Link to Tableau Storyboard: HERE
Link to Code HTML: Attached Separately

LEGO: An Analysis of How Sets are Valued

Explanation of Data, Methods, Artistic Decisions, Visualizations:

The Dataset

- "Sets.csv" is a free-to-use dataset from Kaggle
    - "Lego Sets and Prices Over Time" by Alex Racape (from 2023)
- Each row corresponds to an individual set
- Columns include Set Name, Set ID, Release Year, LEGO Theme & Subtheme, Packaging Type, Number of Instructions, Number of Minifigs, Number of Pieces, Retail Availability, Rating (out of 5), Retail Price in USD, Number Owned, Quantity, Current Price in USD
- Total Quantity = number of times set was sold on Bricklink in last 6 months
- Current Price = most recent price the set was sold for on Bricklink
- NOTE: This dataset is very limited, and certain metrics are less valid than preferred (more to come about this in later analysis)

Filtering

- Filtered Data by Release Year later than 1999
    - LEGO Star Wars began in 1999, and given its prominence as the top-selling theme, the most accurate analysis relevant to today would begin at that point
- For this project, rows with missing values for price or piece count were removed
    - Left with only 2222 sets, ranging from 2005-2023
- Columns renamed for convenience and clarity

New Metrics

- Price Growth Rate
    - The percent change in price of the set between Resale Price and List (Retail) Price
- numYears
    - Change in price of sets will usually be larger if the set has been retired for longer
    - numYears value given by 2023 (expected year of latest Bricklink sale) - Release Year
    - Useful for normalizing Price Growth Rate to account for time

- Minifig/Price
    - The ratio of the number of minifigures in a set to its retail price
- Price/Piece
    - The ratio of the number of pieces in a set to its retail price

External Data

- CPI Data
    - Obtained via Fred St. Louis
    - Downloaded tabulated data of CPI Index for all Urban Consumers: All Items in US City Average
    - Converted to datetime→Year format
    - Year-Over-Year Inflation calculated by percent change in CPI for each pair of consecutive years between 2005-2023
    - Average Annual Inflation found by averaging Year-Over-Year Inflations for the 18 years (~2.52%)
    - NOTE: CPI for each year taken on January 1st
        - The period from January 1, 2022 to January 1, 2023 more closely represents what happened in 2022
        - 2023-2024 data not available yet on Fred St. Louis
        - Year shifted by 1, have data for 17 of the 18 year cycles
    - Merged into sets.csv by 'Year' column
- S&P 500 Data
    - Obtained via Macrotrends.net
    - Using 'Year Close' from 2004 as present value, 'Year Close' from 2023 as future value
        - 'Year Close' 2004 is closest value to start of 2005
    - Determined average growth rate r using present→future value conversion (accounting for compounded growth over time), where PV=Year Close 2004, FV=Year Close 2023, n years=18
    - r~7.91%
- Dow Jones (DJ) Data
    - Obtained via Macrotrends.net
    - Using 'Year Close' from 2004 as present value, 'Year Close' from 2023 as future value
        - 'Year Close' 2004 is closest value to start of 2005
    - Determined average growth rate r using present→future value conversion, where PV=Year Close 2004, FV=Year Close 2023, n years=18
    - r~7.20%

Additional Cleaning

- Yearly Growth Rate
    - Computes present→future value conversion between Retail Price and Resale Price of LEGO set
    - For simplicity, sets that released in 2023 (our "current year") are assumed to not have increased in monetary value, especially if the set is still on store shelves
        - Yearly Growth Rate assigned to 0 for these cases
- Theme Growth Rate
    - Take the mean value of 'Yearly Growth Rate' for each theme (*Star Wars, City, Harry Potter*, etc…)

Visualization

*See link interactive visualizations at link at top of this document

Part 1:

- We compare a subset of LEGO themes and their average annual price growth rates (average growth rate in price among all LEGO sets belonging to given theme) to the average growth rates found for CPI, S&P 500, and Dow Jones indices
- Originally chose to view the LEGO themes with the highest average growth rate in price among sets
    - Decided instead to focus on themes with the most LEGO sets released during the given time interval (2005-2023), as investing in these themes more closely resembles the diversified, longer-term investing expected in a formal stock index fund
    - Additionally, themes such as *Minions: The Rise of Gru* possess higher increases in value per set because so few sets were actually released. Due to smaller supply, these naturally become more valuable per year on average, but may not be a stable investment since there is little variety of sets relative to themes like *Star Wars*
- NOTE: The limited data set is a factor here; had missing values not been removed, we'd expect to see themes like Harry Potter with highest number of sets released
    - The Resale Price column from the original Kaggle dataset indicates the most recent price a LEGO set was sold for on Bricklink at the time of the dataset's creation
        - **This is NOT an average resale price**
        - The most recently sold set could have been overvalued or undervalued

        - This is a potential explanation for the negative average price growth rate of certain themes in the bar chart
        - More complete data would be useful in improving this visualization's accuracy
- Of the LEGO themes displayed, only *Star Wars* beats the index funds in annual average growth rate, and only by small margins
    - In general, stock index funds are more diversified and less volatile than LEGO markets alone, so investing in LEGO should be recommended for those who are truly passionate and knowledgeable about that specific space

Part 2:

- We look for correlations between Minifig/Price Ratio and Price/Piece Ratio with the number of sets owned for each set in the dataset
- Both trendlines have negligible R-Squared values, implying that these metrics are only weakly related to how much a LEGO set is bought
- Although the relationships are still weaker than expected, the results are not surprising—for most LEGO customers, the price per piece and number minifigures per price are not significant factors in the decision of whether or not to buy a set
    - That being said, the discrete number of minifigures and pieces in a set are likely to be stronger determinants for consumer behavior, as minifigures appreciate more than the LEGO build itself and largest piece counts are often assumed to correspond to more complicated building experiences

Part 3:

- Instead of studying the impacts of ratios, we pivot to the raw number of minifigures and pieces in a LEGO set, and how those values are related to the retail price (List Price).
- A much stronger relationship is observed in these cases. List Price is positively correlated with both number of minifigures in a set and number of pieces in a set, but more so with the number of pieces (R-Squared=0.875).
- These results are expected; sets with more individual pieces would require marginally higher costs to produce, and thus would be charged at higher prices
    - Unique character molds or accessories would also lead to higher production costs, especially if those parts are only used for one individual set

Part 4:

- We track general trends across LEGO sets of all themes from 2005-2023.

- The color scheme of the curve is associated with the average list price of LEGO sets for that year
- Avg. Minifig/Price:
    - On average, this value has decreased over time, especially as average retail prices have risen
    - The higher list price could be result of increased consumer demand, inflated piece counts, and general inflation
    - Although the net change is negative, the exact trend likely differs by theme (ex: LEGO Star Wars sets including less minifigures per the price than LEGO Harry Potter sets)
- Avg. Price/Piece
    - There has been dramatic fluctuation, yet this value in 2023 has returned to roughly the same as it was in 2005
- Avg. Rating
    - The average rating of LEGO sets (measured out of 5 stars) has decreased significantly in recent years (2019–) after remaining steady beforehand
    - This metric could be subject to response bias:
        - Most LEGO consumers do not rate sets online
        - People are more likely to leave a review for a product they are unhappy with
        - Some proportion of LEGO consumers buy sets no matter what, and many do not read the ratings prior to making purchase decisions
- Avg. Number of Minifigures
    - The average number of figures in a LEGO set was down for over a decade before slightly rising again in the last couple years
        - This could be partly due to the increase in the number of sets produced after 2004, when LEGO became increasingly more popular
        - The minifigures often accumulate the most monetary value relative to the rest of the LEGO set in the resale market, encouraging LEGO to include more figures and thus charge higher prices
        - Some consumers only buy sets for the minifigures
    - More and more digital television content is released yearly, especially with the increase in online streaming services
        - This has led to the introduction of many new pop culture characters that consumers look to buy

Part 5:

- The pie chart visualizes the total number of resold sets on Bricklink per LEGO theme during this time period

- See the interactive link at the top of this document to hover over pie slices or view specific LEGO themes from the side legend
- Allows for comparison of various LEGO themes and how they share the after market space
- As expected, the third-party demand is highest for LEGO Star Wars
- Again, due to limited data available, this does NOT paint the full picture, as Bricklink is generally geared towards Star Wars in particular and sites like eBay, Etsy, etc… provide other large markets for LEGO resale

Training a Model

- How much attention does LEGO pay to its sets once they retire and move through third-party markets?
    - If LEGO had a better idea regarding LEGO sets and their value after retirement, would that influence which sets are produced going forward?
- Refer again to attached HTML file for code
- Using scikit_learn library in Python
- One-Hot Encoding:
    - With 'Theme' and 'Packaging' presumed to be relevant features to predict future Resale Price (target), need to represent those columns numerically
    - Get_dummies creates a numbering system to associate each unique value of the column with a number
    - Allows for all features to be treated as numeric variable types
- NOTE: struggled with converting 'packaging' to float type instead of boolean type
    - Could be a problem with original dataset structure
- Ran a Random Forest Regression w/ a 30-70 split between train and test data
- Results: Not Strong (R-Squared=0.55)
    - Right now, this is not a strong model for predicting future Resale Price given characteristics of a LEGO set upon its release on shelves
    - Suffering from limited free data available
    - Expect only minimal increase in model accuracy after tuning hyperparameters
    - The idea behind the model is still valuable→if LEGO could use the model to learn how consumers value LEGO sets years after release, the company could factor that into its set design

Conclusion:

- The initial results of this project are not groundbreaking, BUT LEGO has the resources to use these ideas in productive ways
- LEGO has access to all proprietary data for past sets, allowing for a larger dataset and better training for a regression/classification model

- LEGO could also use customer data (not accessible to the public) to train the model
- Can help LEGO make future decisions for designing sets if they can see which elements are closely related to value
- Would advise LEGO to scrape resale data from eBay, Etsy, other sites beyond Bricklink
    - Remember, the ‘Resale Price’ column in the free online dataset only indicates the most recent resale price for a set on Bricklink, which is subject to bias & noise
    - Taking the average resale prices over a given time interval for each set gives a better representation of the increase in monetary value
- Ultimate Vision: When designing new sets, LEGO can enter in expected piece counts, minifigure counts, theme, retail price, other details to then estimate/forecast what the set’s value will be *x* years later, allowing them to take advantage knowing what LEGO consumers like

Hi, I’m a college student working on project to forecast LEGO set future resale prices. It may be valuable to LEGO team. My data is limited, but want to share ideas + analysis w/ you in case it helps.