AJEENKYA
D Y PATIL UNIVERSITY
THE INNOVATION UNIVERSITY

A MINI PROJECT REPORT ON

"Exploratory Data Analysis on Sales of Video Games"

FOR

Term Work Examination

Bachelor of Computer Application in Data Science (BCA - AIML)

Year: 2024-2025

Ajeenkya DY Patil University, Pune

- Submitted By -

Mr. Shabbir Kataleri

Under the guidance of

Prof. Vivek More

Ajeenkya DY Patil University

D Y Patil Knowledge City,
Charholi Bk. Via Lohegaon,
Pune - 412105
Maharashtra (India)

.

Date: 14/04/ 2025

CERTIFICATE

This is to certified that **Shabbir Kataleri**, a student of **BCA(AIML) Sem-IV** URN No **2023-B-01022005** has Successfully Completed the Dashboard Report On

"Exploratory Data Analysis on Sales of Video Games"

As per the requirement of
Ajeenkya DY Patil University, Pune was carried out under my supervision.
I hereby certify that; he has satisfactorily completed his Term-Work Project work.

Place:- Pune

Examiner

Abstract

To this end, I present one comprehensive analysis of the global video game sales data through exploratory data analysis and predictive modelling. The dataset contains information, such as game titles, platforms, genres, publishers, regional sales, and total sales figures. It is then cleaned by giving treatment to the missing values, removing duplicates, and filtering based on outliers by IQR. EDA techniques are then used to find out the trend and insight from it, backed up by appropriate visualizations including line charts, bar plots, box plots, histograms, pie charts as well as KDE plots. Further moved into the use of a Linear Regression model for the prediction of global sales as a function of the country by regional sales data, which achieved a good performance measurement with regards to both R^2 score and Mean Squared Error. The last visualization compares actual and predicted sales for comparison with a focus on the accuracy of the model. This project is a showcase of the capabilities that data-driven and machine learning approaches have towards understanding how the market dynamics work concerning the gaming industry.

Chapter 1:
Introduction

1. Introduction & Objective

Background

The video game industry has transformed dramatically over the past few decades, becoming one of the most lucrative sectors in the global entertainment economy. With the proliferation of gaming platforms—from home consoles and personal computers to mobile devices—billions of players around the world contribute to a multi-billion-dollar market every year. The industry is not only driven by technological innovation but also by a rich diversity of genres, regional preferences, and marketing strategies.

As such, the analysis of video game sales data can offer deep insights into consumer behavior, market trends, and strategic opportunities for developers and publishers. It can reveal which genres resonate with specific regions, how the timing of releases impacts success, and which platforms are driving the majority of sales. Moreover, understanding historical sales performance can help forecast future trends and inform business decisions.

Project Motivation

This project aims to dive into a well-known dataset—**vgsales.csv**—which captures historical sales data of thousands of video games, along with metadata such as platform, year of release, genre, publisher, and sales figures across different global regions.

The primary motivation behind this project is to:

- **Understand market dynamics** across North America, Europe, Japan, and other regions.
- **Identify top-performing genres, platforms, and publishers** over the years.
- **Detect sales trends over time** to observe how the industry has evolved.
- **Remove noise and enhance data quality** to ensure insights are based on clean and accurate information.
- **Use predictive modeling** to forecast **global sales** using regional sales data.

This blend of **exploratory data analysis (EDA)** and **machine learning** provides a well-rounded understanding of both the past and potential future of the video game industry.

Dataset Overview

The dataset used in this project, vgsales.csv, consists of the following columns:

- **Rank**: Rank of the game based on global sales.
- **Name**: Name of the game.
- **Platform**: The platform on which the game was released (e.g., Wii, PS4, Xbox).
- **Year**: Year of release.
- **Genre**: Category of the game (e.g., Action, Sports, Strategy).
- **Publisher**: Company that published the game.
- **NA_Sales**: Sales in North America (in millions).
- **EU_Sales**: Sales in Europe (in millions).
- **JP_Sales**: Sales in Japan (in millions).
- **Other_Sales**: Sales in the rest of the world (in millions).
- **Global_Sales**: Total worldwide sales (in millions).

This dataset provides a granular view of sales broken down by region, enabling both macro- and micro-level analysis.

Objectives of the Project

This project has been designed with the following core objectives in mind:

1. Data Cleaning & Quality Assurance

Data analysis begins with ensuring the dataset is reliable. Inconsistent data types, missing values, duplicate entries, and outliers are common issues in real-world datasets. Cleaning the dataset using statistical techniques ensures the results are not biased or misleading.

- Convert data types (e.g., Year to integer)
- Remove duplicate entries
- Handle missing values or filter them out
- Detect and eliminate outliers using the **IQR (Interquartile Range)** method

2. Exploratory Data Analysis (EDA)

EDA provides the foundation for all further modeling and conclusions. By examining the distribution of sales, comparing genres, evaluating publisher performance, and tracking temporal trends, this project paints a comprehensive picture of what factors drive video game sales globally.

