Google Data Analytics Professional Certificate

Capstone Project Análisis de Datos de Sets de LEGO

Francisco Cerna Fukuzaki | fcernaf@gmail.com

16/03/2023

Introducción

El objetivo de este documento es demostrar los conocimientos adquiridos a lo largo de los cursos ofrecidos por la plataforma Coursera para obtener el certificado de Google Data Analytics Professional Certificate. Como proyecto final o Capstone Course se eligió como tema los sets de LEGO. La base de datos fue obtenida de OpenIntro, la cual contiene 1304 observaciones y 14 variables. El proceso de análisis exploratorio de datos (Exploratory Data Analysis, EDA) se realizó a las variables `theme`, `pieces`, `price` y `year`. Como resultado se obtuvo la variación y covariación de las variables y se representó en histogramas, diagrama de barras, scatterplots y boxplots.

Keywords: Google Data Analytics Professional Certificate, Data Analysis Process, R programming language, Capstone, Exploratory Data Analysis (EDA).

Proceso de Análisis de Datos - Data Analysis Process

1. Ask

El trabajo tiene como tema principal realizar un análisis exploratorio de datos a los Set de LEGO. El problema que se está tratando de resolver es la evolución del precio de los Sets a lo largo del tiempo. Mi audiencia es todas aquellas personas que les guste los LEGO.

2. Prepare

En este proyecto final se utilizó como base de datos las ventas de Set de LEGO publicadas en OpenIntro, la cual pertenece a un artículo publicado en Journal of Statistics and Data Science Education y fue utilizada para construir un modelo de Regresión Lineal Múltiple. Así mismo, al pertenecer a un artículo, se puede validar la veracidad e integridad de los datos.

Esta base de datos contiene 1304 observaciones y 14 variables. Se evaluaron cuatro de las trece variables disponibles en el dataset. Las variables consideradas fueron:

- `theme`: Variable cualitativa.
- `pieces`: Variable cuantitativa.
- `price`: Variable continua.
- `year`: Variable discreta.

3. Process

Para el desarrollo, se utilizó el lenguaje de programación R. Por lo que se procedió a declarar las dependencias o paquetes necesarios y visualización de los datos, análisis descriptivo de los datos y limpieza de datos.

Instalación y declaración de paquetes

Se utilizaron paquetes para visualización de datos y análisis descriptivo.

```
install.packages("ggplot2")
install.packages("skimr") # Paquete para análisis descriptivo.
install.packages("dplyr")
```

Luego, se procedió a declarar los paquetes.

```
library(ggplot2)
library(skimr)
library(dplyr)
```

La base de datos se encuentra en un archivo con formato CSV. El contenido fue almacenado en un dataframe.

```
lego_dataset <- read.csv(file="lego_population.csv")
```

Para validar que se almacenó correctamente en el dataframe, se visualizó los primeros registros del dataframe.

```
head(lego_dataset)
```

```
##   item_number                 set_name   theme pieces price amazon_price year
## 1       41916    Extra Dots - Series 2    DOTS    109  3.99         3.44 2020
## 2       41908    Extra Dots - Series 1    DOTS    109  3.99         3.99 2020
## 3       11006     Creative Blue Bricks Classic     52  4.99         4.93 2020
## 4       11007    Creative Green Bricks Classic     60  4.99         4.93 2020
## 5       41901   Funky Animals Bracelet    DOTS     33  4.99         4.99 2020
## 6       41902 Sparkly Unicorn Bracelet    DOTS     33  4.99         4.99 2020
##      ages pages minifigures packaging weight unique_pieces  size
## 1 Ages_6+    NA          NA Foil pack   <NA>             6 Small
## 2 Ages_6+    NA          NA Foil pack   <NA>             6 Small
## 3 Ages_4+    37          NA       Box   <NA>            28 Small
## 4 Ages_4+    37          NA       Box   <NA>            36 Small
## 5 Ages_6+    NA          NA Foil pack   <NA>            10 Small
## 6 Ages_6+    NA          NA Foil pack   <NA>             9 Small
```

Análisis descriptivo de los datos

Luego de que la base de datos fue cargada en un dataframe. Se procedió a realizar el análisis descriptivo.

Como se aprecia en el resultado de la ejecución del código de la sección anterior, el dataframe contiene las trece variables. Por lo que se filtró las variables consideradas para este proyecto final.

```
lego_dataframe <- select(lego_dataset, theme, pieces, price, year)
```

```
skim(lego_dataframe)
```

Table 1: Data summary

Name	lego_dataframe
Number of rows	1304
Number of columns	4
Column type frequency:	
character	1
numeric	3
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
theme	270	0.79	2	21	0	40	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
pieces	42	0.97	389.45	603.50	1.00	79.00	198.00	455.00	6020.00	
price	239	0.82	46.35	58.91	1.99	14.99	29.99	49.99	699.99	
year	0	1.00	2019.00	0.82	2018.00	2018.00	2019.00	2020.00	2020.00	

A partir del resultado, se puede apreciar que las variables `pieces` y `price` tienen valores vacíos.

```
summary(lego_dataframe)
```

```
##     theme               pieces           price            year
##  Length:1304        Min.   :   1.0   Min.   :  1.99   Min.   :2018
##  Class :character   1st Qu.:  79.0   1st Qu.: 14.99   1st Qu.:2018
##  Mode  :character   Median : 198.0   Median : 29.99   Median :2019
##                     Mean   : 389.4   Mean   : 46.35   Mean   :2019
##                     3rd Qu.: 455.0   3rd Qu.: 49.99   3rd Qu.:2020
##                     Max.   :6020.0   Max.   :699.99   Max.   :2020
##                     NA's   :42       NA's   :239
```

Luego de ejecutar las funciones de análisis descriptivo, no se pudo identificar si la variable cualitativa `theme` tiene valores vacíos. Por lo que, se procedió a contabilizar las observaciones o registros vacíos.

```
sum(is.na(lego_dataframe$theme))
```

```
## [1] 270
```

Con esto, se pudo identificar que existen 270 registros vacíos para la variable `theme`.

Limpieza de datos

Se procedió a limpiar los registros donde al menos una variable sea vacía en el dataframe.

```
lego_dataframe_limpio <- na.omit(lego_dataframe)
```

```
skim(lego_dataframe_limpio)
```

Table 4: Data summary

Name	lego_dataframe_limpio
Number of rows	1003
Number of columns	4
_____	
Column type frequency:	
character	1
numeric	3
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
theme	0	1	2	21	0	40	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
pieces	0	1	441.76	653.98	1.00	101.00	229.00	500.50	6020.00	
price	0	1	47.82	60.16	3.49	14.99	29.99	59.99	699.99	
year	0	1	2019.03	0.82	2018.00	2018.00	2019.00	2020.00	2020.00	

```
summary(lego_dataframe_limpio)
```

```
##     theme               pieces           price             year
##  Length:1003        Min.   :   1.0   Min.   :  3.49   Min.   :2018
##  Class :character   1st Qu.: 101.0   1st Qu.: 14.99   1st Qu.:2018
##  Mode  :character   Median : 229.0   Median : 29.99   Median :2019
##                     Mean   : 441.8   Mean   : 47.82   Mean   :2019
##                     3rd Qu.: 500.5   3rd Qu.: 59.99   3rd Qu.:2020
##                     Max.   :6020.0   Max.   :699.99   Max.   :2020
```

La cantidad de observaciones fue reducida de 1304 a 1003 registros.

