

UNIVERSIDAD OBERTA DE CATALUNYA

**ESTUDIOS DE INFORMÁTICA, MULTIMEDIA Y TELECOMUNICACIONES.
TIPOLOGIA Y CICLO DE VIDA DE LOS DATOS.
PRACTICA2- . LIMPIEZA Y ANÁLISIS DE DATOS.**

**APELLIDOS: ZAMBRANO ZAMBRANO
NOMBRES: LÍDER EUCLIDES**

RESPUESTAS A LAS PREGUNTAS DE LA PRACTICA 2 – LIMPIEZA Y ANÁLISIS DE DATOS.

DESCRIPCIÓN DEL DATASET.

1. ¿POR QUÉ ES IMPORTANTE Y QUÉ PREGUNTA/PROBLEMA PRETENDE RESPONDER?

El juego de datos que hemos seleccionado para nuestra PRACTICA, lo hemos tomado de la página de data.world, aquí el enlace: <https://data.world/earino/churn>, es un conjunto de datos con el nombre churn.csv que recoge datos de una empresa de telecomunicaciones y que permitirá ver a los clientes que han abandonado la compañía y quienes no, es importante este conjunto de datos porque nos permite conocer las circunstancias en que los usuarios han abandonado la compañía mediante la observación de las variables de estudio y las condiciones que genera el algoritmo de clasificación (específicamente árboles de decisión) a aplicar en vista que trabajaremos como variable predictiva con una variable categórica.

Churn (en español abandono) es la variable predictiva, y por medio del análisis de esta variable podemos conocer la respuesta a la siguiente pregunta **¿Quiénes abandonan la compañía de telecomunicaciones y quienes no?**

Por otro lado nos permite comprender un modelo de resolución de problemas en las empresas; ya que mediante este análisis podremos entender las causas por las que los clientes se dan de baja a ciertos servicios de telefonía.

El conjunto de datos seleccionado cuenta con 5000 filas y 18 columnas; el nombre de las columnas están en inglés pero la hemos traducido al español para una mejor comprensión de los datos, quedando los nombres de las variables de esta manera:

"abandono", "longitud de cuenta", "plan_internacional", "plan de correo de voz", "numero de mensajes de correo", "minutos por dia", "llamadas por dia", "total de carga diaria", "total de minutos", "total de llamadas", "total de carga", "minutos de noche", "llamadas de noche", "carga de noche", "minutos internacionales", "llamadas internacionales", "total de cargo internacional", "reclamos".

CHURN (Abandono): Indica si el usuario abandona o no la compañía.

ACCOUNTLENGTH (Longitud de cuenta): Hace referencia a la duración de las cuentas.

INTERNATIONALPLAN (Plan internacional): Indica si el cliente registra un plan con llamadas y mensajes al exterior.

VOICEMAILPLAN (Plan de correo de voz): Indica si el cliente registra un plan con correo de voz.

NUMBERVMAILMESSAGES (Numero de mensajes de correo electrónico): Indica el número de mensajes por correo que ha realizado el cliente.

TOTALDAYMINUTES (total de minutos al día): Indica el número de minutos que el cliente habló durante el día.

TOTALDAYCALLS (total de llamadas diarias): Indica el número de llamadas que el usuario realizó diariamente.

TOTALDAYCHARGE (Cargo total por día): Indica el monto que el usuario cancelará por llamadas o mensajes realizados durante el día.

TOTALEVEMINUTES (Total de minutos): Indica el número total de minutos que el usuario ha hablado.

TOTALEVECALLS (Total de llamadas): Indica el número total de llamadas que el usuario ha realizado.

TOTALEVECHARGE (Total de cargas): Indica el monto que el usuario cancelará por llamadas o mensajes realizados.

TOTALNIGHTMINUTES (Minutos por la noche): Indica el número de minutos hablados durante la noche.

TOTAL NIGHT CALLS (Llamadas por la noche): Indica el número de llamadas efectuadas por el usuario durante la noche.

TOTAL NIGHT CHARGE (Cargo total nocturno): Indica el número de cargas que el usuario hace en la noche para realizar llamadas o enviar mensajes.

TOTAL INTL MINUTES (Minutos internacionales): Indica el número de minutos que el usuario ha hablado con familiares o amigos al extranjero.

TOTAL INTL CALLS (Llamadas internacionales): Indica el número de llamadas que el usuario ha realizado al llamar al extranjero.

TOTAL INTL CHARGE (Total de cargo internacional): monto económico que el usuario tendrá que cancelar por los mensajes y llamadas al exterior.

CUSTOMER SERVICE CALLS (Llamadas de servicio al cliente): Indica el número de llamadas que el usuario realizó al personal de servicio al cliente de la compañía para dar a conocer algún reclamo.

2. INTEGRACION Y SELECCION DE LOS DATOS DE INTERES A ANALIZAR.

De las 18 variables no vamos a necesitar las variables ACCOUNTLENGTH, VOICEMAILPLAN, NUMBERVMAILMESSAGES, TOTALDAYCHARGE, TOTALEVEMINUTES, TOTALEVECALLS, TOTALEVECHARGE, TOTAL NIGHT CHARGE, TOTAL INTL CHARGE, ya que estas variables no las consideramos significativas para la predicción.

Seleccionamos exclusivamente las columnas que necesitamos

```
data <- data[, c(1, 3, 6, 7, 12,13, 15,16,18)]
```

Las columnas seleccionadas en orden ascendente son abandona, plan_internacional, minutos_dia, llamadas_dia, minutos_noche, llamadas_noche, minutos_internacionales, llamadas_internacionales, reclamos.

3. LIMPIEZA DE DATOS

3.1. ¿LOS DATOS CONTIENEN CEROS O ELEMENTOS VACIOS? ¿CÓMO GESTIONARIAS CADA UNO DE ESTOS CASOS?

¿LOS DATOS CONTIENEN CEROS O ELEMENTOS VACIOS?

Nuestra data si contiene ceros, pero como sabemos el cero no siempre hace referencia a un valor perdido, en algunos contextos puede representar datos vacíos o no definidos, sin embargo en otros contextos como el ejemplo de nuestro conjunto de datos se trata de valores legítimos.

Nuestra data original no ha contenido valores vacíos, sin embargo para tratar a los valores vacíos mediante las técnicas de imputación de datos, hemos manipulado nuestra data para que existan valores vacíos y así poder tratarlos.

La técnica de imputación que hemos considerado es la de completar los valores vacíos mediante la media aritmética en las variables minutos_dia y llamadas_dia.

Conocemos que existen muchas técnicas de imputación de datos como la de KNN, imputación por medias que hemos aplicado, imputación por regresión y la misma imputación manual, entre otras.

¿CÓMO GESTIONARIAS CADA UNO DE ESTOS CASOS?

Primero cargamos nuestra data.

```
data<-read.csv("./churn.csv",header=T,sep=",")  
attach(data)
```

Modificamos los nombres a los campos.

```
names(data) <- c("abandona", "longitud_cuenta", "plan_internacional", "plan_correo_voz", "mensajes_por_correo", "minutos_dia", "llamadas_dia", "total_carga_dia", "total_minutos", "total_llamadas", "total_carga", "minutos_noche", "llamadas_noche", "total_carga_noche", "minutos_internacionales", "llamadas_internacionales", "total_cargo_internacional", "reclamos")
```

Seleccionamos las columnas con las que vamos a trabajar.

```
data <- data[, c(1, 3, 6, 7, 12, 13, 15, 16, 18)]
```

La función `dim()` nos permite conocer exactamente el número de filas y columnas de nuestro dataset.

```
dim(data)
```

```
## [1] 5000    9
```

Observamos 5000 registros y 9 variables, existían 18 variables pero hemos seleccionado solo las variables que necesitamos.

Es de gran interés saber si tenemos muchos valores nulos (campos vacíos) y la distribución de valores por variables. Es por ello recomendable empezar el análisis con una visión general de las variables. Mostraremos para cada atributo la cantidad de valores perdidos mediante la función `summary`.

El nuestro conjunto de datos original no existían valores vacíos; pero hemos eliminado algunos valores para tratarlos.

```
summary(data)
```

```
## abandona    plan_internacional minutos_dia    llamadas_dia
## No :4293    no :4527             Min.   : 0.0    Min.   : 0
## Yes: 707    yes: 473             1st Qu.:143.7  1st Qu.: 87
##                                     Median :180.1  Median :100
##                                     Mean    :180.3  Mean    :100
##                                     3rd Qu.:216.2  3rd Qu.:113
##                                     Max.    :351.5  Max.    :165
##                                     NA's    :2      NA's    :2
## minutos_noche llamadas_noche minutos_internacionales
## Min.   : 0.0    Min.   : 0.00    Min.   : 0.00
## 1st Qu.:166.9    1st Qu.: 87.00    1st Qu.: 8.50
## Median :200.4    Median :100.00    Median :10.30
## Mean    :200.4    Mean    : 99.92    Mean    :10.26
## 3rd Qu.:234.7    3rd Qu.:113.00    3rd Qu.:12.00
## Max.    :395.0    Max.    :175.00    Max.    :20.00
##
## llamadas_internacionales reclamos
## Min.   : 0.000    Min.   :0.00
## 1st Qu.: 3.000    1st Qu.:1.00
## Median : 4.000    Median :1.00
## Mean    : 4.435    Mean   :1.57
## 3rd Qu.: 6.000    3rd Qu.:2.00
## Max.    :20.000    Max.    :9.00
##
```

```
# Revisamos nuevamente valores vacios
colSums(is.na(data))
```

```
##          abandona      plan_internacional      minutos_dia
##              0              0              2
##      llamadas_dia      minutos_noche      llamadas_noche
##              2              0              0
## minutos_internacionales llamadas_internacionales      reclamos
##              0              0              0
```

Vemos que existen 2 valores vacíos en la variable minutos_dia y 2 valores vacíos en la variable llamadas_dia.

Ahora vamos a tratar los valores vacíos por la media aritmética en las variables discretas.

```
# Tomamos la media para valores vacios de la variable "minutos_dia" y "llamadas_dia"
data$minutos_dia[is.na(data$minutos_dia)] <- mean(data$minutos_dia,na.rm=T)
data$llamadas_dia[is.na(data$llamadas_dia)] <- mean(data$llamadas_dia,na.rm=T)
```

Revisamos nuevamente la data, para verificar los cambios, y vemos que ya no existen valores vacíos.

```
# Revisamos nuevamente valores vacios
colSums(is.na(data))
```

```
##          abandona      plan_internacional      minutos_dia
##              0              0              0
##      llamadas_dia      minutos_noche      llamadas_noche
##              0              0              0
## minutos_internacionales llamadas_internacionales      reclamos
##              0              0              0
```

3.2 IDENTIFICACION Y TRATAMIENTO DE VALORES EXTREMOS

Para identificar los valores extremos hemos utilizado un diagrama de cajas y la función boxplots.stats ().

INICIAMOS CON LA COLUMNA MINUTOS_DIA

El proceso que hemos ejecutado es el siguiente: calculamos los cuartiles de la columna minutos_dia

```
quantile(data$minutos_dia)
```

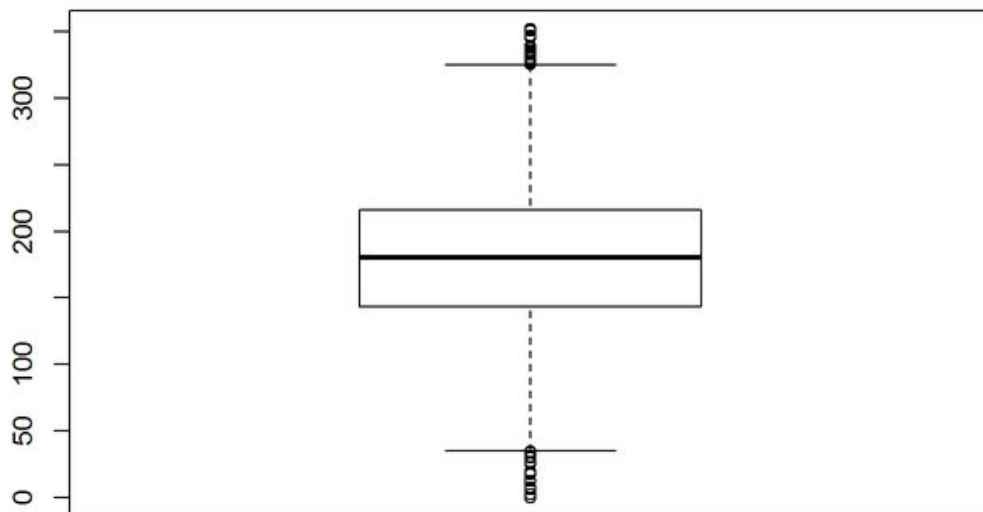
```
##      0%    25%    50%    75%   100%  
##    0.0 143.7 180.1 216.2 351.5
```

El primer dato es el valor mínimo y el último es el valor máximo, quiere decir que para esta variable hay clientes que no hablan ningún minuto al día; y también hay usuarios que hablan hasta 351.5 minutos al día. (No habrá valores mayores a 351.5 ya que este es el valor máximo de la columna minutos_dia)

Los otros 3 datos son los cuartiles, el cuartil 1 (143.7), el cuartil 2 (180.1) y el cuartil 3 (216.2); (el cuartil 2 que es 180.1 es la media).

Podemos ver los datos de la columna minutos_dia en un diagrama de cajas, para ello trabajamos con el comando boxplot().

```
boxplot(data$minutos_dia)
```



Ahora vamos a verificar si hay datos atípicos para esto necesitamos el rango intercuartil que se lo obtiene con el comando IQR.

El rango intercuartil (IQR) es la distancia entre el primer cuartil (Q1) y el tercer cuartil (Q3). El 50% de los datos está dentro de este rango.

```
IQR(data$minutos_dia)
```

```
## [1] 72.5
```

El rango intercuartil es 72.5

Para saber si hay algún dato atípico necesitamos un rango; y lo calculamos así:

(El rango intercuartil multiplicado por 1.5 veces este mismo rango y sumado el primer cuartil para encontrar el valor máximo del rango; a su vez el rango intercuartil lo multiplicamos por 1.5 y restamos el primer cuartil y encontramos el valor mínimo del rango)

```
MIN<- (143.7-1.5*72.5)  
MAX<- (143.7+1.5*72.5)
```

Ahora vamos a observar el rango que nos muestra el MIN y el MAX.

```
range(MIN,MAX)
```

```
## [1] 34.95 252.45
```

Si los datos de la columna minutos_dia; se encuentran dentro de este rango (34.95 y 252.45) significa que no hay datos atípicos, pero si hay un dato por fuera de este rango ese dato será atípico.

Veamos el rango de la columna de minutos_dia.

```
range(data$minutos_dia)
```

```
## [1] 0.0 351.5
```

Observamos que hay datos fuera del rango de MIN y MAX, eso quiere decir que tenemos datos atípicos.

Pero ¿cuáles son esos valores que están fuera del rango de MIN y MAX?; los podemos observar con la función boxplots.stats.

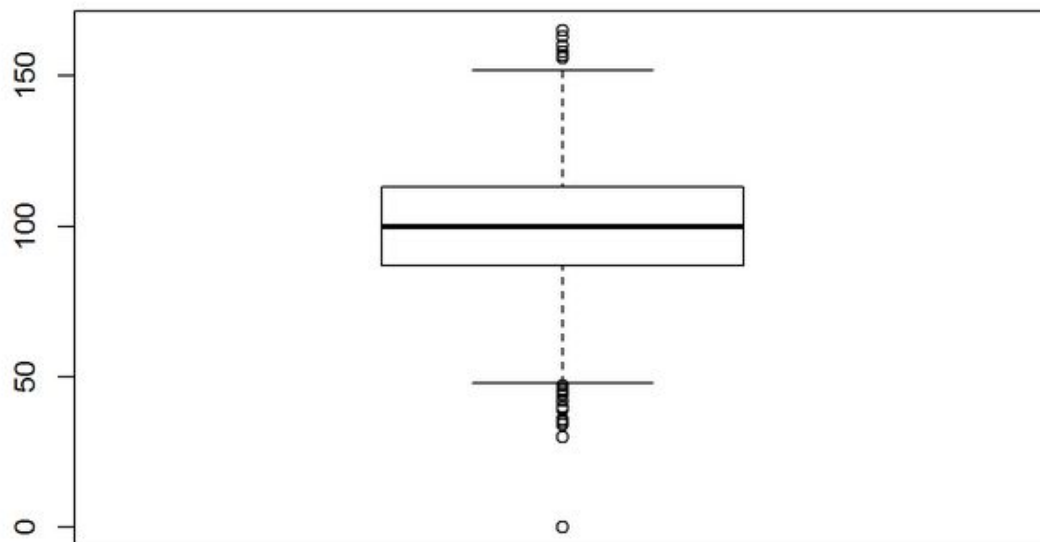
```
boxplot.stats(data$minutos_dia)$out
```

```
## [1] 332.9 337.4 326.5 350.8 335.5 30.9 34.0 334.3 346.8 12.5 25.9
## [12] 0.0 0.0 19.5 329.8 7.9 328.1 27.0 17.6 326.3 345.3 2.6
## [23] 7.8 18.9 29.9 338.4 326.1 325.4 351.5 332.1 325.5 6.6 34.5
## [34] 7.2
```

COLUMNA LLAMADAS_DIA

Graficamos con el comando boxplot y observamos valores fuera del bigote superior e inferior.

```
boxplot(data$llamadas_dia)
```



Calculamos los cuartiles

```
quantile(data$llamadas_dia)
```

```
## 0% 25% 50% 75% 100%
## 0 87 100 113 165
```

Calculamos el rango intercuartil


```
IQR(data$llamadas_dia)
```

```
## [1] 26
```

Calculamos el valor mínimo y máximo del nuevo rango para comparar con los valores extremos.

```
MIN<- (87-1.5*26)  
MAX<- (87+1.5*26)
```

Imprimimos los valores mínimos y máximos.

```
range(MIN,MAX)
```

```
## [1] 48 126
```

Estos datos nos indican que si los valores están entre 48 y 126 no son valores atípicos, fuera de estos sí; luego observamos los valores que están fuera de este rango.

