Tarea #: 1
Tema: Exploración de datos Y PCA
Fecha entrega: 11:59 pm 11 de Septiembre de 2024
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)
Bogotá	103.5	7.18	10.5	32	52	48	18
Medellín	44.1	2.57	11.2	31	53	47	7.5
Cali	22.4	2.23	13.8	30	52	48	4.2
Barranquilla	16.8	1.23	12.4	29	51	49	3.1
Cartagena	10.5	1.03	10.9	30	51	49	2.8
Bucaramanga (test)	7.3	0.58	9.2	33	52	48	1.5
Pereira	6.2	0.48	12	32	52	48	1.3
Cúcuta (test)	5.1	0.76	16.3	28	51	49	1.2
Ibagué (test)	4.8	0.53	13.4	31	52	48	1.1
Santa Marta	4	0.52	11.6	29	51	49	0.9
Manizales	3.8	0.43	10.7	32	53	47	0.8
Villavicencio	3.5	0.5	13	30	51	49	0.8
Pasto	3.2	0.45	12.9	31	52	48	0.7
Montería	3	0.49	13.5	29	51	49	0.7
Valledupar	2.8	0.47	14.8	28	51	49	0.6
Neiva	2.5	0.35	14.1	30	52	48	0.6
Popayán	2.3	0.33	15.2	31	52	48	0.5

Armenia	2.1	0.3	13.3	32	53	47	0.5
Sincelejo	2	0.28	16.5	29	51	49	0.5
Tunja	1.8	0.25	10	31	52	48	0.4
Florencia	1.7	0.2	17.5	28	51	49	0.4
Riohacha	1.5	0.22	15.7	27	51	49	0.3
Quibdó	1.3	0.13	18.2	26	52	48	0.3
San Andrés	1.2	0.08	14	27	50	50	0.2
Yopal	1.1	0.15	11.5	29	51	49	0.2
Leticia	1	0.05	13.6	26	51	49	0.1
Arauca (test)	0.9	0.08	12.2	29	51	49	0.1
Mocoa	0.8	0.04	15	28	52	48	0.1
Mitú	0.7	0.01	20	25	51	49	0.05
Puerto Carreño (test)	0.6	0.01	22	24	50	50	0.05

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.

GDP (USD Billion):

Media: 8.75

Mediana: 2.65

Desviación Estándar Poblacional: 19.579713140561246

Population (Millions):

Media: 0.7309999999999999

Mediana: 0.39

Desviación Estándar Poblacional: 1.3300936057285593

Unemployment Rate (%):

Media: 13.833333333333334

Mediana: 13.45

Desviación Estándar Poblacional: 2.8955521446215093

Average Age:

Media: 29.433333333333334

Mediana: 29.0

Desviación Estándar Poblacional: 1.9093337988826249

Women (%):

Media: 51.5

Mediana: 51.0

Desviación Estándar Poblacional: 0.7637626158259734

Men (%):

Media: 48.5

Mediana: 49.0

Desviación Estándar Poblacional: 0.7637626158259734

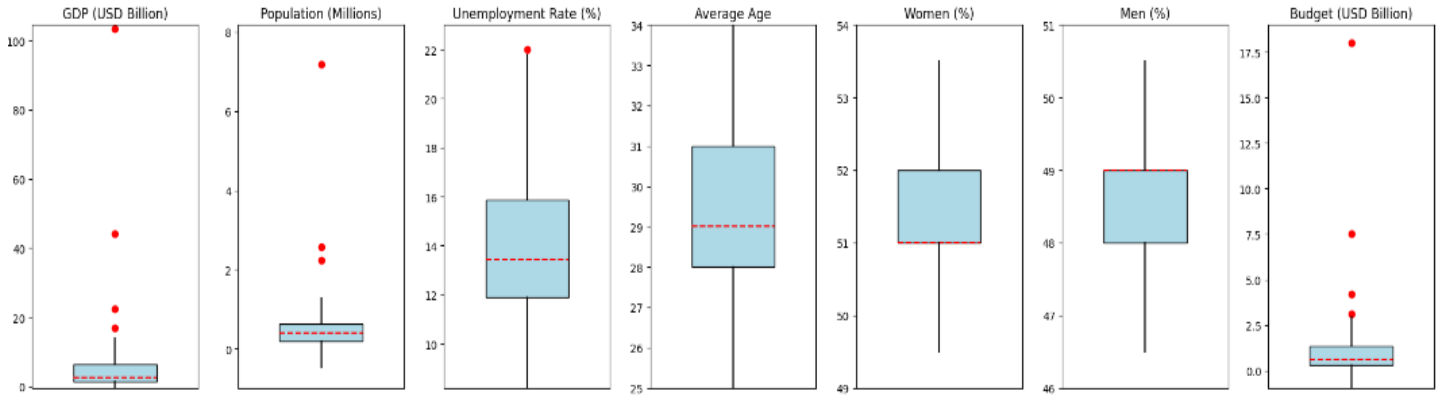
Budget (USD Billion):

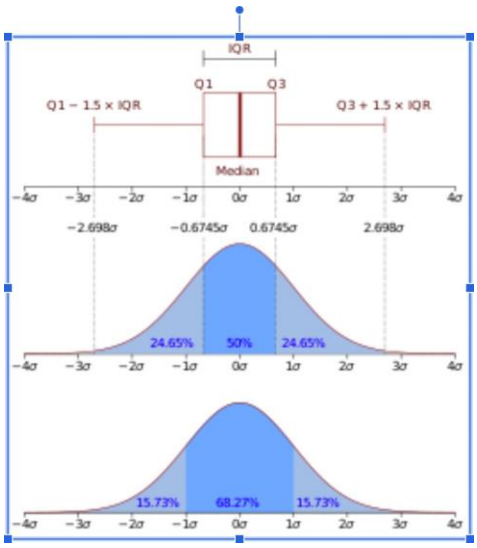
Media: 1.6499999999999997

Mediana: 0.6

Desviación Estándar Poblacional: 3.3931794333142284

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

La covarianza entre GDP (USD Billion) y Population (Millions) es : 25.79571666666666

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

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

$$\mathrm{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

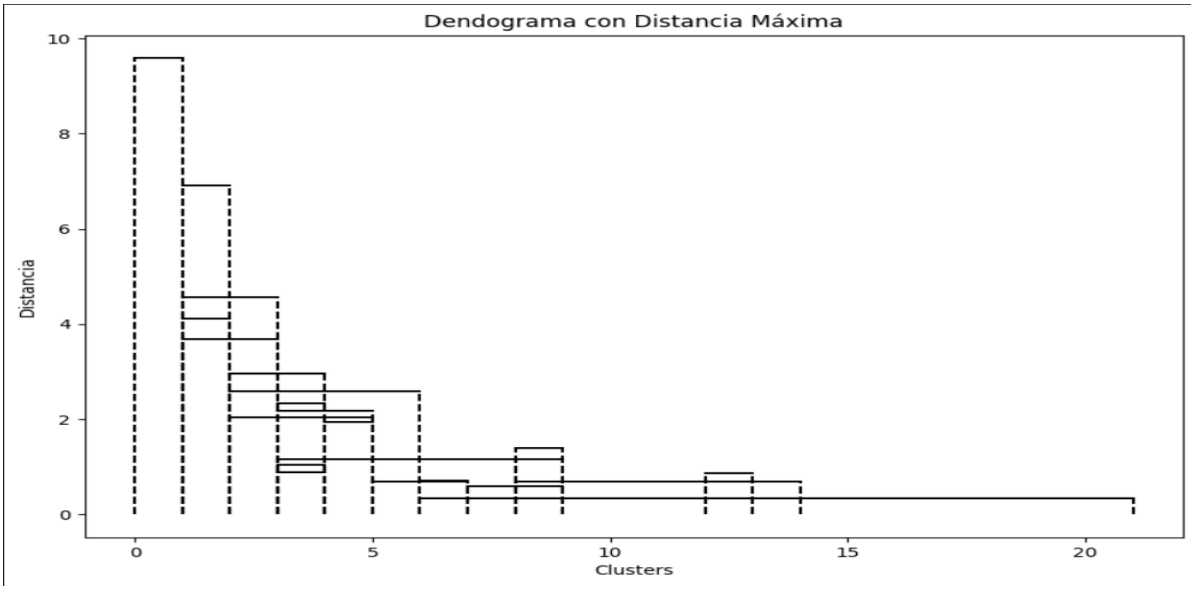
La covarianza entre GDP (USD Billion) y Population (Millions) es : 0.9905104636501415

1.5. Explica la relación entre covarianza y correlación.

La **covarianza** mide cómo varían juntas dos variables y puede ser positiva (ambas aumentan o disminuyen juntas) o negativa (una aumenta mientras la otra disminuye), pero su magnitud depende de las unidades de las variables. La **correlación**, en cambio, estandariza esta medida, proporcionando un valor entre -1 y 1 que indica la fuerza y dirección de la relación lineal, sin depender de las unidades. Así, la correlación ofrece una interpretación más clara y comparable de la relación entre variables.

