Steam Store Games Exploratory Data Analysis

Hazar Belge
Galatasaray University, Computer Engineering
Istanbul, Turkey
Introduction to Data Analysis

I. Introduction

Steam is a video game digital distribution service by Valve. It was launched as a standalone software client in 2003 as a way for Valve to provide automatic updates for their games and expanded to include games from third-party publishers. Steam has also expanded into an online web-based and mobile digital storefront.

In this paper, we will look into a dataset about different video games with their different attributes like price, owner count, playtime, achievements count etc.

After describing the parameters and data in a comprehensive manner, we will ask a high-level research question and a low-level technical questions. Finally, we will have a description of the methods we intend to use to answer these questions.

II. Steam Dataset

This dataset is published by Nik Davis on Kaggle. It includes 27075 different video games and 18 different attributes for each of them from Steam. The dataset was last updated in May 2019, so I removed the games that were released in 2019 to prevent them from misleading the year-to-year comparison. After this removing process we have 24862 video games in total. So, let's take a look at some of attributes of dataset.

A. Achievements

Achievements are designed to increase a player's enjoyment of the video game and elicit a sense of satisfaction. When we think about it in this context, can we say that every developer company who tries to be successful has put achievements in its game? Let's see what the situation is.

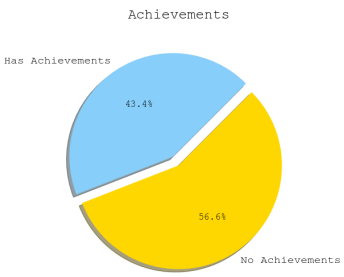

Fig. 1. How many games have achievements?

As we can see from Fig. 1., although the majority of video games prefer to put achievements in their games, the rate of those who do not (43.4%) is not low.

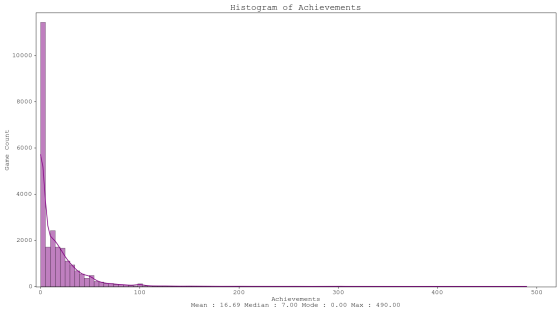

Fig. 2. Histogram of Achievements.

When we look at Fig. 2., we can easily see that video games do not want to increase the number of achievements. As the number of achievements decreases, the number of games belonging to that achievement number increases.

B. Age Rating

Age Rating is a somewhat complex/indirect feature. Video games do not directly choose their age rating. The combination of genre, subject and content of the video games that they created makes that age restriction necessary. That's why developers have 3 options. Reducing adult content and lowering or resetting the age limit. Not reducing adult content and ignoring the high age limit. Those who try to find a way by lowering the adult content and age restriction by a certain degree.

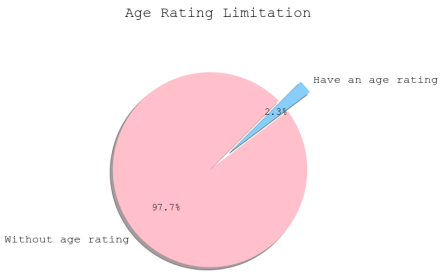

Fig. 3. How many games have age rating?

We can clearly see that the vast majority did not choose to use age-restricted content in their games (Fig. 3.).

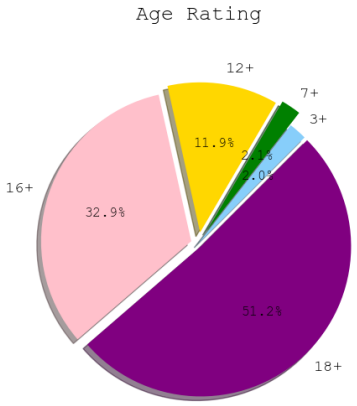

Fig. 4. Age Ratings Pie Chart.

However, we can say that those who choose to use adult content prefer to have more or high-level content and do not care about the age limit (Fig. 4.).

