TURTLE GAMES: CUSTOMER SALES/TRENDS

Analysis Report

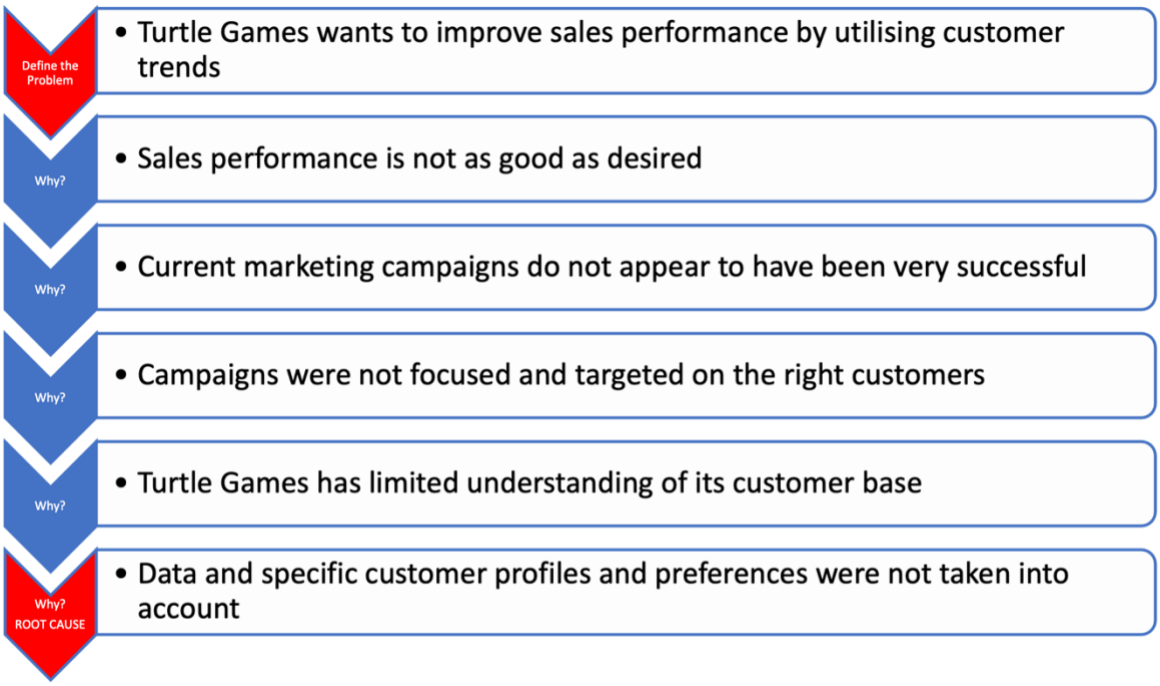
FAO: Marketing Director

Claire Lawrence
Data Analyst

Introduction

The Business Problem

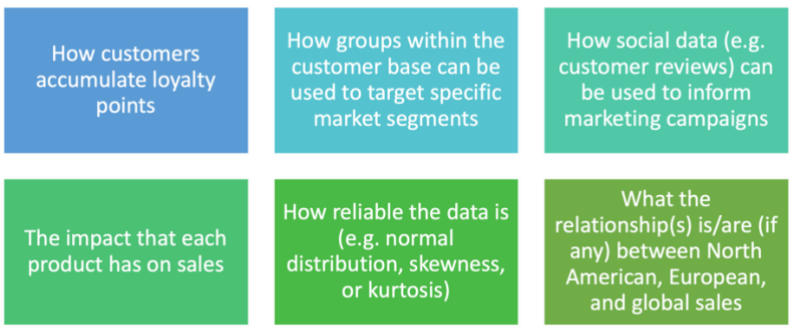
Turtle Games is a games manufacturer/retailer with a global customer base, selling their own and others' products (video games, books, board games, toys). The business problem is defined below:



Report Purpose

This report analyses customer demographics, reviews and sales[1] to inform new marketing development strategies.

Main Questions



[1] turtle_reviews.csv, turtle_sales.csv.

