Abstract

The video game industry has grown exponentially over the last decade. In 2020 alone the gaming industry generated $155 Billion in revenue [1] and it is projected that it will generate more than $260 Billion in revenue by 2025. One of the leading platforms which facilitates this industry is the Steam software by Valve corporation. This project aims to understand the relationship between the sales of video games on the Steam platform and the attributes which contribute to them. The analysis can give game developers and publishers insight into what makes certain games popular and allow them to make games that will have a larger possibility to stand out in the market.

Introduction

Steam is a digital distribution service for video games, primarily focused on the PC market. It is the largest distributor of games with over 75% of the market share. With major tech giants like Google, Amazon, and Meta entering the space and industry leaders like Microsoft consolidating their presence with larger acquisitions, the PC video gaming industry is booming. Millions are spent on developing a successful video game. Before companies put their resources into the development of a video game they need to have a strong vision of whether the video game will be a success. To predict whether the game will be successful or not is becoming complex to understand as it is governed by several factors like platform support, genre, paid/free, languages supported, etc.

Problem Statement

In order to survive in the playing field and create games that can be successful, it is important to understand the factors which affect the popularity and sale of the game. The question we set out to answer - "*Is it possible to predict how well the game will sell based on its characteristics?*". Another question the analysis aims to answer - "*What are the common characteristics that define the best-selling games?*"

Target Audience

Game developers and publishers are a strong group of consumers for our analysis. By understanding the factors that affect the sales of a video game, game developers and publishers would be able to find ways to stimulate customers to make a purchase. The analysis will help video game developers to manage and maintain quality and at the same time remain relevant in the market. It will help them examine new games to be produced in terms of user reviews, users' most preferred genre, top platforms preferred by gamers, etc.

Another set of target audience are the customers. The advent of digitalization has brought advantages and disadvantages for both developers and consumers; increased availability of

games leads to bigger markets and potentially more consumers, however at the same time consumers face the potential of becoming overwhelmed by the sheer number of available games. Our analysis can help customers to select games based on the most favoured genre of other players, top user reviews of a particular game, etc.

Previous Works

Previous works on video game sales such as ‘Predicting Global Video-Game Sales’ [2] estimate the game sales based on critics and user reviews. They consider these variables by looking at how successful a game is in terms of these reviews. However, they have not considered genre for their analysis. We consider genre to be an important factor for sales prediction as there is a correlation between the player’s interest towards a genre variety and the game industry’s revenue [3].

Works such as ‘Empirical Analysis on Sales of Video Games: A Data Mining Approach’ studies factors that make the sales of video games become a blockbuster. This study uses dataset of video game titles to estimate the effect of dependent variables to estimate the lifetime unit sales of games released in the US. The techniques used included the kNearest Neighbour (k-NN), Random Forest, and Decision Tree [4]. Another analysis ‘Video Game Sales Analysis — Visualization and Regression’ [5] predict possible sales using Scikit learn in Python.

The paper ‘Factors that Impact Video Game Sales’ explores trends in past video game sales by fitting a linear regression. It looks for what kind of variables had an impact on past global sales and if those influential variables change when the focus is shifted to sales of three specific regions [6].

Previous works on Video game sales such as ‘Analysis of Video Game Sales’ [7] estimate the game sales based on only two numerical features - average rating and years since release. The paper ‘The Impact Of Platform On Global Video Game Sales’ examines video game sales by platform in the global market [8]. However, these studies focus on limited features in their dataset such as platform or average rating but did not consider other attributes such as genre category, price, controller support etc. In our project, we analyze the different factors which contribute to making a game popular on steam.

Data Collection, Preparation, and Description

Web Scraping

Since our analysis is based on the games available on steam, we needed to collect the information available on games listed on the platform.

There are 3 main sources from which we have collected this information. Each of them is described below:

1. Steam

Steam is a digital marketplace for video games developed by Valve corporation. The provide game developers and publishers a means to digitally publish their software. It is the largest platform for PC gaming currently. Each game hosted on the platform has a unique app ID associated with it and the game webpage has all relevant information about the game available on it.

2. Steamspy

Steamspy is a website that collects information about games from steam and other external sources. The website also collects other statistics like the number of owners of a game, youtube statistics, average player base over the lifetime of the game, etc. The number of owners field is of key interest to us, as we want to analyze what factors of a game affect this parameter. However, Steam does not publish this data publicly, therefore most analysts and websites, including steamspy, who provide such information have their own methods to reach to a figure for number of owners for a game.

3. Metacritic

Metacritic is a website that aggregates reviews for films, TV shows, music albums, and video games from critics as well as users. We look only at the PC score for the games in our analysis. There are 2 scores presented for each game Metascore and Userscore. The metascore is generated from critic reviews, such as major publications in the gaming industry. The userscore is generated from user reviews directly. The metascore falls in a range of 0 to 100 while the userscore is in a range from 0 to 10.

There are 4 stages to collecting the data, the pipeline for these stages are given below:

1. Collect list of games and their Steam app IDs from steamspy
2. Collect game details from the steam store using the app IDs
3. Collect additional information for each game from steamspy using app ID
4. Scrape metacritic scores for games by generating the metacritic URL with the game name

We will now describe each stage in detail below.

1. Get List of Games

The first step in the data collection process is to get a list of games to scrape their data. We collect this data from steamspy.com. Using their APIs we can send a request to get the game name and their steam app ID. We generate a csv file with these 2 parameters. A sample set of data from this table is given below.

