FACTORS AFFECTING THE SALES OF VIDEOGAMES

Big Data Fundamentals Coursework

Contents

Chapter 1

Introduction

Data analytics is more important in the video game industry than ever before, with studios using data to decide what kind of games to produce based on what will sell best, to tweak existing games in order to increase how much time players invest in it and to improve the effectiveness of marketing for upcoming releases.

Some of the factors that affect the sales performance of a game include the genre of the videogame, the platform it is released on and the publisher/development studio that releases the game. Additionally, the market to which the game is targeted at and sold to also has an impact, with games released in one market not necessarily performing as successfully in another.

The process of adapting a game from one market to another is known as localisation. This process involves multiple aspects, consisting of the following:

- **Translation**: the translation of the game's text and voice acting from the original language into the new market's language.
- **Editing and Proofreading**: correcting spelling and semantic errors and checking the consistency of the translation.
- **Integration**: integrating the newly translated materials into the game by changing and adjusting the User Interface to make sure everything displays correctly.
- **Regional adaptation**: For instance, in Germany references can't be made to the Nazis or fascist imagery, whilst in Australia all references to drugs and alcohol need to be removed from your game.

With an average translation cost of 0.07 to 0.15 Euros per word and with game scripts of 100's of thousands of words, localising a videogame from one market to another can be very expensive, and, in the case of a localised games that don't meet sales expectation, very costly.

In light of this, my aim in this report is to analyse the data on what effect genre, platform and publisher have on a games global sales, as well as to determine whether it is possible to predict a game's sales in one region based on pre-existing sales data from another region.

Chapter 2

Dataset – Videogame Sales from 1982 to 2017

The data used for this project is freely available from Kaggle, with the original source of the data being from vgchartz.com, a site with lots of tables and charts of videogame related data. This specific table collects data on videogame sales from 1982 to 2017.

Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82

Table 2.1 First 3 rows of the videogame dataset

The dataset is held in a two-dimensional Pandas dataframe with dimensions of 16598 x 11. It contains several columns, each representing the following:

- Rank – the game's position in the overall sales rankings
- Name – the title of the game
- Platform – the system on which the game was released, eg. Wii or NES
- Year – the year in which the game was released
- Genre – the genre of the game released, eg. Sports or Action
- Publisher – the publishing company that released the game
- NA_Sales, EU_Sales, JP_Sales, Other Sales – these are the sales for each respective region (in millions)
- Global_Sales – this is the total worldwide sales, also in millions

When analysing the data, I initially started by checking for any columns containing null values as well as checking to see if there were any extreme values. I found that both the Year and Publisher column contained several rows with null values, so I discarded these to clean up the table. Then I plotted a graph of the number of games released and sold each year:

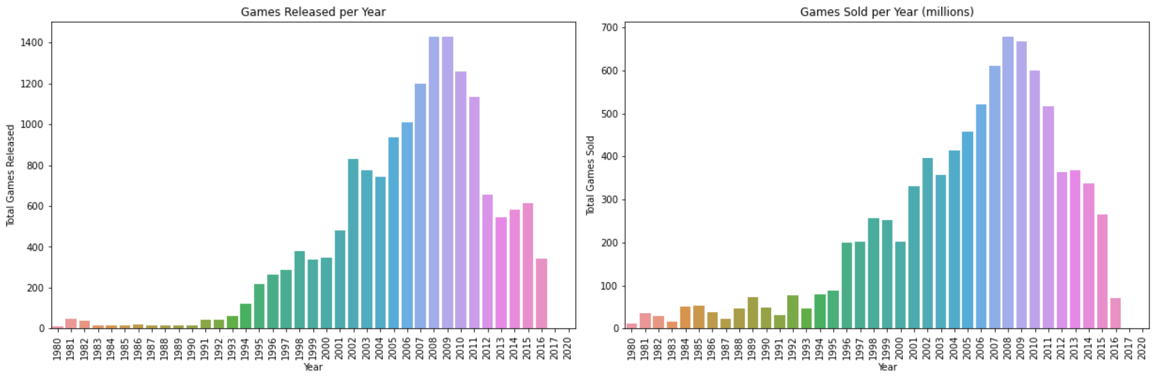

Fig 2.1 Graphs of Number of Games Released and Number of Games Sold each year from 1982 to 2017

From these charts it seems that the number of unique games released and sold each year has increased since 1982 up to a peak in 2008/2009, with the number of new games released and sold dropping by over 50% since then.

