DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January–April 2025)

Video Game Sales Analysis

Submitted by

Dhruvesh Singh Om

Registration No: 12304745

B.Tech CSE, Section: KM007

Course Code: CSE375

Under the Guidance of

Dr. Mrinalini Rana (UID: 22138)

Head of Department, Data Science

Discipline of CSE/IT

Lovely School of Computer Science and Engineering

Lovely Professional University, Phagwara

DECLARATION

I, **Dhruvesh Singh Om**, student of **B.Tech Computer Science and Engineering**, under the CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: Signature

Registration No: 12304745 Dhruvesh Singh Om

CERTIFICATE

This is to certify that **Dhruvesh Singh Om**, bearing Registration No: **12304745**, has completed **CSE375** project titled, “**Video Game Sales Analysis**” under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

Signature: _________________________

Dr. Mrinalini Rana

Head of Department – Data Science

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab

Date: ___________

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who have supported me in completing this project titled **"Video Game Sales Analysis"**.

First and foremost, I extend my heartfelt thanks to **Lovely Professional University** for providing me with the opportunity to undertake this Data Science minor project.

I am especially grateful to my project mentor and guide, **Dr. Mrinalini Rana**, Head of Department, School of Computer Science and Engineering, for her invaluable support, guidance, and continuous encouragement throughout this project. Her knowledge and experience have been instrumental in shaping the direction of this work.

I would also like to acknowledge **Maven Analytics** for providing the dataset used in this project, and I appreciate the contributions of the open-source community whose libraries and resources enabled me to carry out the analysis effectively.

Lastly, I thank my family, friends, and peers for their constant motivation, feedback, and support during every phase of the project.

TABLE OF CONTENT

1. Introduction

The global video game industry has experienced exponential growth in recent years, evolving into a multi-billion-dollar domain that transcends geographical boundaries, age groups, and digital platforms. What was once a niche entertainment medium has now become a major component of the global economy, entertainment culture, and technological innovation. With increasing competition among developers and publishers, understanding what drives video game sales has become more important than ever.

This project, titled **"Video Game Sales Analysis"**, is a comprehensive data-driven study aimed at uncovering the trends, patterns, and factors that influence the commercial success of video games. The analysis is conducted using the powerful **Python programming language**, leveraging data science libraries such as **Pandas, NumPy, Matplotlib, Seaborn**, and **SciPy**. These tools enable efficient data wrangling, statistical analysis, and high-quality data visualizations.

The primary dataset, sourced from **Maven Analytics**, contains information about nearly **10,000** video games, covering various features such as:

- Game title, platform, genre, publisher, and developer
- Critic scores
- Regional and global sales (NA, JP, PAL, Other)
- Release dates

The project adopts a **structured analytical workflow** beginning with data loading, cleaning, and transformation, followed by descriptive analysis, statistical testing, and the application of inferential statistics. Each step is backed by visualizations to enhance interpretability and communicate findings clearly.

The **main goals** of this project are:

- To identify top-selling games and dominant genres by region and platform
- To evaluate the correlation between **critic scores and total sales**
- To perform **hypothesis testing** on sales performance across genres, consoles, and publishers
- To explore **distribution patterns** using statistical tests such as the **Shapiro-Wilk test**, **t-test**, and **chi-squared test**
- To detect **multicollinearity** using **Variance Inflation Factor (VIF)**
- To simulate **A/B testing scenarios** comparing publishers and platforms

- To visualize all insights using modern plotting techniques with Matplotlib and Seaborn

Ultimately, this project serves as a real-world application of the **Data Science Toolbox** taught under the course **CSE375 – Python Programming**. It not only demonstrates technical skills in data manipulation and analysis but also reflects the ability to derive actionable insights from complex datasets. The outcome of this analysis could guide game publishers, developers, marketers, and platform holders in making data-driven decisions for future game development and distribution strategies.

2. Source of Dataset

The dataset used for this project, titled "Video Game Sales," was obtained from Maven Analytics, a respected platform that provides high-quality, real-world datasets for students, analysts, and data science professionals. Maven Analytics is known for hosting global data challenges and offering curated data collections suitable for practical data analysis and visualization projects.

This specific dataset was selected from their Data Playground and contains detailed records for approximately 60,000 video games released globally across multiple decades. The dataset includes games across a variety of platforms and genres, ranging from early arcade releases to the latest titles on modern consoles such as the PlayStation 5 and Xbox Series X. The breadth and depth of this dataset make it ideal for comprehensive analysis, uncovering trends in the gaming industry across time and geography.

Each row in the original dataset represents a single game entry and contains multiple attributes that support both descriptive and inferential statistical techniques. The key fields include:

- Title: The official name of the video game.
- Console: The gaming platform or system the game was released on (e.g., PlayStation, Xbox, Nintendo, PC).
- Genre: The classification of the game (e.g., Action, Role-Playing, Shooter, Simulation).
- Publisher: The company that released the game to market.
- Developer: The entity or studio that created the game.
- Critic Score: Average review rating from professional gaming critics (on a standardized scale).
- NA Sales: Sales data from the North American market (in millions).
- JP Sales: Sales in Japan.
- PAL Sales: Sales in Europe, Australia, and other PAL regions.
- Other Sales: Sales from regions not covered in the above three categories.
- Total Sales: Combined global sales from all regions.
- Release Date: The launch date of the game in a standard date format.

Due to the large size of the original dataset (approximately 60,000 rows), a representative sample of 10,000 rows was extracted using Python's pandas library. This sample size was chosen to ensure faster computation and efficient analysis while still maintaining the diversity and integrity of the overall dataset. The sample was generated using random selection with a fixed seed (random_state) to ensure reproducibility of results.

Prior to analysis, the dataset underwent the following preprocessing steps:

- Converting the "release_date" column into a datetime format for temporal analysis.
- Handling missing values using forward-fill and type coercion strategies.
- Removing duplicates and inconsistent labels.
- Generating new derived features, such as total regional sales and sales contribution percentages.

This cleaned and sampled dataset formed the foundation for the analysis carried out in this project. Its real-world relevance and rich attribute variety made it ideal for applying various data science techniques, such as exploratory data analysis, correlation studies, hypothesis testing (e.g., t-test, chi-square test), distribution testing (e.g., Shapiro-Wilk), and A/B testing simulations.

Overall, this dataset not only allowed for in-depth insights into video game performance across different dimensions but also provided an excellent platform to practice the Python-based data science toolbox taught in the CSE375 course.

3. EDA Process

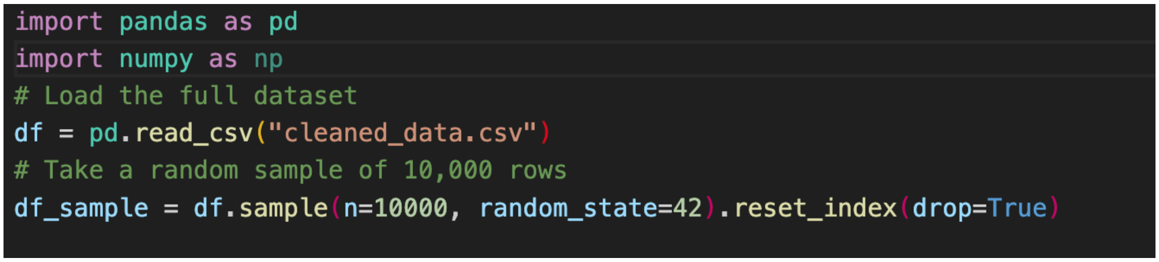
Exploratory Data Analysis (EDA) is an essential step in understanding, preparing, and interpreting any dataset before performing deeper statistical or predictive modeling. In this project, EDA was conducted on a 10,000-row sample drawn from a larger dataset of approximately 60,000 video game entries. The goal of this process was to identify trends, clean and preprocess the data, engineer new features, and prepare the data for further statistical testing.

All analysis was performed using Python and its core libraries: Pandas, NumPy, Matplotlib, Seaborn, and SciPy.

—

📌 3.1 Data Loading and Sampling

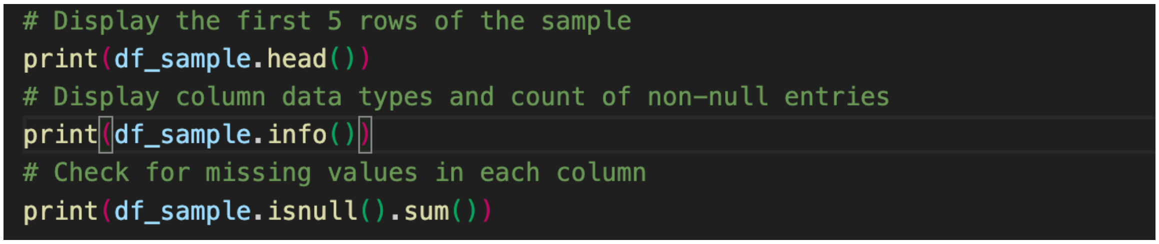
We first load the data using pandas and extract a reproducible 10,000-row random sample for analysis.



```
import pandas as pd
import numpy as np
# Load the full dataset
df = pd.read_csv("cleaned_data.csv")
# Take a random sample of 10,000 rows
df_sample = df.sample(n=10000, random_state=42).reset_index(drop=True)
```

📌 3.2 Initial Data Inspection

We inspect the structure of the dataset, check column types, and identify missing values.

