The Evolution of Complexity in LEGO Sets

Edward Beach [*1] **Patricia Schlegel** [*2]

Abstract

This study aims to determine if the production complexity of LEGO sets has been increasing since the beginning of LEGO production. Using statistical methods such as regression models and cluster analysis, our analysis reveals a consistent upward trend in set complexity, with a notable acceleration observed in recent years. Our findings provide insights into the evolving nature of LEGO set designs.

1. Introduction

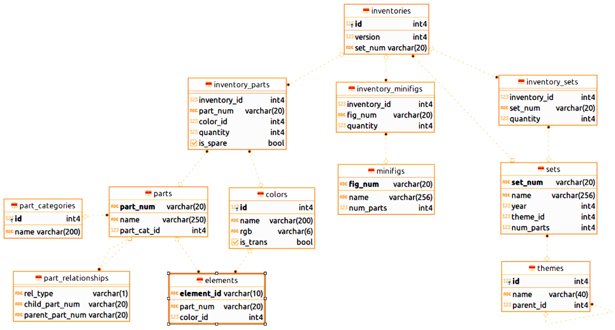
As LEGO enthusiasts and researchers, we wanted to work with a LEGO dataset. **?** observed that different features of LEGO, such as set size, number of colors, and minifigures, have consistently increased over the years. We started to wonder, if we could quantify the complexity of LEGO sets in terms of production, not in terms of the intricacy of production processes, but in the diversity and richness of the final products and say that it increased over the years.

LEGO sets are culturally important, known and enjoyed by many children in our society. Some individuals even continue to engage with LEGO into adulthood, collecting sets for fun or investment (Dobrynskaya & Kishilova, 2018). In that way, the LEGO company has expanded to become the world's largest toy manufacturer. A significant portion of this success is attributed to the company's strategic licensing agreements, particularly with influential transmedia franchises like Star Wars and Harry Potter (Mazzarella & Hains, 2019). Understanding evolution of production complexity adds valuable insights into the evolving dynamics of entertainment and creativity, as a higher production complexity likely results in the creation of more complicated LEGO sets in terms of building.

In the following sections, we detail the methods and data used for our complexity analysis. This involves a comprehensive exploration of the Rebrickable dataset, encompassing details on LEGO sets released between 1949 and 2024. We describe the formulation of the complexity metric, its normalization, and the subsequent linear and exponential regressions modeling for predicting mean complexity per year until 2030. Additionally, a k-means clustering analysis is detailed, outlining the optimal number of clusters based on the complexity metric. Our findings consistently indicated a noticeable upward trend in set complexity over the years.

2. Data and Methods

Our project used the Rebrickable dataset, spanning LEGO Sets from 1949 to 2024. The dataset is organized into several CSV files (see Figure 1) that contain details on 17.077 sets, 35.408 parts, 13.546 minifigures, 144 themes, 251 colors, and 66 part categories. The data can also be found in our git repository, which contains the code and details on our data preprocessing, analysis, and plots. Data preprocessing and data analysis were performed using Python version 3.9.5 and regression and cluster analysis in the library sklearn version 1.3.2.



Imagesource: Inverted image from
https://rebrickable.com/downloads/

Figure 1. Structure of the Rebrickable dataset

The initial data processing involved merging several classes into two comprehensive datasets – one focused on parts in different sets and there characteristics and the other on minifigures in different sets. We filtered the LEGO themes for the main themes to prevent the inclusion of every minor

[*]Equal contribution [1]Matrikelnummer 5451904, edward.beach@student.uni-tuebingen.de, MSc Cognitive Science [2]Matrikelnummer 5480232, patricia.schlegel@student.uni-tuebingen.de, MSc Cognitive Science.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2023/24 (Module ML4201). Style template based on the ICML style files 2023.

or niche theme, ensuring that our datasets primarily encapsulate the broader and more significant themes within the LEGO catalog. To maintain relevance, the datasets were filtered to include only data up to the year 2023, ensuring the incorporation of completed years. While examining the dataset, we noted a considerable fluctuation in the number of sets in the database during the early years, specifically in instances where the set count was below ten (1949, 1950, 1953, 1959, 1960), and there were even years with no releases (1951 and 1952). This observed volatility had the potential to significantly influence the complexity score, introducing fluctuations in the data. Notably, this trend stabilizes in the later years, where after 1960, there is a consistent release of 20 sets or more, surpassing 50 after 1975, and exceeding 100 after 1988. To ensure a more reliable analysis, we excluded years before 1961, aiming to avoid the impact of early-year volatility on the complexity assessment.

