Project Report

Lego Sales

Samuel Acker

1. Introduction

Lego has been around for almost a hundred years and has been a staple in the toy section while also maintaining a massive presence in the adult collectable market. There are lego sets that come in all sizes, shapes, complexity, and different themes. How does the price of Lego sets react under specific parameters? Does the number of pieces cause a higher price for a set? Maybe it's the number of minifigures that will cause prices to increase? It could be the theme being a big factor in price, what themes have the highest average price? Which themes have the lowest average price? Age of the set could have significance, what are the average prices throughout the years?

In this project I plan to find what variables correlate positively and negatively with secondary market price and Retail Price of lego sets. I will do this by joining data scraped from BrickSet.com with data on secondary Lego market prices from Mendeley.com.

2. Data

This project uses data from two sources: BrickSet list of Retail prices for each set, and Mendely's dataset of secondary market prices.

2.1 *Retail Prices*

I scraped data from BrickSet which contained 'Id', 'Theme', 'SetName', 'USRetailPrice', 'Pieces', and 'Minifigs'. Across 64 pages on Brickset.com I wrote a web crawling to scrape the data from all these pages and got a little over 10,000 data instances. This data was saved as 'Brickset-List.csv', and this could be found in my project submission. The web crawling script is contained as 'Brickset_Scrape', and could be found in my project submission. To clean the data I will have to make NaNs in 'Minifigs' to zero.

2.2 *Secondary Market Prices*

The website data.mendeley.com allows users to post data and this data set is a culmination of Lego sets and their secondary market price for each month from Jan 2018 - Apr 2019. The price taken is from brickpicker.com. This data contains each set's: 'Id', 'Theme', 'Name', '# of pieces', and '# of minifigures', for price columns they're labeled '1/1/18', '2/1/18', and so until '4/1/19'. There are 2322 Lego sets in this data spanning releases from 1981-2014. To clean this data, the price columns originally show up in python like '2018-01-01 00:00:00', so I will drop the 0s from the end. The data was saved as 'DiB whole sample Jan 2018 - Apr 2019 prices.xlsx', and could be found in my project submission.

2.3 *Merging Retail and Secondary Market Prices*

Since both datasets include the Set Id it was easy to merge the data together using the Id from each set. I did an inner merge because I want to compare all the sets that are common between both datasets for consistent analysis between retail and secondary market prices. This inner join limits the Retail Price data because it has 8000 more instances in the data than the Secondary market data. I named the merged data BigBrick and this is the data that

will be referenced in the code.

Table 1 Data Dictionary

Column	Type	Source	Description
Id	Numeric	Both	The Lego ordering system for each release
Theme	Text	Both	The theme each set is apart of
Secondary Market Price	Numeric	Mendeley	The secondary market price for each set throughout the months for 2018-2019
USRetailPrice	Numeric	BrickSet	Price for retail purchase
Pieces	Numeric	Both	Number of pieces each set contains
SetName	Text	Both	The name each set is released as
# Of Minifigures	Numeric	Both	Number of Minifigures each set contains
Year of release	Numeric	Mendeley	This is the year in which the lego set was released

3. Analysis

3.1 *Themes*

To find out the themes with the highest average price I created a data frame called theme_avg_price and grouped by 'Theme' and calculated the mean for 'USRetailPrice'. For secondary market prices I have to extract the monthly price data using a for loop, and then I calculated the mean for the sets with consolidated monthly data. I plotted both using a matplot bar chart.

