Hazel Chan

July 3, 2022

LSE Course 3: Predictive Analytics

Assignment 3 Report

Table of Contents

[Background 2](#_Toc107780470)

[Approach & Visualisation 2](#_Toc107780471)

[A3\_Week\_1.ipynb 2](#_Toc107780472)

[A3\_Week\_3.ipynb 7](#_Toc107780473)

[A3\_Week\_4.R 12](#_Toc107780474)

[A3\_Week\_5.R script 14](#_Toc107780475)

[A3\_Week\_6.R script 16](#_Toc107780476)

[Predictions 18](#_Toc107780477)

# Background

Turtle Games is a game manufacturer and retailer that sells their own products and products manufactured by other companies. They offer three types of product range: Lego, various toys and games, and video games for their global customers. As a global company, they have an objective of improving overall sales performance based on data analysis of price, customer sentiment and forecast of global sales. This report will detail the methodology used to conduct analysis, insights and predictions of sales.

# Approach & Visualisation

There are three aspects to analyse the dataset and improve sales performance. Firstly, simple and multiple linear regression functions in python are used to build a pricing model for lego products based on pieces and customer age. R tidyverse package is used to analyse the age group who is most likely to leave reviews and highest price point age group of 25 or above are willing to purchase. Secondly, Natural Language Processing is used to conduct sentiment analysis to understand feedback from customers who purchased various toys and games. Thirdly, multiple linear regression model in R is used to predict total global sales of all video games for next financial year based on Europe and North America sales.

## A3\_Week\_1.ipynb

To predict the optimal price for lego products with 8000 pieces, the lego pricing trend needs to be identified before creating simple and multiple linear regression models for prediction.

Using describe function in python, the price ranges from USD 2.27 to max USD1104 with a mean of USD65.

Table

Description automatically generated

Using seaborn package to create histogram and boxplot, the price distribution is strongly skewed to the right with a long tail on the positive side. Majority of the lego is priced between USD 20 to 70, while outliers represent expensive products ranging from USD 180 to USD 1100.

Chart

Description automatically generated with medium confidence

Chart, calendar

Description automatically generated

Graphical user interface, text, application

Description automatically generatedA subset with price as y independent variable and pieces as x dependent variable is created to build simple linear regression model. The subset is then split into train (70%) and test (30%) sets to validate accuracy of the model.

Train data set is fitted with simple linear regression model.

Graphical user interface

Description automatically generated with low confidence

A scatter plot is created to visualize relationship between price and pieces of train set.

There is a positive relationship with a strong R-squared value of 0.76. This means the increase in pieces explains 76% variation, or increase, of price.

Chart, scatter chart

Description automatically generated

Graphical user interface, text, application, email, Teams

Description automatically generatedWith the intercept and coefficient value, the predict function is used to predict price of 8000 pieces lego. According to the training model, the optimal price is USD797.

Chart, scatter chart

Description automatically generatedTo validate the model, the test set is used to predict the price, which resulted in USD 783 which is similar to USD 797.

Multiple linear regression model is built to understand whether age of 30 variable affects the price of 8000 pieces lego product. As subset is created to include three variables, price, pieces and age.

Graphical user interface, text, application, Teams

Description automatically generated

Model is built based on the train with a strong R-squared meaning pieces and age explains 76% of price variation. The predicted price for 8000 pieces of lego purchased by 30 year old is USD783.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text

Description automatically generated with medium confidence

The test set is fitted into the model predicting optimal price at USD 814 which is similar to that of the train set.

Graphical user interface, text, application, email

Description automatically generated

## A3\_Week\_3.ipynb

Natural language processing is applied to analyse customer reviews of various toys and games and identify areas for improvement to satisfy customer and maximise sales.

Using python, game\_reviews dataset is imported and viewed. A subset with only full review text is created, since summary is too short and cleaned by removing missing values, convert all text to lowercase, remove punctuation marks and duplicates. See notebook cell 1 to 18. Graphical user interface, text, application

Description automatically generated

Chart, bar chart

Description automatically generatedText

Description automatically generated with low confidenceNext, text is pre-processed using tokenisation to calculate the frequency distribution of words. Each sentence is split into individual tokens, added into one list using for loop and plotted. As there are many stopwords such as “and”, “the” affecting results, they are removed then, visualise the clean tokens with WordCloud. Words with higher frequency are larger such as “game”, “play”, “great”.

Polarity score is generated using textblob library and plotted onto histogram and boxplot. Overall, majority of reviews are slightly positive with polarity score mainly between 0 to 0.25.

