Multivariate Analysis in Video Games sales *

Álvaro Novillo *Universidad Carlos III*
Paolo Salvatore Lodato Olano *Universidad Carlos III*

In this article, we perform several dimensionality reduction techniques and clustering algorithms on a video game sales dataset available on Kaggle (https://www.kaggle.com/datasets/gregorut/videogamesales/data). Specifically, we use Principal Component Analysis (PCA) and Multidimensional Scaling (MDS) to reduce the dimensionality of the dataset. We also perform kmeans clustering technique to our dataset to identify sales patterns in different videogame genres. The article discusses the advantages and limitations of each technique and provides insights into the video game market based on the analysis.

Keywords: PCA,MDS,kmeans,Cluster, Videogames, Sales

About the dataset

The dataset under consideration contains information on video games with sales greater than 100,000 copies between 1980 and 2016. The dataset includes 11,493 unique game sales, detailing the name, year of release, genre, platform, and sales figures across numerous regions.

The dataset contains the following fields:

- **Rank** - Ranked by overall sales
- **Name** - Name of each videogame
- **Platform** - The games platform
- **Year** - Year of Release
- **Genre** - Genre of Game
- **Publisher** - Publisher of Game
- **NA_Sales** - Sales in NA (per Million)
- **EU_Sales** - Sales in EU (per Million)
- **JP_Sales** - Sales in JP (per Million)
- **Other_Sales** - Sales in ROW[1] (per Million)
- **Global_Sales** - Total worldwide sales (per Million)

Data Preprocessing

The dataset contains 11 variables, including quantitative variables like sales figures across various regions (NA_Sales, EU_Sales, JP_Sales, Other_Sales, and Global_Sales), the release year, and the

*Replication files are available on the author's Github account (https://github.com/AlvaroNovillo). **Current version**: diciembre 13, 2023; **Corresponding author**: alvanovi@ucm.es.

[1]Net Sales (ROW) means the gross amount billed or invoiced on sales by Company and its Affiliates and Sublicensees of Licensed Products, less the following: (a) customary trade, quantity, or cash discounts and commissions to non-affiliated brokers or agents to the extent actually allowed and taken; (b) amounts repaid or credited by reason of rejection or return; (c) to the extent separately stated on purchase orders, invoices, or other documents of sale, any taxes or other governmental charges levied on the production, sale, transportation, delivery, or use of a Licensed Product which is paid by or on behalf of Company; (d) outbound transportation costs prepaid or allowed and costs of insurance in transit; and (e) allowance for bad debt that is customary and reasonable for the industry and in accordance with generally accepted accounting principles. ("Net Sales (ROW) Definition," n.d.)

rank of the game based on overall sales. Additionally, it includes multi-state categorical variables like the genre, platform, and publisher of the game. To conform with the desired format, which requires at least two binary variables, we will filter out the video games of recent years and focus on titles that we are already acquainted with. Moreover, we will limit our research to two primary platforms, namely, Xbox One and PS4.

Table 1: Top five videogames, according to the sales ranking, that we are going to work with

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
34	34	Call of Duty: Black Ops 3	PS4	2015	Shooter	Activision	5.77	5.81	0.35	2.31	14.24
78	78	FIFA 16	PS4	2015	Sports	Electronic Arts	1.11	6.06	0.06	1.26	8.49
93	93	Star Wars Battlefront (2015)	PS4	2015	Shooter	Electronic Arts	2.93	3.29	0.22	1.23	7.67
102	102	Call of Duty: Black Ops 3	XOne	2015	Shooter	Activision	4.52	2.09	0.01	0.67	7.30
110	110	Fallout 4	PS4	2015	Role-Playing	Bethesda Softworks	2.47	3.15	0.24	1.10	6.96
222	222	FIFA 17	PS4	2016	Sports	Electronic Arts	0.28	3.75	0.06	0.69	4.77

