02450: Introduction to Machine Learning and Data Mining

Project 1

Group 65

Authors:
Gabriel Luo (s221400)
Christian Lund (s203962)
Nil Palau (s222953)

October 4, 2022

Group Member Contributions

	Section 1	Section 2	Section 3	Section 4	Exam Questions
s221400	30%	30%	40%	30%	33.3%
s203962	30%	40%	30%	40%	33.3%
s222953	40%	30%	30%	30%	33.3%

Table of Contents

1. A description of your data set.

The dataset was extracted from the website Kaggle. The data represents a collection of the ratings and statistics of over 16000 unique video games since 1985. The data also contains information on the sales in various regions over the world, as well as overall global sales. It also shows the scores the game received from both critics and users. Other attributes such as the genre and publisher are also included.

The overall problem of interest is to predict one of the attributes based on other available information. For example, being able to predict the score for the video game received based on the year of release, sales, genre, and publisher. By building a machine learning model, video game publishers/developers may be able to utilize the resulting trends to determine what genres future games should be focused on, or which region's sales have the greatest impact in terms of overall profit and how well the game is received.

A reference of where the data was obtained can be obtained through the following link.
https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings

In terms of summarizing a previous analysis of the data, consider the dataset linked below:
https://www.researchgate.net/publication/297754899_The_Impact_Of_Platform_On_Global_Video_Game_Sales

This specific dataset hasn't been used in a previous analysis, however, a publication titled "The Impact of Platform on Video Game Sales" was found to use similar data. The paper looked at global video game sales by platform from 2006 to 2011. They concluded that Nintendo was by far the most dominant in the period, with the Wii and DS having by far the highest and second highest average number of sales ($8.37 million and $5.06 million). In addition to this, they also had the highest single-selling games, with the Wii having 15 games exceeding $20 million in annual revenue. This is more than any other platform's (other than the DS) single highest selling game.

We would like to be able to predict how well a game would sell in the Japan region based on its sales in other regions, platform, critic score, rating and more. We know that most games are released before in regions like North America, so by creating a model that can predict this, companies can save money on commercializing by predicting their potential gains in a region.

Based on the regression matrix, we would like to create a model to predict the sales in Japan based on all the other attributes including other regions' sales, scores given by both users and critics and attributes such as the platform (e.g we could expect a Nintendo game to sell more than other ones). For that reason, we think that all the other attributes will play an important role.

In the classification task, we will predict what critic score a game received based on the sales, user score, platform, genre, etc. To create "classes" for this, we are going to divide the scores into 3 different ranges, for example bad, average and good. We will then predict which range each game falls into.

We removed global sales, number of critic reviews, number of user reviews and names. We removed global sales as it was just a sum of the other sales. Other than being redundant, this also would have made it too easy for our model to predict how many sales each game got in Japan. We removed the number of reviews and names of the games as we found them irrelevant for our model.

We also multiplied the values for the sales by 1000, so that instead of being in millions the sales were represented in thousands. This was done as most of the games had less than 1 million, making the value of the attribute smaller in scale than the rest of the attributes, which could produce some bias in the future when we train our models (a huge difference in scale between attributes can bias our model's learning). The user score value was multiplied by 10 to match the score of the critic scores for more consistency.

2. A detailed explanation of the attributes of the data.

Table 1 below provides a summary of the attributes and their properties.

Table 1: Summary of the attributes and their properties

Attribute	**Discrete/continuous**	**Nominal/Ordinal/Interval/Ratio**
Platform	Discrete	Nominal
Year of Release	Discrete	Interval
Genre	Discrete	Nominal
Publisher	Discrete	Nominal
NA Sales	Continuous	Ratio
EU Sales	Continuous	Ratio
JP Sales	Continuous	Ratio
Other Sales	Continuous	Ratio
Critic Score	Discrete	Ordinal
User Score	Discrete	Ordinal
Developer	Discrete	Nominal
ESRB Rating	Discrete	Nominal

Note that the values for sales were set as continuous because even though we could say that the possible values for them are finite, the difference between the number of games sold can be infinitesimal from one row to another, thus we can consider all sales continuous.

Around 9000 of the data points were missing critic and user scores. These games were omitted from the dataset as future analysis relied on using these scores for analysis, and even after omitting these data points, there was still sufficient data to perform analysis on. The same is true for the ESRB ratings, but luckily most of the games which didn't include either the critic score or user score also didn't include ESRB ratings, and so only a few more data points were removed. There was an additional few games which were missing data points for the year, publisher, and genre. To simplify the reduction, the .dropna() function was used in Python so that all data points which had missing values in one of the attributes were removed from the data set. The result is that just under 7000 data points remained, which is sufficient for further model development.

