PES UNIVERSITY

ANALYSIS ON VIDEO GAME SALES

Kevin Arulraj : PES1201700659

ABSTRACT

I will be carrying out an Exploratory Data Analysis on a dataset containing a detailed list of video games. I aim to extract meaningful information regarding the current state of the video games and find correlations between the variables influencing the industry.

INTRODUCTION

The video game industry is increasing in size year by year. It is gradually becoming the market leader in the entertainment industry with applications ranging from education to healthcare. As such, there is a need for comprehensive analysis about the state of the industry, the various factors contributing to a video game's sales and so on.

There is a large amount of information that can be extracted and put to use from this dataset. What genre of games sell well in Japan, how fast is the industry growing, the success of the various consoles etc. are all questions we hope to answer using this dataset.

DATASET

The dataset is collected from Kaggle. It is a large dataset consisting of 16700 video games over a timespan of 20 years with 17 columns of data for each. The columns consist of the genre of the game, its sales in various regions, the platform for which it was sold, and user and critic ratings.

DATA CLEANING

There was a lot of pre-processing required before we could use the dataset and extract meaningful information from it. The columns User_Count and Critic_Count contained large numbers of missing values and were mostly useless and were thus dropped.

In some places, the User_Score was not yet determined('tbd') and we had to reset it to 0. The empty values in User_Score and Critic_Score were also filled with 0's to ensure accurate computations. Similarly, the empty values in Developer and Rating were filled with 'Unknown'.

There was an issue due to multiple rows present for a single game being present for each platform it was released on. This interfered with our calculations and thus had to be processed. The sales of the game on each platform was summed, the scores given by critics and users were averaged to ensure correct calculations. The other rows were dropped and a new row containing the more appropriate data was added.

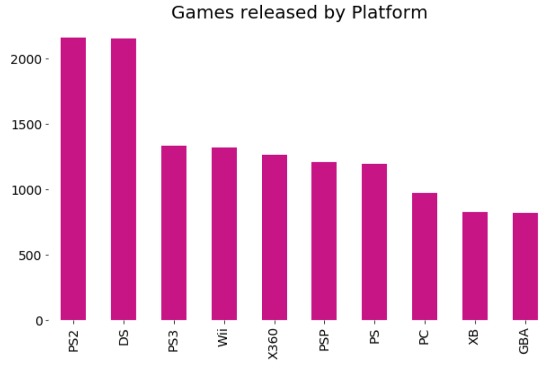
TYPE OF DISTRIBUTION

As can be seen from the below histogram, video game sales are a highly skewed distribution, with the majority of video games barely breaking even with a few outliers being global blockbusters.

ANALYSIS AND INFERENCES

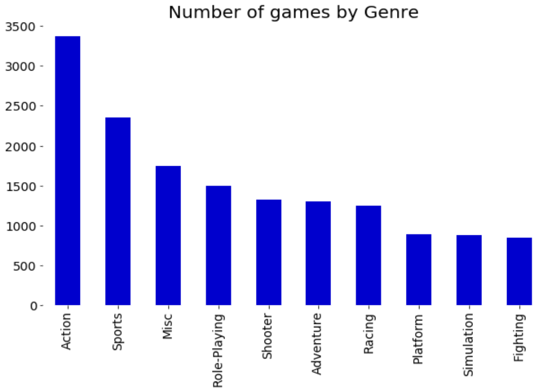
1. Number of Games Released by Platform

As we can see from the graph, the highest number of games were sold for the PS2 and the DS. The least number of sales is for the GBA. The PS2 was a massive success and sold in unprecedented numbers thus showcasing a large number of titles.



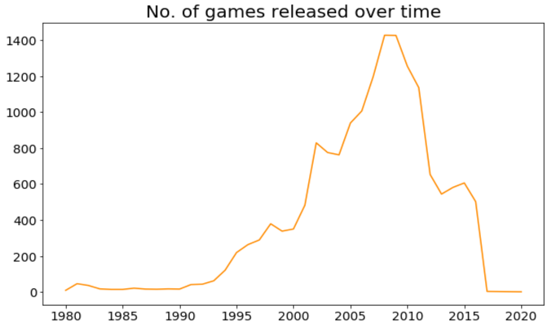
2. Number of games by Genre

Action games were the most developed followed by sports games. The least number of games developed were for the platform, simulation and fighting categories.



