CDS 101 – Final Project Report

Werner Wyss - Kanita Haque - Hussein Hamdan

December 13, 2025

Contents

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,
```

```r
    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)

```r
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
```

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

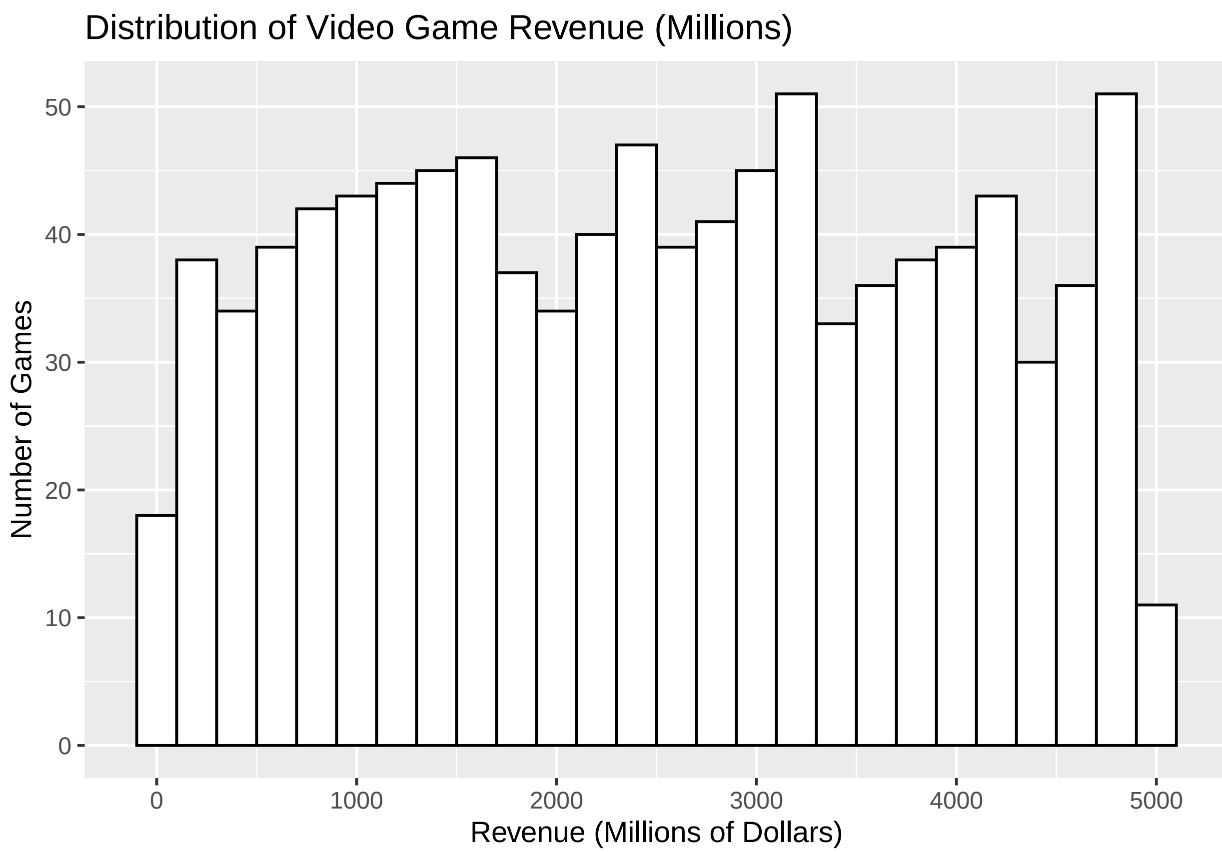
```
##
##  Cross-Platform          Mobile Nintendo Switch              PC     PlayStation
##             168             158             158             174             175
##            Xbox
##             167
```

```r
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") +
```

