Ali Kumral
30586

17/01/2024

CS210 Course Project Report

1-) <u>Introduction</u>

Steam is the largest platform for purchasing and playing games on the PC platform today. Launched on September 12, 2003, Steam has gained a lot of users since then and I am one of them. I opened my Steam account on February 28, 2016, and I have accumulated a lot of data on this platform over the past 8 years.

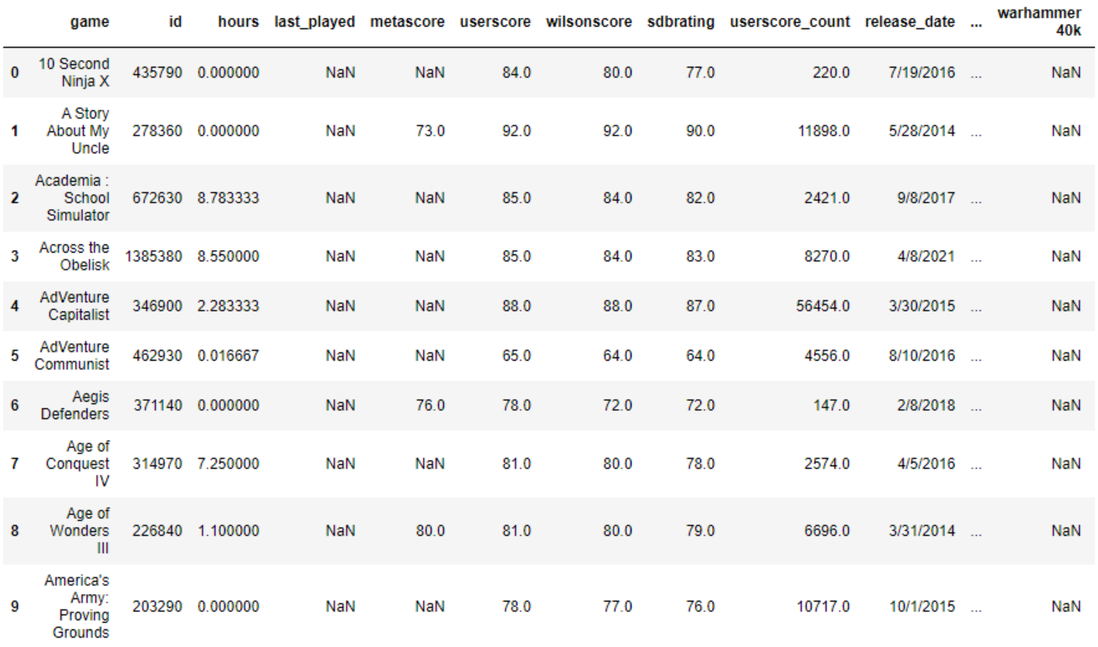
In this project, I will seek answers to my hypothesis and research questions by using my data on Steam. Moreover, to achieve these goals, I will use learning methods such as regression and random forest methods. More information about these models and their results can be found in parts 5 and 6.

Coding part of every graph, model and method used in this project can be found in the Python and Jupyter Notebook files, which are located in the GitHub repository where this PDF is located.

2-) <u>Data Collection</u>

Steam's API was used to obtain the data used in this project. 4 different data (user badges, user library, recent games, and user info) files were pulled from Steam API as Json files. To see these Json files or examine the API code, you can examine the "Steam API & Databases" file in the GitHub repository. The important information in the obtained Json files was combined and turned into csv files. These files are named “steam-library” and “steam-wishlist”.

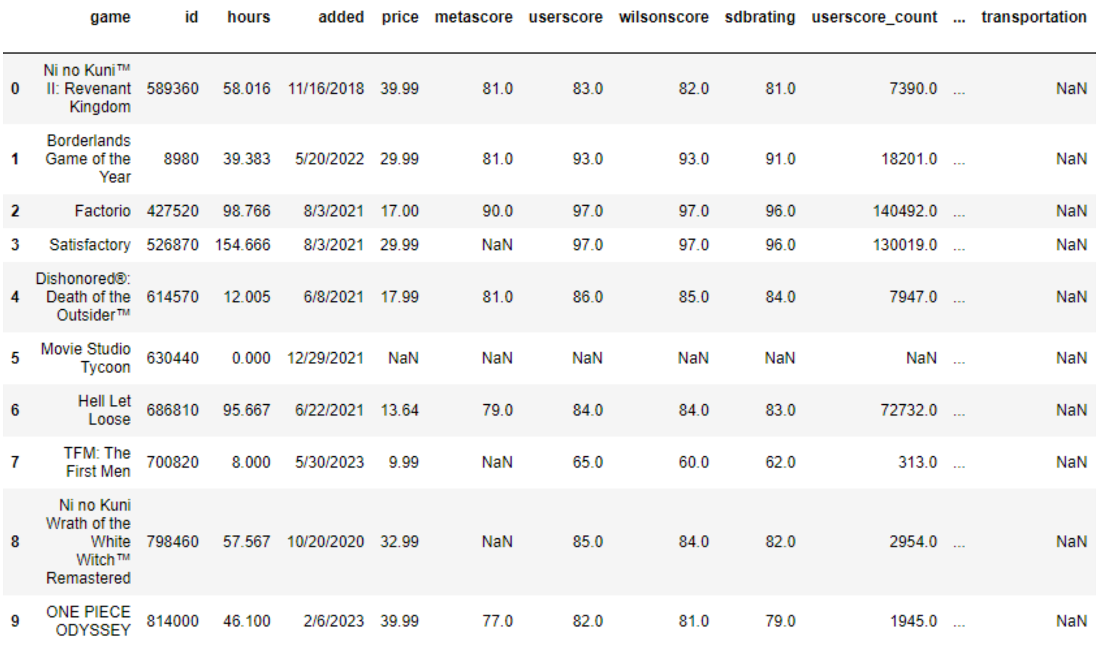
2.1 Steam-library csv file contains information about the games my account has and the playing time, ratings, and genre of these games. The file has 388 rows and 418 columns.



