

Clasificación no supervisada

Los estudios de clasificación no supervisada buscan definir, para una determinada característica (suceso elemental), un conjunto de grupos de observaciones (sucesos) con valores cercanos. Dichos grupos, denominados clusters, permitirán que, a partir de los valores de las características (sucesos elementales) que configuran un suceso, se pueda asignar el mismo como perteneciente a uno de ellos.

Cada cluster queda definido a través de sus parámetros por lo que es imprescindible disponer de una muestra de sucesos para poder determinarlos. Dicha muestra incluirá los valores del suceso elemental o característica para el cuál se está buscando la clasificación.

La clasificación no supervisada o clustering se denomina no supervisada porque **los valores que definen los diferentes clusters, o clases, se definen durante el mismo proceso de clasificación**, mientras que en el caso de la clasificación supervisada las clases han sido previamente definidas.

Las técnicas de clusterización más conocidas y utilizadas son:

- K-medias (K-means, a partir de aquí vamos a utilizar el nombre en inglés)
- Clustering jerárquico aglomerativo
- DBSCAN



Clasificación no supervisada

Ejercicio: Se tiene un conjunto de calificaciones académicas, pertenecientes a dos grupos de alumnos, mañana y tarde, formadas por dos notas: Teoría y Laboratorio. Las calificaciones tanto de Teoría como de Laboratorio tendrán valores de 0 a 5, donde 5 será la mayor calificación posible y 0 la menor. Establecer las características, los sucesos elementales, definir los posibles valores de los clusters y explicar qué haría la clusterización.

Las características son las calificaciones de teoría y laboratorio y el grupo al que pertenece el alumno.

Los sucesos elementales son: {t0, t1, t2, t3, t4, t5, l0, l1, l2, l3, l4, l5, gm, gt}

La clusterización, en función de los valores de las calificaciones de teoría y de laboratorio permitirá saber si el alumno pertenece al grupo de mañana o de tarde.

Clasificación no supervisada

La obtención de los clusters a partir del algoritmo **K-means** sigue un proceso de 2 a n pasos:

A. El paso A se puede separar a su vez en 3 subpasos:

1. **Selección del número K de clusters**, en los que se van a agrupar los datos y los centroides que representarán los mismos. Se eligen de forma arbitraria por parte del usuario. Los centroides serán los puntos medios del grupo de puntos (sucos) que componen el cluster.
2. **Cálculo de la distancia euclíadiana** de cada punto a cada uno de los centroides definidos.
3. **Asignación de los puntos (sucos) a los clusters**. Se asigna cada punto a cada cluster teniendo en cuenta las distancias a cada centroide. Se asigna el punto al centroide que esté más cercano

Clasificación no supervisada

Ejercicio: A partir de la muestra de calificaciones de Teoría y Laboratorio: 1. {4, 4}; 2. {3, 5}; 3. {1, 2}; 4. {5, 5}; 5. {0, 1}; 6. {2, 2}; 7. {4, 5}; 8. {2, 1}. Realizar el paso A.1 del algoritmo K-means:

El primer paso es la **selección del número K de clusters**, en los que se van a agrupar los datos y los centroides que representarán los mismos. Se eligen de forma arbitraria por parte del usuario. Los centroides serán los puntos medios del grupo de puntos (sucos) que componen el cluster.

Al tratarse de ocho puntos lo más lógico es pensar que hay un número reducido de clusters por lo que se van a tomar 2 clusters.

Los centroides, teniendo en cuenta los valores de la muestra las características, se toman arbitrariamente como:

C1 (0,1) y C2 (2,2)

Clasificación no supervisada

Ejercicio: A partir de la muestra 1. {4, 4}; 2. {3, 5}; 3. {1, 2}; 4. {5, 5}; 5. {0, 1}; 6. {2, 2}; 7. {4, 5}; 8. {2, 1}. Realizar el paso A.2 del algoritmo K-means:

El segundo paso es el cálculo de las distancias euclídeas de los puntos a los centroides:

$$\text{Punto 1, (4,4): } d_{1C_1} = \sqrt{\sum_{i=1}^2 (p_i - q_i)^2} = \sqrt{(0-4)^2 + (1-4)^2} = 5$$

$$d_{1C_2} = \sqrt{(2-4)^2 + (2-4)^2} = 2,83$$

$$\text{Punto 2, (3,5): } d_{2C_1} = \sqrt{(0-3)^2 + (1-5)^2} = 5 \quad d_{2C_2} = \sqrt{(2-3)^2 + (2-5)^2} = 3,16$$

$$\text{Punto 3, (1,2): } d_{3C_1} = \sqrt{(0-1)^2 + (1-2)^2} = 1,41 \quad d_{3C_2} = \sqrt{(2-1)^2 + (2-2)^2} = 1$$

$$\text{Punto 4, (5,5): } d_{4C_1} = \sqrt{(0-5)^2 + (1-5)^2} = 6,4 \quad d_{4C_2} = \sqrt{(2-5)^2 + (2-5)^2} = 4,24$$

$$\text{Punto 5, (0,1): } d_{5C_1} = \sqrt{(0-0)^2 + (1-1)^2} = 0 \quad d_{5C_2} = \sqrt{(2-0)^2 + (2-1)^2} = 2,23$$

$$\text{Punto 6, (2,2): } d_{6C_1} = \sqrt{(0-2)^2 + (1-2)^2} = 2,23 \quad d_{6C_2} = \sqrt{(2-2)^2 + (2-2)^2} = 0$$

$$\text{Punto 7, (4,5): } d_{7C_1} = \sqrt{(0-4)^2 + (1-5)^2} = 5,66 \quad d_{7C_2} = \sqrt{(2-4)^2 + (2-5)^2} = 3,61$$

$$\text{Punto 8, (2,1): } d_{8C_1} = \sqrt{(0-2)^2 + (1-1)^2} = 2 \quad d_{8C_2} = \sqrt{(2-2)^2 + (2-1)^2} = 1$$

Clasificación no supervisada

Ejercicio: A partir de la muestra 1. {4, 4}; 2. {3, 5}; 3. {1, 2}; 4. {5, 5}; 5. {0, 1}; 6. {2, 2}; 7. {4, 5}; 8. {2, 1}. Realizar el paso A.3 del algoritmo K-means:

El tercer paso es la asignación de los puntos (sucesos) a los clusters. Se asigna cada punto a cada cluster teniendo en cuenta las distancias a cada centroide. Se asigna el punto al centroide que esté más cercano

Realizamos una matriz con las distancias de cada punto a cada centroide.

Punto	1	2	3	4	5	6	7	8
c1	5	5	1.41	6.4	0	2.23	5.66	2
c2	2.83	3.16	1	4.24	2.33	0	3.61	1

Realizamos una matriz con la asignación de cada punto a cada cluster (más cercano, 1)

Punto	1	2	3	4	5	6	7	8
c1	0	0	0	0	1	0	0	0
c2	1	1	1	1	0	1	1	1

Las definición de un conjunto de clusters a partir de la técnica K-means sigue un proceso de 2 a n pasos:

- B. El paso B se puede separar a su vez en 3 subpasos. Los dos últimos son como los del Paso A y el primero es
 1. Recálculo de los centroides, Ahora ya no se hace de forma arbitraria sino que sobre lo que se ha obtenido en la primera iteración (para la segunda iteración) o de lo obtenido en la iteración anterior (para cualquier otra) se recalculan los centroides dando a los mismos el valor medio de los puntos asignados a dicho centroide.

Clasificación no supervisada

Ejercicio: Realizar el paso B.1 del algoritmo K-means para la búsqueda de clusters del ejercicio 2.

El primer subpaso del paso B, que es el distinto, es el recálculo de los centroides:

Teniendo en cuenta la matriz de asignaciones vemos que solo un punto, el quinto (0,1) ha sido asignado al primer cluster, cuyo centroide coincide con el punto $C_1=(0,1)$ por lo que el primer cluster, formado únicamente por ese punto, no varía su centroide, que coincide con el punto.

En el segundo cluster están el resto de los puntos por lo que hay que recalcular el centroide de este cluster haciendo la media de los puntos que en él se encuentran, y que es igual a:

$$c_2 = \left(\frac{4 + 3 + 1 + 5 + 2 + 4 + 2}{7}, \frac{4 + 5 + 2 + 5 + 2 + 5 + 1}{7} \right) = (3, 3.43)$$

Ejercicio: Realizar el paso B.2 del algoritmo K-means para la búsqueda de clusters del ejercicio 2.

El segundo paso es el cálculo de las distancias euclídeas de los puntos a los centroides. La primera distancia no ha cambiado porque el centroide es el mismo.

$$P1, (4,4): d_{1C_2} = \sqrt{(3 - 4)^2 + (3.43 - 4)^2} = 1.15$$

$$P2, (3,5): d_{2C_2} = \sqrt{(3 - 3)^2 + (3.43 - 5)^2} = 1.57$$

$$P3, (1,2): d_{3C_2} = \sqrt{(3 - 1)^2 + (3.43 - 2)^2} = 2.46$$

$$P4, (5,5): d_{4C_2} = \sqrt{(3 - 5)^2 + (3.43 - 5)^2} = 2.54$$

$$P5, (0,1): d_{5C_2} = \sqrt{(3 - 0)^2 + (3.43 - 1)^2} = 3.86$$

$$P6, (2,2): d_{6C_2} = \sqrt{(3 - 2)^2 + (3.43 - 2)^2} = 1.74$$

$$P7, (4,5): d_{7C_2} = \sqrt{(3 - 4)^2 + (3.43 - 5)^2} = 1.86$$

$$P8, (2,1): d_{8C_2} = \sqrt{(3 - 2)^2 + (3.43 - 1)^2} = 2.63$$

Ejercicio: Realizar el paso B.3 del algoritmo K-means:

El tercer paso es la asignación de los puntos (sucesos) a los clusters. Se asigna cada punto a cada cluster teniendo en cuenta las distancias a cada centroide. Se asigna el punto al centroide que esté más cercano

Realizamos una matriz con las distancias de cada punto a cada centroide.

Punto	1	2	3	4	5	6	7	8
c1	5	5	1.41	6.4	0	2.23	5.66	2
c2	1.15	1.57	2.46	2.54	3.86	1.74	1.86	2.63

Realizamos una matriz con la asignación de cada punto a cada cluster (más cercano, 1)

Punto	1	2	3	4	5	6	7	8
c1	0	0	1	0	1	0	0	1
c2	1	1	0	1	0	1	1	0

eu.boculatv.com is sharing your screen Stop Sharing Hide

Clasificación no supervisada

Ejercicio: Si la clusterización no hubiera concluido con la realización del paso B.3 realizar los pasos del algoritmo K-means que sean necesarios para completarla. Indicar cuántos pasos son.

El primer subpaso de las iteraciones es el recálculo de los centroides:

En esta iteración hay que calcular los dos centroides, el primero es:

$$c_1 = \left(\frac{1 + 0 + 2}{3}, \frac{2 + 1 + 1}{3} \right) = (1, 1.33)$$

En el segundo centroide recalculado es:

$$c_2 = \left(\frac{4 + 3 + 5 + 2 + 4}{5}, \frac{4 + 5 + 5 + 2 + 5}{5} \right) = (3.6, 4.2)$$

A continuación calculamos de las distancias euclídeas de los puntos a los centroides:

$$P1, (4,4): d_{1C_1} = \sqrt{(1-4)^2 + (1.3-4)^2} = 4.03 \quad d_{1C_2} = \sqrt{(3.6-4)^2 + (4.2-4)^2} = 0.45$$

$$P2, (3,5): d_{2C_1} = \sqrt{(1-3)^2 + (1.3-5)^2} = 4.2 \quad d_{2C_2} = \sqrt{(3.6-3)^2 + (4.2-5)^2} = 1$$

$$P3, (1,2): d_{3C_1} = \sqrt{(1-1)^2 + (1.3-2)^2} = 0.7 \quad d_{3C_2} = \sqrt{(3.6-1)^2 + (4.2-2)^2} = 3.41$$

$$P4, (5,5): d_{4C_1} = \sqrt{(1-5)^2 + (1.3-5)^2} = 5.44 \quad d_{4C_2} = \sqrt{(3.6-5)^2 + (4.2-5)^2} = 1.61$$

$$P5, (0,1): d_{5C_1} = \sqrt{(1-0)^2 + (1.3-1)^2} = 1.04 \quad d_{5C_2} = \sqrt{(3.6-0)^2 + (4.2-1)^2} = 4.82$$

$$P6, (2,2): d_{6C_1} = \sqrt{(1-2)^2 + (1.3-2)^2} = 1.2 \quad d_{6C_2} = \sqrt{(3.6-2)^2 + (4.2-2)^2} = 2.72$$

$$P7, (4,5): d_{7C_1} = \sqrt{(1-4)^2 + (1.3-5)^2} = 4.74 \quad d_{7C_2} = \sqrt{(3.6-4)^2 + (4.2-5)^2} = 0.89$$

$$P8, (2,1): d_{8C_1} = \sqrt{(1-2)^2 + (1.3-1)^2} = 1.05 \quad d_{8C_2} = \sqrt{(3.6-2)^2 + (4.2-1)^2} = 3.58$$

Asignamos los puntos al nuevo centroide que esté más cercano

Realizamos una matriz con las distancias de cada punto a cada centroide.

