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**: noviembre 15, 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 actuality, the average ranking of games within our dataset stands at 9373.

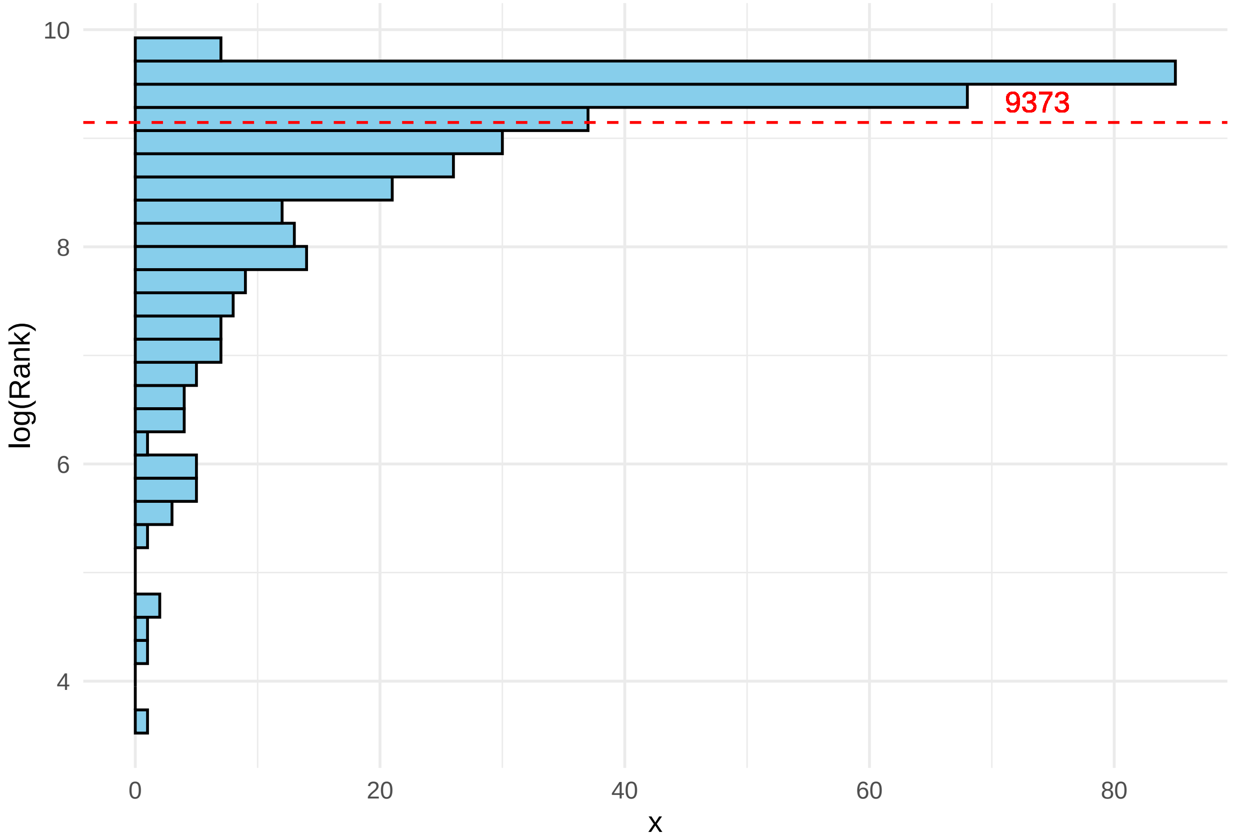


Figure 1: Distribution of the log-transformed Rank values. The red dashed line represents the median of the distribution

