Predicting Future Outcomes

Antonio Arantes

July 2024

Index

1. Background of the business scenario

Turtle Games is a global game manufacturer and retailer that produces and sells its products and those from other companies. Their diverse product range includes books, board games, video games, and toys. The company collects extensive sales and customer review data to enhance overall sales performance by analyzing customer trends. Our team of data analysts has been contracted to address critical business questions for Turtle Games. Specifically, we will explore how customers engage with and accumulate loyalty points, how to segment customers for targeted marketing, and how to leverage text data from customer reviews to inform marketing strategies and business improvements.

2. Analytical approach

The team has been provided with one dataset: *turtle_reviews*. The first step after importing the dataset was to sense-check the data. The dataset comprises 2000 rows and 11 columns, with no missing values or duplicates. The following cleaning tasks were performed:

Action	Objective
Drop columns *language* and *platform.*	Both columns only have one value, so they were unnecessary for the analysis.
Capitalize *education* values.	Improve the consistency of the data.
Convert the *product* into a string.	The product works as categorical data, referencing what was reviewed.
Change column names.	Improve the consistency of the data.

Table 1 - Cleaning Tasks

There were no outliers in the columns *Age, Spending_score,* and *Remuneration.* 266 outliers were detected in the column *Loyalty_points*, but they were kept to avoid misrepresenting the high values of that column and the clients with those values.

Each row represents a product review made by a client, and we have several repeated combinations of Age, Remuneration, Spending Score, Gender, and Education.

	gender	age	remuneration	spending_score	loyalty_points	education	product	review	summary
70	Male	72	40.18	55	1322	Phd	9530	AGE APPROPIATE. G'SON LOVED	Five Stars
270	Male	72	40.18	55	1322	Phd	9560	I bought this for an adult crafter. While she ...	Not Enough Yarn
470	Male	72	40.18	55	1322	Phd	9560	Great to use with kids who need support with s...	Great to use with kids who need support with s...
670	Male	72	40.18	55	1322	Phd	9597	Love these magnets, they are adorable! Just a ...	Too cute!
870	Male	72	40.18	55	1322	Phd	1581	I have always loved this game. It is as much o...	One of the best!
1070	Male	72	40.18	55	1322	Phd	2521	i haven't had a full chance to play with it ye...	WoA review
1270	Male	72	40.18	55	1322	Phd	6646	It's just beautiful work. And the box is also ...	Awesome

Figure 1 - Example of the repeated variables.

We can confirm that the loyalty points are kept the same across the complete set of characteristics despite different products. That indicates that loyalty points are indexed to each client and not to each review as expected. To avoid putting different weights and repeating ourselves while analyzing our clients, a data frame that grouped the five characteristics that define each client was grouped, assuming that no two persons have the same set of five factors. Two new columns were also created; one is an identifier of the client called *client_id*, and another is a column called *number_of_sales*, which is the number of rows each customer appears and, consequently, the number of products sold to that client.

	age	remuneration	spending_score	gender	education	number_of_sales	client_id	loyalty_points
0	17	13.94	40	Female	Postgraduate	1	1	233
1	17	18.86	98	Male	Phd	1	2	774
2	17	27.06	4	Male	Phd	1	3	45
3	17	27.06	92	Male	Phd	7	4	1042
4	17	35.26	54	Female	Postgraduate	1	5	797

Figure 2 - First five rows of the clients data frame.

Most of our analyses were done in this table. If we look at the reviews table with repeated loyalty points, we give extra weight to clients who make more than one purchase. Even when we cluster our customers, if we don't create this data frame, we will have inconsistent results since we don't have 2000 clients. The *client* data frame is made of 782 unique customers.

Some functions were created to plot graphics quickly. For linear regression, we used the OLS method; for decision trees, a Decision Tree Regressor; and for clustering, K-means. We employed Vader and TextBlob for sentiment analysis to determine which provided better results, enabling a comprehensive evaluation of customer trends and insights.

3. Visualisation and notebook structure

The Jupyter notebook was structured to have a clean look and to identify the section we are analysing quickly. The colour scheme is based on the Turtle Games colour to be consistent with the brand identity. That can be seen in the section titles of the notebook and on the graphs, mainly on the title, x-axis label, y-axis label, and some bar and line charts.

