Project Overview

What defines a Games success? A review of STEAM games

4770 Final project overview

DATASET: https://www.kaggle.com/datasets/fronkongames/steam-games-dataset

PREMISE: Extract information based on a Games performance, User reviews, pricing, and overall “Success” using data mining techniques

GOALS:

1. Define and quantify what makes a game a “success”

2. Compare paid versus free games

3. Analyze genre tags and compare game types

4. Visualize these results and explore what they mean

How to meet these goals

Goal 1: retrieve a dataset off of kaggle containing steam games verify its collection process and the tools used to scrap the data

Goal 2: look for metrics such as # of reviews, avg player rating(positive or negative), player count Download/purchase numbers and any other relative data we'll load these into mySQL

Goal 3: Connect the data to jupyter labs and tableau in order to query and visualize the data found on mySQL

Goal 4: explore the data and record the code used to query the data and the different calculated fields in tableau

Goal 5: clean the data and prepare it for modeling remove null values , add column paid or free

Goal 6: k means clustering

Goal 7: regression and prediction

Goal 8: final analysis answering the question and goals we have defined

Goal 9: put all of this together in tableau and slides and begin final report

Final report

CAP4770 Final Project Report:
A data comparison of Free vs Paid games

Team Members: Damien Teston, Deep Dabhi, Andres Gal'lino

Problem Statement and Background

The goal of this project is to analyze whether free games are enough to fulfill players based on player engagement data, review statistics, and popularity metrics. We are also ultimately looking to find out what free games have to offer and how they compete with paid games. The motivation behind this research is due to the gaming industry having recently been filled with more free-to-play models with a good amount of success, which led to our interest in wanting to look into this comparison overall through the lens of data analysis. We wish to solve the question of whether these games can compete with paid titles in delivering the same experience to the player. By applying predictive modeling techniques, this project seeks to understand what free games offer, how they compete with paid games, and whether free games can provide fulfillment comparable to paid games. For collecting our dataset, we decided to use the Steam platform's game analytics. Along with this, we have chosen to use Kaggle for a convenient way to find these statistics all at once.

Data Collection and Preprocessing

For this project, we decided to collect video game statistics from Steam since steam is the largest gaming platform out there we thought it perfect to be our sampling pool.

Using Kaggle we found a dataset fitting our needs we looked for a large quantity of games the Data set has 111,446 games with around 23,242 being free games and the rest being paid games we had to keep this in mind during our data presentation as to not have skewed results in the comparison charts.

List of attributes in the data set:

```
----
Index(['Name', 'Date', 'Estimated_Owners_Range', 'Peak CCU', 'Age Required',
       'Price', 'null', 'DiscountDLC count', 'About the game',
       'Supported languages', 'Full audio languages', 'Reviews',
       'Header image', 'Website', 'Support url', 'Support email', 'Windows',
```

```
       'Mac', 'Linux', 'Metacritic score', 'Metacritic url', 'User score',
       'Positive', 'Negative', 'Score rank', 'Achievements', 'Recommendations',
       'Notes', 'Average playtime forever', 'Average playtime two weeks',
       'Median playtime forever', 'Median playtime two weeks', 'Developers',
       'Publishers', 'Categories', 'Genres', 'Tags', 'Screenshots', 'Movies',
       'Release_Year', 'is_free'],
      dtype='object')
```

For processing this data it was uploaded into MySQL and accessed through Juypterlabs which is where a lot of the cleaning was primarily done.

During this transition the values for some of the fields shifted so that had to be edited as it was causing errors whenever we tried to fetch game names and instead returned the release dates.

For the cleaning, null values were removed and many different fields were changed to fit the proper value they represent such as release dates going by year.

In the dataset we had to determine what fields we needed to look at and review so a calculated field was made in tableau after connecting to the MySQL database using Price as a base to count the number of games that were “free” and those that cost money were “paid”. This was how we visualized the two types of games.