C. Category

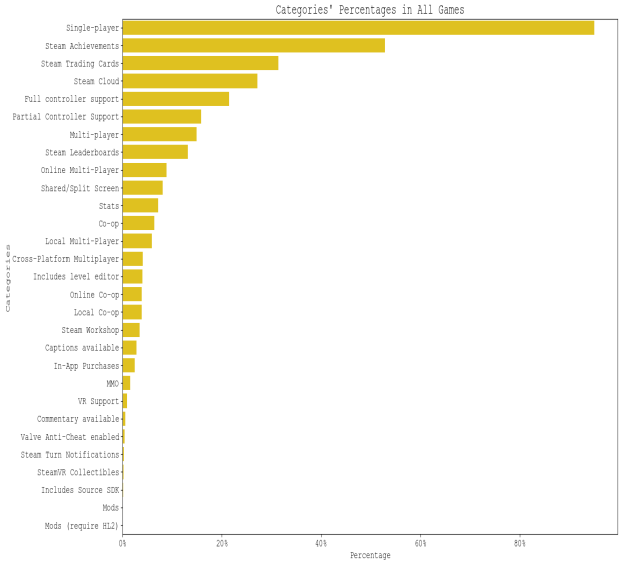

Fig. 5. Categories' percentages of all games.

Fig.5. shows us that the majority of video games are in the single-player category. While multi-player games are in the minority, the single-player category is followed by games with achievements and games where you can win and trade cards.

D. English Support

Supporting the English language, which is the most spoken language in the world, also plays a big role in increasing the number of people you can reach.

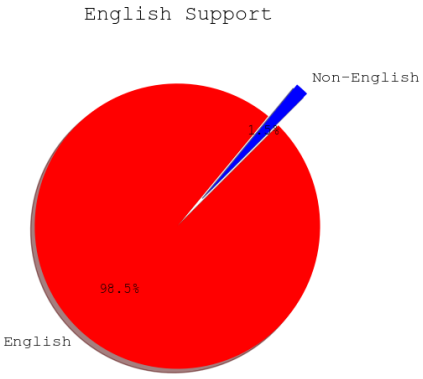

Fig. 6. How many games have English support?

Fig. 6. shows us that 98.5% of 24862 video games, or 24489, support the English language.

E. Genre

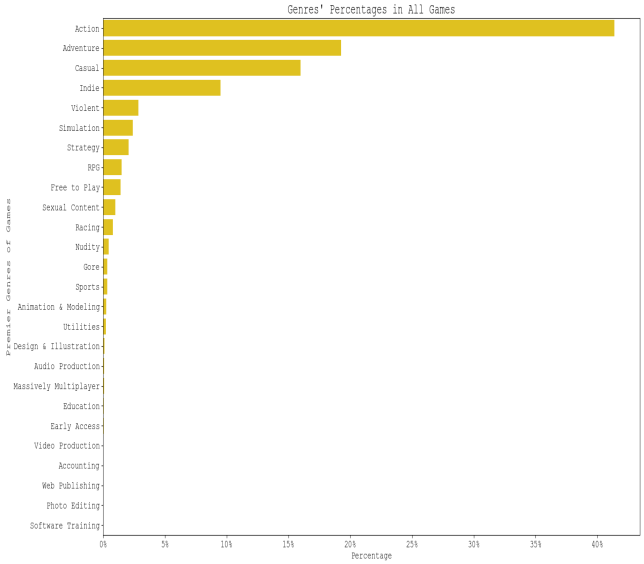

Fig. 7. Genres' percentages of all games.

Fig. 7. shows us that the majority of video games are in the action genre. The action genre is followed by adventure, casual, indie and violent games, respectively.

F. Owner Count

Owner Count is the main attribute of the steam dataset. Naturally, we can easily predict that as the owner count increases, the number of video games belonging to that count will decrease. In this exploratory data analysis, we examine why video games that reach 2M+ owners and 1M-2M owners sell this much.

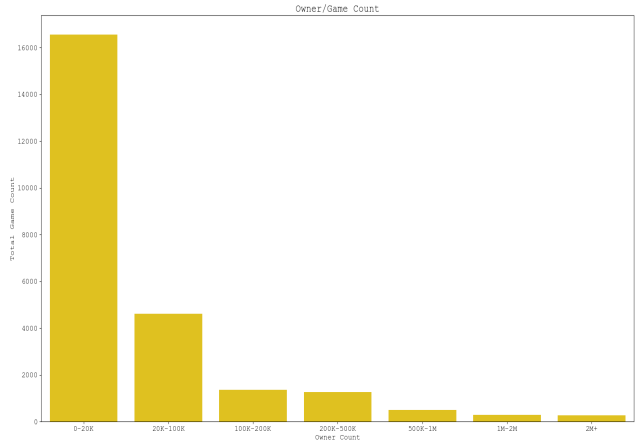

Fig. 8. Distribution of owners count in all games.

