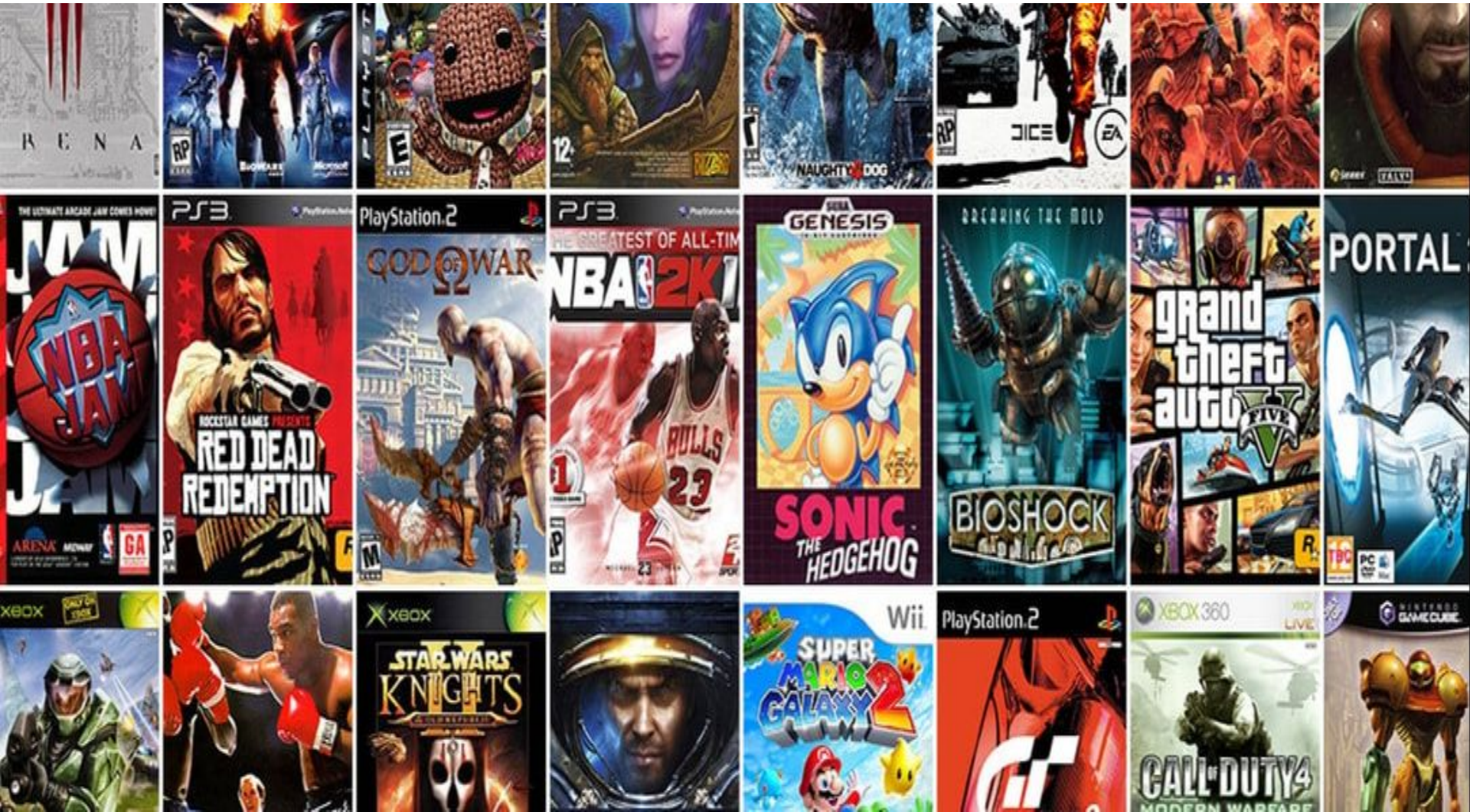
Video Game Sales Analysis

Group 29:

- Kushal Therokar
- Vamshi Krishna Perabathula
- Khiem Doan


ARENA
NBA JAM
PS3
ROCKSTAR GAMES PRESENTS
RED DEAD REDEMPTION
PlayStation 2
GOD OF WAR
PS3
NBA 2K1
SEGA GENESIS
SONIC THE HEDGEHOG
BREAKING THE MOLD
BIOSHOCK
grand theft auto V
PORTAL
XBOX
XBOX
STAR WARS KNIGHTS
Wii
SUPER MARIO GALAXY 2
PlayStation 2
XBOX 360
LIVE
CALL OF DUTY 4
MODERN WARFARE
NINTENDO GAMECUBE

Introduction and Objectives

1

The project aims to analyze a dataset of video game sales to derive insights beneficial for stakeholders in the gaming industry.