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

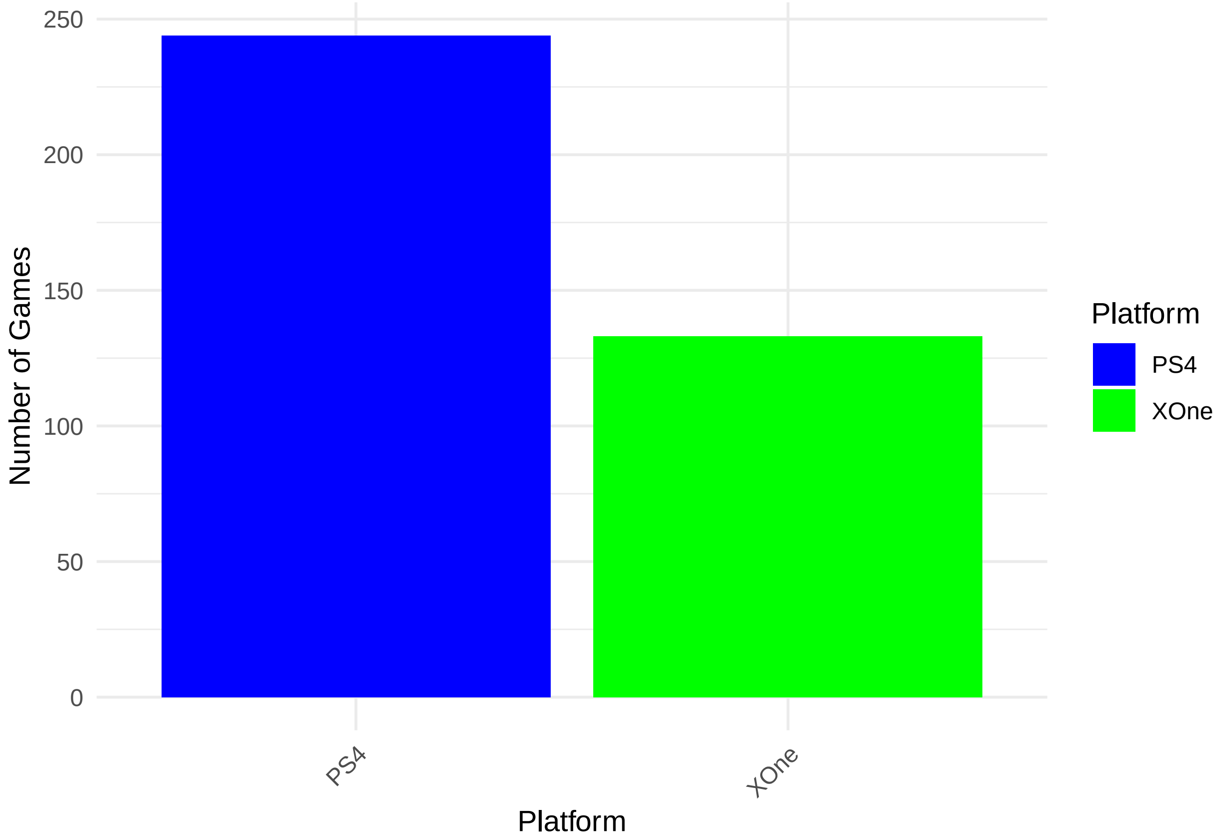


Figure 2: Number of games of each platform inside the dataset

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 ϵ an arbitrary small constant.