G. Platform

The platform on which the video game will be played is very important. While it is known that Windows-supporting machines are the majority today, developing games only for Mac or Linux is a project that will fail as a result. In addition, developing games only for Windows and leaving Mac or Linux users out will cause loss of users up to a certain level, although not as much as the opposite.

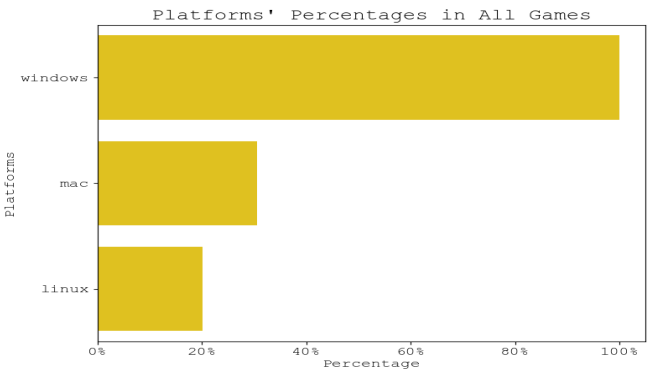

Fig. 9. Platforms' percentages of all games.

As we can easily see from Figure 9, almost 100% of 24862 video games support the Windows platform (99.97%). The Windows platform is followed by Mac with 30.59% and Linux with 20.07%.

H. Positive Rate

The positive rate attribute was synthesized from the existing positive_reviews and negative_reviews columns in the dataset. It is a value between 0 and 1 obtained by dividing the number of positive reviews by the total number of reviews.

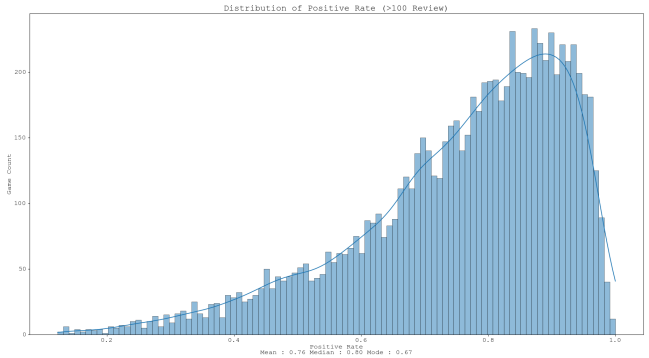

Fig. 10. Histogram of Positive Rate.

The positive rate distribution is concentrated around 0.8, although it does not appear on the histogram due to the fact that bins=100 and a value between 0 and 1, its mode is equal to 0.6667, mean is 0.7583 and median is 0.7959.

I. Price

Perhaps the most important factor that comes to mind when it comes to game sales is price. While some developers try to make the game free to reach more people and generate income with microtransactions, other companies want to remove the costs directly from the sale.

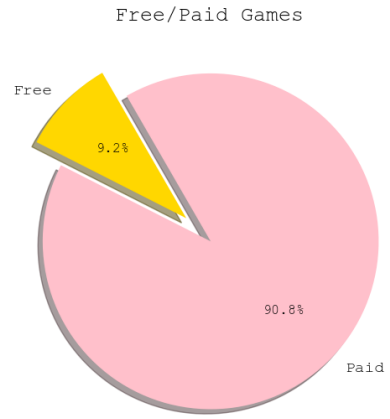

Fig. 11. How many games are paid?

As we can see in Figure 11, the percentage of developers making the game free is only 9.2%.

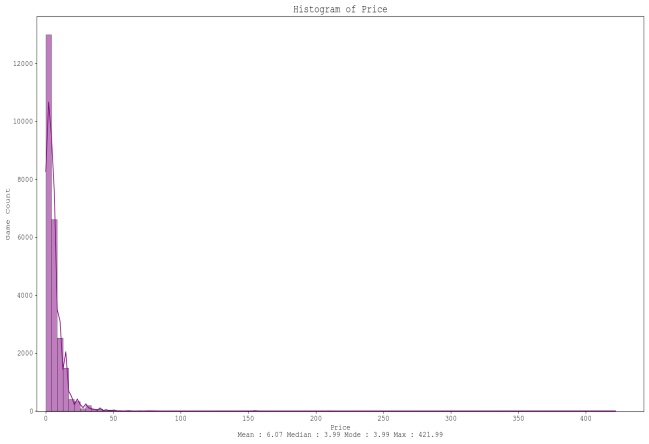

Fig. 12. Histogram of Price.

