INST414 Sprint 2 Report - Seamus Sullivan

Project Title: Video Game Ratings, Sales, and Genre Analysis
Research Question: How do critic and user review scores relate to video game sales across different genres?
Track: ADSP
Link to GitHub Repository:
https://github.com/seamusgsullivan/Video-Game-Ratings-Sales-and-Genre-Analysis
(This repository is different from the one used in Sprint 1, as my research question changed. However, it still follows the Cookiecutter project structure.)

Data Acquisition & Description

Confirmation of Data Access

I successfully obtained my primary dataset, "Video Game Sales with Ratings" by Rush Kirubi, from Kaggle (https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings). I downloaded the CSV file through Kaggle's "Download" button and loaded it into a Jupyter Notebook using the Pandas library (pd.read_csv()). The dataset loaded correctly with all columns and values visible. No access issues occurred.

Source and Collection Methodology

The dataset was compiled by Rush Kirubi, who combined data from two main sources:

- VGChartz, which tracks video game sales by platform and region.
- Metacritic, which provides critic and user review scores.

According to the dataset's Kaggle description, Kirubi's work was motivated by Gregory Smith's earlier web scrape of VGChartz video game sales (https://www.kaggle.com/datasets/gregorut/videogamesales), which served as the foundation for the sales data. Kirubi then extended the dataset by performing a separate web scrape of Metacritic, adding review-based variables such as critic and user scores, review counts, developer, and ESRB rating. The author also acknowledged using a public GitHub repository that assisted with the scraping process. The resulting dataset combines commercial sales data with review information for thousands of video games.

Temporal and Spatial Coverage

- Temporal coverage: The dataset covers games released from approximately 1980 to 2016, based on the Year_of_Release field. Three games list 2017 as the release year, and one lists 2020. Several entries do not include a release year.
- Spatial coverage: The dataset has a global scope, with regional sales fields for North America (NA_Sales), European Union (EU_Sales), Japan (JP_Sales), and everywhere else (Other_Sales), plus worldwide sales (Global_Sales).

Unit of Analysis

Each row/entry represents a single video game title released on a specific platform. For example, "FIFA 14" on Xbox 360 and "FIFA 14" on PlayStation 3 are recorded as separate entries in the dataset. This setup was kept on purpose because review scores can be different (and usually are) for each platform. Keeping these separate rows makes the data more accurate and helps show how scores and sales relate across different genres.

Sample Size

The raw dataset contains 16,719 rows/entries. For my project, I will only use rows that include non-missing values for Genre, Global_Sales, Critic_Score, and User_Score, since these are key to my research question. After initially cleaning the data, the initial cleaned dataset contains 8,099 rows/entries.

Variable Inventory

Variable Name	Data Type	Description	Relevance to Research Question	Missing Data %
Name	Text	Title of the video game	—	0.01%

Platform	Text, Categorical	Gaming platform (e.g. PS3, Wii, DS)	—	0.00%
Year_of_Release	Integer, Continuous	Year the game was released	—	1.61%
Genre	Text, Categorical	Game genre (e.g. Action, Role-Playing, Sports)	★ Feature candidate - grouping variable	0.01%
Publisher	Text, Categorical	Company that published the game	—	0.32%
NA_Sales	Float, Continuous	Game sales in North America (in millions of units)	—	0.00%
EU_Sales	Float, Continuous	Game sales in the European Union (in millions of units)	—	0.00%
JP_Sales	Float, Continuous	Game sales in Japan (in millions of units)	—	0.00%
Other_Sales	Float,	Game sales in	—	0.00%

	Continuous	the rest of the world, excluding NA, EU, and JP (in millions of units)		
Global_Sales	Float, Continuous	Total worldwide sales (in millions of units)	★ Target variable	0.00%
Critic_Score	Float, Continuous	Aggregate critic score from Metacritic (0-100)	★ Feature candidate - predictor variable	51.33%
Critic_Count	Integer, Continuous	Number of critic reviews used to compute Critic_Score	—	51.33%
User_Score	Float, Continuous	Aggregate user score from Metacritic (0-10)	★ Feature candidate - predictor variable	54.60% (initially 40.10% before converting “tbd” entries to null in final cleaning)
User_Count	Integer, Continuous	Number of user reviews used to compute User_Score	—	54.60%

Developer	Text, Categorical	Company or team that developed the game	—	39.61%
Rating	Text, Categorical	ESRB rating (e.g. E for everyone, T for teen, M for mature)	—	40.49%

Secondary Data Source

I will not use a secondary dataset for this project. The primary dataset by Rush Kirubi already includes everything I need, such as sales numbers, critic and user review scores, and game genres. It gives enough information to explore my research question without adding any other data sources.

