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 que 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): **data/cities.csv**

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 $\sum \frac{X_i}{n}$

PIB (Billones de USD)	8.75
Población (Millones)	0.731
Tasa de desempleo (%)	13.833
Edad promedio	29.233
Mujeres (%)	51.5
Hombres (%)	48.5
Presupuesto (Billones de USD)	1.65

Mediana $\frac{n+1}{2}$

PIB (Billones de USD)	2.65
Población (Millones)	0.39
Tasa de desempleo (%)	13.45
Edad promedio	29

Mujeres (%)	51
Hombres (%)	49
Presupuesto (Billones de USD)	0.6

Desviación estándar $\sqrt{\frac{\sum (X_i - \widetilde{X})^2}{n-1}}$

PIB (Billones de USD)	19.914
Población (Millones)	1.352
Tasa de desempleo (%)	2.945
Edad promedio	2.238
Mujeres (%)	0.776
Hombres (%)	0.776
Presupuesto (Billones de USD)	3.451

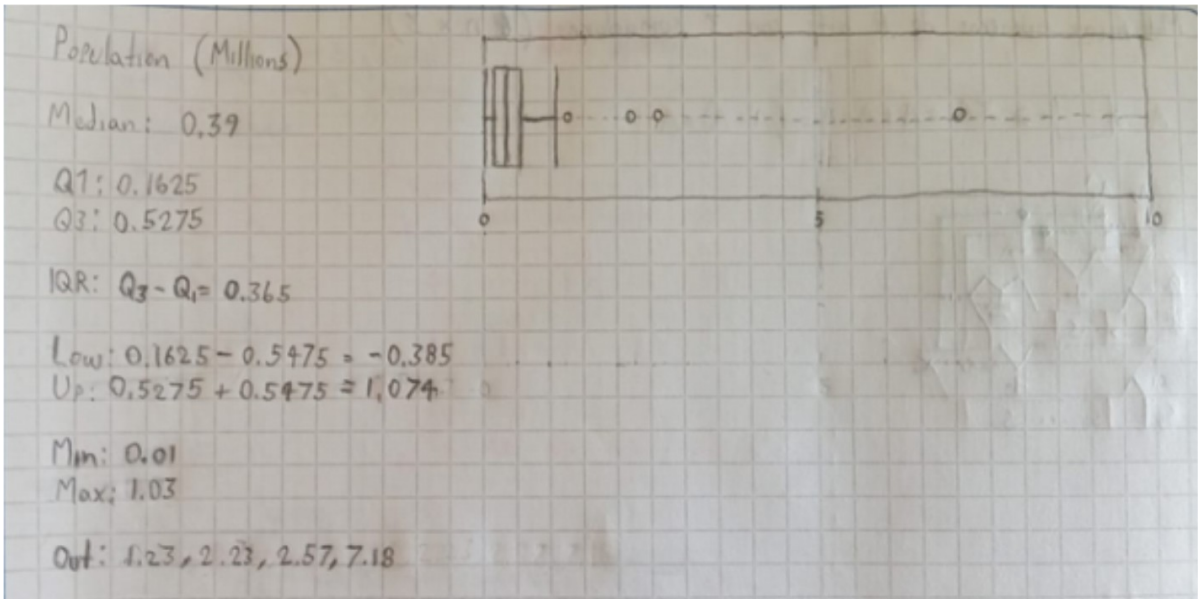
Moda

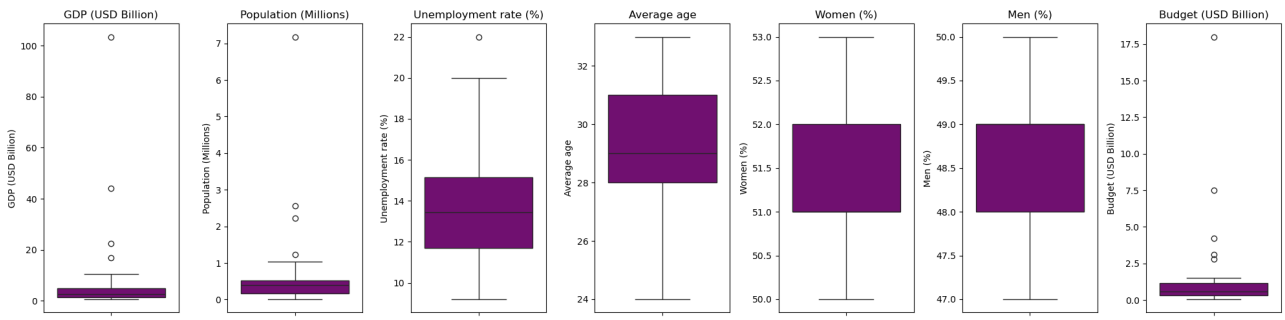
Ciudad	Arauca
PIB (Billones de USD)	0.6
Población (Millones)	0.01
Tasa de desempleo (%)	9.2
Edad promedio	29
Mujeres (%)	51
Hombres (%)	49
Presupuesto (Billones de USD)	0.1
etiqueta	2
entrenamiento	Sí

Repeticiones de la moda

Ciudad	1
PIB (Billones de USD)	1
Población (Millones)	2
Tasa de desempleo (%)	1
Edad promedio	6
Mujeres (%)	14
Hombres (%)	14
Presupuesto (Billones de USD)	3
etiqueta	10
entrenamiento	21

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

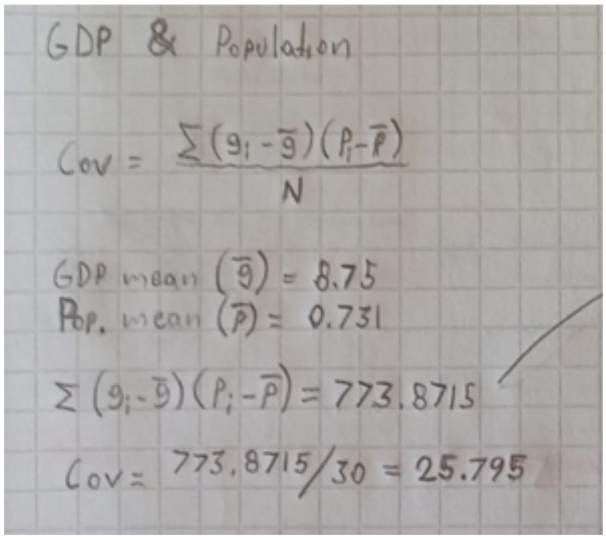




1.3. Cual es la covarianza entre las 2 variables X1, X2

X1 = PIB
X2 = Población

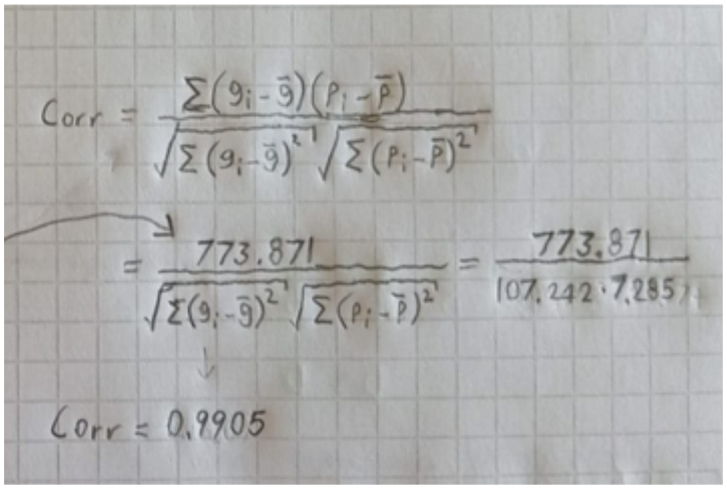
$$\frac{\sum (X_i - \widetilde{X})(Y_i - \widetilde{Y})}{N}$$



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

x1 = PIB
x2 = Población

$$\frac{\sum (X_i - \widetilde{X})(Y_i - \widetilde{Y})}{\sqrt{\sum (X_i - \widetilde{X})^2 \sum (Y_i - \widetilde{Y})^2}}$$