> En RStudio Cloud se puede visualizar los gráficos sin problemas. Sin embargo, al exportar los resultados en un archivo con extensión PDF no se pueden apreciar los gráficos debido a que la variable categórica `theme` tiene caracteres extraños debido a los símbolos de marcas registradas, como TM o R. Por lo que, se retira los caracteres extraños.

```
lego_dataframe_limpio$theme <- gsub(" ", "", lego_dataframe_limpio$theme)
```

4. Analyze

Para la etapa de Análisis, se procedió a revisar los valores únicos, varianza y covarianza de las variables.

Valores únicos

Se obtuvo los valores únicos de la variable cualitativa `theme`. Esta variable contiene los nombres de los Sets de LEGO que fueron vendidos.

```
length(unique(lego_dataframe_limpio$theme))
```

```
## [1] 40
```

Existen 40 nombres de Set de LEGO.

```
unique(lego_dataframe_limpio$theme)
```

```
##  [1] "DOTS"                "Classic"             "DUPLO®"
##  [4] "Friends"             "Disney"              "City"
##  [7] "Unikitty!"           "NINJAGO®"            "Star Wars"
## [10] "Minecraft"           "Marvel"              "Creator 3-in-1"
## [13] "Batman"              "THE LEGO® MOVIE 2"   "Technic"
## [16] "Speed Champions"     "BrickHeadz"          "LEGO® Frozen 2"
## [19] "LEGO® Super Mario"   "Harry Potter"        "Hidden Side"
```

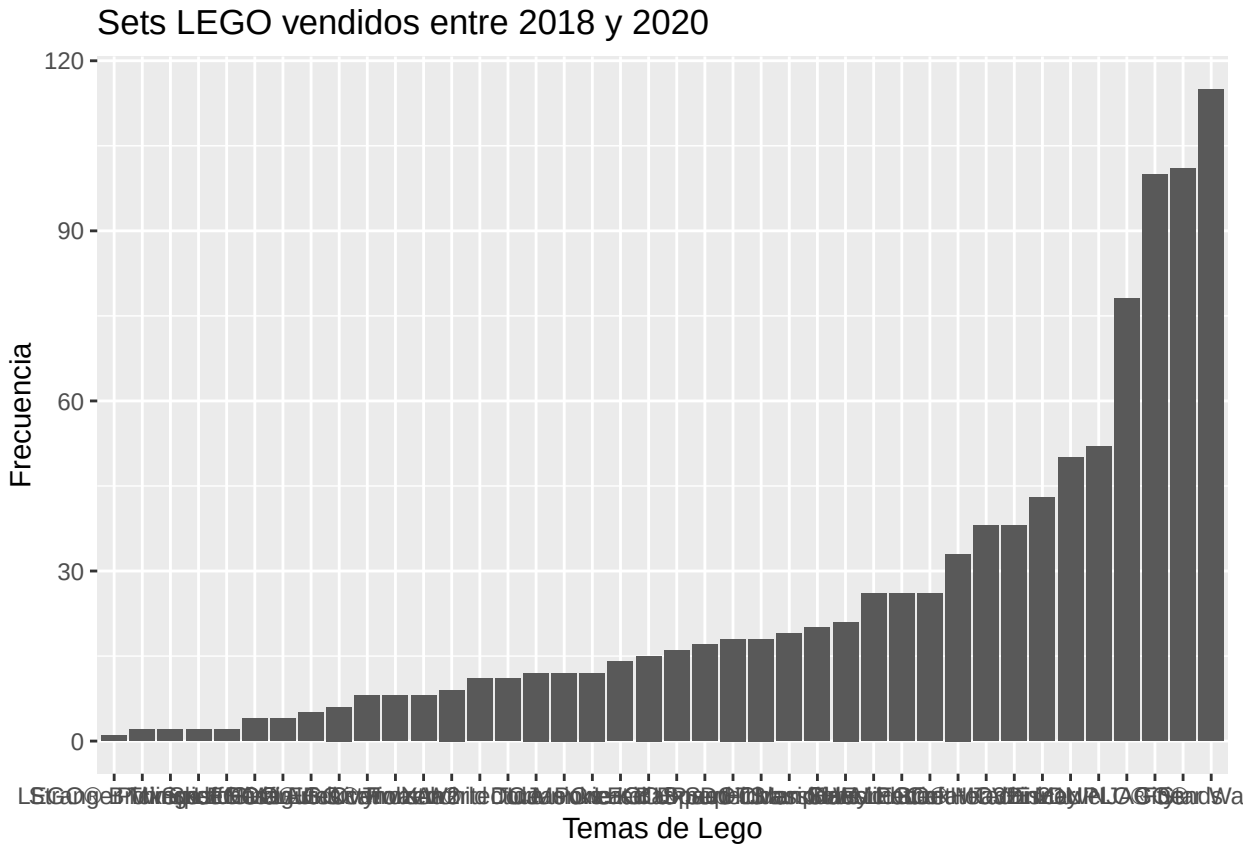
```
## [22] "Trolls World Tour"    "Minions"              "Powerpuff Girls"
## [25] "Jurassic World"       "Overwatch®"           "Spider-Man"
## [28] "Juniors"              "DC"                   "Architecture"
## [31] "Ideas"                "Creator Expert"       "LEGO® Art"
## [34] "Powered UP"           "Stranger Things"      "Monkie Kid"
## [37] "Xtra"                 "Minifigures"          "LEGO® Brick Sketches"
## [40] "LEGO® Education"
```

Distribución de datos

La distribución de los datos permitirá analizar mejor las variables.

Primero se procede a analizar la variable `theme` en un diagrama de barras.

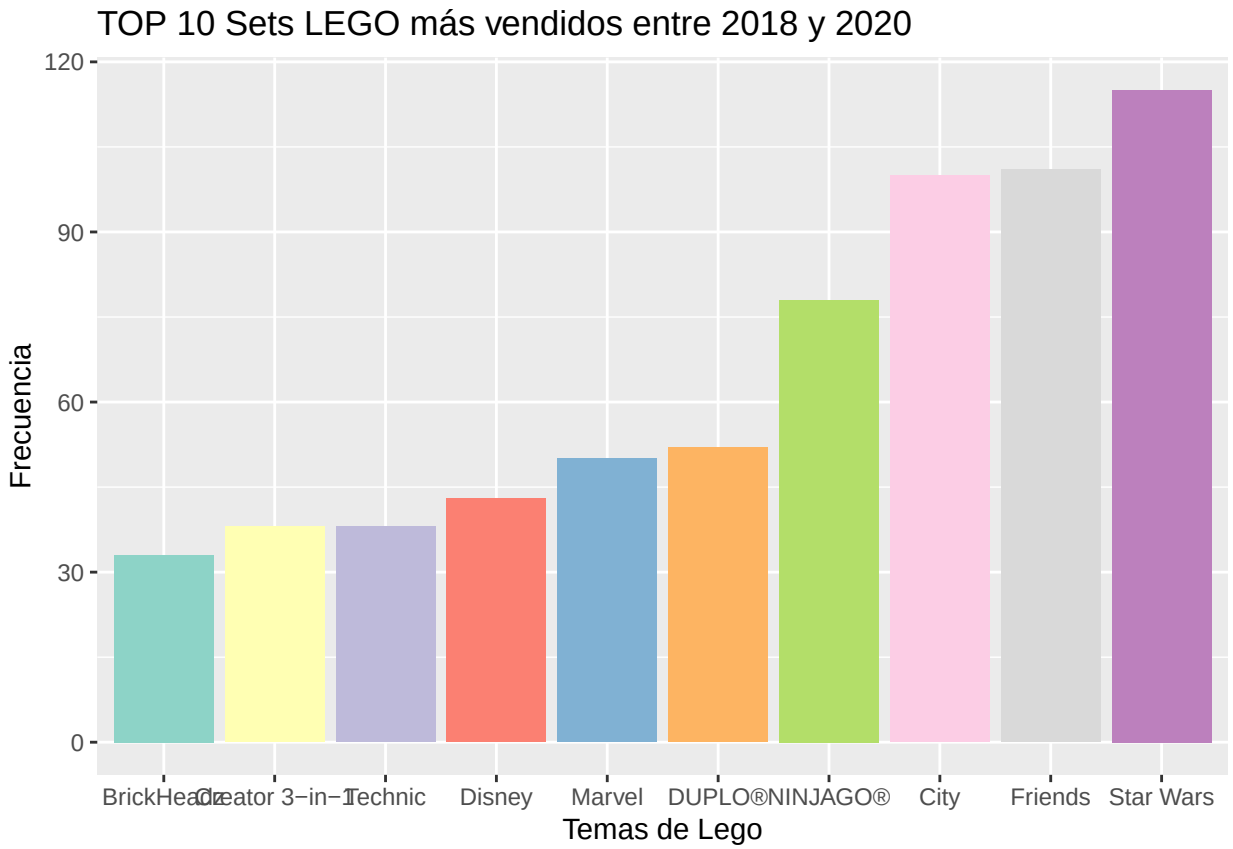
```
lego_dataframe_limpio %>%
  count(theme) %>%
  arrange(desc(n)) %>%
  mutate(theme = reorder(theme, n)) %>%
  ggplot(aes(x = theme, y = n)) +
  geom_bar(stat = "identity") +
  labs(x = "Temas de Lego", y = "Frecuencia", title = "Sets LEGO vendidos entre 2018 y 2020")
```