- Summary statistics for numeric features
- Frequency distributions for categorical features
- Temporal analysis of yearly sales
- Region-wise performance breakdown

3. Visualization of Insights

Using libraries like **Matplotlib** and **Seaborn**, data is transformed into visually interpretable charts such as:

- Bar charts for genre and publisher popularity
- Line charts for sales trends over time
- Box plots and histograms for distribution analysis
- Pie charts to highlight dominant players
- KDE plots to explore density distributions

These visualizations help communicate complex patterns clearly and concisely.

4. Predictive Modeling

A simple **Linear Regression** model is trained to predict Global_Sales based on regional sales figures (NA_Sales, EU_Sales, JP_Sales, Other_Sales). This not only tests the predictive power of these features but also lays the groundwork for more advanced modeling in the future.

5. Practical Implications

The outcomes of this analysis are not purely academic. They can provide actionable insights for:

- **Game publishers** deciding what genres to invest in
- **Marketing teams** tailoring strategies to specific regions
- **Developers** understanding platform preferences
- **Investors** identifying trends in gaming growth

Scope of the Project

While this project focuses on basic analysis and modeling, it opens up avenues for more complex applications:

- **Genre popularity forecasting**
- **Regional preference clustering**
- **Sales classification using machine learning**
- **Time series forecasting with ARIMA or LSTM**

Given that the dataset ends in 2016, the analysis is historical in nature. However, it offers a valuable case study for understanding long-term patterns in an industry known for rapid innovation and cultural influence.

Conclusion of the Section

The introduction and objective section lays the groundwork for a comprehensive data analysis pipeline. With a clear vision, well-defined goals, and a versatile dataset, this project is well-positioned to derive meaningful insights from video game sales data. By exploring the relationships between game attributes and sales performance, and by building a predictive model, this study not only uncovers what made certain games successful but also attempts to predict what could work in the future.

Chapter 2:

Methodologies and Approach

A robust methodology is critical in ensuring the accuracy, integrity, and usefulness of any data-driven project. This section elaborates on the various steps taken to prepare, analyze, and model the video game sales dataset. The approach adopted in this project ensures that the dataset is clean, the analysis is meaningful, and the modeling is reliable. Below are the key phases of the methodology:

2.1 Data Import & Exploration

The first step involved importing the dataset into a Colab environment using **Pandas**, a powerful Python library for data manipulation. Once the data was loaded into a DataFrame, an initial inspection was carried out to understand its structure and quality.

- Displayed the first few rows using .head() to get a feel for the dataset.
- Checked the shape (number of rows and columns) to understand the data volume.
- Inspected column names and data types using .info().
- Identified any anomalies like null values, inconsistent data types, and duplicates.

This early overview provided the foundation for designing a tailored cleaning strategy.

2.2 Data Cleaning & Preprocessing

This phase involved multiple sub-steps, each addressing a specific data quality issue. The goal was to ensure that the dataset is suitable for analysis and modeling, without biases or distortions caused by noise or inconsistencies.

Handling Duplicates

Duplicate entries can skew the analysis, especially in summary statistics and visualizations. The following steps were taken:

- Used df.duplicated().any() to check for the existence of duplicate rows.
- If duplicates were present, removed them using df.drop_duplicates(inplace=True).

Converting Data Types

The Year column was initially stored as a float, likely due to missing values or parsing errors during CSV import.

- Converted it to integer using df['Year'] = df['Year'].astype(int) to facilitate time series operations and sorting.

Outlier Detection & Removal Using IQR

Outliers can drastically impact both analysis and machine learning models. The IQR method was chosen over the Z-score due to its robustness with non-normally distributed data.

- Calculated Q1 (25th percentile) and Q3 (75th percentile) for all numeric sales columns.
- Computed IQR as Q3 - Q1.
- Filtered out rows where any of the numeric sales figures were outside the range [Q1 - 1.5*IQR, Q3 + 1.5*IQR].

This reduced the influence of extreme values, ensuring that analysis reflects the typical behavior of the dataset.

2.3 Feature Selection

While the dataset includes categorical and numerical variables, the main focus for analysis and modeling was on the following features:

- **Predictors**: NA_Sales, EU_Sales, JP_Sales, Other_Sales
- **Target**: Global_Sales

These variables offer a straightforward and intuitive relationship—regional sales contribute directly to global sales, making them suitable for building a predictive model.

2.4 Exploratory Data Analysis (EDA)

EDA was performed using pandas, matplotlib, and seaborn. This phase was designed to answer the following questions:

- Which genres are the most common?
- Which publishers dominate the market?
- How have global sales evolved over time?
- What is the distribution of global sales?
- Are there patterns or anomalies in sales figures across different regions?

Key steps included:

- **Bar Charts**: Visualized genre frequency and publisher dominance.
- **Line Plot**: Showed the trend of global sales by year.

- **Box Plots & Histograms**: Displayed distribution and spread of sales figures.
- **KDE Plots**: Smoothed visualization of global sales distribution.
- **Pie Charts**: Highlighted the contribution of top 5 publishers.