Table 2. List of games

appid	name
570	Dota 2
730	Counter-Strike: Global Offensive
578080	PUBG: BATTLEGROUNDS
1063730	New World
440	Team Fortress 2
304930	Unturned
271590	Grand Theft Auto V
550	Left 4 Dead 2
230410	Warframe
105600	Terraria
252490	Rust
1245620	ELDEN RING

2. Scrape Game Information from Steam

Once we have the list of steam app IDs we can use them to retrieve the game information from steam using their web REST API. We generate the URL and using a python script we store all the information returned by the API. Taking Dota 2 as an example we create the following URL - http://store.steampowered.com/api/appdetails/?appids=570

The response of this request is a dictionary with the following fields -
['type', 'name', 'steam_appid', 'required_age', 'is_free', 'controller_support', 'dlc', 'detailed_description', 'about_the_game', 'short_description', 'fullgame', 'supported_languages', 'header_image', 'website', 'pc_requirements', 'mac_requirements', 'linux_requirements', 'legal_notice', 'drm_notice', 'ext_user_account_notice', 'developers', 'publishers', 'demos', 'price_overview', 'packages', 'package_groups', 'platforms', 'metacritic', 'reviews', 'categories', 'genres', 'screenshots', 'movies', 'recommendations', 'achievements', 'release_date', 'support_info', 'background', 'content_descriptors','metacritic_user_score']

3. Scrape Additional Information from Steamspy

We collect further information about the games from steamspy. We do this by using the app IDs again and generating a steamspy URL. An example URL for steamspy scraping is as follows - https://steamspy.com/api.php?request=appdetails&appid=570

From the response to this request we extract the information from the following fields -

['score_rank', 'positive', 'negative', 'owners', 'average_forever', 'average_2weeks', 'median_forever', 'median_2weeks', 'price', 'initialprice', 'discount', 'ccu', 'youtube_stats']

4. Web Scrape Metacritic scores

We finally scrape the metacritic critic and user scores from metacritic.com. Unlike Steam and steamspy where we could directly retrieve the information using steam app ID. However, these IDs don't help us in the case of metacritic as they use the video game name as part of the URL. Hence, in order to get the maetacritic scores we format the name of the games by replacing spaces and special characters with hyphens. For example, In the case of Dota 2, the metacritic URL is as follows - https://www.metacritic.com/game/pc/dota-2

Based on the webpage returned we use beautifulsoup package in python to extract the review scores.

Data Cleaning

We followed the methods outlined below to clean the dataset and transform the variables-

1. Remove unnecessary columns

   The columns that we will not use in our analysis have been dropped. These columns, for example, have been removed for the following reasons:

   - score_rank, fullgame column are removed as they contain too many missing values
   - descriptions column are removed as they will not be used in our analysis.
   - price and discount columns are removed as they were effective only during a certain period of time when there were sales on video games.
   - ccu is removed because it shows concurrent players at one moment in time.

2. Handle missing observations

   Missing values are handled for the below columns -

   - platform: There are only a few records(66) that are missing the platform. We discovered that they are old games with low ratings, so they have been removed.
   - metacrtic_user_score: To limit the number of missing records, this attribute is scraped from the metacritic website and used to create a new feature custom_user_rating. It offers user ratings in a more comprehensive way. It's a mix of metacritic user ratings and steam user reviews. If metacritic user score isn't present, the custom user score will have an overall positive rating; otherwise, it will contain the mean of the overall positive rating and metacritic user score.
   - supported_languages : With the understanding that each game support at least one language, missing values are replaced with English.
   - Categories: This column is used to create the categories count column, which keeps track of how many categories the game fits into. With the understanding that a game would belong to at least one category, missing values are substituted with 1.

3. Datatype conversion and content parsing

Almost every column in our dataset require further processing before it can be used for analysis. We handled the columns in this part by extracting the relevant data and transforming it into the appropriate data types. We have also derived new columns based on existing columns.

- platform: It contains True or False values in string format for each of the three platforms namely Windows, MAC, and Linux. We took the data for each platform and divided it into three columns: Window platform, Mac platform, and Linux platform, each of which has a value of 1 if the game is supported by it and 0 if it is not.
- Owner: This column is made up of the lower and upper bound of an estimation for the number of owners for each title. This column can be modified in several different ways. We opted to calculate the mid-point and save it in the column Total_owners based on the analysis we'll be doing moving forward. As a result, we have 13 distinct values for this column.
- total_owner_cat(Derived Column) : A new column total_owner_cat is created by clustering Total_owners into 3 categories.
  10000 - 150000 owners in category 0
  350000 - 3.5 Million owners in category 1
  7.5 Million - 150 Million number of owners in category 2.
- initialprice: For instance, the initialprice in raw data is 2999, whereas the game's actual pricing is 29.99 USD. After dividing by 100, initialprice is converted to its right form.
- release_date: The game's release date is available in a variety of formats. A few examples are provided below :
  {'coming_soon': False, 'date': '9 Jul, 2013'}
  {'coming_soon': False, 'date': '4/nov./2020'}
  {'coming_soon': False, 'date': '2 févr. 2021'}
  Because the format isn't consistent, we've essentially retrieved the game's release year and placed it in a new column called release_year.
- overall_positive_rating(Derived Column): The number of positive ratings garnered by a game is listed in the positive column. Similarly, the total number of bad ratings a game has earned is listed as negative. The aggregate of both is added to a new column called total rating. The ratio of positive ratings to total ratings is a new feature overall_positive_rating.
- supported_languages: Observed that supported languages have NaN values. Replacing rows with NaN values with 1 under the assumption that the game should support at least 1 language.
- categories: Identified that the column categories have 74 missing values. If the categories column is null replacing with 1 else number of categories
- Controller_support: The controller_support column has either has value full or its blank. A full indicates that the game is fully supported by the controller, whereas a blank indicates that the game is not supported. Full is replaced with 1 and blank with 0.
- genre: Cleaning Genres column and One Hot encoding
- genre_count: genre_count store total number of genres supported by a game.

