

ICD_ProyectoFinal

Laura del Pino Díaz

15/12/2016

Contents

Introducción	2
Las bases de datos	2
Australian (Australian Credit Approval)	2
Estudio de las variables numéricas	3
Estudio del número de valores perdidos	3
Estadísticos principales	3
Test de normalidad	4
Información gráfica	11
Estudio de las correlaciones	16
Estudio de las variables categóricas.	18
Estudio de los valores perdidos	18
Estudio de los estadísticos principales	18
Wizmir (Weather of Izmir)	25
Hipótesis previas	25
Variables numéricas.	25
Estudio de los valores perdidos	26
Test de normalidad	26
Estudio de los principales estadísticos	26
Estudio de correlación	28
Regresión	30
Problema	30
Regresores elegidos	30
Modelo de regresión simple	30
Modelo de regresión lineal múltiple	32

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 elaborarán distintos modelos de regresión con el objetivo de predecir la temperatura media.

Las bases de datos

En este apartado estudiaremos las bases de datos *Australian Credit Approval*(abreviado *australian*) para el problema de clasificación y *Weather of Izmir*(abreviado *wizmir*) para el problema de regresión.

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 número de observaciones, 690, y los diferentes tipos de variables 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. Este estudio lo haremos en dos partes: una parte dedicada a las variables numéricas y otra parte dedicada a las variables categóricas.

```
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")

numerical_australian <- australian[, c(-1, -4, -8, -9, -11, -12,
  -15)]
categorical_australian <- australian[, c(1, 4, 8, 9, 11, 12)]
output_australian <- australian[, 15]
```

Estudio de las variables numéricas

Son varios los parámetros que nos permiten conocer con más detalle la base de datos aunque no se tenga un nombre descriptivo para cada una de las variables, ejemplo de esto es el número de valores perdidos, la media, mediana, moda y cuartiles. Así mismo se hace necesario comprobar algunas asunciones que hacen determinados algoritmos que emplearemos en este problema de clasificación, como son los test de normalidad del conjunto de datos (para el algoritmo LDA y QDA) y la igualdad de las varianzas (para el algoritmo LDA).

Estudio del número de valores perdidos

En la página web de la base de datos de *Australian Credit Approval* se indica que la base de datos tiene valores perdidos, en esta sección vamos a comprobar qué variables tienen esos valores perdidos.

```
numerical_australian_na <- apply(is.na(numerical_australian),  
  2, sum)  
numerical_australian_na
```

```
A2  A3  A5  A6  A7  A10 A13 A14  
0   0   0   0   0   0   0   0
```

Puesto que en las variables numéricas no hay valores perdidos, podemos seguir el estudio con las variables en este estado sin necesidad de imputar valores.

Estadísticos principales

```
numerical_stats <- summary(numerical_australian)  
numerical_std <- apply(numerical_australian, 2, sd)  
numerical_stats <- rbind(numerical_stats, numerical_std)  
numerical_stats
```

```
          A2           A3  
"Min.    : 16   " "Min.    : 0   "  
"1st Qu.:1942  " "1st Qu.: 15  "  
"Median  :2629  " "Median  : 125 "  
"Mean    :2697  " "Mean    : 1187 "  
"3rd Qu.:3525  " "3rd Qu.: 665 "  
"Max.    :8025  " "Max.    :26335 "  
numerical_std "1554.55973203261" "3069.11004226953"  
          A5           A6  
"Min.    : 1.000  " "Min.    :1.000  "  
"1st Qu.: 4.000  " "1st Qu.:4.000  "  
"Median  : 8.000  " "Median  :4.000  "  
"Mean    : 7.372  " "Mean    :4.693  "  
"3rd Qu.:10.000  " "3rd Qu.:5.000  "  
"Max.    :14.000  " "Max.    :9.000  "  
numerical_std "3.68326478743128" "1.9923160695339"  
          A7           A10  
"Min.    : 0.0   " "Min.    : 0.0  "  
"1st Qu.: 5.0   " "1st Qu.: 0.0  "  
"Median  : 35.0  " "Median  : 0.0  "  
"Mean    : 453.4  " "Mean    : 2.4  "  
"3rd Qu.: 219.8  " "3rd Qu.: 3.0  "  
"Max.    :14415.0 " "Max.    :67.0  "  
numerical_std "1387.90032404432" "4.862940034227"  
          A13          A14
```

```

"Min.    :   0   "  "Min.    :     1.0   "
"1st Qu.:  80   "  "1st Qu.:     1.0   "
"Median : 160   "  "Median :     6.0   "
"Mean   : 184   "  "Mean   : 1018.4   "
"3rd Qu.: 272   "  "3rd Qu.:    396.5   "
"Max.   :2000   "  "Max.   :100001.0   "
numerical_std "172.159273536299" "5210.10259830269"

```

Como podemos ver las variables con mayor varianza son la variable A2,A3,A7 y A14 debido a que los rangos que puede tomar dicha variable son mayores.

Algunos algoritmos como LDA y QDA realizan suposiciones sobre las variables que reciben a la entrada. En el caso LDA supone que los conjuntos que discriminan tienen la misma varianza y son poblaciones normalmente distribuidas.

Test de normalidad

Establezcamos como hipótesis nula que todas las variables numéricas pertenecen a una misma distribución. Para comprobarlo realizaremos el test no paramétrico de Kruskal Wallis.

```
kruskal.test(numerical_australian)
```

```

Kruskal-Wallis rank sum test

data: numerical_australian
Kruskal-Wallis chi-squared = 2713.3, df = 7, p-value
< 2.2e-16

```

Puesto que el p-value es menor que 0.05 rechazamos la hipótesis de que todas las variables siguen la misma distribución. Necesitamos encontrar aquellas variables que sí que sigan una distribución normal, para ello someteremos a todas al test de Shapiro-Wilk.

```
numerical_australian_shapiro <- apply(numerical_australian, 2,
shapiro.test)
numerical_australian_shapiro
```

\$A2

```

Shapiro-Wilk normality test

data: newX[, i]
W = 0.96297, p-value = 3.452e-12

```

\$A3

```

Shapiro-Wilk normality test

data: newX[, i]
W = 0.42625, p-value < 2.2e-16

```

\$A5

```
Shapiro-Wilk normality test
```

```
data: newX[, i]
W = 0.95306, p-value = 5.121e-14
```

\$A6

Shapiro-Wilk normality test

```
data: newX[, i]
W = 0.78023, p-value < 2.2e-16
```

\$A7

Shapiro-Wilk normality test

```
data: newX[, i]
W = 0.34139, p-value < 2.2e-16
```

\$A10

Shapiro-Wilk normality test

```
data: newX[, i]
W = 0.53306, p-value < 2.2e-16
```

\$A13

Shapiro-Wilk normality test

```
data: newX[, i]
W = 0.82069, p-value < 2.2e-16
```

\$A14

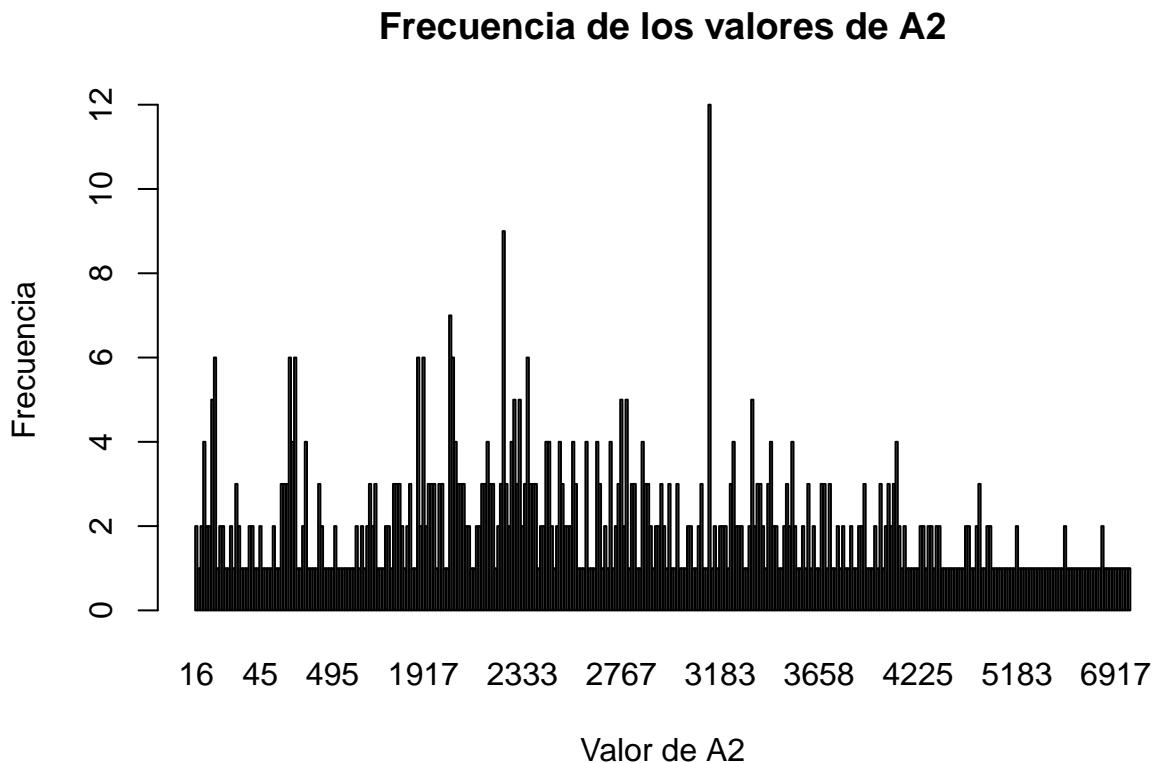
Shapiro-Wilk normality test

```
data: newX[, i]
W = 0.16985, p-value < 2.2e-16
```

Dado que todos los p-value de todas las variables es menor que 0.05 deducimos que ninguna de las varibales sigue una distribución normal por lo que no se espera que ni el algoritmo LDA ni el algoritmo QDA funcionen bien para la clasificación.

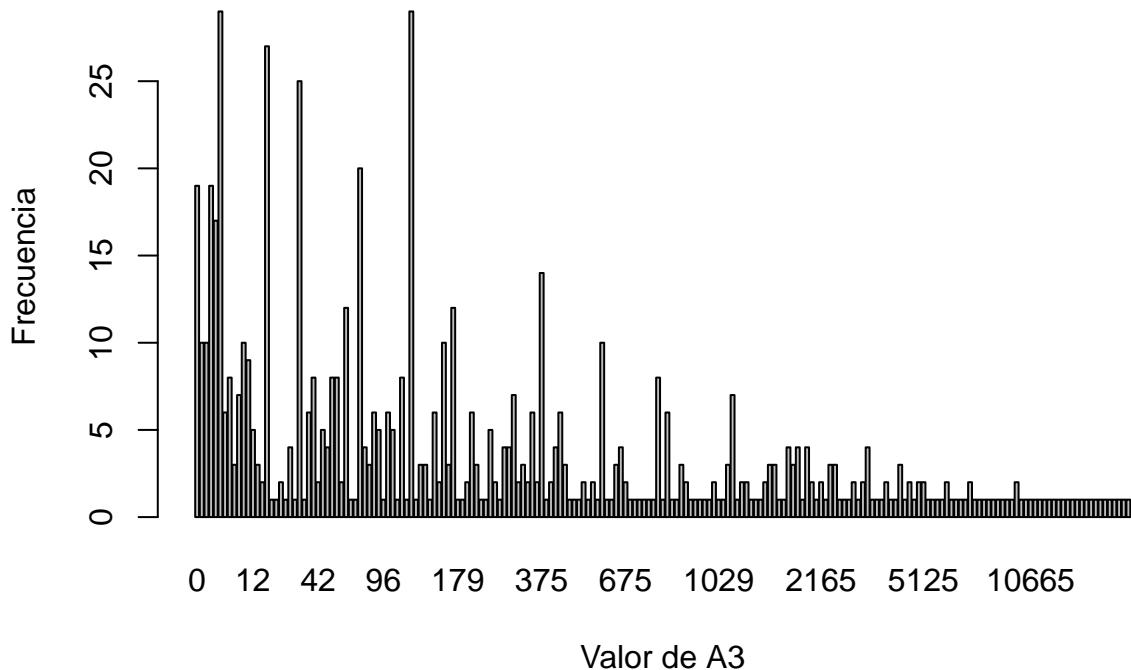
A continuación se muestran los gráficos de barras de todas las variables numéricas:

```
barplot(table(numerical_australian[, 1]), main = "Frecuencia de los valores de A2",
       xlab = "Valor de A2", ylab = "Frecuencia")
```



