Introduction:

Turtle Games is a company offering a wide range of entertainment products. They have come to us to better understand their sales figures and determine the optimal sales strategy moving forward.

Our analysis helps turtle games identify the best pricing, take a closer look at customer sentiment and offer predictions of their future sales figures.

Week 1:

There are no missing values in our data set, no further action on the part is required. By comparing the min/max of the data we can see that piece count ranges from one to 7541 and the targeted age between 0 and 30.

I created a bar plot to observe trends between the median list price and the targeted age group. From it, we can see that the most expensive products are targeted to small children between the ages of 4-8 . Product prices between the ages of 24-30 was as well more elevated than for other age groups.

We need to create test and train data to evaluate how our model performs. We first need to train our model on the train data set and see if it can accurately predict values in the test data set. This allows us to determine an accurate model.

When conducting a linear regression, the p value for piece count is below 0.01 signifying it is statistically significant. This is not the case for age. An increase in 1 lego piece increases the price by 0.1$. The r2 value is 0.76 for both the simple and multiple linear regression signalling a strong correlation, yet age doesn’t seem to affect this association. The r2 values were similar in both the train and test regressions showing the good accuracy of our model.

The mean absolute error shows us the mean of absolute distance between actual and predicted values. The MAE is largely above 0, meaning there are large variances between predicted and actual values.

The optimal price for lego products with 8000 pieces is 792.6$. The optimal price for products with 8000 pieces geared to 30 year olds is 793.3$

Week 3:

My first word cloud was very uninformative and hard to identify the most frequent words that appear in the review as it is cluttered with stop words. I then plotted a new word cloud where I removed those stop words and was able to produce an informative graph.

The most frequent appearing words are the largest. The words game, one, play/player, card, really and great appear the most frequently. From really and great we can infer that many reviewers had strong polar opinions on the product. The words card and game inform us on what type of product is being sold and reviewed . From other large it is hard to extrapolate a significant meaning, however by graining down to smaller appearing words, we can understand the data further.

The bell curve shows that most reviews are positive, however, the overall sentiment seems to be relatively neutral, with the majority of the views possessing a polarity score between 0 and 0.25.

Many reviews among the 20 most positive polarity reviews are simple yet contain words such as awesome. The principal categories of reviews among these top 20 are about the condition of the product, how the product was exactly as advertised and fitted the customer’s needs, they also focus on games, boardgames and cards.

Among the top 20 negative reviews conversely the principality of the negative comments are to do with the quality of the product, how it was not what they expected and that the product was not entertaining.

Week 4:

To ascertain which age group submits the most reviews I created a scatterplot. From it we can observe that toys designed for those between the ages of 5-10 submit vastly more reviews than anyone else. The number of reviews per each other age category remained comparable, with a slight dip in reviews for products designed for those between 10-15 years of age.

To answer the question of what is the most expensive lego set purchased by customers who are at least 25 years old, I created a boxplot. The graph shows that the most expensive item sold at this age group is just slightly more expensive than 250 dollars. The mean price of products sold is 75 dollars with values beyond 110 dollars being outliers.

Week 5:

There were no missing values in the data set, no further action on that part is required. I converted all values under genre to lowercase, then I merged the columns genre and platform.

To visualise the skewness of the data I created a density plot. From that graph we can establish that the data is right skewed with the majority of data entries being close to 0.

To study a correlation that could predict to a certain degree global sales, I decided to observe the relationship between years and global sales. For this, I created a scatterplot with a linear regression line going through it. The regression line indicates a slight decrease in sales over time, however this line may be misleading due to outliers, most like sales after 2016, as the scatterplot shows that there is a positive association between an increase in time and an increase in global sales. Global sales, are steadily increasing until the early mid 2000’s where they significantly decreased.

Week 6

When conducting a multiple linear regression, both NA’s and EU’s sales p values are below 0.01 meaning they significantly predict Global sales. As well the r2 value of the multiple linear regression is 0.46 showing a very high correlation. An increase of 1$ in NA and EU sales shows an increase of 1.15$ and 1.35$ respectively.

The optimal price for forecasted global sales of lego products are 5055$, 5058$ and 5060$ in 2021, 2022, and 2023 respectively.

Conclusion

I suggest including further variables to improve the accuracy of our analysis and predictions. I recommend using these results to direct Turtle Game’s sales strategy. However, until a holistic understanding of the influencing factors can be ascertained I suggest proceeding with caution.