# Turtle Games

## An analysis of Customers & Sales

Will Burton

2024

# Appendices



## Business Context

Turtle Games, a global game retailer and manufacturer, offers a diverse range of products including books, board games, video games, and toys. The company aims to enhance its sales performance by leveraging data on sales and customer information to better understand and capitalise on customer and market trends.

Our team of data analysts has been tasked with analysing this data to inform strategic decisions and drive growth for Turtle Games. We distilled the context down to one business question **How can Turtle Games use data to enhance customer loyalty and sales?**

Further exploration of the business context can be found in (Appendix 1**A**) and a SWOT analysis (Appendix 1**B**).

## Analytical Approach

The core business question is split into two parts:

**Part 1**: *Who are Turtle Games' customers and what are their habits?*

- Customer Insights (Appendix **2**)
- Customer Segmentation (Appendix **3**)
- Customer Feedback (Appendix **4**)

**Part 2**: *What do sales data reveal about market trends and performance?*

- Platform Performance (Appendix **5**)
- Sales Trends (Appendix **6**) & Predictive Modelling (Appendix **7**)

Please see (Appendix 1**C**) for the import and cleaning approach in Python and (Appendix) **1D** for R.

In Python, we used advanced analytical techniques to explore Turtle Games' dataset, focusing on multiple linear regression (MLR) , K-means clustering, and natural language processing (NLP).

- Utilising libraries like Scikit-learn, we executed multiple linear regression to identify potential variables influencing customer loyalty points, revealing spending score and total income as significant predictors.

- For customer segmentation, K-means clustering found some distinct customer groups based on spending_score offering the opportunity for targeted marketing strategies.
- Additionally, employing NLP tools such as NLTK, Text Blob and VADER we analysed customer reviews to extract sentiment and thematic trends, providing invaluable insights into customer satisfaction and preferences.

In R, our analytical strategy involved applying multiple regression analysis and conducting a thorough platform performance analysis.

- Using the ggplot2 and dplyr packages enabled a useful platform analysis, visualizing sales performance across gaming platforms. This allowed us to focus on platforms that significantly contribute to global and regional sales. Which in turn allows for strategic emphasis and potential to capitalise on market trends and optimize sales performance.
- Using the lm() function, we extended our investigation into sales data, finding correlating variables and using multiple regression again to quantify the impact of regional sales on global sales. This analysis confirmed the strong influence of NA and EU sales on global figures, allowing for potential future resource allocation decisions.

## Visualisation and Insights

Both Python and R presented different outputs with regards to insights and their corresponding visualisations. Colour vision deficiency friendly palettes have been used where possible and interactivity added to allow for situations where there are too many data points to be covered by colour alone.

In Python our focus again was on MLR, K-means and NLP :

- Pairplots helped us see which variables were best connected to our key metric of spending score. This allowed us to focus on the positive linear relationship between total income, and loyalty points early and effectively. Scatterplots too proved key for linear regression analysis, and they highlighted the predictive model's ability to forecast loyalty points with 83% accuracy, allowing Turtle Games to anticipate customer needs and behaviours.
- K means clustering relies on helpful visual statistical tools like The Elbow and Silhouette methods to help define segment numbers. Once defined, K Means uses scatter plots to easily define customer segments which allowed for discovery of several potentially underutilised customer groups.
- For NLP word clouds allowed us to get an immediate sense of the overall positive sentiment with frequent positive words like "*fun*," "*love*," "*good*," "*great*," and "*play*. Histograms too were especially useful for seeing the spread of distribution for sentiment scores across our models. Scatter plots revealed that the longer the feedback is, the more positive it tends to be.

In R we used visualisations to help evaluate our MLR models and to highlight key platform sales insights:

- Throughout our analysis in R, boxplots, histograms and scatterplots were key to understanding the shape and distribution of the dataset. In particular boxplots allowed us to immediately see the distribution of the sales data showing that most sales were relatively low but a few very successful selling games skewed the data considerably. A stacked barplot was a very successful way of

interpreting platform data, detailing the success of specific platforms (e.g., Wii, Xbox 360, PS3) and the variance in regional preferences.

- Finally, with the regression analysis, scatter plots proved invaluable again allowing for an accessible way to visualise accuracy and errors. Allowing us to explore relationships between regional and global sales, underscoring the model's strong predictive power and the implications for targeting marketing and development efforts.

## Patterns and Predictions

**Customer Insights** – An MLR model using total_income and spending_score allows us to explain and predict loyalty_points accurately.
However, the presence of heteroscedasticity and outliers which skew the data will need to be addressed.
Finally, the focus should be on improving the quality data that already correlates to loyalty_points, rather than adding new variables.

**Customer segmentation** – Corelating nicely with our earlier analysis we observed that high loyalty and spending scores are strongly correlated, indicating that engaged customers significantly contribute to revenue. Use customer segments identified to focus on loyalty programs, recognition and identify barriers to increased spending. Please consider the individual cluster analyses in (Appendix **3** of this report) for full details.

**Customer feedback** – While the report showed an overall positive sentiment, attention needs to be paid to harnessing recommendations from positive/neutral feedback. Also consider investing in multilingual NLP tools and proactively investigating negative customer feedback.

**Platform performance** – Our stacked bar chart gave a useful understanding of the spread and distribution of global platform sales. Close attention must be paid to shifting market preferences, such as the emerging dominance of PlayStation in Europe, to adapt marketing and development strategies accordingly. Focus on emerging international sales trends outside of NA and EU markets.

**Sales Trends & Predictive modelling** - Multiple linear regression models show a strong predictive relationship between regional sales and global sales, with an 97% accuracy. However, heteroscedasticity and outliers must be addressed again as they skew the data quite considerably.

* Please find all insights and recommendations in the appendices.

# **Appendix 1 -** Introduction

a) - Let's apply the Five Why's approach to problem solving, the main question must revolve around the business problem.

Problem Statement: **Turtle Games needs to improve overall sales performance by understanding their customer base and sales trends.** Here I assume that there is a decline in sales that Turtle Games wants to address.

*Why is sales performance not meeting expectations?*

- There is limited understanding of customer behaviour and engagement, particularly with areas such as loyalty point accumulation.

*Why is there a limited understanding of customer behaviour?*

- Inadequate data on customer engagement and loyalty points, hindering effective segmentation and targeted marketing.

*Why is there inadequate data on customer engagement?*

- Possibly fragmented or outdated data collection systems, and underutilisation of text data like social media reviews.

*Why are data collection systems fragmented?*

- Historical prioritisation of other areas resulting in a company that doesn't prioritise data driven solutions.

Why has is there not a focus on data driven solutions.

- Our team has been brought in to address this.

By addressing the root causes found through the Five Whys analysis, Turtle Games can implement targeted strategies to improve their understanding of customer trends and preferences, leading to more effective sales performance and business growth.

b) - SWOT analysis

| | | |
|---|---|---|
| **Strengths** | Strong brand recognition and reputation in the gaming industry. | Diverse portfolio of innovative gaming products catering to various demographics. |
| **Weaknesses** | Limited understanding of customer behaviour, data and engagement. | Reliance on traditional marketing strategies with limited personalisation or awareness of customer segments. |

| | | |
|---|---|---|
| **Opportunities** | Utilising customer trends to drive sales and enhance customer loyalty. | Community engagement with a process of reading and analysing reviews already in place. |
| **Threats** | Competition from established gaming companies and indie developers in an already saturated market. | Vulnerability to disruption in the gaming industry due to an inability to adapt quickly to changing consumer preferences and market dynamics driven by data insights. |

## c) - Python

Import and sense checking routine - Below is a summary of this process please consult the attached notebook for discussions on the output of these codes.

**Importing and sense checking**

- First I import the data using *read_csv()* and examine the top of the data frame for immediate comprehension

```
# Import the data set.
reviews_raw = pd.read_csv

# View the DataFrame.
reviews_raw.head()
```

- After importing the .cvs into the notebook I explicitly check for missing values using *isna()* followed by sense checking data types and column names using *info()*

```
# Check for missing values.
reviews_raw.isna().sum()
```

```
# Explore the data.
reviews_raw.info()
```

- Next I consider basic descriptive statistics of both numeric and categorical data using *describe()* in order to get a sense of the size of the data

```
# describe the numerical data with basic statistics
reviews_raw.describe().round(2) # round() to limit clutter
```

```
# Describe the descriptive data with basic statistics
reviews_raw.describe(include=['O'])
```

- Before cleaning I will visualise, with box plots for example, any data that immediately warrants it before cleaning.

