Media

GDP (USD Billion)	8.750000
Population (Millions)	0.731000
Unemployment rate (%)	13.833333
Average age	29.233333
Women (%)	51.500000
Men (%)	48.500000
Budget (USD Billion)	1.650000

Mediana

GDP (USD Billion)	2.65
Population (Millions)	0.39
Unemployment rate (%)	13.45
Average age	29.00
Women (%)	51.00
Men (%)	49.00
Budget (USD Billion)	0.60

Desviación Estándar

GDP (USD Billion)	19.914433
Population (Millions)	1.352832
Unemployment rate (%)	2.945052
Average age	2.238893
Women (%)	0.776819
Men (%)	0.776819
Budget (USD Billion)	3.451187

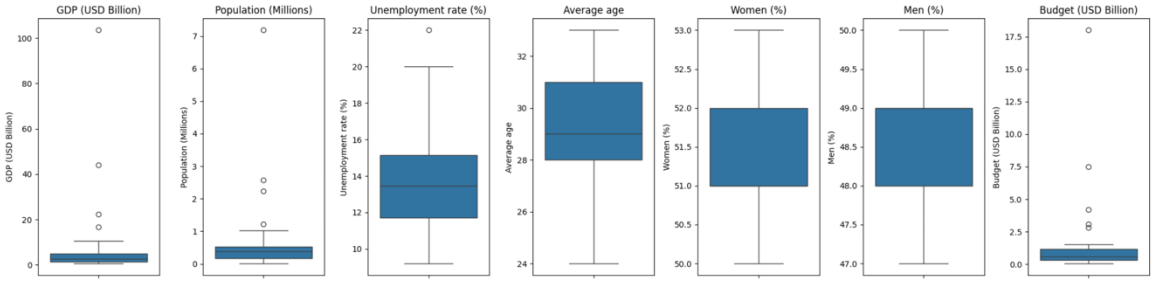
Moda

GDP (USD Billion)	0.60
Population (Millions)	0.01
Unemployment rate (%)	9.20
Average age	29.00
Women (%)	51.00
Men (%)	49.00
Budget (USD Billion)	0.10

Repeticiones

```
GDP (USD Billion)          1.0
Population (Millions)      2.0
Unemployment rate (%)      1.0
Average age                6.0
Women (%)                 14.0
Men (%)                   14.0
Budget (USD Billion)       3.0
```

1.2. Dibujar un boxplot a mano. Utilizando los datos de la tabla 1 y las siguientes proporciones



1.3 Matriz de Covarianza

	GDP (USD Billion)	Population (Millions)	Unemployment rate (%)	Average age	Women (%)	Men (%)	Budget (USD Billion)
GDP (USD Billion)	396.584655	26.685224	-19.123103	15.087931	4.243103	-4.243103	68.613448
Population (Millions)	26.685224	1.830154	-1.342793	1.104931	0.279138	-0.279138	4.639603
Unemployment rate (%)	-19.123103	-1.342793	8.673333	-5.232184	-0.993103	0.993103	-3.481379
Average age	15.087931	1.104931	-5.232184	5.012644	1.224138	-1.224138	2.762069
Women (%)	4.243103	0.279138	-0.993103	1.224138	0.603448	-0.603448	0.734483
Men (%)	-4.243103	-0.279138	0.993103	-1.224138	-0.603448	0.603448	-0.734483
Budget (USD Billion)	68.613448	4.639603	-3.481379	2.762069	0.734483	-0.734483	11.910690

1.4 Matriz de Correlación

	GDP (USD Billion)	Population (Millions)	Unemployment rate (%)	Average age	Women (%)	Men (%)	Budget (USD Billion)
GDP (USD Billion)	1.000000	0.990510	-0.326060	0.338398	0.274281	-0.274281	0.998327
Population (Millions)	0.990510	1.000000	-0.337033	0.364803	0.265616	-0.265616	0.993730
Unemployment rate (%)	-0.326060	-0.337033	1.000000	-0.793518	-0.434092	0.434092	-0.342523
Average age	0.338398	0.364803	-0.793518	1.000000	0.703845	-0.703845	0.357464
Women (%)	0.274281	0.265616	-0.434092	0.703845	1.000000	-1.000000	0.273964
Men (%)	-0.274281	-0.265616	0.434092	-0.703845	-1.000000	1.000000	-0.273964
Budget (USD Billion)	0.998327	0.993730	-0.342523	0.357464	0.273964	-0.273964	1.000000

1.5 Explique la relación entre covarianza y correlación

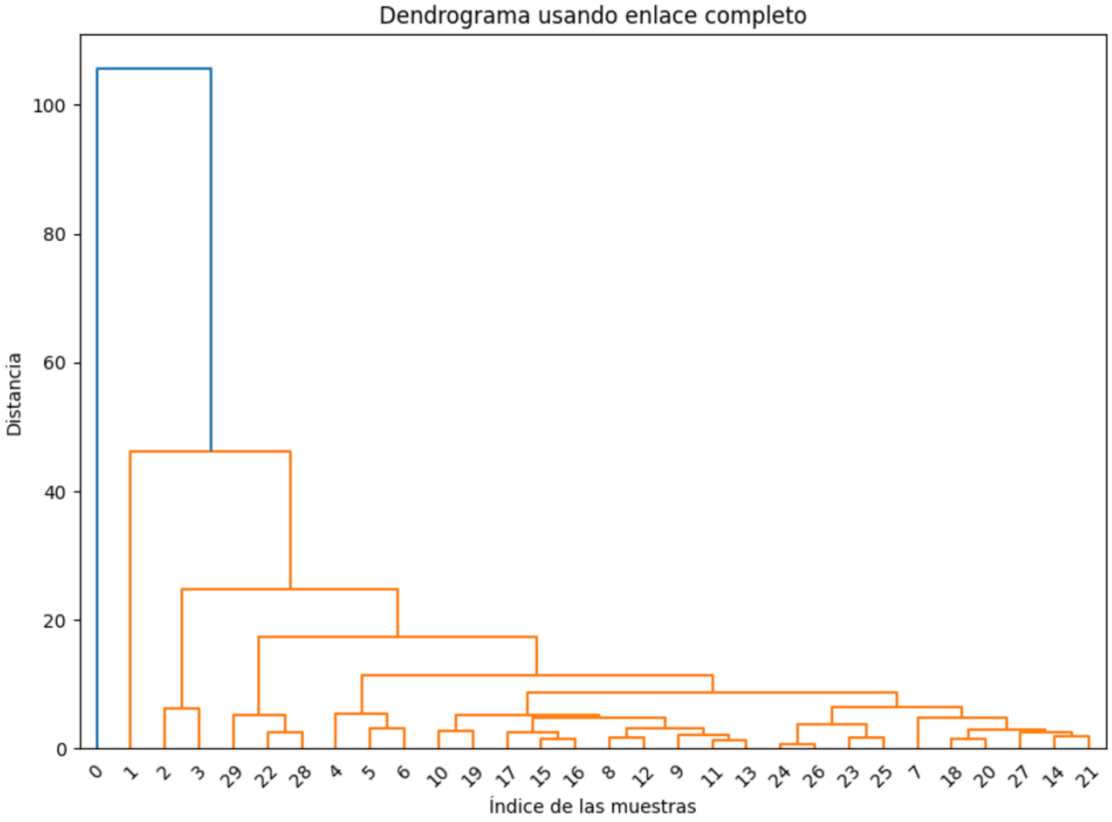
Se entiende que la correlación implementa la covarianza, pero dan valores estadisticos diferentes, en primera parte la covarianza, identifica cuando varía las variables, esto tiene

valores como positivo y negativo, que nos dicen que los valores crecen o decrecen, o sea, las variables no están normalizadas. En el caso de la correlación nos ayuda a detectar dirección y fuerza, de manera lineal. Esta compara dos variables, y están entre valores de -1 y 1.

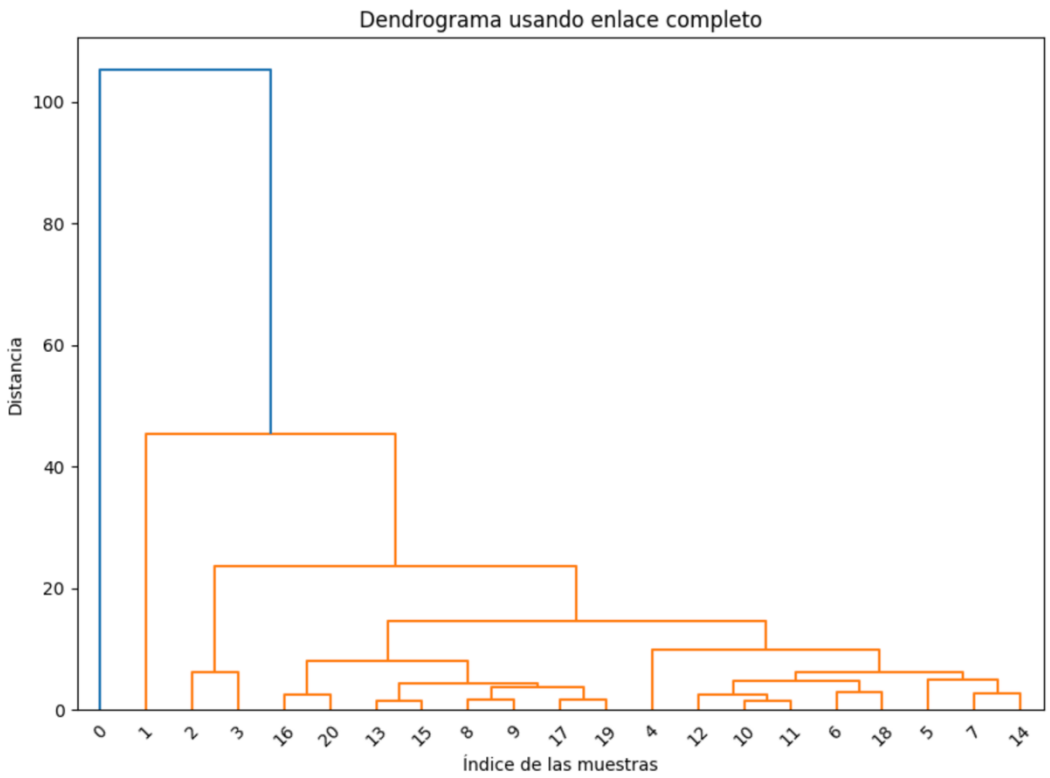
La covarianza indica la dirección de la relación, mientras que la correlación mide tanto la dirección como la intensidad de la relación, eliminando la influencia de las unidades de medida.