Upon closer examination of the charts, I also noticed that there was a game apparently released in 2020, despite the dataset only containing games released up until 2017. I extracted the row containing the year 2020 and found that 'Imagine: Makeup Artist' had been mistakenly labelled as coming out in 2020 when it was in fact released in 2009. After replacing the year with the correct value, I also dropped the columns from 2017 as it seemed to be incomplete. Finally, after cleaning the data I decided to truly begin my analysis by plotting the top 10 games in terms of global sales.

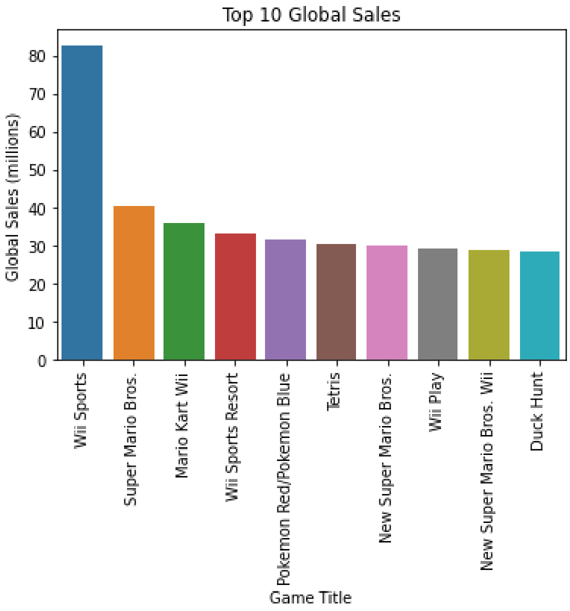

Fig 2.2 Bar Chart of Top 10 Games in the World by Sales

Wii Sports seems to be the most popular game worldwide by a large margin, with every other game in the top ten seeming to be by the same Publisher, Nintendo. I wanted to check if this was simply due to Nintendo releasing the most games or if the chart reflected the actual quality of their games instead, so I plotted charts of the top 10 publishers by releases and sales to compare.

