Assignment 3: Predicting Future Outcomes for Turtle Games

Background and Business Problem Context

Turtle Games is a game manufacturing and retailing company which operates globally. The range of products offered by the company ranges from video and board games to books and toys. As its main business objective, Turtle Games is concerned with improving overall sales performance.

The Report employs an analytical approach which is informed by the business questions aimed to gauge the effectiveness of Turtle Games sales performance.

1. Reviews:

1.1. How do customers accumulate loyalty points?

1.2. How specific groups within the customer base can be used to target specific market segments?

1.3. How can social data be used to inform marketing campaigns?

2. Sales:

2.1. What impact does each product have on sales?

2.2. How reliable is the data used to make assumptions about sales relationships?

2.3. What is/are the relationship(s) between North American, European and global sales?

The Report has analysed the available data to make recommendations, communicate insights into customer trends and inform the sales and marketing strategy of Turtle Games decision-makers.

Analytical Approach

The Report used Python and R programming tools to analyse key metrics and trends to inform decision-making.

Reviews [Python]

In the initial steps, the Report loaded the 'turtle_reviews.csv' in Python Notebook and explored the dataset from which a separate 'reviews' DataFrame was created and further analysed with descriptive statistics. The 'drop()' function enabled to eliminate all unnecessary to accentuate only the information particularly relevant to discern the relationships in the customer base demographics and determine customer attitudes towards the products.

In [9]:
```
# Rename the column headers.
reviews_new_col = pd.DataFrame(reviews_drop)
reviews_new_col = reviews_new_col.rename({'remuneration (k£)': 'remuneration',
                                          'spending_score (1-100)': 'spending score',
                                          'loyalty_points': 'loyalty points'}, axis=1)
reviews_new_col.head()
```

Out[9]:

	gender	age	remuneration	spending score	loyalty points	education	product	review	summary
0	Male	18	12.30	39	210	graduate	453	When it comes to a DM's screen, the space on t...	The fact that 50% of this space is wasted on a...
1	Male	23	12.30	81	524	graduate	466	An Open Letter to GaleForce9*:\n\nYour unpaint...	Another worthless Dungeon Master's screen from...
2	Female	22	13.12	6	40	graduate	254	Nice art, nice printing. Why two panels are f...	pretty, but also pretty useless
3	Female	25	13.12	77	562	graduate	263	Amazing buy! Bought it as a gift for our new d...	Five Stars
4	Female	33	13.94	40	366	graduate	291	As my review of GF9's previous screens these w...	Money trap

1.1. How do customers accumulate loyalty points?

The Report performed the ***linear regression*** function and visualised the regression line to evaluate the statistical relationship between different sets of variables from the DataFrame. In the linear regression model, the 'loyalty points' column was assigned to the dependent variable measured in the model against three sets of dependent variables, i.e., 'spending', 'remuneration' and 'age'.

```
# Independent variable.
x = reviews_new['spending score']

# Dependent variable.
y = reviews_new['loyalty points']
```

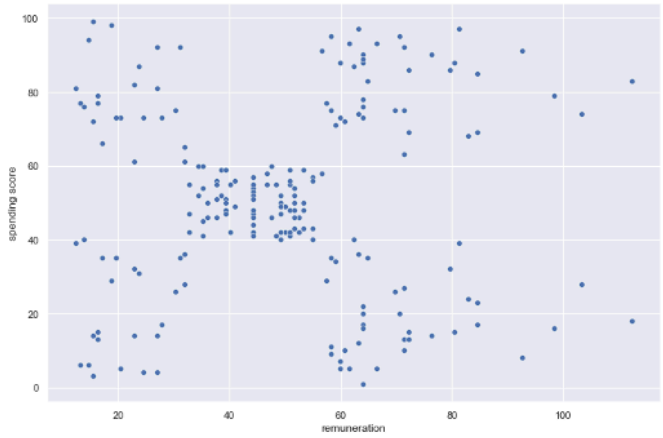
To minimize the square of errors or the distance between the predicted and actual values, the Report employed the ***Ordinary Least Squares (OLS)*** method.

```
# OLS model and summary.
f = 'y ~ x'
reviews_model = ols(f, data = reviews_new).fit()

reviews_model.summary()
```

1.2. How specific groups within the customer base can be used to target specific market segments?

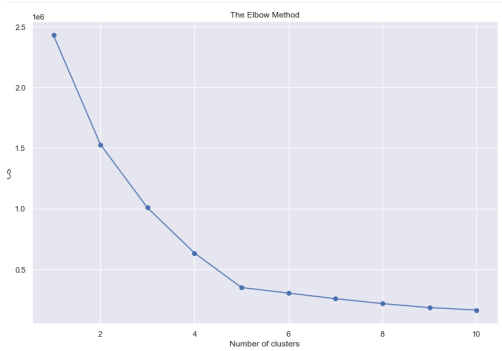
To identify the specific groups in the customer base, the Report built and tested the K-means clustering. The Elbow and Silhouette methods helped determine the optimal number of clusters to fit the k-means predicted model.

```
In [308]: # Visualise the data
          ax = sns.scatterplot(x='remuneration', y='spending score',
                               data=df2_drop)
```



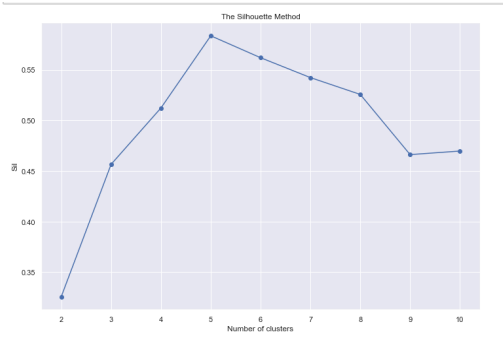
The Elbow Method

The graph starts to straighten out at k = 5.



