ALLIANCE UNIVERSITY

Private University established in Karnataka State by Act No.34 of year 2010
Recognized by the University Grants Commission (UGC), New Delhi

Project Report

Bachelor of Computer Applications

Semester II

Statistics to Data Science

By:

Dakshraj Singh Chandawat
Reg No.- 2411021240050
Class- BCA-A

Department of Computer Applications

Alliance University

Chandapura-Anekal Main Road, Anekal

Bengaluru-562106

April 2025

Comprehensive Analysis of Video Game Sales Data

Introduction

About the Dataset

The dataset employed in this project captures a wide spectrum of information related to the global video game industry. It encompasses data from 1980 to 2020 and includes key attributes for each video game entry such as the game title, the platform on which it was released, the year of release, genre classification, and publisher. Most importantly, it provides detailed numerical values for regional sales across North America (NA), Europe (EU), Japan (JP), and other regions, along with the calculated global sales.

The richness of this dataset makes it an ideal candidate for conducting in-depth data analysis. It serves as a lens to observe how consumer interests, platform evolution, and genre popularity have transformed the video game landscape over four decades. By analyzing the patterns embedded in the data, we can derive significant insights that inform current and future trends in game development and marketing.

Background Context

The video game industry is one of the most rapidly evolving and financially significant segments of the global entertainment market. From pixelated arcade games to lifelike virtual simulations, the journey has been nothing short of revolutionary. With the widespread availability of home gaming consoles, handheld devices, and PCs, gaming has become a mainstream activity enjoyed by diverse demographics worldwide. This shift has brought about intense competition among developers and publishers, each striving to secure a strong foothold in the market.

Understanding the driving forces behind game sales — such as which genres people love, which platforms are most profitable, and which regions contribute the most to global revenues — is vital for stakeholders ranging from game designers to business strategists.

Project Overview

This project presents a comprehensive analytical exploration of a global video game sales dataset using Python-based data science tools. The focus lies in applying effective data preprocessing, exploratory data analysis (EDA), and visual storytelling to reveal hidden trends, behavioral patterns, and regional preferences in the gaming industry.

We begin by inspecting and cleaning the dataset to ensure accuracy and completeness. This is followed by visual exploration using libraries such as Seaborn and Matplotlib to uncover trends in sales and releases. Furthermore, statistical techniques are applied to understand correlations between different sales regions and identify the top-selling entities across genres, platforms, and publishers.

The strength of this project lies not only in identifying which games or companies succeeded but also in understanding why they did so. This understanding provides a framework that can guide future innovations and marketing decisions in the gaming sector.

Project Goal

The overarching aim of this project is to use a data-driven approach to dissect and understand the core components of the video game market. By leveraging structured sales data, we aim to:

- Uncover the most influential genres, platforms, and publishers based on total and regional sales.
- Evaluate the historical growth and decline of the industry in terms of sales and game releases.
- Examine the extent to which sales in different regions correlate with one another and with global sales.
- Identify the most commercially successful video games of all time and explore the reasons behind their success.
- Derive valuable business insights and actionable recommendations that can serve as a strategic guide for stakeholders in the gaming industry.

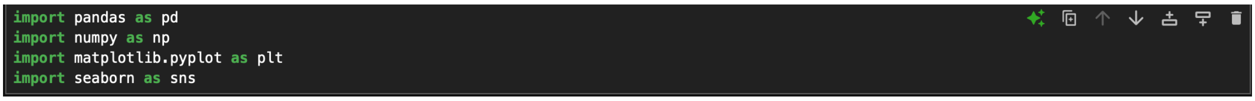
Objectives

To accomplish the above goals, the project is structured around the following detailed objectives:

1. **Data Acquisition and Cleaning:** Read the dataset, understand its structure, and remove any incomplete or inconsistent records to ensure clean and usable data.
2. **Preliminary Analysis:** Use summary statistics to gain an initial understanding of the data's scope and scale.
3. **Genre-Level Analysis:** Identify the frequency of different genres and assess their contribution to global and regional sales.
4. **Platform-Level Insights:** Determine which platforms have generated the highest sales and hosted the most successful games.
5. **Publisher Performance:** Analyze which publishers have dominated the market based on global sales figures.
6. **Time Series Analysis:** Track how the number of game releases and total sales have changed over time.
7. **Correlation Evaluation:** Use statistical correlation to explore relationships among sales figures across different regions.
8. **Visualization:** Employ data visualization techniques to effectively communicate findings.
9. **Derive Conclusions:** Synthesize insights to support decision-making and future forecasting.

Code with Explanations:

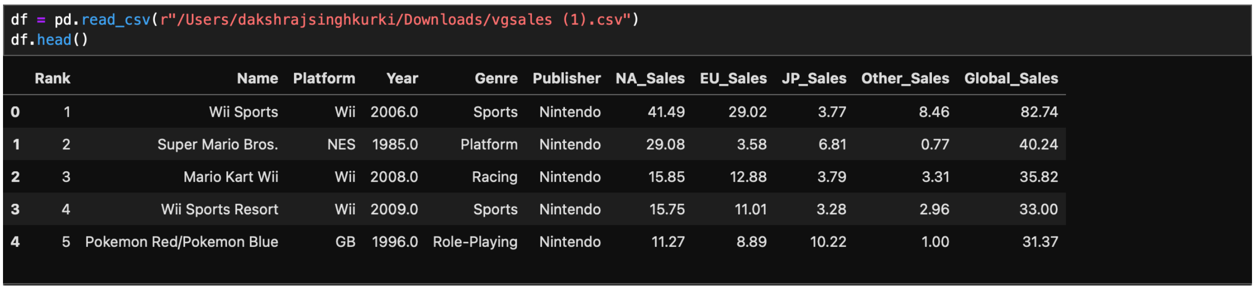
1.Importing Libraries:



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

These are essential Python libraries for data manipulation and visualization. pandas is used to load and structure the dataset, numpy handles numerical operations, matplotlib.pyplot is used for plotting, and seaborn creates aesthetically pleasing and informative graphics.

2.Loading the Dataset:

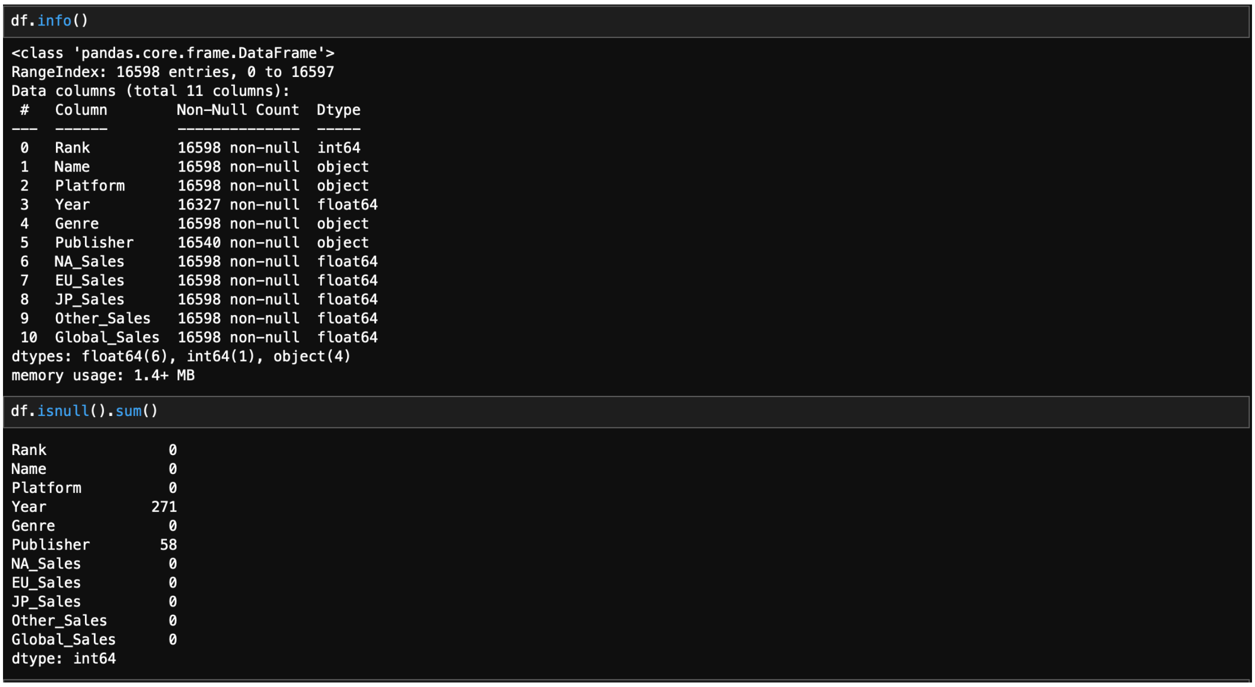


```
df = pd.read_csv(r"/Users/dakshrajsinghkurki/Downloads/vgsales (1).csv")
df.head()
```

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37

The dataset is loaded from a CSV file using `pandas.read_csv()` and displayed using `.head()` to preview the first five rows. This helps in understanding the structure and content of the dataset.

