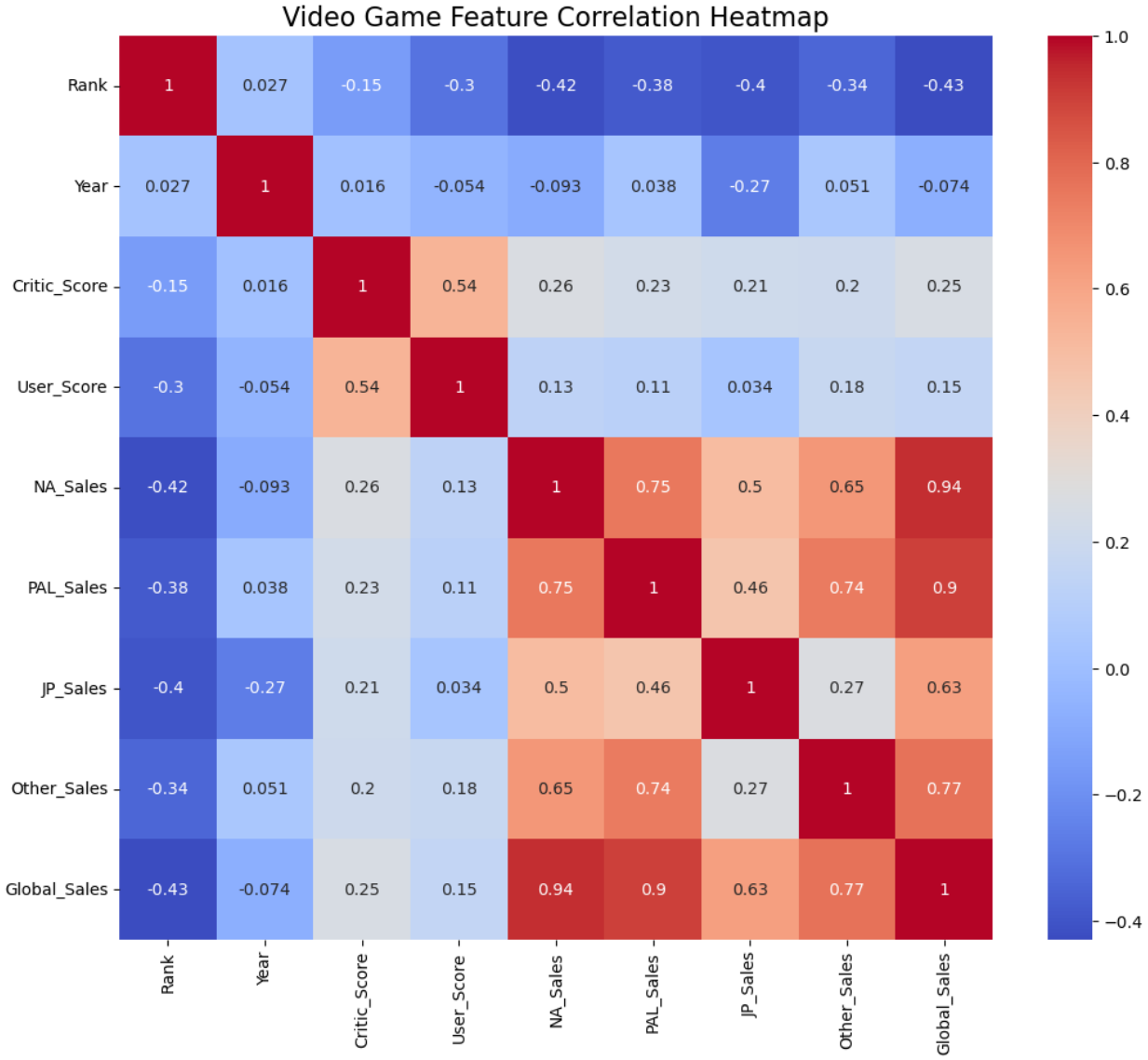
3. Explain the graph (e.g. what is x axis, what is y axis)

The x axis of the graph is each individual association rule and the y axis represents both the lift and confidence displayed side by side for comparison.

4. Harvest Highlights

The main highlights of this graph is to show the lift and confidence relationships in each discovered Association rule using producers and developers. We can see that the relationships between lift and confidence vary a lot and don't seem to have a distinct pattern.

[Question 5]: Which attributes have the highest correlation with critic scores?



1. Purpose

The purpose of this graph is to visualize what numeric features of the dataset have the highest correlation with good critic scores.

