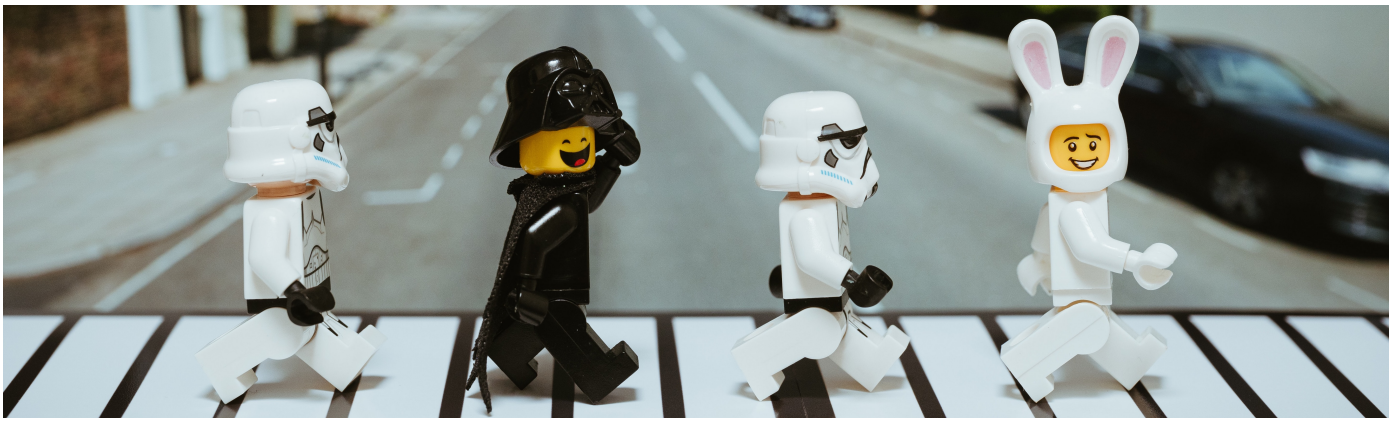
Introduction



Let's look at Lego sets!

Lego is a household name across the world, supported by a diverse toy line, hit movies, and a series of successful video games. In this project, we are going to explore a key development in the history of Lego: the introduction of licensed sets such as Star Wars, Super Heroes, and Harry Potter.

It may not be widely known, but Lego has had its share of ups and downs since its inception in the early 20th century. This includes a particularly rough period in the late 90s. As described in [this article (https://www.businessinsider.com/how-lego-made-a-huge-turnaround-2014-2?r=US&IR=T)](), Lego was only able to survive due to a successful internal brand (Bionicle) and the introduction of its first licensed series: Star Wars. In the instructions panel are the two questions you will need to answer to complete this project.

Before diving into our analysis though, let's become familiar with the two datasets that will help you with this project:

datasets/lego_sets.csv

- **set_num:** A code that is unique to each set in the dataset. ***This column is critical, and a missing value indicates the set is a duplicate or invalid!***
- **set_name:** A name for every set in the dataset (note that this can be the same for different sets).
- **year:** The date the set was released.
- **num_parts:** The number of parts contained in the set. ***This column is not central to our analyses, so missing values are acceptable.***
- **theme_name:** The name of the sub-theme of the set.
- **parent_theme:** The name of the parent theme the set belongs to. Matches the `name` column of the `parent_themes` csv file.

datasets/parent_themes.csv

- **id:** A code that is unique to every theme.
- **name:** The name of the parent theme.
- **is_licensed:** A Boolean column specifying whether the theme is a licensed theme.

From here on out, it will be your task to explore and manipulate the existing data until you are able to answer the two questions described in the instructions panel. Feel free to add as many cells as necessary. Finally, remember that you are only tested on your answer, not on the methods you use to arrive at the answer!

Note: *If you haven't completed a DataCamp project before you should check out the Intro to Projects (https://projects.datacamp.com/projects/33) first to learn about the interface. In this project, you also need to know your way around* `pandas` *DataFrames and it's recommended that you take a look at the course Data Manipulation with pandas (https://www.datacamp.com/courses/data-manipulation-with-pandas).*

Task 1: What percentage of all licensed sets ever released were Star Wars Themed?

Task 2: In which year was Star Wars not the most popular licensed theme?

Task 3: How many unique sets were released each year (1955-2017)?

