Parcial Estadistica 3: Objetivo validar el conocimiento del aprendizaje supervisado y no supervisado. Todos los puntos utilizaran los siguientes datos

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **City** | **Population (millions)** | **GDP (billion USD)** | **Annual Budget (billion USD)** | **Debt (billion USD)** | **Inflation (%)** | **Capital** | **Cluster Label** |
| Mexico City | 9.2 | 411 | 16 | 8 | 4.8 | Yes | 0 |
| São Paulo | 12.4 | 337 | 15.2 | 9.5 | 5.1 | No | 3 |
| Buenos Aires | 15 | 112 | 8.1 | 5.6 | 113.4 | Yes | 0 |
| Rio de Janeiro | 6.7 | 172 | 8.7 | 7.3 | 5.1 | No | 2 |
| Bogotá | 7.4 | 95 | 5.2 | 2.7 | 9.2 | Yes | 1 |
| Lima | 10.7 | 94 | 4.5 | 3.2 | 7.9 | Yes | 3 |
| Santiago | 6.8 | 166 | 6.1 | 4.8 | 6.6 | Yes | 1 |
| Caracas | 2.9 | 30 | 2.5 | 3 | 398.4 | Yes | 3 |
| Medellín | 2.5 | 40 | 3.3 | 2 | 9.2 | No | 3 |
| Quito | 2.8 | 26 | 2.1 | 1.5 | 3.7 | Yes | 0 |
| Guadalajara | 5.2 | 50 | 3.9 | 2.2 | 4.8 | No | 0 |
| Monterrey | 4.9 | 93 | 5 | 3.4 | 4.8 | No | 1 |
| Curitiba | 3.5 | 44 | 2.6 | 2.1 | 5.1 | No | 3 |
| Porto Alegre | 4.3 | 55 | 3 | 2.2 | 5.1 | No | 0 |
| Salvador | 2.9 | 42 | 3.3 | 2 | 5.1 | No | 2 |
| Recife | 1.6 | 26 | 1.9 | 1.4 | 5.1 | No | 0 |
| Fortaleza | 3.9 | 34 | 2.5 | 2.1 | 5.1 | No | 2 |
| Brasília | 4.7 | 66 | 5.8 | 3 | 5.1 | Yes | 1 |
| Belo Horizonte | 2.7 | 57 | 4.2 | 2.9 | 5.1 | No | 0 |
| Havana | 2.2 | 15 | 1.2 | 0.8 | 40 | Yes | 0 |
| Santo Domingo | 3.4 | 30 | 1.7 | 1.1 | 4.8 | Yes | 3 |
| Panama City | 1.5 | 60 | 2.5 | 1.8 | 2.3 | Yes | 3 |
| San Salvador | 1.1 | 20 | 1.2 | 1 | 4.7 | Yes | 3 |
| La Paz | 0.8 | 10 | 0.6 | 0.7 | 2.9 | Yes | 0 |
| Santa Cruz | 1.5 | 12 | 0.8 | 0.5 | 2.9 | No | 3 |
| Asunción | 0.7 | 12 | 0.9 | 0.6 | 4 | Yes | 1 |
| Maracaibo | 1.6 | 14 | 1.1 | 0.8 | 398.4 | No | 3 |
| Córdoba | 1.5 | 21 | 1.7 | 1 | 113.4 | No | 1 |
| Rosario | 1.3 | 18 | 1.4 | 0.9 | 113.4 | No | 3 |
| Managua | 1 | 8 | 0.7 | 0.5 | 7.9 | Yes | 1 |
| Montevideo | 1.8 | 45 | 2.2 | 1.7 | 4.7 | Yes | 3 |

[datos\_parcial\_202402](https://docs.google.com/spreadsheets/u/0/d/1Jzds2vQXM2CMPl-eG3-vaY-leMQwrmDuh07pp7MqJnE/edit)

Punto 1(10) incluir las siguientes actividades de exploracion de datos:

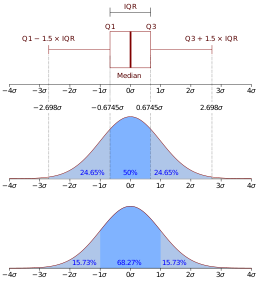
* calcular la correlacion entre population y GDP. Cual es el numerador y denominador