3. Exploring the Data Structure:



```
df.info()
```

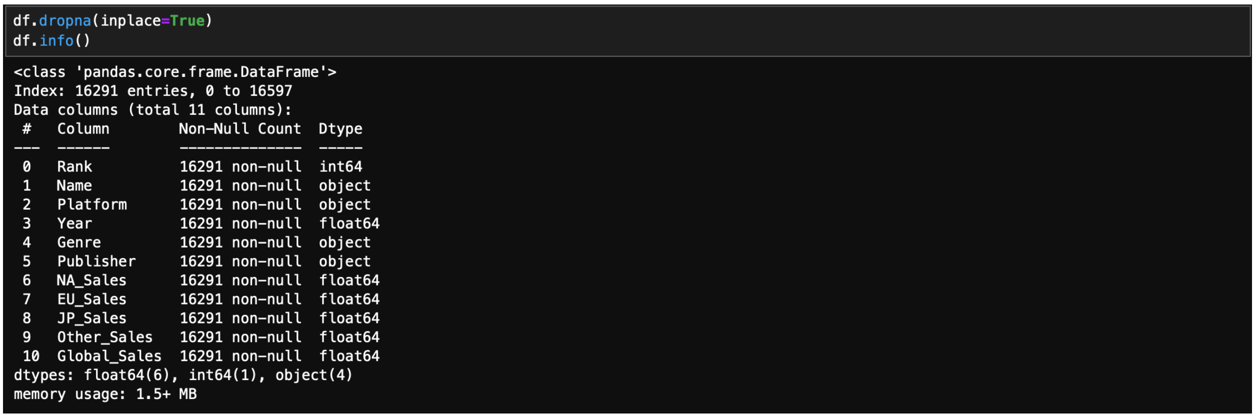
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16598 entries, 0 to 16597
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Rank          16598 non-null  int64  
 1   Name          16598 non-null  object 
 2   Platform      16598 non-null  object 
 3   Year          16327 non-null  float64
 4   Genre         16598 non-null  object 
 5   Publisher     16540 non-null  object 
 6   NA_Sales      16598 non-null  float64
 7   EU_Sales      16598 non-null  float64
 8   JP_Sales      16598 non-null  float64
 9   Other_Sales   16598 non-null  float64
 10  Global_Sales  16598 non-null  float64
dtypes: float64(6), int64(1), object(4)
memory usage: 1.4+ MB
```

```
df.isnull().sum()
```

```
Rank              0
Name              0
Platform          0
Year            271
Genre             0
Publisher        58
NA_Sales          0
EU_Sales          0
JP_Sales          0
Other_Sales       0
Global_Sales      0
dtype: int64
```

df.info() provides an overview of the data types and non-null counts for each column. df.isnull().sum() calculates the number of missing values in each column. Identifying missing data is crucial for ensuring the quality of analysis.

4. Cleaning the Data:

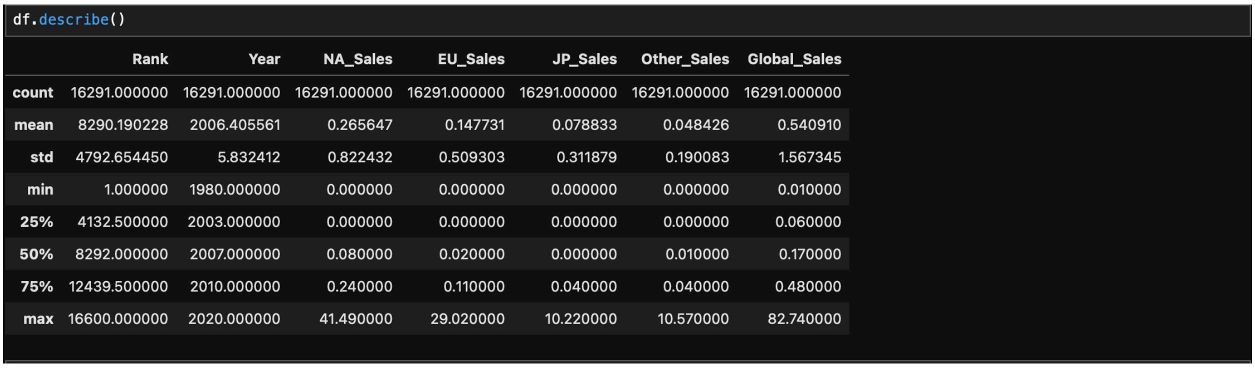


```
df.dropna(inplace=True)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 16291 entries, 0 to 16597
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Rank          16291 non-null  int64  
 1   Name          16291 non-null  object 
 2   Platform      16291 non-null  object 
 3   Year          16291 non-null  float64
 4   Genre         16291 non-null  object 
 5   Publisher     16291 non-null  object 
 6   NA_Sales      16291 non-null  float64
 7   EU_Sales      16291 non-null  float64
 8   JP_Sales      16291 non-null  float64
 9   Other_Sales   16291 non-null  float64
 10  Global_Sales  16291 non-null  float64
dtypes: float64(6), int64(1), object(4)
memory usage: 1.5+ MB
```

Rows with missing values are dropped using `dropna(inplace=True)`. This is a necessary step to maintain the integrity of subsequent analyses and visualizations. The `info()` function is called again to confirm the removal of incomplete entries.

5. Descriptive Statistics:

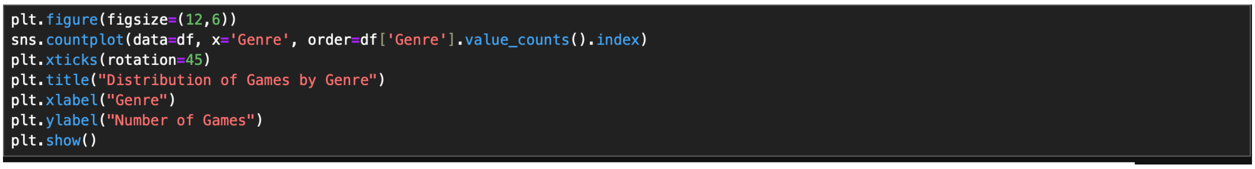


