Data Science Insight Into the History of Video Games
23 Heatran

Ayşe Orkan
181180055

Çiya Baran Öner
181180056

Mert Sağır
181180061

Mustafa Nesin
181180054

ABSTRACT

In this study, various analyzes were made on datasets related to video games. These analyzes include video games sales, genres, gaming frequencies and gaming platforms.

1. INTRODUCTION

The purpose of this study is to grasp the history of video games with data science and make some insights and predictions. These predictions and information are useful to both gamers and developers to gain insight into video games. We have carried out various studies using the data we have for this purpose.

2. METHODOLOGY

2.1 Predicting the Score

One of the questions we sought to answer in the project was what score a newly released game would get from critics or users. Initially we thought we could use regression to find the answer to this. For this, we first examined the correlations between the features on the data set we have using a heatmap. However, we concluded that using a machine learning algorithm would not be very useful in this question, since there is no other column where the score information has sufficient correlation. Finally, in this question, we found it reasonable to use the average method over other numeric columns in the datasets. For instance, the score of a new game to be published by a company was highly likely to be close to the average score of the games it had previously published.

2.2 Linear Regression

Linear Regression is a supervised Machine Learning model where the model finds the most appropriate linear line between the independent and dependent variable. With linear regression, when sales figures of games are available for various regions , sales of the region without sales information for a new game can be estimated relative to other regions.

3. DATASET

3.1 Video Game Sales with Ratings

Video Game Sales with Ratings [1] is one of the data sets that we use extensively. The dataset is also quite popular on Kaggle. It consists of 16 columns and 11,563 unique rows. The columns in the dataset are as follows;

- Name: The name of the game,
- Platform: Console that running the game,
- Year of release,
- Genre: Game's category,
- Publisher,
- Game sales in North America,
- Game sales in Europe,
- Game sales in Japan,
- Game sales in other continents,
- Total sales in the world,
- Average critic score,
- Number of critics,
- Average user score,
- Number of users who rated,
- Developer of the game,
- The ESRB rating of the game (E.g. Everyone, Teen, Adults Only…).

After importing the dataset, we made some adjustments to the dataset so that we could use it in our questions. First, we examined how many null values in each column with the help of the following heatmap.

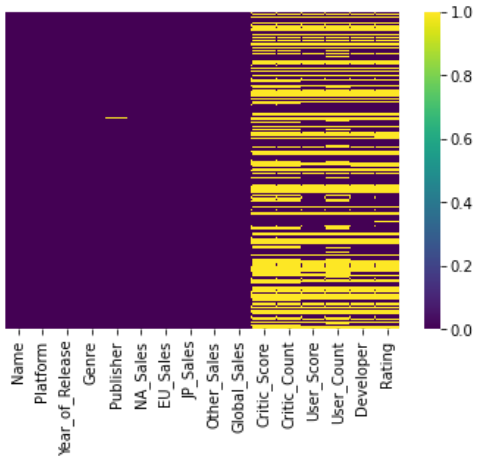


Figure 1. Heatmap of the null values on the Video Game Sales with Ratings dataset

By looking at the distribution of null values in the columns, we saw that almost half of the data in the columns we will use for score prediction is null. However, considering that there are approximately 1200 rows in the data set, we decided that half of this number, 6000, is sufficient to answer our questions. As a result, we deleted the rows that were null in the columns we will use.

One of the problems we encountered in this dataset was the presence of string values in the User Score column, which we expected to be a floating point number. In the explanations on Kaggle, it was stated that some User Score values were written as tbd in the data set because it was not determined yet. We also filtered these values for the sake of the data clarity.

Another issue was that the User Score value was out of 10, while the Critic Score value was out of 100. For this, we scaled the User Score by multiplying it's column by 10.

3.2 Video Games Sales 2019

Video Games Sales 2019 [4] is another dataset that we use. It consists of 23 columns and 37102 unique rows. It has additional features such as the URL of the game page and the URL of the game's cover image.

