Predicting future outcomes

Analysing historical data can uncover plenty of insights for organisations and one of the best benefits of it is that we can build predictive models based on the historical data and help businesses plan and manage the risks better. This report contains a thorough description of the data analysis performed for the client, Turtle Games.

Context.

Turtle Games is a game manufacturer and retailer that sells its own products as well as products manufactured by other companies, such as books, video games, toys, board games, etc. The client has been collecting the data from sales as well as customer reviews from a variety of sources. Their objective is to improve their overall sales performance. They asked us to help uncover the insights from the data, namely: how customers collect loyalty points, how they should target specific groups of customers, how the social networks data can be used to inform marketing strategies. They want to know how each product affects the overall sales, and finally, how North American and European sales affect global profit.

Analytical approach.

The data from the marketing department, 'turtle_reviews.csv' file was wrangled in Python and for the sales data, R script was utilised to examine the 'turtle_sales.csv' file. In both cases, the raw data first had to be carefully examined and cleaned, if necessary. The 'turtle_reviews' file was first imported and loaded in Jupyter notebook, the columns were checked and the unnecessary columns, 'language' and 'platform', were dropped. The resulting data frame columns had to be renamed for the ease of wrangling. In R studio, the procedure was similar, namely: the file was imported, columns were viewed and some of the redundant columns were removed ('Ranking', 'Year', 'Genre' and 'Publisher'), data types were checked and I ensured that there were no null values that could potentially skew the data.

Regarding the choice of libraries, in Jupyter notebook, the following libraries were imported: statsmodels (for statistical analysis), from scikit-learn (sklearn): 'train-test-split library' for simple linear and multiple linear regression models; from 'linear_model', 'LinearRegression' was imported for the regression models; for visualisations – matplotlib.pyplot library was used. For the further NLP analysis, the nltk library was imported with the following packages: word_tokenize, stopwords, TextBlob and WordCloud. As for the R script, the following packages were utilised: tidyverse, dplyr, ggplot2. Then 'moment' package was imported for the Shapiro-Wilk test and 'psych' library

to build a correlation plot.

In order to identify how the Turtle Games clients collect loyalty points, the multiple regression model was built with the 'loyalty points' as a dependent variable (y) while age, income and spending scores were set as independent variables (X). The R-squared or the coefficient of determination indicate the proportion of the total variation in y (dependent variable) that is explained or accounted for by X (independent variables). The more independent variables there are, the higher the R-squared value is. The R squared value of age, income and spending score showed 0.839, which means that 84% of the variability in loyalty points is based on the combined independent variables – age, income and spending score.

However, if we look at the coefficients of independent variables, the coefficient of 'age' only accounts for 11.06 increase while 'income' and 'score' seem to be more important variables in predicting and explaining the loyalty points (34.01 and 34.18 respectively). For each one-unit increase in income, the estimated increase in 'loyalty points' is approximately 34.01.

When building a train-test model, the data set was divided in 80-20 proportion. That is, 80% of the data was tested (with 1600 observations) and the remaining 20% of the observations were used to train the model. The OLS regression results now show t-values (age – 11.313, income – 62.004, spending score – 67. 253). The R squared is 0.842 in the test model. The model was then tested with some random values. I set the new_age to be 47, income – 15.20, and the spending score – 95. The predicted outcome was 2081. According to the model, a person with the above data could have 2081 loyalty points.