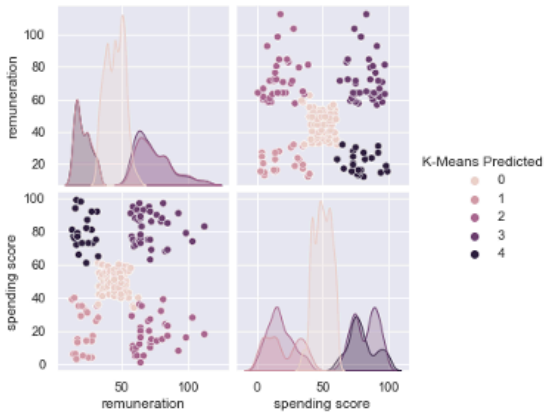
The Silhouette Method

The graph reaches a peak at k = 5.



The fitting of the K-Means Predicted to the optimal number of clusters was displayed in the pairplot with the five distinct clusters derived from the correlation between remuneration and spending scores. The histograms reveal that the dataset is non-normally distributed.

```
# Apply the final model:
## The optimal number of clusters if 5 as determined by the Elbow and Silhouette metho
kmeans = KMeans(n_clusters = 5, max_iter = 15000, init='k-means++', random_state=0).fit
clusters = kmeans.labels_
x['K-Means Predicted'] = clusters

# Plot the predicted.
ax = sns.pairplot(x, hue='K-Means Predicted', diag_kind= 'kde')
```



1.3. How can social data be used to inform marketing campaigns?

The Report applied ***Natural Language Processing (NLP)*** to understand and interpret customer attitudes. The ***tokenisation*** function was applied to the both 'review' and 'summary' columns to plot the ***WordCloud*** images to break sentences into individual words.

```
# Tokenise the words.
df4['tokens_review'] = df4['review'].apply(word_tokenize)
df4['tokens_summary'] = df4['summary'].apply(word_tokenize)

df4 = df4.reset_index()

# Preview data.
df4.head()
```

	index	review	summary	tokens_review	tokens_summary
0	0	when it comes to a dms screen the space on the...	the fact that 50 of this space is wasted on ar...	[when, it, comes, to, a, dms, screen, the, spa...	[the, fact, that, 50, of, this, space, is, was...

To improve the accuracy of the visualization, all the ***alpanum***, or tokens that are neither alphabets nor numbers (e.g., punctuation marks, etc.), were filtered out alongside the English ***stopwords***. A new WordCloud was created from the list of tokens without stop words.

```
# Delete all the alpanum.
tokens1 = [word for word in all_tokens if word.isalnum()]
```

```
# Remove all the stopwords

# Import the stop word list.
from nltk.corpus import stopwords

# Create a set of English stop words.
english_stopwords = set(stopwords.words('english'))
```

```
# Create a filtered list of tokens without stop words.
tokens2 = [x for x in tokens1 if x.lower() not in english_stopwords]

# Define an empty string variable.
tokens2_string = ''
for value in tokens2:
    # Add each filtered token word to the string.
    tokens2_string = tokens2_string + value + ' '
```

To identify the most frequent word, the **frequency distribution** function FreqDist() was applied. The new DataFrame created from the Counter class was used to plot a histogram with the 15 most common words captured in the social data.

```
# View the frequency distribution.
fdist1 = FreqDist(tokens2)

# Preview the data.
fdist1
```

```
FreqDist({'stars': 466, 'five': 381, 'game': 319, 'great': 295, 'fun': 218, 'love': 93
, 'good': 92, 'four': 58, 'like': 54, 'expansion': 52, ...})
```

```
# Import the Counter class.
from collections import Counter

# Generate a DataFrame from Counter.
counts = pd.DataFrame(Counter(tokens2).most_common(15),
                      columns=['Word', 'Frequency']).set_index('Word')

# Preview data.
counts
```

Finally, the Report reviewed the ***polarity and sentiment scores*** to determine whether both columns are normally distributed and display the 20 negative and positive sentiments expressed by customers online.

Sales [R / R Studio]

2.1. What impact does each product have on sales?

Since the sales department prefers R to Python, the Report embarked on the exploratory data analysis in R Studio to explore and prepare insights into the sales relationships. Having imported the 'turtle_sales.csv' file into the R Studio and created a 'sales_data' framework, the Report then used ***subsetting*** to remove unnecessary columns.

```
# Print the data frame.
sales_data <- read.csv('turtle_sales.csv')
View(sales_data)

# Create a new data frame from a subset of the sales data frame.
# Install and import necessary packages.
install.packages("dplyr")
library(dplyr)

# Remove unnecessary columns.
sales_data_new <- subset(sales_data, select = c(Product, Platform, NA_Sales, EU_Sales, Global_Sales))

# View the data frame.
View(sales_data_new)
```