2

The research also seeks to understand regional preferences in gaming and how various factors influence game sales.

3

Primary objectives include identifying popular game types, profitable gaming platforms, and trends over time.

Data Exploration

Overview of the Dataset

- The dataset contains information on game titles, platforms, release years, genres, sales in different regions, and user and critic scores.
- Sales data across different regions, critic and user scores, and other relevant game information are included.
- Some missing values are present in the 'Publisher', 'Developer', and 'Rating' fields.

Overview of our data

```
overview = vgsales_data.info()
overview
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6894 entries, 0 to 6893
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Name             6894 non-null   object
 1   Year_of_Release  6894 non-null   int64
 2   Genre            6894 non-null   object
 3   Publisher        6893 non-null   object
 4   NA_Sales         6894 non-null   float64
 5   EU_Sales         6894 non-null   float64
 6   JP_Sales         6894 non-null   float64
 7   Other_Sales      6894 non-null   float64
 8   Global_Sales     6894 non-null   float64
 9   Critic_Score     6894 non-null   int64
 10  Critic_Count     6894 non-null   int64
 11  User_Score       6894 non-null   float64
 12  User_Count       6894 non-null   int64
 13  Developer        6890 non-null   object
 14  Rating           6826 non-null   object
dtypes: float64(6), int64(4), object(5)
memory usage: 808.0+ KB
```

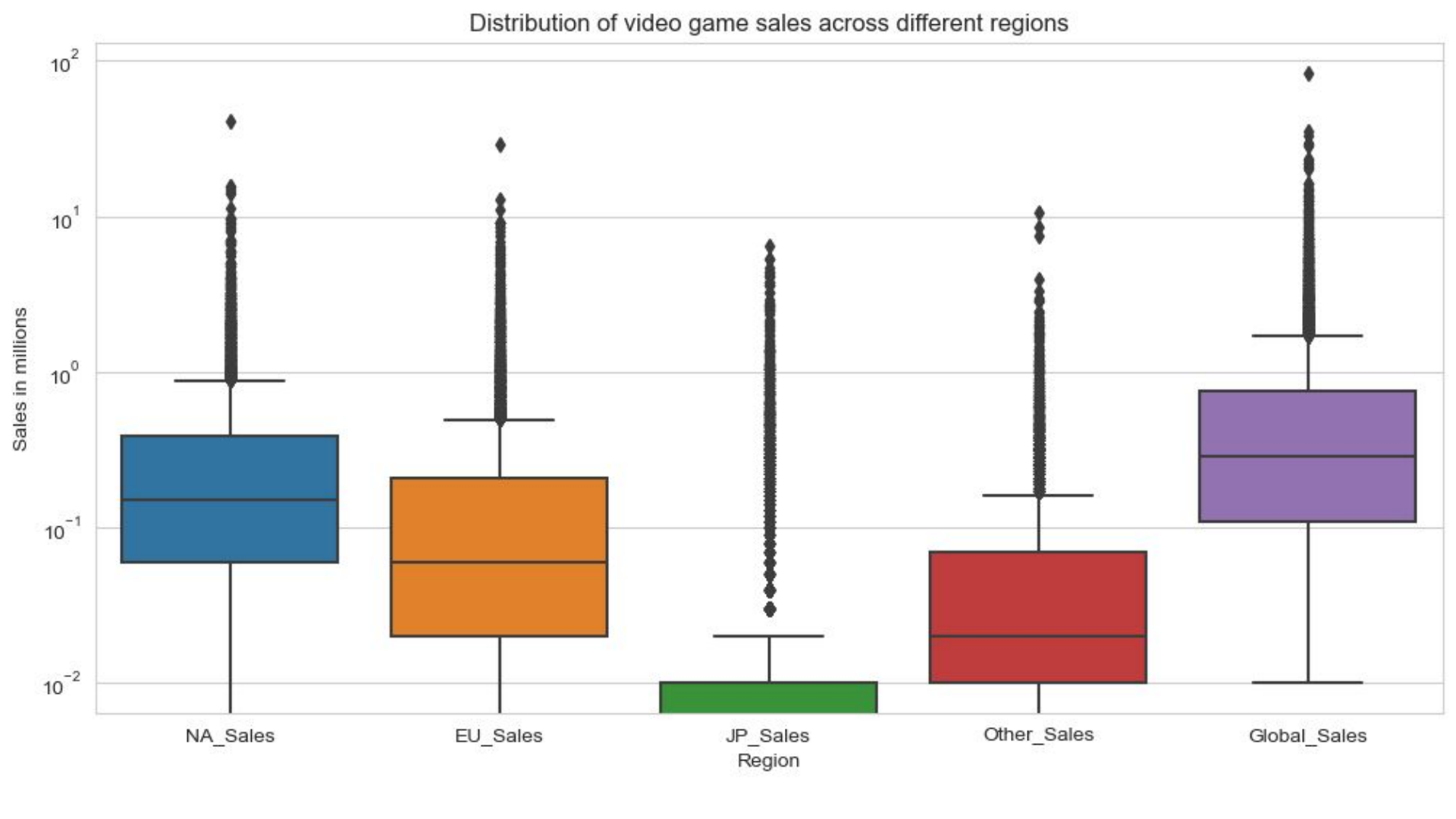
The dataset contains 6,894 entries and 15 columns, with various data types including strings, integers and decimals.