The analysis started with exploratory data analysis. Several univariate and bivariate charts were produced to evaluate and understand the information about Turtle Games clients accurately. Some variables were assessed by sales and clients.

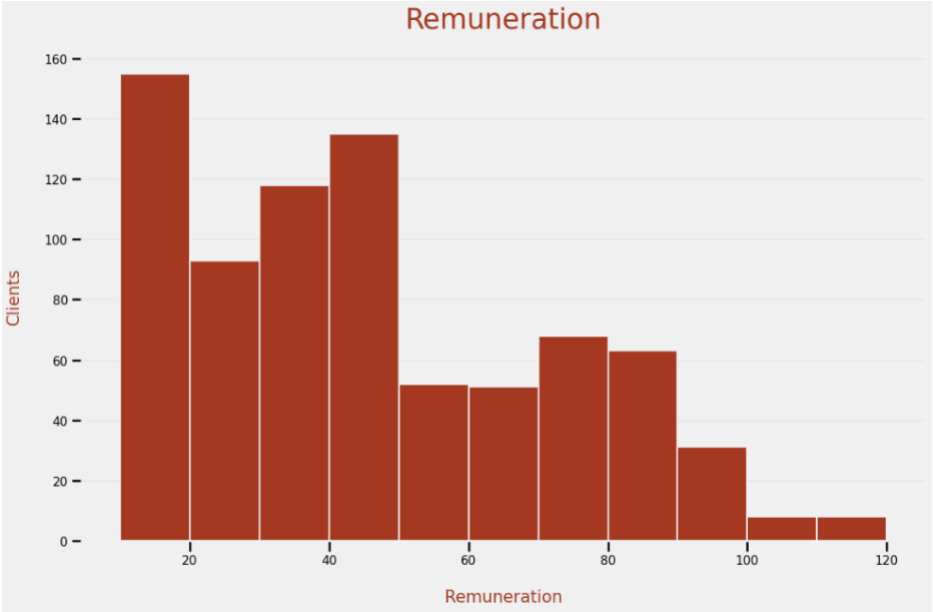

Figure 3 - Example of univariate analsysis.

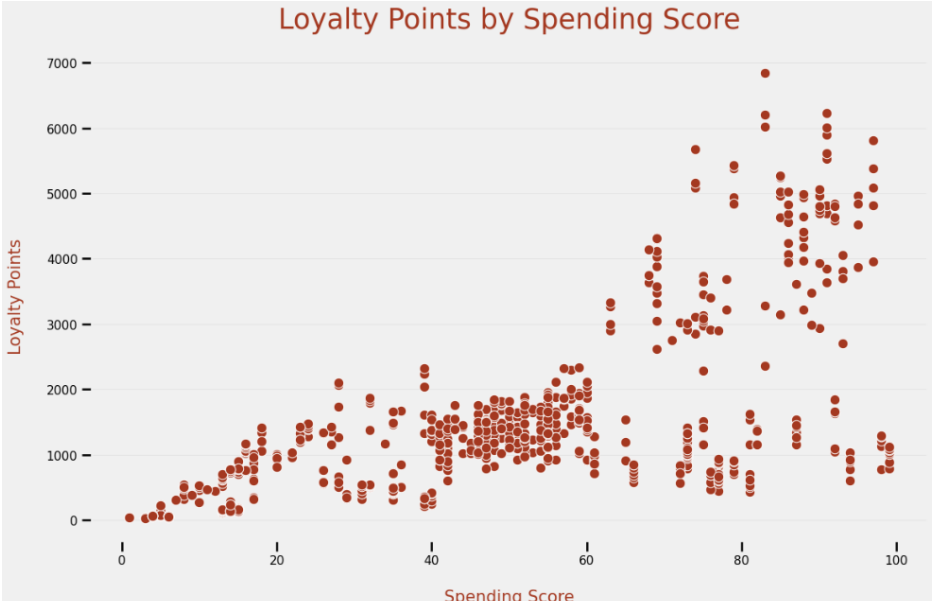

Figure 4 - Example of bivariate analysis.

After the exploratory data analysis, several algorithms were used on the data to analyze patterns and trends, get insights, and encounter the business objectives (more information can be consulted in the Appendix). The following table shows the machine learning techniques and the business objectives each helps to solve.

