The Anatomy of Video Games: A Comprehensive Analysis of the Video Game Industry

Project Authors

- Moyi Li (2136213)
- Jingjing Dong(2161424)
- Zhikai Li(2138341)

Summary of research questions and result

Question 1: How are the different game genres distributed, and which genre of video games is the most popular of all time?

Description: We are going to figure out the proportional distribution of genres in terms of their numbers of video games. The analysis will not distinguish the game release time, release platform, and sales region, but only focus on the single variable genre.

Result:

The Action genre is the most popular, making up about one fifth of all games, followed by Role-Playing and Shooter genres. Fighting is the least popular genre. The remaining genres are more evenly distributed, with nearly half of all games falling into these categories.

Question 2: How sales in each region vary for the top 5 publishers of all games?

Description: By analyzing our dataset, we observe that the regions where game sales are divided into North America, Europe, Japan, and other regions of the

world. To identify the top 5 publishers, we will consider the ones that have published the highest number of video games. Our goal is to examine the sales distribution of each of these publishers.

Results:

The top 5 publishers are Ubisoft, Sega, Nintendo, Electronic Arts, and Activision.

All five publishers have a large percentage of sales in North America and Europe, and a smaller percentage of sales in the Japan region and other regions.

Question 3: Are we able to create an accurate model to predict the sales of games based on our data with multiple features, such as by using the platform of the game, the publisher of the game, and the genre of the game? If we can do that, what is the most important feature that influences the sales of the game? If we cannot, why?

Description: Logistically speaking, we consider the platform, publisher, publishing year and game genre to be the primary factors that may influence game sales. The objective of this question is to examine whether we can predict game sales with high accuracy and, if so, which factor influences sales the most. If we cannot create a model to make an accurate prediction of the sales of games, what are the potential reasons behind and what does this mean?

Result:

By comparing and observing the result from our training model. We found out we are not able to create an accurate model to predict the sales of games based on our data. The result shows that the prediction of our model is only somewhat accurate, which means that through our data, we can not accurately predict the global sales of the game in the world only through the game publisher, publishing platforms, game genres and the publishing time. The reasons behind this are that the proportion of low-sales games in data is too large, which affects the accuracy of model training. Also the features used for model training, except for the column of game sales, do not have a strong linear relationship with the global sales of games, so it is difficult to predict accurately through these features.

Question 4: What is the relationship between a video game's production excellence and total sales?

Description: For this problem, we would like to consider the video game's production excellence in terms of the meta_score and user_score. We are also going to figure out which type of score is more accurate in assessing the relationship with sales.

Result:

From analyzing both the regression line chart and the calculation for evaluating the linear regression model, there is a weak relationship between video games' production excellence and total sales.

Question 5: How have the players' preferences in various genres of video games changed over time?

Description: For this question, we will define the player's preferences based on the total global_sales. We would like to investigate the variation of the sales on different genres from 1996 to 2015. In this problem, we plan to investigate the trend and any potential causes for such trends.

Results:

We chose the top three genres for analyzing the changement in Global Sales. The results show that in total, Shooter games drastically decreased but Action games and Role-play games have fluctuated a lot but overall slightly decreased.

Motivation

With the development of technology, video games have become an important part of people's entertainment. We cannot deny how video games have become a part of people's lives and thus affect them. Video games are not only one of the ways in which people are entertained, but they are also becoming an important part of economic development, as the gaming trade accounts for a large part of the world's economic development. The relation between the year of release of

the same game, the platform of release, the different types of video games, and the audience of the game are all directions worth analyzing and thinking about.

Additionally, video game companies are involved in the job market as well. As learning computers and programming becomes a popular trend, it is becoming worthwhile for university students like us to think about how games will affect the market, both before and after graduation. Therefore, our main motivation for this project is to analyze how different categories, publishing platforms, years, and other factors affect the direction of the game's audience in order to gain a better understanding of the game's audience. This will facilitate our analysis of the gaming economy and our future career direction.

Dataset

We are going to conduct research based on two datasets. For investigating most of the research questions, we are going to primarily based on the first dataset Video Game Sales. The dataset includes 16598 rows and 11 columns, including the name, releasing platform, year, genre, publisher, sales in global and each region, and rank of overall sales.

The second dataset Metacritic Games includes 19992 rows and 6 columns, including variables name, platform, release time, a short summary for each game, and meta_score and user_score for evaluating each game. Specifically, We will mainly use meta_score and user_score for evaluating the production excellence. These two columns are from the well-known game review website Metacritic. The meta_score represents the weighted average of scores given by selected critics or publications, while the user_score represents the average of scores given by users who have submitted ratings and reviews on the website.

In investigating the research questions, we are going to merge these two dataset based on the names of the video games to collect information for both sales and scores for production excellence.

URL for Video Game Sales:

https://www.kaggle.com/datasets/gregorut/videogamesales

URL for Metacritic Games:

https://www.kaggle.com/datasets/henrylin03/metacritic-games-user-reviews-and-metascores

Method

Question 1: Which genre of video games is the most popular of all times?

Since we are only focusing on the single variable genre, we used the pie chart with a new library in Plotly to visually display data for all the genres of different games where our data exists. After groupby the table by genre, we imported the plotly.graph_objects to create an interactive pie chart. It automatically shows the genre name, specific counts of games, and the percentage of that genre while hovering. After that, we can intuitively see which genre has the highest proportion with the overall proportional distribution, with the overall distribution of all genres.

Question 2: How sales in each region vary for the top 5 publishers of all games?

First of all, we can use the filtering in Pandas to get the five game manufacturers that sell the most games. Instead of using geopandas to create geomaps, we used a horizontally stacked bar chart to show the proportional distribution for each top publisher. The top five publishers are shown on the y-axis, and the percentage of sales in each region is shown on the x-axis. The sales in each region are shown by the different stack colors of each bar. This can intuitively show the proportional distribution of games sold by different game manufacturers in different regions around the world.

Question 3: Are we able to create an accurate model to predict the sales of games based on our data, such as by using the platform of the game, the publisher of the game, and the genre of the game? If we can do that, what is the most important feature that influences the sales of the game? If we cannot, why?

Because different kinds of factors or characteristics of the games themselves might affect the sales of games, we want to try whether we can accurately predict the sales of a game with multiple features through training with our data. We determined the features that we thought might affect the sales before the game is released, such as the game genres, the platform for the game release, game publishing time and the game publisher. We filtered our data. After allocating the part we train and test the data and converting the non-numeric values in our column, we established our machine learning model to do our regression problem. We chose to use a sequential model in Keras to achieve our challenge goal, and also wanted to use this model to achieve more accurate machine learning prediction. After adding the hidden layer and output layer, we conducted training. We adjusted the number of times to iterate over the entire training dataset and the number of samples that are used by the model each time to observe whether our training will have the phenomenon of underfitting or overfitting. At the same time, we also adjusted the complexity of our model according to the final test results and the time spent in training. We know the difference between our prediction after training and the real data by comparing the mean square error.

Question 4: What is the relationship between a video game's production excellence and total sales?