1.7 Dendograma

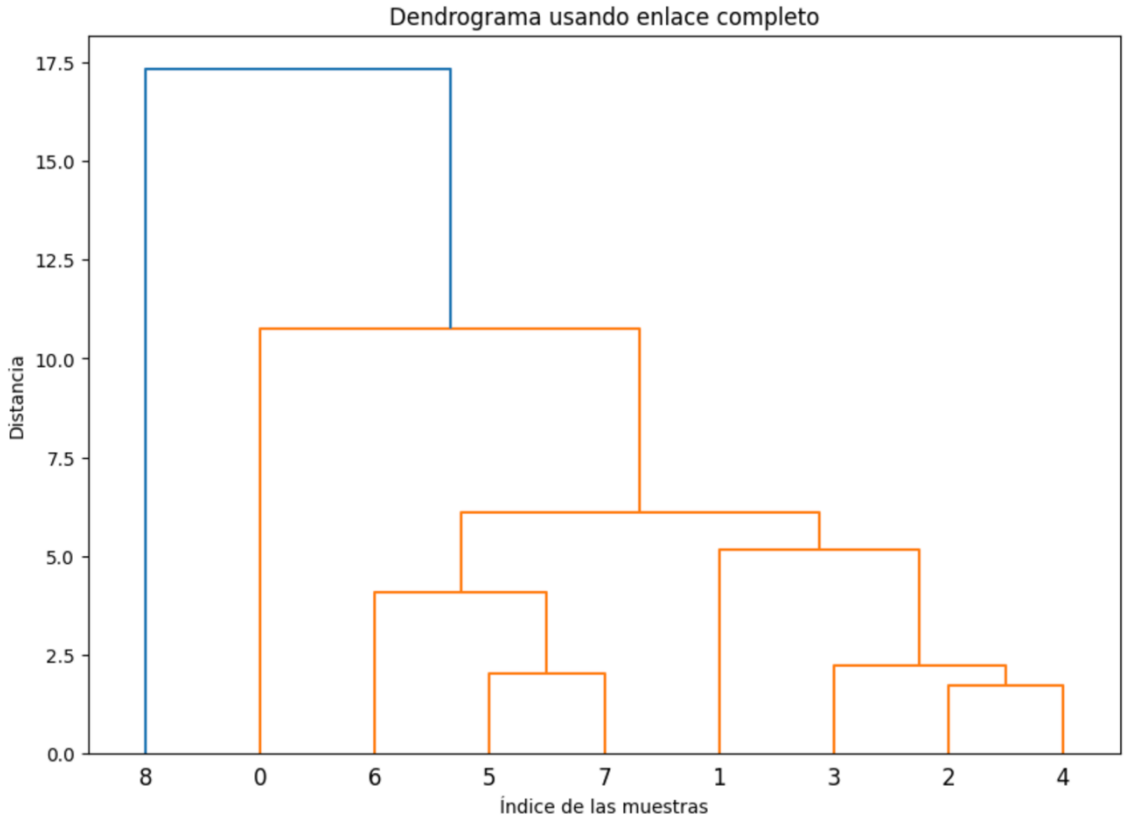
Con el data set completo



Con los de training



Con los de testing



2.1 Matriz de covarianza

	GDP (USD Billion)	Population (Millions)
GDP (USD Billion)	396.584655	26.685224
Population (Millions)	26.685224	1.830154

2.2 Cuales son los eigenvalues

```
[3.98380395e+02 3.44139937e-02]
```

2.3 Cual es la varianza explicada por cada eigenvalues

```
Varianza total: 398.4148093103449
Varianza explicada por cada componente:
[9.99913623e-01 8.63772954e-05]
```

2.4 Cual es el valor del eigenvector

```
[[ 0.99774346 -0.06714158]
 [ 0.06714158  0.99774346]]
```

2.5 Cual es la matriz proyectada

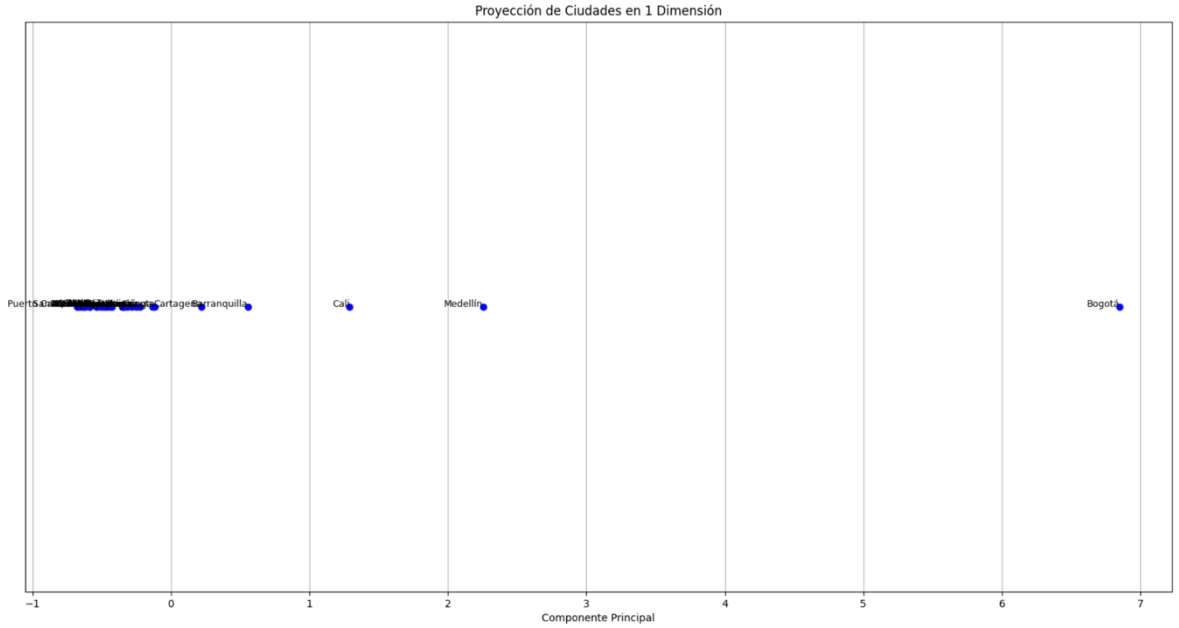
```
Matriz proyectada:
[[ 5.15381062  4.51267902]
 [ 1.89419666  1.25826945]
 [ 0.77124478  1.0776375 ]
 [ 0.43540102  0.34671042]
 [ 0.10426971  0.21828795]
 [-0.08151144 -0.10829742]
 [-0.14261314 -0.17953839]
 [-0.1845329   0.03427014]
 [-0.21143043 -0.13723108]
 [-0.25270164 -0.14198907]
 [-0.26743634 -0.20881496]
 [-0.27919023 -0.15527711]
 [-0.29700158 -0.19175488]
 [-0.30517403 -0.16106384]
 [-0.31637521 -0.17538061]
 [-0.33772008 -0.2643675 ]
 [-0.34892126 -0.27868428]
 [-0.36062723 -0.30050236]
 [-0.3667326  -0.31516205]
 [-0.37843857 -0.33698013]
 [-0.38605832 -0.37414373]
 [-0.39524035 -0.3584553 ]
 [-0.40997505 -0.42528119]
 [-0.41759479 -0.46244479]
 [-0.41915707 -0.40959276]
 [-0.42930076 -0.48426287]
 [-0.4328822  -0.46141605]
 [-0.43999715 -0.49107835]
 [-0.44660732 -0.51323934]
 [-0.45170312 -0.51289643]]
```

2.6 Cual es el error o diferencia entre la matriz proyectada

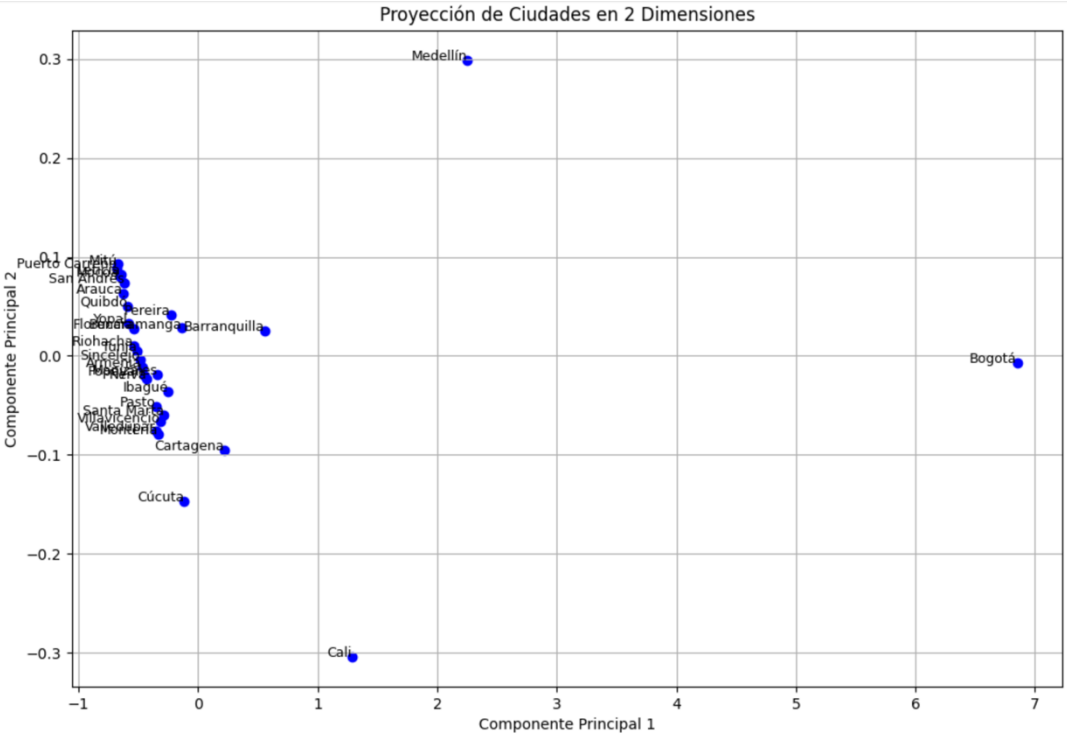
GDP (USD Billion)	Population (Millions)	
0	98.346189	2.667321
1	42.205803	1.311731
2	21.628755	1.152363
3	16.364599	0.883290
4	10.395730	0.811712
5	7.381511	0.688297
6	6.342613	0.659538
7	5.284533	0.725730
8	5.011430	0.667231
9	4.252702	0.661989
10	4.067436	0.638815
11	3.779190	0.655277
12	3.497002	0.641755
13	3.305174	0.651064
14	3.116375	0.645381
15	2.837720	0.614368
16	2.648921	0.608684
17	2.460627	0.600502
18	2.366733	0.595162
19	2.178439	0.586980
20	2.086058	0.574144