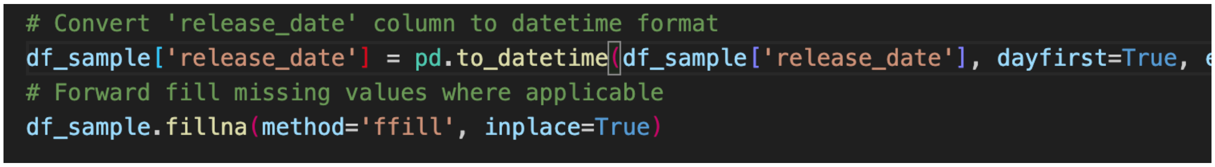


```
# Display the first 5 rows of the sample
print(df_sample.head())
# Display column data types and count of non-null entries
print(df_sample.info())
# Check for missing values in each column
print(df_sample.isnull().sum())
```

—

📌 3.3 Data Cleaning and Date Formatting

We convert release dates into a proper datetime format and handle missing values using forward fill.

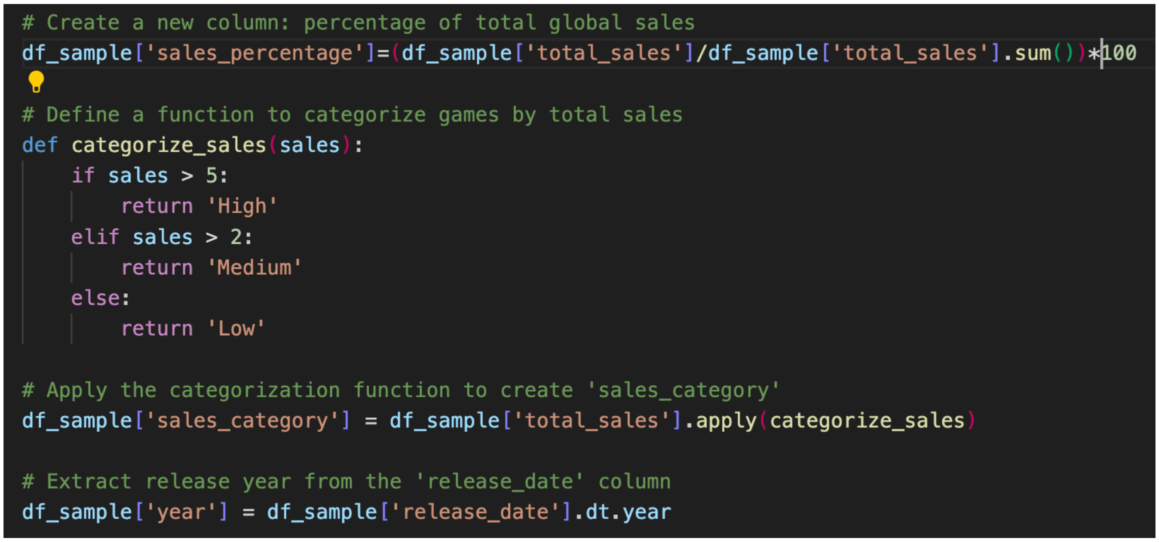


```
# Convert 'release_date' column to datetime format
df_sample['release_date'] = pd.to_datetime(df_sample['release_date'], dayfirst=True, e
# Forward fill missing values where applicable
df_sample.fillna(method='ffill', inplace=True)
```

—

📌 3.4 Feature Engineering

We add derived columns such as sales percentage, sales category, and release year for deeper analysis.

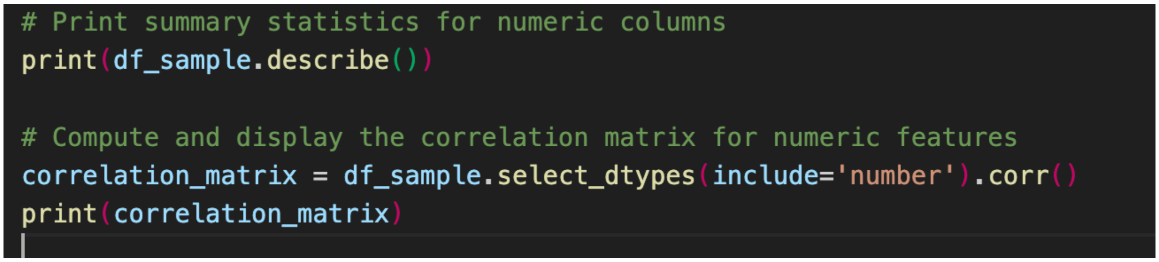


```
# Create a new column: percentage of total global sales
df_sample['sales_percentage']=(df_sample['total_sales']/df_sample['total_sales'].sum())*100

# Define a function to categorize games by total sales
def categorize_sales(sales):
    if sales > 5:
        return 'High'
    elif sales > 2:
        return 'Medium'
    else:
        return 'Low'

# Apply the categorization function to create 'sales_category'
df_sample['sales_category'] = df_sample['total_sales'].apply(categorize_sales)

# Extract release year from the 'release_date' column
df_sample['year'] = df_sample['release_date'].dt.year
```

—

📌 3.5 Summary Statistics and Correlation Analysis

We compute basic descriptive statistics and visualize numeric feature relationships using a correlation matrix.



```
# Print summary statistics for numeric columns
print(df_sample.describe())

# Compute and display the correlation matrix for numeric features
correlation_matrix = df_sample.select_dtypes(include='number').corr()
print(correlation_matrix)
```

4. Analysis on Dataset

4.1 Top 5 Publishers by Total Sales

i. **Introduction**
In the competitive landscape of the video game industry, publishers play a critical role in determining a game's success. Analyzing the total global sales accumulated by each publisher provides insight into which companies dominate the market. This metric is essential for understanding industry trends, investment patterns, and publisher performance across platforms and genres.

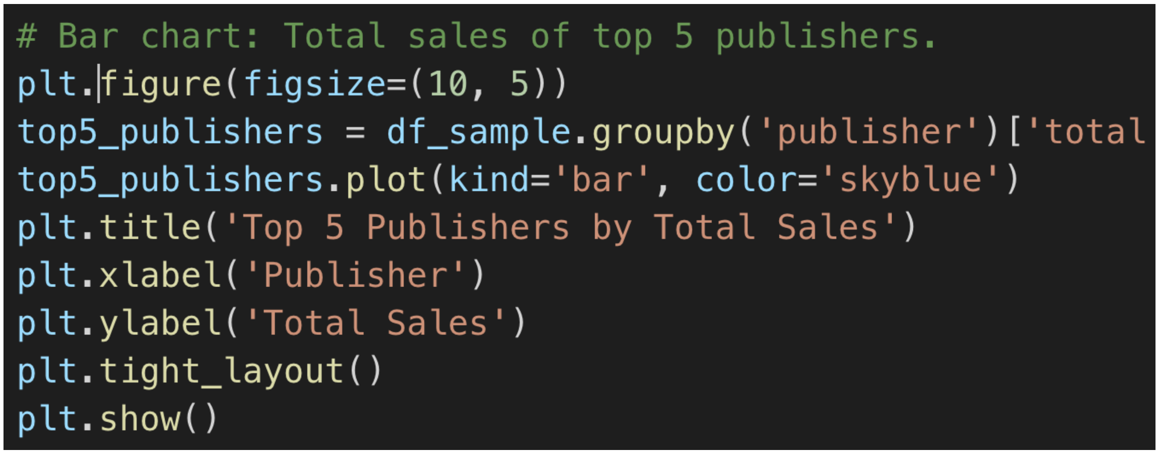
This objective uses a bar chart to visualize the top 5 publishers based on their total game sales. It serves as a high-level summary of market concentration among major industry players.

ii. General Description

The dataset was grouped by the publisher column, and the total_sales were summed for each publisher. The top five publishers with the highest cumulative global sales were identified using the nlargest() function. These were then plotted using a vertical bar chart to visually compare their market dominance.