OLS Regression Results

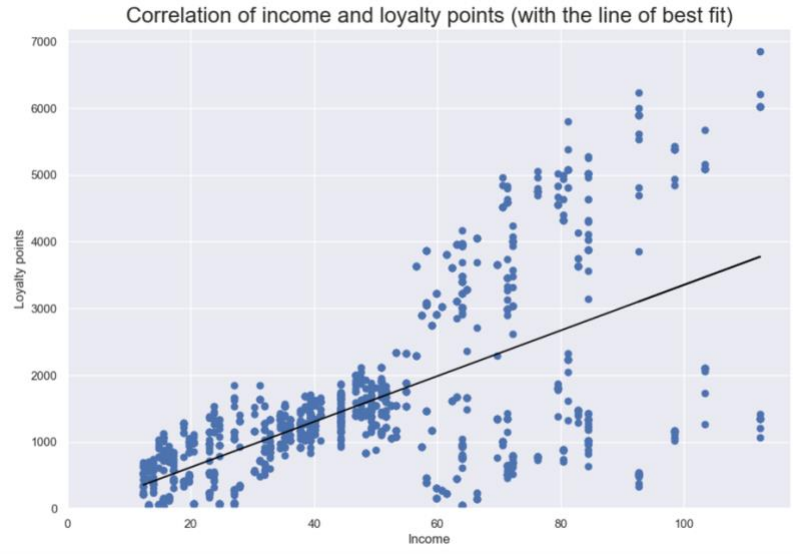
Dep. Variable:	loyalty_points	R-squared:	0.842
Model:	OLS	Adj. R-squared:	0.842
Method:	Least Squares	F-statistic:	2846.
Date:	Mon, 24 Jul 2023	Prob (F-statistic):	0.00
Time:	12:55:03	Log-Likelihood:	-12246.
No. Observations:	1600	AIC:	2.450e+04
Df Residuals:	1596	BIC:	2.452e+04
Df Model:	3		
Covariance Type:	nonrobust		

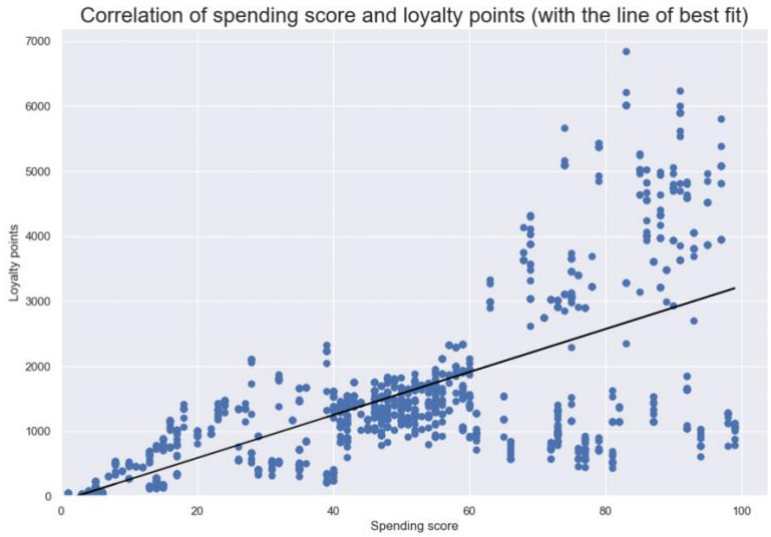
	coef	std err	t	P>|t|	[0.025	0.975]
const	-2197.0105	58.134	-37.792	0.000	-2311.037	-2082.984
age	11.0137	0.974	11.313	0.000	9.104	12.923
income	34.2457	0.552	62.004	0.000	33.162	35.329
score	33.9681	0.505	67.253	0.000	32.977	34.959

Omnibus:	14.722	Durbin-Watson:	2.050
Prob(Omnibus):	0.001	Jarque-Bera (JB):	15.856
Skew:	0.189	Prob(JB):	0.000360
Kurtosis:	3.308	Cond. No.	377.

Visualisations and insights.

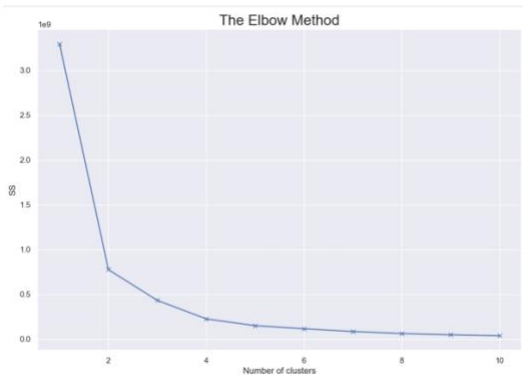
With a multiple linear regression model, it is possible to create a plane showing all three variables used. However, for the purposes of this analysis, I decided to build simple linear regression models with each of the variables – age, income and spending score. As expected, 'age' variable does not play such a significant role in predicting the loyalty scores. However, income and spending scores did show a weak positive correlation between the variables.

