COMPONENT ONE

PREDICTING VIDEO GAME SALES PERFORMANCE USING SUPERVISED AND UNSUPERVISED LEARNING TECHNIQUES

ABSTRACT

This project explores the application of various supervised and unsupervised learning techniques to predict the sales performance of video games. The Artificial Neural Network (ANN) Model emerged as the best and most accurate model outperforming the other models.

For the clustering analysis Hierarchical Clustering performed better with lower DBI Score and higher silhouette score compared to the K-means Clustering, especially for the 'NA_Sales', 'EU_Sales' combination, which indicates better clustering quality.

1.0 INTRODUCTION

In previous years forecasting of sales have been done traditionally by using human expertise, however with machine learning it is now possible to predict more accurate sales by leveraging sophisticated algorithms (Jain et al., 2023). The video game industry is a competitive market that is characterized by technological advances, and customer preferences. Hence, projections about sales are critical decisions to make strategic decisions. (Affan, Vishwakarma, and Kumari, 2024).

2.0 METHODOLOGY

The emergency vehicles identification dataset containing 16416 rows and 16 columns containing video games sales details, was analysed using different supervised learning techniques, including Polynomial, Linear, Random Forest Regression and ANN Algorithms using Mean Squared Error(MSE),Coefficient of determination(R^2) ,Root mean squared error (RMSE), Mean absolute error (MAE), unsupervised learning techniques like K-Means, Hierarchical Clustering using Silhouette coefficient and Davis Bouldin Index(DBI) as the evaluation metrics were executed using the Jupyter notebook environment.

3.0 ANALYSIS AND RESULTS

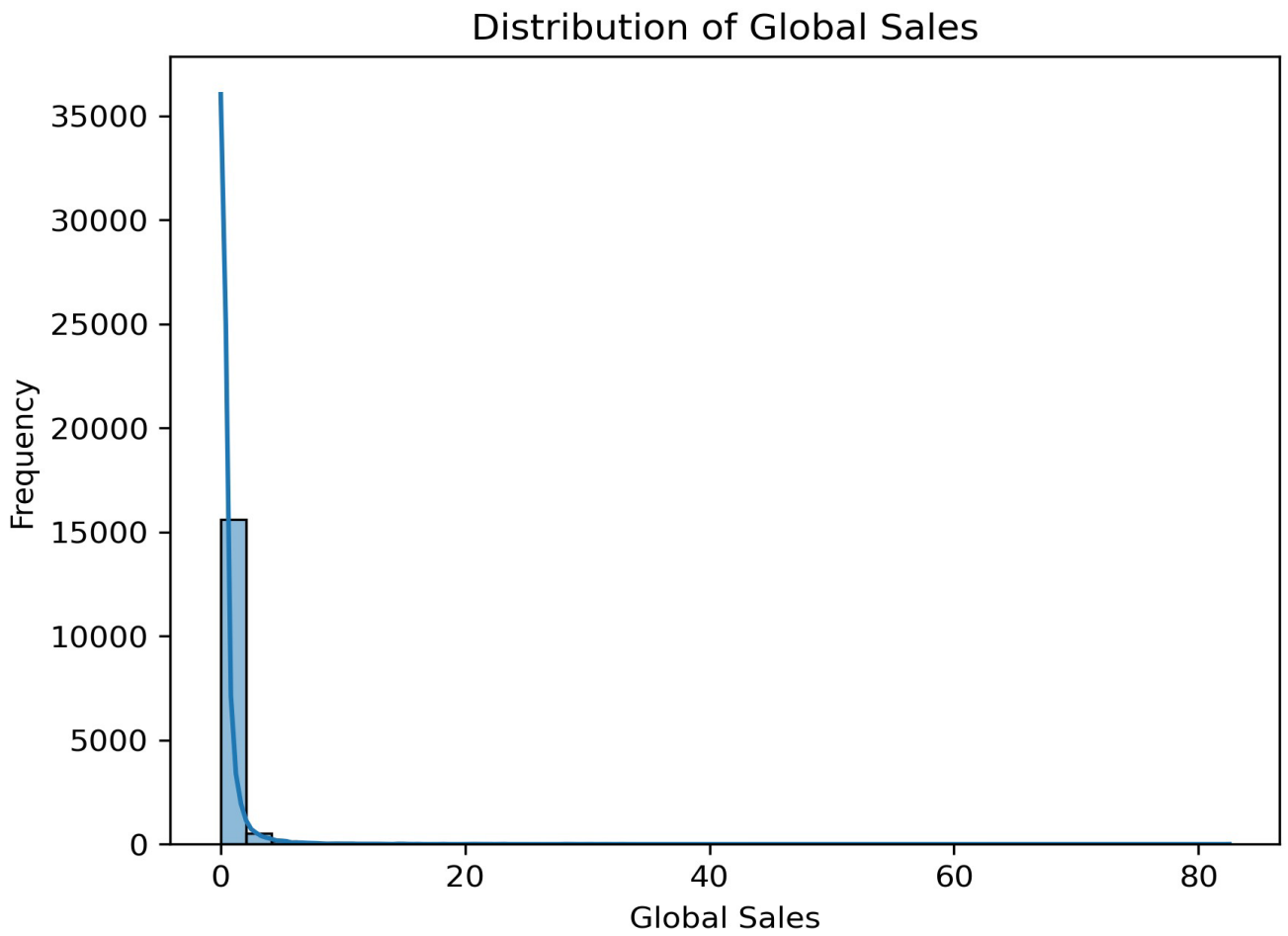

Figure 1: Distribution of Global Sales

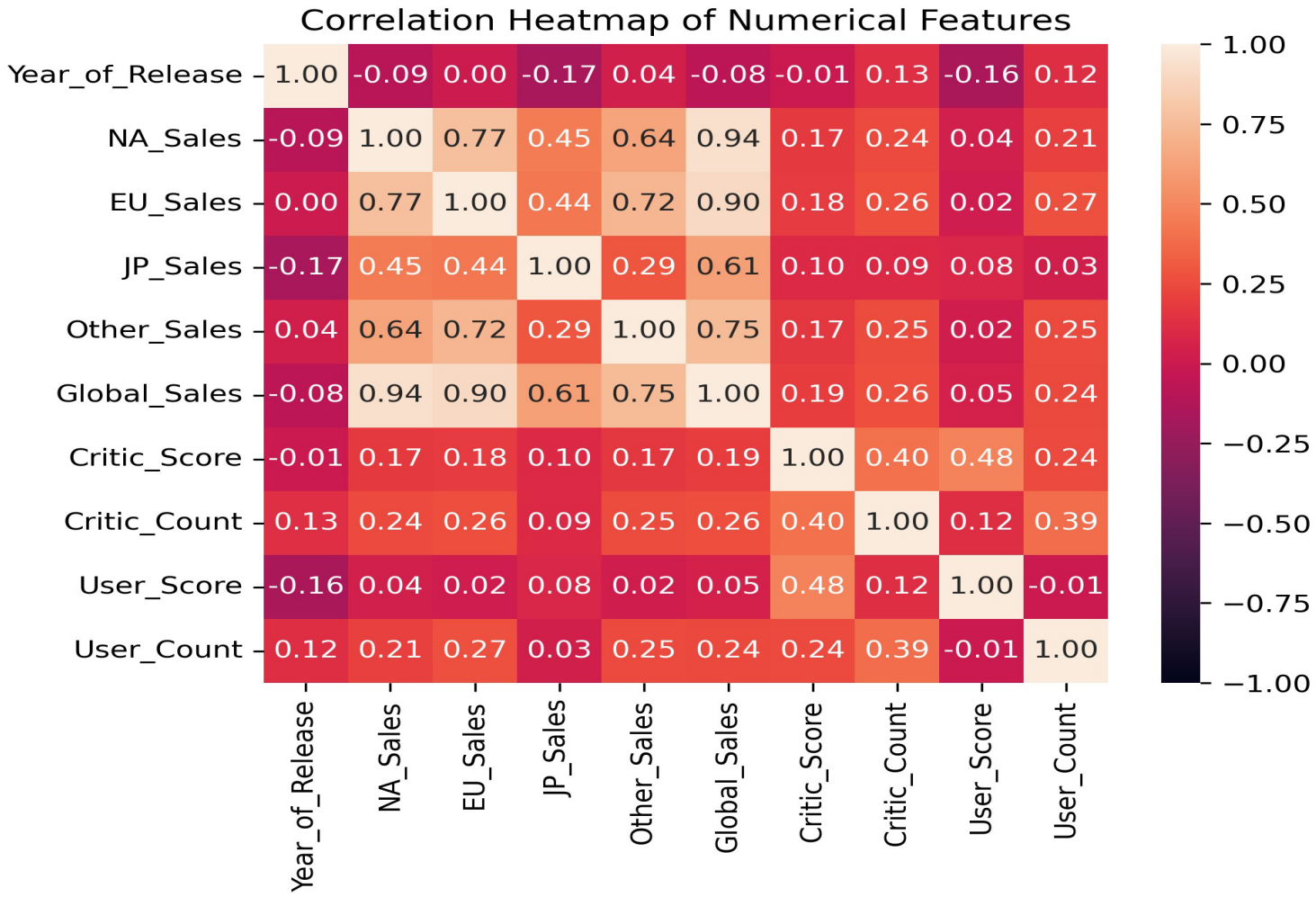

Figure 2: Numerical Features Correlation Map

Figure 1 reveals that majority of video games have very low global sales. In **Figure 2** the correlation plot shows significant numerical features influencing global sales, with a strong positive correlation between NA_Sales and Global_Sales (0.77), weak positive between Critic_Count and Global_Sales (0.12).

3.1 LINEAR vs POLYNOMIAL REGRESSION-SINGLE NUMERICAL FEATURES Table 1: NA_Sales Evaluation Metrics

Linear	**Polynomial(Degree = 2)**
R^2: 0. 9290	R^2: 0.8332
MSE: 0.3005	MSE: 0.7066
RMSE: 0.5482	RMSE: 0.8406
MAE: 0.2023	MAE: 0.2105

