Turtle Games is a global games manufacturer and retailer of both its own products as well as selling products manufactured by other companies. An analysis has been undertaken to answer some key questions (listed below) which will allow for strategies to be produced to improve sales performance.

- How customers accumulate loyalty points;
- How customer base groups can be used to target specific market segments;
- Social data usage in marketing campaigns;
- Impact of product on sales;
- Data Reliability
- Any relationship between N.American, European, and global sales.

Approach / Findings using Python & R

The Sklearn library alongside other packages (**Figure 1**) was heavily used within the Technical to perform Regression, create models and depict this in visualisations. Regression Analysis was used to understand the how loyalty points are accumulated. The data set was first prepped to ensure uniformity and unnecessary columns were dropped to improve interpretability and data quality. A simple linear regression model was used to see if a relationship exists between the independent variables; age, remuneration and spending with the dependent variable; loyalty points of which then Multiple Regression could be conducted. The OLS Model was used to estimate a linear regression model and to fit a linear equation to observed data. **Figure 2** shows that age does not relate to loyalty points therefore it was not further explored. Overall it can be seen in **Figure 4** that a strong positive correlation exists between remuneration, spending scores and loyalty points which concludes that spending at Turtle games needs to occur in order to accumulate loyalty points.

K-means clustering was carried out to group similar customer points to target marketing segments. The elbow and silhouette method were used to identify the K value (**Figure 5**) which both returned an output of 5. This figure was used alongside the values of 4 to create pairplots with Seaborn and fit different models to ensure all possible clusters were accounted for. **Figure 6**, shows the 5 clusters visualised. Overall the 5 clusters identify customer sub groups.

To understand more about approaches to marketing, NLP was applied to customer reviews to analyse customer sentiment. The libraries were first imported (**Figure 7**) and a new df was created focusing on the review and summary columns. Pre-processing tasks such as Tokenisation and Normalisation were carried out including the removal of empty rows(**Figure 8**) and transformation to lowercase. The Lambda function was used in conjunction with the apply() and join () methods to convert each element within the string to lowercase as well as join the words of each element into a string. This allowed for efficiency as it was achieved within one code rather than separate ones. Following from this the punctuation was removed (**Figure 9**) to reduce the complexity of the data and allow for efficient analysis. The NLTK Library was heavily used with NLP, beginning with tokenising the data using the Punkt module. It is essential to perform this step to understand the key words used in customer reviews. Identifying key words and calculating the sentiment score allows us to understand how customers feel about Turtle Games. The frequency of the words was calculated to view the most used words and was then visualised using a bar chart and word cloud (**Figures 10 and 11**). The sentiment polarity and subjectivity were then calculated to analyse deeper into the overall sentiment of customers primarily using Textblob (**Figure 12**). It was important to do this as certain words identified during tokenisation such as “well” and “would”

are not clear in terms of context. Both of these words can be used to express negative review or positive, therefore by calculating the polarity we can understand more about the context.Generate_polarity and generate_subjectivity functions were applied to both the summary and review columns. This allowed for scores to be calculated to figure out if the overall sentiment is positive, negative or neutral. The scores were then displayed using a histogram to see whether the comments are biased to a particularly strong sentiment. The usage of a histogram chart aids in assessing the bias visually. Following from these the top 20 positive and negative comments were derived and collated into a new data frame. This will not only help understand what customers love and in turn identify Turtle Game strong points but also allow improvement strategies to made using the negative comments.

A series of Histograms, scatterplots and boxplots were created using ggplot, qplot and tidyverse in R to gain insights into sales data. As can be seen in a few examples **Figures 13-15**, the sales generated between different genres across different continents. The visualisations created as a result can help inform various aspects of the business such as the marketing department. For example shooter games may be bought by more people in Europe than N.America which then leads on to heavily directing shooter game campaigns in N.America to increase sales.

EDA techniques were used to clean and manipulate the data within R so that the reliability of the data could be determined using statistical analysis. The total sales per product was derived through the group by and sum function, and in particular global sales was the main focus. Global sales is a combination of both NA, EU and other sales it was the best fit to understand total sales generated by each product (**Figure 16**). The code grouped all global sales by product and then calculated the total generated. Descriptive statistics such as the mean, min, and max of each product was then calculated to create visualisations. The secondary part of this involved testing the normality of the data using qq plots and the Shapiro Wilkes Test across each sales column (**Figure 17 -18**). As can be seen within the qqplot the points deviate upwards from the red line, indicating that the data is not normally distributed. This was also confirmed within the Shapiro Wilkes test. Outliers were determined through these visualisations and can be seen more clearly in the comparisons between various sales data (**Figure 19**). In addition to this, a relationship was identified between the different sales regions. They are strongly positively correlated to one another.

Takeaway

More data needs to be collected on the exact products that are being sold, the marketing streams already in use and the age demographic of the customer base, for example the effectiveness of the current marketing campaigns could be measured across the regions and different age groups to analyse response rate.