	GDP (USD Billion)	Population (Millions)
21	1.895240	0.578455
22	1.709975	0.555281
23	1.617595	0.542445
24	1.519157	0.559593
25	1.429301	0.534263
26	1.332882	0.541416
27	1.239997	0.531078
28	1.146607	0.523239
29	1.051703	0.522896

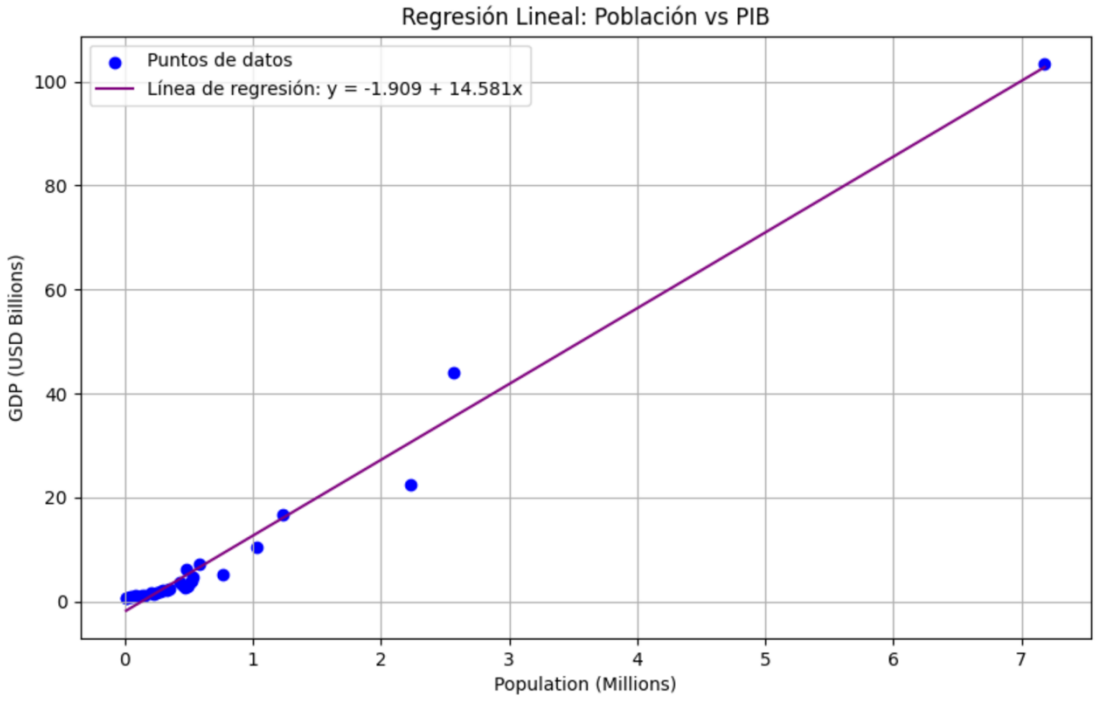
2.7 Ciudades en 1 dimensión



2.8 Ciudades en 2 dimensiones

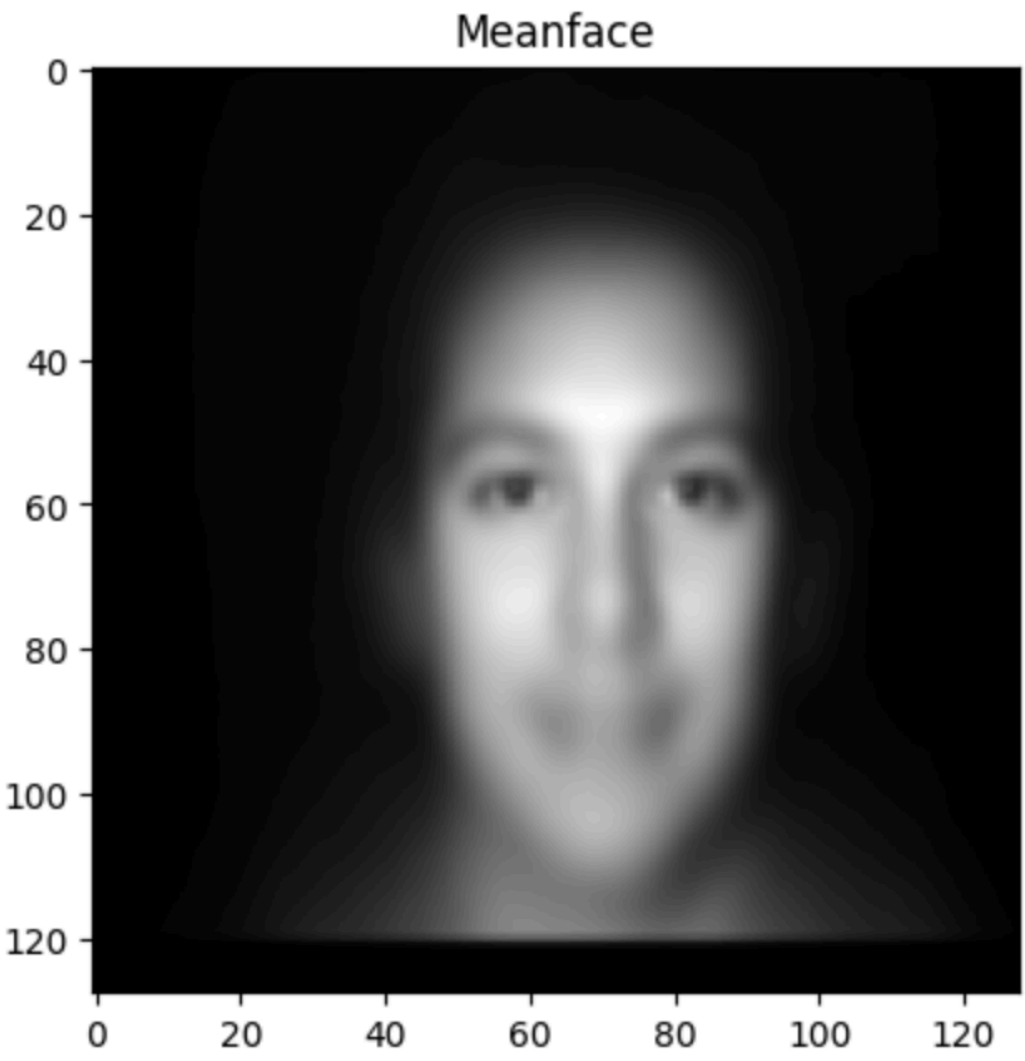


3. Regresión lineal




```
Ecuación de la regresión lineal: y = -1.909 + 14.581x
```

4. MeanFace

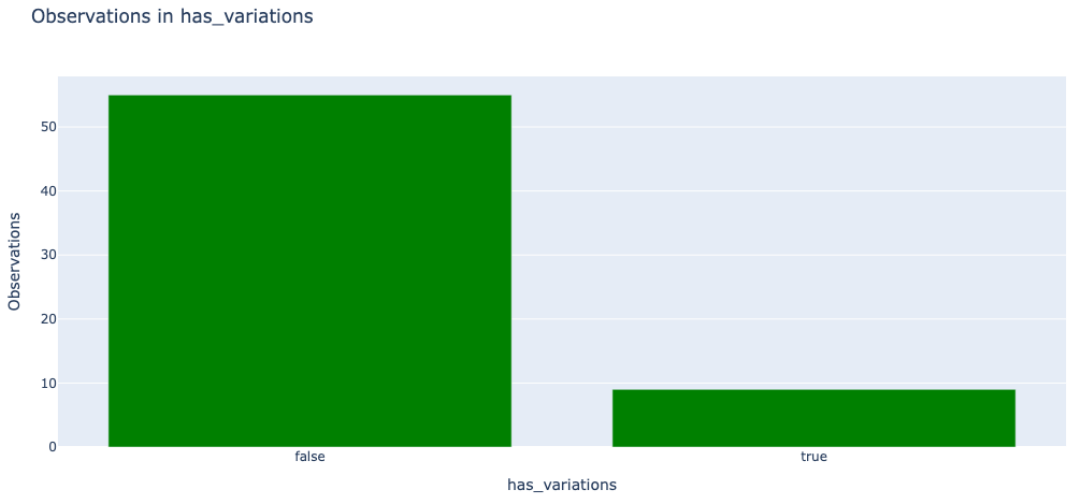


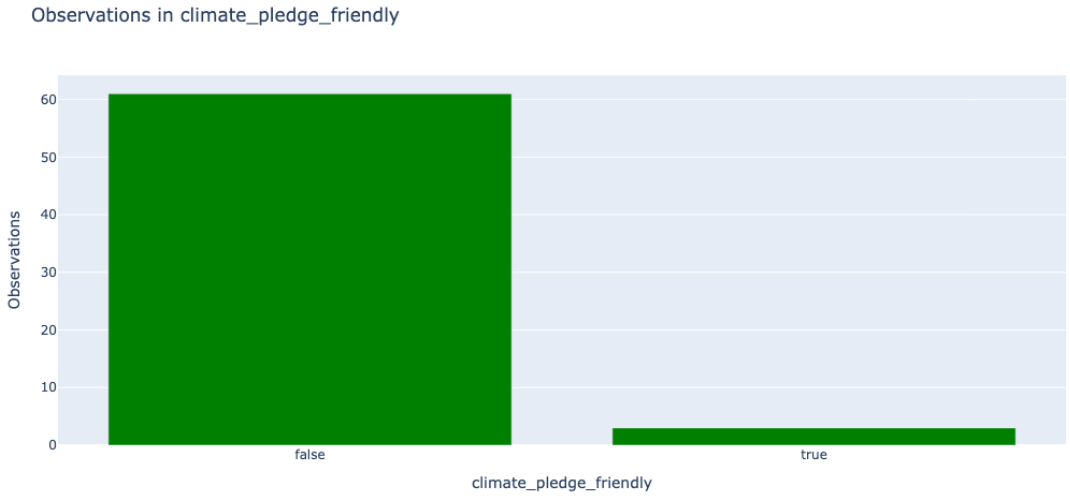
4.2 Centrar los datos, utilizar PCA. ¿Cuántos componentes se deben utilizar para mantener el 95% de las características?. Crear una tabla para mostrar las primeras 5 caras utilizando, la mean face + los datos reconstruidos utilizando la primera componente, después con 3 componentes, después con las primeras 20 componentes, después con las componentes que explican el 95% de la varianza y por último con el numero de componentes que tiene el 99% de la varianza. ¿Qué se puede concluir de los resultados?


