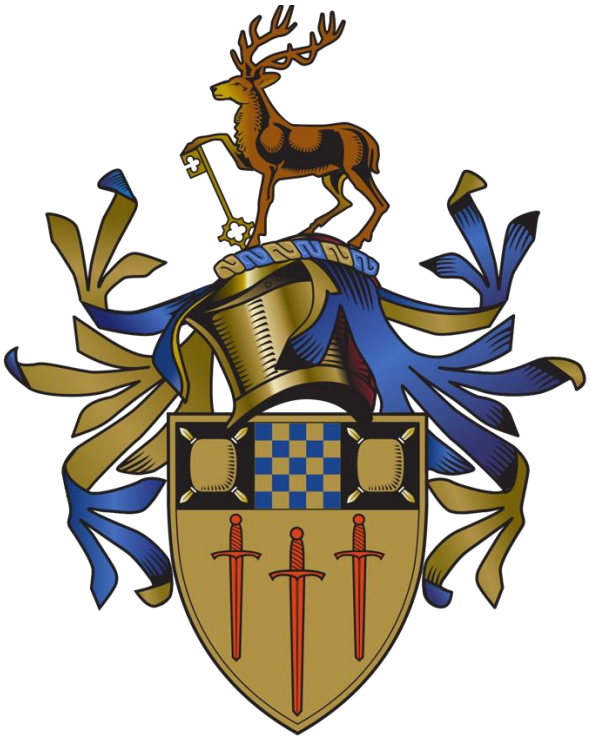
University of Surrey

Faculty of Engineering and Physical Science

Department of Computer Science

COM3018 Practical Business Analytics

Video Game Trend Analysis Report



Group Name: Nantendo

Group members:

Anastasia Anichenko,	Manuel Bradicic,	Matthew Samm
Felix Olesen,	Michal Sitarz,	Alex Williams

Guildford, 28th November 2022

Declaration of Originality

We confirm that the submitted work is our own work. No element has been previously submitted for assessment, or where it has, it has been correctly referenced. We have clearly identified and fully acknowledged all material that is entitled to be attributed to others (whether published or unpublished) using the referencing system set out in the programme handbook.

We agree that the University may submit my work to means of checking this, such as the plagiarism detection service Turnitin® UK and the Turnitin® Authorship Investigate service. We confirm that I understand that assessed work that has been shown to have been plagiarised will be penalised.

Table of Contents

1 Introduction 1

1.1 Problem Definition 1

1.2 Dataset 2

1.3 Hypothesis and Objectives 2

1.4 Methodology 3

2 Methods 5

2.1 Additional Data 5

2.2 Data Cleaning 6

2.2.1 Not Available Values 6

2.2.2 Detecting and Removing Outliers 6

2.3 Field Transformations 7

2.3.1 Field Encoding 7

2.3.2 Derivation of New Fields 7

2.4 Exploratory Analysis 8

3 Models 11

3.1 Regression 12

3.1.1 Relationship of Regional Demographics on Sales 12

3.1.2 What affects the sales? 13

3.1.3 What do franchises tell us about sales? 14

3.2 Neural Network (NN) 14

3.3 Random Forest 15

3.4 Time Series Sequential Classification (TSC) 15

3.5 Clustering 16

3.5.1 Additional Pre-processing 16

3.5.2 K-Means Clustering 17

3.5.3 Hierarchical Clustering 18

4 Results 19

4.1 Regression 19

4.1.1 Linear Regression: Demographics 19
4.1.2 Linear Regression: Importances 21
4.1.3 Logistic regression 23
4.2 Neural Networks 25
4.3 Random Forests 26
4.4 Time Series 27
4.5 Clustering 30
4.5.1 Hierarchical Clustering: Dendrograms 30
4.5.2 Hierarchical Clustering: Scatter Plots 31
4.5.3 Hierarchical Clustering: Dendrograms 35
4.5.4 Hierarchical Clustering: Scatter Plots 36
5 Discussion 40
5.1 Which Game to Make? 40
5.2 Do Franchises Affect Game Sales? 40
5.3 When Should the Game be Released? 41
5.4 Do Regions Affect Sales? 42
6 Conclusion 45
6.1 Future Improvement 45
7 Bibliography 47
Appendix A – Project Plan 49

Table of Figures

Figure 1 Global Video Games Sales over Years 1

Figure 2 Median Age (2021) 6

Figure 3 Game Frequency over Years in Database 9

Figure 4 (a) Average Global Sales per Platform (b) Average Global Sales per Genre 9

Figure 5 (a)Average Global Sale per Rating (b) Games per Global Sales bin 10

Figure 6 (a) Average Japan gender distribution, affecting Japanese Sales (b) Average gender distribution in 'other' regions affecting the sales 19

Figure 7 (a) Average European Age Distribution, affecting European Sales (b) Average age distribution in 'Other' regions affecting the sales 20

Figure 8 Average Age Distribution in North America against North America Sales 21

Figure 9 Strengths of Linear Model for Japan 22

Figure 10 (a) Logistic Classifier, probability of in Franchisese over GlobalSales(x) (b) ROC curve 24

Figure 11 (a)GlobalSales over Predictions (b) NN accuracy and loss over epochs 25

Figure 12 Random Forest model for sales prediction 26

Figure 13 Time Series Graphs for Platforms 28

Figure 14 Time Series Regional Graphs 29

Figure 15 Time Series Japan Sales trend for Role-Playing platforms 29

Figure 16 Time Series- Neural Network (a) Scoring History (b) Actual vs Predicted 30

Figure 17 (a) Silhouette Index Plot for Strategy (b) Cluster graph for Strategy Genre 31

Figure 18 (a) Cluster Graphs for Role Playing games (b) Cluster Graphs for Genre Strategy 31

Figure 19 Dendrogram for Role Playing Genre 35

Figure 20 (a) Scatter plot of clusters for regional Japan and North America sales for action genre (b) Scatter plot of clusters for regional EU sales and North America sales for action genre 36

Figure 21 (a) Scatter plot of clusters for regional Japan and North America sales for role playing genre (b)Scatter plot of clusters for regional EU and North America sales for role playing genre 37

Figure 22 (a) Scatter plot of clusters for regional Japan and North America sales for shooter genre (b) Scatter plot of clusters for regional EU and North America sales for shooter genre 38

Figure 23 Project Plan.....49

Summary of Tables

Table 1 Summary of Models.....12

Table 2 Linear Regression Training R^2 results 22

Table 3 Overall Importance of Features in the Linear Model for Japan.....23

Table 4 Confusion Matrix.....24

Table 5 Random Forest optimal paths to the best sales.....27

Table 6 Profile Summary for Best and Worst Strategy Clusters 33

Table 7 Profile Summary for Best and Worst Role Playing Clusters.....35

Table 8 Summarising Significant Cluster Profile Details for Action Games Genre 37

Table 9 Summarising Significant Cluster Profile Details for Role Playing Genre.....38

1 Introduction

For the following group project, our team has decided to take a role as an up-and-coming video game company, Nantendo. We want to find trends in game sales to pick up what type of game to produce and where to sell it.

1.1 Problem Definition

Throughout the past decades, video games have been increasingly entertaining people all around the world. The yearly revenues generated globally by video games are around $155 billion, with over 2 billion users [1].

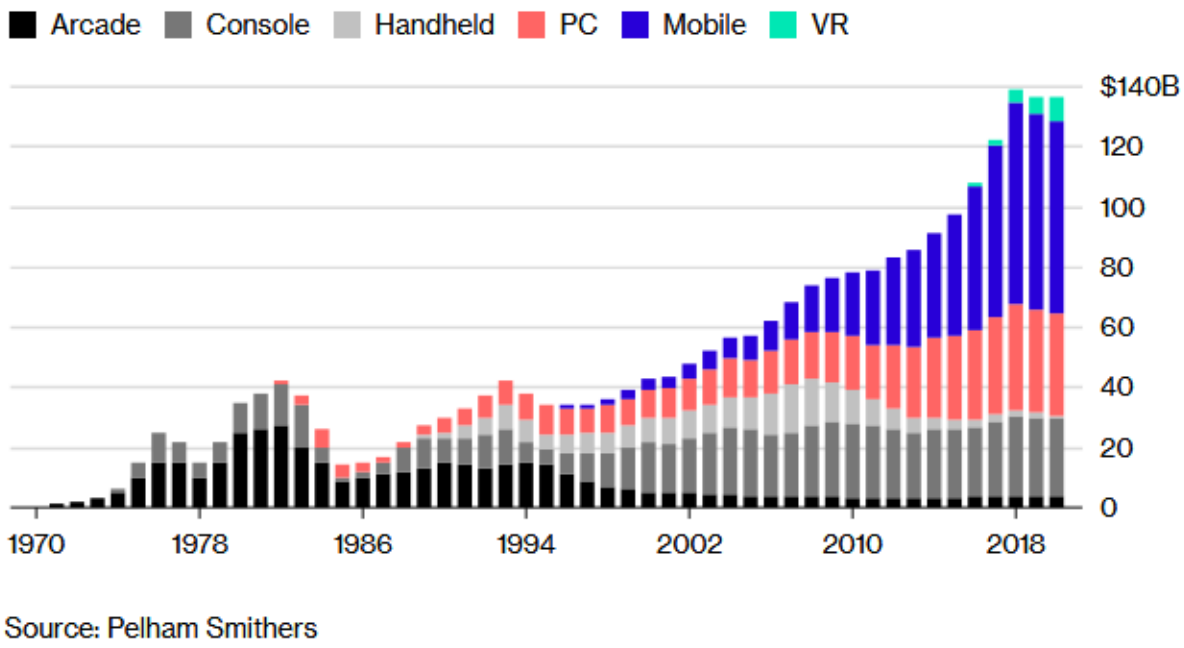


Figure 1 Global Video Games Sales over Years [2]

With the largely growing industry, it might be often difficult to enter the market as a start-up. Many new games are often overshadowed by the number of games constantly released and ones produced by large studios. Furthermore, large companies often have their own data analytics teams working on data that is not publicly available [3], giving them an advantage over smaller companies that do not have past experiences and data from the market. It is especially done when a company owns smaller companies and many games, therefore they have a lot of data on what works with their games and what doesn't. There are different studies also done by using different data, such as forecasting game sales based on YouTube trends [4], which type of game it is [5] and more. Therefore, we decided on taking a games sales dataset

from Kaggle [6] and analysing trends in the data. Equipped with the knowledge identified in the data, it will be possible for us to design and distribute a new game.

1.2 Dataset

The dataset contains over 11,000 game entries which span across various genres and platforms. The games are also spread out over the years, ranging from 1980 to 2020. With that knowledge, we should be able to make predictions and find correlations between how sales are affected. The dataset is comprised of the following fields:

- Name of the game
- Platform
- Year of Release
- Genre
- Publisher
- Sales across different regions
- User and Critic Scores
- User and Critic Counts
- Ratings
- Global Sales
- Regional Sales (North America, Europe, Japan and Other)

Furthermore, we found datasets containing gender [7] and age demographics [8], which can be combined on the year of release column in the games dataset we are using. With that, we will have the demographics of the region for when the game got released. The data from the game's dataset was scraped from Metacritic [9] . For the missing data entries, we decided to do some data scraping to fill the missing values. Finally, the data that we contain can be used to produce some new data. We used the names from the dataset to check if a game is in a franchise or not.

1.3 Hypothesis and Objectives

The main objective of this study is to find trends which maximise revenue for a game, and thus determine what type of game will be most profitable to make. Our assumption is that the games that are well reviewed are going to be sold more, and hence generate more profit. However, we believe there are other factors that can contribute to how well the game is sold,

as usually the reception of the game cannot be controlled beforehand. Our hypothesis is that *the sales will be increased by trending genres in regions, sold on specific platforms*. In other words, we believe that the type of game produced and sold on specific platforms will perform better than other genres. The prediction is that action or role-playing games, especially ones designed by large companies (i.e., AAA games), will have the highest sales as they will have the biggest reach, most money on marketing, large budget and large development teams working on the games. We think there will also be a correlation between whether a game is in a franchise or not, as usually a community of players is built behind such games over time.

Likewise, we believe that different regions will have diverse genre preferences which will also affect the sales. Different cultures will have varied gaming cultures within them. For instance, Northern America might have a higher popularity of action games, whereas Japan might have a higher sale of role-playing games. Likewise, we think that countries with younger populations will have more game sales, as well as populations with higher male populations will have higher sales, especially for action games. The regional trends can be identified by utilising the regional data of sales, with the combined data of regional demographics. This will allow us to make assumptions which regions are suited for the game we will want to produce. This can affect the marketing and sale of those games in those regions. Finally, the times that the games are released should also have an impact on how the game sells. This is since people should be more likely to buy games before Christmas, or right before summer.

1.4 Methodology

In our Project Plan, we identified which techniques we were going to use and how the workload would be distributed among the team. We divided ourselves into pairs and worked on different techniques in that way. We first completed data pre-processing and cleaning (further discussed in *Methods*), after which all different groups investigated different modelling techniques to identify trends. The initial plan was to look at the following models:

- Regression
- Clustering
- Neural Networks and Time Series

The use for regression is that it can allow us to explore the dataset, identify trends in the dataset, and then dive deeper into which features impact the sales. Secondly, our data has a lot of dimensions, so with clustering we can identify which groups of features affect the sales and it allows us to exclude our biases and assumptions when trying to find relations. Finally, neural

networks were used to identify unseen patterns and correlations in the data, going a step beyond linear regression models. The models and their outcomes are further discussed later. After performing some initial tests on these models, we also concluded that some didn't perform as well as we expected so we also tried looking into Random Forests as a contingency plan, which proved to be more successful given that we have a lot of categorical data.