Punto	1	2	3	4	5	6	7	8
c1	4.02	4.18	0.67	5.43	1.05	1.2	4.74	1.05
c2	0.45	1	3.41	1.61	4.82	2.72	0.89	3.58

Realizamos una matriz con la asignación de cada punto a cada cluster (más cercano, 1)

Punto	1	2	3	4	5	6	7	8
c1	0	0	1	0	1	1	0	1
c2	1	1	0	1	0	0	1	0

Ejercicio: Si la clusterización no hubiera concluido con la realización del paso B.3 realizar los pasos del algoritmo K-means que sean necesarios para completarla. Indicar cuántos pasos son.

Como ha habido un cambio de un punto, tenemos que volver a recalcular los centroides, será el cuarto subpaso de las iteraciones :

En esta segunda iteración hay que calcular los dos centroides, el primero es:

$$c_1 = \left(\frac{1+0+2+2}{4}, \frac{2+1+2+1}{4} \right) = (1.25, 1.5)$$

En el segundo centroide recalculado es:

$$c_2 = \left(\frac{4+3+5+4}{4}, \frac{4+5+5+5}{4} \right) = (4, 4.75)$$

El segundo paso de la segunda iteración (o quinto paso) es el cálculo de las distancias euclídeas de los puntos a los centroides:

$$P1, (4,4): d_{1C_1} = \sqrt{(1.25 - 4)^2 + (1.5 - 4)^2} = 3.72 \quad d_{1C_2} = \sqrt{(4 - 4)^2 + (4.75 - 4)^2} = 0.75$$

$$P2, (3,5): d_{2C_1} = \sqrt{(1.25 - 3)^2 + (1.5 - 5)^2} = 3.91 \quad d_{2C_2} = \sqrt{(4 - 3)^2 + (4.75 - 5)^2} = 1.03$$

$$P3, (1,2): d_{3C_1} = \sqrt{(1.25 - 1)^2 + (1.5 - 2)^2} = 0.56 \quad d_{3C_2} = \sqrt{(4 - 1)^2 + (4.75 - 2)^2} = 4.07$$

$$P4, (5,5): d_{4C_1} = \sqrt{(1.25 - 5)^2 + (1.5 - 5)^2} = 5.13 \quad d_{4C_2} = \sqrt{(4 - 5)^2 + (4.75 - 5)^2} = 1.03$$

$$P5, (0,1): d_{5C_1} = \sqrt{(1.25 - 0)^2 + (1.5 - 1)^2} = 1.34 \quad d_{5C_2} = \sqrt{(4 - 0)^2 + (4.75 - 1)^2} = 5.48$$

$$P6, (2,2): d_{6C_1} = \sqrt{(1.25 - 2)^2 + (1.5 - 2)^2} = 0.9 \quad d_{6C_2} = \sqrt{(4 - 2)^2 + (4.75 - 2)^2} = 3.4$$

$$P7, (4,5): d_{7C_1} = \sqrt{(1.25 - 4)^2 + (1.5 - 5)^2} = 4.45 \quad d_{7C_2} = \sqrt{(4 - 4)^2 + (4.75 - 5)^2} = 0.25$$

$$P8, (2,1): d_{8C_1} = \sqrt{(1.25 - 2)^2 + (1.5 - 1)^2} = 0.9 \quad d_{8C_2} = \sqrt{(4 - 2)^2 + (4.75 - 1)^2} = 4.25$$



En el tercer paso (sexto) asignamos los puntos al nuevo centroide que esté más cercano

Realizamos la matriz con las distancias de cada punto a cada centroide.

Punto	1	2	3	4	5	6	7	8
c1	3.72	3.91	0.56	5.13	1.35	0.9	4.45	0.9
c2	0.75	1.03	4.07	1.03	5.48	3.4	0.25	4.25

Y realizamos la matriz con la asignación de cada punto a cada cluster (más cercano, 1)

Punto	1	2	3	4	5	6	7	8
c1	0	0	1	0	1	1	0	1
c2	1	1	0	1	0	0	1	0

Hemos comprobado que la matriz de asignaciones no ha cambiado,

Punto	1	2	3	4	5	6	7	8
c1	0	0	1	0	1	1	0	1
c2	1	1	0	1	0	0	1	0

por lo que ya hemos calculado los clusters en los que se pueden agrupar los datos.

El primer cluster está formado por los sucesos 3, 5, 6 y 8 y centrado en el punto

C1 (1.25, 1.5)

El segundo cluster está formado por los sucesos 1, 2, 4 y 7 y centrado en el punto C2 (4, 4.75)

Para ver si los grupos de Mañana y Tarde tienen diferencias en el aprendizaje habría que ver si los clusters obtenidos coinciden con los alumnos de cada grupo. 

Las definición de un conjunto de clusters a partir de la técnica de **Clusterización Jerárquica Aglomerativa** sigue un proceso de 2 a n pasos, que se repetirá hasta que quede un solo cluster:

- A. **Paso A: Obtener la matriz de distancias euclidianas entre clusters.** En este paso se calculará la matriz de distancias de cuyos valores serán las distancias de cada cluster al resto de los clusters.
- B. **Paso B: Unir los dos clusters más próximos.** En este paso se ordenarán las distancias obtenidas y se generará un nuevo cluster uniendo los dos clusters más próximos.

¿Qué entendemos por proximidad entre clusters? 

Según lo que entendamos por proximidad entre clusters tendremos un algoritmo de clusterización jerárquica aglomerativa distinto.

En la primera iteración cada punto individual será considerado como un cluster.

Ejercicio: A partir de la muestra de datos 1. {0.89, 2.94}; 2. {4.36, 5.21}; 3. {3.75, 1.12}; 4. {6.25, 3.14}; 5. {4.1, 1.8}; 6. {3.9, 4.27}

Realizar el paso A del algoritmo de clusterización jerárquica aglomerativa:

El primer paso de cada iteración es el cálculo de la **matriz de distancias euclídeas entre todos los clusters**, que en el caso de la primera iteración es el cálculo de las distancias entre todos los puntos, porque cada punto es un cluster.

Hay que darse cuenta que la distancia 1 a 2 es la misma que la 2 a 1, así que lo que tenemos son combinaciones de 6 elementos, puntos, tomados de 2 en 2.

$$C_6^2 = \frac{6!}{2!(6-2)!} = \frac{6.5}{2} = 15$$

A partir de la muestra...

1. {0.89, 2.94}; 2. {4.36, 5.21}; 3. {3.75, 1.12}; 4. {6.25, 3.14}; 5. {4.1, 1.8}; 6. {3.9, 4.27}

Realizar el paso A del algoritmo de clusterización jerárquica aglomerativa:

$$d_{12} = \sqrt{(0.89 - 4.36)^2 + (2.94 - 5.21)^2} = 4.15 \quad d_{23} = \sqrt{(4.36 - 3.75)^2 + (5.21 - 1.12)^2} = 4.13$$

$$d_{13} = \sqrt{(0.89 - 3.75)^2 + (2.94 - 1.12)^2} = 3.39 \quad d_{24} = \sqrt{(4.36 - 6.25)^2 + (5.21 - 3.14)^2} = 2.8$$

$$d_{14} = \sqrt{(0.89 - 6.25)^2 + (2.94 - 3.14)^2} = 5.36 \quad d_{25} = \sqrt{(4.36 - 4.1)^2 + (5.21 - 1.8)^2} = 3.42$$

$$d_{15} = \sqrt{(0.89 - 4.1)^2 + (2.94 - 1.8)^2} = 3.41 \quad d_{26} = \sqrt{(4.36 - 3.9)^2 + (5.21 - 4.27)^2} = 1.05$$

$$d_{16} = \sqrt{(0.89 - 3.9)^2 + (2.94 - 4.27)^2} = 3.29 \quad d_{45} = \sqrt{(6.25 - 4.1)^2 + (3.14 - 1.8)^2} = 2.53$$

$$d_{34} = \sqrt{(3.75 - 6.25)^2 + (1.12 - 3.14)^2} = 3.21 \quad d_{46} = \sqrt{(6.25 - 3.9)^2 + (3.14 - 4.27)^2} = 2.61$$

$$d_{35} = \sqrt{(3.75 - 4.1)^2 + (1.12 - 1.8)^2} = 0.76$$

$$d_{36} = \sqrt{(3.75 - 3.9)^2 + (1.12 - 4.27)^2} = 3.15$$



$$d_{56} = \sqrt{(4.1 - 3.9)^2 + (1.8 - 4.27)^2} = 2.48$$

A partir de la muestra...

1. {0.89, 2.94}; 2. {4.36, 5.21}; 3. {3.75, 1.12}; 4. {6.25, 3.14}; 5. {4.1, 1.8}; 6. {3.9, 4.27}

Realizar el paso A del algoritmo de clusterización jerárquica aglomerativa:

Calculamos la matriz de distancias:

Punto	1	2	3	4	5	6
1	0					
2	4.15	0				
3	3.39	4.13	0			
4	5.36	2.8	3.21	0		
5	3.41	3.42	0.76	2.53	0	
6	3.29	1.05	3.15	2.61	2.48	0

Tipos de algoritmos de Clasificación Jerárquica Aglomerativa según la definición de proximidad entre Clusters:

- A. **MIN.** Define la proximidad entre dos Clusters como la distancia que haya entre los dos puntos más cercanos de los dos clusters. Produce clusters contiguos, en los que cada punto está más cercano al menos a un punto en su cluster que a cualquier otro punto en otro cluster. Se denomina también **Single Link**.
- B. **MAX.** Define la proximidad entre dos Clusters como la distancia que haya entre los dos puntos más lejanos de los dos clusters. Se denomina también **Complete Link**.
- C. **Group Average (Media del Grupo).** Define la proximidad entre dos Clusters como la media de distancias entre todas las parejas que se puedan formar con puntos de los dos clusters.

$$\text{proximidad} (C_i, C_j) = \frac{\sum_{j=1}^m \text{proximidad} (x_i, y_j)}{m \cdot n}$$

A partir de la muestra... 1. {2.89, 2.94}; 2. {4.36, 5.21}; 3. {3.75, 1.12}; 4. {5.25, 3.14}; 5. {4.1, 1.8}; 6. {3.9, 4.27}. Realizar el paso B del algoritmo de clusterización jerárquica aglomerativa utilizando el algoritmo con la definición de proximidad MIN:

Primera Iteración: Si tomamos la matriz de distancias entre los clusters, teniendo en cuenta que en la primera iteración cada punto constituye un cluster.

Punto	1	2	3	4	5	6
1	0					
2	4.15	0				
3	3.39	4.13	0			
4	5.36	2.8	3.21	0		
5	3.41	3.42	0.76	2.53	0	
6	3.29	1.05	3.15	2.61	2.48	0

Los dos clusteres más próximos son 3 y 5. Por lo que el primer cluster, C1, es el formado por esos dos puntos.

Como no tenemos un solo cluster pasamos a la segunda iteración.

Realizar la segunda iteración del algoritmo de clusterización jerárquica aglomerativa:

Realizamos el paso A. Cálculo de la matriz de distancias entre clusters.

Los datos son ahora:

1. {0.89, 2.94}; 2. {4.36, 5.21}; 4. {6.25, 3.14}; 6. {3.9, 4.27} y C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}}

Punto	1	2	4	6	C1-3	C1-5
1	0				↳	
2	4.15	0				
3	3.39	4.13			0	
4	5.36	2.8	0		3.21	
5	3.41	3.42	2.53		0	0
6	3.29	1.05	2.61	0	3.15	2.48

La distancia entre clusters, es ahora entre los cuatro puntos 1, 2, 4 y 6, y el cluster 1 formado por los puntos 3 y 5, en la iteración anterior, y en consecuencia la distancia entre los puntos 3 y 5 ahora es 0 porque están en el mismo cluster.

