Price Analysis of Games at The Epic Game Store

Authors: Akshita, Bella, and Shey

Section: CSE163 AJ

Table of Contents

1. Summary of questions and results

1. **How accurately can we predict the price of a game based on the various features present in the dataset? Which one is the most influential/accurate feature for predicting the price?**

   We use logistic regression model of machine learning to predict how features affect the price of the games and compare the accuracies of each feature. We also use bar graphs to visualize the accuracies of different features and provide a better understanding of the comparison between different features. We also compared the train accuracy and the test accuracy by combining all the features.

   Based on these scores and graphs, we can determine the genre feature that predicted the affordability of the video game in the testing dataset most accurately.

2. **Who are the top 3 game publishers? Which three game publishers publish the most expensive games for the company? What is the average price of a game for each game publisher?**

   After combining the datasets and filtering the columns we need, we used the groupby function and other functions to solve these three questions. We determined the top 3 game publishers, which are Ubisoft, THQ Nordic, and Devolver Digital. PublisherSquare Enix, Square Enix, and Ubisoft Entertainment published the most expensive games for the company. We also determined the average price of a game for the total of 373 game publishers. The result will be shown in the form of a series.

   In addition, we made a bar plot and a line plot to compare the over-time trends of numbers of games for different game genres from 2018 to 2022. Based on the charts, we can better determine the development trend of different game genres and have a basic understanding of each genre. We can see the difference in the number of games between different genres of game. In this way, we will be able to predict the future development trend of each game genre.

3. **How have game prices changed over 2018 to 2022 for different game genres? Compare trends in genres over time. Has the pandemic affected price trends for video games?**

For this question, we did some data analysis and made a visualization for it. We used seaborn library to make a scatter plot for showing the trend of price change from 2018 to 2022 with different game genres. We compared the trends based on the genres of Shooter, Horror, Strategy-Survival, Action Adventure, and Racing.

Based on the plots, we can find the trend and distribution pattern of game price change for five genres from 2018 to 2022 and make an analysis of it. Based on the chart, we can see the price change of each genre in these years.

Action Adventure is the most dominant genre and Racing is the least dominant genre. In the year 2018 to year 2022, the price of the Strategy-Survival genre games has been increasing. The price of other genre games seems to be constant over this period of time.

2. Motivation

The answers to the research questions above would be beneficial to both businesses and consumers. It makes price predictions for games for the benefit of customers. They can plan their budgets and determine the best time to purchase a game by knowing if the video game is affordable. By compiling statistics on game publishers and their games, businesses can compare data with other businesses and make adjustments to improve sales. By examining trends over time, people can learn how accidents or special events (such as epidemics) affect sales and devise strategies to avoid losing a lot of money.

3. Dataset

The dataset that our group is going to use is related to the games available on Epic Games Store. The Epic Games Store is a digital video game storefront for Microsoft Windows and macOS. And, we found this dataset on Kaggle. In this dataset, there are four csv files and the total column number is 70. Our group is going to join different files to one dataset file. In the dataset, we can get all the information about the games and apps in the Epic Games Store. We can know information such as the price of the game and when the game

was released. With these useful data and information in the dataset, our group has the confidence to make good analysis about the Epic Game Store.

URL of the dataset:

https://www.kaggle.com/datasets/ramjasmaurya/epic-games-store?select=epic-games-demo.csv

4. Method

General computations: Data combining and Cleaning. Combine all our data frames into a single dataframe. This will allow us to work with video games, apps and demo games. This will allow us to overcome our challenge of working with messy data.

- How accurately can we predict the price of a game based on the various features present in the dataset? Which features improve our accuracy for prediction?

  To answer this research question we would: 1) Clean the data we intend to use. 2) Choose an appropriate machine learning model to test accuracy 3) Decide which features are affordable by calculating the median values of price 4) Make a model that predicts the price by using features of genre, year, top critic average, features of game, critics recommend. 5) Get the accuracy scores and make bar graphs based on the accuracy data. 6) Compare how accurate our predicted prices are to the actual prices and compute an accuracy score. 7). Generate a dictionary to compare the result between test accuracy and train accuracy in terms of combining all the features. All these computations will help us understand how accurately we can predict the price of games based on factors like features, type, genre, cpu, gpu, memory, storage and publisher (specific columns used to build the model). This research question will also help overcome our challenge of using machine learning. We will also be using new libraries in python to perform logistic regression, because most of the columns used for prediction are categorical variables.