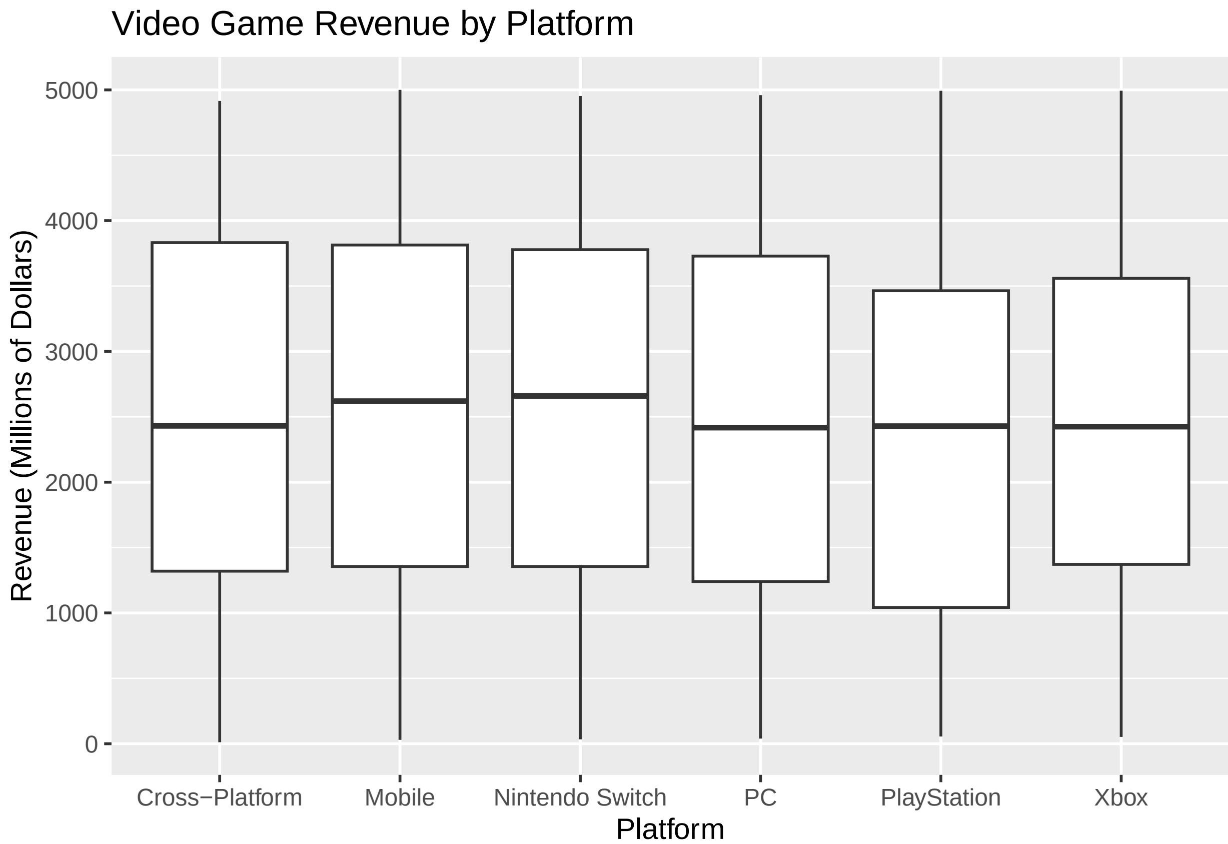
```r
  labs(
    title = "Distribution of Video Game Revenue (Millions)",
    x = "Revenue (Millions of Dollars)",
    y = "Number of Games"
  )
```



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.