2. Methodology

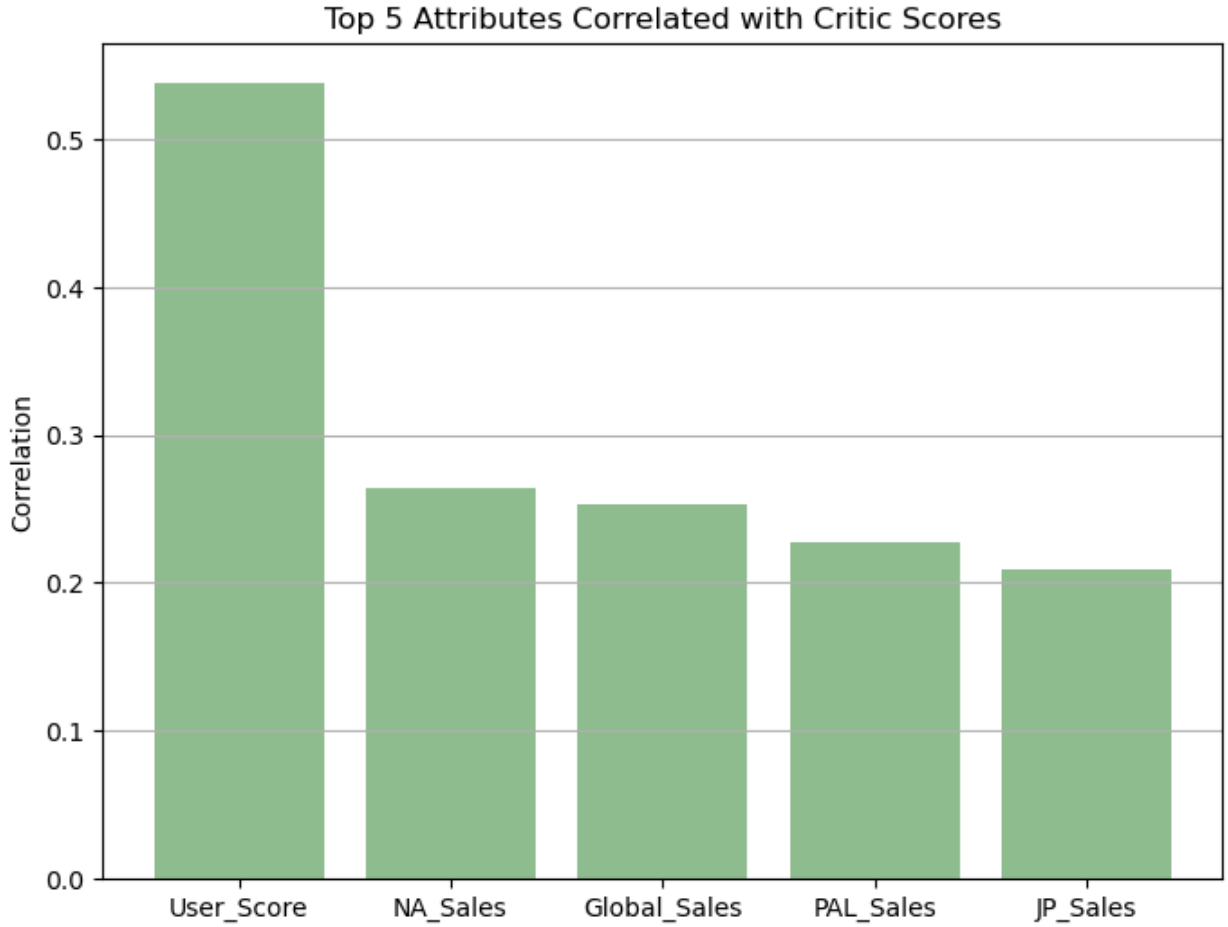
Filter the dataframe to only contain numeric values. Then use seaborn sns.heatmap() and df.corr() to visualize the correlation heatmap.

3. Explain the graph (e.g. what is x axis, what is y axis)

This graph contains all of the data set's numeric attributes and their correlations with each other as well as a key to represent the color correlation values.

4. Harvest Highlights

We can see that user scores are highly correlated with critic scores, along with various attributes representing video game sales.



1. Purpose
2. Methodology

We first store the correlation in a dataframe in descending order and use that to find the top 5 attributes correlated with critic scores. We then use this information to create a bar graph visualizing these values.

