Lego Insights and Investment Application

Joshua Lowe
University of Canterbury
joshpaullowe@gmail.com

Yaquub Ali
University of Canterbury
abdiy2333@gmail.com

Yazeed Almutairi
University of Canterbury
almutairiyazeed22@gmail.com

1 Introduction

1.1 LEGO

LEGO is a globally recognised brand producing plastic pieces that can be assembled to form a wide range of models. LEGO is popular across a wide range of ages which stems from its ability to be themed. From Star Wars to Harry Potter, LEGO engages people with vastly different interests which has allowed it to grow into a multi-billion-dollar business (Fortune Business Insights, 2021). This project will leverage data from multiple sources to provide insight into the market dynamics of LEGO and provide an application to help investors make informed decisions about which sets to invest in.

1.2 Goals

LEGO's partnerships with entertainment franchises contributes to its success by resonating with global audiences (Fortune Business Insights, 2021). This has allowed LEGO to become a viable investment option for investors (Dobrynskaya & Kishilova, 2022). Through data analysis, the resale prices of LEGO sets can be predicted based on a variety of features. Cluster analysis will be used to group LEGO sets based on features and user ratings. Then, through machine learning, a Price Prediction Model will be created to predict the resale price of sets based on features such as piece count, theme, subtheme and minifigure count. Techniques such as random forest, linear regression and XGBoost will be utilised to create predictions of the average yearly return of a Lego set.

A key objective of this project is to use Amazon Web Services (AWS) as a cloud-based data storage to house all datasets required for the entire project. AWS provides scalability and security ensuring large volumes of data can be efficiently stored and accessed (Singh, 2024). Additionally, Dash by Plotly will be utilised to develop and interactive dashboard that integrates with AWS (Jean-Michel D, 2018). The dashboard will serve as the front-end interface allowing users to predict the investment potential of LEGO sets. Additionally, knowledge embedding will be used to make the dashboard more user friendly (Linjuan et al., 2022). This technique allows Large Language Models (LLMs) such as ChatGPT to access the results of the analysis conducted throughout this project. This data pipeline ensures a streamlined, end-to-end workflow where data is pulled from AWS, analysed, and presented through dash. This approach ensures the project is scalable and efficient.

1.3 Constraints

The success of the project depended on obtaining high-quality, consistent data from multiple sources. The data came in a mixture of structured, semi-structured and unstructured data. It ranged from downloadable comma separated value files (CSVs) to APIs. This required careful implementation and cleaning. The most significant obstacle faced was gathering retail and resale price data, which took much longer than expected. It would have been ideal to gather data from websites such as Brickeconomy, Bricklink and Brickinsights, but this was not possible due to paywalls, private API keys or a limited number of API calls. Many of these sites also did not allow scraping using CSS selectors. When attempts were made to do this, it was instantly blocked. We reached out to some websites that had private API keys, however many of these sites did not respond.

2 Data

The five V's of Big data are Volume, Variety, Velocity, Veracity and Value (Hiba et al., 2015). The data sets used in this project are relatively small in size, however the framework still provides a structured approach to understanding the data. The data came from three sources, Brickset, Brickowl and Rebrickable (Brickset Home Page, 2024; Rebrickable - Build with LEGO, 2024; Brick Owl - LEGO Marketplace, 2024; Tosic, 2021)

2.1 Volume

Volume refers to the amount of data which is stored (Hiba et al., 2015). This can range from bytes all the way up to petabytes. More data typically leads to more issues with storage and analysis however, in the case of machine learning, more data is typically better (Sarker, 2021). The Brickset data contains 15,634 rows and 36 columns which is large enough for analysis. It has a substantial number of missing values posing a challenge to get into a format that is suitable for analysis. The Rebrickable dataset has multiple datasets with varying sizes. The two sets that were mainly used for analysis contain information about the sets and the themes. The "output_sets" dataset has 23,131 rows and six columns and the themes dataset contains 465 rows with three columns. None of the Rebrickable datasets contain missing values. The Brickowl data only has 5304 rows and there are no missing values. The overall Volume of the datasets is large enough to create insightful analysis but not big enough to case issues for running analysis on the local devices used throughout the project.

