Assignment
CS989 Big Data Fundamentals

Predicting the sales of video games

Parth Varangaonkar

201851829

Table of Contents

Introduction

Since the first ever game was released back in 1958, it opened up a new field for not only computer engineers but for many other individuals from different backgrounds. Ever since the business of video games was started, it has only seen an upward graph in every term. The history starts with the rise of the classic console game ‘Mario Bros (1983)’ to the most recent release of ‘Red dead redemption 2 (2018)’. There have been a large number of games that have been published and developed in the recent times, however only a few have been able to make their mark on the industry. This entertainment industry has not only caught the attention of the children, it also has drawn people from all the ages to engage with video games. With the rise of technology, there has been the use of artificial intelligence within the games, development of Virtual Reality games . Video games have captured a big chunk of the market and become a big billion dollar industry in the past two decades.

This report aims to explain the dataset of the video game sales and show the proper analysis of the games that have been published between the year 2000-2016. The rationale behind choosing this particular range of year was to show the current trends of the gaming industry. This report will also explore Hierarchical clustering and fit the linear regression model on the dataset.

Introduction to the dataset

The dataset used for analysis is The Video Game sales(1) with their user and critic ratings. It covered a range of years from 1980 to 2016. It has the variables namely 'Name', 'Platform', 'Year_of_Release', 'Genre', 'Publisher', 'NA_Sales', 'EU_Sales', 'JP_Sales', 'Other_Sales', 'Global_Sales', 'Critic_Score', 'Critic_Count', 'User_Score', 'User_Count', 'Developer', and 'Rating'(appendix 1). The sales of the games is indicated in the number of million units a particular game sold. The dataset was chosen because of the varied variety of questions that could be potentially be answered with it.

Problem definition

Video games are a wide research area which can be explored into with a depth of immense knowledge. ‘The Guardian’ newspaper had an interview with the English actor, Andy Serkis where he quoted

“*Every age has its storytelling form, and video gaming is a huge part of our culture. You can ignore or embrace video games and imbue them with the best artistic quality. People are enthralled with video games in the same way as other people love the cinema or theatre. Over time, I think perceptions will change.*” *(2)*

While exploring the dataset there were a number of potential problems that could have been picked. It offers a wide range of questions that can be answered using the video games dataset. the report will study and answer the below listed questions:

1. Which genre of the games sells the highest?
2. Trends for the sales of games in the selected dataset.
3. Is there an increase in the sales over the years?
4. Does any country has a greater sales than the others?
5. Predict the sales over the next few years using linear regression

Challenges and problems

One of the biggest problems that was encountered during the analysis is finding the suitable unsupervised method to implement on the variables. There are 14471 rows and 16 columns to establish a relationship. The selection of the method to select an supervised method is difficult and finally the Hierarchical clustering comes in play.

Another problem that was faced during the analysis was implementing the supervised method to help predict the sales of the games during the year 2000-2016. Even after going through a lot of different articles and websites, it was finally decided to fit the linear regression model. However, there was no useful result at the end of the supervised method.

Methodology used

The video game dataset was explored using the pandas and the numpy libraries in python. First of all the general metadata was explored to decide the variables to be used. After that, the number of columns and the other relevant data was extracted from the dataset. This subset was further explored and then used to analyse the rest of data. ‘Ygames’ has 14471 rows and 16 columns. The null values were checked and left the way they were because there were a large number of missing values. They were only removed when using the data to predict the sales of video games. There was an entry in the year 2020 which was erroneous. This had to be removed since this game was originally released in the year 2009. The error was confirmed from the IGN website(3). Many of the entries were verified by referring to this website. The Ygames data was further divided into smaller subsets to be used for the analysis. After the removal of the single data, the analysis part was carried out.

Data analysis

The data was analysed initially analysed in a general manner. The general statistics about the subset were discovered to further explore the data.

Which genre of the games sells the highest?

