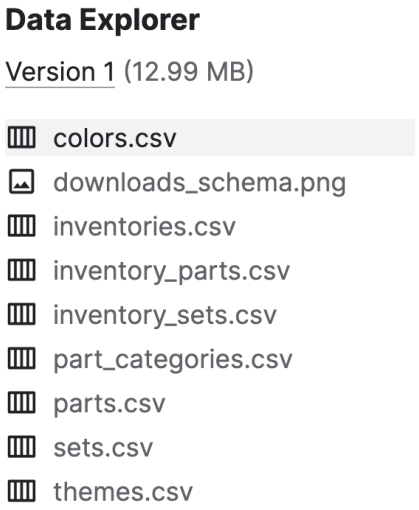
112-1 Database - HW1

LEGO is a popular brand of toy building bricks. They are often sold in sets in order to build a specific object. Each set contains a number of parts in different shapes, sizes, and colors. This database contains information on which parts are included in different LEGO sets. It was originally compiled to help people who owned some LEGO sets already figure out what other sets they could build with the pieces they had. Source: https://www.kaggle.com/datasets/rtatman/lego-database/data

HW1 requests:

1. Launch a database service with PostgreSQL engines or other DBMS.
2. Create a "lego" database
3. In the lego database, create tables for eight .csv files provided by the data source with public schema, and set a suitable primary key set for each table (you can refer to the schema provided by the data source). **Due to some issues with the data, please ignore the foreign key settings for "part_num" column.**



4. Try to write queries in Query Tool to
    a. extract the name of the set and name of the theme of all the LEGO sets published in 2017 Example of a LEGO set.
    b. extract the total number of LEGO sets in each year from 1950 to 2017, in descending order of total number of LEGO sets
    c. extract the name of the most popular theme, defined by the number of sets in the themes. Example of a theme.
    d. extract the average number of parts in a set for each theme, with the name of the theme and the average number of parts per set. In ascending order of average number of parts in a set
    e. find out the name of the colors that are most used in the unique LEGO parts, and list the top 10. Example of a LEGO part. Unique means if we have 10 green Plate 1 x 1, it should be counted as 1 green part. If we have 10 green

Plate 1 x 1 and 10 green Plate 2 x 2, then it should be counted as 2 green part.

   f. find out the name of the colors that are most used in the LEGO parts, for each theme, and list the top 1 for each theme (please provide the name of the theme, too). Different from the “e”, if we have 10 green Plate 1 x 1, it should be counted as 10 green parts. **Hint: Quantity**

5. Please **describe** the detailed process, including launching the DB, creating the new tables, setting the attributes, data type, and keys, and the query process with query results in a PDF file, following the format we provided. ~~(extra 10pts for redesigning the table schema and detailed description)~~ Due to the holidays, we haven’t had a chance to introduce Ch7, so this sentence is removed.
6. submit the PDF file to the E3 system HW1 section by **2023/10/26 23:59**

HW1 format:

1. The process of creating the “lego” databases (can be screenshot and/or SQL/non-SQL statements with text explanation) (8pts)
   Ans:

2. The process of importing eight required .csv files into lego database (can be screenshot and/or SQL/non-SQL statements with text explanation). Please include/describe the data type and keys of the imported table in your screenshot, SQL statements, and explanations (32pts) Note: if you import these eight tables with the same process (but of course the data type and keys are different), you can directly reference the duplicated steps if you have described them in the previous table. 如果表格輸入部分有重複的步驟，你可以直接寫同xxx表格輸入的步驟x~y。(update: 2023/10/15)
   Ans:

3. The **SQL statements** and **output results** of 4a (10pts). If the SQL statements or output results are not provided, you will not get the points. Note: if the result table contains too many rows to be included in the document, you can upload the result table to any cloud service, and put the link of the whole results table with particle content with the total number of rows in your result table in the document. 如果資料太長，可以只放部分結果以及總資料筆數，然後附上一個連結連到完整的results檔案。(update: 2023/10/15)
   Ans:

4. The SQL statements and output results of 4b (10pts)
   Ans:

5. The SQL statements and output results of 4c (10pts)
   Ans:

6. The SQL statements and output results of 4d (10pts)
   Ans:

7. The SQL statements and output results of 4e (10pts)

Ans:

8. The SQL statements and output results of 4f (10pts)
   Ans:

HW1 Q&A:

Q: Do we have to use pgAdmin to access PostgreSQL?
A:
No. You do not need to use pgAdmin to access PostgreSQL. You can use other tools to access your database with PostgreSQL.

Q: I want to create an attribute with a double data type, but I get an error message saying: " type double doesn't exist". Why?
A:
You can check the official manual for PostgreSQL. Maybe double precision is the data type you need in the PostgreSQL environment?

Q: I get an error message saying that the "Big5" encoding is not supported in the PostgreSQL environment. What can I do?
A:
Maybe you can try to convert the encoding, from "Big5" to "utf8" by using the other software. For example, VS code, Notepad++, etc. After encoding conversion, you may try to import the file again.

Q: I noticed there are rows in the 'num_parts' column with a value of -1. Should we include these in our calculations? (update: 2023/10/20)
A:
Yes, for the time being, please **INCLUDE** the rows with -1 in the 'num_parts' column for your calculations. We currently do not have a definitive understanding of what this value represents. Additionally, if you have insights or findings about what -1 might represent in 'sets.csv', please include them in your homework submission.

	set_num	name	year	theme_id	num_parts
11645	Vancouver-1	LEGO Store Grand Opening Exclusive Set, Oakrid...	2012	408	-1
1683	240-1	Wooden Storage Box Large, Empty	1967	383	-1
6545	66392-1	Duplo Cars Super Pack 3 in 1 (5816, 5817, 5818)	2012	506	-1
1837	2852726-1	Gyroscopic Sensor for Mindstorms NXT	2011	259	0
11144	DKIDEASBOOK-1	The LEGO Ideas Book	2011	501	0
10152	8805-0	Minifigure Series 5 [Random Bag]	2011	540	0