1. Background

Turtle games has a global customer base and have a business objective of improving overall sales performance. With this in mind , I have been instructed to extract insights from existing data that will enable Turtle games to meet this business objective.

2. Analytical approach

2.1. Steps used to Clean the data with R and Python

In order to answer the business questions this analysis utilised Python and R to clean, explore, manipulate and visualise the data depending on the particular business question.

The first step in any analysis is to import clean and sense check the data. The process followed in both R and Python to clean the data:

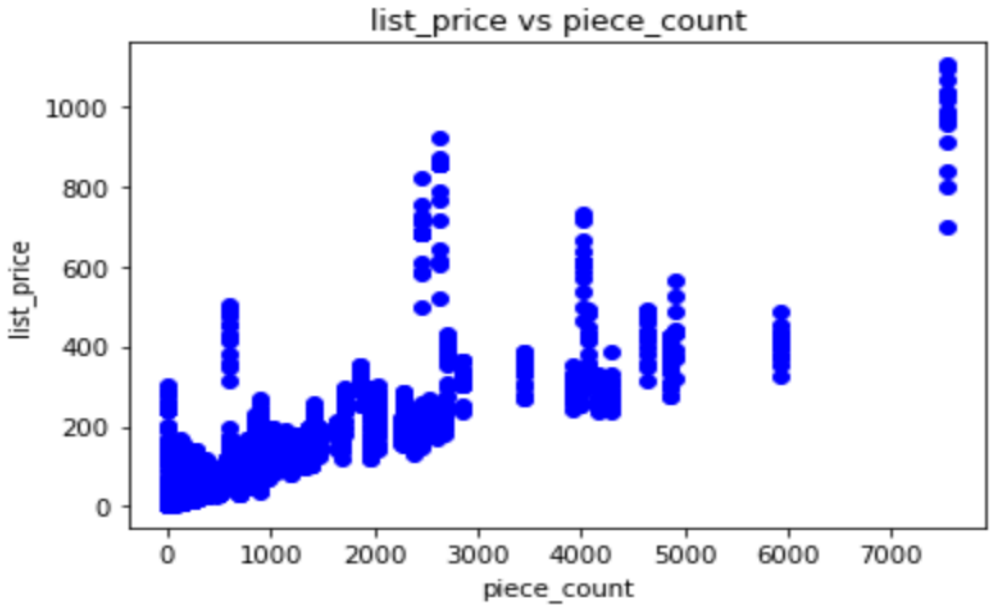
a. Viewing the data frame
b. Checking the dimensions of the data frame
c. Check for errors in data types
d. View descriptive statistics to see the spread, maximum values, minimum values and mean of the values in the dataset
e. Check for missing values and either remove or replace depending on the data and analysis needs
f. Remove unnecessary variables that do not add value
g. Clean string variables: change column values to lower case for uniformity, change column names to title etc.

3. Explore, analyse and visualise the data to answer the business questions

3.1. What price should be for an 8000 piece Lego set for an intended customer group of 30 year olds?

To answer this question we use the piece count variable to predict the list price using a simple linear regression model. To improve the accuracy of the prediction a multiple linear regression model is also fitted with the addition of age as a second predictor variable. During exploration of the data we ensure that the dependent and independent variables are linearly related as is seen in Figure 1 this was the case for the variables used in the model.

