


Reporte Análisis de Datos de Fuentes Diversas

Elaborado por: Marcos Ulises Sánchez.

Docente: Msc. Arlen López.

18 de septiembre de 2024

Descripción de la fuente y almacenamiento de datos

El dataset utilizado contiene información sobre el café en diferentes países durante varios años. Fue extraído de Kaggle en el siguiente enlace: [Worldwide Coffee Habits Dataset](https://www.kaggle.com/datasets/colinhealy/worldwide-coffee-habits-dataset)  [\(kaggle.com\)](https://www.kaggle.com/datasets/colinhealy/worldwide-coffee-habits-dataset). Las variables incluyen:

- **Country:** El nombre del país donde se recolectaron los datos.
- **Year:** El año de recolección de los datos, desde el año 2000 a 2023.
- **Coffee Consumption (kg per cápita per year):** Es la cantidad de café que consume una persona anualmente
- **Average Coffee Price (USD per kg):** El precio promedio del kilogramo de café en dólares americanos.
- **Type of Coffee Consumed:** Café más popular en los diferentes países.
- **Population (millions):** La población estimada de cada país en millones de habitantes.

En total, encontramos 10000 registros almacenados en un archivo de tipo CSV.

Análisis Descriptivo

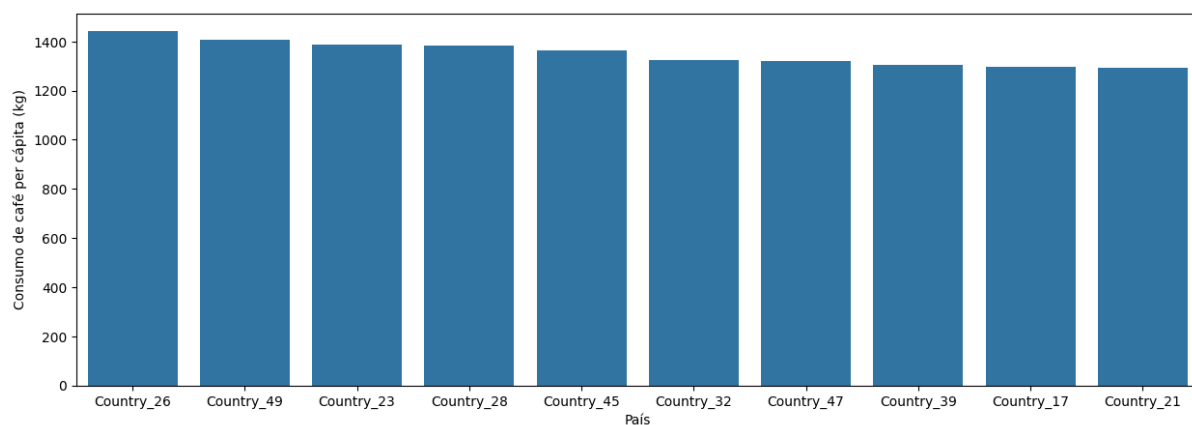
	Type of Coffee	Frequency
0	Latte	2071
1	Cappuccino	2001
2	Mocha	1984
3	Americano	1975
4	Espresso	1969

	Year	Coffee Consumption (kg per capita per year)	Average Coffee Price (USD per kg)	Population (millions)
count	10000.000000	10000.000000	10000.000000	10000.000000
mean	2011.666900	6.061865	9.461891	75.167120
std	6.911695	2.313427	3.151403	43.023176
min	2000.000000	2.000385	4.000742	1.002494
25%	2006.000000	4.070743	6.728261	37.465847
50%	2012.000000	6.094491	9.458371	75.021943
75%	2018.000000	8.061127	12.136285	112.595868
max	2023.000000	9.999399	14.997053	149.995850