Figure 1

```
# Install the statsmodels package.
!pip install statsmodels

# Import the necessary libraries.
import numpy as np
import pandas as pd

# Visualisation
import seaborn as sns
import matplotlib.pyplot as plt

# The statsmodels
import statsmodels.api as sm
from statsmodels.formula.api import ols


# Note: Indicates situations that aren't necessarily exceptions.
import warnings
warnings.filterwarnings('ignore')
```

Figure 2

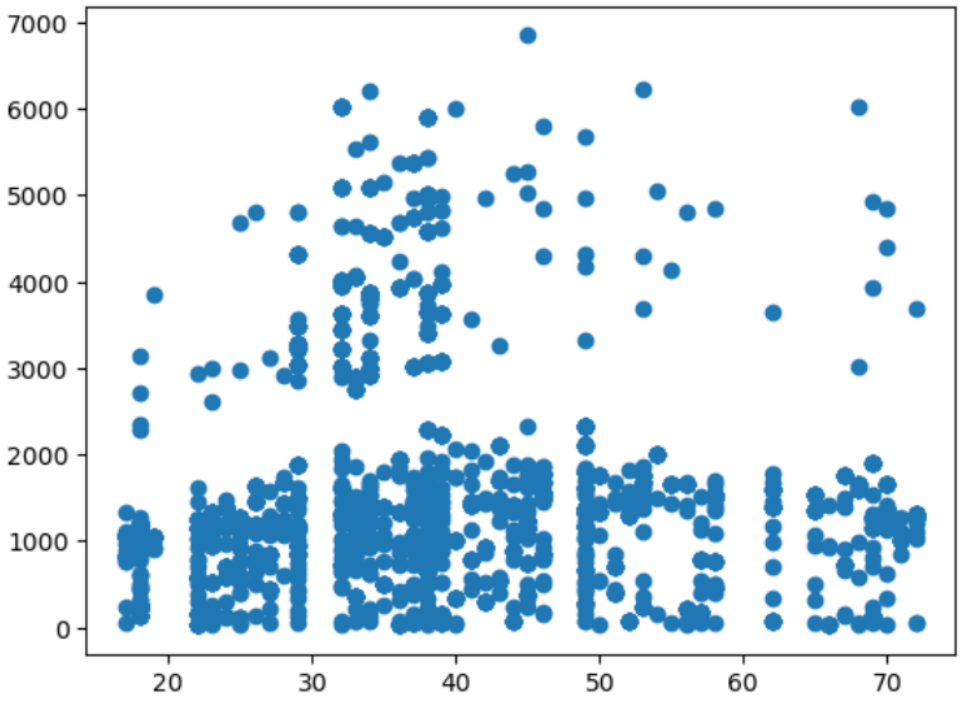


Figure 3

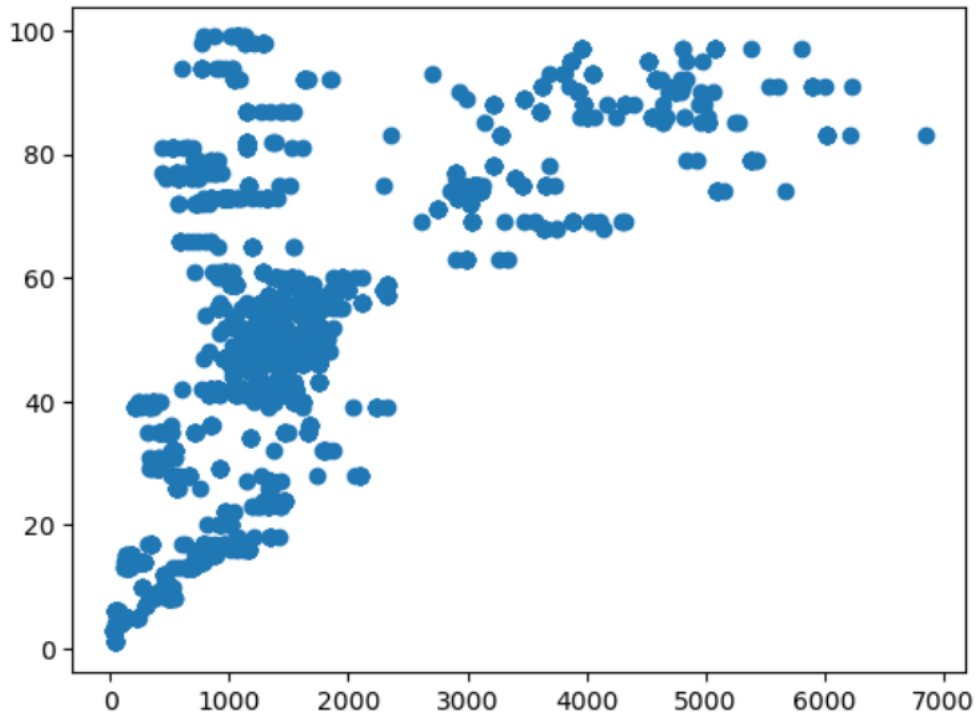


Figure 4

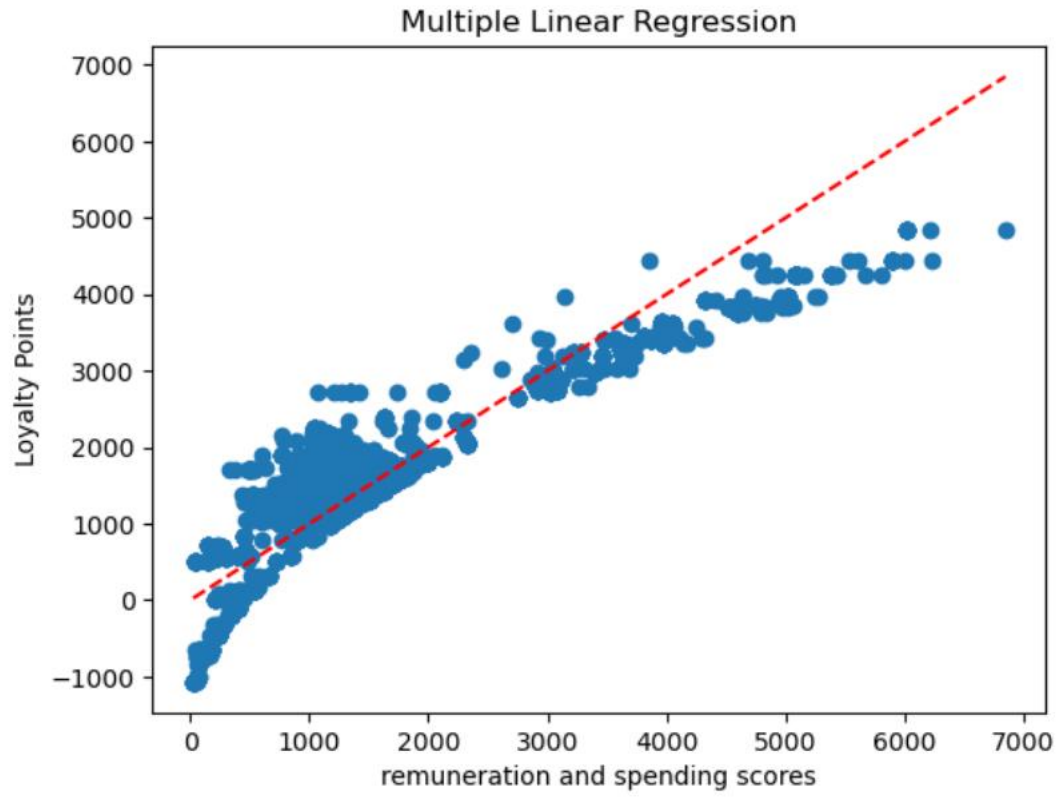
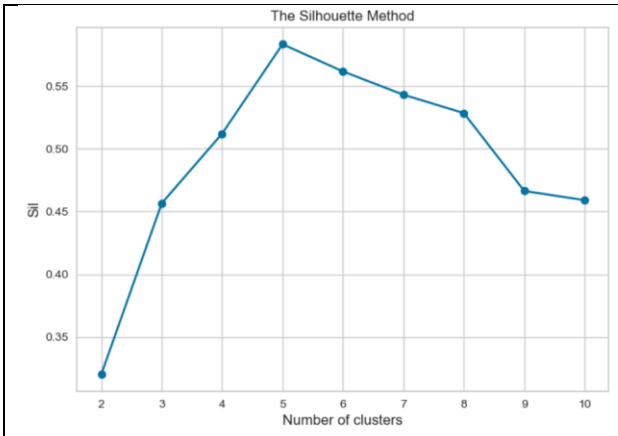


Figure 5



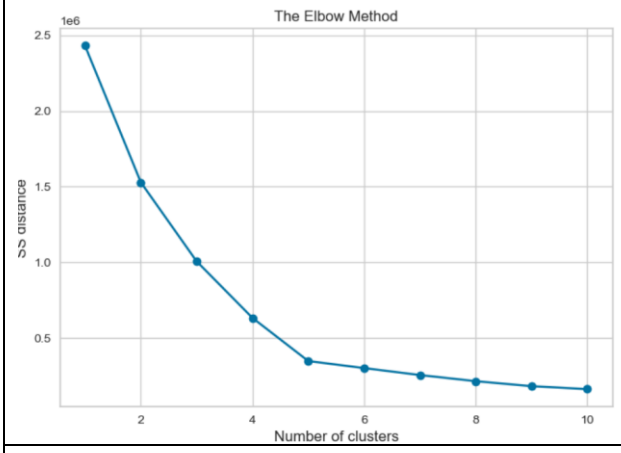


Figure 6

Figure 7

```
# Import all the necessary packages.
import pandas as pd
import numpy as np
import nltk
import os
import matplotlib.pyplot as plt
!pip install wordcloud
!pip install textblob

# nltk.download ('punkt').
# nltk.download ('stopwords').

from wordcloud import WordCloud
from nltk.tokenize import word_tokenize
from nltk.probability import FreqDist
from nltk.corpus import stopwords
from textblob import TextBlob
from scipy.stats import norm

# Import Counter.
from collections import Counter

import warnings
warnings.filterwarnings('ignore')
```

Figure 8

```
# Drop the empty rows.
df3.dropna(subset=['summary'], inplace=True)

# View the shape of the DataFrame.
df3.shape
```

Figure 9

```
# Remove punctuation.
df3['summary'] = df3['summary'].str.replace('[^\w\s]','')

# Preview the result.
df3['summary'].head()
```

