Group L Final Report

Abstract

In this project, we conduct a study that examines the relationship between the number of pieces and the mean recommended price of 740 Lego sets, as well as the relationship between 10 different theme groups among Lego sets and their corresponding mean recommended price. This project opens new ways for LEGO to alter their price based on characteristics of the set, and helps customers predict the expense of a Lego set. We use a simple linear regression model to analyze the relationship between pieces and price, and a multiple linear regression model to evaluate whether theme is a confounder in this relationship. We also use ANOVA to examine the relationship between theme and price. Our analysis finds statistically significant evidence of an association between pieces and price, and between theme and price, with theme not found to be a confounder in the first relationship.

Introduction

Lego is a line of plastic constructions that are manufactured by The Lego Group, one of the largest toy companies in the world to this date. In this project, we present a modern, data-driven approach to analyzing the price of Lego sets using a data set about LEGO building brick groups.

The phenomenon we study is how Lego sets' price can be affected by certain characteristics of the set, including the number of pieces and the set's theme. Our focused research questions include: Does the number of pieces in a Lego set affect its price? And do different Lego themes have different levels of expense associated with them? We think that our research topic is relevant to business intelligence, since it will give Lego insight into how it wants to alter its price. This project is also beneficial to customers, as by analyzing the dataset even more in-depth, customers will be able to see which characteristics would usually make a set more expensive.

We plan to study this topic by conducting two hypothesis tests using a linear regression model and the Analysis of Variance (ANOVA). For our primary hypothesis, we hypothesize that there will be a positive linear relationship between the mean recommended price and the mean number of pieces of Lego sets in this data set, meaning more pieces will be associated with a higher price. For our secondary hypothesis, we hypothesize that themes affect the mean recommended price of Lego sets, whether the effects are positive or negative.

In our project, we create graphs to visualize our findings, checked the conditions for the tests (including the residual plots), and include a fitted linear regression models. In the following sections, we describe the data set and our activity, and conclude with the indications of our findings.

Data

The population of interest in our study is all LEGO sets that were released in the United States from 2018 to 2020. We use a modified version of data originally published by Peterson and Zeigler (2021). The data was collected from the website brickset.com, a database of information on LEGO sets. The variable Pages, which was not used in our analysis, is the only variable not taken from brickset. We considere the variables Price, Pieces, and LargerThemes for our analysis. Price, our response variable, is the price in USD that LEGO recommends for each set. Pieces is one of our explanatory variables, and counts the total pieces in each set. LargerThemes is our other explanatory variable, which we formed from Theme, which recorded the product line each set was released in. The original variable Theme has 41 levels, so we made the new variable by grouping associated themes into 10 buckets. These 10 buckets are: "Kids Theme", "Freestyle Theme", "Lego Universe Theme", "MoviesThemes","Video Games Theme", "Other", "Toys With Remote Theme" "Objects Theme", "Home Decor Theme", "ArchitectureTheme." We select sets that we think fit into each group based on the LEGO official description on Lego.com. For example, we group the sets that were based on movies or TV serieslike "Harry Potter", "Minions", "Powerpuff Girls", "Jurassic World" into "MoviesTheme." Toys that have an engine and can run are put into "Toys With Remote Theme". Lego also released an unique line of pictures made from Lego (similar to a puzzle set), so we put that in "Home Decor Theme." There may be some overlapping among in the groups, but we think that overlapping section is minimal.

There are missing values for our variables of interest in the original data, so we remove all missing values. In total, 565 observations are excluded, which is about 40% of the original data. Below here, we include the Table 1, which summarizes important relevant statistics for our study.

Table 1

	Overall Study Population (n=740)
Variable	**Median (Q1, Q3)**
Price (in USD)	29.99 (14.99, 59.99)
Number of Pieces	235 (117.75, 518.25)
Theme Type	**Mean (Proportion)**
Architecture	41 (5.5%)
Freestyle	56 (7.6%)
Home Decor	14 (1.9%)
Kids	52 (7%)
Lego Universe	230 (31.1%)
Movies	205 (27.7%)
Objects	17 (2.3%)
Other	50 (6.8%)
Toys with Remote	31 (4.2%)
Video Games	44 (5.9%)

Here are some data visualizations we made to demonstrate the relationships between the response variable and explanatory variables, and the data distributions in both groups.