Original
Original
Original
Original
Original
1 Componente
1 Componente
1 Componente
1 Componente
1 Componente
3 Componentes
3 Componentes
3 Componentes
3 Componentes
3 Componentes
20 Componentes
20 Componentes
20 Componentes
20 Componentes
20 Componentes
155 Componentes (95%)
155 Componentes (95%)
155 Componentes (95%)
155 Componentes (95%)
155 Componentes (95%)
1 Componentes (99%)
1 Componentes (99%)
1 Componentes (99%)
1 Componentes (99%)
1 Componentes (99%)

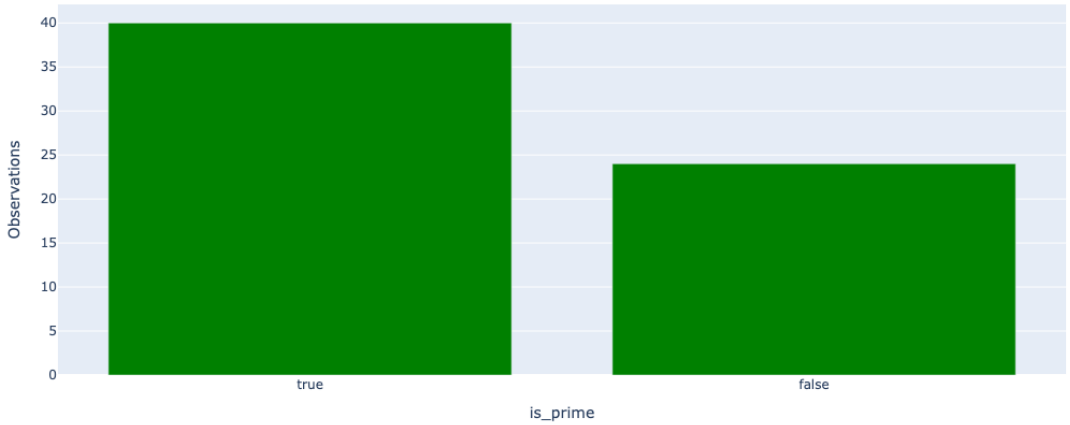

1 Componente
1 Componente
1 Componente
1 Componente
1 Componente

5.1.1 Para las variables categóricas un gráfico de barras. Categoría numero de observaciones.

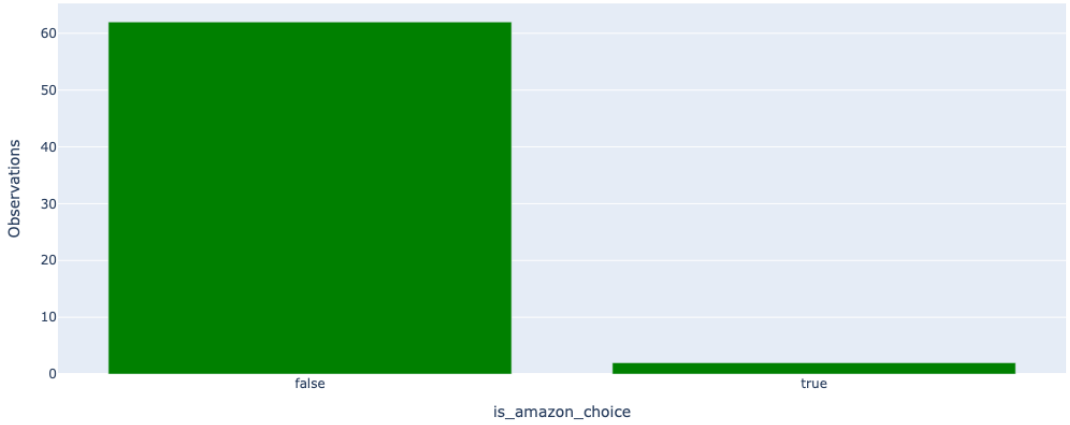


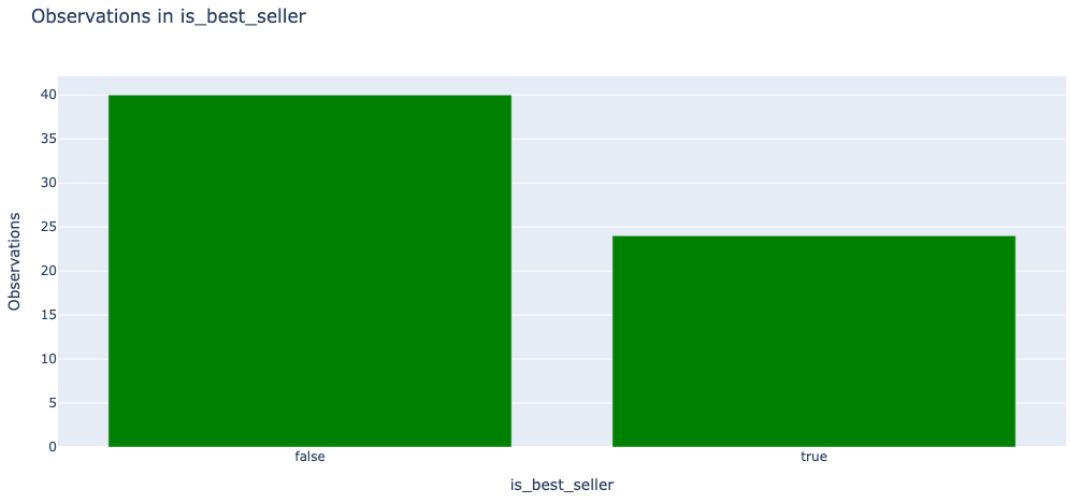


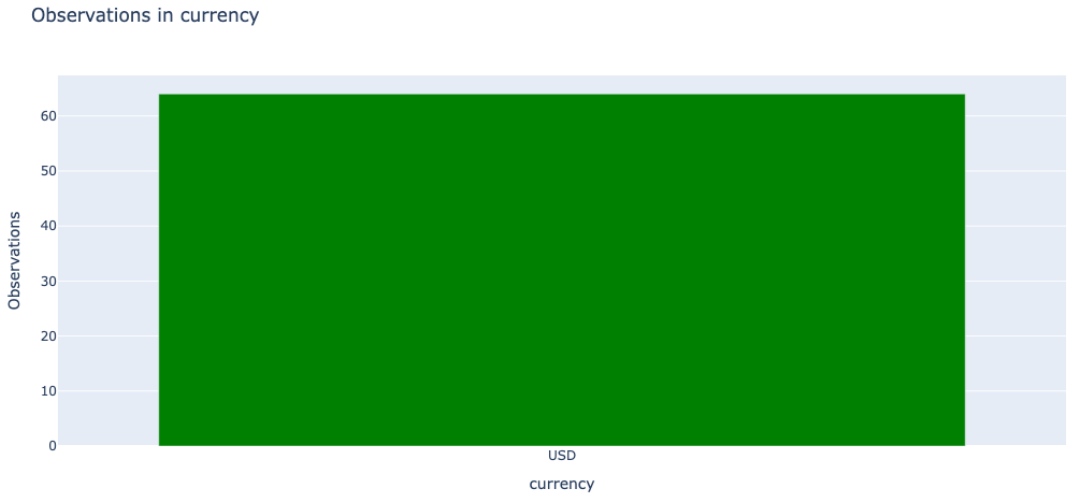
Observations in is_prime



Observations in is_amazon_choice

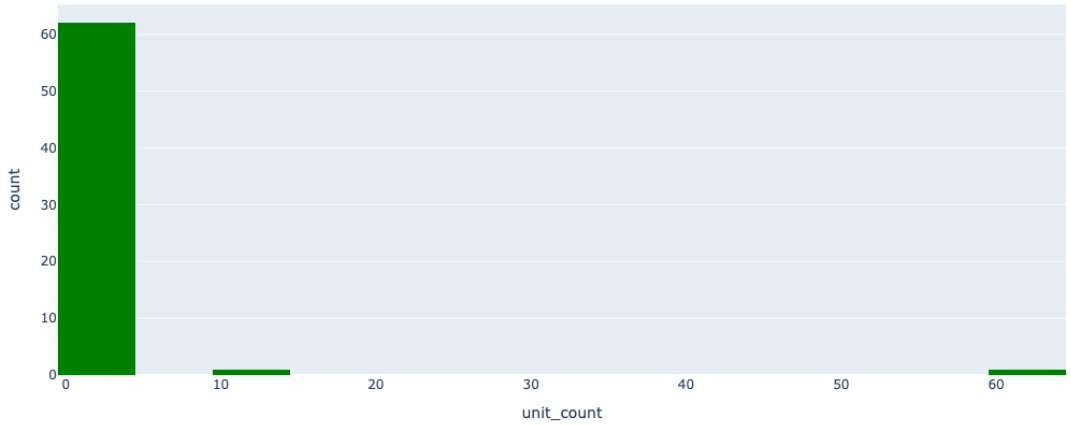






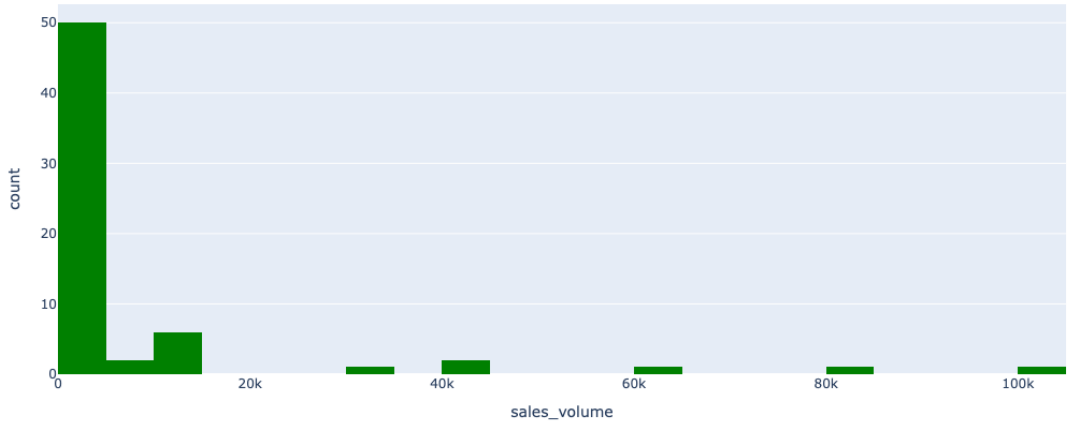
5.1.2. Para las variables numéricas crear histogramas. Listar los productos que están más lejos de 5 estándares de desviación, y serían considerados outliers. Hacer test de si es una distribución normal o no.