- custom_user_score: It is a new feature that offers user ratings in a more comprehensive way. It's a mix of metacritic user ratings and steam user reviews. If metacritic user score isn't present, the custom user score will have an overall positive rating; otherwise, it will contain the mean of the overall positive rating and metacritic user score.

4. Handle duplicate games

   All AppIDs should be unique, and any rows with the same ID should be treated as duplicates. As a result, the duplicate steam id rows need to be removed.

5. Filter records based on condition

   Since we're trying to figure out what factors lead to a popular game/larger number of owners, we've excluded games with less than 100 ratings, reasoning that a popular game will have a rating of more than 100.

Data Dictionary

Data dictionary for raw data is added in Appendix A.

Exploratory Data Analysis

We performed EDA primarily to see what data can reveal beyond the formal modeling or hypothesis testing task and provide a better understanding of data set variables and the relationships between them [9]. We have used the SweetViz library, an open-source Python library that generates beautiful, high-density visualizations to kickstart EDA with just two lines of code [10].

Variable Summary and Status

At the top of the report, we see summary information on the data set and the variables (columns) we fed into it. In this case, we passed in a data set with 47 columns after performing data cleaning. Of those, 14 were treated as Numeric, 32 as Categorical, and 1 text. The summary information also tells us that there were 8729 observations in the data (this means rows in the dataset).

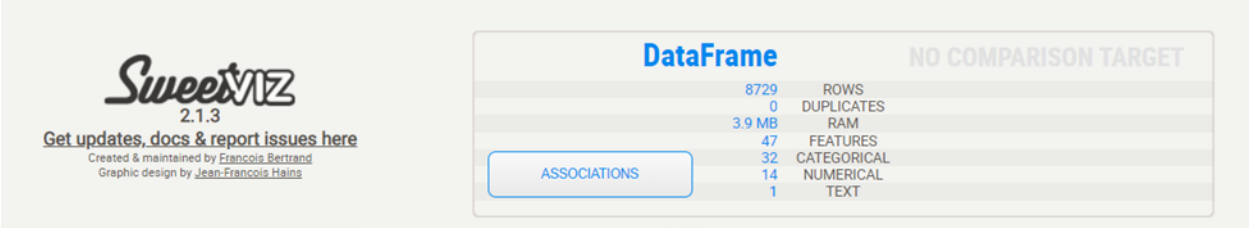

Figure 1. SweetVIZ Summary Information

Variable Analysis

As we can see in Figure 2, almost 98% of games have an age requirement as value 0. We assume age 0 as games having no requirement. There are few games that have age requirement of above 18 - less than 1%.

More than 80% of games are free and only about 20% of games are paid out of the game data we have collected.

For games that require controller support, we see about 75% of games do need it.

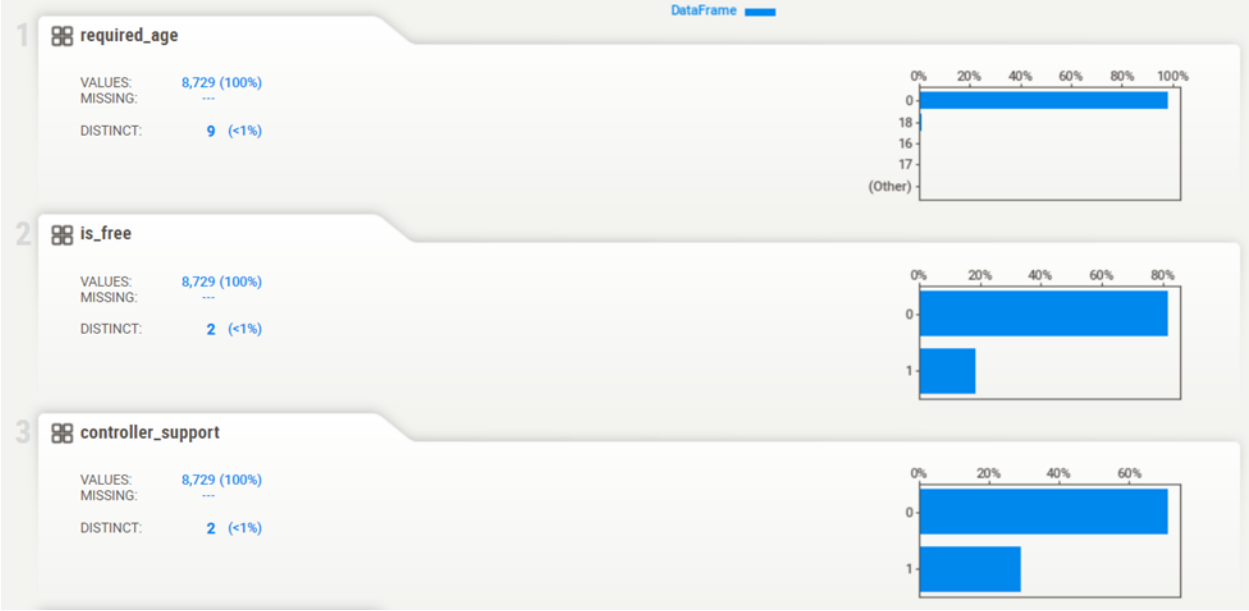

Figure 2. Summary Statistics - required_age, is_free, controller_support

We have three categorical columns- Window_platform, Mac_platform and Linux_platform. Figure 3. shows a plot of the most widely used platform among these three. We can clearly see windows is most used by gamers followed by mac and Linux platform is least used among all three platforms.