2.2 Variety

Variety is the diversity of data types. This refers to unstructured, semi-structured and structured data, as well as the column types of each dataset (i.e integers or strings). The Rebrickable data is structured, and a schema can be found online meaning joins between various datasets within it can be performed. Both the Brickowl and Brickset datasets do not have a schema but do come in a tabular format with rows and columns, making them semi structured. Across all the datasets there is a variety of data types, particularly string, integer and date types. The varied nature of these datasets means careful data integration is required.

2.3 Velocity

Velocity refers to the speed at which the data is generated. The velocity of the LEGO data is relatively low. All datasets are updated periodically rather than in real-time. However, the dashboard implemented for this project gets input from the user. This means there is some degree of dataflow. When users are using the dashboard, it is essential that the experience is smooth and efficient. There are various techniques to reduce latency and make the front-end user friendly. Machine learning algorithms should be executed in advance, meaning that when a user requires an output form the algorithm the response will be instant.

2.4 Veracity

Veracity is the quality of the data. Missing values, outliers, variance and imbalances all impact the Veracity of the data. Both Brickset and Brickowl contain a large number of missing values. For

Brickset the columns “Launch Data" and “Exit Date” nearly two-thirds of the rows contain missing values. These missing values could have a significant effect on the results if not taken into consideration and dealt with accordingly. For all columns (variables) of the Brickset data and the Brickowl data, the percentage of outliers is less than 1%. This is relatively low and likely has little effect on results of the analysis. The Rebrickable data has a slightly higher outlier percentage with the number of pieces columns having 1.68% outliers. This is again relatively small and will have little effect on the results. None of the columns in any of the datasets display near zero variance. This is to be expected due to the huge variation in values expected when dealing with price data, piece count, set number, theme type, etc. The datasets do contain some classification imbalance, some themes have a lot more sets than others. For instance, “Star Wars” has 822 compared to “The Simpsons” which only has two. This can have a significant effect on the results of the analysis because if the two sets of “The Simpsons” have a dramatic price increase then it would indicate this theme is a strong predictor for resale value but in reality, it's too small of a sample size to be a statistically meaningful indicator.

3.4 Value

Value refers to the actionable insights that can be derived from the data. The combination of Brickset, Rebrickable, and Brickowl data has imperfections, but offers substantial value for exploring LEGO trends and predicting price. The Brickset data in particular had data on retail and resale price allowing for price prediction to be carried out. Each dataset has unique aspects that when used in conjunction provide a detailed dataset to perform analysis on.

3 Methodology

3.1 Data Pipeline

Extract, Transform and Load (ETL) is a data engineering process used to collect data from multiple sources, clean it and then load it into a target system for further analysis/storage (Souibgui et al., 2019). This process ensures that the data is organised, cleaned and structured properly so it can be analysed for meaningful insights. Amazon Web Service (AWS) is a cloud service provided created Amazon to assist with ETL. This allows for an efficient data pipeline which is essential for creating a scalable application. The ETL process overview can be seen in Diagram 1.

