Tarea #: 1
Tema: Exploración de datos, PCA y regresión básica
Fecha entrega: 03/26/2025 11:55 PM
Objetivo: Utilizar conceptos estadísticos para entender la relación entre las variables de una base de datos. Adicionalmente, utilizar python como herramienta de exploración de datos y validación de hipótesis.
Entrega: Crear un repositorio en su github personal. Dentro del proyecto debe existir una carpeta llamada tarea 1, dentro debe tener una carpeta doc con este documento incluyendo todas las respuestas y los gráficos. Adicionalmente, debe existir una carpeta src con el código del notebook utilizado. Debe adicionar la cuenta jdramirez como colaborador del proyecto y enviar un email antes de q se termine el dia indicando el commit desea le sea calificado.

1. Utilizas el siguiente set de datos para calcular paso por paso (mostrar procedimiento y fórmulas):

City	GDP (USD Billion)	Population (Millions)	Unemployment Rate (%)	Average Age	Women (%)	Men (%)	Budget (USD Billion)	initial label	training
Bogotá	103.5	7.18	10.5	32	52	48	18	2	Yes
Medellín	44.1	2.57	11.2	31	53	47	7.5	3	Yes
Cali	22.4	2.23	13.8	30	52	48	4.2	2	Yes
Barranquilla	16.8	1.23	12.4	29	51	49	3.1	3	Yes
Cartagena	10.5	1.03	10.9	30	51	49	2.8	1	Yes
Bucaramanga	7.3	0.58	9.2	33	52	48	1.5	2	No
Pereira	6.2	0.48	12	32	52	48	1.3	1	Yes
Cúcuta	5.1	0.76	16.3	28	51	49	1.2	1	No
Ibagué	4.8	0.53	13.4	31	52	48	1.1	3	No
Santa Marta	4	0.52	11.6	29	51	49	0.9	3	Yes
Manizales	3.8	0.43	10.7	32	53	47	0.8	2	Yes
Villavicencio	3.5	0.5	13	30	51	49	0.8	0	No

Pasto	3.2	0.45	12.9	31	52	48	0.7	1	No
Montería	3	0.49	13.5	29	51	49	0.7	3	Yes
Valledupar	2.8	0.47	14.8	28	51	49	0.6	2	Yes
Neiva	2.5	0.35	14.1	30	52	48	0.6	3	Yes
Popayán	2.3	0.33	15.2	31	52	48	0.5	1	Yes
Armenia	2.1	0.3	13.3	32	53	47	0.5	0	Yes
Sincelejo	2	0.28	16.5	29	51	49	0.5	1	Yes
Tunja	1.8	0.25	10	31	52	48	0.4	2	Yes
Florencia	1.7	0.2	17.5	28	51	49	0.4	2	Yes
Riohacha	1.5	0.22	15.7	27	51	49	0.3	3	No
Quibdó	1.3	0.13	18.2	26	52	48	0.3	1	Yes
San Andrés	1.2	0.08	14	27	50	50	0.2	2	Yes
Yopal	1.1	0.15	11.5	29	51	49	0.2	0	Yes
Leticia	1	0.05	13.6	26	51	49	0.1	3	Yes
Arauca	0.9	0.08	12.2	29	51	49	0.1	2	No
Mocoa	0.8	0.04	15	28	52	48	0.1	0	No
Mitú	0.7	0.01	20	25	51	49	0.05	2	Yes
Puerto Carreño	0.6	0.01	22	24	50	50	0.05	0	No

Tabla tomada del DANE https://www.dane.gov.co/files/operaciones/PIB/departamental/anex-PIBDep-TotalDepartamento-2022pr.xlsx.

1.1. ¿Cuál es la media, mediana y desviación estándar?, y la moda y los valores repeticiones de la moda para los datos categóricos.

MEDIA

GDP	8.750000
Population	0.731000
Unemployment Rate	13.435333
Average Age	29.233333
Women (%)	51.500000
Men (%)	48.500000
Budget	1.650000
Initial Label	1.700000

MEDIANA

GDP	2.65
Population	0.39
Unemployment Rate	13.45
Average Age	29.00
Women (%)	51.00
Men (%)	49.00
Budget	0.60

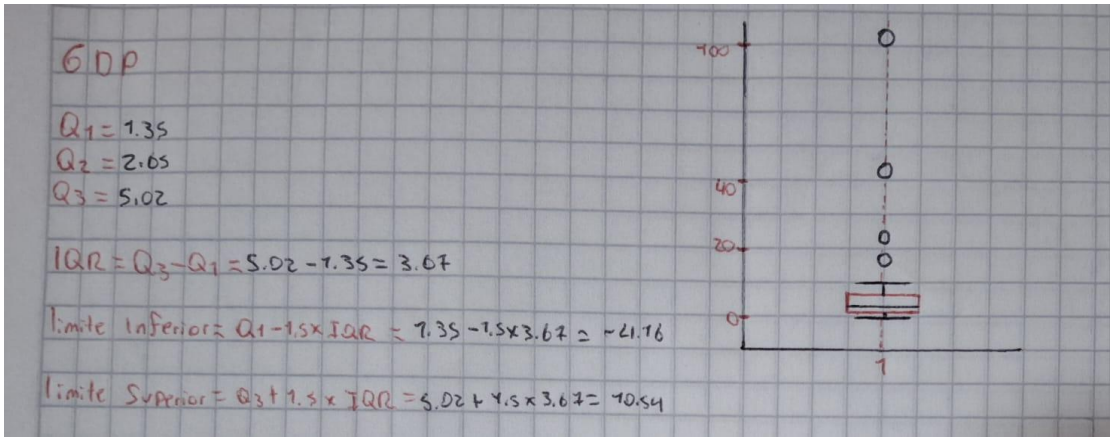
DESVIACION ESTANDAR

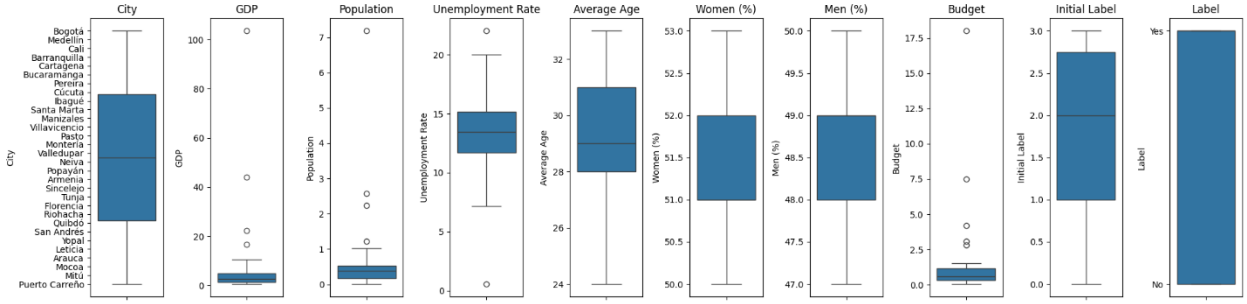
GDP	19.914433
Population	1.352832
Unemployment Rate	3.869034
Average Age	2.238893
Women (%)	0.776819
Men (%)	0.776819
Budget	3.451187

MODA

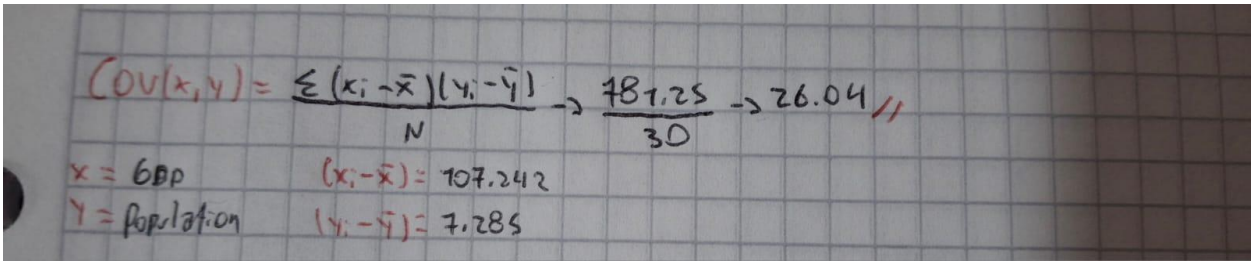
City	Arauca
GDP	0.6
Population	0.01
Unemployment Rate	0.58
Average Age	29.0
Women (%)	51.0
Men (%)	49.0
Budget	0.1

