

LSE Data Analytics Online Career Accelerator  
DA301: Advanced Analytics for Organisational Impact  
Assignment 3  
Megan Bilas  
4 July 2022

## Background/context of the business

Turtle Games is a games manufacturer and retailer. To improve overall global sales performance, this analysis provides recommendations on:

- The optimal price Lego products should be sold for based on the number of Lego pieces in the set and the age of the customer that the product is most likely to be purchased by
- The customer group that will most likely leave a review on the products they have purchased
- The most expensive product purchased by a particular group of customers
- The general sentiment of customers across all e-store products
- The predicted global sales for the next financial year

## Analytical approach

The analytical approach to addressing the key business objectives is summarised in Table 1.

### Pre-processing

All data files were imported into the respective Python and R programs as csv files and libraries needed for the analyses were loaded (Table 1). Following this step, data files were explored to determine the number of observations, columns available, data types present, and summary statistics. The data files were also assessed for any missing values. The specific steps used to prepare and run each analysis are grouped by business objectives below.

### Determine the optimal price at which they should sell Lego products based on the number of Lego pieces in the Lego set and the age of the customer that the product is most likely to be purchased by

The relationship between `piece_count` and `list_price` was examined through simple linear regression. The dependent variable was defined as `list_price` and the independent variable was defined as `piece_count`. The data was split into train and test data sets. Linear regression was run on both subsets. The R-squared, intercept, and coefficient values of the train and test data were compared, alongside scatterplot visualisations (Figures 4-7). All values and visualisations closely matched, signifying the strength of the model.

**Figure 4. Linear regression R-squared values**

Train Subset (n=8,583)	Test Subset (n=3, 678)
0.764	0.736

*R-squared indicates the proportion of variance in the dependent variable that can be explained by the independent variable*

**Figure 5. Linear regression intercept values**

Train Subset (n=8,583)	Test Subset (n=3, 678)
17.022	18.043

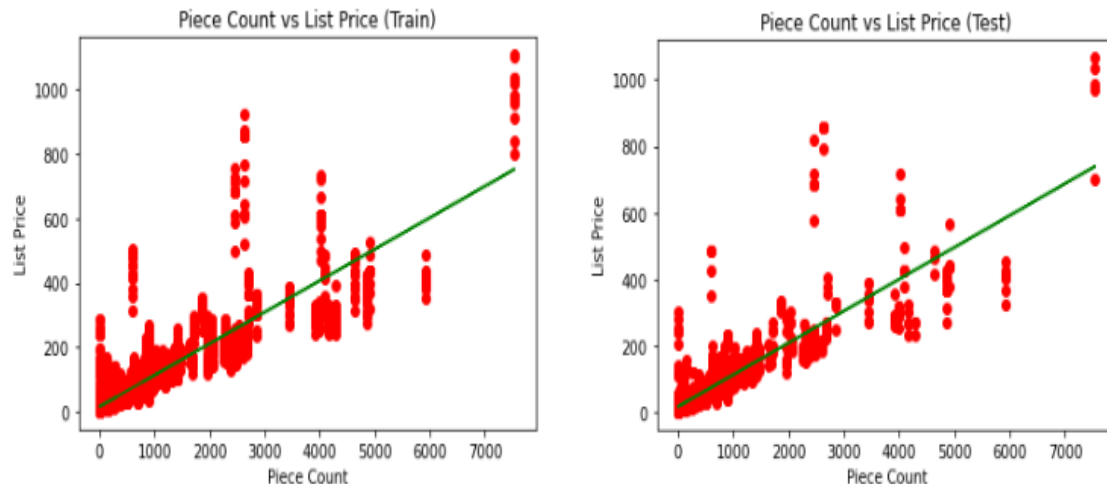
*Intercept indicates where the function crosses the y-axis*

**Figure 6. Linear regression coefficient values**

Train Subset (n=8,583)		Test Subset (n=3, 678)	
0.097		0.096	

*The coefficient value signifies the among the dependent variable changes for a unit increase in the independent variable*

**Figure 7. Linear regression scatterplot comparisons**



The relationship between piece\_count, ages, and list\_price was examined through multiple linear regression (MLR). The dependent variable was defined as list\_price and the independent variables were defined as piece\_count and ages. Data was split into train and test data sets. The R-squared, intercept, and coefficient values of the train and test data were compared (Figures 8-10). All values closely matched, signifying the strength of the model.

**Figure 8. MLR R-squared values**

Train Subset (n=8,583)		Test Subset (n=3, 678)	
0.764		0.736	

*R-squared indicates the proportion of variance in the dependent variable that can be explained by the independent variables*

**Figure 9. MLR intercept values**

Train Subset (n=8,583)		Test Subset (n=3, 678)	
16.840		16.840	

*Intercept indicates where the function crosses the y-axis*

**Figure 10. MLR coefficient values**

Train Subset (n=8,583)		Test Subset (n=3, 678)	
piece_count	ages	piece_count	ages
0.097	0.011	0.097	0.011

*The coefficient value signifies the among the dependent variable changes for a unit increase in the independent variables*

Multicollinearity and heteroscedasticity were also assessed but were not present, thus providing greater confidence in the model.