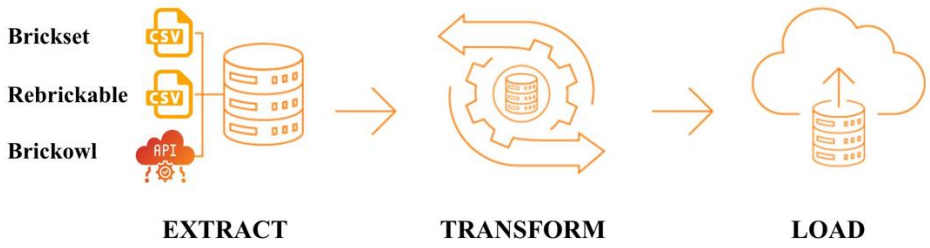

Diagram 1. *Diagram of Extract, Transform and Load Process*

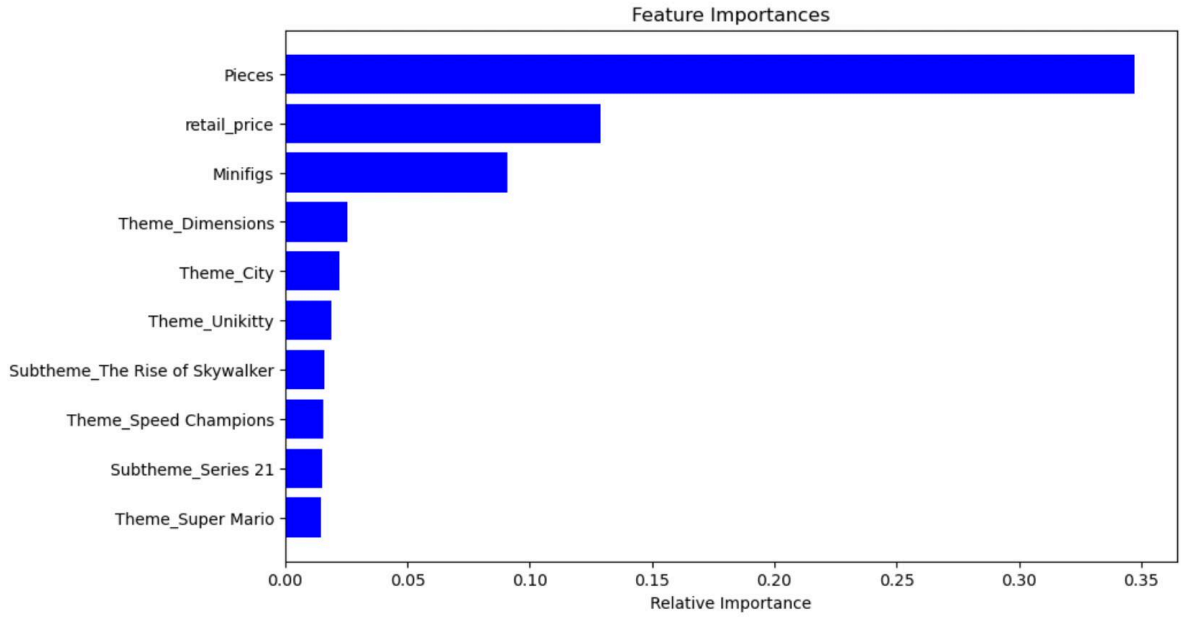

Figure 3: *Feature importance in predicting percentage growth for LEGO*

In figure 4, we observe the top 10 LEGO themes by their predicted percentage growth per year, which indicates how much these themes have increased in value over time.

Speed Champions takes the top spot with the highest average annual growth, followed closely by Jurassic World and Minecraft. These themes appear to consistently appreciate, suggesting strong demand in the resale market.

Other themes such as Creator Expert, Ideas, The Hobbit, The Lord of The Rings and Pirates of the Caribbean also exhibit notable growth, reinforcing their potential as solid investments for LEGO collectors.

Based on the RMSE of 0.276 the data means the predicted percentage growth is off by roughly 0.27% and the R^2 score of 0.998 means 99% of the variance in the data is explained. Which means this model is predicting percentage growths per year extremely well for themes.

However, it is important to consider the dataset's variability, as the themes Monkie Kid and Ghostbusters both have fewer than three entries in the data, which could skew their average percentage growth. Such a small sample size may not provide a reliable picture of long-term trends, and the high growth rates observed could be driven by outliers or specific sets rather than a consistent pattern across the theme.