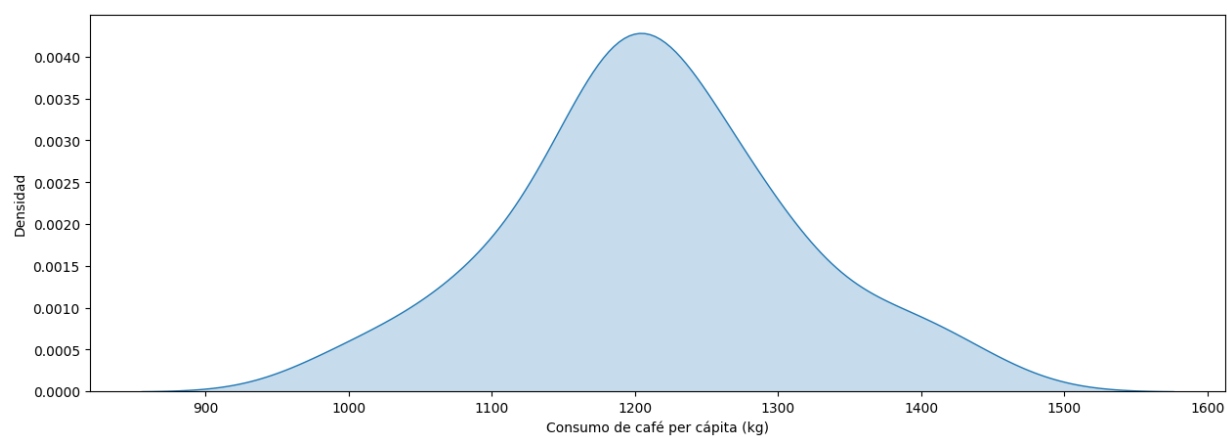
Mostramos las estadísticas descriptivas usando Python, también generamos dos tablas de distribución de frecuencias para ver cuantos registros por país hay y cuantos registros por tipo de café hay presentes. Se tienen registros de 50 países, no mostramos la tabla que se puede consultar en el notebook adjunto. Mostramos la distribución de frecuencia de los tipos de café.

	Type of Coffee	Frequency
0	Latte	2071
1	Cappuccino	2001
2	Mocha	1984
3	Americano	1975
4	Espresso	1969

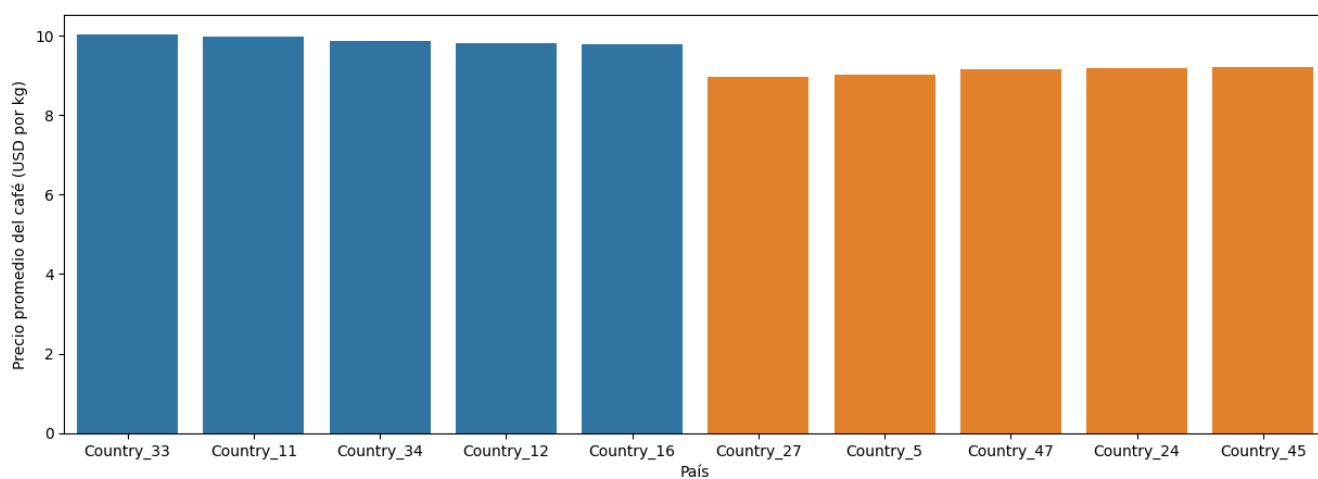
Notaremos también, que el consumo per cápita de café es una variable que no cambia mucho. En el top 10:

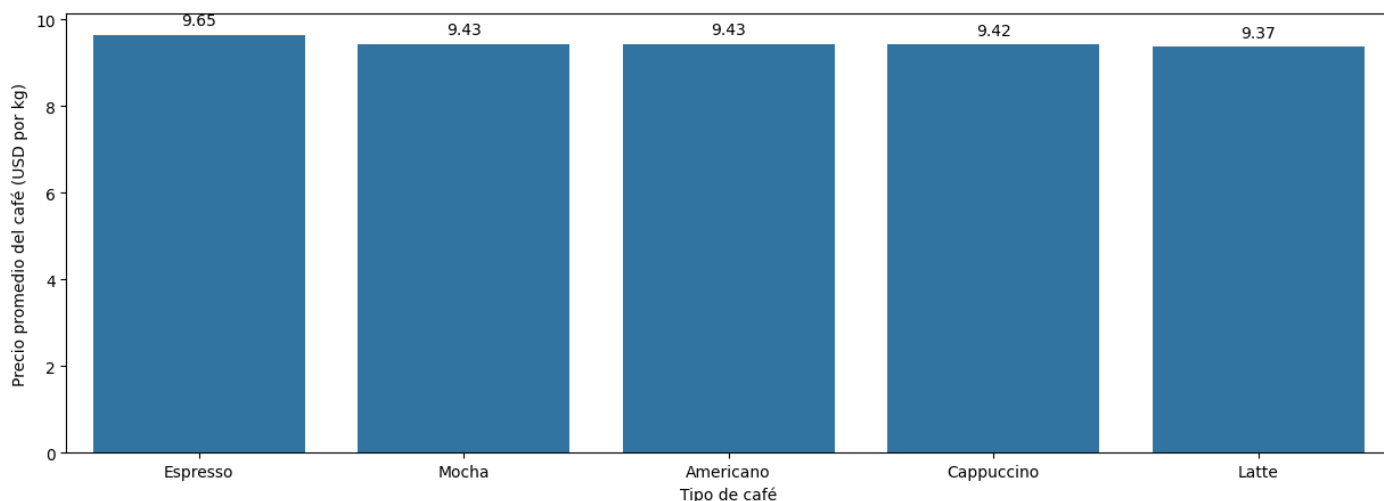


Entre los distintos tipos de café, el consumo total registrado es demasiado similar entre los países con mayor consumo total. También tenemos la distribución:



Para analizar el precio promedio del café (tomamos todos los tipos en cuenta), tomamos los extremos para ver qué tanto distan entre sí.



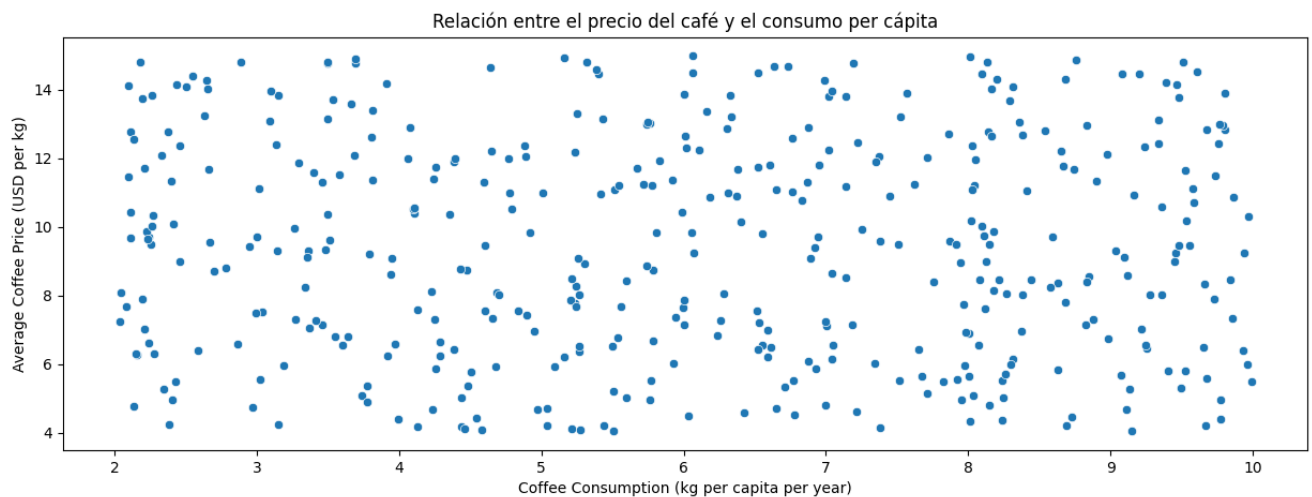


Patrones y relaciones encontradas

Podemos observar dos patrones importantes, según la lógica, pues, los datos pueden decir lo contrario o llevarnos a resultados que no son estadísticamente significativos.

Si comparamos el consumo per-cápita y el precio promedio del kg del café, deberíamos poder encontrar una correlación entre estas variables en un modelo de regresión lineal simple, en el que el consumo es la variable dependiente y el precio la independiente. Esto, según la teoría, pero cuando realizamos nuestra regresión lineal esto no parece ser así, es decir, que probablemente el consumo sea afectado por valores que no tenemos en consideración, cómo la cultura del país o el acceso a productos sustitutos.