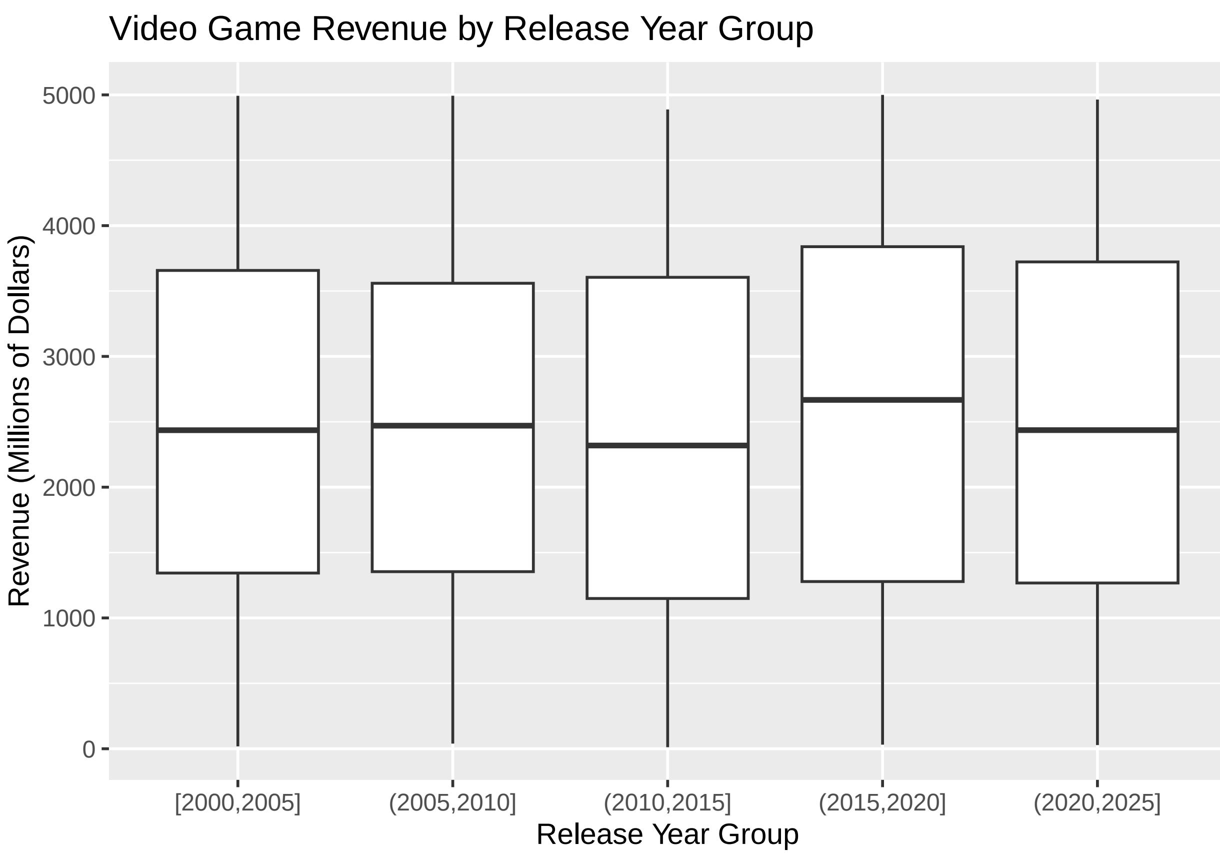
```r
# Boxplot of revenue by platform (in millions)
ggplot(gaming_clean, aes(x = Platform, y = Revenue_Millions)) +
  geom_boxplot() +
  labs(
    title = "Video Game Revenue by Platform",
    x = "Platform",
    y = "Revenue (Millions of Dollars)"
  )
```



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.