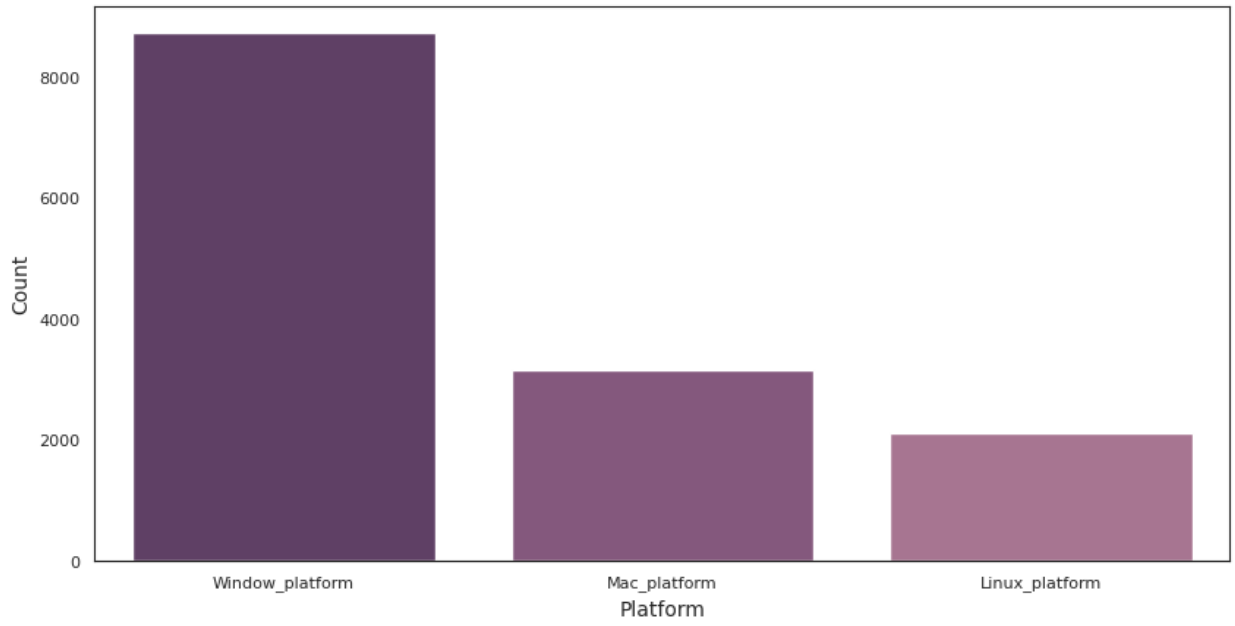

Figure 3. Summary Statistics - Game count by platform

Below we have a graph of top selling games by ratings. Counter-strike is the top game having the highest ratings. The other top selling games are PUBG and DOTA 2.

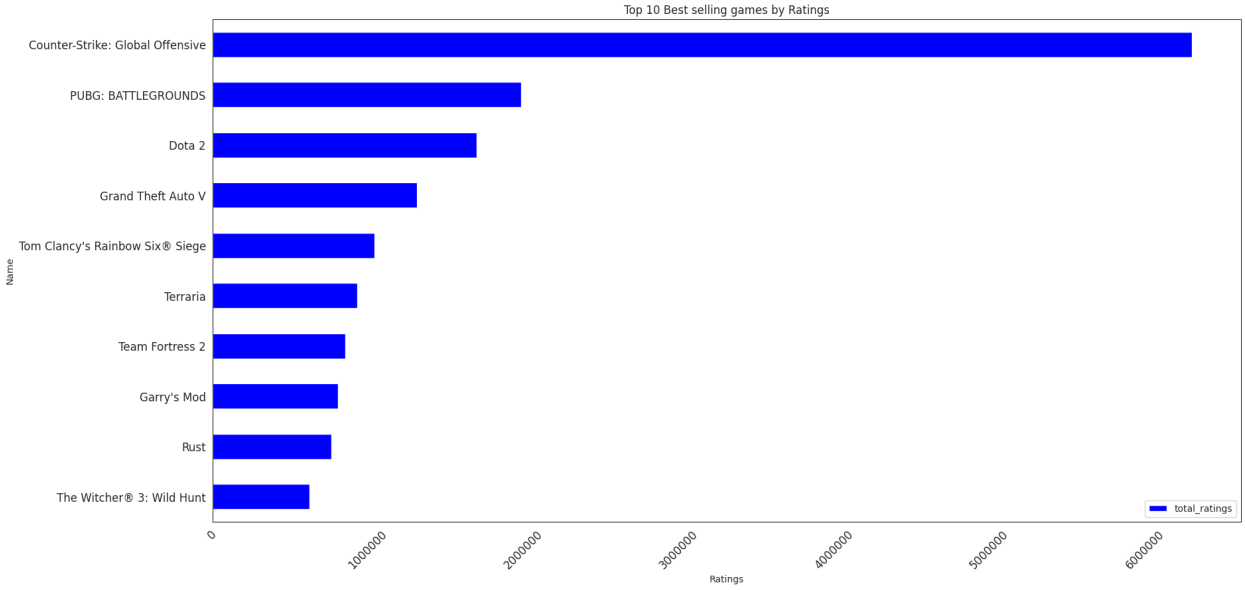

Figure 4. Summary Statistics - Top games by Ratings

Another plot of our categorical features is the Genre feature. We have total of 13 genres in our dataset. Below we have a barplot of genres the game belongs to and their total counts. The most common genre is the indie genre indicating they are games made by independent or

smaller development teams without the financial support of a larger publisher. Another common genre is the action and adventure genre among games.