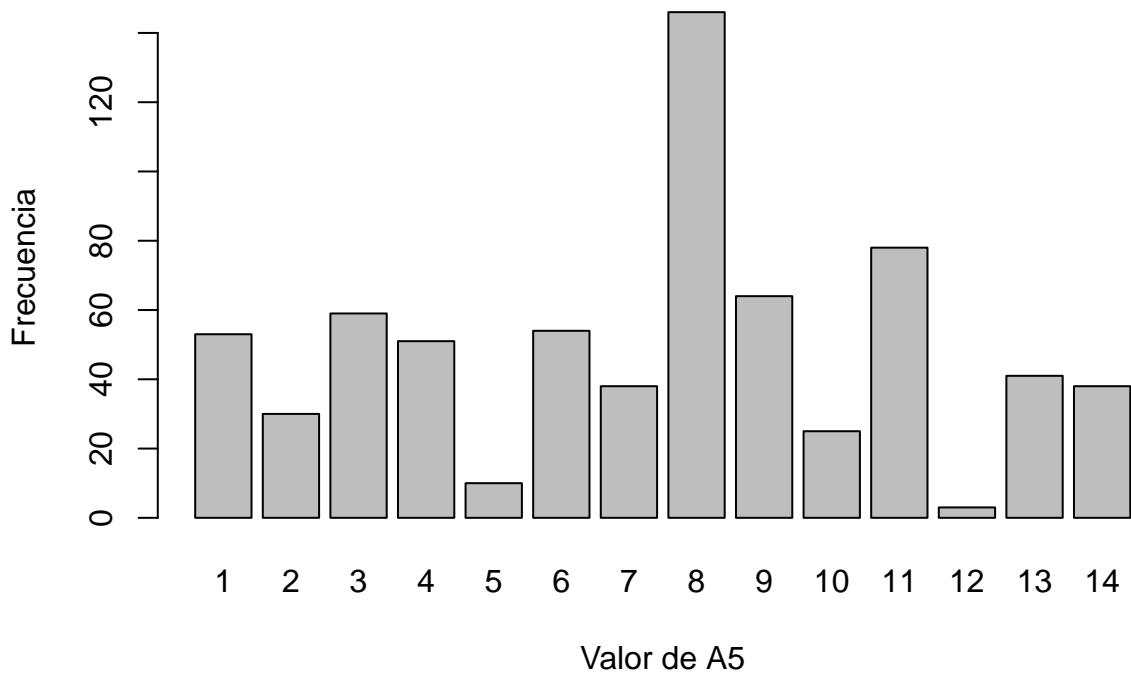
```
barplot(table(numerical_australian[, 2]), main = "Frecuencia de los valores de A3",
       xlab = "Valor de A3", ylab = "Frecuencia")
```

Frecuencia de los valores de A3



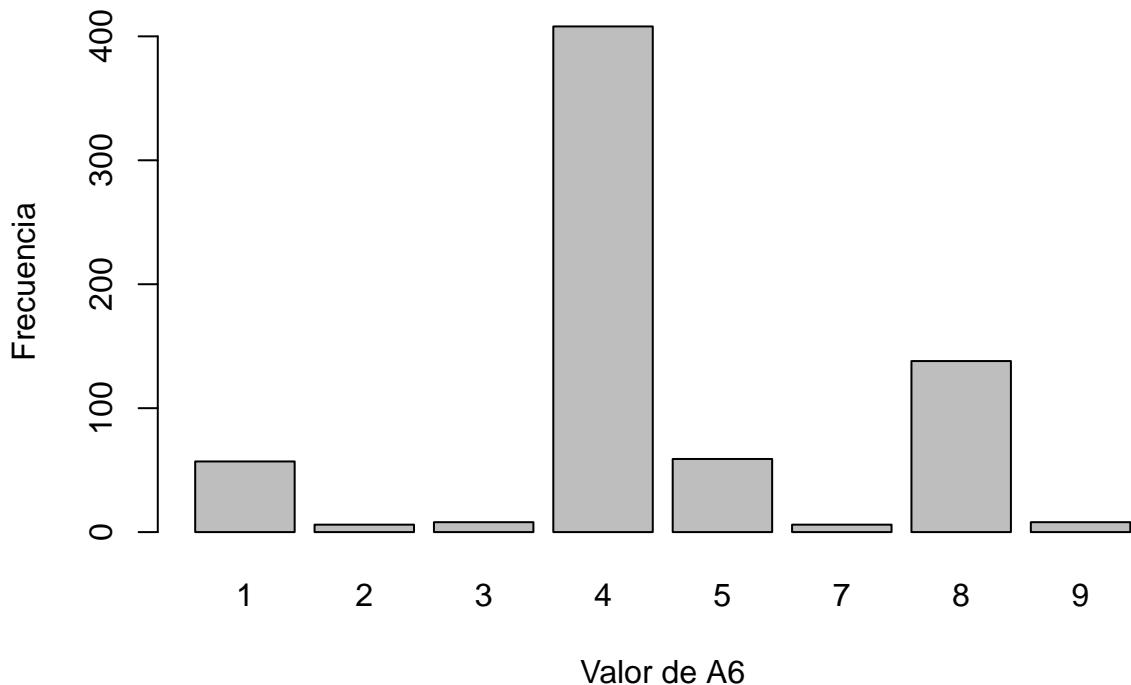
```
barplot(table(numerical_australian[, 3]), main = "Frecuencia de los valores de A5",
       xlab = "Valor de A5", ylab = "Frecuencia")
```

Frecuencia de los valores de A5



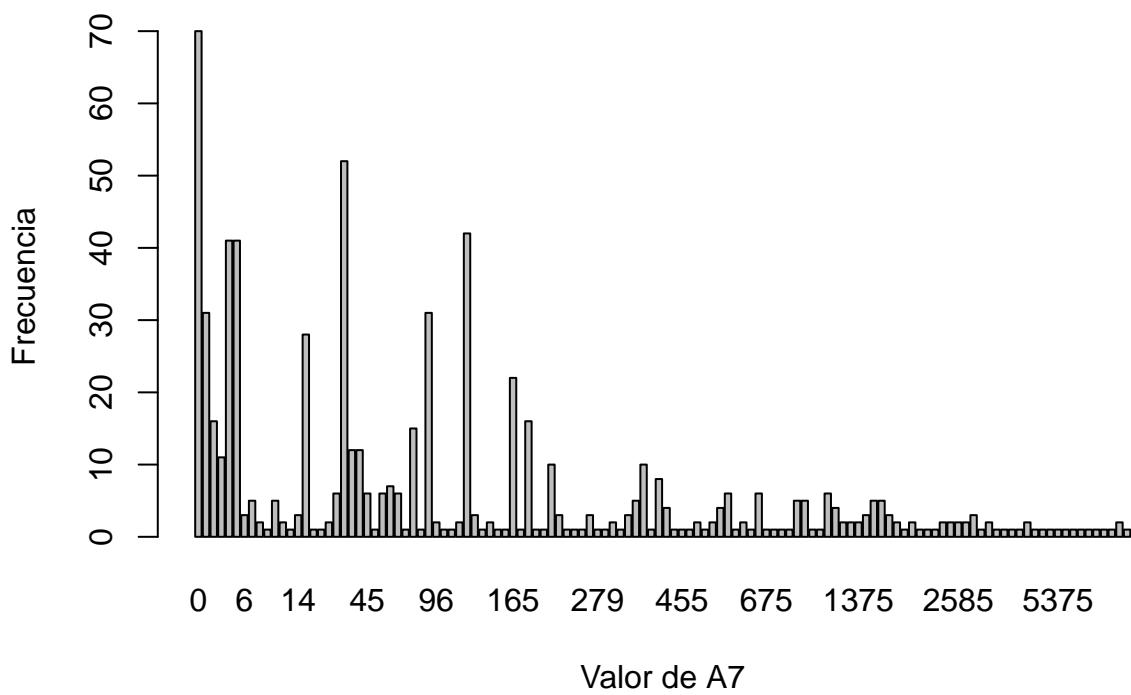
```
barplot(table(numerical_australian[, 4]), main = "Frecuencia de los valores de A6",
       xlab = "Valor de A6", ylab = "Frecuencia")
```

Frecuencia de los valores de A6



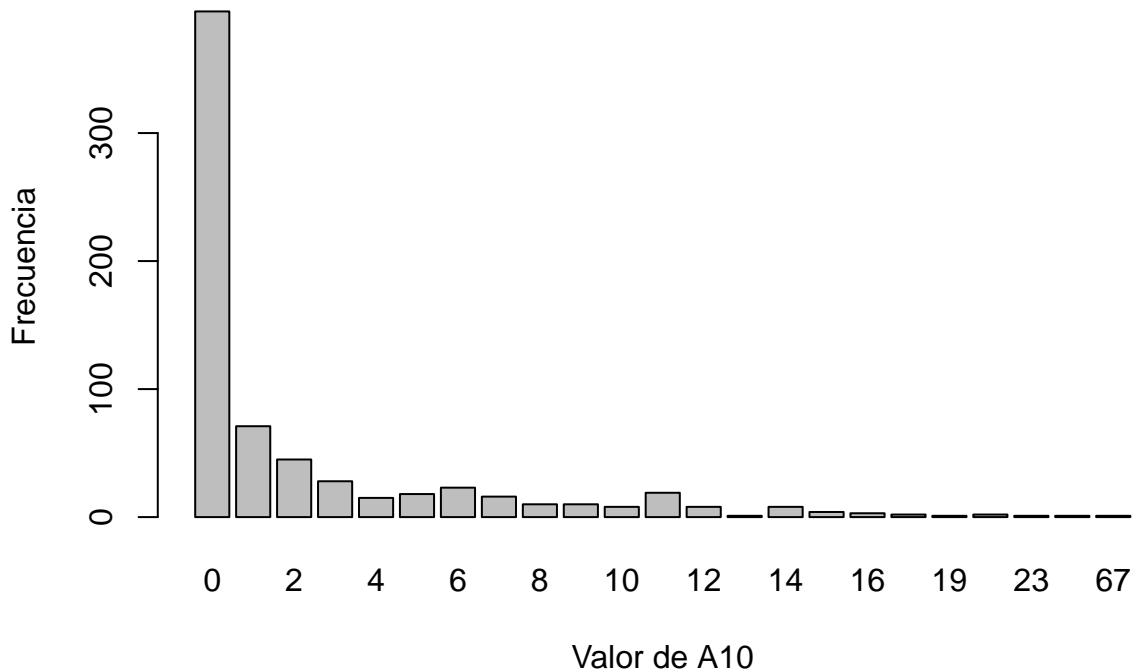
```
barplot(table(numerical_australian[, 5]), main = "Frecuencia de los valores de A7",
       xlab = "Valor de A7", ylab = "Frecuencia")
```

Frecuencia de los valores de A7



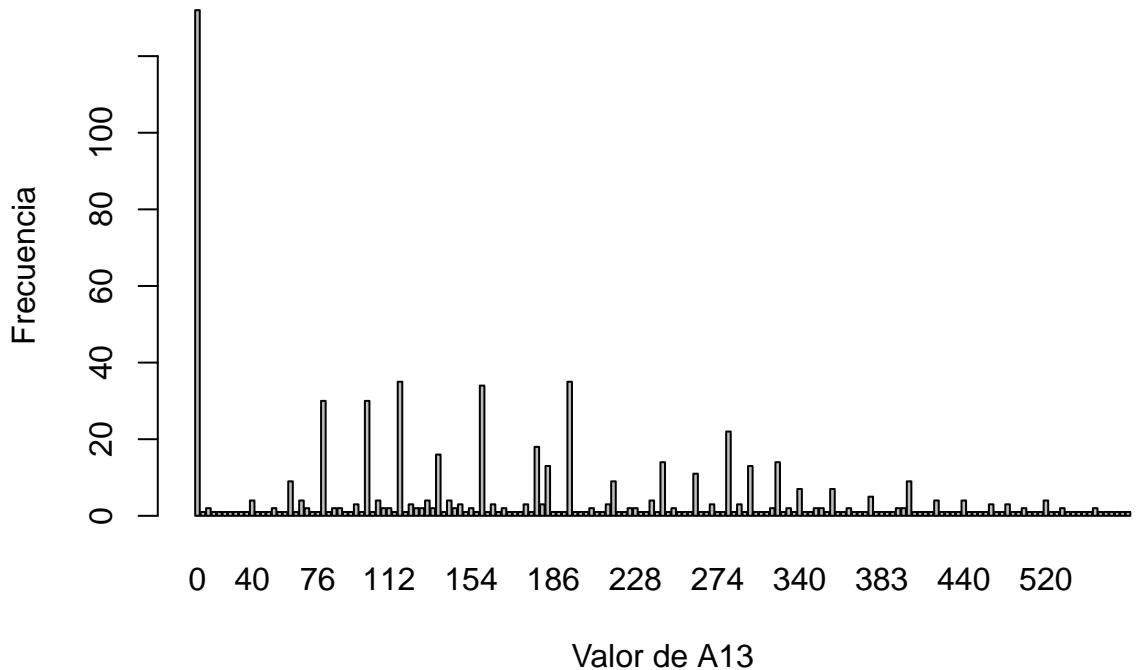
```
barplot(table(numerical_australian[, 6]), main = "Frecuencia de los valores de A10",
       xlab = "Valor de A10", ylab = "Frecuencia")
```

