Data Visualisation and the Web
Ali Fawzi

Individual Project Report

Introduction

In this report I will be analyzing, reconstructing and tidying up data so that I can show clear data visualisations that will be used to answer my research question(s). The data I have decided to analyse is video game sales with ratings from Vgchartz and Metacritic. There are millions of games being sold every year all around the world. My motivation for choosing this dataset was to find out the exact ratings and sales of game being sold each year. In addition, I wanted to see which types of games are very popular, generate most sales, and which ones have the most demand.

Research Topic

My research questions include:

- Are newer games more popular than older games?
- What is the most popular gaming genre?
- What is the most popular game?

I have chosen the following research questions as I believe they will help me find out which type of games are generally the most popular, and why are millions of them able to be sold each year.

Data

I found my data on Kaggle.com by looking for data sets on video games. The data set is in a csv format on Microsoft Excel. It is in this format so that the creator of the data set could easily import and export the data into tables to store the tabular data. The steps I took to pre-process the data included importing the necessary libraries, organising the data into separate columns to be processed. In addition, I got rid of unnecessary that would not factor in my data analysis and data visualisations. The data I deleted includes 3 rows of games produced in 2017 when the data is only up to the year 2020, and I deleted game data that mentioned that the game would be published 2020 which when researched was false and irrelevant as the data only goes up to games produced in 2016 as the latest year. Also, I noticed that quite a few columns had gaps in them. One of them being the critic score column. I made sure to not use this data so that it doesn't produce inconsistencies in data visualisations.

https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings - [Date Accessed:o 21 Nov 2019, 3:20 PM]

Exploratory and explanatory data visualisation

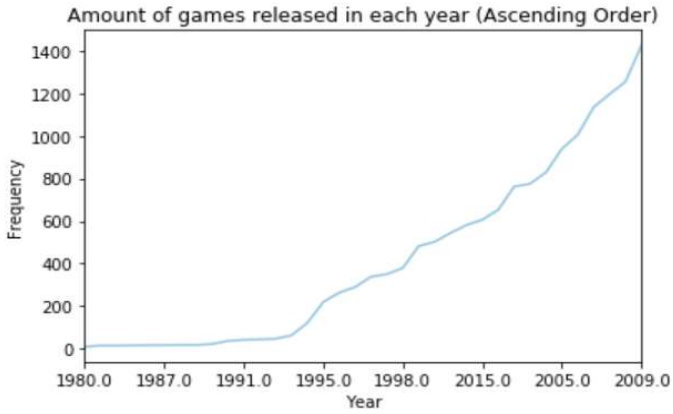
In order to successfully create clear visual representations of data, I had to conceptualise my variables. The variables of interest include:

- Year (int64 – Ordinal): This data type allowed me to gain an understanding of whether the production of games is increasing or decreasing.

- Genre (Categorical – Nominal): This variable allowed me to understand which gaming genres are the most popular.
- EU, NA, JP, other, and Global Sales (Numerical – Ratio): This variable has allowed me to view and compare the sales between different games.
- Game names (Categorical – Nominal): This variable is used to specify which exact games have had good sales and ratings.

	Year	NaSales	EuSales	JpSales	OtherSales	GlobSales
count	16446.000000	16715.000000	16715.000000	16715.000000	16715.000000	16715.000000
mean	2006.484616	0.263377	0.145060	0.077617	0.047342	0.533649
std	5.877050	0.813604	0.503339	0.308853	0.186731	1.548104
min	1980.000000	0.000000	0.000000	0.000000	0.000000	0.010000
25%	2003.000000	0.000000	0.000000	0.000000	0.000000	0.060000
50%	2007.000000	0.080000	0.020000	0.000000	0.010000	0.170000
75%	2010.000000	0.240000	0.110000	0.040000	0.030000	0.470000
max	2016.000000	41.360000	28.960000	10.220000	10.570000	82.530000