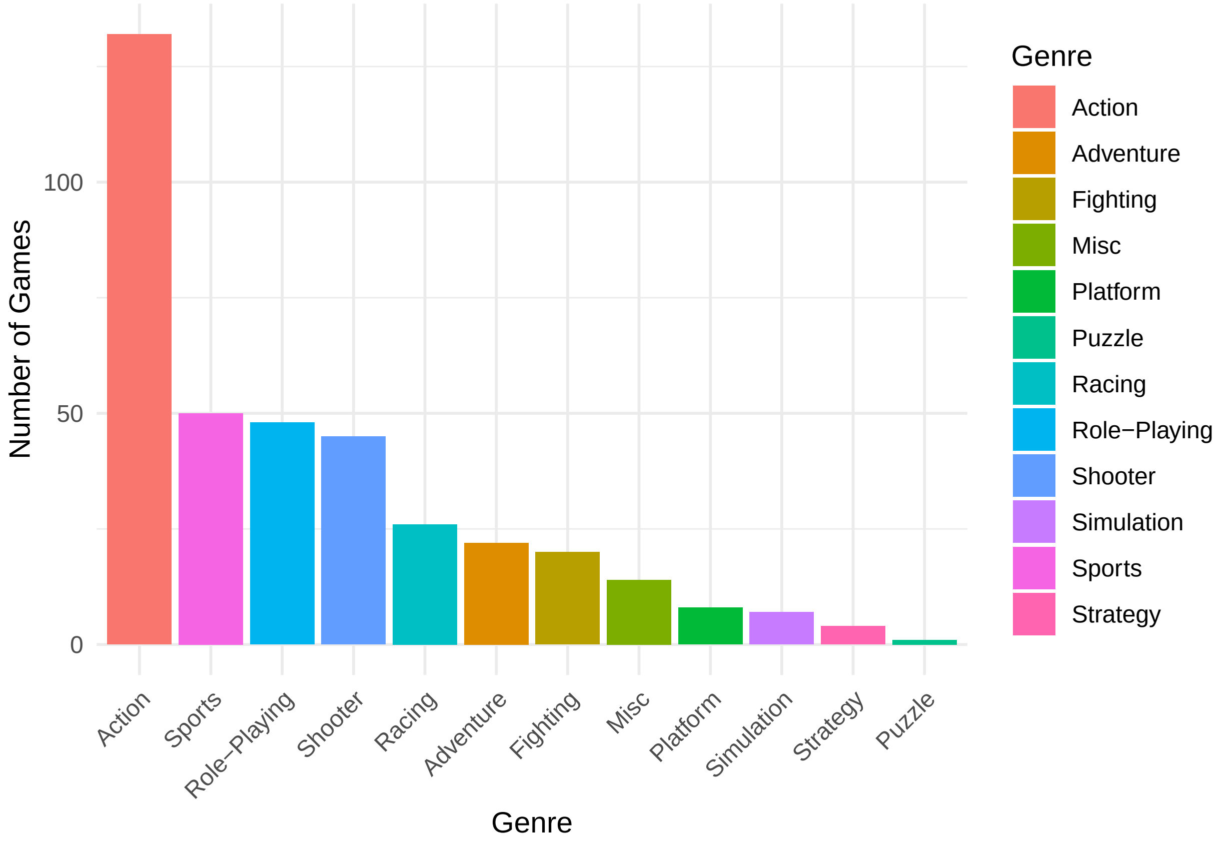


Figure 3: Barplot of the amount of games of each genre

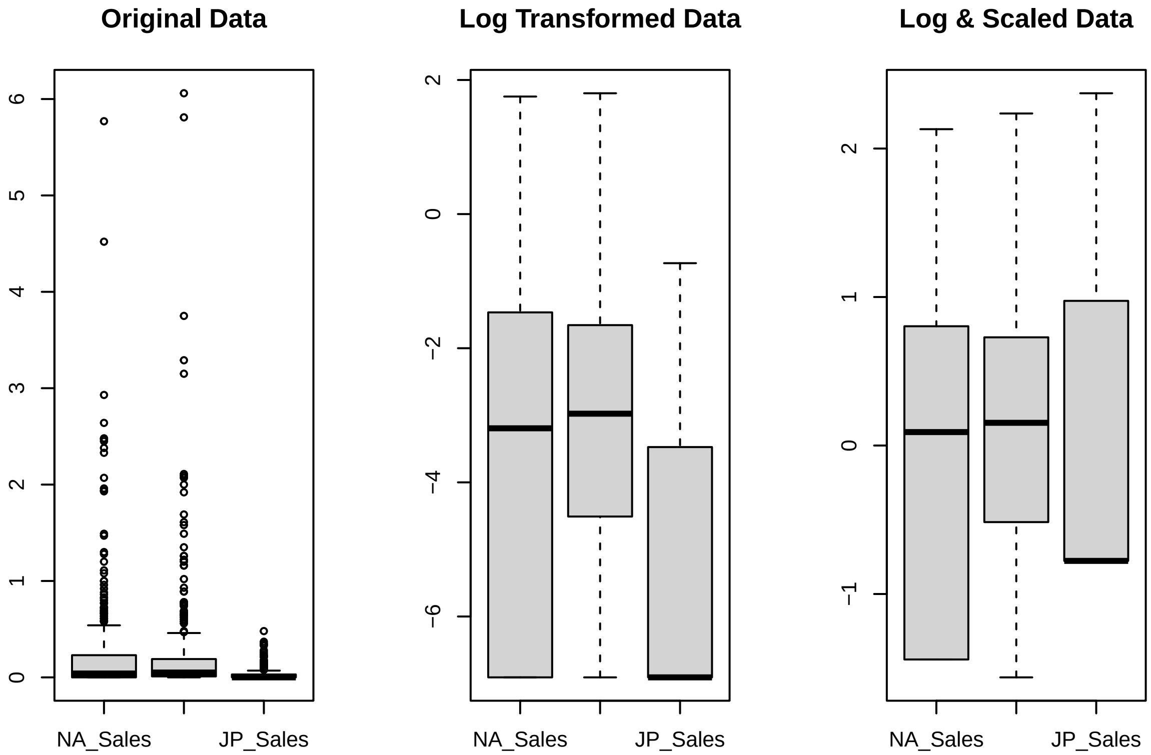


Figure 4: Boxplots of the different transformations applyied to the sales data. From left to right, the original data, the log-transformed data and the log transformed and scaled data

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 West is 0.394, with an even lower correlation of 0.299 with the North American market, as illustrated in Fig. 4. 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 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 utilised in the Japanese video game industry are of great significance. Numerous Japanese games are designed with a primary focus on the local market, giving rise to gameplay and cultural elements that may not resonate as strongly with Western or North American audiences. Therefore, these games may not be successful beyond their intended audience in the East, resulting in a larger sales gap and a weaker association with these regions.

On the contrary, the sales in North America and Europe are more correlated as these regions share similar western cultures and comparable marketing strategies. Conversely, Japan presents a more distinct market, with its citizens' tastes significantly differing from those of western cultures. Fig. 4 displays a correlation matrix using a heatmap to visualize this relationship.