```
df.describe()
```

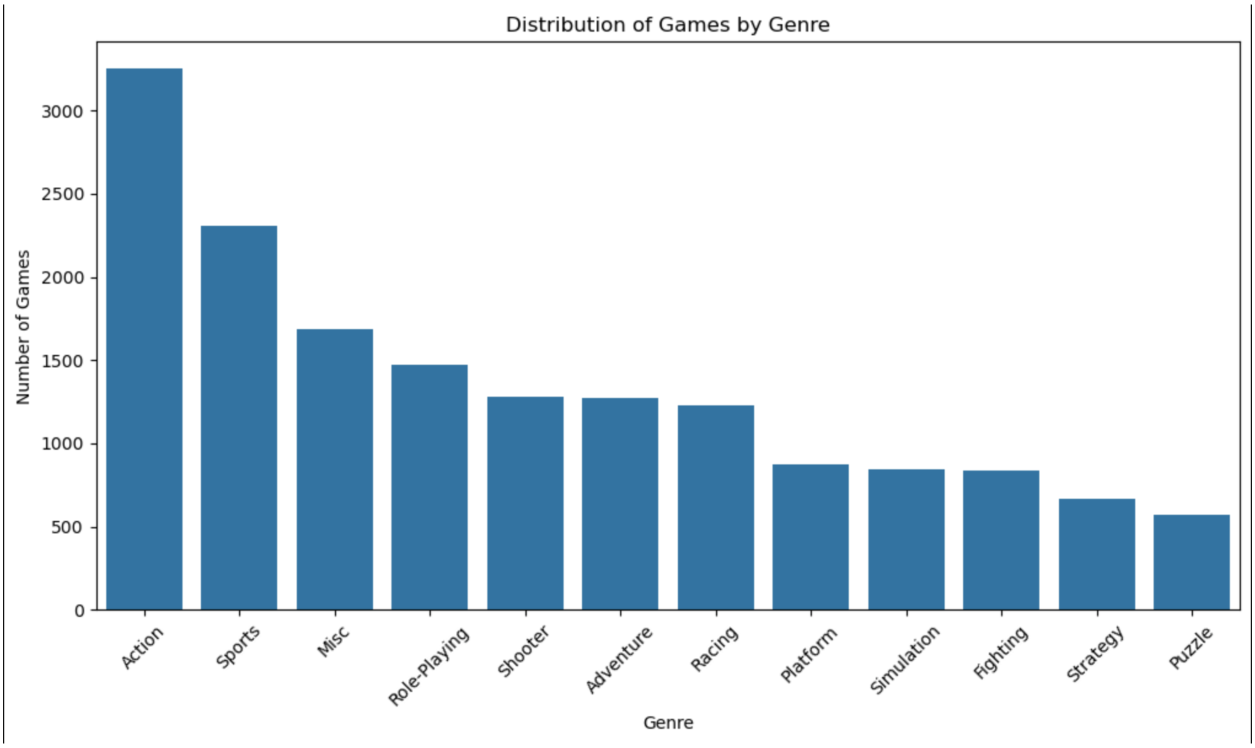
	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
count	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000
mean	8290.190228	2006.405561	0.265647	0.147731	0.078833	0.048426	0.540910
std	4792.654450	5.832412	0.822432	0.509303	0.311879	0.190083	1.567345
min	1.000000	1980.000000	0.000000	0.000000	0.000000	0.000000	0.010000
25%	4132.500000	2003.000000	0.000000	0.000000	0.000000	0.000000	0.060000
50%	8292.000000	2007.000000	0.080000	0.020000	0.000000	0.010000	0.170000
75%	12439.500000	2010.000000	0.240000	0.110000	0.040000	0.040000	0.480000
max	16600.000000	2020.000000	41.490000	29.020000	10.220000	10.570000	82.740000

This command provides summary statistics for all numerical columns, including count, mean, standard deviation, and quartiles. It gives a quick overview of the distribution and spread of values such as sales figures and release years.

6.Genre Distribution:

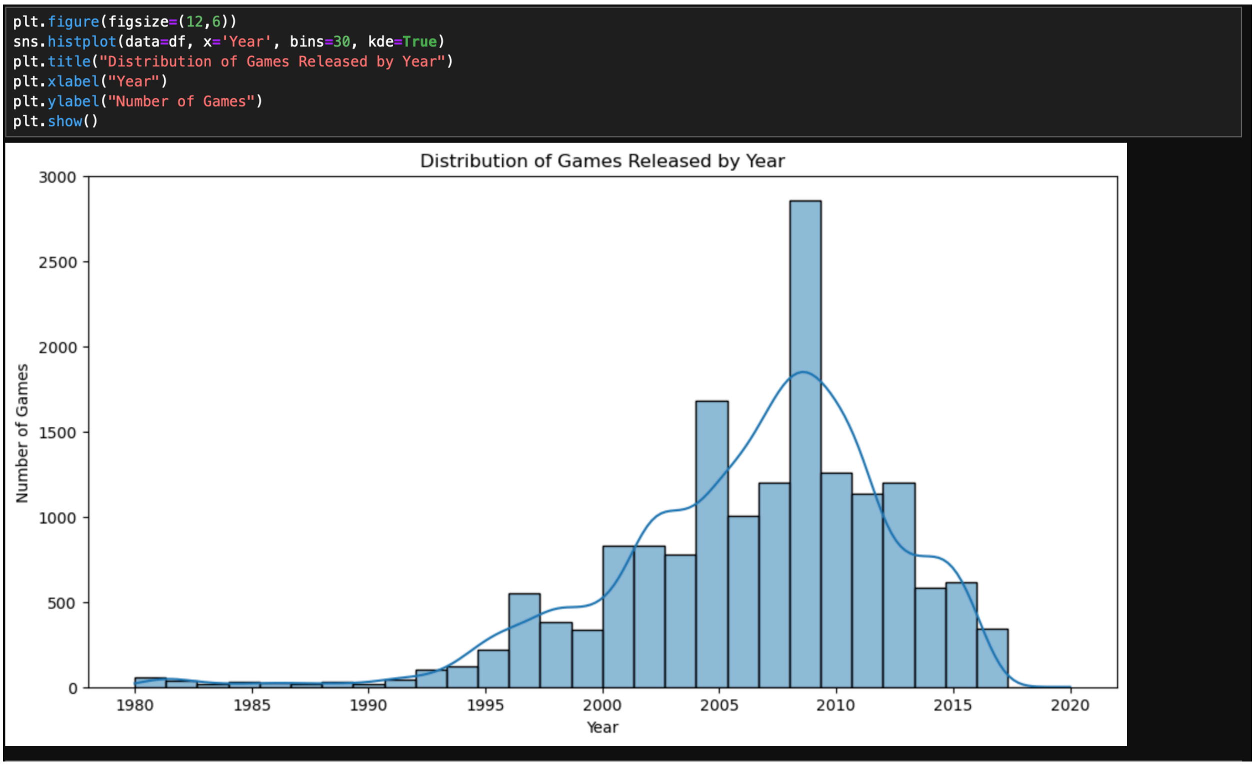


```
plt.figure(figsize=(12,6))
sns.countplot(data=df, x='Genre', order=df['Genre'].value_counts().index)
plt.xticks(rotation=45)
plt.title("Distribution of Games by Genre")
plt.xlabel("Genre")
plt.ylabel("Number of Games")
plt.show()
```



This visualization shows the number of games per genre using a bar chart. Genres are sorted in descending order of frequency. This plot helps identify which types of games are most commonly released.

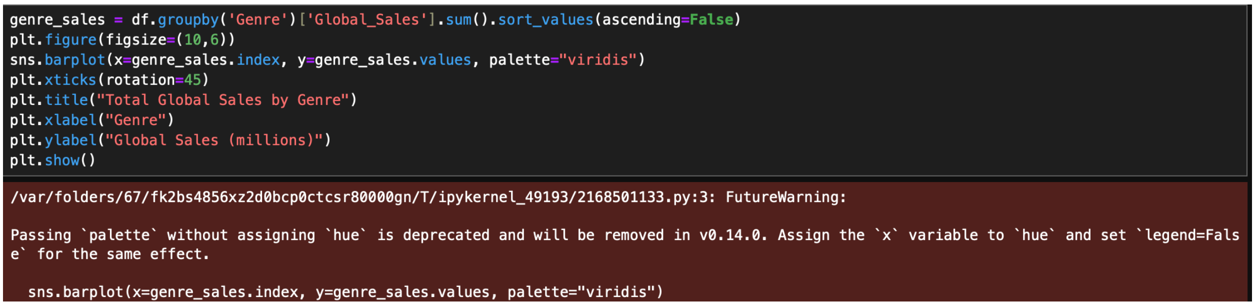
7.Games Released by Year:



```
plt.figure(figsize=(12,6))
sns.histplot(data=df, x='Year', bins=30, kde=True)
plt.title("Distribution of Games Released by Year")
plt.xlabel("Year")
plt.ylabel("Number of Games")
plt.show()
```

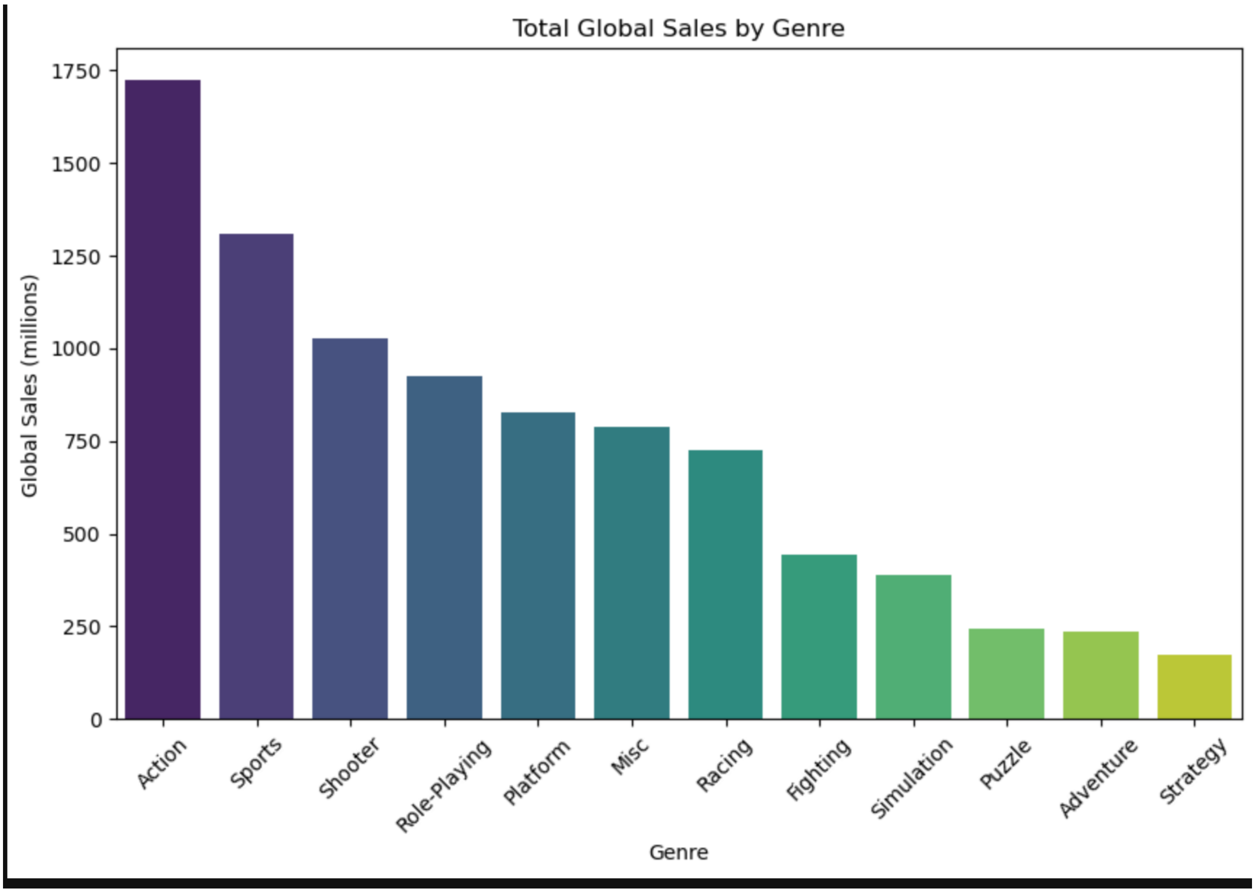
A histogram shows the distribution of game releases over the years. The KDE (Kernel Density Estimation) overlay provides a smoothed representation of the distribution, helping identify trends in industry growth.

8.Total Global Sales by Genre:



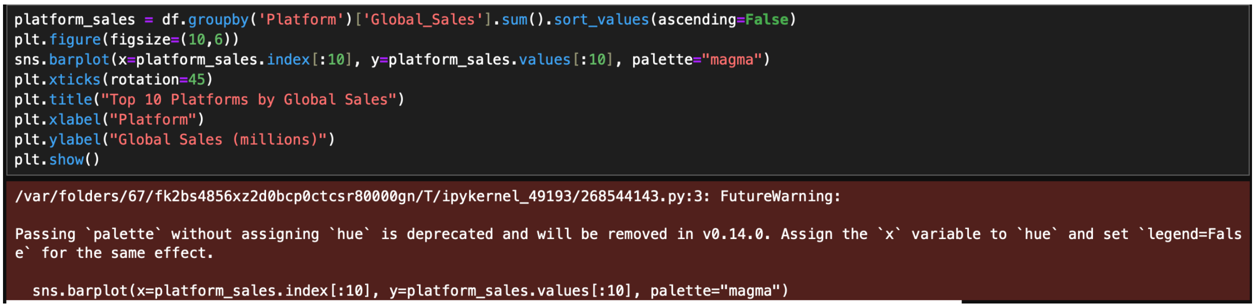
```python
genre_sales = df.groupby('Genre')['Global_Sales'].sum().sort_values(ascending=False)
plt.figure(figsize=(10,6))
sns.barplot(x=genre_sales.index, y=genre_sales.values, palette="viridis")
plt.xticks(rotation=45)
plt.title("Total Global Sales by Genre")
plt.xlabel("Genre")
plt.ylabel("Global Sales (millions)")
plt.show()
```

```
/var/folders/67/fk2bs4856xz2d0bcp0ctcsr80000gn/T/ipykernel_49193/2168501133.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=Fals
e` for the same effect.

