ICD_ProyectoFinal

Laura del Pino Díaz

15/12/2016

Contents

Introducción

En este proyecto vamos a realizar un análisis de dos bases de datos: la base de datos de la aprobación de créditos en australia (australian credit approval) y el tiempo atmosférico de la ciudad de Izmir (wizmir). A partir de este análisis de los datos se realizará un estudios de modelos de clasificación con la base de datos de la aprobación de los créditos para determinar si se le pueden conceder o no el crédito. Mientras que con la base de datos del tiempo se elaaborarán distintos modelos de regresión con el objetivo de predecir la temperatura media.

Las bases de datos

En este apartado estudiaremos en la medida de lo posible las bases de datos asignadas para cada uno de los problemas.

Australian (Australian Credit Approval)

La base de datos *australian credit approval* tiene 15 atributos de los cuales actúan como predictores 14.

Los atributos de esta base de datos en particular no tienen un nombre descriptivo que te permita conocer que es lo que representan los datos por razones de confidencialidad, tal y como se detalla en la página de UCI. Lo que si conocemos es el tipo de variable que componen la base de datos y el intervalo o valores que puede tomar cada variable y se enlistan a continuación.

- A1 nominal {0, 1}
- A2 real [16.0,8025.0]
- A3 real [0.0,26335.0]
- A4 nominal {1, 2, 3}
- A5 entero [1,14]
- A6 entero [1,9]
- A7 real [0.0,14415.0]
- A8 nominal {0, 1}
- A9 nominal {0, 1}
- A10 entero [0,67]
- A11 nominal {0, 1}
- A12 nominal {1, 2, 3}
- A13 entero [0,2000]
- A14 entero [1,100001]
- Class nominal {0,1}

Dado la no descriptividad de los nombres no podemos realizar hipótesis previas sobre la base de datos. Por lo que procedemos a realizar un estudio de los principales estadísticos de cada variables.

Estudio de los principales estadísticos del conjunto de datos Australian

Variables numéricas

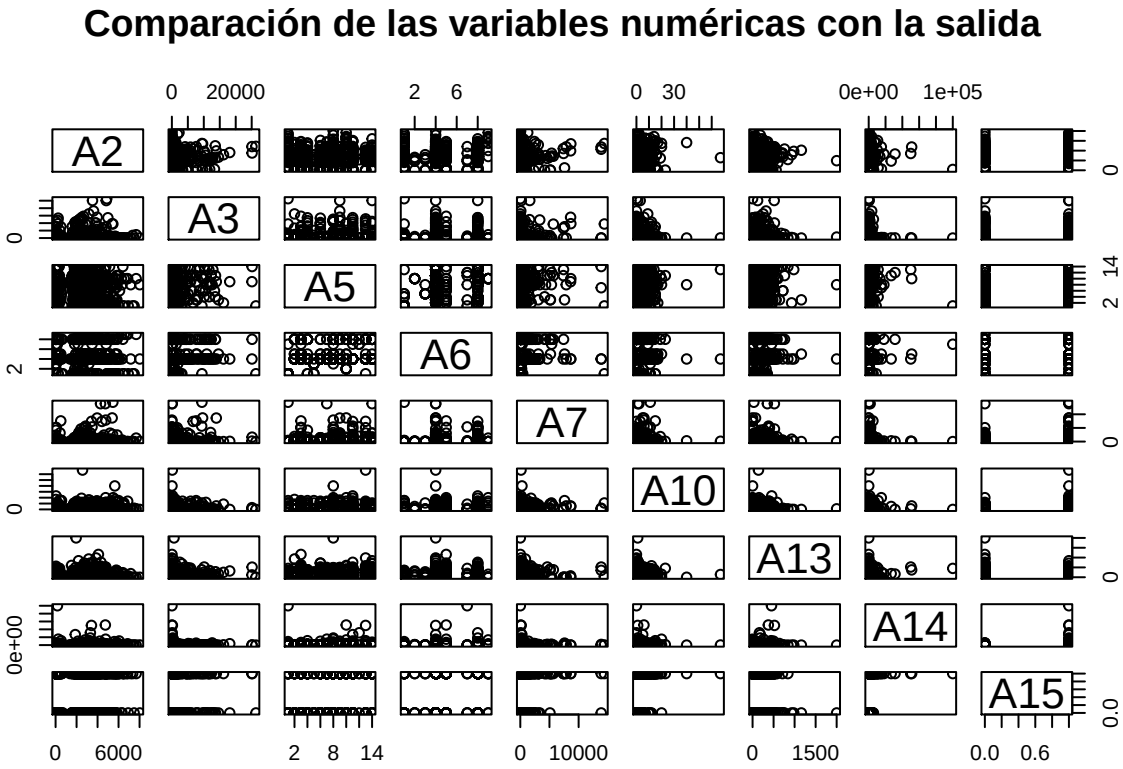
Cargamos la base de datos y miramos los cuartiles, valores máximos y desviación standart de las variables numéricas.

```
australian <- read.csv("./AustralianClassification/australian/australian.dat",
    comment.char = "@", header = FALSE)
names(australian) <- c("A1", "A2", "A3", "A4", "A5", "A6", "A7",
    "A8", "A9", "A10", "A11", "A12", "A13", "A14", "A15")
```

```
# Solo las clases numéricas
numerical_stats <- summary(australian[, c(-1, -4, -8, -9, -11,
    -12, -15)])
numerical_std <- apply(australian[, c(-1, -4, -8, -9, -11, -12,
    -15)], 2, sd)
```

Como podemos ver las variables con mayor varianza son la variable A2,A3,A7 y A14 que tienen su valor en las unidades de millar. Vamos a comparar las variables con la salida que está en la variable A15