4. EXPERIMENTS

4.1 Predicting the Score

For this question, we used the Video Game Sales with Ratings dataset. After making the relevant adjustments in the dataset, we used the following heatmaps to see the correlation between columns. We will look at what features we can use regression based on these heatmaps.

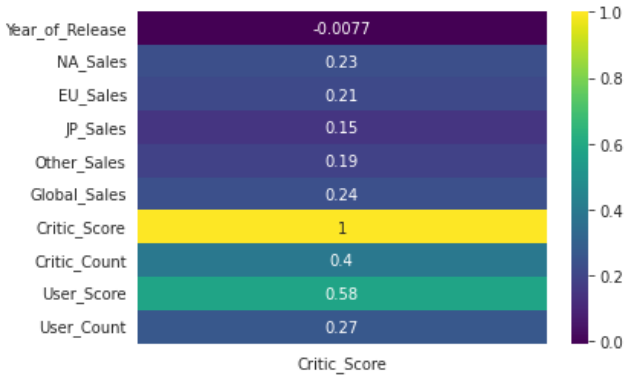


Figure 2. Heatmap showing correlation between Critic Score and other continuous features

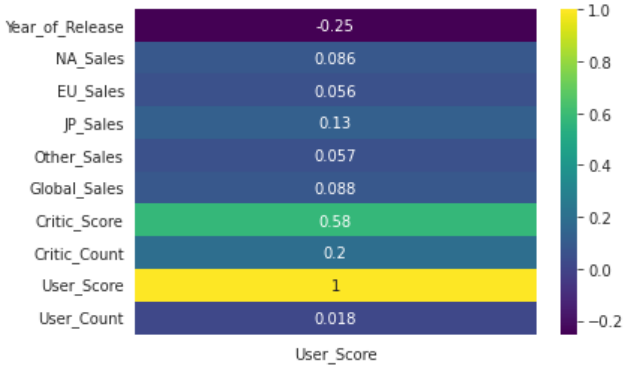


Figure 3. Heatmap showing correlation between User Score and other continuous features

Based on these heatmaps, we saw that there was no significant correlation between the score columns and the other columns. So, we cannot apply regression on this dataset based on these columns.

There are two types of scores to predict. These are Critic Score and User Score values. Let us plot both columns into the same histogram plot to see the relationship between these two values.

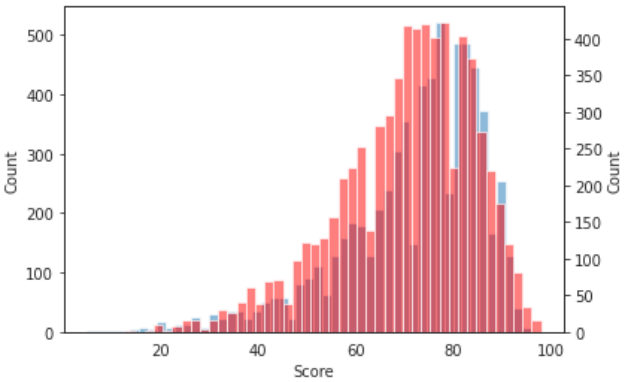


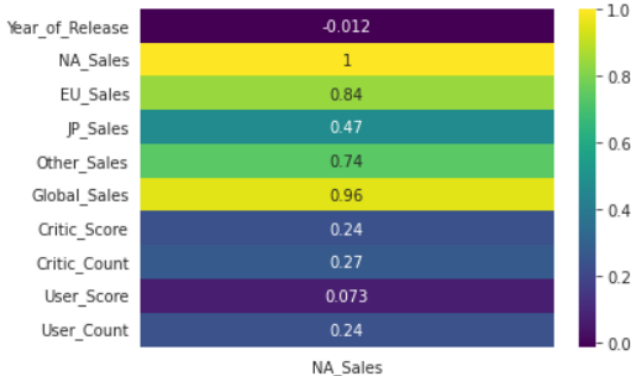
Figure 4. User Score (blue) and Critic Score (red) in the same histogram plot

As expected, there is a similarity between the two values. Only the critics are just more 'stingy' than the users when it comes to rating the games. Regression can be used between these two

values, but it goes against the nature of this question as they both represent the score. Let us try to predict the Critic Score for this question.

