# Mind the Brick

**Lego as Alternative Investment: Predicting the Future Value of LEGO Sets Using Machine Learning.**

Yorran van der Slik

Student Number: 2030421

May 13th, 2019

Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Master of Science in Data Science & Society.

Thesis committee:

Dr. ir. M. Van Otterlo
Dr. M. Van Zaanen

Tilburg University

School of Humanities and Digital Sciences

Department of Cognitive Science & Artificial Intelligence

Tilburg, The Netherlands

May 2019

# Preface

Over the past few months, I was lucky enough to re-indulge myself into the wonderful world of LEGO, as part of this work that concludes my study at Tilburg University. Not only did it bring back fond memories of a visit to LEGO Land as a kid, but it also brought me new LEGO bricks and an hours-long lasting smile after watching the 2017 LEGO Batman Movie. None of this would have been possible without the assistance, guidance, and support from several individuals whom I would like to thank.

First and foremost, I would like to thank my supervisor during the thesis trajectory, Martijn van Otterlo. I would not have come-up with LEGO as a topic, but thanks to Martijn's passion for the brick and guidance throughout the process, LEGO has earned an eternal sweet spot in my heart. Second, I would like to thank Brick Insight founder, Linus Bohman. Without his work on Brick Insight and the corresponding API, this research would be lacking several features. Also, Linus' MOCs on Flickr are worth a look and an inspiration for any LEGO fan. Thirdly, I would like to thank my family and friends for their continuous support throughout my educational adventures. It has been quite the ride.

I hereby state that this research has solely been written and conducted by me. The author is solely responsible for the content of this paper, including any mistakes, opinions, and interpretations. The copyright rests with the author. The Department of Cognitive Science & Artificial Intelligence, School of Humanities and Digital Sciences, and Tilburg University in general, cannot be held liable for the content of the author's thesis.

Yorran van der Slik

May 13th,2019

Star Wars & LEGO, including all names and other related items, are registered trademarks and/or copyrights of Lucasfilm Ltd., and/or the LEGO GROUP or their respective trademark and copyright owners. This thesis is not endorsed by the LEGO GROUP and/or Lucasfilm Ltd.

# Abstract

Investors increasingly look for new possibilities to gain increased financial returns. One way of doing so, is by exploring the possibility of investing in alternative investments. Recently, LEGO, yielding above market average returns, has been linked in literature with alternative investments. However, it is unclear which LEGO sets yield great financial returns. This study sets out to examine the financial performances of LEGO as an alternative investment and studies the possibility to predict the future value of LEGO. It does so by incorporating a broader data science perspective using various data sources, as it does not just evaluate LEGO's overall performance, but evaluates on a set and theme specific level. Additionally, this study includes different Machine Learning methods that allow for future value prediction of LEGO sets. The results of this study suggest that *Brickheadz*, *Avatar* and, *Ninja* themed LEGO products yield the greatest average annual returns, whereas on a product-specific level, space and superheroes related products yield above-average financial returns. Additionally, this research proposes a two-degree Polynomial model that predicts the future value of a LEGO set that allows investors and collectors alike to determine the value of LEGO sets used as an alternative investment. This study contributes to the small body of literature that categorizes LEGO as an alternative investment and sets a baseline for future LEGO value predictions.


*Keywords:* alternative investments, LEGO, prediction, machine learning.

# Table of Content

**Preface** ..... 1

**Abstract** ..... 2

**Table of Content** ..... 3

**Chapter One: Introduction** ..... 4

**Chapter Two: Literature Review** ..... 6

- **Background on LEGO** ..... 6
- **Investments** ..... 8

**Chapter Three: Data and Methodology** ..... 14

- **Data Sources and Collection Methodology** ..... 14
- **Feature Selection** ..... 16
- **Missing Data** ..... 18
- **Data Partitioning** ..... 20
- **Overview of Software** ..... 21
- **Algorithms and Evaluation** ..... 22

**Chapter Four: Results** ..... 24

- **Financial Performance of LEGO Products as Alternative Investment** ..... 24
- **Predicting Future Value of LEGO Products** ..... 28

**Chapter Five: Discussion and Conclusion** ..... 31

- **Managerial Implications** ..... 32
- **Limitations and Further Research** ..... 33
- **Conclusion** ..... 34

**Reference** ..... 35

# Chapter One: Introduction

The LEGO Company, founded in 1916, recently celebrated its 100th birthday. The company, originally founded as a carpentry shop, only began manufacturing wooden toys in 1932, while the LEGO brick that we know the company for today, was not introduced until 1949. The plastic moulded bricks, available in more than 140 countries, are a long-term favourite among children and parents, arguably for the creativity that is inherent to LEGO. LEGO is not just one of the world's largest toy producers, it is also the brand behind multiple theme parks, the main character in multiple LEGO movies, and the innovator of a completely new way of learning. Asides from being invented by a toy manufacturer after the Second World War, LEGO is not just a toy for children. Increasingly, we witness adults who remain interested in the Danish blocks and keep spending large sums of money on newly developed sets and spend hours on specialized internet forums to discuss LEGO. Also, second-hand LEGO is being traded at many internet platforms, and some prices are way higher than the original retail price, with products frequently selling for more than triple their initial retail value (Andrews, 2017). However, it is partially unclear why certain LEGO sets are traded for these premiums.

These premiums have gone unnoticed to research for a long time, and LEGO sets have only recently been linked in academic research as an alternative investment (Dobrynskava & Kishilova, 2018). Alternative investments are investments not made in cash, bonds or equity, but rather in collectibles, often lacking an immediate open trading platform (Skully, 2007). Despite being less common investments, the alternative investment market is thriving (KPMG, 2018; Vaitkune, 2018). However, it remains unclear if LEGO could be categorized as an alternative investment. Regardless, LEGO seems to outperform other investments with LEGO sets yielding 18.5 % (p. 15, Dobrynskava & Kishilova, 2018) between 1987 and 2015, outperforming the S&P 500. Even more interesting, these market returns sustained over the financial crisis when many other investments saw negative returns (Schroders, 2017). As LEGO is not a publicly traded company, investing in LEGO has to be done by more traditional means, such as second-hand internet platforms, making extensive financial analysis difficult.

Much remains to be discovered about the premiums for which those LEGO sets are being traded. It is especially unknown which LEGO themes and sets are most likely to result in profitable future returns and what the characteristics are of these LEGO sets, as Dobrynskava and Kishilova (2018) only looked at the overall financial returns.

Therefore, current research is two-fold. It first evaluates the financial performance of LEGO themes and sets as an alternative investment. Afterward, this research continues with the main goal of this research, as it aims at defining a machine learning algorithm that predicts the future value of a LEGO set. Per this aim, the main research question for this thesis is as followed:

*"To what extent can we predict the future value of LEGO investments given the LEGO set characteristics?*

For academics, this research shows value in the field of alternative investment. To the best of my knowledge, not much has been written on LEGO as an alternative investment, with the exception of Dobrynskava and Kishilova (2018). Additionally, in research by Cummings and Zhang (2016) aiming to find the variables influencing the later value of an alternative investment, LEGO has not been mentioned. Media has covered LEGO as investment in the past (Hyde, 2015; Bradley, 2016), but the number of articles spiked after the study by Dobrynskava and Kishilova (2018) was published (Burton, 2019; Popina, 2019; Visser, 2019), suggesting increased attention from the greater public. Therefore, current research gives a potentially fruitful new avenue for researchers and investors interested in LEGO as an alternative investment. This research differs from the existing work by Dobrynskava and Kishilova (2018) who calculated the overall financial returns, as it uses a broader perspective and looks at the set and theme specific financial returns. For practitioners and collectors alike, the outcome of this research may show of value, as it provides valuable insight into LEGO sets and themes yielding above market financial returns, and helps them determine whether alternative investments in LEGO are a good opportunity for diversification in their portfolio and for future returns. Additionally, this research enables investors and collectors to predict the future value of a LEGO set, helping them evaluate prospects in alternative investments.

Before being able to measure the financial performance of LEGO as alternative investment and predict the future value of LEGO sets successfully, this research starts with a literature review elaborating on alternative investments in chapter two. This is followed by chapter three, which presents methodology and data collection methods of this research. This data will then be used to analyse the financial returns of LEGO sets as an alternative investment and to create a supervised Machine Learning algorithm that will predict the future value of a LEGO set, which will be presented in chapter four, the findings. Chapter five then discussed the findings, limitations and presents further research possibilities.

# Chapter Two: Literature Review

This chapter covers existing literature that this research builds upon. The first section covers the LEGO Company and its products, as well as research to changes in LEGO products. The second section is an examination of existing research on alternative investments, based on which a framework is purposed. The final section of this chapter covers measurements of future returns for investments. The literature provided in this chapter sets the framework and boundaries for current research.

## Background on LEGO

Ole Kirk Christiansen (1891 – 1958), a carpenter born in Billund, Denmark, established what is known today as the LEGO Company (Lauwaert, 2008). Originally, Christiansen was into manufacturing traditional Danish houses and furniture, but later changed to the manufacturing of smaller objects and wooden toys. After the Second World War, the company reformed and started experimenting with plastic, ultimately resulting in the plastic brick we know today as LEGO (LEGO.com, 2017). The name LEGO is a contraction of the Danish words '*leg godt'*, which translate to '*to play well'* and is still in use today. Despite facing ultra-demanding customers, such as ten-year old boys (Robertson & Breen, 2013), having financial setbacks (The Guardian, 2018) and operating in a branch with few entry-barriers and many substitutes, LEGO is currently among the most recognizable brands in the world (Kauflin, 2017). Additionally, LEGO is the most valuable brand in the toy industry, outperforming Fisher-Price and Bandai-Namco (Brand Finance Toys, 2018).

To remain competitive, LEGO's management adjusted the strategy to changing markets, by adopting a user-driven innovation culture. This change marks a crucial shift in the design of new LEGO products by including the key stakeholder into the design, namely the consumer (Lauweart, 2008; Ind, 2014; Robertson & Green, 2013). Additionally, LEGO's management noticed a change in consumers. Where the LEGO bricks lost interest from kids in the age range of 7 to 17, they won back in the age group of 17 – 47 as a result of *'adults staying younger for longer'* (Lauwaert, 2008, p. 18). Research into this last group found a fan base, which was described by Baichtal and Meno (2011) as a cult, with people re-creating gigantic battleships with LEGO and even dressing up as a LEGO mini-figures. Because of the popularity among adults and user-driven innovation, LEGO launched various new products, including a new line featuring models of architecturally significant buildings, which include the Empire State Building and Berliner Brandenburg Gate (Antorini et al., 2012).

