

ICD_ProyectoFinal

Laura del Pino Díaz

15/12/2016

Contents

| | |
|--|----------|
| Introducción | 2 |
| Las bases de datos | 2 |
| Australian (Australian Credit Approval) | 2 |
| Estudio de los principales estadísticos del conjunto de datos Australian | 2 |
| Variables numéricas | 2 |
| Variables categóricas. | 12 |
| Wizmir (Weather of Izmir) | 17 |
| Hipótesis previas | 17 |
| Variables numéricas. | 17 |

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 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)
numerical_stats
```

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

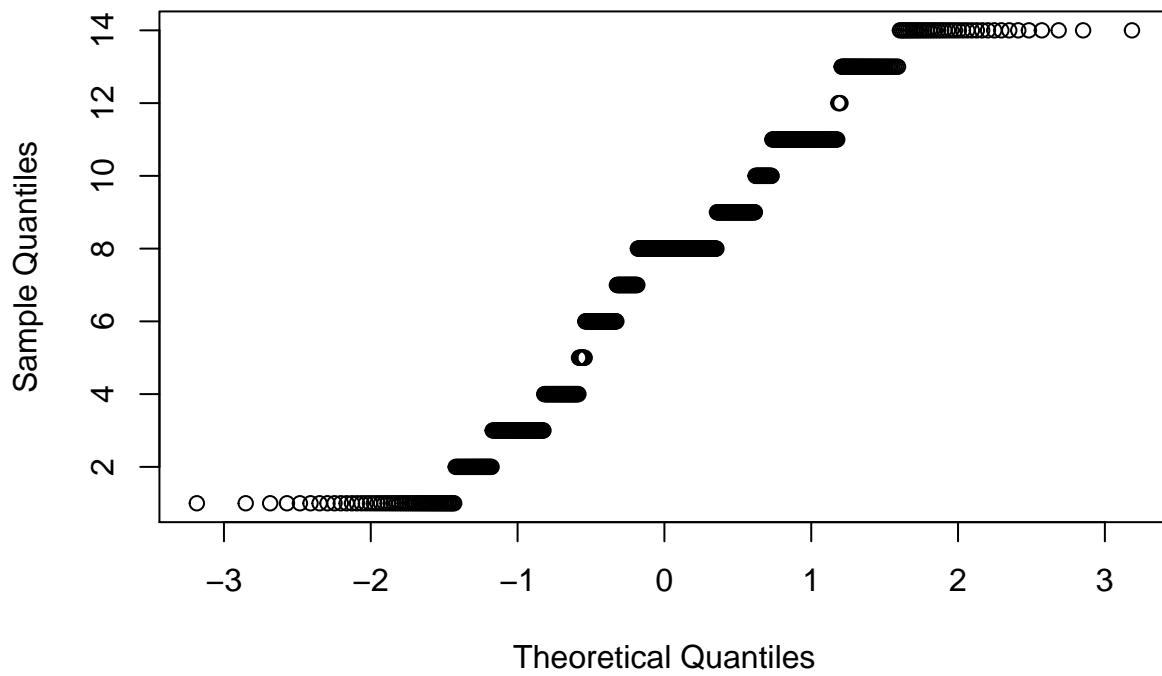
| A2 | A3 | A5 | A6 | A7 |
|-------------|-------------|-------------|----------|-------------|
| 1554.559732 | 3069.110042 | 3.683265 | 1.992316 | 1387.900324 |
| A10 | A13 | A14 | | |
| 4.862940 | 172.159274 | 5210.102598 | | |

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. La comparación de varianzas nos ayuda a la hora de comprobar que variables tienen una varianza similar, hecho que resulta interesante para algoritmos como LDA donde se tiene como suposición que la varianza es la misma o muy similar. Por esto ante un modelo para LDA deberíamos considerar las variables A5 y A10 que tienen un valor muy similar, o las variables A2 y A7.

La otra hipótesis que necesita el algoritmo LDA es que las variables siguen una distribución normal. Para mostrar esta característica con las variables A5, A10, A2 y A7 realizaremos un diagrama Q-Q y consideraremos que siguen una distribución normal si los gráficos son similares a una línea recta.

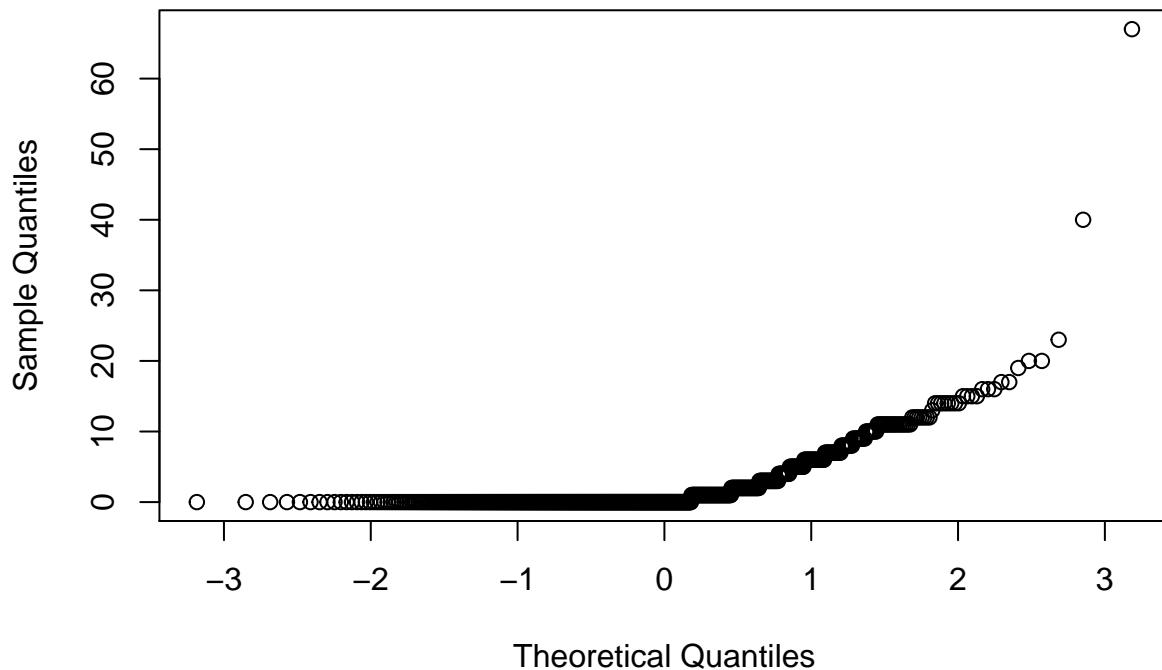
```
qqnorm(australian[, 5], main = "QQ para la variable A5")
```

QQ para la variable A5