	game	id	hours	last_played	metascore	userscore	wilsonscore	sdbrating	userscore_count	release_date	...	warhammer 40k
0	10 Second Ninja X	435790	0.000000	NaN	NaN	84.0	80.0	77.0	220.0	7/19/2016	...	NaN
1	A Story About My Uncle	278360	0.000000	NaN	73.0	92.0	92.0	90.0	11898.0	5/28/2014	...	NaN
2	Academia : School Simulator	672630	8.783333	NaN	NaN	85.0	84.0	82.0	2421.0	9/8/2017	...	NaN
3	Across the Obelisk	1385380	8.550000	NaN	NaN	85.0	84.0	83.0	8270.0	4/8/2021	...	NaN
4	AdVenture Capitalist	346900	2.283333	NaN	NaN	88.0	88.0	87.0	56454.0	3/30/2015	...	NaN
5	AdVenture Communist	462930	0.016667	NaN	NaN	65.0	64.0	64.0	4556.0	8/10/2016	...	NaN
6	Aegis Defenders	371140	0.000000	NaN	76.0	78.0	72.0	72.0	147.0	2/8/2018	...	NaN
7	Age of Conquest IV	314970	7.250000	NaN	NaN	81.0	80.0	78.0	2574.0	4/5/2016	...	NaN
8	Age of Wonders III	226840	1.100000	NaN	80.0	81.0	80.0	79.0	6696.0	3/31/2014	...	NaN
9	America's Army: Proving Grounds	203290	0.000000	NaN	NaN	78.0	77.0	76.0	10717.0	10/1/2015	...	NaN

Figure 1. Raw data of the Steam-library csv

2.2 Steam-wishlist csv file contains the games that are wishlisted in my account. This file contains more or less similar type (columns) parameters as other file (steam-library file) for games. However, unlike my library, this file contains the games in my wishlist. The file has 40 rows and 226 columns. The values in the "hours" column in this file were taken from the HowLongToBeat website in order to be more detailed and accurate.



	game	id	hours	added	price	metascore	userscore	wilsonscore	sdbrating	userscore_count	...	transportation
0	Ni no Kuni™ II: Revenant Kingdom	589360	58.016	11/16/2018	39.99	81.0	83.0	82.0	81.0	7390.0	...	NaN
1	Borderlands Game of the Year	8980	39.383	5/20/2022	29.99	81.0	93.0	93.0	91.0	18201.0	...	NaN
2	Factorio	427520	98.766	8/3/2021	17.00	90.0	97.0	97.0	96.0	140492.0	...	NaN
3	Satisfactory	526870	154.666	8/3/2021	29.99	NaN	97.0	97.0	96.0	130019.0	...	NaN
4	Dishonored®: Death of the Outsider™	614570	12.005	6/8/2021	17.99	81.0	86.0	85.0	84.0	7947.0	...	NaN
5	Movie Studio Tycoon	630440	0.000	12/29/2021	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
6	Hell Let Loose	686810	95.667	6/22/2021	13.64	79.0	84.0	84.0	83.0	72732.0	...	NaN
7	TFM: The First Men	700820	8.000	5/30/2023	9.99	NaN	65.0	60.0	62.0	313.0	...	NaN
8	Ni no Kuni Wrath of the White Witch™ Remastered	798460	57.567	10/20/2020	32.99	NaN	85.0	84.0	82.0	2954.0	...	NaN
9	ONE PIECE ODYSSEY	814000	46.100	2/6/2023	39.99	77.0	82.0	81.0	79.0	1945.0	...	NaN

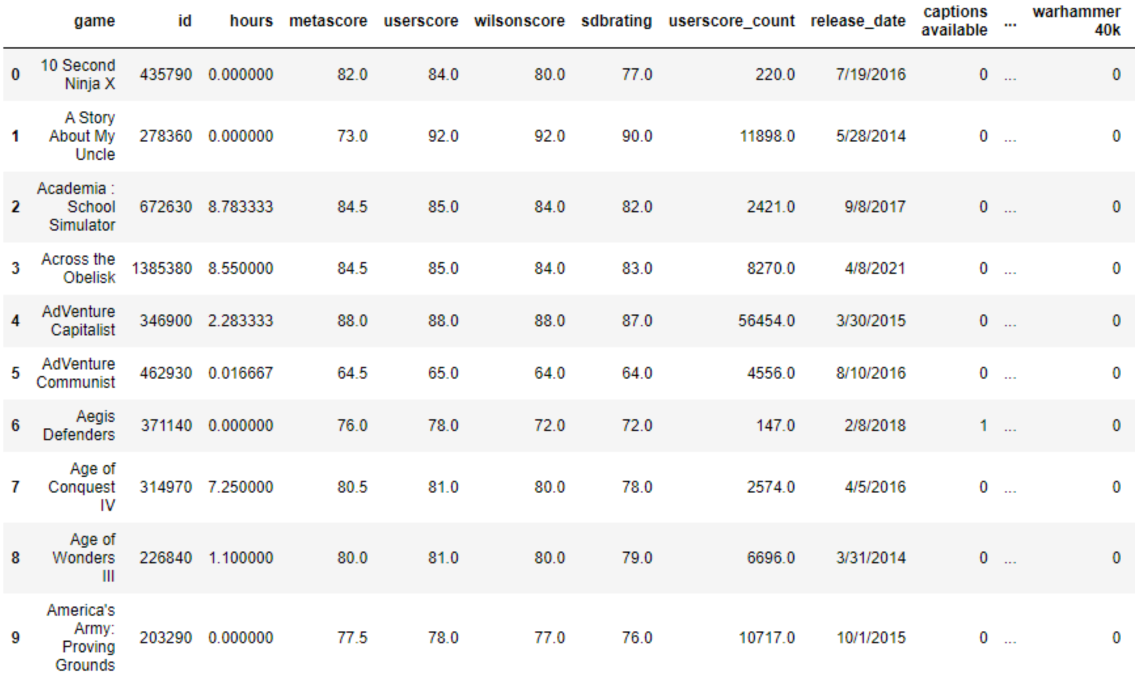
Figure 2. Raw data of the Steam-wishlist csv

