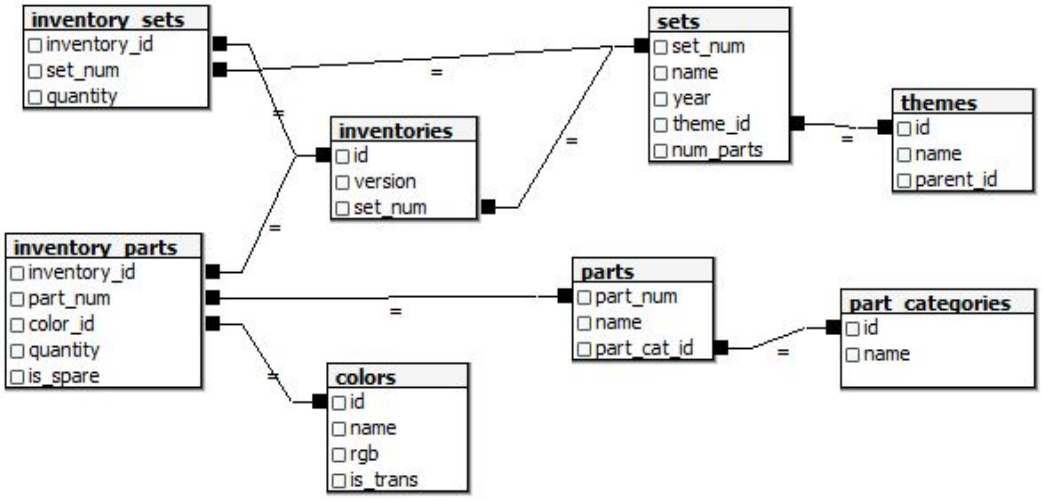
LEGO

Dataset and motivation

Our dataset contains the LEGO Parts/Sets/Colors and Inventories of every official LEGO set in the Rebrickable database. The data was collected on July 2017. LEGO is a popular brand of toy building bricks. They are often sold in sets which contains a number of parts in different shapes, sizes and colors. Each set can build a specific object such as gardens and cars. 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. This is a pretty rich data. The picture show the relationship between each table.



Actual task definition/Research Question

By analyzing the dataset, we are trying to figure out the following questions

- How have the size of sets changed over time?
- What colors are associated with witch themes?
- Could you predict which theme a set is from just by the bricks it contains?
- What sets have the most-used pieces in them?
- What sets have the rarest pieces in them?
- Have the colors of LEGOs included in sets changed over time?

Quality of cleaning

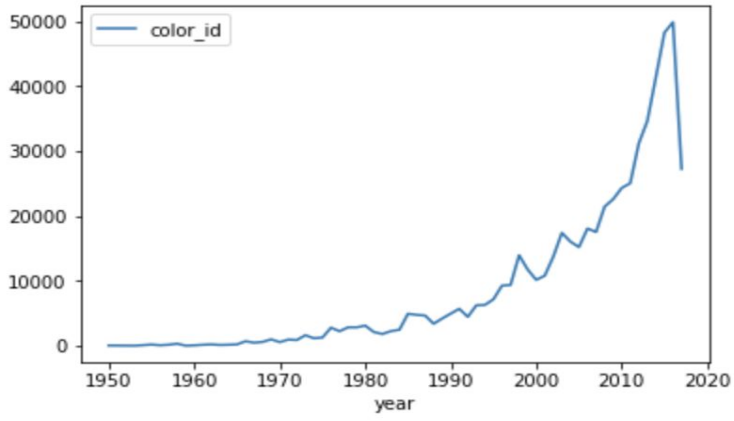
Most of the tables are pretty clean. There are two tables contain some missing values. One is colors and the other one is themes. In colors table, some of the color name are write as unknown. To ensure the integrity of data, we decided to keep these “unknown” in the table instead of deleting them. The themes table has two columns which are “name” and “parent_id”. There are 111 missing values on “parent_id” column. We decided to ignore them because the “name” column can also be the key attribute of this table. Those missing values will not influence our analyze.

Visualization

How the colors of LEGOs included in sets changed over time?

