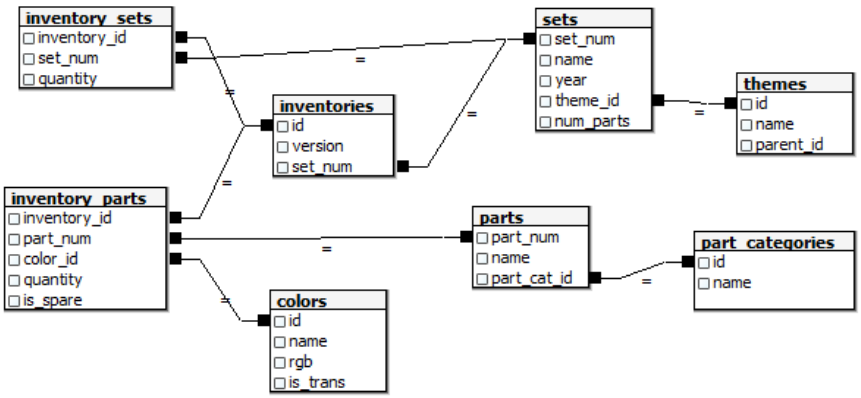
Lego Data Exploratory Data Analysis
Project 2 Proposal

Michael Hurth, Didi Dunn, Uthman Alibalogun
https://github.com/UC-Berkeley-I-School/lego_eda.git

The primary dataset we plan to analyze is the lego dataset available on kaggle at https://www.kaggle.com/datasets/rtatman/lego-database?resource=download. This data exists in a schema that consists of 8 tables. The tables contain information about lego sets and their part inventories, colors, categories, and themes. An imdb dataset with movie titles, genres, and release dates will be used in conjunction with the lego theme data to investigate the relationships between movie releases and corresponding lego set releases. The imdb data set is available at https://www.imdb.com/interfaces/. The imdb dataset will be joined to the lego set using regex expressions to link the movie titles.

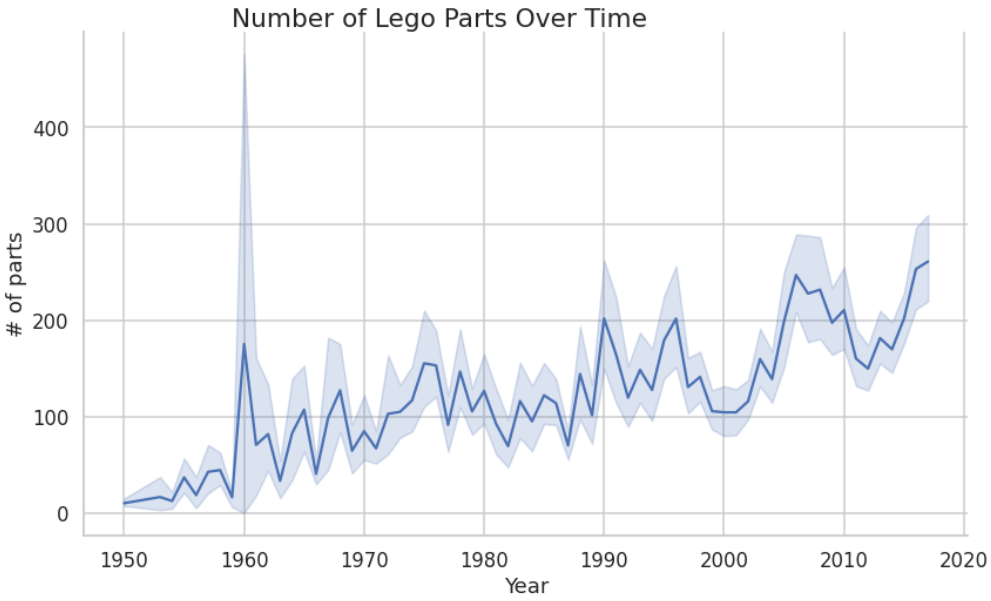


Initial plots, figures, or tables

- Shape of the lego dataset

DataFrames	Source	Row Count	Column Count
colors	lego	135	4
parts	lego	25993	3
sets	lego	11673	5
themes	lego	614	3
inventories	lego	11681	3
inventory_parts	lego	580251	5
inventory_sets	lego	2846	3
part_categories	lego	57	2
imdb_titles	imdb	9717486	9

- Years covered? The dataset ranges from 1950 to 2017.



Initial Exploratory Data Analysis Questions

- How have the use of colors changed over time (Total pieces, Fractions)?
- How have part counts per set changed over time?
- What are the most popular themes of all time and by year?
- How are the themes related to movie releases
- How have the uses of spare parts changed through time and how are they related to different sets
- What are the largest and smallest sets available?
- What are the genres of movies have the most themes (e.g. Family, Action, Horror)?
- Does of the popularity of a movie franchise correlate with the popularity of a lego theme?
- Do the themes have easily identifiable clusters of colors, piece types, piece counts, and price?

Some of the variables (column names) you intend to explore and what kind of insights you expect to glean

We are interested in exploring how lego sets have changed over time and if that there are any interesting findings. Some of the key fields we'd like to explore include the following.

Key Fields:

- colors
- themes

- quantity
- is_spare
- set:name
- year
- part:name
- theme:name
- num_parts

For example, on the num_parts (number of parts) field, we did a little bit of initial analysis and saw that in the 1950s, the number of parts in a given lego set was relatively small and the trend is that the number of pieces in a set has increased over the years. However, from the plot, we can also see that in certain decades, such as the 1960s, the spread of the number of parts in any given set was really large with some sets having upwards of 400 parts and some sets having far fewer parts. While this is just some preliminary analysis, we plan to explore this data in further detail, looking at other shifts over the years and trying to connect this to the IMDB dataset.

When tying in the IMDB dataset, we might look at how certain major movie titles might have impacted lego design in a given time period. For example, how might the original Star Wars trilogy in the late 1970s and early 1980s have impacted lego set designs? Were there new themes, colors, and sets introduced in that timeframe that correspond with the Star Wars trilogy? How have other major movie releases over from 1950 to the present impacted Lego set designs?

What we plan to cover in the final report and how we plan to organize it

In our final report, we will start by introducing the lego dataset and its schema. In addition, we plan to also explore the IMDB dataset and see if the lego datasets correspond to particular movie releases, so we will also introduce the IMDB dataset and its schema. We plan to document our decisions throughout our analysis and use what we're learning in class about descriptive statistics to summarize and support our conclusions about the data.

To begin, we will provide an overview of the dataset, including its structure and contents. We will then present the results of our exploratory data analysis, answering the initial questions we posed and highlighting any interesting trends or patterns we identified. Next, we will dive deeper into the relationship between Lego set releases and corresponding movie releases, examining the impact of major movie titles on Lego set designs over time. We will present our findings in a clear and organized manner, using visualizations and descriptive statistics to help illustrate our points. Finally, we will conclude with a summary of our key findings and recommendations for future research. We will discuss the implications of our analysis and highlight any areas where further investigation would be beneficial.