3-) Data Cleaning and Standardization

In this project, the Data Cleaning part consists of 2 parts. The first part constitutes the steps taken to correct the problem that may arise when analyzing the raw data. The second part consists of data cleaning steps to optimize the model in the machine learning section. The detailed data and codes of this process can be found in the Jupyter Notebook file in the GitHub repository.

3.1 First Part

As seen in Figure 1 and Figure 2, there are parameters in the csv file that will make the data analysis and machine learning part difficult. In order to eliminate this difficulty, the NaN values in the csv file have been replaced with 0 and the "x" values with 1. Thus, 0 indicates the absence of that parameter in that game and 1 indicates its presence in that game as well. These 0 and 1 assignments are valid only for genres. To correct other NaN values (for example, Meta Score values with NaN values), the average of similar parameters is filled in.



	game	id	hours	metascore	userscore	wilsonscore	sdbrating	userscore_count	release_date	captions available	...	warhammer 40k
0	10 Second Ninja X	435790	0.000000	82.0	84.0	80.0	77.0	220.0	7/19/2016	0	...	0
1	A Story About My Uncle	278360	0.000000	73.0	92.0	92.0	90.0	11898.0	5/28/2014	0	...	0
2	Academia : School Simulator	672630	8.783333	84.5	85.0	84.0	82.0	2421.0	9/8/2017	0	...	0
3	Across the Obelisk	1385380	8.550000	84.5	85.0	84.0	83.0	8270.0	4/8/2021	0	...	0
4	AdVenture Capitalist	346900	2.283333	88.0	88.0	88.0	87.0	56454.0	3/30/2015	0	...	0
5	AdVenture Communist	462930	0.016667	64.5	65.0	64.0	64.0	4556.0	8/10/2016	0	...	0
6	Aegis Defenders	371140	0.000000	76.0	78.0	72.0	72.0	147.0	2/8/2018	1	...	0
7	Age of Conquest IV	314970	7.250000	80.5	81.0	80.0	78.0	2574.0	4/5/2016	0	...	0
8	Age of Wonders III	226840	1.100000	80.0	81.0	80.0	79.0	6696.0	3/31/2014	0	...	0
9	America's Army: Proving Grounds	203290	0.000000	77.5	78.0	77.0	76.0	10717.0	10/1/2015	0	...	0

Figure 3. Standardized data of the Steam-library csv

3.2 Second Part

The basis of data standardization in this part is to make the model used in the machine learning part more accurate and optimal. In order to achieve this, it was necessary to create two new data frames using the intersecting parameters of two different csv files (steam-library & steam-whislist). It is critical to perform this step because the test and train data must contain the same parameters. With this step, two different data frames containing different rows (games), but the same columns (parameters) were obtained.

4-) Hypothesis and Research Questions

My hypothesis is that among the games in the wishlist database, the predicted hours of the games in the genre I like the most (that is, the genre in which I have the most playing time) should be higher than the others.

My research questions are,

- Among the games I own, which genre contains the most games in terms of quantity?
- Among the games I own, which genre has been played the most?
- What is the percentage of backlogged games in my library?
- Is there a correlation between played/backlogged games and these games' review rating?
- Is there a correlation between played/backlogged games and these games' popularity?

5-) Findings and Results of Research Questions

Even if it is your own data, it is important to examine and analyze it to understand the database. Research Questions can help us at this point. Finding answers to these questions is of great importance for the project.