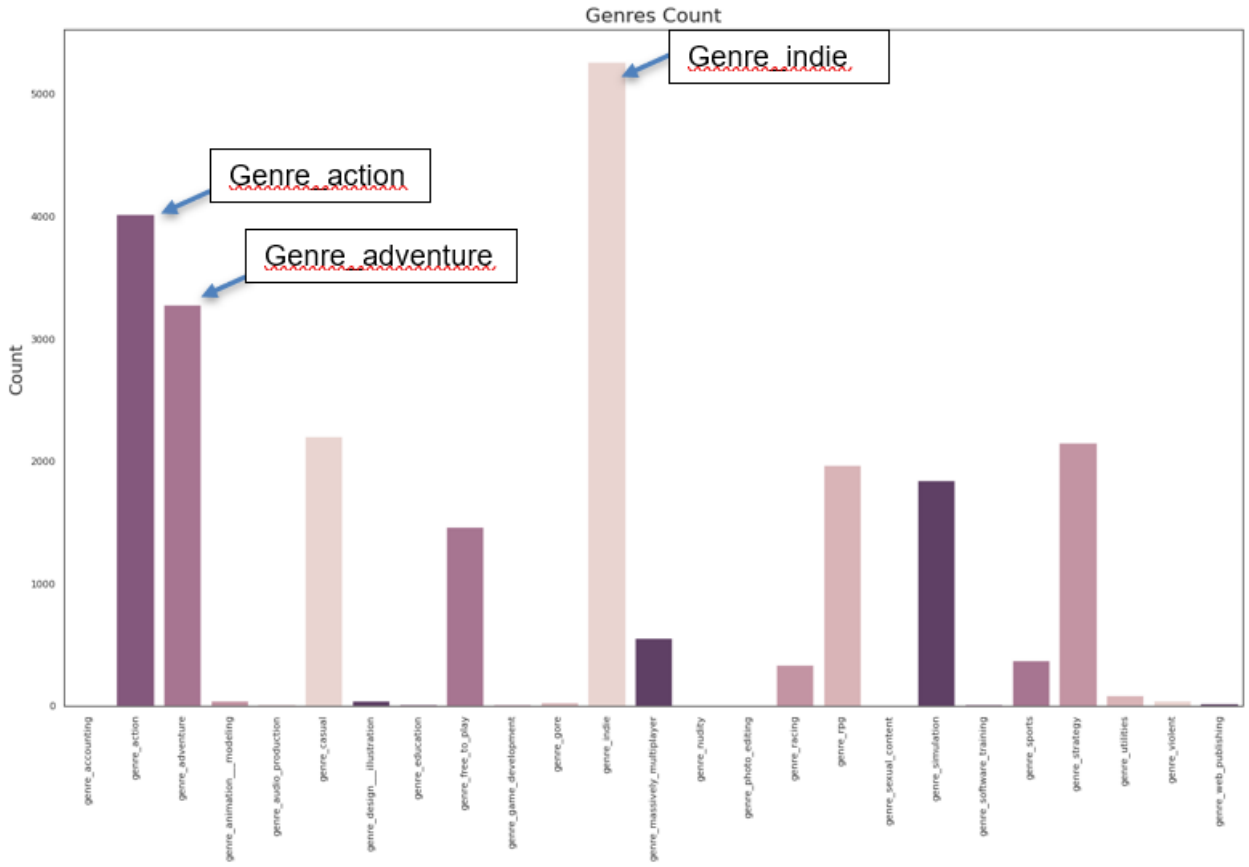

Figure 5. Summary Statistics - Genre by count

Analysis and Results

K-means Clustering

Total Owner Clusters

We have used K-means clustering to cluster our feature *total_owners*. k-means clustering tries to group similar kinds of items in form of clusters. It finds the similarity between the items and groups them into these clusters [11]. It starts with random centroids across the feature space and iteratively calculates the distance of the other feature points from the centroid, and groups the points that are closer to the centroid.

Elbow Method

Elbow is one of the most famous methods by which we can select the right value of k and boost model performance. Hence, according to the below elbow method, we can see the optimal number of clusters of 3 for our total_owners feature.

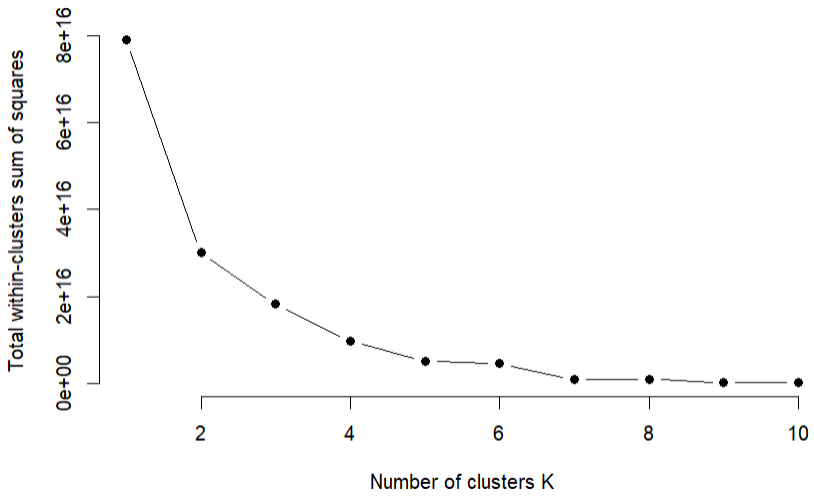

Figure 6. K-Means Clustering for k=1 to k=10 to create cluster of Total_Owners

The 3 clusters can be labeled as the following -

High level of ownership - High level of ownership would be the gamers who play games on steam platform frequently. Some players play games almost every day and are likely to explore different games available. So they would likely purchase games and their ownership would be a lot higher.