- Who are the top 3 game publishers? Which three game publishers publish the most expensive games for the company? What is the average price of a game for each game publisher?

To answer this research question we would: 1) Combine four datasets and drop the null values 2) Filter the columns we want 3) Group our data by all the unique
publishers in the dataset. 4) Determine the top 3 game publishers who published the most numbers of games and 3 game publishers that publish the most expensive games 5) Compute the average price for each group of publishers . All these computations will help us understand on average how games for each publisher are priced. For these computations we will focus on the publisher, price columns, and numbers of games. These computations will help us overcome the challenge of multiple datasets.

- How have game prices changed over 2018 to 2022 for different game genres?

To answer this research question we would: 1) Filter out our years of interest. 2) Group our data by five main genres for each year. 3) Plot our results in a chart to show trends in prices over time and use different colors to distinguish different genres on the graph. 5) Compare trends from the chart. All these computations help us visualize the trends in game prices over time across all genres. We will focus on the release date, price and genre columns to answer this research question. These computations will be related to our challenge of dealing with messy and multiple datasets to extract what we want for the trends.

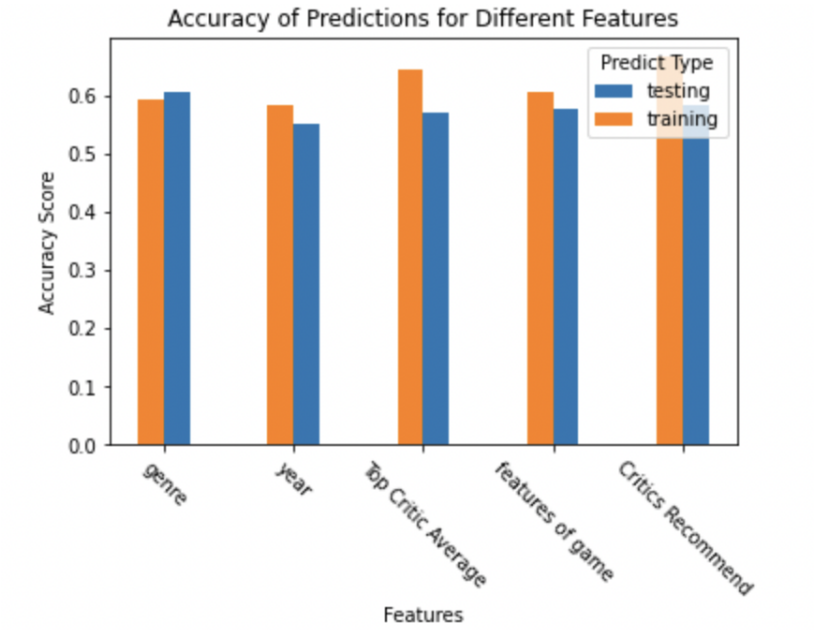
5. Results

1. **How accurately can we predict the price of a game based on the various features present in the dataset? Which one is the most influential/accurate feature for predicting the price?**

For these questions, we used machine learning to solve and we used bar graphs to show our results for comparison of accuracy between different features.
Based on our results, we can determine that these two features are the most influential separately in test and train accuracy. Critics Recommend and Genre are the most accurate features for predicting price in a test dataset.

```
({'Critics Recommend': 0.6661367249602543,
  'Top Critic Average': 0.643879173290938,
  'features of game': 0.6057233704292527,
  'genre': 0.5930047694753577,
  'year': 0.5834658187599364},
 {'Critics Recommend': 0.5822784810126582,
  'Top Critic Average': 0.569620253164557,
  'features of game': 0.5759493670886076,
  'genre': 0.6075949367088608,
  'year': 0.5506329113924051})
```

- First Dictionary: **Training Accuracy**
- Second Dictionary**: Testing Accuracy**



Every time the test and train accuracy will generate different values randomly, but the result pattern is the same. The most accurate feature for predicting is the same. In terms of the results, we can conclude that genre and critics recommend can be the most accurate factors in predicting the price of the games for this company.

