COMPONENT TWO: SALES PERFORMANCE OF VIDEO GAMES

ABSTRACT

As the number of gamers continue to grow worldwide, the video games market grows as well. Analyzing what factors influence the global and regional markets is very important to various stakeholders in the sector. This Kaggle videogames dataset allows the analysis of various factors and relationships in the videogames market and gives a better understanding on how certain variables influence sales and also identifies various patterns in the market. The analysis showed that the American market is the strongest market for video games sales, there isn't a strong single determinant of video games sales but a combination of critics and user opinion is a decent metric. Video games are classified better based on the platforms they are released under and they form better clusters based on ratings. Other factors including price of the video games should be studied to get a better understanding of the video games market

INTRODUCTION

The video game industry is a fast-growing, highly competitive market with billions of gamers worldwide. Statista (2021) estimated the numbers of gamers worldwide at 3.2 billion. Predicting video game sales performance and understanding the factors that influence it is important for game developers, publishers, game console manufacturers, and video game retailers. The Kaggle's video game sales dataset provides a valuable resource for analyzing and investigating video game sales patterns globally. The purpose of this study is to examine how various factors predict global sales of video games, assess the impact of critic and user ratings on sales in different regions, classify and group video games data based on categorical variables. By answering these research questions, this study aims to contribute to the understanding of the video game market and provide insights for game developers and publishers.

METHODOLOGY

The methodology entails downloading the dataset from Kaggle, cleaning the dataset, examining the variables, and applying correlation and various regression methods to see which variables best predict sales globally and in various regions. Various regression methods are also applied to a combination of the critic and user variables to determine how they influence sales in various regions. The dataset's pertinent categorical variables will be used for classification, and the best variable to characterize the groups formed will be chosen using internal and external evaluation metrics. The analysis will be carried out in a Jupyter Notebook. The study will give game publishers and developers information about the video game market.

RESULTS

WHICH OF THE VARIABLES IN THE VIDEO GAME DATASET OR A COMBINATION OF THEM BEST PREDICTS "GLOBAL SALES" OF VIDEO GAMES AND WHY? PROVIDE QUANTITATIVE JUSTIFICATIONS FOR YOUR ANSWERS.

Understanding factors that influence global sales of video games can be a of huge benefits to game manufacturers, console makers, and game retailers.

The following variables are considered for this exercise as individual variables and as a combination

- **REGIONAL SALES:** to understand what markets perform better and contribute to the global sales, informing global success expectations based on performance in certain regions.

- **CRITICS AND USERS SCORE AND COUNTS:** This is useful to understand how other users and critics experience and reviews affect the sales performance of games
- **PLATFORM, GENRE AND RATING:** This is important to understand if users buy games based on certain platforms they're available on, the genre of the game, or how the game is rated.

In understanding how these variables best predict Global sales, a correlation heatmap is first generated to see the relationship between these values with global sales

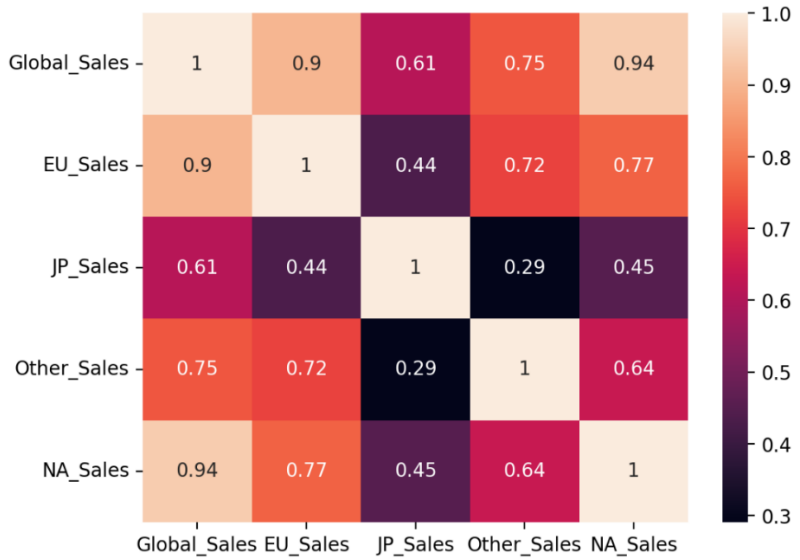

Figure 1:Heatmap showing the relationship between global sales and regional sales