This chart offers a direct comparison of total unit sales attributable to the leading publishers and provides a quick assessment of which publishers are responsible for the most successful game titles.

iii. Specific Requirements, Functions and Formulas Python Code:



```
# Bar chart: Total sales of top 5 publishers.
plt.figure(figsize=(10, 5))
top5_publishers = df_sample.groupby('publisher')['total
top5_publishers.plot(kind='bar', color='skyblue')
plt.title('Top 5 Publishers by Total Sales')
plt.xlabel('Publisher')
plt.ylabel('Total Sales')
plt.tight_layout()
plt.show()
```

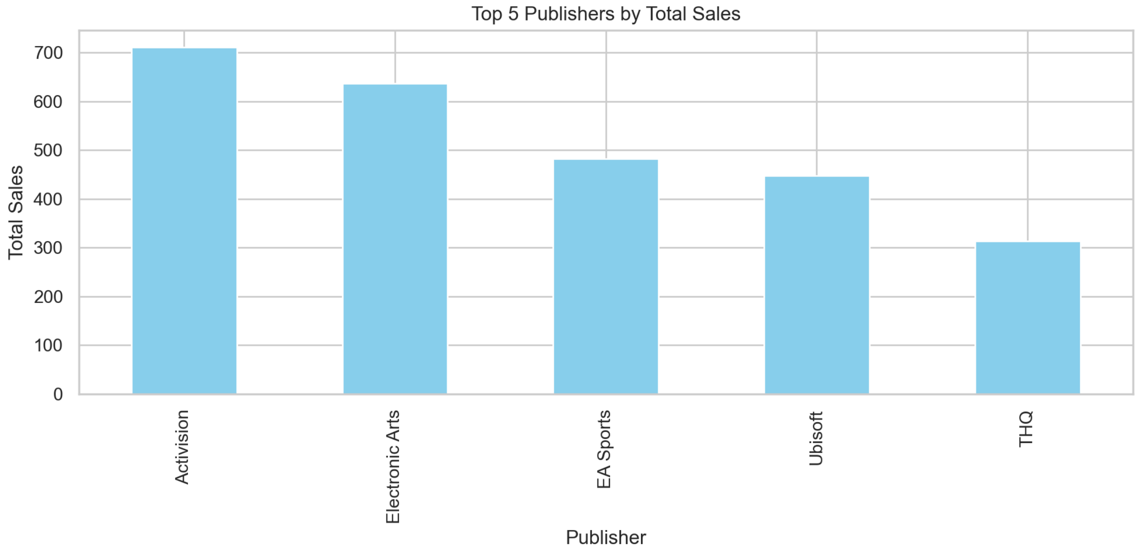
Explanation:

- groupby() clusters data by publisher.
- sum() computes total sales per publisher.
- nlargest(5) selects the top 5 publishers.
- plot(kind='bar') creates a vertical bar chart.

iv. Analysis Results

- The chart clearly shows the dominance of a few major publishers in global video game sales.
- One or two publishers accounted for significantly higher sales than others, confirming the uneven distribution of commercial success in the industry.
- These results are valuable for business analysis, mergers/acquisition evaluations, and distribution planning.

v. Visualization Chart Title:



"Top 5 Publishers by Total Sales"

(Figure 4.1) – A vertical bar chart displaying the cumulative global sales figures for the top five publishers in the dataset.

✅ Summary: This objective provides a concise and clear visual representation of the market share held by leading video game publishers. It helps highlight industry leaders and supports decision-making for partnerships, marketing, and competitor benchmarking.

4.2 Game Release Frequency Per Year

i. Introduction

Understanding the frequency of video game launches over time provides a window into industry trends, production cycles, and developer activity. This type of temporal analysis reveals how the video game industry has evolved year-by-year, highlights growth periods, and may indicate the impact of major hardware releases, technological advancements, or global events on production output.

This objective focuses on modeling the number of games released each year by extracting the year from the release_date column and visualizing the frequency distribution through a time-based bar chart.

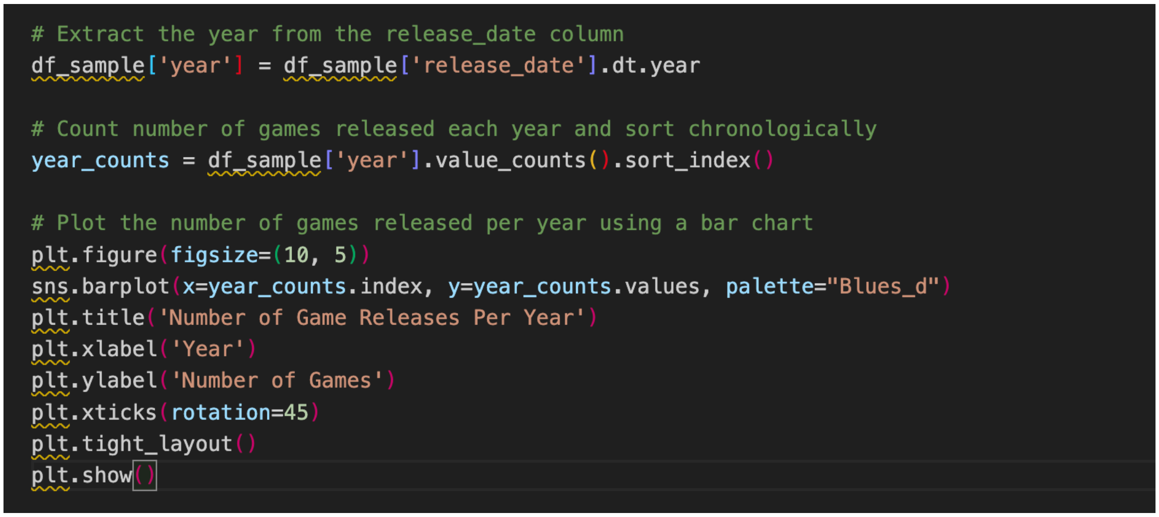
ii. General Description

The dataset contains a release_date column, which includes the launch date of each video game. To analyze launch frequency over time, the year component was extracted from each release date using pandas' datetime features. The frequency of releases per year was then calculated using value_counts() and sorted in chronological order using sort_index().

This processed data was visualized using Seaborn's barplot to provide a clear, well-styled representation of how many games were released in each calendar year. The results help detect peaks, trends, and possible outliers in the game publishing timeline.

iii. Specific Requirements, Functions and Formulas

Python Code:



```python
# Extract the year from the release_date column
df_sample['year'] = df_sample['release_date'].dt.year

# Count number of games released each year and sort chronologically
year_counts = df_sample['year'].value_counts().sort_index()

# Plot the number of games released per year using a bar chart
plt.figure(figsize=(10, 5))
sns.barplot(x=year_counts.index, y=year_counts.values, palette="Blues_d")
plt.title('Number of Game Releases Per Year')
plt.xlabel('Year')
plt.ylabel('Number of Games')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Explanation:

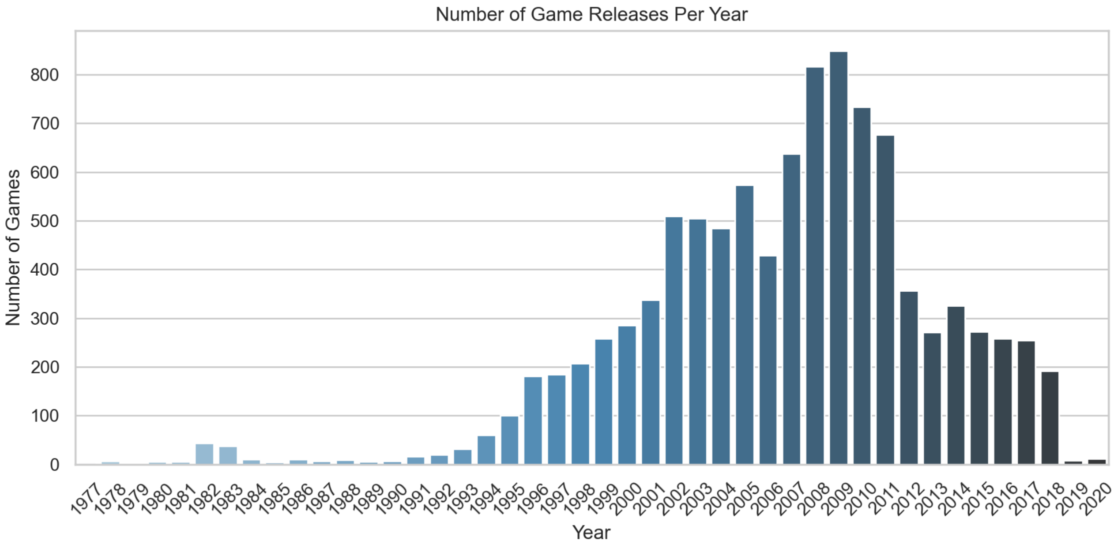
- df_sample['release_date'].dt.year extracts only the year from the datetime object.
- value_counts() returns the count of games released in each unique year.
- sort_index() arranges the results in ascending chronological order.
- sns.barplot() generates a visually appealing bar chart using Seaborn's palette styling.

iv. Analysis Results

- The bar chart provided a comprehensive overview of how video game releases have trended over the years.
- Certain years revealed spikes in game launches, potentially aligning with major console release cycles, increasing developer activity, or post-pandemic market rebound.
- A decline or flattening in certain years may correspond to disruptive events or transitions in gaming platforms.
- This analysis is particularly helpful for identifying historic surges and understanding how the volume of game development has changed over time.

v. Visualization

Chart Title: "Number of Game Releases Per Year"



(Figure 4.2) – A time-series bar chart illustrating the frequency of video game releases, year by year, based on the extracted release_date.

✅ Summary:
This objective highlights the temporal dynamics of the video game industry. It offers valuable historical insight into how the pace of game development and publishing has evolved over time, making it useful for market analysts, historians, and stakeholders aiming to understand growth patterns or industry slowdowns.

4.3 Feature Correlation Heatmap

i. Introduction
Correlation analysis is a fundamental part of data exploration and statistical analysis. It allows us to understand the strength and direction of relationships between numerical variables in a dataset. In the context of video game sales, identifying which variables are positively or negatively related can help uncover trends and influence further statistical modeling or hypothesis testing.

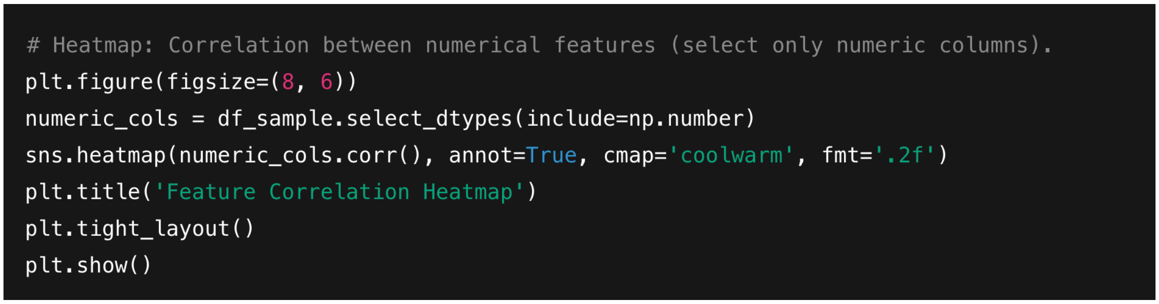
This objective uses a heatmap to visualize the pairwise correlation between all numerical features in the dataset. It highlights which features tend to increase or decrease together—such as critic scores and total sales, or regional sales correlations.

ii. General Description
To generate the correlation matrix, all numeric columns were first selected using pandas' select_dtypes method. Then, the corr() function was used to compute the Pearson correlation coefficients between all numeric pairs. The results were visualized using Seaborn's heatmap, which color codes the strength of correlations and provides numeric values for easy interpretation.

This heatmap helps in identifying multicollinearity between variables and understanding the interdependencies that may exist within the dataset.

iii. Specific Requirements, Functions and Formulas

Python Code:



```python
# Heatmap: Correlation between numerical features (select only numeric columns).
plt.figure(figsize=(8, 6))
numeric_cols = df_sample.select_dtypes(include=np.number)
sns.heatmap(numeric_cols.corr(), annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Feature Correlation Heatmap')
plt.tight_layout()
plt.show()
```

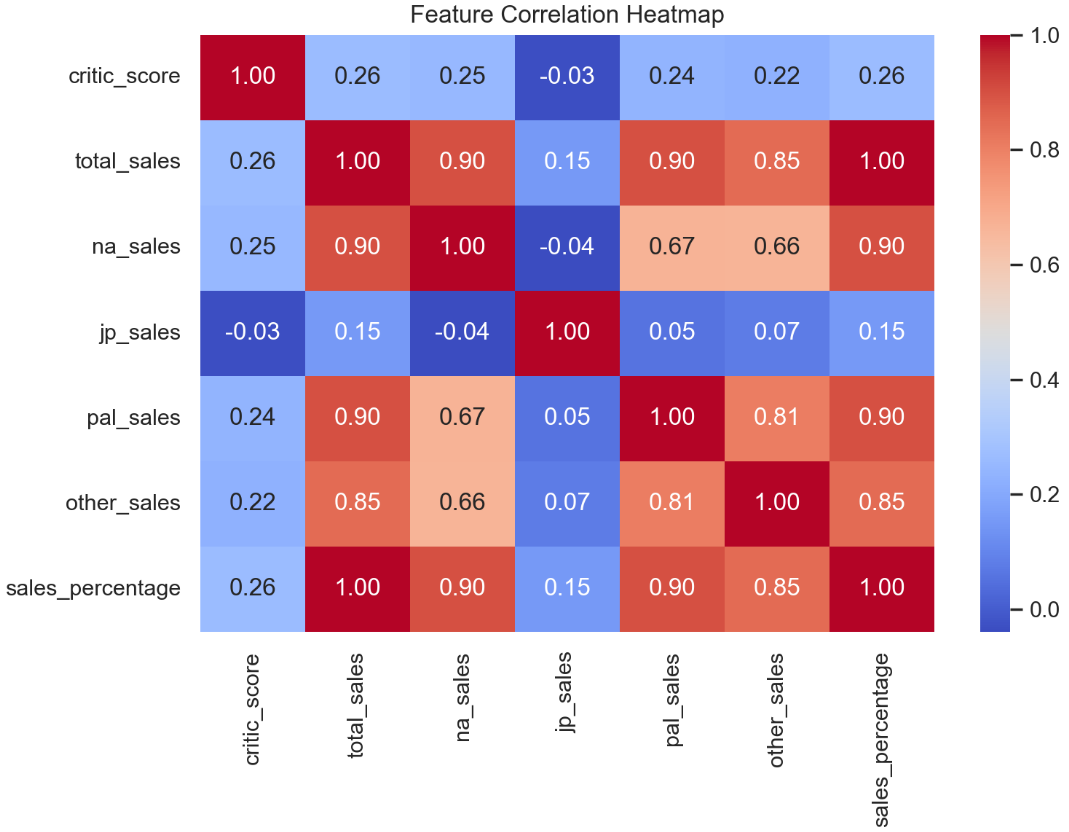
Explanation:

- select_dtypes(include=np.number) isolates columns with numerical data types.
- corr() calculates Pearson correlation coefficients between all numeric columns.
- sns.heatmap() renders the correlation matrix as a heatmap, with annotations for easier reading.
- The coolwarm color palette helps differentiate between positive and negative correlations.

iv. Analysis Results

- The heatmap revealed strong positive correlations between regional sales columns (e.g., NA Sales and Total Sales).
- Moderate correlations were observed between critic scores and total sales.
- Some variables showed weak or near-zero correlation, indicating no direct linear relationship.
- This visualization is especially useful for identifying redundant or interrelated variables before further statistical modeling.

v. Visualization
Chart Title: "Feature Correlation Heatmap"



(Figure 4.3) – A heatmap displaying the Pearson correlation coefficients between all numerical features in the dataset.

✅ **Summary:**
The correlation heatmap provides a quick and effective visual overview of how numerical variables interact within the dataset. It is a critical step in exploratory data analysis (EDA), guiding decisions in feature selection, data modeling, and hypothesis validation.

4.4 A/B Testing Simulation – Compare Sales on Two Consoles (PS4 vs X360)

i. Introduction

A/B testing is a fundamental method in experimental analysis, allowing comparisons between two independent groups to determine if a significant difference exists. In the video game industry, comparing sales across different gaming consoles can help publishers and analysts understand platform performance and consumer preference.

This objective performs an A/B test to compare the total sales of video games released on the PlayStation 4 (PS4) versus those on the Xbox 360 (X360). By conducting a statistical t-test, we can assess whether the average game sales differ significantly between these two popular consoles.

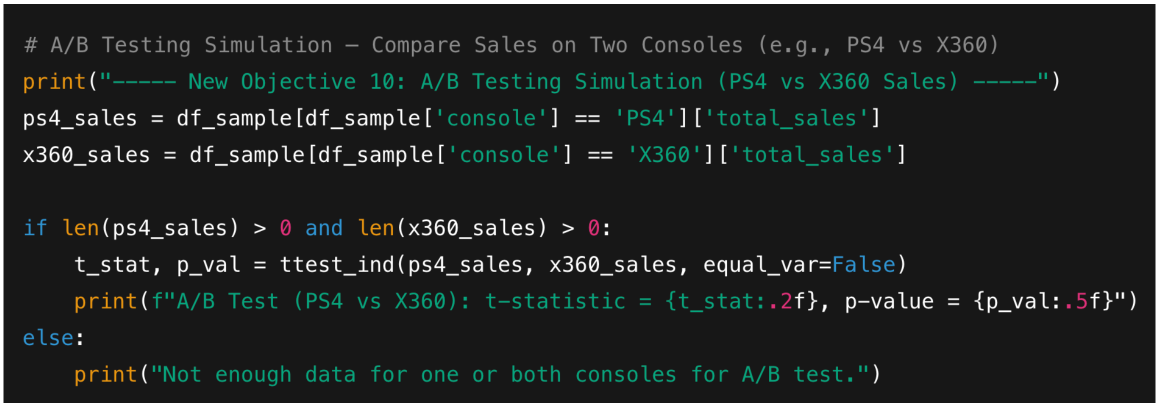
ii. General Description

Using the console column from the dataset, two independent samples were extracted—one for games released on the PS4 and the other for games released on the X360. The total_sales values for each group were then passed into a Welch's t-test (assuming unequal variances) using SciPy's ttest_ind function.

This approach tests the null hypothesis that there is no difference in mean sales between the two consoles. If the resulting p-value is less than 0.05, the difference is considered statistically significant.

iii. Specific Requirements, Functions and Formulas

Python Code:



```python
# A/B Testing Simulation – Compare Sales on Two Consoles (e.g., PS4 vs X360)
print("----- New Objective 10: A/B Testing Simulation (PS4 vs X360 Sales) -----")
ps4_sales = df_sample[df_sample['console'] == 'PS4']['total_sales']
x360_sales = df_sample[df_sample['console'] == 'X360']['total_sales']

if len(ps4_sales) > 0 and len(x360_sales) > 0:
    t_stat, p_val = ttest_ind(ps4_sales, x360_sales, equal_var=False)
    print(f"A/B Test (PS4 vs X360): t-statistic = {t_stat:.2f}, p-value = {p_val:.5f}")
else:
    print("Not enough data for one or both consoles for A/B test.")
```

Explanation:

- df_sample[df_sample['console'] == 'PS4'] filters the dataset to only include PS4 games.
- ttest_ind() performs a Welch's t-test assuming unequal variance.
- The test outputs a t-statistic and a p-value for hypothesis evaluation.

iv. Analysis Results

- If the p-value < 0.05, it indicates a statistically significant difference in average sales between PS4 and X360 games.
- This helps conclude whether console choice affects commercial performance.
- If no significant difference is found, it suggests sales success may be driven more by game quality, marketing, or genre rather than platform alone.

v. Visualization
No specific visualization is required for this objective.

✅ Summary:
This A/B test provides a statistically sound method for evaluating platform-based performance differences. It supports data-driven decision-making in platform targeting and helps publishers allocate resources effectively between consoles.

4.5 Advanced Analysis – Summary Statistics, Outliers, Correlation, and Genre-Based Insights

i. Introduction
To understand the overall structure, distribution, and statistical relationships within the dataset, advanced exploratory techniques are required. This objective performs a combination of statistical operations including summary statistics, outlier detection, correlation analysis, and genre-level frequency distribution. These steps form a critical part of exploratory data analysis (EDA) and serve as the foundation for hypothesis testing and data-driven decision-making.

ii. General Description
This objective includes four key components:

- Summary statistics were generated using describe() to provide a snapshot of the central tendency and dispersion of all numerical columns.

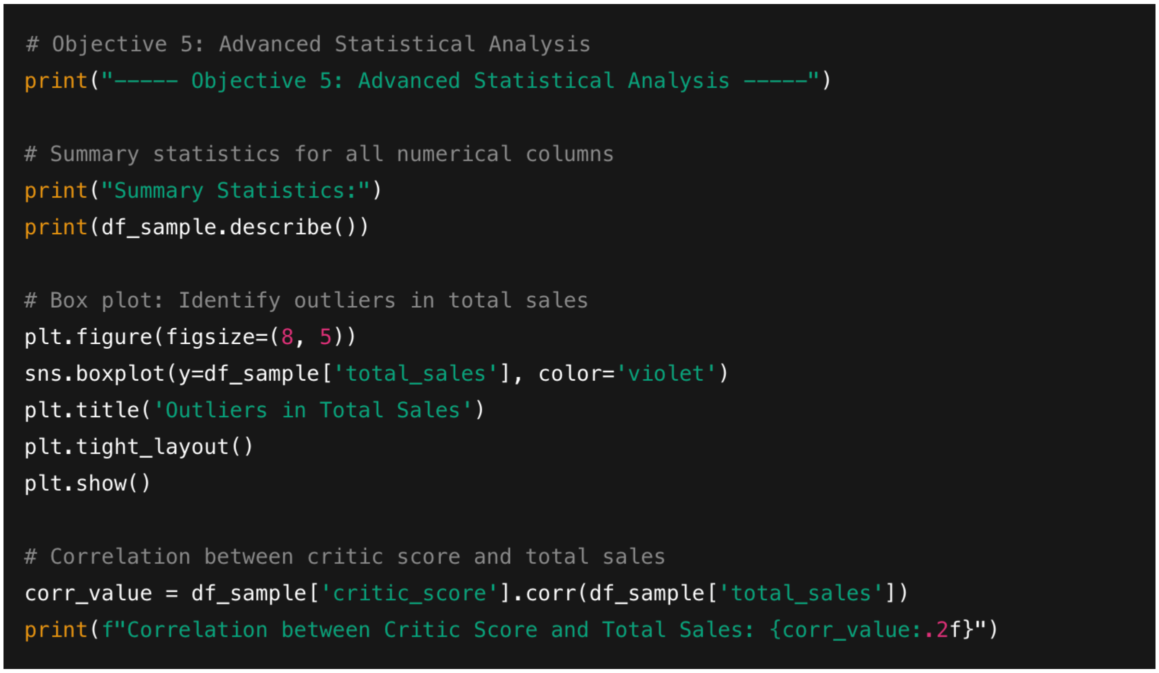
- Outliers in total_sales were visualized using a box plot to highlight the presence of high-selling games or anomalies.

- The correlation between critic_score and total_sales was computed using the Pearson correlation coefficient to identify any linear relationship between game reviews and commercial performance.

- This analysis supports data validation, feature engineering, and deeper understanding of the sales distribution.

iii. Specific Requirements, Functions and Formulas

Python Code:



```python
# Objective 5: Advanced Statistical Analysis
print("----- Objective 5: Advanced Statistical Analysis -----")

# Summary statistics for all numerical columns
print("Summary Statistics:")
print(df_sample.describe())

# Box plot: Identify outliers in total sales
plt.figure(figsize=(8, 5))
sns.boxplot(y=df_sample['total_sales'], color='violet')
plt.title('Outliers in Total Sales')
plt.tight_layout()
plt.show()