In the histogram that we can see in Figure 12, the majority of the games are between $0 and $50, but the video games that can be called Triple-A with the maximum value of $421 have extended the histogram to the right in the x-axis.

J. Released Date

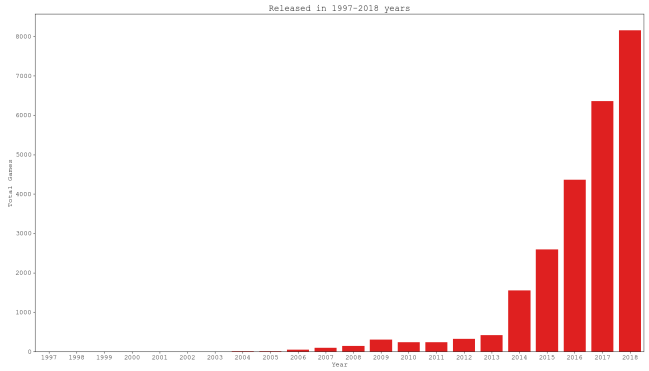

Fig. 13. Video Games Per Year.

From Figure 13, we can clearly see that the number of video games released every year is constantly increasing.

III. Questions about the Steam Data

We need the high-level research question to analyze the data and parameters. Then we need low-level technical questions to support this high-level question. With all these questions, we can reach the result more easily.

A. The High Leve Research Question: What are the attributes that affect game sales the most?

While developing video games, the sales numbers of the game and the income we can generate from it are very important. That's why we want our game to sell well, but it is not certain what are the attributes that make this sale big. There are many factors that will affect game sales. While Flappy Bird with almost no budget makes millions, the game with a budget of millions of dollars may not be liked at all and may not make a profit. Which brings us to our high-level research question that may reveal the solution to this problem. "What are the attributes that affect game sales the most? Of course, we also need subquestions to answer this question.

B. Low Level Technical Questions

- Does having more achievements affects the game sales positively?
- Do higher age ratings affect the game sales negatively?
- Does having english support affects the game sales positively?
- Do genres affect the game sales?
- Do categories affect the game sales?
- Does developers' name affects the game sales?
- Does having more platform support affects the game sales positively?
- Do higher prices affect the game sales negatively?
- Does publishers' name affects the game sales?
- Do release dates affect the game sales?
- Does having more reviews affects the game sales positively?

IV. Hypothesis Tests And The Methods

A. Hypothesis Tests

- "Does having more achievements affects the game sales positively?" This question is about a relationship between two different data.
  Null Hypothesis: There isn't any relation between achievements and owners count.
  Alternative Hypothesis: There is a relation between achievements and owners count.

- "Do higher age ratings affect the game sales negatively?" This question is about a relationship between two different data.
  Null Hypothesis: There isn't any relation between age ratings and owners count.
  Alternative Hypothesis: There is a relation between age ratings and owners count.

- "Does having english support affects the game sales positively?" This question is about a relationship between two different data.
  Null Hypothesis: There isn't any relation between having english support and owners count.
  Alternative Hypothesis: There is a relation between having english support and owners count.

- "Do genres affect the game sales?" This question is about a relationship between two different data.
  Null Hypothesis: There isn't any relation between genres and owners count.
  Alternative Hypothesis: There is a relation between genres and owners count.

- ”Do categories affect the game sales?” This question is about a relationship between two different data.
  Null Hypothesis: There isn’t any relation between categories and owners count.
  Alternative Hypothesis: There is a relation between categories and owners count.

- ”Does developers’ name affects the game sales?” This question is about a relationship between two different data.
  Null Hypothesis: There isn’t any relation between developers’ name and owners count.
  Alternative Hypothesis: There is a relation between developers’ name and owners count.

- ”Does having more platform support affects the game sales positively?” This question is about a relationship betwwen two different data.
  Null Hypothesis: There isn’t any relation between having more platform support and owners count.
  Alternative Hypothesis: There is a relation between having more platform support and owners count.