Determine the customer group that will most likely leave a review on the products they have purchased

Following cleaning steps to remove outliers (see Table 1), the data was aggregated by age using dplyr and purrr. This allowed the number of reviews to be plotted in ggplot by age.

Determine the most expensive products purchased by a particular group of customers

Data was filtered to customers aged 25 and over. 'List\_price' was plotted on ggplot to identify the price of the most expensive product.

Determine the general sentiment of customers across all e-store products

The data for 'games\_reviews.csv' was cleaned according to the steps outlined in Table 1 to enable it to be run through the nltk packages for natural language processing. Polarity scores were generated to identify the top 20 positive and negative reviews. The distribution of the sentiment was visualised in kde and boxplots.

Determine the predicted global sales for the next financial year

The data for 'games\_sales' was assessed to determine whether it was normally distributed through 'qqnorm' plots (Figures 11-13) and evaluation of skewness values (Figure 14). Because it was outside the scope of this study, the data was not normalised. However, it is highly recommended that this be done in the future before running further statistical analyses.

**Figure 11. Qqnorm plot- Global sales**

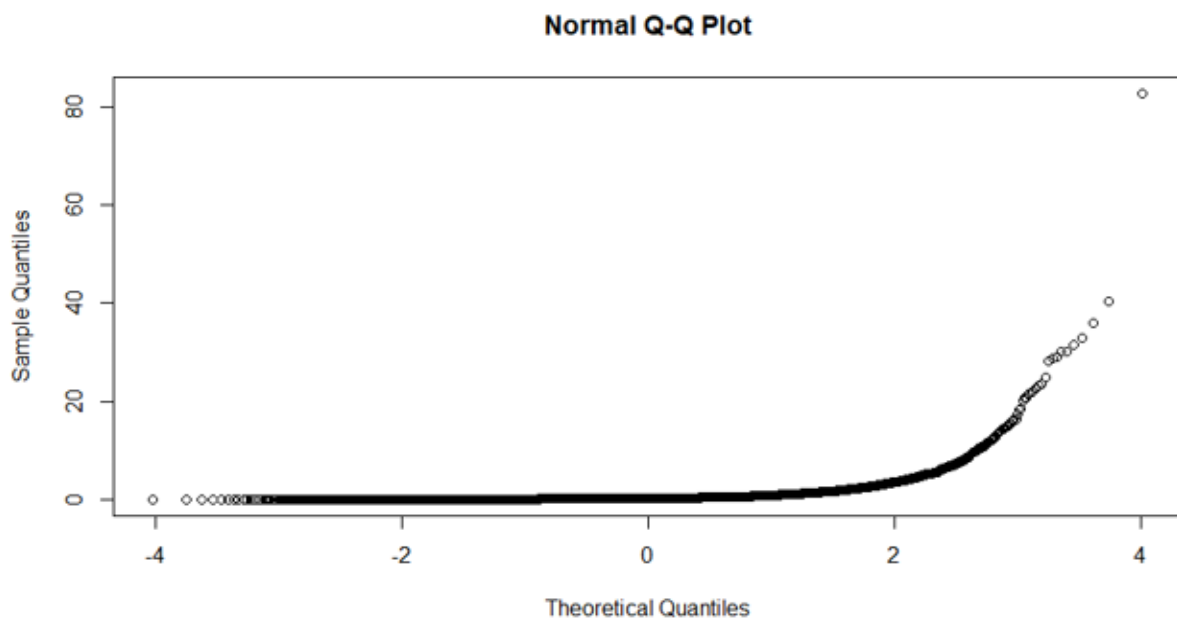


Figure 12. QQnorm plot – North American sales

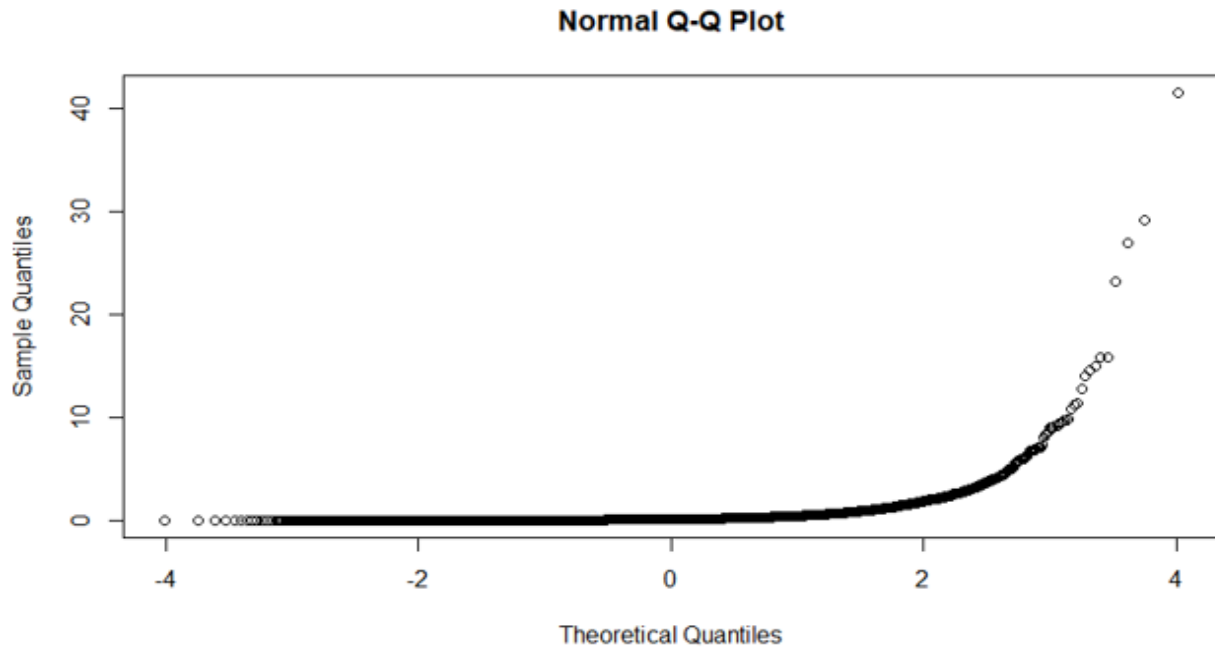
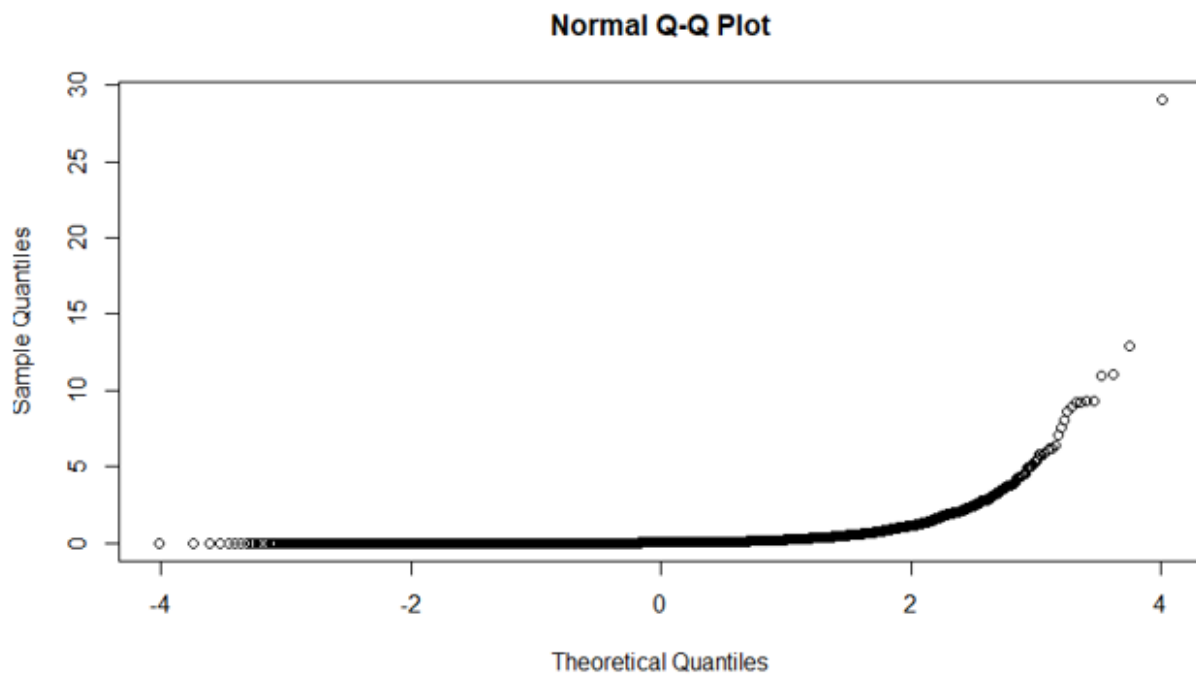


