BACKROUND OF THE BUSSINESS

Turtle Games is a game manufacturer and retailer with a global customer base. The Company manufactures and sells its products, along with selling external products produced by other manufacturers selling similar products. Its products include books, board games, video games, and toys. The data provided is collected from sales as well as customer reviews. By utilising consumer trends, Turtle Games seeks to increase overall sales performance by analysing social media reviews and understanding customer trends to optimise their sales strategies. This will be achieved by analysing customer reviews, analysing what products generate the most sales, evaluating if the data provided is reliable, exploring the correlation between sales and Continents, and drawing better predictions.

ANALITYCAL APPROACH

Exploration of data allows the deeper understating of the data set (turtle reviews), making it easier to navigate and use. The better the data is understood, the better the analysis will be.

● Necessary libraries were imported to enable analysis, the libraries include (numpy, pandas, matplotlib.pyplot, seaborn, statsmodels.api, statsmodels.formula.api, sklearn.metrics, nltk.tokenize, wordcloud, nltk, nltk.corpus, and textblob).

● Data frame was analysed to look closely at the data type, information, and missing values. Descriptive statistics offered insight into what each field represented and helped discover missing and duplicate data which were removed to enable better analysis.

● The Sales and Customer review data was subsetted to keep relevant columns to enable easier analysis, in addition columns were also renamed to separate them from the previous columns that have not been updated.

● To enable natural language processing, the review text was converted to lowercase, punctuations and stop words were removed, and a resert.index was applied to allow the words to be tokenized into lists. In the Sales data, due to Product IDs having numeric texts and being continuous by nature, they were converted to factors.

● Platforms in the sales data were categorised by console family (PS1, PS2 are considered PlayStation consoles) and console type (Handheld, Home, or dedicated) to allow for easier visualisation and reading.

● The sales products were categorised, the data frame was changed into a “long form”, and a left join was used to add categories that are going to be valuable in the analysis, such as Console Family

DATA ANALYSIS

● A graph was plotted with a linear regression for the following variables; Spending Score, Renumeration and Age against loyalty points, this was done determine relationship between variables and loyalty points.

● K-means clustering elbow and silhouette method was used to determine the different types of key customer segments present in generating sales.

● Clean reviews were tokenised into lists and strings, and word clouds were generated to determine (in broad terms) most frequent words. Polarity and subjectivity scores were computed and visualised in order to generate frequency distributions that depict the distribution of customer sentiment.

● To determine how products influence sales across console families, I created charts faceted by Console family and segmented by sales region (North America, Europe, Global).

● Data reliability was determined by testing for normality using kurtosis, skewness, Shapiro-Wilk tests, and qqplots. Multiple tests were used to resolve discrepancies and justify normalcy.

● A multiple regression model based on North American and European sales was tested to predict Global Sales.

VISUALISATIONS AND INSIGHTS

● To enable a more straightforward visualisation of the data and observe the relationship between variables, a scatter plot was created to represent the variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. The scatter graph allowed the identification of outliers in the data more efficiently as we as the overall trend and relationship between variables. The strength of the relationship was also visualised. In addition, a line of best fit was drawn on the graph to see if there was any correlation between variables. Furthermore, Visuals have been annotated whenever possible with titles, subtitles, legends, and axes to enable easy reading.

●Identifying similar/dissimilar trends and outliers across console families is viewed using multivariate faceted scatterplots, which give a comprehensive overview of multivariate relationships across various categories.

● A word cloud is another visualisation used in this analysis. It is a graphic representation of text data in the form of tags, which are typically single words whose importance is shown by their size and colour. There is an ever-increasing need to analyse vast amounts of text produced by these systems as unstructured data in the form of text continues to experience unprecedented growth, especially within the field of social media. By visualising the word frequency in the text as a weighted list, a Word Cloud is an excellent tool for aiding in the visual interpretation of a text and is helpful in quickly gaining insight into the most important elements in a given text relating to turtle game reviews.

● To illustrate the multivariate relationships between Remuneration and Spending Score and to visualise the ideal k-means predicted clusters, pair plots were used in the k-means clustering process.