2 Methods

2.1 Additional Data

One of the steps we took was to add data to the original dataset. The first method we focused on was finding additional data that we can use. As mentioned in the introduction, our hypothesis is that the regions where the games are being sold might have an impact on the sales. The datasets that we found contain the age median over the years (as it can be seen in *Figure 2*) and the gender distribution for a country over the years.

The main issue with this was that it is by countries and we need it by region, in order to fit the regional sales we have. Therefore, we found a dataset [10] which maps the countries to regions, and with some data manipulation we were able to group the age and gender data by region, and then join it with the games data by the year. In that case, each game will have a corresponding field which will identify what the demographics was for that year. We also had to remove some NA values for the continents as some of the older data was missing (around 1950s). However, the lack of that data doesn't have much significance as most of the games came later when the data was already there. The data also provides the future predictions for both of these datasets, so if there was a correlation, it could be predicted how the game will do in the future based on the regional demographics.

The second approach was to scrape the MetaCritics website in order to find the missing data and get some additional information. The rows had plenty of missing values, so we scraped for them in order to not have to drop so many games. For instance, critic scores NA values were decreased by over 10% and year of release for around 50%. Furthermore, the original dataset only contained the release years, but we wanted to obtain the release dates as well. Hence, we used the scraper to get the release dates, which were used for the Time Series analysis, as the years themselves were not enough for it.

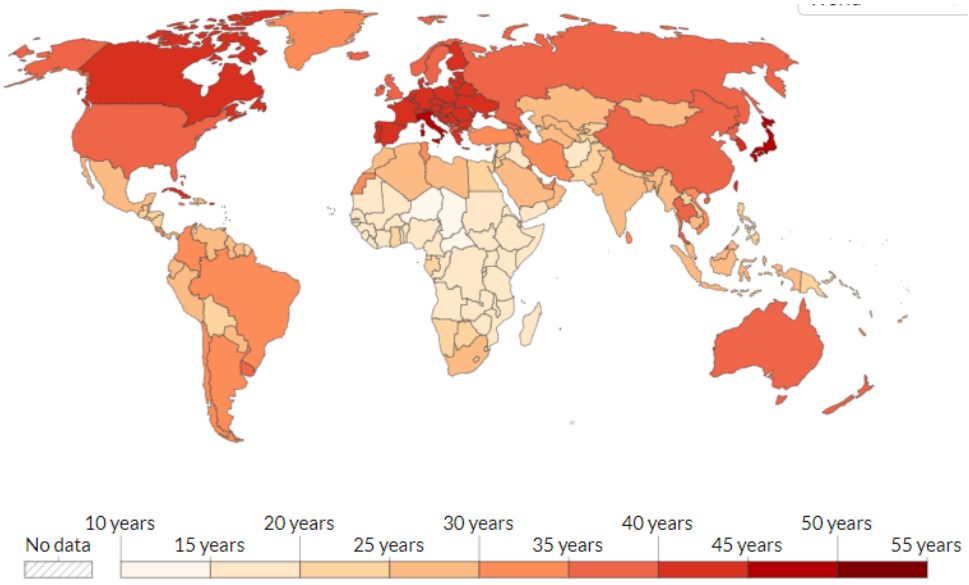


Figure 2 Median Age (2021) [8]

2.2 Data Cleaning

2.2.1 Not Available Values

Our chosen dataset initially contained 9895 not available (NA) values. One of the challenges was identifying NA values as a lot of them were stored as a variety of different characters including empty strings, “tbd”, “N/A” and “RP”. Once identified we unified these values to the same NA value in R for ease of use. They are important to identify and handle as it can affect the outcome of our models by limiting the amount of data models can be trained on. As mentioned above our main approach to handling NA values consisted of scraping the web for the missing data. This reduced the amount of NA values by over 10% for critic score and around 50% for year of release. Unfortunately, the addition of the release dates also increased the total number of NA values in the dataset as a lot of games seemed to lack those. We accepted this increase as the release date data would only be used for Time Series analysis and could be dropped for all other models as we saw fit.

2.2.2 Detecting and Removing Outliers

A large portion of data cleaning involved detecting and removing data outliers. To be able to remove the outliers, the data was first sorted into symbolic and numerical fields, then the numerical fields were separated further into discrete and continuous fields. We then applied the chi-squared test method for detecting outliers to the continuous data. This test determined the goodness of fit of each field based on an outlier confidence value that we manually tuned.

Once the outliers were identified and plotted, they were then removed from the dataset. We made the decision to remove the outliers rather than impute or cap the outlier values as we believed that this would skew our sales data in a way that wouldn't provide any extra intrinsic value once applied to the models down the line.

2.3 Field Transformations

2.3.1 Field Encoding

One method of encoding that we used was one-hot encoding. Lots of the fields in our dataset were categorical/symbolic and thus have no meaningful order. We used one-hot encoding for both our genre field, and the platform field. E.g. The genre of game could be "Racing" or "Action" which do not have an objective order. We transformed this into multiple fields, one for each category. A value of one is assigned to the field corresponding to the category the record is in; all other fields are assigned zero.

When encoding our sales fields, we had two approaches, we could normalise the fields so that they lie between 0 and 1 or we could use binning. These encoding methods transform the data into suitable values that the machine learning algorithms can work with and give meaning to.

2.3.2 Derivation of New Fields

A metric that was not covered by the dataset that we thought would carry insight into a game's performance is its inclusion in an existing game franchise. This is based on the anecdotal evidence that games such as Call of Duty, FIFA, and Pokémon tend to get high sales irrelevant of the popular opinion of the game. This new value for each game was a simple true/false that indicated a game's inclusion in a franchise, no note was made of which franchise a game was deemed to be a part of.

Games were grouped into perceived franchises based on the similarity of their titles. This was done by a script which first removed common words such as 'the', 'and', 'a', etc. before comparing each word in the title with every other game title in the dataset. From this an array of related games was formed for each game, with some having no relation and others having over 20 related games. At this stage, we had to decide how many related games would earn a group the rank of franchise; the number of games that we agreed upon was 3 related titles. This gave the script enough precision to remove outliers, but also keep in game franchises that only

have a few games in them. The final count was 4700 games categorised as belonging to a franchise, roughly a quarter of the games in the dataset.

It is important to note that there is a degree of error in the franchise detection. Due to the method of searching for related games, there were inevitably some false negatives and false positives. This means that while the majority of franchise analysis will be accurate, there is a greater chance for anomalous data trends.

2.4 Exploratory Analysis

For the data exploration, we investigated the data distributions, in order to visualise how the data is spread in different features. This exploration has been done to identify which data can be used for training the models, and to give us insight into some basic trends.

The first step which we did was to check the number of games which were released over the years. Even though the dataset contains some newer data, it only contains a couple of entries, and therefore, the number of games released peaks around 2011 and then steadily decreases (see *Figure 3*). The expectation would be for the data to keep increasing consistently. The most possible reason is due to the fact that most new data was not scraped, and therefore makes the newer data less reliable to work with, in the context of time series. Furthermore, this implies that the trends which will be found might have some variety from the current ones as they might be outdated and the global sales for those periods will be lower than the previous ones, and thus, this should be considered in the evaluation.

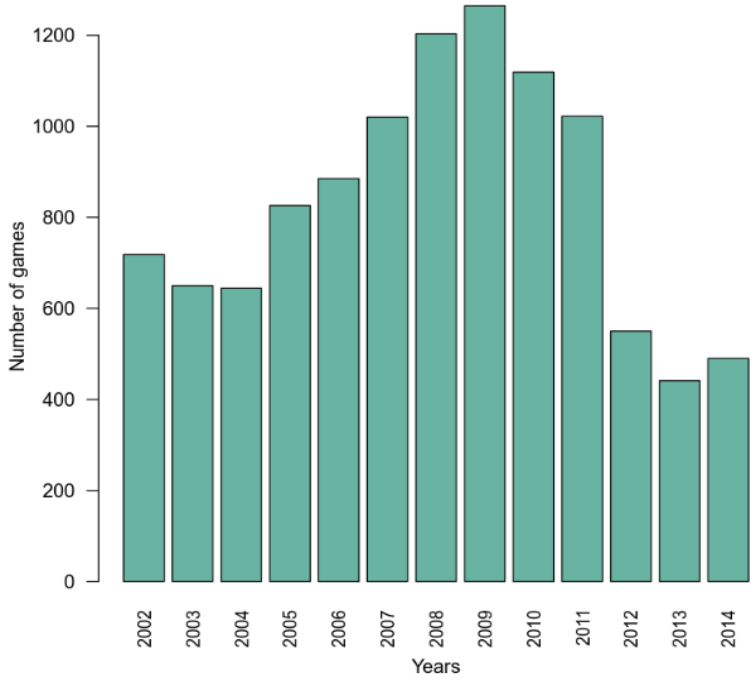
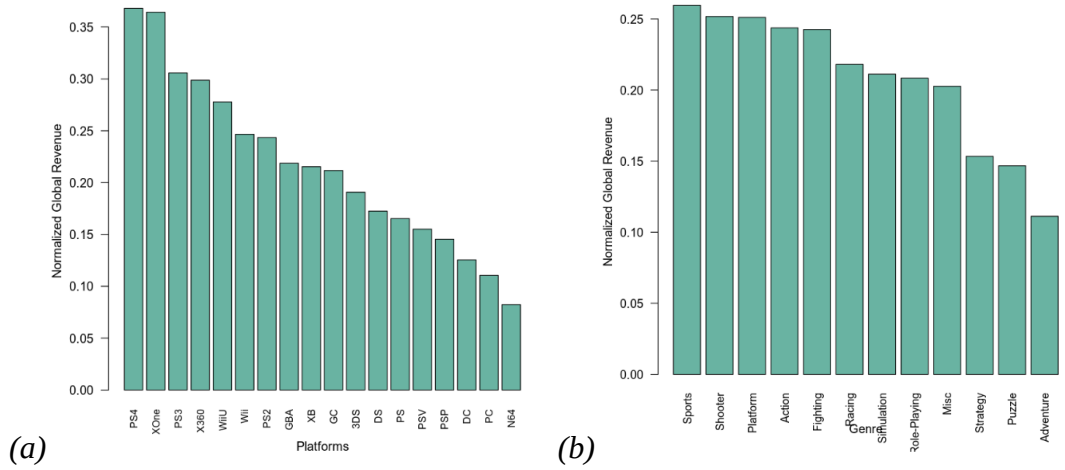


Figure 3 Game Frequency over Years in Database

Afterwards, we turned our attention to investigating which genres and which platforms influenced the global sales the most (*Figures 4(a),(b)*, respectively).

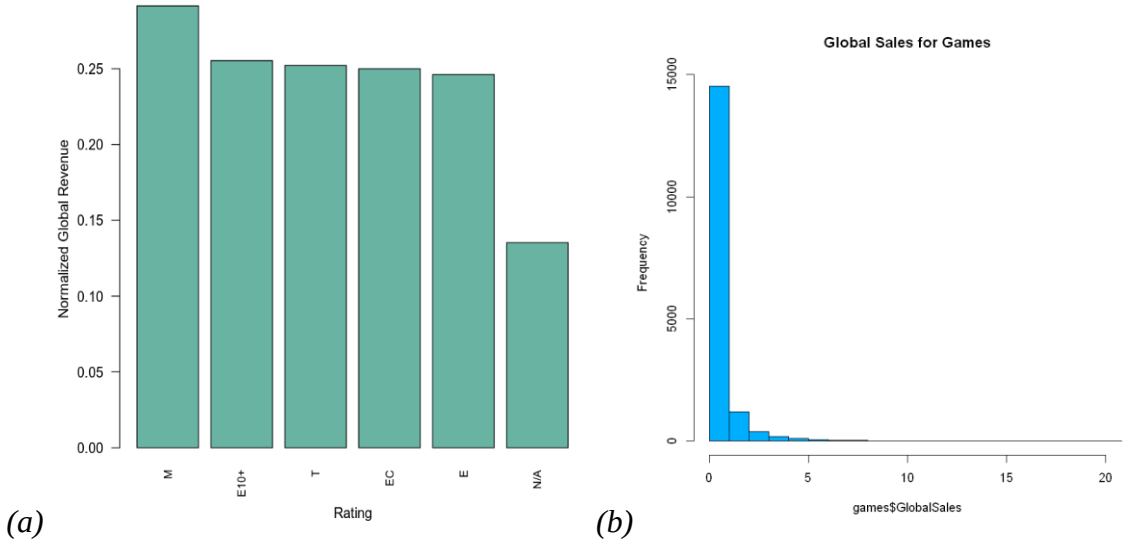


(a) *(b)*

Figure 4 (a) Average Global Sales per Platform (b) Average Global Sales per Genre

To get the results we grouped the games by platforms/genres and then summarised the mean global sales to compare the platforms and genres. From our analysis, we concluded that from our data the best sold games on average were console games, and more specifically newer generation consoles (as this is before PlayStation and Xbox games were released) like PlayStation 4 and Xbox One, followed by the older generation consoles. Interestingly, we found that the main genres contributing to global sales were sports games. The sports games

were closely followed by shooter and platform games. Action games, which we predicted in our hypothesis to be the best selling games, were only the 4th best one, and the role playing games were much worse.