Moreover, we also calculate the accuracy after combining all the features. We can see the pattern that the training accuracy is higher than the test accuracy. In addition, based on the dictionary of accuracy, we state that there are more than half numbers of correctly classified data points that are close to the total data points, which implies that our model may not be the most accurate one to predict if the video game is affordable since it achieved half of the accuracy in testing.

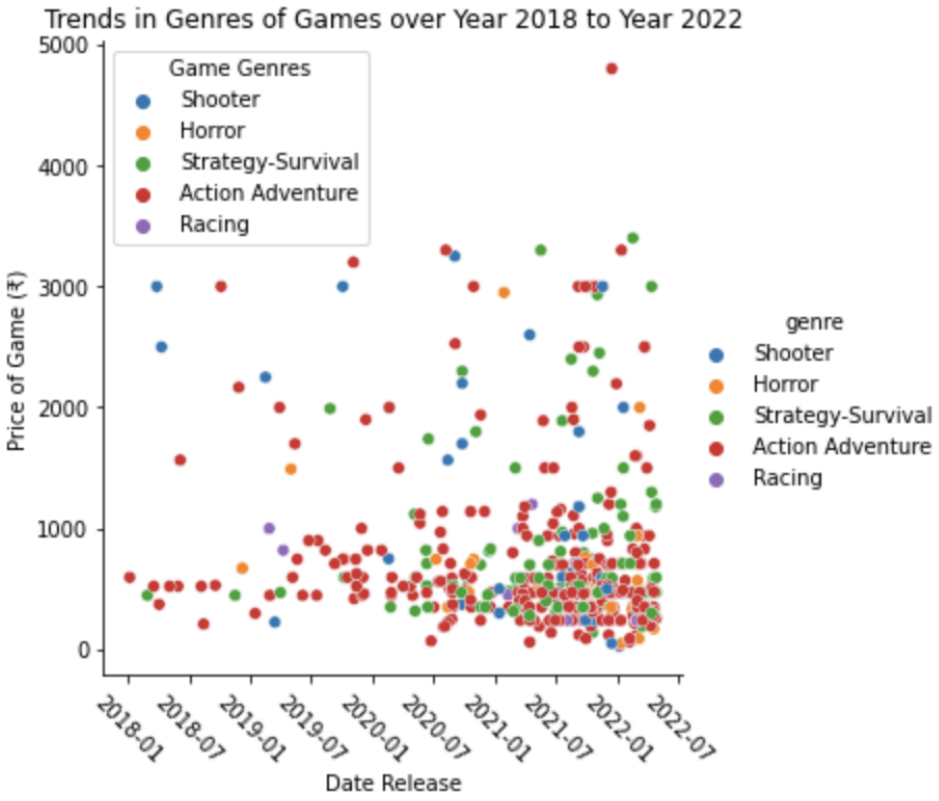
Specifically, it has a test accuracy of about 0.5254237288135594 and a training accuracy of about 0.6949152542372883.

2. **Who are the top 3 game publishers? Which three game publishers publish the most expensive games for the company? What is the average price of a game for each game publisher?**

   For this part, we just used the groupby function and other functions to solve these three questions. We determined the top 3 game publishers, which are Ubisoft, THQ Nordic, and Devolver Digital. PublisherSquare Enix, Square Enix, and Ubisoft Entertainment published the most expensive games for the company. We also determined the average price of a game for the total of 373 game publishers. The result will be shown in the form of a series.

3. **How have game prices changed over 2018 to 2022 (pandemic) for different game genres? Compare trends in genres over time.**