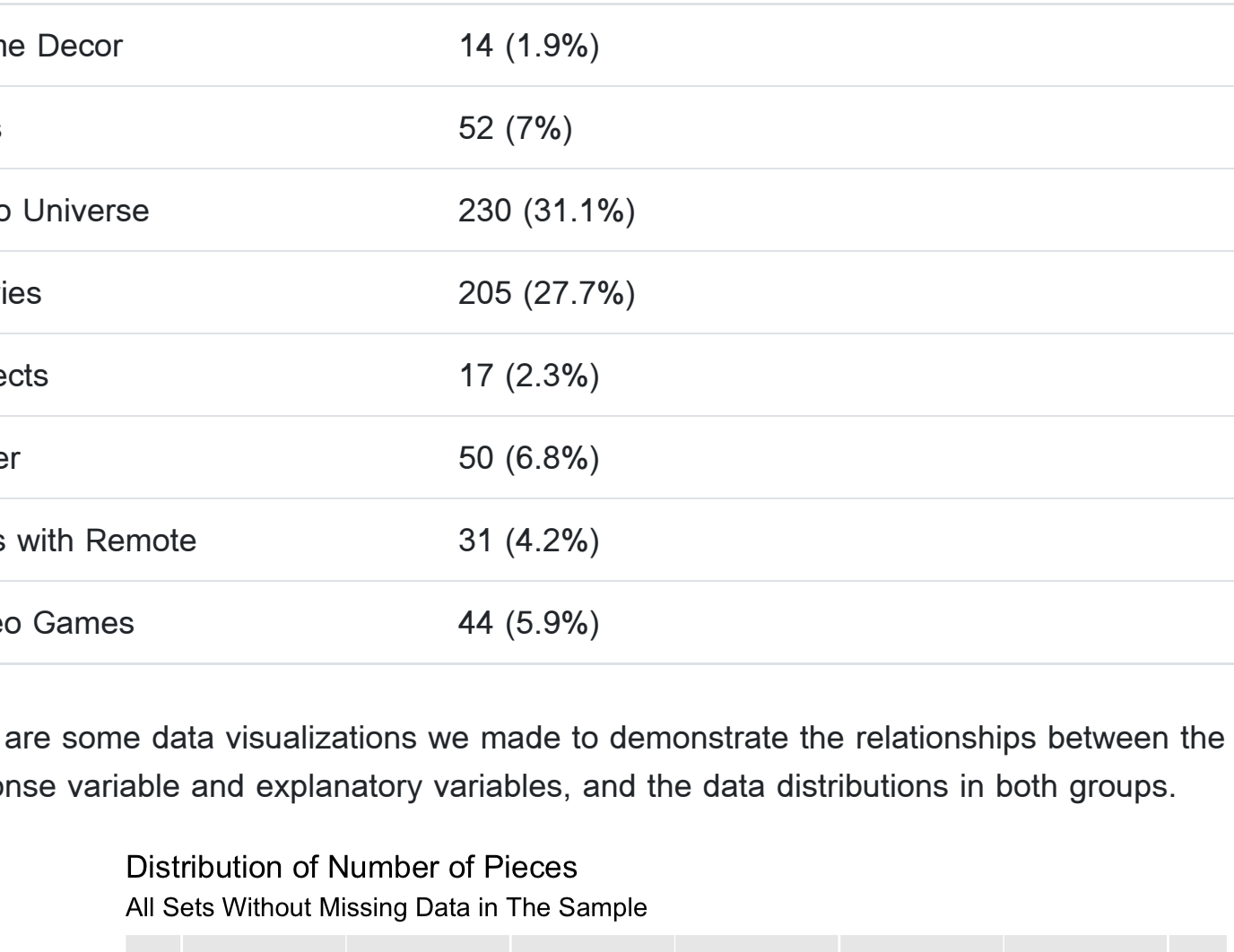

Figure 1: This is a plot that describes the distribution of number of pieces among all Lego sets in our sample.

Figure 2: This is a plot that describes the frequency of theme groups among all Lego sets in our sample.

Results

Results for Primary Hypothesis

For our statistical analysis for our primary hypothesis, we want to analyze whether there is a linear relationship between the mean number of pieces in a lego set and the mean price of our lego set. In order to achieve this, we first create a hypothesis test. We hypothesize that the mean number of pieces in Lego sets affects the lego sets' mean recommended price. Our null hypothesis states that there was no association between mean number of pieces and mean price of lego sets. The alternative hypothesis states that there is an association between the average number of pieces and the mean price of lego sets.

We fit a linear regression model into the relationship between Price and Pieces. The y-intercept of this equation is approximately $10.84, meaning the price for a lego set with 0 pieces will be $10.84. The slope is approximately $0.0812, meaning that for every one lego piece increase in a set, the price will be $.0812 higher.

