Steam Games Success Analysis

YASAMIN HOSSEINZADEH SANI

DATA SCIENCE AND BIG DATA ANALYSIS PROJECT

Department of Computer Engineering - Data Science
University of Pavia, Italy

25th January 2025

Introduction

THIS PROJECT ANALYZES HOW GENRE, PRICING STRATEGY, AND USER REVIEWS IMPACT A GAME'S SUCCESS. SUCCESS IS MEASURED THROUGH:

- 📊 **SALES PERFORMANCE**
- ⏳ **PLAYTIME**
- ⭐ **USER RATINGS**

Hypothesis : The genre, pricing strategy, and user reviews collectively influence a game's success on the Steam platform, measured by sales performance, playtime, and user ratings, with unique patterns and relationships that vary depending on these factors.

Project Overview : Steps Taken

1. Environment Setup
2. Data Upload
3. Data Cleaning & Preprocessing
4. Feature Engineering
5. Exploratory Data Analysis (EDA)
6. MapReduce Jobs
7. MongoDB Integration
8. Machine Learning with PySpark
9. Model Evaluation

1 Environment Setup

1 Installed & Configured:

✅ Hadoop (HDFS) – For distributed data storage.
✅ Apache Spark (Local Mode) – Used for data processing on a single machine.
✅ MongoDB – NoSQL database for storing processed data.

2 Virtual Machine Setup:

💻 Configured a virtual machine with all required tools.
📌 Installed Hadoop, Spark, and MongoDB for data handling.

3 Storage & Integration:

🔗 HDFS used for storing raw and processed datasets.
🔗 MongoDB integrated for querying structured data efficiently.

2 Data Upload

Storage in HDFS:

- 🗄️ Hadoop Distributed File System (HDFS) was used to store large datasets for efficient processing.
- 🔹 Used hdfs dfs -put command to upload files into HDFS.

```bash
hdfs dfs -put /local/path/dataset.csv /steam_data/
```

```
ubuntu@dsbda-vm:~$ hdfs dfs -ls /steam_data/
Found 7 items
drwxr-xr-x   - ubuntu supergroup          0 2025-02-19 18:11 /steam_data/cleaned_feature_engineered_unified.csv
drwxr-xr-x   - ubuntu supergroup          0 2025-02-19 18:11 /steam_data/cleaned_steam.csv
drwxr-xr-x   - ubuntu supergroup          0 2025-02-19 18:11 /steam_data/feature_engineered_steam.csv
drwxr-xr-x   - ubuntu supergroup          0 2025-02-19 18:11 /steam_data/output_mapreduce
drwxr-xr-x   - ubuntu supergroup          0 2025-02-19 18:11 /steam_data/output_mapreduce2
drwxr-xr-x   - ubuntu supergroup          0 2025-02-19 18:11 /steam_data/output_mapreduce3
-rw-r--r--   1 ubuntu supergroup    5815053 2025-02-19 18:11 /steam_data/steam.csv
ubuntu@dsbda-vm:~$
```

Contents of my HDFS

3️⃣ Data Cleaning & Preprocessing

💻 Framework: Apache Spark (PySpark)

Utilized PySpark DataFrames instead of Pandas for distributed data processing.

🗄️ Storage & Processing: Hadoop HDFS

Data was stored and retrieved from HDFS, ensuring scalability for large files.

📂 Cleaned data saved in HDFS:

hdfs://localhost:9000/steam_data/cleaned_steam.csv

3 Data Cleaning & Preprocessing

✅ Standardized Column Names:

Renamed steam_appid or appid to app_id for consistency.

✅ Removed Duplicates:

Ensured no duplicate entries exist in the dataset.

✅ Trimmed Whitespace:

Removed extra spaces from all string columns.

✅ Cleaned & Converted Price Values:

Converted prices to float format.
Replaced empty prices with null values.

✅ Fixed Numeric Columns:

Converted positive_ratings, negative_ratings, average_playtime, and price to numeric types.

✅ Dropped Invalid Rows:

Removed rows with all null values.
Ensured critical columns (app_id, name) don’t have missing values.

4️⃣ Feature Engineering

1️⃣ Objective:

✔️ Convert raw game data into meaningful features that improve analysis.
✔️ Prepare dataset for predictive modeling and hypothesis testing.

