PRIMERA PARTE

CORRELACIÓN

Pequeño debate inicial:

Podrían normbrar RELACIONES de la “vida cotidiana”?

Podrían nombrar relaciones directas?
Podrían nombrar relaciones indirectas?

Podrían mencionar algun caso especial en el cual haya una clara RELACION pero la CORRELACION fallaría)

Desde el punto de vista Predictivo, que nos indica exactamente la **CORRELACION**?

Data Mining

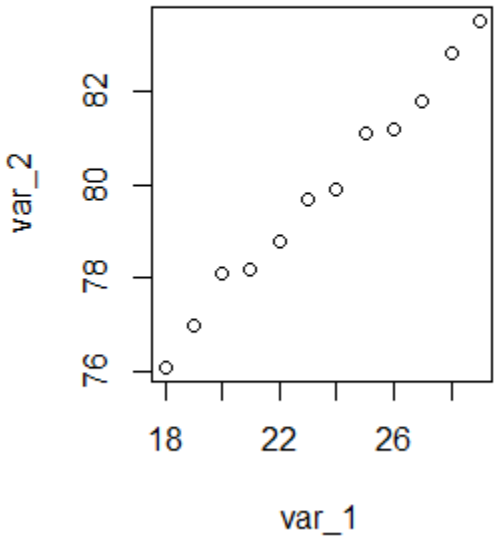
Para realizar este ejercicio se utilizarán los datos del archivo “Correlacion.csv”
Para ello, es posible tanto importar dicha planilla, o cargar los datos manualmente de las seis variables como vectores.

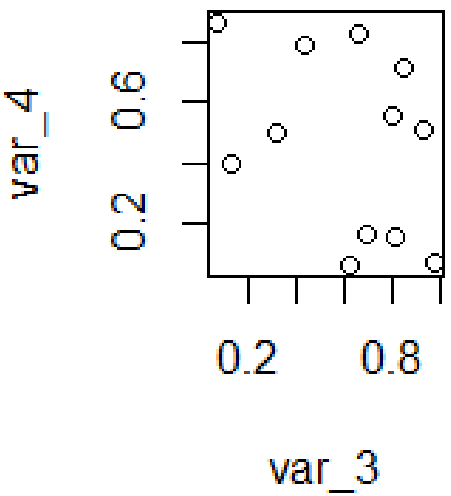
Nosotros la importaremos:

```
data_cor <- read.csv("/Data/DATASETS/Correlacion.csv",header=T,sep=",")
var_1 <- as.numeric(data_cor$Var_1)
var_2 <- as.numeric(data_cor$Var_2)
var_3 <- as.numeric(data_cor$Var_3)
var_4 <- as.numeric(data_cor$Var_4)
var_5 <- as.numeric(data_cor$Var_5)
var_6 <- as.numeric(data_cor$Var_6)
var_7 <- as.numeric(data_cor$Var_7)
```

La primera aproximación que podemos realizar para ver si existe una correlación entre un par de variables, es visual, con el comando ya conocido plot

```
> plot(var_1,var_2)
> plot(var_3,var_4)
```





Podríamos ahora realizar un análisis analítico buscando los índices de correlación.

Recordemos que los índices de correlación cercanos a 1 muestran una **correlación directa**, y aquellos cercanos a -1 una **correlación indirecta**.

Se puede deducir aquí que una correlación cercana a cero, muestra una **correlación aleatoria**.

```
>corre_v1_v2 <- cor(var_1, var_2)
> corre_v1_v2
>[1] 0.9943661
```

Este resultado nos muestra que var_1 y var_2 se encuentran fuertemente correlacionadas de manera directa.

```
> corre_v3_v4 <- cor(var_3, var_4)
> corre_v3_v4
[1] -0.3781641
```

Este resultado nos muestra que entre var_3 y var_4 no existe prácticamente ningún tipo de correlación, es casi aleatoria.

Data Mining

Explorar una por una las variables de un set de datos de gran volumen puede ser muy tedioso y muy poco práctico. Existe sets de datos que superan las 100.000 variables ampliamente.

