CO4_LSE_DA_301 Advanced Analytics for Organisational Impact

Assignment: Predicting future outcomes

Report

Jennifer Cheng
4-24-2023

1. Background/context of the business (100 words):

Briefly describe the business scenario (context) and the business problem you aim to solve for Turtle Games.

I am part of a team of data analysts at Turtle Games, a game manufacturer and retailer with a global customer base. My company manufactures and sells our products, along with sourcing and selling products manufactured by other companies. Our product range includes books, board games, video games, and toys. Turtle Games collects data from sales as well as customer reviews. Our business objective is to improve overall sales performance by utilising customer trends.

To achieve our goals, we have developed an initial set of questions to explore depth:

1. How customers accumulate loyalty points?
2. How groups within the customer base can be used to target specific market segments?
3. How social data (e.g. customer reviews) can be used to inform marketing campaigns the impact that each product has on sales?
4. How reliable the data is (e.g. normal distribution, skewness, or kurtosis)?
5. What the relationship(s) is/are (if any) between North American, European, and global sales?

2. Analytical approach (350 words)

Describe the approach taken to import, clean, and analyse data in Python and R. Include a detailed and insightful description of the processes you used and the decisions you made during analysis, such as the choice of libraries, functions, and variables. Ensure that the description of the steps taken to prepare data for analysis is clear, well-organised, and relevant to Turtle Games's objectives.

To make recommendations for Turtle Games, I've analysed its sales data and performed the following in Python and R:

1. Make predictions with regression.
2. Make predictions with clustering.
3. Analyse customer sentiments with reviews.
4. Visualise data to gather insights.

Once data was loaded in Python, it was examined, cleaned, and transformed - wrangled the review data- clear that some columns were irrelevant to the analysis and therefore deleted. I've checked that there are no missing data in the Loyalty Points and Product column and then looked at the customers' information in relation to loyalty. I created a new csv file with the cleaned dataframe. So, I performed three regression plots with spending, remuneration, and age against loyalty.

Next, I prepared the data for clustering by creating a new dataframe with the necessary columns only, remuneration and spending score. Then proceed to plot the new dataframe. Using the Silhouette and Elbow methods to determine the optimal number of clusters for k-means clustering. Then evaluated the two, three, and four clustered k based on the insights from the Silhouette and Elbow methods.

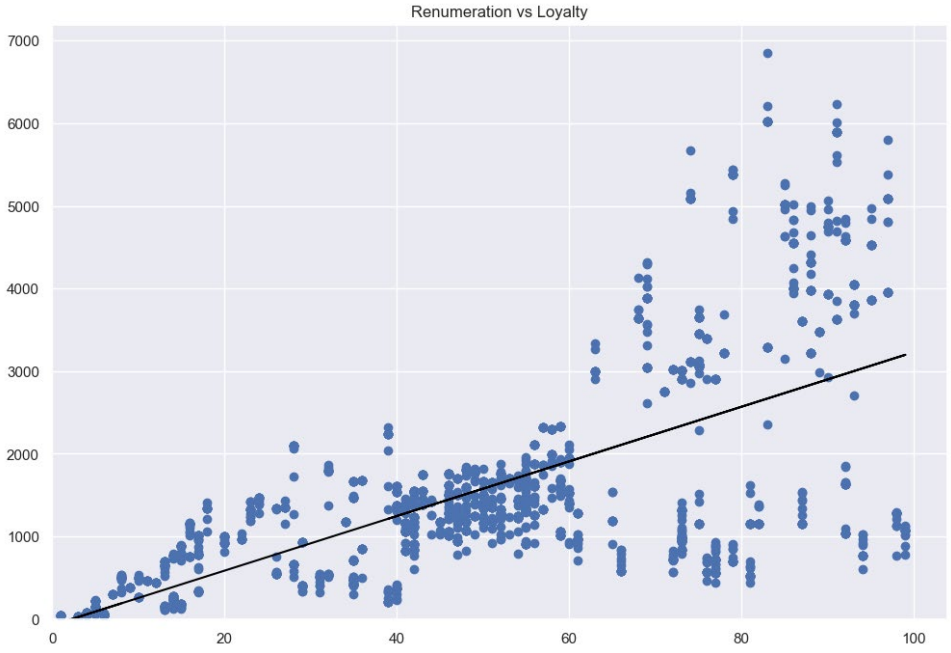
Then I looked at the sentiment on the customers' reviews and summaries. To clean the data, first transform all the text into lower case, and then replace the punctuation in the columns and tokenise the data to create wordclouds. Next was to examine the frequency distribution and polarity of the review and summaries without alphanumeric characters and stopwords. Last was to review the polarity and sentiment and twenty positive and negative reviews.

Then, I moved on to R and analysed the sales data. The first step is to clean the data and create a new dataframe without the redundant columns (Ranking, Year, Genre, Publisher). Proceed to create plots to examine the data with a normal distribution, skewness, and Q-Q plot to determine the correlation. Also, the linear regression predicted global, North America, and Europe sales.