Table 2 below shows a summary of statistics for chosen attributes.

Table 2: Summary of statistics for chosen attributes

Attribute	**Mean**	**Median**	**Min**	**Max**	**Standard Deviation**
Year of Release	2007.44	2007.00	1985.00	2016.00	4.21
NA Sales (in thousands)	394.48	150.00	0.00	41360.00	0.97
EU Sales (in thousands)	236.09	60.00	0.00	28960.00	0.69
JP Sales (in thousands)	64.16	0.00	0.00	6500.00	0.29
Other Sales (in thousands)	82.67	20.00	0.00	10570.00	0.27
Critic Score (out of 100)	70.27	72.00	13.00	98.00	13.69
User Score (out of 100)	71.85	75.00	5.00	96.00	1.44

Note that the mode for attributes is not included as the mode for the sales was 0 and didn't provide much information. This makes sense as it is unlikely that any two games would have the exact same number of sales, but it is likely that multiple games may not have sold anything in the Japanese region for example. As such, having a mode of 0 doesn't provide any insightful information on the attributes or the data and was therefore removed.

3. Data visualization(s) based on suitable visualization techniques including a principal component analysis (PCA).

In Figure 1 below we can see the box plots of the sales across the regions we will consider, including our target, the sales in Japan. As we can see, we have many outliers, which would be known as AAA games, games created by large companies that, regardless of the quality, always sell large amounts.

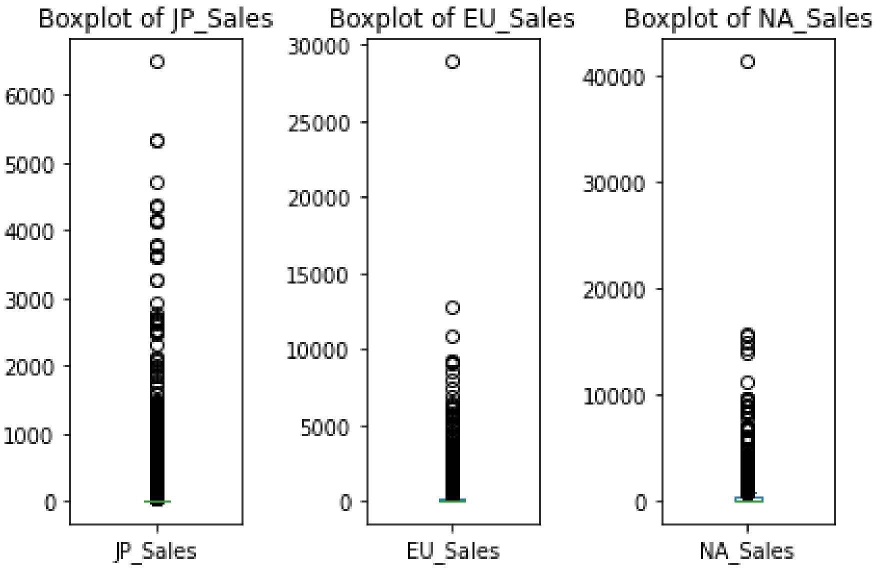

Figure 1: Box plots of sales

We can clearly observe the presence of outliers. Using quantiles (percentiles) we will remove the top and bottom 5%. After doing this removal of outliers, we can see the resulting boxplot data in Figure 2 below.

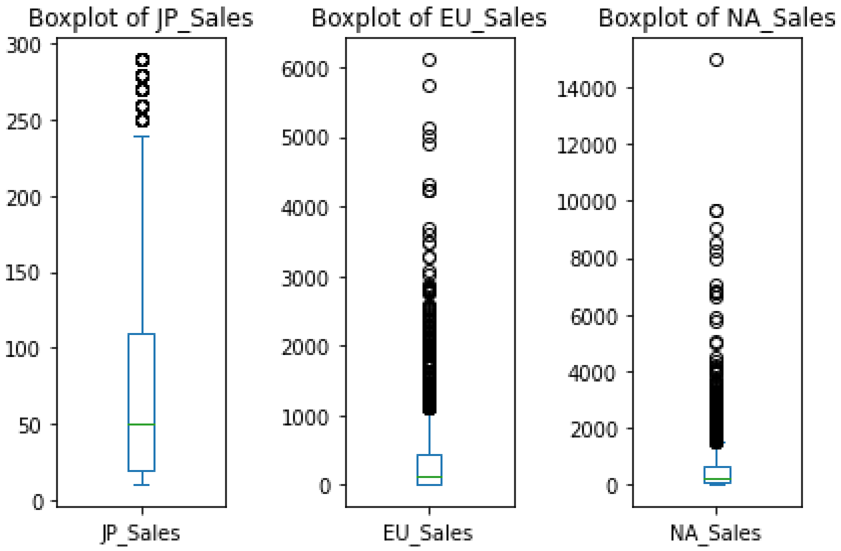

Figure 2: Box plot of sales after the removal of outliers

In the Appendix, we can see the density distribution of all the attributes we can find in the dataset.