In [129]:

```
# Use this cell to begin your analyses, and add as many cells as you would like!

#importing pandas library
import pandas as pd
```

In [130]:

```
#make a dataframe of two csv files.
df = pd.read_csv('./lego_sets.csv')
p_theme = pd.read_csv('./parent_themes.csv')
```

In [131]:

```
#head for viewing top 5 rows
df.head()
```

Out[131]:

	set_num	name	year	num_parts	theme_name	parent_theme
0	00-1	Weetabix Castle	1970	471.0	Castle	Legoland
1	0011-2	Town Mini-Figures	1978	NaN	Supplemental	Town
2	0011-3	Castle 2 for 1 Bonus Offer	1987	NaN	Lion Knights	Castle
3	0012-1	Space Mini-Figures	1979	12.0	Supplemental	Space
4	0013-1	Space Mini-Figures	1979	12.0	Supplemental	Space

In [106]:

```
1 p_theme.head()
```

Out[106]:

	id	name	is_licensed
0	1	Technic	False
1	22	Creator	False
2	50	Town	False
3	112	Racers	False
4	126	Space	False

In [132]:

```
1 #check the shape of dataset
2 print(df.shape)
3 print(p_theme.shape)
```

```
(11986, 6)
(111, 3)
```

In [133]:

```
1 #making a new dataframe and join that two csv files according to their columns
2 main_df = df.merge(p_theme,left_on = 'parent_theme',right_on='name')
3 main_df.head()
```

Out[133]:

	set_num	name_x	year	num_parts	theme_name	parent_theme	id	name_y	is_licens
0	00-1	Weetabix Castle	1970	471.0	Castle	Legoland	411	Legoland	Fa
1	00-2	Weetabix Promotional House 1	1976	NaN	Building	Legoland	411	Legoland	Fa
2	00-3	Weetabix Promotional House 2	1976	NaN	Building	Legoland	411	Legoland	Fa
3	00-4	Weetabix Promotional Windmill	1976	126.0	Building	Legoland	411	Legoland	Fa
4	00-7	Weetabix Promotional Lego Village	1976	NaN	Building	Legoland	411	Legoland	Fa

In [134]:

```
#shape of the main dataframe
main_df.shape
```

Out[134]:

```
(11986, 9)
```

In [135]:

```
#dropping the column "name_y" because of same datapoints
main_df= main_df.drop(columns='name_y')
main_df.head()
```

Out[135]:

	set_num	name_x	year	num_parts	theme_name	parent_theme	id	is_licensed
0	00-1	Weetabix Castle	1970	471.0	Castle	Legoland	411	False
1	00-2	Weetabix Promotional House 1	1976	NaN	Building	Legoland	411	False
2	00-3	Weetabix Promotional House 2	1976	NaN	Building	Legoland	411	False
3	00-4	Weetabix Promotional Windmill	1976	126.0	Building	Legoland	411	False
4	00-7	Weetabix Promotional Lego Village	1976	NaN	Building	Legoland	411	False

In [136]:

```
#now we'll make a licensed dataframe for only licensed themes.
licensed = main_df[main_df.is_licensed == True]
licensed.head()
# licensed.shape
```

Out[136]:

	set_num	name_x	year	num_parts	theme_name	parent_theme	id	is_licensed
3493	10018-1	Darth Maul	2001	1868.0	Star Wars	Star Wars	158	True
3494	10019-1	Rebel Blockade Runner - UCS	2001	NaN	Star Wars Episode 4/5/6	Star Wars	158	True
3495	10026-1	Naboo Starfighter - UCS	2002	NaN	Star Wars Episode 1	Star Wars	158	True
3496	10030-1	Imperial Star Destroyer - UCS	2002	3115.0	Star Wars Episode 4/5/6	Star Wars	158	True
3497	10123-1	Cloud City	2003	707.0	Star Wars Episode 4/5/6	Star Wars	158	True

In [137]:

```
#check the star wars theme in "parent_theme" column
licensed[licensed.parent_theme=='Star Wars']
```

Out[137]:

	set_num	name_x	year	num_parts	theme_name	parent_theme	id	is_licensed
3493	10018-1	Darth Maul	2001	1868.0	Star Wars	Star Wars	158	True
3494	10019-1	Rebel Blockade Runner - UCS	2001	NaN	Star Wars Episode 4/5/6	Star Wars	158	True
3495	10026-1	Naboo Starfighter - UCS	2002	NaN	Star Wars Episode 1	Star Wars	158	True
3496	10030-1	Imperial Star Destroyer - UCS	2002	3115.0	Star Wars Episode 4/5/6	Star Wars	158	True
3497	10123-1	Cloud City	2003	707.0	Star Wars Episode 4/5/6	Star Wars	158	True
...	...	...	...	...	...	...	...	...
4097	VP-12	Star Wars Co-Pack of 7121 and 7151	2000	2.0	Star Wars Episode 1	Star Wars	158	True
4098	VP-2	Star Wars Co-Pack of 7110 and 7144	2001	2.0	Star Wars Episode 4/5/6	Star Wars	158	True
4099	VP-3	Star Wars Co-Pack of 7131 and 7151	2000	2.0	Star Wars Episode 1	Star Wars	158	True
4100	VP-4	Star Wars Co-Pack of 7101 7111 and 7171	2000	3.0	Star Wars Episode 1	Star Wars	158	True
4101	VP-8	Star Wars Co-Pack of 7130 and 7150	2000	NaN	Star Wars Episode 4/5/6	Star Wars	158	True