Changes in LEGO's assortment go further than new additions to the collection aimed at adults. Instead of having solely generic themes, such as City, LEGO now includes themes that feature more trend products, for example related to movies. Additionally, themes are now *"increasingly based on conflicts",* such as Harry Potter and Pirates of the Caribbean (Bartneck et al., 2013, p. 8). These conflict-based LEGO products also have effect on the overall products in the assortment. In today's LEGO sets, almost 30% contain at least one weapon of some sort and this number is increasing (Bartneck et al., 2016). Additionally, the inclusion of new themes – both conflict and non-conflict related -, also leads to the creation of new special LEGO pieces and mini-figures, such as Harry Potter himself. The variety of one of the most crucial parts of a LEGO set, being the mini-figure, is increasing steadily and "*if the current [increasing] trend continues, then soon every Mini-figure in every set will be unique"* (Bartneck et al., 2013, p. 8). Furthermore, in-line with the themes - also the mini-figure faces show some form of 'violence', with the number of happy faces decreasing and angry faces rapidly increasing (Bartneck et al., 2013). Thus, arguably LEGO sets are turning violent, which is on par with a global trend of toys becoming more violent. Similar results have been found with competitor Playmobil (Bartneck et al., 2016).

In general, independent of the increased violence, LEGO sets have become more complex. In researching all LEGO sets that were produced between 1955 and 2015, Bartneck and colleagues (2018) found that LEGO sets have become bigger with the biggest LEGO set now being the UCS Millennium Falcon featuring over 7,500 LEGO pieces, beating the Hogwarts Castle and Taj Mahal (Smith, 2018). Additionally, despite seeing a dip around the financial crisis, LEGO sets have become more specialized (Carron, 2016), arguably because of the new themes. Besides more complex, the sets also display more colour variation, with colour variety increasing 2.4% per year (Bartneck et al., 2018).

When a new set is launched through mainstream channels such as (online) retailers, sets are picked up by consumers and collectors alike. However, once a set is not available through the main stream channels anymore, for example because it is discontinued or only sold as limited edition, the price of that set tends to jump up significantly on the second-hand market (Dobrynskava & Kishilova, 2018). This has not gone unnoticed, especially with media covering high-value sales (f.e. Bradley, 2016; Burton, 2019), leading to collectors looking for ways to use their LEGO as an investment.

## Investments

Following collectors that see LEGO as a future investment, current research argues that LEGO is an alternative investment. Before going in-depth on alternative investments and why LEGO fits the characteristics of alternative investments, this research first establishes what a traditional investment is.

***Traditional Investments***. As a definitive term of traditional investments is lacking in academic literature, this literate section looks at existing classifications of investments. The most general classification of investments is based on the nature of the claim of the investor. Anson and collages (2010) distinguish two types. The first type of investment is a *debt instrument*, where the issuer agrees to repay the borrowed amount over a certain period and pays interest for this service. The second type is an *equity instrument*. In this type of investment, the investor becomes partly owner of the company they bought shares of and is therefore entitled to a part of the earnings of that company. In a wide context, all traditional investment are investments which offer returns in the form of dividend or interest and are traded continuously on a public financial market (Hietbrink, 2017).

As investors are looking to maximize their risk-adjusted earnings, they often combine traditional assets to create a stable and overall profitable portfolio of assets. The term risk is of importance to investors, as already in 1952, Markowitz concluded that investors should include different products in their portfolio to minimize risk. Typically, investment portfolios are made up of cash, bonds and, equities. However, increasingly in an aim to diversify portfolios, we see new types of product in investors' portfolios, which include alternative investments.

***Alternative Investments.*** In contrast, alternative investments are investments not made in cash, bonds or equity, but rather in collectibles. In a broader sense, alternative investments are assets which do not have an immediate trading platform (Skully, 2007) and are assets that '*require a premium return because they are illiquid*' (Scholes, 2000, p.17). Additionally, alternative investments have no intrinsic value, nor do they generate cash flow, making it hard to evaluate (Skully, 2007). Some researchers argue that the term 'alternative' is not relevant anymore, as many of these investments have '*become increasingly mainstream',* as some of them have been regularized (Wilson, 2007, p. 1). Investors looking for improved returns on their portfolio, but also those looking for improved risk spread, are increasingly looking at alternative investment (AON, 2016).

Often, these investments are made on preference by the investor and not solely for financial purposes (McAndrews, 2010). The reasons for buying alternative assets are wide-spread and include an individual's passion for a specific product or a more philanthropic nature such as the status gained by owning a collectable item (Stewart, 2016). This philanthropic reasoning of the purchase of an alternative investment was already described by Belk in 1995. Belk (1995) argued that alternative investments are bought because investors prefer to acquire inessential consumer goods that are rare or unique. Because of this uniqueness, investors have the opportunity to enjoy the investment. Therefore, alternative investments are also referred to as emotional assets (Sanning et al., 2006; Jurek & Stafford, 2015). Additionally, in-line with this research, emotional assets are purchased for their future returns and as a way of diversifying the investor's portfolio (Briere et al., 2014; 2015; Cummings & Zhang, 2016).

According to Baghai and colleagues (2015), the number of alternative investments options has grown twice as fast as traditional asset classes. However, in current literature, there is no unambiguous list of assets that classify as emotional assets and/or alternative investments. For example, Yau and colleagues (2007) did an attempt in categorizing alternative investment into two broad categories and many subcategories, but did not include tangible assets. The categorization made by Dimson and Speanjers (2014), only included tangibles, making a distinction between stamps, art and, musical instruments, whereas Skully (2007) distinguishes seven different groups of tangibles. Yet another categorization was purposed by Lee (2015), including land and alcoholic beverages, but a definitive classification of alternative investments is lacking. However, there is plenty of research on different potential alternative investments, ranging from arts and stamps to wines. Combining different categorizations of alternative investments in literature, this research gives a non-exhaustive list of alternative investments that will be used to classify LEGO as an alternative investment. This research so far distinguishes 3 categories, being *1)* collectibles, *2)* usables and *3)* consumables.

The first category of alternative investments are collectible assets and include many of the better known alternative investments, such as art and paintings. Arguably, art is a fruitful investment that benefits from the uniqueness of individual products (Mei & Moses, 2002; Campbell, 2009). One of the first calculations on alternative investments by Anderson (1974) already showed that art gave an above average rate of returns on investment. Over the years, much has been written on the returns of art (Campbell, 2005; Campbell, 2009), with an average return over the past 25 years of 5.2% (Renneboog &

Spaenjers, 2013). Just like art, other collectibles, such as stamps and antiques show value as an alternative investment. These products are rare, show no supply elasticity and are often influenced by their age, giving them the features of collectible (Campbell, 2009; Graddy & Margolis, 2011). Additionally, diamonds and gems can be categorized as a collectible alternative investment. Renneboog & Spaenjers (2012) found that diamonds and other gems outperformed stocks between 1999 and 2010 by returning 6.4 % annually and showed positive covariance with stock returns, underlining the importance of collectible products as an alternative investment.

The second category includes the so-called usable products. In the categorization of current research, usable products are products that can be used, but present the highest value when they remain unused and well-maintained. Often these assets are relatively rare with only a couple of assets produced. Examples of this category include toys such as Matchbox cars and musical instruments (Campbell, 2008; Graddy & Margolis, 2011; Bucur, 2016). The most promising and beloved alternative investment asset are classic cars (Barclays, 2012). As shown, investors have increasingly included automobiles as an investment in their portfolio and it is becoming a leader in the alternative investment arena (Ebeling, 2012; Pembrey. 2016). Between 2007 and 2016, investments in classical cars have shown superior annual returns, although dependent on the state of the car (Martin, 2016). Additionally, Hierbrink (2017) found that classic sports cars yield 11.52 % annually, with a special Lamborghini Countach LP 4000 Berlinetta yielding as much as 461 %.

The third category contains consumable products and includes the frequently cited alternative investment in alcoholic beverages (f.e. Sanning et al., 2006; Masset et al., 2010). One of the limitations of investing in alcoholic beverages is that they may be perishable and once consumed, have no value at all. Hence, this category is one-time consumable. The quality of many alcoholic beverages, such as wine, are highly correlated with geographical factors, such as the weather. This makes for a unique batch every year, making a good wine rare. Because of this, good wines show high returns with the top wines yielding 17,4 % annually (Fogarty, 2006). Because wines are a well-known alternative investment and show great returns, researchers have tried to predict the future value of wines. Based on costly wines from the French Bordeaux region, Ashenfelter (2005) predicted the value of wines on a variety of variables, suggesting that the year of production was most important. Even though the research was a good attempt of creating a prediction model, resulting in a 5% error rate (RMSE = $ 440.67), Ashenfelter's (2005) research was lacking a broad data perspective, something that can be solved with current data-driven research.

Nonetheless of above-purposed categorization, all alternative investments have some distinguishable features, setting them apart from traditional investments. Firstly, alternative investments are products that are tangible and have a certain use from origin, whether it is consumption or transportation (Stewart, 2016). Secondly, the market on which these investments are traded on less regulated and developed markets (Skully, 2007). Thirdly, these investments, from origin, have a form of emotion and because they are heterogeneous goods they are collectible (Campbell, 2007). Lastly, alternative investments do not generate a constant cash flow and only return financial value when they are sold, making their valuation directly linkable to supply and demand (Ollagnon, 2012). The same applies to LEGO, being a tangible product originally made to use (*'play with'*), traded on second-hand online markets and off-line gatherings, and becoming a collectible for fans and investors alike. Also, LEGO products do not give a continuous cash flow, but rather only once when sold. Thus, LEGO products match the requirements of being categorised as alternative investment. Nevertheless, until recently it has not been researched as an alternative investment (Dobrynskava & Kishilova, 2018).

***Attractiveness of Alternative Investments.*** Independent of this self-categorization of collectibles, usables and, consumables, the value of all alternative investments is determined by the exclusivity (rarity) of the products, as well as the unique features and state of the product. Additionally, none of the above-mentioned products generate a constant cash flow, with the exception when it is sold or lend. Therefore, one could question the attractiveness of alternative investments. As stated before, the attractiveness of an investment is dependent on the hand by the future return of the asset and on the other hand by the risk of the investment (Schumacher, 2016). In general, investors try to balance their portfolio with risky and safe assets (Kolm et al., 2014). Hence, investors looking to maximize their returns should consider the risks and investors looking at alternative investments for diversification should look for assets that show minimal correlation with stock market returns (Burton & Jocabsen, 1999). However, even though the general returns of alternative investments out-performs traditional investments, investors should consider the costs of trading in these markets and the dangers and pitfalls (Dimson & Spaenjers, 2014). Additionally, many alternative investments require long-term storage space and bare the risks of storage, such as theft and damage (Campbell, 2007).

***Measuring Success of Alternative Investments.*** Investment decisions are often made based on future returns in comparison to taken risk (Brealey et al., 2012). As traditional assets normally generate cash flow, this cash flow can be used to evaluate investments. For example, one popular method to measure the investment's profitability is the often-cited Return on Investment (ROI) metric, which is the financial return (profit) divided by the investment cost. In general, when the ROI is positive, the investment is probably worthwhile, depending on the risk taken (e.g. Browne, 1998; Erdogmus, Favaro & Strigel, 2004; Bodnar & Schmid, 2007; Ichsani & Suhardi, 2015). However, ROI and many derivatives of this metric, have limitations which undermine its usage as metric for alternative investments (Baker & Filbeck, 2013). Additionally, many standard measures implicit a cash flow in an investment, which is not the case with most alternative investment (Knight Frank, 2018), nor for LEGO products (Dobrynskava & Kishilova, 2018).