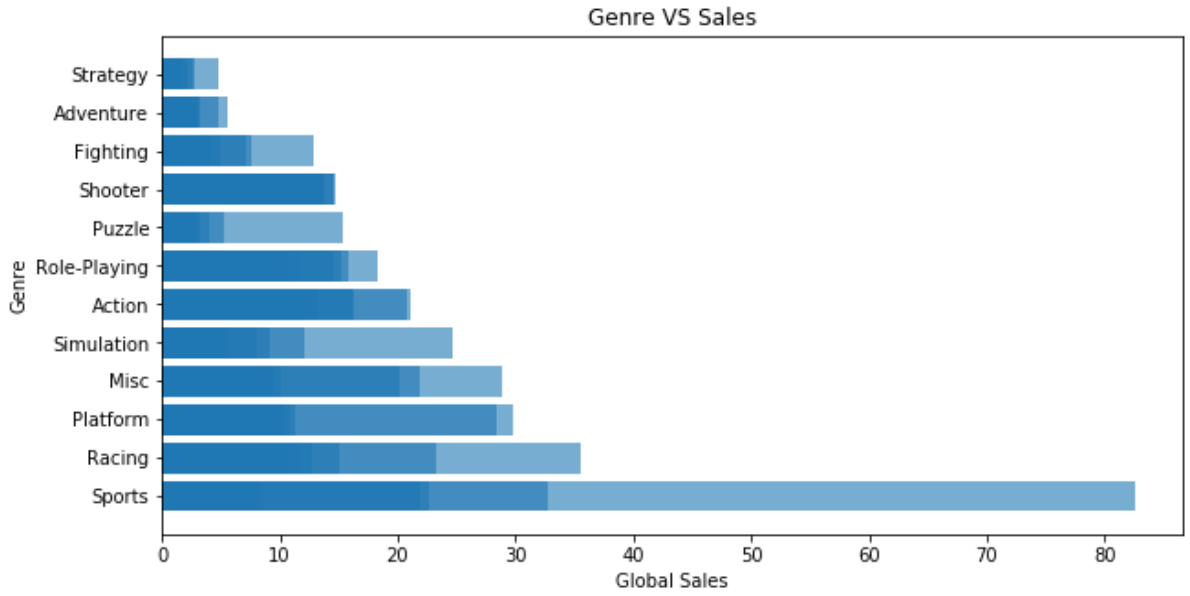

Figure 1: Genre vs sales graph

The graph in *figure 1* clearly explains the relationship between the genre and the sales of the particular genre. In the past two decades it is clear that the sports genre has the highest sales among all. The second highest selling genre is Racing but it is nowhere near the sports genre. The density of the sales is clear from the graph that most of the genres have been able to sell games near about 20-25 million copies or below that number. The least selling is the strategy games and the reason that can possibly support this statement is that games are supposed to be played for entertainment purpose. There are a very few number of gamers who prefer strategic games. However, there is an interesting fact to look at in this graph. When we look at the bar for the sales of shooter games, it clearly suggests that over the years it has been more consistent than the other genres. It only supports that the different sequels of the games such as Call of Duty, Battlefield are highly anticipated among the gamers.

The consistency of the Shooter genre can be better understood by the scatter plot of the same in *figure 2*.

Figure 2:Scatter plot for the genre vs global sales relationship

In the above graph, it is clear that shooter genre has a better consistency in sales when compared to the other genres. The sports genre has only one value that has sold the highest number of copies which is around 85 million copies sold globally.

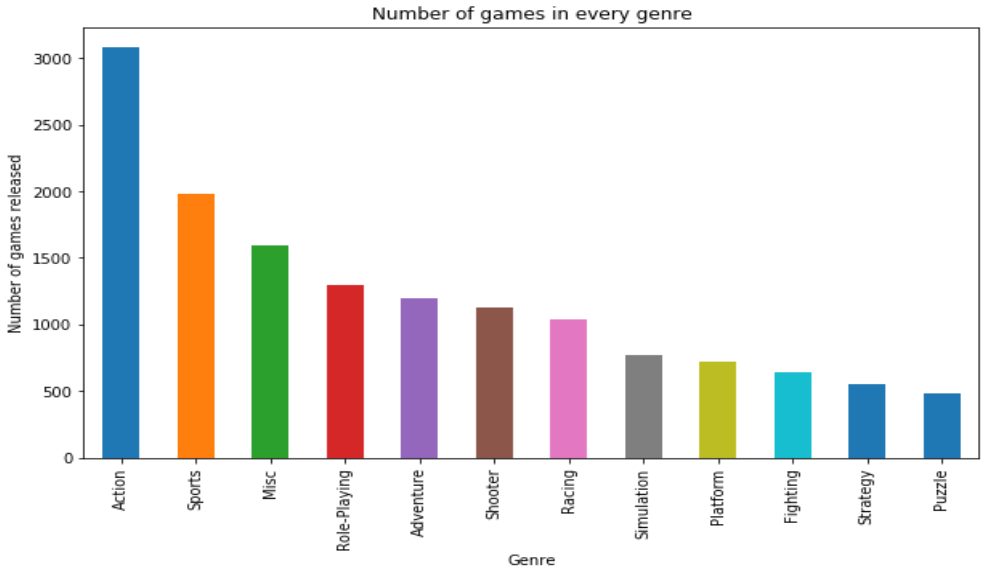

Figure 3: Games in every genre

When we look at the *figure 3*, it depicts that the number of games developed in the Action genre is the highest. Sports has the second highest number of games during the years 2000-

2016. The puzzle genre had the least number of games. Even though strategy games were the second to the last in production, this genre hasn't sold as per the expectations.

Trends for the sales of games in the selected dataset

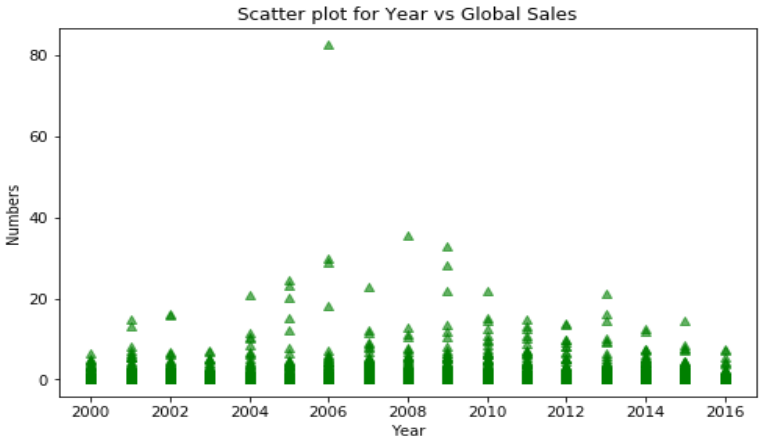

Figure 4: Scatter plot for Yearly Sales

The trends in the past two decades from this dataset don't show a very bright future for the sales of the video games. The sales show a downward graph in the future. The sales are almost uniformly distributed among the years, as shown in *figure 4.* The sports game that has the highest number of copies sold during the last two decades was in the year 2006. The density of the sales is the highest supposedly in the year 2011.