Figure 10

Figure 11

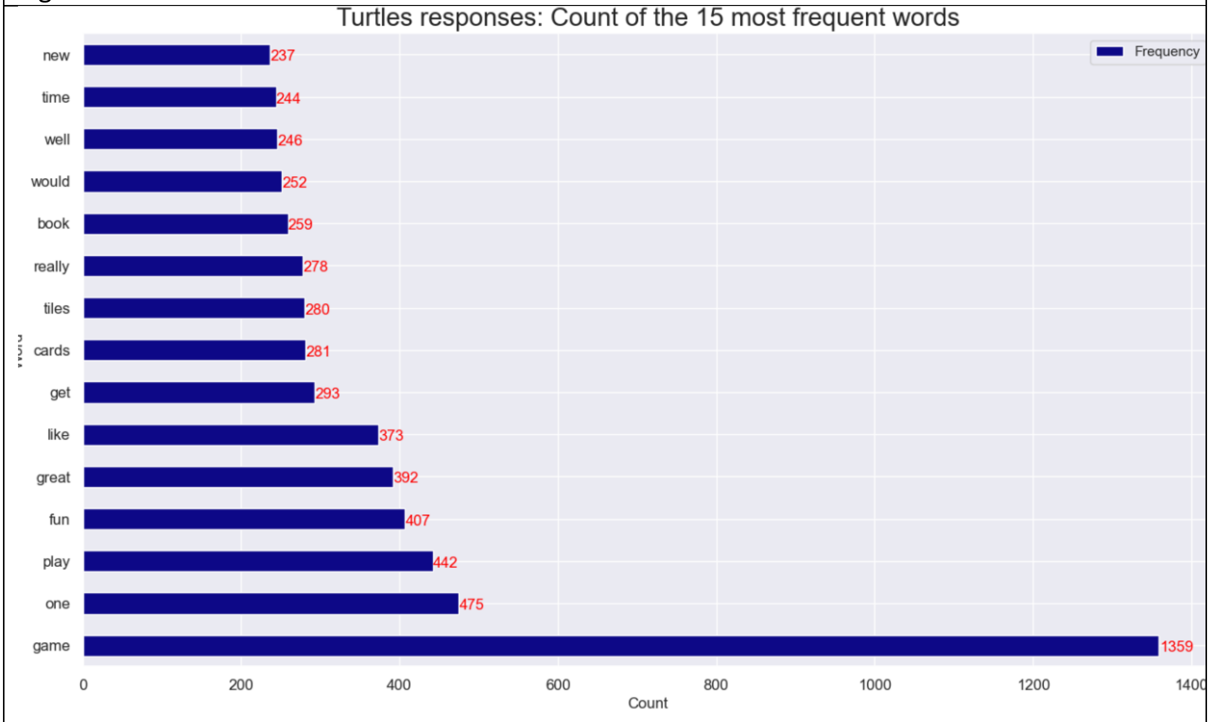


Figure 12

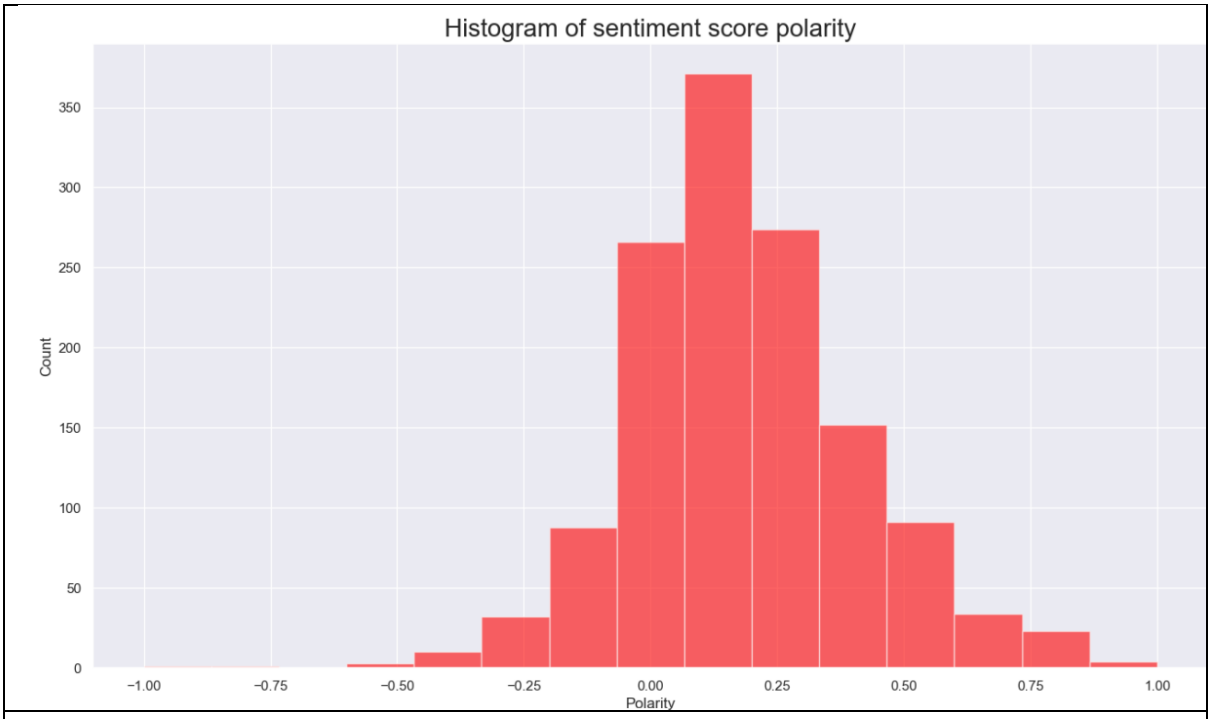


Figure 13

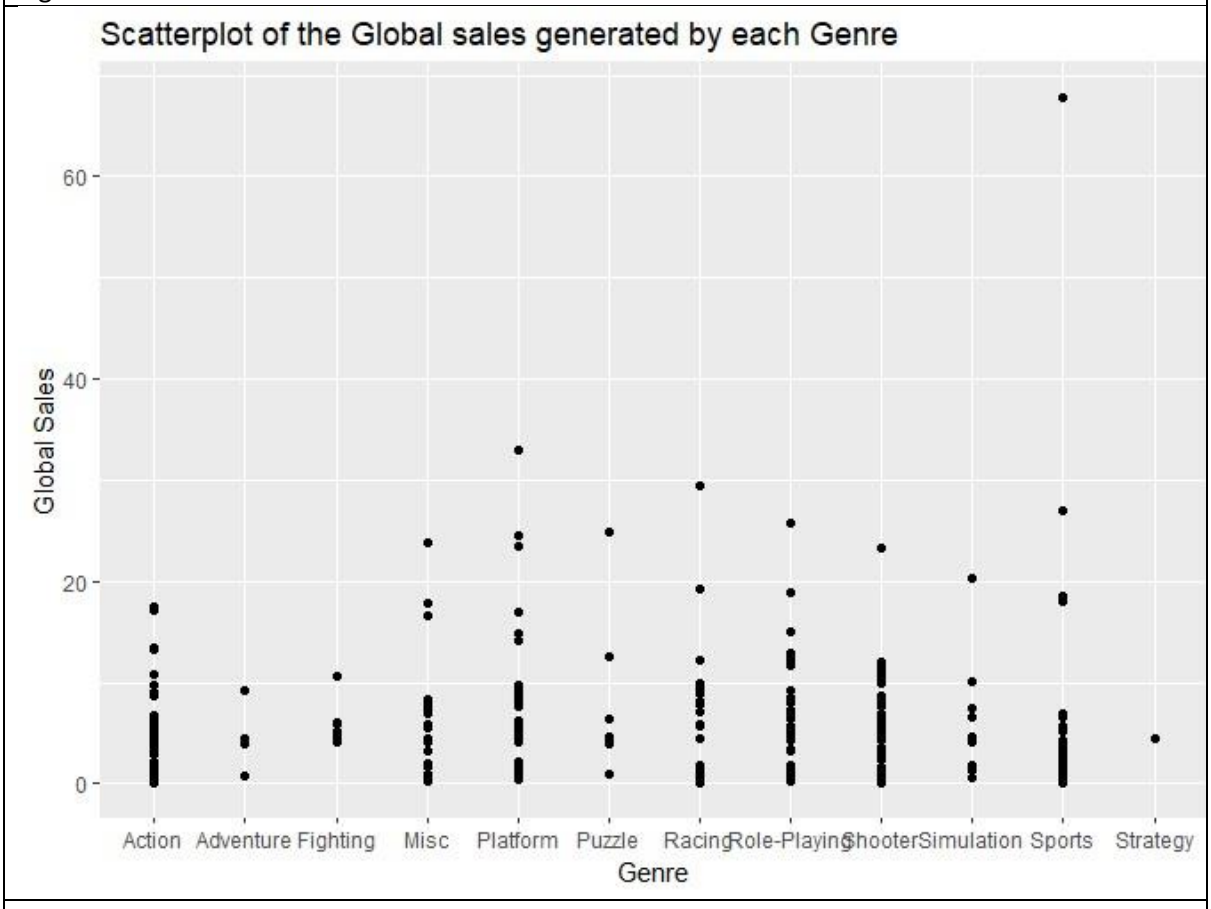


