From Bricks to Bucks: Exploring Influential Variables in Sales Patterns of Lego Sets

Samuel Love, 84107034

Table of contents

Abstract

This project explored the influential variables in sales patterns of Lego bricks (a popular worldwide brand of plastic construction toys). It aimed to ascertain if it were possible to accurately predict the secondary prices of Lego sets using a defined set of characteristics, and to discover if Lego sets based on licensed intellectual properties differ from Lego original sets. This would be useful information for anyone interested in predicting the future value of their sets.

This project consisted of two parts; examining trends over time of various Lego characteristics, and using supervised learning to training machine learning algorithms to produce predictive models. The 5 main objectives were; Time-series analysis, producing visualisations for data cleaning and analysis, k-means clustering, predictive modeling linear regression, and feature ranking focusing on historical price data.

Four datasets from three sources were used; JB, D15, D18, MP. The data was retrieved, cleaned, joined, re-cleaned, subsetted, clustered, trained and then tested. The data sets were then joined to make a data set Merged. The machine learning algorithms were linear regression, elastic net, random forest, and support vector machine. Research limitations included small, biased data, and unverifiable assumptions.

The results showed that Lego sets are increasing in quantity and complexity over time. Increased complexity in Lego sets correlates with an increase in price. There was potential for organic clustering of the data. Historical price variables are very significant predictors of current prices. However, models without historic price variables performed better on average.

Introduction

This project pertains to Lego, a worldwide brand of plastic construction toys produced by Danish company The Lego Group (The LEGO Group, 2023). The project and supervision is provided by the University of Canterbury School of Mathematics and Statistics.

The Lego Group was established in 1932 (The LEGO Group, 2023). The Second World War (1939-1945) included German occupation of Denmark from 1940 until 1945 (Unknown Author, 2012). The war significantly strained every countries resources. Successful companies relied on innovation and creative thinking to overcome difficulties. The Lego Group initially produced wooden toys, which continued until 1960. Between 1949 and 1963, plastic bricks made from cellulose acetate were introduced. They already accounted for half of The Lego company's output by 1951. However, this plastic would deform with heat and change over time. Acrylonitrile butadiene styrene was introduced in 1963 and is still used to this day (Lauwaert, 2008).

Innovation has continued with the variation in Lego sets which are categorised into themes. LEGO either created original themes such as Lego City, or licensed famous intellectual properties such as Star Wars. The Lego Group consistently releases new sets and retires old ones. As such there is a wide variety of Lego, most of which is traded in secondary markets.

My main research question is: 'Can we accurately predict the secondary prices of Lego sets using a defined set of characteristics?' My secondary research question is: 'Do Lego sets based on licensed intellectual properties differ from Lego original sets?'

This project consists of two parts. Firstly we will examine trends over time of various Lego characteristics to develop an understanding of The Lego Group's creativity. Then we will train several machine learning algorithms using supervised learning to produce predictive models. The models will train and test on two subsets of data to predict prices in 2019 and 2022. The main difference is predictions on the 2019 subset includes historical price variables (recommended retail price, 2015 secondary prices, and 2018 secondary prices). These models will develop our understanding of the feasibility of predicting secondary prices and which Lego set features are significant predictors. This would be useful for anyone who collects or invests in Lego sets that wants to predict the future value of their sets.

To achieve this goal, objectives include:

- Time-series analysis based on the year that sets were released.
- Visualisation to help with data cleaning and for analysis of time-series data, clustering effects, and model metrics.
- K-means clustering on two standardised subsets of data to explore any organic groups in the data.
- Predictive modelling on two standardised subsets of data using four algorithms; linear regression (LR), elastic net (EN), random forest (RF), and support vector machine (SVM), as well as a null model.

- Feature ranking performed by EN and RF primarily to see if historical price data is a useful good predictor of current prices.

The main challenges were defining codeable parameters and then finding reliable secondary market price data.

The Shiny package from R Studio is the primary tool to create an interactive dashboard of visualisations (Hosted Dashboard). The dashboard has three main sections.

There are four datasets from three sources, discussed below.

Data

JB

The main dataset (**JB**) is provided by Jason Bryer (2023) in the JBryer/brickset R package, a package with tools to interact with the Brickset API. This dataset contains 19,409 observations of 36 variables of contextual Lego data (appendix table 3). Notable variables include the year each set was released (from 1970 to 2023 inclusive), the number of pieces in each set, the number of minifigures in each set, the US recommended retail price, and the theme of each set. The unique identifier variable, SetID, is present for every row.

There is 65.6% data present and 34.4% missing. The majority of missingness is seen in approximately the first third of the observations. This logically reflects the increase in scope of collected data over the years as Brickset grew. However, there is consistent missingness of variables *subtheme*, *pieces*, and *minifigs*. The missingness of *pieces* is unexpected as every set should have a number of *pieces* included, unlike *subtheme* and *minifigs* which are not necessary for every set.

This data will be useful for creating visualisations that show trends over time. It also serves as an excellent frame of reference for the price data as it provides many variables that can be used for machine learning. The only price data in this dataset is RRP so there are three supplementary datasets that contain secondary market pricing data.

D15 & D18

The main secondary price dataset (**D15** & **D18**) is provided by Victoria Dobrynskaya (2021), generated for the paper: LEGO - The Toy of Smart Investors. **D15** contains 2332 rows of 11 variables (appendix table 4), whilst **D18** contains 2332 rows of 22 variables (appendix table 5). This data was parsed from Brickpicker and represents the average of the 30 most recent transactions on Ebay for each set. **D15** contains yearly average prices for new and used sets in 2015, whilst **D18** contains monthly average prices for sets in 2018 and the first quarter of 2019. This data will be useful for visualising changes over time and machine learning. Common variables have the same number of unique values, indicating that these dataframes are near identical.

D15 has 99.9% data present and less than 0.1% missing. Variables *#.of.minifigure* and *Secondary.market.prices.of.new.sets.in.2015* are the only ones with missingness. **D18** has 98.9% data present and 1.1% missing. Monthly price variables comprise this missingness. Interestingly, the missingness is from the 5th monthly price variable onwards and is consistent across observations.

Looking closer we see that the first four monthly price variables are character type variables but all contain the value 0 for these observations. If these observations are missing due to user error (possible since the data was parsed from Brickpicker directly and is not available

there now), then this missingness is missing completely at random. However, it is possible that this data is missing not at random if there were no sales for the specific sets each month. This is more likely as some of the months do have price data. This type of missingness can introduce bias, however, since most of these N/As will be omitted when generating yearly averages, effects will be minimal.

MP

Another supplementary dataset (MP) is provided by mrpantherson (2023). This dataset contains 5075 rows of 9 variables (appendix table 6). The secondary prices for the year 2022 were scraped across different sellers whilst other parameters were sourced from Rebrickable.

