Econometric 8320: Jordan Wheeler
Tools for Data Analysis May 2, 2018

Final Project Write Up

1 Introduction

For the final project in Economics 8320 we were told to use the data analysis and Python techniques/tools we learned throughout the semester to solve a problem. The specific problem had to come from one of three data sites and the data on these sites had to be analyzed and interpreted. The first phase of the final project was to write Python scripts to scrape the website that we chose. The next phase was to clean and organize the scraped data. Lastly, after cleaning the data, we used plotting and statistical methods to analyzed the data.

2 Research Questions

The first question I wanted to look into was if I could model the price of a lego set by the number of pieces the lego set had. It seems reasonable that the price would be influenced by the number of pieces since the number of pieces would correlate to the amount of ”raw” material used. To evaluate this question, I used a simple linear regression model to see if the number of pieces could explain the variance in the price.

The second question I wanted to look into was if the average price of a lego set has increased from 1997 to 2017. I would assume that the price has increase due to inflation of the US economy. To evaluate this question, I used a bar chart to visually see the increase.

3 Data Scraping

The data source that I used was the Brickset website. I used a spider to retrieve the Lego Set Name, Price, Number of Pieces, Number of Minifigures, and Year of all the lego sets sold on the Brickset website from year 1997 to 2017. I saved all the data retrieved into a CSV file which I then used as my data frame.

This process was somewhat tedious. I created a Python script for each year (.py files of the scrape attached in folder) since I could not figure out how to extract the year that each lego set was sold. I made sure that each script wrote into the same CSV file so I would not have to deal with combining data frames later in the process.

4 Data Cleaning

This part of the project was by far the most time consuming. Once I had scraped the data into a CSV file, I opened the CSV file as a Pandas data frame in Python to examine the layout of the data. Right away I noticed a few things that were formatted right.

The first thing that I noticed was the prices scraped from the website included both USD and Euro. This was an issue since I need to extract the USD price. To do this, I used the Regex package to find the first number of the ”Price” column (the USD was always before the Euro). If a lego set only had a Euro price, then I made sure that I removed that lego set from the data frame.

The next thing that I noticed was that during the scraping process, sometimes the number of pieces in a lego set was inserted as the price. I tried for a few hours to look into the HTML code of the website and look into my scraping scripts but could not seem to find a reason why some entries did this. To solve this problem, I got rid of any row that did not have a decimal in the ”Price” column (I noticed that all the prices should have a decimal, so if it did not, then it was a case of a bad scrape).

The last thing that I had to do to clean up the data frame was to get rid of any rows that did not have a number in the ”Pieces” column. Since I was trying to model the price of a lego set by the number of pieces, if a lego set did not have any recorded pieces, I could not use that lego set in my model building.

Once I had accomplished these data cleaning tasks, my data frame was ready for some plotting and statistical modeling.