Figure 1. Relationship between the variables

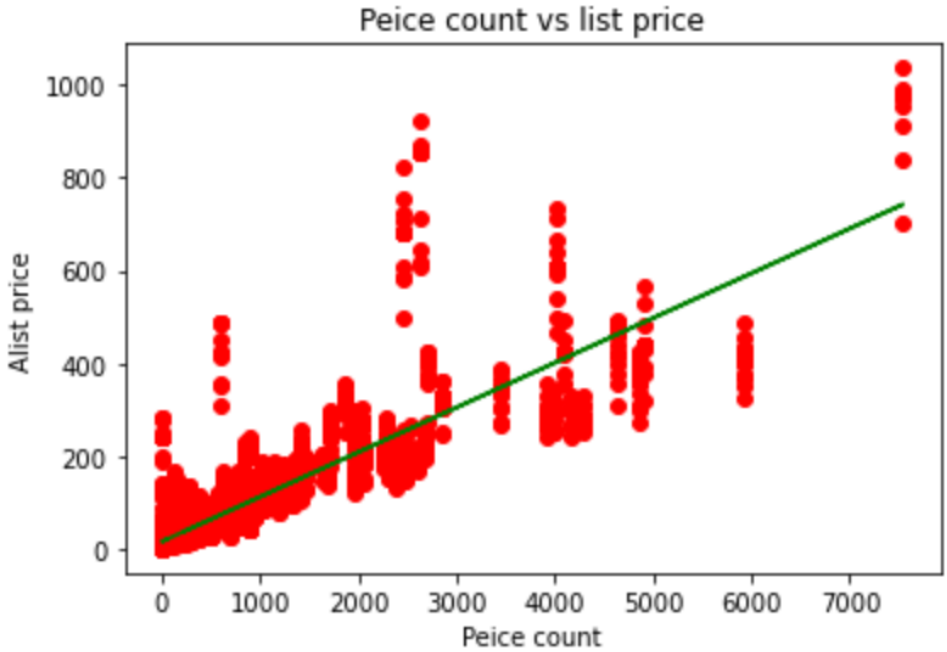


3.2. Evaluating the Models

Evaluation results show that the simple linear regression model correctly predicts 70% of the observed values and has a Mean Absolute Error of 20.08 which is much closer to zero when considering the spread of the data suggesting that this is a robust model.
Similarly The MLR model has a strong accuracy value of 75.70% and MAE of 20.08 suggesting that we can be confident in the accuracy of the predictions.

3.3. Linear regression and Multiple linear regression model predictions

Figure 2. Linear regression



The simple regression model predicts the price of an 8000 piece Lego set at $786.26. The Multiple linear regression model (using age as an additional independent variable) predicts an almost identical price of $787.06.

Even though the relative robustness and accuracy of the models provide valuable insights we have to also consider the spending habits of customers in this age group. Figure 3, shows that the highest price of any purchase for this age group is $259.87 and that only occurred the once. Figure 4. Shows that the most popular price of products purchased by this age group is $36.59.

Considering these insights it is clear that the price predicted by the models does not align with the purchase habits of the age group. This suggests that there are many factors that impact pricing. As such I recommend that further analysis is conducted in the future to look at other factors that impact pricing such as product features. Doing so will allow for the optimum price to be set thus helping towards attainment of the business objective.

Figure 3.

