Video Game Sales Analysis

Likun Zhang
ISBD, Renmin University of China

May 9, 2023

1 Introduction

The Video Game Sales Dataset provides up-to-date information on the sales performance and popularity of various video games worldwide. The data are collected by Kaggle from VGChartz, a video game sales tracking website. The data includes the name, platform, year of release, genre, publisher, and sales in North America, Europe, Japan, and other regions. It also features scores and ratings from both critics and users, including average critic score, number of critics reviewed, average user score, number of users reviewed, developer, and rating. This comprehensive and essential dataset offers valuable insights into the global video game market and is a must-have tool for gamers, industry professionals, and market researchers.

In this report, we aim to summarize the history of the global game sales market and predict its future development.

2 Methods

2.1 Data Management

The data consist of 16719 rows and 16 columns. The details of all covariates are given in Table 1. In the data cleaning step, we group the same video game released on different platforms together and add up their salesl; we removed rows with "Genre" equals "Idea Factory" or "Sony Computer Entertainment" and "Year of Release" later than 2016 for the data sparsity. Finially we end up with a dataset with unique game name of 11325 rows.

2.2 Exploratory Data Analysis

Figure 1 shows the total global video game sales by year. The dataset collects the game sales information since year 1980. We can see that most data points are collected after year 2000.

Figure 2 shows the global video game sales by year from 1980 to 2016 by genre. From this plot, we may speculate that some genres of games share the similar development trend. Note that the high variation before year 2000 dues to the sparsity of data points.

To better capture development trends, we cluster 12 genres of games, based on their average global sales from 2000 to 2016 using Dynamic Time Warping (DTW). Dynamic Time Warping is an algorithm used for measuring similarity between two temporal sequences that may vary in speed. It is useful in many domains such as speech recognition, data mining and financial markets. After the clustering we end up with 4 clusters; the details are given in Figure 3.

It is reasonable to suspect that the game market development is associated with the development of hardwares. For example, the high performace of computers faciliates PC