Frecuencia de los valores de A10



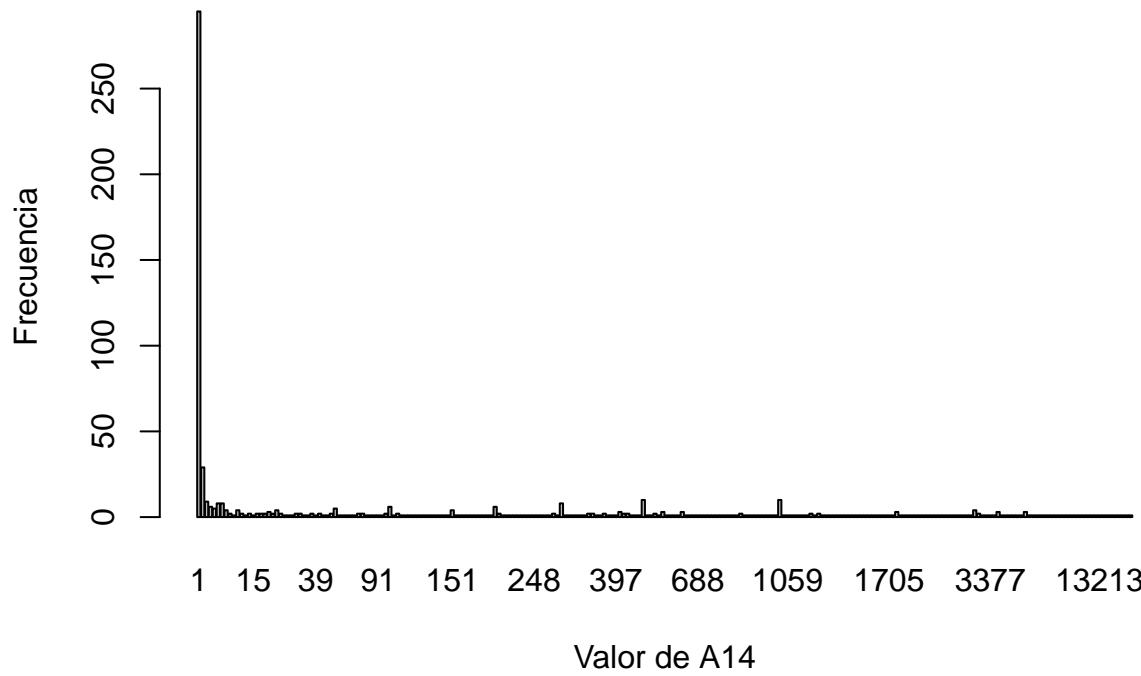
```
barplot(table(numerical_australian[, 7]), main = "Frecuencia de los valores de A13",
       xlab = "Valor de A13", ylab = "Frecuencia")
```

Frecuencia de los valores de A13



```
barplot(table(numerical_australian[, 8]), main = "Frecuencia de los valores de A14",
       xlab = "Valor de A14", ylab = "Frecuencia")
```

Frecuencia de los valores de A14



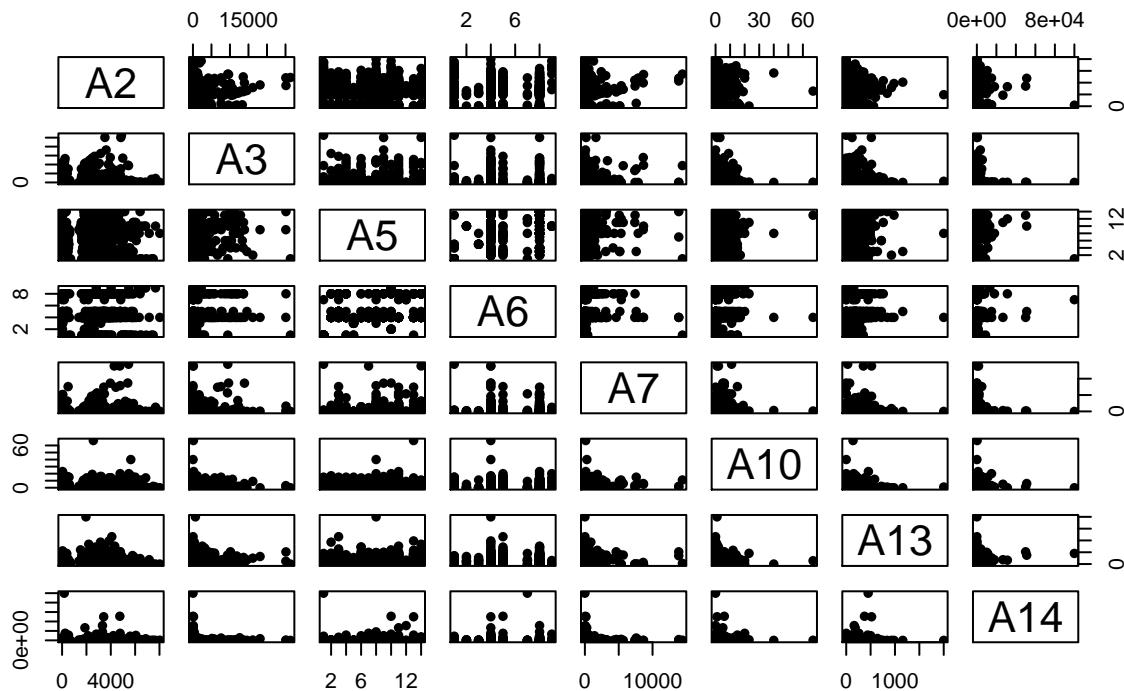
Como ya habíamos visto de forma numérica, por las gráficas no se puede decir que ninguna de las variables numéricas sea una distribución normal.

Información gráfica

Vamos a mostrar los valores que toman cada una de las variables y a compararlas entre ellas con un scatter plot.

```
pairs(numerical_australian, main = "Comparación de las variables numéricas con la salida",  
      pch = 16)
```

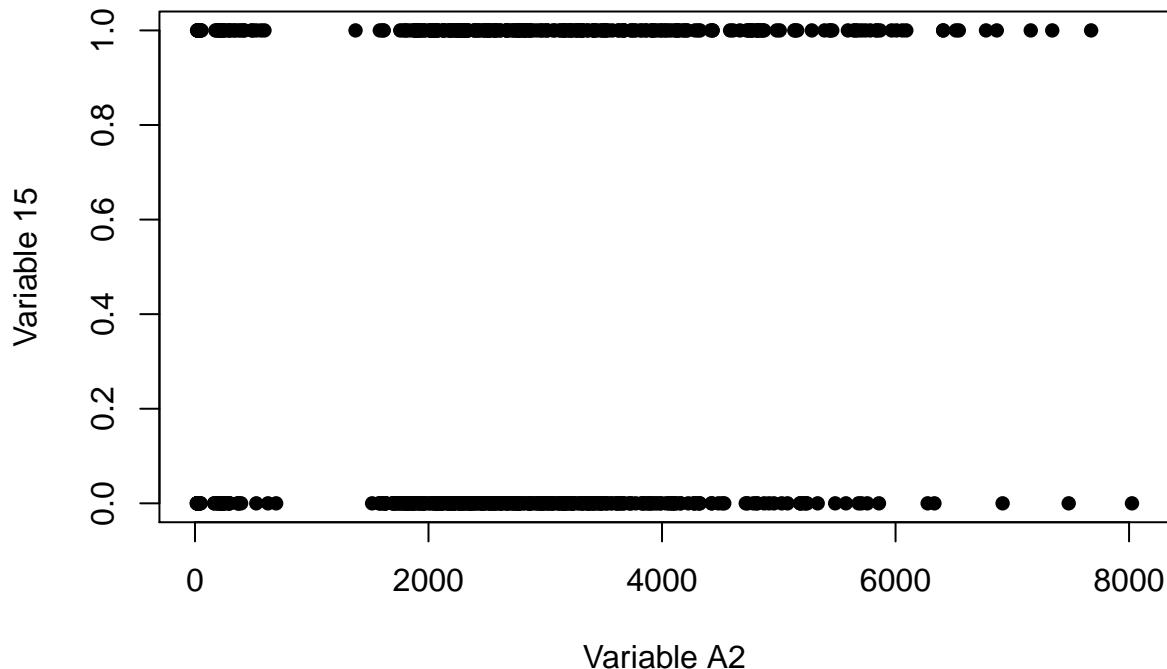
Comparación de las variables numéricas con la salida



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.

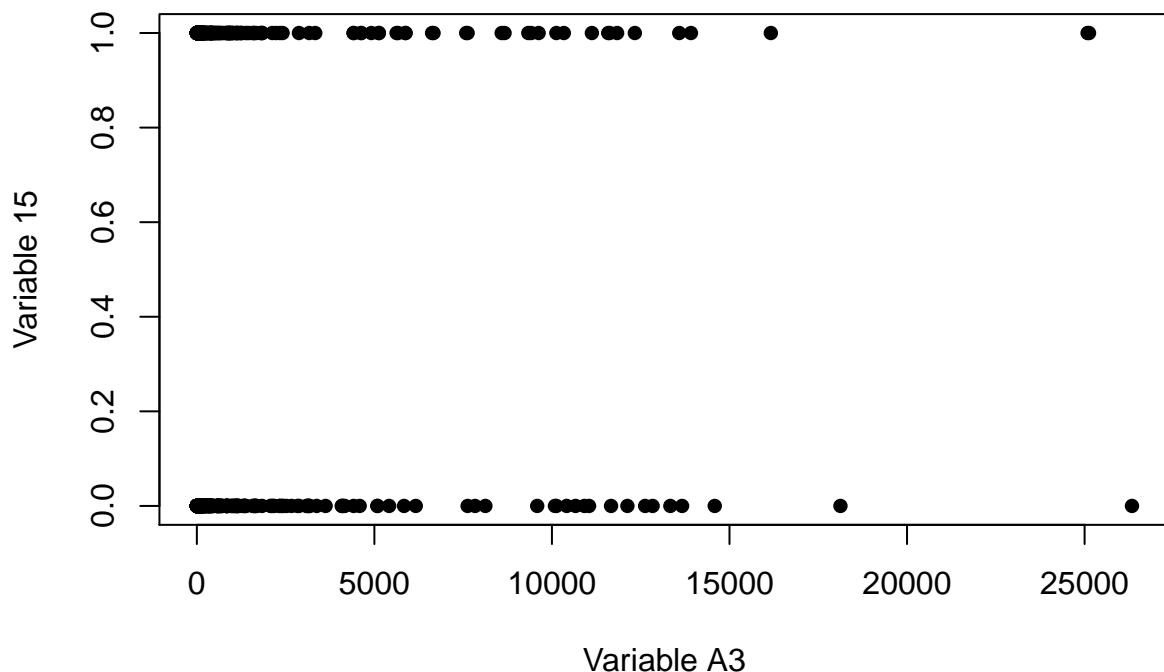
```
plot(australian[, 2], australian[, 15], main = "Comparación A2 con la salida",
      pch = 16, xlab = "Variable A2", ylab = "Variable 15")
```

Comparación A2 con la salida



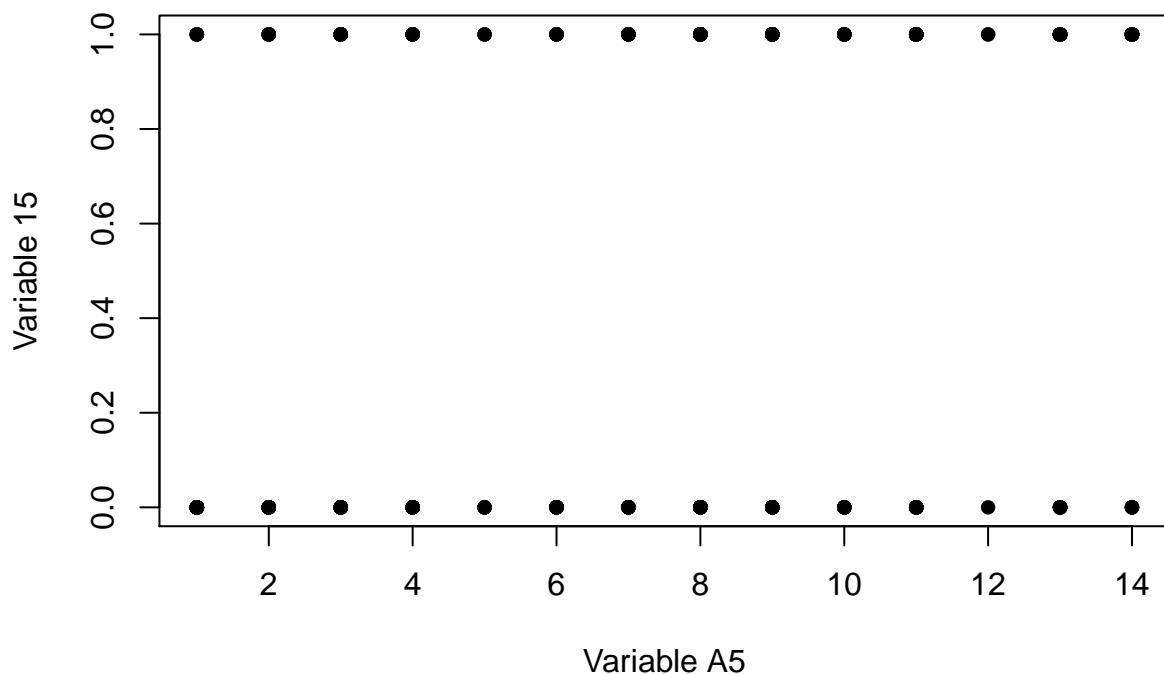
```
plot(australian[, 3], australian[, 15], main = "Comparación A3 con la salida",
      pch = 16, xlab = "Variable A3", ylab = "Variable 15")
```