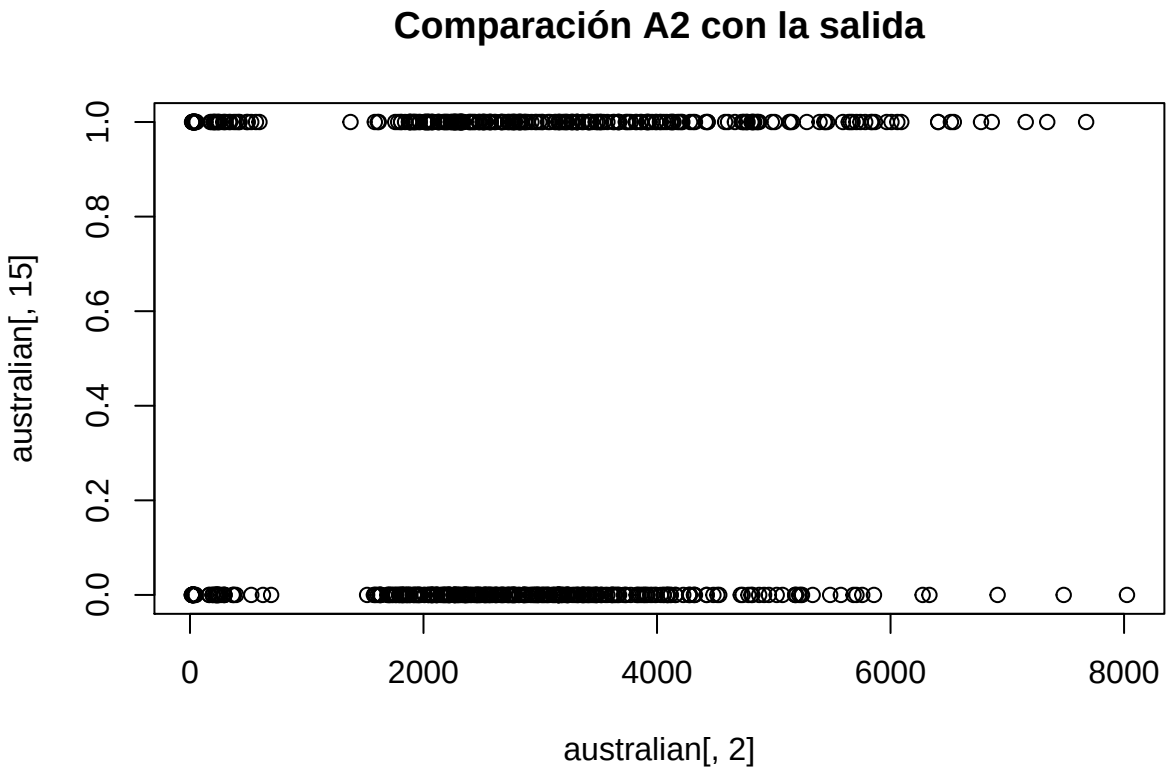
```
pairs(australian[, c(-1, -4, -8, -9, -11, -12)], main = "Comparación de las variables numéricas con la s
```



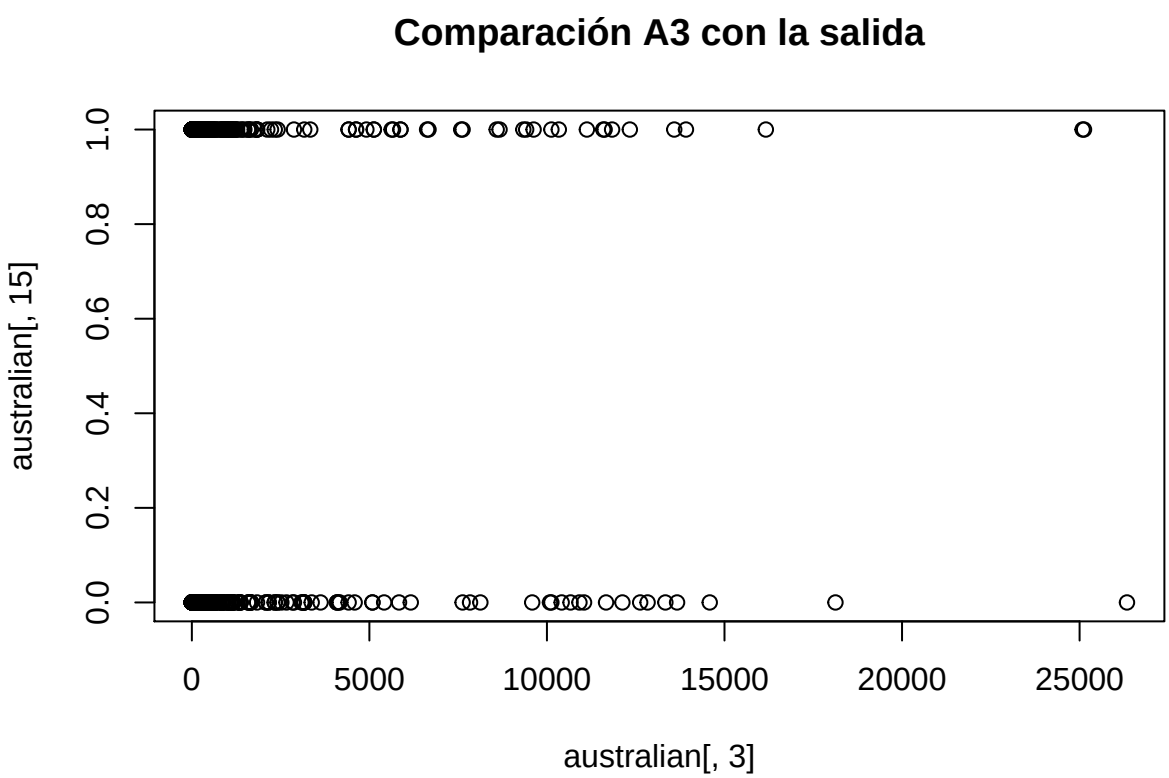

```
# TODO legend
```

Puesto que es una base de datos para clasificación con dos clases tiene sentido que en todas ellas aparezcan las dos columnas. Pero verlas así en pequeño no nos permite deducir si esa variable aporta mucho o poco a la salida, por lo que vamos a realizar unos plots para analizar mejor los datos.

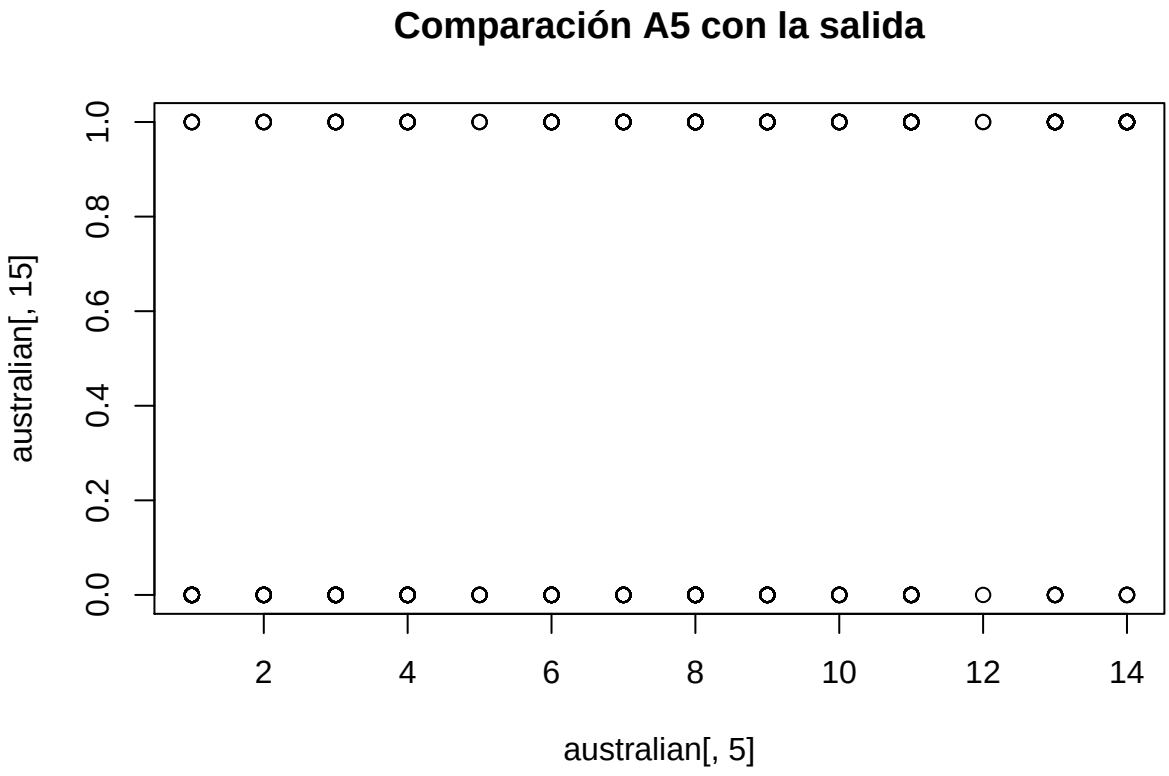
```
plot(australian[, 2], australian[, 15], main = "Comparación A2 con la salida")
```