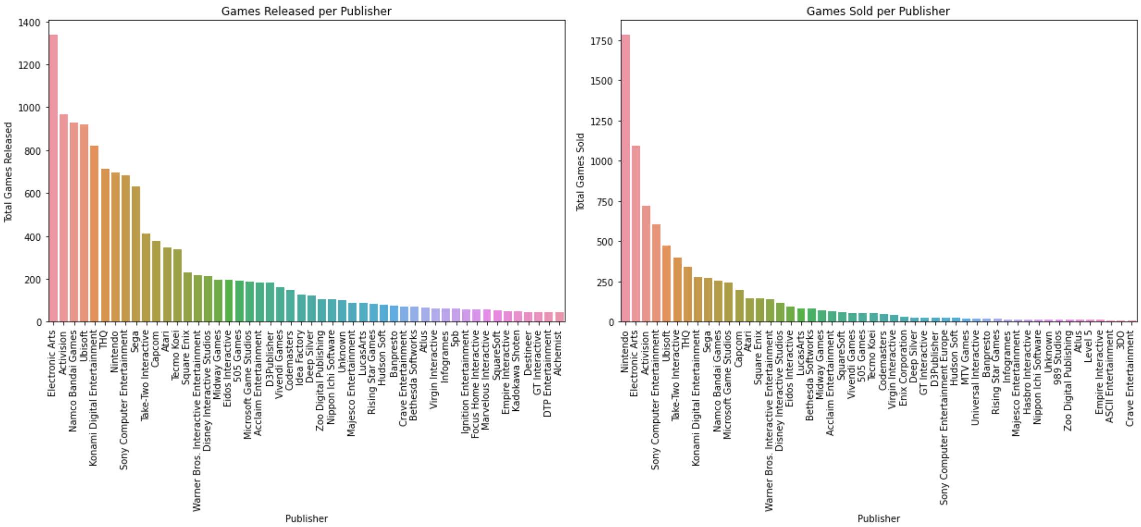

Fig 2.3 Graphs of Number of Games Released and Number of Games Sold by each Publisher

Surprisingly Nintendo is only the 7th highest publishing company in terms of the number of titles released, however it is overwhelmingly the company with the best selling games, beating out all 6 of the companies above it in terms of releases, with over 3750 million units sold, with the second place company Electronic Arts barely passing 1000 million sales.

To see if Nintendo dominated the top 10 games in every individual region, I decided to check if each region followed the same or a similar trend to the global trend.