Para facilidad visual de los nombres de Sets de LEGO, se procedió a visualizar los 10 primeros Sets.

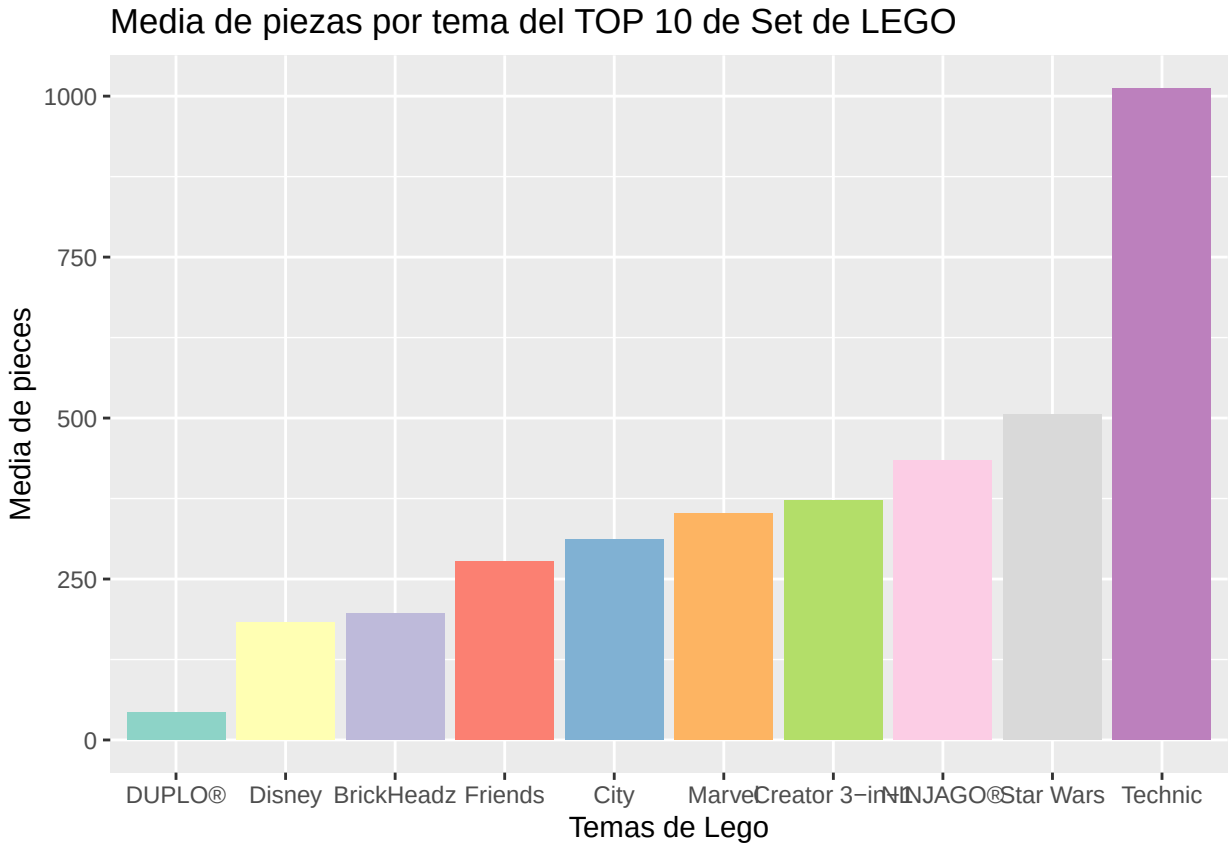
```
lego_dataframe_limpio %>%
  count(theme) %>%
  arrange(desc(n)) %>%
  head(10) %>%
  mutate(theme = reorder(theme, n)) %>%
```

```
  ggplot(aes(x = theme, y = n, fill = as.factor(theme))) +
  scale_fill_brewer(palette="Set3") +
  geom_bar(stat = "identity") +
  labs(x = "Temas de Lego", y = "Frecuencia", title = "TOP 10 Sets LEGO más vendidos entre 2018 y 2020")
  theme(legend.position = "none")
```



Se obtiene la media de la cantidad de piezas por el top 10 de los temas de Set de LEGO.