Figure 13. QQnorm plot – European sales



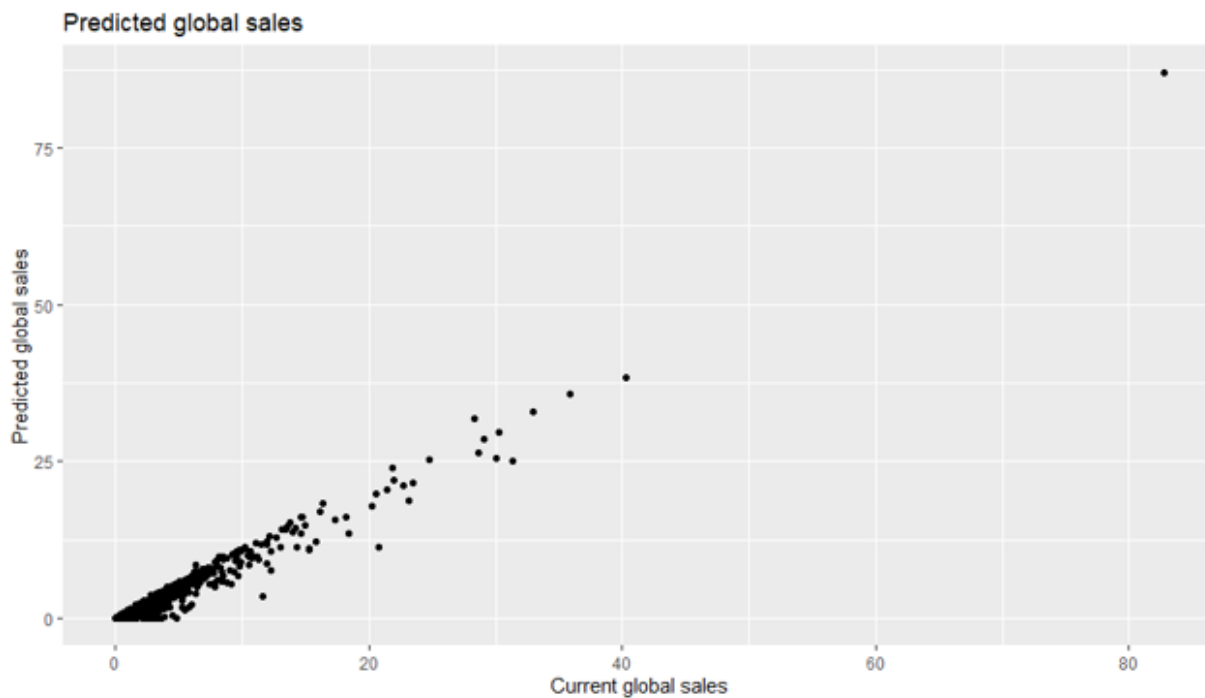
**Figure 14. Skewness values**

Global sales	North American sales	European sales
17.399	18.798	18.874

*These values indicate a very high positive skew*

MLR models were created to predict global sales based on Europe and North American sales. The model using both Europe and North American sales to predict global sales was used based on its high R-squared value. A linear relationship was found between current global sales and predicted global sales (Figure 16).

**Figure 16. Relationship between current and predicted global sales**



**Table 1. Analytical approach summary**

Data set used	Business objective	Business question	Statistical software	Libraries	Method	Additional checks/cleaning
lego	Determine the optimal price at which Turtle Games should sell Lego products based on the number of Lego pieces in the Lego set and the age of the customer that the product is most likely to be purchased by	What price should be set for the Lego sets that have 8,000 Lego pieces?	Python	Numpy: used to statistical calculations Pandas: used for data manipulation and analysis Statsmodels: used to get OLS to fit a regression line and to calculate variance inflation factor Sklearn: used to calculate simple and multiple linear regression Matplotlib: used for plotting Seaborn: used for plotting	Simple linear regression: models the relationship between two continuous variables	
lego	Determine the optimal price at which Turtle Games should sell Lego products based on the number of Lego pieces in the Lego set and the age of the customer that the product is most likely to be purchased by	What price should be set for all the Lego sets that have 8,000 Lego pieces and are most likely to be purchased by customers who are 30 years old?	Python	Numpy: used to statistical calculations Pandas: used for data manipulation and analysis Statsmodels: used to get OLS to fit a regression line and to calculate variance inflation factor Sklearn: used to calculate simple and multiple linear regression Matplotlib: used for plotting Seaborn: used for plotting	Multiple linear regression: uses two or more independent variables to predict the outcome of a dependent variable	Data split into train and test subsets to avoid overfitting: when a statistical model fits so closely with its training data that the algorithm cannot make accurate predictions on unseen data. Split 70% train; 30% test – rationale: we want the training data set to be as large as possible while retaining enough observations to have a reasonable sized test data set