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

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

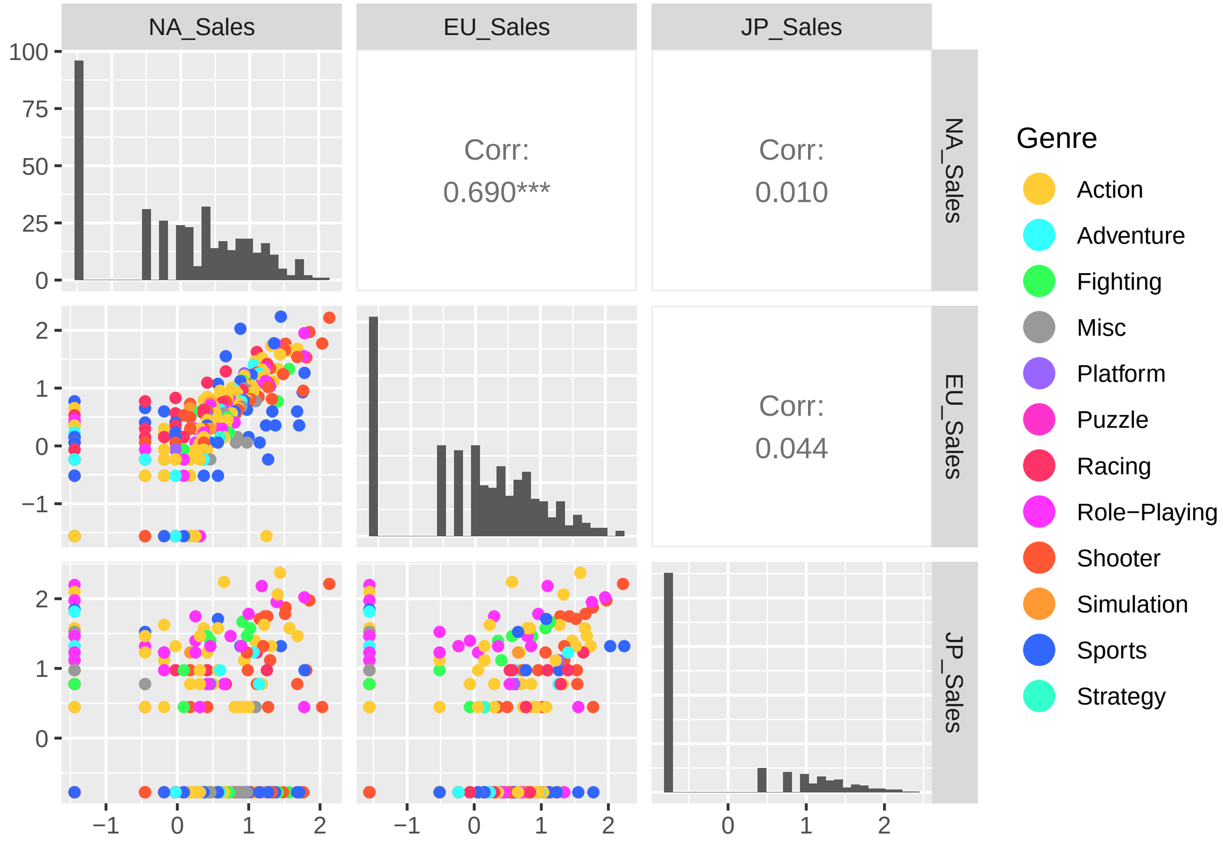


Figure 5: Correlation Plot of Video Game Sales in Different Regions

investigation. It is evident from the table that the genre of Role-Playing Games (RPGs) enjoys significantly greater popularity in Japan as compared to North America and Europe. Strikingly, our research 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 above the diagonal line (y = x), 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

