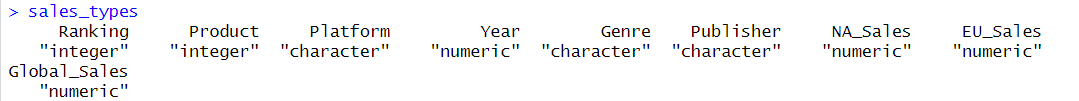
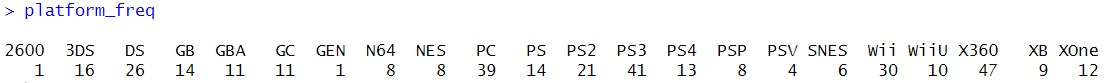
**LSE\_DA301 Assignment Report – Predicting Future Outcomes**

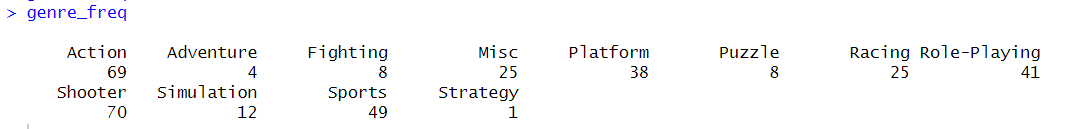
Turtle Games is a game manufacturer and retailer who sells books, board games, video games and toys produced by themselves and other companies. They have a global customer base, and are looking to increase global sales by utilising customer trends. This report intends to cover how the analysis was completed, with the attached video presentation containing the analysis that can help Turtle Games increase their global sales.

To begin with, the relevant R packages and data were installed and loaded. A null count was carried out on the data set to check for null values and any nulls were omitted from the data set. The data types were checked to confirm all values were stored as expected, and some exploratory analysis was completed to look at the number of platforms, genres and publishers there are.



The number of products per platform, genre and publisher was also explored. From this, we can see that X360, PS3 & PC were the three most common platforms, with shooter, action & sports being the most popular genres. Nintendo, Electronic Arts and Activision were the three most frequent publishers.





A screenshot of a computer screen

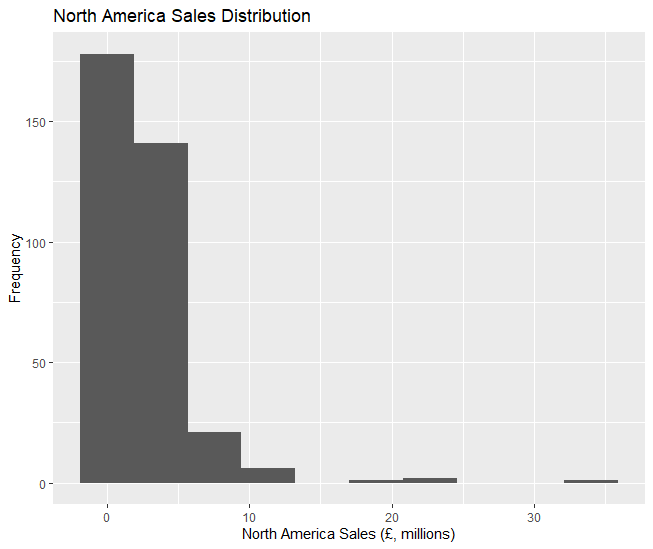
Description automatically generated

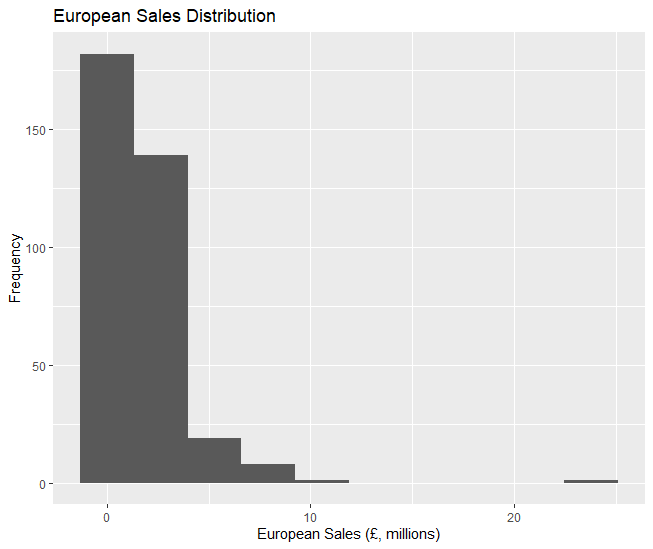
A new data frame, was created removing the unnecessary columns to explore the sales and was summarised to get the min, median, max and quartile values for each column.

