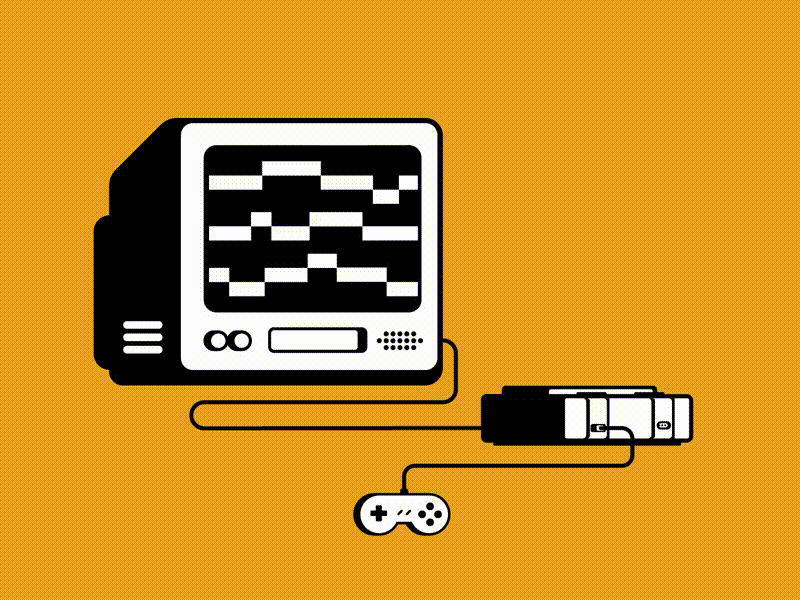
**Chart, waterfall chart

Description automatically generated**

****

**Executive Summary:**

**Problem Summary:**

This project aims to investigate popular game genres in Japan and their impact on game firms for clients targeting the Japanese market. The Japanese gaming industry is one of the largest and most important in the world, and understanding market trends and characteristics can help companies stay ahead of changes and maximize sales.

**Data:**

The dataset contains a diverse selection of games with varying sales performance, popularity, and longevity scraped from Japanese video game sales websites

**Analysis:**

Linear mixed-effects regression (LMER) models were used to analyze the relationship between weekly sales and various predictor variables, fitted using three LMER models in R with the Bobyqa optimization method. Models include the game title as a random effect, which provides the best fit for exploring popular game genres in Japan and their implications for game companies. Its insights can assist game companies in making informed decisions about game development and marketing in Japan.

**Key Findings:**

Based on our analysis, we found that games with genres Adventure (ADV), Role-Playing (RPG), and Miscellaneous (ETC) have positive estimates, while genres such as Education, Lifestyle (HOB), Racing (RCE), Strategy (SLG), and Sports (SPT) have negative estimates. Additionally, being in the top 10 has a statistically significant effect on weekly sales.

**Problem Definition & Significance:**

**Who is the target client for this project, and what business problem are you trying to address for this client? Why is this an interesting or important problem? Present some brief industry statistics or background research to make a strong case for your target problem.**

This project's target customer is a game development business or publisher aiming to increase their presence in the Japanese market.

The purpose of this project is to investigate popular game genres in Japan and their ramifications for game firms. We examine game survival times and sales over time, considering numerous aspects such as game characteristics, genre, publisher, time of release, and market circumstances.

For various reasons, this is an intriguing and significant problem:

* With a market value of more than $21 billion, Japan's gaming industry is one of the largest and most important in the world.
* Companies may improve their chances of success by studying popular genres and market trends in the gaming business.
* Long-term success requires staying ahead of market changes, and analyzing the characteristics that contribute to a game's sales and durability may help organizations alter their approach.
* Understanding the nuances of the Japanese gaming market may also help businesses tailor their marketing strategies and promotional efforts to better connect with their target audience and maximize sales.

**Prior Literature:**

**How have others tried to address this problem and with what outcomes?**

* Famitsu, a well-known Japanese video game magazine, has utilized sales data and trends to determine popular game genres (Famitsu).
* Exploring survival rates of companies in the UK video-games industry: An empirical study
* Estimating Video Game Success using Machine Learning

**Data Source/Preparation:**

**Where did you source your data?**

Data has been scraped from the internet using a tool called “selector Gadget” in conjunction with an R package (“rvest”) from the following websites:

* Perfectly-nintendo
* Resetera

**What variables did it contain and how were they measured?**

*current\_ranking:* The current sales ranking of the game within the specified week.

*prev\_ranking:* The sales ranking of the game in the previous week.

*platform:* The gaming platform on which the game is available (e.g., WIU, 3DS).

*game\_title:* The title of the game.

*top\_10:* A binary indicator (0 or 1) denoting whether the game is in the top 10 sales rankings for the week.

*genre:* The genre of the game (e.g., RCE for Racing, FTG for Fighting).

*publisher:* The company that published the game (e.g., Nintendo).

*release\_date:* The date when the game was released.

*price:* The price of the game in the given currency.

*weekly\_sales:* The number of units sold during the specified week.

*total\_sales:* The cumulative number of units sold since the game's release.

*market\_share:* The game's share of the total market sales for the specified week, expressed as a percentage.

*ltd\_sales:* Lifetime-to-date sales (i.e., total sales up to the end of the specified week).

*sales\_ratio:* The ratio of weekly sales to total sales.

*sales\_yoy:* The year-over-year sales growth for the game, expressed as a percentage.

*year:* The year in which the specified week falls.

*week\_no:* The week number within the year.

*start\_dt:* The start date of the specified week.

*end\_dt:* The end date of the specified week.

*weekly\_count:* A counter variable indicating the number of times the game has appeared in the dataset (e.g., 1 for the first week, 2 for the second week).