When we look at the other columns, the feature that is most related to the Critic Score seems to be Developer and Publisher. We will calculate the average scores of each developer and publisher. Thus, when a developer or publisher releases a new game, its expected score will be the average of the scores of their previous games.

4.2 Linear Regression



Based on the heatmap, a significant correlation is seen between the North American sales column and the other sales columns. Thus, with linear regression, North American sales can be estimated relative to sales in other regions.

5. RESULTS

5.1 Predicting the Score

We concluded that the feature that is most related to the Critic Score is Developer and Publisher. So we can predict the score of the new game from 2 different features.

5.1.1 Score of the Developer's New Game

The developers with the highest and lowest average scores are given in the tables below.

Table 1. Average scores of developers with a Critic Count greater than 15

Rank	Developer	Score
1.	DMA Design	97.0
2.	Irrational Games, 2K Marin	96.0
3.	Digital Extremes, 2K Marin	94.0
4.	Bungie Software	93.6
5.	DMA Design, Rockstar North	93.0
	…	…
1023.	Red Tribe	29.0
1024.	Santa Cruz Games	28.0
1025.	AI	26.0
1026.	Neko Entertainment	23.0
1027.	Idol FX	22.0

Here, we see that especially the games developed by DMA Design are highly rated by Critics. Therefore, we can predict that the score will be higher than 90 in a new game that DMA Design will develop. On the other hand, for example, we see Red Tribe was rated low by Critics. Therefore, we can predict that the new game that Red Tribe will release will have about a score of 30.

5.1.2 Score of the Publisher's New Game

The publishers with the highest and lowest average scores are given in the tables below.

Table 2. Average scores of publishers with a Critic Count greater than 15

Rank	Publisher	Score
1.	Valve	96.0
2.	Valve Software	93.0
3.	2D Boy	90.0
4.	SquareSoft	89.2
5.	Devolver Digital	88.0
	…	…
191.	Mad Catz	49.0
192.	Phantom EFX	44.0
193.	Telltale Games	40.0
194.	Evolved Games	38.0
195.	Avanquest	31.0

Here, we see that especially the games published by Valve are highly rated by Critics. Therefore, we can predict that the score will be higher than 90 in a new game that Valve will release. On the other hand, for example, we see Telltale Games was rated low by Critics. Therefore, we can predict that the new game that Telltale Games will release will have about a score of 40.

5.2 Which Continent Plays Which Game More?

Let us find the 5 most played games in North America, Europe and Japan.

5.2.1 Most Played Games in North America

After sorting the sales values in North America in descending order, we got a result like the one below.

Table 3. 5 most played games in North America

Rank	Game	Sales (in millions)
1.	Wii Sports	41.36
2.	Super Mario Bros.	29.08
3.	Duck Hunt	26.93
4.	Tetris	23.20
5.	Mario Kart Wii	15.68

5.2.3 Most Played Games in Europe

After sorting the sales values in Europe in descending order, we got a result like the one below.

Table 4. 5 most played games in Europe

Rank	Game	Sales (in millions)
1.	Wii Sports	28.96
2.	Mario Kart Wii	12.76
3.	Nintendogs	10.95
4.	Wii Sports Resort	10.93
5.	Brain Age: Train Your Brain in Minutes a Day	9.20