Comparación A3 con la salida



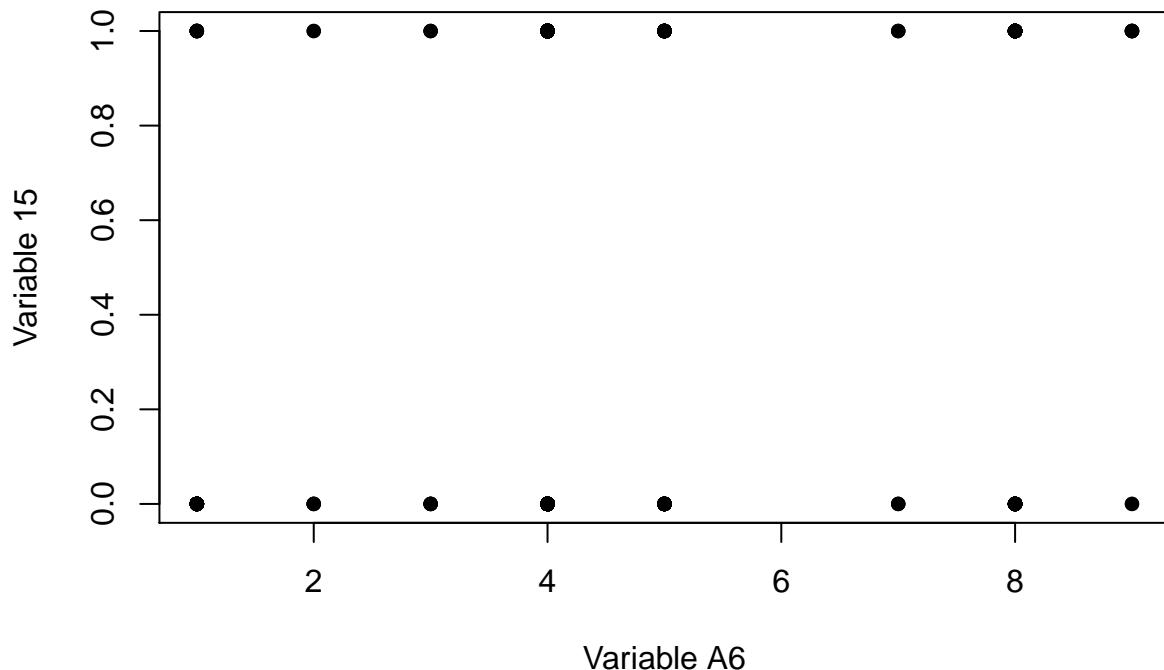
```
plot(australian[, 5], australian[, 15], main = "Comparación A5 con la salida",
      pch = 16, xlab = "Variable A5", ylab = "Variable 15")
```

Comparación A5 con la salida



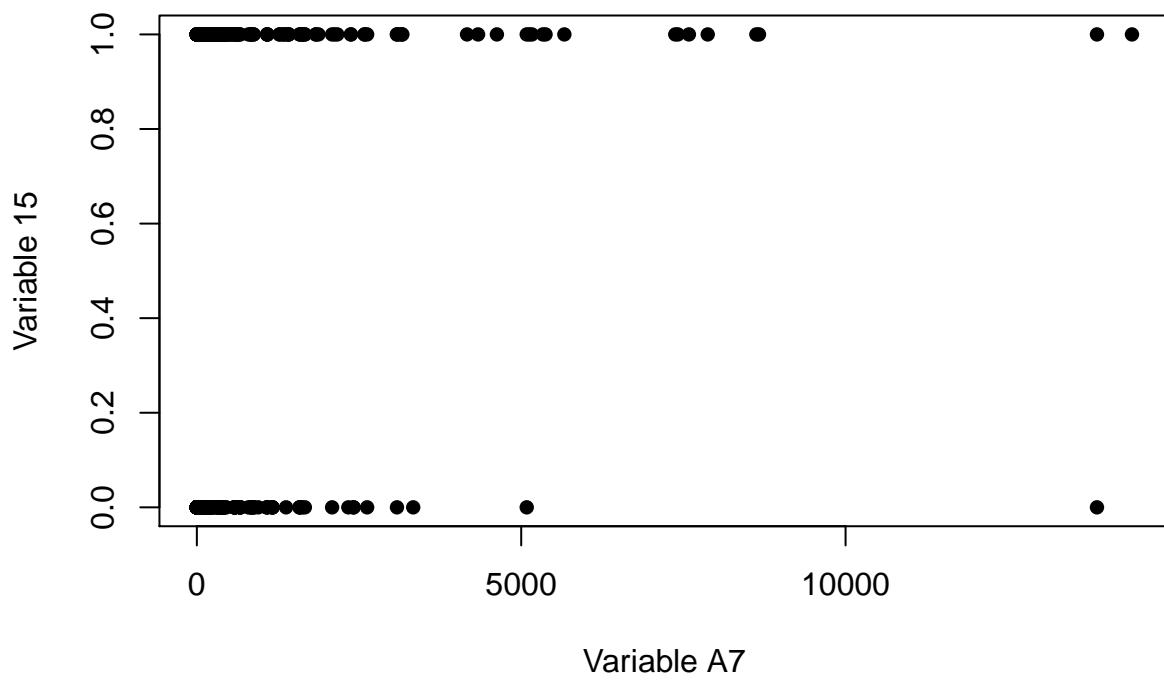
```
plot(australian[, 6], australian[, 15], main = "Comparación A6 con la salida",
      pch = 16, xlab = "Variable A6", ylab = "Variable 15")
```

Comparación A6 con la salida



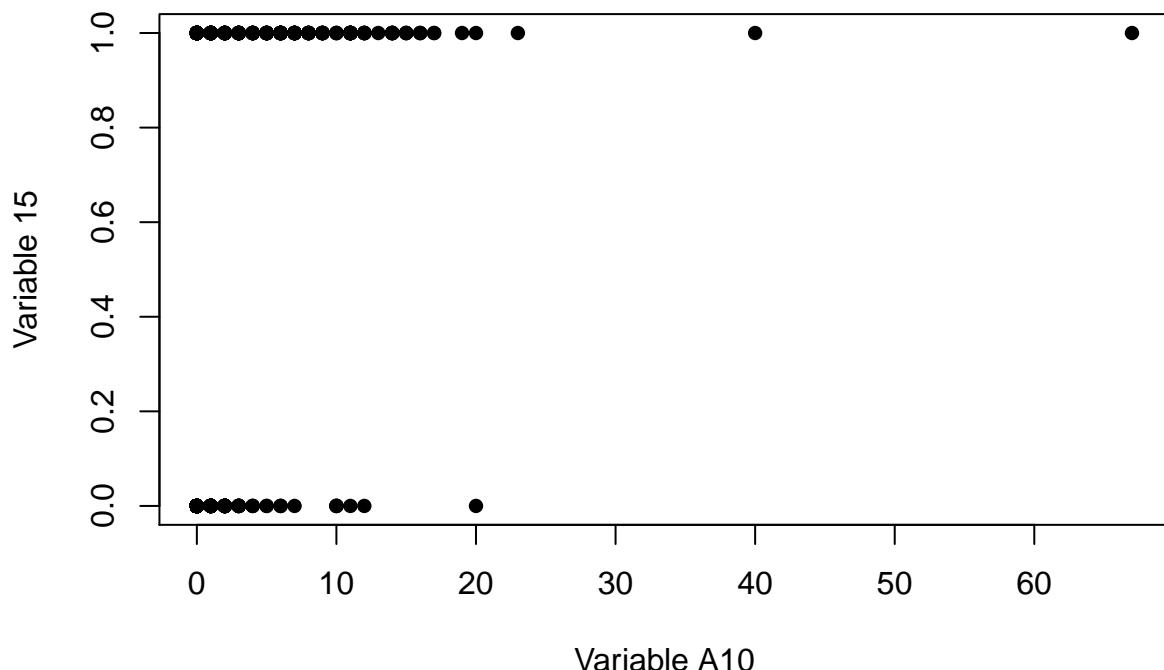
```
plot(australian[, 7], australian[, 15], main = "Comparación A7 con la salida",
      pch = 16, xlab = "Variable A7", ylab = "Variable 15")
```

Comparación A7 con la salida



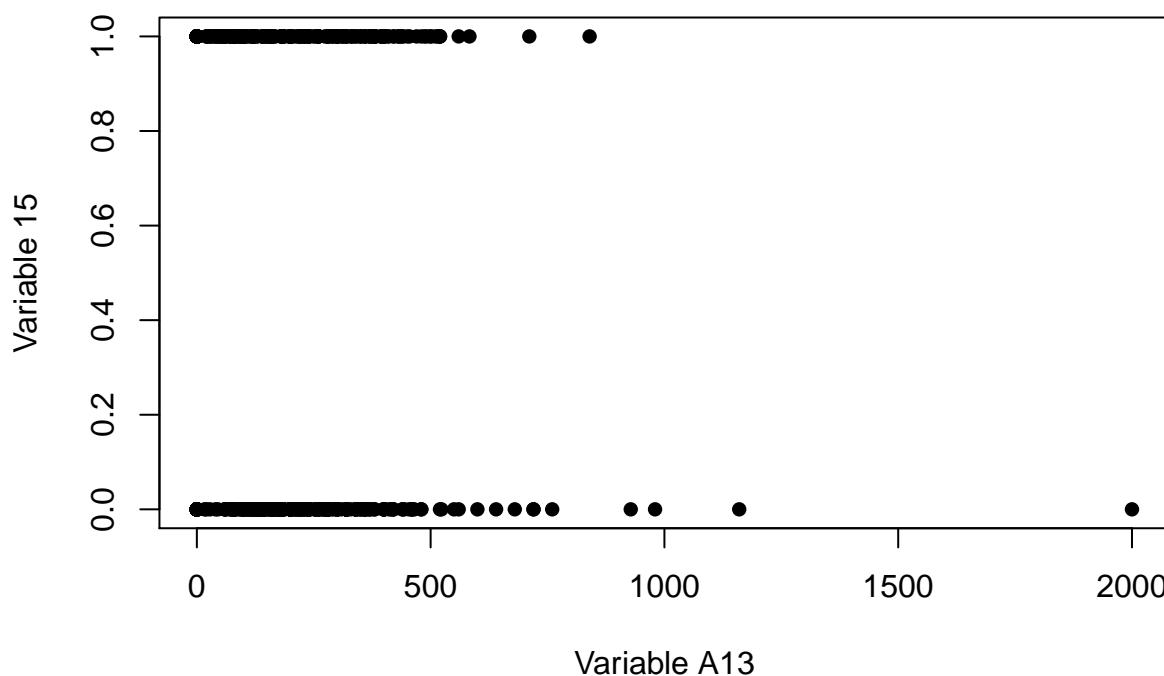
```
plot(australian[, 10], australian[, 15], main = "Comparación A10 con la salida",
      pch = 16, xlab = "Variable A10", ylab = "Variable 15")
```