	Business Objectives		
Machine Learning Technique	**Understand how customers engage with loyalty points**	**Segment Customers**	**Leverage text data from customers' reviews**
Multiple Linear Regression (OLS)	X		
Decision Tree (Decision Tree Regressor)	X		
Clustering (K-Means clustering)		X	
Sentimental Analysis (TextBlob)			X
Sentimental Analysis (Vader)			X

Table 2 - Machine Learning Techniques and Business Objectives.

4. Patterns, trends, and insights

In analyzing the loyalty points data of Turtle Games clients, the distribution exhibits significant skewness (skewed right) and kurtosis (heavy-tailed). Most clients possess fewer than 1500 points, indicating potential for increased retention and recurrent purchases. Despite a wide range (25 to 6847 points), the interquartile range (IQR) reveals that half of the clients have points within a 957-point span. These insights suggest retaining clients to enhance sales and leverage loyalty program benefits.

In analyzing the accumulation of loyalty points among Turtle Games clients, three primary factors emerge as significant contributors: remuneration, spending score, and age. Remuneration and spending score carry more weight compared to age. Turtle Games can enhance client retention and loyalty by strategically focusing on these factors. Implementing targeted incentives aligned with remuneration levels and spending behaviours can encourage more frequent and higher-value purchases. Moreover, personalized marketing strategies tailored to different age demographics can optimize engagement and loyalty program effectiveness. This approach ensures that resources are efficiently utilized to maximize client satisfaction and long-term profitability.

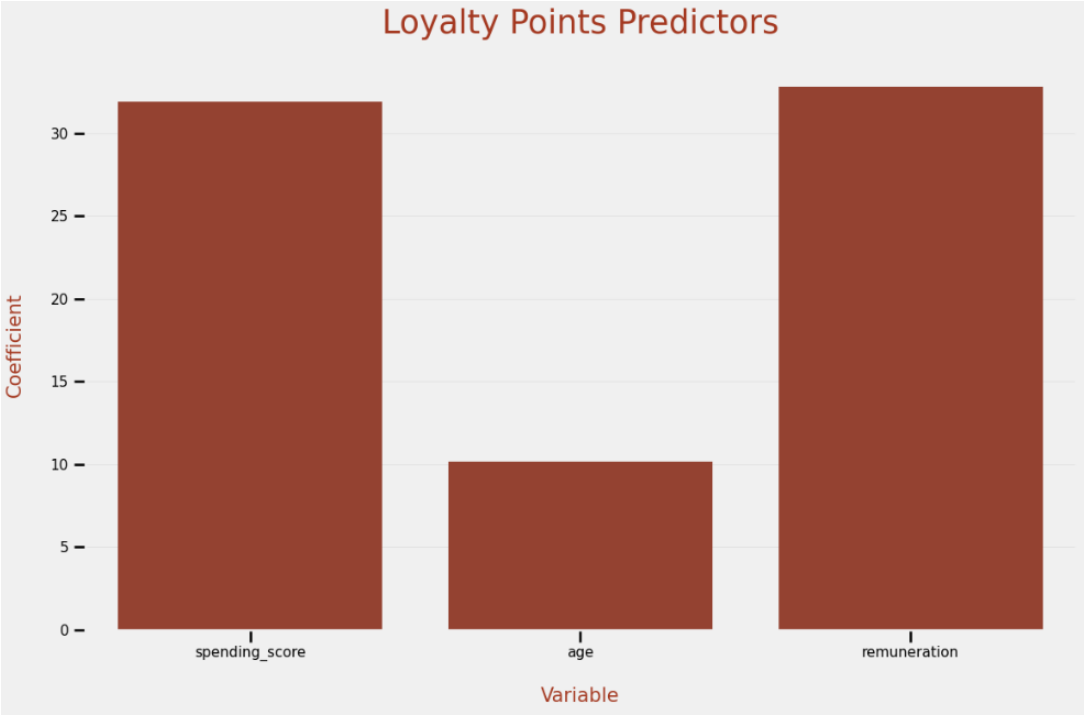

Figure 5 - An increase of 1 unit in each of those variables will bring the following increase in Loyalty Points – Remuneration (+32.84), Spending Score (+31.92), and Age (+10.15).