```
qqnorm(australian[, 10], main = "QQ para la variable A10")
```

QQ para la variable A10

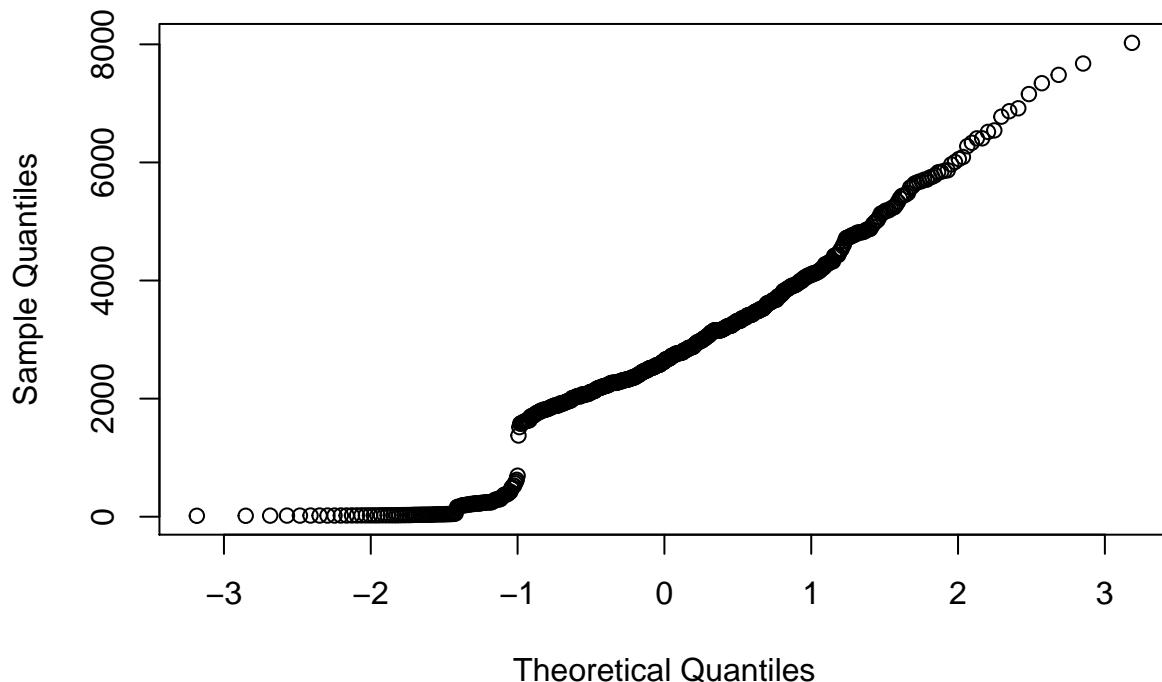


variable A5 si parece seguir una distribución normal, pero la variable con la que comparte la varianza A10 no la sigue.

La

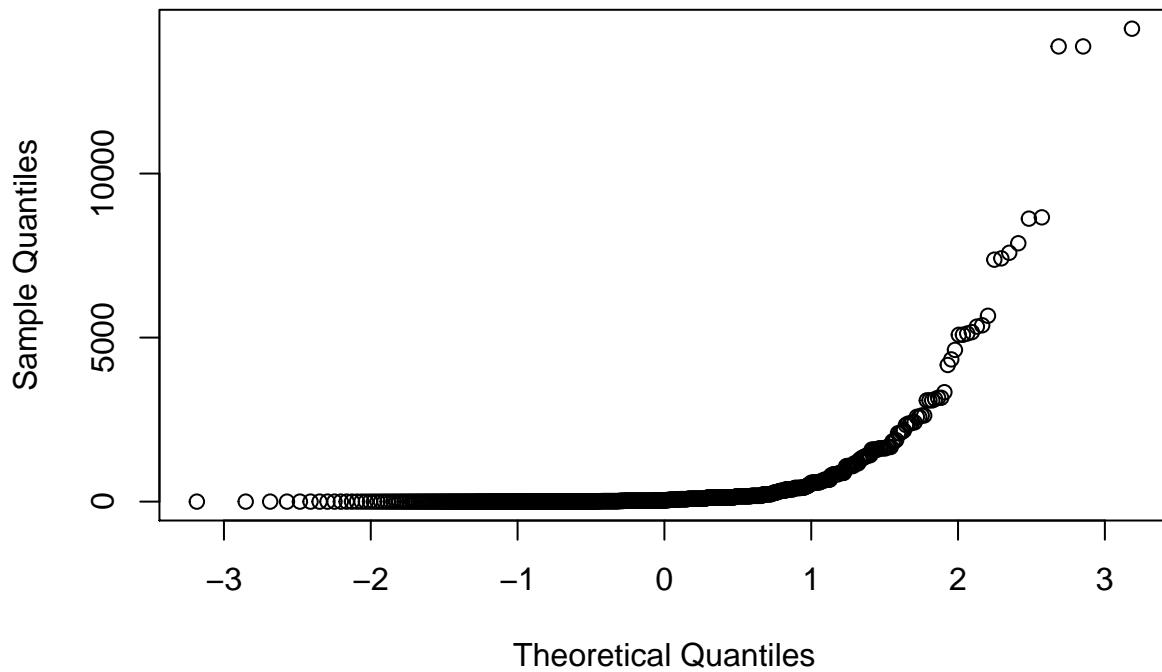
```
qqnorm(australian[, 2], main = "QQ para la variable A2")
```

QQ para la variable A2