Comparación A10 con la salida



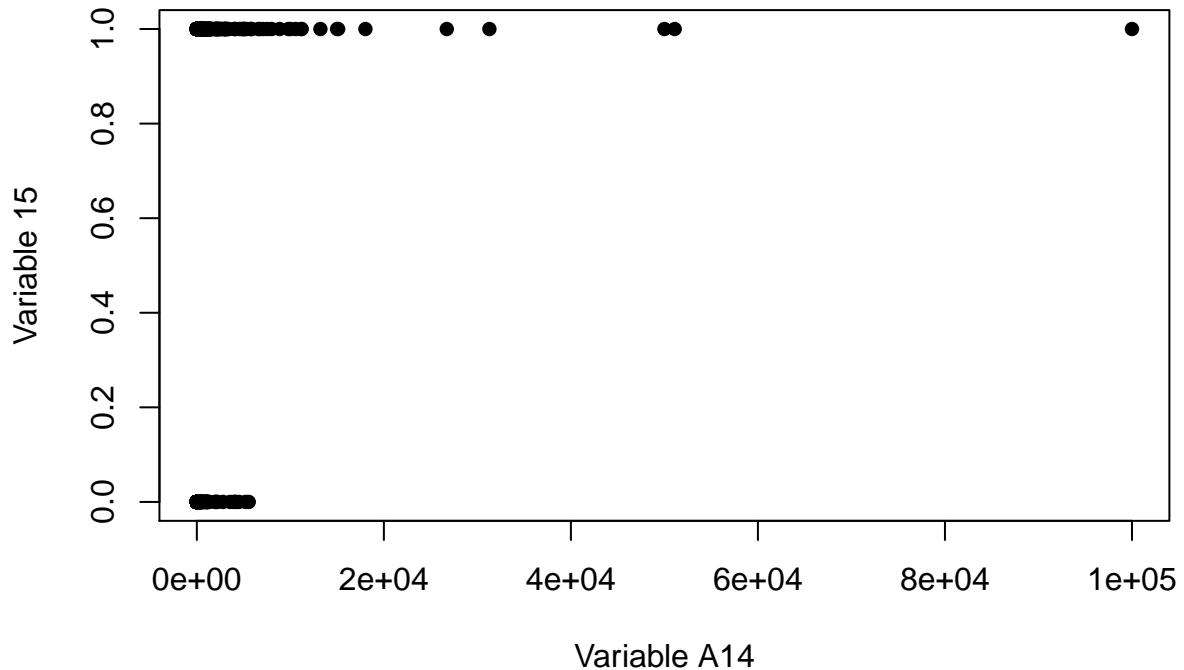
```
plot(australian[, 13], australian[, 15], main = "Comparación A13 con la salida",
      pch = 16, xlab = "Variable A13", ylab = "Variable 15")
```

Comparación A13 con la salida



```
plot(australian[, 14], australian[, 15], main = "Comparación A15 con la salida",
      pch = 16, xlab = "Variable A14", ylab = "Variable 15")
```

Comparación A15 con la salida



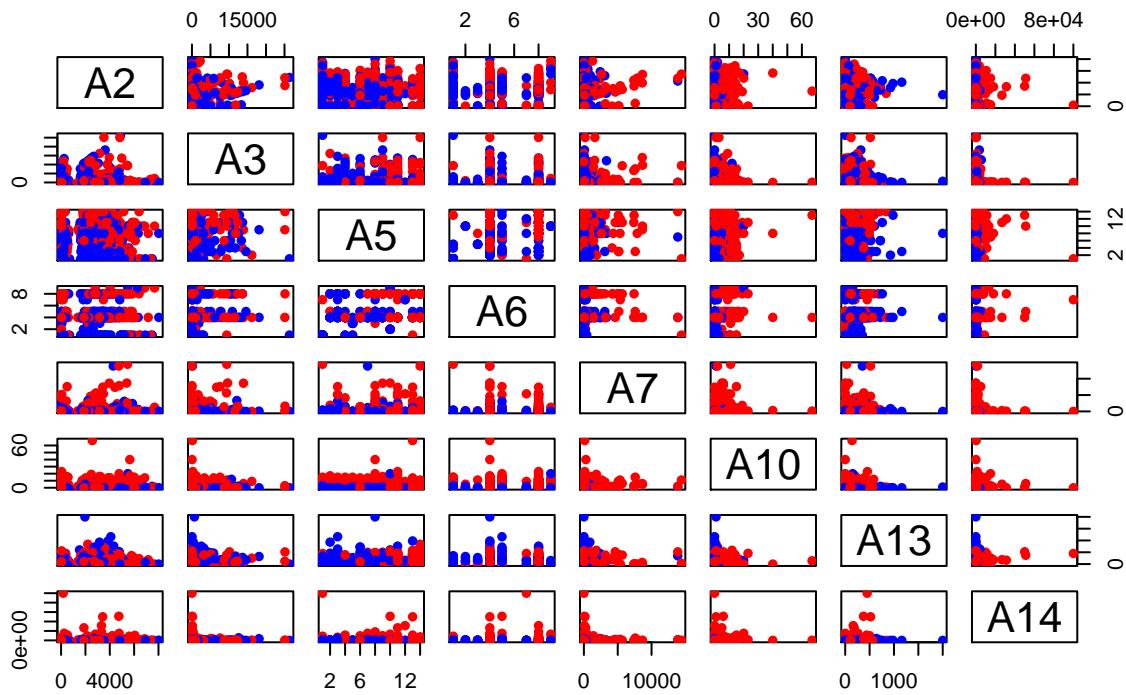
No podemos decir que ninguna de las variables numéricas sea realmente discriminante para la salida. Sin embargo, no descartaría la variable 14 para que intervenga un modelo de clasificación, porque para valores altos solamente da salida positiva para la clase 1.

Estudio de las correlaciones

Volveremos 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 el dataset australiano", col = ifelse(australian[, 15] == 1, "red", "blue"), pch = 16)
```

Comparación de las variables numéricas entre ellas



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.

Vamos a comprobar de forma numérica que no existe esta correlación, para ello calcular la correlación entre todos los pares de variables

```
cor(numerical_australian, method = "pearson")
```

	A2	A3	A5	A6
A2	1.000000000	0.014978547	-0.05862121	-0.001506815
A3	0.014978547	1.000000000	0.05232107	0.065151425
A5	-0.058621212	0.052321072	1.00000000	0.402283761
A6	-0.001506815	0.065151425	0.40228376	1.000000000
A7	0.073636240	0.149224719	0.09780098	0.077021503
A10	0.117647696	-0.009713481	0.15016586	0.098840728
A13	-0.025026165	-0.014903065	0.08813968	0.070661763
A14	0.002460758	-0.035580640	0.03073527	0.064840849
	A7	A10	A13	A14
A2	0.073636240	0.117647696	-0.025026165	0.002460758
A3	0.149224719	-0.009713481	-0.014903065	-0.035580640
A5	0.097800977	0.150165862	0.088139683	0.030735269
A6	0.077021503	0.098840728	0.070661763	0.064840849
A7	1.000000000	0.104332170	0.005644437	-0.018071625
A10	0.104332170	1.000000000	-0.119808064	0.063692439
A13	0.005644437	-0.119808064	1.000000000	0.065608872
A14	-0.018071625	0.063692439	0.065608872	1.000000000

Como ya pensábamos nos existe correlación significativa entre ningún par de variables, pero llama la atención la correlación del 0.4 entre la variable A5 y la variable A6.

Estudio de las variables categóricas.

Las variables categóricas 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 y el valor más frecuente o moda. Pero antes que nada vamos a estudiar si el conjunto de los datos categóricos contienen valores pedidos

Estudio de los valores perdidos

```
categorical_australian_na <- apply(is.na(cbind(categorical_australian,
  output_australian)), 2, sum)
categorical_australian_na
```

A1	A4	A8
0	0	0
A9	A11	A12
0	0	0
output_australian		
0		

Las variables categóricas no tienen valores perdidos, dado que en la página web de esta base de datos en UCI indica que si que tiene valores perdido, asumimos que la copia de la base de datos, obtenida de Prado, que tenemos se le han imputado los valores perdidos antes de proporcionárnosla a los alumnos.

Estudio de los estadísticos principales

Como mencionamos anteriormente los estadísticos que nos interesan en estas variables categóricas son: el mínimo, el máximo, los cuartiles y la moda. Por suerte para nosotros todos estos valores los podemos obtener con un solo comando excepto la moda que la obtendremos aparte.

```
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"

```

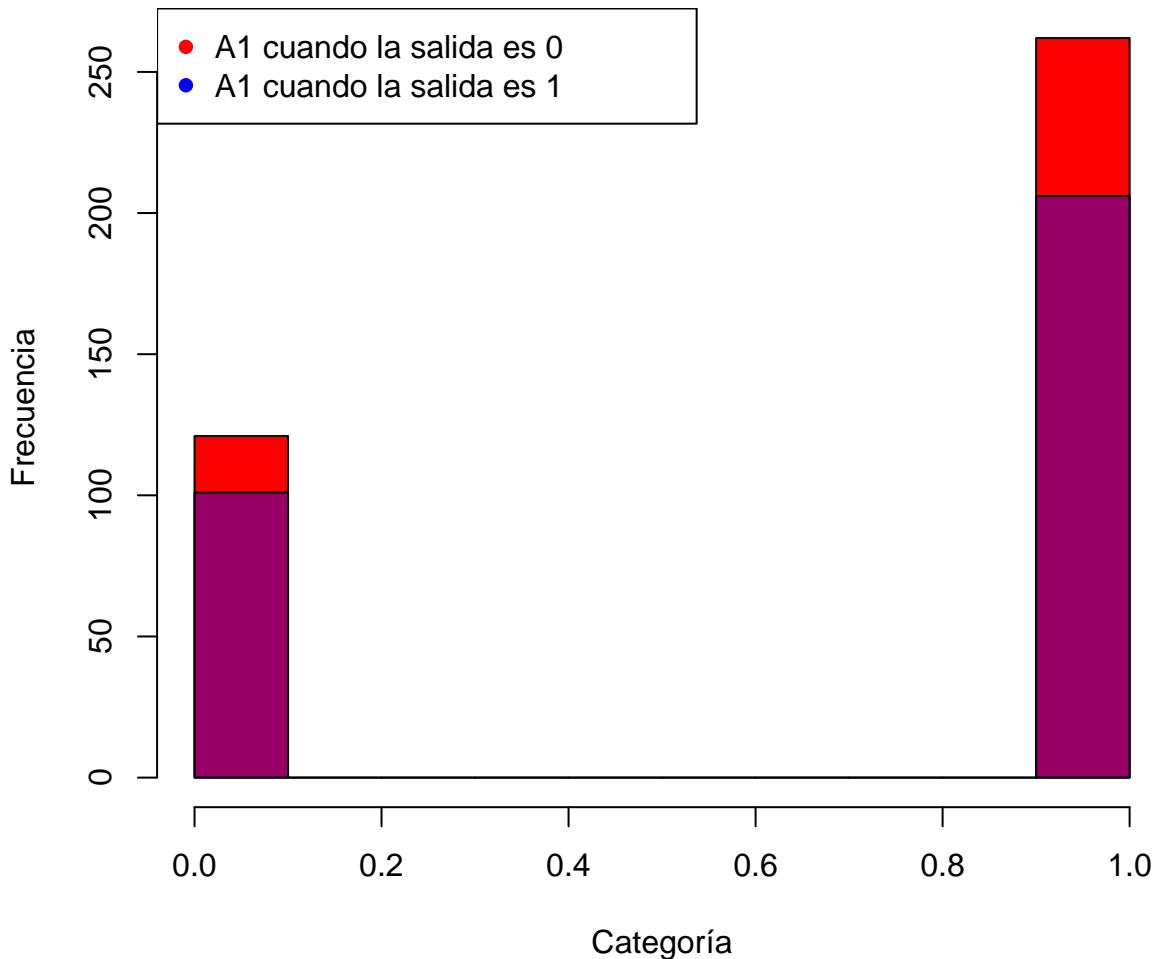
La información anterior nos ilustra como se distribuye cada una de las variables pero no como se relacionan con la salida, para ello dibujaremos gráficos en donde comparamos la frecuencia de cada valor con respecto al valor que toma la salida.