Data set used	Business objective	Business question	Statistical software	Libraries	Method	Additional checks/cleaning
						<p>Multicollinearity: when there are strong correlations between two or more independent variables; assessed by calculating the variance inflation factor (VIF) of the independent variables</p> <p>Heteroscedasticity: when there are uneven variances in the residuals, which can result in biased and skewed results; assessed by running Breusch-Pagan test</p>
lego	Determine the customer group that will most likely leave a review on the products they have purchased	Which age group submits the most reviews?	R	Tidyverse: packages used to read-in, explore, transform, and visualise data	Data aggregation	A qplot boxplot revealed two outliers above 300, which were removed by filtering the data
lego	Determine the most expensive products purchased by a particular group of customers	What is the cost of the most expensive Lego set purchased by customers who are at least 25 years old?	R	Tidyverse: packages used to read-in, explore, transform, and visualise data	Data sub-setting by filtering for those aged 25 and over	
games_reviews	Determine the general sentiment of customers across all e-store products	What is the general sentiment of customers across all products?	Python	Numpy: used to statistical calculations Pandas: used for data manipulation and analysis	Natural language processing	<p>Missing values removed for reviewText and data extracted from this column</p> <p>Sentences converted to lowercase, punctuation</p>



Data set used	Business objective	Business question	Statistical software	Libraries	Method	Additional checks/cleaning
				NLTK: used for symbolic and natural language processing in English		removed, duplicates dropped, and index reset  Text converted into tokens and stopwords removed
games_reviews	Determine the general sentiment of customers across all e-store products	Based on the popularity of the sentiment, what are the top 20 positive and top 20 negative reviews?	Python	Numpy: used to statistical calculations Pandas: used for data manipulation and analysis NLTK: used for symbolic and natural language processing in English	Natural language processing	Missing values removed for reviewText and data extracted from this column  Sentences converted to lowercase, punctuation removed, duplicates dropped, and index reset  Text converted into tokens and stopwords removed  Data sorted by descending positive and negative scores
games_sales	Determine the predicted global sales for the next financial year	What are the predicted global sales for the next financial year based on current sales from Europe and North America?	R	Tidyverse: packages used to read-in, explore, transform, and visualise data	Multiple linear regression: uses two or more independent variables to predict the outcome of a dependent variable	R-squared values compared for different models to choose model with highest R-squared value  Relationship between chosen model and current global sales plotted to examine linearity

## Visualisation and insights

Insights are grouped by key business questions below.

What price should be set for the Lego sets that have 8,000 Lego pieces?

Turtle Games should charge \$782.94.

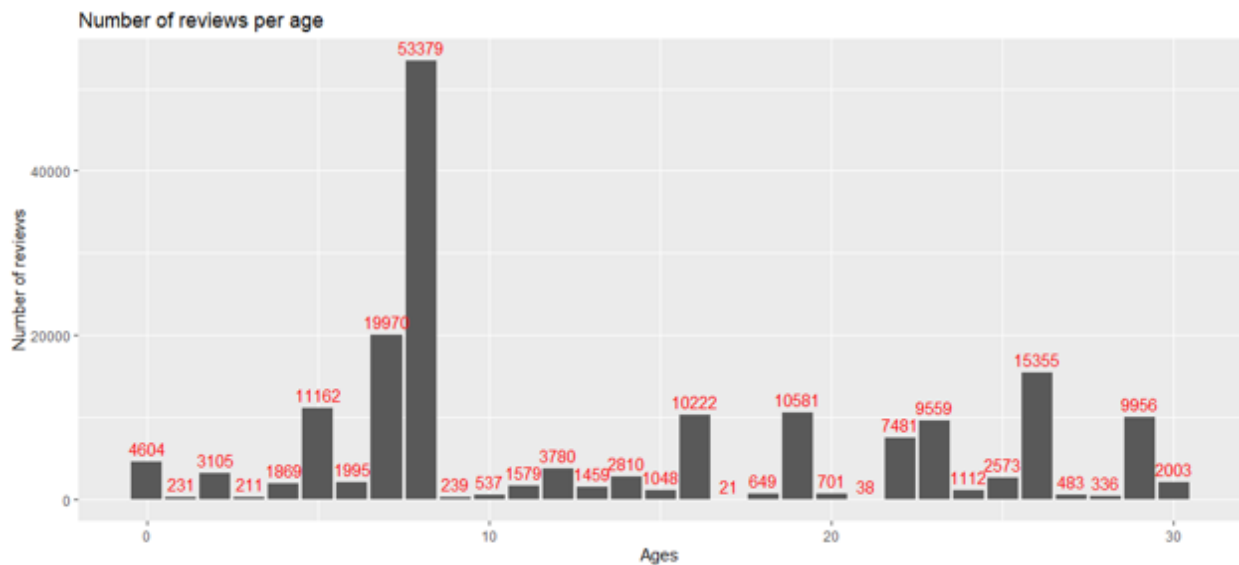
What price should be set for all the Lego sets that have 8,000 Lego pieces and are most likely to be purchased by customers who are 30 years old?

Turtle Games should charge \$796.79.

Which age group submits the most reviews?

