**Tarea #**: 1

**Realizado por:** Santiago Cárdenas Franco
**Tema**: Exploración de datos
**Fecha entrega**: 11:59 pm Agosto 21 de 2023
**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 ):

| x1 | x2 | x3 |
|---:|---:|---:|
| 4 | 4 | 28 |
| 2 | 3 | 24 |
| 2 | 4 | 30 |
| 3 | 5 | 32 |
| 1 | 3 | 18 |
| 3 | 6 | 41 |
| 3 | 6 | 44 |
| 0 | 1 | 5 |
| 1 | 3 | 18 |
| 0 | 0 | 1 |
| 5 | 9 | 62 |
| 1 | 2 | 17 |
| 2 | 3 | 24 |
| 1 | 3 | 19 |
| 3 | 6 | 42 |
| 4 | 8 | 56 |
| 4 | 8 | 56 |
| 3 | 6 | 44 |
| 5 | 9 | 64 |
| 1 | 2 | 17 |

| | | |
|---|---|---|
| 1 | 2 | 17 |

1.1. ¿Cuál es la media, mediana y desviación estándar?, y la moda y los valores repeticiones de la moda para los datos categóricos.

1.1

Media → Cantidad de datos = 21 ⇒ $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{N}$

$\bar{x}_1 = \frac{4+2+2+3+1+3+3+0+1+0+5+1+2+1+3+4+4+3+5+1+1}{21}$

$\bar{x}_1 = \frac{49}{21} \approx 2,33$

$\bar{x}_2 = \frac{4+3+4+5+3+6+6+1+3+0+9+2+3+3+6+8+8+6+9+2+2}{21}$

$\bar{x}_2 = \frac{93}{21} \approx 4,428$

$\bar{x}_3 = \frac{28+24+30+32+18+41+44+5+18+1+62+17+24+19+42+56+56+44+64+17+17}{21}$

$\bar{x}_3 = \frac{659}{21} \approx 31,380$

Mediana → Se ordenan los valores de menor a mayor, Para impares = $Me = \frac{n+1}{2}$ ⇒ $Me = 11$ (Posición)

$x_1$ = 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5

$Q_1$ $Q_3$

$Me_{x_1} = 2$

$x_2$ = 0, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 5, 6, 6, 6, 6, 8, 8, 9, 9

$Me_{x_2} = 4$

$x_3$ = 1, 5, 17, 17, 17, 18, 18, 19, 24, 24, 28, 30, 32, 41, 42, 44, 44, 56, 56, 62, 64

$Me_{x_3} = 28$

Moda → La Moda es el dato que más se repite, se usará la organización de los datos anteriores para identificar la moda.

$Mo_{x_1}$ = 1 ⇒ Se repite 6 veces

$Mo_{x_2}$ = 3 ⇒ Se repite 5 veces

$Mo_{x_3}$ = 17 ⇒ Se repite 3 veces

$x_2$ = 0, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 5, 6, 6, 6, 6, 8, 8, 9, 9

$Me_{x_2} = 4$

$x_3$ = 1, 5, 17, 17, 17, 18, 18, 19, 24, 24, 28, 30, 32, 41, 42, 44, 44, 56, 56, 62, 64

$Me_{x_3} = 28$

Moda → La Moda es el dato que más se repite, se usará la organización de los datos anteriores para identificar la moda.

$Mo_{x_1}$ = 1 ⇒ Se repite 6 veces

$Mo_{x_2}$ = 3 ⇒ Se repite 5 veces

$Mo_{x_3}$ = 17 ⇒ Se repite 3 veces

Para $x_1$, $x_2$ y $x_3$ ⇒ Tienen en común el dato 1 una sola vez

Para $x_1$ y $x_2$ ⇒ Tiene en común los datos 0 (una vez), 1 (una vez), 2 (tres veces), 3 (cinco veces), 4 (Cuatro dos veces), y el 5 (una vez), la moda entre estos dos es 3 que se repite 5 veces.

Desviación Estandar → $S = \sqrt{\frac{\sum_{i=1}^{n} (x-\bar{x})^2}{N}}$

$S_{x_1} = \sqrt{\frac{(1,66)^2+(-0,33)^2+(-0,33)^2+(0,66)^2+(-1,33)^2+(0,66)^2+(0,66)^2+(-2,33)^2+(-1,33)^2+(-2,33)^2+(2,66)^2+(-1,33)^2+(-1,33)^2+(-1,33)^2 \dots}{21}}$

$\dots \sqrt{\frac{(0,66)^2+(1,66)^2+(1,66)^2+(0,66)^2+(2,66)^2+(-1,33)^2+(1,33)^2}{21}} \approx \sqrt{\frac{48,96}{21}} \approx \sqrt{2,331} \approx 1,527$

$$S_{x_2} = \sqrt{\frac{(-0{,}428)^2+(-1{,}428)^2+(-0{,}428)^2+(0{,}571)^2+(-1{,}428)^2+(1{,}571)^2+(1{,}571)^2+(-3{,}428)^2+(-1{,}428)^2 \dots}{21}}$$

$$\dots\sqrt{\frac{(-4{,}428)^2+(4{,}571)^2+(-2{,}428)^2+(-1{,}428)^2+(-1{,}428)^2+(1{,}571)^2+(3{,}571)^2+(3{,}571)^2+(1{,}571)^2+(0{,}571)^2 \dots}{21}}$$