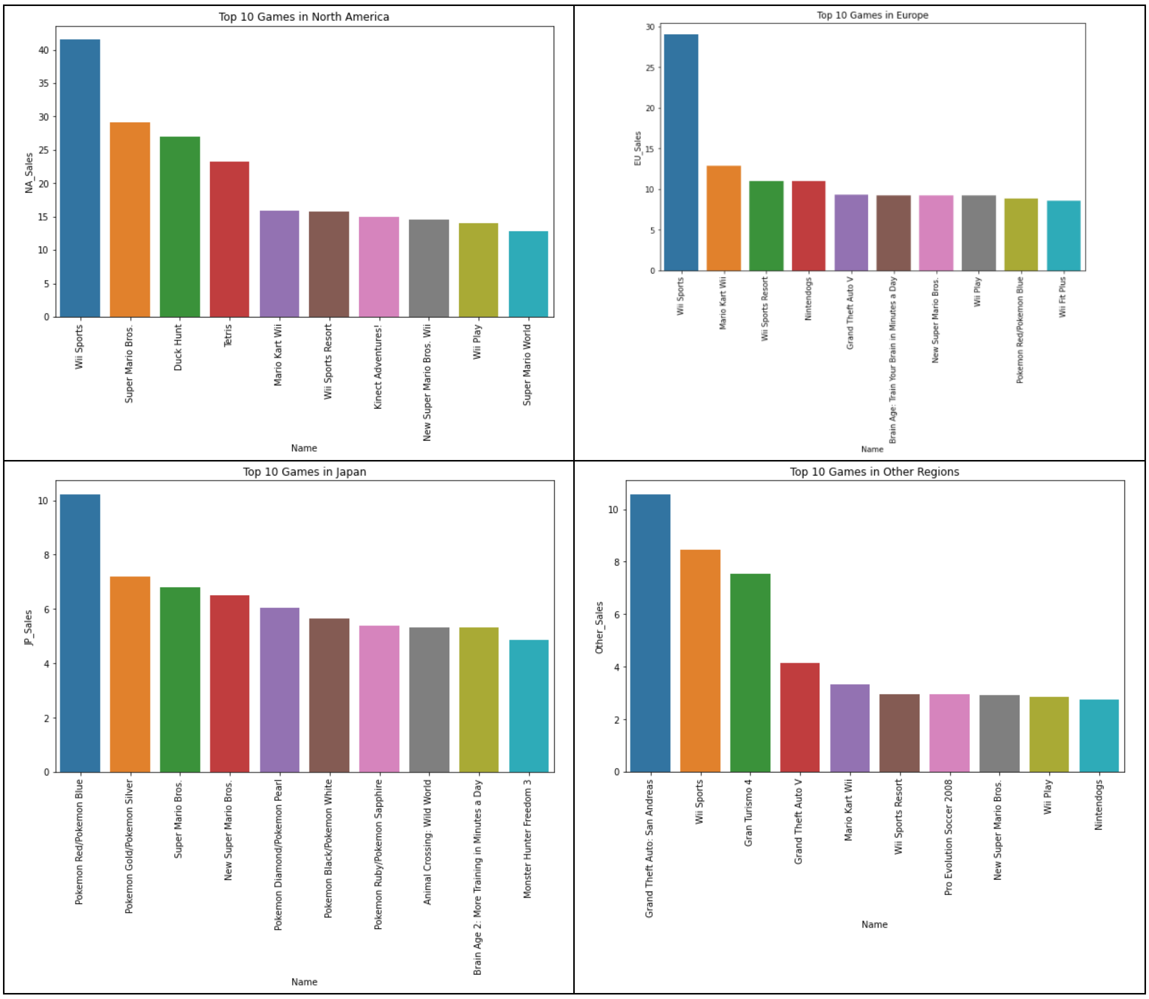

Fig 2.4 Bar Chart of Top 10 Games in each region by Sales

The global pattern seems to mostly hold for both North America and Europe, with the majority of the games in the global top 10 also appearing in the global top 10. Sales in Other regions seem to partly follow the trend, however the best-selling game in these regions doesn't even appear in the global top 10, although it does contain some of the top 10. Japan however breaks the trend completely, only sharing one game with the global top sales, with all the other games being a top 10 best seller in Japan alone. Curious about this I plotted a chart of the games sold each year, separated by region, to see if they followed similar sales patterns.

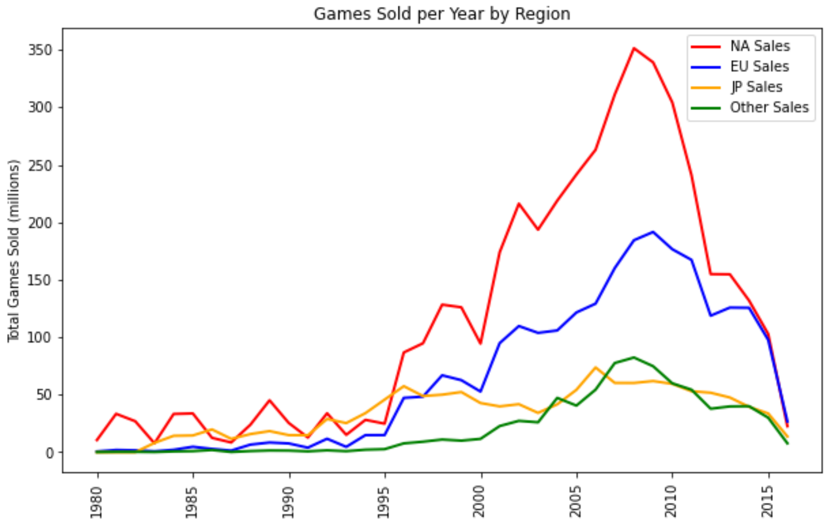

Fig 2.5 Chart of Game Sales per Year, separated by Region

From the chart of sales over time it seems that North America, Europe and Other regions all follow the same rough trend from 1995 onwards, with crests and troughs appearing in similar places for each, albeit with different magnitudes. Japan however seems to follow a pattern all of its own that appears unaffected by the other regions. As this could be a major problem to predicting the sales performance of games across regions, I created a heatmap of the sales for each region, in order to see if there was any correlation present at all.