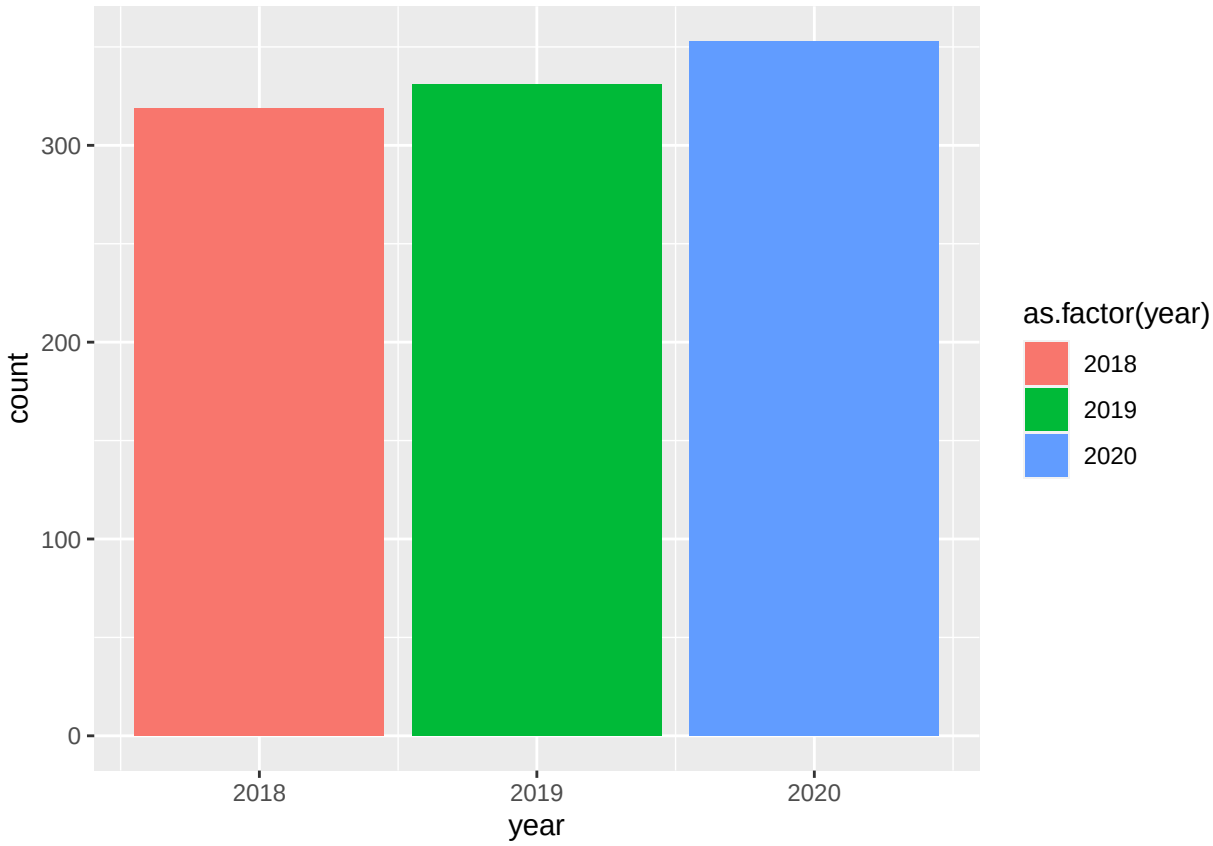
```
top10_themes <- lego_dataframe_limpio %>%
  count(theme) %>%
  arrange(desc(n)) %>%
  head(10) %>%
  pull(theme)

media_por_theme <- lego_dataframe_limpio %>%
  filter(theme %in% top10_themes) %>%
  group_by(theme) %>%
  summarize(media_pieces = mean(pieces))

ggplot(media_por_theme, aes(x = reorder(theme, media_pieces), y = media_pieces, fill = as.factor(reorde
  scale_fill_brewer(palette="Set3") +
  geom_bar(stat = "identity") +
  labs(x = "Temas de Lego", y = "Media de pieces", title = "Media de piezas por tema del TOP 10 de Set
  theme(legend.position = "none")
```



Del gráfico de barras se puede concluir que a pesar de que el tema Star Wars es el que tiene mayor cantidad de ventas, no es el tema con mayor cantidad de piezas. Sino que es Technic el que tiene mayor cantidad de piezas a pesar de estar en el octavo puesto del TOP de temas más vendidos.

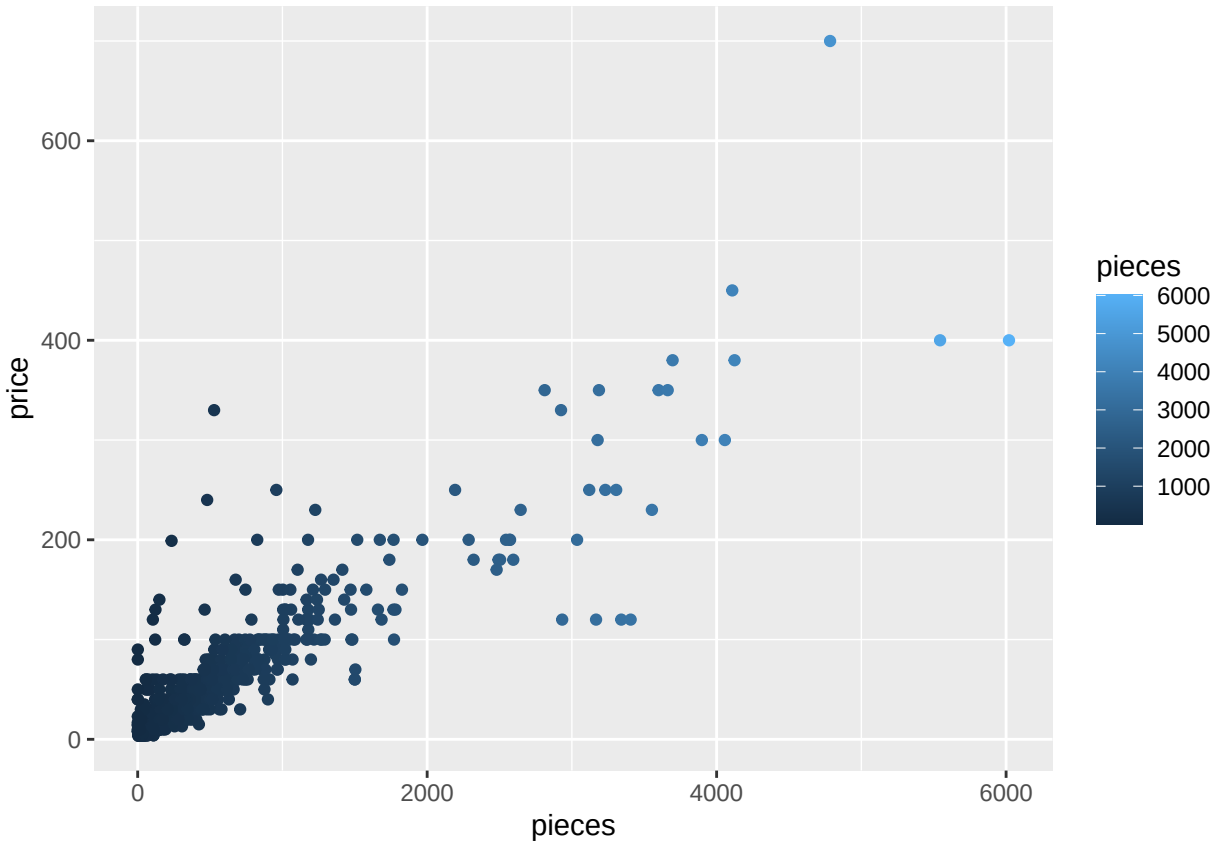
Para tener un detalle de la cantidad de temas vendidos se agregará la variable `year`.

```
ggplot(data = lego_dataframe_limpio) +
  geom_bar(mapping = aes(x = year, fill=as.factor(year)))
```



A partir del gráfico de barras anterior, se puede concluir que existe una relación lineal entre las ventas de Sets de LEGO y los años.

```
ggplot(data = lego_dataframe_limpio, mapping = aes(x = pieces, y = price, color=pieces)) +
  geom_point()
```