Analytical Approach

Tool	Action	Process	Rationale	Relevant Observations
Python	**Imported libraries (numpy, pandas, matplotlib, seaborn, statsmodels, sklearn, scipy, nltk, textblob, wordcloud, counter.**		Carry out various analyses.	
	Data imported into DataFrame. Sense-checked.	Using head(), info(), describe().	Import, check, understand data.	
	Null values checked.	Isnull().sum()	Prevent impacting analysis.	None found.
	Outliers checked.	Identify upper/lower limits of numerical variables, DataFrame filtered.	Prevent impacting analysis.	New DataFrame without outliers created.
	Data cleaned.	Columns dropped/renamed.	Clean data.	
	Linear regression	Spending, remuneration, age plotted.	See if variables are good predictors of loyalty points.	Poor models, best with original data.
	Multiple linear regression (three/two variables, spend_score >= 60)	All variables plotted.	See if variables are good predictors of loyalty points.	Poor models. Spend score <60 did not show improved result.
	Customer profiles examined (inc. top/bottom points quartiles).	Histograms, bar plots. Quartile data subset into new DataFrames.	Understand customer profiles.	Average profile identified. Points vs. spending score do not correlate.
	K-means clustering.	Remuneration/spending score plotted, ideal cluster number tested, model fitted, clusters interpreted using describe(), bar plots.	Identify customer categories to predict behaviour/inform marketing.	7 clear clusters identified, trends analysed. Very effective model.
	Cluster predictions merged with review data.	Merged into new DataFrame.	For later analysis to relate customer/review to cluster.	

	Top words in reviews identified.	Duplicates removed, words tokenised, stopwords removed, word clouds and count plots generated.	Top words to inform marketing.	Positive words featured heavily. 'Book' was common.
	Sentiment analysis of reviews.	Polarity scores generated, histograms plotted.	Identify sentiment for marketing.	Leaning towards positive sentiment.
	Top 20 positive/negative reviews identified.	Top/bottom 20 reviews into DataFrames.	Understand reviews, link to products.	Largely accurate, some errors i.e. 107 reviewed as toy, not video game.
R	**Imported libraries tidyverse, plotly, moments, psych**		Carry out various analyses.	
	Data imported into dataframes, sense-checked.	Using View, str, dim, as_tibble, typeof, class	Understand data, check usability.	
	Product variable changed to factor. Check for null values.	Using mutate/table functions.	Prevent negative impact on analysis.	No null values.
	Basic trends explored including outliers.	Basic plots (histograms, scatterplots), sum calculations, upper/lower limits calculated.	Understand trends/data.	Outliers identified/filtered. Global_Sales includes non EU/NA sales.
	Top-selling products explored.	Summed/filtered dataframes created, interactive barplots generated.	Understand top 10 products.	Wii platform popular, product 107 especially. Only video game data present (limitation).
	Normality of dataset checked (original & without outliers).	QQPlots, Shapiro-Wilk test, skewness, kurtosis.	Determine reliability.	Heavy tails, positive skew, not normally-distributed.
	Linear regression (original/sum data).	Various models created. Log applied using mutate() function. Plots used to view output.	Determine a good model to predict Global_Sales.	Models using sum data were best.
	Multiple linear regression on sum data.	Model built. Predicted values compared to actual.	Model to predict Global_Sales.	Accurate/usable e.g. actual = 67.85, predicted = 68.06.

Patterns, Trends and Insights

Accumulation of Loyalty Points

- There is no clear relationship between any/all variables (spending score, income and age) and a customer's loyalty points. Also, a customer with low income may buy an expensive product and therefore obtain more points.
- There appears to be a linear relationship between income/spending score. A higher-income customer is predicted to have a higher spending score.