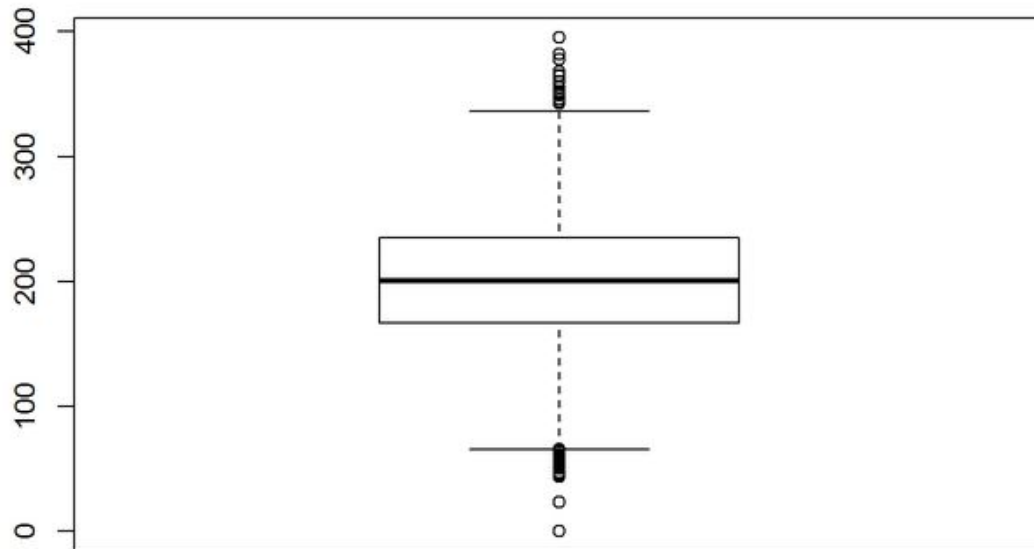
```
boxplot.stats(data$llamadas_dia)$out
```

```
## [1] 47 158 47 163 36 40 158 165 30 42 0 45 0 45 160 156 35  
## [18] 42 158 157 45 44 44 44 40 47 34 39 156 156 157 44 46 160  
## [35] 39
```

COLUMNA MINUTOS_NOCHE

Graficamos con el comando boxplot y observamos valores fuera del bigote superior e inferior.

```
boxplot(data$minutos_noche)
```



Calculamos los cuartiles

```
quantile(data$minutos_noche)
```

```
##      0%   25%   50%   75%  100%
##    0.0 166.9 200.4 234.7 395.0
```

Calculamos el rango intercuartil

```
IQR(data$minutos_noche)
```

```
## [1] 67.8
```

Calculamos el valor mínimo y máximo del nuevo rango para comparar con los valores extremos.

```
MIN<- (166.9-1.5*67.8)
MAX<- (166.9+1.5*67.8)
```

Imprimimos los valores mínimos y máximos.

```
range(MIN,MAX)
```

```
## [1] 65.2 268.6
```

Estos datos nos indican que si los valores están entre 65.2 y 268.6 no son valores atípicos, fuera de estos sí; luego observamos los valores que están fuera de este rango.

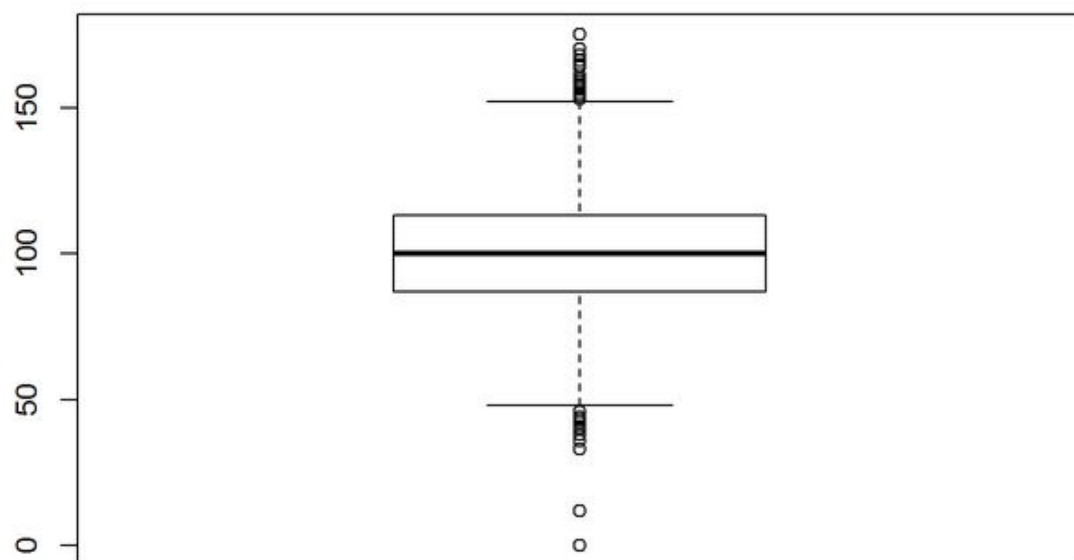
```
boxplot.stats(data$minutos_noche)$out
```

```
## [1] 57.5 354.9 349.2 345.8 45.0 342.8 364.3 63.3 54.5 50.1 43.7  
## [12] 349.7 352.5 23.2 63.6 381.9 377.5 367.7 56.6 54.0 64.2 344.3  
## [23] 395.0 350.2 50.1 53.3 352.2 364.9 61.4 47.4 381.6 50.9 46.7  
## [34] 359.9 65.2 59.5 0.0 355.1 60.3
```

COLUMNA LLAMADAS_NOCHE

Graficamos con el comando boxplot y observamos valores fuera del bigote superior e inferior.

```
boxplot(data$llamadas_noche)
```



Calculamos los cuartiles

```
quantile(data$llamadas_noche)
```

```
## 0% 25% 50% 75% 100%  
## 0 87 100 113 175
```

Calculamos el rango intercuartil

```
IQR(data$llamadas_noche)
```

```
## [1] 26
```

Calculamos el valor mínimo y máximo del nuevo rango para comparar con los valores extremos.

```
MIN<-(87-1.5*26)
```

```
MAX<-(87+1.5*26)
```

Imprimimos los valores mínimos y máximos.

```
range(MIN,MAX)
```

```
## [1] 48 126
```

Estos datos nos indican que si los valores están entre 48 y 126 no son valores atípicos, fuera de estos sí; luego observamos los valores que están fuera de este rango.

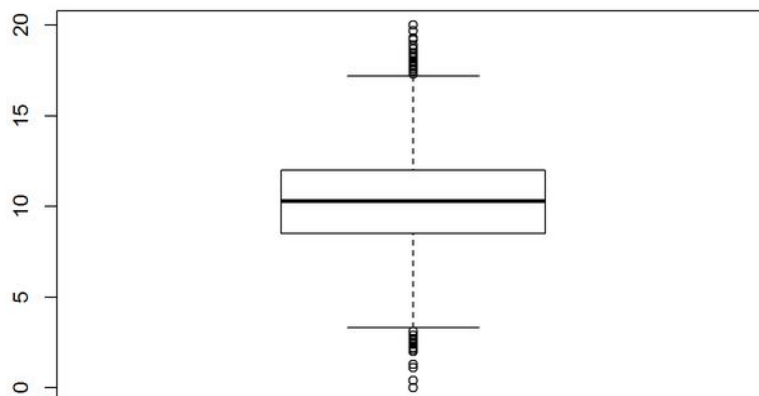
```
boxplot.stats(data$llamadas_noche)$out
```

```
## [1] 46 42 44 42 153 175 154 158 155 157 157 154 153 166 33 155 156  
## [18] 38 36 156 164 153 40 168 161 159 160 170 158 154 42 41 159 38  
## [35] 46 155 42 12 46 165 43 155 0
```

COLUMNA MINUTOS_INTERNACIONALES

Graficamos con el comando boxplot y observamos valores fuera del Bigote superior e inferior.

```
boxplot(data$minutos_internacionales)
```



Calculamos los cuartiles.

```
quantile(data$minutos_internacionales)
```

```
##    0%   25%   50%   75%  100%  
##   0.0   8.5  10.3  12.0  20.0
```

Calculamos el rango intercuartil

```
IQR(data$minutos_internacionales)
```

```
## [1] 3.5
```

Calculamos el valor mínimo y máximo del nuevo rango para comparar con los valores extremos.

```
MIN<-(8.5-1.5*3.5)  
MAX<-(8.5+1.5*3.5)
```

Imprimimos los valores mínimos y máximos.

```
range(MIN,MAX)
```

```
## [1] 3.25 13.75
```

Estos datos nos indican que si los valores están entre 3.25 y 13.75 no son valores atípicos, fuera de estos sí; luego observamos los valores que están fuera de este rango.

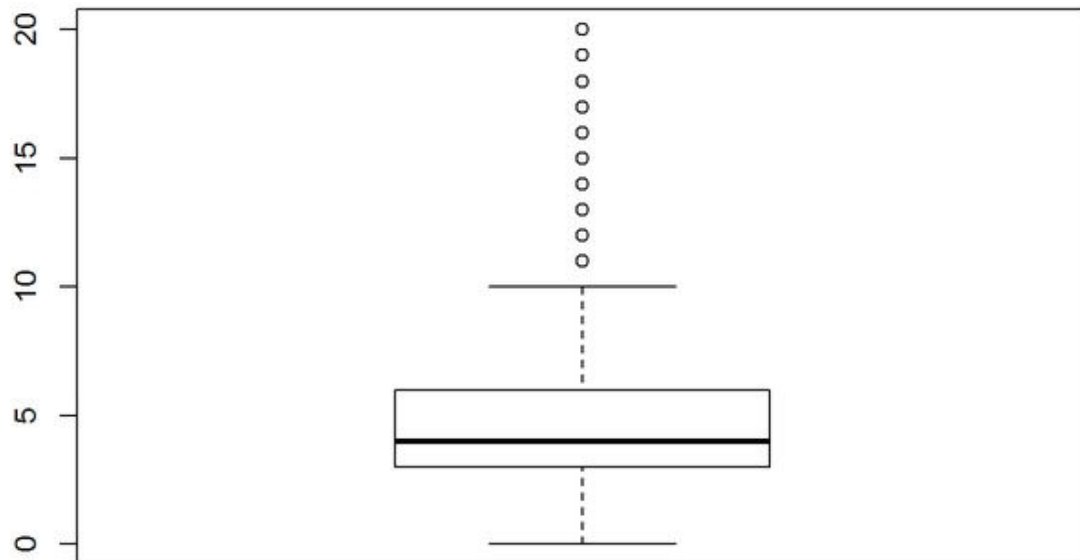
```
boxplot.stats(data$minutos_internacionales)$out
```

```
## [1] 20.0 0.0 17.6 2.7 18.9 0.0 18.0 2.0 0.0 17.5 17.5 18.2 0.0 0.0  
## [15] 1.3 0.0 0.0 0.0 2.2 18.0 0.0 17.9 0.0 17.3 17.3 18.4 2.0 17.8  
## [29] 2.9 3.1 17.6 17.3 2.6 0.0 0.0 18.2 0.0 18.0 1.1 0.0 18.3 0.0  
## [43] 0.0 2.1 2.9 17.5 2.1 2.4 2.5 0.0 0.0 17.8 18.9 0.0 18.7 0.0  
## [57] 3.1 0.4 19.3 19.2 2.2 0.0 19.7 0.0 0.0 17.8 18.5 0.0 1.1 2.0  
## [71] 17.7 19.7
```

COLUMNA LLAMADAS_INTERNACIONALES

Graficamos con el comando `boxplot` y observamos valores fuera del bigote superior.

```
boxplot(data$llamadas_internacionales)
```



Calculamos los cuartiles

```
quantile(data$llamadas_internacionales)
```

```
##    0%   25%   50%   75%  100%  
##     0     3     4     6    20
```

Calculamos el rango intercuartil

```
IQR(data$llamadas_internacionales)
```

```
## [1] 3
```

Calculamos el valor mínimo y máximo del nuevo rango para comparar con los valores extremos.

```
MIN<- (3-1.5*3)  
MAX<- (3+1.5*3)
```

Imprimimos los valores mínimos y máximos.

```
range(MIN,MAX)
```

```
## [1] -1.5  7.5
```

Estos datos nos indican que si los valores están entre -1.5 y 7.5 no son valores atípicos, fuera de estos sí; luego observamos los valores que están fuera de este rango.

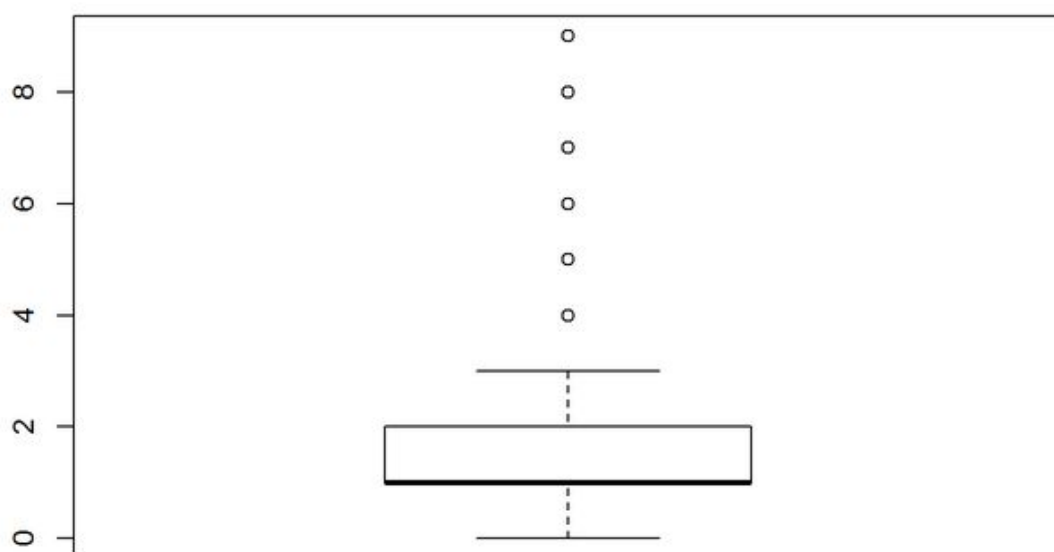
```
boxplot.stats(data$llamadas_internacionales)$out
```

```
## [1] 19 15 11 12 13 11 12 11 13 12 11 11 18 11 12 13 12 12 11 15 13 15 11
## [24] 11 14 13 11 13 13 12 11 14 15 18 12 13 11 14 11 12 14 15 12 11 16 11
## [47] 11 11 11 15 11 14 11 11 12 13 11 11 16 13 11 13 11 15 11 12 13 18 12
## [70] 12 12 11 13 11 13 14 20 17 13 12 19 13 12 16 11 16 16 11 11 12 15 16
## [93] 12 12 11 11 11 15 11 11 12 13 11 13 18 13 11 11 12 17 11 12 11 11 11
## [116] 11 11 16
```

COLUMNA RECLAMOS

Graficamos con el comando boxplot y observamos valores fuera del bigote superior.

```
boxplot(data$reclamos)
```



Calculamos los cuartiles

```
quantile(data$reclamos)
```

```
##    0%   25%   50%   75%  100%  
##     0     1     1     2     9
```

Calculamos el rango intercuartil

```
IQR(data$reclamos)
```

```
## [1] 1
```

Calculamos el valor mínimo y máximo del nuevo rango para comparar con los valores extremos.

```
MIN<-(1-1.5*1)  
MAX<-(1+1.5*1)
```

Imprimimos los valores mínimos y máximos.

```
range(MIN,MAX)
```

```
## [1] -0.5  2.5
```

Estos datos nos indican que si los valores están entre -0.5 y 2.5 no son valores atípicos, fuera de estos sí; luego observamos los valores que están fuera de este rango.

```
boxplot.stats(data$reclamos)$out
```

```
## [1] 4 4 4 5 5 5 4 4 4 4 4 4 4 4 4 4 5 5 4 5 4 4 5 4 4 4 4 4 5 4 4 7 4 4 4  
## [36] 4 4 5 4 4 4 4 4 5 4 7 4 9 5 4 4 5 4 4 5 5 4 6 4 6 5 5 5 6 5 4 4 5 4 4  
## [71] 7 4 6 5 4 4 4 6 4 4 5 4 4 4 4 4 5 5 6 5 4 4 4 5 4 4 4 4 5 5 4 4 4 4  
## [106] 6 4 5 4 6 4 4 4 4 4 4 4 4 6 4 4 4 4 8 4 4 5 4 4 4 6 5 5 7 4 4 5 4 4  
## [141] 5 4 4 5 7 4 4 5 7 4 4 4 4 8 6 4 4 5 5 5 4 4 5 4 4 4 4 4 4 4 4 4 4 5 6  
## [176] 4 5 4 4 5 5 4 6 4 4 4 9 6 4 5 5 4 6 4 4 5 4 4 4 5 5 6 4 5 4 4 4 4 5 4  
## [211] 4 4 5 4 5 6 4 4 5 4 4 4 5 4 4 4 4 4 5 7 6 5 6 7 5 5 4 6 4 4 4 4 5 6 7  
## [246] 4 4 4 5 5 5 4 4 4 5 6 5 5 4 4 4 4 4 4 4 4 5 4 4 5 5 4 4 5 4 5 4 4 4 5  
## [281] 5 4 4 6 6 4 5 5 4 4 5 4 5 4 5 4 4 4 4 4 4 4 4 4 4 4 6 4 4 4 4 5 4 4 4 5  
## [316] 5 4 4 5 5 5 4 4 7 4 4 5 5 5 6 4 4 4 4 4 4 4 4 4 4 4 4 7 7 4 6 4 4 4 4 4  
## [351] 4 4 4 4 7 5 4 4 4 5 4 4 6 4 4 5 4 4 6 5 5 6 5 6 4 4 4 4 4 5 4 4 5 4 6  
## [386] 4 6 4 5 4 4 4 6 4 4 4 4 4 5
```


En todas las variables observamos que existen datos fuera del rango MIN y MAX; pero los valores no los podemos considerar datos extremos, por ejemplo observamos que estos valores pueden darse en la columna minutos_día; ya que observamos que apenas hay 2 usuarios de 5000 que no registran minutos al día, y también hay usuarios que han hablado hasta 350 minutos al día: lo que corresponde a hablar 6 horas diarias aproximadamente (6 horas multiplicado por 60 minutos = 360 minutos), y esto es real en los usuarios más jóvenes en la telefonía celular, además de NO sobrepasar los 720 minutos que corresponden a las 12 horas en el día.