*no\_of\_weeks:* The total number of weeks since the game's release.

**Which variables (DV, IV) did you select for your analysis and why?**

I selected the following variables for the analysis:

**Dependent Variable (DV):**

*weekly\_sales*: This variable represents the weekly sales of video games, which is the main focus of our analysis. We want to understand the factors that influence the weekly sales of video games.

**Independent Variables (IV):**

*total\_sales:* The total sales of a video game can be a good indicator of its popularity, which may affect its weekly sales.

*market\_share:* The market share of a game can help us understand how it performs compared to other games in the market, which may influence weekly sales.

*price:* The price of a video game can affect consumers' purchasing decisions and, consequently, its weekly sales.

*week\_no:* The week number can help us understand any seasonality or trends in weekly sales.

*top\_10:* This binary variable indicates whether a game is in the top 10 sales, which can provide insights into the popularity of a game and its potential impact on weekly sales.

*genre:* The genre of a video game can influence the preferences of consumers and, therefore, its weekly sales.

*publisher:* The publisher of a video game can influence its marketing and promotion, which may affect its weekly sales.

**How did you clean the data?**

* Checked for and dealt with missing values.
* Converted columns to appropriate data types (e.g., factors, numeric).
* Removed non-numeric values from the 'price' and 'weekly\_sales' columns and replaced them with NAs or appropriate values.
* Imputed missing values in the 'price' column with the median price.
* Removed outliers in the 'price' column by calculating the threshold for outliers and removing rows with price values above the threshold.

(The outliers removed are not the actual games but the bundles)

* Filtered out missing and non-finite values in the 'price' column.
* Rescaled numeric predictor variables (total\_sales, market\_share) using the scale function.
* Created interaction terms and random effects for linear mixed models based on game\_title, genre, and publisher.

**Variable Choice:**

**What are your predictors of interest? Explain the rationale for these predictors.**

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Expected Sign** | **Rationale** |
| Total Sales | Positive | Higher total sales indicate a game's popularity, which could lead to higher weekly sales. |
| Market Share | Positive | A larger market share suggests that the game is well-received, resulting in higher sales. |
| Price | Negative | Higher-priced games may deter some customers, potentially leading to lower weekly sales. |
| Week number | Negative | Sales generally decline over time, so later weeks may see lower sales. |
| Publisher | Mixed | Popular Publisher games tend to sell more |
| Top 10 | Positive | Being in the Top 10 list signals that the game is popular, which may boost weekly sales. |
| Genre | Mixed | Different genres may appeal to different audiences, leading to varying effects on sales. |

**Descriptive Analysis & Data Visualizations:**

**What patterns/trends do you see in your data? What do you infer from these trends?**

we can observe some patterns/trends in the data:

Ranking:

The current ranking ranges from 1 to 30, with a mean of 15.44.

The previous ranking ranges from 0 to 54, with a mean of 10.4.

The presence of 0 in the previous ranking indicates that some games were newly introduced in the ranking list.

Top 10:

The mean value of 0.6627 suggests that, on average, about 66% of the games in the dataset have been in the top 10 at some point.

Genre, Platform, and Publisher:

There is a variety of genres, platforms, and publishers, indicating a diverse selection of games in the dataset.

Release date:

The release dates range from 2010-03-04 to 2023-03-17, with a mean around 2018-07-17.

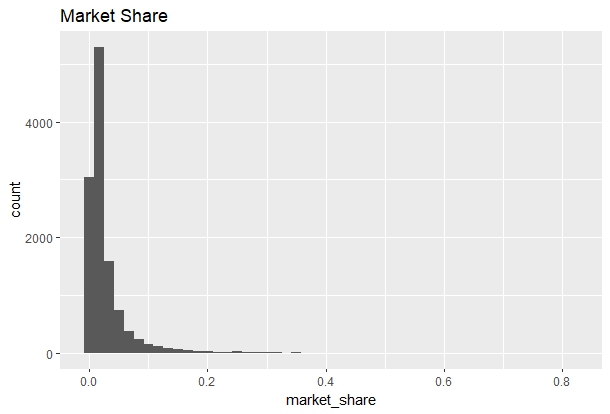
This suggests that the data includes games released over a 13-year span.

Chart, histogram

Description automatically generated

Total sales:

The total sales range from 92.2 to 1498.2, with a mean of 391, indicating a wide variation in game popularity and sales performance.



Market share:

Market share ranges from 0 to 0.8181, with a mean of 0.0336, showing that some games have a significantly larger share of the market than others.

Chart, histogram

Description automatically generated

Price:

Game prices range from 1.111 to 72.709, with a median price of 5.980 and a mean price of 6.031. The price distribution is very balanced, with 75% of the games costing below 6.980.

Year and Week number:

The data spans from 2015 to 2023, covering various years and week numbers (1 to 53).

Number of weeks:

The number of weeks for which a game has been in the dataset ranges from 1 to 305, with a mean of 62.89, indicating varied longevity for different games in the dataset.