Realizar el paso B del algoritmo de clusterización jerárquica aglomerativa utilizando el algoritmo con la definición de proximidad MIN:

Los datos son ahora:

1. {0.89, 2.94}; 2. {4.36, 5.21}; 4. {6.25, 3.14}; 6. {3.9, 4.27} y C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}}

Punto	1	2	4	6	C1-3	C1-5
1	0					
2	4.15	0				
3	3.39	4.13			0	
4	5.36	2.8	0		3.21	
5	3.41	3.42	2.53		0	0
6	3.29	1.05	2.61	0	3.15	2.48

Los dos clusteres más próximos son 2 y 6. Por lo que el segundo cluster, C2, es el formado por esos dos puntos.

Realizar la tercera iteración del algoritmo de clusterización jerárquica aglomerativa:

Realizamos el paso A. Cálculo de la matriz de distancias entre clusters.

Los datos son ahora:

1. {0.89, 2.94}; 4. {6.25, 3.14}, C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}}, y C2 {2. {4.36, 5.21}; 6. {3.9, 4.27}}

Punto	1	4	C1-3	C1-5	C2-2	C2-6
1	0					
2	4.15				0	
3	3.39		0		4.13	
4	5.36	0	3.21		2.8	
5	3.41	2.53	0	0	3.42	
6	3.29	2.61	3.15	2.48	0	0

La distancia entre clusters, es ahora entre los dos puntos 1, 4 y los clusters, 1 formado por los puntos 3 y 5; y 2 identificado en la iteración anterior, y en consecuencia la distancia entre los puntos 2 y 6 ahora es 0 porque están en el mismo cluster.

Realizar el paso B del algoritmo de clusterización jerárquica aglomerativa utilizando el algoritmo con la definición de proximidad MIN:

Los datos son ahora:

1. {2.89, 2.94}; 2. {4.36, 5.21}; 4. {5.25, 3.14}; 6. {3.9, 4.27} y C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}}

Punto	1	4	C1-3	C1-5	C2-2	C2-6
1	0					
2	4.15				0	
3	3.39		0		4.13	
4	5.36	0	3.21		2.8	
5	3.41	2.53	0	0	3.42	
6	3.29	2.61	3.15	2.48	0	0

Los dos clusteres más próximos son C1 y C2. Por lo que el tercer cluster, C3, es el formado por esos dos clusters.

Como no tenemos un solo cluster pasamos a la cuarta iteración.

Realizar la cuarta iteración del algoritmo de clusterización jerárquica aglomerativa:

Realizamos el paso A. Cálculo de la matriz de distancias entre clusters.

Los datos son ahora:

1. {0.89, 2.94}; 4. {6.25, 3.14} y C3 {C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}}, C2 {2. {4.36, 5.21}; 6. {3.9, 4.27}}}

Punto	1	4	C3-C1-3	C3-C1-5	C3-C2-2	C3-C2-6
1	0					
2	4.15				0	
3	3.39		0		0	
4	5.36	0	3.21		2.8	
5	3.41	2.53	0	0	0	
6	3.29	2.61	0	0	0	0

La distancia entre clusters, es ahora entre los dos puntos 1, 4 y el cluster 3 formado por los clusters 1 y 2 y en consecuencia la distancia entre los clusters 1 y 2 ahora es 0 porque están en el mismo cluster.

Realizar el paso B del algoritmo de clusterización jerárquica aglomerativa utilizando el algoritmo con la definición de proximidad MIN:

Los datos son ahora:

1. {0.89, 2.94}; 4. {6.25, 3.14} y C3 {C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}}, C2 {2. {4.36, 5.21}; 6. {3.9, 4.27}}}}

Punto	1	4	C3-C1-3	C3-C1-5	C3-C2-2	C3-C2-6
1	0					
2	4.15				0	
3	3.39		0		0	
4	5.36	0	3.21		2.8	
5	3.41	2.53	0	0	0	
6	3.29	2.61	0	0	0	0

Los dos clusteres más próximos son el punto 4 y C3. Por lo que el cuarto cluster, C4, es el formado por esos dos clusters.

Y ya si que tenemos un solo cluster C5 que será el formado por el C4 y el punto 1.

A partir de la muestra... 1. {2.89, 2.94}; 2. {4.36, 5.21}; 3. {3.75, 1.12}; 4. {5.25, 3.14}; 5. {4.1, 1.8}; 6. {3.9, 4.27}. Realizar el paso B del algoritmo de clusterización jerárquica aglomerativa utilizando el algoritmo con la definición de proximidad MAX:

Primera Iteración: Si tomamos la matriz de distancias entre los clusters, teniendo en cuenta que en la primera iteración cada punto constituye un cluster.

Punto	1	2	3	4	5	6
1	0					
2	4.15	0				
3	3.39	4.13	0			
4	5.36	2.8	3.21	0		
5	3.41	3.42	0.76	2.53	0	
6	3.29	1.05	3.15	2.61	2.48	0

Los dos clusteres más próximos son 3 y 5. Por lo que el primer cluster, C1, es el formado por esos dos puntos.

Realizar la segunda iteración del algoritmo de clusterización jerárquica aglomerativa:

Realizamos el paso A. Cálculo de la matriz de distancias entre clusters.

Los datos son ahora:

1. {0.89, 2.94}; 2. {4.36, 5.21}; 4. {6.25, 3.14}; 6. {3.9, 4.27} y C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}}

Punto	1	2	4	6	C1-3	C1-5
1	0					
2	4.15	0				
3	3.39	4.13			0	
4	5.36	2.8	0		3.21	
5	3.41	3.42	2.53		0	0
6	3.29	1.05	2.61	0	3.15	2.48

La distancia entre clusters, es ahora entre los cuatro puntos 1, 2, 4 y 6, y el cluster 1 formado por los puntos 3 y 5, en la iteración anterior, y en consecuencia la distancia entre los puntos 3 y 5 ahora es 0 porque están en el mismo cluster.

Realizar el paso B del algoritmo de clusterización jerárquica aglomerativa utilizando el algoritmo con la definición de proximidad MAX:

Los datos son ahora:

1. {0.89, 2.94}; 2. {4.36, 5.21}; 4. {6.25, 3.14}; 6. {3.9, 4.27} y C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}}

Punto	1	2	4	6	C1-3	C1-5
1	0					
2	4.15	0				
3	3.39	4.13			0	
4	5.36	2.8	0		3.21	
5	3.41	3.42	2.53		0	0
6	3.29	1.05	2.61	0	3.15	2.48

Los dos clusteres más próximos son 2 y 6. Por lo que el segundo cluster, C2, es el formado por esos dos puntos.

Realizar la tercera iteración del algoritmo de clusterización jerárquica aglomerativa:

Realizamos el paso A. Cálculo de la matriz de distancias entre clusters.

Los datos son ahora:

1. {0.89, 2.94}; 4. {6.25, 3.14}, C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}}, y C2 {2. {4.36, 5.21}; 6. {3.9, 4.27}} y

Punto	1	4	C1-3	C1-5	C2-2	C2-6
1	0					
2	4.15				0	
3	3.39		0		4.13	
4	5.36	0	3.21		2.8	
5	3.41	2.53	0	0	3.42	
6	3.29	2.61	3.15	2.48	0	0

La distancia entre clusters, es ahora entre los dos puntos 1, 4 y los clusters, 1 formado por los puntos 3 y 5; y 2 identificado en la iteración anterior, y en consecuencia la distancia entre los puntos 2 y 6 ahora es 0 porque están en el mismo cluster.

Realizar el paso B del algoritmo de clusterización jerárquica aglomerativa utilizando el algoritmo con la definición de proximidad MAX:

Los datos son ahora:

1. {0.89, 2.94}; 4. {6.25, 3.14}, C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}}, y C2 {2. {4.36, 5.21}; 6. {3.9, 4.27}}

Punto	1	4	C1-3	C1-5	C2-2	C2-6
1	0					
2	4.15				0	
3	3.39		0		4.13	
4	5.36	0	3.21		2.8	
5	3.41	2.53	0	0	3.42	
6	3.29	2.61	3.15	2.48	0	0

Los dos clusteres más próximos son 4 y C2. Por lo que el tercer cluster, C3, es el formado por esos dos clusters.

1. {0.89, 2.94}; 4. {6.25, 3.14}, C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}}, y C2 {2. {4.36, 5.21}; 6. {3.9, 4.27}}}
¿Por qué es 2.8 la menor proximidad entre clusters con el algoritmo MAX?:

Juan J. Cuadrado

Punto	1	4	C1-3	C1-5	C2-2	C2-6
1	0					
2	4.15				0	
3	3.39		0		4.13	
4	5.36	0	3.21		2.8	
5	3.41	2.53	0	0	3.42	
6	3.29	2.61	3.15	2.48	0	0

La proximidad entre clusters con MAX se define como distancia que haya entre los dos puntos más lejanos de los dos clusters.

distancia (1-4)=5.36

distancia (C1-C2) = distancia ({3. {3.75, 1.12}; 5. {4.1, 1.8}} - {2. {4.36, 5.21}; 6. {3.9, 4.27}}) =
máximo (distancia (3,2), distancia (3,6), distancia (5,2), distancia (5,6)) =
máximo (4.15, 3.15, 3.41, 2.48) = 4.15

distancia (C1-1) = distancia ({3. {3.75, 1.12}; 5. {4.1, 1.8}} - 1.{0.89, 2.94}) =
máximo (distancia (3,1), distancia (5,1)) = máximo (3.39, 3.41) = 3.41

distancia (C1-4) = distancia ({3. {3.75, 1.12}; 5. {4.1, 1.8}} - 4. {6.25, 3.14}) =
máximo (distancia (3,4), distancia (5,4)) = máximo (3.21, 2.53) = 3.21

distancia (C2-1) = 4.15 distancia (C2-4) = 2.8



Realizar la cuarta iteración del algoritmo de clusterización jerárquica aglomerativa:

Realizamos el paso A. Cálculo de la matriz de distancias entre clusters.

Los datos son ahora:

1. {0.89, 2.94}; C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}} y C3 {C2 {2. {4.36, 5.21}; 6. {3.9, 4.27}}, 4. {6.25, 3.14}}

Punto	1	C1-3	C1-5	C3-C2-2	C3-C2-6	C3-4
1	0					
2	4.15			0		
3	3.39	0		4.13		
4	5.36	3.21		0		0
5	3.41	0	0	3.42		2.53
6	3.29	3.15	2.48	0	0	0

La distancia entre clusters, es ahora entre 4 el punto y los clusters, 1 formado por los puntos 3 y 5; y 3 identificado en la iteración anterior, y en consecuencia la distancia entre el punto 4 y el cluster 2 0 porque están en el mismo cluster.

Realizar la cuarta iteración del algoritmo de clusterización jerárquica aglomerativa:

Realizamos el paso A. Cálculo de la matriz de distancias entre clusters.

Los datos son ahora:

1. {0.89, 2.94}; C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}} y C3 {C2 {2. {4.36, 5.21}; 6. {3.9, 4.27}}, 4. {6.25, 3.14}}

Punto	1	C1-3	C1-5	C3-C2-2	C3-C2-6	C3-4
1	0					
2	4.15			0		
3	3.39	0		4.13		
4	5.36	3.21		0		0
5	3.41	0	0	3.42		2.53
6	3.29	3.15	2.48	0	0	0

1 con 3 (3.39) y 5 (3.41): distancia max 3.41/1 con 2 (4.15), 6 (3.29) y 4 (5.36): distancia max 5.36

La distancia C1-C2 es 4.13

Los dos clusteres más próximos son el punto 1 y C1. Por lo que el cuarto, C4, es el formado por esos dos clusters: C4 {1. {0.89, 2.94}; C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}}}}

Y ya si que tenemos un solo cluster C5 que será el formado por el C3 y el C4.

A partir de la muestra... 1. {2.89, 2.94}; 2. {4.36, 5.21}; 3. {3.75, 1.12}; 4. {5.25, 3.14}; 5. {4.1, 1.8}; 6. {3.9, 4.27}. Realizar el paso B del algoritmo de clusterización jerárquica aglomerativa utilizando el algoritmo con la definición de proximidad **Group Average**:

$$\text{proximidad} (C_i, C_j) = \frac{\sum_{i=1}^m \text{proximidad} (x_i, y_j)}{m*n}$$

Primera Iteración: Si tomamos la matriz de distancias entre los clusters, teniendo en cuenta que en la primera iteración cada punto constituye un cluster.