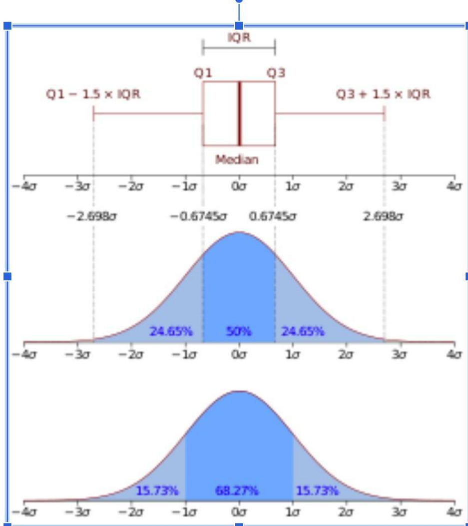
$$\dots\sqrt{\frac{(-2{,}428)^2+(-2{,}428)^2}{21}} \approx \sqrt{\frac{143{,}93}{21}} \approx \sqrt{6{,}853} \approx 2{,}618$$

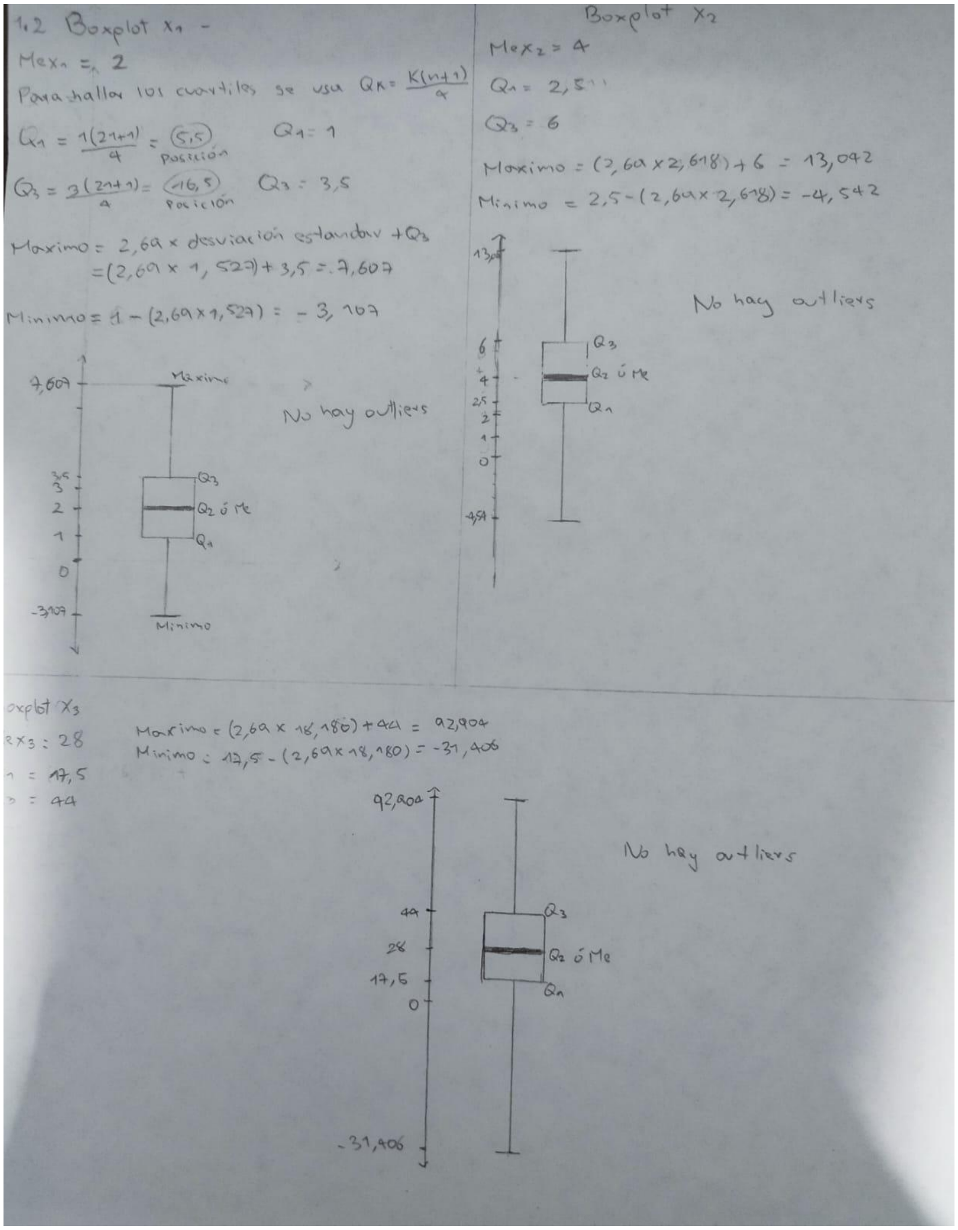
$$S_{x_3} = \sqrt{(-3{,}380)^2+(-7{,}380)^2+(-1{,}380)^2+(0{,}619)^2+(-13{,}380)^2+(0{,}619)^2+(12{,}619)^2+(-26{,}380)^2+(-13{,}380)^2+(-30{,}380)^2 \dots}$$

$$\dots\sqrt{(30{,}619)^2+(-14{,}380)^2+(-7{,}380)^2+(-12{,}380)^2+(10{,}619)^2+(24{,}619)^2+(24{,}619)^2+(12{,}619)^2+(32{,}619)^2+\dots}$$

$$\dots\sqrt{(-14{,}380)^2+(-14{,}380)^2} \approx \sqrt{\frac{6940{,}76}{21}} \approx \sqrt{330{,}512} \approx 18{,}180$$