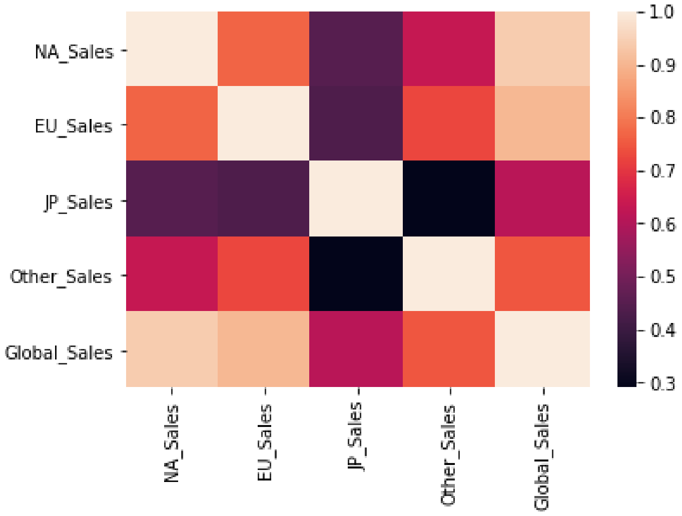

Fig 2.6 Heatmap showing correlation in sales data for each region

Judging from the heatmap it seemed by hypothesis was correct with European and North American sales showing a strong correlation of 80-90%, with Other Sales having a fairly strong correlation of ~75% with Europe, although it also has a somewhat lower correlation of ~60% with North America.

Unsurprisingly Japanese sales show the lowest correlation with other regions, with less than 40% correlation with North America and Europe and less than 30% with the Other regions. This suggests that either the relationship in sales between Japan and other regions is either more complex or non-existent. In order to further explore this, I decided to carry out some basic machine learning.

Chapter 3

Unsupervised Analysis – Clustering

O'Reilly's Python Data Science Handbook defines Unsupervised Machine Learning as so:

> *Unsupervised learning* involves modelling the features of a dataset without reference to any label and is often described as "letting the dataset speak for itself." These models include tasks such as *clustering* and *dimensionality reduction.* Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more succinct representations of the data. (VanderPlas, 2017, p.322)

The aim for using Unsupervised Analysis in this project was to find any groups present in the dataset, with each group ideally representing a specific region. The dataset is not ideal for unsupervised analysis as all the numerical data are stored in float format and exist on a continuous scale, making it difficult to cluster. To help ease this, each sales value has been rounded to the nearest whole number, reducing the number of unique sales values from 1735 down to a much more manageable 81.

A multidimensional matrix of European, Japanese and Other Sales data was extracted from the dataset, while the American Sales were transformed into a one-dimensional array in order to see if the sales performance in North America could be predicted based off another region's data.

This newly simplified data will be tested with two different clustering method, Hierarchical and K-Means.

Hierarchical

Hierarchical clustering is the name for a group of clustering algorithms that group datapoints into similar clusters by repeatedly merging or splitting them, with the hierarchy of clusters being represented as a tree (dendrogram).

The Agglomerative Clustering function used on this data set carries out hierarchical clustering by using a bottom-up approach, with each observation initially in a cluster by itself with clusters being successively merged into as few clusters as possible. The linkage criteria determine the how the merge strategy is carried out:

- **Ward:** This minimises the sum of squared differences within all clusters.
- **Maximum linkage:** This minimises the maximum distance between observations of pairs of clusters.
- **Average linkage:** This minimises the mean of the distances between all observations of pairs of clusters.
- **Single linkage:** This minimises the distance between the closest observations of pairs for each c Because the problem isn't well suited to clustering, all linkage and affinity strategies

As we are yet unsure which clustering method would best suit the data, I used all possible combinations available with Scikit-Learn to make sure I would not miss any positive results.