**Cleaning the data**

- In this project, with no missing values, there is only basic cleaning to do. The data is well kept overall.
- I create a new data frame keeping only the relevant columns by using the *drop()* command

```
# Drop unnecessary columns and create new df 'reviews' to preseve original
reviews = reviews_raw.drop(columns=['language', 'platform'])
```

- Whilst then renaming certain columns with *rename()* to avoid any potential confusion and finally sense check the new data frame to ensure the changes have gone through

```
# Rename the column headers to avoid confusion with parenthesis
reviews = reviews.rename(columns={'remuneration (k£)': 'total_income',
                                  'spending_score (1-100)': 'spending_score',
                                  'product': 'product_code'})

# View column names.
reviews.info()
```

- Before saving the dataframe I check for duplicates using *duplicated()*. Something I will do again later in the process as you will see when we use natural language processing.

```
# Checking for duplicates
duplicate_rows = reviews[reviews.duplicated()]

# Display duplicate rows
print("Duplicate rows:")
print(duplicate_rows)

# Check the number of duplicate rows
print("Number of duplicate rows:", len(duplicate rows))
```

- Lastly I save the data using *to_csv()* to create a hard copy back up

```
# Create a CSV file as output.
reviews.to_csv('reviews.csv', index=False)
```

## d) R - Importing and sense checking - Below is a summary of this process please consult the attached notebook for discussions on the output of these codes.

- We install R libraries as we go through the cod but initially I include our main cleaning library of tidyverse and g*g*plot2 for initial visualisations.

```
# Install and import packages
library('tidyverse')
library('ggplot2')
```

- The data is then imported, and a new data frame is created for cleaning. The new dataframe is then viewed.

```
# Import the dataset.
sales_raw <- read.csv(file.choose(), header=T)

# Print the dataframe.
head(sales_raw)
```

**Cleaning the data**

- First we make a subset of the relevant columns by dropping irrelevant ones.

```
# Create a new dataframe from a subset of the sales dataframe.
# Remove unnecessary columns, keeping only Turtle Game relevant ones.
sales_clean <- select(sales_raw, -Ranking, -Year, -Genre, -Publisher)
```

- Then, by checking for missing values and duplicates, we can minimise the chance of inaccurate data and therefore inaccurate insights.

```
# Check for missing values.
sapply(sales_clean, function(x) sum(is.na(x))) # Counts NAs for each column.
# There is potential to remove the observations with missing values.
# However, we will remove duplicate rows if there are any.
sales_clean <- unique(sales_clean)
```

- We view the descriptive statistics for the data to get an overall impression of the shape and distribution of the data.

```
# View the descriptive statistics and check the data types are accurate.
summary(sales_clean)
```

- R has a very useful package which provides a comprehensive report on statistics, this is optional, and a lot of this analysis is done manually in R Studio.

```
# (Optional) Using DataExplorer to generate a report.
# install.packages("DataExplorer")
# library(DataExplorer)
# create_report(sales_clean) # Remove '#' for the report
```

- Finally, an option to download the file for separate viewing is supplied as requested by the Turtle Games sales team.

```
# Optional download of cleaned sales .csv
write.csv(sales_clean, file = "sales_clean.csv", row.names = FALSE)
```

# Appendix 2 - Customer Insights: **Linear regression**

**# Customer Insights: Linear regression to examine loyalty_points**

Turtle games is primarily concerned with loyalty points at this stage and the main question is **How do customers engage with and accumulate loyalty points?** "*Loyal customers can be an economic source of income for companies as well as affecting people around them with their advice and incentives, enabling the business to gain new customers at less cost. Therefore, establishing customer loyalty provides a great competitive advantage, customer losses are prevented, revenue streams are secured and new customers.*
*are easier to reach.[1]"* Kahraman 2020

For this analysis we will examine the relationships between the remaining variables to the main column in question, loyalty_points with linear regression.

Initial exploration and descriptive statistics involved importing necessary libraries such as NumPy, Pandas, Matplotlib, Seaborn, and Statsmodels, along with performing sense checking on the review's dataset.

The dataset contained 2000 entries with no missing values or duplicates across 11 columns, including both numerical and categorical data types.

Basic statistics, such as mean, median, quartiles, and standard deviation, were computed for numerical columns like age, total income, spending score, loyalty points, and product codes.

Exploratory visualisations, including box plots for numerical variables and a summary of categorical data, were generated to gain insights into the distribution and characteristics of the dataset.

[1] Arslan, I Kahraman. "THE IMPORTANCE of CREATING CUSTOMER LOYALTY in ACHIEVING SUSTAINABLE COMPETITIVE ADVANTAGE - ProQuest." Www.proquest.com/Openview/B713fccb5e654cec5a2a678a93b67e89/1?Pq-Origsite=Gscholar&Cbl=4371414 , Aug. 2020

By printing a histogram, we can see similar data to what the exploratory box plot showed us, namely that some customers have extremely high points, but most are towards the lower end of the spectrum as the graph skewed to the left.

The relevant numeric variables to explore are spending_score, total_income and age so a pairplot and heatmap is used to examine any immediate visual relationship. Both results show us that spending_score and total_income are likely to influence loyalty_points but that age won't.

- One important thing to note here in this early exploratory visualisation is that spending_score and total_income do not correlate well together, they might not suffer from Multicollinearity. This is an early indication that they could prove useful when used together in any multiple regression models.

**Simple linear regression**

The first exploration is using simple linear regression regarding **spending_score** on the x axis and loyalty_points on the y axis. Loyalty_points will always be the dependant variable for this analysis as it is the variable I am trying to predict.

- While a statistically significant P Value result is good the R squared shows us that only 45% of the model's variation can be explained by the two variables. This is not enough to be reliable so additional variables should be considered.

The second exploration concerns **total_income** and loyalty points.

- Again, a strong P Value renders the model statistically significant but a weak R Squared of 38% hinders the model's practical usage. The addition of more variables feels like the most practical solution.

The third and final variable **Age** proves to have a nonlinear relationship, which we saw earlier in the pairplot and heatmap. Coupled with the fact that only 2% R Squared suggests that hardly any of the model's predictions will be valuable it is decided to no longer continue with the analysis.

**Multiple regression analysis**

Having seen that more variables are needed to achieve a higher R Squared score multiple linear regression is employed using the two variables that showed promise. As such total_income, spending_score and loyalty_points will be plotted to explore any potential multiple relationship.

Having checked for multicollinearity and heteroscedasticity we find the model to be serviceable with an R Squared of %83.

VIF values around 1 suggesting variables are not correlated and multicollinearity does not exist in the regression model.

- comparing this result to the heatmap of correlation at the start shows a 0.0056 correlation between total_income and spending_score. So, this is expected.

However, heteroscedasticity does seem to exist because of such a low LM test P Value being effectively zero.

- This is not a surprise as we saw a similar pattern in the initial visual explorations of out linear regression tests as seen below.

The statistical measures of Mean Absolute Error and Mean Squared Error provide a reliable source of information on the model's accuracy.

> Mean of loyalty_points: 1578 - Mean Absolute Error: 401

> Standard Deviation of loyalty_points: 1283 - Square Root of the Mean Square Error (RMSE): 530

- The errors are not insignificant but as the data is in many thousands the MAE and MSE do not ruin it by any means.
- Overall, the model performs well with manageable errors compared to the dataset's variability. However, there is still room for improvement and the model would benefit from more accurate data.

Finally, we visualise the accuracy and functionality of the model by plotting an actual vs predicted loyalty points scatter plot. This allows us to see that while many of the points in the middle cluster well along the line, indicating good predictions, the head and tail of the points get further away and therefore get less accurate.

Visualising the residuals provides a good insight into the performance of the model. While it is not a perfect bell curve, it is encouraging to see that peak is almost perfectly centred around 0, suggesting most of our predictions are on average correct. The left tail shows a decent descent where the right tail starts to falter, this shows us that the model tends to overestimate the loyalty points and the range isn't as close to zero as one would like.

Then a tool is made so that stakeholders can input their own data into the model to test it.

**The key takeaways from our regression analysis**

**Insights:**

- As we saw in the heatmap and the pairplot right at the start of the analysis, there are only two meaningful variables when it comes to exploring loyalty_points. These are spending_score and total_income. These variables have a positive correlation, offering us the basic conclusion that if you have more total_income and spending_score this will be reflected in higher loyalty_points. This suggests that the other variables can be dropped when it comes to exploring further predictions with loyalty points.
- Individually however, the R Squared of spending_score (45%) and total_income(38%) to loyalty_points were too modest and would not have provided a good working model. By combing the two together in a multiple regression analysis we have a much healthier R Squared of 83%. Meaning that 83% of the observed variation can be explained by the model's inputs.

**Recommendations:**

- *Data Issues*: High MAE, MSE the presence of heteroscedasticity are legitimate issues that need to be addressed. Rather than adding new variables to the model Turtle games should focus on increasing the accuracy and size of the variables that do already contribute to the model. Namely, focus on spending_score, total_income and of course loyalty points. Better data will lead to better models.
- *Customer Retention:* By identifying the factors that lead to higher loyalty points, Turtle Games can create strategies to improve customer retention. They could offer special promotions or loyalty reward thresholds, to customers who are at risk of churning but have high spending scores or incomes.
- *Predictive Power:* The model itself can be used for predictive analytics, helping Turtle Games forecast future loyalty points accumulation based on expected changes in customer income and spending. Especially relevant as we see more global cost of living crises appear.

Overall, the tool allows for a sophisticated approach to marketing that is proactive rather than reactive, potentially leading to higher customer engagement, better customer retention, and a healthier bottom line for The Turtle Games.