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 librerias. Vamos a crear 6 grupos, es decir, k=6 (clusters).

Cargar el resultado de la ciudad del datset de testing y la ciudad q es mas cercana al centroide. En la competencia de kaggle. https://www.kaggle.com/t/fb4269a7c52845488efdd718afe03847



1.7. Calcula el resultado de un dedograma utilizando la distancia maxima en en python.

Se calculo la distancia máxima para hacer el dendograma, sin utilizar librerías

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

Matriz de Covarianza:
[1.0344827586206895, 1.0246659968794567]
[1.0246659968794567, 1.0344827586206895]

2.2. Cuales son los eigenvalues

Eigenvalues: [2.059148755500146, 6.660626602072578e-16]
Eigenvectors:
[[0.7071067811865475, 0.7071067811865475], [0.7071067811865476, 0.7071067811865476]]

2.3. Cuál es la varianza explicada por el eigenvalue.

Varianza Explicada por el Eigenvalue: [0.9999999999999998, 3.234650524534241e-16]

2.4. Cual es el valor del eigenvector

Valor del Eigenvector Principal: [0.7071067811865475, 0.7071067811865475]

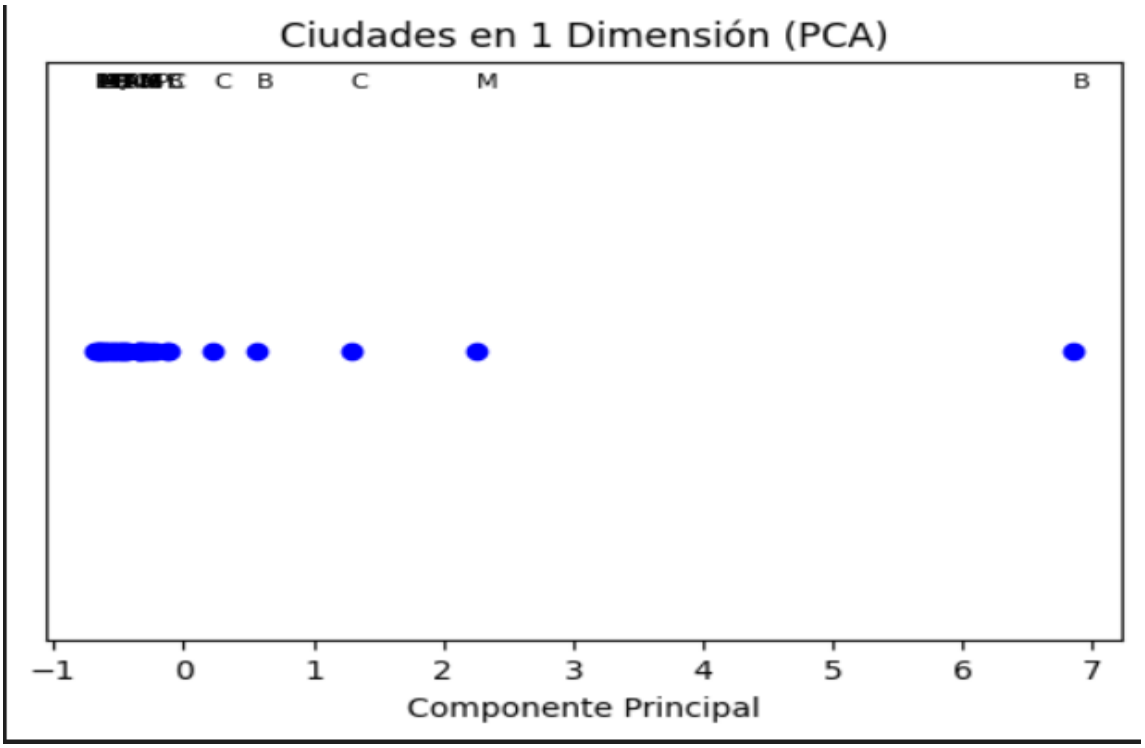
2.5. Cuál es la matriz proyectada.

Matriz Proyectada:
[6.850254902970839, 2.2542914886585734, 1.289860705902846, 0.5559990515941917, 0.22215487242440524, -0.22552843029010897, -0.28371492170438956, -0.33878373663741, -0.3124044862854518, -0.349819854129487, -0.33577783268351447, -0.35363312033553873, -0.42826201546057013, -0.44611730311259434, -0.46928880894761804, -0.4835326709566296, -0.5067041767916534, -0.5368966933496633, -0.5334871082696897, -0.58855592320271, -0.6187484397607199, -0.5851463381227364, -0.6419199455957436, -0.6740190952567795, -0.6776305208997921, -0.13264056638697524, -0.11640000323926344, -0.24950729837728886, -0.6544590150647684, -0.629582716689758]

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

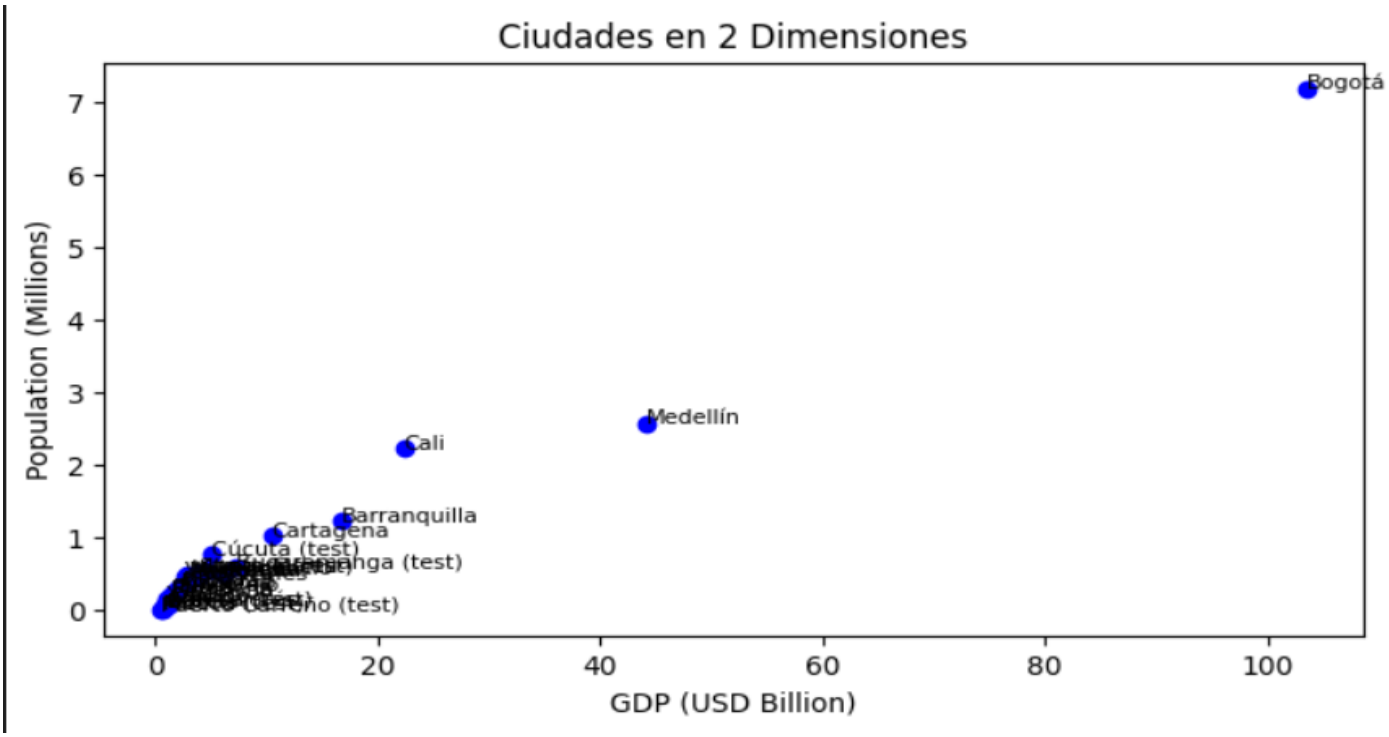
Error de Reconstrucción: 0.2846860904957457