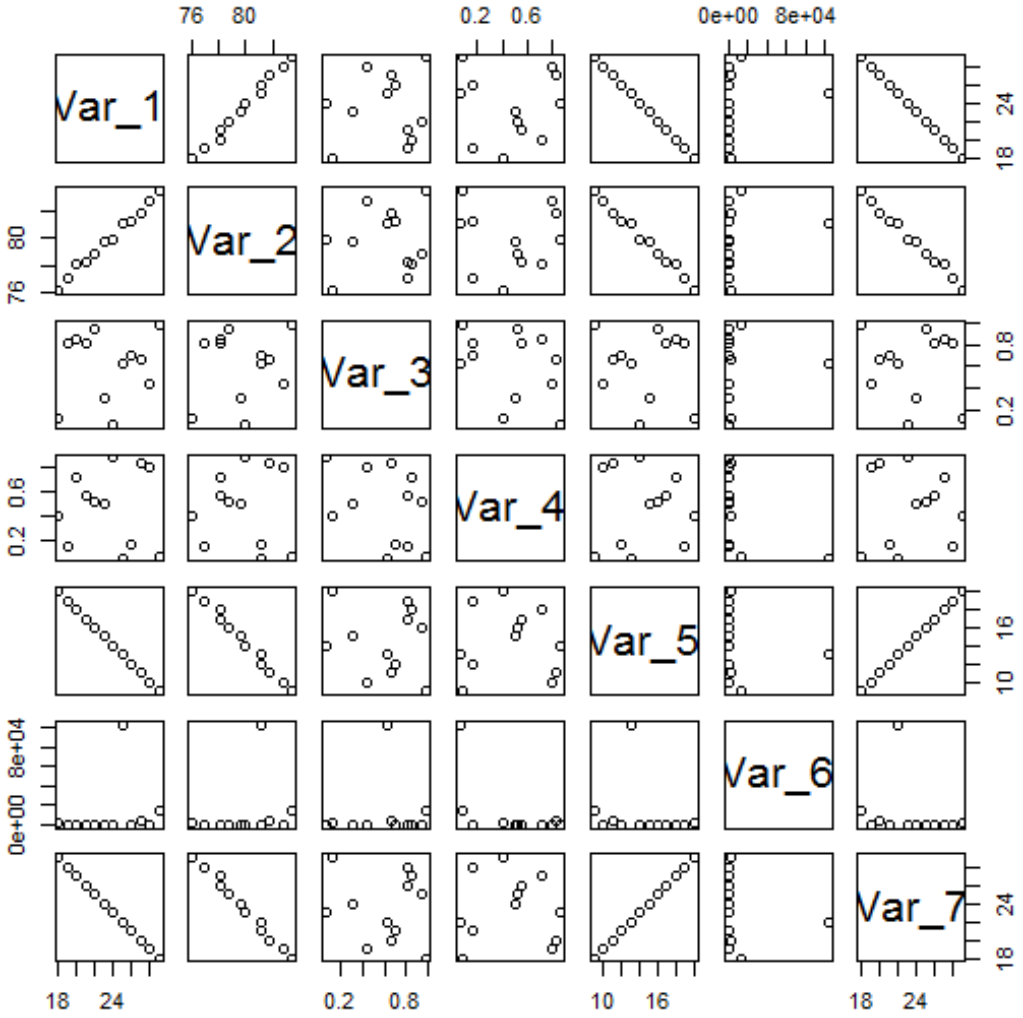
Para ello R nos presenta dos funciones sumamente útiles:

1. Matriz de Correlación Gráfica

```
>pairs(data_cor)
```

Nos muestra una matriz gráfica de correlación de TODAS las variables

Notar que la función “pairs” se corre contra el set de datos completo.