Since we are using a linear regression line analysis ($E[Y|x] = \beta_0 + \beta_1 x$), we want to see if there is a statistically significant relationship between the number of pieces (the y value or explanatory variable) and mean recommended price of Lego sets (the x value or response variable). The p-value is the probability that the estimated slope in our fitted regression line appears in real life if the null were true. We define that statistically significant relationship occurs if the p-value is smaller than .05, which means that we then are able to conclude that the β_1 value is unusually large under the null and we can reject our null hypothesis.

This is how our hypothesis tests are setup:

$H_0 : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

In order to test for significance, we construct a hypothesis test using permutation. We permute the data set by using the infer package in R, which randomly assigns the number of pieces to the 740 lego sets in our data sets, without altering the prices. We permute the two variables 5000 times in R.

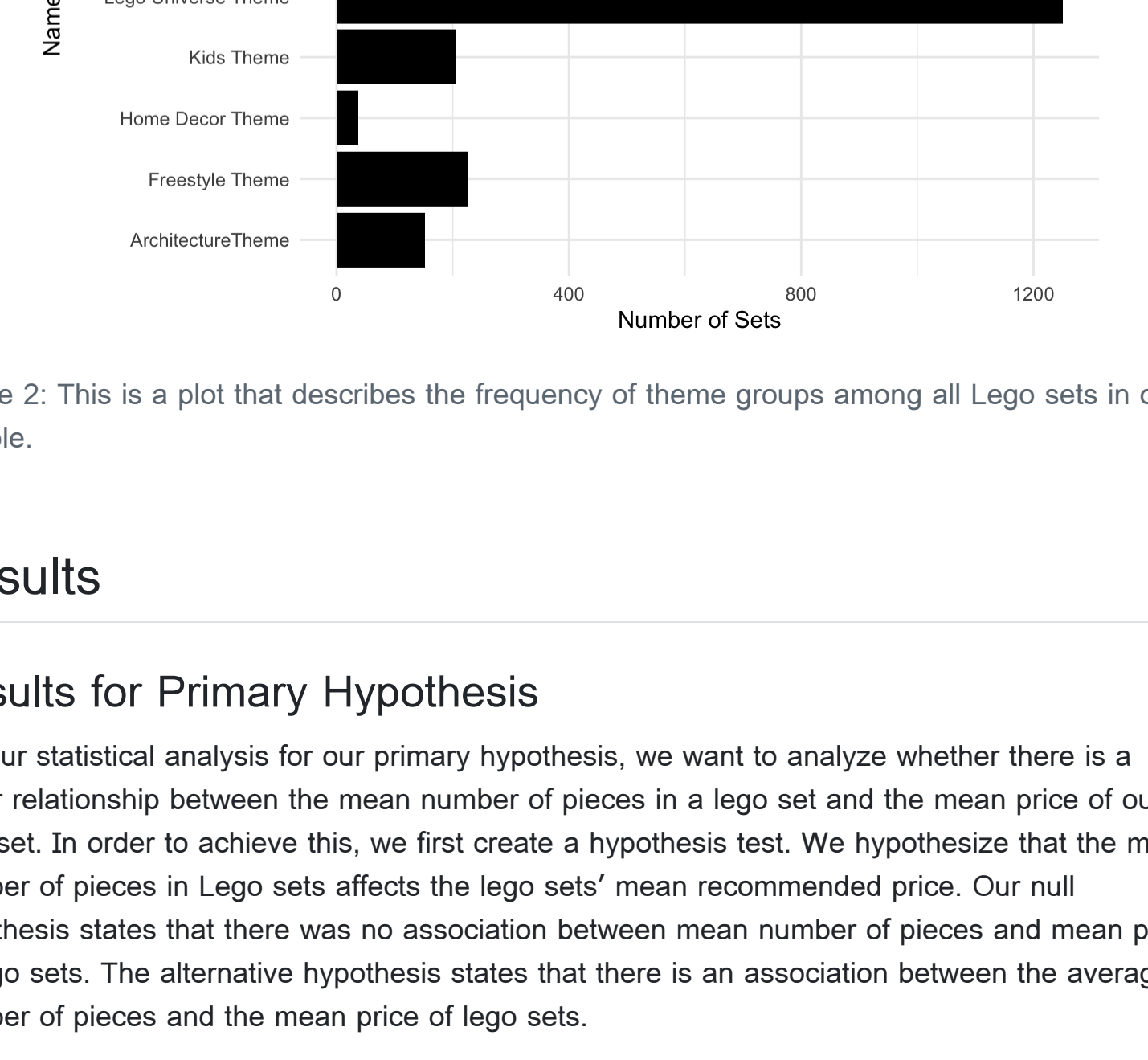

Figure 3: This figure shows the null distribution and p-value (signified by the red line)

For our results we get our p-value = 0. We think that R rounds it to this value, meaning that the p-value is smaller than 0.05. Hence, we reject the null hypothesis and are able to conclude that the mean recommended price and the number of pieces in a Lego set has a statistically significant relationship with each other.