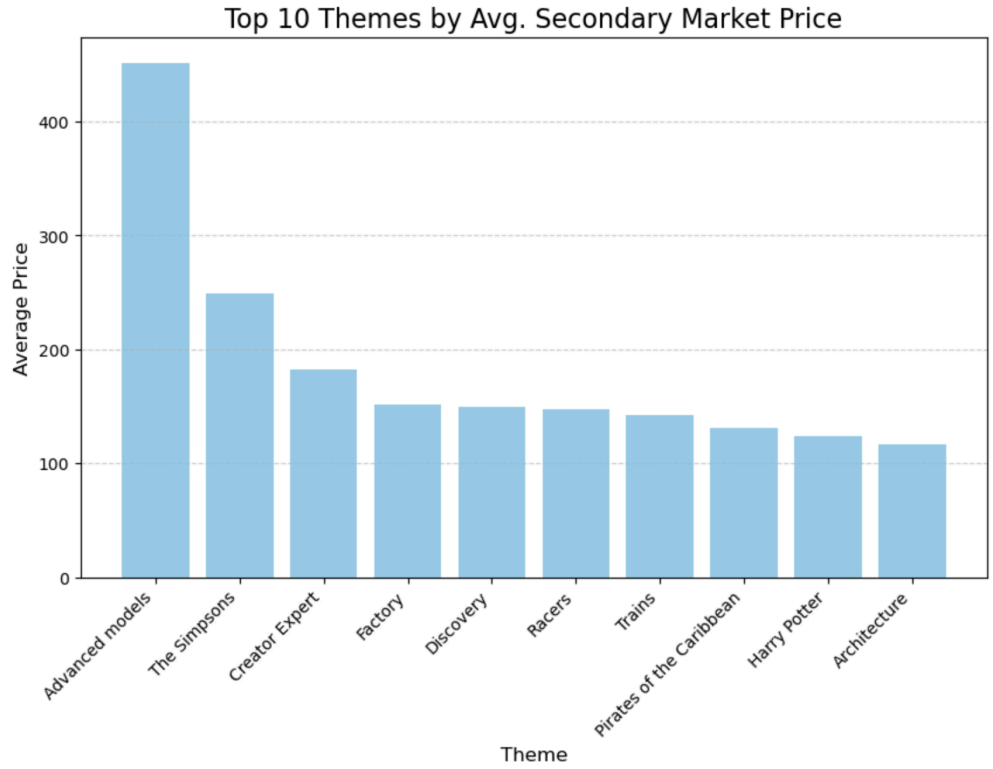


Figure 1 Bar Chart of Top ten themes by Avg. Secondary Market Price

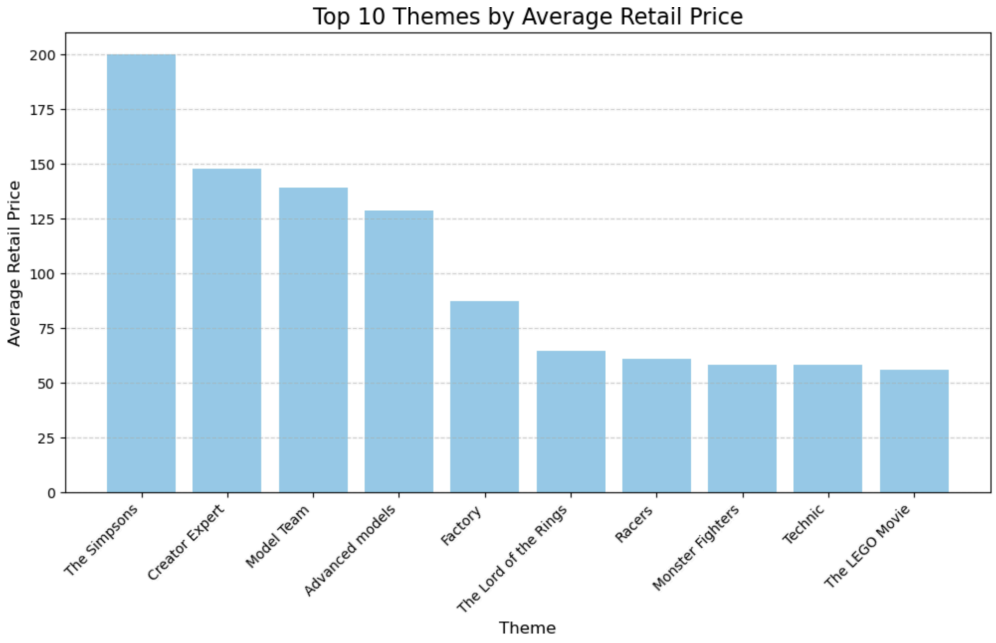


Figure 2 Bar Chart of Top ten Themes by Avg. Retail Price

Looking at these two charts we can find similar high priced Themes such as; 'The Simpsons', 'Creator Expert', and 'Factory'. Interestingly both charts feature two different movie franchises, Figure 1 contains 'Pirates of The

Caribbean' and 'Harry Potter', while Figure 2 includes 'The Lord of The Rings' and 'The Lego Movie'. This shows a difference between what movie theme's are held as more expensive for the secondary and retail market.

