

21-sept-2013

Exploración del primer micro conjunto de datos que tenemos de pepsi.

Makefile: pandoc -s -V geometry:margin=0.7in -V lang=spanish 1_exploratorio.md -o Exploratorio.pdf

Archivos

- Dcd : Info de clientes. 99+1 variables
- Indcli: Archivo resumen del cliente en el día. 32+1 variables
- Indpro: Archivo x producto, x cliente, x ruta de las cosas vendidas o no. Este es el archivo +imp. 17+1 cols
- Indtie: Info de las horas. 7+1 campos

Indpro

En este archivo están las ventas o no-ventas por producto. R sí cachea bien las 18 columnas, que son:

Nombre original	Nombre nuestro	Descripción
PRO_CLAVE	prod.id	clave del producto
CLI_CLAVE	cliente	clave del cliente corporativa; pa q sea id, concatenar con sucursal
RUT_CLAVE	ruta.id	clave de ruta corporativa
ID_SUC	centro.id	sucursal corporativa; centro de distribución de papitas
FECHA_TRANS	fecha.trans	fecha de la transacción (cuando le venden papitas)
FECHA_CARGA	fecha.carga	fecha en que se esta realizando la carga (cuando sale camionetita)
COM_CLAVE	cia.id	clave de compañía INASE: sabritas (1), quaker. . .
DHP_IMPORTE	vta.impo	Venta Real importe
DHP_UNIDADES	vta.units	Venta Real unidades
BAN_VTA	dummy.vta	Bandera que indica si el producto tuvo venta o no
DHP_INVINI	invent.0	Inventario Inicial
DHP_UNIDEV	u.devol	Unidades devueltas
DHP_UNIROT	u.rot	Unidades rotadas
DHP_UNICAP	u.capt	Unidades Capturadas
DHP_UNISUG	u.sug	Unidades Sugeridas
DHP_UNIFAL	u.falt	Unidades Faltantes
DHP_PRECTE	precio	Precio Cliente

```
setwd("/Users/PandoraMac/Documents/David/pepsi2/Datos_R/")
nombres<-c("carpeta", "prod.id", "cliente", "ruta.id", "centro.id", "fecha.trans", "fecha.carga", "cia.id", "vta.imp
s<-as.data.frame(read.table("R_indpro.txt.mod", sep="|", header=FALSE,
```

```
colClasses=c(rep("character",8),"numeric",rep("integer",8),"numeric"))
names(s)<-nombres
head(s)
```

	carpeta	prod.id	cliente	ruta.id	centro.id	fecha.trans	fecha.carga	cia.id	vta.imp	vta.units	dummy.vta
1	A3490435	695012	I000015919	0001	349	2013-04-18	2013-04-18	30	17.52	3	1
2	A3490435	065713	I000015919	0001	349	2013-04-18	2013-04-18	30	10.84	2	1
3	A3490435	017315	I000015919	0001	349	2013-04-18	2013-04-18	30	16.26	3	1
4	A3490435	583201	I000015919	0001	349	2013-04-18	2013-04-18	30	6.09	1	1
5	A3490435	357231	I000015919	0001	349	2013-04-18	2013-04-18	30	5.83	1	1
6	A3490435	637023	I000015919	0001	349	2013-04-18	2013-04-18	30	12.50	2	1

	invent.0	u.devol	u.rot	u.capt	u.sug	u.falt	precio
1	NA	0	0	3	0	0	5.84
2	NA	0	0	2	0	0	5.42
3	NA	0	0	3	0	0	5.42
4	NA	0	0	1	0	0	6.09
5	NA	0	0	1	0	0	5.83
6	NA	0	0	2	0	0	6.25

Primero, filtramos por compañía:

```
> table(s$cia.id)
```

	1	10	11	14	2	3	30	31
442155	68	2177	514	53385	14507	65986	11627	

```
s<-subset(s,s$cia.id=="1")
```

La cía 1 es Sabritas. Así que sólo tenemos 442K transacciones.

