Lego Sets' Prices

Angela Zhang - Duong Do - Vicky Xu

Abstract

This project examines if there are associations between LEGO prices & themes, and LEGO prices & certain age groups, based on a data set posted on brisket.com. The significance of these questions lies in understanding how different factors can drive up the prices of LEGO sets. To answer our research questions, an ANOVA parametric hypothesis test and an ANOVA randomization-based hypothesis test were conducted. The analysis revealed from the hypotheses tests that there are associations between LEGO price and themes as well as LEGO price and age groups. Our findings can help educate LEGO enthusiasts, producers, or retailers on market trends and developing pricing strategies.

Introduction

LEGO has long been enjoyed by adults and children alike. With a long history of abundant creativity and great variance, LEGO sets are a valuable source of information for statistical studies. The characteristics of LEGO sets such as the number of unique pieces, the theme, and the target age group can provide important insights into consumer preferences and behavior. Understanding how these features influence the price of LEGO sets can be valuable information for both LEGO manufacturers and consumers.

This statistical study focuses on examining the relationship between the price of LEGO sets and their unique characteristics. The research questions centered around whether there is an association between the price of LEGO sets and their themes, targeted age groups, and the number of unique pieces. Additionally, the study aimed to determine if the association was positive or negative, and how strong the relationship was.

The findings of this study can have significant implications for LEGO manufacturers, as they can use this information to make strategic decisions about product development and pricing. Additionally, consumers can use this information to make more informed purchasing decisions. Overall, this study adds to our understanding of the toy industry and provides insights into the complex relationship between LEGO sets and their prices. By studying the unique characteristics of LEGO sets, we gain a deeper appreciation for the creative potential of these beloved toys.

Data

To answer our research questions, we investigated 1304 LEGO sets with 14 characteristics documented and posted on brickset.com from Jan. 1, 2018 to Sept. 11, 2020.

Click here to access the data set.

We focus on 4 main variables of the data set:

- `price`: the price of the LEGO set (measured in US dollars or USD), numerical
- `unique_pieces`: the number of unique pieces in the LEGO set (measured in pieces), numerical
- `ages`: the target age group of the LEGO set, categorical
- `theme`: the theme of the LEGO set, categorical

Table 1

	Overall Study Population (n=1304)
unique pieces (piece)	101.5 (50.0, 177.8)
price (USD)	29.99 (14.99, 59.99)
Star Wars™	114 (8.74%)
Friends	101 (7.75%)
City	100 (7.67%)

Data Wrangling

Initial observations reported 239 missing values for `price`, 46 missing values for `unique_pieces`, 111 missing values for `ages`, and 270 missing values for `theme`. Before conducting any data analysis, we decided to omit any observations with unreported `price` and `unique_pieces`.

We chose to only look into three most popular themes (Star Wars, Friends, and City).

```
# A tibble: 3 x 2
  theme      count
  <chr>      <int>
1 City         100
2 Friends      101
3 Star Wars    114
```