2.7. Pintar todas las ciudades en 1 dimension.



Las letras que se ven poco son las iniciales de la ciudad

2.8. Utilizar python par apintar todas las ciudades en 2 dimensiones



3. PCA

Cargar el data set de caras que está en la carpeta datos de la tarea 2 (ver notebook https://github.com/jdramirez/UCO_ML_AI/blob/master/src/notebook/PCA.ipynb):

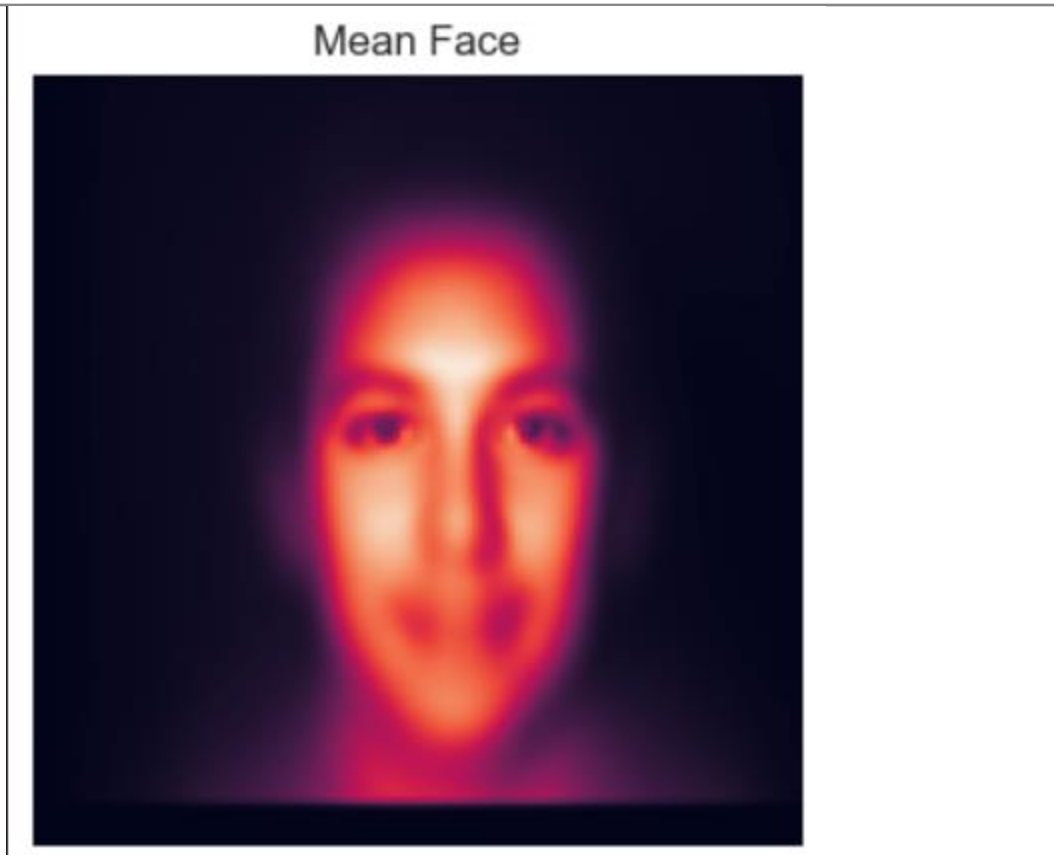
Las siguientes caras son parte del data set q se utilziara para aprender PCA.
Training (1000 faces to train):

1855,4729,3954,2886,3168,4943,2288,2872,5059,2618,3365,1432,5092,4140,1600,4372,3157,208
5,1264,4716,3533,3701,4524,1290,2415,2627,3391,2243,4988,5066,4386,2071,2875,2049,4944,41
78,3953,2881,1638,1852,3739,4381,3998,2076,3396,2244,5061,2620,1899,1297,2412,3706,4523,1
263,4711,3534,1607,4375,3150,2082,3362,1435,5095,4147,4986,5068,4388,2843,3991,2629,1890,
4718,1864,4972,3965,3159,2616,2424,2040,3192,4185,5057,2272,2888,3166,1631,4343,1403,417
1,2286,3354,4515,3730,3502,1255,4727,1609,3962,4975,4149,3708,1863,1897,1299,2844,3996,20
78,3398,4981,3505,1252,4720,4512,3737,1404,4176,2281,3353,3161,1636,4344,4182,5050,2275,2
047,3195,2423,2611,3763,4546,4774,3551,2483,4310,1662,3135,3909,3307,4122,1450,1696,2013,
2221,3797,2645,4780,2477,4921,3338,3936,1239,1837,4579,2448,2810,5209,4787,2470,3790,264
2,2226,5003,1691,2014,2828,3300,4125,4919,1457,4317,1665,3132,4773,3556,2484,3764,4541,28
17,2219,1830,2689,3569,3931,4328,4926,1468,5035,1495,2210,2022,5207,2446,3594,4583,2674,3
560,4745,1237,4577,1839,2680,3752,4113,1461,3336,3104,3938,4321,1653,3799,2479,1698,2821,
3907,3309,4910,4548,1806,3103,4326,1654,4114,1466,4928,3331,4570,2687,3755,3567,4742,123
0,4584,2673,2441,3593,2025,2819,5200,5032,1492,2217,3558,1801,1459,4917,4319,3900,2228,28
26,4789,1298,1896,3399,4980,2079,2845,3997,4148,4974,1608,3963,3709,1862,2046,3194,4183,5
051,2274,2610,2422,4513,3736,3504,4721,1253,3160,4345,1637,4177,1405,2280,3352,1865,4719,
3158,3964,4973,4389,2842,3990,5069,4987,2628,1891,4170,1402,2287,3355,3167,2889,4342,163
0,3503,4726,1254,4514,3731,2425,2617,4184,5056,2273,2041,3193,3952,2880,1639,4179,4945,18
53,3738,2048,2874,4710,1262,3535,3707,4522,3363,5094,4146,1434,4374,1606,3151,2083,3397,2
245,5060,4380,2077,3999,1296,2413,2621,1898,5058,2873,2619,4728,1854,4942,2289,3169,3955,
2887,2626,1291,2414,4387,2070,3390,2242,5067,4989,4373,1601,3156,2084,3364,5093,4141,143
3,3700,4525,4717,1265,3532,2440,3592,4585,2672,1493,5033,2216,2818,2024,5201,1467,4929,41
15,3330,3102,1655,4327,3566,1231,4743,4571,2686,3754,2827,2229,4788,1800,3559,4318,3901,1
458,4916,4576,1838,2681,3753,3561,1236,4744,3939,3105,1652,4320,1460,4112,3337,2023,5206,
1494,5034,2211,4582,2675,2447,3595,3308,4911,3906,4549,1807,2478,3798,1699,2820,1664,431
6,3133,3301,4918,1456,4124,3765,4540,4772,3557,2485,3791,2643,4786,2471,1690,2829,2015,22
27,5002,3568,1831,2688,4927,1469,3930,4329,2218,2816,2220,5005,1697,2012,4781,2476,3796,2
644,4775,3550,2482,3762,1809,4547,3306,1451,4123,1663,4311,3908,3134,2449,2811,5208,3937,
4920,3339,1836,4578,1238,1944,4638,3079,2997,3845,4852,2399,2963,5148,2709,3274,4051,518
3,1523,4263,1711,2194,3046,4607,1375,3422,3610,4435,1381,2504,2736,2352,3280,5177,4899,42
97,2160,2158,2964,4069,4855,2990,3842,1729,1943,3628,4290,2167,3889,2355,3287,5170,2731,1
988,1386,2503,3617,4432,4600,1372,3425,4264,1716,2193,3041,3273,5184,1524,5179,4897,4299,
3880,2952,2738,1981,4609,1975,4863,3048,3874,2707,2535,3083,2151,5146,4094,2363,3077,299
9,4252,1720,4060,1512,3245,2397,4404,3621,3413,4636,1344,1718,3873,4058,4864,3619,1972,19
86,1388,2169,3887,2955,3289,4890,3414,4631,1343,4403,3626,4067,1515,3242,2390,3070,4255,1
727,5141,4093,2364,3084,2156,2532,2700,3672,4457,1919,1317,4665,2592,3440,1773,4201,3818,
3024,3216,1541,4033,1787,2102,2330,5115,2754,3686,4691,2566,4830,3229,3827,1328,4468,192
6,2559,2901,4696,2561,2753,3681,2337,5112,1780,2939,2105,3211,1546,4808,4034,1774,4206,30
23,1310,4662,2595,3447,3675,4450,2906,2308,1921,2798,3478,3820,4239,1579,4837,1584,5124,2