Sin embargo si tuviéramos un valor mayor a 720 minutos, ese dato correspondería a un valor extremo para esta columna.

En la columna llamadas_día vemos que estos valores también pueden darse, ya que pueden haber personas que hagan más de 126 llamadas en el día considerando su situación laboral, el valor máximo de llamadas es 165 y si lo multiplicamos por 2.5 (que serían el tiempo en minutos que duraría cada llamada) da como resultado 412.5 minutos; un valor superior a los 350 minutos que una persona hablaría según el cálculo de la variable anterior.

En las otras variables al tratarse de minutos y llamadas sucede el mismo caso, por lo que tampoco lo consideramos como valores extremos a aquellos datos que salen del rango de MIN y MAX.

Y finalmente la columna reclamos, vemos que aunque no es muy normal que los usuarios hayan hecho hasta 9 llamadas al servicio de llamadas por reclamos, sin embargo esta situación puede darse, ya que pueden existir clientes muy inconformes ya sea por tratarse de que el servicio se ha caído o por anular algún otro servicio o presentar alguna otra inconformidad.

4. ANALISIS DE LOS DATOS.

4.1. SELECCIÓN DE LOS GRUPOS DE DATOS QUE SE QUIEREN ANALIZAR/COMPARAR (PLANIFICACION DE LOS ANALISIS A APLICAR)

A continuación, se seleccionan los grupos dentro de nuestro conjunto de datos que pueden resultar interesantes para analizar y/o comparar.

```
# Agrupacion por plan internacional
grupol_plan <- data[, c(1,2)]
data_sinplan <- grupol_plan[grupol_plan$plan_internacional == "no",]
data_plan_int <- grupol_plan[grupol_plan$plan_internacional == "yes",]
```

El comando summary nos muestra los resultados de cada grupo del plan internacional.

```
summary(data$plan_internacional)
```

```
##    no  yes  
## 4527  473
```

```
# Agrupacion por churn (abandona)  
grupo2_churn <- data[, c(1,2)]  
data.NOabandona <- grupo2_churn[grupo2_churn$abandona == "No",]  
data.SIabandona <- grupo2_churn[grupo2_churn$abandona == "Yes",]
```

El comando summary nos muestra los resultados de cada grupo de la variable abandona.

```
summary(data$abandona)
```

```
##    No  Yes  
## 4293  707
```

Para agrupar por churn (abandona) y plan internacional; utilizaremos el comando table.

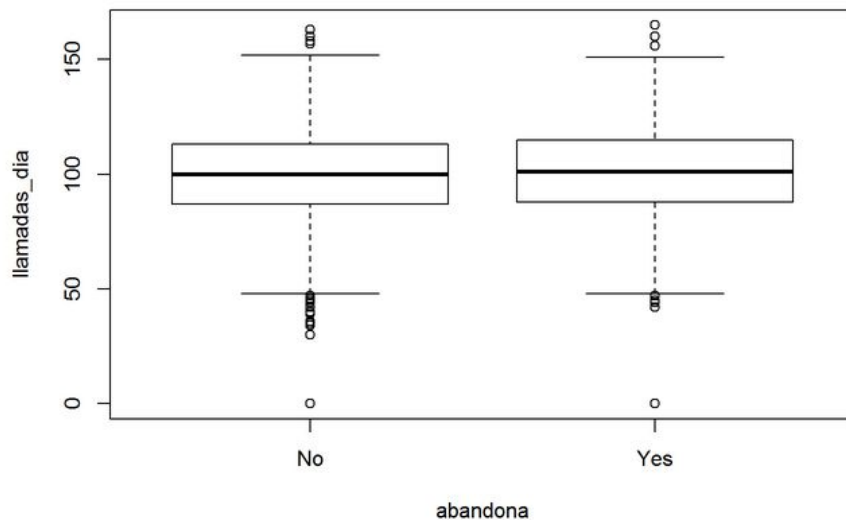
```
table(data$abandona,data$plan_internacional)
```

```
##  
##           no  yes  
##    No  4019  274  
##    Yes   508  199
```

Para agrupar por llamadas al dia y churn utilizamos un gráfico, para facilitar la comprensión de los datos.

```
# Agrupacion por llamadas al dia  
data2 <- data[, c(1,4)]  
plot(data2)
```

4.2.



COMPROBACION DE LA NORMALIDAD Y HOMOGENEIDAD DE LA VARIANZA.

Para la comprobación de que los valores que toman nuestras variables cuantitativas provienen de una población distribuida normalmente, utilizaremos la prueba de normalidad de Anderson-Darling.

Así, se comprueba que para cada prueba se obtiene un p-valor superior al nivel de significación prefijado $\alpha = 0,05$ o $\alpha = 0,10$. Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal.

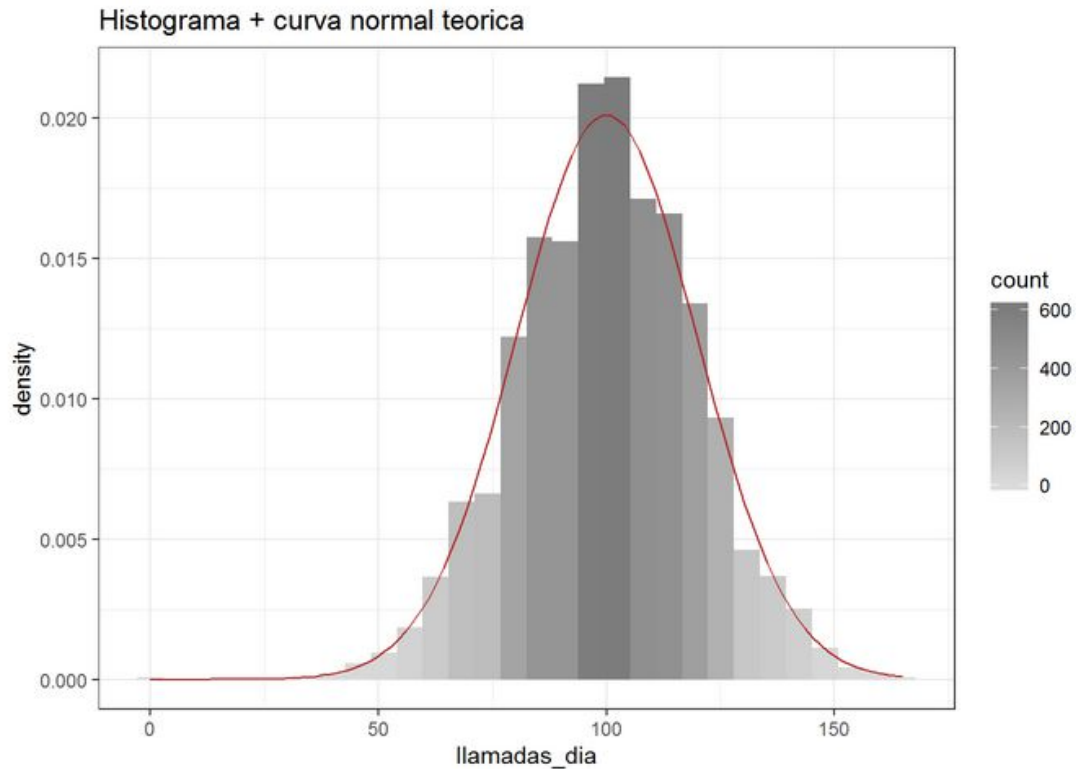
```
library(nortest)
alpha = 0.05
col.names = colnames(data)
for (i in 1:ncol(data)) {
  if (i == 1) cat("Variables que no siguen una distribucion normal:\n")
  if (is.integer(data[,i]) | is.numeric(data[,i])) {
    p_val = ad.test(data[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      # Format output
      if (i < ncol(data) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}
```

```
## Variables que no siguen una distribucion normal:
## llamadas_dia, llamadas_noche,
## minutos_internacionales, llamadas_internacionalesreclamos
```

VEAMOS LA GRAFICA DE LA COLUMNA LLAMADAS_DIA

EN ESTE EJERCICIO CONSIDERAMOS QUE LA HIPOTESIS NULA ES: H_0 = "La muestra proviene de una población con distribución normal".

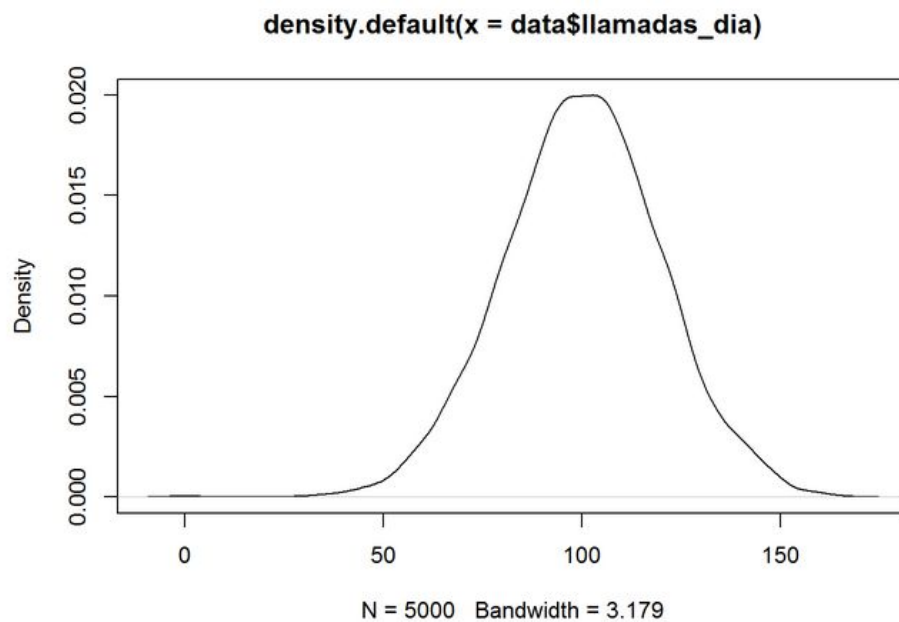
Observemos el Histograma + la Curva Normal teórica



```
library(ggplot2)
ggplot(data = data, aes(x = llamadas_dia)) +
  geom_histogram(aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
               args = list(mean = mean(data$llamadas_dia),
                           sd = sd(data$llamadas_dia))) +
  ggtitle("Histograma + curva normal teorica") +
  theme_bw()
```

Ahora podemos observar la curva de la variable llamadas_dia que no sigue una distribución normal.

```
plot(density(data$llamadas_dia))
```



Finalmente aplicamos la prueba de Anderson – Darling

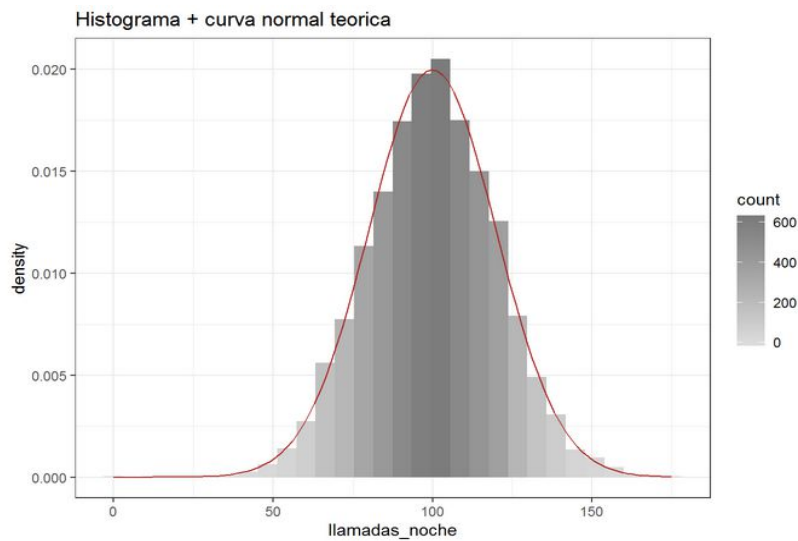
```
ad.test(data$llamadas_dia)$p.value
```

```
## [1] 0.01729142
```

A pesar de que en la gráfica observamos una figura bastante simétrica, vemos que al aplicar la prueba de Anderson - Darling el valor que obtenemos es menor con relación al 5% de referencia, por lo tanto nuestra conclusión es que rechazamos la hipótesis nula (H_0).

VEAMOS LA GRAFICA DE LA COLUMNA LLAMADAS_NOCHE

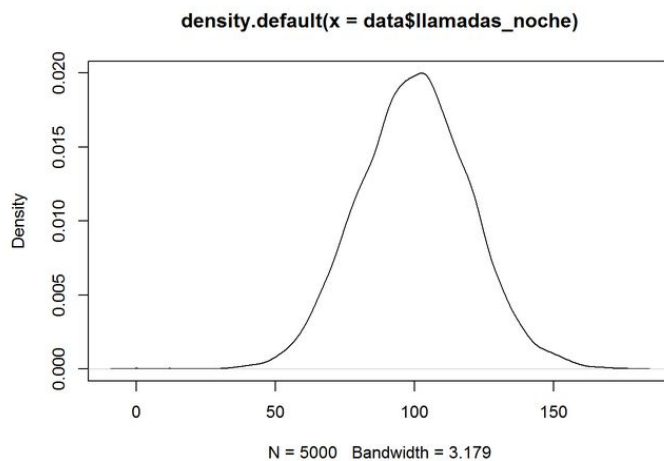
Observemos el Histograma + la Curva Normal teórica



```
library(ggplot2)
ggplot(data = data, aes(x = llamadas_noche)) +
  geom_histogram(aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
               args = list(mean = mean(data$llamadas_noche),
                           sd = sd(data$llamadas_noche))) +
  ggtitle("Histograma + curva normal teorica") +
  theme_bw()
```

Ahora podemos observar la curva de la variable llamadas_noche que no sigue una distribución normal.

```
plot(density(data$llamadas_noche))
```



Aplicamos la prueba de Anderson - Darling

```
ad.test(data$llamadas_noche)$p.value
```

```
## [1] 0.03766821
```

Como el valor obtenido aplicando la prueba de Anderson-Darling se acerca a 5%, ya que nos da un resultado de 3.7% aplicamos la prueba de Kolmogorov-Smirnov con el comando Lillie para comparar y verificar los valores de la probabilidad.

```
lillie.test(data$llamadas_noche)$p.value
```

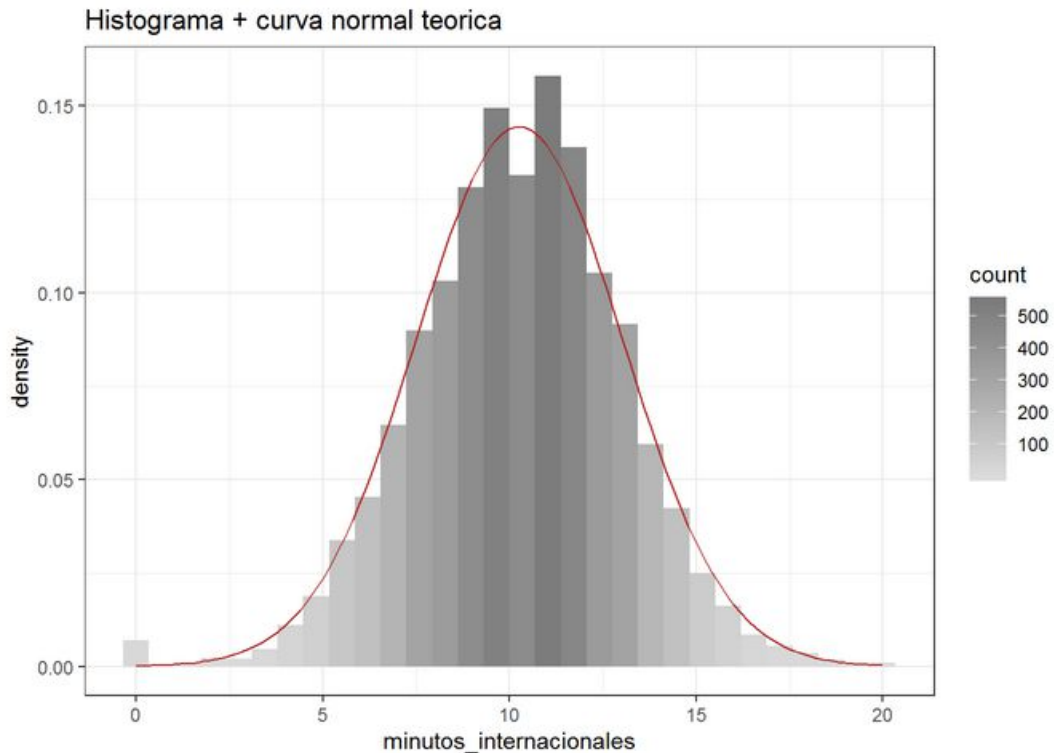
```
## [1] 0.004817295
```

Con esta otra prueba no tenemos duda, tenemos un valor bien por debajo del 5%.

A pesar de que en la gráfica observamos una figura bastante simétrica, vemos que al aplicar la prueba de Anderson - Darling y la prueba de Kolmogorov-Smirnov el valor que obtenemos es menor con relación al 5% de referencia, por lo tanto nuestra conclusión es que rechazamos la hipótesis nula (H_0).