5 Question One

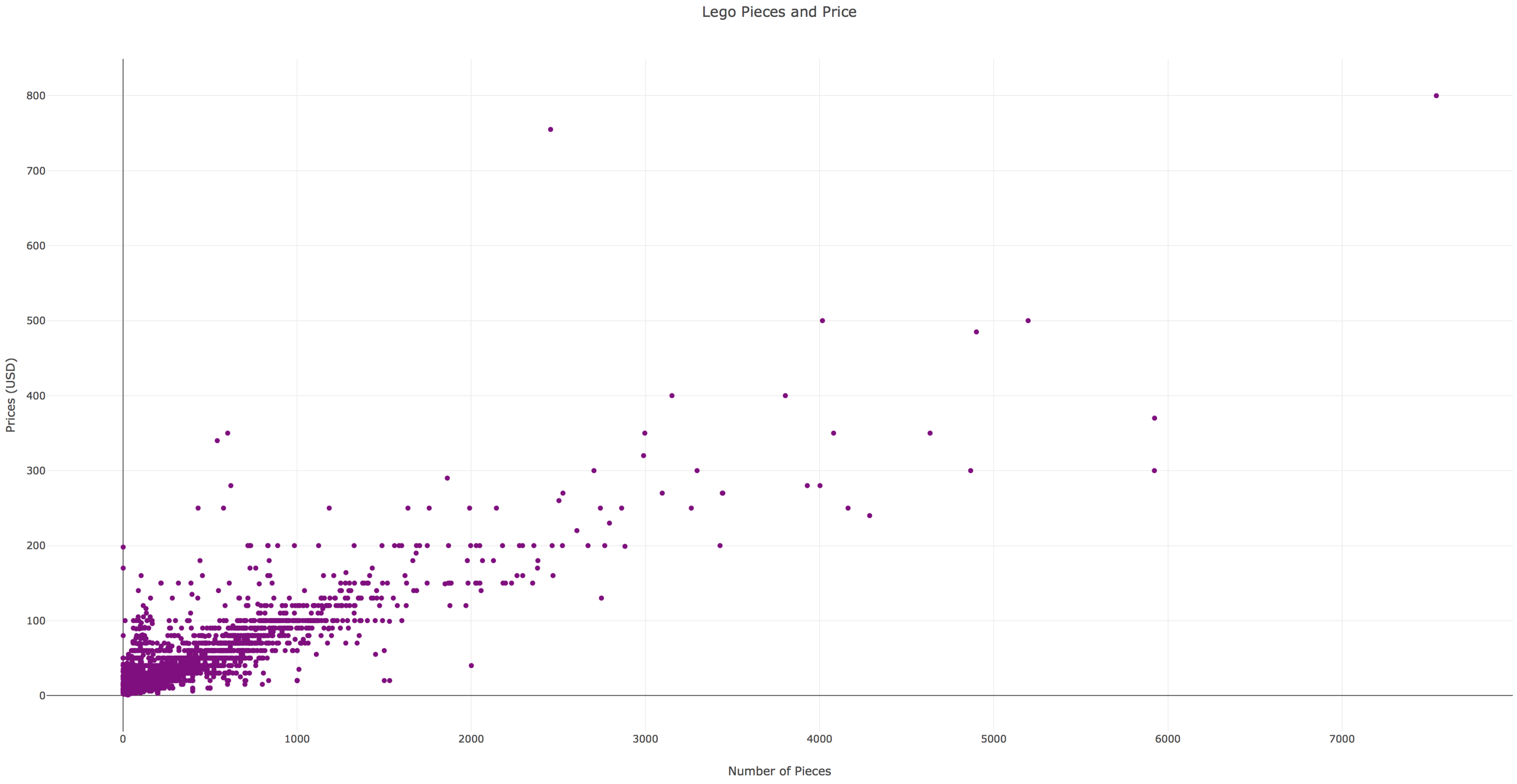
For this question I used my clean dataset and to create a regression model where price was the dependent variable and the number of pieces was the independent variable. The following table is a summary of the regression model created:

Dep. Variable:	Price Amount	**R-squared:**	0.750
Model:	OLS	**Adj. R-squared:**	0.750
Method:	Least Squares	**F-statistic:**	1.526e+04
Date:	Wed, 02 May 2018	**Prob (F-statistic):**	0.00
Time:	21:47:38	**Log-Likelihood:**	-22933.
No. Observations:	5095	**AIC:**	4.587e+04
Df Residuals:	5093	**BIC:**	4.588e+04
Df Model:	1		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.7419	0.359	27.122	0.000	9.038	10.446
Pieces	0.0852	0.001	123.533	0.000	0.084	0.087

