Turtle Games: Predicting future outcomes

Introduction

Turtle Games is a game manufacturer and retailer with a global customer base. The company manufactures and sells its own products, along with sourcing and selling products manufactured by other companies. Its product range includes books, board games, video games, and toys. The company collects data from sales as well as customer reviews. Turtle Games has a business objective of improving overall sales performance by utilising customer trends.

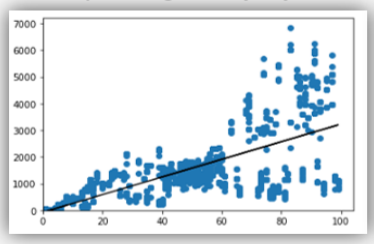
They would like to know:

- How customers accumulate loyalty points.
- How groups within the customer base can be used to target specific market segments.
- How social data (customer reviews) can be used to inform marketing campaigns the impact each product has on sales.
- How dependable the data is.
- What the relationship(s) is/are (if any) between North American, European, and global sales.

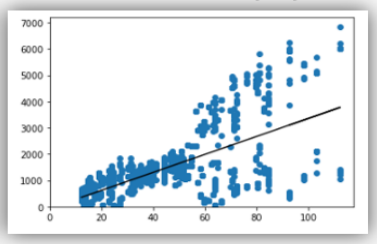
Method and Results

When looking at the possible correlation between the independent variables and dependent, which is loyalty points, a stronger correlation is on spending and remuneration. However, age is having less impact on loyalty points increase. Overall, when looking at the R-sq. that describes total variability of y explained by the variability of x, we can see that the total percentage per each case is not even close to 50%, age and loyalty points having a 0.002%. Also, when looking at the coefficient of x that describes the slope of the regression line by how much the response variable y changes when x changes by one unit, we can see that particularly on age/loyalty case we get a negative number, showing no relationship. The P value for both remuneration and spending score vs loyalty points are zero, thus, being highly statistically significant at explaining the variation in the dependent variable.

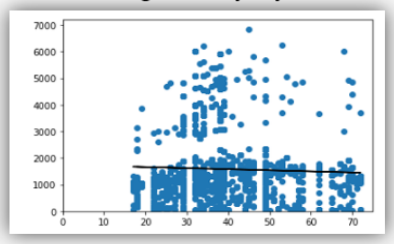
spending vs. loyalty



remuneration vs. loyalty



age vs. loyalty



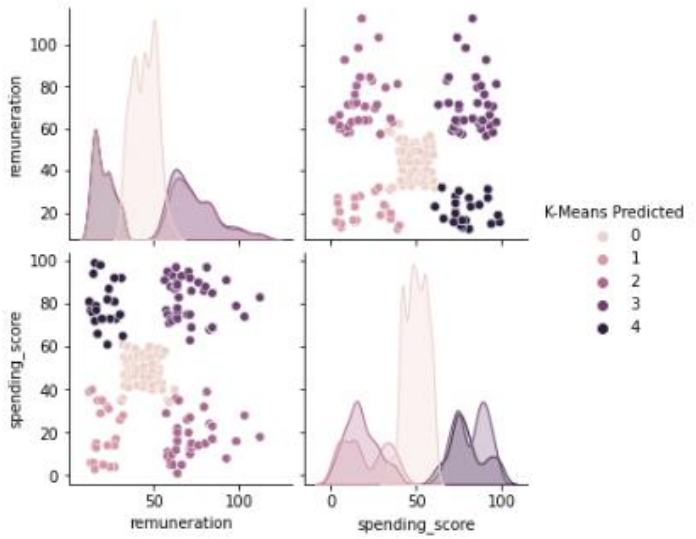
Notes:

R-squared:	0.452
x	33.0617
P	0.000

R-squared:	0.38
x	34.1878
P	0.000

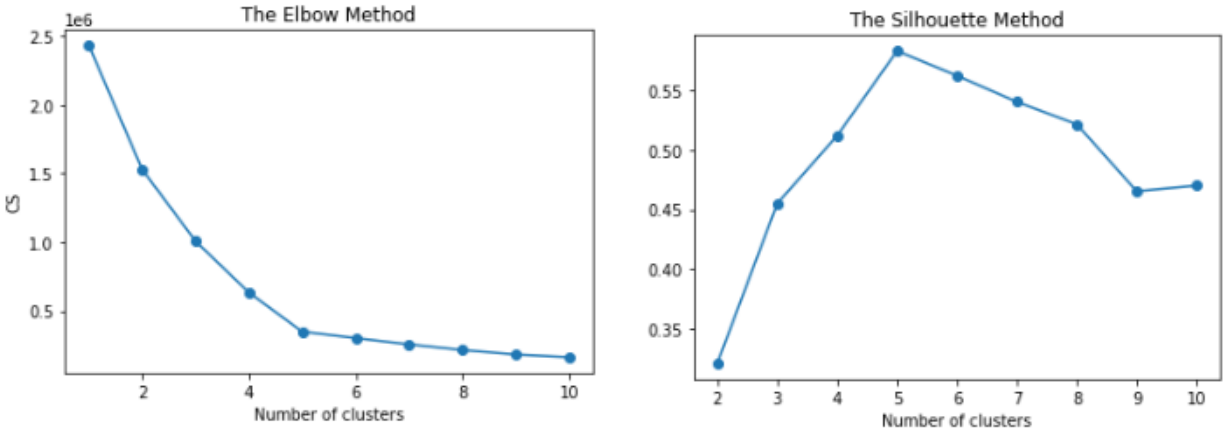
R-squared:	0.002
x	-4.0128
P	0.058

Furthermore, to target specific market segments, remuneration and spending scores were plot using a pairplot to find the correlation and clusters. A visibility of potential five clusters was detected.



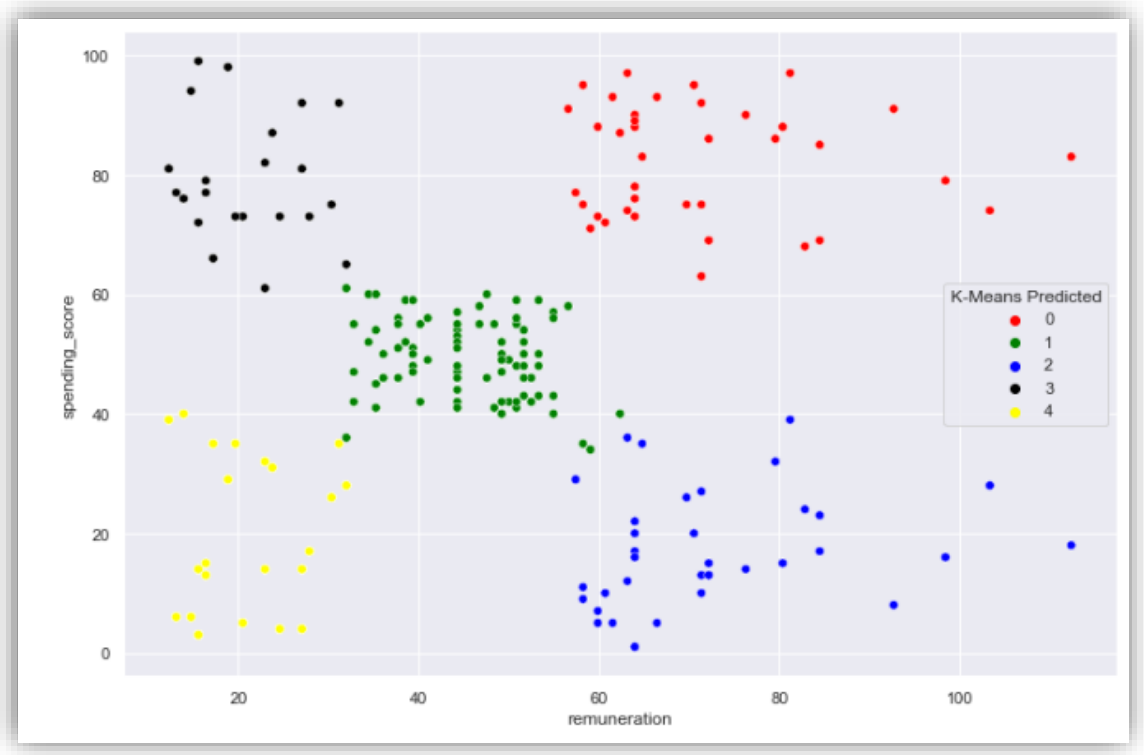
Silhouette and Elbow method was used to help find the optimal number of clusters for k-means clustering. When analysing the elbow method, it can be clearly seen that the value metric initially changes rapidly and then is slowing down. When the point of rapid change stops, this is when the elbow point can be detected. In this case scenario the optimal number of clusters is showing five.

By looking at the silhouette method, which, represents the mean silhouette coefficient over points, the maximum point indicates five.