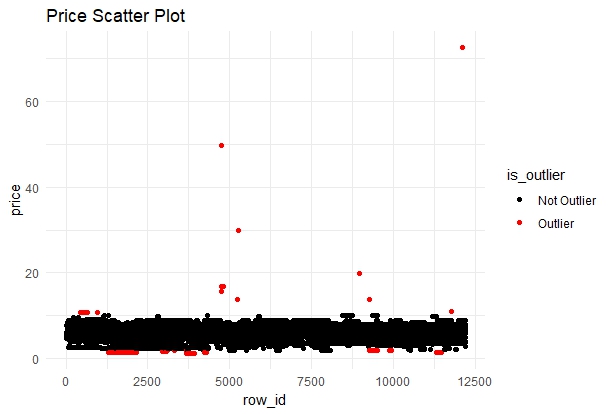
Based on these trends, we can infer the following:

***There is a diverse selection of games in the dataset, covering a wide range of genres, platforms, publishers, and release dates.***

***Sales performance varies significantly among games, with some having much higher total sales and market shares than others.***

***The dataset includes games with different levels of popularity and longevity, as indicated by their rankings, total sales, and the number of weeks they have been in the dataset.***

**Outliers:**



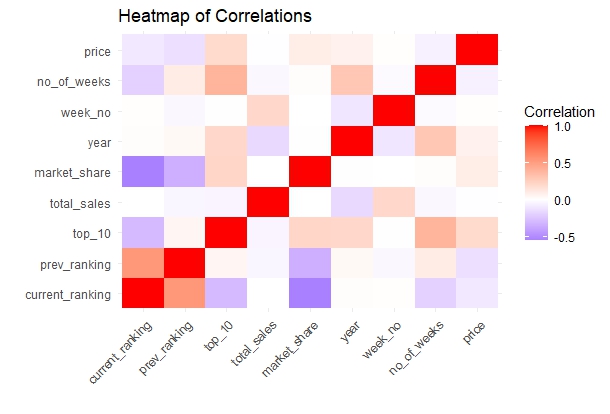
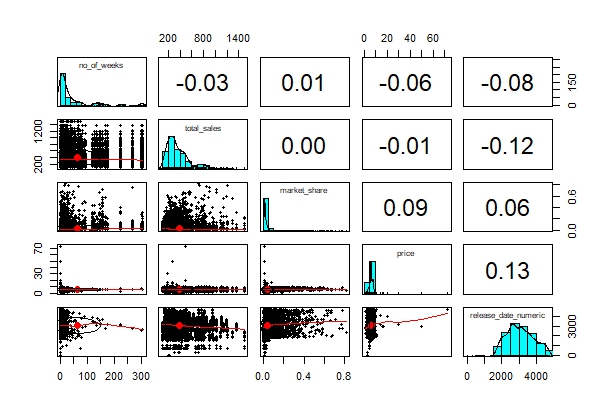
The summary statistics show that the price variable contains outliers from the ggplot.

The scatter plot depicts the price variable distribution by classifying it as Outlier or Not Outlier to provide a clear view of how many outliers are there in this data. Outlier rows have been removed as a result, ensuring that the analysis is not biased.

**`**

**Assumptions:**

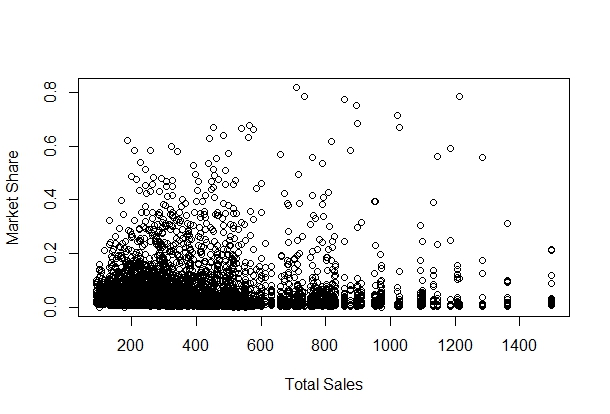
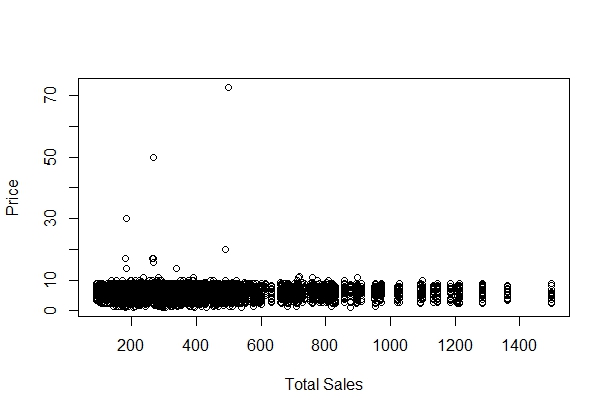
**Linearity between predictors:**

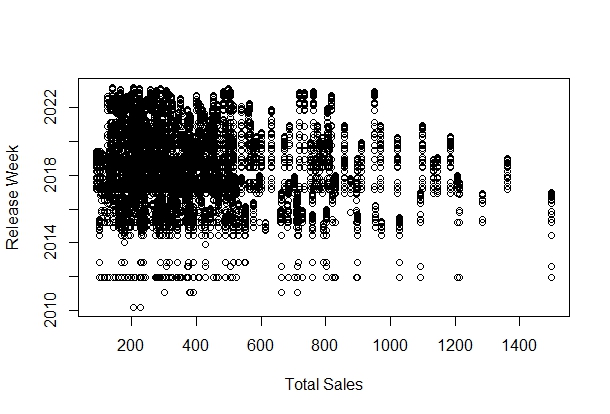
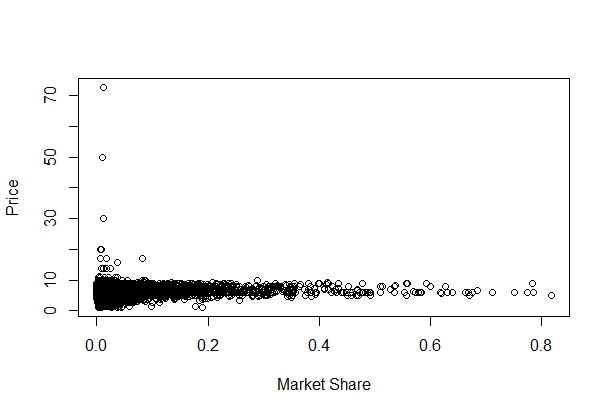
****

