Data Analytics Capstone

Analysis of Video Game Sales

BHN1 Task 3: Project Report

Jared C. Plaisance

Western Governors University

05/06/2024

Table of Contents

A. Project Overview

My capstone project is aimed at unraveling the dynamics of video game sales. The project seeks to address the question of how video game sales vary across different platforms, genres, and regions over time and what factors contribute to the success of top-selling video games in each market. This is vital given the increasingly competitive landscape of the video game industry, where companies are continually striving to navigate the complexities of consumer preferences, technological advancements, and market trends.

The scope of the project consists of an analysis of different factors that influence video game sales. This includes examining platform preferences, genre preferences, and publisher preferences. By diving into these categories, the project aims to identify the key drivers that contribute to sales performance.

To achieve these objectives, the project utilizes a data-driven approach by leveraging a large dataset for sales data on video games. The dataset includes a wide range of variables, such as platform availability (e.g., PlayStation, Xbox, Wii), genre classifications (e.g., action, adventure, sports), and publisher classifications (e.g., Nintendo, Electronic Arts, Activision). Through data collection and analysis, the project aims to uncover insights that can inform strategic decision-making within the video game industry.

In terms of methodology, the project adopts a structured approach inspired by the CRISP-DM framework, which guides the various stages of the data analysis process. This includes data understanding, data preparation, exploratory data analysis, modeling, evaluation, and deployment. By following this methodology, the project ensures a systematic approach to data analysis, allowing for the identification of meaningful patterns and trends within the data.

In addition to leveraging statistical analysis techniques, the project also utilizes the Python programming language and various libraries such as Pandas, NumPy, Matplotlib, Seaborn, scikit-learn, and SciPy for data manipulation, visualization, and modeling. These tools enable the project to effectively handle and analyze large volumes of data, uncovering valuable insights that can inform strategic decision-making within the video game industry.

B. Project Execution

The execution of the project closely followed the plan developed in Task 2. The project plan's overall objectives and deliverables remained consistent, following a straightforward methodology. In terms of the project planning methodology, the project followed a traditional waterfall approach, where each phase of the project was completed sequentially before moving on to the next.

The project timeline and milestones underwent some modifications to accommodate changes in the speed of completion. The overall duration of the project was shortened, and certain milestones were adjusted to reflect the iterative nature of the analysis process. For example, evaluation and statistical modeling took less time than anticipated due to the cleanliness of the data source. As a result, subsequent phases of analysis were adjusted, resulting in a shorter end date.

Project Timeline

Milestone	Duration	Anticipated Start Date	Anticipated End Date
Initiation	1 Day	4/22/2024	4/23/2024
Data Collection	1 Day	4/23/2024	4/24/2024
Data Analysis	3 Days	4/24/2024	4/27/2024
Evaluation	4 Hours	4/27/2024	4/27/2024
Statistic Modeling	3 Hours	4/27/2024	4/27/2024
Visualization	2 Hours	4/28/2024	4/28/2024

Documentation	2 Hours	4/28/2024	4/29/2024
Feedback	1 Day	4/29/2024	4/30/2024
Presentation	1 Day	4/30/2024	5/1/2024

C. Data Collection Process

Our data selection and collection process began with thorough planning to ensure that we obtained a comprehensive dataset aligned with our project objectives. We identified a single, reliable source from which to gather the necessary data. This source, Kaggle, was chosen based on its availability for extensive and uncopyrighted information relevant to our analysis. However, as we progressed with the data collection process, we encountered various challenges that required us to adapt our approach. One significant challenge was the inconsistency in data formatting and quality within the dataset.

To address this issue, we developed data cleaning scripts for the specific inconsistencies encountered. These scripts enabled us to standardize the format of the data and ensure its integrity for further analysis. Additionally, we encountered occasional gaps in the data, which necessitated techniques to fill missing values while minimizing bias. Despite these challenges, we remained committed to obtaining a high-quality dataset that would serve as a solid foundation for our analysis.

Throughout the data collection phase, we also encountered unplanned data governance issues, primarily stemming from discrepancies in data definitions and terminology within the primary data source. To mitigate these issues, we conducted thorough data validation exercises. This involved scrutinizing the dataset to identify inconsistencies and ambiguities, which were then addressed through clarification of data definitions. By developing clear data governance protocols, we ensured that data was interpreted and used consistently throughout the project, hence guaranteeing consistency and reliability.

C.1 Advantages and Limitations of Data Set

The dataset utilized for our analysis presented various advantages as well as limitations that shaped the scope of our study. One notable advantage was its comprehensive coverage of the video game industry, encompassing a wide array of variables such as game genres, platforms, publishers, and regional data. This allowed for a thorough exploration of different parts of the industry, facilitating a more holistic understanding of its dynamics. Additionally, the dataset had a large sample size, providing a solid foundation for statistical analysis and strengthening the reliability of our findings. Its inclusion of historical data spanning multiple years further enabled an in-depth study, facilitating the identification of trends and patterns over time within the industry.

However, the dataset was not without its limitations. One notable constraint was the presence of incomplete or missing data, which could potentially introduce bias and undermine the accuracy of certain analyses. Additionally, while the dataset covered various aspects of the video game industry, it may have lacked certain variables or dimensions that could have provided deeper insights. Quality issues such as data entry errors or reporting biases also posed challenges, necessitating careful validation and verification of the data to ensure its reliability. Furthermore, the static nature of the dataset meant that it did not capture real-time or dynamic changes occurring within the industry, limiting our ability to assess emerging trends or recent developments accurately.

Despite these limitations, the dataset's open-access nature and reasonable level of consistency in data formatting and structure were notable strengths. Its accessibility facilitated the transparency and reproducibility of our analysis, allowing others to verify our findings and build upon our work. By acknowledging both the advantages and limitations of the dataset, we were able to conduct a more informed analysis, drawing meaningful conclusions while being mindful of the data's constraints.

D. Data Extraction and Preparation Process

Our data extraction and preparation processes ensured the acquisition of clean, structured data suitable for analysis. To begin, we downloaded the sales data named "vgsales.csv" from the video game sales dataset on Kaggle. Once the dataset was obtained, we started data cleaning and preprocessing. This involved a series of steps to address inconsistencies, errors, and missing values within the data. Techniques such as filling or dropping missing values, standardizing data formats, and data transformation were applied to enhance data quality and usability.

In executing these processes, we relied on the Python programming language and its associated libraries, particularly Pandas and NumPy, due to their versatile capabilities in data manipulation and analysis. Statistical methods were also employed to identify and address data quality issues effectively. Visualization tools such as Matplotlib and Seaborn were utilized to visualize data distributions, patterns, and relationships, aiding in exploratory data analysis and the validation of preprocessing steps. These tools and techniques were chosen for their compatibility with our dataset and the objectives of our project, enabling us to efficiently transform raw data into actionable insights.