Data Quality Assessment & Cleaning

Missing Data

Some variables in the dataset have missing values. The percentage of missing values for each variable is shown in the Variable Inventory table. Most missingness appears to be systematic, especially for review scores, since older games often lack Metacritic data. For handling missing data, I dropped rows that were missing Genre, Global_Sales, Critic_Score, or User_Score, because all are essential for my analysis (as explained in the Sample Size section). Other missing values were left as-is since they are not critical for my research question. As a result, all remaining rows include the key variables needed to study how critic and user scores relate to sales across genres, so that my analysis uses complete and relevant data. No additional imputation or deletion was needed, since all rows in the cleaned dataset now have values for the key variables - Genre, Global_Sales, Critic_Score, and User_Score. This approach is appropriate because my analysis focuses on the relationship between review scores, genre, and global sales.

By keeping only rows with complete values for these key variables, I make sure that the results are based on accurate and relevant data without introducing bias from imputation.

Data Quality

I checked for outliers in the numeric variables used in my analysis (Global_Sales, Critic_Score, and User_Score) using boxplots. I found some outliers - games that have extremely high sales compared to most others. These are genuine extreme values, not errors, so they were kept. For duplicates, I found that the dataset has no duplicate rows/entries. After cleaning the dataset to include only rows with non-missing Genre, Global_Sales, Critic_Score, and User_Score:

- Genre: All entries match standard genre categories. There are no formatting or conflicting information issues.
- Global_Sales: All values are non-negative. Extremely high sales values exist but are genuine and retained.
- Critic_Score and User_Score: The two scores originally use different scales (Critic_Score 0-100, User_Score 0-10) and User_Score was stored as an object because some entries were non-numeric (e.g., 'tbd'), unlike Critic_Score. To make them comparable and usable for analysis, User_Score was first converted to numeric and then scaled to 0-100.

No other measurement issues, proxy variables, or definitional changes were identified in the variables relevant to my analysis. I also found that there are no inconsistencies in formatting, conflicting information, or data entry errors in the key variables used.

Feature Engineering (Initial)

I did not create any new variables. However, I transformed one variable to make the data more consistent. The User_Score column was originally stored as an object because some entries were listed as "tbd" (to be determined). These "tbd" entries were converted to null values and removed along with other null/missing data during the earlier cleaning step (explained in the Sample Size and Missing Data sections). This issue was discovered after the initial cleaning, however, so the earlier cleaning process was rerun to make sure that all "tbd" entries were removed and the column could be converted to numeric format and analyzed properly. Afterward, I rescaled User_Score from a 0-10 range to a 0-100 range to match Critic_Score. This makes the two variables directly comparable and easier to visualize together in later analysis. No other feature engineering was done yet.

Data Cleaning Pipeline

Original dataset size: 16,719 rows

- Step 1: Inspected missing values for all variables to identify where null entries occurred and understand overall data completeness.
- Step 2: Dropped rows missing any of the key variables - Genre, Global_Sales, Critic_Score, or User_Score - since all are essential for analysis.
- Step 3: Identified that User_Score was stored as an object due to non-numeric "tbd" entries. These entries were converted to null values, and the column was transformed into numeric format. The User_Score values were then scaled from a 0-10 to a 0-100 range to match Critic_Score. After this fix, the missing value percentages were recalculated, and User_Score's missing percentage aligned with User_Count's, confirming the correction (the Variable Inventory table was updated accordingly).
- Step 4: Reran the earlier cleaning process (Step 2) to remove the new null entries and make sure that all remaining rows had complete data for the key variables.
- Step 5: Checked for outliers using boxplots and verified that extreme values (especially in sales) were genuine.
- Step 6: Verified there were no duplicate rows or inconsistent formatting in key variables.

Final cleaned dataset size: 7,017 rows

Critical Thinking Question

My cleaning decisions can introduce bias by changing which games are included in my analysis. Since I removed all rows missing Genre, Global_Sales, Critic_Score, or User_Score, older or less popular games without review data were excluded. This means the cleaned dataset is biased toward newer or higher-profile games that have both critic and user reviews on Metacritic. As a result, the findings about how scores relate to sales may not fully represent the entire gaming market, especially older titles or those released before review aggregation became common.

Code Note:

All code for the **Data Acquisition & Description** and **Data Quality Assessment & Cleaning** sections can be found in Sprint_2_Cleaning.ipynb, which is located in the “notebooks” folder of my GitHub repository. All outputs should already be included in the notebook, but you can rerun the cell to reproduce them if needed.