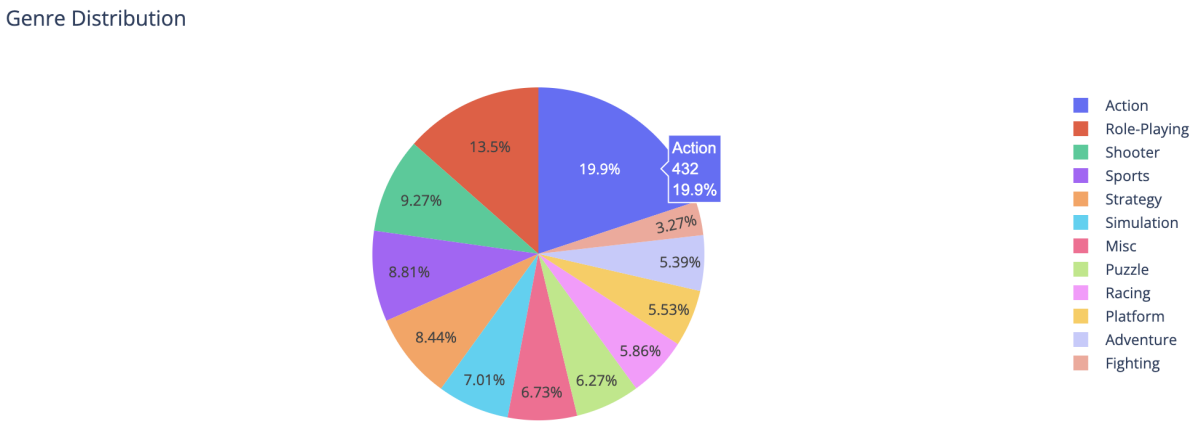
To solve this problem, we need to involve multiple data sets. The two dataset are merged by game names and platforms since both these two dataset have these two variables. After merging the dataset, we conducted a linear analysis of the two different scores and the actual game sales. We first used all the combined data to plot the regression line but the results are difficult to view since the points are all clustered at the bottom and the line is not obvious. We then filtered the data to ensure that worldwide sales were between 0.5 and 10, allowing us to clearly see the regression pattern. The results are shown in subplots to better compare the difference between meta score vs global sales and users score vs global sales. Finally, the coefficient of determination of different regression models is calculated by using a linear regression model from sklearn.linear_model for better analyzing the regression level.

Question 5: How have the players' preferences in various genres of video games changed over time?

We investigated the player's preferences based on the total global_sales column from our data, since the number of sales directly reflects the player's preference. In this problem, we created a line graph with the game publishing time as the x variable, global_sales as the y variable, and genres differentiated by different colored lines. Since there are more than 10 different genres, it is difficult to view all of them with the lines clustered together. Thus, we selected the top 3 genres from observing the pie chart. Moreover, since the data ended in 2015 and we plan to analyze the data for video games in 20 years, we filtered the year range to be 1996 to 2015. Finally, by using seaborn and matplotlib, the line chart was created. The line chart could directly show how players' preferences for top 3 game genres change from 1996 to 2015 for answering the research problem five.

Results

Question 1:



Overall, the pie chart provides insight into the distribution of different game genres and highlights the most and least popular categories. According to the pie chart, the Action genre is the most popular, accounting for approximately 20% of all games. This is followed by Role-Playing and Shooter genres, which make up about 11% and 9% of all games, respectively. On the other hand, Fighting has the lowest proportion at 3.3%.

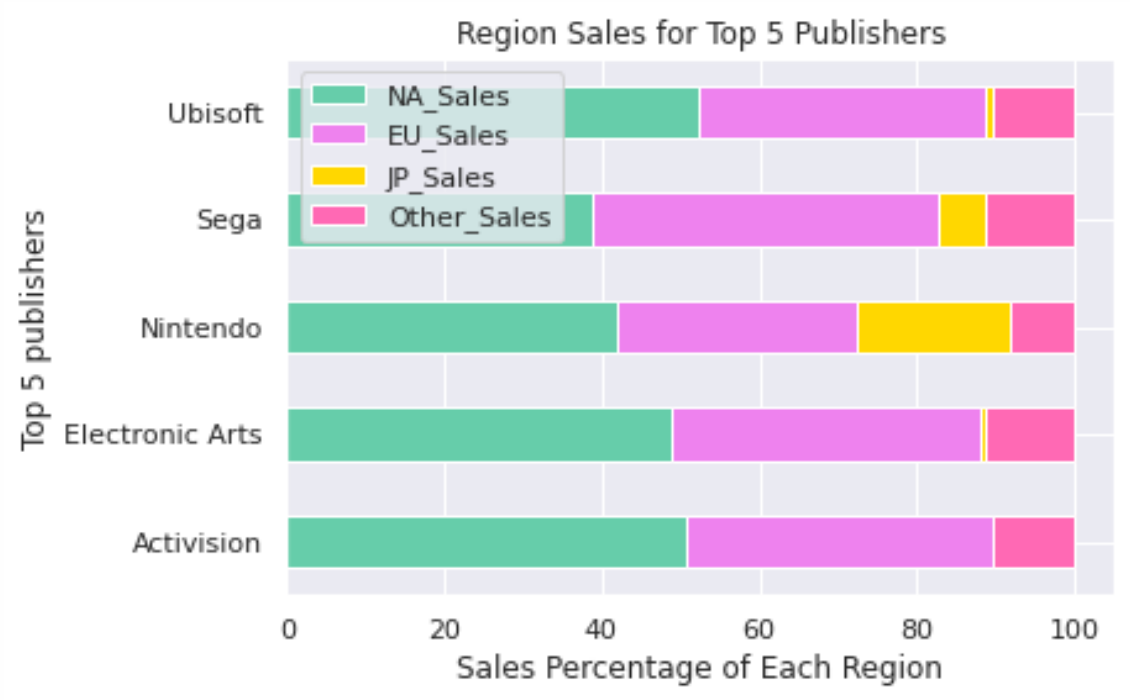
The remaining genres, including Sports, Strategy, Simulation, Misc, Puzzle, Racing, Platform, and Adventure, are more evenly distributed, ranging from about 5% to 8%. These genres collectively account for slightly over half of all games. It suggests that

there is still significant demand for games across a wide range of genres, and that game developers may still benefit from catering to a diverse range of player interests.

It is worth noting that the top four genres, Action, Role-Playing, Shooter, and Sports collectively make up half of all games, indicating a preference for these types of games among players. It may indicate a preference for games with fast-paced, sports-like, interactive gameplay and immersive storylines, which are often hallmarks of these genres. Therefore, for game publishers, targeting these genres, especially Action, may be able to potentially attract more players. However, the significant attention given to these types of games also implies that there is a large number of games in these categories, making it more difficult for publishers to stand out in a crowded market.

Meanwhile, the Fighting genre's low proportion suggests that it may be less popular or less in-demand among gamers. It may be due to a simple shift in player preferences or a lack of innovations in Fighting games. However, this also highlights that there are fewer games in this category, making it possible for innovative gameplay and game design to make a significant impact in this genre.

Question 2:



We discovered that the five leading publishers in video game sales are Ubisoft, Sega, Nintendo, Electronic Arts, and Activision. These publishers predominantly generate sales in North America and Europe, with lower sales figures in Japan and other regions.

Among these publishers, Nintendo and Sega have the most prominent sales in Japan, while others have very few sales. Instead of Nintendo, the rest of the publishers all have over 80% sales in North America and Europe, where the sales in each region are roughly the same. Across all of the top 5 publishers, sales from other regions comprise approximately 10% of their total sales.