Tables simples

- carpeta
- prod.id. poquitos productos tienen más de 5K registros.

```
> length(table(s$prod.id))
[1] 710
> min(table(s$prod.id))
[1] 1
> max(table(s$prod.id))
[1] 16317
> sum(table(s$prod.id)>5000) #sólo de sabritas son los productos con más de 5000 ventas, btw
[1] 38
> sum(table(s$prod.id<10))
[1] 442155
```

- cliente.id. Nos dijeron que la clave del cliente cambia con el centro de distribución, así que:

```
> head(table(s$cliente,s$centro.id))
```

		454	759
I000000001	8110	2755	
I000000002	17647	9203	
I000000007	4	0	
I000000008	0	6	
I000000017	1	0	

```

I000000018      85      0
> nrow(s) #totales, eran 590419
[1] 442155
> length(paste(s$centro.id,s$cliente.id))
[1] 442155
s$cliente.id<-paste(s$cliente,s$centro.id,sep="-")

■ ruta.id

> length(table(s$ruta.id)) #383 en total... la mayoría son de papitas
[1] 230
> plot(table(s$ruta.id))

```

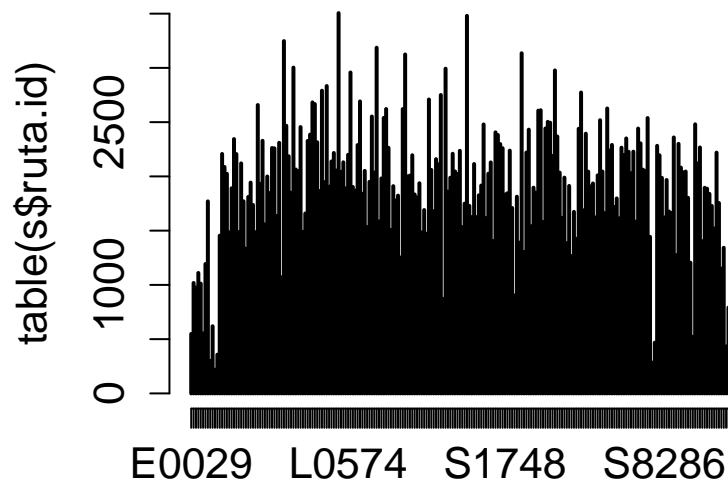


Figura 1: rutas

```

■ centro.id

> table(s$centro.id) #en los datos originales hay ootro centro, que no distribuye papitas

454      759
289823 152332

■ fecha.trans, fecha.carga Transferencia <= carga
■ vta.imp

summary(s$vta.imp)
      Min.      1st Qu.      Median      Mean      3rd Qu.      Max.      NAs
-33860.00    20.24     29.50     41.93     47.20    89530.00    5189

#hist(s$vta.imp)

```

Hay muchos negativos. Habrá que ver q onda con ellos.

```
■ vta.units
```

```
summary(s$vta.units)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NAs
-5024.000	4.000	5.000	7.728	9.000	10800.000	4951

```
plot(s$vta.units)
```

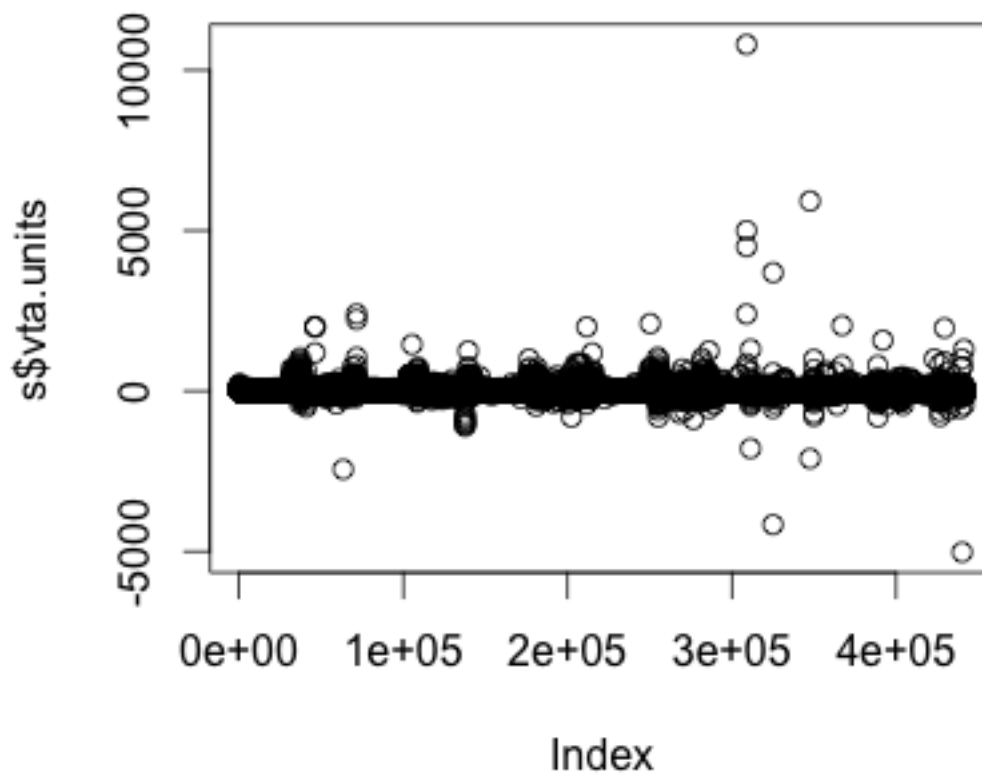


Figura 2: unidades

```
■ dummy.vta
```

```
table(s$dummy.vta)
```

0	1
7364	434791

El 2 % no fue venta...

