Turtle Games Business Problems

Background/context of the business

Turtle Games is a global game manufacturer and retailer which sells books, board games, video games, and toys. The company collects sales and customer review data and wishes to utilize this data to improve sales performance by analyzing customer trends.

In this case, Turtle Games wants to understand the following questions.
1. Do customers with different age/remuneration/spending score accumulate loyalty points differently?
2. How to best segment the customer base?
3. How can customer reviews be used to optimize marketing campaigns?
4. How are the sales of each product?
5. How reliable is the data?
6. What is the relationship between sales in each region?

Analytical approach

A. Customer review data.

I used Python to clean and analyze customer review data by following these steps.

Preparing the data set and libraries
1. Import the necessary libraries, including:
    - numpy and pandas for data wrangling
    - matplotlib and seaborn for data visualization
    - statsmodel for regression modeling
    - sklearn for clustering
    - nltk for natural language processing
2. Load and explore the data set using *info* and *dtype* function, and check if there are any missing values with the *isna* function.

Question 1. Do customers with different age/remuneration/spending score accumulate loyalty points differently?

1. Remove columns that are not necessary for the analysis and change the column names to make them easier to reference.

2. Use the statsmodel function to identify linear relationship between loyalty points and age/remuneration/spending scores by plotting linear regression.

Question 2. How to best segment the customer base?

1. Create a new dataframe containing only the remuneration and spending score columns.
2. Use pairplot and scatterplot to identify possible clutters.
3. Use the Silhouette and Elbow methods to determine the optimal number of k-mean clustering.
4. Fit and plot the final model using the most optimal number of clusters.

Question 3. How can customer reviews be used to optimize marketing campaigns?

1. Subset only the necessary columns, namely *review* and *summary*.
2. Normalize the data by
   - converting the data to lowercase
   - joining the words in each column into one paragraph,
   - removing punctuations
   - dropping duplicates
3. Make a copy of the dataframe and apply tokenization.
4. Use wordcloud to visualize the frequency of each word.
5. Create frequency distribution and remove stopwords and alphanumeric characters.
6. Create another wordcloud and identify the 15 most common words and their sentiment polarity.
7. Plot histograms and review the sentiment score of each column.

B. Sales Data

I used R to clean and analyze sales data by following these steps.

Question 4. How are the sales of each product?

1. Import the data and import the necessary libraries, including:
   - Tidyverse
2. Remove redundant columns.
3. Create scatterplots and histograms to gain initial insights into the sales data.
4. Use group by and aggregate to find the sales per product
5. Visualize the data to gain more insights.

Question 5. How reliable is the data?

1. Determine the normality of the dataset using Q-Q plots.
2. Perform a Shapiro-Wilk test.
3. Determine the skewness and Kurtosis.

Question 6. What is the relationship between sales in each region?

1. Create a linear regression and a multiple linear regression model to determine the relationship between sales in each region and visualize the data.
2. Predict the global sales based on provided values.

Visualization and insights

Question 1. Do customers with different age/remuneration/spending score accumulate loyalty points differently?
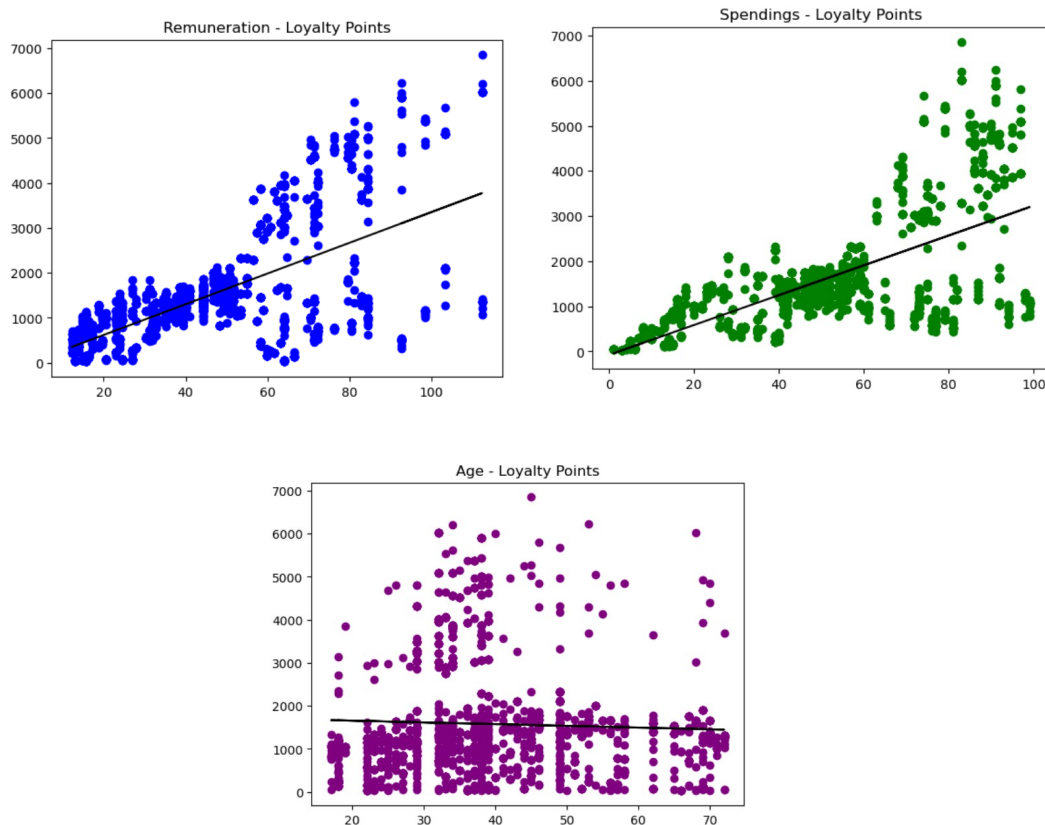

