**Project 2 Proposal - Lego**

Adara Liao

Alexandra Iov

Rachael Burns

**Summary**

The Lego toy captures the imagination of both children and adults around the world and have delighted generations for more than 80 years. Like it’s young customers, Lego has matured, especially in how they select and market Lego sets. In recent years, Lego has entered the realm of games, TV and movie merchandising and further raised its profile through films featuring characters built entirely from Lego blocks. With the growing and enduring popularity of Lego and its importance in shaping childhood experiences, we will examine Lego data to ask burning questions on how the composition of Lego sets have changed over time, which colors are trending over time and how the company has performed financially. We used the publicly-available Lego dataset from Kaggle (covering data on lego parts from 1950), data on top grossing films for each year from Box Office Mojo, and financial information from Lego’s company website (covering high-level financials from 2003).

**Links**

<https://www.kaggle.com/rtatman/lego-database>

<http://www.boxofficemojo.com/yearly/chart/?yr=2017&view=releasedate&view2=domestic&sort=gross&order=DESC&&p=.htm> (going to have to scrape data of top 100 grossing films for each year)

Financial data shared on github playground: [github-playground](https://github.com/MIDS-INFO-W18/github-playground)/lego\_usd.csv

Link to denormalized data file on Google Drive: <https://drive.google.com/open?id=1ne6tNwhymh4rXcjc-g9N6ti-16Fz9QgK>

**Dataset Structure**

* Kaggle dataset was 8 separate files
* Denormalized dataset has 21 columns and 580,251 entries.
  + Some missing values.
  + 364 part IDs were one to many between tables, signified with a, b, c etc appended on to the part ID – we decided to use regular expression matching and concatenate the part names for all matches to handle these cases.
* Financial data contains 27 rows with 15 entries
* Movie data has 100 entries fojr each year since 1980 with 2 crucial columns

**Initial Exploration**

* How has the complexity of sets, in terms of number of colors and number of pieces, changed over time?
* Which colors have risen or fallen in popularity?
* How has the frequency of pink pieces and sets with a high proportion of pink changed over time? How might this reflect changes in attitudes towards marketing to girls?
* How many themes correlate with top box office films? How has that number changed over time?
* What was the box office rank each film for which Lego created a set? What was the average rank? How has that changed throughout the years?
* How has revenue and profitability changed over time?

Profitability

We can see that profitability increased sharply in 2006 even with modest increases in revenue. It drops as soon as revenue growth declines. It seems that profitability accelerates with change in revenue rate of change.

[Note: need to flip the time periods in x-axis]