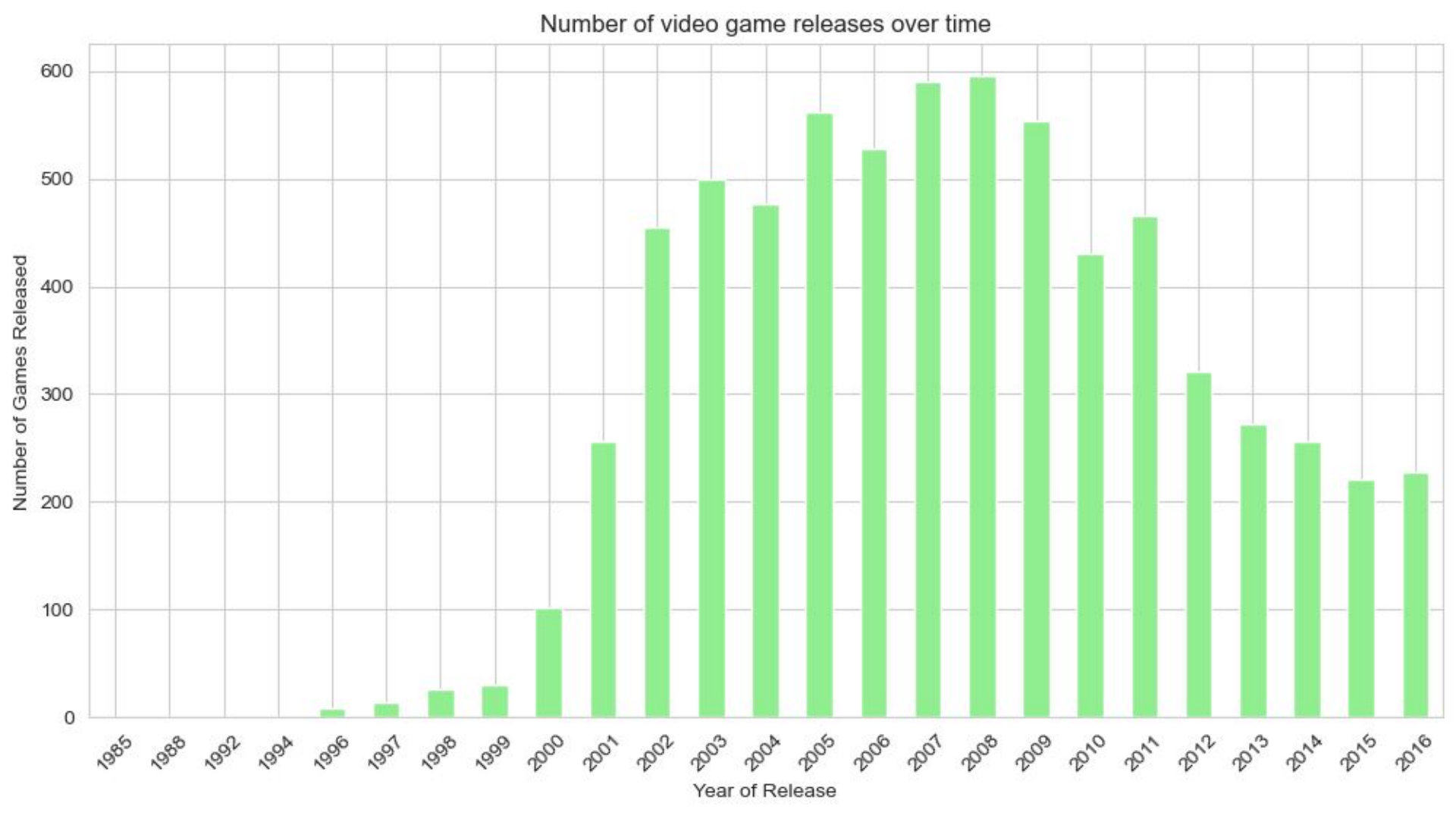
```
missing_values = vgsales_data.isnull().sum()
missing_values
```

```
Name                0
Year_of_Release     0
Genre               0
Publisher           1
NA_Sales            0
EU_Sales            0
JP_Sales            0
Other_Sales         0
Global_Sales        0
Critic_Score        0
Critic_Count        0
User_Score          0
User_Count          0
Developer           4
Rating             68
dtype: int64
```

