Shiv Chandaria
Assignment 3
Word Count: 1030

DA 301 – Turtle Games Analysis

Turtle Games is looking to improve their overall sales performances through analyzing data they have collected from sales and customer reviews. To achieve this, they want to better understand the reliability of the data, customers loyalty, different customer groups, reviews, and relationship between sales across products and regions. A successful analysis will reveal information to inform Turtles Games decisions to boost sales performance.

While Turtle Games has contextualized the problem well, before the analysis I would ask to clarify how data like loyalty points, spending score, and summary were calculated. I would also find out how the data was collected, ensuring its accurate, and that it was collected ethically, with privacy in mind, as sensitive information like remuneration is included in the dataset.

Analytical Approach

Before working with the data, I familiarized myself with the metadata.

I began my analysis using python. After importing all the necessary Python dictionaries, I imported the 'turtle_reviews.csv', and began cleaning data by removing columns that were redundant for my analysis. Once I was content with the data structure, to help Turtle Games understand loyalty points, I proceeded to conduct three simple linear regressions to understand relationships between Loyalty Points, and Age, Remuneration and Spending Score. These three independent variables represent all possible regressions, providing an extensive analysis. For the simple linear regression, I split the data into train and test sets, such that the regression was not overfitted to the data.

As the marketing department wanted to make use of remuneration and spending scores, I proceeded to explore if we could divide up customers into different segments with these variables using clustering techniques. To determine the optimal number of clusters I used both the elbow and silhouette techniques, which both suggested 5. Utilizing K-means clustering I grouped customers into 5 useful groups for marketing to use, and made a visualization demonstrating this.

To help inform future marketing campaigns, I also conducted a sentiment analysis on customer reviews and summaries. Using natural language processing I set out on finding the most popular words used in reviews and summaries, and found top 20 positive and negative reviews received. To do this I changed the reviews to not include stop words and punctuation to make the results of the analysis more readable and helpful. I also created boxplots showing the general sentiment of reviews.

I conducted the rest of the analysis on R, at request of Turtle Games and to take advantage of R's powerful data analytic tools. After importing, exploring (through looking at summaries, min, max and, mean) and cleaning the 'turtles_sales.csv' dataframe, I began

creating plots as part of my exploratory data analysis. At request of Turtle Sales, I paid specific attention to platform and NA and EU sales, and grouped sales by platform to create some bar charts.

During my exploration of the data I realized that some products had data on several rows, split by platform. As I wanted to know the sales of each game regardless of platform in different regions, I grouped the data by 'product_id'. Next, I checked the normality of the sales data set using Q-Q plots, the Shapiro-Wilk Test, and determined the Skewness and Kurtosis of the sales data.

Finally, the sales department wants to understand if there was a relationship between North American, European and Global sales. From the metadata we know that NA Sales and EU sales feed into Global Sales. I ran a simple linear regression to see if there was a relation between sales in the NA and EU region. As this revealed there was not collinearity between the two variables, I ran a multiple linear regression where NA sales and EU sales were independent variables, and global sales was the dependent variable. The model created was promising and I used it to predict global sales based on provided NA and EU sale values.

Visualizations, Insights, Patterns and Predictions

From my analysis, I created visualizations to highlight the insights, patterns and predictions I made. For the simple linear regressions, I created three scatter plots showing relations between loyalty points and the independent variables. With these graphs, I also plotted the line of best fit from the regression to help others understand how strong the linear relationships are. These graphs clearly show the weak linear relation loyalty points has with remuneration and spending score, and lack of a relation with age.

When conducting the k-clustering, I noted that a scatter plot would clearly highlight the five different groups created. As a result, I created a scatter plot showing remuneration and spending score where each group had a different colored point on the plot to help viewer differentiate.

The visuals for the sentiment analysis presented challenging choices. While a word cloud was created for the most common words, I felt the graphic was too busy and hard to interpret easily. Hence, I opted to ignore this word cloud, and created a simple table showing the most common words. Similarly, the most negative and positive reviews were presented in simple textual format. To help Turtle games understand the general sentiment of their reviews I did create a simple boxplot which shows that in general reviews are quite positive.

To show how sales from different platforms did on the NA and EU regions, I created two bar graphs. The choice and design of bar graphs makes it easy for viewers to compare how platforms did in the regions. A quick glance on the bar plot reveals that the PS3, X-Box 360 and PS3 are the most popular platforms, but the PS3 is proportionality a lot more popular in Europe compared to the USA.

I extracted data on the normality of the sales columns and simply put them into a textual table form for use during my presentation. All three sales data was not normally distributed

with a strong positive skew. The heads were also heavy, showing a leptokurtic tail. To show the lack of linear relation between NA and EU sales I created a similar scatter plot to the loyalty point one, with a line of best fit. I lastly presented my predictions for global sales in a table. This format of presentation is simple, and get the point out easily. When presenting my predictions I do put emphasis that the model has a great adjusted r-squared of 0.9664.