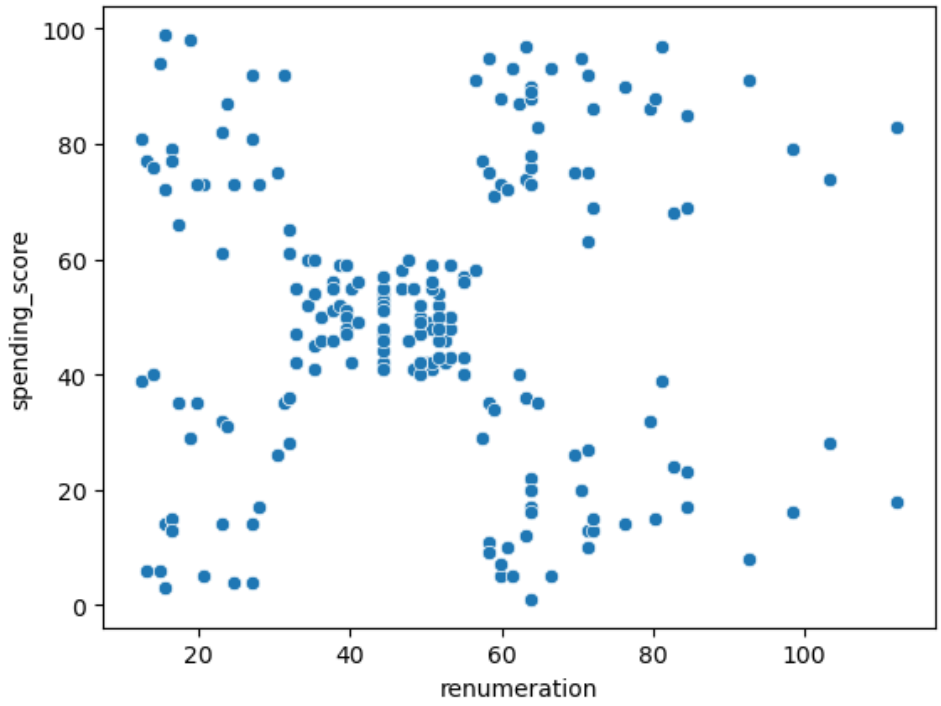
3. Visualisation and insights (350 words)

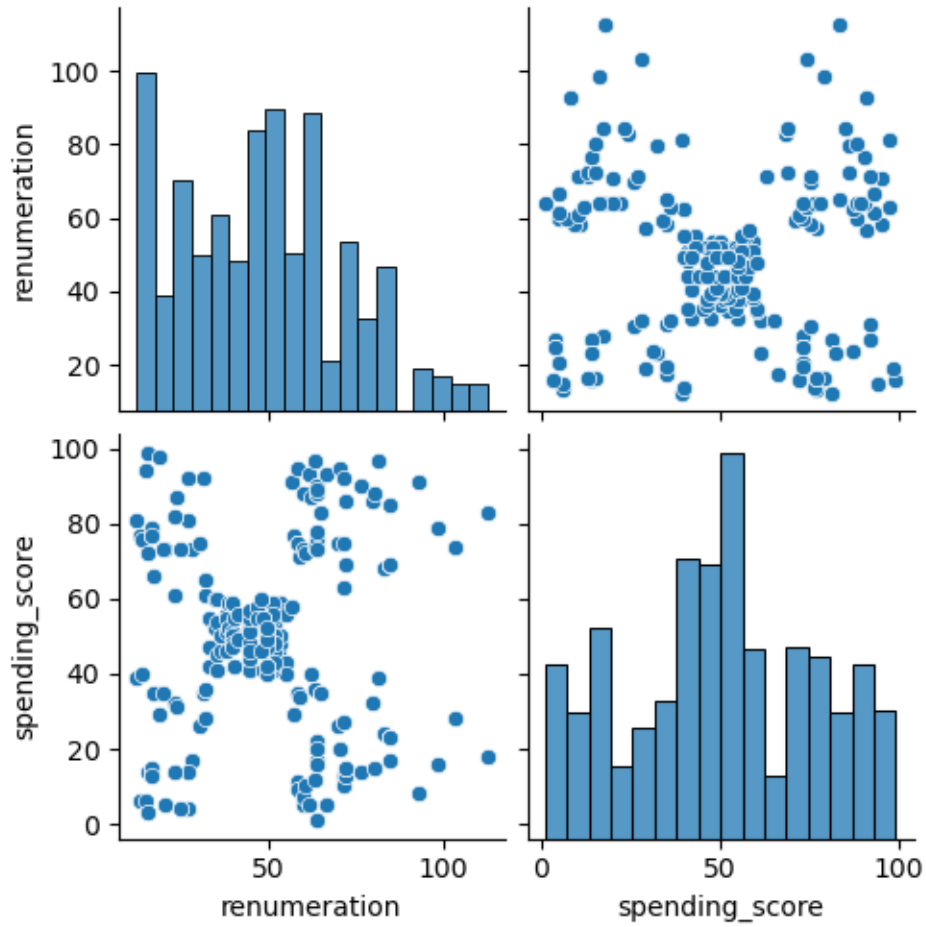
Describe the rationale for the selected visualisations, and ensure that the interpretations of the visualisation outputs are detailed, insightful, and relevant to the business objectives

