CDS 101 – Final Project Report

Werner Wyss - Kanita Haque - Hussein Hamdan

December 08, 2025

Contents

1. Problem Definition **1**

2. Data Acquisition & Description **2**

3. Data Cleaning & Preprocessing **2**

4. Exploratory Data Analysis (EDA) **4**

5. Visualization Quality and Storytelling **8**

6. Modeling Approach **9**

7. Model Implementation & Evaluation **9**

8. Conclusions & Recommendations **9**

9. Code Quality & Reproducibility **9**

10. References **9**

Appendix (Optional) **9**

1. Problem Definition

The main question of this project is: Which factors help predict the revenue of a video game? Using the Gaming Industry Trends dataset, we investigate how platform, genre, release year, and other features relate to a game's financial performance.

This question is interesting because the gaming industry continues to grow, and understanding the features associated with higher revenue can help studios, marketers, and developers make informed decisions. Predictive insights can also help identify which types of games tend to perform well in the market.

The objective of this analysis is to build linear models that use game characteristics to estimate revenue and to identify which variables are most strongly associated with higher or lower revenue. Our goal is not to

perfectly predict sales but to understand which factors contribute meaningfully to revenue patterns in the dataset.

Key assumptions include:

- The dataset is accurate and represents industry trends fairly.
- Revenue is influenced by the variables included (platform, genre, release year, etc.).
- Linear modeling is appropriate for exploring these relationships.

2. Data Acquisition & Description

The dataset is called Gaming Industry Trends and the source for the dataset is Kaggle. We obtained the dataset through the download link on the Kaggle website. The main features would include game title, genre, platform, release year, and revenue, along with possible additional fields describing popularity or rating. The approximate number of rows and columns would be 1000 rows and 11 columns. Some biases to take into consideration would be sampling bias. The dataset would subject to sampling bias due to the limited amount of data given (only 1000 rows). This constraint can exclude large portions of the gaming industry, particularly small studios, niche genres, and emerging platforms.

As a result, the observed trends may reflect more mainstream titles.

3. Data Cleaning & Preprocessing

The analysis required us to perform data cleaning operations which verified the accuracy and usability of all information in the dataset. The first step involved changing multiple column names because their original names included periods which made them hard to understand. The new names improved the dataset's readability.

The analysis required us to eliminate all entries containing missing essential data points including revenue and genre and platform and release year. The games require these essential variables to perform analysis. The analysis excluded all entries containing invalid data points because real-world data requires positive values for players and revenue.

The process involved transforming multiple columns into their appropriate data formats. The analysis required us to convert Genre and Platform into factor variables because they represent categories and Release_Year into an integer data type.