- ”Do higher prices affect the game sales negatively?” This question is about a relationship between two different data.
  Null Hypothesis: There isn’t any relation between price and owners count.
  Alternative Hypothesis: There is a relation between price and owners count.

- ”Does publishers’ name affects the game sales?” This question is about a relationship between two different data.
  Null Hypothesis: There isn’t any relation between publishers’ name and owners count.
  Alternative Hypothesis: There is a relation between publishers’ name and owners count.

- ”Do release dates affect the game sales?” This question is about a relationship between two different data.
  Null Hypothesis: There isn’t any relation between released date and owners count.
  Alternative Hypothesis: There is a relation between released date and owners count.

- ”Does having more reviews affects the game sales positively?” This question is about a relationship between two different data.
  Null Hypothesis: There isn’t any relation between having more reviews and owners count.
  Alternative Hypothesis: There is a relation between having more reviews and owners count.

V. Multivariable Study

So far we have only examined our univariate data and thought about how we might find answers to our questions. Now we can move on to multivariate analysis, which we accept as interdependent, and look for answers to our questions in the previous section. Let’s first look at the correlation heatmap.

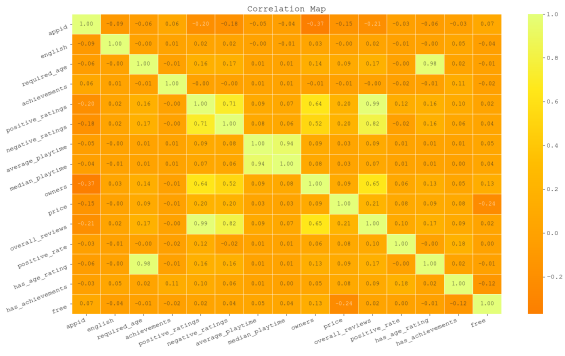

Fig. 14. Correlation Heatmap.

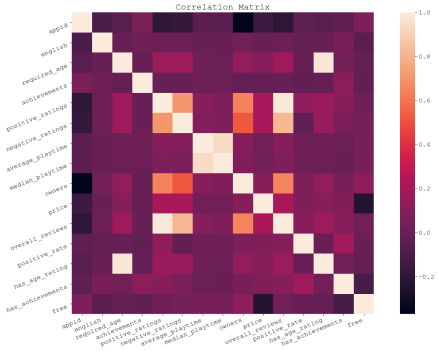

Fig. 15. Correlation Heatmatrix.

By looking at the heatmap, we can say that the variables that the owners count is more dependent on are overall_reviews, positive and negative reviews. In addition, it can be noticed that the achivements and price factors are also slightly dependent on this issue. Let’s look at the scatter and regression plots related to the variables in the table. Let’s examine their respective graphs. After that, we will observe how much non-quantitative variables depend by looking at theirs Box and Whiskers methods.

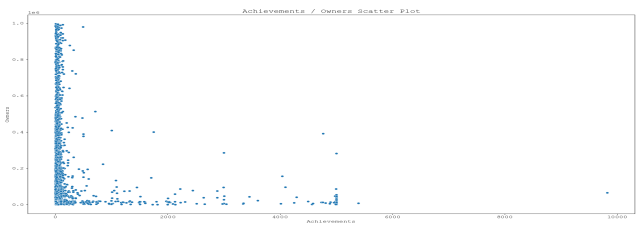

Fig. 16. Scatter Plot of Achievements/Owners.

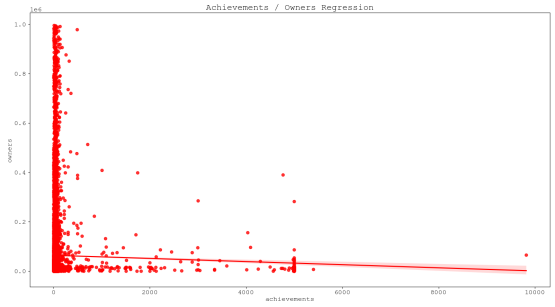

Fig. 17. Regression of Achievements/Owners.

Based on the scatter and regression graphs, we can say that the achievements column has a negative effect on the number of owners. In addition, the has_achievements column shows that the presence of achievements has a positive effect on the number of owners. Therefore, we can say that the presence of achievements in a game, but not using it more than necessary, is the situation that will affect the sales of the game in the most positive way.

