**ASSIGNMENT 3: Predicting Future Outcomes**

Analysis Objectives

The main objective of this analysis is to provide the key stakeholders Turtle Games with insights and recommendations into which products they should be targeting their marketing on to increase sales globally. To achieve this, data from sales figures and customers will be imported, cleaned and wrangled to allow easy to understand visualisations to be created, which will then be presented to the stakeholders, allowing them to make informed decisions based on the presentation.

This report and presentation will propose any actionable insights that are found from the analysis.

Data collection and wrangling process

The CSV files provided have been imported into 2 programming languages, Python and R. Refer to the uploaded files accompanying this report to see the code used.

The data cleaning process included:

* Checking and replacing null or NA values where necessary
* Checking the data types and changing as required
* Renaming and dropping columns of data that are not required
* Aggregating and filtering the data to prepare it for visualisations
* Summarising the data to preform descriptive and statistical analysis

The data has been plotted on graphs and in visualisations for two reasons. To sense check and perform descriptive analysis, which has enabled the data to be analysed for normality and define the number of clusters, for example. It has also been used to present aesthetically pleasing, simple diagrams to be presented to the stakeholders.

Accumulation of Loyalty Points and Groups within the Customer Base

There is a correlation between the spending score and loyalty points, and remuneration and loyalty points. The customers can be split into 5 distinct groups when comparing their spending scores against their remunerations. These groups can be classified as; Low Spend, low remuneration; high spend, low remuneration; low spend, high remuneration; and high spend, high remuneration and average spend, average remuneration. There are 2 groups of interest here, a group of customers on under 30k who have a high spending score, and the customers who are on over 60k and have a high spending score. It may be worth analysing what marketing the first of those groups are receiving, and see if the same marketing strategy needs to be applied to the customers earning under 30k, who have a low spending score. We should also continue to target the high earners.

Using social data to inform marketing campaigns

Lots of positive sentiment from the customer reviews for young children (8 and 9 year olds). Books and sticker books mainly. Also, there is very good reviews of the anger management book for both adults and children. One review is from a therapist so this endorsement could be used as part of an advertisement for this product. Some of the negative sentiment includes the use of words such as ‘difficult’ and ‘boring’. It is worth noting that none of the most popular words are negative though.

The impact each product has on sales

A major factor in of each product on sales is the genre and publisher. For example, from the data EA Entertainment make very popular/high selling sports games and Activision are very popular in the shooter genre. As well as the genre and publisher, the platform has a large impact on sales. Xbox 360, PS3 and PC are the most popular platforms.

How reliable is the data?

The skewness values of greater than 1 for the NA, EU and Global Sales indicates that the data is skewed to the right. The QQ-plot also supports this skewness. This is also visible from the histograms, but these make sense as most products fall into the lower sales figures of under £10M for Global Sales. The data is not normally distributed so the residuals have been verified by plotting, after the linear regression model of NA vs EU sales was created. There is no pattern in the residual plot, therefore the relationship between the NA and EU sales is linear. The correlation between NA and EU sales are positively correlated as the correlation value is greater than 0. As it is closer to 1 than 0 square rooting any of the data to remove the skewness could affect this correlation. There are also outliers in all 3 variables, where some of the product sales are very large. It is assumed that these sales figures are correct.

Relationship between different sales regions and Global Sales

There is a relationship between EU and NA sales for each product. Popular products are overall, popular in both regions. There are 3 products that are a lot more popular in North America than in Europe. This is also evident from the Adjusted R-Squared value of the linear regression model. These 3 products are all Nintendo products, and the majority of the sales are from the 1980’s.

Insights and recommendations

Business recommendations from this analysis are:

* There is a section of the customer base that are the lower earners but have a high spending score. Determine if there is a reason for this (ie. marketing campaign) and if there is then apply the same strategy to the other low earners.
* Target marketing at high earners as a large number of these have a high spending score
* The anger management product has very good reviews including a review from a professional. This product(s) could be marketed as being effective
* Customers are very happy with the toys and sticker books for under 10’s.
* Certain publishers are more popular for some genre’s than others. Marketing campaigns should be targeted by this.
* Certain platforms generate much larger sales than others. These should also be targeted.
* Targeting European and North American markets will be most beneficial as this is where the majority of Global sales are made currently