The dataset became ready for advanced analysis after we finished all necessary cleaning operations which ensured data consistency and accuracy. The cleaning process protects our analysis from future errors which results in better reliability for our visualizations and models.

```r
library(dplyr)

# Read in the raw CSV
gaming_raw <- read.csv("gaming_industry_trends.csv")

# Rename messy column names with periods to cleaner names with dashes
gaming_raw <- gaming_raw %>%
  rename(
    Game_Title              = Game.Title,
    Release_Year            = Release.Year,
    Revenue_Millions        = Revenue..Millions...,
```

```
    Players_Millions        = Players..Millions.,
    Peak_Concurrent_Players = Peak.Concurrent.Players,
    Metacritic_Score        = Metacritic.Score,
    Esports_Popularity      = Esports.Popularity,
    Trending_Status         = Trending.Status
  )

# Drop any possible rows with missing key values
gaming_clean <- gaming_raw %>%
  filter(
    !is.na(Revenue_Millions),
    !is.na(Genre),
    !is.na(Platform),
    !is.na(Release_Year)
  ) %>%
# Remove any possible invalid numeric values
 filter(
    Revenue_Millions > 0,
    Players_Millions >= 0,
    Peak_Concurrent_Players >= 0,
    Metacritic_Score >= 0,
    Metacritic_Score <= 100
  ) %>%
# Convert variables to appropriate types
 mutate(
    Genre                       = as.factor(Genre),
    Platform                    = as.factor(Platform),
    Esports_Popularity          = as.factor(Esports_Popularity),
    Trending_Status             = as.factor(Trending_Status),
    Release_Year                = as.integer(Release_Year)
  )

# Final check of the cleaned dataset
summary(gaming_clean)
```

```
##   Game_Title             Genre               Platform    Release_Year
##  Length:1000        Action  :122   Cross-Platform :168   Min.   :2000
##  Class :character   Sports  :116   Mobile         :158   1st Qu.:2006
##  Mode  :character   Strategy:116   Nintendo Switch:158   Median :2012
##                     Fighting:103   PC             :174   Mean   :2012
##                     Shooter :100   PlayStation    :175   3rd Qu.:2018
##                     Horror  : 96   Xbox           :167   Max.   :2024
##                     (Other) :347
##   Developer          Revenue_Millions  Players_Millions Peak_Concurrent_Players
##  Length:1000        Min.   :  11.43   Min.   :  0.53   Min.   : 0.11
##  Class :character   1st Qu.:1276.19   1st Qu.: 52.01   1st Qu.:12.97
##  Mode  :character   Median :2476.13   Median :107.04   Median :26.41
##                     Mean   :2483.02   Mean   :103.50   Mean   :31.60
##                     3rd Qu.:3677.80   3rd Qu.:155.63   3rd Qu.:46.02
##                     Max.   :4999.79   Max.   :199.98   Max.   :96.62
##
##  Metacritic_Score Esports_Popularity  Trending_Status
##  Min.   : 50.00   No :493            Declining:326
```

```
##  1st Qu.: 62.00   Yes:507           Rising   :335
##  Median : 76.00                     Stable   :339
##  Mean   : 74.99
##  3rd Qu.: 87.00
##  Max.   :100.00
##
```

4. Exploratory Data Analysis (EDA)

```
library(ggplot2)

summary(gaming_clean[, c("Revenue_Millions",
                         "Players_Millions",
                         "Peak_Concurrent_Players",
                         "Metacritic_Score",
                         "Release_Year")]
        )
```

```
##  Revenue_Millions  Players_Millions Peak_Concurrent_Players Metacritic_Score
##  Min.   :  11.43   Min.   :  0.53   Min.   : 0.11          Min.   : 50.00
##  1st Qu.:1276.19   1st Qu.: 52.01   1st Qu.:12.97          1st Qu.: 62.00
##  Median :2476.13   Median :107.04   Median :26.41          Median : 76.00
##  Mean   :2483.02   Mean   :103.50   Mean   :31.60          Mean   : 74.99
##  3rd Qu.:3677.80   3rd Qu.:155.63   3rd Qu.:46.02          3rd Qu.: 87.00
##  Max.   :4999.79   Max.   :199.98   Max.   :96.62          Max.   :100.00
##   Release_Year
##  Min.   :2000
##  1st Qu.:2006
##  Median :2012
##  Mean   :2012
##  3rd Qu.:2018
##  Max.   :2024
```