There is 100% data present in **MP**. The 47 year the sets were released across, covers a wide range that is comparable to **JB**.

Merged

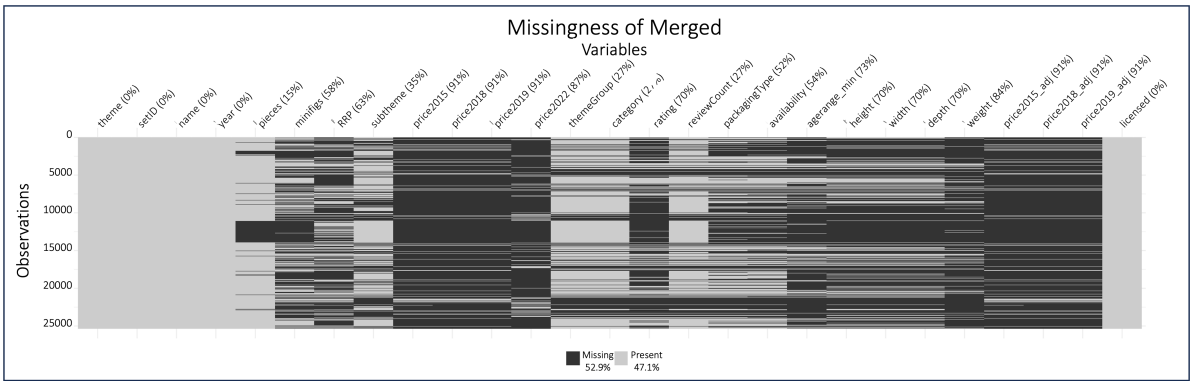
The result of joining the four datasets by the column setID into one is the dataset referred to as **Merged**. **Merged** has 26324 observations of 29 variables (appendix table 7). This is 6915 more than the **JB** dataset. As there are 2322 observations in **D15** and **D18**, and 5075 observations in **MP**, there are only 482 *setID*s that were common amongst the datasets. This caused high rates of missingness.

There is 52.9% missingness in **Merged**. This can be attributed to the already high 36.2% missing data in **JB** in conjunction with the relatively small datasets **D15**, **D18**, and **MP** (12.5% and 27.5% of **JB** respectively). Additionally, the D's price variables have 91% missingness and, as they are transformed to columns that are adjusted for inflation, the effect of this missingness is doubled.

Looking at this correlation of missingness, we see that the method choice between Kendall, Pearson, and Spearman makes no noticeable difference to the visualisation. This is interesting as many variables have skewed data which is more suited to Spearman's correlation as it handles ordinal or non-normally distributed continuous data. This is likely a result of having many discrete variables.

Grouping variables by OLO, GW, and HC all reveal that width, height, and depth have highly correlated missingness, as do *availability* and *packaging type*, and *themeGroup*, *category*, and *reviewCount*. This is also apparent looking at the missingness graph. The first group have inherited correlated missingness from **JB** whilst the other two groups had no missingness in **JB** and so have gained correlated missingness during the joining of the dataframes. Another group with highly correlated missingness are the price 2015, 2018, and 2019 variables. This is again due to the missingness introduced when joining. There are no major issues with missingness in **Merged**, other than its sheer quantity.

Chart 1: Missingness of Merged



The boxplots of numeric data show many potential outliers. However, when considering the nature of the variables with outliers, there are justifications for these observations to be legitimate. Variables *pieces*, *minifigs*, and *reviewCount* are highly skewed with many low counts that vastly reduce the IQR. *weight* is similarly skewed whilst *depth* is less so. A manual inspection of notable outliers reveals that they result from unusual Lego sets rather than poor data quality.

The rising-value chart shows nothing of concern but does visualise the skewness of most variables. However, this chart is more useful for continuous variables.

Additions

Additional columns for prices adjusted to inflation have been added to **Merged**. Inflation multipliers are $USD yearly averages as per the Federal Reserve Bank of Minneapolis (2023) calculator. Price variables for 2015, 2018, and 2019 were inflated to 2022 prices. This standardisation will reduce noise from any disparity between price variables. Of note however, is that RRP was not inflated.

Another addition was a column for licensing categorisation for themes. This was scraped from a Wikipedia table (Wikipedia contributors (2023)) then joined to **Merged**. Manual investigations were needed due to a large proportion of unmatched themes. This resolved to 100% present *licensed* variable which is categorised into *Original*, *licensed*, and *Hybrid*. This will be used when investigating whether licensing has any noticeable effects.

Limitations

There are some limitations to consider when visualising and modeling this data. The data could be biased data as popular sets can be overrepresented creating sampling bias. Additionally, the data collectors arbitrarily chose which sets to track, creating selection bias. There is lack of context with the limited documentation of **MP**. Generating 2019 price averages is misleading

as only the first four months of the year are available. Also, averages were the most recent 30 transactions on eBay which is misleading for sets with different numbers of transactions.

There are relatively small sample sizes with **D15** and **D18** about 12.5%, and **MP** about 25% the size of **JB**. This introduced missingness into **Merged** which affects the usefulness of visualisations since omitting missing observations from graphs can create incorrect impressions. It also affects the machine learning process since it necessitates further cleaning to remove N/A's before training models.

Assumptions about the data were difficult to verify. They include $USD, as-new quality of sets, and consistent characteristic variables. Finally, the changing landscape is not well represented with static data. Changes could include effects from seasons, popularity, or retirement phases. eBay Inc. (2020) reported that "spending on LEGO sets increased 89 per cent since the launch of Channel 9's LEGO Masters".

MP setIDs had an odd problem of having -x appended where x represents different sets with the same ID number. This is an artefact of earlier years not having unique setIDs meaning these sets are typically unrelated. The vagueness of the collection process makes verification difficult and reduces data trustworthiness.

Method

Data Retrieval & Cleaning

Data retrieval and processing takes place in the file **DATA601Data.qmd**. The datasets are all downloadable csv files. Several user-defined-functions assist with the cleaning process. Cleaning the raw datasets facilitates the joining process. This requires consistent variable names and types, and adjusted missing observations.

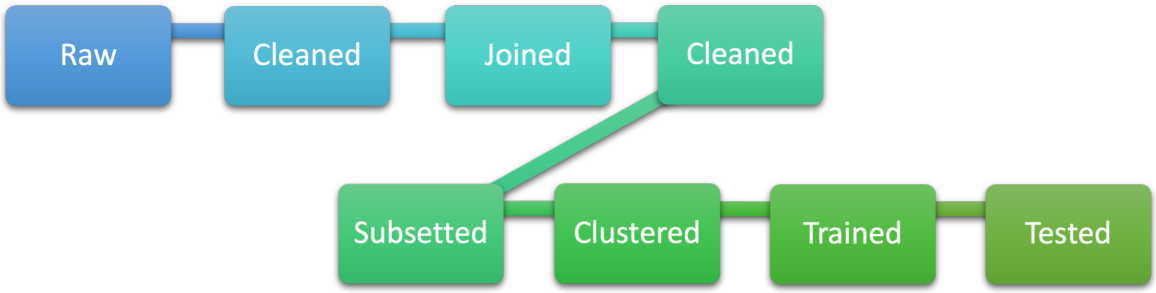
Cleaning **JB** involves renaming, dropping, re-categorising, and mutating variables. Some variables (for example URLs) are omitted, whilst others have missing observations correctly categorised. The cleaned dataframe has 19,409 observations of 23 variables (decreased from 36). Missingness is reduced to 28.9% (from 36.2%).

