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. The article discusses the advantages and limitations of each technique and provides insights into the video game market based on the analysis.

Keywords: PCA, 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 rank of the game based on overall sales. Additionally, it includes multi-state categorical variables

*Replication files are available on the author's Github account (https://github.com/AlvaroNovillo). **Current version**: diciembre 12, 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.)

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

In Table 1. the top five selling games for 2015 and 2016, in PS4 ans Xbox One are shown. As we can see, the first one, which is Call Of Duty: Black Ops 3 is among the top 50 best selling games of the dataset (in PS4).

Examining the distribution of the filtered games rank, as seen in Fig. 1, considering its skewness, it can be confirmed that the vast majority of games released during this time period did not have a significant impact on the industry. In fact, the average ranking of games within our dataset stands at 9373.

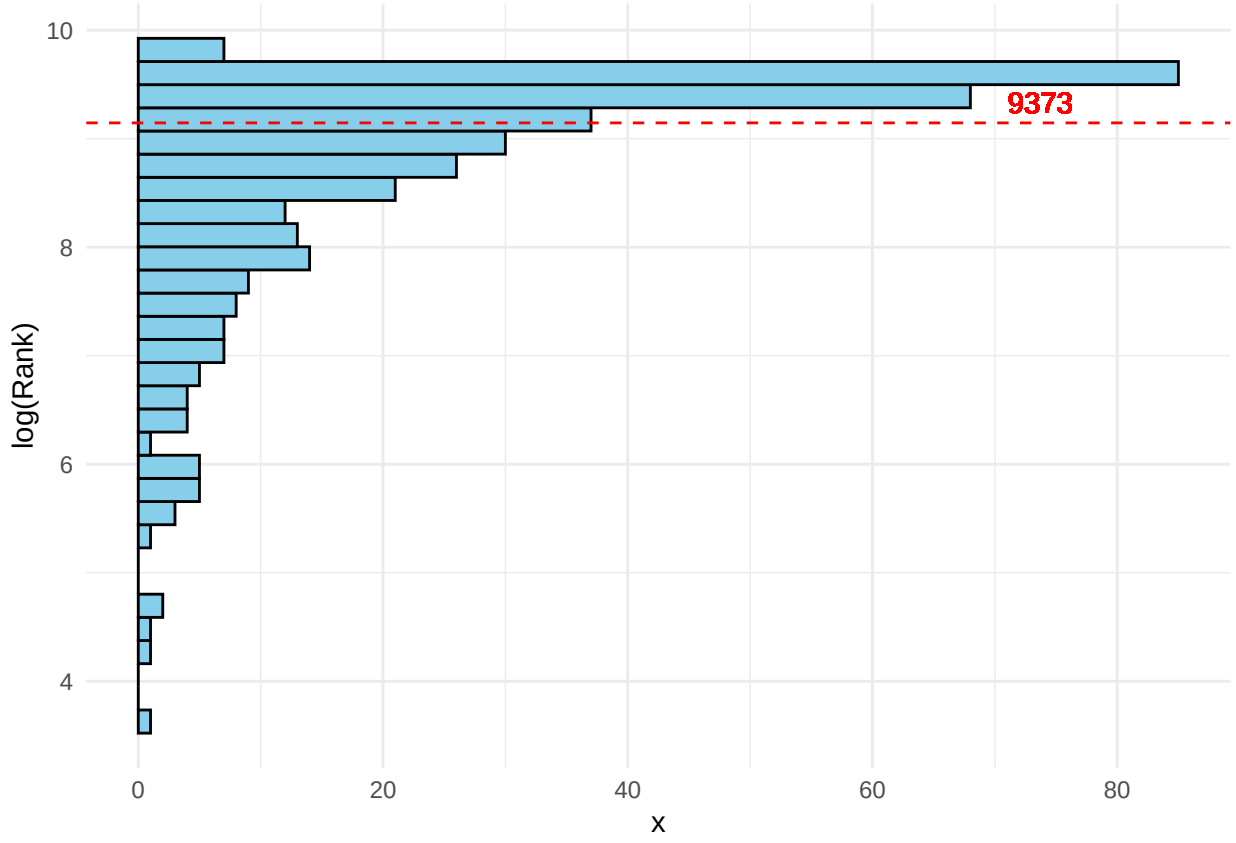


Figure 1: Distribution of the log-transformed Rank values. The red dashed line represents the median of the distribution

Figs. 2 and 3 allow us to explore the basic features of our dataset, informing us of the amount of games from each platform, and the amount of games of each genre. In our dataset, the mayority of the sold games are from PS4, and the most popular genre is Action, followed by Sports, Role-Playing and Shooter