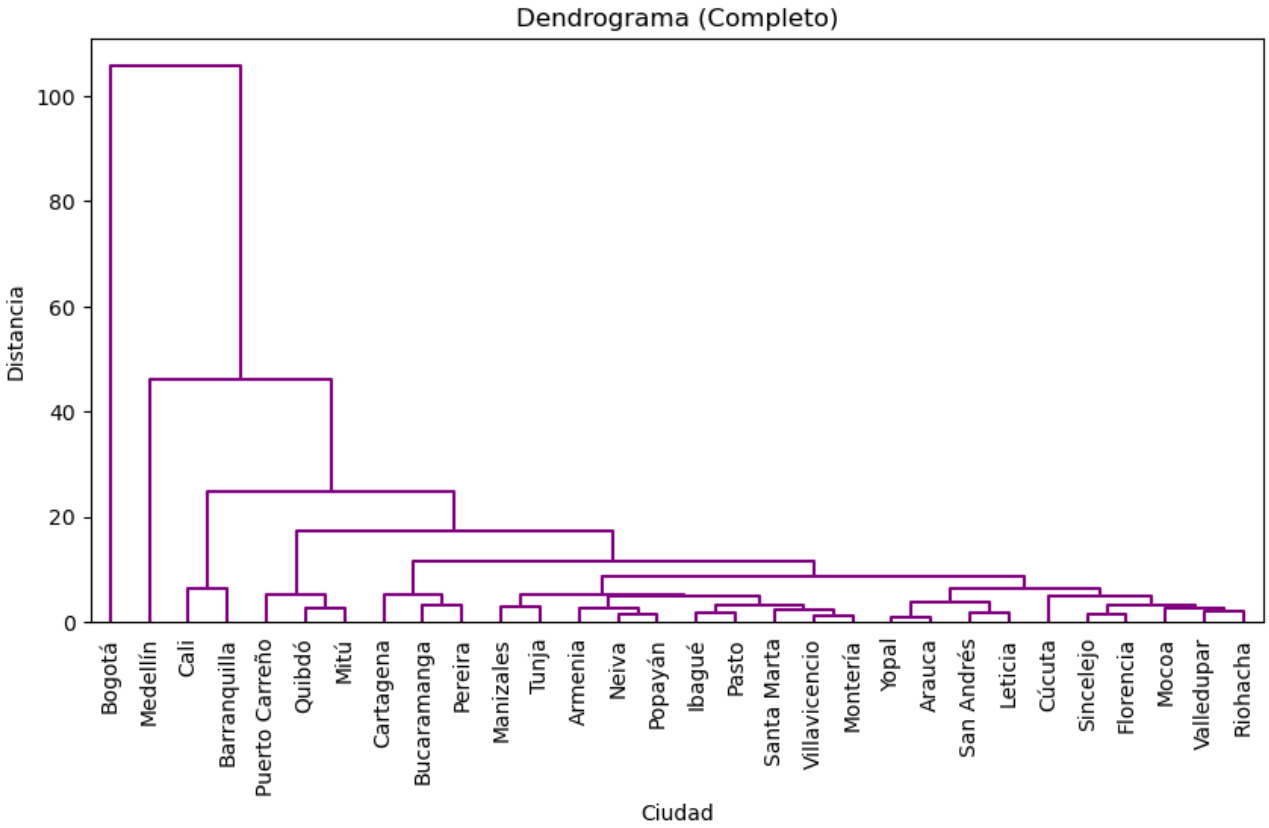
1.5. Explica la relación entre covarianza y correlación.

Ambas son útiles para examinar la relación entre dos variables. La covarianza indica la dirección de la relación lineal entre ambas variables. La correlación indica algo similar, pero además dice con qué fuerza las dos variables están relacionadas entre sí. En términos simples, podría decirse que la correlación es una versión estandarizada de la correlación.

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).

Ciudad	Etiqueta
Bogotá	2
Medellín	2
Cali	0
Barranquilla	3
Cartagena	3
Pereira	0
Santa Marta	3
Manizales	0
Montería	3
Valledupar	3
Neiva	0
Popayán	0
Armenia	0
Sincelejo	1
Tunja	0
Florencia	1
Quibdó	1
San Andrés	3
Yopal	3
Leticia	3
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

	PIB	Población
PIB	396.584	26.685
Población	26.685	1.830

2.2. Cuales son los eigenvalues

Eigenvalue 1: 398.380
Eigenvalue 2: 0.034

2.3. Cuál es la varianza explicada por el eigenvalue.

Varianza explicada por el CP1: 0.999
Varianza explicada por el CP2: 8.634e-5

2.4. Cual es el valor del eigenvector

Eigenvector 1: < 0.998, -0.067 >
Eigenvector 2: < 0.067, 0.998 >

2.5. Cuál es la matriz proyectada.

Componente principal 1	Componente principal 2
94.97	0.07
35.39	-0.54
13.72	0.58
8.07	-0.04
1.77	0.18
-1.46	-0.05
-2.56	-0.08
-3.64	0.27
-3.95	0.06
-4.75	0.11
-4.96	0.03
-5.25	0.12
-5.56	0.09
-5.75	0.15
-5.95	0.14
-6.26	0.04
-6.46	0.03
-6.66	0.02
-6.77	0
-6.97	-0.01
-7.07	-0.06
-7.27	-0.02
-7.47	-0.1
-7.58	-0.14
-7.67	-0.07
-7.78	-0.16

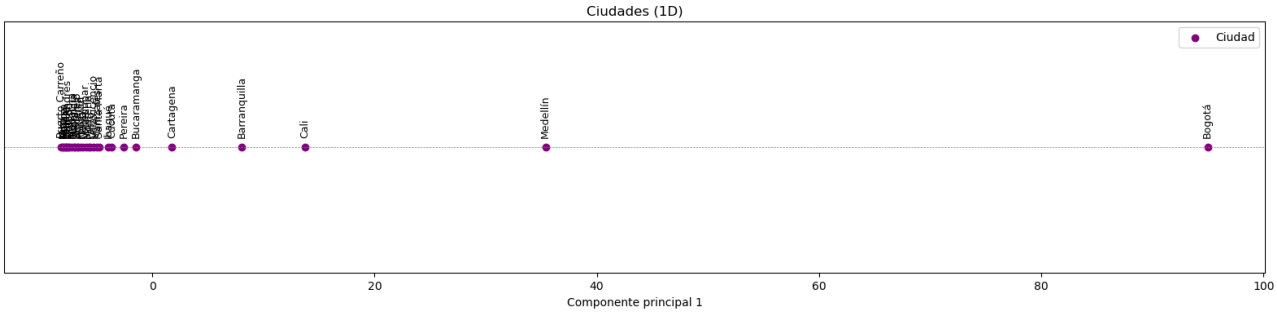
-7.88	-0.12
-7.98	-0.16
-8.08	-0.18
-8.18	-0.17

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

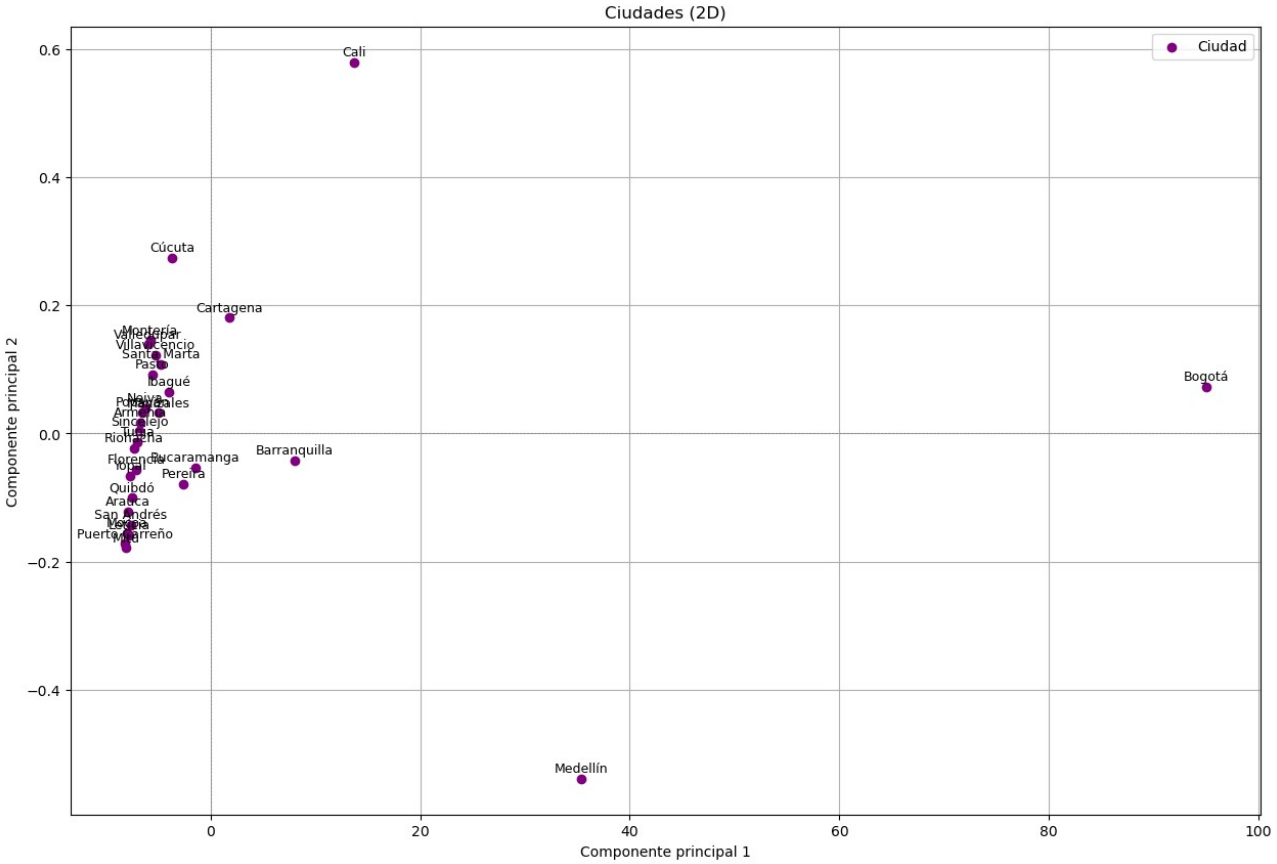
PIB	Población
8.53	7.11
8.71	3.11
8.68	1.65
8.73	1.27
8.73	0.85
8.76	0.63
8.76	0.56
8.74	0.49
8.75	0.47
8.75	0.41
8.76	0.4
8.75	0.38
8.76	0.36
8.75	0.34
8.75	0.33
8.76	0.31
8.76	0.3
8.76	0.28
8.77	0.28
8.77	0.26
8.77	0.26
8.77	0.24
8.77	0.23
8.78	0.22
8.77	0.22

