Lego analysis

Introduction

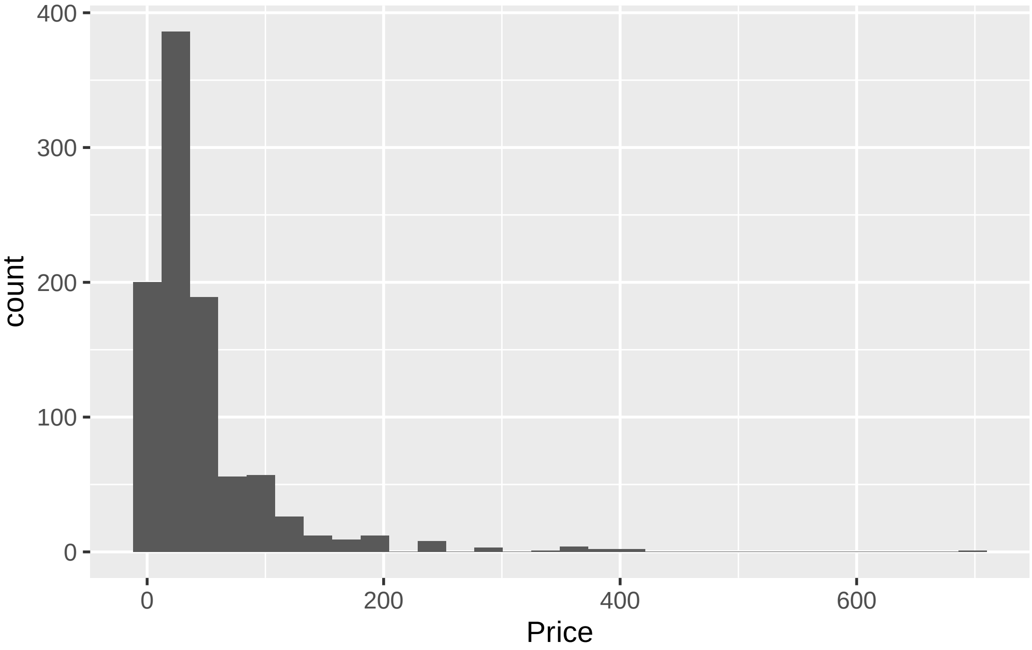
Lego Group is the largest toy company in the world. This study aims to understand what properties of a Lego set affect the price of Lego sets. Legos are extremely popular amongst younger children, and in recent years the popularity has spread to adults as well. Our project investigates whether factors such as a Lego set's availability (retail vs. exclusive, etc) and how many pieces a Lego set contains are related to a Lego set's price.

To answer this question we used a dataset from Peterson & Zeigler (2021) that contains the information of the Lego sets produced between January 1, 2018 and September 11, 2020. The dataset contains data for variables such as: the set's name, its targeted age group, its price, the number of pieces in the set, the type of packaging used, its availability, etc. For this study, we will be focusing on how the number of pieces contained in each set and the set's availability affects the Lego set's price.