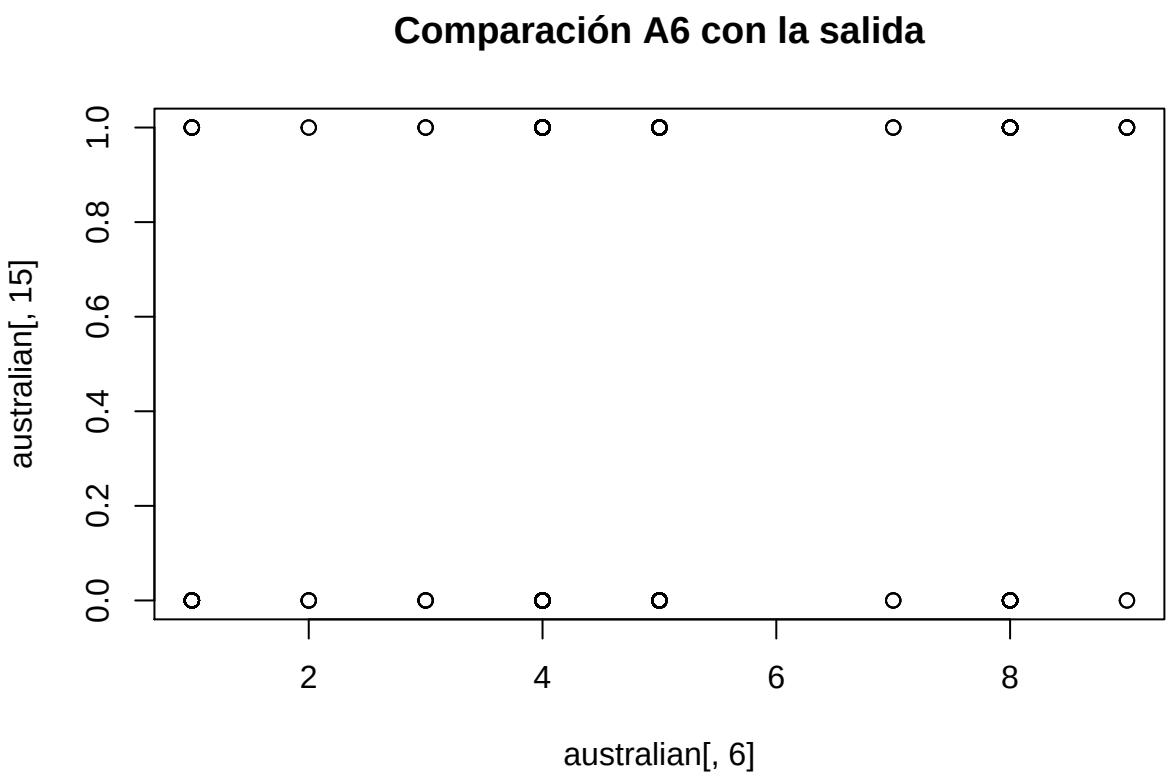

```
plot(australian[, 3], australian[, 15], main = "Comparación A3 con la salida")
```



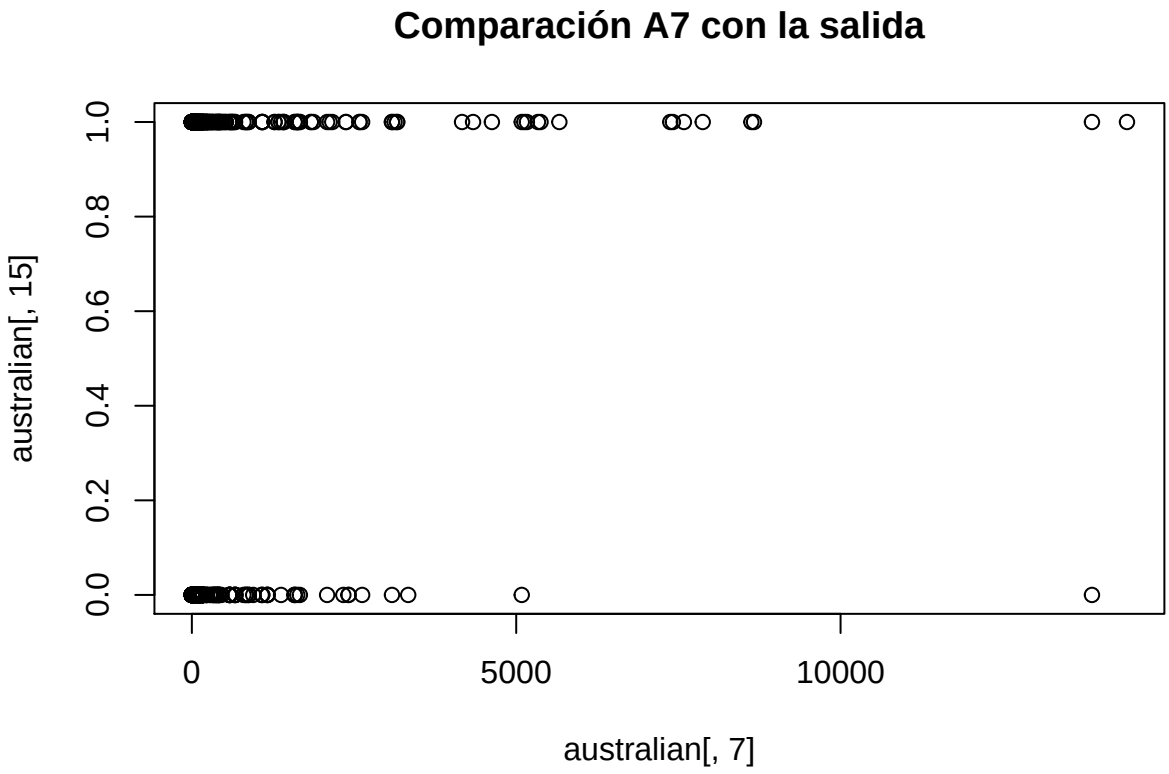

```
plot(australian[, 5], australian[, 15], main = "Comparación A5 con la salida")
```