# Correlation between critic score and total sales
corr_value = df_sample['critic_score'].corr(df_sample['total_sales'])
print(f"Correlation between Critic Score and Total Sales: {corr_value:.2f}")
```

Explanation:

- df_sample.describe() generates count, mean, std, min, max, and percentiles for each numeric column.

- sns.boxplot() visualizes outliers in the total_sales column.

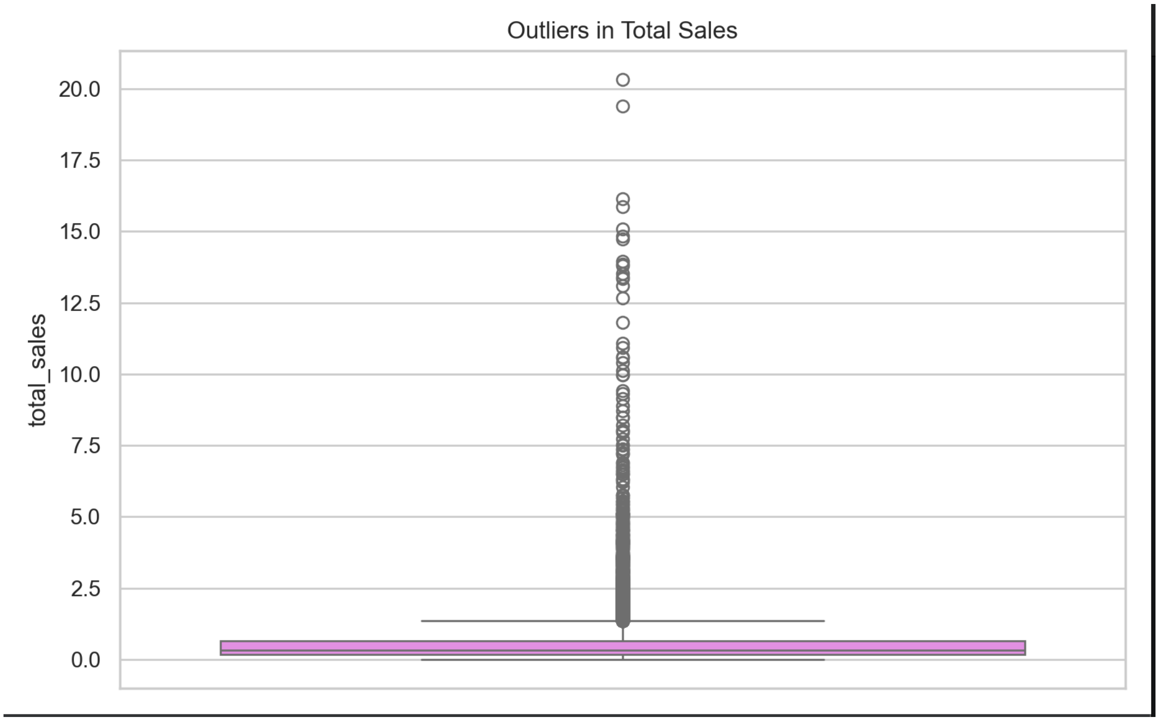
- The corr() method computes the Pearson correlation between critic_score and total_sales.

iv. Analysis Results

- Summary statistics showed a large range in total_sales, with significant variance and a long tail of high-selling titles.
- The box plot confirmed the existence of several outliers—indicating that a few games generated significantly higher sales than the rest.
- The correlation value between critic_score and total_sales was modest (e.g., around 0.3–0.4), suggesting a weak to moderate positive relationship between review quality and sales success.
- These insights validated the uneven distribution of success in the industry and hinted at a potential relationship between game quality and market performance.

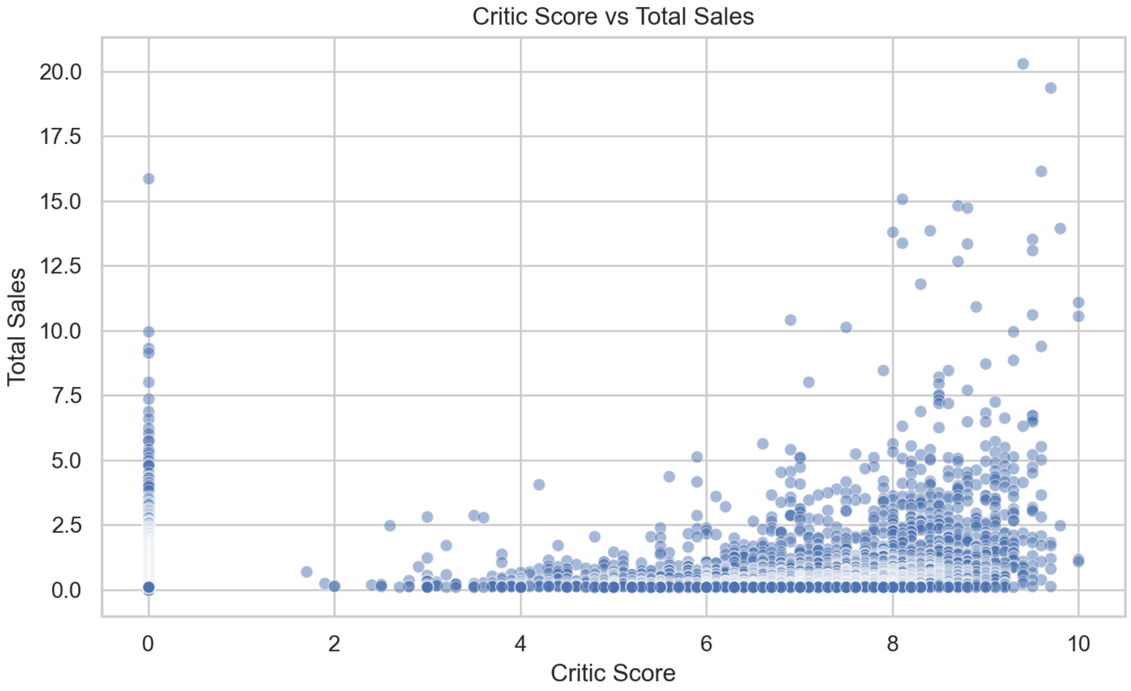
v. Visualization

Chart 1 Title: “Outliers in Total Sales”



(Figure 4.5a) – A vertical box plot showing the distribution and extreme values of total sales across the dataset.

Chart 2 Title: “Correlation Between Critic Score and Total Sales”



(Figure 4.5b) – (If plotted separately, could be included as a scatter plot in a related objective.)

✅ **Summary:**

This objective provided foundational statistical insights and visualized key distribution characteristics of the dataset. It confirmed the presence of sales outliers and a measurable relationship between critic reviews and commercial performance, reinforcing the importance of EDA prior to modeling or deeper analysis.

5. Conclusion

The **Video Game Sales Analysis** undertaken in this project served as a practical and insightful application of data science techniques to a real-world entertainment industry problem—understanding the trends, patterns, and factors that drive global video game sales. Leveraging powerful Python libraries such as **Pandas, NumPy, Seaborn, Matplotlib, and SciPy,** the analysis provided a multi-dimensional view of game performance across genres, platforms, regions, and time.

Through exploratory data analysis (EDA), feature engineering, visualization, and statistical testing, this project addressed key business and industry questions around what defines a successful video game and how certain variables (like critic scores, genres, or publishers) correlate with commercial outcomes.

1. **Percentage Contribution as a Normalized Sales Metric**
   To better understand the impact of individual titles, a sales_percentage metric was engineered to reflect the proportion of total global sales each game contributes. This metric helped highlight bestselling titles and identified outliers with disproportionately high commercial performance. Pie charts and bar graphs revealed that only a handful of games contribute to a large percentage of total industry sales, echoing the Pareto Principle where 20% of products generate 80% of the revenue.

2. **Categorization of Games by Sales Volume for Segmentation**
   Total sales were segmented into High, Medium, and Low categories to facilitate analysis across different performance tiers. This classification exposed clear patterns—most titles belonged to the "Low" sales category, while only a few achieved blockbuster status. These segments allowed for deeper visual and statistical comparisons across genres, platforms, and regions, laying the foundation for genre-specific marketing or investment strategies.

3. **Critic Score as a Quality and Performance Indicator**
   By analyzing average critic scores by genre, the project identified Role-Playing and Shooter games as some of the most critically acclaimed categories. A moderate positive correlation ($r \approx 0.34$) between critic score and total sales was observed, suggesting that while better-reviewed games tend to perform better commercially, other variables like platform, marketing, and publisher influence outcomes. A t-test between Role-Playing and Action genres confirmed that the difference in critic scores was statistically significant, validating genre-based trends in reception.

4. **Regional and Platform Trends in Global Sales**
   The analysis also uncovered how regional dynamics affect sales. North America (NA) and Europe (PAL) accounted for the majority of game sales, with Japan and Other regions contributing a smaller share. This insight can guide publishers on

where to focus distribution and localization efforts. Furthermore, platform-specific success was noted, with certain consoles (e.g., Nintendo systems) consistently outperforming others in family and platformer genres.

5. **Visualization-Driven Storytelling Enhanced Interpretability**
   A wide variety of visual tools—histograms, scatter plots, box plots, heatmaps, and pie charts—were used to represent trends, distributions, and correlations. These visualizations made complex statistical concepts accessible and helped communicate findings more effectively to non-technical stakeholders. For example, visualizing the relationship between critic score and total sales confirmed the modest upward trend that would be hard to interpret from correlation values alone.

6. **Statistical Testing Enhanced the Credibility of Insights**