Cleaning **D15** and **D18** is very similar. The price variables require separate handling. We select the price variable from **D15** that represent 'new' sets on the secondary market, and aggregate the monthly price variables from **D18**. There was one mismatched setID between the two datasets, which manual intervention resolved. The cleaned dataframes both have 2322 observations of 9 and 8 variables (down from 11 and 22 respectively). There is missingness of <0.1% and 0.4% respectively.

Cleaning **MP** was much the same. Of note, *theme* and *subtheme* variables required parsing from a single variable. *setID*'s with -1 were assumed to be standard so were converted accordingly, leaving other suffixes as is. The cleaned dataset has 5075 observations of 7 variables (down from 9). Missingness has been introduced in the new column *subtheme* and the price2022 column. This missingness is relatively high at 71% and 36% respectively leading to overall missingness of 15.3%. The *subtheme* missingness is expected, however, the high missingness of significantly reduces the amount of useful data for training models.

Cleaning **Merged** involved manual intervention to fully populate *licensed* and consolidation of theme names.

Subsets of secondary price datasets were created for clustering and machine learning. The cleaned 2019 subset (resembling **D15** and **D18** combined) has 2,277 observations of 12 variables, whilst the cleaned 2022 subset (resembling **MP**) has 3,241 observations of 7 variables. *licensed* was included in both subsets, whilst *subtheme* was excluded from the 2022 subset for excessive missingness.



Visualising

Visualisations are used at every step of answering the research questions. Summaries and missingness plots visualise both the raw data and the cleaned data (including the subsets). **Merged** has additional visualisations that provide information on variable distribution, missingness correlation, outliers, and homogeneity.

The primary exploratory visualisations are split into four groups:

- *Categories* displays the number of sets released each year, and categorical variable distributions for each year.
- *Dimensions* displays numerical variable distributions based on the *licensed* category.
- *Prices* displays in depth price variables comparisons. *Price Boxplots* and *Average Prices by Year* compare the price variables with each other. *SetID Price by Year* examines different prices for individual sets. *Price vs Numeric* shows scatterplots of numeric and price variables. *Price vs Categorical* shows boxplots of categorical and price variables.
- *Clusters* visualises the effects of clustering the price variables into two or three clusters. The 2019 subset is clustered on its four price variables whilst the 2022 subset is only clustered on *price2022*.

The subsets are visualised in *Machine Learning Models*, as are line graphs and boxplots of metrics, line graphs of hyperparameters, and bar graphs of features. These show the best performing model for each bootstrap, so are distributions of useful models.

Machine Learning

The caret package by Kuhn (2023) "streamline(s) the process for creating predictive models". The documentation is extensive and was used to develop the machine learning process for this project.

The goal is to use supervised learning to predict prices for 2019 and 2022. Since most algorithms require datasets with complete cases, the subsets have no missingness. The 2019 subset has

2,277 observations of 12 variables. The 2022 subset has 3,241 observations of 7 variables. Comparisons of subsets will reveal the effects of price variables (*RRP*, *price2015*, and *price2018*) that are present only in the 2019 subset.

In preparation for training, the categorical variables *theme* and *licensed* were one-hot-encoded, with *theme* being grouped into 'superthemes' to reduce cardinality. All variables were centred and scaled to allow equitable treatment by the algorithms, enabling fair comparisons. Data is split into training and testing sets using a 75/25 ratio. The training set was used to generate 20 bootstraps to enable training distributions of each model that will give a clearer indication of performance.

Using a range of algorithms increases the chances of finding a successful model. Selected algorithms:

- Linear regression (LR)
- Elastic net (EN)
- Random forest (RF)
- Support vector machine (SVM)

A null model was included as a comparative baseline.

There are three metrics suitable for regression machine learning problems:

- R-squared (R^2) measures the proportion of variance in the dependent variable that is predictable from the independent variables.
- Root Mean Square Error ($RMSE$) measures the standard deviation of the residuals.
- Mean Absolute Error (MAE) measures the average absolute difference between the observed outcomes and the predictions.

EN, RF, and SVM have tune-able hyperparameters. These were tuned using a Cartesian grid search to methodically find the best performing model.

- EN uses alpha to control the mixing ratio between L1 (lasso) and L2 (ridge) regularization penalties, and lambda to control the regularization strength.
- RF uses mtry to specify the number of variables to be randomly sampled as candidates at each split when building trees.
- SVM uses C to balance the correct classification of training examples against maximising the decision function's margin, and sigma to define how far the influence of a single training example reaches.

EN grid tried 5 values for alpha and 10 for lambda. RF grid tried 4 values for mtry. SVM grid tried 4 values for C and 10 for Sigma.

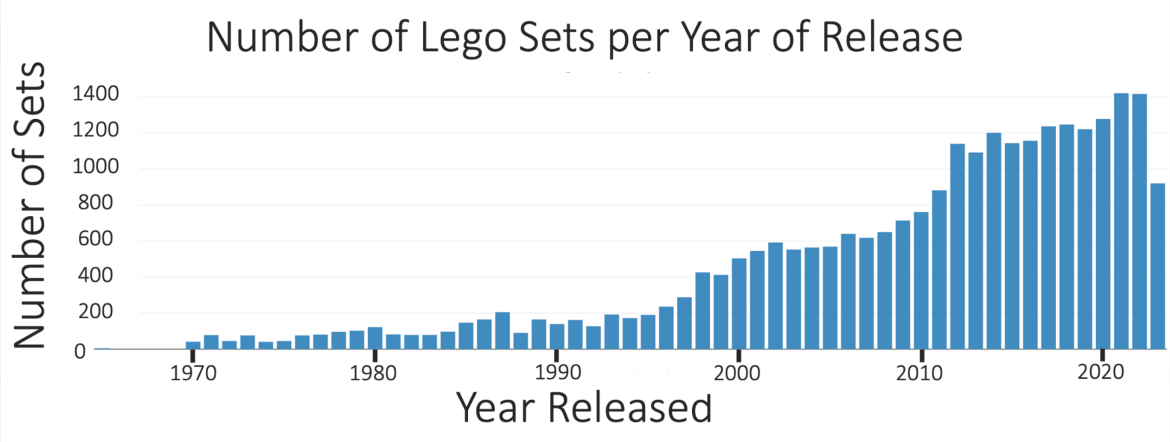
This results in training 50 EN models, 4 RF models, and 40 SVM models on each bootstrap. With 20 bootstraps, we train 1885 models in total on each subset of data. This is computationally expensive but achievable with such small datasets. Only EN and RF have built-in feature selection.