8.78	0.21
8.78	0.2
8.78	0.2
8.78	0.19
8.78	0.18

2.7. Pintar todas las ciudades en 1 dimensión.

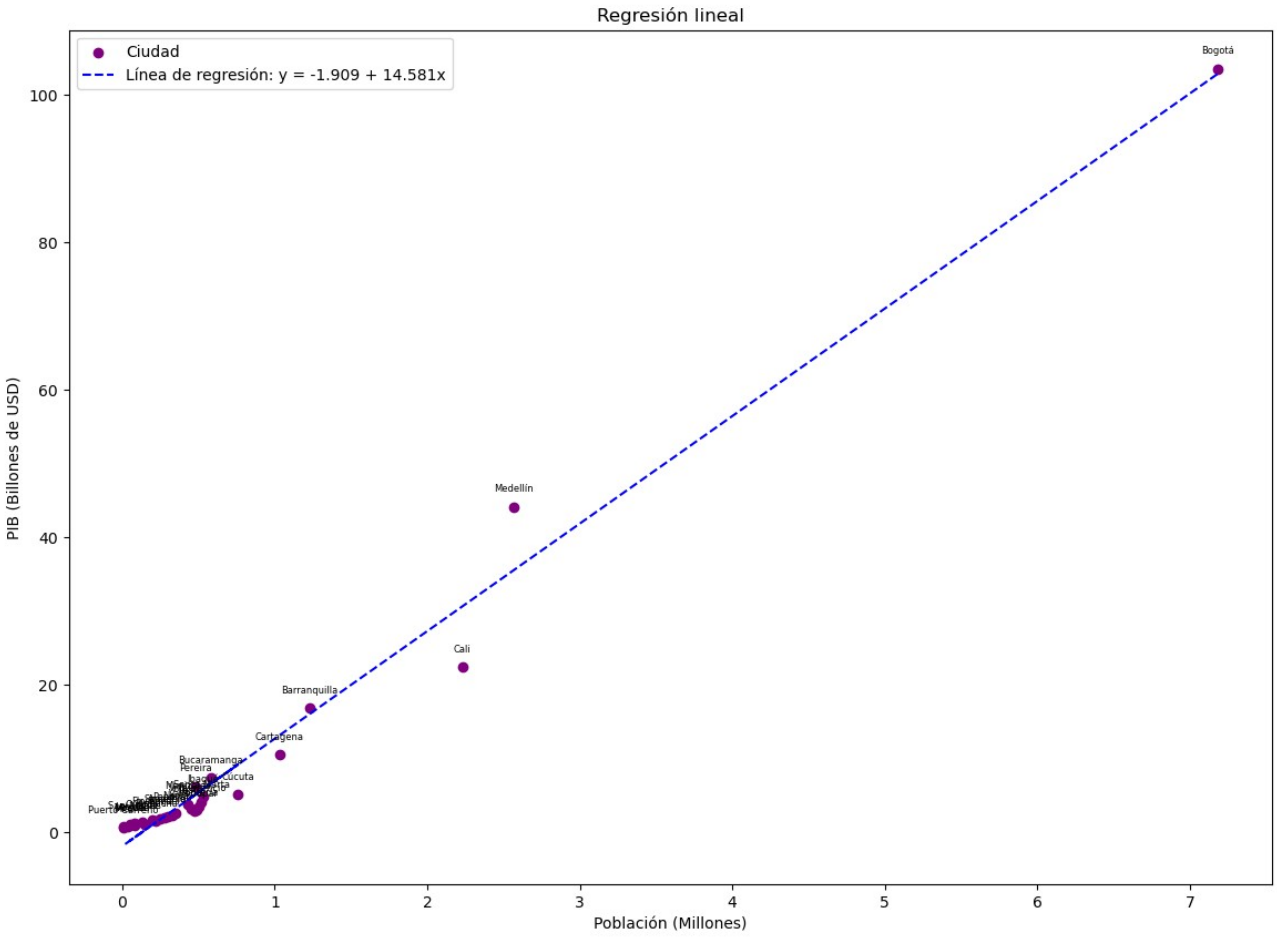


2.8. Utilizar python para pintar todas las ciudades en 2 dimensiones



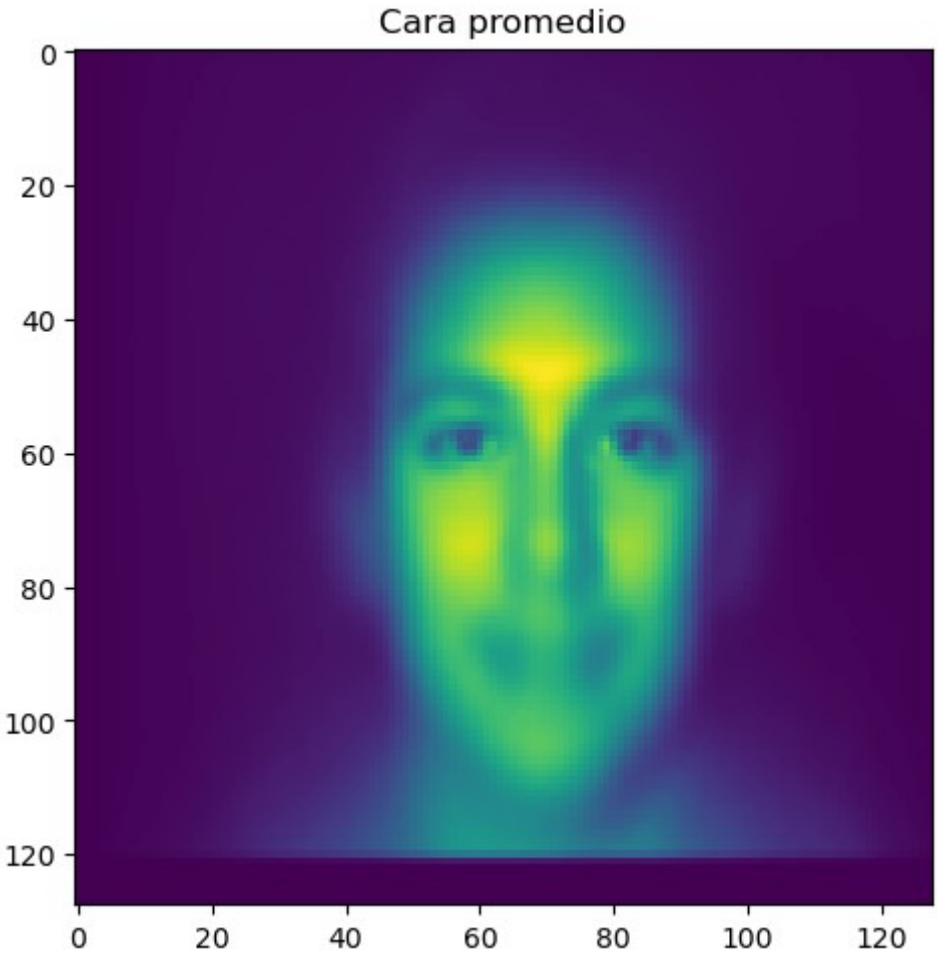
3. Regression. Utiliza las variables GDP (USD Billion) y Population (Millions) para crear una regresión. X es la población, y es el GDP.

B1: 14.581
B0: -1.909



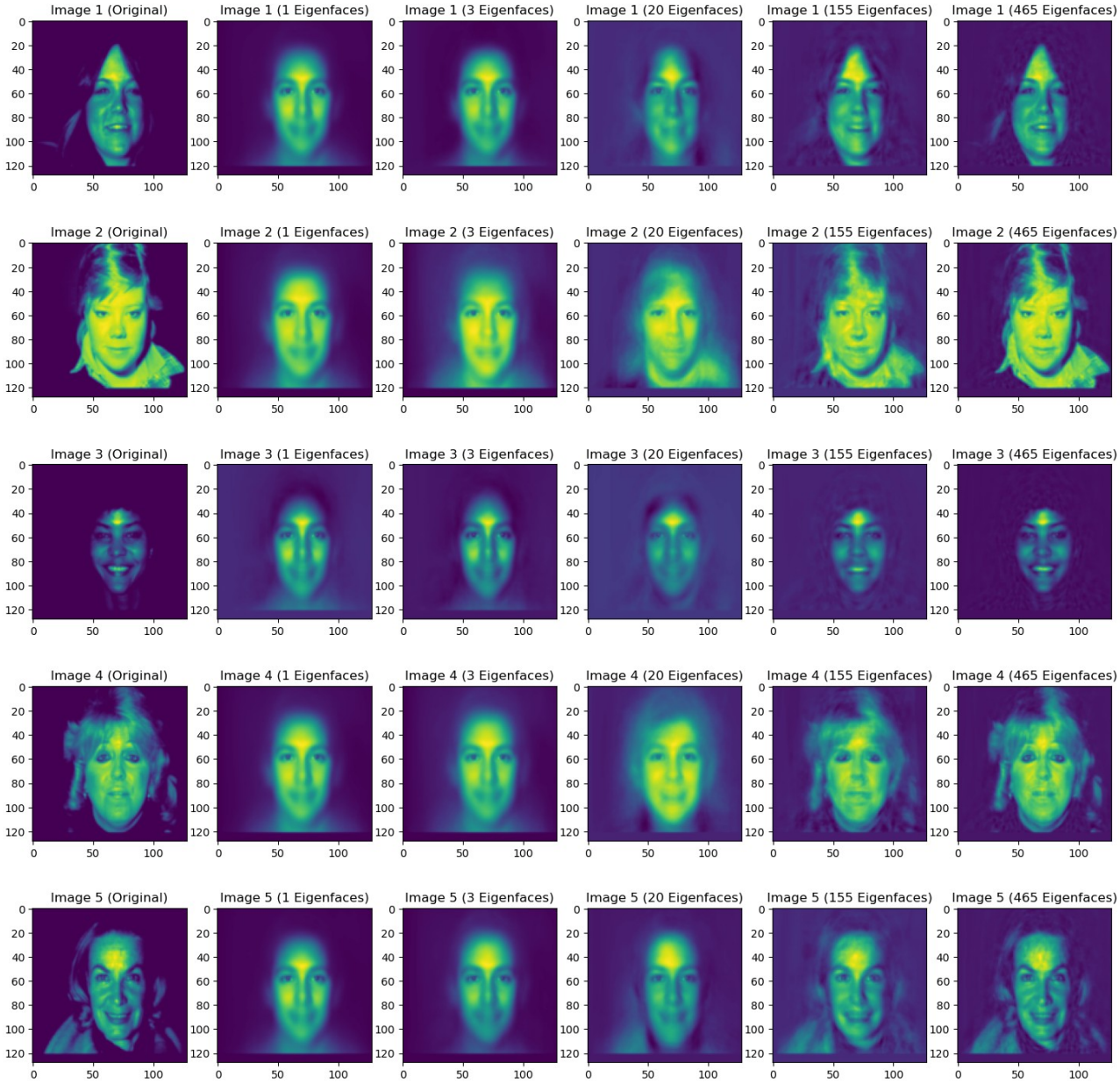
4. PCA

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



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?