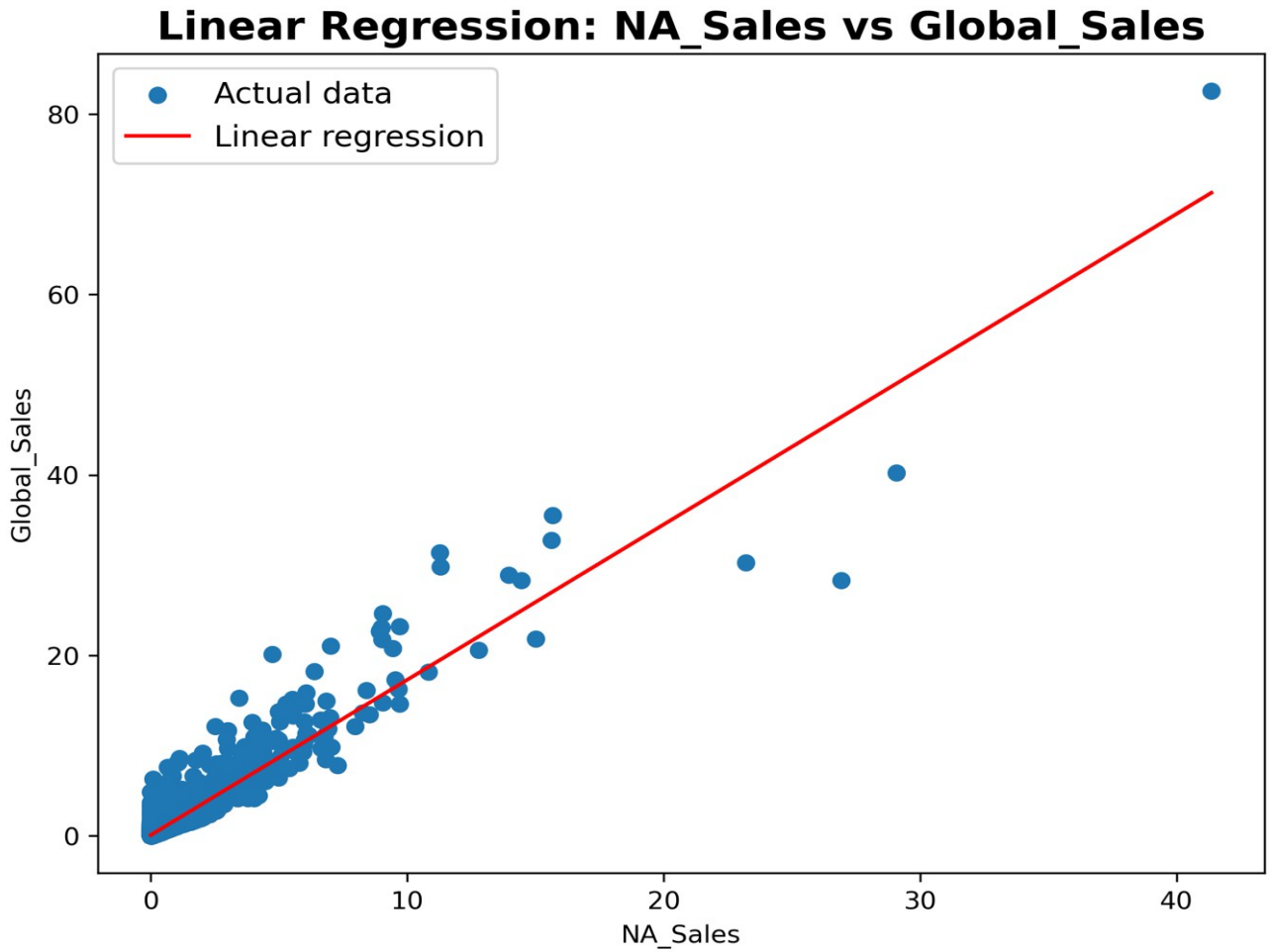

Figure 3: Linear Regression NA_Sales vs Global_Sales

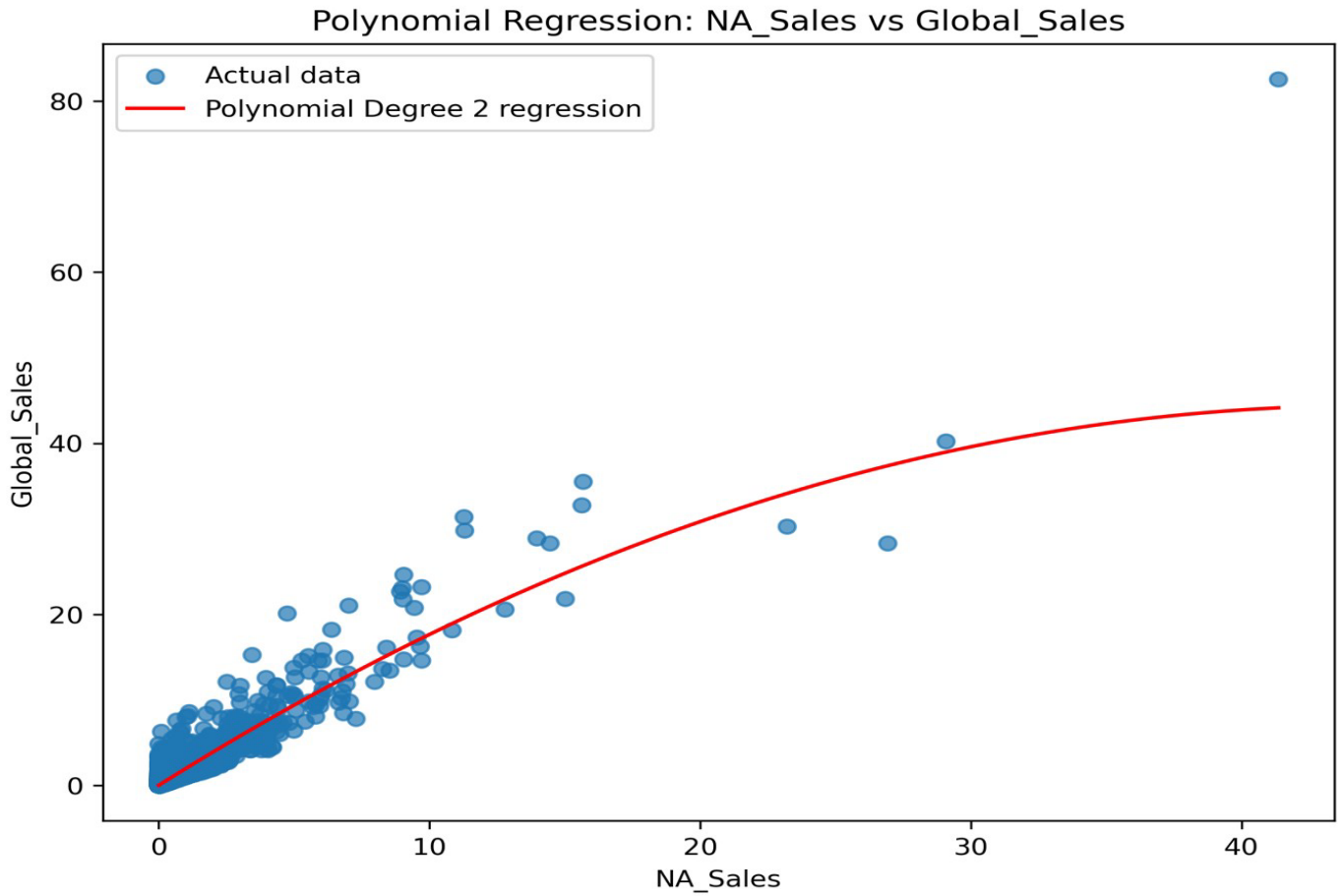

Figure 4: Polynomial Regression NA_Sales vs Global_Sales

In Table 1 the R^2 of the linear regression (0. 9290) is higher than polynomial (0.8332). The MSE of Linear (0.3005) is lower than Polynomial (0.7066).

3.1.2 EU_Sales

Table 2: EU_Sales Evaluation Metrics

Linear Model	**Polynomial Model (Degree = 2)**

R^2: 0.9235	R^2: 0.9210
MSE: 0.3241	MSE: 0.3346
RMSE: 0.5693	RMSE:
MAE: 0.2448	MAE: 0.2456

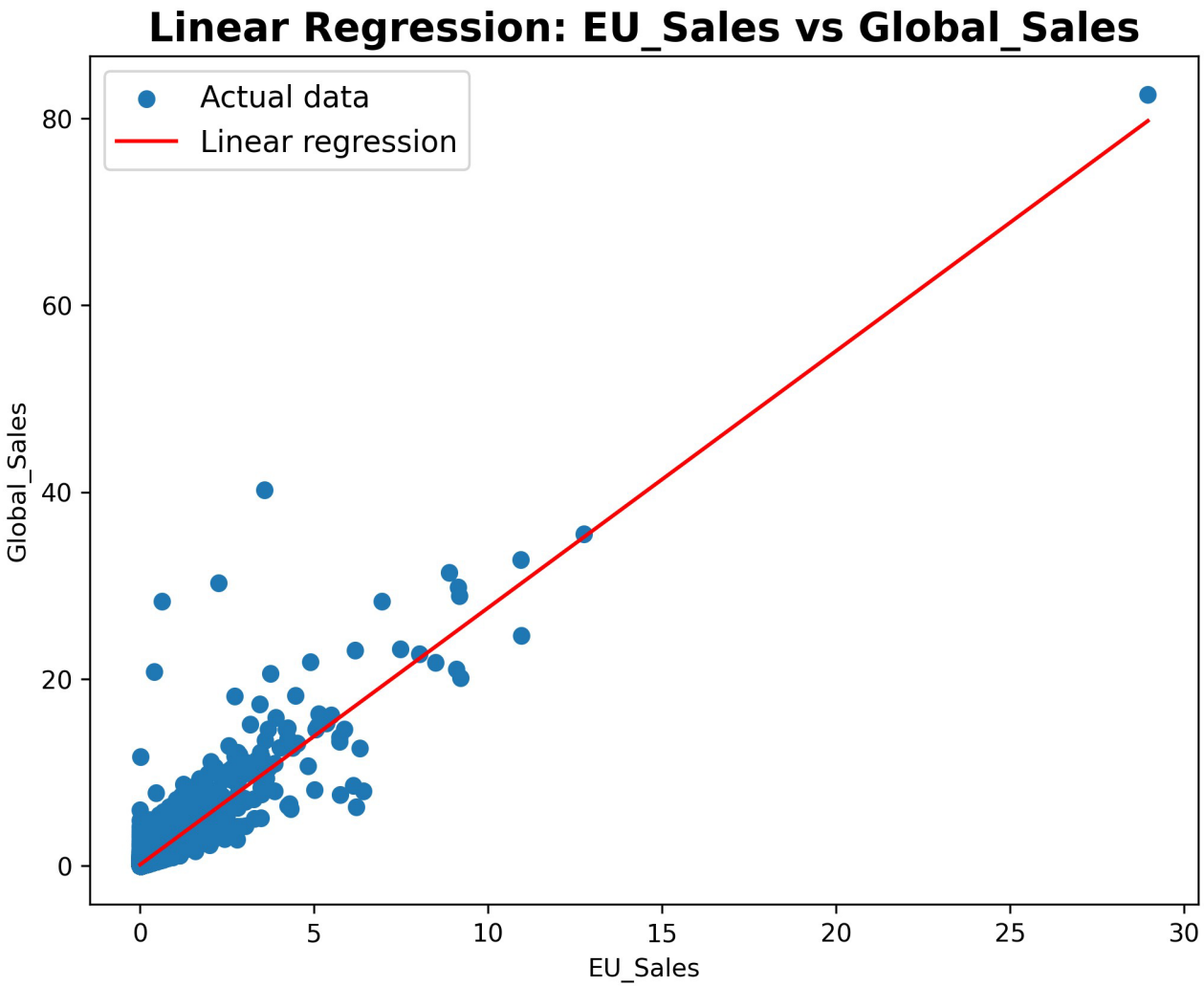

Figure 5: Linear Regression EU_Sales vs Global_Sales

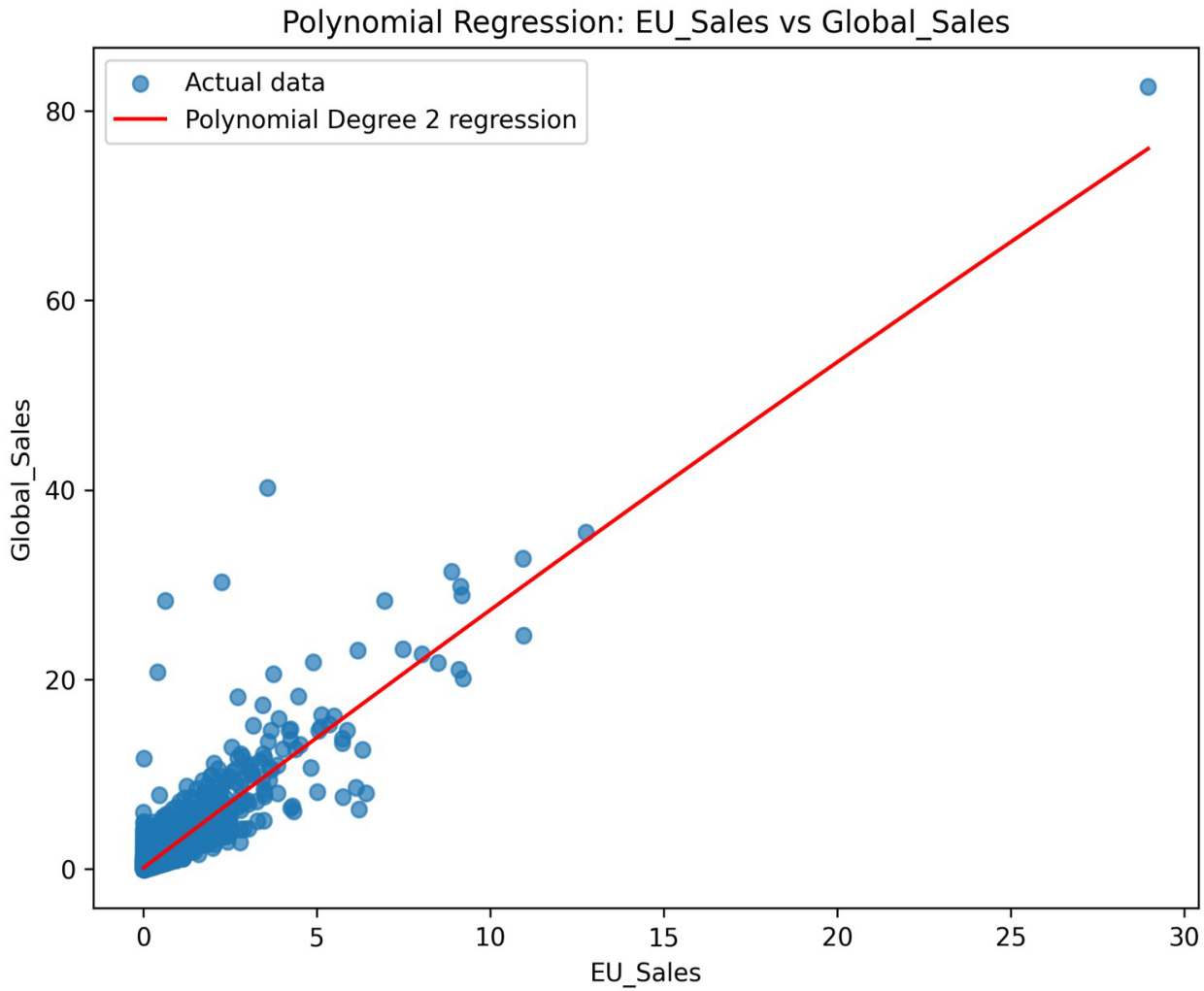

Figure 6: Polynomial Regression NA_Sales vs Global_Sales

In Table 1 the R^2 of the linear regression (0.9235) is the best fit with higher than polynomial (0.9210). The MSE of Linear (0.3241) is lower than that of Polynomial (0.3346). Figure 5 & 6 further demonstrates this.