There are some missing values in the dataset

- Publisher: 1 missing value
- Developer: 4 missing values
- Rating: 68 missing values


Distribution of video game sales across different regions
Sales in millions
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
NA_Sales
EU_Sales
JP_Sales
Other_Sales
Global_Sales
Region


Number of video game releases over time
Number of Games Released
0
100
200
300
400
500
600
1985
1988
1992
1994
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Year of Release

Data Cleaning and Transformation

Data Cleaning

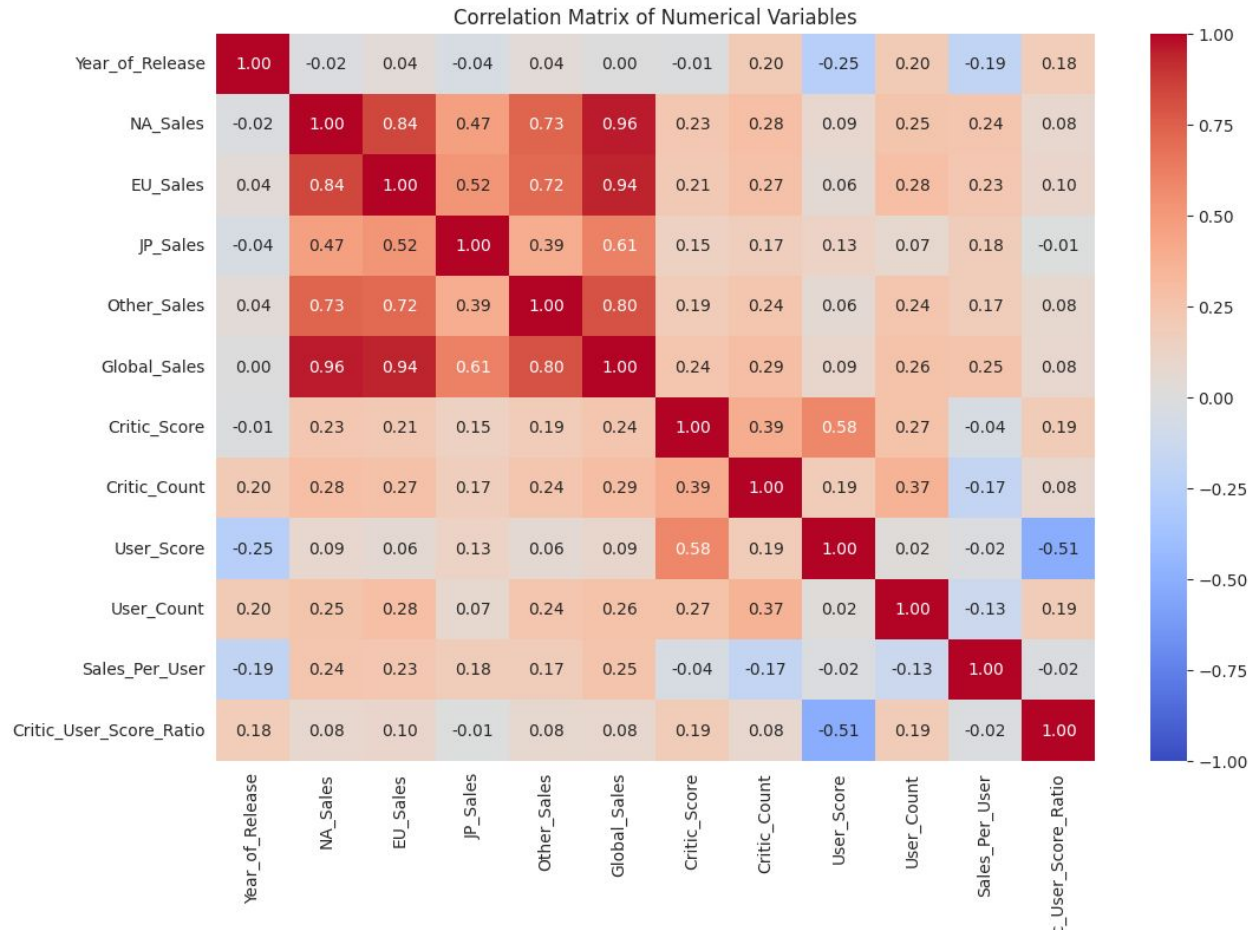
- Handling missing values in the 'Publisher', 'Developer', and 'Rating' fields
- Removing duplicate entries
- Correcting inconsistent data formats

Data Transformation

- Creating additional metrics like 'Sales Per User'
- Standardizing numerical variables
- Converting categorical variables into numerical representations

Descriptive Visualization

Heatmap of Correlation Matrix



Key Findings:

- *Global Sales and Regional Sales:*
  There are strong positive correlations between global sales and sales in individual regions like North America, Europe, Japan, and others

- *Critic Score and User Score:*
  Moderate positive correlation between critic scores and user scores, indicating that games rated highly by critics also tend to be rated highly by users.