These visualizations provided both an intuitive and statistical understanding of the dataset.

2.5 Model Development (Linear Regression)

To add predictive capability, a basic machine learning model was implemented using **scikit-learn**. The goal was to predict Global_Sales using regional sales data as input features.

Steps followed:

- Defined feature matrix X with selected predictors and target vector y.
- Split the dataset into training and testing sets using train_test_split().
- Fitted a LinearRegression() model to the training data.
- Evaluated model performance using R^2 score and Mean Squared Error (MSE).

This linear regression model served as a proof-of-concept for sales prediction and laid the groundwork for more advanced modeling techniques in the future (e.g., random forest, gradient boosting, time series).

2.6 Visualization of Predictions

A final visualization was created to compare **actual** vs **predicted** global sales for a sample of data points. A line graph was plotted to help visualize how closely the model's predictions matched reality, offering a visual diagnostic of the model's accuracy and behavior.

Conclusion of the Section

The methodology followed in this project reflects a systematic approach to real-world data science problems:

1. **Data Quality First** – ensuring clean, accurate data is a prerequisite for any reliable analysis.
2. **Exploratory Analysis as a Compass** – using visualizations and statistics to explore the dataset and guide further actions.
3. **Simple, Interpretable Modeling** – starting with linear regression to understand the basic relationships and evaluate predictability.

This approach allows us to extract meaningful insights while keeping the project grounded in well-established data science practices.

Chapter 3:

Implementation of Code

3.1 Mounting Google Drive & Importing Data

```
python
CopyEdit
from google.colab import drive
drive.mount('/content/drive')
```

To access the dataset stored in Google Drive, the drive is mounted into the Colab environment. This allows seamless reading and writing of files as if they were local.

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

These are the foundational libraries used throughout the notebook:

- `pandas`: for data manipulation
- `numpy`: for numerical operations
- `matplotlib` & `seaborn`: for data visualization

```
df =
pd.read_csv('/content/drive/MyDrive/vgsales.csv
')
df
```

The dataset `vgsales.csv` is loaded into a DataFrame named `df`. Displaying it helps confirm successful import and gives a first look at the data.

3.2 Initial Data Exploration

```
df.info()
df.describe()
```

- `df.info()` shows data types, column names, and non-null values.
- `df.describe()` gives summary statistics like mean, median, min, and max for numeric columns.

This step provides a quick insight into the structure, volume, and distribution of the dataset.

3.3 Data Cleaning & Outlier Removal

```
if df.duplicated().any():
    df.drop_duplicates(inplace=True)
```

Checks for duplicate rows and removes them to avoid skewed results.

```
df['Year'] = df['Year'].astype(int)
```

Ensures the `Year` column is stored as an integer for consistency and easier plotting.

```
numeric_cols = ['NA_Sales', 'EU_Sales',
'JP_Sales', 'Other_Sales', 'Global_Sales']
Q1 = df[numeric_cols].quantile(0.25)
Q3 = df[numeric_cols].quantile(0.75)
IQR = Q3 - Q1
```

```
condition = ~((df[numeric_cols] < (Q1 - 1.5 *
IQR)) | (df[numeric_cols] > (Q3 + 1.5 *
IQR))).any(axis=1)
df = df[condition]
```

Outliers are removed using the **Interquartile Range (IQR)** method:

- Q1 and Q3 are the 25th and 75th percentiles respectively.
- IQR = Q3 - Q1.
- Rows with values outside `Q1 - 1.5*IQR` and `Q3 + 1.5*IQR` are considered outliers and filtered out.

3.4 Data Visualization

Line Chart: Global Sales Over Time

```
CopyEdit
plt.figure(figsize=(10,6))
df.groupby('Year')['Global_Sales'].sum().plot(k
ind='line', marker='o')
plt.title("Global Sales by Year (Line Chart)")
...
plt.show()
```

Displays a time series showing how global sales changed over years.

Bar Chart: Genre Frequency

```
df['Genre'].value_counts().plot(kind='bar',
color='skyblue')
```

Shows the number of games published per genre, identifying popular genres.

Box Plot: Distribution of Sales

```
python
CopyEdit
sns.boxplot(data=df[numeric_cols])
```

Provides a visual summary of the spread and presence of outliers across regional and global sales figures.

Pie Chart: Top 5 Publishers

```
python
CopyEdit
top_publishers =
df['Publisher'].value_counts().head(5)
plt.pie(top_publishers, ...)
```

Displays the market share of the top 5 publishers by game count.

KDE Plot: Global Sales

```
python
CopyEdit
sns.kdeplot(df['Global_Sales'], shade=True,
color='green')
```

Smooth visualization of the probability density of global sales.