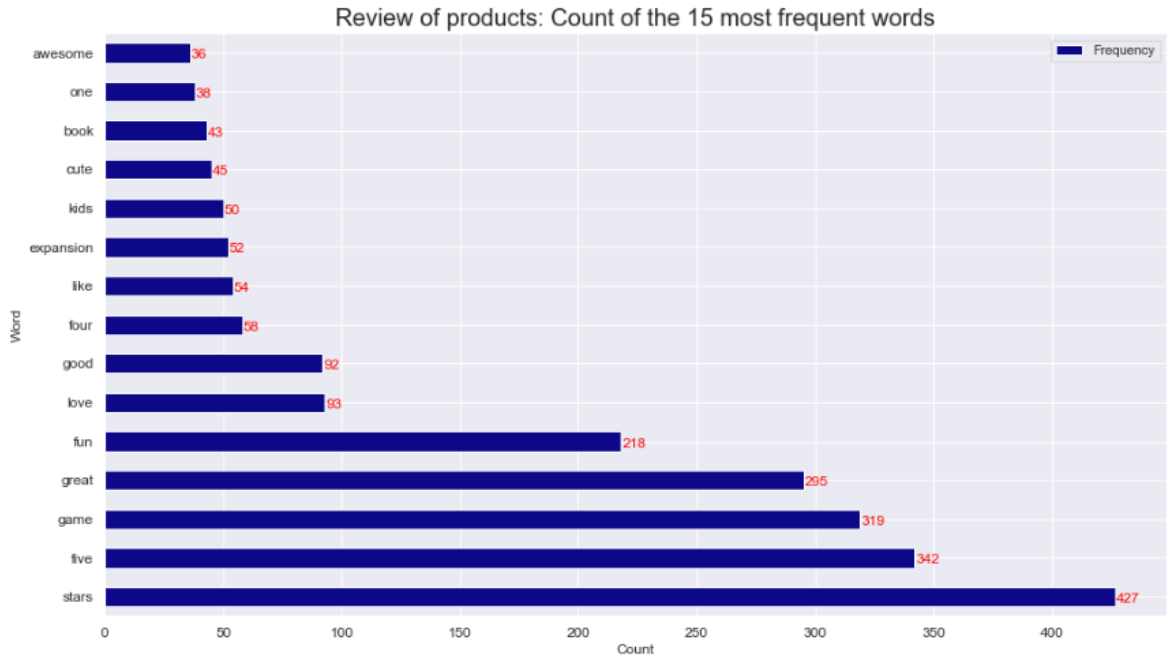
An investigation has been made on testing three values for k and check whether the accuracy improves. It has been concluded that the most optimal value is five as it provides clearer separation between groups.

By fitting the final model and looking at the number of observations per predicated class, cluster one shows to have the biggest number of observations, followed by 0,2,4 and 3. Therefore, the final plot shows the following interpretation:

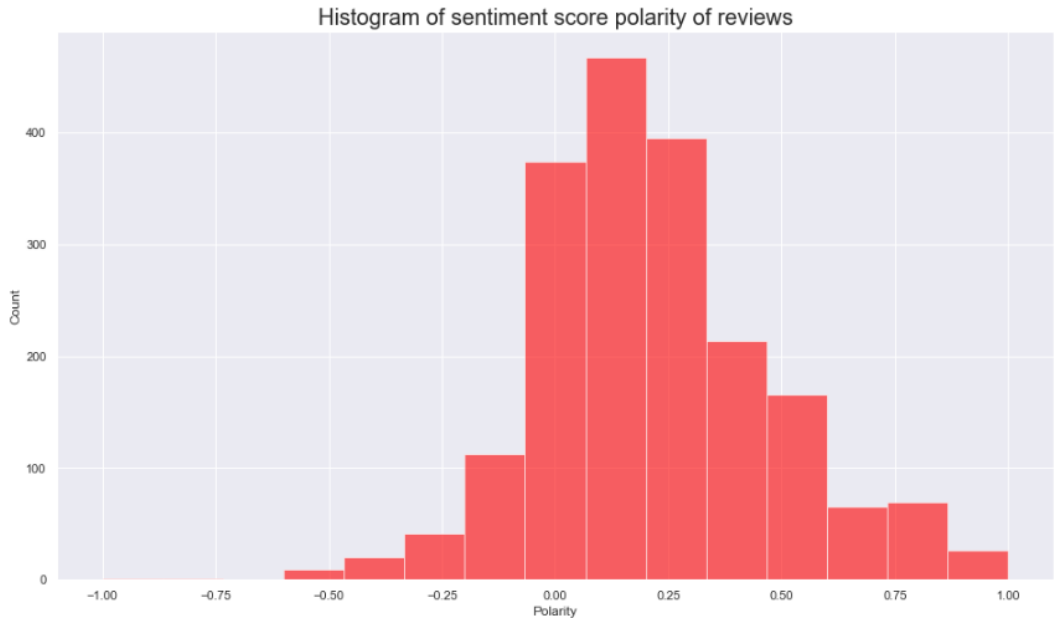


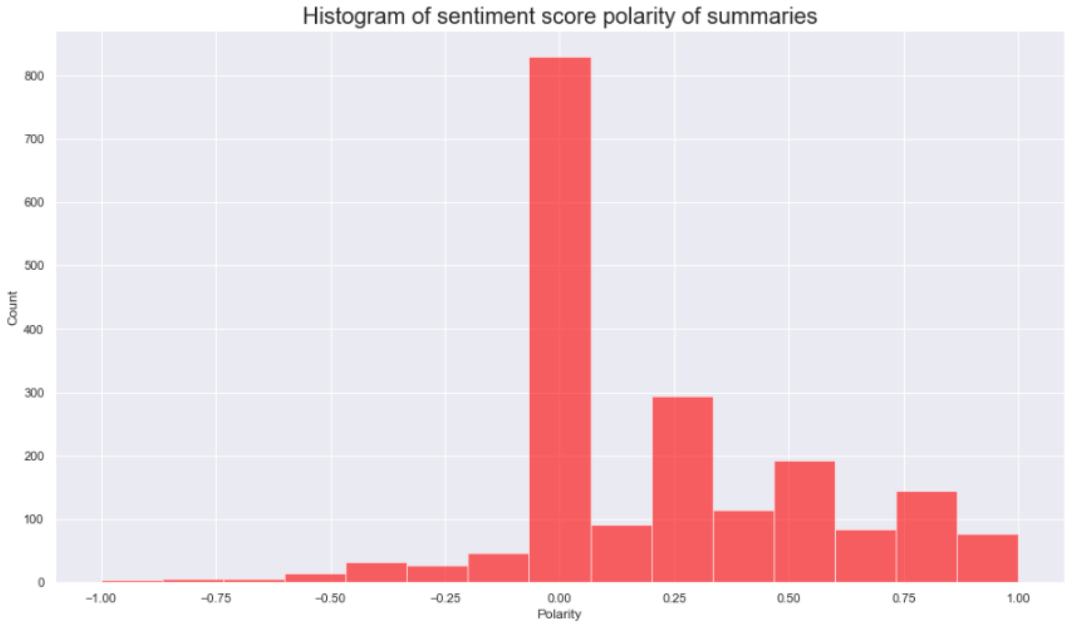
- Green cluster with the biggest number of data points shows average income and average spending group.
- Red cluster shows higher income and higher spending score.
- Blue cluster - higher income and lower spending score.
- Yellow cluster - lower income and lower spending score.
- Black cluster - lower income and higher spending score.

To support the marketing department, NLP was applied to measure the sentiment of the customers and analyse the most frequently used words. The data must go through pre-processing to bring the most accurate results. Review and Summary columns text has been changed to lowercase, punctuations and duplications were then removed. Moreover, the text has been split into individual words and it has also been filtered by English stopwords, which, do not bring any significance to the data. This technique has helped identify the most popular used words and by looking at top fifteen words, seven words presented a positive emotion. It is also important to notice that there were no negative words detected. ‘Stars’ and ‘Five’ were the most popular words.



By looking at the sentiment score polarity on a histogram, both summary and review are slightly skewed to the right and there is a trend for more positive sentiment. Review distribution looks more symmetric. However, by looking at the summary, there is a substantial high sentiment polarity of score zero.