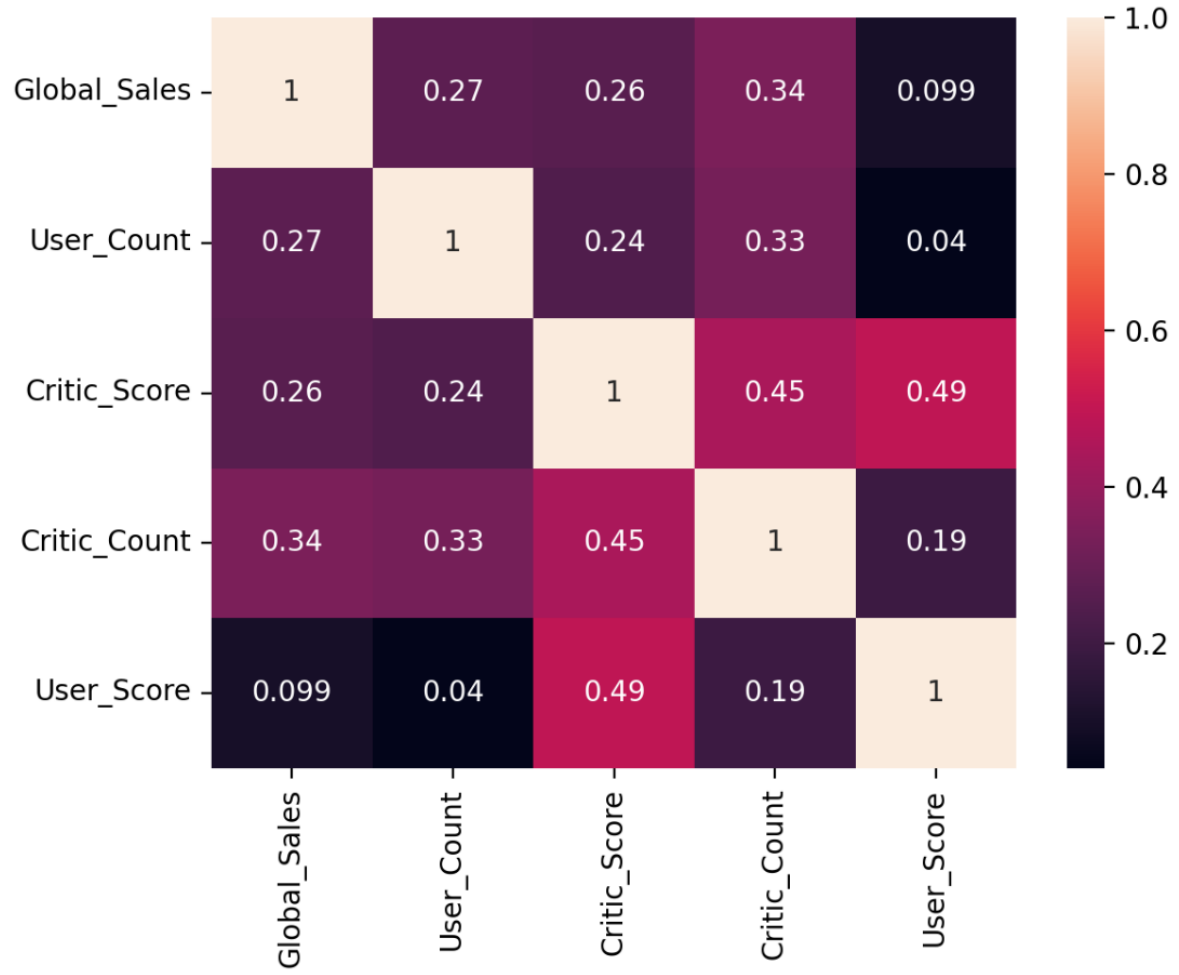

Figure 2:Correlation heatmap showing the relationship between users/critics with Global sales

There North American sales at 0.94 has the highest positive correlation with the Global sales, which is an early indicator that this is the best predictor of Global sales. Games that sell well in North America are likely to do well globally. This is followed by European sales, the rest of the word combined, then Japan sales.