VEAMOS LA GRAFICA DE LA COLUMNA MINUTOS_INTERNACIONALES

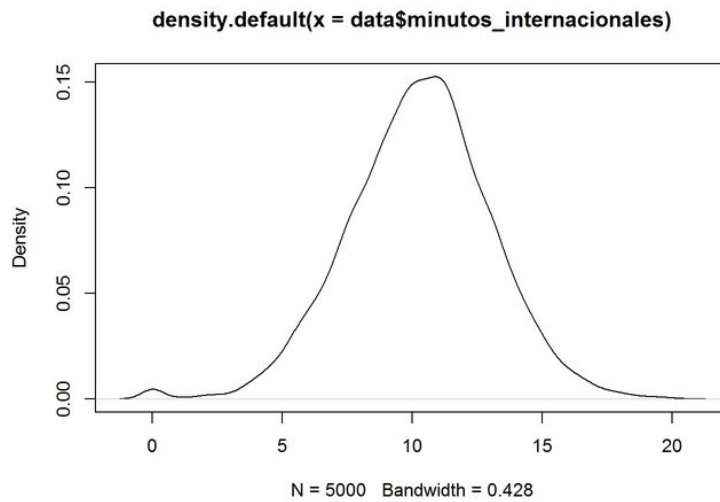
Observemos el Histograma + la Curva Normal teórica



```
library(ggplot2)
ggplot(data = data, aes(x = minutos_internacionales)) +
  geom_histogram(aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
               args = list(mean = mean(data$minutos_internacionales),
                           sd = sd(data$minutos_internacionales))) +
  ggtitle("Histograma + curva normal teorica") +
  theme_bw()
```

Ahora podemos observar la curva de la variable minutos_internacionales que no sigue una distribución normal.

```
plot(density(data$minutos_internacionales))
```



Finalmente aplicamos la prueba de Anderson – Darling.

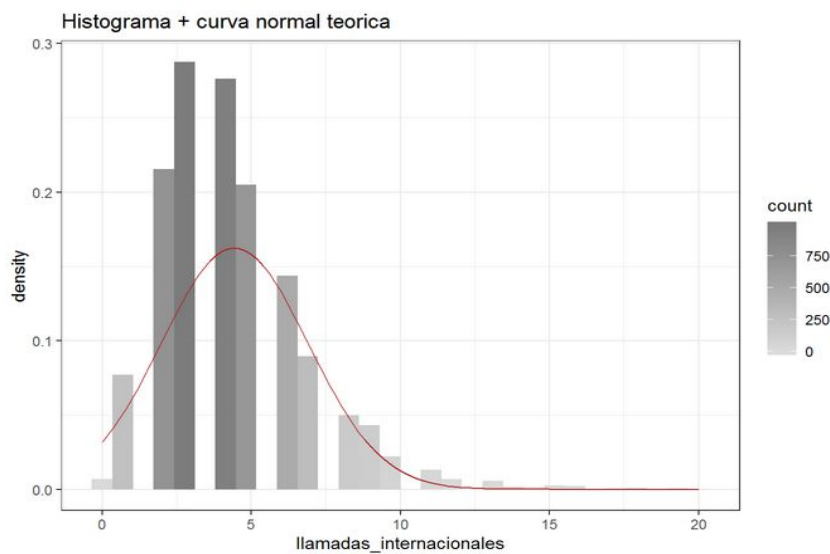
```
ad.test(data$minutos_internacionales)$p.value
```

```
## [1] 9.278932e-09
```

En la gráfica observamos una figura un poco simétrica, vemos también que al aplicar la prueba de Anderson - Darling el valor que obtenemos es bastante menor con relación al 5% de referencia, por lo tanto nuestra conclusión es que rechazamos la hipótesis nula (H_0).

VEAMOS LA GRAFICA DE LA COLUMNA LLAMADAS_INTERNACIONALES.

Observemos el Histograma + la Curva Normal teórica

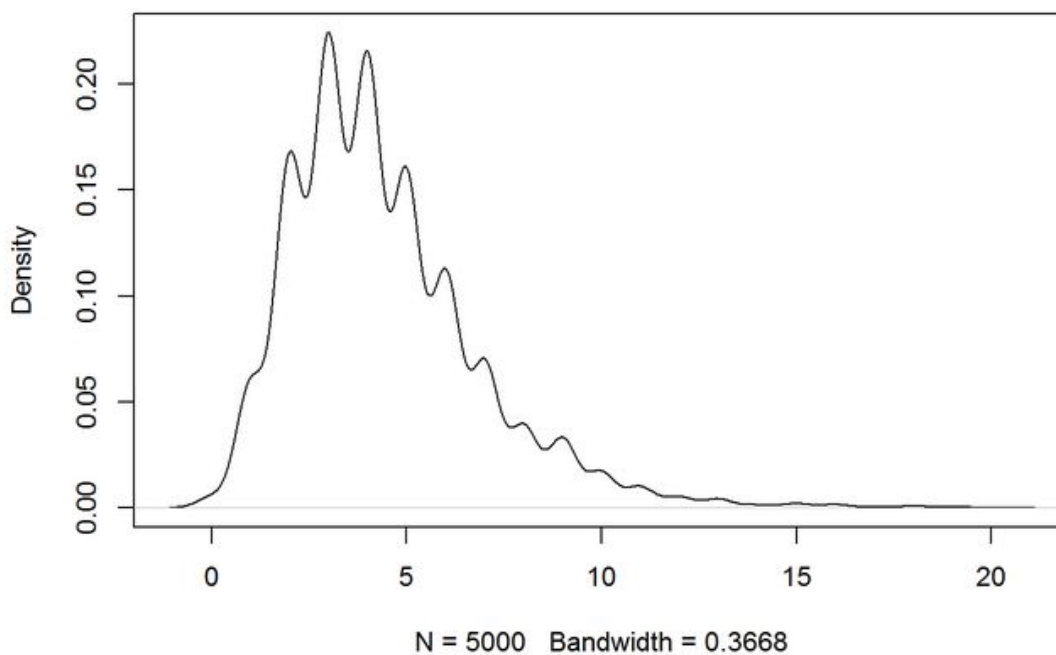



```
library(ggplot2)
ggplot(data = data, aes(x = llamadas_internacionales)) +
  geom_histogram(aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
               args = list(mean = mean(data$llamadas_internacionales),
                           sd = sd(data$llamadas_internacionales))) +
  ggtitle("Histograma + curva normal teorica") +
  theme_bw()
```

Ahora podemos observar la curva de la variable llamadas_internacionales que no sigue una distribución normal.

```
plot(density(data$llamadas_internacionales))
```

density.default(x = data\$llamadas_internacionales)



Aplicamos la prueba de Anderson – Darling.

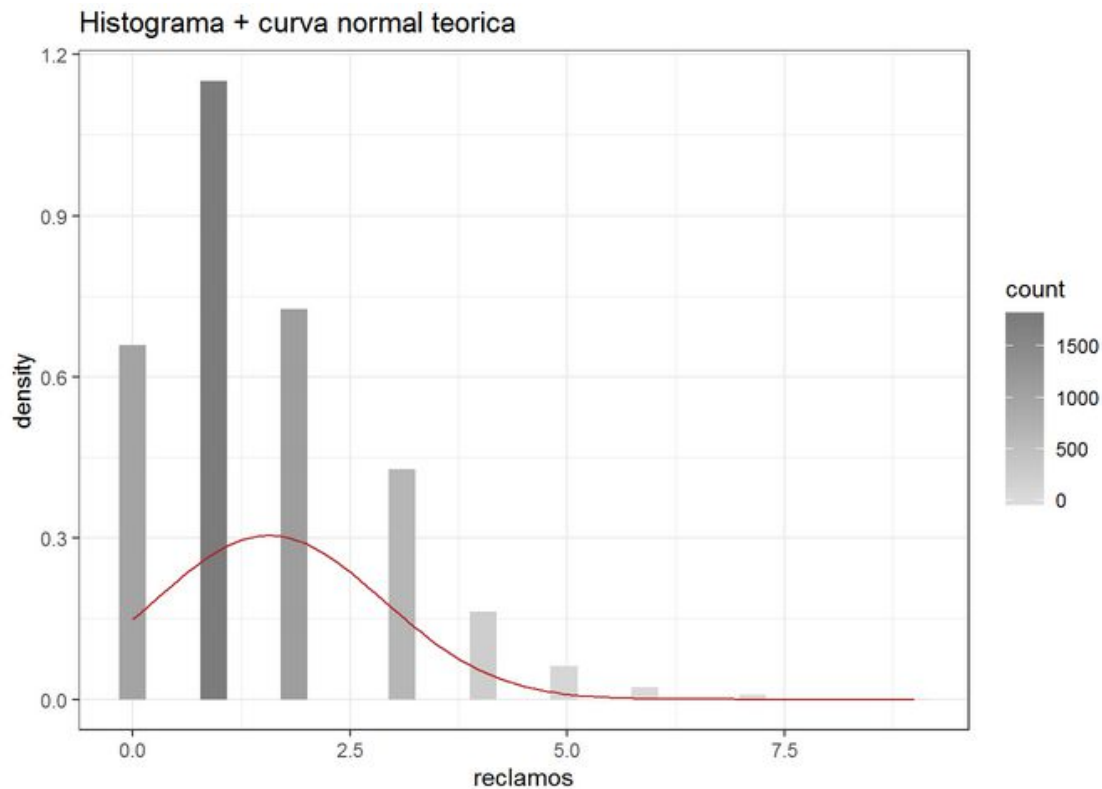
```
ad.test(data$llamadas_internacionales)$p.value
```

```
## [1] 3.7e-24
```

En la gráfica observamos una figura ASIMETRICA, vemos también que al aplicar la prueba de Anderson - Darling el valor que obtenemos es bastante menor con relación al 5% de referencia, por lo tanto nuestra conclusión es que rechazamos la hipótesis nula (H_0).

VEAMOS LA GRAFICA DE LA COLUMNA RECLAMOS

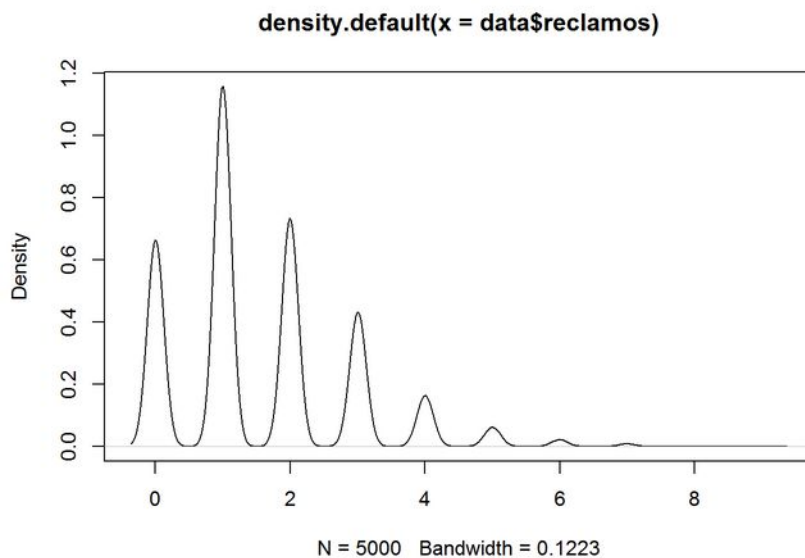
Observemos el Histograma + la Curva Normal teórica



```
library(ggplot2)
ggplot(data = data, aes(x = reclamos)) +
  geom_histogram(aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
               args = list(mean = mean(data$reclamos),
                           sd = sd(data$reclamos))) +
  ggtitle("Histograma + curva normal teorica") +
  theme_bw()
```

Ahora podemos observar la curva de la variable reclamos que no sigue una distribución normal.

```
plot(density(data$reclamos))
```



Finalmente aplicamos la prueba de Anderson – Darling.

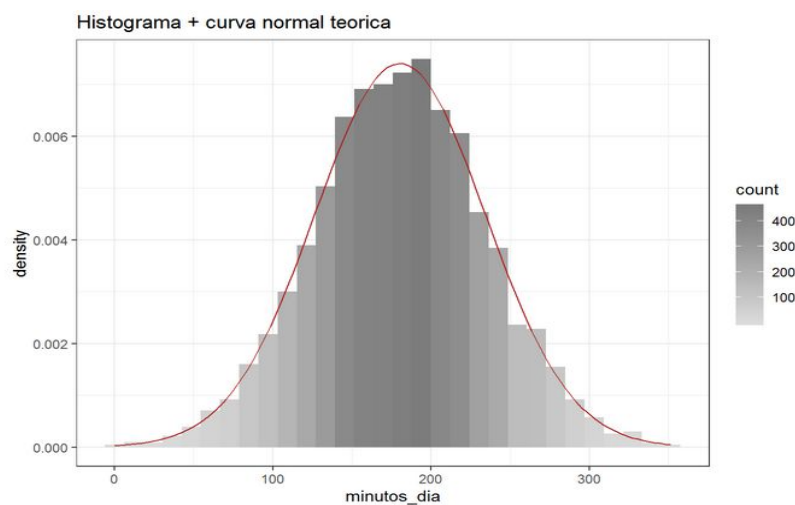
```
ad.test(data$reclamos)$p.value
```

```
## [1] 3.7e-24
```

En la gráfica observamos una figura ASIMETRICA, vemos también que al aplicar la prueba de Anderson - Darling el valor que obtenemos es bastante menor con relación al 5% de referencia, por lo tanto nuestra conclusión es que rechazamos la hipótesis nula (H_0).

GRAFICA DE LA COLUMNA MINUTOS_DIA

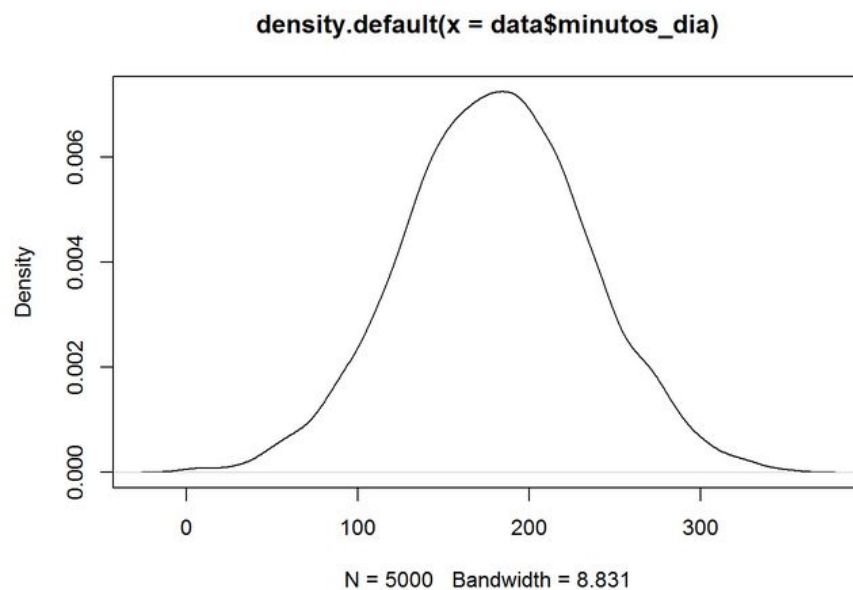
Observemos el Histograma + la Curva Normal teórica



```
library(ggplot2)
ggplot(data = data, aes(x = minutos_dia)) +
  geom_histogram(aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
               args = list(mean = mean(data$minutos_dia),
                           sd = sd(data$minutos_dia))) +
  ggtitle("Histograma + curva normal teorica") +
  theme_bw()
```

Ahora podemos observar la curva de la variable minutos_dia que sigue una distribución normal.

```
plot(density(data$minutos_dia))
```



Aplicamos la prueba de Anderson – Darling.

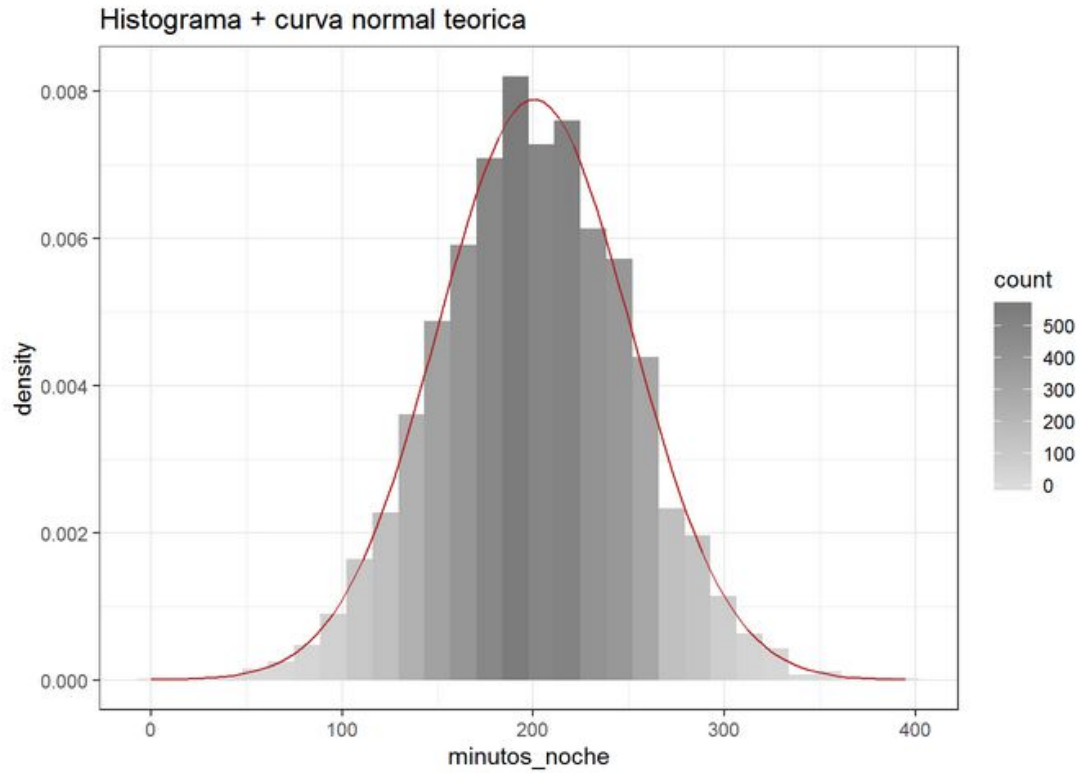
```
ad.test(data$minutos_dia)$p.value
```

```
## [1] 0.8813471
```

En la gráfica observamos una figura bastante simétrica, vemos también que al aplicar la prueba de Anderson - Darling el valor que obtenemos es bastante mayor con un 88% con relación al 5% de referencia, por lo tanto nuestra conclusión es que NO rechazamos la hipótesis nula (H_0), pues la rechazaremos siempre que sea menor al valor referencial (5%).