Para mantener el 95% de la varianza se necesitan 155 componentes.
Para mantener el 99% de la varianza se necesitan 465 componentes.

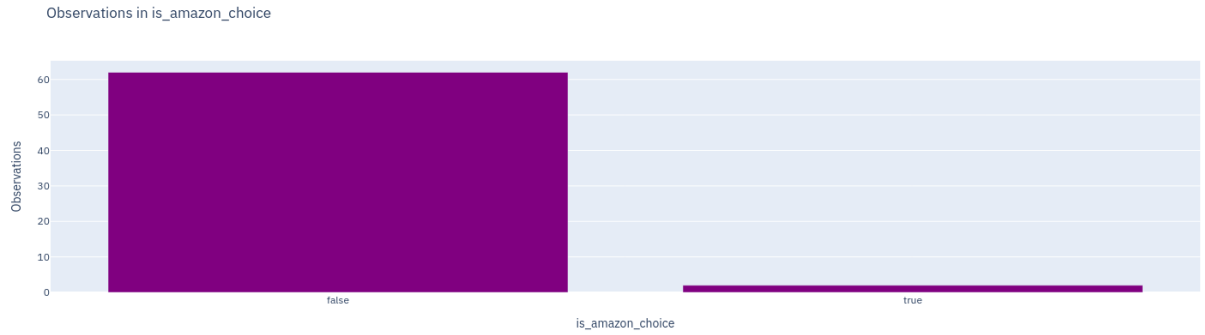


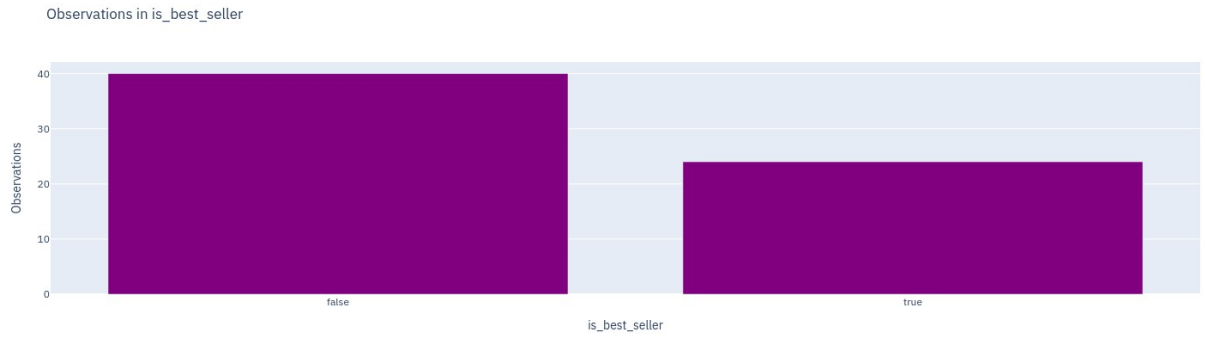
Al aplicar una reducción de dimensionalidad, se están tomando los componentes que aportan más características a la imagen, por lo tanto, entre más componentes se utilicen mejor será la calidad de la imagen.

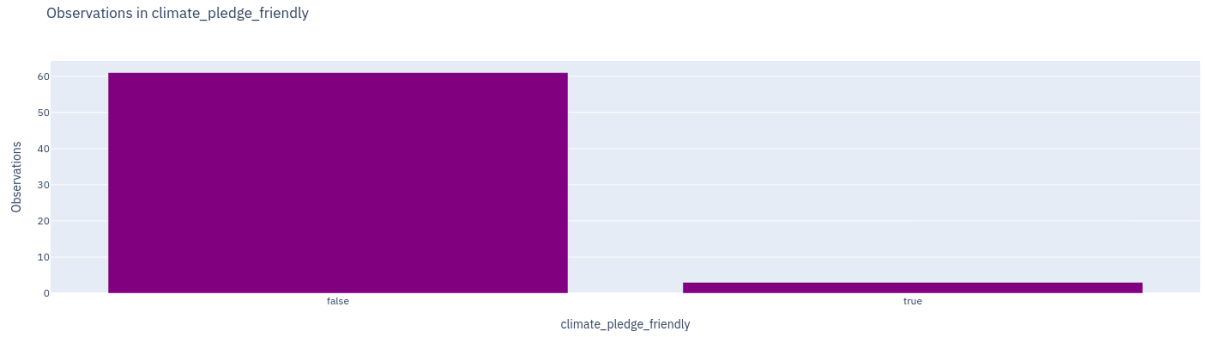
5. Utilizando el dataset del amazon data/amazon_products.csv crear: Utilizar la librería de plotly.

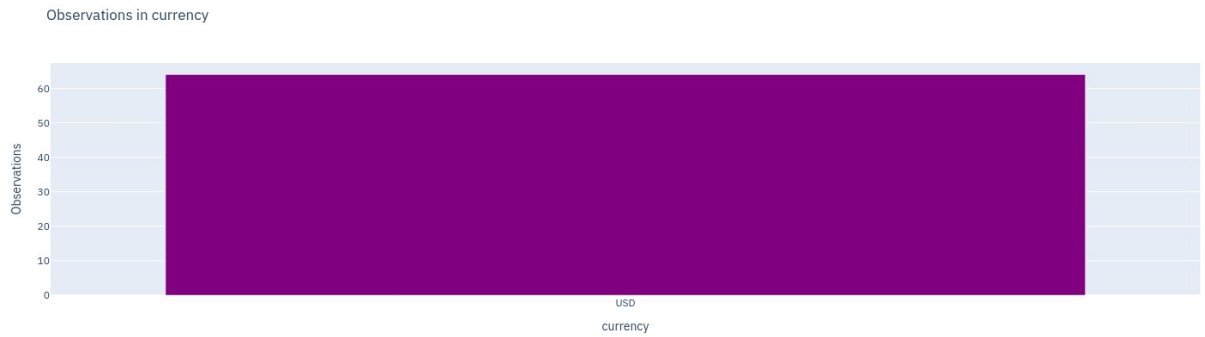
5.1. Distribución de cada variables:

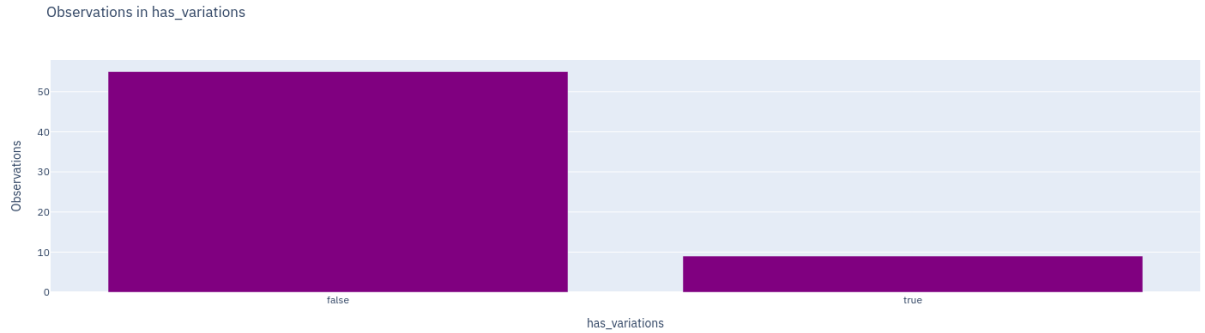
5.1.1. Para las variables categóricas un gráfico de barras. Categoría numero de observaciones.