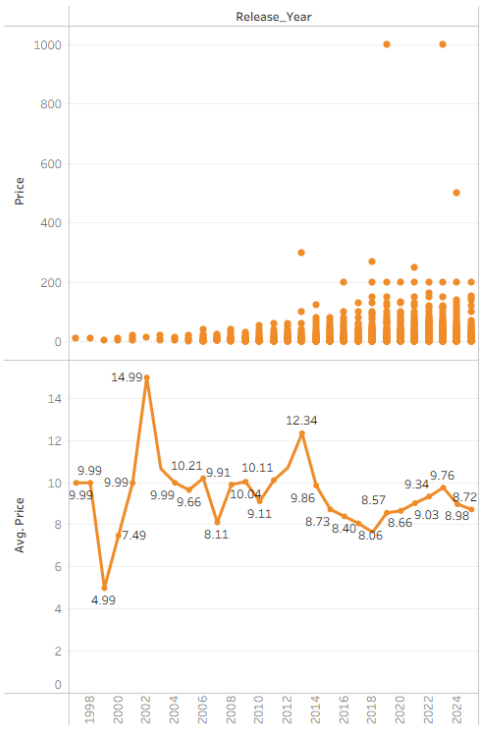
Later, a new column was added to the csv to signify the game’s type. That way, database queries for the types would be easier.

Another calculated field we needed was review ratio. This was a product of negative and positive reviews left on a game, which was used to gauge a users average enjoyment of a game. Prior studies show that Steam reviews can be effectively used to analyze player sentiment and engagement (Guzsvinecz & Szűcs, 2024).

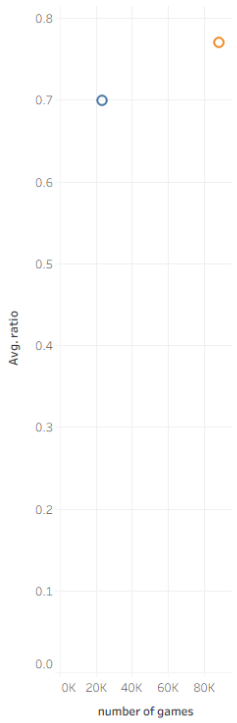
Exploring Data/Data Modeling

Below is one of the first tableau dashboards we made. We used this as an analysis for the changing average of price over the years, showing that the average price of a random paid game stays pretty constant and only fluctuates every few years before returning to around the 10.00 mark. Another key depiction we wanted to discern here is that even despite the difference in sum, the two categories of games are not that far off when it comes to user review. We can note that paid games slightly score above free games.

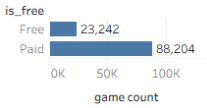
average price per year



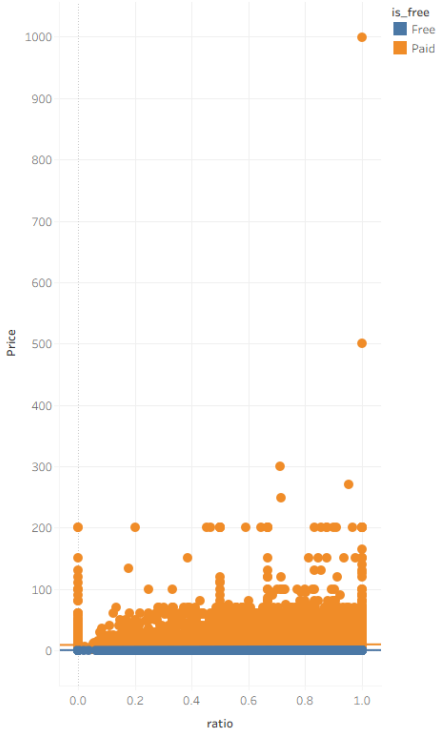
review average



game count



Price/review line regression



Ratio vs. Price. Color shows details about is_free. Details are shown for Name. The view is filtered on is_free, which keeps Free and Paid.