GRAFICA DE LA COLUMNA MINUTOS_NOCHE

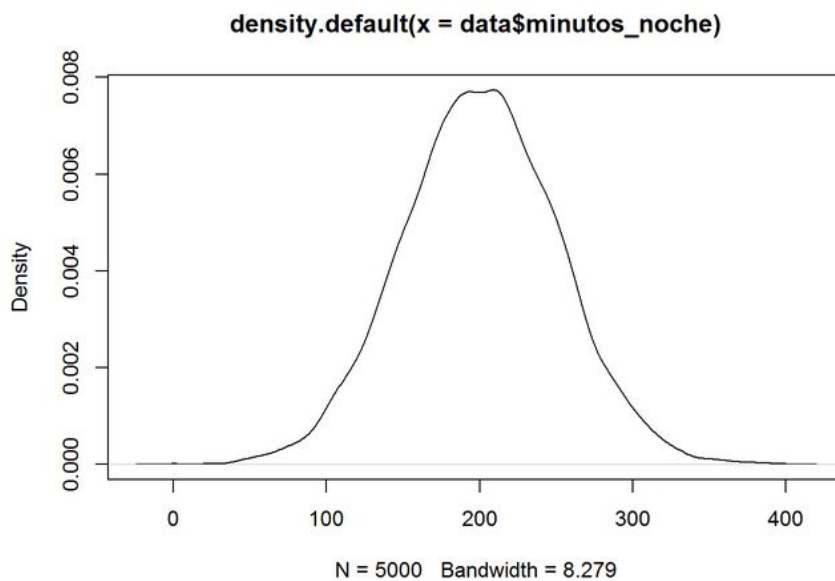
Observemos el Histograma + la Curva Normal teórica



```
library(ggplot2)
ggplot(data = data, aes(x = minutos_noche)) +
  geom_histogram(aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
               args = list(mean = mean(data$minutos_noche),
                           sd = sd(data$minutos_noche))) +
  ggtitle("Histograma + curva normal teorica") +
  theme_bw()
```

Ahora podemos observar la curva de la variable minutos_noche que sigue una distribución normal.

```
plot(density(data$minutos_noche))
```



Aplicamos la prueba de Anderson – Darling.

```
ad.test(data$minutos_noche)$p.value
```

```
## [1] 0.9031988
```

En la gráfica observamos una figura bastante simétrica, vemos también que al aplicar la prueba de Anderson - Darling el valor que obtenemos es bastante mayor con un 90% con relación al 5% de referencia, por lo tanto nuestra conclusión es que NO rechazamos la hipótesis nula (H_0), es decir que aceptamos que la muestra proviene de una distribución normal.

PRUEBA DE HOMOGENEIDAD

(H_0 = “Las varianzas poblacionales son iguales”)

Primero aplicamos Test de Bartlett que permite contrastar la igualdad de varianza en 2 o más poblaciones sin necesidad de que el tamaño de los grupos sea el mismo. Es más sensible que el test de Levene a la falta de normalidad.

VARIANZA POR TEST DE BARTLETT ENTRE MINUTOS_DIA Y ABANDONA.

```
bartlett.test(minutos_dia ~ abandona, data = data)
```

```
##  
## Bartlett test of homogeneity of variances  
##  
## data: minutos_dia by abandona  
## Bartlett's K-squared = 147.2, df = 1, p-value < 2.2e-16
```

Luego vamos a ver la Varianza por test de Fligner-Killeen.

```
fligner.test(minutos_dia ~ abandona, data = data)
```

```
##  
## Fligner-Killeen test of homogeneity of variances  
##  
## data: minutos_dia by abandona  
## Fligner-Killeen:med chi-squared = 228.9, df = 1, p-value < 2.2e-16
```

Y finalmente para estas dos columnas probamos la Varianza por el test de Levene.

Se requiere instalar la librería lawstat.

```
library(lawstat)
```

```
## Warning: package 'lawstat' was built under R version 3.6.2
```

```
levene.test(data$minutos_dia, data$abandona)
```

```
##  
## Modified robust Brown-Forsythe Levene-type test based on the  
## absolute deviations from the median  
##  
## data: data$minutos_dia  
## Test Statistic = 240.64, p-value < 2.2e-16
```

En los tres casos el valor de la probabilidad (p-value) es un número menor a 0.05 (5%); por lo que rechazamos la hipótesis nula. (Exactamente el p-value en decimal es 0.000000000000000022)

VARIANZA ENTRE RECLAMOS Y ABANDONO POR EL TEST DE LEVENE.

```
library(lawstat)  
levene.test(data$reclamos, data$abandona)
```

```
##  
## Modified robust Brown-Forsythe Levene-type test based on the  
## absolute deviations from the median  
##  
## data: data$reclamos  
## Test Statistic = 256.48, p-value < 2.2e-16
```

También observamos que el p-value es un número muy pequeño, menor al 5% referencia; por lo que rechazamos la hipótesis nula.

VARIANZA POR TEST DE BARTLETT ENTRE LLAMADAS_DIA Y ABANDONA.

```
bartlett.test(llamadas_dia ~ abandona, data = data)

##
## Bartlett test of homogeneity of variances
##
## data:  llamadas_dia by abandona
## Bartlett's K-squared = 3.5789, df = 1, p-value = 0.05852
```

En vista que el p-value es igual al 5% y un poco mayor de hecho; NO rechazamos la hipótesis nula, por lo tanto concluimos de que existe homogeneidad de varianzas, es decir que aceptamos que la varianza de ambas muestras son iguales.

4.3. APLICACIÓN DE PRUEBAS ESTADISTICAS PARA COMPARAR LOS GRUPOS DE DATOS. EN FUNCION DE LOS DATOS Y EL OBJETIVO DEL ESTUDIO, APLICAR PRUEBAS DE CONTRASTE DE HIPOTESIS, CORRELACIONES, REGRESIONES, ETC. APLICAR AL MENOS TRES METODOS DE ANALISIS DIFERENTES.

PRIMERA PRUEBA ESTADISTICA:

PRUEBA DE CONTRASTE DE HIPOTESIS MEDIANTE LA COMPARACION DE LA TENDENCIA CENTRAL.

Análisis de relación entre una variable cuantitativa y otra cualitativa.
Tenemos dos variables que vamos a comparar:

X= ABANDONA (cualitativa)
Y= LLAMADAS DIA (cuantitativa)

Veamos la estructura de los datos de las dos variables a comparar.

```
data2 <- data[, c(1,4)]
summary(data2)
```

```
##  abandona    llamadas_dia
## No :4293   Min.    :  0
## Yes: 707   1st Qu.: 87
##           Median :100
##           Mean   :100
##           3rd Qu.:113
##           Max.   :165
```


Veamos también los primeros 6 registros de ambas columnas.

```
head(data2)
```

```
##   abandona llamadas_dia
## 1      No         110
## 2      No         123
## 3      No         114
## 4      No          71
## 5      No         113
## 6      No          98
```

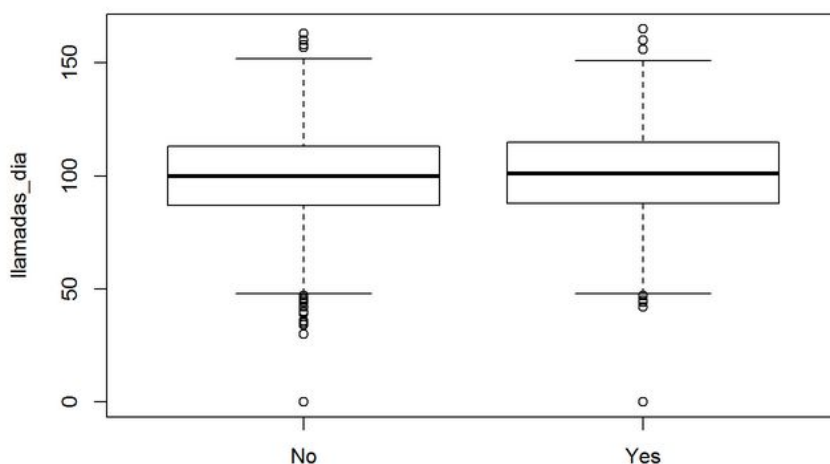
Partimos de la hipótesis de que ambas variables son independientes, formulado en una pregunta quedaría:

¿Los clientes que abandonan la compañía son independientes al número de llamadas que realizan diariamente?

Para determinar si estas dos variables (cualitativa y cuantitativa) están relacionadas o son independientes; analizaremos sus medias, es decir que si sabemos que la media o la mediana de las llamadas que realiza el cliente son iguales para quienes abandonen o no abandonen la compañía; podemos afirmar que ambas variables son independientes, es decir que quienes abandonan la compañía no dependen del número de llamadas que realizan; si por el contrario se prueba que las medias o las medianas son diferentes podemos concluir que ambas variables están relacionadas, y que por lo tanto los clientes que abandonan la compañía dependen del número de llamadas telefónicas que realizan a diario.

Veamos una gráfica.

```
plot(data2)
```



Podemos ver que la media entre los clientes que abandonan la compañía y la media de los clientes que no abandonan la compañía es la misma; de modo que NO rechazamos la hipótesis nula (H_0 = la media del primer grupo es igual a la del segundo grupo); por lo tanto concluimos que quienes abandonan la compañía NO dependen de las llamadas telefónicas diarias.

Aplicamos una prueba T-STUDENT para comprobar la igualdad de las medias de los dos grupos.

```
t.test(llamadas_dia ~ abandona, data=data, var.equal = TRUE)
```

```
##
## Two Sample t-test
##
## data: llamadas_dia by abandona
## t = -1.1064, df = 4998, p-value = 0.2686
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -2.4676379 0.6871328
## sample estimates:
## mean in group No mean in group Yes
##          99.90753          100.79778
```

Observamos que la prueba T-STUDENT nos permite conocer la media de los dos grupos, donde confirmamos que son iguales, además el p-value del 26% que es superior al 5% de referencia nos indica que las muestras si son iguales, de modo que no rechazamos la hipótesis nula.

* SEGUNDA PRUEBA ESTADISTICA: PRUEBA DEL CHI CUADRADO

La prueba de Chi-cuadrado en R es un método estadístico que se utiliza para determinar si dos variables categóricas tienen una correlación significativa entre ellas.

Su fórmula estadística está dada por:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Dónde: O= se refiere a las frecuencias observadas E= frecuencias esperadas.

Las dos variables categóricas que vamos a comparar son: Abandona y plan internacional.

Hipótesis de partida= Las dos variables en estudio son independientes.

Observamos las columnas con las que vamos a trabajar en esta prueba estadística.

```
data2 <- data[, c(1,2)]
summary(data2)
```

```
## abandona plan_internacional
## No :4293 no :4527
## Yes: 707 yes: 473
```

Con la función `summary` encontramos un resumen de los datos de ambas columnas, donde tenemos que en la columna “abandona” 4293 clientes NO abandonan la compañía y 707 si lo hacen.

En la columna plan internacional que 4527 personas NO tienen plan internacional y 473 si lo tienen, cabe indicar que en este resumen las columnas no están relacionadas.

Ahora veamos una tabla de frecuencias con los datos de ambas columnas relacionados, estos datos serían las frecuencias observadas.

TABLA DE FRECUENCIAS OBSERVADAS. (Se ha agregado las frecuencias marginales para las filas y columnas con el comando `addmargins`)

```
table(data2)
```

```
##          plan_internacional
## abandona  no  yes
##      No 4019 274
##      Yes  508 199
```

```
tabla<-table(data2)
addmargins(tabla)
```

```
##          plan_internacional
## abandona  no  yes  Sum
##      No 4019 274 4293
##      Yes  508 199 707
##      Sum 4527 473 5000
```

Los datos obtenidos en esta tabla nos dice que 4019 clientes que no tienen plan internacional NO abandonan la compañía, y 274 clientes que registraban plan internacional tampoco abandonaron la compañía.

Por el contrario 508 clientes que no tenían plan internacional SI abandonaron la compañía y 199 usuarios que tenían plan internacional también abandonaron la compañía.

Ahora podemos observar las frecuencias esperadas, redondeadas a 2 decimales.

TABLA DE FRECUENCIAS ESPERADAS.

```
esperados<-chisq.test(tabla)$expected
b<-round(esperados,2)
b
```

```
##      plan_internacional
## abandona      no      yes
##      No  3886.88 406.12
##      Yes   640.12  66.88
```

Presentamos los datos de la tabla de frecuencias observadas en porcentajes, redondeados a 3 decimales.

```
porc<-(prop.table(tabla))
a<-round(porc,3)
a
```

```
##      plan_internacional
## abandona      no      yes
##      No  0.804 0.055
##      Yes 0.102 0.040
```

APLICAMOS LA PRUEBA ESTADÍSTICA DEL CHI CUADRADO.

```
chisq.test(x = tabla)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tabla
## X-squared = 333.19, df = 1, p-value < 2.2e-16
```

OBTENEMOS COMO RESPUESTA A LA PRUEBA DEL CHI CUADRADO EL VALOR DE: 333.19.

Ahora para aceptar o rechazar la hipótesis nula (H_0 = ambas variables son independientes) obtenemos el valor crítico tabular de chi-cuadrado, al 95% con 1 grado de libertad.

```
qchisq(0.95,1)
```

```
## [1] 3.841459
```

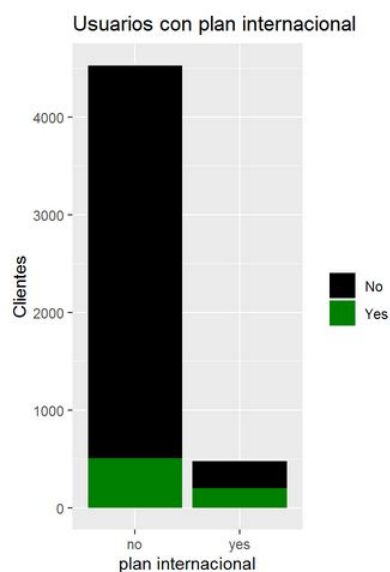
La respuesta del valor teórico con un nivel de riesgo del 5% y con 1 grado de libertad es: 3.84.

CONCLUSION: Nuestro valor experimental (333.19) es mayor al valor teórico (3.84), por lo que rechazamos la hipótesis nula y asumimos la hipótesis alternativa, lo que quiere decir que ambas variables están relacionadas.

- Ahora veamos un gráfico de barras, para visualizar la relación entre la variable plan internacional y la variable abandona. (En la primera barra observamos la relación inversa y en la segunda barra la relación directa).

```
if(!require(ggplot2)){  
  install.packages('ggplot2', repos='http://cran.us.r-project.org')  
  library(ggplot2)  
}  
if(!require(grid)){  
  install.packages('grid', repos='http://cran.us.r-project.org')  
  library(grid)  
}  
if(!require(gridExtra)){  
  install.packages('gridExtra', repos='http://cran.us.r-project.org')  
  library(gridExtra)  
}
```

```
grid.newpage()  
plotbyplan1<-ggplot(data,aes(plan_internacional,fill=abandona))+geom_bar() +labs(x="plan internacional", y="Clientes") +  
guides(fill=guide_legend(title=""))+ scale_fill_manual(values=c("black", "#008000"))+ggtitle("Usuarios con plan internacional")  
grid.arrange(plotbyplan1,ncol=2)
```



En el gráfico de barras y la información de la tabla de frecuencias observadas; vemos que la gran mayoría de usuarios no tienen plan internacional, son exactamente 4527 usuarios que no tienen plan internacional, de los cuales solo 508 abandonan la compañía pero 4019 no lo hicieron; por otro lado vemos que existen 473 usuarios que tienen plan internacional y 274 no abandonaron la compañía y 199 de ellos sí lo hicieron.

TERCER METODO DE ANALISIS.

Árboles de clasificación y regresión.

Modelo de clasificación. Árbol de decisión C5.0

Nuestro objetivo es crear un Árbol de decisión que permita analizar qué tipo de usuarios de nuestra data ha tenido probabilidades de abandonar la compañía. Por lo tanto, la variable por la que clasificaremos es el campo "abandona".

Para la futura evaluación del Árbol de decisión, es necesario dividir el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba.

Seleccionamos la columna 1 donde se encuentra nuestra variable dependiente, y luego seleccionamos las demás columnas donde se encuentran las variables que se relacionan con la variable "abandona".

```
new_data<-data[, c(1, 2, 3, 7, 9)]
set.seed(666)
y <- new_data[,1]
X <- new_data[,2:5]
```

Observamos los nombres de las columnas.

```
names(new_data)
```

```
## [1] "abandona"                "plan_internacional"
## [3] "minutos_dia"             "minutos_internacionales"
## [5] "reclamos"
```

Calculamos a cuantas filas corresponde dos tercios de los datos ($2 \cdot 5000 / 3 = 3333$) y dividimos "manualmente" el conjunto.

```
trainX <- X[1:3333,]
trainy <- y[1:3333]
testX <- X[3334:5000,]
testy <- y[3334:5000]
```

CREACION DEL MODELO, CALIDAD DEL MODELO Y EXTRACCION DE REGLAS.