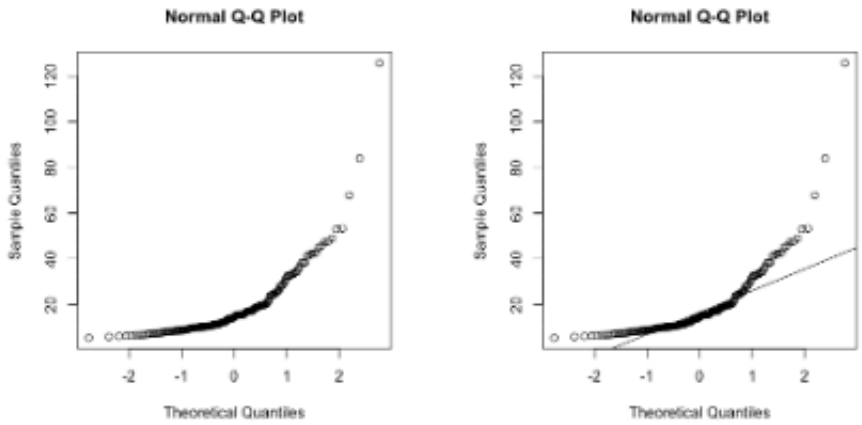
The group_by() was used to create a new dataset based on the sum aggregated values of sales based on Product. To develop a more comprehensive analytical model, the Report also produced the combined sales dataframe which summarised the aggregated or total values of global and regional sales.

```
## 2a) Use the group_by and aggregate functions.
# Group data based on Product and determine the sum per Product.
sales_per_product = sales_data_new %>% group_by(Product) %>%
  summarise(Total_NA_Sales = sum(NA_Sales),
            Total_EU_Sales = sum(EU_Sales),
            Total_Global_Sales = sum(Global_Sales),
            .groups = 'drop')

# View the data frame.
View(sales_per_product)
# Explore the data frame.
summary(sales_per_product)

### Create a column combining total NA_Sales, EU_Sales and Global_Sales per product
combined_sales_per_product = sales_data_new %>% group_by(Product) %>%
  summarise(Total_Sales = sum(NA_Sales, EU_Sales, Global_Sales),
            .groups = 'drop')
```

2.2. How reliable is the data used to make assumptions about sales relationships?

To explore and explain the normality of the dataset, the Report was processed with the following steps:

- Normal ***Q-Q Plots*** with qqnorm() and qqline() functions, to observe the linearity and measure how close to normal the standardised residuals are.



- The ***Shapiro-Wilk test***, to observe the Statistic value W;

```
        Shapiro-Wilk normality test

data:  combined_sales_per_product$Total_Sales
W = 0.71023, p-value < 2.2e-16
```

- ***Skewness and Kurtosis***, to determine the symmetry of the distribution vie checking the relative and combined size of the two tails, respectively;

```
> ## 3c) Determine Skewness and Kurtosis
> # Skewness and Kurtosis.
> skewness(combined_sales_per_product$Total_Sales)
[1] 3.141885
> kurtosis(combined_sales_per_product$Total_Sales)
[1] 18.5417
```

- The ***correlation*** function, to calculate the correlation between Product and Total Sales.

```
> ## 3d) Determine correlation
> # Determine correlation.
> cor(combined_sales_per_product$Product, combined_sales_per_product$Total_Sales)
[1] -0.5876374
```

2.3. What is/are the relationship(s) between North American, European and global sales?

As the department of sales is particularly concerned with better understanding the relationship between North America, Europe and global sales, the Report investigates all possible patterns in the simple data by using a simple and multiple linear regression model.

- In the ***simple linear regression model***, the three relationships were explored which yielded the individual forecasting Models 1–3.

```
## 2a) Determine the correlation between columns
# Create a linear regression model on the original data.
cor(sales_per_product)
View(cor(sales_per_product))

### (1) Test the relationship between Global Sales and NA Sales
plot(sales_per_product$Total_NA_Sales, sales_per_product$Total_Global_Sales)

model1 <- lm(Total_Global_Sales ~ Total_NA_Sales, data = sales_per_product)
summary(model1)

### (2) Test the relationship between Global Sales and EU Sales
plot(sales_per_product$Total_EU_Sales, sales_per_product$Total_Global_Sales)

model2 <- lm(Total_Global_Sales ~ Total_EU_Sales, data = sales_per_product)
summary(model2)

### (3) Test the relationship between NA Sales and EU Sales
plot(sales_per_product$Total_NA_Sales, sales_per_product$Total_EU_Sales)

model3 <- lm(Total_EU_Sales ~ Total_NA_Sales, data = sales_per_product)
summary(model3)
```

The forecasting accuracy and strength of the regression models were tested by visualisations, descriptive statistics via summary() function and the calculation of the sum of squares error (SSE).

```
### MODEL 1
# Plot the residuals.
plot(model1$residuals)
# Specify the coefficients and add a line of best fit.
abline(coefficients(model1))

# Calculate the sum of squares error (SSE) to determine the accuracy of the forecasting model.
SSE1 = sum(model1$residuals^2)
SSE1
```

- The ***multiple linear regression model*** tested the relationship between Global Sales and two independent variables, i.e., Sales in North America and the European Union. The accuracy and significance of Model 4 were tested by visualisations and descriptive statistics.

```
# 3. Create a multiple linear regression model

# Select only numeric columns from the original data frame.
sales_data_numeric <- subset(sales_data, select = c(NA_Sales, EU_Sales, Global_Sales))
View(sales_data_numeric)

# Multiple linear regression model.
cor(sales_data_numeric)
View(cor(sales_data_numeric))
model4 <- lm(Global_Sales ~ NA_Sales+EU_Sales, data = sales_data_numeric)

# Descriptve statistics
summary(model4)
```