Some key observations include:

* *A positive trend between the number of weeks on the market and total sales, as games available for a longer time typically accumulate more sales.*
* *No clear relationship between the number of weeks on the market and market share or price, suggesting that other factors influence these variables.*
* *A positive relationship between total sales and market share, as games with higher sales are likely to have a larger market share.*
* *No clear linear relationship between total sales, market share, or price, as various factors can influence these variables.*

**Multicollinearity:**

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

*In conclusion, the most obvious example of multicollinearity is the relationship between total sales and market share. The remaining variable pairings do not have strong linear correlations, indicating that multicollinearity is not a major problem for these pairs.*

*As the project plan to build an LMER model, these are resistant to violations of the independence and normality assumptions, they can be sensitive to violations of the homoscedasticity and linearity assumptions. All regression models, including LMER models, must make the assumption of no multicollinearity.*

**Models:**

**How did you statistically model your data?**

In this analysis, the dependent variable, weekly sales, was initially modeled using a linear regression model. However, due to the presence of outliers and violations of assumptions such as normality, homoscedasticity, multi-collinearity, auto-correlation, and independence, an LMER model was employed instead. *(ref: appendix)*.

This approach proved effective in identifying popular games specific to each genre. The analysis considered fixed and random effects of publishers and game titles, and two types of LMER models were built. Data transformation techniques, including the scale off function in R and log transformation of price and weekly\_sales variables, were applied to improve model fit and compare the performance of different models in addressing the problem statement.

The data were fitted using three LMER models: one based on the game title, one based on the genre, and one based on the publisher. Models were fitted in R using the "lmer" function, with the optimization method set to "bobyqa" and the maximum number of function evaluations set at 1e5.

The Bobyqa optimization technique was utilized to fit statistical models. This algorithm is a mathematical tool that assists in determining the optimum answer to a problem, in this case, fitting the models. The technique works by starting with an initial guess of the answer and then incrementally improving it until the optimal solution is obtained.

**Present at most 3-4 models, but you must carefully choose your models.**

After investigating all eight models in the project, the following three were selected for their ability to better understand the relationship between game sales and predictors such as genre, publisher, time of release, and market conditions:

Original Model with the genre as a random effect (fit\_genre): Provides insights into the influence of genre on weekly sales without variable scaling.

Model with scaled data and genre as a random effect (fit\_genre\_sc): Offers a clearer understanding of relationships by controlling for the potential influence of different units and scales.

Model with log-transformed data (price & weekly\_sales): Accounts for variations and relationships among the factors, with log transformations applied to price and weekly\_sales variables.

These models were chosen based on their capacity to account for variations and relationships among the factors being analyzed, enabling more informed decision-making.

**Which model is best and why?**

**A picture containing chart

Description automatically generated Text

Description automatically generated** **Text

Description automatically generated**

Upon evaluating the three models under consideration, we compared their goodness-of-fit statistics, such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), to determine the most suitable model for our analysis. It is worth noting that lower AIC and BIC values indicate a superior model fit:

Original Model with genre as a random effect (fit\_genre): AIC: 105019.300, BIC: 105160.000

Model with scaled data and genre as a random effect (fit\_genre\_sc): AIC: 105054, BIC: 105096

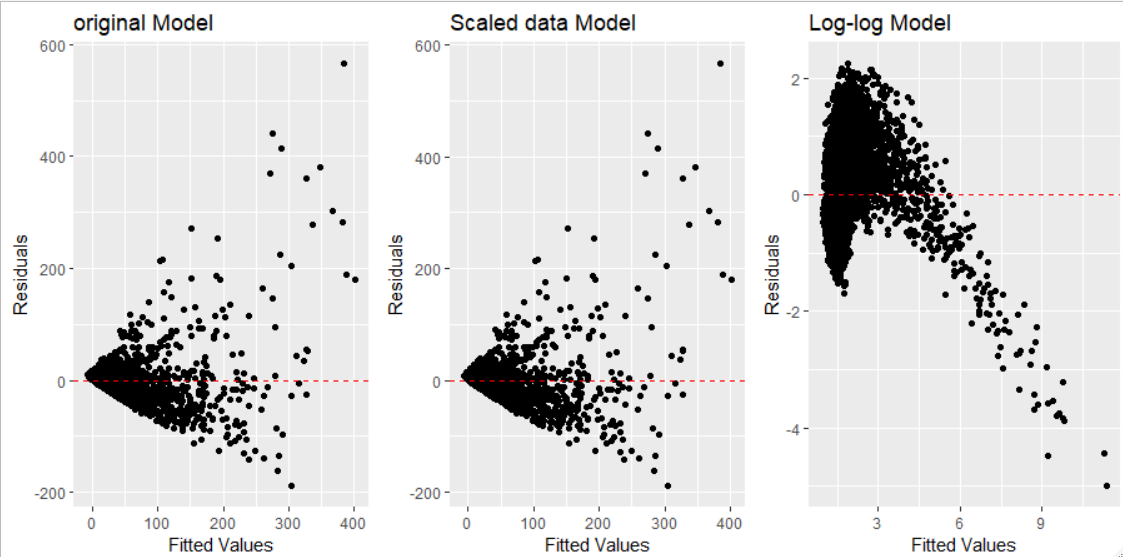
**Model with log-transformed data and genre as a random effect (fit\_genre\_log): AIC: 25637.340, BIC: 25689.170**

Based on these statistics, the fit\_genre\_log model, which incorporates log-transformed data for price and weekly\_sales, demonstrates the best fit among the three alternatives. This model provides a comprehensive understanding of the relationships between weekly sales and various factors, taking into account the nuances of different genres.