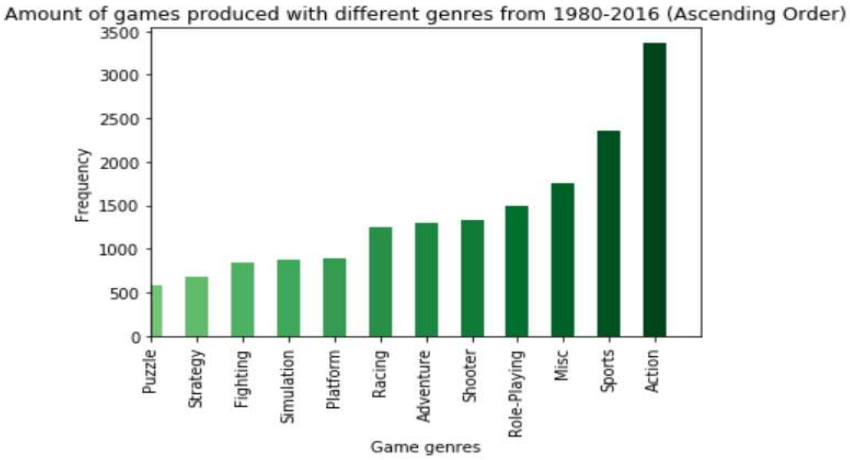
The above table summarises the data I am analysing. This tables allows to understand the overall amounts of data provided as well as the minimum and maximum values. Based on the table I can see that the data only contains information on games released from the year 1980 to 2016. As well as that, it shows the highest and lowest sale values that were produced. The highest amount of global sales produced was 82.53 million units. This shows that one game managed to get 82.53 million units sold which demonstrates that this game was very popular when it was released. This game with the highest number of units sold was Wii Sports.



The visualisation above show a line graph representation for the amount of games released in each year. We can see that the data used is Categorical since its key variable is qualitative. Also, we can see that the data is Ordinal as it follows an order for each year. I used a line graph to represent this data as it would clearly show the increase in frequencies between each year as I have ordered the years from lowest frequency of games released to the highest amount of games released. With regards to my research question, “Are newer games more popular than older games?”, we can make a case that the older games may have been less popular due to less demand from the potential customer. More games started to be produced in the 2000s. A BBC article mentioned that the increase in demand

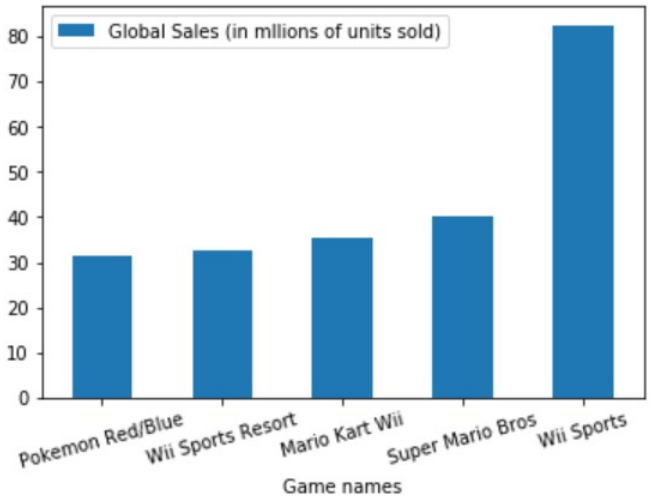
for gaming was due to the development and advancements of technologies. Therefore, we can say that developers produced more games as their technologies had vastly improved. However, we can see that in the year 2015 there were less games less games being produced than there were in 2005. This could be due to a halt/a slower rate of development in gaming technology which means less game production.

https://www.bbc.co.uk/news/technology-46746593 - [Date Accessed: 12th December 5:30 PM]