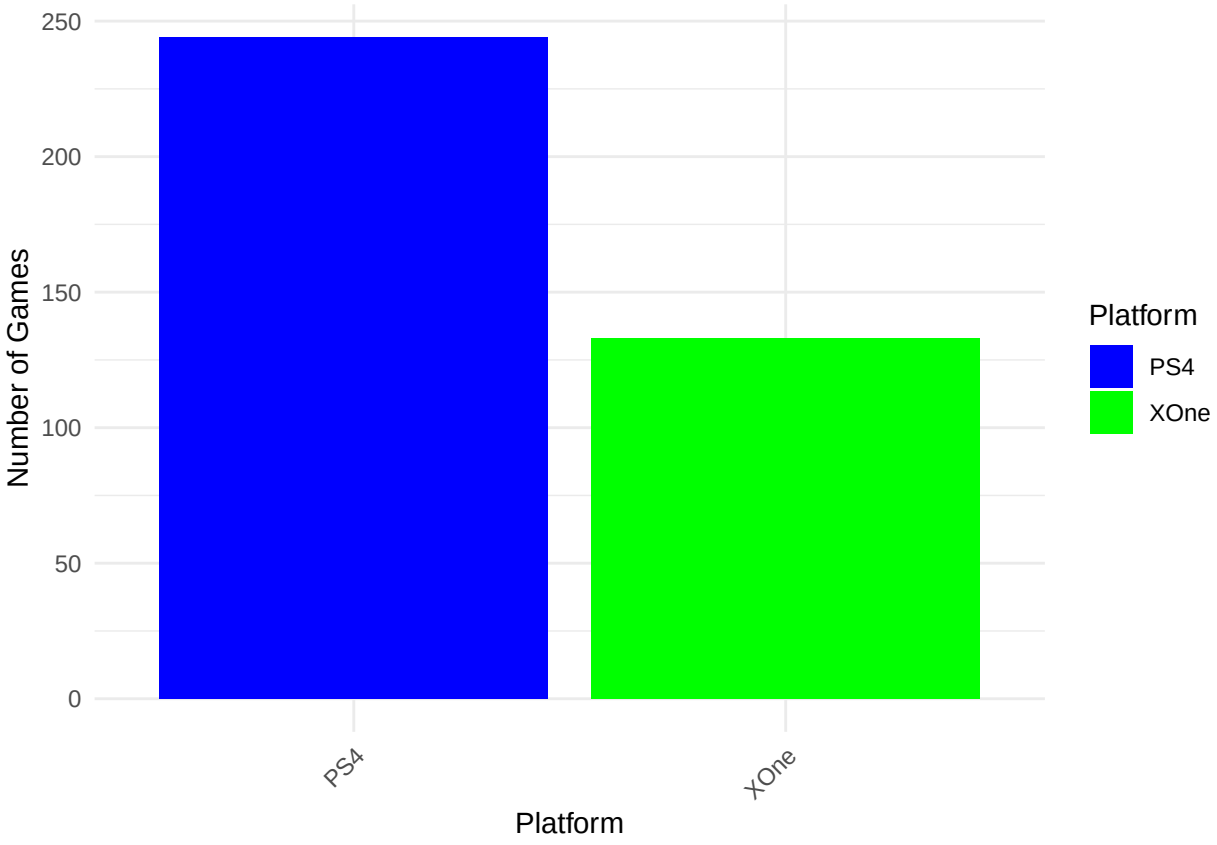


Figure 2: Number of games of each platform inside the dataset

From the previous plot we can see the Rank variable is right skewed. For positive skewness we can apply a log transformation[2], which would help to normalize the distribution of the variables. Many statistical methods, including linear regression and analysis of variance, assume that the residuals are normally distributed. Normalizing our data would also be useful to help achieve zero mean and unit variance. In Fig. 4, we can visualize these transformations performed on the sales variables.

Since the ranking is solely determined by overall sales figures, it is worthwhile investigating whether the top-selling game in certain regions differs from that of others. Our expectation is that the best-selling games in Japan will differ from those sold in the West. To do such analysis, we will start by computing the *correlation matrix* of the sales in the different regions.

The *correlation matrix* can be computed as follows: $\text{Cor}(X, Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$, where $\text{Cov}(X, Y)$ is the covariance between variables X and Y, and σ_X and σ_Y are their respective standard deviations. The correlation matrix provides a comprehensive view of the linear relationships between vari-

[2]Since the variables referring to the sales in our dataset presents zeros, we have applied Box-Cox technique (See Sakia (1992)) to identify the appropriate transformation for our case (obtaining $\lambda \approx 0$), leading to the application of $log(x + \epsilon)$ transformation, being $\epsilon > 0$ an arbitrary small constant.