2️⃣ Features Created & Transformed:

✅ Genre One-Hot Encoding 🎮
✅ Pricing Strategy 💰
✅ User Sentiment Score ⭐
✅ Game Success Metric 📈

3️⃣ Output Data:

📂 Feature-engineered dataset stored in HDFS:
hdfs://localhost:9000/steam_data/feature_engineered_steam.csv

✅ Genre One-Hot Encoding 🎮

✅ Original Format:

The genres column contains multiple genres separated by ";" (e.g., "Action;Adventure;RPG").

✅ Transformation Applied:

Extracted all unique genres from the dataset.
Created binary (0/1) columns for each genre.
If a game belongs to a genre, it gets "1", otherwise "0".

```
>>> df.select("genres", "is_Action", "is_Adventure", "is_RPG", "is_Simulation").show(5)
+--------------------+---------+------------+------+-------------+
|              genres|is_Action|is_Adventure|is_RPG|is_Simulation|
+--------------------+---------+------------+------+-------------+
|              Action|        1|           0|     0|            0|
|              Action|        1|           0|     0|            0|
|          Simulation|        0|           0|     0|            1|
|Adventure;Casual;...|        0|           1|     0|            0|
|Action;Adventure;RPG|        1|           1|     1|            0|
+--------------------+---------+------------+------+-------------+
only showing top 5 rows
```

✅ Pricing Strategy 💰

✅ Original Format:

- The price column contains raw numerical values in USD (e.g., 0.00, 4.99, 19.99, etc.).
- Some prices are missing or empty, which were handled in data cleaning.

✅ Transformations Applied:

Grouped games into four pricing categories to analyze trends.

```
>>> df.groupBy("price_tier").count().show()
+----------+-----+
|price_tier|count|
+----------+-----+
|      High|  324|
|       Low|19592|
|    Medium| 4599|
|      Free| 2560|
+----------+-----+
```

✅ User Sentiment Score ⭐

✅ Original Format:

- Two separate columns:
  - positive_ratings → Number of users who liked the game.
  - negative_ratings → Number of users who disliked the game.

✅Transformation Applied:

- Calculated sentiment score using the formula:

$$\text{sentiment\_score} = \frac{\text{positive\_ratings}}{\text{positive\_ratings} + \text{negative\_ratings}}$$

- If a game has no ratings, the score is set to NULL to avoid division errors.

```
>>> df.select("positive_ratings", "negative_ratings", "sentiment_score").show(5)
+----------------+----------------+------------------+
|positive_ratings|negative_ratings|   sentiment_score|
+----------------+----------------+------------------+
|          9442.0|          2881.0| 0.766209526900917|
|            14.0|            15.0|0.4827586206896552|
|           676.0|           794.0|0.4598639455782313|
|           249.0|            31.0|0.8892857142857142|
|           492.0|           235.0|0.6767537826685007|
+----------------+----------------+------------------+
only showing top 5 rows
```

✅ Game Success Metric 📈

✅ Success Metric Formula (Weighted Score) :

$$\text{success\_metric} = (\text{positive\_ratings} \times 0.5) + (\text{average\_playtime} \times 0.3) + (\text{owners\_numeric} \times 0.2)$$

positive_ratings	average_playtime	owners_numeric	success_metric
9442.0	322.0	3500000	704817.6
14.0	0.0	10000	2007.0
676.0	0.0	150000	30338.0
249.0	3.0	150000	30125.4
492.0	654.0	150000	30442.2

only showing top 5 rows

5 Exploratory Data Analysis (EDA)

✅Objective of EDA :

- ✓ Identify patterns in game performance, pricing, genres, and reviews.
- ✓ Understand factors influencing success metric.
- ✓ Detect missing values, outliers, and correlations.

✅Key Insights from EDA :

✓Top Genres by Success 📊:

```
+--------------------+------------------+
|genre               |avg_success       |
+--------------------+------------------+
|Massively Multiplayer|125395.6846473029 |
|Free to Play        |120873.09571596244|
|Action              |40516.43568848189 |
|RPG                 |33086.53976798144 |
|Animation & Modeling|30741.81518987341 |
|Strategy            |29765.517499046906|
|Design & Illustration|25853.335632183906|
|Adventure           |22447.010367859624|
|Utilities           |20761.47739726027 |
|Simulation          |20708.316153253763|
+--------------------+------------------+
only showing top 10 rows
```

```
exploded_df.groupBy("genre").agg(mean("success_metric").alias("avg_success")).orderBy(desc("avg_success")).show(10)
```