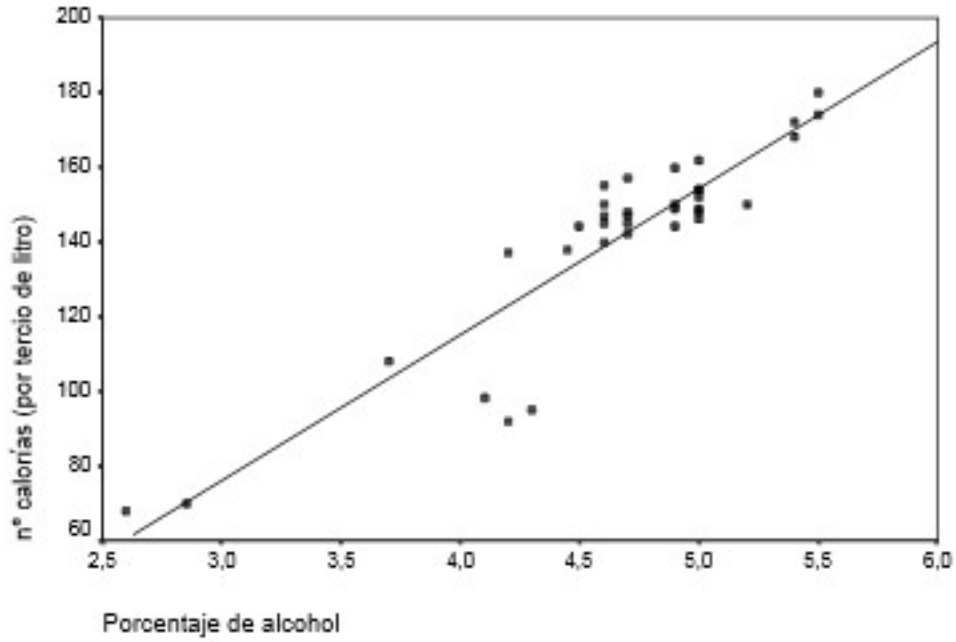


SEGUNDA PARTE

REPASO

REGRESIÓN LINEAL MÚLTIPLE

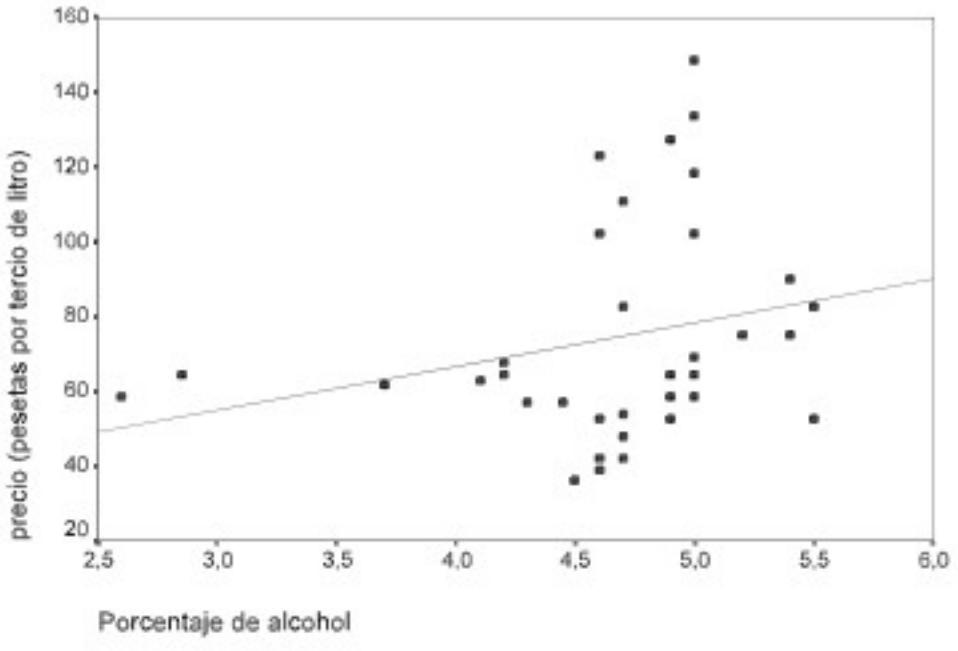
Regresión Lineal Simple: Repaso de Conceptos



$Y_i = -33{,}77 + 37{,}65\, X_i$

nº de calorías = - 33,77 + 37,65 (% de alcohol)

R2 = 0.83



precio = 20,16 + 11,61 (% de alcohol)

$R^2 = 0{,}06$

¿Como podemos leer el R^2 en los dos gráficos anteriores?

¿Como podemos leer el R^2 en los dos gráficos anteriores?

En el primero de los gráficos, podemos afirmar que si conocemos el porcentaje de alcohol en una cerveza, podemos inferir la cantidad de calorías que ésta contiene con un 83% de certeza.

¿Como podemos leer el R^2 en los dos gráficos anteriores?

En el primero de los gráficos, podemos afirmar que si conocemos el porcentaje de alcohol en una cerveza, podemos inferir la cantidad de calorías que ésta contiene con un 83% de certeza.

En el segundo de los gráficos, podemos afirmar que si conocemos el porcentaje de alcohol en una cerveza, podemos inferir el precio de la misma, solo con un 6% de certeza.

¿Como podemos leer el R^2 en los dos gráficos anteriores?

En el primero de los gráficos, podemos afirmar que si conocemos el porcentaje de alcohol en una cerveza, podemos inferir la cantidad de calorías que ésta contiene con un 83% de certeza.

En el segundo de los gráficos, podemos afirmar que si conocemos el porcentaje de alcohol en una cerveza, podemos inferir el precio de la misma, solo con un 6% de certeza.

Modelo de Regresión Lineal Simple

$$Y_i = B_0 + B_1 X_i$$

Modelo de Regresión Lineal Múltiple

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Pruebas de Significación

De manera simplificada, diremos que estas pruebas sirven para contrastar la hipótesis nula de que un **coeficiente de regresión vale cero** en la población. Niveles menores a 0.05 indican que debemos rechazar dicha hipótesis.

Recordemos, que un coeficiente de regresión igual a cero, indica ausencia total de relación lineal. Por lo tanto un "*Sig*" muy pequeño, rechazará esa hipótesis y nos indicará que **variables son relevantes para generar el modelo.**

Trabajar con el dataset: "Regresion Multiple"

En este ejemplo, Vemos que la variable "GRADUADOS" es la más significativa de todas, al igual que "DESEMPLEO", Mientras que "INGRESOS" y "PERIODO" no lo son.

```
data_reg <- read.csv("/Data/DATASETS/Regresion_Multiple.csv", header=T,sep=",")

model_reg <- lm(ENROL ~ DESEMPLEO + GRADUADOS + INGRESOS + PERIODO, data_reg)

summary(model_reg)
```

Qué utilidad importante encontramos a este modelo más allá de su valor PREDICTIVO?