As can be seen, not many of the different attributes are normally distributed. At first sight, we can discard all the attributes that do not follow a normal distribution: Platform, Genre, Publisher, Rating and Developer. Most of them were encoded and hence do not express a continuous variable.

On the other hand, we can see that the sales attributes do not follow a normal distribution either since most of the games from our data did not sell over a million copies, while a few games sold up to 10+ million. It can be concluded that only the attributes which express scores (either given by users or critics) do follow a normal distribution, with a remarkably high mean in both cases. This can clearly be seen as shown in Figure 3 below. In the JP sales, we can see that the left side is significantly denser than the right side of the mean, making it not normally distributed. The user score, however, is very evenly distributed around its mean showing that it is clearly normally distributed.

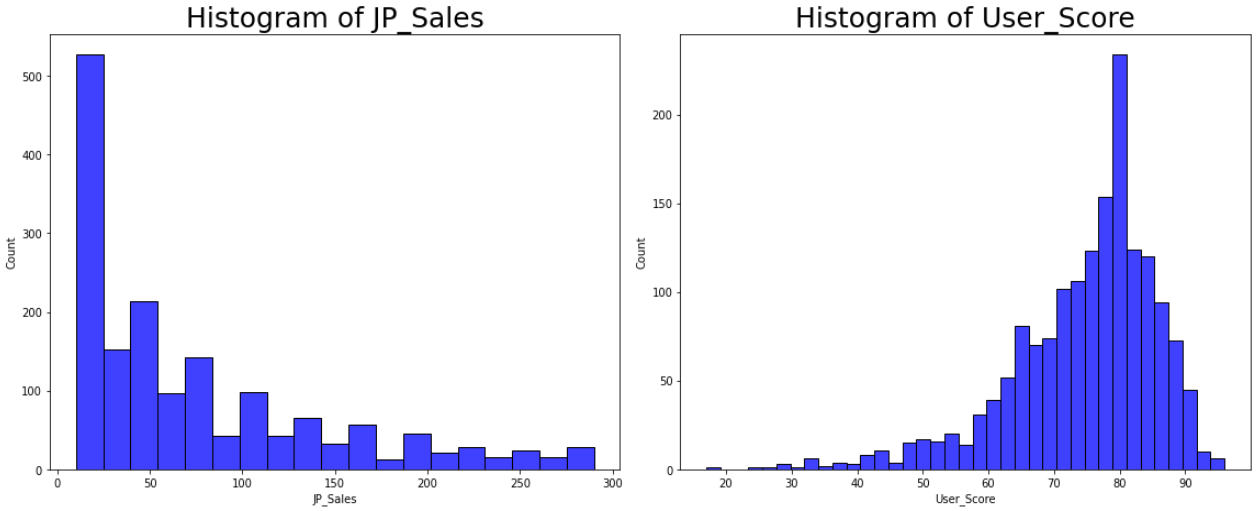

Figure 3: Histogram of JP sales and user score to check normal distribution

The correlation matrix for our data and its attributes is shown in Figure 4 below.