Histogram of unit_count

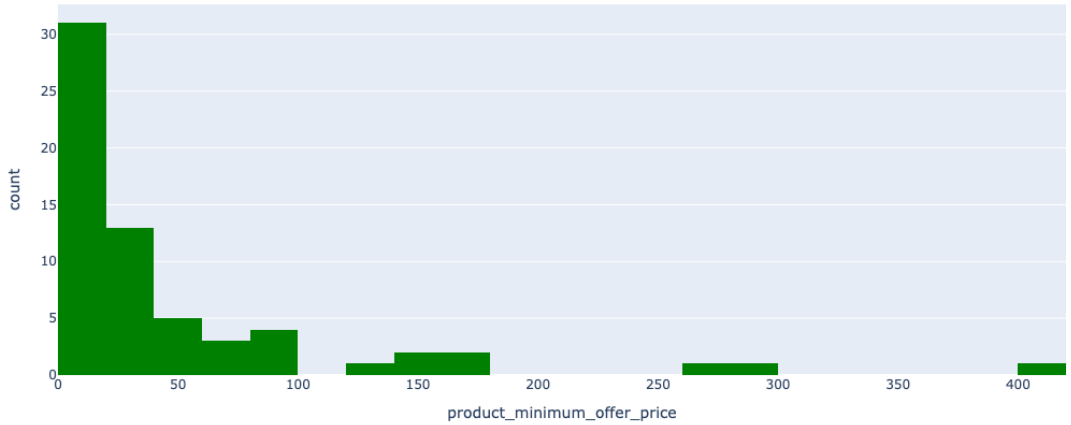


Histogram of unit_price

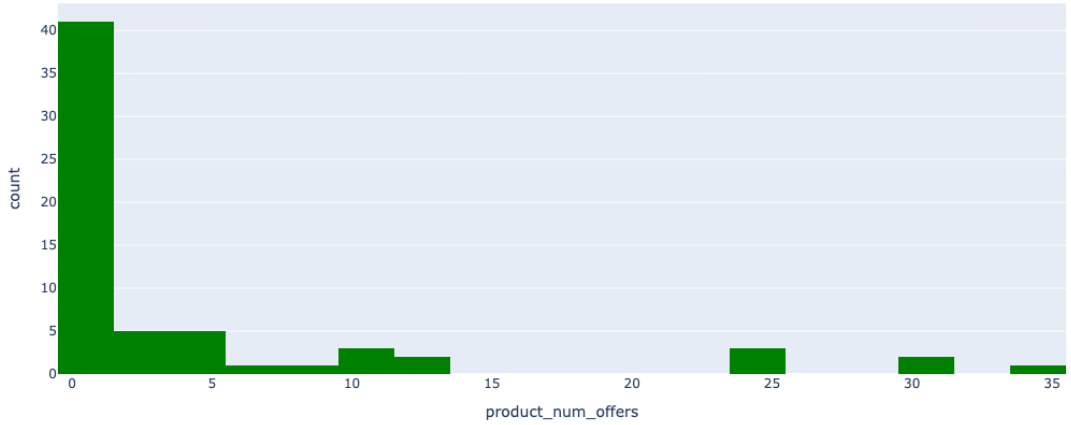
Histogram of sales_volume



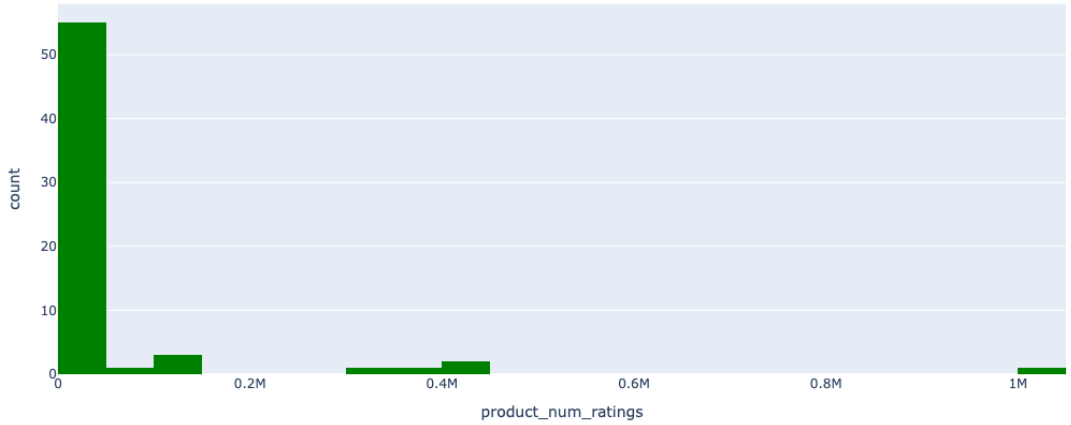
Histogram of product_minimum_offer_price