Table 2 Bottom 10 Themes for Avg. Secondary Market and Retail Price

	Secondary Market	Retail Price
1.	Life of George -$6.12	Seasonal -$9.04
2.	LEGO Universe -$13.79	LEGO Universe -$9.99
3.	Books -$15.18	Make and Create -$10.00
4.	Seasonal -$18.70	Power Functions -$10.99
5.	Power Functions -$21.43	Miscellaneous -$14.21
6.	Miscellaneous -$26.10	HERO Factory -$14.64
7.	Make and Create -$26.92	Bionicle -$18.73
8.	Legends of Chima -$27.62	Books -20.00
9.	Hero Factory -$30.13	Friends -$26.68
10.	Prince of Persia -$34.07	Cars -$28.26

These bottom ten themes reveal a trend for lower priced themes, themes lacking complexity such as 'Seasonal', 'Make and Create', and 'Books', and themes from medium popular properties like 'Legends of Chima', 'Prince of Persia', and 'Bionicles' will be more inexpensive than other sets. This makes it clearer to understand the top 10 popular themes because it includes massive popular properties such as 'The Simpsons', 'Harry Potter', and 'lord of The Rings', while also highlighting the themes with more complexity such as 'Advanced Models', 'Technic', and 'Creator Expert' as being more expensive.

The secondary market and retail avg prices for the themes contain similar trends in both markets for what makes a theme expensive, complexity in building, existing popular property, and what makes a theme inexpensive, simple in building, semi popular existing property.

3.2 *Number of Pieces*

Naturally one would think the more pieces in a Lego set, the higher the cost to make, and thus price will be higher for the set. To explore this premonition I wanted to find the correlation between Price and number of pieces and plot it on a scatter plot.