301,2133,3485,2557,4492,2765,3471,1326,4654,1928,4466,3643,2791,1570,4002,3227,3829,3015,1742,4230,3688,2568,1789,2930,3816,3218,4801,1917,4459,1319,3012,1745,4237,4839,1577,4005,3220,4461,3644,2796,3476,1321,4653,4495,2762,3482,2550,2908,2134,1583,5123,2306,3449,1910,4806,1548,4208,3811,2339,2937,4698,1389,1987,3288,4891,3886,2954,2168,4865,4059,1719,3872,3618,1973,3085,2157,5140,4092,2365,2701,2533,4402,3627,3415,1342,4630,3071,1726,4254,1514,4066,3243,2391,1974,4608,3875,3049,4862,4298,3881,2953,4896,5178,2739,1980,1513,4061,3244,2396,2998,3076,1721,4253,3412,1345,4637,4405,3620,2534,2706,5147,4095,2362,3082,2150,2991,3843,1728,4854,4068,1942,3629,2965,2159,1373,4601,3424,3616,4433,3272,1525,4057,5185,1717,4265,2192,3040,2354,3286,5171,4291,3888,2166,1387,2502,2730,1989,5149,2962,2708,4639,1945,4853,2398,2996,3844,3078,2737,1380,2505,4296,2161,2353,3281,4898,5176,1710,4262,2195,3047,3275,1522,4050,5182,3611,4434,1374,4606,3423,3483,2551,4494,2763,5122,1582,2307,2135,2909,4004,4838,1576,3221,3013,4236,1744,3477,4652,1320,4460,3645,2797,2936,2338,4699,1911,3448,4209,3810,4807,1549,1929,4467,3642,2790,3470,4655,1327,3014,3828,4231,1743,4003,1571,3226,2132,5125,1585,2300,4493,2764,3484,2556,3219,4800,3817,1318,1916,4458,2569,3689,1788,2931,4207,1775,3022,3210,4035,1547,4809,3674,4451,4663,1311,2594,3446,2752,3680,4697,2560,1781,2104,2938,2336,5113,3479,1920,2799,1578,4836,3821,4238,2309,2907,2331,5114,1786,2103,4690,2567,2755,3687,4664,1316,2593,3441,3673,4456,1918,3217,4032,1540,4200,1772,3025,3819,2558,2900,3826

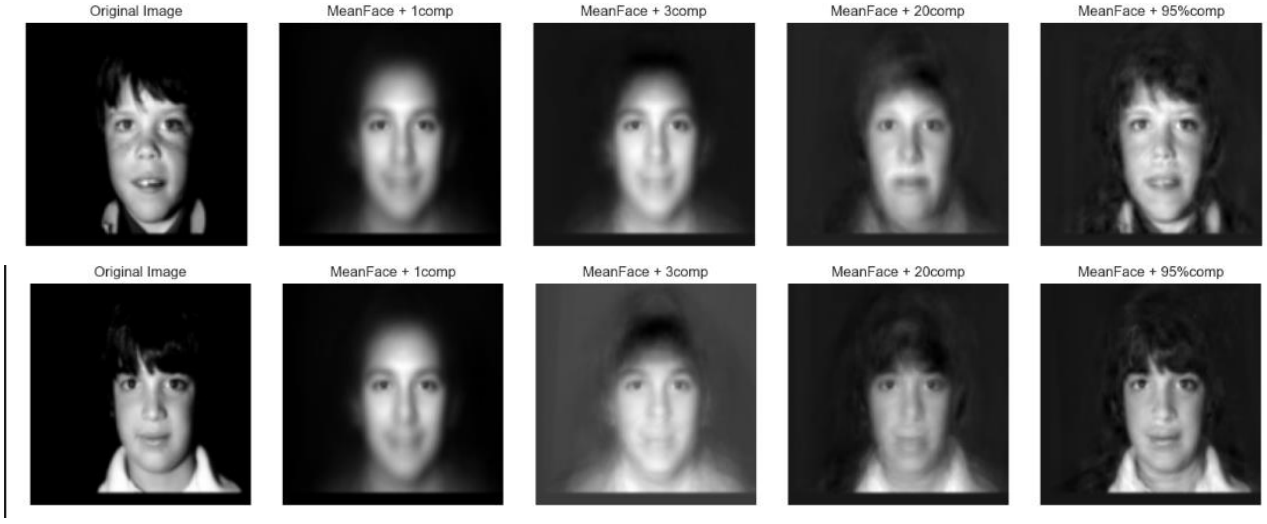
Testing (300 faces):
'4831,3228,4469,1927,1329,5109,2922,2748,4679,1905,4813,3038,3804,2777,4480,3497,2545,2121,2313,5136,1596,4222,1750,3007,3235,4010,1562,3651,2783,4474,4646,1334,3463,3803,1768,4028,4814,1902,3669,2589,2119,2925,4641,1333,3464,3656,2784,4473,3232,4017,1565,4225,1757,3000,2314,5131,1591,2126,3490,2542,2770,4487,1934,4648,3009,3835,4822,2913,5138,1598,2779,3499,4021,1553,3204,3036,4213,1761,2580,3452,4677,1305,4445,3660,2574,4683,2746,3694,5107,2322,2110,1795,4489,2128,2914,4019,4825,1759,3832,3658,1933,2117,1792,5100,2325,2741,3693,2573,4684,4442,3667,2587,3455,4670,1302,3031,4214,1766,4026,1554,3203,2371,5154,4086,3091,2143,2527,2715,1356,4624,3401,3633,4416,1958,3257,2385,1500,4072,1732,4240,3859,3065,1993,2518,3892,2940,4885,3866,2188,4871,3268,4429,1967,1369,1735,4247,3062,3250,2382,1507,4849,4075,3634,4411,1351,4623,3406,2712,2520,2978,3096,2144,2376,5153,4081,3439,1960,1538,4876,5198,3861,4278,4882,2349,3895,2947,1994,1969,4427,3602,3430,1367,4615,3868,2186,3054,1703,4271,1531,4043,5191,3266,2172,4285,5165,2340,3292,2724,2516,1393,3259,4840,2985,3857,1358,1956,4418,2529,4088,2971,2511,1394,2723,5162,2347,3295,2949,2175,4282,4878,1536,4044,5196,3261,2181,3053,1704,4276,3437,1360,4612,4420,3605,2976,3098,2378,1951,3408,4249,2982,3850,4847,1509,1758,3833,4824,4018,3659,1932,4488,2915,2129,2586,3454,1303,4671,4443,3666,1555,4027,3202,3030,1767,4215,5101,2324,2116,1793,2572,4685,2740,3692,1599,5139,2912,3498,2778,4649,1935,4823,3834,3008,2747,3695,2575,4682,2111,1794,5106,2323'
Utiliza solo las caras de entrenamiento para los siguientes puntos:

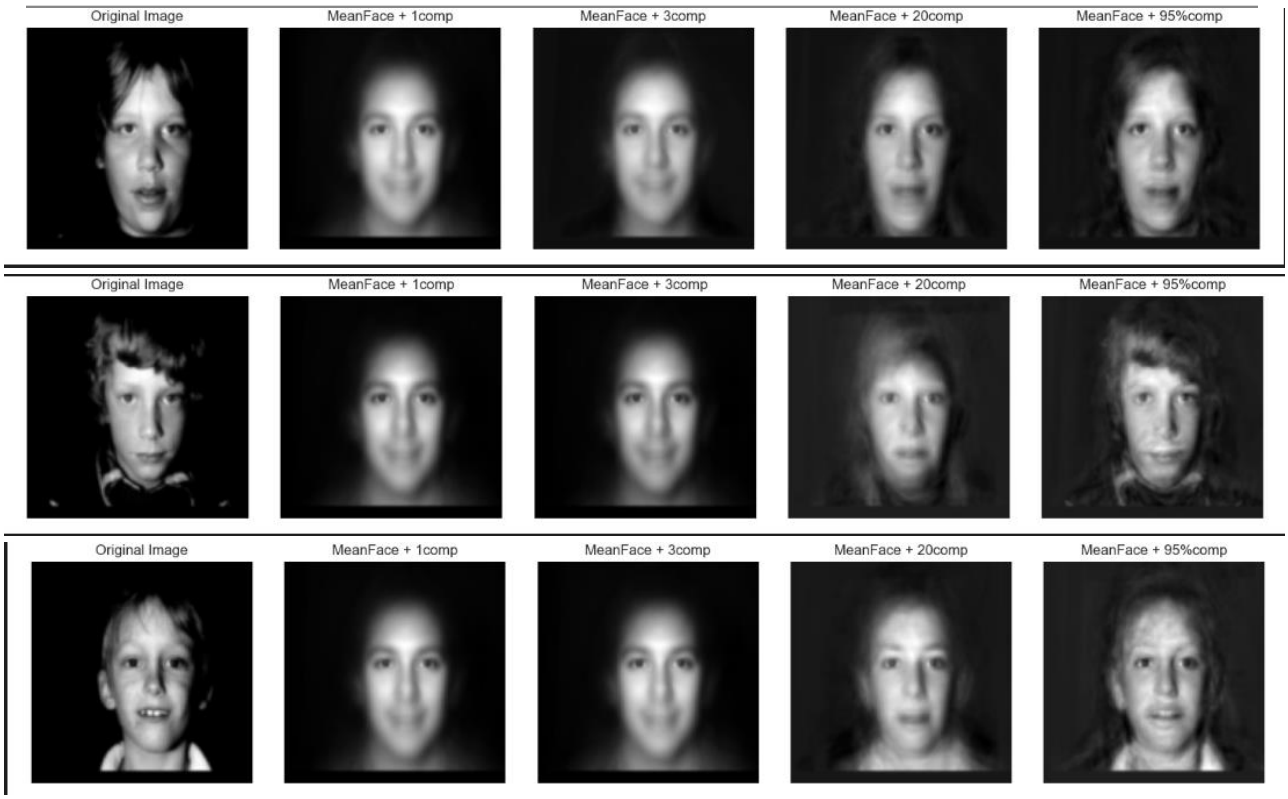
1. Calcular la mean face. Que es la cara con el promedio de los pixeles y visualizarla.



2. Centrar los datos, utilizar PCA. ¿Cuántos componentes se deben utilizarpara 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 puedeconcluir de los resultados?

   Datos de las cara con todos los componentes 1,3,20,95 %





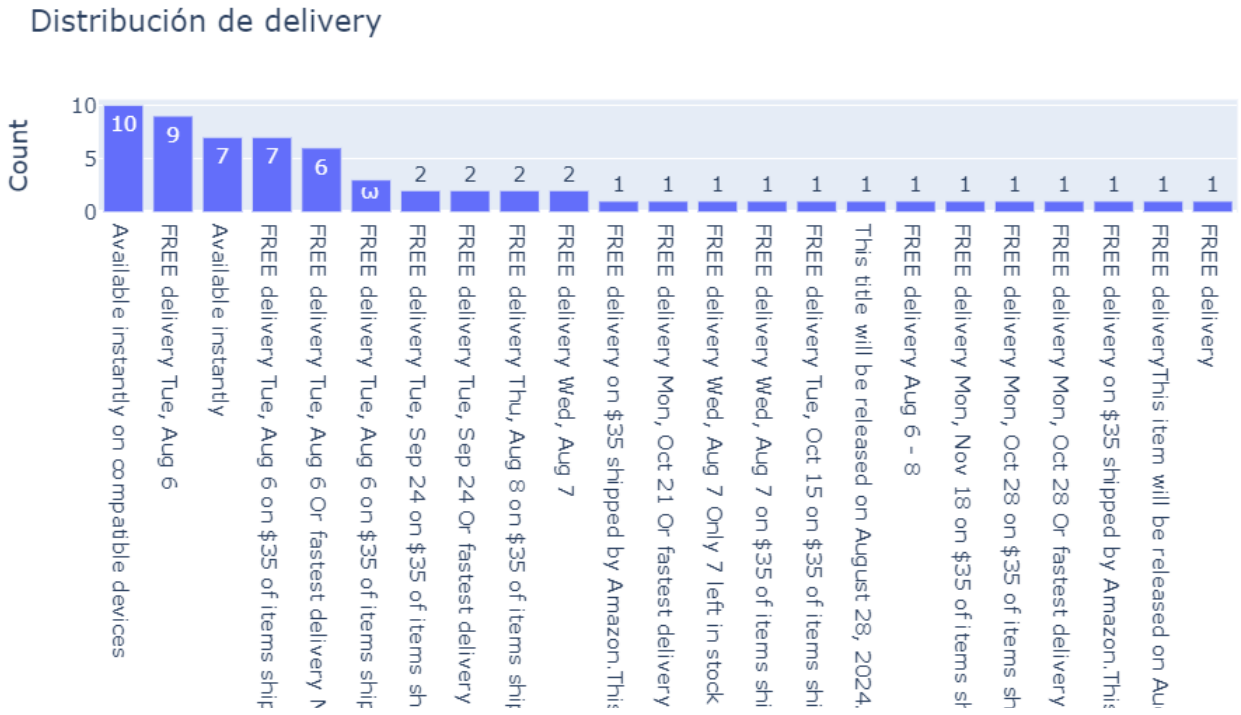
Datos de las caras con todos los componentes

Se puede concluir con los datos que y de las imágenes reontruidad que hasta los 10 componentes se tienen rasgos importantes como los ojos nariz ya después de los diez componentes se va teniendo un poco mas claridad en rasgos como el cabello y gestos. Se puede decir que la información se mantiene muy bien a pesar de eliminar varias dimensiones el pca nos ayuda a mantener la mayor cantidad de varianza para que no se pierdan rasgos e informacion mas importante de las imágenes.

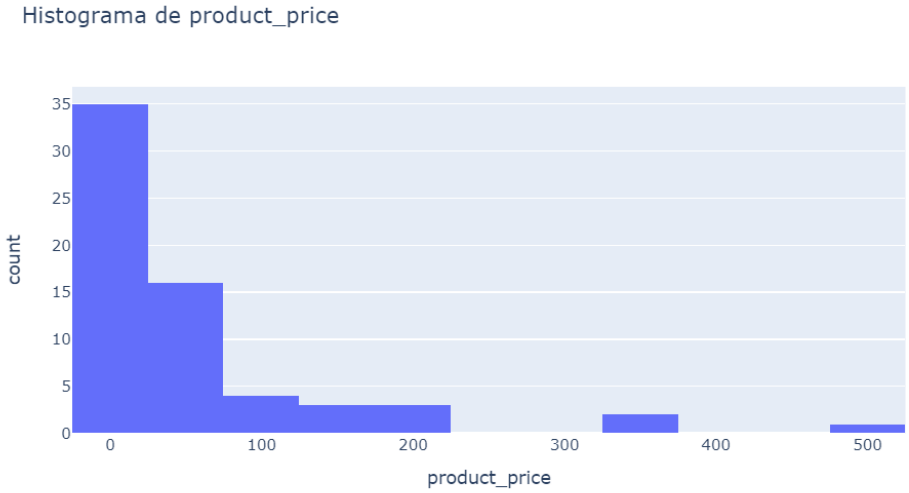
Utiliza los datos de testing. Y envia un archivo a kagle de los datos de testing con la primera componente. Recuerde que testing no puede ser utilizado para aprender PCA. https://www.kaggle.com/t/e125b8f15bb0480188059e6346e53522

4. Utilizando el dataset del amazon data/amazon_products.csv crear: **Utilizar la librería de plotly**.
   4.1. Distribución de cada variable:
      4.1.1. Para las variables categóricas un gráfico de barras. Categoría número de observaciones.