Results

Categories

We see a significant increase in the number of sets released each year. The earliest year in the data, 1970, has nearly the fewest sets at 41. The year 2000 is the first with more than 500 sets whilst 2012 is the first over 1000 and 2021 has the most sets at 1420. There is no missingness of *setID*s.

Chart 2: Sets Per Year



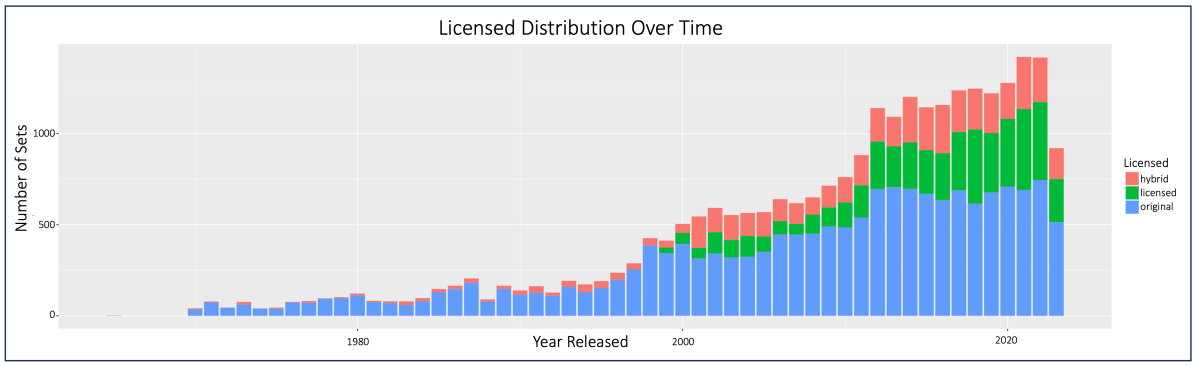
There are 162 unique *themes* making the graph difficult to interpret. However, we do see a fairly consistent spread of themes each year with the main difference over time being an increase in numbers. There is no missingness of themes.

There are 968 *subtheme*s so we can visualise the top 50 (by count). This is similar to the themes, just slightly more diverse for the different years. The higher granularity accounts for this. There is 34.2% missingness of *subtheme*s.

*themeGroup*s having only 16 groups produces a more interpretable graph. Early years (1965-1975) contain high proportions of 'technical' and 'vintage' theme groups whilst later years have higher proportions of *licensed* and 'miscellaneous' groups. Other groups remain fairly consistent over time. There is 26.2% missingness of theme groups.

Theme IPs shows that *licensed* themes date back to the 1970s, but have steadily increased, especially after the year 2000. Recent years consist of roughly 40% themes either being fully or in part *licensed*.

Chart 3: Licensing Categories



Categories also show diversification from the year 2000. There is 26.2% missingness of category.

Packaging type is predominantly box or bag with more diversification after the year 2010. However, there is significant missingness of 50.1%.

Availability is predominantly retail, with increasing retail-limited in recent years. Again, there is significant missingness of 51.7%.

Dimensions

The Lego set dimension variables can visualise any differences of licensing. Dimensions visualised are *pieces*, *minifigs*, *height*, *width*, *depth*, *weight*, *rating*, *reviewCount*, and *age*.

Looking at the full dataset, we see a positive trend in *pieces* and negative trends in *depth*, *weight*, and *age*. Other variables have consistent trend lines over time. *height*, *width*, *depth*, and *weight* have many more variable spread after 1995, making trend lines more sensitive to changes.

Licensed themes show consistent distributions starting in 1999. There are negative trends in *height*, *width*, *reviewCount*, and *age*. Weight data is available from 2004 onwards.

Original themes populate the earlier years. The distributions look very similar to those of the full dataset. There are increasing trends of *height* and *width*, and again negative trends in *depth*, *weight*, and *age*.

There are more Hybrid themes in later years than original themes. There are positive trends in *rating* and *age*, with negative trends in *width*, *depth*, and *weight*.

There are differences in the trends and distributions of licensing. However, it proves difficult to attribute these to differences to the effects of licensing alone due to noise generated from missing data and fewer historical sets. An analysis of price data may be more enlightening.

Prices

Price Boxplots of *RRP*, *price2015*, *price2018*, *price2019*, and *price2022*, show heavily skewed distributions. There are many outliers above the $500 mark, even with an increased tolerance (IQR multiplier). *RRP* has no values above $1000, whilst the others all include values over $1500. Boxplots of standardised prices (adjusted to 2022) reveal that *price2015* has the highest ranging outliers but a very similar distribution to *price2022*. *price2018* and *price2019* are similar with each other which have smaller interquartile ranges. When centred and scaled, *RRP* has the largest distribution.

Chart 4: Price Variable Boxplots

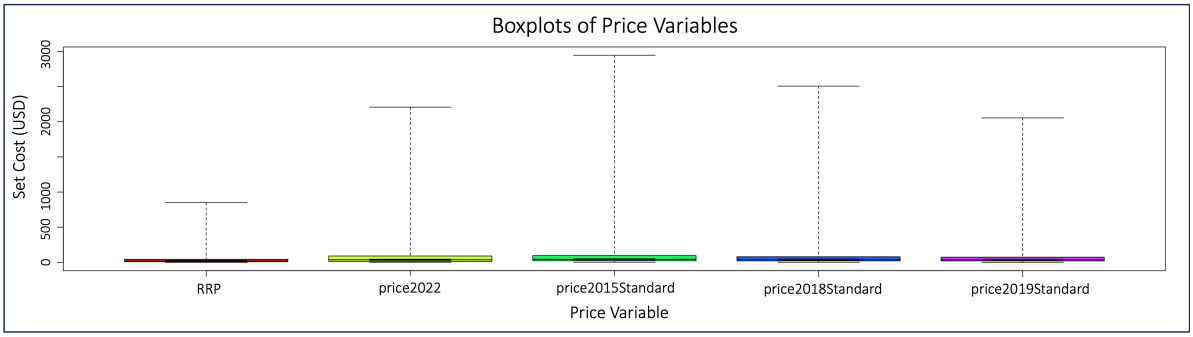
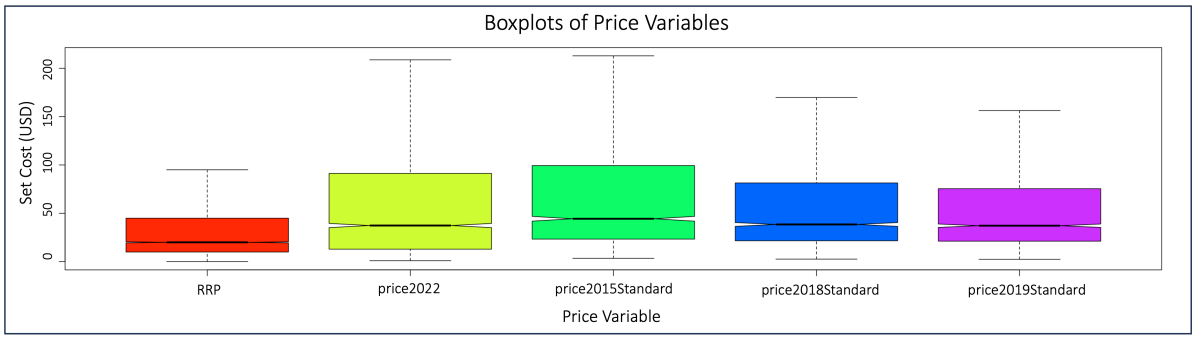