Initially, we indeed predicted that the majority of game sales would come from North America and Europe. Considering the populations of these regions (580 million in North America and 746 million in Europe), the roughly similar ratio on the plot aligns with our initial predictions. The comparatively lower sales in Japan (with a population of 125 million) can be explained by its smaller population size as well. However, we were surprised to find that the NA and EU regions accounted for a significantly higher percentage of sales than expected. Even Nintendo, a Japanese company, only represented around 10% of sales in Japan. We also expected countries like China to make up a larger portion of sales in the "other regions", as they have a higher population of 1.4 billion.

The result can be interpreted for the following reasons. Firstly, as we are only focusing on the top 5 publishers in the world, such a small data size does not have wide applicability. Secondly, we investigated the data's origins and found the video games in the table were only defined as console games of following lists. Thus, considering China as a developing country and that the number of Chinese console game players was only 15.9 million in 2021, the small proportion of other regions can be explained.

Question 3:

```
Epoch 1/150
408/408 - 2s - loss: 742.6533 - mse: 742.6533 - val_loss: 1.7360 - val_mse: 1.7360 - 2s/epoch - 4ms/step
Epoch 2/150
408/408 - 1s - loss: 2.6194 - mse: 2.6194 - val_loss: 1.7287 - val_mse: 1.7287 - 709ms/epoch - 2ms/step
Epoch 3/150
408/408 - 1s - loss: 2.5769 - mse: 2.5769 - val_loss: 1.6897 - val_mse: 1.6897 - 737ms/epoch - 2ms/step
408/408 - 1s - loss: 2.4272 - mse: 2.4272 - val_loss: 1.4864 - val_mse: 1.4864 - 802ms/epoch - 2ms/step
Epoch 149/150
408/408 - 1s - loss: 2.3777 - mse: 2.3777 - val_loss: 1.4848 - val_mse: 1.4848 - 744ms/epoch - 2ms/step
Epoch 150/150
408/408 - 1s - loss: 2.3964 - mse: 2.3964 - val_loss: 1.4804 - val_mse: 1.4804 - 721ms/epoch - 2ms/step
102/102 [==============================] - 0s 1ms/step
Test set mean squared error: 1.4803521131012367
```

After we continued to test, adjust, test, observe, and adjust for a lot of times, we can confirm that the global sales prediction of our model after training is not very accurate. This is a surprising and interesting result. While the data set we first use has more than 16000 rows,the mean square error of our prediction and real data fluctuates between 1 and 2, occasionally more than 3. (Note that the unit of game sales in our data is millions). The mean square error between 1 and 2 seems to be not large for the game

with high sales volume of our data set. For example, the first Wii Sports sold 82 million, and the second Super Mario Bros sold more than 40 million. However, it cannot be ignored that about 85% of the game sales in our data set are less than 1 million, which means that our machine prediction is not accurate for most of the examples in our data.

In order to further improve the prediction accuracy, we have tried many methods, including but not limited to changing the neurons in hidden layers, adding more hidden layers, adjusting the number of times to iterate over the training dataset, and adjusting the number of samples that are used by the model each time. We can see from the training process of our model that the differences between predicted and real output gradually decrease with the number of training, which means that it is true that the model can learn to predict something from the dataset.

At the same time, we control the number of times of training to ensure that there is no overlapping. In the process, our predicted results become relatively stable than before, but still the best result does not break below 1.0. Therefore, we used the regression model from the sklearn that we learned from class and assignment to test ourselves privately. Through the same features and labels, we find that the mean square error ratio is a little larger than our sequential model most of the time, which means it is more inaccurate.

This is expected because the model we use is relatively more complex, but it means that it is likely that there are other potential reasons that lead to inaccurate prediction. We think one of the biggest potential reasons is that in our data set, the data of low-sales games accounts for too much. As mentioned above, 85% of the games have not sold more than 1 million globally. It is also worth noting that there is also a part of the data in our data set with a global sales volume of 0.01 million, which is likely to be an important factor affecting machine learning and model training. Thus, we try to use the filtered dataset after this dataset and our second dataset merge. Only more than 2000 rows in this dataset have filtered the popularity of the game at the same time, because it only contains the data corresponding to the two datasets. We also added the name, meta scores and user scores column as our features, because we considered that the name of the game when it was released might indeed become a factor for players to buy the game. For example, the name of a famous game sequel can guarantee a certain number of sales. We also added meta score and user score, because we believe that the score after the game is released can affect subsequent players' desire to buy the game.

For the test of the second data set, because the data set is smaller, we adjusted the number of training times, and the number of samples that were used by the model each

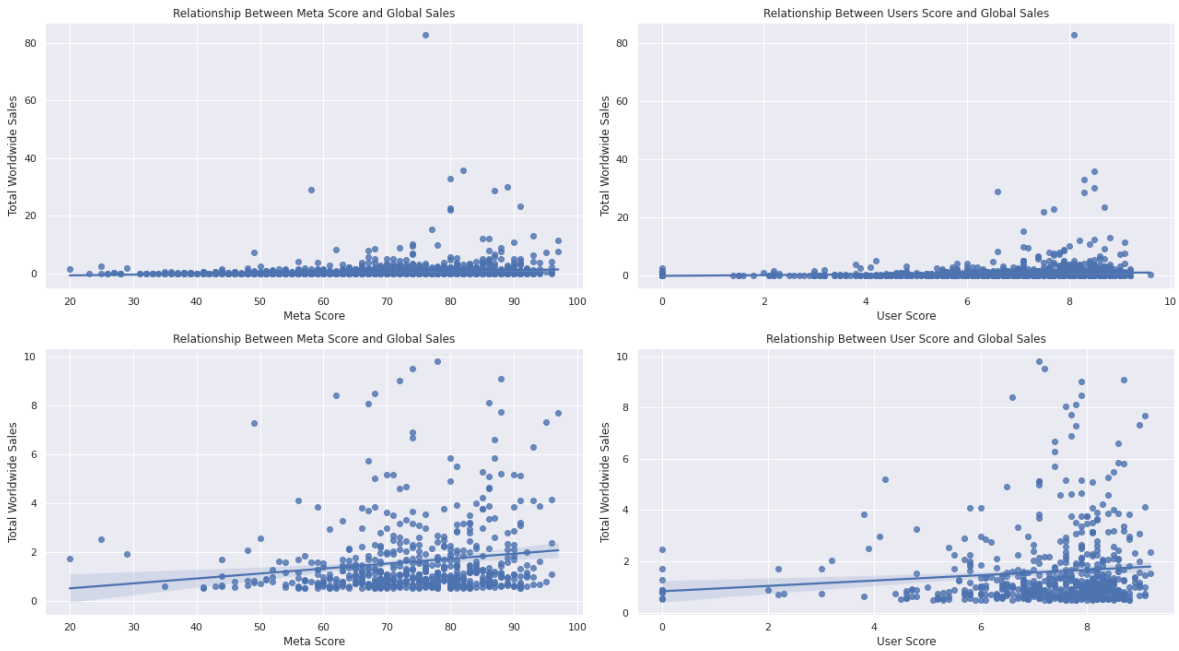
time, and we also added one more hidden layer to try to increase the accuracy of prediction. After many tests, observations and adjustments of our model, the actual improvement of the prediction results is not ideal. The prediction results are more inaccurate and unstable than before. We analyzed the reason for this is that this dataset has fewer samples than before, and also because it still contains a large number of low-sales games. In research question 4, we measured the game score and the number of games sold. The result is that their relationship is relatively weak, so the improvement of the prediction results is very limited. We also tested it with sklearn's expression model, and the results were similar and also less accurate.

```
Epoch 1/80
200/200 - 1s - loss: 133.4660 - mse: 133.4660 - val_loss: 2.7100 - val_mse: 2.7100 - 1s/epoch - 6ms/step
Epoch 2/80
200/200 - 0s - loss: 15.5434 - mse: 15.5434 - val_loss: 4.5489 - val_mse: 4.5489 - 422ms/epoch - 2ms/step
Epoch 3/80
200/200 - 0s - loss: 12.6807 - mse: 12.6807 - val_loss: 2.7520 - val_mse: 2.7520 - 423ms/epoch - 2ms/step

Epoch 78/80
200/200 - 0s - loss: 9.5678 - mse: 9.5678 - val_loss: 1.8586 - val_mse: 1.8586 - 422ms/epoch - 2ms/step
Epoch 79/80
200/200 - 0s - loss: 9.5670 - mse: 9.5670 - val_loss: 1.8583 - val_mse: 1.8583 - 429ms/epoch - 2ms/step
Epoch 80/80
200/200 - 0s - loss: 9.5669 - mse: 9.5669 - val_loss: 1.8503 - val_mse: 1.8503 - 435ms/epoch - 2ms/step
13/13 [==============================] - 0s 2ms/step
Test set mean squared error for main analysis using final_result: 1.8502745732109016
Test set mean squared error with sklearn: 6.67878475
```