Our approach to data extraction and preparation was guided by the recognition of common challenges associated with real-world datasets, such as inconsistencies, missing values, and outliers. By adopting a systematic approach to data cleaning and preprocessing, we ensured the reliability and validity of our analysis results. The use of appropriate tools and techniques facilitated the transformation of raw data into meaningful insights, which helped lay a solid foundation for our subsequent data analysis and interpretation.

E. Data Analysis Process

E.1 Data Analysis Methods

While analyzing the data, we employed a combination of descriptive and inferential statistical methods along with exploratory data analysis techniques. Descriptive statistics were used to summarize the main characteristics of the dataset, providing insights into the central tendency and distribution of key variables such as sales figures, genres, platforms, and publishers. This included measures such as mean, which helped to understand the overall trends and patterns present in the data. Exploratory data analysis techniques, including data visualization using histograms, box plots, and scatter plots, were employed to visually explore relationships and patterns within the dataset. This allowed for the identification of potential trends or outliers.

In addition to descriptive analysis, inferential statistical methods were used to test hypotheses and infer relationships between variables. This included techniques such as hypothesis testing (e.g., t-tests, ANOVA) to compare means across different groups or categories. These inferential methods allowed us to draw conclusions about the population based on the data, providing deeper insights into the factors influencing video game sales and market dynamics.

E.2 Advantages and Limitations of Tools and Techniques for Analysis

The tools and techniques utilized for data analysis provided a framework for exploring and deriving insights from the dataset. Their versatility allowed for a wide range of analytical tasks, from data manipulation to modeling, enabling a comprehensive examination of the data. Python, with libraries like Pandas and NumPy, were useful resources due to their functionalities and adaptability for analyzing the data. These tools helped facilitate the data preprocessing, which allowed us to address inconsistencies, missing values, and outliers effectively.

The limitations we faced during this project were visualizing the data into insightful graphs for us to use. This process was challenging due to trying to make the code less repetitive while keeping the particular features for each graph. We also found it difficult to find insightful statistical tests that aligned with our hypothesis. Through rigorous trial and error, we finally figured out the information that we were looking for.

Despite these challenges, the advantages of these tools enabled a thorough and insightful analysis of the dataset. By using the strengths of these tools while mitigating challenges, we were able to derive meaningful insights to inform decision-making processes within the video game industry.

E.3 Application of Analytical Methods

In applying the analytical methods outlined in Part E1 to the data, we followed a systematic step-by-step approach to ensure accuracy and reliability in our analysis. Initially, we conducted a descriptive statistic to summarize key characteristics of the dataset, such as the mean, using Pandas in Python. This provided an overview of the data's sales values across regions. Then, we employed exploratory data analysis (EDA) techniques, including histograms, box plots, and scatter plots generated using Matplotlib and Seaborn libraries. Through visualization, we analyzed the data distribution and identified potential outliers or patterns, which were essential for understanding underlying trends.

We then conducted hypothesis testing by using statistical tests like t-tests or ANOVA from the SciPy library. Should the assumptions not be met, we applied appropriate transformations as alternatives, ensuring the validity of our analysis.

By verifying the assumptions or requirements for each analytical method employed, we maintained the integrity and validity of our data analysis process, ultimately deriving meaningful insights and decision-making processes within the video game industry.

F. Data Analysis Results

F.1 Statistical Significance

In evaluating the success of our data analytics project, a significant aspect lies in examining the output of our analytical methods and the significance of the derived results. First, considering the t-test values obtained for comparing sales between different regions and genres, we observe compelling statistical evidence supporting significant differences. For instance, when comparing sales between the Action genre in North America (NA) and Europe (EU), the t-statistic of 8.78 and a p-value of approximately 2.11e-18 indicate a substantial disparity in sales figures between these regions. Similarly, when contrasting sales between the Action genre for NA and Japan (JP), the remarkably high t-statistic of 21.12 and a minuscule p-value of 7.59e-96 further emphasize significant discrepancies in sales. Furthermore, the comparison of sales between the Action genre for EU and JP yields a notable t-statistic of 14.42 and an exceedingly low p-value of 1.84e-46, reinforcing substantial variations in sales across regions.

In evaluating sales between the top publishers across different regions, the analysis extends beyond t-tests to include ANOVA tests, providing insights into the variance in sales attributed to various publishers. The F-statistic values, along with their corresponding p-values, reveal significant differences in sales between top publishers across NA, EU, and JP. For instance, the comparison of sales between top publishers in NA yields an F-statistic of 57.41 and a p-value of 3.38e-25, indicating substantial variability in sales attributed to different publishers in the region. Similarly, the F-statistic values for comparisons in EU and JP regions underscore significant variations in sales attributable to top publishers, with p-values indicating the statistical significance of these differences.

Furthermore, analyzing sales between top platforms across different regions provides additional insights into market dynamics. The F-statistic values obtained from ANOVA tests, along with their associated p-values, highlight notable discrepancies in sales between top platforms across NA, EU, and JP. Notably, the comparisons reveal

statistically significant variations in sales attributed to different platforms within each region, reflecting the diverse preferences and market dynamics shaping sales trends.

F.2 Practical Significance

The practical significance of our data analytics solution extends beyond the statistical findings, as it offers actionable insights that can shape strategic decisions within the video game industry. Through our analysis of sales data spanning different regions, genres, publishers, and platforms, we provide valuable guidance for businesses seeking to navigate and thrive in this competitive landscape. For instance, our examination of sales trends across regions highlights specific market dynamics, enabling companies to tailor their marketing and distribution strategies to specific geographic areas more effectively. By identifying regions with more demand for particular genres or platforms, companies can allocate resources strategically, maximizing market penetration and revenue generation.

Our analysis facilitates informed investment decisions by showcasing genre-specific trends and consumer preferences. With this knowledge, video game developers and publishers can prioritize resources towards genres with high growth potential, aligning their product portfolios with evolving market demands. This strategic alignment not only enhances profitability but also fosters long-term sustainability by ensuring that investments are directed towards areas of maximum return. Additionally, our insights into top-performing publishers and platforms offer opportunities for collaboration and partnership, enabling companies to leverage existing market dominance to expand their reach and diversify revenue streams.

F.3 Overall Success

The overall success and effectiveness of our project can be assessed through various aspects of how it delivered actionable insights and empowered decision-making within the video game industry. Firstly, our project demonstrated success in effectively leveraging data analytics techniques to extract valuable insights from a complex dataset. By employing a combination of descriptive statistics and inferential analysis, we were able to uncover meaningful patterns and trends in sales data across different regions, genres, publishers, and platforms. This laid the foundation for informed decision-making and strategic planning.

