ANALISIS DE DATOS LEGO®

- Karen Raczkowski
- Juan Ignacio Ribet
- Fabian Sarmiento

inventories

id	int4
version	int4
set_num	varchar(20)

inventory_parts

inventory_id	int4
part_num	varchar(20)
color_id	int4
quantity	int4
is_spare	bool

inventory_minifigs

inventory_id	int4
fig_num	varchar(20)
quantity	int4

inventory_sets

inventory_id	int4
set_num	varchar(20)
quantity	int4

minifigs

fig_num	varchar(20)
name	varchar(256)
num_parts	int4

sets

set_num	varchar(20)
name	varchar(256)
year	int4
theme_id	int4
num_parts	int4

part_categories

id	int4
name	varchar(200)

parts

part_num	varchar(20)
name	varchar(250)
part_cat_id	int4

colors

id	int4
name	varchar(200)
rgb	varchar(6)
is_trans	bool

part_relationships

rel_type	varchar(1)
child_part_num	varchar(20)
parent_part_num	varchar(20)

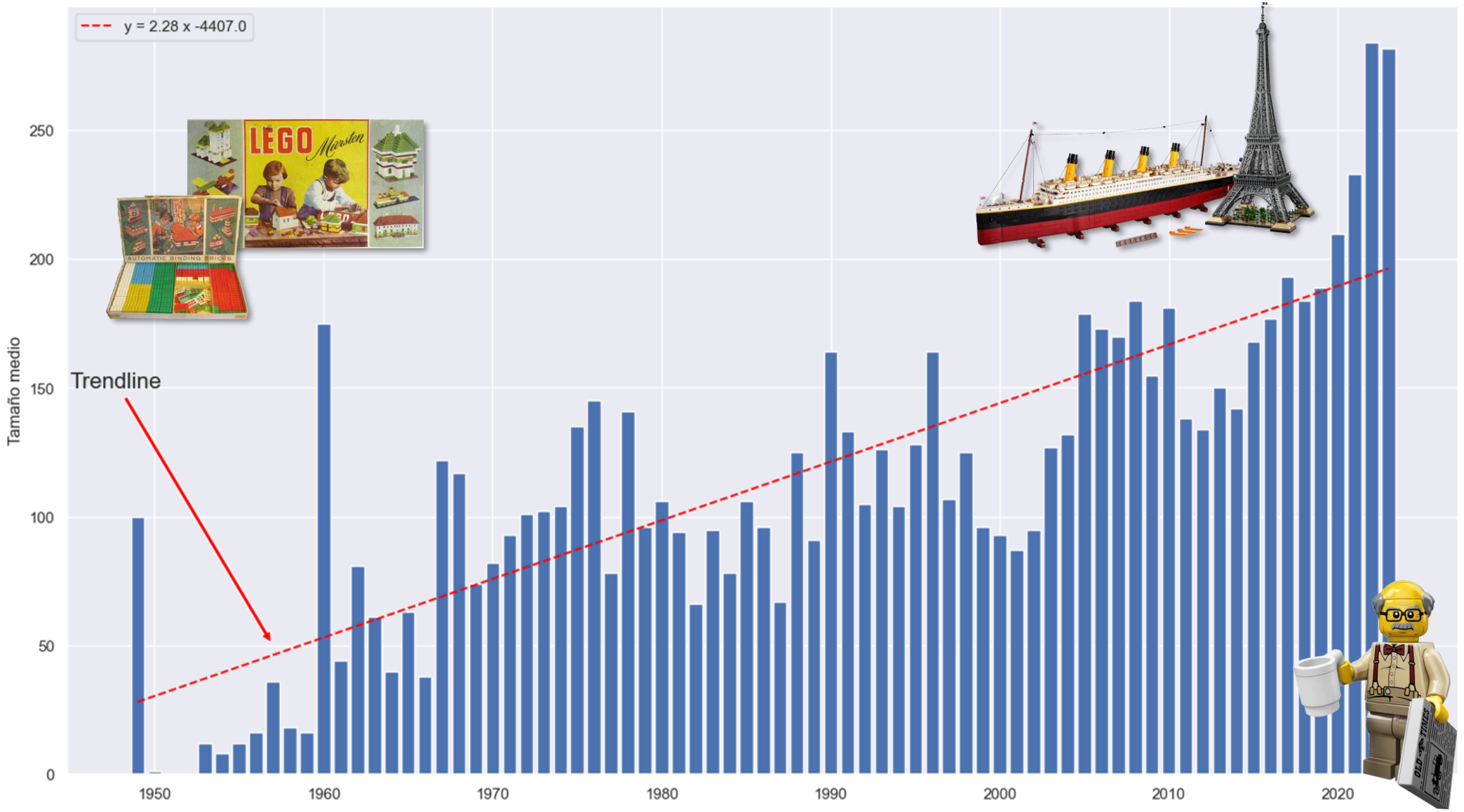
elements

element_id	varchar(10)
part_num	varchar(20)
color_id	int4

themes

id	int4
name	varchar(40)
parent_id	int4

¿Cómo evolucionaron los sets de lego en tamaño a través de los años?



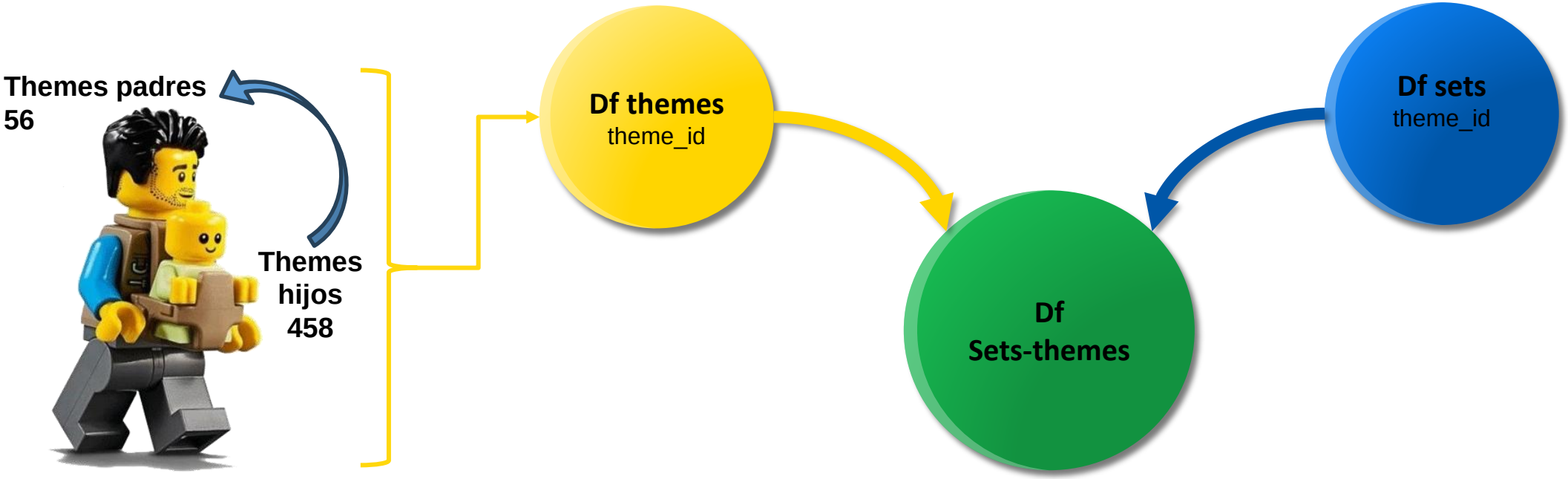
¿Podría predecir a qué temática pertenece un *set* basado en el contenido de este?

