This project was developed as part of the selection process for the Data Business Enabling Analyst position within the LEGO Design: Capacity, Delivery & Analytics department. The primary objective was to showcase an end-to-end analytical workflow, demonstrating my ability to translate raw data into a clear business strategy.

The assignment required analyzing a fictional, small, dataset of toy prototypes to deliver actionable recommendations to the Head of Product Design. The task involved formulating a precise research question, defining target KPIs, and synthesizing the findings into a presentation of maximum five slides supported by three to four key visualizations.
Additionally, I was required to draft a specific note to the stakeholder to address the business impact of the analysis.

To execute this, I utilized Python for data validation, cleaning, and statistical correlation checks, followed by Power BI for dashboarding and visual storytelling. *(Note: The dataset used contains fictional data provided specifically for this case study).*

ToyCo Prototype Analysis

- **Stakeholder**: Head of Product Design

-Note for Stakeholder: These findings will help you select the launch strategy that maximizes Long-Term Engagement and builds Brand Loyalty with both children and parents. Additionally, the analysis identifies niche opportunities to reposition underperforming products, ensuring we maximize ROI on our R&D spend.

- **Research Question**: Within the 5 new toy prototypes, which one should be released to the global market to maximize Long-Term Engagement?

- **Targeted KPI for the Analysis**: Retention Rate (RepeatPlay variable)

Cleaning & Statistics with Python

	Toy	AgeGroup	FunRating	Difficulty	RepeatPlay	Region
80	Toy A	3–5	3	3	No	East
84	Toy A	3–5	3	3	No	East

Checking for null values and duplicates I actually found a duplicate row, specifically row 80 & 84.

81	Toy E	6-8	2	2	Yes	East
82	Toy A	3-5	3	3	No	East
83	Toy C	3-5	1	3	No	West
84	Toy D	6-8	3	4	Yes	North
85	Toy E	6-8	4	2	Yes	East
86	Toy A	3-5	3	3	No	East
87	Toy A	9-11	4	4	No	East

I then went to look and highlight the rows on excel, since the dataset was small and it was just 2 rows, for a more immediate visualization.

My past Python project:
https://github.com/tancripdpm/data-analysis-and-machine-learning-class-project

	AgeGroup_Numeric	FunRating	Difficulty
AgeGroup_Numeric	1.000000	0.233784	0.048012
FunRating	0.233784	1.000000	-0.149451
Difficulty	0.048012	-0.149451	1.000000
RepeatPlay_Numeric	-0.097714	0.128178	0.063915

	RepeatPlay_Numeric
AgeGroup_Numeric	-0.097714
FunRating	0.128178
Difficulty	0.063915
RepeatPlay_Numeric	1.000000

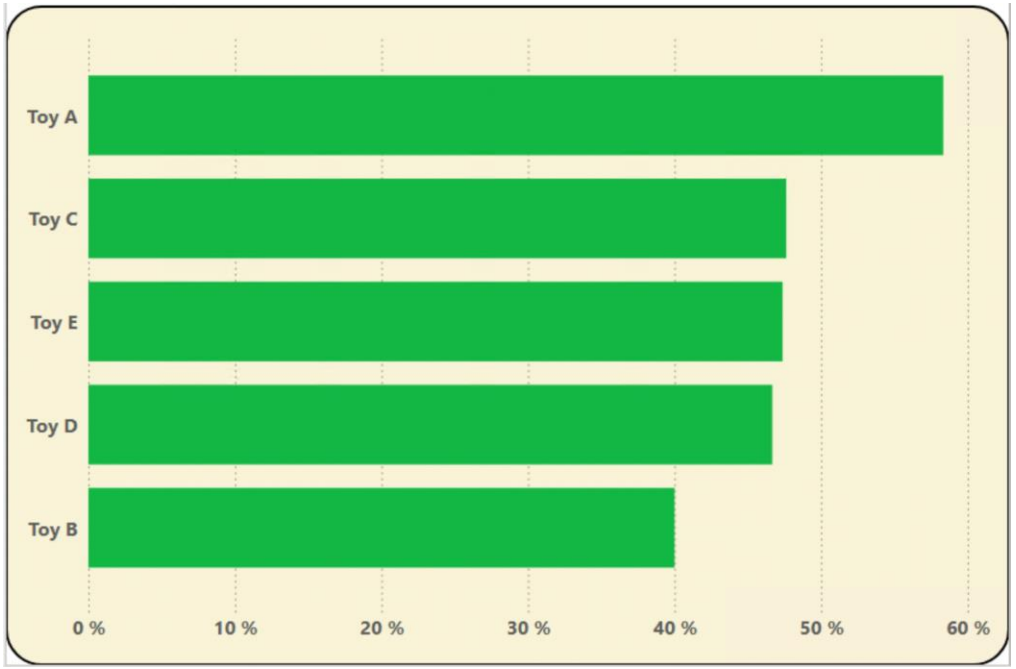
All columns of the dataset appear to not be correlated with each other, this means, for example, that difficulty does not influence (significantly) the fun rating. (at least in this first analysis)