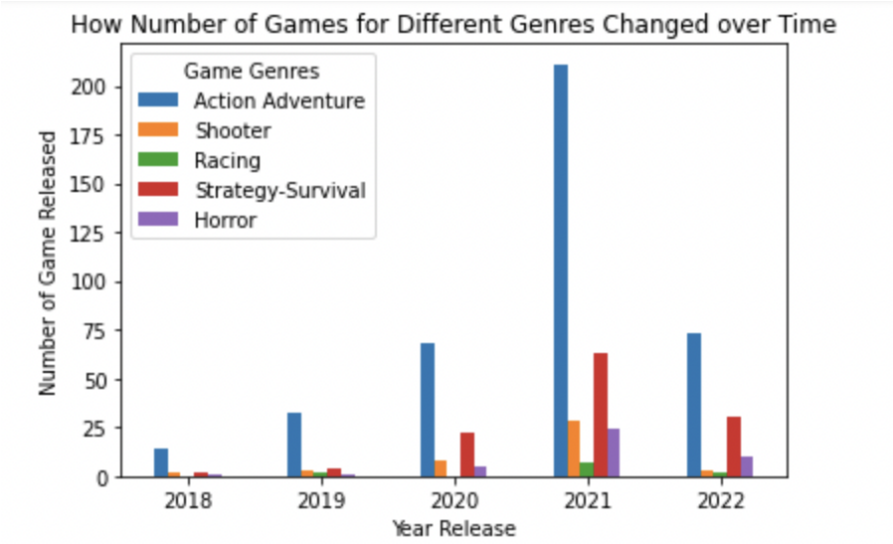
   For this question, we did some data analysis and made a visualization for it. We used seaborn library to make a scatter plot for showing the trend of price change from 2018 to 2022 with different game genres. We compared the trends based on five popular genres: Shooter, Horror, Strategy-Survival, Action Adventure, and Racing.

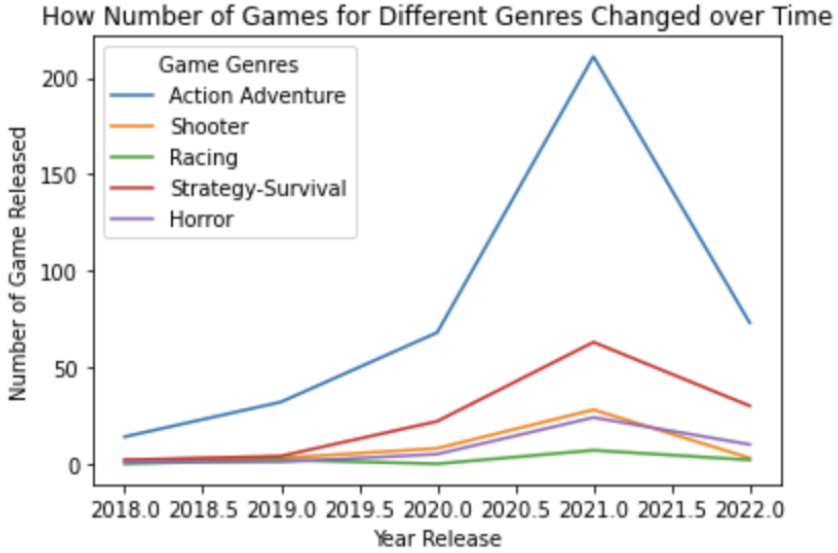


Based on the plots, we can determine that most game genres' prices are placed from 0 to 1000 from 2018 to 2022. Moreover, at the beginning, only action adventure has a price, but more game genres start to give price to the games in 2019. Based on this finding, our group suggests that this may be due to the game market growing from 2019 to 2022.

Besides these findings, we also find that Strategy-Survival and Action Adventure have the highest price of games,which is above 3000. Moreover, Racing doesn't seem to have as high prices as other game genres. The possible reason behind this might be that this game genre is not as popular as other game genres or the capital of this game is low.

Then, we made a bar chart and a line plot comparing the over-time trends of numbers of games for different game genres from 2018 to 2022.





Based on the charts, we can find that the genre of action adventure has the most number of games every year. In 2021, there is a peak for the numbers of games of all five game genres. We may conclude that the reason behind this trend is due to the effect of the pandemic. The company published more games since many people stayed at home and they had the need for new games for entertainment. Moreover, based on the two charts, we can determine that the most popular game genres are action adventure and strategy-survival. Racing has the same and stable trend from 2018 to 2022, which means that people don't have a strong interest in racing games.