```
qqnorm(australian[, 7], main = "QQ para la variable A5")
```

QQ para la variable A5

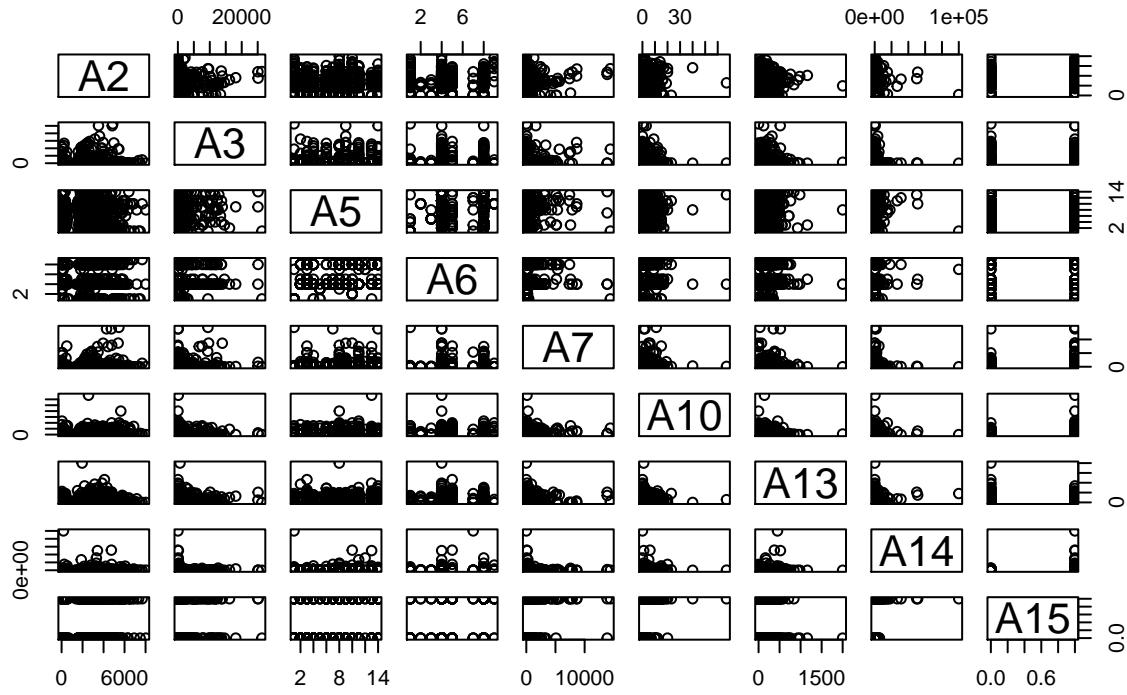


La variable A2, puede seguir una distribución normal pero el punto de ruptura que tiene me desconcierta.

Sin embargo queda claro que la variable A7 no sigue una distribución normal.

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

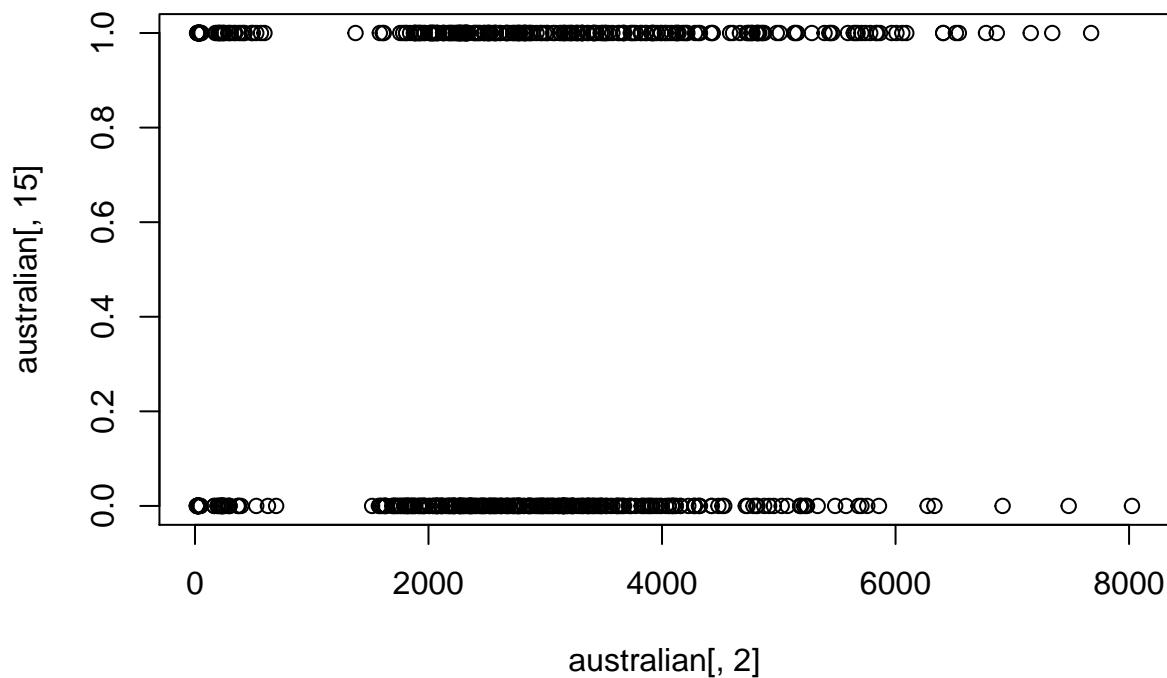
Comparación de las variables numéricas con la salida



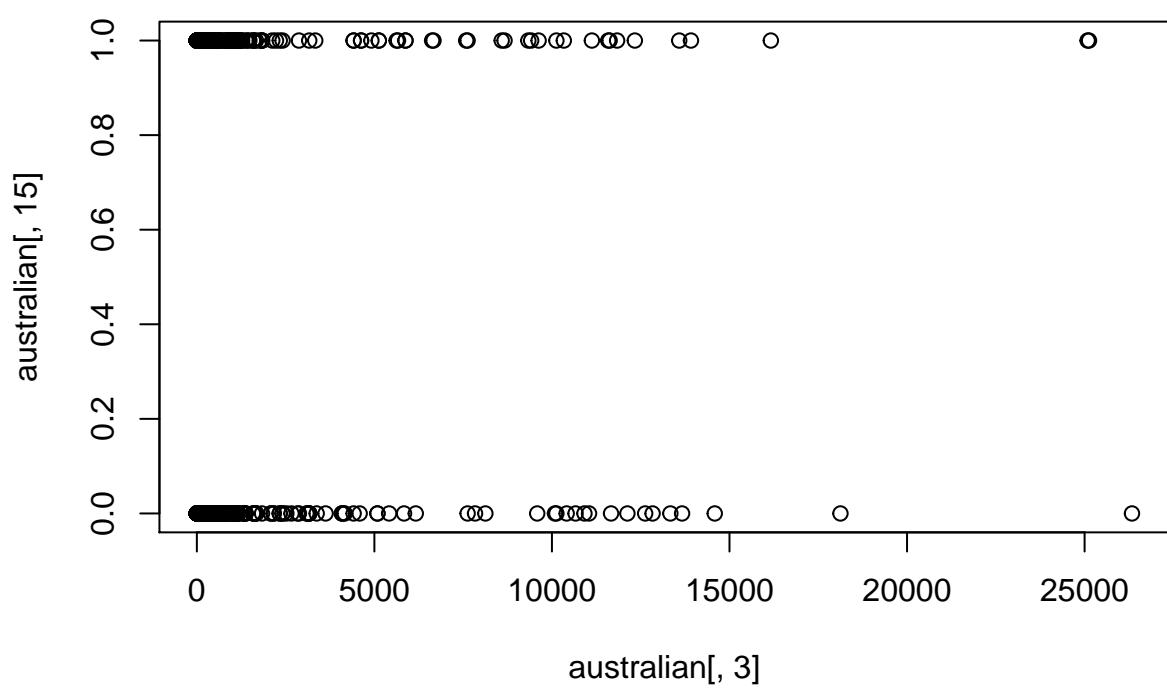
Puesto que es una base de datos para clasificación con dos clases tiene sentido que en todas ellas aparezcan