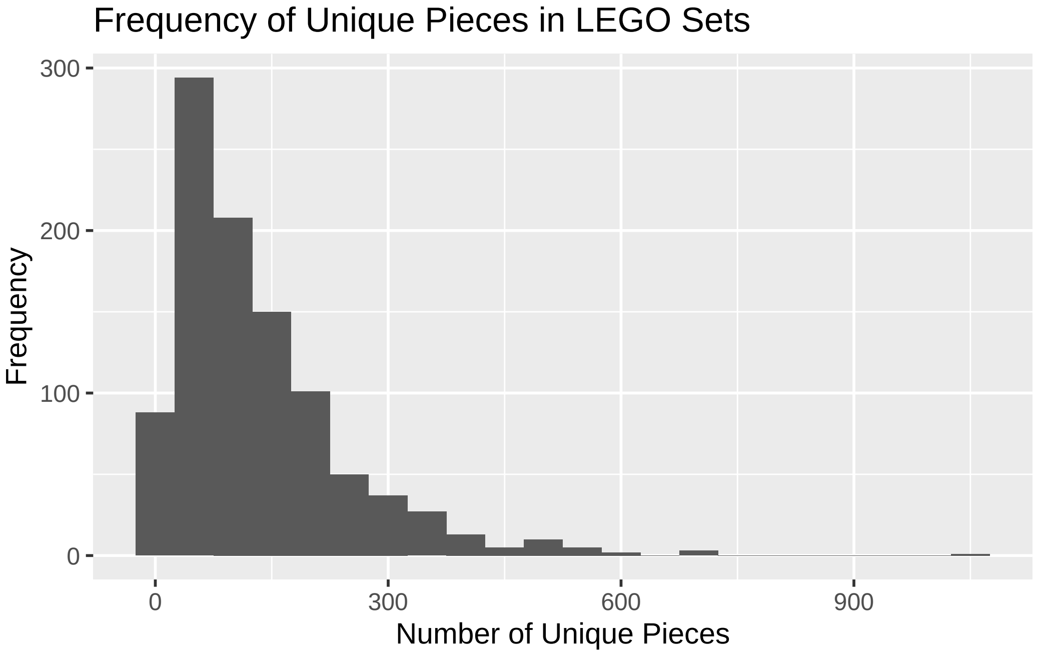
There are 114 observations themed Star Wars, 101 observations themed Friends, and 100 observations themed City.

Since there were many overlaps in the documented age groups (for example, Ages_4+, Ages_4-99, Ages_4-7, etc.), we collapsed them into 3 major categories:

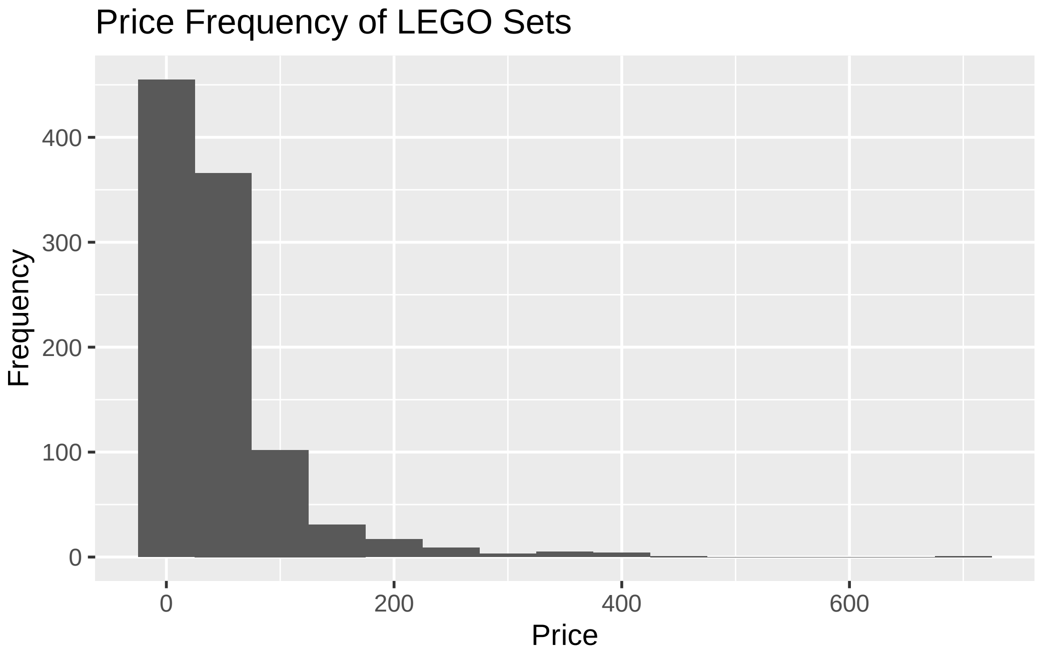
```
# A tibble: 3 x 2
# Groups:   age_group [3]
  age_group      n
  <chr>      <int>
1 Ages_1.5-6   211
2 Ages_16+      78
3 Ages_6-16    705
```

There are 211 observations in group Ages_1.5-6, 705 observations in group Ages_6-16, and 78 observations in group Ages_16+.