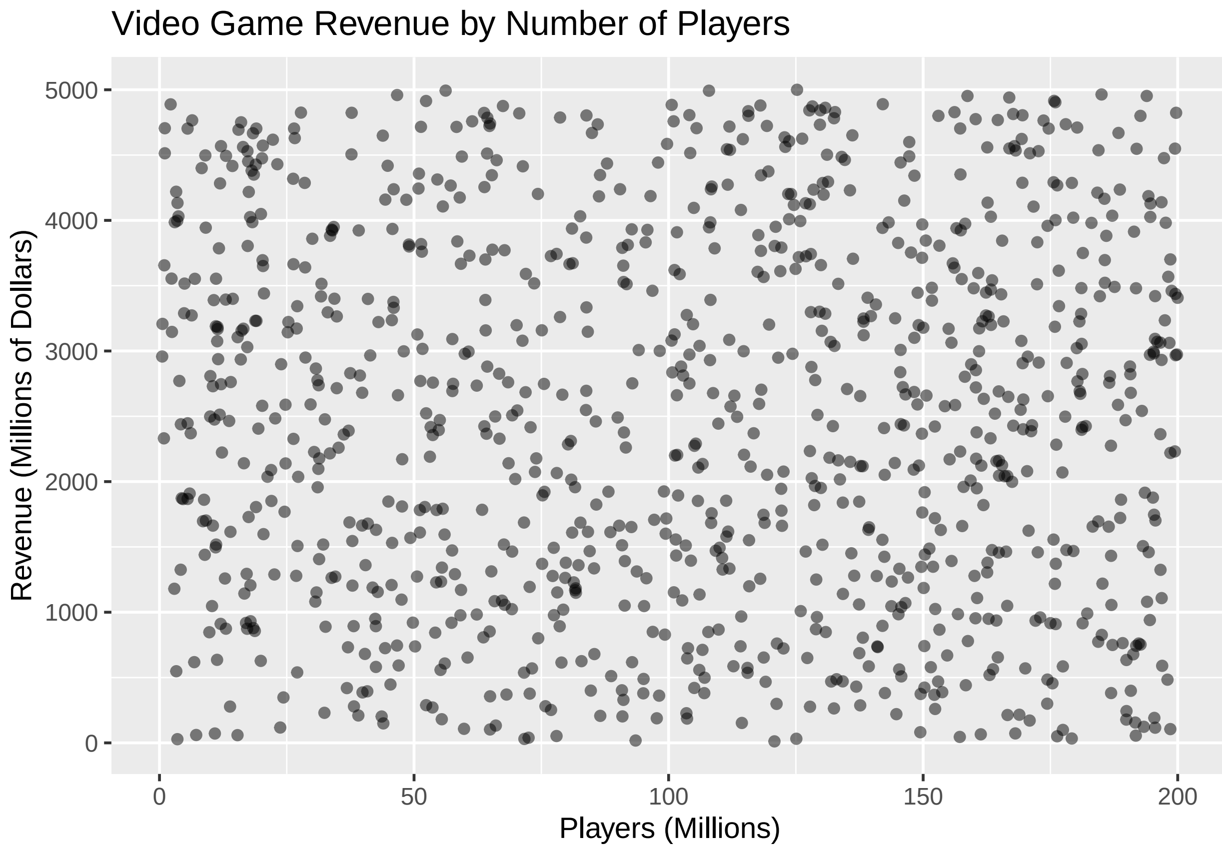
```r
# Boxplot of revenue vs. release year
gaming_clean$Year_Group <- cut(
  gaming_clean$Release_Year,
  breaks = seq(2000, 2025, by = 5),
  include.lowest = TRUE
)

ggplot(gaming_clean, aes(x = Year_Group, y = Revenue_Millions)) +
  geom_boxplot() +
  labs(
    title = "Video Game Revenue by Release Year Group",
    x = "Release Year Group",
    y = "Revenue (Millions of Dollars)"
  )
```



Boxplot: Revenue vs. Release Year

The boxplot presents the distribution of video game revenue across grouped release years. Each box reflects the spread of revenue values within a five year interval, highlighting the median and variability of earnings. The height and range of the boxes vary, indicating inconsistent revenue patterns over time. No clear progression or decline is observed across across the year groups, suggesting that the release period does not directly influence a game's financial success. Revenue values are widely scattered within each group, showing that high and low earnings occur in all time frames without a predictable trend.

```r
# Scatterplot of revenue by number of players (in millions)
ggplot(gaming_clean, aes(x = Players_Millions, y = Revenue_Millions)) +
  geom_point(alpha = 0.5) +
  labs(
    title = "Video Game Revenue by Number of Players",
    x = "Players (Millions)",
    y = "Revenue (Millions of Dollars)"
  )
```



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.

Boxplots were employed to compare revenue across categorical variables such as platform and release year group. Boxplots are well suited for this purpose because they display medians, variability, and outliers within each category. The visualization of platforms showed that distributions differ in both central tendency and spread, suggesting that platform choice may influence revenue. The release year boxplot revealed wide variability across all intervals, with no consistent upward or downward trend, indicating that timing alone does not guarantee financial success.