We also suspect that the themes of the Lego sets can be a confounder in the relationship between Price and Pieces, which means that any change in the recommended price that we observed can be attributed to the theme of the set, and not the number of pieces. Hence, to examine this, we fit a multiple regression model into the relationship between Price, Pieces, and LargerThemes, with the first as a response variable and the latter two as explanatory variables.

This linear regression model looks at the relationship between the theme of a Lego set (the first independent variable), the price of the Lego set (the dependent variable), and the number of pieces in a Lego set (the second independent variable). LargerThemes is also a nominal variable, and Price and Pieces are continuous variables. In the equation for a linear regression line, our y value is the dependent variable which is the price that is based off of the theme of the Lego set and the price of the Lego set, and our y-intercept, which is the ground value for price, is 3.07047 dollars. R also randomly chooses the reference level as Architecture Theme. This means that a Lego set with Architecture Theme and 0 pieces would be 3.07047 dollars on average.

We have 11 slopes which each represent how the Lego sets' price may vary according to the number of pieces and among different themes. Each Theme slope describes how the price changes if the set falls in that Theme group. For example, we will interpret two estimated slopes, and this will take into account two themes, Home Decor Theme (slope: -37.95) and Freestyle Theme (slope: -3.32) . This means that the Home Decor Theme set will be 37.95 dollars cheaper than the price of a regular Lego set, and the Freestyle Theme will be about 3.32 dollars cheaper than the regular Lego set.

In this new multiple regression line, we find that Price still has a positive relationship with Pieces, however, the slope coefficient of Pieces is now 0.0853, instead of the original 0.081994. This means that for each piece increase in the set, the recommended price will increase by $0.0853 on average. We wonder if this is enough to conclude that Larger Themes is a confounder and that it distorts our initial findings. By informal rule, we determine that if the slope changes by 10% (in either direction), then the newly introduced variable is a confounder to the relationship[1].

We want to find if the new slope changes by 10% (in either direction):

$(0.0853 - 0.081994)/0.081994 * 100 = 4.03$

The results show that the slope coefficient only changes by 4.03%, so by informal rule, we conclude that LargerThemes is not a confounder, and the relationship between the recommended price of a set and the number of pieces in that set remains a linear, positive one.

Results for Secondary Hypothesis

Our secondary hypothesis is that the mean recommended price of Lego sets is not the same for all theme groups, or there is some variability among all themed Lego sets. To test for variability, we use an Analysis of Variance, or ANOVA. An ANOVA test is a type of statistical test used to determine if there is a statistically significant difference between two or more categorical groups, and in our case, we want to apply ANOVA to see if there's a price difference among 10 themes of Lego.

Among the 10 themes, the theme Architecture is used as our reference level for this analysis, as this is the reference level chosen by R. We conduct an F-test to calculate the p-value and conclude our hypothesis test. We assume that the observations in the samples are independent from each other, and the variance among and within groups is about the same, which makes the data set qualified for this approach.

This is how our hypothesis are set up:

H_0: The mean price is the same for each level of LargerThemes $\mu_1 = \mu_2 = \ldots = \mu_{10}$

H_A: The mean price is not the same for each level of LargerThemes, i.e. at least one mean is different from the others

In R, we calculate the F statistics from the observed data using the infer package. Subsequently, we generate a simulation under the null by using permutation 5000 times. This means that R randomly assigns 10 Lego themes to the 740 lego sets in our data sets, without altering the prices. The purpose of this simulation is to observe how the themes and the price of Lego sets interact, if they truly don't have any association with each other. Finally, we calculate the probability that our observed F statistics exist in real life if the null were true, which is presented by a p-value.