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

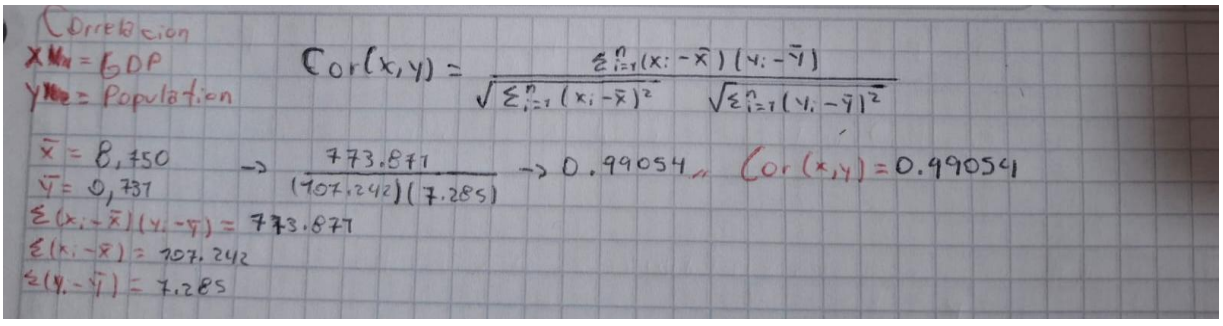




1.3. Cual es la covarianza entre las 2 variables X1, X2



1.4. Cuál es la correlación entre la variable x1 y x2 (Calcularla a mano). Correlación puede ser escrita también como:



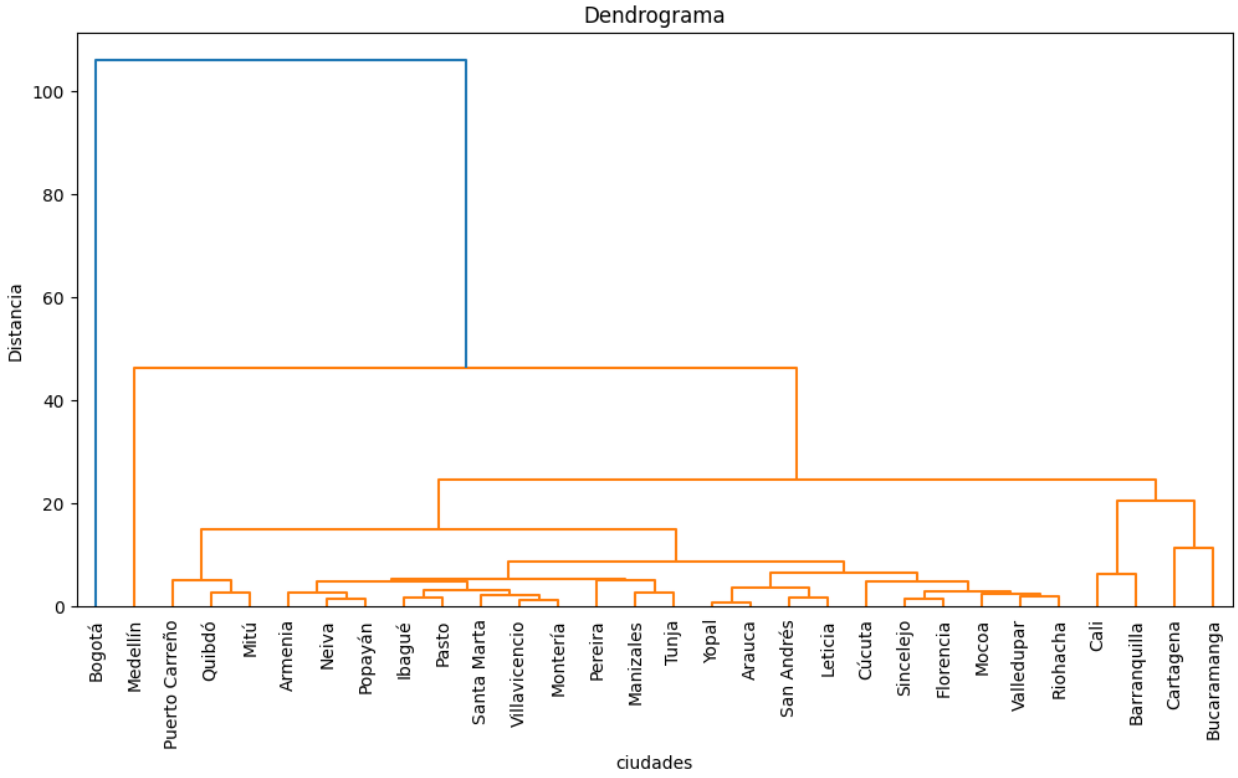
1.5. Explica la relación entre covarianza y correlación.

La covarianza indica la dirección de la relación entre las dos variables.
La correlación indica tanto la dirección como la fuerza de la relación, y es más útil para interpretar y comparar.

1.6. Calcule el resultado del algoritmo K-means sobre este set de datos a mano como lo hicimos en excel o con python sin utilizar librerías. Vamos a crear 4 grupos, es decir, k=4 (clusters).

City	new-label
Bogotá	2
Medellín	2
Cali	0
Barranquilla	3
Cartagena	3
Pereira	0
Santa Marta	3
Manizales	0
Montería	3
Valledupar	1
Neiva	0
Popayán	0
Armenia	0
Sincelejo	1
Tunja	0
Florencia	1
Quibdó	1
San Andrés	1
Yopal	3
Leticia	1
Mitú	1

1.7. Calcula el resultado de un dendrograma utilizando la distancia máxima (complete) en python.



2. PCA. Utilizar los datos de la tabla 1, para calcular PCA y reducir la dimensionalidad de 2 dimensiones a 1. Para este ejercicio se debe utilizar las variables GDP (USD Billion) y Population (Millions) para crear un vector con una sola dimensión.

2.1. Cual es la matriz de covarianza

	GDP	Population
GDP	396.584655	26.685224
Population	26.685224	1.830154

2.2. Cuales son los eigenvalues

Eigenvalue 1: 3.98380395e+02

Eigenvalue 2: 3.44139937e-02

2.3. Cuál es la varianza explicada por el eigenvalue.

CP1: 9.99913623e-01

CP2: 8.63772954e-05

2.4. Cual es el valor del eigenvector

eigenvector 1: 0.99774346 -0.06714158
eigenvector 2: 0.06714158 0.99774346

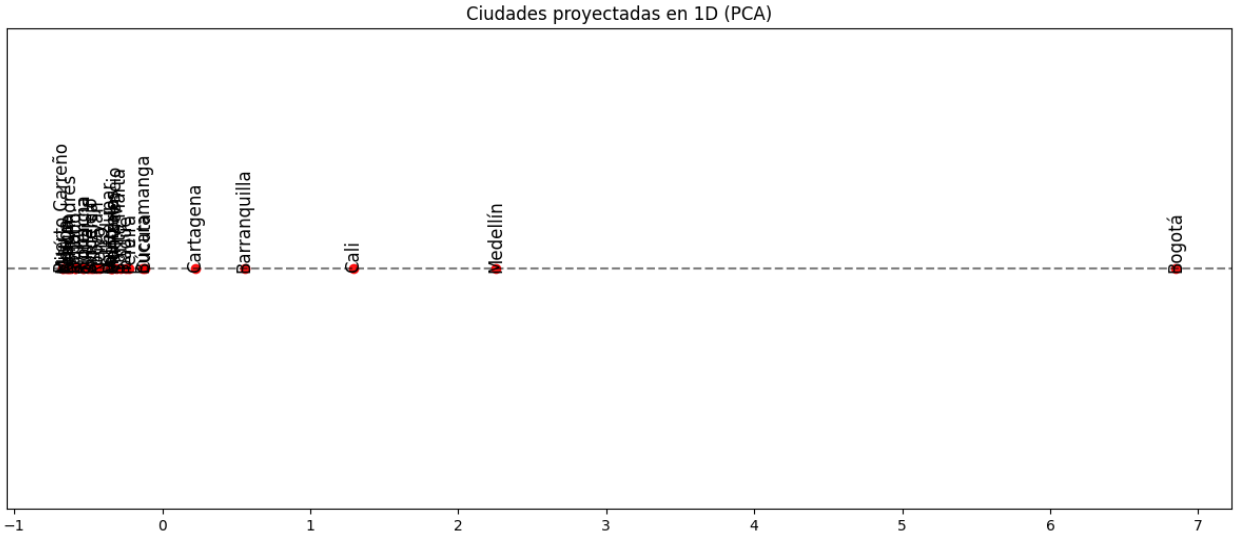
2.5. Cuál es la matriz proyectada.

```
[[ 9.49691887e+01  7.27833176e-02]
 [ 3.53937046e+01 -5.38604462e-01]
 [ 1.37198434e+01  5.79134943e-01]
 [ 8.06533849e+00 -4.26156942e-02]
 [ 1.76612638e+00  1.80827538e-01]
 [-1.45686639e+00 -5.33039783e-02]
 [-2.56109835e+00 -7.92225914e-02]
 [-3.63981652e+00  2.74001310e-01]
 [-3.95458212e+00  6.46627867e-02]
 [-4.75344830e+00  1.08398612e-01]
 [-4.95903973e+00  3.20300160e-02]
 [-5.25366286e+00  1.22014531e-01]
 [-5.55634298e+00  9.22698303e-02]
 [-5.75320601e+00  1.45607884e-01]
 [-5.95409753e+00  1.39081330e-01]
 [-6.26147756e+00  3.94945871e-02]
 [-6.46236908e+00  3.29680329e-02]
 [-6.66393202e+00  1.64640442e-02]
 [-6.76504919e+00  3.22333254e-03]
 [-6.96661213e+00 -1.32806562e-02]
 [-7.06974356e+00 -5.64536716e-02]
 [-7.26794942e+00 -2.30704874e-02]
 [-7.47354085e+00 -9.94390836e-02]
 [-7.57667228e+00 -1.42612099e-01]
 [-7.67174671e+00 -6.60558994e-02]
 [-7.77823522e+00 -1.59116088e-01]
 [-7.87599531e+00 -1.22469626e-01]
 [-7.97845532e+00 -1.55665207e-01]
 [-8.08024392e+00 -1.78883354e-01]
 [-8.18001826e+00 -1.72169196e-01]]
```

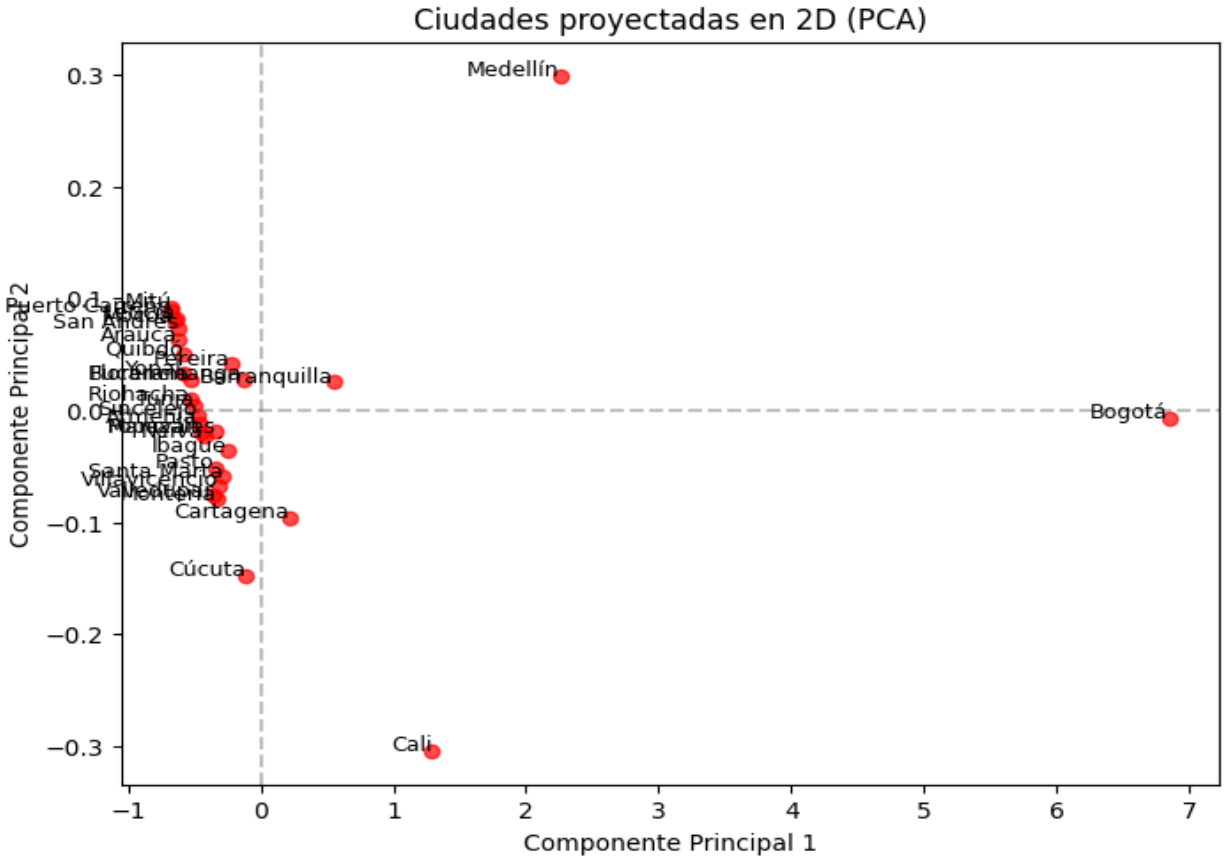
2.6. Cual es el error o diferencia entre la matriz proyectada