Punto	1	2	3	4	5	6
1	0					
2	4.15	0				
3	3.39	4.13	0			
4	5.36	2.8	3.21	0		
5	3.41	3.42	0.76	2.53	0	
6	3.29	1.05	3.15	2.61	2.48	0

Realizar la segunda iteración del algoritmo de clusterización jerárquica aglomerativa:

Realizamos el paso A. Cálculo de la matriz de distancias entre clusters.

Los datos son ahora:

1. {0.89, 2.94}; 2. {4.36, 5.21}; 4. {6.25, 3.14}; 6. {3.9, 4.27} y C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}}

Punto	1	2	4	6	C1-3	C1-5	1	2	4	6	C1-3	C1-5
1	0						0					
2	4.15	0					4.15	0				
C1-3	3.39	4.13			0		3.4	3.78			0	
4	5.36	2.8	0		3.21		5.36	2.8	0		2.87	
C1-5	3.41	3.42	2.53		0	0	3.4	3.78	2.87		0	0
6	3.29	1.05	2.61	0	3.15	2.48	3.29	1.05	2.61	0	2.82	2.82

$$\text{proximidad } (p_1, C_1) = \frac{\sum_{j=1}^2 \text{proximidad } ((p_1, p_3), (p_1, p_5))}{2 * 1} = \frac{3.39 + 3.41}{2} = 3.4$$

La distancia entre clusters, es ahora entre los cuatro puntos 1, 2, 4 y 6, y el cluster 1 formado por los puntos 3 y 5, en la iteración anterior, pero ahora cambian las distancias porque se hace con la media

Juan J. Cuadrado

Reto de actividades Teoría - Bb Collaborate

Juan José Cuadrado is sharing content

Clasificación no supervisada

Realizar el paso B del algoritmo de clusterización jerárquica aglomerativa utilizando el algoritmo con la definición de proximidad Group Average:

Los datos son ahora:

1. {0.89, 2.94}; 2. {4.36, 5.21}; 4. {6.25, 3.14}; 6. {3.9, 4.27} y C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}}

Punto	1	2	4	6	C1-3	C1-5
1	0					
2	4.15	0				
C1-3	3.4	3.78			0	
4	5.36	2.8	0		2.87	
C1-5	3.4	3.78	2.87		0	0
6	3.29	1.05	2.61	0	2.82	2.82

Los dos clusteres más próximos son 2 y 6. Por lo que el segundo cluster, C2, es el formado por esos dos puntos.

Buscar cualquier cosa Juan José Cuadrado 11:36 01/12/2020

Clasificación no supervisada

Juan J. Cuadrado

Realizar la tercera iteración del algoritmo de clusterización jerárquica aglomerativa:

Realizamos el paso A. Cálculo de la matriz de distancias entre clusters.

Los datos son ahora:

1. {0.89, 2.94}; 4. {6.25, 3.14}; C1{3. {3.75, 1.12}; 5. {4.1, 1.8}, C2{2. {4.36, 5.21}; 6. {3.9, 4.27}}}

Punto	1	4	C1-3	C1-5	C2-2	C2-6	1	4	C1-3	C1-5	C2-2	C2-6
1	0						0					
C2-2	4.15				0		3.72				0	
C1-3	3.39		0		4.13		3.4		0		3.3	
4	5.36	0	3.21		2.8		5.36	0	2.87		2.7	
C1-5	3.41	2.53	0	0	3.42		3.4	2.87	0	0	3.3	
C2-6	3.29	2.61	3.15	2.48	1.05	0	3.72	2.7	3.3	3.3	3.3	0

$$\text{proximidad } (C_1, C_2) = \frac{\sum_{j=1}^2 \text{proximidad } ((p_3, p_2), (p_3, p_6), (p_5, p_2), (p_5, p_6))}{2*2} = \frac{4.13 + 3.15 + 3.42 + 2.48}{4} = 3.3$$

La distancia entre clusters, es ahora entre los dos puntos 1 y 4, y los dos clusters 1 y 2, pero ahora cambian casi todas las distancias porque se hace con la media.

Realizar el paso B del algoritmo de clusterización jerárquica aglomerativa utilizando el algoritmo con la definición de proximidad Group Average:

Los datos son ahora:

1. {0.89, 2.94}; 2. {4.36, 5.21}; 4. {6.25, 3.14}; 6. {3.9, 4.27} y C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}}

Punto	1	4	C1-3	C1-5	C2-2	C2-6
1	0					
C2-2	3.72				0	
C1-3	3.4		0		3.3	
4	5.36	0	2.87		2.7	
C1-5	3.4	2.87	0	0	3.3	
C2-6	3.72	2.7	3.3	3.3	3.3	0

Los dos clusteres más próximos son 4 y C2. Por lo que el segundo tercer cluster, C3, es el formado por esos dos clusters.

Como no tenemos un solo cluster pasamos a la cuarta iteración.

Realizar la tercera iteración del algoritmo de clusterización jerárquica aglomerativa:
Realizamos el paso A. Cálculo de la matriz de distancias entre clusters.

Los datos son ahora:

1. {0.89, 2.94}; 4. {6.25, 3.14}; C1{3. {3.75, 1.12}; 5. {4.1, 1.8}, C2{2. {4.36, 5.21}; 6. {3.9, 4.27}}}

Punto	1	C1-3	C1-5	C3-2	C3-6	C3-4	1	C1-3	C1-5	C3-2	C3-6	C3-4
1	0						0					
C2-2	4.15			0			4.27			0		
C1-3	3.39	0		4.13			3.4	0		3.15		
4	5.36	3.21		2.8		0	4.27	3.15		0		0
C1-5	3.41	0	0	3.42		2.53	3.4	0	0	3.15		3.15
C2-6	3.29	3.15	2.48	0	0	2.61	4.27	3.15	3.15	0	0	0

$$\text{proximidad } (C_1, C_3) = \frac{\sum_{i=1}^3 \sum_{j=1}^2 \text{proximidad } ((p_3, p_2), (p_3, p_6), (p_3, p_4), (p_5, p_2), (p_5, p_6), (p_5, p_4))}{2*3} \\ = \frac{4.13 + 3.15 + 3.21 + 3.42 + 2.48 + 2.53}{6} = 3.15$$

La distancia entre clusters, es ahora entre los dos puntos 1y 4, y los dos clusters 1 y 2, pero ahora cambian casi todas las distancias porque se hace con la media.

Realizar el paso B del algoritmo de clusterización jerárquica aglomerativa utilizando el algoritmo con la definición de proximidad Group Average:

Los datos son ahora:

1. {0.89, 2.94}; 2. {4.36, 5.21}; 4. {6.25, 3.14}; 6. {3.9, 4.27} y C1 {3. {3.75, 1.12}; 5. {4.1, 1.8}}

Punto	1	C1-3	C1-5	C3-2	C3-6	C3-4
1	0					
C3-2	4.27			0		
C1-3	3.4	0		3.15		
C3-4	4.27	3.15		0		0
C1-5	3.4	0	0	3.15		3.15
C3-6	4.27	3.15	3.15	0	0	0

Los dos clusteres más próximos son C1 y C3. Por lo que el segundo cuarto cluster, C4, es el formado por esos dos clústeres.

Y ya si que tenemos un solo cluster C5 que será el formado por el C4 y el punto 1.

Datos Anómalos

La identificación de datos anómalos es fundamental para realizar un buen análisis de datos.

Los datos anómalos pueden ser:

- **Datos erróneos**, procedentes de errores de medida, que deben ser eliminados porque conducirán a un análisis de datos con conclusiones erróneas. La detección de outliers es a menudo parte del preprocesado de datos, específicamente de la limpieza de datos.
- **Datos correctos**, con mucha significación, que se apartan de lo normal y que deben ser analizados con mucho detenimiento porque pueden conducir a hallazgos importantes.

Los datos anómalos suelen ser entre el 1% y el 3% de los datos.

Un evento que aparezca una vez de cada mil ocurriría miles de veces cuando se consideran millones de datos (o millones de veces si se considerasen billones de datos)

Un evento compuesto por varias características individuales puede ser anómalo aunque los valores individuales de cada característica no lo sean.



Datos Anómalos

Los estudios de identificación se sucesos anómalos o outliers **buscan encontrar y catalogar como outliers aquellos sucesos que sean muy diferentes del resto de los que componen la muestra estudiada.**

Para dar una medida de lo anómalo que es un suceso se ha definido la variable puntuación o **grado de outlier** (outlier score) o grado de anomalía. La manera de definir el grado de outlier depende de la técnica empleada.

El grado de outlier se fija arbitrariamente por el analista de datos teniendo en cuenta el estudio que está realizando.

Datos Anómalos

La manera de medir el grado de outlier depende de la técnica empleada.

Se puede establecer una **clasificación de las técnicas de identificación de outliers** basándose en dos criterios:

El primer criterio de clasificación de las técnicas de detección de outliers está basado en **si se dispone o no de un modelo definido en un análisis previo de los datos**. Por ejemplo si se tiene una clasificación o regresión previa de los datos, los datos que se ajusten al modelo serán considerados normales mientras que los que no se ajusten serán outliers.

Datos Anómalos

Se puede establecer una **clasificación de las técnicas de identificación de outliers** basándose en dos criterios:

El segundo criterio de clasificación dependerá de la **técnica empleada** en el análisis:

- **Estadísticas.** Se han desarrollado un gran número de técnicas estadísticas de identificación de outliers.
 - **Ordenación. Caja y Bigotes.**
 - **Dispersión. Desviación Típica.**
 - **Regresión. Error estándar de los residuos.**
- **Basadas en la proximidad.** Buscan sucesos que están muy separados del resto de sucesos, se basan en la definición de distancias. **K-vecino más próximo.**
- Basadas en la densidad. Las que estén en una zona espacial en la que hay menos densidad de sucesos que la media observada en esa muestra.
- Basadas en clusters. Se aplican a sucesos previamente clusterizados.

Datos Anómalos

La identificación de sucesos anómalos a partir de la técnica **Caja y Bigotes** sigue un proceso de 4 pasos:

1. **Determinación del grado de outlier** o distancia a la que un suceso (punto) se considera un outlier. Se elige arbitrariamente.

2. **Se ordenan los datos y se obtienen los cuartiles** utilizando la ecuación:

$$\tilde{x}_c = x_{[nc]+1} \text{ si } nc \notin \mathbb{N} \quad [nc] \text{ parte entera de } nc$$

$$\tilde{x}_c = \frac{x_{[nc]} + x_{[nc]+1}}{2} \text{ si } nc \in \mathbb{N}$$

3. **Se calculan los límites del intervalo para los valores atípicos** utilizando la ecuación:

$$(Q_1 - d(Q_3 - Q_1), Q_3 + d(Q_3 - Q_1))$$

4. Se identifican los outliers como los valores que quedan fuera del intervalo calculado en el paso 3.

Datos Anómalos

Se tiene la siguiente muestra de siete valores de resistencia y densidad para diferentes tipos de hormigón {resistencia, densidad}: {3, 2; 3.5, 12; 4.7, 4.1; 5.2, 4.9; 7.1, 6.1; 6.2, 5.2; 14, 5.3}. Realizar un análisis de identificación de outliers de la resistencia utilizando el método basado en las medidas de ordenación, método de caja y bigotes.

Determinación del grado de outlier o distancia a la que un suceso (punto) se considera un outlier. Se elige arbitrariamente. Elegimos $d=1.5$.

Se ordenan los datos y se obtienen los cuartiles utilizando la ecuación:

Los valores ordenados son: {3, 3.5, 4.7, 5.2, 6.2, 7.1, 14}

$$nc = 7 \cdot \frac{1}{4} = 1.75 \notin \mathbb{N} \rightarrow \tilde{x}_1 = x_{1+1} = x_2 = 3.5$$

$$nc = 7 \cdot \frac{3}{4} = 5.25 \notin \mathbb{N} \rightarrow \tilde{x}_3 = x_{5+1} = x_6 = 7.1$$

Se calculan los límites del intervalo para los valores atípicos utilizando la ecuación:

$$(Q_1 - d(Q_3 - Q_1), Q_3 + d(Q_3 - Q_1)) = (3.5 - 1.5(7.1 - 3.5), 7.1 + 1.5(3.6))$$

Datos Anómalos

Se tiene la siguiente muestra de siete valores de resistencia y densidad para diferentes tipos de hormigón {resistencia, densidad}: {3, 2; 3.5, 12; 4.7, 4.1; 5.2, 4.9; 7.1, 6.1; 6.2, 5.2; 14, 5.3}. Realizar un análisis de identificación de outliers de la resistencia utilizando el método basado en las medidas de ordenación, método de caja y bigotes.

$$(3.5 - 1.5(7.1 - 3.5), 7.1 + 1.5(3.6)) = (-1.9, 12.5)$$

Se identifican los outliers como los valores que quedan fuera del intervalo calculado en el paso 3.

