Final Project: Report

Spencer Whan (041039602), Edward Madaire (041034306), Nathan Mitchell (041046777),

Tim Callahan (041044836*)* & Harryson Belizaire (040971897)

23F_BUS0007_010 Strategic Business Intelligence

Professor: Swapnil Marutirao Kangralkar

Due Date: December 7th, 2023

Introduction:

The dataset we have chosen is data from Kaggle that relates to and was acquired from the popular digital video game distribution platform "Steam" developed by the Valve corporation. It includes public data that is displayed via steam pages related to users, games, and recommendations. This data will conveniently help to display how often users leave reviews, locate the top rated games, and find the choices for titles worth promotion and advertisements. This report will display the significance of simply analyzing public information to help with decision making.

For the simplicity of analyzing the data in Power BI, we took the 3 received CSV files and compiled them into a single XLSX. This means that "games.csv," "recommendations.cvs," and "users.csv" were all compiled into "Steam_DB.xlsx." No other changes were made to the data, this was done for the sake of simplicity in our personal analysis and calculations.

Description of the Dataset:

To make things less confusing, the following descriptions will be separated into groups based off of the original file formatting that was received.

games.csv:

- 'app_id' - Unique application ID.
- 'title' - Title of the game.
- 'date_release' - Release date of game, recorded in Year-Month-Day.
- 'win' - True if accessible on a Windows operating system, otherwise false.
- 'mac' - True if accessible on a Macintosh operating system, otherwise false.
- 'linux' - True if accessible on a Linux operating system, otherwise false.

- ‘rating’ - Positive/negative rating of a game, written as a string. There a total of 9 different ratings, stated in descending order:
    - Overwhelming Positive
    - Very Positive
    - Mostly Positive
    - Positive
    - Mixed
    - Negative
    - Mostly Negative
    - Very Negative
    - Overwhelmingly Negative
- ‘positive_ratio’ - Ratio of positive review status, number out of 100.
- ‘user_reviews’ - Number of reviews left on the game.
- ‘price_final’ - The current listed price of the game.
- ‘price_original’ - Original price on game release date.
- ‘discount’ - Discounted price due to the game being on sale.
- ‘steam_deck’ - True if steam deck compatible, otherwise false.

recommendations.csv:

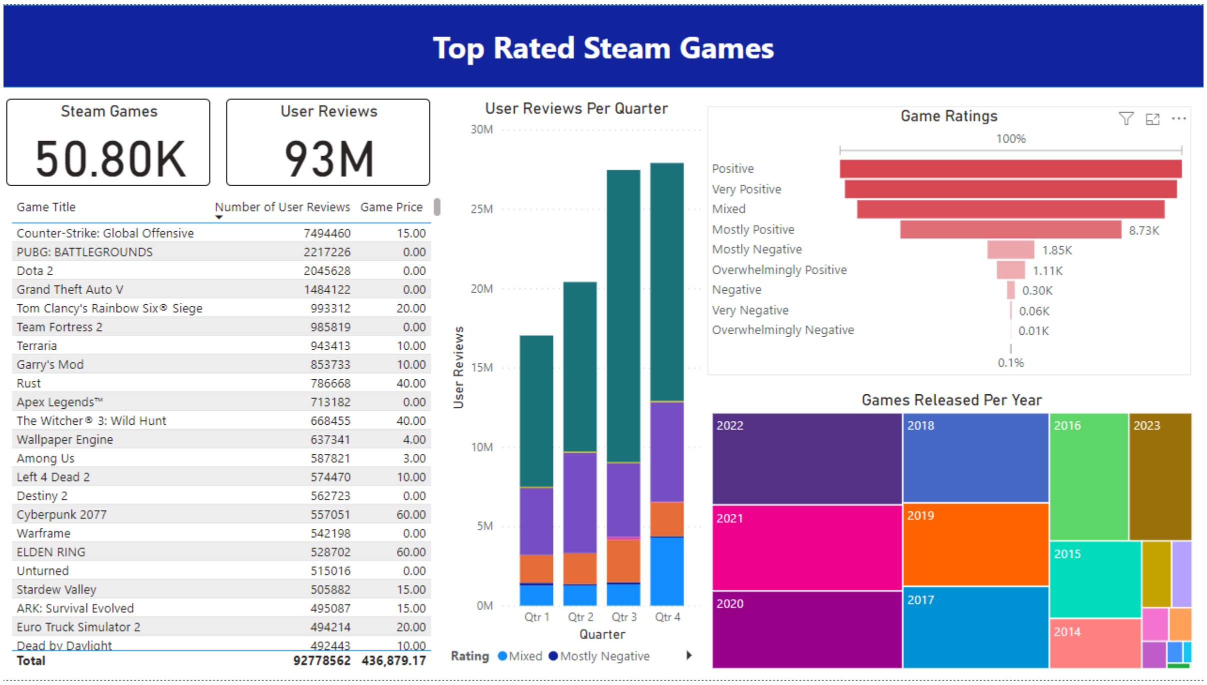
- ‘app_id’ - Unique application ID.
- ‘helpful’ - The amount of time the recommendation is marked as helpful.
- ‘funny’ - The amount of time the recommendation is marked as funny.
- ‘date’ - Date of recommendation, recorded in Year-Month-Day.
- ‘is_recommended’ - True if the review is marked as recommended, false if not.