The second heat map shows a positive correlation coefficient for all the values, but these weak values indicate that while they have an individual positive correlation with global sales, they are not strong predictors of Global sales. The critic count shows the highest positive correlation indicating the number of critics who choose to review a game is fairly important

Regression analysis is done on all of the regional sales against global sales

REGION	Best Regressor	MSE	RMSE	MAE	R-2 SCORE
North America Sales	Linear Regression	0.2875	0.5362	0.2025	0.9304
Europe Sales	Linear Regression	0.2995	0.5472	0.2402	0.9275
Other Region Sales	Support Vector regression	0.7221	0.8498	0.2578	0.8251

Japan Sales	Gradient Boosting	2.4140	1.5537	0.5191	0.4152

The north American sales is the best predictor of Global sales with the lowest error the highest r2 score.

Regression on the user and critics scores and count against global sales

	Best Regressor	MSE	RMSE	MAE	R-2 SCORE
User Score	Gradient Boosting	4.0775	2.0193	0.5891	0.0122
User Count	Linear Regression	4.0071	2.0018	0.5692	0.0293
Critic Score	Gradient Boosting	3.8977	1.9743	0.5136	0.0558
Critic Count	Gradient Boosting	3.7182	1.9283	0.5159	0.0993

Though the critic count shows the highest score of these variables, it is not a good predictor of global sales.

Multiple regression on all the user and critic variables

Variable	Best Regressor	MSE	RMSE	MAE	R-2 SCORE
All user and critic variables	Gradient Boosting	3.6594	1.9130	0.4640	0.1135

The combination of the values together still isn't a strong enough predictor of global sales

Platform, Genre and ratings

Variable	Best Regressor	MSE	RMSE	MAE	R-2 SCORE
Platform	Random Forest Regression	6.0405	2.4577	0.6883	0.0114
Genre	Linear Regression	6.0890	2.4676	0.7042	0.0035
Ratings	Random Forest regression	6.0774	2.4652	0.7111	0.0054
Combination of all 3 (Multiple regression)	Random Forest Regression	5.9170	2.4325	0.6893	0.0316

Rating performs best among the variables but it is still not a strong predictor of global sales. Even the combination of all the variables still does not perform strongly enough.

The strongest predictor of global sales from the analysis done is the sales in North America. Outside of the sales values, the strongest single indicator is the r2 score of 0.09 means it explains 9% of the variation in the dependent variable. The strongest combination is all the sales column which is

expected as this is they make up the summation of global sales. The strongest combination outside of sales is the combination of critics and users scores and counts as it has an r2 score of 0.1135, explaining 11% of the variation in Global sales.

WHAT EFFECT WILL THE NUMBER OF CRITICS AND USERS AS WELL AS THEIR REVIEW SCORES HAVE ON THE SALES OF VIDEO GAMES IN NORTH AMERICA, EU AND JAPAN?