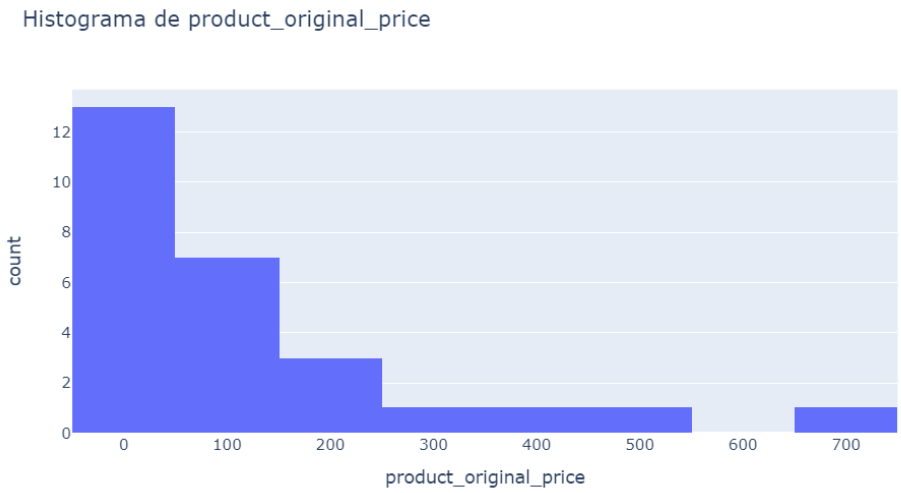
      Se colocaran solo gráficos con valor aquí en el código se encontraran todos:



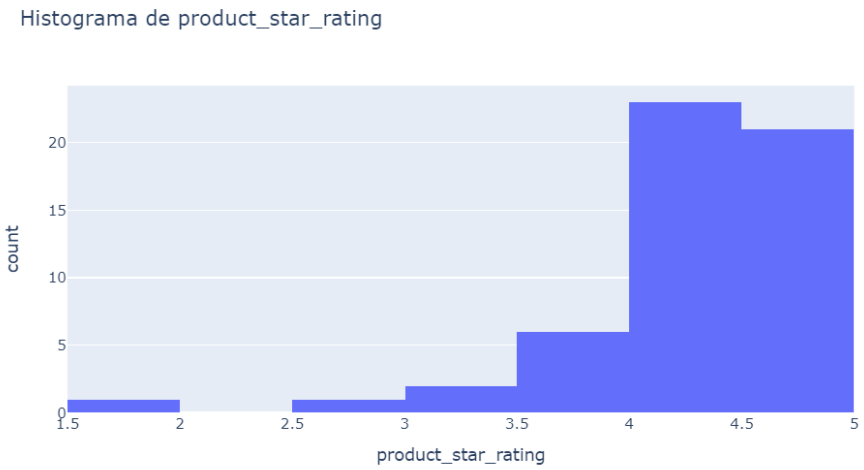
4.1.2. Para las variables numéricas crear histogramas. Listar los productos que están más lejos de 5 estándares de desviación, y serían considerados outliers. Hacer test de si es una distribución normal o no.



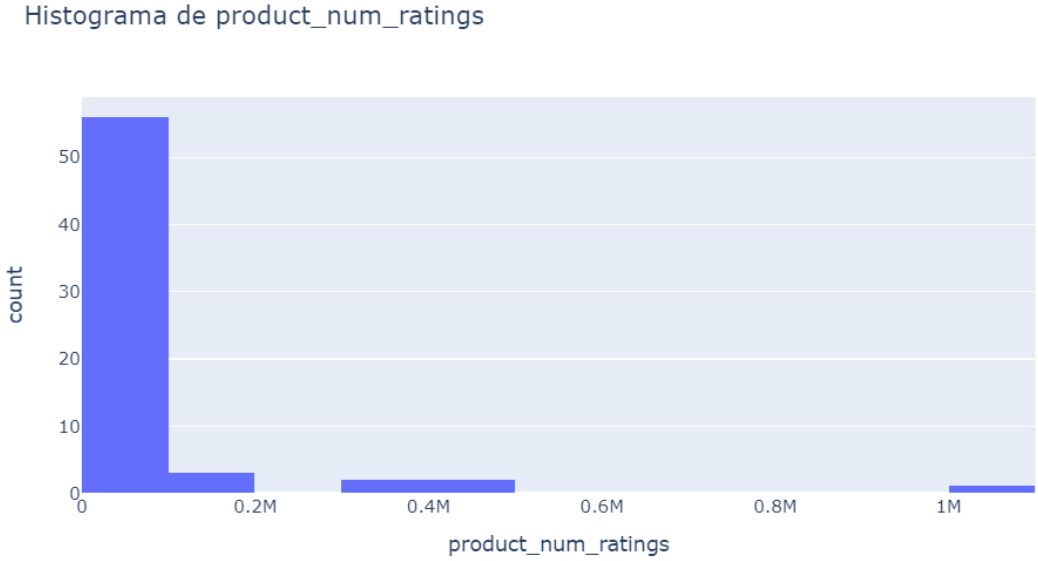
Test de Shapiro-Wilk para product_price: Estadístico=0.6241337060928345, p-valor=1.5730359723131748e-11



Test de Shapiro-Wilk para product_original_price: Estadístico=0.7093768119812012, p-valor=5.2496334319585e-06



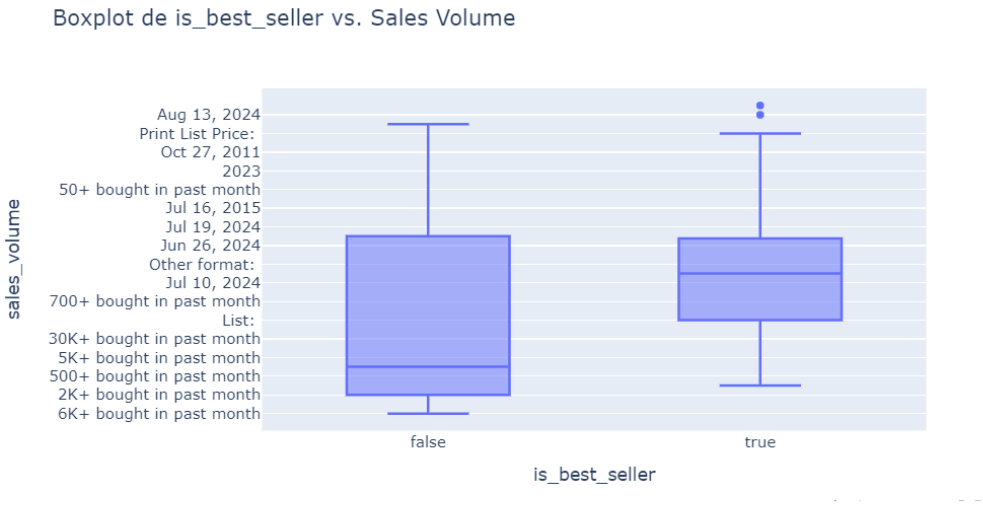
Test de Shapiro-Wilk para product_star_rating: Estadístico=0.7836762070655823, p-valor=1.6984222384053282e-07