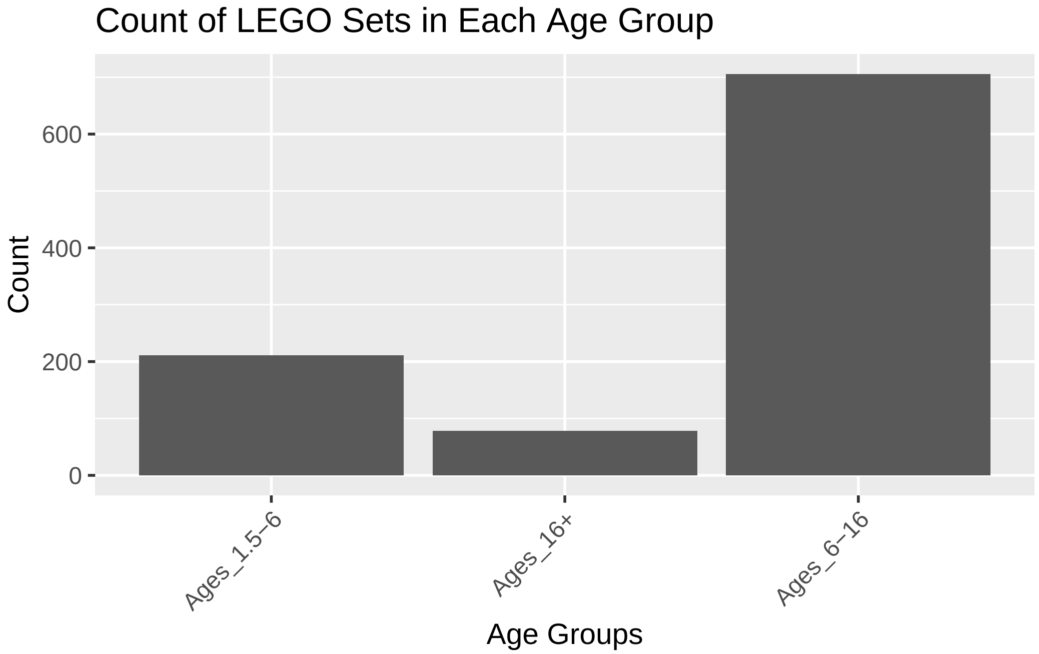
Data Visualizations



This graph shows the frequencies of unique pieces in our data set. We can see that the histogram is skewed to the right suggesting that there aren't many LEGO sets with the number of unique pieces that are extremely large.