- In addition, the ***prediction*** function with confidence interval was calculated to test the forecasting power of the linear multiple regression model.

```
> given_values1 <- data.frame(NA_Sales=c(34.02), EU_Sales=c(23.80))
> predict(model4, newdata=given_values1)
       1
71.46857
> ###Predict the values with confidence interval
> predict(model4, newdata=given_values1, interval = 'confidence')
       fit      lwr      upr
1 71.46857 70.16242 72.77472
```

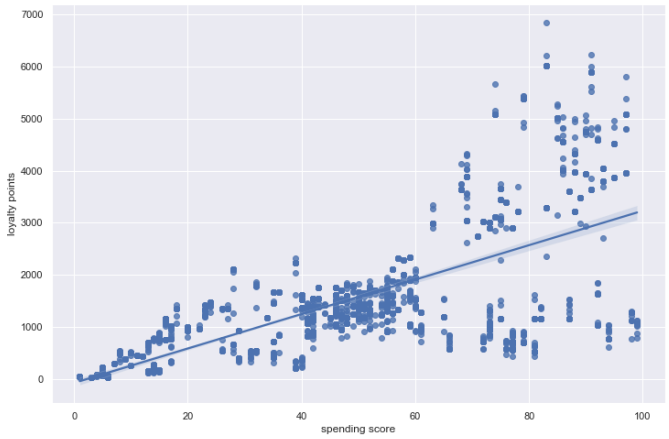
Visualisation and Insights

Reviews [Python]

1.1. How do customers accumulate loyalty points?

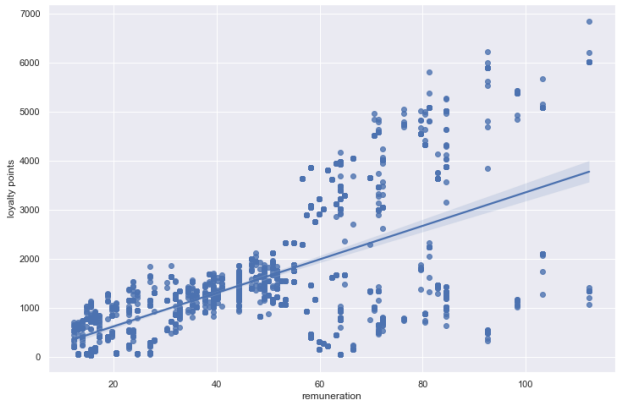
The produced OLS Regression summarizes the information on the accuracy, or fit, of the three computed regression models. The strongest statistical relationship exists between ‘loyalty’ and ‘spending’ variables, with the R-squared value of 0.452 and P-value of 0.102 being comparably high. It could be suggested, therefore, that the customer’s spending nature and behaviour affect the accumulation of loyalty points.

Spending Score (x) vs Loyalty Points (y)



As indicated by the OLS models and regression plots, the other two independent variables – i.e., 'remuneration' and 'age' – have significantly less impact on the accumulation of loyalty scores as reflected in the lower R-squared values of 0.379 and 0.001. In both cases, low R-squared values suggest that neither income per year nor age variables could explain the variation in the loyalty accumulation.

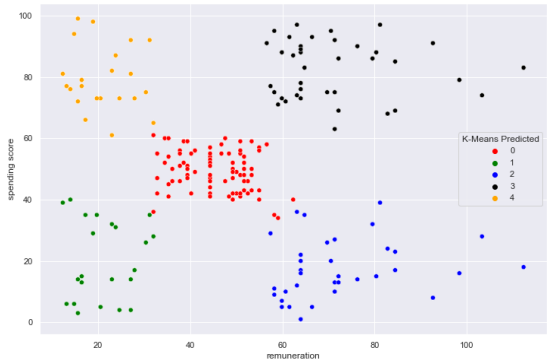
Remuneration (x) vs Loyalty Points (y)



Age (x) vs Loyalty Points (y)

1.2. How specific groups within the customer base can be used to target specific market segments?

The clustering methods lowed to distinguish the five specific groups of customers based on the attributed remuneration scores. The K-Means Predicted model revealed that there no explicit assumptions aba out positive correlation can be made about the impact of salaries per year on the spending score. That is, higher annual salaries do not necessarily translate into higher spending scores.



	remuneration	spending score
0	12.30	39
1	12.30	81
2	13.12	6
3	13.12	77
4	13.94	40

In that, Turtle Games analytical departments should take caution when extrapolating the wider insights about the market segment from the specific groups in the sample customer dataset.

1.3. How can social data be used to inform marketing campaigns?

The WorldCloud images generated using the tokenisation function reflect the most common attitudes expressed in the online customer reviews and summaries. While a large number of the words captured in the WordCloud include words from which no evaluation of negative/positive sentiments could be discerned (e.g., “game”, “play”, etc.), the elimination of alpanum and English stopwords enabled to create a more nuanced representation of customer sentiments.