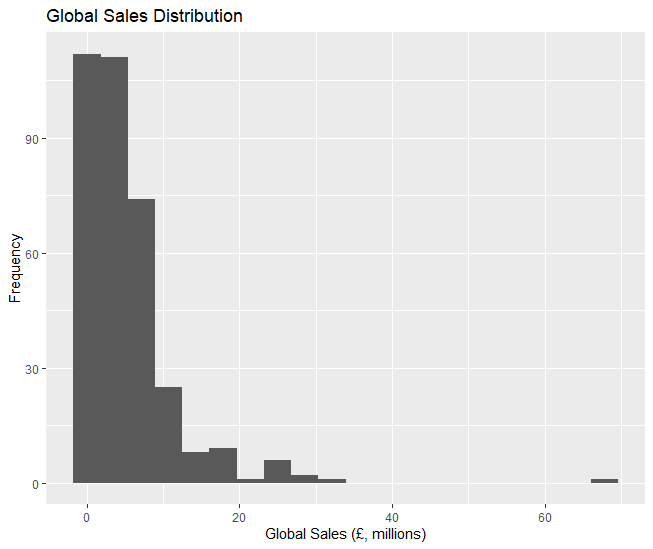
A screenshot of a computer screen

Description automatically generated

Histograms showing the distribution of sales in North America, Europe and overall (Global) were then plotted using ggplot2.







Visual analysis suggests that the North American, European and Global data does not follow a normal distribution, and has a steep positive (right) tail. This can be further reinforced by looking at the Q-Q plots below. In each case, the data points bend upwards from the reference line at each end, suggesting positive skew.

A graph of a graph

Description automatically generatedA graph with numbers and lines

Description automatically generatedA graph with numbers and lines

Description automatically generated

Furthermore, the Shapiro-Wilk normality test produces a very small p value for each sales region, suggesting that the data significantly departs from the normal distribution.

A white background with black text

Description automatically generatedA white background with black text

Description automatically generated

A white background with black text

Description automatically generated

A kurtosis and skewness value is also calculated for each region. The skewness value is relatively large for each region in the positive direction, again suggesting that the data does not follow a normal distribution and is positively skewed. The kurtosis values suggest that the tails are short with a steep central region, again demonstrating what was shown in the histograms above.

A close up of a white background

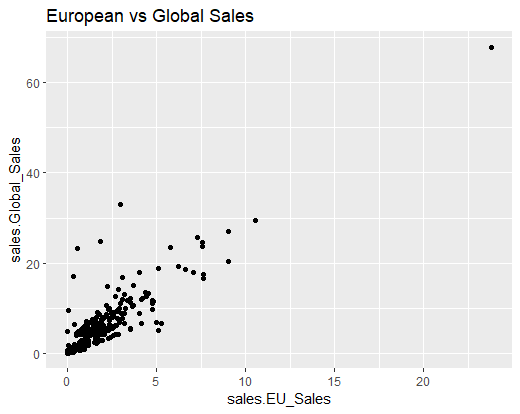
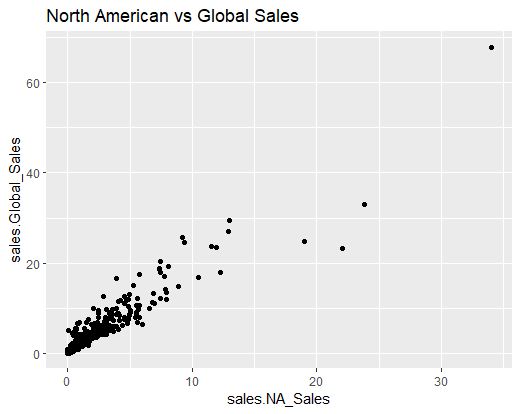
Description automatically generatedA close-up of a computer screen

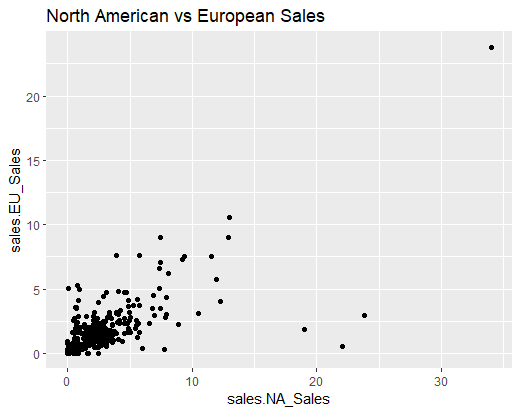
Description automatically generated

A close-up of a white background

Description automatically generated

A regression model was then built to explore the relationship between North American, European and Global sales. To begin with, simple plots of North America vs Global, European vs Global and North American vs European were produced.





We can see from each plot that a general linear trend is followed. This means we can carry out linear regression on the model. The linear regression model gives the below summary:

A screenshot of a computer

Description automatically generated

The residuals for the min, 1Q, 3Q and Median are all small, meaning the model has a good fit. The max residual value is slightly larger, but should not affect the quality of the model to an extent that warrants change. This is backed up by the Multiple R-Squared value of 0.9687, showing a high 96.87% of Global sales can be described by the model.