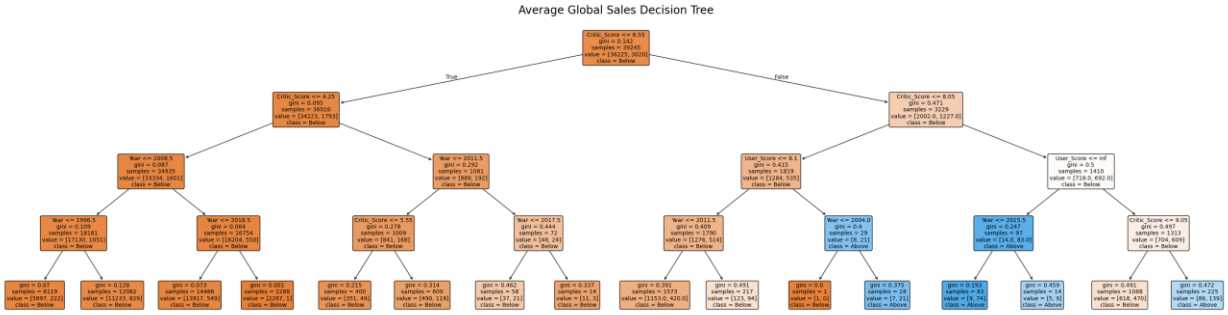
3. Explain the graph (e.g. what is x axis, what is y axis)

The x-axis of the graph represents the top 5 attributes correlated with critic scores. The y-axis represents the correlation values.

4. Harvest Highlights

This shows that the attributes with the highest correlation with critic scores are user scores, NA sales, global sales, PAL sales, and JP sales.

[Question 6]: Can we predict whether a video game's sales will be above or below average based on platform, developer, and year?



1. Purpose

The purpose of this graph is to visualize if a video game's global sales will be above or below average depending on its platform, developer, and year.

2. Methodology

To do this, we first find the mean of global sales in the data set and make a new column containing the above/below average value of each video game. We then use LabelEncoder(), and train_test_split from sklearning to encode the data and split it into testing data and training data. After that, we use DecisionTreeClassifier with max_depth=4 to train the model. After the model is trained, we plot the tree to visualize the results.

3. Explain the graph (e.g. what is x axis, what is y axis)

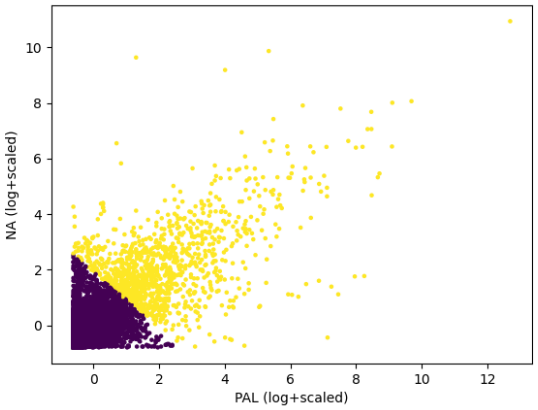
The graph shows four levels of decisions that can be made involving a video game's platform, developer, and year. Based on these decisions, the tree displays whether or not a game's global sales will be above or below average.

