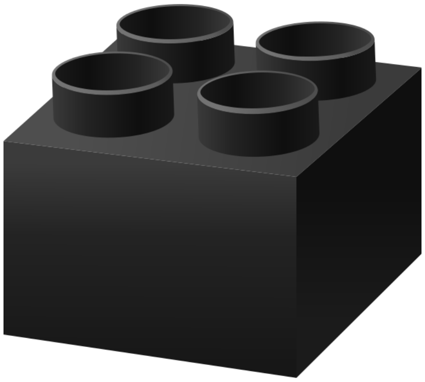
LEGO

DESCRIPTIVE DATA ANALYSIS



Report by Hyerim Hwang
Pratt Institute

Introduction

A LEGO dataset has been always a curiosity to me after analyzing it for the group project performed in last semester. Since LEGO has a wide variety of parts, sets, themes, etc, and they all have individual names and ids, I barely took selected variables to create new tables using Rstudio. I was not good at Rstudio, so I created charts and dashboards in Tableau Public after merging LEGO's tables in Rstudio. At that time, I focused more on LEGO's themes and actual customers' reviews on the official LEGO site. I had a look at the generall changes on LEGO's sets over time through the dataset on this report.

Data Source

A LEGO fan website Rebrickable offers a database of Parts/Sets/Colors and Inventories of every official LEGO set through fetching API or downloading csv files (Rebrickable). An elaborate and complex database, based on the relational schema, has multiple datasets that are composed of each table that contains all factors relate to LEGO's Parts/Sets/Colors and Inventories (Fig. 1). Each stitched line represents how each table relates to each other.