Figure 14

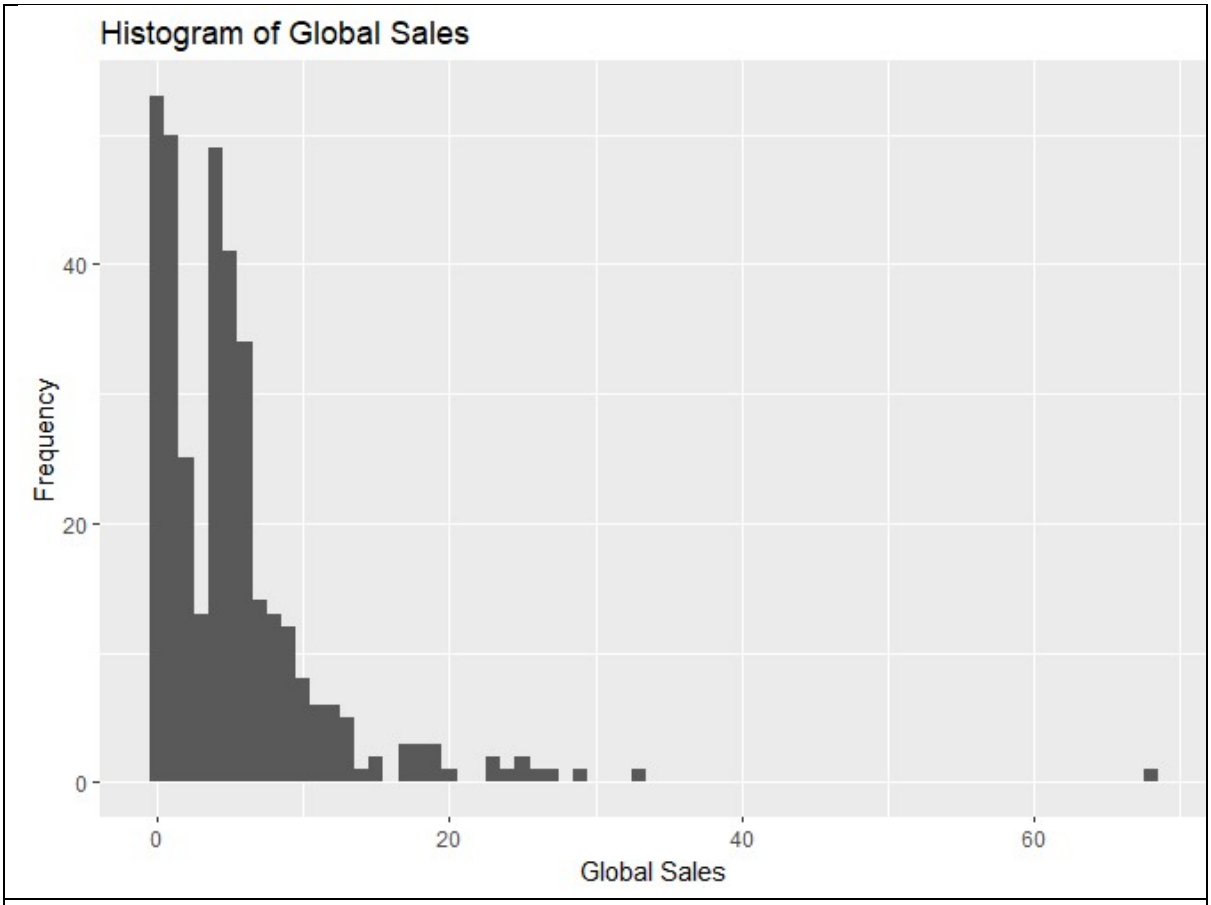

Histogram of Global Sales
Frequency
40
20
0
0
20
40
60
Global Sales

Figure 15