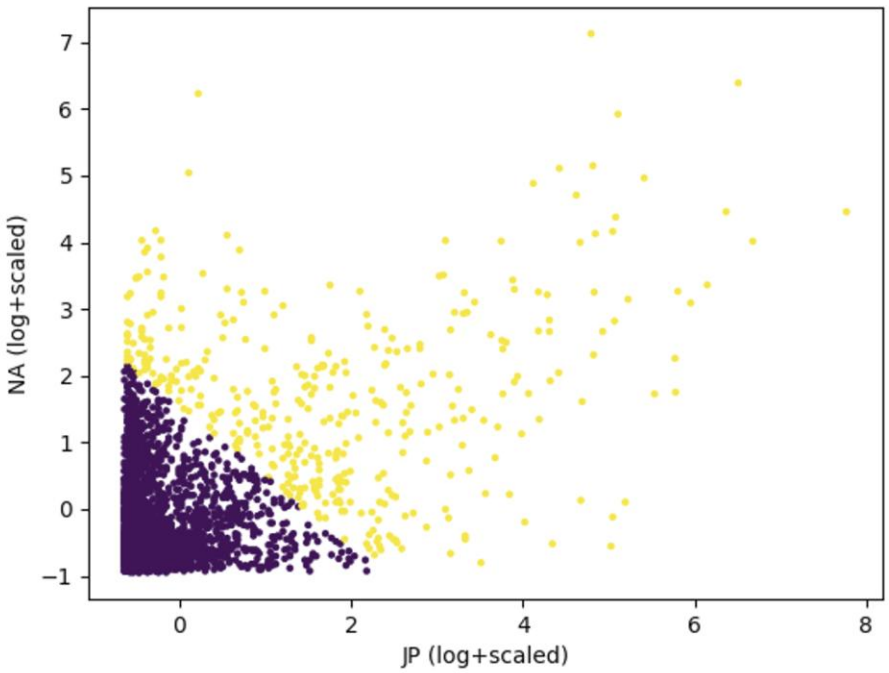
4. Harvest Highlights

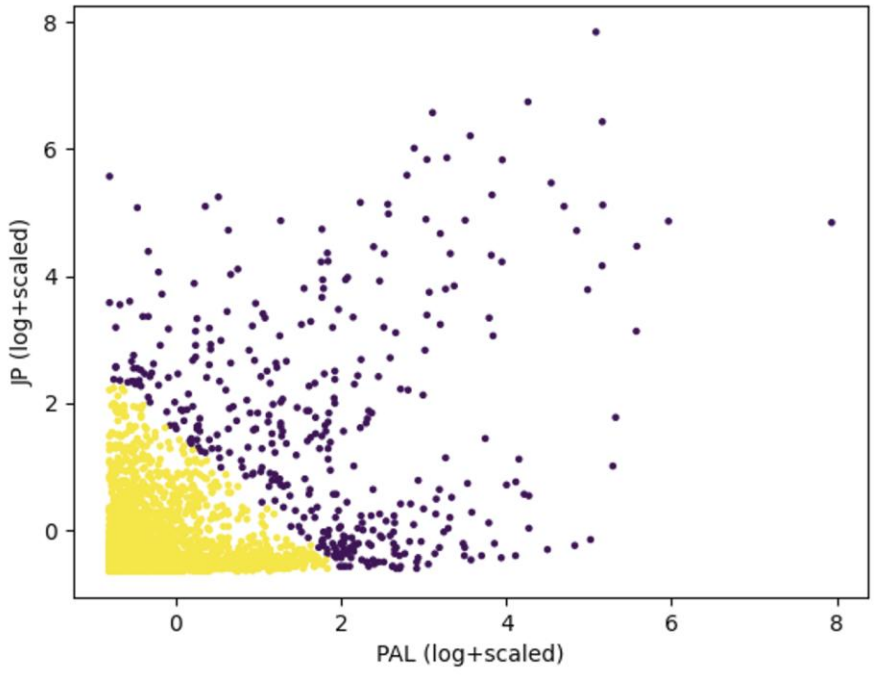
The decision tree shows that video games with higher critic and user scores are more likely to have above average global sales.

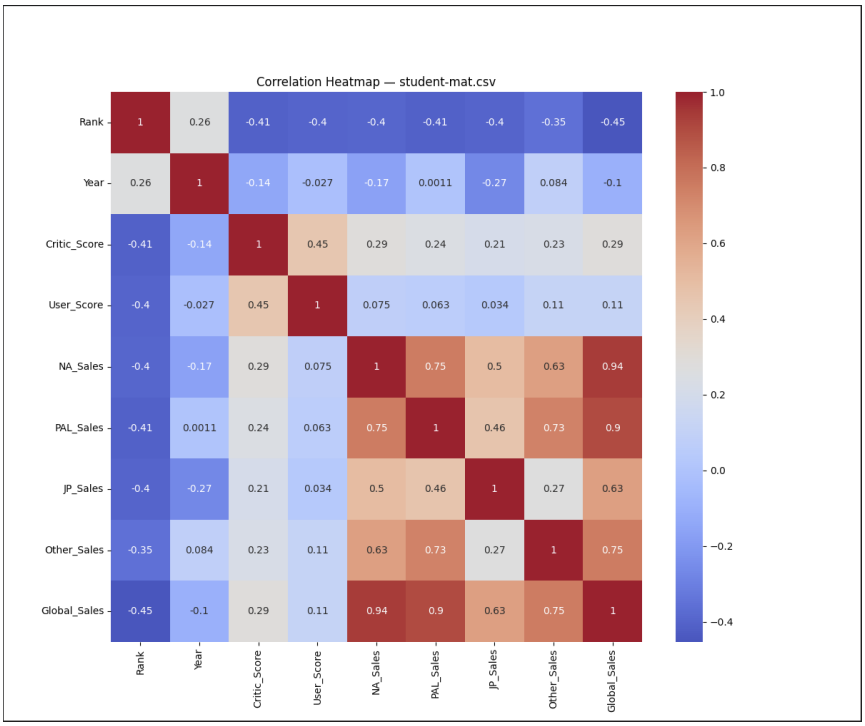
[Part 3: Max Shuford] for Student C

[Question 7]: Based on Regional Sales, can we determine a meaningful cluster of titles?