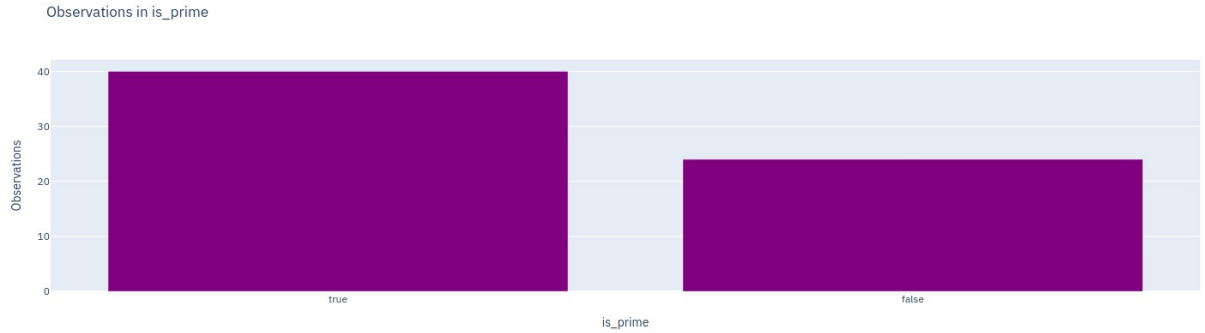




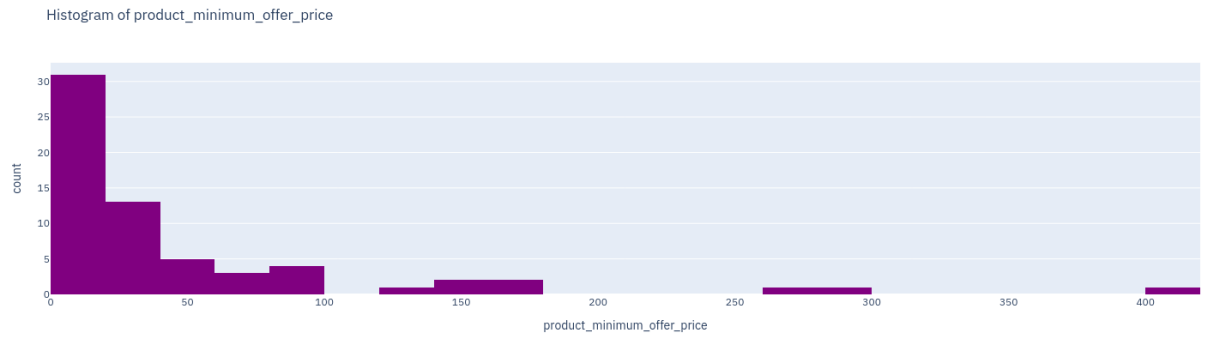


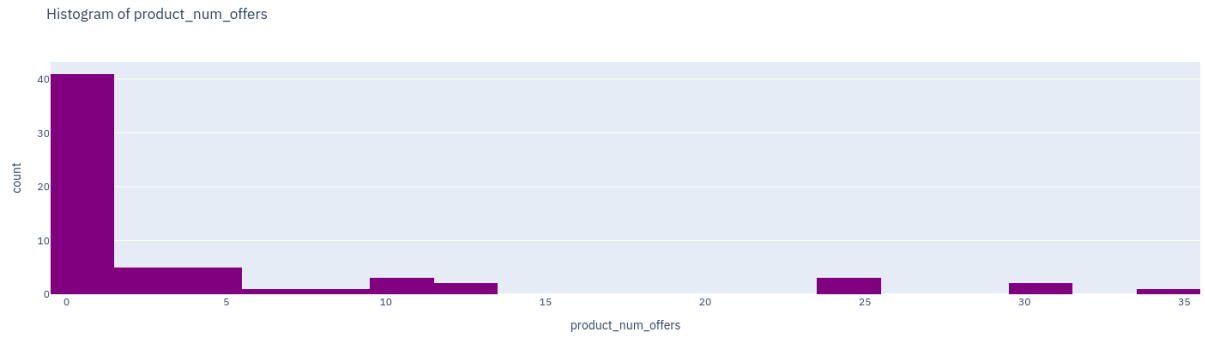




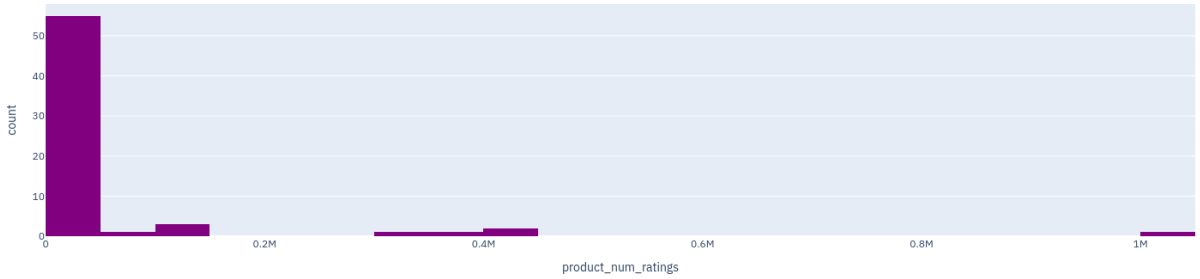


5.1.2. Para las variables numéricas crear histogramas. Listar los productos que están más lejos de 5 estándares de desviación, y serían considerados outliers. Hacer test de si es una distribución normal o no.