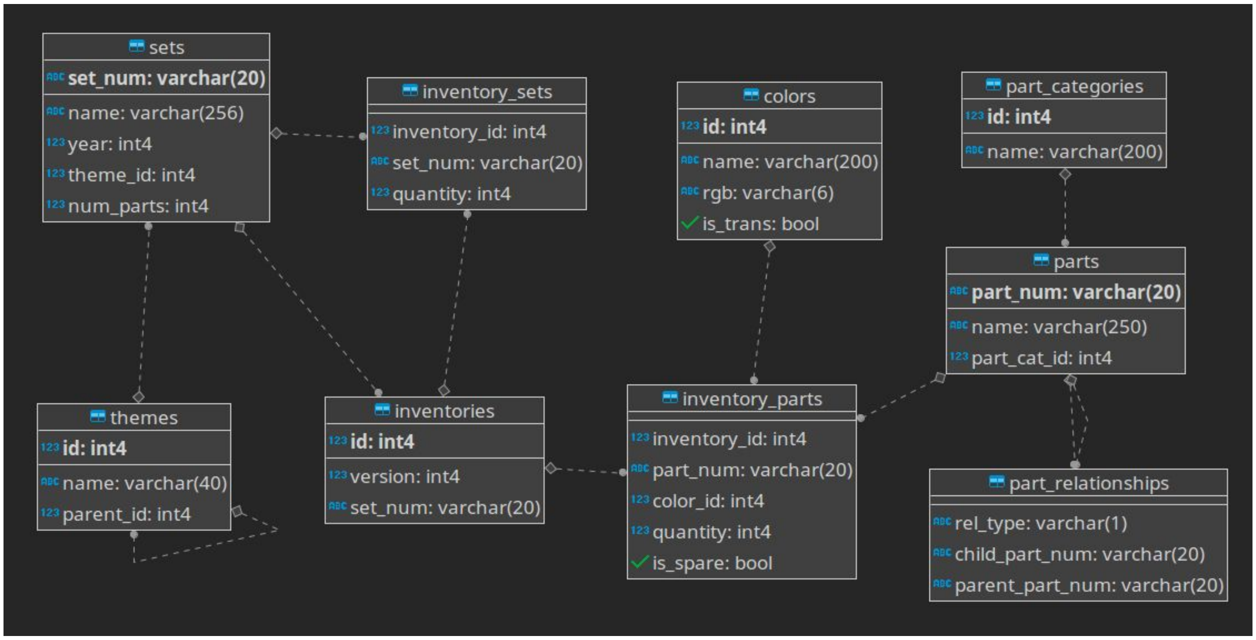


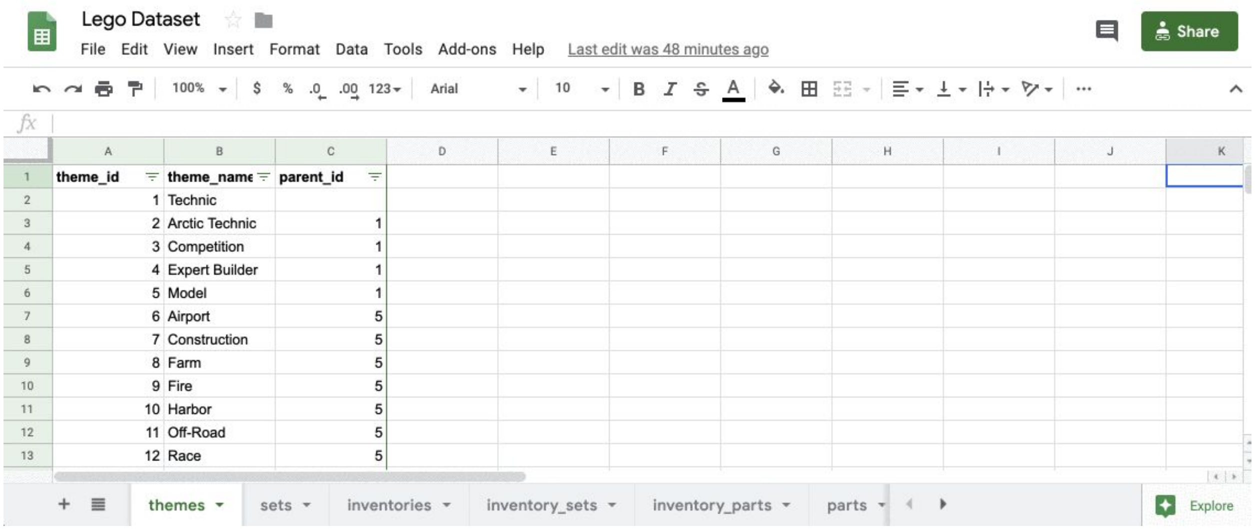
Fig. 1: The information is stored in different tables.

There is a title with an icon on the top that contains the name of the table, and a second-placed variable is the "primary key" which always is an integer that designated to uniquely identify all table records and it indicates "the unique value for each row" (Primary key). Some table doesn't contain the primary key, but it does have, at least, another table's primary key, in this case, we call it as "foreign key" when one table has other primary keys. Through any foreign key, more than two databases can be merged.

One of the limitations of the dataset will be the separation of data. Since nine separated tables contain unique variables, it was impossible to take all variables that need to take lots of merging actions with multiple tables.

Data Cleaning and Preparing

To see all datasets at a glance, it was a prerequisite to open them into Google sheets and learn what is in the datasets(Fig. 2).



theme_id	theme_name	parent_id
1	Technic	
2	Arctic Technic	1
3	Competition	1
4	Expert Builder	1
5	Model	1
6	Airport	5
7	Construction	5
8	Farm	5
9	Fire	5
10	Harbor	5
11	Off-Road	5
12	Race	5

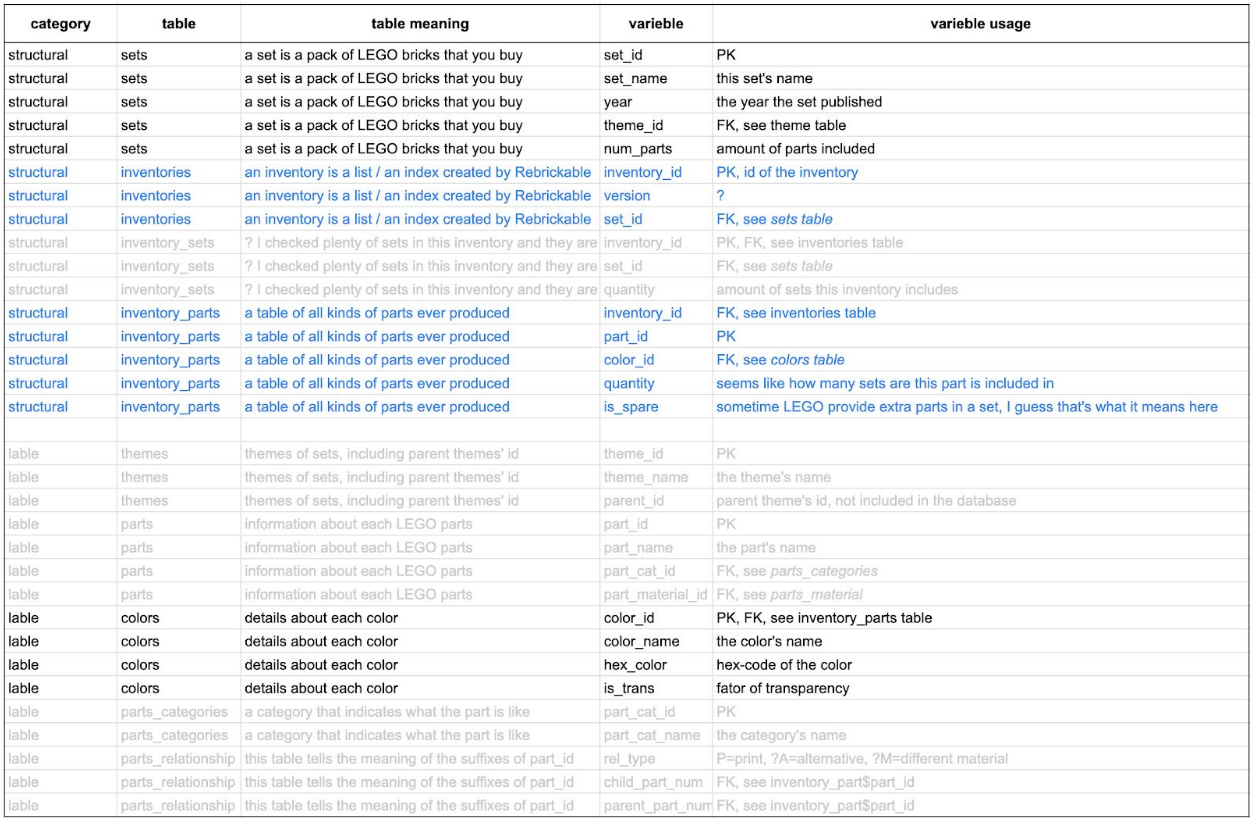
Fig. 2: Google Sheets were used to examining nine csv files. Excel or other programs would also work well with the csv file.