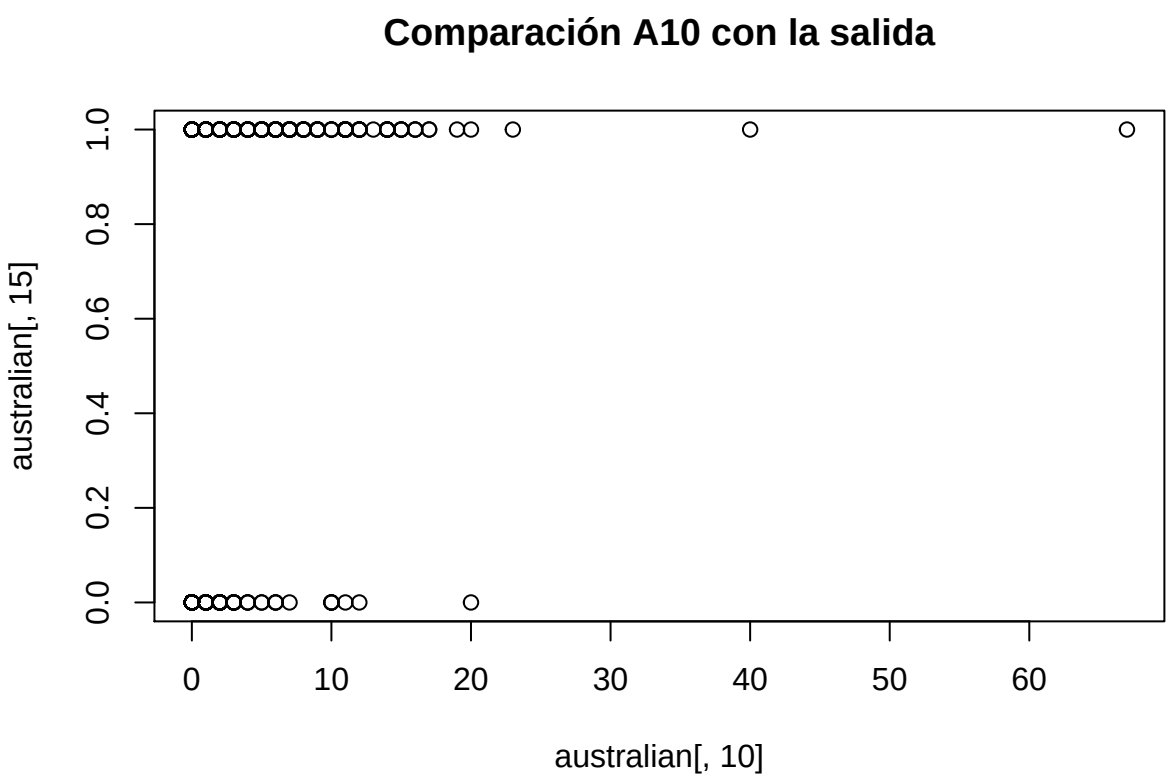

```
plot(australian[, 6], australian[, 15], main = "Comparación A6 con la salida")
```



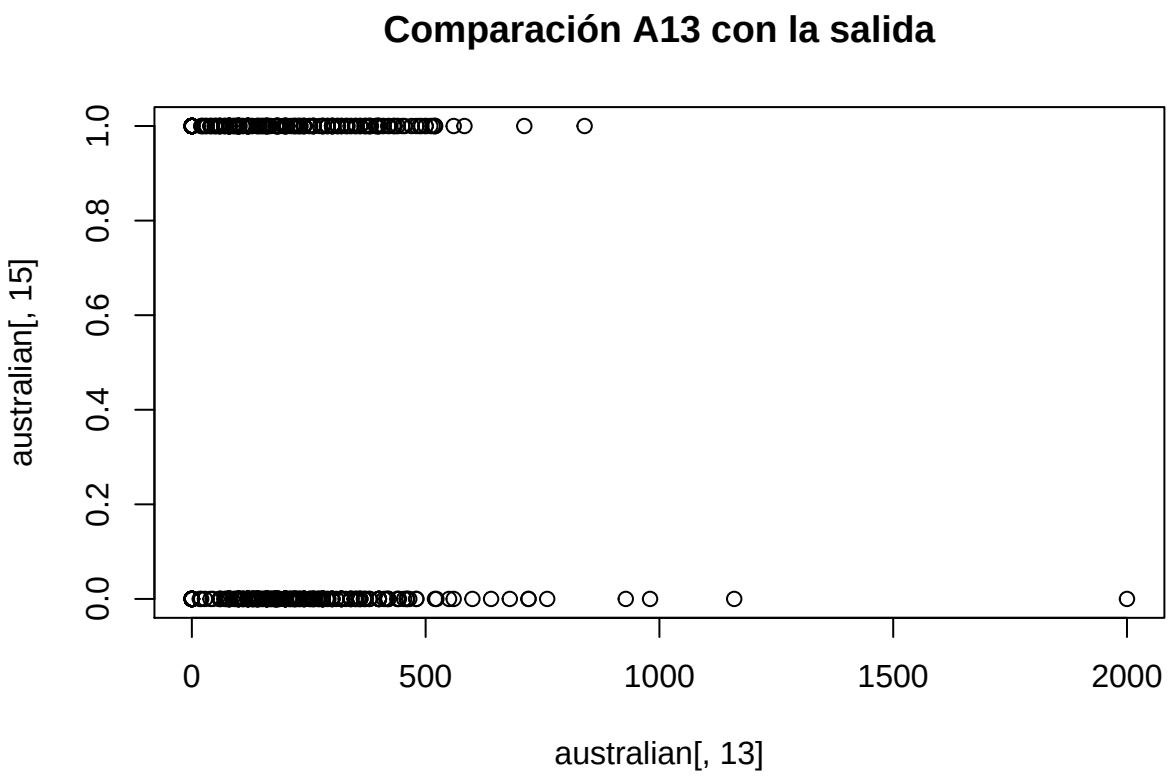

```
plot(australian[, 7], australian[, 15], main = "Comparación A7 con la salida")
```