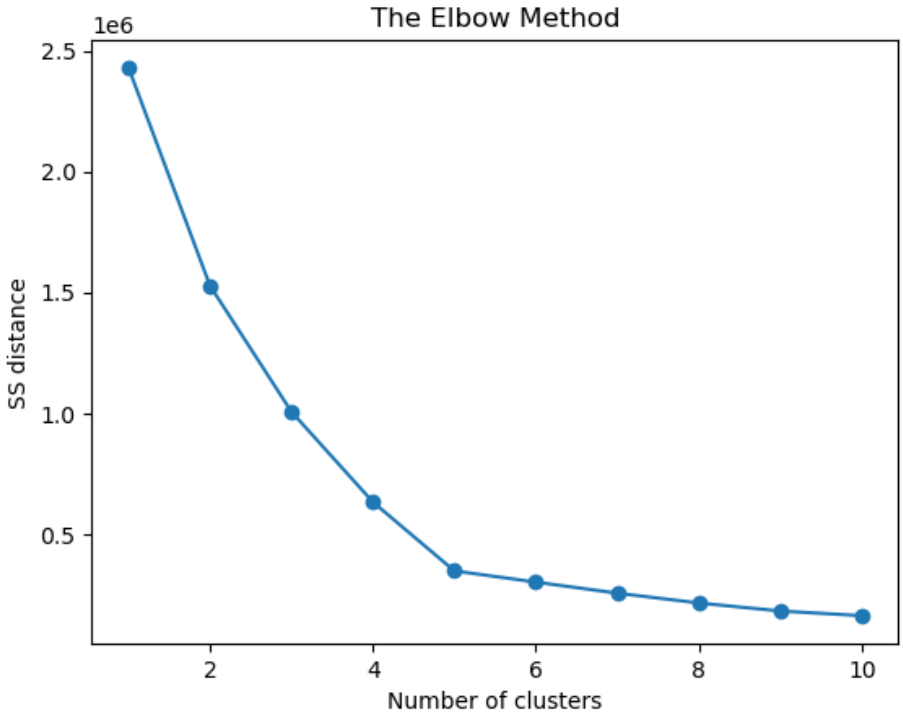
First, to find out the correlation between the customers and loyalty points. Then, I looked at the customer's spending, remuneration and age. This will give a profile to the customers who are more loyal to Turtle Games.



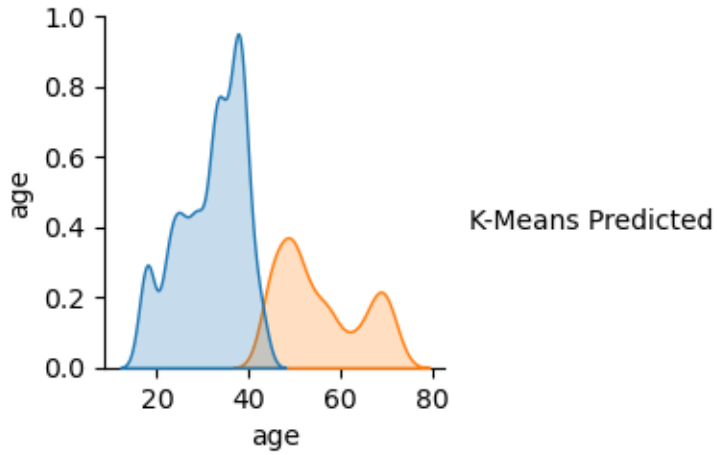
Next, I examined the groups within the customer base's remuneration and spending to target specific market segments.

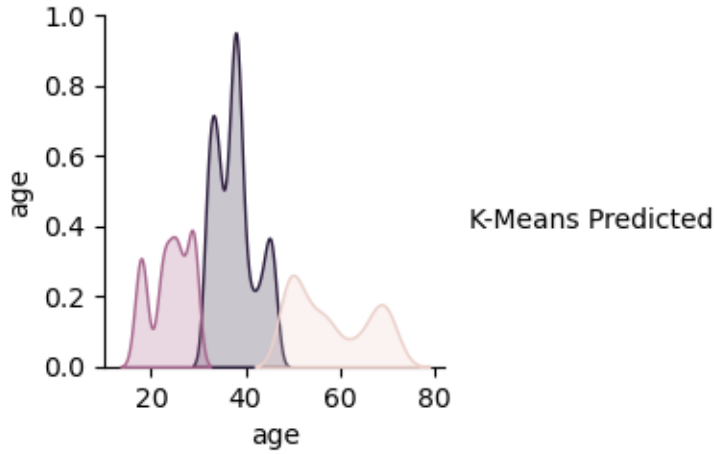

spending_score
renumeration
100
80
60
40
20
0
20
40
60
80
100

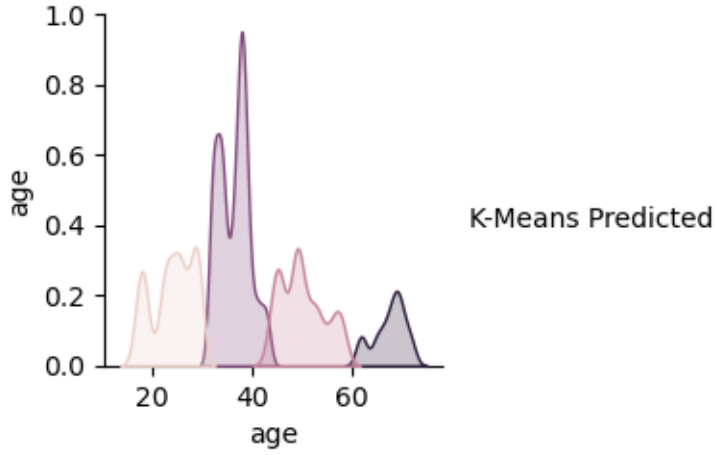

renumeration
spending_score
renumeration
spending_score
100
80
60
40
20
0
50
100


The Elbow Method
1e6
2.5
2.0
1.5
1.0
0.5
SS distance
2
4
6
8
10
Number of clusters


The Silhouette Method
0.68
0.66
0.64
0.62
0.60
0.58
Sil
2
3
4
5
6
7
8
9
10
Number of clusters