El outlier es el punto $x_7 = 14$ porque no se encuentra dentro del intervalo de valores normales.

Datos Anómalos

La identificación de sucesos anómalos a partir de la técnica de la Desviación Típica sigue un proceso de 5 pasos:

1. Determinación del grado de outlier o distancia a la que un suceso (punto) se considera un outlier. Se elige arbitrariamente.

2. Se obtiene la media aritmética utilizando la ecuación:

$$\bar{x}_a = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i}$$

3. Se obtiene la desviación típica utilizando la ecuación:

$$s_a = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x}_a)^2}{\sum_{i=1}^n f_i}}$$

4. Se calculan los límites del intervalo para los valores atípicos utilizando la ecuación.

$$(\bar{x}_a - ds_a, \bar{x}_a + ds_a)$$

5. Se identifican los outliers como los valores que quedan fuera del intervalo calculado en el paso 4.

Datos Anómalos

Para la muestra del ejercicio anterior realizar un análisis de identificación de outliers de la densidad utilizando el método basado en las medidas de dispersión, desviación típica.

1. **Determinación del grado de outlier o distancia a la que un suceso (punto) se considera un outlier.** Se elige arbitrariamente. Elegimos $d=2$. Tendremos dentro del intervalo el $\left(1 - \frac{1}{4}\right) 100 = 75\%$ de los datos

2. **Se obtiene la media aritmética** utilizando la ecuación:

$$\bar{x}_a = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i} = \frac{2 + 12 + 4.1 + 4.9 + 6.1 + 5.2 + 5.3}{7} = 5.66$$

3. **Se obtiene la desviación típica** utilizando la ecuación:

$$s_a = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x}_a)^2}{\sum_{i=1}^n f_i}} = \sqrt{\frac{(2 - 5.66)^2 + 40.2 + 2.43 + 0.58 + 0.19 + 0.21 + 0.13}{7}} = \sqrt{8.16} = 2.86$$

4. **Se calculan los límites del intervalo para los valores atípicos** utilizando la ecuación.

$$(\bar{x}_a - ds_a, \bar{x}_a + ds_a) = (5.66 - 2.286, 5.66 + 2.286)$$



Datos Anómalos

Para la muestra del ejercicio anterior realizar un análisis de identificación de outliers de la densidad utilizando el método basado en las medidas de dispersión, desviación típica.

$$(5.66 - 2.286, 5.66 + 2.286) = (0, 11.38)$$

Se identifican los outliers como los valores que quedan fuera del intervalo calculado en el paso 4.

El outlier es el punto $x_2 = 12$ porque no se encuentra dentro del intervalo de valores normales.

Datos Anómalos

La identificación de sucesos anómalos a partir de la técnica de la Regresión sigue un proceso de 5 pasos:

1. **Determinación del grado de outlier** o distancia a la que un suceso (punto) se considera un outlier. Se elige arbitrariamente (normalmente 3 ó 4)
2. **Se obtiene la regresión lineal** utilizando las ecuaciones:

$$s_{xy} = \frac{\sum_{i=1}^n \sum_{j=1}^m f_{ij} x_i y_j}{\sum_{i=1}^n f_i} - \left(\frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i} \right) \cdot \left(\frac{\sum_{j=1}^m f_j y_j}{\sum_{j=1}^m f_j} \right); r_{xy} = \frac{s_{xy}}{s_x s_y}; b = \frac{s_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x} \quad a = \bar{y} - b \bar{x}$$

3. **Se obtiene el error estándar de los residuos** utilizando la ecuación:

$$s_r = \sqrt{\frac{\sum_{i=1}^n (y_i - y_{ci})^2}{n}}$$

4. **Se calculan los límites del intervalo para los valores atípicos** utilizando la ecuación.

$$ds_r$$

5. **Se identifican los outliers como aquellos tales que** $|y_i - y_{ci}| > ds_r$

Datos Anómalos

Para la muestra del ejercicio anterior realizar un análisis de identificación de outliers para la regresión de ambas, densidad en función de la resistencia, variables utilizando el método del error estándar de los residuos.

3. **Se obtiene el error estándar de los residuos** utilizando la ecuación:

$$s_r = \sqrt{\frac{\sum_{i=1}^n (y_i - y_{ci})^2}{n}}$$

Primero calculamos y_{ci} para todos los puntos. $y_{ci} = 6.015 - 0.057x_i$

$$y_{c1} = 6.015 - 0.057(3) = 5.84$$

$$y_{c2} = 6.015 - 0.057(3.5) = 5.82$$

$$y_{c3} = 6.015 - 0.057(4.7) = 5.75$$

$$y_{c4} = 6.015 - 0.057(3.2) = 5.72$$

$$y_{c5} = 6.015 - 0.057(7.1) = 5.61$$

$$y_{c6} = 6.015 - 0.057(6.2) = 5.66$$

$$y_{c7} = 6.015 - 0.057(14) = 5.22$$

Y se tiene un s_r :

$$s_r = \sqrt{\frac{\sum_{i=1}^n (y_i - y_{ci})^2}{n}} = \sqrt{\frac{(2 - 5.84)^2 + \dots + (5.3 - 5.22)^2}{7}} = 2.85$$

Datos Anómalos

Para la muestra del ejercicio anterior realizar un análisis de identificación de outliers para la regresión de ambas, densidad en función de la resistencia, variables utilizando el método del error estándar de los residuos.

4. Se calculan los límites del intervalo para los valores atípicos utilizando la ecuación.

$$ds_r = 2.285 = 5.7$$

5. Se identifican los outliers como aquellos tales que $|y_i - y_{ci}| > ds_r$

El punto (3.5, 12) se identifica como outlier ya que $|12 - 5.82| = 6.18 > 5.7$

e - Google Chrome

Datos Anómalos

La identificación de sucesos anómalos a partir de la técnica **K-vecinos** sigue un proceso de 2 pasos:

- A. El paso A se puede separar a su vez en 2 subpasos:

1. Determinación del grado de outlier o distancia a la que un suceso (punto) se considera un outlier. Se elige arbitrariamente.
2. Determinación del número de orden, o K, del vecino más próximo para el que un suceso tiene que tener el grado de outlier para que el suceso sea considerado un outlier. Se elige arbitrariamente.

Datos Anómalos

Se tiene un conjunto de calificaciones académicas, pertenecientes a un grupo de alumnos. Las calificaciones tanto de Teoría como de Laboratorio tendrán valores de 0 a 5, donde 5 será la mayor calificación posible y 0 la menor. La muestra está formada por las siguientes cinco calificaciones (sucesos): 1. {4, 4}; 2. {4, 3}; 3. {5, 5}; 4. {1, 1}; 5. {5, 4}. Realizar el paso A del algoritmo K-vecinos para la búsqueda de datos anómalos.

Determinación del grado de outlier o distancia a la que un suceso (punto) se considera un outlier. Se elige arbitrariamente.

Como se elige arbitrariamente, teniendo en cuenta las características del problema y de la muestra tomamos una distancia euclídea de $d=2.5$.

Determinación del número de orden, o K, del vecino más próximo para el que un suceso tiene que tener el grado de outlier para que el suceso sea considerado un outlier.

Como se elige arbitrariamente, teniendo en cuenta las características del problema y de la muestra tomamos una el 3er vecino o suceso más próximo. $K=3$.

Datos Anómalos

La identificación de sucesos anómalos a partir de la técnica K-vecinos sigue un proceso de 2 pasos:

B. El paso B se puede separar a su vez en 3 subpasos:

1. Cálculo de las distancias euclídeas entre todos los puntos.
2. Ordenación por distancias de los vecinos de cada punto hasta llegar al K definido.
3. Identificación de los outliers como aquellos sucesos cuyo K vecino se encuentre a una distancia mayor que el grado de outlier definido.

Datos Anómalos

Realizar el paso B del algoritmo K-vecinos para la identificación de outliers del ejercicio anterior.

Cálculo de las distancias Euclídeas entre todos los puntos. Calculamos la distancia de cada punto con el resto de los puntos de la muestra.

$$\text{Puntos 1-2, } \{\{4, 4\}, \{4, 3\}\}: d_{12} = \sqrt{\sum_{i=1}^2 (p_i - q_i)^2} = \sqrt{(4-4)^2 + (4-3)^2} = 1$$

$$\text{Puntos 1-3, } \{\{4, 4\}, \{5, 5\}\}: d_{13} = \sqrt{(4-5)^2 + (4-5)^2} = 1.41$$

$$\text{Puntos 1-4, } \{\{4, 4\}, \{1, 1\}\}: d_{14} = \sqrt{(4-1)^2 + (4-1)^2} = 4.24$$

$$\text{Puntos 1-5, } \{\{4, 4\}, \{5, 4\}\}: d_{15} = \sqrt{(4-5)^2 + (4-4)^2} = 1$$

$$\text{Puntos 2-3, } \{\{4, 3\}, \{5, 5\}\}: d_{23} = \sqrt{(4-5)^2 + (3-5)^2} = 2.24$$

$$\text{Puntos 2-4, } \{\{4, 3\}, \{1, 1\}\}: d_{24} = \sqrt{(4-1)^2 + (3-1)^2} = 3.61$$

$$\text{Puntos 2-5, } \{\{4, 3\}, \{5, 4\}\}: d_{25} = \sqrt{(4-5)^2 + (3-4)^2} = 1.41$$

$$\text{Puntos 3-4, } \{\{5, 5\}, \{1, 1\}\}: d_{34} = \sqrt{(5-1)^2 + (5-1)^2} = 5.66$$

$$\text{Puntos 3-5, } \{\{5, 5\}, \{5, 4\}\}: d_{35} = \sqrt{(5-5)^2 + (5-4)^2} = 1$$

$$\text{Puntos 4-5, } \{\{1, 1\}, \{5, 4\}\}: d_{45} = \sqrt{(1-5)^2 + (1-4)^2} = 5$$

Datos Anómalos

Realizar el paso B del algoritmo K-vecinos para la identificación de outliers del ejercicio anterior.

Identificación de los outliers como aquellos sucesos cuyo K vecino se encuentre a una distancia mayor que el grado de outlier definido.

La grado de outlier es una distancia de 2.5 y el K elegido es el 3, según esto tenemos:

- Punto 1: Distancia al tercer punto más próximo, punto 3: 1.41
- Punto 2: Distancia al tercer punto más próximo, punto 3: 2.24
- Punto 3: Distancia al tercer punto más próximo, punto 2: 2.24
- Punto 5: Distancia al tercer punto más próximo, punto 2: 1.41
- **Punto 4: Distancia al tercer punto más próximo, punto 5: 5**

En consecuencia el único punto cuya distancia a su 3er vecino más próximo es mayor de 2.5 es el Punto 4 (1, 1), por lo que debe ser considerado outlier.

Datos Anómalos

La identificación de sucesos anómalos utilizando la **Densidad Relativa** y la técnica **Local Outlier Factor, LOF** sigue un proceso de 3 pasos:

- A. **Determinación del grado de outlier de cada punto mediante el Cálculo de la densidad, d, de cada punto.** Hay diferentes definiciones de la densidad. Una de las más utilizadas es:

$$\text{densidad } (x_i, K) = \left(\frac{\sum_{x_j \in N(x_i, K)} \text{distancia}(x_i, x_j)}{\text{cardinal } N(x_i, K)} \right)^{-1}$$

El paso A implica 4 subpasos

1. **Determinación del número de orden, o K, del vecino más próximo que se va a utilizar para calcular la densidad de cada punto.** Se elige arbitrariamente.
2. **Cálculo de las distancias entre cada punto y el resto de los puntos.** Distancia Manhattan, para dos dimensiones:

$$\text{distancia } (x_i, x_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}|$$

3. **Cálculo del cardinal o tamaño del conjunto N para cada punto.** N es el conjunto que contiene los vecinos cuya distancia a x_i es igual o inferior a la del K vecino más próximo.
4. **Cálculo de la densidad, d, de cada punto.** Al calcularse la densidad en cada punto como el inverso de la media de la distancia de los K vecinos más próximos, esta técnica está muy relacionada con la de la proximidad.

Datos Anómalos

Se tiene un conjunto de calificaciones académicas, pertenecientes a un grupo de alumnos. Las calificaciones tanto de Teoría como de Laboratorio tendrán valores de 0 a 5, donde 5 será la mayor calificación posible y 0 la menor. La muestra está formada por las siguientes cinco calificaciones (sucesos): 1. {4, 4}; 2. {4, 3}; 3. {5, 5}; 4. {1, 1}; 5. {5, 4}. Realizar el paso A.1 del algoritmo LOF para la búsqueda de datos anómalos.

