TURTLE GAMES
PREDICTING FUTURE OUTCOMES

Dilan Koc

Introduction

Turtle games, as a global company, aims to improve overall sales performance based on data analysis of loyalty points, customer sentiment, and global sales.
This report outlines the approach used and insights observed in the data analytics process of predicting future outcomes.

Business Questions

- How do customers accumulate loyalty points?
- How useful are remuneration and spending scores data?
- Can social data be used in marketing campaigns?
- What is the impact on sales per product?
- The reliability of the data?
- If there are any possible relationships in sales between North America, Europe, and Global sales.

Data wrangling and exploration

The datasets were imported into a Jupyter Notebook and R Programming from a .csv file for analysis. The shape and data type are analysed before determining variables of interest. Also, descriptive statistics and exploratory visuals are used to better understand the distribution of the datasets. During the exploration, some issues were identified, and datasets were prepared to be analysed:

- Unnecessary columns, for this analysis, dropped.
- Column names have changed.
- Duplicates removed.
- No empty columns were found.
- Visuals of the variables and statistical results show that the data set does not follow a normal distribution.

PYTHON

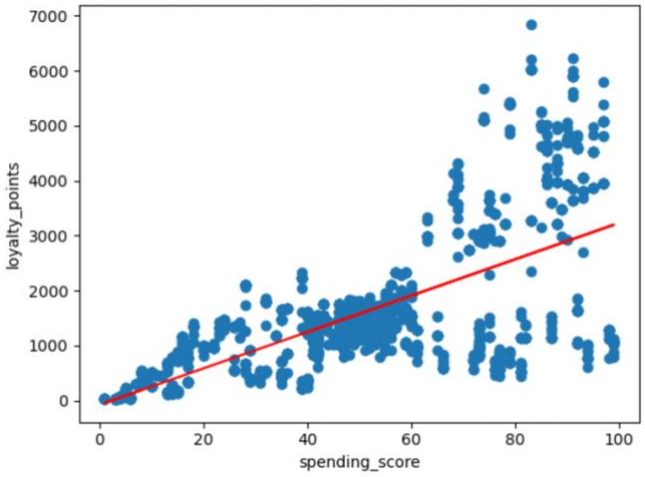
Week 1

- **Q: How do customers accumulate loyalty points?**

To accumulate loyalty points (dependent variable) of a customer, a relationship needs to be identified in the given dataset. To achieve this, regression analysis will be used. Each variable will be used separately in simple linear regression to find out which variable has an impact.

OLS Simple Linear Regression model will be used to explain the relationship.

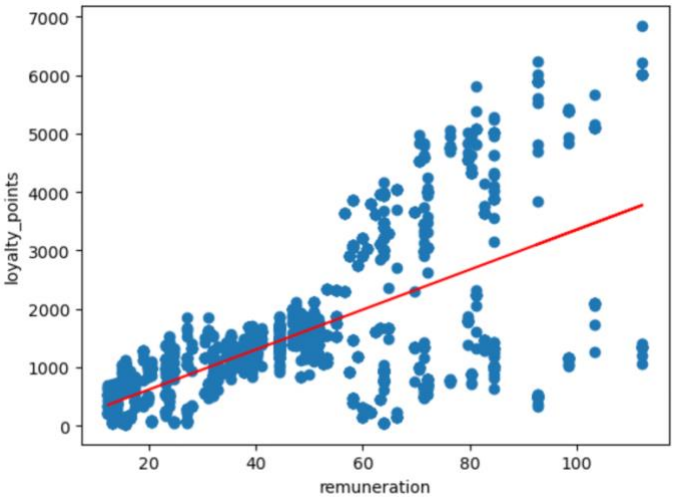
1) **Spending score versus loyalty points.**



- OLS Regression results show that the variables are not statistically strong enough to allow predictions. R-squared value confirms that 45% of the variability observed in the target variable is explained by the regression model.

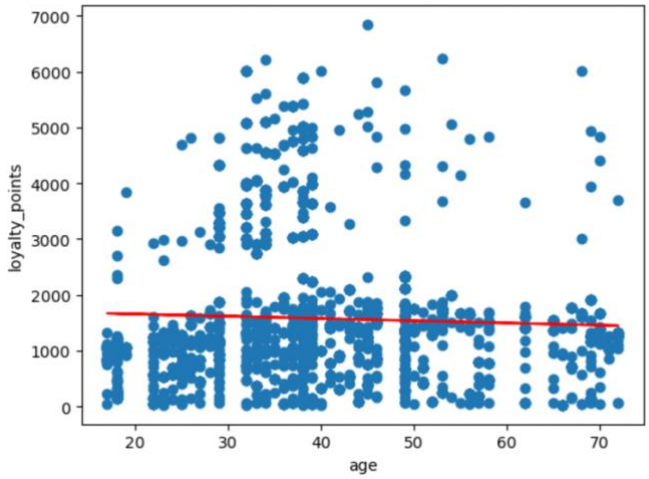
- Converting the variables by using the sqrt () and log () transformation improved the R-squared value to 67% and showed that there is a positive relationship between the variables but still is not statistically significant to rely on.