5.2.3 Most Played Games in Japan

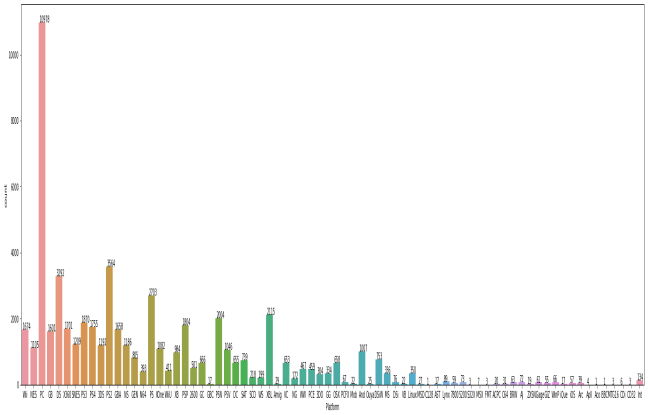
After sorting the sales values in Japan in descending order, we got a result like the one below.

Table 5. 5 most played games in Japan

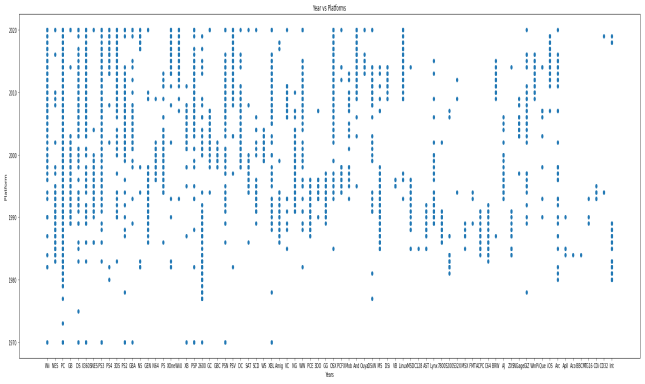
Rank	Game	Sales (in millions)
1.	Pokemon Red/Pokemon Blue	10.22
2.	Pokemon Gold/Pokemon Silver	7.20
3.	Super Mario Bros.	6.81
4.	New Super Mario Bros.	6.50
5.	Pokemon Diamond/Pokemon Pearl	6.04

5.3 Which platform should be focused on for the new games from the publisher's perspective?

Platform is one of the most important things for gamers. A platform can be so popular or comfortable for users but it doesn't mean that it will be used forever. For example, once PS1 was the most popular platform but it almost does not exist now. Thus, publishers have to give weight to some platforms more than others. So let's examine what happened over the years.

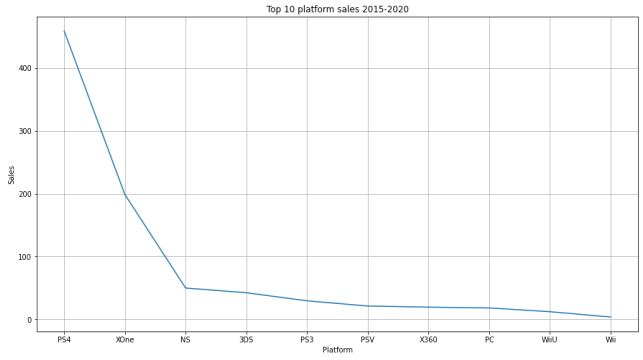


This graph shows which platforms the games are most often produced for.For example, PC is the first, PS2 is the second and DS is the third. However, DS and PS2 don't use it very much at the moment.



This graph shows which platform has been used for how many years. As we can see, many of them are not using it at the moment. We decided to use the years 2015-2020 to make the most correct decision.

Therefore, we removed datas from before 2015 and printed out the 10 most preferred platforms for 2015-2020.



As we can see, the best seller platform for games is the PS4. XOne is in the second order. So publishers could shape their preferences according to the chart.

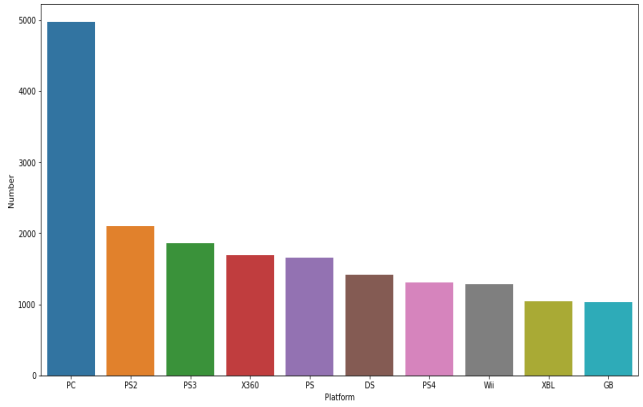
5.4 What are North American sales relative to sales in other regions?