Cuando nos conviene usar solo las variables relevantes y dejar de lado las que no lo son?

TERCERA PARTE

DISTRIBUCIÓN DE FRECUENCIAS

En primer lugar vamos a presentar una técnica sumamente útil para diferentes procedimientos en el Análisis de Datos, que es la generación de números aleatorios.

Los números aleatorios nos permitirán realizar dos tareas fundamentales:

a. Simulaciones
b. Contrastar resultados y compararlos.

El comando de R “sample” nos permitirá generar números aleatorios de manera muy sencilla:

```
> ?sample
```

En primer lugar vamos a presentar una técnica sumamente útil para diferentes procedimientos en el Análisis de Datos, que es la generación de números aleatorios.

Los números aleatorios nos permitirán realizar dos tareas fundamentales:

a. Simulaciones
b. Contrastar resultados y compararlos.

El comando de R “sample” nos permitirá generar números aleatorios de manera muy sencilla:

```
> ?sample
```

Por ejemplo podemos simular arrojar una moneda al aire 10 vecesy ver si es “cara” o “ceca”

```
> monedas <-  sample(0:1,10,replace=T)
```

Ej 1. Puede describir esta línea de comando que hace exactamente?

Ej 2. Repita la misma línea de comando al menos 5 veces más. ¿Qué conclusiones puede sacar?

Otros ejemplos interesantes, que nos ayudarán en futuros temas de probabilidades:

Tire un dado 20 veces:

```
> dados <- sample(1:6,20,replace=T)
```

Ej: Puede generar ahora números aleatorios para jugar al Loto el día de mañana?
Dato: Recordar que el Loto varía entre los números 0 y 49 (recordar que para el Loto hay que sacar 6 numeros diferentes y no pueden estar repetidos)

CUARTA PARTE

PRE-PROCESAMIENTO DE DATOS

Distribución de Frecuencias:

Vamos a graficar ahora nuestra distribución de frecuencias del experimento de arrojar una moneda al aire 10 veces:

```
> plot(table(monedas))
```

Ej: Volver a realizar este mismo experimento, pero en lugar de arrojar al aire una moneda 10 veces, hacerlo 100 veces, luego 1.000 veces, 10.0000 y por último 1.000.000

¿Qué conclusiones puede sacar?

Pre-Procesamiento de Datos

El Pre-Procesamiento de Datos es una etapa crítica en el ciclo de un proyecto de Data Mining.

Suele ser considerada la etapa que más tiempo requiere, llegando a representar el 70% delesfuerzo total.

Importante: Estas tareas son de índole estadística, y debe realizarlas el analista en Data Mining.

Importante: No se debe confundir con el proceso de Data Warehouse.

Pre-Procesamiento de Datos

Las tareas más importantes de Pre-Procesamiento de datos son:

1**.Detectar errores de carga:** Ejemplos clásicos son:

- Ingresos: -100
- Género: 4 (Codificando 1 para Hombres y 2 para Mujeres)
- Género: Masculino, Embarazado: SI

2. **Datos vacíos:** Muchos algoritmos pueden soportar datos vacíos sin problema, pero otros no.
Más allá de soportarlos o no, un dato vacío puede quitar mucha información valiosa y generar ruido.

IMPORTANTE: Considerar como se los trata: NA / 0 / Remover / Reemplazar por valor más usado / reemplazar por valor medio.

3. **Outliers:** Dependiendo el tipo de estudio que estamos llevando a cabo, el tratamiento de outliers es crítico.

4**. Normalización:** Reducción de escala para crear un rango más pequeño. Considerar outliers antes.

5**. Selección de Variables:** Uno de los pasos más interesantes y laboriosos a nivel matemático; estadístico.

6. **Data Transformation**: Crear nuevas variables a partir de data existente.
Ej: Tiempo de permanencia = Fecha de Ingreso – Fecha de Baja
Malware hex extraction.

Data Consistente:

Data en la cual valores nulos, valores especiales, errores obvios y outliers son corregidos o eliminados .