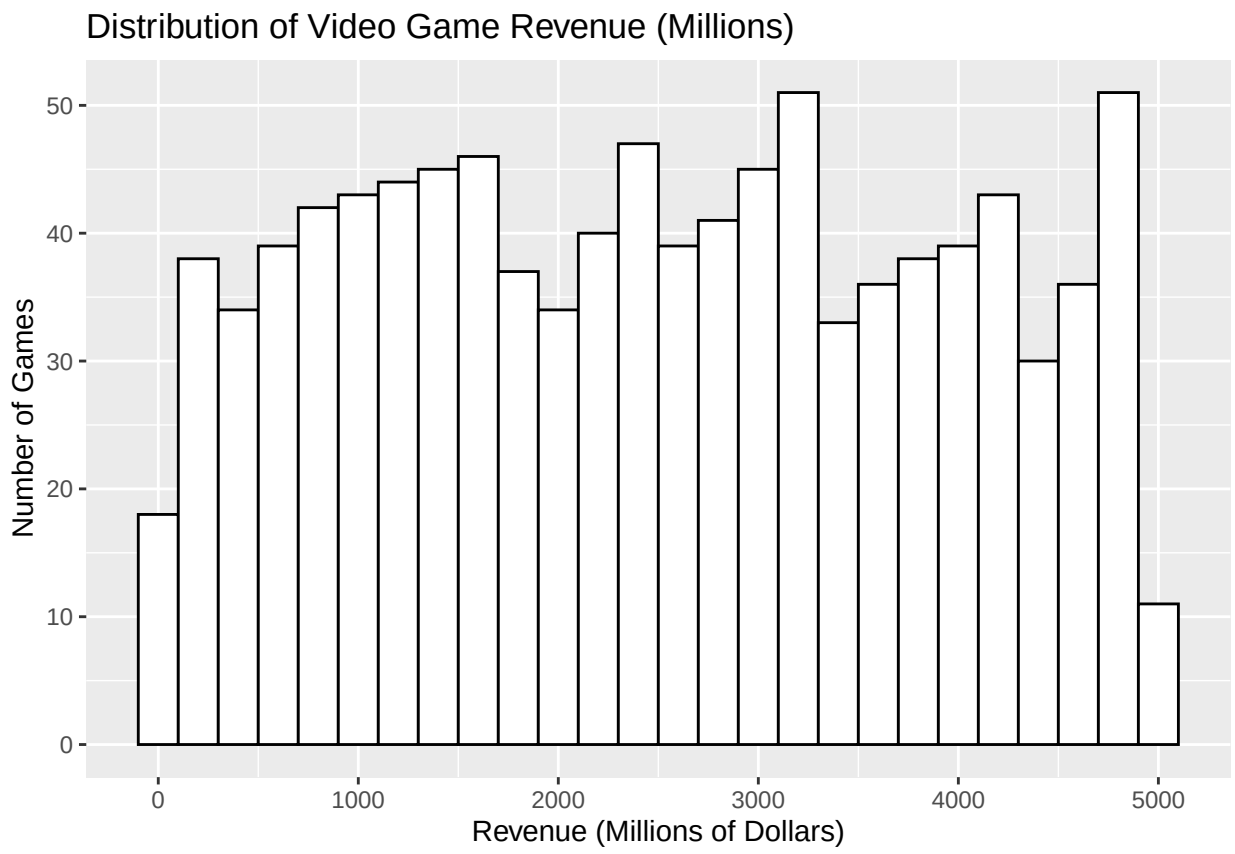
```
# Frequency tables for main categorical variables
table(gaming_clean$Platform)
```

```
##
##  Cross-Platform          Mobile Nintendo Switch              PC     PlayStation
##             168             158             158             174             175
##            Xbox
##             167
```

```
table(gaming_clean$Genre)
```

```
##
##     Action  Adventure   Fighting     Horror     Racing        RPG    Shooter
##        122         87        103         96         95         78        100
## Simulation     Sports   Strategy
##         87        116        116
```