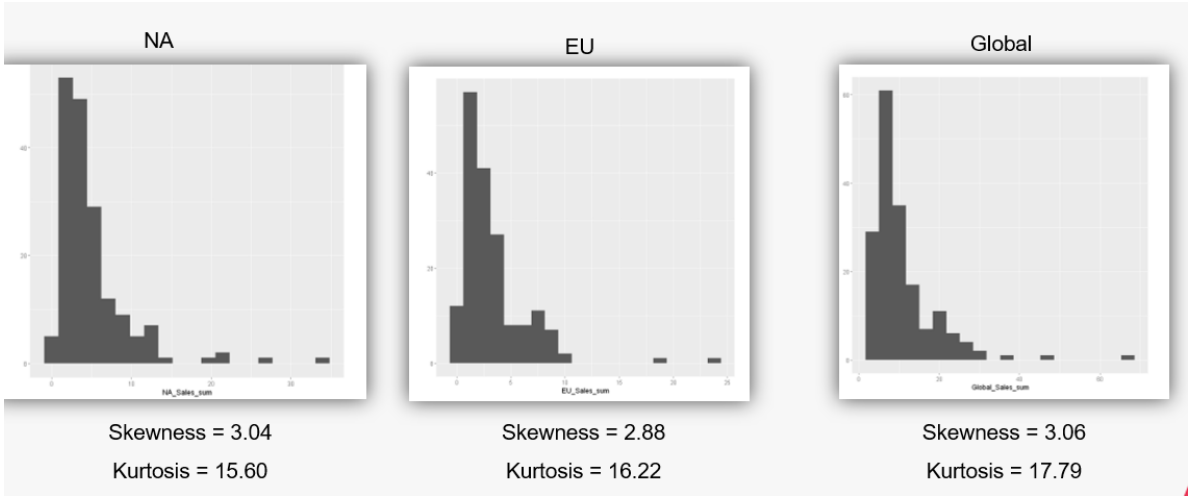


An analysis was made by looking at top twenty positive/negative reviews. When analysing negative reviews and summaries, all of them relate to product dissatisfaction rather than customer service. This helps to understand where the focus should be. Also, when exploring top twenty negative and checking the subjectivity of whether reviews are fact-based and objective or opinion-based and subjective, the result showed a more popular opinion-based score. It is important to priorities fact-based and act from there.

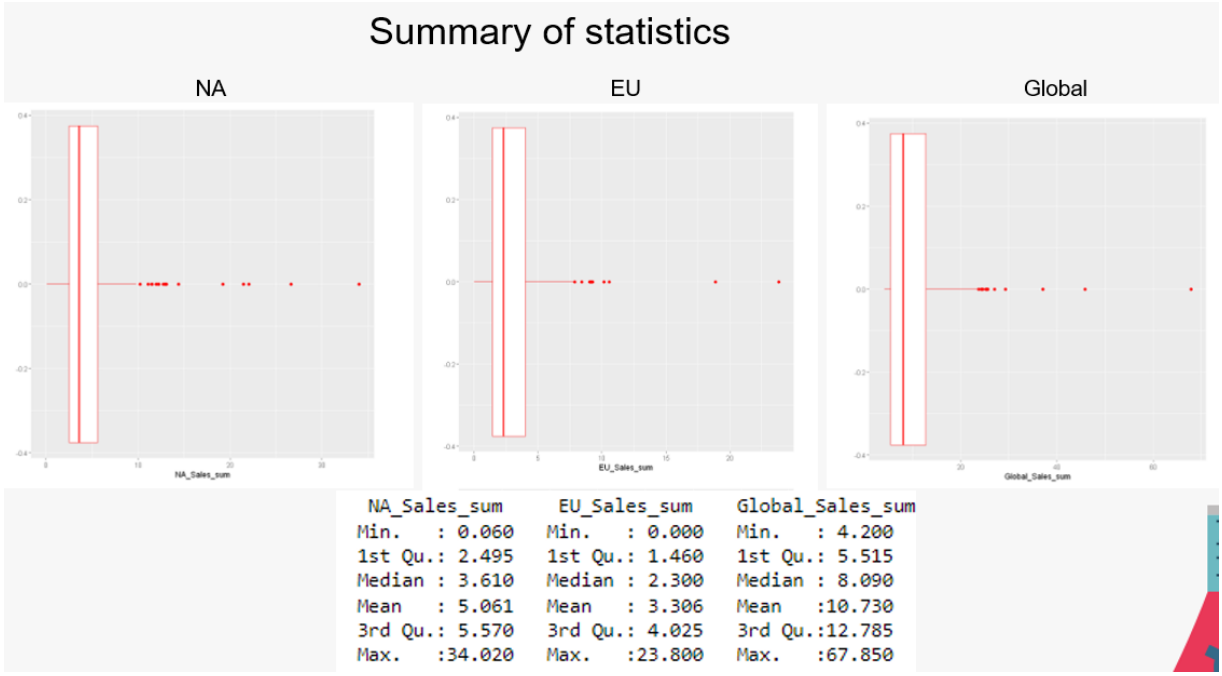
Furthermore, it is crucial to detect why there is such a high value score of 0.0 for sentiment polarity on summaries and investigate the accuracy of these. It was found that all starred reviews have been classified as 0.0 along with a sizeable number of positive reviews.