Exploratory Data Analysis

Univariate Analysis

I first looked at the key numeric variables: global sales, critic scores, and user scores. Most games sell under 1 million copies, with a few very high-selling games creating a right-skewed distribution for global sales (Figure 1). Critic scores are roughly normally distributed, mostly between 60 and 80 (Figure 2), and user scores generally fall between 60 and 90, showing that players tend to give positive reviews (Figure 3). Summary statistics match these patterns. The median global sales is low, showing that most games sell modestly, while the mean is slightly higher because of the outliers. Critic and user scores have moderate spread, with standard deviations around 13-14 points. Looking at game counts by genre, Action, Sports, and Shooter games are the most common, while Puzzle, Adventure, and Strategy games appear less often (Figure 4). The summary statistics for global sales by genre show that Adventure and Strategy games generally have lower median sales and less variation compared to other genres, while genres like Sports, Miscellaneous, and Platform have higher sales and more spread.

These univariate analyses show that most games sell modestly, but there is still enough variation in sales and review scores to explore their relationship. They also indicate that while critic and user scores vary, most games receive fairly positive reviews, which may limit the ability to detect strong differences in sales based on review scores alone. Additionally, since some genres are much more common than others, and some tend to have higher sales and more variation while others have lower sales and less variation, genre could affect the patterns seen between review scores and sales.

Bivariate/Multivariate Analysis

I next looked at how variables relate to each other and to global sales. Scatterplots and hexbin plots show that critic scores have a weak positive relationship with sales - higher-scoring games tend to sell a little more (Figures 5-6). User scores have an even weaker link to sales (Figures 7-8), so player ratings alone do not explain much about game success. The correlation matrix confirms this. Both critic and user scores only show a weak positive correlation with sales (Figure 9). Boxplots of sales by genre show that Adventure and Strategy games have less variation in sales compared to other genres (Figure 10). The pairplot shows a slight positive relationship between critic scores and sales, but user scores still have low correlation with sales (Figure 11). A boxplot of games with sales over 5 million shows that even high-selling games mostly cluster near the lower end of that range (Figure 12).

Overall, these analyses indicate that the features - critic scores, user scores, and genre - relate differently to the target variable (global sales). Critic scores are modestly related to global sales, while user scores are less predictive. Genre also plays a role, as discovered in the Univariate Analysis, with some genres showing higher sales and more variation, while others show lower sales and less variation. However, no single feature fully explains variations in sales.

Visualization Best Practices

All figures include clear titles, labeled axes, and captions that explain what the graph shows. I used chart types that match the data: histograms for distributions, scatterplots for relationships, and boxplots to compare groups. I kept font sizes and figure sizes readable.

Surprising Findings

I did not expect to see that quite a few games with low critic or user scores still sold relatively well, and that quite a few games with high critic or user scores did not sell very much. These patterns challenge my initial assumption that higher review scores would consistently correspond to higher sales. Additionally, the relationship between review scores and sales is weaker than I anticipated. This suggests that reviews alone are not a reliable predictor of game success, and other factors, such as genre, likely play an important role in determining game success.

Data Limitations Discovered

There are some limitations to this data. I can't fully answer why some games sell more than others because important factors like marketing budget, franchise popularity, or production costs are not included. These are variables I wish I had, as they would help show what drives sales beyond review scores and genre. Some genres, like Puzzle, have smaller sample sizes, which makes comparisons across genres less reliable. Overall, these limitations mean that while I can explore relationships between review scores, genre, and sales, I cannot account for all the factors that influence a game's performance.

Track-Specific Focus

I used a Random Forest model to explore how features relate to global sales. Critic and user scores were treated as features, and the model showed that critic scores were slightly more important than user scores for predicting sales (Figure 13). This indicates some predictive signal, but it is relatively weak. I also grouped sales into Low, Medium, and High tiers. Most games fall into the Low tier, revealing a class imbalance in the data (Figure 14). This imbalance means that predictions for higher-sales games may be less reliable because they are underrepresented in the dataset. Overall, critic and user scores provide some information about sales, but their predictive power is limited. Other factors not included in the dataset, such as marketing or franchise popularity, likely play a bigger role in determining which games sell more.

Code Note:

All code for the **Exploratory Data Analysis** section can be found in Sprint_2_EDA.ipynb, which is located in the "notebooks" folder of my GitHub repository. All outputs should already be included in the notebook, but you can rerun the cell to reproduce them if needed.