NA (log+scaled)
PAL (log+scaled)
0
2
4
6
8
10
12


NA (log+scaled)
JP (log+scaled)
7
6
5
4
3
2
1
0
−1
0
2
4
6
8


JP (log+scaled)
PAL (log+scaled)
8
6
4
2
0
0
2
4
6
8



[Your explanation: 1. Purpose, 2. Methodology, 3. Explain the graph (e.g. what is x axis, what is y axis), 4. Harvest Highlights.]

1. Purpose

The purpose of this question was to see if we can see any grouping of sales between Japan, European, and American games sales. The purpose of these groupings was to identify the bulk games vs the outlier games and the trends between regions.

2. Methodology

To analyze patterns in video game performance, I began by cleaning the dataset and removing entries with missing values in the sales columns used for clustering. Since raw sales data are heavily skewed and contain extreme outliers, I applied a log transformation followed by standardization to normalize the distribution of each feature. I selected JP_Sales and NA_Sales as the primary attributes for clustering, as these regions exhibit distinct market behaviors and provide useful insight into regional demand patterns. After preprocessing, I used the K-Means algorithm and evaluated different values of k using the silhouette score to determine the optimal number of clusters. The highest silhouette score was achieved at k = 2, indicating that the dataset naturally separates into two well-defined groups. This preprocessing and evaluation pipeline allowed me to create meaningful and interpretable clusters that reflect real differences in regional game sales performance.

3. Results

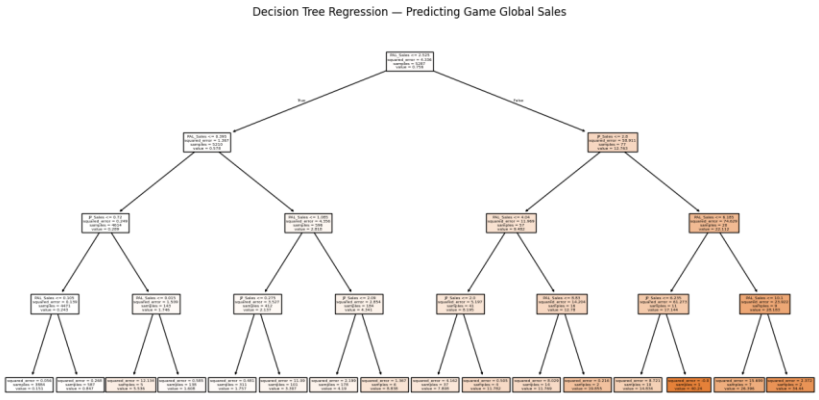
I performed K-Means clustering using two sales regions at a time. The scatterplots of PAL vs. JP Sales, PAL vs. NA Sales, and JP vs. NA Sales all produced a consistent clustering structure, where the algorithm identified two distinct groups of video games. Across all three visualizations, the first cluster represents the large majority of titles that sell at relatively low levels across each region. The second cluster contains a much smaller subset of high-performing