Missing Values (NA)

NA ← Not Available

```
age <- c(23, 16, NA)
mean(age)
is.na(age)
mean(age, na.rm = TRUE)
```

El valor no se encuentra presente, aunque su clase puede ser conocida sin problema

```
person_1 <- c(21,6.0)
person_2 <- c(42,5.9)
person_3 <- c(0,5.7)
person_4 <- c(21,NA)
persons <- rbind(person_1,person_2,person_3,person_4)
persons
         [,1] [,2]
person_1   21  6.0
person_2   42  5.9
person_3   0  5.7
person_4   21   NA
```

```
         [,1] [,2]
person_1   21  6.0
person_2   42  5.9
person_3   0  5.7
person_4   21   NA

complete.cases(person)
## [1] TRUE TRUE TRUE FALSE

person_new <- as.data.frame(na.omit(persons))
```

TENER SIEMPRE CUIDADO CUANDO UN VALOR ES CERO. CHEQUEAR QUE LOS CEROS SEAN BIEN RECONOCIDOS EN R.

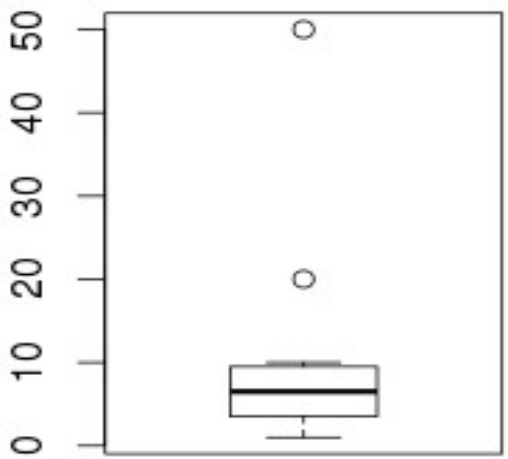
OUTLIERS

Los outliers no son necesariamente errores.
Hay que decidir con criterio si quitarlos o no.

```
x <- c(1:10, 20, 50)
boxplot.stats(x)$out
[1] 20 50

x_no_outliers <- x[!x %in% boxplot.stats(x)$out]

boxplot(x)
```