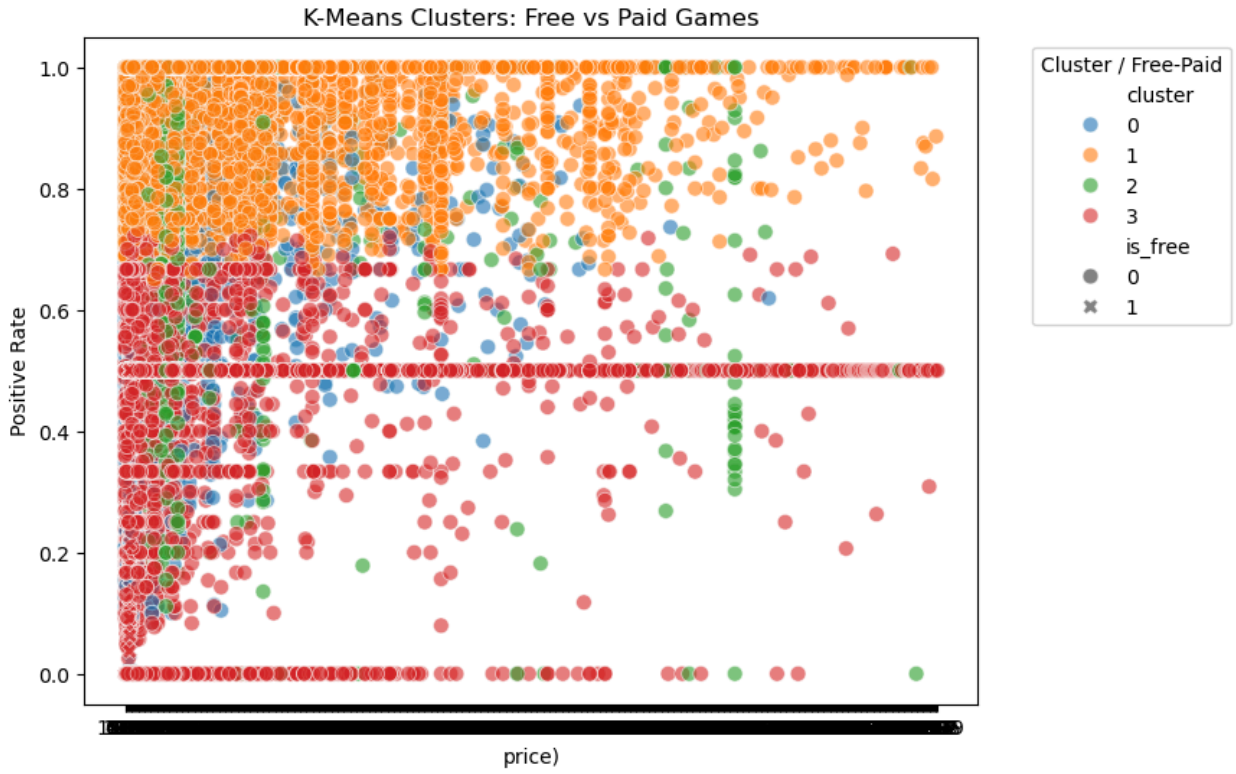
To determine whether price is related to player satisfaction, a correlation and regression analysis was conducted between game price and positive review ratio.

P-value: < 0.0001
Equation: Price = 0.830651*ratio + 8.50677

Coefficients

Term	Value	StdErr	t-value	p-value
ratio	0.830651	0.176845	4.69706	< 0.0001
intercept	8.50677	0.143353	59.3414	< 0.0001

This showed that games with higher review ratios did turn out to usually be a paid game. It is important to note that the slope is very modest which suggests that despite price having some sway, other factors such as genre or DLC also strongly influence reviews. In a recent study, Guzsvinecz and Szűcs (2024) analyzed Steam game reviews and found that sentiment patterns vary significantly across major game genres.