Se crea el Árbol de decisión usando los datos de entrenamiento:

```
model <- C50::C5.0(trainX, trainy, rules=TRUE )
summary(model)
```

```
##
## Call:
## C5.0.default(x = trainX, y = trainy, rules = TRUE)
##
##
## C5.0 [Release 2.07 GPL Edition]      Wed Jan 01 22:15:37 2020
## -----
##
## Class specified by attribute `outcome'
##
## Read 3333 cases (5 attributes) from undefined.data
##
## Rules:
##
## Rule 1: (2604/128, lift 1.1)
##   plan_internacional = no
##   minutos_dia <= 264.4
##   reclamos <= 3
##   -> class No   [0.950]
##
## Rule 2: (1015/152, lift 1.1)
## Rule 3: (1015/152, lift 1.1)
## Rule 4: (57, lift 6.8)
##   plan_internacional = yes
##   minutos_internacionales > 13.1
##   -> class Yes  [0.983]
##
## Rule 5: (102/13, lift 6.0)
##   minutos_dia <= 160.2
##   reclamos > 3
##   -> class Yes  [0.865]
##
## Rule 6: (211/84, lift 4.1)
##   minutos_dia > 264.4
##   -> class Yes  [0.601]
##
## Default class: No
##
##
## Evaluation on training data (3333 cases):
##
##           Rules
##   -----
##   No      Errors
##
##      6  313 ( 9.4%)  <<
##
##
##   (a)  (b)  <-classified as
##   ----  ----
##   2753   97   (a): class No
##   216   267   (b): class Yes
```

```
## Attribute usage:
##
## 99.55% minutos_dia
## 87.76% reclamos
## 79.84% plan_internacional
## 75.97% minutos_internacionales
##
##
## Time: 0.0 secs
```

Observamos que el algoritmo se ha aplicado a 3333 casos con 5 variables;

REGLA 1: En la regla 1 observamos a la "class No" con un 95% lo que nos indica que la mayoría de los usuarios no abandonan la compañía, observamos que de 2604 casos analizados por el algoritmo solo un total de 128 usuarios abandonan la compañía.

Este 95% que se queda en la compañía, no tienen registrados un plan internacional, y los minutos de llamadas al día son \leq a 264.4 y sus reclamos son \leq a 3.

REGLA 2: En la regla 2 observamos a la "class No" con un 93.5% lo que nos indica que la mayoría de los usuarios no abandonan la compañía, observamos que de 2475 casos analizados por el algoritmo solo un total de 160 usuarios abandonan la compañía.

Este 93.5% que se queda en la compañía, los minutos de llamadas al día son \leq a 264.4; además sus minutos internacionales son \leq a 13.1 y sus reclamos son \leq a 3.

REGLA 3: En la regla 3 observamos a la "class No" con un 88.9% lo que nos indica que la mayoría de los usuarios no abandonan la compañía, observamos que de 1896 casos analizados por el algoritmo un total de 210 usuarios abandonan la compañía.

Este 88.9% que se queda en la compañía, los minutos de llamadas al día son mayores a 160.2, pero son menores o iguales a 264.4

REGLA 4: En la regla 4 observamos a la "class Yes" con un 98.3% lo que nos indica que la mayoría de los usuarios abandonan la compañía, observamos a 57 casos analizados por el algoritmo, con un lift alto de 6.8.

Estos usuarios si tienen registrados un plan internacional, y sus minutos internacionales son mayores a 13.1

REGLA 5: En la regla 5 observamos a la "class Yes" con un 86.5% lo que nos indica que la mayoría de los usuarios abandonan la compañía; de los 102 casos analizados 13 NO abandonan la compañía. Los minutos al día de estos usuarios son \leq a 160.2 y sus reclamos son $>$ a 3.

REGLA 6: En la regla 6 observamos a la "class Yes" con un 60.1%; lo que nos indica que la mayoría abandona la compañía, de 211 casos 84 no abandonan la compañía.

Estos usuarios registran minutos al día > 264.4.

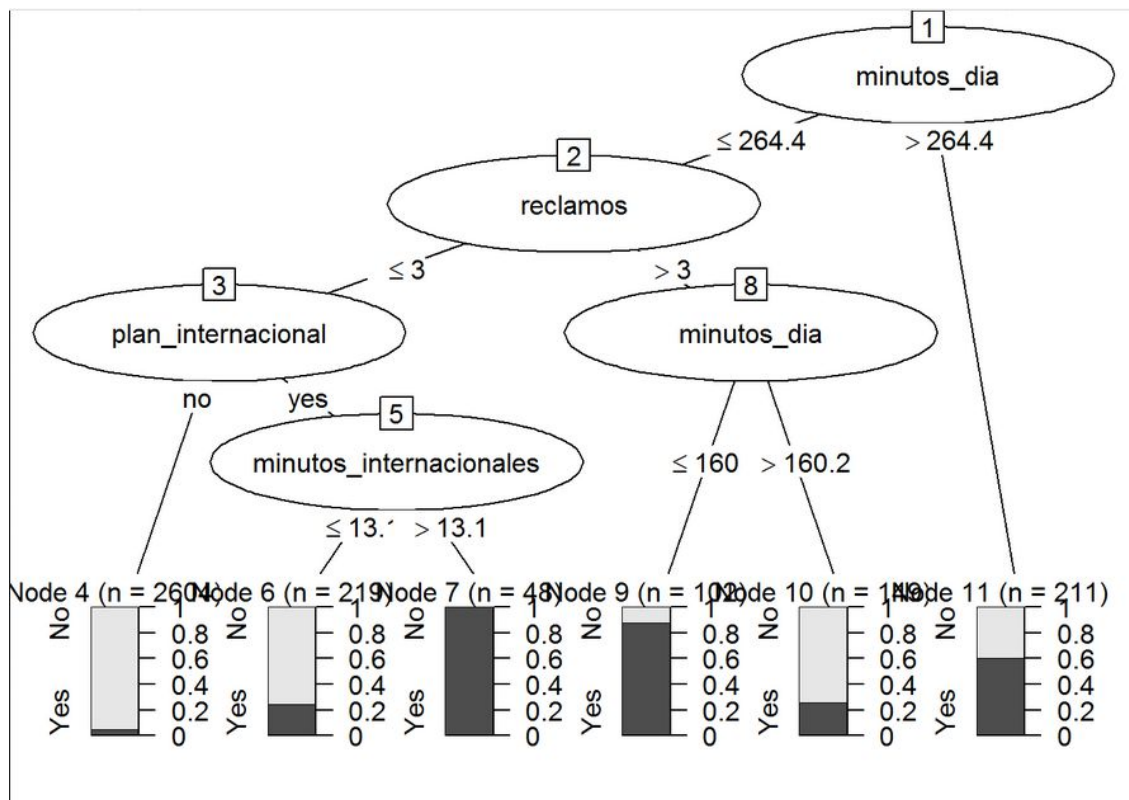
De acuerdo a la estimación del error del modelo se tiene que existe un 9.4% de error que corresponden a 313 casos.

A continuación mostramos el Árbol obtenido.

```
model <- C50::C5.0(trainX, trainy)
print(model)
```

```
##
## Call:
## C5.0.default(x = trainX, y = trainy)
##
## Classification Tree
## Number of samples: 3333
## Number of predictors: 4
##
## Tree size: 6
##
## Non-standard options: attempt to group attributes
```

```
plot(model)
```



N= indica el número de muestras.

El algoritmo ha considerado que a partir de responder las diferentes variables que ha estado seleccionando incluido un valor que ha considerado oportuno podemos determinar quiénes se van y quienes no de la compañía.

El mejor corte que se ha podido hacer es a partir de preguntar si los minutos que hablan los usuarios son menores o iguales a 264.4 minutos al día; si las personas hablan 264.4 minutos al día o menos; se les realiza otra pregunta: ¿sus llamadas por reclamos son menores o iguales que 3?; si la persona registra 3 o menos reclamos; se le formula otra pregunta ¿tienen plan internacional?; si no tiene plan internacional, la mayoría no abandona la compañía en un 95% (donde n=2604); un 5% si lo hará.

Ahora si las personas tienen plan internacional: se les formularia esta pregunta ¿cuantos minutos hablan con familiares o amigos en el extranjero?; si hablan 13.1 minutos o menos la mayoría no abandona la compañía en un 78%; un 22% si lo hará (donde n=219); si hablan más de 13.1 minutos el 100% abandona la compañía (donde n=48).

Si las personas registran más de 3 reclamos y hablan 160.2 minutos o menos, el 90% abandona la compañía un 10% no lo hará (donde n=102); si las personas registran más de 3 reclamos y hablan más de 160.2 minutos al día el 78% no abandona la compañía; el 22% si lo hará. (Donde n=149).

Finalmente: si las personas hablan más de 264.4 minutos al día, el 60% abandona la compañía y el 40% no lo hará (donde n=211).

VALIDACION DEL MODELO CON LOS DATOS RESERVADOS.

Una vez tenemos el modelo, podemos comprobar su calidad prediciendo la clase para los datos de prueba que nos hemos reservado al principio.

```
predicted_model <- predict( model, testX, type="class" )
print(sprintf("La precision del Arbol es: %.4f %", 100*sum(predicted_model == testy) / length(predicted_model)))
```

```
## [1] "La precision del Arbol es: 90.8218 %"
```

Tenemos a nuestra disposición el paquete gmodels para obtener información completa sobre los resultados que ha podido predecir nuestro modelo.

```
if(!require(gmodels)){
  install.packages('gmodels', repos='http://cran.us.r-project.org')
  library(gmodels)
}
```

```
CrossTable(testy, predicted_model, prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE, dnn = c('Resultado real', 'Prediccion'))
```

```
##
##
##   Cell Contents
## |-----|
## |                      N |
## |      N / Table Total |
## |-----|
##
##
## Total Observations in Table:  1667
##
##
##               | Prediccion
## Resultado real |          No |          Yes | Row Total |
## -----|-----|-----|-----|
##           No |      1394 |          49 |      1443 |
##           |      0.836 |      0.029 |           |
## -----|-----|-----|-----|
##           Yes |       104 |         120 |       224 |
##           |      0.062 |      0.072 |           |
## -----|-----|-----|-----|
## Column Total |      1498 |         169 |      1667 |
## -----|-----|-----|-----|
##
##
```

En este apartado vemos la información más completa; sobre los resultados que ha podido predecir nuestro modelo.

Donde vemos que el modelo ha dicho que es falso para 1394 casos cuando realmente eran falso; y ha dicho que es falso a 104 casos que realmente eran verdadero; el modelo también ha dicho que ha sido verdadero a 49 casos que realmente eran falsos, y ha dicho que es verdadero a 120 casos que realmente eran verdaderos.

5. REPRESENTACION DE LOS RESULTADOS A PARTIR DE TABLAS Y GRAFICAS.

SE LOS HA ELABORADO EN CADA ITEMS.

6. RESOLUCION DEL PROBLEMA. A PARTIR DE LOS RESULTADOS OBTENIDOS ¿CUÁLES SON LAS CONCLUSIONES? ¿LOS RESULTADOS PERMITEN RESPONDER AL PROBLEMA?

1. En la PRUEBA DE CONTRASTE DE HIPOTESIS MEDIANTE LA COMPARACION DE LA TENDENCIA CENTRAL; aplicando la prueba T-STUDENT comprobamos que las variables “abandona” y “llamadas_dia” son independientes; ya que las medias de los clientes que abandonan la compañía y la media de los clientes que no abandonan la compañía es la misma; de modo que NO rechazamos la hipótesis nula (H_0 = la media del primer grupo es igual a la del segundo grupo); por lo tanto concluimos que quienes abandonan la compañía NO dependen de las llamadas telefónicas diarias.

2. En el análisis de la prueba del CHI CUADRADO, comprobamos que existe relación de las muestras de las variables “abandona” y “plan internacional”, ya que los resultados fueron para el valor experimental (333.19) y para el valor teórico (3.84), como el valor experimental es $>$ al valor teórico, llegamos a la conclusión de rechazar la hipótesis nula y asumimos la hipótesis alternativa, lo que quiere decir que ambas variables se encuentran relacionadas.

3. En la predicción del modelo del Árbol del c50, el paquete gmodels para obtener información completa sobre los resultados que ha podido predecir nuestro modelo; observamos que el modelo ha dicho que 1394 usuarios no abandonarían la compañía y realmente no abandonaron la compañía; y de 104 casos el modelo ha dicho que no abandonarían la compañía cuando si lo hicieron; por el contrario el modelo ha dicho que 49 personas abandonaban la compañía cuando no lo hicieron, y de 120 personas el modelo predijo que abandonaban la compañía y realmente si lo hicieron; esta conclusión le da al algoritmo una precisión de 90.82% donde $n = 1667$.

4. La gran mayoría no abandona la compañía en la siguientes condiciones: CASO1 si los usuarios no tienen plan internacional, CASO2 si los usuarios hablan más de 160.2 y menos de 264.4 minutos al día; CASO3 si los reclamos de los usuarios han sido menores a 3; CASO4 si los minutos internacionales que habla el usuario son menores o iguales a 13.1.

Los resultados nos han permitido comprender y responder al problema.

7. CODIGO: HAY QUE ADJUNTAR EL CODIGO. PREFERIBLEMENTE EN R, CON EL QUE SE HA REALIZADO LA LIMPIEZA, ANALISIS Y REPRESENTACION DE LOS DATOS. SI LO PREFERÍS, TAMBIEN PODEIS TRABAJAR EN PHYTON.

title: 'Tipologia y ciclo de vida de los datos - Limpieza y analisis de los datos'

author: "Autor: Lider E. Zambrano Zambrano"

date: "Diciembre 2019"

output:

html_document:

highlight: default

number_sections: yes

```

theme: cosmo
toc: yes
toc_depth: 2
includes:
  in_header:
pdf_document:
  highlight: zenburn
  toc: yes
word_document: default
---
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```

LIMPIEZA DE LOS DATOS.

Primero cargamos nuestra data.

```

```{r message= FALSE, warning=FALSE}
data<-read.csv("./churn.csv",header=T,sep=",")
attach(data)
```

```

Modificamos los nombres a los campos.

```

```{r message= FALSE, warning=FALSE}
names(data)
c("abandona","longitud_cuenta","plan_internacional","plan_correo_voz","mensajes_por
_correo","minutos_dia","llamadas_dia","total_carga_dia","total_minutos","total_llamad
as","total_carga","minutos_noche","llamadas_noche","total_carga_noche","minutos_int
ernacionales", "llamadas_internacionales", "total_cargo_internacional","reclamos")
```

```

Seleccionamos las columnas que necesitamos

```

```{r message= FALSE, warning=FALSE}
data <- data[, c(1, 3, 6, 7, 12,13, 15,16,18)]
```

```

La función dim() nos permite conocer exactamente el número de filas y columnas de nuestro dataset.

```

```{r}
dim(data)
```

```

Observamos 5000 registros y 9 variables, existían 18 variables pero hemos seleccionado solo las variables que necesitamos.

TRATAMIENTO DE CEROS O ELEMENTOS VACIOS

Es de gran interes saber si tenemos muchos valores nulos (campos vacios) y la distribucion de valores por variables. Es por ello recomendable empezar el analisis con una vision general de las variables. Mostraremos para cada atributo la cantidad de valores perdidos mediante la funcion summary.

El nuestro conjunto de datos original no existian valores vacios; pero hemos eliminado algunos valores para tratarlos.

```
```{r}
summary(data)
```
```

```
```{r}
Revisamos nuevamente valores vacios
colSums(is.na(data))
```
```

Vemos que existen 2 valores vacios en la variable minutos_dia y 2 valores vacios en la variable llamadas_dias.

Ahora vamos a tratar los valores vacios por la media aritmetica en las variables discretas.

```
```{r}
Tomamos la media para valores vacios de la variable "minutos_dia" y "llamadas_dia"
data$minutos_dia[is.na(data$minutos_dia)] <- mean(data$minutos_dia,na.rm=T)
data$llamadas_dia[is.na(data$llamadas_dia)] <- mean(data$llamadas_dia,na.rm=T)
```
```

Revisamos nuevamente la data, para verificar los cambios, y vemos que ya no existen valores vacios.

```
```{r}
Revisamos nuevamente valores vacios
colSums(is.na(data))
```
```

IDENTIFICACION Y TRATAMIENTO DE VALORES EXTREMOS

Para identificar los valores extremos hemos utilizado un diagrama de cajas y la función boxplots.stats().

* COLUMNA MINUTOS_DIA

Primero calculamos los cuartiles

```
```{r}
quantile(data$minutos_dia)
```
```

El primer dato es el valor minimo y el ultimo es el valor maximo, quiere decir que para esta variable hay clientes que no hablan ningun minuto al dia; y tambien hay usuarios que hablan hasta 351.5 minutos al dia. (no habra valores mayores a 351.5 ya que este es el valor maximo de la columna minutos_dia)

Los otros 3 datos son los cuartiles, el cuartil 1 (143.7), el cuartil 2 (180.1) y el cuartil 3 (216.2); (el cuartil 2 que es 180.1 es la media).

Podemos ver los datos de la columna minutos_dias en un diagrama de cajas, para ello trabajamos con el comando boxplot y observamos que existen valores fuera de los bigotes superior e inferior.

```
```{r}
boxplot(data$minutos_dia)
```
```

Ahora vamos a confirmar la existencia de datos atipicos para esto necesitamos el rango intercuartil que se lo obtiene con el comando IQR.

El rango intercuartil (IQR) es la distancia entre el primer cuartil (Q1) y el tercer cuartil (Q3). El 50% de los datos está dentro de este rango.

```
```{r}
IQR(data$minutos_dia)
```
```

El rango intercuartil es 72.5

Para confirmar la existencia de un dato atipico necesitamos un rango; y lo calculamos asi: (El rango intercuartil multiplicado por 1.5 y sumado el primer cuartil para encontrar el valor maximo del rango; a su vez el rango intercuartil lo multiplicamos por 1.5 y restado el primer cuartil encontramos el valor minimo del rango)

```
```{r}
MIN<-(143.7-1.5*72.5)
MAX<-(143.7+1.5*72.5)
```
```

Ahora vamos a observar el rango que nos muestra el MIN y el MAX.

