Reporte de Conversión de Clientes

Introducción

El presente documento contiene información del *Bank Marketing Data Set*, cuya fuente es el artículo **A Data-Driven Approach to Predict the Success of Bank Telemarketing** publicado por S. Moro, P. Cortez y P. Rita en 2014. La información se utiliza con fines ilustrativos de cómo pueden extraerse elementos de interés, especialmente en la estratificación de las edades y estado civil. Es importante mencionar, ya que la información es pública y de libre acceso para el público a través del repocitorio de la University of California, Irvine.

Para el presente caso se realiza un Análisis Exploratorio de Datos (EDA por sus siglas en *inglés*), considerando los grupos de edad y el estado civil de los clientes del banco, que han sido contactados para la colocación de productos financieros.

Se analiza como variable principal la **Tasa de Conversión**, con el objetivo de validar los segmentos de mayor y menor conversión dentro de la campaña de *call center* que realiza el banco.

En este documento se integran los *scripts* de R para que el usuario pueda replicar los resultados sin inconvenientes, para que pueda validar la veracidad de los resultados y afirmaciones que se plasman en el reporte.

Análisis Exploratorio de Datos

Antes de comenzar propiamente el análisis exploratorio de datos, es imporante definir el término **Tasa de Conversión** en función de su significado y la forma en que se calcula.

La **Tasa de Conversión** es un indicador que permite medir y analizar los resultados de una estrategia de *marketing*, ya que es el resultado de dividir el número de ventas realizadas (*conversiones*) entre el número total de clientes contactados.

Formula Tasa de Conversión

$$TC = \frac{\mathrm{C}}{\mathrm{N}}$$

TC : Tasa de Conversión.

C : Conversiones.

N : Total Clientes.

Este indicador refleja el *rendimiento* de cómo va la campaña de *marketing*. Permite mantener un monitoreo de los resultados, así como su evaluación constante por un periodo definido. De esta forma, se puede contrastar la **Tasa de Conversión** con los objetivos definidos, evaluar las variaciones que se presenten para tomar medidas oportunas y/o refomular las estrategias, canales de venta, productos y calidad del servicio que se están utilizando para conseguir los objetivos propuestos.

Una vez definido el indicador principal de este análisis, corresponde revisar las variables que se incluyen en el conjunto de datos del *Bank Marketing Data Set*. La información se integra de 41,118 registros y 21 variables, de las cuales 5 son de tipo entero, 5 de tipo decimal o flotante y 11 de tipo texto.

```r
# Definir la librerías de trabajo
library(dplyr)
library(ggplot2)

# Cargar el conjunto de datos del Bank Marketing Data Set
df <- read.csv('bank-additional-full.csv', sep=';', header=T)

# Motrar la estructura de la información
print(str(df))
```

```
'data.frame':   41188 obs. of  21 variables:
 $ age           : int  56 57 37 40 56 45 59 41 24 25 ...
 $ job           : chr  "housemaid" "services" "services" "admin." ...
 $ marital       : chr  "married" "married" "married" "married" ...
 $ education     : chr  "basic.4y" "high.school" "high.school" "basic.6y" ...
 $ default       : chr  "no" "unknown" "no" "no" ...
 $ housing       : chr  "no" "no" "yes" "no" ...
 $ loan          : chr  "no" "no" "no" "no" ...
 $ contact       : chr  "telephone" "telephone" "telephone" "telephone" ...
 $ month         : chr  "may" "may" "may" "may" ...
 $ day_of_week   : chr  "mon" "mon" "mon" "mon" ...
 $ duration      : int  261 149 226 151 307 198 139 217 380 50 ...
 $ campaign      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ pdays         : int  999 999 999 999 999 999 999 999 999 999 ...
 $ previous      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome      : chr  "nonexistent" "nonexistent" "nonexistent" "nonexistent" ...
 $ emp.var.rate  : num  1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...
 $ cons.price.idx: num  94 94 94 94 94 ...
 $ cons.conf.idx : num  -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...
 $ euribor3m     : num  4.86 4.86 4.86 4.86 4.86 ...
```

```
 $ nr.employed   : num  5191 5191 5191 5191 5191 ...
 $ y             : chr  "no" "no" "no" "no" ...
NULL
```

De igual forma se hace una revisión de la información para confirmar que no existan valores NA o NULL.