Therefore, Burton and Jacobs (1999) suggest three methods for measuring the future returns on alternative investments, by *averaging,* creating a *hedonic regression*, and creating a *repeated sales regression*. The first method, averaging over the returns, is one of the most basic approaches, but allows for a quick comparison and overview. Regardless of its drawbacks, the metric in general scores well to allow investors to select products, such as with paintings (Burton & Jacobs, 1999). The second method of deriving a measurement for alternative investments is a hedonic regression, which is a regression on investment specific attributes, such as age and purchase price. However, hedonic regression models typically include many variables which lead to various problems such as multi-collinearity making analysis complex (Burton & Jacobs, 1999; Masset & Henderson, 2010). The third method is that of repeated sales regression, which is frequently cited in asset investments, such as real estate (Jones, 2010; Peng, 2012; Nagaraja, Brown & Wachter, 2014). The underlying idea is to consider goods that are very similar and that have been traded at least twice during the period under consideration. Based on this, the return between these two moments can be calculated. However, this method is very prone to outliers as these investments are not sold often and re-occurring. Additionally, the method relies heavily on the definition of '*similar*' products (Masset & Henderson, 2010).

Out of the methods proposed, this research focusses on using the *average annual return* of a LEGO set, using the average price on the second-hand market, while averaging over the number of year between the year of release and the years of sale. This method has been used with success before for various alternative investments, including wines (Masset & Henderson, 2010) and diamonds (Small, Smith & Small, 2012). Additionally, in a

predictive research to stamps Dimson and Spaenjers (2011) were able to successfully predict values within a 5% error margin. The formula of the annual returns based on the average observed price is as follows:

$$Average\ annual\ return_i = \frac{Average\ Price\ Second\ Hand\ Market_i - \ Retail\ Price_i}{Year\ of\ Sale_i - Year\ of\ Release_i}$$

Additionally, investors prefer a measurement which compares the returns with potential returns elsewhere, the so-called '*cost of capital'* (Brealey et al., 2012; Vlaović-Begović et al., 2013). To compare the returns of LEGO investment, this research uses the average return of the S&P 500 at this time, as suggested by Gollier and Weitzman (2010). The cost of capital used in this research is 5.1% as per Yardeni and colleagues (2019). With this cost of capital, the minimal required return value can be calculated, also known as the Net Present Value, or NPV (Magni, 2009; Breley et al., 2012). The NPV calculation has been used in practice and research before and is arguably one of the better methods to evaluate investments (Kort, 1990; Ross, 1995; Götze et al., 2008; Magni, 2009). The NPV starts negative with the investment in the LEGO set (purchased at retail price) before adding the incoming cash flow, being the average price of a LEGO product on the second-hand market, and discounts this with the cost of capital. As LEGO investments only have a one-time return in cash flow, the formula is as follows:

$$Net\ Present\ Value_i = \ -Retail\ Price_i + \frac{Average\ Price\ Second\ Hand\ Market_i}{1.051^{(Year\ of\ release_i - 2019)}}$$

This formula will be used to compare LEGO sets that have a positive NPV (and thus are potential fruitful investments) with products that do not have a positive NPV to see how these products differ.

# Chapter Three: Data and Methodology

The previous section provided a summary of academic literature related to the LEGO, alternative investments and evaluation methods. This chapter builds upon that knowledge and introduces the data and methodology of this research. It starts with an overview of the data sources and methods used to collect the data. This is followed by a description of the final data frame, which will be used later to predict the values. After this, features are analysed, giving some initial insights into their predictive power. Based on these features, missing data is analysed and data is partitioned. This chapter ends with an explanation of the software and models used to predict the future value of LEGO sets.

## Data Sources and Collection Methodology

This research builds upon various data sources. The first data source is a collection of files from rebrickable.com. The collection of files, all in .csv format, form the start of creating a dataset that will allow for prediction and gives an overview of all existing LEGO set IDs, including games and books. Based on this, the theme of every particular set could be added. As the initial theme was quite vague, such as '*Lion Knights'*, all sub-themes were altered to their parent theme, which is the theme under which the product is marketed on the box. In the above-mentioned case, the product falls in the '*Castle'* parent theme. This reduced the number of themes from 624 sub-themes to 147 parent-themes, from now on themes.

After creating the theme, the most frequent brick-colour per LEGO set was added using a combination of different files from rebrickable.com. In this process, the number of rows grew to 410.000 as there was a row for every brick in every LEGO set. Hence, after finding the most frequent colour per set. To reduce the complexity and data abundance, the file was reduced to one row per set again. With this step, the data now had the most frequent colour name and ID for every LEGO set. However, for some, - mostly older - LEGO sets, detailed information on the contents of the LEGO set is missing and the main colour could not be given. Additionally, the number of unique sets is enlarged as certain '*value-packs'* (a combination of various LEGO sets in the same theme) get a unique product ID. Additionally, the data contains product IDs of LEGO related books and (board) games, further enlarging the total number of unique product IDs. Concluding, sourced from the first data source are the set ID, year, number of parts, the parent theme, and the most common colour per LEGO set.

The second data source for this research is an API from Brick Insight (brickinsight.com), which is currently still closed to the public but was made available to the researcher after direct contact with the Brick Insight founder, Linus Bohman. Brick Insight combines reviews of LEGO products from over 35 different websites, creating arguably the most complete overview of reviews on LEGO products. In total, Brick Insight includes 35,000 reviews on 7,326 unique LEGO products on a 0-100 scale with a mean of 79.96 (SD = 10.43), including many unique LEGO sets. Additionally, this API includes the recommended retail price in USD (M = $ 28.95, SD = $ 41.16), which will be used in a later stage to measure to the price increase. Alternatively, this research considered using the Brickset.com API for collecting reviews. However, since the Brick Insight data also includes the reviews from Brickset.com, further action in this API were refrained. The second source provided the number of mini-figures, the recommended retail price, and the review rating for 7326 LEGO products.

The third data source is the Bricklink API. Bricklink.com is a platform for LEGO fans and traders alike, and services as market-place. For every LEGO set traded on Bricklink, the sales price is saved for 6 months, taking into consideration whether sets are sold as used or new. After research and browsing the Bricklink market place, it became apparent that some products labelled as '*used*' miss LEGO pieces, instructions, a box, or are damaged. Therefore, current research only uses the average prices of LEGO sets that were labelled as new. Before being allowed to collect data from the Bricklink API, one has to become an active user and LEGO trader by participating in several sales and receiving positive feedback from vendors and clients. To become an active user, the researcher bought two LEGO sets to his taste, one being the Fire Chief's Car released in 1978 (LEGO set ID 602) in-line with the authors favourite TV-show, Chicago Fire. The second LEGO set bought was the LEGO Architecture Shanghai Skyline, released in 2018 (LEGO set ID 21039). The latter was sold again to become an active seller, receive positive feedback and eventually obtain access to the Bricklink API. Additionally, the Bricklink API limits the possibility to extract large volumes of data, hence substantial processing was conducted to collect the required data, using an extensive API GET call that constantly re-authenticates using different IP-addresses by means of the Rauth Package by Countryman (2015). In total, the Bricklink API provided average prices and the number of sales of 9529 LEGO sets, with an average of $ 90.82 (SD = $218.77). Concluding, both the average price of a LEGO set being traded on the second-hand market and the number of sales per LEGO set were collected using the Bricklink.com API.

## Feature Selection

After merging the various source using the unique set ID, a 10-wide stacked data frame was created using the following features, as shown in Table 1. First, the data frame contained the unique product ID for every LEGO product, followed by the year of release. The data contains several old LEGO products, with the oldest being launched in 1949. However, most products were released relatively recent, with the average release year being 2003. The most recent LEGO products in the data were released in 2019.

The third column contains the number of parts for every LEGO product. Some products only have 1 part, such as books, whereas the biggest set based on the number of bricks contains over 7.500 parts, being the LEGO Star Wars Millennium Falcon. For every product, the theme was converted to the parent theme and is in the fourth column as a categorical value. The fifth column of the data frame also contains a categorical value, being the most frequent colour for that LEGO set. The sixth column contains the number of mini-figures per LEGO set, with many sets containing 0 or 1 mini-figure, the mean is 1.37, with the maximum number of LEGO mini-figures in a LEGO set being 36. This number is represented as an integer, as half mini-figures are non-existent. The number of mini-figures does not include animals, such as parrots and farm animals.

The seventh column contains the recommended retail price in US dollars represented as a float, with the most expensive LEGO product also being the LEGO Star Wars Millennium Falcon. The eight column contains the review rating on a 0 – 100 scale, represented as an integer. There are several LEGO products scoring 100 points as a review, including the 2018 Space Shuttle Transporter (31091-1) and Arctic Supply Aircraft (60196-1). The ninth column is the number of products sold over the last six months of that LEGO set on bricklink.com. How often a product is sold, can influence the price paid on the market place with market dynamics. The most frequently traded product over the last six months is the LEGO Batman Disco Batman (30607-1), being traded 397 times. The final column contains the predictor value, being the average market price of the LEGO product on the second-hand market, denoted as a float. The most expensive LEGO set on the second-hand market, traded in the last six months, is a small police car (6600-1) released in 1978, consisting of just 16 parts and 1 mini-figure.

**Table 1.** *Descriptive Statistics Predicting Features.*

| | Year | Number of Parts | Theme | Frequent Colour | Mini Figures | Retail Price | Rating | Sold | Market Price |
|---|---|---|---|---|---|---|---|---|---|
| Variable | Numeric | Categorical | Categorical | Categoric | Numeric | Numeric | Numeric | Numeric | Numeric |
| Datatype | Integer | Integer | String | String | Integer | Float | Integer | Integer | Float |
| Mean | 2003 | 156.27 | - | - | 1.37 | 28.95 | 79.96 | 15.88 | 90.82 |
| Min | 1949 | 1 | - | - | 0 | 0 | 19 | 0 | 0.00 |
| Max | 2019 | 7541 | - | - | 36 | 799.99 | 100 | 397 | 5000.00 |
| NAs | 0 | 0 | 0 | 6359 | 2224 | 6773 | 7384 | 1861 | 1862 |
| NA % | - | - | - | 43.5% | 15.2% | 46.4% | 50.5% | 12.7% | 12.7% |
| Source | Rebrickable.com | | | | BrickInsight.com | | | Bricklink | |

The data was further explored using a correlation matrix in NumPy and SPSS 24 to measure correlation between features. The results of the Pearson correlation matrix, as shown in Table 2, showed various significant relations. Among them is a positive association between the year of release of a LEGO and the number of parts (r = .346, p = < .001). Additionally, recommended retail price strongly correlates with number of parts (r = .761, p = < .001). The recommended retail price further showed, a correlation with the number mini-figures (r = .310, p = < 0.001). Controversial to literature that suggests that product reviews positively affect the price (Noone & McGuire, 2014; Macik & Macik, 2016), the rating of a LEGO product does only correlate moderately with the average price on the second-hand market (r = .122, p = < .001). The categorical features, being the theme and colour of a LEGO set, were omitted from this matrix.

