From Bricks to Bucks

Exploring Influential Variables in Sales Patterns of Lego Bricks

By Sam Love

Project Overview

The University of Canterbury School of Mathematics and Statistics is providing supervision for this project involving Lego, a global brand of plastic construction toys.

This project has two aims; to examine trends in Lego characteristics over time, and to train machine learning algorithms to produce predictive models of prices in 2019 and 2022.

The main research question is,

‘Can we accurately predict the secondary prices of Lego sets using a defined set of characteristics?’

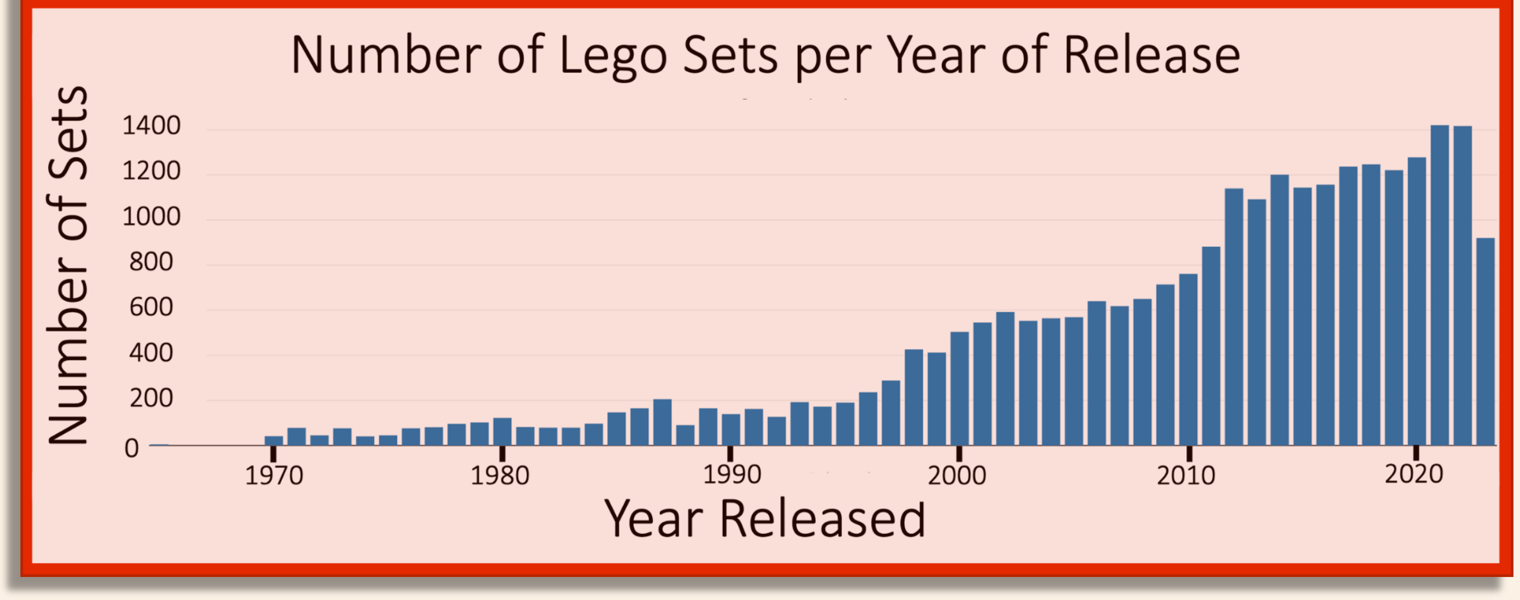
A sub-question is,

‘Do Lego sets based on licensed intellectual properties differ from Lego original sets?’

To achieve this goal, objectives include; Time Series Analysis of Lego characteristics, Data Visualisation to assist with data cleaning and display relevant statistics, Cluster Analysis to identify organic patterns in the data, Predictive Modeling of secondary market prices (2019 and 2022) via supervised learning, and Feature Ranking to determine the significance of each characteristic.