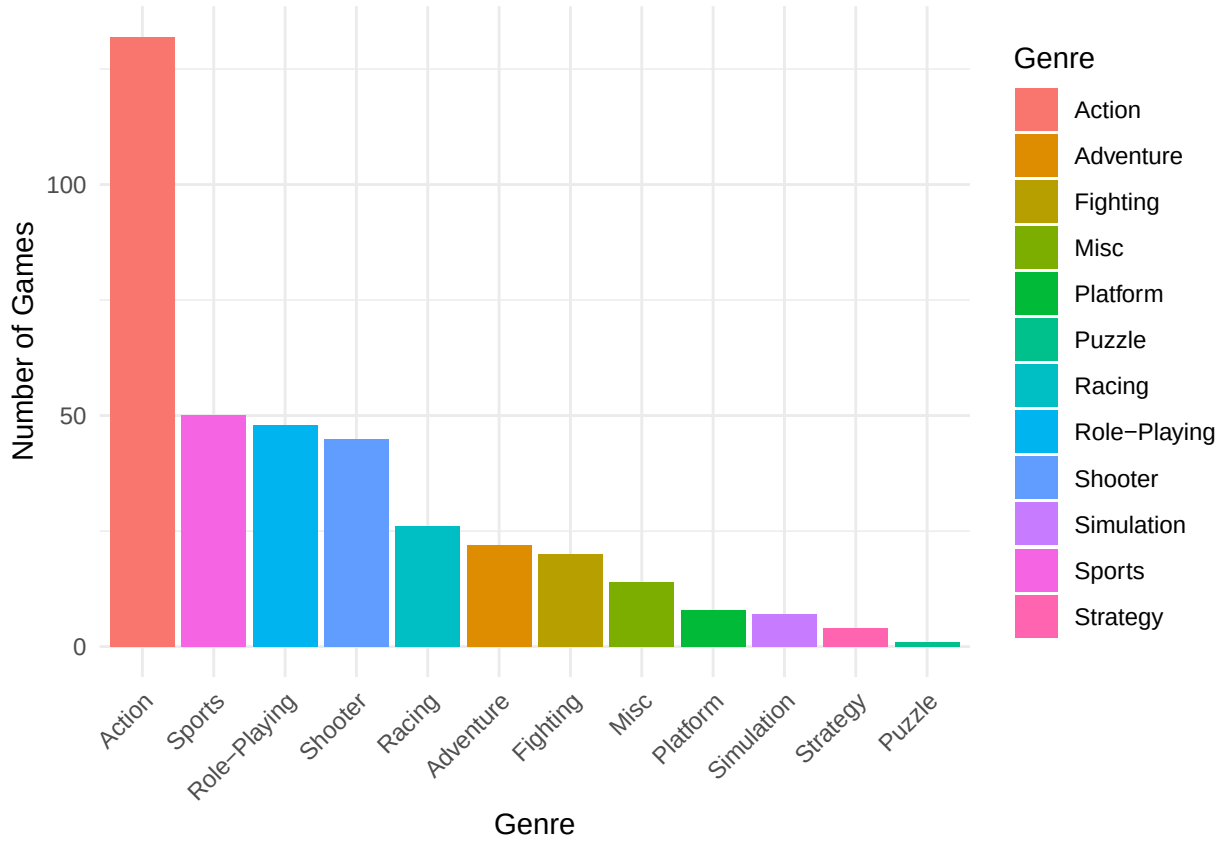


Figure 3: Barplot of the amount of games of each genre

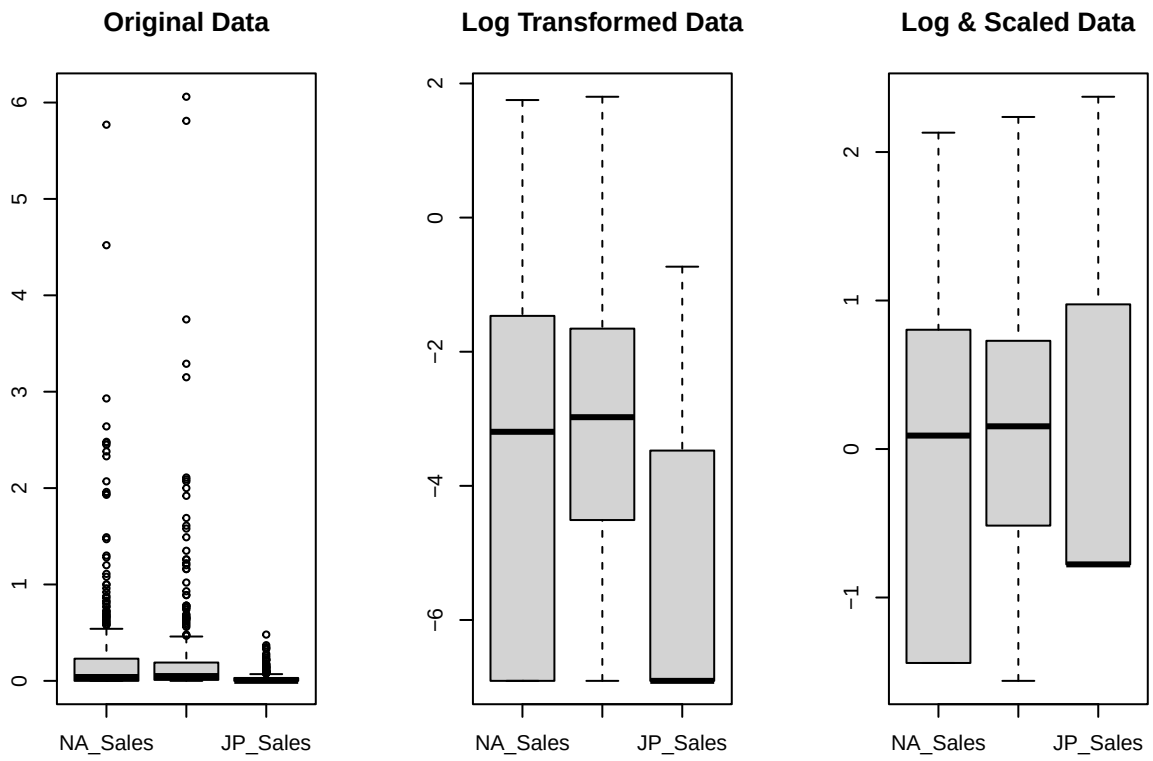


Figure 4: Boxplots of the different transformations applyied to the sales data. From left to right, the original data, the log-transformed data and the log transformed and scaled data

Table 2: Covariance matrix of sales

	NA_Sales	EU_Sales	JP_Sales
NA_Sales	1.00	0.690	0.010
EU_Sales	0.69	1.000	0.044
JP_Sales	0.01	0.044	1.000

ables in a dataset. It is useful for identifying patterns, understanding dependencies, and detecting multicollinearity in statistical analyses.

Doing so (Table 2), we can see that the correlation between the sales in Japan and the Western market is 0.044, with an even lower correlation of 0.01 with the American market, as illustrated in Fig. 5. This raises compelling questions about the underlying factors that contribute to these correlations. It is clear that several key factors highlight the significant differences between the Oriental and Western video game industries, leading to this low correlation.

First and foremost, the contrast in gaming preferences between regions plays a key role. As seen in the intercorrelation measurements, there is some correlation in the sales in Europe and NA. In the West, specifically in North America, action and shooter games are incredibly popular. However, the Japanese market favours Role Playing Games (RPGs), which differs greatly from the Western market. As a result of these diverging gaming genres, differing sales patterns naturally occur, ultimately contributing to the observed low correlation.

The marketing and localization strategies employed in the Japanese video game industry hold immense importance. Many Japanese games prioritize the local market, leading to gameplay and cultural elements that might not strongly resonate with Western or North American audiences. Consequently, these games might struggle to achieve success beyond their intended Eastern audience, resulting in a larger sales gap and weaker association with these regions. In contrast, sales in North America and Europe tend to be more closely linked due to their shared Western cultures and comparable marketing approaches. Conversely, Japan represents a distinct market, with its citizens' preferences significantly differing from those of Western cultures.

It is also worth noting that given the nature of the variables, they all present a positive correlation. To ensure the low *intercorrelation* between our sales data, we have computed different correlation measures[3], presented in Table 3.

Table 3: Correlation measurements of the sales in the different regions

	q1	q2	q3	q4	q5	q6
q	0.365	0.378	0.277	0.424	0.581	0.319

As noted earlier when examining the correlation between pairs of markets, the intercorrelation between the three markets is low because of the aforementioned socio-cultural factors.

To delve deeper into the differences in the market, Table 4 presents a comprehensive analysis of the percentage distribution of sales across the top three genres within diverse regions under investigation. It is evident from the table that the Role-Playing Games (RPGs) enjoys significantly greater popularity in Japan as compared to North America and Europe. Strikingly, our research

[3]The computed metrics are the ones that have been sugested in class. See Grané (n.d.)

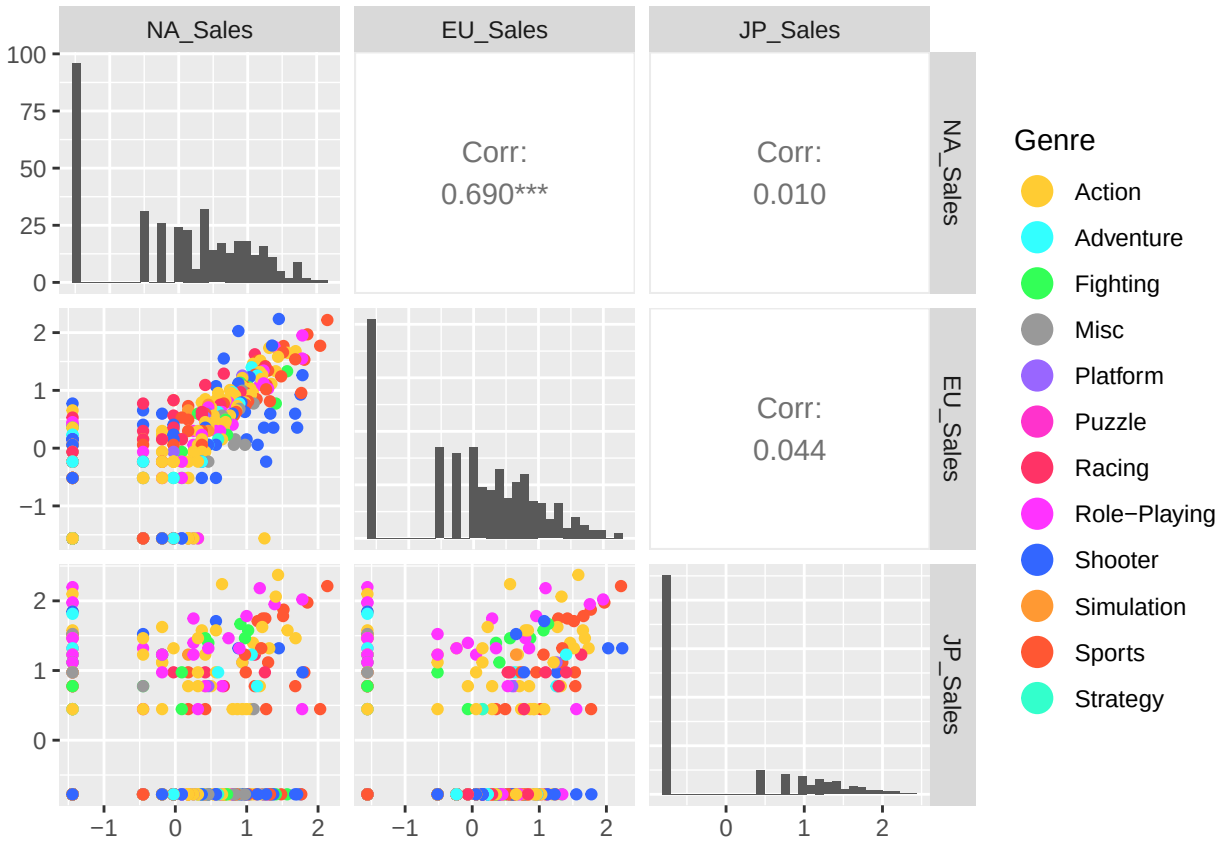


Figure 5: Correlation Plot of Video Game Sales in Different Regions

reveals that Action games emerge as the most prevalent genre in Japan, accounting for a substantial portion of the region's total sales, encompassing 35.28% of the market share. This results can also be found in Fig. 5, where we can see almost all Role-playing and Action games are above the diagonal line (y = x) in the last row plots (where Japanese sales is the y axis), indicating the higher popularity of this genres in the Japanese market

Table 4: Percentage distribution of sales for the top three genres in different regions

Genre	Percentage_NA_Sales	Percentage_EU_Sales	Percentage_JP_Sales
Action	20.79	23.19	35.28
Role-Playing	11.75	11.81	27.26
Shooter	35.69	29.99	15.38

Principal Component Analysis (PCA)

After an initial exploration and necesary preprocessing of the dataset, we are now ready to conduct Principal Component Analysis (PCA) to reduce the problem's dimensionality.