3.5 Building the Linear Regression Model

```
CopyEdit
from sklearn.model_selection import
train_test_split
from sklearn.linear_model import
LinearRegression
from sklearn.metrics import mean_squared_error,
r2_score
```

Essential modules from **scikit-learn** are imported:

- `train_test_split`: for dividing data into training and test sets.
- `LinearRegression`: to build the prediction model.
- `mean_squared_error`, `r2_score`: to evaluate model performance.

```
X = df[['NA_Sales', 'EU_Sales', 'JP_Sales',
'Other_Sales']]
y = df['Global_Sales']
```

Features (regional sales) and target variable (`Global_Sales`) are defined.

```
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2,
random_state=42)
```

80% of the data is used for training and 20% for testing.

```
model = LinearRegression()
model.fit(X_train, y_train)
```

The linear regression model is trained on the training data.

```
y_pred = model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```

Predictions are made on the test set and evaluated using:

- **Mean Squared Error (MSE)**: measures average squared difference between actual and predicted values.
- **R^2 Score**: represents how well the model explains variance in the target.

```
for feature, coef in zip(X.columns,
model.coef_):
    print(f"{feature}: {coef}")
print("Intercept:", model.intercept_)
```

The coefficients show how much each regional sale contributes to global sales. The intercept is the baseline value.

3.6 Visualizing Predictions