Correlation between Retail Price and Number of Pieces: 0.93

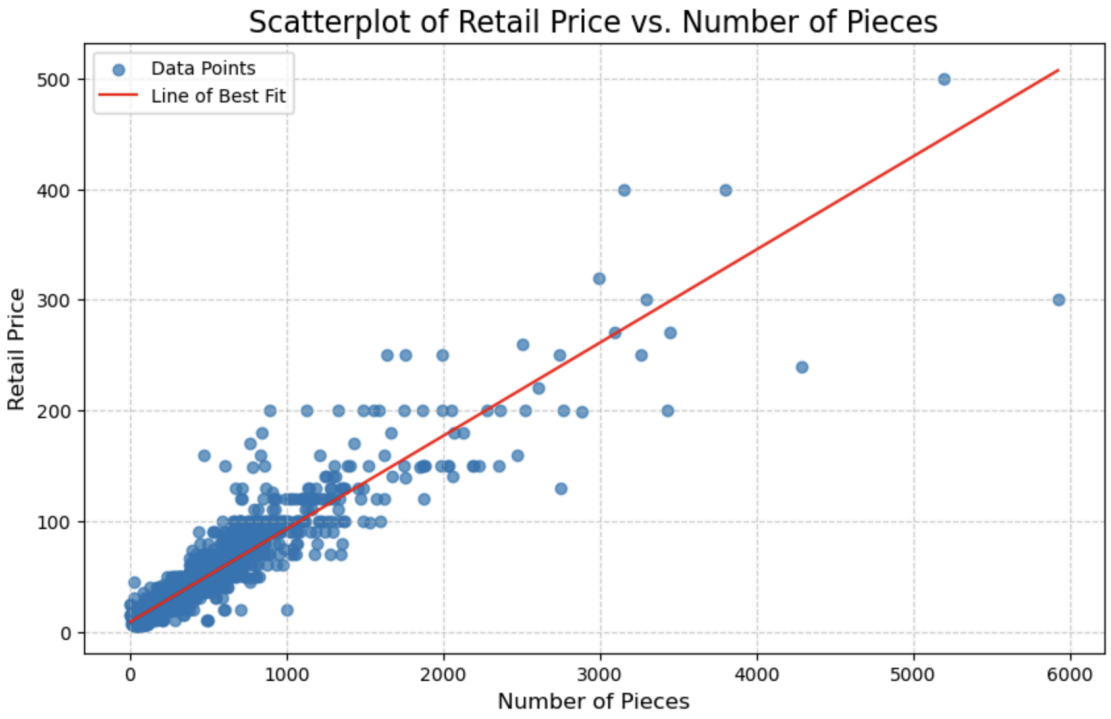
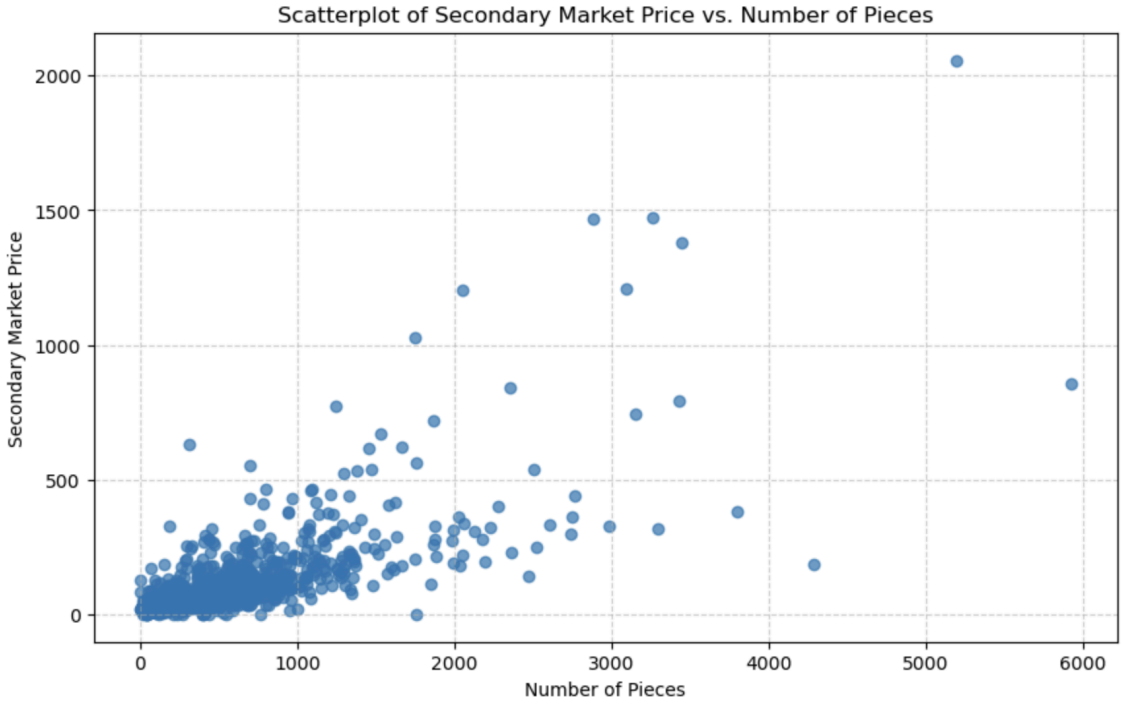


Figure 3 Correlation and Scatter Plot of Retail Price and Number of Pieces

Figure 3's Scatter Plot illustrates a positive linear relationship between the retail price and number of pieces. A Correlation of .93 also confirms our visual estimate of the number of pieces being a significant variable in the retail pricing for Legos.




```
Correlation between Secondary Market Price and Number of Pieces: 0.76
```

Figure 4 Correlation and Scatter Plot of Secondary Market Price and Number of Pieces

Looking at the scatter plot in figure 4 we again see a positive linear relationship between the secondary market price and number of pieces, but with a small amount of outliers at the 500 dollars and 2000 pieces mark. The correlation for secondary market price and number of pieces is a strong .76 but isn't an infallible measure.

Number of pieces is a stronger predictor for the retail price than in the secondary market because it has a higher correlation closest to 1. This could be because when buying second hand the cost of pieces isn't factored into the price as it is for retail pricing, and the secondary price could be more dependent on themes or just overall popularity of the set itself.

3.3 *Number of MiniFigures*

Lego sets all contain pieces but not all have minifigures, I wanted to see how this affected the price for both markets. I visualized the secondary market and retail price with number of minifigures in a scatterplot and found correlation between these variables.