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





1.2 Boxplot $X_1$ -

$Me_{X_1} = 2$

Para hallar los cuartiles se usa $Q_K = \frac{K(n+1)}{4}$

$Q_1 = \frac{1(21+1)}{4} = 5{,}5$ posición $\quad Q_1 = 1$

$Q_3 = \frac{3(21+1)}{4} = 16{,}5$ posición $\quad Q_3 = 3{,}5$

Maximo = 2,69 × desviación estandar + $Q_3$
$= (2{,}69 \times 1{,}527) + 3{,}5 = 7{,}607$

Minimo $= 1 - (2{,}69 \times 1{,}527) = -3{,}107$



Boxplot $X_2$

$Me_{X_2} = 4$

$Q_1 = 2{,}5$

$Q_3 = 6$

Maximo $= (2{,}69 \times 2{,}618) + 6 = 13{,}042$

Minimo $= 2{,}5 - (2{,}69 \times 2{,}618) = -4{,}542$



Boxplot $X_3$

$Me_{X_3} = 28$

$Q_1 = 17{,}5$

$Q_3 = 44$

Maximo $= (2{,}69 \times 18{,}180) + 44 = 92{,}904$

Minimo $= 17{,}5 - (2{,}69 \times 18{,}180) = -31{,}406$



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

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

1,3

Covarianza entre $x_1$, $x_2$ $\Rightarrow$ $Cov(x_1, x_2) = \dfrac{\sum_i (x_{1i} - \bar{x}_1) * (x_{2i} - \bar{x}_2)}{N}$

$x_1 = 4, 2, 2, 3, 1, 3, 3, 0, 1, 0, 5, 1, 2, 1, 3, 4, 4, 3, 5, 1, 1$

$x_2 = 4, 3, 4, 5, 3, 6, 6, 1, 3, 0, 9, 2, 3, 3, 6, 8, 8, 6, 9, 2, 2$

$$Cov(x_1, x_2) = \frac{(1{,}66 * (-0{,}428)) + (-0{,}333 * (-1{,}428)) + (-0{,}333 \times (-0{,}428)) + (0{,}667 \times 0{,}571) + (-1{,}333 \times (-1{,}428))}{21} \ldots$$

$$\ldots \frac{(0{,}667 \times 1{,}571) + (0{,}667 \times 1{,}571) + (-2{,}333 \times (-3{,}428)) + (-1{,}333 \times (-1{,}428)) + (-2{,}333 \times (-4{,}428)) + (2{,}667 \times 4{,}571))}{21} \ldots$$

$$\ldots \frac{(-1{,}333 \times (-2{,}428)) + (-0{,}333 \times (-1{,}428)) + (-1{,}333 \times (-1{,}428)) + (0{,}667 \times (1{,}571)) + (1{,}667 \times 3{,}571) + (1{,}667 \times 3{,}571)}{21} \ldots$$

$$\ldots \frac{(0{,}667 \times 1{,}571) + (2{,}667 \times 4{,}571) + (-1{,}333 \times (-2{,}428)) + (-1{,}333 \times (-2{,}428))}{21} \approx \frac{75}{21} \approx 3{,}571$$

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

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

1,4

Correlación entre $x_1$ y $x_2$ $\Rightarrow$ $Cor(x_1, x_2) = \dfrac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1) * (x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2} \cdot \sqrt{\sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2}}$

Como la formula de covarianza tiene el mismo númerador que correlación se enfocará en el denominador

$\sqrt{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2} \approx \sqrt{48{,}96}$ → Es $\sqrt{48{,}96}$ debido al númerado dentro de raiz cuadrada de la desviación estandar

$\approx 6{,}997$

$\sqrt{\sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2} \approx \sqrt{143{,}93} \approx 11{,}997$

$Cor(x_1, x_2) = \dfrac{75}{6{,}997 \times 11{,}997} \approx 0{,}9$

1.5. Explica la relación entre covarianza y correlación.

**Respuesta:** La covarianza es de 3.571 y la correlación es de 0.9, se puede decir que al ser la correlación tan alta y positiva indican que ambas variables crecen en el mismo sentido, y a la para si una crece la otra también, la covarianza como está de dispersos los datos y con ayuda de la correlación establece que ambas variables tienen una misma dispersión de los datos.

1.6. Calcule el resultado del algoritmo K-means sobre este set de datos a mano como lo hicimos en excel. Vamos a crear 3 grupos, es decir, k=3 ( clusters).

| **x1** | **x2** | **x3** | **Random** |
|---|---|---|---|
| 4 | 4 | 28 | 0 |
| 2 | 3 | 24 | 0 |
| 2 | 4 | 30 | 0 |
| 3 | 5 | 32 | 1 |
| 1 | 3 | 18 | 1 |
| 3 | 6 | 41 | 2 |
| 3 | 6 | 44 | 1 |
| 0 | 1 | 5 | 2 |
| 1 | 3 | 18 | 2 |
| 0 | 0 | 1 | 2 |
| 5 | 9 | 62 | 1 |
| 1 | 2 | 17 | 1 |
| 2 | 3 | 24 | 2 |
| 1 | 3 | 19 | 0 |
| 3 | 6 | 42 | 2 |
| 4 | 8 | 56 | 1 |
| 4 | 8 | 56 | 0 |
| 3 | 6 | 44 | 1 |
| 5 | 9 | 64 | 1 |
| 1 | 2 | 17 | 2 |
| 1 | 2 | 17 | 2 |

