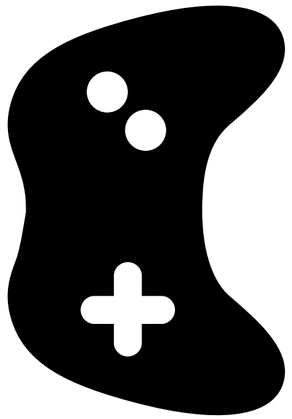
TEAM-2

Amit Talmale
Amit Saraswat
Keshav Pareek
Aaditya Saxena

MACHINE LEARNING



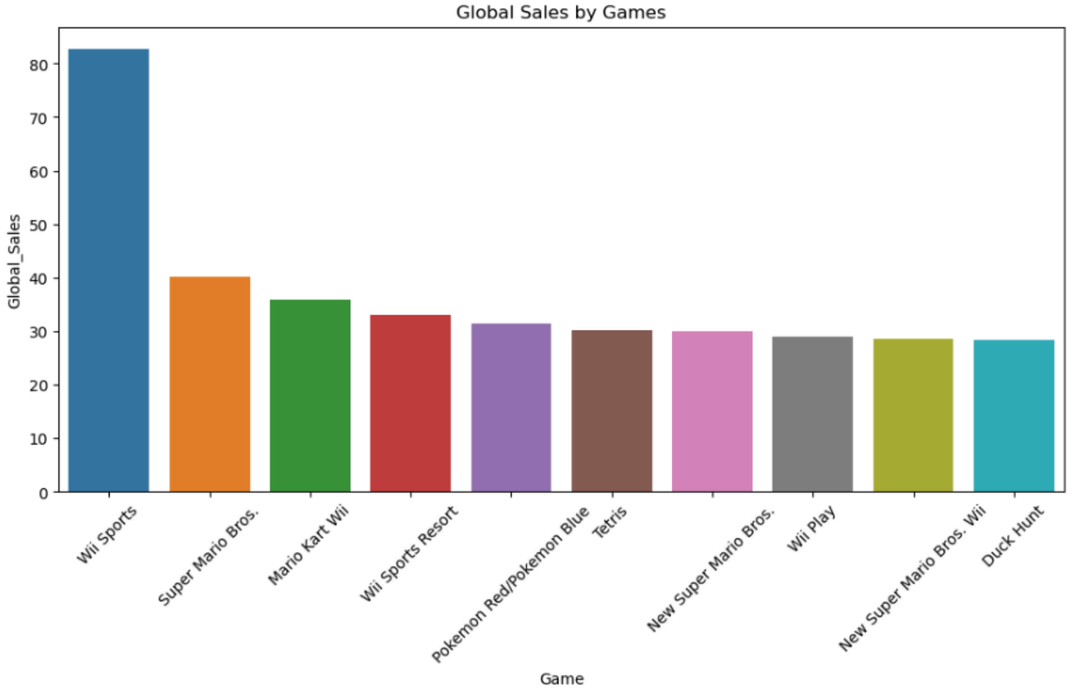
VIDEO GAMES SALES

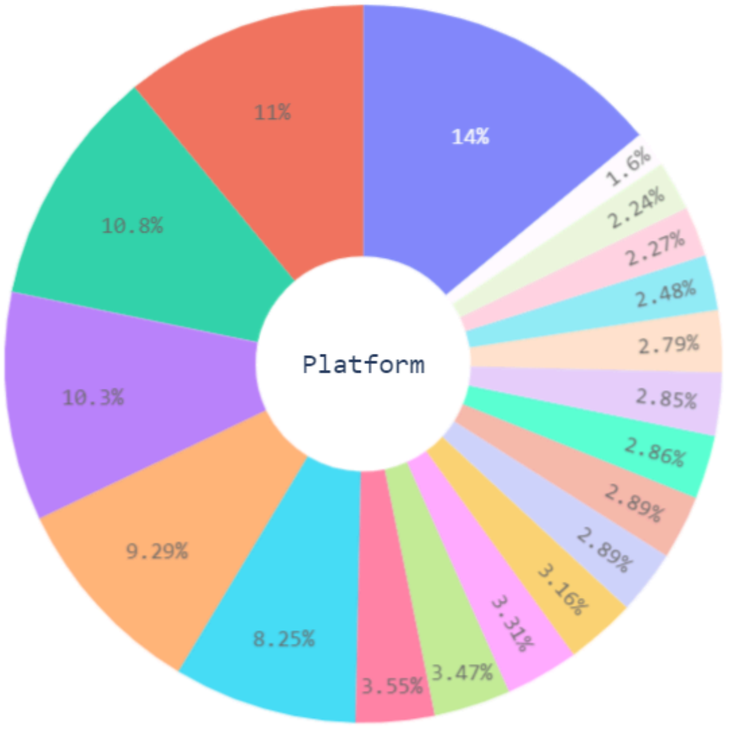
The video game market has experienced unprecedented growth in recent years, attracting both users and capital alike. With the exponential rise in the number of video game products available, implementing effective marketing strategies to maximize product revenue and assisting users in selecting products for optimal benefits have become crucial challenges.