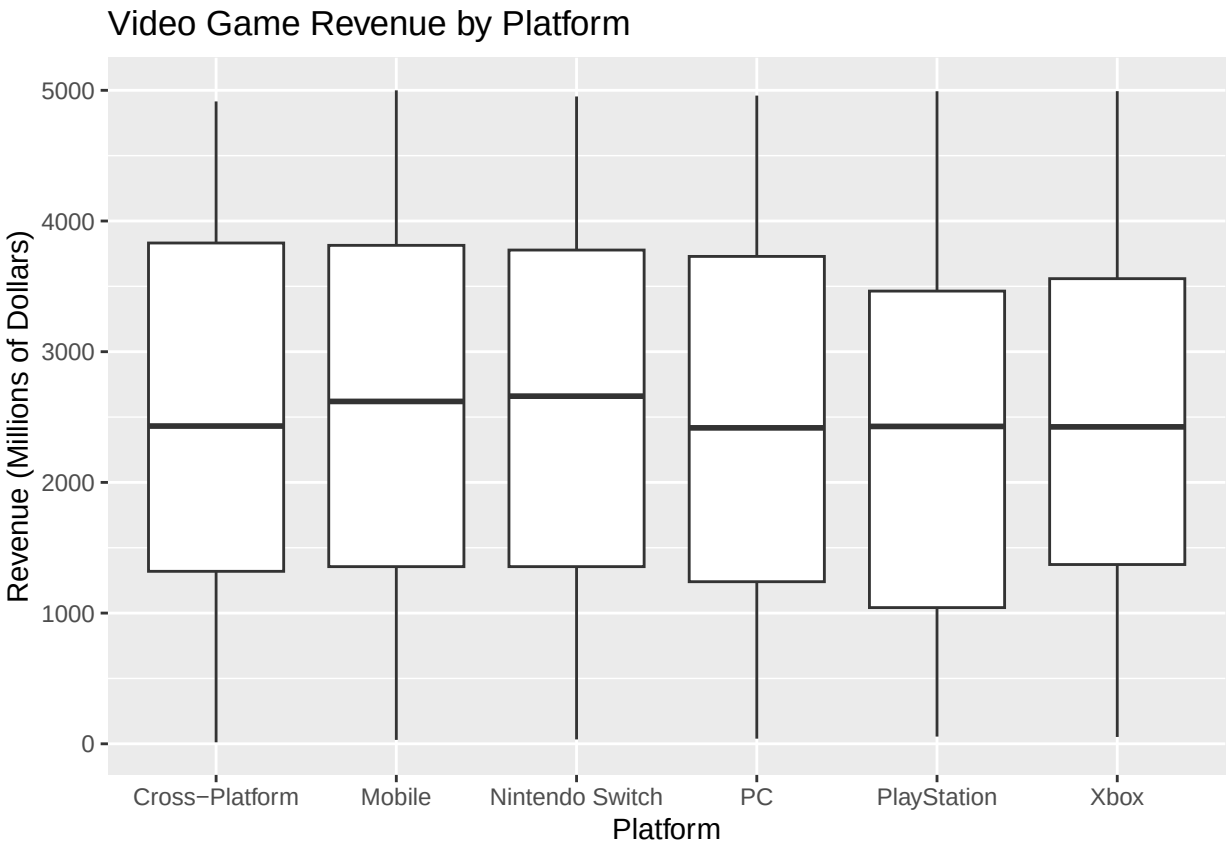
```r
# Histogram of game revenue (in millions)
ggplot(gaming_clean, aes(x = Revenue_Millions)) +
  geom_histogram(binwidth = 200, color = "black", fill = "white") +
  labs(
    title = "Distribution of Video Game Revenue (Millions)",
    x = "Revenue (Millions of Dollars)",
    y = "Number of Games"
  )
```



Histogram: Distribution of Video Game Revenue

The histogram presentation demonstrates how video game revenue distributes across the complete dataset. The majority of games generate revenue between 1,000 million dollars and 4,000 million dollars. The revenue of most games stays within this range but two games generate either very low or extremely high revenue. The data indicates that video games generate either substantial financial success or average revenue levels.

```r
# Boxplot of revenue by platform (in millions)
ggplot(gaming_clean, aes(x = Platform, y = Revenue_Millions)) +
  geom_boxplot() +
  labs(
    title = "Video Game Revenue by Platform",
    x = "Platform",
    y = "Revenue (Millions of Dollars)"
  )
```

