Tutorial 6 Worksheet AY 22/23 Sem I

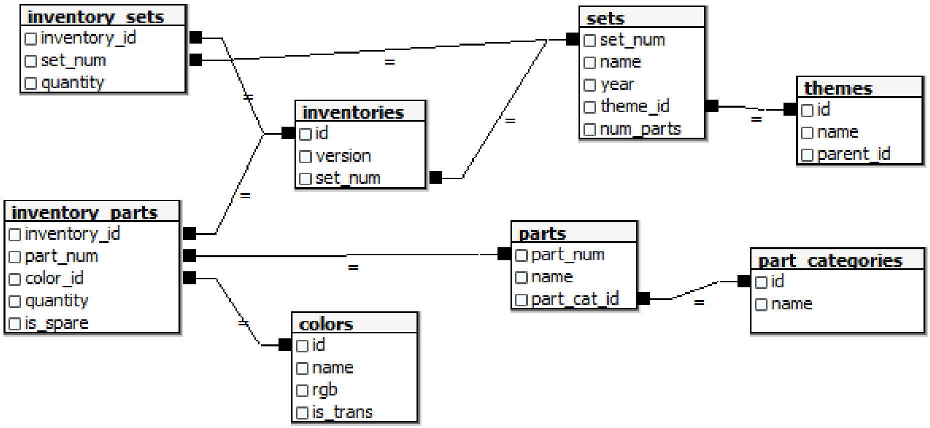
DSA2101

Lego Dataset

The dataset that we shall work on, can be downloaded from Canvas. This dataset contains the LEGO Parts/Sets/Colors and Inventories of every official LEGO set in the Rebrickable database. These files were current as of July 2017.

LEGO is a brand of toy building bricks. LEGO bricks are often sold in "sets", which allow the owner to build specific objects, for instance an X-wing fighter from Star Wars Episode IV, or Bilbo's home from The Hobbit. Each set can be classified under a theme. Each set contains parts, which have a part number, and a category. Parts also differ in terms of their colour. There is also an inventory of sets and version numbers.

The dataset is provided in eight csv files. Here is the schema that identifies how the files are linked together.



Use the tables to answer the following questions:

1. Consider the tables in `sets.csv` and `themes.csv`. Suppose we extract 3 major themes:
   - Star Wars: `id` or `parent_id` between 158 and 185.
   - Lord of the Rings (LotR): `id` between 561 and 569.
   - Superheros: `id` between 482 and 493.

   Re-create the following table, which shows the number of different sets in each decade:

major_theme	[1950,1960)	[1960,1970)	[1970,1980)	[1980,1990)	[1990,2000)	[2000,2010)	[2010,2020]
Other	132	303	564	1030	1622	3317	3837
Star Wars	0	0	0	0	13	237	342
Superhero	0	0	0	0	0	30	206
LotR	0	0	0	0	0	0	40

2. Still working with `sets.csv` and `themes.csv`, which are the themes that do not have any associated sets?

3. When we perform a left join between `sets` and `inventories` tables, the resulting table has more rows than the original `sets` table. Why? Identify the cause for this.

4. The `rgb` column in the `colors` table contains the RGB specification in hexadecimal format. Reshape the table into the following format:

id	name	is_trans	channel	hex	dec
-1	Unknown	f	R	00	0
-1	Unknown	f	G	33	51
-1	Unknown	f	B	B2	178
0	Black	f	R	05	5
0	Black	f	G	13	19
0	Black	f	B	1D	29

5. Collapse each colour in the tidy data frame from Q4, to contain the predominant channel(s) in each named LEGO colour. The dominant channel for each named colour is computed by comparing the values for each channel R,G, and B, and keeping only the ones with the largest value.

name	dom_col
Magenta	R
Trans-Purple	B
Medium Lime	G
Trans-Yellow	R
Medium Violet	B
Fabuland Brown	R
Speckle Black-Silver	RGB
Bright Light Orange	R
Very Light Gray	R
Sand Green	G

6. Which sets (from `sets.csv`) do not have any Bricks (from `part_categories.csv`) in them?