Principal Component Analysis (PCA) constitutes a vital stage in our data analysis pipeline for several reasons. It enables us to reduce the dimensionality of our dataset by transforming the initial variables into a set of independent principal components, which capture the essential information with fewer variables. This reduction becomes particularly valuable when dealing with datasets of high dimensionality, as it allows for more manageable and comprehensible analyses.

In our Principal Component Analysis, all avaliable sales variables will be taken into account. Thus, beside from the sales of the three principal regions at study, we will also consider the Sales in ROW[4], and the Global sales (per Million)

Table 5: Summary of PCA

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.873	1.016	0.559	0.294	0.246
Proportion of Variance	0.702	0.206	0.063	0.017	0.012
Cumulative Proportion	0.702	0.908	0.971	0.988	1.000

As seen by the results (Table 5), 90.8% of the total variability in the dataset has been explained by the two first components. Thus, these two variables can accurately represent the data. By using the package `factoextra` we can create a Scree plot (Fig. 6) that visualizes our decision on the number of components or factors retained in the analysis.

By examining the eigenvector that corresponds to the chosen variables, in this instance, the two highest eigenvalues' eigenvectors, we can provide an interpretation and significance to the selected variables.

The first principal component **PC1** seems to be influenced by each variable, having higher (negative) loadings for NA_Sales, EU_Sales, Other_Sales, and Global_Sales. This component highlights that these five criteria simultaneously fluctuate. Thus, if sales in a particular region

[4]See "Net Sales (ROW) Definition" (n.d.)

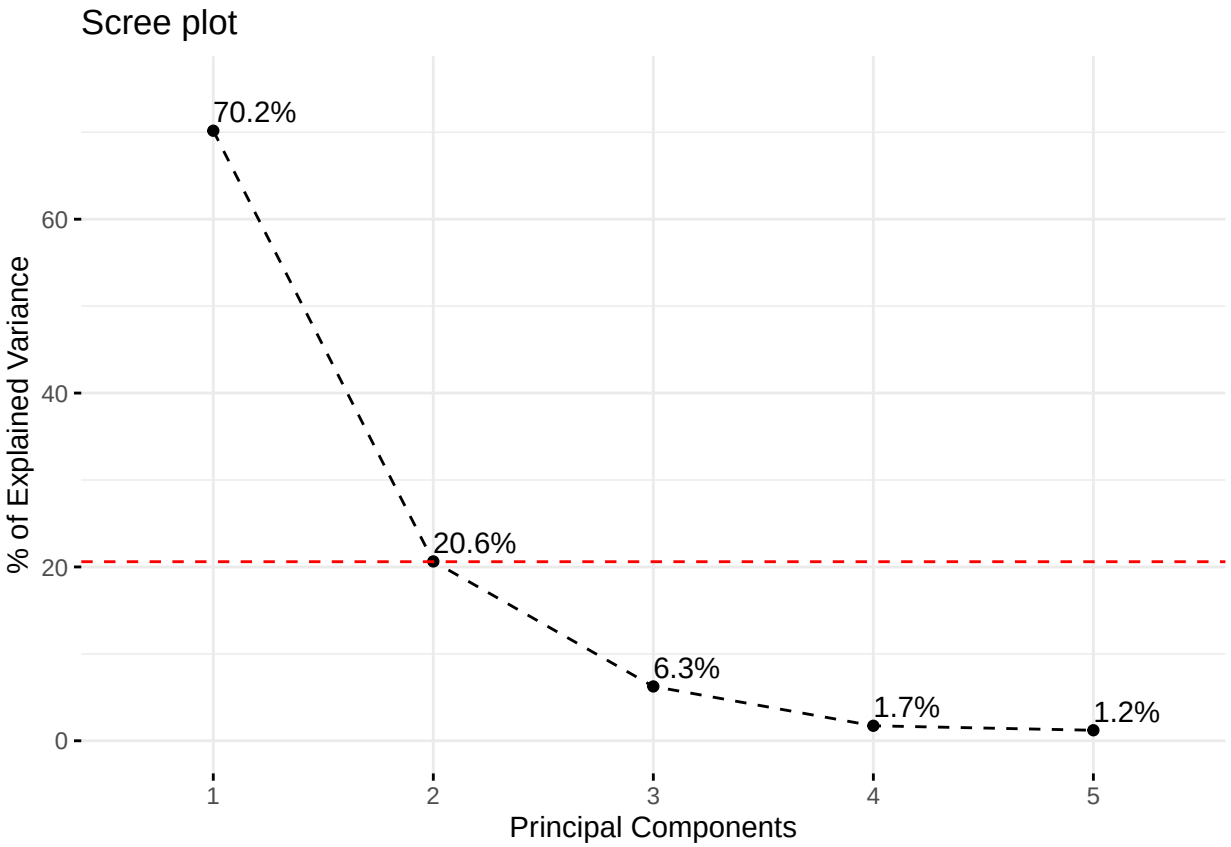


Figure 6: Scree plot visualizing the number of components retained in the analysis. The red dashed line represents the cutoff selected

Table 6: Variables contribution to the first two principal components

	PC1	PC2
NA_Sales	-0.48	-0.18
EU_Sales	-0.48	-0.14
JP_Sales	-0.10	0.96
Other_Sales	-0.52	-0.05
Global_Sales	-0.51	0.14

increase, they tend to increase in the other regions, thereby increasing the Global Sales and Sales in the ROW. Sales in Japan also increase, but at a lower rate, due to the previously mentioned low correlation of the Japanese market with respect to the others. It can be understood as a measure of the total amount of sales

The second principal component **PC2** seems to represent a pattern primarily related to sales in Japan, distinguishing it from sales in other regions.

Fig. 7 presents the principal components analysis of leading game genres. The games situated towards the left in the plot are indicative of the highest sales volumes, while those positioned towards the top mainly pertain to games predominantly sold in Japan. Shooter games (illustrated as blue crosses) are the most widely sold games worldwide, with some examples proving particularly profitable in the Japanese market, based by the points located in the top-left quadrant. As previously highlighted, certain action games (represented as yellow triangles) and role-playing games (denoted by pink squares) have a predominant presence in the Japanese market, with some exclusively marketed within this region (located notably in the top-right quadrant of the plot).

To conclude the PCA analysis, we can determine how much each variable is represented in a given component. To do so, we will implement the *square cosine* technique. Mathematically, the Cos2 for a variable or category in a given component is calculated as the squared cosine of the variable's/category's coordinates on that component.

