SESIÓN 2: PROFILING

Una de las mayores preocupaciones en todas las empresas es evitar la fuga de clientes, este es el caso del sector bancario. Para poder tener un conocimiento fiable y poder adoptar políticas de prevención, se han listado las bajas ocurridas en un periodo de tiempo de terminado. Para ellas se ha recogido la información que se dispone del ex cliente: sus características y posición en el año anterior a la baja y el cambio ocurrido hasta antes de tres meses de la baja. La información se completa con una muestra aleatoria de clientes que no han sido baja en el periodo considerado, para los que se recoge la misma información. Con los datos recogidos se ha formado el fichero "churn.txt", conteniendo las siguientes variables:

```
> names(churn)
 [1] "Baja"                "edatcat"             "sexo"
 [4] "antig"               "Nomina"              "Pension"
 [7] "Debito_normal"       "Debito_aff"          "VISA"
[10] "VISA_aff"            "MCard"               "Amex"
[13] "Total_activo"        "Total_Plazo"         "Total_Inversion"
[16] "Total_Seguros"       "Total_Vista"         "dif_resid"
[19] "oper_caj_Libreta"    "oper_ven_Libreta"    "dif_CC"
[22] "dif_Libreta"         "dif_Plazo"           "dif_Ahorro"
[25] "dif_Largo_plazo"     "dif_Fondos_inv"      "dif_Seguros"
[28] "dif_Planes_pension"  "dif_Hipoteca"        "dif_Prest_personales"
```

1. Lea este fichero y efectúe un "summary" de los datos. ¿Detecta algún error o inconsistencia?. Si es así, corríjalo.

   A simple vista, se puede ver que la variable sexo que se presupone categórica de dos valores, tiene uno 'HOMBRE' y uno 'No informado'. Pareciera que 'No informado' debería ser 'MUJER'.

   ```
   > d <- read.csv(file="churn.txt", sep=" ")
   > d <- transform(d, sexo_new= ifelse(sexo=="No informado", "MUJER",
   "HOMBRE"))
   > d$sexo <- d$sexo_new
   > d$sexo_new <- NULL
   ```

   Fuera de eso no se parece observar nada más.

2. Especifique cuál es la variable de respuesta y cuáles son las explicativas y el tipo de todas ellas.

   La variable de respuesta es la de interés, así que en este caso es "Baja", dado que lo que nos interesa saber/predecir en el futuro es que usuarios se darán de baja. Según la numeración de las variables.

   Variable de respuesta: Baja

   Variables categóricas: sexo, Nomina, Pension, Debito_normal, Debito_aff, VISA, VISA_aff, MCard, Amex, dif_resid, edatcat

   Variables cuantitativas: antig, Total_activo, Total_Plazo, Total_Inversion, Total_Seguros, Total_Vista, oper_caj_Libreta, oper_ven_Libreta, dif_CC, dif_Libreta, dif_Plazo, dif_Ahorro, dif_Largo_plazo, dif_Fondos_inv, dif_Seguros, dif_Planes_pension, dif_Hipoteca, dif_Prest_personales

   Edad, que debería ser cuantitativa, es una variable categórica porque en este caso es "categoría de edades", es decir, hay 7 grupos diferentes donde encajan todas las edades