	GDP	Population
0	8.530811	7.107217
1	8.706295	3.108604
2	8.680157	1.650865
3	8.734662	1.272616
4	8.733874	0.849172
5	8.756866	0.633304
6	8.761098	0.559223
7	8.739817	0.485999
8	8.754582	0.465337
9	8.753448	0.411601
10	8.759040	0.397970
11	8.753663	0.377985
12	8.756343	0.357730
13	8.753206	0.344392
14	8.754098	0.330919
15	8.761478	0.310505
16	8.762369	0.297032
17	8.763932	0.283536
18	8.765049	0.276777
19	8.766612	0.263281
20	8.769744	0.256454
21	8.767949	0.243070
22	8.773541	0.229439
23	8.776672	0.222612
24	8.771747	0.216056
25	8.778235	0.209116
26	8.775995	0.202470
27	8.778455	0.195665
28	8.780244	0.188883
29	8.780018	0.182169

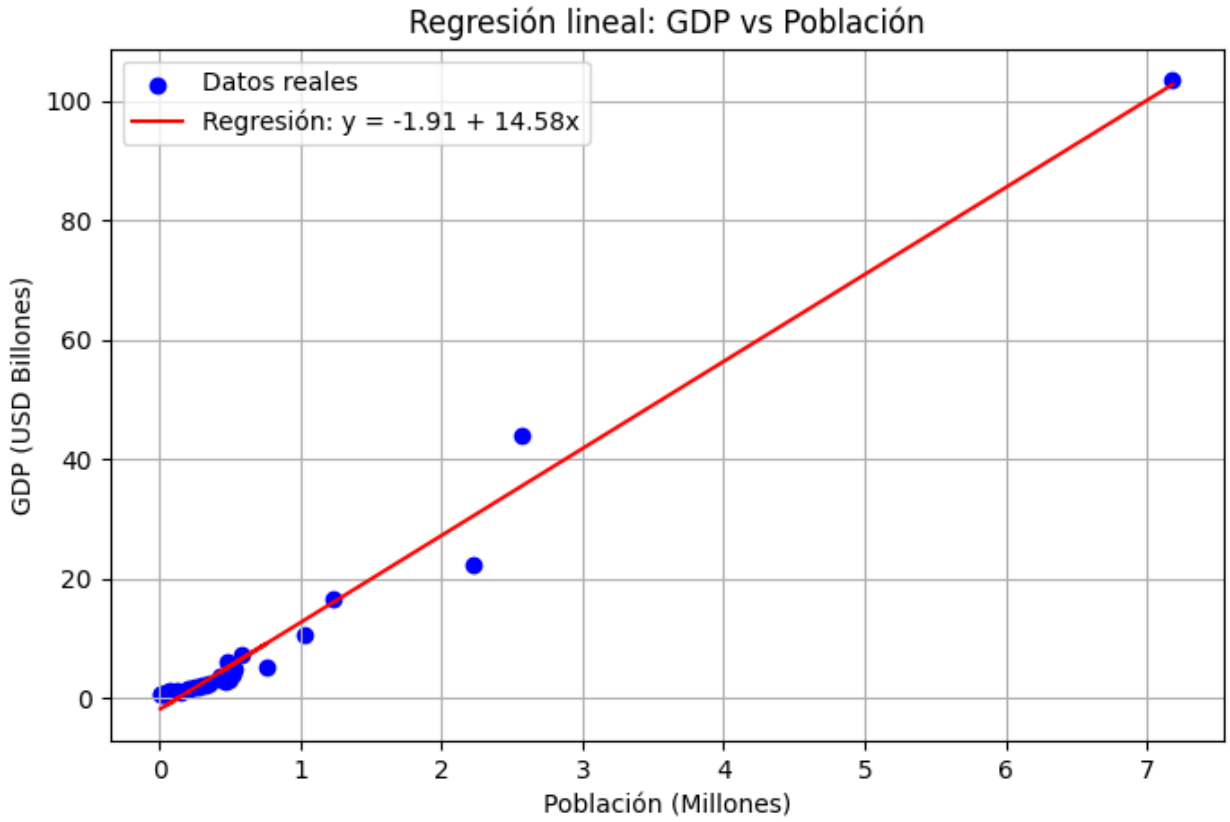
2.7. Pintar todas las ciudades en 1 dimensión.



2.8. Utilizar python para pintar todas las ciudades en 2 dimensiones,



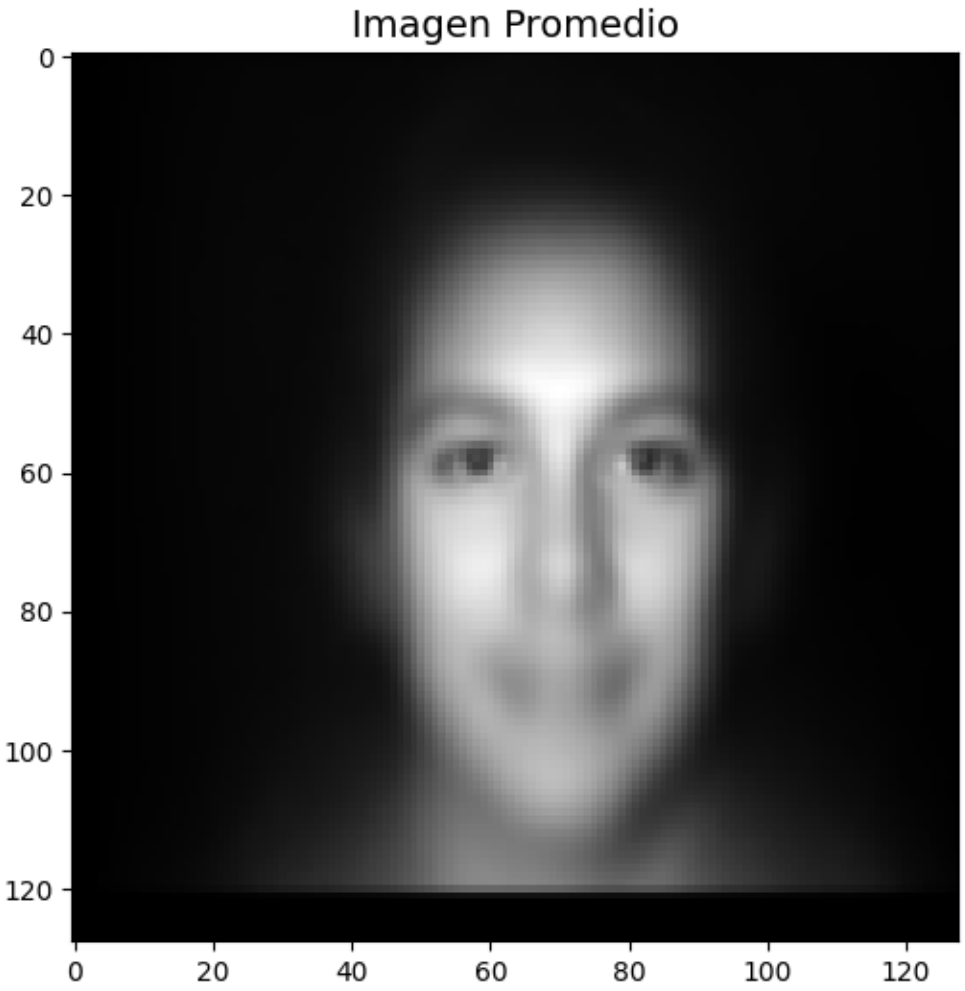
3. Regression. Utiliza las variables GDP (USD Billion) y Population (Millions) para crear una regresión. X es la población, y es el GDP.
   3.1. Calcular b0, b1 sin librerías.
   b0 (intercepto): -1.91
   b1 (pendiente): 14.58
   3.2. Graficar la línea y los puntos



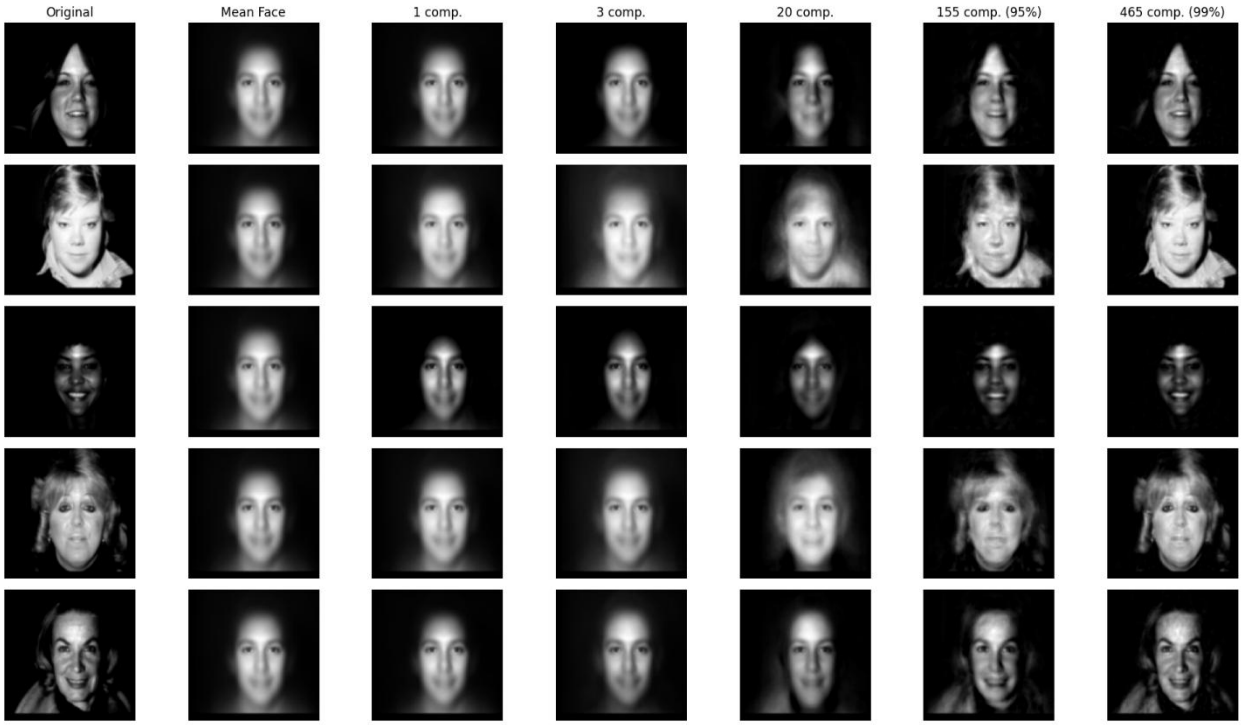
4. PCA

Utiliza solo las caras de entrenamiento para los siguientes puntos:

1. Calcular la mean face. Que es la cara con el promedio de los pixeles y visualizarla.

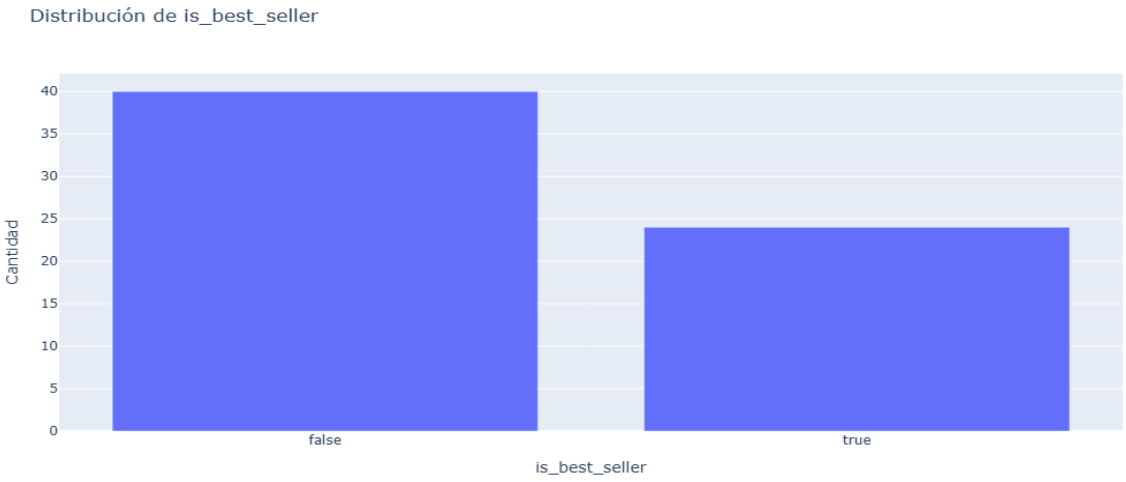


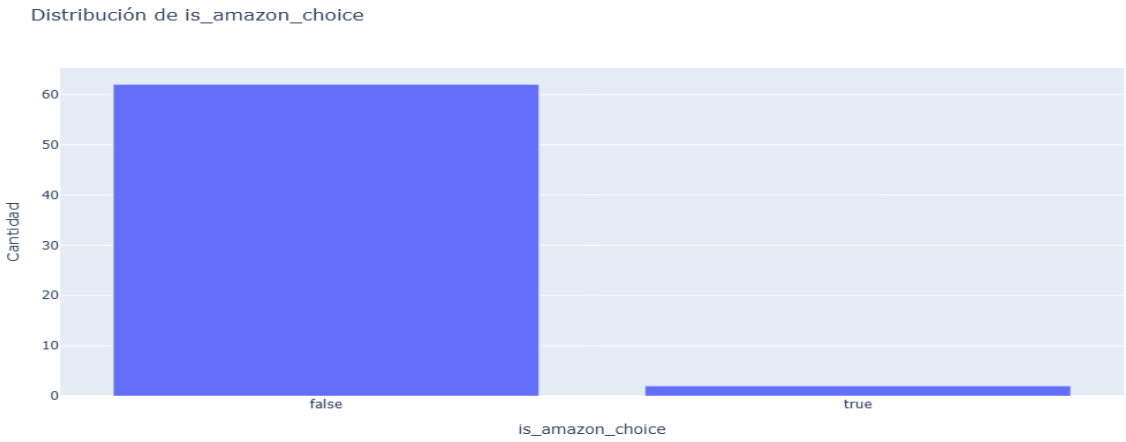
2. Centrar los datos, utilizar PCA. ¿Cuántos componentes se deben utilizar para mantener el 95% de las características?. Crear una tabla para mostrar las primeras 5 caras utilizando, la mean face + los datos reconstruidos utilizando la primera componente, después con 3 componentes, después con las primeras 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?

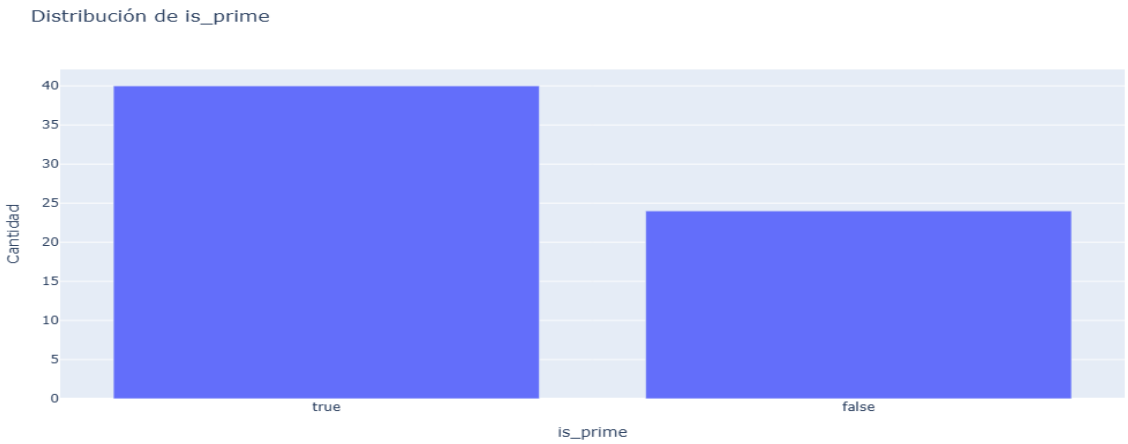


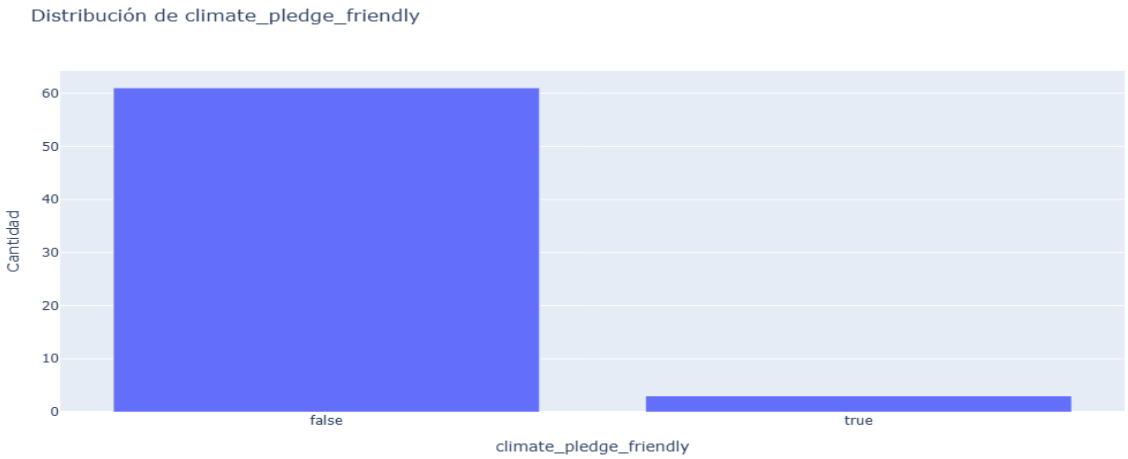
Se están tomando los componentes principales que aportan más características a la imagen, entre más componentes principales se utilicen, mejor será la calidad de la imagen reconstruida. En este sentido, el PCA optimiza el equilibrio entre eficiencia y precisión, haciendo posible trabajar con grandes volúmenes de datos de manera simplificada sin comprometer demasiado la calidad.

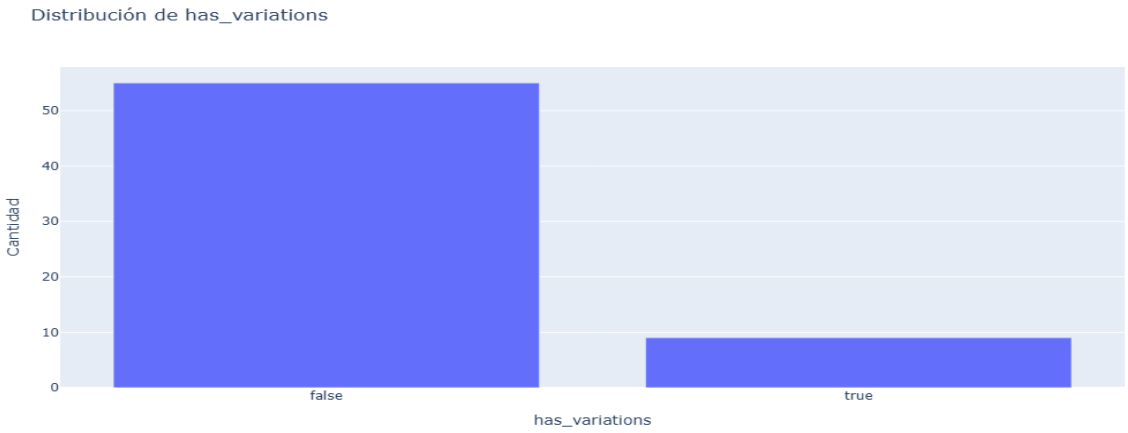
5. Utilizando el dataset del amazon data/amazon_products.csv crear: **Utilizar la librería de plotly**.
    5.1. Distribución de cada variables:
        5.1.1. Para las variables categóricas un gráfico de barras. Categoría numero de observaciones.