Refined Problem Statement & Analytical Plan

Revised Problem Statement

Shortly after starting Sprint 2, I decided to change my original research question. When reviewing the discussion section of the Kaggle dataset I am using, I found that someone had already explored almost the same research question I initially planned to do in Sprint 1, which was: Can we predict how well a video game will sell before or shortly after release, using features such as genre, platform, publisher, release year, and critic/user ratings? Because of that, I thought it would be better to take a different approach and focus on a question that still connects to the dataset but hasn't been explored by others. I also realized that including too many features could make the project harder to manage and interpret. So, I simplified the focus to a smaller set of key variables that are directly measurable and meaningful. My new research question is: How do critic and user review scores relate to video game sales across different genres?

This question stays consistent with my overall goal of understanding what factors might influence game sales but keeps the analysis more focused and realistic. I kept this same research question after doing my exploratory data analysis (EDA), because the results supported its importance. The patterns I found - such as a weak but visible relationship between critic review scores and sales, and noticeable differences across genres - supported my decision to keep the same research question. There was no need to change it further after EDA.

Updated Analytical Approach

Based on what I found in EDA, the relationship between review scores and sales appears weak and possibly nonlinear, with a slightly stronger connection for critic scores than for user scores. Because of this, tree-based models such as Random Forest and Gradient Boosting are the most suitable techniques. These models can capture complex patterns and interactions between variables better than a simple linear regression, which struggled to fit the data well in EDA. I may still include a linear regression baseline for comparison. For evaluation metrics, I will use R^2 (to measure how much variation in sales the model explains) and RMSE (to measure the average size of prediction errors). These metrics are appropriate because sales is a continuous variable, and together they show both how well the model fits and how accurate its predictions are. For data splitting, I will use an 80/20 train-test split to evaluate model performance on unseen data. If I add more complex models later, I may introduce a validation set or use cross-validation to tune hyperparameters more reliably. For feature selection, I will focus on a small, meaningful set of variables: critic score, user score, and genre. I will use correlation checks, feature importance rankings, and model interpretability results to confirm that these features are meaningful and useful. Additionally, since I already created sales tiers (Low, Medium, and High) in EDA to explore class imbalance, I plan to extend this by building a classification model that predicts these tiers. This may reveal clearer patterns than trying to predict exact sales numbers. In all modeling and analysis steps, I will emphasize interpretability over accuracy - the goal is to understand whether higher review scores are linked to higher sales, and whether that relationship differs across genres.

Challenges & Mitigation Strategies

So far, one obstacle I've encountered is that critic and user scores only weakly correlate with global sales. This emerged during EDA, where I noticed the relationship between review scores and sales is modest for critic scores and even weaker for user scores. To address this in Sprint 3, I will explore these relationships using tree-based models and sales tiers to identify any patterns, differences across genres, and surprising trends, rather than focusing on perfect prediction. If the primary regression approach shows low predictive power, my backup plan is to use a classification model based on sales tiers (Low, Medium, High), which may reveal clearer patterns.

Another concern from EDA is class imbalance, since most games fall in the Low sales tier. To mitigate this, I will consider balancing methods for classification models, such as adjusting class weights or resampling. Finally, to make my analysis clear and easy to follow, I will keep all work organized and well-commented inside Jupyter notebooks so others can understand or reproduce the results.

Progress Tracking & Next Steps

Updated Timeline

- Sprint 2 accomplishments: Cleaned the dataset✅, explored missing values✅, created visualizations as part of exploratory data analysis (EDA)✅, summarized distributions✅, identified weak relationships between review scores and sales✅, found several differences in sales across genres✅, created sales tiers✅, and ran preliminary Random Forest models✅.

- Sprint 3 rough plan: I plan to fully implement predictive models, including Random Forest and Gradient Boosting for regression and classification. I will evaluate performance using R^2, RMSE, and classification metrics. I will also document all steps clearly in Jupyter notebooks and verify that my selected features (critic score, user score, and genre) are useful and meaningful.
- Sprint 4 rough plan: I hope to uncover how critic and user scores are linked to global sales across different genres, identifying which genres show that higher review scores correspond more strongly (or less strongly) to higher sales. I will compare patterns between critic and user reviews, note any surprising trends (such as high-selling games with low review scores), and summarize key insights about how review scores relate to sales. Visualizations and clear summaries will support these findings, making it easier to understand patterns and differences across genres.

Self-Assessment

- I am currently on track with the timeline and have completed all Sprint 2 goals.
- The biggest risk right now is that review scores may not strongly predict sales, which could limit the predictive power of some models, but my analysis can still provide useful insights into patterns and differences across genres.
- I don’t need any extra support or resources right now.