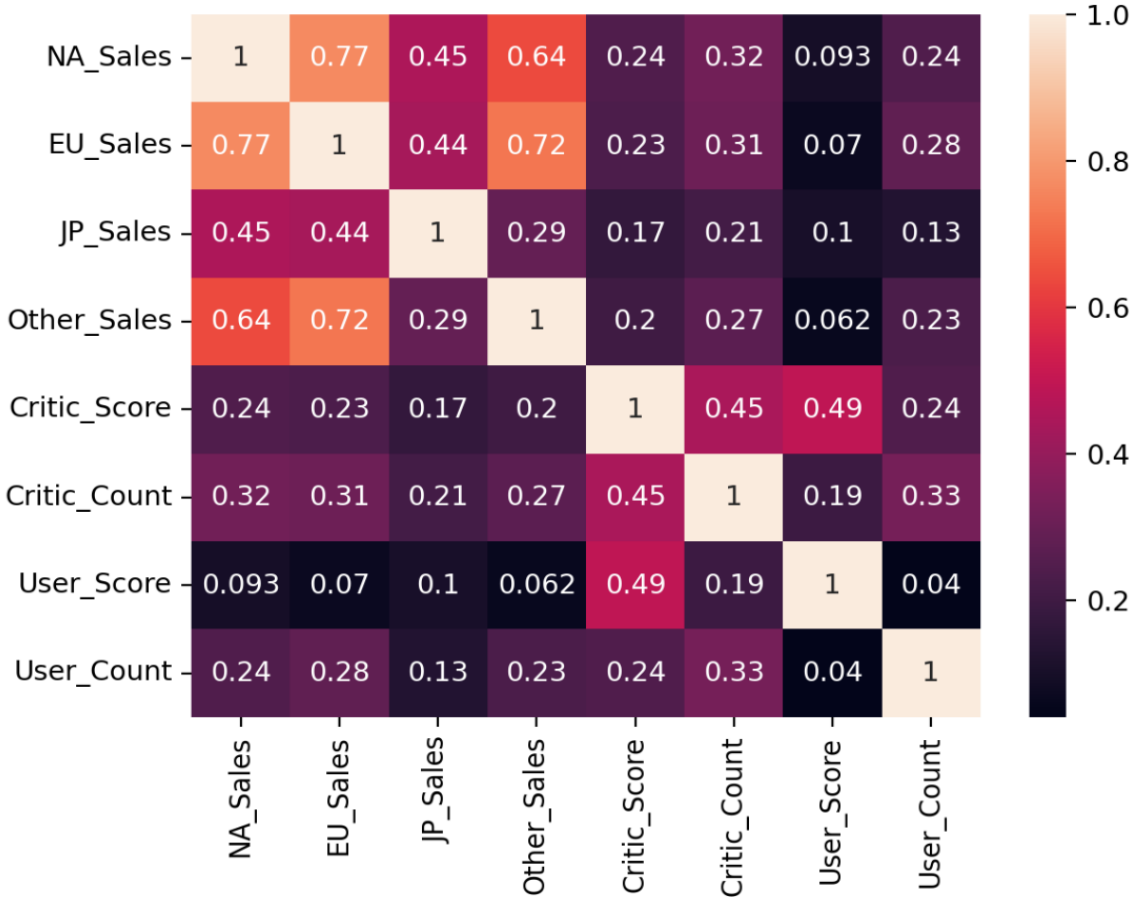

Figure 3: Correlation heatmap for critic and user scores against regional sales

There is positive correlation between the critics and users and their counts and scores against the sales in the various regions. But all of these results are generally weak and suggest they do not have much of an effect on the various sales regions. But the critic count still shows the most effect of the listed variables with a correlation coefficient of 0.32 against North America sales.

All of the critic and user variables are combined to do a multiple regression against the various regions

Target Variable	Best Regressor	MSE	RMSE	MAE	R-2 SCORE
NA Sales	Gradient Boosting	0.9261	0.9624	0.2545	0.1020
EU Sales	Multiple Linear Regression	0.4474	0.6689	0.1791	0.0781
Japan	Multiple Linear Regression	0.0913	0.3022	0.1184	0.0581

Looking at the best effect which is on North America, the R-squared score of 0.1020 indicates that only 10.20% of the variation in the dependent variable can be explained by the combination of users and critics. This means that there are other factors that are influencing the regional sales of video games that are not accounted for in the model.

WHAT PROPELLED THE CHOICE OF YOUR REGRESSOR FOR THIS TASK? APTLY DISCUSS WITH QUANTITATIVE REASONS!