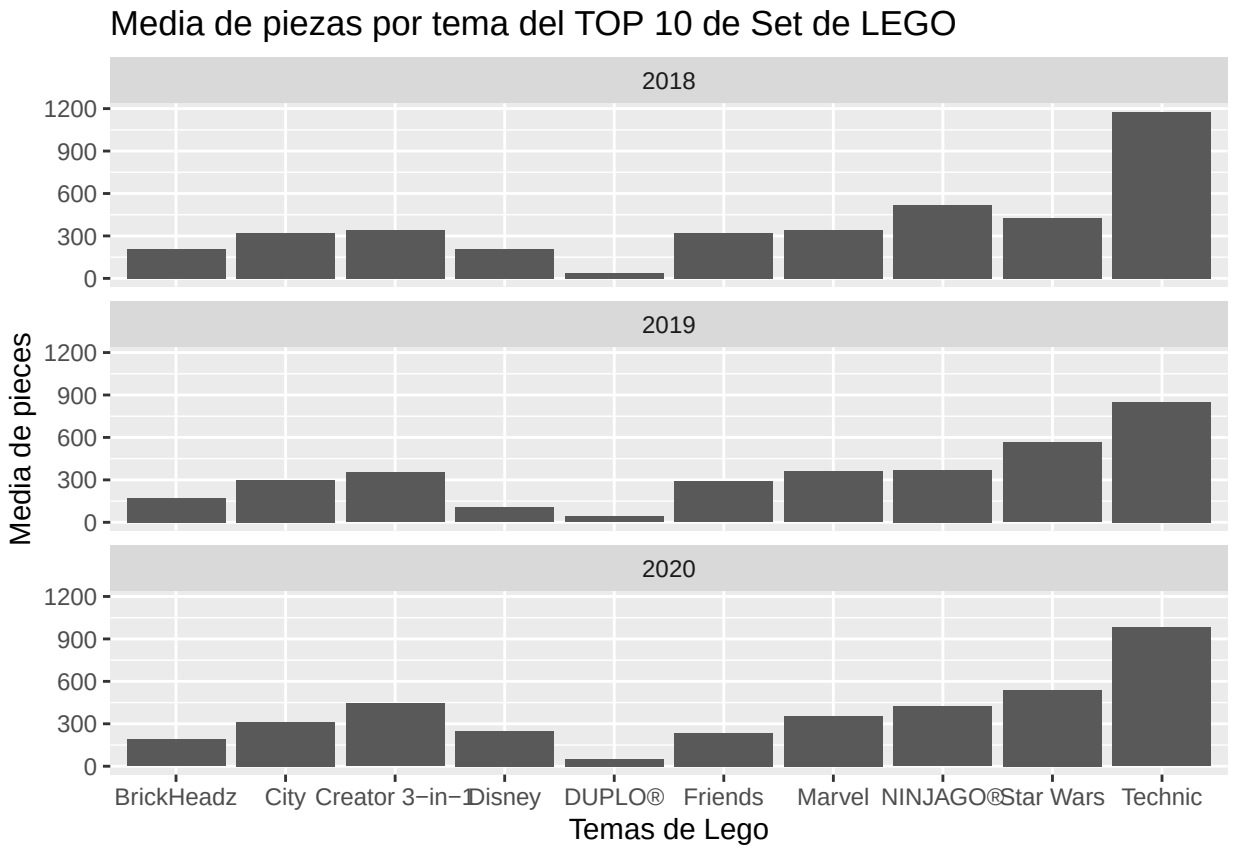
Se puede observar que existe una relación entre el Set tenga más piezas, el precio aumentará. Al mismo tiempo, se aprecia que la mayor cantidad de ventas se ubica en los Sets con precios menores a USD 100.00.

A partir del TOP 10 de temas, se obtendrá la cantidad de ventas por año.

```
# Seleccionar el top 10 de themes
lego_dataframe_limpio_top10_themes <- lego_dataframe_limpio %>%
  count(theme) %>%
  arrange(desc(n)) %>%
  slice(1:10) %>%
  select(theme)

# Filtrar los datos para el top 10 de themes y calcular la media de pieces por año
media_por_theme_year <- lego_dataframe_limpio %>%
  filter(theme %in% lego_dataframe_limpio_top10_themes$theme) %>%
  group_by(theme, year) %>%
  summarise(media_pieces = mean(pieces)) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'theme'. You can override using the
## `.groups` argument.
```

```
# Generar el gráfico con facet_wrap
ggplot(media_por_theme_year, aes(x = theme, y = media_pieces)) +
  geom_bar(stat = "identity") +
  labs(x = "Temas de Lego", y = "Media de pieces", title = "Media de piezas por tema del TOP 10 de Set
  facet_wrap(~ year, nrow = 3)
```

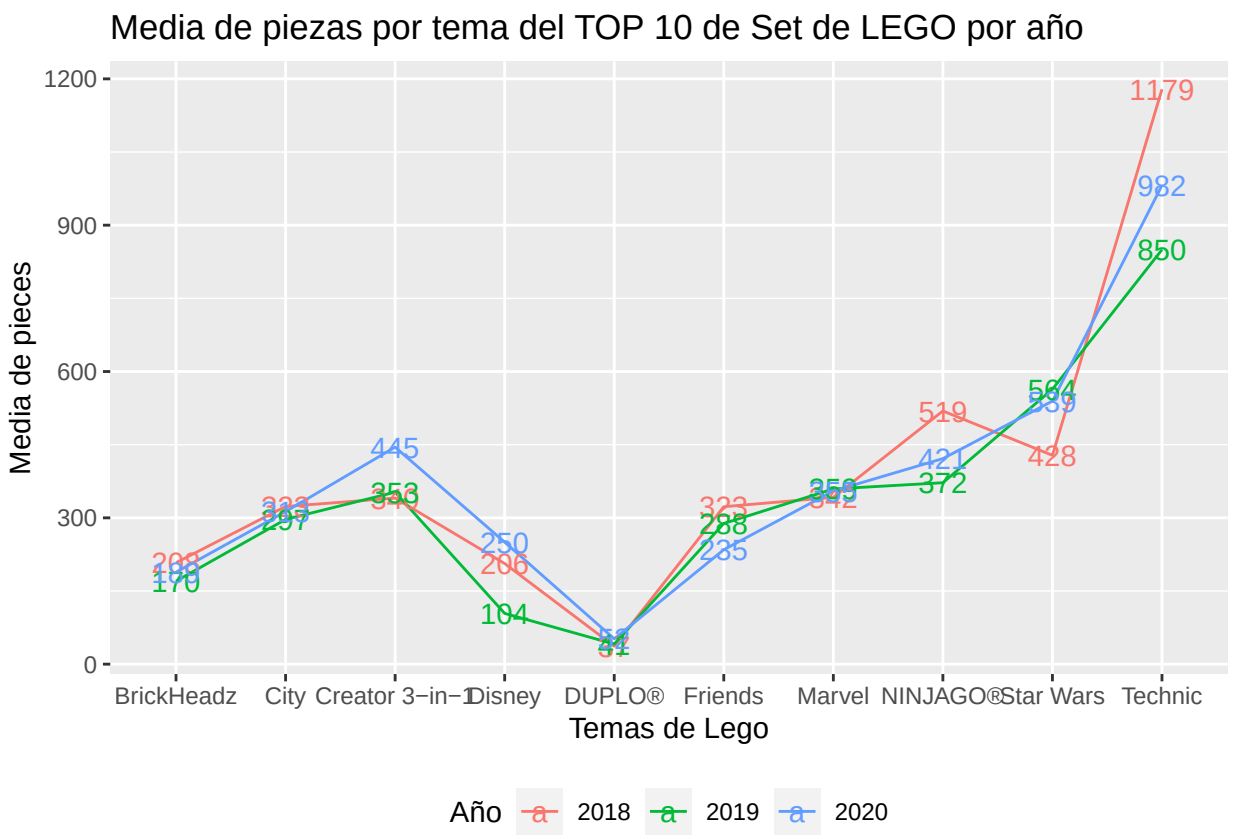


Los datos mostrados anteriormente se colocan en un gráfico de líneas para poder apreciar mejor la evolución del promedio de piezas por tema por año.

```
# Calcular la media de la variable pieces por cada valor único de la variable theme y año
media_por_theme_year <- lego_dataframe_limpio %>%
  filter(theme %in% top10_themes) %>%
  group_by(theme, year) %>%
  summarize(media_pieces = mean(pieces))
```

```
## `summarise()` has grouped output by 'theme'. You can override using the
## `.groups` argument.
```

```
# Crear un gráfico de línea
ggplot(media_por_theme_year, aes(x = theme, y = media_pieces, group = year, color = as.factor(year))) +
  geom_line() +
  geom_text(aes(label=sprintf("%.f", media_pieces))) +
  labs(x = "Temas de Lego", y = "Media de pieces", title = "Media de piezas por tema del TOP 10 de Set d
  scale_color_discrete(name = "Año") +
  theme(legend.position = "bottom")
```