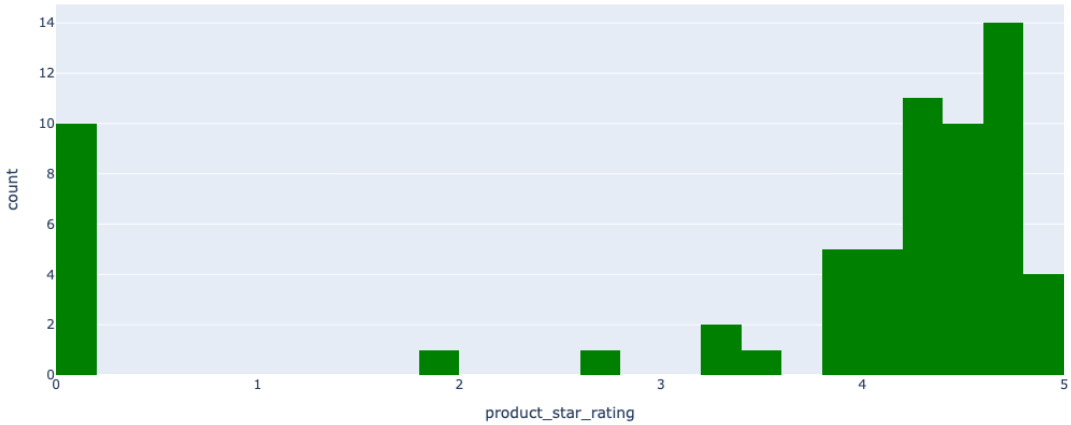
Histogram of product_num_offers



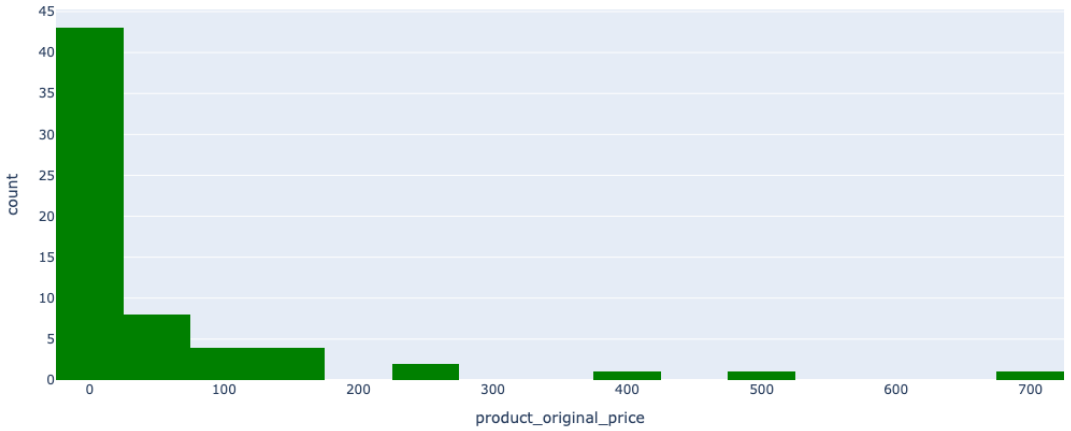
Histogram of product_num_ratings



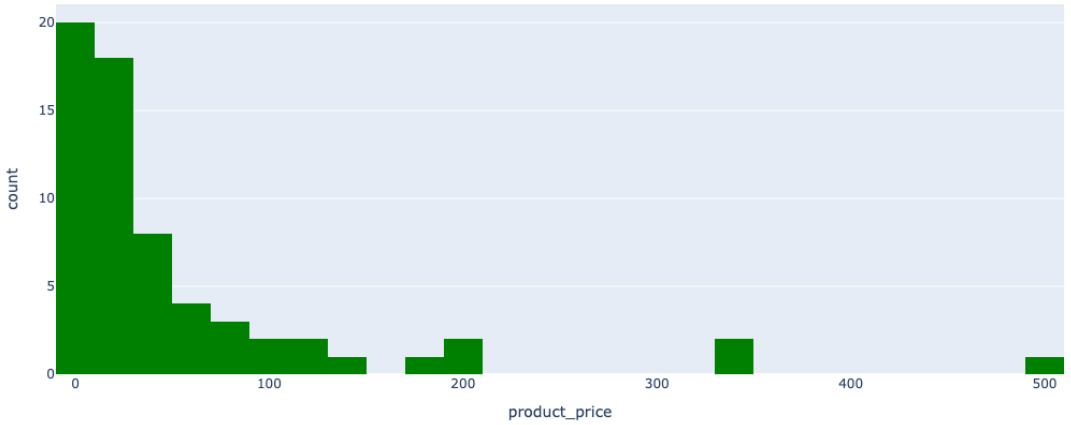
Histogram of product_star_rating



Histogram of product_original_price



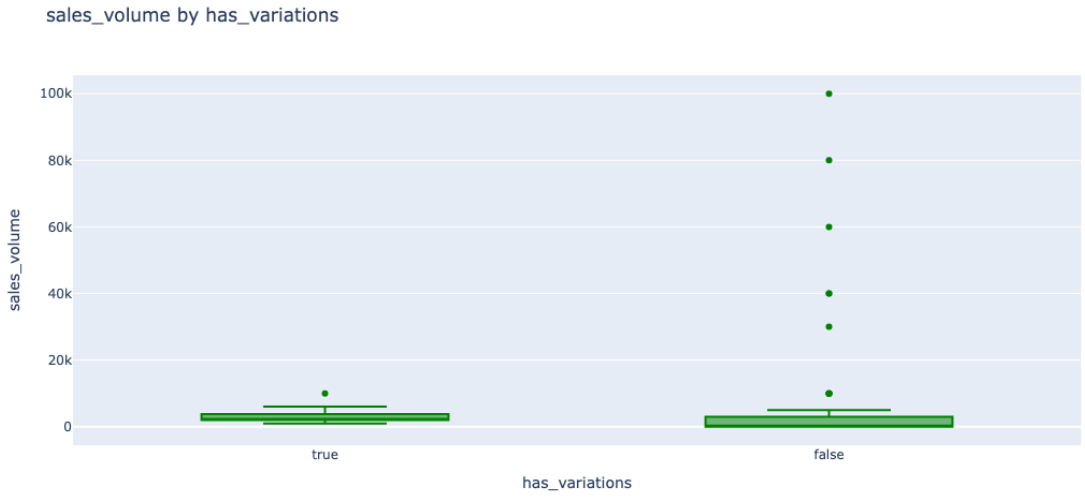
Histogram of product_price



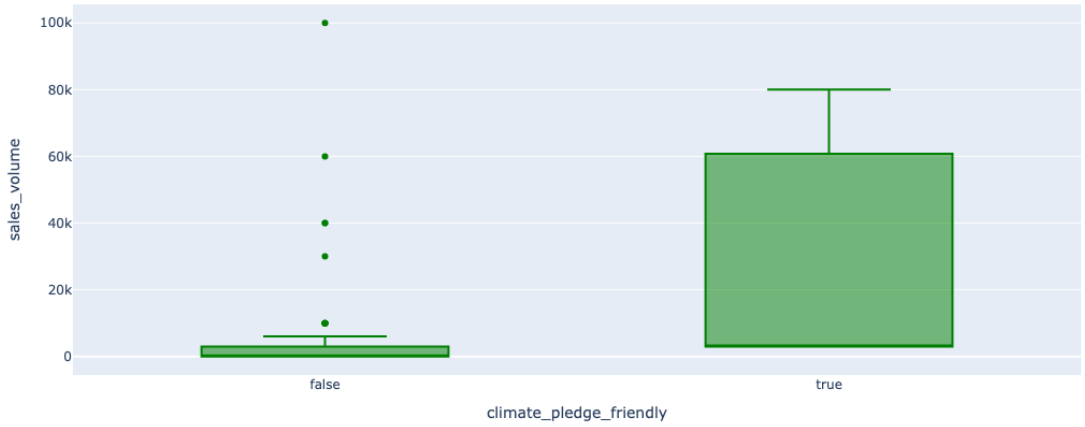
```
Shapiro-Wilk test for normality:

product_price: W=0.6241, p=0.0000
  -> Not normal

product_original_price: W=0.5015, p=0.0000
  -> Not normal

product_star_rating: W=0.6529, p=0.0000
  -> Not normal

product_num_ratings: W=0.3746, p=0.0000
  -> Not normal

product_num_offers: W=0.5428, p=0.0000
  -> Not normal

product_minimum_offer_price: W=0.6508, p=0.0000
  -> Not normal

sales_volume: W=0.4319, p=0.0000
  -> Not normal

unit_price: W=0.2447, p=0.0000
  -> Not normal

unit_count: W=0.1460, p=0.0000
  -> Not normal
```

```
Outliers found in product_original_price:
          asin  product_original_price
13  B0CGTD5KVT                   699.0

--------------------------------------------------

Outliers found in product_num_ratings:
          asin  product_num_ratings
55  B07Y8SJGCV              1015448

--------------------------------------------------

Outliers found in sales_volume:
          asin  sales_volume
29  B0D5FZGY8W        100000

--------------------------------------------------

Outliers found in unit_price:
          asin  unit_price
25  B0CS12LZLS        1.91
37  B0CV4FQPY1        2.05

--------------------------------------------------

Outliers found in unit_count:
          asin  unit_count
29  B0D5FZGY8W        60.0

--------------------------------------------------
```

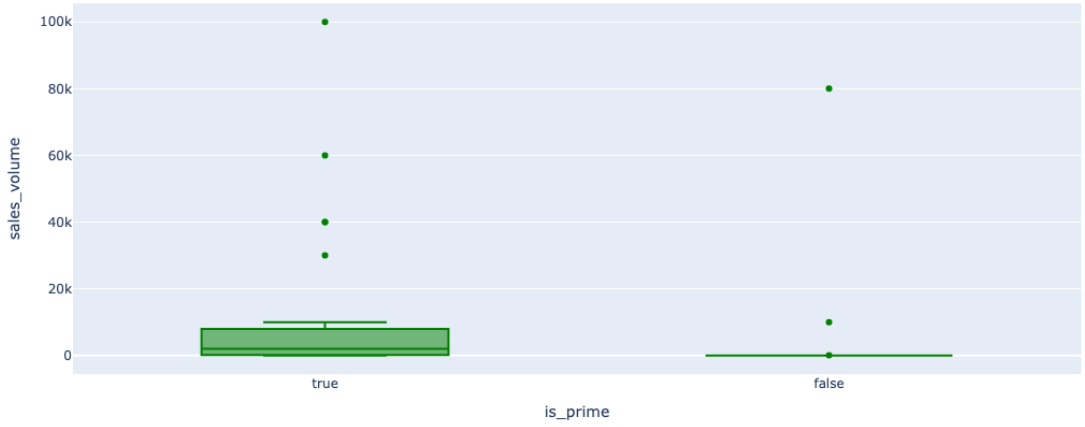
5.2.1. Variables categóricas debes crear un boxplot. Explique cómo interpreta el gráfico