las dos columnas. Pero verlas así en pequeño no nos permite deducir si esa variable aporta mucho o poco a la salida, por lo que vamos a realizar unos plots para analizar mejor los datos.

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

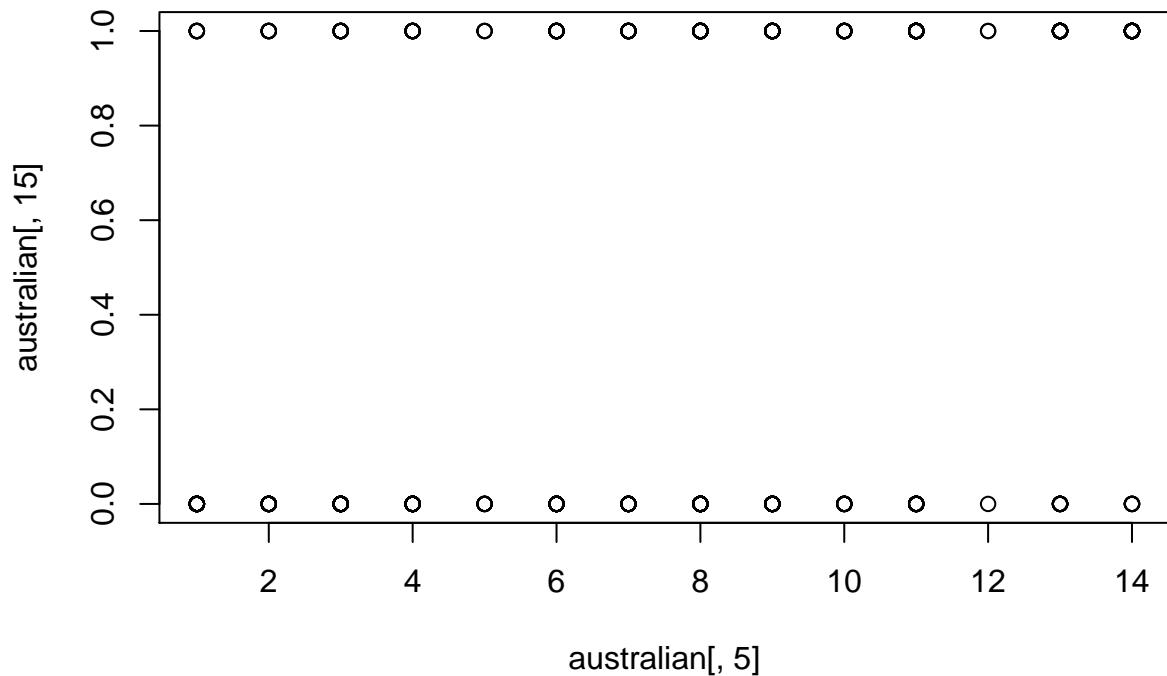
Comparación A2 con la salida



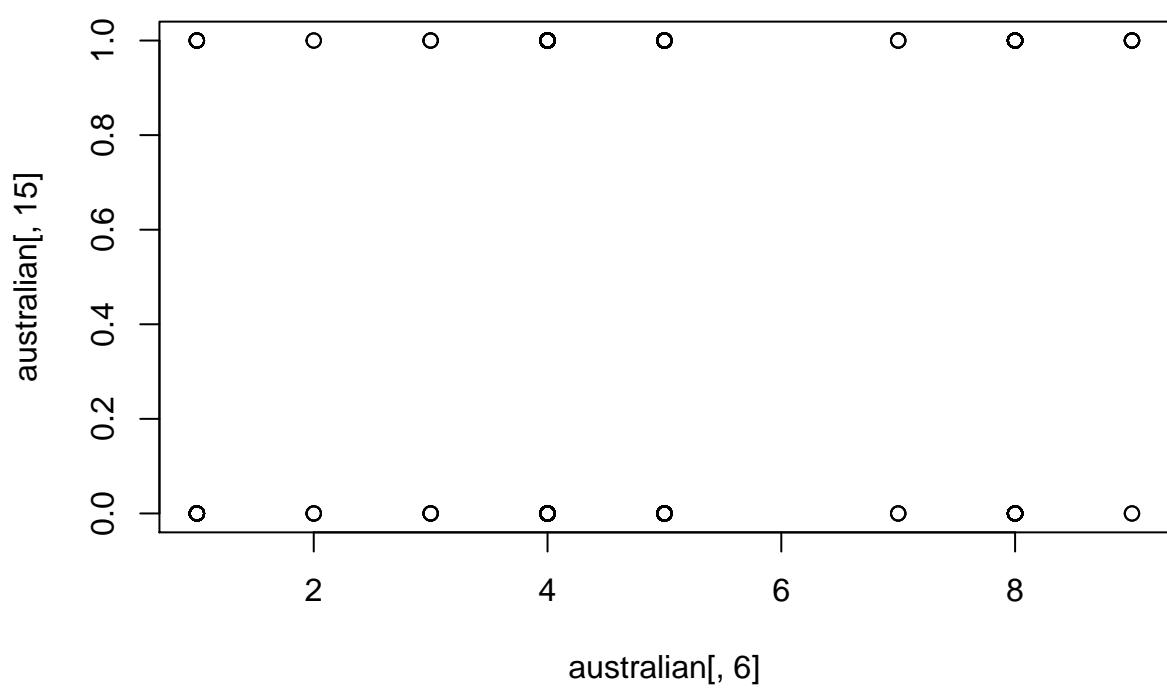
Comparación A3 con la salida



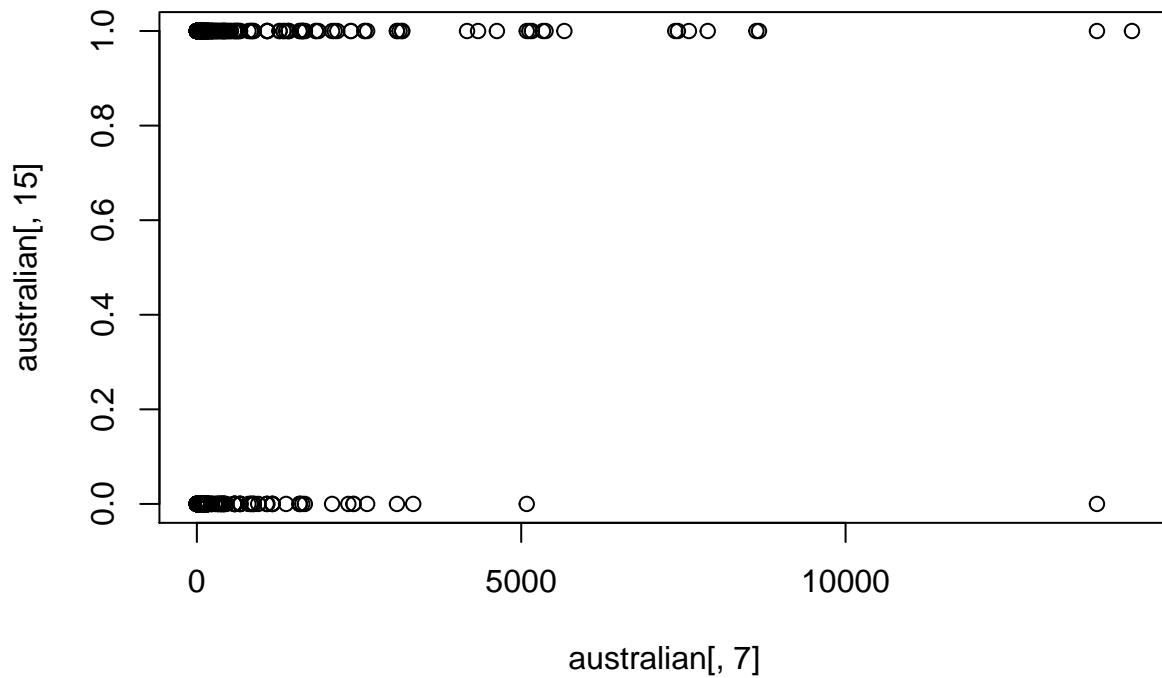
Comparación A5 con la salida



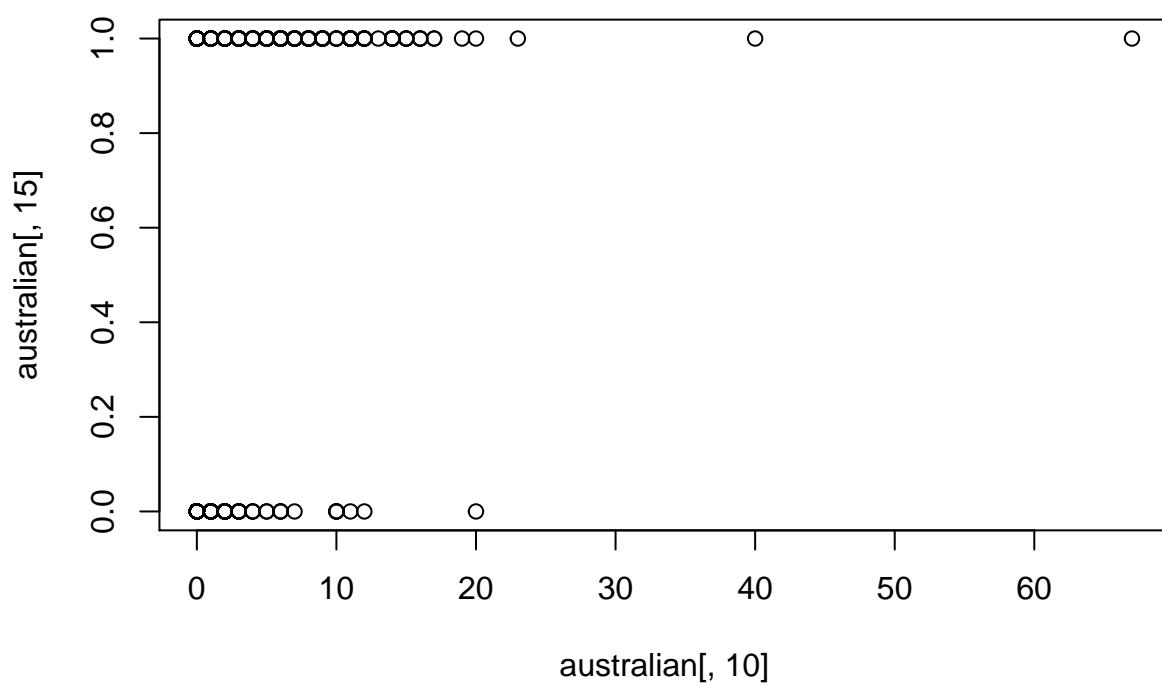
Comparación A6 con la salida



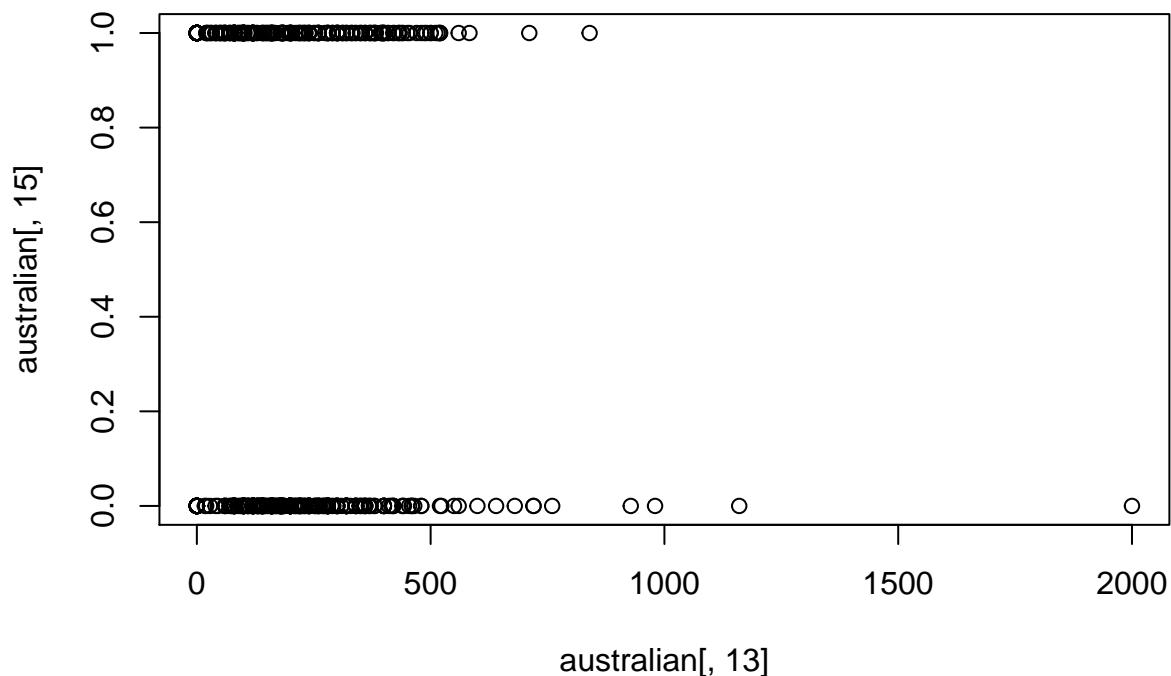
Comparación A7 con la salida