Therefore, we can conclude that through our data set, we cannot create an accurate model to predict the sales of games based on our data with features that are not the sales of games. The main reason behind this is that low-sales games take up a large proportion of our data set. Our result means that the name, publisher, gene, publishing year of the game can affect the sales volume of the game to some extent, but we cannot predict the sales volume of a game because of these features. The popularity of a game cannot be determined by these factors.

Question 4:




```
coefficients of determination for meta score before filtering: 0.01641231114089503
coefficients of determination for user score before filtering: 0.008957905181321557
coefficients of determination for meta score after filtering: 0.023935902156589806
coefficients of determination for user score after filtering: 0.010689561739393061
```

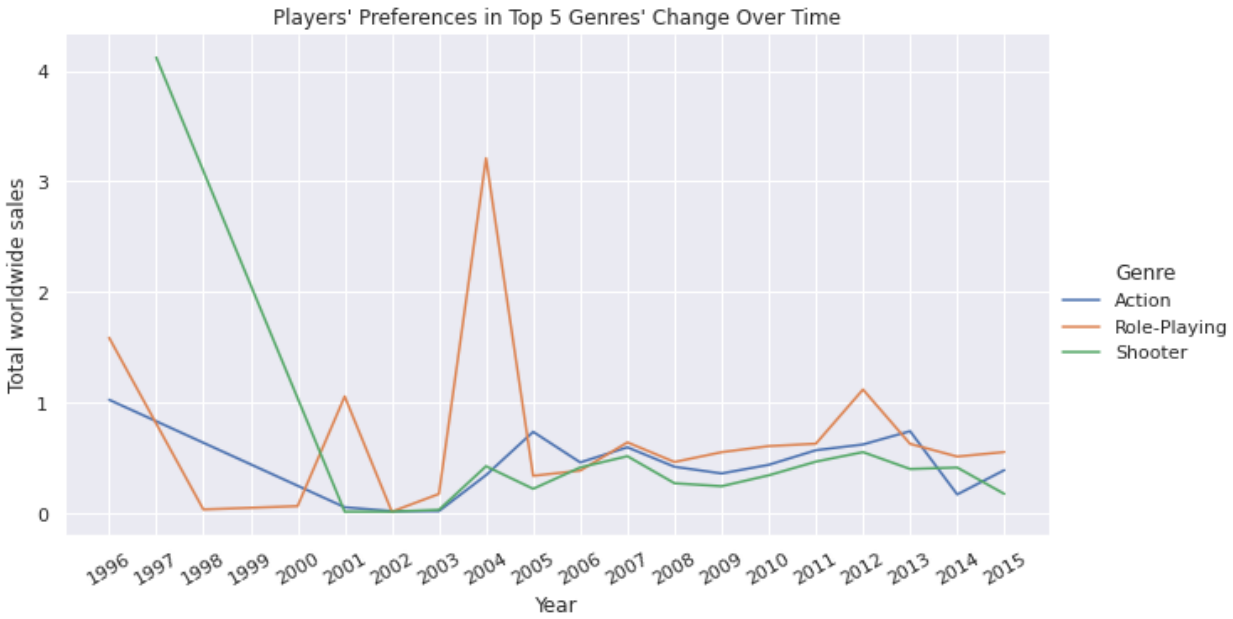
The two plots located at the top of the graph are regression line plots with all game data selected for analysis. From the graphs, we can see that a large number of points are clustered in the range of 0-20 for global sales, causing us to view an unclear regression line. The regression line obtained for the full data set has no obvious regression trend, indicating that there is no close relationship between Meta Score or User Score and global sales. We want to further exclude the effect of higher and lower values on the regression line, and show the regression relationship with a clearer regression line graph. Thus, we filtered all the data with global sales between 0.5 and 10 and plotted the two regression lines shown in the plot at the bottom of the figure. These two plots have a clearer regression line, but still do not show a closer regression relationship. In order to accurately determine the level of regression, we calculated the coefficient of determination. The coefficient of determination is between 0.0 and 1.0. In general, the higher the coefficients of determination, the better the model fits the data. From our screenshots, we can see that the coefficient of determination is small, even less than 0.1, both before and after filtering. This shows that there is a weak regression relationship between Meta score and User score on global sales.

These findings are surprising because our group initially believed that high scores from critics and users correlate to greater worldwide sales of video games. This research, however, indicates that this is not always the case. The diverse character of the video

game business could be one reason for this absence of correlation. There are numerous game categories, platforms, and styles, and what attracts reviewers and users may or may not transfer to economic success. Furthermore, other variables such as marketing, release timing, and availability on various platforms may have a larger influence on worldwide sales than review ratings.

These findings have significant consequences for game developers and publishers. While receiving high ratings from reviewers and users is a source of satisfaction, it does not always convert into increased sales. Instead, developers and publishers should concentrate on variables such as marketing strategies, releasing time, and releasing platform, which may have a larger influence on worldwide sales. Moreover, rather than depending extremely on review ratings as a sign of possible success, they may need to consider the unique features of their particular game and intended audience.

Question 5:



As shown in the line chart, from 1996 to 2015, the Shooter game dramatically decreased which became the game genre with lowest global sales of the top three genres. As for Role-playing and action games, there's also a slight decrease. Overall, the Action game is less volatile and tends to develop smoothly; the Shooter game decreases the most and tends to develop smoothly after 2005; the Role-playing game

tends to fluctuate more. When analyzing the peak, both Shooter and Action game reached its peak before 2000 while the role-playing game reached its peak in 2004.