In the linear regression model, the test score is 0.59 and the train score is 0.76. If the train score is higher than the test score, it is a sign of overfitting. Overfitting is the problem that a generated analysis overfits a particular dataset and therefore can not fit new data that is not included in that dataset.

We tried to see if our linear regression model made an accurate prediction. For this, we entered the sales data we have (12.76,3.79,3.29). Our model's estimate is 15.53 for North America. The actual sales figure is 15.68.

5.5 On which platforms are the games played the most?

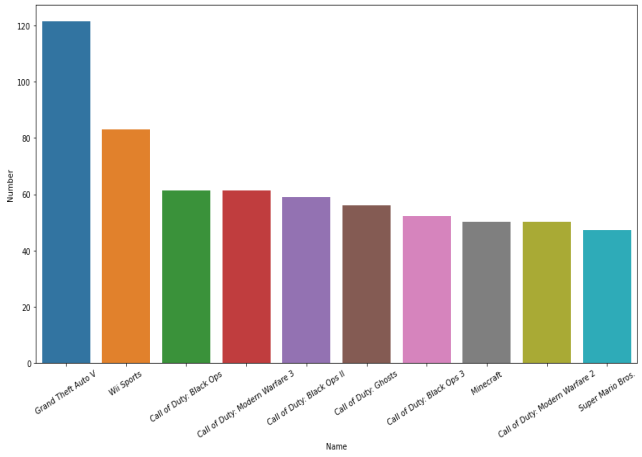
For developers, it is very important to know which games are played on which platforms. Because the same game comes out on different platforms which cause different sales numbers. Also the sales numbers of the platforms are different. Nearly everyone has a PC but not everyone has a PS3 etc. So developers should know on which platform each game is played the most.



This graph shows on which platforms are the games played the most. As you can see, the PC is the main platform for playing games. But if we look between gaming consoles then PS2 is higher than the others. That is mainly because the PS2 is the most sold gaming console of all time.

5.6 What are the most played games over time?

Hundreds of millions of games are sold every year around the world. We can generally determine the success of a game by how much that game sells. Therefore, it is important to know which games are the most played from the past to the present.



As you can see in the graph above, Grand Theft Auto 5 is by far the most played game according to the Video_Games_Sales_as_at_22_Dec_2016 database. Wii sports is in second place after it. And the rest of the games on the list are closer playtimes to each other after these two.

6. CONCLUSIONS

In this study, using data science discipline and python language, information about video games that can be useful to players, developers and game companies is extracted from datasets. This information includes the most sold and played games, the most played game types, the platforms with the most games, the most preferred platforms, the sales of the regions, developer and publisher scores, and which platform the publishers prefer for the new game. The results obtained can also be used to create market strategies. Players can consider game scale when choosing a platform, and their sales when choosing a game. While developers and game companies develop games for the most preferred platforms, they can also develop games according to the most preferred game types.

We did research on data science, practiced various techniques and learned to work together for a common goal. Each member of the group contributed to the completion of this study by doing their best.

7. REFERENCES

[1] Kaggle.com. 2016. Video Game Sales with Ratings. [online] Available at: <https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings> [Accessed 15 January 2022]

[2] https://en.wikipedia.org/wiki/Linear_regression

[3] https://en.wikipedia.org/wiki/Overfitting

[4] Kaggle.com. 2019. Video Games Sales 2019. [online] Available at: <https://www.kaggle.com/ashaheedq/video-games-sales-2019/metadata> [Accessed 25 January 2022]