Comparación A10 con la salida

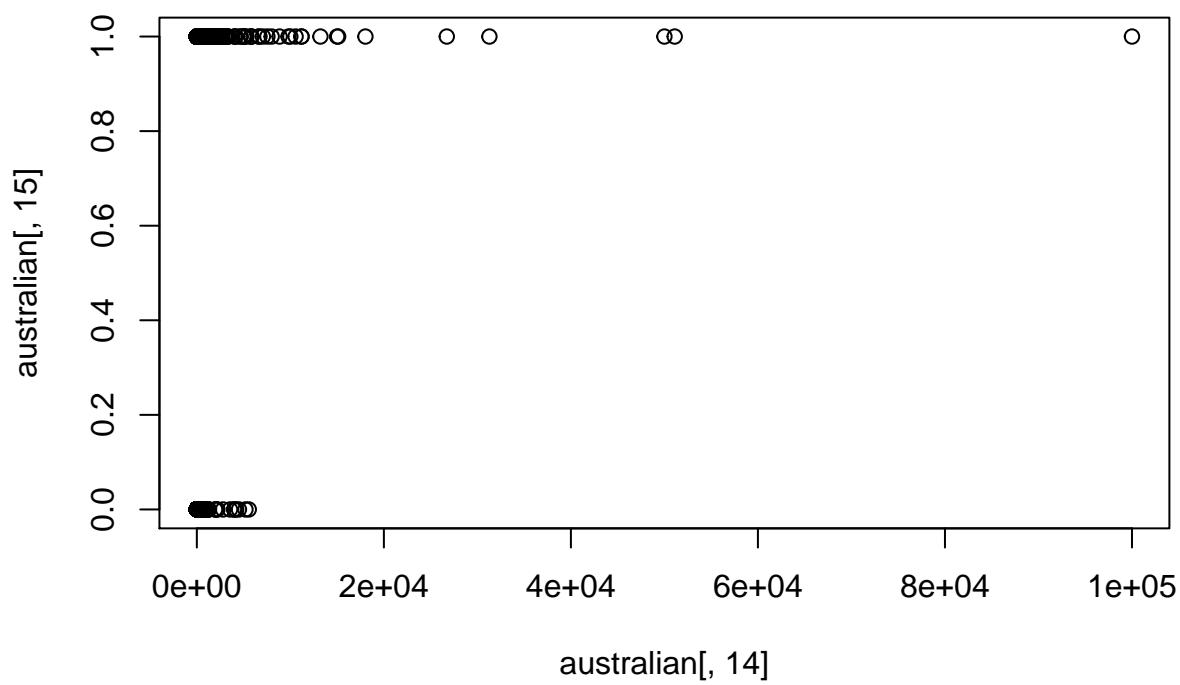


Comparación A13 con la salida



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

Comparación A15 con la salida

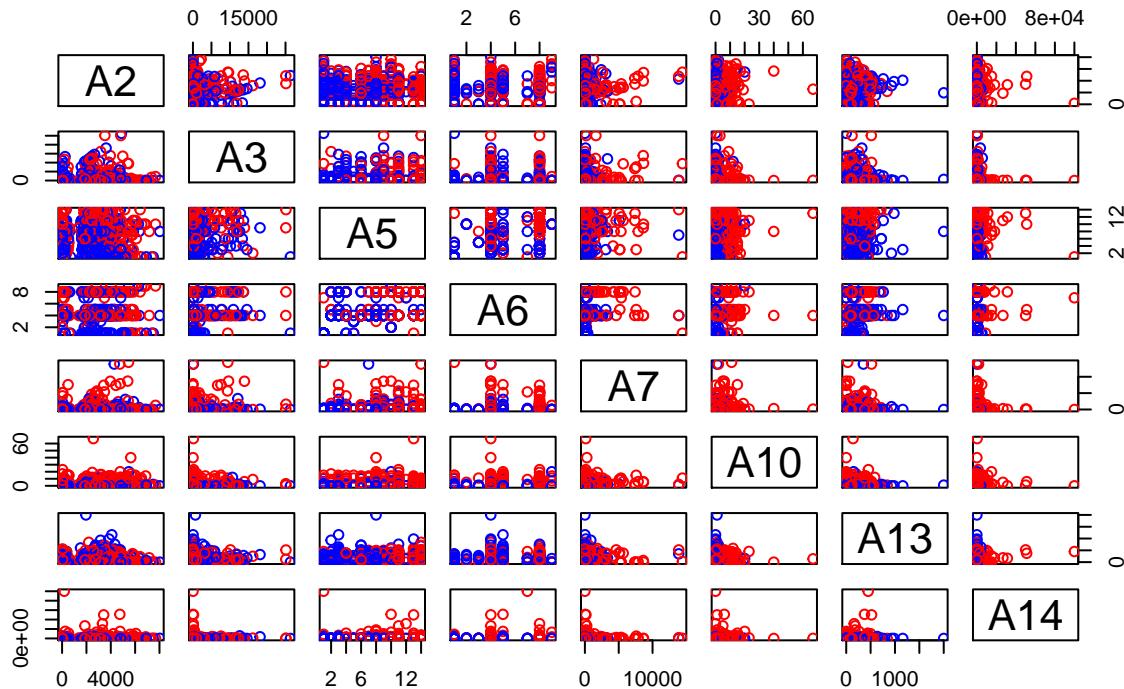


No podemos decir que ninguna de las variables numéricas sea realmente discriminante para la salida. Sin embargo, no descartaría la variable 14 para que intervenga un modelo de clasificación, porque para valores altos solamente da salida positiva para la clase 1.

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 entre ellas",  
      col = ifelse(australian[, 15] == 1, "red", "blue"))
```