It is interesting to note that the different genres of game reach their peaks in different time periods. This suggests that the popularity of game genres tends to fluctuate over time and is not necessarily stable. Another interesting finding is that Role-playing games are more unstable than the other two types of games. This could be because Role-playing games have a bigger niche audience than Shooters or Action games, so their popularity is more influenced by market trends and shifts in customer preferences. The growth of other game categories, such as Combat games, which have become increasingly popular in recent years, could explain the decrease in sales of Shooter games. Another factor that may be discouraging some consumers from purchasing shooter games is the increasing worry about violence in video games and its possible effect on players.

Overall, the findings of this study indicate that game genres' popularity fluctuates over time, and that revenue patterns in various genres can be affected by a variety of variables. Understanding these patterns and factors may be critical for future game developers and marketers seeking to build and push great games.

Impact and Limitations

Question 1:
The potential impact of the results is on the game's distribution merchants. Specifically, those game merchants who have not yet identified a game genre may benefit from our analysis. Publishers who do not understand the market and rush headlong into the game space may be hurt. The analysis may also exclude or harm certain groups, particularly those who prefer genres that are less represented in the chart. They may feel overlooked or underserved by the industry, leading to a lack of diversity in the gaming market.

One potential limitation of the analysis is that it only represents the gaming industry at a specific time range. The data from the latest years may change as new games are released and player preferences shift. Additionally, the chart only includes games that have been categorized into specific genres, which may not capture the nuances of individual games or sub-genres.

Another potential bias is that the data only include the distribution of genres, without including any sales information. In that way, It is possible that there is no direct correlation between the number of games and their sales. For instance, the Fighting genre with the lowest number of games may not necessarily have the lowest sales figures. As a result, individuals seeking to gain insights into game sales should not rely solely on the results of this analysis. Instead, they should consider other research questions to understand the relationship between game genres and sales.

Question 2:

Firstly, for the top five game publishers, the plot highlights the importance of NA and EU as major markets for video game sales. Thus, it is reasonable for marketing departments to allocate more funds to promoting advertising resources in these two regions. In addition, all the publishers can learn from the top 5 publishers' marketing successes and industry trends, such as increasing the advertising of the NA and EU regions in a similar manner. However, this may also cause harm to companies that are other than the top 5 on the stacked bar plot because the plot does not survey all game publishers, and there may be companies that mainly operate in Japan or other regions that use this conclusion, resulting in wrong publicity and thus suffering revenue harm. Therefore, for other game makers, we do not recommend using this conclusion.

Focused on the limitation, upon further investigation, we discovered that the data in the table "vgsales" was collected through web scraping from website VGChartz, but VGChartz stopped producing sales estimates for video game software after 2018. Consequently, the data on software sales comes only from official shipment/sales data provided by developers and publishers. This may result in some inconsistencies between the recorded data and actual sales figures, making the findings presented in the stacked bar plot surprising.

Additionally, we learned that the original data set included a column for PAL sales, which refers to a television broadcasting system used in many countries. However, since this data was not included in table "vgsales," the percentage of sales attributed to PAL using regions may not be entirely accurate.

Question 3:

Our results for the third research question can have implications for many game developers or private game studios who worry that some certain types or themes of

games are unpopular or hard to sell. Especially for small game studios, it is normal to worry that they are not well-known game publishers, resulting in poor sales of games. Our results show that no matter the name of the game, genre, publishing time, publisher or publishing platform, these factors cannot determine the number of games sold. Even meta score and player score still cannot. Many games with high ratings are not selling well, at the same time, many games with low scores still sell a lot. These results can solve the concerns of game developers or private studios: because there is no lack of players in any type of game sold on any kind of the platform, developers and private studios can focus on the specific game quality in order to sell more.

There can exist bias in our results and data, since although the data set we use is sufficient with samples, it still cannot represent the entire game market. In addition, the machine learning model we use for testing and training is not very complex and it does not guarantee 100% accuracy. There may be other methods, or model training can achieve better results.

For this research result, one of the biggest potential limitations for this research result, and probably one of the reasons that affects our results, is that most of the game sales samples in the dataset that we used are very small, and the smallest accurate unit in our data regarding the number of sales is 0.1 million, which is not very accurate. Therefore, this may not only affect the results of the training model, but also bring our results many limitations, because the range of sales considered is too large. If a game publisher can be sure that the number of games sold is within a certain range, it is likely that for them, the publishing platform, year, and genre of the game will still have a great impact on the number of games sold. In addition, our results only focus on the global sales. If it is in the designated region, some factors may also become very influential. For example, in Japan, the popularity of games published on game consoles is much higher than that published on PC platforms. Therefore, for some game producers who only face or focus on specific regions, our results are not suitable for their reference.

Question 4:

Our results for the fourth research question have important implications for game developers and publishers who rely only on high ratings from critics and users as a criterion for success. Our findings suggest that there is no clear relationship between review scores and global sales, and that other variables such as marketing, release timing, and platform availability may have a greater impact on sales. In addition, our results may be beneficial to game developers and publishers who are struggling in the marketplace. By focusing on the unique characteristics of their games and target

audiences, they may be able to create a marketing strategy that resonates with consumers and leads to increased sales.

On the other hand, individuals or companies that rely heavily on review scores to make decisions, such as investors or game industry analysts, may be excluded or potentially harmed by our analysis. They may need to consider factors other than review scores when making investment or market analysis decisions.

There may be some bias in our data that may affect our results. For example, our analysis is limited to video games released on major gaming platforms, which may not be representative of the video game market as a whole. Also, our data is partially missing after merging, which especially affects the data of ratings, so the conclusions obtained may not be precise.

The limitations of our analysis should be taken into consideration when interpreting our results. While our findings suggest that there is no clear relationship between review scores and global sales, it is possible that this relationship may exist in other markets or for specific types of games.

Question 5:

The potential implication of the result of question five is that the game developers and marketers should be aware that the popularity of game genres can fluctuate over time and is not always stable. They should consider the factors that may impact the sales of various game genres and modify their strategy appropriately. For example, if a specific game genre is losing appeal, they may want to shift their emphasis to more popular game genres.

Players of various game categories may benefit from these results as well since they can use these findings to predict the trends in the gaming industry and make more informed decisions about which game to buy. However, those who have a strong interest in specific game genres, might be excluded from the or harmed by these results if their favorite games are declining in popularity.

There are several potential biases in the data that may impact the results of this analysis. For example, because the data only contains sales from the top three game genres, it may not be representative of sales from smaller game genres.

The limitations of the results should also be considered. For example, the data only covers a limited time period (1996-2015), and trends in game genre popularity may have shifted since then. Moreover, the initial dataset contains more than 10000 data but

after merging the data, there are only 2000 left. The losing of the dataset could also lead to imprecise results.

Challenge goals

Machine Learning: Because the dataset we selected has a large amount of data, as well as many different variables and columns, we want to know whether we can make an accurate prediction model through these things. For the challenge goal, we want to use the machine training model we didn't learn from the class. We choose to use the sequential model from the Keras library that allows us to create neural networks in our layers for training several different variables to be our features that we believe are likely to be important influencing factors to make our prediction model with different labels, and make comparisons after completion. At the same time, in order to use this model well, we also learned, extended and used different methods and programs from the model. We also learned how to infer our training by observing the specific situation of training in our model, while it is iterating over the training dataset. Through the observation results, we adjusted the parameters, used layers, and the number of neurons in the layers in each layer. We have followed the content of machine learning learned in the classroom, and also expanded the content in the classroom and extracurricular learning.