Limitations:

- *Total (i.e. lifetime) points would allow for better predictive models and more reliable prediction of customer behaviour.*
- *It is not clear how spending score is allocated.*

Recommendations:

- *Further analysis/modelling using lifetime customer points.*
- *Linear regression model predicting income with spending score (score allocation system needs clarifying).*

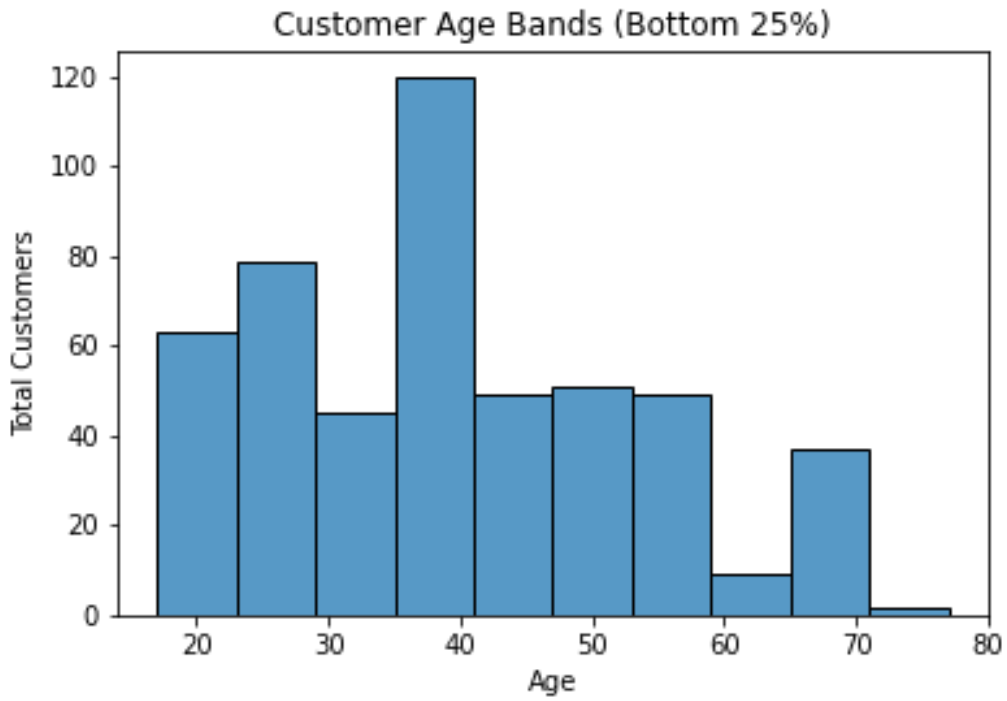
Market Segments to Target

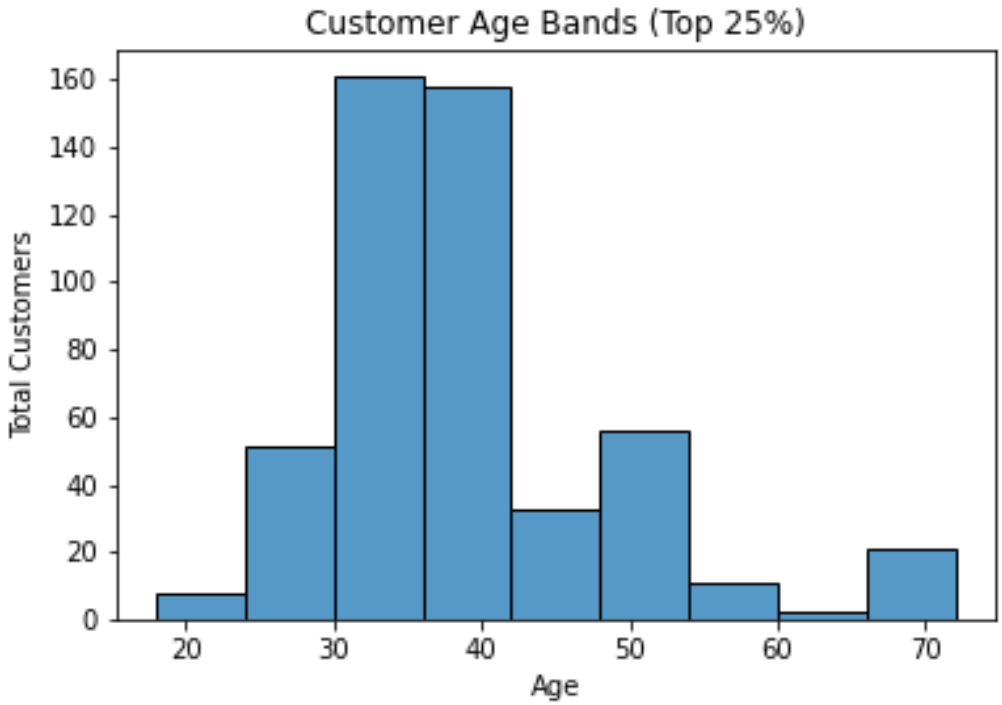
The average customer is:

- 39.5 years old;
- Earns £48k;
- Slightly more likely to be female;
- A graduate or PhD;
- Buying for children.