5 Exploratory Data Analysis (EDA)

✔ Pricing Strategy & Success 💰

```
df.stat.corr("price", "success_metric")
```

```
+----------+------------------+----------+
|price_tier|avg_success       |game_count|
+----------+------------------+----------+
|Free      |90104.42081218273 |2561      |
|High      |78963.76130030959 |323       |
|Medium    |45433.76877582085 |4599      |
|Low       |14068.524761370012|19592     |
+----------+------------------+----------+
```

✔ Top Genres by Average Playtime 📊:

```
exploded_df.groupBy("genre").agg(mean("
average_playtime").alias("avg_playtime"
)) \

.orderBy(desc("avg_playtime")).show(10,
truncate=False)
```

```
+--------------------+-----------------+
|genre               |avg_playtime     |
+--------------------+-----------------+
|Massively Multiplayer|725.4840940525588 |
|Free to Play        |554.4377934272301 |
|Photo Editing       |429.5833333333333 |
|RPG                 |277.04825986078885|
|Strategy            |193.16545939763628|
|Simulation          |154.24085483249902|
|Adventure           |151.6718173661649 |
|Action              |144.01663446190037|
|Racing              |142.220703125     |
|Web Publishing      |136.21428571428572|
+--------------------+-----------------+
only showing top 10 rows
```

5 Exploratory Data Analysis (EDA)

✔ User Sentiment & Reviews ⭐

```
df.groupBy("sentiment_score").agg(mean(
"success_metric").alias("avg_success"))
 .orderBy(desc("avg_success")).show(10)
```

```
+------------------+-----------+
|   sentiment_score| avg_success|
+------------------+-----------+
|0.8587102445738107|3.04389367E7|
|0.8679519627192155|1.63289502E7|
|0.5046315318430526|1.52549734E7|
|0.9381067983233773|   7260488.0|
|0.9028501246698102|   7147261.4|
|0.9178798179969125|   7115024.0|
| 0.956772781556948|   3185587.1|
|0.7025678471461604|   3167481.6|
|0.8452187962100882|   3155521.0|
|0.9676488334287702|   3126379.0|
+------------------+-----------+
only showing top 10 rows
```

✔ Price Range Analysis💰:

```
df.withColumn("price_range",
when(col("price") < 10, "<10")
              .when((col("price") >= 10) &
(col("price") < 30), "10-30")
              .otherwise("30+")) \

.groupBy("price_range").agg(mean("average_pl
aytime").alias("avg_playtime")) \

.orderBy(desc("avg_playtime")).show(truncate
=False)
```

```
+-----------+------------------+
|price_range|avg_playtime      |
+-----------+------------------+
|30+        |1229.3746130030959|
|10-30      |247.6617423419509 |
|<10        |113.79534064743329|
+-----------+------------------+
```

6 MapReduce Jobs

1 Overview of MapReduce Jobs:

✓ Used Hadoop MapReduce to process large-scale Steam game data.
✓ Applied three MapReduce jobs to analyze genres, pricing, and reviews

2 MapReduce Jobs

✓The price tier and pricing strategy might correlate with a game success matrice
✓Certain game genres might lead to higher success based on playtime, ratings, or sales
✓Positive and negative ratings, alongside sentiment analysis likely impact the game's success

```
hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.4.1.jar
-input hdfs://localhost:9000/steam_data/feature_engineered_steam.csv
-output hdfs://localhost:9000/steam_data/output_pricing_success
-mapper mapreduce1_mapper.py
-reducer mapreduce1_reducer.py
```

✅Pricing Strategy & Success 💰

📌 Objective: Determine if pricing affects success metrics.

📌 Map: Assigns games to price tiers (Free, Low, Medium, High).

📌 Reduce: Computes average playtime & success metric per price tier.

📌 Output: A summary showing how price affects success.

```
ubuntu@dsbda-vm:~$ hdfs dfs -ls /steam_data/output_mapreduce
Found 2 items
-rw-r--r--   1 ubuntu supergroup          0 2025-02-07 11:17 /steam_data/output_mapreduce/_SUCCESS
-rw-r--r--   1 ubuntu supergroup         57 2025-02-07 11:17 /steam_data/output_mapreduce/part-00000
ubuntu@dsbda-vm:~$ hdfs dfs -cat /steam_data/output_mapreduce/part-00000
Free    90139.61
High    78720.12
Low     14068.52
Medium  45433.77
ubuntu@dsbda-vm:~$
```

✅Genre-Based Success Analysis 🎮

📌 Objective: Identify genres that lead to higher success (sales, playtime, ratings).

📌 Map: Extracts genres and their corresponding success metrics.