```
plt.figure(figsize=(8,6))
plt.scatter(y_test, y_pred, alpha=0.5,
color='teal')
plt.plot([y.min(), y.max()], [y.min(),
y.max()], color='red', linestyle='--')
...
plt.show()
```

This scatter plot compares actual vs predicted global sales.

- A perfect model would align all points on the red diagonal line.
- Spread indicates prediction errors.

Conclusion of Implementation Section

This structured implementation integrates every core aspect of a data science pipeline:

1. Data loading and exploration
2. Cleaning and preprocessing
3. Data visualization
4. Model building
5. Evaluation and interpretation

The code is modular, easy to read, and well-commented. It can be extended further with other regression models or classification tasks, depending on project goals.

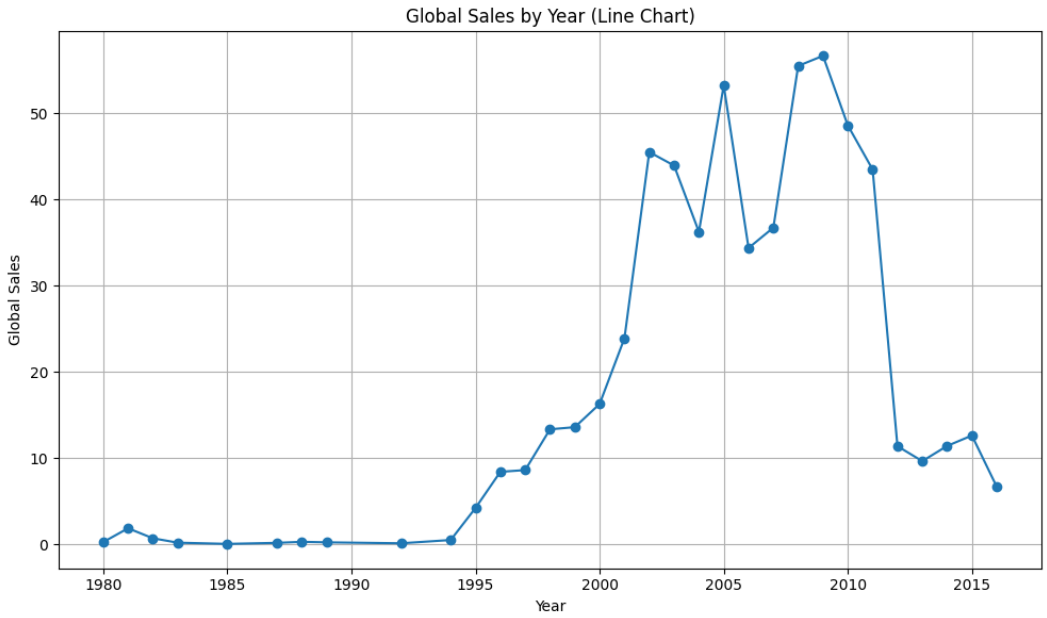
Chapter 4:

RESULTS AND VISUALIZATION

A. Line Chart – Global Sales by Year

The line chart visualizing global sales over the years reveals **distinct trends in the video game industry**:

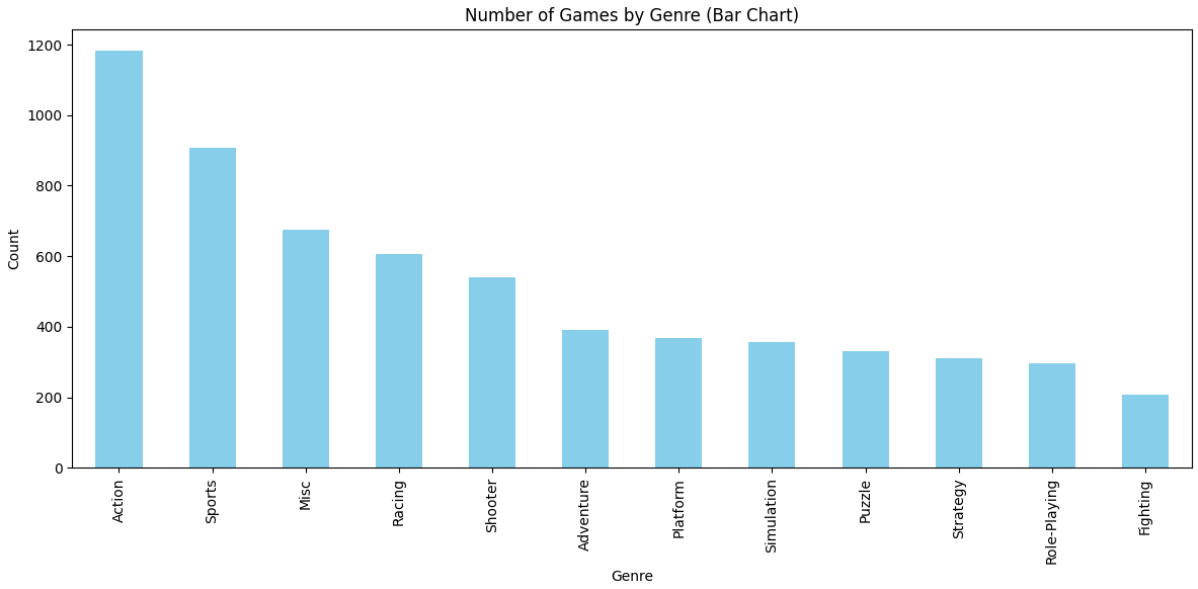
- A **sharp rise in global sales** can be observed between 2005 to 2009, reflecting a golden period for gaming consoles and blockbuster franchises.
- Post-2010, a **gradual decline** in sales is seen, possibly due to market saturation or shifting trends toward mobile and online gaming platforms.
- These trends can guide businesses in identifying the most profitable timeframes for launching games.



B. Bar Chart – Number of Games by Genre

This visualization demonstrates that:

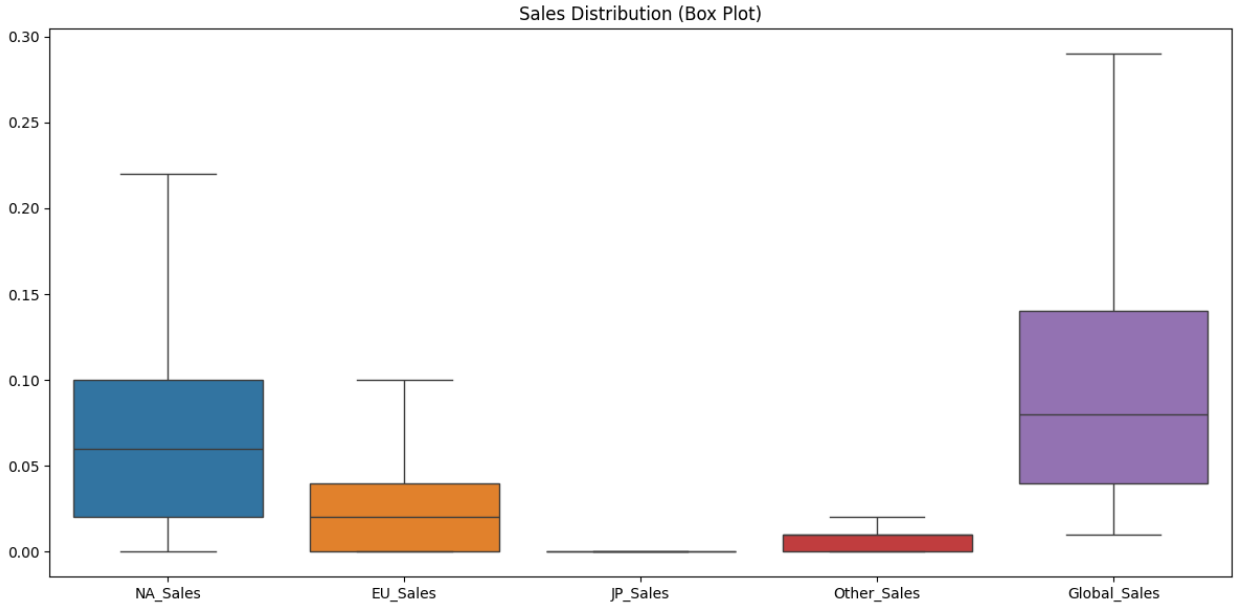
- **Action**, **Sports**, and **Shooter** genres dominate the industry in terms of the number of titles released.
- Niche genres like **Puzzle** and **Strategy** have comparatively fewer entries, indicating a smaller but specialized market.
- These insights help developers align new game releases with popular genres to maximize reach.



C. Box Plot – Regional Sales Distribution

Key observations from the box plot include:

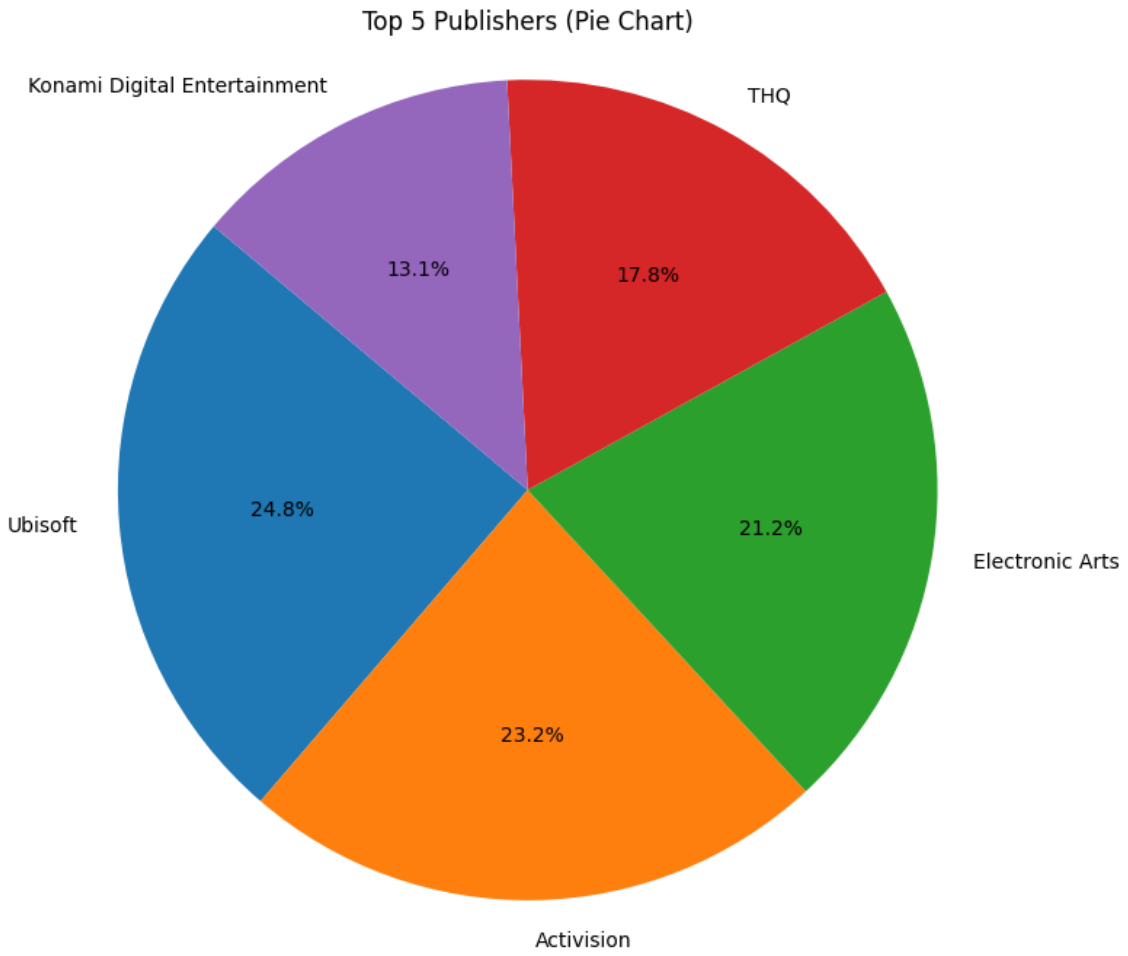
- Sales in **North America and Europe** have wider distributions and higher medians compared to other regions.
- **Japan and Other regions** exhibit lower and more compact distributions.
- The presence of **outliers in each region** is expected due to high-selling franchise games like Pokémon, FIFA, or GTA.
- This distribution helps businesses understand market reach and regional dominance.