Different regression models perform differently depending on the nature of the relationship pf the variables being compared. I used various regression models in analyzing the data for various independent variables against global sales and also when analyzing multiple independent variables against regional sales. In analyzing independent variables against the global sales, the linear regression outperformed all other regressors indicating it is the best model for this task when doing regression for a single variable with the following results posted as the best when looking at NA Sales against global sales

REGION	Best Regressor	MSE	RMSE	MAE	R-2 SCORE
North America Sales	Linear Regression	0.2875	0.5362	0.2025	0.9304

```
The evaluation metrics for all the regressors for all regional Sales combined

Linear Regression - Mean Squared Error: 0.0000, Root Mean Squared Error: 0.0053, Mean Absolute Error: 0.0030, R-squared Score:
1.0000
Random Forest Regression - Mean Squared Error: 0.6990, Root Mean Squared Error: 0.8361, Mean Absolute Error: 0.0397, R-squared
Score: 0.8307
Gradient Boosting Regression - Mean Squared Error: 0.5854, Root Mean Squared Error: 0.7651, Mean Absolute Error: 0.0453, R-squa
red Score: 0.8582
Lasso Regression - Mean Squared Error: 0.0274, Root Mean Squared Error: 0.1655, Mean Absolute Error: 0.0542, R-squared Score:
0.9934
Support Vector Regression - Mean Squared Error: 0.0056, Root Mean Squared Error: 0.0748, Mean Absolute Error: 0.0743, R-squared
Score: 0.9986
K Neighbors Regression - Mean Squared Error: 0.9083, Root Mean Squared Error: 0.9531, Mean Absolute Error: 0.0397, R-squared Sc
ore: 0.7800
```

In doing various multiple regressors for all the sales against global sales. The multiple linear regressor performs perfectly well with a MSE of 0.0000 and R2 score of 1.000 indicating it is the best at predicting linear relationships which is what exists between all regional sales and global sales, and that the other regressors are more suited to complex relationships.

In analyzing multiple variables with complex relationships against global sales and regional sales, the gradient boosting regressor outperformed the other regressors with lowest error rate and highest coefficient of determination of 11% with the detailed results as follows when looking at all critics and users against global sales:

Variable	Best Regressor	MSE	RMSE	MAE	R-2 SCORE
All user and critic variables	Gradient Boosting	3.6594	1.9130	0.4640	0.1135

Also when looking at users and critics against regional sales, Gradient boosting still performs better than other regressors. Screenshot of results of all regressors posted below. Result of Gradient Boosting Regression - Mean Squared Error: 0.9261, Root Mean Squared Error: 0.9624, Mean Absolute Error: 0.2545, R-squared Score: 0.1020.

```
The evaluation metrics for all critics and users against NA Sales

Linear Regression - Mean Squared Error: 0.9415, Root Mean Squared Error: 0.9703, Mean Absolute Error: 0.2914, R-squared Score:
0.0871
Random Forest Regression - Mean Squared Error: 0.9577, Root Mean Squared Error: 0.9786, Mean Absolute Error: 0.2716, R-squared
Score: 0.0714
Gradient Boosting Regression - Mean Squared Error: 0.9261, Root Mean Squared Error: 0.9624, Mean Absolute Error: 0.2545, R-squa
red Score: 0.1020
Lasso Regression - Mean Squared Error: 0.9579, Root Mean Squared Error: 0.9787, Mean Absolute Error: 0.2810, R-squared Score:
0.0712
Support Vector Regression - Mean Squared Error: 0.9985, Root Mean Squared Error: 0.9992, Mean Absolute Error: 0.2453, R-squared
Score: 0.0318
K Neighbors Regression - Mean Squared Error: 1.0345, Root Mean Squared Error: 1.0171, Mean Absolute Error: 0.2818, R-squared Sc
ore: -0.0030
```

USE ALL THE RELEVANT CATEGORICAL VARIABLES IN THE VIDEO GAME DATASET AS THE TARGET VARIABLE AT EACH INSTANCE AND DETERMINE WHICH OF THE VARIABLES PERFORMED BEST IN CLASSIFYING THE DATASET. EXPLAIN YOUR FINDINGS

I considered the Platform, Genre and Ratings as the relevant variables classified them against the variables influenced by user experience at each instance which are the Critic and user score and counts and the global sales variable which represents the summation of all regional sales.