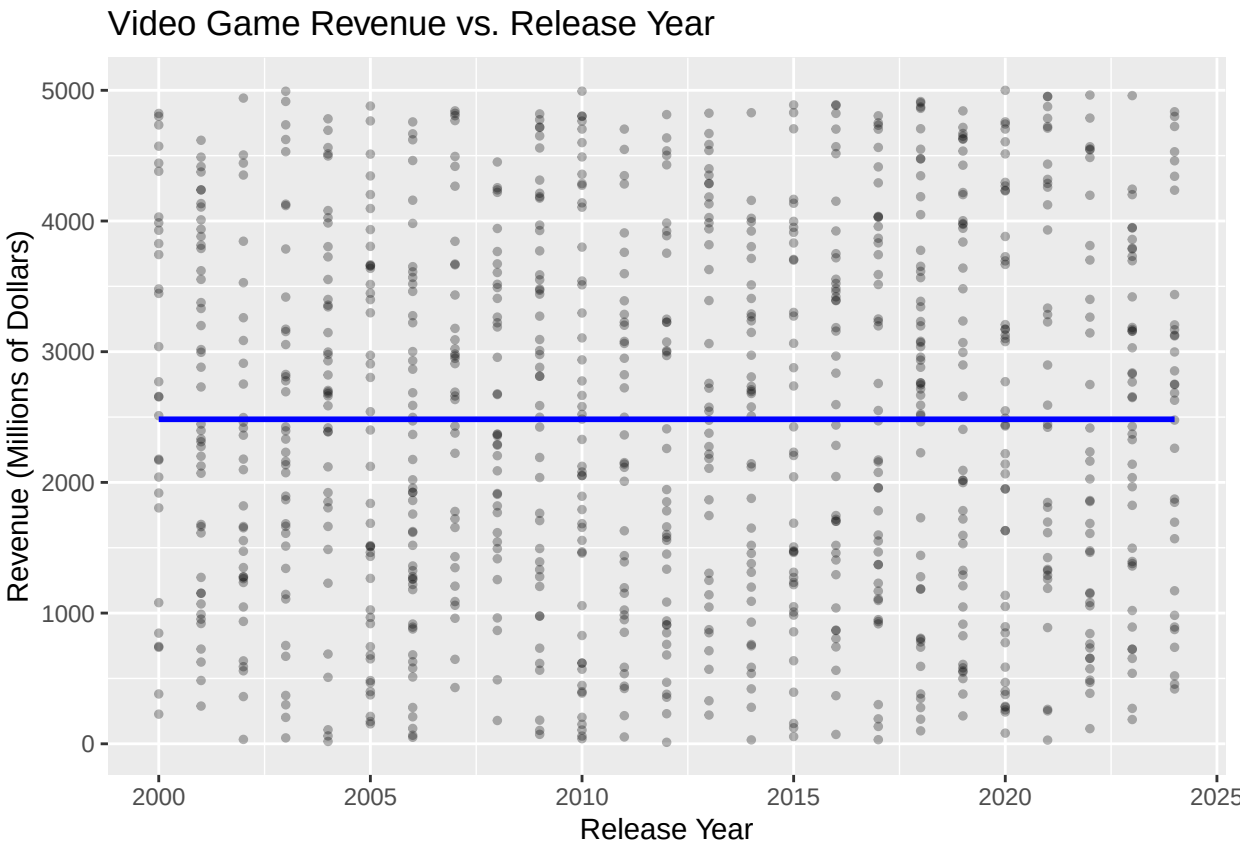


Boxplot: Video Game Revenue by Platform

The boxplot presentation enables users to evaluate revenue performance between different gaming platforms. The revenue distribution for each platform appears as a box in the graph. The median revenue value appears as a line inside each box while the box height indicates the extent of revenue variation. The platforms demonstrate different revenue levels and revenue distribution patterns. The analysis reveals which gaming platforms generate higher total revenue.

```
# Scatterplot of revenue vs. release year (in millions)
ggplot(gaming_clean, aes(x = Release_Year, y = Revenue_Millions)) +
  geom_point(alpha = 0.3, size = 1) +
  geom_smooth(se = FALSE, color = "blue") +