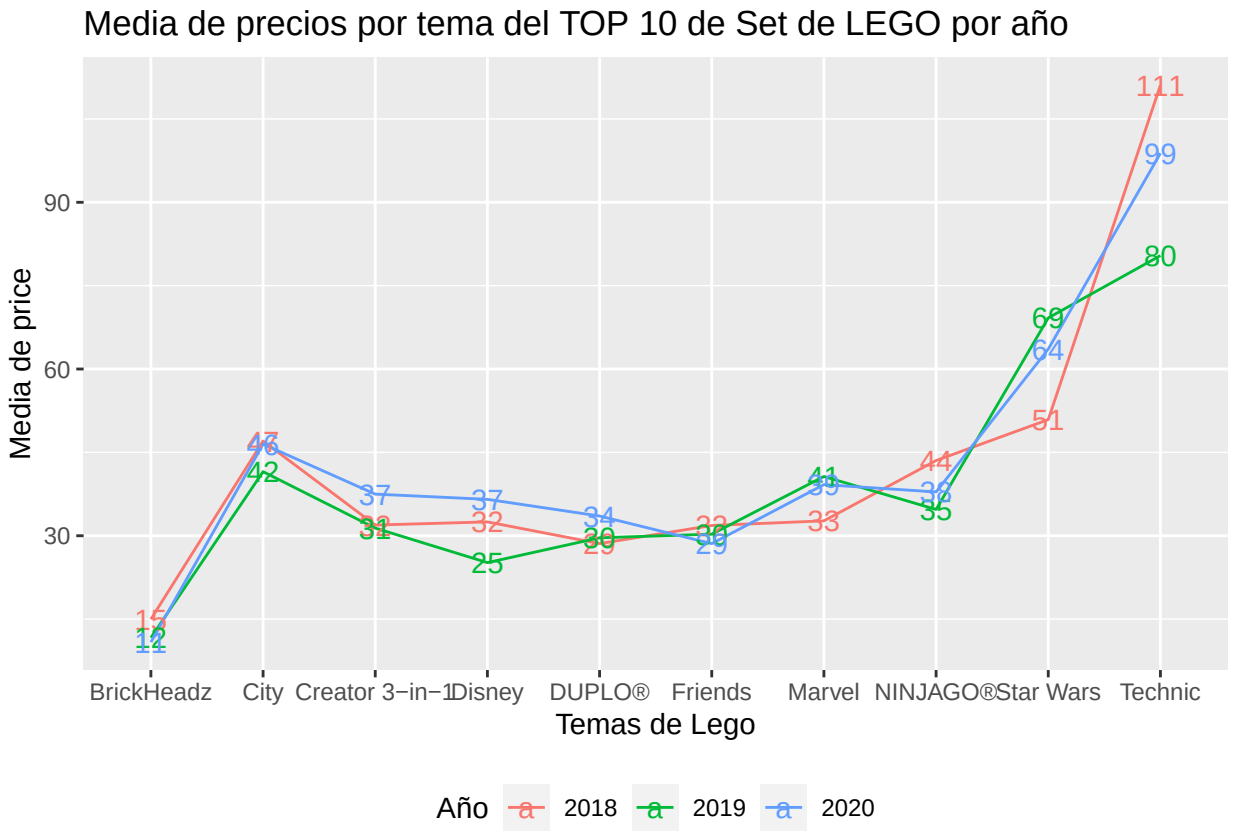


Se realiza el mismo gráfico para el caso de promedio de precio.

```
# Calcular la media de la variable price por cada valor único de la variable theme y año
media_por_theme_year <- lego_dataframe_limpio %>%
  filter(theme %in% top10_themes) %>%
  group_by(theme, year) %>%
  summarize(media_prices = mean(price))
```

```
## `summarise()` has grouped output by 'theme'. You can override using the
## `.groups` argument.
```

```
# Crear un gráfico de línea
ggplot(media_por_theme_year, aes(x = theme, y = media_prices, group = year, color = as.factor(year))) +
  geom_line() +
  geom_text(aes(label=sprintf("%.f", media_prices))) +
  labs(x = "Temas de Lego", y = "Media de price", title = "Media de precios por tema del TOP 10 de Set
  scale_color_discrete(name = "Año") +
  theme(legend.position = "bottom")
```



A partir del gráfico se puede ver que las líneas por cada año son parecidas tanto para el promedio de piezas como el promedio de precio.

Finalmente, se hace una comparación entre las variables `pieces` y `price` por `year`.

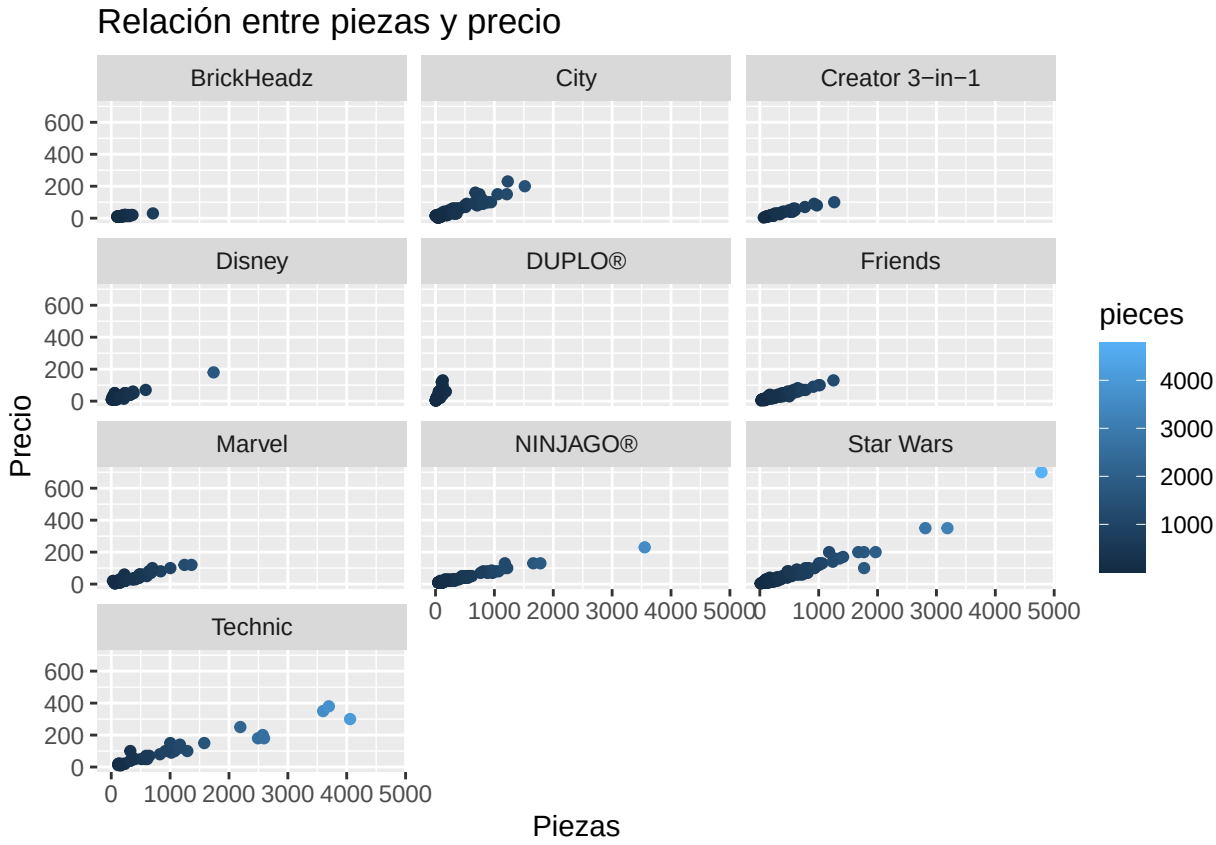
```
lego_dataframe_limpio %>%
  filter(theme %in% top10_themes) %>%
  group_by(year, theme) %>%
  summarize(media_price = mean(price), media_pieces = mean(pieces))
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 30 x 4
## # Groups:   year [3]
##     year theme          media_price media_pieces
##    <int> <chr>                <dbl>        <dbl>
##  1  2018 BrickHeadz            15.0         208.
##  2  2018 City                  47.0         323.
##  3  2018 Creator 3-in-1        31.9         340.
##  4  2018 DUPLO®                28.6          36.8
##  5  2018 Disney                32.5         206.
##  6  2018 Friends               31.8         323.
##  7  2018 Marvel                32.7         342.
##  8  2018 NINJAGO®              43.5         519.
##  9  2018 Star Wars             50.9         428.
## 10  2018 Technic              111.         1179.
## # ... with 20 more rows
```