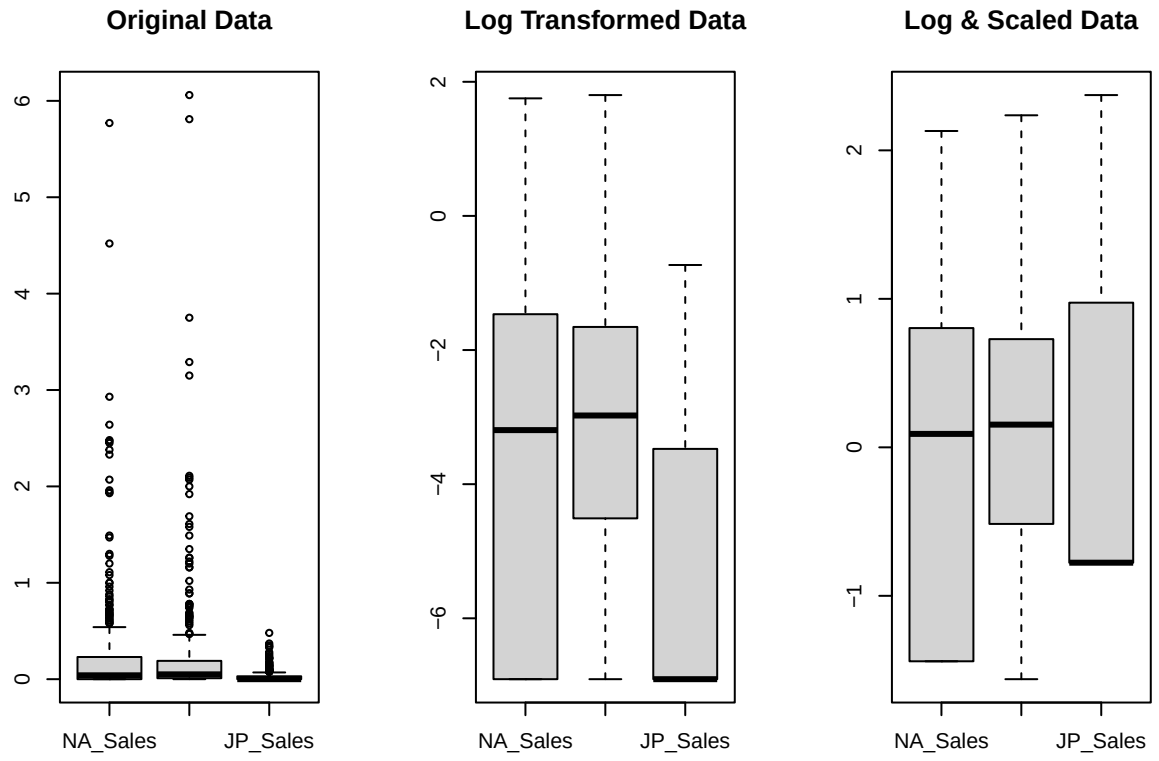

Figure 1: Boxplots of the different transformations applyied to the sales data. From left to right, the original data, the log-transformed data and the log transformed and scaled data

We recall from the previous part of the analysis that the Rank variable is right-skewed. For positive skewness, we can apply a logarithmic transformation[2]. Many statistical methods, includ-

[2]Since the variables related to sales in our dataset have zeros, we have used the Box-Cox technique (see Sakia (1992)) to identify the appropriate transformation for our case (obtaining $\lambda \approx 0$), which leads to the application of the $log(x+\epsilon)$ transformation, where $\epsilon > 0$ is an arbitrary small constant

ing linear regression and analysis of variance, assume that the residuals are normally distributed. Normalising our data would also be useful to achieve zero mean and unit variance. In Figure 1 we can visualise these transformations performed on the sales variables.

PCA

After an initial exploration and necessary pre-processing of the dataset, we applied Principal Component Analysis to our sales data, reducing the dimensionality of the problem to two dimensions. The results are shown in Fig 2, where we can see that the games on the left of the plot are those with the highest sales volumes, while those on the top are mainly games that are predominantly sold in Japan. Shooter games (shown as blue crosses) are the best-selling games worldwide, with some examples proving particularly profitable on the Japanese market, as indicated by the points in the upper left quadrant. As highlighted in the previous part of the analysis, certain action games (represented by yellow triangles) and role-playing games (represented by pink squares) have a predominant presence in the Japanese market, with some being marketed exclusively in this region (notably in the top right quadrant of the graph).

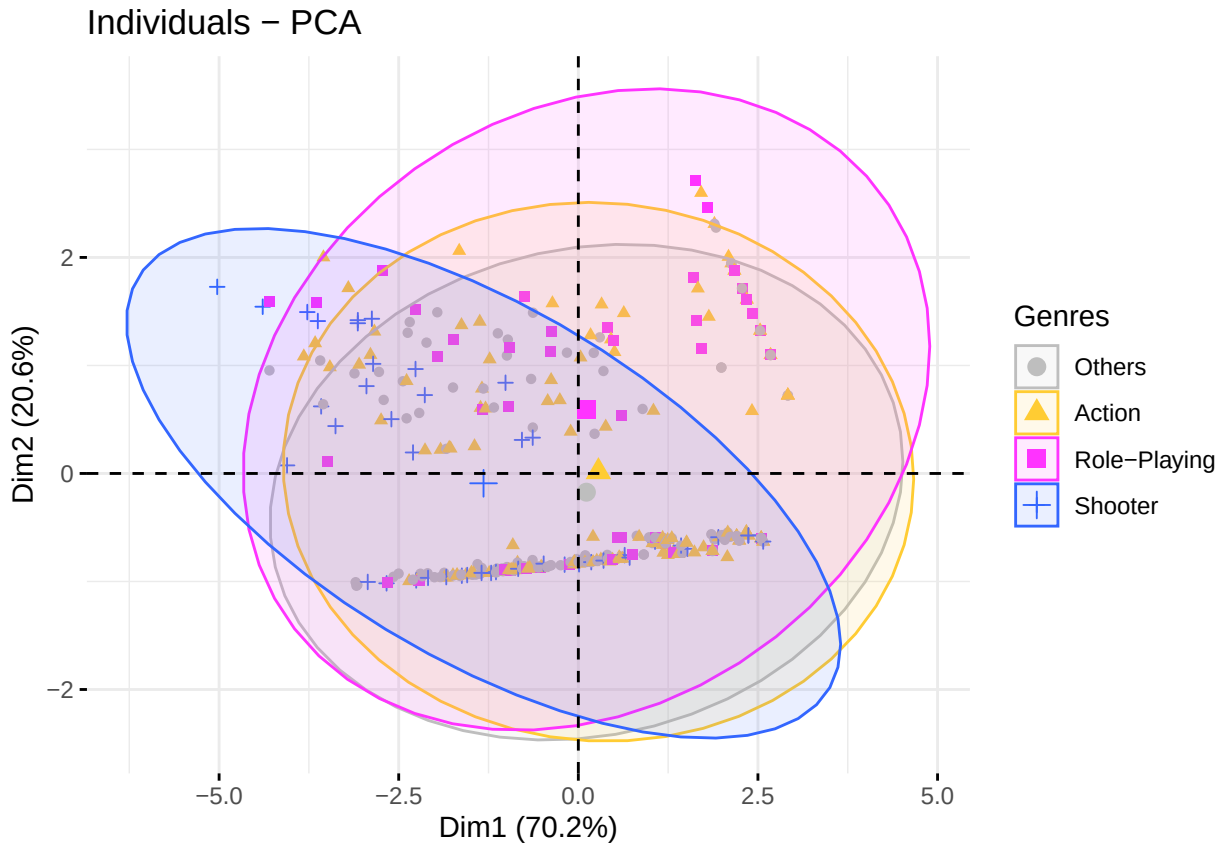

Figure 2: Principal Component Analysis (PCA) plot displaying the top selling game genres' distribution with respect to the two principal components. Points in yellow corresponds to Action games, those in pink to Role-Playing games, those in blue to Shooters, and the rest of the genres are visualized in grey.