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.

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.

```
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
```

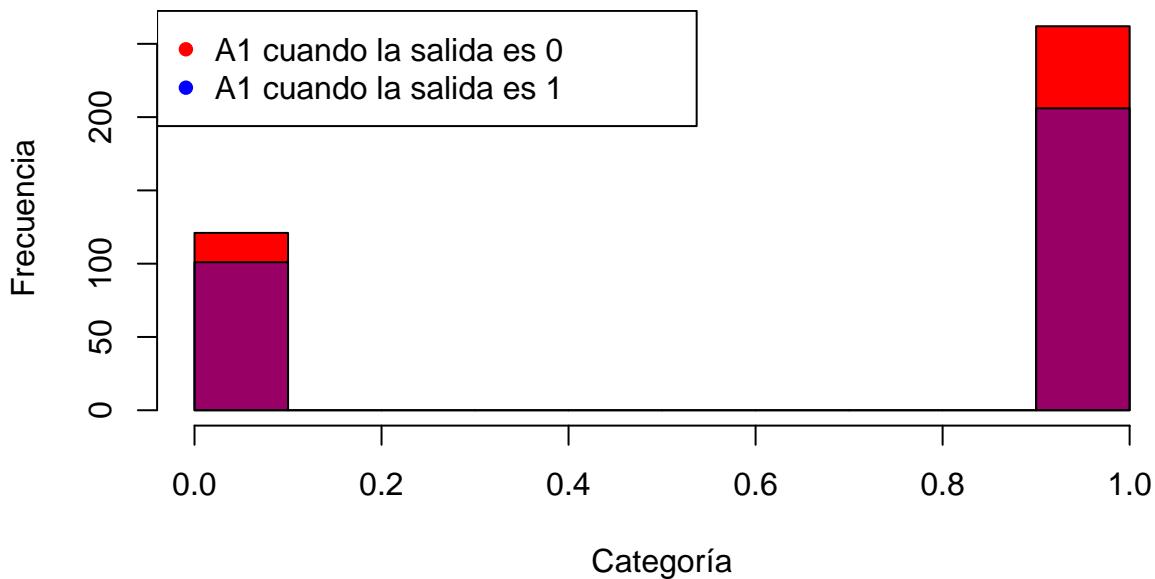
| A1 | A4 | A8 |
|-----------------|------------------|---------------------|
| "Min. :0.0000 | " "Min. :1.000 | " "Min. :0.0000 " |
| "1st Qu.:0.0000 | " "1st Qu.:2.000 | " "1st Qu.:0.0000 " |
| "Median :1.0000 | " "Median :2.000 | " "Median :1.0000 " |
| "3rd Qu.:1.0000 | " "3rd Qu.:2.000 | " "3rd 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.