Chart 5: Price Variable Boxplots without outliers



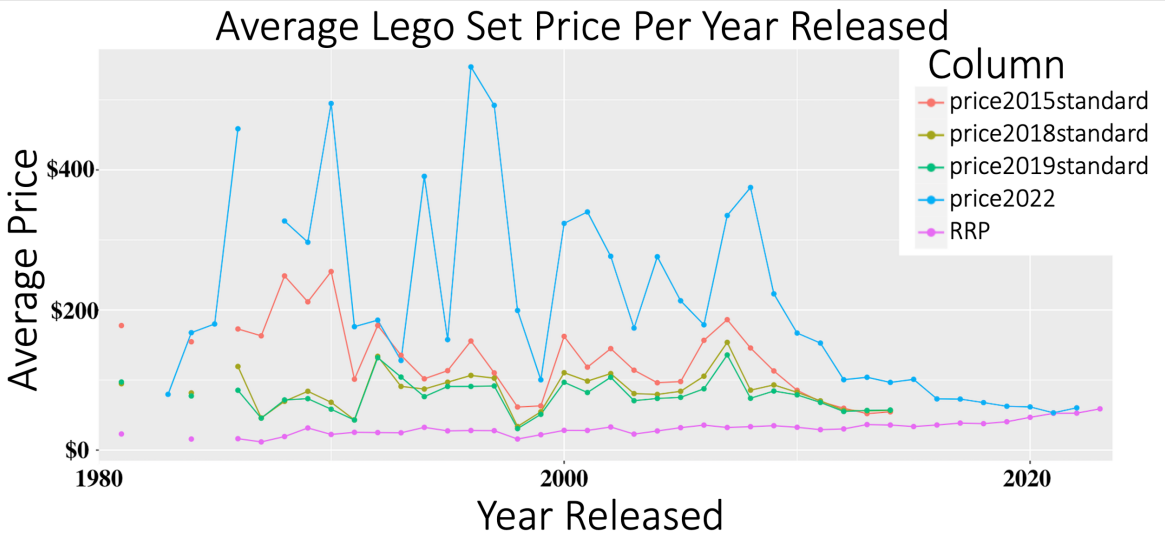
This indicates that whilst most sets remained at or just below their RRP price, some drastically increased in price. There are limitations however, including varying amounts of data for each variable and RRP not being adjusted for inflation. We could further investigate the sets that have the highest secondary market prices to develop our understanding.

SetID Price by Year shows steadily increasing amounts of sets in later years. It is difficult to determine any general patterns in prices for sets in a given year, however we can see which sets have multiple price data points and how these compare.

Average Price by Year shows that standardised 2018 prices have the same trend as 2019 but are generally slightly higher. Standardised 2015 prices roughly follow the same trend, except the late 1980s having significantly higher prices. Interestingly the 2015 prices are generally higher than both 2018 and 2019. The 2022 price averages follow the same pattern but far with significantly higher averages. This is likely due to the different data collection methods and incentives using separate data subsets for machine learning.

RRP has much more steady averages with less variation per year. The outliers identified on the boxplots cause much more variation in the secondary market price of sets, although there are interesting fluctuations in secondary price depending on the year released. For example the years 1998 and 2013 have roughly equal average prices between RRP and the other variables, unlike 1992 and 2007 which have peaks in average secondary price where there is no significant difference in RRP of those years.

Chart:



Price vs Numeric and **Price vs Categorical** show that numeric variables (*pieces*, *minifigs*, *height*, *width*, *depth*, *weight*, *rating*, *reviewcount*, and *age*) and categorical variables (*theme*, *subtheme*, *themegroup*, *licensed*, *category*, *packagingtype*, *availability*) are only fully present for *RRP*. Other price variables have limited characteristic data (*pieces*, *minifigs*, *age*, *theme*, *subtheme*, and *licensed*).

Every variable except *reviewcount* has a positive correlation with *RRP*. Boxplots of *theme* and *themegroup* show significant variation in pricing, for instance *themeGroup* 'Educational' having a lower quartile higher than the majority of other variables higher quartile. The most significant *packagingtype* is 'Box', and *availability* is 'Educational'.

The three price variables from **D15** and **D18** have positive correlations with *pieces*, *minifigs*, and *age*. Not unexpectedly, they also have similar small variation of *theme*, *subtheme*, and *licensed*. The price variable from **MP** has different characteristic data, but also variation less significant than *RRP*.

There are no significant differences in all price variables depending on a sets licensing group. The highly skewed data creates compact boxplots that are difficult to interpret.

Clusters

The two subsets of data used for machine learning can first be clustered. "Two simple ways in determining the (optimal) number (of) clusters are the Elbow method and Silhouette method" (SAPUTRA et al., 2020). Both methods indicate the optimal number of clusters to be two to three. K-means clustering is appropriate since "K-means is the simplest clustering algorithm in a partitional-based clustering where each cluster will be completely separated" (SAPUTRA et al., 2020).

The 2019 subset clusters on four price variables (*RRP*, *price2015*, *price2018*, and *price2019*), whilst the 2022 subset clusters on one variable (*price2022*). Clustering this data causes significant class inbalance, lessening the effectiveness of the categorical variable (*licensed* and *theme*) visualisations.

The 2019 subset with two clusters has 189 observations in *cluster 1* and 2088 in *cluster 2*. There is a fairly natural split of the data seen on the bimodal plots of price variables against themselves. There are significant differences in all the numeric variables except *yearReleased* and *age*. However, the categorical variables show similar distributions across clusters.