The Cos2 values range between 0 and 1, where:

- A low Cos2 value indicates that the variable/category is not well represented by the component.
- A high Cos2 value signifies a strong representation of the variable/category on the component.

Fig. 8 combines a biplot of the attributes with the computed cos2 score, from which we can extract that all variables that we considered when doing PCA are strongly represented in both of the principal components selected.

Multidimensional scaling (MDS)

Multidimensional Scaling (MDS) is a 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

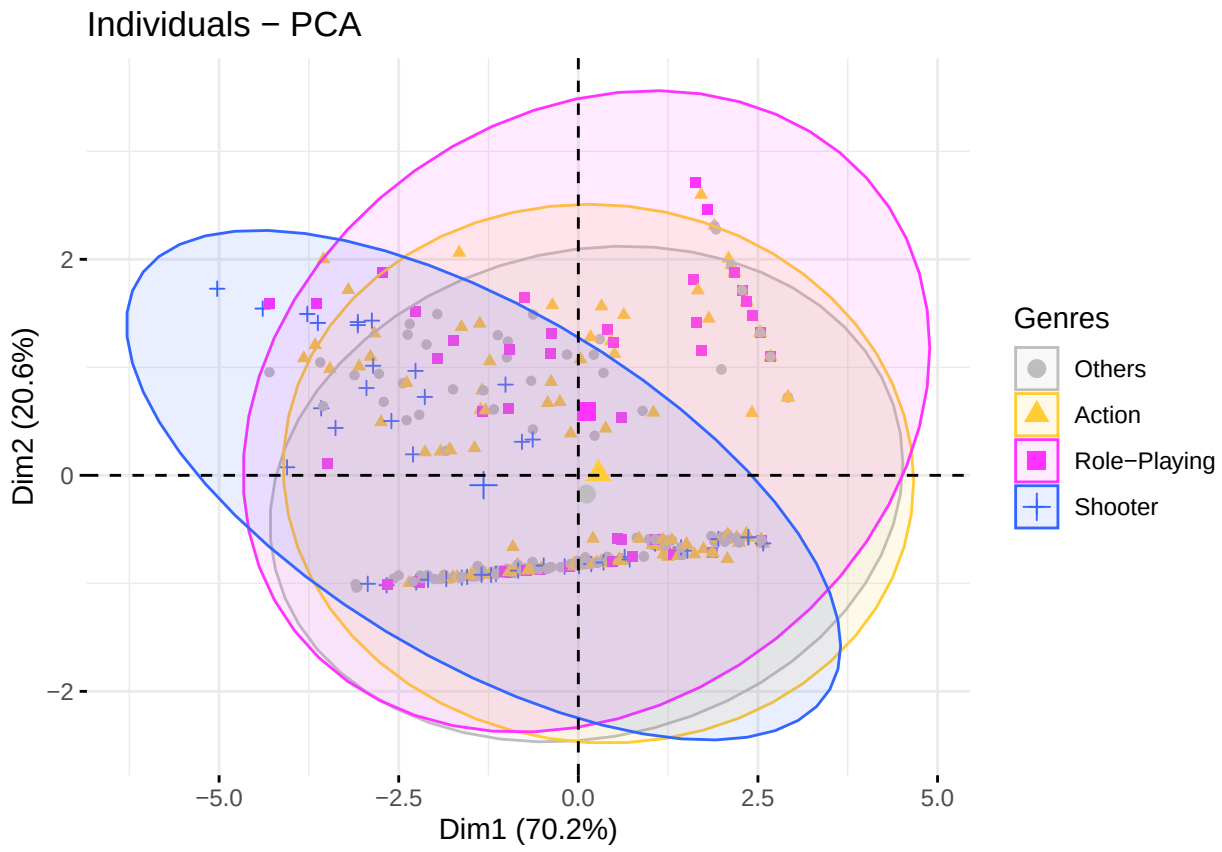


Figure 7: Principal Component Analysis (PCA) plot displaying the top selling game genres' distribution with respect to the two principal components. Points in yellow corresponds to Action games, those in pink to Role-Playing games, those in blue to Shooters, and the rest of the genres are visualized in grey.

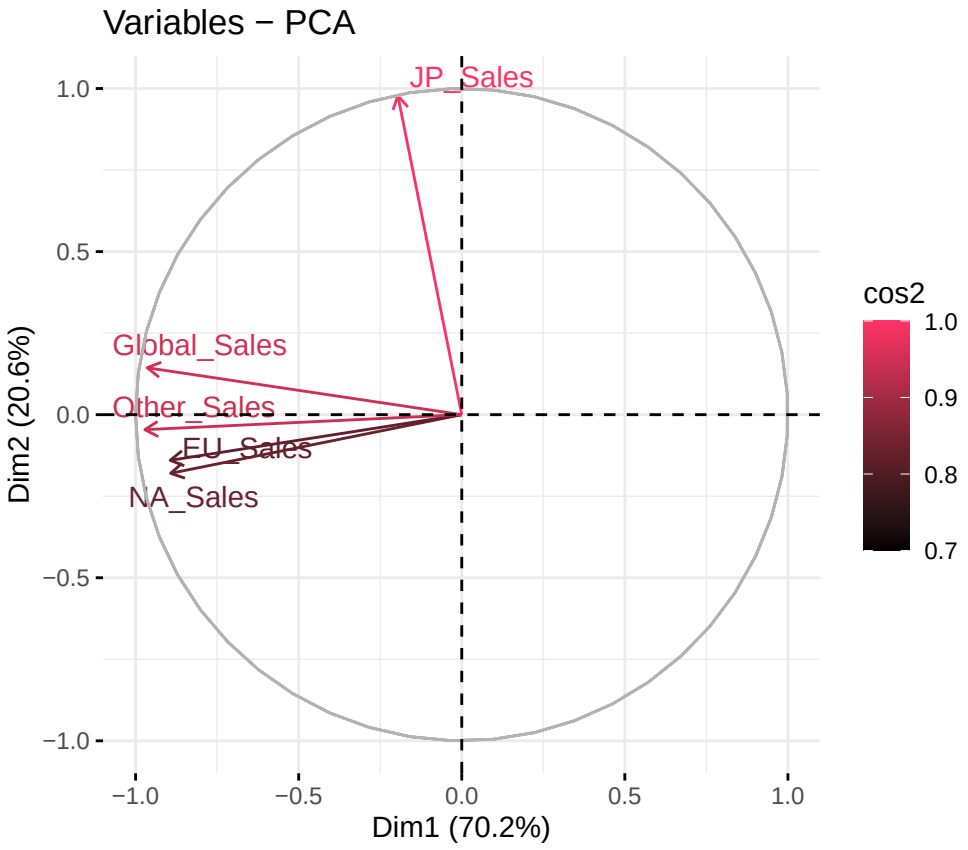
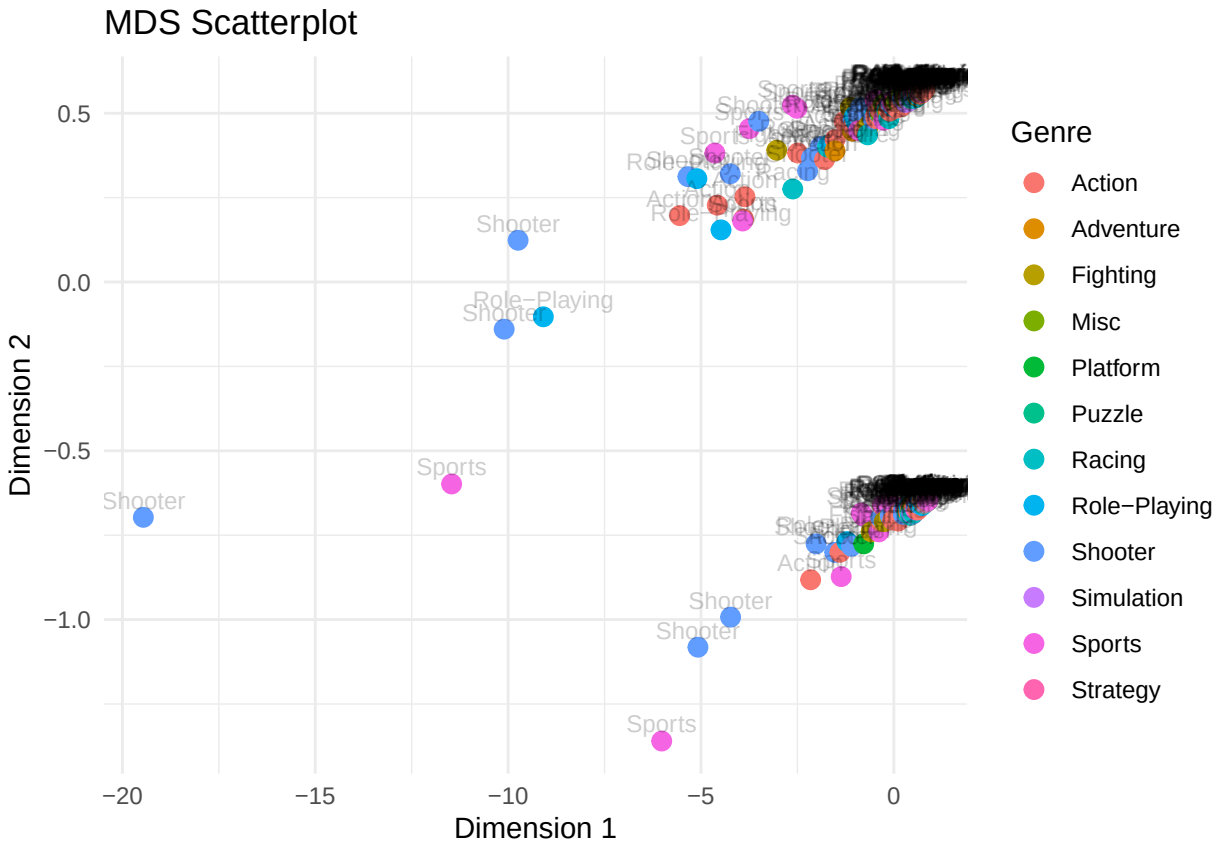