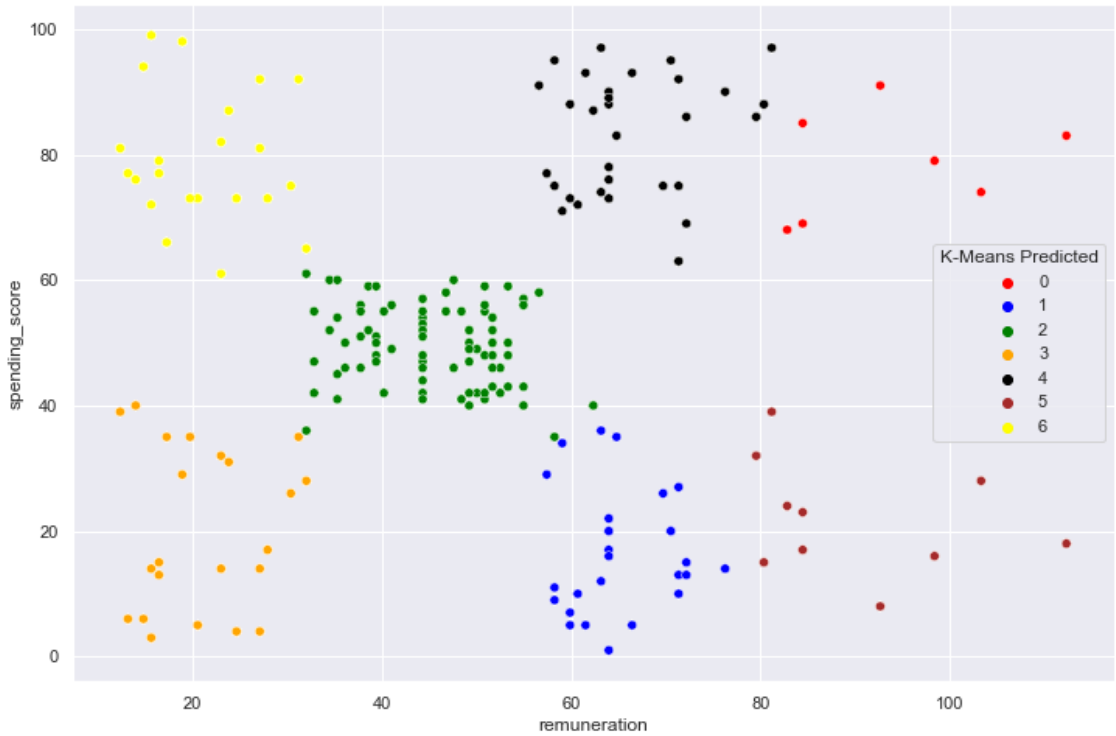
The top and bottom 25% of customers points totals were reviewed. They shared the above characteristics - differences were only:

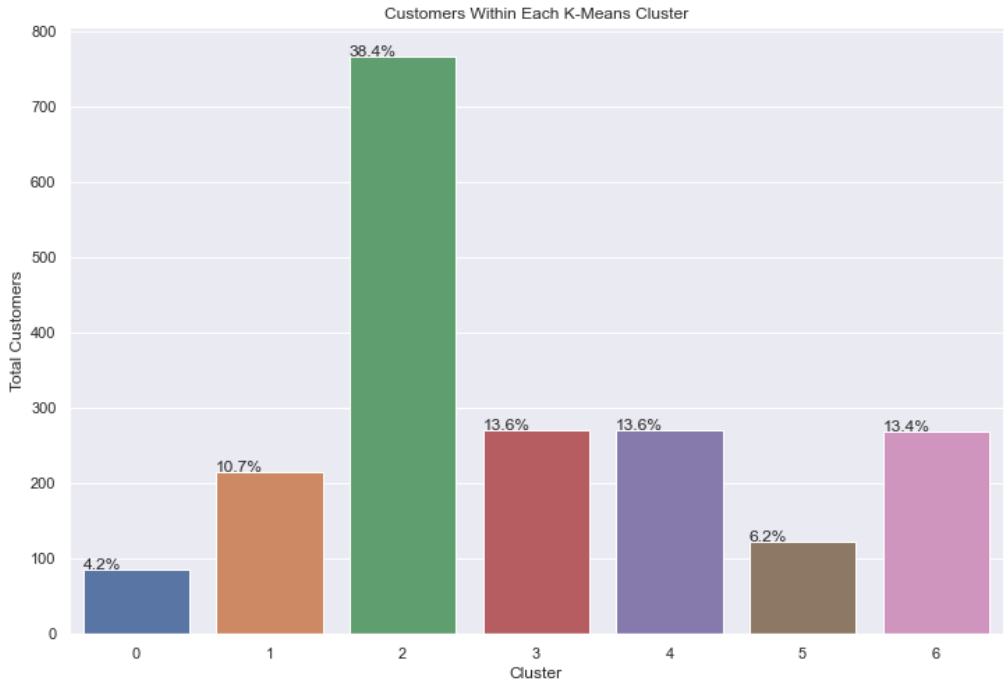
- Income (top 25% average £68k, bottom 25% average £33k);
- Age distribution (top 25% largely aged 30s/40s, bottom 25% more broadly between 20s and 70s).





The clustering of customer characteristics suggest the largest category is cluster 2 (see charts below), followed by clusters 3, 4 and 6:





Clusters 2, 3 and 4 follow the income/spend linear pattern discussed earlier:

- Cluster 2 income = £32k to £62K;
- Cluster 3 income = £12.3k to £32k;
- Cluster 4 income = £57k to £81k.

Evidently, a large proportion with lower income have a high spend (cluster 6).

Recommendations:

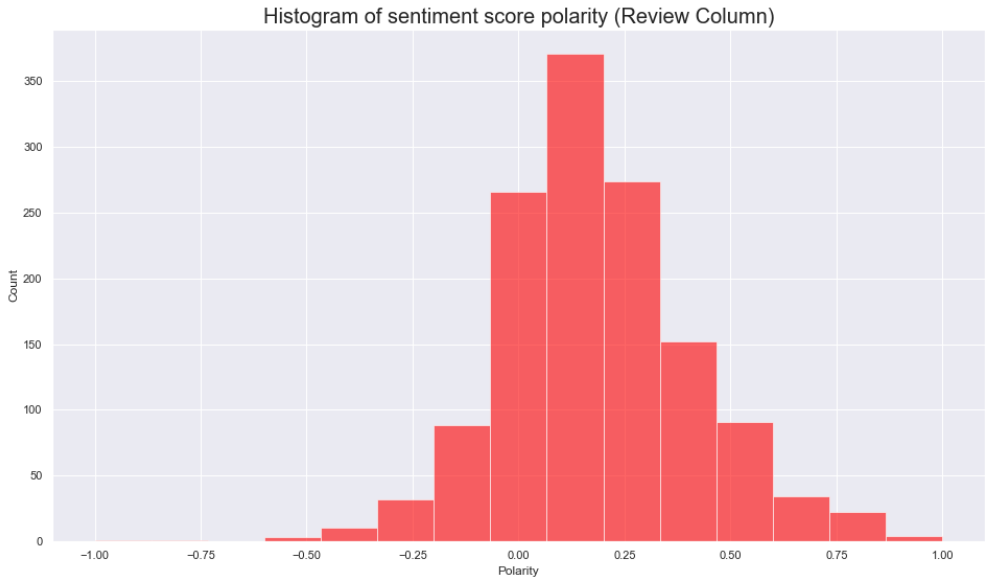
- *Target customers within stated demographic for marketing campaigns (e.g. Facebook).*
- *Focus advertising towards 3x specified clusters (primarily £32k to £62k, the most common cluster).*
- *Further analysis to determine which products each segment purchases for directed targeting.*

Social Data Analysis

Common words within reviews imply positiveness and, often, family-focus:

The presence of 'book' is interesting – this suggests books are commonly bought. With 301 appearances, this may be an important product type.