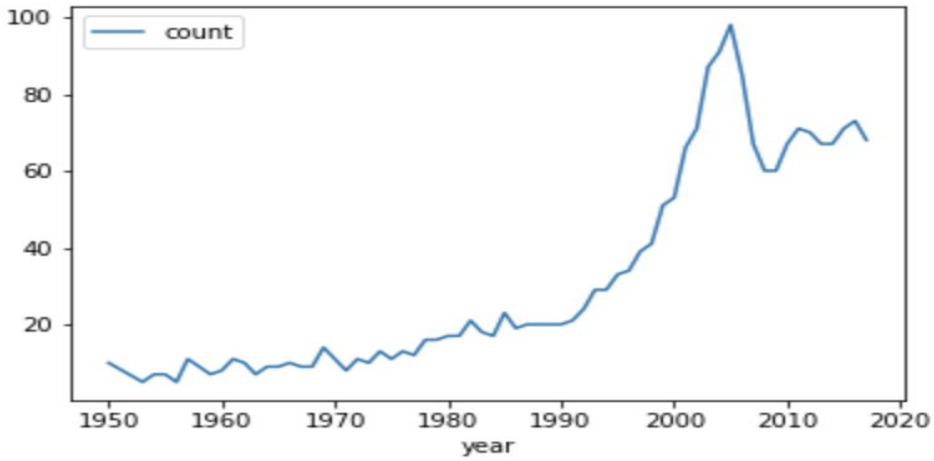
	name_x	rgb	is_trans	inventory_id	part_num	color_id	quantity	is_spare	version	set_num	name_y	year	theme
0	Unknown	0033B2	f	80	belvfair6	-1	1	f	1	5827-1	Royal Coach	2000	319
1	Unknown	0033B2	f	80	belvfem26	-1	1	f	1	5827-1	Royal Coach	2000	319
2	Unknown	0033B2	f	80	belvmale13	-1	1	f	1	5827-1	Royal Coach	2000	319
3	Blue	0055BF	f	80	33211	1	2	f	1	5827-1	Royal Coach	2000	319
4	Blue	0055BF	f	80	33212	1	2	f	1	5827-1	Royal Coach	2000	319

	year	color_id
0	1950	46
1	1953	18
2	1954	98
3	1955	193
4	1956	83

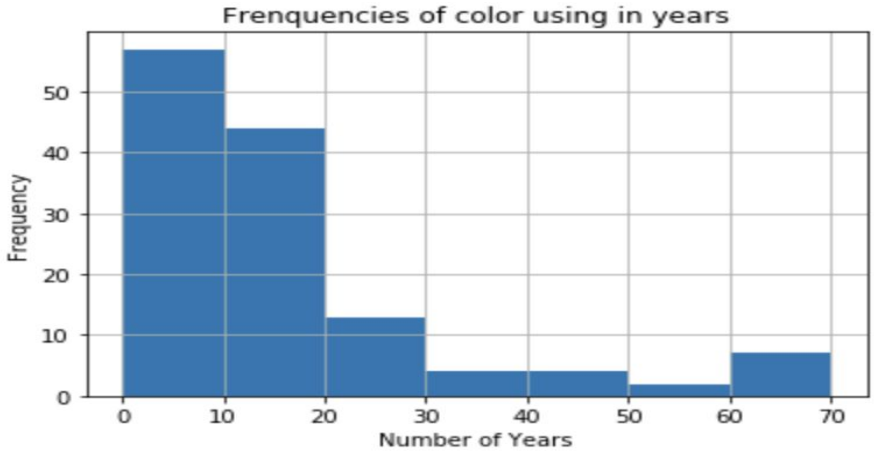


To see the color of sets changed over time, we firstly merge the table of color and set. Then select “year” and “color_id” columns. To see how many colors were using in each year, we count the number of colors and then plot them into the graph. This graph show the number of colors were used every year, include the repeat colors. From the graph, we conclude that the number of pieces color has been using in

each year were increased to the peak of 50000 on around 2015 and decreased sharply to almost 30000 in these years. This is because the pieces size were decreased in 2015.



The graph shows the number of unique color that were used every year. The x-axis shows years and y-axis show numbers. We conclude that Lego has 131 unique colors. The number of unique color increased to the peak of 98 in 2005 and started to decrease.