Multidimensional scaling (MDS)

Multidimensional Scaling (MDS) is another dimensionality reduction technique that visualizes the pairwise dissimilarity or similarity between data points. It is particularly useful for datasets containing both qualitative and quantitative data, as MDS can handle various types of input distances, including those based on categorical variables.

Unlike Principal Component Analysis (PCA), which works well with quantitative variables, MDS is versatile and applicable to mixed datasets. MDS results provide a low-dimensional representation that preserves the original dissimilarity or similarity structure, making it valuable for revealing patterns, clustering, and interpreting relationships in diverse datasets with a mix of variable types.

The `dist` function in R, by default, uses Euclidean distance for numerical variables. However, when categorical variables are present, dist converts them into binary indicators (dummy variables) and calculates the dissimilarity based on these indicators. The Jaccard distance, which measures the dissimilarity between two sets, is commonly used for binary data.

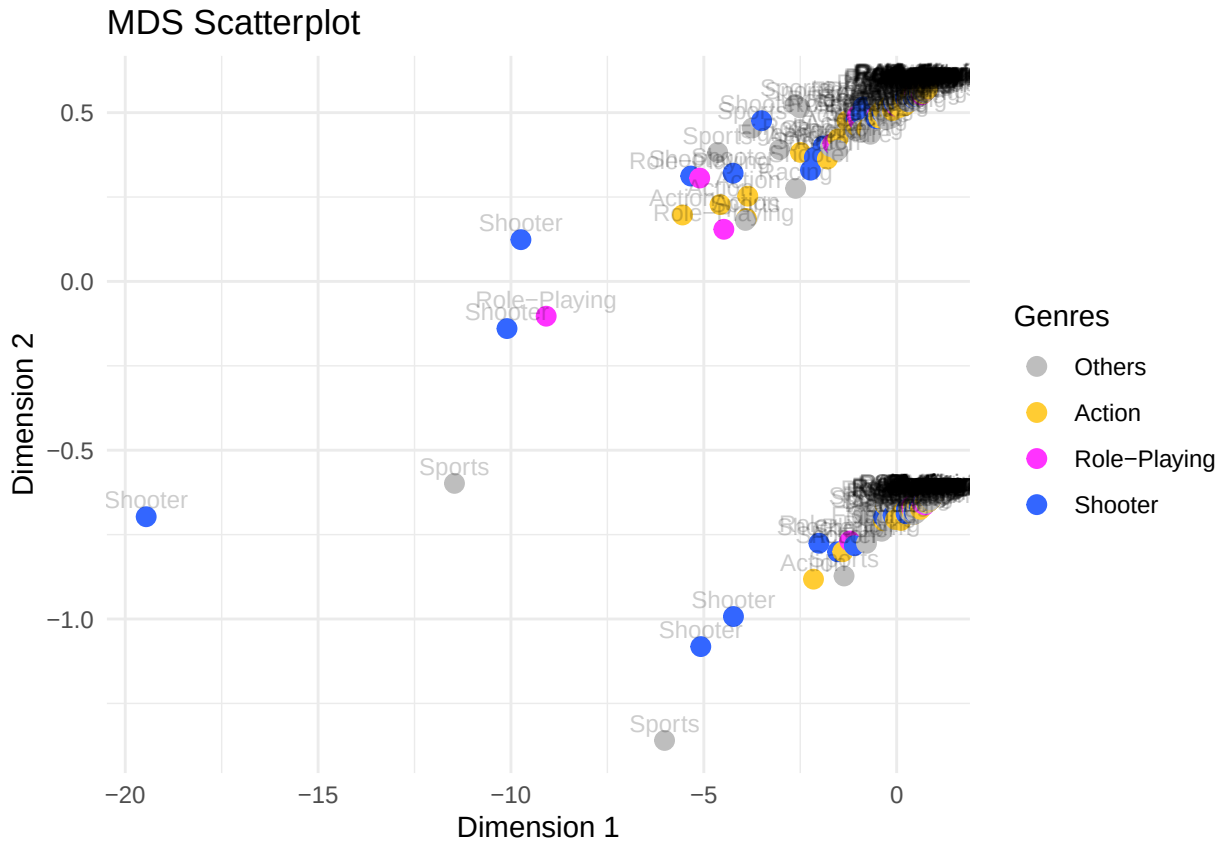

Figure 3: Biplot showing MDS performed using the Euclidean distance

Nevertheless, Euclidean distance is not inherently suitable for categorical variables. The Euclidean distance metric assumes a continuous numerical scale and relies on the notion of geometric distances in a continuous space. Categorical variables, on the other hand, represent distinct categories without a natural ordering or continuous progression.

Alternatively, Gower's distance is designed to handle mixed data types and offers more flex-

ibility in handling both numerical and categorical variables. The `daisy` function in the `cluster` package is often used for this purpose.

We choose to use Gower's distance over Euclidean distance in the context of mixed data because Gower distance is specifically designed to handle datasets that include a combination of numerical, categorical, and ordinal variables. When working with diverse types of variables, such as continuous measurements, categorical labels, or ordinal rankings, traditional distance measures like Euclidean may not be appropriate due to their assumptions about data types.