After analyzing the factors contributing to loyalty points, a clustering process was applied using remuneration and spending score, the two most influential variables. This analysis segmented Turtle Games clients into five distinct groups. By focusing on these key variables, Turtle Games can better understand and target its customer base, enhancing loyalty and driving sales.

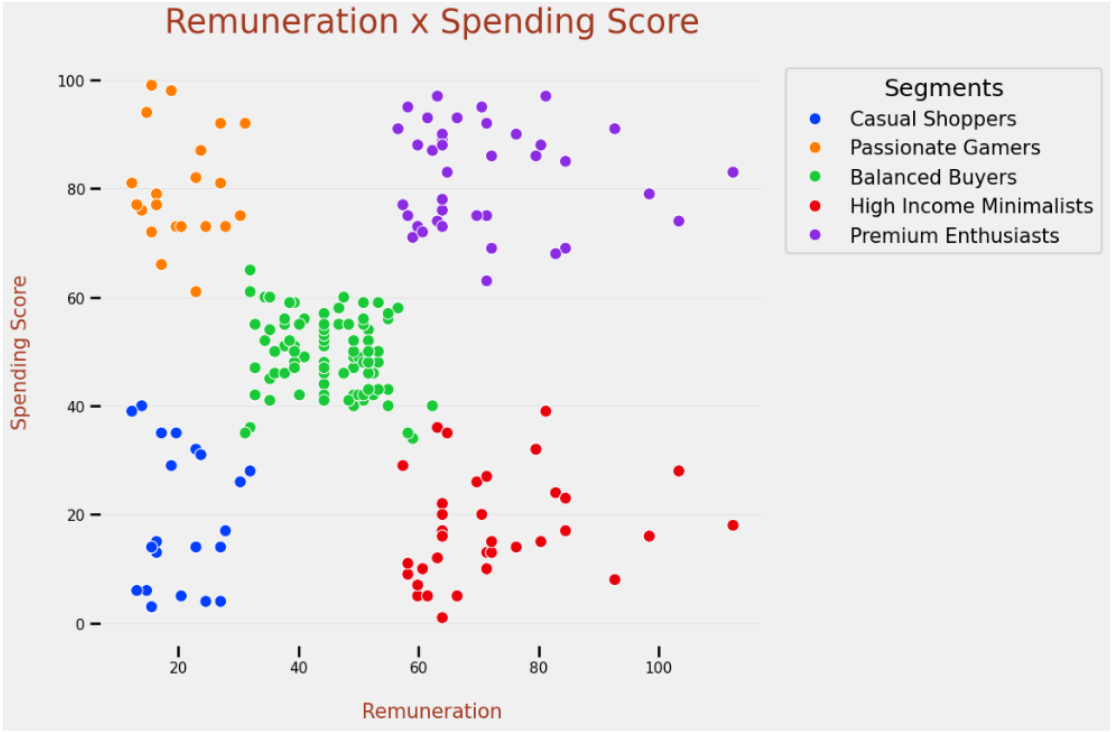

Figure 6 - Remuneration x Spending Score and the correspondent clusters.

Segment	Remuneration	Spending Score
Casual Shoppers	Low	Low
Passionate Gamers	Low	High
Balanced Buyers	Medium	Medium
High Income Minimalists	High	Low
Premium Enthusiasts	High	High

Table 3 - Turtle Games Segments.

Segment	Share	Age	Number of Sales	Loyalty Points
Casual Shoppers	16.9%	41.7	1.98	246.64
Passionate Gamers	16.2%	36.19	2.04	980.52
Balanced Buyers	36.1%	40.29	2.81	1388.88
High Income Minimalists	15.3%	40.48	2.75	953.19
Premium Enthusiasts	15.5%	38.11	2.94	4197.02

Table 4 - Characteristics of each segment.