Consequently, we recommend utilizing the fit\_genre\_log model to explore the popularity of game genres in Japan and their implications for game companies. The insights gleaned from this model can assist game companies in making well-informed decisions regarding game development and marketing strategies, ensuring that they effectively cater to the gaming community.

**Quality Checks:**

**How do you know that your analysis is trustworthy? Test of assumptions? Robustness checks**

****

From the Fitted Values vs Residuals for the original and the Scaled data Model we can infer that the spread of the residuals are increasing as the fitted value increases. That is spread is not constant (Heteroskedasticity).Also, for the original and the Scaled data model there are possible outliers.

 From Log-Log model for the residual vs Fitted plot we can say that when the fitted values are low the residuals are negative, positive when the fitted value is in the middle and negative when the fitted value is large. Since the sign of the residual’s changes based on the magnitude of the fitted values, this may indicate that the model is mis specified or that there is some other issue with the mode

**Text, letter

Description automatically generated**  **Text, letter

Description automatically generated**

**Text

Description automatically generated**

**In this analysis, we have employed the Kolmogorov-Smirnov (KS) test to assess the normality of the residuals for three different models: fit\_genre, fit\_genre\_sc, and fit\_genre\_log. The test results indicate that the residuals for all three models do not adhere to a normal distribution, as evidenced by the low p-values obtained in each case.**

**Recommendations:**

**What recommendations do you have for your client, based on your analysis?**

We can identify distinct genres with larger intercepts from the random effects output of the three models (fit\_genre\_sc , fit\_genre, fit\_genre\_log). This may suggest enhanced sales potential. Here are a few genres that stand out:

* RPG (Role-Playing Games): In both the scaled and log-transformed models, RPGs show a positive random impact, indicating that this genre does well in terms of weekly sales.
* ETC (Other/Undefined Genres): In the scaled model, the ETC genre has a positive random impact, indicating that games in this category may have some sales potential. However, because the ETC category is not well-defined, it is critical to further investigate the characteristics of these games.

Based on these insights:

* Develop and sell RPG games: To leverage on existing market demand, focus on generating creative and entertaining RPG games. RPG titles should be promoted via focused marketing initiatives such as trailers, influencer collaborations, and in-game events.
* Investigate and capitalize on popular ETC games: Examine the features of successful ETC games to uncover trends and components that appeal with customers. Incorporate these insights into future game development and create targeted marketing campaigns to promote similar games.
* The customer may optimize their game development and marketing efforts by focusing on these specific genres and employing customised techniques, resulting in greater sales and market share.

**Appendix:**

**Other Research Investigations:**

A picture containing table

Description automatically generatedPost doing the analysis and building the model related to the problem statement, we have researched other models as mentioned below:

**Original Model with the game title as a random effect (fit\_title)**

**Original Model with the publisher as a random effect (fit\_pub)**

**Original Model without market share as fixed effect (fit\_wms)**

**Original Model with genre and publisher as random effect (fit\_gp)**

**Model with scaled data with publisher as random effect (fit\_pub\_sc)**

**Recommendations from these models:**

Our analysis suggests that ranking among the top 10 games sold is more crucial for revenue growth than sales volume growth.

Only a select few game genres significantly influence weekly sales. The "genreHOB" category, for instance, has a positive correlation with revenue, whereas the "genreRCE," "genreSPT," and "genreSTG" categories have a negative correlation with sales volume.

To improve sales, we recommend picking game genres that have a positive relationship with sales and concentrating on ranking among the top 10.

**R Code:**

library(readr)  
library(dplyr)  
library(lme4)  
library(ggplot2)  
library(readxl)  
library(ggcorrplot)  
library(mice)  
library(mitools)  
library(psych)  
library(ggplot2)  
library(car)  
library(gridExtra)  
library(nortest)  
  
#Importing data  
data <- read\_excel("latest\_games\_update\_2.xlsx")  
  
str(data)  
summary(data)  
  
# Check for missing values  
sum(is.na(data))  
  
# Convert columns to the appropriate data types  
data$platform <- as.factor(data$platform)  
data$genre <- as.factor(data$genre)  
data$publisher <- as.factor(data$publisher)  
  
# Remove yen symbol from price column  
data$price <- gsub("¥", "", data$price)  
  
# Find non-numeric values in the 'price' column  
non\_numeric\_price <- which(!is.na(as.numeric(data$price, warn = FALSE)) == FALSE)  
data$price[non\_numeric\_price]  
  
# Replace non-numeric values in the 'price' column with NAs or appropriate values  
data$price[non\_numeric\_price] <- NA # Replace with NA  
  
# Check unique values in price column  
unique(data$price)  
  
# Check if there are any missing values in price column  
any(is.na(data$price))  
summary(data$price)  
table(is.na(data$price))  
  
# Impute missing price values with the median  
median\_price <- median(data$price, na.rm = TRUE)  
data$price[is.na(data$price)] <- median\_price  
  
# Verify that there are no more missing values in price column  
sum(is.na(data$price))  
  
# Find non-numeric values in the 'weekly\_sales' column  
non\_numeric\_weekly\_sales <- which(!is.na(as.numeric(data$weekly\_sales, warn = FALSE)) == FALSE)  
data$weekly\_sales[non\_numeric\_weekly\_sales]  
  
