Curso: Estadistica 3
Profesor: Jose Daniel Ramirez Soto

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):

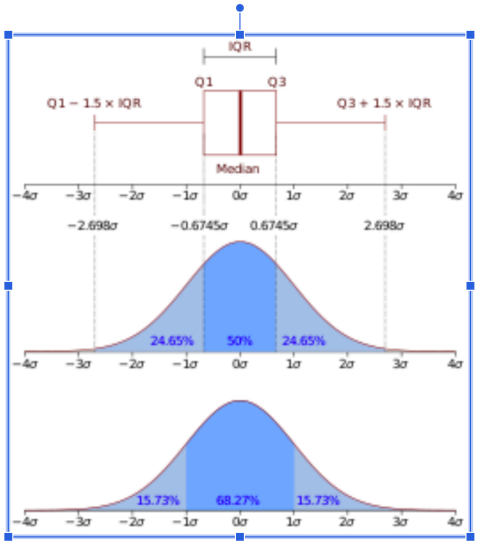
City	GDP (USD Billion)	Population (Millions)	Unemployment Rate (%)	Average Age	Women (%)	Men (%)	Budget (USD Billion)	initial label	training
Bogotá	103.5	7.18	10.5	32	52	48	18	2	Yes
Medellín	44.1	2.57	11.2	31	53	47	7.5	3	Yes
Cali	22.4	2.23	13.8	30	52	48	4.2	2	Yes
Barranquilla	16.8	1.23	12.4	29	51	49	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.

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

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

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

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

1.5. Explica la relación entre covarianza y 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).

Cargar el resultado de la ciudad del dataset de testing y la ciudad que es más cercana al centroide. En la competencia de kaggle.
https://www.kaggle.com/t/7434d84fc6964966b9979d5adc379964

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
   2.2. Cuales son los eigenvalues
   2.3. Cuál es la varianza explicada por el eigenvalue.
   2.4. Cual es el valor del eigenvector
   2.5. Cuál es la matriz proyectada.
   2.6. Cual es el error o diferencia entre la matriz proyectada
   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.
   3.2. Graficar la línea y los puntos

4. 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 dataset que se utilizara para aprender PCA.
Training (1300 faces to train):