# **Appendix 3 -** Customer Segmentation: **K Means clustering**

**# Customer Segmentation: K Means clustering**

The main question is **How customers can be segmented into groups, and which groups can be targeted by the marketing department?** As a paper on K Means argues, "*The partnership between businesses and consumers is increasingly crucial with technological growth. It is necessary to manage this relationship for the company's future growth* [2]" Nandapala 2020

I believe spending score to be the most useful variable here because we're looking for customers that contribute the most to purchasing within the company. As such we will mostly look for groups around this variable. The method we will use is K Means clustering which sorts data into distinct groups by finding common patterns, helping to identify natural clusters in the data, much like organising items into separate bins based on their similarities..

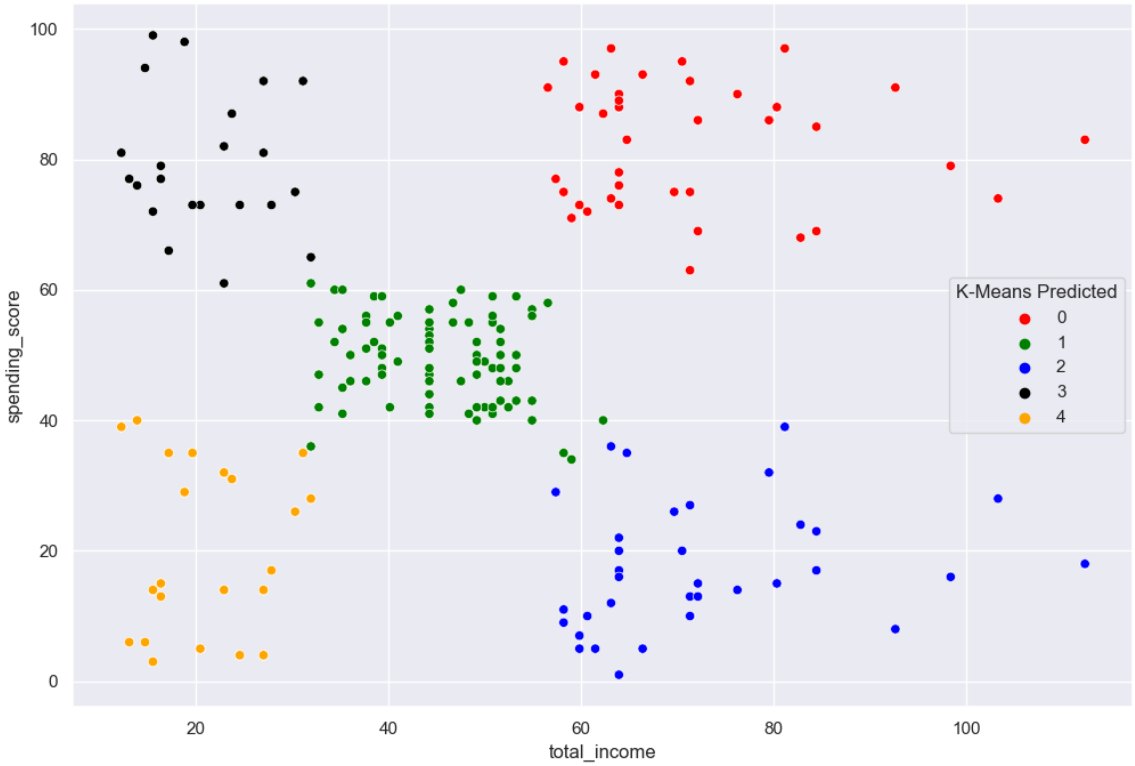
Module 1 showed us that total_income, spending_score and loyalty_points were all linked so we will bear that in mind.

The clustering process is linear and logical so please view the notebook for the detailed process. Insights and particular further explanations will be shown here.

**Spending Score vs. Income:** This comparison will help identify spending patterns across different income levels, indicating which income groups contribute the most to purchasing within the company. It can reveal segments of high-spending, high-income customers as well as other income segments' spending behaviours.

The elbow and silhouette methods both confirm that clustering of 5 would be most appropriate.

[2] Nandapala, E.Y.L , and K.P.N Jayasena. "The Practical Approach in Customers Segmentation by Using the K-Means Algorithm." https://Ieeexplore.ieee.org/Abstract/Document/9342639 , IEEE, 2020 . Accessed 2 Mar. 2024.




**Group 1 - 774 customers - Central Market Segment (39%)**

- *Summary:* Dominates the middle range for both income and spending, making it the core segment.
- *Key Insight:* Represents the average customer profile with balanced income and spending.
- *Key Recommendation:* Continue to engage this segment with diverse product offerings and loyalty programs that resonate with their average spending and income levels.

**Group 0 - 356 customers - High Income, High Spenders (18%)**

- *Summary:* High-value customers with both high income and high spending scores.
- *Key Insight:* Despite being smaller in number, their high engagement levels make them critical for revenue.
- *Key Recommendation:* Target with exclusive offers, premium services, and personalised marketing to enhance their experience and loyalty.

**Group 2 - 330 customers - Untapped Potential (17%)**

- *Summary:* Higher income but not maximising spending potential, indicating room for growth.
- *Key Insight:* This segment has the means but lacks the motivation to spend more.
- *Key Recommendation:* Identify barriers to higher spending and introduce targeted incentives to convert their potential into actual spending.

**Group 4 - 271 customers - Low Priority Segment (14%)**

- *Summary:* The least potential among the groups, contributing the smallest portion to total observations.
- *Key Insight:* Shows minimal engagement and spending.
- *Key Recommendation:* Offer occasional, value-focused promotions to gently boost their spending without significant resource allocation.

**Group 3 - 269 customers - Loyal Low Earners (13%)**

- *Summary:* Exhibits loyalty through spending despite lower income levels, suggesting value-driven purchasing behaviour.

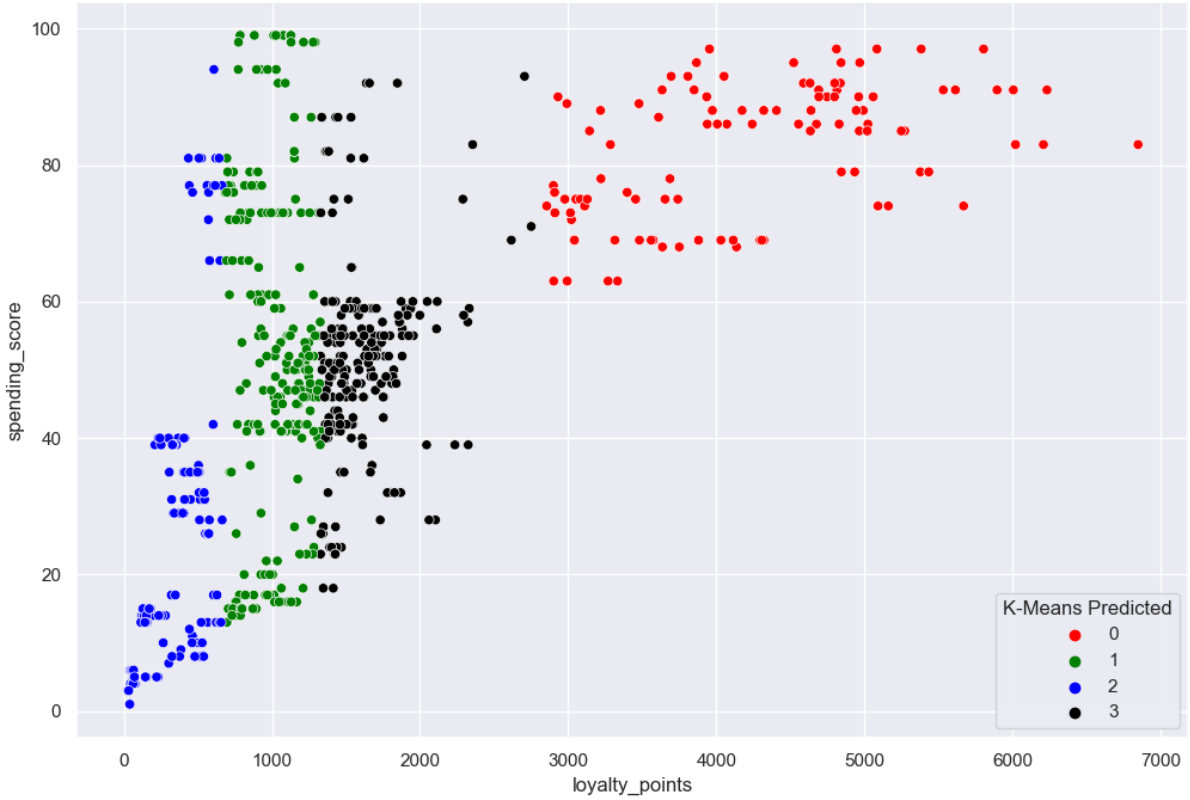
- *Key Insight:* Their loyalty, potentially influenced by age or other demographic factors, suggests long-term customer value as their financial situation improves.
- *Key Recommendation:* Foster this loyalty with rewards and recognition programs, ensuring they feel valued and have reasons to increase spending if their income grows.