Figure 8: Visualization of the quality of representation (Cos2) of rows/columns from the results of Principal Component Analysis (PCA). The color gradient represents the strength of representation, with black indicating low representation, orange indicating moderate representation, and red indicating high representation.

revealing patterns, clustering, and interpreting relationships in diverse datasets with a mix of variable types.

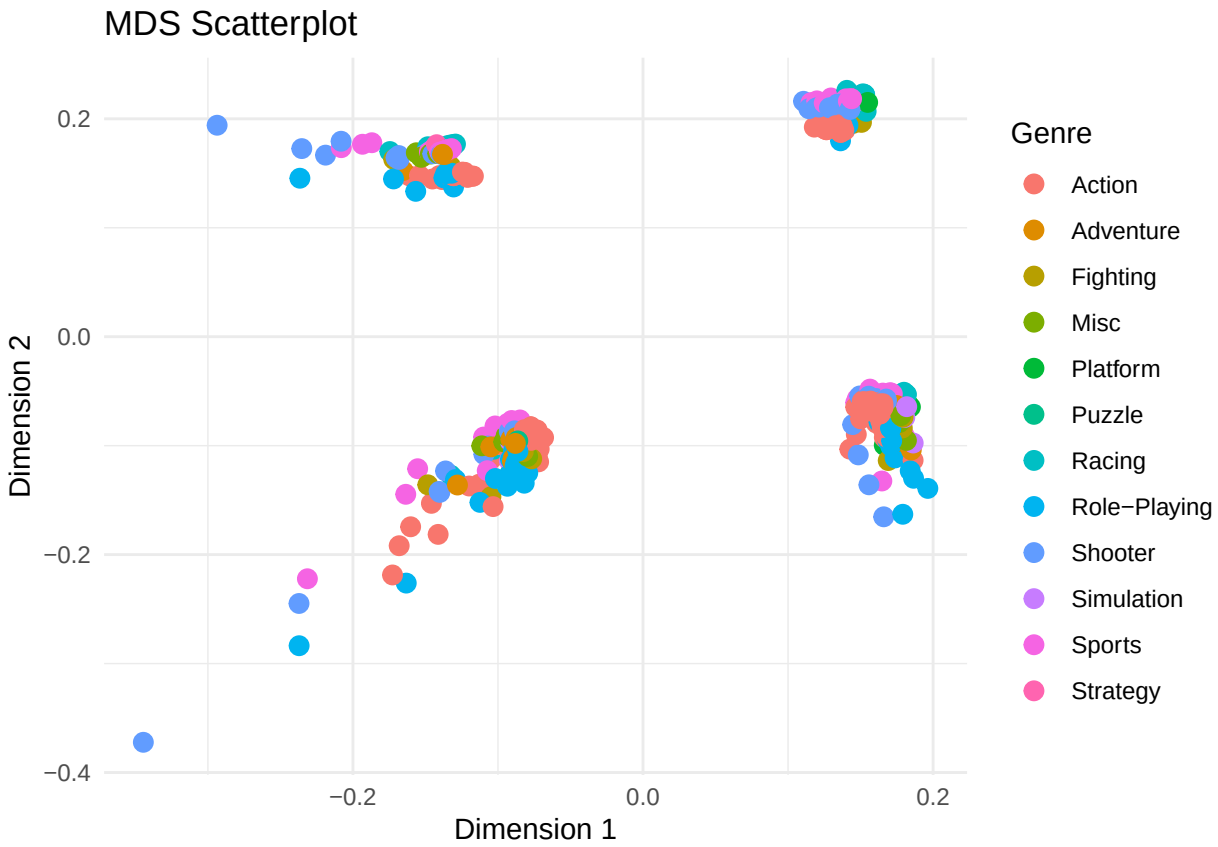
The `dist` function in R, by default, uses Euclidean distance for numerical variables. However, when categorical variables are present, dist converts them into binary indicators (dummy variables) and calculates the dissimilarity based on these indicators. The Jaccard distance, which measures the dissimilarity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



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 flexibility 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.



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.