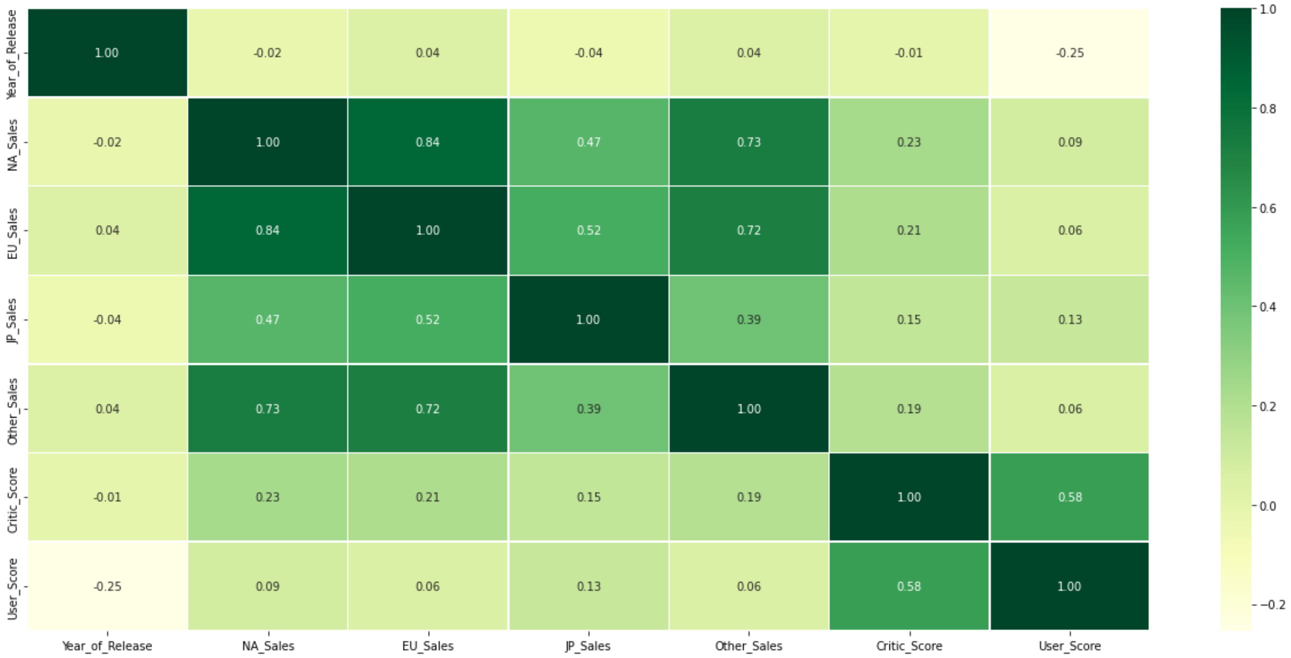

Figure 4: Correlation matrix

In general, our variables are very highly correlated. The sales variables are especially highly correlated. This is to be expected as games that sell well in one region tend to sell well across the world. Furthermore, the critic score is positively correlated with all sales while the user score is barely correlated at all. This suggests people are more likely to listen to the opinions of critics than other users when deciding whether to buy a game or not. Lastly, user score is inversely correlated to the year of release. This is very interesting as it shows users have been giving worse scores over time, suggesting people's standards are growing faster than the quality of games.

Seeing that there are high correlation values between the different market statistics with the Japanese market, we could conclude that creating a model to predict the performance of a game in the region of Japan will be highly feasible.

Before carrying out the PCA analysis, we removed all our encoded attributes. This included genre, platform, publisher, developer and rating. We did this as all their values were random numbers in ascending order from 1 and therefore their value had no significance. In addition, because the scales for the attributes were widely different, the data was standardized using the mean vs the standard deviation.

From the plot shown in Figure 5 below, the first PCA component accounts for 87% of the explained variance. The plot also includes a threshold line which indicates the value of 90%. The second PCA component accounts for much less variation, around 12%, but it brings the cumulative explained variance to around 98% for the first two components. This is useful as it means that 98% of the 8-dimensional data that was started can be represented with only two axes. From the plot, it can be seen that the remaining 6 PCA components account for little and make a relatively insignificant impact on the total explained variance when combined with the other components before it.