The random forest classifier performed better than the other classifiers in classifying the data as it posted the highest accuracy across the three variables used in the classification. The Platform and ratings variable posted similar scores with both their accuracy at 0.69.

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.54	0.49	0.51	814
1	0.66	0.71	0.68	765
2	0.99	1.00	0.99	799
3	0.72	0.81	0.76	818
4	0.52	0.45	0.48	795
accuracy			0.69	3991
macro avg	0.68	0.69	0.69	3991
weighted avg	0.68	0.69	0.69	3991

Confusion Matrix for Rar

```
[[401 153   3  70 187]
 [ 99 543   2  45  76]
 [  2   0 797   0   0]
 [ 54  31   3 659  71]
 [190 102   2 140 361]]
```

Figure 4: Random forest Classification report and confusion matrix for the Rating Variable

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.75	0.90	0.82	305
1	0.98	1.00	0.99	297
2	0.32	0.19	0.24	295
3	0.61	0.67	0.64	274
4	0.63	0.66	0.64	285
5	0.75	0.81	0.78	313
6	0.81	0.92	0.86	293
7	0.37	0.26	0.30	314
8	0.58	0.58	0.58	309
9	0.76	0.89	0.82	281
10	0.64	0.64	0.64	281
11	0.82	0.95	0.88	281
12	0.48	0.38	0.43	318
13	0.86	0.95	0.90	292
14	0.58	0.47	0.52	287
15	0.59	0.61	0.60	285
16	0.81	0.88	0.84	326
accuracy			0.69	5036
macro avg	0.67	0.69	0.68	5036
weighted avg	0.67	0.69	0.67	5036

Confusion Matrix for Random Forest Classifier:

```
[[275   0   3   0   0   3   0   1   4   1   4   1   1   1   7   0   4]
 [  0 297   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [ 18   1  57  26  22  21  11  19  10   7  16  13  26   5   5  32   6]
 [  5   0   4 184  20   2   9  14   4   2   2   2  12   2   1   7   4]
 [  6   2   3  12 187   5   4  13   6   2  11   2   9   2   6   9   6]
 [  2   0   5  10   0 253   0   4   4   8   4   4   4   3   2   4   6]
 [  1   1   1   8   3   0 270   0   0   0   4   0   3   2   0   0   0]
 [  9   0  25  21  11   8  14  81  27   6  18   4  34   4  18  29   5]
 [  6   1   5   4   5   5   5  13 178  11   3  10  11   7  27   4  14]
 [  5   0   3   1   0   2   0   0   6 249   2   2   2   2   1   1   5]
 [ 13   0   9   6   9   5   4   6  14   5 180   2   9   0   7  10   2]
 [  2   0   3   3   1   1   0   0   0   1   0 266   0   3   0   1   0]
 [  7   0  24  12  11  14  12  28  21   6  16   4 121   7  15  12   8]
 [  1   0   2   1   0   0   0   0   2   4   1   1   1 278   1   0   0]
 [  9   0  18   6   8   6   1  10  25  16  13   7  10   3 136  10   9]
 [  7   0  13   8  18   3   1  29   6   4   4   2   7   3   6 174   0]
 [  3   0   2   1   0  11   1   3   1   6   2   4   1   2   3   0 286]]
```

Figure 5:Random forest classification report and confusion matrix for the platform variable

The accuracy for classification by ratings was 0.69, while the accuracy for classification by platform was 0.69 as well but platform has a larger sample size. The precision, recall, and F1-score for classification by platform were higher for most of the classes than the ones obtained for classification by ratings. Additionally, the confusion matrix for classification by platform is more balanced and has fewer misclassifications.