Correlation between Retail Price and Number of Minifigures: 0.39

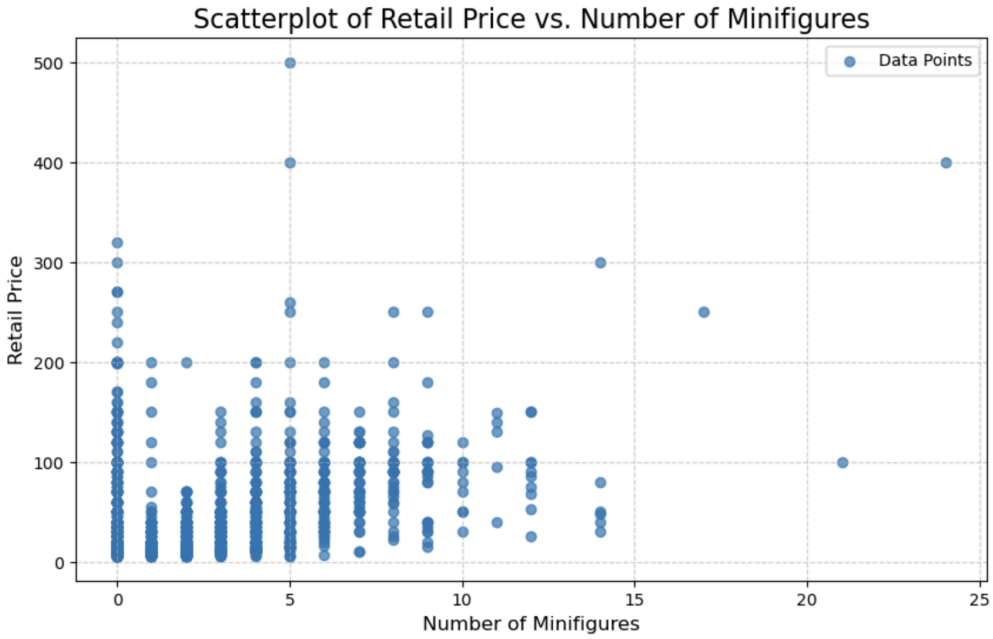
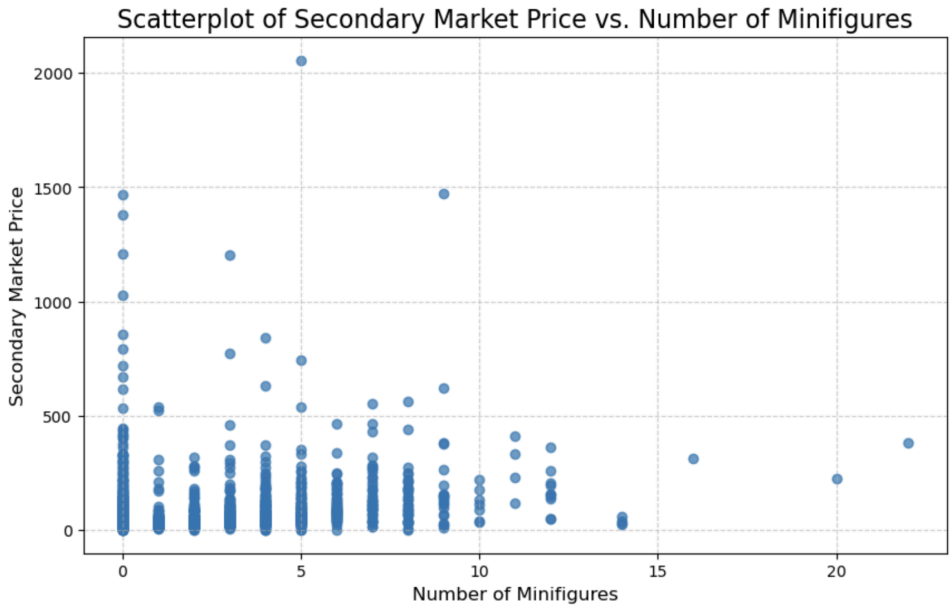


Figure 5 Correlation and Scatter Plot of Retail Price and Number of Minifigures

In this scatter plot we can see a small positive relationship for number of minifigures and retail price, with outliers skewing the relationship. The correlation(.39) is smaller than the general threshold of .5 meaning there is little correlation between these variables



Correlation between Secondary Market Price and Number of Minifigures: 0.19

Figure 6 Correlation and Scatter Plot of Secondary Market Price and Number of Minifigures

In Figure 6 the scatter plot depicts a negative relationship for secondary market price and number of

minifigures. Interestingly, we can observe significantly higher prices for sets with 0 minifigures. Looking at the correlation we see no correlation with a value of .19.