Se organiza toda la tabla de acuerdo a la columna Random de manera decreciente

| **x1** | **x2** | **x3** | **Random** |
|---|---|---|---|
| 3 | 6 | 41 | 2 |
| 0 | 1 | 5 | 2 |
| 1 | 3 | 18 | 2 |
| 0 | 0 | 1 | 2 |
| 2 | 3 | 24 | 2 |
| 3 | 6 | 42 | 2 |
| 1 | 2 | 17 | 2 |
| 1 | 2 | 17 | 2 |
| 3 | 5 | 32 | 1 |
| 1 | 3 | 18 | 1 |
| 3 | 6 | 44 | 1 |

| | | | |
|---|---|---|---|
| 5 | 9 | 62 | 1 |
| 1 | 2 | 17 | 1 |
| 4 | 8 | 56 | 1 |
| 3 | 6 | 44 | 1 |
| 5 | 9 | 64 | 1 |
| 4 | 4 | 28 | 0 |
| 2 | 3 | 24 | 0 |
| 2 | 4 | 30 | 0 |
| 1 | 3 | 19 | 0 |
| 4 | 8 | 56 | 0 |

Después de organizar se determina centroides (Para lo centroides se realiza el promedio de acuerdo al grupo) y distancias (Se utiliza la fórmula euclidiana) y se determina el nuevo label de acuerdo a la distancia esta se comparan y la menor determina el grupo

| x1 | x2 | x3 | Random | centroide_1 | centroide_2 | centroide_3 | distancia_c1_2 | distancia_c2_1 | distancia_c3_0 | Nuevo label iteración 1 |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 6 | 41 | 2 | 1.375 | 2.875 | 20.625 | 20.67720665 | 1.131923142 | 9.74063653 | 1 |
| 0 | 1 | 5 | 2 | | | | 15.79705273 | 37.59030793 | 26.7447191 | 2 |
| 1 | 3 | 18 | 2 | | | | 2.654595073 | 24.40350897 | 13.56760848 | 2 |
| 0 | 0 | 1 | 2 | | | | 19.88207421 | 41.67770687 | 30.82661188 | 2 |
| 2 | 3 | 24 | 2 | | | | 3.434657916 | 18.40601125 | 7.555130707 | 2 |
| 3 | 6 | 42 | 2 | | | | 21.66326095 | 0.1767766953 | 10.72753467 | 1 |
| 1 | 2 | 17 | 2 | | | | 3.747916088 | 25.53000685 | 14.6860478 | 2 |
| 1 | 2 | 17 | 2 | | | | 3.747916088 | 25.53000685 | 14.6860478 | 2 |
| 3 | 5 | 32 | 1 | 3.125 | 6 | 42.125 | 11.68532734 | 10.17503071 | 0.938083152 | 0 |
| 1 | 3 | 18 | 1 | | | | 2.654595073 | 24.40350897 | 13.56760848 | 2 |
| 3 | 6 | 44 | 1 | | | | 23.63888481 | 1.879162047 | 12.70747811 | 1 |
| 5 | 9 | 62 | 1 | | | | 41.98269733 | 20.18740325 | 31.03675241 | 1 |
| 1 | 2 | 17 | 1 | | | | 3.747916088 | 25.53000685 | 14.6860478 | 2 |
| 4 | 8 | 56 | 1 | | | | 35.84057582 | 14.04568439 | 24.90140558 | 1 |
| 3 | 6 | 44 | 1 | | | | 23.63888481 | 1.879162047 | 12.70747811 | 1 |
| 5 | 9 | 64 | 1 | | | | 43.95505517 | 22.15922494 | 33.01030142 | 1 |
| 4 | 4 | 28 | 0 | 2.6 | 4.4 | 31.4 | 7.90865823 | 14.29269919 | 3.698648402 | 0 |
| 2 | 3 | 24 | 0 | | | | 3.434657916 | 18.40601125 | 7.555130707 | 2 |
| 2 | 4 | 30 | 0 | | | | 9.462921061 | 12.34022893 | 1.574801575 | 0 |
| 1 | 3 | 19 | 0 | | | | 1.67238602 | 23.41540625 | 12.58093796 | 2 |
| 4 | 8 | 56 | 0 | | | | 35.84057582 | 14.04568439 | 24.90140558 | 1 |

Ahora se cambia la etiqueta Random por el de la iteración 1 y se vuelve a organizar toda la tabla de mayor a menor

| x1 | x2 | x3 | Nuevo label interacción 1 |
|---|---|---|---|
| 0 | 1 | 5 | 2 |
| 1 | 3 | 18 | 2 |
| 0 | 0 | 1 | 2 |

| | | | |
|---|---|---|---|
| 2 | 3 | 24 | 2 |
| 1 | 2 | 17 | 2 |
| 1 | 2 | 17 | 2 |
| 1 | 3 | 18 | 2 |
| 1 | 2 | 17 | 2 |
| 2 | 3 | 24 | 2 |
| 1 | 3 | 19 | 2 |
| 3 | 6 | 41 | 1 |
| 3 | 6 | 42 | 1 |
| 3 | 6 | 44 | 1 |
| 5 | 9 | 62 | 1 |
| 4 | 8 | 56 | 1 |
| 3 | 6 | 44 | 1 |
| 5 | 9 | 64 | 1 |
| 4 | 8 | 56 | 1 |
| 3 | 5 | 32 | 0 |
| 4 | 4 | 28 | 0 |
| 2 | 4 | 30 | 0 |