Distribución de is_best_seller
Cantidad
0
5
10
15
20
25
30
35
40
false
true
is_best_seller


Distribución de is_amazon_choice
Cantidad
0
10
20
30
40
50
60
false
true
is_amazon_choice

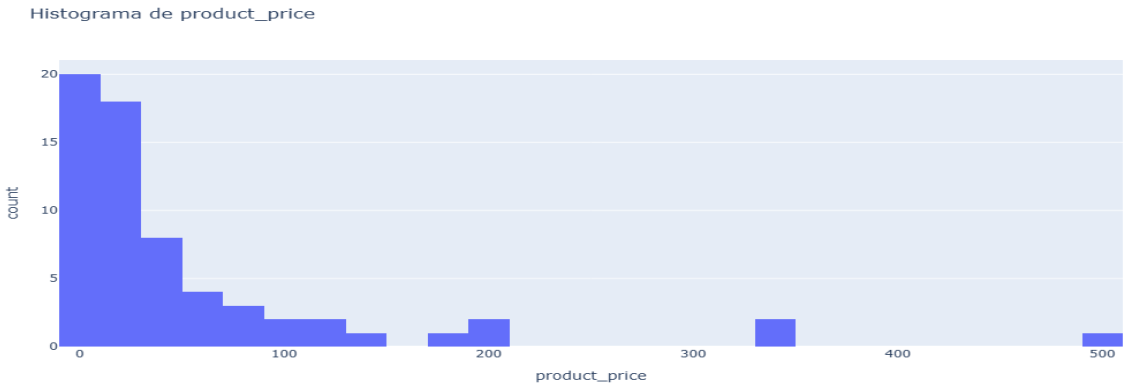

Distribución de is_prime
Cantidad
0
5
10
15
20
25
30
35
40
true
false
is_prime

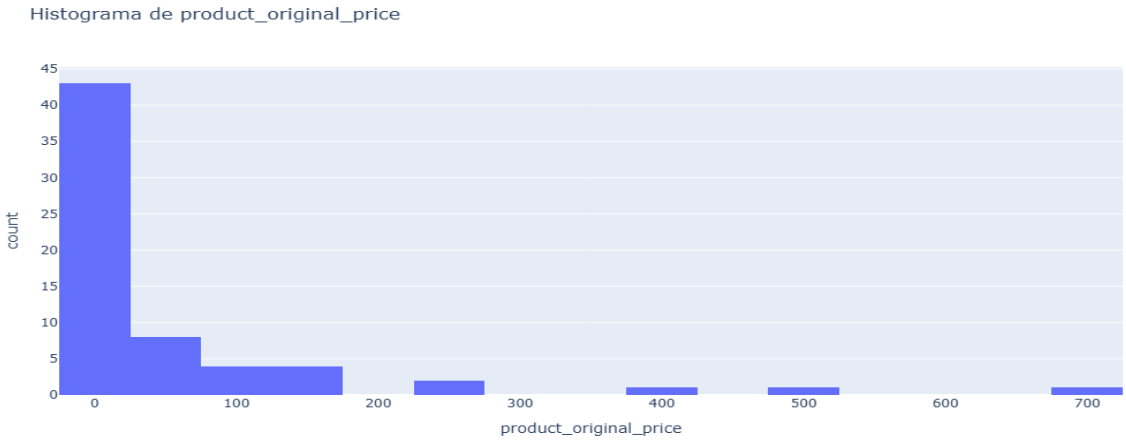

Distribución de climate_pledge_friendly
Cantidad
60
50
40
30
20
10
0
false
true
climate_pledge_friendly

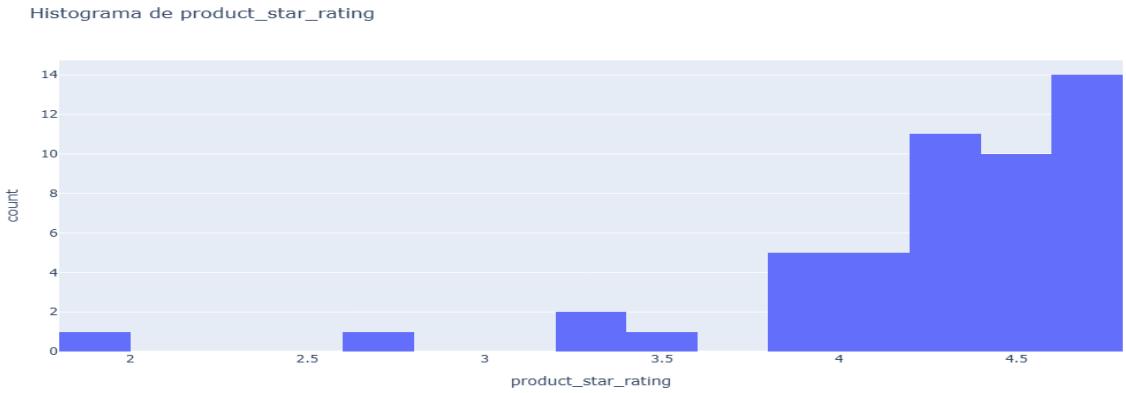

Distribución de has_variations
Cantidad
50
40
30
20
10
0
false
true
has_variations

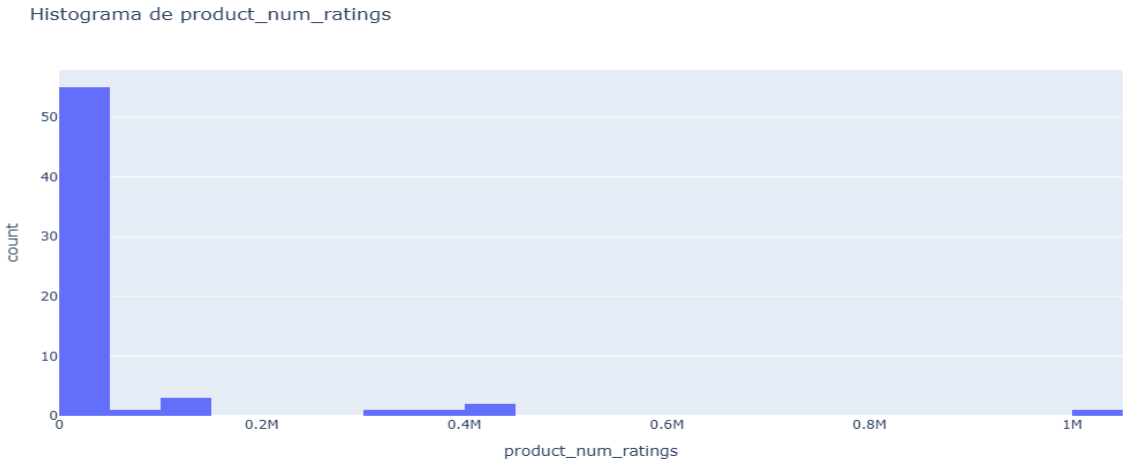

Distribución de currency
Cantidad
60
50
40
30
20
10
0
USD
currency

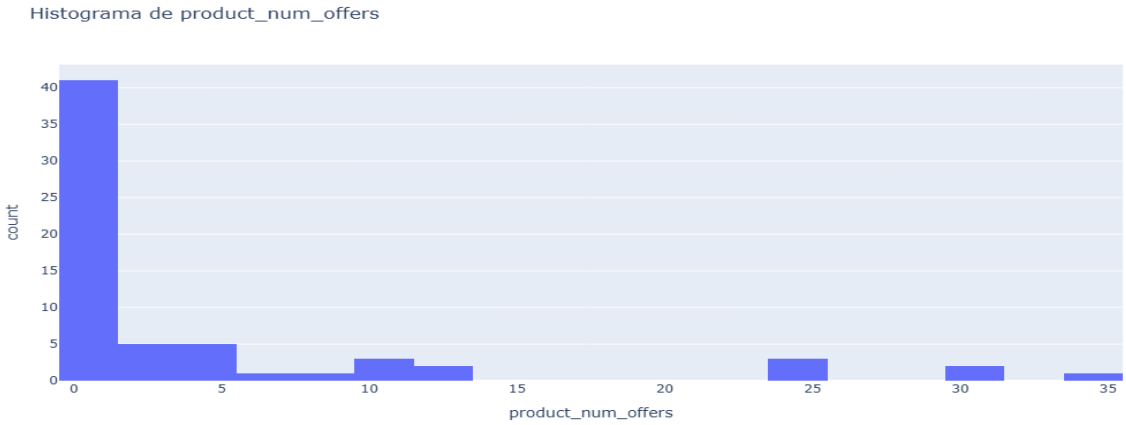
Para las variables numéricas crear histogramas. Listar los productos que están más lejos de 5 estándares de desviación, y serían considerados outliers. Hacer test de si es una distribución normal o no.