  sns.barplot(x=genre_sales.index, y=genre_sales.values, palette="viridis")
```



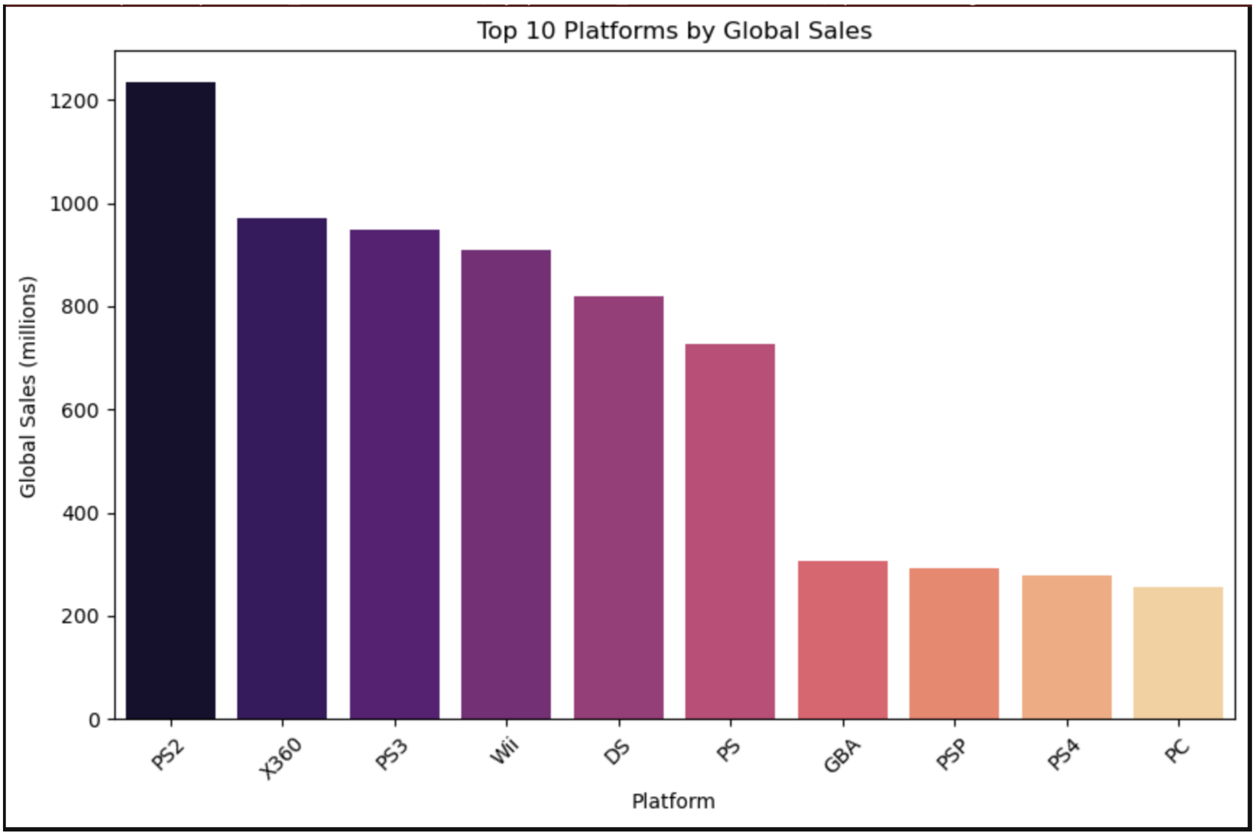
This bar plot shows total global sales per genre. By aggregating sales data, it provides insight into which game genres are most profitable.

9.Platform Sales Analysis:



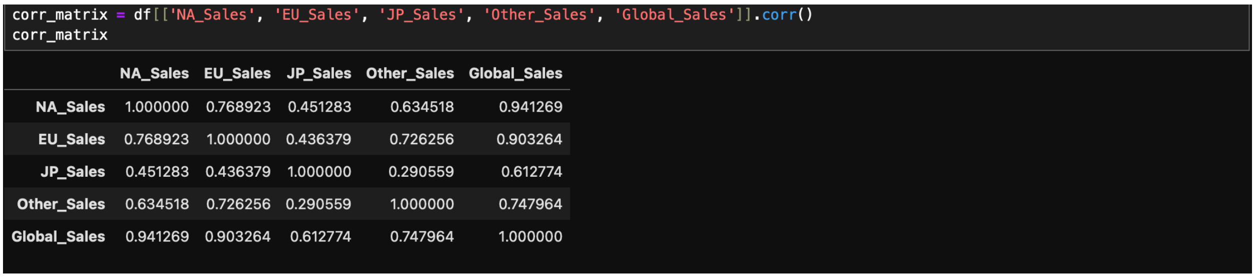
```
platform_sales = df.groupby('Platform')['Global_Sales'].sum().sort_values(ascending=False)
plt.figure(figsize=(10,6))
sns.barplot(x=platform_sales.index[:10], y=platform_sales.values[:10], palette="magma")
plt.xticks(rotation=45)
plt.title("Top 10 Platforms by Global Sales")
plt.xlabel("Platform")
plt.ylabel("Global Sales (millions)")
plt.show()
```

```
/var/folders/67/fk2bs4856xz2d0bcp0ctcsr80000gn/T/ipykernel_49193/268544143.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=Fals
e` for the same effect.

  sns.barplot(x=platform_sales.index[:10], y=platform_sales.values[:10], palette="magma")
```



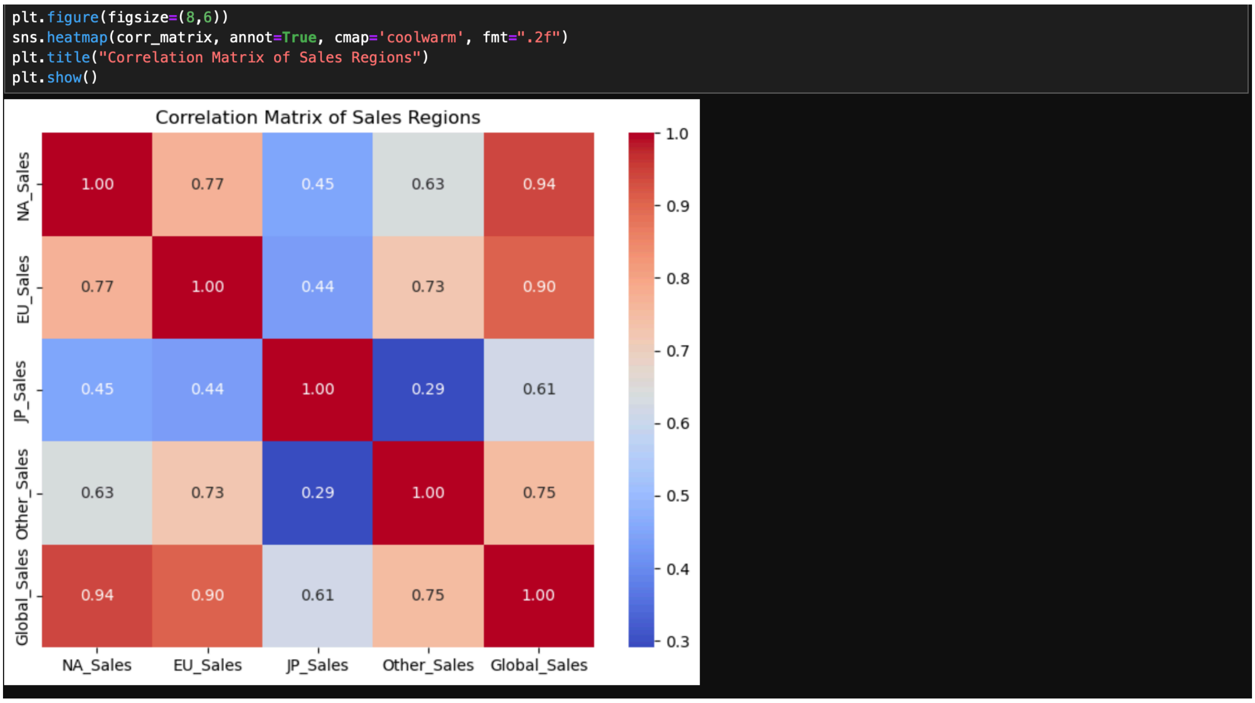
This plot displays the ten most successful gaming platforms in terms of global sales. This is useful for identifying which systems dominated the market.

10.Correlation Matrix:



```
corr_matrix = df[['NA_Sales', 'EU_Sales', 'JP_Sales', 'Other_Sales', 'Global_Sales']].corr()
corr_matrix
```

	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
NA_Sales	1.000000	0.768923	0.451283	0.634518	0.941269
EU_Sales	0.768923	1.000000	0.436379	0.726256	0.903264
JP_Sales	0.451283	0.436379	1.000000	0.290559	0.612774
Other_Sales	0.634518	0.726256	0.290559	1.000000	0.747964
Global_Sales	0.941269	0.903264	0.612774	0.747964	1.000000



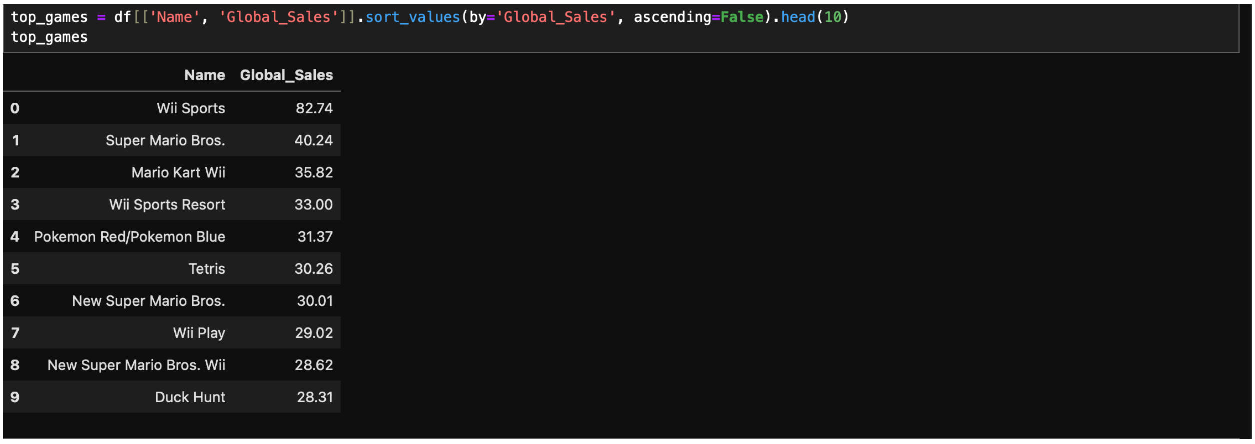
```
plt.figure(figsize=(8,6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Correlation Matrix of Sales Regions")
plt.show()
```

Correlation Matrix of Sales Regions

	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
NA_Sales	1.00	0.77	0.45	0.63	0.94
EU_Sales	0.77	1.00	0.44	0.73	0.90
JP_Sales	0.45	0.44	1.00	0.29	0.61
Other_Sales	0.63	0.73	0.29	1.00	0.75
Global_Sales	0.94	0.90	0.61	0.75	1.00

A correlation matrix with heatmap visualization shows the linear relationships between sales regions and global sales. High values indicate strong correlations, which is useful in understanding market dependencies.

11.Top Selling Games:

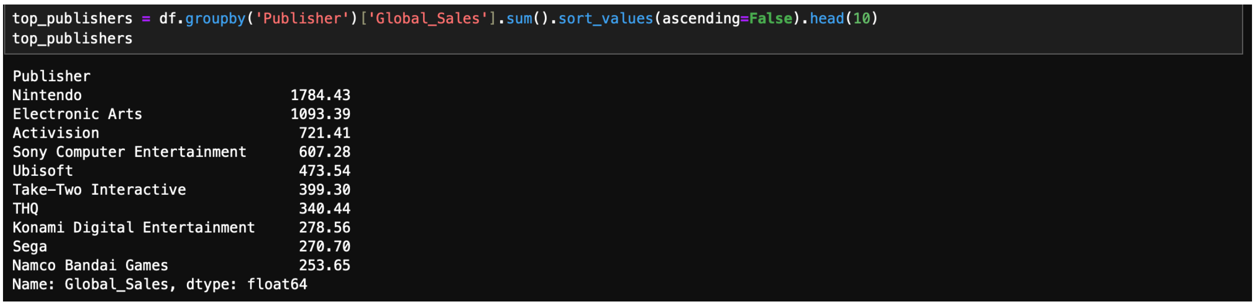


```
top_games = df[['Name', 'Global_Sales']].sort_values(by='Global_Sales', ascending=False).head(10)
top_games
```

```
                       Name  Global_Sales
0                Wii Sports         82.74
1         Super Mario Bros.         40.24
2            Mario Kart Wii         35.82
3         Wii Sports Resort         33.00
4  Pokemon Red/Pokemon Blue         31.37
5                    Tetris         30.26
6     New Super Mario Bros.         30.01
7                  Wii Play         29.02
8  New Super Mario Bros. Wii        28.62
9                 Duck Hunt         28.31
```

This code identifies and prints the top 10 best-selling video games globally. It helps spotlight major hits and their impact on the industry.

12.Top Publishers:

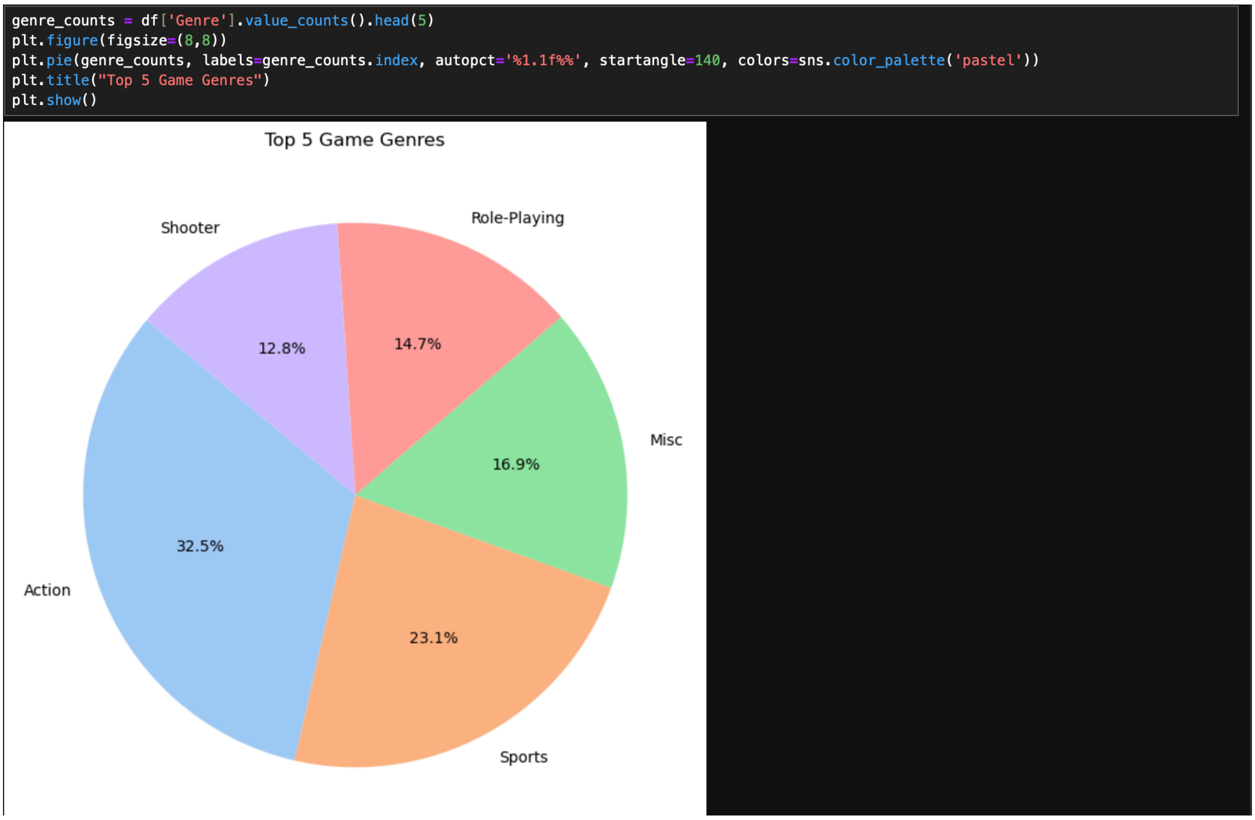


```
top_publishers = df.groupby('Publisher')['Global_Sales'].sum().sort_values(ascending=False).head(10)
top_publishers
```