- Preparación del dataset completo

- Análisis exploratorio

- Ingeniería de features

Preparación del dataset completo



	set_num	name	year	theme_id	num_parts	img_url	parent_id	themes_parent_nam
0	001-1	Gears	1965	1	43	https://cdn.rebrickable.com/media/sets/001-1.jpg	0	Technic
1	002-1	4.5V Samsonite Gears Motor Set	1965	1	3	https://cdn.rebrickable.com/media/sets/002-1.jpg	0	Technic
2	1030-1	TECHNIC I: Simple Machines Set	1985	1	210	https://cdn.rebrickable.com/media/sets/1030-1.jpg	0	Technic
3	1038-1	ERBIE the Robo-Car	1985	1	120	https://cdn.rebrickable.com/media/sets/1038-1.jpg	0	Technic
4	1039-1	Manual Control Set 1	1986	1	39	https://cdn.rebrickable.com/media/sets/1039-1.jpg	0	Technic
...	...	...	...	...	...	...	...	...
21292	M20-2566-10	Modulex Box - 10x Window 1 x 6 x 6	1963	716	10	https://cdn.rebrickable.com/media/sets/m20-256...	0	Modulex
21293	M20-2575-10	Modulex Box - 10x Window 1 x 7 x 5	1963	716	10	https://cdn.rebrickable.com/media/sets/m20-257...	0	Modulex
21294	M20-2576-10	Modulex Box - 10x Window 1 x 7 x 6	1963	716	10	https://cdn.rebrickable.com/media/sets/m20-257...	0	Modulex
21295	M20-2586-10	Modulex Box - 10x Window 1 x 8 x 6	1963	716	10	https://cdn.rebrickable.com/media/sets/m20-258...	0	Modulex
21296	WEETABIX1-1	Weetabix Castle	1970	414	471	https://cdn.rebrickable.com/media/sets/weetabi...	411	Legoland

97 rows × 8 columns

Preparación del dataset completo

‘sets-themes’

Name
Year
theme_id

'set_
num'

Datos de interés para predecir a qué tema pertenece un set

num_partr
img_url
parent_id

‘themes
_parent
_name’

Por cada valor único de: 'id + version + set_num', se indica a que temática padre pertenece

‘inventories’

	id	version	set_num	themes_parent_name
0	1	1	7922-1	Sports
1	3	1	3931-1	Friends
2	4	1	6942-1	Space
3	15	1	5158-1	Service Packs
4	16	1	903-1	System

36299 rows × 4 columns

Preparación del dataset completo

Inventories_inventory_parts

id	version	set_num	themes_parent_name	part_num	color_id	quantity_part	is_spare	rgb

inventories — id — **inventory_parts**

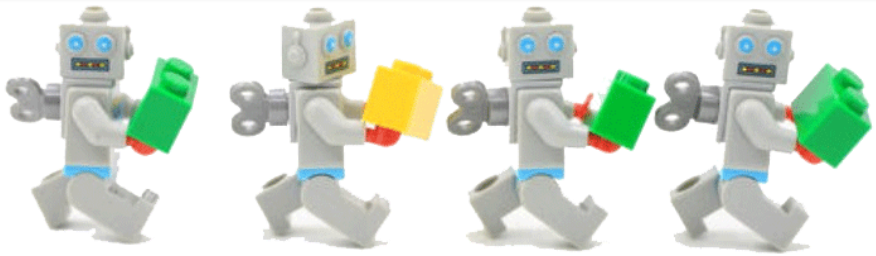
Eliminamos la columna ‘img_url’

Eliminamos NaN: minifigs

1083524 rows × 8 columns

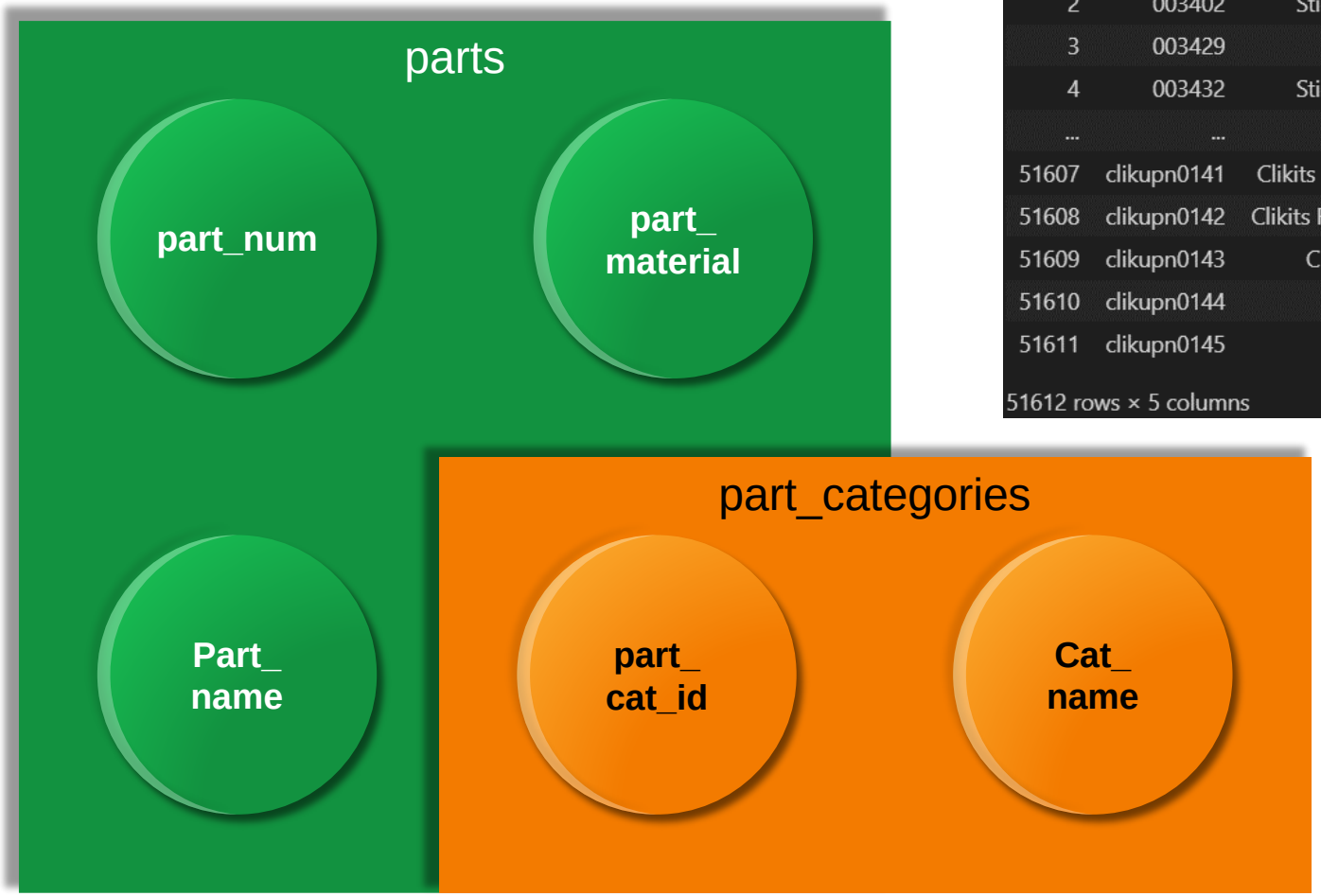
colors

color_id — Rgb



Preparación del dataset completo

Df ‘parts_cat’
Obtenemos un dataset que contiene información sobre el material y la categoría de cada tipo de pieza.



	part_num	name	part_cat_id	part_material	cat_name
0	003381	Sticker Sheet for Set 663-1	58	Plastic	Stickers
1	003383	Sticker Sheet for Sets 618-1, 628-2	58	Plastic	Stickers
2	003402	Sticker Sheet for Sets 310-3, 311-1, 312-3	58	Plastic	Stickers
3	003429	Sticker Sheet for Set 1550-1	58	Plastic	Stickers
4	003432	Sticker Sheet for Sets 357-1, 355-1, 940-1	58	Plastic	Stickers
...	...	...	...	...	...
51607	clikupn0141	Clikits Connector, Ring 10 x 10 x 1 for Pencil...	48	Plastic	Clikits
51608	clikupn0142	Clikits Rectangle 30 x 12 with 2 Slots & Tabs ...	48	Plastic	Clikits
51609	clikupn0143	Clikits Circle 10 x 10 (Pencil Holder Base)	48	Plastic	Clikits
51610	clikupn0144	Clikits Container Cube Drawer	48	Plastic	Clikits
51611	clikupn0145	Clikits Container Cube Drawer Unit	48	Plastic	Clikits