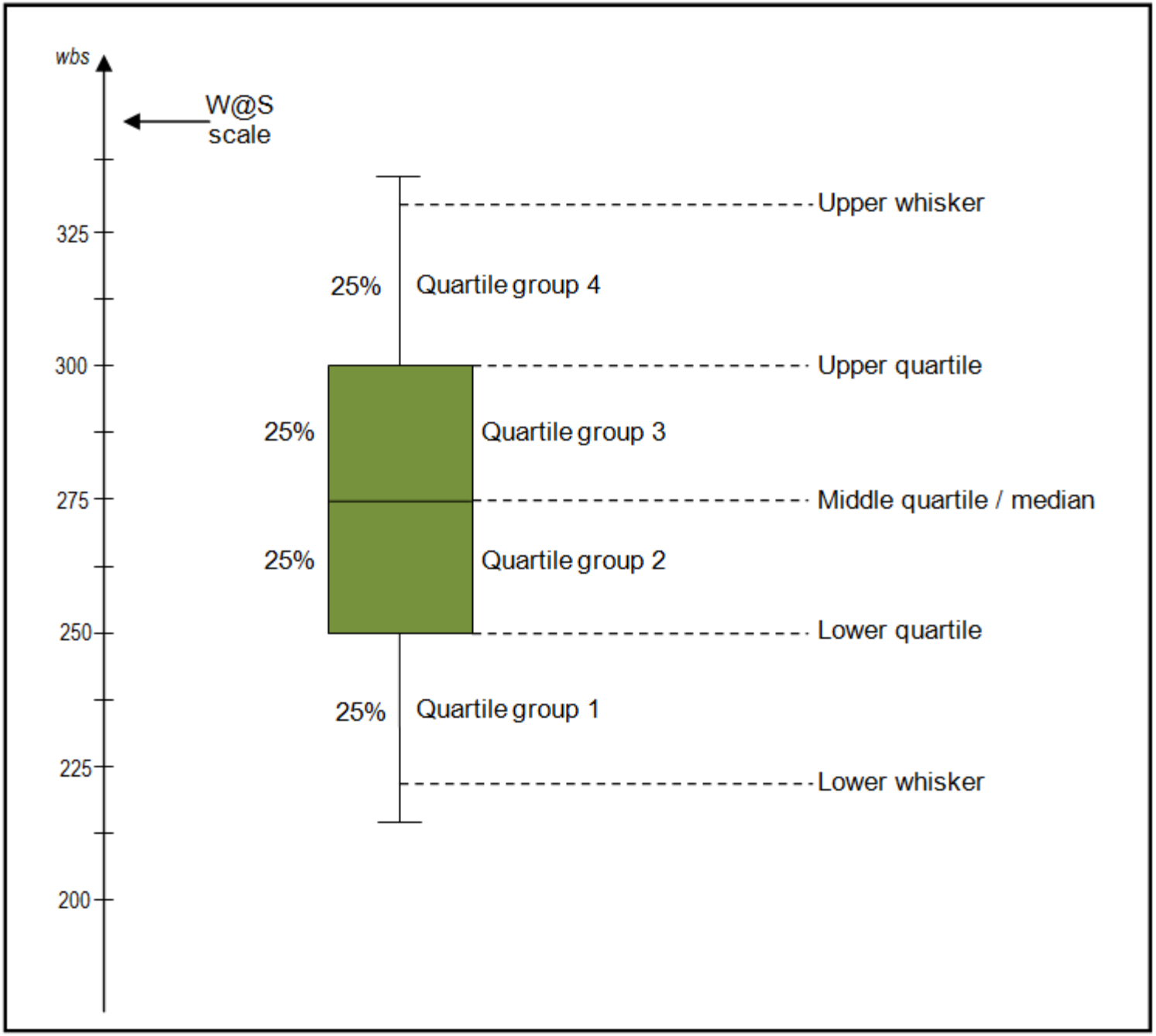


Outlier =
Cualquier valor que se ubica mas alla de una vez y media la longitud de la caja, tanto por encima como por debajo.

Por debajo:
Q1 – 1.5 × IQR

Por arriba:
Q3 + 1.5 × IQR

Interquartil = Q3 – Q1
(tamaño de la caja)


wbs
W@S
scale
325
300
275
250
225
200
Upper whisker
25%
Quartile group 4
Upper quartile
25%
Quartile group 3
Middle quartile / median
25%
Quartile group 2
Lower quartile
25%
Quartile group 1
Lower whisker

Este data-set tiene outliers?

10.2, 14.1, 14.4. 14.4, 14.4, 14.5, 14.5, 14.6, 14.7, 14.7, 14.7, 14.9, 15.1, 15.9, 16.4

1. Busco la Media.
El set de datos, tiene 15 puntos, por lo tanto: (15+1)/2 = 8
Entonces Q2 = 14.6

2. Busco Q1 y Q3
Hay 7 valores por debajo de Q2 y 7 por encima
Q1 es el cuarto valor y Q3 es el decimosegundo
Q1 = 14.4
Q3 = 14.9

3. Calculo IQR = 14.9 – 14.4 = 0.5

4. Calculo estadisticos
Outliers below: Q1 – 1.5x IQR
Outliers above: Q3 + 1.5x IQR

14.4 - (1.5 x 0.5) = 13.65
14.9 + (1.5 x 0.5) = 15.65

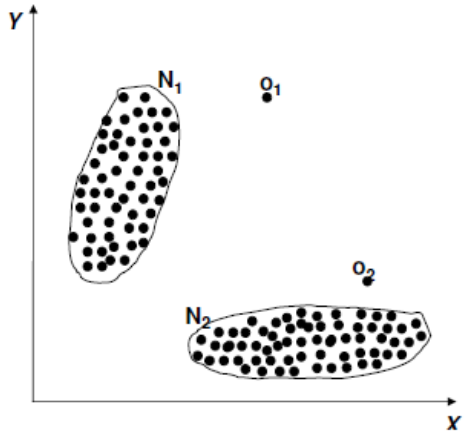
5. Busco Outliers

Menos a 13.65 y Mayores a 15.65: “10.2” , “15.9” , “16.4”

Anomalías - Primera clasificación

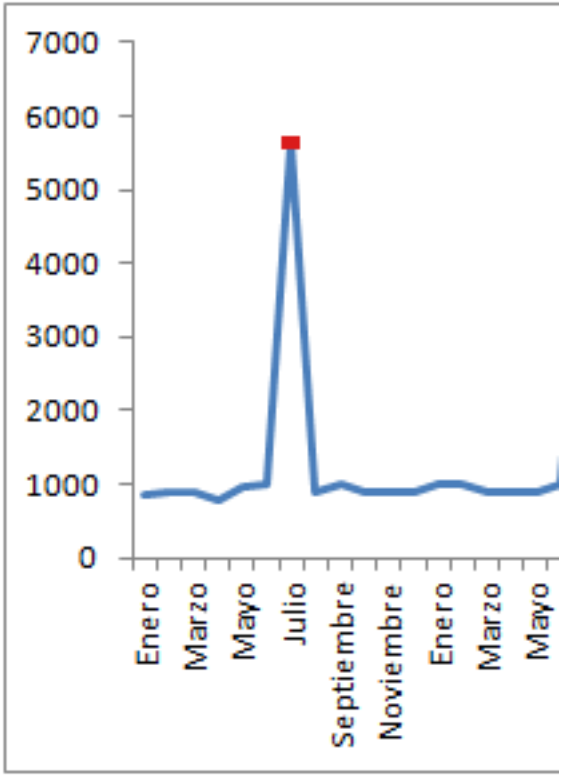
Tipos de Anomalías

1. Anomalías de Punto
2. Se crean nubes de comportamiento “normales” y se estudian desviaciones de la misma. A esas anomalías se las conoce como “outliers”