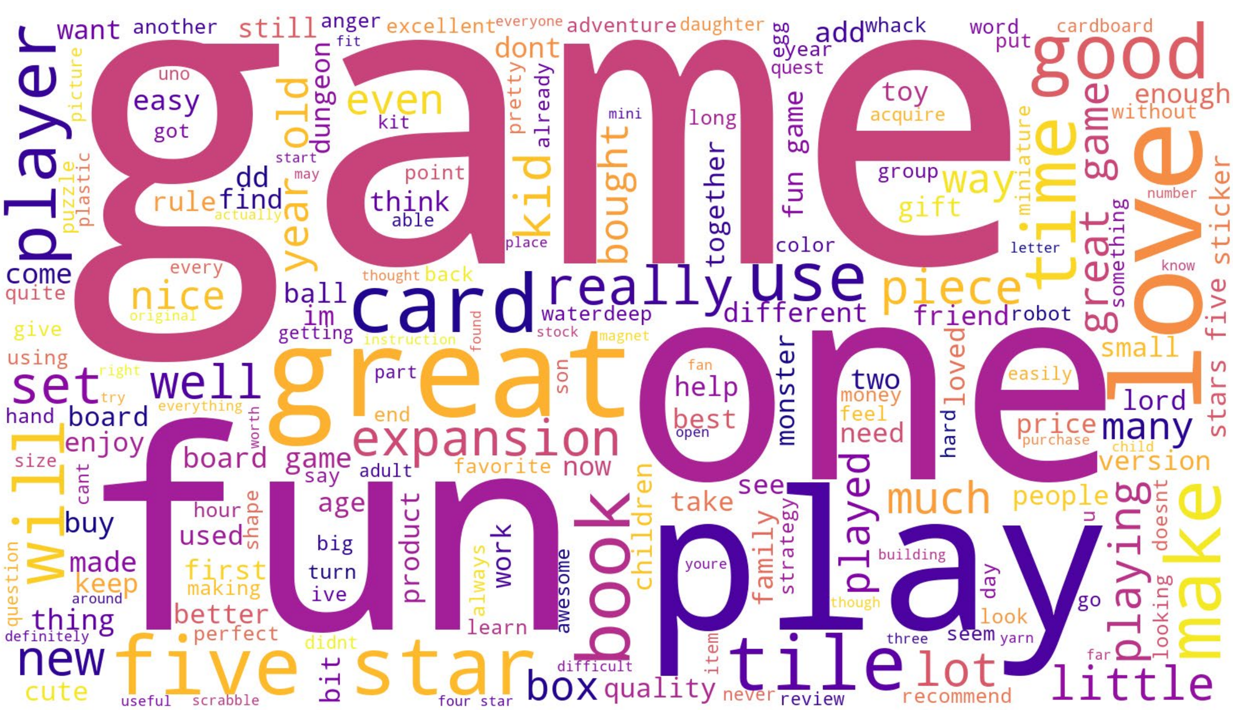


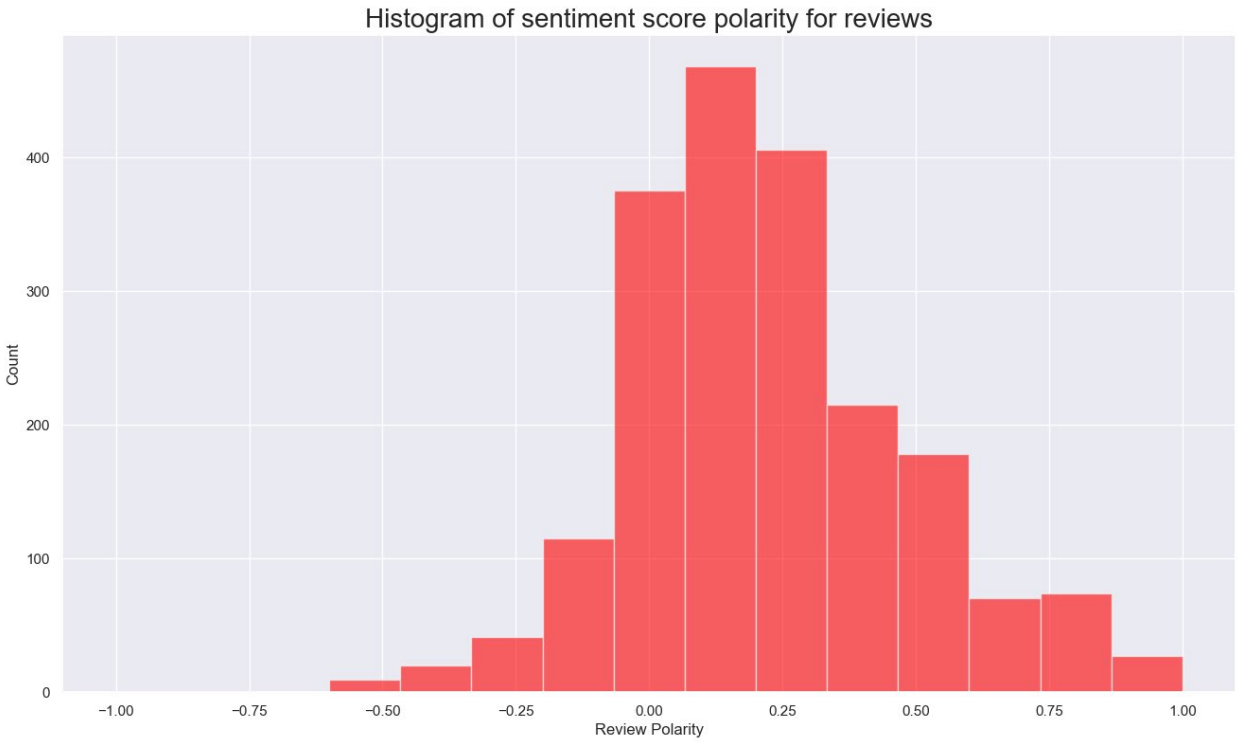