> **Diferencias entre NA y NULL**
>
> Generalmente tienden a confundirse ambos tipos de valores nulos cuando se analiza información con valores faltantes, sin embargo, cada uno tiene una naturaleza distinta. Por ejemplo **NULL** representa en R el objeto nulo, es decir, valores que no están definidos al momento de obtener los resultados de una función. Por su parte **NA** es una constante lógica que representa el indicador de valor faltante.

Como se puede observar en las líneas de código inferiores, el conjunto de datos no muestra la existencia de valores nulos o faltantes en ninguna de las variables que lo integran.

```
# Identificar si existen valores Na en el conjunto de datos
print(paste0('Total de registros NA: ', sum(is.na(df))))
```

```
[1] "Total de registros NA: 0"
```

```
# Identificar si existen valores Null en el conjunto de datos
print(paste0('Total de registros NULL: ', sum(is.null(df))))
```

```
[1] "Total de registros NULL: 0"
```

Ya identificadas las variables y validando que no existan valores faltantes o nulos, se procede a caluclar la Tasa de Conversión, modificando la variable **y** a través de un flujo condicional, dado que contiene una confirmación de si el cliente ha adquirido algún producto financiero (valor 1) o no (valor 0).

```r
# Categorizar variable y
df <- df %>%
  mutate(y=ifelse(y=='no', 0, 1))

# Confirmar que el tipo de variable sea entero
df$y <- as.integer(df$y)
```

Con esta variable modificada se puede calcular el número de conversiones y dividr por el total de registros que tiene la tabla, dando como resultado la **Tasa de Conversión Total** que se obtiene de toda la campaña de *marketing*.

```r
# Calcular el total de conversiones y total de clientes
conversions <- sum(df$y)
customers <- nrow(df)

# Obtener la tasa de conversión
conversion_rate <- round((conversions/customers)*100, 2)
print(paste0('Tasa de Conversión Total: ', conversion_rate, '%'))
```

```
[1] "Tasa de Conversión Total: 11.27%"
```

Con esta información ahora es posible identificar los estratos de edad para revisar cuáles son los estratos de mayor o menor nivel de conversión.

Para efectuar la estratificación, se agrupa la información de acuerdo a 7 diferentes estratos *Teens (de 17 a 20 años)*, *(20,30] años*, *(30,40] años*, *(40,50] años*, *(50,60] años*, *(60,70] años* y *+70 años*. Con base en estas agrupaciones, se calculan dos variables: el número total de clientes por cada grupo (TotalCount) y el número de conversiones (NumberConversions). Por último, con base en ambas variables *sumarizadas*, se calcula la tasa de conversión para cada estrato.

```
# Estratos de edad (Teens (17-19), 20-30, 30-40, 40-50, 50-60, 60-70 y +70)
conversionAgeGroups <- df %>% group_by(AgeGroup=cut(age, breaks=append((17),
seq(20, 70, by=10)))) %>%
  summarize(TotalCount=n(), NumberConversions=sum(y)) %>%
  mutate(ConversionRate=round(NumberConversions/TotalCount, 2)*100)
```