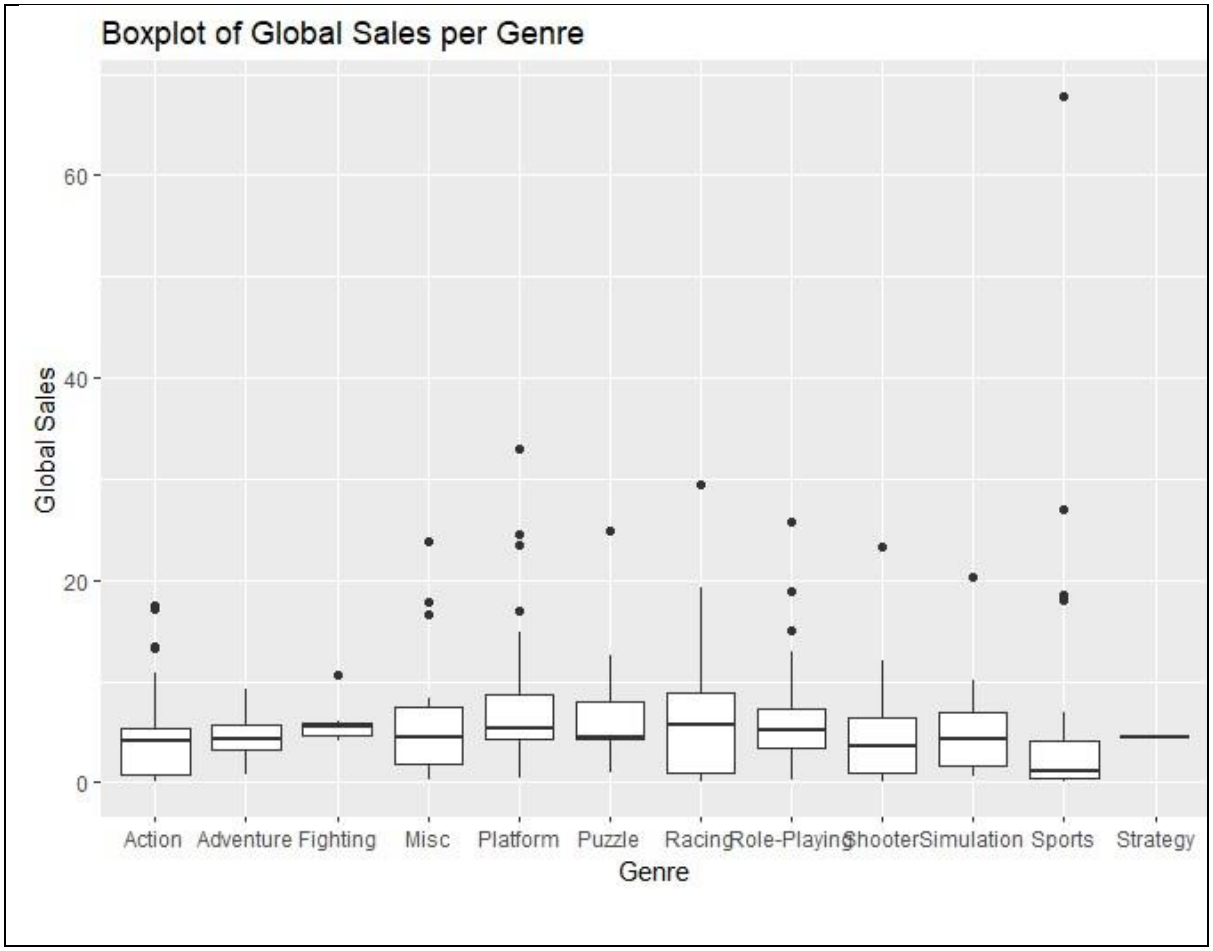


Figure 16

```
sales_summary <- sales_df %>%
  group_by(Product) %>%
  summarize(total_sales = sum(Global_Sales))
```