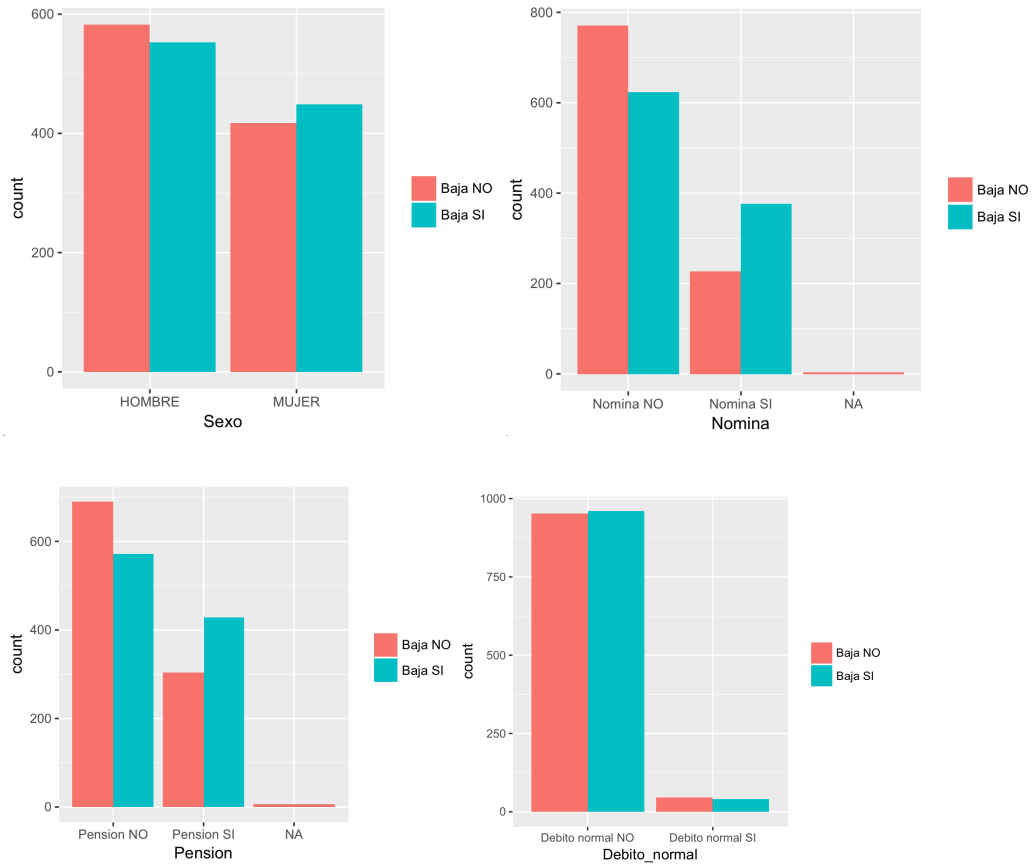
de los clientes.

3. Efectúe una gráfica de los datos; un diagrama de barras para las variables categóricas y un histograma para las variables continuas.

Para cada uno de los tipos de variables he decido añadir siempre la variable de respuesta *Baja*. Así voy haciendo una primera inspección de cómo se comportan los datos.

Diagrama de barras:

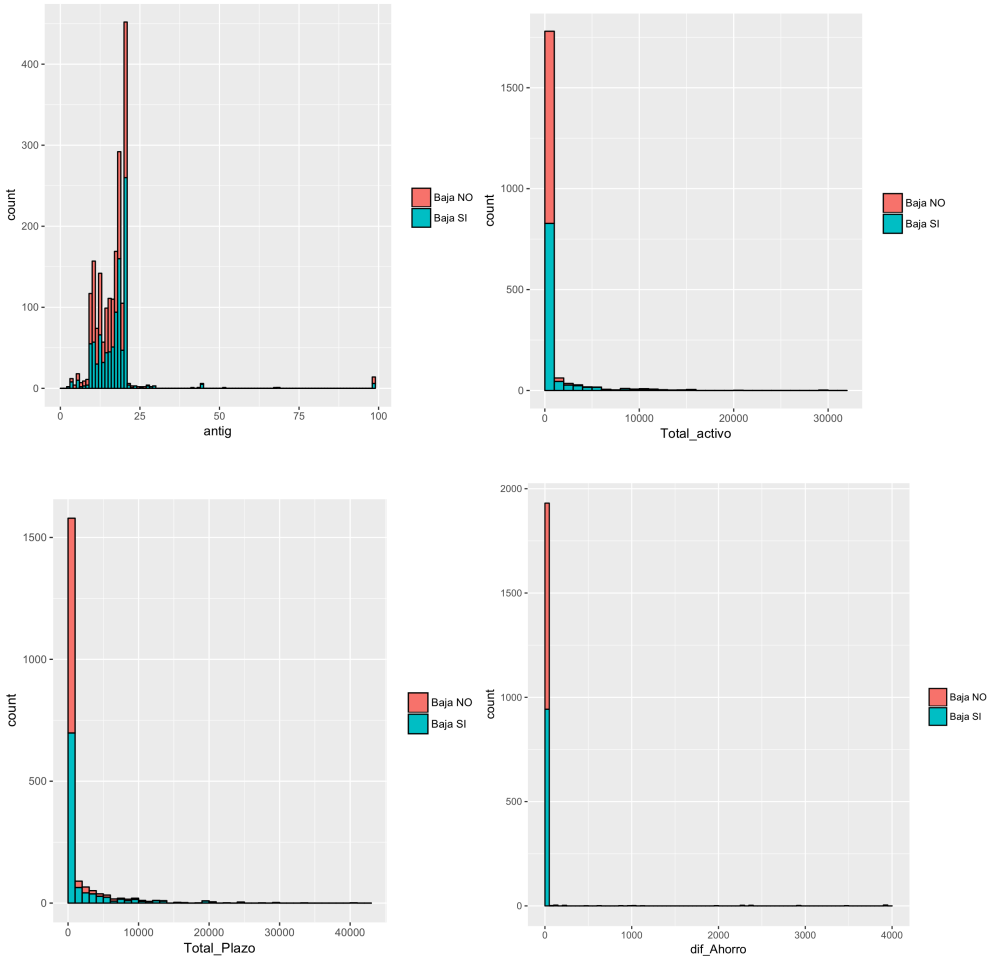
```
> mysubset <- subset(d, select=c("sexo","Baja"))
> ggplot(mysubset, aes(mysubset$Baja, fill=mysubset$sexo) ) +
geom_bar(position="dodge") + scale_x_discrete("Bajas") +
theme(legend.title=element_blank())
```