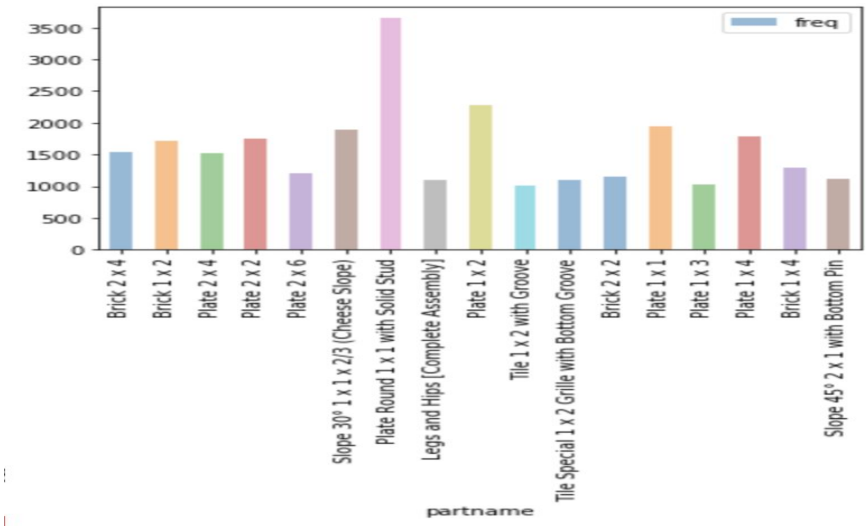


This graph is about the number of color that were using in years. From the graph, we can see that there are almost 8 colors were using more that 60 years and around 60 colors were just started to use in these 10 years. We conclude that Lego is adding more colors in these 20 years.

What sets have the most-used pieces in them?

This research question might Help people who owned some LEGO sets figure out what other sets they could build with the pieces they had. To analyze this problem, we firstly merged the table of parts and sets. Then we tried to find the pieces that were used in more than 1000 sets.

partname	freq
Tile 1 x 2 with Groove	1003
Plate 1 x 3	1020
Legs and Hips [Complete Assembly]	1092
Tile Special 1 x 2 Grille with Bottom Groove	1102
Slope 45° 2 x 1 with Bottom Pin	1117
Brick 2 x 2	1151
Plate 2 x 6	1196
Brick 1 x 4	1286
Plate 2 x 4	1515
Brick 2 x 4	1537
Brick 1 x 2	1711
Plate 2 x 2	1748
Plate 1 x 4	1792
Slope 30° 1 x 1 x 2/3 (Cheese Slope)	1890
Plate 1 x 1	1950
Plate 1 x 2	2282
Plate Round 1 x 1 with Solid Stud	3650



This graph show all of the parts that were used by more than 1000 sets. The number of them is 17. The most frequent one is Plate Round 1*1 with Solid Stud which was used in 3650 sets. From the graph, we conclude that most of the frequency parts were using in 1000 to 1500 sets. Two of them were used by more than 2000 sets.

setname	freq
4 × 4 Fire Truck	17.0
Car	17.0
City Fire Accessory Set	17.0
Fire ATV	17.0
Fire Boat	17.0
Fire Car	17.0
Fire Helicopter	17.0
Fire Hovercraft	17.0
Fire Motorcycle	17.0
Fire Plane	17.0
Fire Rescue	17.0
Fire Station	17.0
Fire Truck	17.0
Fireboat	17.0
Firefighter	17.0
Gahlok-Kal	17.0
Kohrak-Kal	17.0
Lehvak-Kal	17.0
Nuhvok-Kal	17.0
Off-Road Fire Rescue	17.0
Off-Road Fire Truck & Fireboat	17.0
Pahrak-Kal	17.0
Tahnok-Kal	17.0

We also conclude which sets contain these 17 pieces, here is the graph shows the list of those sets. There are 23 sets that contain the most frequent parts. By seeing these, those people who owned most of these sets could built other set with the parts they have.