- ‘hours’ - The amount of hours the user has prior to making the recommendation.
- ‘user_id’ - The unique user ID of the reviewer/creator.
- ‘review_id’ - The unique ID of the review.

users.csv:

- ‘user_id’ - Unique user ID.
- ‘products’ - Number of products owned by the user.
- ‘reviews’ - Number of reviews left by the user.

Type of Graphs and Observations:



Top Rated Steam Games

Steam Games

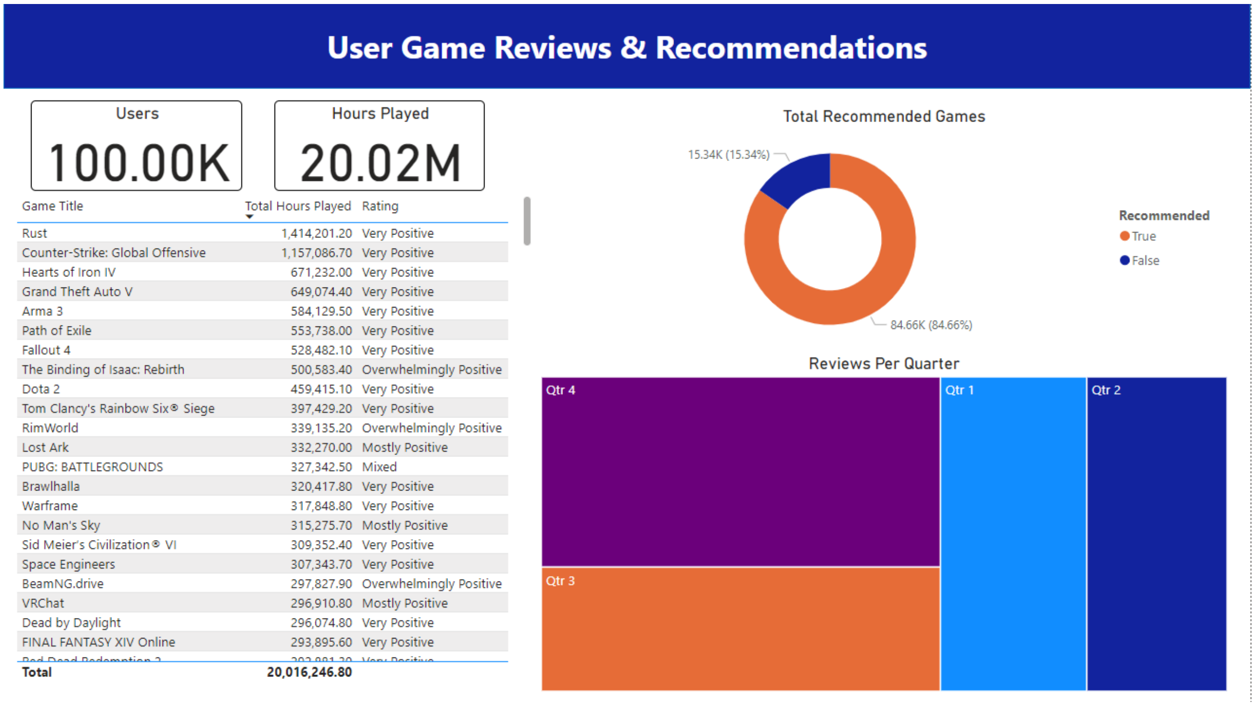
50.80K

User Reviews

93M

Game Title	Number of User Reviews	Game Price
Counter-Strike: Global Offensive	7494460	15.00
PUBG: BATTLEGROUNDS	2217226	0.00
Dota 2	2045628	0.00
Grand Theft Auto V	1484122	0.00
Tom Clancy's Rainbow Six® Siege	993312	20.00
Team Fortress 2	985819	0.00
Terraria	943413	10.00
Garry's Mod	853733	10.00
Rust	786668	40.00
Apex Legends™	713182	0.00
The Witcher® 3: Wild Hunt	668455	40.00
Wallpaper Engine	637341	4.00
Among Us	587821	3.00
Left 4 Dead 2	574470	10.00
Destiny 2	562723	0.00
Cyberpunk 2077	557051	60.00
Warframe	542198	0.00
ELDEN RING	528702	60.00
Unturned	515016	0.00
Stardew Valley	505882	15.00
ARK: Survival Evolved	495087	15.00
Euro Truck Simulator 2	494214	20.00
Dead by Daylight	492443	10.00
Total	**92778562**	**436,879.17**