D. Pie Chart – Top 5 Publishers

This pie chart displays the **top 5 publishers** in terms of game count:

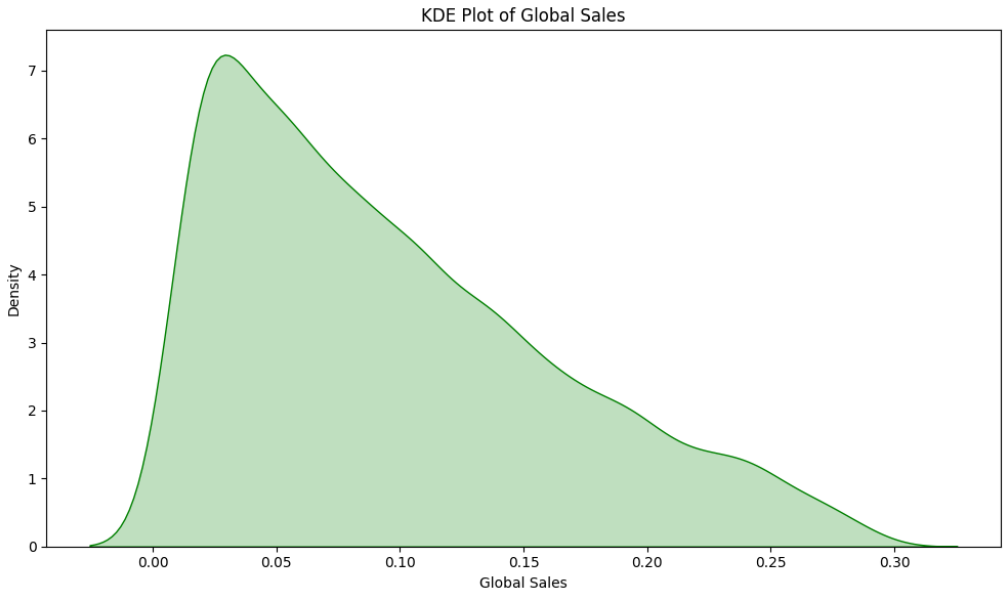
- Publishers like **Electronic Arts**, **Nintendo**, and **Activision** dominate the chart.
- They collectively occupy a large portion of the market, reflecting brand loyalty and marketing power.
- Identifying such leaders helps understand trends in publishing and market influence.



E. KDE Plot – Global Sales Distribution

The KDE (Kernel Density Estimation) plot shows:

- A **right-skewed distribution**, where most games have **low global sales**, and only a few perform exceptionally well.
- The long tail indicates **blockbuster hits** with extremely high global sales, a common trait in entertainment industries.
- Such patterns indicate that **only a small fraction of games generate massive revenue**, while

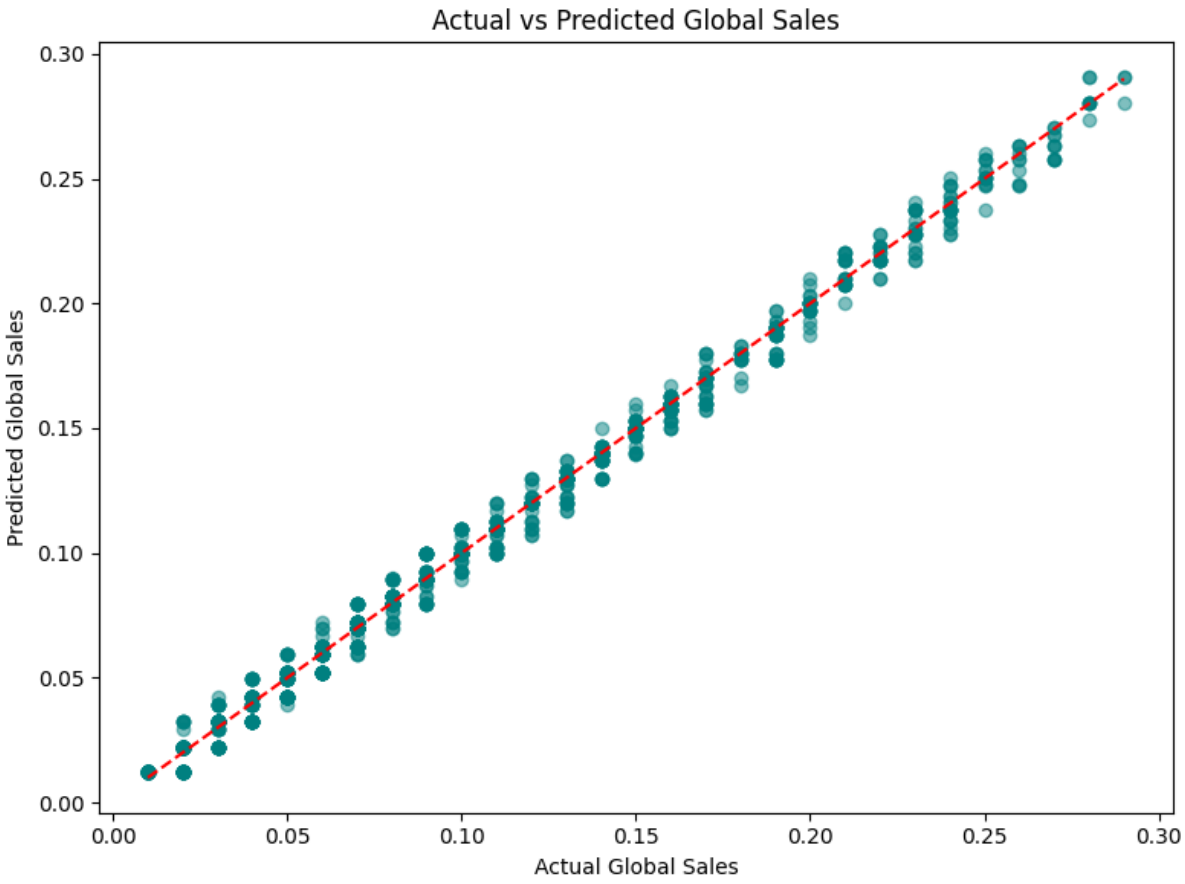


Results:

The linear regression model yielded impressive results in predicting global sales based on regional sales. Key performance metrics and model parameters are as follows:

- **Mean Squared Error (MSE):** 2.82e-05 – This low value indicates that the average squared difference between actual and predicted global sales is minimal, reflecting high prediction accuracy.
- **R^2 Score:** 0.9941 – The model explains over **99.4%** of the variance in global sales, which signifies an excellent fit and confirms that regional sales are highly predictive of total global sales.
- **Model Coefficients:**
  - NA_Sales: 1.0041
  - EU_Sales: 1.0049
  - Other_Sales: 0.7227
  - JP_Sales: 0.0
  - **Intercept:** 0.0021

These coefficients indicate the contribution of each regional sales figure to the global sales. Interestingly, the JP_Sales coefficient is 0.0, suggesting that it had no statistical weight in the model's predictions—possibly due to multicollinearity or zero variance in the subset used.



Actual vs Predicted Global Sales Plot

The scatter plot in the image visualizes the comparison between **actual** and **predicted** global sales:

- Each teal dot represents a data point, plotted with actual global sales on the x-axis and predicted values on the y-axis.
- The **red dashed line** is the ideal line where predicted = actual. The fact that the dots closely follow this line confirms the model's high accuracy.
- The clustering of points along the line suggests consistent predictions across the entire range, with no significant over- or under-prediction.