The visualisation above shows a vertical bar graph representation for the amount of games produced that have different genres within the yearly timespan of 1980-2016 in ascending order. We can see that the data used is Categorical since its key variable is qualitative. Also, we can see that the data is Nominal; the reason for this is because the data has no intrinsic order. I visualised the data through a vertical bar because it would be much clearer and presentable to show the data this way so we can see the clearly which types of games are being least produced, and which ones are being produced the most. We can clearly see that puzzle games are being produced the least as it is at the beginning of the graph since it has the lightest shade of green. Also, the graph demonstrates that Action games are being produced the most as there are about 3500 of these types of games being produced from 1980 to 2016. However, we cannot completely say that Action games must be getting produced the most at present; the reason for this is because this data graph is from 1980 to 2016 which means that it could be Action games may have had more demand in one of the years and perhaps a few other which is why so many were being produced, but another genre type could have had more productions in another year. Therefore, with regards to my research question, “What is the most popular gaming genre”, we can say that based on the bar graph that action games are overall the most popular since there was more production for them but, we cannot conclude that action games are the most popular of all time since for some of years between the timespan, this result could have been different. Perhaps for one of the years in the timespan, sports games could have had the most production.

Data Visualisation and the Web
Ali Fawzi



The key variables for the above visualisation are the game names and the global sales. The name of the games are qualitative data which means they are categorical. The type of data is nominal since there is no intrinsic order. Global sales are numerical variables which means that the data is quantitative. The data type is ratio since the data can be manipulated and the difference between 2 of the values has no meaning. The visualisation above shows that I have created a bar graph to represent the top 5 globally sold games from 1980-2016. I chose to visualise this data through a vertical bar chart since bar charts make it very easy to compare variables and spot relations with data as it is very clear to analyse. Judging by the graph, we can see that the games with the most global sales is Wii Sports. Therefore, with regards to my research question, “What is the most popular game”, we can say that the Wii Sports was the most popular game since it had 82.53 million units sold. However, there could be some disagreements with this as there are different results with JP sales, EU sales, and NA sales. However, this is the game that was recognised the most globally due to its global sales

Conclusion and evaluation

Judging by my 1st graph, you can see that there was a greater demand and interest in making games in the 2000s than there was in the late 1980-1999. This was due to the development in technology. As technology was being enhanced, the interest for both producing and purchasing games also increased. As for the second graph, we can see that the amount of different action games being produced was the highest. By looking at the third graph we can make an argument that Wii Sports was globally, the most popular game produced when looking at games from 1980-2016. This may not be the case places such as Japan. Therefore, taking into consideration all the visualisations, we can conclude that sports games produced in the 2000s have the best overall sales and ratings. This is because sports games are very popular since the genre was 2nd off the top 5 games, and Wii Sports had the most global sales. In addition, to the fact that this game was produced int the 2000s. However, I had to delete some data as it was either false or too small, this could have slightly impacted my findings. I have learnt that generally sports games, made in the 2000s are one of the most popular gaming genres. They will only continue to grow in popularity.

Even though I produced great visualisation that represent my dataset very well, I do believe that I could have made the line graph clearer by ordering the years in ascending order to make it look a lot clearer. Also, in the future, I will make sure create more variations of data visualisations to further analyse and interpret my data.

Additional References

- https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.plot.bar.html – [Date Accessed: 7th November 8:14 PM]
- https://www.samplequestions.com/types-of-research-questions.html – [Date Accessed: 11th December 6:30 PM]
- https://en.wikipedia.org/wiki/Pok%C3%A9mon_(video_game_series) – [Date Accessed: 12th December 2:07 PM]
- https://www.computerhope.com/issues/ch001356.htm – [Date Accessed: 12th December 3:29 PM]
- https://medium.com/datadriveninvestor/data-preprocessing-for-machine-learning-188e9eef1d2c – [Date Accessed: 12th December 5:59 PM]
- https://rpubs.com/vh42720/vgsales – [Date Accessed: 13th December 3:10 AM]
- https://nces.ed.gov/nceskids/help/user_guide/graph/whentouse.asp – [13th December 10:21 AM]
-