📌 Reduce: Aggregates the average success score per genre.

📌 Output: A ranked list of genres by success.

```
ubuntu@dsbda-vm:~$ hdfs dfs -cat /steam_data/output_mapreduce2/part-00000
Action  40516.44,144.02,1581.93,6.14
Adventure       22447.01,151.67,869.81,6.12
Casual  10874.13,85.07,273.91,4.11
Early Access    10746.85,81.30,388.05,7.04
Gore    9779.49,47.65,323.68,5.97
Indie   15318.09,112.85,540.15,5.05
Massively Multiplayer   125395.68,725.48,3468.11,4.32
RPG     33086.54,277.05,1346.53,6.94
Racing  17230.03,142.22,663.79,7.05
Simulation      20708.32,154.24,819.66,7.21
Sports  16031.16,115.52,519.48,7.88
Strategy        29765.52,193.17,863.86,6.94
Violent 9872.38,48.14,294.76,6.23
ubuntu@dsbda-vm:~$
```

Genre Success_Metric Avg_Playtime Avg_Positive_Ratings Avg_Sales(Price)

✅User Sentiment & Success ⭐

📌 Objective: Understand how positive & negative ratings impact game success.

📌 Map: Extracts positive & negative ratings alongside sentiment score.

📌 Reduce: Computes correlation between sentiment & success.

📌 Output: Sentiment analysis results showing the relationship between reviews & game success.

```
ubuntu@dsbda-vm:~$ hdfs dfs -cat /steam_data/output_mapreduce3/part-00000
Correlation between Positive Ratings and Success: 0.73
Correlation between Negative Ratings and Success: 0.70
Correlation between Sentiment Score and Success: 0.04
ubuntu@dsbda-vm:~$
```

7 MongoDB Integration

1 Objective:

- ✓ Store processed game data in MongoDB for fast querying and analysis.
- ✓ Perform aggregations & lookups to extract insights on success, pricing, and reviews

Steps Taken:

1 Converted MapReduce outputs into JSON using a Python script.

```
2025-02-25T01:05:33.042+0100 I
> show dbs
admin           0.000GB
config          0.000GB
local           0.000GB
netflix         0.003GB
netflix_data    0.152GB
netflix_db      0.003GB
steam_analysis  0.000GB
steam_games     0.129GB
>
```

```
Free    90139.61
High    78720.12
Low     14068.52
Medium  45433.77
```

2 Store JSON Data in MongoDB:

```
mongoimport --db steam_games --collection game_ratings --file mapreduce_output.json --jsonArray
```

3 Queried MongoDB for insights using find(), sort(), and aggregations.

7️⃣ MongoDB Integration

✅ 1. success metric for different price tiers sorted in descending order.

```
> db.game_ratings.find().sort({ value: -1 }).limit(5).pretty();
{
        "_id" : ObjectId("67a6658dd5ed6f524d99f152"),
        "key" : "Free",
        "value" : 90139.61
}
{
        "_id" : ObjectId("67a6658dd5ed6f524d99f14f"),
        "key" : "High",
        "value" : 78720.12
}
{
        "_id" : ObjectId("67a6658dd5ed6f524d99f151"),
        "key" : "Medium",
        "value" : 45433.77
}
{
        "_id" : ObjectId("67a6658dd5ed6f524d99f150"),
        "key" : "Low",
        "value" : 14068.52
}
>
```

7️⃣ MongoDB Integration

✅ 2. top 5 game genres sorted by success metric.

```
> db.genre_success.find().sort({ success_metric: -1 }).limit(5).pretty();
{
        "_id" : ObjectId("67a66412d5ed6f524d99f11e"),
        "genre" : "Massively Multiplayer",
        "success_metric" : 125395.68,
        "average_playtime" : 725.48,
        "positive_ratings" : 3468.11,
        "negative_ratings" : 4.32
}
{
        "_id" : ObjectId("67a66412d5ed6f524d99f118"),
        "genre" : "Action",
        "success_metric" : 40516.44,
        "average_playtime" : 144.02,
        "positive_ratings" : 1581.93,
        "negative_ratings" : 6.14
}
{
        "_id" : ObjectId("67a66412d5ed6f524d99f11f"),
        "genre" : "RPG",
        "success_metric" : 33086.54,
        "average_playtime" : 277.05,
        "positive_ratings" : 1346.53,
        "negative_ratings" : 6.94
}
{
        "_id" : ObjectId("67a66412d5ed6f524d99f123"),
        "genre" : "Strategy",
        "success_metric" : 29765.52,
        "average_playtime" : 193.17,
        "positive_ratings" : 863.86,
        "negative_ratings" : 6.94
}
{
        "_id" : ObjectId("67a66412d5ed6f524d99f119"),
        "genre" : "Adventure",
        "success_metric" : 22447.01,
        "average_playtime" : 151.67,
        "positive_ratings" : 869.81,
        "negative_ratings" : 6.12
}
>
```