The project's effectiveness is evident in its ability to translate data-driven insights into actionable recommendations with tangible business impact. For instance, our analysis identified emerging market trends, enabling video game companies to reallocate resources, prioritize investments, and capitalize on untapped opportunities. Whether it was guiding marketing strategies, informing product development decisions, or facilitating partnerships with top-performing publishers and platforms, our project played a pivotal role in shaping how business strategies should be aligned with market dynamics.

Furthermore, the success of the project can be measured by its possible contribution to enhancing operational efficiency and driving competitive advantage within the video game industry. By providing timely and accurate insights, our project can enable companies to streamline operations, optimize resource allocation, and stay ahead of market trends.

G. Key Takeaways

G.1 Conclusion Summary

The analysis conducted in our project yielded several key conclusions that provide valuable insights into the video game industry's dynamics and market trends. Firstly, our examination of sales data across different regions highlighted significant variations in consumer preferences and market demand. Specifically, comparisons between North America (NA), Europe (EU), and Japan (JP) revealed distinct sales patterns, indicating the importance of tailoring strategies to specific geographic markets. Also, our genre-specific analysis identified trends in genre popularity, allowing companies to prioritize investments and resource allocation accordingly. Additionally, insights into top publishers and platforms showed the competitive landscape within the industry, offering opportunities for collaboration and strategic partnerships to expand market reach and revenue streams. Overall, our analysis emphasizes the importance of data-driven decision-making in navigating the complexities of the video game market, enabling companies to capitalize on emerging opportunities and stay ahead of evolving consumer trends.

G.2 Effective Storytelling Support

The tools and graphical representations chosen for visually communicating findings in our analysis were selected to support effective storytelling by enhancing clarity, facilitating comprehension, and engaging the audience. Firstly, tools such as Matplotlib and Seaborn were utilized for data visualization due to their versatility and customization options, allowing us to create visually appealing graphs and charts that effectively convey complex information. By selecting appropriate chart types, such as bar plots, line plots, and scatter plots, we were able to present data in a clear and concise manner, making it easier for the audience to interpret and understand the key insights.

The graphical representations employed in our analysis were designed to highlight trends, patterns, and relationships within the data, thereby enhancing storytelling by providing visual context and narrative structure. For example, histograms and box plots were used to illustrate the distribution of sales data and identify outliers, helping to convey the spread of sales figures across different categories.

Also, the use of color and labeling in our graphical representations served to emphasize key points and draw attention to important findings, enhancing storytelling by guiding the audience's focus and reinforcing key messages. By incorporating visual elements that are both aesthetically pleasing and informative, our chosen tools and graphical representations not only support effective communication of findings but also create a compelling narrative that resonates with the audience.

G.3 Recommended Courses of Action

Based on our analysis of the video game industry, we recommend two strategic courses of action to capitalize on emerging opportunities and address key challenges. Firstly, leveraging insights from our examination of regional sales trends, we advocate for a strategic approach to market expansion. This involves prioritizing regions with untapped potential and growing consumer demand, such as emerging markets in Asia-Pacific and Latin America. By allocating resources towards expanding their presence in these regions, companies can enhance market penetration and capture market share effectively.

Secondly, in light of our genre-specific analysis revealing shifting consumer preferences and evolving market trends, we propose a proactive approach to genre diversification and innovation. Video game companies should prioritize investments in new and emerging genres while exploring opportunities for hybrid genres that blend elements of traditional genres to appeal to broader audiences. Creating a culture of innovation and creativity within development teams can lead to the creation of

groundbreaking titles that resonate with consumers and differentiate companies from competitors.

By pursuing these courses of action, video game companies can position themselves for sustained growth and competitiveness in an increasingly dynamic and competitive market landscape. These recommendations, grounded in our analysis of sales data and market trends, provide actionable insights to guide strategic decision-making and drive business success in the evolving video game industry.

H. Panopto Presentation Link

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=af0cfbb8-d6cf-4fed-9da6-b16b001ab91a

I. Appendix A

I.1 GitHub Link to Jupyter Notebook File

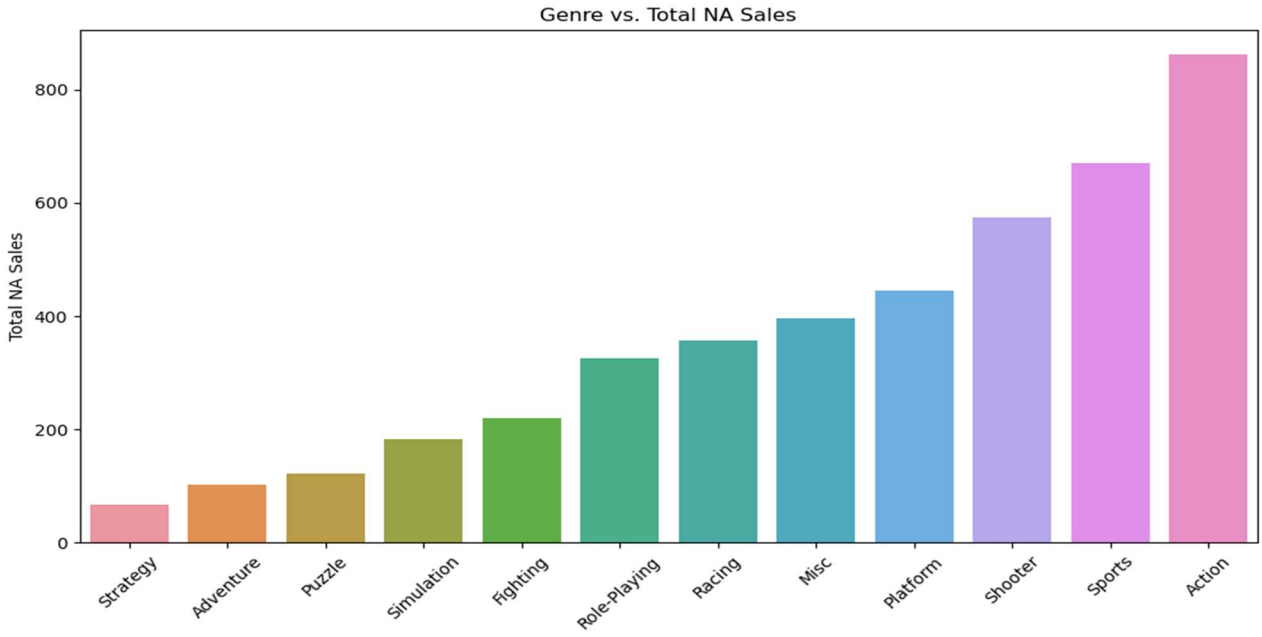
https://github.com/jaredplaisance/vgsales