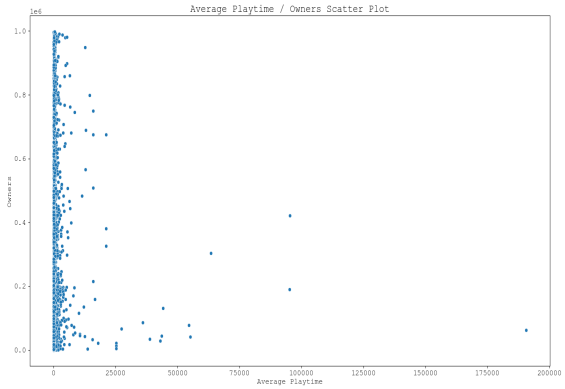

Fig. 18. Scatter Plot of Average Playtime/Owners.

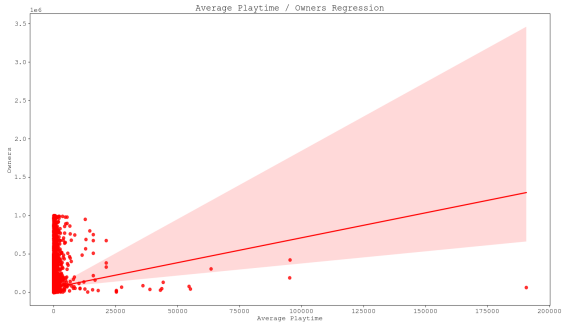

Fig. 19. Regression of Average Playtime/Owners.

Based on the scatter and regression graphs, we can say that the average_playtime column has a positive effect on the number of owners. However, the distribution of the majority of the games is in the part where the average playtime is close to 0. Therefore, we can say that a high average playtime will not adversely affect game sales.

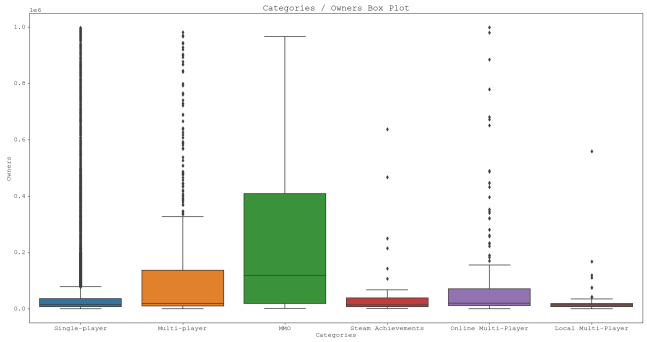

Fig. 20. Box Plot of Categories/Owners.

Based on the Categories/Owners box plot made for the 6 most existing categories, we can say that the MMO category has a wider spread and its median is superior to the other categories. We can say that the reason for this is that the games in the MMO category are usually free games. The Single Player category, which is the most common category in our dataset, tells us something different. It is impossible to say how this category can directly affect the sales of a game due to the presence of so many. Although we say that the median value is in under-owned games, we have a lot of outliers. The Multi Player category, which is the third category to be examined, is somewhere between Single Player and MMO. Although its median and distribution is similar to the Single Player category with low sales, it is more similar to MMO with its maximum value and outliers.

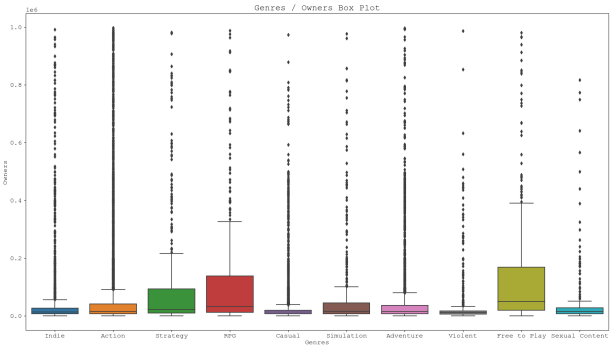

Fig. 21. Box Plot of Genres/Owners.

Based on the Genres/Owners box plot made on the top 10 existing genres, we can say that Free to Play and RPG genres (the fact that there is a similar data for that category, since there are genres in the MMO category, shows the accuracy of our graphics) have a wider spread and median is superior to other

genres. We can say that this is due to the fact that these types of games are mostly free games. The most common genre in our dataset, the action genre, tells us something different. (Since the genre in the Single Player category is action in general, the fact that there is a similar data for that category shows the accuracy of our graphics.) It is impossible to say how this genre can directly affect the sales of a game due to the presence of too many. Although we say that the median value is in under-owned games, we have a lot of outliers. Strategy and Adventure genres, which can be examined as the 3rd, are somewhere between Action and Free To Play/RPG. Although its median and distribution are similar to the action genre in places with low sales, Free To Play/RPG is more similar with its maximum value and outliers.