Overall, the model performs remarkably well and is reliable for forecasting global sales based on regional data.

Chapter 5:

CONCLUSIONS AND FUTURE SCOPE

Conclusion

This project set out to explore and analyze a real-world dataset on video game sales and to develop a predictive model capable of estimating global sales based on regional sales figures. Through careful data preprocessing, exploratory data analysis, and the implementation of a Linear Regression model, we have successfully met these objectives.

The analysis revealed several insightful patterns in the gaming industry:

- **Sales Trends Over Time:** From the line plot, we observed that global video game sales have experienced noticeable fluctuations over the years, reflecting shifts in technology, consumer behavior, and market dynamics.
- **Genre Distribution:** The bar chart demonstrated the dominance of certain genres like Action, Sports, and Shooter, suggesting strong user preferences and trends that can influence future game development and marketing strategies.
- **Sales Distribution:** The box plot highlighted the presence of outliers and the wide range of sales figures across regions, underlining the disparity in performance among video games.
- **Top Publishers:** The pie chart showcased the market leadership of certain publishers, giving a snapshot of competitive dynamics in the industry.
- **Sales Distribution Density:** The KDE plot visualized how most global sales are clustered in lower ranges, with fewer games achieving blockbuster success.

From a modeling standpoint, the **Linear Regression model** demonstrated outstanding performance. The **R^2 score of 0.9941** indicates that regional sales (NA, EU, JP, and Others) are exceptionally strong predictors of total global sales. The model's low **Mean Squared Error** reinforces this high accuracy. Additionally, the scatter plot of predicted vs. actual sales affirmed that the model closely follows the true values, validating its reliability.

An important observation during model interpretation was that JP_Sales had no significant impact (coefficient = 0). This could imply either a lack of variance in Japanese sales data within our sample or possible multicollinearity among features. Further diagnostics would be needed to confirm this.