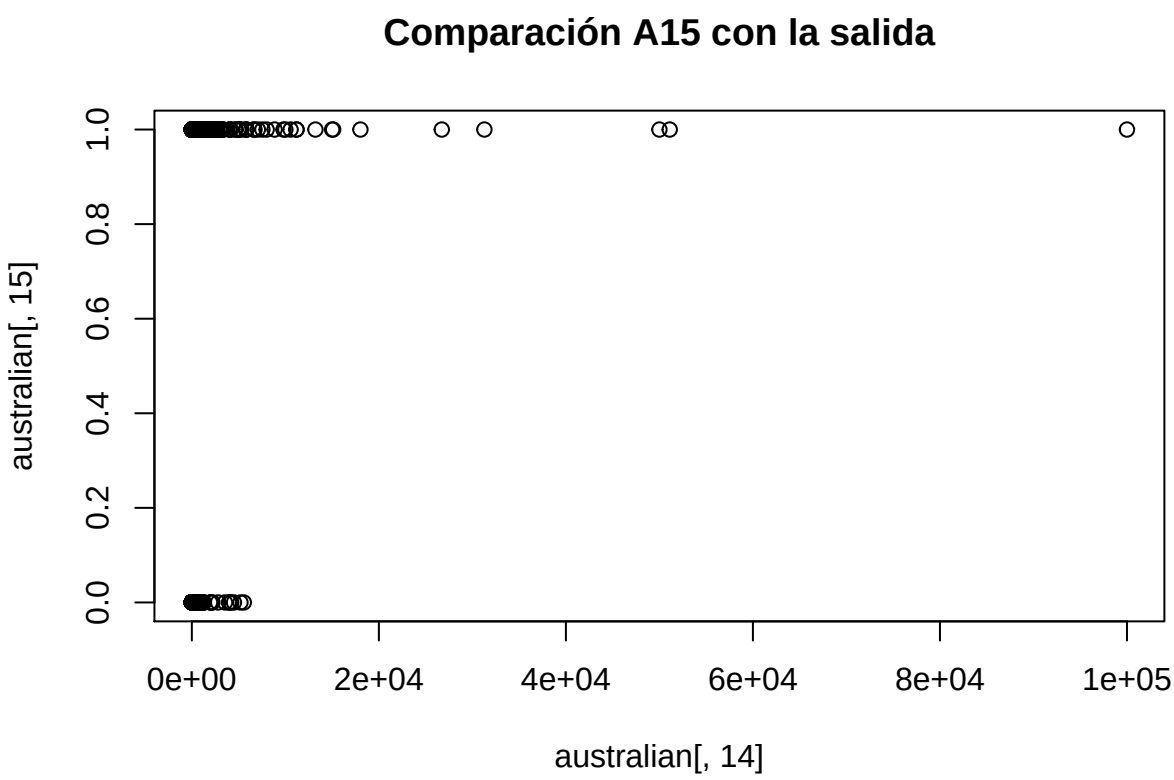

```
plot(australian[, 10], australian[, 15], main = "Comparación A10 con la salida")
```




```
plot(australian[, 13], australian[, 15], main = "Comparación A13 con la salida")
```



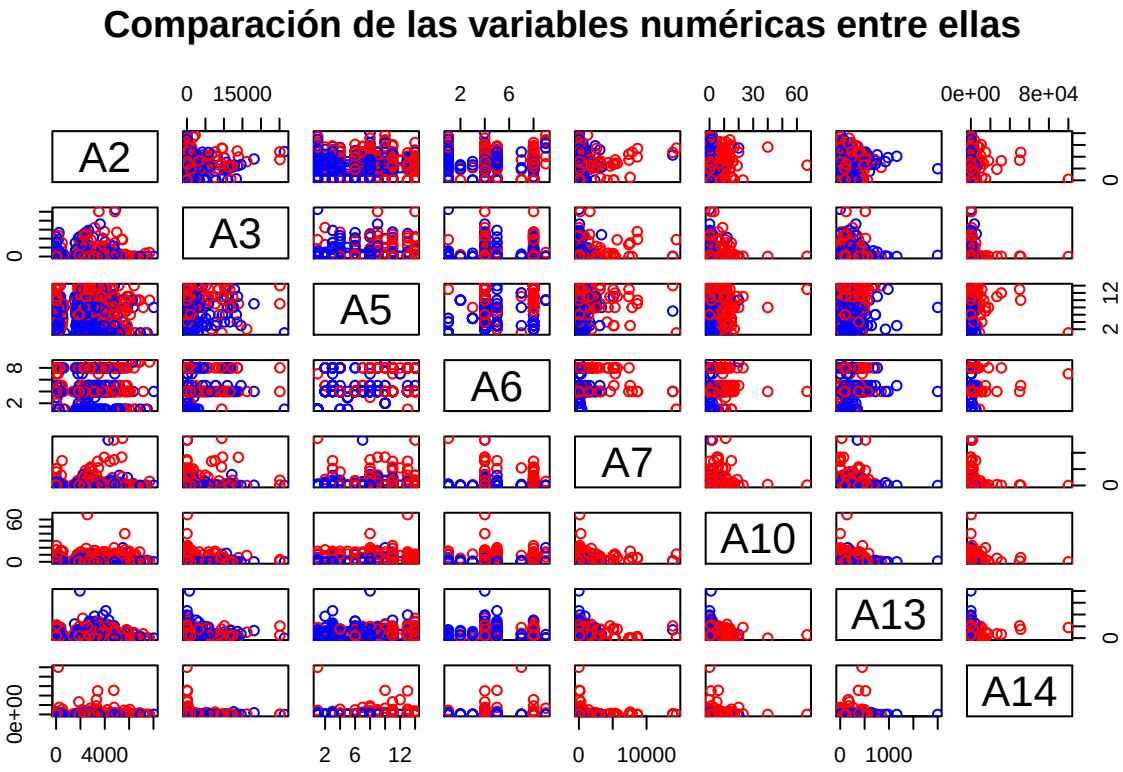

```
plot(australian[, 14], australian[, 15], main = "Comparación A15 con la salida")
```



No podemos decir que ninguna de las variables numéricas sea realmente discriminante para la salida. Sin embargo, no descartaría la variable 14 para que intervenga un modelo de clasificación, porque para valores altos solamente da salida positiva, es decir con valor 1.

Volvermos a representar las variables ahora dejando la variable de salida fuera

```
pairs(australian[, c(-1, -4, -8, -9, -11, -12, -15)], main = "Comparación de las variables numéricas en
    col = ifelse(australian[, 15] == 1, "red", "blue"))
```



Aparentemente no hay relación entre las variables, lo que parece curioso es que cualquiera de las variables que interacciona con la variable 6 forma una nube de puntos que recuerda a un histograma.