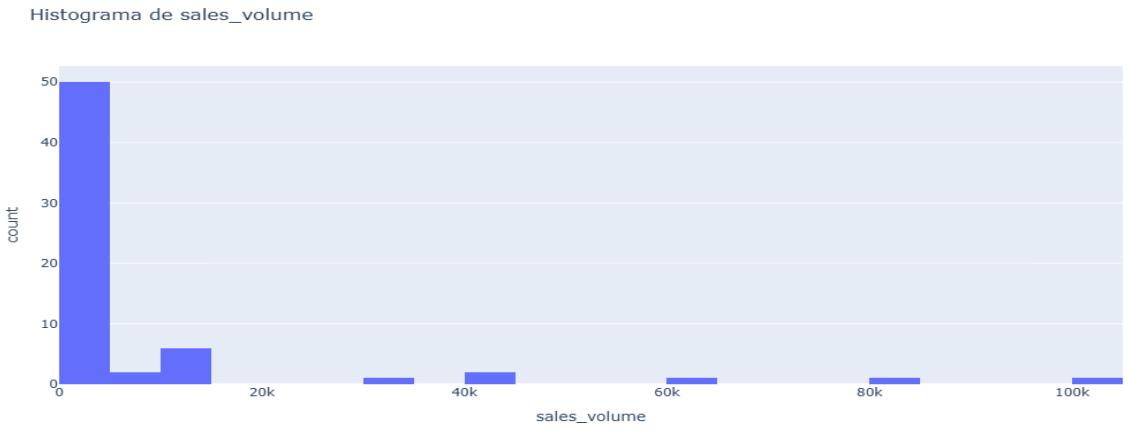


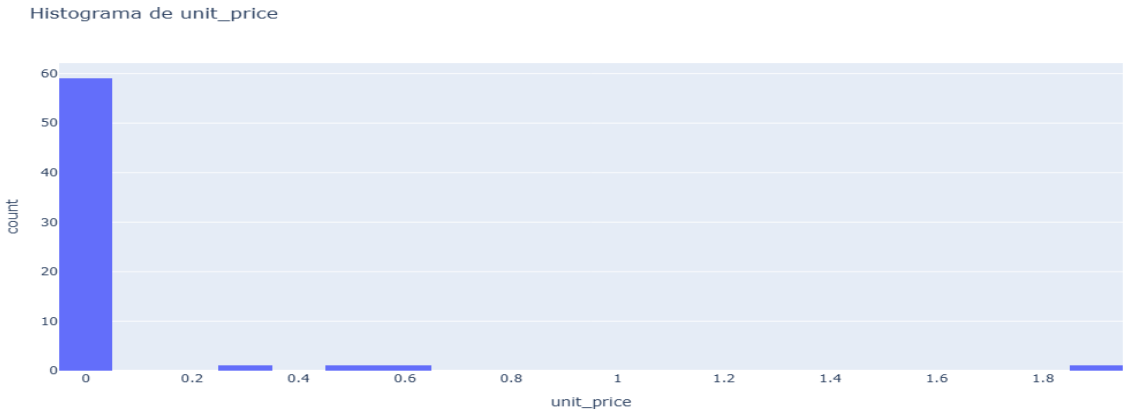


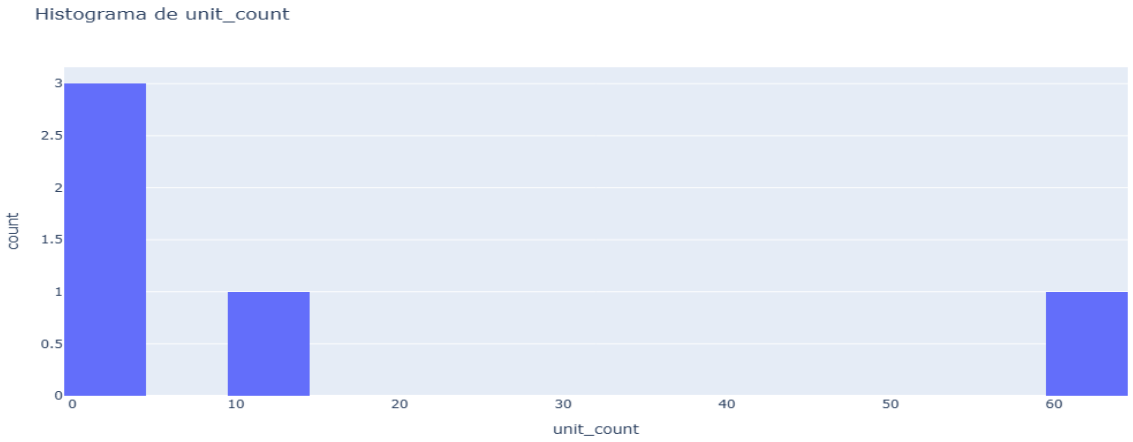

Histograma de product_num_ratings
count
product_num_ratings
0
10
20
30
40
50
0
0.2M
0.4M
0.6M
0.8M
1M


Histograma de product_num_offers
count
product_num_offers
0
5
10
15
20
25
30
35
40


Histograma de product_minimum_offer_price
count
product_minimum_offer_price
0
5
10
15
20
25
30
0
50
100
150
200
250
300
350
400


Histograma de sales_volume
count
sales_volume
0
10
20
30
40
50
0
20k
40k
60k
80k
100k


Histograma de unit_price
count
unit_price
0
10
20
30
40
50
60
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8


Histograma de unit_count
count
unit_count
0
0.5
1
1.5
2
2.5
3
0
10
20
30
40
50
60

Outliers en product_original_price:

asin	product_original_price
B0CGTD5KVT	699.0

Outliers en product_num_ratings:

asin	product_num_ratings
B07Y8SJGCV	1015448

Outliers en sales_volume:

asin	sales_volume
B0D5FZGY8W	100000.0

Outliers en unit_price:

asin	unit_price
B0CS12LZLS	1.91
B0CV4FQPY1	2.05

Test de Shapiro-Wilk para product_price: p-value = 1.573030422468821e-11
-> No es normal

Test de Shapiro-Wilk para product_original_price: p-value = 2.0741459881388667e-13
-> No es normal

Test de Shapiro-Wilk para product_star_rating: p-value = 1.6984222005346795e-07
-> No es normal

Test de Shapiro-Wilk para product_num_ratings: p-value = 4.9289427177438526e-15
-> No es normal

Test de Shapiro-Wilk para product_num_offers: p-value = 8.132068399110867e-13
-> No es normal

Test de Shapiro-Wilk para product_minimum_offer_price: p-value = 4.580891980997434e-11
-> No es normal

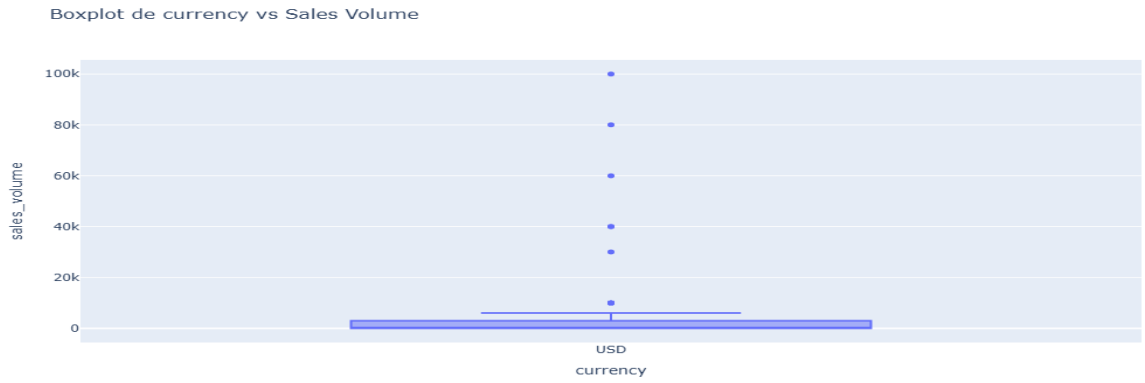
Test de Shapiro-Wilk para sales_volume: p-value = 2.5402611702158015e-14
-> No es normal

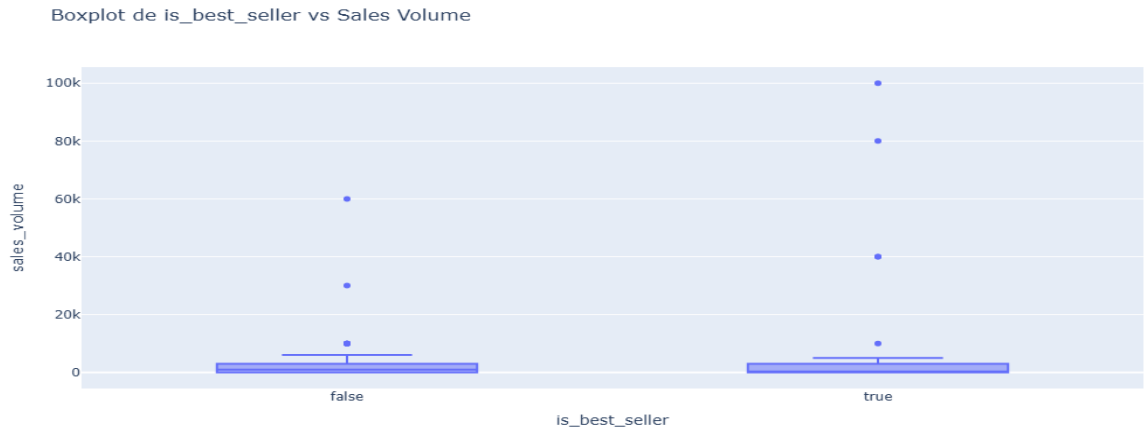
Test de Shapiro-Wilk para unit_price: p-value = 1.8278396662200194e-16
-> No es normal

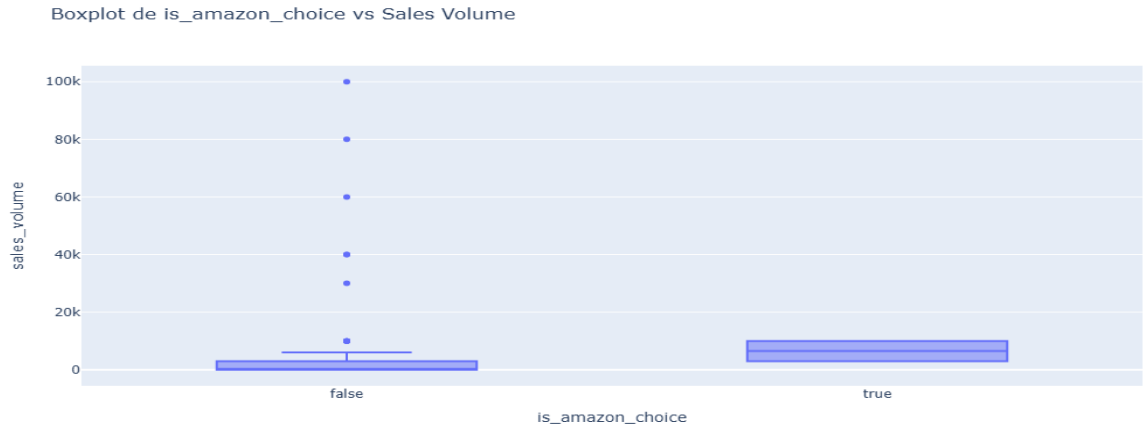
Test de Shapiro-Wilk para unit_count: p-value = 0.0067438087882866465
-> No es normal

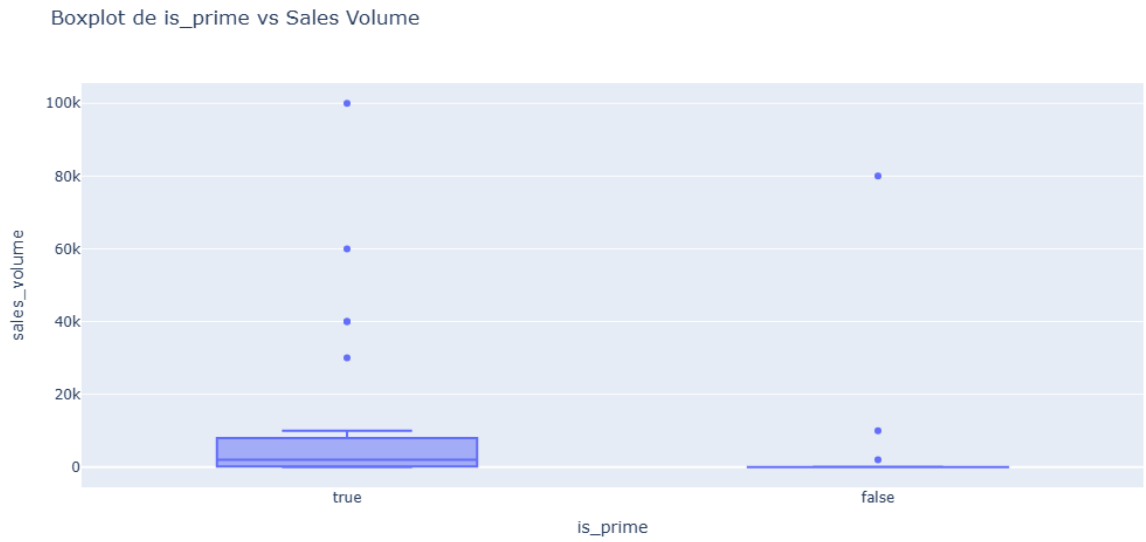
5.2. Gráfico de la relación de cada variable con respecto al sales_volume (convertir a numero):