The visualization above shows the relationships between loyalty points and age/remuneration/spending score. As we can see, remuneration and spending positively correlate with loyalty points. This means customers with higher remuneration or spending scores will likely accumulate more loyalty points. However, that is not the case for age. From the visualization, we can see that the trendline is relatively flat. This means that the relationship between age and loyalty point accumulation is not obvious and significant.
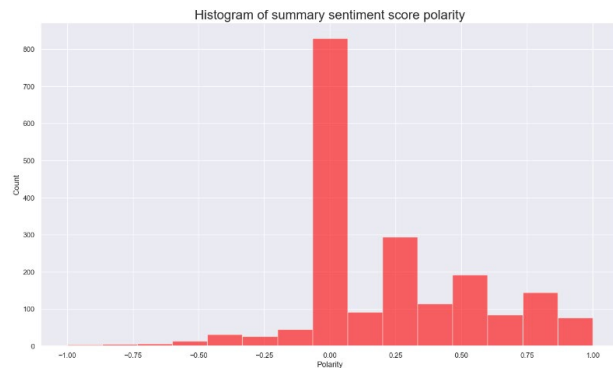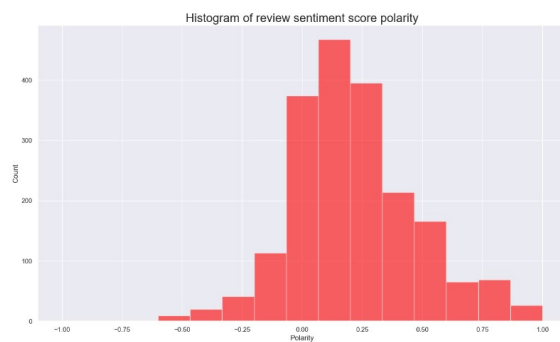
Question 2. How to best segment the customer base?
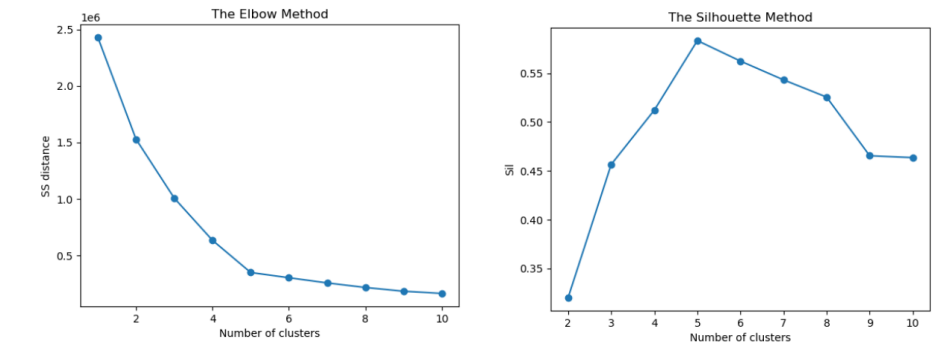
Remuneration - Spending Score 5 Clusters
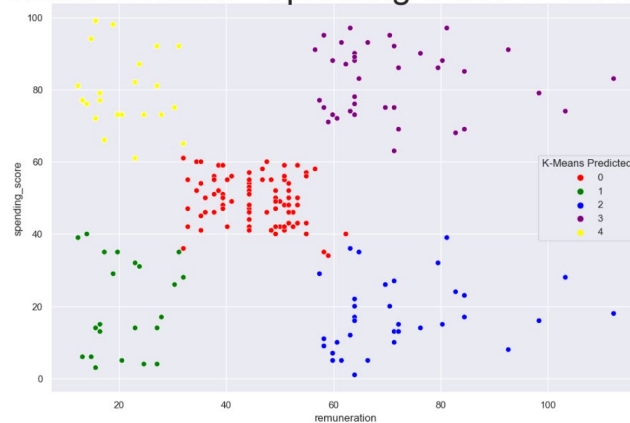
The Elbow and Silhouette method suggests dividing the customer into 5 groups based on their remuneration and spending score is best. Using 5 clusters will ensure that the customer in each cluster have high homogeneity, and market segmentation will be more accurate.