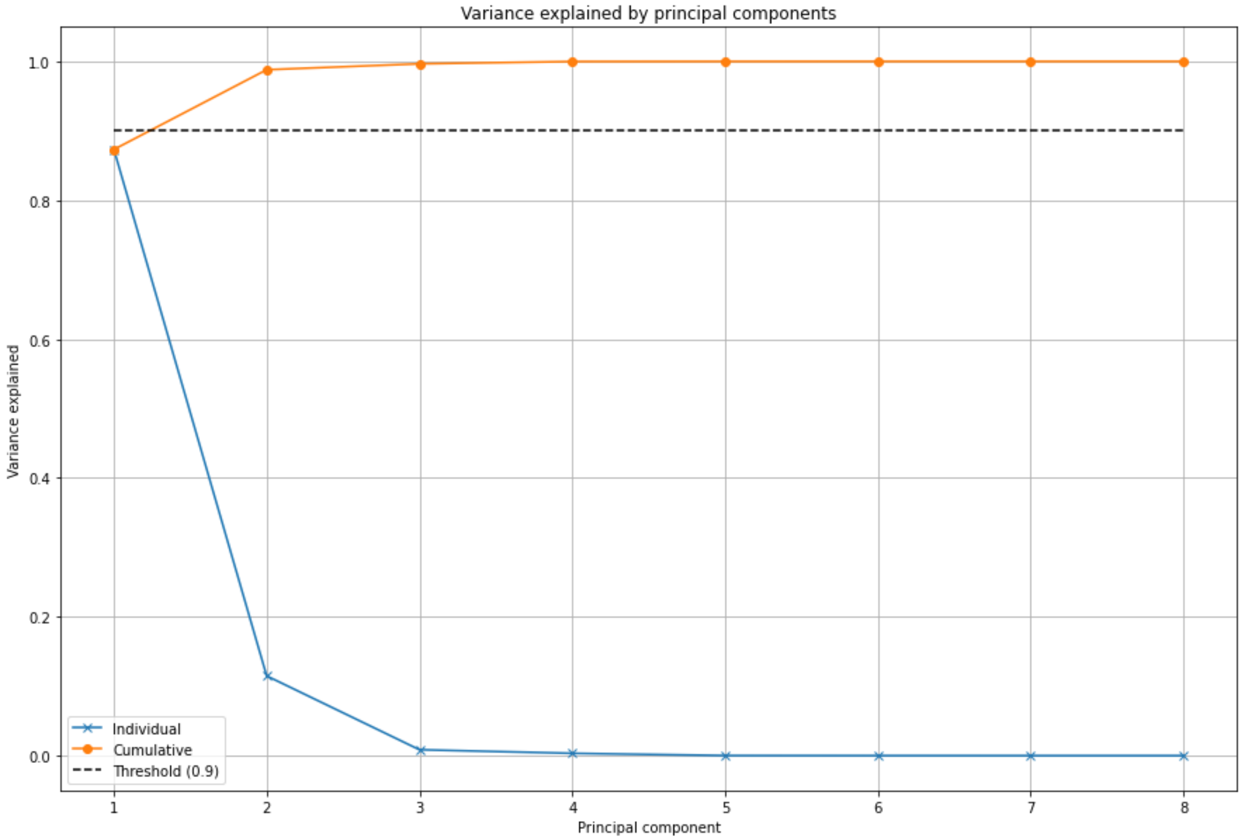

Figure 5: Plot of variance explained by the principal components from PCA

As such, the amount of variation explained as a function of the number of PCA components is shown in the figure and shows that the first two PCA components can account for around 98% of the data.

The histogram shown in Figure 6 below shows the principal directions of the first two PCA components. Only the first two are shown here as from above it was determined that the first two components account for 98% of the data, which was deemed acceptable in terms of representing the data set.