The highest correlation number (although still very weak) is beetween AgeGroup and FunRating, meaning older children tend to appreciate the prototypes more than the younger ones.

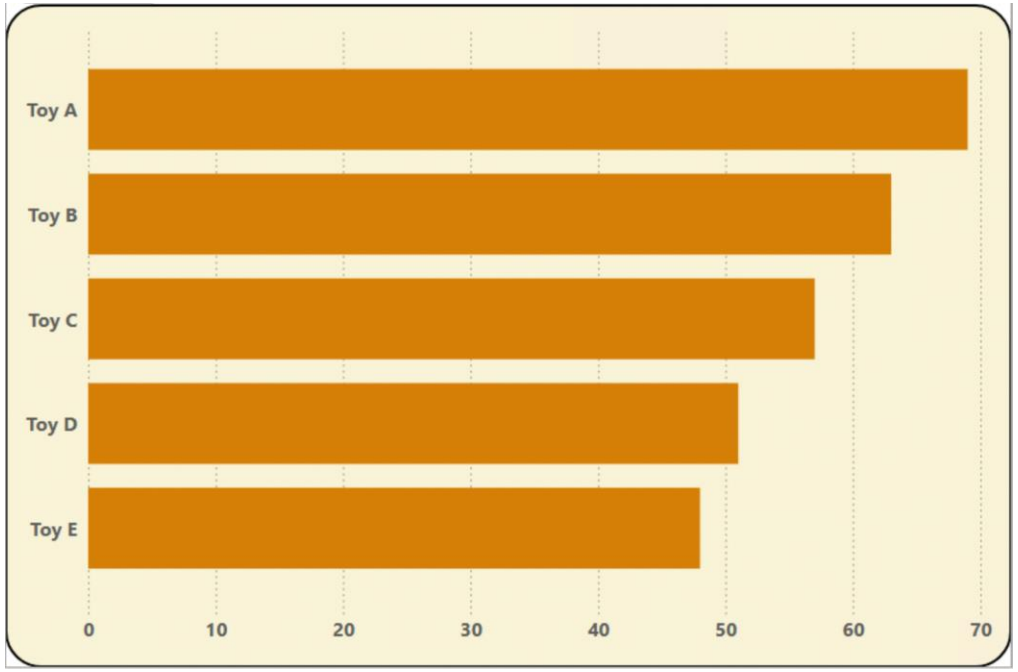
Link to the whole Python code: statistics.ipynb

Power BI Visualizations

% Repeat Play by Toy



FunRating by Toy



As seen in Python's correlation analysis, fun rating does not influence Repeat Play, meaning a toy rated very fun (such as Toy B) can be at the same time, the toy children want to play with again the least.
However, in our particular case we have a clear winner: **Toy A** not only has the highest percentage of repeat play, but also the highest fun rating, meaning is the prototype to be released to maximize **Long-Term Engagement** (answering our research question) but also **Short-Term Engagement.**