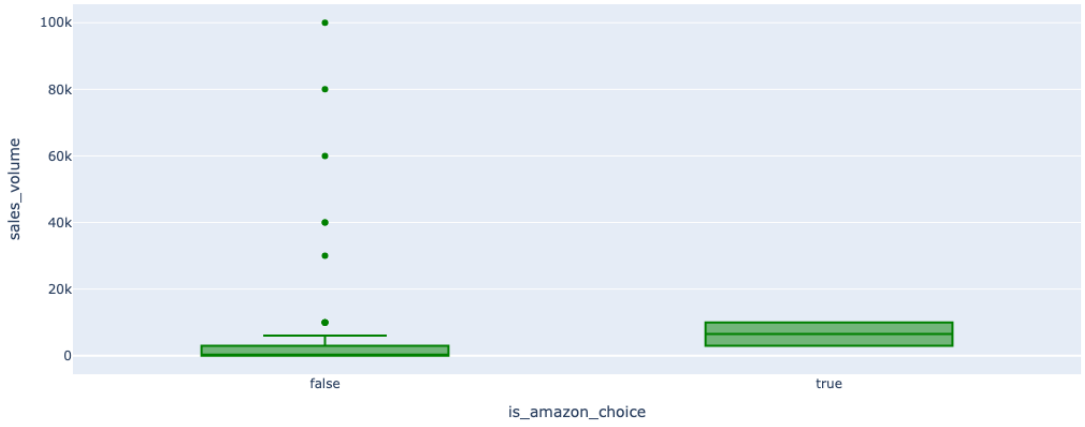
sales_volume by climate_pledge_friendly

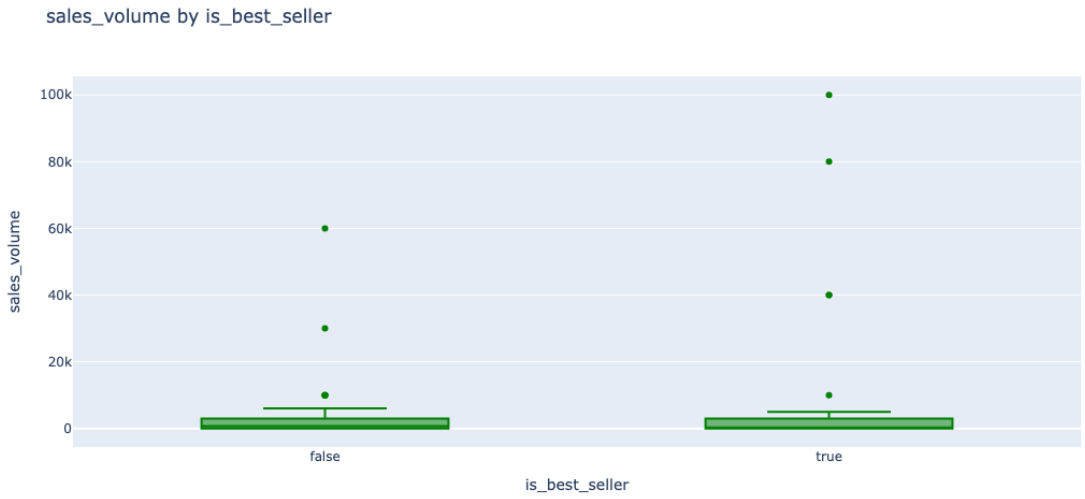


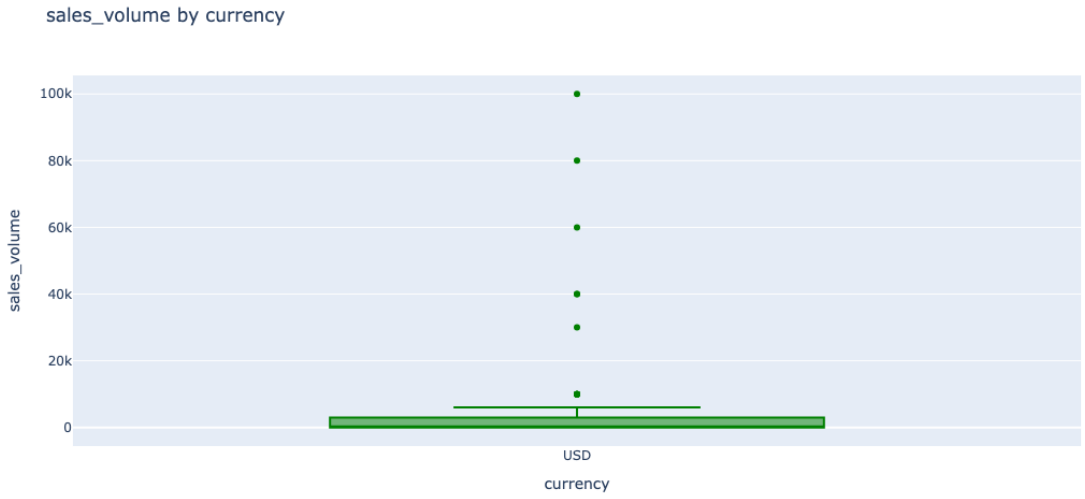
sales_volume by is_prime



sales_volume by is_amazon_choice

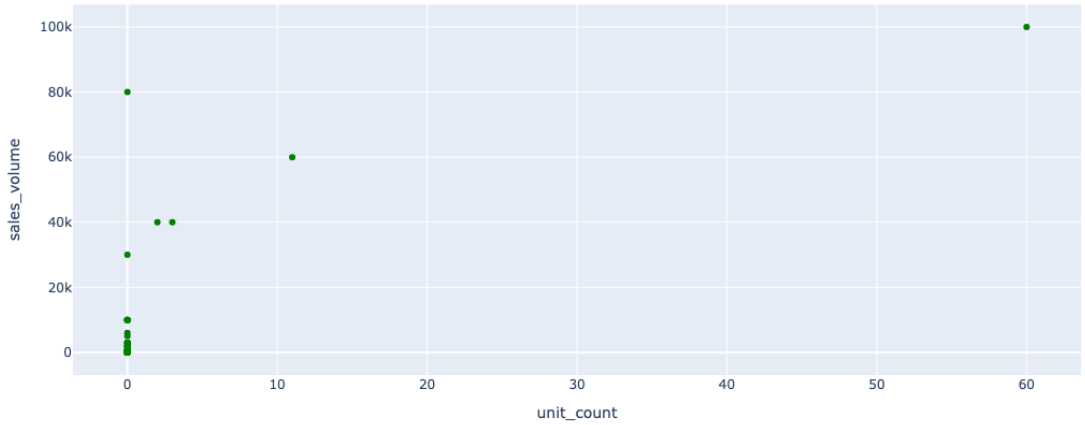






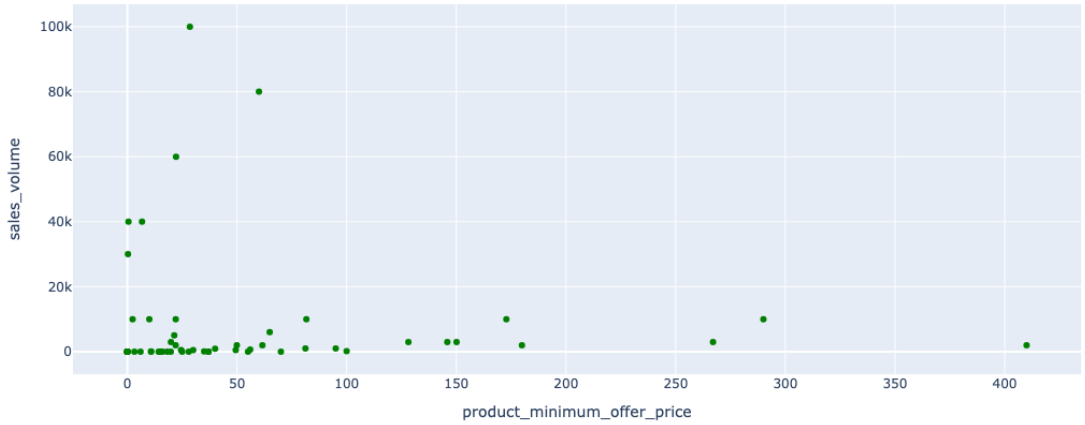
5.2.2. Variables numéricas vas a crear un scatter plot.Explique cómo interpreta el gráfico

unit_count vs sales_volume

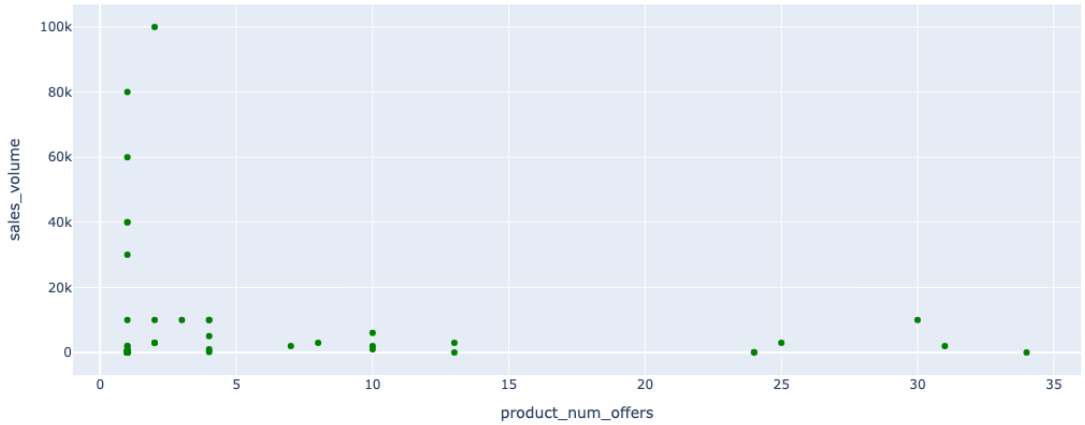


unit_price vs sales_volume

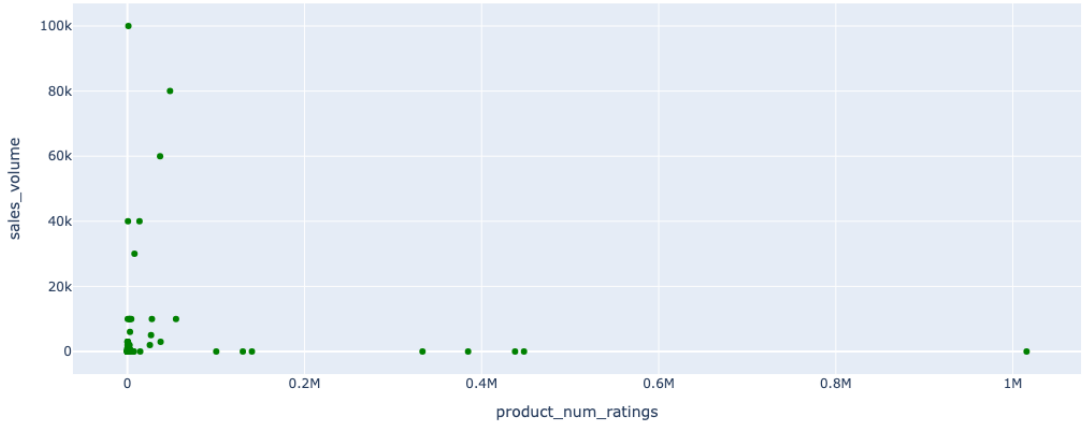
product_minimum_offer_price vs sales_volume