```

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)

```

Frecuencia de la variable A1



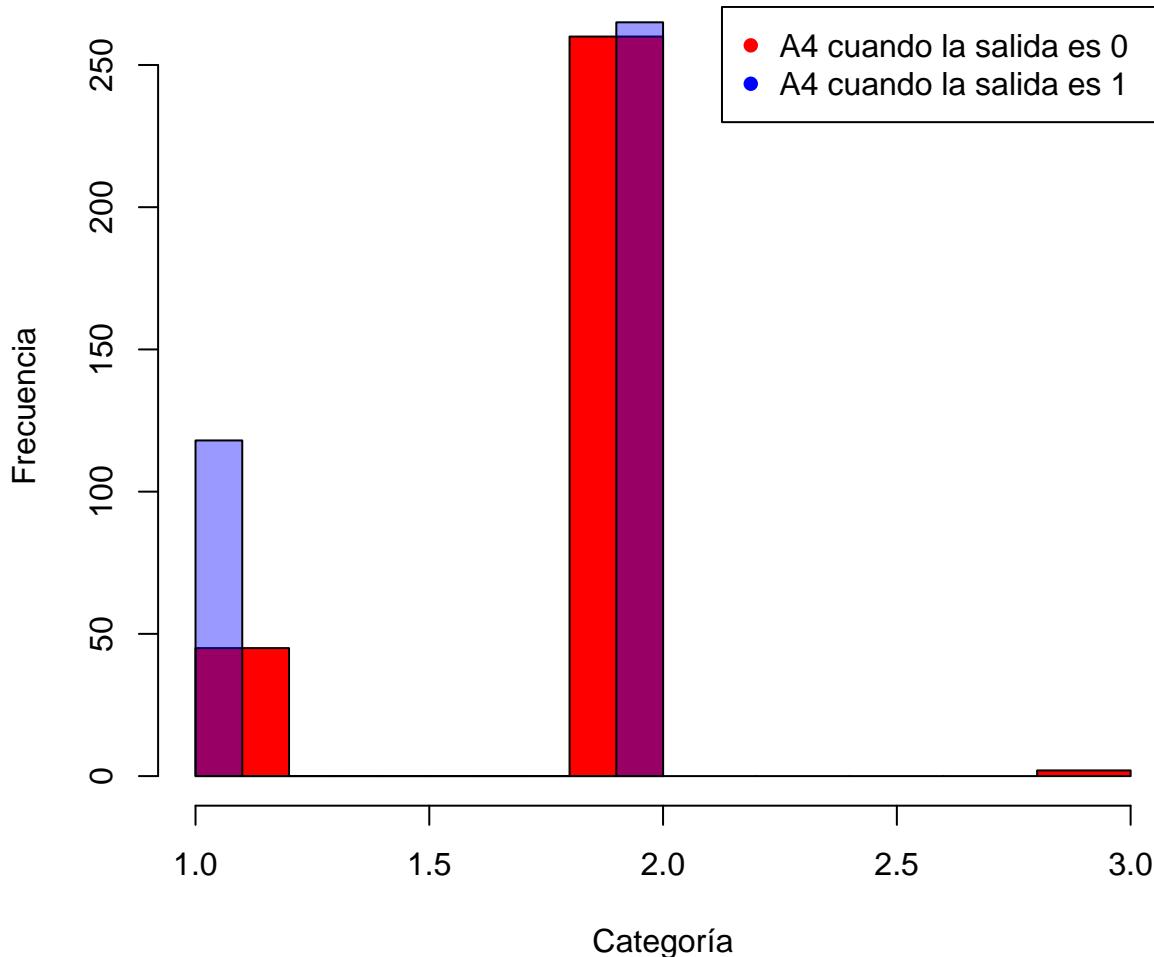
```

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 = 0.5, col = c("red", "blue"), pch = 16)

```

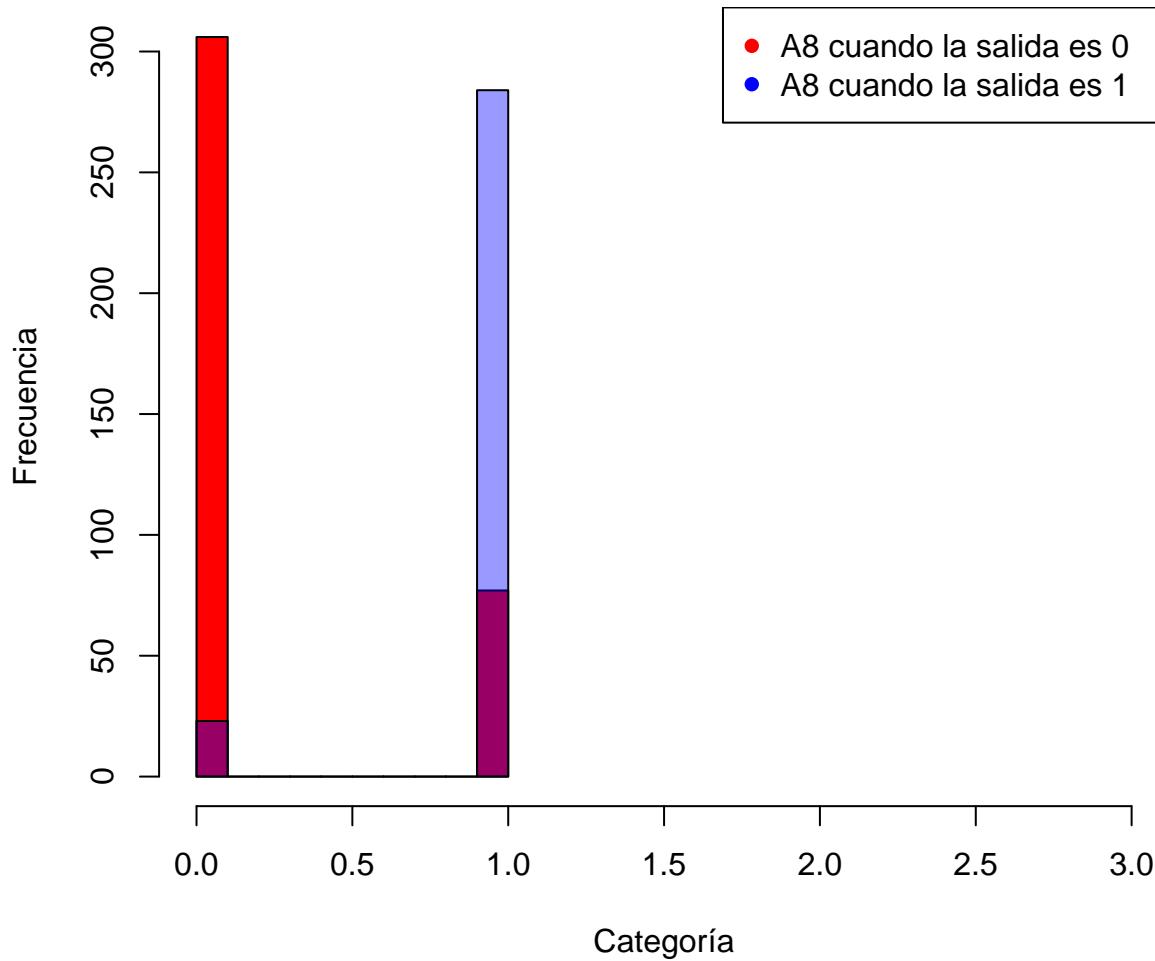
```
text.width = 0.8, col = c("red", "blue"), pch = 16)
```

Frecuencia de la variable A4



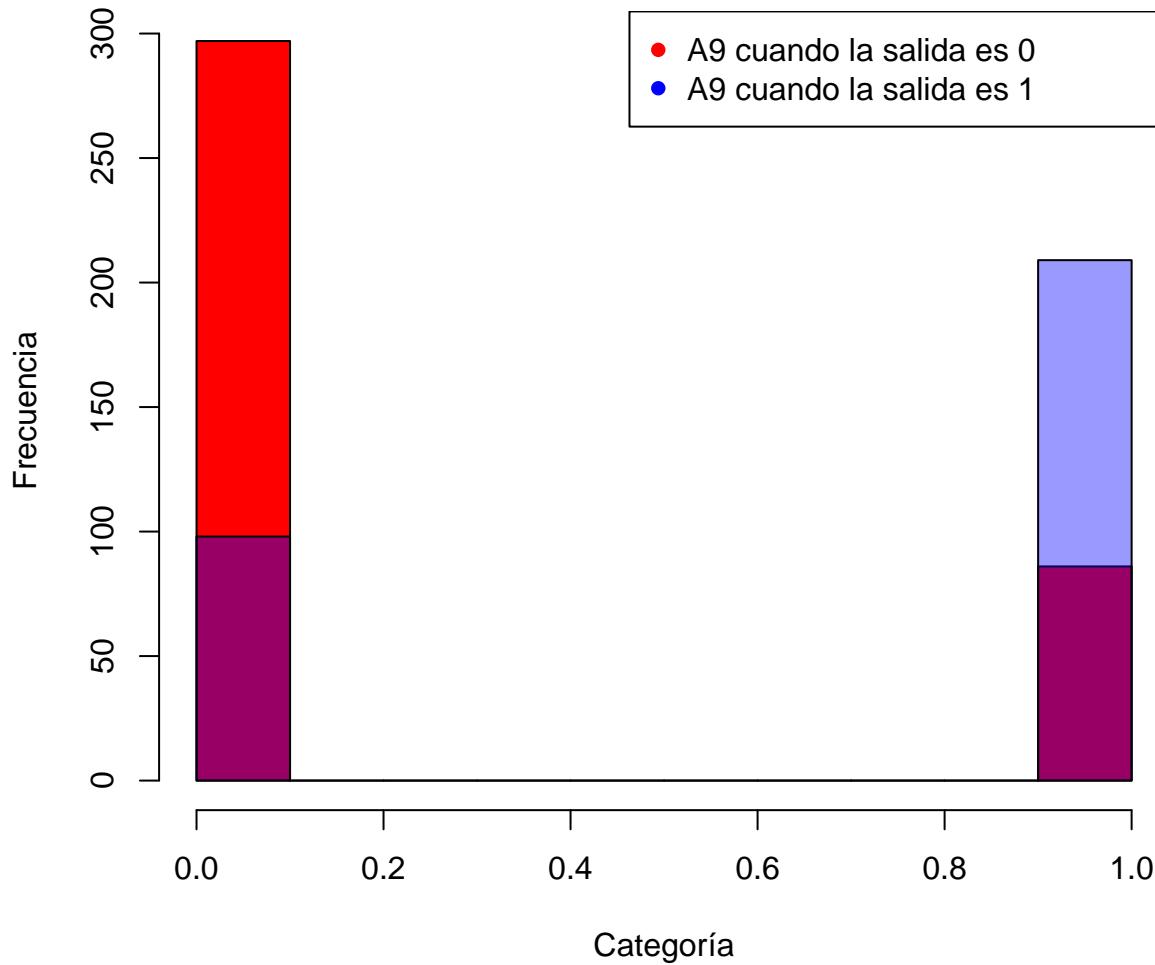
```
hist(australian[which(australian[, 15] == 0), 8], col = rgb(1,
0, 0, 1), main = "Frecuencia de la variable A8", ylab = "Frecuencia",
xlab = "Categoría", xlim = c(0, 3))
hist(australian[which(australian[, 15] == 1), 8], col = rgb(0,
0, 1, 0.4), add = T)
legend("topright", legend = c("A8 cuando la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