MAE	0.075
RMSE	0.276
R^2	0.998

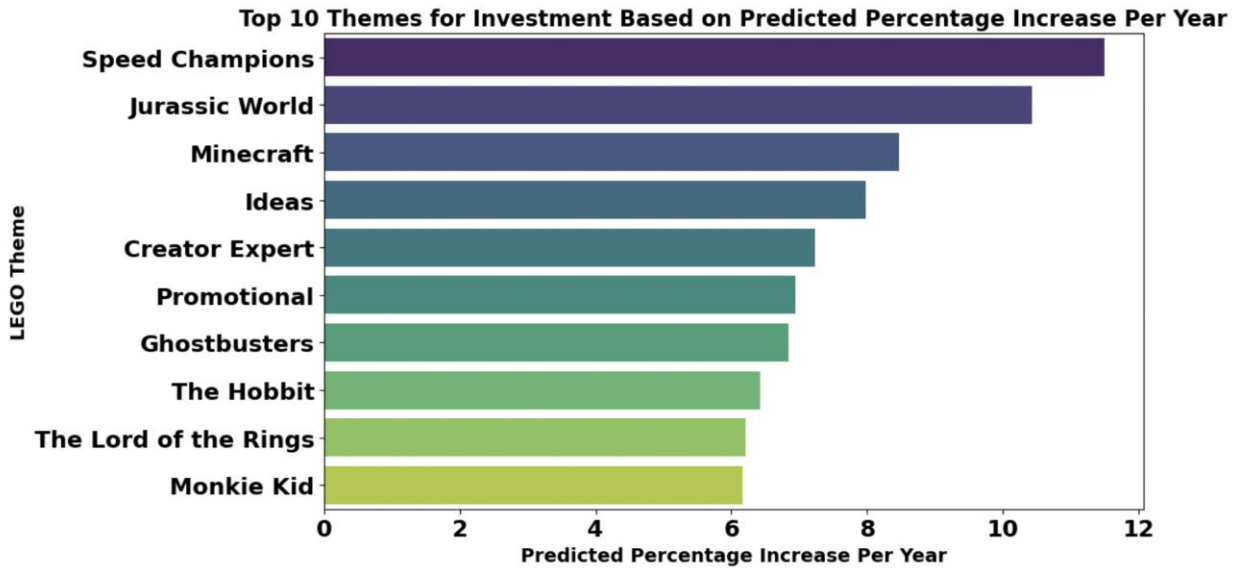

Figure 4: *Top 10 LEGO themes by predicted percentage growth per year*

In figure 5, we see the top 10 subthemes with the highest average percentage growth per year. As we can see this graph was dominated by speed champions with 4 out of the 10 subthemes belonging to speed champions. With those being Mclaren, Ford, Porsche, Cheverolet. Jurassic World and Ninjago both had multiple subthemes with those being Fallen Kingdom, Sons of Garamadon, Legend of Isla nubar, hunted and the hands of time.

Based on the RMSE of 0.408 the data means the predicted percentage growth is off by roughly 0.40% and the R^2 score of 0.997 means 99% of the variance in the data is explained. Which means this model predicts percentage growths per year extremely well for subthemes.

To ensure a more reliable analysis, subthemes with fewer than 10 entries were filtered out. This adjustment was necessary because, without this filter, all of the top 10 subthemes initially had less than 5 entries, which could have significantly skewed the results. By focusing only on subthemes with a larger number of entries, we obtain more dependable data, making the findings more representative of overall trends in the LEGO secondary market.