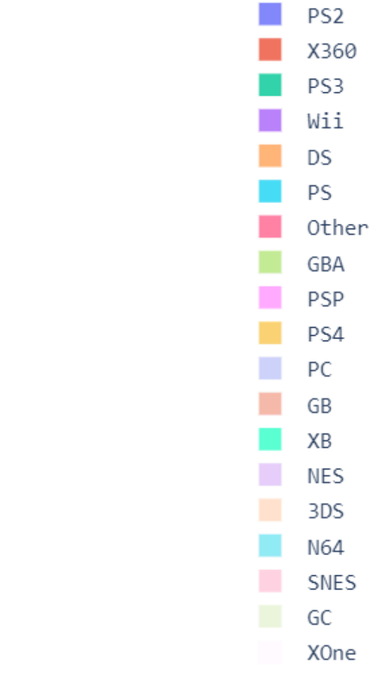
USD 242.39 billion in 2023

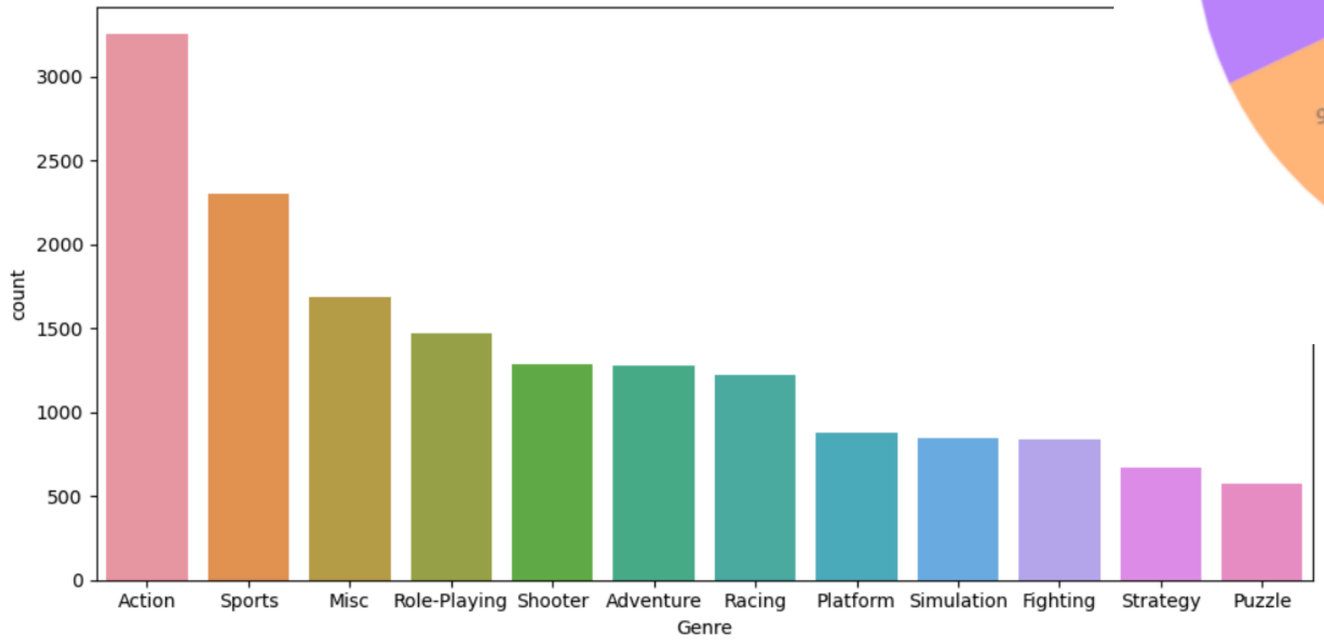
Dataset

Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33
5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1	31.37
6	Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.26
7	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.38	9.23	6.5	2.9	30.01
8	Wii Play	Wii	2006	Misc	Nintendo	14.03	9.2	2.93	2.85	29.02
9	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.59	7.06	4.7	2.26	28.62
10	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31
11	Nintendogs	DS	2005	Simulation	Nintendo	9.07	11	1.93	2.75	24.76
12	Mario Kart DS	DS	2005	Racing	Nintendo	9.81	7.57	4.13	1.92	23.42
13	Pokemon Gold/Pokemon Silver	GB	1999	Role-Playing	Nintendo	9	6.18	7.2	0.71	23.1
14	Wii Fit	Wii	2007	Sports	Nintendo	8.94	8.03	3.6	2.15	22.72
15	Wii Fit Plus	Wii	2009	Sports	Nintendo	9.09	8.59	2.53	1.79	22
16	Kinect Adventures!	X360	2010	Misc	Microsoft Game Studios	14.97	4.94	0.24	1.67	21.82
17	Grand Theft Auto V	PS3	2013	Action	Take-Two Interactive	7.01	9.27	0.97	4.14	21.4
18	Grand Theft Auto: San Andreas	PS2	2004	Action	Take-Two Interactive	9.43	0.4	0.41	10.57	20.81
19	Super Mario World	SNES	1990	Platform	Nintendo	12.78	3.75	3.54	0.55	20.61
20	Brain Age: Train Your Brain in Minutes a Day	DS	2005	Misc	Nintendo	4.75	9.26	4.16	2.05	20.22
21	Pokemon Diamond/Pokemon Pearl	DS	2006	Role-Playing	Nintendo	6.42	4.52	6.04	1.37	18.36
22	Super Mario Land	GB	1989	Platform	Nintendo	10.83	2.71	4.18	0.42	18.14
23	Super Mario Bros. 3	NES	1988	Platform	Nintendo	9.54	3.44	3.84	0.46	17.28
24	Grand Theft Auto V	X360	2013	Action	Take-Two Interactive	9.63	5.31	0.06	1.38	16.38
25	Grand Theft Auto: Vice City	PS2	2002	Action	Take-Two Interactive	8.41	5.49	0.47	1.78	16.15
26	Pokemon Ruby/Pokemon Sapphire	GBA	2002	Role-Playing	Nintendo	6.06	3.9	5.38	0.5	15.85
27	Pokemon Black/Pokemon White	DS	2010	Role-Playing	Nintendo	5.57	3.28	5.65	0.82	15.32
28	Brain Age 2: More Training in Minutes a Day	DS	2005	Puzzle	Nintendo	3.44	5.36	5.32	1.18	15.3
29	Gran Turismo 3: A-Spec	PS2	2001	Racing	Sony Computer Entertainment	6.85	5.09	1.87	1.16	14.98


Global Sales by Games
Global_Sales
Game
Wii Sports
Super Mario Bros.
Mario Kart Wii
Wii Sports Resort
Pokemon Red/Pokemon Blue
Tetris
New Super Mario Bros.
Wii Play
New Super Mario Bros. Wii
Duck Hunt


Platform
14%
11%
10.8%
10.3%
9.29%
8.25%
3.55%
3.47%
3.31%
3.16%
2.89%
2.89%
2.86%
2.85%
2.79%
2.48%
2.27%
2.24%
1.6%

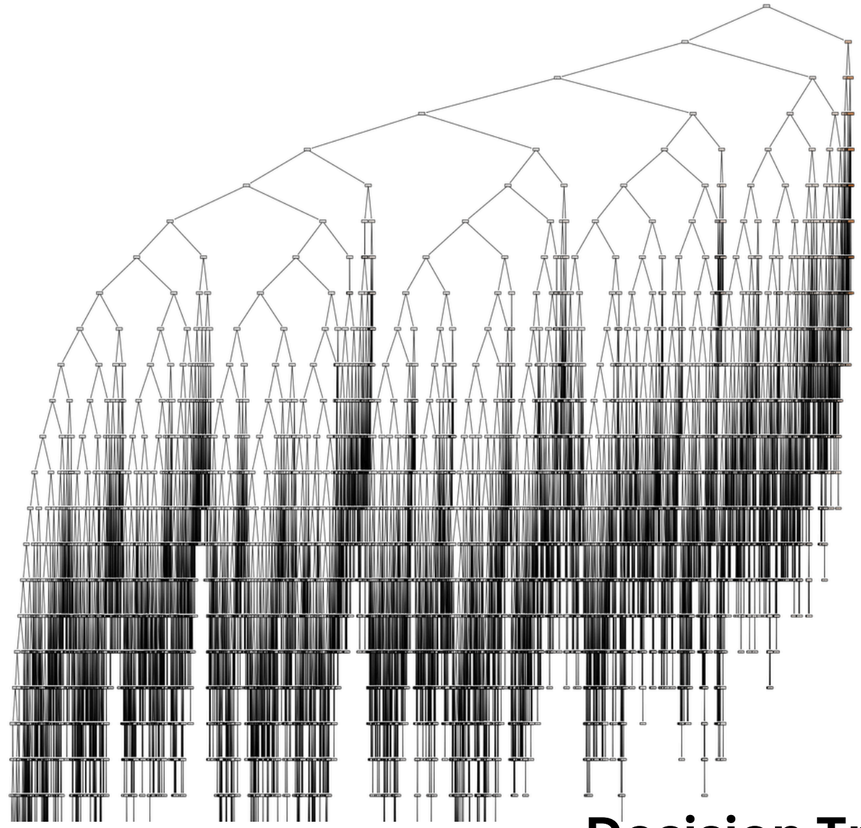

PS2
X360
PS3
Wii
DS
PS
Other
GBA
PSP
PS4
PC
GB
XB
NES
3DS
N64
SNES
GC
XOne

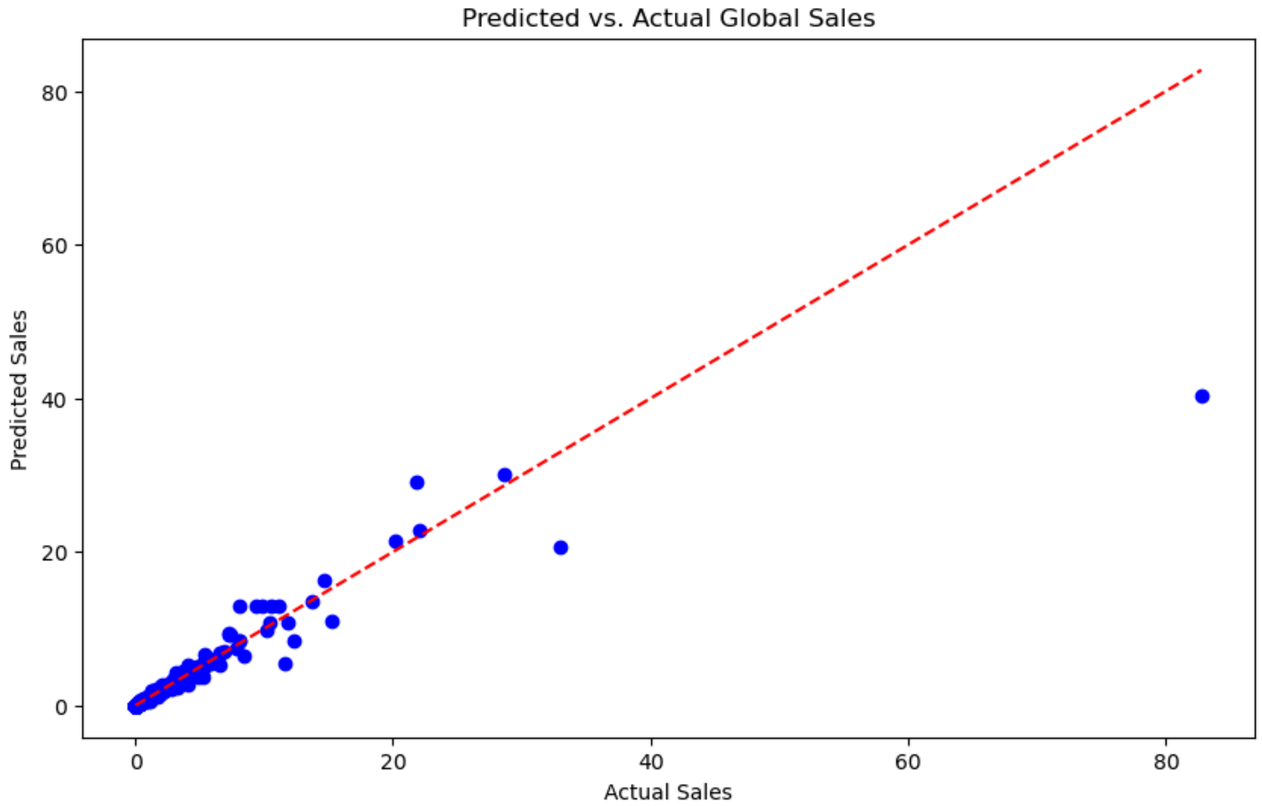

count
Genre
Action
Sports
Misc
Role-Playing
Shooter
Adventure
Racing
Platform
Simulation
Fighting
Strategy
Puzzle

Decision Tree

A decision tree in machine learning is a predictive modeling tool that's used to map out possible outcomes and decisions based on input data. It resembles a flowchart where each internal node represents a "test" on an attribute (e.g., whether a feature is above or below a certain value), each branch represents the outcome of that test, and each leaf node represents a class label or a decision.

1. Data Loading and Preprocessing:
    - The dataset ('vgsales.csv') is loaded into a pandas DataFrame.
    - Missing values are filled with the mean of each column.
2. Feature Selection and Target Definition:
    - Features ('Year', 'NA_Sales', 'EU_Sales', 'JP_Sales', 'Other_Sales') and the target variable ('Global_Sales') are selected.
3. Train-Test Split:
    - The dataset is split into training and testing sets (80% training, 20% testing) using train_test_split().
4. Model Initialization and Training:
    - A Decision Tree Regressor model is initialized with a fixed random state.
    - The model is trained on the training data using fit().
5. Decision Tree Visualization:
    - The trained decision tree is visualized using plot_tree() to understand its structure.
6. Prediction and Evaluation:
    - Predictions are made on the test data using predict().
    - The R^2 score is calculated to evaluate the model's performance.