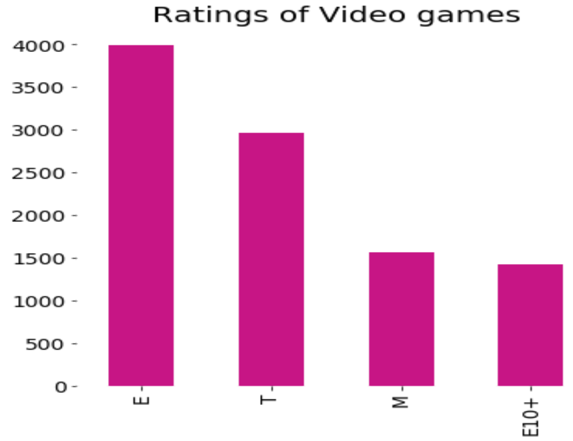
3. Number of games released over time

The number of games released has shown a consistent increase till about 2010 and then a sudden drop after 2010. The highest number of games sold in a year was in 2010. This goes along well with the market trend that is focusing more on larger and big-budget video games rather than multiple small ones. i.e. quality over quantity.



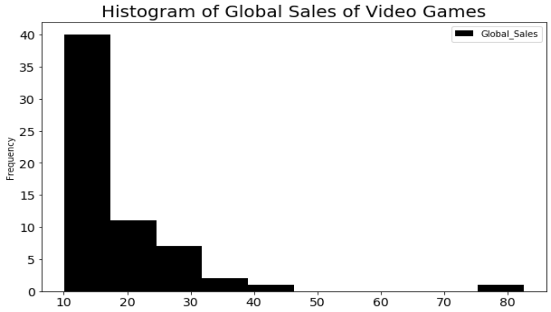
4. Number of games by their ESRB Rating

The highest number of games are given the E rating. E stands for Everyone and developers tend to release games that can be played by the maximum number of people in order to maximise revenue.



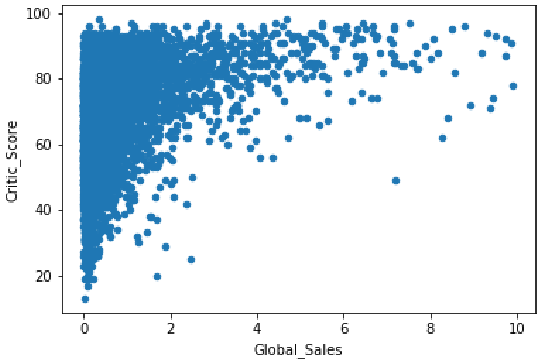
5. Global Sales of Video Games

From this histogram, we can see that video game sales are a highly right skewed distribution. Very few games manage to sell more than 20 units globally with major blockbusters selling nearly 4 times as much.



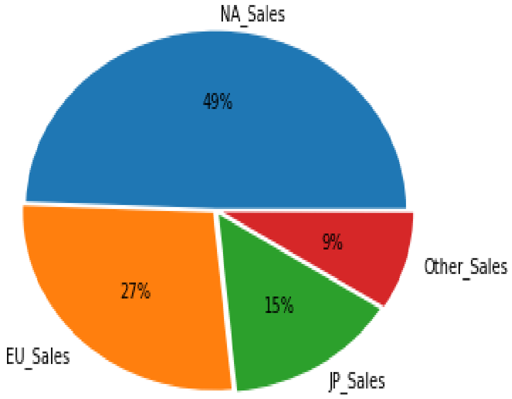
6. Critic Score v/s Global Sales

In this scatter plot, we see that there is a very weak positive correlation between critic scores and global sales. Games that don’t sell well (i.e the vast majority) have a wide range of critic scores but the games that sell well invariably have a good score given by critics barring a few outliers.



