

# Predict Customer's preferred Brand

Presentación: Luis Sánchez Peña  
[sanchezlsp@gmail.com](mailto:sanchezlsp@gmail.com)

Código fuente y presentación:  
<https://github.com/meetsCode/Mod2T3Luis>

25 de Septiembre de 2017

# Customer's preferred Brand

## Introducción

Blackwell Electronics'CTO Danielle Sherman quiere que completemos la encuesta hecha por el departamento de marketing.

La encuesta busca conocer la marca preferida por los clientes pero parte quedó incompleta. Se desea conocer la marca preferida de todos los clientes encuestados, incluyendo los incompletos.

Debe entregarse un informe con los métodos usados y las salidas devueltas por R.

# Customer's preferred Brand

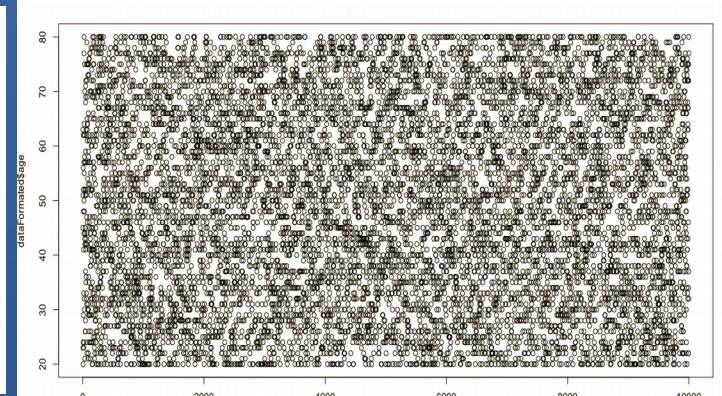
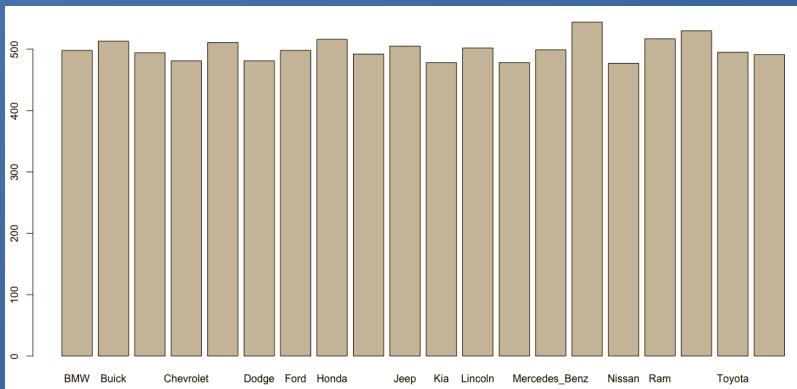
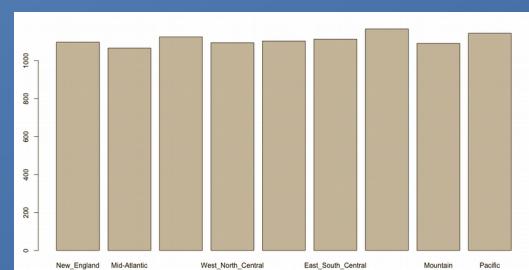
## Fuente de datos

- La empresa entrega dos ficheros resultantes de la encuesta:
  - Un CSV con 10.000 instancias completas: tabla1.
  - Una hoja de cálculo tipo Excel con 5.000 instancias defectuosas (falta el campo “Brand”): tabla2.
- En uno de los documentos tenemos las leyendas de cada campo.

# Customer's preferred Brand

## Fuente de datos

- El fichero completo parece ser una encuesta perfecta. Tiene las instancias perfectamente distribuidas en todos los campos.
- ¡El fichero incompleto también tiene todas las instancias perfectamente distribuidas! ¿-?
- No se observa correlación directa entre datos numéricos.
- ¿Es posible una encuesta perfecta?



# Customer's preferred Brand

## Análisis realizados

Buscamos el mejor modelo para completar la tabla2.