Decision Tree Model Evaluation:
R^2 Score: 0.842756804328362

- The R^2 score, also known as the coefficient of determination, measures how well the model explains the variability of the target variable.
- An R^2 score of 1 indicates a perfect fit, where the model perfectly predicts the target variable.
- An R^2 score of 0 means that the model does not explain any of the variability of the target variable.
- Negative R^2 values indicate that the model is worse than simply predicting the mean of the target variable.

In this case, an R^2 score of 0.842 indicates that the model explains approximately 84.2% of the variability in the global sales of video games based on the selected features. This suggests that the model is performing well and is able to capture a significant portion of the variation in the target variable.

SVM (support vector machine)

SVM stands for Support Vector Machine, and it's a powerful supervised learning algorithm used for classification and regression tasks in machine learning.

Problem statement : we aim to develop a predictive model to forecast global sales of video games based on various features such as platform, genre, and publisher. Additionally, we'll explore the conversion of this regression task into a binary classification task based on a predefined threshold for sales.

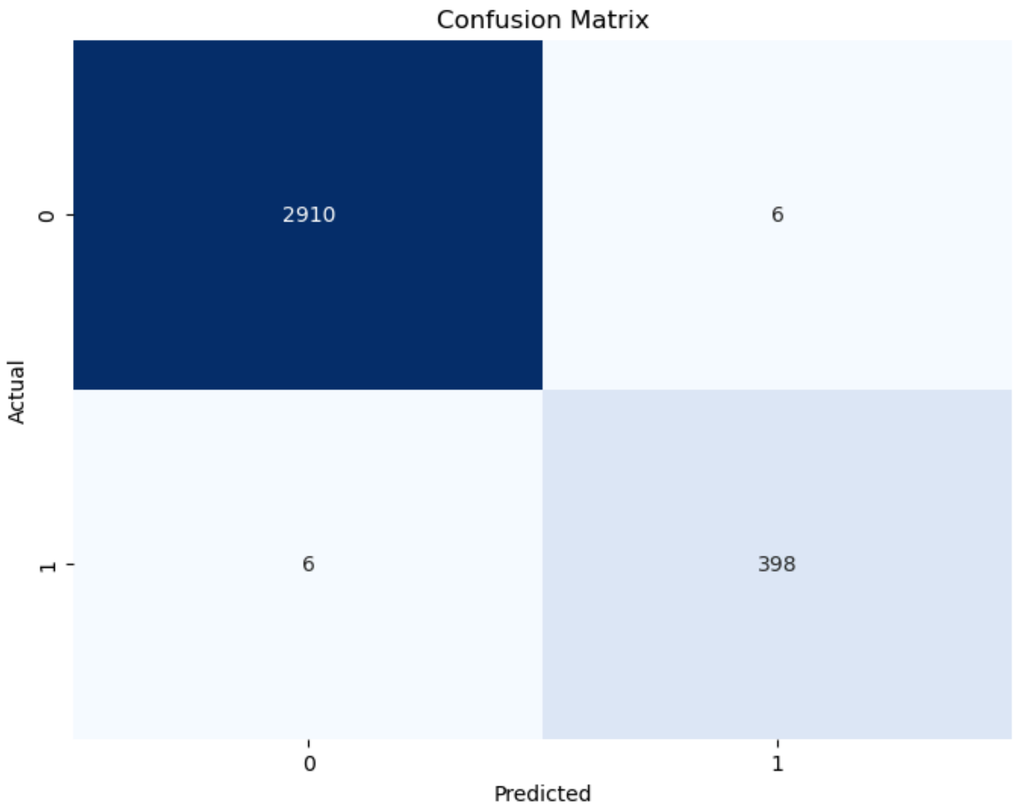
1. Loads the dataset.
2. Encodes categorical variables using LabelEncoder.
3. Splits the dataset into features (X) and the target variable (y).
4. Splits the dataset into training and testing sets..
5. Makes predictions on both the training and testing sets.
6. Evaluates the model using root mean squared error (RMSE).
7. Visualizes the actual vs predicted sales.

Result

```
Train RMSE: 0.7997838035269803
Test RMSE: 1.6193541355534162
Train Accuracy: 0.9950293718933574
Test Accuracy: 0.9963855421686747
Classification Report for Test Data:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2916
           1       0.99      0.99      0.99       404

    accuracy                           1.00      3320
   macro avg       0.99      0.99      0.99      3320
weighted avg       1.00      1.00      1.00      3320

Confusion Matrix:
[[2910    6]
 [   6  398]]
```

- RMSE (Root Mean Square Error): The RMSE values for both the training and testing sets are relatively low, indicating that the model's predictions are close to the actual values.
- Accuracy: The accuracy scores for both the training and testing sets are exceptionally high, close to 1. This suggests that the model's classification predictions match the actual labels with high accuracy.
- Classification Report: The precision, recall, and F1-score for both classes (0 and 1) are high, indicating good performance in classifying instances into the respective classes.

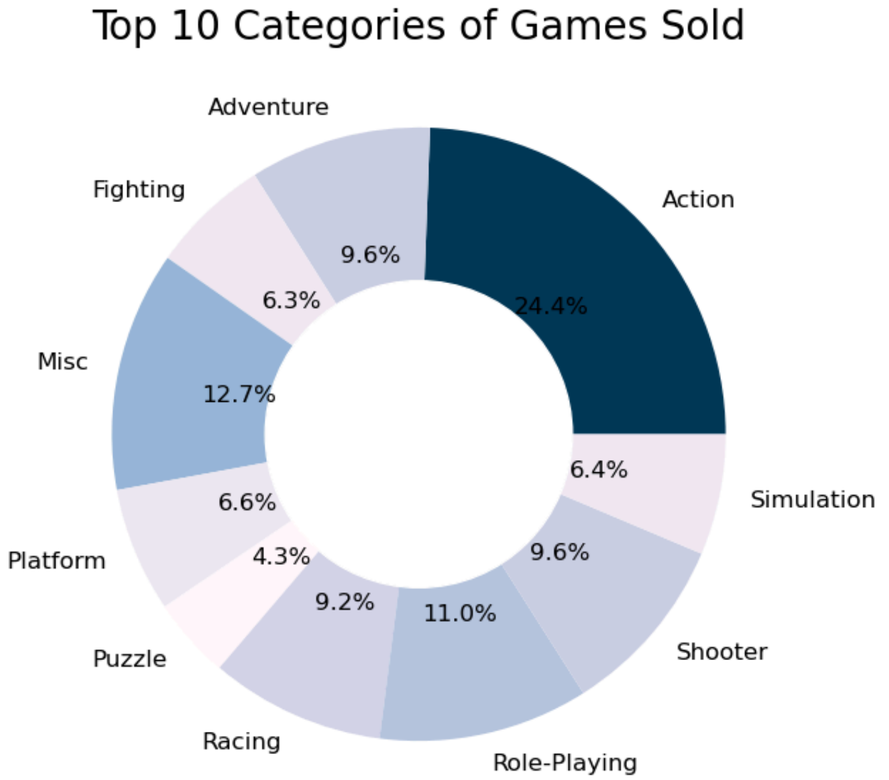
Overall, based on these metrics, the model seems to perform very well on both the regression (RMSE) and classification (Accuracy, Precision, Recall, F1-score) tasks.