MAE	0.075
RMSE	0.408
R^2	0.997

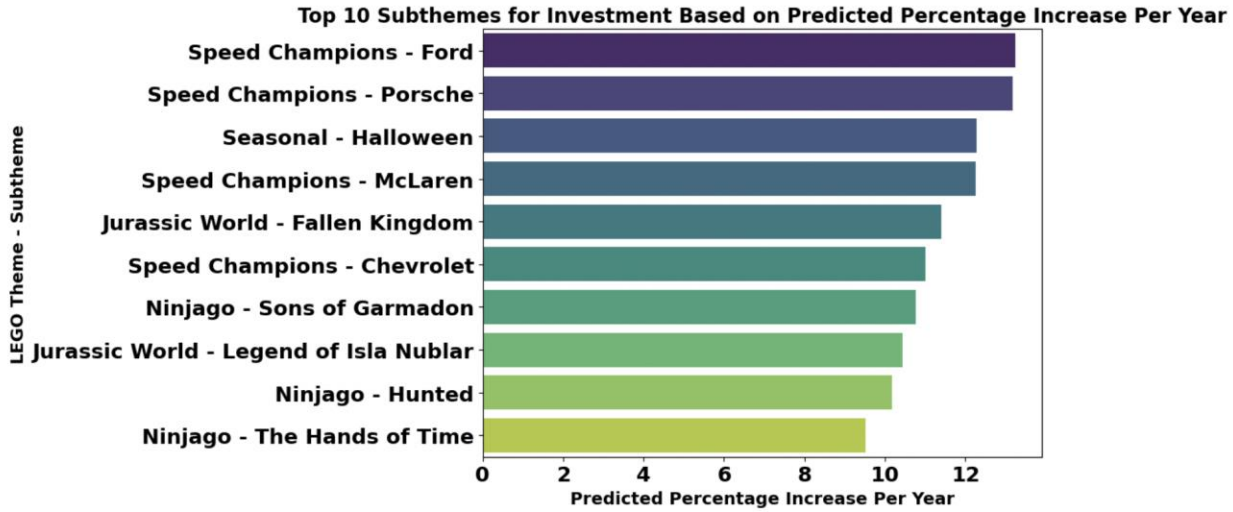

Figure 5: *Top 10 Lego subthemes by predicted percentage growth per year*

Cluster Analysis

This section focuses on the results of our clustering analysis and is aimed at identifying distinct investment profiles within the LEGO market. We analyse features like yearly price growth, resale price, number of pieces, and years in retirement to uncover which variables are most influential in distinguishing between different clusters of LEGO sets. The analysis includes a series of visualizations, which gives a clear overview of trends and patterns within the data, giving beneficial insights into how specific LEGO sets perform in the resale market. We will then evaluate the effectiveness of the KMeans algorithm in categorizing LEGO sets into meaningful clusters. We will also touch on the decision to use Kmeans, along with the challenges that came with it, and the methods explored to determine how many cluster groups we will use.

From figure 6 and 7 below, we can see that there are some clear differentiation factors between our clusters. Figure 6 shows the relationship between the number of pieces in a LEGO set and its retail price in USD, with points color-coded to represent the different clusters identified through the Kmeans algorithm. Figure 7 shows general summary statistics for each cluster. We can see that cluster 2 is dominated by high-priced and large-piece sets, and clearly stands out with sets that have an average of 1402 pieces, a mean resale value of $412.16 USD, and a retail price of $140.34 USD. Cluster 1 and 3 look to have relatively similar piece counts at around 200 and retail prices at around $26, however, cluster three has a significantly higher average resale price at $109.23 compared to cluster 1's $43.74. This could potentially mean that cluster 1 and 3 mostly include more generally affordable sets, with decent investment potential, though not as premium as cluster 2.

Figure 6: *Scatter Plot of Retail prices against pieces per cent, by cluster.*

Cluster Name	Average Price Growth (%)	Average Resale Price (USD)	Average Pieces	Average Years in Retirement	Average Retail Price
Cluster 1: Low Price, Moderate Growth	4.08%	$43.74	203.38	6.15	$26.25
Cluster 2: Large Premium Sets, High Growth	6.90%	$412.16	1401.82	8.55	$140.34
Cluster 3: Retired Collectibles, Modest Growth	3.81%	$109.23	207.67	18.66	$27.96

Figure 7*: Summary Statistics Per Cluster*

From figure 8, we can see how the distribution of top themes varies across clusters. Cluster 1 shows a more balanced distribution of themes, indicating that sets from a wide variety of themes are present here. This cluster seems to have a mix of sets from both popular themes and more niche ones. In Cluster 0, Collectable Minifigures dominate with 488 sets, followed by City (357 sets) and Star Wars (347 sets). These themes are widely popular and probably attract a wide range of collectors and investors. Cluster 1 looks to be dominated by Star Wars, along with a small portion of sets from Collectable Minifigures, Marvel Super Heroes and Harry Potter. Like Cluster 1, Cluster 3 also has a relatively even distribution of themes, however with the majority being from Star Wars, and Town. These sets might not see as much price growth as others, for reasons that will be discussed later in the report, but they continue to be popular for buyers.