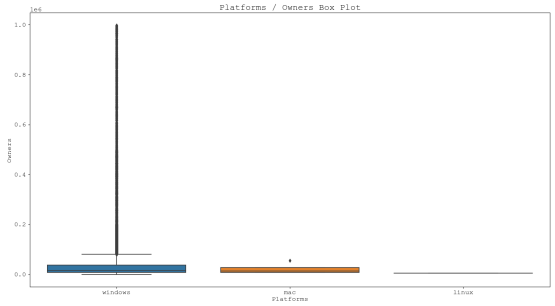

Fig. 22. Box Plot of Platforms/Owners.

Based on the Platforms/Owners box plot, we can say that the windows platform has a clear advantage over other platforms. However, this superiority does not have a positive effect on game sales. The fact that the game market is mostly carried out on this platform leaves us without comment on what effect it has on game sales.

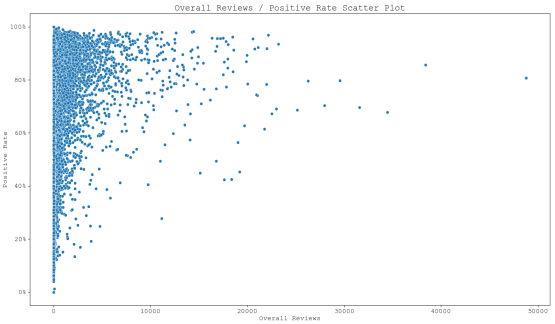

Fig. 23. Scatter Plot of Overall Reviews/Positive Rate.

Based on the Overall Reviews/Positive Rate scatter plot, the more users review a game, the more positive these reviews show. It is inevitable that more popular games sell more, so more people review the game. Now let's examine the effect of the positive rate column on game sales.

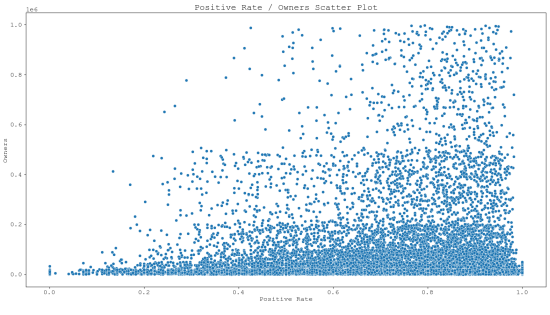

Fig. 24. Scatter Plot of Positive Rate/Owners.

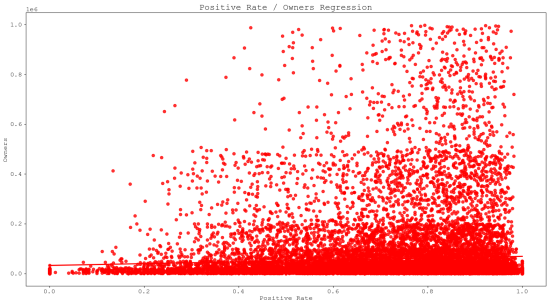

Fig. 25. Regression of Positive Rate/Owners.

Based on the scatter and regression graphics, we can say that the positive_rate column has a positive effect on game sales. The more positive reviews a game gets, the more likely it is to sell more games.

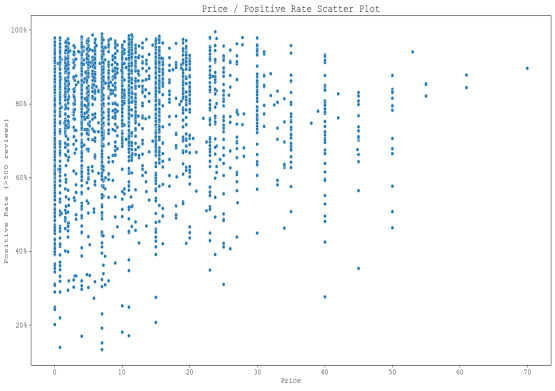

Fig. 26. Scatter Plot of Positive Rate/Price.