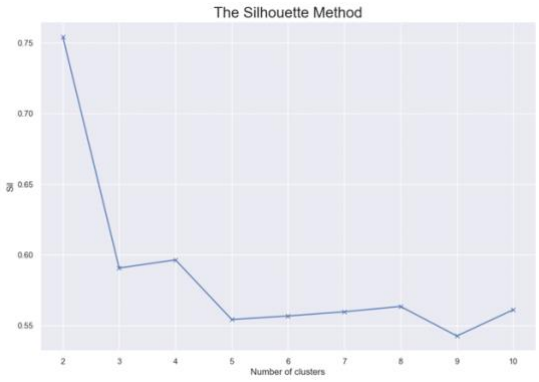




Having identified that remuneration and loyalty points are strongly correlated, it was decided to try and group the customers using k-means clustering technique. Clustering is a type of unsupervised machine learning. We needed to identify groups within the customer base that can be used to target specific market segments. Clustering is a means of data reduction. We try to reduce the data to a smaller number of clusters. The fewer clusters we have, the greater data reduction is achieved. However, with fewer clusters, we might get greater heterogeneity within those groups with a larger ss value. There has to be a compromise when choosing the number of clusters that could give us greater homogeneity within each cluster, but at the same time, good data reduction.

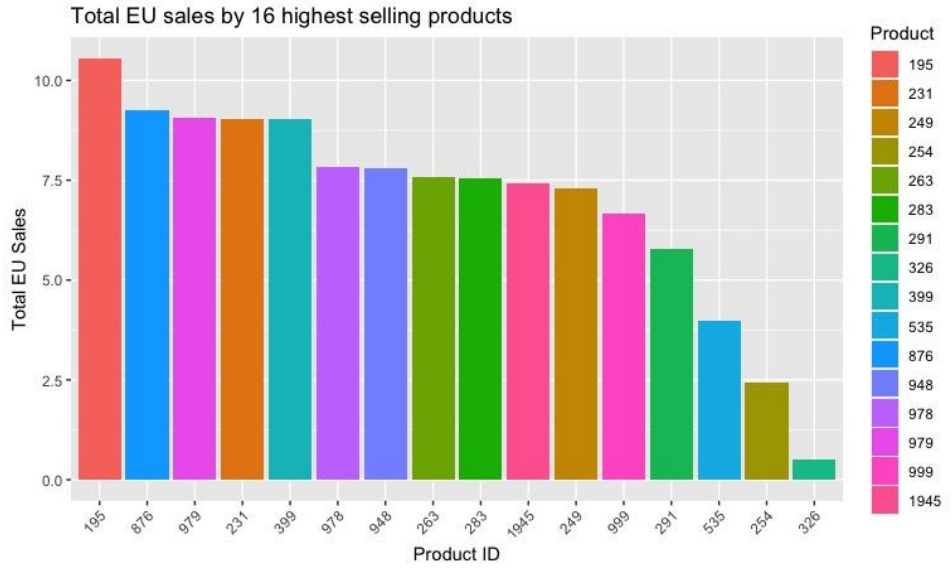
To choose the optimal number of clusters, Elbow and Silhouette methods were used. In the Elbow method, we look at the SS (sum of squares) factor, and when the WSS (within – cluster sum of squares) start to diminish, it is visible on the plot. Here we could either choose 3 or 4. I had to cross-check with the silhouette method, and it is evident here that the silhouette score of the k=4 is closer to 1, which is why 4 clusters seem to be the best option.