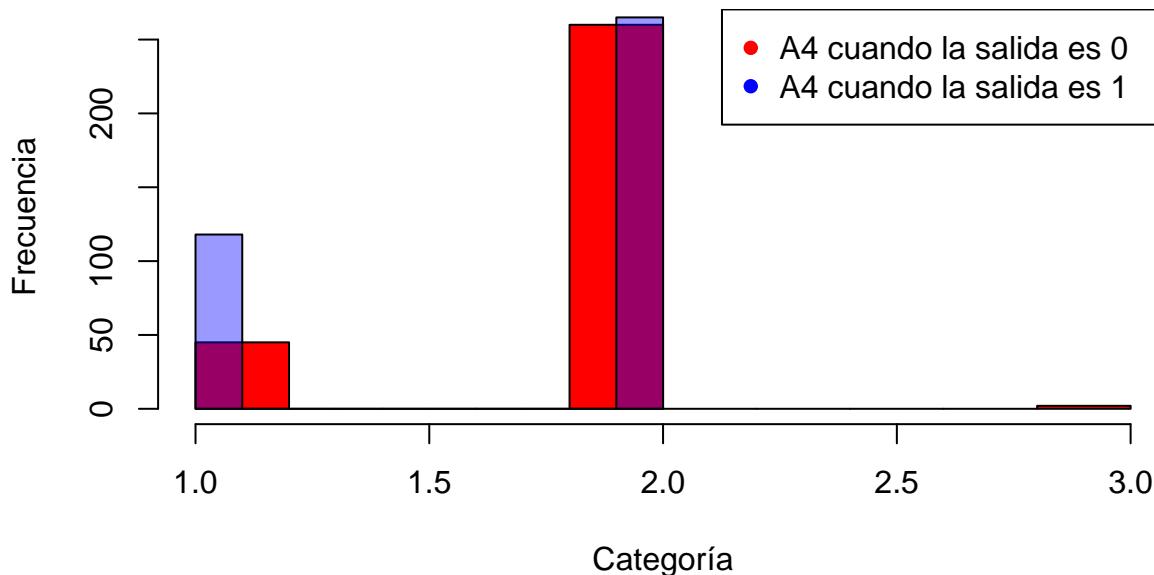
```
hist(australian[which(australian[, 15] == 0), 1], col = rgb(1,
0, 0, 1), main = "Frecuencia de la variable A1", ylab = "Frecuencia",
xlab = "Categoría")
hist(australian[which(australian[, 15] == 1), 1], col = rgb(0,
0, 1, 0.4), add = TRUE)
legend("topleft", legend = c("A1 cuando la salida es 0",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

Frecuencia de la variable A1



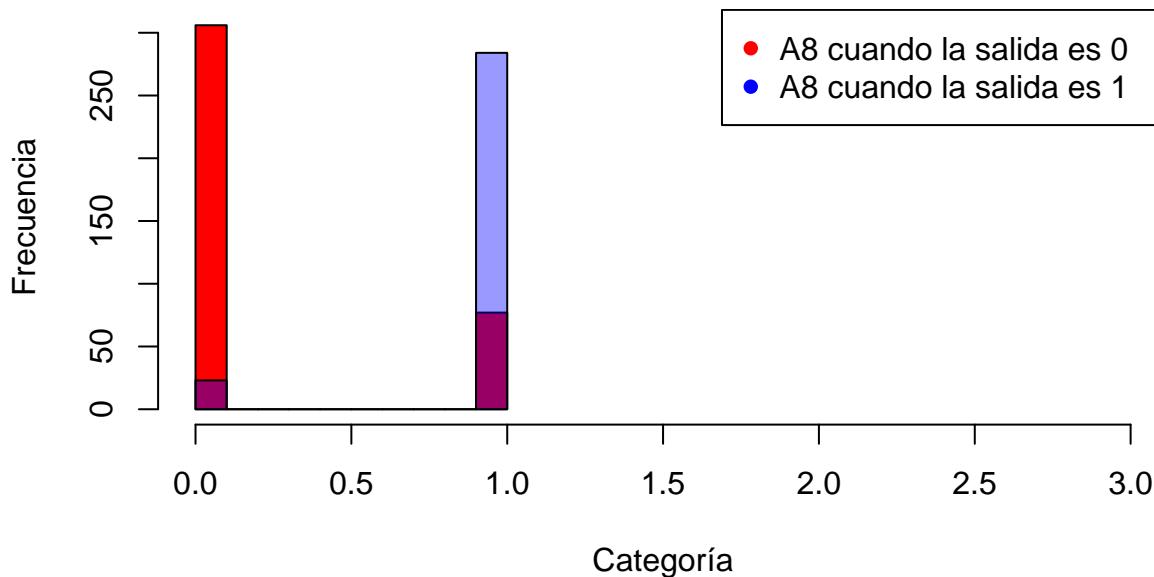
```
hist(australian[which(australian[, 15] == 1), 4], col = rgb(1,
0, 0, 1), main = "Frecuencia de la variable A4", ylab = "Frecuencia",
xlab = "Categoría")
hist(australian[which(australian[, 15] == 0), 4], col = rgb(0,
0, 1, 0.4), add = T)
legend("topright", legend = c("A4 cuando la salida es 0", "A4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



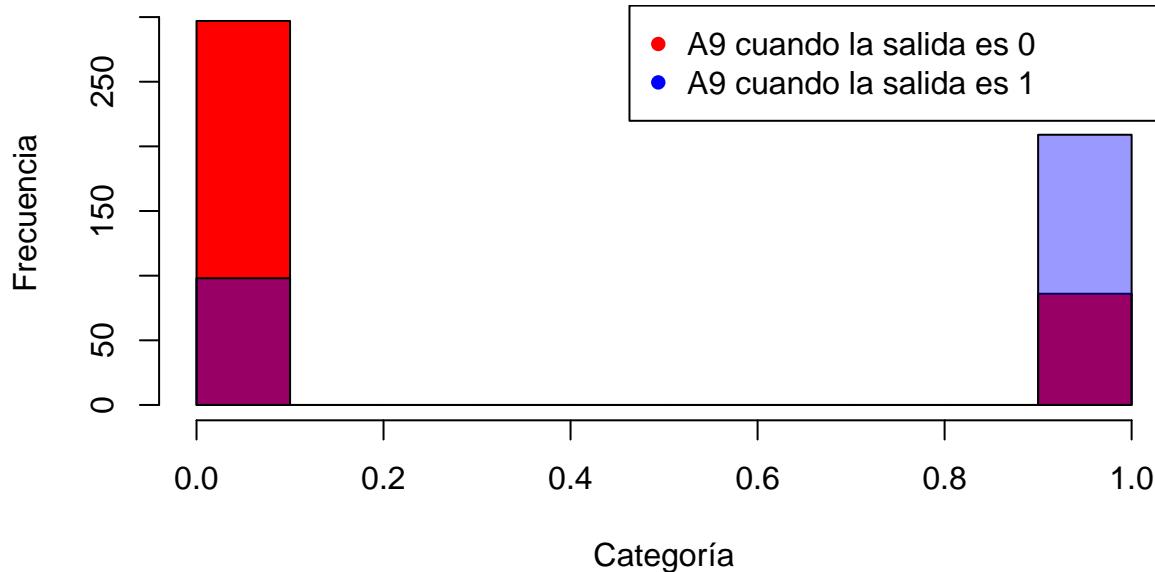
```
hist(australian[which(australian[, 15] == 0), 9], col = rgb(1, 0, 0, 1), main = "Frecuencia de la variable A9", ylab = "Frecuencia", xlab = "Categoría")
```

```

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

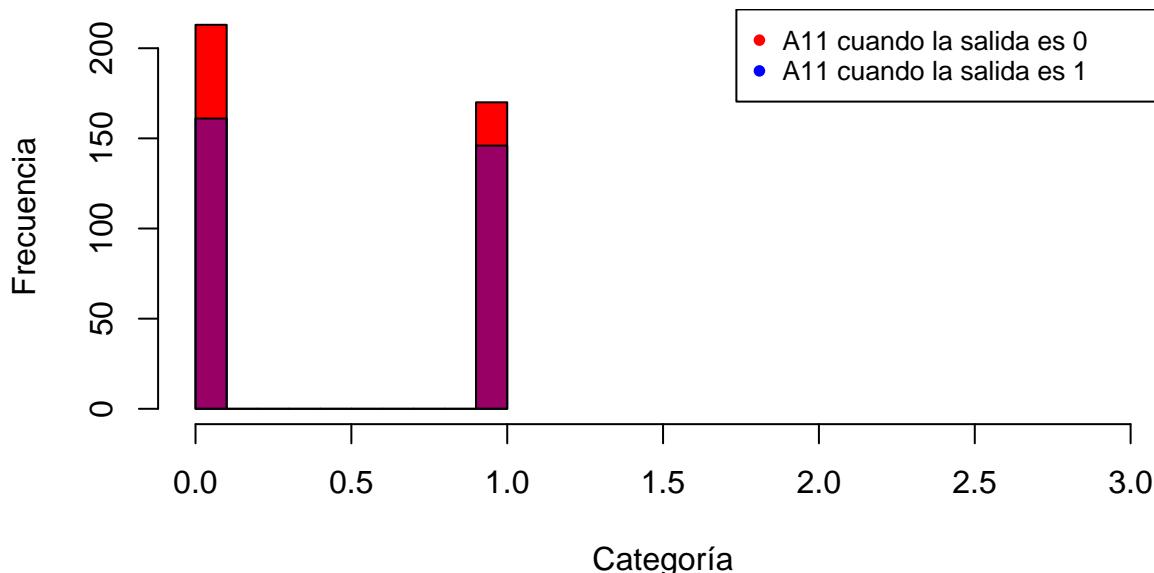