● To compare medians across various console families and identify outliers, boxplots were used.

RESULTS

Linear Regression plots for Spending Score vs Loyalty points, Renumeration vs Loyalty points and Age vs Loyalty points.

```
# Plot the graph  with a regression line
sns.regplot( x= "SS", y = "loyalty_points", data = reviews)

plt.show()
```

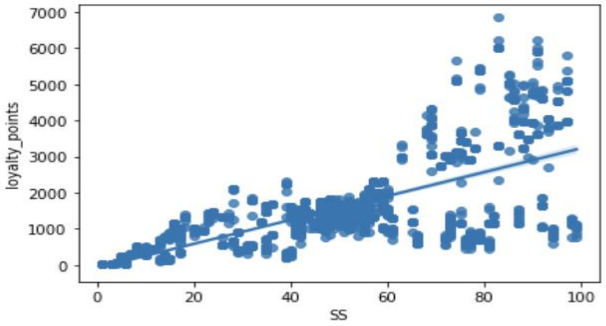

Fig 1. Spending Score vs loyalty points

```
# Plot the graph  with a regression line
sns.regplot( x= "rmtn", y = "loyalty_points", data = reviews)

plt.show()
```

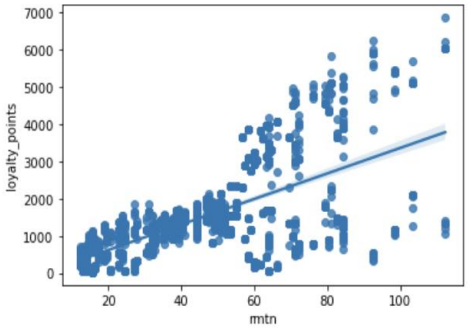

Fig 2. Renumeration vs Loyalty points

```
# Plot the graph  with a regression line
sns.regplot( x= "age", y = "loyalty_points", data = reviews)

plt.show()
```

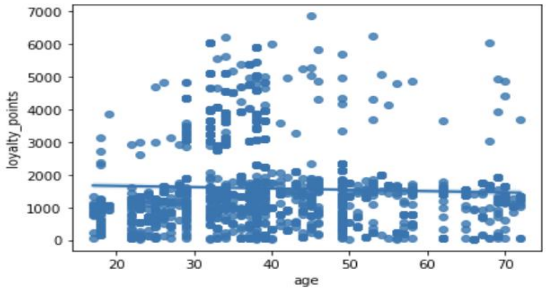

Fig 3. Age vs Loyalty points

From the Linear regression plots it can be observed that the Loyalty points have a positive relationship with Renumeration and Spending Score but a negative relationship with age.