(a) *(b)*

Figure 5 (a)Average Global Sale per Rating (b) Games per Global Sales bin

Finally, we had a look at our final categorical feature, the ratings. The ratings represent the age at which the game can be played. We thought it would be useful to look at, as it usually represents how many people can play the game. On one side, games with higher maturity rating will result in less of the younger people buying the games, but on the other side, games with lower maturity rating will result in older population buying less of them as they might be targeted towards the younger players. The analysis of the mean global sales per rating, has yielded that the highest revenues on average were of the M rating (see *Figure 5(a)*), meaning that the games for the age over 17 are sold better.

3 Models

Model	Columns	Column Transforms	Description
Linear Regression - Demographics	- Regional Sales - Age Distribution - Gender Distribution	- Normalised data - Mean of age and gender distributions per year	Predicting best fit line and finding trends between regional demographics and the regional sales
Linear Regression - All fields	- Global Sales - Ratings - Platforms - Genres - InFranchise - Age Distribution - Gender Distribution	- Normalised data - Rating, Platforms and Genres one hot encoded - InFranchise is a dummy variable (0 or 1) - Sales was transformed with a log function	Predicting the global and regional sales based on all of the numeric data and picking out the most important predictors.
Logistic Regression	- Global Sales - InFranchise	- Normalised data - InFranchise is a dummy variable (0 or 1)	Classifying games in franchises based on the global revenue per game.
Neural Network (NN)	- Global Sales - Genre - inFranchise	- Tried Normalising the Global sales, when that didn't work then we tried with original values. When that also failed, split games into high and low sellers. - One hot encoding of genre	Using neural networks to predict the global sales based on genre and franchise.
K-Means Clustering	- Global Sales - Regional Sales - Platforms - Genres - InFranchise - Ratings	- Normalised global and regional sales. - Rating: ordered encoding. - InFranchise is a dummy variable (0 or 1) - Platforms and Genres one hot encoded	K-Means clustering with an optimal k value based on the silhouette index was run against the associated fields. Clustering was run on subsets split between the one-hot encoded genres from the original dataset. Platforms were also one-hot encoded and used in the summary evaluation of each cluster.
Hierarchical Clustering	- Global Sales - Regional Sales	- Normalised global and regional sales.	Hierarchical clustering with number of clusters based on

	- Platforms - Genres - InFranchise - Ratings	- Rating: ordered encoding. - InFranchise is a dummy variable (0 or 1) - Platforms and Genres one hot encoded	manually analysing dendrograms. Clustering was run on subsets split between the one-hot encoded genres from the original dataset. Platforms were also one-hot encoded and used in the summary evaluation of each cluster.
Random Forest	- Global Sales - Ratings - Platforms - Genres - InFranchise - Age Distribution - Gender Distribution - Names - Developers	- inFranchise is a dummy variable (0 or 1) - Platforms and genres are one hot encoded. - log(GlobalSales) used as well as normal GlobalSales	Using Random Forests to predict global sales and identify importance of different attributes.
Time Series	- Global Sales - Regional Sales - Release Date	- Filtered by Genre or Platform	Use Time Series Sequential Classification to investigate trends linked to sales and the time of year a game is released. Also investigate trending genres and platforms over the years.

Table 1 Summary of Models

3.1 Regression

For our statistical analysis with regression, we were exploring various correlations between different features in our dataset. Given that the majority of our dataset is categorical data, rather than continuous we looked only at some specific relationships between features affecting the global sales, and tried to determine what could affect it.

3.1.1 Relationship of Regional Demographics on Sales

During our data pre-processing we added additional data to the original game's dataset. The motivation behind it was to identify whether the demographics of a region affect the sales of games. In our hypothesis, we expected there to be a pattern. Firstly, over the years the game culture was more catered towards the male population, and so we predicted that the gender ratio will have an impact on how well some of the games are sold. Likewise, we made an assumption that younger populations will have higher sales, as our prediction is that usually video games are played by teenagers and people in their 20s.

Given that our dataset contains regional sales, we wanted to go specifically into regions and compare the correlations between different regions. To achieve that, we split the dataset into the outcome field, which in this case was the regional sales, and the predictor, which is either age or gender demographics. The initial problem was that all the data from the same year had the same demographic value. For that reason, we grouped them together and took the average, i.e. we were looking at how global sales are affected by the average demographic for a year. We used the points to generate a scatter plot and then calculated the best-fit line. We had to make small adjustments in the pre-processing steps in order to remove a higher percentage of outliers, as they skewed the best-fit line extensively, due to highly abnormal sales. In order to evaluate the linear regression, we split the dataset into a train and test set using a holdout method (with a 70:30 split). Using the test set, we evaluated the model performance by calculating mean absolute error (MAE), root means squared error (RMSE) and R^2. With the first two, we will be able to assess how different the predicted values are from the actual values in our dataset. As per [10] R^2 allows us to measure statistically how far away the scattered points are from the best fit line, and verify if the line explains the variability of the data around its mean With the two metrics, we will be able to evaluate the performance of the model, verify the accuracy of the detected correlations and compare the models.

3.1.2 What affects the sales?

The second technique we looked at with regression was training a linear model on all of the numerical features in our dataset. The reason behind it was that we can go over various variables and evaluate their strength on the sales based on their coefficients. This will indicate the variables' importance on the output. We focused on predicting the global sales, as well as the regional sales, as we expect there might be some differences between them. Our data largely consists of categorical data, such as Ratings, Genres and Platforms, which are largely important for our analysis. Therefore, as mentioned previously, they were one-hot encoded so that they can be used in the linear model. We also took it a step further, and we experimented with some transformation functions, such as logarithmic, exponential and polynomial (with various degrees). Just as in the above-mentioned method, we used R^2, RMSE and MAE to evaluate and compare the models.

3.1.3 What do franchises tell us about sales?

Another regression analysis which was observed in this work is the logistic classifier. As per the business hypothesis, we expected that the games which are in the franchise will have greater sales. Since the characteristic of franchises in our dataset is a dummy variable, it takes the values 0 or 1 to indicate if the game is in a franchise or not (see 2.2 Data Cleaning for more), logistic classifier seemed like the right choice. The intention behind this statistical method is to find the model to describe the correlation between global revenue and games in franchises. In the data processing part we are using only two columns: 'inFranchise' and 'GlobalSales', after calculating the threshold, and the mean of the games in the franchises, we were able to create a model. Predicting a logistic classification model allowed us to create a confusion matrix with relevant statistical values. A confusion matrix is a method used to describe the performance of a classification algorithm by checking TP, TN, FN, and TP, calculating accuracy, false positive rate and false negative rate. These values will provide us with some data on how accurate and reliable our model is. Furthermore, plotting a Receiver Operator Characteristic (ROC) curve shows the diagnostic ability of our binary-categorical classifier. Having a ROC curve shows us the trade-off between sensitivity (or TPR) and specificity (FPR), also it allows us to measure the distance between the points and the ideal clinical discriminator for choosing the ideal points.

3.2 Neural Network (NN)

In the project plan we outlined that we would train a Neural Network on the dataset, using past data to predict the performance of a new game based on inputs like genre, platform and region demographics. The aim of this was to evaluate how well a new format of the game would perform in a region based on what genre it is and what platform it gets released on, allowing for a greater development focus in that area. Furthermore, it would help us uncover trends/relationships in the data that we could not find in the previous models.

We made the decision to use neural networks due to the increased probability of reflecting changes in the data that other models may generalise. We attempted multiple different approaches to implement the neural network. Our first attempt was to use Regression. First, we normalised Global Sales, so all values were between zero and one. We then used Root Mean Squared Error (RMSE) as our evaluation metric. Our next attempt was to use Classification, we started by binning the Global Sales into different percentiles. Then each of these bins

became a class in our classification problem. We also tried splitting the data into high and low-selling titles to more closely group the data and allow the model to more easily detect patterns.

3.3 Random Forest

After realising the underwhelming performance of our neural network models, we decided to experiment with Random Forests to see if they would be more fit for purpose. The aim remained much the same as with neural networks; creating a model to predict sales values based on a given input in order to lead development focus on a specific area. However, Random Forests have the added benefit of being comparatively simpler, leading them to better generalising the data trends.

We evaluated two different sets of input values, one of all values, excluding those related to specific regions (unfortunately this invalidated the joining of the other datasets), and one with all values, excluding region specific data as well as user scores, critic scores, user count and critic count. By comparing these approaches, we hoped to gather insight into whether user reviews would be necessary for our planned estimation of sales data. We used RMSE and node importance as our evaluation metrics for this section.

3.4 Time Series Sequential Classification (TSC)

We also investigated time series sequential classification. We used this to investigate the trends linked to global sales and the time of year a game was released. This would allow Nantendo to optimise their sales by choosing when in the year to release their game. We used Release Date, a field we generated with data scraping, to create a time series graph and decompose this to identify general trends and seasonal trends.

We used the non-pre-processed data as we didn't want our sales fields normalised or our filter fields (genre or platform) to be one-hot encoded. However, we removed NA rows and reduced it to global sales less than 0.5 to remove outliers. We removed all fields apart from the release date and sales so that the time series could focus on plotting trends of sales over time. We filtered this by either Platform or Genre depending on which trend we wanted to observe.

The first trend we looked at was the platforms over time, the aim of this was to allow us to see which platforms would give Nantendo better sales. Whilst using domain knowledge, as shown in *Figure 4(a)*, we know that older platforms like arcade games have slowly generated

less revenue over the years and newer generations of consoles become more popular over time; It would be useful to find which consoles in the same generation are trending in which regions.

We also used time series to find trending genres over the years as in our hypothesis we stated that the main factor that will affect game sales will be the trending genres in different regions. Due to the dataset being very heavily dependent on the time/release date of games, this model is very suitable for the problem. This would allow Nantendo to predict which genre of game was becoming more popular and which genres were dying per region in recent years. As per our hypothesis we believe that games like action games will be more popular in America, whereas RPGs will be more popular in Japan. In order to get the data for this, we scraped release dates from the website Metacritic.

We then use a deep neural network to look at time series and predict in future what sales a certain genre or platform will receive. Like linear regression we split the dataset into a train and test set using a 70:30 split. Furthermore we evaluated the neural networks performance by calculating the root mean squared error and R^2.

3.5 Clustering

3.5.1 Additional Pre-processing

We had already performed basic pre-processing on the games dataset like unifying NA values, removing outliers and normalising numeric values. When we began working with clustering we quickly realised this was not enough as we were not seeing interpretable clusters. So using the cyclical nature of CRISP-DM we went back and re-analysed the data. In order to improve our data we ended up doing the following things:

We encoded the rating in order so that we would have additional numeric values to include as predictors for clustering.

We dropped the UserScore, UserCount, CriticScore and CriticCount columns. We can only perform clustering on data that does not contain any NA values, as a result each row that contained NA's was dropped. The 4 columns mentioned previously have the largest amount of NA values (even after scraping). For example the CriticScore column has 8582. This meant that when we kept only complete cases of rows a lot of records were removed simply due to the lack of data in these columns.

We didn't include any of the demographics data or Year of Release in the predictors either. When we included these columns our clusters became extremely distinct because so much discrete data that is also strongly correlated caused the clustering algorithms to focus on them.

When giving the clustering models their predictors we filtered the data by genre. We came to this approach after a number of experiments with various predictors and data. Splitting the data by genre gave us the most meaningful clusters to analyse both for Kmeans and Hierarchical clustering. We believe one of the reasons for this is because we do not have a very large number of predictors that we used as the majority of the data is categorical and clustering works only on numeric data.

3.5.2 K-Means Clustering

As stated in our introduction, our decision to use k-means clustering stemmed from the fact that our dataset had a very high dimensionality. Therefore, k-means clustering was used as an exploratory analysis into the dataset to see which features impacted sales the most and what possible unseen trends could be found between fields. K-means clustering provided the added benefit of being a form of unsupervised learning so there wasn't bias towards a required output field that needed to be focused on. We identified some possible trends in the data that allowed us to recognise high-selling genres and the fields that gave the best relationships to global sales per cluster.

The initial idea of clustering was to use k-means clustering as it is a popular clustering method to work with. We focused clustering on all of the sales fields along with ratings and the field specifying if a game was franchised. The data was then split into subsets based on the one-hot encoded genre fields. We initially intended to run clustering on the columns including regional demographic data but due to how the demographic dataset was joined with the original games dataset, all of it became discrete and unusable. Therefore the dimensionality of the data was reduced by quite a large amount. This was a deviation from the project plan as we had hoped to involve more fields from the dataset during clustering. To try and get more meaningful insight from the dataset, we decided to look at the fields associated with high-selling clusters in genres. Their proportions were then analysed with respect to the best and worst performing clusters per genre. This provided some useful evaluations that will be discussed further in the evaluation of this model.

To evaluate the optimal number of clusters for each genre subset, we ran the silhouette index method against them and plotted cluster graphs with their respective k values per genre. We then plotted the clusters onto graphs to get a visual representation of the degree to which the clusters were similar to each other in each graph. Afterward, we printed a series of values describing all of the constituent parts of the clusters. These values can be seen in the tables for k-means clustering in the results section of the document.

3.5.3 Hierarchical Clustering

Aside from K-means clustering which is one of the most popular approaches we also tried hierarchical clustering. The benefit of hierarchical clustering is that we do not need a predefined number of clusters. On top of that we hoped that a second clustering approach would solidify or expand upon our findings from K-means clustering. We used dendrograms to decide where to cut the clusters based on the maximum height between nodes when placing a horizontal line. This is because the height distance between nodes is an indication of the similarity of clusters. We analysed the dendrograms and selected the optimal cluster number for each genre by performing clustering on a subset of the data that only contains records of a specific genre.