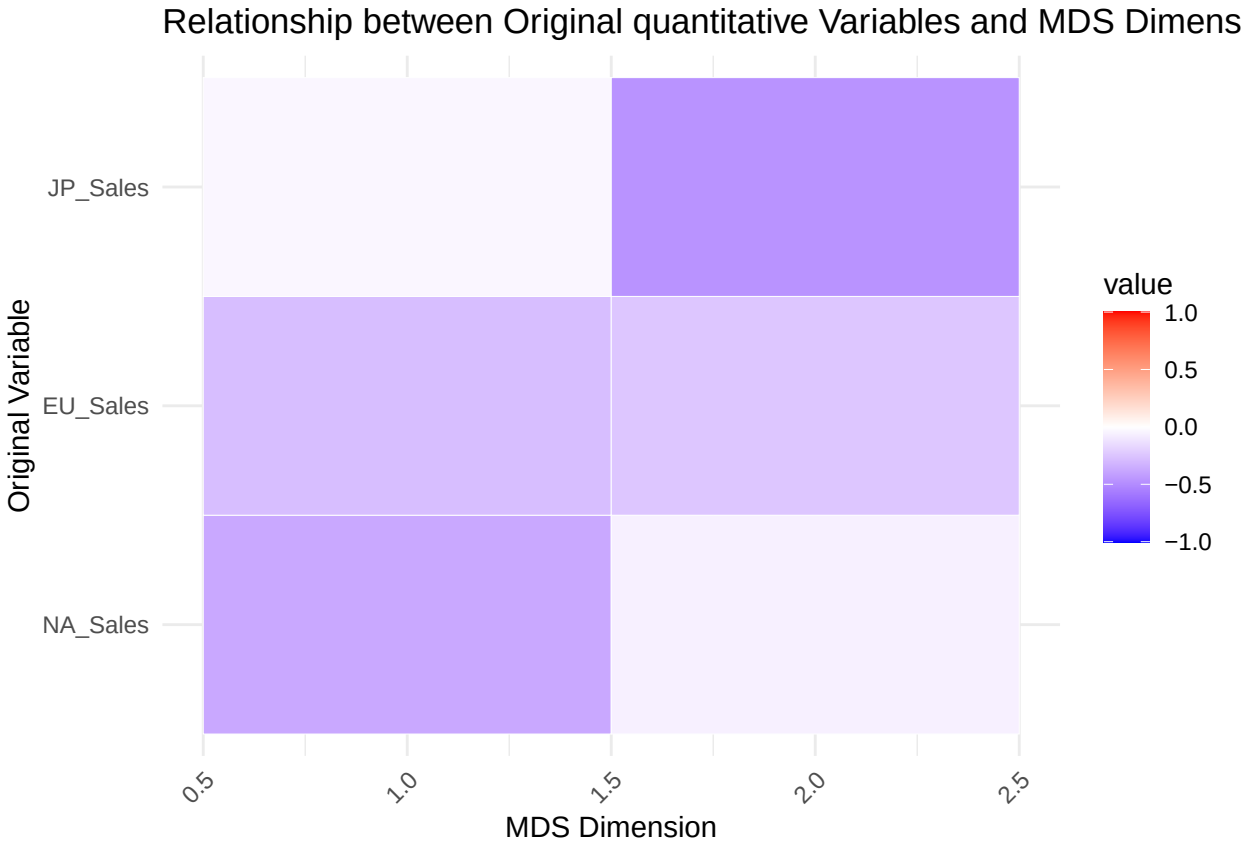
The Gower distance $d_G(x, y)$ between two data points x and y is computed as:

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

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 appropiate 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.



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, and performing Principal Component Analysis, 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:

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

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:

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

where $|S_j|$ 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:

```
# 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(data, centers = i, nstart = 10)
  wss[i] <- kmeans_model$tot.withinss
}
```

Fig. 9 contains the elbow plot of the model created above. Given the results, we consider that using $K = 3$ clusters will be the best option to faciliate the interpretation of the results.

Fig. 10 visualizes the clusters found with respect with the previously found Principal Components.

Based on Fig. 10, we can derive an interpretation for the clusters found. The cluster labelled as 1, represents games with high or moderate sales volume and predominantly sold outside Japan (with some exceptions that seem to be popular in the Orient). The second cluster (2) includes games with the highest sales volume in Japan, with some exceptions regarding games with the highest sales volume in the West. Finally, the third (3) cluster contains the games with the lowest sales in our data set, both in Japan and in the West.

By inspecting the genre distribution in each cluster, Fig. 11, and based on the aforementioned classification, we can have a visual representation of which is the most sold genre.

As previously mentioned, shooter games have the highest sales, given their predominance within the second (2) cluster. Furthermore, this finding is in line with what we saw in Figure 7, where we saw that shooter games are the ones that amount to the highest sales, with some cases being predominantly sold in Japan. It is interesting to see how heterogeneous the three defined clusters are, although the sales trends are clearly visible in the results.