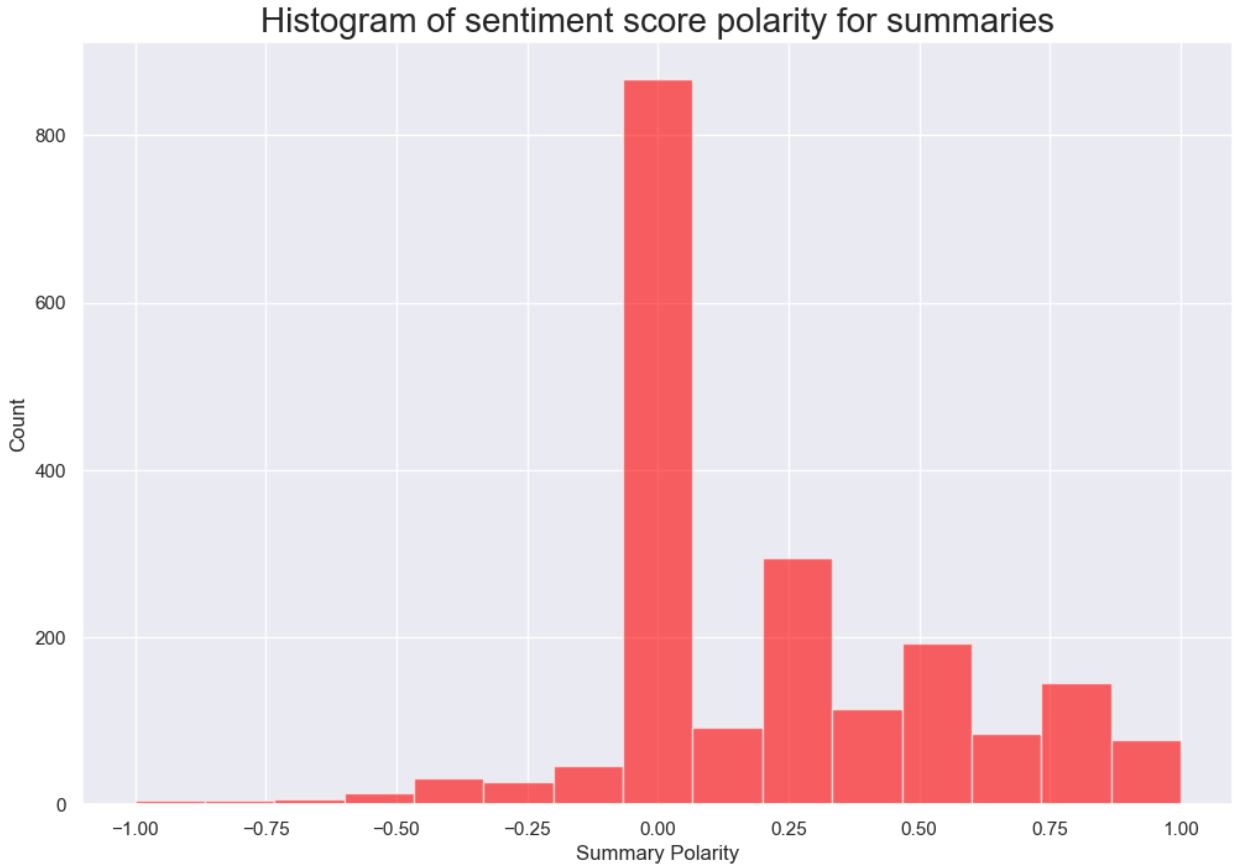
Sentiment analysis of customer reviews and summaries.