Number of minifigures is a metric that isn't a strong indicator for price, and in fact based on the findings it isn't an indicator for a higher price. Between the markets, retail prices show more correlation than secondary market prices with the number of minifigures, this could be because the cost of producing more minifigures is included in the retail price while secondary market price is found without cost of production. Another reason why price and number of minifigures aren't correlating is because the quality aspects of a minifigure matter more than the quantity, and so this is why we don't have a relationship with the number of minifigures, as we do with the number of pieces.

3.4 *Year Released*

How old a Lego Set could be a factor in secondary market prices, older rarer sets logically seem like it would cost more to purchase than a newer, more available set. Year released doesn't seem as relevant to analyze for retail price, because you won't be able to buy sets at retail price from 2004 respectively in the present, but we can look at an overall trend for retail pricing over the year.

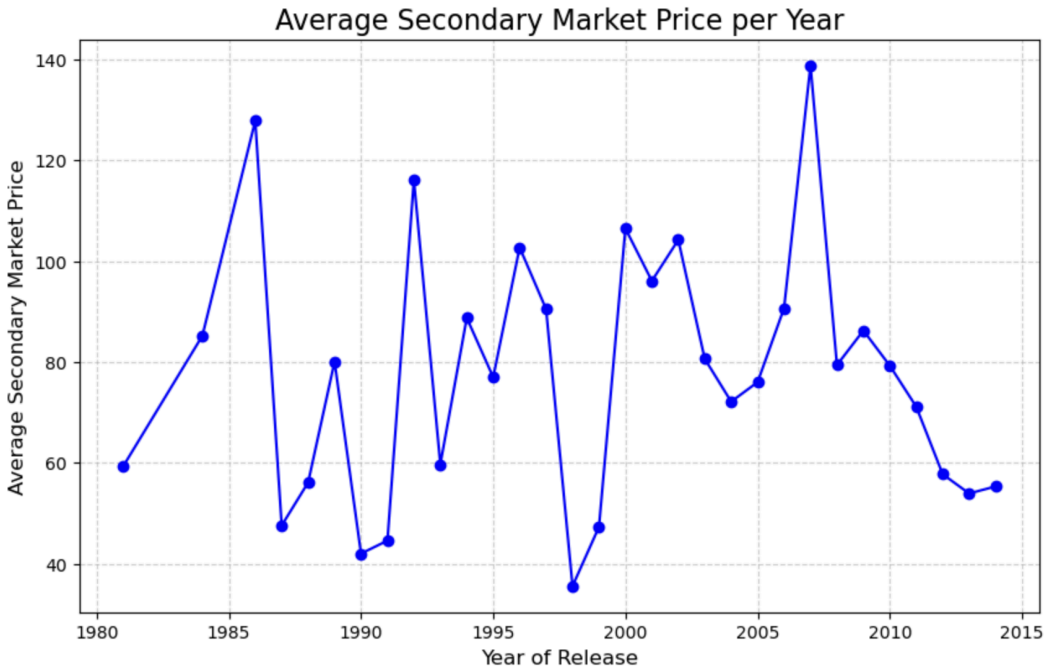


Figure 7 Line chart of Avg. Secondary Market Price per Year of Release

The line chart in figure 7 reveals a ton of variance in average price for each year, with high highs ,and low lows. This lack of a trend in avg. price over the years leads me to think years old is not an indicator of an expensive lego set. The avg. price per year is instead highly dependent on what massively popular set was released.

Table 3 Top 3 sets by Average Secondary Market Price in 2007

Set Name	Avg. Secondary Market Price
Ultimate Collector's Millennium Falcon	$2057.11
Cafe Corner	$1203.60
Eiffel Tower	$792.10

Table 4 Top 3 sets by Average Secondary Market Price in 1998

Set Name	Avg. Secondary Market Price
Flying Ninja Fortress	$127.76
Samurai Stronghold	$44.62
Ninja Surprise	$30.30

Diving deeper into the years with the highest (2007) and lowest (1998) avg. secondary market price we can find obvious differences between table 3 and 4. Table 3 explains the high value for 2007, there was a massive release in the Ultimate Collector's Millennium Falcon, many Star Wars fans dream of owning it, and so it is easy to understand why it is so expensive. All three of those sets are extremely complex with thousands of pieces and in turn they fetch exorbitant prices. While in table 4 it illuminates why 1998 has such a low avg. price, these are simple sets lacking an existing popular property theme.