I found some facts that needed to be done before conducting analysis and placed it in a list as shown below.

1. The variables of the dataset need to be described and listed.
2. The tables are needed to find how to merge the tables within R Studio.
3. The table needs to be created which will contain all datasets to figure out the hypothesis of the report, the general changes in LEGO's sets over time specially in colors.

Variables

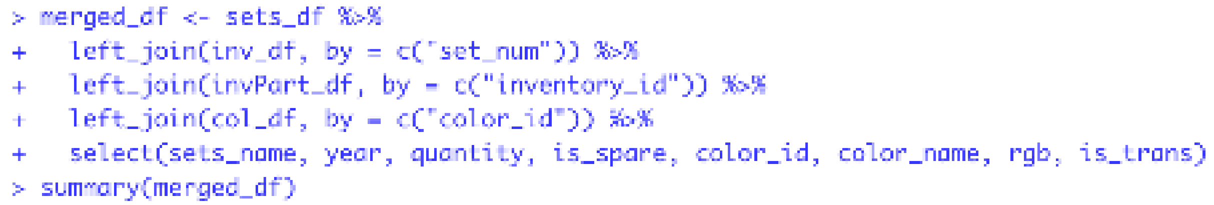
To understand the datasets, there are two ways to categorize the variables by structures and labels (See the first column named "category" in Fig. 3). The structure is a higher-level of dataset covering the core infrastructure that LEGO has. The label is a secondary-level of the dataset contains more detailed information. The black-colored variables are the dataset of sets and colors that precisely I need, and the blue-colored variables are required to merge datasets of sets and colors.



category	table	table meaning	varieble	varieble usage
structural	sets	a set is a pack of LEGO bricks that you buy	set_id	PK
structural	sets	a set is a pack of LEGO bricks that you buy	set_name	this set's name
structural	sets	a set is a pack of LEGO bricks that you buy	year	the year the set published
structural	sets	a set is a pack of LEGO bricks that you buy	theme_id	FK, see theme table
structural	sets	a set is a pack of LEGO bricks that you buy	num_parts	amount of parts included
structural	inventories	an inventory is a list / an index created by Rebrickable	inventory_id	PK, id of the inventory
structural	inventories	an inventory is a list / an index created by Rebrickable	version	?
structural	inventories	an inventory is a list / an index created by Rebrickable	set_id	FK, see *sets table*
structural	inventory_sets	? I checked plenty of sets in this inventory and they are	inventory_id	PK, FK, see inventories table
structural	inventory_sets	? I checked plenty of sets in this inventory and they are	set_id	FK, see *sets table*
structural	inventory_sets	? I checked plenty of sets in this inventory and they are	quantity	amount of sets this inventory includes
structural	inventory_parts	a table of all kinds of parts ever produced	inventory_id	FK, see inventories table
structural	inventory_parts	a table of all kinds of parts ever produced	part_id	PK
structural	inventory_parts	a table of all kinds of parts ever produced	color_id	FK, see *colors table*
structural	inventory_parts	a table of all kinds of parts ever produced	quantity	seems like how many sets are this part is included in
structural	inventory_parts	a table of all kinds of parts ever produced	is_spare	sometime LEGO provide extra parts in a set, I guess that's what it means here
lable	themes	themes of sets, including parent themes' id	theme_id	PK
lable	themes	themes of sets, including parent themes' id	theme_name	the theme's name
lable	themes	themes of sets, including parent themes' id	parent_id	parent theme's id, not included in the database
lable	parts	information about each LEGO parts	part_id	PK
lable	parts	information about each LEGO parts	part_name	the part's name
lable	parts	information about each LEGO parts	part_cat_id	FK, see *parts_categories*
lable	parts	information about each LEGO parts	part_material_id	FK, see *parts_material*
lable	colors	details about each color	color_id	PK, FK, see inventory_parts table
lable	colors	details about each color	color_name	the color's name
lable	colors	details about each color	hex_color	hex-code of the color
lable	colors	details about each color	is_trans	fator of transparency
lable	parts_categories	a category that indicates what the part is like	part_cat_id	PK
lable	parts_categories	a category that indicates what the part is like	part_cat_name	the category's name
lable	parts_relationship	this table tells the meaning of the suffixes of part_id	rel_type	P=print, ?A=alternative, ?M=different material
lable	parts_relationship	this table tells the meaning of the suffixes of part_id	child_part_num	FK, see inventory_part$part_id
lable	parts_relationship	this table tells the meaning of the suffixes of part_id	parent_part_num	FK, see inventory_part$part_id

Fig. 3: A column: table refers to the name of tables, and what is the meaning of it was placed next to it. The column: category and variable usage were created to examine the data.