The 2019 subset with three clusters has 19 observations in *cluster 1*, 1949 in *cluster 2*, and 309 in *cluster 3*. The split of the data is less natural with *cluster 1* consisting of the tail end of most extreme price values. The differences in all the numeric variables remain significant. Variables *yearReleased* and *age* have more noticeable differences. The categorical variables again show similar distributions across clusters.

The 2022 subset with two clusters has 3102 observations in *cluster 1* and 139 in *cluster 2*. There are significant differences in numeric variables and similar categorical variable distributions across clusters.

The 2022 subset with three clusters has 3019 observations in *cluster 1*, 212 in *cluster 2*, and 10 in *cluster 3*. This class inbalance shows clustering of the most extreme price values. The variable differences are consistent with previous.

There is clear evidence that the subset data may be suited to clustering. Due to the significant data skew, clustering into two groups captures the organic split between expensive complicated sets and cheaper regular sets. Clustering into three groups causes more significant class inbalance and is likely more artificial than organic.

Cluster groups can be used in machine learning to test whether they play a significant role in predicting prices, via feature selection ranking.

Machine Learning Models

The best performing models are those with the highest R^2 and lowest $RMSE$ and MAE. There are 14 features for predicting *price2019* and 9 features for predicting *price2022*.

The trained models for predicting *price2019* all perform better than the null model. However, there is no consistently best model. SVM has the highest R^2 and lowest $RMSE$ whilst RF has lowest MAE. EN and LR have near identical metric distributions. EN also has consistent performance on the test data, whereas RF and SVM improve R^2 and $RMSE$ but worsen MAE scores.

Table 1: Average 20219Price Prediction Metrics

Model	R2_Train	RMSE_Train	MAE_Train	R2_Test	RMSE_Test	MAE_Test
EN	0.16	0.97	0.07	0.17	0.97	0.08
LR	0.16	0.97	0.07	NA	NA	NA
Null	0.00	0.99	0.56	NA	NA	NA
RF	0.08	0.99	0.02	0.26	0.94	0.09
SVM	0.21	0.96	0.09	0.31	0.92	0.11

EN has low values of λ which reduces the EN model to an OLS regression. α is predominantly 0.75 or 1, meaning that a Lasso regression is preferable to a Ridge regression for this data. The optimal *mtry* hyperparameter for RF is 6 across all bootstraps. This is slightly higher than the default 1/3 of features, causing potentially reduced bias but increased variance. The SVM models also have consistent hyperparameters across all bootstraps. C is 2.0 while σ is consistently close to 0 meaning the models potentially overfit the training data.

EN ranks the top six features as *price2018*, *price2015*, *RRP*, *cluster*, *pieces*, and *yearReleased*, with *price2018* by far the most significant feature. The next three have similar importance. RF has similar top 6 features *price2018*, *cluster*, *price2015*, *pieces*, *RRP*, and *yearReleased*. *price2018* is still by the most significant feature, however each subsequent feature is relatively much more significant than those from EN. Additionally, the features *age* and *minifigs* have more significance than in EN.

licensed, and *theme* have minimal significance as predictors of *price2019*.

The trained models for predicting *price2022* all perform better than the null model. Again, there is no consistently best model. EN has the highest R^2 and lowest $RMSE$ whilst RF has the lowest MAE. All three models improve their R^2 and $RMSE$ scores, whilst EN has consistent MAE and the others have worse MAE.

Table 2: Average 2022 Price Prediction Metrics

Model	R2_Train	RMSE_Train	MAE_Train	R2_Test	RMSE_Test	MAE_Test
EN	0.54	0.71	0.25	0.62	0.62	0.25
LR	0.54	0.71	0.25	NA	NA	NA
Null	0.00	0.99	0.53	NA	NA	NA
RF	0.18	0.97	0.08	0.55	0.71	0.23
SVM	0.43	0.83	0.17	0.60	0.64	0.21

EN has low values of λ which reduces the EN model to an OLS regression. α is predominantly 0.25 or 0, meaning that a Ridge regression is preferable to a Lasso regression for this data. The optimal *mtry* hyperparameter for RF is 5 or 6 across all bootstraps. This is slightly higher than the default 1/3 of features, causing potentially reduced bias but increased variance. The SVM models also have consistent hyperparameters across all bootstraps. C is 2.0 while σ is consistently close to 0.25.

Both EN and RF rank the top three features as *cluster*, *pieces*, and *yearReleased*.

Features *licensed*, and *theme* have minimal significance as predictors of either *price2019* or *price2022*.

Conclusions

Analysis of **Categories** and **Dimensions** shows that Lego sets are increasing in quantity and complexity over time.

Analysis of **Prices** shows that increased complexity in Lego sets tends to increase the price. Additionally, price distributions were fairly equal, regardless of the Lego theme licensing.

Analysis of **Clusters** shows that numeric variables show significant differences from clustering but categorical variables remain homogeneous. This indicates minimal effects from licensing since it does not significantly change depending on price.

Analysis of **Machine Learning Models** shows that historical price variables are very significant predictors of current prices. However, the models predicting 2022 prices had much better metrics than the models predicting 2019 prices. This means the historical price variables may have added noise to the data causing less confident predictions. Licensing had very little significance as a predictor of price.

Both research questions can be confidently answered.

It is entirely possible to produce machine learning models that accurately predict Lego. All models tested in this project outperformed the null model. The degree of accuracy achieved can be improved with hyperparameter tuning. Future work could also improve accuracy by exploring which type of algorithms perform best on Lego data.

The effects from Lego set licensing on pricing and model performance are minimal. There was no significant difference produced from licensing, meaning Lego original themes can compete with established intellectual properties. This is a testament to Legos global popularity and longevity.

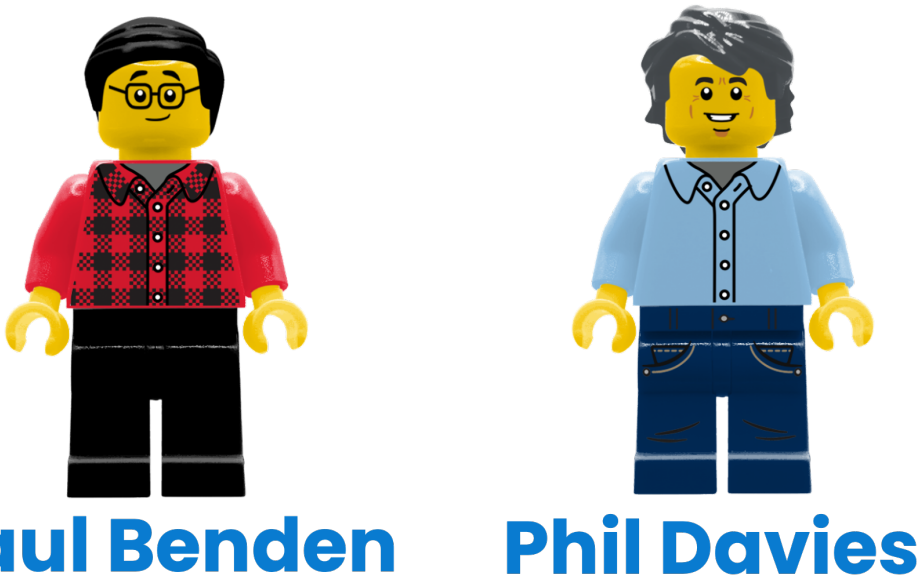
Future work

This project predominantly taught me an understanding of the comprehensive process of machine learning using real world data. There is much room for improvement and further investigations with this project. Larger, more comprehensive datasets could facilitate more reliable data analytics, at the cost of decreased processing speeds, especially when training models. The data used was heavily skewed, meaning transforming may create more normalised data which is preferable for analysis. Incorporating validation error would help identify any overfitting to the training data, which is more likely with smaller datasets.