- *Sales and Review Counts:*
  Popular games, those with higher sales, tend to get more reviews from both critics and users.

- *Critic to User Score Ratio:*
  The difference between critic and user scores doesn't strongly depend on other factors in the dataset since the Critic_User_Score_Ratio has low correlation with most other variables.

Descriptive Visualization

Bar Plot of Game Releases by Genre

Key Findings:

- *Action and Sports:*

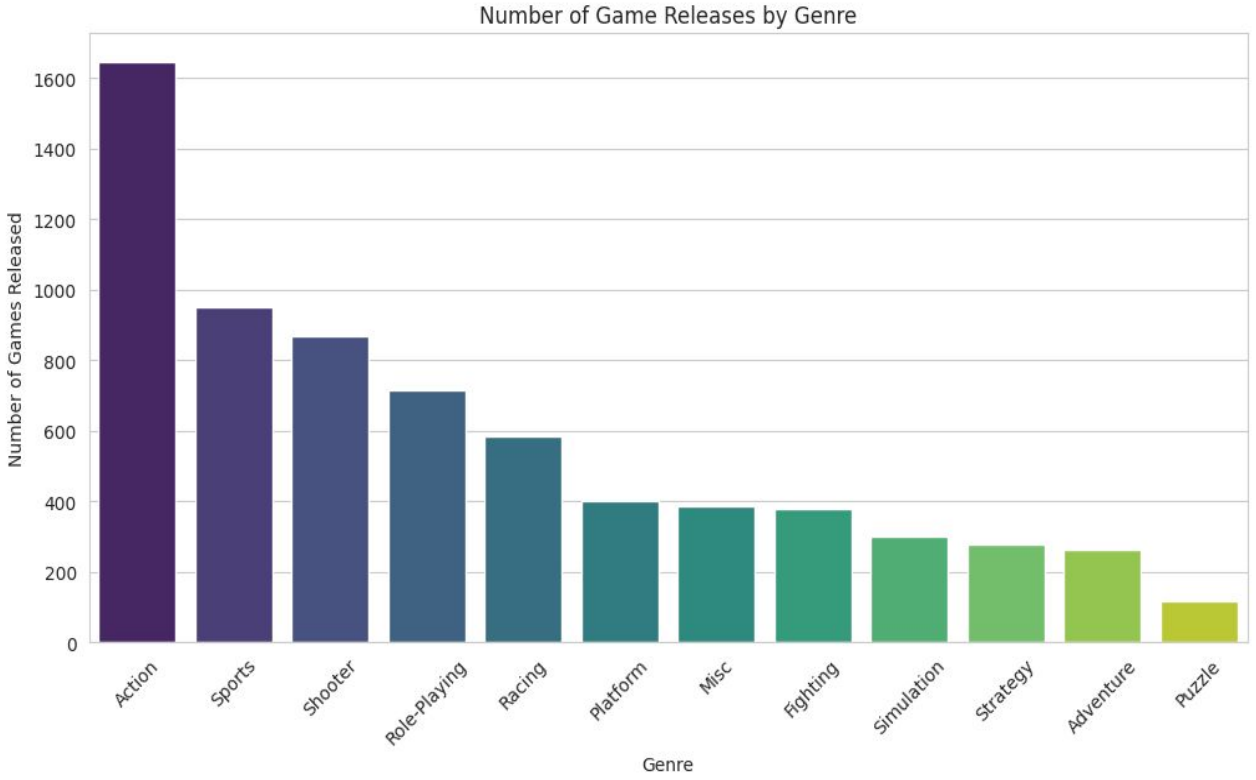
  These genres have the highest number of releases, indicating their popularity among developers and publishers.

- *Role-Playing and Shooter:*

  These genres also have a significant number of releases, showcasing their appeal to a broad audience.