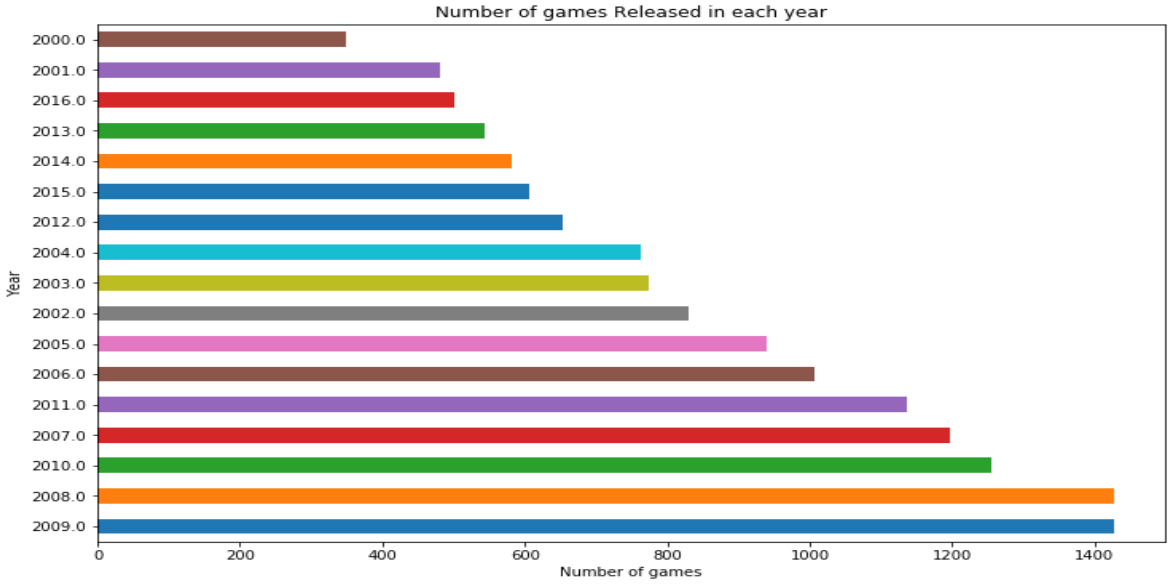

Figure 5: Number of games per year

Looking at the graph in figure 5, it is clear that the highest number of games were released in the year 2008 and 2009. One of the trends that can be noticed in the above graph is that the sales are almost every previous/next year the number of games released has been similar. Another trend that can be implied from the analysis is that the number of games published during the recent years has decreased. It can be supported by the fact that the development of a quality game requires a big budget. This can be a contributing factor to the reason there are a fewer number of games in the year 2016.

Considering the number of games every platform, the highest number was produced for the PS2 and DS. These have almost the same number of games for the respective platforms, as shown in *figure 6*. Over the time, it is clearly visible that the older generation of consoles has a lower number of games available in the graph. While the newer generation of consoles are not very far from the top spot, PC games have a low number of games since it has a wider approach to the audience.

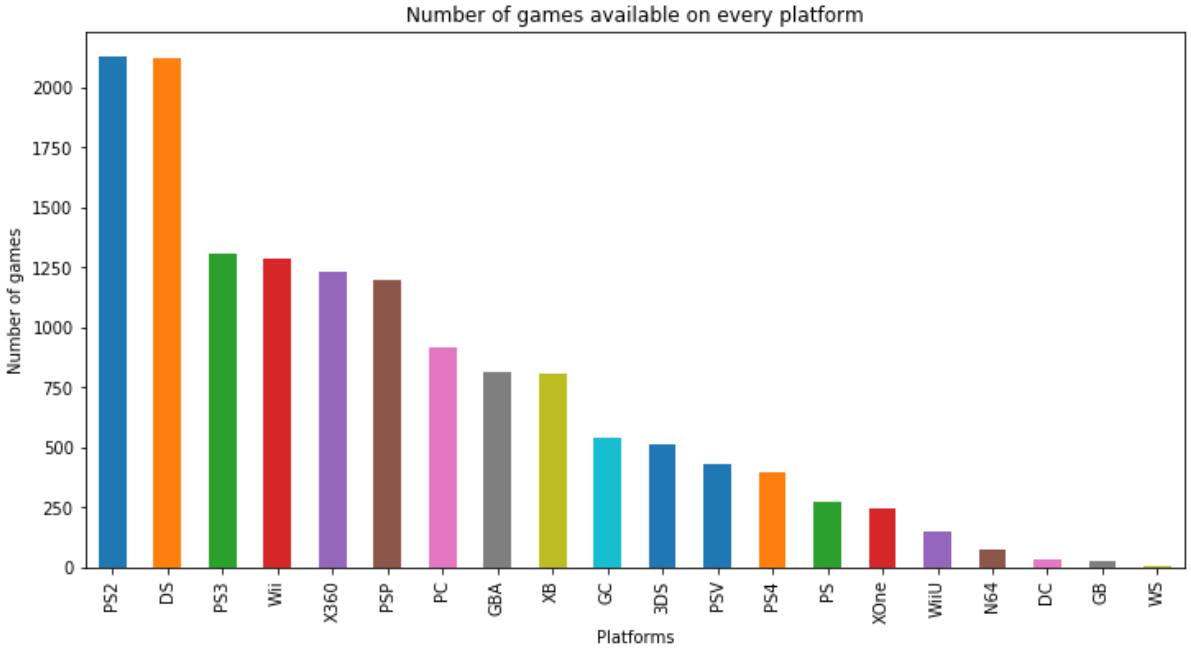

Figure 6: Games on every platform

Now, let us explore through the sales and genre in the past two decades split into two decades. This was done to understand the shift of players from the platforms as well as the genres that have been in production lately. It can be used to explain a broader prospect for the future of games.

