Report

Turtle Games is a game manufacturer and retailer that has hired me as a data analyst to analyse their data sets to help with their business development. The main business objective is to improve overall sales performance by determine the new pricing for their Lego products, while also identifying current customer sentiments on their games, and also predicting sales for video games for next year. This report aims to go through my procedure as I apply exploratory, descriptive and predictive analysis on three data sets from the business and in the end point out my suggestions and predictions from the insights I gathered.

I started off with importing the necessary libraries, and accessing the Lego data set by importing it into Python and then cleaned and wrangled it to prepare for my analysis, which is applying predictive models, namely simple and multi-linear regression models with Python. After reviewing the information of the data set, I checked for missing values and removed duplication, then I created simple visualisations for exploratory analysis; to find out more about the main columns I will be using, 'list_price', 'piece_count' and 'ages'. Then I split the data into training and testing subsets with 70% and 30% ratio for both my linear regression models. I also used the Lego data set for my exploratory analysis on customer behaviours with R, of which I started off by importing 'tidyverse' library and the data file it into R and then start my process cleaning and wrangling it during preparation. I decided to remove values equal to zero from 'ages' column, as the data there represents target customer age and with infantry toys, the reviews will be submitted by the parents and therefore I do not know their age. I then created basic visualisation for exploratory analysis with objective of finding trends in customer reviews by their age and also purchasing behaviours. For the games review data set, I mainly used it for sentiment analysis by applying natural language processing model. I started off with cleaning and manipulating the data set to decide which columns to use and which to ignore. I decided to focus on three columns, 'reviewText', 'reviewerName' and 'summary'. I removed missing values from the data set and then transform all the data from 'reviewText', which is the main column I will apply sentiment analysis on, into lowercase, and then remove all punctuation, remove duplication, apply tokenisation to split each word as one token and eliminate stop words for more precise analysis. The last data set I will be working with is games sales, of which I will only use R for my analysis. First, I import necessary libraries for wrangling the data and visualisation, then I imported the data file to create the data set. I quickly checked for information on the summary of the data, finding out the required columns with the data I want to focus on. My next step is to clean and manipulate the data set for preparation for

visualisation. I determined missing values, change values under 'Genre' column to lowercase and created a new column by merging values under 'Genre' and 'Platform'. Also, because of the unit of count for the numeric values are under millions, I converted them to thousands to make it easier for visualisation, and also removed outliers above 3000 ($1000s) and any null values. My next step for my analysis is to visualise and evaluate the skewness of the data, which I will discuss further later during my explanation of insights. The last stage of my analysis is to apply prediction model with R to the data set for prediction of global sales of the video games. I checked the statistic summary of the data and created a temporary new data set only containing numeric values, then I checked for correlations between all the variables. Next, I applied multi-linear regression model to the data to determine optimal prediction for sales.

Now I will discuss the visualisations I created and all the outcomes and insights I gathered from my exploratory, descriptive and prescriptive analysis. First, I will discuss my regression model for the Lego data set. From the initial exploratory visualisations, I discovered that price trend positively to the number of pieces of the product, and products with highest list price are targeted at infants, young children and young adults. Both my models have not achieved the best R-squared values and mean absolute error values, with averaging 75% and 20 respectively, meaning the models are not perfect but I believe they are acceptable to use for analysis. I used them to predict sale price for product with 8,000 pieces resulting around $826 and predict sale price for product with 8,000 pieces that are most likely purchased by 30-year-old customers resulting around $825. Next, when using R to create basic visualisations, I created bar chart to find out which age group of customers submitted the most reviews, and found out target customers age 19 submitted the most reviews, and the overall curve suggests most reviews are left by teenagers age group and young adults. Then I created a scatterplot to find the most expensive product purchased by customers above target age 25 years old, and found out it is with target age 29 years old, with the price at $260, and in fact, this age group contain all of the products with price above $200. My next stage is to apply sentiment analysis using a natural language processing model I created to determine whether the reviews from customers are more positive or negative towards the products. After preparing the data, I created a word cloud to visualise the tokens extracted from the data, by first glance all words with most emphasise appeared to me positive but it required further analysis. My next step is to generate a polarity score to determine the positiveness of all reviews. I created a histogram to visualise the polarity score results, majority of the score is between 0 and 0.25, and it showed that most sentiments are more positive than negative as the skewness is negative. Lastly, I

extracted 20 top positive and negative reviews for a further analysis to check if it matches with the results from the polarity score. I can draw that from positive reviews, customers are happy with the condition of the product, matching their requirements, perfect to use as a gift, best game, great addition, and overall performance meeting their expectations. While from negative reviews, most of the negative reviews are based on the service of the business over the quality of the product. The last part of my analysis is creating visualisations on sales of games of the business and apply prediction model to predict sales. I used R for creating the visualisations to evaluate the skewness of the data, I decided to focus on three columns that contain sales from Europe, North America and globally. I created a histogram and boxplot respectively for all three variables and found out all have positive skewness. The visualisations still showed that are many outliers, but majority of values range between 0 and 1000 (in $1000s), and by comparing it with the original data set, a conclusion can be drawn that while there are games that sells great, over half of the games still performed poorly while in comparison. I then created two scatterplots aiming to find the correlation between sales in North America and Europe, and ultimately see how it can help me to predict global sales. I also added the genre of games to further break down the values for clearer views, and in conclusion both variables have a positive correlation with global sales, meaning that if the business can keep improving sales in those two regions, it is highly likely it will increase global sales. Next, I applied multi-linear regression model to predict the global sales for all games following my previous analysis, from the statistics summary I can tell that the model has a 96.5% R-squared value and also very high t-values for both North America and Europe sales variables, meaning the model will be highly accurate for prediction using two high relative variables for global sales. Lastly, I added the predictions to the original data set for convenience for future analysis.

In conclusion, to draw up my analysis, I would like to make recommendations to the business for its objectives. First, Turtle Games should re-shape the pricing structure of its Lego products, preferably with more information provided for a more accurate prediction model. Secondly, the business should put in more effort in customer service and aftercare services to improve satisfaction of their customers, this will lead to a more positive sentiment towards the products and the business overall. Lastly, the business should reconsider what types of video games to stock; this is be break down to only focusing on certain gaming platforms, the latest released games or games of a particular genre. This will help improving sales performance overall and also save on expenses.