51612 rows × 5 columns

Preparación del dataset completo

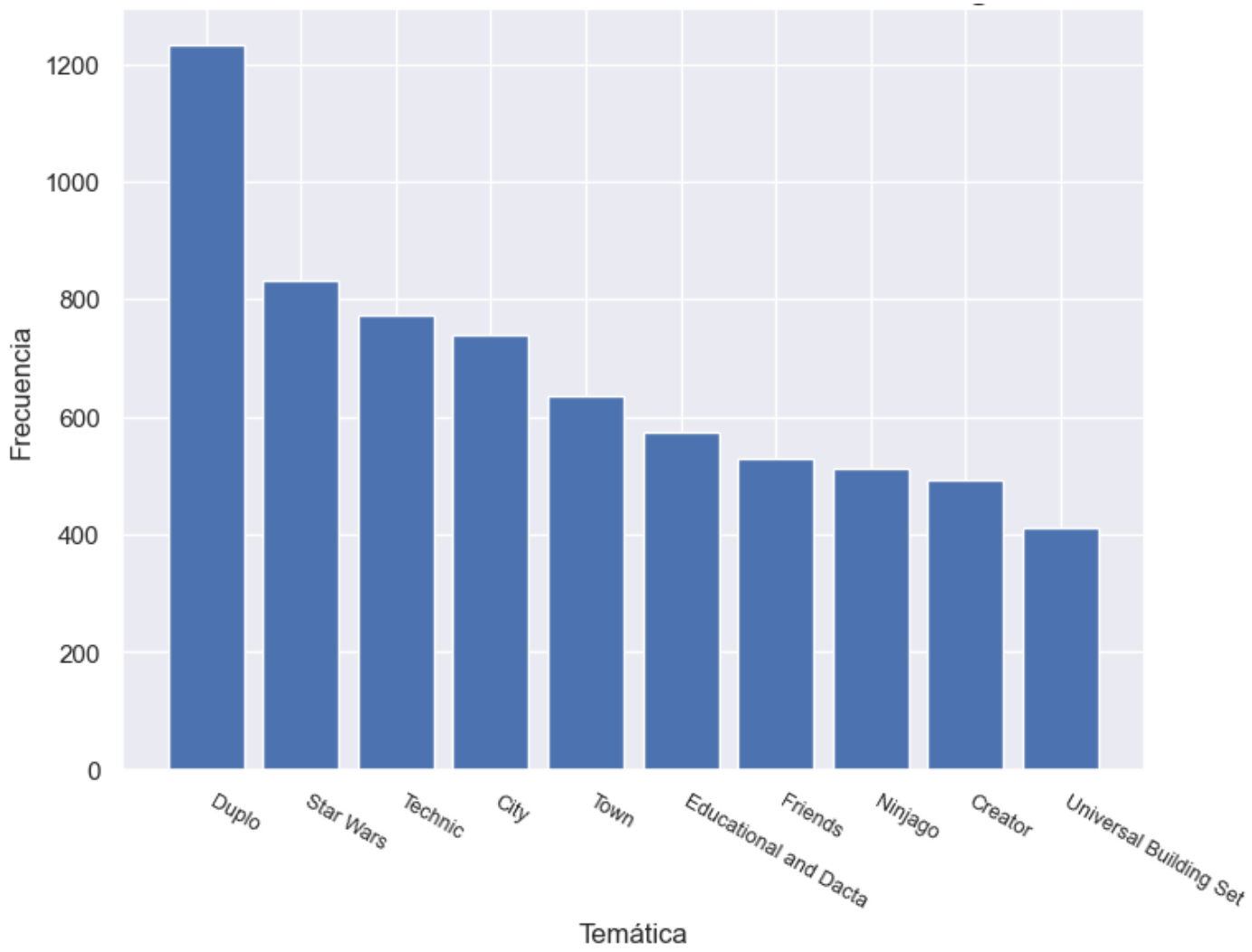
Dataset completo

El dataset completo contiene información sobre el contenido de los sets (colores, materiales, categorías de las piezas).

Cada observación pertenece a un tipo de pieza, la variable ‘quantity_part’ indica cuántas de esas piezas hay en cada set.

Preparación del dataset completo

- Verificamos las temáticas que más se repiten en los sets de Lego para así trabajar sólo con ellas y reducir la cantidad de clases.



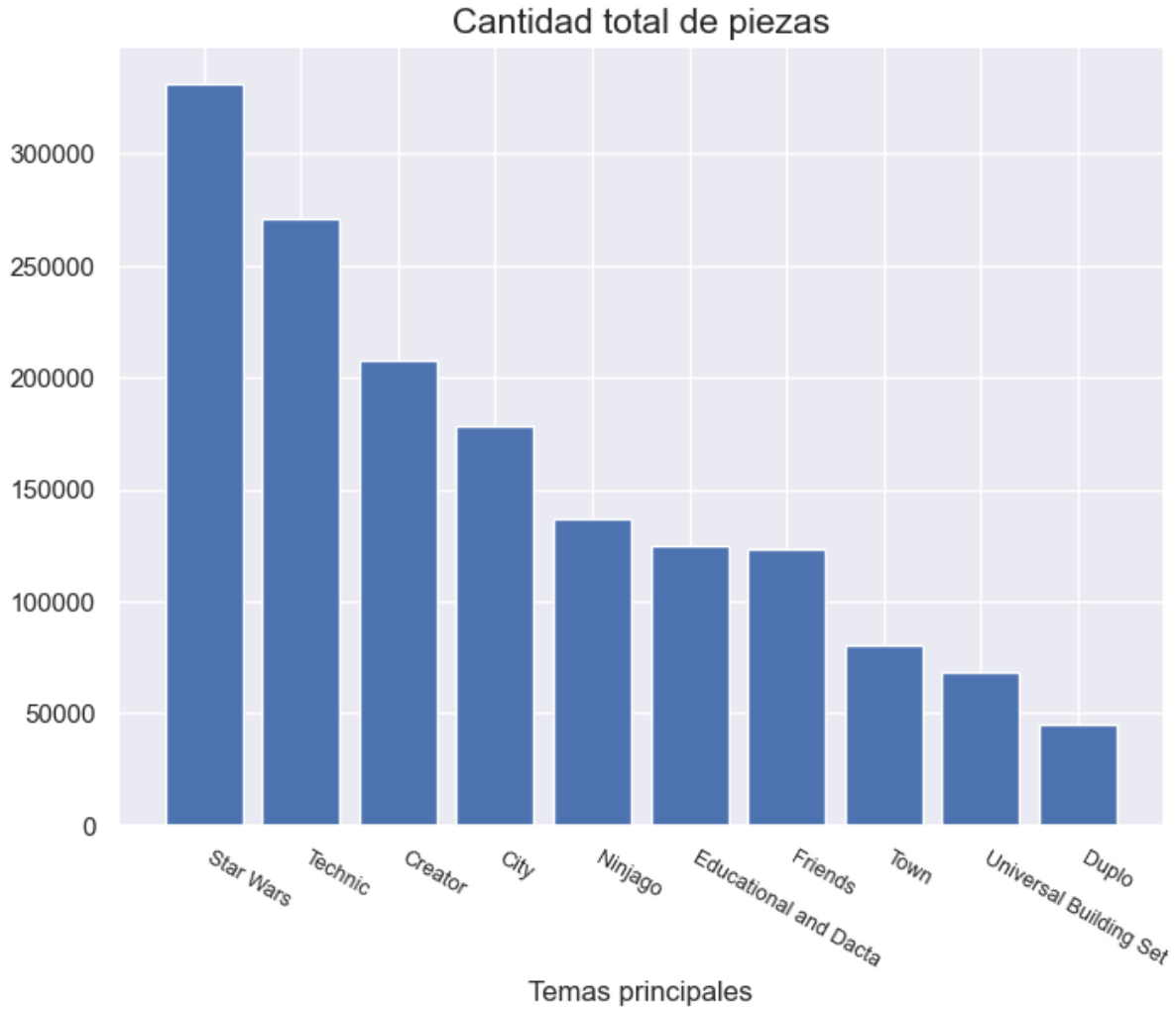
Análisis exploratorio

- Variables categóricas nominales:
  - set_num: número de identificación único para cada set
  - themes_parent_name: nombre de la temática padre
  - rgb: valor RGB aproximado (color) para cada tipo de pieza
  - part_material: el material de cada tipo de pieza
  - cat_name: categoría de cada tipo de pieza

- Variable numérica:
  - quantity_part: cantidad de piezas de cada tipo en cada set

set_num	themes_parent_name	quantity_part	rgb	part_material	cat_name
3931-1	Friends	1	#FCFCFC	Plastic	Minifig Accessories
41135-1	Friends	2	#FCFCFC	Plastic	Minifig Accessories
1821-1	Town	4	#9BA19D	Plastic	Minifig Accessories
1821-1	Town	1	#F2CD37	Plastic	Minifig Accessories
3040-1	Universal Building Set	1	#FCFCFC	Plastic	Minifig Accessories