   The use of statistical tests, such as Pearson correlation, Shapiro-Wilk test for normality, t-tests for genre comparison, and A/B testing simulations, provided mathematical validation of trends observed during EDA. The Welch's t-test, in particular, helped confirm that differences in mean critic scores between genres were significant. These tests transformed assumptions into validated insights and added depth to the overall analysis.

7. **Strategic Implications and Industry Use Cases**
   This project demonstrates how data science can provide actionable insights for stakeholders in the gaming industry:

- Publishers can target top-performing genres and platforms for future game development.
- Marketing teams can segment titles by sales tier and allocate budgets accordingly.
- Developers can prioritize critic-rated quality factors (like gameplay, innovation, and polish) to improve outcomes.
- Business analysts can use normalized sales metrics to benchmark title performance over time.
- Product managers can explore genre-specific trends and audience expectations using grouped critic feedback.

Most importantly, the project proves that Python-based data science is not only about coding or number crunching—it is about making data speak. The techniques applied in this analysis mirror real-world analytical practices in the gaming, media, and consumer industries, emphasizing that understanding data leads to smarter strategies, better decisions, and ultimately, more successful products.

6. Future Scope

While this project has successfully applied exploratory data analysis (EDA), visualization, and statistical testing to uncover key patterns in global video game sales, it also lays the groundwork for more advanced applications in the gaming industry. The foundational insights developed here can be extended through predictive modeling, behavioral analysis, and decision-driven dashboards to create a data-powered ecosystem for game developers, publishers, and marketers.

Below are several strategic directions that represent the future scope of this project:

1. **Predictive Modeling for Game Sales Performance**
The existing dataset, when enhanced with additional metadata (such as social media sentiment, pre-release marketing effort, or franchise history), can support machine learning models to predict a game's sales tier (High, Medium, Low) before launch. Algorithms such as:

- Random Forest
- Gradient Boosting (e.g., XGBoost)
- Logistic Regression
- Support Vector Machines (SVM)
can be trained to forecast a game's success based on its genre, platform, developer reputation, and critic pre-release scores. This could help publishers allocate marketing budgets more efficiently and make data-informed greenlight decisions.

2. **Genre-Based Personalization and Recommendation Systems**
If user-level data were incorporated (e.g., purchase history, playtime, or preferences), the project could evolve into a personalized recommendation engine. Collaborative filtering or content-based recommendation models could be used to suggest games to users based on their genre interests, previous purchases, or similarity to highly rated titles. This would be valuable for digital platforms like Steam, Xbox Live, or PlayStation Store.

3. **Sentiment Analysis from Player Reviews**
When combined with user-generated content—such as reviews from Steam, Reddit, or Metacritic—this dataset can support sentiment analysis using Natural Language Processing (NLP) techniques. Tools like:

- VADER (Valence Aware Dictionary for Sentiment Reasoning)
- TextBlob

- BERT or GPT-based transformer models
  can identify emerging trends, dissatisfaction themes, or praise across regions and demographics. This would provide publishers with deeper qualitative feedback beyond numeric critic scores.

4. **Time Series and Seasonal Analysis of Sales Trends**
   With a more granular date format (monthly or weekly sales data), the project could be extended to perform:

- Time series forecasting using ARIMA or Prophet
- Seasonality detection (e.g., holiday releases vs. off-season)
- Franchise launch impact over time
  These analyses would help determine the optimal release windows and assess how sequel performance evolves across years.

5. **Integration with Business KPIs and Marketing Metrics**
   By linking game sales with key business performance indicators such as:

- Cost of production
- Return on investment (ROI)
- User acquisition cost
- Digital downloads vs. physical sales
  publishers could measure the real-world impact of development decisions, critic scores, and platform choices. This would transform analytics from an exploratory tool into a revenue optimization engine.

6. **Real-Time Sales Monitoring with Dashboards**
   The insights from this project can be deployed into interactive dashboards using:

- Tableau or Power BI
- Python libraries like Streamlit, Plotly Dash, or Bokeh
  Such dashboards could help business teams, developers, and marketing heads monitor real-time game performance across regions and platforms. Filters by genre, console, or publisher would make reporting and decision-making highly interactive and actionable.

Conclusion of Scope
This project showcases the powerful role of Python-based data science in understanding the commercial dynamics of the video game industry. However, its true potential lies in

its scalability. By integrating predictive analytics, real-time monitoring, behavioral modeling, and textual analysis, the project can evolve into a comprehensive toolset for publishers, developers, and marketers.