Dashboard #1



User Game Reviews & Recommendations

Users

100.00K

Hours Played

20.02M

Game Title	Total Hours Played	Rating
Rust	1,414,201.20	Very Positive
Counter-Strike: Global Offensive	1,157,086.70	Very Positive
Hearts of Iron IV	671,232.00	Very Positive
Grand Theft Auto V	649,074.40	Very Positive
Arma 3	584,129.50	Very Positive
Path of Exile	553,738.00	Very Positive
Fallout 4	528,482.10	Very Positive
The Binding of Isaac: Rebirth	500,583.40	Overwhelmingly Positive
Dota 2	459,415.10	Very Positive
Tom Clancy's Rainbow Six® Siege	397,429.20	Very Positive
RimWorld	339,135.20	Overwhelmingly Positive
Lost Ark	332,270.00	Mostly Positive
PUBG: BATTLEGROUNDS	327,342.50	Mixed
Brawlhalla	320,417.80	Very Positive
Warframe	317,848.80	Very Positive
No Man's Sky	315,275.70	Mostly Positive
Sid Meier's Civilization® VI	309,352.40	Very Positive
Space Engineers	307,343.70	Very Positive
BeamNG.drive	297,827.90	Overwhelmingly Positive
VRChat	296,910.80	Mostly Positive
Dead by Daylight	296,074.80	Very Positive
FINAL FANTASY XIV Online	293,895.60	Very Positive
Red Dead Redemption 2	293,881.20	Very Positive
Total	**20,016,246.80**	

Dashboard #2

The dashboards utilize the following visuals:

- Table: Displays list of games with supporting data.
- Stacked column chart: Displays the user reviews per quarter.
- Funnel chart: Displays the game rating percentiles.
- Treemap: Displays games released per year and reviews/recommendations per quarter.
- Donut chart: Displays recommended game percentiles.
- Cards: Displays Steam games, user reviews, users, and hours played.

Observations:

While analyzing the data via the dashboards, we observed a few unique trends. Q3 and Q4 received higher amounts of reviews and recommendations compared to Q1 and Q2 , with Q4 receiving the most. Q4 takes place over the busy holiday season where users are open to purchasing and playing games more frequently - leading to a higher number of reviews/recommendations.

In general, Steam offers games with very good ratings. Of the 9 categories for game ratings (please refer to the "Description of the dataset" section), The top 4 ratings are Positive, Very Positive, Mixed, and Mostly Positive. In terms of marketing, games that receive ratings within these categories are promoted frequently to increase user play time (hours) and sales. Although the dataset is limited in terms of user data and does not include sales, utilizing the total play time of a game is a good way to validate a games popularity and quality (positive or negative).

Descriptive Statistics:

The dataset that we found was stated to be clean. After investigation and analysis, we determined that this statement was true.

The same as done previously, the following descriptive statistics will be separated into groups based off of the original file formatting that was received. Only descriptive statistics that appear significant or useful will be included in this portion of the report to avoid redundancy and useless information.

games.csv:

98.43% of games have Windows operating system accessibility.

25.6% of games have Macintosh operating system accessibility.

17.78% of games have Linux operating system accessibility.

100% of games are steam deck compatible.

- positive_ratio:

    - Mean: 77
    - Median: 81
    - Mode: 100
- user_reviews:
    - Mean: 1826
    - Median: 48
    - Mode: 10
- price_final:
    - Mean: 8.6
    - Median: 4.99
    - Mode: 0

recommendations.csv:

- helpful:
    - Mean: 3
- funny:
    - Mean: 1
- hours:
    - Mean: 200.16
    - Median: 99.2
    - Mode: 0.3