```
■ invent.0 summary(s$vta.units)
```

```
> summary(s$invent.0)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NAs
   NA      NA      NA    NaN     NA     NA  442155
```

■ u.devol

```
summary(s$u.devol)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NAs
 0.000  0.000  0.000  0.025  0.000 100.000  27432
plot(s$u.devol)
```

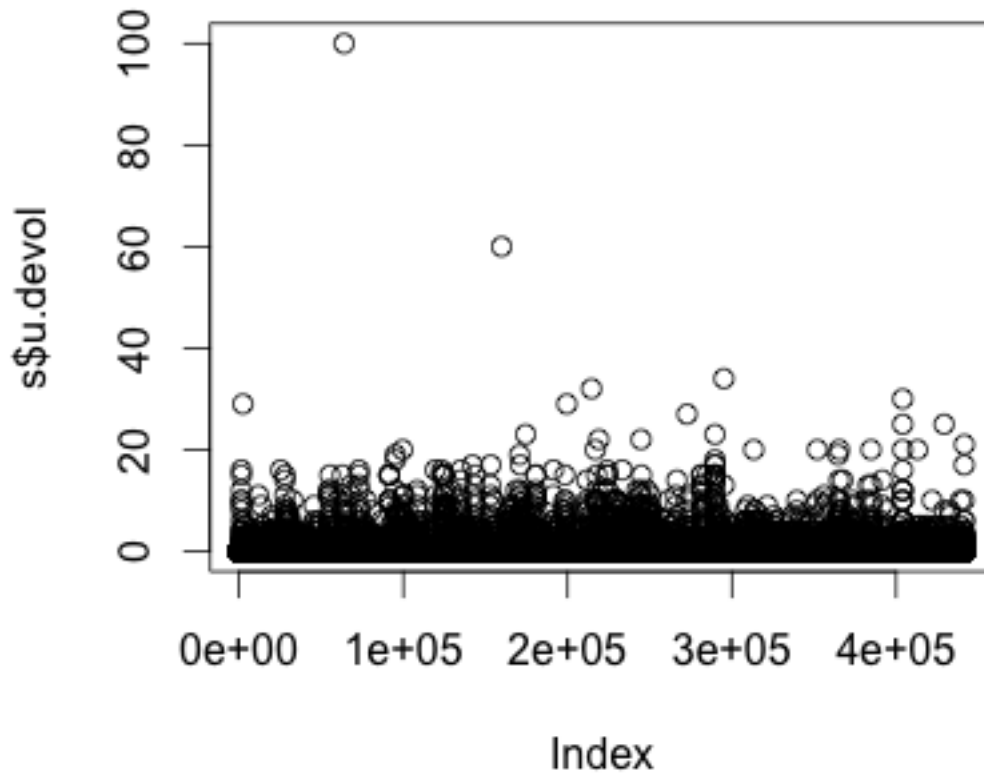
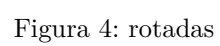


Figura 3: devueltas

“u.rot”

```
summary(s$u.rot)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NAs
 0.000  0.000  0.000  0.891  0.000 5024.000  15458
plot(s$u.rot)
```

■ u.capt