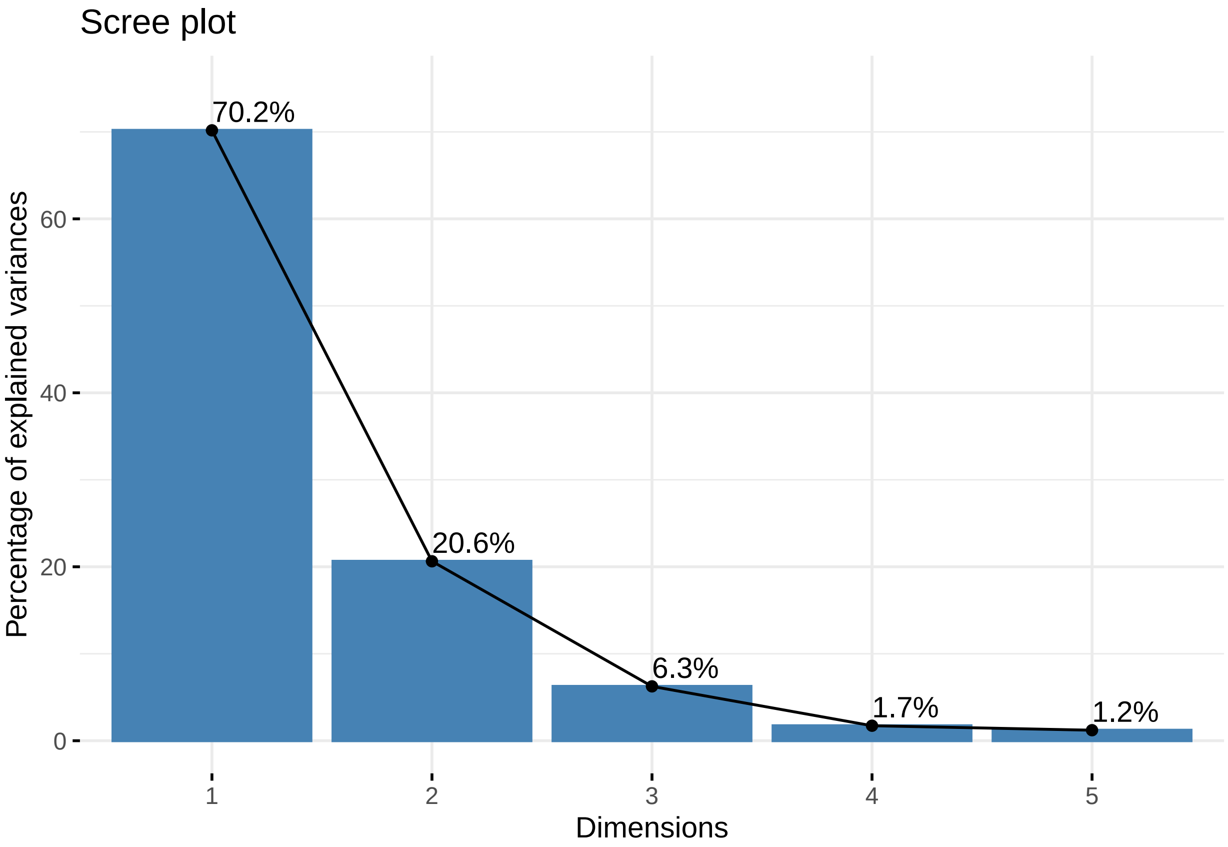
After scaling our data, we are now prepared to perform Principal Component Analysis (PCA). Scaling is a crucial preprocessing step as it ensures that each variable contributes equally to the analysis by standardizing their scales. PCA is particularly sensitive to the scale of variables, and standardizing them helps prevent variables with larger scales from dominating the analysis.