Average level of ownership - The players who enjoy playing games but are not everyday players fall into this category. They are the ones who like playing and trying out different games hence they own the game but do not go to the extent of owning a large number of games.

Low level of ownership - The players who play games less frequently will fall into this category. These are the players who play only in their leisure time and are not keen on owning or purchasing games that much.

Linear Regression

In this study, we are attempting to identify the elements that influence game sales. The game's owners are hence our dependent variable. SteamSpy is unable to obtain accurate data from Steam due to privacy concerns, therefore we can only get the owners' lower and upper bounds. We used the midpoint of these bounds for our analysis. However, because the range is fairly large, this resulted in only 13 distinct values. So we were hesitant to perform the linear regression, and after examining the assumptions, we discovered that we couldn't do so with the dependent variable owner.
As a result, we chose to perform an ordinal logistic regression with three levels of ownership: low, average, and high, as described in the next section.

The following is the result of linear regression, and because the first assumption, Mean-Zero Error, is violated, we cannot use linear regression to forecast the outcome because it is fatally severe.

Summary of linear regression

We used multivariate linear regression to begin our analysis. We fitted the model with the dependent variable Total_owners.

```
Call:
lm(formula = Total_owners ~ ., data = game_regression)

Residuals:
      Min        1Q    Median        3Q       Max
-79480314   -447896   -135256    183107 106519158

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.381e+06  1.265e+06  -1.092  0.27497
lang               4.187e+04  4.615e+03   9.073  < 2e-16 ***
anime              2.571e+04  4.743e+05   0.054  0.95678
design            -8.487e+04  4.832e+05  -0.176  0.86056
game_dev          -2.060e+05  7.493e+05  -0.275  0.78339
multiP             2.115e+05  1.312e+05   1.611  0.10711
sexual            -3.260e+05  7.910e+05  -0.412  0.68020
sports            -1.607e+05  1.344e+05  -1.195  0.23199
violent           -2.723e+05  4.298e+05  -0.634  0.52636
action             1.307e+05  7.550e+04   1.731  0.08353 .
race              -3.878e+05  1.374e+05  -2.823  0.00476 **
adventure         -3.730e+04  8.203e+04  -0.455  0.64935
casual            -2.051e+05  7.760e+04  -2.643  0.00822 **
free               1.792e+05  1.259e+05   1.423  0.15468
mac               -1.527e+04  6.432e+04  -0.237  0.81234
win                3.519e+05  1.255e+06   0.280  0.77925
linux              1.664e+05  7.134e+04   2.333  0.01969 *
cat_count          6.612e+04  8.546e+03   7.737 1.13e-14 ***
genre_count        1.016e+05  5.724e+04   1.775  0.07587 .
user_age           5.637e+04  1.034e+04   5.451 5.13e-08 ***
age                7.322e+04  7.981e+03   9.175  < 2e-16 ***
custom_user_score  4.860e+04  1.688e+04   2.880  0.00399 **
is_free            3.178e+05  1.057e+05   3.007  0.00265 **
initialprice       1.230e+03  2.484e+03   0.495  0.62063
controller_support -1.957e+05  6.070e+04  -3.224  0.00127 **
indie              -3.400e+05  7.973e+04  -4.265 2.02e-05 ***
rating             2.407e+01  2.847e-01  84.539  < 2e-16 ***
web_publish        -3.463e+05  5.760e+05  -0.601  0.54772
strategy           -1.909e+05  8.475e+04  -2.252  0.02434 *
simulate           -2.795e+04  8.189e+04  -0.341  0.73288
movie              -1.652e+06  2.179e+06  -0.758  0.44819
gore               -4.709e+04  5.255e+05  -0.090  0.92859
edu                -3.314e+05  7.302e+05  -0.454  0.64996
utilities          -9.573e+04  2.935e+05  -0.326  0.74431
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2167000 on 8694 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4999,    Adjusted R-squared:  0.498
F-statistic: 263.3 on 33 and 8694 DF,  p-value: < 2.2e-16
```

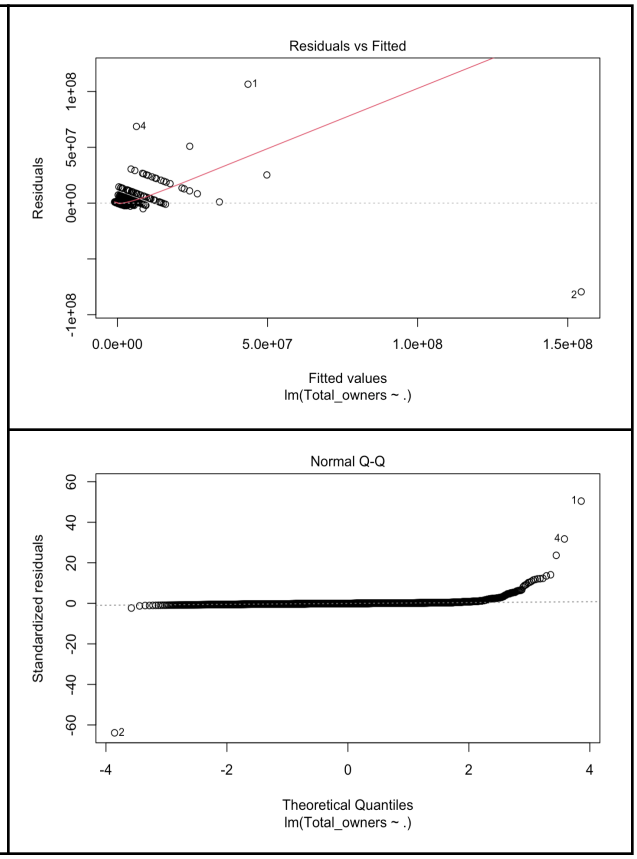

Figure 7. Multivariate linear Analysis result and assumptions check

Despite the fact that the linear regression model has a good adjusted R square value (0.4999) and F-statistic (263.3), we cannot draw any conclusions from it because the linear regression model must meet the following assumption criteria:

Table 2. OLS Model Assumptions

Assumption	Description	How to check	Consequences	Severeness
Mean-Zero Error	Conditional expectation of errors on independent variables is zero	Residual Plot	Biased and inconsistent estimates	Fatal
Uncorrelated Error	Residuals should not be autocorrelated	Residual Plot	Bloated or deflated standard error	Severe
No Perfect Multicollinearity	There should be a minimal linear relationship between the independent variables	VIF Test	Bloated standard error	Fixable
Homoskedasticity	The variance of the residuals should be constant	Residual Plot	Incorrect Standard Error	Fixable
Normal Error	Residuals are normally distributed	Q-Q Plot	Incorrect sampling distribution	Fixable

Looking at the residuals plot, we can observe that the data points do not symmetrically scatter about the reference line y=0, which is highly alarming. In reality, all of the data points are stacked above the reference line. This plainly demonstrates that mean-zero error has been violated, and biased error has resulted. Multi-variate linear regression cannot be used to make any predictions. The residuals do not fall on the 45-degree reference line, indicating that the normality assumption is also violated, as shown by the Q-Q plot

Ordinal Logistic Regression

As described in our previous section, we have done our analysis using Ordinal Logistic Regression with dependent variable total owners category which is categorical and ordered. It is clustered into 3 categories with the given range and with order 0<1<2. Here 0, 1, and 2 depict the low, average, and high levels of ownership. The purpose of the analyses is to discover which variable(s) has the most effect on the total owner's category.

Table 3. Distribution of Classes

Category	Level of ownership	Range (No of owners)	Number of games
0	Low	10000 - 150000	5041
1	Average	350000 - 3.5 Million	3543
2	High	7.5 Million - 150 Million	145

Below is the predictor variables that are selected to conduct the analyses: Genre animation, Genre design, Genre game development, Genre sexual, Genre sport, Genre violent, Genre Race, Genre Adventure, Windows Platform: Runs on Windows OS, Initial Price: Price when the video game was released, Controller support

The hypothesis of the ordinal regression is as follows -

1. Null Hypothesis H0: There are no statistically significant factors between variables that influence the total owner's category
2. Alternate Hypothesis H1: There is at least one statistically significant factor between the variables that influence the total owner's category

We fitted the proportional odds logistic regression model and calculated the p-value. The highlighted in red are the variables that were found to be significant in our model.

	Value	Std. Error	t value	p value
anime	0.05929988	4.507653e-01	1.315538e-01	8.953372e-01
design	-0.68132912	4.531894e-01	-1.503409e+00	1.327336e-01
game_dev	0.19611428	6.959495e-01	2.817938e-01	7.781016e-01
sexual	-13.18665650	1.268156e-07	-1.039829e+08	0.000000e+00
sports	0.12141636	1.137060e-01	1.067810e+00	2.856063e-01
violent	0.11844106	3.265894e-01	3.626605e-01	7.168585e-01
race	-0.23658681	1.200786e-01	-1.970266e+00	4.880791e-02
adventure	-0.17875120	4.600623e-02	-3.885369e+00	1.021745e-04
win	12.93346844	1.784139e-02	7.249136e+02	0.000000e+00
initialprice	0.02339494	1.876210e-03	1.246925e+01	1.098502e-35
controller_support	0.26772937	4.972953e-02	5.383711e+00	7.296578e-08
0|1	13.51571004	1.784139e-02	7.575479e+02	0.000000e+00
1|2	17.34711530	8.518575e-02	2.036387e+02	0.000000e+00

Figure 8. Result for Ordinal Regression

From this model, The significant categorical variables like the Windows platform can be interpreted as any video game that runs on the Windows platform is associated with a higher likelihood of having more owners as opposed to non-windows platforms.
Similarly, the significant continuous variable initial price can be interpreted as: with one unit increase in initial price the log of odds of having a higher number of users increases by 0.02 keeping all other variables const.
Since there is at least one variable that is statistically significant, the null hypothesis (H0) is rejected and the alternative hypothesis (H1) is accepted.

Model Justification

Since the Ordinal Logistic Regression model has been fitted, we need to check the assumptions to ensure that it is a valid model. The assumptions of the Ordinal Logistic Regression are as followed

1. The dependent variable is ordered.
   We have selected the number of owners as our dependent variable. It has been categorized into 3 levels namely 0 (10000 to 75000), 1(150000 to 750000), and 2(1500000 to 150000000) with 0<1<2.
2. One or more independent variables are either continuous, categorical, or ordinal
   This assumption is satisfied as our independent variable is a combination of both continuous and categorical variables.
3. No multicollinearity
   We have not tested for multicollinearity as all independent variables are categorical except the initial price.
4. Proportionality odds
   This assumption essentially means that the relationship between each pair of outcome groups has to be the same. If the relationship between all pairs of groups is the same, then there is only one set of coefficients, which means that there is only one model. If this assumption is violated, different models are needed to describe the relationship between each pair of outcome groups.

We used the brant test to evaluate this assumption, and we discovered that the parallel assumption holds true because the p-values for all variables are greater than alpha (0.05). There is also an Omnibus variable in the result, which effectively stands for the entire model and is greater than 0.05. As a result, the proportional odds assumption is upheld, and the model is valid for this dataset.