Is free	0	1	
cluster			
0	19858	4173	
1	30365	2150	
2	1683	0	
3	27695	14076	

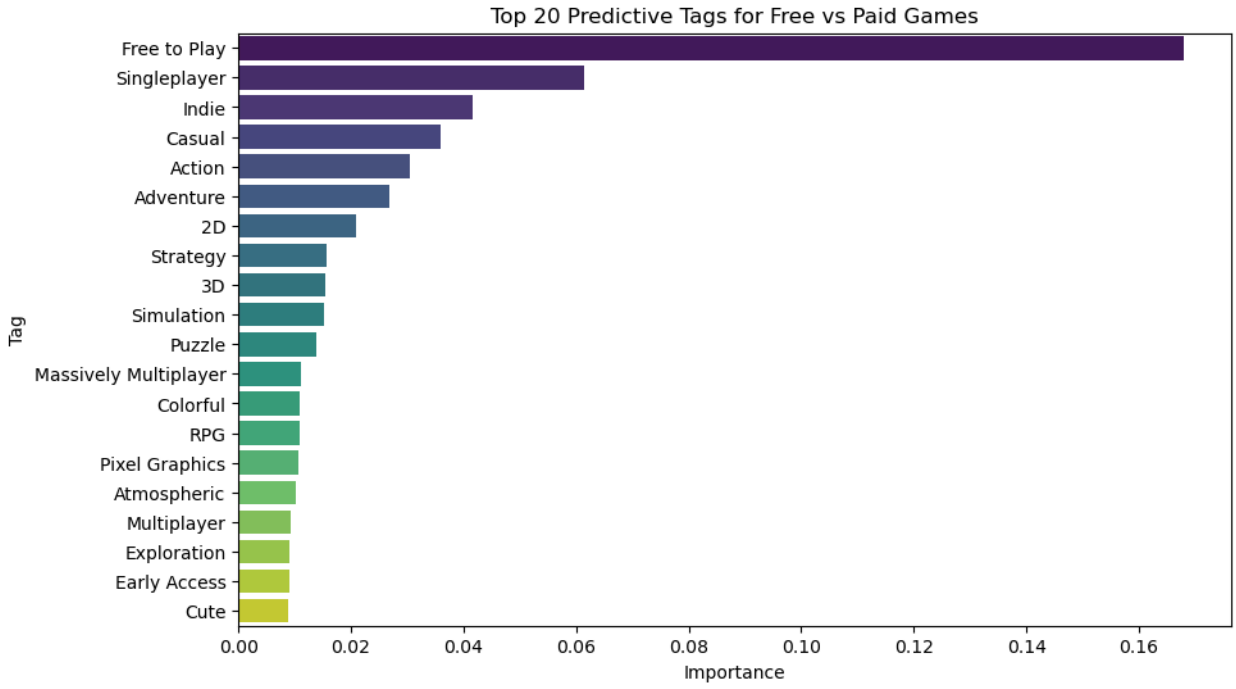
In our analysis of Steam games, we used K-Means clustering to group games based on price, positive review rates, and release year. The goal was to compare free and paid games and see if patterns emerge beyond just cost. The clustering produced four groups, each with distinct characteristics. Any missing or null values were filled with 0.5 to avoid calculation errors.

Cluster 3 contains most of the free games along with some very low-priced paid games. While these games share low prices, their positive review rates vary, showing that some free games are extremely popular while others receive mixed feedback. Cluster 0 and Cluster 1 are mostly paid games, with cluster 1 representing higher-priced or widely praised titles. Cluster 2 is a small group of paid games that are outliers, likely due to unusual price points, release years, or unique reception.

The analysis shows that free games can achieve high popularity despite having no cost, and some low-priced paid games behave similarly to free games in terms of user reviews. Paid games display more diversity, indicating that price alone doesn't determine user reception. By examining these clusters, we can clearly see trends in game popularity, success, and user engagement, highlighting which free or paid games stand out based on their reviews and overall reception rather than just their price.