- *Puzzle and Strategy:*

  These genres have fewer releases compared to others, which might reflect a more niche audience or specific gameplay preferences



Descriptive Visualization

Scatter Plot of Global Sales vs. User Score

Key Findings:

- *Broad Distribution:*

  Games with a wide range of user scores have achieved varying levels of global sales. There isn't a clear linear relationship between user scores and sales

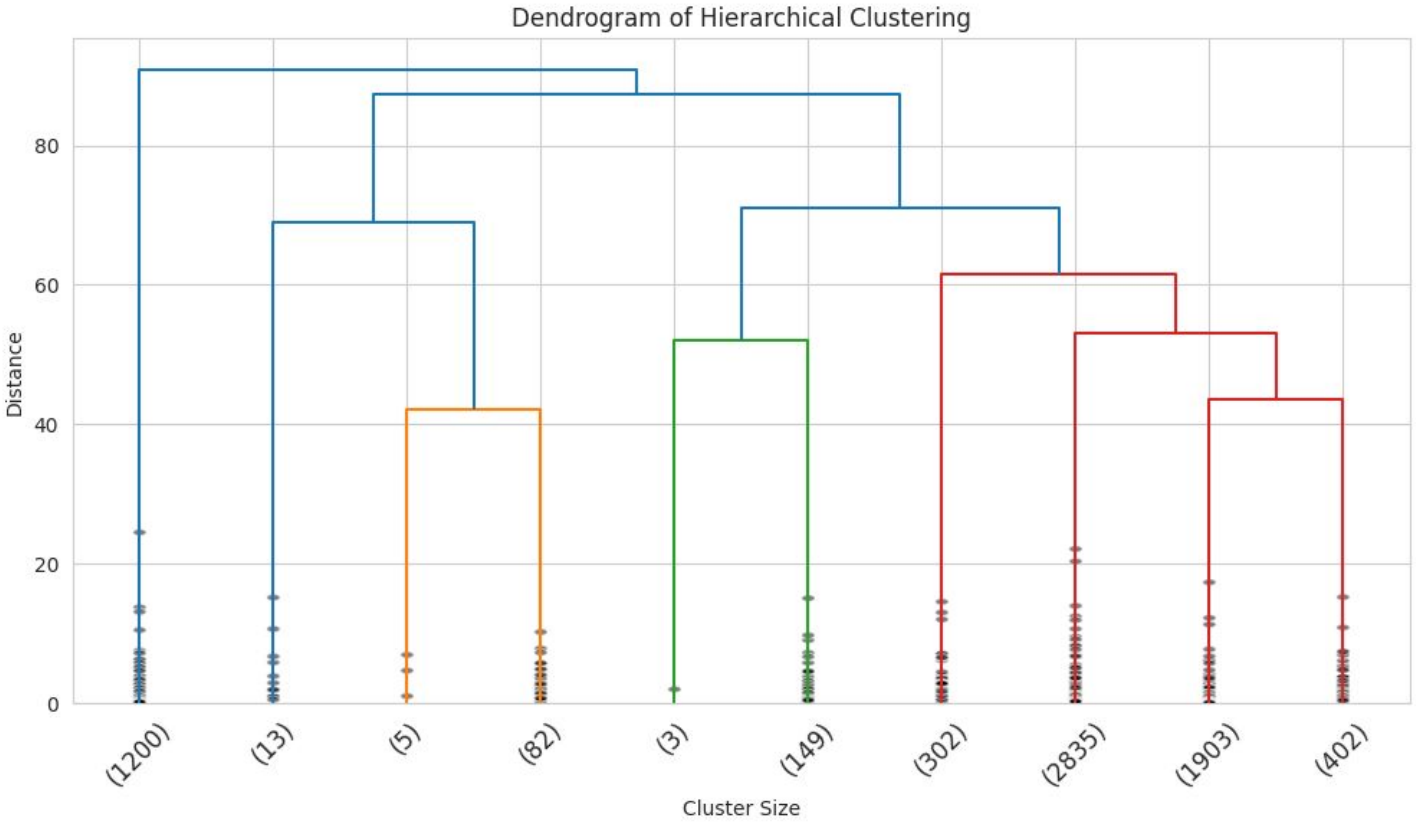
- *High Sales at Different Scores:*

  There are games with both high and low user scores that have achieved high global sales, indicating that user scores are not the sole determinant of sales success.

- *Concentration at Lower Sales:*

  A large number of games, regardless of user score, have lower global sales, which is typical given the competitive nature of the gaming industry.