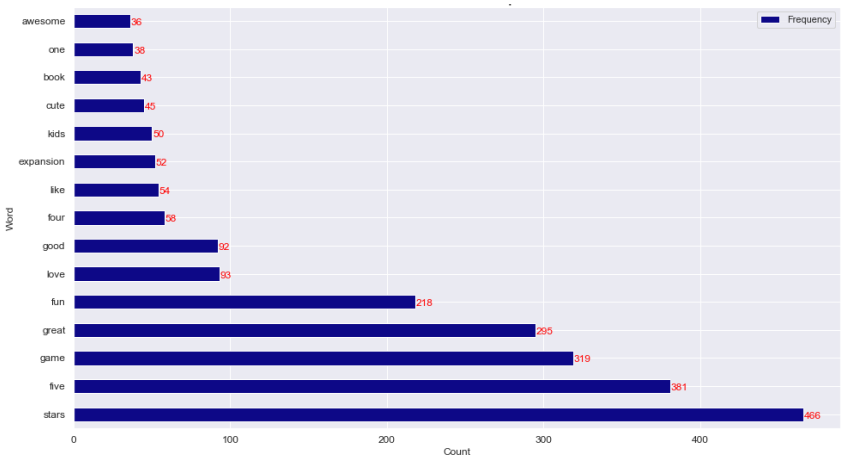
REVIEW WorldCloud

SUMMARY WorldCloud

Based on the WorldCloud created for the 'review' and 'summary' and the count of the 15 most frequent words, it could be argued that customer attitudes towards the Turtle Games products are mainly positive. For example, the sentiments shared by customers include among others such evaluative characteristics as "five stars", "like/love", "good", and "awesome" which might be indicative of higher satisfaction with product quality.

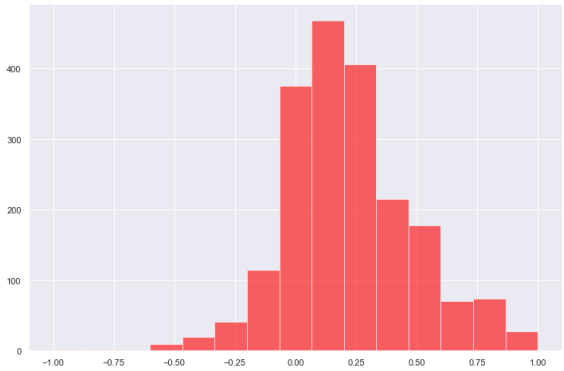
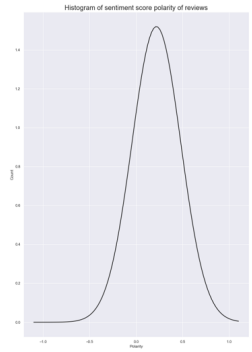
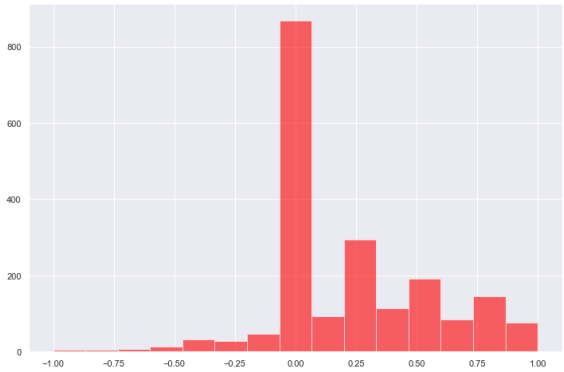
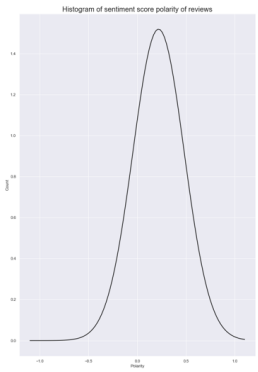
WorldCloud without stopwords

Top-15 most frequent words



Word	Frequency
stars	466
five	381
game	319
great	295
fun	218
love	93
good	92
four	58
like	54
expansion	52
kids	50
cute	45
book	43
one	38
awesome	36

From the histogram of polarity and sentiment score polarity, it is evident that online reviews submitted by customers who purchased and used the products deviate from a normal distribution pattern and seem to be slightly asymmetric and skewed to the left. Similarly, the summarises of the customer reviews follow a left-skewed distribution.

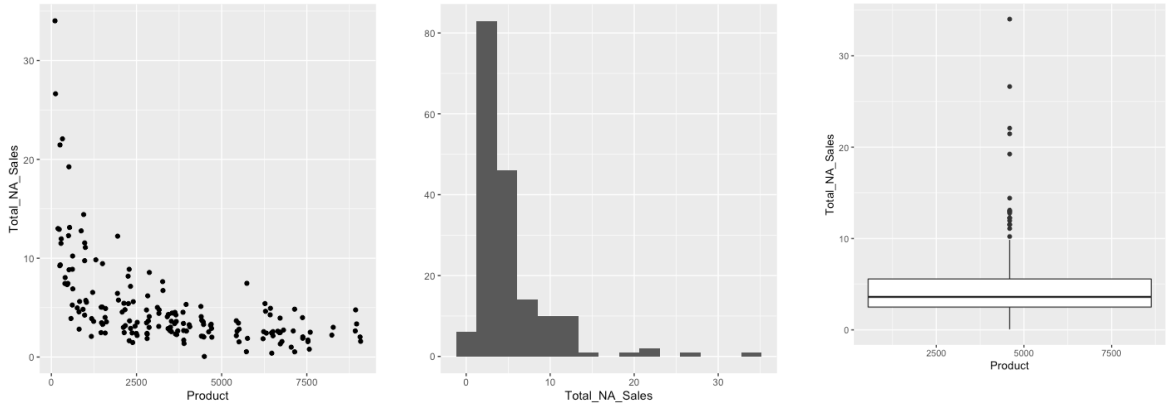
	Histogram of polarity	Histogram of sentiment score
REVIEW		
SUMMARY		