Figure 3.1 below displays the effectiveness of each linkage criterion:

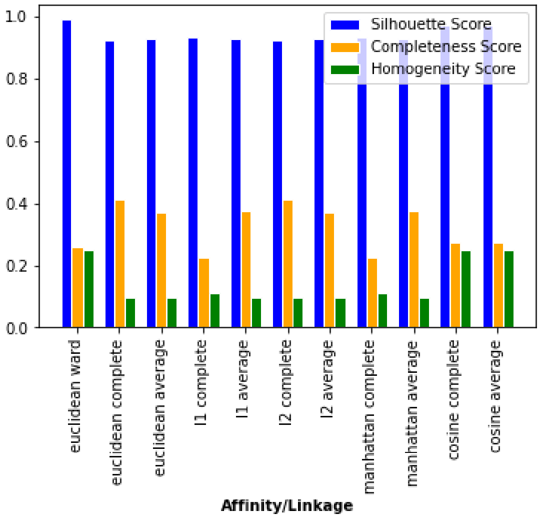

Fig 3.1 Scoring metrics for each Affinity/Linkage Combination

The three metrics measured work like so:

- **Silhouette score** is used to calculate the effectiveness of a clustering technique. Its value ranges from -1 to 1, with 1 signifying well assigned and clearly distinguished clusters and -1 representing clusters that have been assigned in the wrong way
- **Completeness score** measures the proportion of items of the same class that are assigned to the same cluster. It ranges from 1 to -1, where 1 means that all items have been assigned the correct cluster.
- **Homogeneity score** is similar to the completeness score where it measures what proportion of a cluster are made up of items of the same type. It ranges from 1 to -1, with 1 representing all items for each cluster are of the same type.

All methods were highly effective at generating distinct clusters, with the lowest silhouette score still remaining above 0.9. The completeness score had much higher variance with Euclidean complete performing best with a completeness score of 0.412, which means less than 50% of items of the same class are correctly assigned to the same cluster. Homogeneity performed far worse with the best linkage (Cosine Complete) only reaching a value of 0.253 suggesting clusters contain only 25% of the same item. Unfortunately, this means that even the best performing method is effective only about a third of the time.

K-Means

K-Means is another type of clustering algorithm that takes a specified number of clusters (n) and clusters data by trying to separate samples into n groups of equal variance. For this dataset it would be expected that 4 clusters would be the optimal number as this is the number of regions used in the table. To check this, I carried out K-Means clustering for 2 to 30 clusters, the results of which are plotted on figure 3.2. From this it seems that 6 clusters perform best, however as the completeness and homogeneity scores for both are still quite low, it is probably best to discount the K-Means results.

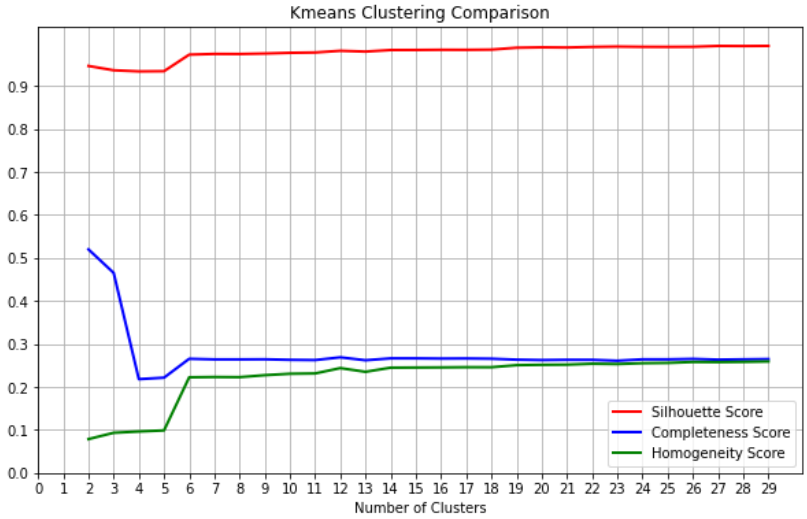

Fig 3.2 Scoring metrics for different numbers of clusters

Chapter 4

Supervised Approach

O’Reilly’s Python Data Science Handbook defines Unsupervised Machine Learning as so:

> *Supervised learning* involves somehow modelling the relationship between measured features of data and some label associated with the data; once this model is determined, it can be used to apply labels to new, unknown data. This is further subdivided into *classification* tasks and *regression* tasks: in classification, the labels are discrete categories, while in regression, the labels are continuous quantities. (VanderPlas, 2017, p.322)

In this case data refers to the sales quantities for each region, while the labels are the regions themselves. In order to prepare the data for supervised analysis, I converted the dataframe from a wide to a long format, converting the multiple regional sales columns into a column for regions and a column for sales. The purpose of the analysis is to determine what effect the Region has on the sales of a game.