Confusion Matrix

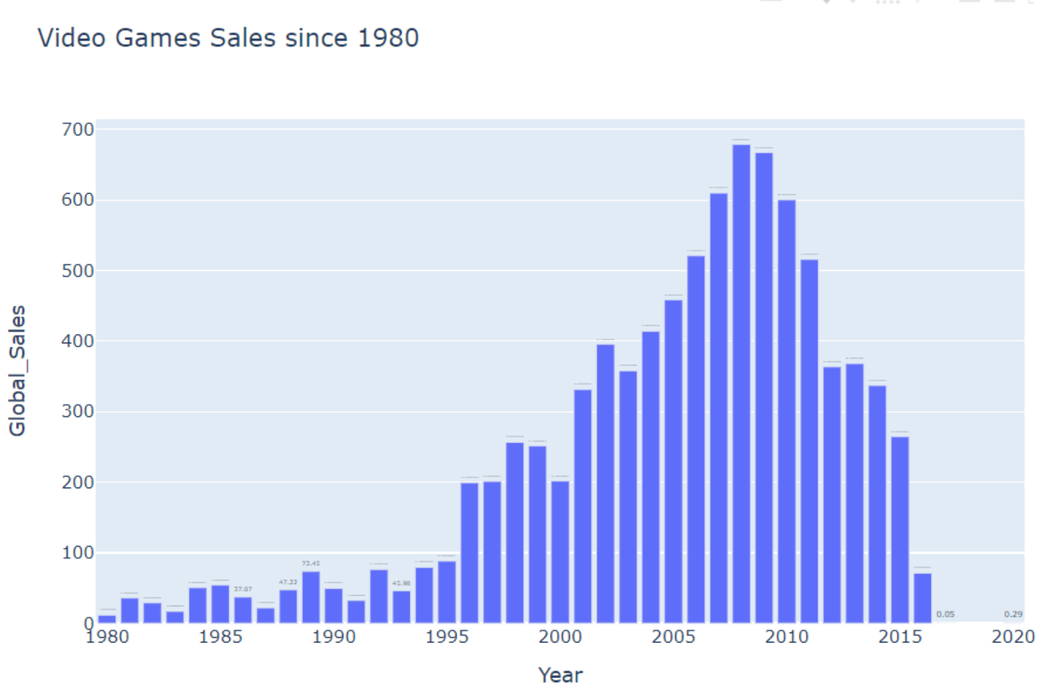
Linear Regrerssion

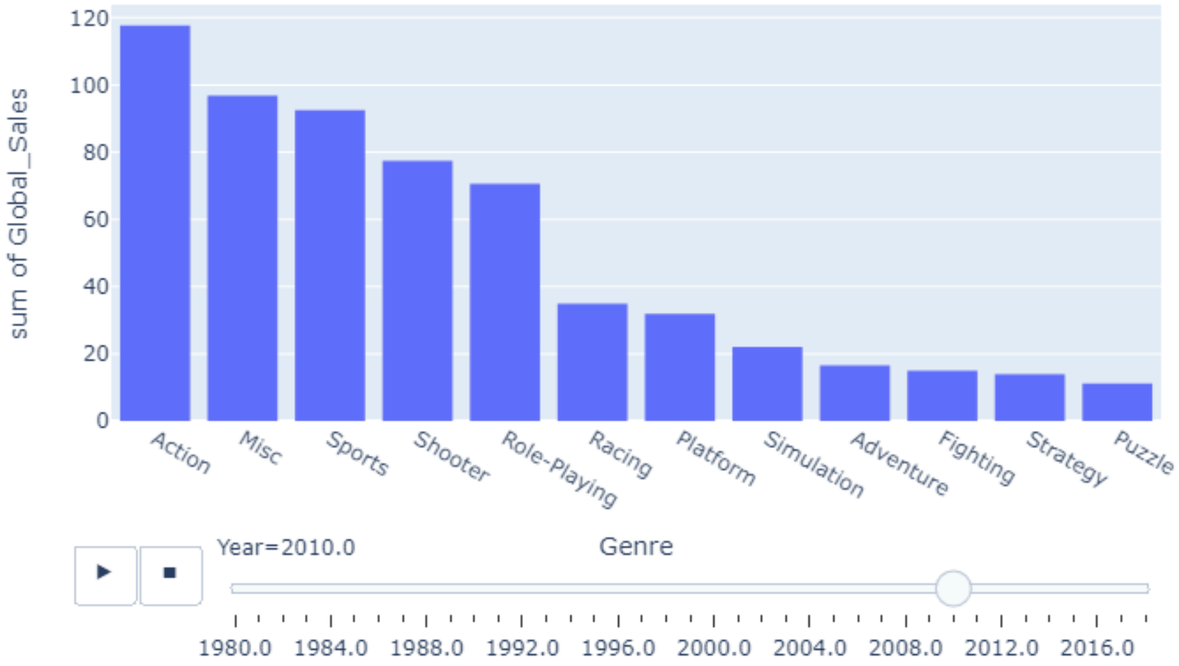
Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship and estimates coefficients to minimize the difference between observed and predicted values. The coefficients represent the impact of independent variables on the dependent variable. Linear regression is widely used for prediction and understanding relationships in various fields.



Abstract

Examines the relationship between various independent variables and global video game sales.
Considers factors such as critic ratings, user ratings, platform, genre, and regional sales data.
It aims to provide insights for game developers to understand market preferences and optimize their strategies.



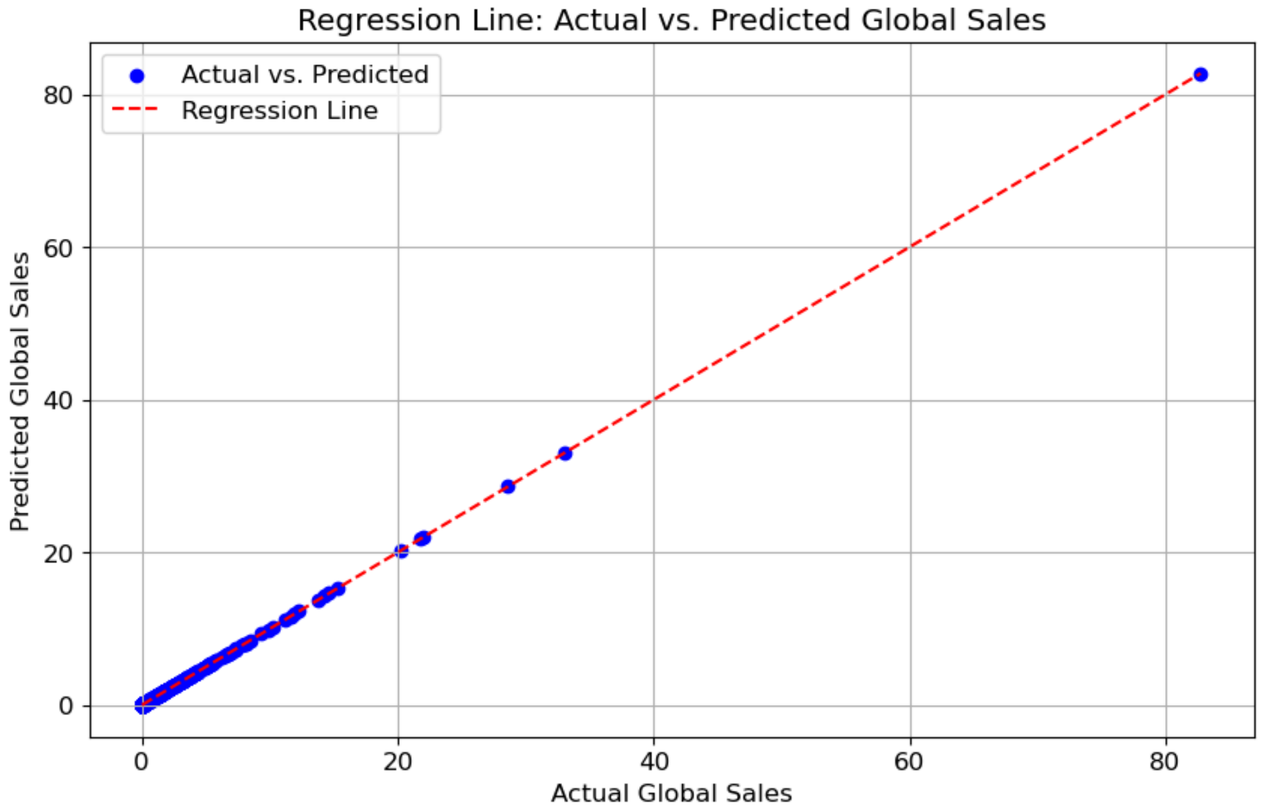


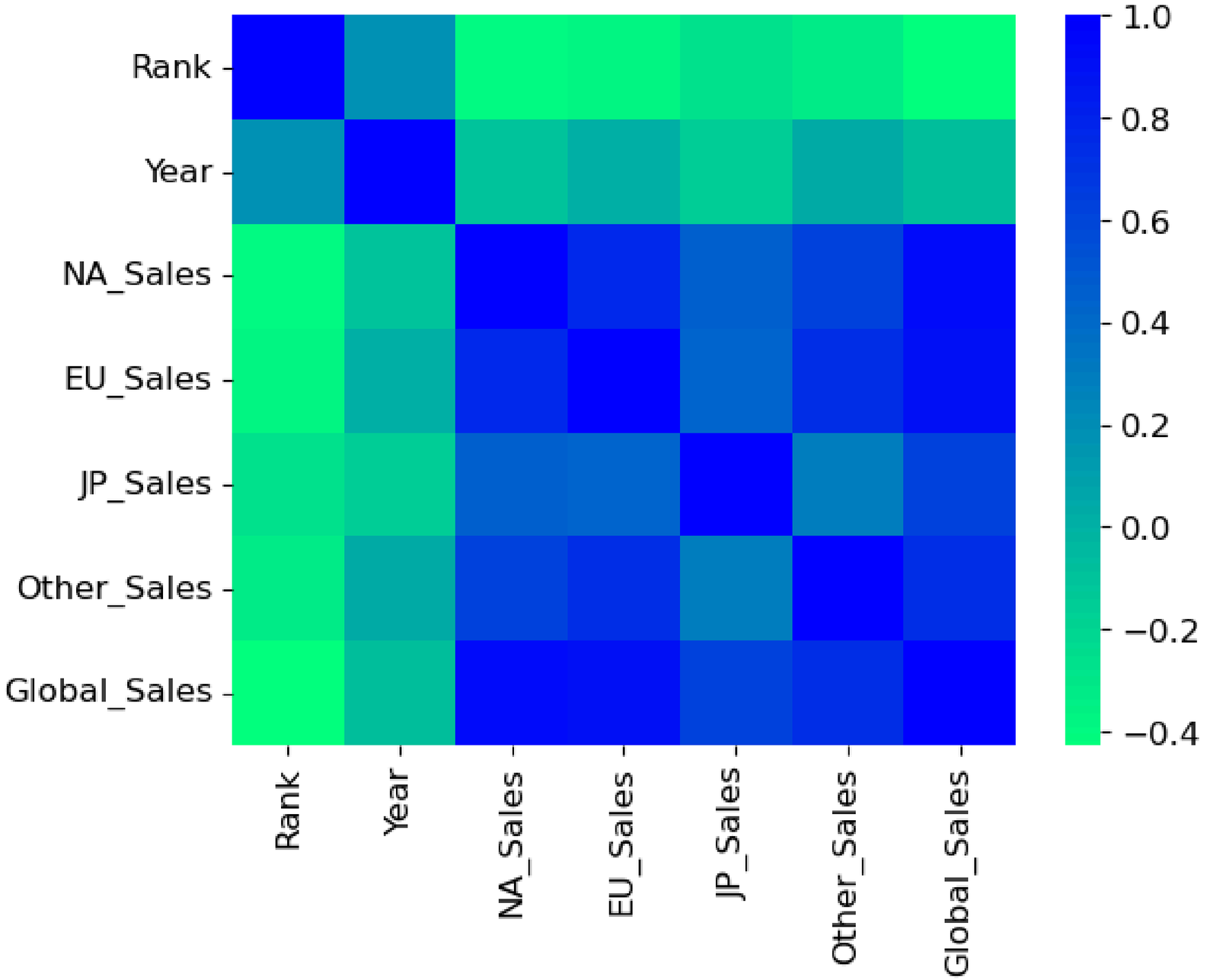
Results

Independent variables : Rank , NA_Sales , EU_Sales, JP_Sales, Other_Sales.
Dependent variable : Global_Sales

RMSE : 0.005342241176650359

Model score : 0.9999933287153024




Rank
Year
NA_Sales
EU_Sales
JP_Sales
Other_Sales
Global_Sales
Rank
Year
NA_Sales
EU_Sales
JP_Sales
Other_Sales
Global_Sales
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Problem Statement

Build a model to predict the global sales of video games based on the provided features.