To analyze the general changes in LEGO's sets over time especially in colors, the variables were selected from "sets_name, year" in sets table, "quantity, is_spare" in inventory table, and "color_id, color_name, RGB, is_trans" in the colors table. The Joining Data in R with dplyr helped how to merge more than two tables within foreign keys, and here is my merged_df data from four datasets (Fig. 4).



```
> merged_df <- sets_df %>%
+   left_join(inv_df, by = c("set_num")) %>%
+   left_join(invPart_df, by = c("inventory_id")) %>%
+   left_join(col_df, by = c("color_id")) %>%
+   select(sets_name, year, quantity, is_spare, color_id, color_name, rgb, is_trans)
> summary(merged_df)
```

Fig. 4: The merged_df table was created by four different datasets, sets_df, inv_df, invPart_df, and col_df.

Outliers

To specify the data to get certain analyses, other variables were not selected such as inventory_sets, themes, parts, parts_categories, and parts_realationship (See all grey-colored texts in Fig. 3). Also some of the variables like "version" in the inventories table, or "part_id" in the inverntory_parts table, they will treat as an outlier since the report doesn't deal with part information of LEGO.

Results and Observations

Overall, Lego sets have been produced in a wide variety of colors and sets from 1949 to 2019. The data visualization obviously seemed that LEGO began to produce more variety of kinds in their production.

Observations 1: Changes in quantities of sets over time

I created a plot graph containing all sets' datasets that helped me understand the general rising pattern which clearly shows that the complexity of LEGO (Fig. 5). However, the graph is clustered under 2,000 pieces and some LEGO sets are located in a much higher position. Also, a gradually increased trend line indicates that the quantity of sets hasn't changed that much.

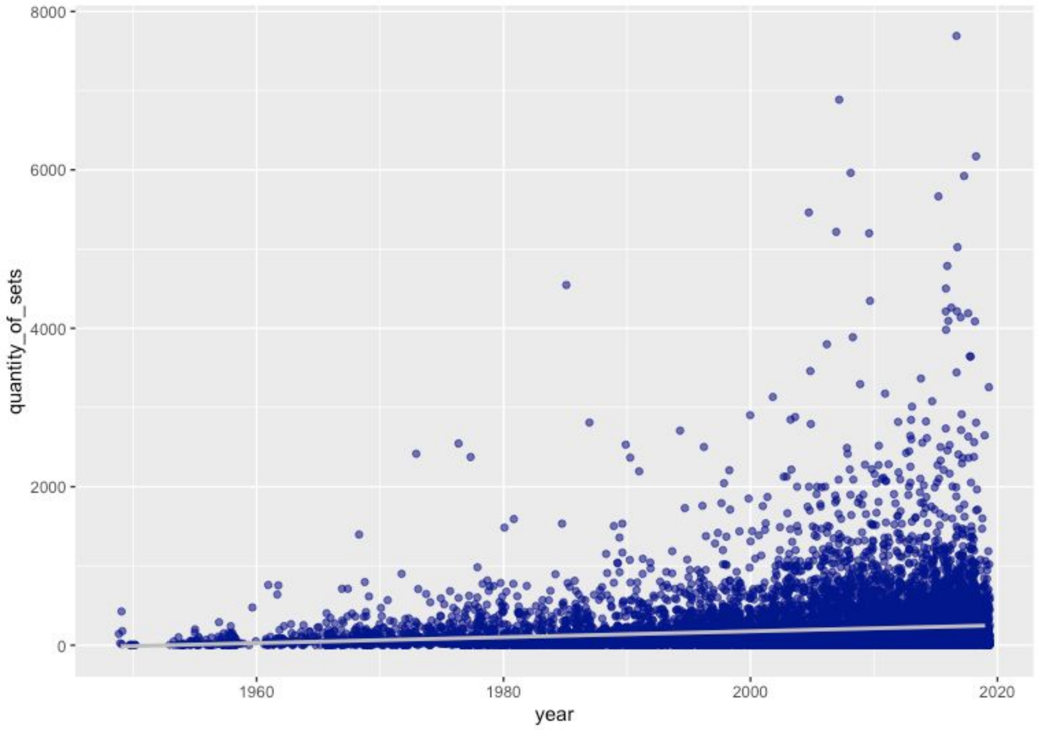
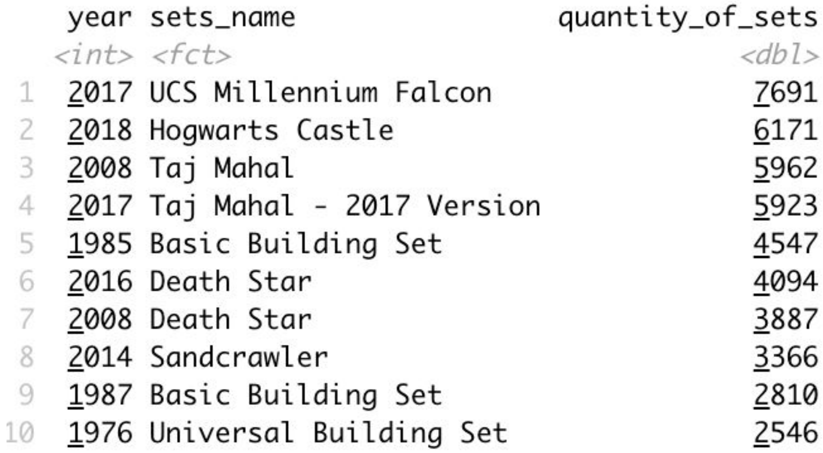


Fig. 5: A lot of points are clustered with less meaning.