	Name	**Year**	**Platform**	**Genre**	**Publisher**	**Region**	**Sales**
0	Wii Sports	2006	Wii	Sports	Nintendo	NA	41.0
1	Super Mario Bros.	1985	NES	Platform	Nintendo	NA	29.0
2	Mario Kart Wii	2008	Wii	Racing	Nintendo	NA	16.0
3	Wii Sports Resort	2009	Wii	Sports	Nintendo	NA	16.0
4	Pokemon Red/Pokemon Blue	1996	GB	Role-Playing	Nintendo	NA	11.0

Table 4.1 Long form version of Table 2.1

After splitting the data into a one-dimensional array containing the Region and a multi-dimensional array containing the sales, the data was once again split into a training part and a test part in the ratio 75:25. This was then used to carry out a logistical regression on the data.

Logistic regression works by determining the probability of a binary event occurring based on a set of inputs. While it ideally would be used to compare two related numerical values to one another, it can still be applied to this question by restructuring the Regions as binary questions, eg. ‘Are these sales figures likely to be from North America?’ or ‘Are these figures likely to be from Japan?’ the value output will be condition with the highest probability of occurring.

However, by examining the classification report, it’s clear that this analysis won’t provide accurate values.

	Precision	recall	f1-score	support
EU	0.19	0.08	0.11	4162
Global	0.47	0.31	0.38	4075
JP	0.11	0.05	0.07	4014
NA	0.29	0.11	0.16	4138
Other	0.27	0.85	0.41	3971
accuracy			0.28	20360
macro avg	0.27	0.28	0.23	20360
Weighted avg	0.27	0.28	0.22	20360

Table 4.2 Classification Report for Logistic Regression

The column headings in the classification report convey the following:

- **Precision**: The ratio of true positives to all data points marked as positive
- **Recall**: The proportion of all positives that were correctly identified as positive
- **F1**: A weighted average between precision and recall

Unfortunately, with the way the data is set up, even in the best scenario, predicting which sales figures are Global, Logistical Regression can’t even correctly identify values half of the time. Other methods of supervised learning, including Decision Tree modelling, K Nearest Neighbour and Naïve-Bayes also had similarly low levels of accuracy and even adjusting the test/train ratio had only minor effects on the accuracy of the predictions

Chapter 5

Reflections and Conclusion

If I was to attempt this assignment again I would have either looked to find other tables that could be combined with the one used to provide more numerical variables for analysis, or instead selected a different table with more than one type of numerical variable as well as sales.

It seems almost certain that better methods could have been used to chosen to analyse the data.

The Hierarchical clustering algorithm showed some promise as it had consistently high silhouette scores and middling homogeneity and completeness scores. Perhaps by better preparing the data or by including other numerical metrics relative to the games it would be much easier for it to accurately cluster the data.

As shown above the K-means clustering algorithm delivered bad outputs, with a terrible homogeneity score and a plummeting completeness score. This is likely due to the need for the labels to have numeric values to allow the distances between them to be compared. However, as the regions don’t lend themselves to being easily assigned numbers relative to one another, the calculated distances don’t reflect the data properly. This could perhaps be improved by using more in-depth regions, where the distances used could be from one country or even city to another.

The supervised regression methods were ill-suited for this analysis as they require multiple numerical values that have some form of relationship between them in order to function. As it stands the connections between the data are too weak, and there is too much noise present for any meaningful conclusions to be drawn.