word cloud Paid

word map Free



Top 20 tags in FREE games (by percentile):

tag	Free_pct	Paid_pct	Diff_pct
Indie	0.234619	0.473725	-0.239106
Free to Play	0.186921	0.003555	0.183366
Action	0.168244	0.343476	-0.175232
Casual	0.144713	0.333438	-0.188725
Adventure	0.134418	0.328513	-0.194095
Singleplayer	0.122261	0.431929	-0.309668
Multiplayer	0.085004	0.083441	0.001563
Strategy	0.081278	0.163691	-0.082413
RPG	0.075739	0.143327	-0.067588
Simulation	0.075053	0.164282	-0.089229
Early Access	0.065640	0.111707	-0.046067
2D	0.061915	0.227535	-0.165620
VR	0.055395	0.053328	0.002066
Massively Multiplayer	0.043434	0.010113	0.033321
Atmospheric	0.041816	0.145036	-0.103220
Shooter	0.035737	0.081054	-0.045317
First-Person	0.035737	0.101029	-0.065292
Puzzle	0.035688	0.146518	-0.110830
Story Rich	0.035296	0.116079	-0.080783
3D	0.034217	0.138667	-0.104449

Top 20 tags in PAID games (by percentile):

tag	Free_pct	Paid_pct	Diff_pct
Indie	0.234619	0.473725	-0.239106
Singleplayer	0.122261	0.431929	-0.309668
Action	0.168244	0.343476	-0.175232
Casual	0.144713	0.333438	-0.188725
Adventure	0.134418	0.328513	-0.194095
2D	0.061915	0.227535	-0.165620
Simulation	0.075053	0.164282	-0.089229
Strategy	0.081278	0.163691	-0.082413
Puzzle	0.035688	0.146518	-0.110830
Atmospheric	0.041816	0.145036	-0.103220
RPG	0.075739	0.143327	-0.067588
3D	0.034217	0.138667	-0.104449
Pixel Graphics	0.031717	0.119446	-0.087728
Colorful	0.025246	0.118076	-0.092830
Story Rich	0.035296	0.116079	-0.080783
Early Access	0.065640	0.111707	-0.046067
Exploration	0.025001	0.108604	-0.083603
Cute	0.023334	0.107511	-0.084177

First-Person	0.035737	0.101029	-0.065292
Arcade	0.024217	0.100464	-0.076247

Top 20 tags in FREE games (by count, descending):

tag	Free	Paid	Diff
Indie	4786	37709	-32923
Free to Play	3813	283	3530
Action	3432	27341	-23909
Casual	2952	26542	-23590
Adventure	2742	26150	-23408
Singleplayer	2494	34382	-31888
Multiplayer	1734	6642	-4908
Strategy	1658	13030	-11372
RPG	1545	11409	-9864
Simulation	1531	13077	-11546
Early Access	1339	8892	-7553
2D	1263	18112	-16849
VR	1130	4245	-3115
Massively Multiplayer	886	805	81
Atmospheric	853	11545	-10692
Shooter	729	6452	-5723
First-Person	729	8042	-7313
Puzzle	728	11663	-10935
Story Rich	720	9240	-8520
3D	698	11038	-10340

Top 20 tags in PAID games (by count, descending):

tag	Free	Paid	Diff
Indie	4786	37709	-32923
Singleplayer	2494	34382	-31888
Action	3432	27341	-23909
Casual	2952	26542	-23590
Adventure	2742	26150	-23408
2D	1263	18112	-16849
Simulation	1531	13077	-11546
Strategy	1658	13030	-11372
Puzzle	728	11663	-10935
Atmospheric	853	11545	-10692
RPG	1545	11409	-9864
3D	698	11038	-10340
Pixel Graphics	647	9508	-8861
Colorful	515	9399	-8884
Story Rich	720	9240	-8520
Early Access	1339	8892	-7553

Exploration	510	8645	-8135
Cute	476	8558	-8082
First-Person	729	8042	-7313
Arcade	494	7997	-7503

The tag analysis of free and paid Steam games reveals clear differences in design, focus and player engagement strategies. Among free games, the most common tags by percentile are Indie (23.5%), Free-to-Play (18.7%), Action (16.8%), Casual (14.5%), and Adventure (13.4%), showing that free titles emphasize accessibility, fast-paced gameplay, and broad audience appeal. Multiplayer also appears at nearly the same rate in both free and paid games, while Massively Multiplayer and VR appear slightly more often in free titles, highlighting the importance of online ecosystems and experimental platforms in the free-to-play market.