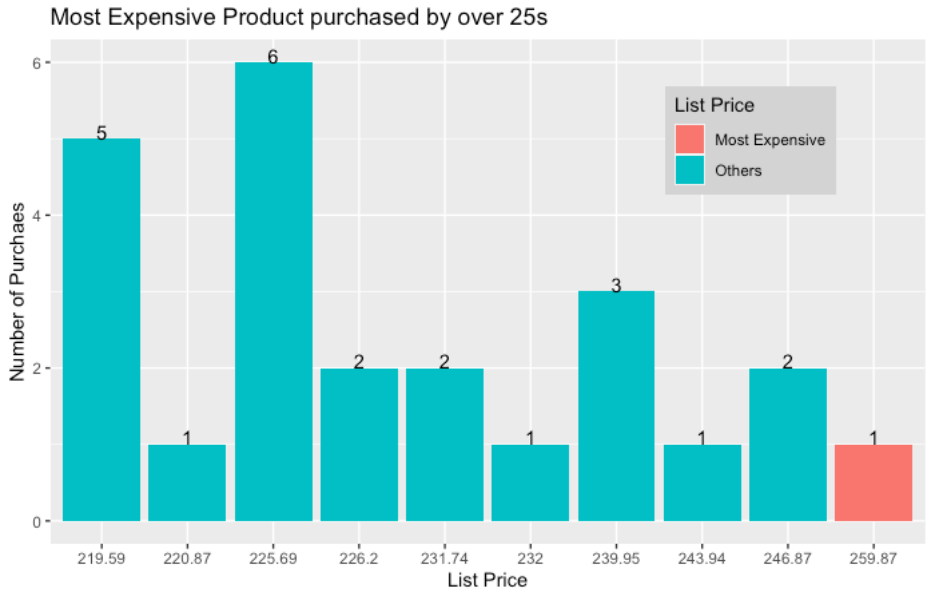
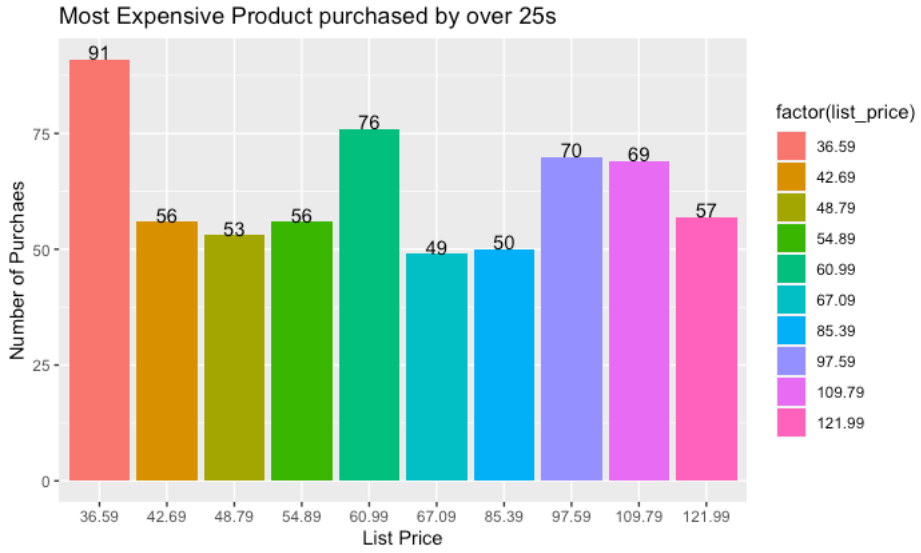


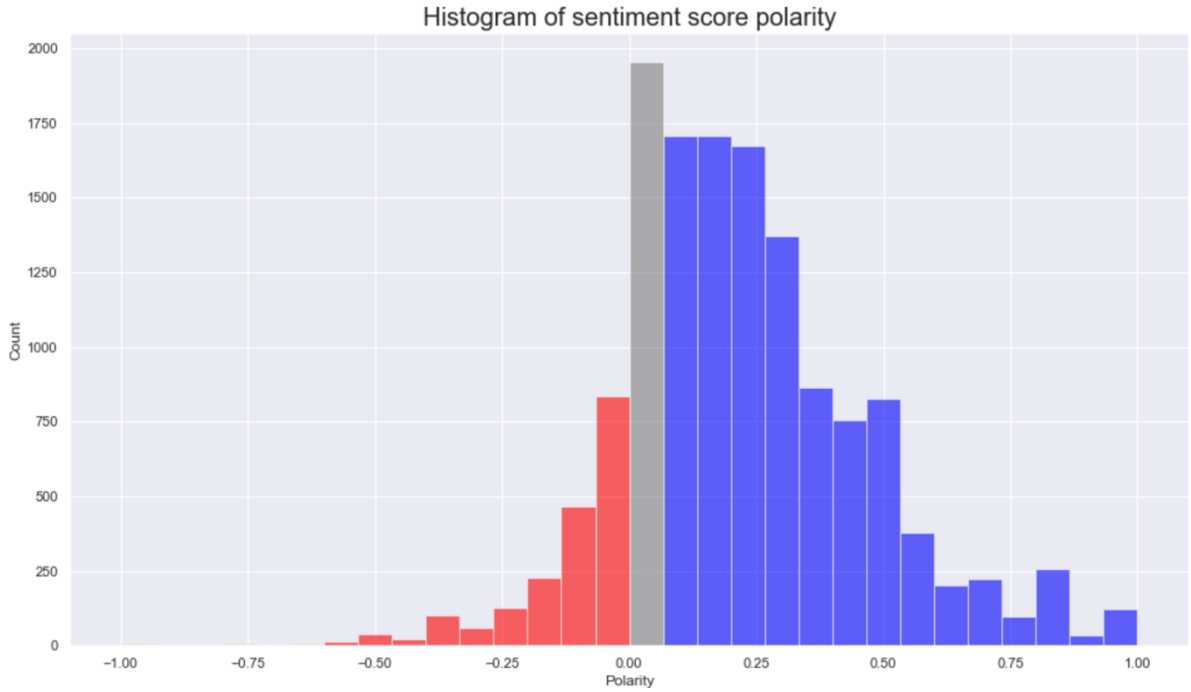
Figure 4.