Se realiza un gráfico de dispersión para las variables `pieces` y `price` por cada `theme`.

```
lego_dataframe_limpio %>%
  filter(theme %in% top10_themes) %>%
  group_by(year, theme) %>%
  ggplot(aes(x = pieces, y = price, color=pieces)) +
  geom_point() +
  facet_wrap(~theme, ncol=3) +
  labs(title = "Relación entre piezas y precio",
       x = "Piezas",
       y = "Precio") +
  xlab("Piezas") +
  ylab("Precio")
```



A partir del gráfico anterior, se aprecia que no todos los temas tienen Sets con gran cantidad de piezas. Los Sets que aparecen con una cantidad superior a las 2000 piezas son los temas NINJAGO, Star Wars y Technic.

Continuando con el análisis, se procedió a revisar del TOP 10 de los temas con mayor cantidad de piezas promedio con el precio promedio por cada año disponible en la base de datos. Para ello, me apoyé de un gráfico de dispersión para los años 2018, 2019 y 2020.

```
# Agrupar datos por tema
temas <- lego_dataframe_limpio %>%
  filter(year == 2018) %>%
  group_by(theme) %>%
  summarize(media_price = mean(price), media_pieces = mean(pieces))

# Seleccionar los 10 temas más comunes
top10_temas <- temas %>%
  top_n(10, media_pieces)
```

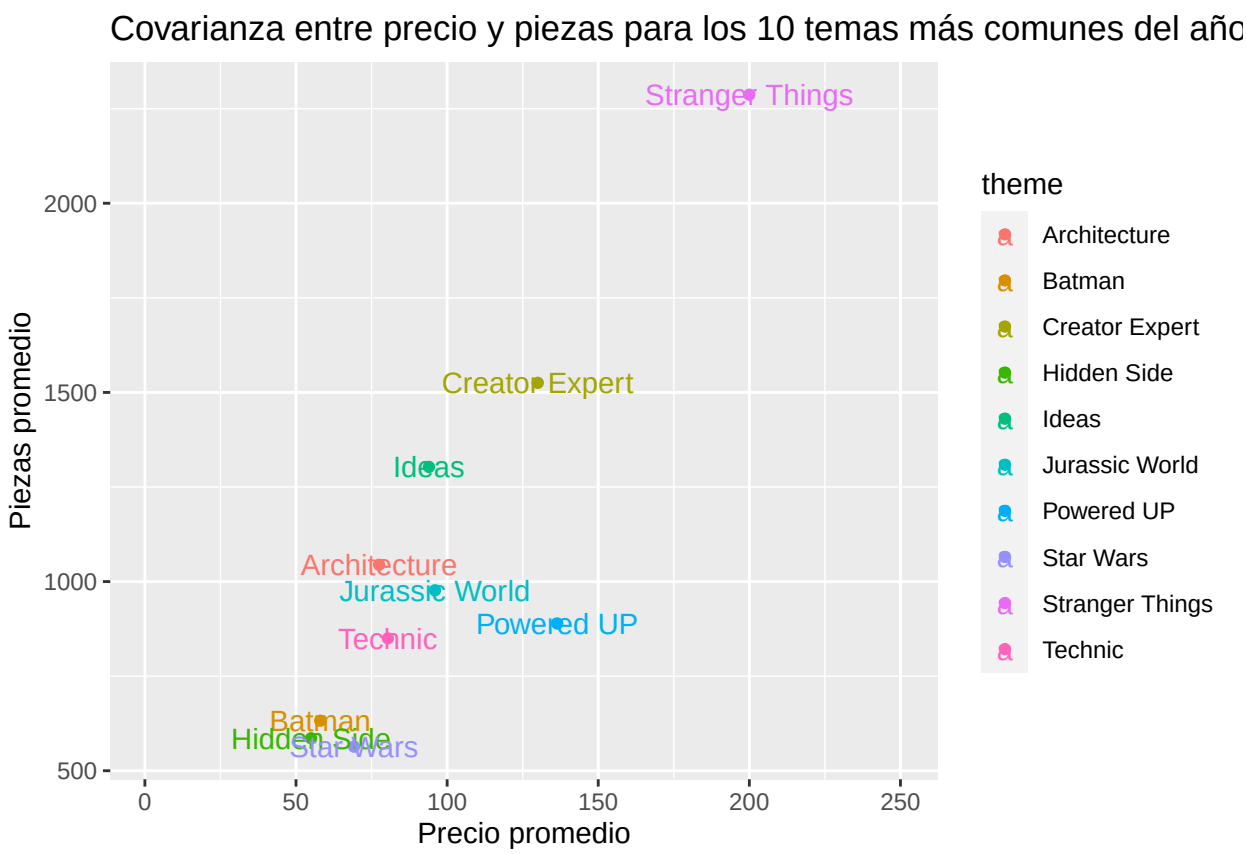
```
# Gráfico de dispersión con tema en el eje X
ggplot(top10_temas, aes(x = media_price, y = media_pieces, color = theme)) +
  geom_point() +
  geom_text(aes(label=theme)) +
  labs(x = "Precio promedio", y = "Piezas promedio",
       title = "Covarianza entre precio y piezas para los 10 temas más comunes del año 2018") +
  xlim(1, 250)
```

Como resultado para el 2018, se identificó que los tres Sets (NINJAGO, Star Wars y Technic) con cantidad de piezas promedio y precio promedio se encuentran aproximadamente entre los USD 50 y el Technic con un precio promedio superior al USD 100. Por lo que, se identifica que este tema es más caro en comparación de los otros dos.