**Table 2.** *Pearson Correlations Among Predicting Features*.

| Feature | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1. Year of Release | 1 | .346** | .069** | - .023* | -.275** | .191** | .004 |
| 2. Number of Parts | .346** | 1 | 0.381** | .761** | -.103** | .149** | .334** |
| 3. Number of Mini-figures | .069** | .381** | 1 | .310** | -.035** | .033** | .149** |
| 4. Recommended Retail Price | -.023* | .761** | 0.310** | 1 | .370** | .220** | .492** |
| 5. Rating Aggregated | -.275** | -.035** | -.035** | .227** | 1 | .070** | .122** |
| 6. Quantity Sold | .191** | .149** | 0.033** | .220** | .070** | 1 | .002 |
| 7. Average Price Market | 0.004 | .334** | 0.149** | 0.492** | .122** | .002 | 1 |

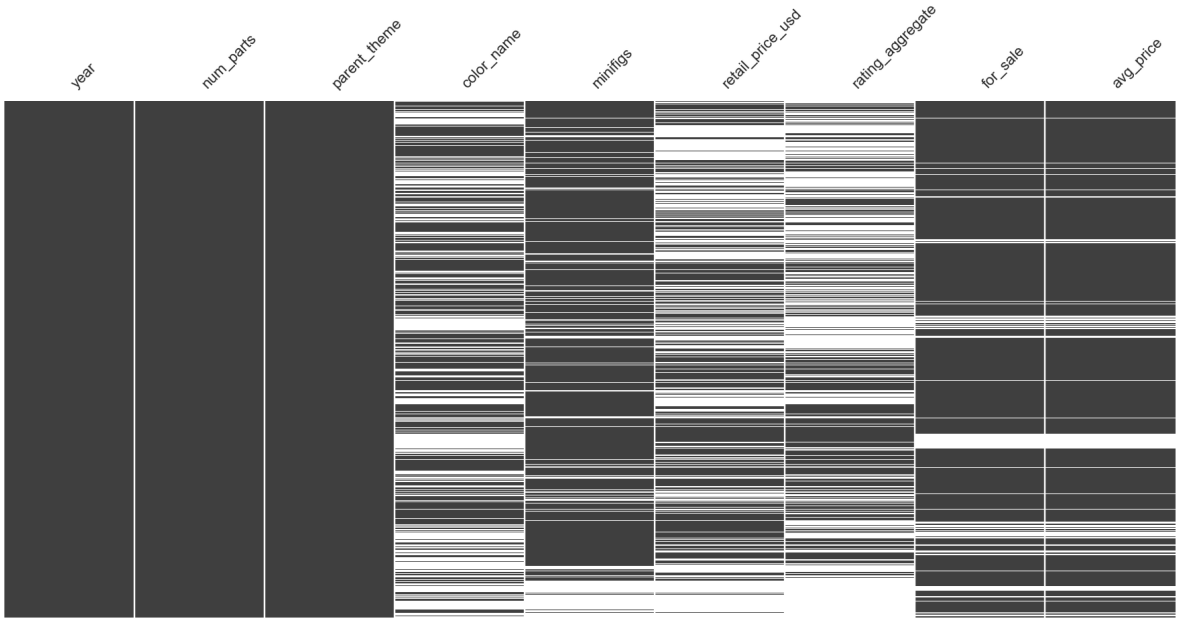
**. Correlation significance at the 0.01 level (2-tailed).

In addition to above mentioned-correlation matrix, the Scikit-learn feature selection algorithm was used, as recommended by Khurana and colleagues (2016). Corresponding with the correlation matrix, the algorithm suggested the recommended retail price as most influential feature for the algorithm that predicts the average price on the second-hand market, followed by the number of parts and number of mini-figures in each set.

## Missing Data

With data collection from various sources, after combining data, many instances are missing as all additional files are joined on the initial comprehensive list of products IDs, which include the value-packs, games and, books, among others. Therefore, out of the original 14608 unique product IDs, data is missing for 7384 reviews, 1862 sale prices, 2224 mini-figures and 6773 recommended retail prices. In total, 18.7 % of all data is missing. However, for certain features this percentage is higher, such as for review ratings (50.5%). The spread of missing data is displayed in Figure 1 for all predicting features, showing an uneven spread of missing data among the various features, challenging further analysis.

**Figure 1.** *Spread of Missing Data.*




The missing data could be categorized as Missing at Random, which is a '*misnomer as missing data is anything but missing at random* (Scheffer, 2002, p.3). This is mostly due to the fact that data comes from various sources, and therefore is dependent on the source.

Various methods to replace the missing values were evaluated, including complete removal, mean substitution and, prediction methods. Complete removal of products where data is missing would significantly reduce the model to 3.500 product IDs. Mean substitution was considered for the number of mini-figures, but as suggested by research, was excluded as this would undermine the model as this method does not preserve the relationships among variables (Roth et al., 1999; Scheffer & Graham, 2002). As several features correlate, a Scikit-Learn predictor algorithm was created for the missing review ratings, missing mini-figures and the missing recommended retail-price, but results were lacking as it had to train on minimal existing features. This model, for example, suggested 3 mini-figures, for a LEGO set that contained 0. Another method to cope with missing data is multiple imputations and has shown in different fields to provide unbiased estimates and delivers accurate parameter estimates, in particular the Multiple Imputations by Chained Equations method, or MICE (Sterne, et al., 2009; Barnaldi & Enders, 2010). MICE is an imputation model that runs multiple regression models for every feature and models each missing value conditionally on the observed features in the data (Azur et al., 2011). As the support for packages that enable MICE in python is limited, this step was performed in R using the MICE package by Van Buuren and Groothuis-Oudshoorn (2010). Van Buuren and Groothuis-Oudshoorn (2010) suggest a missing data threshold of 5%, whereas the data in this research is missing in 18.7 % of the cases, with an even bigger spread on the missing data. As many products in the data set have missing values, the tested accuracy on the imputation is very low, scoring a mere 24.7% accuracy on the test set. Due to time constraints in this research, the researcher was unable to optimize the MICE model in the hope for better accuracy. With data imputations not delivering proficient data, continuing with this data would significantly undermine the later design model, as the imputed data is likely to be incorrect. The decision was therefore made to drop the products that had missing data for the following features.

First, as current research is aimed at predicting the returns of a LEGO set using supervised Machine Learning, the predictor value has to be present to train the algorithm later on. Therefore, all products that had missing data or no available sales data in this column were deleted, omitting 5051 products from this research. Secondly, the original recommended retail price has to be present as well in order to calculate the returns. However, to the best of my knowledge, currently there is no source that has an unambiguous recommended retail price, as this varies per country and year, further challenging the measurement of '*future returns'*. Therefore, solely products with a

recommended retail price are used in this research, omitting another 1862 data instances. Thirdly, as after removing these instances, there were still products without multiple predictor values, the decision was made to exclude all products without a review.

After deleting these products, missing values for categorical features, being the theme and the most frequent colour of the LEGO set, were recoded as '*None*', making a distinct categorical value. Once the missing values were removed and recoded, the categorical values were transformed to numeric categorical values using Scikit-Learn label encoder (Pedregosa et al., 2011), making the None values -1 values.

To minimize the effect of outliers, the features were scaled to minimize the effect of distances in a feature (for example, the number of mini-figures ranges from 0 to 36). Two types of scaling that come with Scikit-Learn were tested, namely min-max scaling and robust scaling, the latter showing best results in existing work (Martinsson et al., 2011; Olson et al., 2016) and current study. The robust Scaler uses a similar method to min-max scaler, but instead of using the actual min and max values, it uses the statistical quartile range, making it more robust to outliers (Géron, 2017).

## Data Partitioning

After handling the missing data, the cleaned and pre-processed dataset has been split using Scikit-Learn (Pedregosa et al., 2011) to a training set consisting of 80% (3996 products) of the total instances and a 20% hold-out (1000) for testing purposes. The Pareto, or 80/20, split is frequently used in Data Science and seems to perform well for many tasks (Macek, 2008), but others have also suggested using 2/3 for training and 1/3 for testing (Dobbin & Simon, 2011) or 85% and 15% for training and testing respectively (Anderson & Mugruder, 2017). The created algorithm can then be tested on the test data set, allowing for a fair evaluation (Géron, 2017). To ensure validity and minimize bias, K-fold cross-validation was used in Scikit-Learn. Cross-validation estimates the expected error of an algorithm by repeatedly running a training and testing procedure on different parts of the data set, hereby preventing testing and training on a contaminated part of the data (Geras & Sutton, 2013). By including K-fold cross-validation, the algorithm can train multiple times without overfitting on the training data (Blum et al., 1999). In line with literature, this research uses 10-fold cross validation (Kohavi, 1995) and averages over the results for all test-results to prevent bias in the data partitioning.

## Overview of Software

Most of the steps in this research were carried out using the Python programming language, originally developed in 1991 by Guido van Rossum as versatile open-source language. The possibility of adding external packages to the Python language make it a real powerhouse, in particular for many data science-related topics (Bressert, 2012). This research uses some commonly used packages in Data Science, including Pandas (McKinney, 2010), NumPy (Travis, 2006) and Scikit-learn (Pedregosa et al., 2011). Table 3 presents an overview of the Python libraries used and their purpose in this research. All Python processing steps were performed in the Jupyter environment, with the exception of API calls. The API calls that required coding were performed in Spyder-IDE, an open-source software development environment for Python originally created by Pierre Raybaut (Spyder Github, 2019). One step was performed using R and the MICE package (Van Buuren and Groothuis-Oudshoorn, 2010), for the purpose of missing data imputations.

**Table 3.** *Python Libraries Used In This Research.*

| Library | Version | Purpose |
|---|---|---|
| NumPy | 1.13.3 | Scientific formulas and data manipulation. |
| Pandas | 0.19.2 | Data mangling, cleaning, viewing, missing data analysis. |
| Scikit-Learn | 0.18 | Machine-learning algorithms, the train & test split and performance metrics. |
| Matplotlib | 3.0.3 | Used for plotting purposes and EDA. |
| Seaborn | 0.9.0 | Used to visualize missing data. |
| Math | 3.7.3 | Various mathematical functions supporting NumPy. |
| Rauth | 0.7.2 | Used to re-authenticate API calls. |
| Multiprocessing | 2.6.2.1 | Enables multi-threading for API processes. |
| JSON | 2.0.9 | Customized reading of JSON API files. |

Additionally, some statistical analysis were performed in SPSS 24. Initially, SPSS was used to perform a Pearson's correlation analysis among predicting features. Later, the program was used to compare LEGO sets that have a positive NPV with those with a negative NPV, using an independent sampled T-test. Additionally, SPSS was used to validate the descriptive statistics performed in Python – specifically NumPy - to ensure validity.