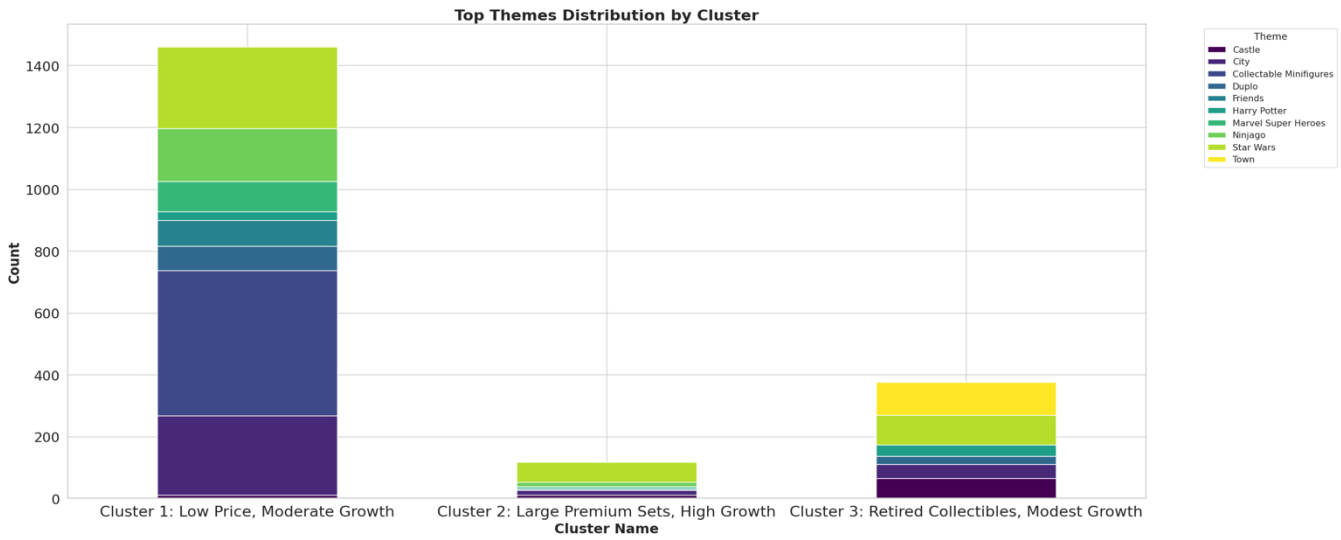

Figure 8: *Stacked Bar Chart of Theme Distribution By Cluster*

Figure 9 shows the price growth percentage per year for each cluster against each other. Figure 10 shows the average price growth per year against years in retirement for each cluster. Figure 9 clearly shows that Cluster 2 has the highest price growth per year at 6.9%, which is close to double cluster 1 and 3, at 4.08% and 3.81%, respectively. As seen earlier in figure 7, the relatively large and premium nature of the sets within cluster 2 could be driving this. Star Wars, Harry Potter, and Marvel Super Heroes are examples of themes that drive this extreme growth, suggesting they are highly sought after in the resale market despite their significantly higher prices. Figure 10 also shows that sets within this cluster have been in retirement for an average of 8.55 years. Cluster 1, with a 4.08% price growth per year, comes second. As seen in figure 9, sets in cluster 1 have on average, been in retirement for the least amount of time at 6.15 years. Cluster 3, on the other hand, with 3.81% yearly price growth, mainly includes sets that have been in retirement for a very long time at an average of 18.66 years.