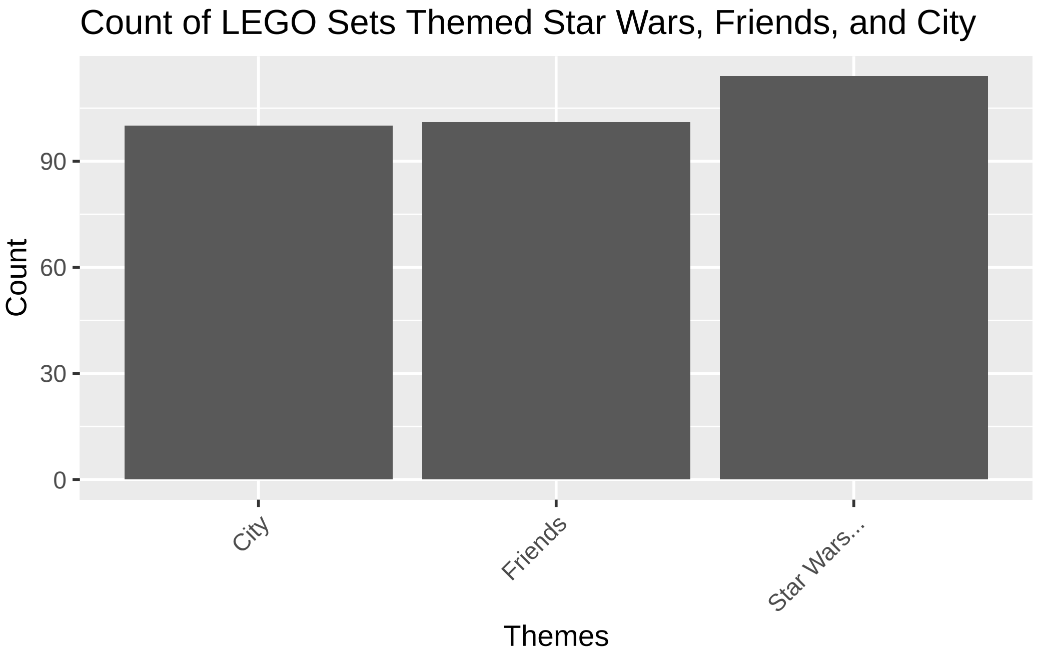


This histogram shows the frequencies of the prices of LEGO sets in this dataset. The histogram is skewed right which suggests that most of the prices are not overly expensive.



This bar graph shows the count of LEGO sets that belong in each of the following age groups:

ages 1.5-6, ages 6-16, ages 16+.



The bar graph shows the three most popular LEGO set themes within this dataset: City, Friends, and Star Wars.

Data Analysis

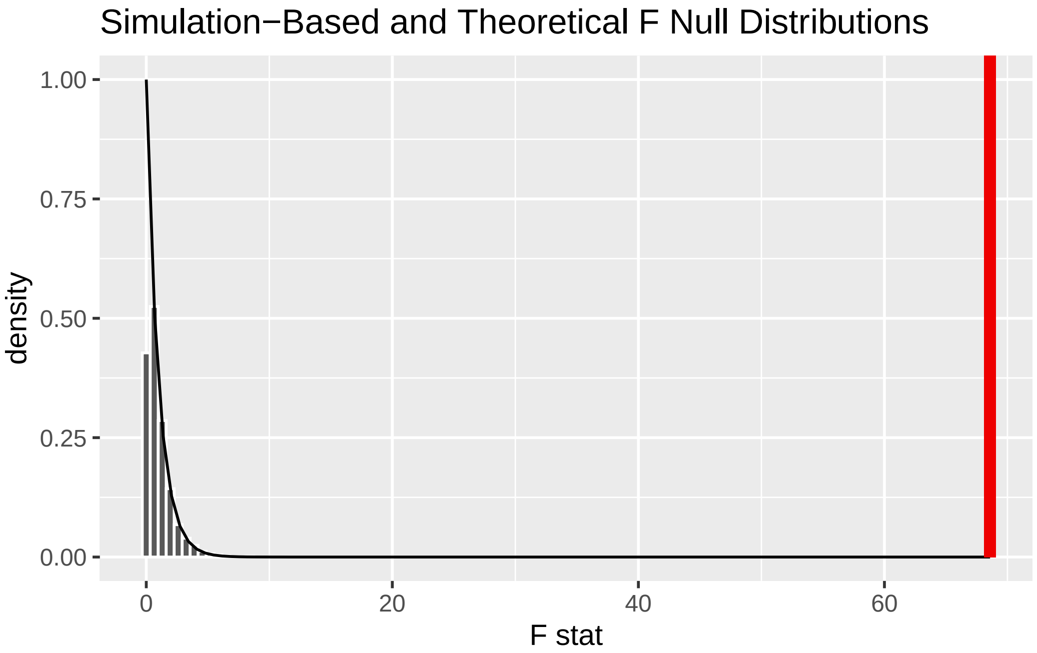
Parametric Hypothesis Test: Price & Age

Null Hypothesis: There is no significant difference among the mean prices of LEGO sets of different age groups.