Determinación del número de orden, o K, del vecino más próximo que se va a utilizar para calcular la densidad de cada punto. Se elige arbitrariamente.

Como se elige arbitrariamente, teniendo en cuenta las características del problema y de la muestra tomamos una el 3er vecino o suceso más próximo. K=3.

Datos Anómalos

Realizar el paso A.2 del algoritmo LOF para la identificación de outliers del ejercicio anterior.

Cálculo de las distancias Manhattan entre todos los puntos. Calculamos la distancia de cada punto con el resto de los puntos de la muestra.

Puntos 1-2, $\{(4, 4), \{4, 3\}\}$: $d_{12} = |x_{11} - x_{21}| + |x_{12} - x_{22}| = |4 - 4| + |4 - 3| = 1$

Puntos 1-3, $\{(4, 4), \{5, 5\}\}$: $d_{13} = |4 - 5| + |4 - 5| = 2$

Puntos 1-4, $\{(4, 4), \{1, 1\}\}$: $d_{14} = |4 - 1| + |4 - 1| = 6$

Puntos 1-5, $\{(4, 4), \{5, 4\}\}$: $d_{15} = |4 - 5| + |4 - 4| = 1$

Puntos 2-3, $\{(4, 3), \{5, 5\}\}$: $d_{23} = |4 - 5| + |3 - 5| = 3$

Puntos 2-4, $\{(4, 3), \{1, 1\}\}$: $d_{24} = |4 - 1| + |3 - 1| = 5$

Puntos 2-5, $\{(4, 3), \{5, 4\}\}$: $d_{25} = |4 - 5| + |3 - 4| = 2$

Puntos 3-4, $\{(5, 5), \{1, 1\}\}$: $d_{34} = |5 - 1| + |5 - 1| = 8$

Puntos 3-5, $\{(5, 5), \{5, 4\}\}$: $d_{35} = |5 - 5| + |5 - 4| = 1$

Puntos 4-5, $\{(1, 1), \{5, 4\}\}$: $d_{45} = |1 - 5| + |1 - 4| = 7$

Datos Anómalos

Realizar el paso A.3 del algoritmo LOF para la identificación de outliers del ejercicio anterior.

Ordenación por distancias de los vecinos de cada punto hasta llegar al K definido, 3, para calcular el N de cada punto.

- Punto 1: Mínima distancia, punto más próximo, el punto 2: 1. Segunda distancia, punto 5: 1. Y por último la distancia al tercer punto más próximo, que es el K elegido es al punto 3: 2. Por lo que $N = 3$.
- Punto 2: Mínima distancia, punto 1: 1. Segunda distancia, punto 5: 2. Distancia al tercer punto más próximo, que es el K elegido es al punto 3: 3. Por lo que $N = 3$.
- Punto 3: Mínima distancia, punto 5: 1. Segunda distancia, punto 1: 2. Distancia al tercer punto más próximo, que es el K elegido es al punto 2: 3. Por lo que $N = 3$.
- Punto 4: Mínima distancia, punto 2: 5. Segunda distancia, punto 1: 6. Distancia al tercer punto más próximo, que es el K elegido es al punto 5: 7. Por lo que $N = 3$.
- Punto 5: Mínima distancia, punto 1: 1. Segunda distancia, punto 3: 1. Distancia al tercer punto más próximo, que es el K elegido es al punto 2: 2. Por lo que $N = 3$.

No siempre N coincide con K. Si hubiéramos tomado $K = 1$, por ejemplo para el punto 1. La mínima distancia, que sería $K = 1$, sería a los puntos 1 y 5, por lo que N sería igual a 2.

Datos Anómalos

Realizar el paso A.3 del algoritmo LOF para la identificación de outliers del ejercicio anterior.

Ordenación por distancias de los vecinos de cada punto hasta llegar al K definido, 3, para calcular el N de cada punto.

- Punto 1: Mínima distancia, punto más próximo, el punto 2: 1. Segunda distancia, punto 5: 1. Y por último la distancia al tercer punto más próximo, que es el K elegido es al punto 3: 2. Por lo que N = 3.
- Punto 2: Mínima distancia, punto 1: 1. Segunda distancia, punto 5: 2. Distancia al tercer punto más próximo, que es el K elegido es al punto 3: 3. Por lo que N = 3.
- Punto 3: Mínima distancia, punto 5: 1. Segunda distancia, punto 1: 2. Distancia al tercer punto más próximo, que es el K elegido es al punto 2: 3. Por lo que N = 3.
- Punto 4: Mínima distancia, punto 2: 5. Segunda distancia, punto 1: 6. Distancia al tercer punto más próximo, que es el K elegido es al punto 5: 7. Por lo que N = 3.
- Punto 5: Mínima distancia, punto 1: 1. Segunda distancia, punto 3: 1. Distancia al tercer punto más próximo, que es el K elegido es al punto 2: 2. Por lo que N = 3.

No siempre N coincide con K. Si hubiéramos tomado K = 1, por ejemplo para el punto 1. La mínima distancia, que sería K = 1, sería a los puntos 1 y 5, por lo que N sería igual a 2.

Si tuviéramos por ejemplo K = 2 y hubiera tres puntos a la misma distancia que fuese la segunda mínima distancia al punto y otro a menor distancia, N sería igual a 4.

Datos Anómalos

Realizar el paso A.4 del algoritmo LOF para la identificación de outliers del ejercicio anterior.

Cálculo de la densidad, d, de cada punto.

$$\text{densidad } (x_i, K) = \left(\frac{\sum_{x_j \in N(x_i, K)} \text{distancia}(x_i, x_j)}{\text{cardinal } N(x_i, K)} \right)^{-1}$$

- P1: $d(x_1, 3) = \left(\frac{\text{distancia}(x_1, x_2) + \text{distancia}(x_1, x_5) + \text{distancia}(x_1, x_3)}{\text{cardinal } N(x_1, 3)} \right)^{-1} = \left(\frac{1+1+2}{3} \right)^{-1} = 0.75$
- P2: $d(x_2, 3) = \left(\frac{1+2+3}{3} \right)^{-1} = 0.5$
- P3: $d(x_3, 3) = \left(\frac{1+2+3}{3} \right)^{-1} = 0.5$
- P4: $d(x_4, 3) = \left(\frac{5+6+7}{3} \right)^{-1} = 0.17$
- P5: $d(x_5, 3) = \left(\frac{1+1+2}{3} \right)^{-1} = 0.75$

Datos Anómalos

La identificación de sucesos anómalos utilizando la **Densidad Relativa** y la técnica **Local Outlier Factor, LOF** sigue un proceso de 3 pasos:

- B. **Cálculo de la densidad relativa media, drm, de cada punto.** Hay diferentes definiciones de la densidad relativa media. Una de las más utilizadas es:

$$\text{densidad relativa media } (x_i, K) = \frac{\text{densidad } (x_i, K)}{\frac{\sum_{x_j \in N(x_i, K)} \text{densidad}(x_j, K)}{\text{cardinal } N(x_i, K)}}$$

Que calcula la proporción entre la densidad en un punto y la media de densidades del conjunto N que define dicho punto a partir del número de orden K. La densidad relativa media tenderá a cero en los outliers.

- C. **Obtención de los outliers,** como aquellos puntos cuya densidad relativa media sea significativamente menor que la del resto de elementos de la muestra. Se pueden establecer diferentes métodos para establecer cuando la drm es significativamente menor.

La densidad relativa, que tiene en cuenta el vecindario del punto, el conjunto N, se utiliza porque si solo se utiliza la densidad absoluta pueden no identificarse correctamente outliers en muestras de datos con regiones de diferentes densidades.

Datos Anómalos

Realizar el paso B del algoritmo LOF para la identificación de outliers del ejercicio anterior.

Cálculo de la densidad relativa media, drm, de cada punto.

$$\text{densidad relativa media } (x_i, K) = \frac{\text{densidad } (x_i, K)}{\frac{\sum_{x_j \in N(x_i, K)} \text{densidad}(x_j, K)}{\text{cardinal } N(x_i, K)}}$$

- P1: $\text{drm } (x_1, 3) = \frac{\text{densidad } (x_1, 3)}{\frac{\text{densidad } (x_2, 3) + \text{densidad } (x_5, 3) + \text{distancia } (x_3, 3)}{\text{cardinal } N(x_1, 3)}} = \frac{0.75}{\frac{0.5 + 0.75 + 0.5}{3}} = 1.29$
- P2: $\text{drm } (x_2, 3) = \frac{0.5}{\frac{0.75 + 0.75 + 0.5}{3}} = 0.75$
- P3: $\text{drm } (x_3, 3) = \frac{0.5}{\frac{0.75 + 0.75 + 0.5}{3}} = 0.75$
- P4: $\text{drm } (x_4, 3) = \frac{0.17}{\frac{0.5 + 0.75 + 0.75}{3}} = 0.26$
- P5: $\text{drm } (x_5, 3) = \frac{0.75}{\frac{0.75 + 0.5 + 0.5}{3}} = 1.29$

Datos Anómalos

Realizar el paso C del algoritmo LOF para la identificación de outliers del ejercicio anterior.

Obtención de los outliers.

Las drm que tenemos para los puntos son: P1: 1.29, P2: 0.75, P3: 0.75, P4: 0.26, P5: 1.29

Por lo que simplemente comparándolas se observa que la drm del punto 4, 0.26, es significativamente inferior que el resto. La media sería 0.86 y la mediana 0.75.

En consecuencia el único punto con una densidad significativamente inferior al resto de los puntos de la muestra es el Punto 4 (1, 1), por lo que debe ser considerado outlier.



Visualización de Datos

¿Qué es visualización?

Formar en la mente una imagen visual de un objeto abstracto.

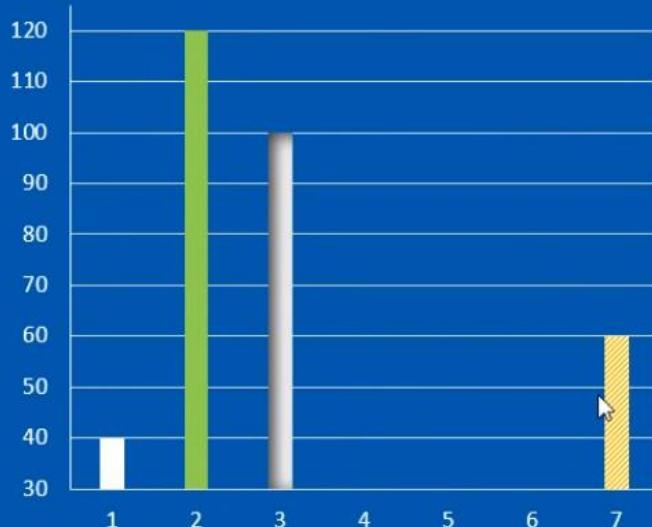
¿Cómo Podemos visualizar los datos?

Datos en bruto

Visualización de Datos

¿Cómo Podemos visualizar los datos?

Con Gráficos

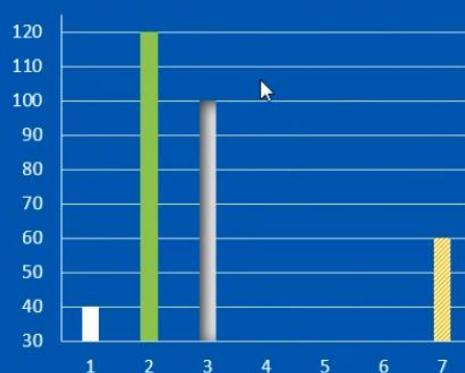


Los mismos datos que en la diapositiva previa

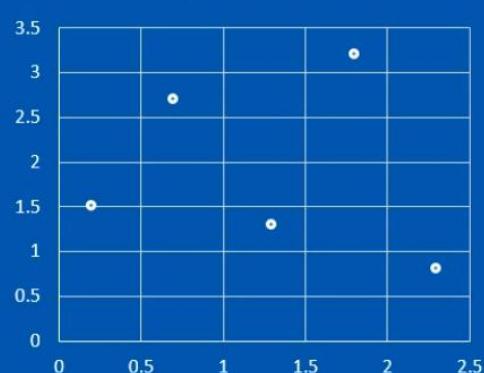
¿Qué es la visualización gráfica de datos?