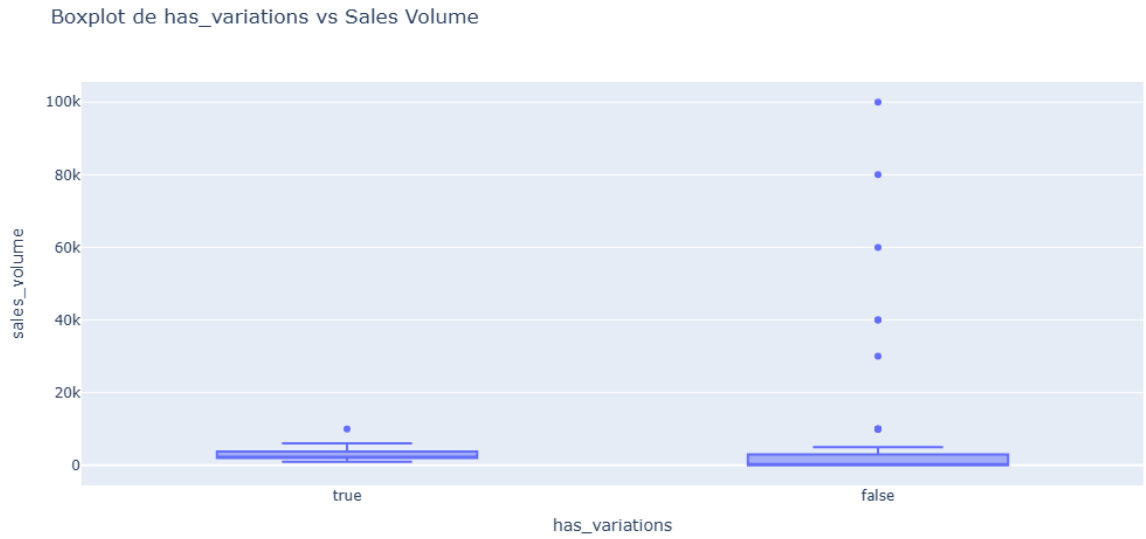
5.2.1. Variables categóricas debes crear un boxplot. Explique cómo interpreta el gráfico





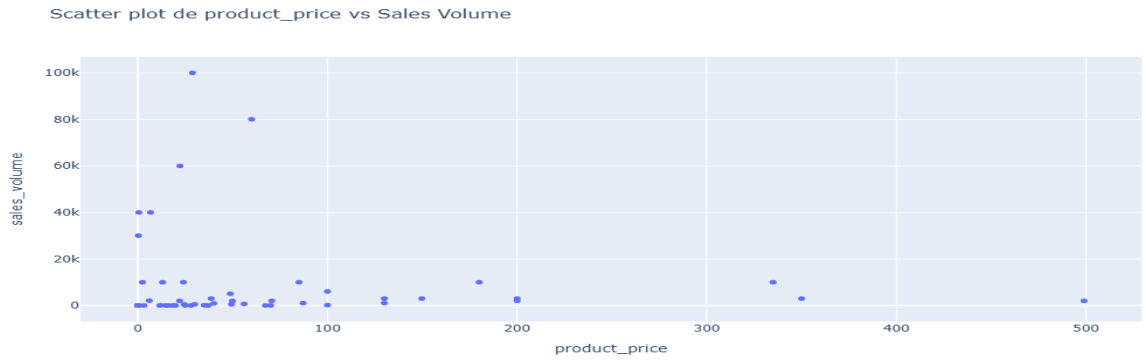


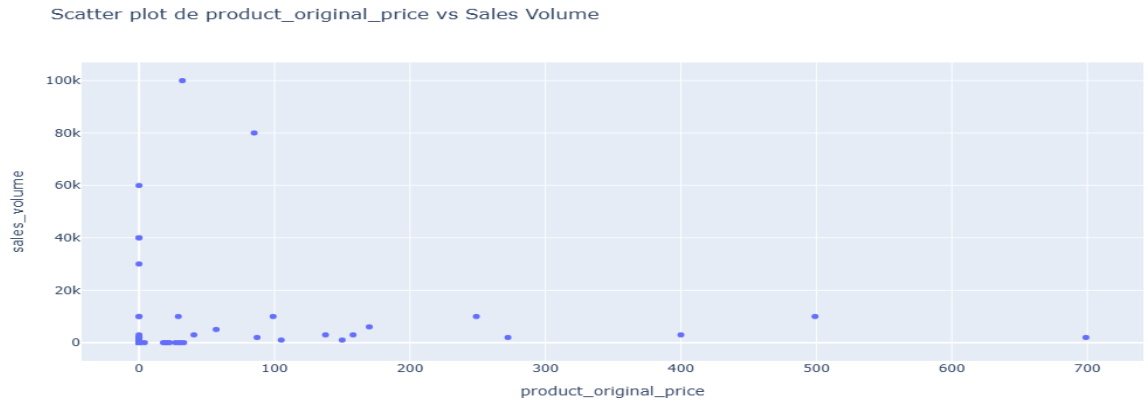


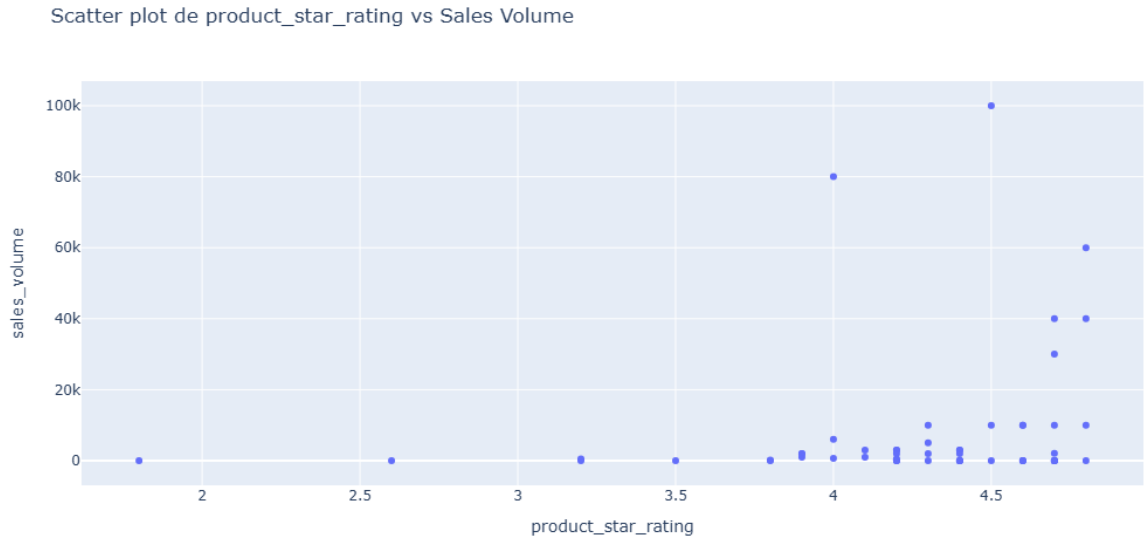


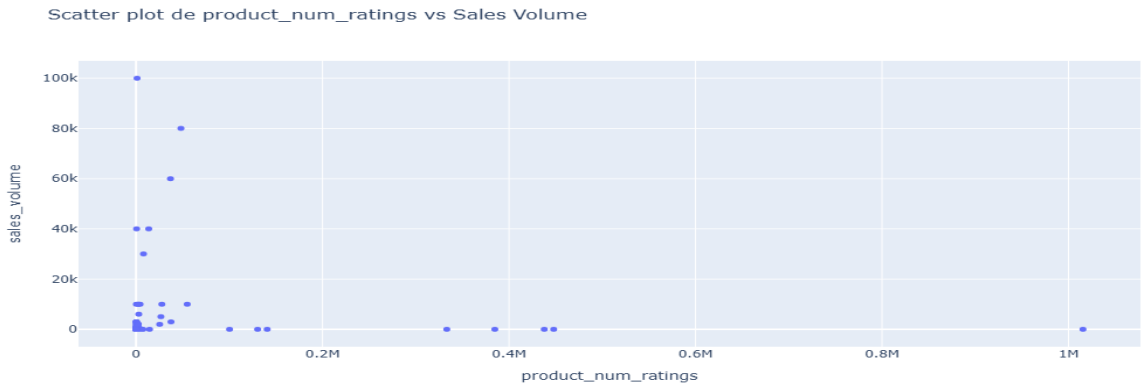
Comparando categorías, se puede identificar si unas presentan volúmenes de ventas consistentemente más altos o con menor dispersión, lo que indica una influencia del factor categórico sobre el desempeño de ventas (productos que son amigables con el medio ambiente y lo que son Amazon Choice).