5.1 Most Frequent Genres

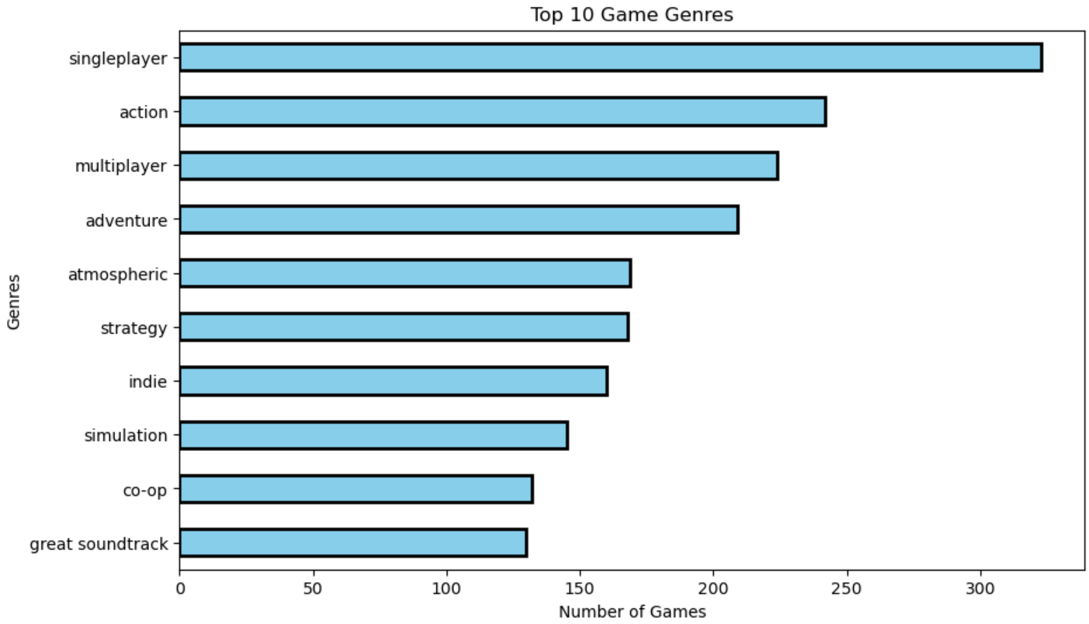


Figure 4. Frequency of the game genres in the Steam-library csv

It is not a surprise that super-genres like Singleplayer and Multiplayer, which include many genres, are at the top.

5.2 Most Played Genres

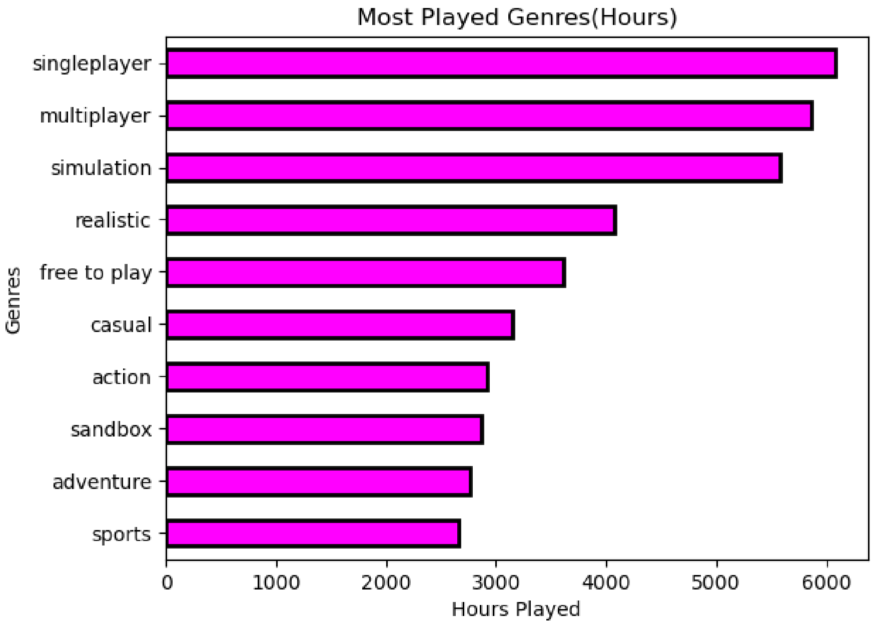


Figure 5. Most played genres in the Steam-library csv

We can say that the genres have changed significantly when it comes to playing the game rather than just buying it. The playtime info seen in the graph will be important for us when testing our hypothesis.

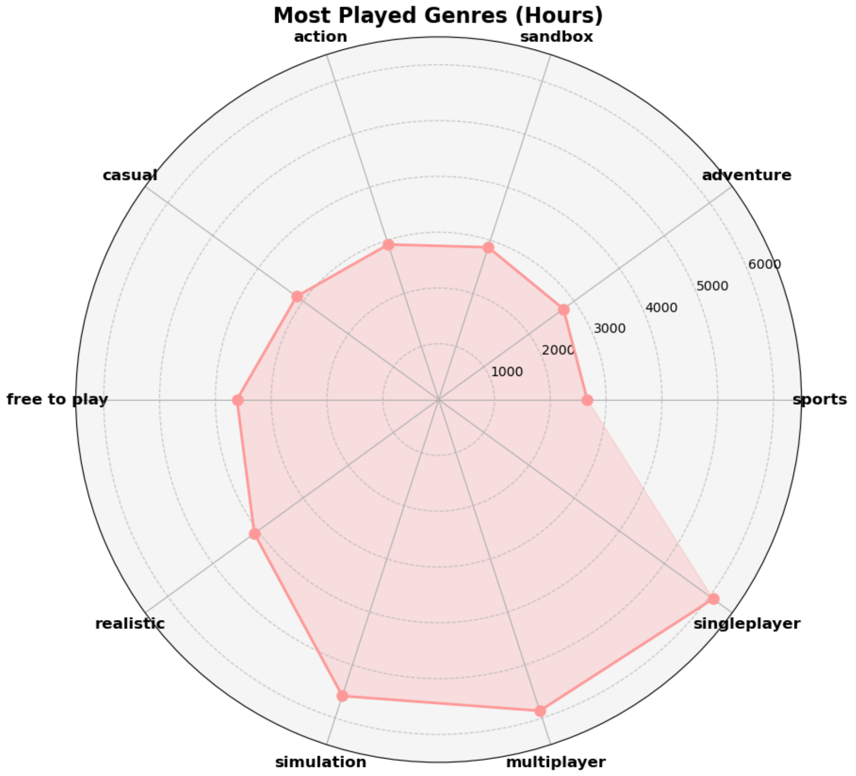


Figure 6. Alternate graph of most played genres in the Steam-library csv

5.3 Percentage of Backlogged Games

Speaking of played games, what about the backlogged games? In gaming terminology, “backlogged” means a game that has been purchased but not played. Hence, here are the played/unplayed game percentages.