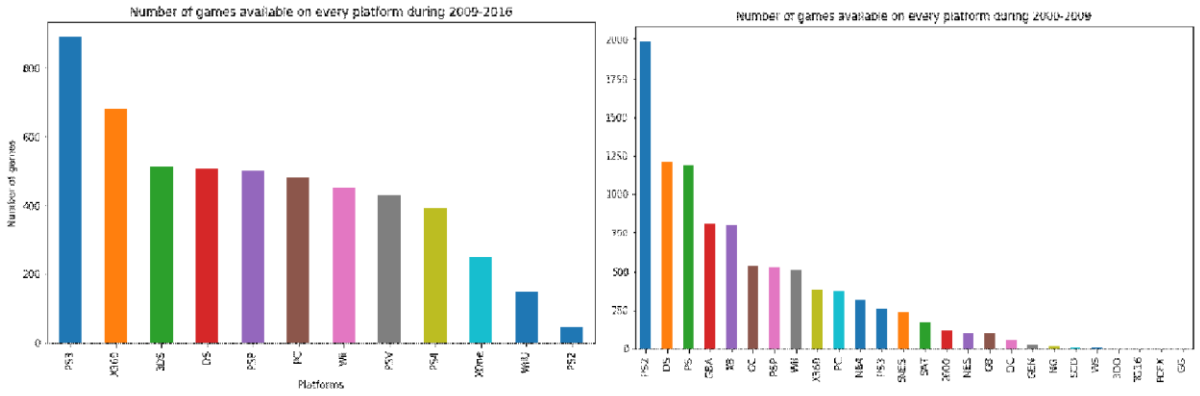

Figure 7: Number of games on platforms in 2000-2009(right) and 2000-2016(left)

The difference is clearly visible in the graphs, *figure 7*, which show that there was a large number of platforms available during the year 2000-2009.. It also depicts a shift of the mass number of users towards the modern technologies with much better and upgraded models. An intriguing fact that rises here is that the number of games during these years for the PC has not increased in a significant numbers. It has been the same rank as in the previous years. The dominant platform here are the Sony PlayStation(PS2 and PS3) platform games which have been the highest during all the years. An abundant number of Platform are missing in the graph of 2009-2016 indicates that these platforms were outdated and then ultimately their production was shutdown.

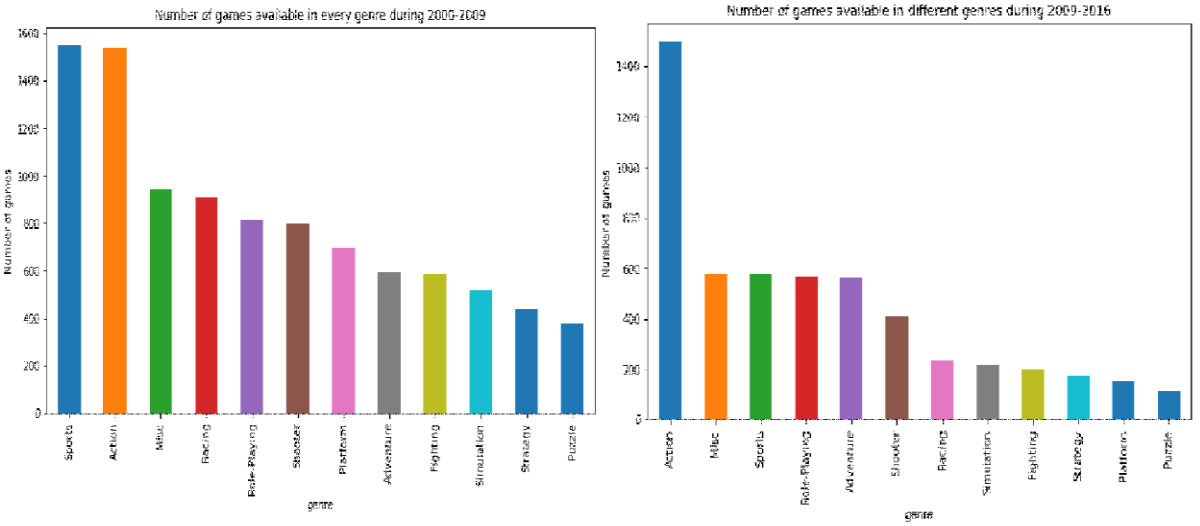

Figure 8: Genres available during 2000-2009(left) and 2009-2016(right)

It is clear from the *figure 8* that Action genre was more in production during 2009-2016. It is clear that action genre was one of the in-demand genres, while sports saw a decline in popularity in the latter part of the years. While the sports was the highest produced genre during 2000-2009, it lost its top spot to action genre and fell down to the third spot behind miscellaneous games. Sports games were reduced to less than half of the number that was produced in the years 2000-2009. The number of games produced has seen a decline in the recent years. The only genre that has not been affected by the fact are the Action games which have almost the same number of games produced even in the second part of the graph. This can explain a shift of mentality of the gamers from the sports genre to the action games. There are a large number of action games that have plenty of multiplayer options which can be played with an opponent from around the world or with a local player as well. In comparison to this part, there are not a large number of sports games which make this option easily accessible to the user or do not support this feature.

Is there an increase in the sales over the years?