7 MongoDB Integration

✅ 3. correlations between different rating metrics and game success..

```
> db.sentiment_correlation.find().pretty();
{
        "_id" : ObjectId("67a664e3d5ed6f524d99f134"),
        "Correlation between Positive Ratings and Success" : 0.73,
        "Correlation between Negative Ratings and Success" : 0.7,
        "Correlation between Sentiment Score and Success" : 0.04
}
>
```

✅ 4. success metrics for the "Action" genre.

```
> db.genre_success.find({ genre: "Action" }).pretty();
{
        "_id" : ObjectId("67a66412d5ed6f524d99f118"),
        "genre" : "Action",
        "success_metric" : 40516.44,
        "average_playtime" : 144.02,
        "positive_ratings" : 1581.93,
        "negative_ratings" : 6.14
}
>
```

8 Machine Learning with PySpark

1 Objective

✓ Predict whether a game is Free (1) or Paid (0) using game features.
✓ Train & compare three classification models:

- Logistic Regression
- Random Forest
- Gradient Boosting (GBT)

2 Data Preparation:

✅ Feature Engineering:

- Added a binary label (is_free) → 1 for Free, 0 for Paid.
- Filled missing values in numeric columns (average_playtime, ratings, etc.).
- Encoded genres as numerical values using StringIndexer.
- Assembled relevant features (playtime, ratings, sentiment_score) into a vector

8 Machine Learning with PySpark

3 Model Training & Testing:

✅ Dataset Split:

- 80% training, 20% testing.
- Cached the datasets for performance improvement.

✅ Trained 3 Models:

- Logistic Regression: A baseline classifier.
- Random Forest: Strong feature selection, handles missing values.
- Gradient Boosting (GBT): More powerful, learns complex relationships.

4 Sample Predictions

✅ Applied each model to the test dataset.

✅ Extracted predictions and probabilities.

8 Machine Learning with PySpark

```
Sample Predictions for Logistic Regression:
+----------------------------------------------+------------------------------------------------------+-------+----------+----------------------------------------+
|name                                          |features                                              |is_free|prediction|probability                             |
+----------------------------------------------+------------------------------------------------------+-------+----------+----------------------------------------+
|Ricochet                                      |[5.0,175.0,2758.0,684.0,0.8012783265543288]           |0      |0.0       |[0.9241413342167257,0.07585866578327427]|
|Counter-Strike: Global Offensive              |[94.0,22494.0,2644404.0,402313.0,0.8679519627192155]|1      |1.0       |[4.1516445835455004E-7,0.9999995848355416]|
|Disciples II: Rise of the Elves               |[11.0,0.0,451.0,108.0,0.8067978533094812]             |0      |0.0       |[0.9251678263068088,0.07483217369319117]|
|HeXen: Beyond Heretic                         |[5.0,50.0,249.0,26.0,0.9054545454545454]              |0      |0.0       |[0.9235854492439323,0.07641455075606773]|
|X2: The Threat                                |[11.0,732.0,212.0,50.0,0.8091603053435115]            |0      |0.0       |[0.922093343439897,0.07790665656010298] |
|Peggle™ Nights                                |[8.0,237.0,368.0,24.0,0.9387755102040817]             |0      |0.0       |[0.9214983256416382,0.07850167435836175]|
|RoboBlitz                                     |[0.0,78.0,43.0,14.0,0.7543859649122807]               |0      |0.0       |[0.9277890497534234,0.07221095024657664]|
|STAR WARS™ Jedi Knight - Jedi Academy™        |[5.0,1254.0,5684.0,330.0,0.9451280345859661]          |0      |0.0       |[0.915533313942608,0.08446668605739205] |
|X-COM: Enforcer                               |[11.0,0.0,146.0,138.0,0.5140845070422535]             |0      |0.0       |[0.9318588786556954,0.06814112134430461]|
|Men of War™                                   |[11.0,0.0,1664.0,257.0,0.8662155127537741]            |0      |0.0       |[0.9232842527319924,0.07671574726800756]|
|Battlestations Pacific                        |[54.0,132.0,1028.0,542.0,0.6547770700636942]          |0      |0.0       |[0.9214313547450727,0.07856864525492735]|
|Civilization IV: Beyond the Sword             |[11.0,108.0,1396.0,66.0,0.9548563611491108]           |0      |0.0       |[0.9209499895893364,0.07905001041066362]|
|Spear of Destiny                              |[5.0,0.0,200.0,36.0,0.847457627118644]                |0      |0.0       |[0.9251875018319033,0.07481249816809665]|
|Warhammer® 40,000: Dawn of War® - Winter Assault|[11.0,213.0,606.0,76.0,0.8885630498533724]          |0      |0.0       |[0.922286316316722,0.07771368368327802] |
|Sherlock Holmes: The Mystery of the Mummy     |[9.0,0.0,59.0,130.0,0.31216931216931215]              |1      |0.0       |[0.936399367574545,0.063600632425455]   |
|Sacred Gold                                   |[41.0,337.0,1489.0,319.0,0.8235619469026548]          |0      |0.0       |[0.9183934890314535,0.08160651096854654]|
|Rayman Raving Rabbids™                        |[17.0,29.0,239.0,80.0,0.7492163009404389]             |0      |0.0       |[0.9256566584246423,0.07434334157535771]|
|Helldorado                                    |[11.0,0.0,53.0,23.0,0.6973684210526315]               |0      |0.0       |[0.9279193099463217,0.07208069005367834]|
|Resident Evil™ 5/ Biohazard 5®                |[17.0,868.0,9250.0,1618.0,0.8511225616488775]         |0      |0.0       |[0.915215020755075,0.08478497924492501] |
|Penumbra: Black Plague Gold Edition           |[721.0,188.0,1396.0,79.0,0.9464406779661017]          |1      |0.0       |[0.7351072535415669,0.2648927464584331] |
+----------------------------------------------+------------------------------------------------------+-------+----------+----------------------------------------+
only showing top 20 rows
```