**Age vs total income:** This relationship has the potential to show us patterns in any income-driven age segments, life stages, and purchasing power. Which feels partially relevant in industry where games need time and money to be played. However, age persists in being nonlinear so the potential for it to be an unhelpful variable remains.

As we saw with the attempted regression to plot the relationship between age and loyalty_points, there doesn't seem to be any relationship between these two variables. As such, this line of questioning will be discounted.

**Loyalty Points vs. Spending Score:** Of course, we expect to see a direct correlation between these two variables, but it would be interesting to see how K means segments these customers exactly. Understanding this link can help prioritise marketing efforts towards retaining and rewarding loyal customers who contribute significantly to purchasing.

The elbow and silhouette methods here suggest that 4 clusters are most appropriate.




This analysis confirms the basic principle which we expect to be true that if you spend more, then you are more loyal. However, you only seem to be able to achieve a loyalty_points score of > 2700 if you have a spending score of > 60.

**Group 0 - 345 customer - High Loyalty, High Spending (17%)**

- *Summary:* Smallest yet high-quality customer segment with high loyalty and spending.
- *Key Insight:* Integral for retention and premium market strategies.
- *Key Recommendation:* Develop exclusive loyalty rewards and upscale product lines for this group.

**Group 1 - 638 customers - Moderate Loyalty and Spending (32%)**

- *Summary:* Largest customer segment with moderate loyalty points and spending scores.
- *Key Insight:* Holds potential for increased brand loyalty and spending.
- *Key Recommendation:* Engage with brand loyalty programs and cross-selling tactics to elevate their spending.

**Group 2 - 423 customers - Low Loyalty, Varied Spending (21%)**

- *Summary:* Second smallest group, indicative of newer or less engaged customers with low loyalty scores.
- *Key Insight:* Represents an opportunity to boost engagement and purchase frequency.
- *Key Recommendation:* Offer loyalty incentives and marketing to encourage return visits and higher engagement.

**Group 3 - 594 customers - Loyal but Lower Spending (30%)**

- *Summary:* Large customer base with high loyalty yet lower spending.
- Key Insight: May represent budget-conscious customers with potential for increased purchase frequency.
- *Key Recommendation:* Provide value-driven offers, bundle deals to encourage more frequent purchases.

**Education vs. Spending Score:** Examining spending behaviour across different education segments can provide insights into the preferences and purchasing patterns of customers with varying educational backgrounds. This information can guide targeted marketing campaigns tailored to specific education segments.

There are clearly clusters to be found here and our initial analysis into how customer's education levels relate to their spending habits has shown promising results, indicating potential patterns worth further exploration. However, we currently face challenges in visually presenting these findings due to the complexity of the data, which limits our ability to draw immediate, clear conclusions.

There are alternatives such as ordinal encoding which can be explored in the future but there will always be a problem of labelling those clusters. For instance, if ordinal encoding found 3 clusters as well, how would we then name those clusters? If one cluster has half its users from basic and half from graduate, that probably wouldn't be useful and as such this line of clustering is discounted.

**Insights:**

- High loyalty and spending scores are strongly correlated, indicating that engaged customers significantly contribute to revenue. This correlates with our earlier regression analysis of the same data.
- K-Means clustering effectively segments customers, enabling targeted marketing and product development strategies. Please revise each segment individually for specific insights and recommendations. Some examples include:
    - There exists a segment with higher income but lower spending, suggesting untapped potential for increased revenue.
    - Lower income segments show considerable loyalty through spending, highlighting the value of targeting these groups with specific strategies.

**Recommendations:**

- Use customer segments identified through K-Means clustering for more refined, data-driven marketing and product development efforts. Please revise each segment individually for specific insights and recommendations. Some examples include:
    - Develop exclusive loyalty programs and upscale product lines for high-value customers to enhance their experience and loyalty.

- Identify barriers to increased spending in higher income segments and design targeted incentives to unlock potential revenue.
- Foster loyalty among lower income customers with rewards and recognition programs, encouraging them to increase their spending if their financial situation improves.
- Continue to research to relationship between education levels and spending habits to find a way to make the data more accessible.

# **Appendix 4 -** Customer Feedback: **NLP**

The main question is **how can text data (e.g. social data such as customer reviews) be used to inform marketing campaigns and make improvements to the business? "***Through customer review mining, businesses can better understand their customers' sentiments, preferences, and pain points. By identifying patterns and trends in the reviews, companies can make informed decisions to enhance their products, services, and overall customer experience.[3]"* Kobiruo 2023

To consider language analysis we will use Natural Language Processing libraries such as TextBlob and VADER to get a balanced view of sentiment analysis. (NLP) uses machine learning to reveal the structure and meaning of text so it is a particularly useful tool in sentiment analysis.

- This exploration includes essential NLP tasks such as text classification, sentiment analysis, and pattern recognition, providing insights into the structure and meaning of language data.

For this work we will be deviating from the columns we used previously and considering the columns 'review' and 'summary' of the dataframe. A review being the customer's review of the product and the summary being its summary.

—

Cleaning is essential to all data work but in particular NLP is sensitive to incorrect data or non-formatted data. We create our new dataframe highlighting the two key columns of reviews and spending.

A concise explanation here - "*When we normalize text, we attempt to reduce its randomness, bringing it closer to a predefined "standard". This helps us to reduce the amount of different information that the computer has to deal with, and therefore improves efficiency [4]*" Diego Lopez 2021

By making all the words lower case, removing punctuation and dropping duplicates we give our model the best possible chance at reading human sentiment correctly.

---

[3] Otebele, Kobiruo . "From Insights to Conversions: The Power of Customer Review Mining in Conversion Optimization - Invesp." Invesp, 2 Aug. 2023, www.invespcro.com/blog/customer-review-mining /. Accessed 2 Mar. 2024.

[4] Yse, Diego Lopez. "Text Normalization for Natural Language Processing (NLP)." Medium, 12 Mar. 2021, www.towardsdatascience.com/text-normalization-for-natural-language-processing-nlp-70a314bfa646. Accessed 2 Mar. 2024.

—

Then we want to tokenise our words. Tokenisation aims to break down the text into smaller elements facilitating easier and faster processing.

At this early stage, plotting word clouds of the review and summary columns gives us a quick indication of what the overall sentiment might be.

- From our initial review there seems to be an overall positive sentiment.

—-

Turtle games has asked for the most frequently used positive and negative words from both columns. So, we continue to clean the data in order to move from a broader view into a more precise one.

This process involves removing alphanumeric characters, stop words, non english words and lemmatising the data.

- Lemmatizing is the process of converting words to their base or root form, simplifying variations of words to an ordinary form to improve text analysis.
    - For example, lemmatizing the words "running," "ran," and "runs" would convert them all to their root form "run."

- The importance of normalisation is highlighted here. "*Normalisation aims to change the textual data into a comparable and uniform format. For example, ensure that all characters are lower case, remove numbers or convert to textual representation, remove punctuation, strip white space, and remove general English stop words. Normalisation processes are based on a detailed and nuanced understanding of grammatical rules and norms*[5]." LinkedIn community 2023

With our data cleaned completely, normalised and ready for NLP we can begin the sentiment analysis. We begin by creating a word cloud of all the words from both columns and it is useful to see that the sentiment remains positive. It does.

Let's continue to do a deeper dive into the specifics of which words are most common whilst considering polarity, how people feel about the products. We use the data frame which contains all the tokens from both columns for the broadest possible overview.

The most frequent 15 words shows us the following insights:

- **Game Enjoyment**: The prominence of words like "*fun*," "*love*," "*like*," "*good*," and "*great*" suggests customers have a positive experience with the games. These terms indicate a strong satisfaction with the entertainment value.
- **Engagement Level**: The word "*play*" being one of the most frequent indicates that the act of playing is a central theme in the feedback. This reflects active engagement with the games.
- **Repetitive Excellence**: The word "*game*" appears the most often, confirming that the feedback is focused on the gameplay experience. The high occurrence of "*time*" shows a repetitive acknowledgment of quality or enjoyment.

—

Now we have explored the data in a broader sense and have a more detailed view of the seemingly positive feeling we can start to use NLP for more precise sentiment analysis.

[5] LinkedIn community. “What Are the Best Practices for Cleaning Data for Natural Language Processing?” Www.linkedin.com, 26 Oct. 2023, www.linkedin.com/advice/0/what-best-practices-cleaning-data-natural-language . Accessed 2 Mar. 2024.

There are many useful libraries and ways of conducting sentiment analysis for our purposes we will consider two. TextBlob and VADER. Using both TextBlob and VADER provides a more comprehensive sentiment analysis, as it combines TextBlob's strengths in handling general texts with VADER's expertise in detecting nuances in social media language, ensuring a broader and more accurate understanding of sentiments across different text sources.

First, **TextBlob** sentiment and subjectivity analysis. Using the same code as before, we begin to create a centralised data frame containing all the relevant columns for our analysis. This centralisation will come in especially useful when we need to reference and compare TextBlob and VADER later on.

Plotting TextBlob histograms allows us to see a quick and easy visualisation that again suggests that sentiment is positive for both review and summaries.