El resto de las gráficas son todas muy parecidas. He ejecutado las líneas para ver el aspecto que tienen y como método de QA sobre los datos. Pero no las he pegado en el documento. Como comentario general, muchos de las gráficas muestran valores NA (como el gráfico de barras de Pension). Esos datos están probablemente mal.

Histograma:

```
> mysubset <- subset(d, select=c("antig","Baja"))
> max_val <- max(mysubset$antig)
> ggplot(mysubset, aes(mysubset$antig, fill=mysubset$Baja)) +
geom_histogram(breaks=seq(0, max_val,1), col="black",
aes(fill=mysubset$Baja)) + theme(legend.title=element_blank()) +
labs(x="Antiguedad", y="count")
```



De forma similar, he efectuado los histogramas para todas las variables pero aquí adjunto solamente 4 de ellos, los que me han parecido más significativos.

4. Efectúe el "profiling" de las bajas (con la función *catdes* de la librería "FactoMineR"). Interprete el resultado.

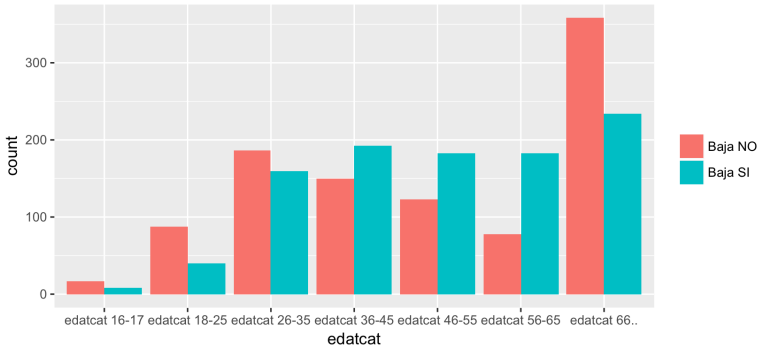
```
> catdes(d, num.var=1, proba=0.05)
```

La primera información que se nos presenta son las variables que mejor explican el comportamiento de la variable de respuesta.

```
                p.value df
edatcat    3.772902e-21  6
VISA       6.501235e-16  2
Nomina     3.564497e-13  2
Debito_aff 1.456320e-10  2
Pension    5.495301e-09  2
```

Eso quiere decir que edatcat (las categorías por edades) es la característica más significativa de todas.

Una rápida inspección al diagrama de barras nos muestra como los datos se distribuyen de forma desigual por todos los rangos de edad. Como el rango 56-65 tiene el mayor porcentaje de clientes que se dieron de baja y como el rango de mayores de 65 tiene el mayor porcentaje de clientes que no se dieron de baja.