## Algorithms and Evaluation

This study used four predictive algorithms, all from the Scikit-Learn package for Python (Pedregosa et al., 2011), with the base seed set to 42 to enable reproducibility. The first is the traditional Linear Regression algorithm, which aims at creating a simple linear line through the data (Varian, 2014). The Linear Regression model serves as a baseline to compare further methods on their error measure.

Second, a Polynomial Regression has also been implemented, supporting more flexible relations in the data compared to simple Linear Regression (Darlington, 1990; Varian, 2014). Additionally, Polynomial Regression can cope better with categorical features (Tutz, 2011), which are included in current research's data. The flexibility of this method comes from the number of added regression coefficients, called the degrees in Scikit-learn (Pedregosa et al., 2011; Géron, 2017). For this research, the number of degrees has been tested in the range of 1 to 10.

Third, this research implements a Random Forest of Regression Tree, which has proven to be the most powerful solution in previous data science research for prediction (Ali et al., 2012; Touw et al., 2012). Random Forest work with a number of decision trees, creating the forest (Pedregosa et al., 2011). For this research, a Random Forest was generated using 1 to 10 trees, as well as several higher values to observe the difference in results and the potential overfitting model. The maximum depth of the tree was tested in the range of 1 to 10, before eventually being maximized to 5.

The fourth model was a Random Forest Classification model, that was used to classify LEGO sets yielding a positive and negative NPV. A Random Forest Classification Tree creates multiple trees, just as with the Regression Tree, that combine into the so-called forest. Classification Trees, especially in ensemble form, are powerful method to classify labels correctly (Géron, 2017).

All models were evaluated using the error measure Root Mean Squared Error (RMSE), which is as the name suggests the root of the MSE, reducing the impact of extreme values. The RMSE formula averages the squared differences between the predicted ($\hat{y}$) and observed value (y), and takes the root of this value. The formula is as follows:

$$RMSE = \sqrt{\sum_{j=1}^{n} \left(y_j - \hat{y_j}\right)^2}$$

Even though the RMSE has been challenged in research as too insensitive to outliers (Willmott & Matsuura, 2005; Willmott et al., 2012), it remains one of the '*best*' overall measures for model performance and suits current research value prediction task and models (Willmott, 1982; Chai & Draxler, 2014a; Géron, 2017). However, current research also looked at alternative error measures.

One alternative to RMSE, for example, is the Mean Absolute Error (MAE) which sums all the absolute values of the errors to obtain a total error and divides this by the number of errors (Willmott & Matsuura, 2005). Even though the use of MAE is challenged, as the use of absolute values is highly undesirable in many mathematical calculations (Chai & Draxler, 2014b), it is supported by many researchers as it less sensitive to outliers (Willmott & Matsuura, 2005; Willmott et al., 2012). For example, as demonstrated by Willmott & Matsuura (2005), whilst the MAE remains at a constant value of 2, the RMSE varies in a range from 2 to 4. Additionally, a combination of metrics is often required, or at least recommended, to assess model performance (Chai & Draxler, 2014a). Therefore, besides the RMSE, this research also includes the MAE for interpretation purposes. The formula for the MAE is as follows:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

The evaluation goal in this research is to minimize the RMSE on the test data, suggesting a better fit to unseen data. As the RMSE is measured in the same scale as the predictor value, there is no standard evaluation comparison. Nonetheless, a lower error is better (Géron, 2017).

# Chapter Four: Results

This chapter has two main sections. In the first section, the financial performance of LEGO products as an alternative investment are presented, based on their average annual return and NPV. Additionally, this first section contains an overview of the top performing LEGO sets based on their average annual return. In the second section, the performance results of algorithms that aim at predicting the future value of a LEGO set are presented.

## Financial Performance of LEGO Products as Alternative Investment

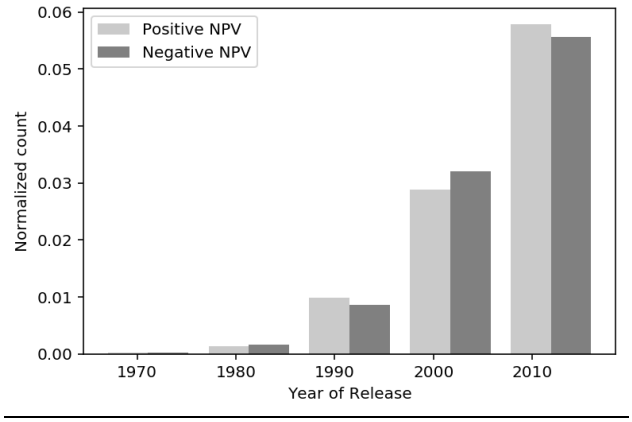
***LEGO Sets Trades.*** In order to analyse the most-popular themes among collectors, this research looks at the number of products sold over the last six months. With the LEGO Batman Disco Batman Pack (30607-1) - which was licenced to LEGO in 2017 after the release of the LEGO Batman Movie (Chun Chen, 2017) - being the most traded LEGO item (397 times), one would expect that Superheroes related items are traded most frequently. After analysing the number of products sold per theme on the second-hand market, the Super heroes theme is indeed among the top five most frequently traded themes. As shown in Table 4, this top 5 is being led by the Star Wars theme, being traded a total of 2,0653 times over the last six months. On second place comes the Super Heroes theme (973 times) and on third place comes LEGO City (784). Additionally, Table 4 shows the most frequently traded items that were sold for a premium over the market-value, marking a positive NPV, something investors require for their investments. This model assumes that the LEGO set was bought when it was released, and uses the observed average price from the second-hand market to calculate the NPV as discussed before in the Literature Review. In the top five most frequently traded products with a positive NPV, the LEGO theme *Friends* is exchanged for *Seasonal*, suggesting that LEGO sets in the Seasonal theme (including many Christmas related products) are sold for a premium.

**Table 4.** *Most Commonly Traded Themes.*

| **Top 5** | **Frequency** | |
|---|---|---|
| # | All LEGO sets | Positive NPV |
| 1. | Star Wars (2,065) | Star Wars (1,490) |
| 2. | Super Heroes (973) | Super Heroes (712) |
| 3. | City (784) | City (448) |
| 4. | Creator (558) | Seasonal (362) |
| 5. | Friends (470) | Creator (357) |

***Descriptive Statistics of LEGO Sets Returning a Positive NPV.*** After calculating the NPV for all 3630 LEGO sets included, the majority (2891 LEGO sets) have a positive NPV. As shown in Figure 2, most LEGO sets were launched after the year 2000, also with most of them yielding a positive NPV. However, sets launched around 1980 and 2000 have the biggest percentage of negative NPV sets, and LEGO sets launched in the 1990s and 2010s more frequently show a positive NPV.

**Figure 2.** *Normalized Distribution of LEGO Sets Released Per Year.*




There are various minimal differences between LEGO sets that have a positive NPV and those that have a negative NPV. Using an independent sample T-test analysis in SPSS, no significant relations were found. There are however minimal differences between the groups' descriptive statistics. To start with, analysing the number of bricks per LEGO sets that had a positive NPV (M = 321.76, SD = 460.29) versus those that had a negative NPV (M = 330.69, SD = 401.20), only a minimal difference was found. Additionally, the number of mini-figures does not considerably change for LEGO sets having a positive NPV (M = 2.25, SD = 2.35) or negative NPV (M = 2.18, SD = 2.69). The retail price of products with a positive NPV is slightly lower (M = 33.73, SD = 42.45) compared to products that have a negative NPV (M = 40.64, SD = 42.23). Additionally, there are two products with a recommended retail price of $ 499.99, both Star Wars related. One of the two, the 2007 Millennium Falcon has a high positive NPV. The second one, the 2016 Death Star does not have a positive NPV, despite having 27 mini-figures and over 4,000 parts.

***Top 5 LEGO Set Investments.*** As there seems to be no difference between LEGO products with a positive NPV and negative NPV based on the descriptive comparison above, this research now looks at the LEGO sets with a positive NPV, yielding the highest average annual return.

Table 5 provides an overview of LEGO products with positive NPVs and yielding the highest average annual return. Out of these LEGO sets, the 1984 Surface Rover is the LEGO set yielding the greatest returns, once sold for $ 0.99, recently sold on the second-hand market for $ 400.00. The second set on the list is the special edition Star Wars Detention Block Rescue. This LEGO set was only available during the 2017 Florida Star Wars Celebration (Ruijter, 2017). Whereas this 220 piece LEGO set was available for $ 39.99, sealed boxes are sold online on average for $ 412.98. Other LEGO products yielding the greatest returns are the Laser Ranger (being another Space related LEGO set). The third item among the LEGO sets yielding the highest average annual returns is another exclusive, namely the 2016 *Brickheadz* Iron Man & Captain America (41492-2). *Brickheadz* are collectible LEGO products, mimicking a character from a movie, and the Iron Man & Captain America set was only sold in 2016 following the release of the *Captain America: Civil War* movie. Even though this set is frequently being traded (and there have only been 4 sales in the last six months in new condition), the average return on this product is still over 570 % annually. Fourth on the list is the 1980 LEGO Classic Town Gas Station (6375-2), featuring an Exxon Gas Station. The final product on the list is the 1989 Laser Ranger, being a spaceship made out of 41 bricks and selling online for $ 148.16.

**Table 5.** *LEGO Sets Yielding the Highest Average Annual Return.*

| | Set ID | Name | Year | Parts | Retail Price | Average Price | Sold # | Average Annual return |
|---|---|---|---|---|---|---|---|---|
| 1 | 6804-1 | Surface Rover | 1984 | 21 | $ 0.99 | $ 400.- | 2 | 975.82 % |
| 2 | CELEB 2017 | Detention Block Rescue Star Wars | 2017 | 220 | $ 39.99 | $ 412.98 | 1 | 466.35 % |
| 3 | 41492 | Iron Man & Cpt. America | 2016 | 174 | $ 39.99 | $ 724.44 | 4 | 570.52 % |
| 4 | 6375-2 | Exxon Gas Station | 1980 | 270 | $ 20.00 | $ 2229.18 | 1 | 283.13 % |
| 5 | 6810-1 | Laser Ranger | 1989 | 41 | $ 2.00 | $ 148.16 | 5 | 243.13 % |

Arguably, this top five contains three classic LEGO products (being Space and Classic Town themes), whereas the newer products in this top five are both Star Wars and Super Hero related. When analysing the average annual return per theme, only one of the themes belonging to top performing products returns, being *Brickheadz*. As shown in Table 6, on

average, the *BrickHeadz* theme is the best performing theme with products released recently in 2016 and 2017, demonstrating an average annual return of 102.52 %. Additionally, the results show that *Brickheadz* are frequently being sold and the original retail price was relatively low. Second on the list is the *Avatar* theme, yielding a 70.53 % average annual return. The Avatar products were released in 2006 and, although only include two different LEGO sets, still sell for a premium, making them a top performer. Third on the list of top performing themes are the 2016-2018 *Speed Champions,* returning on average 57.49 % annually, and is traded frequently with 746 *Speed Champions* LEGO sets being sold over the last six months. Fourth and fifth on the list of best-performing themes are *Ninja* and *Sculpture*s, with 46.76 % and 44.49 % respectively. The Sculptures theme contains architectural and technological wonders, including one of the biggest LEGO sets ever produced, the 2008 Taj Mahal (10189-1) containing 5922 pieces, and the 2003 LEGO Wright Flyer (10124-1), a LEGO aeroplane based on the design of the 1903 Kitty Hawk by the Wright Brothers.