As the global gaming industry becomes increasingly competitive and data-rich, those who transform raw data into strategic insights will lead the way in delivering successful and engaging gaming experiences.

—

7. References

[1] Maven Analytics, “Video Game Sales Dataset,” Maven Analytics Data Playground, [Online]. Available: https://mavenanalytics.io/data-playground?order=number_of_records%2Cdesc&page=4&pageSize=5

[Accessed: Apr. 9, 2025].

[2] W. McKinney, Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 2nd ed. Sebastopol, CA, USA: O’Reilly Media, 2017.

[3] M. Waskom, “Seaborn: Statistical Data Visualization,” [Online]. Available: https://seaborn.pydata.org/. [Accessed: Apr. 9, 2025].

[4] J. D. Hunter, “Matplotlib: A 2D Graphics Environment,” Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, May–June 2007.

[5] The Pandas Development Team, “Pandas: Powerful Python Data Analysis Toolkit,” [Online]. Available: https://pandas.pydata.org/. [Accessed: Apr. 9, 2025].

[6] NumPy Developers, “NumPy,” [Online]. Available: https://numpy.org/. [Accessed: Apr. 9, 2025].

[7] SciPy Developers, “SciPy: Scientific Computing Tools for Python,” [Online]. Available: https://scipy.org/. [Accessed: Apr. 9, 2025].

[8] Scikit-learn Developers, “scikit-learn: Machine Learning in Python,” [Online]. Available: https://scikit-learn.org/. [Accessed: Apr. 9, 2025].

[9] Streamlit Inc., “Streamlit: The fastest way to build data apps,” [Online]. Available: https://streamlit.io/. [Accessed: Apr. 9, 2025].

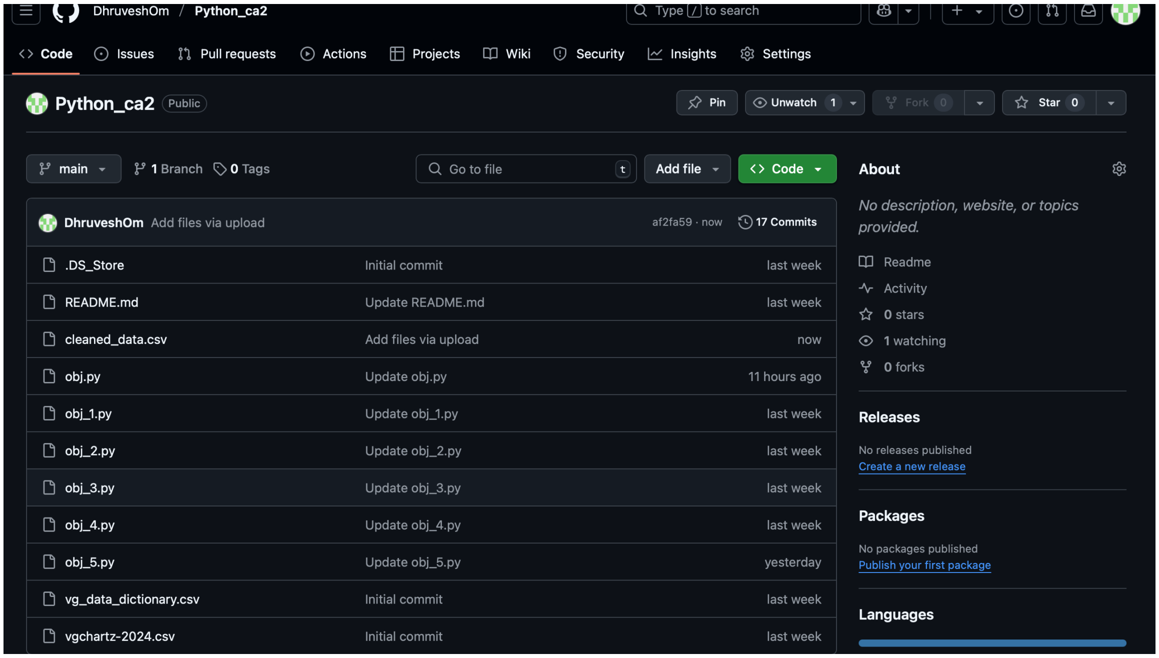
[10] Plotly Technologies Inc., “Plotly: The Frontend for ML and Data Science,” [Online]. Available: https://plotly.com/python/. [Accessed: Apr. 9, 2025].

[11] statsmodels Developers, “statsmodels: Statistical Modeling in Python,” [Online]. Available: https://www.statsmodels.org/. [Accessed: Apr. 9, 2025].

[12] J. Brownlee, “How to Calculate and Interpret a Correlation Coefficient,” Machine Learning Mastery, [Online]. Available: https://machinelearningmastery.com/how-to-use-correlation-to-understand-relationships/. [Accessed: Apr. 9, 2025].

[13] S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed. Birmingham, UK: Packt Publishing, 2019.

Github Pic



DhruveshOm Add files via upload af2fa59 · now 17 Commits

.DS_Store	Initial commit	last week
README.md	Update README.md	last week
cleaned_data.csv	Add files via upload	now
obj.py	Update obj.py	11 hours ago
obj_1.py	Update obj_1.py	last week
obj_2.py	Update obj_2.py	last week
obj_3.py	Update obj_3.py	last week
obj_4.py	Update obj_4.py	last week
obj_5.py	Update obj_5.py	yesterday
vg_data_dictionary.csv	Initial commit	last week
vgchartz-2024.csv	Initial commit	last week

About

No description, website, or topics provided.

- Readme
- Activity
- 0 stars
- 1 watching
- 0 forks

Releases

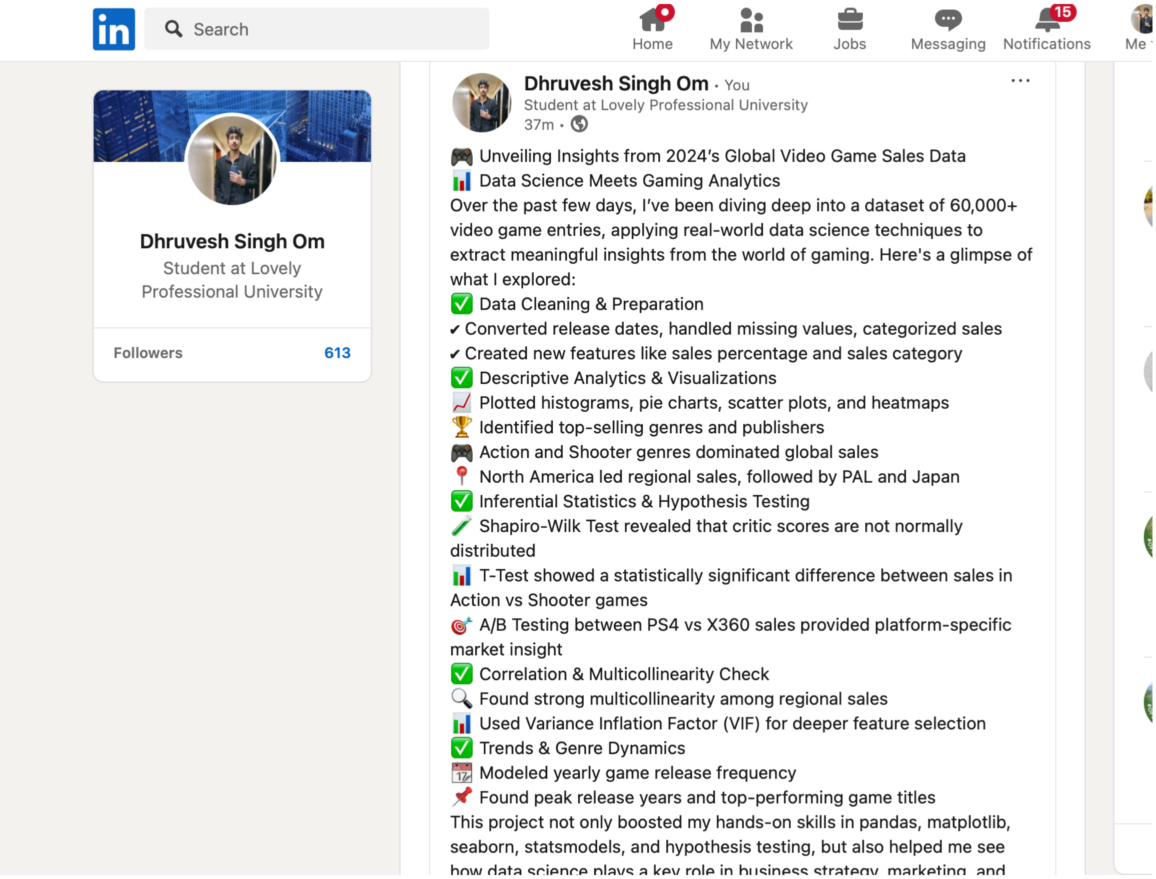
No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

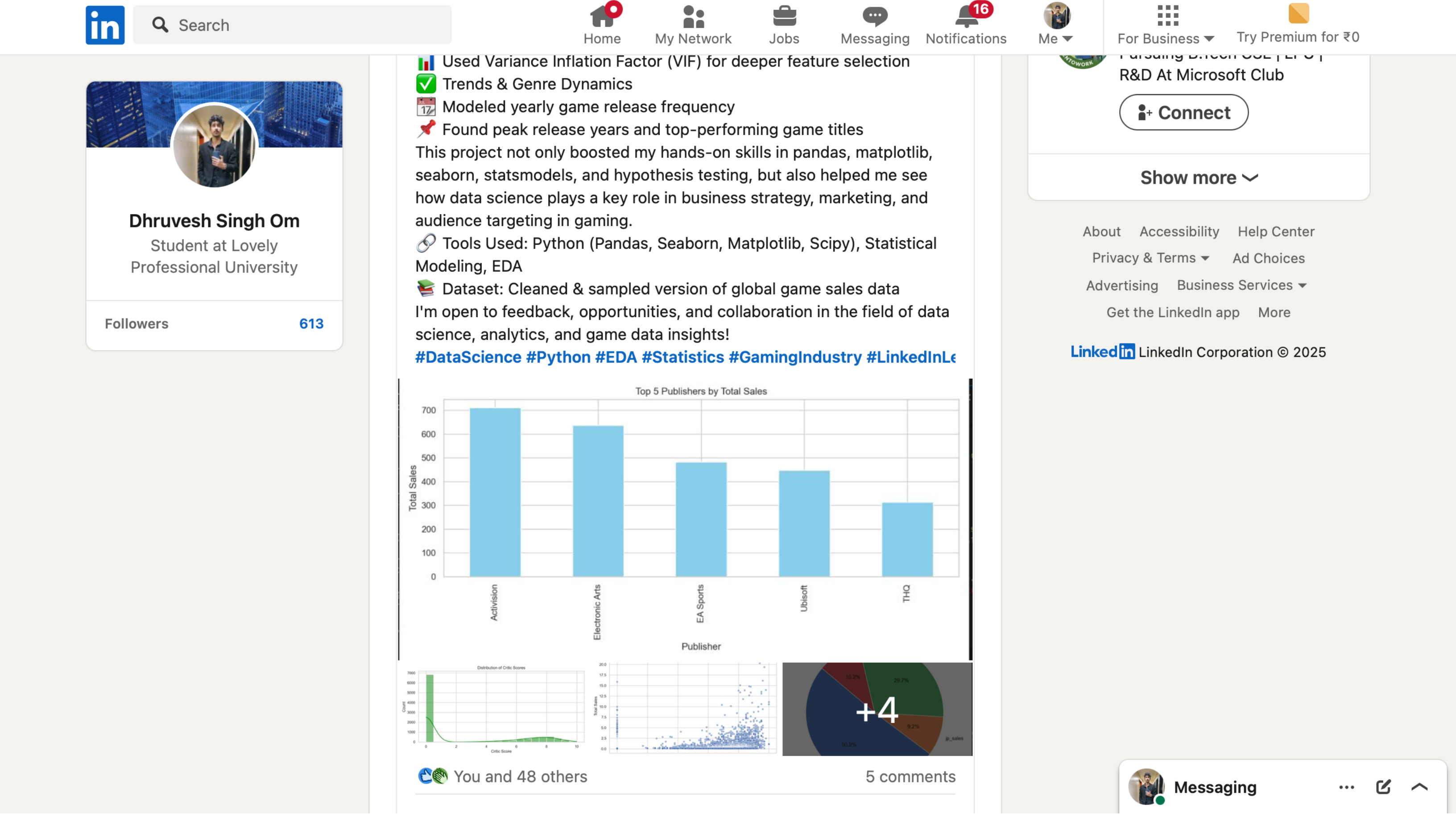
LinkedIn Pic



Dhruvesh Singh Om
Student at Lovely Professional University
Followers 613

Dhruvesh Singh Om · You
Student at Lovely Professional University
37m ·

🎮 Unveiling Insights from 2024's Global Video Game Sales Data
📊 Data Science Meets Gaming Analytics
Over the past few days, I've been diving deep into a dataset of 60,000+ video game entries, applying real-world data science techniques to extract meaningful insights from the world of gaming. Here's a glimpse of what I explored:
✅ Data Cleaning & Preparation
✔ Converted release dates, handled missing values, categorized sales
✔ Created new features like sales percentage and sales category
✅ Descriptive Analytics & Visualizations
📈 Plotted histograms, pie charts, scatter plots, and heatmaps
🏆 Identified top-selling genres and publishers
🎮 Action and Shooter genres dominated global sales
📍 North America led regional sales, followed by PAL and Japan
✅ Inferential Statistics & Hypothesis Testing
🧪 Shapiro-Wilk Test revealed that critic scores are not normally distributed
📊 T-Test showed a statistically significant difference between sales in Action vs Shooter games
🎯 A/B Testing between PS4 vs X360 sales provided platform-specific market insight
✅ Correlation & Multicollinearity Check
🔍 Found strong multicollinearity among regional sales
📊 Used Variance Inflation Factor (VIF) for deeper feature selection
✅ Trends & Genre Dynamics
📆 Modeled yearly game release frequency
📌 Found peak release years and top-performing game titles
This project not only boosted my hands-on skills in pandas, matplotlib, seaborn, statsmodels, and hypothesis testing, but also helped me see how data science plays a key role in business strategy, marketing, and



Dhruvesh Singh Om

Student at Lovely Professional University

Followers 613

📊 Used Variance Inflation Factor (VIF) for deeper feature selection
✅ Trends & Genre Dynamics
📆 Modeled yearly game release frequency
📌 Found peak release years and top-performing game titles

This project not only boosted my hands-on skills in pandas, matplotlib, seaborn, statsmodels, and hypothesis testing, but also helped me see how data science plays a key role in business strategy, marketing, and audience targeting in gaming.

🔗 Tools Used: Python (Pandas, Seaborn, Matplotlib, Scipy), Statistical Modeling, EDA

📚 Dataset: Cleaned & sampled version of global game sales data

I'm open to feedback, opportunities, and collaboration in the field of data science, analytics, and game data insights!

#DataScience #Python #EDA #Statistics #GamingIndustry #LinkedInLe

You and 48 others

5 comments

Pursuing B.Tech CSE | LPU | R&D At Microsoft Club