Comparé los métodos knn y Random Forest (RF) ambas con distintas semillas estadísticas y cross-validation (10)

Los resultados fueron:

modelo	semilla	k	Acc train	Kappa train	Acc test	Kappa test
KNN998	998	15	0.5989332	0.0653966	0.6166467	0.1063153
KNN1234	1234	19	0.6004803	0.0603191	0.6030412	0.0613679
RF1234	998		0.9224516	0.8352687	0.9231693	0.8373113
RF998	1234		0.9224898	0.8352859	0.5166066	-0.024852

# Customer's preferred Brand

## Reflexión

- ¿Tiene sentido usar RF para completar una tabla y luego inferir relaciones?
  - Si la tabla 1 ya da buenos resultados ¿Para qué más datos?
  - Si la tabla1 no aproxima bien.¿Por qué usarla? Solo amplificará el error al llenar la tabla2.

¡Mejor volver a la pregunta!

¿Qué marca quieren nuestros clientes?

# Customer's preferred Brand

## Reflexión

¿Qué marca quieren nuestros clientes?

Respuesta: La que más compran.

# Customer's preferred Brand

## Reflexión

Los datos que necesito para responder a esa pregunta son los de ventas.

# Customer's preferred Brand

## 2<sup>a</sup> Reflexión

¿Para qué sirve la encuesta?

Para saber quiénes son nuestros  
clientes.

# Customer's preferred Brand

Quiénes compran qué

Para poder identificar a nuestros clientes he hecho dos análisis:

- Árbol de decisión (J48 con WEKA y C50 con R).
- Análisis visual.

# Customer's preferred Brand

## Quiénes compran qué

Decission tree J48 en WEKA:

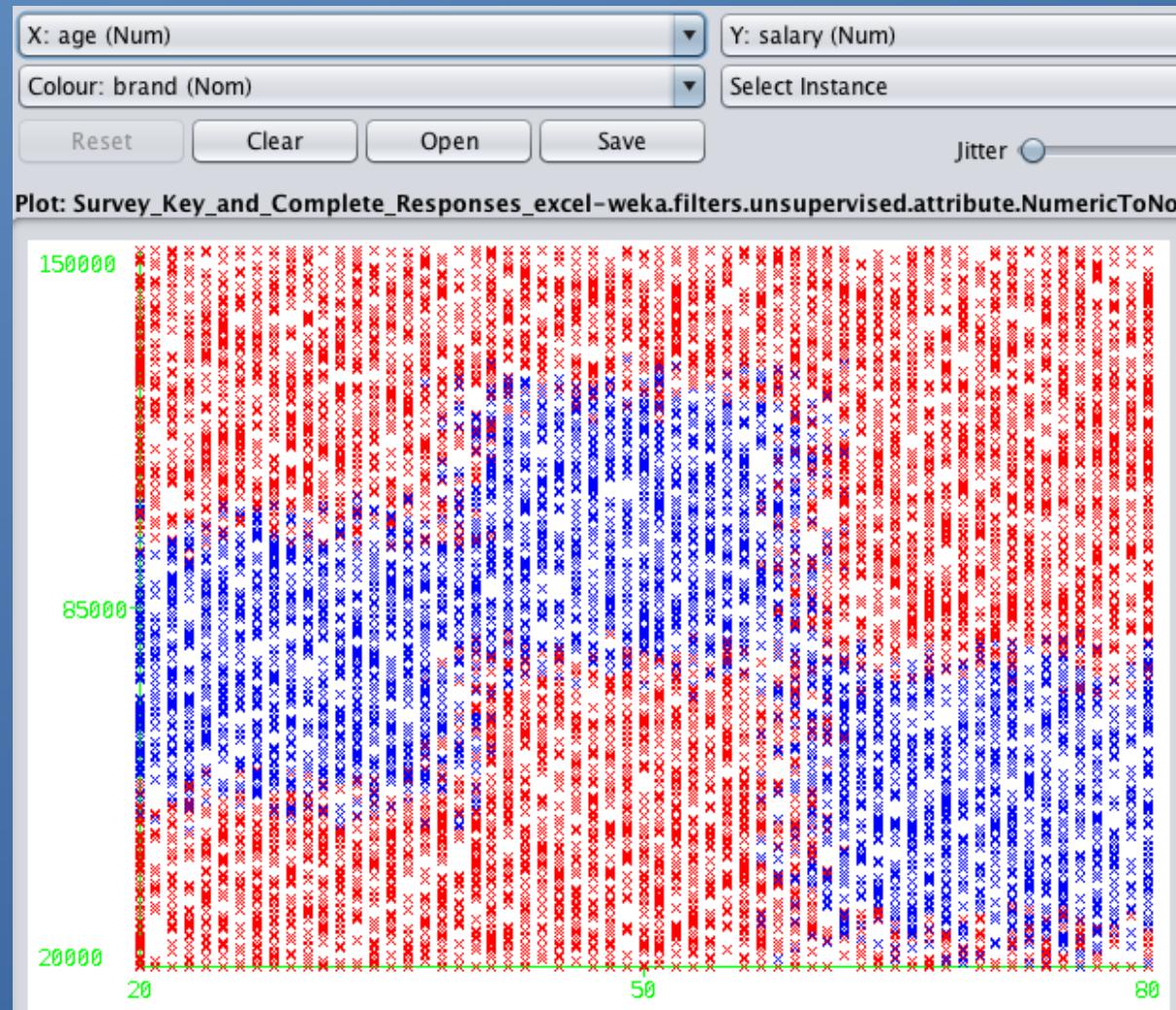
```
salary <= 125058.22
|   salary <= 45570.48
|   |   age <= 59: 1 (1340.0/23.0)
|   |   age > 59
|   |       salary <= 30848.55
|   |       |   salary <= 24972.87: 1 (148.0/39.0)
|   |       |   salary > 24972.87
|   |       |       age <= 61
|   |       |       |   salary <= 25660.98: 0 (3.0/1.0)
|   |       |       |   salary > 25660.98: 1 (11.0)
|   |       |       age > 61: 0 (158.0/45.0)
|   |       salary > 30848.55: 0 (348.0/14.0)
salary > 45570.48
|   salary <= 100191.79
|   |   age <= 40
|   |       salary <= 51922.38: 1 (160.0/52.0)
|   |       salary > 51922.38: 0 (1308.0/98.0)
|   |   age > 40
|   |       age <= 59
|   |       |   salary <= 76745: 1 (718.0/101.0)
|   |       |   salary > 76745: 0 (567.0/40.0)
|   |       age > 59
|   |       |   salary <= 76556.47: 0 (778.0/75.0)
|   |       |   salary > 76556.47: 1 (646.0/42.0)
|   |   salary > 100191.79
|   |       age <= 38: 1 (591.0/42.0)
|   |       age > 38
|   |           age <= 58: 0 (594.0/75.0)
|   |           age > 58
|   |               age <= 59
|   |               |   salary <= 118339.09: 0 (22.0/4.0)
|   |               |   salary > 118339.09: 1 (8.0/1.0)
|   |               age > 59: 1 (666.0/21.0)
salary > 125058.22: 1 (1934.0/36.0)
```