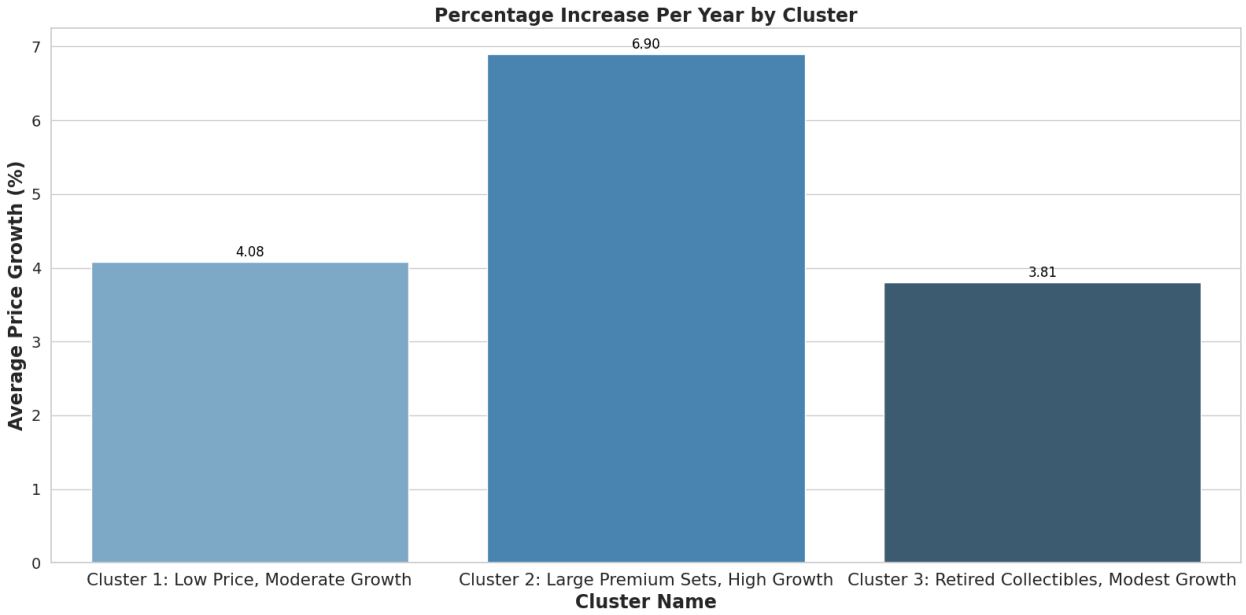

Figure 9: *Percentage Increase Per Year By Cluster*

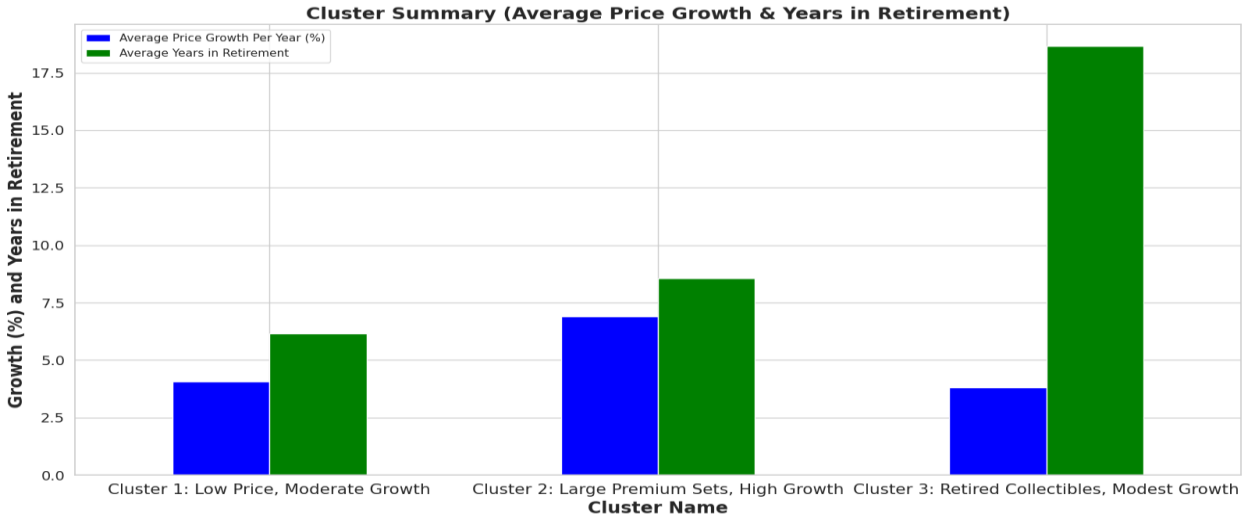

Figure 10: *Average Price Growth Per Year Against Years In Retirement By Cluster*

As certain clusters contained disproportionately larger amounts of data than others, we wanted to get a general measure of which themes performed the best across all our clusters. From Figure 11, we can see that the majority of the top-performing themes seem to belong to either popular TV shows, movies, and video games. For example, we can see that the theme ‘Speed Champions’ dominates with what looks to be around 13% annual price growth, followed closely by ‘Jurassic World’ and ‘Minecraft’. Looking through the rest of the themes on the left we also see other fan favourites such as ‘Star Wars’ and ‘The Lord Of The Rings’. Overall, I believe that cluster analysis performed very well on our data and did a good job of segmenting our clusters based on unique, and easily understandable characteristics.