Alternative Hypothesis: There is a statistically significant difference among the mean prices of LEGO sets of different age groups.

We evaluate the hypothesis test with a significance level of 0.05.

```
# A tibble: 1 x 1
  p_value
    <dbl>
1       0
```



```
Analysis of Variance Table

Response: price
                      Df  Sum Sq Mean Sq F value    Pr(>F)    
as.factor(age_group)   2  428277  214138  68.586 < 2.2e-16 ***
Residuals            991 3094084    3122                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the parametric hypothesis test for age group and price, we can reject the null hypothesis and conclude that there’s a difference between the mean prices of LEGO sets of different age groups, since the p-value is less than the chosen significance level (<.001). These are the findings for our primary results, associated with a multiple regression model below.

See Figure 1

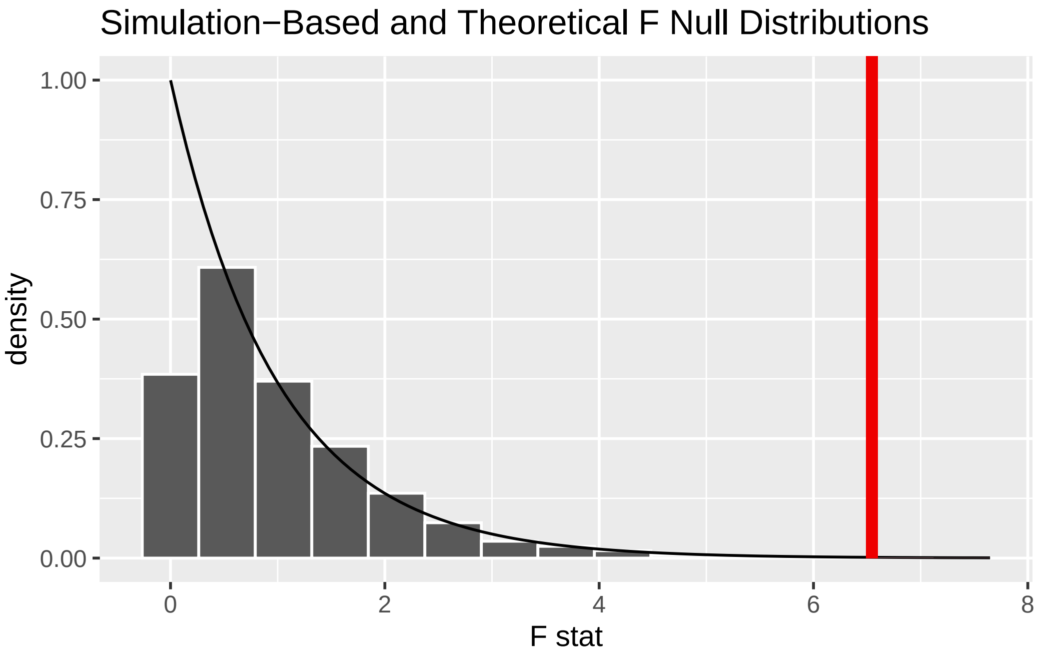
Randomization-Based Hypothesis Test: Price & Theme

Null Hypothesis: There is no difference among the mean prices of lego sets of different themes.

Alternative Hypothesis: There is a difference among the mean prices of lego sets of different themes.

We evaluate the hypothesis test with a significance level of 0.05.

```
# A tibble: 1 x 1
  p_value
    <dbl>
1   0.001
```



Based on the ANOVA randomization-based hypothesis test for price and theme, we can reject the null hypothesis because of the small p-value that is less than the chosen significance level, and conclude that there is a difference among the mean prices of LEGO sets of different themes. These are the findings for our secondary results and they were not adjusted for multiple testing.

Multiple Regression Model: Age Groups, Price, and Unique Pieces