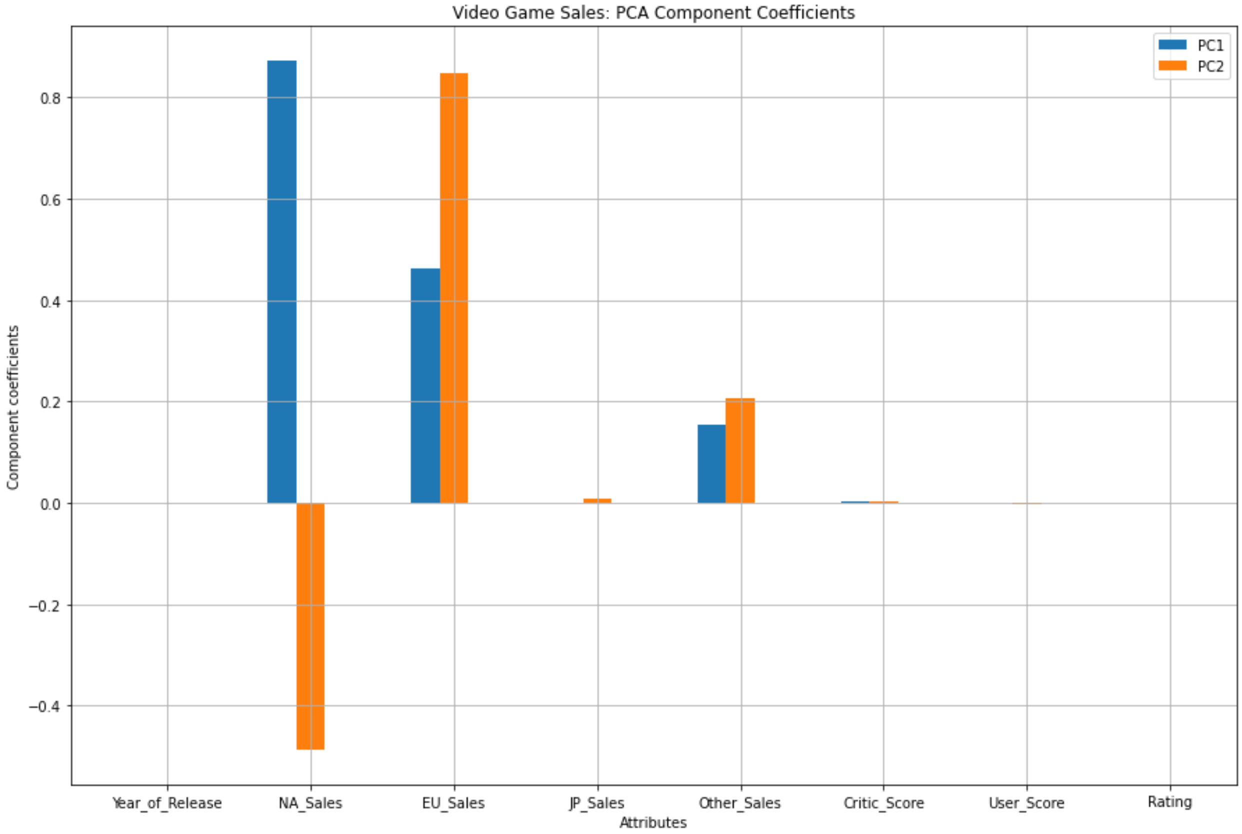

Figure 6: Histogram showing the coefficients of the attributes as applied to the first two PCA components

From the plot, the coefficients for Year of Release, JP Sales, Critic Score, User score, and Rating are very small and have little effect on the two principal components. The coefficients for NA Sales, EU Sales, and Other Sales have a much more significant impact. For the first principal component, the values from NA sales, EU sales, and other sales will have a significant effect. In addition, high values in the attributes mentioned above would guarantee a positive projection onto the first component since all the coefficients are positive. The attributes of NA Sales, EU Sales, and Other Sales will also have significance for the projection onto the second component. However, unlike the first component, the coefficients aren't all negative, meaning that depending on which values are higher the projection onto the second component may vary between being positive or negative. For example, a high value for NA Sales and a low value for both the EU and Other Sales would result in a negative projection.

The plot shown in Figure 7 below shows the data when projected onto the first two PCA components. By projecting onto these two components, the data is visualized more easily compared to when the data is left with eight dimensions. Since it was found that the explained variance from the first two components was around 98%, the projection onto this plane will not lose much information. The other benefit of using a PCA visualization is that it makes it easier to see any groupings or trends in the data. One thing to note is that a sample of 500 points was used for the plot as plotting all the points proved to be too messy and made it difficult to see anything since all the points were on top of each other.