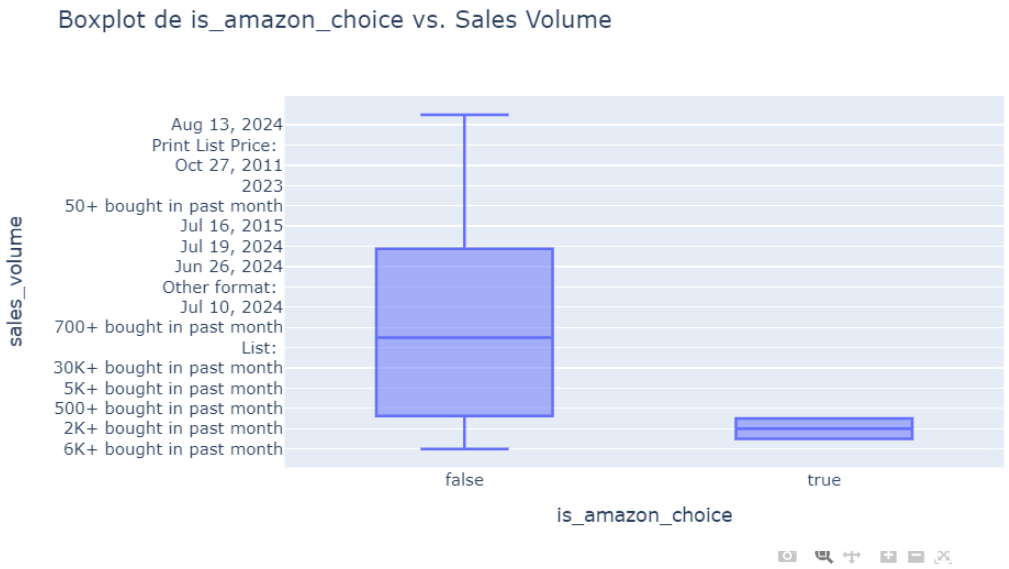


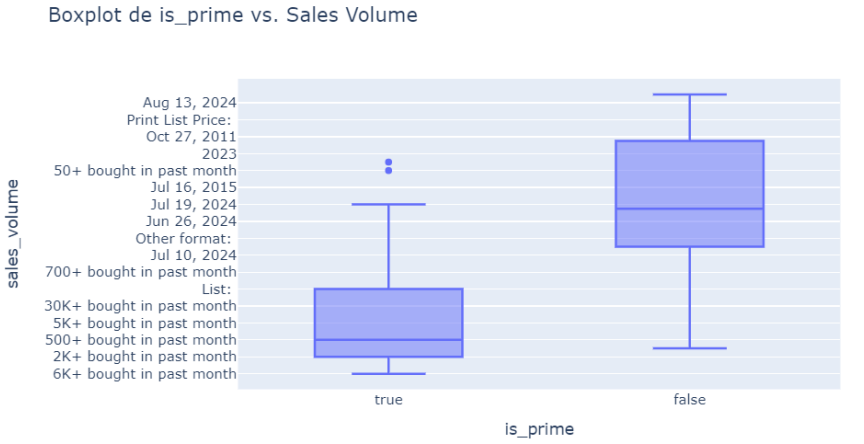
Ninguno de los datos cumple para ser una distribución normal

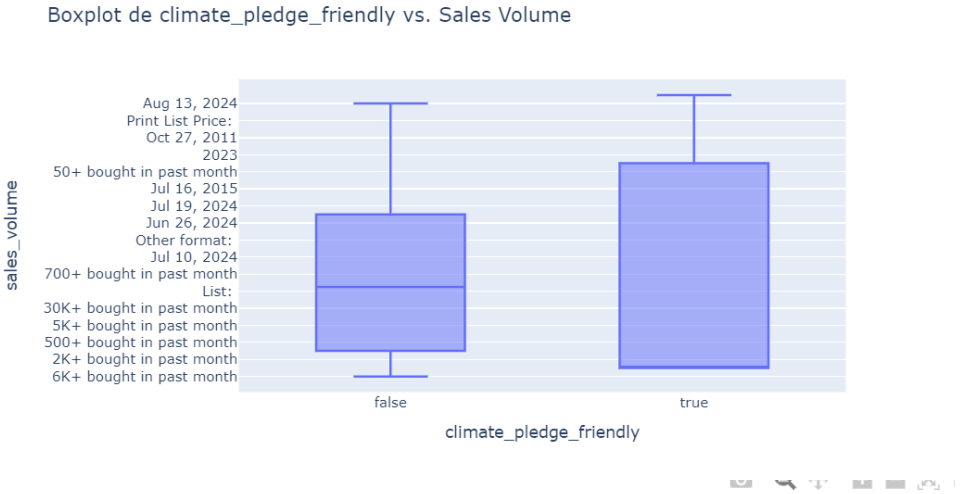
4.2. Gráfico de la relación de cada variable con respecto al `sales_volume (convertir a numero)`:

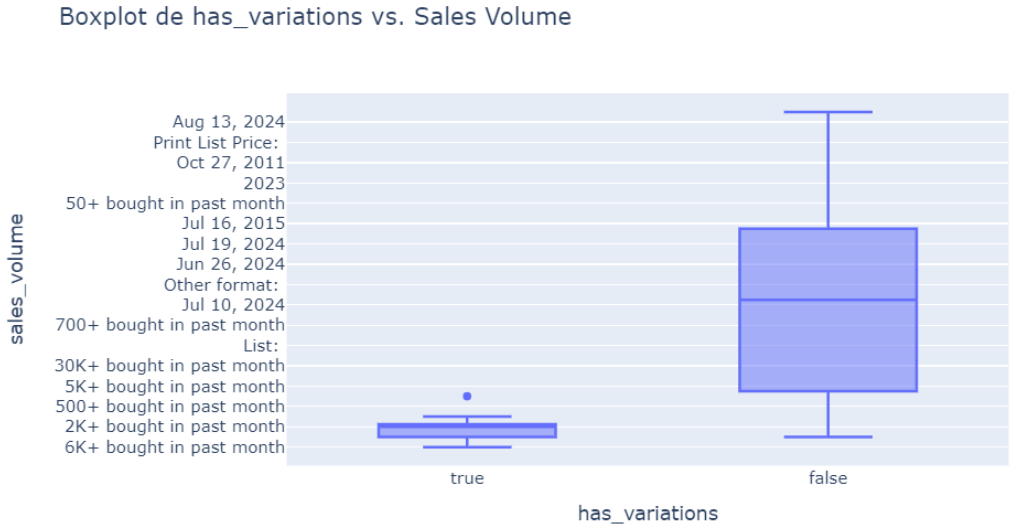
4.2.1. Variables categóricas debes crear un boxplot. Explique cómo interpreta el gráfico


Boxplot de is_best_seller vs. Sales Volume
sales_volume
Aug 13, 2024
Print List Price:
Oct 27, 2011
2023
50+ bought in past month
Jul 16, 2015
Jul 19, 2024
Jun 26, 2024
Other format:
Jul 10, 2024
700+ bought in past month
List:
30K+ bought in past month
5K+ bought in past month
500+ bought in past month
2K+ bought in past month
6K+ bought in past month
false
true
is_best_seller

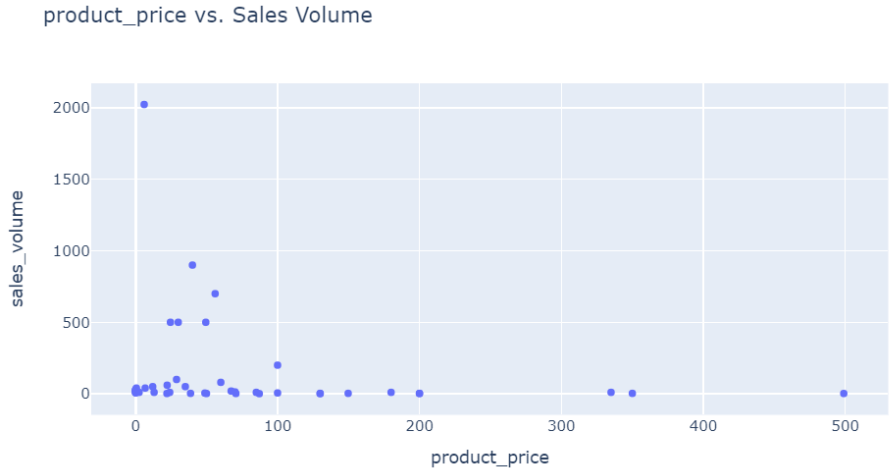

Boxplot de is_amazon_choice vs. Sales Volume
sales_volume
Aug 13, 2024
Print List Price:
Oct 27, 2011
2023
50+ bought in past month
Jul 16, 2015
Jul 19, 2024
Jun 26, 2024
Other format:
Jul 10, 2024
700+ bought in past month
List:
30K+ bought in past month
5K+ bought in past month
500+ bought in past month
2K+ bought in past month
6K+ bought in past month
false
true
is_amazon_choice