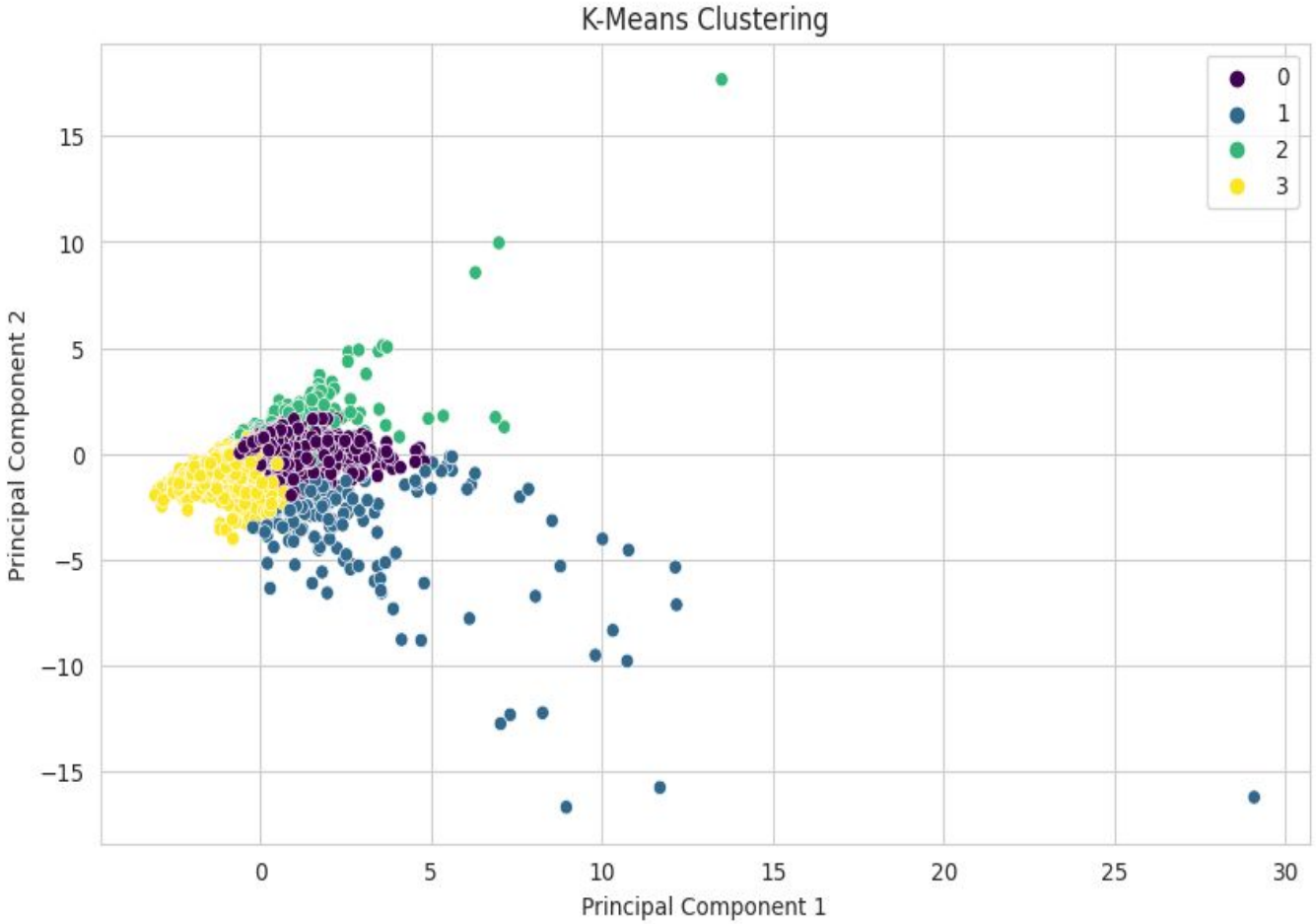
Dendrogram of Hierarchical clustering



Key Findings:

- The longer the line (distance), the more different the groups are from each other.
- The short lines show data points or groups that are very similar.
- Different color branches showing different clusters:
  - Blue branch is the largest cluster.
  - Orange and Green branches are middle-sized clusters.
  - Red branch is the smallest cluster.
- The numbers at the bottom are the size of each cluster.