```
Publisher
Nintendo                        1784.43
Electronic Arts                 1093.39
Activision                       721.41
Sony Computer Entertainment      607.28
Ubisoft                          473.54
Take-Two Interactive             399.30
THQ                              340.44
Konami Digital Entertainment     278.56
Sega                             270.70
Namco Bandai Games               253.65
Name: Global_Sales, dtype: float64
```

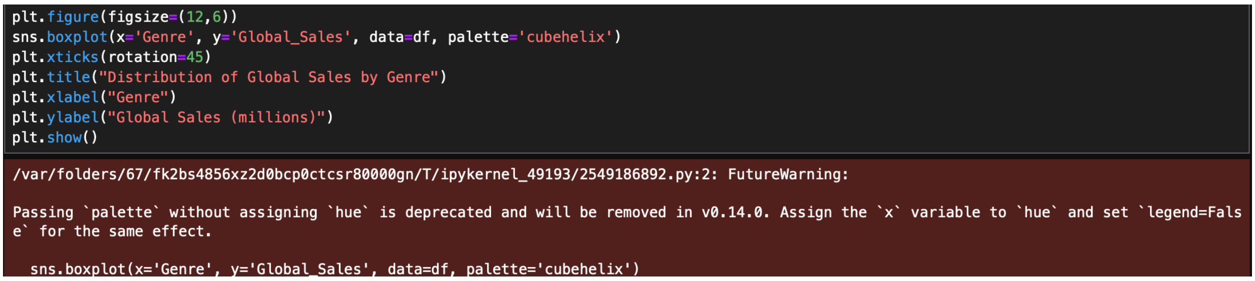
Similar to the top games, this snippet ranks publishers by total global sales. It offers insights into which companies have dominated the gaming market over time.

13.Genre Distribution Pie Chart:



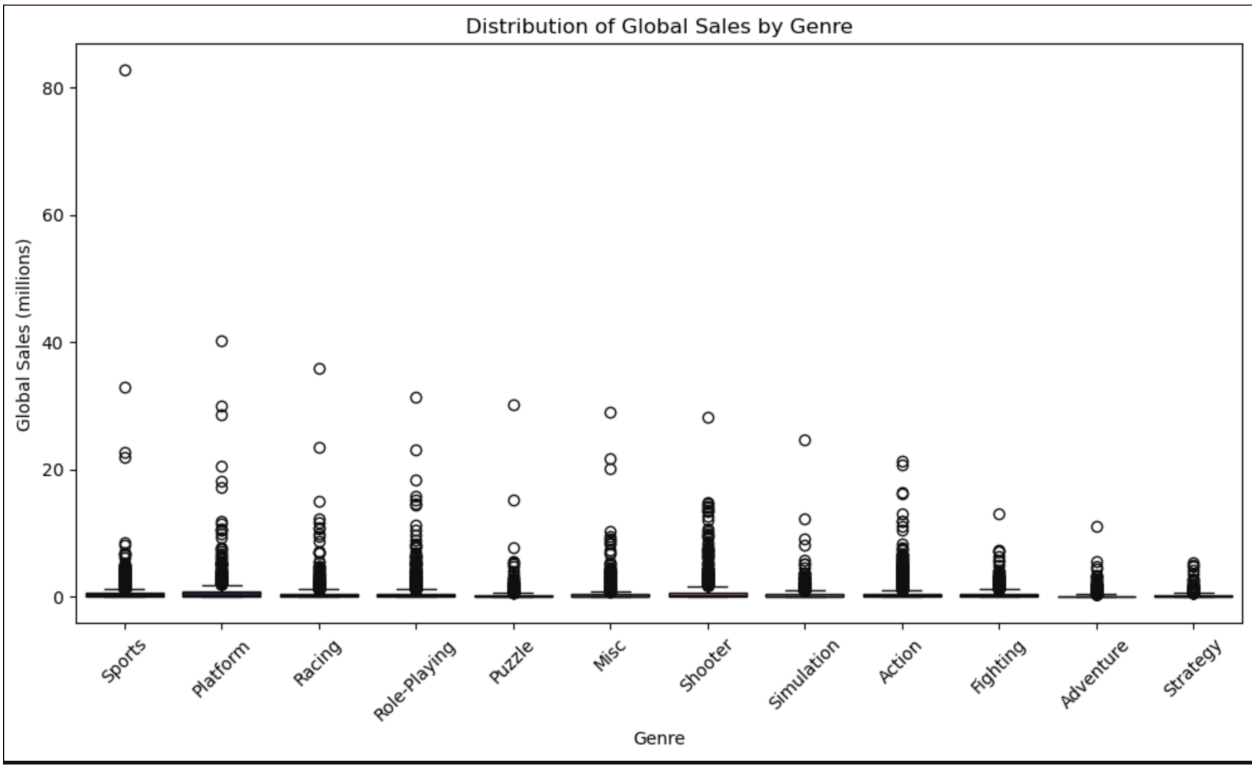
```
genre_counts = df['Genre'].value_counts().head(5)
plt.figure(figsize=(8,8))
plt.pie(genre_counts, labels=genre_counts.index, autopct='%1.1f%%', startangle=140, colors=sns.color_palette('pastel'))
plt.title("Top 5 Game Genres")
plt.show()
```

14.Boxplot of Global Sales by Genre:



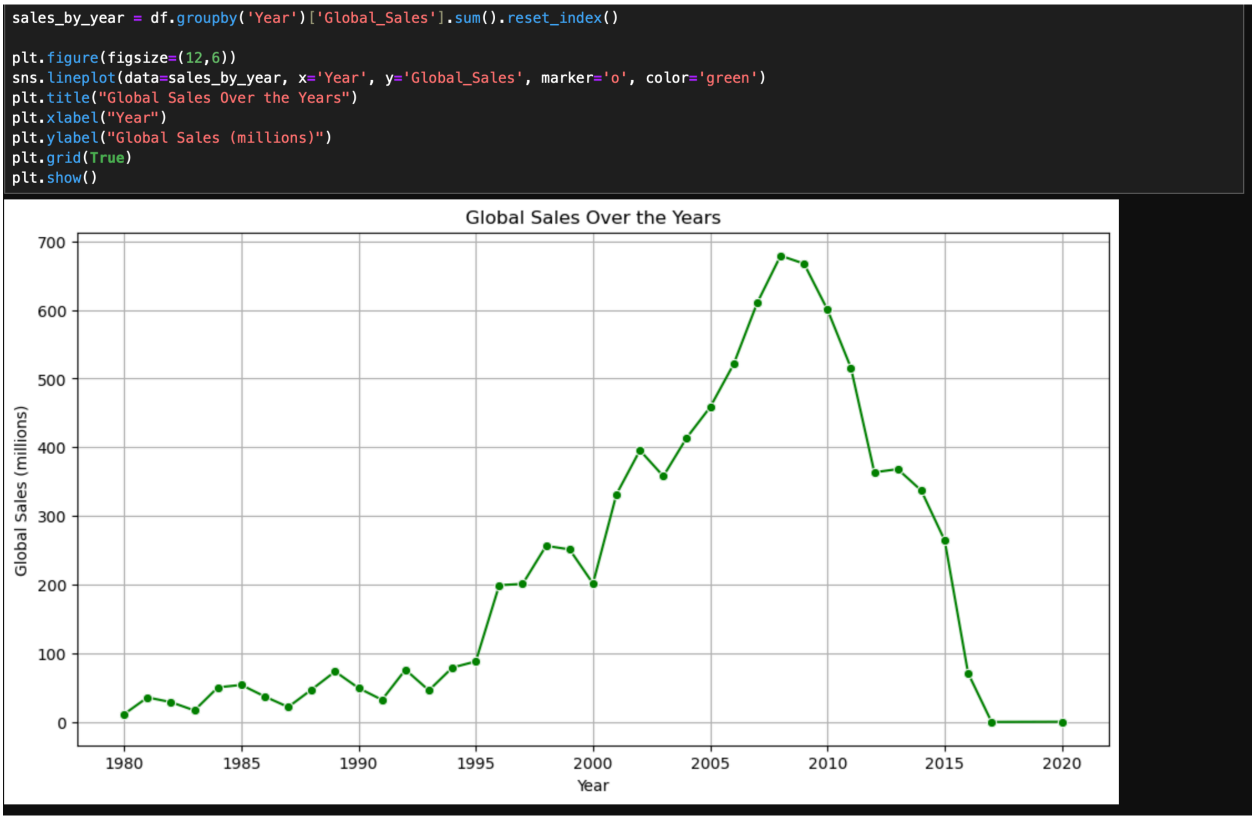
```
plt.figure(figsize=(12,6))
sns.boxplot(x='Genre', y='Global_Sales', data=df, palette='cubehelix')
plt.xticks(rotation=45)
plt.title("Distribution of Global Sales by Genre")
plt.xlabel("Genre")
plt.ylabel("Global Sales (millions)")
plt.show()
```

```
/var/folders/67/fk2bs4856xz2d0bcp0ctcsr80000gn/T/ipykernel_49193/2549186892.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=Fals
e` for the same effect.