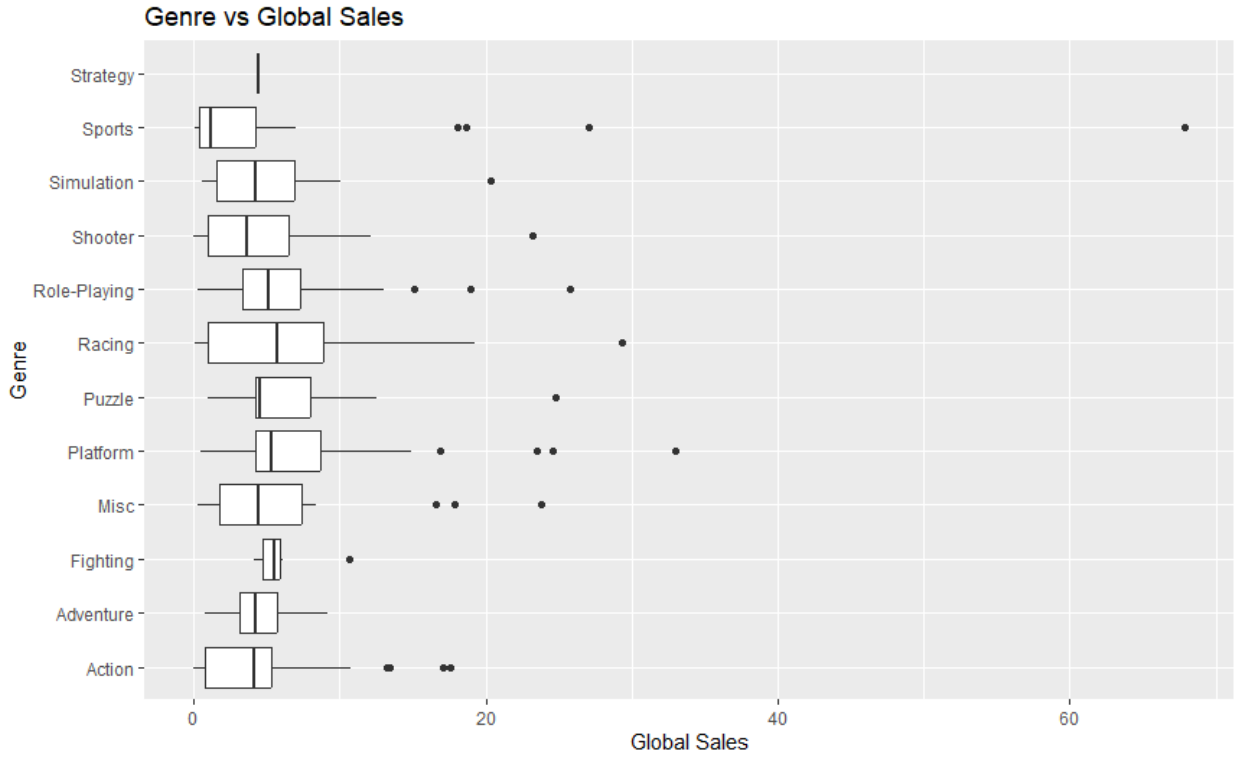

and
it
that
but
in
for
the
game
of the
is a
one
or
to
a great
this game
them
time
also
play
for the
just
with the
if you
really
board game
player
more
some