Análisis exploratorio



	min	max	mean	median
City	1	1,949	241.14	151
Creator	1	5,962	421.32	200
Duplo	1	387	36.35	25
Educational and Dacta	1	4,900	218.02	83
Friends	1	2,074	232.76	95
Ninjago	1	6,182	267.74	62
Star Wars	1	7,663	397.96	165
Technic	1	4,134	350.42	121
Town	2	2,059	125.67	66
Universal Building Set	1	1,201	166.97	52

Análisis exploratorio

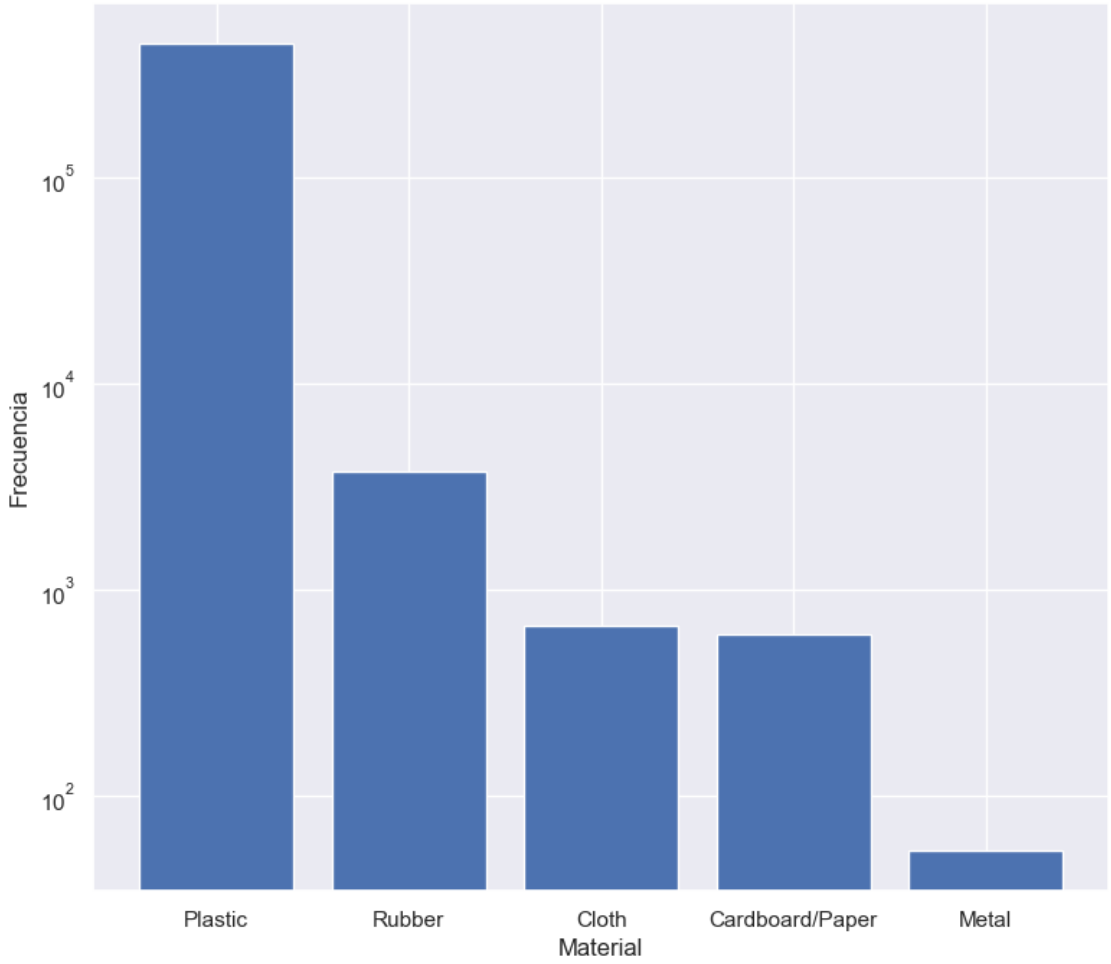
- El dataset contiene **5 materiales** distintos y **129 colores** distintos.

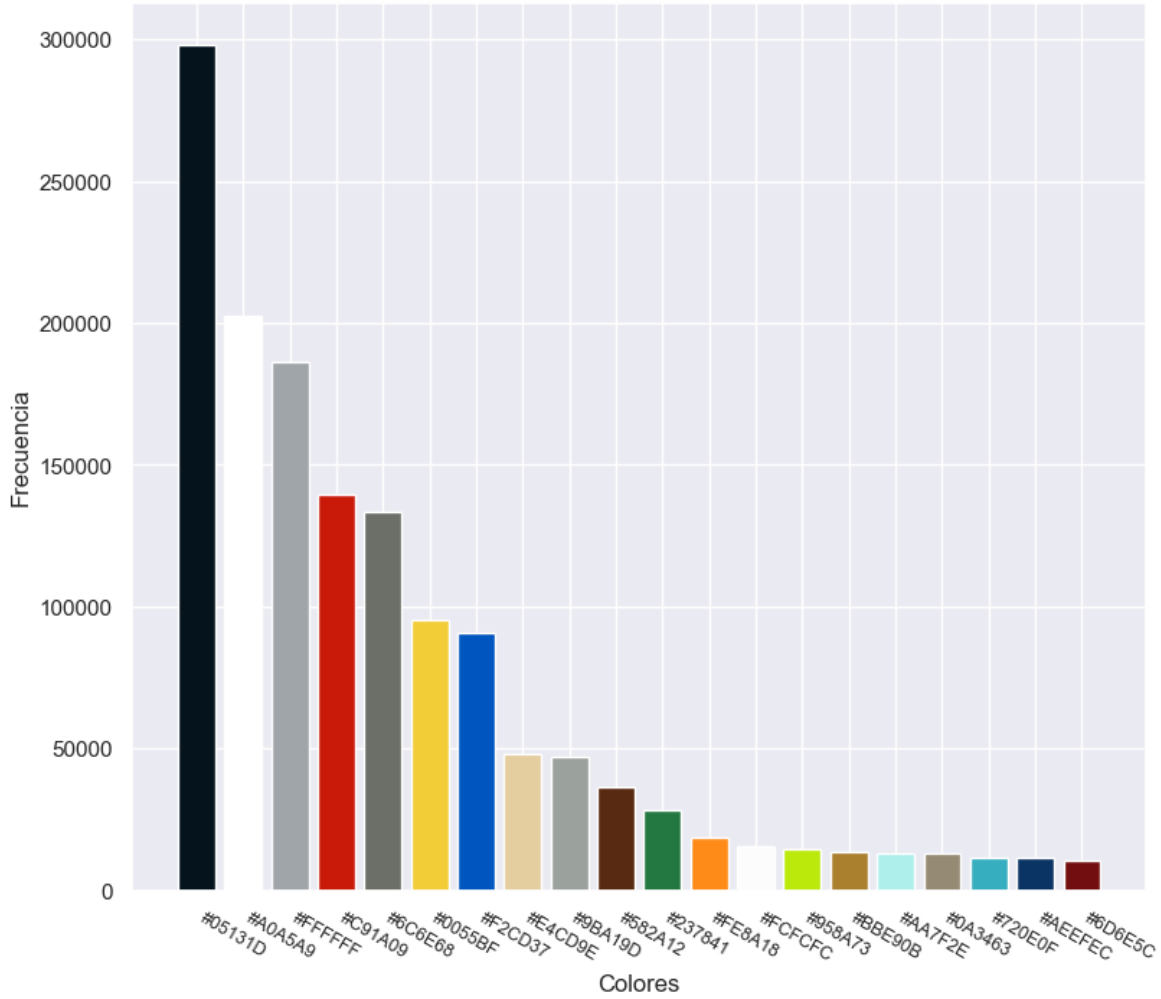
Material	Cantidad de tipos de piezas	Cantidad de piezas
Plastic	449,687	1,548,932
Rubber	3,728	15,796
Cloth	677	856
Cardboard/Paper	612	621
Metal	55	201

RGB	Cantidad de tipos de piezas	Cantidad de piezas
#05131D	78,785	298,123
#A0A5A9	54,980	202,726
#FFFFFF	56,820	186,177
#C91A09	41,746	139,590
#6C6E68	38,217	133,440
#0055BF	20,247	95,518
#F2CD37	28,105	90,864
#E4CD9E	12,293	48,047
#9BA19D	11,994	47,166
#582A12	11,345	36,403