**Table 6.** *LEGO Themes Yielding the Highest Average Annual Return Including Average Descriptive Statistics per Theme.*

| | Theme Name | Products in Theme | Total Theme Sold | Average Number of Bricks | Average Retail Price Theme | Average Annual Return |
|---|---|---|---|---|---|---|
| 1 | Brickheadz | 13 | 977 | 116.08 | $ 12.30 | 102.52 % |
| 2 | Avatar | 2 | 14 | 562.00 | $ 40.00 | 70.53 % |
| 3 | Speed Champions | 21 | 746 | 362.05 | $ 32.85 | 57.49 % |
| 4 | Ninja | 18 | 97 | 150.43 | $ 18.17 | 46.76 % |
| 5 | Sculptures | 17 | 792 | 2038.41 | $ 147.82 | 44.49 % |

Despite having two products in the top five products yielding the greatest annual return, the *Space* theme is only 20$^{th}$ on the list of top-performing themes overall, returning 34.65 %. Also, the Star Wars theme only demonstrates a 17.78 % average return annually, and is thus not among the top-performing themes.

## Predicting Future Value of LEGO Products

Now that the best-performing LEGO products in terms of their average annual returns and positive NPVs are presented, this research continues with exploring the possibility to predict the future value of LEGO sets. To do so, four different algorithms were tested.

***Linear Regression.*** The first algorithms, Linear Regression, acts as a baseline model. As presented in Table 7, it scored an average RMSE of 163.72 and an MAE of 78.80 on the test set, whilst the RMSE on the training set was 174.98 on average and the corresponding explained variance of the algorithm was average ($R^2 = 0.49$). To improve this score, several features were removed from the Linear Regression model, worsening the RMSE on the test set. For example, after removing the *year of release* feature, the RMSE on both the test and training set increased. Additionally, a Ridge Regression was tested, yielding similar results.

**Table 7.** *Linear Regression Model Performance*.

| **Predictor** | **Performance** | | | |
|---|---|---|---|---|
| | RMSE Train-set | RMSE Test-set | MAE Test-set | $R^2$ test-set |
| Linear Regression | 174.98 | 163.72 | 78.80 | 0.49 |
| – Year of Release | 183.28 | 167.94 | 78.32 | 0.46 |
| Ridge Regression | 183.28 | 167.94 | 78.32 | 0.46 |

***Polynomial Regression.*** After analysing the results of the Linear Regression model and the spread of data, additional degrees were added, making the model a Polynomial Regression model (Darlington, 1990; Varian, 2014). The RMSEs of 2 and 3 degrees Polynomial Regression are lower compared to the simple Linear Regression but the model also shows to be prone to overfitting. Table 8 displays the RMSEs, MAE and, $R^2$ of a Polynomial Regression model with different degrees.

**Table 8.** *Polynomial Regression Model Performance.*

| **Number of Degrees** | **Performance** | | | |
|---|---|---|---|---|
| # | RMSE Train-set | RMSE Test-set | MAE Test-set | $R^2$ test-set |
| 1. | 174.98 | 163.72 | 78.80 | 0.49 |
| 2. | 133.95 | 116.91 | 59.92 | 0.74 |
| 3. | 112.77 | 161.44 | 62.45 | 0.51 |
| 4. | 93.66 | 2887.66 | 289.68 | -0.94 |
| 5. | 55.64 | 83836 | 6084.89 | -131,668.67 |

As Table 8 suggests, the Polynomial Regression with only one degree scores the same as a Linear Regression. Since the Polynomial Regression with only one 1 degree automatically becomes a Linear Regression, replicating the baseline scores. The Polynomial Regression scores best on the test-data with two degrees, resulting in an average RMSE of 116.91 and MAE of 59.92, with a relatively well-fitting model ($R^2 = 0.74$). When adding more degrees, the model over-fits drastically, resulting in an increased RMSE on the test-set for a four-degree model (RMSE = 2887.66, $R^2 = -0.94$) and five-degree model (RMSE = 83836, $R^2 = -131{,}668.67$). The Polynomial Regression with 2 and 3 degrees beats the baseline Linear Regression model. The two-degree Polynomial Regression holds an intercept of $ - 128.37, and, based on the features multiplies 45 coefficients created by the algorithm, calculates the future value of the LEGO set.

***Random Forest Regression Tree.*** The third algorithm used to predict the future value of LEGO sets is a Random Forest, which generates multiple Regression Trees (Breiman, 2001; Liaw et al., 2002). For various number of trees (so-called estimators) in the Random Forest, the performance was measured. Out of all the models, the model with 25 estimators scored best on the test set, resulting in an RMSE of 151.70 and an MAE of 58.00. This means that the model beats the MAE baseline, but does not come close to the RMSE baseline. Within the first 25 estimators, the biggest improvements on the RMSE scores are made. When growing the Random Forest to 250 estimators and more, the RMSE on the training set decreases steadily towards 100, as shown in Table 9. Despite the training set RMSEs decreasing when the number of trees increased, the RMSE on the test set fluctuates around 155.00. Thus, adding more estimators does not improve the RMSE, due to the limited number of predictor features.

**Table 9.** *Subset of Random Forest Predication Model Performance.*

| **Number of Trees** | **Performance** | | |
|---|---|---|---|
| # | RMSE Train-set | RMSE Test-set | MAE Test-set |
| 5. | 100.67 | 164.30 | 61.37 |
| 10. | 102.80 | 152.17 | 58.66 |
| 25. | 105.50 | 151.70 | 58.00 |
| 50. | 104.10 | 155.16 | 58.43 |
| 100. | 102.34 | 154.80 | 57.98 |
| 250. | 101.90 | 155.65 | 58.40 |

***Classification Tree.*** As these RMSE scores are relatively high compared to the mean average price, this research also looks at the possibility to predict whether a LEGO set matches the minimal profitability requirements by investors, the NPV. Hence, the fourth model used is a classification model, specifically a Random Forest Classifier, that uses an ensemble learning method for classification trees. For this research, initially 5 distinct classification classes were created, depending on the financial performance of a LEGO set.

The first class contained LEGO sets having both a negative NPV and a Negative average annual return (classified as double negative). The second class included the products with only a negative NPV. The third class is the neutral class where the NPV was zero. Products scoring a positive NPV were classified as positive or double positive, depending on their score, making a total of 5 distinguishable labels for the algorithm

As classification models cannot be compared using RMSE, this model was tested using the accuracy score. The Random Forest was tested with various numbers of trees in the forest and a depth maximum of 6. As shown in Table 10, the classifier yielded a 66.5 % accuracy on the test set for 250 trees, suggesting that the algorithm in 66.5% of the cases could correctly predict the financial class of a LEGO predict. When adding more trees, the accuracy on the test set did not show improvement. Also, the results of the training set were relatively stable over the number of trees.

**Table 10.** *Subset of Random Forest Classification Model Performance.*

| **Number of Trees** | **Performance** | |
|---|---|---|
| # | Accuracy Train-set | Accuracy Test-set |
| 10. | 71.34 % | 65.51 % |
| 100. | 71.44 % | 65.49 % |
| 250. | 71.39 % | 66.47 % |
| 500. | 71.34 % | 66.42 % |

As a five-class model is rather unambiguous, another two-class Random Forest model was created, with the first label being a negative or a neutral NPV and the second label being a positive NPV. Using 10-fold cross-validation, this model, on average, scored an accuracy of 83.60 % in a robust model ($F1$ = 0.94), doing better than the five-class model and allowing for a more certain classification between future LEGO sets returning a positive or negative NPV.

# Chapter Five: Discussion and Conclusion

Before looking into the possibility to predict the future value of LEGO sets, this research first looked into the financial performance of LEGO sets as an alternative investment. Overall, the LEGO products being traded on the second-hand market show an average annual return of 21.34 %, which is slightly higher than the average return by Dobrynskava and Kishilova (p. 15, 2018), who reported a return of 18.5 % annually. Overall, LEGO sets outperform the average market return, markings its potential as alternative investment. This research also looked at the average annual returns per theme, showing that *Space* and Star Wars theme are not among the top performing themes, despite having separate products that are yielding substantial returns. Per this finding, this study contradicts Baichtal & Meno (2011) who argued that *Space* was a top performing theme. Rather, based on the findings in this research, the *BrickHeadz* theme is the best performing overall theme.

This is however in line with the Maciorowski brothers (2016), who suggested that other '*geeky'* themes are popular and are therefore likely to be traded for premiums. The results of this research support this notion. Even though the *Space* theme was not a top performing theme overall, it does deliver products that out-perform most other LEGO sets, including the 1984 Surface Rover, returning on average 975.82 % annually, outperforming alternative investment in classic cars, reaching 461 % annually (Hietbrink, 2017). Additionally, Star Wars and Heroes related LEGO sets, arguably *geeky*, are among the most commonly traded products on the second-hand LEGO market, marking their popularity.

The main purpose of this research was to investigate the prospect of predicting the future value of LEGO sets given the set characteristics using Supervised Machine Learning and thereby answer the main research question: *"To what extent can we predict the future value of LEGO investments given the LEGO set characteristics?*

This study reveals that a two-degree Polynomial Regression is able to predict the future value of a LEGO set with an RMSE of $ 116.91 and MAE of $ 59.92. This RMSE is relatively high compared to the average price on the second-hand market ($ 68.51). Regardless, the model fits the data relatively well ($R^2 = 0.74$), but is still under-performing to comparable existing work by Ashenfelter (2005), who predicted the value of wine ($R^2 = 0.82$). Additionally, this research classifies LEGO sets that yield a positive NPV, a minimal requirement by investors, with 83.60 % accuracy. Again, the study by Ashenfelter (2005), showed a better performing model (94.5 %). However, wine and LEGO are not interchangeable, and therefore current research sets a baseline for future LEGO predictions.

## Theoretical Implications

This research has several implications for academic literature. In contrasts to current research on alternative investment, this research suggests that LEGO indeed matches the characteristics of alternative investments. Based on research to alternative investments, current research argues to combine different alternative investments and introduces a classification with three main categories, being *1)* collectibles, *2)* usables and *3)* consumables, where LEGO would classify as a usable (*'to play with'*) alternative investment. Additionally, this research contributes to the analysis of LEGO as an alternative investment, by presenting the average annual return per theme, supplementing the research by Dobrynskava and Kishilova (2018) who only presented the overall financial returns. This study contradicts existing work by Maciorowski & Maciorowski (2016), and Baichtal & Meno (2011) who argued that space-related themes are best performing, whereas this study shows that the *Brickheadz* theme is yielding the best returns.