Turtle Games can implement different strategies for each type of segment according to the characteristics seen in Tables 3 and 4:

- *Casual Shoppers (16.9%)* – With low remuneration and spending scores, these clients make fewer purchases. Strategies include offering frequent small rewards and affordable game promotions to boost engagement and spending.
- *Passionate Gamers (16.2%)* — Despite low remuneration, they allocate a significant portion of their budget to their interests. This group is essential for Turtle Games since they can become Premium Enthusiasts with higher remunerations. Tailored loyalty programs, exclusive content, and promotions on popular games can further increase their engagement.
- *Balanced Buyers (36.1%)* – This largest segment has medium remuneration and spending scores and is a reliable group. Personalized offers, enhanced shopping experiences, and flexible payment options could drive loyalty and higher spending.
- *High Income Minimalists (15.3%)* – With high remuneration but low spending scores, these clients are selective buyers. Exclusive high-end products, limited-time offers, and emphasis on product quality can entice them to spend more.
- *Premium Enthusiasts (15.5%)* — High remuneration and spending scores make this segment of Turtle Games the most essential. VIP programs, premium rewards, early access to releases, and personalized communications can strengthen their loyalty and spending.

Sentiment analysis on customer reviews and product summaries reveals that nearly 90% of comments are positive, while less than 10% are negative. Notably, passionate gamers show a 7.34% negative summary share, which is significant given

their potential as a key target for future growth. Analyzing average review scores uncovers a critical pattern: negative summary scores correlate with low loyalty points and single, non-repeated purchases. To address this, Turtle Games should offer incentives for negative experiences, such as discounts on future purchases, personalized support, or additional loyalty points. These strategies can help turn negative experiences into customer retention and satisfaction opportunities, driving repeat business and long-term loyalty.

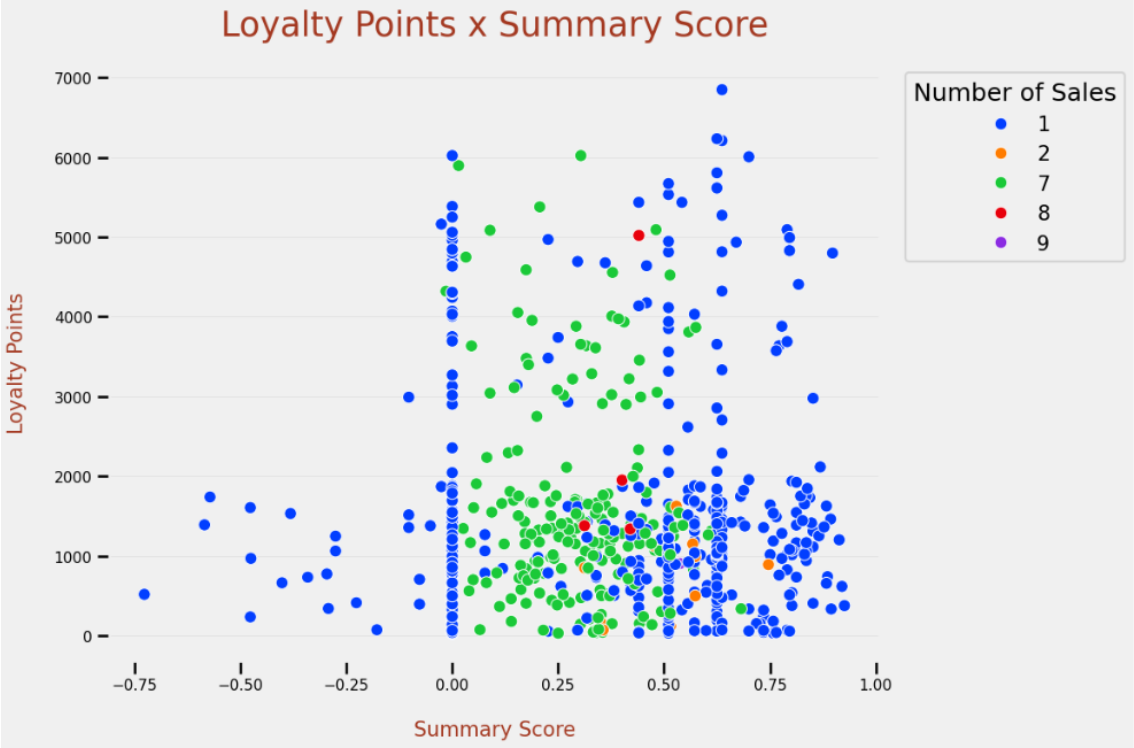

Figure 7 - Summary score by loyalty points and number of sales.

Appendices

Appendix A – Business questions

The business questions clarify objectives, guide data collection, and focus analysis, ensuring solutions are directly aligned with addressing specific business problems effectively and efficiently.