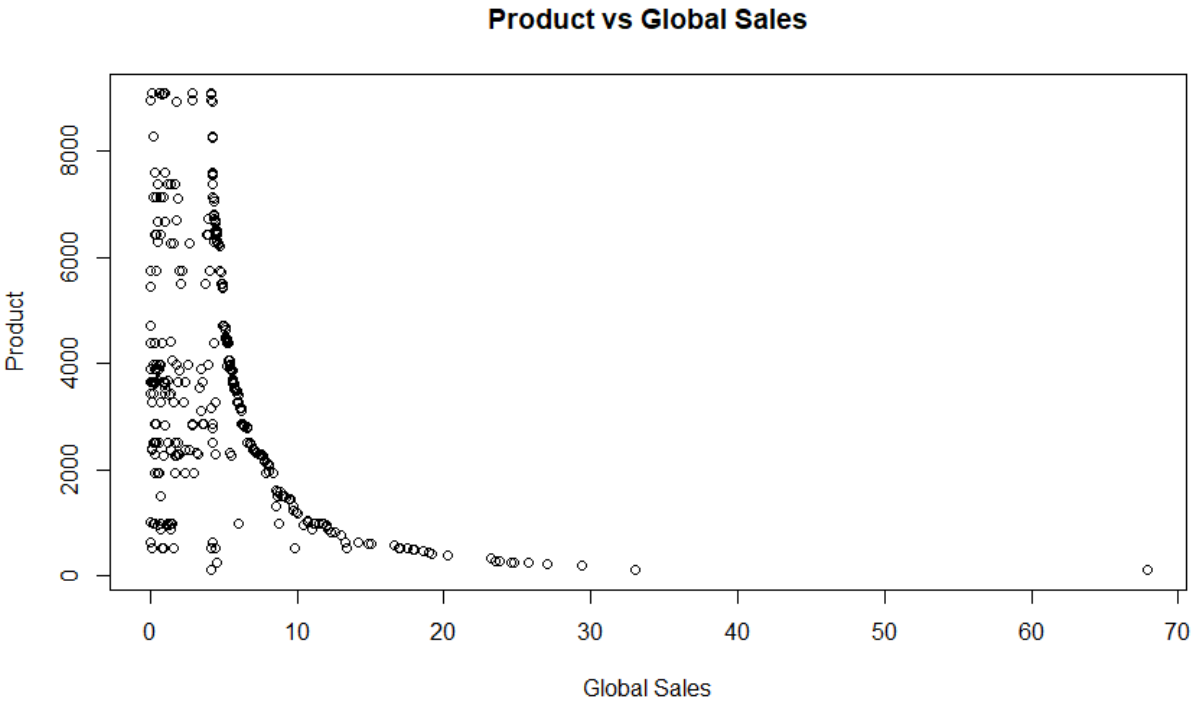
`

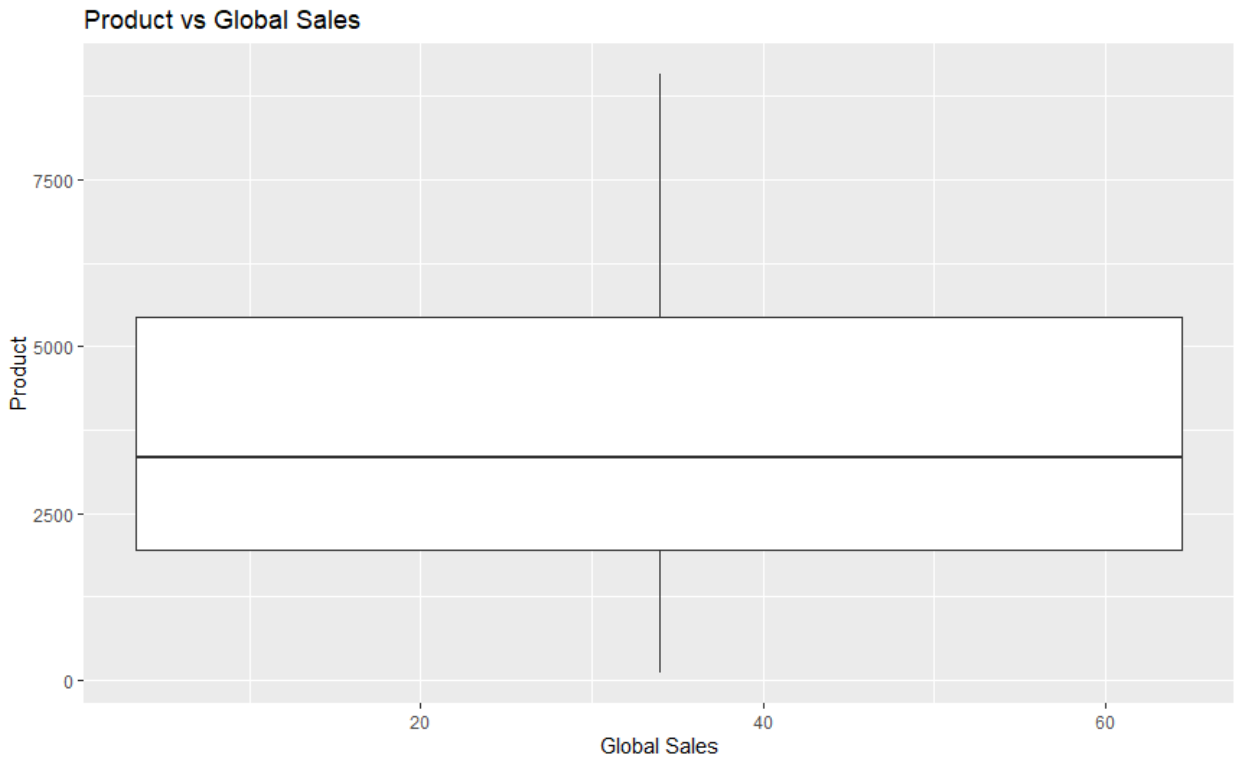

game
fun
one
play
great
love
card
good
player
five star
tile
book
make
use
really
expansion
piece
time
well
set
will
nice
kid
bought
year old
little
lot
played
much
great game
even
easy
toy
new
box
quality

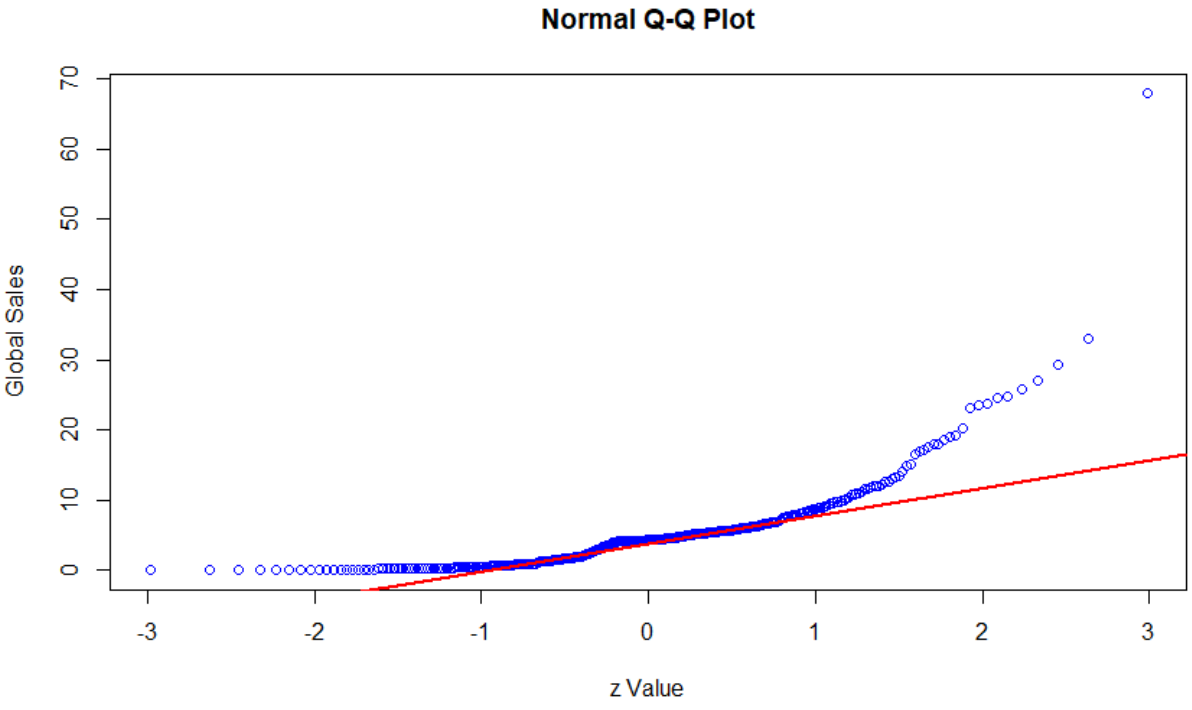

Histogram of sentiment score polarity for reviews
Count
0
100
200
300
400
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Review Polarity