3.1.3 JP_Sales

Table 3: JP_Sales Evaluation Metrics

Linear Model	**Polynomial Model (Degree = 2)**
R^2: 0.3177	R^2: 0.3226
MSE: 0.3241	MSE: 0.3346
RMSE: 0.5693	RMSE: 0.5784
MAE: 0.2448	MAE: 0.2456

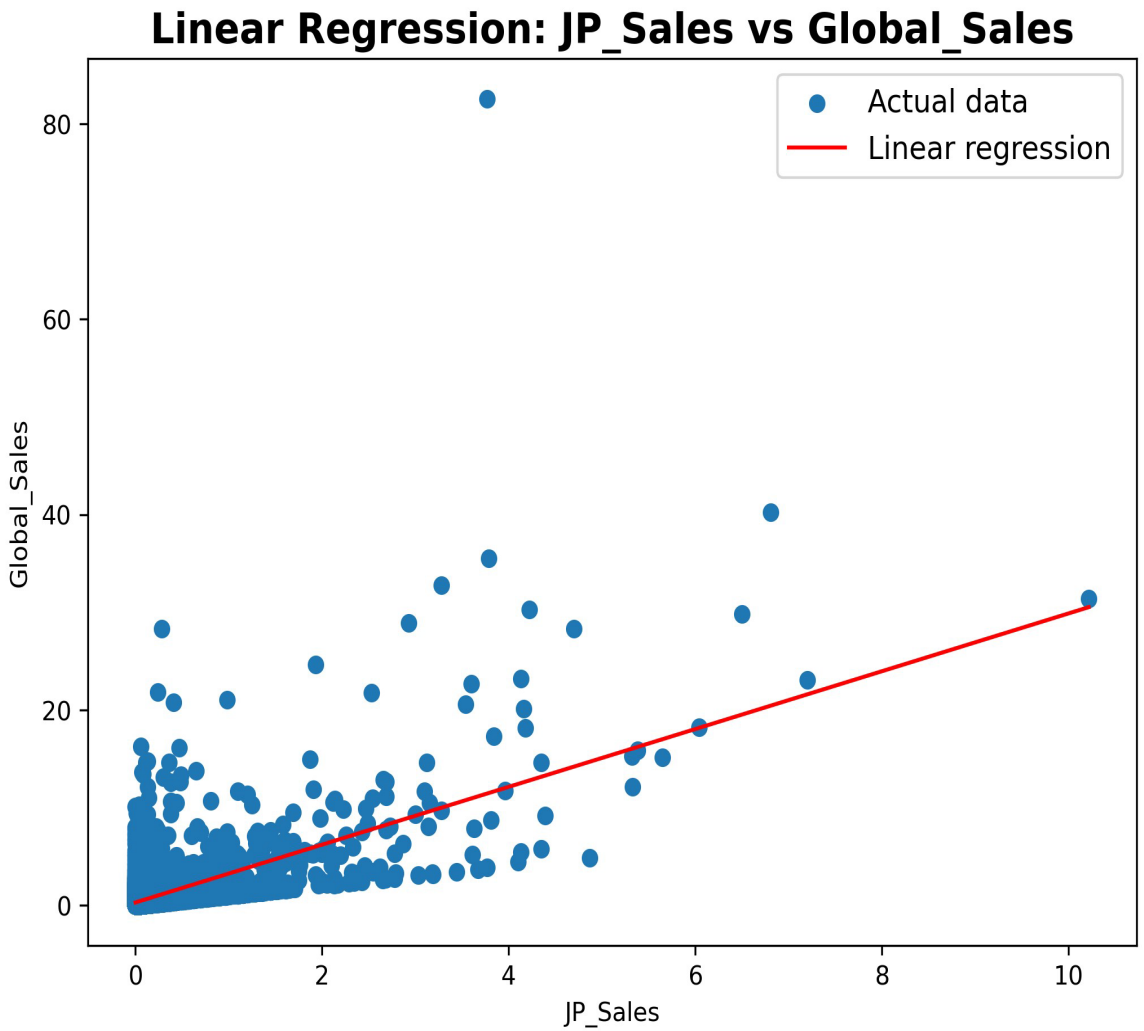

Figure 7: Linear Regression JP_Sales vs Global_Sales

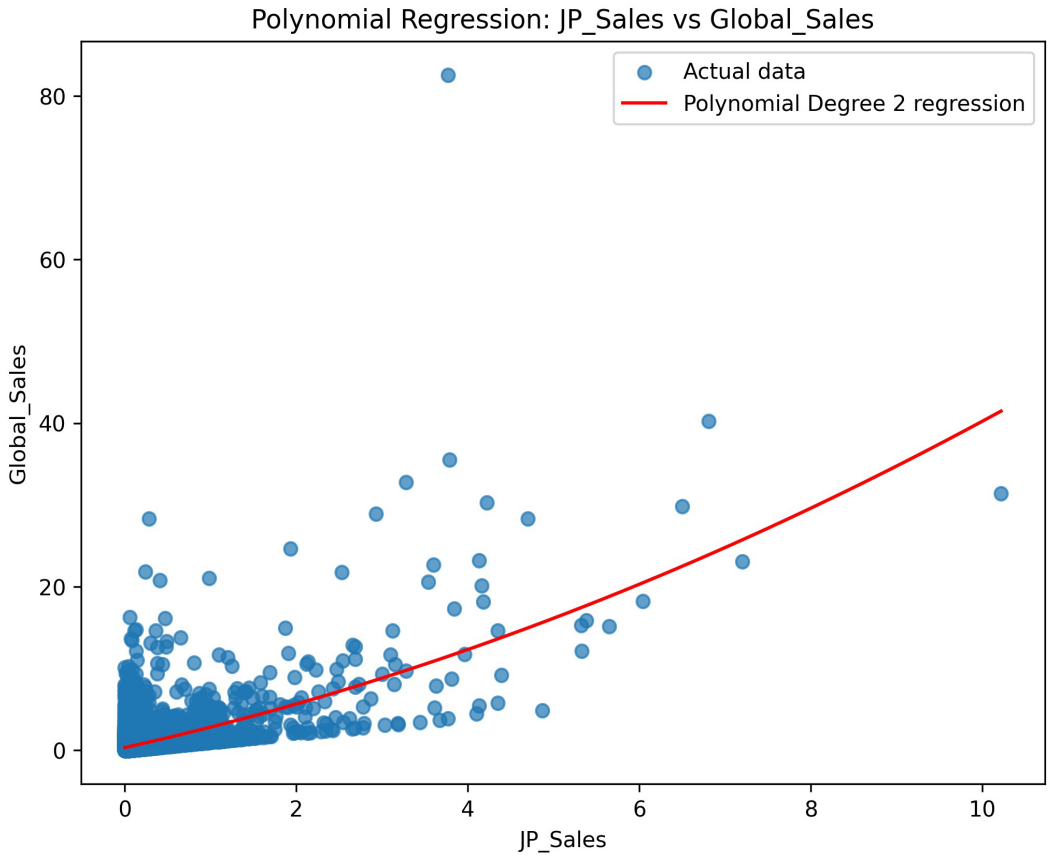

Figure 8: Polynomial Regression JP_Sales vs Global_Sales

The polynomial regression demonstrates a stronger predictive relationship between "JP_Sales" and Global Sales, with higher R^2 score (0.3226) and lower error metrics than the linear model (Table 2).

3.1.4 Other_Sales

Table 4: Other_Sales Evaluation Metrics

Linear	**Polynomial (Degree = 2)**
R^2: 0.7050	R^2: 0.6004
MSE: 1.2493	MSE: 1.6922
RMSE: 0.5693	RMSE: 1.3008
MAE: 1.1177	MAE: 0.2902

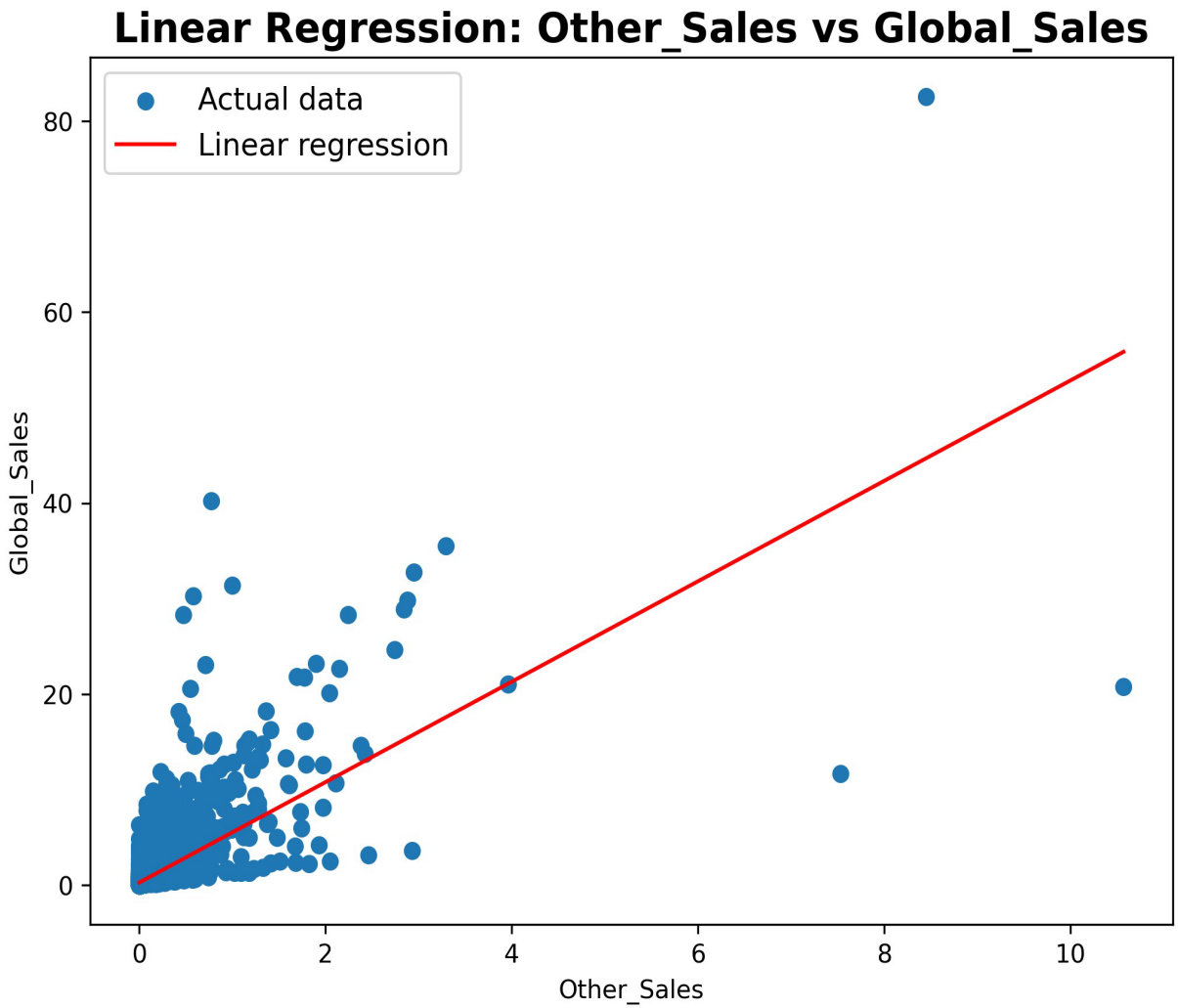

Figure 9: Linear Regression Other_Sales vs Global_Sales

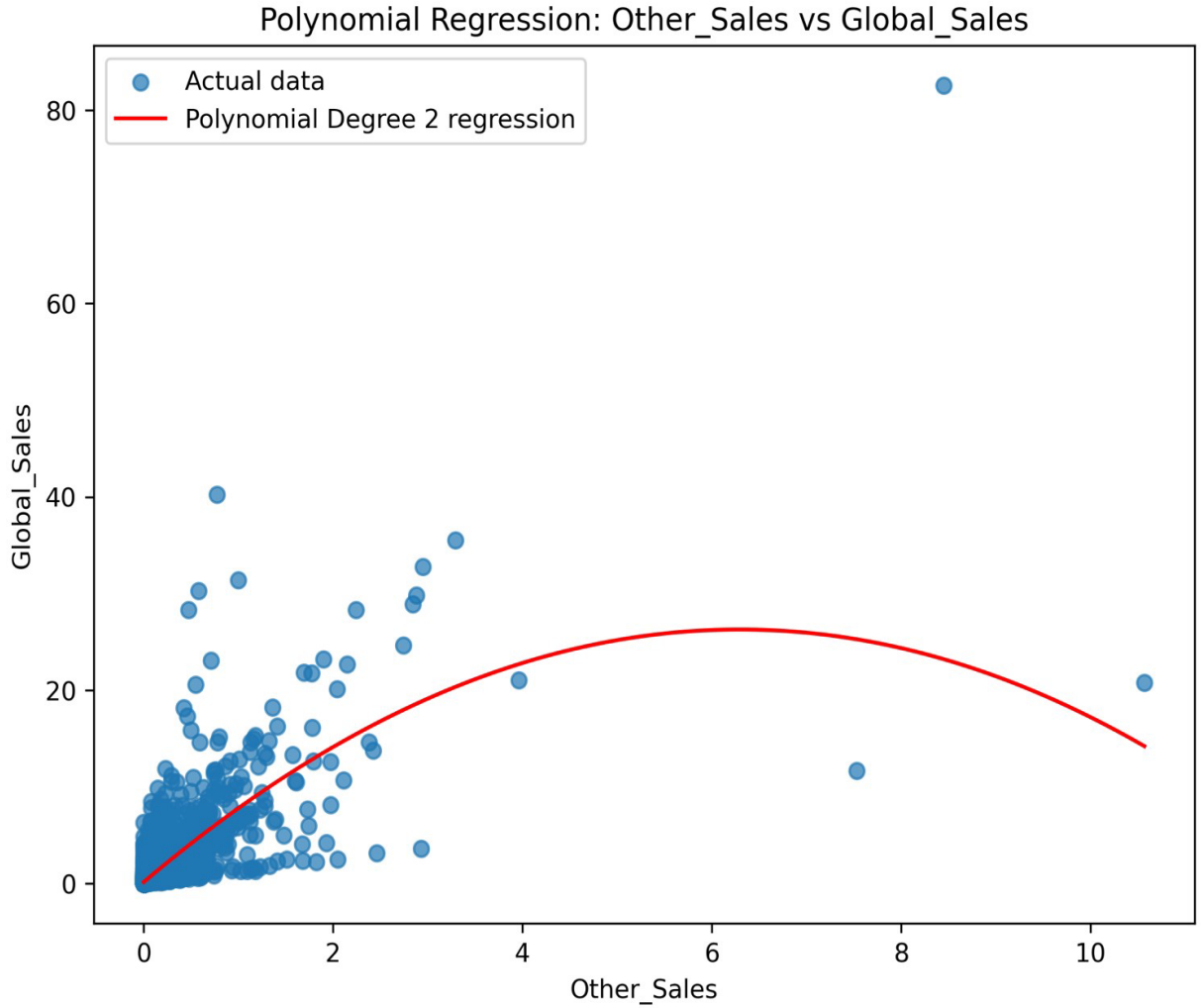

Figure 10: Polynomial Regression Other_Sales vs Global_Sales

In Table 4 the linear regression has higher R^2(0.7050) with and lower MSE compared to polynomial (0.6004).