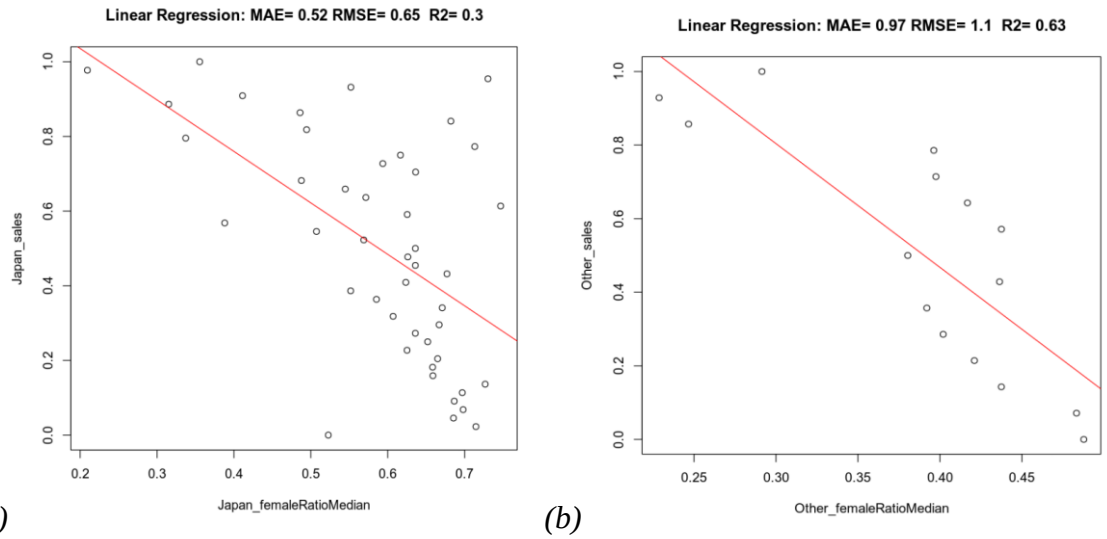
Our evaluation method for hierarchical clustering consisted of two parts. We decided not to create cluster plots but instead create scatter plots so that we can hopefully find new relationships in the data. For each set of clusters (per genre) we plotted them on a scatter plot with colours corresponding to each individual cluster. Determining the x and y axis of this scatter plot required some experimentation. We initially tried mapping Ratings against Global Sales but this had little value. This is because even though we encoded ratings in order they are still discrete, so it was hard to visualise significant relationships. We then tried mapping various regional sales data against each other and this showed far better plots with some interesting clusters to analyse. We also made sure to denormalize all of the values before plotting them. Once we had the scatter plots we profiled the clusters by summarising some of the most important details such as percentage of games in franchises or the percentage of games in platforms. The majority of these summaries were in percentages as all clusters contain a different amount of records and we wanted a way of comparing different clusters.

4 Results

4.1 Regression

4.1.1 Linear Regression: Demographics

Comparing the regional sales to the regional demographics yielded mixed results, as every region produced varied results from each other. However, some of them do follow a similar trend, albeit with a weak correlation. Japan, Europe and Other regions (all except Northern America), follow a similar pattern with regards to the gender ratio and the age median. In *Figures 6 (a) and (b)*, both of the regression lines fit the data in a similar pattern; when the population has a lower female percentage, the video games sell better on average.

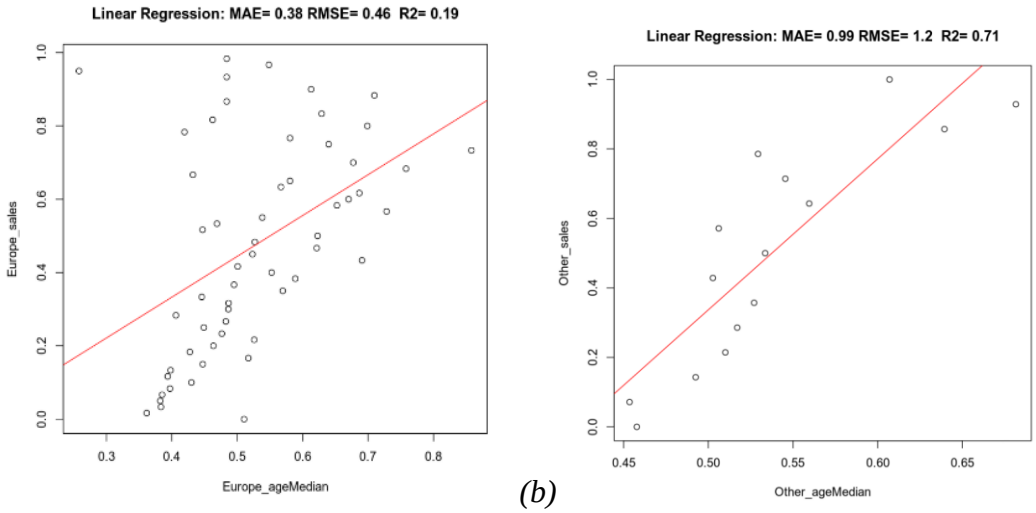


(a) *(b)*

Figure 6 (a) Average Japan gender distribution, affecting Japanese Sales (b) Average gender distribution in 'other' regions affecting the sales

The main difference between the two graphs is the variety of data. Other dataset contains less data points, and therefore the model is easier to predict. The range of data is also smaller and thus makes it difficult to compare directly to the other regions. So even though R2 shows statistical significance, The MAE and RMSE show that it does not perform well predicting test data points. On the other hand, Japan has a weaker correlation, as depicted by R2, but the model performs much better on the test dataset, making smaller errors. In general, from the gender distribution, it can be concluded that there is some correlation, even if it is not very strong, between the gender distribution in the population and the sales of video games for that region. We tried fitting the data with polynomial functions, but there was no improvement in performance, most likely due to the spread of the data.

Similarly, the analysis into the regional age distribution yielded similar results. As shown in *Figures 7 (a) and (b)*, the overall trend is that on average, games sold in regions with higher age median have performed better. Once again, using polynomial functions has increased R^2, but it sacrificed the performance of the model. Most likely, using the polynomial curves overfits the data as there are not that many data points and some of them might be outliers.



(a) *(b)*

Figure 7 (a) Average European Age Distribution, affecting European Sales (b) Average age distribution in 'Other' regions affecting the sales

Finally, one region which performed badly was Northern America (see *Figure 8*). The data around the mean is largely spread out and performing regression on it doesn't predict much. Using a parabola would not make accurate predictions as the data around the mean would be largely varied.. No matter which function was used, the data doesn't show much, and that can be proven by an extremely low R^2 value, showing there is no correlation between the predicted regression line. Therefore, the conclusion can be drawn that the age median doesn't affect the North America sales.

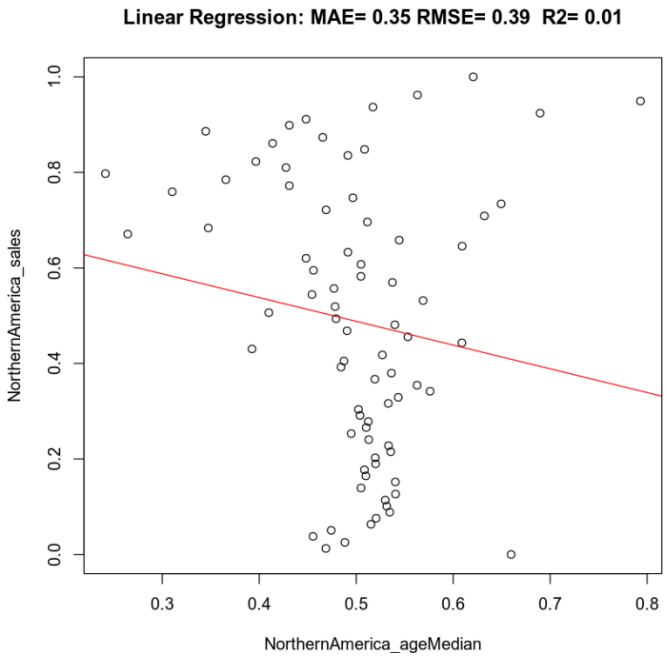


Figure 8 Average Age Distribution in North America against North America Sales

4.1.2 Linear Regression: Importances

Training the linear model on all of the available numerical data had a subpar performance. Table 2, summarises the results and the steps taken to evaluate the linear model with R^2. Initially, we were looking solely at global sales and we used all the data we had available. The performance of the model was extremely low, and there didn't seem to be any correlation between the features when combined together, at least no linear correlation. We tried, removing features that were the least important (based on the coefficients from the model), usually cutting the last couple but we found that it didn't yield any better results and sometimes it would even decrease the correlation.

Therefore, we started experimenting with transformation functions. More specifically we looked at polynomial and logarithmic functions. We went through a range of degrees in the polynomial curve, but it did not perform any better than the initial model. We haven't included it in Table 2, but we tried transforming the continuous inputs with the same transformation functions, but those did not affect the output either. However, using a logarithmic transformation on the output (sales) has had a massive improvement in the correlation of the linear model. As it can be seen in Table 2, we run it on both the Global sales, but also regional

sales. Even though the correlations and performance are still weak, throughout, they are not non-existent anymore.

	Linear + All data	*Linear + Some data*	*Polynomial (degree = 3) + All data*	*Logarithmic + Some data*	**Logarithmic + All data**
Global	0.06	0.05 - 0.06	0.05	0.22-0.23	**0.23**
Europe				0.26-0.27	**0.28**
North America				0.40-0.43	**0.44**
Japan				0.35-0.37	**0.37**
Other				0.33	**0.33**

Table 2 Linear Regression Training R^2 results

With that, we were able to look at what the strongest contributors to the output were based on the coefficients from the model. Even though the R^2 and importance value vary significantly throughout the regions, there is a major similarity between all of them. In all of the cases, ratings had the highest importance on sales (see *Figure 9* and *Table 2*).

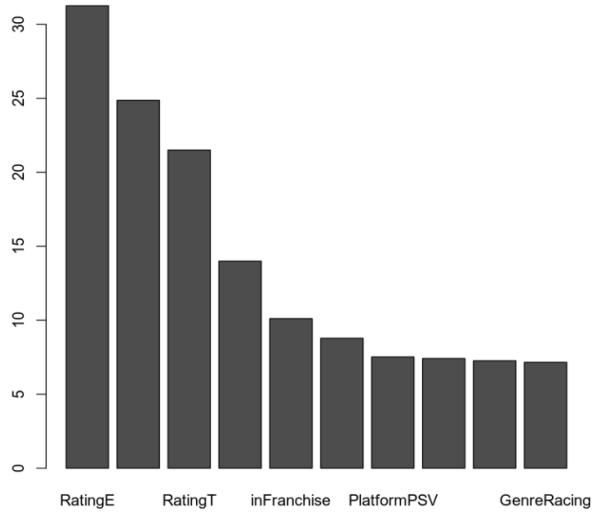


Figure 9 Strengths of Linear Model for Japan

As *Table 3* suggests, from this result, ratings affect sales the most in every region. Unfortunately, due to the data being weakly correlated and the statistical importance being low

in the linear models, other methods should be used to identify how different features affect global sales. It seems like the linear model is not enough to predict those correlations.

Feature	Overall Importance
RatingE	31.3
RatingE10	24.9
RatingT	21.5
RatingM	14.0
inFranchise	10.1
GenreRolePlaying	8.8
PlatformPSV	7.5

Table 3 Overall Importance of Features in the Linear Model for Japan

4.1.3 Logistic regression

The results of the correlational analysis are set out in *Figure 10.* It can be seen that the data is spread across the apsis for both y (probability of being in franchise) being 0 and 1, hence no correlation can be observed directly from the graph without using a machine learning method. Simply, we calculated a 'threshold' for the value of 'inFranchise', over which the class is inFranchise and under which it is not, {0,1} (See Equation 1).

$$p(inFranchise) = \{ \text{ 1 when > 0.36; 0 when y < 0.36 } \}$$

Equation 1 probability of in Franchises

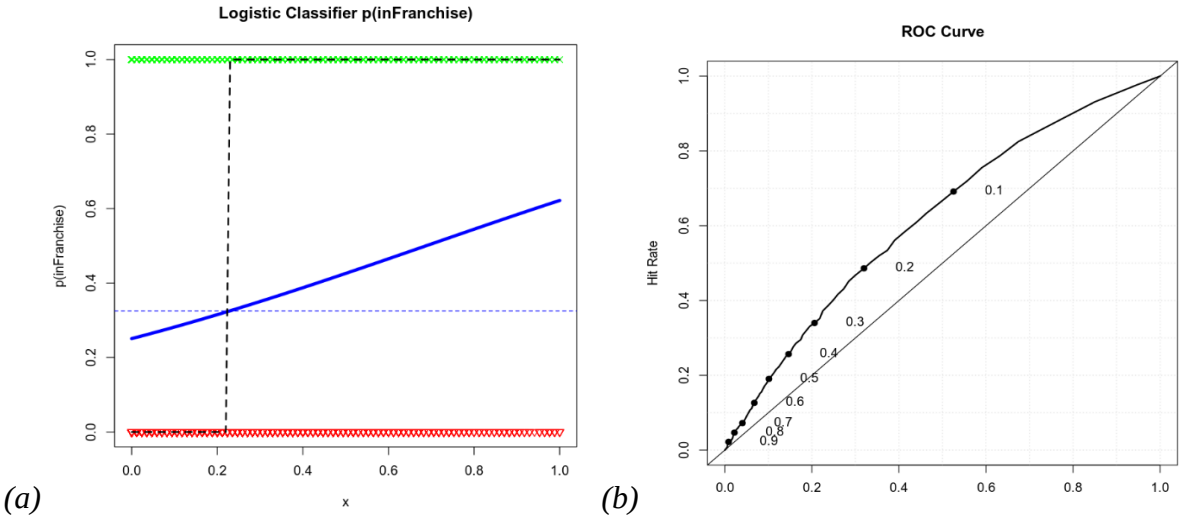


Figure 10 (a) Logistic Classifier, probability of in Franchisese over GlobalSales(x) (b) ROC curve

Furthermore, a confusion matrix was defined in table 4. The model produced the accuracy of 66.87%, true positive rate (TPR) 85.13% and false positive rate(FPR) of 15%.