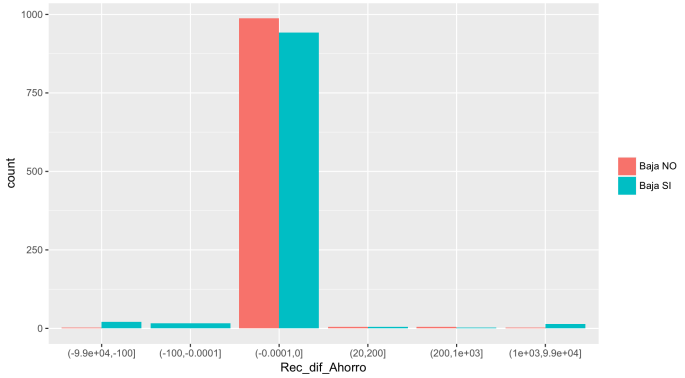


5. Represente visualmente la relación de las variables explicativas con la variable de respuesta; para ello discretize las variables continuas (esto es, recodifíquelas según un cierto número de intervalos; tenga en cuenta el significado especial del valor 0 a la hora de establecer los intervalos de recodificación) y represente mediante barplots el *porcentaje de baja* de las modalidades de las variables categóricas (tanto las categóricas originales como las continuas recodificadas).

Esto es parecido a lo que ya hice en el ejercicio 4, falta recodificar los valores especiales (0), trabajaré con la variable *dif_Ahorro*, haciendo un subset y recodificando los valores

```
> churn2$Rec_dif_Ahorro = cut(churn$dif_Ahorro, breaks=c(-99000,-
100,-0.0001,0,20,200,1000,99000))
> mysubset <- subset(churn2, select=c("Rec_dif_Ahorro","Baja"))
> ggplot(mysubset, aes(mysubset$Rec_dif_Ahorro, fill=mysubset$Baja)
) + geom_bar(position="dodge") + scale_x_discrete("Rec_dif_Ahorro")
+ theme(legend.title=element_blank())
```

Con lo cual queda:



6. Suponga que quiere analizar la compra de un producto a partir del barrio de residencia (alto o bajo) (indicador del poder adquisitivo del cliente). En un primer análisis de obtiene la siguiente tabla:

	Compra SI	Compra NO	Total
Clase alta	20	373	393
Clase baja	6	316	322

En su opinión, ¿el poder adquisitivo del cliente, tiene alguna influencia sobre la compra o no del producto? (Responda sólo calculando las probabilidades, sin realizar la prueba de hipótesis de igualdad entre ambas probabilidades).

Pues según esa tabla el 5.09% de los clientes de clase alta compran el producto, mientra

que tan solo el 1.86% de los clientes de clase baja lo hacen. Sí se podría decir que tiene cierta influencia.

Un empleado senior de la compañía nos sugiere profundizar más en el análisis y tener en cuenta la edad de los clientes. Cruzando por edad (adulto o joven) los dos tipos de barrio mencionados, obtenemos las siguientes tablas:

ADULTOS	Compra SI	Compra NO	Total
Clase alta	3	176	179
Clase baja	4	293	297
JOVENES	Compra SI	Compra NO	Total
Clase alta	17	197	214
Clase baja	2	23	25

Con estos nuevos datos, obtenemos:
Adultos clase alta: 1.67%
Adultos clase baja: 1.34%
Jóvenes clase alta: 7.94%
Jóvenes clase baja: 8%

¿Tenía razón el empleado de que era conveniente tener en cuenta la edad?. ¿Cuál de los dos factores, el barrio de residencia o la edad, es el determinante en la compra del producto en cuestión?

Es mucho más determinante la edad, dado que en ambos casos (clase baja o alta) hay un índice significativamente mayor de jóvenes que compran el producto que adultos. En el caso de la clase baja, los índices de la primera tabla son tan bajos, porque el número de adultos de clase baja es muy elevado y ellos son los que menos compran el producto, el número de jóvenes de la misma clase es mucho inferior, no consiguiendo contrarestar el peso de los adultos y subir el porcentaje.