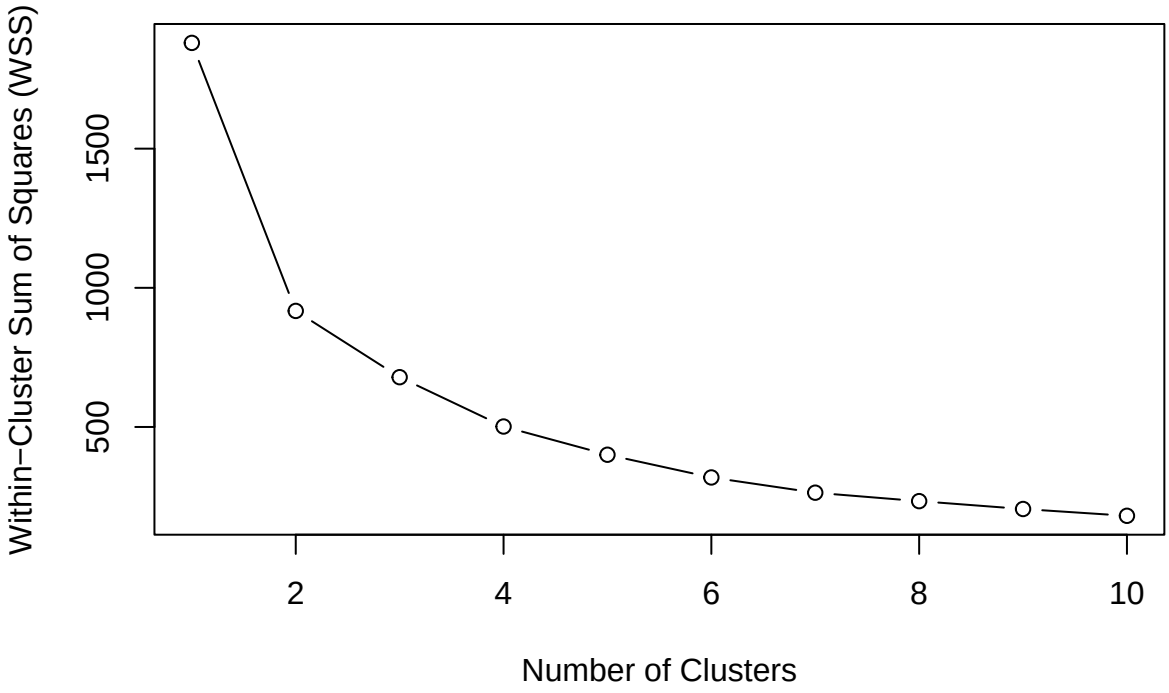


Figure 9: Elbow plot for the K-means clustering algorithm

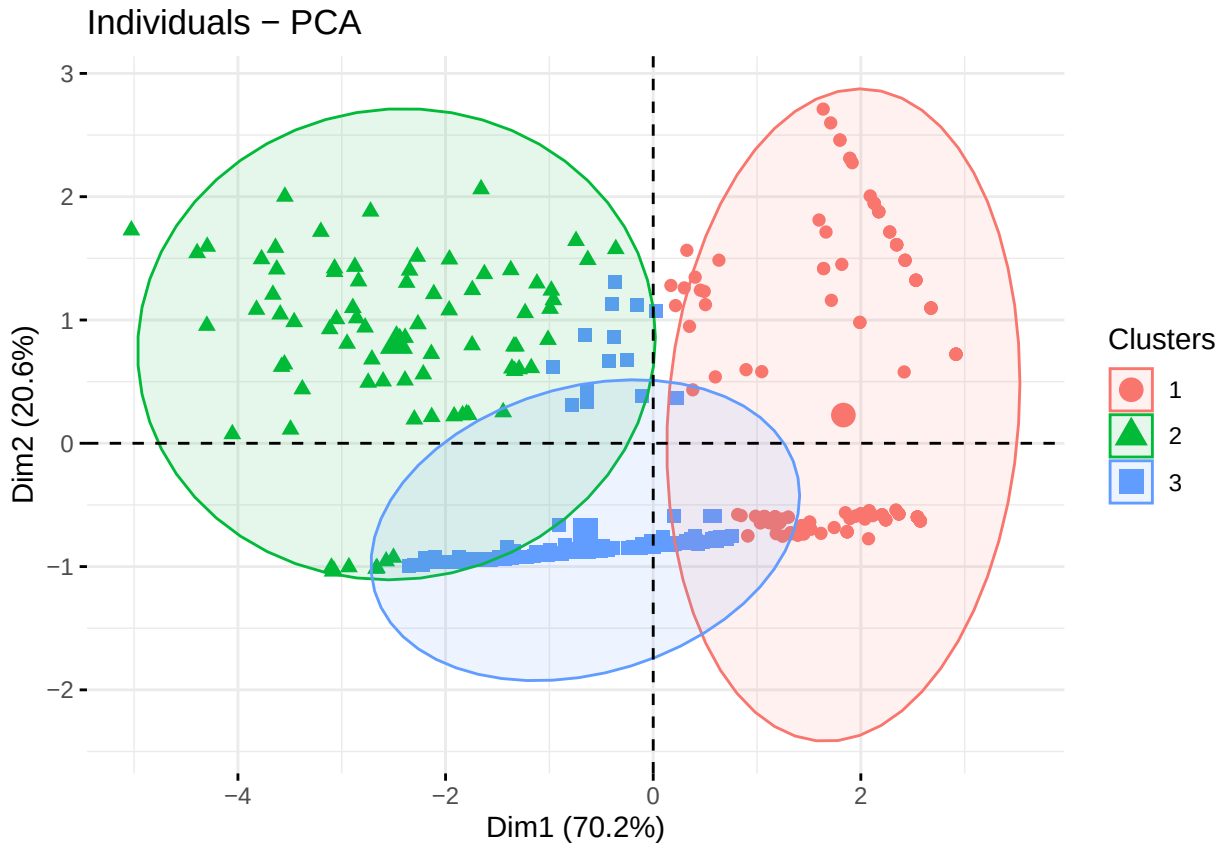


Figure 10: Visualization of the clusters with respect to the Principal Components. Points with a higher size corresponds to the cluster centers of each group

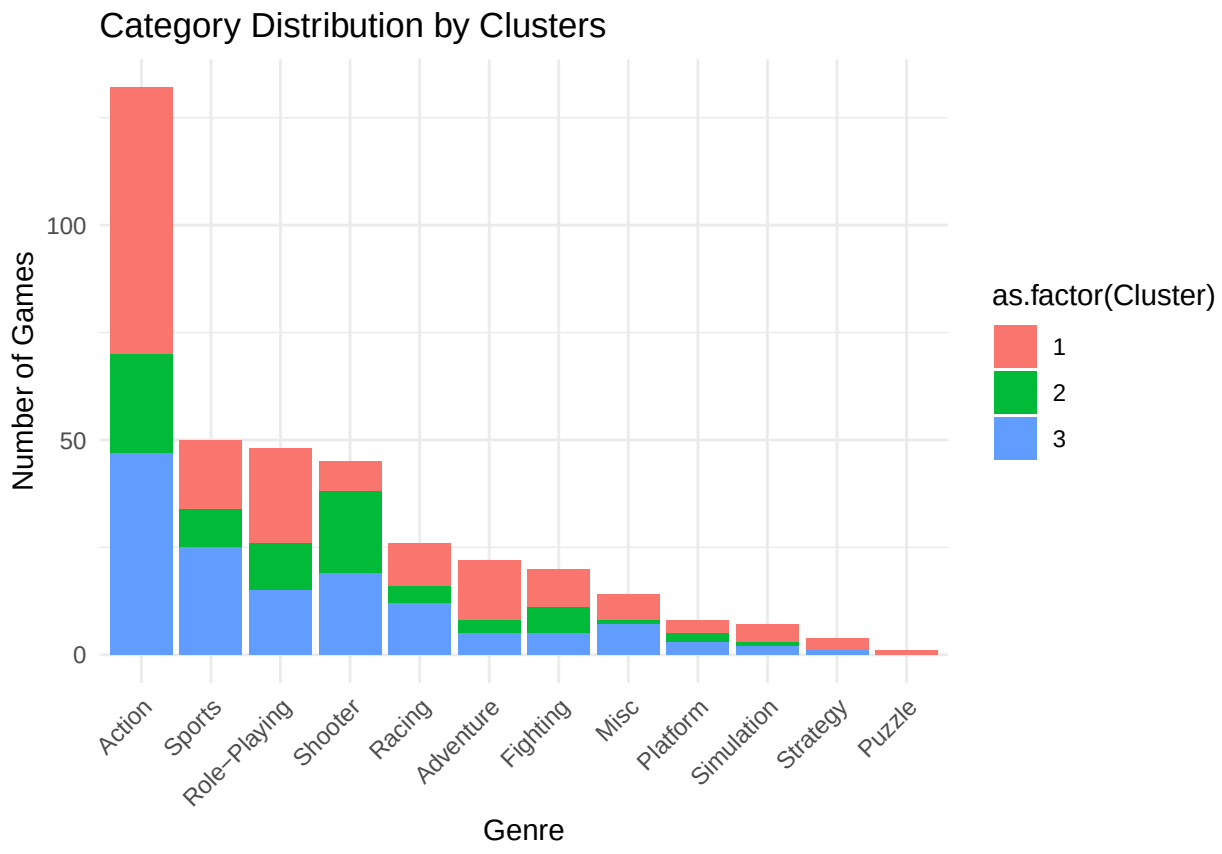


Figure 11: Genre distribution in each cluster.

References

Grané, Aurea. n.d. "Multidimensional Datasets." https://aulaglobal.uc3m.es/mod/resource/view.php?id=4751246.

"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.