2) **Remuneration versus loyalty points**



- Both log () and sqrt () functions are used to improve the accuracy of the variables. The highest R-squared value is 39%, which is not enough to allow predictions between those variables.

3) **Age versus loyalty points.**



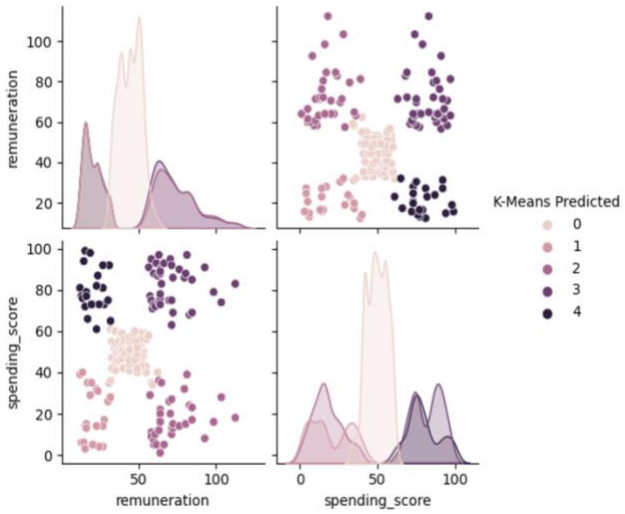
- The age variable also has no impact or relationship on the loyalty points; therefore it cannot be used for predictions. The R-squared value for this model is 0%.

As simple linear regression did not give the desired outcome, by adding all the variables together to explain loyalty points, a multiple linear regression model is used. That way R-squared value confirms that 83% of the variability observed in the target variable is explained by the multiple linear regression model.

Upon request, this model can be used to allow predictions for the stakeholders.

Week 2

- **Q: How useful are remuneration and spending scores data?**



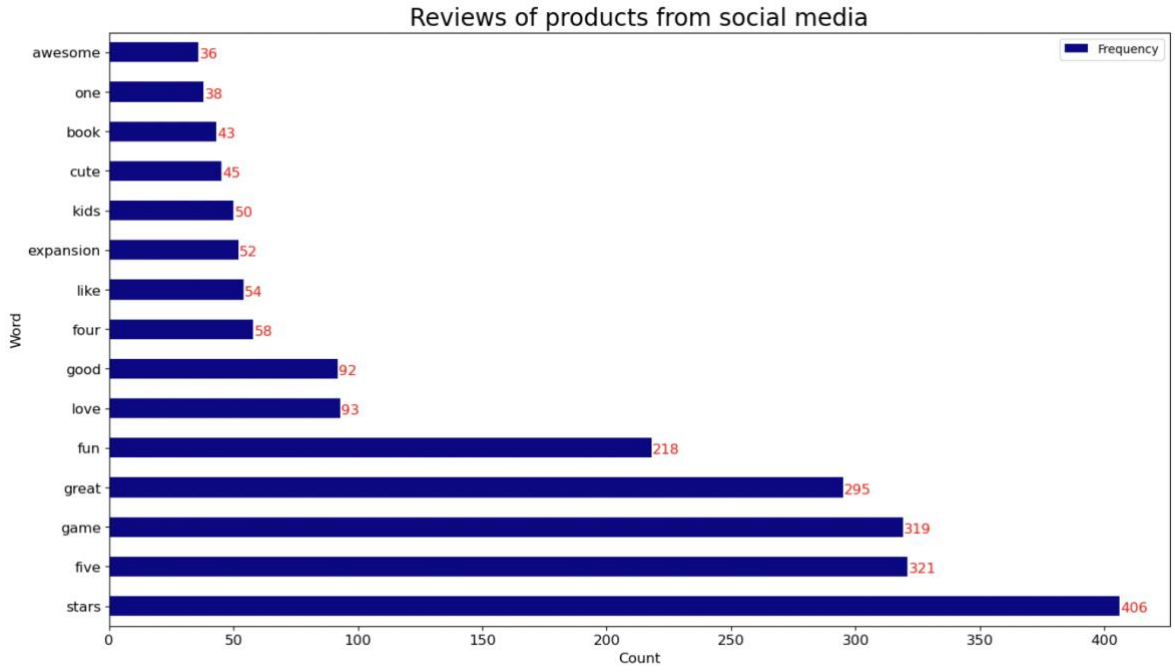
Explain the usefulness of the remuneration and spending scores, variables plotted to gain insights. Visual check proved that the values can be grouped.

K-means clustering will help to identify the groups, and to identify how many groups can be found in the dataset both the Elbow and Silhouette method will be used.

- The Elbow and Silhouette methods proved that there are five different groups.
- Those groups can be used to target customers to increase their sales.

Week 3

- **Q: Can social data be used in marketing campaigns?**



Sentiment analysis will help to analyse, how satisfied turtle games customers are. Frequency distribution and polarity scores of the sentiment analysis can help us reveal the scores of customer satisfaction.

