Assignment Report Course 3

Introduction

To improve the overall sales performance Turtle Games wants to understand:

- How customers accumulate loyalty points.
- How customers can be grouped to specific segments
- How social data can be used to inform marketing campaigns
- How the products impact sales.
- Whether the data is reliable
- What the relationship between the sales regions is.

Cleaning the data

The data was loaded into a Jupyter Notebook and first analyzed based on the structure and data types. There are 352 rows and 9 columns in the sales.csv file and 2000 rows and 11 columns in the review files. Some descriptive statistics were deployed to understand the distribution of the data. To simplify the workflow the unneeded columns were dropped from the review data. There are no missing values that needed to be replaced.

How do customers accumulate loyalty points?

To answer this question the reviews data frame was analyzed. Tests for linearity on different variables against the loyalty points were performed. The result of the analysis is that the data is not normally distributed for any of the variables. Age, Annual Salary, and Spending Scores were analyzed and there is a high correlation between spending behavior and loyalty points. Therefore, the conclusion to the first question is that customers are likely to acquire loyalty points by having a high spending score.
To follow up on this, I would recommend clustering the customer based on the spending score and check if there are any relationships between education, positive review scores or gender.

How can customers be segmented?

The second objective was approached by subsetting the data frame to salary and spending score as the two variables would define the customer segmentation. After visualizing the data, it was visible that there are 5 clusters. To verify this, the elbow and silhouette methods were applied and in both cases five clusters turned out to be the ideal number. Two more tests were performed with 8 and 4 respectively but the cluster size is most evenly distributed with 5 clusters. The central cluster way always the largest one with an annual salary between 31.000-62.000£ and aged between 34-61 years old.

How can social data be used to inform marketing campaigns?

To answer this question, NLP was applied on the review’s dataset. The cleansing activities for the text consisted of subsetting the data and tokenize the words so that they can be analyzed. Moreover, the text was converted into string and punctuations as well as stop words were removed to produce more meaningful insights. The overall consent is positive, and the common words are ‘family’, ‘children’ & ‘gift’. As an action these comments could be analyzed more in-depth to understand why the customers liked it. Finally, a sentiment score analysis was performed to understand whether customers like or dislike the products. Overall, the scores tend to be more positive, and the score seems to be consistent with the top and bottom 20 comments.

How do products impact sales?

This workflow was performed in R, where the sales data was analyzed in detail. After sense-checking the data a boxplot was created to compare the sales across the different sales regions. The dataframe was subsetted and melted so that all three sales columns could be compared with each other and the result showed that North America has the higher average sales compared to EU. To answer the core question, how products impact sales, the data was summarized by product with the sum of global sales. The result of the top 10 products were very close after the 6th spot, therefore only the top 5 were shown. Product 107, 515, 123, 254 and 195 are driving the most sales globally.

How reliable is the data?

To answer this question, the data was first aggregated into the sum of sales across all the regions. After evaluating the data visually, it was visible that the data is not normally distributed, there is a high density in the center and a lot of outliers for all of the subsets. After a plotting of the residuals in qqplots it was even more visible that the data is not well suited for statistical tests. A Shapiro test as well as the test for kurtosis and skewness did nor result in better numbers. Therefore, the data is not fit to be used to perform inferential statistics but nevertheless, can still provide valuable insights.

What is the relationship between the sales number in the different regions?

A scatterplot was created to show the relationship between NA-EU, NA-Global & EU-Global. This was done with ggplot and a line of best fit was added to determine the type of relationship. Unsurprisingly the data Is highly correlated since the Global Sales data is dependent on both the NA and EU sales. The relationship between NA and EU Sales is positive but has less predictive power.
A multiple linear model was created to predict Global Sales with the NA and EU sales data and the model is almost a perfect fit with an R-squared of 0.96. After performing some tests on predicted vs observed values, the model’s predictive power is confirmed.

While the model is a great fit, I would complement this with other indicators for example annual inflation or customer sentiment, as Sales from the regions are naturally correlated to the Global Sales figure.