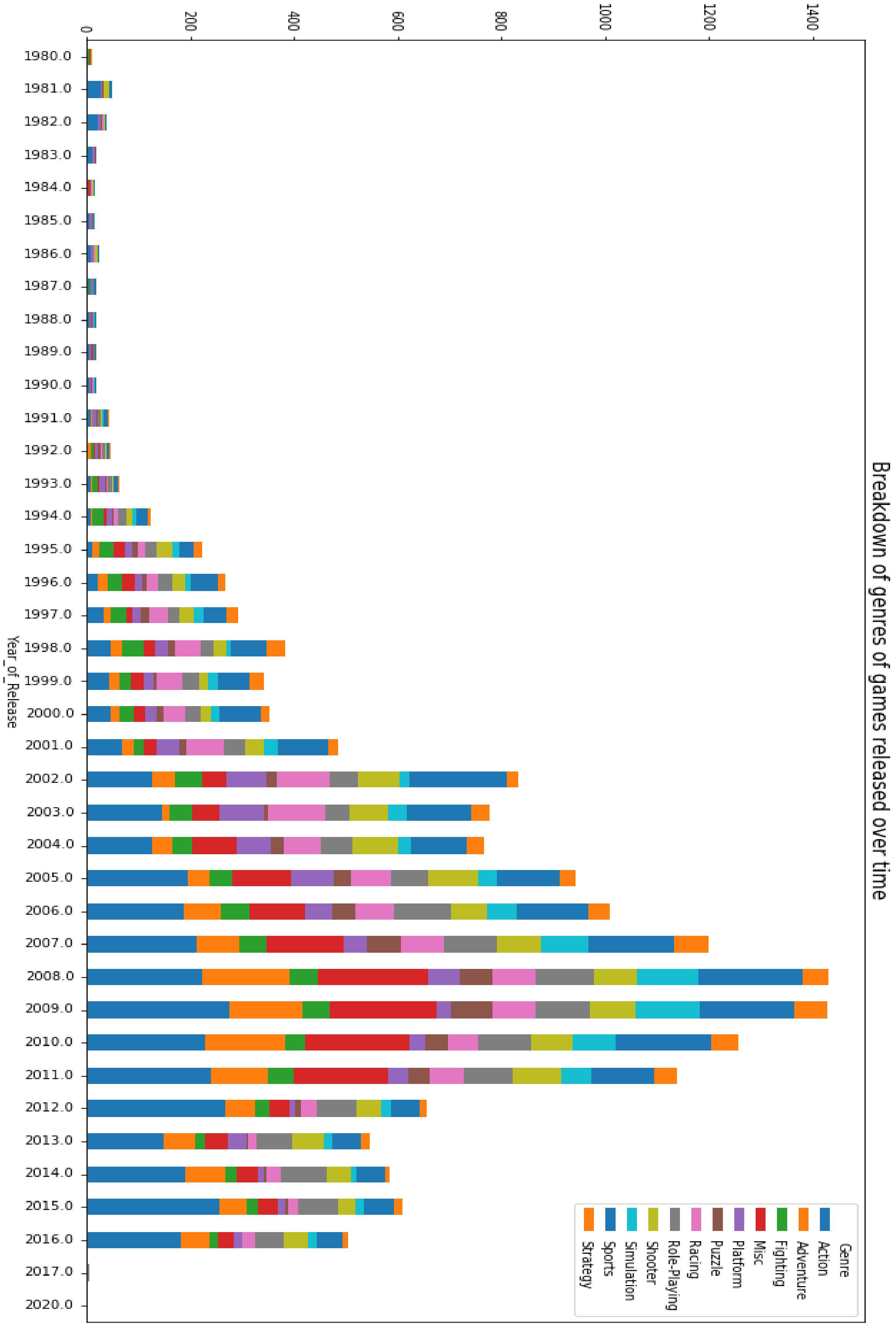
7. Sales by Region

Looking at the pie chart, we can straight away say that the most sales come from the North American region followed by Europe and Japan all of which sell more than the rest of the world combined.