Se calcula la iteración 2

| **x1** | **x2** | **x3** | **Nuevo label interacción 1** | **centroide_1** | **centroide_2** | **centroide_3** | **distancia_c1_2** | **distancia_c2_1** | **distancia_c3_0** | **Nuevo label iteración 2** |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 5 | 2 | 1 | 2.2 | 16 | 11.11035553 | 46.6973996 | 25.3990376 | 2 |
| 1 | 3 | 18 | 2 | | | | 2.154065923 | 33.5095602 | 12.23837317 | 2 |
| 0 | 0 | 1 | 2 | | | | 15.19341963 | 50.7852023 | 29.47503652 | 2 |
| 2 | 3 | 24 | 2 | | | | 8.10185164 | 27.51164526 | 6.227180564 | 0 |
| 1 | 2 | 17 | 2 | | | | 1.019803903 | 34.6358863 | 13.35830994 | 2 |
| 1 | 2 | 17 | 2 | | | | 1.019803903 | 34.6358863 | 13.35830994 | 2 |
| 1 | 3 | 18 | 2 | | | | 2.154065923 | 33.5095602 | 12.23837317 | 2 |
| 1 | 2 | 17 | 2 | | | | 1.019803903 | 34.6358863 | 13.35830994 | 2 |
| 2 | 3 | 24 | 2 | | | | 8.10185164 | 27.51164526 | 6.227180564 | 0 |
| 1 | 3 | 19 | 2 | | | | 3.104834939 | 32.52138719 | 11.25956384 | 2 |
| 3 | 6 | 41 | 1 | 3.75 | 7.25 | 51.125 | 25.36611914 | 10.22940003 | 11.12554618 | 1 |
| 3 | 6 | 42 | 1 | | | | 26.352222951 | 9.240704789 | 12.1151879 | 1 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 6 | 44 | 1 | | | | 28.32737192 | 7.272594104 | 14.09885732 | 1 |
| 5 | 9 | 62 | 1 | | | | 46.67161878 | 11.0856044 | 32.40027435 | 1 |
| 4 | 8 | 56 | 1 | | | | 40.529494543 | 4.938686566 | 26.27630957 | 1 |
| 3 | 6 | 44 | 1 | | | | 28.32737192 | 7.272594104 | 14.09885732 | 1 |
| 5 | 9 | 64 | 1 | | | | 48.64401299 | 13.053376 | 34.37699489 | 1 |
| 4 | 8 | 56 | 1 | | | | 40.529494543 | 4.938686566 | 26.27630957 | 1 |
| 3 | 5 | 32 | 0 | 3 | 4.333333333 | 30 | 16.36581804 | 19.27149774 | 2.108185107 | 0 |
| 4 | 4 | 28 | 0 | | | | 12.49959999 | 23.35359983 | 2.260776661 | 0 |
| 2 | 4 | 30 | 0 | | | | 14.15061836 | 21.44506062 | 1.054092553 | 0 |

Ahora se cambia la etiqueta iteración 1 por la etiqueta iteración 2 y se vuelve a organizar toda la tabla de mayor a menor

| **x1** | **x2** | **x3** | **Nuevo label iteración 2** |
|---|---|---|---|
| 0 | 1 | 5 | 2 |
| 1 | 3 | 18 | 2 |
| 0 | 0 | 1 | 2 |
| 1 | 2 | 17 | 2 |
| 1 | 2 | 17 | 2 |
| 1 | 3 | 18 | 2 |
| 1 | 2 | 17 | 2 |
| 1 | 3 | 19 | 2 |
| 3 | 6 | 41 | 1 |
| 3 | 6 | 42 | 1 |
| 3 | 6 | 44 | 1 |
| 5 | 9 | 62 | 1 |
| 4 | 8 | 56 | 1 |
| 3 | 6 | 44 | 1 |
| 5 | 9 | 64 | 1 |
| 4 | 8 | 56 | 1 |
| 2 | 3 | 24 | 0 |
| 2 | 3 | 24 | 0 |
| 3 | 5 | 32 | 0 |
| 4 | 4 | 28 | 0 |
| 2 | 4 | 30 | 0 |

Se calcula la iteración 3

| **x1** | **x2** | **x3** | **Nuevo label** | **centroide _1** | **centroide _2** | **centroide _3** | **distancia _c1_2** | **distancia _c2_1** | **distancia _c3_0** | **Nuevo label iteración 3** |
|---|---|---|---|---|---|---|---|---|---|---|