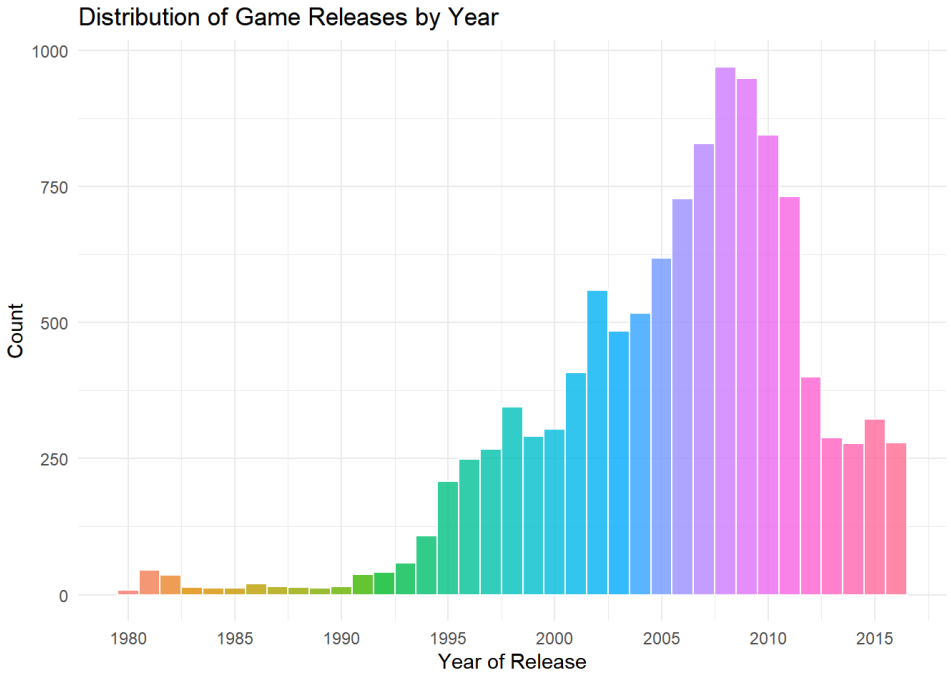


Figure 1: Total global video game sales by year from 1980 to 2016

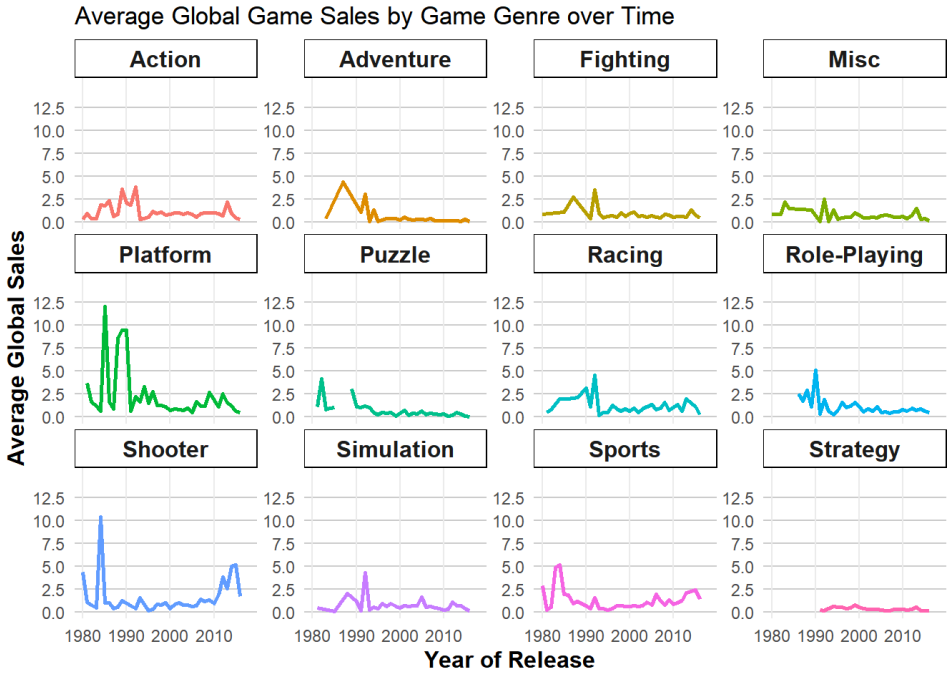
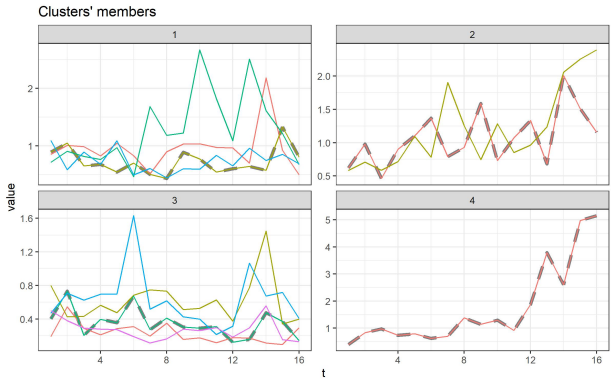


Figure 2: Global video game sales by year from 1980 to 2016 by genre

```
Cluster 1 contains the following genres:
[1] "Action"       "Fighting"     "Platform"     "Role-Playing"
Cluster 2 contains the following genres:
[1] "Racing" "Sports"
Cluster 3 contains the following genres:
[1] "Adventure"  "Misc"        "Puzzle"     "Simulation" "Strategy"

Cluster 4 contains the following genres:
[1] "Shooter"
```

(a) Clusters information



(b) Clusters details

Figure 3: 12 generes of games clustered based on DTW

games with complex features and elaborate visual, which was hard to imagine 20 years ago. Hence We use a CPU and GPU Performances Dataset downloaded from Kaggle to help capture the hardwares development over the two decades and better predict the future global game sales market. We average the frequency of CPUs and GPUs released by year. See Figure 4.

2.3 Model Fitting

One important feature of our task is that we have to handle multiple time series simultaneously and those series are correlated with each other. Instead of analyze each series separately, we use Vector Autoregressive (VAR) model. The VAR model is a multivariate time series model that relates current observations of a variable with past observations of itself and past observations of other variables in the system. It is used to capture the relationship between multiple quantities as they change over time. VAR models generalize the single-variable (univariate) autoregressive model by allowing for multivariate time series.

In brief, given a k-dimensional time series $y_t = (y_{1t}, \ldots, y_{kt})$, a VAR model of order p