609 rows × 8 columns

In [138]:

```
#check the null values in licensed dataframe
licensed.isnull().sum()
```

Out[138]:

```
set_num         153
name_x          153
year              0
num_parts       577
theme_name      153
parent_theme      0
id                0
is_licensed       0
dtype: int64
```

In [139]:

```
#dropping the null values from licensed dataframe considering the "set_num"
licensed = licensed.dropna(subset=['set_num'])
```

In [140]:

```
#recheck the null values
licensed.isnull().sum()
```

Out[140]:

```
set_num           0
name_x            0
year              0
num_parts       515
theme_name        0
parent_theme      0
id                0
is_licensed       0
dtype: int64
```

In [141]:

```
#make a new dataframe for starwars that contains the only starwars theme from "parent_t
star_wars = licensed[licensed['parent_theme']=='Star Wars']
```

In [119]:

```
#finding the percentage of the licensed sets ever released for starwars them.
the_force = (star_wars.shape[0]/licensed.shape[0]*100)
```

Ans: task:1

In [146]:

```
#value in percentage:
print('Star Wars licensed sets ever released is {} %'.format(the_force))
```

```
Star Wars licensed sets ever released is 51.653944020356235 %
```

so , there's 52% of all licensed sets ever released were star wars themed according to our data exploration part

In [147]:

```
#second task
```

In which year was Star Wars not the most popular licensed theme?

In [121]:

```
licensed.isnull().sum()
```

Out[121]:

```
set_num           0
name_x            0
year              0
num_parts       515
theme_name        0
parent_theme      0
id                0
is_licensed       0
dtype: int64
```

In [149]:

```
#making new dataframe for sorted values of year
t_df_01 = licensed.sort_values('year')
t_df_01.head()
```

Out[149]:

	set_num	name_x	year	num_parts	theme_name	parent_theme	id	is_licensed
3702	7161-1	Gungan Sub	1999	379.0	Star Wars Episode 1	Star Wars	158	True
3705	7171-1	Mos Espa Podrace	1999	NaN	Star Wars Episode 1	Star Wars	158	True
3690	7140-1	X-wing Fighter	1999	271.0	Star Wars Episode 4/5/6	Star Wars	158	True
3685	7130-1	Snowspeeder	1999	NaN	Star Wars Episode 4/5/6	Star Wars	158	True
3684	7128-1	Speeder Bikes	1999	93.0	Star Wars Episode 4/5/6	Star Wars	158	True

In [152]:

```
#groupby the "year" and "parent_theme" columns and aggregate sum function and reset the
t_df_02 = task_2.groupby(['year','parent_theme']).sum().reset_index()
t_df_02.head()
```

Out[152]:

	year	parent_theme	num_parts	id	is_licensed
0	1999	Star Wars	1384.0	2054	13
1	2000	Disney's Mickey Mouse	405.0	1940	5
2	2000	Star Wars	2580.0	4108	26
3	2001	Harry Potter	1284.0	2706	11
4	2001	Star Wars	2949.0	2212	14

In [153]:

```
#last dataframe for that sorting values according to "is_licensed" column and dropping
t_df_03 = t_df_02.sort_values('is_licensed',ascending=False).drop_duplicates(['year'])
t_df_03.head()
```

Out[153]:

	year	parent_theme	num_parts	id	is_licensed
82	2017	Super Heroes	13123.0	34704	72
76	2016	Star Wars	6934.0	9638	61
67	2015	Star Wars	11410.0	9164	58
59	2014	Star Wars	8293.0	7110	45
47	2012	Star Wars	6769.0	6794	43

Ans :: Task :: 2

so , from "t_df_03" dataframe we can get the result.

in 2017 star wars wasnt the best theme.

In [154]:

```
# Task :: 3
```

How many unique sets were released each year (1955-2017)?