The final step involved creating a consolidated dataset by grouping data using the set number. The grouped dataset included pertinent details for each set such as release year, theme, total number of part and number of different parts, minifigure quantities, color diversity, category variety, counts of unique parts (quantity of one in the set), and the proportion of unique to not unique parts within each set.

After data preparation, we started to explore the data to discern trends in LEGO set features over the years, with the goal of replicating previous findings (**?**). For each year we plotted the number of released sets, mean part count per set, mean different parts per set, mean minifigures per set, themes per year, mean part categories per set, mean colors per set, the number of different colors, and mean unique parts per set. Additionally, we employed a logarithmic transformation on the y-axis to identify any upward or downward trends (see Figure 2). We can see that all the numbers are increasing over the years, which replicated the findings of **?** and led us to suspect that the complexity LEGO sets, in terms of production, could be increasing as well.

The following deeper exploration involved analyzing the mean proportion of unique parts to non-unique parts per set per year. Additionally, the ten most used themes between 2000 and 2023 were identified using different criteria (number of sets, number of parts, most different parts) and furthermore determined the most used ten colors in those themes. Predictions for these metrics and the number of released sets and mean part count per set until 2030 were made using linear regression. These analyses can be found in the exploratory analysis file on github.

After that we got an intuition about the dataset and moved to the main analysis. We introduce a complexity metric for each set. Because we want to examine the complexity in terms of production complexity and diversity of set features, we used factors such as the total number of parts and number of different parts, the different part categories, the number of unique parts, the number different colors, and number of minifigures. This complexity metric aimed to quantify the intricacy of production, considering the varied materials/colors, number of parts that have to be produced and different manufacturing processes required. We therefore propose the following formula for the production complexity:

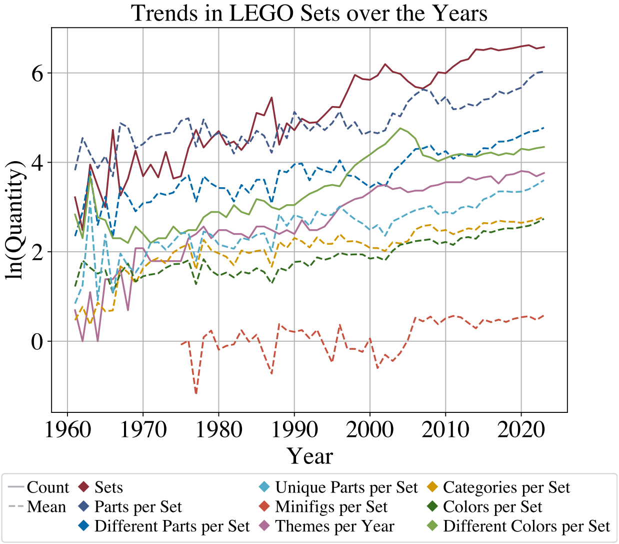


Figure 2. Trends of different features of LEGO sets over the years

$$Comp = Parts + Diff + Cat + Uniq + Col + Figs$$

where

$$\begin{aligned}
Comp &: \text{Complexity}\\
Parts &: \text{Total number of parts}\\
Diff &: \text{Number of different parts}\\
Cat &: \text{Number of different categories of the parts}\\
Uniq &: \text{Number of unique parts}\\
Col &: \text{Number of different colors}\\
Figs &: \text{Number of minifigures}
\end{aligned}$$