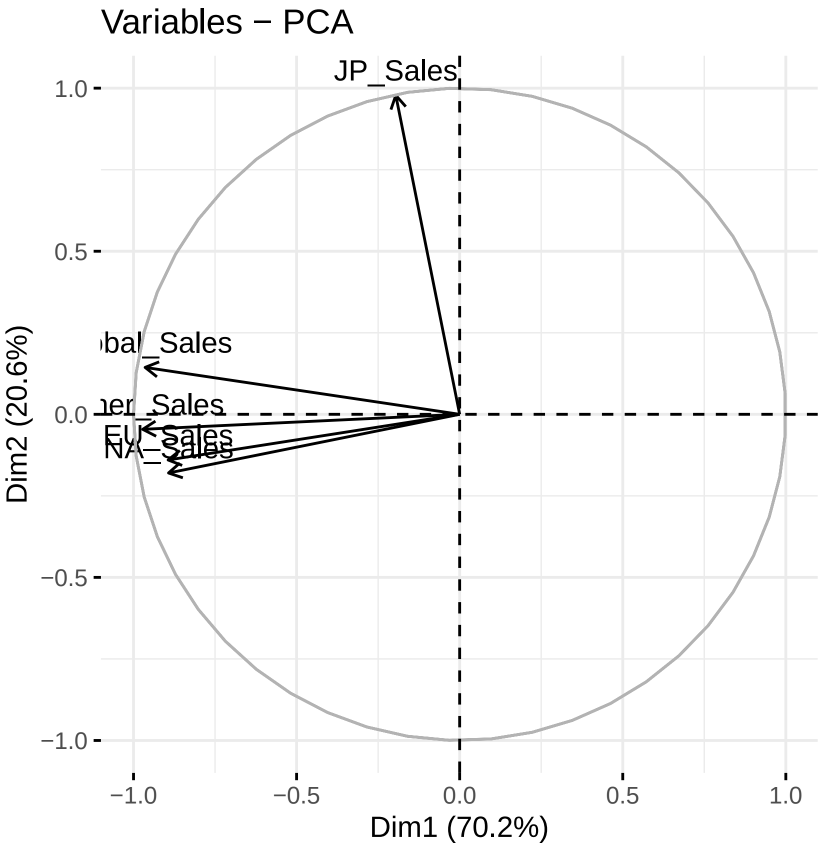
Principal Component Analysis (PCA) is a crucial step in our data analysis pipeline for several reasons. PCA allows us to effectively reduce the dimensionality of our dataset by transforming the original variables into a set of uncorrelated principal components, therefore addressing multicollinearity, capturing the essential information with fewer variables. This reduction is particularly valuable when dealing with high-dimensional datasets, enabling more manageable and interpretable analyses. Furthermore, PCA helps reveal underlying patterns and relationships within the data by highlighting the variables that contribute most to the observed variability.