```

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)

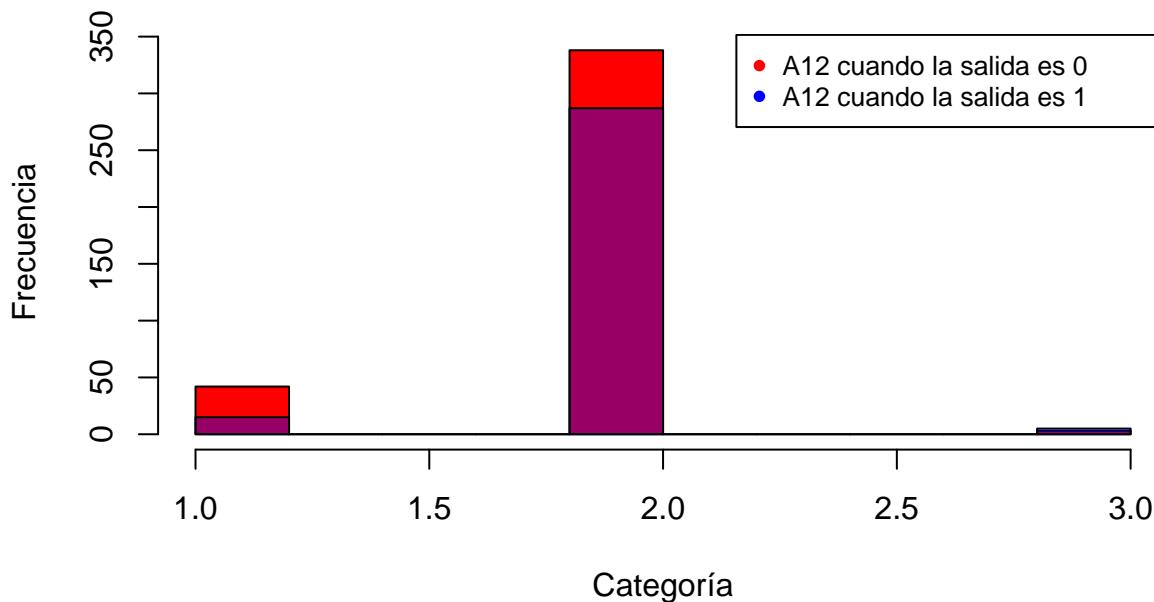
```

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 di 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 que el 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 atmosférico, 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 nombres de las variables son descriptivos lo que nos permite formular hipótesis antes de visualizar los datos.

Hipótesis previas

Las hipótesis previas nos permiten 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 qué 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 temperatura, 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 los principales estadísticos del conjunto:

```
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. | :0.000000 | Min. :29.26 | Min. : 2.300 |
| 1st Qu. | :0.000000 | 1st Qu.:29.85 | 1st Qu.: 7.100 |
| Median | :0.000000 | Median :29.95 | Median : 7.300 |
| Mean | :0.09257 | Mean :29.97 | Mean : 7.197 |
| 3rd Qu. | :0.000000 | 3rd Qu.:30.08 | 3rd Qu.: 7.600 |
| Max. | :7.600000 | 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
```

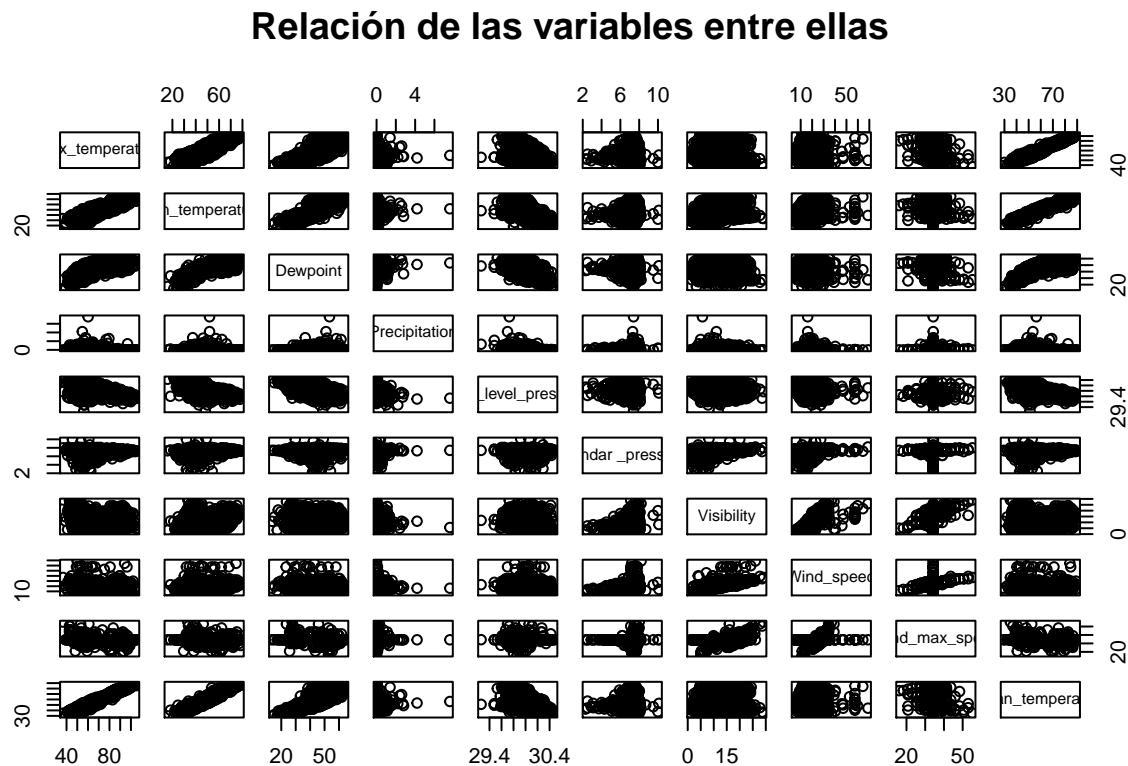
| | Max_temperature | Min_temperature | Dewpoint |
|--|------------------|--------------------|-------------------|
| | 15.9267131 | 13.2260798 | 9.3445446 |
| | Precipitation | Sea_level_pressure | Standar _pressure |
| | 0.3528008 | 0.1676890 | 0.6849532 |
| | Visibility | Wind_speed | Wind_max_speed |
| | 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, me 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")
```



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.

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 la variable de rocío supone alguna mejora.