More Visualizations

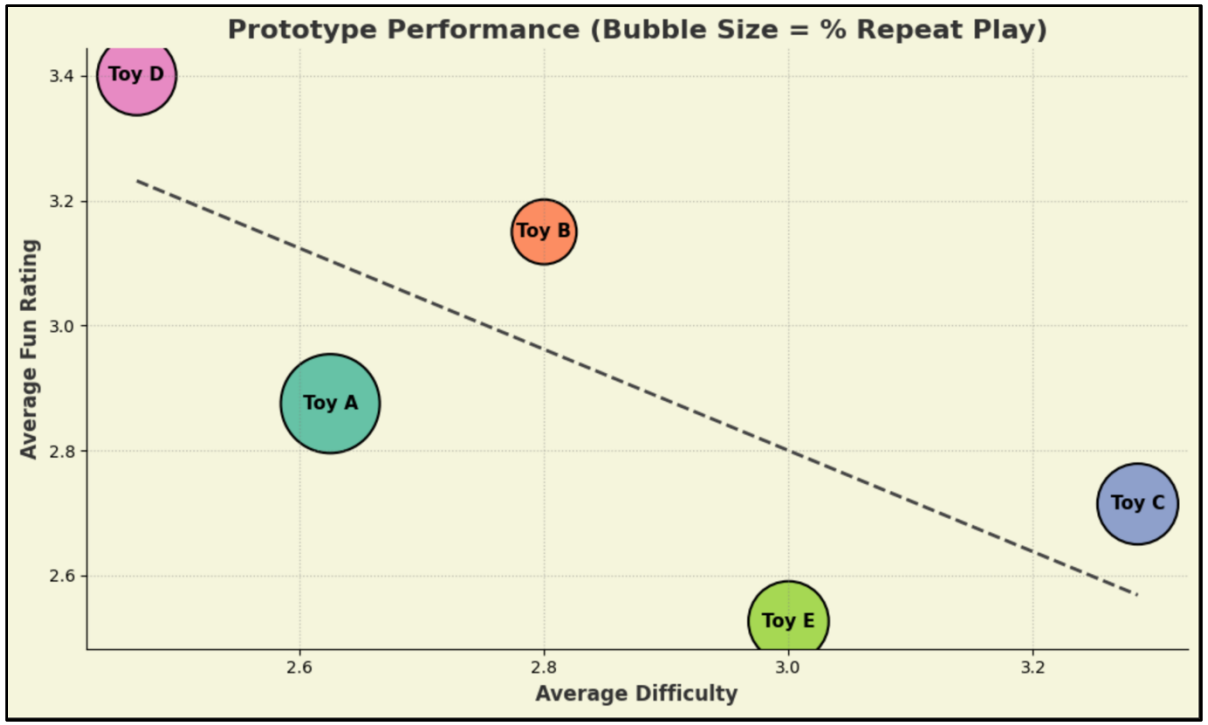
Toy A is the most successful in general, but in an analysis considering different age groups the result changes completely: it is the least favorite among children aged between 6 and 8.

This is the type of insights you can acquire just through visualizations: as previously seen in Python, difficulty and fun rating are weakly correlated with each other, however we can clearly see a trend here where toys with increased difficulty are more likely to be rated less fun (while the bubble size, repeat play, appears to be unaffected by the incresed difficulty).

My past PowerBI project:
Pizza Chian Simulation (Quick Overview).pdf

% Repeat Play by AgeGroup

AgeGroup_Numeric	Toy A	Toy B	Toy C	Toy D	Toy E	Total
1 3-5	81,82 %	20,00 %	50,00 %	66,67 %	42,86 %	56,76 %
2 6-8	16,67 %	50,00 %	42,86 %	33,33 %	60,00 %	41,38 %
3 9-11	57,14 %	42,86 %	50,00 %	33,33 %	42,86 %	45,45 %
Total	58,33 %	40,00 %	47,62 %	46,67 %	47,37 %	48,48 %



Actionable Insights and Recommendations

- **Engagement beats “fun”**: A high fun rating does not imply a high engagement rate. If an obvious choice, such as Toy A, with a high fun rate and a high engagement rate wouldn’t exists, the next best choice would be Toy C, even tho it has a lower fun rating then Toy B.
- **Averages hide Opportunities**: While Toy A is on average the best Toy among all age groups, for the 6-8 age group is actually the one performing the worst, while toy E performs the best in this category. Not releasing Toy E could lead to a missed “niche market” opportunity, such as the age group 6-8.
- **Difficulty is not a Barrier**: While difficulty, at the end, does influence fun rating, it does not influence Repeat play, meaning a ”simple toy” doesn’t lead to a longer engagement rate.