1855,4729,3954,2886,3168,4943,2288,2872,5059,2618,3365,1432,5092,4140,1600,4372,3157,2085,126
4,4716,3533,3701,4524,1290,2415,2627,3391,2243,4988,5066,4386,2071,2875,2049,4944,4178,3953,2
881,1638,1852,3739,4381,3998,2076,3396,2244,5061,2620,1899,1297,2412,3706,4523,1263,4711,3534
,1607,4375,3150,2082,3362,1435,5095,4147,4986,5068,4388,2843,3991,2629,1890,4718,1864,4972,39
65,3159,2616,2424,2040,3192,4185,5057,2272,2888,3166,1631,4343,1403,4171,2286,3354,4515,3730,
3502,1255,4727,1609,3962,4975,4149,3708,1863,1897,1299,2844,3996,2078,3398,4981,3505,1252,472
0,4512,3737,1404,4176,2281,3353,3161,1636,4344,4182,5050,2275,2047,3195,2423,2611,3763,4546,4
774,3551,2483,4310,1662,3135,3909,3307,4122,1450,1696,2013,2221,3797,2645,4780,2477,4921,3338
,3936,1239,1837,4579,2448,2810,5209,4787,2470,3790,2642,2226,5003,1691,2014,2828,3300,4125,49
19,1457,4317,1665,3132,4773,3556,2484,3764,4541,2817,2219,1830,2689,3569,3931,4328,4926,1468,
5035,1495,2210,2022,5207,2446,3594,4583,2674,3560,4745,1237,4577,1839,2680,3752,4113,1461,333
6,3104,3938,4321,1653,3799,2479,1698,2821,3907,3309,4910,4548,1806,3103,4326,1654,4114,1466,4
928,3331,4570,2687,3755,3567,4742,1230,4584,2673,2441,3593,2025,2819,5200,5032,1492,2217,3558
,1801,1459,4917,4319,3900,2228,2826,4789,1298,1896,3399,4980,2079,2845,3997,4148,4974,1608,39
63,3709,1862,2046,3194,4183,5051,2274,2610,2422,4513,3736,3504,4721,1253,3160,4345,1637,4177,
1405,2280,3352,1865,4719,3158,3964,4973,4389,2842,3990,5069,4987,2628,1891,4170,1402,2287,335
5,3167,2889,4342,1630,3503,4726,1254,4514,3731,2425,2617,4184,5056,2273,2041,3193,3952,2880,1
639,4179,4945,1853,3738,2048,2874,4710,1262,3535,3707,4522,3363,5094,4146,1434,4374,1606,3151
,2083,3397,2245,5060,4380,2077,3999,1296,2413,2621,1898,5058,2873,2619,4728,1854,4942,2289,31
69,3955,2887,2626,1291,2414,4387,2070,3390,2242,5067,4989,4373,1601,3156,2084,3364,5093,4141,
1433,3700,4525,4717,1265,3532,2440,3592,4585,2672,1493,5033,2216,2818,2024,5201,1467,4929,411
5,3330,3102,1655,4327,3566,1231,4743,4571,2686,3754,2827,2229,4788,1800,3559,4318,3901,1458,4
916,4576,1838,2681,3753,3561,1236,4744,3939,3105,1652,4320,1460,4112,3337,2023,5206,1494,5034
,2211,4582,2675,2447,3595,3308,4911,3906,4549,1807,2478,3798,1699,2820,1664,4316,3133,3301,49
18,1456,4124,3765,4540,4772,3557,2485,3791,2643,4786,2471,1690,2829,2015,2227,5002,3568,1831,
2688,4927,1469,3930,4329,2218,2816,2220,5005,1697,2012,4781,2476,3796,2644,4775,3550,2482,376
2,1809,4547,3306,1451,4123,1663,4311,3908,3134,2449,2811,5208,3937,4920,3339,1836,4578,1238,1
944,4638,3079,2997,3845,4852,2399,2963,5148,2709,3274,4051,5183,1523,4263,1711,2194,3046,4607
,1375,3422,3610,4435,1381,2504,2736,2352,3280,5177,4899,4297,2160,2158,2964,4069,4855,2990,38
42,1729,1943,3628,4290,2167,3889,2355,3287,5170,2731,1988,1386,2503,3617,4432,4600,1372,3425,
4264,1716,2193,3041,3273,5184,1524,5179,4897,4299,3880,2952,2738,1981,4609,1975,4863,3048,387
4,2707,2535,3083,2151,5146,4094,2363,3077,2999,4252,1720,4060,1512,3245,2397,4404,3621,3413,4
636,1344,1718,3873,4058,4864,3619,1972,1986,1388,2169,3887,2955,3289,4890,3414,4631,1343,4403
,3626,4067,1515,3242,2390,3070,4255,1727,5141,4093,2364,3084,2156,2532,2700,3672,4457,1919,13
17,4665,2592,3440,1773,4201,3818,3024,3216,1541,4033,1787,2102,2330,5115,2754,3686,4691,2566,
4830,3229,3827,1328,4468,1926,2559,2901,4696,2561,2753,3681,2337,5112,1780,2939,2105,3211,154
6,4808,4034,1774,4206,3023,1310,4662,2595,3447,3675,4450,2906,2308,1921,2798,3478,3820,4239,1
579,4837,1584,5124,2301,2133,3485,2557,4492,2765,3471,1326,4654,1928,4466,3643,2791,1570,4002
,3227,3829,3015,1742,4230,3688,2568,1789,2930,3816,3218,4801,1917,4459,1319,3012,1745,4237,48
39,1577,4005,3220,4461,3644,2796,3476,1321,4653,4495,2762,3482,2550,2908,2134,1583,5123,2306,
3449,1910,4806,1548,4208,3811,2339,2937,4698,1389,1987,3288,4891,3886,2954,2168,4865,4059,171