Sentiment analysis demonstrated a positive leaning within reviews (see below):



Limitations:

- *The top 20 positive/negative review identification was largely accurate, with some identified errors.*

Recommendations:

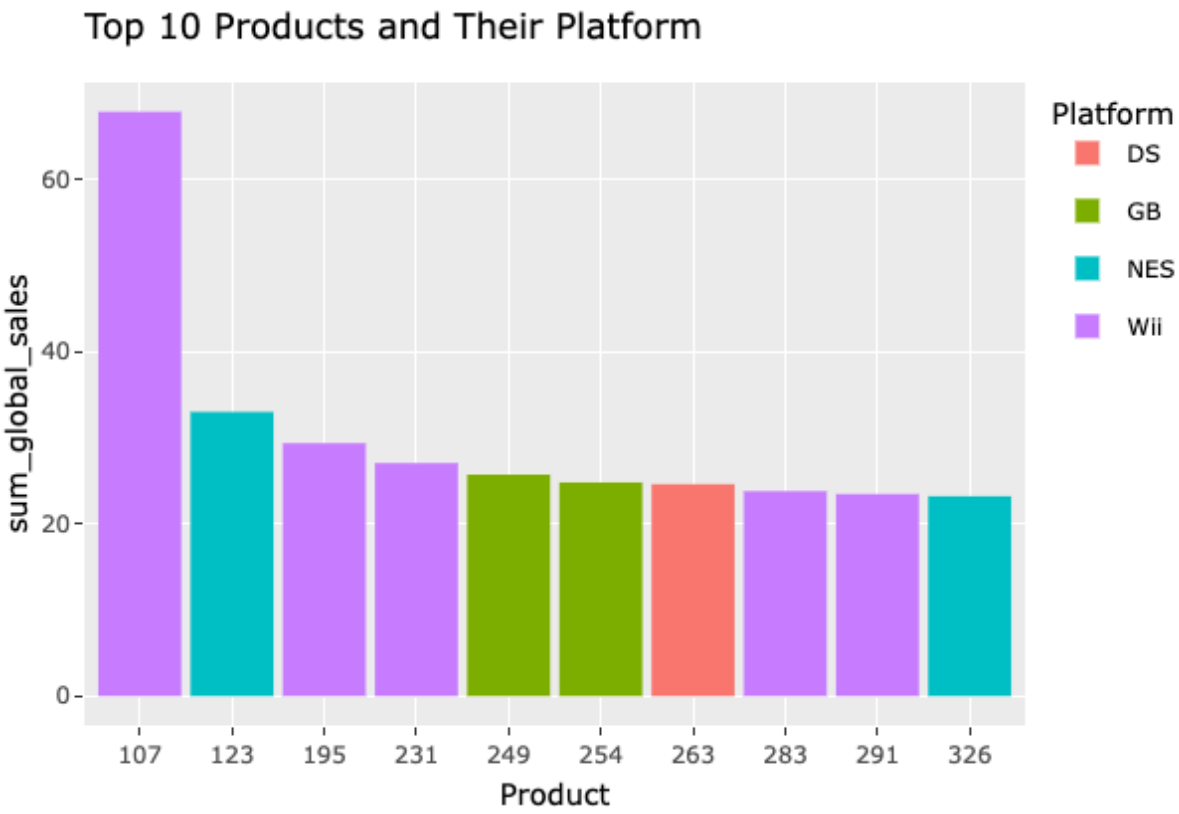
- *Consider associating reviews with products. Customers who gave poor reviews could be contacted for remediation.*
- *Further analysis to determine how books rank in sales/whether to advertise them more.*
- *Target marketing at those with children (reviews suggest products often bought for kids).*

Impact of Products on Sales

Available sales data only includes video games. Findings suggest:

- Product 107 (Wii platform/'sports' genre) is by far the top seller across all regions. It's suggested the COVID pandemic may be a factor here;
- The most popular products differ by region. However, products 195, 231 and 123 are universally popular;
- Globally, 'shooter', 'action' and 'sports' games are most popular, on the X360, PS3 and PC.