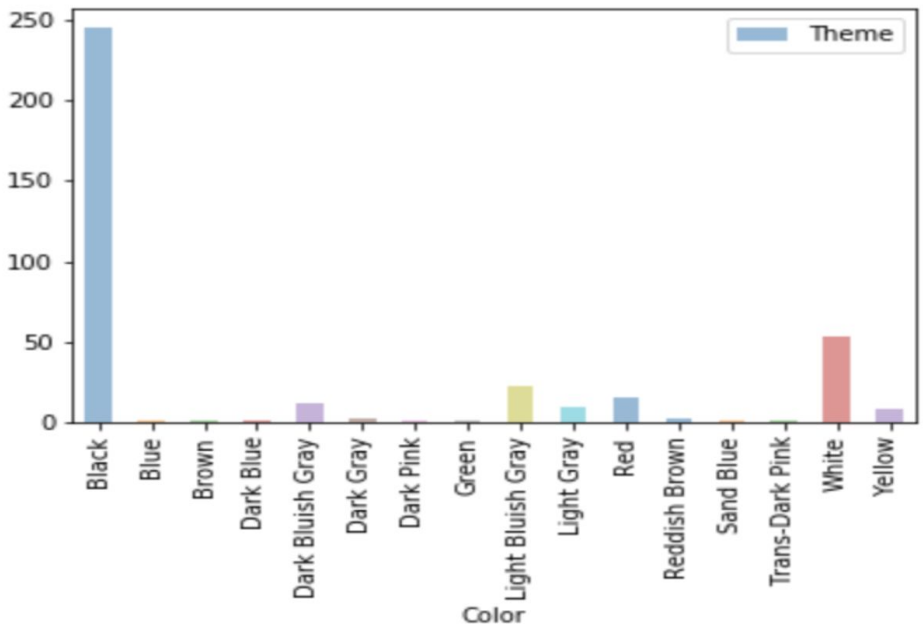
What colors are associated with which themes?

By searching the answer of this question, we merge the table of colors and themes. We found that both of these two tables contain huge amount of data. Each theme contains many sets which have a lot of pieces with different colors. What we firstly did is to count the number of each color that was used in every theme.

Theme	Color	Count
12V	Black	576
	Red	326
	Light Gray	314
	Blue	127
	Yellow	121
	White	119
	Trans-Clear	56
	Dark Gray	25
	Trans-Red	13
	[No Color]	11
	Trans-Yellow	9
	Brown	4
	Green	1
	Royal Blue	1
4 Juniors	White	10
4.5V	Black	920
	Red	631
	Blue	331
	Light Gray	327
	Yellow	281
	White	210
	Trans-Clear	83
	Dark Gray	18
	Brown	6
	Green	6

There are 379 themes that were counted, the table above show the partial data as an example.

	Color	Theme
0	Black	245
1	Blue	1
2	Brown	1
3	Dark Blue	1
4	Dark Bluish Gray	12
5	Dark Gray	2
6	Dark Pink	1
7	Green	1
8	Light Bluish Gray	23
9	Light Gray	10
10	Red	16
11	Reddish Brown	3
12	Sand Blue	1
13	Trans-Dark Pink	1
14	White	53
15	Yellow	8