Approach

USING KNN

1) compute the distance of new samples with the original ones.

2) choose the K past examples that are closest to the new example.

3) calculate the average of K nearest neighbors.

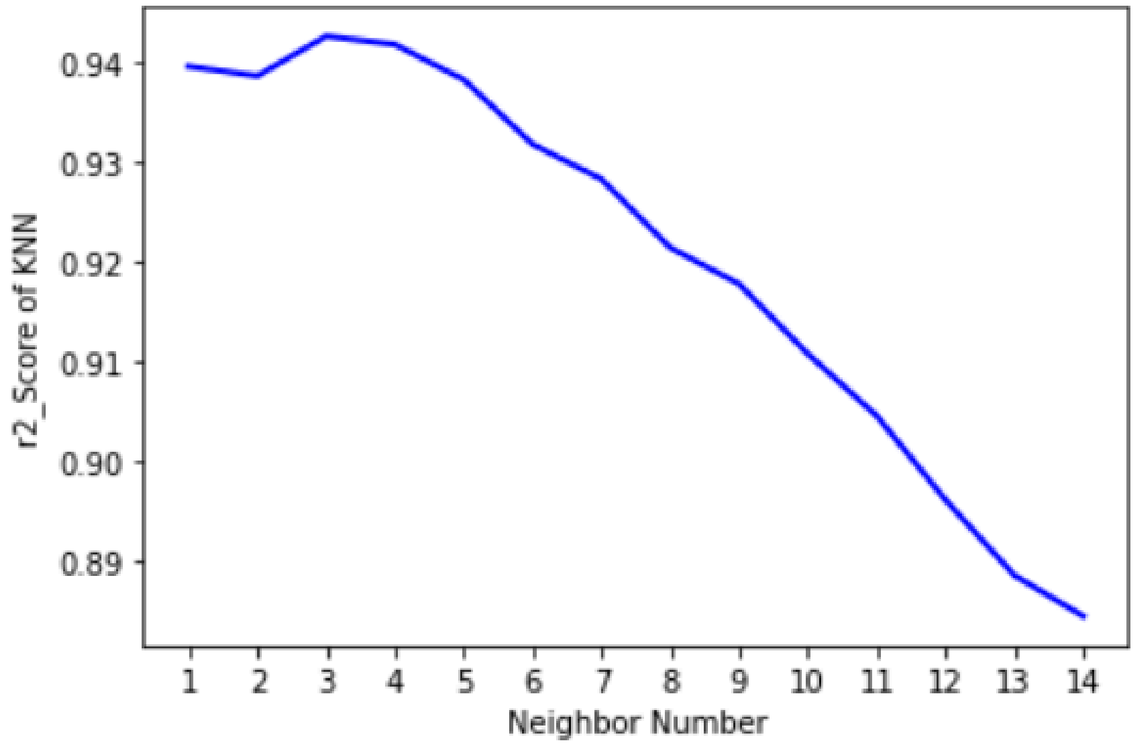
The overall accuracy of this model is about 85%, in a research paper which is relatively high and can be used to generate an accurate prediction.

- **Compute the distance of new samples with the original ones.**

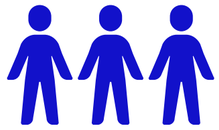
- **Each point on the plot corresponds to the R^2 score of the KNN regressor for a specific value of k.**

- **The overall r2_score of this model is about 0. 93.**

- **This signifies that your KNN model is effectively capturing 93% of the relationship between the features and the target variable. This is a strong performance, suggesting the model can make accurate predictions on unseen data (X_test).**



Problem Statement



Classify video game sales (low, medium, high) based on their features (genre, platform, publisher, year) considering potential class imbalance in the sales data.

Approach

USING Random Forest Classifier

- **Data Preprocessing and Feature Engineering**
- Creates a new categorical feature "Sales_category" based on the sales amount in the "Global_Sales" column.
- Drop the "Global_Sales" column as it's no longer needed after creating the category.
- It uses one-hot encoding to convert categorical features (Genre, Platform, Publisher) into numerical features for the machine learning model.
- Combines the encoded features with the remaining numerical features (Year, Sales_category) into a single data frame.

- **Data Balancing:**
- Analyze the class distribution in the "Sales_category" using Counter to identify potential imbalances.
- Applies RandomOverSampler to address class imbalance by creating synthetic samples for the minority class(es).
- This helps ensure the model doesn't get biased towards the majority class.

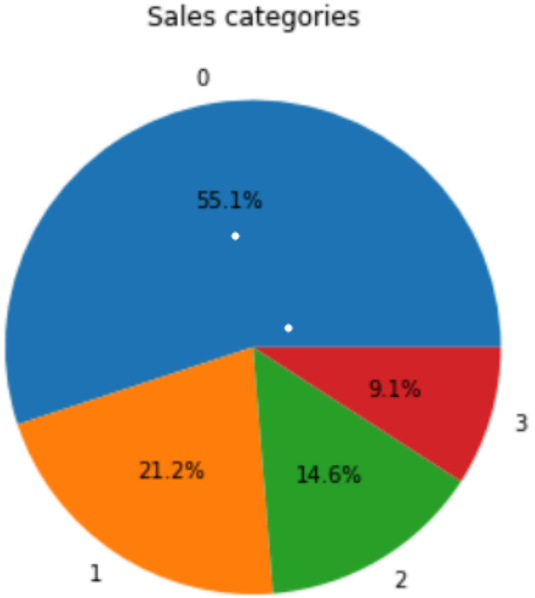
- **Model Building and Evaluation:**
- Splits the preprocessed and balanced data into training and testing sets.
- Trains a Random Forest Classifier model on the training data. Random Forest is a good choice for imbalanced datasets.
- Evaluate the model's performance on the unseen testing data using an accuracy score.

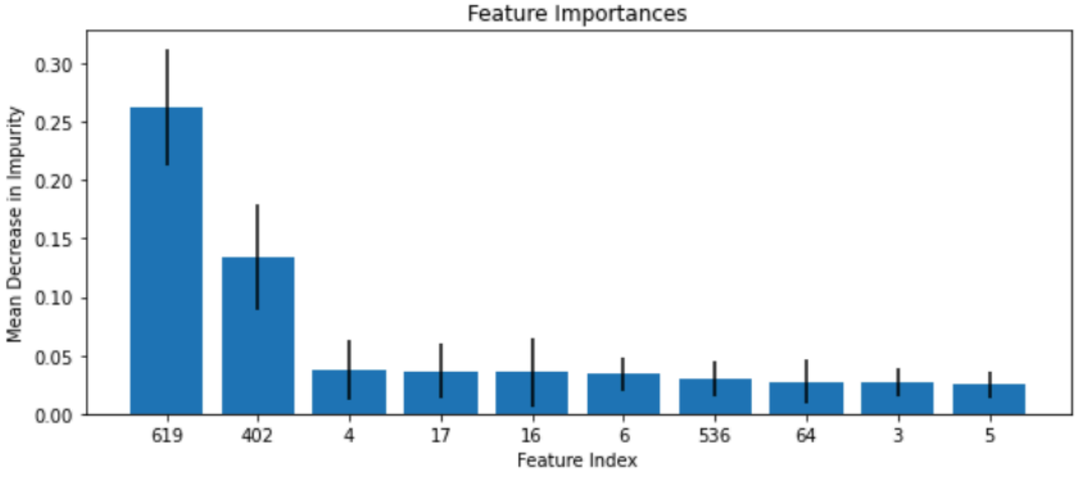
```
vgClassification.head()
```

	Genre_Action	Genre_Adventure	Genre_Fighting	Genre_Misc	Genre_Platform	Genre_Puzzle	Genre_Racing	Genre_Role-Playing
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	0
3	0	0	0	1	0	0	0	0
4	0	0	1	0	0	0	0	0