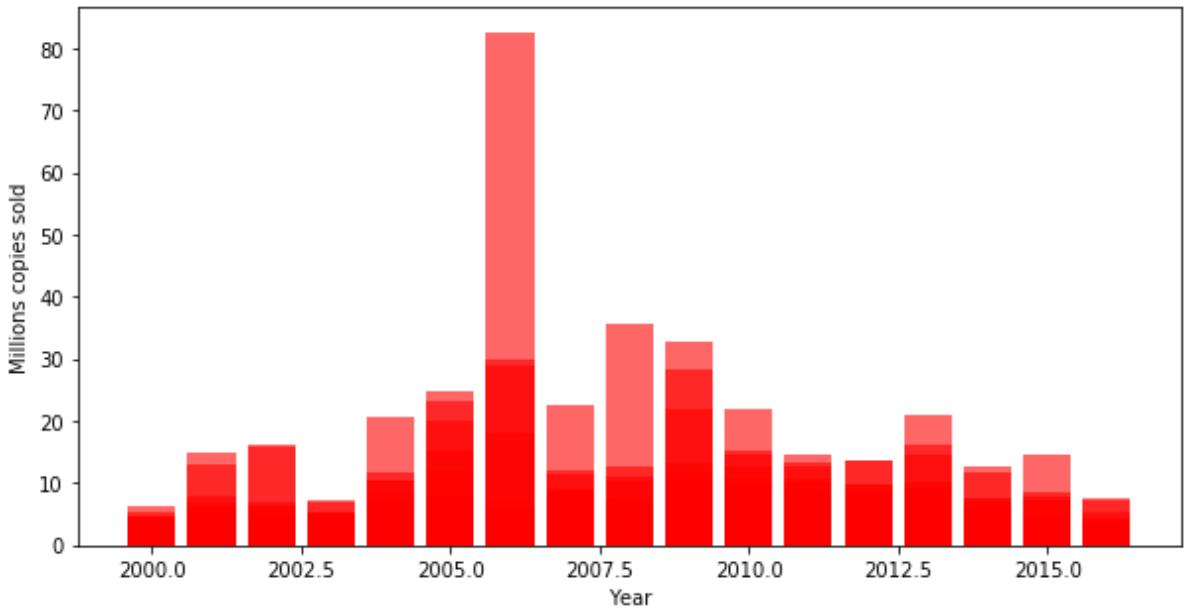

Figure 9: Global sales over the years

To check the sales, the global sales had to be plotted against the year of release so as to obtain the graph as in *figure 9.* The darker region shows the larger number of games that sold a similar number of copies across the globe in the same year. It can be observed that the number of games sold was highest in the years 2005-2009. The fact that the graph shows a unsteady downward slope for the sales can be a concern for the industry in the future. When compared this data to the most recent games released, 'Spider-Man(2018) on PS4', it broke the sales records by selling 3.3 million copies in 3 days. Since the game was only released on PS4 platform, it can be considered a big success and signifies a great future for action games in the future for the gaming industry. Earlier, the record was held by 'God of War(2018)' which sold 3.1 million copies in the first three days of its release*(3)*. It is evident from the fact that the future of games will be even more interesting to keep an eye on.

Does any country has a greater sales than the others?

When this question supposed to be answered, the grouping of the years and the sales according to the different regions of the world was one of the important tasks that was needed to be performed. Hence, the code*(4)* groups the years of release with the sums of the video game sales of every year with respect to the regions. There are a range of years that have been added which could not be removed even after making the subset of the data. To plot the

graph all of the years from the *games dataset* had to be taken into consideration because using the subset Ygames gave an error or the data was only visible until the year 2008 which happened in the case of the Japanese sales graph. This graph in *figure 10* shows the data for the rest of the world excluding the North American sales, Japanese sales and the European union sales.

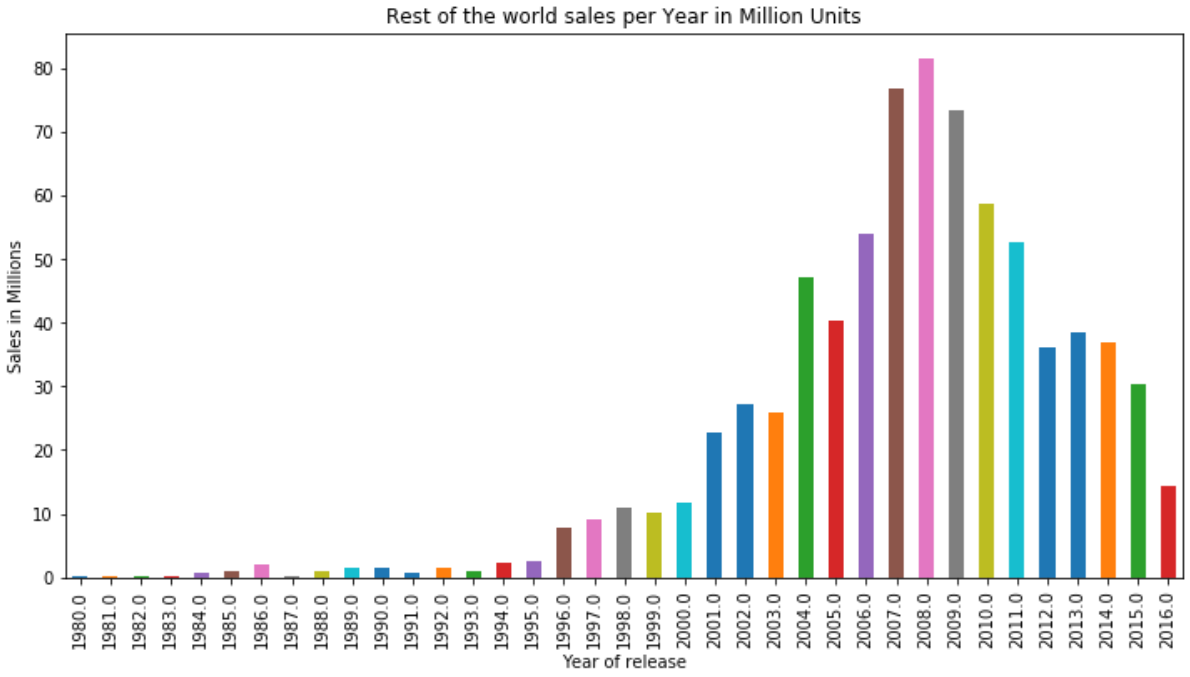

Figure 10: Rest of the world sales over the years

Over the years, it can be noticed that the sales has been jumping up in a significant numbers. The sales for the rest of the world, *figure 10*, shows an inconsistent performance after the year 2009. Even though the copies sold over the years have been decreased, this is not the only aspect that can be noticed with in the graph. 2016 year has a low number because it has only half of the games that were released in the year. If the data had been for whole of the year, it could have shown a different figure for the year 2016. With the likes of 'Call of Duty: Infinite Warfare' which was released in the same year and 'Grand Theft Auto V' which was released in the fall of 2013, were some of the biggest selling games of the year 2016*(5).*