- Interestingly, when testing for subjectivity, we see TextBlob suggest that there are a large number of negative summaries that it detects as objective. This could be because when writing summaries of reviews, users write shorter and more direct sentences and TextBlob believe that to present as objective.

Now we will do the same analysis but using **VADER.**

- TextBlob is primarily used for simple, rule-based natural language tasks and sentiment analysis, while VADER specialises in analysing sentiments in social media texts by considering slang, emoticons, and intensifiers. So, it will be an interesting comparison.

VADER does not perform a subjectivity analysis like TextBlob, but we are able to plot corresponding histograms of review and summary sentiment analysis. We do this using the Compound Score from VADER, which is an overall sentiment score.

VADER see’s the review sentiment more positively than TextBlob as evidenced by the next graphic, showing all of the sentiment scores side by side.

- Review VADER being so positive perhaps shows the difference in technique compared to TextBlob when it comes to language analysis. With VADER specialising in analysing sentiments in social media perhaps we are seeing it understand the subtleties of the review better than TextBlob.

—

Next we create a dataframe to include all relevant sentiment analysis columns created thus far. This allows us to easily draw what we need to.
Here's what each column represents:

1. *index*: The index of the review entry.
2. *review*: The original text of the review.
3. *review*_normalised: The review text after normalisation, which may include lowercasing, lemmatization, and removal of stop words.
4. *review_tokens_normalised*: The review text tokenized and normalised.
5. *review_polarity*: The polarity score of the review, indicating the sentiment polarity (positive, negative, or neutral) using TextBlob.
6. *review_vader_compound*: Just the overall 'compound' sentiment result.
7. *review_vader*: The sentiment analysis results of the review using Vader, providing scores for negative, neutral, and positive sentiment, as well as a compound score.
8. *review_subjectivity*: The subjectivity score of the review, indicating how subjective or objective the text is.
9. *summary*: The original text of the summary.
10. *summary_normalised*: The summary text after normalisation, similar to review_normalised.
11. *summary_tokens_normalised*: The summary text tokenized and normalised.
12. *summary_polarity*: The polarity score of the summary, using TextBlob.
13. *summary_vader_compound*: Just the overall 'compound' sentiment result

14. *summary_vader*: The sentiment analysis results of the summary using VADER.
15. *summary_subjectivity*: The subjectivity score of the summary.

—-

Then we consider the top 20 positive/negative reviews/summaries from both language models.

This analysis is fairly straightforward with the exception of VADER. For the positive VADER reviews the text has to be truncated as the most positive reviews are exceptionally long. VADER works well with social media in particular, so it seems to have worked better with longer pieces of text than TextBlob.

- To get around this the user is able to download a .csv of the most positive reviews to view on a program of their choosing.
- Summaries of each of the models and their insights are provided in the notebook.

—

**Main Insights from the individual TextBlob and VADER Analysis:**

- **Design and Complexity**: Customers have expressed that some games are either too intricate or poorly designed, which complicates usability and diminishes the user experience.
- **Quality Assurance**: There is a notable issue with the quality of products, suggesting inconsistencies with customer expectations and the actual standards of the items received.
- **Value Perception**: A disconnect between product pricing and perceived value has been identified, leading to customer dissatisfaction.
- **Educational Impact**: Products with educational value receive positive feedback, indicating success in this particular aspect.
- **Customer Experience Gap**: Negative sentiments such as "boring" and "disappointing" highlight a gap between the marketed product experience and the reality.
- **Creative Engagement**: Products that support creative engagement, particularly those that involve crafting and have tangible high-quality results, are well-regarded.
- **Design Functionality vs. Aesthetics**: A preference for functional design over aesthetic elements suggests that customers prioritise the usability and playability of games.
- **Replayability and Engagement**: Customers favour games that offer varied and engaging content that can be enjoyed over multiple play sessions.

**Recommendations**:

- **Enhance Usability**: Simplify game mechanics where possible and ensure that the complexity level is appropriate for the target audience.
- **Improve Quality Control**: Implement stricter quality checks to ensure products meet a high standard that aligns with customer expectations.
- **Reassess Pricing Strategy**: Evaluate product pricing to ensure it reflects the value and quality of the games, minimising the gap between customer expectations and reality.
- **Expand Educational Offerings**: Capitalise on the positive reception of educational products by expanding this range.
- **Promote Craft, Creativity and Education**: Continue to develop products that encourage creativity, offering clear instructions and high-quality materials.
- **Prioritise Functional Design**: Revise product designs to focus more on functionality and interactive elements, potentially reducing the emphasis on non-essential artwork.
- **Adjust Game Complexity**: Provide a range of complexity within games or offer adjustable difficulty levels to cater to a broader audience.
- **Innovate for Replayability**: Introduce dynamic content, expansion packs, or updates to existing games to maintain engagement and encourage repeated play.

---

General sentiment distribution visualisations

First we must extract the negative, neutral, positive and compound summary scores from VADER for both the review and the summary columns respectively. This allows us to do direct comparisons to TextBlob while treating them as separate variables.

- For instance, we can then explore review length vs compound score which shows us that longer reviews tend to be more positive according to the more sensitive model of VADER.
- The same is mostly true for summary however worth noting is a large spike in the neutral zone which could suggest the neutral way in which consumers tend to summarise their reviews. This correlates with summaries as being seen as more objective by TextBlob.

Then we create simple word clouds of the three VADER classification options of positive, negative and neutral words to see VADER's interpretation.

Visualising a boxplot of compound sentiment distribution allows us to see a new visual representation of data that we have already explored. It is interesting however just to note how positive the inter quartile range of the reviews is and how almost all the negative reviews are treated as outliers.

Finally, there is a basic pie chart of sentiment of the boxplot above exploring VADER's interpretation of sentiment distribution between the two columns of review and summaries. Those outliers that we saw earlier on the negative end of the review compound score only amount to 6.2% of the total reviews.

---

**Main Insights from the entire NLP analysis:**

- *Positive Sentiment Prevalence*: Analysis reveals a generally positive sentiment in customer reviews and summaries, with frequent positive words like "*fun*," "*love*," "*good*," "*great*," and "*play*" are among the most mentioned.
- *Necessity of Data Cleaning*: Proper cleaning and normalisation of text data are crucial for accurate sentiment analysis, enhancing the model's ability to interpret human sentiment correctly.
- *Use of NLP Tools*: Utilising both TextBlob and VADER for sentiment analysis provides a broad and nuanced understanding of sentiments, catering to general texts and reviews specifically.
- *Potential Challenges in Language Diversity*: Current NLP systems have limitations in handling multiple languages and long contexts, which can be significant for businesses with a diverse international customer base such as Turtle Games.
    - As explored well here "*They* (NLP models) *are also not very good at tasks that require reasoning, common sense, grounding, and physical intuition. Existing systems are too English centric and their performance tends to degrade in other languages – this is particularly relevant in Europe which is highly multilingual (e.g., there are 24 official languages in the EU). Training and running these models are computationally very demanding and the current transformer-based architectures do not scale very well to long contexts*.[6]" Williams 2024
    - We saw that with our own data when TextBlob's 15th best summary which was "only ok at best". A strong mix of neutral and positive words on their own but a phrase that's clearly negative to a natural English speaker.

[6] Williams, Jonathan. "Challenges in Natural Language Processing Require Coordination across a Large Scientific Network." European Lab for Learning & Intelligent Systems, 30 Jan. 2024, www.ellis.eu/news/challenges-in-natural-language-processing-require-coordination-across-a-large-scientific-network . Accessed 2 Mar. 2024.

- Which also leads to the next point, since Turtle Games is an international company there will need to be some consideration towards non-English-speaking customers and their non-English sentiments.

**Recommendations:**

- *Leverage Positive Feedback:* Potentially emphasise the frequently mentioned positive words in marketing campaigns to highlight what customers appreciate most about the products.
- *Address neutral and positive Sentiments:* Analyse the context of neutral and positive sentiments to identify areas for improvement and address specific customer concerns in product development and marketing strategies.
- *Customer feedback analysis:* Reach out to customers who have submitted negative reviews and summaries and start conversations which could lead to constructive conclusions. The longer the review the more positive it tends to be, so you need to be proactive in start conversations with negative review/summary submitters.
- *Invest in Multilingual NLP:* Given the international presence, invest in or develop NLP tools that can accurately analyse sentiment in multiple languages to better understand and cater to the diverse customer base.

# **Appendix 5 –** Sales data: **Platform Performance**

**Cleaning and manipulating data using R.**

We are asked to explore the impact on sales per product_id. Grouping data by product ID proved problematic two main reasons: First, product IDs tend to be arbitrarily assigned numbers which show no real promise when comparing them to other variables and we currently don't know which ID refers to which game. Secondly there proved to be no meaningful relationship between ProductID and sales data.

Platform however proved to be a useful variable and therefore shall be used to explore the sales data in the notebook. The new question is **What is the impact on sales per Platform**?

This is crucial because "*gaming on multiple platforms is now more common too, with 68% of all console gamers playing games on 3 or more devices as of Q2 2020*.[7]" Morris 2020 So a detailed understanding of console trends is vital to capturing as much market share as possible.