click to scroll output; double click to hide

5 rows × 621 columns





K-Mean Clustering

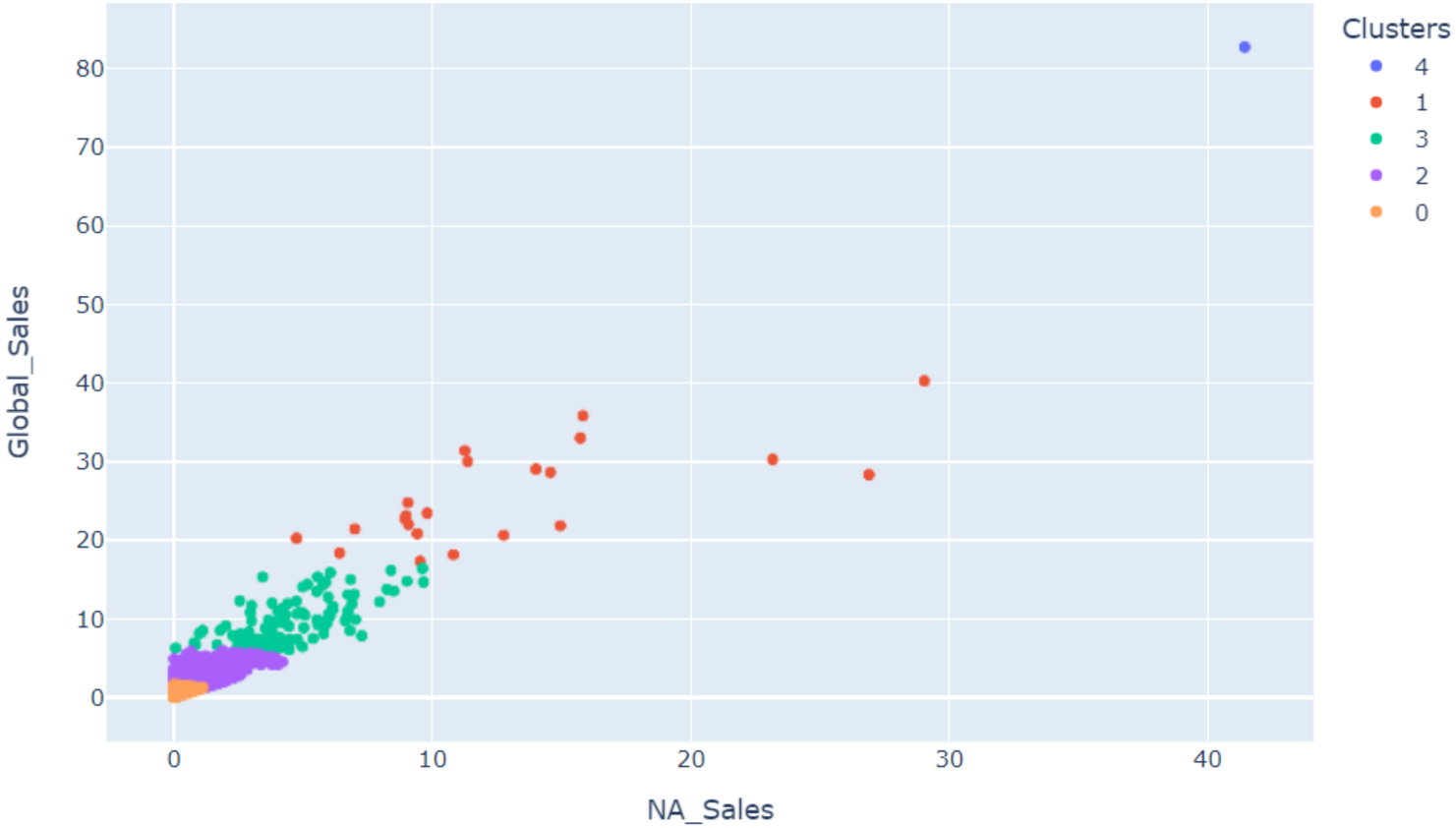
```
KMeans
KMeans(n_clusters=5)
```

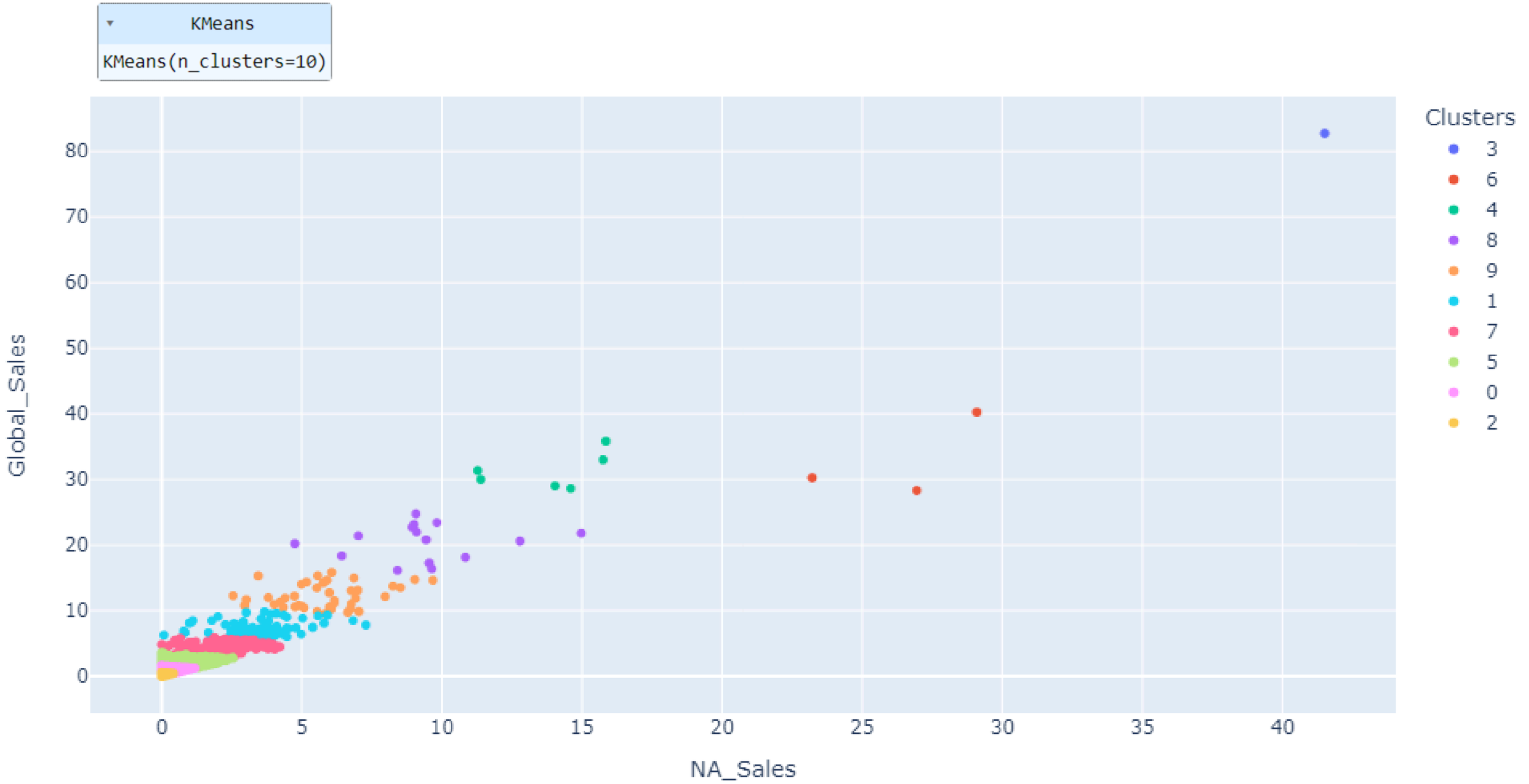
	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Clusters
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74	4
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	2
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82	2
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00	2
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	2

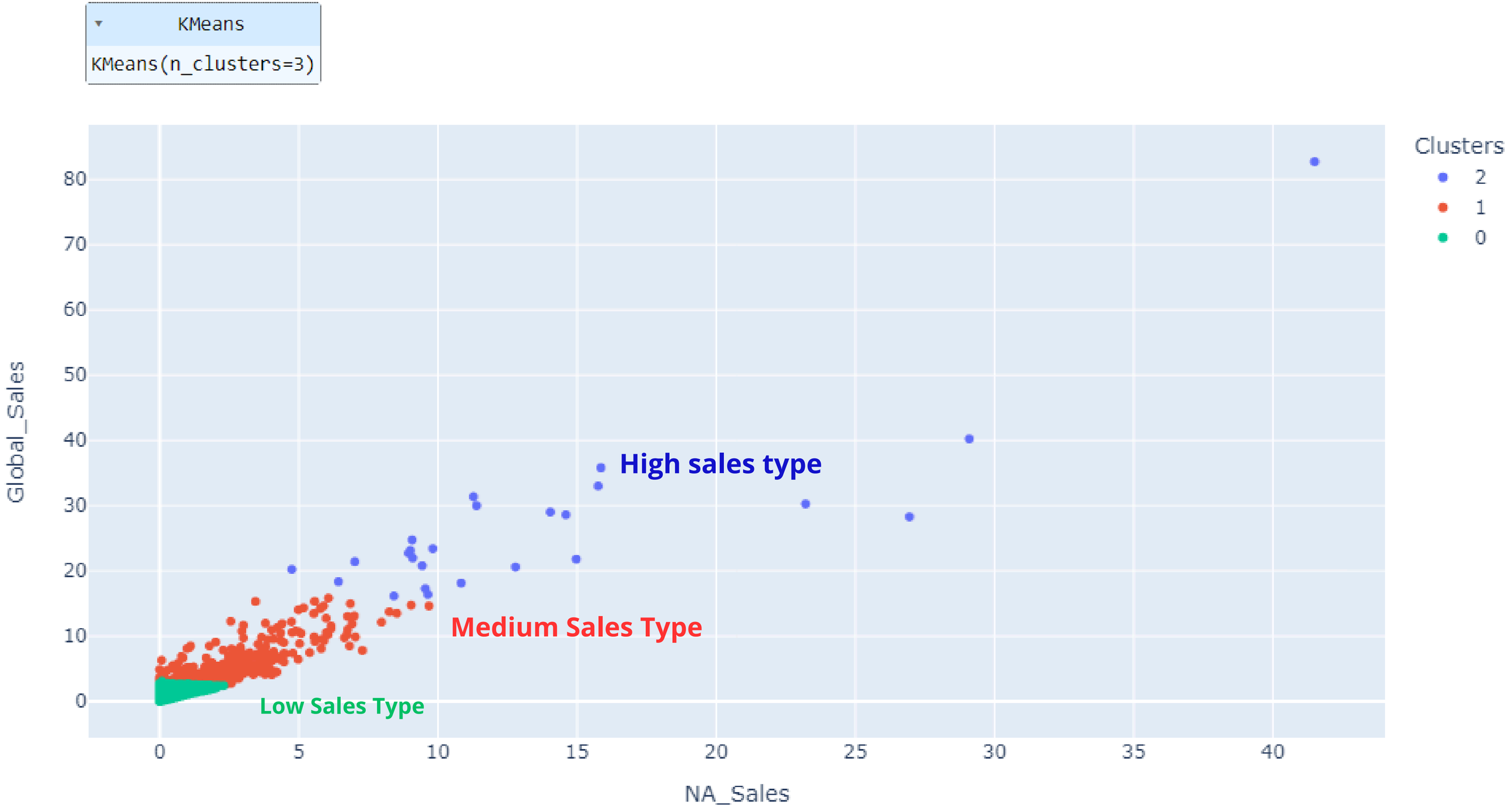
Problem Statement: Identifying groups or clusters of games with similar sales patterns.