```
```{r}
range(MIN,MAX)
```
```

Si los datos de la columna minutos_dia; se encuentran dentro de este rango (34.95 y 252.45) significa que no hay datos atipicos, pero si hay un dato por fuera de este rango ese dato sera atipico.

Veamos el rango de la columna de minutos_dia.

```
```{r}
range(data$minutos_dia)
```
```

Observamos que hay datos fuera del rango de MIN y MAX, eso quiere decir que tenemos datos atipicos.

Pero ¿cuales son esos valores que estan fuera del rango de MIN y MAX?; los podemos observar con la funcion boxplot.stats.

```
```{r}
boxplot.stats(data$minutos_dia)$out
```
```

* COLUMNA LLAMADAS_DIA

Graficamos con el comando boxplot y observamos valores fuera del vigote superior e inferior.

```
```{r}
boxplot(data$llamadas_dia)
```
```

calculamos los cuartiles

```
```{r}
quantile(data$llamadas_dia)
```
```

Calculamos el rango intercuartil

```
```{r}
IQR(data$llamadas_dia)
```
```

Calculamos el valor minimo y maximo del nuevo rango para comparar con los valores extremos.

```
```{r}
MIN<-(87-1.5*26)
MAX<-(87+1.5*26)
```
```

Imprimimos los valores minimos y maximos.

```
```{r}
range(MIN,MAX)
```
```

Estos datos nos indican que si los valores estan entre 48 y 126 no son valores atipicos, fuera de estos si; luego observamos los valores que estan fuera de este rango .


```
```{r}
boxplot.stats(data$llamadas_dia)$out
```
```

* COLUMNA MINUTOS_NOCHE

Graficamos con el comando boxplot y observamos valores fuera del vigote superior e inferior.

```
```{r}
boxplot(data$minutos_noche)
```
```

calculamos los cuartiles

```
```{r}
quantile(data$minutos_noche)
```
```

Calculamos el rango intercuartil

```
```{r}
IQR(data$minutos_noche)
```
```

Calculamos el valor minimo y maximo del nuevo rango para comparar con los valores extremos.

```
```{r}
MIN<-(166.9-1.5*67.8)
MAX<-(166.9+1.5*67.8)
```
```

Imprimimos los valores minimos y maximos.

```
```{r}
range(MIN,MAX)
```
```

Estos datos nos indican que si los valores estan entre 65.2 y 268.6 no son valores atipicos, fuera de estos si; luego observamos los valores que estan fuera de este rango .

```
```{r}
boxplot.stats(data$minutos_noche)$out
```
```

*COLUMNA LLAMADAS_NOCHE

Graficamos con el comando boxplot y observamos valores fuera del vigote superior e inferior.

```
```{r}
boxplot(data$llamadas_noche)
```
```

calculamos los cuartiles

```
```{r}
quantile(data$llamadas_noche)
```
```

Calculamos el rango intercuartil

```
```{r}
IQR(data$llamadas_noche)
```
```

Calculamos el valor minimo y maximo del nuevo rango para comparar con los valores extremos.

```
```{r}
MIN<-(87-1.5*26)
MAX<-(87+1.5*26)
```
```

Imprimimos los valores minimos y maximos.

```
```{r}
range(MIN,MAX)
```
```

Estos datos nos indican que si los valores estan entre 48 y 126 no son valores atipicos, fuera de estos si; luego observamos los valores que estan fuera de este rango .

```
```{r}
boxplot.stats(data$llamadas_noche)$out
```
```

* COLUMNA MINUTOS_INTERNACIONALES

Graficamos con el comando boxplot y observamos valores fuera del vigote superior e inferior.

```
```{r}
boxplot(data$minutos_internacionales)
```
```

calculamos los cuartiles

```
```{r}
quantile(data$minutos_internacionales)
```
```

Calculamos el rango intercuartil

```
```{r}
IQR(data$minutos_internacionales)
```
```

Calculamos el valor minimo y maximo del nuevo rango para comparar con los valores extremos.

```
```{r}
MIN<-(8.5-1.5*3.5)
MAX<-(8.5+1.5*3.5)
```
```

Imprimimos los valores minimos y maximos.

```
```{r}
range(MIN,MAX)
```
```

Estos datos nos indican que si los valores estan entre 3.25 y 13.75 no son valores atipicos, fuera de estos si; luego observamos los valores que estan fuera de este rango .

```
```{r}
boxplot.stats(data$minutos_internacionales)$out
```
```

* COLUMNA LLAMADAS_INTERNACIONALES

Graficamos con el comando boxplot y observamos valores fuera del vigote superior.

```
```{r}
boxplot(data$llamadas_internacionales)
```
```

calculamos los cuartiles

```
```{r}
quantile(data$llamadas_internacionales)
```
```

Calculamos el rango intercuartil

```
```{r}
IQR(data$llamadas_internacionales)
```
```

Calculamos el valor minimo y maximo del nuevo rango para comparar con los valores extremos.

```
```{r}
MIN<-(3-1.5*3)
MAX<-(3+1.5*3)
```
```

Imprimimos los valores minimos y maximos.

```
```{r}
range(MIN,MAX)
```
```

Estos datos nos indican que si los valores estan entre -1.5 y 7.5 no son valores atipicos, fuera de estos si; luego observamos los valores que estan fuera de este rango .

```
```{r}
boxplot.stats(data$llamadas_internacionales)$out
```
```

* COLUMNA RECLAMOS

Graficamos con el comando boxplot y observamos valores fuera del vigote superior.

```
```{r}
muestra<-c(6,7)
boxplot(data$reclamos)
```
```

calculamos los cuartiles

```
```{r}
quantile(data$reclamos)
```
```

Calculamos el rango intercuartil

```
```{r}
IQR(data$reclamos)
```
```

Calculamos el valor minimo y maximo del nuevo rango para comparar con los valores extremos.

```
```{r}
MIN<-(1-1.5*1)
MAX<-(1+1.5*1)
```
```

Imprimimos los valores minimos y maximos.

```
```{r}
range(MIN,MAX)
```
```

Estos datos nos indican que si los valores estan entre -0.5 y 2.5 no son valores atipicos, fuera de estos si; luego observamos los valores que estan fuera de este rango .

```

```{r}
boxplot.stats(data$reclamos)$out
```

```

En todas las variables observamos que existen datos fuera del rango MIN y MAX; pero los valores no los podemos considerar datos extremos, por ejemplo observamos que estos valores pueden darse en la columna minutos_día; ya que observamos que apenas hay 2 usuarios de 5000 que no registran minutos al día, y también hay usuarios que han hablado hasta 350 minutos al día: lo que corresponde a hablar 6 horas diarias aproximadamente (6 horas multiplicado por 60 minutos = 360 minutos), y esto es real en los usuarios más jóvenes en la telefonía celular, además de NO sobrepasar los 720 minutos que corresponden a las 12 horas en el día.

Sin embargo si tuviéramos un valor mayor a 720 minutos, ese dato correspondería a un valor extremo para esta columna.

En la columna llamadas_día vemos que estos valores también pueden darse, ya que pueden haber personas que hagan más de 126 llamadas en el día considerando su situación laboral, el valor máximo de llamadas es 165 y si lo multiplicamos por 2.5 (que serían el tiempo en minutos que duraría cada llamada) da como resultado 412.5 minutos; un valor superior a los 350 minutos que una persona hablaría según el cálculo de la variable anterior.

En las otras variables al tratarse de minutos y llamadas sucede el mismo caso, por lo que tampoco lo consideramos como valores extremos a aquellos datos que salen del rango de MIN y MAX.

Y finalmente la columna reclamos, vemos que es normal que los usuarios hayan hecho hasta 9 llamadas al servicio de llamadas, ya sea por tratarse de anular algún servicio o presentar otras inconformidades por el servicio.

ANALISIS DE LOS DATOS

SELECCIÓN DE LOS GRUPOS DE DATOS QUE SE QUIEREN ANALIZAR/COMPARAR

A continuación, se seleccionan los grupos dentro de nuestro conjunto de datos que pueden resultar interesantes para analizar y/o comparar.

```

```{r}
Agrupacion por plan internacional
grupo1_plan <- data[, c(1,2)]
data_sinplan <- grupo1_plan[grupo1_plan$plan_internacional == "no",]
data_plan_int <- grupo1_plan[grupo1_plan$plan_internacional == "yes",]
```

```

El comando summary nos muestra los resultados de cada grupo del plan internacional.

```

```{r}
summary(data$plan_internacional)
```

```

```

```{r}
Agrupacion por churn (abandona)
grupo2_churn <- data[, c(1,2)]
data.NOabandona <- grupo2_churn[grupo2_churn$abandona == "No",]
data.SIabandona <- grupo2_churn[grupo2_churn$abandona == "Yes",]
```

```

El comando summary nos muestra los resultados de cada grupo de la variable abandona.

```

```{r}
summary(data$abandona)
```

```

Para agrupar por churn(abandona) y plan internacional; utilizaremos el comando table.

```

```{r}
table(data$abandona,data$plan_internacional)
```

```

Para agrupar por llamadas al dia y churn utilizamos un grafico, para facilitar la comprension de los datos.

```

```{r}
Agrupacion por llamadas al dia
data2 <- data[, c(1,4)]
plot(data2)
```

```

COMPROBACION DE LA NORMALIDAD Y HOMOGENEIDAD DE LA VARIANZA

Para la comprobación de que los valores que toman nuestras variables cuantitativas provienen de una población distribuida normalmente, utilizaremos la prueba de normalidad de Anderson-Darling.

Así, se comprueba que para que cada prueba se obtiene un p-valor superior al nivel de significación prefijado $\alpha = 0,05$ o $\alpha = 0,10$. Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal.

```

```{r}
library(nortest)
alpha = 0.05
col.names = colnames(data)
for (i in 1:ncol(data)) {
 if (i == 1) cat("Variables que no siguen una distribucion normal:\n")
 if (is.integer(data[,i]) | is.numeric(data[,i])) {
 p_val = ad.test(data[,i])$p.value
 if (p_val < alpha) {
 cat(col.names[i])
 # Format output
 if (i < ncol(data) - 1) cat(", ")
 if (i %% 3 == 0) cat("\n")
 }
 }
}
```

```

```
}
}
}
'''
```

* VEAMOS LA GRAFICA DE LA COLUMNA LLAMADAS_DIA.

EN ESTE EJERCICIO CONSIDERAMOS QUE LA HIPOTESIS NULA ES: H_0 = "La muestra proviene de una poblacion con distribucion normal".

Observemos el Histograma + la Curva normal teorica

```
'''{r}
library(ggplot2)
ggplot(data = data, aes(x = llamadas_dia)) +
  geom_histogram(aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
               args = list(mean = mean(data$llamadas_dia),
                           sd = sd(data$llamadas_dia))) +
  ggtitle("Histograma + curva normal teorica") +
  theme_bw()
'''
```

Ahora podemos observar la curva de la variable llamadas_dia que no sigue una distribución normal.

```
'''{r}
plot(density(data$llamadas_dia))
'''
```

Aplicamos la prueba de Anderson - Darling

```
'''{r}
ad.test(data$llamadas_dia)$p.value
'''
```

A pesar de que en la grafica observamos una figura simetrica o bastante simetrica, vemos que al aplicar la prueba de Anderson - Darling el valor que obtenemos es menor con relacion al 5% de referencia, por lo tanto nuestra conclusion es que rechazamos la hipotesis nula (H_0).

* VEAMOS LA GRAFICA DE LA COLUMNA LLAMADAS_NOCHE

Observemos el Histograma + la Curva Normal teórica

```
'''{r}
library(ggplot2)
ggplot(data = data, aes(x = llamadas_noche)) +
  geom_histogram(aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
```

```
stat_function(fun = dnorm, colour = "firebrick",
  args = list(mean = mean(data$llamadas_noche),
    sd = sd(data$llamadas_noche))) +
ggtitle("Histograma + curva normal teorica") +
theme_bw()
```

```

Ahora podemos observar la curva de la variable llamadas\_noche que no sigue una distribución normal.

```
```{r}
plot(density(data$llamadas_noche))
```

```

Finalmente aplicamos la prueba de Anderson - Darling.

```
```{r}
ad.test(data$llamadas_noche)$p.value
```

```

Como el valor obtenido aplicando la prueba de Anderson-darling se acerca a 5%, ya que nos da un resultado de 3.7% aplicamos la prueba de Kolmogorov-Smirnov con el comando Lillie para comparar y verificar los valores de la probabilidad.

```
```{r}
lillie.test(data$llamadas_noche)$p.value
```

```

Con este resultado no tenemos duda, tenemos un valor bien por debajo del 5%.

A pesar de que en la grafica observamos una figura simetrica o bastante simetrica, vemos que al aplicar la prueba de Anderson - Darling y la prueba de Kolmogorov-Smirnov el valor que obtenemos es menor con relacion al 5% de referencia, por lo tanto nuestra conclusion es que rechazamos la hipotesis nula (Ho).

## \* VEAMOS LA GRAFICA DE LA COLUMNA MINUTOS\_INTERNACIONALES

Observemos el Histograma + la Curva Normal teórica

```
```{r}
library(ggplot2)
ggplot(data = data, aes(x = minutos_internacionales)) +
  geom_histogram(aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
    args = list(mean = mean(data$minutos_internacionales),
      sd = sd(data$minutos_internacionales))) +
  ggtitle("Histograma + curva normal teorica") +
  theme_bw()
```

```



```
'''
```

Ahora podemos observar la curva de la variable minutos\_internacionales que no sigue una distribución normal.

```
'''{r}
plot(density(data$minutos_internacionales))
'''
```

Finalmente aplicamos la prueba de Anderson - Darling

```
'''{r}
ad.test(data$minutos_internacionales)$p.value
'''
```

En la grafica observamos una figura un poco simetrica ,vemos tambien que al aplicar la prueba de Anderson - Darling el valor que obtenemos es bastante menor con relacion al 5% de referencia, por lo tanto nuestra conclusion es que rechazamos la hipotesis nula (Ho).

\* VEAMOS LA GRAFICA DE LA COLUMNA LLAMADAS\_INTERNACIONALES.

Observemos el Histograma + la Curva Normal teórica

```
'''{r}
library(ggplot2)
ggplot(data = data, aes(x = llamadas_internacionales)) +
 geom_histogram(aes(y = ..density.., fill = ..count..)) +
 scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
 stat_function(fun = dnorm, colour = "firebrick",
 args = list(mean = mean(data$llamadas_internacionales),
 sd = sd(data$llamadas_internacionales))) +
 ggtitle("Histograma + curva normal teorica") +
 theme_bw()
'''
```

Ahora podemos observar la curva de la variable llamadas\_internacionales que no sigue una distribución normal.

```
'''{r}
plot(density(data$llamadas_internacionales))
'''
```

Aplicamos la prueba de Anderson - Darling

```
'''{r}
ad.test(data$llamadas_internacionales)$p.value
'''
```

En la grafica observamos una figura ASIMETRICA ,vemos tambien que al aplicar la prueba de Anderson - Darling el valor que obtenemos es bastante menor con relacion al

5% de referencia, por lo tanto nuestra conclusion es que rechazamos la hipotesis nula (Ho).

#### \* VEAMOS LA GRAFICA DE LA COLUMNA RECLAMOS

Observemos el Histograma + la Curva Normal teórica

```
```{r}
library(ggplot2)
ggplot(data = data, aes(x = reclamos)) +
  geom_histogram(aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
    args = list(mean = mean(data$reclamos),
      sd = sd(data$reclamos))) +
  ggtitle("Histograma + curva normal teorica") +
  theme_bw()
```
```

Ahora podemos observar la curva de la variable reclamos que no sigue una distribución normal.

```
```{r}
plot(density(data$reclamos))
```
```

Aplicamos la prueba de Anderson - Darling

```
```{r}
ad.test(data$reclamos)$p.value
```
```

En la grafica observamos una figura ASIMETRICA ,vemos tambien que al aplicar la prueba de Anderson - Darling el valor que obtenemos es bastante menor con relacion al 5% de referencia, por lo tanto nuestra conclusion es que rechazamos la hipotesis nula (Ho).

#### \* VEAMOS LA GRAFICA DE LA COLUMNA MINUTOS\_DIA

Observemos el Histograma + la Curva Normal teórica

```
```{r}
library(ggplot2)
ggplot(data = data, aes(x = minutos_dia)) +
  geom_histogram(aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
  stat_function(fun = dnorm, colour = "firebrick",
    args = list(mean = mean(data$minutos_dia),
      sd = sd(data$minutos_dia))) +
  ggtitle("Histograma + curva normal teorica") +
  theme_bw()
```
```

```
'''
```

Ahora podemos observar la curva de la variable minutos\_dia que sigue una distribución normal.

```
'''{r}
plot(density(data$minutos_dia))
'''
```

Aplicamos la prueba de Anderson - Darling

```
'''{r}
ad.test(data$minutos_dia)$p.value
'''
```

En la grafica observamos una figura bastante simetrica, vemos tambien que al aplicar la prueba de Anderson - Darling el valor que obtenemos es bastante mayor con un 88% con relacion al 5% de referencia, por lo tanto nuestra conclusion es que NO rechazamos la hipotesis nula (Ho), es decir aceptamos que la muestra proviene de una distribucion normal.