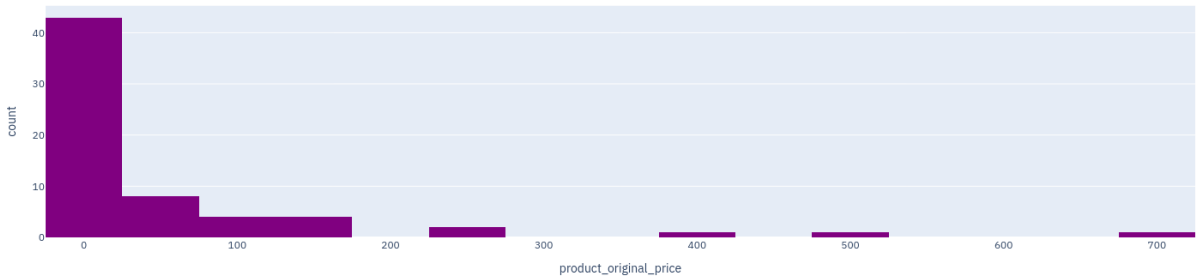




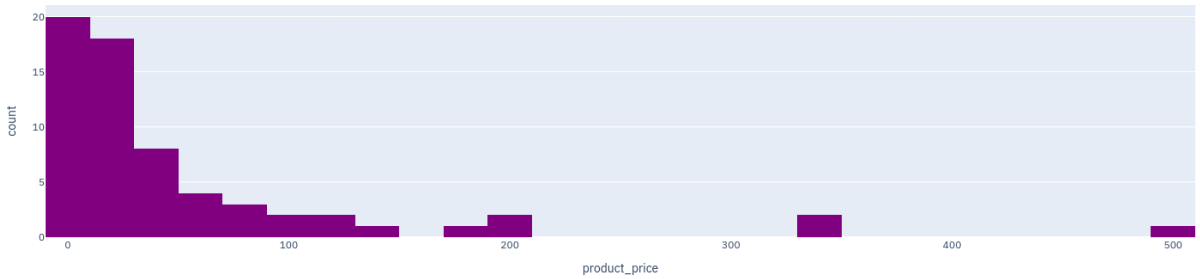
Histogram of product_num_ratings



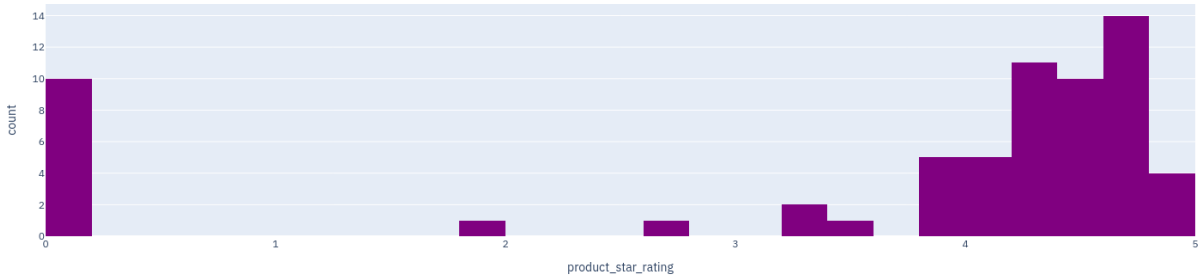
Histogram of product_original_price

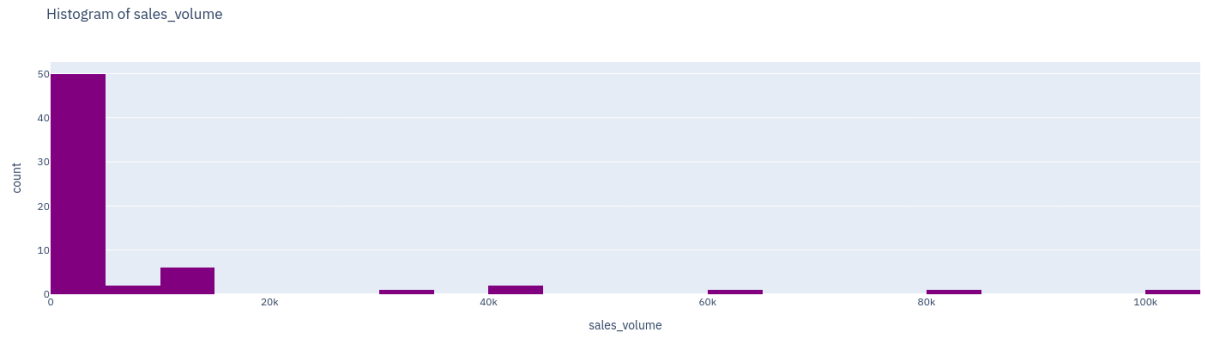


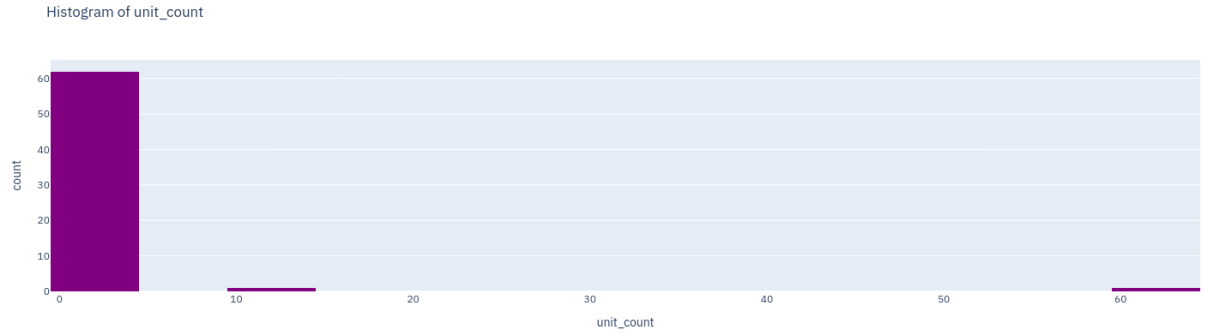
Histogram of product_price

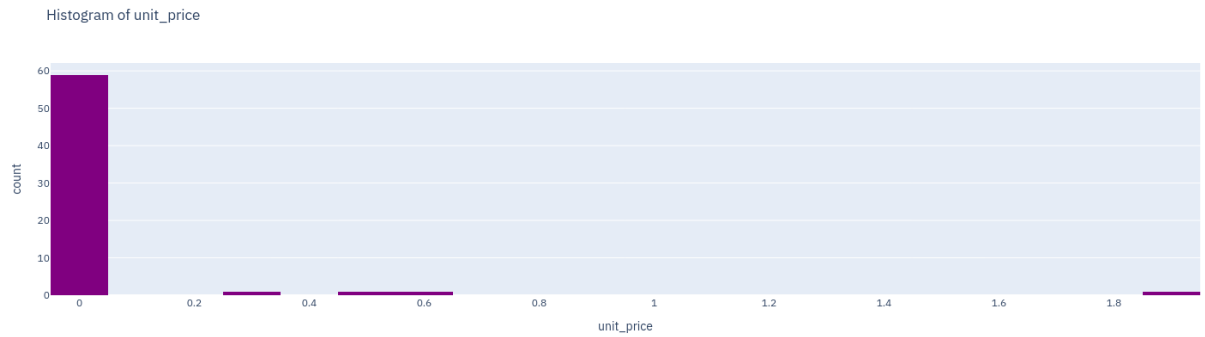


Histogram of product_star_rating









Outliers en product_original_price:

	asin	product_original_price
13	B0CGTD5KVT	699.0

Outliers en product_num_ratings:

	asin	product_num_ratings
55	B07Y8SJGCV	1015448

Outliers en sales_volume:

	asin	sales_volume
29	B0D5FZGY8W	100000

Outliers en unit_price:

	asin	unit_price
25	B0CS12LZLS	1.91
37	B0CV4FQPY1	2.05

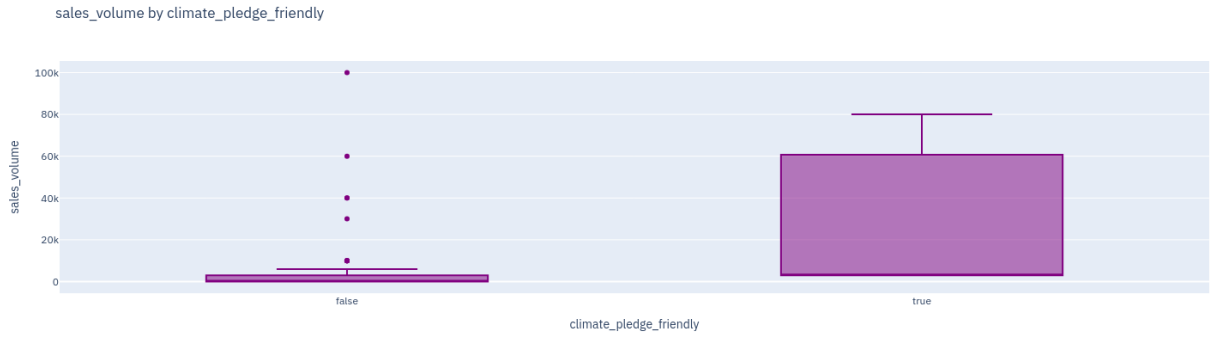
Outliers en unit_count:

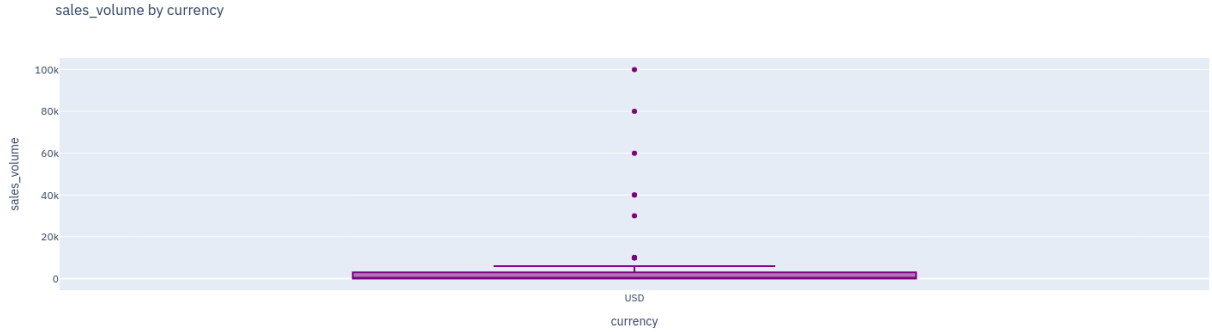
	asin	unit_count
29	B0D5FZGY8W	60.0

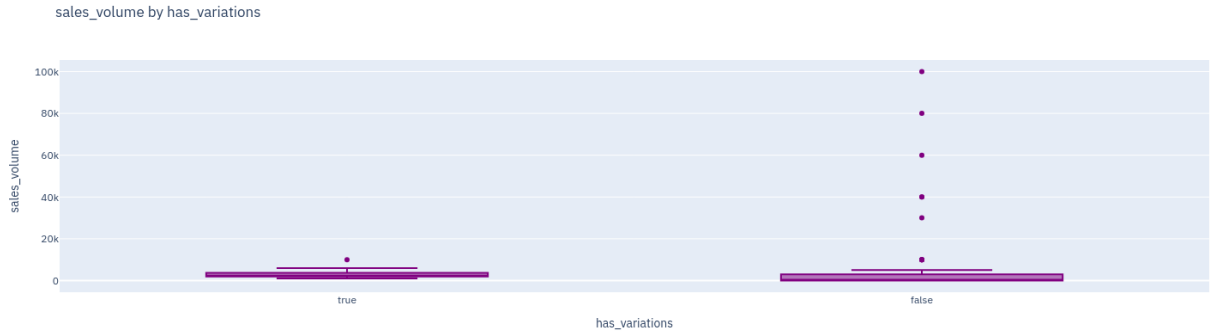
Ninguna distribución es normal según la prueba de Shapiro-Wilk.

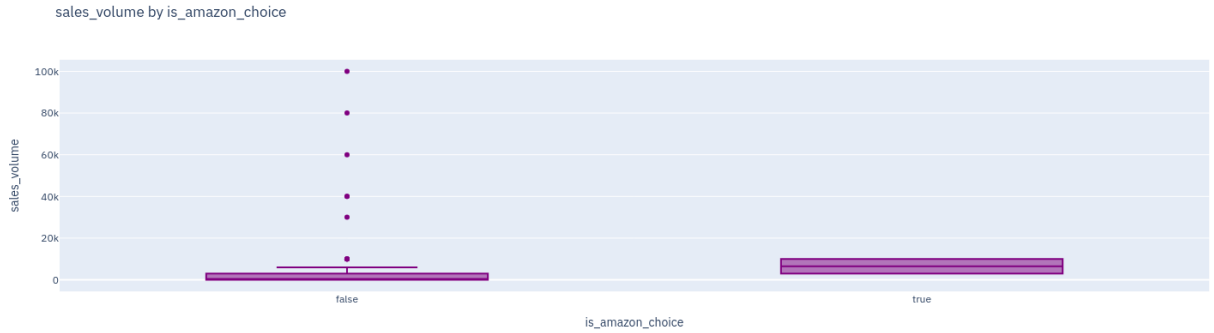
5.2. Gráfico de la relación de cada variable con respecto al sales_volume (convertir a numero):

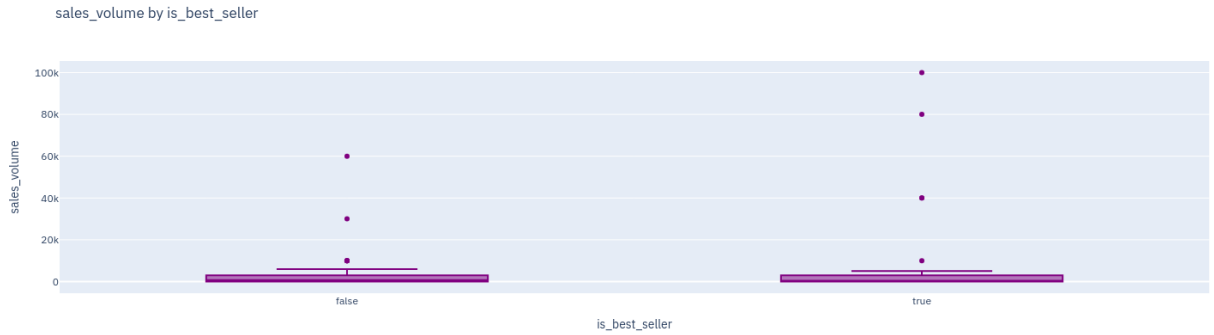
5.2.1. Variables categóricas debes crear un boxplot. Explique cómo interpreta el gráfico