Exploratory data analysis

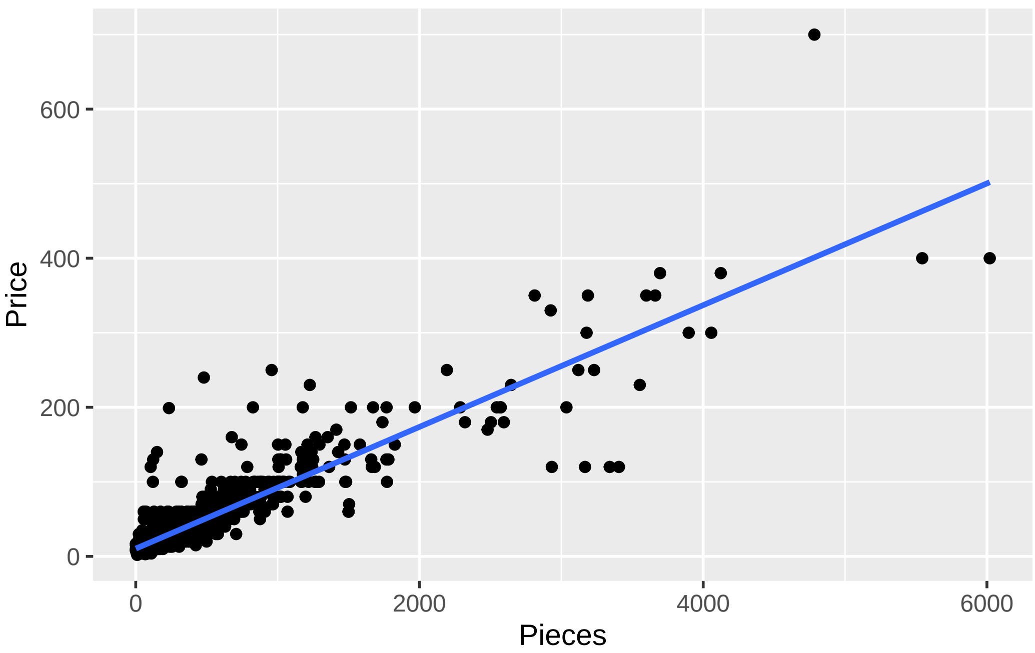
Price distribution:

This figure is a histogram of the distribution of the prices of legos. This distribution appears to be extremely right skewed, with most of the prices being lower than 200. There is an outlier price around 700.



Price and pieces:

This figure shows the relationship between our numerical outcome variable Price and our numerical explanatory variable number of pieces. There appears to be a strong positive relationship between the number of pieces and price. In other words, as the number of pieces increases, there is an associated increase in the price of lego. In accordance with the positive relationship, the slope of the linear regression model is 0.082.



```
                Estimate  Std. Error  t value     Pr(>|t|)
(Intercept) 10.37614974 0.994211514 10.43656 3.100974e-24
Pieces       0.08166235 0.001279477 63.82479 0.000000e+00
```

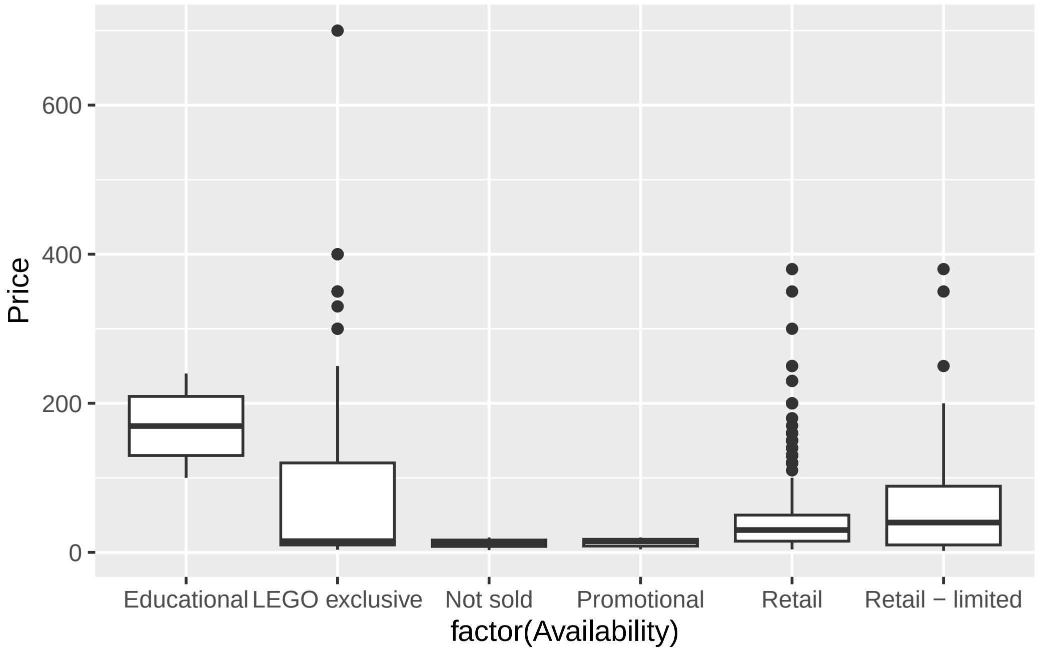
$$\widehat{Price} = \hat{\beta}_0 + \hat{\beta} \cdot Pieces$$
$$\widehat{Price} = 10.37614974 + 0.08166235 \cdot \ Pieces$$