Frecuencia de la variable A8



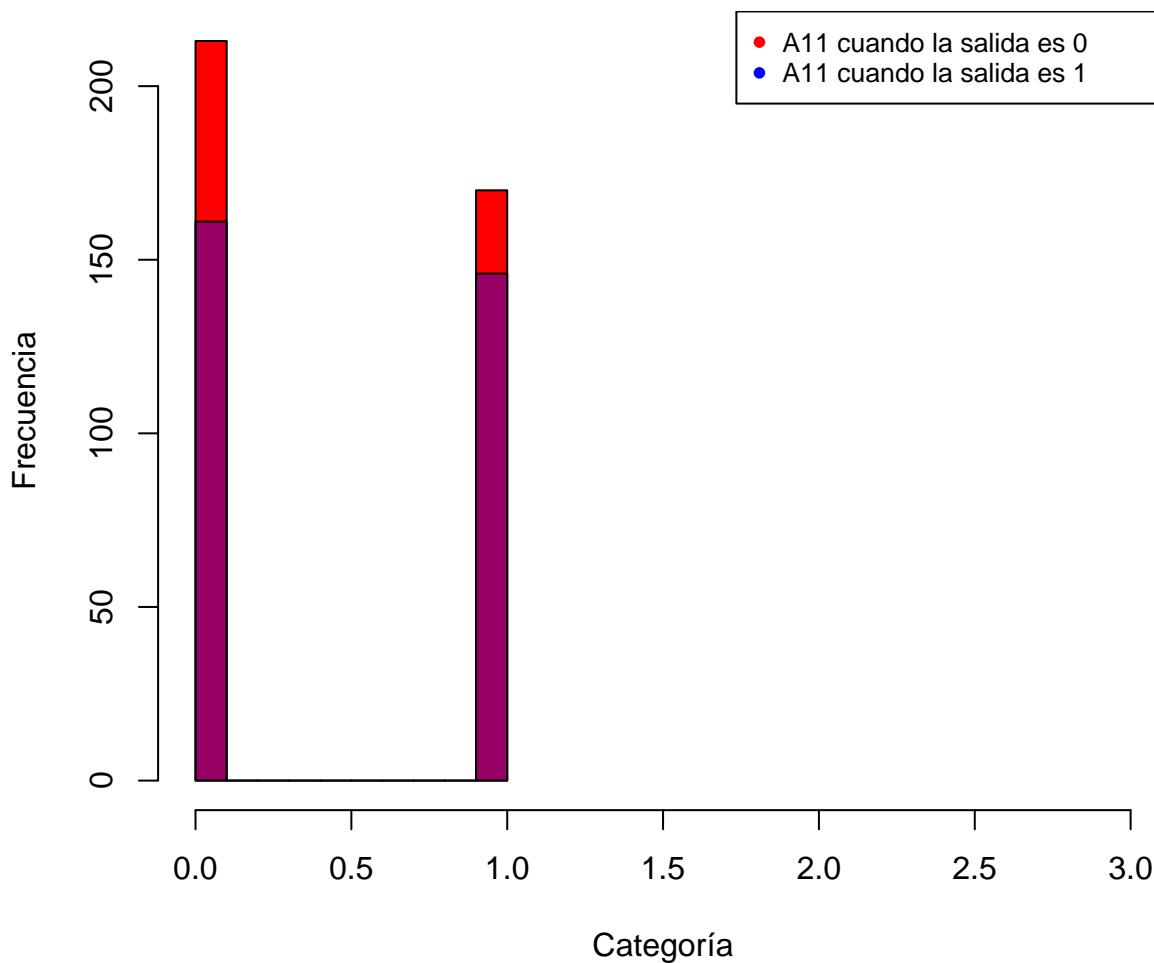
```
hist(australian[which(australian[, 15] == 0), 9], col = rgb(1,
0, 0, 1), main = "Frecuencia de la variable A9", ylab = "Frecuencia",
xlab = "Categoría")
hist(australian[which(australian[, 15] == 1), 9], col = rgb(0,
0, 1, 0.4), add = T)
legend("topright", legend = c("A9 cuando la salida es 0", "A9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

Frecuencia de la variable A9



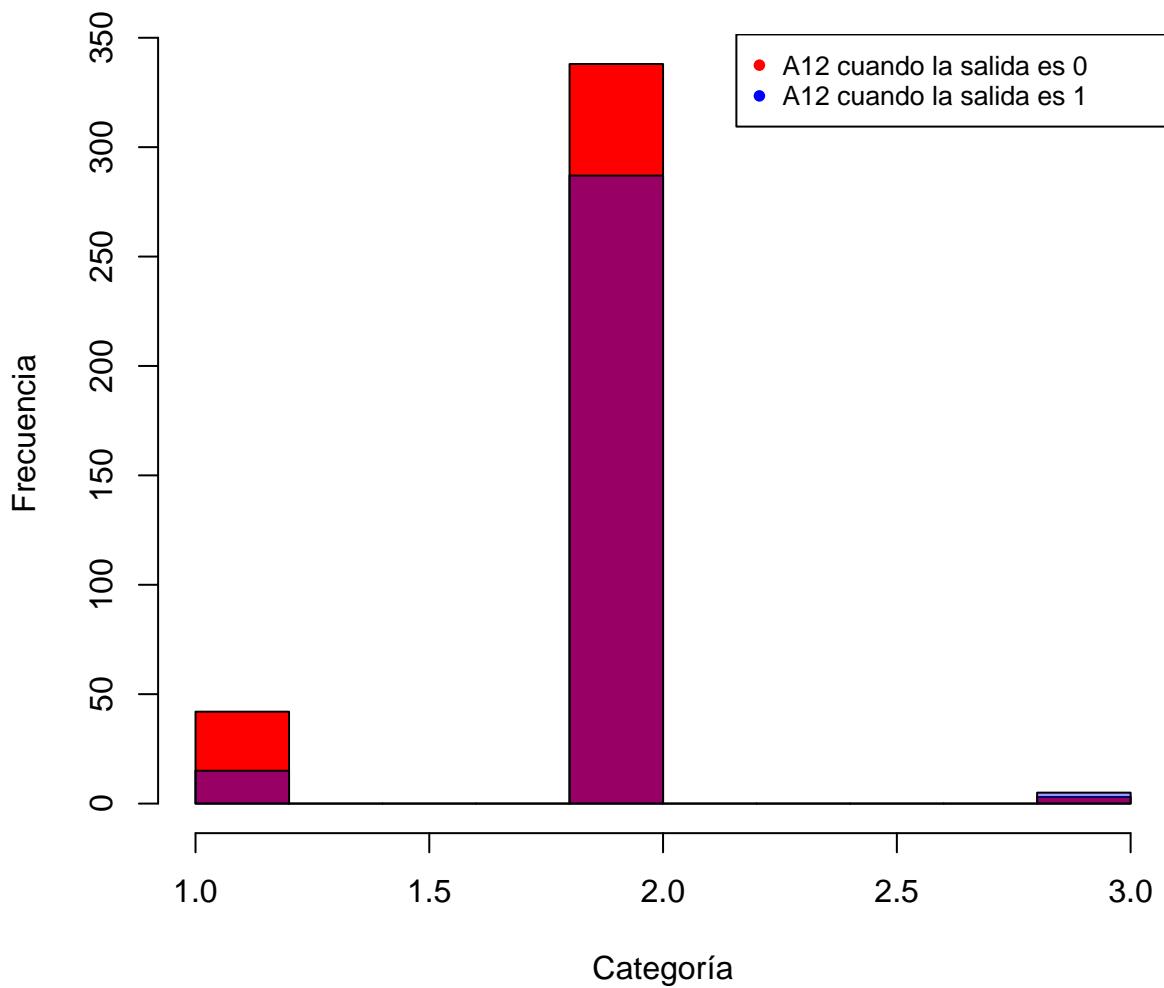
```
hist(australian[which(australian[, 15] == 0), 11], col = rgb(1,
  0, 0, 1), main = "Frecuencia de la variable A11", ylab = "Frecuencia",
  xlab = "Categoría", xlim = c(0, 3))
hist(australian[which(australian[, 15] == 1), 11], col = rgb(0,
  0, 1, 0.4), add = T)
legend("topright", legend = c("A11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

Frecuencia de la variable A11



```
hist(australian[which(australian[, 15] == 0), 12], col = rgb(1,
  0, 0, 1), main = "Frecuencia de la variable A12", 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)
```

Frecuencia de la variable A12



Viendo las gráficas anteriores, vemos que la gran mayoría de las variables no discriminan bien la salida, pero las variables A8 y A9 si que discriminan bien la mayoría de los casos.

De este estudio solo podemos concluir que pueden actuar como buenos discriminantes las variables categóricas A8 y A9 así como la variable A14. A primera vista no esperaría buenos resultados de ningún modelo, excepto del KNN que al ajustarse mejor a los datos puede ser más flexible.

Wizmir (Weather of Izmir)

La base de datos de *Weather of Izmir* contiene datos referentes a distintas variables relacionadas con el tiempo aatmosférico, a saber:

- Temperatura máxima (Max_temperature) real[36.7,105.0]
- Temperatura mínima (Min_temperature) real[15.8,78.6]
- Rocío (Dewpoint) real[13.6,64.4]
- Precipitación (Precipitation) real[0.0,7.6]
- Presión a nivel del mar (Sea_level_pressure) real[29.26,30.48]
- Presión normal (Standard_pressure) real[2.3,10.1]
- Visibilidad (Visibility) real[0.92,29.1]
- Velocidad del viento (Wind_speed) real[4.72,68.8]
- Velocidad máxima del viento (Max_wind_speed) real[16.11,55.24]
- Temperatura media (Mean_temperature) real[29.4,89.9]

El objetivo con esta base de datos es calcular la temperatura media a partir de las demás variables de datos.

Al contrario que el conjunto de datos anterior, los nombre de las variables son descriptivos lo que nos permite formular hipótesis antes de visualizar los datos.

Hipótesis previas

Las hipótesis previas nos permite son una primera aproximación del modelo de datos, nos permite crear los primeros modelos a partir de los cuales iterar para obtener un mejor resultado.

1. La temperatura media es un modelo lineal en el que intervienen la temperatura mínima y la temperatura máxima. Posiblemente $0.5 \times \text{temperatura mínima} + 0.5 \times \text{temperatura máxima}$, por la propia definición de media.
2. Por la ley física que relaciona la temperatura y la presión, a mayor presión mayor temperatura. No se espera que esta ley se cumpla por completo ya que está estipulada para gases de volumen constante, por esto tanto la presión como la presión a nivel del mar tienen que ver en cierta medida con la temperatura.
3. La velocidad del viento y la velocidad máxima del mismo, no intervienen o lo hacen en una medida despreciable, puesto que son factores que intervienen más en la sensación térmica que en la propia temperatura.
4. El rocío y las precipitaciones, tienen que ver más como consecuencia de la temperatura que como factor generador de la temperatura, no se descarta su intervención pero se espera que sea mínima.
5. Sobre la visibilidad, no sabemos que comportamiento tendrá puesto que en ocasiones hay poca visibilidad a causa de bancos de niebla que se generan por las altas temperaturas, pero en ciertas zonas del planeta puede ser por polvo en suspensión traído por el viento, que genera que haya mayor temperatuda, por ello como actuará esta variable en el modelo es todo un misterio.

Ahora con estas hipótesis previas, procedemos a estudiar las variables, que en este caso son solo numéricas, con respecto de la salida.

Variables numéricas.

En primer lugar vamos a cargar el dataset.

```
wizmir <- read.csv("./WizmirRegression/wizmir/wizmir.dat", header = FALSE,
  comment.char = "@")
names(wizmir) <- c("Max_temperature", "Min_temperature", "Dewpoint",
  "Precipitation", "Sea_level_pressure", "Standar _pressure",
  "Visibility", "Wind_speed", "Wind_max_speed", "Mean_temperature")
```

Tras ello vamos a obtener el número de valores perdidos que tenemos en el conjunto de datos.

Estudio de los valores perdidos

```
wizmir_na <- apply(is.na(wizmir), 2, sum)  
wizmir_na
```

	Max_temperature	Min_temperature	Dewpoint
	0	0	0
Precipitation	Sea_level_pressure	Standar _pressure	
	0	0	0
Visibility	Wind_speed	Wind_max_speed	
	0	0	0
Mean_temperature			
	0		

Esta base de datos no tiene valores perdidos, lo que nos facilita el trabajo.

Test de normalidad

En este caso, el problema no es un problema de clasificación por lo que los algoritmos no hacen suposiciones de los datos de entrada como el caso anterior que necesitaba que los datos estuvieran normalmente distribuidos, por ello no vamos a realizar un estudio de la normalidad de los datos.

Estudio de los principales estadísticos

La obtención de los principales valores estadísticos la obtenemos, como hemos visto en los casos anteriores, mediante la función summary.

```
summary(wizmir)
```

	Max_temperature	Min_temperature	Dewpoint
Min.	: 36.70	Min. :15.80	Min. :13.60
1st Qu.	: 59.00	1st Qu.:40.10	1st Qu.:41.30
Median	: 70.70	Median :50.00	Median :48.20
Mean	: 72.22	Mean :50.74	Mean :46.62
3rd Qu.	: 87.10	3rd Qu.:62.20	3rd Qu.:53.60
Max.	:105.00	Max. :78.60	Max. :64.40
	Precipitation	Sea_level_pressure	Standar _pressure
Min.	:0.00000	Min. :29.26	Min. : 2.300
1st Qu.	:0.00000	1st Qu.:29.85	1st Qu.: 7.100
Median	:0.00000	Median :29.95	Median : 7.300
Mean	:0.09257	Mean :29.97	Mean : 7.197
3rd Qu.	:0.00000	3rd Qu.:30.08	3rd Qu.: 7.600
Max.	:7.60000	Max. :30.48	Max. :10.100
	Visibility	Wind_speed	Wind_max_speed
Min.	: 0.92	Min. : 4.72	Min. :16.11
1st Qu.	: 6.56	1st Qu.:16.10	1st Qu.:34.28
Median	:10.50	Median :19.81	Median :34.28
Mean	:11.16	Mean :19.81	Mean :34.28
3rd Qu.	:15.40	3rd Qu.:23.00	3rd Qu.:34.28
Max.	:29.10	Max. :68.80	Max. :55.24
	Mean_temperature		
Min.	:29.40		
1st Qu.	:49.60		
Median	:60.00		
Mean	:61.51		
3rd Qu.	:75.20		

Max. :89.90

Exceptuando las variables de precipitación (Precipitation) y la presión normal (Standar_pressure) podemos decir que todas las variables se mueven en el mismo rango de 20-100 aproximadamente por lo que si se generase un modelo que no tuviese dichas variables podríamos evitar el paso intermedio de normalizar las variables para igualar el rango.

