Analysis of Lego from 1950-2017

Abstract— **This paper analyses how Lego sets, and prices have evolved since the 1950s. Lego sets have become more colourful and Lego has collaborated with many companies to create sets of things people love. While the average size of a Lego set has increased, so have the prices, with a lot more sets being considered overpriced in the 2010s compared to the 1950s. The number of pieces a set has is the factor that influences price the most, challenging sets are also more expensive than easier ones. Predictive models were built to help us understand how Lego decides on prices of sets.**

1. Introduction

Lego was founded in 1932 and gained popularity in the 1970s. In 2017 Lego was the world's largest toy company, valued at approximately $7.6 billion [1] with some reports claiming Lego is a better investment than gold. Their goal to inspire and promote creativity using technology like robotics and popular themes such as Star Wars makes for valuable play time for both adults and children. A lot goes into making a Lego set behind the scenes that we do not usually take into consideration such as acquiring new licenses for themes such as Harry Potter and Lord of the Rings or designing the set and instructions which can end up taking months or longer depending on the size of the set.

Lego prices can be quite controversial with some sets selling for over $1000 such as the Millennium Falcon. Analysis on how sets and prices have changed and how prices are decided might help us understand if they are worth it, as well as provide useful insight from a business perspective for Lego and other companies who might be interested in future collaborations. A lot of research has gone into benefits of playing with Lego, some examples include lowering anxiety and stress, development of problem solving and spatial awareness skills, promotes creativity and experimentation as well as focus and concentration [2], making it a popular approach in therapy to develop social communication skills in children [3].

2. Data and Analytical Questions

I. Data

The data used for this analysis were sourced from Kaggle.com. First is a Lego database composed of 8 datasets ranging from 57 to 580,251 rows and 2 to 7 columns made up of discrete, continuous, and Boolean data along with strings. The database includes information on Lego sets such as number of pieces, colour's of pieces as well as theme name and release year from 1950 to 2017. All 8 datasets can be merged through mutual columns shown in the schema.

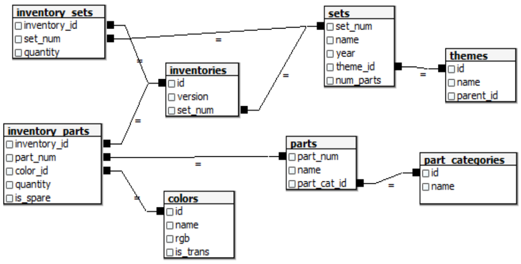

Figure 1: Schema of datasets in the Lego database

The second dataset contains 12261 rows and 11 columns comprised of discrete, continuous data as well as categorical and strings. This dataset provides information on the ratings as well as listed price of Lego sets for various countries including the UK and US. The Lego database and the second dataset can be merged on several columns making it suitable to answer more research questions. Not all the data is conserved when merging the Lego database and the second dataset since they do not contain the exact same Lego sets, however, enough is available to conduct the analysis.

II. Analytical Questions

The aims of the research questions are to provide information on how Legos have changed over the years but also provide useful and interesting insight from a business and customers point of view. The research questions aimed to answer are:

- How have Legos evolved over the years?
- How have prices of Lego sets changed over the years? How are prices different between countries?
- Is Lego overpriced?
- What factors determine the price of a Lego set?

3. Analysis

I. Data Preparation

All datasets in the Lego database can be merged through linked columns, however, columns that can be merged have different names in each dataset, for example, the **themes.csv** dataset has a column named **id**, which can be linked to the **sets.csv** dataset on the **themes_id** column. All mutual columns have been renamed to allow datasets to be merged. While all datasets can be merged, it can result in a very messy final dataset, therefore, only datasets which help answer the analytical questions are used and merged. Datasets have also been grouped and aggregated by specific features to further help with the analysis, an example of this is grouping a dataset by theme and calculating the average number of parts or number of unique sets per theme.

In the Lego database, the dataset **colors.csv** contains 2 colour's which can be considered null values, *"[No Color]"* and *"Unknown"*, after inspecting Lego pieces with these colour's through the part_num column, they represent unique items such as stickers or mini-figures and were not removed from the database. The second dataset contained a lot of null values, all numeric null values were replaced with the mean value of that column, null values in the categorical column **review_difficulty** are replaced with *'Average'* and rows missing a theme name were dropped since there were only 3.

Using the second dataset, we can compare prices of Lego sets between countries, merging **sets.csv** to the second dataset on **set_num** allows us to analyze price change over the years as well. Since both datasets do not contain same Lego sets, data is lost when merging, leaving us with 4429 rows. From this, there are 398 unique Lego sets. For a fair comparison between certain countries, only 160 unique Lego sets are used since they are priced in each country. As the US has the highest number of unique Lego sets, it is used to analyze prices of sets as well as build linear regression models.

