Multivariate Analysis in Video Games sales *

Álvaro Novillo *Universidad Carlos III*
Polo *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 06, 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.

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.

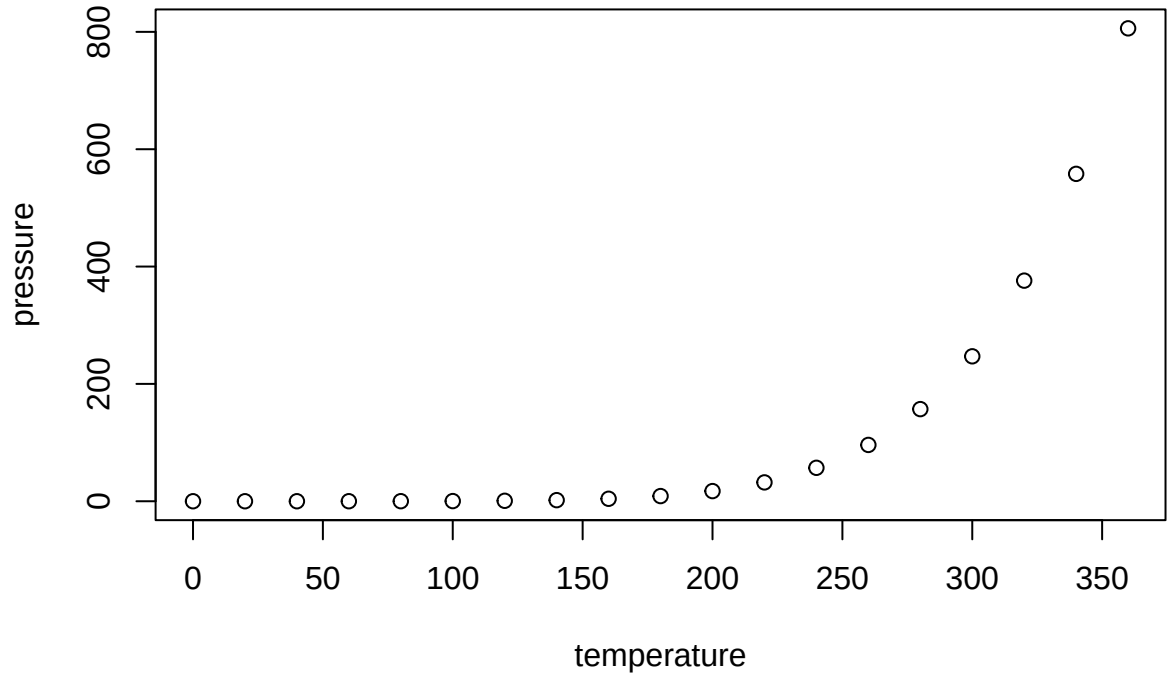
```
df = read.csv("vgsales.csv")

# Filter for games in 2015 and 2016, and on the PS4 or PC platform
filtered_df <- df[df$Year %in% c(2015, 2016) & df$Platform %in% c("PS4", "XOne"), ]
head(filtered_df)
```

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

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.