8 Machine Learning with PySpark

```
Sample Predictions for Random Forest:
+----------------------------------------------+-----------------------------------------------------+-------+----------+-----------------------------------------+
|name                                          |features                                             |is_free|prediction|probability                              |
+----------------------------------------------+-----------------------------------------------------+-------+----------+-----------------------------------------+
|Ricochet                                      |[5.0,175.0,2758.0,684.0,0.8012783265543288]          |0      |0.0       |[0.8725250899221109,0.1274749100778892] |
|Counter-Strike: Global Offensive              |[94.0,22494.0,2644404.0,402313.0,0.8679519627192155]|1      |0.0       |[0.5196960046294327,0.48030399537056734]|
|Disciples II: Rise of the Elves               |[11.0,0.0,451.0,108.0,0.8067978533094812]            |0      |0.0       |[0.8881167989465208,0.11188320105347918]|
|HeXen: Beyond Heretic                         |[5.0,50.0,249.0,26.0,0.9054545454545454]             |0      |0.0       |[0.8881167989465208,0.11188320105347918]|
|X2: The Threat                                |[11.0,732.0,212.0,50.0,0.8091603053435115]           |0      |0.0       |[0.8963928900885774,0.10360710991142263]|
|Peggle™ Nights                                |[8.0,237.0,368.0,24.0,0.9387755102040817]            |0      |0.0       |[0.8963928900885774,0.10360710991142263]|
|RoboBlitz                                     |[0.0,78.0,43.0,14.0,0.7543859649122807]              |0      |0.0       |[0.8881167989465208,0.11188320105347918]|
|STAR WARS™ Jedi Knight - Jedi Academy™        |[5.0,1254.0,5684.0,330.0,0.9451280345859661]         |0      |0.0       |[0.8963928900885774,0.10360710991142263]|
|X-COM: Enforcer                               |[11.0,0.0,146.0,138.0,0.5140845070422535]            |0      |0.0       |[0.8881167989465208,0.11188320105347918]|
|Men of War™                                   |[11.0,0.0,1664.0,257.0,0.8662155127537741]           |0      |0.0       |[0.8881167989465208,0.11188320105347918]|
|Battlestations Pacific                        |[54.0,132.0,1028.0,542.0,0.6547770700636942]         |0      |0.0       |[0.9501433309598681,0.04985666904013194]|
|Civilization IV: Beyond the Sword             |[11.0,108.0,1396.0,66.0,0.9548563611491108]          |0      |0.0       |[0.8881167989465208,0.11188320105347918]|
|Spear of Destiny                              |[5.0,0.0,200.0,36.0,0.847457627118644]               |0      |0.0       |[0.8881167989465208,0.11188320105347918]|
|Warhammer® 40,000: Dawn of War® - Winter Assault|[11.0,213.0,606.0,76.0,0.8885630498533724]         |0      |0.0       |[0.8963928900885774,0.10360710991142263]|
|Sherlock Holmes: The Mystery of the Mummy     |[9.0,0.0,59.0,130.0,0.31216931216931215]             |1      |0.0       |[0.8881167989465208,0.11188320105347918]|
|Sacred Gold                                   |[41.0,337.0,1489.0,319.0,0.8235619469026548]         |0      |0.0       |[0.8963928900885774,0.10360710991142263]|
|Rayman Raving Rabbids™                        |[17.0,29.0,239.0,80.0,0.7492163009404389]            |0      |0.0       |[0.8881167989465208,0.11188320105347918]|
|Helldorado                                    |[11.0,0.0,53.0,23.0,0.6973684210526315]              |0      |0.0       |[0.8881167989465208,0.11188320105347918]|
|Resident Evil™ 5/ Biohazard 5®                |[17.0,868.0,9250.0,1618.0,0.8511225616488775]        |0      |0.0       |[0.8725250899221109,0.1274749100778892] |
|Penumbra: Black Plague Gold Edition           |[721.0,188.0,1396.0,79.0,0.9464406779661017]         |1      |1.0       |[0.0,1.0]                                |
+----------------------------------------------+-----------------------------------------------------+-------+----------+-----------------------------------------+
only showing top 20 rows
```