The summary of the dataset had a fact that the top 3 of the list, “Basic Building Set”, “Universal Building Set” and “Hogwarts Castle” got enormous pieces out of all (Fig. 6).



```
    year sets_name                    quantity_of_sets
   <int> <fct>                                   <dbl>
 1  2017 UCS Millennium Falcon                    7691
 2  2018 Hogwarts Castle                          6171
 3  2008 Taj Mahal                                5962
 4  2017 Taj Mahal - 2017 Version                 5923
 5  1985 Basic Building Set                       4547
 6  2016 Death Star                               4094
 7  2008 Death Star                               3887
 8  2014 Sandcrawler                              3366
 9  1987 Basic Building Set                       2810
10  1976 Universal Building Set                   2546
```

Fig. 6: The type of building sets listed on the top 3.

To track the top 10 sets movements, I filtered seven of them to see their trend over time. The 'Death Star', ‘Sandcrawler’, 'UCS Millennium Falcon', 'Hogwarts Castle' are quite recent released LEGO sets, according to Fig. 7. The 'Fire Station' must be steady seller LEGO sets and the 'Basic Building Set' and 'Universal Building Set' don’t produce any more these days. It seems to show that the more quantity of LEGO sets used to related to building sets and is influenced by users’ needs or trends.

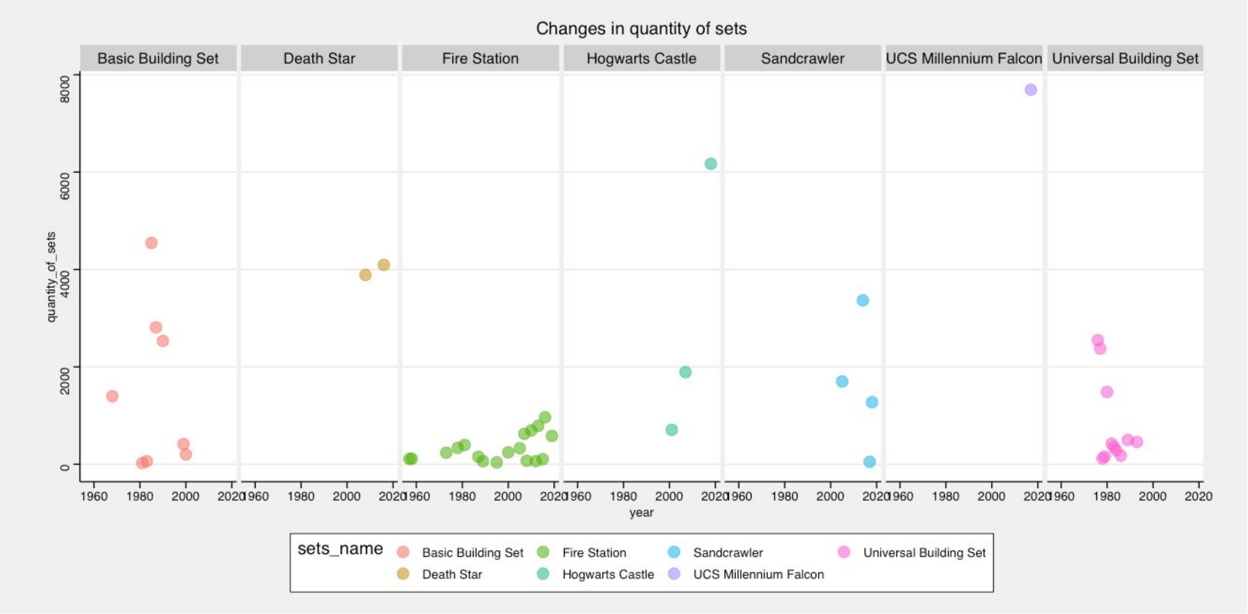


Fig. 7: Top seven the most have pieces in the sets were selected.

Observations 2: Changes in colors over time

To see how LEGO produces more and more colors over time(1929 - 2019), I created a chart to see their rising complexity using the count of color_id on the y-axis (Fig. 8). Since the dataset contains the RGB code, I tried to find how I can connect with the color code and point, but it didn't work out.