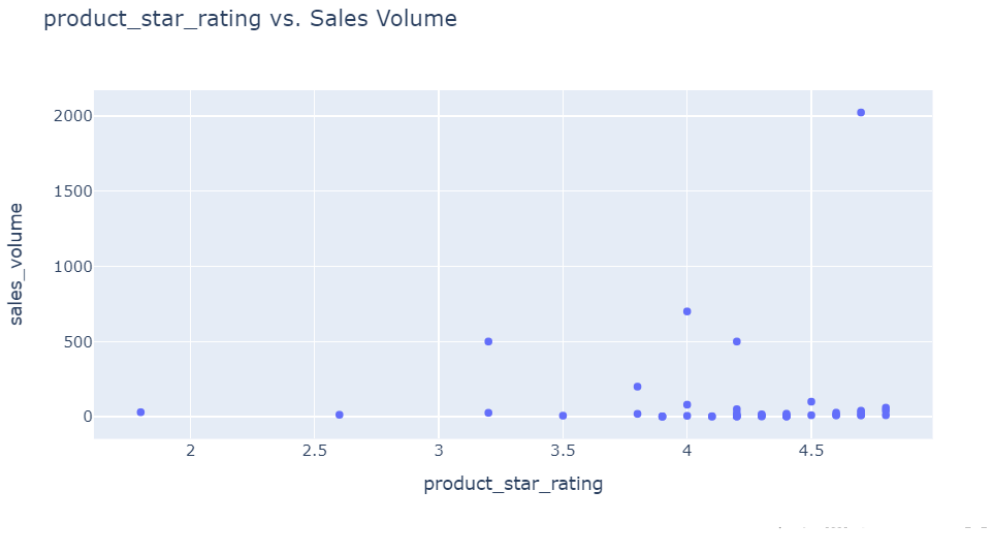

Boxplot de is_prime vs. Sales Volume
sales_volume
Aug 13, 2024
Print List Price:
Oct 27, 2011
2023
50+ bought in past month
Jul 16, 2015
Jul 19, 2024
Jun 26, 2024
Other format:
Jul 10, 2024
700+ bought in past month
List:
30K+ bought in past month
5K+ bought in past month
500+ bought in past month
2K+ bought in past month
6K+ bought in past month
true
false
is_prime

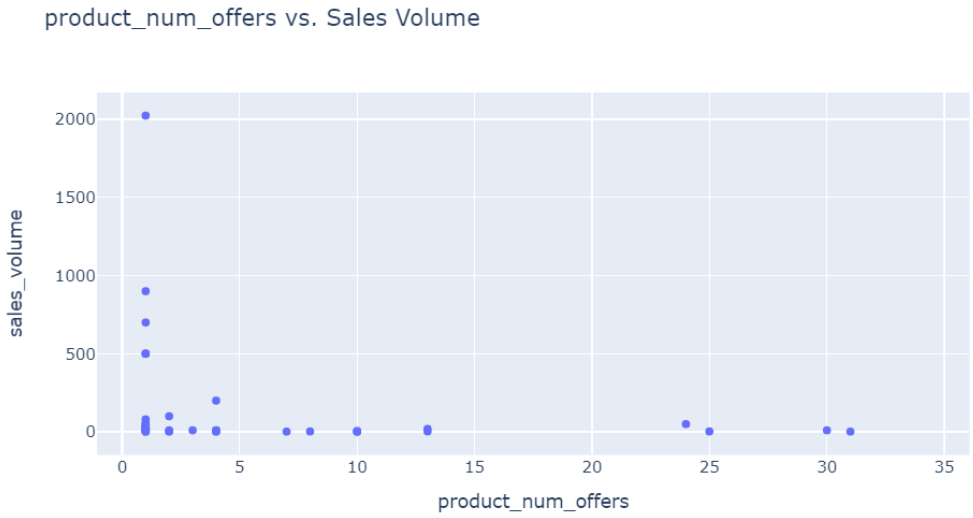




4.2.2. Variables numéricas vas a crear un scatter plot.Explique cómo interpreta el gráfico





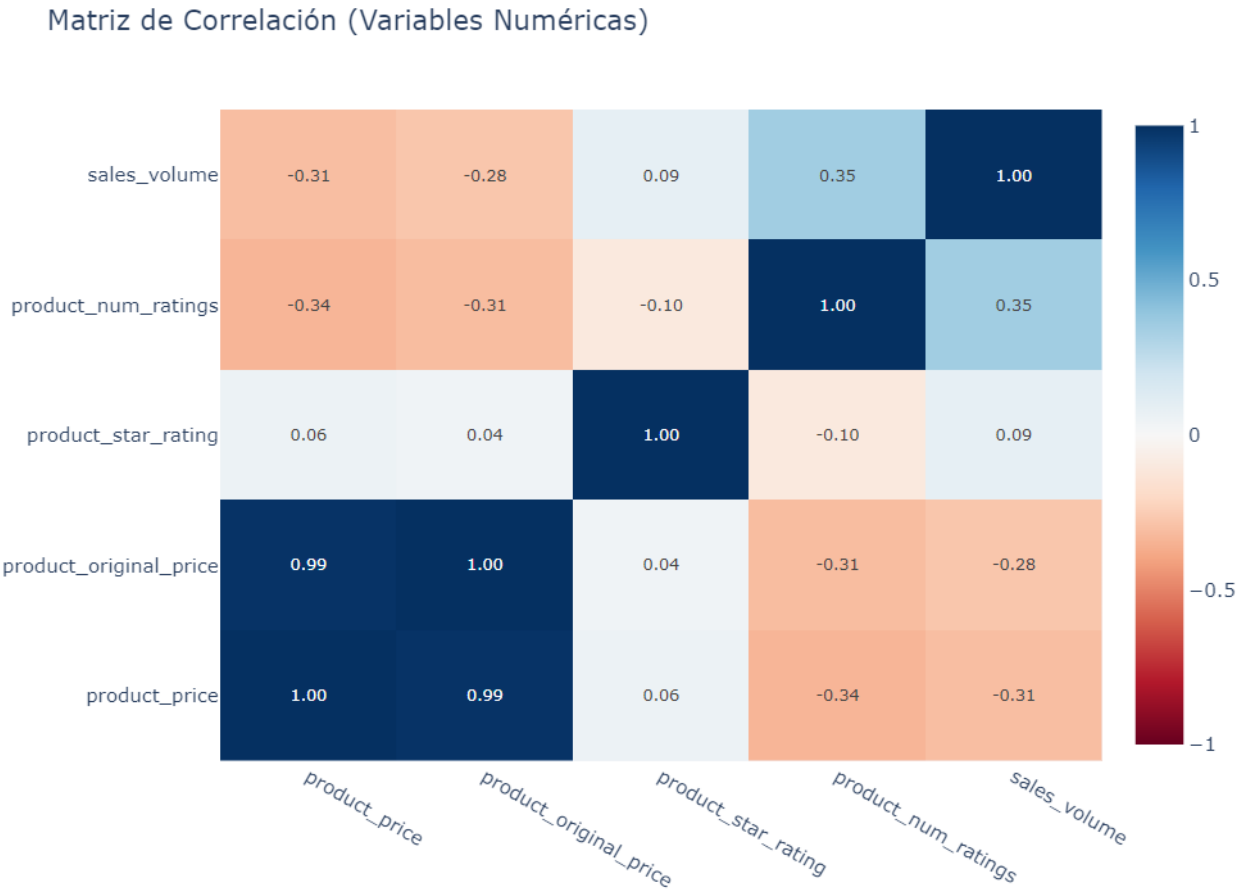


Estos son algunos de los gráficos

4.3. Matriz de correlación.

4.3.1. 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.

Se



Se nota que hay muy poca relacion entre las variables solo una salió del 99% de correlación de resto todas las variables están por debajo del 40% .

4.3.2. 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?

Al convertir los dummy en el código me genera un error