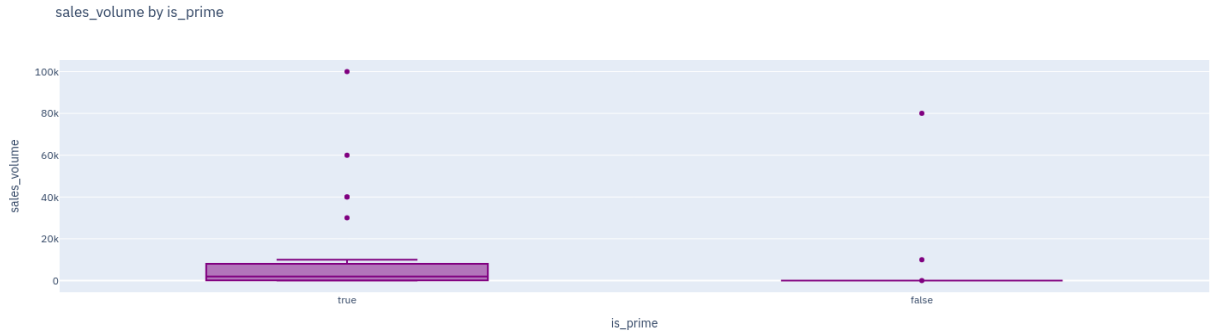








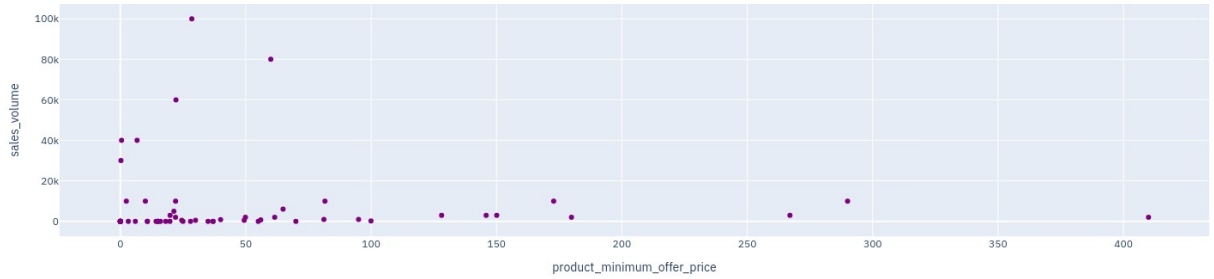




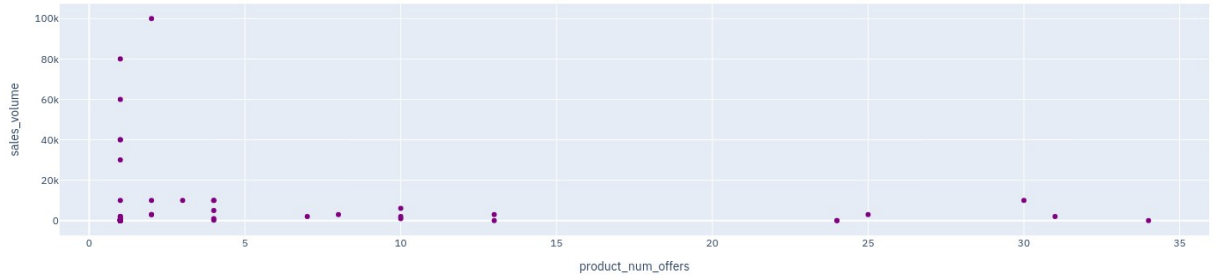
En general se aprecia que el volumen de ventas está demasiado alejado de la media en cada variable categórica (A excepción de de los productos que son amigables con el medio ambiente y lo que son Amazon Choice). Esto posiblemente se debe a que el volumen de ventas tiene una escalada demasiado grande en comparación con la cantidad de datos presentes en cada variable categórica.

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

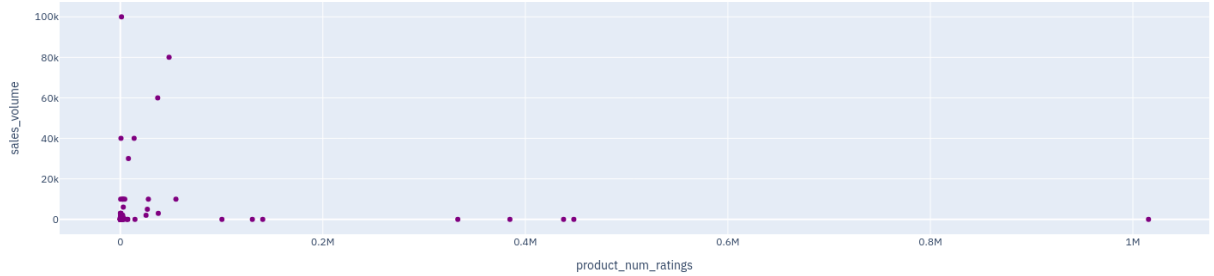
product_minimum_offer_price vs sales_volume