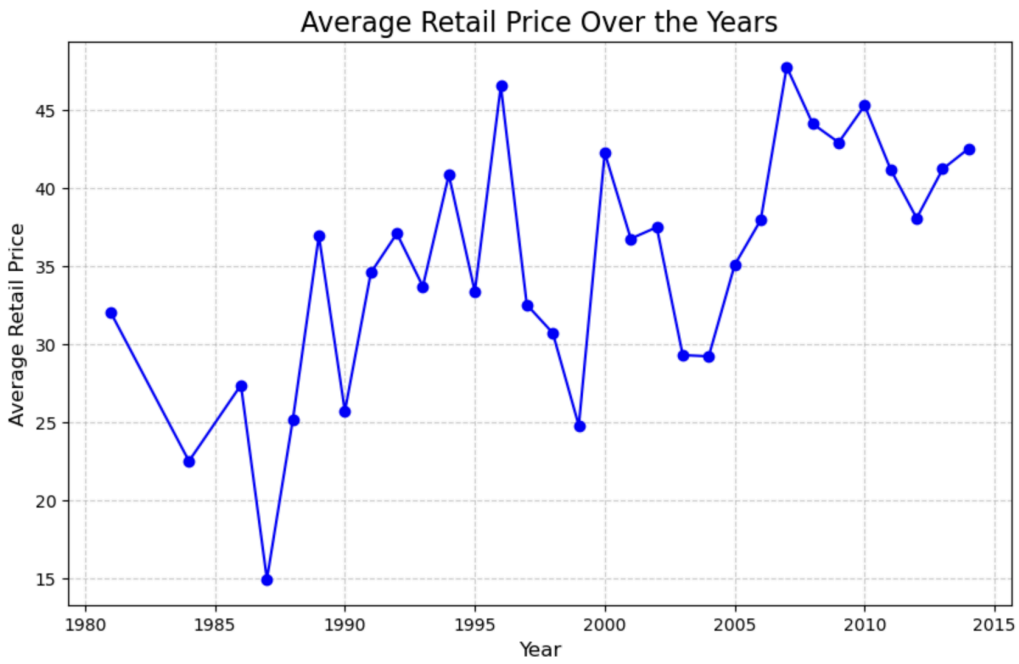


Figure 8 Line chart of Average Retail Price per Year of Release

Over the years we see variance for the average retail price, but overall there is a linear positive trend for price as year increases. This rise in prices could be due to inflation or overall cost of production increasing as years go on.

Overall for the secondary market I found no trend in year of release affecting avg price, but instead finding years with more complex sets, and massive pop culture tie-ins being a distinguishing factor between a year having a high or low avg price.

4. Conclusion

In this project, I analyzed four aspects of Lego's Retail and Secondary market sales: Number of Pieces and Minifigures, Release Date, and Theme affect the price of each Lego set. In short, from the analysis questions in the proposal, I found the following results.

1. **What themes make the price higher or lower?**

   For both the Retail and Secondary Market prices, the trend was the same. Themes that are complex and or are an existing popular property (e.g The Simpsons) will be higher priced, than themes that are simple and or are from existing semi-popular properties (e.g Prince of Persia) will be lower.

2. **How does the number of pieces in a set affect the price?**

   For both the Retail and Secondary Market prices, they had similar trends. Number of pieces has a very strong correlation(.93) with a higher retail price, and a strong correlation(.76) for a higher secondary market price. For both markets the more pieces there are in a set, the more likely the set will have a higher price.

3. **How does the number of minifigures in a set affect the price?**

   In the Retail market there was little correlation(.39) between price and number of minifigures, and in the secondary market there was little to no correlation(.19) between these variables. For retail more minifigures could raise production costs, causing a higher price. Both markets show the number of minifigures has little effect on prices.

4. **What year has the highest and lowest average price?**

   In the secondary market the highest avg price per year was 2007 being around $140, and the lowest year being 1998 around $38. In the secondary market there is no trend for year of release affecting price. The retail market shows avg. prices rising steadily over the years, leaving the $25 range and entering $40+.

The limitations to this project are; the data only contains Lego sets from 2014 or earlier leaving out nearly ten years of data, and because of the inner merge the data is capped for what appears only in the Secondary Market and Retail data, leaving around 8000 lego sets out of retail price analysis.