Los gráficos de datos muestran gráficamente cantidades medidas (datos de características o sus relaciones) por medio del uso combinado de (alguno de) puntos, líneas, un sistema de coordenadas, números, símbolos, palabras, sombras y color. [Tufte, 1983]

Datos Observados = Frecuencia



Relaciones entre datos, observadas o descubiertas



VISUALIZACIÓN DE DATOS

¿Cuándo comenzó la visualización de datos?

Con la excepción de los mapas, la visualización gráfica de datos comenzó no hace mucho tiempo, entre 1750 y 1800. En consecuencia fue sobre 200 años atrás, con los trabajos de Playfair, Lambert, Minard y Marey.

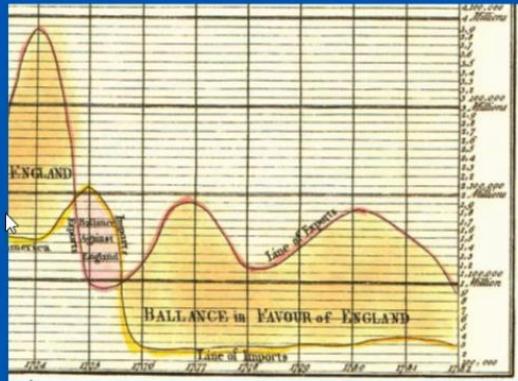
Juan J. Cuadrado Gallego

Mapa de los caminos de Yu, China



1100 A.C.

Importaciones/exportaciones
Norte America-Inglaterra



W. Playfair 1782

Gráficos de datos: Clasificación

Antes de ver que factores impactan en la calidad de los los gráficos de datos y cómo mejorar dicha calidad vamos a ver una nueva clasificación los gráficos de datos:

- Gráficos de Ejes
- Gráficos de Círculos
- Gráficos de Mapas
- Gráficos de Matrices
- Gráficos de Redes
- Gráficos de Dibujos

Vamos a ver los factores de calidad específicos para algunos gráficos de ejes y algunos factores generales.

También veremos como se puede realizar investigación sobre la calidad de las visualizaciones de datos.

Juan J. Cuadrado Gallego

Diagrama de Barras

Se presentan datos cualitativos y cuantitativos discretos.

Cada valor de la característica es representado mediante una barra cuyo altura es la frecuencia absoluta o relativa del valor.

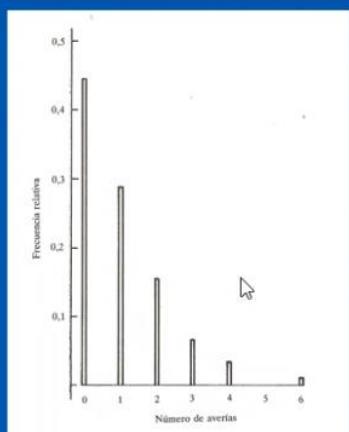
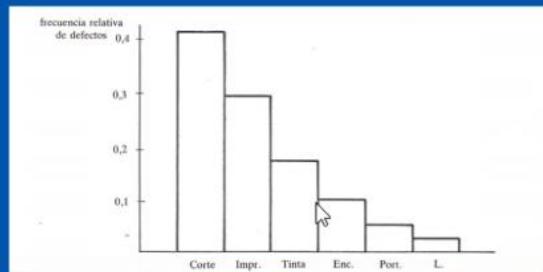


Diagrama de Pareto.

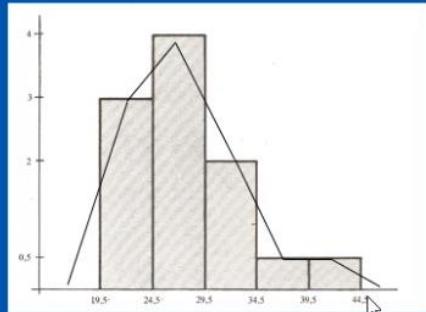
Las clases se ordenan de izquierda a derecha por el valor de su f_r



Histograma

Se representan datos continuos agrupados.

Cada rectángulo representa una clase cuya base es proporcional a la amplitud del intervalo



VISUALIZACION DE DATOS

Representar un histograma para los datos de los radios de los satélites menores de Urano:
13, 15, 16, 20, 20, 22, 27, 29, 30, 33, 34, 42

- b) Respuesta: 13, 15, 16, 20, 20, 22, 27, 29, 30, 33, 34, 42

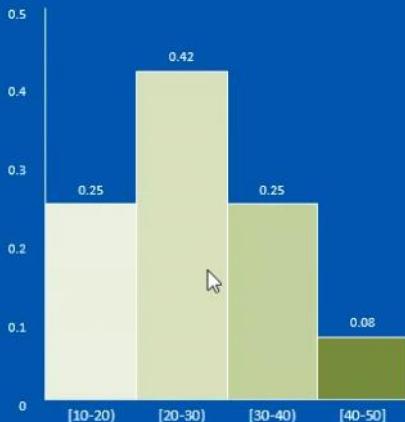


Diagrama de tallo y hojas de Tukey

Se representan datos cuantitativos. Es como un histograma cuyas barras son los datos.

Dibujar una línea vertical. El dígito de la clase (rama) se muestra a la izquierda y las unidades (hojas) a la derecha. La rama define la clase y las hojas sus frecuencias.

Representar, utilizando un diagrama de tallos y hojas, los datos de los radios de los stelites menores de Urano: 13, 15, 16, 20, 20, 22, 27, 29, 30, 33, 34, 42

Solución:

1	3	5	6
2	0	0	2
3	0	3	4
4	2		

Diagrama de Tukey de Caja y Bigotes.

Se representan datos cuantitativos continuos.

Se ordenan los datos y se obtienen los cuartiles.

Se dibuja un rectángulo que termina en Q1 y Q3.

La posición de la mediana Q2 es indicada con una línea. Los límites del intervalo son calculados para los outliers y se conectan con el rectángulo mediante una línea.



Es fácil identificar los outliers como esos puntos que están fuera de la línea.

Representar, utilizando un diagrama de caja y bigotes, los datos de los radios de los satélites menores de Urano: 13, 15, 16, 20, 20, 22, 27, 29, 30, 33, 34, 42

Respuesta:

$$Q_1=18 \quad Q_2=24,5 \quad Q_3=31,5$$

$$(Q_1 - 1,5(Q_3 - Q_1), Q_3 + 1,5(Q_3 - Q_1)) = (18 - 1,5(31,5 - 18), 31,5 + 1,5 \cdot 13,5)$$

$$(-2,25, 51,75) \equiv (0, 51,75)$$



Scatter Diagram y línea de regresión

Para pares de datos los gráficos permiten extraer más conocimiento que simplemente presentar la información, permitiendo razonar sobre ellos.

Si tenemos la regresión lineal de cuatro grupos de datos: 1. {10, 8.04; 8, 6.95; 13, 7.58; 9, 8.81; 11, 8.33; 14, 9.96; 6, 7.24; 4, 4.26; 12, 10.84; 7, 4.82; 5, 5.68}; 2. {10, 9.14; 8, 8.14; 13, 8.74; 9, 8.77; 11, 9.26; 14, 8.1; 6, 6.13; 4, 3.1; 12, 9.13; 7, 7.26; 5, 4.74}; 3. {10, 7.46; 8, 6.77; 13, 12.74; 9, 7.11; 11, 7.81; 14, 8.84; 6, 6.08; 4, 5.39; 12, 8.15; 7, 6.42; 5, 5.73}; 4. {8, 6.58; 8, 5.76; 8, 7.71; 8, 8.84; 8, 8.47; 8, 7.04; 8, 5.25; 19, 12.5; 8, 5.56; 8, 7.91; 8, 6.89}.

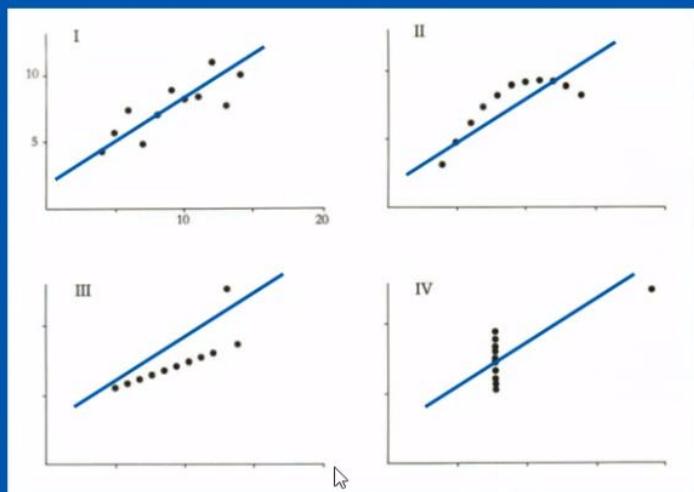
Todos los grupos dan el mismo modelo de regresión lineal, con la misma correlación:

$$\begin{aligned} n &= 11; \bar{x} = 9; \bar{y} = 7,5 \\ y &= 3 + 0,5x \\ r &= 0,82 \end{aligned}$$

No hemos incrementado mucho nuestro conocimiento sobre los cuatro grupos de datos. Vamos a ver que sucede cuando utilizamos un modelo gráfico como un diagrama de puntos.

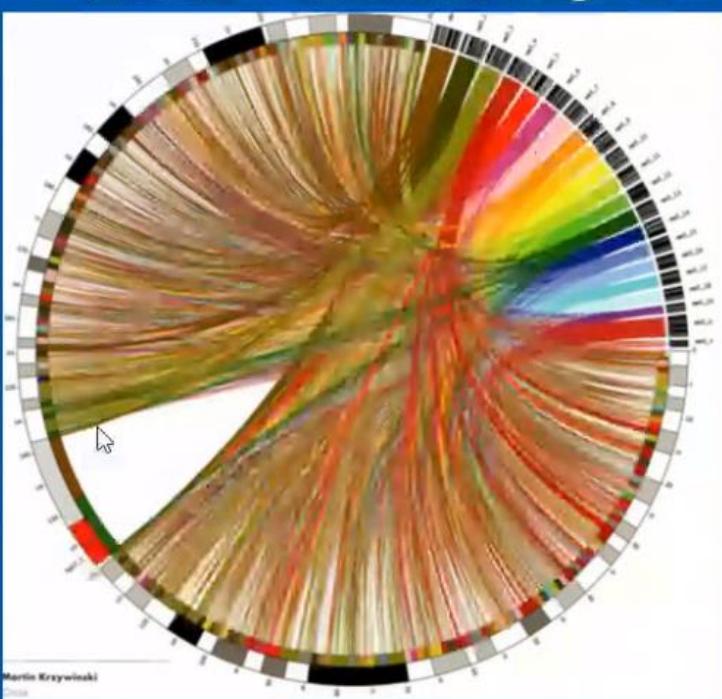
Scatter Diagram and Regression Line

Realizamos el scatter diagram y la regression line de los datos previos.



Los gráficos permiten mucho más que presentar información, permiten razonar sobre ella.

Relación de los cromosomas en el genoma



Visualización de Datos: Factores de calidad de los gráficos

¿Qué es calidad?

La adecuación de un producto o servicio a sus características específicas.

¿Cuáles son las características específicas de un gráfico?

Dar al observador una más fácil comprensión de la mayor cantidad de información posible, sobre datos observados o sobre relaciones en los datos observadas o descubiertas, en el menor tiempo posible.

¿Qué es la calidad de los gráficos de datos?

Conseguir dar al observador una comprensión más fácil de la mayor cantidad de información, sobre los datos observados o relaciones observadas o descubiertas en los datos, en el menor tiempo posible.

¿Cómo puede ser obtenida la calidad en los gráficos de datos?

Con tres principios básicos:

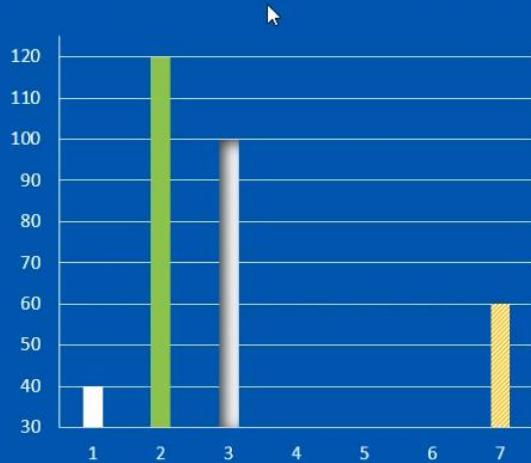
1. Mínimos elementos gráficos
2. Buen diseño y situación de los elementos.
3. Correspondencia exacta con la información textual



Gráficos de Ejes: Factores de Calidad

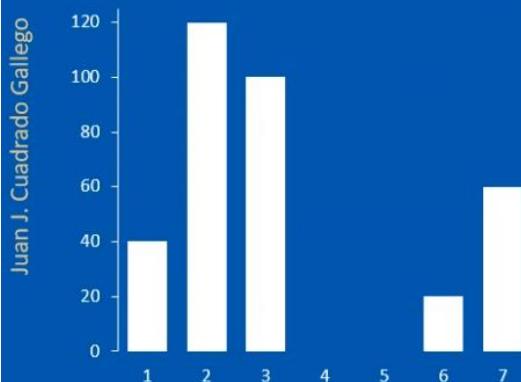
Recordar los principios: mínimos elementos, buen diseño, y datos verdaderos

Diagramas de Columnas/Barras. En un eje se representan los diferentes valores de los datos, y en el otro sus frecuencias, porcentajes, etc. Valores discretos.