8. Number of games released by year and their breakdown by category

This bar graph shows the number of games being released each genre broken down by genre. Even as the number of games released per year increases, the proportion of games in a particular genre remains more or the less the same each year with action games consistently being the largest genre by far.



Breakdown of genres of games released over time

HYPOTHESIS TESTING

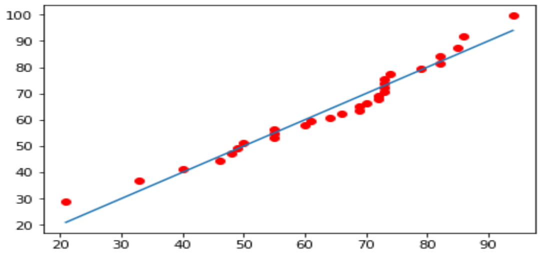
Z-Test

We will compare the general Critic Scores of games for the Xbox 360 and PS4 console. We will be performing as we have a large normally distributed dataset.

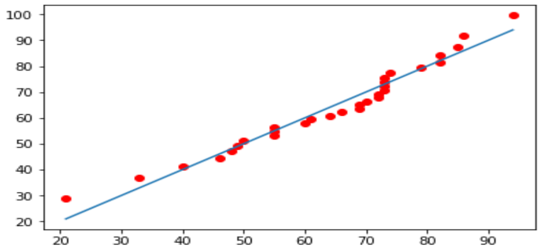
Null Hypothesis: H_O – Critic Scores of Xbox 360 games are better than those of the PS4.

Alternate Hypothesis: H_A – Critic Scores of Xbox 360 games are lesser than or equal to those of the PS4.

Taking one random sample of 30 critic scores each from a population of critic scores of the games in each console, and checking for normality, we can see that the scores are normally distributed.



Normality Check for scores from Xbox360 games



Normality Check for scores from PS4 games

Mean of scores taken from random sample of Xbox games = 64.8

Mean of scores taken from random sample of PS4 games = 70.2

Standard deviation of scores of Xbox games = 16.1

Standard deviation of scores of Ps4 games = 12.7

Computing Z score for the difference of the means we get Z = -2.513

Computing P value from Z score we get P = 0.006

With this P value we can reject Null Hypothesis and accept the Alternate Hypothesis. So we conclude that PS4 games generally have a better score than Xbox games.

Chi-Squared Test for Goodness of Fit

We will be testing our dataset using the Chi-Squared test for goodness of fit to see if the sample data distribution is consistent with the given population distribution

Null Hypothesis: H_O – The data is consistent with the specified distribution.

Alternate Hypothesis: H_A – The data is not consistent with the specified distribution.

We take one random sample of 300 values from the dataset's platform column and compute the Chi-Squared test statistic and the p-value.

We calculated the relative frequencies of each genre in the sample and population. We computed the proportions and then calculated the chi-squared test statistic and the p-value.

We observed that the p-value consistently remained above 0.2 for taken samples. Since we only reject H_O when p-value<=0.05 we can conclude that both the null and alternate hypotheses are plausible.

Conclusions

1. From Graph 1, we can infer that the PS2 and DS consoles had the most number of games released for their platform. They were released during the gaming boom of the 90's and the early 00's and sold the most.

2. We can infer that the games that come under Action category sell the most copies by looking at Graph 2. People love action games in which they can do fantastical moves and this can be seen by the large number of action games being produced.

3. From Graph 4, we can infer that the games that have an E rating, appealing to all of the player base has sold most copies and is followed by games that are likely to interest teens. We can say that the game developers are targeting teens for their sales.

4. We can conclude that there is a very weak positive correlation between the Critic Score and the games' sales from Graph 6.

5. From Graph 7, we can see that the highest number of games released were in 2009-2010 and in every year the Action genre had the most number of games.

6. Graph 8 shows that the region with the highest sales is the North American region, and it is almost equal to the sales in the rest of the world.

7. From the Z-Test performed above we can conclude that the Null Hypothesis : Critic Scores of the XBox Games are better than the PS4 Games can be rejected.

8. By performing a goodness of fit test we can say that the taken samples are consistent with the population