Figure 17

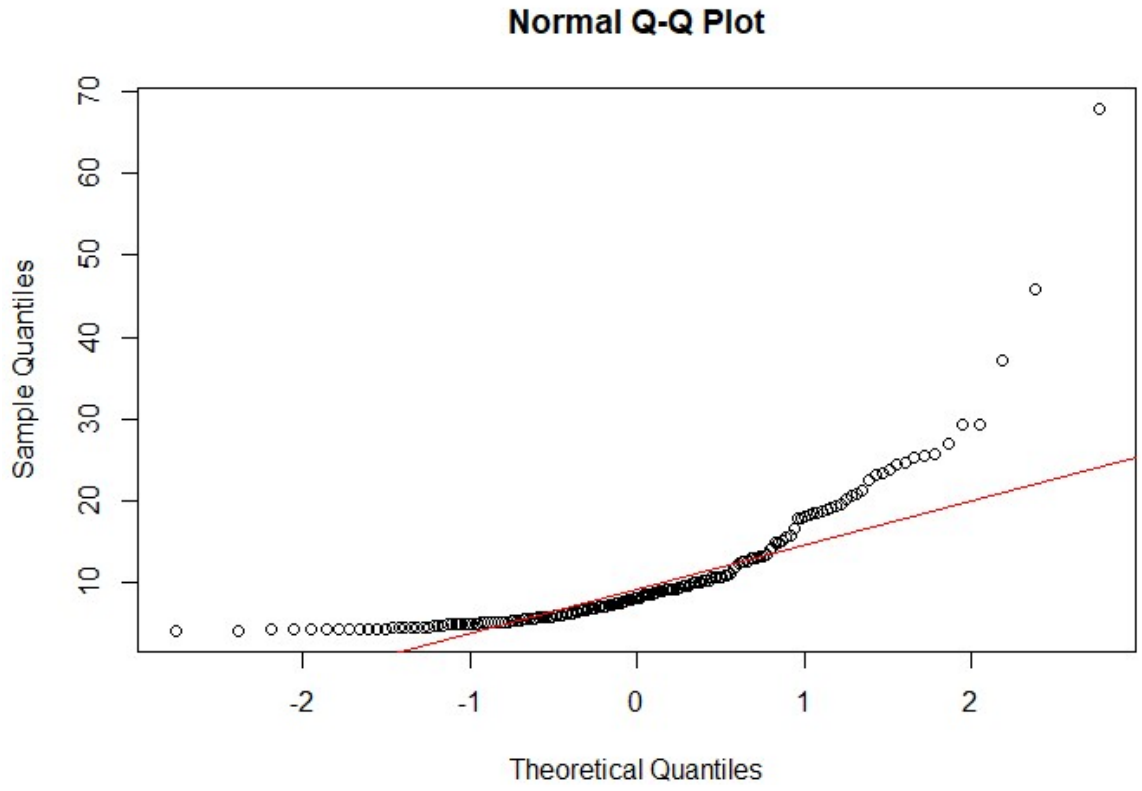

Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
10
20
30
40
50
60
70
-2
-1
0
1
2