Gráficos de Ejes: Factores de Calidad

Recordar los principios: mínimos elementos, buen diseño, y datos verdaderos



1. Barras demasiado estrechas y muy separadas (50%)

2. Colores, texturas o sombras que distraen

3. Línea base en 0

4. Elementos mínimos

5. Ejes con los nombres de variables y gráfico con nombre

Gráficos de Ejes: Factores de Calidad

Recordar los principios: mínimos elementos, buen diseño, y datos verdaderos



1. Cuando las características son nominales u ordinales quizás es mejor un gráfico de Pareto

2. Cuando los nombres de las características son muy largos quizás es mejor un diagrama de barras

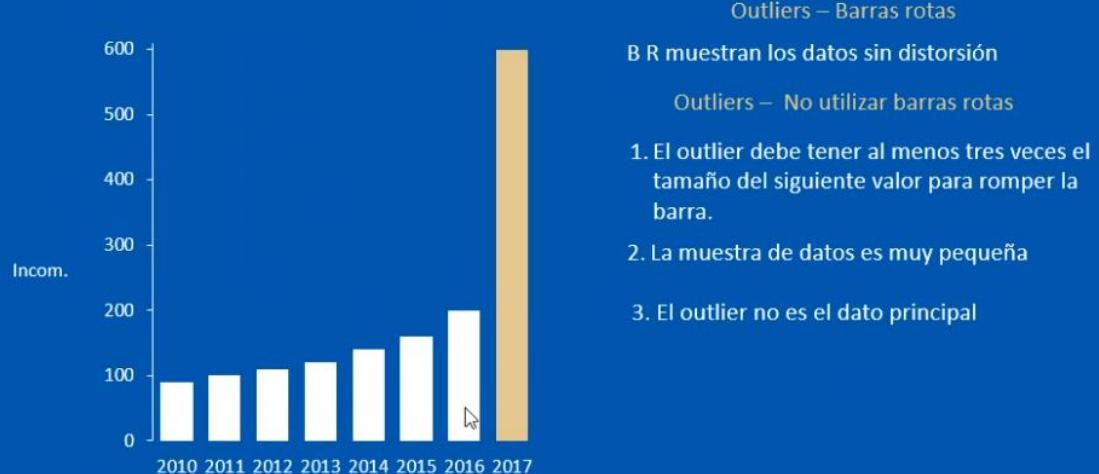
3. Algunas veces es necesario resaltar un valor específico

4. Las líneas verticales y de escala deben ser evitadas

5. Las fechas o los valores alfabéticos deben tener su propio orden

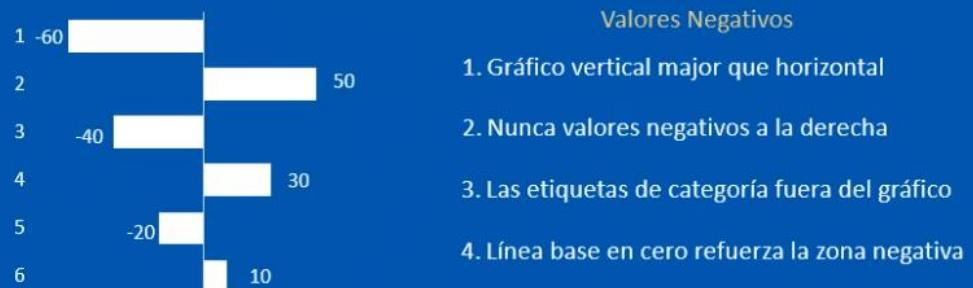
Gráficos de Ejes: Factores de Calidad

Recordar los principios: mínimos elementos, buen diseño, y datos verdaderos



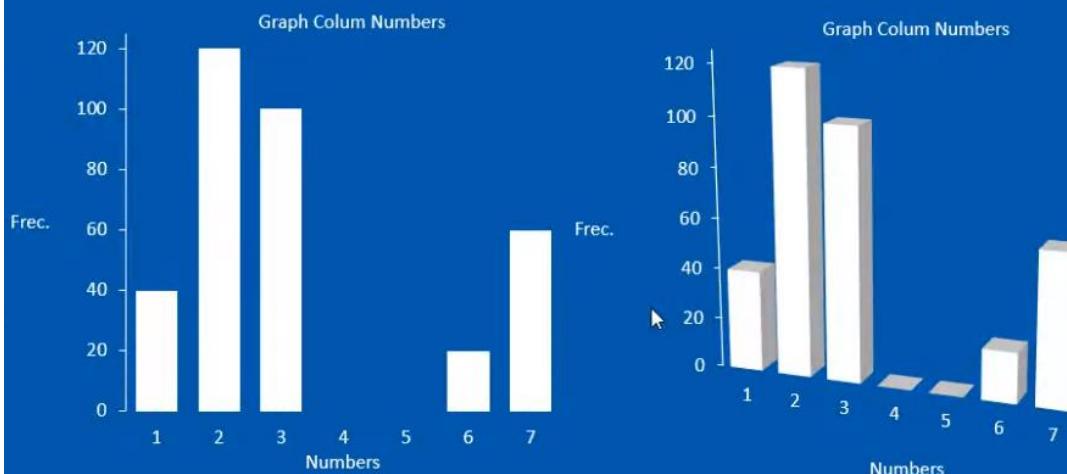
Gráficos de Ejes: Factores de Calidad

Recordar los principios: mínimos elementos, buen diseño, y datos verdaderos



Gráficos de Ejes: Factores de Calidad

Recordar los principios: mínimos elementos, buen diseño, y datos verdaderos

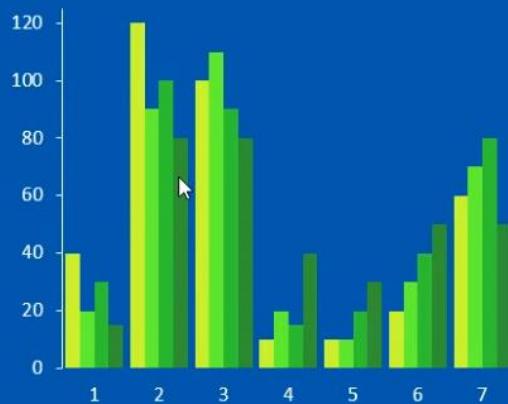


¿Es mejor en 3D? ¿Hay más información? ¿Más clara?

Gráficos de Ejes: Factores de Calidad

Recordar los principios: mínimos elementos, buen diseño, y datos verdaderos

Multiples barras – Colores y Leyendas



1. No alto contraste en las barras

2. Sombras de claro a oscuro

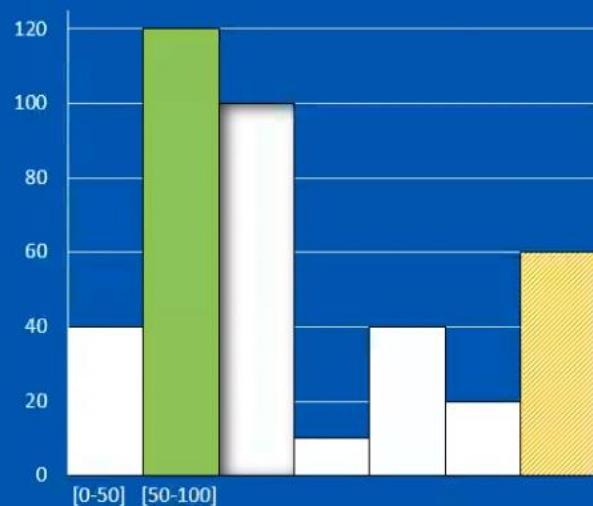
3. No más de cuatro o cinco categorías

3. Legendas en el gráfico y en el mismo orden

Gráficos de Ejes: Factores de Calidad

Recordar los principios: mínimos elementos, buen diseño, y datos verdaderos

Histograma. En un eje los valores, en el otro las frecuencias

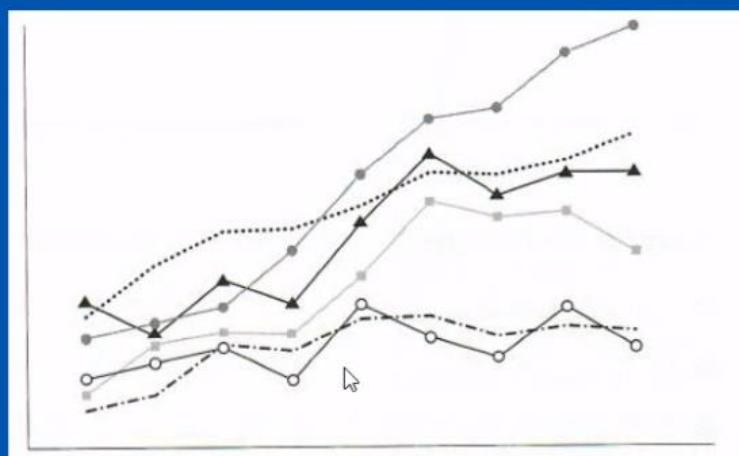


Se aplican todas las reglas de calidad del diagrama de barras

Gráficos de Ejes: Factores de Calidad

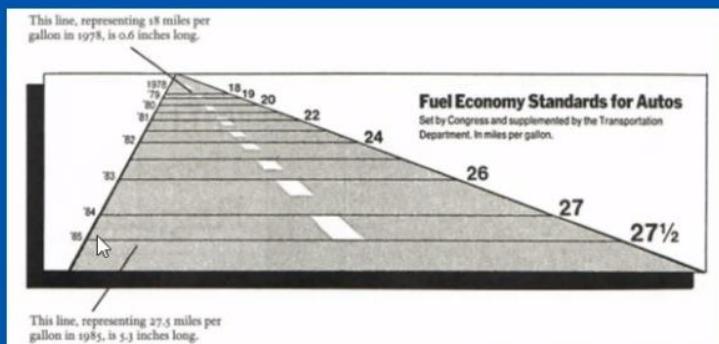
Recordar los principios: mínimos elementos, buen diseño, y datos verdaderos

Gráficos de línea. Para valores continuos



No spaghetti graficos. Cuatro líneas máximo

¿Cómo se puede desarrollar investigación sobre la calidad de las visualizaciones de datos?



$$\text{size of effect show in data} = \frac{27.5 - 18}{18} \times 100 = 53\%$$

$$\text{size of effect in graphic} = \frac{5.3 - 0.6}{0.6} \times 100 = 783\%$$

¿Cómo se puede desarrollar investigación sobre la calidad de las visualizaciones de datos?

Distorsión gráfica de los datos [En desarrollo]

$$\text{Data Graphical Distortion} = 1 - \frac{\text{Data increase}}{\text{Graphic increase}}$$

Lie Factor:

No normalizado. Valores entre 1 (Verdad) y ∞ (Maxima mentira)

Distorsión gráfica de los datos :

Normalizada. Valores entre 0 (sin distorsión) y 1 (Máxima distorsión) ($\times 100$ in percentage)

Ejemplo: $\text{data graphical distortion} = \frac{53}{783} = 0.06$ (1-0.06=0.94 of distorsión)

94% of distorsión

¿Cómo se puede desarrollar investigación sobre la calidad de las visualizaciones de datos?

Ocultación gráfica de los Datos [En desarrollo]

$$Data\ Graphical\ Occultation = 1 - \frac{Data\ elements}{Graphic\ elements}$$

Data Ink Ratio:

Normalizada. Marco y elementos de descripción de los datos no son considerados.

Data Graphical Occultation:

Normalized. Marco y elementos de descripción de los datos son considerados. Valores entre 0 (without occultation) y 1 (Maximum occultation)

Otra cuestión es graphical completeness, si todos los datos están en el gráfico
Para medir eso se debe definir otra variable...