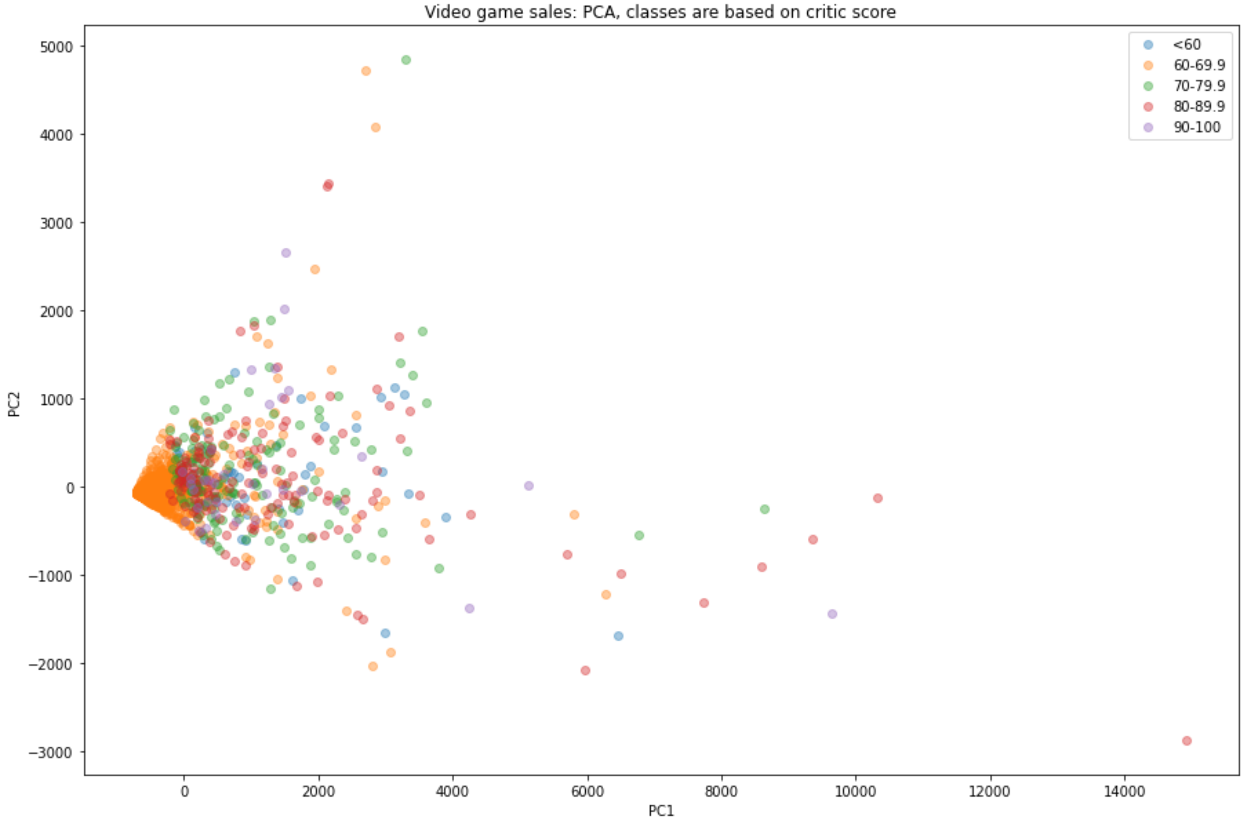

Figure 7: Plot of data on the two primary PCA components

The data points were classified based on the critic score ratings, and the different classes are shown in different colours as indicated by the legend on the plot. The critic score ratings were chosen to match our classification goal as mentioned earlier in the report. The different classes were divided in increments of 10 for the value of the score, except for the lowest group where all the scores below 60 were grouped together. This was done based on the distribution graph for the critic score shown in the Appendix which shows that the data is centred around a mean of around 70. If values below 60 were also grouped into increments of 10, most of the classes under 60 would have very few points and wouldn't add much insight.

From the plot, most of the games with a score between 60 and 70 are grouped together in the bottom left corner. The rest of the score classes are relatively randomly distributed, seemingly diverging from the group made by the games with scores between 60 and 70. This shows some insight that the classification for the games based on critic scores can be done relatively easily if determining whether it is in the range of 60 to 70. Otherwise, the plot shows that it will be relatively difficult to classify games in any other ranges for the critic score since they are so widely distributed.

4. A discussion explaining what you have learned about the data.

When first looking into the data, it was realized that some of the columns contained many null values, so deleting these columns before deleting all the rows with nulls was the first crucial step to take into consideration. After seeing the high correlation between the sales in different regions, it was decided that it could be feasible to try to predict how a game would perform if launched in a new region, based on the performance of other regions. A real application that could become an interesting tool for game studios.

It has been observed that some regions’ sales are greatly affected/correlated by the reviews given by critics of the sector, while in others, there has not been any correlation at all. Some consistency from the critics was also observed since the year of release of the video game does not correlate with the received score, which could suggest that all games were objectively evaluated.

Based on the PCA plot, it can be seen that most of the data can be represented with only two dimensions. When projected onto this plane, there is a grouping of games with critic scores ranging from 60-70 that can be seen. Other than that, it is difficult to distinguish any other trends or groupings in relation to the critic scores.

To conclude, we believe that the implementation of a first learning algorithm will be successful since we possess a great amount of quality data with a high correlation between attributes, making the development of an accurate algorithm for predicting the sales in the region of Japan feasible.

Problems

Problem 1. Option D. To see this, consider each other parameters listed. For the time of day, because there is a measurable distance between objects (ie the difference between 12:00 and 3:00 can be measured) and that zero does not mean an absence of what is measured, it is an interval. For traffic lights, it is a quantitative value and since zero represents an absence of traffic lights, the parameter is a ratio. Running over is a ratio because it represents a quantity of run-over accidents and zero represents an absence of these accidents. For the congestion level, it is more qualitative, there is no definite quantitative value associated with it. The congestion level can still be ranked so it is ordinal.

Problem 2. Option A. To solve this, the following equation is used

$$d_\infty(x,y) = \max\{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_M - y_M|\}$$

The equation represents the max-norm distance, or the p-distance when p = infinity. The values for x_{14} and x_{18} can be plugged into the equation as shown below

$$d_\infty(x,y) = \max\{|26 - 10|, |0 - 0|, |2 - 0|, \ldots, |0 - 0|\} = 7.0$$

The outputted value is 7, which matches option A.

Problem 3. Option A. The variance explained by a certain number of components can be calculated by the following equation:

$$\frac{\sum_{i=0}^{K} \sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2}$$

For option A, the singular values for the first 4 principal components are 13.9, 12.47, 11.48, and 10.03. By plugging these values into the Equation above, the following evaluation can be computed:

$$\frac{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} = 0.8668$$

The variance explained is around 86.68%, which is greater than 0.8, so statement A is true.

