**Videogames: Stats and the Pandemic?**

*In this project, we seek to find interesting statistics from videogames.*

*Such as the trends in gaming popularity and perhaps*

*how the pandemic can influence the sales.*

*Non Pronanun & Sirak Yohannes*

Summary

• How has the genre trends evolved over the years and does COVID19 have any effect on the performance of videogame sales?

• The most popular genres remain mostly unchanged, even if the sales do fluctuate.

• Gaming has mainly an upward trajectory sales-wise and from the results that we found, COVID19 has a strong impact on the sales.

• RESEARCH QUESTION2

• Mincraft was the largest game of all time.

• Action games are the most popular genre, first person is second.

Motivation

Videogames are one of the most enjoyed pastimes for people and it is also holds some of the fastest growing communities within it – especially since COVID hit and people have been led to find new hobbies within their homes as a result of the pandemic, and one of these hobbies was videogames. However – with thousands of videogames on the market – it can be challenging to find a point of entry to the gaming world, and it can even dissuade people from getting into gaming altogether, so we seek to find ways that we can help people understand the demographics of the videogame community. This way, people with no prior knowledge of the gaming community are able to see what other people are enjoying and could find a place that they would enjoy and fit in. Furthermore, we want to know how the COVID pandemic has affected the community of videogames as well, so we want to find ways that we can analyze how big of an impact the pandemic had on videogames.

Data Setting

The main dataset that we use for this project is the videogame sales dataset, linked below. It contains 11,000 rows and 11 columns of videogame sales and statistics in each year, accurate up to 2015.

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

Method

Before starting, we inspected the dataset to have a greater understanding of what the dataset is able to tell us and how we can use it to help find us find answers to our research questions. In each of the research question parts, we need different parts of the dataset, so we started by identifying what we need for each of the parts of the research questions and from that we cleaned the data for each of the parts accordingly.

For research question one.

Part 1 – Machine learning to predict sales in 2020.

• Cleaned the data up to the parts that we needed: the sales and year.

• Made a scatterplot representing each videogame and their sales in each year.

• Used a linear regression as our model to predict the sales in 2020.

• Plotted the regression line onto the graph and found the actual number of sales through a google search.

Part 2 – Trends of videogame genres throughout the 2000s and 2010s.

• Cleaned the data up to the parts that we needed: the sales, the genre, and the year.

• Created a plot with 4 subplots that we can use for the year with 5-year intervals between them – 2000, 2005, 2010, and 2015.

• Plotted each of the subplots out, where each of the subplots were also sized relative to their sales in that year.

Results

Research Question 1, Part 1.

Based on the regression, we were projected to sell 55.4 million copies of videogames in 2020; however, the real number was closer to 140.4 million copies sold. This suggests that based on the growth trend of the 2010 – assuming this continues for future growth – the COVID pandemic had a massive impact on the sales. The result is understandable, the COVID limited people to their homes and it is natural that videogames would be a performing well, given that it allows people to explore a virtual world. We hypothesized that videogames were impacted by the pandemic, but we didn’t think that the impact was this big, because this could mean that the pandemic led to the videogame market growing by 3 times.

Research Question 1 – Part 2.

Based on this, we can conclude that the genres of videogames have remained relatively constant throughout the years, suggesting that although mainstream media may point to a year being the year of a genre, the general trend points to the fact that the popular genres have remained unchanged. Furthermore, the results show that the 4 most popular genres – platform, shooter, role-playing, and action – held almost a 60% market share of videogames in the 2 decades. However, it is important to note that videogame genres can be ambiguous.

Impacts and Implications

From our results, we can conclude that the COVID19 pandemic did have a large effect on the sales of videogames in 2020, and that the genres of videogames have remained largely unchanged throughout the years. However, we can’t say that COVID19 caused this surge in sales because correlation doesn’t mean causation. People that would benefit from these findings would mainly be people who are looking to get into videogames but are unsure where they should start – as this would give them some form of information in order to enter the community. Perhaps people who are looking to see if the gaming market is profitable will be able to see the trends of that, and how COVID can become an opportunity for them to potentially make a game. There are also certainly limitations to this project. Since this is a dataset from Kaggle, and it is a “centralized” dataset, the data may contain inaccurate information so there is no way to tell that all of the data that is in this dataset is 100% true. In addition, the genres of videogames can be ambiguous and games that arch over different genres can be misrepresented. For example, GTA 5 is both a role-playing game, an action game, as well as a shooter, but it’s labeled as action – so there are areas where this dataset is not perfect.

Challenge Goals

The first challenge goal that we had was to be able to use external, more advanced plotting packages to create plots that are able to tell us more than just two dimensions. Where normally, plots are used to tell the correlation between x and y, but we want to be able to make plots that are able to give us another piece of information. In this case, it would be the genre\_trends graph, that shows the sales, genre distribution and time in the same plot. The second is to find correlation between this data and COVID19 using machine learning algorithms. This dataset doesn’t tell us explicitly about how COVID19 affects the sales of videogames, so we want to be able to relate something that is overarching like COVID19 to something that is more niche like videogame sales – thus our goal is to find the effect that COVID19 had on videogame sales through the use of machine learning algorithms to predict the expected growth versus the actual numbers of COVID19.

Plan Evaluation

Predicted – Actual

Data cleaning and analysis (2 hours – 2 hours)

This one is not exact because we thought that we would need to do the cleaning once, but we had to do some cleaning before starting the implementations and then we had to clean again for the needs of each of the implementations that we had set.

Research Question 1 (2-3 hours – 5 hours)

It took longer than we expected because learning how to use new libraries together was time consuming because of the mismatch in the parameters that some of the functions in each library requires. Then, implementing the parts needed a lot of time to debug and make sure that the graphs were showing the right data and that they were readable. Though, we believe that if we were to recreate this assignment for another topic, it’d be closer to our prediction.

Research Question 2 (2-3 hours)

During the development process, I explored various prototypes for both of my codes in order to determine the most effective representation for the information. After careful consideration, I ultimately chose to utilize a Barnhart and a line chart. These visual elements were selected to effectively convey the comprehensive data and enhance the overall presentation. By employing the Barnhart and line chart, I aimed to create a visually engaging and informative display that would effectively communicate the intended information.

Report (2 hours – 2 hours)

We didn’t really track the time it took for this one because we wrote it along the way of implementing the coding portions of this project. However, it was close to our prediction because although there were some parts we can copy from the proposal, we rewrote most of the parts, so it reflected more with the actual work that we did.

Testing

We made a test file called “test.csv” which holds the first 150 rows of the videogame sales dataset, we were able to calculate the correct statistics by hand due to the fact that the dataset was manageable. It worked correctly on most of the parts apart from predicting the sales in 2020, as there is not enough data to form an accurate linear regression line to represent the trajectory of videogame sales throughout the years. The test results for the graph are in the images folder, but the call for them in main was deleted to prevent overwriting with the results. In addition, I made a testing method to test if the program is counting the dataset properly using the test dataset and comparing the results to what we got by counting it by hand.

Collaboration

We used the plotly website to learn about the different graphs and how to use them, as well as Stack Overflow to help with errors that we encountered while working on libraries that were not class materials.