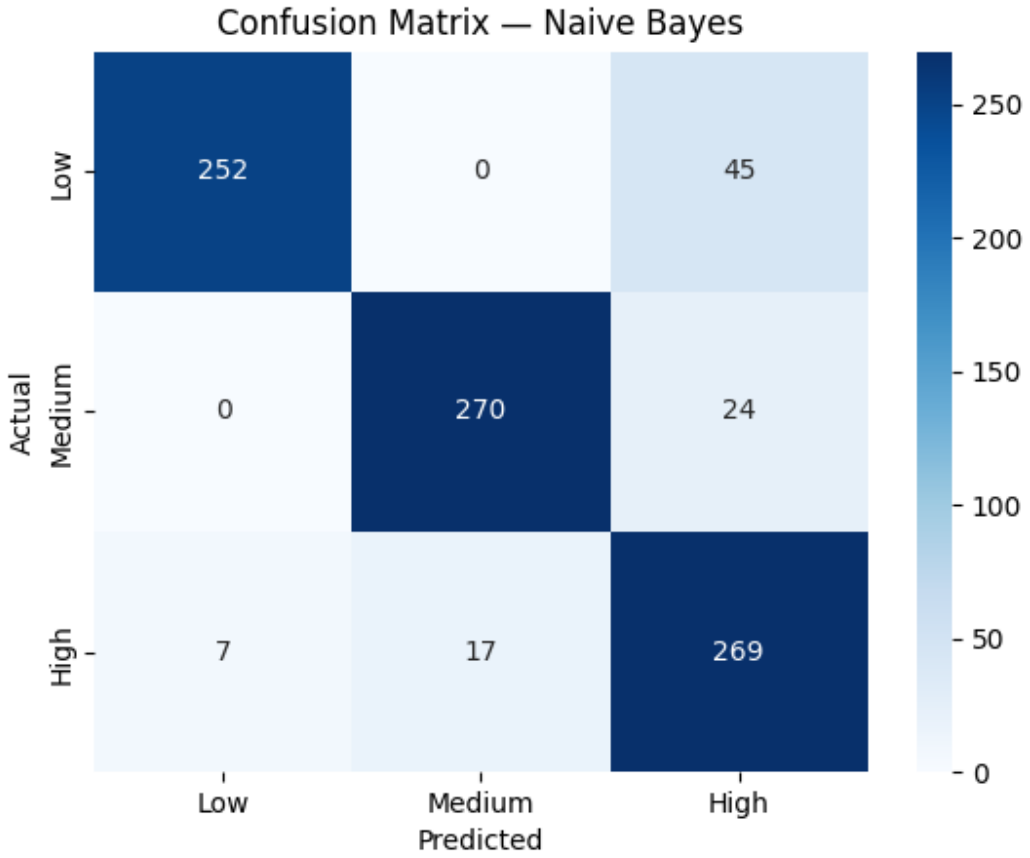
titles that achieve significantly stronger sales, forming a clear separation even after log scaling. This pattern suggests that sales behavior is dominated by a small number of blockbuster games that perform well across multiple regions, while the majority of games sell modestly. The consistent separation across different regional pairings reinforces that k = 2 is the most meaningful and stable number of clusters for this dataset. Overall, the clustering results effectively highlight the divide between globally successful titles and the broader market of lower-selling games.

4. Highlights

the K-Means algorithm consistently separated the dataset into two dominant groups. The first cluster contains the overwhelming majority of titles, characterized by low to moderate sales in each region, reflecting the typical performance of most games released on the market. In contrast, the second cluster isolates a much smaller subset of high-performing, top-selling games that achieve significantly stronger sales across multiple regions. This divide highlights the highly skewed nature of the video game industry, where only a limited number of blockbuster franchises generate large international sales while most titles remain regionally limited or modest in performance.

[Question 8]: Can we predict a game global sales based on its regional sales?





[Your explanation: 1. Purpose, 2. Methodology, 3. Explain the graph (e.g. what is x axis, what is y axis), 4. Harvest Highlights.]

1. Purpose
The purpose of this analysis is to evaluate whether regional video game sales can be used to predict a game's overall commercial performance. Specifically, the study investigates two related questions:
(1) whether regional sales can accurately estimate a game's global sales total, and
(2) whether a game can be classified into a sales performance tier—low, medium, or high—based on its sales in North America, Europe, Japan

2. Methodology
For regression analysis, a Decision Tree Regressor was trained using regional sales as predictors and global sales as the target variable, with a train–test split of 70/30. This model provides an interpretable structure that highlights which regions contribute most to predicting worldwide sales totals. To support a classification task, the continuous global sales variable was discretized into three equally sized tiers—Low, Medium, and High sellers—using quantile binning. A Gaussian Naive Bayes classifier was then trained to predict these tiers from the same regional sales inputs. Finally, both models were combined into a unified framework in which Naive

Bayes provides a categorical sales tier prediction, and the Decision Tree provides a numerical global sales estimate.

3. Results

The Decision Tree Regressor trained on log-transformed regional sales achieved meaningful predictive structure, revealing that PAL and JP sales were the strongest determinants of global sales. The regression tree produced interpretable sales partitions, with predicted global sales ranging from under one million units for low-performing titles to over twenty million units for top-selling games.