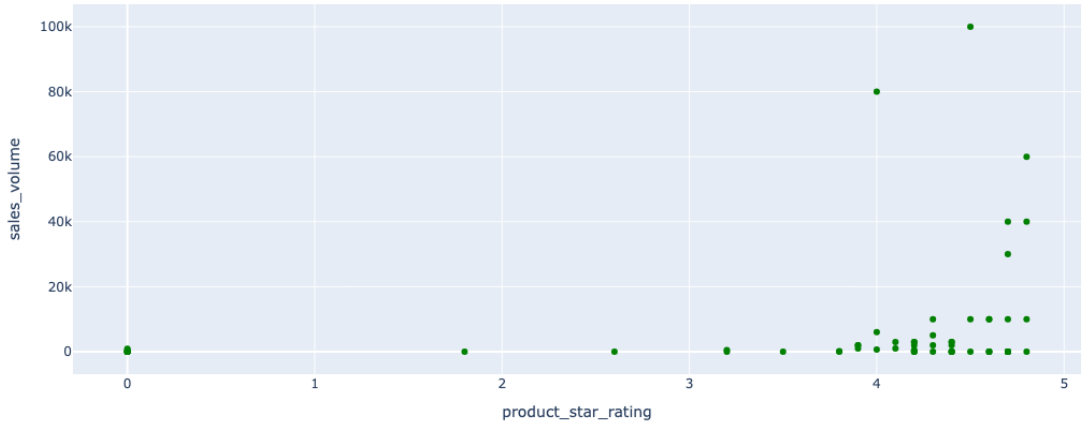
product_num_offers vs sales_volume

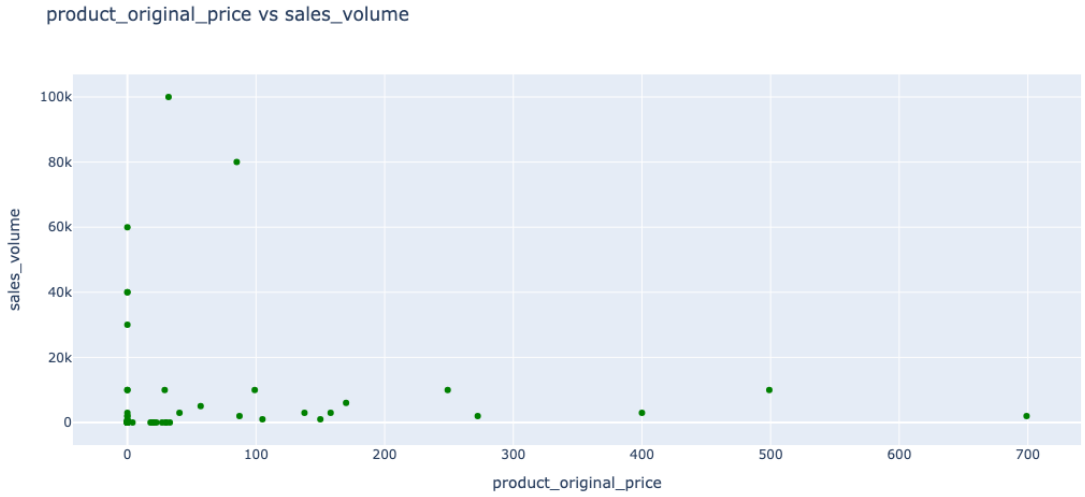


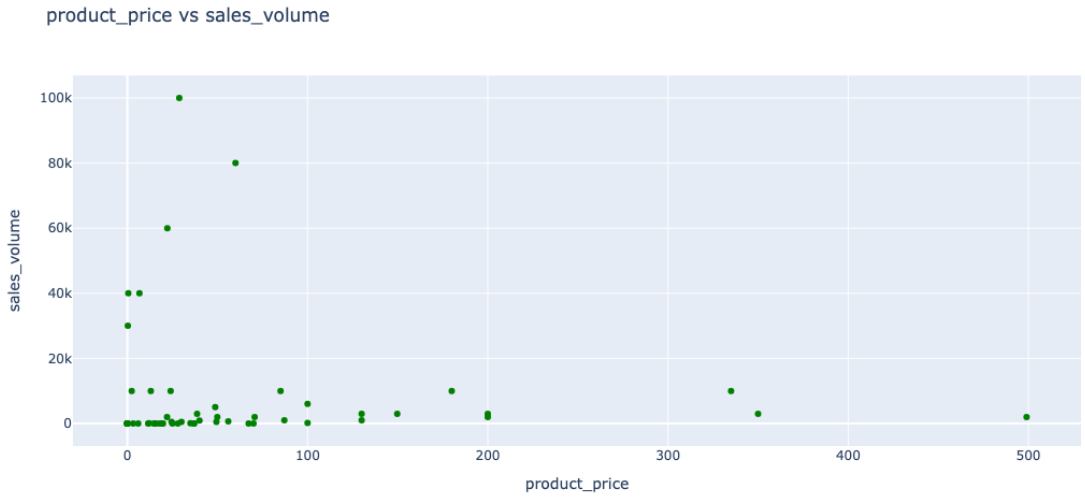
product_num_ratings vs sales_volume



product_star_rating vs sales_volume

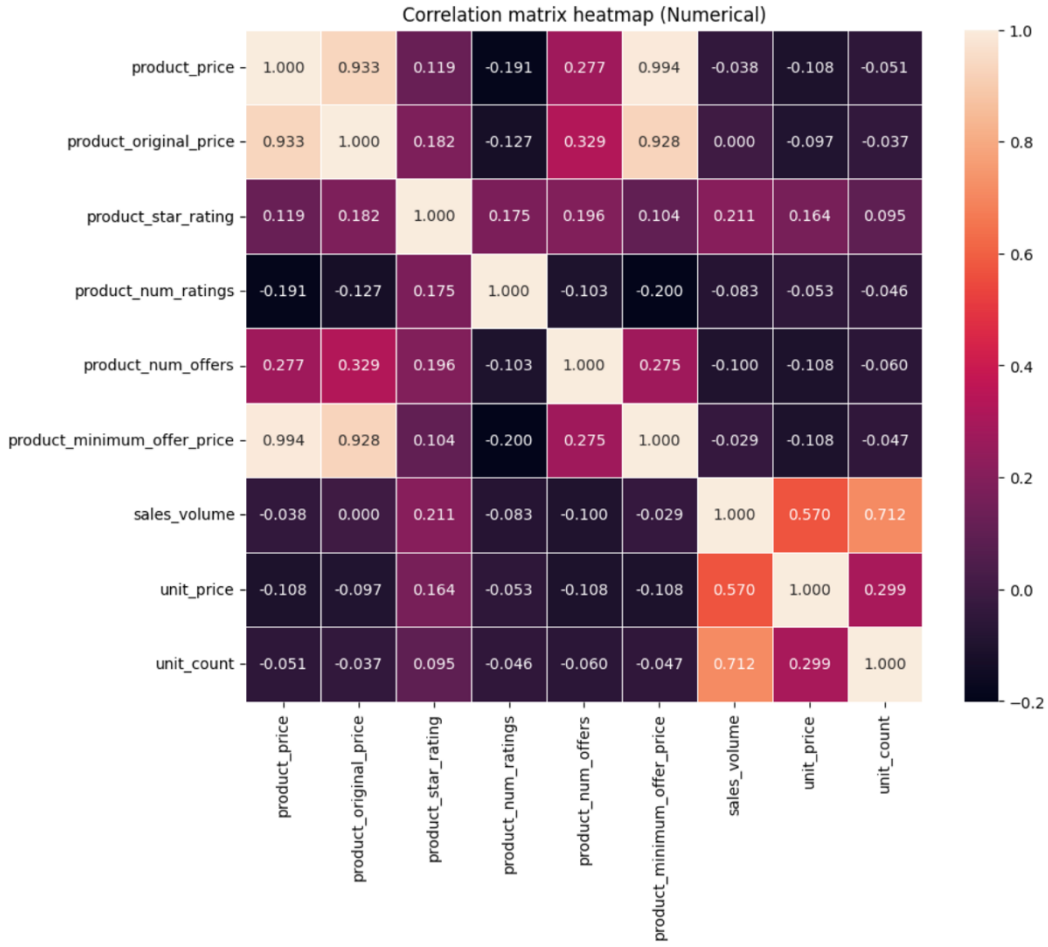




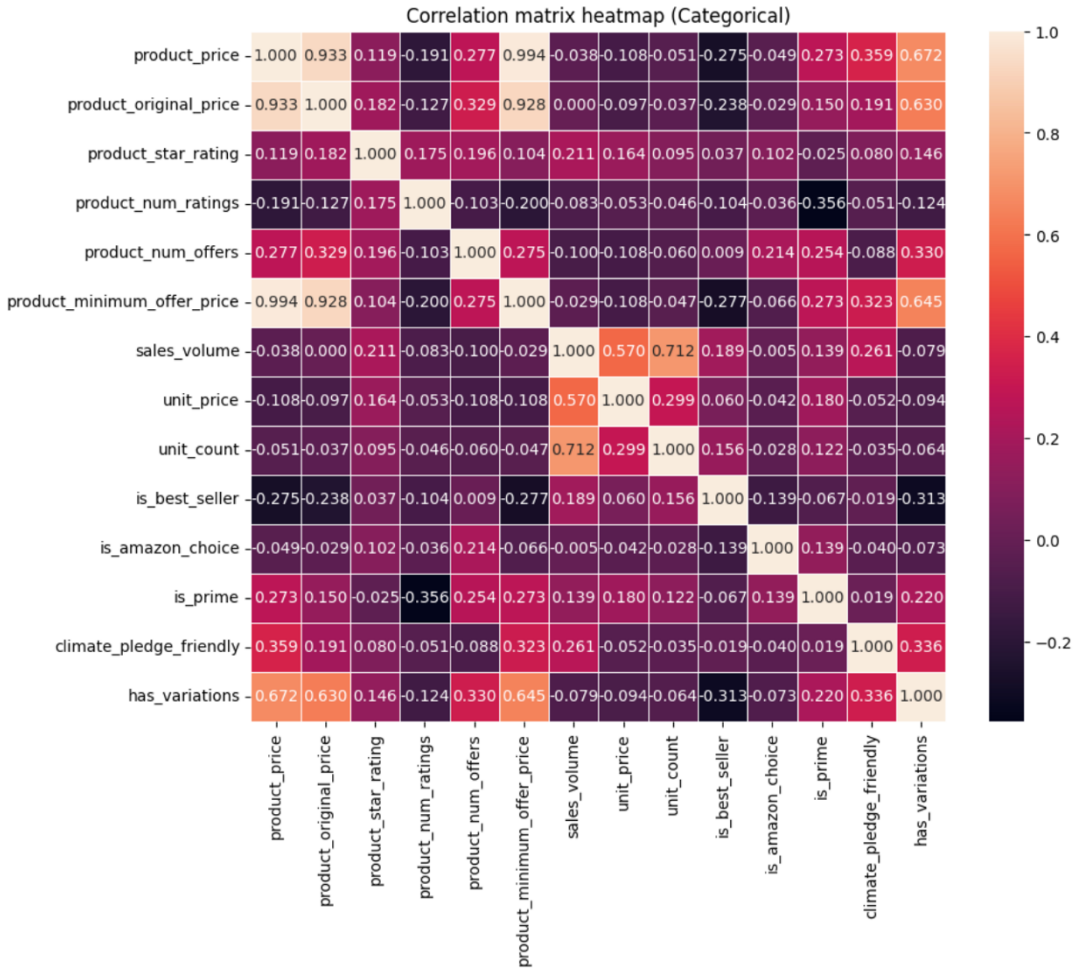


5.3.1. Cree la matriz de correlación, cuales son las variables más importantes para explicar la variabilidad de las sales_volume. Explique por qué el coeficiente es negativo o

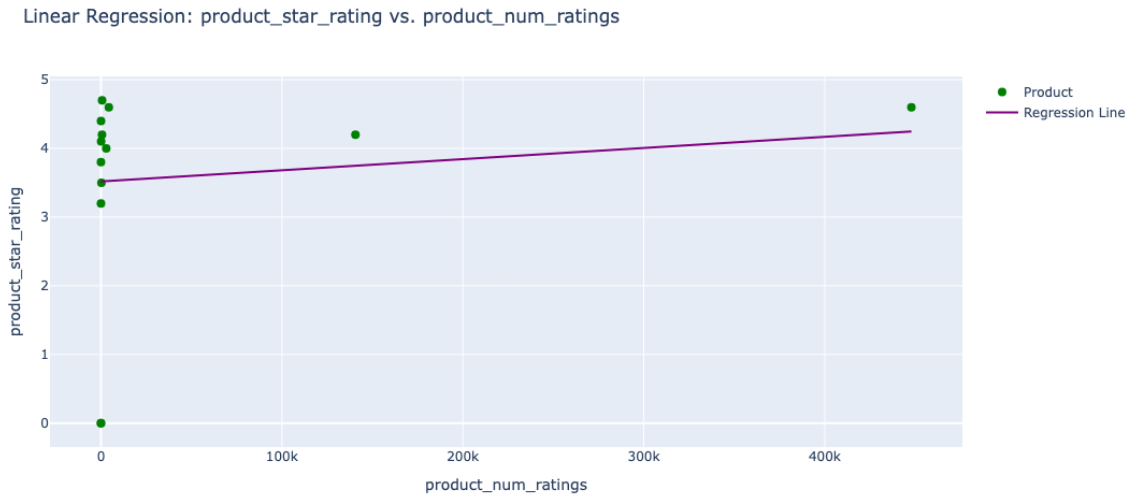
positivo.



5.3.2. Cree las dummy variables para todas las variables categóricas y genere la matriz de correlación nuevamente. ¿Cuál es el valor de variable categórica con mayor correlación?



5.3.3. Utilizar python para imputar los valores nulos con la media. Después dividir los datos en train y test. Por ultimo hacer una regresión entre x que es product_num_ratings y y product_star_rating qué es la calificación. Cual es el coeficiente b1 y b0. Describir resultados.


Linear Regression: product_star_rating vs. product_num_ratings
product_star_rating
product_num_ratings
0
1
2
3
4
5
0
100k
200k
300k
400k
Product
Regression Line