Además, aparece el tema Creator Expert con un precio promedio y cantidad de piezas superior al resto de temas.

```
# Agrupar datos por tema
temas <- lego_dataframe_limpio %>%
  filter(year == 2019) %>%
  group_by(theme) %>%
  summarize(media_price = mean(price), media_pieces = mean(pieces))

# Seleccionar los 10 temas más comunes
top10_temas <- temas %>%
  top_n(10, media_pieces)

# Gráfico de dispersión con tema en el eje X
```

```
ggplot(top10_temas, aes(x = media_price, y = media_pieces, color = theme)) +
  geom_point() +
  geom_text(aes(label=theme)) +
  labs(x = "Precio promedio", y = "Piezas promedio",
       title = "Covarianza entre precio y piezas para los 10 temas más comunes del año 2019") +
  xlim(1, 250)
```

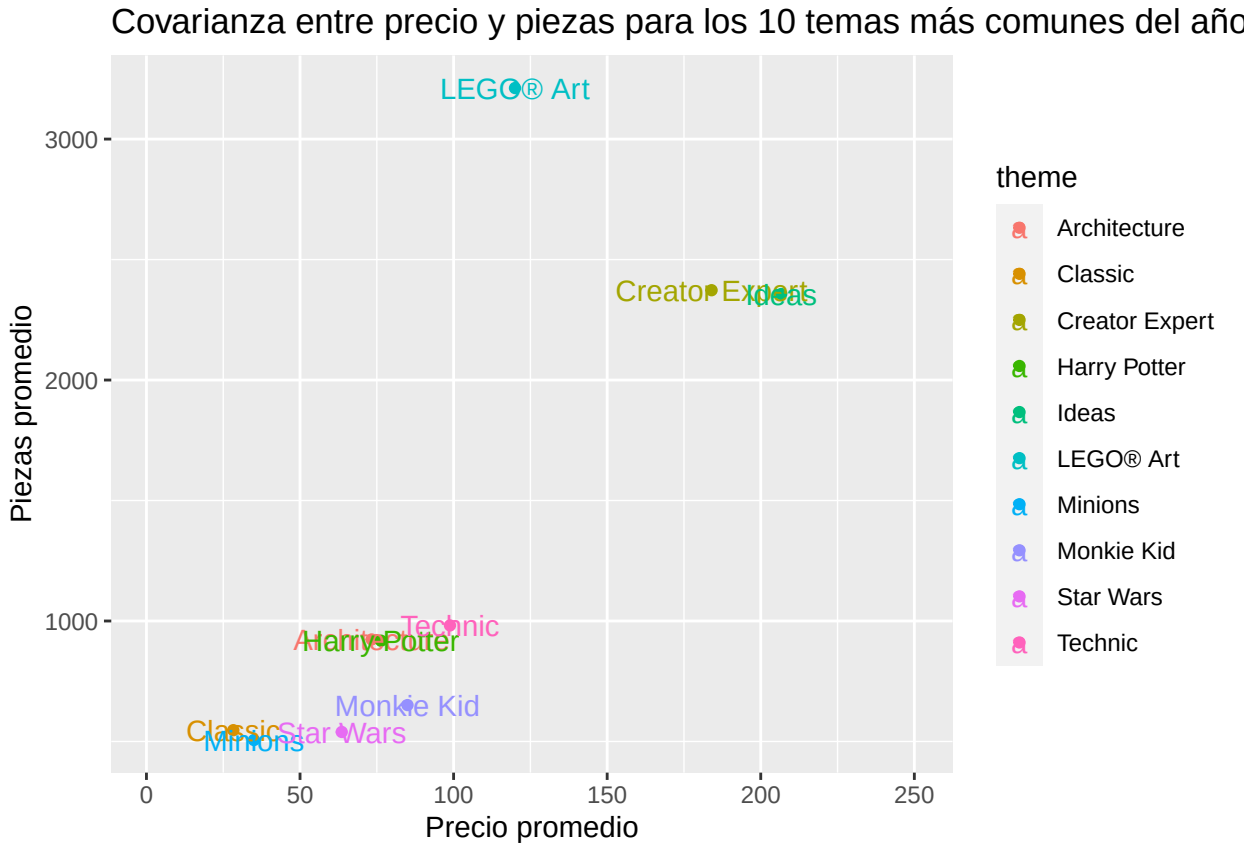


Para el año 2019, el precio promedio de Creator Expert bajó y el precio promedio del tema Stranger Things se encuentra bordeando los USD 200.

Respecto a los temas Star Wars y Technic siguen apareciendo en el gráfico. Sin embargo, ya no aparece NINJAGO. Esto no quiere decir que las ventas de NINJAGO hayan desaparecido, simplemente no aparecen en el TOP 10 de piezas promedio de ventas de Set de LEGO del 2019.

```
# Agrupar datos por tema
temas <- lego_dataframe_limpio %>%
  filter(year == 2020) %>%
  group_by(theme) %>%
  summarize(media_price = mean(price), media_pieces = mean(pieces))

# Seleccionar los 10 temas más comunes
top10_temas <- temas %>%
  top_n(10, media_pieces)

# Gráfico de dispersión con tema en el eje X
ggplot(top10_temas, aes(x = media_price, y = media_pieces, color = theme)) +
  geom_point() +
  geom_text(aes(label=theme)) +
```

```
  labs(x = "Precio promedio", y = "Piezas promedio",
       title = "Covarianza entre precio y piezas para los 10 temas más comunes del año 2020") +
  xlim(1, 250)
```



Para el 2020, el tema LEGO Art con un precio promedio aproximado de USD 125 ofrece una gran cantidad de piezas promedio en comparación del resto de temas.

Además, se aprecia que los temas Creator Expert e Ideas tienen un precio promedio mayor a USD 175. Y se aprecia que el tema Ideas supera el precio promedio de Creator Expert. Ambos tienen una cantidad de piezas promedio similar.

Respecto a los temas Star Wars y Technic, se mantienen con un precio promedio menor a USD 100.

Conclusiones

Al finalizar el análisis, he llegado a la conclusión que las ventas por temas son muy similar a lo largo de los tres años considerados dentro del proyecto. Existe poca variación de ventas entre los años 2018, 2019 y 2020. Esto se pudo visualizar en el gráfico de líneas con el TOP 10 de temas más vendidos.

Así mismo, para las personas que inviertan en la compra de un set de LEGO, los temas de Creator Expert tienen un precio promedio alto, pero al mismo tiempo una cantidad promedio de piezas. Pero si se desea adquirir un set con un precio promedio menor y gran cantidad de piezas la mejor opción sería LEGO Art.

Para aquellas personas que son coleccionistas, los temas Star Wars, NINJAGO y Technic lideran las ventas entre los años 2018 y 2020. Por lo que, se puede deducir que se mantendrán estos temas y la posibilidad que puedan incrementar su colección.

5. Share y Act

Los resultados obtenidos estarán presentados en un archivo con formato PDF generado por RStudio Cloud.

El archivo PDF y el código fuente (Rmd) serán compartidos en mi cuenta de GitHub fcernafukuzaki.

- Link al archivo con los resultados PDF
- Link al código fuente

Referencias

- Anna D. Peterson & Laura Ziegler (2021) Building a Multiple Linear Regression Model With LEGO Brick Data, Journal of Statistics and Data Science Education, 29:3, 297-303, DOI: 10.1080/26939169.2021.1946450
- Data Sets. (n.d.). https://www.openintro.org/data/index.php?data=lego_population
- 7 Análisis exploratorio de datos (EDA). (n.d.). https://es.r4ds.hadley.nz/an%C3%A1lisis-exploratorio-de-datos-eda.html
- Google Data Analytics Professional Certificate. (2023). Coursera. https://www.coursera.org/professional-certificates/google-data-analytics