3.2 REGRESSION MODELS USING MULTIPLE NUMERICAL FEATURES

In table 5 below the linear regression model using multiple input features ('NA_Sales', 'JP_Sales', 'Other_Sales', 'Year_of_Release', 'User_Count', 'Critic_Count') gave an R^2 (0.9759), signifying high positive correlations between these features and Global Sales.

Table 5: Multiple Numerical Features Evaluation Metrics

R^2:	0.9759
MSE:	0.1020
RMSE:	0.3194
MAE:	0.0977

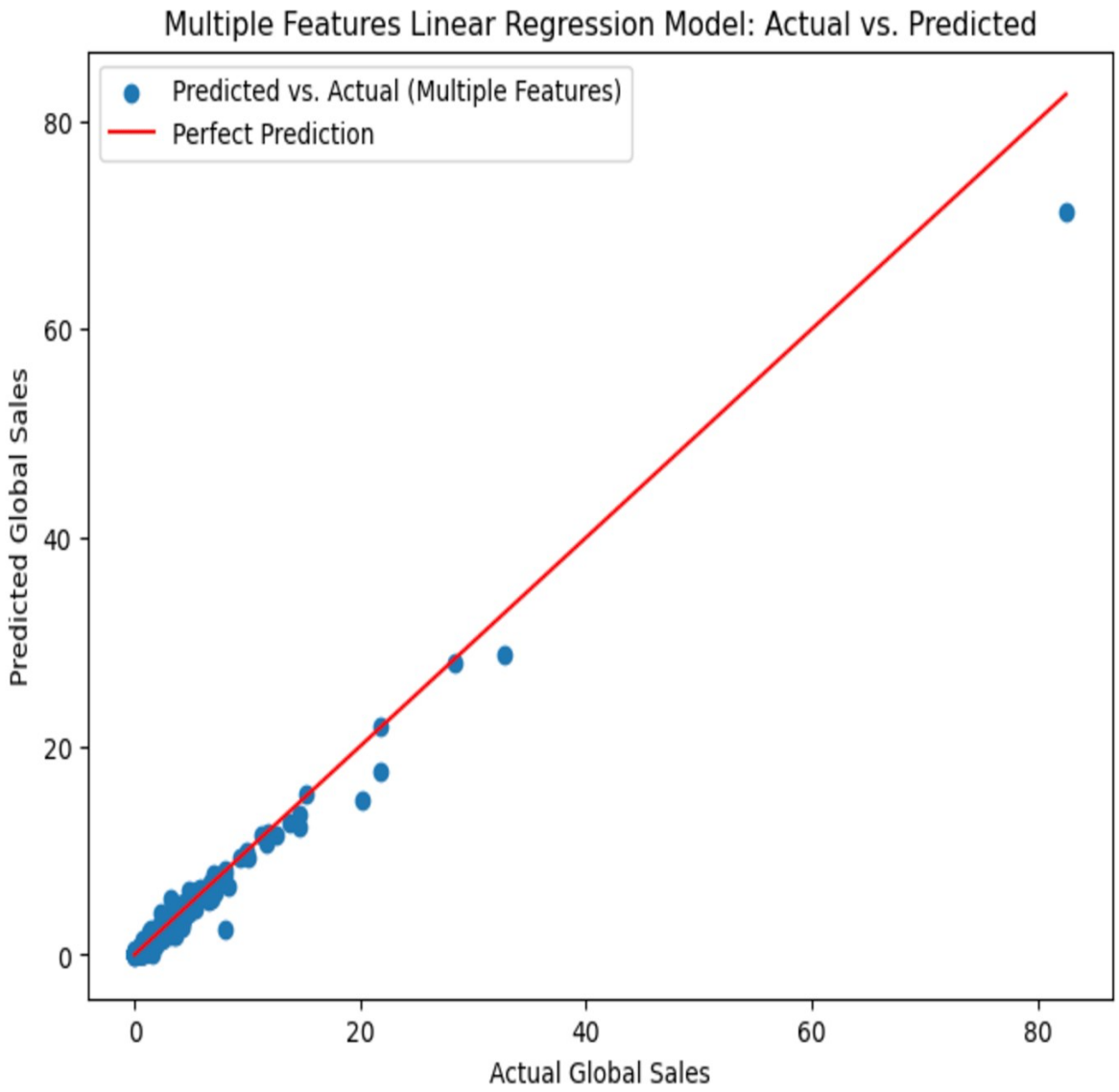

Figure 11: Multiple Features Linear Regression

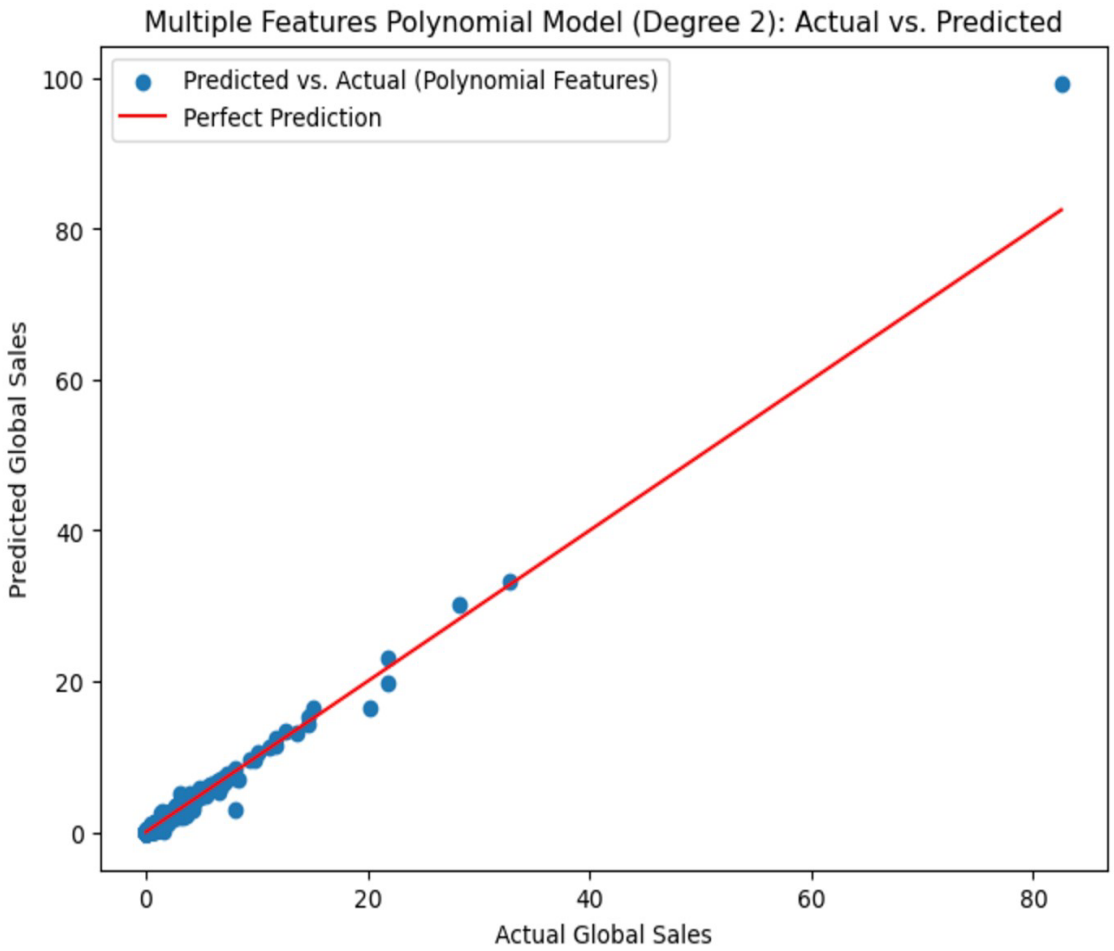

Figure 12: Multiple Features Polynomial Regression

3.3 RANDOM FOREST REGRESSION (CATEGORICAL & NUMERICAL FEATURES)

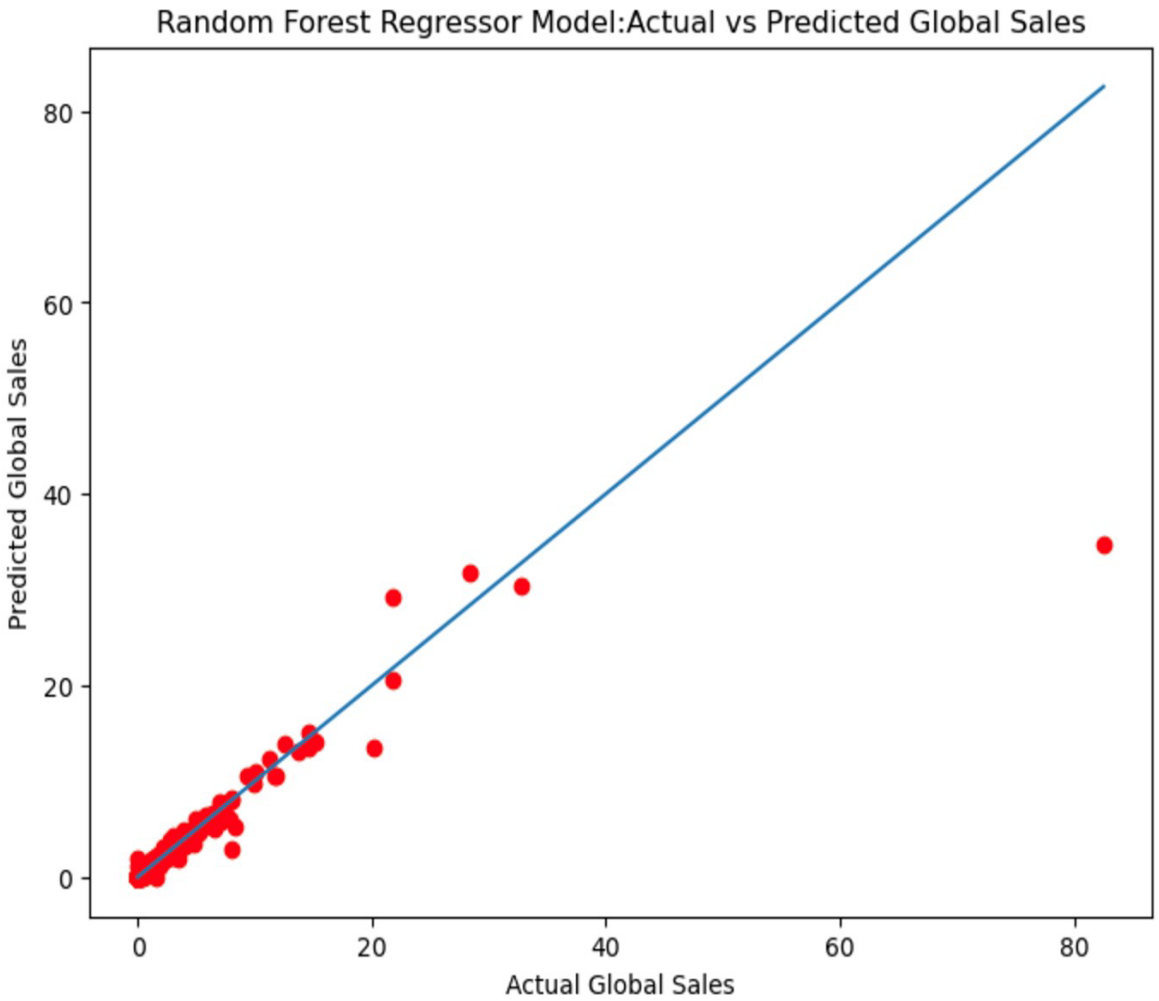

Figure 13: Random Forest Regressor Model

Table 6: Random Forest Regressor Evaluation Metrics

(R^2):	**0.8195**
MSE:	0.7643
RMSE:	0.8742
MAE:	0.0716

In **Figure 13** it is observed that the model's predictions are close to the actual values, with lower errors on average. **Table 6** also indicates Random Forest Regressor has an MSE of 0.7643 and R^2 Score of 0.8195.

3.4 ARTIFICIAL NEURAL NETWORK

Table 7: ANN Architecture

Parameters	**Details**
Model	1
Number of Layers	3
Neurons per Layer	64
Dropout Rate	10 %
Output Dimension	1
Input Dimensions	15
Optimizer	Adam (Learning rate: 0.001)
Loss Function	Mean Squared Error(MSE)
Batch size	None
Epochs	200
Validation strategy	10% (0.1)
Early stopping	20

Activation Function	ReLU(Hidden Layers), Linear(Output Layer)

To improve predictive accuracy, hyperparameter tuning was conducted across Models 2, 3, and 4.