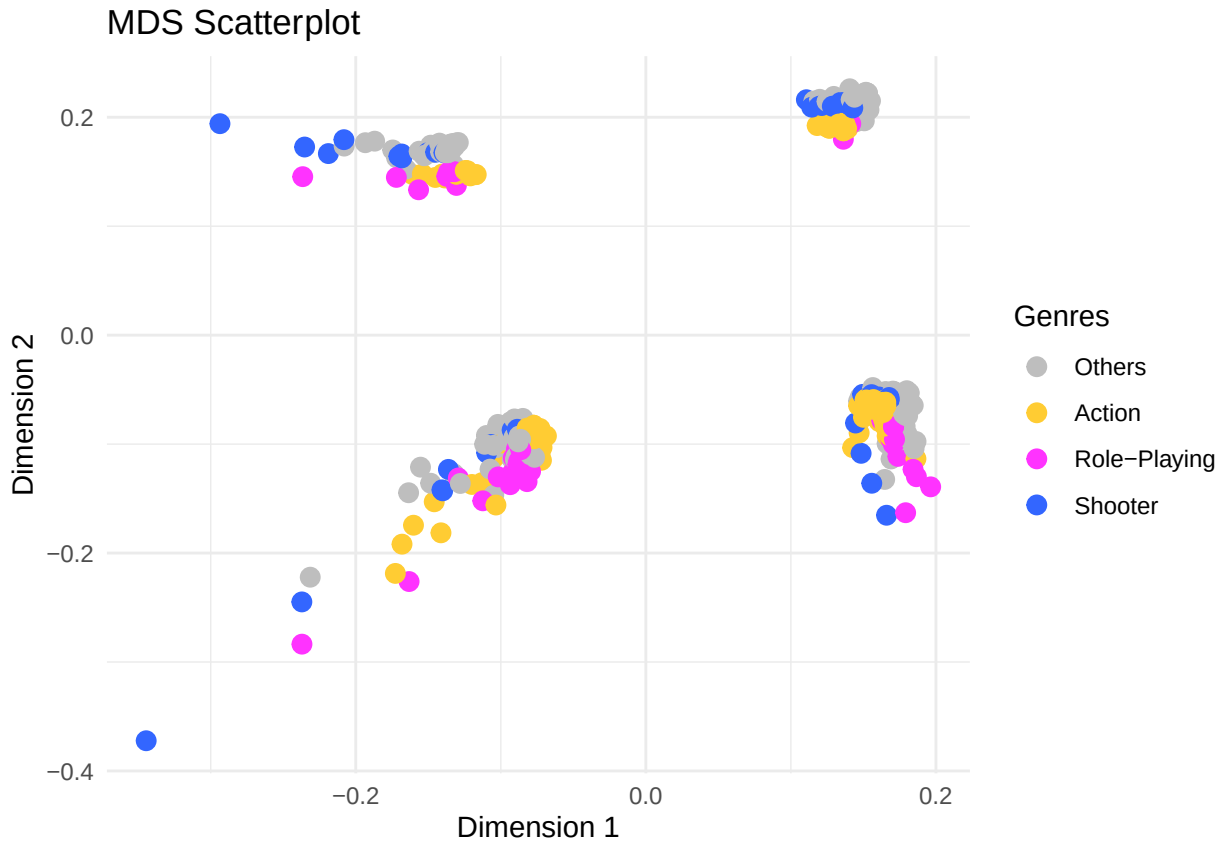

Figure 4: Biplot showing MDS performed using the Gower's distance

Gower distance provides balanced treatment for categorical variables, ensuring dissimilarity is computed based on shared categories. These attributes make Gower distance a robust and flexible option for dissimilarity measurement, particularly in real-world scenarios with heterogeneous data types.

The Gower distance $d_G(x, y)$ between two data points x and y is computed as:

$$d_G(x, y) = \frac{\sum_{i=1}^{n} w_i \cdot s_i(x, y)}{\sum_{i=1}^{n} w_i}$$

where $s_i(x, y)$ represents the dissimilarity measure for each variable, w_i denotes the weight assigned to each variable, and n is the total number of variables. This formula accommodates different variable types and scales, providing a comprehensive dissimilarity metric for mixed datasets.

The variability explained by using Gower's distance is 0.5970888 while originally using the default distance we got a result of 0.9586581. Although the explained variability is lower using

Gower's distance, we know it's appropriate to use it either way due to the nature of our data, which is mixed.

In the heatmap displayed below, we aim to elucidate the intricate relationship between the original quantitative variables and the MDS dimensions. The color intensity in the heatmap reflects the correlation or loadings, offering insights into how each variable contributes to the different dimensions.