In a separate section at the bottom of the R notebook users can find details of the issues that arise when trying to use Product as a main variable.

We deal with the large amount of platform options by installing Plotly. This enables plots to be interactive and display the many platforms neatly.

Visualising the relationship between the sales data is similar to before but now, with Platform included, we have a whole new dimension to explore.

We begin by creating a new dataframe with the min, mean and max of sales data as requested by Turtle Games.

[7] Morris, Tom. "Everything to Know about the Console War." GWI, 12 Nov. 2020, www.blog.gwi.com/trends/games-console-war .

Then we create a new dataframe from our clean data grouping the sum of the regional and global sales data by platform. Now let's explore the visualisations of that data.

**Scatter plots:**

- Plotting regional sales against global sales and colouring in the points allows us to see that distribution more effectively.
- The plot is interactive as well so as to navigate the many platforms without cluttering the screen.
- Importantly, we calculate the total percentage that each regional platform makes up of the global sales.
    - For instance, NA's bestselling platform is the Xbox360. The Xbox360 is responsible for around 60% of all global sales but in the EU it is half that, around 30% of global total sales.
    - The NA plot shows us that platforms tend to have 3 main groups.
        - A tight cluster below 50 million sales
        - A small breakout group between 50 and 100 million sales
        - And the outliers above 150
    - The EU doesn't tend to follow this pattern, instead it has a main cluster below 50 million sales and then a small group of outliers above 60 million.

**Barplot:**

- A barplot allows us to have a macro level view of the data so we can see overall sales. This is useful to get a broader overview of sales data by platform.
- The Wii has the highest total global sales among all platforms, represented by a tall pink bar.
- The Xbox 360 and PS3 follow, with significant sales, shown by large orange and blue bars, respectively.
- The DS, a handheld console, also shows high sales, with a cyan bar.

**Histograms:**

- Boxplots show us the same information again but really help to further highlight the distributions without our plots.
- All boxplots show that right skew with the medians appearing towards the bottom of the box.
    - NA shows very clearly those two high outliers and a broad IQR. This is similar to the Global sales suggesting a large variability of sales within that IQR. EU's IQR is a bit narrow which in turn might indicate less variability.
    - This variability could be explained by population differences between the regions. I would be interesting to know the relative differences in the size of customer base for the regions.

**Normality testing**

When data is normally distributed, it simplifies analysis and interpretation because many statistical methods are designed with this assumption in mind.

QQ plots help to show us a visual representation of normality in the dataset by plotting the quantiles of the data against the quantiles of normal distribution. We're essentially testing to see if it fits an expected pattern. Given the tests we have done so far and the outliers we have seen, it seems unlikely.

- NA shows us that most of the plot is normally distributed. However, there are significant outliers at the top and the tail of the plot suggesting that it sways from normality at both ends.

- EU does the same except that the line is less steep, which might indicate the variance is less than NA, that the values of NA or more extreme.
- Global sales show the same pattern as we have seen previously too.

In order to perform a Shapiro-Wilk test we first install moments

- The Shapiro-Wilk test is another test of normality using null hypothesis and the statistical measure of P-Values. Essentially we're looking to see if the distribution follows a bell curve pattern of normal distribution by offering the null hypothesis of **there is no difference between your distribution and a normal distribution.**
- In this particular case, we are looking for a high P Value of above 0.05 in order to not reject the null hypothesis that the data is distributed normally.
- We are also looking for a W value to be as close to 1 as possible, which also indicates normality.

**NA results** - W = 0.78271, P-value = 0.0002716,

- The W value is promising at 0.78.
- However, the P value is well below 0.05 suggesting that we reject the null hypothesis and say that the data is not normally distributed.
- This is not a surprise given previous results.

**EU results** - W = 0.78087, P-value = 0.0002549

- Remarkably similar to the NA results, we must reject a normal distribution.

**Global results** - W = 0.82775, P-value = 0.001406

- Similar again, even though the P Value is indeed higher than the regional P Values it is still below the threshold of 0.05.

**Skewness and Kurtosis are useful measures of distribution too.**

- Skewness and kurtosis are statistical tools that help us understand the shape of our data distribution. Skewness focuses on the symmetry of the data, showing us if the distribution leans more towards the left or right of the average. A skewness close to 0 indicates a symmetrical distribution, which is ideal for a normal distribution.
- Kurtosis, on the other hand, tells us about the peak's sharpness and the tails' thickness. A kurtosis value near 3 suggests a distribution similar to the normal distribution, with a moderate peak and tails.

**NA results** - *Skewness* = 1.543152 and *Kurtosis* = 4.569459

- These results show us that there is a right skew on the NA data. We have seen this already, most of the games sold are lower on the spectrum but there are a few exceptional outliers that push the data's average up at the end.
- The Kurtosis is far from 3 and suggests that there is a high peak to our curve and also thick tails at the end. Both at the lower and upper end, which we saw in our QQ plots as well.

**EU results** - *Skewness* = 1.415835 and *Kurtosis* = 3.825959

- The results are broadly the same but slightly less indicating that EU sales are ever so slightly more normally distributed.

**Global results** - *Skewness* = 1.262187 and *Kurtosis* = 3.536568

- Again, the results are broadly the same but slightly less indicating that Global sales are ever so slightly more normally distributed than the regional plots.

Finally, we determine correlation between the variables, this must be done through numeric variables and as such we will determine if the regional sales columns relate to each other and to the global sales data.

- NA to Global sales = **0.9588786**
- EU to Global Sales =**0.956914**
- NA to EU Sales = **0.8657527**

All sales data shows values close to one and unsurprisingly confirm that they correlate to each other. In simple terms If a platform does well in NA or EU it tends to do well everywhere else as well.

**Choosing the best plot to represent the data.**
All the plots have been useful and have each provided excellent detail and insights into the relationships between the sales columns. The best plot would be one that encapsulates all the insights we've found so far and shows an immediately visual summary of the data. A stacked barplot hits all of those criteria extremely well.

First we create a new column that discovers the remaining percentage of global sales that NA and EU don't account for.

- We put it in the sales_summary dataframe.
- Then create the interactive stacked bar chart in Plotly by assigning traces to each variable we want highlighted.
- What the plot shows us is the individual regions total percentage contributions to the overall sales. You can either use the Y axis scales or simply hover over a bar to see precise percentages.
- Notable insights are:
    - The top selling platforms globally are Wii, Xbox360 and PS3. All three have surprisingly even contributions from EU and NA to global sales. It is hard to draw further conclusions without knowing the relative customer populations of these regions, but it would be interesting to know how much of their relative populations are being used.
    - PC is dominated by EU 65%, a clear marketing preference there and signals perhaps a push to introduce or at least encourage NA to take up PC gaming.
    - Bars like the Atari 2600 need context to be understood. [8] *Wikipedia Contributors 2019* The 2600 was made by an American company and released in NA in 1977. It had a staggered release in the EU starting a year later in 1978 but it took even longer for some countries like France who saw it released in 1982. Games and consoles move quick in the industry and this lag would have impacted sales in the EU.
    - The 'Other_Global_Pct' column indicates the rest of the world's contributions to global sales which aren't from NA or EU. For most platforms, this percentage ranges from 10% to 40%, which is significant and shows that other regions play a key role in the global gaming market. "*Global player numbers will reach 3.38 billion, with emerging regions driving player growth.*[9]" Wijman 2023
    - When comparing the Nintendo platforms from older generations (NES, SNES, N64) to newer ones (Wii, WiiU), there is an apparent shift with the newer consoles showing a more balanced contribution between NA and EU, whereas the older consoles were more NA-centric.
        - There seems to be an emerging change in PlayStation preference. What starts out as a strong NA showing with the original 41% market share compared to EU's 31% eventually ends in EU dominance. The PSV only

[8] Wikipedia Contributors. "Atari 2600." Wikipedia, Wikimedia Foundation, 6 Oct. 2019, www.en.wikipedia.org/wiki/Atari_2600

[9] Wijman, Tom . "New Free Report: Explore the Global Games Market in 2023." Newzoo, 8 Aug. 2023, www.newzoo.com/resources/blog/explore-the-global-games-market-in-2023?utm_campaign=2023-08-GMRF-GGMR+2023+launch+article&utm_source=GGMR+2023+Press+Release&utm_term=GGMR+2023+Press+Release&utm_content=GGMR+2023+Press+Release . Accessed 3 Mar. 2024.

holds %13 share of sales compared to Europe's 45% and the rest of the world's 42% Interestingly NA has not picked up on the PSV in the way that the rest of the world and EU has. That warrants further investigation as PlayStation, traditionally performs well better in NA as evidenced by previous PlayStation.
    - This could indicate shifting market dynamics or a growing preference for the PlayStation platform in Europe as evidenced by this article noting the particular "*collapse*[10]" Yin-Poole 2023 for Xbox in Europe in favour of the Play Station V.