By addressing how reliable sales data is, an exploration was made using R tool by plotting scatterplots, histograms, and boxplots. The sales were grouped per Product ID, to remove the noise and get a cleaner plot.

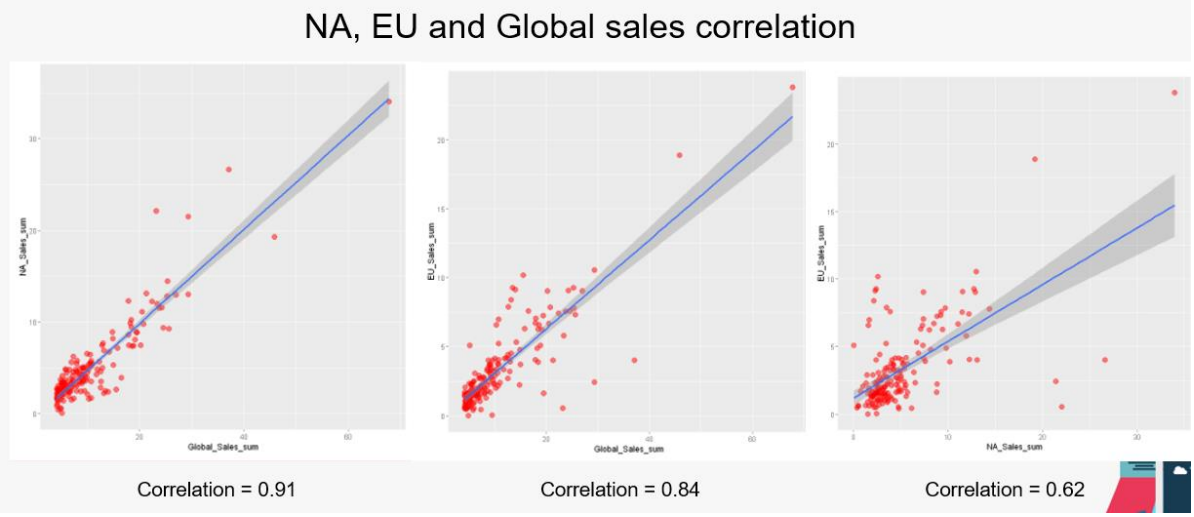
The histograms by regions showed a positively skewed data along with a few outliers detected. Measures of shape were applied. Hence, skewness confirms that all sales data is positively skewed and kurtosis being greater than three, meaning that data is leptokurtic or a heavy-tailed distribution with many outliers.



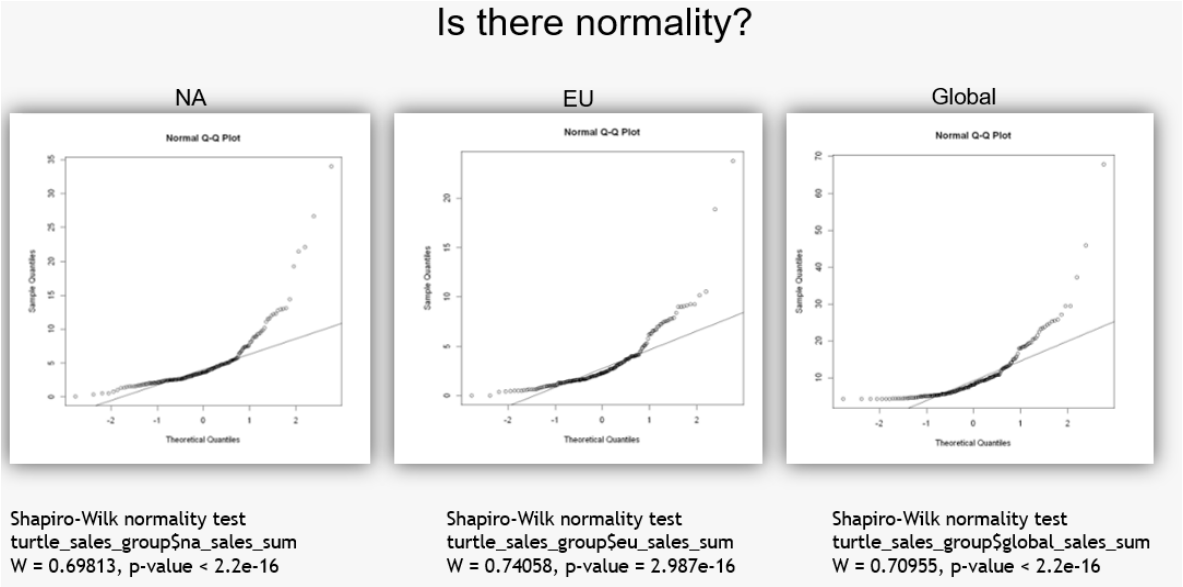
Also, boxplots were added to see the outliers from a different view along with a better understanding of summary of statistics. It can be clearly seen that EU is having the smallest number of sales, followed by NA and Global performing the best.