# Replace non-numeric values in the 'weekly\_sales' column with NAs or appropriate values  
data$weekly\_sales[non\_numeric\_weekly\_sales] <- NA # Replace with NA  
  
# Convert columns to the appropriate data types  
data$platform <- as.factor(data$platform)  
data$genre <- as.factor(data$genre)  
data$publisher <- as.factor(data$publisher)  
data$price <- as.numeric(data$price)  
data$weekly\_sales <- as.numeric(data$weekly\_sales)  
  
# Create plots to visualize the distribution of the variables  
ggplot(data, aes(x = total\_sales)) + geom\_histogram(bins = 50) + ggtitle("Total Sales")  
ggplot(data, aes(x = market\_share)) + geom\_histogram(bins = 50) + ggtitle("Market Share")  
ggplot(data, aes(x = price)) + geom\_histogram(bins = 50) + ggtitle("Price")+scale\_x\_continuous(limits = c(0, 13))  
  
# Calculate mean and standard deviation  
mean\_price <- mean(data$price, na.rm = TRUE)  
sd\_price <- sd(data$price, na.rm = TRUE)  
  
# Define the threshold for outliers (mean + 2 \* standard deviation)  
threshold <- mean\_price + 2 \* sd\_price  
  
# Add row number to the data  
data <- data %>%  
 mutate(row\_id = row\_number())  
  
# Create the scatter plot  
ggplot(data, aes(x = row\_id, y = price, color = is\_outlier)) +  
 geom\_point() +  
 ggtitle("Price Scatter Plot") +  
 scale\_color\_manual(values = c("Not Outlier" = "black", "Outlier" = "red")) +  
 theme\_minimal()  
  
# Find the exact outlier rows  
outlier\_rows <- which(data$is\_outlier == "Outlier")  
outlier\_rows  
  
df\_new <- data[-c(230,379,434,468,559,616,650,963,1143,1173,1292,2788,2789,2817,4323,4735,4742,4743,4766,4801,5248,5249,5251,5880,5881,5910,8453,8485,8510,8543,8570,8604,8634,8701,8725,8757,8970,8971,9266,9267,9358,9387,9473,9506,11187,11188,11209,11210,11246,11247,11278,11279,11309,11789,12121), ]  
  
df\_new <- data[-c(3421), ]  
  
data<- df\_new  
  
# Filter out missing and non-finite values  
data\_filtered <- data %>%  
 filter(!is.na(price) & is.finite(price))  
  
# Create the plot without limits  
ggplot(data\_filtered, aes(x = price)) +  
 geom\_histogram(bins = 50) +  
 ggtitle("Price") + scale\_x\_continuous(limits = c(0,11))  
geom\_vline(aes(xintercept = threshold), color = "red", linetype = "dashed", size = 1) +  
 annotate("text", x = threshold, y = Inf, label = "Outliers Threshold", vjust = 2, hjust = 0, color = "red")  
  
# Create scatter plots to check linearity between predictors  
data$release\_date\_numeric <- as.numeric(data$release\_date - min(data$release\_date, na.rm = TRUE), units = "days")  
pairs.panels(data[, c("no\_of\_weeks", "total\_sales", "market\_share", "price", "release\_date\_numeric")])  
  
# Create scatter plots to check homoscedasticity between predictors  
plot(data$total\_sales, data$market\_share, xlab = "Total Sales", ylab = "Market Share")  
plot(data$total\_sales, data$price, xlab = "Total Sales", ylab = "Price")  
plot(data$total\_sales, data$release\_date, xlab = "Total Sales", ylab = "Release Week")  
plot(data$market\_share, data$price, xlab = "Market Share", ylab = "Price")  
plot(data$market\_share, data$release\_date, xlab = "Market Share", ylab = "Release Week")  
plot(data$price, data$release\_date, xlab = "Price", ylab = "Release Week")  
  
# Select relevant features  
data\_1 <- select(data, weekly\_sales,game\_title, genre, no\_of\_weeks, total\_sales, market\_share, price, year, week\_no, top\_10)  
#including publisher column  
data\_2 <- select(data, weekly\_sales,publisher,game\_title, genre, no\_of\_weeks, total\_sales, market\_share, price, year, week\_no, top\_10)  
  
# Rescale numeric predictor variables  
data\_scaled <- data\_1  
data\_scaled$total\_sales <- scale(data$total\_sales)  
data\_scaled$market\_share <- scale(data$market\_share)  
  
#scaling for data including publisher  
data\_scaled\_pub <- data\_2  
data\_scaled\_pub$total\_sales <- scale(data$total\_sales)  
data\_scaled\_pub$market\_share <- scale(data$market\_share)  
  
#Linear model  
lm\_fit <- lm(weekly\_sales ~ total\_sales + market\_share + price + week\_no + top\_10, data = data\_1)  
summary(lm\_fit)  
vif(lm\_fit)  
  
# Check model assumptions  
plot(lm\_fit, which = 1) # Residuals vs Fitted  
plot(lm\_fit, which = 2) # Normal Q-Q plot  
plot(lm\_fit, which = 3) # Scale-Location  
plot(lm\_fit, which = 4) # Cook's distance  
  
#based on genre  
fit\_genre <- lmer(weekly\_sales ~ market\_share + price + week\_no + top\_10 + genre + (1| genre),  
 data = data\_1, control = lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e5)))  
vif(fit\_genre) #for genre the data needs to be scaled to address convergence issues with the model  
summary(fit\_genre)  
  
#based on scaled data  
fit\_genre\_sc <- lmer(weekly\_sales ~ market\_share + price + week\_no + top\_10 + (1| genre),  
 data = data\_scaled, control = lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e5)))  
  