Sales [R / R Studio]

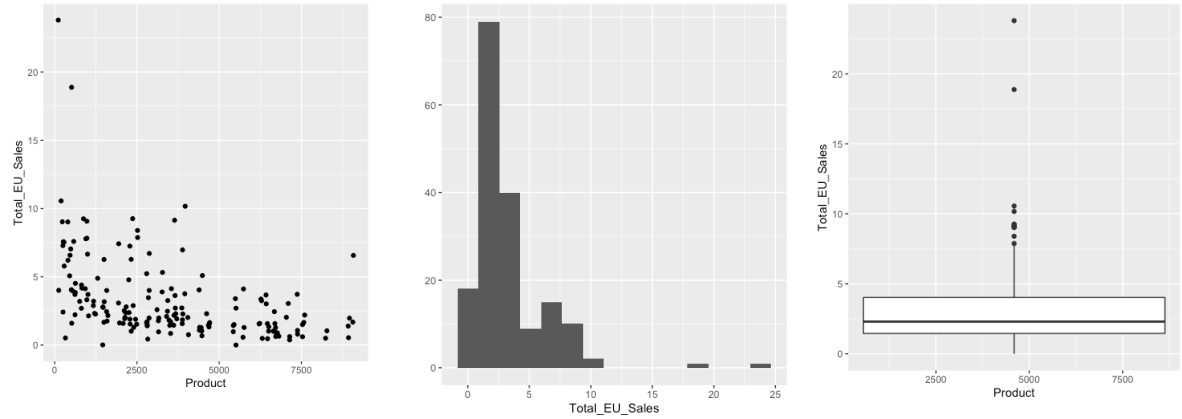
2.1. What impact does each product have on sales?

To determine the initial insights into the data, a separate set of scatterplots, histograms and boxplots were created for the sales in the world, North America and Europe. The same visualisations were subsequently improved by taking the aggregated data based on the sum per Product. As displayed in the table below, the resulting aggregate impact of each product on sales contains several outliers and has a right-skewed distribution function.

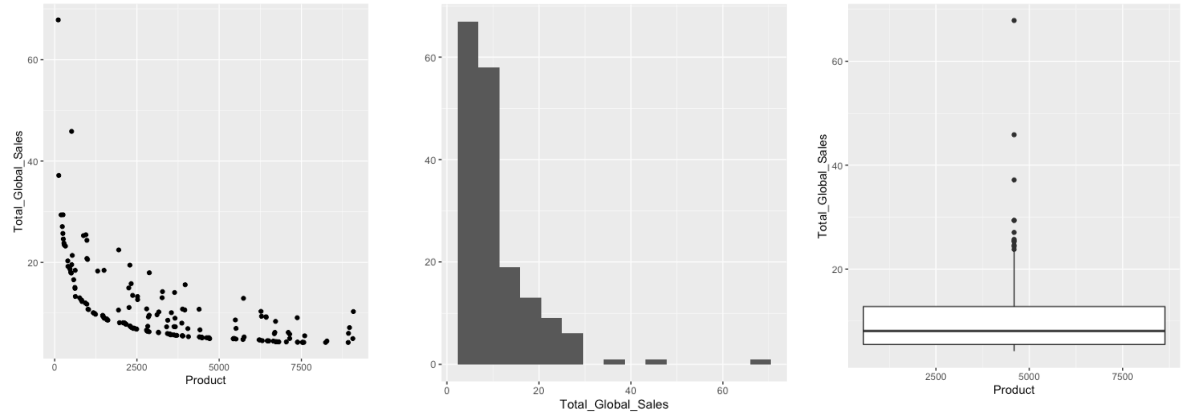
Sales in North America (x) vs Product (y)



Sales in Europe (x) vs Product (y)



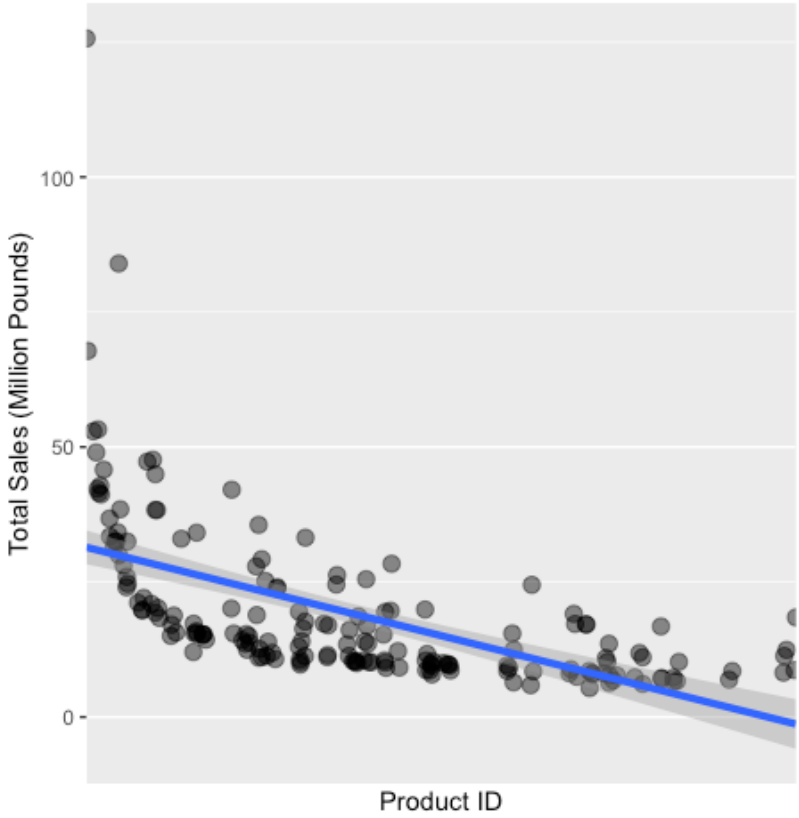
Global Sales (x) vs Product (y)