2. Anomalías Contextuales

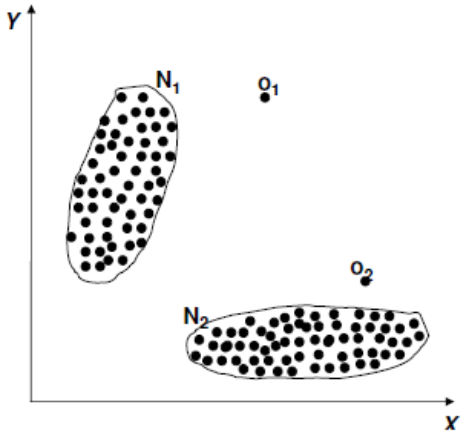
Se debe estudiar un patrón de contexto respecto a los outliers encontrados



Anomalías - Primera clasificación

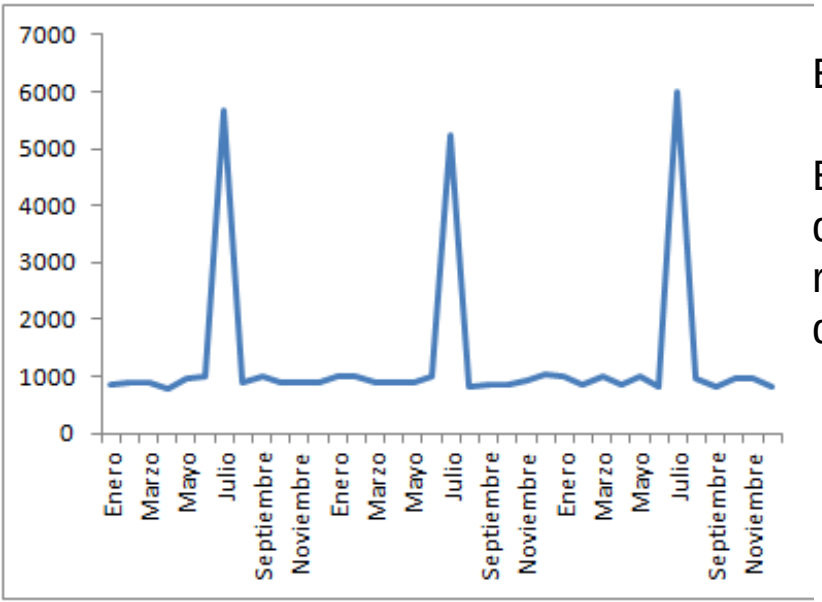
Tipos de Anomalías

1. Anomalías de Punto
2. Se crean nubes de comportamiento “normales” y se estudian desviaciones de la misma. A esas anomalías se las conoce como “outliers”

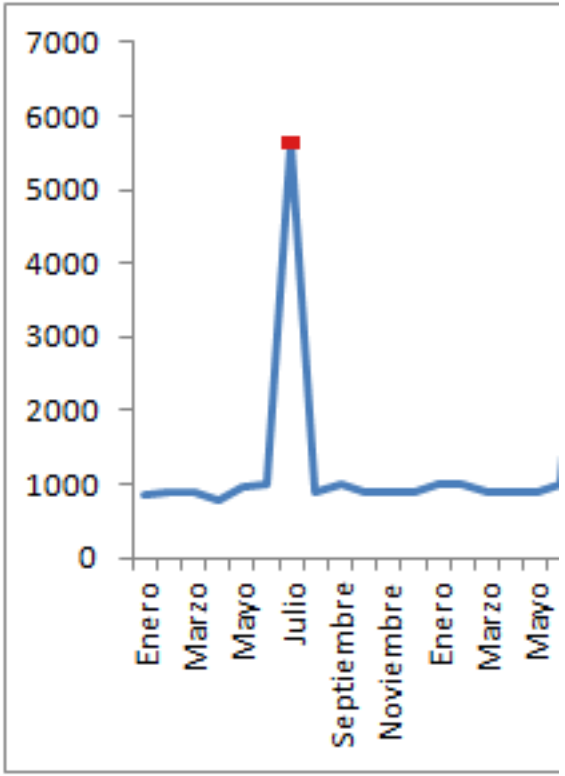


2. Anomalías Contextuales

Se debe estudiar un patrón de contexto respecto a los outliers encontrados



El punto en cuestión, NO era una anomalía. En la figura de la derecha, el punto rojo se encuentra descontextualizado.