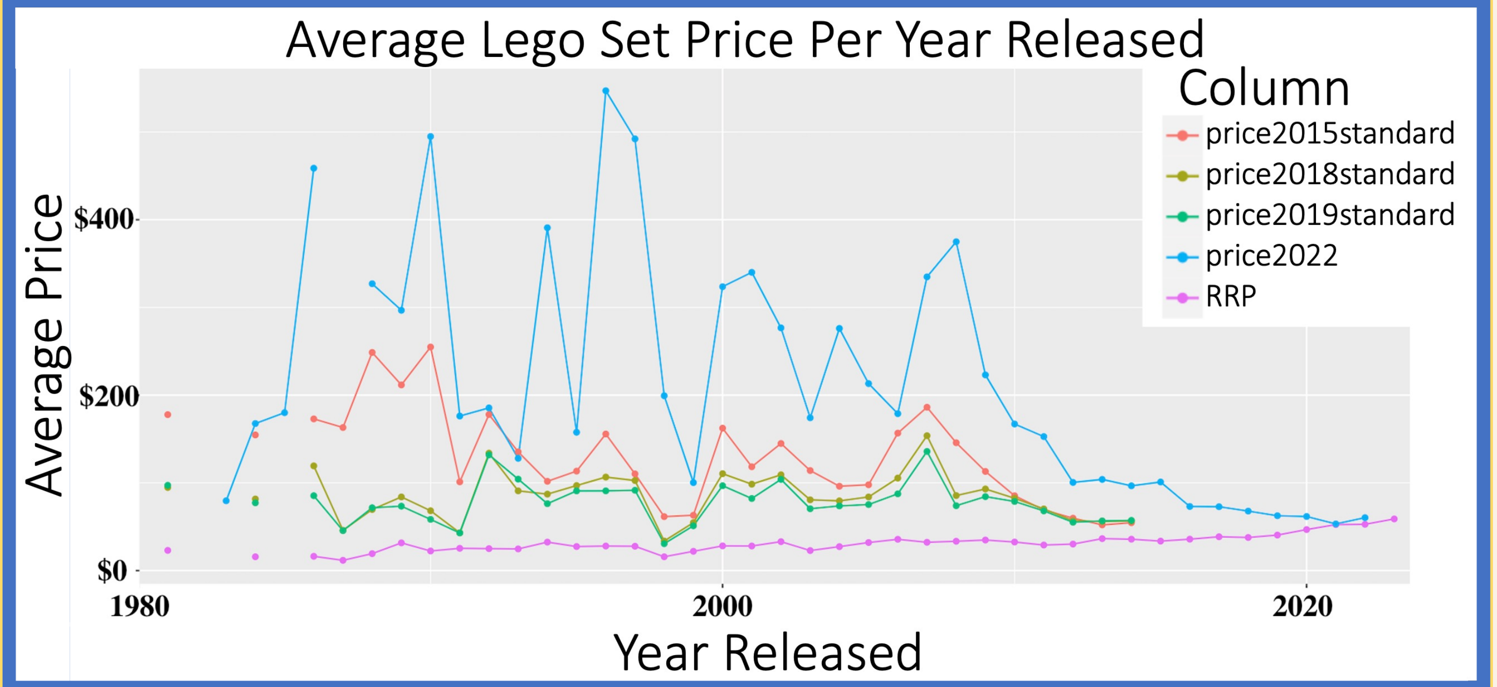
Data

The data for this project consists of four datasets retrieved as csv files from three sources. The main dataset (JB) has 19,409 **observations of 36 variables with missingness of 34.4%.** The main price dataset (D15) has 3,322 **observations of 11 variables with missingness of <0.1%.** The main price dataset (D18) has 3,322 **observations of 22 variables with missingness of 1.1%.** The supplementary price dataset (MP) has 5,075 **observations of 9 variables with 100% present data.**



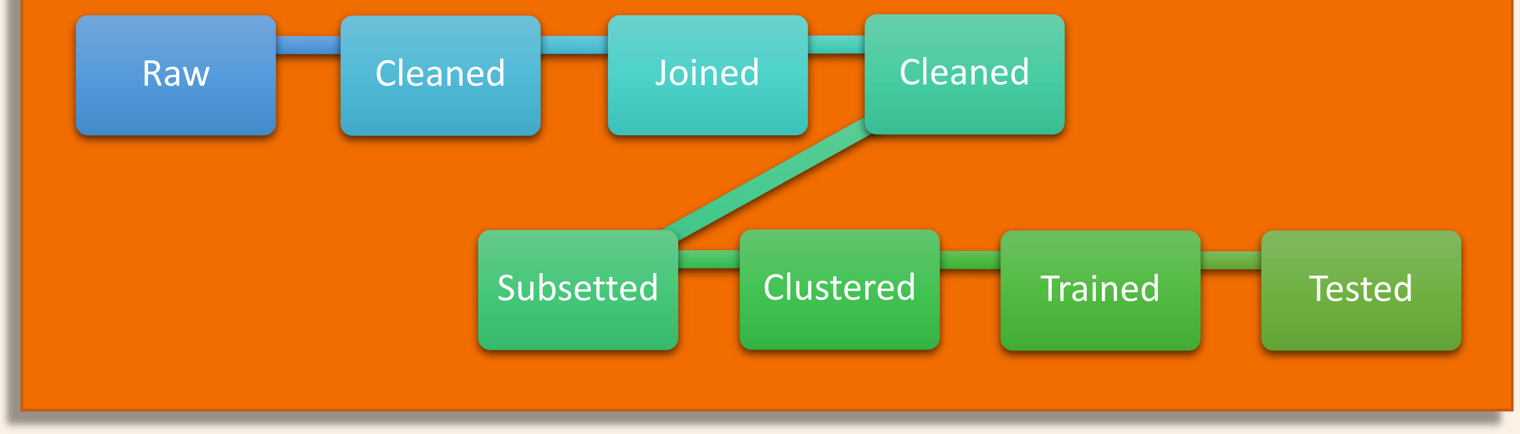
Results

Increasing numbers of sets released each year. Most variables have increasing diversity over time. Price variables positively correlate with increasing complexity of dimension variables. Best model based on performance metrics (MAE, R2, RMSE) varies. Price variables are the most significant predictors for 2019 price.



Methodology

Data cleaning involved manipulating and renaming variables to enable a successful join operation when creating the main dataset **Merged. Merged** has 26,324 observations of 31 variables with 50.4% missingness. Supplementary variables include prices adjusted for inflation and categorical theme licensing. **Visualisations were generated for each raw and cleaned dataset, as well as various variables present in Merged with an emphasis on Price variables. Subsets focusing on price2019 and price2022 with 100% present data were visualised and used for cluster analysis and machine learning. Five models (null, LR, EN, RF, SVM) were trained on 20 bootstraps of both subsets using a 75/25 train/test split.**



Discussion

Lego sets in the data increase in both quantity and complexity over time. Price variables on average have high variation over time. Price variables are the best predictors for 2019 prices, whilst the best predictors for 2022 prices (in the absence of other price variables) are cluster group, number of pieces, and year released. There is no obvious best performing model, meaning this data is tricky to predict. Whether sets are from licensed intellectual properties or not, does not show any significant differences in either characteristics, price, or price predictions. Brickeconomy.com is an excellent example of an extended application of price prediction models.

DATA601 2023
Supervisor Phil Davies & Paul Benden