**Insights:**

- Product ID as a variable for sales analysis proves not to be useful due to its arbitrary assignment and lack of identifiable game information. However, users are able to explore the approach to ProductID at the end of the R workbook.
- Platform-based sales analysis introduces a new dimension to explore, with interactive plots enhancing visualisation.
- Histograms Scatter plots and bar plots continue to show comparable results to what we saw with the initial exploratory visualisations.
    - Most platforms have game sales under significant figures, with outliers indicating exceptional success on certain platforms like Xbox360 and Wii.
- Normality tests, including QQ plots and Shapiro-Wilk tests, confirm sales data does not follow a normal distribution, potentially due to significant outliers.
    - Specifically, skewness and kurtosis analysis indicate a right skew in the data, with a higher peak and thicker tails than a normal distribution.
- Correlation analysis confirms a strong relationship between regional and global sales, suggesting platforms successful in NA or EU tend to perform well globally.
- We choose a stacked barplot as the best visualisation method is based on its ability to show the total percentage contributions of individual regions to overall sales, offering clear insights into market dynamics and platform performance.
    - Significant insights include the balanced contributions of NA and EU to global sales for top-selling platforms, the dominant share of EU in PC gaming, and the historical and regional context influencing sales, such as the Atari 2600's staggered release affecting EU sales.
    - The analysis also notes shifting market preferences, such as the emerging dominance of PlayStation in Europe and the significant role of other regions contributing to global sales, highlighting the importance of considering all markets in strategic planning.

# **Appendix 6 –** Sales data: **Sales Trends**

**EDA using R.**

The first step is to load the appropriate libraries and packages. For this early stage of the analysis, we'll be using tidyverse for data manipulation and ggplot2 for visual exploration.

Turtle Games has asked us to focus on sales data, so we drop unnecessary columns until we have a new dataframe, which we call *sales_clean*.

[10] Yin-Poole, Wesley. "Xbox Series X and S Sales Have Collapsed in Europe." IGN, 22 Nov. 2023, www.ign.com/articles/xbox-series-x-and-s-sales-have-collapsed-in-europe . Accessed 2 Mar. 2024.

```
    Product       Platform           NA_Sales          EU_Sales       Global_Sales
 Min.   : 107   Length:352         Min.   : 0.0000   Min.   : 0.000   Min.   : 0.010
 1st Qu.:1945   Class :character   1st Qu.: 0.4775   1st Qu.: 0.390   1st Qu.: 1.115
 Median :3340   Mode  :character   Median : 1.8200   Median : 1.170   Median : 4.320
 Mean   :3607                      Mean   : 2.5160   Mean   : 1.644   Mean   : 5.335
 3rd Qu.:5436                      3rd Qu.: 3.1250   3rd Qu.: 2.160   3rd Qu.: 6.435
 Max.   :9080                      Max.   :34.0200   Max.   :23.800   Max.   :67.850
```

Descriptive statistics reveals many insights:

- The high minimum and maximums in the sales columns across the board suggest that there is a wide range of success in games sold.
- For instance, in North America the first quartile results suggest that 25% of games sold are sold for approximately £0.5 million pounds. Indeed, the third quartile result of £3.13 million shows us that there are a few highly successful games outside of the maximum. These outliers are the reason there is such a jump from the third quartile to the maximum.
- This pattern remains true for both Europe and unsurprisingly, global sales too. The dataset skews to the right. This means a small number of games have extremely high sales, while the majority have moderate to low sales.
- The dataset contains 352 entries with a broad product range suggesting many games.
- Adding the two maxes of the regional sales columns gives us approximately £58 million. This is not the full £67.85 of the global sales so there is revenue from either other countries or other items that we are not party too.
    - Even though NA and EU combined do explain around 85% of total sales it is worth noting that all insights will also only be 85% of the full explanation.
    - Perhaps in future, all countries can be included in another column.

The option to download the cleaned sales.csv is provided as well as the option to present a full report by DataExplorer.

**Initial visualisations**

Let us do some initial exploratory visual analyses to see how this might be represented.

Initial visualisations

- A **histogram** shows us what we saw from the descriptive statistics. Most games appear to have sold for less than £10 million in NA and EU.
    - Plotting the histograms on top of each other allows us to see comparable distributions. With NA in green and EU in red you can see that EU has more games that sold for less and NA has the majority of games that sold extremely well, those outliers from before.

- **Boxplots** will suffer from those outliers, so we create a series of boxplots that remove outliers in order to have a more accessible plot. Simply turn “outline = TRUE” to in
    - Even with the outliers removed there is a significant gap between the 3rd quartiles and the max for all boxplots.
    - You can also see on the Y scales for the two regional sales columns that the NA is selling much larger quantities of games.

- Finally let’s consider **scatterplots** which will examine the relationships between two variables. Given that we’re looking at sales data, let’s compare the sales columns.
    - The data fits well with our previous insights and observations with evidence of many outliers towards the right-hand side for each plot.
    - There are positive linear relationships in all plots suggesting that each sales column influences one and other. Which is to be expected. Games that sell well in EU sell well in NA and therefore sell well globally.

- The relationship between EU and global sales is partially strong linear line of best fit which suggests that the EU might be a good predictor of how games will do globally.

**Key Insights:**

- High variability in game sales, with a wide range between minimum and maximum sales across regions.
- Sales data is right-skewed, indicating a few games achieve extremely high sales while the majority have moderate to low sales.
- North American and European sales combined account for approximately 85% of total global sales, suggesting other regions also contribute significantly.
- Initial visual analysis using histograms shows most games sell for less than £10 million in North America and Europe, with North America having more high-selling outliers.
- Boxplots, even with outliers removed, reveal a significant gap between the third quartiles and the maximum, indicating the presence of remarkably successful games.
- Scatterplots demonstrate positive linear relationships between regional sales columns, suggesting that games selling well in one region tend to sell well globally.
- The relationship between European sales and global sales is particularly strong, hinting at the potential predictive value of European sales for global performance.

# Appendix 7 – Sales Data: Predictive Modelling

**Exploring relationships in the sales data**

We have already seen that our dataframe is highly correlated from previous tests but let's look at the original dataframe using a correlation heat map allows us to see it in a different visual way.

First we drop the non-numeric columns, check for missing values and remove duplicates in order to increase accuracy of the model.

- There are two rows in the column Year with missing values, as they contribute an insignificant sum to the sales column we could have removed them however, the whole column itself proves to be irrelevant so we don't need to.

As expected the sales columns correlate very highly with one and other and interestingly there are no significant relationships between year and ranking. So, we drop them from our dataframe leaving only the original, cleaned sales data. However further analysis could be considered to find out why ranking doesn't seem to have an effect on sales, when instinctively it feels as though it should.

We create a simple linear regression model to explore the relationships between the sales data.

- **NA to EU sales**
  - EU_Sales Coefficient (1.18953): Each unit increase in EU_Sales increases NA_Sales by 1.19 units.
  - R-squared (0.4991): 50% of NA_Sales variability is explained by EU_Sales.
  - P-value (< 2e-16): The relationship between EU_Sales and NA_Sales is statistically significant.
    - 50% needs to be better, more variables should be added.

- **Global to NA sales**
    - NA_Sales Coefficient (1.71837): Each unit increase in NA_Sales increases Global_Sales by 1.72 units.
    - R-squared (0.8745): 87.45% of Global_Sales variability is explained by NA_Sales.
    - P-value (< 2e-16): Indicates a significant statistical relationship between NA_Sales and Global_Sales.
        - 87% is a much healthier percentage.

- **Global to EU sales**
    - EU_Sales Coefficient (2.71593): Each unit increase in EU_Sales increases Global_Sales by 2.72 units.
    - R-squared (0.7705): 77.05% of Global_Sales variability is explained by EU_Sales.
    - P-value (< 2e-16): Demonstrates a highly significant statistical relationship between EU_Sales and Global_Sales.
        - 77% is still good but could have improvement.

There are still large outliers as we have seen from previous models involving the sales data which we may need to address.

- This could involve removing them or using tools such as sqrt to square root the numbers. Discussions will be needed with Turtle Games as to how best to proceed.
- All models are statistically significant.
- All models show positive relationships.

**Now let's consider multiple linear regression.**
We plot the two regional sales categories of NA and EU against Global and run the model.

- **NA_Sales Coefficient** (1.15562): Indicates a strong positive effect on Global_Sales; each unit increase in NA_Sales increases Global_Sales by 1.16 units.
- **EU_Sales Coefficient** (1.34128): Also shows a positive impact on Global_Sales, with each unit increase in EU_Sales increasing Global_Sales by 1.34 units.
- **Multiple R-squared** (0.9687): Suggests that 96.87% of the variability in Global_Sales is explained by NA_Sales and EU_Sales together, indicating a very strong model fit.
- **P-value** (< 2e-16 for both coefficients): Both NA_Sales and EU_Sales have a highly significant statistical relationship with Global_Sales.
- **Intercept** (0.22404): The baseline level of Global_Sales when both NA_Sales and EU_Sales are zero.

The results are very promising indeed and indicate an extremely strong model.

However first we must check for **multicollinearity** by installing the *CAR* package. Multicollinearity occurs when independent variables in a regression model are highly correlated. This can cause problems in estimating the coefficients accurately.

- With both NA_Sales and EU_Sales having a **VIF** of approximately 1.991094, this indicates that there is some correlation between these variables, but it's not high enough to be overly concerned about severe multicollinearity impacting the regression model.