Variables categóricas.

Las variables catergóricas merecen un estudio propio puesto que estadísticos como la media o la desviación típica no tienen un valor interesante ya que no tiene sentido si tenemos las clases 1,2,3 y que la clase media sea 2,2. Por ello los estadísticos que vamos a usar en esta sección son los cuartiles, el valor más frecuente.

```
categorical_stats <- summary(australian[, c(1, 4, 8, 9, 11, 12)])
categorical_stats <- categorical_stats[-4, ]
```

Para obtener el valor más frecuente o moda del conjunto haremos uso de la siguiente función:

```
Mode <- function(x) {
    ux <- unique(x)
    ux[which.max(tabulate(match(x, ux)))]
}
```

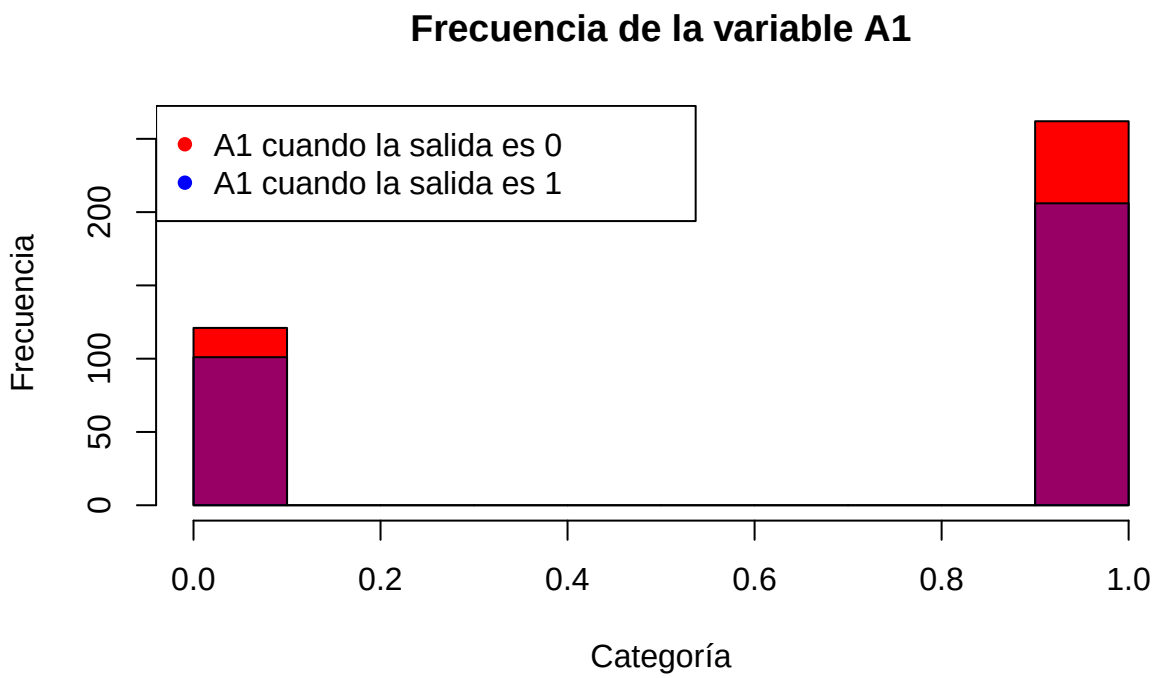
Realizando la moda sobre cada una de las variables categóricas tenemos:

```
categorical_stats <- rbind(categorical_stats, apply(australian[,
    c(1, 4, 8, 9, 11, 12)], 2, Mode))
categorical_stats
```

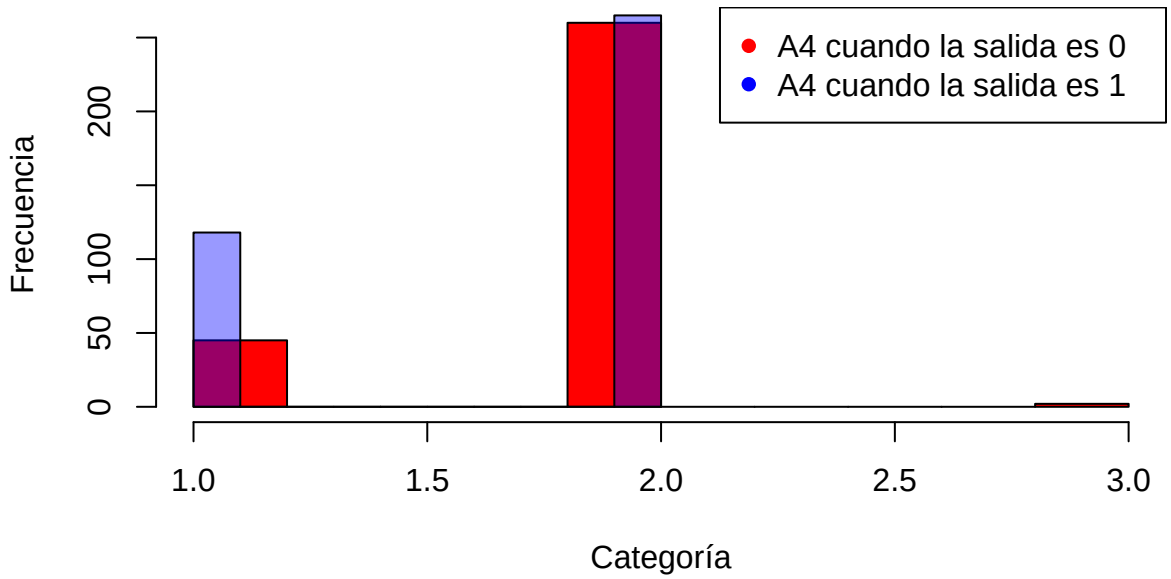
```
       A1                 A4                A8
 "Min.   :0.0000  " "Min.   :1.000  " "Min.   :0.0000  "
 "1st Qu.:0.0000  " "1st Qu.:2.000  " "1st Qu.:0.0000  "
 "Median :1.0000  " "Median :2.000  " "Median :1.0000  "
 "3rd Qu.:1.0000  " "3rd Qu.:2.000  " "3rd Qu.:1.0000  "
 "Max.   :1.0000  " "Max.   :3.000  " "Max.   :1.0000  "
 "1"                "2"               "1"
       A9                A11               A12
 "Min.   :0.0000  " "Min.   :0.000  " "Min.   :1.000  "
 "1st Qu.:0.0000  " "1st Qu.:0.000  " "1st Qu.:2.000  "
 "Median :0.0000  " "Median :0.000  " "Median :2.000  "
 "3rd Qu.:1.0000  " "3rd Qu.:1.000  " "3rd Qu.:2.000  "
 "Max.   :1.0000  " "Max.   :1.000  " "Max.   :3.000  "
 "0"                "0"               "2"
```

La información anterior nos ilustra como se distribuye cada una de las variables pero no como se relacionan con la salida, para ello dibujaremos gráficos en donde comparamos la frecuencia de cada valor con respecto al valor que toma la salida.