- The most common 15 words are illustrated above.
- Polarity scores validate that 90% of the reviews are neutral and positive.

R – Programming

The analysis will be finalised by using R programming.

Week 4

- **Q: What is the impact on sales per product?**

Using the product ID to declare the impact on sales could only be done by sorting the columns, as visually would be overwhelming.

- Product 107 sold the most both in North America and Europe.

Week 5

- **Q: The reliability of the data?**

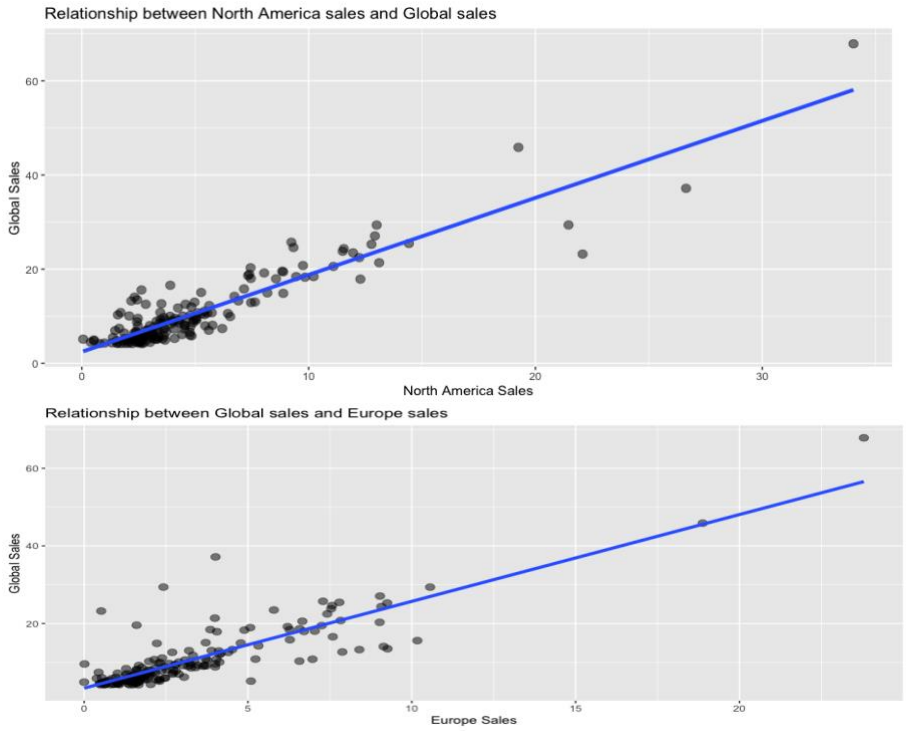
Creating a histogram and running some statistical tests would give the result of how reliable the dataset is.

- The histogram visuals identified that the sales columns do not have a normal distribution and are skewed to the right (positively skewed).
- Skewness and kurtosis statistical tests provide evidence of highly skewed data on sales columns and outliers.
- Shapiro-Wilk test with the small p-value also suggests that the data is not normally distributed.
- Data normalisation might help to transform the dataset.
- Outliers can be replaced with the average value.

Therefore, outliers which also skew the distribution of the dataset should be further investigated to find out if they are the real number or some error while collecting the data.

Week 6

- **Q: If there are any possible relationships in sales between North America, Europe, and Global sales?**



Using another statistical test on sales data, a strong positive correlation was determined in the sales columns. Simple linear and multiple linear regression will be used to identify any relationship if exists.

- A strong correlation was identified between Europe sales – Global sales and North America sales – Global sales and proved by the statistical results.
- Correlation between Europe and North America with a low R-squared value of 50% does not give enough confidence to allow predictions.
- Multiple linear regression with all sales data with 97% confidence, the target variable (Global Sales) can be explained by the regression model.
- In any case, both models allow us to make predictions.

Conclusion

- Identifying how customers accumulate loyalty points can be explained and predicted with multiple linear regression with the highest accuracy.

- Creating five distinct groups with clustering can give an idea to the marketing team to target customers in order to increase overall sales. Such as:

    - Customers with the highest income and low spending (Cluster 2).
    - Customers with the highest income and high spending (Cluster 0).
    - Customers with low income and highest spending (Cluster 3).

- Further analysis is needed on sentiment analysis to find out about what, and which products bring a positive, negative, and neutral review. So, the marketing team can focus on their strength and weak point in their next marketing campaigns. Also, could be useful for the publisher to develop their product or service.

- Product 107 brings more sales, in North America £34M and Europe £24M. Turtle games could focus on this product in the next marketing campaign.

- Certainly, there is a strong correlation between Global sales, North America sales and Europe sales. The dataset shows that sales in Europe are slightly lower than in North America. Targeting Europe in marketing could be a good opportunity to increase sales.

- Further analysis is required to investigate outliers which may cause inaccurate results.