In contrast, paid games are strongly dominated by Indie (47.4%) and Singleplayer (43.2%), followed by Action, Casual, and Adventure, indicating a stronger focus on solo, structured experiences. Paid games also show much higher representation in Story Rich, Atmospheric, Puzzle, Exploration, and visual style tags such as 2D, 3D, Pixel Graphics, and Colorful, which suggest greater emphasis on narrative depth and artistic presentation. Raw tag counts further reinforce this pattern, as paid games outnumber free games by tens of thousands across most categories, reflecting the larger paid catalog.

Overall, while both markets share core genres, free games prioritize live-service, multiplayer, and accessibility, whereas paid games emphasize narrative depth, visual identity, and single-player immersion.

Results and Interpretation

This project began with a simple core question: Are free-to-play games good enough compared to paid games? Based on our analysis, the data shows that free-to-play games often deliver equal or greater long-term value and success than paid games. Free games consistently maintain larger and more stable player bases over time, while many paid games experience a noticeable decline after their initial launch period.

One of the biggest factors influencing long-term success was game features rather than price alone. Games that support multiplayer modes and receive regular updates with new content tend to retain players at a much higher rate. This helps explain why many free-to-play titles remain popular for long periods, while paid games—especially single-player titles—often peak early and decline once most users complete the content.

The data also revealed clear genre-based trends. Free-to-play models perform especially well in genres such as FPS, MOBAs, and team-based multiplayer games, where competition,

replayability, and social interaction drive long-term engagement. In contrast, paid games tend to perform better in RPGs, story-driven titles, and single-player experiences, where players are often willing to pay upfront for structured content and narrative depth.

Although free-to-play games provide strong value, many rely heavily on in-game purchases and microtransactions to generate profit. This shows that while players may enter for free, revenue is often driven by optional spending rather than initial price.

Application to Real Life

How can consumers and game developers use this data?

For consumers:

This data helps players make smarter purchasing decisions by showing that high-quality and long-lasting gaming experiences do not always require paying upfront. Players can use these insights to identify high-value free games, explore new genres with less financial risk, and better understand which types of games are more likely to stay active over time.

For game developers and studios:

Developers can use this data to gain insight into what drives long-term game success. Understanding which genres perform best under free-to-play versus paid models helps studios make better pricing strategy decisions. The trends also support more effective targeted marketing, improved content update planning, and stronger player retention strategies. Additionally, platforms such as Steam can use this information to improve recommendation algorithms and better promote games based on player behavior.

References and Citations

DATASET: https://www.kaggle.com/datasets/fronkongames/steam-games-dataset

SteamDB. (2024). *Steam statistics and game trends*. https://steamdb.info

- Used to support player counts, game popularity, and long-term engagement trends.

Valve Corporation. (2024). *Steam platform and recommendation system overview*. https://store.steampowered.com

- Supports how Steam uses engagement data, tags, and algorithms.

Thomas, I. (2022, October 6). *How free-to-play and in-game purchases took over the video game industry*. CNBC. https://www.cnbc.com/2022/10/06/how-free-to-play-and-in-game-purchases-took-over-video-games.html

- Article used to back F2P models.

Guzsvinecz, T., & Szűcs, J. (n.d.). *Length and sentiment analysis of reviews about top-level video game genres on the steam platform*. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0747563223003060

Juypter code

Upload code

```
import pandas as pd
from sqlalchemy import create_engine, text

#csvpath
csv_path = r"C:\Users\Xtest.DESKTOP-A1VKC8M\Downloads\archive\games.csv"

# Read CSV
df = pd.read_csv(csv_path)
print(df.head())

# Connect to MySQL
engine = create_engine("mysql+pymysql://root:copra^outrage1@localhost/steam_db")

#Drop the table if it exists
TABLE_NAME = "steam_games"
with engine.begin() as conn:
    conn.execute(text(f"DROP TABLE IF EXISTS `{TABLE_NAME}`"))

# Create table with LONGTEXT columns
columns = ",\n".join([f"`{col}` LONGTEXT" for col in df.columns])
create_sql = f"CREATE TABLE `{TABLE_NAME}` (\n{columns}\n)"
with engine.begin() as conn:
    conn.execute(text(create_sql))

print(f"Table `{TABLE_NAME}` created with all LONGTEXT columns.")

# Insert CSV data into MySQL table
df.to_sql(name=TABLE_NAME, con=engine, if_exists='append', index=False, chunksize=500)
print(f"Inserted {len(df)} rows into `{TABLE_NAME}`.")
```