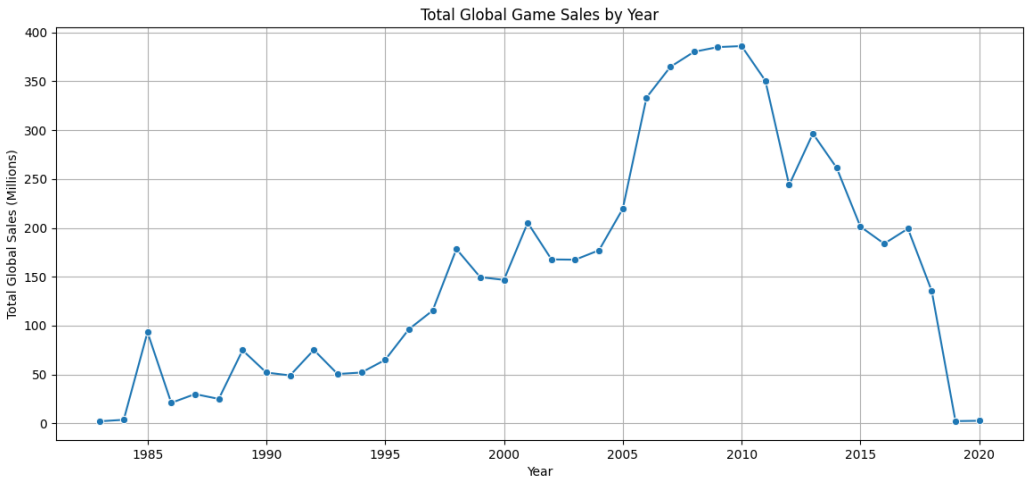
For the classification task, the Naive Bayes model demonstrated high accuracy in predicting whether a game fell into the Low, Medium, or High global sales tier. The classifier achieved an overall accuracy of 89%, with especially strong precision for the High-selling tier (0.97) and Low-selling tier (0.94). The Medium category showed slightly lower precision (0.80), reflecting a greater overlap in regional sales patterns for mid-range titles. The confusion matrix further confirmed that most misclassifications occurred between Medium and High sellers, while Low sellers were consistently well-identified.

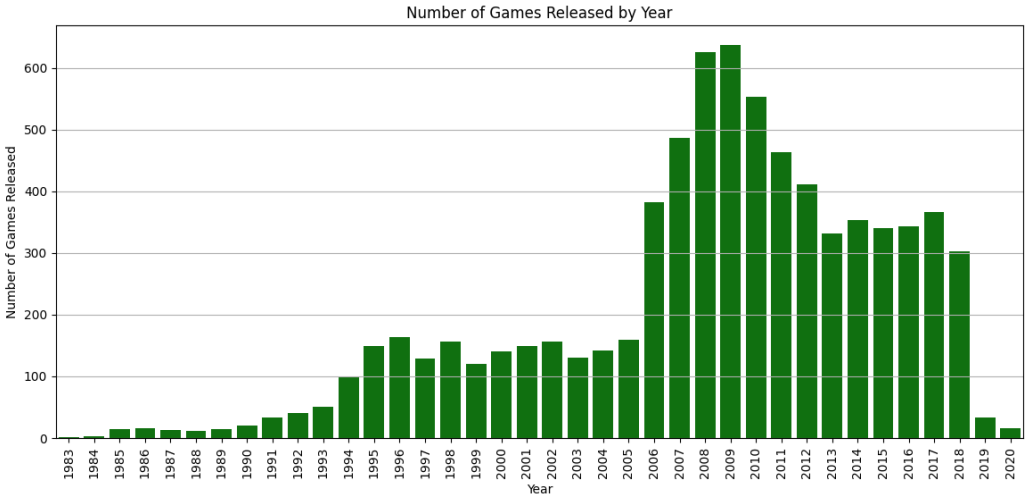
Taken together, the models show that regional sales provide both a reliable numerical estimate of global performance and a strong categorical signal that supports effective classification.

4. Harvest Highlights

Titles with strong PAL and JP sales were consistently associated with high global totals, as reflected in the structure of the Decision Tree Regressor. Second, the Naive Bayes classifier revealed that global sales naturally fall into three distinct tiers, and that these tiers are highly predictable from regional performance alone. The consistency of the results across both the regression and classification tasks highlights a clear pattern, regional markets are not isolated but instead act as strong predictors of worldwide performance. The analysis reinforces the idea that commercial success in one region often correlates with broader international appeal.

[Question 9]: What do the trends in game sales each year based on global sales show?





[Your explanation: 1. Purpose, 2. Methodology, 3. Explain the graph (e.g. what is x axis, what is y axis), 4. Harvest Highlights.]