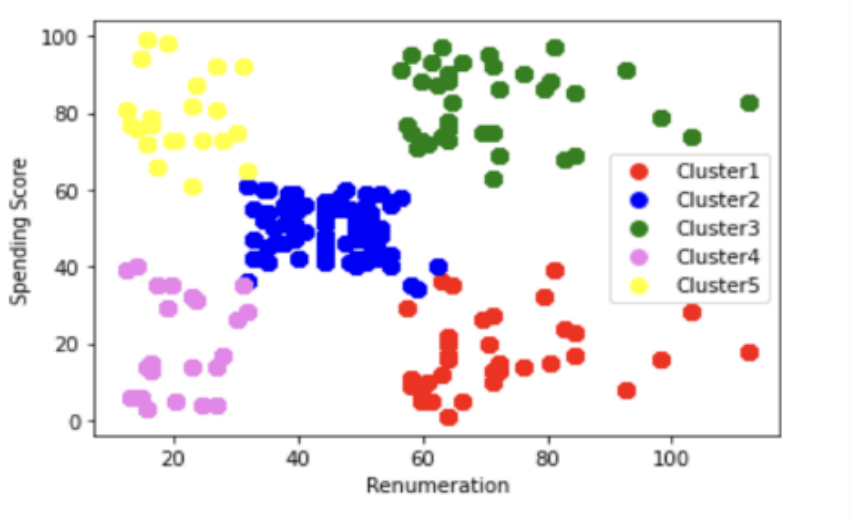

Fig 4. K-means Clustering

Figure 4 shows that there are 5 clusters in different colours, and on the visualisation, there is a key showing what colour represents what cluster. From the reading the visualisation, cluster 1 has a high Remuneration and low Spending Score, cluster 2 has a high density with a low Remuneration and high Spending Score, cluster 3 has a high Remuneration and high Spending Score, cluster 4 has a low Remuneration and low Spending Score, cluster 5 has a low Remuneration and high Spending Score. Therefore, a budget should be allocated to all clusters. However, more attention should be paid to cluster 2 and 3, and cluster 1 could be more cost-effective. Cluster 4 and 5 won't generate sales. In the long run, Turtle Games should focus on maintaining current clients as it is less expensive than acquiring new ones.

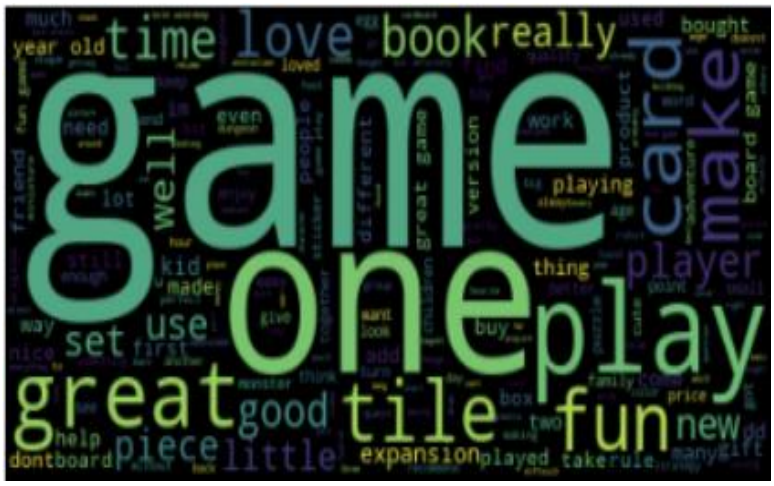

Fig 5. Review Word cloud

On the world cloud graphic produced above the prominent word is game meaning a lot of the reviews were about games produced by Turtle Games.

Out[93]:

	summary	summary_polarity
21	the worst value ive ever seen	-1.000000
208	boring unless you are a craft person which i am	-1.000000
829	boring	-1.000000
1166	before this i hated running any rpg campaign d...	-0.900000
1	another worthless dungeon masters screen from ...	-0.800000
1620	disappointed	-0.750000
144	disappointed	-0.750000
793	disappointed	-0.750000
631	disappointed	-0.750000
363	promotes anger instead of teaching calming met...	-0.700000
890	bad qualityall made of paper	-0.700000
885	too bad this is not what i was expecting	-0.700000
178	at age 31 i found these very difficult to make	-0.650000
518	mad dragon	-0.625000
101	small and boring	-0.625000
1115	disappointing	-0.600000
1015	disappointing	-0.600000
805	disappointing	-0.600000
1804	disappointing	-0.600000
1003	then you will find this board game to be dumb ...	-0.591667

Fig 6. 20 negative reviews

The top 20 negative reviews summarise how customers are dissatisfied about Turtle Games products.

- Products being boring
- Products being difficult to set up
- Poor value
- Product being a disappointment