| | | | iteración 2 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 5 | 2 | 0.75 | 2 | 14 | 9.086390923 | 46.69732996 | 22.92073297 | 2 |
| 1 | 3 | 18 | 2 | | | | 4.13067791 | 33.5095602 | 9.765244493 | 2 |
| 0 | 0 | 1 | 2 | | | | 13.17431213 | 50.78524023 | 26.99555519 | 2 |
| 1 | 2 | 17 | 2 | | | | 3.010398645 | 34.63582863 | 10.87014259 | 2 |
| 1 | 2 | 17 | 2 | | | | 3.010398645 | 34.63582863 | 10.87014259 | 2 |
| 1 | 3 | 18 | 2 | | | | 4.13067791 | 33.5095602 | 9.765244493 | 2 |
| 1 | 2 | 17 | 2 | | | | 3.010398645 | 34.63582863 | 10.87014259 | 2 |
| 1 | 3 | 19 | 2 | | | | 5.105144464 | 32.52138719 | 8.784076502 | 2 |
| 3 | 6 | 41 | 1 | 3.75 | 7.25 | 51.125 | 27.38726894 | 10.22940003 | 13.58528616 | 1 |
| 3 | 6 | 42 | 1 | | | | 28.37362331 | 9.240704789 | 14.57257699 | 1 |
| 3 | 6 | 44 | 1 | | | | 30.34901152 | 7.272594104 | 16.55173707 | 1 |
| 5 | 9 | 62 | 1 | | | | 48.69355707 | 11.0856044 | 34.87348563 | 1 |
| 4 | 8 | 56 | 1 | | | | 42.55070505 | 4.938686566 | 28.74299915 | 1 |
| 3 | 6 | 44 | 1 | | | | 30.34901152 | 7.272594104 | 16.55173707 | 1 |
| 5 | 9 | 64 | 1 | | | | 50.66618695 | 13.053376 | 36.84779505 | 1 |
| 4 | 8 | 56 | 1 | | | | 42.55070505 | 4.938686566 | 28.74299915 | 1 |
| 2 | 3 | 24 | 0 | 2.6 | 3.8 | 27.6 | 10.12731455 | 27.51164526 | 3.736308338 | 0 |
| 2 | 3 | 24 | 0 | | | | 10.12731455 | 27.51164526 | 3.736308338 | 0 |
| 3 | 5 | 32 | 0 | | | | 18.38647601 | 19.27149774 | 4.578209257 | 0 |
| 4 | 4 | 28 | 0 | | | | 14.51077186 | 23.35359983 | 1.469693846 | 0 |
| 2 | 4 | 30 | 0 | | | | 16.17289399 | 21.44506062 | 2.481934729 | 0 |

Desde este punto no se realiza más iteraciones debido a que si se ve la iteración 2 ordenada de mayor a menor y la iteración 3, son complementamente iguales, es más si se reemplaza la iteración 2 con la 3 no se tiene que ordenar porque ya están ordenadas, y como se dijo anteriormente la iteración 2 ordenada es igual a la iteración 3 este ciclo se vuelve a repetir infinitamente.

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 X1, y X2 y crear un vector con una sola dimensión.

2.
PCA entre $x_1$ y $x_2$

$$X = \begin{array}{cc} x_1 & x_2 \end{array}$$
$$X = \begin{bmatrix} 4 & 4 \\ 2 & 3 \\ 2 & 4 \\ 3 & 5 \\ 1 & 3 \\ 3 & 6 \\ 3 & 6 \\ 0 & 1 \\ 1 & 3 \\ 0 & 0 \\ 5 & 9 \\ 1 & 2 \\ 2 & 3 \\ 1 & 3 \\ 3 & 6 \\ 4 & 8 \\ 4 & 8 \\ 3 & 6 \\ 5 & 9 \\ 1 & 2 \\ 1 & 2 \end{bmatrix}$$

1) Centrar y estandarizar

$\mu_1 = 2,33$ $\quad \sigma_1 = 1,527$

$\mu_2 = 4,428$ $\quad \sigma_2 = 2,618$

La desviación estandar y la media sale del punto 1.1

$$X_{estandarizada} = \begin{array}{cc} x_{1n} & x_{2n} \end{array}$$
$$X_{estandarizada} = \begin{bmatrix} 1,091 & -0,163 \\ -0,218 & -0,545 \\ -0,218 & -0,163 \\ 0,436 & 0,218 \\ -0,872 & -0,545 \\ 0,436 & 0,600 \\ 0,436 & 0,600 \\ -1,527 & -1,309 \\ -0,872 & -0,545 \\ -1,527 & -1,691 \\ 1,745 & 1,745 \\ -0,872 & -0,927 \\ -0,218 & -0,545 \\ -0,872 & -0,545 \\ 0,436 & 0,600 \\ 1,091 & 1,363 \\ 1,091 & 1,363 \\ 0,436 & 0,600 \\ 1,745 & 1,745 \\ -0,872 & -0,927 \\ -0,872 & -0,927 \end{bmatrix}$$

$\mu_{1n} = 0$ $\quad \sigma_{1n} = 1$

$\mu_{2n} = 0$ $\quad \sigma_{2n} = 1$

2.1. Cual es la matriz de covarianza

2) Matriz de covarianza

$$\Sigma = \begin{bmatrix} \sigma^2_{x_{1n}} & Cov(x_{1n}, x_{2n}) \\ Cov(x_{2n}, x_{1n}) & \sigma^2_{x_{2n}} \end{bmatrix} = \begin{bmatrix} 1 & 0,892 \\ 0,892 & 1 \end{bmatrix}$$

$Cov(x_{1n}, x_{2n}) = 0,892$

$Cov(x_{2n}, x_{1n}) = 0,892$

2.2. Cuales son los eigenvalues

3) Eigen values $\rightarrow \det(\Sigma - \lambda I) = 0$

$$\det\left(\begin{bmatrix} 1 & 0{,}892 \\ 0{,}892 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = \det\begin{pmatrix} 1-\lambda & 0{,}892 \\ 0{,}892 & 1-\lambda \end{pmatrix} \Rightarrow (1-\lambda)^2 - 0{,}892^2 \Rightarrow \dots$$

$$\dots \Rightarrow 1 - 2\lambda + \lambda^2 - 0{,}797 = 0 \Rightarrow 0{,}203 - 2\lambda + \lambda^2$$

$$\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \Rightarrow \lambda = \frac{+2 \pm \sqrt{4 - 4 \times 0{,}203 \times 1}}{2 \times 1} \Rightarrow \lambda = \frac{2 \pm \sqrt{4 - 0{,}812}}{2} \Rightarrow \lambda = \frac{2 \pm 1{,}785}{2}$$

$$\boxed{\begin{aligned} \lambda_1 &= 1{,}892 \\ \lambda_2 &= 0{,}107 \end{aligned}}$$ Eigen values.