	False	**True**
False (not in franchise)	4602	519
True (in franchise)	1993	469

Table 4 Confusion Matrix

The data presented in *Figure 10 (a) and (b)*, tells us that we are able to predict if the game is in franchises based on its sales with an accuracy of 67%. It also tells us that although there is a small correlation between those two parameters (global sales and in franchise), we still cannot base our business decision on that characteristic. However, the reason this is a good thing is because it reveals the potential in sales of the games which are not owned by a studio, thus increasing our chances on the market.

The results obtained with this method don't have high statistical significance and the correlations are fairly weak, so they can be used to drive the rest of the analysis and make some shallow assumptions about the data correlations, but not much more than that. Large portion of our dataset is not continuous, and hence, we decided to not dive deeper into the regression models and focused on investigating other techniques as it proved not to be the best model for

our dataset. With the other methods, we can explore the effects of genres, ratings and platforms more.

4.2 Neural Networks

At first, we approached the task with a regression model. Our first model produced a RMSE of 2.7% and failed to identify even any slight relationships in the data. In response to this, and the patterns seen in the results, we decided to change our approach to a categorisation model.

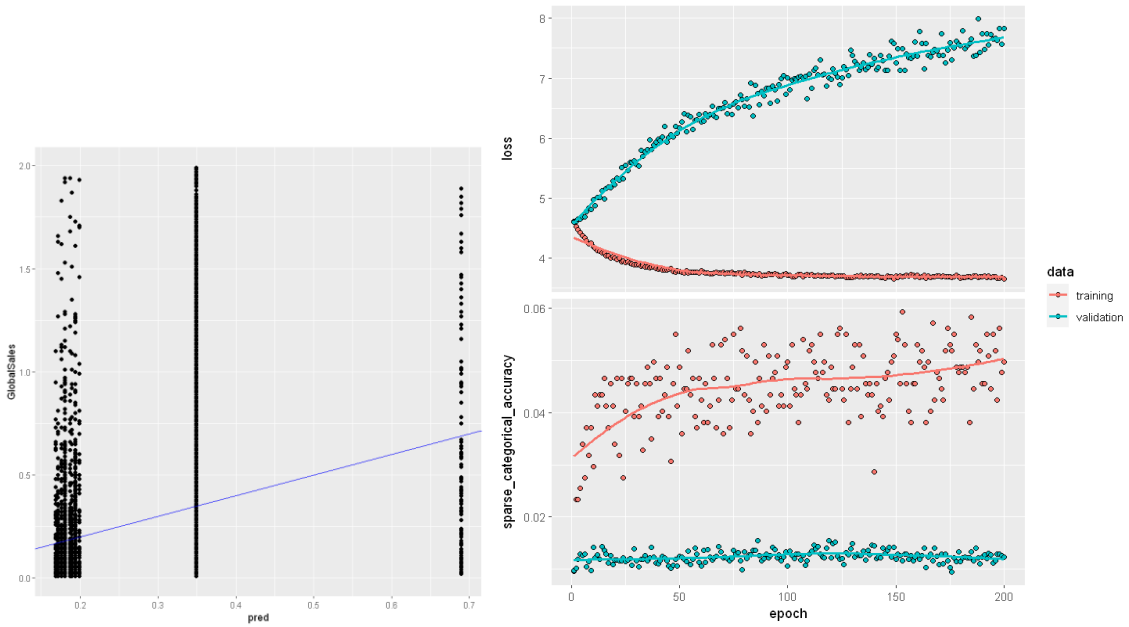


Figure 11 (a)GlobalSales over Predictions (b) NN accuracy and loss over epochs

The categorisation model performed worse in our initial runs, where it achieved scores between 1 and 1.5%. With a bit of tuning, we managed to achieve a result of 6.1% which gave us a small degree of hope that we could get a usable model, however that was the apparent limit of the models we tested; it did not appear feasible to predict more accurately using neural networks in their current implementations. It can be seen in *Figure 11(b)* that despite the training loss decreasing, and the accuracy increasing, the opposite is true for the validation data.

We theorised that the models were struggling to accurately predict values due to the large discrepancy in the volume of games around certain sales points: around half of the games in the dataset have global sales values that are between 0 and 1 million sales, whereas there are only around 850 that have global sales values that are above 2 million (See *Figure 5(b)*). When

testing training the model with the dataset split into low and high selling games, we saw a massive reduction in model performance when training with the high valued games.

The Neural Network models we attempted had very low accuracy. There are a few reasons that we believe led to this result: there was a relatively low correlation between the immediately available fields and the amount of global sales, making it much harder to predict sales with any accuracy; a lot of game sales are driven by currently popular genres and titles, however there was no real way of representing this change in popular interest in the dataset due to the sales being lifetime sales and not broken down per year. It is due to this time-based variance that made time series more effective as, while limited, it was able to see some trends based on the date's games were published. Ultimately, we decided to move onto random forests to carry out the task.

4.3 Random Forests

There are two different performance indicators that arose from Random Forests. The first is the ability to predict the sales performance of a game, and the second is the importance of the nodes in the tree. By predicting sales for a game, we are able to obtain a rough estimate of how well a game will perform and thus whether it will be beneficial to work on. In addition to this, the importance of the nodes allows us additional insight into what categories are important in determining the sales performance of a game.

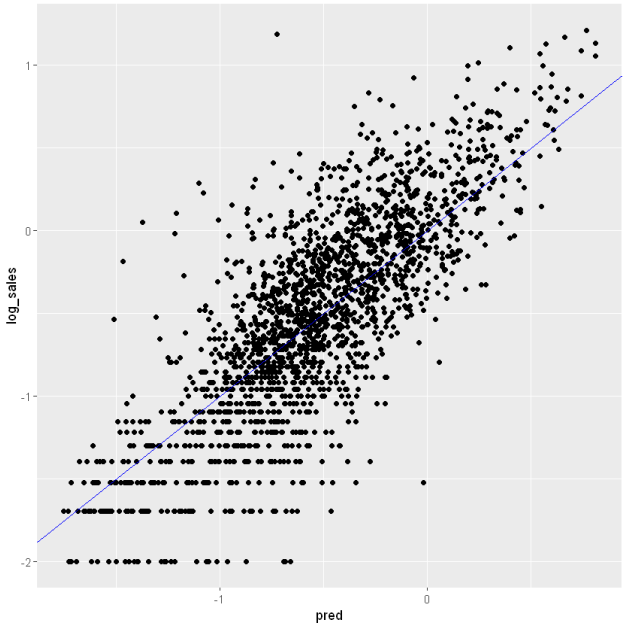


Figure 12 Random Forest model for sales prediction

Rather intuitively, it can be seen that the critic score, user score, and the number of respective ratings are very important in relation to sales. Despite these being values that we have no control over, we can still influence their direction by gauging public opinion to advertisement, beta testing and game announcements. The most instantly relevant important values are those that we can control before or during game development; after the user and critic scores, it can be seen that the game being featured on PC, being a part of a franchise and the name of the game are all of high importance. A game being featured on PC is ranked the 2nd highest out of all important values, therefore we can conclude that it would be vital to release a game on this platform alongside games consoles.

In addition to the predictive capabilities of the Random Forest, we were able to analyse the tree to see what categories gave a good global sales value (Table 5)

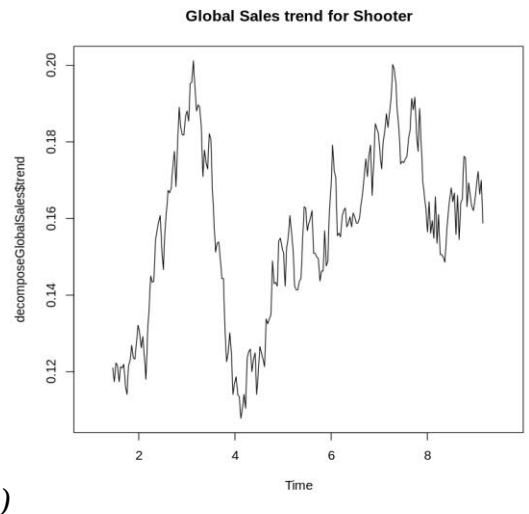
Condition	Prediction (GlobalSales)
Other_femaleRatioMedian <= 50.18 & GenreShooter <= 0.5 & GenreStrategy > 0.5 & PlatformPS2 > 0.5 & RatingM <= 0.5 & inFranchise > 0.5	4.68
NorthernAmerican_ageMedian > 37.75 & Other_ageMedian > 20.85 & GenreStrategy > 0.5 &PlatformPS2 > 0.5 & PlatformPS3 <= 0.5 & inFranchise > 0.5	4.68
Other_femaleRatioMedian <= 50.1832752227783 & Other_femaleRatioMedian <= 50.137134552002 & GenreMisc > 0.5 & PlatformPS2 > 0.5 & RatingM <= 0.5 & inFranchise > 0.5	2.4675

Table 5 Random Forest optimal paths to the best sales

4.4 Time Series

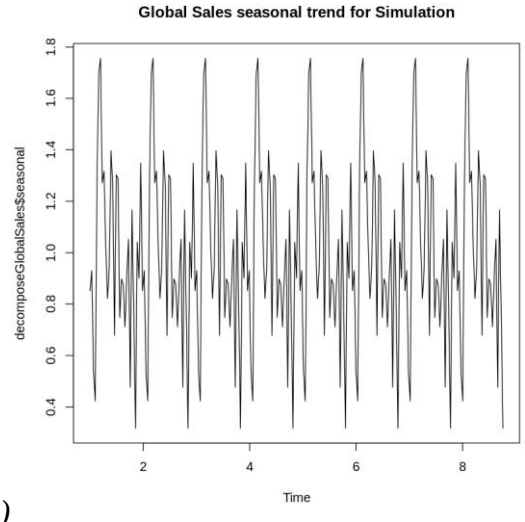
Looking at the trends of platforms we didn't find any useful trends, we discuss further why we think platforms were not a useful field to analyse in the Discussion Section.

However, using the genre showed us interesting trends in the time series graphs. As seen in both *Figures 13 (a) and (b)* the trend in global sales over the years for both the genre "Shooter" the genre "Strategy" are very similar, they both overall have an increase in Global Sales over time.

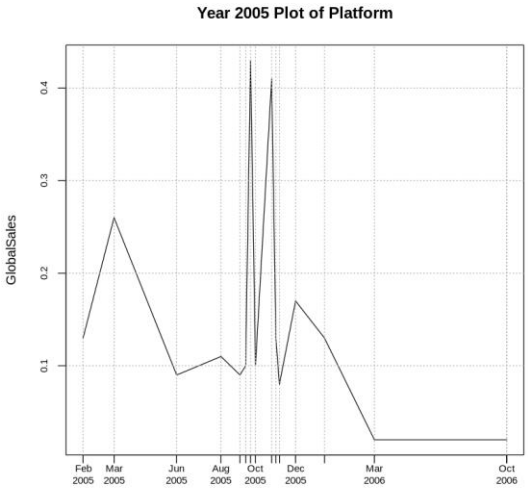


(a)

(b)



(c)



(d)

Figure 13 Time Series Graphs for Platforms

However, in general the global sales for all games genres decreased. This could mean two things about the dataset, either the dataset is unbalanced and has more games released in earlier years or that the newer games have had less time to gain sales.