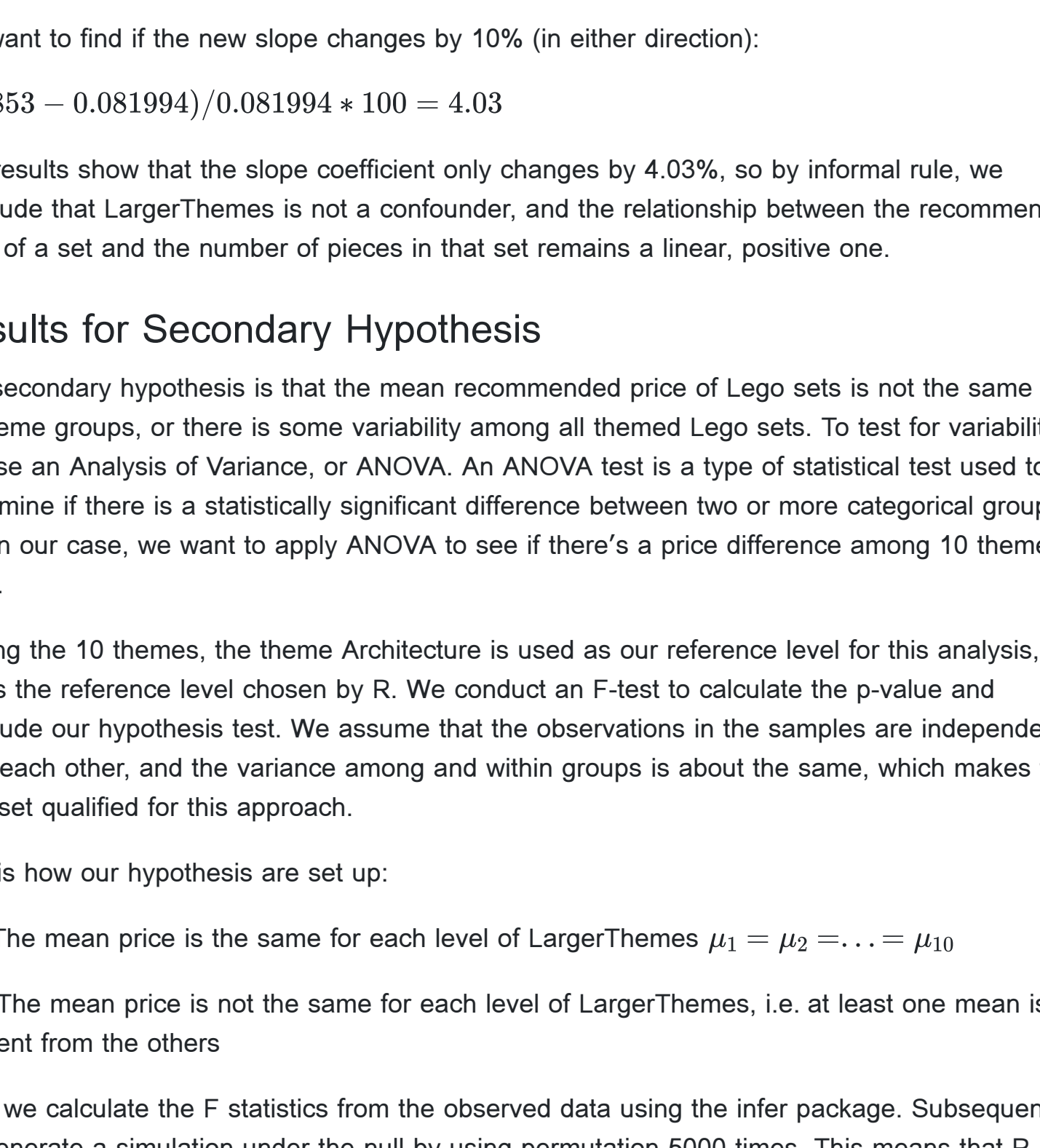

Figure 4: This figure shows the null distribution and observed F-statistic

The results in R give an F statistic of 9.084, and a p-value so small that R automatically rounds to 0. This means that if the null is true, the probability that we would observe such an F score is nearly 0. Since this p-value is smaller than 0.05, we reject the null and conclude that there's variability in recommended price among themed Lego sets.

Conclusion

Our analysis of the relationship between the number of pieces and price of a Lego set found that there is statistically significant evidence of a positive association between pieces and price, and that theme is not a confounding variable. Additionally, we found evidence of some association between the variable LargerThemes and Price, though we are unable to discern for which categories the mean price is different. Due to missing values, we excluded about 40% of the observational units in the original data set from our analysis, which affects the generalizability of our findings. The analysis could be improved by using a different method to deal with missing values, and by further refining our regression models for our primary hypothesis, which were somewhat limited by the methods of analysis we have learned so far. Additionally, the data set does not include data on Lego "mystery sets", so our results are not relevant to these products. Our research does not imply that Lego sets are priced differently based on theme, just that the mean price is not the same for all themes for some reason. We also did not investigate the possibility of pieces being a confounder in the relationship between theme and price. Our research has found that there is a positive association between pieces and price, and some association between theme and price.

Footnotes

1. Multivariable Methods, https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_multivariable/bs704_multivariable_print.html#:~:text=As%20noted%20earlier%2C%20some%20investigators,adjusting%20for%20the%20potential%20confounder. ↩