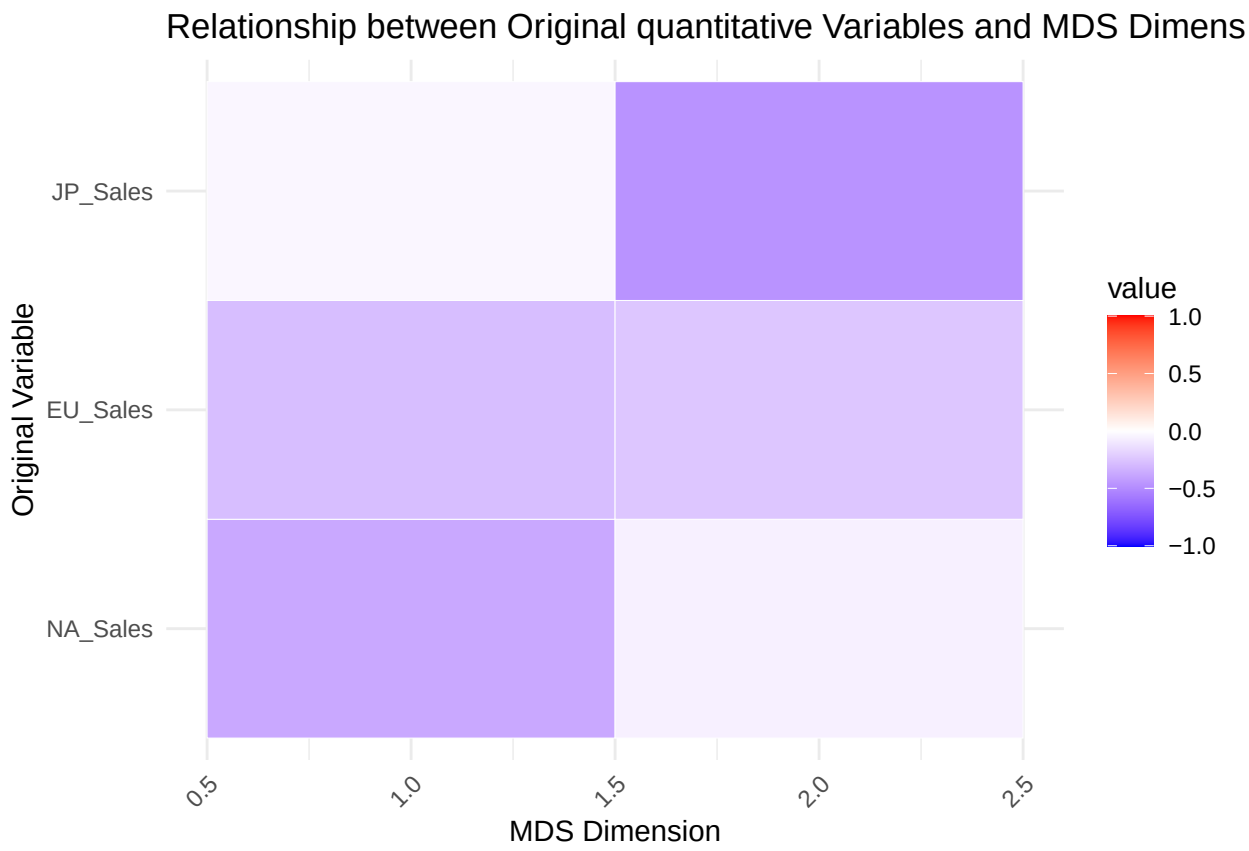

Figure 5: Correlation matrix between the quantitative variables and the MDS dimensions

As we know, in the calculation of Grower's distance there is more weight on the categorical values rather than the quantitative, which is one of its main drawbacks. Nevertheless, we can see the first dimension being more related to how well the game sold in western cultures, while the second dimension has more correlation with Japan sales.

Cluster analysis

After the initial exploratory data analysis, performing Principal Component Analysis and MDS, we could delve deeper in our analysis including *cluster analysis*.

For cluster analysis, there are various methods you can apply to group similar instances together. Common techniques include K-means clustering, hierarchical clustering, and density-based clustering like DBSCAN. In our case, it is convenient to apply *K-means clustering* given the easy application of the algorithm, and the scaling and data pre-processing applied.

K-means clustering is an unsupervised machine learning algorithm that aims to partition n

data points into k clusters. The algorithm works by minimizing the sum of squared distances between the data points and their respective cluster centroids.

The process involves the following steps:

- 1. Randomly select k data points as the initial cluster centroids.
- 2. Assign each data point x_i to the nearest cluster centroid μ_j based on the Euclidean distance.

$$d\left(x_i, \mu_j\right) = \sqrt{\sum_{n=1}^{N} \left(x_{i,n} - \mu_{j,n}\right)^2}$$

  where N is the number of dimensions/features 1 .
- 3. The objective of K-means is to minimize the sum of squared distances within each cluster, which can be expressed as:

$$WSS = \sum_{j=1}^{k} \sum_{x_i \in S_j} \left\| x_i - \mu_j \right\|^2$$

  where S_j is the set of data points in cluster j, μ_j is the centroid of cluster j, and k is the total number of clusters 1 .
- 4. The cluster centroids are updated by taking the mean of all data points assigned to that cluster:

$$\mu_j = \frac{1}{\left|S_j\right|} \sum_{x_i \in S_j} x_i$$

  where $\left|S_j\right|$ is the number of data points in cluster j.

To select the number of clusters, we will use the *elbow* method, which consists of running the algorithm with a varying k and calculating a cost function for each run. Then the cost values are plotted against k values and we choose k at the turning point (called "elbow").

This algorithm can be applied in R using the *cluster* library as follows:

```
library(cluster)
#Prepare PC to perform kmeans
pca_comp = as.data.frame(pca_result$x[,1:2])
# Initialize empty vector to store within-cluster sum of squares
wss <- vector()

# Vary the number of clusters from 1 to 10 and compute the total within-cluster sum
 ↪  of squares
for (i in 1:10) {
  kmeans_model <- kmeans(pca_comp,centers = i,nstart = 10)
  wss[i] <- kmeans_model$tot.withinss
}
```

Note that we have applied the kmeans algorithm to the previously found principal components. This ensures that the implementation of the Euclidean distance is appropriate as the input variables are uncorrelated (orthogonal) numerical values.

Fig. 6 contains the elbow plot of the model created above. Given the results, we consider that using $K = 4$ clusters will be the best option to facilitate the interpretation of the results. We will later see that this is both the most convinient choice, and the most adequate when performing kmeans using MDS Dimensions.