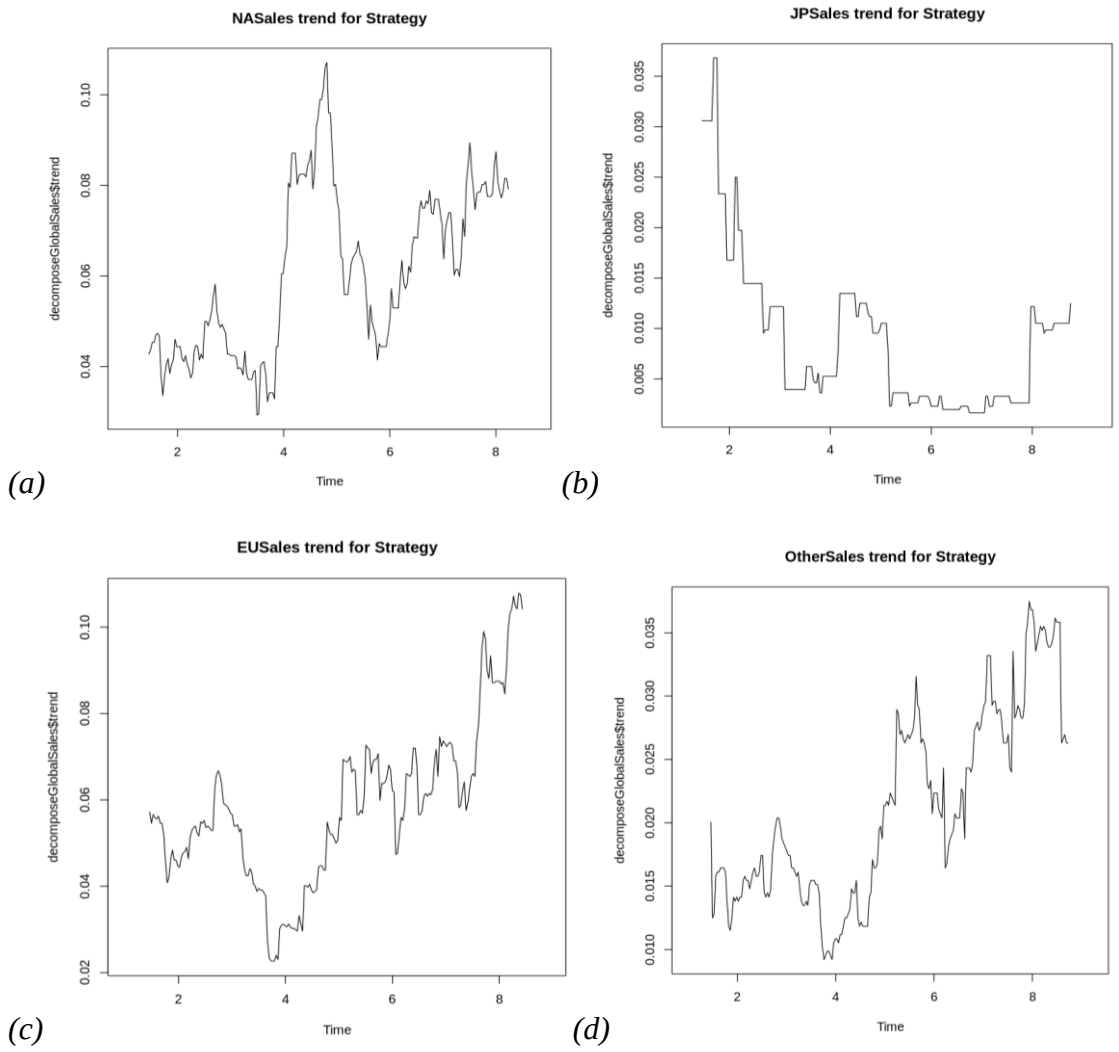


Figure 14 Time Series Regional Graphs

Looking at sales per region, we can see an interesting trend. In *Figure 14* we can see that for the genre "Strategy" all regions apart from Japan have an increase in Global Sales over time. Japan has a significant decrease in sales and the scales of the sales axis is significantly less as well.

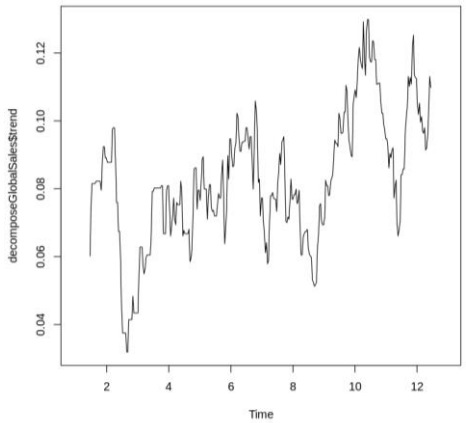


Figure 15 Time Series Japan Sales trend for Role-Playing platforms

Figure 15 shows that the best genre for the Japan Region is Role-Playing, not only does it have a positive trend over time, the scale of the axis is a lot bigger. This conforms to our hypothesis where we stated that action games would be more popular in North America and RPGs being more popular in Japan.

When looking at the seasonal trends, an interesting observation we made is that there is a clear peak in global sales towards the end of each year, an example of this is shown *in Figure 13(c)* for the genre Simulation. Plotting out the year in a separate graph reveals that they peak around the months of October / November. This means that in the dataset the highest-selling games are sold around this time of year.

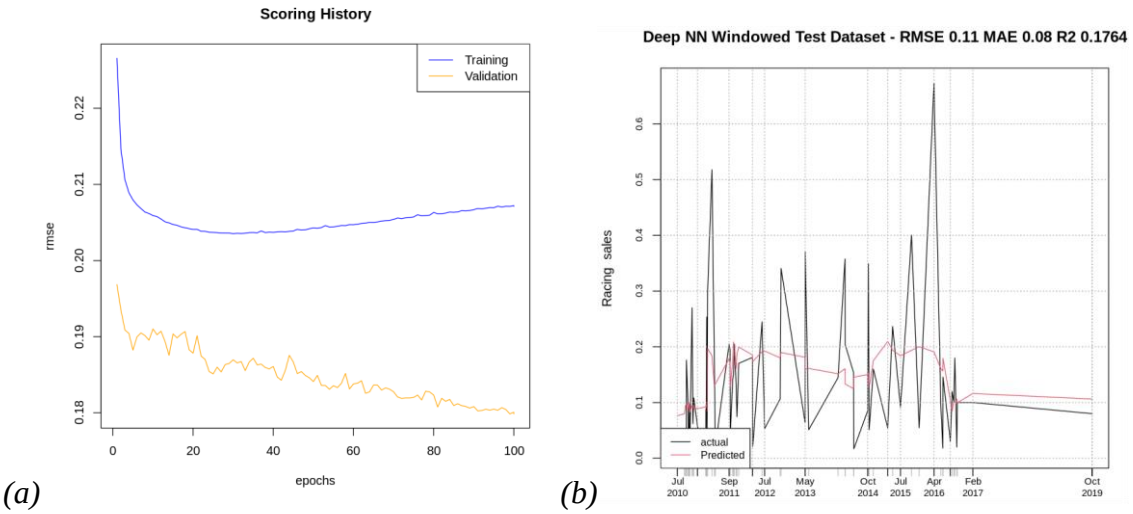


Figure 16 Time Series- Neural Network (a) Scoring History (b) Actual vs Predicted

The deep neural network for predicting the time series model achieved an RMSE of 0.11 and an R squared value of 0.1764 (as shown in *Figure 16*). Although these values are low (see *Figure 16(a), (b)*). shows that towards the final epochs the model learnt to generalise more as the training RMSE increased and the validation RMSE decreased.

4.5 Clustering

4.5.1 Hierarchical Clustering: Dendrograms

Before we were able to form the dataset into clusters, the optimal number of clusters needed to be identified. This was done using the silhouette index method which we ran against each genre subset. See *Figure 17(a)* as an example graph of the silhouette method for games under the strategy genre and *Figure 17(b)* for the associated cluster graph.

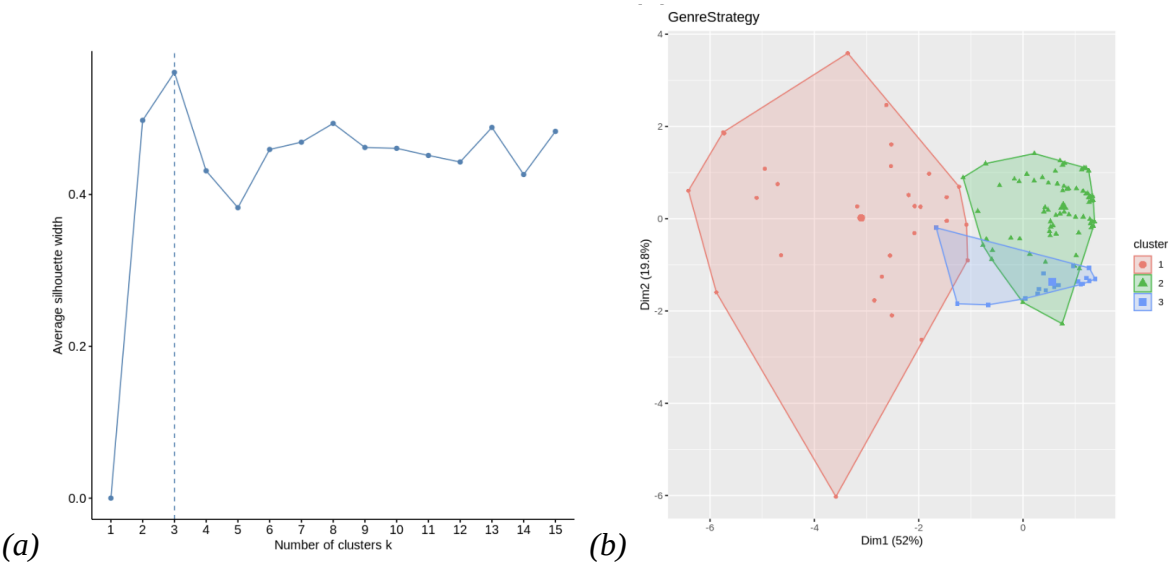


Figure 17 (a) Silhouette Index Plot for Strategy (b) Cluster graph for Strategy Genre

Figure 17(a) depicts the plot of the silhouette index for strategy games. The number of clusters corresponding with the highest silhouette index for sports was 3 clusters. All of the values of k per genre ranged between 2 and 5 clusters.

4.5.2 Hierarchical Clustering: Scatter Plots

After looking at all of the constituent parts of the best and worst performing clusters in every genre, strategy and role playing games seemed to perform the best with respect to values most heavily relating to global sales. Cluster plots of these two genres can be seen *in Figure 18 (a) and (b).*

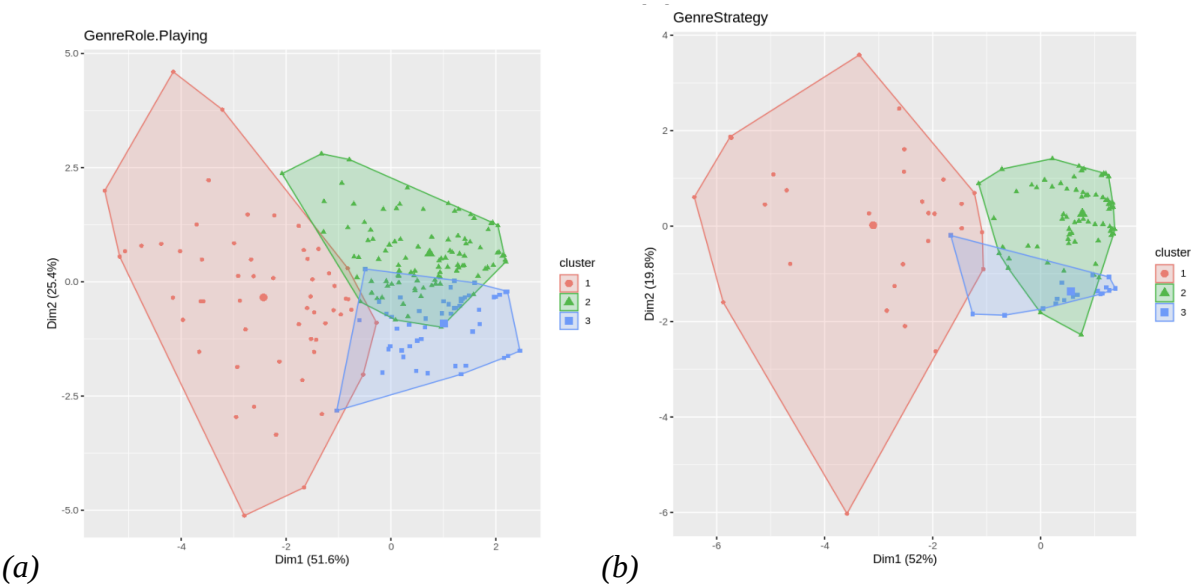


Figure 18 (a) Cluster Graphs for Role Playing games (b) Cluster Graphs for Genre Strategy

In both cases of the Role playing and Strategy genres (See *Figure 18)*, cluster number 1(red) was the best performing cluster and cluster number 2 (green) was the worst performing cluster. According to the analysis of the best and worst clusters of the strategy genre, strategy games had the highest proportion of mean global sales of all of the best performing clusters for every genre at a normalised value of 0.981. The strategy genre also had the highest proportion of a rating that was present among all of the best clusters in every genre, which was a rating of T that constituted 37.9% of all of the high-selling games in the cluster. There was also no substantial visible difference in the frequency of related platforms of high-selling games in the best and worst clusters, leading us to believe that the platform field did not have a large impact on the success of the strategy genre. Specific data values from the profile summary for the strategy can be found in *Table 6* below.

Cluster Profile Fields	**Strategy Best Cluster**	**Strategy Worst Cluster**
Total number of games in cluster	29	102
Average Global Sales	0.981	0.894
Percentage of games with high global sales	100 %	38.2%
Percentage of games with high North American sales	13.8%	1.96%
Percentage of games with high Japanese sales	3.45%	0.98%
Percentage of games with high EU sales	3.45%	0.98%
Percentage of games with high sales in other regions	10.3%	3.92%
Games in franchises with high sales	17.2%	13.7%
Total games in franchises	17.2%	36.2%

Frequency of platforms with associated high-selling games	PlatformDS: 5 PlatformPC: 2 PlatformPS2: 2 PlatformPS3: 8 PlatformPSP: 2 PlatformWii: 6 PlatformX360: 4	Platform3DS: 2 PlatformDS: 4 PlatformPC: 1 PlatformPS2: 2 PlatformPS3: 9 PlatformPSP: 4 PlatformPSV: 1 PlatformWii: 7 PlatformX360: 9
E Rated high-selling games	31.03%	34.3%
EC Rated high-selling games	0%	0%
E10+ Rated high-selling games	20.6%	19.6%
T Rated high-selling games	37.9%	30.4%
M Rated high-selling games	10.3%	15.7%

Table 6 Profile Summary for Best and Worst Strategy Clusters

According to the analysis of the best and worst clusters of the Role playing genre, Role playing games featured the best cluster relative to the highest proportion of high-selling games at 100% and the highest number of games in the cluster compared to other clusters from different genres at percentages of 100% as well. Games rated T and E for everyone also were shown to be quite high in proportion in the best performing cluster at percentages of 33.4% for both T and E. Consoles were demonstrated to be quite impactful in role playing game sales as well with specifically PlayStation and Xbox games having the highest frequency in the high-selling games bracket. More information on the profile summary of the role playing clusters can be seen in *Table 7* below.