4. **What is the general sentiment of customers across our products? And the top 20 positive and negative reviews.**

To answer this question I used sentiment analysis which is a natural language processing (NLP) technique used to determine whether sentiment is positive, negative or neutral. The NLTK package in Python was employed to prepare, transform and analyse the data.

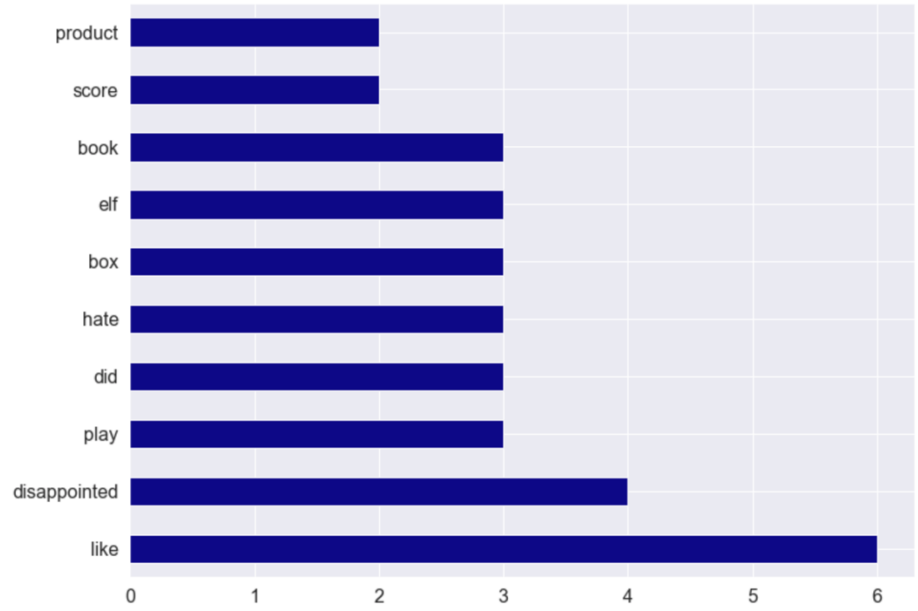
Figure 5. Polarity score



After cleaning and transforming the data we extract the polarity and subjectivity scores. The polarity score ranges between -1 and 1 (-1 being extremely negative and 1 being extremely positive). From Figure 4 we can observe that the general sentiment is positive. The data suggests that customers are generally satisfied with our products.

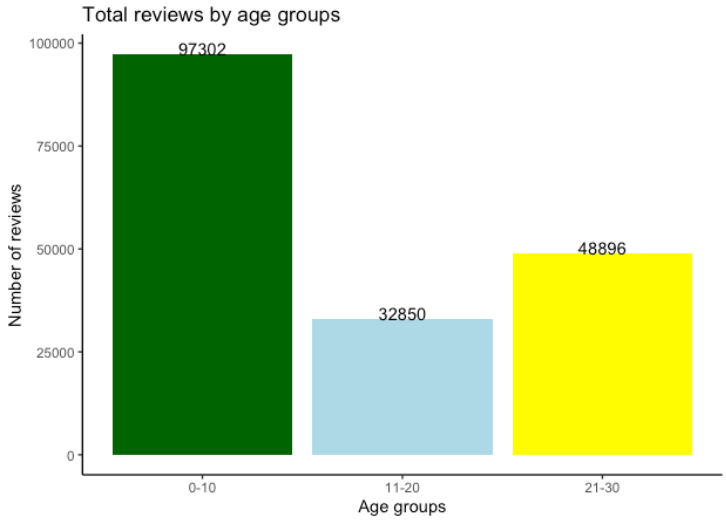
Having extracted the top 20 positive and negative reviews I used the Spacy library to generate named entities to isolate product details that appear in these reviews. Words that appear in the most negative reviews include box, elf and book suggesting that customers are dissatisfied with specific products features these words relate to. I recommend further research in the future using the entire dataset of reviews and linking the named entities directly to products. Such insights will give the business the data it needs to improve products and in turn general customer sentiment. Improvements in customer sentiment supports the business objective of increasing overall sales.