```
summary(s$u.capt)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NAs
  0.000   4.000   5.000   7.643   9.000 5921.000  26854
plot(s$u.capt)
```

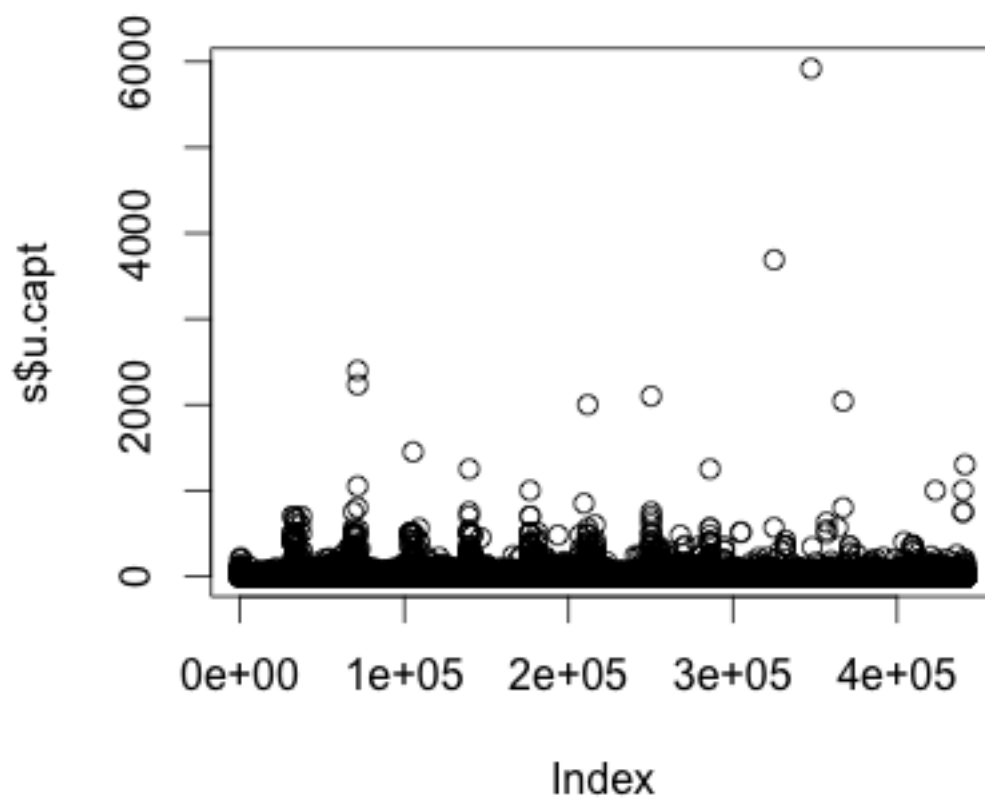


Figura 5: capturadas

■ u.sug

```
summary(s$u.sug)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NAs
    0         0         0         0         0         0  27590
```

■ u.falt

```
summary(s$u.falt)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NAs
    0         0         0         0         0         0  27590
```

■ precio

```
summary(s$precio)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NAs
  0.43   4.22   5.06   6.12   5.90  161.80  31945
plot(s$precio)
```

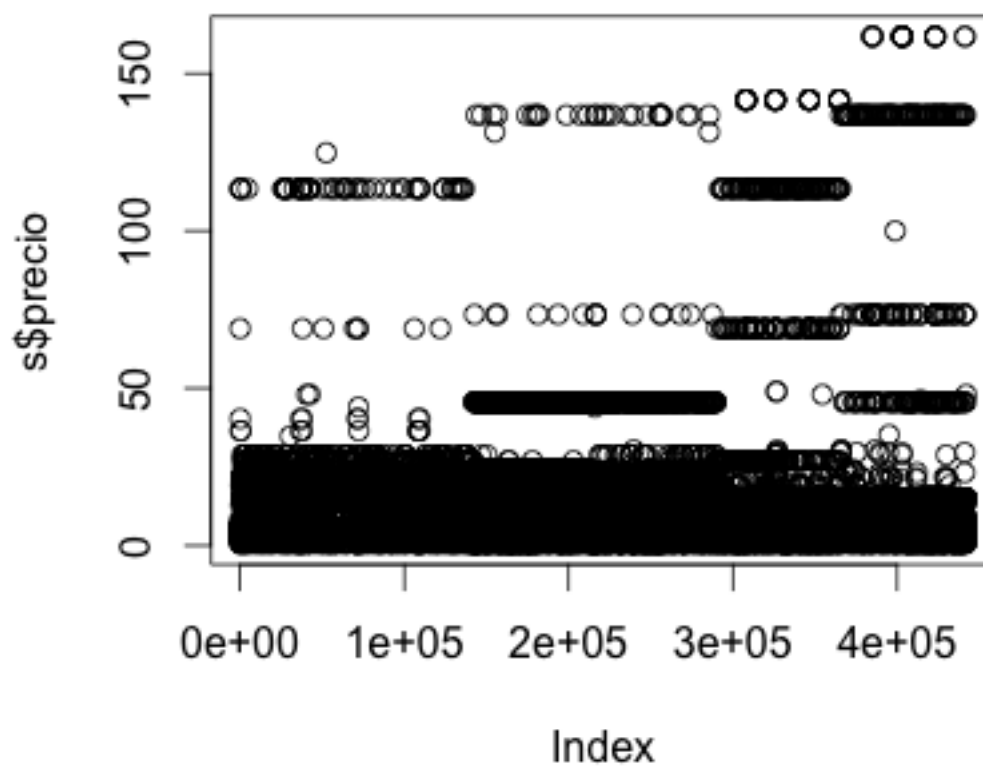


Figura 6: precio

OJO: Qué onda con unidades capturadas y no sé qué tanto?