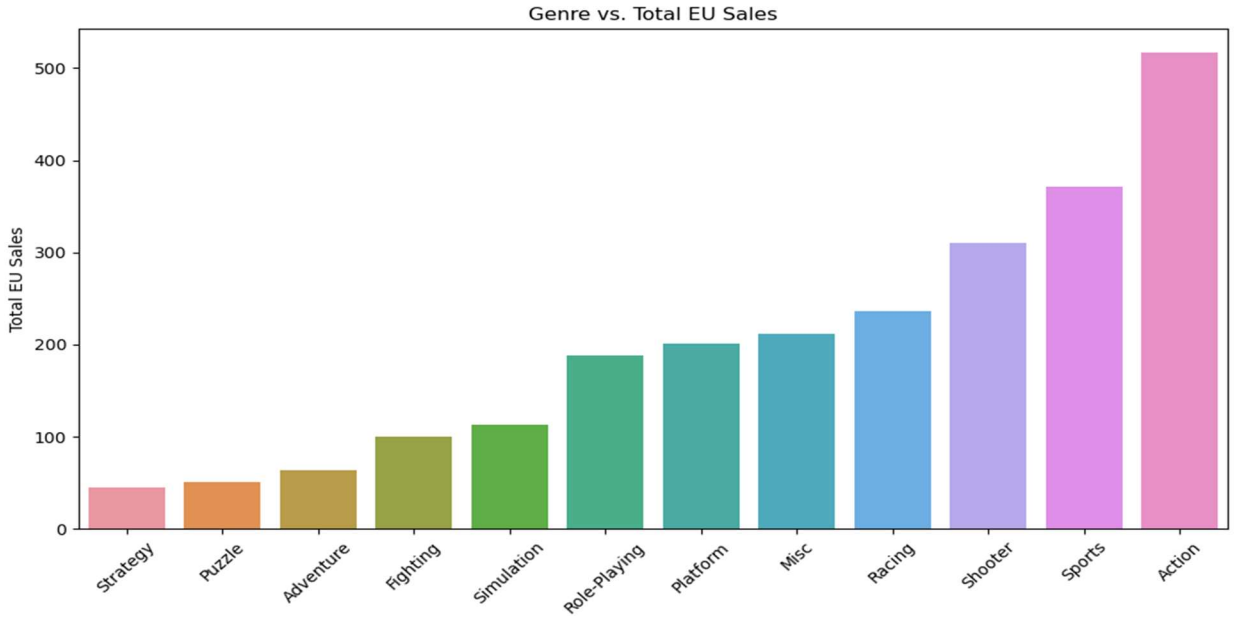
I.2 Kaggle Link to Video Game Sales Dataset

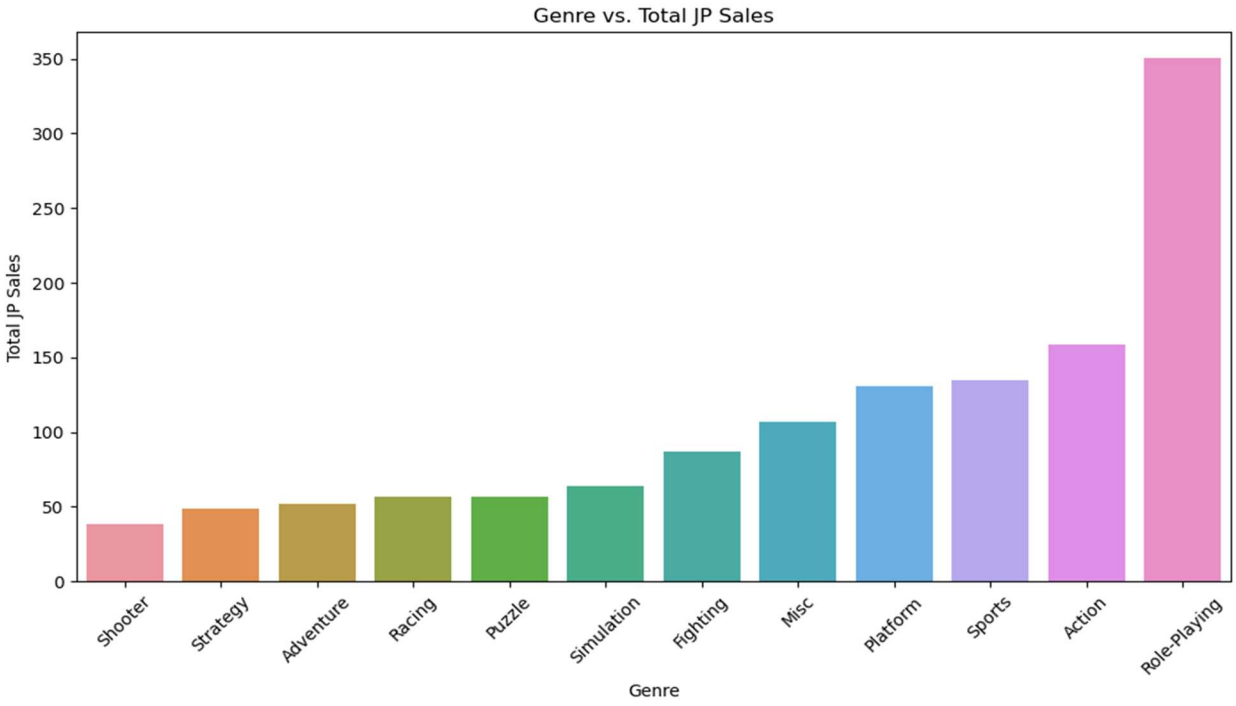
https://www.kaggle.com/datasets/gregorut/videogamesales/data