The most reviews are submitted for those aged 8, with 53,379 reviews (Figure 17).

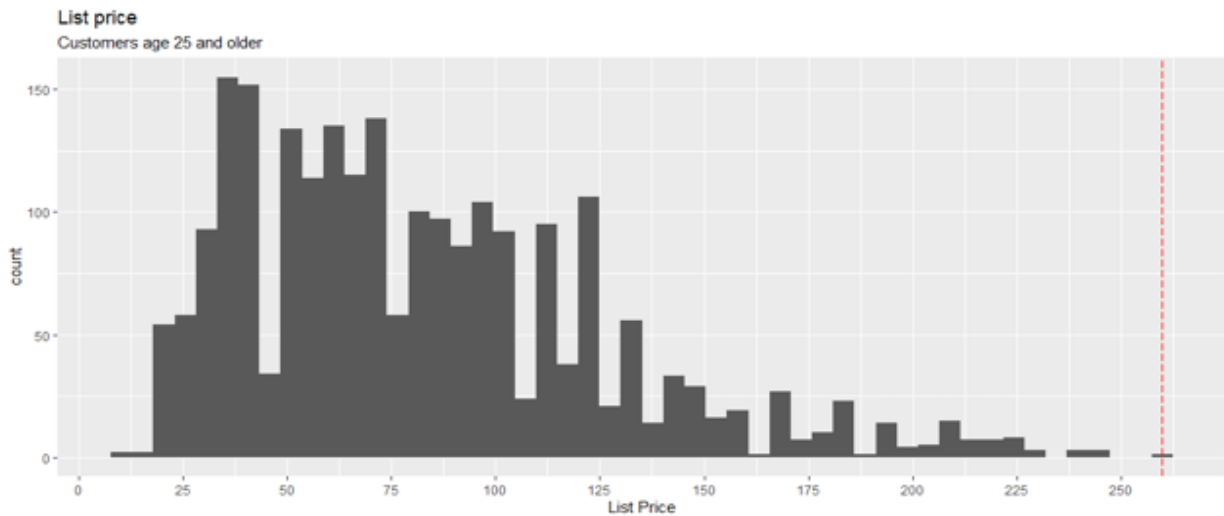
**Figure 17. Reviews by age**



What is the cost of the most expensive Lego set purchased by customers who are at least 25 years old?

The cost of the most expensive Lego set purchased by customers who are at least 25 years old is \$259.87 (Figure 18).

**Figure 18. List price – subset: those aged 25 and over**

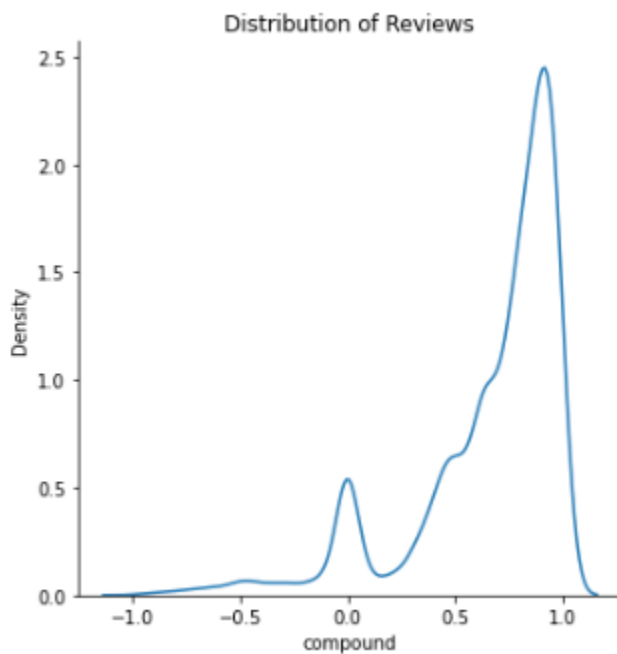


*The data has already been filtered to those aged 25 and older, so there are limited visualisations that can be used to show the single variable of interest (`list_price`). The red line indicates the most expensive product, at the end of the distribution.*

What is the general sentiment of customers across all products?

The general sentiment of customers is positive, as indicated by a high peak in Figure 19 that is close to 1. There is also a smaller peak of neutral reviews (0.0) and fewer negative reviews (represented by negative values).