While there is obviously a relationship between sales performance across regions, unfortunately I was unable to uncover exactly what it is with the data at hand. Perhaps a more knowledgeable data scientist with access to more data on the subject will be able to discover the intricacies of this relationship at a future date.

Appendix A

Environment

Language: Python 3.7.4

IDE: Jupyter 6.0.0

Dataset source: https://www.kaggle.com/rishidamarla/worldwide-video-game-sales

Packages used:

- pandas
- matplotlib.pyplot
- seaborn
- sklea

References

- Michael Waskom (2020) Overview of seaborn plotting functions, Available at: https://seaborn.pydata.org/tutorial.html (Accessed: October 2020).
- william007 (2017) How to Prevent Overlapping x-axis :abels in sns.countplot, Available at: https://stackoverflow.com/questions/42528921/how-to-prevent-overlapping-x-axis-labels-in-sns-countplot (Accessed: October 2020).
- Game Spy (2013) Imagine: Makeup Artist, Available at: http://ds.gamespy.com/nintendo-ds/imaginemakeup-artist/ (Accessed: October 2020).
- Data to Fish (2020) How to Replace Values in Pandas DataFrame, Available at: https://datatofish.com/replace-values-pandas-dataframe/ (Accessed: October 2020).
- nbecker (2019) pandas groupby sort descending order, Available at: https://stackoverflow.com/questions/27018622/pandas-groupby-sort-descending-order/36316186 (Accessed: October 2020).
- Yan Holtz (2017) #122 Multiple lines chart, Available at: https://python-graph-gallery.com/122-multiple-lines-chart/ (Accessed: October 2020).
- Data for Everybody (2020) How to use Pandas, Matplotlib and Seaborn to draw pie charts (or their alternatives) in Python?, Available at: https://www.dataforeverybody.com/matplotlib-seaborn-pie-charts/ (Accessed: October 2020).
- DataScience Made Simple (2020) Reshape wide to long in pandas python with melt() function, Available at: https://www.datasciencemadesimple.com/reshape-wide-long-pandas-python-melt-function/ (Accessed: October 2020).
- Aurélien Géron (2017) Hands on Machine Learning with SciKit-Learn & TensorFlow, 2nd edn., United States of America: O'Reilly.
- Jake VanderPlas (2017) Python Data Science Handbook: Essential Tools for Working with Data, 2nd edn., United States of America: O'Reilly.
- Sanghamitra Bandyopadhyay, Sriparna Saha (2013) Unsupervised Classification, Berlin: Springer.
- Olivier Gaudard (2019) #11 Grouped barplot, Available at: https://python-graph-gallery.com/11-grouped-barplot/ (Accessed: November 2020).
- Ian Dransfield (2016) The Future of Data Analysis, Available at: https://www.gamesindustry.biz/articles/2016-03-24-the-future-of-data-analysis (Accessed: November 2020).
- Ari Vivekanandarajah (2018) How data analytics software is changing the video game industry, Available at: https://seleritysas.com/blog/2018/12/14/data-analytics-software-video-game-industry/#:~:text=The%20information%20is%20particularly%20important,system%20can%20kill%20player%20engagement.&text=Data%20analytics%20software%20is%20particularly,time%20information%20on%20player%20activities. (Accessed: November 2020).
- Nikolay Bondarenko (2018) How video game localization works and how much it costs in 2018, Available at: https://www.gamasutra.com/blogs/NikolayBondarenko/20180914/326495/How_video_game_localization_works_and_how_much_it_costs_in_2018.php (Accessed: November 2020).
- Scikit-Learn (2020) 2.3. Clustering, Available at: https://scikit-learn.org/stable/modules/clustering.html (Accessed: November 2020).
- Giuseppe Bonaccorso (2018) Mastering Machine Learning Algorithms, USA: Packt Publishing.