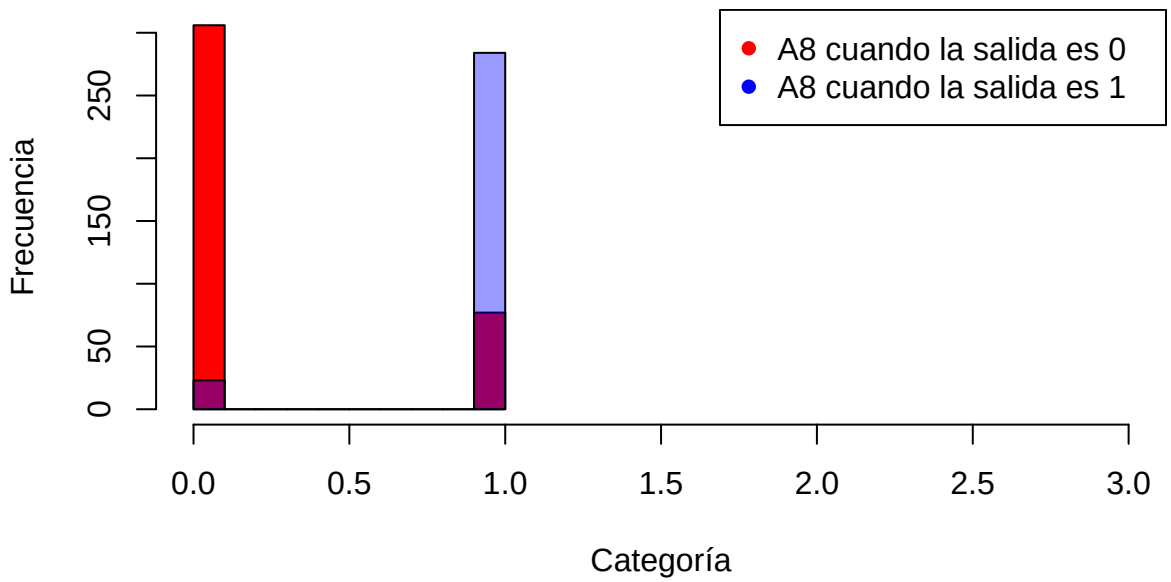
```
hist(australian[which(australian[, 15] == 0), 1], col = rgb(1,
    0, 0, 1), main = "Frecuencia de la variable A1", ylab = "Frecuencia",
    xlab = "Categoría")
hist(australian[which(australian[, 15] == 1), 1], col = rgb(0,
    0, 1, 0.4), add = TRUE)
legend("topleft", legend = c("A1 cuando la salida es 0", "A1 cuando la salida es 1"),
    text.width = 0.5, col = c("red", "blue"), pch = 16)
```




```
hist(australian[which(australian[, 15] == 1), 4], col = rgb(1,
    0, 0, 1), main = "Frecuencia de la variable A4", ylab = "Frecuencia",
    xlab = "Categoría")
hist(australian[which(australian[, 15] == 0), 4], col = rgb(0,
    0, 1, 0.4), add = T)
legend("topright", legend = c("A4 cuando la salida es 0", "A4 cuando la salida es 1"),
    text.width = 0.8, col = c("red", "blue"), pch = 16)
```

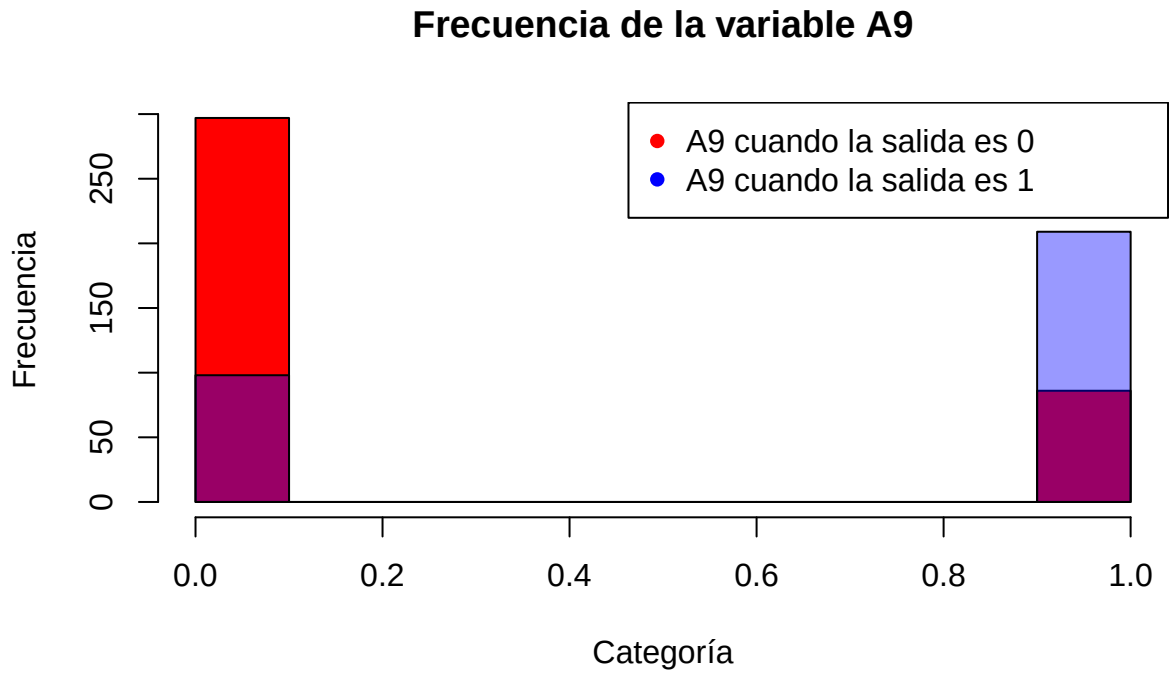



```
hist(australian[which(australian[, 15] == 0), 8], col = rgb(1,
    0, 0, 1), main = "Frecuencia de la variable A8", ylab = "Frecuencia",
    xlab = "Categoría", xlim = c(0, 3))
hist(australian[which(australian[, 15] == 1), 8], col = rgb(0,
    0, 1, 0.4), add = T)
legend("topright", legend = c("A8 cuando la salida es 0", "A8 cuando la salida es 1"),
    text.width = 1.2, col = c("red", "blue"), pch = 16)
```