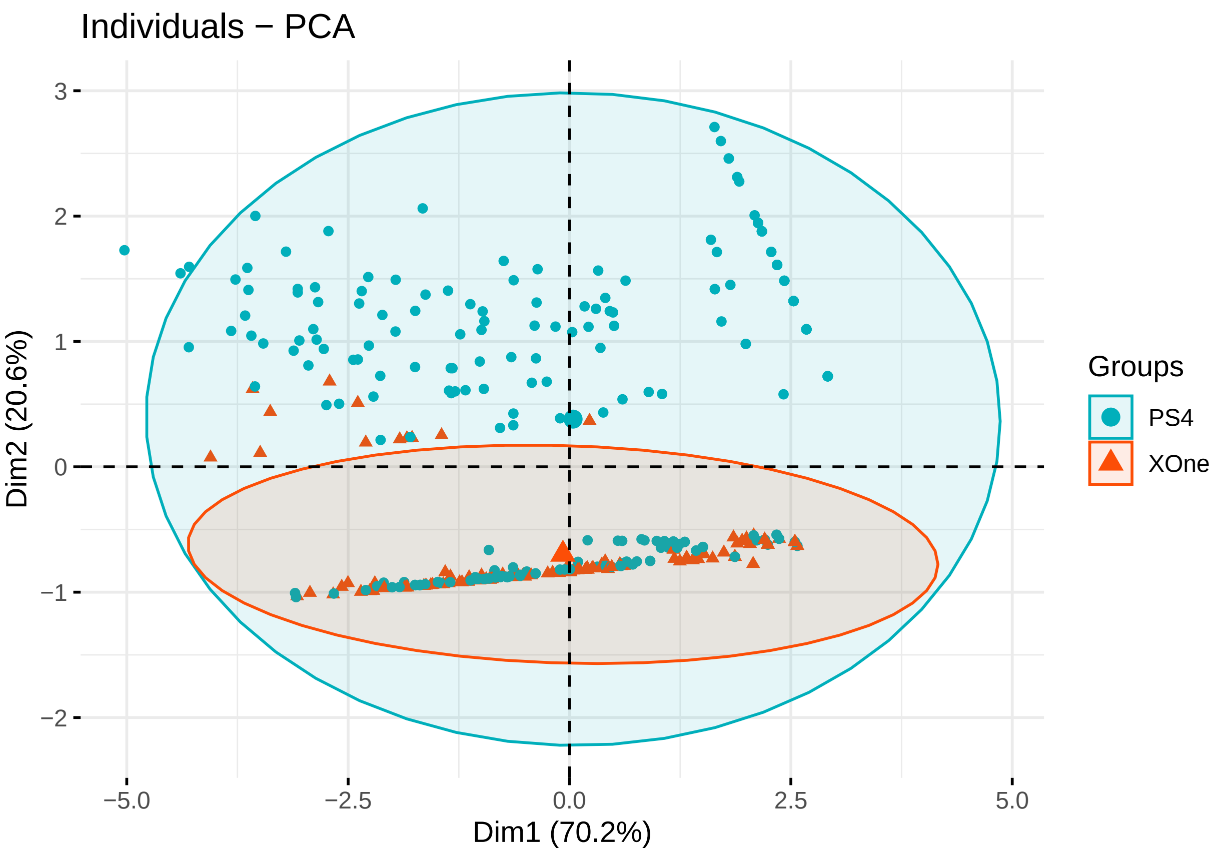
Table 5: Summary of PCA

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.873205	1.015724	0.5590688	0.293816	0.246016
Proportion of Variance	0.701780	0.206340	0.0625100	0.017270	0.012100
Cumulative Proportion	0.701780	0.908120	0.9706300	0.987900	1.000000