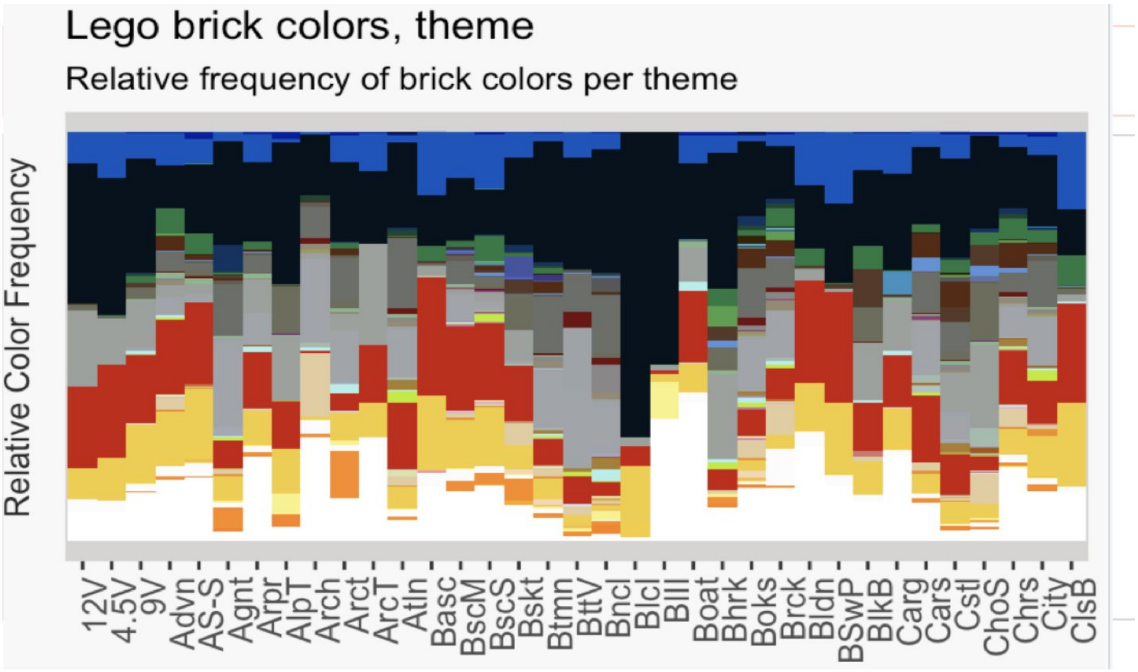


Next, to see which color is the most popular one, we find the most frequent colors for each theme and then count the number of them as the most frequent one in each theme. The above table and graph show the result we got. There are 16 unique colors as the most frequent color in 379 themes. Black, as the most popular color, becomes 245 themes main color. White, gray and red are also two of the most popular colors.

	Theme	Color	Count
0	12V	Black	0.338227
1	12V	Red	0.191427
2	12V	Light Gray	0.184381
3	12V	Blue	0.074574
4	12V	Yellow	0.071051
5	12V	White	0.069877
6	12V	Trans-Clear	0.032883
7	12V	Dark Gray	0.014680
8	12V	Trans-Red	0.007634
9	12V	[No Color]	0.006459
10	12V	Trans-Yellow	0.005285
11	12V	Brown	0.002349
12	12V	Green	0.000587
13	12V	Royal Blue	0.000587

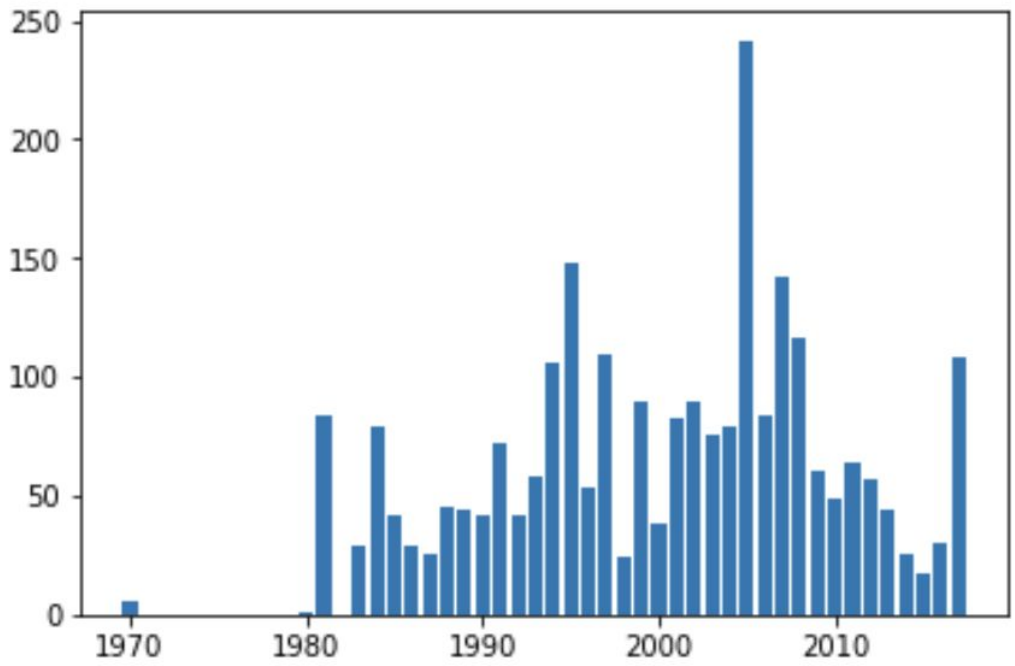
We also count the radio of each color in every theme. By doing this, we count the total number of frequencies and divided it by each color's frequency. The table above shows the example of each color ratio in theme 12V.

To see the data clearly, we draw the graph below to show them. The x-axis show the name of each theme, the length of color stands for the ration of them in each theme. From the graph, we can clearly see that black weigh too much on most of the themes. Moreover, the second important color is white. Also there exist some themes that only contain one single color.



