Predicting Future Outcomes: Report

By Yi Hu Huang

Turtle Games is a game company manufacturer and retailer, meaning it created and sells its own products, in addition to sourcing and selling other companies' products in their stores. The company sells a wide range of products, including books, board games, video games, and toys. For this assignment, I am working as a data analyst that has been contracted by Turtle Games to find out how to improve overall sales performance by using customer trends that the company can find in data collected from their sales and customers.

For this task, I have used Python and R to analyse the data. After first importing the necessary data and converting it into a Data Frame, I cleaned the data by making sure there were no missing values. Once this was confirmed, I looked at the Data Frame and dropped columns which I deemed would be unnecessary or irrelevant for the assignment. For example, for the *reviews* Data Frame, I dropped three columns – language, platform and education. I dropped language because it was irrelevant as all reviews were in English; I dropped platform because also all reviews were on Web; and I dropped education because I did not think this would impact much or have enough significance in analysing trends – for example, someone might have a high-level of education but low renumeration, so between these two variables, renumeration would mean the most. I then created OLS Regression tables with different variables to see if there was significant correlation between the two variables. Firstly, there was positive significant correlation between a customer's spending score and their number of loyalty points. This correlation makes sense, as the more a customer would spend at Turtle Games, the higher their loyalty points. The scatter plot does show some anomalies wherein people with a high spending score do not have that many loyalty points (see Figure 1) – a potential reason for these cases could be that these customers are coming in often and purchasing items, but they are purchasing low-cost items, as loyalty points is converted from the monetary value of the customer's purchase.

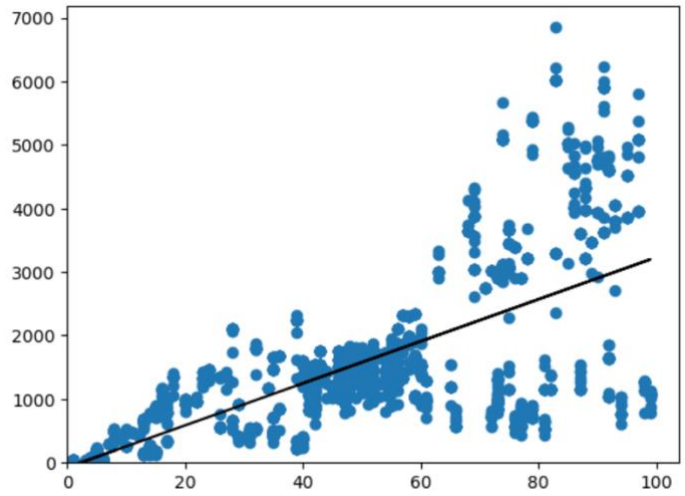

Figure 1

The comparison of renumeration vs loyalty points also had a positive significant correlation (see Figure 2). Similarly, this correlation is not much of a surprise, as people who would have more money to spend, would probably be able to spend more on items that Turtle Games sells. People who have a higher income per year will more willingly spend money on their wants as they have a higher sense of security when they earn more, when compared against

a customer who might earn a lower income thus would have to think carefully about spending their money on items that Turtle Games sells which are more for leisure and pleasure than for necessary living purposes.

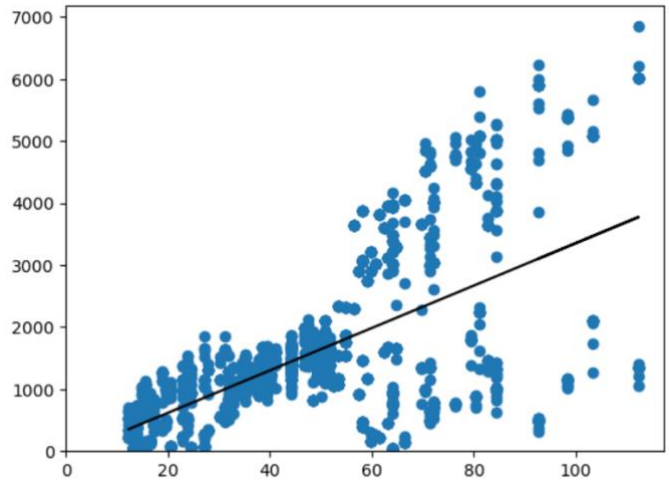

Figure 2

Lastly, I also compared whether there was any correlation between age and loyalty points. After performing the OLS model on the data, the regression table showed that there was no significant correlation between these two variables. The scatter plot below will help visualise that there is no correlation at all between age and number of loyalty points (see Figure 3).

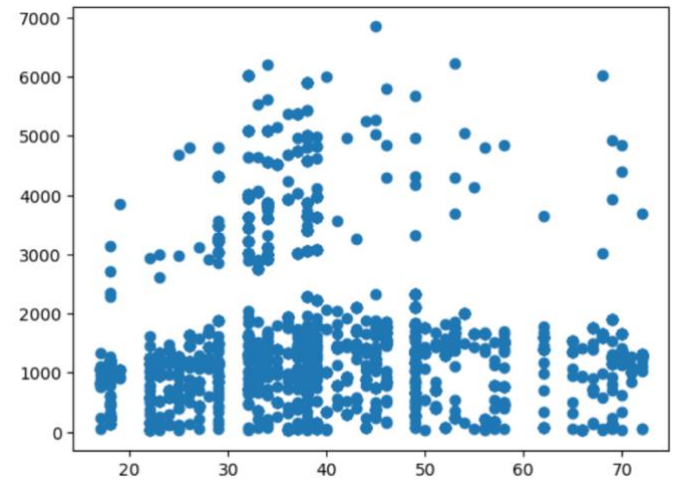

Figure 3

Seeing as the variables renumeration and spending score are the variables which will help the most in figuring out Turtle Games' customers, I created a cluster graph with five clusters as this was deemed the optimal number of clusters according to the Elbow and Silhouette methods (see Figure 4 on next page). Customers in cluster 0 were not only high-spenders at Turtle Games, but also had a high income – people in this cluster could be assumed to be big spenders, have a high passion for games and books, could potentially have a family and that is why they're spending money at Turtle Games for their kids, could be single and able to spend their income on items for their own satisfaction or hobbies. Customers in cluster 1