#Log-log model  
  
fit\_genre\_log <- lmer(log(weekly\_sales) ~ market\_share + log(price) + week\_no + top\_10 + (1| genre),  
 data = data\_1, control = lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e5)))  
  
# Random effects for fit\_genre\_sc  
ranef(fit\_genre\_sc)  
  
# Random effects for fit\_genre\_log  
ranef(fit\_genre\_log)  
  
---------------------------------------Research investigations-----------------------------------------------------------  
   
#based on game title  
fit\_title <- lmer(weekly\_sales ~ market\_share + price + week\_no + top\_10 + genre + (1| game\_title),  
 data = data\_1, control = lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e5)))  
summary(fit\_title)  
  
#based on publisher  
fit\_pub <- lmer(weekly\_sales ~ market\_share + price + week\_no + top\_10 + genre + (1| publisher),  
 data = data\_2, control = lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e5)))  
summary(fit\_pub)  
  
#game and publisher  
fit\_gp <- lmer(weekly\_sales ~ market\_share + price + week\_no + top\_10 + (1| publisher) + (1| genre),  
 data = data\_2, control = lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e5)))  
summary(fit\_gp)  
  
#without market share  
fit\_wms <- lmer(weekly\_sales ~ price + (week\_no \* week\_no) + top\_10 + genre + (1| publisher) + (1| genre),  
 data = data\_2, control = lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e5)))  
summary(fit\_wms)  
  
Models\_data<-stargazer(fit\_title,fit\_genre,fit\_pub,fit\_gp, fit\_wms, type='text')  
  
Models\_data<-stargazer(fit\_title,fit\_genre,fit\_pub,type='html',out='a.html')  
  
#building models by scaling the variables  
fit\_title\_sc <- lmer(weekly\_sales ~ market\_share + price + week\_no + top\_10 + genre + (1| game\_title),  
 data = data\_scaled, control = lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e5)))  
  
  
fit\_pub\_sc <- lmer(weekly\_sales ~ market\_share + price + week\_no + top\_10 + genre + (1| publisher),  
 data = data\_scaled\_pub, control = lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e5)))  
  
  
Models\_scaled\_data<-stargazer(fit\_title\_sc, fit\_genre\_sc, fit\_pub\_sc, type = 'text')  
Models\_scaled\_data<-stargazer(fit\_title\_sc, fit\_genre\_sc, fit\_pub\_sc, type = 'html',out='b.html')  
  
  
# Calculate residuals and fitted values  
residuals\_title <- residuals(fit\_title)  
fitted\_values\_title <- fitted(fit\_title)  
  
residuals\_genre <- residuals(fit\_genre)  
fitted\_values\_genre <- fitted(fit\_genre)  
  
residuals\_pub <- residuals(fit\_pub)  
fitted\_values\_pub <- fitted(fit\_pub)  
  
#scaled genre  
residuals\_genre <- residuals(fit\_genre\_sc)  
fitted\_values\_genre <- fitted(fit\_genre\_sc)  
  
residuals\_pub <- residuals(fit\_pub\_sc)  
fitted\_values\_pub <- fitted(fit\_pub\_sc)  
  
#log-log model  
residuals\_genre\_log <- residuals(fit\_genre\_log)  
fitted\_values\_genre\_log <- fitted(fit\_genre\_log)  
  
# Create the individual residual vs. fitted plots for each model using ggplot()  
plot\_title <- ggplot() +  
 geom\_point(aes(x = fitted\_values\_title, y = residuals\_title)) +  
 geom\_hline(yintercept = 0, linetype = 2, color = "red") +  
 labs(x = "Fitted Values", y = "Residuals") +  
 ggtitle("Title Model")  
  
plot\_genre <- ggplot() +  
 geom\_point(aes(x = fitted\_values\_genre, y = residuals\_genre)) +  
 geom\_hline(yintercept = 0, linetype = 2, color = "red") +  
 labs(x = "Fitted Values", y = "Residuals") +  
 ggtitle("Genre Model")  
  
plot\_pub <- ggplot() +  
 geom\_point(aes(x = fitted\_values\_pub, y = residuals\_pub)) +  
 geom\_hline(yintercept = 0, linetype = 2, color = "red") +  
 labs(x = "Fitted Values", y = "Residuals") +  
 ggtitle("Publisher Model")  
  
# Combine the three plots into one using grid.arrange()  
grid.arrange(plot\_title, plot\_genre, plot\_pub, ncol = 3)  
  
# Calculate residuals and fitted values for scaled data models  
residuals\_title\_sc <- residuals(fit\_title\_sc)  
fitted\_values\_title\_sc <- fitted(fit\_title\_sc)  
  
residuals\_genre\_sc <- residuals(fit\_genre\_sc)  
fitted\_values\_genre\_sc <- fitted(fit\_genre\_sc)  
  
residuals\_pub\_sc <- residuals(fit\_pub\_sc)  
fitted\_values\_pub\_sc <- fitted(fit\_pub\_sc)  
  
# Create the individual residual vs. fitted plots for each scaled model using ggplot()  
plot\_title\_sc <- ggplot() +  
 geom\_point(aes(x = fitted\_values\_title\_sc, y = residuals\_title\_sc)) +  
 geom\_hline(yintercept = 0, linetype = 2, color = "red") +  
 labs(x = "Fitted Values", y = "Residuals") +  
 ggtitle("Scaled Title Model")  
  
plot\_genre\_sc <- ggplot() +  
 geom\_point(aes(x = fitted\_values\_genre\_sc, y = residuals\_genre\_sc)) +  
 geom\_hline(yintercept = 0, linetype = 2, color = "red") +  
 labs(x = "Fitted Values", y = "Residuals") +  
 ggtitle("Scaled Genre Model")  
  
  
plot\_pub\_sc <- ggplot() +  
 geom\_point(aes(x = fitted\_values\_pub\_sc, y = residuals\_pub\_sc)) +  
 geom\_hline(yintercept = 0, linetype = 2, color = "red") +  
 labs(x = "Fitted Values", y = "Residuals") +  
 ggtitle("Scaled Publisher Model")  
  