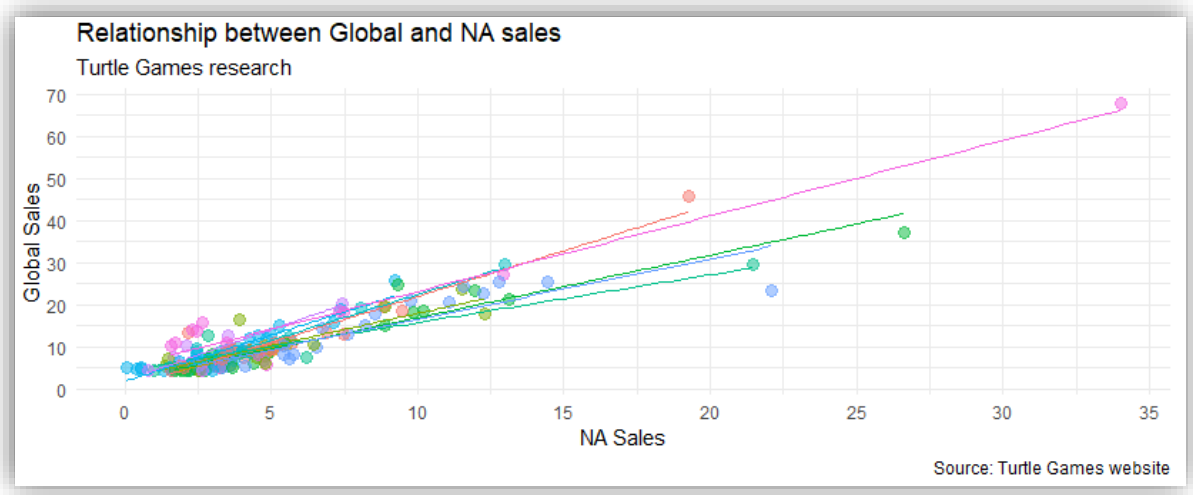
Correlation function was applied to detect a strong correlation between NA/Global sales, followed by EU/Global and a smaller result of .62 for NA/EU. It can be explained that the high relation is due to global sales including both EU and NA regions and hence, providing inaccuracy.