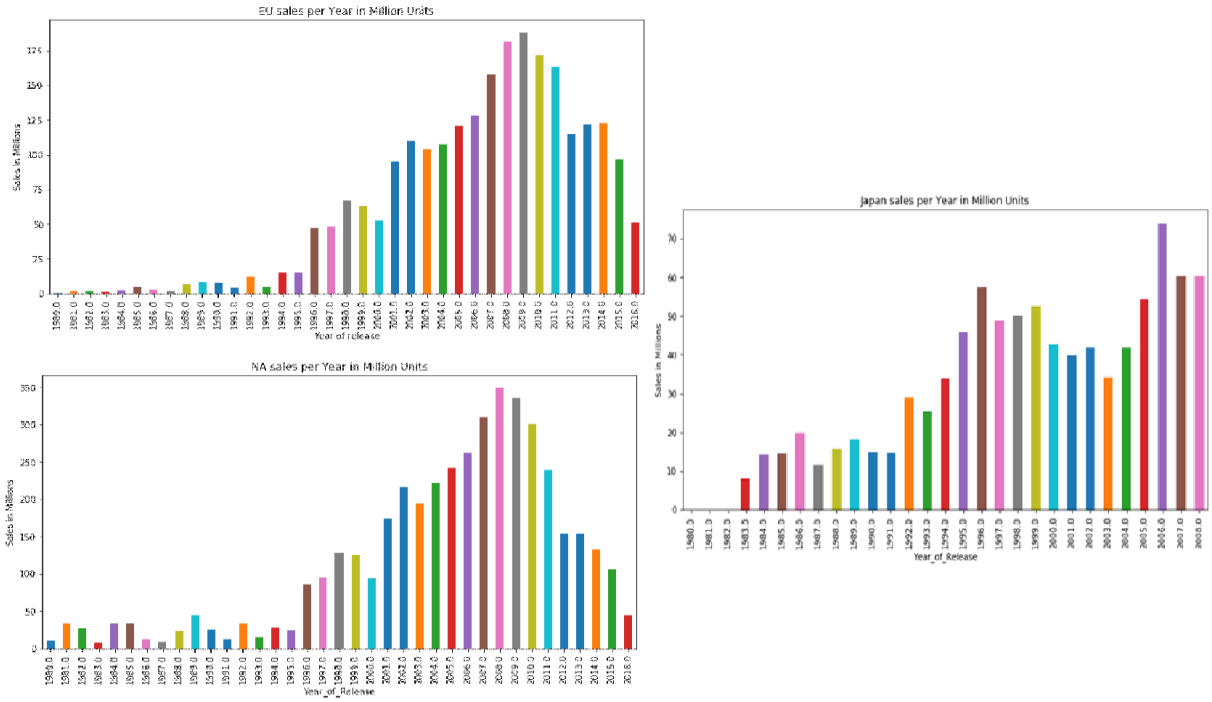

Figure 11: Sales in EU(Top left), North America(Bottom left) and Japan(Right)

The data in *figure 11* represents a wide range of sales in the different regions. One of the notable things is that Japan shows a potential market to work upon. It can be considered as one of the leading buyers of video games in the future. In comparison to the other regions Japan has more consistent sales over the years, while North America and European Union show a similar trend of sales. The sales in japan are only until the year 2008. Even on comparison within the same bracket of years, Japan comes up as a potential market in the future of video games.

Predict the sales over the next few years using linear regression

In this part of the analysis we will explore the depths of Unsupervised method used which is Hierarchical clustering and the supervised method, linear regression used to predict the global sales.

Supervised method: Linear regression

Linear regression can be explained as the model to fit the data and use the relationship between the variables to predict something. To apply the linear regression model, the dataset had to be divided into a smaller subset. Since there were a lot of null values and float data type values, the null values had to be removed and the data was converted to integer type values. The subset '*ucrid*' contained the variables Global Sales, Critic scores and the user scores after the removal of null values and modifying them to integer values. The data used

in clustering analysis was first plotted to check for any outliers which are shown in the *figure 12.*

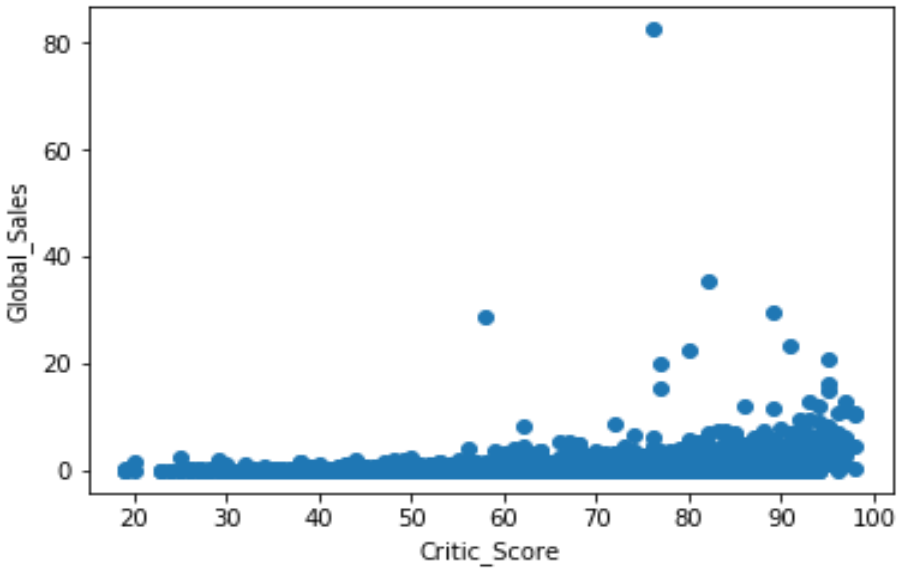

Figure 12: Outlier for clustering

It can be seen that the data has one clear outlier in this case. Since it had to be removed, the code*(7)* was used to remove the anomaly. The next graph in *figure 13* shows how the data looked after modifying the subset.