8 Machine Learning with PySpark

```
Sample Predictions for Gradient Boosting:
2025-02-08 12:19:28,525 WARN scheduler.DAGScheduler: Broadcasting large task binary with size 1069.1 KiB
+--------------------------------------------+--------------------------------------------------------+-------+----------+----------------------------------------+
|name                                        |features                                                |is_free|prediction|probability                             |
+--------------------------------------------+--------------------------------------------------------+-------+----------+----------------------------------------+
|Ricochet                                    |[5.0,175.0,2758.0,684.0,0.8012783265543288]             |0      |0.0       |[0.9094170584733002,0.09058294152669977]|
|Counter-Strike: Global Offensive            |[94.0,22494.0,2644404.0,402313.0,0.8679519627192155]    |1      |1.0       |[0.06309043118724922,0.9369095688127508]|
|Disciples II: Rise of the Elves             |[11.0,0.0,451.0,108.0,0.8067978533094812]               |0      |0.0       |[0.9036369582845397,0.09636304171546028]|
|HeXen: Beyond Heretic                       |[5.0,50.0,249.0,26.0,0.9054545454545454]                |0      |0.0       |[0.8989339485586882,0.10106605144131175]|
|X2: The Threat                              |[11.0,732.0,212.0,50.0,0.8091603053435115]              |0      |0.0       |[0.9157152170391678,0.0842847829608322] |
|Peggle™ Nights                              |[8.0,237.0,368.0,24.0,0.9387755102040817]               |0      |0.0       |[0.925344461475079,0.07465553852492102] |
|RoboBlitz                                   |[0.0,78.0,43.0,14.0,0.7543859649122807]                 |0      |0.0       |[0.9173715201775755,0.0826284798224245] |
|STAR WARS™ Jedi Knight - Jedi Academy™      |[5.0,1254.0,5684.0,330.0,0.9451280345859661]            |0      |0.0       |[0.9186241079830941,0.08137589201690587]|
|X-COM: Enforcer                             |[11.0,0.0,146.0,138.0,0.5140845070422535]               |0      |0.0       |[0.9066235668406375,0.0933764331593625] |
|Men of War™                                 |[11.0,0.0,1664.0,257.0,0.8662155127537741]              |0      |0.0       |[0.9036369582845397,0.09636304171546028]|
|Battlestations Pacific                      |[54.0,132.0,1028.0,542.0,0.6547770700636942]            |0      |0.0       |[0.9262997013839114,0.07370029861608862]|
|Civilization IV: Beyond the Sword           |[11.0,108.0,1396.0,66.0,0.9548563611491108]             |0      |0.0       |[0.9109056159018526,0.0890943840981474] |
|Spear of Destiny                            |[5.0,0.0,200.0,36.0,0.847457627118644]                  |0      |0.0       |[0.9020400187462726,0.0979599812537274] |
|Warhammer® 40,000: Dawn of War® - Winter Assault|[11.0,213.0,606.0,76.0,0.8885630498533724]          |0      |0.0       |[0.9157152170391678,0.0842847829608322] |
|Sherlock Holmes: The Mystery of the Mummy   |[9.0,0.0,59.0,130.0,0.31216931216931215]                |1      |0.0       |[0.9050198933758331,0.0949801066241669] |
|Sacred Gold                                 |[41.0,337.0,1489.0,319.0,0.8235619469026548]            |0      |0.0       |[0.9229473664396484,0.07705263356035164]|
|Rayman Raving Rabbids™                      |[17.0,29.0,239.0,80.0,0.7492163009404389]               |0      |0.0       |[0.9154238840876152,0.08457611591238479]|
|Helldorado                                  |[11.0,0.0,53.0,23.0,0.6973684210526315]                 |0      |0.0       |[0.9005759420634809,0.0994240579365191] |
|Resident Evil™ 5/ Biohazard 5®              |[17.0,868.0,9250.0,1618.0,0.8511225616488775]           |0      |0.0       |[0.924785903589273,0.07521409641072696] |
|Penumbra: Black Plague Gold Edition         |[721.0,188.0,1396.0,79.0,0.9464406779661017]            |1      |1.0       |[0.018909164554028222,0.9810908354459718]|
+--------------------------------------------+--------------------------------------------------------+-------+----------+----------------------------------------+
only showing top 20 rows
```