The top 10 products and their platforms are shown below:



Top products/genres - EU:	107, 195, 231, 399 Action, Sports, Shooter
Top products - North America:	107, 123, 326, 254 Shooter, Platform[2], Action

Limitations:

- *Full sales data (inc. dates) needed to fully understand sales (not only video games).*

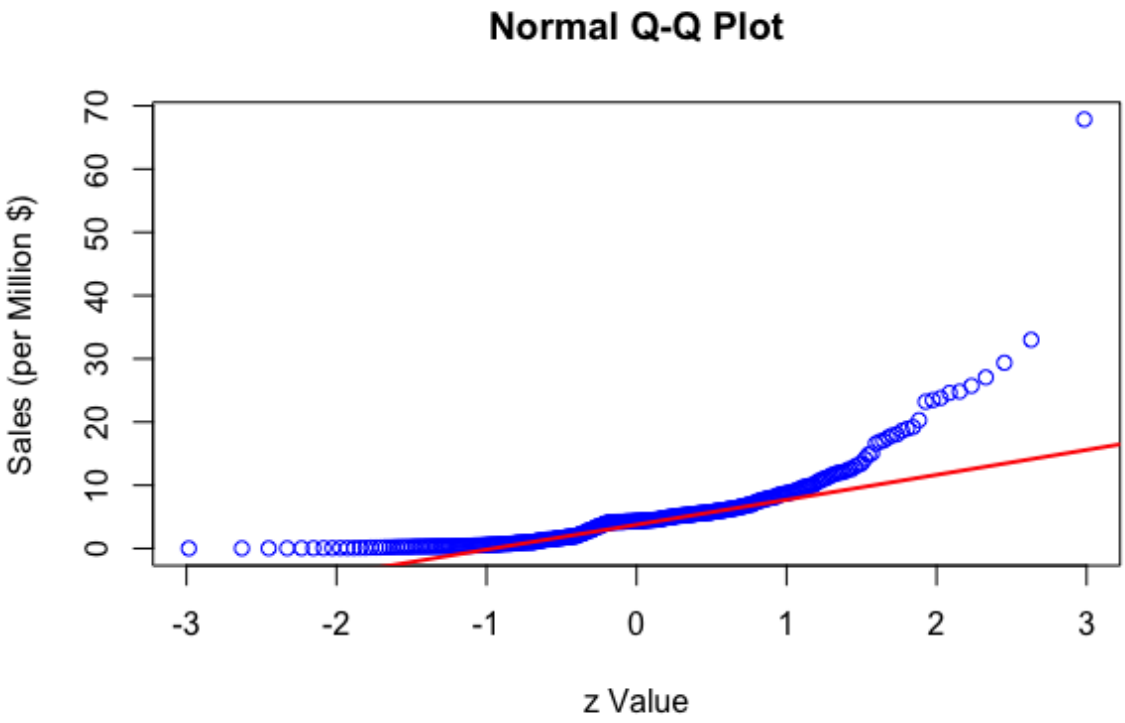
Recommendations:

- *Focus advertising on Shooter/Action/Sports games.*
- *Identify why 107 sold so many to inform strategy.*
- *Further data/analysis on products/their cost to understand popularity and corresponding purchasers.*

Data Reliability

The data is not normally-distributed and is positively skewed (see below). Sales levels are higher than the model predicts at the lowest and highest ends, suggesting perhaps a large price gap between products. However, all sales are not represented. Including all sales could improve reliability.

[2] Recording error?



Limitations:

- *Not all sales data available.*

Recommendation:

- *Obtain full data for further/more accurate analysis.*

Relationship Between Regional Sales

North America (£885.62m) accounts for 47% of sales, while the EU accounts for 31% (£578.61m). 22% were outside of these regions.

A multiple linear regression model was built to predict global sales based upon EU and North American sales with good accuracy:

Actual Global Sales Value (£m)	Predicted Global Sales Value (£m)
68.06	**67.85**
4.91	**4.32**
26.63	**23.21**

Recommendation:

- *Utilise model to predict global sales.*