2.3. Cuál es la varianza explicada por el eigenvalue.

$$\boxed{\begin{aligned} \lambda_1 &= 1{,}892 \\ \lambda_2 &= 0{,}107 \end{aligned}}$$ Eigen values.

Para calcular las varianzas de los eigen Values, se determina la suma total de los eigen values

$$\lambda_T = \lambda_1 + \lambda_2 \Rightarrow \lambda_T = 1{,}892 + 0{,}107 \Rightarrow \lambda_T = 1{,}999$$

Varianza explicada para $\lambda_1 = (1{,}892/1{,}999) * 100\% = 94{,}65\%$

$(1{,}892/2) * 100\% = 94{,}64\%$

$\lambda_2 = (0{,}107/1{,}999) * 100\% = 5{,}35\%$

$(0{,}107/2) * 100\% = 5{,}36\%$

2.4. Cual es el valor del eigenvector

Eigen vector

Se tomará el eigenvalue $\lambda_1$ debido a que es el mayor, tomando un 94,64% de los datos

$$\begin{bmatrix} 1 & 0{,}892 \\ 0{,}892 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda_1 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

1. $x_1 + 0{,}892x_2 = 1{,}892x_1 \Rightarrow 0{,}892x_2 = 0{,}892x_1$
2. $0{,}892x_1 + x_2 = 1{,}892x_2 \Rightarrow 0{,}892x_1 = 0{,}892x_2$

$\Downarrow$

$x_1 = x_2$

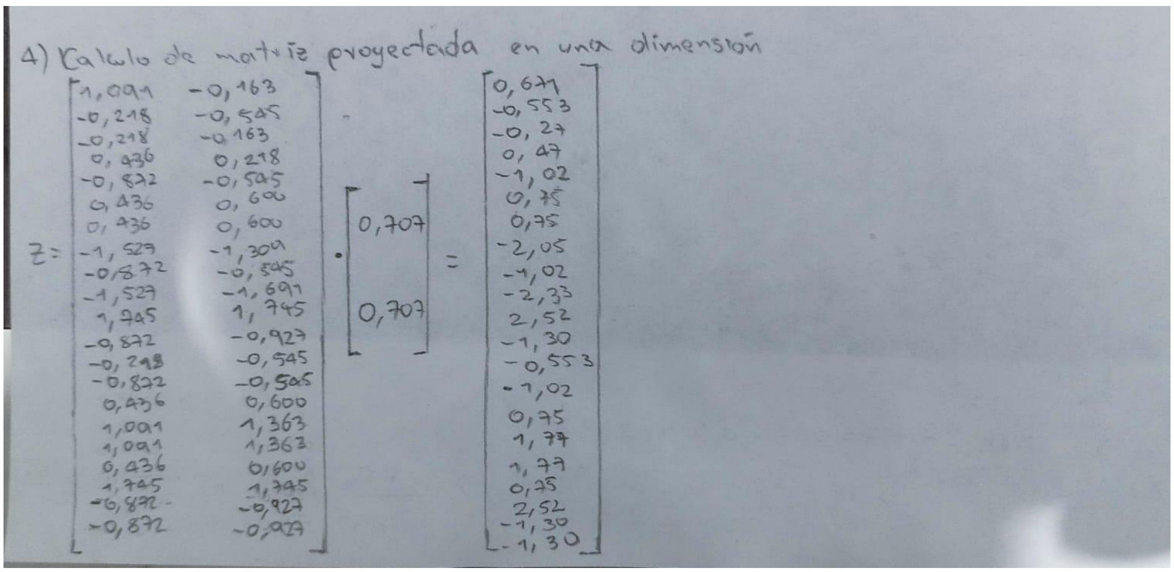
$x_2 = x_1$

$$V_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \text{se verifica } \begin{bmatrix} 1 & 0{,}892 \\ 0{,}892 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = 1{,}892\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1{,}892 \\ 1{,}892 \end{bmatrix} = 1{,}892\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

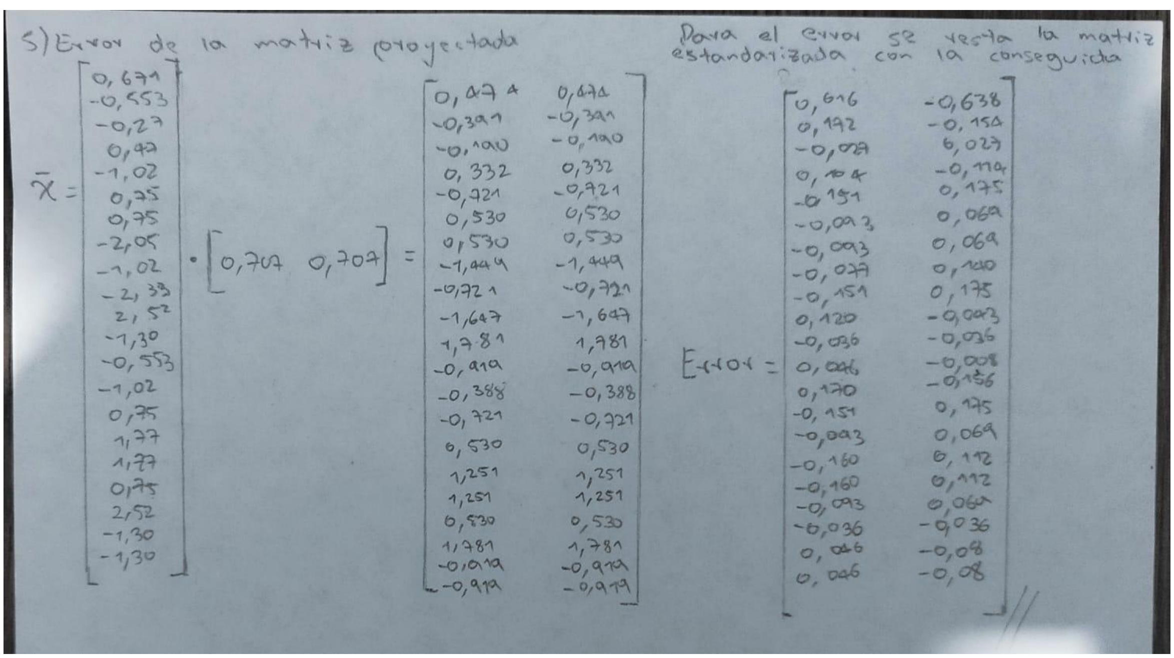
$$\|V_1\| = \sqrt{x_1^2 + x_2^2} = 1 \Rightarrow \sqrt{2x_1} \Rightarrow \sqrt{\tfrac{1}{2}} = x_1 \Rightarrow x_1 = 0{,}707$$