Identify distinct segments or clusters of video games based on their sales patterns across different regions (NA, EU, JP, Other). The objective is to group games together that exhibit similar sales behavior, regardless of their genre, platform, or publisher.

Target Feature:
Sales data (Global_Sales, NA_Sales, EU_Sales, JP_Sales, Other_Sales)




KMeans
KMeans(n_clusters=10)
Clusters
3
6
4
8
9
1
7
5
0
2
Global_Sales
NA_Sales
0
10
20
30
40
50
60
70
80
0
5
10
15
20
25
30
35
40


KMeans
KMeans(n_clusters=3)
Clusters
2
1
0
High sales type
Medium Sales Type
Low Sales Type
Global_Sales
NA_Sales
0
10
20
30
40
50
60
70
80
0
5
10
15
20
25
30
35
40

Using the K-Means algorithm, the data is divided into 3 categories: low, medium, and high.

Low Sales Type

5 most frequent Platform in the low category is DS, PS2, Wii, PS3, and PS5

5 Most frequent Genre in the low category are: Action, Sports, Misc, Role-Playing, Adventure.

5 most frequent Publisher in the low category are: Electronic Arts, Namco Bandai Games, Activision, Ubisoft, Konami Digital Entertainment.

High sales type

5 most frequent Platform in the high category are Wii, DS, GB, NES, PS2.

5 Most frequent Genres in the high category are: Platform, Action, Sports, Role-Playing, Misc.

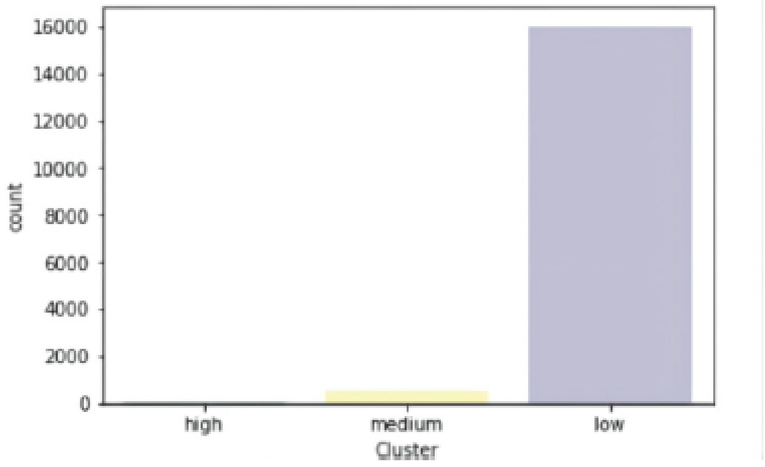
5 Most frequent Publishers in the high category are Nintendo, Take-Two Interactive, and Microsoft Game Studios.

Medium Sales Type

5 most frequent Platform in the medium category are PS2, X360,PS3, PS, DS.

5 most frequent Genres in the medium category are: Action, Shooter, Sports, Role-Playing, Platform.

5 most frequent Publisher in the medium category are: Nintendo, Electronic Arts, Sony Computer Entertainment, Activision, Ubisoft.



Conclusion

- Action and Sports categories are the most sold game genres across the world.
- Nintendo, PS2, DS are the most frequently used gaming platforms
- Most frequent publisher are Electronic Arts, Nintendo, Sony Computer Entertainment

Ridge Regression

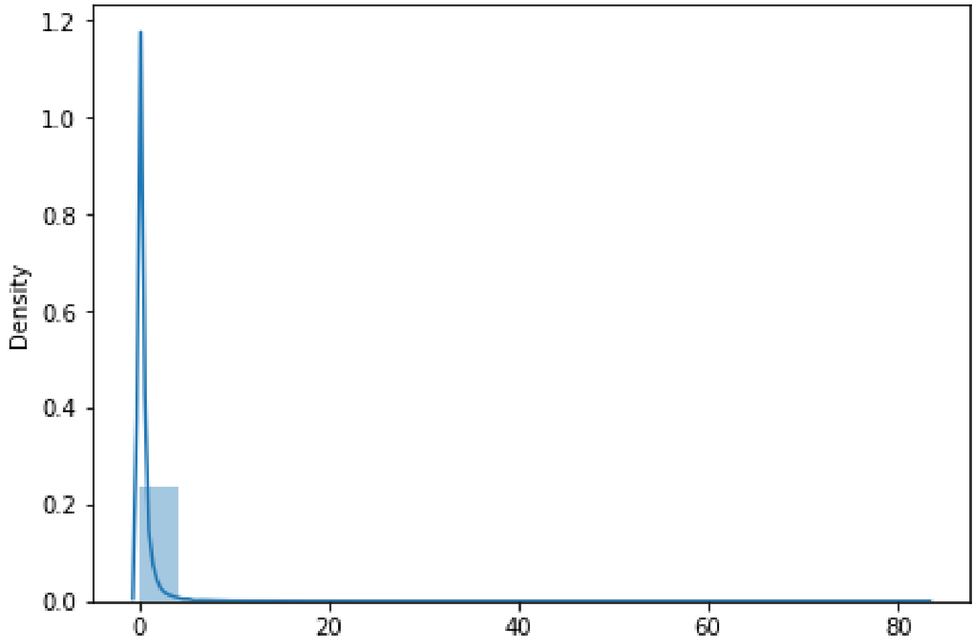
Problem Statement:

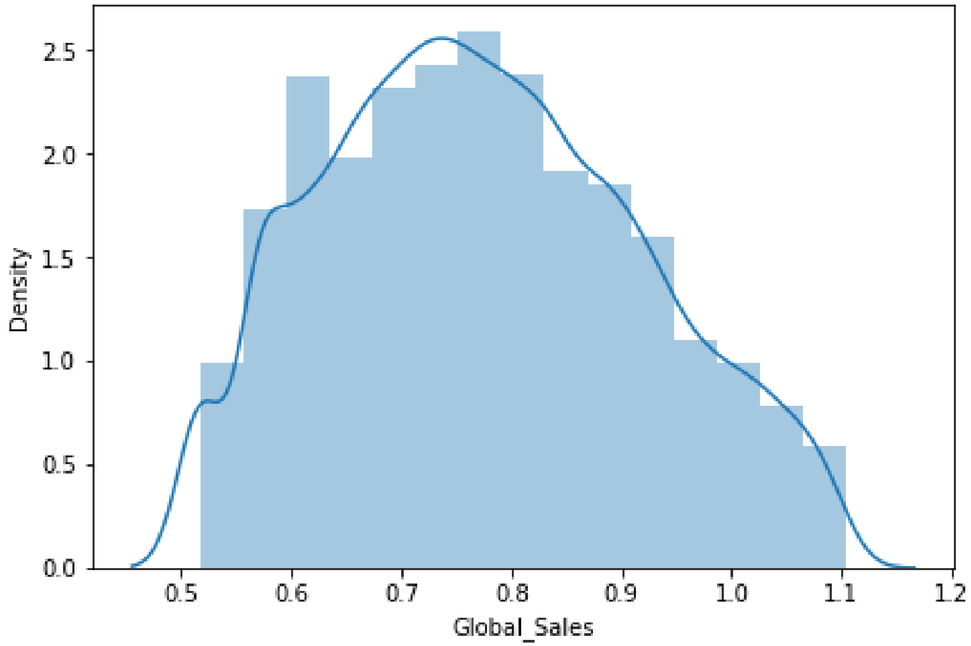
Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L^2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large. This results in predicted values being far away from the actual values.
The cost function for ridge regression is

$$\min(||Y - Xw||^2 + \alpha||w||^2)$$

where Y is the target (Global Sales), X is the feature matrix, α is the penalty term and w is a vector of weights to be computed. The higher the values of
α, the bigger the penalty is. Therefore the magnitude of coefficients w is reduced.

We now inspect the distribution of the (Global Sales) target Y. We see that the distribution is right-skewed. We decide to take only the left tail of the distribution, which contains most of the entries. Moreover, we apply a transformation to the data in order to eventually deal with a normal distribution.