- How can Turtle Games utilize spending score and remuneration data to create targeted marketing campaigns?
- What insights can be drawn from the distribution of loyalty points?
- What strategies can Turtle Games implement to increase loyalty points accumulation among different customer segments?
- How can sentiment analysis of customer reviews inform Turtle Games' marketing strategies and product development efforts?
- How can Turtle Games leverage customer review data to improve product offerings and customer satisfaction?

Appendix B – Recommendations for future work

There are some recommendations for the Turtle Games data collection structure to enhance the data analysis done:

- *Sales data* - Ensure the data frame contains comprehensive sales data, not just customer reviews. This data should include transaction details, purchase dates, quantities, and prices.
- *Accurate client identification* - Implement a system for precise client ID identification to increase efficiency and improve analysis accuracy. This identification will eliminate the need for assumptions and ensure reliable data segmentation.
- *Product information enhancement* - Collect more detailed information about the products sold for deeper analysis. This data should include product categories, discount information, products bought together, product characteristics, price sold, and stock-keeping units, among other aspects.
- *More client information* – Information about demographic and location variables about the clients could drive better insights.

Implementing these recommendations will provide a more comprehensive and actionable data set, enabling Turtle Games to make better-informed decisions to drive sales and customer satisfaction.

Appendix C – Measures related to Loyalty Points

Measure	Value
Skewness	1.66
Kurtosis	5.31
Range	25 – 6847
Difference between maximum and minimum	6822
Interquartile Range	957
Variance	1724604
Standard Deviation	1313.24
Mean	1497.404
Medium	1187
Mode	1014

Table 5 - Measures related to Loyalty Points

Appendix D – Linear Regression

The first step was applying a simple linear regression to Age, Number of Sales, Remuneration, and Spending Score. The Age and Number of Sales presented a low R-squared, while Remuneration and Spending Score were considered moderate predictors.

After that, three different multiple linear regression models were tried: one with Remuneration and Spending Score, another with those two plus Age and Number of Sales, and a third with Remuneration, Spending Score, and Age. This last one presented the best measures of success.

OLS Regression Results

Dep. Variable:	loyalty_points	R-squared:	0.810
Model:	OLS	Adj. R-squared:	0.809
Method:	Least Squares	F-statistic:	1105.
Date:	Sat, 06 Jul 2024	Prob (F-statistic):	6.58e-280
Time:	16:14:44	Log-Likelihood:	-6074.9
No. Observations:	782	AIC:	1.216e+04
Df Residuals:	778	BIC:	1.218e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1999.7279	86.314	-23.168	0.000	-2169.163	-1830.293
spending_score	31.9166	0.778	41.026	0.000	30.389	33.444
age	10.1522	1.514	6.706	0.000	7.181	13.124
remuneration	32.8354	0.813	40.388	0.000	31.239	34.431

Omnibus:	8.164	Durbin-Watson:	1.930
Prob(Omnibus):	0.017	Jarque-Bera (JB):	8.089
Skew:	0.243	Prob(JB):	0.0175
Kurtosis:	3.113	Cond. No.	344.

Figure 8 - OLS summary of the model built with spending score, remuneration and age.

As seen in this report, those coefficients were used to evaluate the relationship between loyalty points and those variables.

The Breusch-Pagan test returned a p-value less than 0.05, which shows that heteroscedasticity is present in the data. A model was created with the squared root of the dependent variable, which solved that problem. The model also increases the adjusted R-squared. That model can be used to predict future clients.

Appendix E – Clustering - Selecting the correct number of clusters

We applied clustering through the k-means algorithm to the remuneration and spending score data to better understand customer segments. After visualizing the scatterplot of these variables, it was evident that five clusters provided a clear and meaningful separation of the data points. We employed two additional techniques to validate this observation: the elbow and silhouette methods. Both methods confirmed that five was the optimal number of clusters, as indicated by a distinct "elbow" in the elbow method plot and a high average silhouette score for the five-cluster solution. This combination of visualization and analytical techniques robustly supports the selection of five clusters for our analysis.