Otra relación obvia es que, mientras mayor sea la población, mayor será el consumo de café reportado. En este caso, si encontramos una relación respaldada por las estadísticas cuando tomamos el logaritmo de la variable consumo.



La primera regresión presentada fue usando solamente datos del 2023, es decir, de corte transversal, que serían diversas observaciones en un período de tiempo determinado. Se intentó con otros años, pero no se pudo llegar a ninguna conclusión.

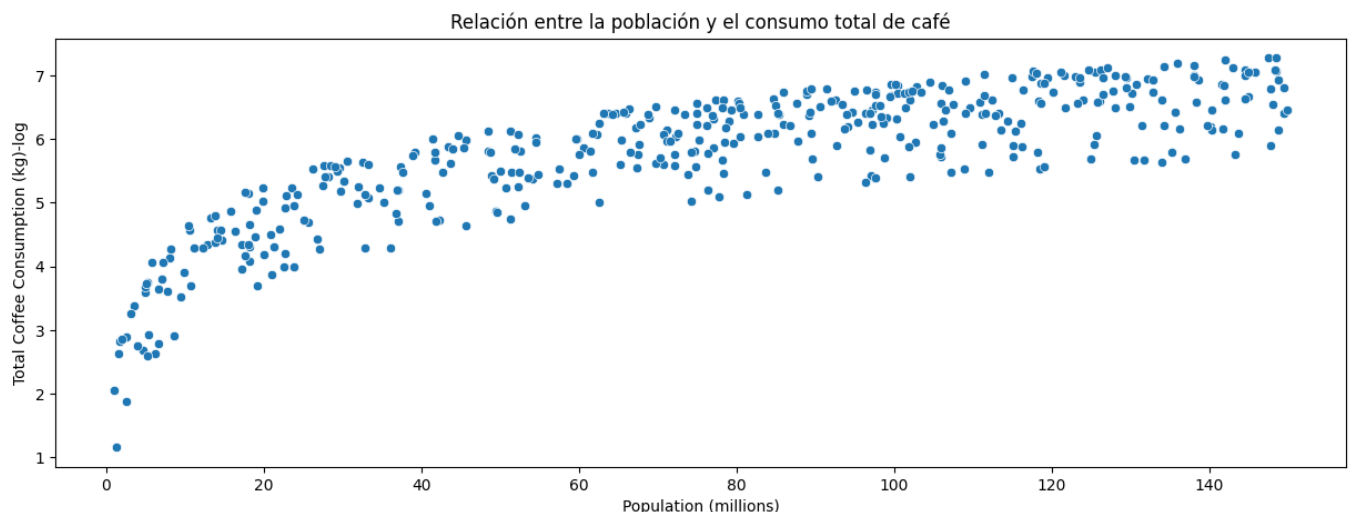
...

OLS Regression Results							
Dep. Variable:	Average Coffee Price (USD per kg)-log			R-squared:	0.001		
Model:	OLS			Adj. R-squared:	-0.001		
Method:	Least Squares			F-statistic:	0.5678		
Date:	Wed, 18 Sep 2024			Prob (F-statistic):	0.452		
Time:	22:40:32			Log-Likelihood:	-174.19		
No. Observations:	419			AIC:	352.4		
Df Residuals:	417			BIC:	360.5		
Df Model:	1						
Covariance Type:	nonrobust						
		coef	std err	t	P> t	[0.025	0.975]
	const	2.2261	0.071	31.406	0.000	2.087	2.365
Coffee Consumption (kg per capita per year)-log		-0.0302	0.040	-0.754	0.452	-0.109	0.049
Omnibus:	72.478	Durbin-Watson:	1.941				
Prob(Omnibus):	0.000	Jarque-Bera (JB):	26.232				
Skew:	-0.392	Prob(JB):	2.01e-06				
Kurtosis:	2.057	Cond. No.	9.11				