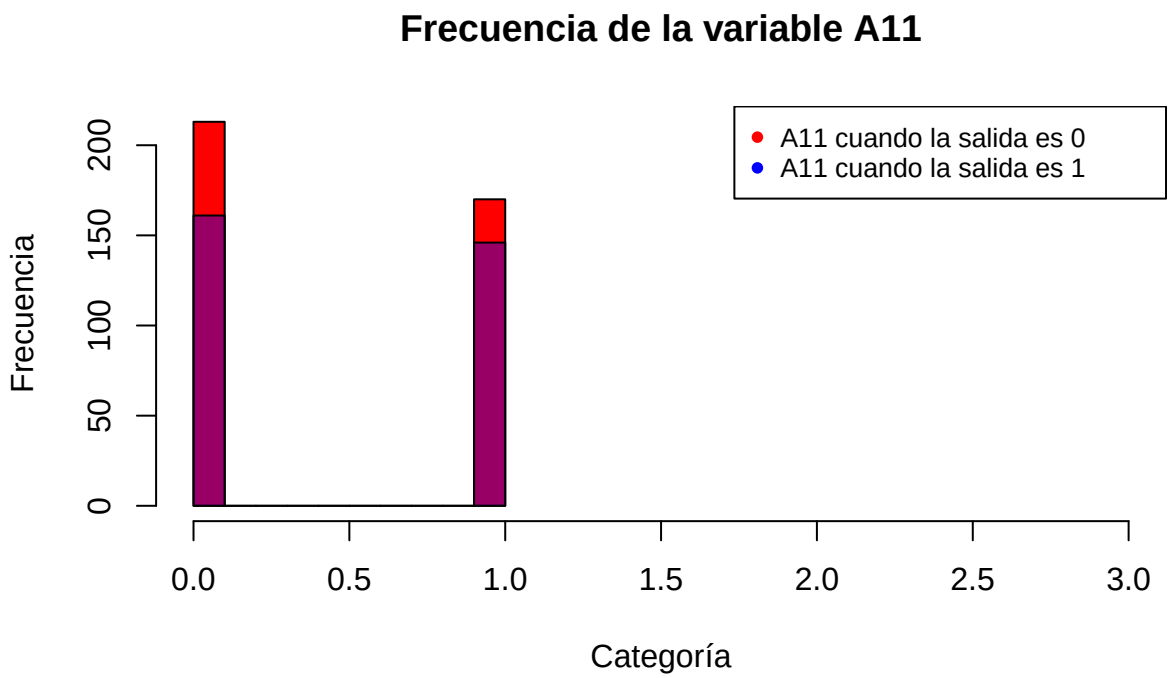



```
hist(australian[which(australian[, 15] == 0), 9], col = rgb(1,
    0, 0, 1), main = "Frecuencia de la variable A9", ylab = "Frecuencia",
    xlab = "Categoría")
```

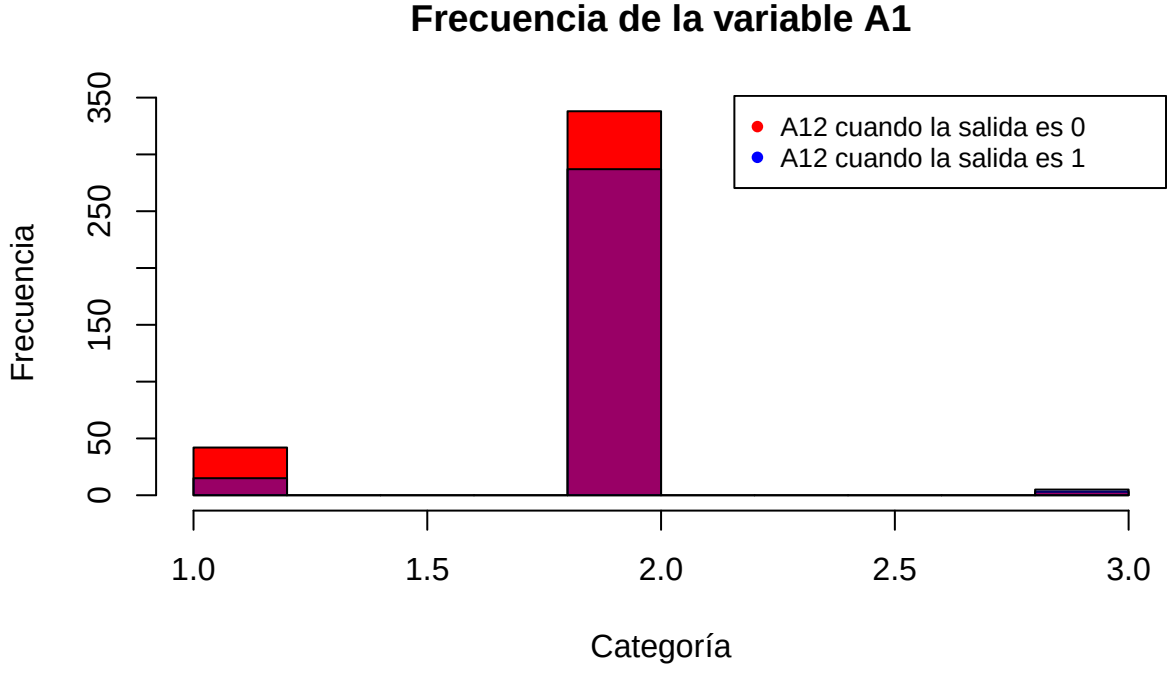
```
hist(australian[which(australian[, 15] == 1), 9], col = rgb(0,
    0, 1, 0.4), add = T)
legend("topright", legend = c("A9 cuando la salida es 0", "A9 cuando la salida es 1"),
    text.width = 0.5, col = c("red", "blue"), pch = 16)
```




```
hist(australian[which(australian[, 15] == 0), 11], col = rgb(1,
    0, 0, 1), main = "Frecuencia de la variable A11", ylab = "Frecuencia",
    xlab = "Categoría", xlim = c(0, 3))
hist(australian[which(australian[, 15] == 1), 11], col = rgb(0,
    0, 1, 0.4), add = T)
legend("topright", legend = c("A11 cuando la salida es 0", "A11 cuando la salida es 1"),
    text.width = 1.2, col = c("red", "blue"), pch = 16, cex = 0.8)
```




```
hist(australian[which(australian[, 15] == 0), 12], col = rgb(1,
    0, 0, 1), main = "Frecuencia de la variable A1", ylab = "Frecuencia",
    xlab = "Categoría")
hist(australian[which(australian[, 15] == 1), 12], col = rgb(0,
    0, 1, 0.4), add = T)
legend("topright", legend = c("A12 cuando la salida es 0", "A12 cuando la salida es 1"),
    text.width = 0.8, col = c("red", "blue"), pch = 16, cex = 0.8)
```



TODO concluir smthing

Wizmir (Weather of Izmir)