Next we must explore **heteroscedasticity** using the *lmtest* package. Heteroscedasticity is where the variance of the errors is not constant across all levels of the independent variables.

- P-value: 1.913e-06, significantly below the 0.05 threshold, suggesting the presence of heteroscedasticity and that the variance is not consistent.
- This is not surprising as we have seen the presence of several outliers on both ends of the spectrum.

So, we note that there is heteroscedasticity in the data and that will need to be addressed by having a discussion with Turtle Games regarding what path they would best prefer to use as there are many options and factors to consider. Until that time, however, let's continue with the model.

**Making predictions with the MLR Model**

Turtle games gave us some figures to evaluate the model and here they are with their results.

a. NA_Sales_sum of 34.02 and EU_Sales_sum of 23.80 - model output = **71.46**
b. NA_Sales_sum of 3.93 and EU_Sales_sum of 1.56 - model output = **6.86**
c. NA_Sales_sum of 2.73 and EU_Sales_sum of 0.65 - model output = **4.25**
d. NA_Sales_sum of 2.26 and EU_Sales_sum of 0.97 - model output = **4.14**
e. NA_Sales_sum of 22.08 and EU_Sales_sum of 0.52 - model output = **26.44**

Given that the model is statistically significant and operating with an R Squared of 97% these outputs are highly accurate. No further modification of the data on outliers is needed however how to continue with heteroscedasticity present should be addressed at some point with Turtle Games.

An interactive prediction interface is made with Shiny App for non-technical stakeholders. This provides a non-technical audience with a means of using the model in the most accessible way.

Finally for the report a visualisation of the model's working capabilities and accuracy is created by making an Actual vs. Predicted Global Sales plot. This interactive scatter plot shows the model's predictions versus the actual global sales. Similar to the previous model we discussed for loyalty points, this graph plots actual versus predicted sales. Ideally, each data point would align with the diagonal line, representing an accurate prediction. As shown, the majority of points are indeed near this ideal line, demonstrating the model's robust predictive capability. Notice the error indicators, which provide a sense of prediction precision.

**Insights:**

- The sales data is highly correlated, confirmed by a correlation heatmap, leading to the removal of non-numeric columns, un-correlated and irrelevant variables such as "Year" and "Ranking" for model accuracy.

- Simple linear regression models reveal significant relationships between regional sales (NA, EU) and global sales, with coefficients indicating how sales in one region predict another:
  - NA to EU Sales: A 1 unit increase in EU sales increases NA sales by 1.19 units, with 50% of NA sales variability explained by EU sales.
  - Global to NA Sales: A 1 unit increase in NA sales increases Global sales by 1.72 units, accounting for 87.45% of Global sales variability.
  - Global to EU Sales: A 1 unit increase in EU sales increases Global sales by 2.72 units, explaining 77.05% of Global sales variability.
  - All models show statistically significant and positive relationships between the sales variables but exploring the combined effect of NA and EU sales to Global is the correct method.

- The exploration of a multiple linear regression model combining NA and EU sales as predictors for Global sales demonstrates a strong model fit, with an R-squared of 0.9687, indicating that 96.87% of the variability in Global sales is explained by NA and EU sales together.

- The presence of multicollinearity and heteroscedasticity was examined, with findings suggesting manageable levels of multicollinearity (VIF ~1.991094) but significant heteroscedasticity, indicating that variance of errors is not constant, which could affect the reliability of the model.

- Predictions made using the model closely align with provided sales data, reinforcing its statistical significance and accuracy despite the noted presence of outliers and heteroscedasticity. However, using the model in the future must come with the aforementioned caveats.

- A model is built using ShinyApp allowing non-technical stakeholders to simply input their numbers into the pop-up box. This makes the process of using the model as accessible as possible.

- An interactive visualisation of predicted values versus actual values is also created in order to best visualise the MLR model for the presentation.

- Recommendations include addressing heteroscedasticity and potentially revisiting the handling of outliers, in consultation with Turtle Games, to refine the model's predictive accuracy.

---

**Overall Sales data analysis key Insights**:

- *Sales Variability*: There's a wide range between the least and most successful games, with sales data right skewed, indicating that a few games achieve very high sales.
- *Regional Sales Contributions*: North American and European sales combined make up approximately 85% of total global sales, highlighting the significance of these regions.
- *Visual Analysis*: Initial exploratory visual analyses (histograms, boxplots, scatterplots) confirm the right-skewness and the existence of outliers indicating exceptional success for certain games and platforms.
- *Platform Analysis*: Platform-based sales analysis reveals outliers like Xbox360 and Wii as exceptionally successful, with interactive plots providing deeper insights into sales distribution.
- *Statistical Findings*: Tests confirm that sales data do not follow a normal distribution, largely due to significant outliers. There's a strong correlation between regional sales, suggesting that success in NA or EU is a good predictor of global success.
- *Model Predictions*: Multiple linear regression models show a strong predictive relationship between regional sales and global sales, with an R-squared of 0.9687.

**Overall Sales data analysis key Recommendations**:

- *Address Heteroscedasticity*: Discuss strategies to manage the non-constant variance of errors (heteroscedasticity), which could affect model reliability and certainly accuracy.
- *Review Outliers*: Consider revisiting the sensitive handling of outliers, which might involve more sophisticated statistical techniques to ensure model accuracy.
- *Enhance Data Collection*: To improve model accuracy and insights, suggest expanding the data collection to include more regions, potentially other influential variables and focus on improving accuracy of current sales variables.
- *Strategic Planning*: Use the insights from platform performance and regional sales contributions for targeted marketing and strategic planning. Focus on platforms and regions demonstrating the highest sales potential to capture as much of a market that will grow to"*$321 billion by 2026* [11]" a new PwC 2023 report says

[11] PwC. "Global Entertainment and Media Outlook 2023–2027." PwC, 21 June 2023, www.pwc.com/gx/en/industries/tmt/media/outlook/insights-and-perspectives.html.

- *Monitor Market Preferences*: Pay close attention to shifting market preferences, such as the emerging dominance of PlayStation in Europe, to adapt marketing and development strategies accordingly

## Bibliography:


Arslan, I Kahraman. "THE IMPORTANCE of CREATING CUSTOMER LOYALTY in ACHIEVING SUSTAINABLE COMPETITIVE ADVANTAGE - ProQuest." *Www.proquest.com/Openview/B713fccb5e654cec5a2a678a93b67e89/1?Pq-Origsite=Gscholar&Cbl=4371414* , Aug. 2020,

Nandapala, E.Y.L , and K.P.N Jayasena. "The Practical Approach in Customers Segmentation by Using the K-Means Algorithm." *https://Ieeexplore.ieee.org/Abstract/Document/9342639* , IEEE, 2020 . Accessed 2 Mar. 2024.

LinkedIn community. "What Are the Best Practices for Cleaning Data for Natural Language Processing?" *Www.linkedin.com*, 26 Oct. 2023, www.linkedin.com/advice/0/what-best-practices-cleaning-data-natural-language . Accessed 2 Mar. 2024.

Morris, Tom. "Everything to Know about the Console War." *GWI*, 12 Nov. 2020, www.blog.gwi.com/trends/games-console-war .

Otebele, Kobiruo . "From Insights to Conversions: The Power of Customer Review Mining in Conversion Optimization - Invesp." *Invesp*, 2 Aug. 2023, www.invespcro.com/blog/customer-review-mining /. Accessed 2 Mar. 2024.

PwC. "Global Entertainment and Media Outlook 2023–2027." *PwC*, 21 June 2023, www.pwc.com/gx/en/industries/tmt/media/outlook/insights-and-perspectives.html.

Wikipedia Contributors. "Atari 2600." *Wikipedia*, Wikimedia Foundation, 6 Oct. 2019, www.en.wikipedia.org/wiki/Atari_2600

.

Williams, Jonathan. "Challenges in Natural Language Processing Require Coordination across a Large Scientific Network." *European Lab for Learning & Intelligent Systems*, 30 Jan. 2024, www.ellis.eu/news/challenges-in-natural-language-processing-require-coordination-across-a-large-scientific-network . Accessed 2 Mar. 2024.

Yin-Poole, Wesley. "Xbox Series X and S Sales Have Collapsed in Europe." *IGN*, 22 Nov. 2023, www.ign.com/articles/xbox-series-x-and-s-sales-have-collapsed-in-europe . Accessed 2 Mar. 2024.

Yse, Diego Lopez. "Text Normalization for Natural Language Processing (NLP)." *Medium*, 12 Mar. 2021, www.towardsdatascience.com/text-normalization-for-natural-language-processing-nlp-70a314bfa646. Accessed 2 Mar. 2024.

Wijman, Tom . "New Free Report: Explore the Global Games Market in 2023." Newzoo, 8 Aug. 2023, www.newzoo.com/resources/blog/explore-the-global-games-market-in-2023?utm_campaign=2023-08-GMRF-GGMR+2023+launch+article&utm_source=GGMR+2023+Press+Release&utm_term=GGMR+2023+Press+Release&utm_content=GGMR+2023+Press+Release . Accessed 3 Mar. 2024.