## Question 3. How can customer reviews be used to optimize marketing campaigns?

Histogram of review sentiment score polarity

Histogram of summary sentiment score polarity

	review	polarity_review
207	booo unles you are patient know how to measure...	-1.000000
181	incomplete kit very disappointing	-0.780000
1773	im sorry i just find this product to be boring...	-0.583333
362	one of my staff will be using this game soon s...	-0.550000
116	i bought this as a christmas gift for my grand...	-0.500000
226	this was a gift for my daughter i found it dif...	-0.500000
229	i found the directions difficult	-0.500000
289	instructions are complicated to follow	-0.500000
300	difficult	-0.500000
1501	expensive for what you get	-0.500000
173	i sent this product to my granddaughter the po...	-0.491667
345	my 8 yearold granddaughter and i were very fru...	-0.446250
531	i purchased this on the recommendation of two ...	-0.440741
305	very hard complicated to make these	-0.439583
421	kids i work with like this game	-0.400000
430	this game although it appears to be like uno a...	-0.400000
490	my son loves playing this game it was recommen...	-0.400000
795	this game is a blast	-0.400000
798	i bought this for my son he loves this game	-0.400000
814	was a gift for my son he loves the game	-0.400000

|	summary	polarity_summary
21	the worst value ive ever seen	-1.000000
207	boring unless you are a craft person which i am	-1.000000
819	boring	-1.000000
1148	before this i hated running any rpg campaign d...	-0.900000
1	another worthless dungeon masters screen from ...	-0.800000
143	disappointed	-0.750000
623	disappointed	-0.750000
785	disappointed	-0.750000
1591	disappointed	-0.750000
361	promotes anger instead of teaching calming met...	-0.700000
875	too bad this is not what i was expecting	-0.700000
880	bad qualityall made of paper	-0.700000
177	at age 31 i found these very difficult to make	-0.650000
100	small and boring	-0.625000
511	mad dragon	-0.625000
797	disappointing	-0.600000
1001	disappointing	-0.600000
1099	disappointing	-0.600000
1773	disappointing	-0.600000
991	then you will find this board game to be dumb ...	-0.591667

The histograms show that the general sentiment towards Turtle Games products are in the positive zone. But to improve the marketing strategy, it is worth exploring what the negative reivews are about annd how can the company fix those problems. For an example, a number of negative comments the manual or instruction being too complicated, the price is too high, and that their expectations were not met.