# Combine the three plots into one using grid.arrange()  
grid.arrange(plot\_title\_sc, plot\_genre\_sc, plot\_pub\_sc, ncol = 3)  
  
#Create the individual residual vs. fitted plots for log-log model  
plot\_genre\_log <- ggplot() +  
 geom\_point(aes(x = fitted\_values\_genre\_log, y = residuals\_genre\_log)) +  
 geom\_hline(yintercept = 0, linetype = 2, color = "red") +  
 labs(x = "Fitted Values", y = "Residuals") +  
 ggtitle("Log-log Model")  
  
#normality Checks  
# Q-Q plot for residuals  
qqnorm(residuals\_title)  
qqline(residuals\_title)  
  
qqnorm(residuals\_genre)  
qqline(residuals\_genre)  
title("Genre Model")  
  
qqnorm(residuals\_pub)  
qqline(residuals\_pub)  
title("Publisher Model")  
  
#Anderson-Darling test for normality  
ad.test(residuals\_title)  
ad.test(residuals\_genre)  
ad.test(residuals\_pub)  
  
qqnorm(residuals\_title\_sc)  
qqline(residuals\_title)  
  
qqnorm(residuals\_genre\_sc)  
qqline(residuals\_genre\_sc)  
title("Genre Model")  
  
qqnorm(residuals\_pub\_sc)  
qqline(residuals\_pub\_sc)  
title("Publisher Model")  
  
  
# Kolmogorov-Smirnov test for residuals\_title\_genre  
ks\_genre <- ks.test(residuals\_genre, "pnorm", mean(residuals\_genre), sd(residuals\_genre))  
print(ks\_genre)  
  
# Kolmogorov-Smirnov test for residuals\_genre\_sc  
ks\_genre\_sc <- ks.test(residuals\_genre\_sc, "pnorm", mean(residuals\_genre\_sc), sd(residuals\_genre\_sc))  
print(ks\_genre\_sc)  
  
  
residuals\_fit\_genre\_log <- residuals(fit\_genre\_log)  
fitted\_genre\_log<- fitted(fit\_genre\_log)  
  
# Kolmogorov-Smirnov test for residuals\_fit\_genre\_log  
ks\_genre\_log <- ks.test(residuals\_fit\_genre\_log, "pnorm", mean(residuals\_fit\_genre\_log), sd(residuals\_fit\_genre\_log))  
print(ks\_genre\_log)  
  
  
  
  
  
#cross validation for genre models based on research statement  
# Set seed for reproducibility  
set.seed(123)  
  
any(is.na(data\_1$weekly\_sales))  
data\_imputed <- data\_1[!is.na(data\_1$weekly\_sales), ]  
  
# Split the dataset into training (80%) and testing (20%) sets  
train\_indices <- createDataPartition(data\_imputed$weekly\_sales, p = 0.8, list = FALSE)  
train\_data <- data\_scaled[train\_indices, ]  
test\_data <- data\_scaled[-train\_indices, ]  
  
# Fit the model on the training set  
fit\_genre <- lmer(weekly\_sales ~ market\_share + price + week\_no + top\_10 + genre + (1| genre),  
 data = train\_data, control = lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e5)))  
  
  
fit\_genre\_sc <- lmer(weekly\_sales ~ market\_share + price + week\_no + top\_10 + (1| genre),  
 data = train\_data, control = lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e5)))  
  
  
fit\_genre\_log <- lmer(log(weekly\_sales) ~ market\_share + log(price) + week\_no + top\_10 + (1| genre),  
 data = train\_data, control = lmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e5)))  
  
  
# Calculate predictions on the testing set  
test\_data$predicted\_sales <- predict(fit\_genre, newdata = test\_data)  
test\_data$predicted\_sales <- predict(fit\_genre\_sc, newdata = test\_data)  
test\_data$predicted\_sales <- predict(fit\_genre\_log, newdata = test\_data)  
  
# Evaluate the performance using mean squared error (MSE) and R-squared  
mse <- mean((test\_data$weekly\_sales - test\_data$predicted\_sales)^2)  
cat("Mean Squared Error:", mse, "\n")  
  
r\_squared <- 1 - mse / var(test\_data$weekly\_sales)  
cat("R-squared:", r\_squared)  
  
a<-rnorm(1000)  
b<-ks.test(a,residuals\_genre\_sc)

**References:**

* *Exploring survival rates of companies in the UK video-games industry : An empirical study - White Rose Research Online*. (n.d.). Exploring Survival Rates of Companies in the UK Video-games Industry : An Empirical Study - White Rose Research Online. <https://eprints.whiterose.ac.uk/113196/>
* Romano, S. (2019, February 18). *Famitsu 2018 Japan video game sales report*. Gematsu. <https://www.gematsu.com/2019/02/famitsu-2018-japan-video-game-sales-report>
* Prasad, Aashish. (2019). Estimating Video Game Success using Machine Learning. 10.13140/RG.2.2.14389.01767.
* Takahashi, D. (2021, July 4). *Newzoo: Game market will hit $200B in 2024*. VentureBeat. <https://venturebeat.com/games/newzoo-game-market-will-hit-200b-in-2024/>
* <https://chat.openai.com>