9,3872,3618,1973,3085,2157,5140,4092,2365,2701,2533,4402,3627,3415,1342,4630,3071,1726,4254,1514,4066,3243,2391,1974,4608,3875,3049,4862,4298,3881,2953,4896,5178,2739,1980,1513,4061,3244,2396,2998,3076,1721,4253,3412,1345,4637,4405,3620,2534,2706,5147,4095,2362,3082,2150,2991,3843,1728,4854,4068,1942,3629,2965,2159,1373,4601,3424,3616,4433,3272,1525,4057,5185,1717,4265,2192,3040,2354,3286,5171,4291,3888,2166,1387,2502,2730,1989,5149,2962,2708,4639,1945,4853,2398,2996,3844,3078,2737,1380,2505,4296,2161,2353,3281,4898,5176,1710,4262,2195,3047,3275,1522,4050,5182,3611,4434,1374,4606,3423,3483,2551,4494,2763,5122,1582,2307,2135,2909,4004,4838,1576,3221,3013,4236,1744,3477,4652,1320,4460,3645,2797,2936,2338,4699,1911,3448,4209,3810,4807,1549,1929,4467,3642,2790,3470,4655,1327,3014,3828,4231,1743,4003,1571,3226,2132,5125,1585,2300,4493,2764,3484,2556,3219,4800,3817,1318,1916,4458,2569,3689,1788,2931,4207,1775,3022,3210,4035,1547,4809,3674,4451,4663,1311,2594,3446,2752,3680,4697,2560,1781,2104,2938,2336,5113,3479,1920,2799,1578,4836,3821,4238,2309,2907,2331,5114,1786,2103,4690,2567,2755,3687,4664,1316,2593,3441,3673,4456,1918,3217,4032,1540,4200,1772,3025,3819,2558,2900,3826,4831,3228,4469,1927,1329,5109,2922,2748,4679,1905,4813,3038,3804,2777,4480,3497,2545,2121,2313,5136,1596,4222,1750,3007,3235,4010,1562,3651,2783,4474,4646,1334,3463,3803,1768,4028,4814,1902,3669,2589,2119,2925,4641,1333,3464,3656,2784,4473,3232,4017,1565,4225,1757,3000,2314,5131,1591,2126,3490,2542,2770,4487,1934,4648,3009,3835,4822,2913,5138,1598,2779,3499,4021,1553,3204,3036,4213,1761,2580,3452,4677,1305,4445,3660,2574,4683,2746,3694,5107,2322,2110,1795,4489,2128,2914,4019,4825,1759,3832,3658,1933,2117,1792,5100,2325,2741,3693,2573,4684,4442,3667,2587,3455,4670,1302,3031,4214,1766,4026,1554,3203,2371,5154,4086,3091,2143,2527,2715,1356,4624,3401,3633,4416,1958,3257,2385,1500,4072,1732,4240,3859,3065,1993,2518,3892,2940,4885,3866,2188,4871,3268,4429,1967,1369,1735,4247,3062,3250,2382,1507,4849,4075,3634,4411,1351,4623,3406,2712,2520,2978,3096,2144,2376,5153,4081,3439,1960,1538,4876,5198,3861,4278,4882,2349,3895,2947,1994,1969,4427,3602,3430,1367,4615,3868,2186,3054,1703,4271,1531,4043,5191,3266,2172,4285,5165,2340,3292,2724,2516,1393,3259,4840,2985,3857,1358,1956,4418,2529,4088,2971,2511,1394,2723,5162,2347,3295,2949,2175,4282,4878,1536,4044,5196,3261,2181,3053,1704,4276,3437,1360,4612,4420,3605,2976,3098,2378,1951,3408,4249,2982,3850,4847,1509,1758,3833,4824,4018,3659,1932,4488,2915,2129,2586,3454,1303,4671,4443,3666,1555,4027,3202,3030,1767,4215,5101,2324,2116,1793,2572,4685,2740,3692,1599,5139,2912,3498,2778,4649,1935,4823,3834,3008,2747,3695,2575,4682,2111,1794,5106,2323

Testing (300 faces):

3037,1760,4212,1552,4020,3205,4444,3661,2581,3453,1304,4676,2924,2118,4815,4029,3802,1769,2588,1903,3668,2127,2315,1590,5130,2771,4486,3491,2543,3657,2785,4472,1332,4640,3465,1756,4224,3001,3233,1564,4016,1904,4678,3805,3039,4812,2923,5108,2749,3234,1563,4011,1751,4223,3006,1335,4647,3462,3650,2782,4475,3496,2544,2776,4481,2312,1597,5137,2120,2180,3052,4277,1705,4045,5197,4879,1537,3260,4421,3604,3436,4613,1361,2722,2510,1395,2174,2948,4283,5163,2346,3294,3409,1950,4846,1508,4248,2983,3851,2379,3099,2977,5164,2341,3293,2173,4284,2517,1392,2725,3431,4614,1366,1968,4426,3603,4042,5190,1530,3267,2187,3055,3869,4270,1702,2528,2970,4089,2984,3856,3258,4841,1957,4419,1359,2521,2713,2377,5152,4080,3097,2145,2979,3251,2383,4074,1506,4848,4246,1734,3063,4622,1350,3407,3635,4410,3894,2946,4883,2348,1995,1961,3438,3860,4279,5199,1539,4877,3632,4417,1959,4625,1357,3400,4241,1733,3064,3858,3256,2384,4073,1501,3090,2142,2370,5155,4087,2714,2526,4870,3269,2189,3867,1368,4428,1966,2519,1992,4884,3893,2941,5018,2833,2659,4768,1814,4902,3915,3129,2666,4591,2454,3586,5215,2030,2202,1487,5027,1641,4333,3116,3324,1473,4101,2692,3740,4565,1225,4757,3572,3912,1679,4905,4139,1813,3778,2498,2834,2008,4750,3575,2695,3747,4562,3323,1474,4106,1646,4334,3111,2205,1480,5020,5212,2037,2453,3581,2661,4596,1825,47

59,3924,3118,4933,2802,1489,5029,2668,3588,1442,4130,3315,3127,1670,4302,3543,2491,4766,4554,3771,2465,4792,3785,2657,5016,2233,2001,1684,4598,2805,2039,4934,4108,1648,3923,3749,1822,2006,1683,5011

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?

Cara original	MeanFace + 1 comp	MeanFace + 3 comp	MeanFace + 10 comp	MeanFace + 95% comp
1				
2				
3				
4				

Utiliza los datos de testing. Y envía un archivo a kagle de los datos de testing con la primera componente. Recuerde que el testing no puede ser utilizado para aprender PCA.
https://www.kaggle.com/t/676b6edd04c64a1cbd83e7b6daf9b59b

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.

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

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

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.

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.