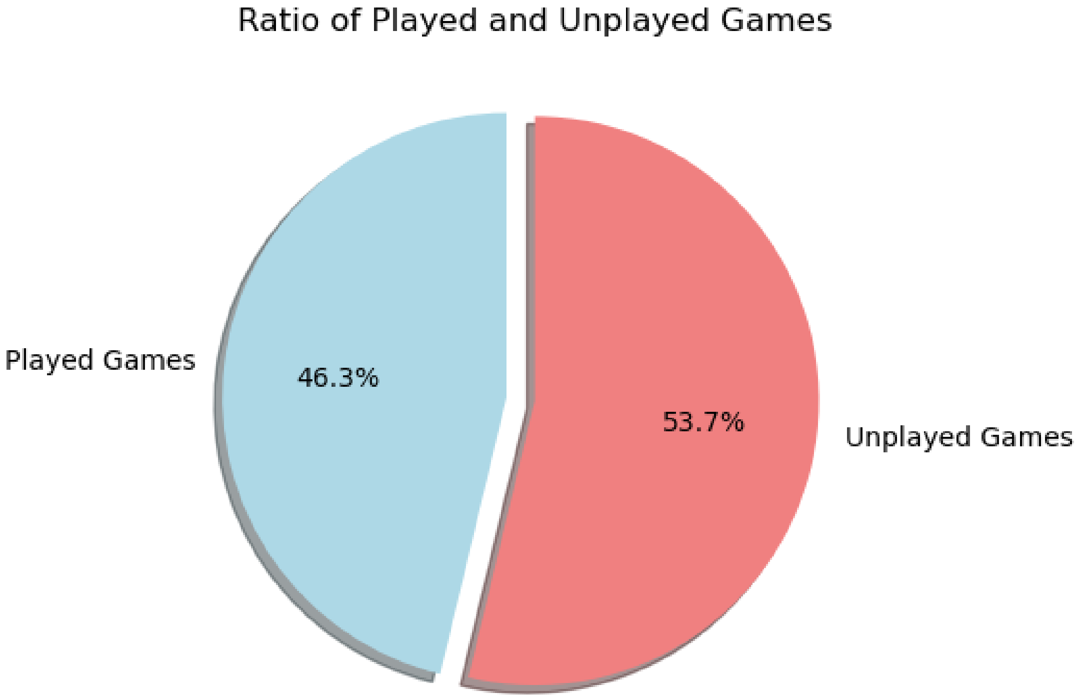


Figure 7. Ratio of Played and Unplayed Games in the Steam-library csv

5.4 Played/Backlogged Games - Review Rating Correlation

We have seen the playing percentage of games in the library, but what makes this difference? In order to understand this, the correlations of played and unplayed games can be looked at. First, let's evaluate the review score of the game on Steam.

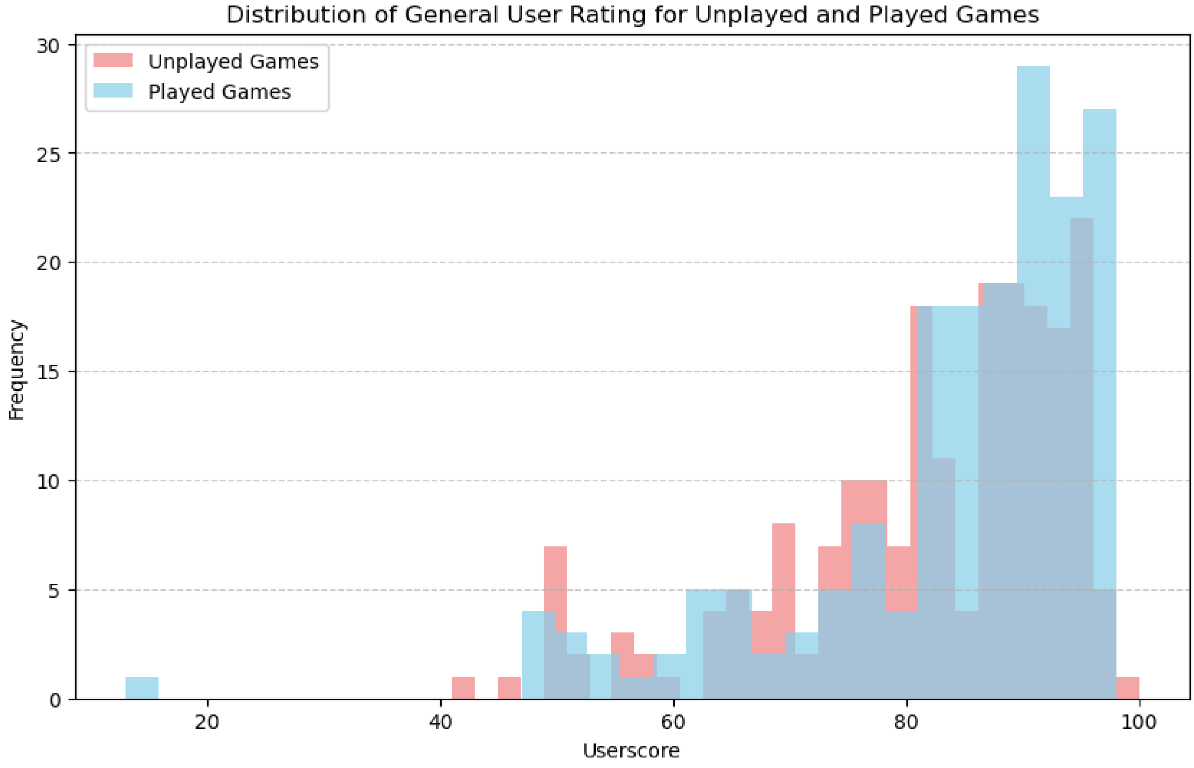
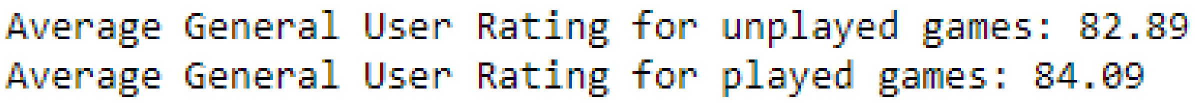