2.2. How reliable is the data used to make assumptions about sales relationships?

Further analysis reveals that the sales dataset does not match the normal distribution as evidenced by the Normal Q–Q plots for the relationship between Product and Total Sales. The absence of strict linearity and the presence of several quantiles distanced from the standard normal distribution indicates that the dataset is not normally distributed.

Relationship between Product ID and Total Sales



Patterns and Predictions

Reviews [Python]

1.1. How do customers accumulate loyalty points?

From the linear regression and OSL analysis, it could be therefore inferred that the accumulation of loyalty points by customers is associated with the amount of the spending score. While high in relative terms when compared to two independent variables against which the regression models are tested, the R-squared value (0.452) indicates that the independent variable does not explain the variation in the dependent variable. Furthermore, a high P-value (0.102) indicates the insignificance of the model which translates into the lack of sufficient evidence to make assumptions about the statistical relationship between spending and loyalty. Hence, other factors such as brand recognition or the level of service perceived by the customer should be taken into consideration before deriving a conclusion about the causal relationship between loyalty and spending.

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.452
Model:	OLS	**Adj. R-squared:**	0.452
Method:	Least Squares	**F-statistic:**	1648.
Date:	Sun, 11 Sep 2022	**Prob (F-statistic):**	2.92e-263
Time:	22:52:49	**Log-Likelihood:**	-16550.
No. Observations:	2000	**AIC:**	3.310e+04
Df Residuals:	1998	**BIC:**	3.312e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-75.0527	45.931	-1.634	0.102	-165.129	15.024
x	33.0617	0.814	40.595	0.000	31.464	34.659

Omnibus:	126.554	**Durbin-Watson:**	1.191
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	260.528
Skew:	0.422	**Prob(JB):**	2.67e-57
Kurtosis:	4.554	**Cond. No.**	122.

1.2. How specific groups within the customer base can be used to target specific market segments?

The clustering method using the remuneration and spending scores revealed that it might. be difficult to pinpoint significant patterns in the correlation between various independent variables. Although specific characteristics enable to derive insight into the sample customer base, the targeting of the specific market segments on the K-Means Predicted clustering might prove a counterproductive strategy for Turtle Games.

1.3. How can social data be used to inform marketing campaigns?

Having tested and visualised the normal distribution of polarity and sentiment score, it could be assumed therefore that positive reviews are more frequent in occurrence because of the right-skewness of the distribution function for both reviews and summaries. Given the predominantly positive attitudes toward the company's products, the marketing campaign implemented the Turtle Games have been successful. To further inform marketing campaigns using the collected social data, it might be worth discerning more in-depth patterns in the sentiments with negative polarity to enhance customer satisfaction and retention rate.

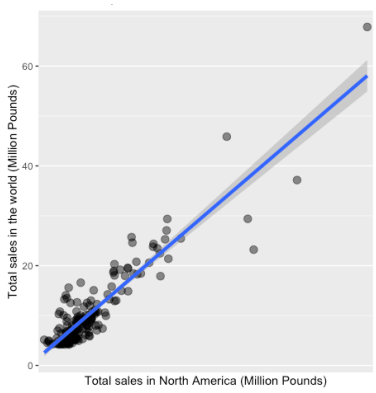
Sales [R / R Studio]

2.2.	How reliable is the data used to make assumptions about sales relationships?

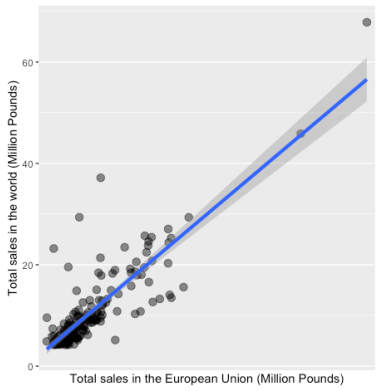
2.3.	What is/are the relationship(s) between North American, European and global sales?

While Product and Total Sales have a negative linear relationship, there is a positive linear relationship between Global Sales, Sales in North America and Europe. That is, an increase in each product negatively affects the aggregate volume of sales, whereas the growth in the volume of gaming products distributed worldwide is directly attributed to the increases he regional sales. However, since the Q–Q plots similarly show the non-normal distribution, it might be argued, that the sample data used to make assumptions about sales is insufficient or incomplete which limits its capacity to yield reliable assumptions.