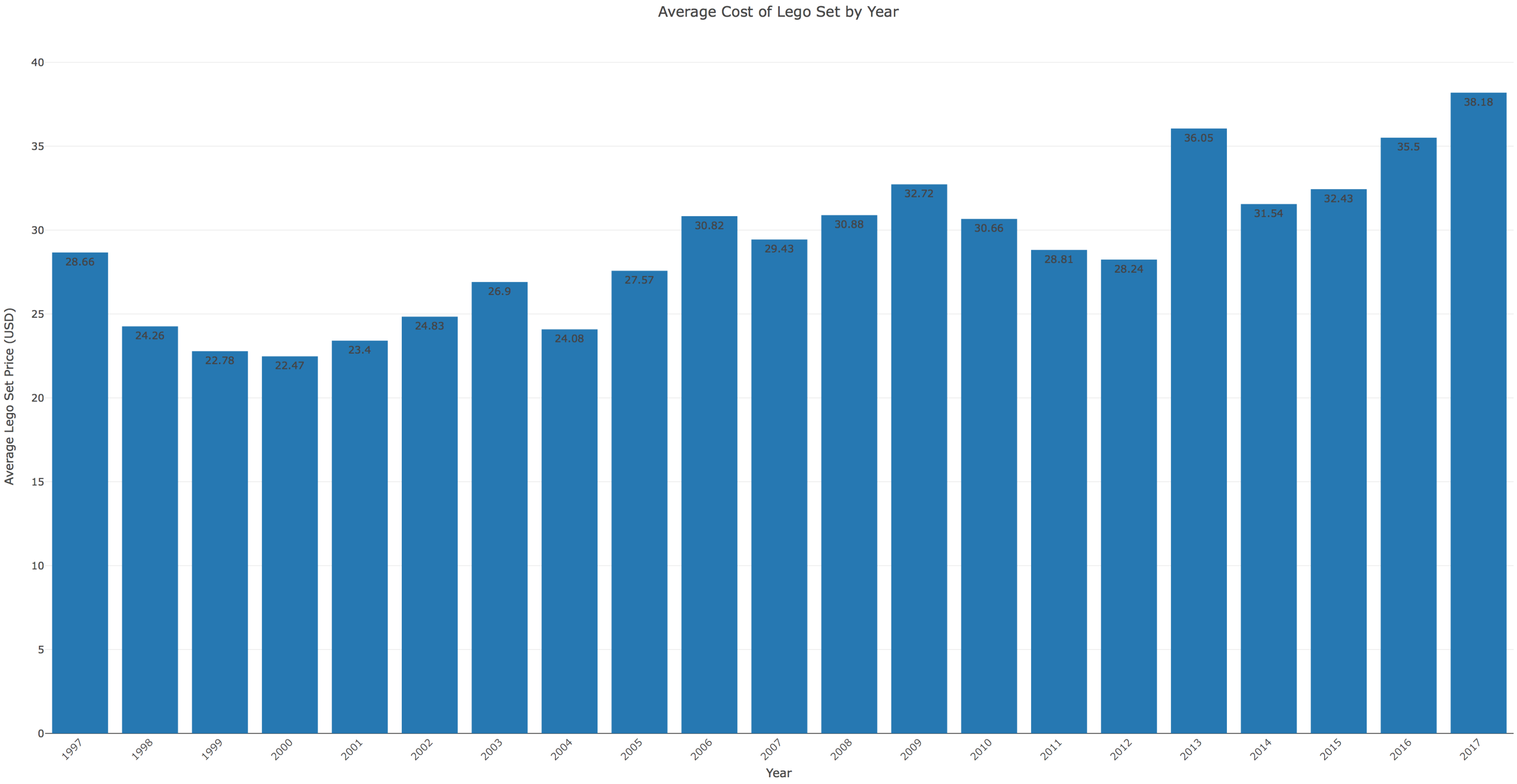
As we can see, the model does fairly well. The R-Squared is .75, which means that 75 percent of the variation in the price can be explained by our model. Also we can see that our independent variable, Pieces, as well as the F statistic for the model itself, are highly significant (F-statistic and t-statistic have the same p-value in a simple linear regression).

To visualize the data a little better and to see this linear relation a little bit better, we can use a scatter plot where the x-axis are the number of pieces and the y-axis is the price. Here is a scatter plot of our data:



6 Question Two

For this question I used a bar chart to evaluate the change in the average price of a lego set throughout the years 1997 to 2017.



As we can see, there is fluctuation (or seasonality) in the average price of a lego set between each year, but, there does seem to be a slight positive trend in the average price of a lego set as the years increase.

7 Policy Implications and Discussion

Now that we have done some analysis on the price of lego sets, this information can be used to help price future lego sets. If a new company comes into the lego set market, they can use a regression model to help them compete against other lego companies. Furthermore, if the lego set market is a competitive market, a company can see where the price of a set ought to be based on the number of pieces, and how much it costed them to make that set, and try to undercut the expected price while also maximizing their profits.

If we had learned the tools and if more time were allowed, it would have been interesting to use a decomposition time series to model the average price of a lego set for the upcoming years. Also it would have been better to do a t-test on the average price to see if there is statistically significant difference from 1997 to 2017 (rather than looking at a bar chart).