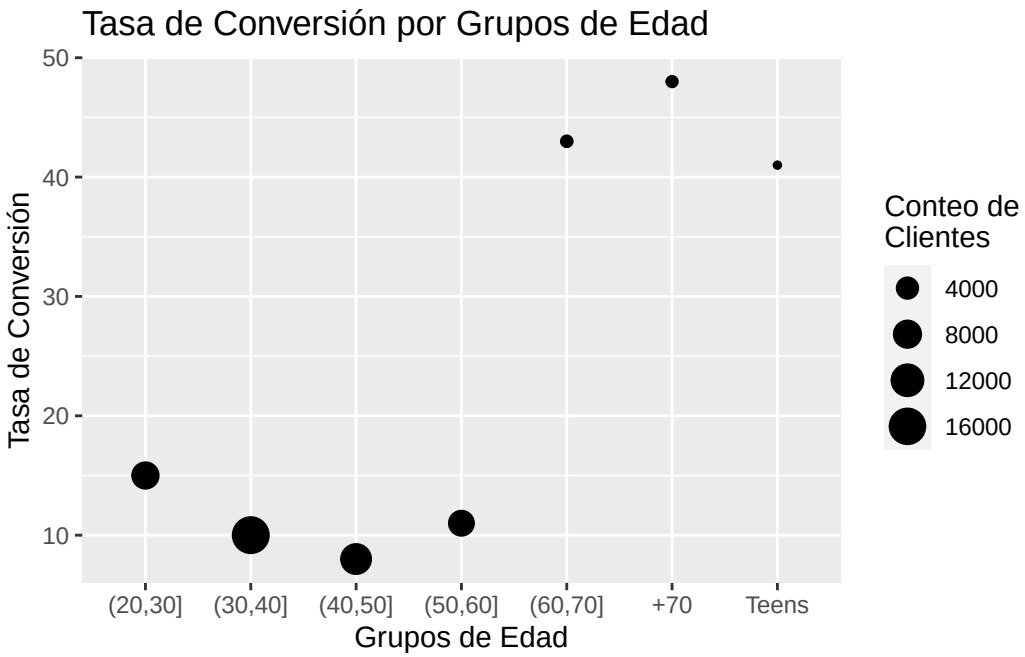
Ya definidos los estratos, se ajustan el tipo de variable de los grupos (AgeGroup), así como las etiquetas para el primer y último estrato.

```
# Renombrar nombres del primer y último grupo
conversionAgeGroups$AgeGroup <- as.character(conversionAgeGroups$AgeGroup)
conversionAgeGroups$AgeGroup[1] <- 'Teens'
conversionAgeGroups$AgeGroup[7] <- '+70'

# Convertir a variables categóricas los grupos de edad
conversionAgeGroups$AgeGroup <- as.factor(conversionAgeGroups$AgeGroup)
```

Por último, se grafican los resultados empleando un gráfico de dispersión modificado, donde el tamaño de las dispersiones está en función de la cantidad de clientes contactados para cada estrato.

```
# Gráfico de dispersión Estratos-Tasa de Conversión
ggplot(conversionAgeGroups, aes(x=AgeGroup, y=ConversionRate)) +
  geom_point(aes(size=TotalCount)) +
  xlab('Grupos de Edad') +
  ylab('Tasa de Conversión') +
  labs(title='Tasa de Conversión por Grupos de Edad', size='Conteo de\nClientes')
```



Se realiza un ejercio similar de segmentación, pero en este nuevo caso se integra la variable Estado Civil para analizar la participación de cada estado civil (Divorciado, Casado, Soltero y Sin Definir).

Los estados civiles para este caso, permiten obtener un nivel adicional de segmentación para los casos en los que la variable **y** es igual a 1, es decir, los resultados que muestra la siguiente gráfica corresponden a las proporciones de clientes, por edad y estado civil, que han contratado un servicio con el banco.

```
# Obtener los grupos por edad y estado civil
conversionAgeMarital <- df %>%
  filter(y==1) %>%
  group_by(AgeGroup=cut(age, append((17), seq(20, 70, by=10))),
           Marital=marital) %>%
  summarize(Count=n(), NumConversions=sum(y)) %>%
  mutate(TotalCount=sum(Count)) %>%
  mutate(ConversionRate=round((NumConversions/TotalCount)*100, 2))
```

```
`summarise()` has grouped output by 'AgeGroup'. You can override using the
`.groups` argument.
```

```r
# Integrar etiquedas Teens y +70
conversionAgeMarital$AgeGroup <- as.character(conversionAgeMarital$AgeGroup)
conversionAgeMarital$AgeGroup[is.na(conversionAgeMarital$AgeGroup)] <- '+70'
conversionAgeMarital$AgeGroup[conversionAgeMarital$AgeGroup=='(17,20]'] <- 'Teens'

# Gráfico de barras para comparar las proporciones de estado civil para cada grupo
ggplot(conversionAgeMarital, aes(x=AgeGroup, y=ConversionRate, fill=Marital)) +
  geom_bar(width=0.5, stat='identity') +
  xlab('Grupos de Edad') +
  ylab('Tasa de Conversión') +
  labs(title='Tasa de Conversión por Grupos de Edad y Estado Civil',
  fill='Estado Civil') +
  scale_fill_discrete(labels=c('Divorciado', 'Casado', 'Soltero', 'Sin Definir'))
```