Figure 18

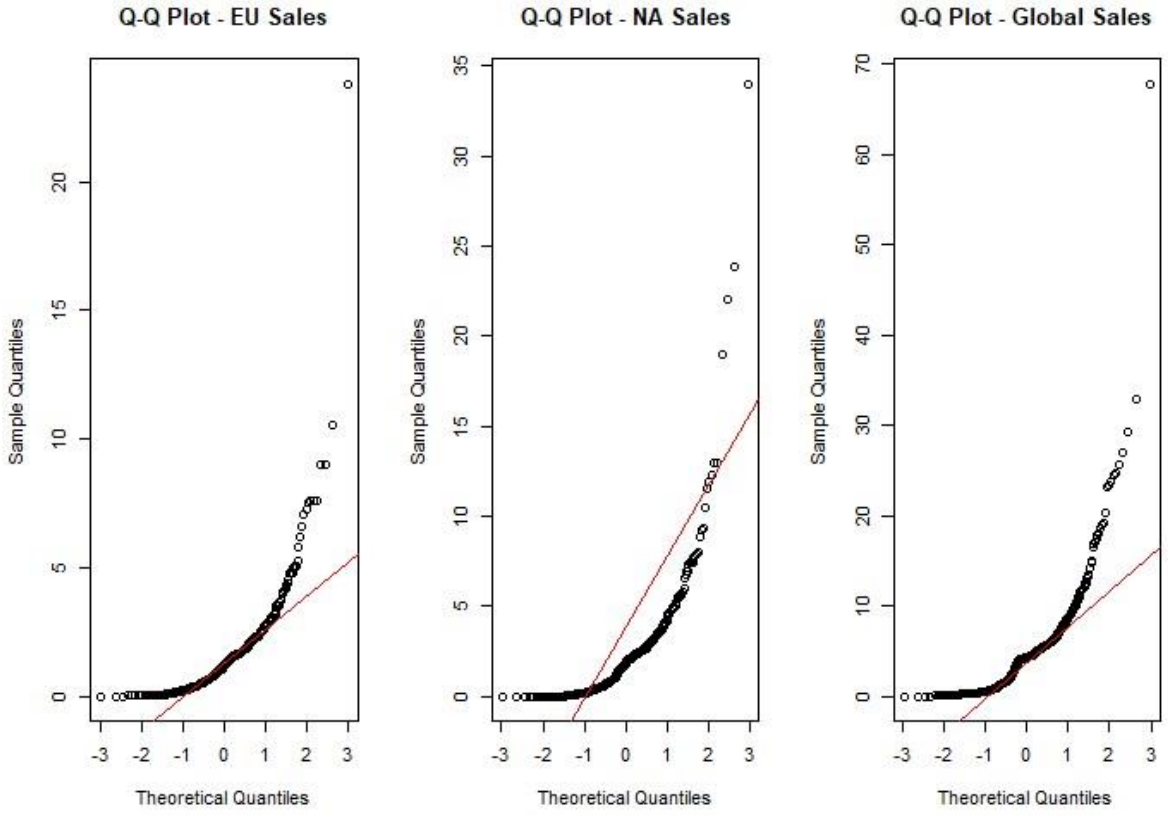
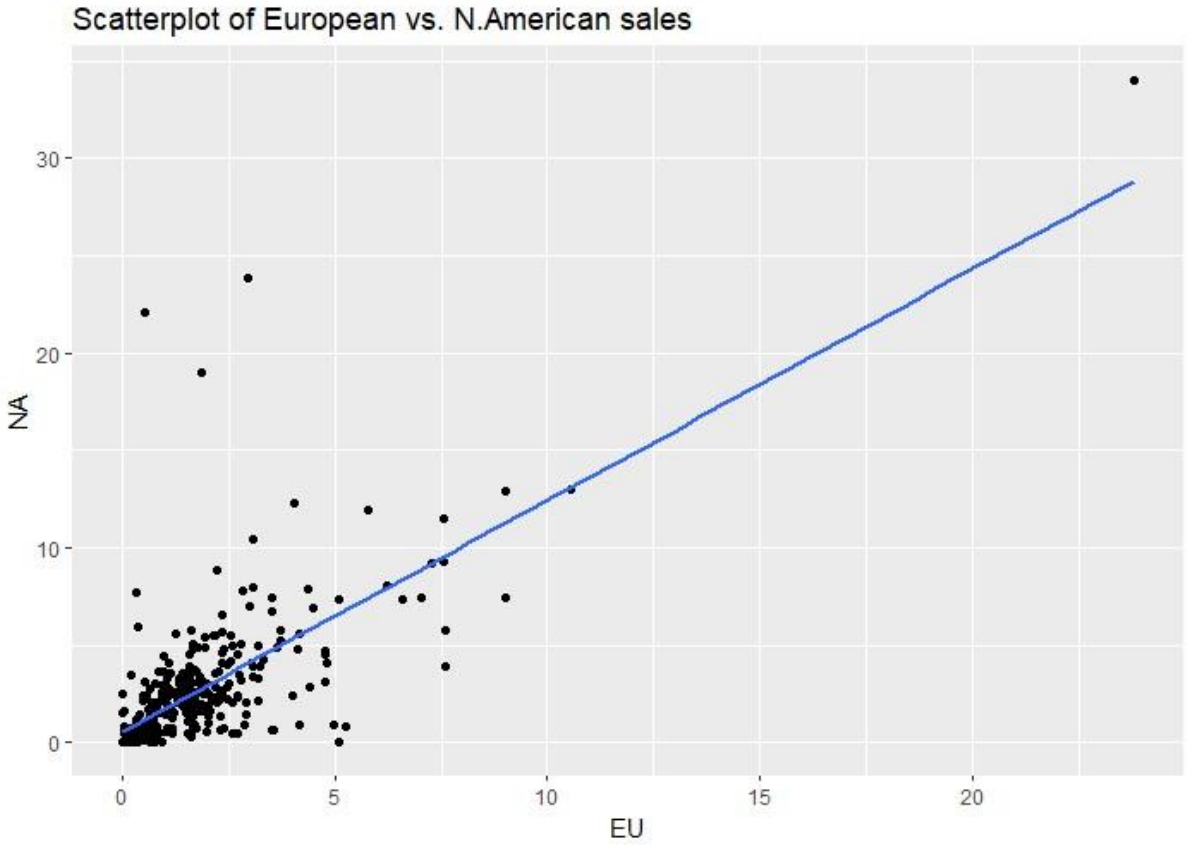

Q-Q Plot - EU Sales
Q-Q Plot - NA Sales
Q-Q Plot - Global Sales
Sample Quantiles
Theoretical Quantiles
-3
-2
-1
0
1
2
3

Figure 19


Scatterplot of European vs. N.American sales
NA
EU
0
10
20
30
0
5
10
15
20