Price and Availability:

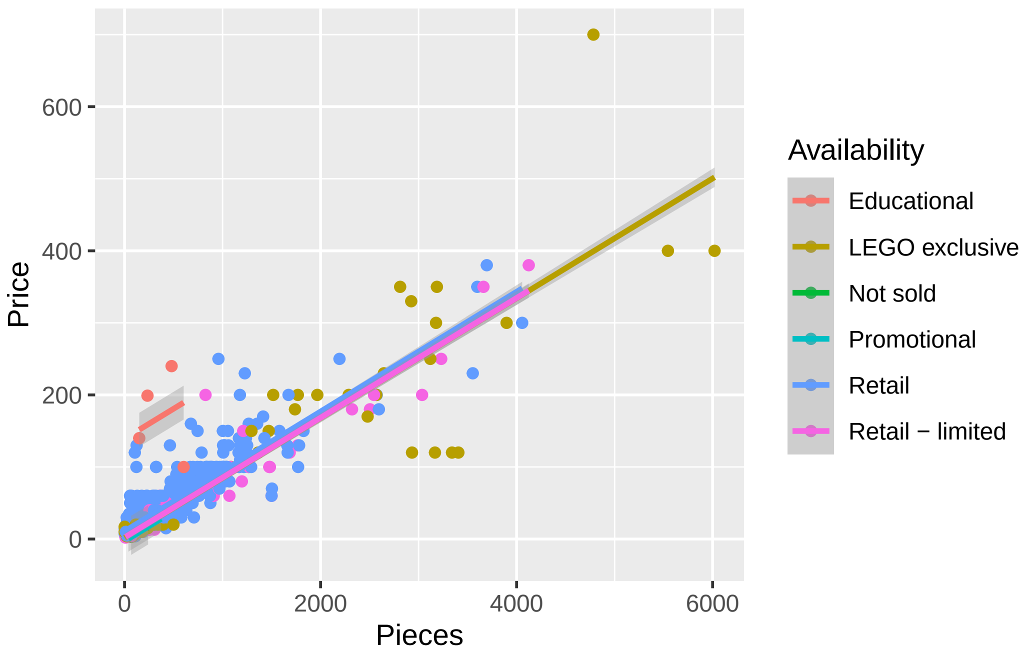
Now we turn to the relationship between price and our categorical explanatory variable which is different availability of legos in Figure. The price of educational legos appears to have a higher price (median around 150) than other kinds of availability (median below 50). While other availability have relatively similar prices, retail-limited are slightly higher (median around 50) among them. The LEGO exclusive category appears to have the most extreme variation, with a number of outliers (a total of 4 from around 300 to around 400) and one extreme outlier (which is around 700) dollars. There are also a number of outliers for Retail (arranged from around 100 to slightly below 400) and Retail-limited (a total of 3 from around 250 to slightly below 400) appearing in the graph.

```
                               Estimate Std. Error   t value     Pr(>|t|)
(Intercept)                   169.70000   28.20481  6.016704 2.527288e-09
```

```
AvailabilityLEGO exclusive    -88.37463   28.88454 -3.059582 2.277876e-03
AvailabilityNot sold         -157.71000   43.08356 -3.660561 2.653585e-04
AvailabilityPromotional      -156.71000   35.35660 -4.432270 1.039713e-05
AvailabilityRetail           -128.82097   28.27506 -4.555993 5.883737e-06
AvailabilityRetail - limited -105.22429   28.99947 -3.628490 3.000833e-04
```



Finally, we examine the relationship between all three variables (pieces, availability, and price), at once, by representing different availability with different colors in the following scatterplot.