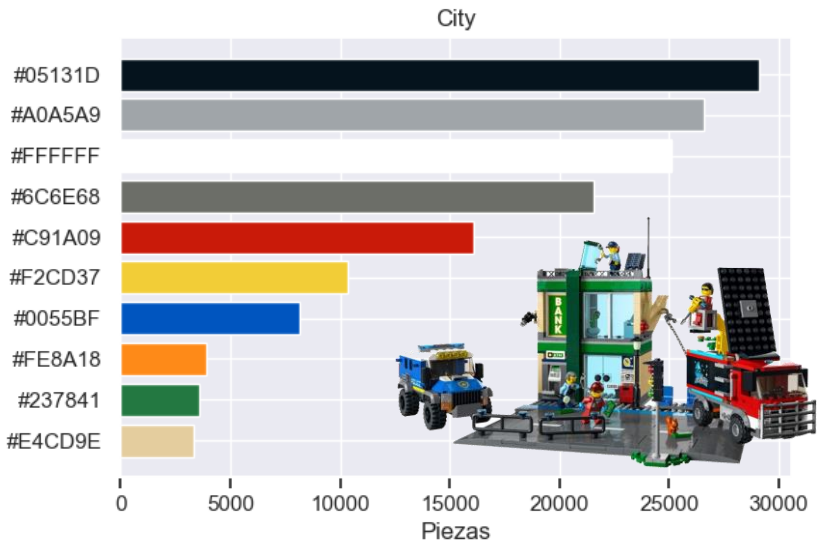
Análisis exploratorio

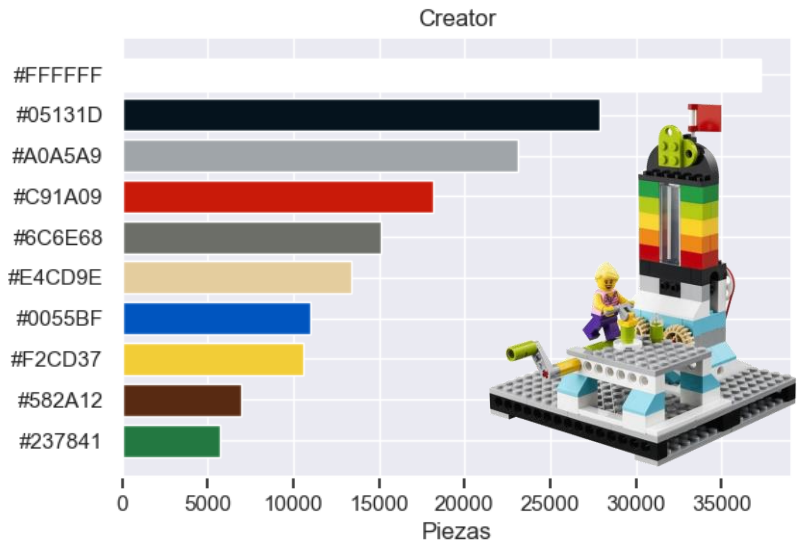
- Cantidad de piezas de cada material y color (20 más comunes):