```
# A tibble: 4 x 7
  term                  estimate std_error statistic p_value lower_ci upper_ci
  <chr>                    <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept                2.14      2.78      0.768   0.443   -3.33     7.60
2 unique_pieces            0.373     0.011    32.9     0         0.351    0.395
3 age_group: Ages_16+     28.0       5.41      5.18    0        17.4     38.7
4 age_group: Ages_6-16    -8.35      3.12     -2.67    0.008  -14.5     -2.21
```

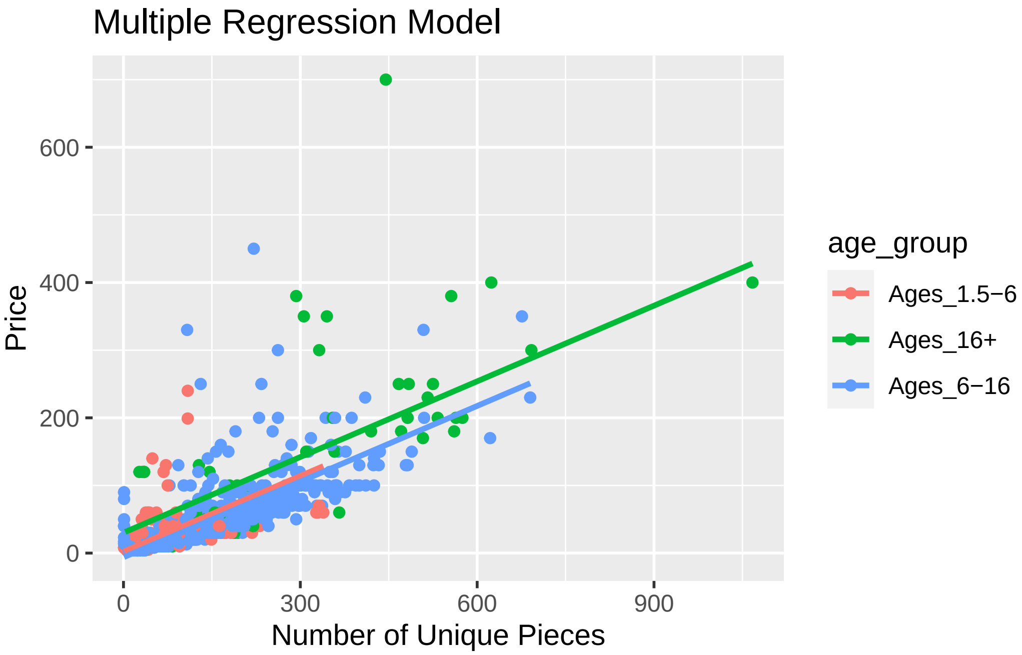


Figure 1: Multiple Regression Lines

To better visualize the association between the number of unique pieces and the price of LEGO sets, we have constructed a bootstrapped multiple regression model with LEGO sets categorized by age groups.

As the model indicates, there is a positive correlation between price and the number of unique pieces. There is also a difference among the intercepts of the three slopes. We can conclude from this model that the more unique pieces there are in a LEGO set, the higher the price. In addition, with the same number of unique pieces, LEGO sets for ages 1.5-6 are $28 less than those for ages 16+, and $8.35 more than those for ages 6-16. These findings are consistent with our parametric hypothesis test on the association between age groups and price.

Conclusion

After conducting an ANOVA parametric hypothesis test and a randomization-based ANOVA test, we found evidence of associations between the mean prices of LEGO sets and both age groups and popular themes. The multiple regression model shows a positive correlation between price and the number of unique pieces, as well as different intercepts for the age groups.

However, our study has some limitations that must be considered. Firstly, there are missing values for some of the variables, which could affect the results. Secondly, we used the lower bound of age groups to categorize them into three major groups, which may lead to potential errors as certain age groups may include adults as well. This means we may have miscalculated the actual sample size for each age group. Lastly, we only examined three popular themes to test the association between price and themes. However, these three themes may not be representative of all the LEGO themes, thus limiting the generalizability of our findings. Future research could address these limitations by finding better ways to deal with missing values, devising more accurate age groupings, and examining a more comprehensive range of themes.