**Tarea 1: Exploración de Datos**

**Esteban Colorado González**

**Universidad Católica de Oriente – Estadística III**

**Tabla de contenido**

[1. Utilizas el siguiente set de datos para calcular paso por paso (mostrar procedimiento y fórmulas) 3](#_Toc128516669)

[1.1. ¿Cuál es la media, mediana y desviación estándar?, y la moda y los valores repeticiones de la moda para los datos categóricos. 3](#_Toc128516670)

[1.2. Dibujar un boxplot a mano. Utilizando los datos de la tabla 1 y las siguientes proporciones. 3](#_Toc128516671)

[1.3. Cuál es la covarianza entre las 2 variables X1, X2 4](#_Toc128516672)

[1.4. Cuál es la correlación entre la variable x1 y x2 (Calcularla a mano). Correlación puede ser escrita también como: 4](#_Toc128516673)

[1.5. Explica la relación entre covarianza y correlación. 4](#_Toc128516674)

[1.6. Calcule el resultado del algoritmo K-means sobre este set de datos. Vamos a crear 2 grupos, es decir, k=2 (2 clusters). 5](#_Toc128516675)

# 1. Utilizas el siguiente set de datos para calcular paso por paso (mostrar procedimiento y fórmulas)

|  |  |  |
| --- | --- | --- |
| **ID** | **X1** | **X2** |
| 1 | 1 | 4 |
| 2 | 1 | 3 |
| 3 | 0 | 4 |
| 4 | 5 | 1 |
| 5 | 6 | 2 |
| 6 | 4 | 0 |

## 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 (X1) = (1 + 1 + 0 + 5 + 6 + 4) / 6 = 2.83

Media (X2) = (4 + 3 + 4 + 1 + 2 + 0) / 6 = 2.33

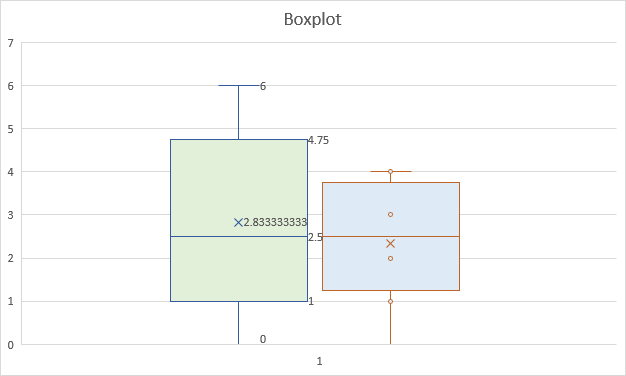
Mediana (X1) = 0, 1, 1, 4, 5, 6 = 1, 4 = (1 + 4) / 2 = 2.5

Mediana (X2) = 0, 1, 2, 3, 4, 4 = 2, 3 = (2 + 3) / 2 = 2.5

Desviación Estándar (X1) = , donde es la media y n es el número de datos = 2.48

Desviación Estándar (X2) = , donde es la media y n es el número de datos = 1.63

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



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

X1’ = (1 + 1 + 0 + 5 + 6 + 4) / 6 = 2.83

X2‘ = (4 + 3 + 4 + 1 + 2 + 0) / 6 = 2.33

**Covarianza (X1, X2)** = ((1 - 2.83) \* (4 - 2.3) + (1 - 2.83) \* (3 - 2.3) + (0 - 2.83) \* (4 - 2.3) + (5 - 2.83) \* (1 - 2.3) + (6 - 2.83) \* (2 - 2.3) + (4 - 2.83) \* (0 - 2.3)) / 6 = -2.611

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

Correlación (X1, X2) = ((1 - 2.83) \* (4 - 2.3) + (1 - 2.83) \* (3 - 2.3) + (0 - 2.83) \* (4 - 2.3) + (5 - 2.83) \* (1 - 2.3) + (6 - 2.83) \* (2 - 2.3) + (4 - 2.83) \* (0 - 2.3)) / (\* )) = -0.77

## 1.5. Explica la relación entre covarianza y correlación.

correlación = covarianza / (desviación estándar de la variable 1 x desviación estándar de la variable 2)

En este caso, si sustituimos los valores conocidos, obtenemos:

correlación = -2.611 / (2.48 x 1.63) = -0.775

El valor de la correlación calculado (-0.775) es muy cercano al valor de la correlación que se nos dio (-0.77). Esto sugiere que la correlación dada es consistente con los valores de covarianza y desviaciones estándar proporcionados. Además, al tener en cuenta las desviaciones estándar de ambas variables, podemos decir que hay una fuerte correlación negativa entre las dos variables, lo que sugiere que cuando una variable aumenta, la otra tiende a disminuir y viceversa, y que la magnitud de esta relación es relativamente alta.

## 1.6. Calcule el resultado del algoritmo K-means sobre este set de datos. Vamos a crear 2 grupos, es decir, k=2 (2 clusters).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **X1** | **X2** | **X1-Mean** | **X2-Mean** | **Random** |
| 2 | 0 | 4 | -2.833333333 | 1.666666667 | 0 |
| 5 | 5 | 1 | 2.166666667 | -1.333333333 | 0 |
| 1 | 1 | 4 | -1.833333333 | 1.666666667 | 1 |
| 3 | 1 | 3 | -1.833333333 | 0.666666667 | 1 |
| 4 | 4 | 0 | 1.166666667 | -2.333333333 | 1 |
| 6 | 6 | 2 | 3.166666667 | -0.333333333 | 1 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **X1** | **X2** | **Iteración 1** | **Iteracion 2** | **New Label** |
| -0.333333333 | 0.166666667 | 8.5 | 12.0625 | 0 |
| -0.333333333 | 0.166666667 | 8.5 | 5.5625 | 1 |
| 0.166666667 | -0.083333333 | 4.5 | 7.0625 | 0 |
| 0.166666667 | -0.083333333 | 2.5 | 4.5625 | 0 |
| 0.166666667 | -0.083333333 | 8.5 | 6.0625 | 1 |
| 0.166666667 | -0.083333333 | 12.5 | 9.0625 | 1 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **X1** | **X2** | **Iteración 1** | **Iteracion 2** | **New Label** |
| -2.166666667 | 1.333333333 | 0.555555556 | 34 | 0 |
| 2.166666667 | -1.333333333 | 25.88888889 | 4.93038E-32 | 1 |
| -2.166666667 | 1.333333333 | 0.222222222 | 25 | 0 |
| -2.166666667 | 1.333333333 | 0.555555556 | 20 | 0 |
| 2.166666667 | -1.333333333 | 24.55555556 | 2 | 1 |
| 2.166666667 | -1.333333333 | 31.22222222 | 2 | 1 |