Cleaning

```
df.rename(columns={"Name": "Date"}, inplace=True)
df.rename(columns={"AppID": "Name"}, inplace=True)
df.rename(columns={"Release date": "Estimated_Owners_Range"}, inplace=True)
df.rename(columns={"Price": "null"}, inplace=True)
df.rename(columns={"Required age": "Price"}, inplace=True)
df.rename(columns={"Peak CCU": "Age Required"}, inplace=True)
df.rename(columns={"Estimated owners": "Peak CCU"}, inplace=True)
```

During the transfer of data some rows got shifted so it was necessary to rename them within juypter

```
# paste and run this in one cell (requires engine from before)
from sqlalchemy.dialects.mysql import LONGTEXT
import math

# make a dtype mapping so every column becomes LONGTEXT in MySQL
dtype_map = {col: LONGTEXT() for col in df.columns}

# safer small chunks and multi-row insert
df.to_sql(
    name="steam_games",
    con=engine,
    if_exists="replace",   # overwrite the table with LONGTEXT columns
    index=False,
    dtype=dtype_map,
    chunksize=500,         # smaller chunk to avoid huge single statements
    method="multi"         # faster multi-row inserts
)

print("✅ uploaded with LONGTEXT for all columns")
```

Imports data back to MySQL

Checks table

```
import pandas as pd

query = "SELECT * FROM steam_games LIMIT 10;"
df_check = pd.read_sql(query, engine)
df_check
```

sql code

```
--1️⃣ Normalize Price column: set nulls or 'null' strings to 0
UPDATE steam_games
SET Price = 0
WHERE Price IS NULL OR Price = 'null';



-- Fill Release_Year for full dates like 'Oct 21, 2008'
UPDATE steam_games
SET Release_Year = YEAR(STR_TO_DATE(`Date`, '%b %d, %Y'))
WHERE `Date` REGEXP '^[A-Za-z]{3} [0-9]{1,2}, [0-9]{4}$';

-- Fill Release_Year for month + year like 'May 2020'
UPDATE steam_games
SET Release_Year = YEAR(STR_TO_DATE(CONCAT('01 ', `Date`), '%d %b %Y'))
WHERE `Date` REGEXP '^[A-Za-z]{3} [0-9]{4}$';

-- Fill Release_Year for numeric years like '2009'
UPDATE steam_games
SET Release_Year = CAST(`Date` AS UNSIGNED)
WHERE `Date` REGEXP '^[0-9]{4}$';
```

This was used to clean the data in the release date column to get a list of release year

contributions

Deep and andres developed the presentation slides and key tableau figures and majorly arranged the final report while providing project analysis and implementation during the project. reviewing the data and models while giving insightful feedback for the result interpretation.

Damien was in charge of data acquisition and preparation this included uploading the data onto the MySQL database and cleaning the dataset before eventually modeling it

Together the team contributed towards project planning, modeling and analysing the data.

This project showed how important clean and well-prepared data is before any analysis can be done. We learned that missing values, inconsistent formatting, and errors in the dataset can strongly affect results if they are not handled correctly.

We also learned how to choose the right analysis methods for different goals. Comparing free and paid games and using clustering helped us understand player behavior in new ways, but it also showed that not every question can be answered with one model.

Another key lesson was the value of visualization. Creating dashboards in Tableau helped us turn large amounts of data into clear visuals that made trends easier to understand and explain.

From a technical perspective, we gained experience connecting different tools together, including MySQL, Python, and Tableau. This helped us understand what a real-world data workflow looks like from start to finish.

Finally, we learned how important teamwork and communication are. Regular feedback, shared editing, and collaboration improved the quality of both the report and the presentation.