Multiple Datasets: Since there are some missing values for some columns, especially for the score column, in our dataset. We decided to include more datasets and then combine and analyze them together so that we can still make data analysis for some of our missing data columns to solve our research question about player preference that involves the sales and scores of the games. At the same time, we also want to introduce more data of other categories for analysis, such as game score, which is divided into media score and player score.

New Library: We want to improve the effect and quality of our data visualization part by using some advanced or interactive visualizations, such as plotly. We will learn by ourselves how to use these libraries for visual display of data.

Work Plan Evaluation

1. Select and compare meaningful and interesting topics that can be used for data analysis (7 hours) **[ALL Finished!]**
    a. Search and compare different datasets, make sure which challenge goals could we reach (together before Feb.11th)
    b. Discuss the proposal and assign each section to different people (Each group member finishing before Feb.15th)
    c. Review, revise and submit the proposal (finished on Feb.15th)

Evaluation: For the first setting up part, we finished everything in detail and on time. There are small problems with the proposal part for which we changed both in coding and final report part.

2. Set up and share the Jetbrains Datalore for creating the prototype (3 hours) **[Finished]**
    a. On Feb.15, we've **already finished** the set up in Jetbrains Datalore, invited all the group members, uploaded the two data sets and already merged the data set together. Thus, we are **fully prepared** for the following part.
    b. According to different research problems, filter data for more detailed analyzing (finished on Feb.20th)
    c. Be prepared for local Python installation (finished before Feb.20th)

Evaluation: There's estimated 2 days delays for filtering the data since we first believe it is easy to do so but actually there's some problems. We asked TA for help with this problem and found out that the datasets are a bit complex. We need to merge 2 columns together instead of 1 column.

3. Responsibilities (each person 10 hours) **[Finished]**
    a. Discuss how we would reach different research problems using specific methods or plots
    b. Assign 5 research problems to different group members (finished before Mar.5th)
        i. For problem 1 (Moyi Li):focus on genre to create the pie chart (finish before Mar.7th)
        ii. For problem 2 (Moyi Li): focus on sales in different regions to create the bar plot, need to filter the data (finish before Mar.8th)
        iii. For problem 3 (Zhikai Li): focus on different factors to predict sales, plan to use machine learning (finish before Mar.8th)

iv. For problem 4 (Jingjing Dong): focus on score and sales, plan to write code on regression (finish before Mar.7th)
v. For problem 5 (Jingjing Dong): focus on the the total number of people favored in different game genre, use plot to display how it changes over time (finish before Mar.8th)

c. Support each other for challenging problems - might discuss through zoom, offline meeting or Jetbrains Datalore (finished before Mar.9th)

Evaluation: the assigning work didn't actually meet the plan since our data cannot make the map. We switched to bar plot instead. Everyone finished the coding part and supported each other for challenging problems before Mar.7th which actually exceeded our plan.

4. Assign and finish report (3 hours) **[Finished]**
    a. Assign and finish different part of the report (finished before Mar.9th)

Evaluation: We finally finished all the parts on Mar.10th which is a bit slower than what we planned. Since we have to analyze each result in detail, do a good evaluation and update changes from the proposal, it took more time to finish the report since there is more work than we expected.

5. Finish the video (before Mar.14th) (3 hours) **[Planning]**

Evaluation: we didn't start making the video now. It might take longer than we estimated.

Testing

In this project, we are mainly using the small dataset small_ds_test.csv to check out whether the plot generated is correct.

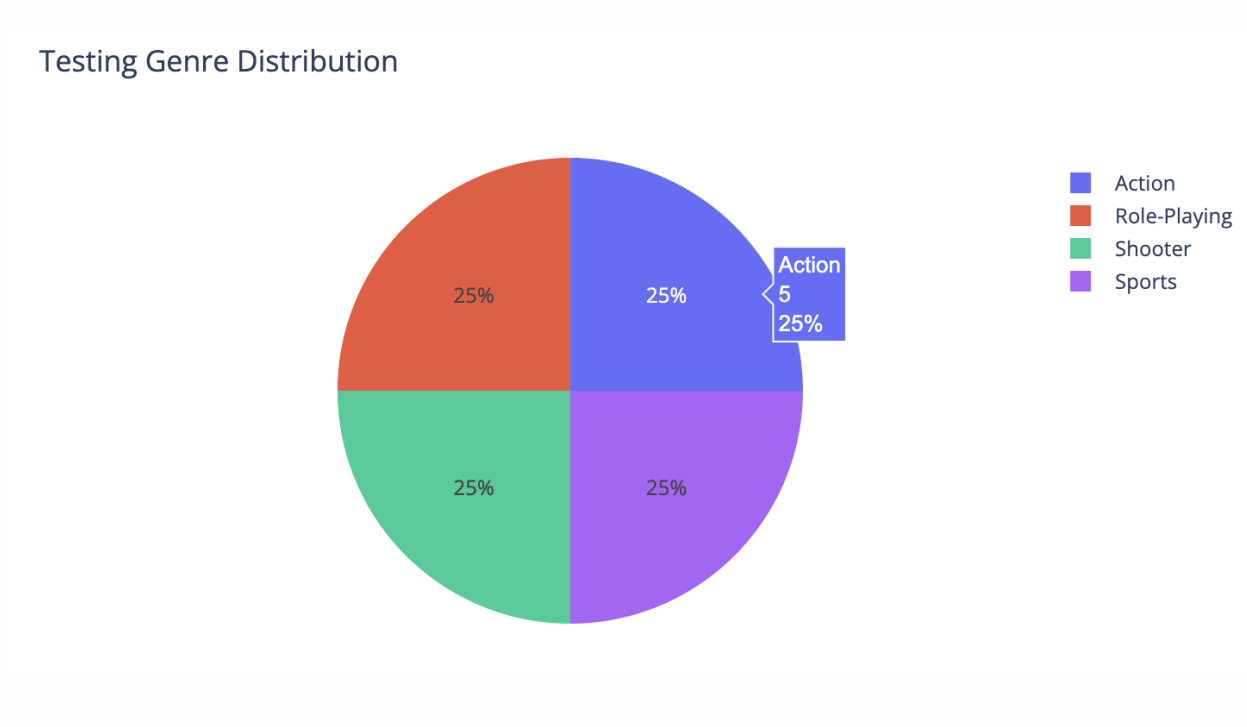
Here is a screenshot of the small_ds_test.csv.