An end goal of this methodology would be a tool that Lego consumers could use to predict prices of sets they own. This would require unsupervised learning, which is feasible as evidenced by the organic clustering potential of this projects data. A realised deployment of this idea is *BrickEconomy* (2023) which "employs statistical analysis and machine learning techniques to predict current and future prices of LEGO sets and minifigures. These predictions are based on analysing sales trends across various secondary markets".

Acknowledgements

Special thanks to my project supervisors, Paul Benden and Phil Davies, for their continued guidance throughout the project.



Thanks to the work produced by Bryer (2023), Dobrynskaya (2021), and mrpantherson (2023) for providing me with the data for this project.

Thanks to the University of Canterbury staff who facilitated my learning and completion of a Masters of Applied Data Science.

Finally, thanks to my family for providing encouragement and advice throughout my studies.

References

BrickEconomy: The economics of LEGO investing. (2023). https://www.brickeconomy.com.

Bryer, J. (2023). *Brickset: An r package to access LEGO set data from rebrickable*. https://rdrr.io/github/jbryer/brickset/f/README.md

Dobrynskaya, V. (2021). *LEGO secondary market price data* (Version 1). Mendeley Data. https://doi.org/10.17632/v9hhs66vm3.1

eBay Inc. (2020). *The LEGO® masters effect: 2 LEGO sets sold every second on eBay as show debuts*. https://www.ebayinc.com/stories/press-room/au/the-lego-masters-effect-2-lego-sets-sold-every-second-on-ebay-as-show-debuts/.

Federal Reserve Bank of Minneapolis. (2023). *Inflation calculator*. https://www.minneapolisfed.org/about-us/monetary-policy/inflation-calculator.

Kuhn, M. (2023). *Caret: Classification and regression training*. https://topepo.github.io/caret/.

Lauwaert, M. (2008). Playing outside the box – on LEGO toys and the changing world of construction play. *History and Technology*, *24*(3), 221–237. https://doi.org/10.1080/07341510801900300

Mahesh, B. (2019). *Machine learning algorithms -a review*. https://doi.org/10.21275/ART20203995

mrpantherson. (2023). *LEGO sets dataset*. Kaggle. https://www.kaggle.com/datasets/mrpantherson/lego-sets

OpenAI. (2023). *ChatGPT-4*. Software available from OpenAI. https://openai.com/

SAPUTRA, D. M., SAPUTRA, D., & OSWARI, L. D. (2020). Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method. *Proceedings of the Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*, 341–346. https://doi.org/10.2991/aisr.k.200424.051

The LEGO Group. (2023). *The LEGO group history*. https://www.lego.com/en-us/aboutus/lego-group/the-lego-group-history.

Unknown Author. (2012). *World war II statistics*. https://web.archive.org/web/20120306144404/http://www.milhist.dk/besattelsen/ww2stat/ww2stat.html.

Wikipedia contributors. (2023). *List of lego themes*. https://en.wikipedia.org/wiki/List_of_Lego_themes.

Appendices

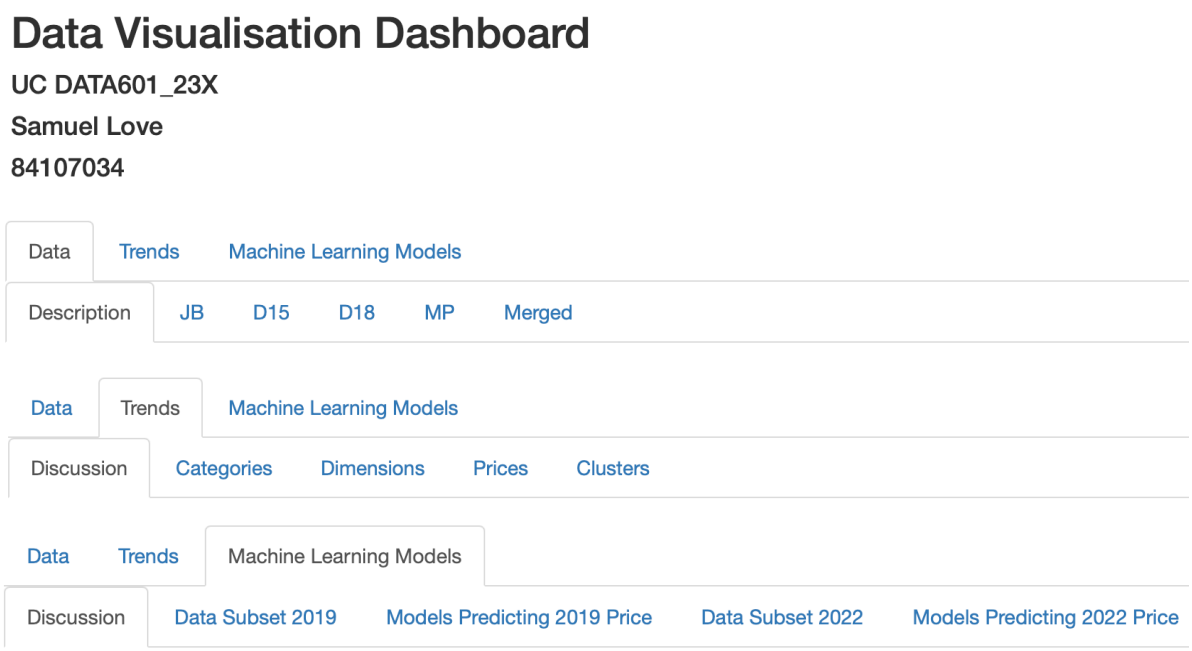
Files

There are four supplementary files. DATA601Data.qmd documents and executes code for processing the data and training and testing the models. global.R, server.R, and ui.R combine to generate the data visualisation dashboard. csv files are generated by DATA601.qmd and read in global.R.