II. Data Derivation

Creating new features from existing ones can benefit the analysis, a new variable was created called **binned_years,** which assigns each Lego set to a specific decade based on its year, this can be very beneficial especially in cases where there is missing or lost information. Combining the Lego database and second dataset resulted in loss of data on Lego sets, when comparing how prices of sets have changed over the years, there are big jumps in average price of Lego set, this is because not all Lego sets are recorded each year in the merged database, hence binning the years allows for a better comparison. The second main feature created is **color_ratio**, this is ratio of colour's used by Lego and is calculated by counting the number of pieces of a specific colour and dividing by the total number of pieces used per decade.

The second dataset contains a lot of categorical columns which can provide more insight on what features affect the price of a Lego set. These columns are **ages**, **theme_name** and **review_difficulty**. Using OneHotEncoding, we can transform these categorical columns into multiple binary columns. Since there are 40 unique themes, 31 unique age groups and 5 review difficulties recorded, we end up with an additional 76 binary columns which can be used as predictors for simple and multiple linear regression, as well as tell us how themes, age groups and Lego set difficulty are correlated with price.

III. Construction of models

Linear regression models are built to predict the price of a Lego set. Inspecting the correlation of variables with price tells us the features which impact the price of a set the most, the implementation of OneHotEncoding provides more information of what impacts the price of a Lego set. For simple linear regression, the variable with the highest correlation to price is chosen as the predictor. For multiple linear regression, the top 10 variables with the highest correlation are chosen as predictors excluding product ID as that is not an important feature of a Lego set, the rest of the variables have a correlation less than 0.12 and are not expected to significantly change the results of the linear regression models.

IV. Validation of results

The second dataset is used to construct models since it contains prices, 75% of the dataset is used to train and 25% is used for testing as well as 10-fold cross validation on the whole dataset to contrast results. Models are compared using their mean square error as well as their R^2 value.

4. Findings, Reflections and Further Work

First, we investigate how Lego sets have changed since the 1950s.

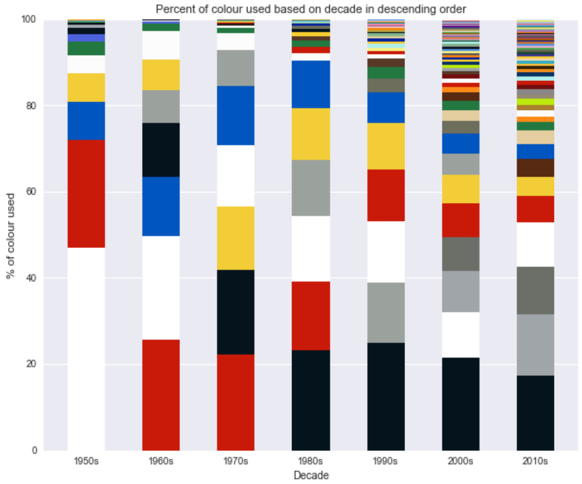

Figure 2: Inspired from Carron, J. (2016). 67 Years of Lego Sets.

Lego sets have become darker since the 1950s, with black and shades of grey being the most used in the 2010s. Lego sets have become more colourful as the number of different colour's used has increased gradually to approximately 100 used in 2010s. These changes in colour's may be affected by the diverse themes added.

Features such as number of sets released, number of pieces and unique themes released each year also provide insight on how Legos have changed.

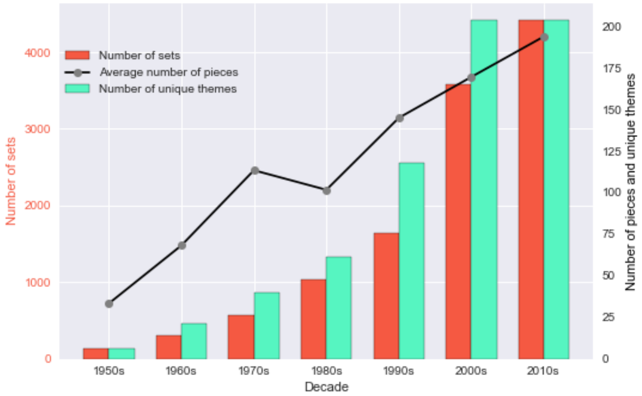

Figure 3: Number of sets, unique themes, and average piece count per decade

Figure 3 shows that the number of sets, number of unique themes and average number of pieces have increased since the 1950s. Lego sets have become larger and are a lot more sets are being made. Until the 2000s, the main Lego themes were created by Lego themselves, such as Technic, which has the highest set count in the 1990s, by the 2000s, Lego branched out more, collaborating with popular companies and creating Lego sets based on scenes from popular movies such as Star Wars and Batman.

Next, we investigate how Lego prices changed each decade and between countries, the prices in the second dataset are all assumed to be in dollars since it was not specified. Merging the datasets leaves us with Lego sets starting from the 1990s, the average price of a Lego set is approximately $10 for all listed countries in the 1990s. Each decade, the price and variance of prices between countries increase significantly, since the average size of set increased, this is expected. Another factor could be the inflation of currency.