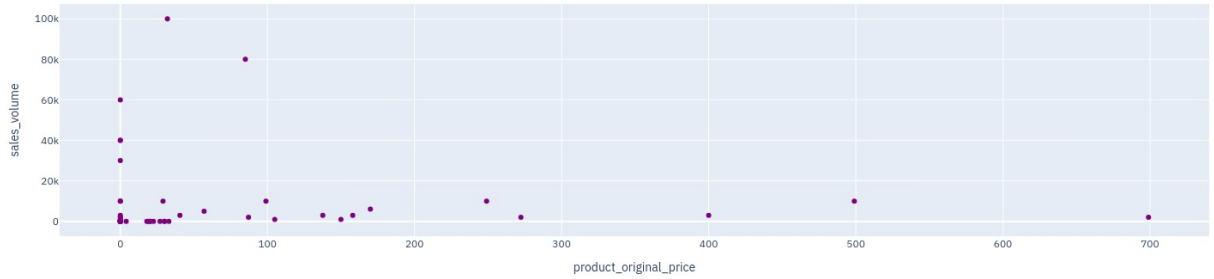
product_num_offers vs sales_volume



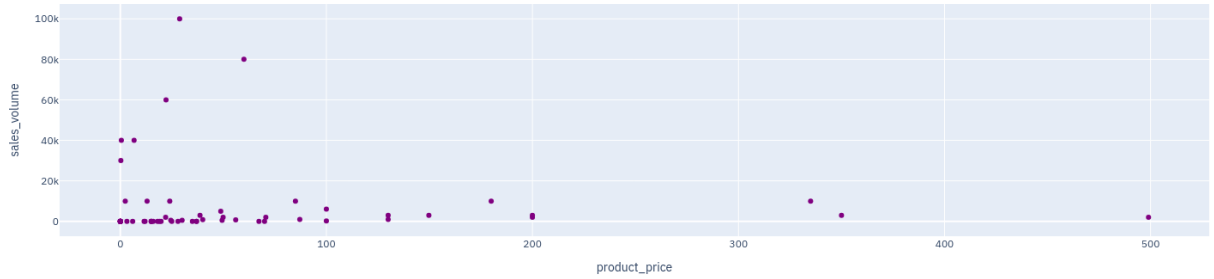
product_num_ratings vs sales_volume

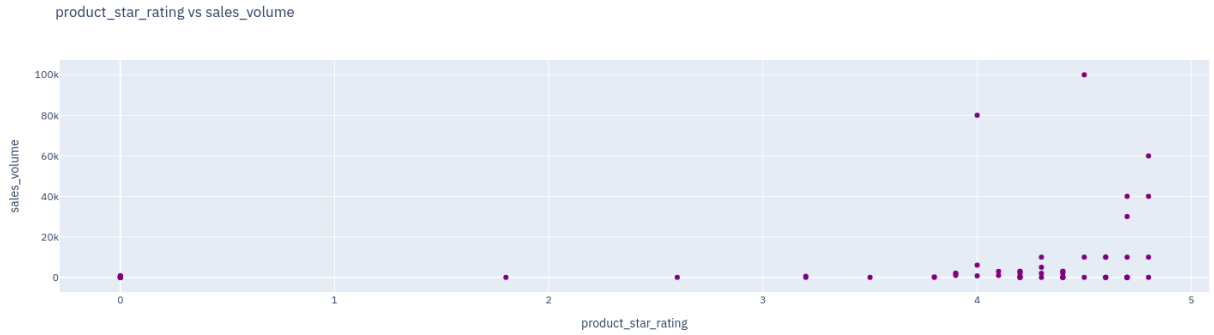


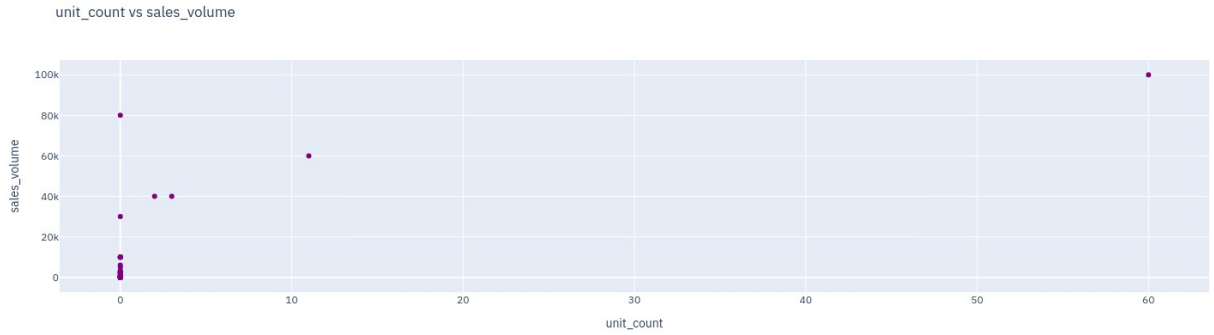
product_original_price vs sales_volume



product_price vs sales_volume



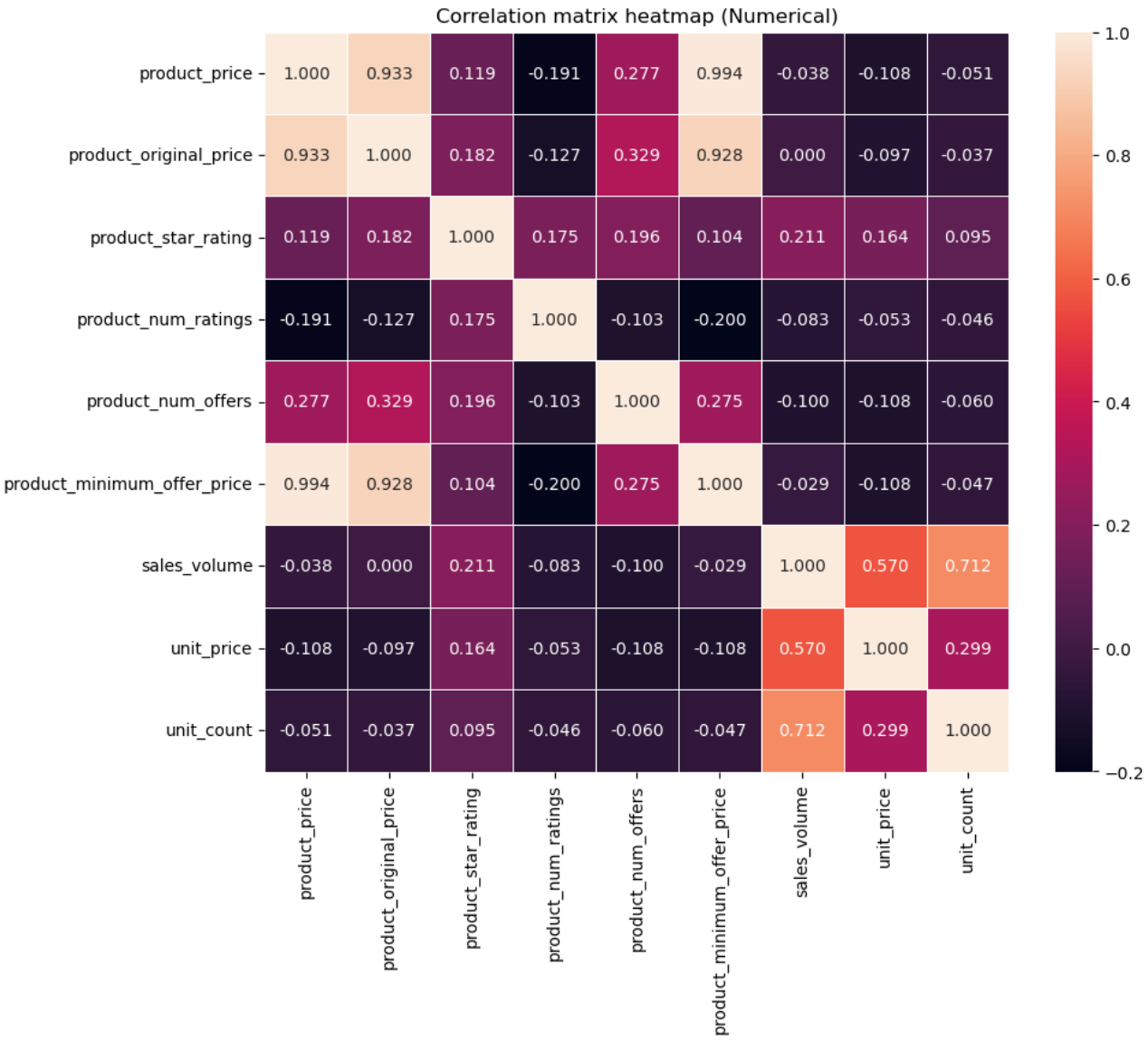




En general se aprecia que el volumen de ventas es alto para las vairables numéricas con un valor bajo, excepto la puntuación por estrellas que tiene un comportamiento inverso.

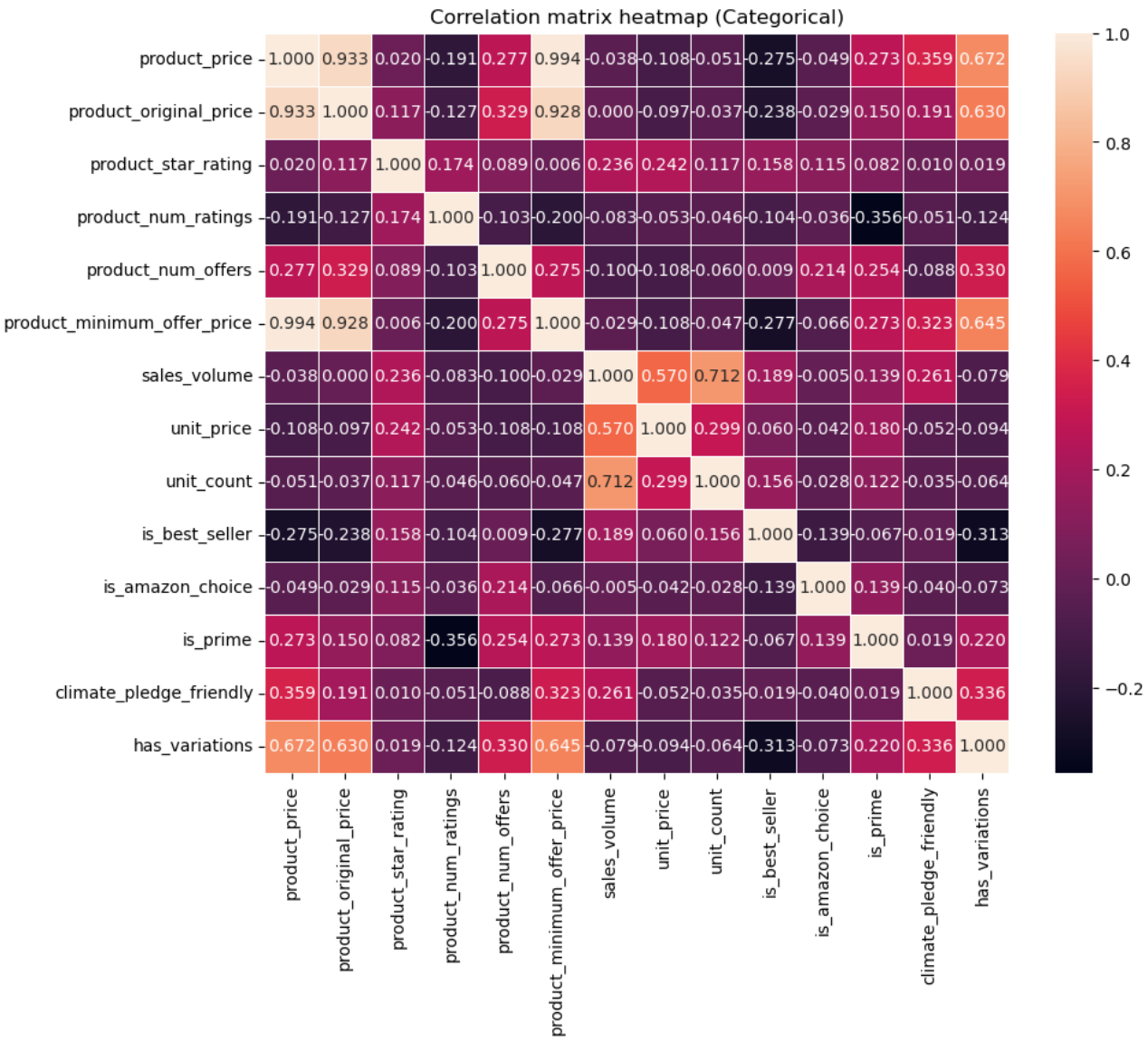
5.3. Matriz de correlación.

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.



Las variables que explican la mayor parte de la variabilidad del volumen de ventas son el precio unitario y el conteo por unidad. Si el coeficiente es positivo es por que se da un relación proporcinal (Si aumenta X, aumenta Y), si es negativo se da una relación inversamente proporcional (Si aumenta X, disminuye Y).

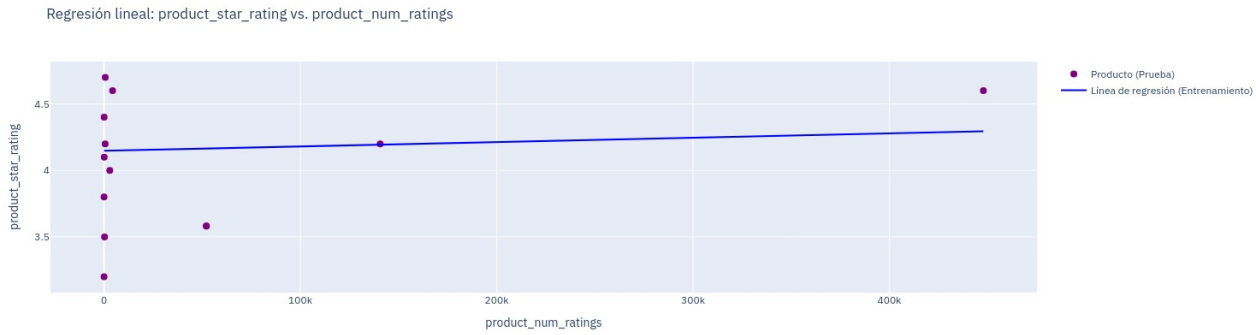
5.3.2. Cree las dummy variables para todas las variables categóricas y genere la matriz de correlación nuevamente. ¿Cuál es el valor de variable categórica con mayor correlación?



La correlación más alta entre variables categóricas se da entre has_variations (Tiene variaciones) y climate_pledge_friendly (Comprometido con el medio ambiente).

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 product_star_rating qué es la calificación. Cual es el coeficiente b1 y b0. Describir resultados.

B1: 3.303e-07
B0: 4.147



Se aprecia que los datos de testing no se alejan demasiado de la predicción (línea de regresión) realizada a partir de los datos de entrenamiento, los cuales representan el 80% de los datos originales.