In summary, this project has proven that data-driven techniques, when applied thoughtfully, can generate valuable insights and highly accurate predictive models, even in dynamic and entertainment-focused domains like the gaming industry.

Future Scope

While the outcomes of this project have been encouraging, there is ample room for enhancement, extension, and further exploration. Here are several promising directions for future work:

*1. **Advanced Machine Learning Models***

Linear regression provides a good baseline, but exploring more advanced algorithms such as **Random Forest**, **XGBoost**, or **Neural Networks** may yield even better predictive performance—especially for more complex, non-linear relationships.

*2. **Incorporating More Features***

Currently, the model is limited to regional sales as features. Including additional variables such as:

- Game rating (ESRB)
- Platform popularity
- Marketing budget
- Number of active players
- Social media mentions or sentiment

…could improve prediction accuracy and offer more holistic modeling of game performance.

*3. **Temporal Forecasting***

Rather than predicting global sales based on existing data, future work can involve **time series forecasting**, predicting sales trajectories year-over-year for upcoming titles based on genre, platform, and publisher.

*4. **Clustering and Recommendation Systems***

Unsupervised learning methods like **K-Means clustering** could be used to segment games into categories (e.g., top-sellers, underperformers, niche), which could support:

- Targeted marketing
- Personalized game recommendations
- Portfolio analysis for game publishers

*5. **Deployment & Dashboarding***

The model and visualizations could be wrapped into a real-time interactive dashboard using tools like **Streamlit**, **Tableau**, or **Power BI**. This would make the insights accessible to non-technical stakeholders like marketers, developers, and investors.

6. ***Addressing Biases and Missing Data***

Further work could explore techniques for handling:

- **Missing values**, especially in the ‘Year’ and ‘Publisher’ columns
- **Bias in data**, such as overrepresentation of Western publishers or mainstream genres, which may skew predictions

7. ***Scaling Up***

If access to larger or more recent datasets becomes available (e.g., through APIs, game review websites, or sales databases), the model could be trained on millions of rows for more generalized predictions across a broader market.