Problem 5. Option A. To solve this, consider the equation to calculate the Jaccard Coefficient, or Jaccard similarity represented by $J(x,y)$

$$J(x,y) = \frac{f_{11}}{K - f_{00}}$$

The equation is composed of three variables. F11 is defined as the number of attributes where x_k=y_k=1. Between the two text documents, there are only two words which appear it both documents: “the” and “words”. Hence, f_{11} = 2. K is defined as the number of attributes, or in this case the vocabulary size. This was defined in the problem as M = 20000 = K. Finally, f_{00} is defined as the number of attributes where x_k=y_k=0, or the number of words which don’t appear in either of the texts. In the first text, there are 8 unique words. In the second text, there are 7 unique words. Together there are 15 words. However, given that it has been established that two words appear in both texts, 15 - 2 = 13 unique words between the two texts. Therefore, the number of words that do not appear in either of these texts is 20000 - 13 = 19,987 = f_{00}. Putting all these values into the equation above gives a result is 2/13, or 0.153846, which matches the value for option A.

Appendix

Figure 8 below shows the histograms for each of the twelve attributes, used to determine if they are normally distributed or not.

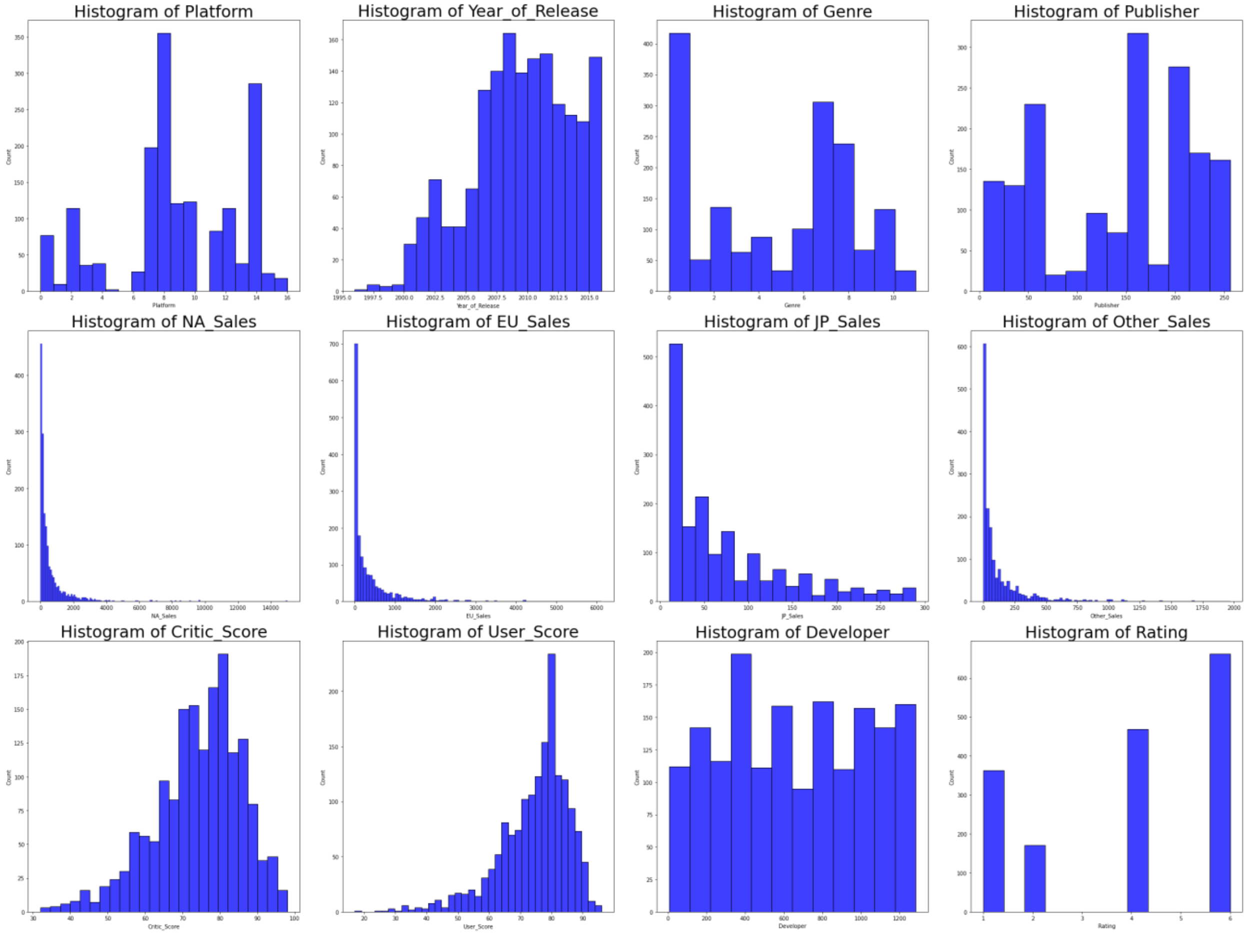

Figure 8: Density histograms for the 12 attributes