Table 8: ANN Hyperparameter Tuning

Parameters	**Details**
Model	2
Number of Layers	3
Neurons per Layer	64
Dropout Rate	30 %
Output Dimension	1
Input Dimensions	15
Optimizer	Adam (Learning rate: 0.001)
Loss Function	Mean Squared Error(MSE)
Batch size	None
Epochs	200

Validation strategy	10% (0.1)
Early stopping	20
Activation Function	ReLU(Hidden Layers), Linear(Output Layer)

Table 9: Hyperparameter Tuning

Parameters	**Details**
Model	3
Number of Layers	3 Hidden Layers
Neurons per Layer	64
Dropout Rate	10 %
Output Dimension	1
Input Dimensions	15
Optimizer	Adam(Learning rate: 0.01)
Loss Function	Mean Squared Error(MSE)
Batch size	None

Number of epochs	200
Validation strategy	10% (0.1)
Early stopping	20
Activation Function	ReLU(Hidden Layers), Linear(Output Layer)

Table 10

Parameters	**Details**
Model	4
Number of Layers	3
Neurons per Layer	128
Dropout Rate	10 %
Output Dimension	1
Input Dimensions	15
Optimizer	Adam (Learning rate: 0.001)
Loss Function	Mean Squared Error(MSE)

Batch size	None
Epochs	200
Validation strategy	10% (0.1)
Early stopping	20
Activation Function	ReLU(Hidden Layers), Linear(Output Layer)

Table 11: ANN Models Evaluation Metrics

Evaluation Metrics	**Model 1**	**Model 2**	**Model 3**	**Model 4**
MSE	0.0439	4.2387	200879.0778	4.2363
R^2	0.9896	-0.0009	-47435.0892	-0.0004

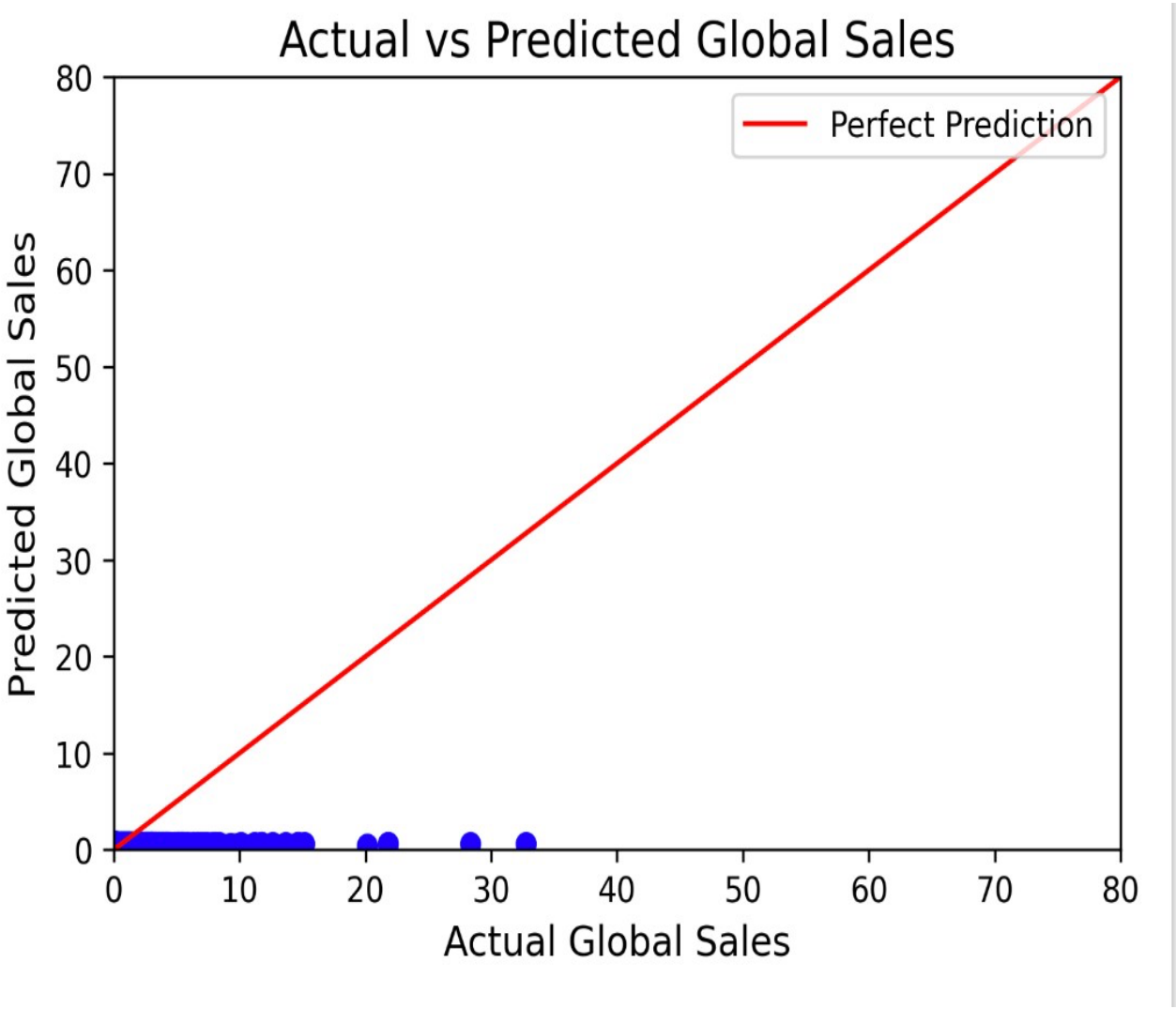

Figure 14: Model Predictions

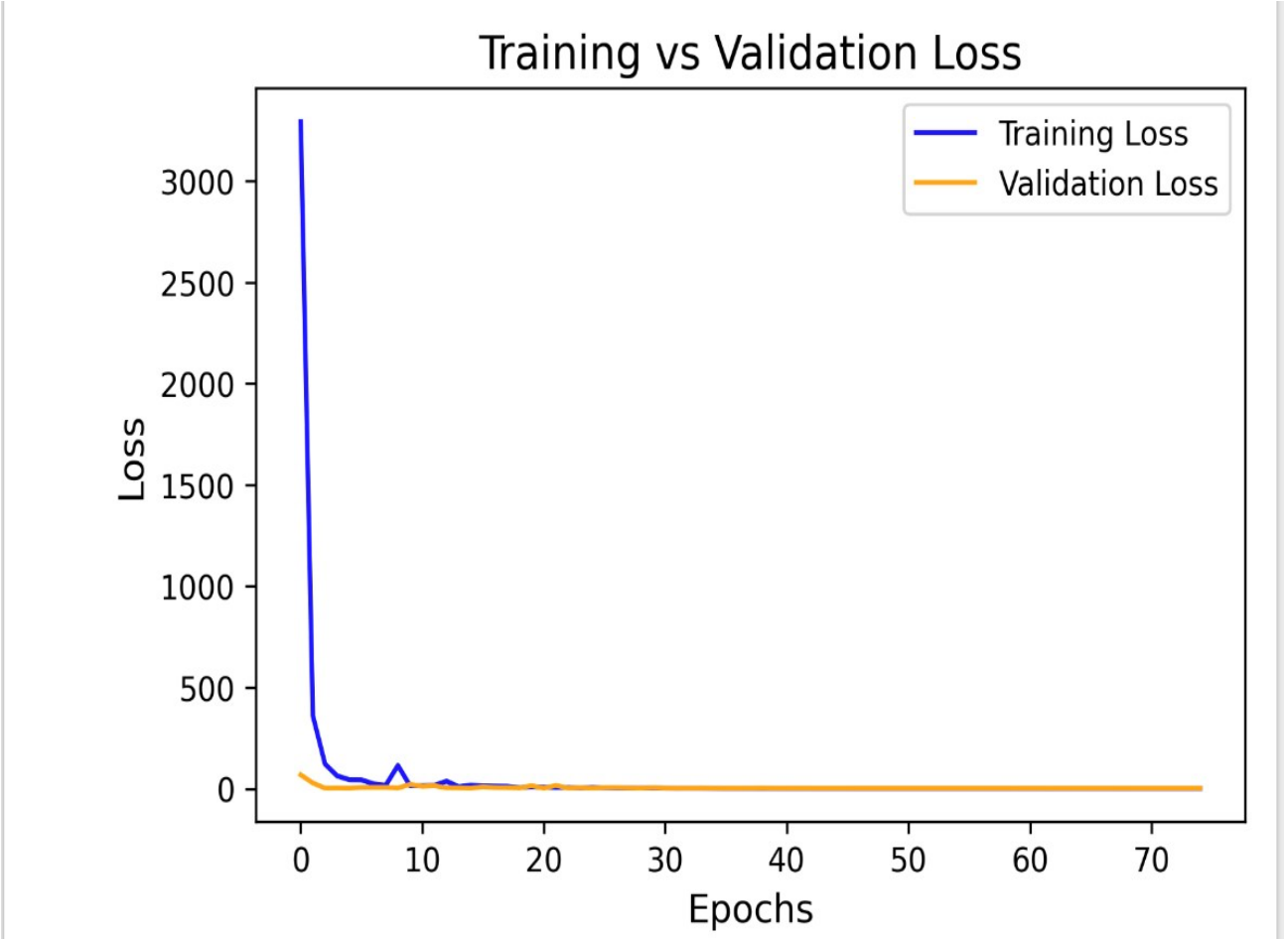

Figure 15: Model Loss

3.5 MODEL COMPARISON

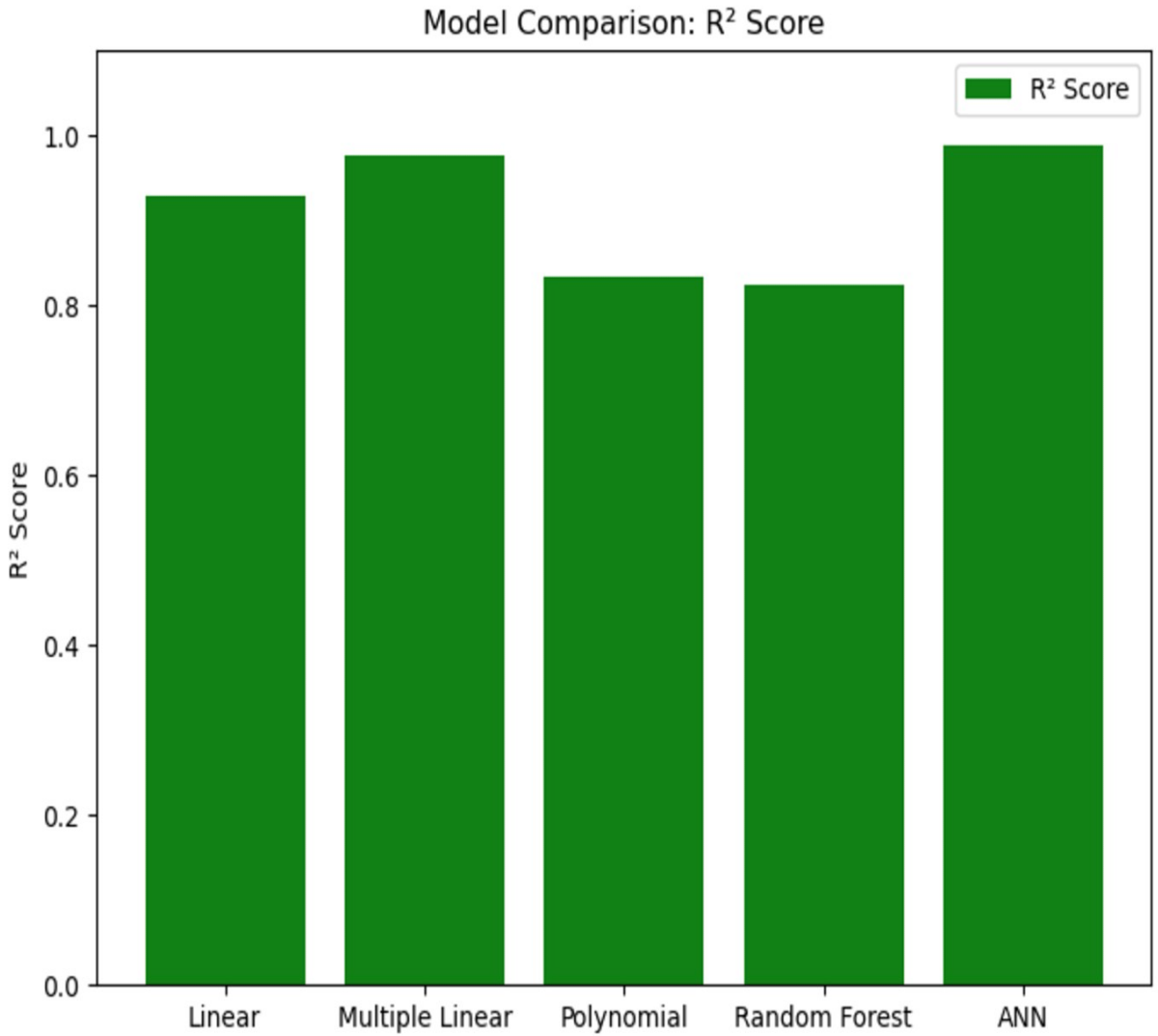

Figure 16: Comparing all models (MSE)