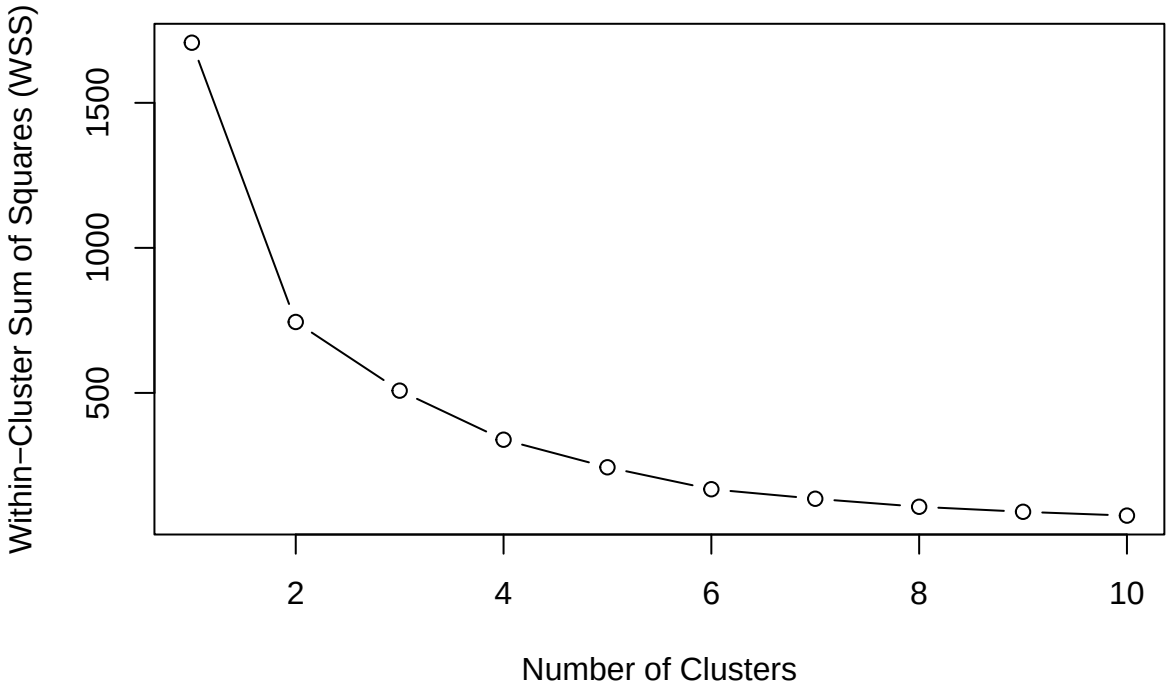

Figure 6: Elbow plot for the K-means clustering algorithm using principal components

Fig. 7 visualizes the clusters found with respect with the previously found Principal Components.

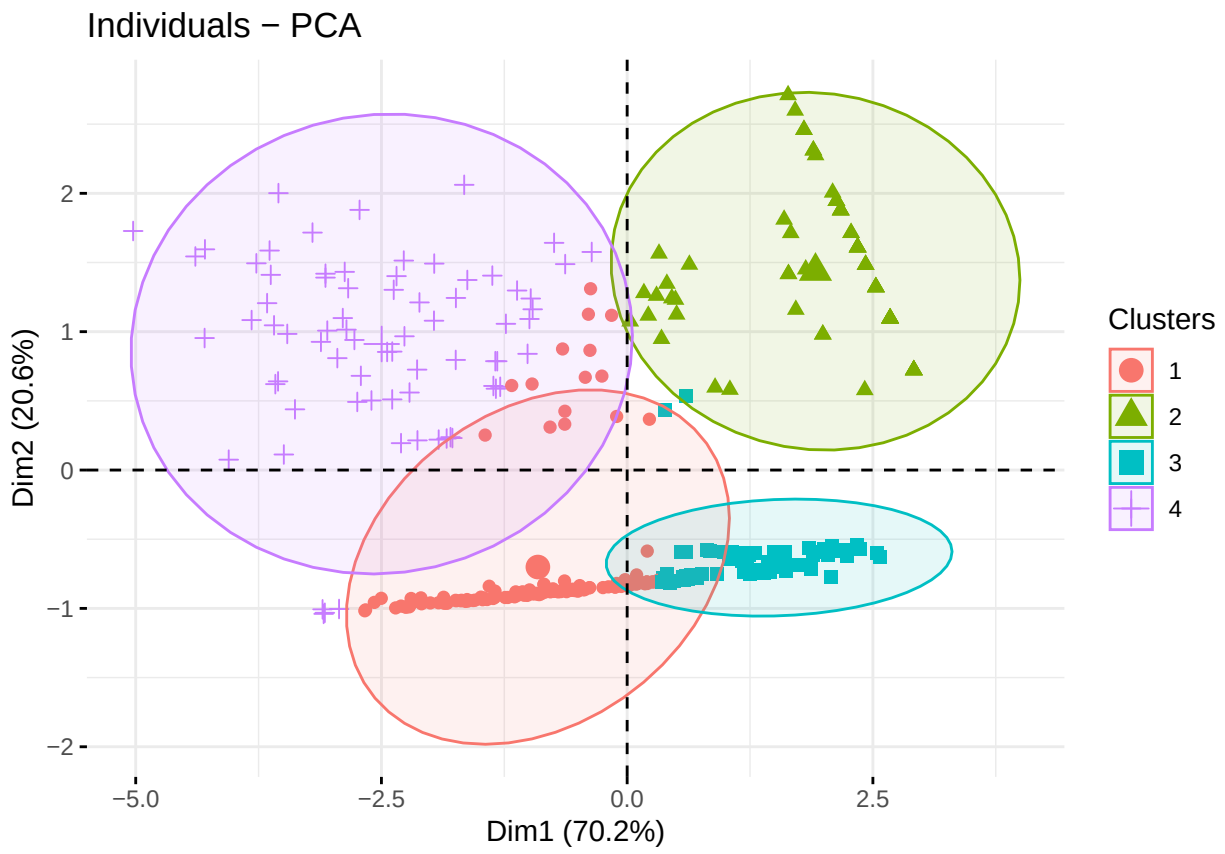

Figure 7: Visualization of the clusters with respect to the Principal Components. Points with a higher size corresponds to the cluster centers of each group

Based on Figure 7 and Table 2, we can derive an interpretation for the clusters found. The cluster labelled (1) contains those games with moderate sales volumes that are mainly sold outside Japan. The second cluster (2) contains games that are mainly sold in Japan. The third cluster (3) contains the games with the lowest sales volume worldwide, and finally cluster (4) represents the games with the highest sales volume worldwide.