Current research also contributes to the field of predicting alternative investments using algorithms - albeit with a high RMSE - which to the best of my knowledge, only has been done by Ashenfelter (2005). It is however likely that these studies are not published after research, as they contain knowledge of financial value. Therefore, current research adds to the new avenue of research on LEGO as an alternative investment, which has gained increased attention in news articles recently (e.g. Burton, 2019).

## Managerial Implications

The subsequently assumed potential limitations notwithstanding, this study suggests two practical implications for investors and possibly collectors who invest in LEGO. First, for the prospecting investor looking for an investment with above-average market returns, this study reveals that they are best of investing in '*geeky*' themes, including *Brickheadz* and heroes related themed LEGO sets, such as *Avatar* and *Ninja*, but not in Star Wars or *Space* related themes. Second, this research proposed a two-class classification model that predicts if a LEGO set will yield a positive NPV, which can be of value to investors. More in-depth, this research proposed a 2-degree Polynomial Linear Regression, which investors can use to predict the future value of LEGO sets given certain LEGO set characteristics, including the year of release, number of bricks and number of mini-figures. Additionally, these findings might be of value for The LEGO Group themselves, who could aim at re-introducing discontinued geeky themes, such as *Avatar* and *Ninja*.

## Limitations and Further Research

This study has gone some way towards enhancing knowledge on future value predictions for LEGO sets using Machine Learning, but is not without limitations and solving them can be valuable avenues for future research.

First, future researchers could assess the possibility of optimizing the prediction model. Current research proposed a prediction model that presents an RMSE of $ 116.91 which may, despite high standard deviations in the data, not be optimum and may benefit from new features that enable enhanced future value prediction. For example, the prediction model may benefit from further experimental feature design. Current research uses the number of LEGO sets sold on the second-hand market as a measurement for popularity of a particular set. However, this does not sufficiently account for the rarity of a LEGO set, which may significantly influence the future value (Koford & Tschoegl, 1998; Maciorowski & Maciorowski, 2016), as is the case with the previously mentioned LEGO Star Wars Detention Block Rescue. However, current data does not support this nor do current data sources currently support this feature.

Second, this research lacks detailed financial information, such as transactions costs, and the timing of LEGO purchases. Due to time constraints and project limitations, current research only uses sales data from the last six months and does not include data on when a LEGO set was purchased. Because of this, current research uses the year of release and the recommended retail price at the time of release to calculate the average annual returns, limiting the financial calculations in this research. Therefore, future researchers might explore the possibility to collect data from investors who keep track of when they purchased a LEGO set, the actual price paid and the corresponding returns.

Third, this research was based on various open data sources. Even though the researcher took great care of ensuring data quality, readers are advised to consider this when interpreting the results of this research. For example, this research only includes transactions from the second-hand marketplace bricklink.com. Future research may consider using a broader range of website. Additionally, the various data sources reported on different LEGO sets, resulting in the omission of 50.5 % of the data instances. Hence, future researchers are advised to improve the number of data instances and data sources to minimize the effect of missing data.

Fourth, current research set out to explore the possibilities to predict the future price of LEGO sets only and therefore neglected individual mini-figures that are sold separately. However, the internet shows several very unique and valuable LEGO mini-figures. For example, the 2013 Mr. Gold (71001-19) valued for up to $ 5,000 and the 2007 Star Wars 30th Anniversary C-3PO Chrome Gold mini-figure, worth over $ 1,500 (Bricklink.com. 2007). Possibly, the most expensive known piece of LEGO is the 14 Karat Gold plated Boba Fett, of which two were given away at the 2010 Comic Con in San Diego. These mini-figures have an estimated price of $ 11,500 (Minifiguresseries.org, 2018). With mini-figures being among the most expensive traded LEGO products, future researchers are advised to consider incorporating mini-figures.

## Conclusion

This study initially set out to examine LEGO sets that show potential as an alternative investment and study the possibility to predict the future value of LEGO sets given the set characteristics. In measuring the financial performance of LEGO sets, this research reveals that, contradicting existing work, *Brickheadz*, Avatar and, Ninja are LEGO themes that yield the greatest average return annually. Further comparison of LEGO sets using their NPV, a popular metric used by investors to evaluate their investment decision, no significant differences were found between LEGO sets that yield a positive NPV and those that do not yield a positive result. Current research proposed a Random Forest Classification that successfully classifies the NPV of unknown LEGO sets as positive or negative. More in-depth, this research tested different Machine Learning algorithms to predict the future value of LEGO sets, given the LEGO sets characteristics. The results of this study show that a two-degree Polynomial Regression model performs best on predicting the future value of a LEGO set

The results of this study contribute to the small body of literature that sees LEGO as an alternative investment. For investors looking at LEGO as an alternative investment, this study marks the beginning of an algorithm that predicts the future value. Even though the financial returns are promising, LEGO remains a specialized product which requires advance knowledge of the market, as not just every Star Wars related product yields great returns. This knowledge is especially needed, as the proposed model in this research shows a relatively high error rate. Therefore, investors are advised to keep an eye out and *'mind the brick'*.

# Reference

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. International Journal of Computer Science Issues (IJCSI), 9(5), 272.

Anderson R. (1974), Paintings as an investment, Economic Inquiry 12(1) : 13-26

Anderson, M. L., & Magruder, J. (2017). Split-sample strategies for avoiding false discoveries (No. w23544). National Bureau of Economic Research.

Anson, M. J., Fabozzi, F. J., & Jones, F. J. (2010). *The handbook of traditional and alternative investment vehicles: investment characteristics and strategies* (Vol. 194). John Wiley & Sons.

Antorini, Y. M., Muñiz, Jr, Albert,M., & Askildsen, T. (2012). Collaborating with customer communities: Lessons from the lego group. MIT Sloan Management Review, 53(3), 73-95. Retrieved from https://search.proquest.com/docview/963962186?accountid=14338

Ashenfelter, O. (2008). Predicting the quality and prices of Bordeaux wine. The Economic Journal, 118(529), F174-F184.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International journal of methods in psychiatric research*, *20*(1), 40-49.

Baichtal, J., & Meno, J. (2011). The cult of LEGO. No Starch Press.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. Journal of school psychology, 48(1), 5-37.

Bartneck C, Moltchanova E (2018) LEGO products have become more complex. PLOS ONE 13(1): e0190651. https://doi.org/10.1371/journal.pone.0190651

Bartneck, C., Ser, Q. M., Moltchanova, E., Smithies, J., & Harrington, E. (2016). Have LEGO Products Become More Violent? PLoS ONE, 11(5), e0155401. | DOI: 10.1371/journal.pone.0155401

Bartneck, Christoph & Obaid, Mohammad & Zawieska, Karolina. (2013). Agents With Faces - What Can We Learn From LEGO Minifigures?.

Bender, J. (2010). LEGO: a love story. John Wiley & Sons.

Blum, A., Kalai, A., & Langford, J. (1999, July). Beating the hold-out: Bounds for k-fold and progressive cross-validation. In COLT (Vol. 99, pp. 203-208).

Bodnar, T., & Schmid, W. (2007). Matrix elliptical contoured distributions versus a stable model: Application to daily stock returns of eight stock markets. Asset Allocation and International Investments, (2007), P. 214.

Bradley. (2016). How to Invest in Legos and Make a Bazillion Dollars. Retrieved from https://www.wealthsimple.com/en-gb/magazine/how-to-invest-in-legos [Accessed 3 May. 2019].

Brand Finance Toys. (2018). 25th ed. [ebook] Brand Finance. Available at: https://brandfinance.com/knowledge-centre/reports/brand-finance-toys-25-2018/ [Accessed 13 Apr. 2019].

Brealey, R. A., Myers, S. C., Allen, F., & Mohanty, P. (2012). Principles of corporate finance. Tata McGraw-Hill Education.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Bressert, E. (2012). SciPy and NumPy: an overview for developers. " O'Reilly Media, Inc.".

Bricklink.com (2007), Bricklink.com – Item Catalog. C-3PO - Chrome Gold (SW 30th Anniversary Edition). https://www.bricklink.com/v2/catalog/catalogitem.page?M=sw0158. [Accessed 5 May. 2019].

Briere, M., Oosterlinck, K., & Szafarz, A. (2015). Virtual currency, tangible return: Portfolio diversification with bitcoin. Journal of Asset Management, 16(6), 365-373.

Browne, S. (1998). The Return on Investment from Proportional Portfolio Strategies. Advances in Applied Probability, 30(1), 216-238. Retrieved from http://www.jstor.org/stable/1427885

Burton, B. (2019). Lego bricks outshine gold bars as investments, study finds. Retrieved from https://www.cnet.com/news/lego-bricks-outshine-gold-bars-as-investments-report-says/ [Accessed 2 May. 2019].

Burton, B. J., & Jacobsen, J. P. (1999). Measuring returns on investments in collectibles. Journal of Economic Perspectives, 13(4), 193-212.

Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. Journal of statistical software, 1-68.

Campbell, R. A. J. (2008). Fine violins as an alternative investment: Strings attached?. Pensions: An International Journal, 13(1-2), 89-96.

Carron, J. (2016). 67 Years of Lego Sets. Retrieved from https://mode.com/blog/lego-data-analysis

Chai, T., & Draxler, R. R. (2014a). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geoscientific model development, 7(3), 1247-1250.

Chai, T., & Draxler, R. R. (2014b). Root mean square error (RMSE) or mean absolute error (MAE)?. Geoscientific Model Development Discussions, 7, 1525-1534.

Chun Chen. (2017). Warner Bros. unveils licensing program for new Lego Batman movie (PHOTOS). Retrieved from https://www.bizjournals.com/losangeles/news/2017/01/19/warner-bros-licensing-program-lego-batman-movie.html [Accessed 4 May. 2019].

Countryman, M. (2015). Rauth Documentation. Release 0.7.2. November 2015.

Cumming, D., & Zhang, Y. (2016). Alternative investments in emerging markets: A review and new trends. *Emerging Markets Review*, *29*, 1-23.

Darlington, R. B. (1990). Regression and linear models (pp. 292-293). New York: McGraw-Hill.

Dimson, E., & Spaenjers, C. (2011). Ex post: The investment performance of collectible stamps. Journal of Financial Economics, 100(2), 443-458.

Dimson, E., & Spaenjers, C. (2014). Investing in emotional assets. Financial Analysts Journal, 70(2), 20-25.

Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. BMC medical genomics, 4, 31. doi:10.1186/1755-8794-4-31

Domingos, P. M. (2012). A few useful things to know about machine learning. Commun. acm, 55(10), 78-87.

Ebeling, A. “Revving Up Your Portfolio with Classic Cars,” Forbes.com, October 5, 2012.

Erdogmus, H., Favaro, J., & Strigel, W. (2004). Return on investment. Ieee Software, 21(3). doi:10.1109/MS.2004.1293068

Fogarty, J. (2008). A Mean-variance Approach to Wine Investment and Summary and Conclusion (Doctoral dissertation, Zugl.: University of Western Australia).

Geras, K., & Sutton, C. (2013, February). Multiple-source cross-validation. In International Conference on Machine Learning (pp. 1292-1300).

Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc.".