  labs(
    title = "Video Game Revenue vs. Release Year",
    x = "Release Year",
    y = "Revenue (Millions of Dollars)"
  )
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

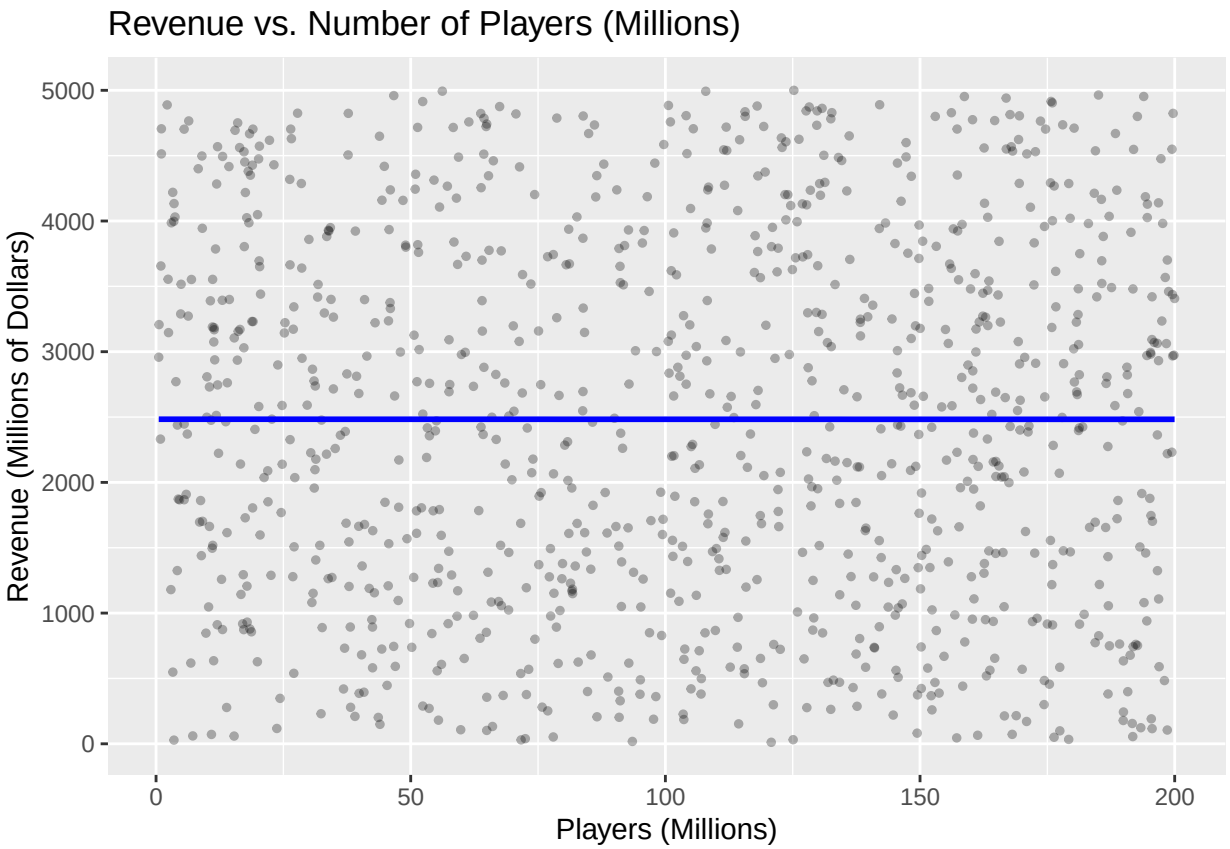


Scatterplot: Revenue vs. Release Year

The scatterplot displays the relationship between game release dates and their corresponding revenue values. The scatterplot contains game data points which follow a blue trend line. The blue trend line maintains a flat position which indicates no existing connection between release year and revenue. The release year of a game does not create a direct link to its revenue generation. The revenue values of games span across all release years without any specific pattern.