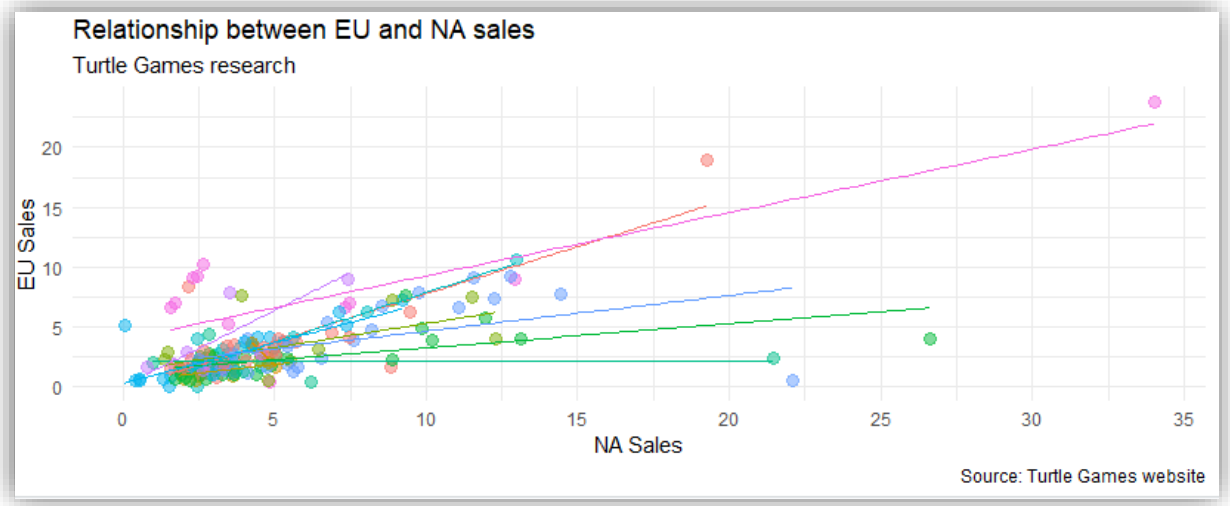


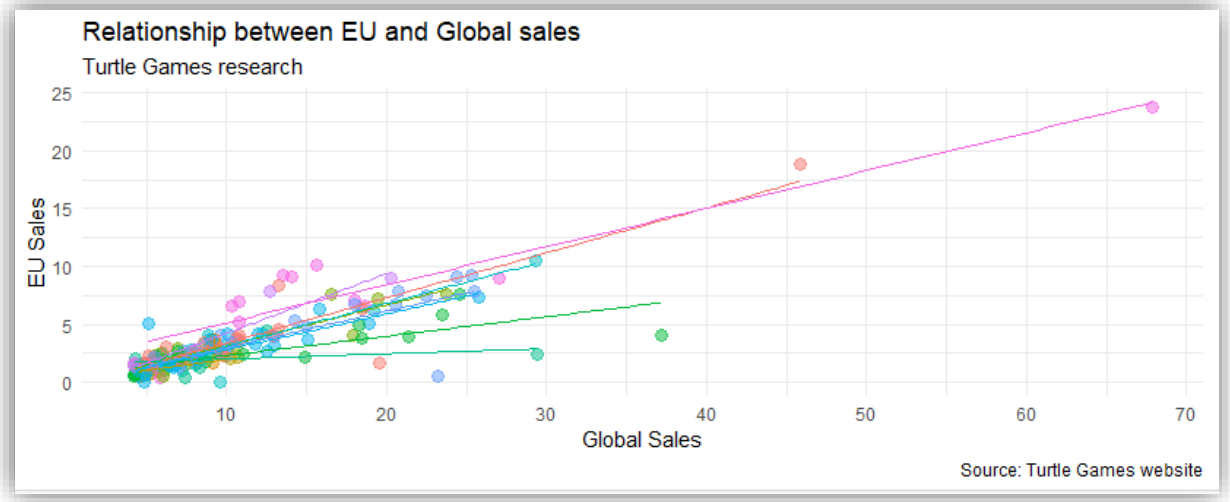
Normality was evaluated using the Shapiro-Wilk test and it was concluded that the data is not normally distributed across all sales regions.



By looking at genres across regions, the most popular are “sports,” “shooter,” “role-playing” and “platform.” There is a strong correlation by regions on “sports,” being positioned in the lowest sales point, but selling the most. However, the plot shows outliers on “sports” and “action” category, having the highest sales point. An investigation should be made onto that and see whether there is such product available online.







Top ten by product shows “sports,” “action” and “shooter” products dominating across all regions.

To better understand the relation between regions, predictive models were applied. Firstly, three separate linear regression models where build to find the best fit between EU/NA, Global/EU, Global/NA. The values of the variables have been mutated to return the natural logarithm and reduce the sum of square errors to strengthen the fit. In turn, there was a strong relation found in Global/EU and Global/NA, showing very low p value, smaller sum of squares error and higher adj. R sq.

An evaluation has been made on creating a multiple linear regression, which showed a much stronger correlation of 97% and low p values, proving significance of global sales with NA and EU region. To test its accuracy, predictions of global sales were made, given NA and EU values. The output showed a good prediction when comparing to the observed values.

However, confidence cannot be stated based on the models provided, due to Global variable including NA and EU regions that distort the results. Therefore, a high accuracy in prediction and correlations is interpreted wrong in that scenario.

Conclusion and Recommendations

In conclusion, there is a strong correlation of spending and remuneration versus loyalty. By looking at spending score versus remuneration, we were able to find five market segments.

Using NLP, the results showed tendency to a positive score. Further analyses are needed to test the 0-score polarity on summary and the subjectivity of the negative reviews by detecting if its personal opinion or factual information. The range of [0,1] would tell us that.

Another NLP model such as VADER could be evaluated on customer reviews and get a distinct perspective on the product performance. VADER not only shows the Positivity and Negativity score but also how positive or negative a sentiment is.

NA and EU sales value must be separated from Global, to get a better sense of the accuracy correlation between performance of the regions.

Lastly, more variables such as product name, pricing, transaction date, manufacturer, stock levels and margins will bring in more accurate outputs. This in turn will help create predictive models such as time series, linear/multiple linear regression, or regression decision trees to estimate the stock sales and pricing strategy.