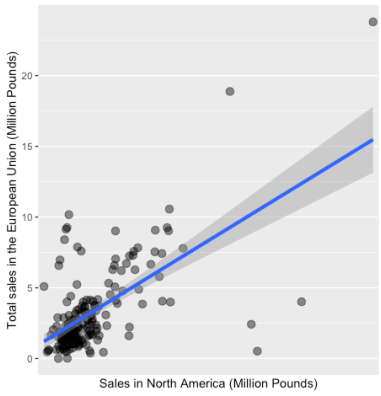
Global Sales vs Sales in North America



Global Sales vs Sales in the European Union



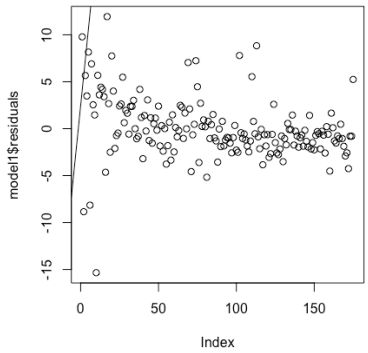
Sales in the European Union vs Sales in North America



The assumption about the limited reliability of the sales is corroborated by the simple regression models which, despite their statistical significance, appear to be inaccurate in terms of forecasting. As observed in the rather high values for the sum of squares error (SSE), the linear regression models based on the Total Sales require a better-fit forecasting model to produce more accurate insights about sales relationships.

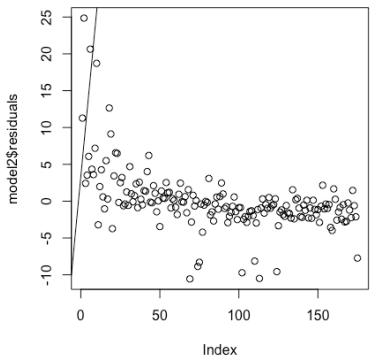
MODEL 1: Total Global Sales ~ Total Sales in North America

The sum of squares error (SSE) = 1845.812



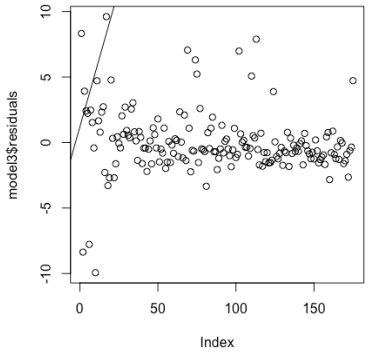
MODEL 2: Total Global Sales ~ Total Sales in Europe

The sum of squares error (SSE) = 3217.935



MODEL 3: Total Sales in Europe ~ Total Sales in North America

The sum of squares error (SSE) = 1016.824

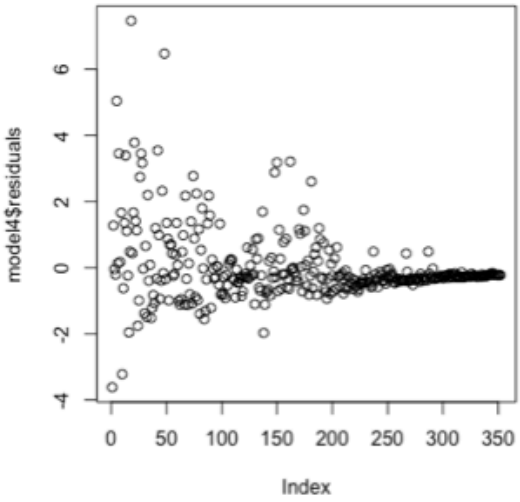


The multiple linear regression model yielded higher accuracy and statistical significance as indicated by the P-value and the adjusted R-squared. From this, it could be inferred that there is a strong positive correlation between Global Sales (dependent variable) and Sales in North America and Europe (independent variable).

MODEL 4: Global Sales ~ Sales in North America + Sales in Europe

P-value = 0.00453

The adjusted R-squared = 0.9685



The predicted values with confidence intervals roughly correspond to the actual values observed in the 'sales_per_product' dataset, which signals the forecasting accuracy of the multiple linear regression model.

PREDICTED VALUE = 71.47 Million Pounds

	Global_Sales = 67.85		
	FIT	**LOWER LIMIT**	**UPPER LIMIT**
NA_Sales = 34.02 **EU_Sales = 23.80**	**71.46857**	**70.16242**	**72.77472**

ACTUAL VALUE = 67.85 Million Pounds

LSE_DA301_assignment_Weeks 4 t... × | sales_per_product ×

Filter

	Product	Total_NA_Sales	Total_EU_Sales	Total_Global_Sales
1	107	34.02	23.80	67.85

Limitations and Further Suggestions

The linear regression and K-Means clustering, added by the profound Natural Language Processing mapping of customer sentiments, revealed several patterns in customer demographics and attitudes. Future marketing campaigns undertaken by Turtle Games should explore the possibilities of social data accessed via open-sourced platforms or by collecting customer feedback or 'willingness to recommend' questionaries.

Having adjusted and tested the accuracy of the regression models built throughout the data analysis, the Report has expressed confidence in the multiple linear regression model on the ground of its goodness of fit, statistical significance and accuracy of predictions. Hence, the Turtle Games stakeholders can curate their marketing and sales strategy to achieve a greater volume of sales on the regional level which will translate into higher sales performance worldwide.