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

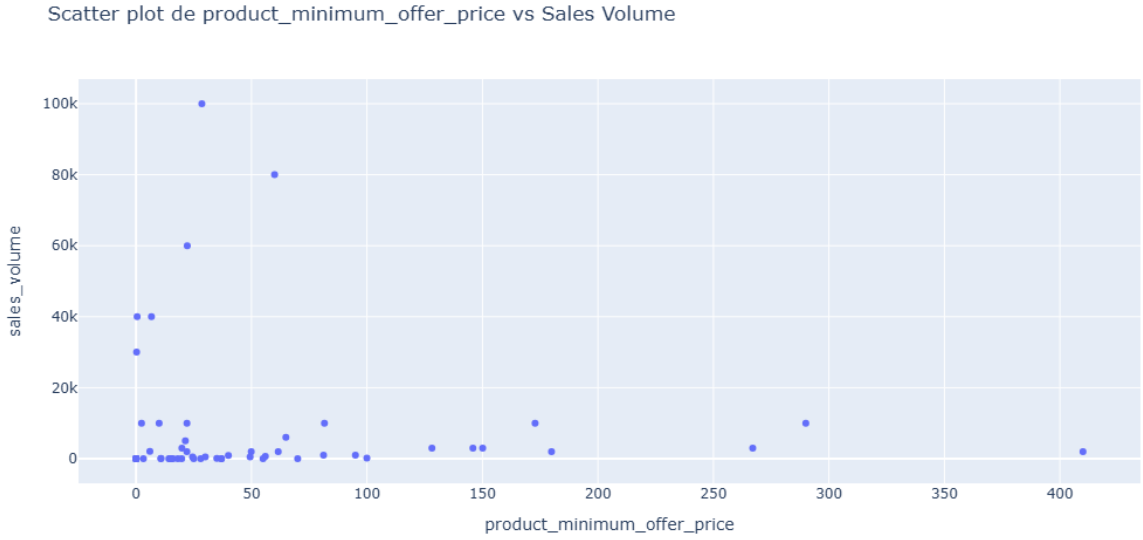


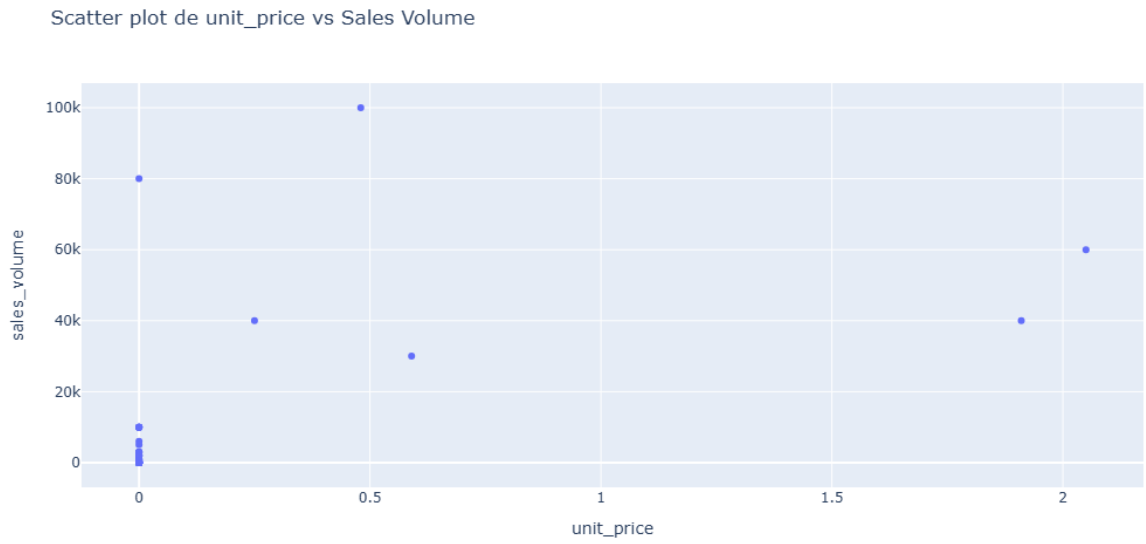


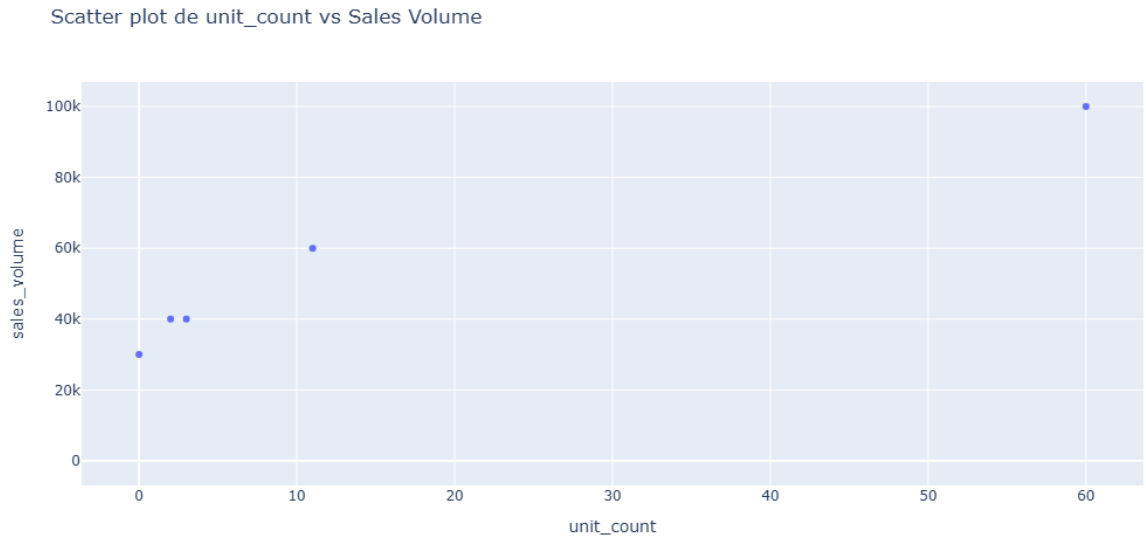



Scatter plot de product_num_ratings vs Sales Volume
sales_volume
100k
80k
60k
40k
20k
0
0
0.2M
0.4M
0.6M
0.8M
1M
product_num_ratings


Scatter plot de product_num_offers vs Sales Volume
sales_volume
100k
80k
60k
40k
20k
0
0
5
10
15
20
25
30
35
product_num_offers


Scatter plot de product_minimum_offer_price vs Sales Volume
sales_volume
100k
80k
60k
40k
20k
0
0
50
100
150
200
250
300
350
400
product_minimum_offer_price

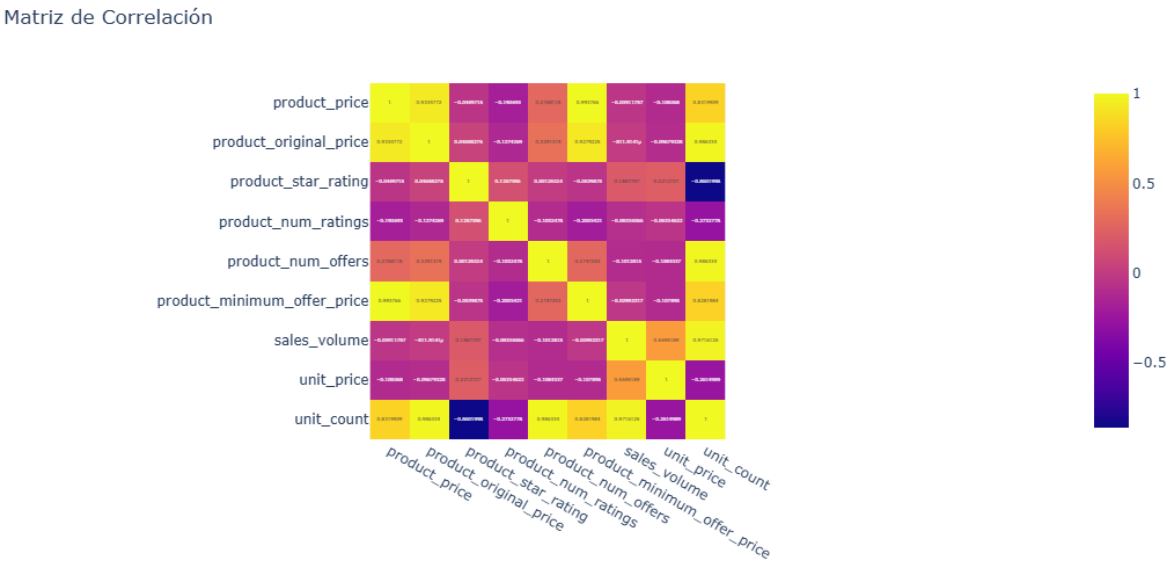




Al interpretar el gráfico, se observa si existe una tendencia en el que el volumen de ventas es alto para las vairables numéricas con un valor bajo

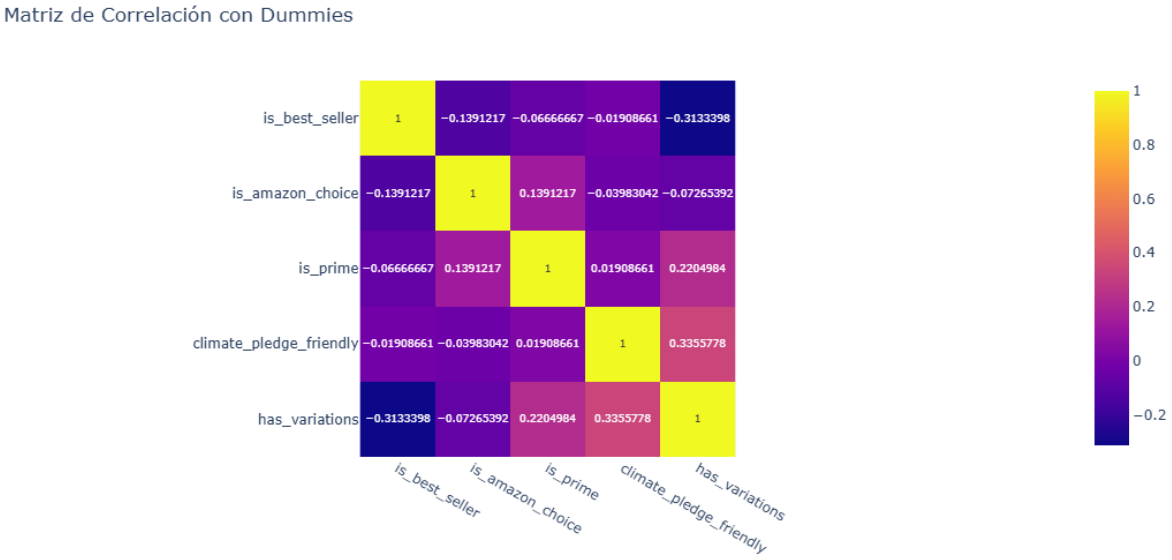
5.3. Matriz de correlación.

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.



El precio unitario y el conteo por unidad son las variables que capturan la mayor parte de la variabilidad del volumen de ventas. El coeficiente es positivo cuando, al aumentar la variable independiente, la dependiente también aumenta; es negativo cuando, al crecer la independiente, la dependiente disminuye.

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?



La mayor correlación entre las variables dummy es de 0.3355778 y se presenta entre "climate_pledge_friendly" y "has_variations".

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.

Coeficiente b0 (intercepto): 4.2559
Coeficiente b1 (pendiente): 0.0000
Se observa que la línea de regresión, obtenida a partir del 80% de los datos de entrenamiento, predice de manera precisa los valores del conjunto de testing. Los datos de testing se ajustan bien a la tendencia marcada por el modelo, lo que indica una buena capacidad de generalización y desempeño del modelo.