For a fair comparison, only prices of the same Lego set are compared between countries. The standard deviation of prices between countries in the 1990s is approximately $7 while in the 2010s it is around $118 with the lowest average prices being in Canada and Denmark and the highest in Czechia according to figure 4.

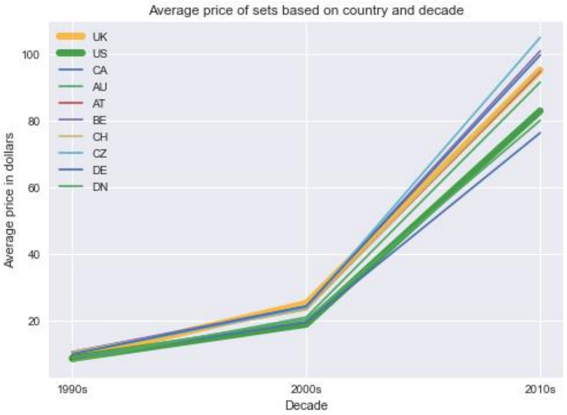

Figure 4

The US had the highest number of recorded Lego sets after merging the Lego database and the second dataset and is used for further analysis on what affects prices of Lego sets as well as the building linear regression models.

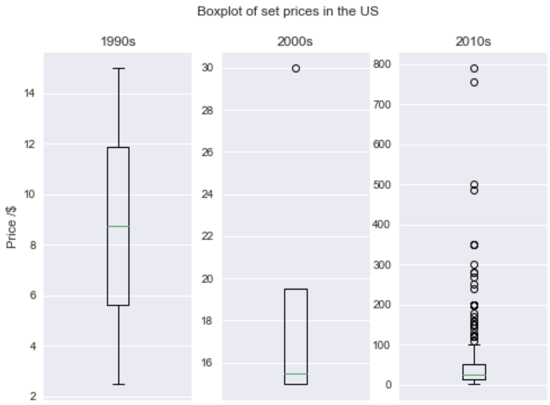

Figure 5

While the dataset does not contain all sets for each year, it is clear there is a noteworthy increase in outliers based on price suggesting Lego Sets can be overpriced but not all are.

Building a linear regression model will provide the expected price of Lego sets which can be compared to the actual price. First, a simple model is built with piece count as the predictor. Using cross validation, the mean square error obtained is 1479 and the average R^2 score of 0.74, telling us how much variance in the price of a set can be explained by the piece count.

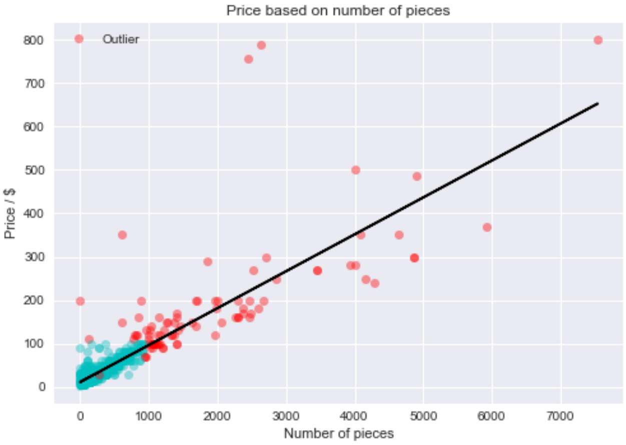

Figure 6: Plot of price based on number of pieces with fitted simple linear regression model

The top 10 highest correlated features were used to build the multiple linear regression model, as shown in table 1. This model gave a mean square error of 1437 and a mean R^2 score of 0.75, making it slightly better than the simple linear regression model. Piece count has the largest effect on price, another significant factor increasing price is if the set is challenging.

Table 1: Top features

Feature	Correlation
piece_count	0.863959
Challenging	0.492232
SERIOUS PLAY ®	0.458823
16+	0.454084
num_reviews	0.424305
6+	0.300810
14+	0.300395
Creator expert	0.256458
11-16	0.158243
9-14	0.127637

As not all Lego sets were available for the analysis, further work would include web scraping all Lego sets, standardizing currencies to disregard inflation as well as analyzing outliers further to see if certain features tend to make a Lego set overpriced.

References

[1] (2017). The annual report on the world's most valuable toy brands. [online] Available at: https://brandfinance.com/wp-content/uploads/1/brand_finance_toys_25_2017_report_locked.pdf

[2] (2019). 10 Incredible Benefits of Playing with Lego® -. [online] Childrenswellnesscentre.co.uk. Available at: https://www.childrenswellnesscentre.co.uk/10-incredible-benefits-of-playing-with-lego-2/ [Accessed 10 Dec. 2020]

[3] Autism.org.uk. (2014). Using Lego therapy with autistic pupils | Network Autism. [online] Available at: https://network.autism.org.uk/good-practice/case-studies/using-lego-therapy-autistic-pupils [Accessed 10 Dec. 2020]

Section	Word count
Abstract	101
Introduction	236
Analytical questions and data	256
Analysis	750
Finding, reflections and further work	599