Figure 6. Named entities negative reviews



In line with above findings on customer sentiment it's also important to understand what segments of our customers most likely to leave reviews after purchasing products. Accordingly after grouping our customers in to three age groups we can observe that (Figure 7) the age group between 0-10 is responsible for more than 97,302 (54%) reviews. Children between 0 and 10 are unlikely to be submitting reviews so we can assume that these reviews are from parents or guardians who buy these products. These insights tell us particular attention paid to the products purchased by these groups to can lead to significant improvement in overall sentiment.

Figure 7. Age group with highest reviews



5. Predicting sales for the next financial year.

To predict sales for the next financial year a multiple linear regression model was fitted using North America and EU sales data to predict Global sales. Data was explored to ensure a linear relationship exists between the predictor variables and global sales as can be seen in Figures 9 and 10

Figure 9.

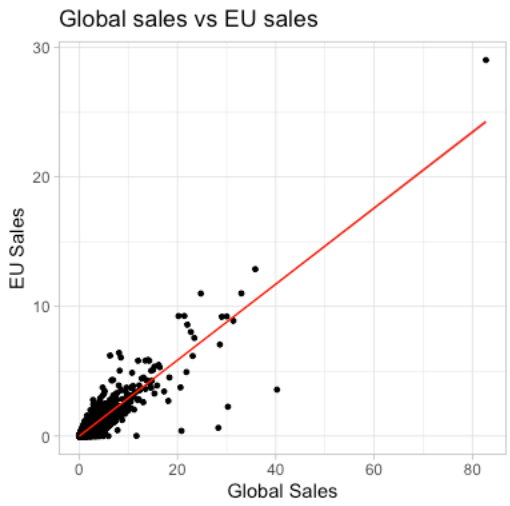
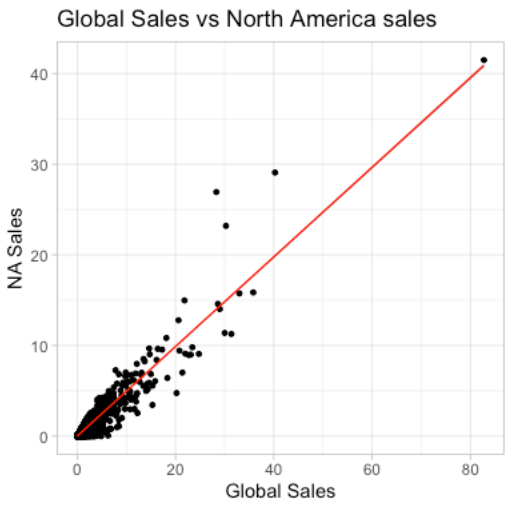


Figure 10.



7.2. Evaluating the model

The Adjusted R-Square shows that 96.4% of the observed values can be explained by the model suggesting that the model is highly accurate and predicted values will be right 96.4% of the time. Further evaluation showed that the model meets all of the assumptions that indicate a robust MLR model such as linearity, normality and homoscedasticity of residual errors.

7.3. Predictions

Figure 11. Predicted values for the 10 highest selling products

	Global_Sales	Predicted_Sales
1	82.74	86.962286
2	40.24	38.306689
3	35.82	35.667618
4	33.00	33.024907
5	31.37	25.008684
6	30.26	29.762309
7	30.01	25.594739
8	29.02	28.600826
9	28.62	26.351932
10	28.31	31.847270

The models prediction for the next financial years sales is shown in figure 11. Due to the model's evidenced accuracy and robustness I am confident that these figures are a reliable reflection of what to expect from sales next year.

I strongly recommend further analysis in the future using similar models to produce more refined and focused insights and predictions. Data held on platforms, publishers and genre can be used in prediction models to identify patterns and trends in specific subsets or regions early.