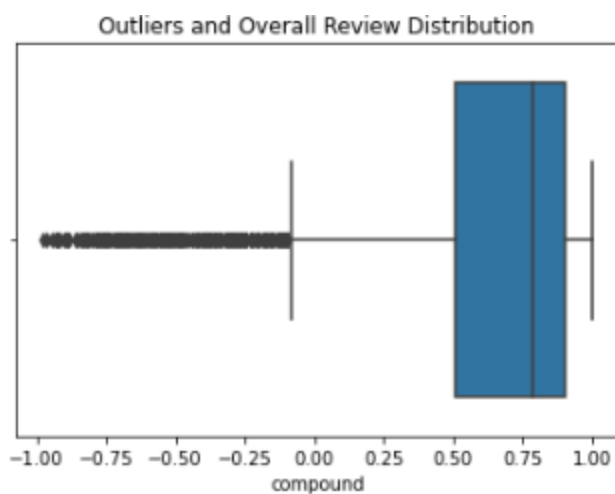
**Figure 19. Distribution of reviews – KDE plot**



*Compound score generated from subtracting negative scores from positive scores. Negative compound sentiments are generally negative, positive compound sentiments are generally positive, and neutral compound sentiments are closer to 0.*

There also many negative outliers (Figure 20). If these outliers were removed, it would push the distribution even closer to having more positive reviews.

**Figure 20. Distribution of reviews - boxplot**



*Compound score generated from subtracting negative scores from positive scores. Negative compound sentiments are generally negative, positive compound sentiments are generally positive, and neutral compound sentiments are closer to 0.*

Based on the popularity of the sentiment, what are the top 20 positive and top 20 negative reviews?

The top 20 positive reviews (Figure 21) reveal that customers are happy with Turtle Games' products as gifts and that they are perceived as being fun. The top 20 negative reviews (Figure 22) reveal that customers are dissatisfied with certain features of the products.

**Figure 21. Top 20 positive reviews**

	neg	neu	pos	compound
adorable	0.0	0.0	1.0	0.4939
cute love	0.0	0.0	1.0	0.8020
good value	0.0	0.0	1.0	0.6486
fine	0.0	0.0	1.0	0.2023
best	0.0	0.0	1.0	0.6369
fascinating	0.0	0.0	1.0	0.5423
entertaining	0.0	0.0	1.0	0.4404
nice gift	0.0	0.0	1.0	0.6908
super great love	0.0	0.0	1.0	0.9217
awesome better	0.0	0.0	1.0	0.7906
awesome great fan	0.0	0.0	1.0	0.8885
fun fun fun	0.0	0.0	1.0	0.8720
wonderful	0.0	0.0	1.0	0.5719
adorable gift	0.0	0.0	1.0	0.7269
perfect holiday	0.0	0.0	1.0	0.7506
gift	0.0	0.0	1.0	0.4404
great entertainment	0.0	0.0	1.0	0.7845
inspiring creativity	0.0	0.0	1.0	0.6597
amazing	0.0	0.0	1.0	0.5859
exciting	0.0	0.0	1.0	0.4939

**Figure 22. Top 20 negative reviews**

	neg	neu	pos	compound
difficult	1.000	0.000	0.0	-0.3612
crazy	1.000	0.000	0.0	-0.3400
limited	1.000	0.000	0.0	-0.2263
missing	1.000	0.000	0.0	-0.2960
disappointment	1.000	0.000	0.0	-0.5106
stupid	1.000	0.000	0.0	-0.5267
gross	1.000	0.000	0.0	-0.4767
scary	1.000	0.000	0.0	-0.4939
disgusting	1.000	0.000	0.0	-0.5267
missing horrible product	0.851	0.149	0.0	-0.6908
nothing great fair	0.840	0.160	0.0	-0.6435
found poor unclear	0.836	0.164	0.0	-0.6249
hate elf disgusting hate life leave shelf alone	0.824	0.176	0.0	-0.9186
waterproof dangerous	0.756	0.244	0.0	-0.4767
small disappointed	0.756	0.244	0.0	-0.4767
broke soon	0.737	0.263	0.0	-0.4215
button lame	0.737	0.263	0.0	-0.4215
wasnt fun	0.730	0.270	0.0	-0.4023
delivery many missing disappointed	0.726	0.274	0.0	-0.6486
glue messy	0.714	0.286	0.0	-0.3612

What are the predicted global sales for the next financial year based on current sales from Europe and North America?

The predicted global sales for the next year are shown for the for the first 20 ranked products in Figure 23. Due to the large product list (n=16,598), it is recommended that the total product list be viewed in the accompanying Rscript.