\* VEAMOS LA GRAFICA DE LA COLUMNA MINUTOS\_NOCHE

Observemos el Histograma + la Curva Normal teórica

```
'''{r}
library(ggplot2)
ggplot(data = data, aes(x = minutos_noche)) +
 geom_histogram(aes(y = ..density.., fill = ..count..)) +
 scale_fill_gradient(low = "#DCDCDC", high = "#7C7C7C") +
 stat_function(fun = dnorm, colour = "firebrick",
 args = list(mean = mean(data$minutos_noche),
 sd = sd(data$minutos_noche))) +
 ggtitle("Histograma + curva normal teorica") +
 theme_bw()
'''
```

Ahora podemos observar la curva de la variable minutos\_dia que sigue una distribución normal.

```
'''{r}
plot(density(data$minutos_noche))
'''
```

Aplicamos la prueba de Anderson - Darling

```
'''{r}
ad.test(data$minutos_noche)$p.value
'''
```

En la grafica observamos una figura bastante simetrica, vemos tambien que al aplicar la prueba de Anderson - Darling el valor que obtenemos es bastante mayor con un 90% con relacion al 5% de referencia, por lo tanto nuestra conclusion es que NO rechazamos la

hipotesis nula ( $H_0$ ), es decir que aceptamos que la muestra proviene de una distribución normal.

\* PRUEBA DE HOMOGENEIDAD ( $H_0$ = "Las varianzas poblacionales son iguales")

Primero aplicamos Test de Bartlett que permite contrastar la igualdad de varianza en 2 o más poblaciones sin necesidad de que el tamaño de los grupos sea el mismo. Es más sensible que el test de Levene a la falta de normalidad.

varianza por test de Bartlett

```
```{r}
bartlett.test(minutos_dia ~ abandona, data = data)
```
```

Luego vamos a ver la Varianza por test de Fligner-Killeen

```
```{r}
fligner.test(minutos_dia ~ abandona, data = data)
```
```

y finalmente para estas dos columnas probamos la Varianza por el test de Levene

```
```{r}
library(lawstat)
levene.test(data$minutos_dia, data$abandona)
```
```

En los tres casos el valor de la probabilidad (p-value) es un número menor a 0.05 (5%); por lo que rechazamos la  $H_0$ .(exactamente el p-value en decimal es 0.000000000000000022)

Veamos la varianza entre reclamos y abandona por el test de Levene

```
```{r}
library(lawstat)
levene.test(data$reclamos, data$abandona)
```
```

También observamos que el p-value es un número muy pequeño, menor al 5% de referencia; por lo que rechazamos la hipótesis nula.

Varianza por test de Bartlett entre llamadas\_dia y abandona.

```
```{r}
bartlett.test(llamadas_dia ~ abandona, data = data)
```
```

En vista que el p-value es igual al 5% NO rechazamos la hipótesis nula, por lo tanto concluimos de que existe homogeneidad de varianzas, es decir que la varianza de ambas muestras son iguales.

## ## PRUEBAS ESTADISTICAS

### ### PRIMERA PRUEBA ESTADISTICA: PRUEBA DE CONTRASTE DE HIPOTESIS MEDIANTE LA COMPARACION DE LA TENDENCIA CENTRAL.

ANALISIS DE RELACION ENTRE UNA VARIABLE CUANTITATIVA Y OTRA CUALITATIVA.

Tenemos dos variables que vamos a comparar:

X= ABANDONA (cualitativa)

Y= LLAMADAS DIA (cuantitativa)

Veamos la estructura de los datos de las dos variables a comparar.

```
```{r message= FALSE, warning=FALSE}
data2 <- data[, c(1,4)]
summary(data2)
```
```

Veamos tambien los primeros 6 registros de ambas columnas.

```
```{r message= FALSE, warning=FALSE}
head(data2)
```
```

Partimos de la hipotesis de que ambas variables son independientes, formulado en una pregunta quedaria:

¿Los clientes que abandonan la compañía son independientes al número de llamadas que realizan diariamente?

Para determinar si estas dos variables (cualitativa y cuantitativa) estan relacionadas o son independientes; analizaremos sus medias, es decir que si sabemos que la media o la mediana de las llamadas que realiza el cliente son iguales para quienes abandonen o no abandonen la compania; podemos afirmar que ambas variables son independientes, es decir que quienes abandonan la compania no dependen del numero de llamadas que realizan; si por el contrario se prueba que las medias o las medianas son diferentes podemos concluir que ambas variables estan relacionadas, y que por lo tanto los clientes que abandonan la compania dependen del numero de llamadas telefonicas que realizan a diario.

Veamos una grafica.

```
```{r}
plot(data2)
```
```

Podemos ver que la media entre los clientes que abandonan la compania y la media de los clientes que no abandonan la compania es la misma; de modo que NO rechazamos la hipotesis nula ( $H_0$ = la media del primer grupo es igual a la del segundo grupo);por lo tanto concluimos que quienes abandonan la compania NO dependen de las llamadas telefonicas diarias.

Aplicamos una prueba T-STUDENT para comprobar la igualdad de las medias de los dos grupos.

```
```{r}
t.test(llamadas_dia ~ abandona, data=data, var.equal = TRUE)
```
```

Observamos que la prueba T-STUDENT nos permite conocer la media de los dos grupos, donde nuevamente concluimos que son iguales, además el p-value del 26% que es superior al 5% de referencia nos indica que las muestras son iguales, de modo que no rechazamos la hipotesis nula.

### ### SEGUNDA PRUEBA ESTADISTICA: PRUEBA DEL CHI CUADRADO

La prueba de Chi-cuadrado en R es un método estadístico que se utiliza para determinar si dos variables categóricas tienen una correlación significativa entre ellas.

Hipotesis de partida= las dos variables en estudio son independientes.

Observamos las columnas con las que vamos a trabajar en esta prueba estadistica.

```
```{r message= FALSE, warning=FALSE}
data2 <- data[, c(1,2)]
summary(data2)
```
```

Con la funcion summary encontramos un resumen de los datos de ambas columnas, donde tenemos que en la columna abandona 4293 clientes NO abandonan la compania y 707 si lo hacen.

EN la columna plan internacional que 4527 personas NO tienen plan internacional y 473 si lo tienen, cabe indicar que en este resumen las columnas no estan relacionadas.

Ahora veamos una tabla de frecuencias con los datos de ambas columnas relacionados, estos serian las frecuencias observadas.

Los datos obtenidos en esta tabla nos dice que 4019 clientes que no tienen plan internacional NO abandonan la compania, y 274 clientes que registraban plan internacional tampoco abandonaron la compania.

Por el contrario 508 clientes que no tenian plan internacional SI abandonaron la compania y 199 usuarios que tenian plan internacional tambien abandonaron la compania.

TABLA DE FRECUENCIAS OBSERVADAS.(Se ha agregado las frecuencias marginales para las filas y columnas con el comando addmargins)

```
```{r message= FALSE, warning=FALSE}
table(data2)
tabla<-table(data2)
addmargins(tabla)
```
```

Ahora podemos observar las frecuencias esperadas, redondeadas a 2 decimales.

TABLA DE FRECUENCIAS ESPERADAS.

```
```{r message= FALSE, warning=FALSE}
esperados<-chisq.test(tabla)$expected
b<-round(esperados,2)
b
```
```

Presentamos los datos de la tabla de frecuencias observadas en porcentajes, redondeados a 3 decimales.

```
```{r message= FALSE, warning=FALSE}
porc<-(prop.table(tabla))
a<-round(porc,3)
a
```
```

Aplicamos la prueba estadística del CHI CUADRADO.

```
```{r message= FALSE, warning=FALSE}
chisq.test(x = tabla)
```
```

OBTENEMOS COMO RESPUESTA A LA PRUEBA DEL CHI CUADRADO EL VALOR DE: 333.19.

Ahora para aceptar o rechazar la hipótesis nula ( $H_0$ = ambas variables son independientes) obtenemos el valor crítico tabular de chi-cuadrado, al 95% con 1 grado de libertad.

```
```{r message= FALSE, warning=FALSE}
qchisq(0.95,1)
```
```

La respuesta del valor teórico con un nivel de riesgo del 5% y con 1 grado de libertad es: 3.84.

CONCLUSION: Nuestro valor experimental (333.19) es mayor al valor teórico (3.84), por lo que rechazamos la hipótesis nula y asumimos la hipótesis alternativa, lo que quiere decir que ambas variables están relacionadas, pero están relacionadas de forma inversa,

es decir que mientras existan menos clientes con plan internacional, aumentaran el numero de clientes que NO abandonaran la compania. (Esta conclusion es para el 90% de las observaciones)

Para un 10% de clientes, las variables estan relacionadas de forma directa: es decir que si aumentan los planes internacionales tambien aumentaran los clientes que NO abandonan la compania.

Ahora veamos un grafico de barras, para visualizar la relacion entre la variable plan internacional y la variable abandona. (En la primera barra observamos la relacion inversa y en la segunda barra la relacion directa).

```
```{r message= FALSE, warning=FALSE}
if(!require(ggplot2)){
  install.packages('ggplot2', repos='http://cran.us.r-project.org')
  library(ggplot2)
}
if(!require(grid)){
  install.packages('grid', repos='http://cran.us.r-project.org')
  library(grid)
}
if(!require(gridExtra)){
  install.packages('gridExtra', repos='http://cran.us.r-project.org')
  library(gridExtra)
}
```

```{r message= FALSE, warning=FALSE}
grid.newpage()
plotbyplan1<-ggplot(data,aes(plan_internacional,fill=abandona))+geom_bar()
+labs(x="plan internacional", y="Clientes")+ guides(fill=guide_legend(title=""))+
scale_fill_manual(values=c("black","#008000"))+ggtitle("Usuarios con plan
internacional")
grid.arrange(plotbyplan1,ncol=2)
```
```

En el grafico de barras y la informacion de la tabla de frecuencias observadas; vemos que la gran mayoria de usuarios no tienen plan internacional, son exactamente 4527 usuarios que no tienen plan internacional, de los cuales solo 508 abandonan la compania pero 4019 no lo hicieron (relacion inversa); por otro lado vemos que existen 473 usuarios que tienen plan internacional y 274 no abandonaron la compania y 199 de ellos si lo hicieron (relacion directa).

### TERCER METODO DE ANALISIS: ARBOLES DE CLASIFICACION Y REGRESION.

Modelo de clasificación: Árbol de decisión C5.0



Nuestro objetivo es crear un Árbol de decisión que permita analizar qué tipo de usuarios de nuestra data ha tenido probabilidades de abandonar la compañía. Por lo tanto, la variable por la que clasificaremos es el campo "abandona".

Para la futura evaluación del Árbol de decisión, es necesario dividir el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba.

Seleccionamos la columna 1 donde se encuentra nuestra variable dependiente, y luego seleccionamos las demás columnas donde se encuentran las variables que se relacionan con la variable "abandona".

```
```{r}
new_data<-data[, c(1, 2, 3, 7, 9)]
set.seed(666)
y <- new_data[,1]
X <- new_data[,2:5]
```
```

Observamos los nombres de las columnas.

```
```{r}
names(new_data)
```
```

calculamos a cuantas filas corresponde dos tercios de los datos ( $2*5000/3=3333$ ) y dividimos "manualmente" el conjunto.

```
```{r}
trainX <- X[1:3333,]
trainy <- y[1:3333]
testX <- X[3334:5000,]
testy <- y[3334:5000]
```
```

##### Creacion del modelo, calidad del modelo y extraccion de reglas

Se crea el Arbol de decision usando los datos de entrenamiento:

```
```{r}
model <- C5.0::C5.0(trainX, trainy,rules=TRUE )
summary(model)
```
```

Observamos que el algoritmo se ha aplicado a 3333 casos con 5 variables;

REGLA 1: En la regla 1 observamos a la "class No" con un 95% lo que nos indica que la mayoría de los usuarios no abandonan la compañía, observamos que de 2604 casos analizados por el algoritmo solo un total de 128 usuarios abandonan la compañía.

Este 95% que se queda en la compañía, no tienen registrados un plan internacional, y los minutos de llamadas al día son  $\leq$  a 264.4 y sus reclamos son  $\leq$  a 3.

REGLA 2: En la regla 2 observamos a la "class No" con un 93.5% lo que nos indica que la mayoría de los usuarios no abandonan la compañía, observamos que de 2475 casos analizados por el algoritmo solo un total de 160 usuarios abandonan la compañía.

Este 93.5% que se queda en la compañía, los minutos de llamadas al día son  $\leq$  a 264.4; además sus minutos internacionales son  $\leq$  a 13.1 y sus reclamos son  $\leq$  a 3.

REGLA 3: En la regla 3 observamos a la "class No" con un 88.9% lo que nos indica que la mayoría de los usuarios no abandonan la compañía, observamos que de 1896 casos analizados por el algoritmo un total de 210 usuarios abandonan la compañía.

Este 88.9% que se queda en la compañía, los minutos de llamadas al día son mayores a 160.2, pero son menores o iguales a 264.4

REGLA 4: En la regla 4 observamos a la "class Yes" con un 98.3% lo que nos indica que la mayoría de los usuarios abandonan la compañía, observamos a 57 casos analizados por el algoritmo, con un lift alto de 6.8.

Estos usuarios si tienen registrados un plan internacional, y sus minutos internacionales son mayores a 13.1

REGLA 5: En la regla 5 observamos a la "class Yes" con un 86.5% lo que nos indica que la mayoría de los usuarios abandonan la compañía; de los 102 casos analizados 13 NO abandonan la compañía. Los minutos al día de estos usuarios son  $\leq$  a 160.2 y sus reclamos son  $>$  a 3.

REGLA 6: En la regla 6 observamos a la "class Yes" con un 60.1%; lo que nos indica que la mayoría abandona la compañía, de 211 casos 84 no abandonan la compañía.

Estos usuarios registran minutos al día  $>$  264.4.

De acuerdo a la estimación del error del modelo se tiene que existe un 9.4% de error que corresponden a 313 casos.

A continuacion mostramos el Arbol obtenido.

```
```{r}
model <- C50::C5.0(trainX, trainy)
print(model)
plot(model)
```
```

n indica el numero de muestras.

El algoritmo ha considerado que a partir de responder las diferentes variables que ha estado seleccionando incluido un valor que ha considerado oportuno podemos determinar quienes se van y quienes no de la compania.

El mejor corte que se ha podido hacer es a partir de preguntar si los minutos que hablan los usuarios son menores o iguales a 264.4 minutos al dia; si las personas hablan 264.4 minutos al dia o menos; se les realiza otra pregunta: ¿sus llamadas por reclamos son menores o iguales que 3?; si la persona registra 3 o menos reclamos; se le vuelve a preguntar ¿tienen plan internacional?; si no tiene plan internacional, la mayoria no abandona la compania en un 95% (donde n=2604); un 5% si lo hará.

Ahora si las personas tienen plan internacional: se les formularia esta pregunta ¿cuantos minutos hablan con familiares o amigos en el extranjero?; si hablan 13.1 minutos o menos la mayoria no abandona la compania en un 78%; un 22% si lo hará(donde n=219); si hablan mas de 13.1 minutos el 100% abandona la compania(donde n=48).

Si las personas registran mas de 3 reclamos y hablan 160.2 minutos o menos, el 90% abandona la compania un 10% no lo hará (donde n=102); si las personas registran mas de 3 reclamos y hablan mas de 160.2 minutos al dia el 78% no abandona la compania; el 22% si lo hará. (donde n=149).

Finalmente: si las personas hablan mas de 264.4 minutos al dia, el 60% abandona la compania y el 40% no lo hará.(donde n=211)

##### Validacion del modelo con los datos reservados

Una vez tenemos el modelo, podemos comprobar su calidad prediciendo la clase para los datos de prueba que nos hemos reservado al principio.

```
```{r}
predicted_model <- predict( model, testX, type="class" )
print(sprintf("La precision del Arbol es: %.4f %%",100*sum(predicted_model == testy)
/ length(predicted_model)))
```
```

Tenemos a nuestra disposicion el paquete gmodels para obtener informacion completa sobre los resultados que ha podido predecir nuestro modelo.

```
```{r}
if(!require(gmodels)){
  install.packages('gmodels', repos='http://cran.us.r-project.org')
  library(gmodels)
}
```
```

```
```{r}
CrossTable(testy, predicted_model,prop.chisq = FALSE, prop.c = FALSE, prop.r
=FALSE,dnn = c('Resultado real', 'Prediccion'))
```
```

En este apartado vemos la información más completa; sobre los resultados que ha podido predecir nuestro modelo.

Donde vemos que el modelo ha dicho que es falso para 1394 casos cuando realmente eran falso; y ha dicho que es falso a 104 casos que realmente eran verdadero; el modelo también ha dicho que ha sido verdadero a 49 casos que realmente eran falsos, y ha dicho que es verdadero a 120 casos que realmente eran verdaderos.

En este apartado vemos la información más completa; sobre los resultados que ha podido predecir nuestro modelo.




Donde vemos que el modelo ha dicho que es falso para 1394 casos cuando realmente eran falso; y ha dicho que es falso a 104 casos que realmente eran verdadero; el modelo también ha dicho que ha sido verdadero a 49 casos que realmente eran falsos, y ha dicho que es verdadero a 120 casos que realmente eran verdaderos.

### **INTEGRANTES DEL GRUPO:**

No se trabajó en grupo, el trabajo realizado es personal.

### **NOMBRE DEL INTEGRANTE:**

ZAMBRANO ZAMBRANO LIDER EUCLIDES

| <b>Contribuciones</b>       | <b>Firma</b>                                                                                |
|-----------------------------|---------------------------------------------------------------------------------------------|
| Investigación previa        | LEZZ<br> |
| Redacción de las respuestas | LEZZ<br> |
| Desarrollo código           | LEZZ<br> |

### **Bibliografía:**

<https://data.world/earino/churn>

[https://rpubs.com/Joaquin\\_AR/218466](https://rpubs.com/Joaquin_AR/218466)

[https://www.kaggle.com/toramky/automobile-dataset#Automobile\\_data.csv](https://www.kaggle.com/toramky/automobile-dataset#Automobile_data.csv)

<https://cran.r-project.org/doc/contrib/Saez-Castillo-RRCmdrv21.pdf>

[http://labrad.fisica.edu.uy/docs/tabla\\_chi\\_cuadrado.pdf](http://labrad.fisica.edu.uy/docs/tabla_chi_cuadrado.pdf)