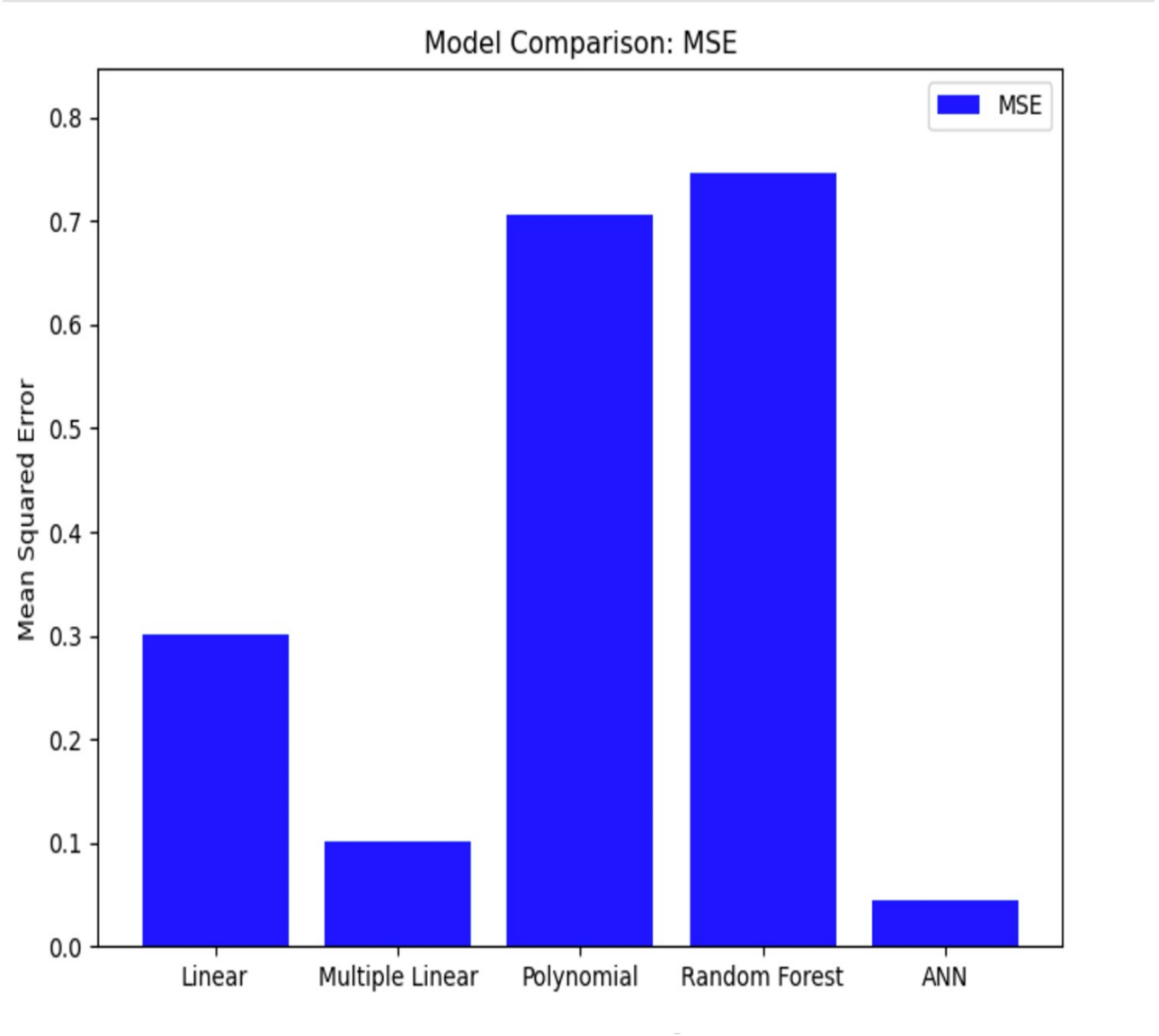

Figure 17: Comparing all models (R^2)

In Figure 15 & 16 it is observed that ANN has the lowest MSE(0.0439) and highest R^2 (0.9896) of all the models.

3.5 K-MEANS CLUSTERING

In Table 12 below the combination of 'NA_Sales', 'EU_Sales' has the lowest DBI (0.6119) and highest Silhouette Coefficient (0.9390).

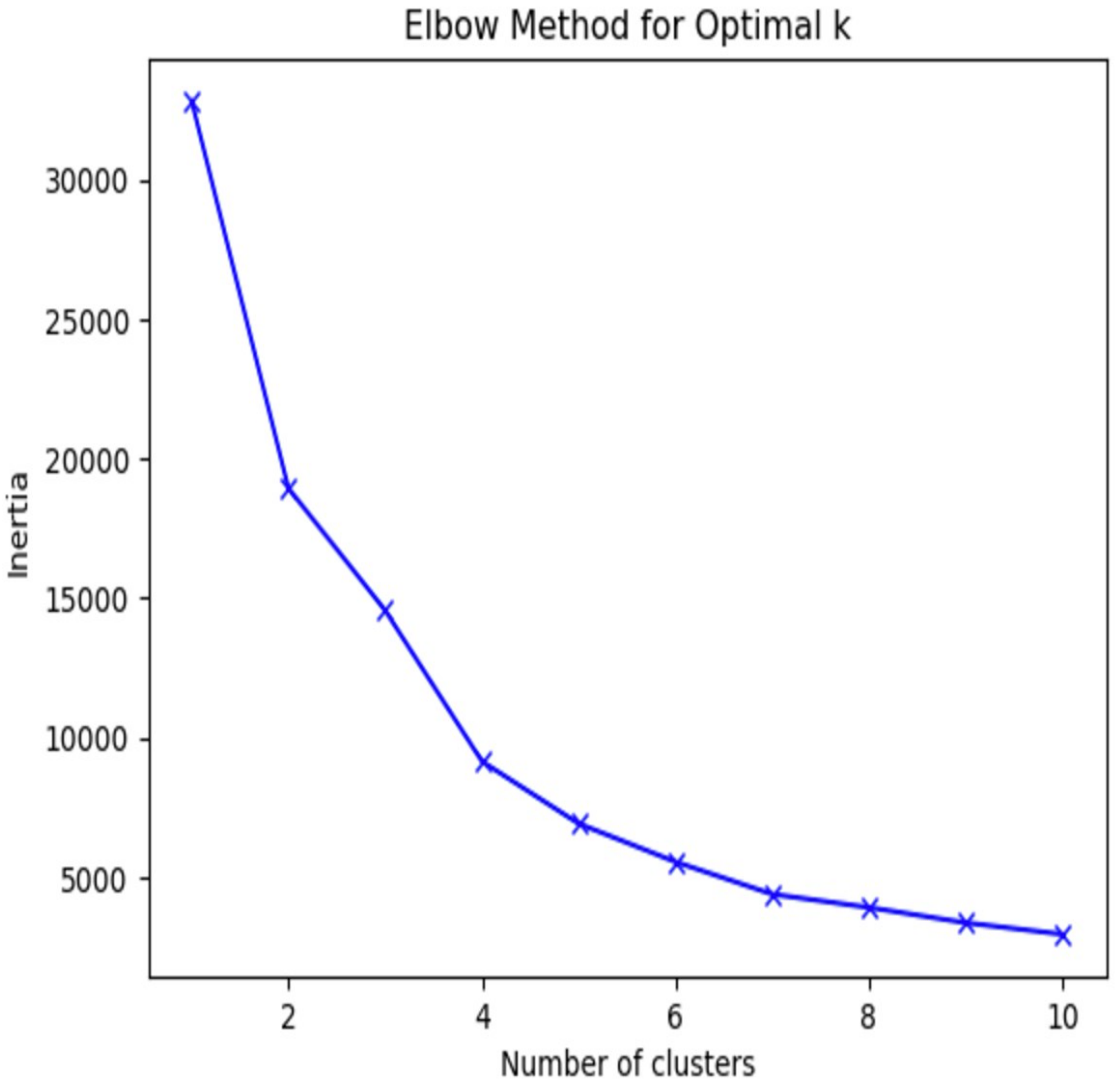
Elbow Method for Optimal k
Inertia
30000
25000
20000
15000
10000
5000
2
4
6
8
10
Number of clusters

Figure 18: Elbow Plot 'NA_Sales' vs 'EU_Sales'

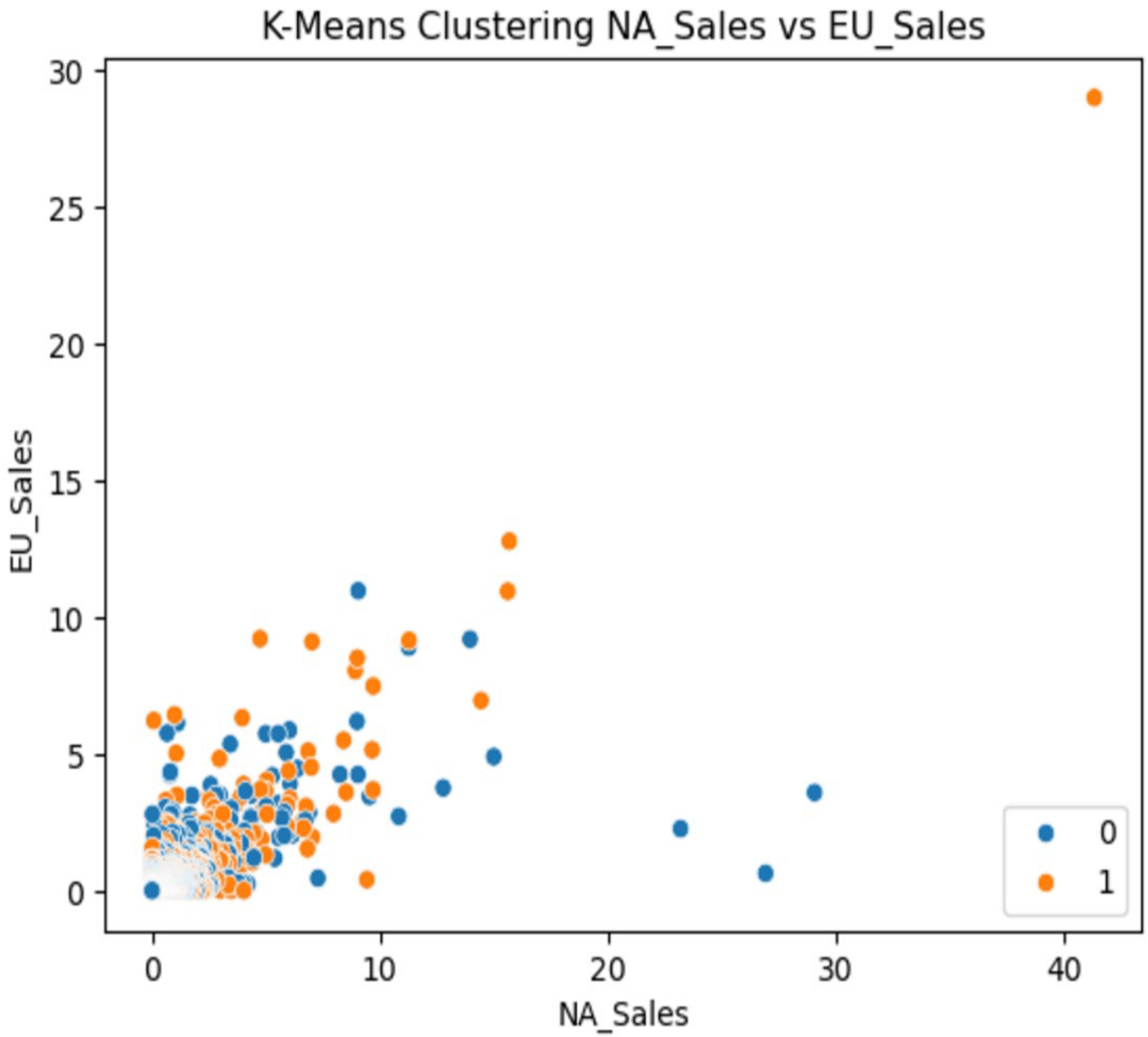

Figure 19: K-means clustering NA_Sales' vs 'EU_Sales'