The complexity metric was then normalized to a range between 0 and 1. This normalization enhances interpretability, where a score of 1 represents the highest complexity and 0 the lowest. Normalization was achieved by applying the formula:

$$Complexity = \frac{Comp - MinComp}{MaxComp - MinComp}$$

where

$Complexity$: Normalized complexity
$Comp$: Complexity
$MinComp$: Minimal complexity across all sets
$MaxComp$: Maximal complexity across all sets

Subsequently, a linear and exponential regression was employed to model the mean complexity per year, providing predictions until 2030. We visually compared the two regression lines and assessed their fit through the examination of the Sum of Squared Residuals (SSR) and Coefficient of Determination (R^2) to find the regression that better fits the data. Regression is a valuable method to find trends in data as it allows for the identification of upward or downward trajectories, providing a mathematical model that captures the relationship between variables. We additionally checked the regressions for overfitting or underfitting using cross-validation, which is a good and simple approach to identify such issues in regression models (Emmert-Streib & Dehmer, 2019).
Additionally, a k-means clustering analysis, guided by the elbow method, was conducted to determine the optimal number of clusters based on the complexity metric.

3. Results

For our analysis, we performed a linear regression ($SSR = 0.002$, $R^2 = 0.724$) and exponential regression ($SSR = 0.002$, $R^2 = 0.794$) for the average complexity per set per year with a prediction until 2030 (see Figure 3). In the cross-validation of the two models we observed that both models show nearly the same mean of the Root Mean Square Error (RMSE) on the test set compared to the training set (linear regression: $RMSE_{training} = 0.005$, $RMSE_{test} = 0.005$; exponential regression: $RMSE_{training} = 0.004$, $RMSE_{test} = 0.003$). Further visualisation, like the plotted distribution of the complexity score, can be found in the regression file on github. We then performed a k-means clustering, that aimed to group the data based on complexity with the elbow method, which set the optimal number of clusters to three. Afterwards we compared the average complexity scores in the different clusters (Cluster 0: 0.009, Cluster 1: 0.06 , Cluster 2: 0.273) and the proportional number of different features in the clusters (see Figure 4), set the complexity level of the clusters from low to high (Cluster 0: Low, Cluster 1: Medium, Cluster 2: High) and visualized the distribution of sets across different clusters (see Figure 5). Further visualisation, like the top ten themes in the different complexity clusters, can be found in the clustering file on github.

4. Discussion & Conclusion

After data preparation, we started to explore the data to discern trends in LEGO set features over the years. We can see that all features of LEGO are increasing over the years (see Figure 2), which led us to suspect that the complexity LEGO sets, in terms of production diversity, could be increasing as well.

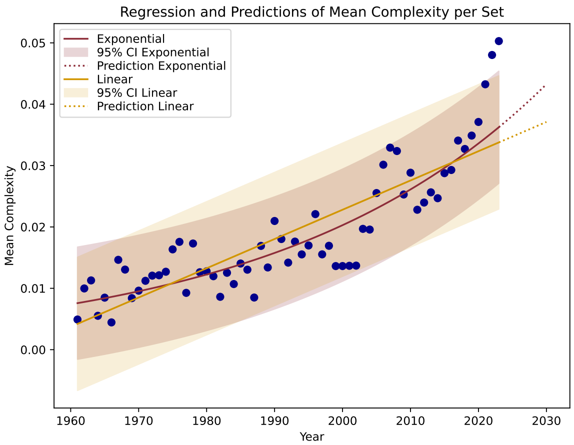


Figure 3. Exponential and linear regression of the mean complexity for a LEGO set per year with predictions until 2030

To test this hypothesis, we started to introduce a complexity measure in terms of production and calculated it for each set. We then calculated a linear and exponential regression for the mean complexity for a set. We can see for both regressions that the mean complexity is increasing and that the prediction of complexity is as well. We can also see that the exponential regression provides a better fit to the data (see Figure 3), this is supported by a lower SSR and a higher R^2 value. We furthermore performed a cross-validation to check for overfitting or underfitting. The results suggest no overfitting in both cases, as the models performed nearly the same on the training data than on unseen test data. The exponential regression model exhibits a little bit lower RMSE values, suggesting superior learning and a better fit to the data. As a result of these considerations, we conclude that the complexity has experienced exponential growth.
In the cluster analysis of the proportion of sets in the different complexity levels, we can see similar findings. The complexity started really low but increased over the years. More and more sets started to be in the clusters with a medium and high complexity, while the number of sets with low complexity are decreasing. Regression and cluster analysis both lead us to the assumption that in recent years, there is a noticeable increase in sets with a high complexity, contributing to an overall upward trend in complexity.
A possible limitation to those findings are that the overall number of sets is increasing over the years as well. Another noteworthy aspect is the impact of the number of released