Test for	X2	df	probability
Omnibus	3.56	11	0.98
anime	0.06	1	0.81
design	0.36	1	0.55
game_dev	0	1	0.99
sexual	0	1	1
sports	0.14	1	0.7
violent	0	1	0.98
race	2.07	1	0.15
adventure	0.01	1	0.93
win	0	1	1
initialprice	0.5	1	0.48
controller_support	0.01	1	0.92

H0: Parallel Regression Assumption holds

Figure 9. Brant Test for Proportionality Odds Assumption

Our data has been divided into two categories: train and test dataset. The train dataset accounts for 80% of the data, while the test dataset accounts for 20%. The test data comprises roughly 1697 observations, while the train data has 7032. We created our model on the train dataset, using the same independent variables that met all of the proportional odds logistic regression assumptions.

To begin, the train dataset is used to make the prediction, and a confusion matrix is generated. For train data, we calculated the misclassification error, which was around 39%. Because our dataset is extremely imbalanced between these three categories, we have more misclassification error data at a high level of ownership. This is understandable because there are a few games that are quite popular. The majority of the games have a low or average degree of ownership.

A similar result can be seen with the test dataset, which has a 40% misclassification error. We may conclude that the model's behavior is consistent because the misclassification error is roughly the same in both the training and test datasets.

We also plotted a confusion matrix to evaluate the performance of the model and calculated precision, recall, and f1 score. The classifier is able to classify classes 0 and 1 but not class 2. This can be explained due to the class imbalance that is evident in the dataset. Compared to classes 0 and 1, video games in class 2 (highly popular) are underrepresented.

```
[1] " ************ Confusion Matrix ************"
    0   1  2
0 985 589 25
1  79 116  8
2   0   0  0
```

Figure 10. Confusion Matrix for test dataset

	precision <dbl>	recall <dbl>	f1 <dbl>
0	0.9257519	0.6160100	0.7397672
1	0.1645390	0.5714286	0.2555066
2	0.0000000	NaN	NaN

3 rows

Figure 11. Precision, Recall, and F1 score for test dataset

Conclusion

This project aims to identify the factors influencing the popularity of video games utilizing features that were collected from the Steam webpage. We performed K means clustering and ordinal regression and classified the video games into 3 categories - low(0), medium(1) and high(2) based on the number of owners. From the above experiments, we can conclude that there is a correlation between features such as adventure genre, racing genre, sexual genre, the initial price of the video games, controller support, windows platform support, and the popularity of the video game. This information can be leveraged for game development reference to select appropriate genres and operating system support to make the games more popular. We have limited our study to those games available on Steam. Future studies could consider utilizing games from multiple platforms to analyze whether the results are consistent or not.

Reference

[1] Investopedia - How the Video Game Industry Is Changing
[2] Predicting Global Video-Game Sales
[3] An Introduction to Videogame Genre Theory. Understanding Videogame Genre Framework
[4] Empirical Analysis on Sales of Video Games: A Data Mining Approach
[5]Video Game Sales Analysis – Visualization and Regression – CriticallyCoding
[6]Factors that Impact Video Game Sales
[7] Analysis of Video Game Sales. Using data from VgChartz.com to predict… | by Jeremy Chow | Modeling (The Data Kind) | Medium
[8](PDF) The Impact Of Platform On Global Video Game Sales
[9]https://www.ibm.com/sg-en/cloud/learn/exploratory-data-analysis#:~:text=EDA%20is%20primarily%20used%20to,for%20data%20analysis%20are%20appropriate
[10] SweetViz Library - EDA in Seconds
[11]https://www.analyticsvidhya.com/blog/2020/10/a-simple-explanation-of-k-means-clustering/#:~:text=k%2Dmeans%20clustering%20tries%20to,groups%20them%20into%20the%20clusters.

Appendix A

The steam_app_data.csv data set contains data on the video game. The set includes 10000 observations and 53 attributes. Each of the columns is defined as follows:

Table 4. Data Dictionary

Features	Description
steam_appid	Unique identifier for each title
is_free	whether the game is free or not
metacritic	critic score retrieved from Metacritic website
controller_ssupport	Values are either Full or blank. Full Controller support means the whole game can be accessed with a controller, this includes all menus.
dlc	DLC means "downloadable content," and refers to features in video games that are downloaded separately from the main game
required_age	Minimum required age according to PEGI UK standards.
detailed_description	description of game
developers	Name (or names) of developer(s). Semicolon delimited if multiple
publishers	Name (or names) of publisher(s). Semicolon delimited if multiple
platforms	Semicolon delimited the list of supported platforms. At most includes: windows; mac; linux
reviews	Textual reviews of each game
categories	Semicolon delimited list of game categories, e.g. single-player;multi-player
genres	Semicolon delimited list of game genres, e.g. action; adventure
recommendations	Total number of recommendations
achievements	Total achievement with highlighted one
release_date	Release date in various formats
metacritic_user_score	user score retrieved from Metacritic website

positive	Total number of positive rating
negative	Total number of negative rating
owners	This column is made up of the lower and upper bound of an estimation for the number of owners for each title. For privacy reasons, SteamSpy can't get exact figures from Steam
average_forever	The total time the user has run this app since adding it to their library. Values are given in minutes.
average_2weeks	The total time the user has run this app in the two-week period leading up to when this data was requested from the API. Values are given in minutes.
median_forever	Median number of players throughout the lifespan of the game
median_2weeks	Median number of players last 2 weeks of the game
price	The current price of the "app" on the Steam storefront, is in US dollars. Free items have a price of 0.
initialprice	The initial price of the game when it was launched.
discount	The current discount on the game at the time of scraping
ccu	concurrent users at the same moment in time
youtube_stats	It contains the total number of views and comments of the top 50 videos uploaded in last week at the time of scraping