```
# Scatterplot of revenue vs. players (both in millions)
ggplot(gaming_clean, aes(x = Players_Millions, y = Revenue_Millions)) +
  geom_point(alpha = 0.3, size = 1) +
  geom_smooth(se = FALSE, color = "blue") +
  labs(
    title = "Revenue vs. Number of Players (Millions)",
    x = "Players (Millions)",
    y = "Revenue (Millions of Dollars)"
  )
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Scatterplot: Revenue vs. Number of Players

The scatterplot examines how revenue levels change when player numbers increase. The plot shows a wide distribution of points which contradicts the expected relationship between player numbers and revenue. The blue trend line maintains a flat position which indicates no existing relationship between player numbers and revenue. The data shows that games with large player bases can generate low revenue while games with smaller player bases can achieve higher revenue. The analysis indicates that player numbers do not effectively predict revenue levels in this particular dataset.

5. Visualization Quality and Storytelling

The visualizations selected for this analysis were chosen to match the structure of the data and the research question.

A histogram was used to examine the distribution of video game revenue, which is a continuous variable. This plot revealed that most games cluster between one and four billion dollars in revenue, with a few extreme outliers. The histogram was appropriate because it highlights skewness and spread, allowing us to assess whether mean or median values better represent the data.

A boxplot was employed to compare revenue across platforms, a categorical variable. Boxplots are well suited for this purpose because they display medians, variability, and outliers within each group. The visualization showed that platforms differ in both central tendency and spread. This could suggest that platform choice may influence revenue, which directly connects to our research question.

Two scatterplots were used to test relationships between continuous predictors and revenue. The first examined released year showed a flat trend line and no clear correlation, which can indicate that newer games do not necessarily earn more. The second scatterplot revealed that large player counts do not guarantee

high revenue. Scatterplots were appropriate here because they allowed us to test for linear or nonlinear associations between continuous variables.

All plots include clear axis labels with units (e.g. "Revenue (Millions of Dollars)"), descriptive titles and legends where necessary. The colors were kept simple, and transparency (alpha) was applied to scatterplots to reduce overplotting and improve readability. Bin widths in the histogram were chosen to balance detail with clarity, and categorical labels in the boxplot were spelled out fully to ensure accessibility and interpretability.

6. Modeling Approach

The problem was framed as a regression task since the target variable, video game revenue, is continuous and measured in millions of dollars. Our objective was not to classify games into categories but to estimate revenue levels based on game characteristics such as platform, genre, release year, players (millions), and critical scores.

We did not implement a heuristic or rule-based model. Instead, our analysis focused directly on regression methods. The choice of regression was appropriate because it allows us to quantify the relationship between multiple predictors and a continuous outcome, while providing interpretable coefficients that indicate the direction and strength of each association.

Our primary modeling approach was multiple linear regression. This model was selected because it can incorporate both categorical predictors (platform, genre, etc) and continuous predictors (players, release year, etc). Encoding categorical variables as factors enabled us to evaluate differences across groups, while continuous variables allowed us to test linear relationships. It was pretty efficient for a dataset that had the size of ours (1000 rows) and offers transparency in interpretation, making it well suited for identifying which factors contribute most strongly to video game revenue.

7. Model Implementation & Evaluation

8. Conclusions & Recommendations

9. Code Quality & Reproducibility

10. References

- CDS 101. Assignment 10: Car Prices. 2025, GMU.
- CDS 102. California Housing Lab. 2025, GMU.

Appendix (Optional)