Cluster Profile Fields	Strategy Best Cluster	Strategy Worst Cluster
Total number of games in cluster	57	102
Average Global Sales	0.962	0.894
Percentage of games with high global sales	100%	38.2%
Percentage of games with high North American sales	7.02%	1.96%
Percentage of games with high Japanese sales	3.51%	0.98%
Percentage of games with high EU sales	1.75%	0.98%
Percentage of games with high sales in other regions	7.02%	3.92%
Games in franchises with high sales	28.1%	13.7%
Total games in franchises	28.1%	36.2%
Frequency of platforms with associated high-selling games	Platform3DS: 1 PlatformDS: 11 PlatformPC: 2 PlatformPS2: 3 PlatformPS3: 14 PlatformPSP: 3 PlatformPSV: 1 PlatformWii: 11 PlatformX360: 11	Platform3DS: 2 PlatformDS: 4 PlatformPC: 1 PlatformPS2: 2 PlatformPS3: 9 PlatformPSP: 4 PlatformPSV: 1 PlatformWii: 7 PlatformX360: 9
E Rated high-selling games	33.4%	36.6%
EC Rated high-selling games	0%	0%

E10+ Rated high-selling games	19.3%	19.6%
T Rated high-selling games	33.4%	32.1%
M Rated high-selling games	14.04%	11.6%

Table 7 Profile Summary for Best and Worst Role Playing Clusters

In general, the clusters showed that games which were rated E have high average sales consistently throughout all game genres. Another observation was that on the global scale, games being in a franchise had little to no effect on the performance on the clusters in all genres. This can be seen when comparing the best and worst clusters in *Table 6* and *Table 7*, where despite all of the games in franchises being high-selling games in the best clusters, a higher proportion of games in the worst cluster were in franchises but the franchised games that had high sales were a fraction of the total franchised games.

4.5.3 Hierarchical Clustering: Dendrograms

Before being able to profile the clusters generated by hierarchical clustering we wanted to choose the optimal number of clusters. Therefore we wrote code to create a dendrogram for each subset of the genre. When making the initial dendrograms we passed in the amount of optimal clusters determined by the Silhouette method which is indicated by the different colours. This was out of interest to see how these two methods compare.

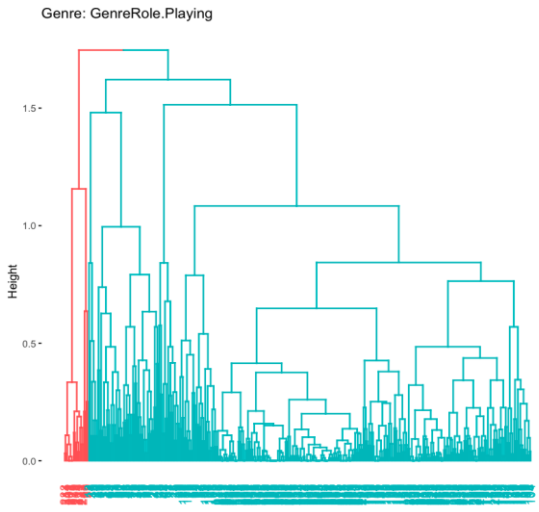


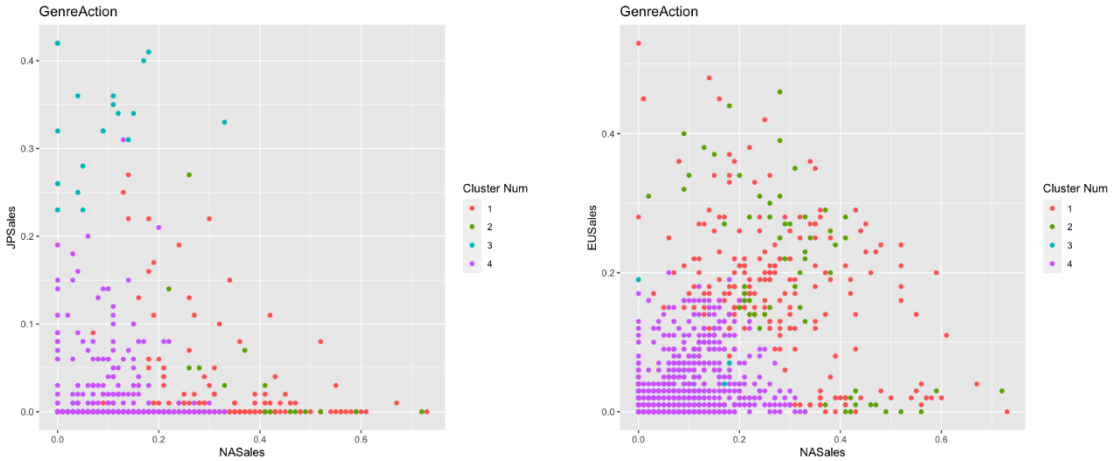
Figure 19 Dendrogram for Role Playing Genre

Figure 19 shows a dendrogram created on a subset of the dataset for the genre Role Playing.

It indicates that the optimal number of clusters selected by the silhouette method may not be the most appropriate for this set of data. We can see the imbalance of records in each cluster, the red one having significantly less records in it. As a result it may be difficult to profile these clusters appropriately and find distinct and meaningful relationships. Based on *Figure 19* a more appropriate number of clusters would be 5. This is because if we placed a horizontal line between height 1.0 and 1.5 we would be intersecting 5 leaves with some of the largest distances to the root nodes. This indicates that those will be 5 fairly distinct clusters since the horizontal distance is the similarity between them. Something to note is we decided where we place this line based on the visualisation rather than numeric values for the sake of simplicity.

4.5.4 Hierarchical Clustering: Scatter Plots

Since we developed a lot of plots for all the different genres, as well as various combinations of predictors to map on *Figure 20 (a), (b)* axis, this section will analyse the most interesting plots.



(a) *(b)*

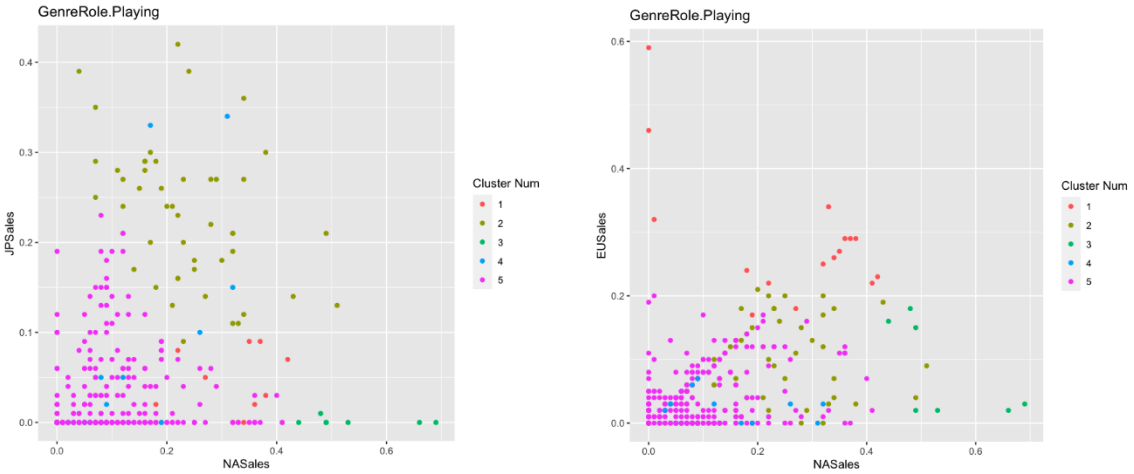
Figure 20 (a) Scatter plot of clusters for regional Japan and North America sales for action genre (b) Scatter plot of clusters for regional EU sales and North America sales for action genre

We analysed the action genres as it had some of the most distinct clusters. The most prominent cluster was distinguished when comparing Japan sales with North America sales. As can be seen on *Figure 20(a)* cluster 3 appears to contain the highest selling games in Japan whilst doing poorly in North America. Based on our analysis of the cluster profiles, cluster 3 has the highest percentage of games in a franchise (60%) and highest percentage of games on Nintendo platforms (45%). It also has the lowest percentage of games on Xbox platforms out of all clusters (15%). We plotted the same cluster data against EU and North America sales and

could see that cluster 3 does not perform well in the EU either. We can see from *Figure 20 (b)* that the remaining clusters are well spread out between both EU and North America. Details of the other significant cluster profile values can be seen in *Table 8* below.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
% games in a Franchise	45.75	56.34	60.00	46.80
% games on Nintendo platforms	34.40	39.44	45.00	34.40
% of games on Xbox platforms	25.40	21.12	15.00	23.07
% of games on PlayStation platforms	36.32	38.03	25.00	23.00

Table 8 Summarising Significant Cluster Profile Details for Action Games Genre



(a) *(b)*

Figure 21 (a) Scatter plot of clusters for regional Japan and North America sales for role playing genre (b)Scatter plot of clusters for regional EU and North America sales for role playing genre

Another genre of interest turned out to be role playing. In *Figure21 (a)* we can again see a very distinct cluster - cluster 3. This appears to be a cluster of games that have high sales in North America and very low sales in Japan. Upon further analysis of the cluster profiles (refer to *Table 9*) we can see that cluster 3 contains a high percentage of games rated M (42.85%). In *Figure 21(b)* cluster 3 is still performing the best in terms of sales in North America but also performs better in the EU with sales values going almost as high as 0.2 compared to most of Japan's sales values being at 0.

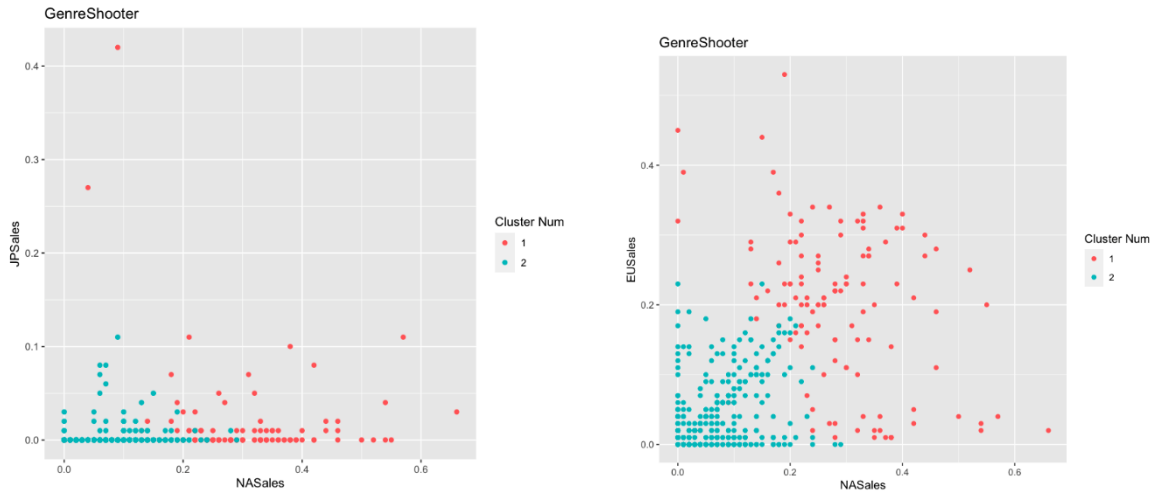
Another interesting observation that can be made is that cluster 2 contains most of the bestselling games in Japan. One of the most prominent features of cluster 2 (similar to what we

had analysed for the action genre) is it has the highest percentage of games in a franchise (56.25%). Cluster 2 also has the highest percentage of games on Playstation platforms (54.16%) and the lowest percentage of both games on Xbox platforms (12.50%) and Nintendo platforms (12.50%).

Cluster 1 is an interesting cluster to profile as it performs poorly in Japan, decently in North America (most sales are between 0.2 and 0.4) and has a couple of outliers that perform really well in the EU see *Figure 21(b)*. Cluster 1 has the highest percentage games with a rating of E in it (62.5%).

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
% games in a Franchise	50.00	56.25	28.57	36.84	48.91
% games on Nintendo platforms	37.50	12.50	28.57	15.79	35.87
% of games on Xbox platforms	31.25	12.50	28.57	26.32	27.17
% of games on PlayStation platforms	31.25	54.16	42.85	47.37	32.97
% of games rated M	18.75	20.83	42.85	15.79	19.93
% of games rated E	62.5	33.33	28.57	42.12	35.14

Table 9 Summarising Significant Cluster Profile Details for Role Playing Genre



(a) *(b)*

Figure 22 (a) Scatter plot of clusters for regional Japan and North America sales for shooter genre (b) Scatter plot of clusters for regional EU and North America sales for shooter genre

Figure 22(a), (b) above are the final cluster plots that we believe are worth mentioning. The cluster profiles themselves did not provide much valuable insight but *Figure 22(a)* shows that the games genre of shooters sells very poorly in Japan with a lot of the records lying on 0 or below 0.1 (with a few outliers). *Figure 22(b)* shows a very different graph of the same clusters but against EU and North America sales and we can see shooters sell far better.