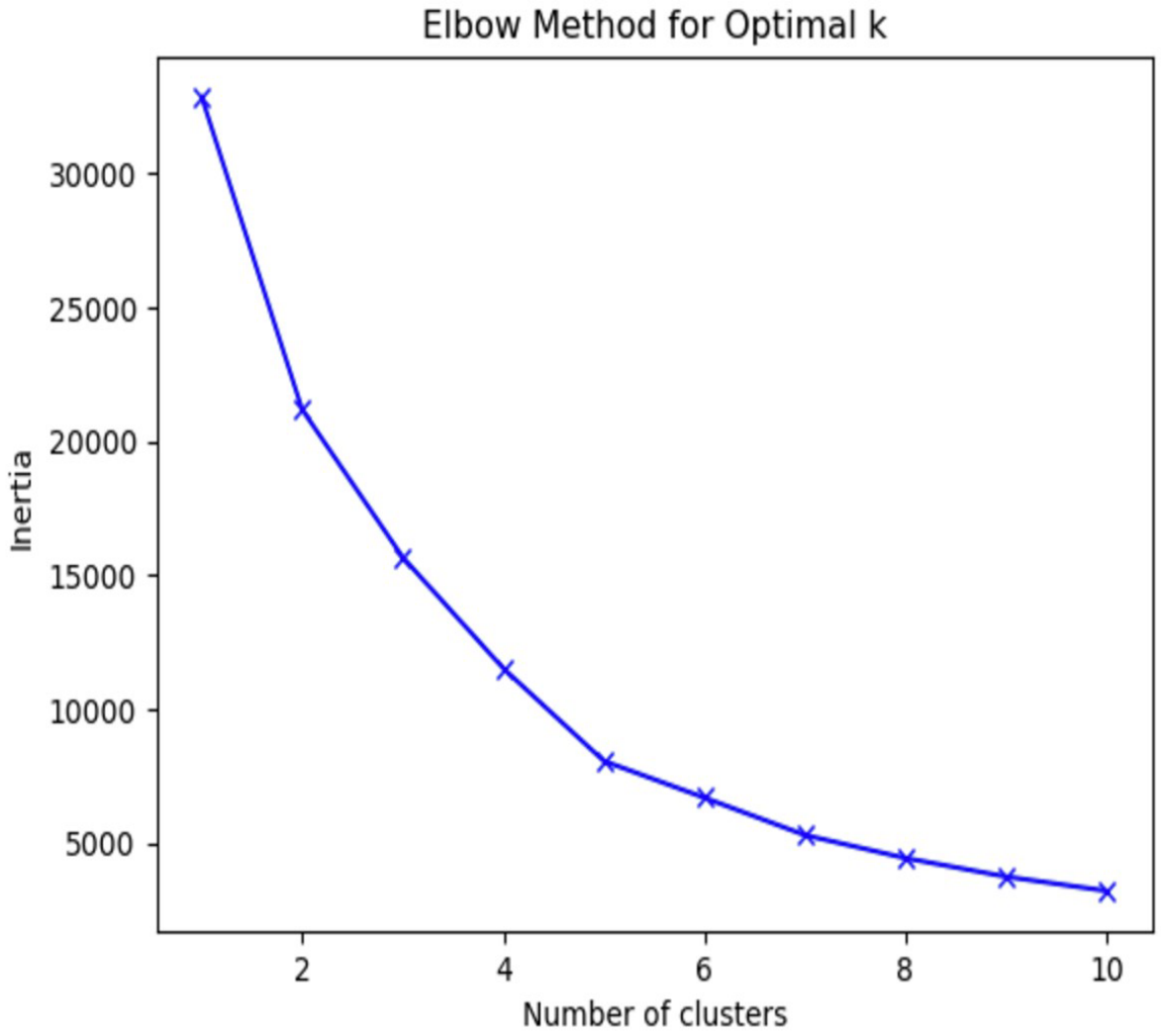

Figure 20: Elbow Plot 'JP_Sales' vs 'Other_Sales'

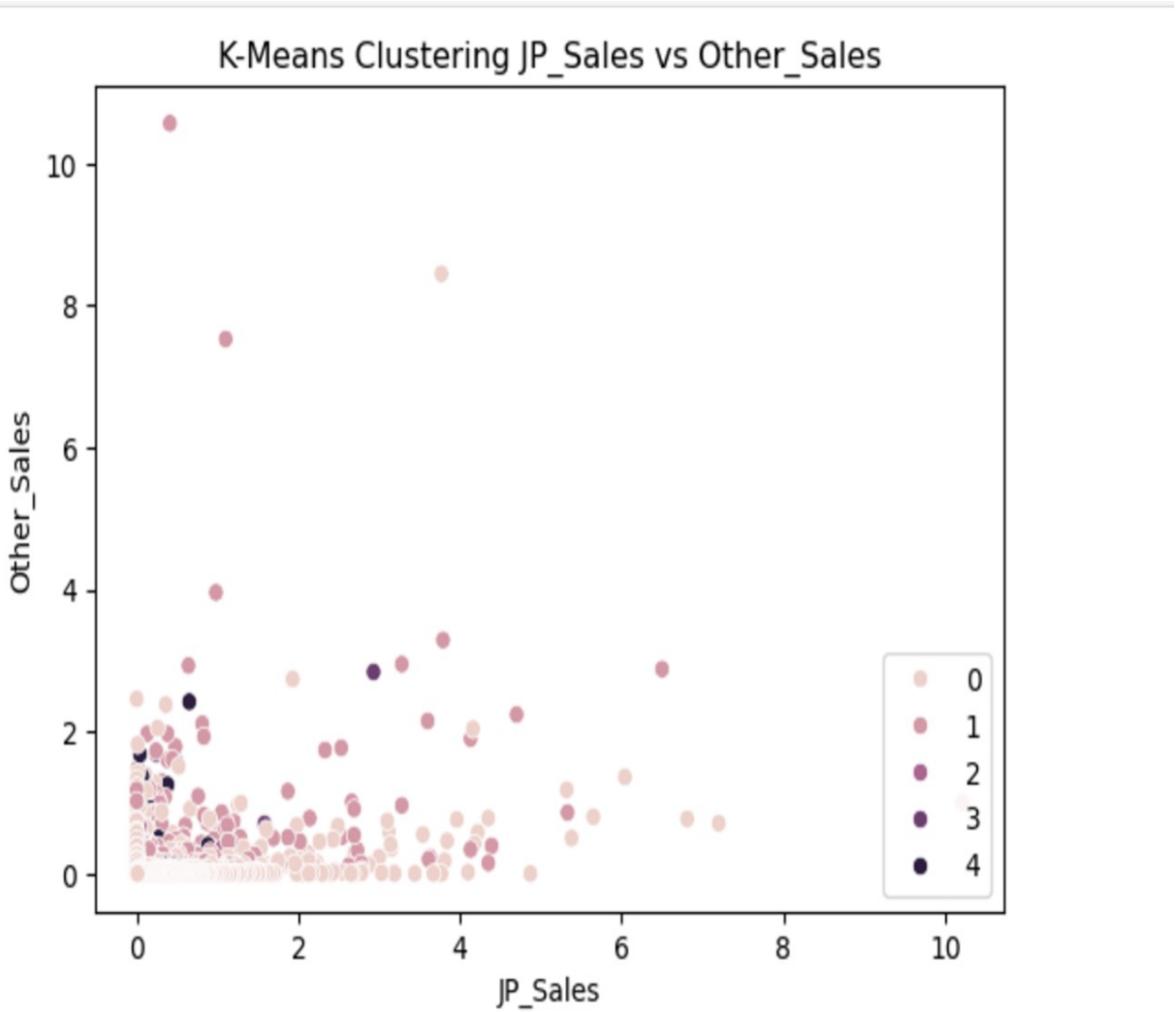
K-Means Clustering JP_Sales vs Other_Sales
Other_Sales
JP_Sales
0
2
4
6
8
10
0
1
2
3
4

Figure 21: K-means clustering JP_Sales' vs 'Other_Sales’

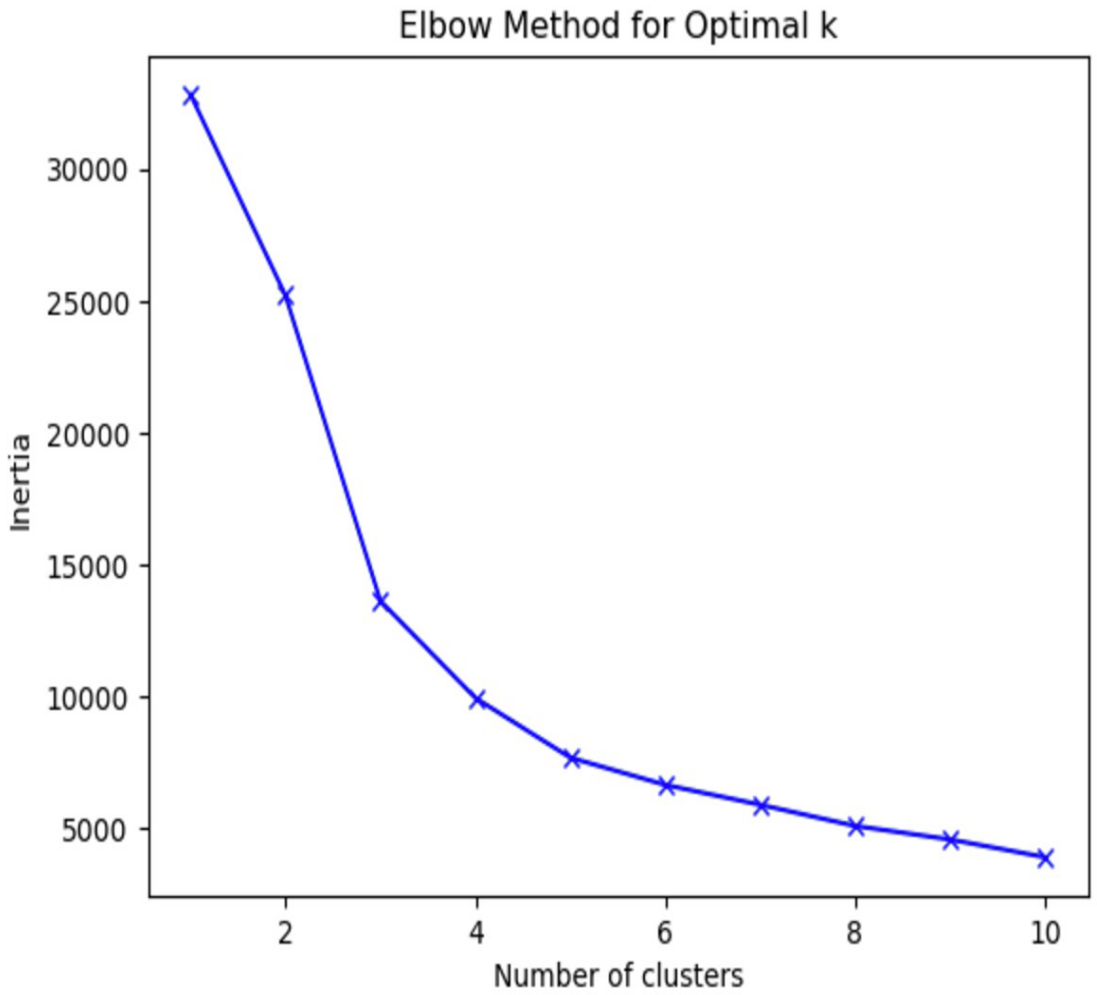

Figure 22: Elbow Plot 'Critic_Score' vs 'User_Score’

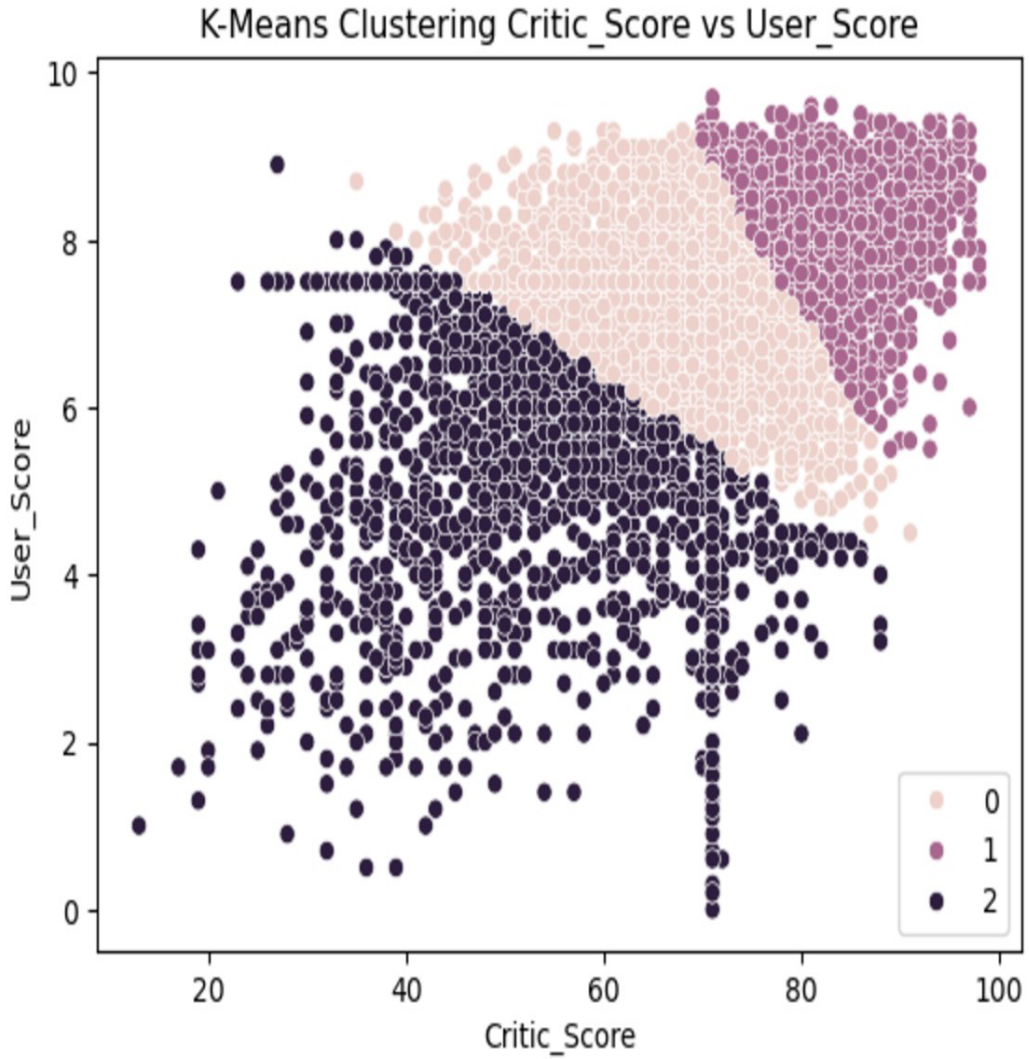

Figure 23: K-means clustering 'Critic_Score' vs 'User_Score'

Table 13: K-Means Evaluation Metrics

	NA_Sales vs EU_Sales	JP_Sales vs Other_Sales	Critic_Score vs User_Score
Optimal k	**2**	**5**	**3**
DBI	0.6119	0.7285	0.7602
Silhouette Coefficient	0.9390	0.6261	0.5712

3.6 HIERARCHIAL CLUSTERING (AGGLOMERATIVE)

In Table 14 we can observe that 'NA_Sales', 'EU_Sales' combination has the lowest DBI Score of 0.0073 and highest Silhouette Score of 0.9900.