According to the Price/Positive Rate scatter plot, the more expensive a game is, the more positive reviews it gets. This is understandable, given that more expensive games are generally of higher quality. Now let's examine the effect of the price column on game sales.

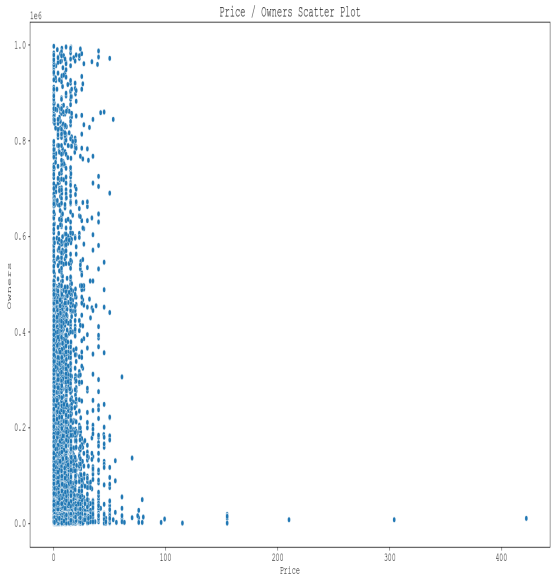

Fig. 27. Scatter Plot of Price/Owners.

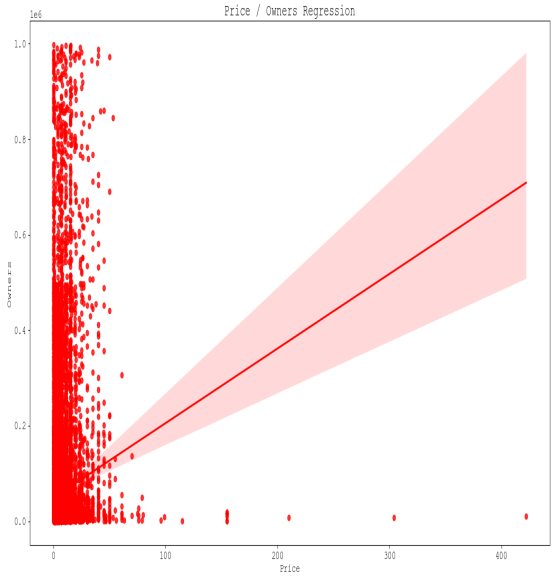

Fig. 28. Regression of Price/Owners.

From the Price/Owners Count scatter and regression plots, we can say that the price column has a huge impact on game sales. And this effect is negative. The more expensive a game is, the fewer copies it sells.

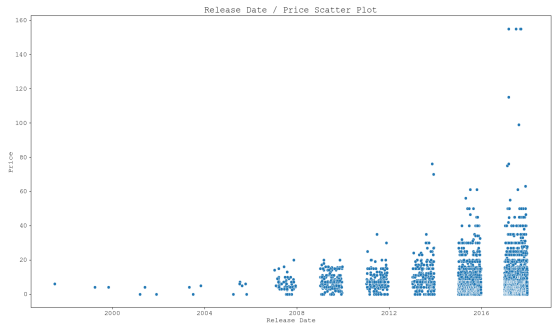

Fig. 29. Scatter Plot of Release Date/Price.

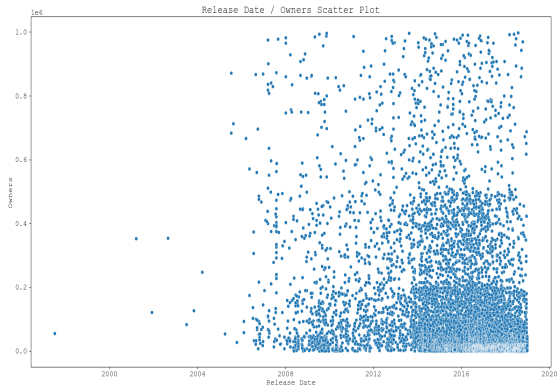

Fig. 30. Scatter Plot of Release Date/Owners.

By looking at the Price/Release Date and Release Date/Owners Count scatter plots, we can say that game prices are increasing as we approach 2020. We can say that the reason for this price increase is the increase in costs and inflation. Although the number of users increases as we approach 2020, we cannot observe that the games released in these years sell more games. We can say that the price column comes into play here.

VI. Conclusion