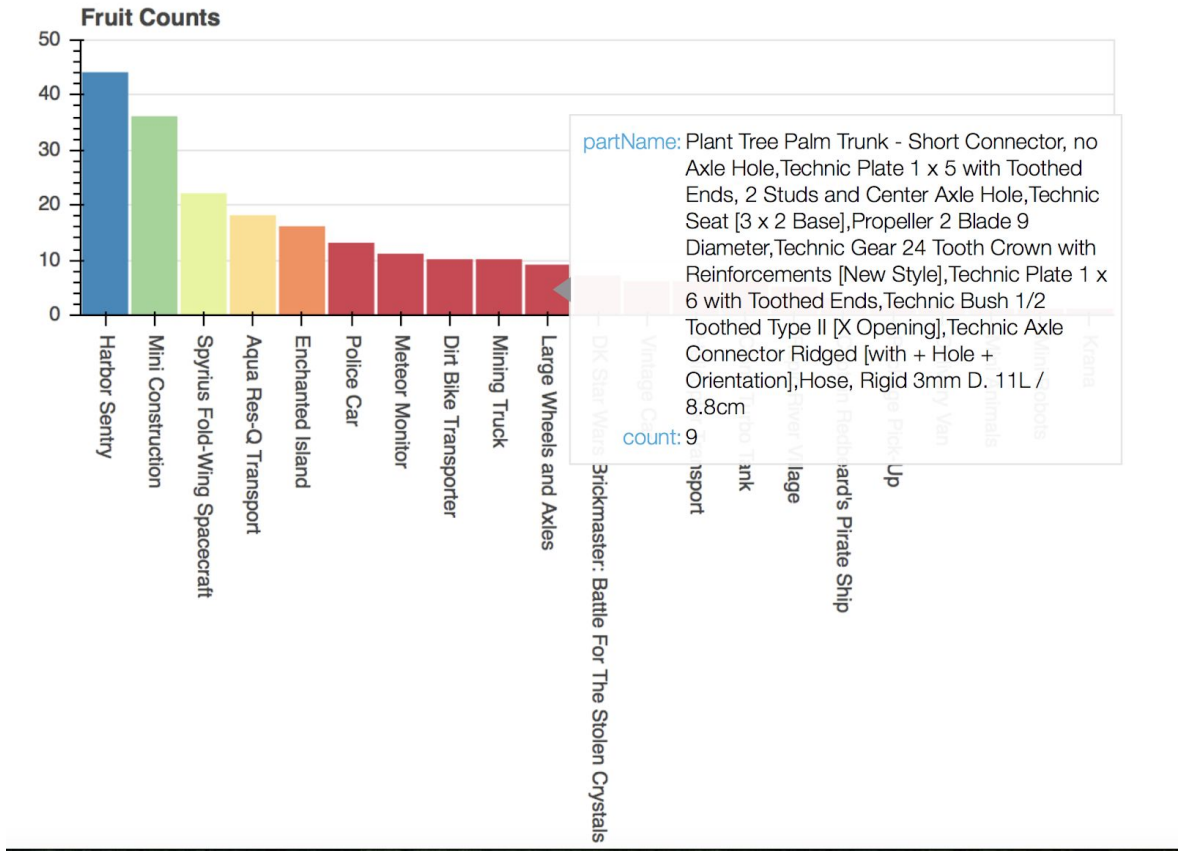
How have the size of set change over time?

By counting the parts for each year and each set and average the parts number by number of sets for each year, we get the graph below. We conclude that the average size of set increase from 1970 to 2005 and decrease after 2005. We think this pattern is same with the pattern above (number of unique color were used over year). The number of unique colors in each set changes cause the number of parts change in each set over year.



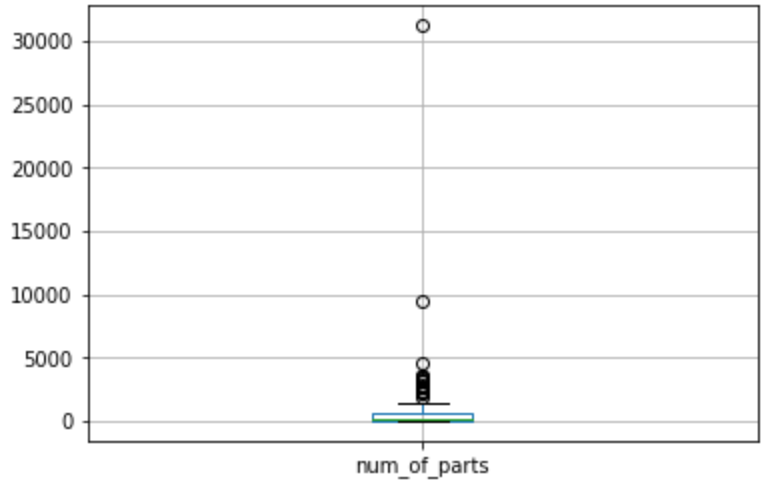
What set have the rarest pieces in them?