Table 2: Descriptive statistics at the cluster level

cluster	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
1	0.574	0.463	-0.596	0.492	0.429
2	-1.157	-1.258	1.147	-1.049	-0.661
3	-0.662	-0.564	-0.755	-0.756	-0.952
4	1.048	1.179	1.174	1.237	1.303

By inspecting the genre distribution in each cluster, Fig. 8, and based on the aforementioned classification, we can have a visual representation of which is the most sold genre.

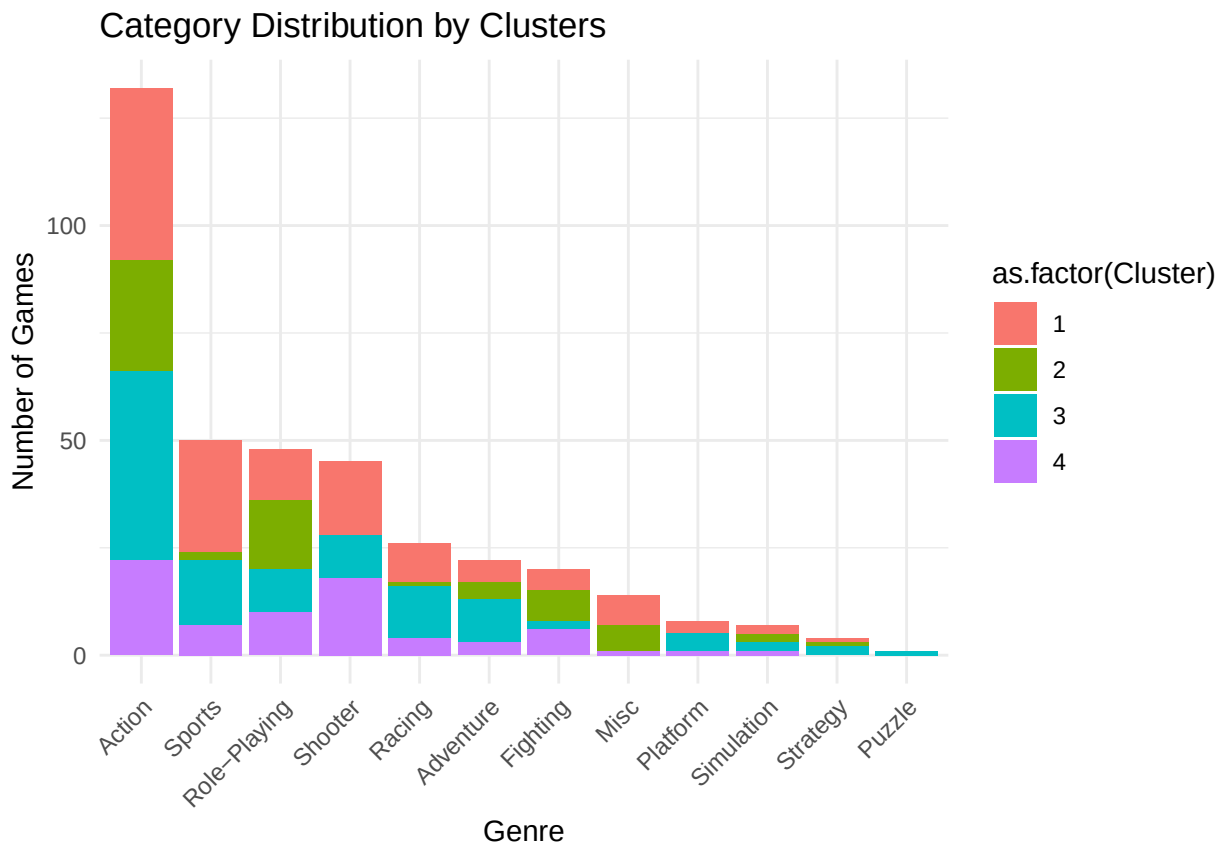

Figure 8: Genre distribution in each cluster.

As mentioned earlier, shooter and action games have the highest sales, given their predominance in cluster (4). We can also see how some of the action and RPG games in our dataset are mainly sold in Japan, given their presence in cluster (2).

As we did with the principal components, we can also apply kmeans clustering algorithm using the obtained dimensions after applying MDS to the dataset.