**Figure 23. Predicted global sales for the top 20 ranked products (USD)**

Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	Global Sales	Predicted Global Sales
1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	82.74	86.96
2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	40.24	38.31
3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	35.82	35.67
4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	33.00	33.02
5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	31.37	25.01
6	Tetris	GB	1989	Puzzle	Nintendo	23.20	2.26	30.26	29.76
7	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.38	9.23	30.01	25.59
8	Wii Play	Wii	2006	Misc	Nintendo	14.03	9.20	29.02	28.60
9	New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.59	7.06	28.62	26.35
10	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	28.31	31.85
11	Nintendogs	DS	2005	Simulation	Nintendo	9.07	11.00	24.76	25.33
12	Mario Kart DS	DS	2005	Racing	Nintendo	9.81	7.57	23.42	21.55
13	Pokemon Gold/Pokemon Silver	GB	1999	Role-Playing	Nintendo	9.00	6.18	23.10	18.74
14	Wii Fit	Wii	2007	Sports	Nintendo	8.94	8.03	22.72	21.17
15	Wii Fit Plus	Wii	2009	Sports	Nintendo	9.09	8.59	22.00	22.10
16	Kinect Adventures!	X360	2010	Misc	Microsoft Game Studios	14.97	4.94	21.82	23.92
17	Grand Theft Auto V	PS3	2013	Action	Take-Two Interactive	7.01	9.27	21.40	20.62
18	Grand Theft Auto: San Andreas	PS2	2004	Action	Take-Two Interactive	9.43	0.40	20.81	11.42
19	Super Mario World	SNES	1990	Platform	Nintendo	12.78	3.75	20.61	19.80
20	Brain Age: Train Your Brain in Minutes a Day	DS	2005	Misc	Nintendo	4.75	9.26	20.22	18.01

## Patterns and predictions

Insights are grouped by business objectives below.

Determine the optimal price at which they should sell Lego products based on the number of Lego pieces in the Lego set and the age of the customer that the product is most likely to be purchased by

The number of pieces in a Lego set is a stronger predictor of list price than age, although both have a linear relationship with list price. Generally, as the number of pieces increases and the age of the target customer increases, the price charged should also rise.

Determine the customer group that will most likely leave a review on the products they have purchased

The greatest number of reviews are left for those between the ages of 8-9. This could be an indicator of the age range customers are buying the most products for and, if so, could help Turtle Games in determining their target consumer profile.

Determine the most expensive products purchased by a particular group of customers

The cost of the most expensive product bought by customer aged 25 and over is \$259.87. However, more customers within this age bracket purchase products that are around \$35.00 and very few purchase products at the highest prices. Turtle Games may wish to examine the distributions in Figure 18 to determine the price point that generates the most sales overall.

Determine the general sentiment of customers across all e-store products

Customer sentiment tends to be high on e-store products. The top 20 positive reviews point toward products being ideal gifts and “fun.” This could help in future marketing campaigns. Meanwhile, the top 20 negative reviews tend to be associated with certain product characteristics, which could help in making product improvements.

Determine the predicted global sales for the next financial year

North American and European sales are strong indicators for global sales. Therefore, closely monitoring sales in these regions at regular intervals will help Turtle Games in predicting the performance of their overall global sales.

## Future actions

The data within the lego and games\_sales data sets is highly skewed. It is recommended that this data be normalised in order to increase the statistical strength of the analyses.

In addition, the accuracy of the insights could be improved by merging the data sets with product information. For example, knowing what products have positive or negative sentiments would help in marketing campaigns and product improvement initiatives.



## Appendix 1. Metadata

### lego.csv

Column	Sample value	Interpretation of columns
ages	19	The age the Lego product caters to
list_price	29.99	Price of the Lego product (in US dollars)
num_reviews	2	Number of reviews
piece_count	277	Number of Lego pieces in the product
play_star_rating	4	Star rating by players/customers (0 to 5 stars)
review_difficulty	0	The difficulty level of the product
country	20	Number of countries the product is sold in

### games\_reviews.csv

Column	Sample value	Interpretation of columns
overall	2	Overall ratings the customer has given the product (out of 5)
verified	FALSE	If the review is verified by Turtle Games, TRUE If the review is not verified by Turtle Games, FALSE
reviewTime	09 22, 2016	Date that the review was submitted in the format of MM DD, YYYY
reviewerID	A1IDMI31WEANAF	Unique ID of the reviewer
reviewerName	Mackenzie Kent	Name of the customer/reviewer
reviewText	When it comes to a DM's screen, the space on the screen itself is at an absolute premium. The fact that 50% of this space is...	Feedback/review submitted by customers who purchased and used the products
summary	The fact that 50% of this space is wasted on art (and not terribly	Summary of the feedback/review

	informative or needed art ...	
unixReviewTime	1474502400	Time taken to write the review (in seconds)  [Note: The time is unix time, which is a way of representing a timestamp as the number of seconds since January 1st, 1970 at 00:00:00 UTC.]
image	—	Image of the product

#### games\_sales.csv

Column	Sample value	Interpretation of columns
Rank	1	World ranking of the game
Name	Wii Sports	Name of the video game
Platform	Wii	Video game console on which the game was launched
Year	2006	Year of launch
Genre	Sports	Genre of the video game
Publisher	Nintendo	Company that published the game
NA_Sales	41.49	Number of games sold in North America (in millions of units)
EU_Sales	29.02	Number of games sold Europe (in millions of units)
Global_Sales	82.74	Total sales in the world (which is a sum of EU_Sales, NA_Sales and online sales) (in millions of units)