J. Appendix B

J.1 Genre vs. Regional Sales Bar Chart and T-Statistic/P-Value

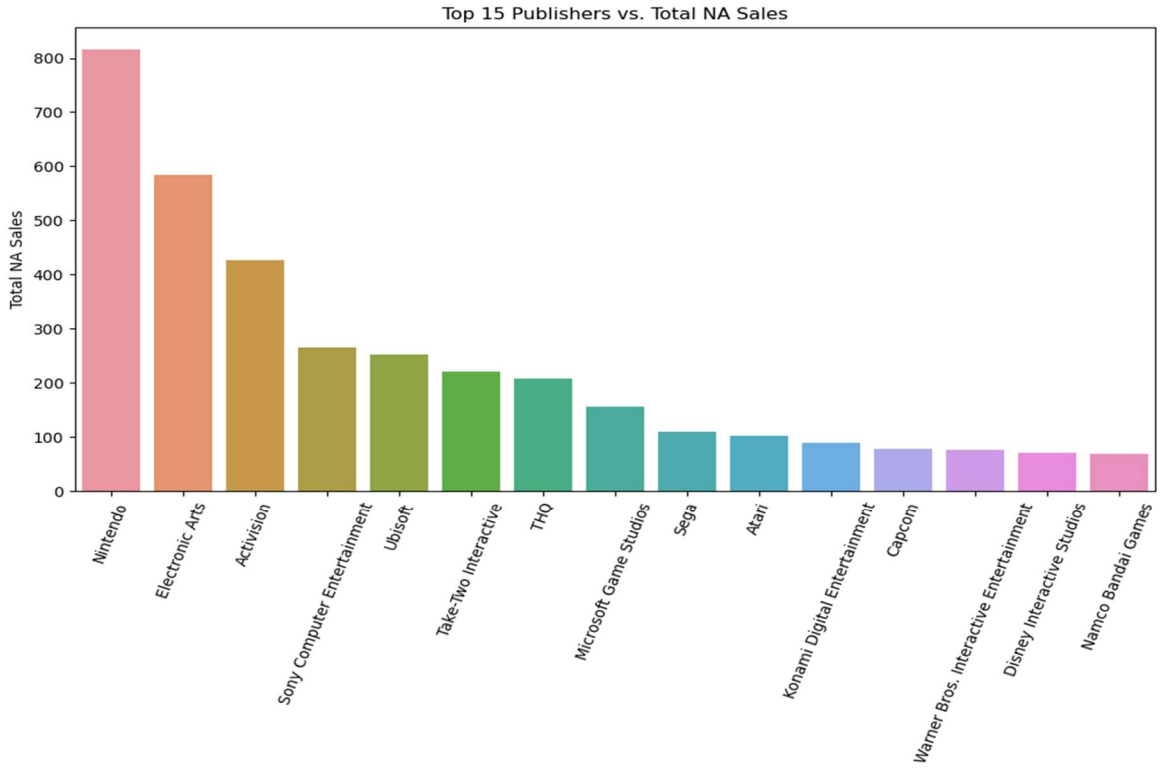


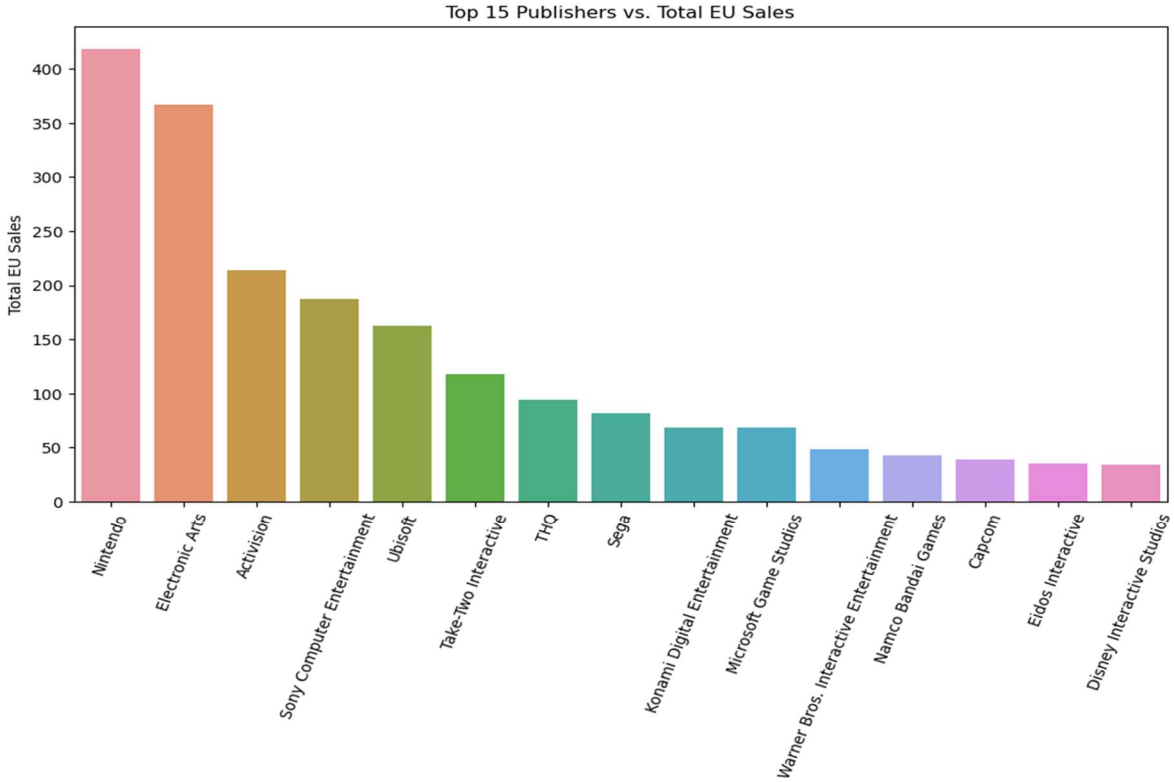


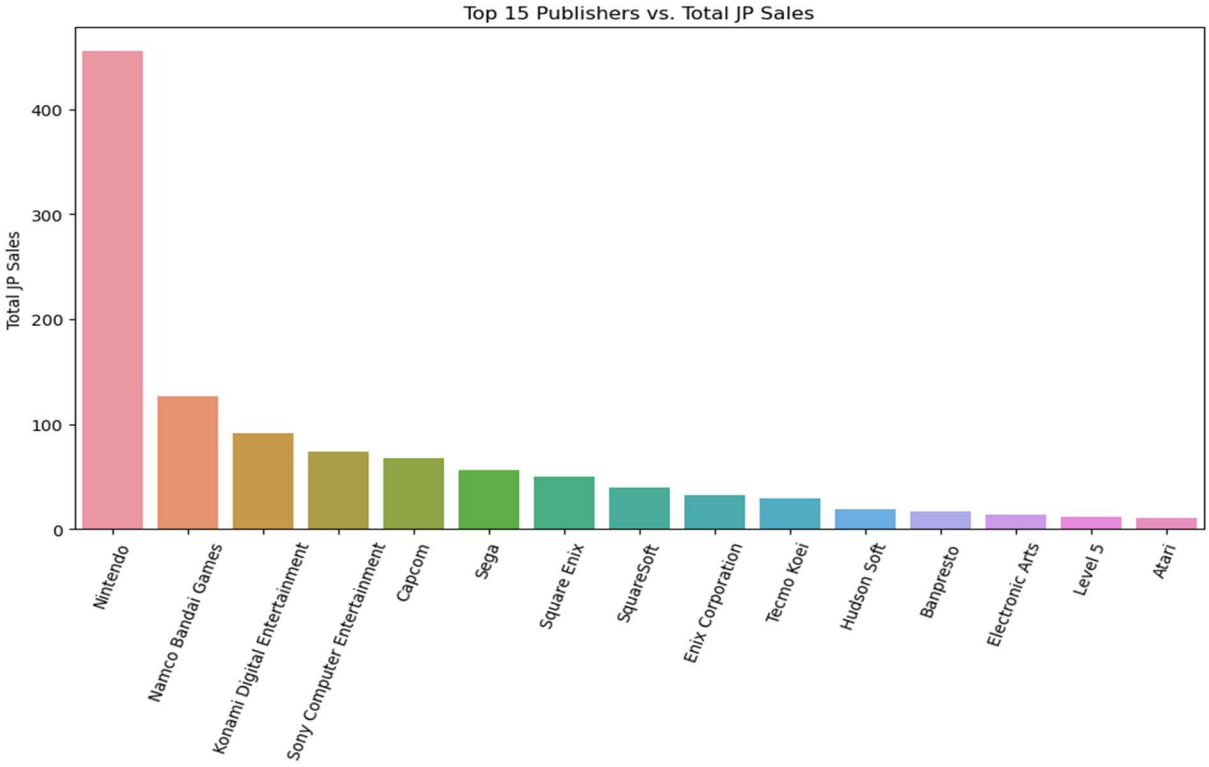



```
Comparison of Sales between Action Genre for NA and EU:
T-statistic: 8.77675736148825
P-value: 2.1129534291944e-18

Comparison of Sales between Action Genre for NA and JP:
T-statistic: 21.116490971849043
P-value: 7.590384980609329e-96

Comparison of Sales between Action Genre for EU and JP:
T-statistic: 14.42383829260818
P-value: 1.840614421008153e-46
```