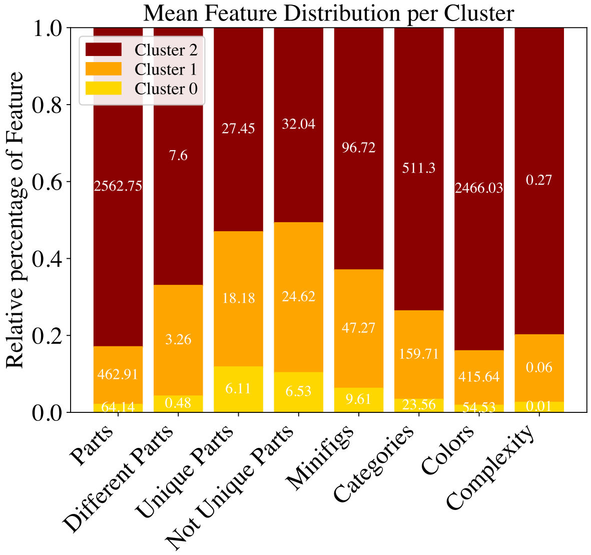


Figure 4. Mean proportion of features in the different clusters

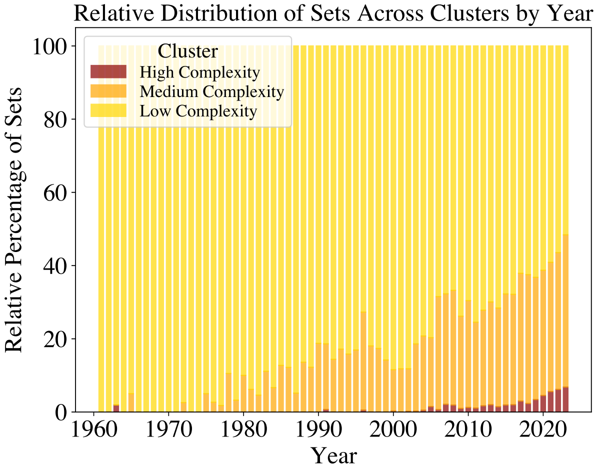


Figure 5. Proportion of sets in the different clusters

sets on the mean complexity in a specific year, particularly evident in earlier years where only a few sets were released. To address this, we made the decision to exclude data before 1961, ensuring that our dataset only includes years with twenty or more released sets. This deliberate exclusion, coupled with the subsequent stabilization in set releases, our data remains interpretable. We especially observe an upward trend in complexity in the recent years (see Figure 3 and Figure 4) where the number of sets and with that the complexity scores already stabilized.

We conclude that the production complexity of LEGO sets experiences exponential growth, characterized by the release of increasingly complex sets and a decrease in simpler ones.

Contribution Statement

Patricia Schlegel prepared the data, calculated the complexity for each LEGO set, made first exploratory analyses and the two regressions of the complexity. Edward Beach performed the k-means clustering and revised the plots for the report. All authors jointly wrote the text of the report.

References

Dobrynskaya, V. and Kishilova, J. Lego-the toy of smart investors. *Available at SSRN 3291456*, 2018.

Emmert-Streib, F. and Dehmer, M. Evaluation of regression models: Model assessment, model selection and generalization error. *Machine learning and knowledge extraction*, 1(1):521–551, 2019.

Mazzarella, S. R. and Hains, R. C. “let there be lego!”: An introduction to cultural studies of lego. *Cultural Studies of LEGO: More than Just Bricks*, pp. 1–20, 2019.