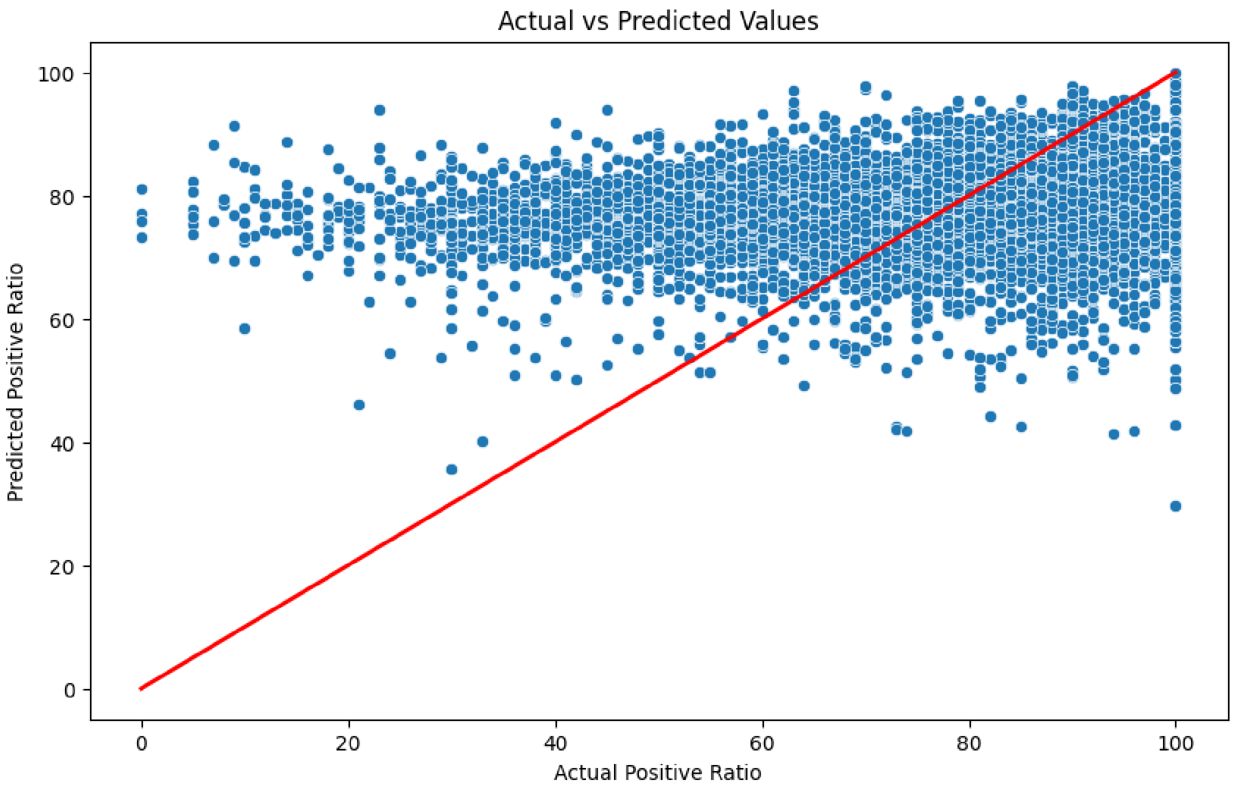
users.csv:

- products:
    - Mean: 122

  - Median: 58
  - Mode: 11
- reviews:
  - Mean: 3
  - Median: 1
  - Mode: 1

Predictive Statistics:

Using machine learning we trained a model to predict the positive ratio of reviews based on price, number of reviews and discount off of the price. To prepare the data to train the model involved reading the csv. file to a dataframe, following this all columns outside of the ones used in the prediction were dropped. After this the data was converted to indicator variables to allow for the model to read and learn from it. Then to finish the preparation of the data the data was separated into x and y axis for the model to differentiate between target variable and input features. A random forest regression model was selected as with the number of factors, the model allows various types of data well, has the capabilities to fit non-linear relationships and is effective with scaling to larger datasets. The method to evaluate the model was selected to be looking at the mean average error of the predicted positive ratio versus the actual positive ratio of each entry in the dataset. The results of the evaluation were a fairly large mean average error however, this is slightly to be expected as we are using publicly available information, due to the fact the majority of user data is protected behind privacy policies. This means that if the model were to be trained using first-party information we could much more accurately pinpoint the factors that cause people to give a game a good or bad rating and thus more simply predict ratings for games.



Conclusion and Recommendations:

Overall, the top categories of all the reviews lean mostly towards the positive side of the scale; therefore, implying that people are more likely to give positive reviews than negative reviews. We have concluded that people are more willing to review something they have enjoyed than something they have not. We also noticed that the gaming market is almost entirely dominated by the Windows operating system, as 98.43% of games listed on steam have Windows operating system accessibility. We recommend for more games to become compatible with other operating systems, as only a small fraction of games listed on steam are compatible with both Macintosh and/or Linux operating systems.

On the predictive aspect, it has proven difficult for us to come to accurate conclusions using predictive statistics based off of the data we have used. This is likely because the data we have obtained is entirely public information, meaning we do not have the significantly important and useful backend information that Valve would have to make more accurate predictions and therefore better decisions/recommendations.