Figure 8. Distribution of General User Rating for Unplayed and Played Games

When we look at the graph, we see that they both have similar distributions. This may indicate that user rating is not an important factor in determining the playability of the game. Let's also look at the numerical values to be sure.



```
Average General User Rating for unplayed games: 82.89
Average General User Rating for played games: 84.09
```

If we evaluate the numerical values, we see that we get similar results there. As a result, we can say that general user rating is not a very important factor in game selection.

5.5 Played/Backlogged Games - Popularity Correlation

Now let's look at another correlation. How does the popularity of the game affect its play rate? In this correlation, I took the popularity parameter as how many people reviewed (user score count in the data frame) the game. In this way, more accurate results can be obtained.

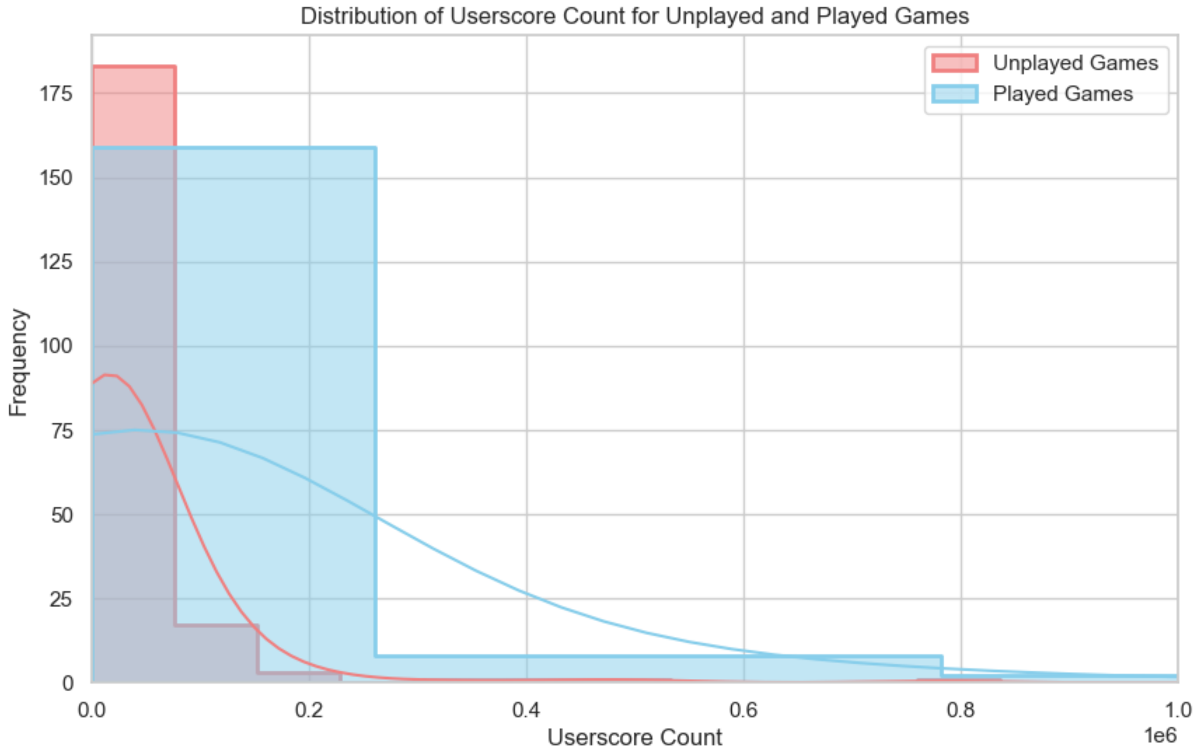
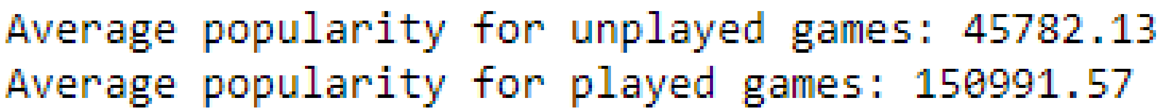


Figure 9. Distribution of Popularity/User Score Count for Unplayed and Played Games

Just by looking at the graph, it can be seen that popularity has an incredible impact on the rate of play of the game. Now let's compare the mathematical data and examine the accuracy of our interpretation.



```
Average popularity for unplayed games: 45782.13
Average popularity for played games: 150991.57
```

The results show that the popularity score of played games is up to 3 times higher than the unplayed ones, which is clear evidence that popularity has a big impact on playing time of the game.

6-) Machine Learning Models and Accuracy Results

It's time to use the data we edited in part 3.2. In this project, two machine learning models, Linear Regression and Random Forest, were used. The purpose of using these models is to predict the playing times of the games in the steam-wishlist data frame by training the steam-library data frame. For the exact prediction values, you can check Jupyter Notebook file in GitHub repository.