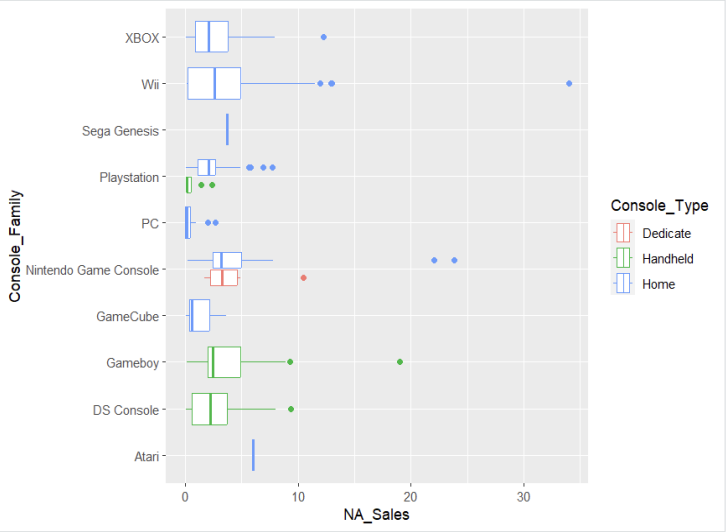

Fig 7. NA Sales by Console type

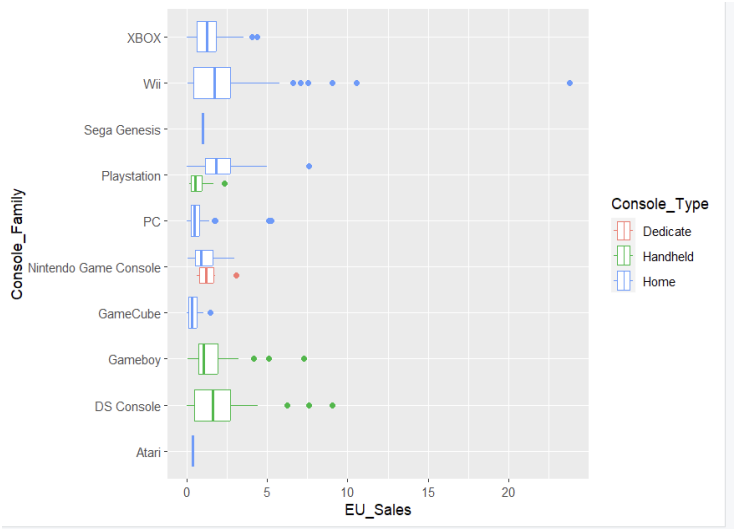

Fig 8. EU Sales by Console Type

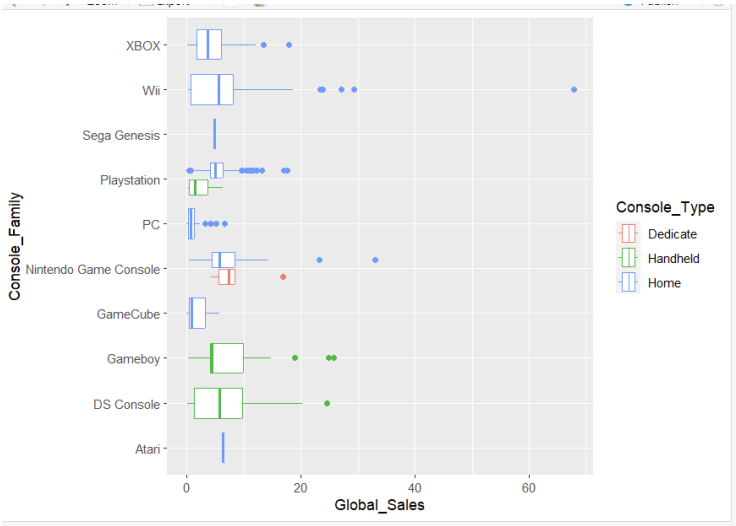

Fig 9. Global Sales by Console Type

From the Box plot, it can be observed that PlayStation, Nintendo game consoles, Xbox, Wii, DS consoles, and PC Games generate the highest sales and Atari, GameCube and Sega Genesis consoles generate the least sales for Turtle Games.