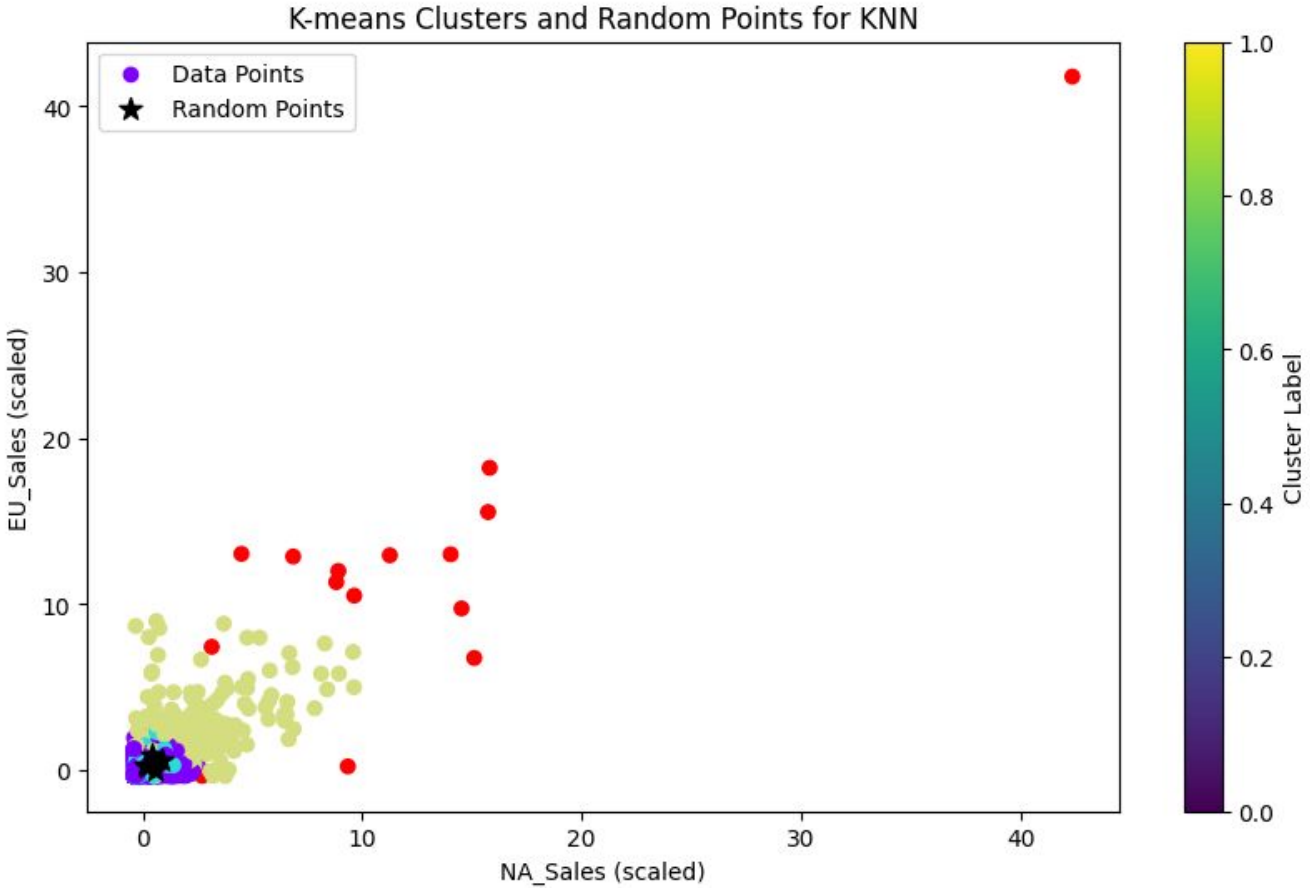
K-Means Clustering



Key Findings:

- **Cluster 0 (Green):** There are a few green dots at the top of the plot, meaning these data points are different from the rest.
- **Cluster 1 (Purple):** This is the biggest cluster in the center. A lot of data points are here. This suggests these points have many shared features.
- **Cluster 2 (Yellow):** This group is near the purple one but towards the top-left. The line between yellow and purple is clear, so they're different but maybe related or similar.
- **Cluster 3 (Blue):** These are spread at the bottom. They go more left to right. They're different from the middle ones (purple and yellow).
- The big group in the middle (purple and yellow) means many of data points are kinda similar.
- The blue ones at the bottom are different in some way. It means they have something in common that the others don't.
- The green ones at the top are unique.

kNN Function



Key Findings:

On the x-axis, we have "NA_Sales (scaled)", and on the y-axis, we have "EU_Sales (scaled)".

- *Data Points:* There's a big cluster of data points at the bottom left corner. This means most sales values are lower and close to each other for both EU and NA.
- *Random Points:* These are shown as red dots. They're spread across the graph, but many are in the middle of the x-axis.
- *Colors:* The color gradient, from purple to green to yellow, represents different clusters or groups. It looks like the data points have been grouped into various clusters based on their similarity.
- *Cluster Label:* The right side has a color bar that shows cluster labels from 0 to 1.

Summary

- Game companies can use the analysis to develop more competitive strategies and better-targeted games.
- Data-driven decision-making can lead to increased sales and market presence for game companies.
- Understanding regional preferences and sales trends is crucial for success in the gaming industry.
- Hierarchical clustering was employed to understand the relationships between various video games based on sales and scores.
- The dendrogram showed the 'distance' at which different games or game categories were related, which can be useful for market segmentation or targeted marketing strategies.
- K-means was applied to segment the games into different clusters, to representing different market segments and consumer preferences.
- A kNN algorithm was built to classify new games based on their nearest existing counterparts in the feature space.