Question 4. How are the sales of each product?
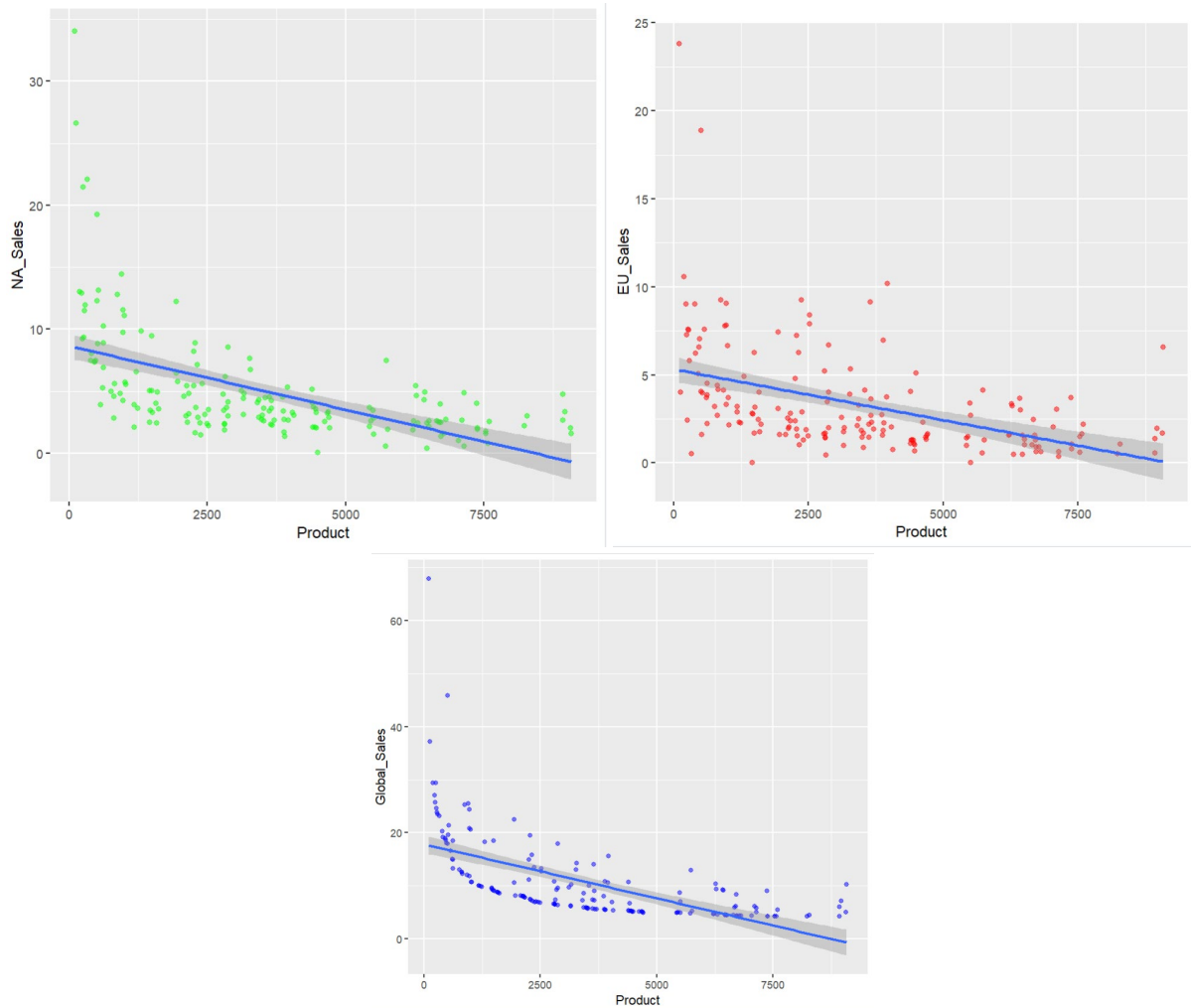
We can see that products with lower product IDs tend to have more sales in all regions. Overall, each product makes 0-20 million pounds. It is worth identifying the products that are not selling well to help consider if those products should be marketed differently.

Question 5. How reliable is the data?
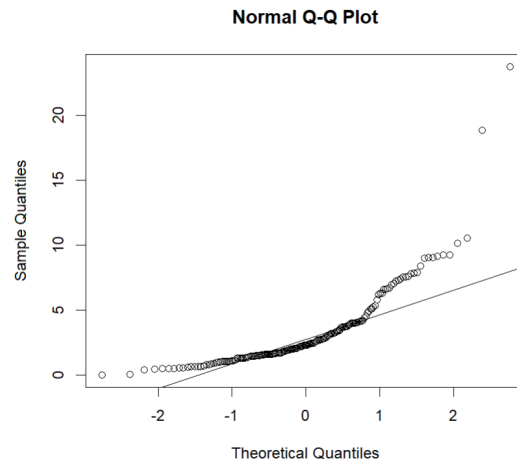
Normal Q-Q Plot

The data is reliable. This conclusion is reached after conducting the Q-Q plots, Shapiro-Wilk test, and identifying skewness. However, the sample size needs to be bigger for us to be able to say for sure if the data is normal.

Question 6. What is the relationship between sales in each region?

Using the multiple linear regression model, we got this formula

Global sales = 1.04242 + 1.13040(NA) + 1.19992(EU)

This means that global sales is positively influenced by sales in EU and North America. The formula above can be used to predict global sales using the sales in the other two regions.

Patterns and predictions

1. Do customers with different age/remuneration/spending score accumulate loyalty points differently?
    Customers with higher remuneration or spending scores will likely accumulate higher loyalty points. There are no clear differences in how customers of different ages accumulate loyalty points.

2. How to best segment the customer base?
    Customers can be best segmented into five groups based on their remuneration and spending score. This ensures that customers in the same group have similar characteristics and behaviors.

3. How can customer reviews be used to optimize marketing campaigns?
    Although the general sentiment towards all products is positive, a number of customers are complaining about three things
    -   The manual/instruction is too complicated
    -   The price is too high
    -   The customers' expectations were not met

    Turtle Games should investigate these claims so that it can optimize its products and marketing strategy.

4. How are the sales of each product?
    Products with lower product IDs tend to have more sales in all regions.
    Typically, each product makes between 0-20 million pounds.

5. How reliable is the data?
    The data is reliable. There needs to be more data to be able to say for sure. The company should continue to collect more data.

6. What is the relationship between sales in each region?
    Sales in each region positively affect each other. This means that if sales in one region increase, sales in other regions are like to increase as well.

    It is worth investing if sales are also influenced by the platform and genre. This is so that the company can spend more budgets selling the genre of products that makes more sales.