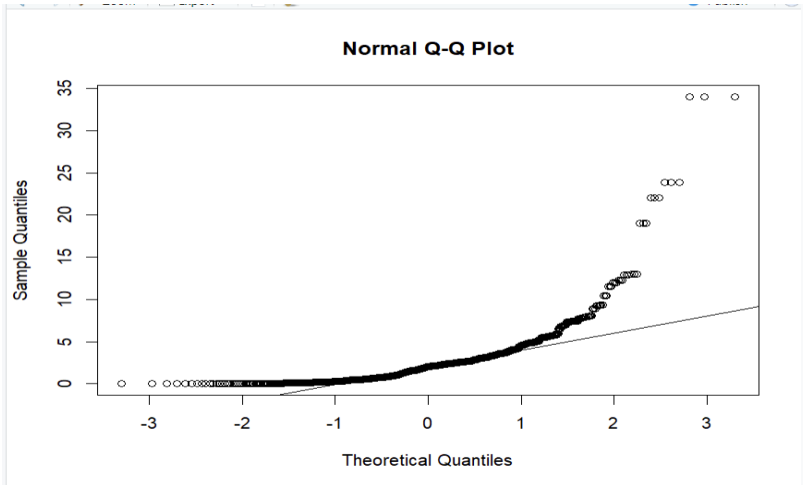

Fig 10. Determine the normality of North America Sales

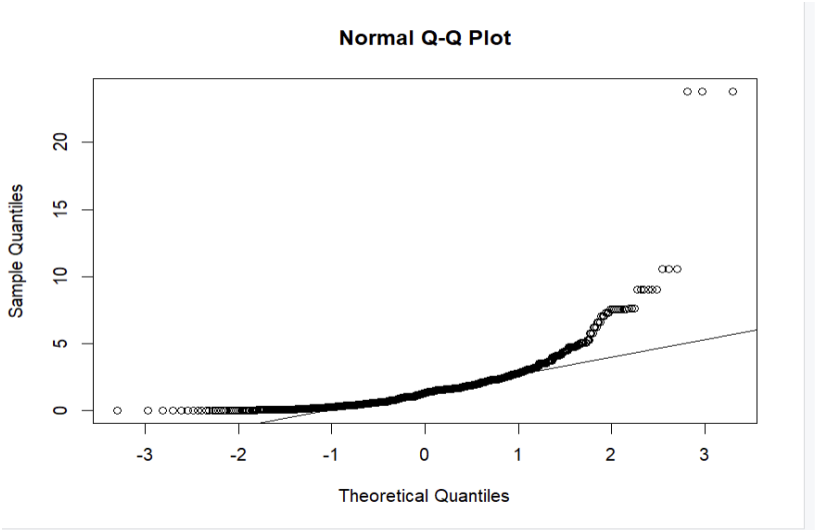

Fig 11. Determine the normality of European Sales

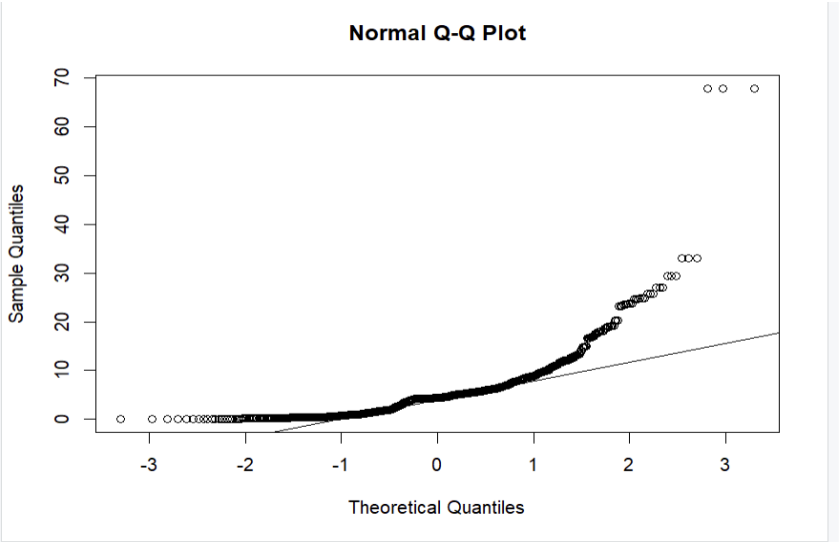

Fig 12. Determine the normality of European Sales

Figures 10,11,12 show the Q-Q plots for North American, European, and Global Sales, respectively. For the global Sales, EU Sales and North America, the Distributions did not match perfectly, as not all quantile points lie along the trend line. The deviation seen is a cause of concern, suggesting that there might have been errors during the Collection of Data.

```
data:  dfPivot$NA_Sales
W = 0.62971, p-value < 2.2e-16

> shapiro.test(dfPivot$EU_Sales)

        Shapiro-Wilk normality test

data:  dfPivot$EU_Sales
W = 0.64615, p-value < 2.2e-16

> shapiro.test(dfPivot$Global_Sales)

        Shapiro-Wilk normality test

data:  dfPivot$Global_Sales
W = 0.68079, p-value < 2.2e-16
```

Fig 13. Shapiro Test

Results obtained from the Shapiro test carried out in R show a P value < 0.05, the null hypothesis that data in the different Sale regions are normally distributed is rejected.

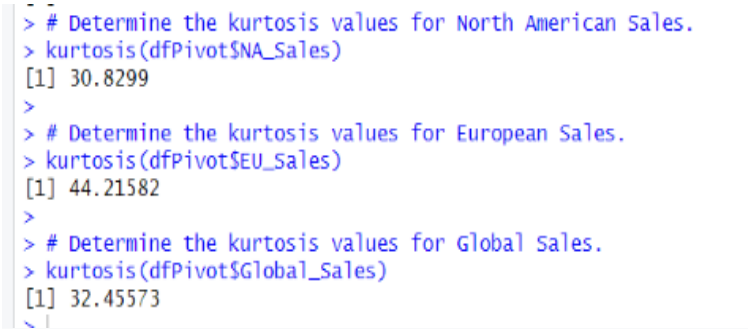

Fig 14. Kurtosis Test

Results from Kurtosis show values greater than 3 from all Sale regions, suggesting a leptokurtic distribution with a higher distribution of outliers than normally distributed data.

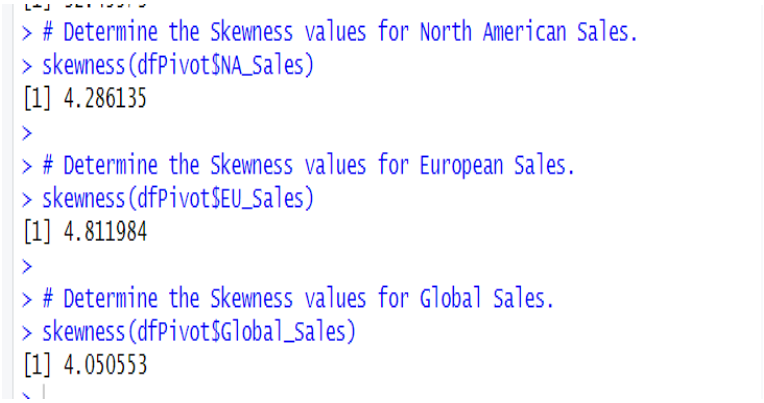

Fig 15. Skewness Test

The skewness levels for all three sales data are greater than 3, suggesting positive skewness and that the distribution is positively right skewed. Kurtosis levels for all sales regions are greater than 3, indicating that the data is exposed to outliers and will produce more extreme outliers than the normal distribution suggesting that the data is not reliable. A high positive correlation was noticed between NA-Sales and EU Sales, indicating that NA Sales and EU Sales contribute to global sales but influence each other only moderately.

RECOMMENDATIONS

● Dissatisfied customers should be contacted, and their reason should be investigated in depth via email or phone call

● Products with faults should be investigated to see if it is the manufacturer's fault or a problem with delivery handling

● Highest sold products should constantly be in stock and products generating less sales should be investigated, and strategies should be made to increase sales, this could be done collecting opinions on products and service from surveys .

●There are duplicate reviews, feedback systems need to be checked to prevent duplicates in collected data

● Utilise sentiment analysis on information retrieved from competitors and social media websites to analyse what thing they are doing to generate sales for their business.

●For more specificity, include additional fields like regions, store names, and distributor suppliers.