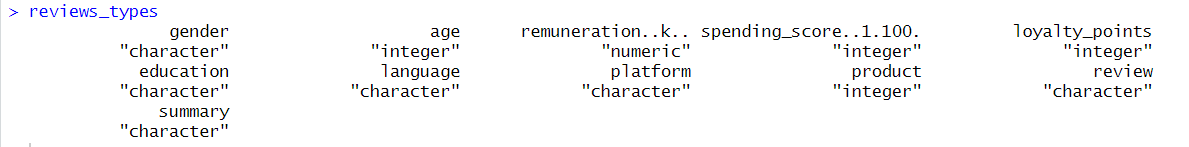
The estimate values show that for every unit increase of NA\_Sales, Global\_Sales increases by 1.15562 units, and for every unit increase of EU\_Sales, Global\_Sales increases by 1.34128 units. This suggests that European sales has a greater impact on Global sales than North American sales.

Global Sales were then predicted using the values provided by Turtle Games:

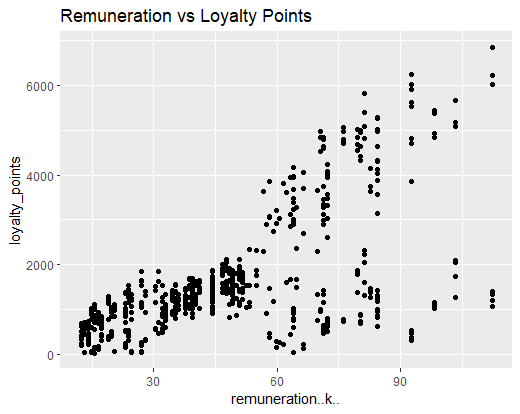
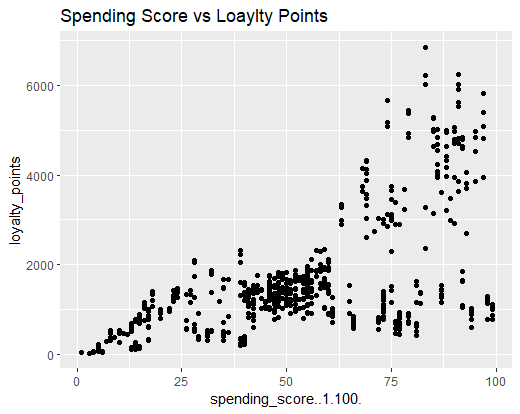
A screenshot of a computer screen

Description automatically generated

The reviews data was then imported to R, the data types checked and any null values omitted.



Plots of spending score and remuneration against loyalty points were then plotted to explore how loyalty points are accrued.



There is a clear relationship between spending score and loyalty points and remuneration and loyalty points. This led to a linear regression model analysis.

A screenshot of a computer

Description automatically generated

The residuals for the model are quite large, meaning the model may not have the closest fit. This is especially true for the large min and max scores. The fact that the median residual value is much smaller than the min and max values suggests that the model is a better fit for the middle values, however not such a good fit for the extreme values. This can be seen graphically in the graphs above – as spending score & remuneration increases, the spread of loyalty points also increases.

The estimates value suggests that as spending score increases by 1 unit, loyalty points increase by 32.9 units (3.s.f). This would suggest that the more the customer spends, the greater the number of loyalty points they accrue.

The data was then imported to Jupyter Notebook to complete further analysis using Python. The data types were checked, along with the describe function to calculate the counts, means and variance figures.

A table with numbers and a few black text

Description automatically generated with medium confidence

A pair plot of the age, remuneration, spending score and loyalty points was then generated to look for patterns that could be used for clustering to group the data.

A collage of blue and white graphs

Description automatically generated

From this, we can see age and loyalty points have no real groups or clustering opportunities against any of the other variables.

The graph suggests that there are 5 clear groups, however the elbow method was used to confirm this.

A graph with a line

Description automatically generated

Producing the elbow method graph shows a clear bend at 5, supporting the fact that there should be 5 clear cluster groups. A K value of 5 was used to create the cluster groups.

A diagram of a number of dots

Description automatically generated with medium confidence

From this, we can see that group 0 have a low remuneration value but high spending score. This means the clients do not earn a high salary, but are willing to spend a significant percentage of their salary buying goods from Turtle Games.

Group 2 also has a similar range for remuneration, with a low salary, however customers in this group spend significantly less than those in group 0. These clients are therefore low salary, low spenders.

Group 3 includes customers with a higher range of salary who also have a high spending score.

Customers of group 4 also have high incomes, however have a lower spending score at Turtle Games.

Customers of group 1 are the ‘average’ customer in terms of spending score and remuneration.

The cluster groups was then added to the original DataFrame.

A screenshot of a computer

Description automatically generated

Sentiment analysis was then carried out , and the sentiment scores and labels were added to the DataFrame.

A graph with blue bars

Description automatically generatedA graph of the sentiment labels against number of reviews was plotted, along with the sentiment score distribution.

A bar graph with different numbers

Description automatically generated with medium confidence A bar graph with different blue bars

Description automatically generated with medium confidence

The sentiment score distribution was also plotted for groups 0, 3 and 4 to help the marketing teams spot patterns they could take advantage of. Finally, a word cloud was created using the most commonly referenced words in the reviews.