Project Brief

Samuel Love, 84107034

Table of contents

Domain

I am working on a project about Lego, a worldwide brand of plastic construction toys produced by The Lego Group since 1949. The project and supervision is provided by the University of Canterbury data science programme. The project title is LEGOscape: A Holistic Data Exploration of Sets, Sentiments, and Systems in the LEGO Universe.

The Lego Group produces sets that are categorised into themes. Themes are created by LEGO (e.g. Lego City) or licensed from popular franchises (e.g. Star Wars). The Lego Group constantly introduces new sets and ideas whilst discontinuing old sets. As such there is a wide variety of Lego in the global market.

A challenge of working with Lego data is producing a focused project due to an abundance of available data and research. However, there remain questions that exploration of Lego data can seek to answer. Projects have high flexibility for goal and scope.

Goal

This project will use machine learning to develop a model for predicting the retail price of Lego sets, dependent on characteristics such as theme, number of pieces, year released, and condition of the set.

The main question is, "Is it possible to accurately predict the retail prices of Lego sets according to a defined set of characteristics?"

The goal of this project is to produce a model that has practical applications for Lego collectors and investors who want to know which characteristics influence the retail price of desired Lego sets.

To achieve this goal, objectives include:

- Time series analysis of Lego characteristics.
- Data visualisation to assist with data cleaning and to display relevant statistics.
- Feature ranking to determine the significance of each characteristic.
- Predictive modeling of Lego sets retail price via supervised learning.

Additionally, there is room to increase the scope of this project by including new objectives such as sentiment analysis for increased context or comparing across themes, especially licensed vs Lego original.

Data

There is an abundance of available Lego data. Notable websites include:

- Rebrickable: Find what sets can be created by combining other sets.
- Brickset: Find relevant information including reviews.
- Bricklink: Trade Lego parts.

Rebrickable provides csv files of their database for parts, sets, and colours. There are 12 files in total, all connected via keys.

Brickset can be interacted with via their API. Jason Bryer, Ph.D. has created an R package jbryer/brickset which contains the legosets dataframe consisting of 18 455 observations of 34 variables including sets, years, themes, pieces, and US retail price.

I will use the United States of America (US) Lego market since it is one of the world's largest markets and the United States Dollar (USD) is commonly used as a standard currency.

Also, since the data includes different time periods, standardising the price by inflation is sensible. Changes to the US Consumer Price Index (CPI) provides the inflation rate. The Federal Reserve Bank of Minneapolis has a chart of annual average CPI and annual percentage change (inflation rate). This will be useful to compare standardised prices for Lego sets.

Plan

My outline is as follows:

Week	Goals
1	Read relevant literature for inspiration
	Plan project goals and methods
	Explore available data
	Refine course notes and setup project files
2	Collect raw data
	Explore raw data
	Clean raw data
3	Prepare the data for plotting
	Produce time series visualisations of the data
	Produce other relevant visualisations of the data
4	Refine the visualisations and make them interactive (Rshiny)
5	Prepare the model
	Begin training the model
B1	Fill gaps
B2	Fill gaps
6	Continue working on the model
	Make the model interactive (if possible)
7	Finish working on the model
8	Draft the report (include literature review)
	Fill gaps/expand scope (time dependent)
9	Prepare deliverables
	Fill gaps/expand scope (time dependent)
10	Prepare deliverables
	Fill gaps/expand scope (time dependent)

This schedule has a logical progression with appropriate time planned for each stage. The last three weeks are designated to summarising the results. If I run into unexpected difficulties, I have time available to resolve them. Alternatively, if everything goes according to schedule, I have time available to increase the scope of the project.

Resources

I will use RStudio for the entirety of this project. R is both my favourite and most practised coding language. Additionally, R is very popular and well-supported with many packages and tutorials at my disposal.

I will specifically make use of Shiny for visualisations and interactive environments, and Quarto for documentation. Time permitting, I will read and discuss other peoples work to give this project broader context. I will also use code snippets and tutorials from various forums and websites.

Additionally, ChatGPT is a useful tool for quickly understanding new concepts, generating generic code, and proofreading. I plan to use it with discretion.

Constraints

This project has no obvious constraints. Consideration is made when using data collected by other people, however, many authors are happy to share their works for educational purposes.