Claves:  
- Salario  
- Edad

# Customer's preferred Brand

## Quiénes compran qué

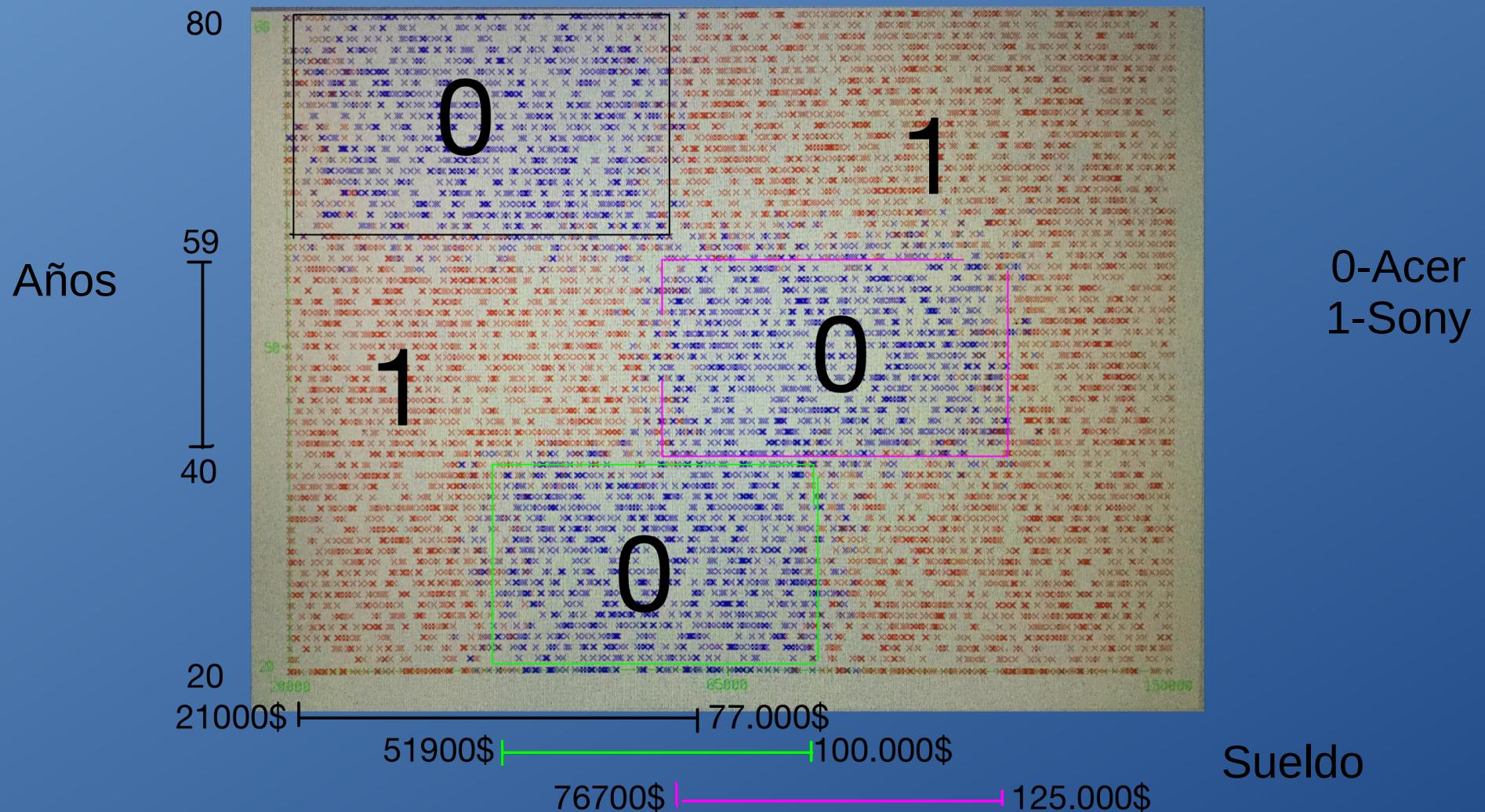
Estudio visual en WEKA:



# Customer's preferred Brand

## Quiénes compran qué

Juntando ambos:



# Customer's preferred Brand Request

- Sería adecuado que nos consultaran antes sobre lo que desean saber para poder escoger los datos y campos adecuados.
- Con los datos actuales se pueden conocer muchas cosas del cliente pero también sería interesante conocer qué clientes no compraron y porqué.
- Es un error pensar que las ventas pasadas se mantendrán en el futuro.
- Es mejor trabajar con valores más abstractos como la imagen de marca que tiene el cliente. Estas se mantienen en el tiempo. También nos permiten buscar productos nuevos de esas u otras marcas que coincidan con el interés del cliente.

# Customer's preferred Brand

## Resumen

Se puede realizar un análisis de los productos más consumidos. Eso nos basta para conocer las marcas con las que trabajar más a corto plazo (y es más barato que una encuesta)

Las ventas pasadas no son suficientes para conocer las ventas futuras. La clave es conocer las necesidades del cliente y qué marca las cubre. ¿Queremos fidelizar a un tipo de cliente?

Una encuesta basada en lo que el cliente valora de cada marca sería una solución a la hora de conocer qué marca vender.

Aún así, con los datos de edad y salario podemos aconsejar mejor al cliente actual, satisfaciendo mejor sus deseos.

Así las cosas desconocemos porqué los extremos en sueldos prefieren un producto más caro que los sueldos medios. Cualquier cosa que añada es pura especulación. Excepción: la elección de los sueldos altos el producto caro.