Goetzmann, W., Renneboog, L. D. R., & Spaenjers, C. (2011). Art and money. American Economic Review, 101(3), 222-226. https://doi.org/10.1257/aer.101.3.222

Gollier, C., & Weitzman, M. L. (2010). How should the distant future be discounted when discount rates are uncertain?. Economics Letters, 107(3), 350-353.

Götze, U., Northcott, D., & Schuster, P. (2008). Investment appraisal. Methods and Models, Berlin, Heidelberg 2008.

Graddy, K., & Margolis, P. E. (2011). Fiddling with value: violins as an investment?. Economic Inquiry, 49(4), 1083-1097.

Hierbrink, S. (2017). Ferrari. An Interesting Investment Option? Doctoral dissertation Financial Management. (2). 1 – 35.

Hyde, D. (2015). Lego a 'better investment than shares and gold'. Retrieved from https://www.telegraph.co.uk/investing/shares/lego-a-better-investment-than-shares-and-gold/. [Accessed 3 May. 2019].

Ichsani, S., & Suhardi, A. (2015). The effect of return on equity (roe) and return on investment (roi) on trading volume. Procedia - Social and Behavioral Sciences, 211, 896-902. doi:10.1016/j.sbspro.2015.11.118

Ind, N. (2014). How participation is changing the practice of managing brands. Journal of Brand Management, 21(9), 734-742.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Jones, C. (2010). House Price Measurement: The Hybrid Hedonic Repeat- Sales Method. Economic Record, 86(272), 95-97.

Jurek, J. W., & Stafford, E. (2015). The cost of capital for alternative investments. The Journal of Finance, 70(5), 2185-2226.

Kauflin, J. (2017). The Most Powerful Brands In 2017. [online] Forbes.com. Available at: https://www.forbes.com/sites/jeffkauflin/2017/02/14/the-most-powerful-brands-in-2017/#751ae750f1f8 [Accessed 13 Apr. 2019].

Khurana, U., Turaga, D., Samulowitz, H., & Parthasrathy, S. (2016, December). Cognito: Automated feature engineering for supervised learning. In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW) (pp. 1304-1307). IEEE.

Knight Frank. (2018). The Wealth Report [Ebook] (13th ed.). London. Retrieved from https://content.knightfrank.com/resources/knightfrank.com/wealthreport/2019/the-wealth-report-2019.pdf [Accessed 20 Apr. 2019].

Koford, K., & Tschoegl, A. E. (1998). The market value of rarity. Journal of Economic Behavior & Organization, 34(3), 445-457.

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

Kolm, P. N., Tütüncü, R., & Fabozzi, F. J. (2014). 60 Years of portfolio optimization: Practical challenges and current trends. European Journal of Operational Research, 234(2), 356-371.

Kort, P. M. (1990). A dynamic net present value rule in a financial adjustment cost model. Optimal Control Applications and Methods, 11(3), 277-282.

KPMG. Unknown Author. (2018). Alternative Investments 3.0 – digitize or jeopardize. Retrieved from https://assets.kpmg/content/dam/kpmg/xx/pdf/2018/01/alternative-investments-3-0-digitize-or-jeopardize.pdf. [Accessed 8 May. 2019].

Lauwaert, M. (2008). Playing outside the box–on LEGO toys and the changing world of construction play. History and Technology, 24(3), 221-237.

Lee, S. H. (2015, May 19). 7 Common Alternative Investments That All Investors Should Know. Retrieved from: https://www.thestreet.com/story/13155301/1/7-common-alternative-investments-that-all-investors- should-know.html [Accessed 17 Apr. 2019].

LEGO.com. (2017). The LEGO History. [online] Available at: https://www.lego.com/en-us/aboutus/lego-group/the_lego_history/1940 [Accessed 13 Apr. 2019].

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Macek, K. (2008). Pareto principle in datamining: an above-average fencing algorithm. Acta Polytechnica, 48(6).

Macik, R., & Macik, D. (2016). Trust and Product/Sellers Reviews as Factors Influencing Online Product Comparison Sites Usage by Young Consumers. Managing Global Transitions, 14(2 (Summer)), 195-215.ariably priced services. Journal of Hospitality & Tourism Research, 38(4), 562-581.

Maciorowski, E., & Maciorowski, J. (2016). The Collectible LEGO Minifigure: Values, Investments, Profits, Fun Facts, Collector Tips. F+ W Media, Inc..

Maciorowski, E., & Maciorowski, J. (2016). The Collectible LEGO Minifigure: Values, Investments, Profits, Fun Facts, Collector Tips. F+ W Media, Inc..

Magni, C. A. (2009). Investment decisions, net present value and bounded rationality. Quantitative Finance, 9(8), 967-979.

Martin, S. G. (2016). The Road Less Traveled: The Case for Collectible Automobiles as an Asset Class. The Journal of Wealth Management, 19(3), 131.

Masset, P., & Henderson, C. (2010). Wine as an alternative asset class. Journal of Wine Economics, 5(1), 87-118.

Masset, P., & Weisskopf, J. P. (2010). Raise your glass: Wine investment and the financial crisis. Available at SSRN 1457906.

McAndrew, C. (2010). The International Art Market 2007-2009: Trends in the Art Trade during Global Recession. European Fine Art Foundation (TEFAF). Belk R. (1995), Collecting as luxury consumption : Effects on individuals and households, Journal of Economic Psychology 16 : 477-490

McKinney, W. (2010, June). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51-56).

Minifigures Series (2018). Unknown Author. The top 10 most expensive LEGO minifugres of all time. - http://minifiguresseries.org/guide/10-most-expensive-lego-minifigures-of-all-time/ . [Accessed 6 May. 2019].

Nagaraja, C., Brown, L., & Wachter, S. (2014). Repeat sales house price index methodology. Journal of Real Estate Literature, 22(1), 23-46.

Noone, B. M., & McGuire, K. A. (2014). Effects of price and user-generated content on consumers' prepurchase evaluations of v

OLLAGNON, O. (2012). Emotional assets as an alternative financial investment: An efficient portfolio diversification solution during financial crises? (Doctoral dissertation, HEC Paris (L'École des hautes études commerciales de Paris)).

Olson, R. S., & Moore, J. H. (2016, December). TPOT: A tree-based pipeline optimization tool for automating machine learning. In Workshop on automatic machine learning (pp. 66-74).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

Pembrey, D. "The Classic Car Is the Investment Star." Financial Times, April 8, 2016.

Peng, L. (2012). Repeat sales regression on heterogeneous properties. The Journal of Real Estate Finance and Economics, 45(3), 804-827.

Popina. (2019). Bloomberg - The Hot New Asset Class Is Lego Sets. Retrieved from https://www.bloomberg.com/news/articles/2019-01-17/lego-collecting-delivers-huge-and-uncorrelated-market-returns [Accessed 2 May. 2019].

Renneboog, L., & Spaenjers, C. (2012). Hard assets: The returns on rare diamonds and gems. *Finance Research Letters*, *9*(4), 220-230.

Renneboog, L., & Spaenjers, C. (2013). Buying beauty: On prices and returns in the art market. Management Science, 59(1), 36-53.

Robertson, D., & Breen, B. (2014). Brick by brick: How LEGO rewrote the rules of innovation and conquered the global toy industry. Crown Business.

Ross, S. (1995). Uses, Abuses, and Alternatives to the Net-Present-Value Rule. Financial Management, 24(3), 96-102. Retrieved from http://www.jstor.org/stable/3665561

Roth, P. L., Switzer III, F. S., & Switzer, D. M. (1999). Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques. Organizational research methods, 2(3), 211-232.

Ruijter, M. (2017). Exclusieve LEGO Star Wars Detention Block Rescue onthuld voor SWC17 - Bouwsteentjes.info. Retrieved from https://www.bouwsteentjes.info/exclusieve-lego-star-wars-detention-block-rescue-onthuld-voor-swc17/ [Accessed 1 May. 2019].

Sanning, Lee W. and Shaffer, Sherrill and Sharratt, Jo Marie, Alternative Investments: The Case of Wine (November 2, 2006). Available at SSRN: https://ssrn.com/abstract=834944 or http://dx.doi.org/10.2139/ssrn.834944

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. Psychological methods, 7(2), 147.

Scheffer, J. (2002). Dealing with missing data. Unknown. Unknown.

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Skully, M. (2007). Alternative investments: definition, importance and risks. JASSA, (3), 35. Availability: https://search.informit.com.au/documentSummary;dn=200711915; ISSN: 0313-5934.

Small, K., Smith, J., & Small, E. (2012). An examination of diamonds as an alternative asset class: Do they have what it takes to make a portfolio sparkle?. The Journal of Wealth Management, 15(3), 67.

Smith, A. (2018). The 25 biggest LEGO sets ever - 2018. [online] The Collector. Available at: https://thecollector.io/features/2017/09/the-25-biggest-lego-sets-ever/ [Accessed 13 Apr. 2019].

Spyder. (2019). spyder-ide/spyder. Retrieved from https://github.com/spyder-ide/spyder [Accessed 25 Apr. 2019].

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj, 338, b2393.

The Guardian, Rupert Neate, 03/06/2018, Lego reports first sales fall in 13 years, saying 'there is no quick fix' - https://www.theguardian.com/lifeandstyle/2018/mar/06/lego-sales-fall-growth-job-cuts [Accessed 13 Mar. 2019].

Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & van Hijum, S. A. (2012). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?. Briefings in bioinformatics, 14(3), 315-326.

Travis E, Oliphant. A guide to NumPy, USA: Trelgol Publishing, (2006).

Tutz, G. (2011). Regression for categorical data (Vol. 34). Cambridge University Press.

Vaitkune, S. (2018). Alternative Investment Thrives In Todays Financial World. Global Ranking & Finance Review. Retrieved from https://www.globalbankingandfinance.com/alternative-investments-thrives-in-todays-financial-world/. [Accessed 11 May. 2019].

Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), 3-28.

Visser, J. (2019). LEGO blijkt betere investering dan aandelen. Retrieved from https://www.metronieuws.nl/in-het-nieuws/2019/02/lego-blijkt-betere-investering-dan-aandelen. [Accessed 3 May. 2019].

Vlaović-Begović, S., Momčilović, M., & Jovin, S. (2013). Advantages and limitations of the discounted cash flow to firm valuation. Škola biznisa, (1), 38-47.

Wilson, N. 2007, ‘Alternatives become ever more main stream’, AIMA Journal, Spring.

Willmott, C. J. (1982). Some comments on the evaluation of model performance. Bulletin of the American Meteorological Society, 63(11), 1309-1313.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 30(1), 79-82.

Willmott, C. J., Robeson, S. M., & Matsuura, K. (2012). A refined index of model performance. International Journal of Climatology, 32(13), 2088-2094.

Wisley, T. O., & Johnson, S. R. (1985). An evaluation of alternative investment hypotheses using non-nested tests. Southern Economic Journal, 422-430.

Yardeni, E. Abbott, J., Johnson, D., Quintana, M. (2019). Stock Market Indicators: 200-Day Moving Averages. Yardeni Research Inc., April 2019, 1-8