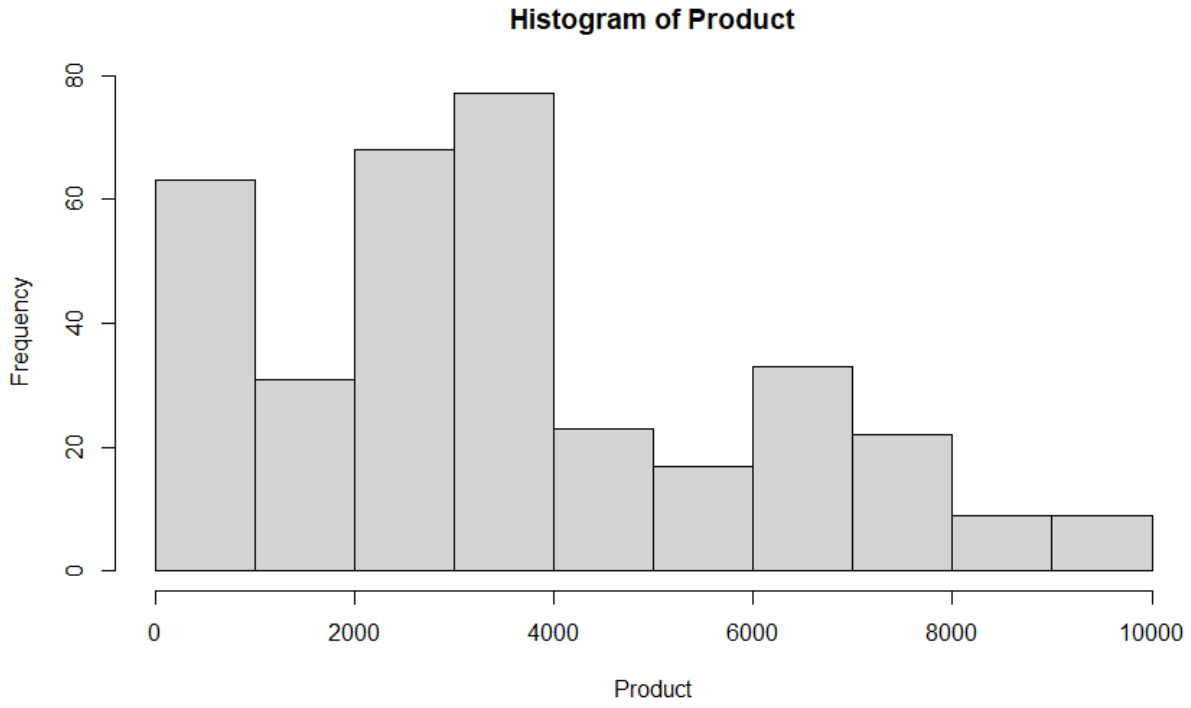

Histogram of sentiment score polarity for summaries
Count
0
200
400
600
800
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Summary Polarity


Genre vs Global Sales
Strategy
Sports
Simulation
Shooter
Role-Playing
Racing
Puzzle
Platform
Misc
Fighting
Adventure
Action
Genre
0
20
40
60
Global Sales


Product vs Global Sales
Product
0
2000
4000
6000
8000
0
10
20
30
40
50
60
70
Global Sales


Product vs Global Sales
Product
7500
5000
2500
0
20
40
60
Global Sales


Normal Q-Q Plot
Global Sales
0
10
20
30
40
50
60
70
-3
-2
-1
0
1
2
3
z Value



4. Patterns and predictions (200 words)

Clearly articulate the patterns you've discovered and the predictions you've made. Describe how they relate to the business scenario, and address Turtle Games's objective

As a business analysis for Turtle Games, it's essential to identify how customers accumulate loyalty points.

The linear regression plots tell us that:

- Spending and loyalty are positively correlated
- Remuneration and loyalty are also positively correlated
- But age and loyalty are negatively correlated.

This means people with higher incomes and who spend more on games are more loyal to Turtle Games. In comparison, loyalty diminishes as they grow older.

Based on the K-mean clustering analysis on remuneration and spending, k=4 (four clusters) might give the best results (groups). The remuneration is closely correlated to the spending; therefore, Cluster 1 for both k=3 and k=4 is the largest group. The number of predicted values per class indicates a better distribution for k=4 than k=3. Through the clusters for k-means clustering it's noted that most popular age of customer is between 30- to 40-year-old.

As for the sentiment analysis, worldcloud and word popularity analysis did not give a lot of formation to the customer's sentiment. Popular words like "game", "great", and "fun" would be useful keywords in future campaigns to attract new customers. The polarity and sentiment analyses revealed that both the customers' reviews are positive, and the summaries are neutral and positive. The 20 negative and positive reviews and summaries did not provide much information because the positive and negative comments were generic.

In terms of sales, the game with the most variety is racing. The other types of games with lots of sales are simulation, shooter, and action. The type of game that is the lowest sales is strategy. Fighting, adventure, and role-play are the other games that generate few sales. The genre of the game with highest range are racing. There was no consistent observation in the sales of the product by product_id. The products with lower numbers have more sales in both the boxplot and histogram analysis. In terms of times, the higher ranking happened between 2000 and 2015.

5. Make recommendations

1. The analysis states that the games' reviews are positive overall. Customers with higher incomes spend more on video games. They should also target their marketing efforts to the late 20's to the 40 year olds with early to mid-30 year olds as highest customer base.
2. Since the comments from the reviews were generic, the company would gain more insights if surveys were asked more specific questions to draw out information.
3. To focus on increasing sales, Turtle Games can provide more products in popular genres such as racing, simulation and shooting.

If you had more time and/or Python and R knowledge, I would conduct the following analysis

1. Also, the price per unit was not available. Therefore the sale data is based on the assumption that the unit price of all the products is the same. If the unit price cost is available, I will conduct an analysis of sales and price.
2. Since video games are generally marketed to males, there are female customers in the data. It would be great to analyse the male-to-female distribution and the relationship between genre and gender.
3. Another analysis to be made is to compare the popular genres and review the comments would be helpful.

Besides these observations from the data analysis, I also want to note that the data ends in 2015 and may be different if it continues to 2023. I would not make a recommendation in 2023 with data that is eight years old.

Also, in the descriptions of Turtle Games, they sell their own products, along with sourcing and selling products manufactured by other companies. However, when doing the analysis, there's no indication of whether the game was developed by Turtle Games or other companies.

Also, the analysis is focused on video games, but Turtle Games carried more than video games. I think the date, source of the products, and type of product lines have to be communicated to Turtle Games in order for them to achieve their objectives understanding the limits of the analysis.

References:

https://stat.ethz.ch/R-manual/R-devel/library/base/html/print.dataframe.html

https://dplyr.tidyverse.org/reference/context.html