1.Trabajando con outliers

2.data(rivers)
3.length(rivers)
4.hist(rivers)
5.boxplot(rivers)
6.
```
boxplot.stats(rivers)$out
rivers_no_outliers <- rivers[!rivers %in% boxplot.stats(rivers) $out]
boxplot(rivers_no_outliers)
```

7.
```
new_rivers_1 <- rivers[rivers < 1250]
   boxplot(new_rivers_1)

   new_rivers_2 <- rivers[rivers < 900]
   boxplot(new_rivers_2)
   hist(new_rivers_2)
```

Edit Rules

```
library(editrules)

people <- read.csv("/Data/data.csv")

(E <- editset(c("age >=0", "age <=
150")))
```

```
record  num1  num2
     1 FALSE FALSE
     2 FALSE FALSE
     3 FALSE FALSE
     4 FALSE  TRUE
     5 FALSE FALSE
```

Pre-Procesamiento de Datos

Normalización:

Distancia Euclídeana:

$$\sqrt{(x1-x2)^2+(y1-y2)^2}$$

$$\sqrt{(a1-a2)^2+(b1-b2)^2+(c1-c2)^2+(d1-d2)^2+....}$$

Ejemplo:

La diferencia de salario entre el empleado 1 y el 4, es de 2000.

La diferencia Euclideana en 3D entre el punto 1 y el punto 4, es de 2000.0005

El salario, es lo que más se considera dentro del set de datos.

Empid	Salary	Age	Experience
1	25000	24	4
2	40000	27	5
3	55000	32	7
4	27000	25	5
5	53000	30	5

$$= \sqrt{(25000-40000)^2+(24-27)^2+(4-5)^2}$$
$$= 15000.00033333333$$

	1	2	3	4	5
1	0.0000000	15000.0003333	30000.0012167	2000.0005000	28000.0006607
2	15000.0003333	0.0000000	15000.0009667	13000.0001538	13000.0003462
3	30000.0012167	15000.0009667	0.0000000	28000.0009464	2000.0020000
4	2000.0005000	13000.0001538	28000.0009464	0.0000000	26000.0004808
5	28000.0006607	13000.0003462	2000.0020000	26000.0004808	0.0000000

Max – Min Normalization

$$B = \left(\frac{(A - \text{minimum value of } A)}{(\text{maximum value of } A - \text{minimum value of } A)}\right) * (D - C) + C$$

$$= \left(\frac{(50000 - 25000)}{(55000 - 25000)}\right) * (1 - 0) + 0$$

$$= 0.83333333333333337$$

Transforma el valor de A en B, comprendido en el rango [C,D]

Nota: Muchos suelen sumarle 1 a los valores mínimos y máximos para evitar los 0, aunque este paso no debería ser crítico.

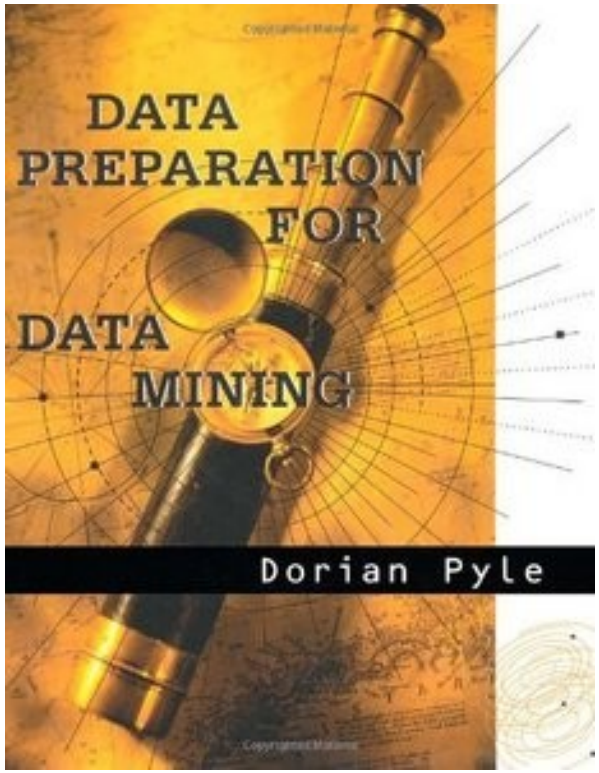
La Normalización por Rangos, no es la única, hay otras. Por ejemplo la z-normalization.

```
zVar <- (b - mean(b)) /
sd(b)
```

Bibliografía

Data Preprocessing for Supervised Leaning

S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas



Ejemplo de una importante Regla de Asociación en el área de la Salud

"Aquellos afiliados con registro de enfermedad hipertensiva por más de 13 meses, que consultaron 28 veces o más a médicos especialistas durante el 2008 y cuya facturación en prácticas de laboratorio fue alta y en productos de farmacia baja constituyen un grupo de riesgo para el año próximo"

Una hipótesis posible es que este nodo resulte de pacientes hipertensos insuficientemente tratados, ya sea en cantidad de dosis o cantidad de fármacos, ya sea por insuficiente prescripción o por falta de adherencia al tratamiento.