Ridge Regression Output

```
MSE:  0.013692314018892262
r2 score:  0.3386062616950603 6
```

Tuning Ridge hyperparameter (grid search)

```
MSE:  0.013667085230912526
r2 score:  0.3398249134417117 3
```

Conclusion

The GridSearch has not helped obtain a bigger r2 score, which is still very small. The main issue concerns the big variance in the predictions for the least profitable titles.
We also used logistic regression in this dataset but that also failed(accuracy 21%).
Several factors impact the Global Sales of well-known video games, which are not stated in our dataset.

- **BUDGET FOR PRODUCTION**
- **ADVERTISING COST**
- **COMPETITION WITH OTHER MORE POPULAR VIDEO GAMES**
- **COST OF THE TITLE IN THE MARKETPLACE**

Problem Statement

1.Problem Statement 1: Sales Prediction Given historical video game sales data, including a feature likely representing time (e.g., Year), develop a model to accurately predict future sales figures for various regions (North America, Europe, Japan, etc.) and globally.

2.Problem Statement 2: Sales Segmentation Analyze historical video game sales data for various regions (North America, Europe, Japan, etc.) and global sales to identify groups (clusters) of games with similar sales patterns. This can be used for market segmentation or understanding regional sales trends.

Approach 1

Predicting Sales with Random Forest Regression

1. Data Split: Divide data into training and testing sets.
2. Pipeline Creation: Combine data preprocessing with the Random Forest model into a single pipeline for efficient processing.
3. Hyperparameter Tuning: Use RandomizedSearchCV to optimize hyperparameters (model settings) of the Random Forest model through random sampling and cross-validation.
4. Model Training: Train the best model (pipeline) using the training data.
5. Model Evaluation: Evaluate the trained model's performance using Mean Squared Error (MSE) for overall prediction accuracy and individual accuracy scores for each sales region.

OUTCOME

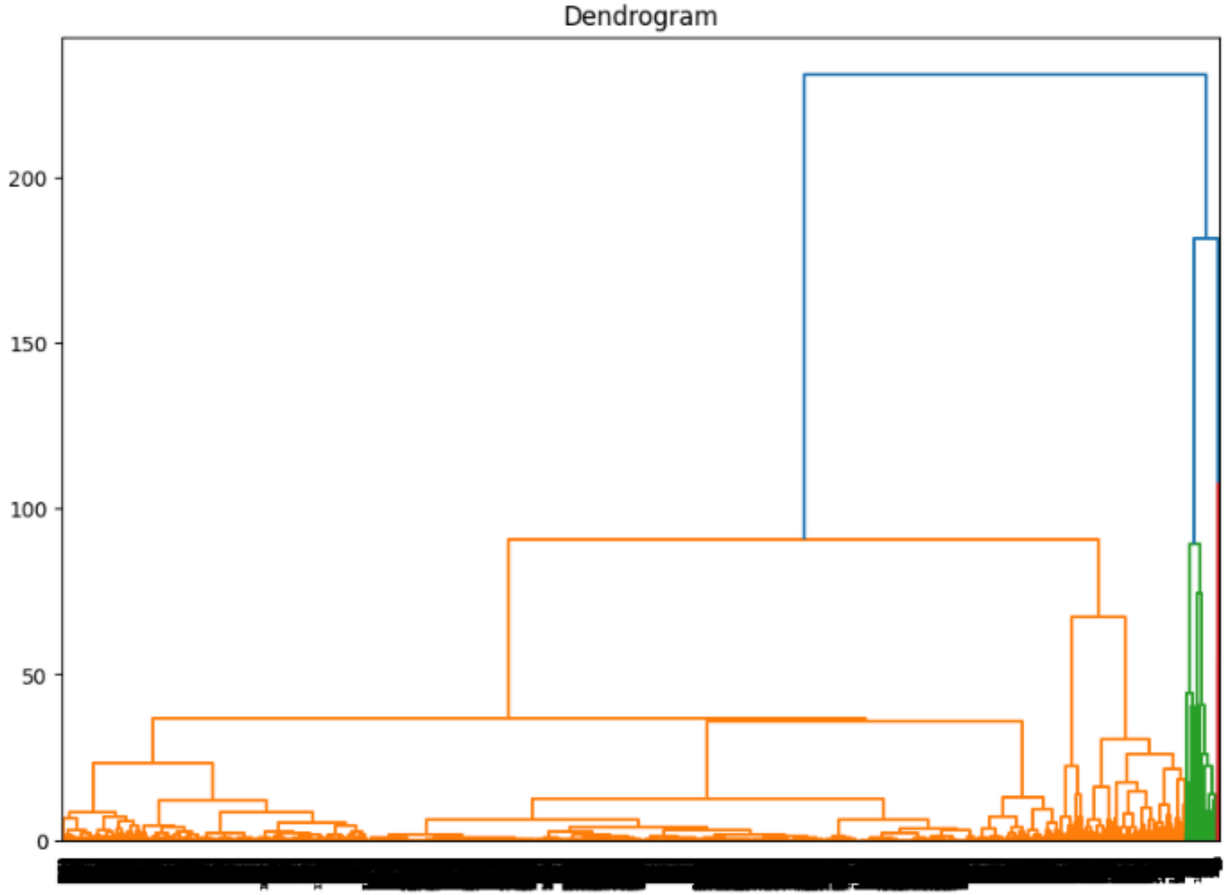
1. Mean Squared Error: 57.65717483673812 %
2. Accuracy for NA_Sales: 87.66626360338573 %
3. Accuracy for EU_Sales: 93.42200725513906 %
4. Accuracy for JP_Sales: 96.46916565900845 %
5. Accuracy for Other_Sales: 98.42805320435308 %
6. Accuracy for Global_Sales: 45.22370012091898 4 %
7. Overall Accuracy: 84.24183796856106 %

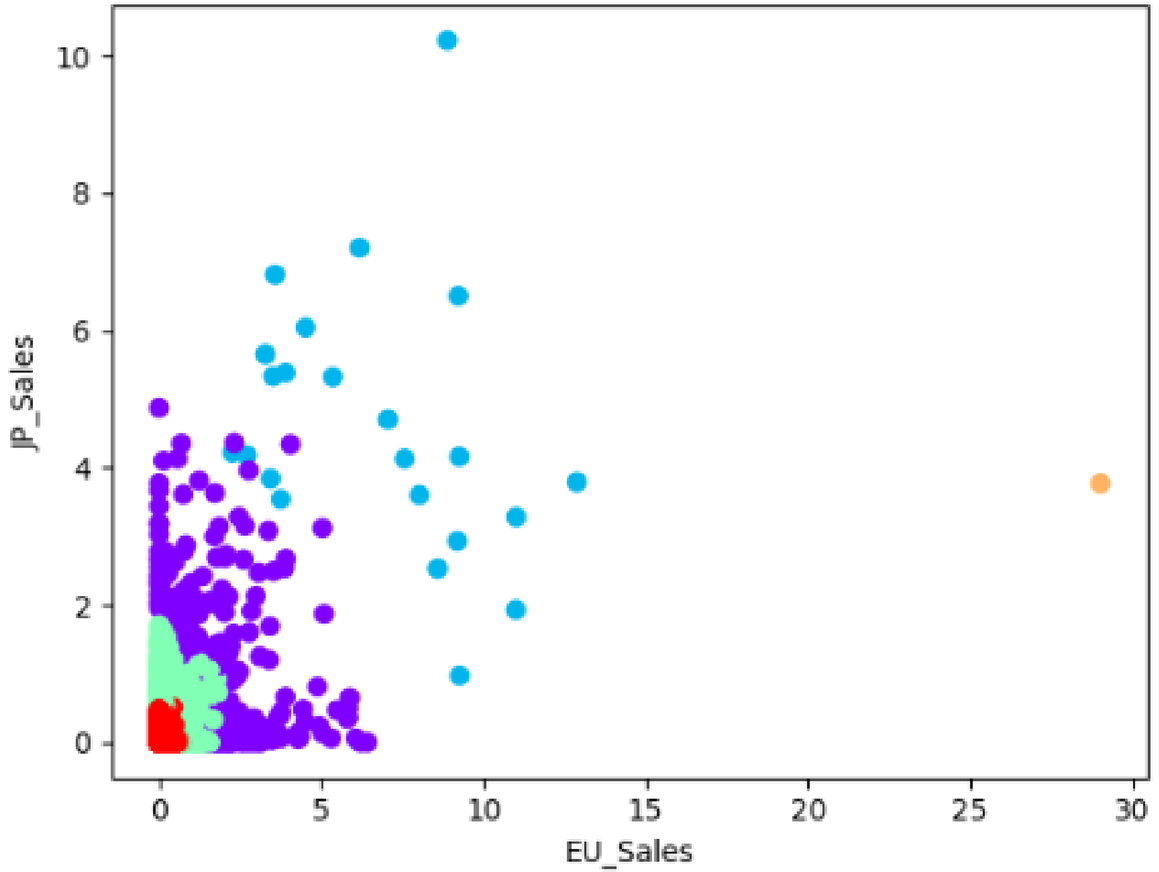
Approach 2

Sales Pattern Discovery

1. **Extract Sales Data: Focuses on sales figures for various regions (EU, JP, etc.) and global sales.**
2. **Standardize Features: Ensures all sales figures are on a similar scale for better clustering.**
3. **Group Similar Sales: Identifies clusters of games with comparable sales patterns across regions using hierarchical clustering.**
4. **Visualize Clusters: Creates a dendrogram to show cluster relationships and a scatter plot to see how regional sales group together.**

OUTCOME


Dendrogram
200
150
100
50
0


JP_Sales
EU_Sales
0
2
4
6
8
10
0
5
10
15
20
25
30

THANK YOU