In [125]:

```
licensed.head()
```

Out[125]:

	set_num	name_x	year	num_parts	theme_name	parent_theme	id	is_licensed
3493	10018-1	Darth Maul	2001	1868.0	Star Wars	Star Wars	158	True
3494	10019-1	Rebel Blockade Runner - UCS	2001	NaN	Star Wars Episode 4/5/6	Star Wars	158	True
3495	10026-1	Naboo Starfighter - UCS	2002	NaN	Star Wars Episode 1	Star Wars	158	True
3496	10030-1	Imperial Star Destroyer - UCS	2002	3115.0	Star Wars Episode 4/5/6	Star Wars	158	True
3497	10123-1	Cloud City	2003	707.0	Star Wars Episode 4/5/6	Star Wars	158	True

In [157]:

```
1 main_df[~main_df['set_num'].isnull()]
```

Out[157]:

	set_num	name_x	year	num_parts	theme_name	parent_theme	id	is_licensed
0	00-1	Weetabix Castle	1970	471.0	Castle	Legoland	411	False
1	00-2	Weetabix Promotional House 1	1976	NaN	Building	Legoland	411	False
2	00-3	Weetabix Promotional House 2	1976	NaN	Building	Legoland	411	False
3	00-4	Weetabix Promotional Windmill	1976	126.0	Building	Legoland	411	False
4	00-7	Weetabix Promotional Lego Village	1976	NaN	Building	Legoland	411	False
...	...	...	...	...	...	...	...	...
11981	8410-1	Swampfire	2010	22.0	Ben 10	Ben 10	270	True
11982	8411-1	ChromaStone	2010	21.0	Ben 10	Ben 10	270	True
11983	8517-1	Humungousaur	2010	14.0	Ben 10	Ben 10	270	True
11984	8518-1	Jet Ray	2010	NaN	Ben 10	Ben 10	270	True
11985	8519-1	Big Chill	2010	20.0	Ben 10	Ben 10	270	True

11833 rows × 8 columns

In [158]:

```
# making a new dataframe for "set_num" that is not containing the null values
# and also make "count" column with 1 value
c_df = main_df[~main_df['set_num'].isnull()]
c_df['count'] =1
c_df.head()
```

```
C:\Users\KEYUR PRAJAPATI\AppData\Local\Temp\ipykernel_15272\3882752604.py:4:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/
stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pand
as.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-v
ersus-a-copy)
  c_df['count'] =1
```

Out[158]:

	set_num	name_x	year	num_parts	theme_name	parent_theme	id	is_licensed	count
0	00-1	Weetabix Castle	1970	471.0	Castle	Legoland	411	False	1
1	00-2	Weetabix Promotional House 1	1976	NaN	Building	Legoland	411	False	1
2	00-3	Weetabix Promotional House 2	1976	NaN	Building	Legoland	411	False	1
3	00-4	Weetabix Promotional Windmill	1976	126.0	Building	Legoland	411	False	1
4	00-7	Weetabix Promotional Lego Village	1976	NaN	Building	Legoland	411	False	1

In [161]:

```
# making this dataframe and groupby the year and using sum function we get the year of
sets_per_year = c_df.groupby(['year']).sum().reset_index()[['year','count']]
sets_per_year.head()
```

Out[161]:

	year	count
0	1950	7
1	1953	4
2	1954	14
3	1955	28
4	1956	12

In [178]:

```
# through this loop, we can get the "year" and "sets" that released.
for index, row in sets_per_year.iterrows():
    print(row['year'],row['count'])
```

```
1950 7
1953 4
1954 14
1955 28
1956 12
1957 21
1958 42
1959 4
1960 3
1961 17
1962 40
1963 18
1964 11
1965 10
1966 89
1967 21
1968 25
1969 69
1970 29
1971 45
1972 38
1973 68
1974 39
1975 31
1976 68
1977 92
1978 73
1979 82
1980 88
1981 79
1982 76
1983 57
1984 76
1985 139
1986 123
1987 209
1988 68
1989 114
1990 85
1991 106
1992 115
1993 111
1994 128
1995 128
1996 144
1997 194
1998 325
1999 300
2000 327
2001 339
2002 447
2003 415
2004 371
2005 330
```

```
2006 283
2007 319
2008 349
2009 403
2010 444
2011 502
2012 615
2013 593
2014 715
2015 670
2016 608
2017 438
```

In [188]:

```
1 sets_per_year[sets_per_year['count']==sets_per_year['count'].max()]
```

Out[188]:

	year	count
62	2014	715

so we have the 2014 year that released 715 sets.