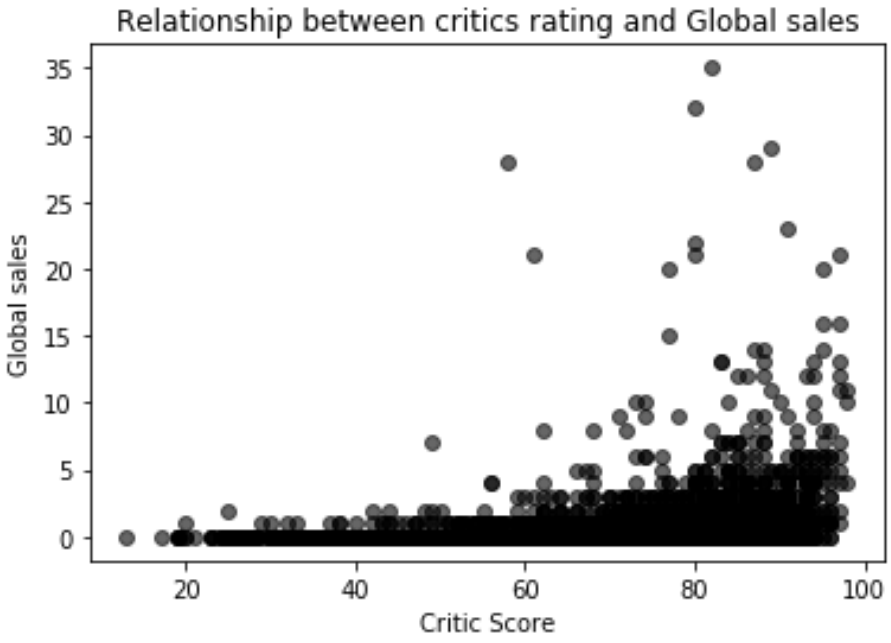

Figure 13: modified data

After this step, it could be easier to put the data into the model. The subset was split into 70:30 ratio where the test size was 30% of *ucrid*. A list of the features and coefficients was obtained as shown in *table 1*.

	Features	Coefficients
0	Global_Sales	1.000000e+00
1	Critic_Score	-1.667900e-16
2	User_Score	3.346446e-17

Table 1: Coefficients and Features

After this, the predict function was used to get a value `1.5052517937476325e-27`. This is the part where the prediction was faltered. At this point it was difficult to comprehend the value and plot the graph for the linear regression model as well.

Unsupervised method: Hierarchical Clustering

Hierarchical clustering or hierarchical clustering analysis can be described as-

"An algorithm that groups similar objects into groups called *clusters*. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other." *(6)*

The first step was to assign the subset '*ucrid*' to a variable U. This was then plotted which gave the same scatter plot as the linear regression model after removing the outlier. The data was then standardized using scale function from the *sklearn package* in Python. The clustering model was used then to make clusters from the variables Global_Sales, User_Score and Critic_Score. The Calinski-Harabaz and Silhouette scores were printed so as to validate and check the consistency over the variables. This confirmed the quality of the clustering method used. The clusters were then plotted to understand the clustering process using dendrograms as in *figure 14*.

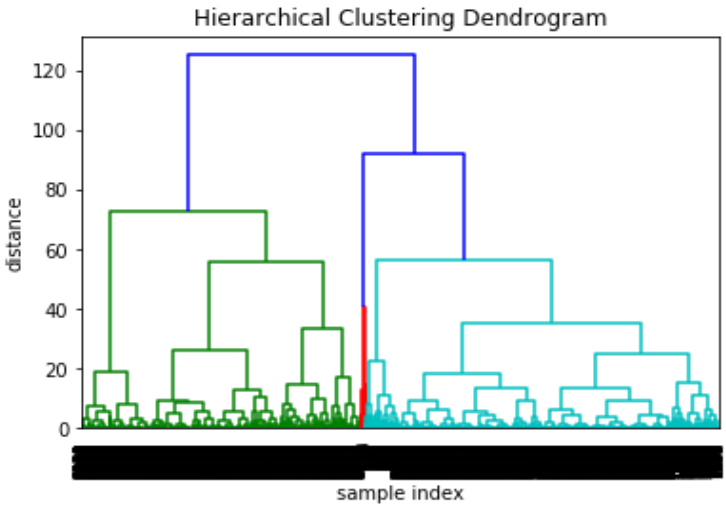

Figure 14: Dendrogram for Clustering the three variables

sTo explain the clustering better, a further method was used to reduce the complexity of the structure. The number of hierarchical clusters were reduced and truncated to show the last 12 points that were clustered using this model. The dendrogram in *figure 15* illustrates the above mentioned.