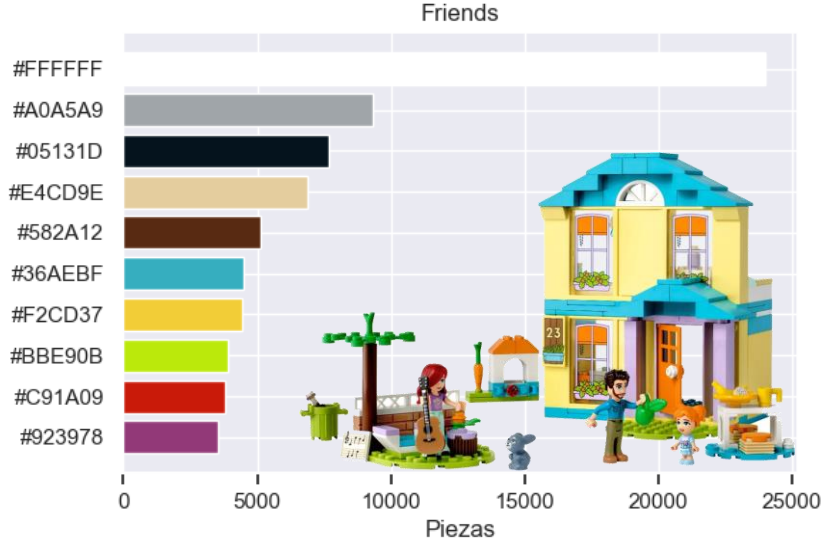


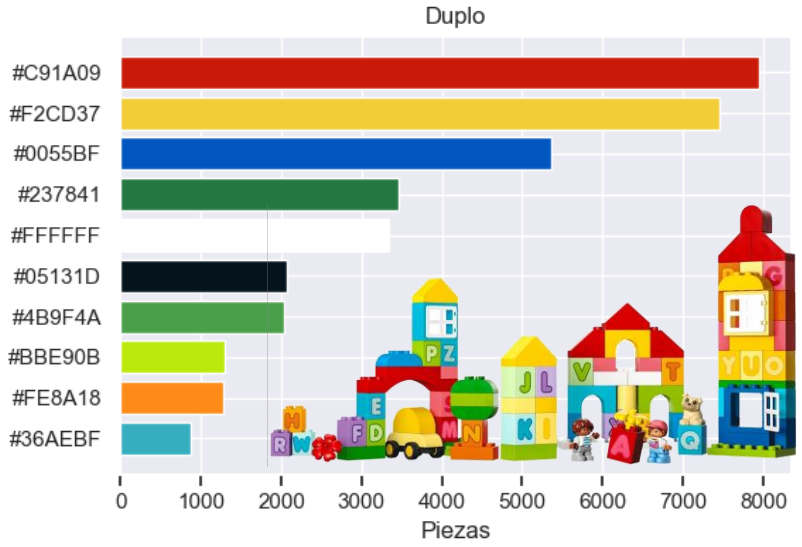


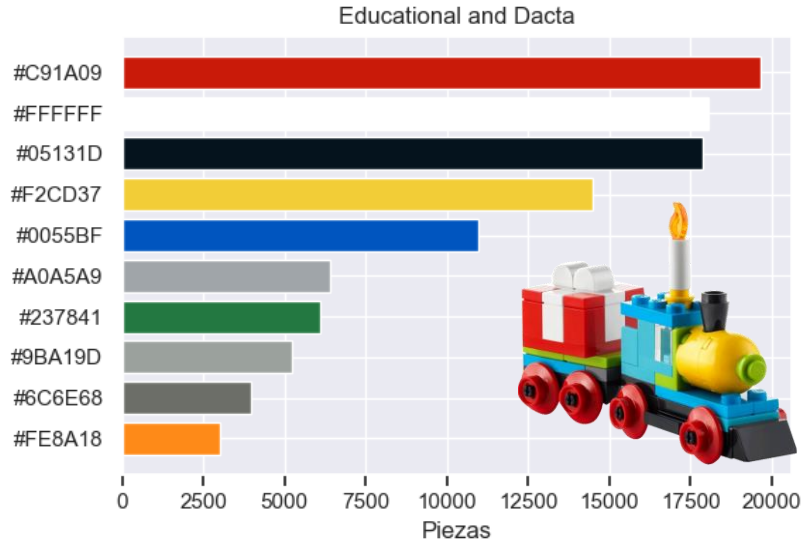
Análisis exploratorio

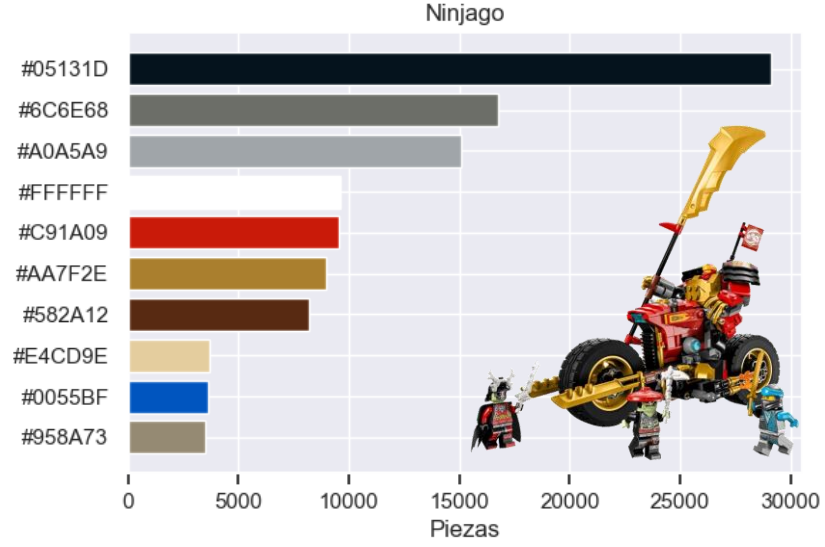




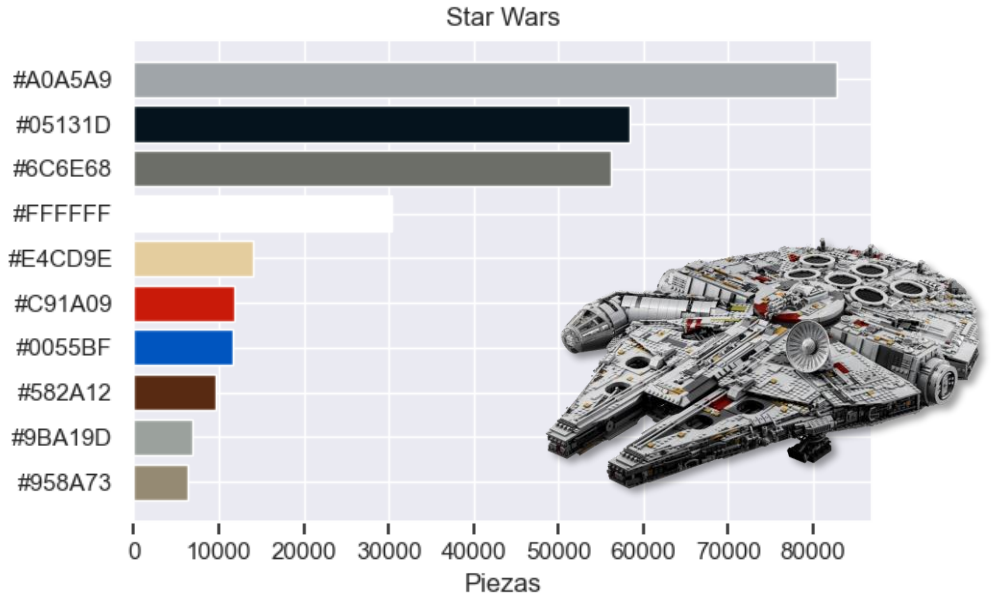


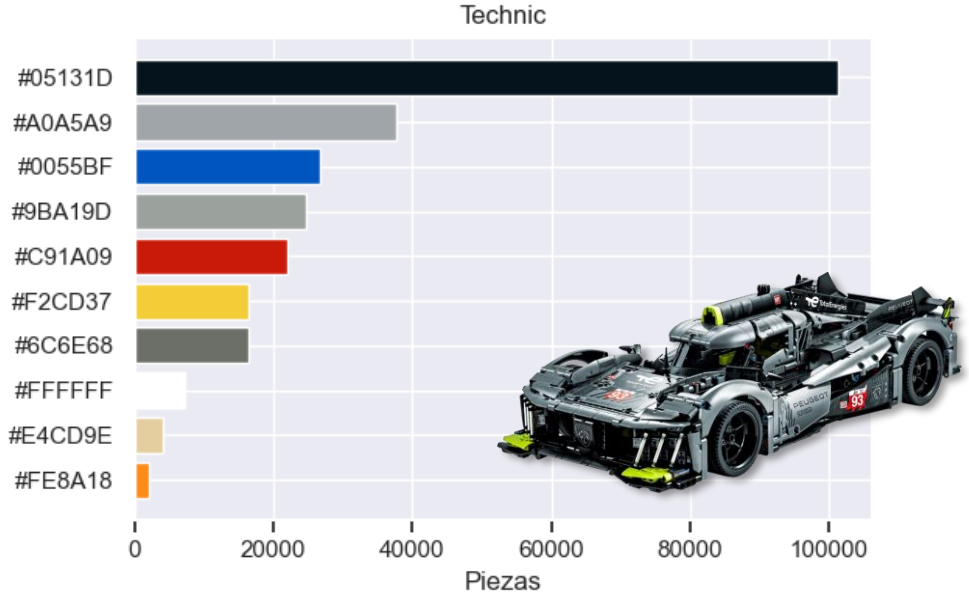


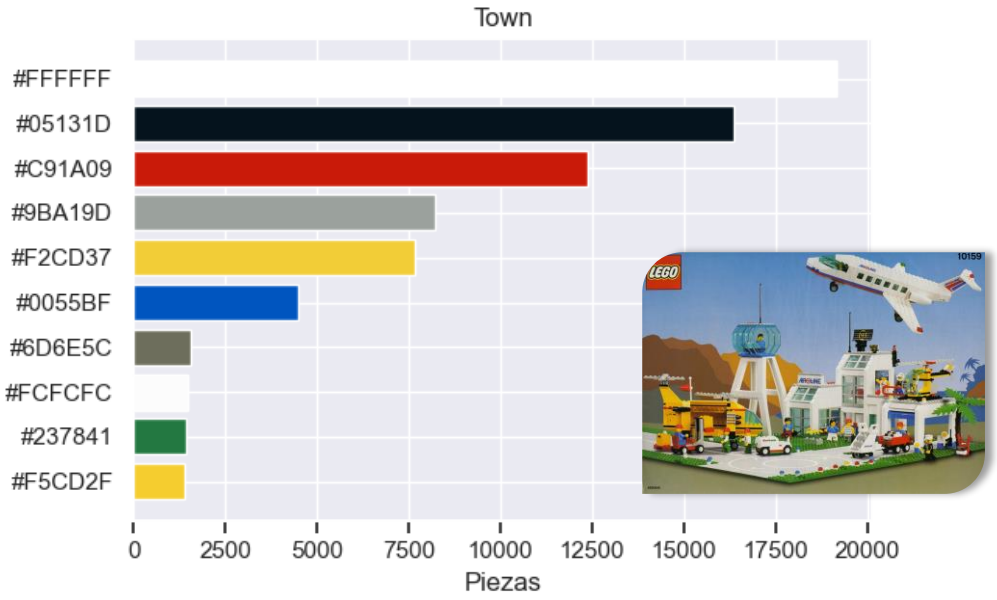


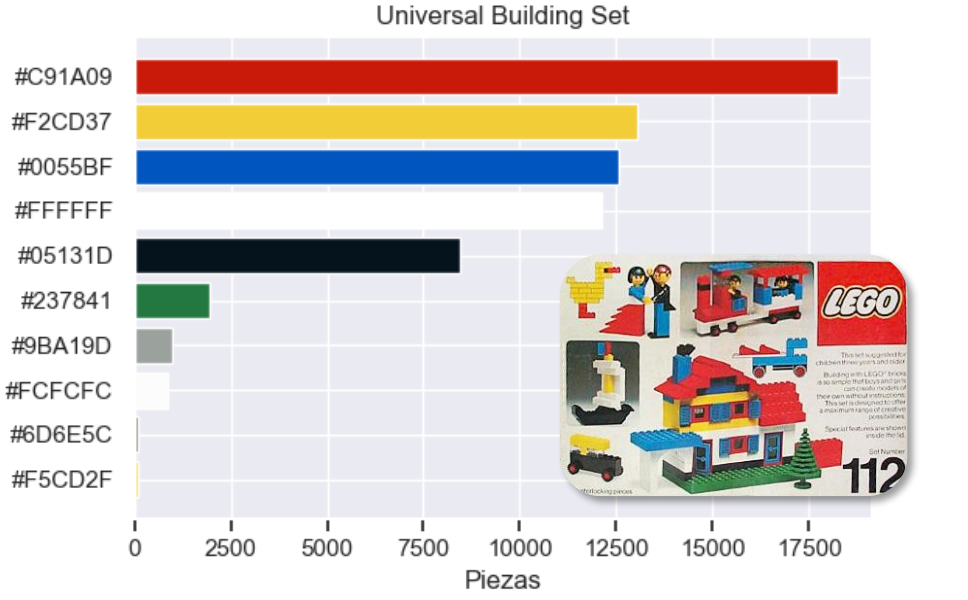


Análisis exploratorio

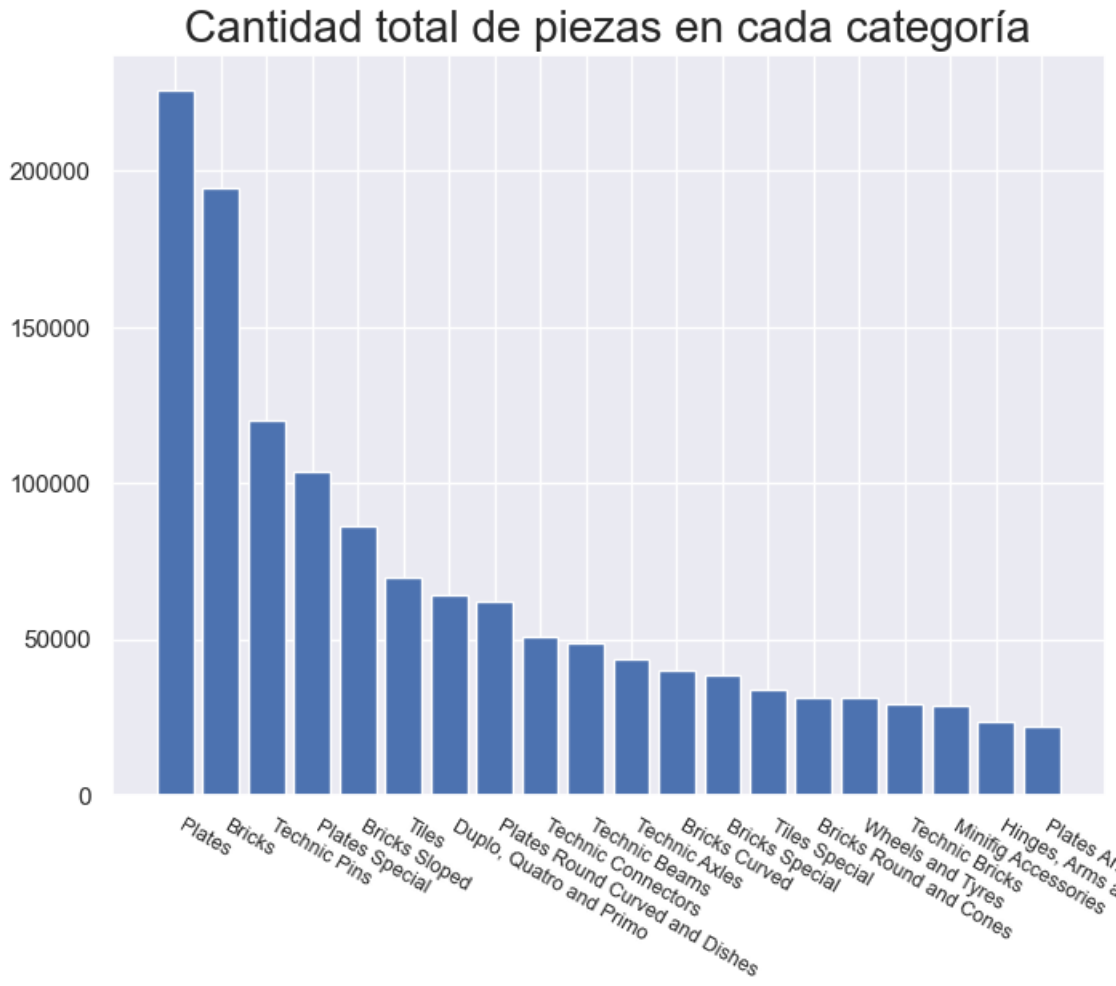








Análisis exploratorio



Ingeniería de features

- Trabajamos con los 10 colores más comunes, todos los materiales y todas las categorías de piezas.

Pasos:

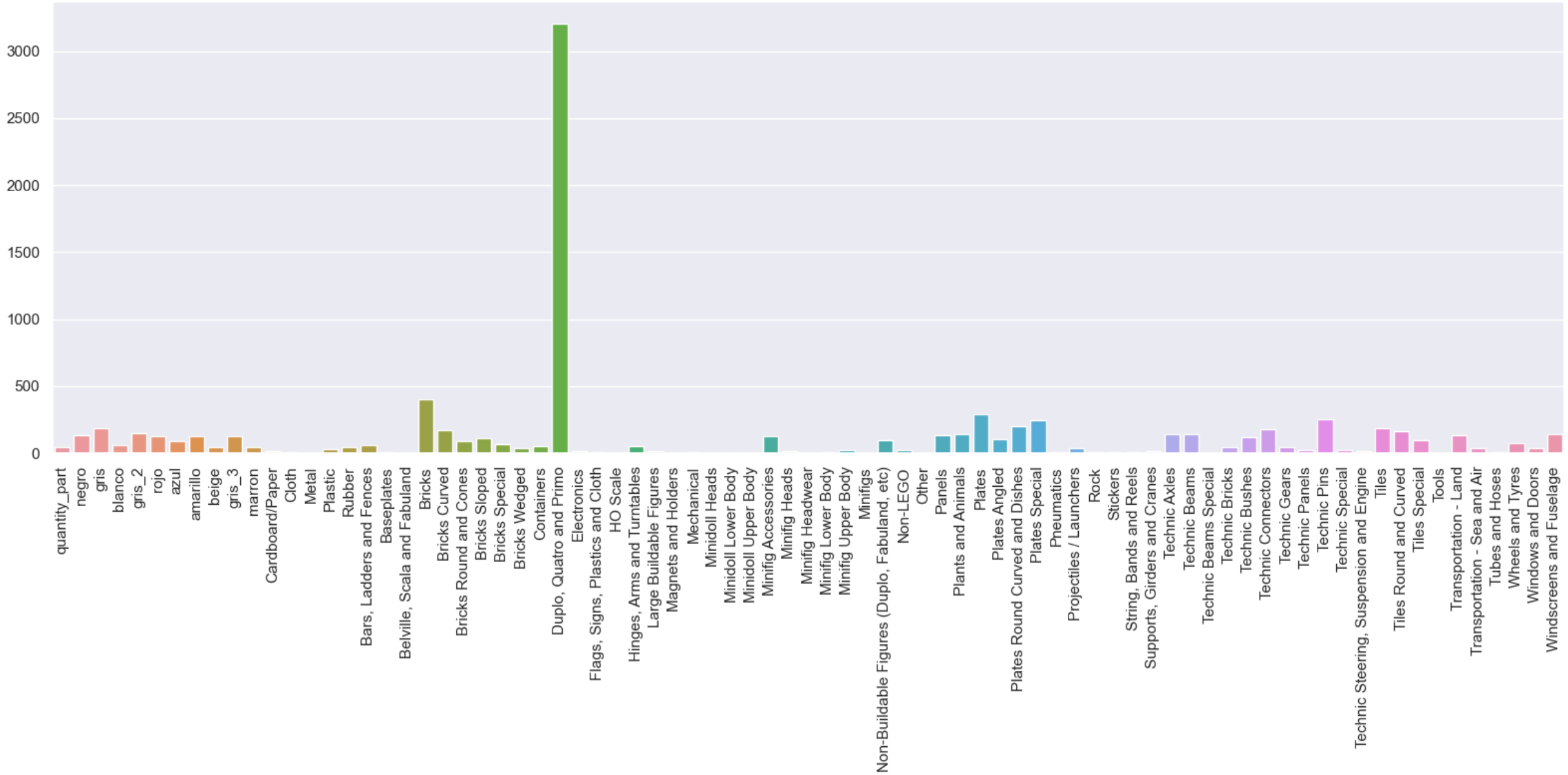
- Creamos dummies para cada categoría.
- Multiplicamos las dummies por ‘quantity_part’ para obtener la cantidad de piezas de cada categoría.
- Agrupamos el dataset sobre la columna ‘set_num’ y sumando todas las piezas.
- Dividimos cada columna por ‘quantity_part’ para obtener las proporciones de piezas en cada set de lego.

	set_num	themes_parent_name	quantity_part	negro	gris	blanco	gris_2	rojo	azul	amarillo	...	Tiles	Tiles Round and Curved	Tiles Special
0	001-1	Technic	43	0.000000	0.000000	0.395349	0.000000	0.302326	0.023256	0.093023	...	0.000000	0.000000	0.000000
1	002-1	Technic	3	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000
2	010-2	Duplo	18	0.000000	0.000000	0.111111	0.000000	0.388889	0.277778	0.222222	...	0.000000	0.000000	0.000000
3	028-1	Duplo	7	0.000000	0.000000	0.285714	0.000000	0.142857	0.285714	0.142857	...	0.000000	0.000000	0.000000
4	030-2	Duplo	29	0.000000	0.000000	0.034483	0.000000	0.379310	0.310345	0.275862	...	0.000000	0.000000	0.000000