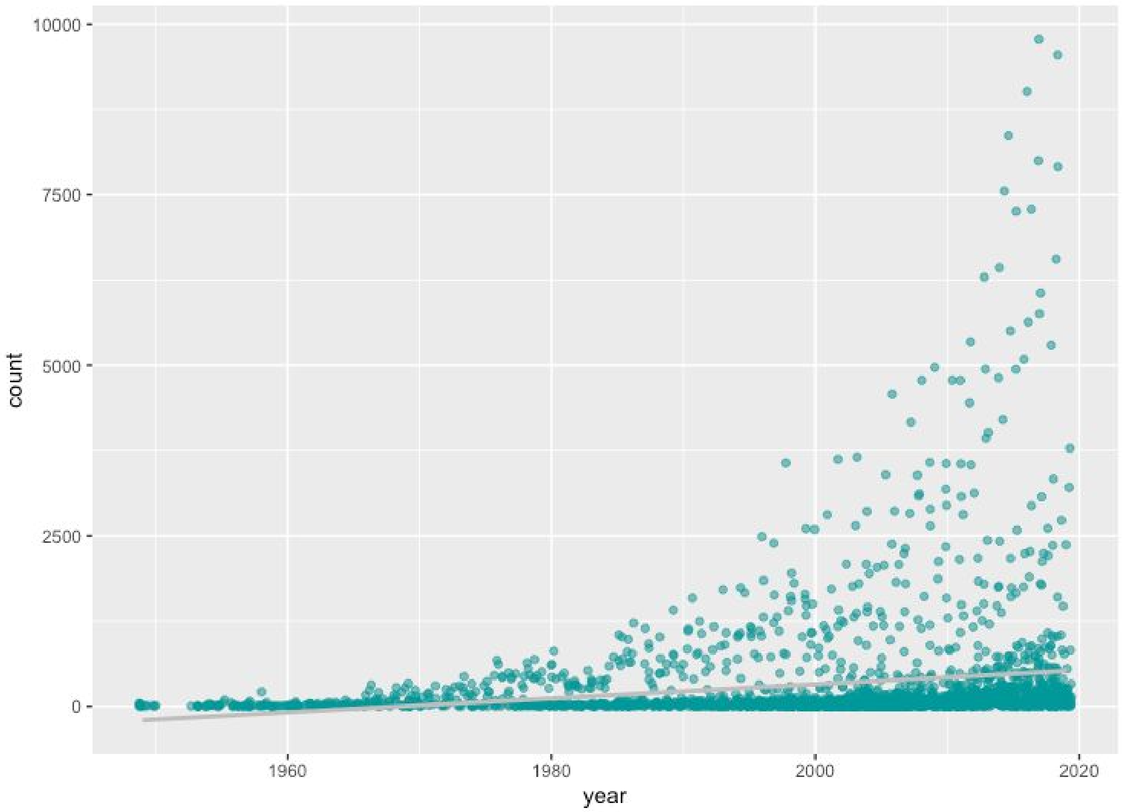
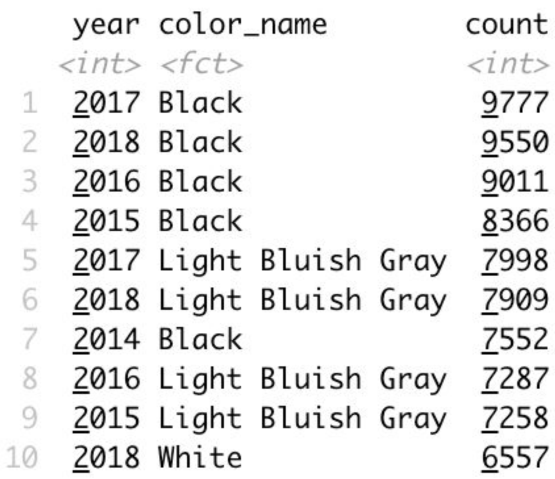


Fig. 8: The points represent the color name, It shows that the color is rising in its variety.

Apparently, black is the most popular color for all sets from 2015 to 2017. It has been produced more than 8,000 times per year (Fig. 9). briefly tells me that monotone color seems to apply to many sets which listed with black, light bluish grey and white. I assumed that LEGO increased their color more and more over time, but the top 10 colors referred to me that they are still popular.



```
    year color_name         count
   <int> <fct>              <int>
 1  2017 Black               9777
 2  2018 Black               9550
 3  2016 Black               9011
 4  2015 Black               8366
 5  2017 Light Bluish Gray   7998
 6  2018 Light Bluish Gray   7909
 7  2014 Black               7552
 8  2016 Light Bluish Gray   7287
 9  2015 Light Bluish Gray   7258
10  2018 White               6557
```

Fig. 9: The Black color is the most prominent color out of all the LEGO bricks.

It clearly shows that black and white colors are the steadily most popular color to use in any sets, and dark bluish-gray and reddish-brown colors are recently rising to frequently use in sets. Since black and dark bluish-gray colors get the most rapid peaks in the table, these two colors' trend will still on the top of the list.