Nos queda por conocer la desviación estándar de los datos:

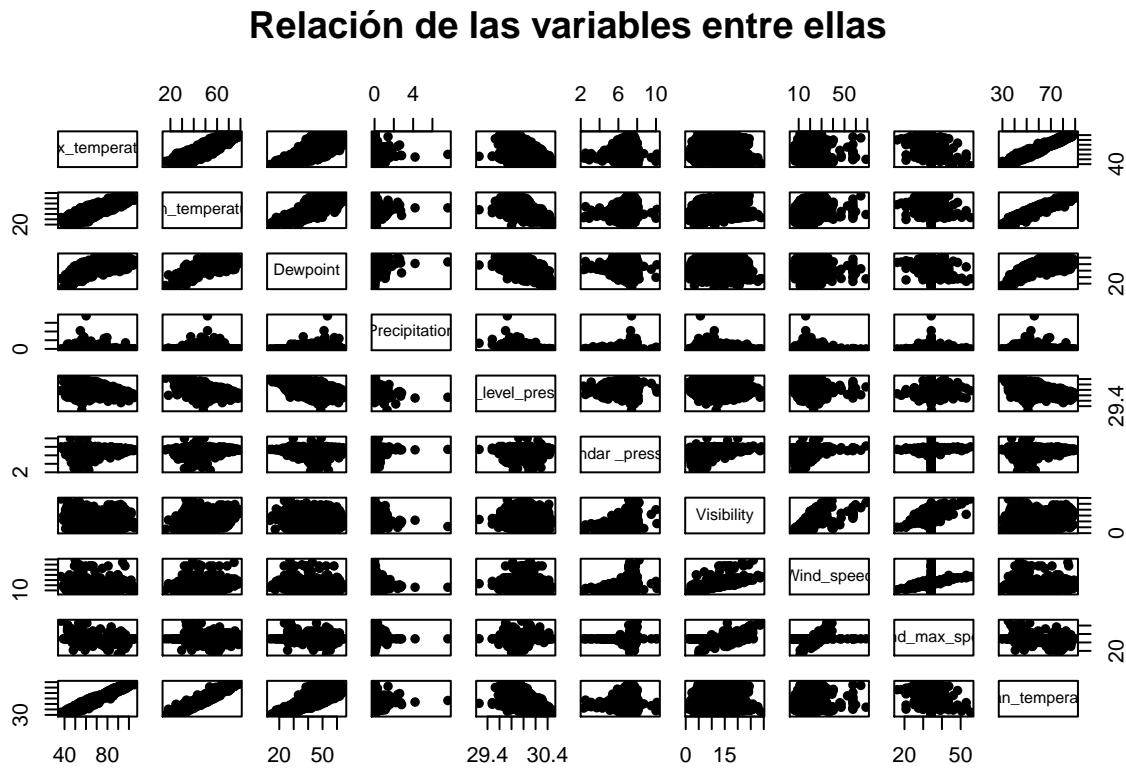
```
wizmir_std <- apply(wizmir[, ], 2, sd)
wizmir_std
```

Max_temperature	Min_temperature	Dewpoint
15.9267131	13.2260798	9.3445446
Precipitation	Sea_level_pressure	Standar _pressure
0.3528008	0.1676890	0.6849532
Visibility	Wind_speed	Wind_max_speed
5.4066647	7.1352505	2.4418256
Mean_temperature		
14.3762319		

Viendo las desviaciones típicas de la temperatura mínima y máxima tener un valor tan cercano al de la temperatura media, que es el valor que tenemos que obtener, hace pensar que podrían compartir distribución.

Si representamos el valor de todas las variables con respecto de la salida obtenemos los siguientes gráficos

```
plot(wizmir, main = "Relación de las variables entre ellas",
     pch = 16)
```



Primero, nos vamos a concentrar en la última fila de la ilustración anterior, en ella están reflejadas todas las variables como variable de “entrada” o variable independiente y la variable de salida como variable independiente. En esta fila podemos ver que las variables de temperatura máxima y mínima tienen una relación lineal con la temperatura media, como era de esperar por la hipótesis 1, pero además esta tendencia lineal también la tiene la variable de rocío que puede ser debido a que el rocío es dependiente de la temperatura mínima, como se ve en la relación entre estas dos variables (y en la relación de la variable rocío con la temperatura máxima también) pero es un detalle a tener en cuenta a la hora de elaborar un modelo.

Otras variables que podrían intervenir en el modelo de una forma, no tan claramente lineal o incluso de orden superior, son la presión a nivel del mar y la velocidad máxima del viento, puesto que dentro de la nube de puntos se puede dibujar una recta decreciente e incluso una curva.

El resto de variables la nube de puntos tiene tal dispersión que podrían no intervenir porque no se ve una función definida que encaje con los datos.

Estudio de correlación

Una forma de comprobar que la variable *Dewpoint* o punto de rocío está relacionada con la temperatura mínima y/o máxima está en la tabla de correlaciones:

```
cor(wizmir[, -10], method = "pearson")
```

	Max_temperature	Min_temperature
Max_temperature	1.0000000	0.89611403
Min_temperature	0.89611403	1.00000000
Dewpoint	0.74937745	0.78634048
Precipitation	-0.19713825	-0.07922557

Sea_level_pressure	-0.51077796	-0.62301824
Standar _pressure	0.09967300	0.17983060
Visibility	0.11244011	0.35256802
Wind_speed	0.08825536	0.25255985
Wind_max_speed	-0.14522657	-0.08434084
	Dewpoint	Precipitation
Max_temperature	0.74937745	-0.19713825
Min_temperature	0.78634048	-0.07922557
Dewpoint	1.00000000	0.02005389
Precipitation	0.02005389	1.00000000
Sea_level_pressure	-0.58739019	-0.15185635
Standar _pressure	0.04351714	0.06492428
Visibility	-0.05894845	-0.04562776
Wind_speed	-0.05782537	-0.01743911
Wind_max_speed	-0.12906064	0.02320726
	Sea_level_pressure	Standar _pressure
Max_temperature	-0.51077796	0.09967300
Min_temperature	-0.62301824	0.17983060
Dewpoint	-0.58739019	0.04351714
Precipitation	-0.15185635	0.06492428
Sea_level_pressure	1.00000000	-0.18339796
Standar _pressure	-0.18339796	1.00000000
Visibility	-0.20288466	0.30903837
Wind_speed	-0.19647381	0.26081990
Wind_max_speed	0.04049549	0.03857268
	Visibility	Wind_speed Wind_max_speed
Max_temperature	0.11244011	0.08825536 -0.14522657
Min_temperature	0.35256802	0.25255985 -0.08434084
Dewpoint	-0.05894845	-0.05782537 -0.12906064
Precipitation	-0.04562776	-0.01743911 0.02320726
Sea_level_pressure	-0.20288466	-0.19647381 0.04049549
Standar _pressure	0.30903837	0.26081990 0.03857268
Visibility	1.00000000	0.76428686 0.23699238
Wind_speed	0.76428686	1.00000000 0.21889408
Wind_max_speed	0.23699238	0.21889408 1.00000000

Las correlaciones que podemos ver del cuadrante anterior son:

- * La temperatura mínima y máxima están correlacionadas entre ellas.
- * El punto de rocío está relacionado tanto con las temperaturas máxima y mínima y en menor medida, y de forma inversa, con la presión a nivel del mar.
- * Tanto la temperatura máxima como mínima están relacionadas de forma inversa con la presión atmosférica a nivel del mar.
- * La visibilidad y la presión estándar tienen una ligera correlación.
- * La visibilidad y la velocidad del viento están fuertemente correlacionadas.

Como conclusión de estos datos diría que el mejor modelo para predecir la temperatura media es un modelo de regresión lineal donde intervengan las variables de temperatura máxima y mínima y habría que estudiar si añadir las variables de rocío y presión a nivel del mar supone alguna mejora.

Regresión

La regresión es el proceso estadístico por el que se estiman la relación entre una o varias variables independientes o predictoras y la variable dependiente, dado un conjunto de datos de entrada. El objetivo de la regresión es obtener un modelo que permita predecir o estimar el valor que tendrá un nuevo dato a su salida, siendo este dato distinto de todos los anteriores que se usaron para crear el modelo. Existen varios algoritmos para realizar regresión, pero en este trabajo solamente utilizaremos la regresión lineal simple, la regresión lineal múltiple y el algoritmo de los k- vecinos más cercanos.

Problema

Utilizaremos regresión para predecir la temperatura media de la ciudad de Izmir que hemos estudiado anteriormente.

Regresores elegidos

En el apartado de análisis de datos anterior pudimos ver que los mejores regresores para este problema son las variables:

- Temperatura máxima
- Temperatura mínima

Pero vamos a realizar una prueba de regresión lineal simple con cinco regresores así que añadiremos a la lista:

- Punto de rocío
- Presión a nivel del mar
- Visibilidad

Modelo de regresión simple

Con el modelo de regresión simple comprobamos que los datos se pueden explicar de la forma de una línea recta usando solamente una variable de predicción. La bondad de este modelo se puede medir por dos coeficientes: R^2 y el RMSE. El primer parámetro nos lo realizar la operación *summary* sobre el modelo obtenido, pero el RMSE es preferible calcularlo a mano por ello definimos la siguiente función *rmse*.

```
rmse <- function(original, algorithm_output) {
  sqrt(sum((original - algorithm_output)^2)/length(algorithm_output))
}
```

Los modelos los crearemos haciendo uso de todos los datos del conjunto de datos

```
lmModel_maxTemp = lm(wizmir$Mean_temperature ~ wizmir$Max_temperature)
lmModel_minTemp = lm(wizmir$Mean_temperature ~ wizmir$Min_temperature)
lmModel_dewPoint = lm(wizmir$Mean_temperature ~ wizmir$Dewpoint)
lmModel_seaPressure = lm(wizmir$Mean_temperature ~ wizmir$Sea_level_pressure)
lmModel_visibility = lm(wizmir$Mean_temperature ~ wizmir$Visibility)
```

Ahora obtendremos los dos parámetros de comparación:

```
summary_maxTemp = summary(lmModel_maxTemp)
rSquared_maxTemp = summary_maxTemp$r.squared
adjusted_r_maxTemp = summary_maxTemp$adj.r.squared
rmse_maxTemp = rmse(wizmir$Mean_temperature, lmModel_maxTemp$fitted.values)
```

```

summary_minTemp = summary(lmModel_minTemp)
rSquared_minTemp = summary_minTemp$r.squared
adjusted_r_minTemp = summary_minTemp$adj.r.squared
rmse_minTemp = rmse(wizmir$Mean_temperature, lmModel_minTemp$fitted.values)

summary_dewPoint = summary(lmModel_dewPoint)
rSquared_dewPoint = summary_dewPoint$r.squared
adjusted_r_dewPoint = summary_dewPoint$adj.r.squared
rmse_dewPoint = rmse(wizmir$Mean_temperature, lmModel_dewPoint$fitted.values)

summary_seaPress = summary(lmModel_seaPressure)
rSquared_seaPress = summary_seaPress$r.squared
adjusted_r_seaPress = summary_seaPress$adj.r.squared
rmse_seaPress = rmse(wizmir$Mean_temperature, lmModel_seaPressure$fitted.values)

summary_visibility = summary(lmModel_visibility)
rSquared_visibility = summary_visibility$r.squared
adjusted_r_visibility = summary_visibility$adj.r.squared
rmse_visibility = rmse(wizmir$Mean_temperature, lmModel_visibility$fitted.values)

rSquared = c(rSquared_maxTemp, rSquared_minTemp, rSquared_dewPoint,
            rSquared_seaPress, rSquared_visibility)
adjRSquared = c(adjusted_r_maxTemp, adjusted_r_minTemp, adjusted_r_dewPoint,
               adjusted_r_seaPress, adjusted_r_visibility)
rmse_lm_results = c(rmse_maxTemp, rmse_minTemp, rmse_dewPoint,
                     rmse_seaPress, rmse_visibility)

linealSimpleModels_comparisonDataFrame = data.frame(rSquared,
                                                    adjRSquared, rmse_lm_results)
row.names(linealSimpleModels_comparisonDataFrame) <- c("MaxTemp",
                                                       "MinTemp", "DewPoint", "SeaPressure", "Visibility")
linealSimpleModels_comparisonDataFrame

```

	rSquared	adjRSquared	rmse_lm_results
MaxTemp	0.95764878	0.95761975	2.957531
MinTemp	0.91902245	0.91896695	4.089580
DewPoint	0.61549043	0.61522688	8.911483
SeaPressure	0.33981129	0.33935879	11.676977
Visibility	0.05107395	0.05042356	13.999502

De esta tabla si miramos la columna de la R^2 (*rSquared*) podemos ver que la temperatura máxima es el mejor regresor del conjunto puesto con un modelo lineal simple explica el 95% de los casos, seguido de cerca por la temperatura mínima que explica el 91% de los casos. El resto de regresores no son tan buenos.

La columna de la R^2 ajustada en este caso no tiene mucho sentido, puesto que es el parámetro de la R^2 ajustada al número de variables del modelo que en este caso es 1, pero se calcula para después poder comparar con el modelo de regresión múltiple.

El RMSE crece a medida que el modelo es peor en la predicción de la salida, siendo prácticamente 3 para el caso de la temperatura máxima y de prácticamente 14 para el caso de la visibilidad que solamente explica el 5% de los casos.

Modelo de regresión lineal múltiple

%Comparar con los modelos anteriores ###Modelo Knn con Cross-validation de 5-particiones ###Comparación KNN con el modelo de regresión lineal múltiple #####Con el modelo obtenido anteriormente #####Con un modelo en el que intervengan todas las variables y que utilice las 5 particiones ###Comparación de algoritmos #####Friedman #####Holms