Chart, histogram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Graphical user interface, text, application, email

Description automatically generatedTo understand reason behind positive sentiment, the top 20 positive reviews are extracted. Using document-term matrix to extract the positive features, most top reviewers thought that games were “awesome”, “perfect” and “wonderful”.

Chart

Description automatically generated

Graphical user interface, text, application

Description automatically generatedTop 20 negative reviews are extracted based on negative polarity score and most reviewers were “disappointed” and mentioned “box”, “book”, “elf”.

Chart, bar chart

Description automatically generated

## A3\_Week\_4.R

R Tidyverse package is used to wrangle and visualise lego data set to understand which age group is most likely to leave reviews and which most expensive price point is purchased by customer above age 25.

The lego data set is imported, viewed using as\_tibble function, and check for missing values. Then, qplot function is used to view distribution of age variable using boxplot. Majority of customers are aged between 10 to 25.

Chart, histogram

Description automatically generated

Column graph identified that age 8 customer left the most reviews.

Chart, histogram

Description automatically generated

A new column is added to age\_df to map the corresponding customer age into 6 age groups and aggregate by age\_group.

Text

Description automatically generated with medium confidence

Column graph identifies customers age 6-10 are most likely to leave reviews, then followed by age 26-30.

Chart, bar chart, histogram

Description automatically generated

Subset with customer older than age 25 is created to identify highest price point accepted. Dataframe is aggregated by age and summarised by maximum price. Column graph indicates age 29 customers purchased the most expensive lego product at USD260, then age 25 customers at USD195.

Chart, bar chart, histogram

Description automatically generated

## A3\_Week\_5.R

Game\_sales dataframe with relevant variables is created and cleaned by identifying missing values and convert genre values to lowercase.

Global sales, North America and Europe sales distribution is visualised using histogram and boxplot of ggplot. All sales distribution is extremely skewed to the right due to extreme outliers. Outliers are included as they represent the best selling products.

Majority of global sales is between 0.06M to 1M units with outlier of 82M, high skewness of 17.34 and heavy kurtosis of 606.75.

Text

Description automatically generatedGraphical user interface, text, application

Description automatically generated

A picture containing text, white

Description automatically generatedChart

Description automatically generated

Majority of NA sales is between 0.01M to 0.6M units with outlier of 82M, high skewness of 18.79 and heavy kurtosis of 651.

Graphical user interface, text, application

Description automatically generatedText, letter

Description automatically generated

Chart, bar chart, histogram

Description automatically generatedA picture containing text, white

Description automatically generated

Majority of EU sales is between 0.01M to 0.27M units with outlier of 29M, high skewness of 18.87 and heavy kurtosis of 758.80.

Text

Description automatically generatedText

Description automatically generated

Chart, histogram

Description automatically generatedA picture containing text, white

Description automatically generated

A picture containing graphical user interface

Description automatically generatedScatterplot shows a strong positive relationship between North America, Europe and Global sales with a correlation coefficient of 0.9 which is very close to 1.

Text

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

## A3\_Week\_6.R

Multiple linear regression model is used to predict global sales based on North America and Europe sales to minimise production waste.

The model is built using the linear regression function where Global\_Sales is inputted as dependent variable and EU\_Sales & NA\_Sales are inputted as independent variables.

Graphical user interface, text, application, email

Description automatically generated

Multiple R-squared 0.96 is very strong and close to 1 meaning North America and Europe sales explains 96% of the variability of Global sales. Three stars next to the variables indicates that they have high significance in the model.

Text

Description automatically generated

Total dataframe is created by aggregating sum of global, NA and EU sales of all games. Predict function generates an outcome of 8340M which is close to actual value of 8920M.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

Description automatically generated

# Predictions

Lego price should increase as pieces and age increase. By identifying optimal price points will enable business to align price with customers perceived value of products, increase change of purchase and maximise sales. Business should price lego with 8000 pieces at USD783 and if customer is age 30, price at USD 814.

Business should also improve product features based on customer reviews to satisfy customer demands with attractive product offerings. Overall review sentiment is slightly positive with polarity score between 0 to 0.25. Based on top negative review analysis, business should improve box packaging, elf image and books to improve polarity score closer to 1.

Customers age between 6-10 left the most 76,120 reviews which may be target customers of lego. Age 29 customers purchased the most expensive lego product at USD260, then age 25 customers at USD195 signifying higher purchasing power. Business should target expensive product range for older customers.

Lastly, based on EU and NA historic sales, global sales of all video games are predicted to reach 8340M units in the next financial year. Business should take forecasted sales into consideration buying and manufacturing units to ensure enough inventory for customer demand.