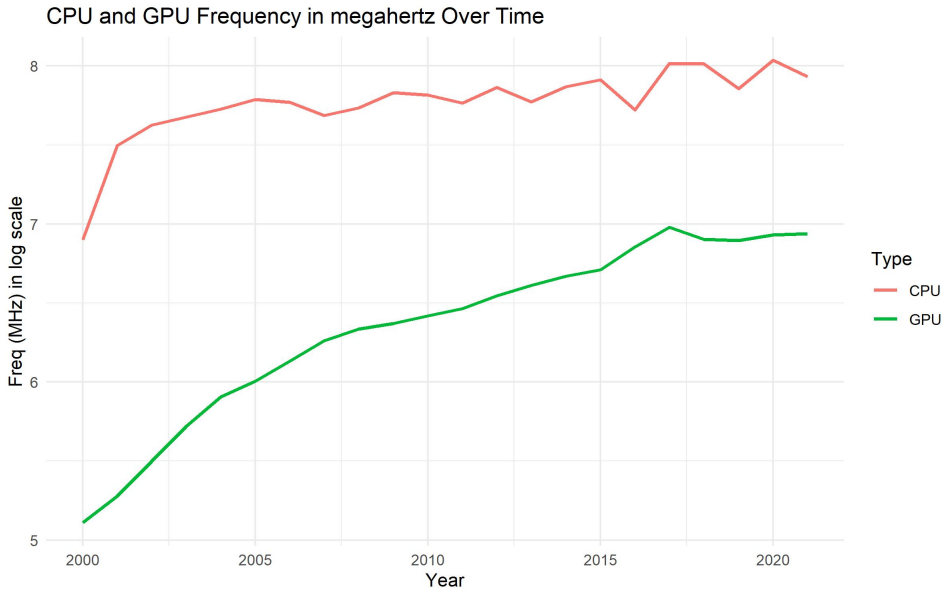


Figure 4: CPUs and GPUs (log) Frequence by year from 2000 to 2021

can be formulated as follows:

$$y_t = \delta + \Theta_i y_{t-1} + .. + \Theta_p y_{t-p} + \epsilon_t.$$

In our model y_t consists of mean global sales of 4 clusters, GPUs frequency and CPUs frequency. We differentiate each single time series once to ensure stationarity and fit a VAR(1) model to our data. Then we pridict the mean global game sale of 4 clusters in the next five years. See Figure 5.

From Figure 5, we see that the future trend of 4 clusters of games are quite different. Cluster 4 ("Shooter") is expected to flourish in the next five years; the future market size of cluster 2 ("Racing" and "Sports") is predicted to expand slowly; the future sales of cluster 1 ("Action", "Fighting", "Platform" and "Role-Playing") and cluster 3 ("Adventure", "Misc", "Puzzle", "Simulation" and "Strategy") are predicted to remain steady.

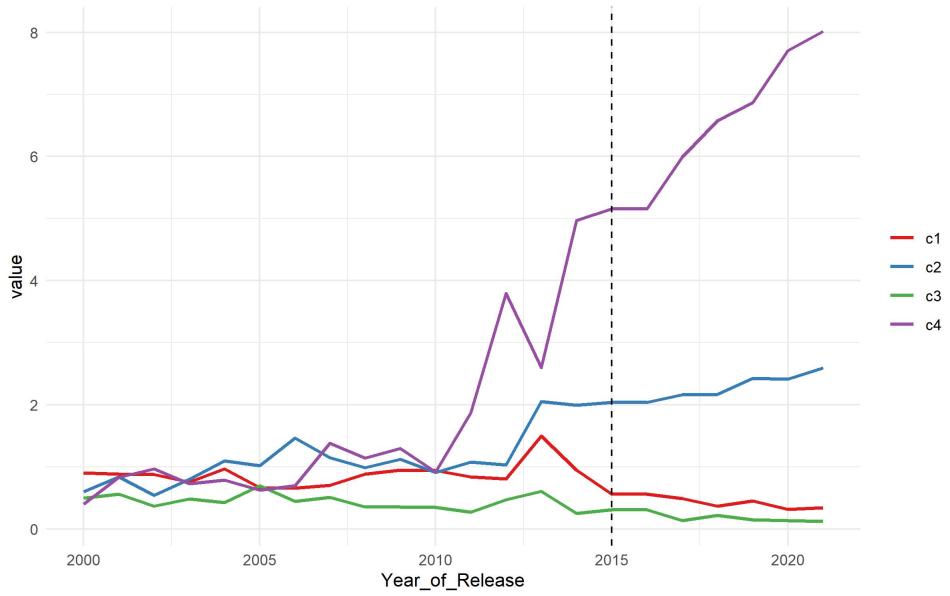


Figure 5: Mean global game sales prediction for 4 clusters

3 Conclusion

- Based on average global sales, We grouped 12 genres of games into 4 cluster
  - 1. "Action" "Fighting" "Platform" and "Role-playing"
  - 2. "Racing" and "Sports"
  - 3. "Adventure" "Misc" "Puzzle" "Simulation" and "Strategy"
  - 4. "shooter"
- In terms of future global game market size, c4 is expected to be the largest, followed by c2, then c1, and finally c3

Table 1: Covariates description

Covariate name	Description
Name	The name of the video game.
Platform	The platform on which the game was released.
Year of Release	The year in which the game was released.
Genre	The genre of the video game.
Publisher	The company responsible for publishing the game.
NA Sales	The sales of the game in North America.
EU Sales	The sales of the game in Europe.
JP Sales	The sales of the game in Japan.
Other Sales	The sales of the game in other regions.
Global Sales	The total sales of the game across the world.
Critic Score	The average score given to the game by professional critics.
Critic Count	The number of critics who reviewed the game.
User Score	The average score given to the game by users.
User Count	The number of users who reviewed the game.
Developer	The company responsible for developing the game.
Rating	The rating assigned to the game by organizations.