J.2 Publisher vs. Regional Sales Bar Chart and T-Statistic/P-Value

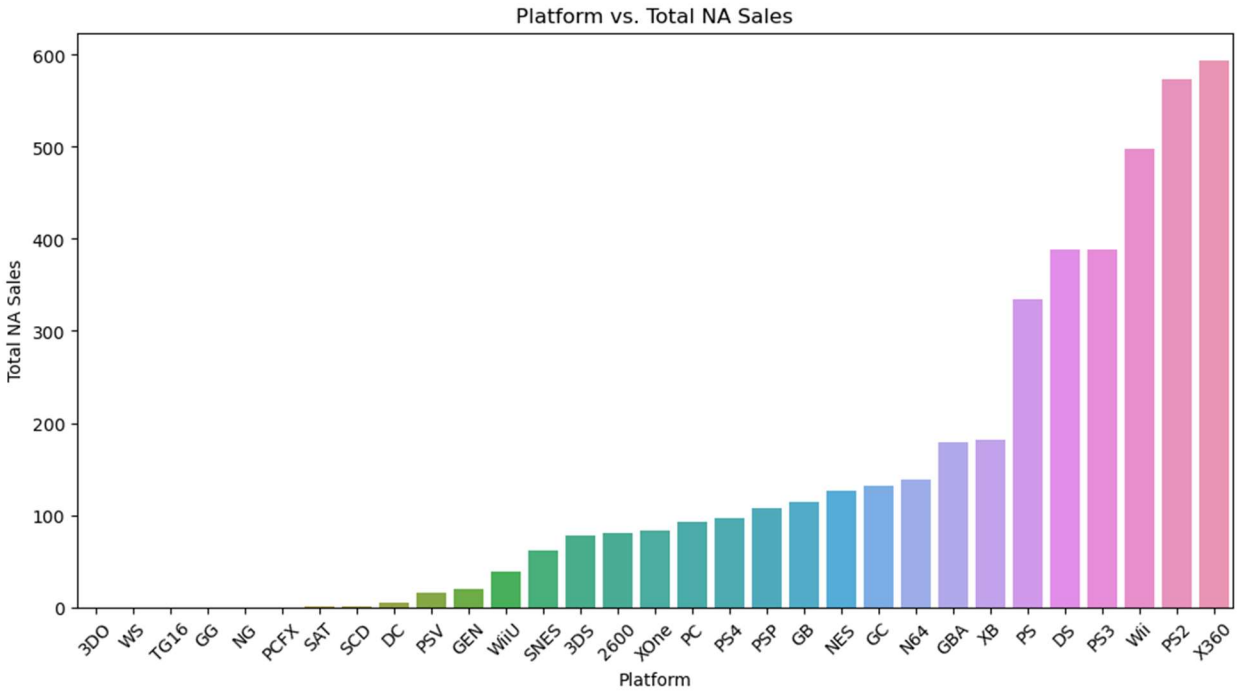


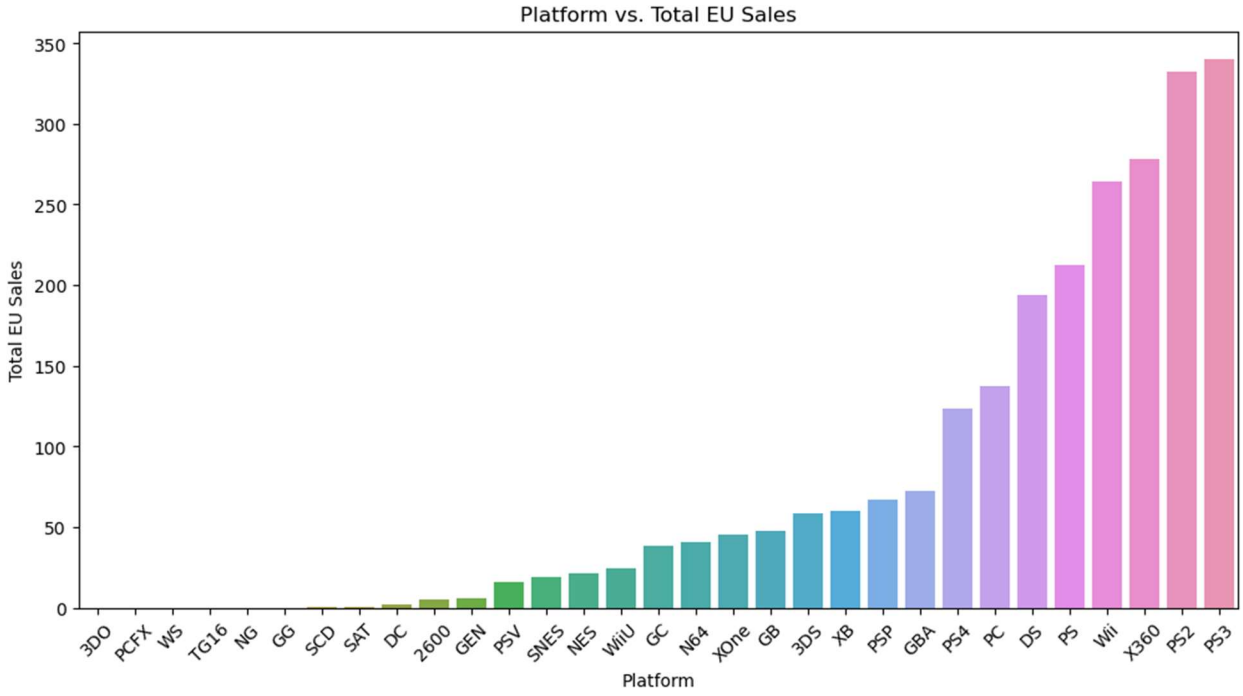


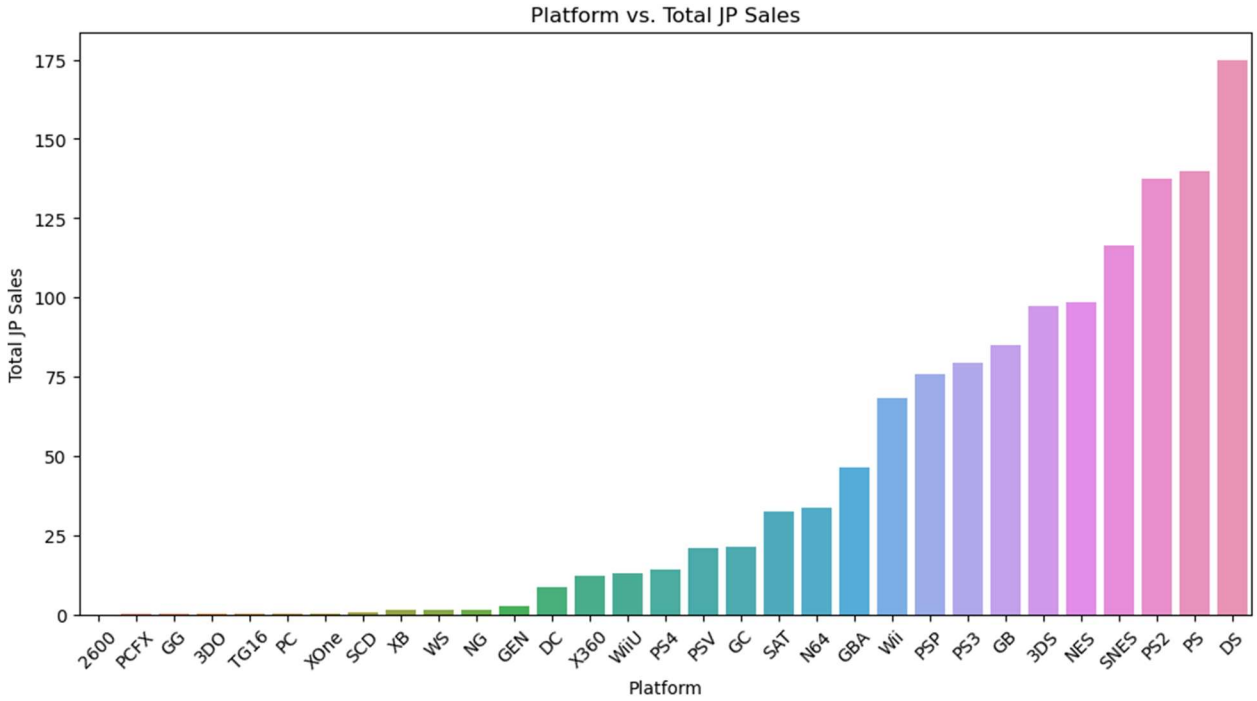



```
Comparison of Sales between Top Publishers in NA:
F-statistic: 57.40901936384362
P-value: 3.3803417518723048e-25

Comparison of Sales between Top Publishers in EU:
F-statistic: 34.469798399552474
P-value: 1.5773828233946234e-15

Comparison of Sales between Top Publishers in JP:
F-statistic: 197.28271192187242
P-value: 3.3638973257641887e-80
```