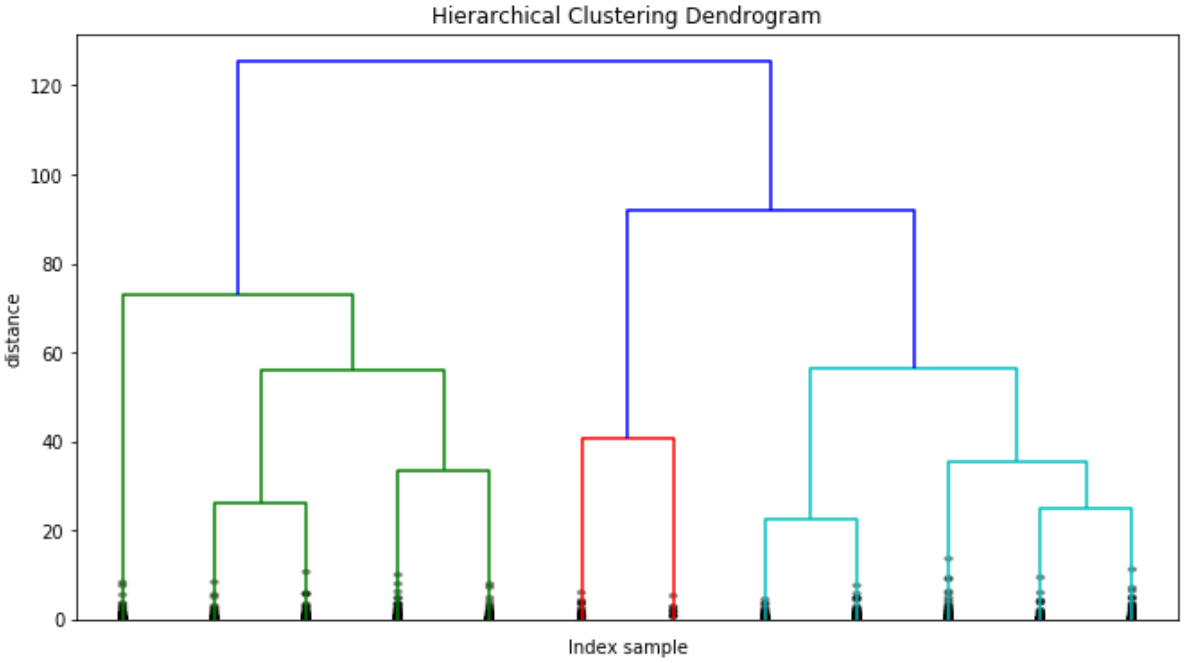

Figure 15: Clustering in the last 12 points

Reflections

After dealing with the video game sales dataset, it can be said that the selection of the unsupervised methods is a major prospect that needs to be further learned and explored. Also, the selection of variables to be used in the clustering was one of the issues.

For the linear regression model, there was a problem with identifying the use of final predicted value. There was a complication while plotting the prediction for the sales as well. It was easy to decide the model to be used for predicting the sales since linear regression uses the input variables to produce an output based on the test model dataset.

Conclusions

The report highlighted the use of different packages in python and the idea of analysing the video games dataset. The sales over the globe were analysed and it can be assumed that in the near future, the gaming industry can be one of the biggest grossing fields. Hierarchical clustering makes sure that the clusters are arranged in a hierarchy. The global sales and the critic scores are a good measure to be applied in the linear regression model. The sales would continue to grow if the games produced have a higher critic rating. This relationship satisfy the fact that the industry will keep on flourishing. However, with the number of games reduced, it will be more focused on the quality of the games, rather than producing abundant number of games.

References

(1) Kaggle.com. (2018). *Video Game Sales with Ratings*. [online] Available at: https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings

(2) Arnott, J. (2018). *Andy Serkis on Enslaved and acting in video games*. [online] the Guardian. Available at: https://www.theguardian.com/technology/gamesblog/2010/nov/08/andy-serkis-enslaved-interview

(3) IGN. (2018). *Imagine: Makeup Artist - IGN.com*. [online] Available at: https://uk.ign.com/games/imagine-makeup-artist

(4) Kaggle.com. (2018). *Video Game Sales EDA | Kaggle*. [online] Available at: https://www.kaggle.com/iliassekkaf/video-game-sales-eda#

(5) Tassi, P. (2018). *The Best-Selling Games Of 2016 Reveal A Few Surprises*. [online] Forbes. Available at: https://www.forbes.com/sites/insertcoin/2017/01/23/the-best-selling-games-of-2016-reveal-a-few-surprises/#2646aff0608e

(6) Displayr. (2018). *What is Hierarchical Clustering? | Displayr.com*. [online] Available at: https://www.displayr.com/what-is-hierarchical-clustering/

(7) Kaggle.com. (2018). *Forecasting Video Game Sales | Kaggle*. [online] Available at: https://www.kaggle.com/jruots/forecast

Appendix

Packages used and environment details

Python version: Python 3.6.5 | Anaconda3-5.2.0 Windows-x86 (64 bit) Development Environment: Jupyter Notebook Copyright © 2018 Project Jupyter

Packages: Numpy, Pandas, Matplotlib, Seaborn, Scipy, Sklearn, pandas.tools.plotting, sklearn.linear_model, sklearn.preprocessing, sklearn.cluster, scipy.cluster.hierarchy

Appendix 1

‘Name’- The name of the game

'Genre'- The genre to which a particular game belongs

'Publisher'- The company which publishes the games that have been made by an internal developer or an external one.

‘NA_Sales'- Sales of the games in North America

'EU_Sales'- Sales of the games in the European Union

'JP_Sales'- Sales of the game in Japan

'Other_Sales'- This variable indicates the sales over the rest of the world excluding Japan, North America and the European Union.

'Global_Sales' – The total sales all over the world

‘Critic_Score’ – the aggregate score by Metacritic.

‘Critic_Count’ – the number of critics to come up with the score

'User_Score' – the aggregate score by the users on the Metacritic website

‘Developer’ – this indicates the company responsible for creating the game

‘Ratings’ – ESRB ratings (Everyone, Teen, Adults Only, etc.)