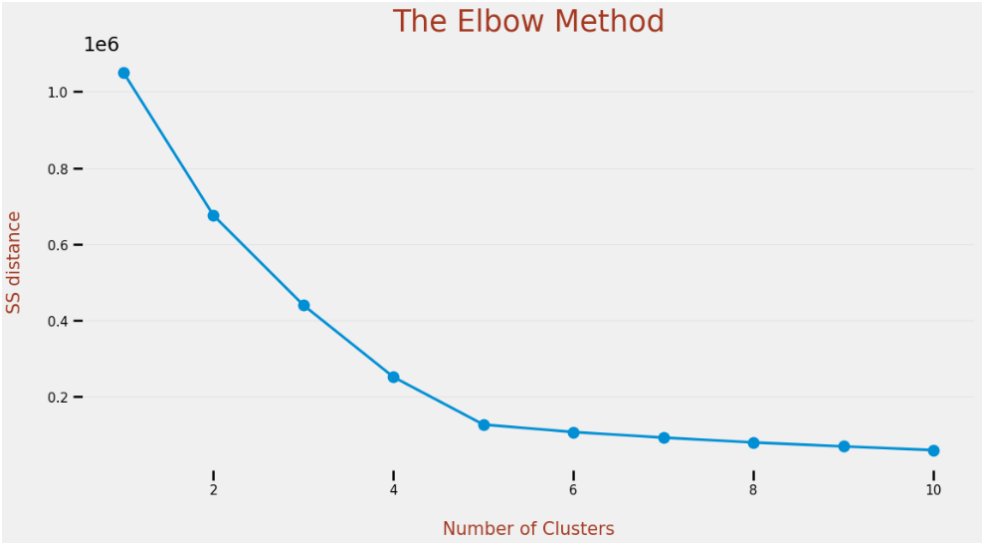

Figure 9 - The Elbow Method.

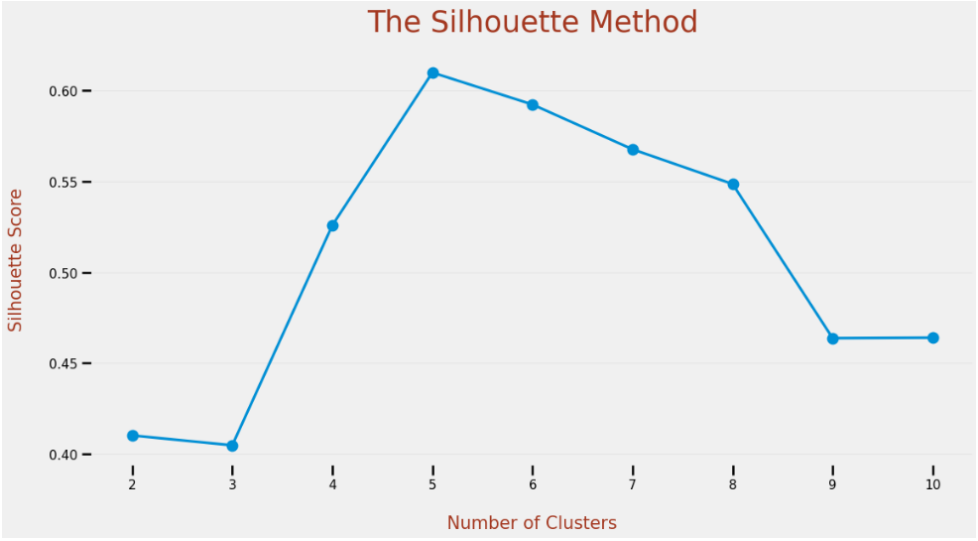

Figure 10 - The Silhouette Method.

Appendix F – Sentimental Analysis

Two sentimental analysis libraries were used – Vader and TextBlob. After all the steps of processing the data, both were employed in the Reviews and the Summaries. To evaluate which model was better, 30 random sentences were selected, and the sentiment of those sentences was manually assessed. It is vital to note that 30 is a tiny sample, and just one evaluation can lead to bias. However, the selection process was done that way due to time constraints and capability.

TextBlob presented 60% accuracy in the reviews and 70% in the summary columns. Vader presented the same for summary and showed 90% for reviews. For that reason, Vader was the tool selected to use for the analysis.

Sentimental Analysis Technique	Column	Accuracy
TextBlob	Reviews	60%
Vader	Reviews	90%
TextBlob	Summary	70%
Vader	Summary	70%

Table 6 - Results of the evaluation of the two techniques.