J.3 Platform vs. Regional Sales Bar Chart and T-Statistic/P-Value

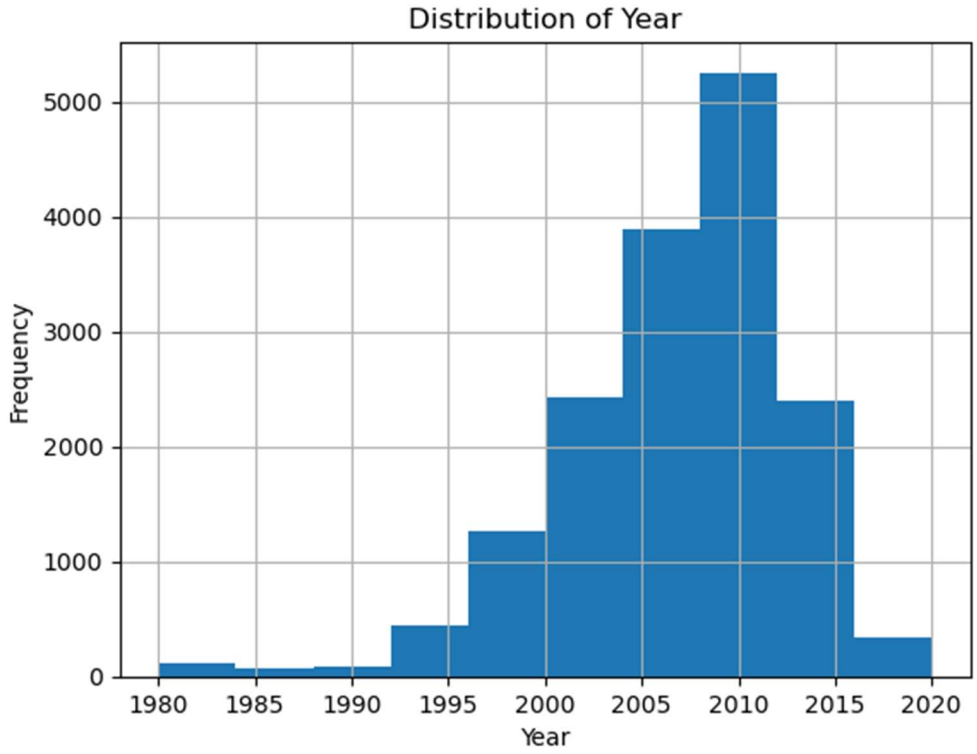




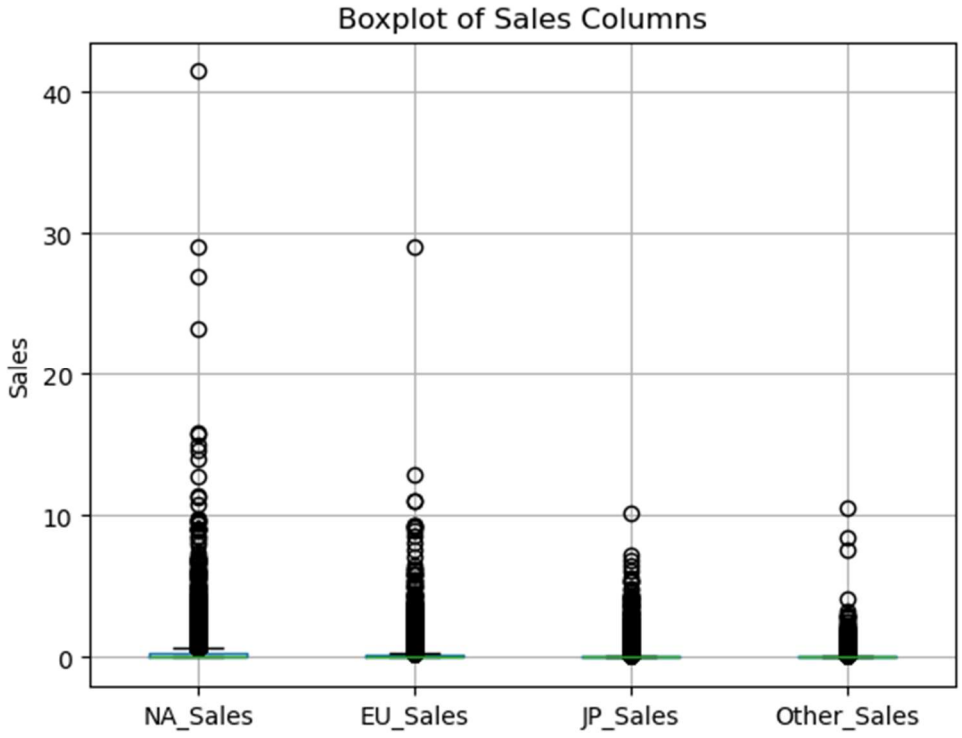



```
Comparison of Sales between Top Platforms in NA:
F-statistic: 15.942212922362598
P-value: 1.2575595312588582e-07

Comparison of Sales between Top Platforms in EU:
F-statistic: 20.504793842821364
P-value: 1.3585854385526925e-09

Comparison of Sales between Top Platforms in JP:
F-statistic: 10.856837637038815
P-value: 1.9687463582561725e-05
```

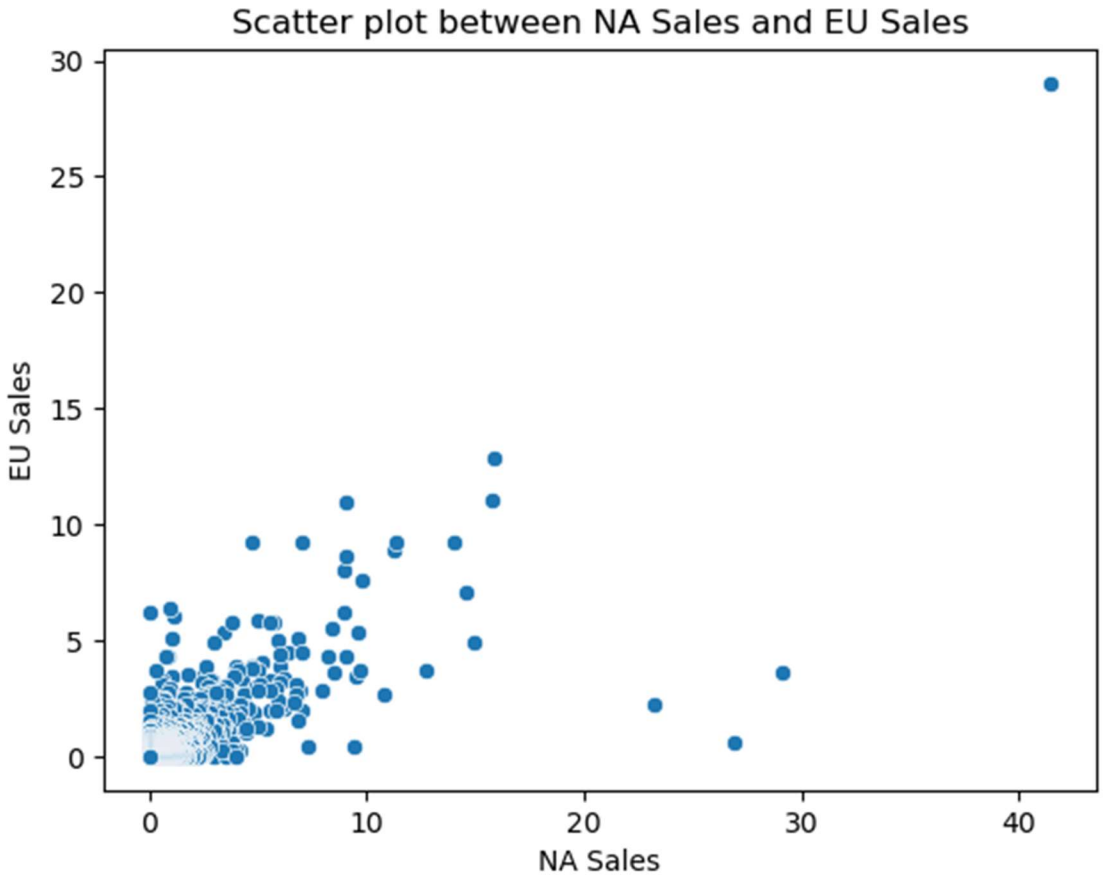
J.4 Frequency of Sales per Year



J.5 Sales Per Region



J.6 Sales Between North America and Europe



K. References

Babb, Jeffry & Terry, Neil & Dana, Kareem. (2013). The Impact Of Platform On Global Video Game Sales. International Business & Economics Research Journal (IBER). 12. 1273. 10.19030/iber.v12i10.8136. https://www.researchgate.net/publication/297754899_The_Impact_Of_Platform_On_Global_Video_Game_Sales

Khaleghi, Kaveh & Lugmayr, Artur. (2012). Video game market segmentation based on user behavior. 283-286. 10.1145/2393132.2393194. https://www.researchgate.net/publication/259287561_Video_game_market_segmentation_based_on_user_behavior

Sacranie, John (2010) "Consumer Perceptions & Video Game Sales: A Meeting of the Minds," The Park Place Economist: Vol. 18 https://digitalcommons.iwu.edu/parkplace/vol18/iss1/12/