9️⃣ Model Evaluation

1️⃣ Evaluation Metrics:

We assessed our models using three key performance metrics:

- ✅ Accuracy → Measures how often the model's predictions are correct.
- ✅ F1 Score → Balances precision and recall, especially useful for imbalanced data.
- ✅ AUROC→ Measures how well the model differentiates between Free & Paid games.

```
Logistic Regression Metrics:
  Accuracy: 0.8986447795380798
  F1 Score: 0.8614908883617144
  AUROC: 0.7596047573744176

Random Forest Metrics:
  Accuracy: 0.9009352929948463
  F1 Score: 0.8889809846102106
  AUROC: 0.7749713861742438

Gradient Boosting Metrics:
  Accuracy: 0.9438824203092193
  F1 Score: 0.943253175098822
  AUROC: 0.8526514053897424
ubuntu@dsbda-vm:/$
```

Conclusions

Based on our analysis, we can conclude the following insights regarding our hypothesis:

✅ 1. Genre Significantly Impacts Success 🎮

- Massively Multiplayer, RPG, and Action games have the highest success metrics.
- Strategy and Adventure games also perform well but with lower average playtime.
- Genre influences both sales and engagement, validating our hypothesis.

📌 Hypothesis Confirmed: Certain genres consistently lead to higher success, measured by sales, playtime, and user ratings.

Conclusions

✅ 2. Pricing Strategy Affects Game Performance 💰

- Free-to-play games dominate in success metrics, attracting high engagement.
- High-priced games also perform well, suggesting premium games can succeed if they offer value.
- Low-priced games have the lowest success, implying pricing strategy matters.

📌 Hypothesis Confirmed: Price tiers correlate with game success, proving that pricing strategy plays a crucial role.

Conclusions

✅ 3. User Reviews & Sentiment Have Mixed Impact ⭐

- Positive and negative ratings both correlate strongly with success (0.73 & 0.70).
- Sentiment score has almost no correlation (0.04) → Raw review count matters more than percentage positivity.
- Popular games receive both good and bad reviews, reinforcing the role of engagement.

📌 Hypothesis Partially Confirmed: While user ratings strongly influence success, sentiment score alone is not a reliable indicator.

Conclusions

✅ 4. Machine Learning Predictions Support Insights 🧠:

- Gradient Boosting (GBT) performed best (94.38% accuracy) in predicting whether a game is Free or Paid.
- Random Forest also performed well, proving that pricing can be predicted from game features.
- This suggests pricing and success patterns can be modeled using AI.

📌 Implication: Developers can use ML models to predict pricing strategies for new games.

Conclusions

🔍 Final Verdict on the Hypothesis :

✅ CONFIRMED: Genre, pricing strategy, and user ratings collectively impact game success on Steam.

- ✓ Genres strongly correlate with success.
- ✓ Pricing strategy influences engagement and revenue.
- ✓ User ratings matter, but sentiment analysis is less predictive.

📌 Data-driven insights from this project can help game developers optimize their pricing, genre selection, and marketing strategies.

THANK YOU