6. Impact and Limitations

1. For the research question of game price trends over time, some other factors other than the epidemic could also affect the price trends over time, but we didn't take those into account. In other words, the reason why the price of the games changed during the year 2018 to year 2022 can be something more than the epidemic. As a result, our conclusions may have limitations to find the real factors behind the graphs.
2. For graphs we made for research question 2 and 3, we only take the five main game genres into account, which are Shooter, Horror, Strategy-Survival, Action Adventure, and Racing. In this way, we actually missed other game genres. This is one limitation since our results only can be applied to these five game genres.
3. In the machine learning part, we only compare the accuracy between five features, which may also cause some bias since we don't take all the features into account. The results of predicting can only be applied to these five features. Moreover, we only predict if the video games are affordable, but we haven't figured out the correct model for predicting the specific price of a game. If we can predict the approximate price of the future game, it can be more helpful and effective in helping people make decisions.

7. Challenge Goals

For this project, we plan to meet four challenge goals, which are **using multiple datasets, having messy data, and applying machine learning.** Among those goals, **machine learning** and **multiple datasets** are the most passionate challenge goals for us.

Our group is going to meet the goal of machine learning because we are going to use different types of models to predict the most popular genre of game in the Epic Game Store and evaluate which model is the best one to predict for our project. We will use models such as decision trees, linear regression, and support vector machines. We will try to compare the results of the models and see which model is the most accurate one.

For the challenge goal of multiple datasets, we have four dataset files provided by the

author. In order to get a detailed and complete analysis, we will join two or more datasets for analysis in the project. With more data and information, we can get more and more results for the final analysis of Epic Game Store by using Python.

During our implementation of the project, we actually met all the challenge goals. In the project, we firstly combine four datasets together and then clean the data by filtering the specific rows and columns we needed. We organized the data by dropping the null values, changing the format of the values for some columns, and creating new datasets for solving research questions. After finishing these parts, we have already reached the goals of **using multiple datasets and having messy data.**

When we tried to solve the first research questions, which needed the implementation of machine learning. We tried to use a supervised model of machine learning to find the five affordable features and then test the accuracy of the price by using these features. We got the test and train accuracy for these features. Every time, the results will be different since the data are split. We made a bar graph for combining the results of each feature so we can compare the results between different features and determine the most influential feature for predicting the game price. Finally, we also use machine learning to compare the accuracy between test and train in terms of combining all the features. That is how we met the goal of using **machine learning**.

8. Work Plan Evaluation

In general, most parts of the work plan are executed accurately in terms of time spent, while others are not. Overall, it's a good work plan because the only part we didn't finish on time was the machine learning part.

The process of development setup, data cleaning, and summary analysis went on smoothly. They didn’t take as much time as we expected, in fact, we spent about 1 hour less on each of them. These are due to the fact that we have a relatively deep understanding of the content of the dataset and we're familiar with dealing with those preparation steps.

The visualization and machine learning portions, on the other hand, took a lot longer than we anticipated.

For the visualization section, our draft code generated a graph that is not visually appealing, so we tried to add some more details, adjust the data used, color, title, and

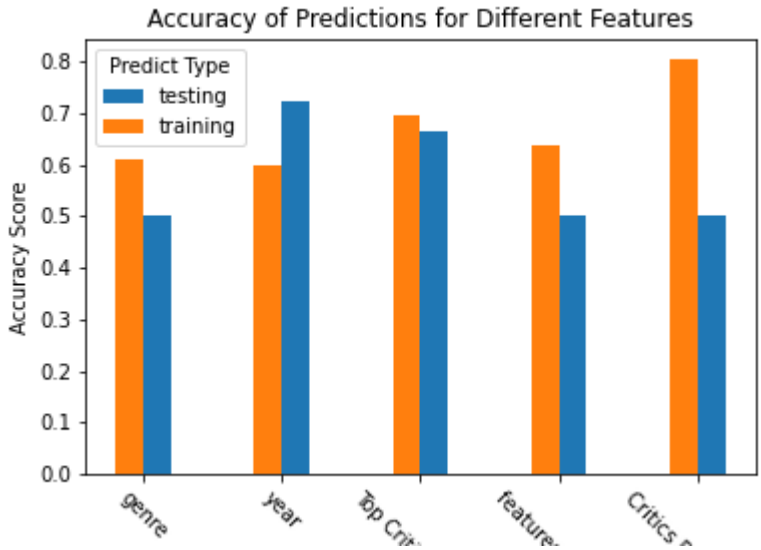
labels of the graph to make it look nicer. It took us 2 hours more than we expected. For the machine learning part, our proposed time is about 7 hours. However, we worked on it for about 2 more hours. The reason is that we are not that familiar with machine learning, and we spent some more time on debugging, finding the appropriate model, and summarizing the results (plotting graphs).