1. Purpose

The purpose of this analysis is to examine how global video game sales have changed over time and to identify meaningful trends in yearly performance. By aggregating total global sales, average sales per game, and the number of games released each year, the goal is to determine how the video game market has evolved and to understand major shifts in consumer demand, industry growth, and market saturation.

2. Methodology

The dataset was filtered to include only entries with valid year and global sales values. The analysis begins by grouping the data by release year and computing three key metrics:
(1) total global sales per year
(2) average global sales per game
(3) the number of games released each year

These metrics were visualized using line plots and bar charts to identify macro-level trends across multiple decades of video game history. Line plots were used to represent total and average sales trends, while a bar chart was used to show yearly game release counts.

3. Results

Total global sales show a strong upward trajectory from the early 2000s through approximately 2008–2009, marking a period of rapid industry growth driven by major console platforms and high-volume franchises. After peaking around 2008, total sales begin a steady decline. This trend could be seen because of a slow move towards digital sales over physical sales or even mobile gaming.

Average global sales per game follow a similar pattern, peaking in the late 2000s before dropping off in later years. This decline is partially explained by the increasing number of games released per year, which surged during the mid-to-late 2000s and resulted in greater market saturation.

In the 2010s, both total and average sales stabilize, with occasional spikes corresponding to hits. These spikes indicate that the market became increasingly dependent on a smaller number of high-performing titles, while the majority of games sold at more modest levels.

4. Harvest Highlights

The analysis of yearly global sales shows several important patterns. Global video game sales increased steadily in the early and mid-2000s and reached their highest point around 2008 and 2009. After that peak, sales began to decline as the industry shifted away from physical game sales and toward digital downloads and online platforms.

The number of games released each year also grew during the mid-2000s, which helped explain why the average sales per game started to drop. As more games entered the market, individual titles sold fewer copies on average. In the later years, global sales became more stable, but large spikes still appeared when major blockbuster games were released.

Overall, the results show that the video game industry moved from a period of rapid growth to a more crowded and competitive market, and eventually toward a system that depends heavily on a few extremely successful titles. This helps explain how both industry trends and consumer behavior have changed over time.

Overview

1. Student 1: Zack Lee
    a. Which regions' sales correlate most strongly with each other, and which ones correlate to global sales the most?
    b. Can we classify video games into low, medium, and high selling games based on their regional sales using KNN?
    c. How have global sales changed over the decades?
2. Student 2: Esther Law
    a. Are certain publishers strongly associated with specific developers?
    b. Which attributes have the highest correlation with critic scores?
    c. Can we predict whether a video game’s sales will be above or below average based on platform, developer, and year?

3. Student 3: Max Shuford
    a. Based on Regional Sales, can we determine a meaningful cluster of titles?
    b. Can we predict a game global sales based on its regional sales?
    c. What do the trends in game sales each year based on global sales show?

Contributions

[Student A, describe your responsibility in this project with details, mention what the methodology you use, what the big challenge you meet.]

My responsibility in this project was to analyze relationships between regional video game sales, build a KNN classification model, and examine how global sales changed over the decades. I focused on identifying which regions correlated most strongly with one another and which had the highest influence on global sales. My methodology included cleaning and preparing the dataset, calculating correlation values, constructing a KNN classifier to categorize games into low, medium, and high sellers, and performing a temporal analysis by grouping global sales by year and visualizing long-term trends. The biggest challenge I faced was handling missing or inconsistent sales data, which required careful filtering and preprocessing before any models or trend analysis could be applied accurately.

[Student B, describe your responsibility in this project with details, mention what the methodology you use, what the big challenge you meet.]

My responsibilities in this project were to use Apriori Analysis to find associations between video game producers and developers, find the top attributes correlated with high critic scores, and create a decision tree using sklearning to determine whether a video game's sales will be above or below average based on platform, developer, and year. A challenge I met was finding the best way to graph the relationships between confidence and lift in my Apriori Analysis. I also had a hard time harvesting highlights from the information I found because there were no distinct patterns I could recognize.

[Student C, describe your responsibility in this project with details, mention what the methodology you use, what the big challenge you meet.]

My main responsibility was completing the temporal analysis and building the Naive Bayes and K-means models. I analyzed global sales trends by year, performed clustering on regional sales, and created a Naive Bayes classifier to place games into sales tiers. My methodology included preparing the data with log transformation and scaling, grouping sales by year, and applying K-means and Naive Bayes to uncover patterns and make predictions.
The biggest challenges I faced were dealing with missing or uneven sales data and selecting the correct modeling approach for each task. Once the data was properly prepared, the models produced clear trends and meaningful insights.