The positive relationship between the number of pieces in a lego and the price of it appears to be very similar between LEGO exclusive, Not sold, Promotional, Retail, and Retail - limited Availability. However, educational availability is an exception.

```
# A tibble: 6 x 3
  Availability          r     N
  <chr>             <dbl> <int>
1 Educational      -0.152     4
2 LEGO exclusive    0.904    82
3 Not sold          0.984     3
4 Promotional       0.951     7
5 Retail            0.897   802
6 Retail - limited  0.953    70
```

According to the correlations between pieces and price under different availability, we can see that only educational availability has a negative correlation between price and pieces. Since the number of legos with educational availability in our data set is only four, we are less confident in generalizing this correlation to the population of all educational lego sets. We tried to fit the data with an interaction model to see if the number of pieces in a model is dependent on the availability category of the set.

```
                                         Estimate  Std. Error    t value
```

```
(Intercept)                          186.09730070 26.81391059  6.9403267
Pieces                                -0.04470975  0.06545505 -0.6830604
AvailabilityLEGO exclusive          -182.37974796 27.00261782 -6.7541506
AvailabilityNot sold                -188.61394404 41.93779816 -4.4974689
AvailabilityPromotional             -182.63913291 31.85064684 -5.7342362
AvailabilityRetail                  -176.71156187 26.83715437 -6.5845864
AvailabilityRetail - limited        -181.85001335 27.06629872 -6.7186879
Pieces:AvailabilityLEGO exclusive      0.12578605  0.06548167  1.9209352
Pieces:AvailabilityNot sold            0.14163610  0.20544723  0.6894038
Pieces:AvailabilityPromotional         0.09439167  0.10047952  0.9394121
Pieces:AvailabilityRetail              0.13150986  0.06548610  2.0082103
Pieces:AvailabilityRetail - limited    0.12496387  0.06552884  1.9070058
                                        Pr(>|t|)
(Intercept)                         7.210128e-12
Pieces                              4.947342e-01
AvailabilityLEGO exclusive          2.491680e-11
AvailabilityNot sold                7.720906e-06
AvailabilityPromotional             1.313153e-08
AvailabilityRetail                  7.514409e-11
AvailabilityRetail - limited        3.145080e-11
Pieces:AvailabilityLEGO exclusive   5.503702e-02
Pieces:AvailabilityNot sold         4.907365e-01
Pieces:AvailabilityPromotional      3.477565e-01
Pieces:AvailabilityRetail           4.490184e-02
Pieces:AvailabilityRetail - limited 5.681935e-02
```

However, the interaction terms we got were all non-significant, meaning that the number of pieces of a set is not dependent on the availability. Therefore, we will use the parallel slope model model to fit our data.

$$\begin{aligned}\widehat{Price} = \hat{\beta}_0 + \hat{\beta}_1 \cdot pieces \ + \hat{\beta}_2 \cdot \mathbb{1}_{\text{LEGO exclusive}}(Availability) \\ +\hat{\beta}_3 \cdot \mathbb{1}_{\text{Not sold}}(Availability) \\ +\hat{\beta}_4 \cdot \mathbb{1}_{\text{Promotional}}(Availability) \\ +\hat{\beta}_5 \cdot \mathbb{1}_{\text{Retail}}(Availability) \\ +\hat{\beta}_6 \cdot \mathbb{1}_{\text{Retail - limited}}(Availability)\end{aligned}$$

Modeling Analysis:

To quantify the relationship between the LEGO recommended price of the set and the total number of pieces in the set and the availability of the set two explanatory variables, we fit a parallel slope model. In this model, the outcome variable was the LEGO recommended price of the set, and the explanatory variable was the total number of pieces in the set and the availability of the set. The regression model for our parallel slope model is shown below

```
                                Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)                    139.22319944 11.982455680  11.618920 2.702707e-29
Pieces                           0.08309966  0.001256073  66.158300 0.000000e+00
AvailabilityLEGO exclusive    -137.44245025 12.284568550 -11.188220 2.112271e-27
AvailabilityNot sold          -139.67044866 18.292004404  -7.635601 5.415101e-14
AvailabilityPromotional       -142.17646292 15.011306036  -9.471292 2.075308e-20
AvailabilityRetail            -128.49484267 12.003420508 -10.704852 2.424254e-25
AvailabilityRetail - limited -137.11140623 12.320379081 -11.128830 3.815521e-27
```