The underestimation of time used for the last two parts put us in a hurry on the last two days. Therefore, it may result in imperfection of these 2 parts. If we have more time, we're willing to take a deeper look and refine the model and results, such as predicting the approximate price of games instead of just predicting if the video game is affordable or not.

9. Testing

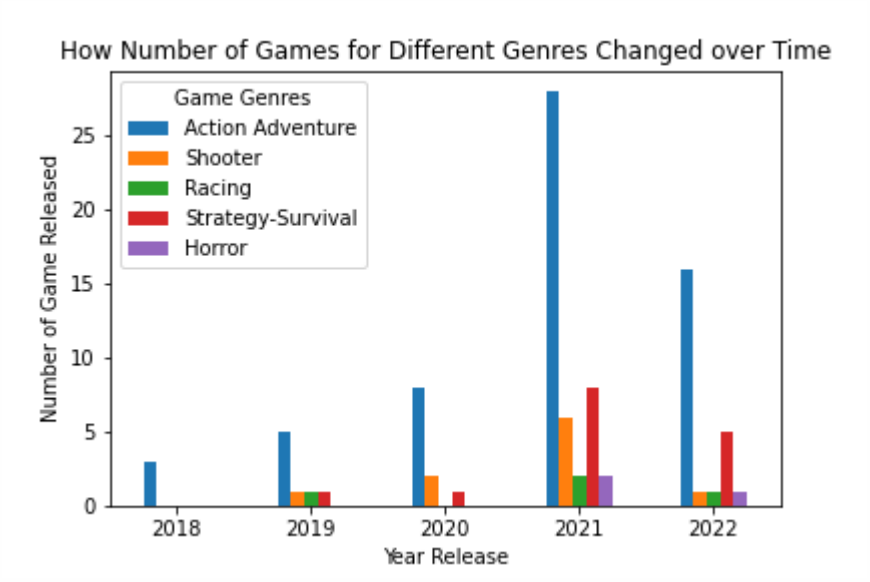
We made one test file for this project.

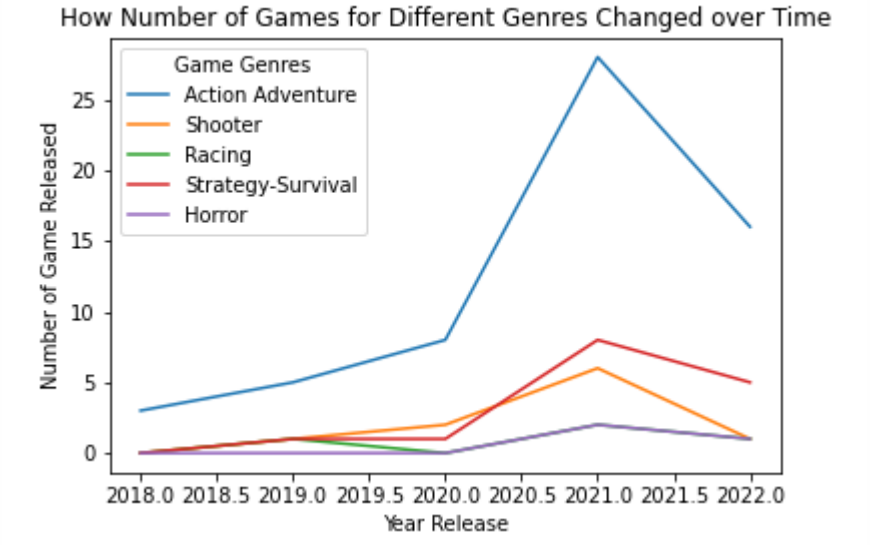
For testing the result of research questions 1, we selected 100 rows from the cleaned data. We used these 100 rows to test the results of using machine learning. The bar graph we got has a similar pattern with the actual graph so we suggest that the testing result of question 1 is accurate.