$$V_1 = [0{,}707 \quad 0{,}707]$$

2.5. Cuál es la matriz proyectada.




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




3. Utilizando el dataset del [proyecto](proyecto) data/CARS.csv crear: **Utilizar la librería de plotly**.

3.1. Distribución de cada variables:

3.1.1. Para las variables categóricas un gráfico de barras. Categoría numero de observaciones.

**Se realizan las gráficas en el notebook en la celda 9.**

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

**Se realizan las gráficas en el notebook en la celda 10.**
**Se realiza el test de distribución normal para cada columna en la celda 11 del notebook.**
**Se listan los modelos de carros que están más lejos de 5 desviaciones estándar en la celda 12 del notebook.**

3.2. Gráfico de la relación de cada variable con respecto a MPG_City:

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

**Se realizan los gráficos (boxplot) con respecto MPG_City y sus respectivos análisis en la celda 14 del notebook.**

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

**Se realizan los gráficos (scatter) con respecto MPG_City y sus respectivos análisis en la celda 15 del notebook.**

3.3. Matriz de correlación.

3.3.1. Cree la matriz de correlación, cuales son las variables más importantes para explicar la variabilidad de MPG_City. Explique por qué el coeficiente es negativo o positivo.

**Se realiza la matriz de correlación y se analizan la variables más importantes para MPG_City en la celda 16 del notebook**

Como se ve en la matriz de correlación se identifica que la variable con mayor relación es MPG_Highway y tiene sentido debido a que la gráfica los datos crecen de manera similar, esto también significa que los vehículos que tienen un buen rendimiento en consumo de combustible en la ciudad tienden a tener también un buen rendimiento en consumo de combustible en autopistas o en carreteras.

Vemos también que MPG_Highway tiene una correlación fuertes con EngineSize(-0.717302), Horsepower(-0.647195), Cylinders(-0.676100) y Weight(-0.790989) al ser MPG_Highway muy similar en comportamiento con MPG_City ambos tienen correlaciones similares obviamente no tienen la misma proporción, pero se pueden sacar las mismas conclusiones para ambas.

Otra correlación bastante fuerte es MPG_City con EngineSize, es una correlación negativa que indica que entre menor sea el tamaño del motor, más combustible ahorra.

Otra correlación bastante fuerte es MPG_City con Cylinders, es una correlación negativa que indica que entre menor sea la cantidad de cilindros, menos combustible gasta.

Otra correlación bastante fuerte es MPG_City con Horsepower, es una correlación negativa que indica que si la potencia del motor aumenta, el consumo de combustible en la ciudad tiende a disminuir, es bastante lógico, ya que entre más potente es el motor generalmente consumen más combustible.

Otra correlación bastante fuerte es MPG_City con Weight, es una correlación negativa que indica que si un auto pesa mucho, su consumo de combustible se ve bastante afectado de manera negativa, es decir consume mucho más debido a su peso.

En si la correlación con mayor valor es MPG_Highway debido a que son muy similares, pero también se tiene EngineSize, Cylinders, Horsepower y Weight.

3.3.2. 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?

**Se generan las dummy variables en la celda 17 del notebook.**

**Se realiza la matriz de correlación y se analiza la variable categórica más importante para MPG_City en la celda 18 del notebook.**

Se observa que la variable categórica con mayor relación a MPG_City es Type_Hybrid con una correlación del 0.561053, es decir que que los autos híbridos tienden a consumir menos gasolina, también se puede observar que está afirmación es correcta con respecto a la gráfica realizada anteriormente.

3.3.3. Cree la matriz de correlación nuevamente removiendo todas los modelos de carro que fueron catalogados como un outlier. (Puede utilizar .query(‘Model in["MDX","TSX 4dr"]’ ). Existe alguna variación en la correlación.

**Se excluyen los outliers de la data de carros y se realiza la matriz de correlación en la celda 19 del notebook.**

Se logra observar que los datos se alteraron, si volvemos a comparar MPG_Highway que tiene una correlación de 0.944083 con MPG_City, antes de eliminar los outliers se observa 0.941021, no es una alteración tan grande, pero para otros datos si puede llegar a modificarlos bastante como a Weight que pasó de -0.737966 a -0.802215, es decir que los outliers estaban actuando como un disminuidor de fuerzas.