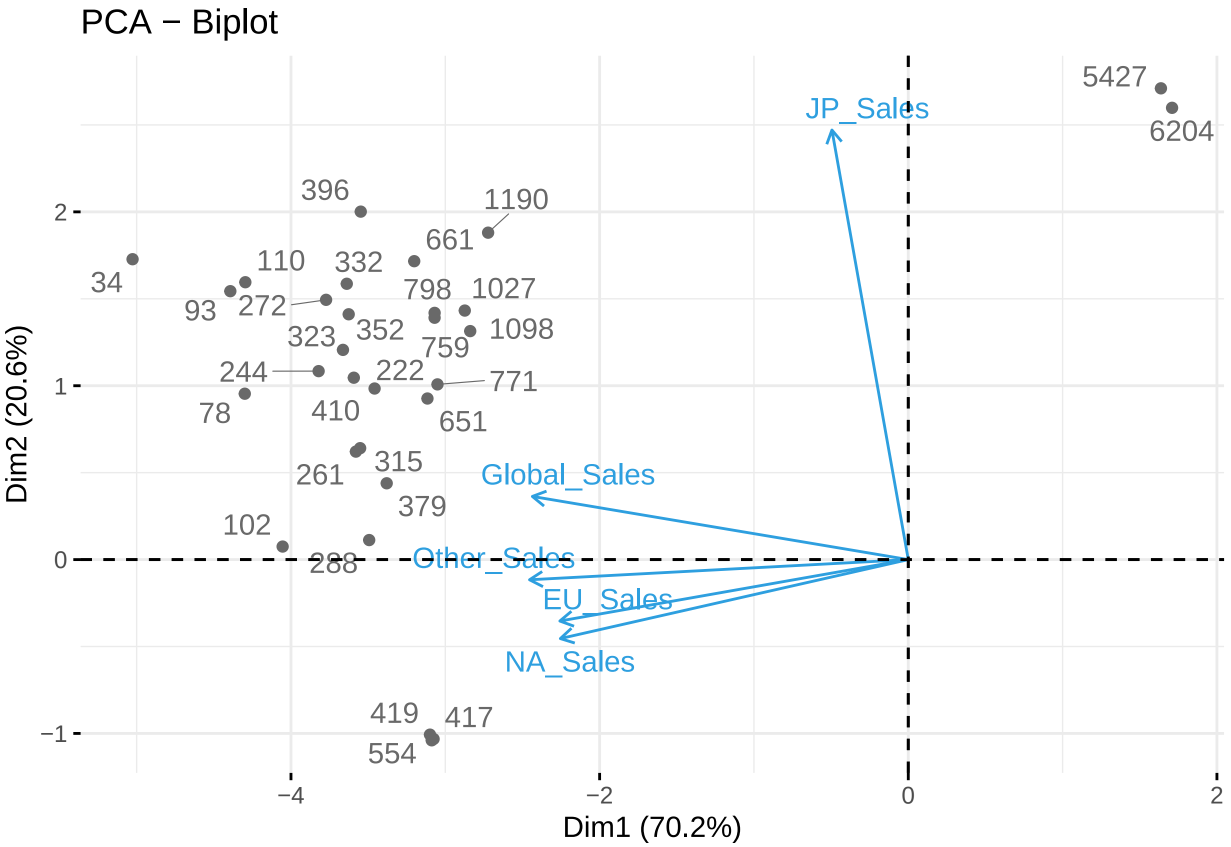
As seen by the results, 0.97 of the total variability in the dataset has been explained by the three first components. The cumulative proportion of variance explained by each of the last two principal component is can be neglected. By using the package `factoextra` we can create a Scree plot which help decide on the number of components or factors to retain in the analysis. It displays the eigenvalues of the principal components or factors in descending order.


Scree plot
70.2%
20.6%
6.3%
1.7%
1.2%
Percentage of explained variances
60
40
20
0
1
2
3
4
5
Dimensions


Variables − PCA
JP_Sales
obal_Sales
her_Sales
EU_Sales
NA_Sales
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
Dim1 (70.2%)
Dim2 (20.6%)



The `fviz_pca_biplot` function in R, part of the `factoextra` package, creates a biplot for Principal Component Analysis (PCA) results. A biplot simultaneously displays both the observations and variables in the same plot, allowing for a quick visual assessment of relationships between them. Due to the large size of our dataset we will display only the top 30 most contributing observations. This is useful because it enables an intuitive exploration of the contribution of variables to each principal component and the relationships between observations in the reduced-dimensional space.



```
# Percentage of variance explained by each principal component
var_explained <- pca_result$sdev^2 / sum(pca_result$sdev^2) * 100

# Cumulative percentage of variance explained
cumulative_var_explained <- cumsum(var_explained)

# Choose the number of components explaining 95% of cumulative variance
num_components <- which(cumulative_var_explained >= 95)[1]

# Retain the selected number of components
selected_components <- pca_result$x[, 1:num_components]

# Print the selected components
print(colnames(selected_components))
```

```
## [1] "PC1" "PC2" "PC3"
```

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.