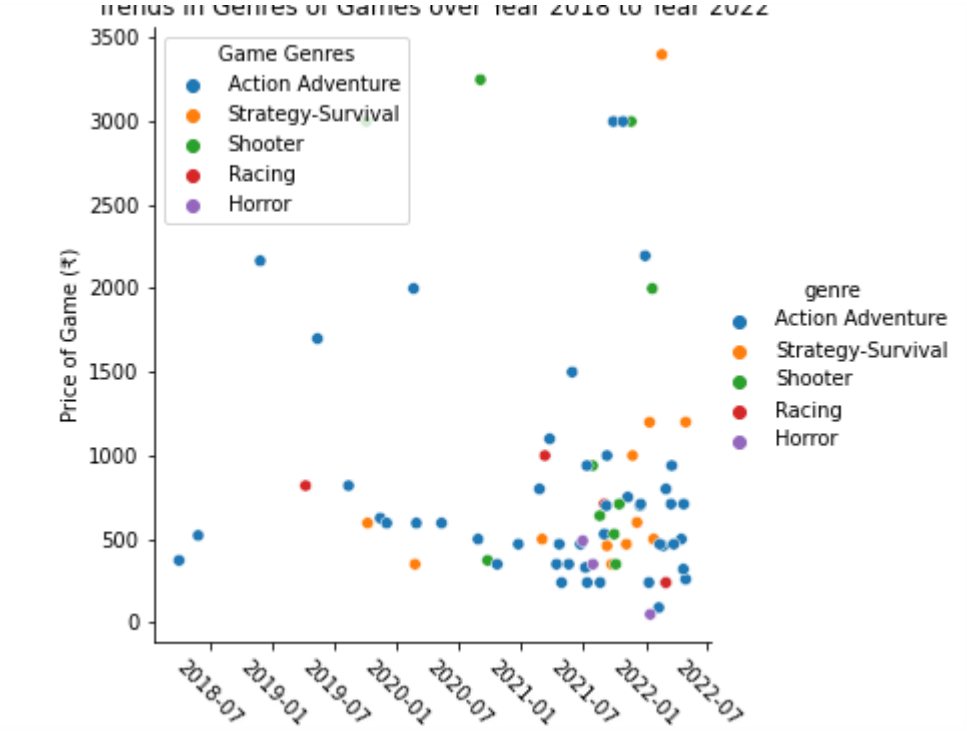


For testing the result of research question 2, we selected 10 rows from the cleaned data after importing the file of cleaning data. We used these 10 rows to test which publishers
are the top 3 game publishers in this company. In this part, we also used the assert equals statements since we need to make sure that the tested results are correct. After we check that the publishers are matching with the actual results, we will compare the actual average prices with the tested average prices. In this way, we can make sure that the actual results we got are correct.

For testing research question 3, we selected 100 rows from the cleaned data. We used these 100 rows to test the results of over time trends of price change of games with different genres. We also made three visualizations as we did in the main file to compare the tested results. The graphs show the similar pattern as the actual graphs did, as a result, we can confirm that our actual results of question 3 are right in the report.


How Number of Games for Different Genres Changed over Time
Game Genres
Action Adventure
Shooter
Racing
Strategy-Survival
Horror
Number of Game Released
0
5
10
15
20
25
2018
2019
2020
2021
2022
Year Release


How Number of Games for Different Genres Changed over Time
Game Genres
Action Adventure
Shooter
Racing
Strategy-Survival
Horror
Number of Game Released
0
5
10
15
20
25
2018.0 2018.5 2019.0 2019.5 2020.0 2020.5 2021.0 2021.5 2022.0
Year Release



Overall, although we used fewer data than the actual data we used in the main file, the results are still similar. As a result, our actual data is reliable and accurate.

10. Collaboration

The other resources we used for this project:

(1) This is an online resource we used to make visualizations (graphs) of the machine learning process.

Data Visualization using Python for Machine Learning and Data science : | by Sanat

(2) Our group got some inspiration from a past cse163 project. Specifically, it gives us an idea of how to use machine learning to predict the price of games with various features in the dataset.

GitHub - RitikShah/cse163-project: Project for CSE 163