6.1 Random Forest Method

Random Forest combines the output of multiple decision trees to reach a single result. In this project 100 estimators are used to calculate the Random Forest Regression.

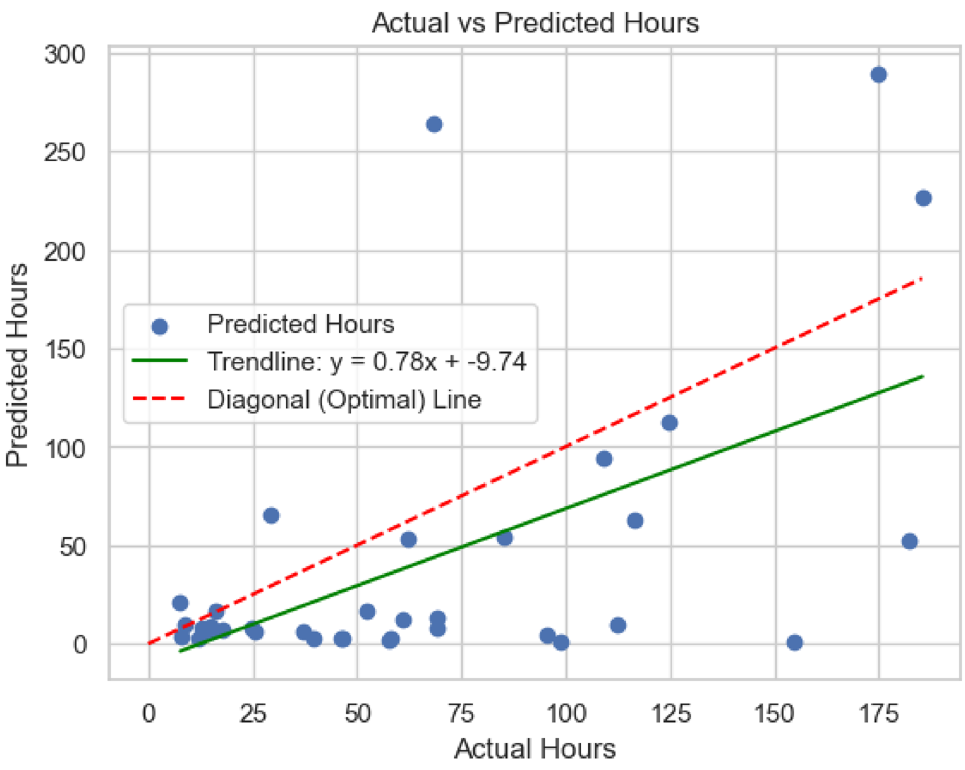


Figure 11. Random Forest Method Results

The red line in the graph shows the optimal line that prediction matches with the actual value. However, we can see that the blue dots are not gathered around the red line, but rather scattered. The green line represents the trendline of the blue dots, which is like a regression of these lines. Even though the green line does not match with the red one, we can say that we got a mediocre estimation method, since it is not too far away from red line. Furthermore, the correlation between 'actual hours' and 'predicted hours' is 0.556, which is not a bad result.

6.2 Linear Regression Method

Linear Regression is one of the simplest yet effective methods among machine learning methods.

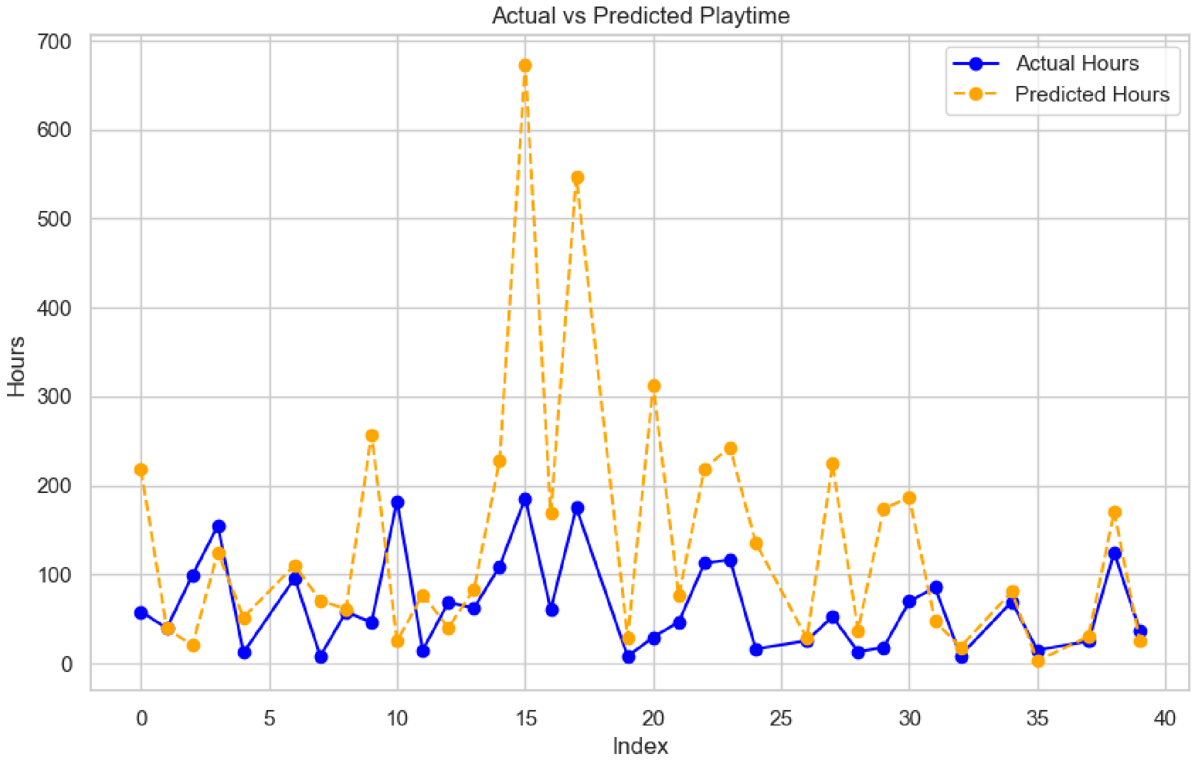
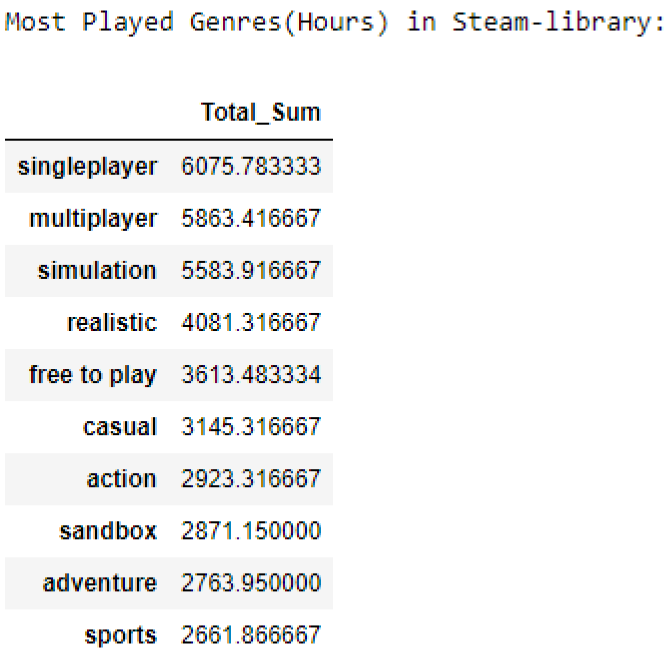