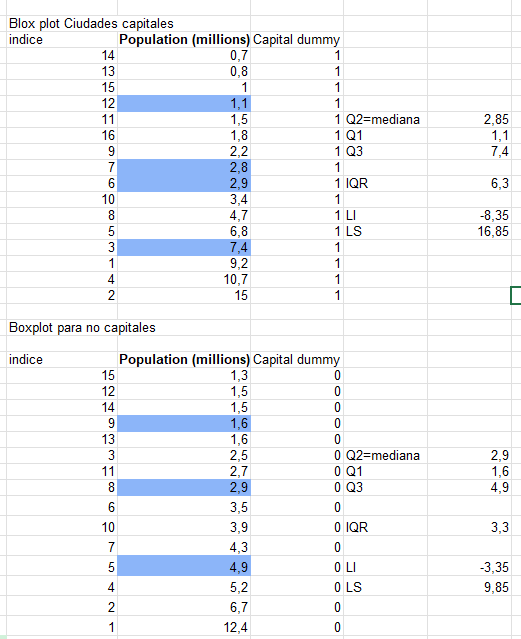
numerador = 217690,3

denominador = 304613,0321

correlación = 0,714645393

* dibujar un boxplot para la variable poblacion para las ciudades capitales y no capitales. Incluir los numeros de Q1, median, Q3 and Q1- 1.5 \* IQR AND Q3+ 1.5 \* IQR



Todo está en el Excel y en el cuaderno, pero en sí lo grafiqué con: 

Punto 2 (15), Cree un cluster donde cada punto es una ciudad utilizando todas las variables incluyendo si es capital o no. La columna cluster label es la asignacion a aleatoria a los 4 clusters. No puede utilizar librerias. Crear 4 clusters utilizando el k-means (Si el k-means toma mas de 3 iteraciones, utiliza la iteracion 3 como la ultima).

1. Cual es el centroide de la iteración actual para el cluster 1, 2 , 3 y 4?

C1= 

C2=

C3= 

C4 =

1. Cual es el nuevo label para los puntos en la iteración 2?

Tabla

Descripción generada automáticamente

1. Cual es el centroide para la iteración 2?

C1=

C2= 

C3=

C4= 

1. Cuantas iteraciones son necesarias para que el k-mean termine?

4

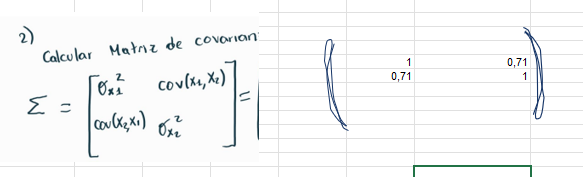
1. Cual es el cluster final para cada punto? . Cuales ciudades estan en el mismo grupo y describa que significa cada centroide.

Lo que dice es que la ciudad teniendo en cuenta todos los datos de las columnas, se pueden agrupar de acuerdo al cluster más cercano

|  |  |  |  |
| --- | --- | --- | --- |
| Grupo 1 | Grupo 2 | Grupo 3 | Grupo 4 |
| São Paulo  Mexico City | Brasília  Rosario  Córdoba  Medellín  Curitiba  Santo Domingo  Panama City  San Salvador  Santa Cruz  Montevideo  Salvador  Fortaleza  Asunción  Managua  Quito  Guadalajara  Porto Alegre  Recife  Belo Horizonte  Havana  La Paz | Buenos Aires  Lima  Bogotá  Monterrey  Rio de Janeiro  Santiago | Caracas  Maracaibo |

Punto 3 (15) Reducir las dimensionalidad de las ciudades de 2 dimensiones a una utilizando PCA.

1. Colocar la matriz de varianza covarianza.



1. Calcular los eigenvalues. Incluir ecuacion cuadratica.

Están en el cuaderno,

X1 = 0.157

X2 = 1.842

1. Cuanta varianza puede explicar el primer eigne vector

X1 explica el 7.85%

X2 explica el 92,1 %

1. Cuales son los eigen vectors

Eigenvector para 0.157

(0.707 , 0.707)

Eigenvector para 1.842

(-0.707 ,- 0.707)

1. Ordene las ciudades de mayor a menor utilizando la primera componente y adicionarlas las primeras 10 la siguiente tabla.

Punto 4 (25) crear una regresion a mano utilizando (X=Poblacion y Y=GDP) los datos de la tabla. Utilizar los datos training que son las ciudades diferentes de (Mexico City,Buenos Aires,

Quito,Guadalajara,Porto Alegre,Recife,Belo Horizonte,Havana,La Paz) como training y las ciudades ((Mexico City,Buenos Aires, Quito,Guadalajara,Porto Alegre,Recife,Belo Horizonte,Havana,La Paz) testing.

1. Calcular b1 (numerador y denominador) y bo (numerador y denominador), para crear la ecuación de la regresión (Solo utilizar train).

B1 = 20,37

B0 = -12,36

1. Cual es el MSE y MAPE para train y test.

MAPE y MSE TRAIN





MAPE y MSE TEST





1. Cual es el R2, el RSS y el TSS.







Punto 5 (15) crea un 3nn utilizando los datos de la misma tabla (X=Poblacion y Y=GDP)

1. Cual es el MSE y MAPE para train y test.



1. Cual modelo funciona mejor

La regresión, tiene un mapa de 0.43