Scatterplots were used to test relationships between continuous predictors and revenue. The first examined number of players and revealed that large player counts do not guarantee high revenue, as the flat trend line suggested no clear correlation. 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

```r
set.seed(123)

n <- nrow(gaming_clean)
train_index <- sample(seq_len(n), size = 0.8 * n)

train_data <- gaming_clean[train_index, ]
test_data  <- gaming_clean[-train_index, ]

game_model <- lm(Revenue_Millions ~ Platform + Genre + Release_Year,
                 data = train_data)

pred_test <- predict(game_model, newdata = test_data)

rmse <- sqrt(mean((test_data$Revenue_Millions - pred_test)^2))

summary(game_model)
```

```
##
## Call:
## lm(formula = Revenue_Millions ~ Platform + Genre + Release_Year,
```

```
##     data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2630.66 -1198.09    18.89  1160.36  2767.56
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1109.0955 14489.5928   0.077    0.939
## PlatformMobile             10.4396   177.5123   0.059    0.953
## PlatformNintendo Switch    89.7919   178.3529   0.503    0.615
## PlatformPC               -139.0185   174.2223  -0.798    0.425
## PlatformPlayStation      -194.4017   174.6627  -1.113    0.266
## PlatformXbox              -51.6885   174.3628  -0.296    0.767
## GenreAdventure             93.8012   220.7362   0.425    0.671
## GenreFighting             317.1953   207.8718   1.526    0.127
## GenreHorror               116.9963   213.0990   0.549    0.583
## GenreRacing               111.4852   211.4191   0.527    0.598
## GenreRPG                  317.2698   237.1109   1.338    0.181
## GenreShooter               38.1569   213.4026   0.179    0.858
## GenreSimulation           189.6699   224.0644   0.846    0.398
## GenreSports               -44.4135   204.4334  -0.217    0.828
## GenreStrategy             107.6987   200.6921   0.537    0.592
## Release_Year                0.6542     7.2034   0.091    0.928
##
## Residual standard error: 1417 on 784 degrees of freedom
## Multiple R-squared:  0.01035,    Adjusted R-squared:  -0.008589
## F-statistic: 0.5464 on 15 and 784 DF,  p-value: 0.9146
```

```
rmse
```

```
## [1] 1420.391
```

8. Conclusions & Recommendations

Our analysis set out to answer the question: Which factors help predict the revenue of a video game? Using the Gaming Industry Trends dataset, we examined platform, genre, release year, player counts, and critical scores to see how these factors to the financial performances of games.

In this dataset, genre shows the clearest differences in revenue, while platform, release year, and player counts do not reliably predict financial success. Critical scores and esports popularity have some influence but are not decisive factors.

Key Findings:

- Platform: Revenue distributions across platforms were broadly similar.
- Genre: Action, Sports, and Shooter titles tended to cluster at higher revenue levels, while Simulation, Strategy, and Horror games were more often associated with lower averages.
- Number of Players did not guarantee high revenue. The trend line was flat.
- Release Year showed wide variability in revenue across all release year groups. No consistent upward or downward trend was observed, indicating that release timing alone did not guarantee high revenue.
- Critical Scores and Esports Popularity had weak to moderate correlations.

Limitations:

- The size of the dataset and the bias towards mainstream provide a narrow angle of the industry. It emphasizes mainstream titles and excludes many indie games, niche games, and emerging platforms.
- There are a lot of key variables missing such as marketing budgets, distribution channels, micro transaction sales, etc. Without these, our models can't capture some of the most important financial influences in the gaming industry.
- Linear regression assumes linear relationships, which may oversimplify complex industry dynamics.

Future Work:

- We could collect a larger and more diverse sample that includes indie titles, niche genres, and emerging platforms. This woud reduce sampling bias and improve generalizability.
- We could explore more interaction effects. Revenue often depends on a combination of factors. Genre alone may not matter, but Genre + Platform might. For instance, Shooters on PC may perform a lot more differently than Shooters on Mobile.

9. Code Quality & Reproducibility

The project code structure enables users to track all processing steps which produce final results. The cleaning section of the code performs two main operations which include renaming columns for clarity and deleting records containing missing or unfeasible data points including negative revenue values and player count values. The analysis results from the model and graphs depend on data that has been validated for accuracy.

The visualizations in this project directly reflect the programming operations that generated them. The Revenue_Millions histogram displays revenue data in 200 million dollar intervals which produces bars that demonstrate most games generate between 1 billion to 4 billion dollars while showing occasional high-value outliers. The boxplot analysis of revenue data by platform platform uses Platform and Revenue_Millions variables directly for its analysis. The scatterplots display trend lines generated by geom_smooth() which maintain a flat pattern throughout the graphs. The data directly produces this visual pattern which demonstrates that release year and player numbers do not effectively predict revenue.

The modeling section containing modeling code presents itself in a straightforward manner. The train-test split becomes reproducible through set.seed(123) and the linear model applies the essential variables which were examined in the graphical analysis. The RMSE calculation uses prediction results to generate an error value which users will encounter when executing the project. The R Markdown file execution sequence produces identical results for dataset cleaning and graph generation and model output. The project becomes both reliable and simple to reproduce because of its well-organized code structure.

10. References

- CDS 101. Assignment 10: Car Prices. 2025, GMU.
- CDS 102. California Housing Lab. 2025, GMU.
- https://www.kaggle.com/datasets/haseebindata/gaming-industry-trends-1000-rows