small_ds_test

Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	NEW_Global_Sales	meta_score	user_score
1	Wii Sports	Wii	1996	Action	Nintendo	20	20	20	20	80	5	5	0.5
3	Mario Kart Wii	Wii	1997	Action	Nintendo	20	20	20	20	80	10	10	1
4	Wii Sports Resort	Wii	1998	Action	Nintendo	20	20	20	20	80	15	15	1.5
7	New Super Mario Bros.	DS	1999	Action	Nintendo	20	20	20	20	80	20	20	2
8	Wii Play	Wii	2000	Action	Activision	20	20	20	20	80	25	25	2.5
12	Mario Kart DS	DS	2001	Role-Playing	Activision	20	20	20	20	80	30	30	3
14	Wii Fit	Wii	2002	Role-Playing	Activision	20	20	20	20	80	35	35	3.5
15	Wii Fit Plus	Wii	2003	Role-Playing	Activision	20	20	20	20	80	40	40	4
28	Brain Age 2: More Training in Minutes a Day	DS	2004	Role-Playing	Electronic Arts	20	20	20	20	80	45	45	4.5
40	Super Smash Bros. Brawl	Wii	2005	Role-Playing	Electronic Arts	20	20	20	20	80	50	50	5
42	Animal Crossing: Wild World	DS	2006	Shooter	Electronic Arts	20	20	20	20	80	55	55	5.5
43	Mario Kart 7	3DS	2007	Shooter	Electronic Arts	20	20	20	20	80	60	60	6
49	Super Mario Galaxy	Wii	2008	Shooter	Sega	20	20	20	20	80	65	65	6.5
54	Super Mario 3D Land	3DS	2009	Shooter	Sega	20	20	20	20	80	70	70	7
61	Just Dance 3	Wii	2010	Shooter	Sega	20	20	20	20	80	75	75	7.5
65	New Super Mario Bros. 2	3DS	2011	Sports	Sega	20	20	20	20	80	80	80	8
69	Just Dance 2	Wii	2012	Sports	Ubisoft	20	20	20	20	80	85	85	8.5
74	Animal Crossing: New Leaf	3DS	2013	Sports	Ubisoft	20	20	20	20	80	90	90	9
74	Animal Crossing: New Leaf	3DS	2014	Sports	Ubisoft	20	20	20	20	80	95	95	9.5
74	Animal Crossing: New Leaf	3DS	2015	Sports	Ubisoft	20	20	20	20	80	100	100	10

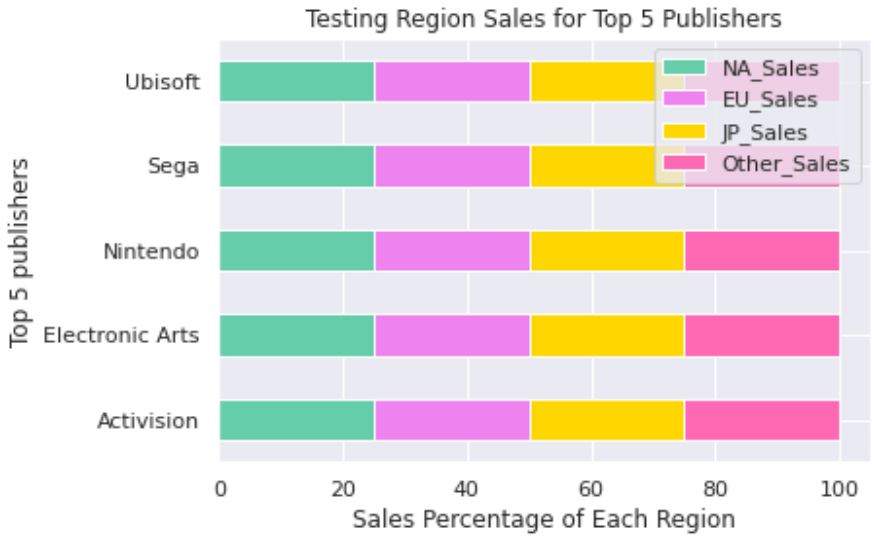
Question 1:

In the small dataset, we manually set the genres of 20 games evenly to Action, Role-Playing, Shooter, and Sports, each with five games. As a result, all four genres are equally distributed in the tested plot, so the method can be proven correct.



Question 2:

Similarly, by assigning the same regional sales to each of the top 5 publishers, we can observe that the distribution of the five publishers is also equally distributed in the testing image. Thus, the method of question 2 can also be proven correct.



Question 3:

Our observation of our model training process can determine whether our model is running correctly or not, and whether there are problems with underfitting or overfitting. As we train, we can see that the loss is the measure of the differences between predicted and real output, which is gradually decreasing in the long run, which means that our model is indeed conducting training. If we observe that after the training has gone through a certain process, the loss will no longer decrease, or will gradually increase, which means that our training is likely to overfit. In order to avoid this, we need to adjust the number of times to iterate over the training dataset by constantly testing to observe what will happen at a specific stage.

In order to judge whether our results are correct, we also used the expression model in sklearn that we have learned and used in class and homework. After running and testing, we found that the results do not differ much from the results of the sequential model, and it is even less accurate for most of the time, which is expected since the training model we use is more complex.

```
Epoch 1/80
200/200 - 1s - loss: 109.7505 - mse: 109.7505 - val_loss: 12.3396 - val_mse: 12.3396 - 1s/epoch - 6ms/step
Epoch 2/80
200/200 - 0s - loss: 16.7663 - mse: 16.7663 - val_loss: 2.2675 - val_mse: 2.2675 - 452ms/epoch - 2ms/step
Epoch 3/80
200/200 - 0s - loss: 13.0553 - mse: 13.0553 - val_loss: 2.6503 - val_mse: 2.6503 - 457ms/epoch - 2ms/step
```

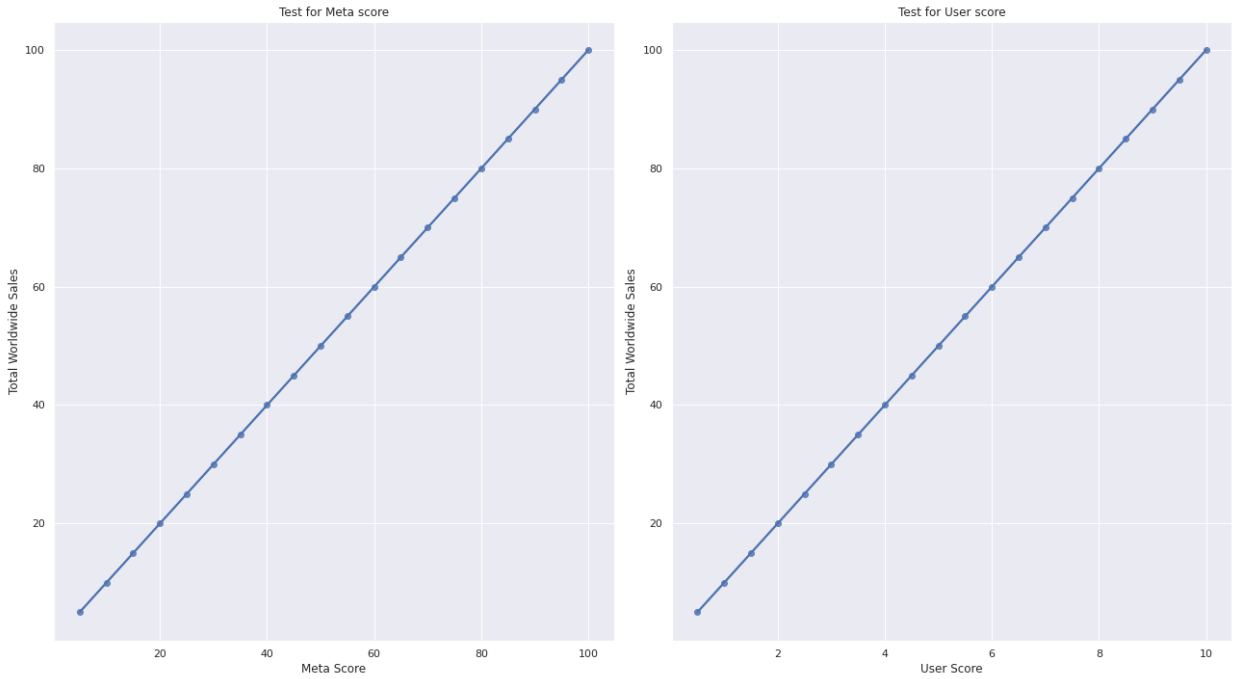
```
Epoch 78/80
200/200 - 0s - loss: 9.4651 - mse: 9.4651 - val_loss: 2.2602 - val_mse: 2.2602 - 436ms/epoch - 2ms/step
Epoch 79/80
200/200 - 0s - loss: 9.4669 - mse: 9.4669 - val_loss: 2.2552 - val_mse: 2.2552 - 430ms/epoch - 2ms/step
Epoch 80/80
200/200 - 0s - loss: 9.4677 - mse: 9.4677 - val_loss: 2.2695 - val_mse: 2.2695 - 433ms/epoch - 2ms/step
13/13 [==============================] - 0s 2ms/step
Test set mean squared error for main analysis using final_result: 2.26947458771945
Test set mean squared error with sklearn: 2.9679195
```