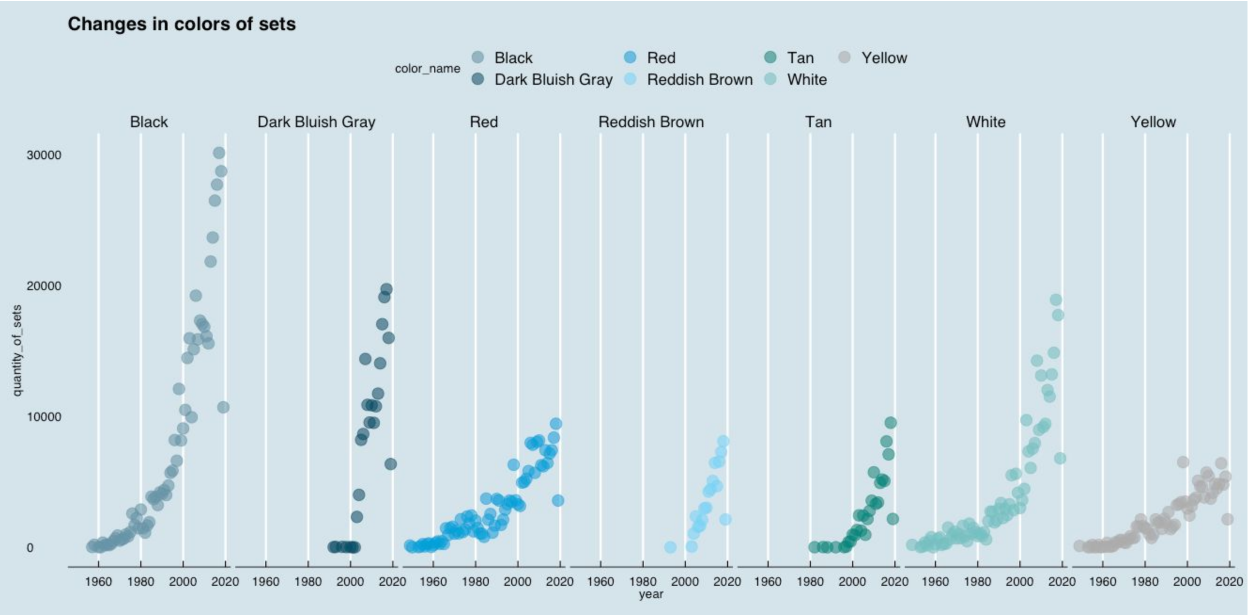
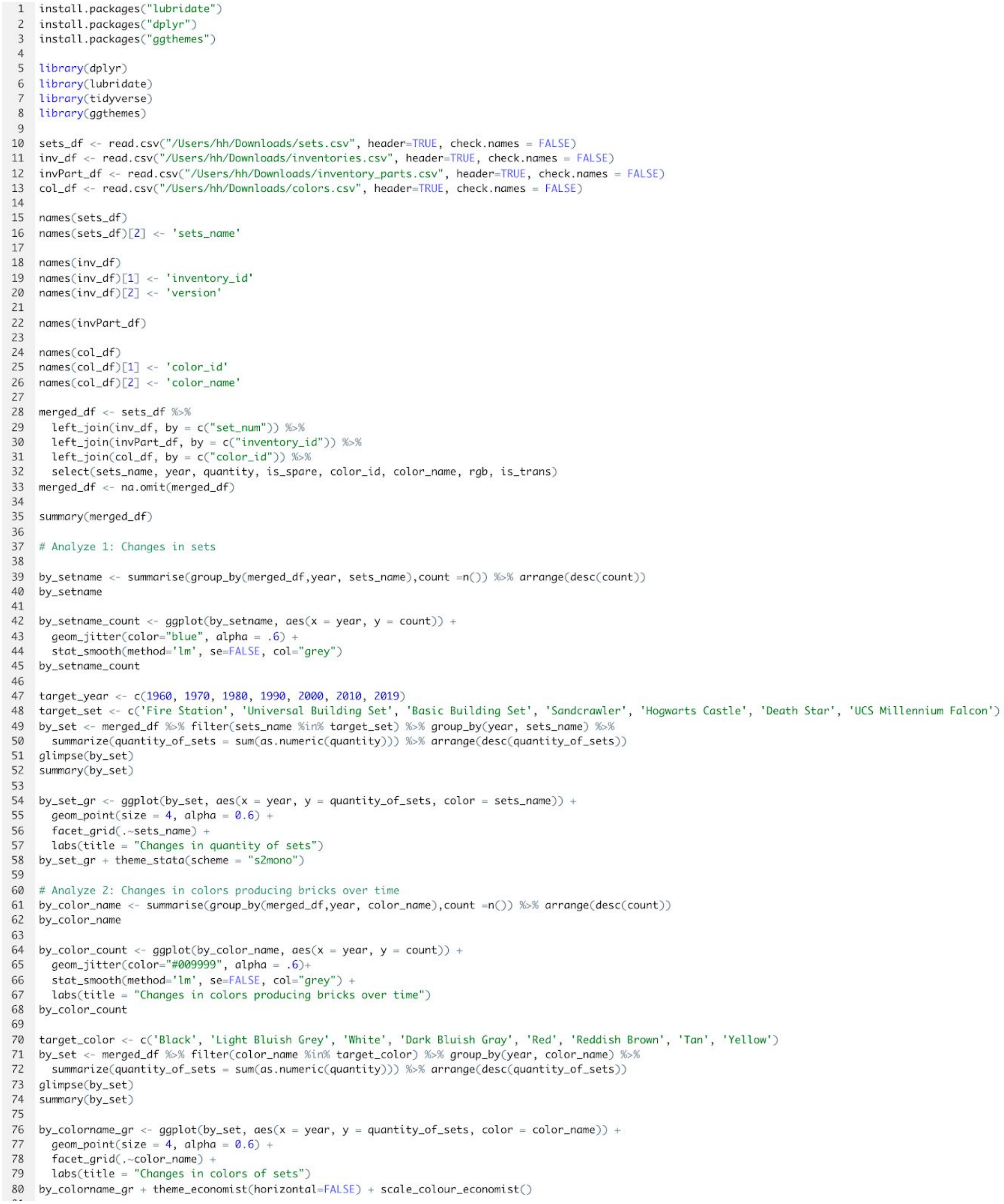


Fig. 10: Facets by the most frequently used colors in LEGO sets.