Hosted dashboard link

Graphs

Dashboard



Tables

Table 3: JB Variables

Name	Type	Unique.Values	Description
setID	integer	19409	Unique identifier for every set
number	character	17997	Secondary identifier
numberVariant	integer	25	Secondary identifier variants
name	character	16196	Name of the set
year	integer	54	Year the set was released
theme	character	158	Theme that the set belongs to
themeGroup	character	16	Supertheme that the set belongs to
subtheme	character	956	Subtheme that the set belongs to
category	character	7	Category that the set belongs to (similar to the theme)
released	logical	2	Indicator for if the set was officially released or not
pieces	integer	1460	Number of pieces in the set
minifigs	integer	33	Number of minifigures in the set
bricksetURL	character	19409	Unique URL to the Brickset page for every set
rating	numeric	30	Average rating for the set (from 0 to 5)
reviewCount	integer	63	Number of reviews for the set
packagingType	character	19	What material the set was shipped in
availability	character	10	Where the set was sold
agerange_min	integer	16	Target age the set is intended for
US_retailPrice	numeric	153	Retail price of the set in the United States of America
US_dateFirstAvailable	Date	978	First date the set was available to purchase in the United States of America
US_dateLastAvailable	Date	2196	Last date the set was available to purchase in the United States of America
UK_retailPrice	numeric	225	Retail price of the set in the United Kingdom
UK_dateFirstAvailable	Date	926	First date the set was available to purchase in the United Kingdom
UK_dateLastAvailable	Date	2067	Last date the set was available to purchase in the United Kingdom
CA_retailPrice	numeric	176	Retail price of the set in Canada
CA_dateFirstAvailable	Date	744	First date the set was available to purchase in Canada
CA_dateLastAvailable	Date	1879	Last date the set was available to purchase in Canada
DE_retailPrice	numeric	172	Retail price of the set in Germany
DE_dateFirstAvailable	Date	513	First date the set was available to purchase in Germany
DE_dateLastAvailable	Date	1251	Last date the set was available to purchase in Germany
height	numeric	247	Height of the set's box in cm
width	numeric	289	Width of the set's box in cm
depth	numeric	281	Depth of the set's box in cm
weight	numeric	1106	Weight of the set's box in kg
thumbnailURL	character	18354	Unique URL to a thumbnail for sets
imageURL	character	18354	Unique URL to an image for sets

Table 4: D15 Variables

Name	Type	Unique.Values	Description
id	character	2322	Unique identifier for every set
theme	character	44	Theme that the set belongs to
name	character	2114	Name of the set
year.of.release	integer	31	Year the set was released
#.of.pieces	integer	808	Number of pieces in the set
#.of.minifigures	integer	17	Number of minifigures in the set
Secondary.market.prices.of.new.sets.in.2015	double	1970	Average second-hand price from Ebay for high quality sets
Secondary.market.prices.of.used.sets.in.2015	double	1447	Average second-hand price from Ebay for low quality sets
Primary.market.price.at.release	double	102	First-hand price at release
age	integer	31	Target age the set is intended for
Size.group.(1.-.Biggest;.4.-.Smallest)	integer	2322	Rough grouping of sets based on physical dimensions

Table 5: D18 Variables

Name	Type	Unique.Values	Description
id	character	2322	Unique identifier for every set
theme	character	44	Theme that the set belongs to
name	character	2114	Name of the set
year.of.release	integer	31	Year the set was released
#.of.pieces	integer	808	Number of pieces in the set
#.of.minifigures	integer	17	Number of minifigures in the set
2018.01.01	character	1529	Average price for the last 30 Ebay transactions in a month
2018.02.01	character	1495	Average price for the last 30 Ebay transactions in a month
2018.03.01	character	1504	Average price for the last 30 Ebay transactions in a month
2018.04.01	character	1464	Average price for the last 30 Ebay transactions in a month
2018.05.01	double	1431	Average price for the last 30 Ebay transactions in a month
2018.06.01	double	1433	Average price for the last 30 Ebay transactions in a month
2018.07.01	double	1439	Average price for the last 30 Ebay transactions in a month
2018.08.01	double	1439	Average price for the last 30 Ebay transactions in a month
2018.09.01	double	1458	Average price for the last 30 Ebay transactions in a month
2018.10.01	double	1416	Average price for the last 30 Ebay transactions in a month
2018.11.01	double	1440	Average price for the last 30 Ebay transactions in a month
2018.12.01	double	1487	Average price for the last 30 Ebay transactions in a month
2019.01.01	double	1444	Average price for the last 30 Ebay transactions in a month
2019.02.01	double	1448	Average price for the last 30 Ebay transactions in a month
2019.03.01	double	1438	Average price for the last 30 Ebay transactions in a month
2019.04.01	double	1436	Average price for the last 30 Ebay transactions in a month

Table 6: MP Variables

Name	Type	Unique.Values	Description
id	character	5075	Unique identifier for every set
name	character	4312	Name of the set
category	character	66	Theme and subtheme that the set belongs to
year	numeric	47	Year the set was released
parts	numeric	1154	Number of pieces in the set
img_link	character	5065	Unique URL to an image for sets
set_link	character	5075	Unique URL to the Brickset page for every set
raw_price	character	2564	Lists of raw prices from various sources
mean_price	numeric	2309	Average of the raw prices

Table 7: Merged Variables

Name	Type	Unique.Values	Description
theme	character	162	Theme of each set
setID	character	23212	Unique identifier for every set
name	character	17685	Name of each set
yearReleased	integer	55	Year each set was released
pieces	integer	1589	Number of pieces in each set
minifigs	integer	34	Number of minifigures in the set
RRP	double	191	Retail price of the set in the United States of America
age	integer	31	Intended user age
subtheme	character	968	Subtheme of each set
price2015	double	1969	Average yearly secondary price in 2015
price2018	double	1984	Average yearly secondary price in 2018
price2019	double	1937	Average yearly secondary price in 2019
price2022	double	2308	Average yearly secondary price in 2022
number	character	17997	Secondary identifier
numberVariant	integer	25	Secondary identifier variant
themeGroup	character	16	Supertheme of each set
category	character	7	Category of each set (similar to themegroup)
released	logical	2	Indicator if a set was officially released
rating	double	29	Average rating for each set (from 0 to 5)
reviewCount	integer	63	Number of reviews for each set
packagingType	character	18	What material each set was shipped in
availability	character	8	Where the set was sold
yearRetired	integer	18	Year each set was retired
height	double	247	Height of the set's box in cm
width	double	289	Width of the set's box in cm
depth	double	281	Depth of the set's box in cm
weight	double	1106	Weight of the set's box in kg
price2015Standard	double	1969	Average yearly secondary price in 2015 adjusted with inflation to 2022 prices
price2018Standard	double	1984	Average yearly secondary price in 2018 adjusted with inflation to 2022 prices
price2019Standard	double	1937	Average yearly secondary price in 2019 adjusted with inflation to 2022 prices
licensed	character	3	Wether the set is licensed from an IP