To further verify that there is no problem with the use of our model, we used a test method. Its features include the sales volume of games in different regions, such as North America and Japan. Then we can find that the predicted results become very accurate, which is expected, because the global sales is composed of these features. This proves that there is no problem with our model, but these features can only be used as a test, because normally it is meaningless to predict its global sales volume if we know the sales of the game in different regions.

```
Epoch 1/150
408/408 - 1s - loss: 15.0205 - mse: 15.0205 - val_loss: 1.5302 - val_mse: 1.530
Epoch 2/150
408/408 - 1s - loss: 1.4226 - mse: 1.4226 - val_loss: 1.0307 - val_mse: 1.0307
Epoch 3/150
408/408 - 1s - loss: 0.7456 - mse: 0.7456 - val_loss: 0.3980 - val_mse: 0.3980
```

```
Epoch 148/150
408/408 - 1s - loss: 0.0304 - mse: 0.0304 - val_loss: 0.0017 - val_mse: 0.0017
Epoch 149/150
408/408 - 1s - loss: 0.0305 - mse: 0.0305 - val_loss: 0.0023 - val_mse: 0.0023
Epoch 150/150
408/408 - 1s - loss: 0.0272 - mse: 0.0272 - val_loss: 0.0015 - val_mse: 0.0015
102/102 [==============================] - 0s 1ms/step
Test set mean squared error with sales column: 0.0014736705600164484
```

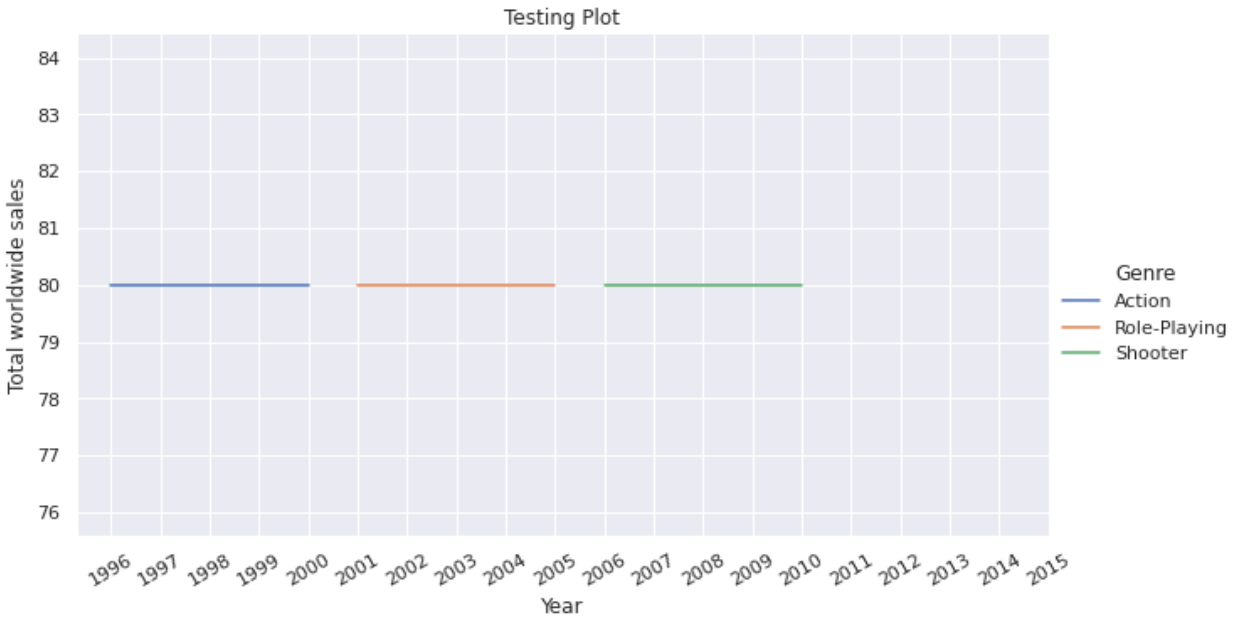
Question 4:




```
meta test: 1.0
user test: 1.0
```

In the small dataset, the meta score are set to be 5, 10, 15, 20…100, user score are set to be 0.5, 1, 1.5, … 10, and new global sales data to be 5, 10, 15, 20, … 100 which means that the coefficient of determination of both chart should be 1 and the line should be a line with second quadrant angle bisector. The chart and calculations meet with the true results. Since we used the same method in testing and actual analysis functions, thus, our result is proved to be correct.

Question 5:



In the small dataset, we set the Action game to have 80 global sales from 1996 to 2000, Role-playing game to have 80 global sales from 2001 to 2005, and Shooter game to have 80 global sales from 2006 to 2010. The plotting for the small data about genre vs year should be three horizontal line segments with different colors that lie on total worldwide sales at 80. The result of this meets with the plot we created. Since the same method is used for testing and true analysis, it is proved that our function written in true analysis is correct.

Collaboration

In this project, we used the online coding platform Datalore. We also accepted debugging and project clarifications from our TA, Ben Zhou. Besides, to finish our challenge goal, we have learned from the following websites.

- Plotly tutorial: https://plotly.com/python/pie-charts/
- Plotly documentation: https://plotly.com/python-api-reference/generated/plotly.graph_objects.pie.html
- Plotly example: https://lifewithdata.com/2022/02/28/how-to-create-a-pie-chart-in-plotly-python/
- Stacked bar chart: https://www.geeksforgeeks.org/stacked-percentage-bar-plot-in-matplotlib/

- Tensorflow Machine learning tutorial: (1) Tensorflow Tutorial for Python in 10 Minutes - YouTube
- Keras Regression model tutorial: Keras regression example - Projectpro
- Tensorflow methods website: tf.keras.Sequential | TensorFlow v2.11.0, tf.keras.layers.Dense | TensorFlow v2.11.0
- Calculating coefficients of determination: https://realpython.com/linear-regression-in-python/
- Reference to interpret Question 2 's result https://www.globenewswire.com/en/news-release/2022/09/20/2519436/28124/en/China-Console-Games-Market-Report-2022-2026-Increase-in-Spending-to-be-Driven-by-On-going-Sales-of-Nintendo-Switch-Sony-PlayStation-and-Microsoft-Xbox.html