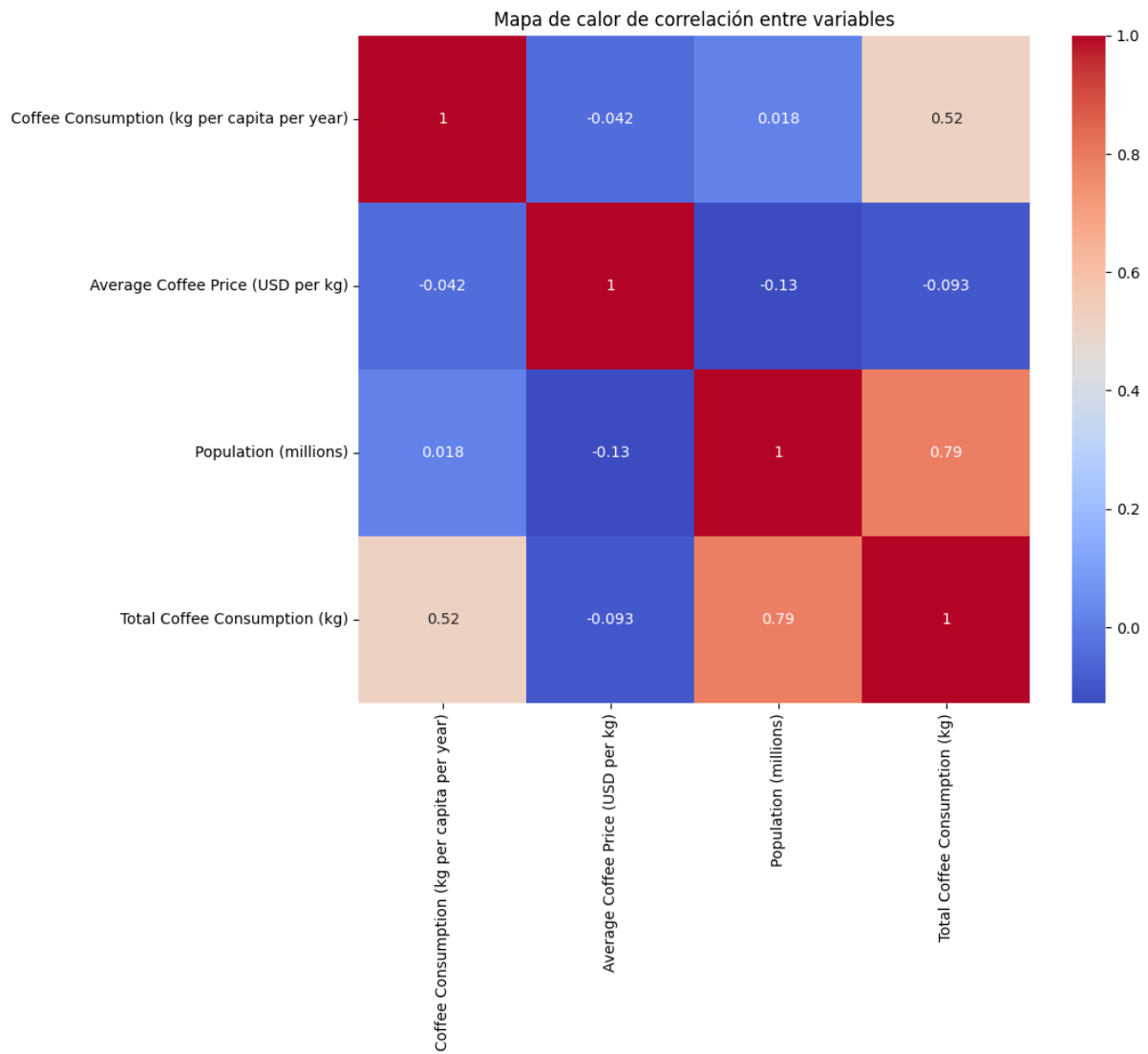
El p-value de 0.452 nos hace que no podamos rechazar la hipótesis nula de coeficiente 0. Este modelo log-log fue con el que obtuvimos un p-value más alto. No se incluyen en el código las otras regresiones que fallaron, para agilizar el proceso de codificación solo se editaba sobre un modelo.

La segunda relación analizada fue entre el logaritmo del consumo y la población. La población se encuentra en millones y será más lógico analizar un cambio porcentual en lugar de unitarios en los precios del café.