  sns.boxplot(x='Genre', y='Global_Sales', data=df, palette='cubehelix')
```



Boxplots help detect the spread and outliers of global sales within each genre. It highlights genres with high variability and exceptional performers.

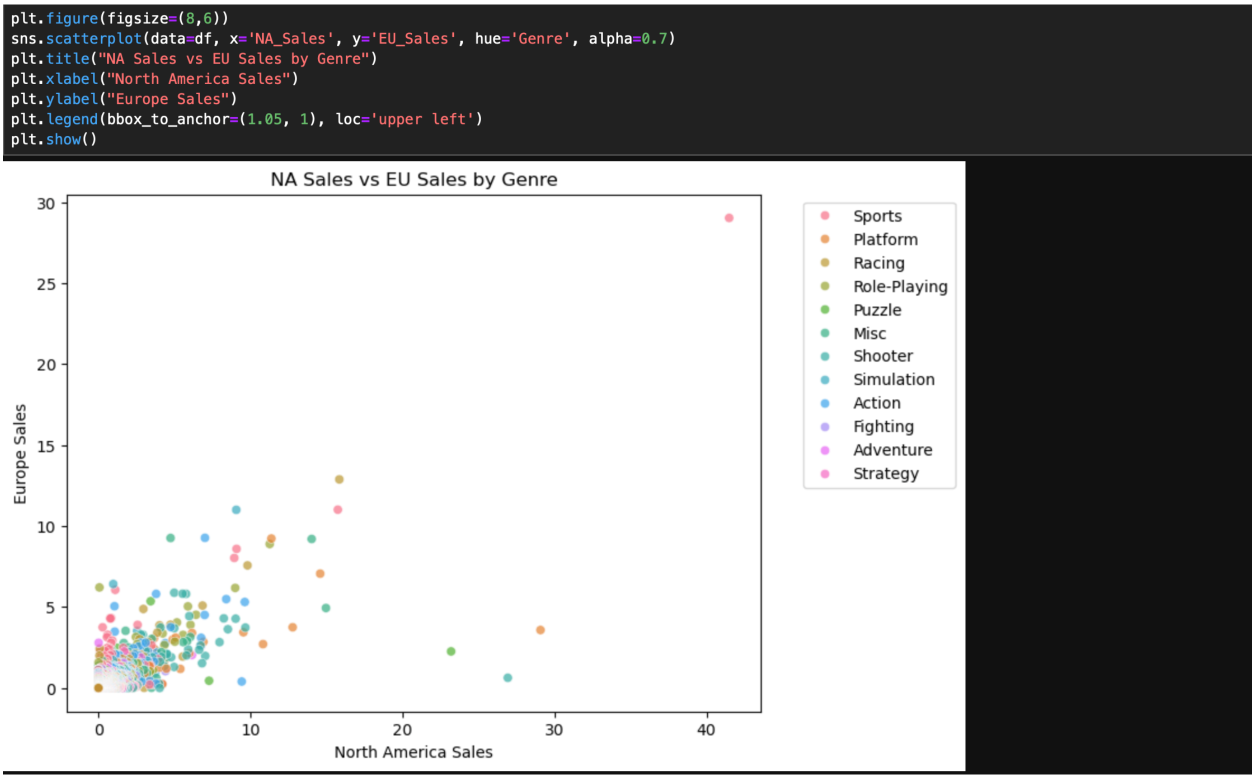
15.Sales Over the Years:



```
sales_by_year = df.groupby('Year')['Global_Sales'].sum().reset_index()

plt.figure(figsize=(12,6))
sns.lineplot(data=sales_by_year, x='Year', y='Global_Sales', marker='o', color='green')
plt.title("Global Sales Over the Years")
plt.xlabel("Year")
plt.ylabel("Global Sales (millions)")
plt.grid(True)
plt.show()
```

This time-series line chart shows the trend in global sales over time. It helps identify peak years and periods of decline or growth in the industry.

16 Regional Sales Comparison:



```
plt.figure(figsize=(8,6))
sns.scatterplot(data=df, x='NA_Sales', y='EU_Sales', hue='Genre', alpha=0.7)
plt.title("NA Sales vs EU Sales by Genre")
plt.xlabel("North America Sales")
plt.ylabel("Europe Sales")
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()
```

Visualizations and Model Evaluation:

Distribution of Games by Genre:

This bar chart reveals the number of games released per genre. The chart shows that the Action genre is the most common, followed by Sports and Miscellaneous. This high frequency indicates strong market demand and developer focus in these areas.

Games Released Over the Years:

The histogram illustrates how the number of video game releases has changed over time. From 1995 to 2009, there is a noticeable upward trend, peaking around 2008–2009. This period reflects a surge in gaming popularity, potentially due to the rise of home consoles and broader internet accessibility.

Total Global Sales by Genre:

A horizontal bar plot visualizes the total global sales across different genres. Shooter and Sports games top the chart, indicating that these genres are not only widely produced but also extremely popular among gamers worldwide.

Top Platforms by Global Sales:

This bar chart displays the top 10 platforms by total global sales. PlayStation, Wii, and Xbox dominate the platform landscape. These consoles have benefited from strong exclusive titles, large player bases, and robust developer ecosystems.

Correlation Matrix of Sales Regions:

A heatmap is used to show correlations between regional sales and global sales. The strongest positive correlation is seen between NA_Sales and Global_Sales (0.94), followed closely by EU_Sales. This highlights how essential North America and Europe are for game success.

Top 10 Best-Selling Games:

A table is generated to show the top 10 best-selling games. Wii Sports leads significantly, followed by classic titles like Super Mario Bros. and Mario Kart Wii. This data emphasizes the importance of brand strength and franchise loyalty.

Top 10 Publishers by Global Sales:

A horizontal bar graph shows cumulative sales for the top publishers. Nintendo stands far ahead, with Electronic Arts and Activision also having substantial influence. These companies have built a strong reputation through consistent releases and marketing.

Pie Chart of Top 5 Genres:

A pie chart highlights the market share of the five most common genres. Action and Sports games dominate, representing the preferences of a wide demographic of players. This visual is effective for understanding the genre distribution at a glance.

Distribution of Global Sales by Genre

Boxplots reveal how sales vary within each genre. While most games have modest sales, each genre contains a few blockbusters with exceptionally high sales. This emphasizes the unpredictable nature of commercial success in gaming.

Global Sales Trend Over the Years

A line graph is used to display the trend of total global sales by year. Sales increased steadily until about 2008 and then began to decline. This could indicate a transition phase in the gaming industry, possibly toward digital distribution or mobile gaming not captured in this dataset.

Regional Comparison: NA vs EU Sales

A scatter plot with color-coded genres compares North American and European sales. The plot reveals a positive linear trend, showing that games popular in NA tend to also perform well in EU. It also helps identify genre-specific regional preferences.

Conclusion and Insights

The video game sales data reveals several key insights into the industry's evolution, regional dynamics, and commercial patterns:

- Nintendo’s Market Leadership: Nintendo has consistently led the market both in terms of individual game success and cumulative publisher sales. Their strategic use of popular franchises and hardware integration has driven their dominance.
- Genre Preferences: Action, Sports, and Shooter genres lead in both the number of games released and total global sales. These genres cater to diverse player interests and often feature high replay value.
- Platform Success: Consoles like PlayStation, Wii, and Xbox have been the most lucrative platforms. This reflects their widespread adoption and strong developer support.
- Regional Sales Trends: North America and Europe are the key contributors to global sales. Games performing well in these regions significantly boost global figures. Japan, while smaller in sales volume, remains vital for certain genres like Role-Playing.
- Sales Trends Over Time: The 2000s marked a golden age for video game releases and sales. Post-2010, although game production continued, total sales declined, possibly due to a shift toward digital and mobile gaming not fully captured in this dataset.
- High Sales Variability: Most games have low sales, while a select few become massive hits. This points to the high-risk, high-reward nature of game development.

Overall, the dataset provides valuable insights for stakeholders in game development, publishing, and marketing. Understanding these patterns can help in making data-driven decisions to maximize future success in the video game industry.

Github link - https://github.com/DakshrajKurki/SDS---Assignment-