Figure 12. Linear Regression Method Results

Just by looking at the graph it can be seen that the model works very well. If we overlook the substantial overestimation in Index 15 and 17, we can say that the model passed the test successfully. In addition to predicting the ups and downs of the graph very well, the model often finds very close values. Moreover, the correlation between 'actual hours' and 'predicted hours' is 0.555. In the next stage, we will test our hypothesis with the values estimated by the model.

7-) Conclusion

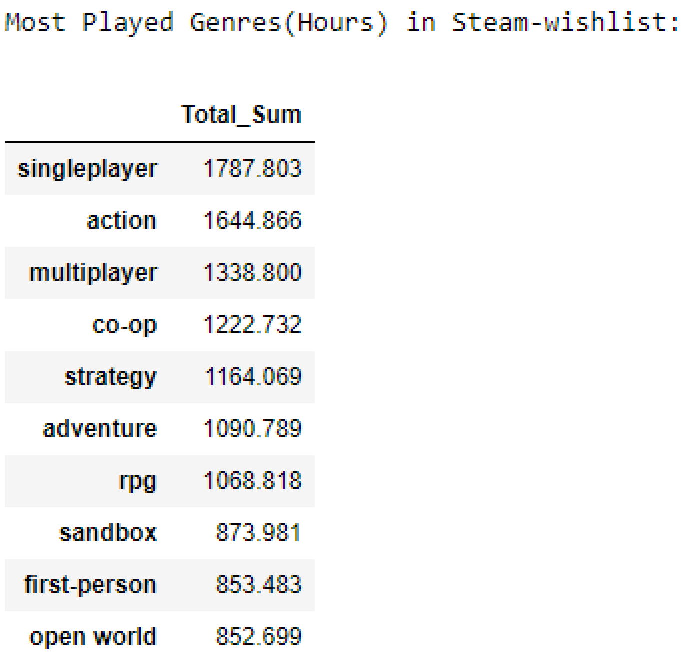
In this final part, I will use the results of the machine learning model (in part 6.2) to test whether our hypothesis (in part 4) is correct. My hypothesis was “among the games in the wishlist database, the predicted hours of the games in the genres I like the most (that is, the genres in which I have the most playing time) should be higher than the others.”. Let’s see if it’s correct or not.



```
Most Played Genres(Hours) in Steam-library:
```

	Total_Sum
singleplayer	6075.783333
multiplayer	5863.416667
simulation	5583.916667
realistic	4081.316667
free to play	3613.483334
casual	3145.316667
action	2923.316667
sandbox	2871.150000
adventure	2763.950000
sports	2661.866667

Figure 13. Most Played Genres(Hours) in Steam-library



```
Most Played Genres(Hours) in Steam-wishlist:
```

	Total_Sum
singleplayer	1787.803
action	1644.866
multiplayer	1338.800
co-op	1222.732
strategy	1164.069
adventure	1090.789
rpg	1068.818
sandbox	873.981
first-person	853.483
open world	852.699

Figure 14. Most Played Genres(Hours) in Steam-wishlist

When the tables are examined, the similarity between the two tables can be seen. Five of the first eight (singleplayer, action, multiplayer, adventure, sandbox) most played genres in the steam-wishlist dataframe match the top genres in the steam-library data frame, which is our main data frame.This clearly indicates that our hypothesis is correct, since hypothesis stated that library hours affects the wishlist hours. Hence the null hypothesis is true.