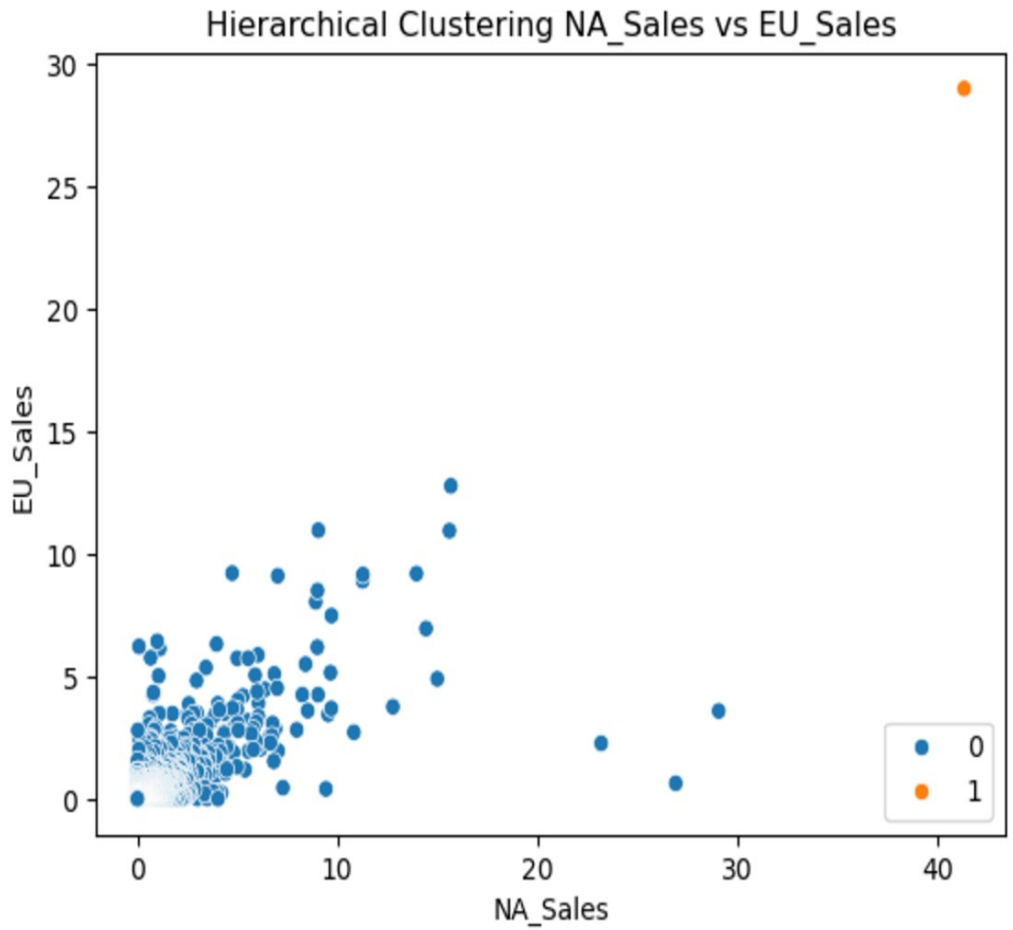

Figure 24: Hierarchical NA_Sales vs EU_Sales

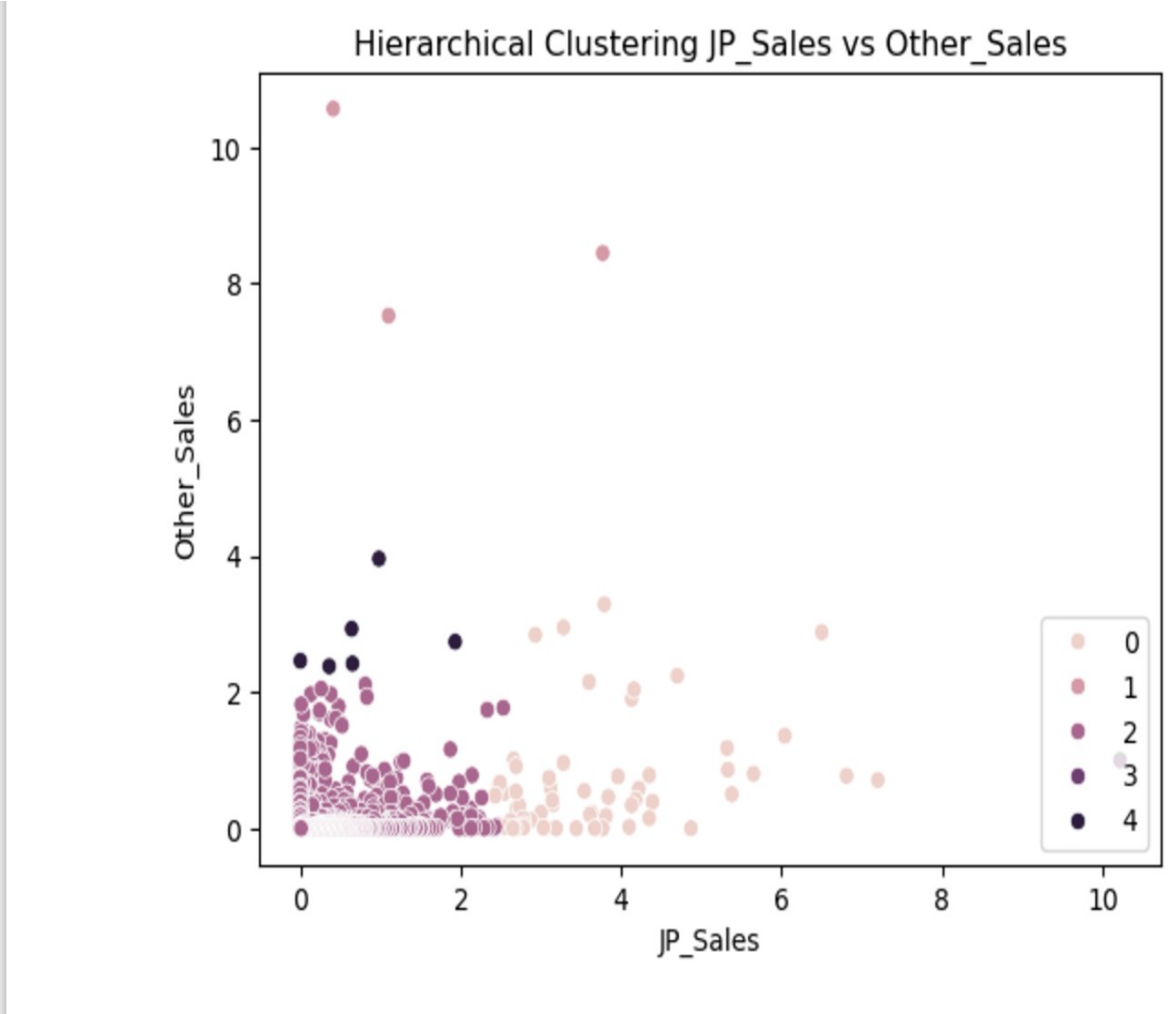

Figure 25: Hierarchical clustering ‘JP_Sales' vs 'Other_Sales’

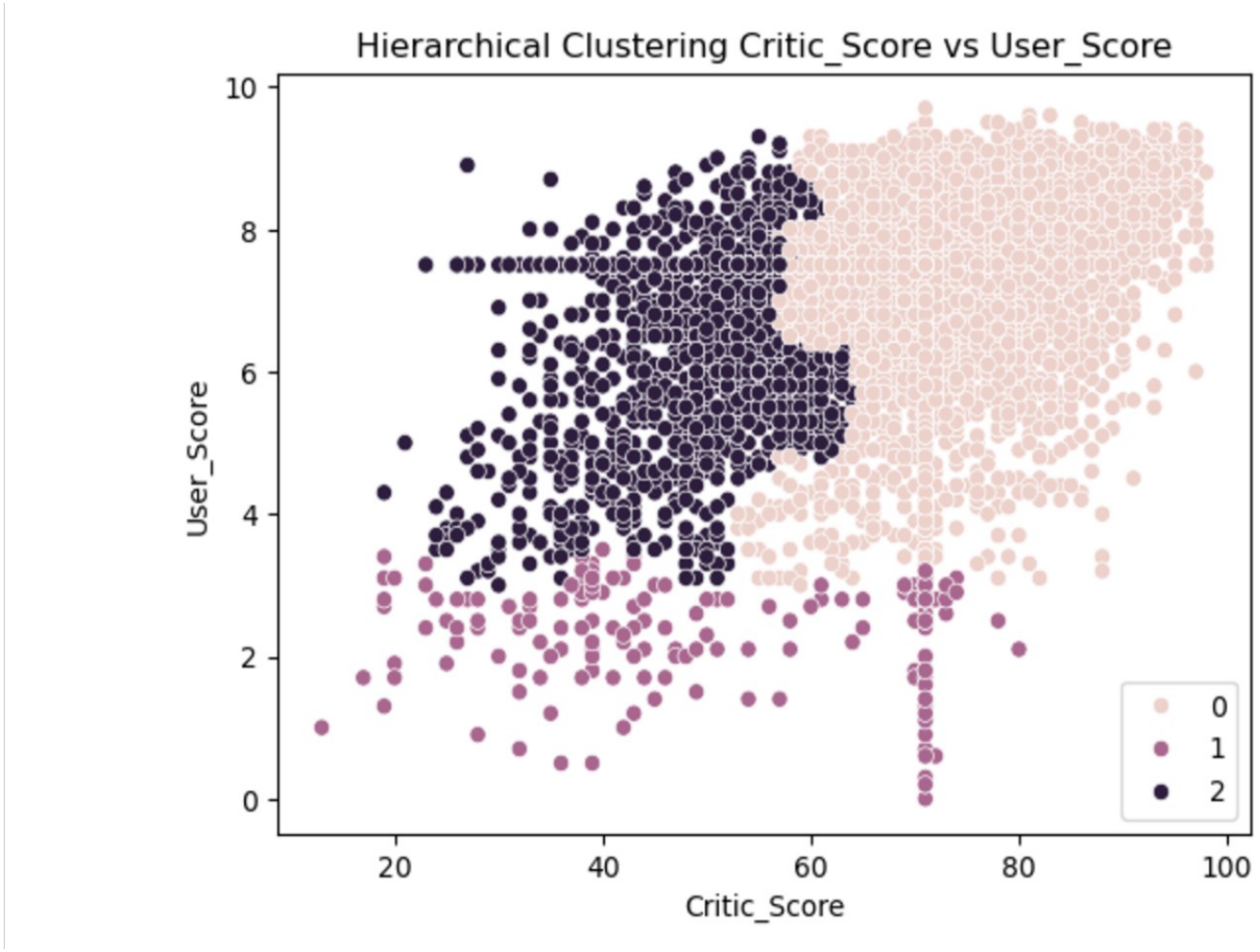

Figure 26: Hierarchical clustering 'Critic_Score' vs 'User_Score'

Table 14: Hierarchial Clustering Evaluation Metrics

HIERARCHIAL	NA_Sales vs EU_Sales	JP_Sales vs Other_Sales	Critic_Score vs User_Score
clusters	**2**	**5**	**3**
DBI	0.0073	0.4162	0.7963
Silhouette Coefficient	0.9900	0.9437	0.6250

4.0 DISCUSSION

From the analysis in Table 13 and 14, we have been able to deduce that among numerical variables, "NA_Sales" provides the best prediction of global sales of video games with a high R^2 value (0. 9290) in simple linear regression. Linear models also outperformed polynomial models for all numerical features except JP_Sales,Critic_Score,User_Score,User_Count and Critic_Count.

The multiple regression models, incorporating several numerical features, demonstrated improved predictive accuracy for global sales of video games achieving R^2 of 0.9708.

The result from Random Forest Regressor using both categorical and numerical features underperformed compared to all other regression models with R^2 of 0.8195.

Artificial Neural Network (ANN) models, incorporating both categorical and numerical features, demonstrated superior performance in predicting global sales of video games compared to all other models. However Model 1 from Table 7 emerged as the most effective. This is reflected in its lower MSE and higher R^2 values, as illustrated in Figures 14 and 15.All other models performed poorly.

Analysis of K-Means clustering using various numerical features combinations, "NA_Sales vs. EU_Sales" provided the best clustering results with the highest Silhouette Coefficient of 0.9390**),** showing well-defined and distinct clusters.

In Table 13 and 14 we can observe that hierarchical clustering outperforms K-Means by producing higher silhouette score (0.9900) and lower DBI (0.0073) for all the combinations except 'Critic_Score' vs 'User_Score'.

4.1 CONCLUSION

In conclusion, ANN model is the best model for predicting global sales of video games with the highest R^2 and lowest MAE while Hierarchical Clustering (Agglomerative), with its high silhouette score and low DBI produces the best clustering using "NA_Sales vs. EU_Sales" combination.

REFERENCES

Jain, H., Vashi, D., Ray, S.K. and Vishal, (2023). *Sales Prediction Using Machine Learning*. In: *Proceedings of the KILBY 100 7th International Conference on Computing Sciences (ICCS 2023)*. https://dx.doi.org/10.2139/ssrn.4495850

Affan, Vishwakarma, S., & Kumari, R., 2024. *Video game sales prediction model using regression model*. Galgotias University. Available at: http://dx.doi.org/10.2139/ssrn.4935036