6732 rows × 81 columns

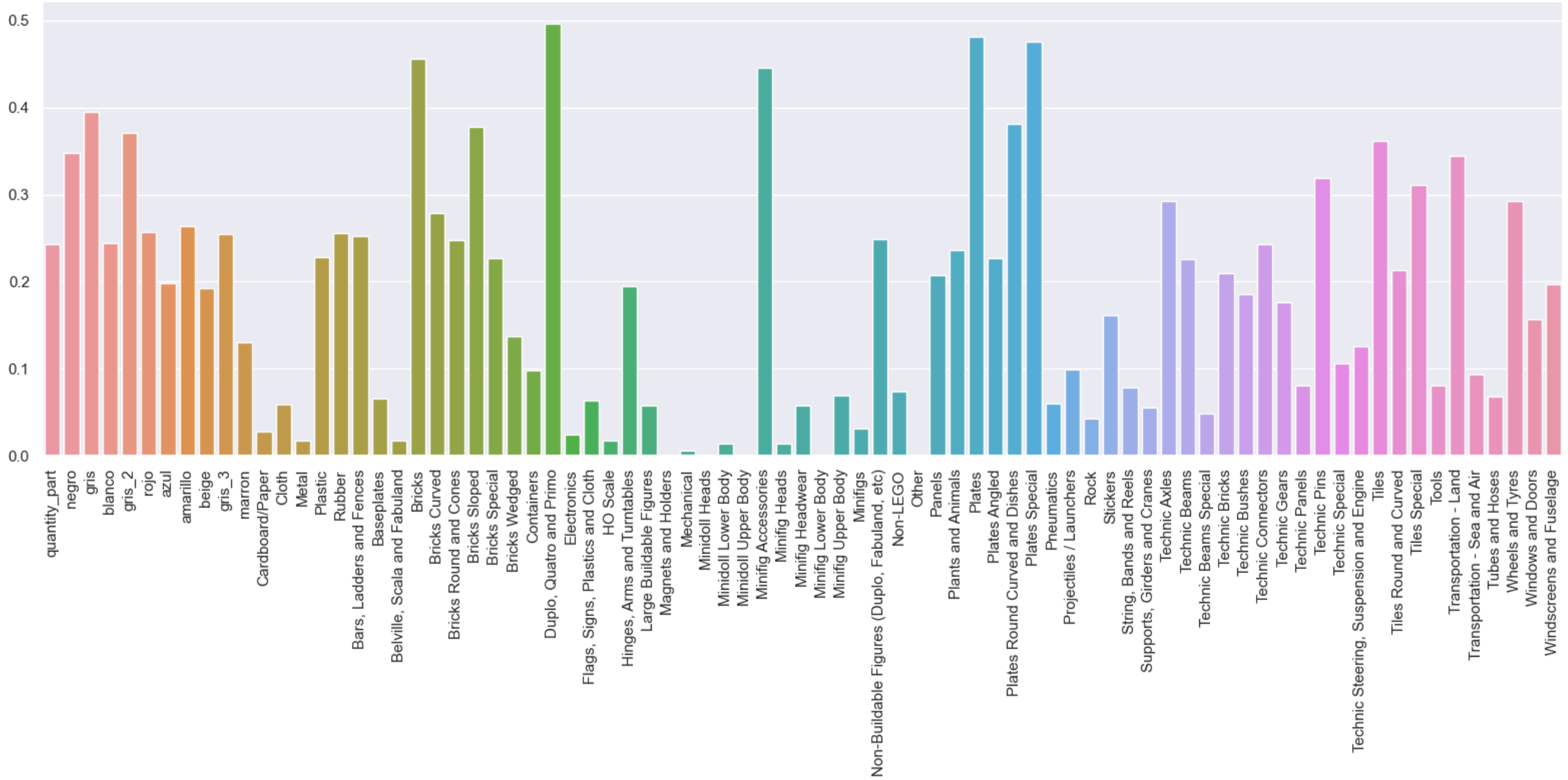
Selección de features

ANOVA



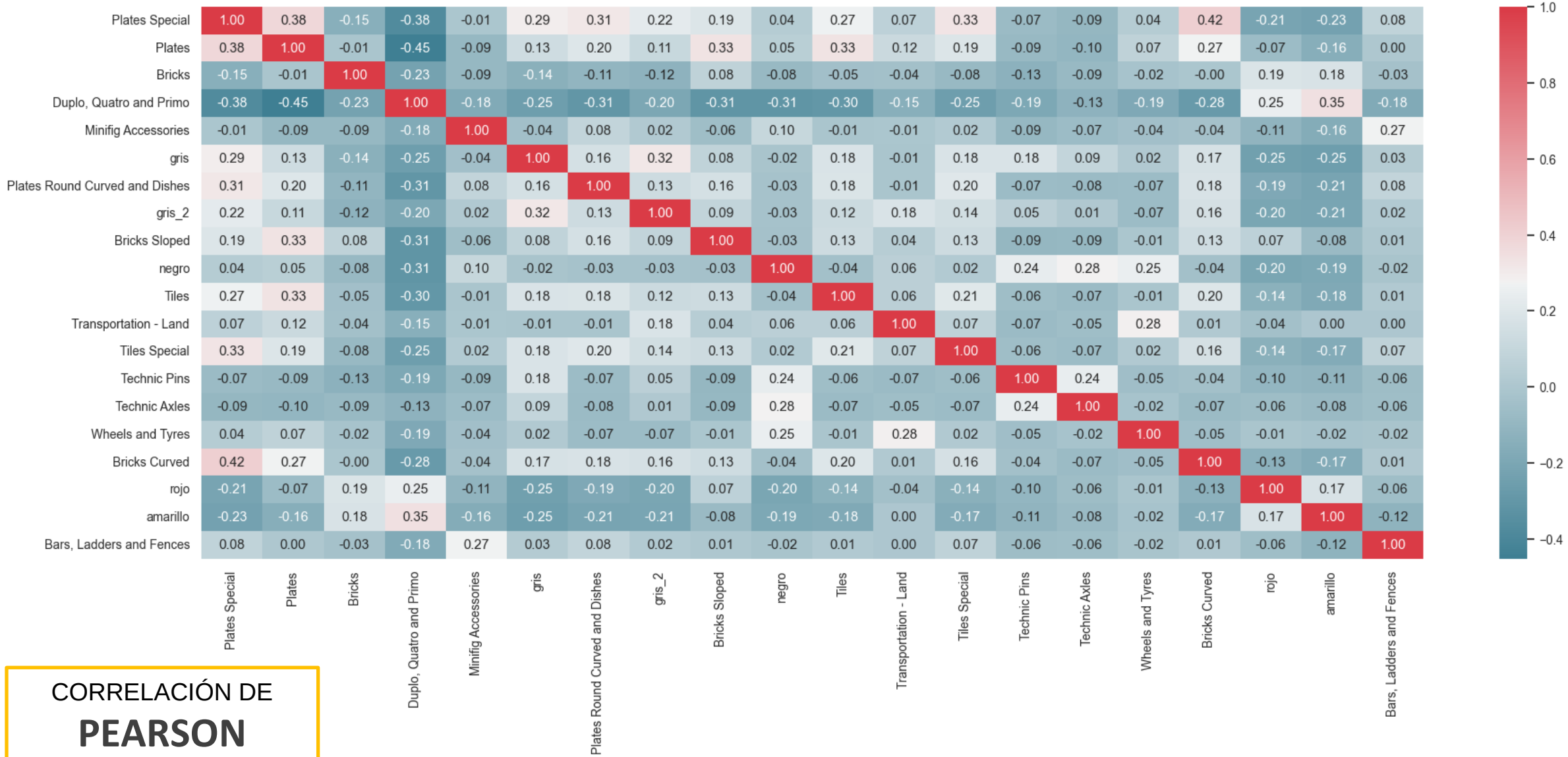
Selección de features

Información mutua



Selección de features

Reducción de features → Información mutua → **20** variables



CORRELACIÓN DE
PEARSON

Selección de features

CORRELACIÓN DE
SPEARMAN

ESPACIO PARA PREGUNTAS

¡MUCHAS GRACIAS!