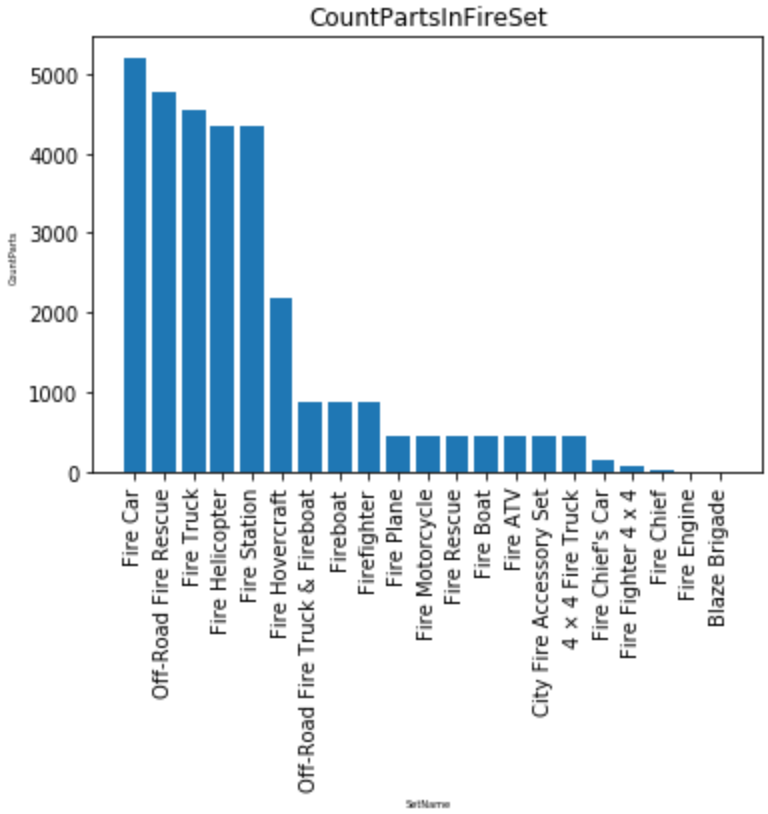
By analyzing this question, we found that there are 21 sets have rarest pieces in them. They are Harbor Sentry, Mini Construction, Spyrius Fold-Wing Spacecraft and Aqua Res-Q Transport and so on. This graph also show the part name which only occurs once in the whole dataset in a specific set. For example, in the Large Wheels and Axles set, there are 9 unique part that only appear in this set and never appear in other sets. The partName clearly describes the name of these parts. After the research, we found that these rarest pieces are some special characters in the set, such as a Tree.



How many parts in each theme?

We check the distribution of parts and we found some points are really looks like outliers. As both graph show, there is a theme which has over 30000 parts. Then we check if this theme (Fire) is outliers. After our check in the data, we found that Fire theme actually has 21 sets and each set have many parts. It is reasonable that Fire theme have over 30000 parts. Also, we search online, we found that Fire actually includes many different sets and each set includes many parts. Therefore, we conclude that there is no outliers in this data.





Modelling activity

Predict theme name by parts it contains:

We use naive bayes to do prediction because the predicted variable is categorical. In this question, we first analyze the question then we decide to use part name a theme contains to predict the theme name. We used categorical response variable to predict categorical variable.

The assumption of naive bayes is assume each response variable is independent and each response variable have equally contributed to predicted variable.

Then we compute the accuracy of the model. The accuracy rate is 0.247. This rate shows that only 24.7% of data can be predicted correctly. This accuracy rate is pretty low and indicates that naive bayes is not fit this data. We think another reason to make the accuracy low is we divided the data into two subset train and test randomly. Some themes appear in the test data may not appear in the train data. Because these themes never appears in the train that makes the prediction result of these theme is 0. But we don't think it is the main reason.

Evaluation

We found that we can not just use the part name to predict the theme name. If we want to increase the accuracy of the prediction, we need more information about parts about each theme. Another finding is naive bayes is not the good for our situation. We may try some clustering algorithms in the future.

Conclusion

By analyzing these six research question, we conclude that Lego has 131 unique colors and the most popular colors are white, black ,red, gray among both sets and parts.In year 2005 has got maximum unique colors which are number of 98. There are 17 parts were used by 1000 sets. The most frequent piece has been used in 3650 sets.