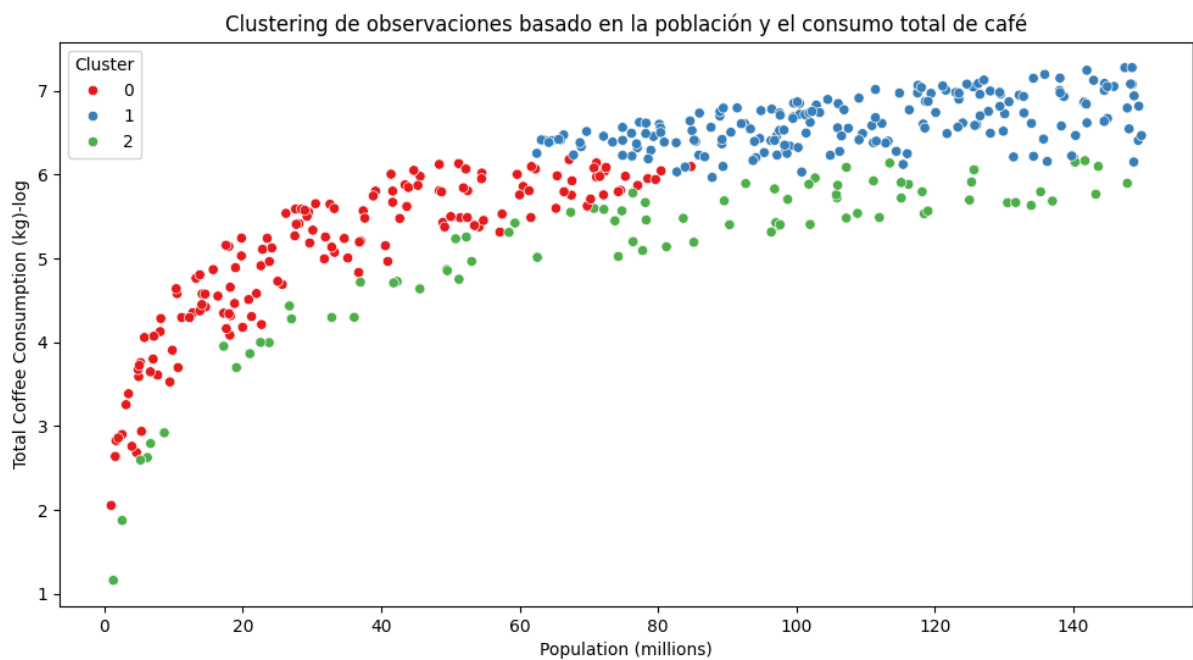
Un valor de R^2 en el que el 66.4% de las variaciones del consumo son explicadas por cambios en la población están bastante bien. El modelo es significativo por los p-values cercanos a 0, y un estadístico de Fisher alto que en relaciones lineales simples no explica más que la existencia de una relación lineal entre población y logaritmo del consumo. Sobre el significado de los coeficientes, se puede inferir que, con un incremento de 1 millón en la población, el consumo de café incrementará en un 2%. También, un estadístico de Durbin Watson cercano a 2 muestra poca presencia de correlación de los errores.



Un mapa de correlación entre variables nos muestra lo ya verificado en la regresión, además de ver que el resto de las variables no tomadas en consideración no están correlacionadas, y que tampoco tendría sentido hacerlo en algunos casos, por ejemplo, consumo total y consumo per-cápita. Este gráfico resalta la fuerte correlación entre la población y el consumo total de café, 0.81. Esta relación intuitiva sugiere que las naciones más pobladas consumen más café en términos absolutos.



Clustering usando K-means



	Coffee Consumption (kg per capita per year)	Average Coffee Price (USD per kg)	Population (millions)	Total Coffee Consumption (kg)
0	0.571082	9.445771	35.419008	212.536707
1	0.662302	8.850299	108.236938	780.401474
2	-0.014235	10.410085	81.027707	230.344780

Para el algoritmo de K-means, se normalizaron todas las variables, para asegurarnos que variables con valores muy grandes no dominen el cálculo de distancias del algoritmo. Se normalizaron:

- Consumo de café per cápita
- Precio promedio de café
- Población
- Consumo total

Dividimos las observaciones en tres clusters, agrupados por su similitud con las variables estandarizadas. Los puntos verdes podríamos decir que representan consumos de bajo a moderados para su población, en el cluster azul están las poblaciones altas y con mucho consumo, que tienen de hecho el menor precio promedio, y finalmente, en el cluster rojo, nos muestran observaciones bajas con un consumo moderado-alto, con un precio de café intermedio.

En esta clase práctica, entendimos primero el comportamiento de los datos, observaciones, columnas, además de cómo estaban distribuidas las variables. Encontramos correlaciones entre variables y también transformamos datos para adaptarse a modelos de machine learning como K-means. En este análisis, se usaron principalmente las observaciones del año 2023, pero, es posible replicar estos patrones para cada año. También, notamos que es poco significativo el tipo de café que se venda, pues el precio promedio no variaba mucho y su consumo era similar.