Prediction for the educational Lego set:

$$\widehat{Price} = 139.223 + 0.083 \cdot \ pieces$$

The intercept ($= 139.223$) represents the LEGO recommended price of the set with educational availability when zero number of pieces is in the set. The estimate for the Pieces coefficient ($= 0.083$) represents that with each additional piece contained in the set, the price of the lego set will increase 0.083, on average. For the AvailabilityLEGO exclusive coefficient, it means that the intercept of the set with LEGO exclusive availability is 137.442 decreased compared to the intercept of the baseline group (educational availability). For the AvailabilityNot sold coefficient, it means that the intercept of the set with not sold availability is 139.67 decreased compared to the intercept of the baseline group. For the AvailabilityPromotional coefficient, it means that the intercept of the set with Promotional availability is 142.176 decreased compared to the intercept of the baseline group. For the AvailabilityRetail coefficient, it means that the intercept of the set with Retail availability is 128.495 decreased compared to the intercept of the baseline group. For the AvailabilityRetail - limited coefficient, it means that the intercept of the set with Retail - limited availability is 137.111 decreased compared to the intercept of the baseline group.

The slope of the regression lines (0.083) were reliably different from 0 (t=66.158, p=0<0.05), indicating a reliably positive relationship between the recommended price of the lego set and the number of pieces in the set. And all the other coefficients are reliably different from 0 (all p_value < 0.05), indicating that other availability of the lego set do have different intercept compared to the baseline group (educational availability).

Discussion

Conclusion:

We found that with the number of pieces increased in the lego set, the price of the set will increase at the same time in most availability categories of the set, except for the education availability, and the slope coefficient for the five availability categories are quite similar. On average, with each additional unit increase on the number of pieces of the set, the LEGO recommended price will increase 0.083. There is no interaction between the number of pieces and the recommended price of the set, so a parallel slope model is best fit for our data.

Overall, our result suggests that the number of pieces and the availability of the lego set are two factors in the recommended price of the set. We think this result is reasonable because more pieces in a lego set means more materials are needed and higher cost for the product. So it's reasonable to have higher prices for the product with higher cost. Moreover, the different price with different availability of the lego set is a marketing strategy for the lego companies. For example, retail-limited has a higher mean price than retail, and the "limited" availability is a way to promote people's spending on the product. The trends found in this study are important for us to understand the lego industry's market strategy and profit.

Limitations:

There are a couple of limitations in our report. First, 336/1304 observations contained missing values in either price, availability, or pieces. That is more than a quarter of the entire data missing, which is a quite sizable amount of data. However, we were still left with 968 rows of observations, which is enough to make meaningful analysis with. A second limitation to our dataset is that there are certain variables that contain very few observations, such as there were only 4 observations of lego sets that are under the educational category in their availability. This limitation made us less confident in generalizing our findings to all lego sets in that category of availability. Lastly, our dataset only contains data from years 2018-2020, which makes our analysis not applicable to current lego sets since it has now been almost three years since 2020.

Further Question:

If we were to continue researching this topic, we would like to work with more data under the education availability. With only 4 observations under this category, we are not confident enough for our conclusion. Extra data might change our current conclusion on how price and availability will be related to each other. Since the data comes from years 2018-2020, it will also be valuable if we have the dataset of legos 2021-2022. In our current dataset, there are too many theme categories but not enough observations within each theme category. It will be useful if we could get more observations on a few specific themes and conduct some future analysis on them. The regression model between the relationships of other different availability such as age and unique pieces could be conducted with price in the future. There also might be some differences if the amazon price is used in data analysis.