seem to be the safe or neutral people – they earn an average income and they are neither very interested or lowly interested in the items that Turtle Games provides. This could be a difficult cluster to target as they are neutral and potentially have no strong opinions on Turtle Games. Cluster 2 groups customers who have a high income but do not spend a lot at the company – this could mean that they are not interested in the items that Turtle Games sells, and they might just buy every so often if Turtle Games sells the one thing they need to buy for themselves or someone else. Cluster 3 represents those who have a low income but still have a high spending score at Turtle Games – these people could be classified as possible game-fanatics or book-lovers, depending on what they purchase, as they are willingly spending a possible large proportion of their income on leisurely items that Turtle Games sells. Additionally, these customers may also be people who are still relatively young, thus at the start of their career and earning a low income, but can afford to purchase a lot and have a high spending score as they still use their parents' money. Those in Cluster 4 are the customers of less interest for Turtle Games, as they do not have a high income nor do they have a high interest in the products that Turtle Games sells, thus they would be hard to target and turn their opinion around on the company and make them purchase more.

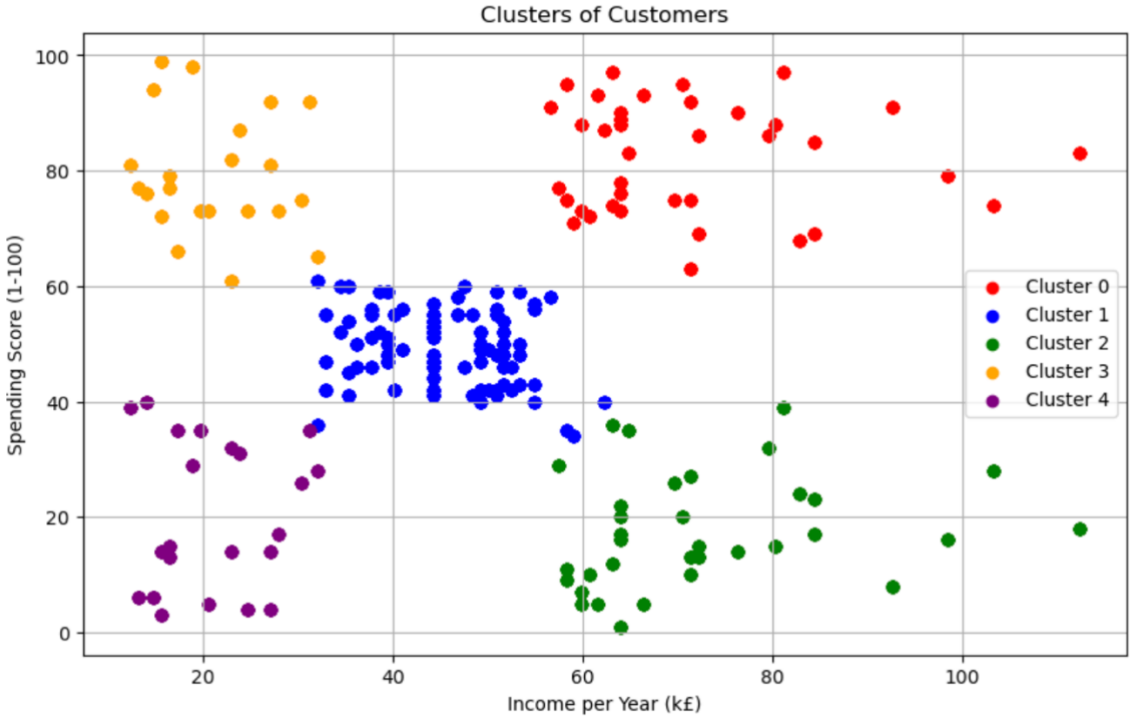

Figure 4

Social data gathered from Turtle Games' reviews did not provide any significant results, as most reviews just emphasised regular terms that would be used for a company that focuses on creating mostly games – when creating a word cloud, the largest terms were words such as fun, game, etc. When creating a sentiment analysis, the overall feel of the reviews was positive as most reviews had a positive score. This analysis on text data can suggest that customers are happy with Turtle Games and have neither any particular complaints nor particular unique points of happiness. The implication of this positive neutrality feel from customers towards Turtle Games could be that the company is in a good place to move upwards.

The only relationship found between North American, European and global sales is that if North American and European sales are bad, global sales will also be bad. Additionally, North American and European sales tend to be similar, which implies that if a product were to sell a lot in North American territories, it would also sell a lot in European territories, and vice versa. Another implication of this statement is that North American and Europe are Turtle Games' most relevant and biggest territories, as their global sales numbers highly depend on their North American and European sales numbers.