Reflection

During the analytics, I expected to see more findings to see how LEGO has been changed from plain bricks to colorful and storytelling sets. The goal of the descriptive data analysis is to figure out the fact that how the sets of LEGO have been changed over time and see if there are good implication of it. Even though the color didn’t get any good insights based on my expectation, the chart of changes in the quantity of sets (Fig. 7) indicates that some popular sets - ‘Basic Building Set’ and ‘Universal Building Set’ started to reduce their pieces and didn’t produce them anymore. On the other hand, ‘Hogwarts Castle’ produced many more pieces after and after, it might be its popularity to make a lot of production. The quantity becomes a major index to its survivability. If any LEGO sets produce fewer pieces than the previous one, LEGO is about to think about its halt production.

Appendix



```r
install.packages("lubridate")
install.packages("dplyr")
install.packages("ggthemes")

library(dplyr)
library(lubridate)
library(tidyverse)
library(ggthemes)

sets_df <- read.csv("/Users/hh/Downloads/sets.csv", header=TRUE, check.names = FALSE)
inv_df <- read.csv("/Users/hh/Downloads/inventories.csv", header=TRUE, check.names = FALSE)
invPart_df <- read.csv("/Users/hh/Downloads/inventory_parts.csv", header=TRUE, check.names = FALSE)
col_df <- read.csv("/Users/hh/Downloads/colors.csv", header=TRUE, check.names = FALSE)

names(sets_df)
names(sets_df)[2] <- 'sets_name'

names(inv_df)
names(inv_df)[1] <- 'inventory_id'
names(inv_df)[2] <- 'version'

names(invPart_df)

names(col_df)
names(col_df)[1] <- 'color_id'
names(col_df)[2] <- 'color_name'

merged_df <- sets_df %>%
  left_join(inv_df, by = c("set_num")) %>%
  left_join(invPart_df, by = c("inventory_id")) %>%
  left_join(col_df, by = c("color_id")) %>%
  select(sets_name, year, quantity, is_spare, color_id, color_name, rgb, is_trans)
merged_df <- na.omit(merged_df)

summary(merged_df)

# Analyze 1: Changes in sets

by_setname <- summarise(group_by(merged_df,year, sets_name),count =n()) %>% arrange(desc(count))
by_setname

by_setname_count <- ggplot(by_setname, aes(x = year, y = count)) +
  geom_jitter(color="blue", alpha = .6) +
  stat_smooth(method='lm', se=FALSE, col="grey")
by_setname_count

target_year <- c(1960, 1970, 1980, 1990, 2000, 2010, 2019)
target_set <- c('Fire Station', 'Universal Building Set', 'Basic Building Set', 'Sandcrawler', 'Hogwarts Castle', 'Death Star', 'UCS Millennium Falcon')
by_set <- merged_df %>% filter(sets_name %in% target_set) %>% group_by(year, sets_name) %>%
  summarize(quantity_of_sets = sum(as.numeric(quantity))) %>% arrange(desc(quantity_of_sets))
glimpse(by_set)
summary(by_set)

by_set_gr <- ggplot(by_set, aes(x = year, y = quantity_of_sets, color = sets_name)) +
  geom_point(size = 4, alpha = 0.6) +
  facet_grid(.~sets_name) +
  labs(title = "Changes in quantity of sets")
by_set_gr + theme_stata(scheme = "s2mono")

# Analyze 2: Changes in colors producing bricks over time
by_color_name <- summarise(group_by(merged_df,year, color_name),count =n()) %>% arrange(desc(count))
by_color_name

by_color_count <- ggplot(by_color_name, aes(x = year, y = count)) +
  geom_jitter(color="#009999", alpha = .6)+
  stat_smooth(method='lm', se=FALSE, col="grey") +
  labs(title = "Changes in colors producing bricks over time")
by_color_count

target_color <- c('Black', 'Light Bluish Grey', 'White', 'Dark Bluish Gray', 'Red', 'Reddish Brown', 'Tan', 'Yellow')
by_set <- merged_df %>% filter(color_name %in% target_color) %>% group_by(year, color_name) %>%
  summarize(quantity_of_sets = sum(as.numeric(quantity))) %>% arrange(desc(quantity_of_sets))
glimpse(by_set)
summary(by_set)

by_colorname_gr <- ggplot(by_set, aes(x = year, y = quantity_of_sets, color = color_name)) +
  geom_point(size = 4, alpha = 0.6) +
  facet_grid(.~color_name) +
  labs(title = "Changes in colors of sets")
by_colorname_gr + theme_economist(horizontal=FALSE) + scale_colour_economist()
```

Reference

Rebrickable, Retrieved from https://rebrickable.com/downloads/

Joining Data in R with dplyr, Retrieved from

Primary Key, Retrieved from

https://www.techopedia.com/definition/5547/primary-key

Joining Data in R with dplyr, Retrieved from

https://rstudio-pubs-static.s3.amazonaws.com/293454_556209d0e42141ab8cb7674644445dcd.html