So the best performing variable is the Platform variable. Looking closely at the diagonal elements of the matrix, the classifier has high accuracy for some classes, such as class 1, where it correctly predicted all of the samples. However, for some other classes, such as class 2, the classifier has lower accuracy, predicting only 57 out of 295 samples correctly.

HOW DID YOU CHECK WHETHER YOUR MODELS DID NOT OVERFIT?

I applied a Regularizer to the random forest classification model for the platform variable and the accuracy value did not change much. The test accuracy was 0.684.

CAN YOUR CLASSIFICATION MODELS BE DEPLOYED IN PRACTICE BASED ON THEIR PERFORMANCES? EXPLAIN.

It should not be deployed in practice. While the model performs reasonably well generally, there are some classes where its classification is very low and in practice that could have significant consequences like class 2 where it only classifies 57 correctly out of 295 with an f1 score of 0.24. But if the misclassification of some items is tolerable in the context it is to be deployed in, then it can be deployed.

IN THE VIDEO GAME DATASET, USE A RELEVANT CATEGORICAL VARIABLE AND OTHER RELEVANT NON-CATEGORICAL VARIABLES TO FORM GROUPS AT EACH INSTANCE. BY EMPLOYING INTERNAL AND EXTERNAL EVALUATION METRICS, DETERMINE WHICH CATEGORICAL VARIABLE BEST DESCRIBES THE GROUPS FORMED.

Two clustering algorithms K-Means and DB Scan are deployed against 6 of the categorical variables, Genre, platform, rating, year or release, publisher and developer. The Rating categorical variable best describes the groups formed. Both K-means and DBSCAN for the ratings had high external evaluation scores (V-measure score of 0.995 and 0.052, Rand Index of 0.998 and 0.064, and Mutual Information of 0.995 and 0.535) as well as high internal evaluation scores (Davies-Bouldin Index 1.841 and 1.125, and Silhouette Coefficient of 0.209 and 0.548.

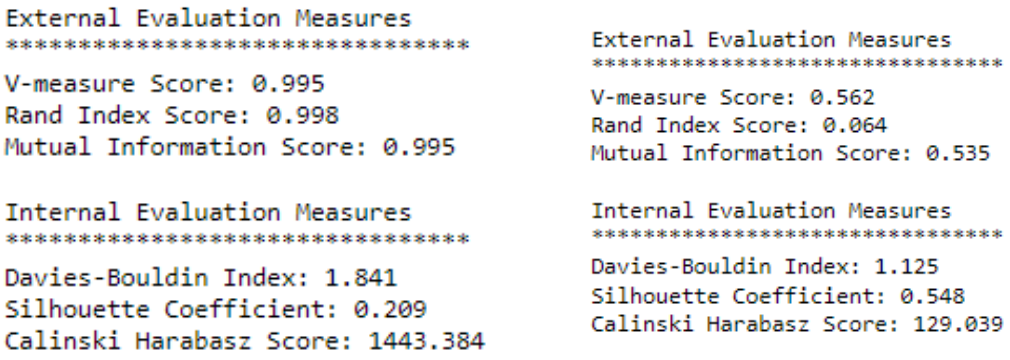

Figure 6: Internal and external evaluation measures for the platform variable using k-means and db scan

CONCLUSION

This analysis provides valuable insights into the videogames market. Video games sales are determined by how well they perform in North America. The market performance of video games is not determined by any one factor outside of regional sales but a combination of multiple factors, some of which aren't included in this study. Number of critics who review video games contribute slightly to the sales performance more than the user scores and reviews. Video games are classified better based on the platforms on which they are released, and while a good model was formed, its inability to sort some classes makes it unsuitable for deployment. Games are better clustered based on their ratings.

The variables considered do not show a strong enough effect on the sales of video games, indicating that more variables like price of video games need to be considered for further studies.

REFERENCES

Statista. (2021). *Number of video gamers worldwide in 2021, by region (in millions).* Retrieved from https://www.statista.com/statistics/293304/number-video-gamers/