5 Discussion

Having explored various techniques of analysing the data, there were some deductions we were able to make in regards to how to create a profitable game and where to distribute it. One of our hypotheses was that the characteristics of a game, such as genre, platform and rating, will have a large impact on the sales, regardless of how well the game is received.

5.1 Which Game to Make?

In our preliminary analysis we have found which games typically sell the best on average in our dataset, but initially we did not find any relations between them. For that reason, our main motivation was to look at clusters which highlight the different factors affecting highest sales, and the patterns between them. Interestingly, we found two main genres that outperformed all of the other ones, strategy and role playing games. Strategy games had the highest mean global sales from all of the clusters and that cluster contained the highest proportions of specific ratings associated with high sales. The main finding was that the teen (T) rated games achieved the highest sales for this genre and platforms had little effect on how well the game sold. Contextually speaking, T rated games were probably higher performing in the cluster on average because they all featured war-like scenarios. Role-playing games performed very well because the proportion of global sales in the cluster and the number of datapoints was the highest. Once again, T rated games performed very well, along with games rated everyone (E). Furthermore, games specifically sold on PlayStation and Xbox platforms performed very well. Analysis from the *4.1 Regression* models has shown that ratings have a high importance and some correlation to the global sales, and hence it might be useful to focus on which ratings performed well overall. In general, the clusters highlighted that games which were rated E and games that were released on console platforms have high sales on average. These findings suggest that if we wanted to design a game within any genre, we could still have potential for designing a well performing game (in terms of global sales) if we make an E rated game for consoles platforms such as Xbox and PlayStation. Equipped with this general knowledge we can narrow down our plan to the above-mentioned types of games, as we can expect them to perform better than others on average.

5.2 Do Franchises Affect Game Sales?

Another important expectation was that games in franchises will dominate the market with higher sales. Firstly, using logistic regression we were hoping to produce a model which would

be able to predict if the game was in franchises depending on the revenue. However, a logistic regression graph showed the distribution of both games in franchises and not in franchises across the whole spectrum. This entails that there are a lot of games which are not in a franchise and they still perform very well on the market. This was further supported when we trained a logistic model. This model got slightly higher than random accuracy, meaning that in most cases it is hard to determine whether it is in a franchise or not based on revenue. When looking at the context of franchise games, there can be some stigma towards games that are sequels and the scrutiny they can be put under through their comparison to the previous game in the franchise. Given that we are planning on making a new game, that is good news for us and other small start-ups, as it will not be known to anyone and it is helpful to know that the franchise doesn't have a large effect. Likewise, clustering has yielded similar results, showing that the best clusters were not correlated to games in franchises. To verify our findings, we used random forests to predict what would constitute the best outcome. Having trained a model with fairly strong correlation and high accuracy, we were able to confirm some of the trends we found in other data. For instance, the trend that was found prominent throughout the bestselling games, was making games on PlayStation and consoles. However, random forests showed that the top combinations are composed of games that are in the franchise. Due to the other significant analysis, this could be due to some regional differences, but in most cases, we found that it does not have a strong effect on the outcome.

5.3 When Should the Game be Released?

In our analysis of the Time Series graphs *(see 3.4Time Series Sequential Classification (TSC))*, we tried looking at trends in platforms however they weren't helpful. We think this was the case because the older platforms in the dataset, games will stop being made after a few years or when the next generation of the console gets released. This means the data points will be sparser as time goes on. Also, in the dataset overall sales of all time are put on the original release date rather than spread over the years. Furthermore, the trends in most platforms won't be helpful to Nantendo as new games will be on the newest generation of consoles; Nantendo would rather make a game for the PS5 rather than an Atari2600.

The results of this research indicated that the increase in global sales for Shooter and Strategy games would be a good idea for Nantendo to make a game in either of these genres. This means that in recent years, strategy and shooter games have become more popular. This is especially relevant when the dataset in general had a decline in global sales over time. In

contrast, the lower sales in Japan suggest that if the game genre was Strategy, that it would perform worse in Japan. This could be interpreted that more effort should be put into advertising it in Japan to make up for the lower interest, and possibly that Strategy games are overshadowed by other genres producing a lower interest.

One interesting finding is that the highest-selling games are sold around September / October. A naïve conclusion to make from this would be that it would be best to sell games around this time as well. However, the big games companies release triple-A games around this time in time for Christmas. To also release a game at this time as a small development company like Nantendo would mean Nantendo would be competing with these large companies and cause their game to be overshadowed by these big releases.

For the deep neural network on the time series, while the R squared isn't perfect, this model is able to generalise the average trend of global sales over the year and ignore outliers to predict the sales of a certain genre at a certain point in time.

5.4 Do Regions Affect Sales?

Furthermore, our original dataset contained regional sales, and thus, we tried to get more regional data so that we can try to find some trends in those regions of interest. We predicted that different regional demographics will have influence on how well the games are being sold, as well as, different regions having impact on which types of games are being sold.

Firstly, *3.1Regression* gave us an insight into how the distribution of age and gender population can affect the game sales. We found that there is some correlation between what the ratio between male and female is in a country. Even though the R^2 analysis has predicted it as being a weak correlation, the model can fairly confidently predict the test set. To further back it up, random forests showed that the demographics were correlated to global sales and change in it had some significant effect, therefore we made some conclusions regarding it. The general trend was that as the male ratio in the population increases, so do the average sales in that region. This information is useful as it allows us to make assumptions on which regions our game is more likely to be sold more in. The main interpretation that we can make out of it, is where the money should be put to advertise the game and distribute them. Most games are sold online nowadays, but some portions are still sold in shops, so selling the games in smart places can result in higher sales. They could be placed primarily in the countries with a high male

population and thus promoting it there could boost the sales of the game. However, there is also a possibility that in those regions the game could be bought regardless, and targeting marketing in countries with lower sales could boost the overall sales. The data can't necessarily tell us which of these would pay off more and there could be other variables affecting the results, but if marketing was targeted in countries where video games are more popular in general, then a new unknown game would probably get a boost. Similar conclusions can be made for the age distribution, where the age distribution is skewed towards the mean being higher. As it was previously mentioned, ratings have a high importance on the sales, and hence this could explain why higher game ratings, such as M and T rated games are most popular according to *2.4 Exploratory Analysis* our data exploration and clustering. The problem with the predictions is that the trend is unlikely to continue, i.e. if the population is too old then the game sales would at some point drop, but we can't conclude that as the data doesn't cover those cases. Finally, determining which regions are the best to sell in, can tell us what content can be placed in a game. Different regions have different laws regarding what can be placed in a game. For instance, in Battlefield 4, China got depicted as the "bad guys" and it got banned. Knowing if a game performs well in a region will let us consider their laws, and if it doesn't we wouldn't have to worry too much about it as it wouldn't impact the sales. To conclude, following these trends can point us in the direction of where to advertise and sell our game to get the most profit from it, but the results show that even though the trends exist, there is a lot of data that doesn't necessarily follow the trend.

Secondly, by looking at the groups of clusters in regional sales, we were able to draw conclusions for which genres do best in which regions. We already analysed what are the global trends, but looking at regional trends can give us more insight into how different games would be received. Previously, we mentioned role playing games being one of the best genres to go for. This analysis has shown us that role playing games in America perform especially well if they are M rated, and in Japan if they are in franchises. Those are some considerations we would have to take if we were to design that type of game. Japan has been shown to have some major differences from the rest of the world. On top of the role playing games, shooters and action games performed very well, but they have very specific characteristics that are not found in other regions. For instance, shooters do not sell well there and action games only perform well when in franchises and mainly on Nintendo. In contrast, the western culture is significantly different, as shooters sell very well there, most likely due to the different culture, and action

games perform badly on platforms and are mainly sold on PCs. E rated games are also very popular in Europe.

Bridging the different analyses together, we can assume that in order to create a bestselling game, there are some trends that we can follow to make the game more appealing across different regions. We have mentioned that role-playing games and strategy games are the two ones that have been found to work very well. After the regional analysis, it can be concluded that making a role playing game would have to be more specific to appeal to the greater public, especially in America and Japan. Focusing on any other famous genre of games, like shooters and actions, also can largely have different receptions in different regions based on different characteristics. Therefore, the game type that seems to be the best option is a strategy game, which would sell well globally and could be deployed on all and any platform to perform well. The rating for it should also be T, as that will most likely maximise the profits.

6 Conclusion

In summary of the findings in the discussion section, we are able to recommend a specific set of games to be made based on what region we hope to distribute the game to and advertise in the most. In terms of creating a game that has the best potential to sell-high globally, a Strategy game, with a maturity rating of Teen (13 and older) and is not involved in a franchise, should provide the optimal chances to reach the higher-selling bracket. Whilst highest selling games were sold during the September to October window, this is from triple-A game studios which would overshadow Nantendo's release, therefore a game release in the summer would be more profitable.

Based on the decision of the upper management team at Nantendo, we may need to focus our efforts towards advertising and emphasising the distribution of our sales in a specific region. The most optimal region for selling our game is North America. From our analysis, games with the strategy genre, with a rating of Mature (17 and older), on any next gen console, performed the best in North America.

In relation to the problem we were working on, these results proved similar to our hypothesis stating that fields apart from critic reviews would be highly impactful on the sales of games globally. This was shown to be the case with the game genre as well as the platform it was sold on. After having analysed our findings from the dataset, we can say with some certainty that a general structure for the game and what restrictions we have to adhere to based on ratings and regions can be defined. Therefore, according to the dataset we selected, we have provided an appropriate solution. However, based on the context of video games and the potential fields to analyse concerning them, our analysis leaves more to be desired. This is discussed more in the future improvements section.

6.1 Future Improvement

As the analysis came to a close, our team decided that with further time and resources given to us by Nantendo, we could narrow down our search by looking at more datasets specific to what constitutes a game rather than the more generalised, global dataset that we initially found. An example dataset that could be useful to magnify our search into what makes a high-selling game could be a dataset pulled from SteamDB [11]. With Steam being a leading video game distributor across the entire world, it provides open access to data on all of its games as well as a wider variety of potential fields to choose from.

Another approach that we could take, given further time and resources assigned to us, would be to scrape for more fields to add to our dataset that describes the game in better detail. These fields include:

- Average playtime
- Number of consoles released on
- Video game price
- Demo Availability
- Popular tags associated with the game

We believe either of these approaches would be able to help narrow down even further on finding out what makes games sell and how we can help point Nantendo in the direction of being a more successful company.

7 Reference List

[1] A. Beattie, “How the Video Game indusrty is changing,” Bloomberg, 2021. [Online]. Available: https://www.investopedia.com/articles/investing/053115/how-video-game-industry-changing.asp.

[2] Phediuk, “History of worldwide video game industry revenue since 1971 , by sector (arcade/console/handheld/PC),” Bloomberg, 2019. [Online]. Available: https://www.resetera.com/threads/history-of-worldwide-video-game-industry-revenue-since-1971-by-sector-arcade-console-handheld-pc-from-bloomberg.96568/.

[3] S. Santosh, “Data Science In Gaming Industry,” Analytics Vidhya, May 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/data-science-in-gaming-industry/.

[4] C. Blomgren, “Forecasting initial sales of video games using Youtube trends,” 2022. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1680040/FULLTEXT01.pdf.

[5] Statista, “Video Games,” Nov 2022. [Online]. Available: https://www.statista.com/outlook/dmo/digital-media/video-games/worldwide.

[6] R. Kirubi, “Video Game Sales with Ratings,” Kaggle, [Online]. Available: https://www.kaggle.com/datasets/rush4ratio/video-game-sales-with-ratings.

[7] H. R. a. M. Roser, “Gender Ratio,” 2019. [Online]. Available: https://ourworldindata.org/gender-ratio.

[8] H. R. a. M. Roser, “Age Structure,” Our World in Data, 2019. [Online]. Available: https://ourworldindata.org/age-structure.

[9] D. Beltekian, “World map region definitions,” Our Wolrd in Data, [Online]. Available: https://ourworldindata.org/world-region-map-definitions.

[10] M. B. Editor, “Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?,” Minitab, 2013. [Online]. Available:

https://blog.minitab.com/en/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit.

[11] “Database of everything on Steam.,” SteamDB, [Online]. Available: https://steamdb.info/.

[12] “Games,” Metacritics, [Online]. Available: https://www.metacritic.com/game.

Appendix A – Project Plan

Refer to Figure 23 Project Plan for our project plan.

Week 1 (Planning)	Project Proposal: • Introduction • Business Analytics Tasks (***Anastasia & Felix***) • Problem Definition & Dataset (***Manuel & Michal***) • Expectations (***Matthew & Alex***) • Project Plan
Week 2 (Data Processing)	Data Pre-processing • Data Scraping – extra data (***Manuel & Michal***) • Merging Datasets (demographics per region, other relevant game stats) (***Manuel & Michal***) • Data transformation (***Matthew & Alex***) Data cleaning (***Anastasia & Felix***) : • Data Scraping – missing data • Verifying correctness of data (identifying outliers, duplicates)
Week 3 (Developing models)	Models • Regression (***Manuel & Michal***) • Clustering (***Anastasia & Felix***) • Neural Networks and TSSC (***Matthew & Alex***) Project Report
Week 4 (Presenting results)	Project Report Presentation & preparation

Figure 23 Project Plan