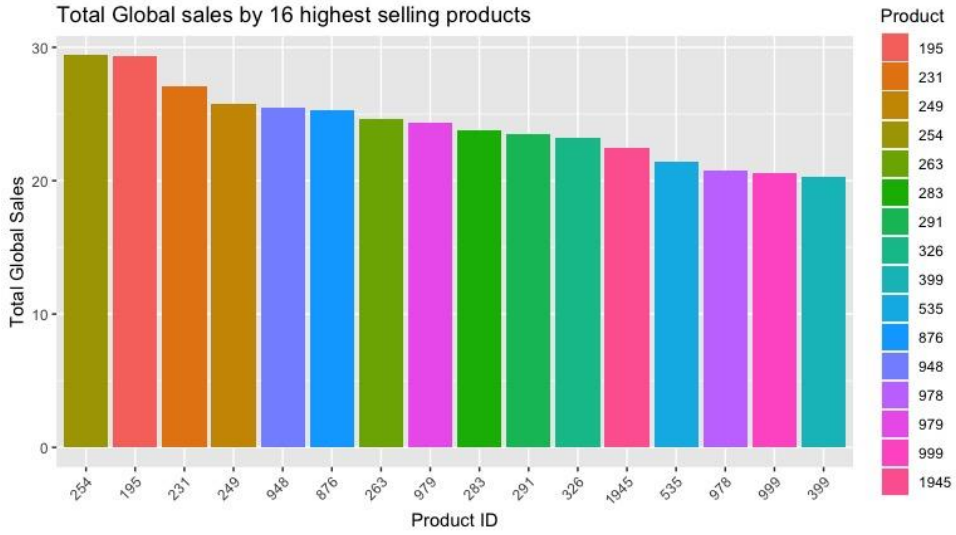


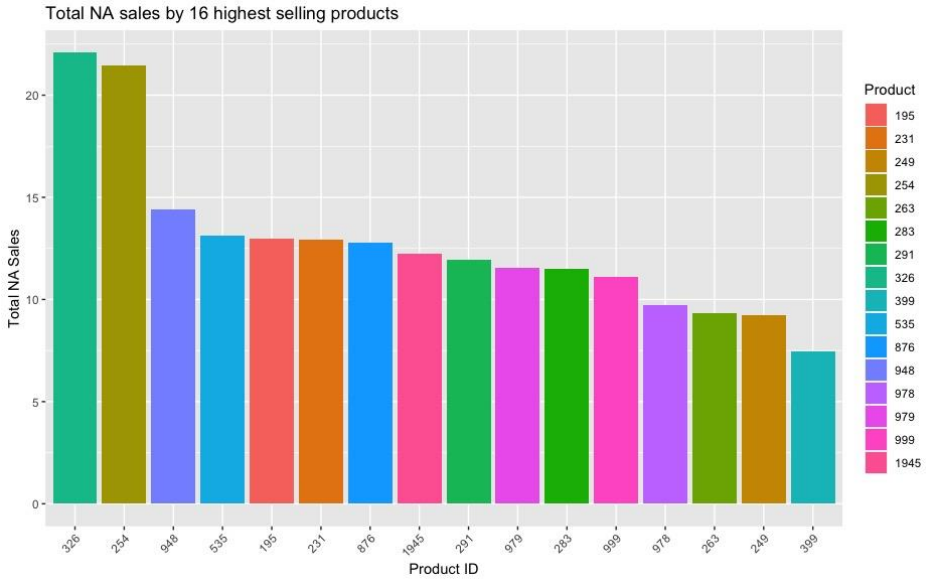


Other important plots and graphs were derived from working with ‘reviews’ and ‘summary’ columns and the most frequently used words were identified here. Looking at the most frequently used words, the word 'game' stands out in both columns with 1359 and 268 times respectively (review and summary columns). 'Play', 'fun', 'great', 'like' also dominate in both columns suggesting that most of the customers enjoy the games selection from Turtle Games and are quite happy with the purchase. The word 'book' was also seen quite often (259 and 42 in both columns) suggesting that books are a quite popular product, but still bought not as often as games.

In R, the most prominent plots that are worth mentioning here are the top selling products. They quite clearly indicate the best-selling products in Europe, Northern American countries and globally. This bar plot would be very important for the sales department of the Turtle Games.







Patterns and predictions.

There were several patterns and insight identified important both to the marketing and sales departments.

Marketing:

1. The higher the income and spending score of the customer, the higher his loyalty score will be. These are the loyal customers that are perhaps bringing the bulk of the revenue. Some loyalty programs or discounts might be introduced to them as a way of encouragement and appreciation.

2. Most of the positive reviews and high frequency words are related to video games mostly. Many of the negative reviews and negative sentiment was in relation to board and card games. They were either unclear or confusing and the quality of the material was bad. Perhaps, clearer instruction and a better consultation should be given to the customers in-store and better descriptions in online stores.

Sales:

1. Highest selling platforms are: Wii, X360, DS, GB, PS2, PS3 and PS4. Highest selling products are 195, 876, 979, 326, 948, 231 across Europe, NA and globally.

2. It is possible to predict the global sales according to Northern American and European sales. Northern American sales account for 91% of the global sales and European sales – 85%.