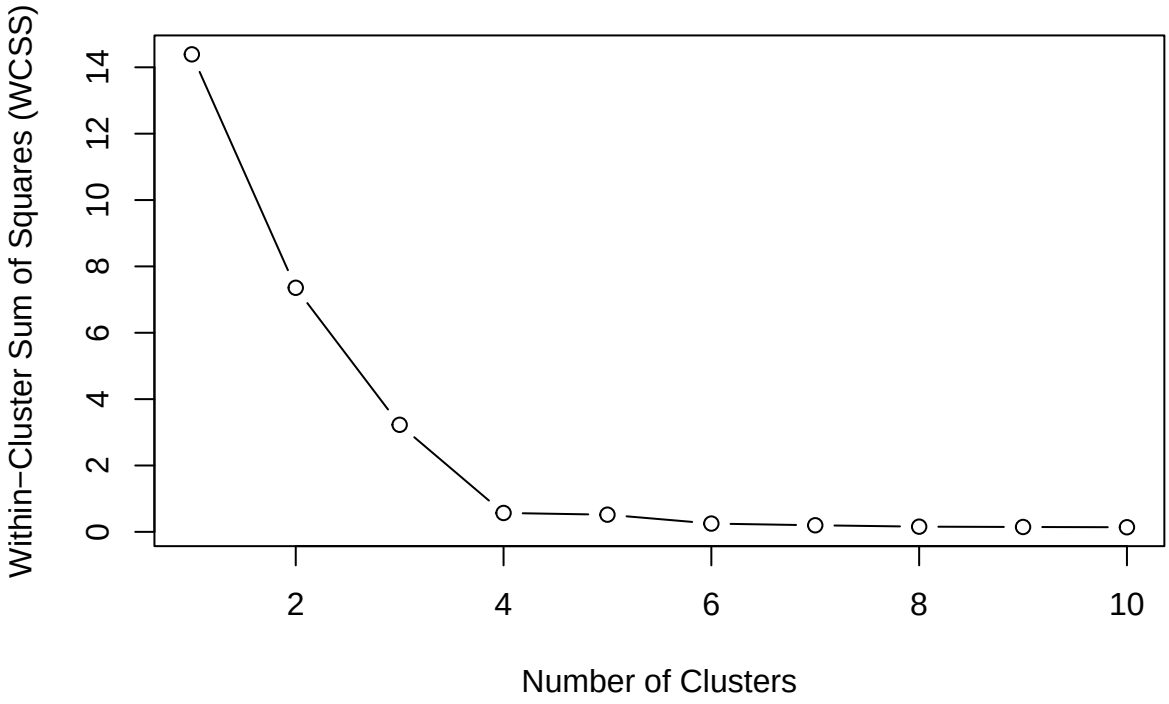

Figure 9: Elbow Plot for K-means Clustering using MDS Dimensions

Using MDS Dimensions to perform kmeans reveals that $k = 4$ cluster division is by far the most appropriate in our case, Fig. 9.

Finally, we can visualize the clusters with respect to the first two dimensions found performing MDS using Gower's distance

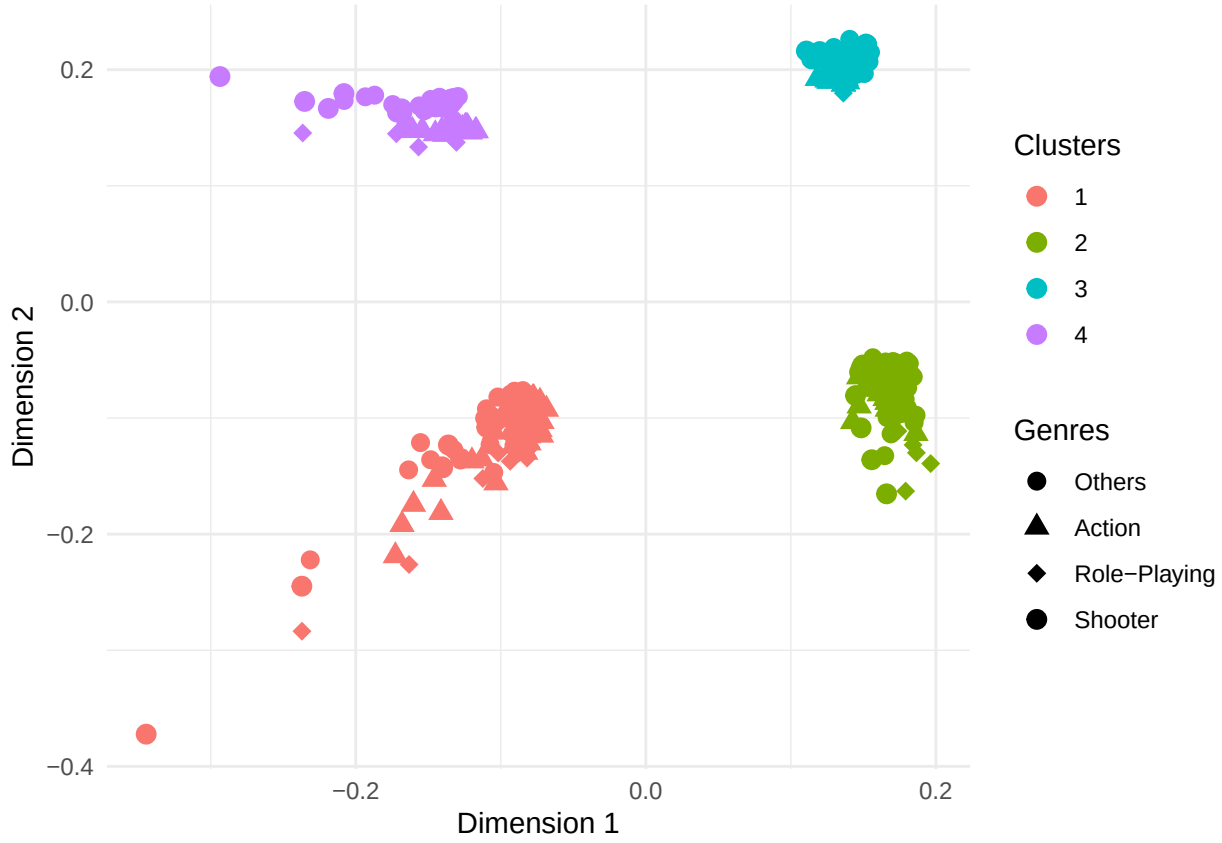

Figure 10: MDS Scatterplot with K-means Clusters

AÑADIR INTERPRETACIÓN

References

"Net Sales (ROW) Definition." n.d. https://www.lawinsider.com/dictionary/net-sales-row.

Sakia, R. M. 1992. "The Box-Cox Transformation Technique: A Review." *Journal of the Royal Statistical Society. Series D (The Statistician)* 41 (2): 169–78. http://www.jstor.org/stable/2348250.