EAD Q2

TGR DATA MINING

Grupo 3

Armenteros López, Ana - ana.armenteros@udc.es

Pérez Paredes, Alexandre - alexandre.perez1@udc.es

facultade de informática da coruña

UNIVERSIDADE DA CORUÑA

Contenido

Figuras

Introducción.

En este trabajo se realiza un análisis descriptivo sobre un conjunto de datos correspondientes a diferentes zonas de una ciudad, además de dos predicciones sobre dichos datos (una con regresión lineal y la otra con regresión logística). Para ello, se ha utilizado "R" (versión 4.1.3) y la librería "rattle". Para mayor comodidad, todos los comandos mencionados a lo largo del trabajo se pueden ver agrupados en Anexo A. Comandos.

Análisis exploratorio y preparación de los datos.

Deseamos hacer una **exploración inicial del conjunto de datos** proporcionado para entender mejor el problema. Así, veremos más claro cómo preparar dichos datos antes de realizar los análisis y predicciones. Se ha otorgado un archivo "datos.txt" con múltiples valores para cada una de sus variables (ver Anexo B. Datos y gráficas adicionales.)

Para poder explorar los datos con más facilidad, los hemos importado a un archivo Excel con la extensión .csv. De esta forma, podemos observar todas las filas con los distintos valores cómodamente. Podemos ver que hay 495 filas que corresponden a cada una de las zonas que queremos analizar (ver Anexo B. Datos y gráficas adicionales.)

El primer paso es **importar la librería "rattle"** en R para obtener las funcionalidades que nos interesan y **leer y guardar los datos** del Excel en una variable llamada "datos":

```
library(rattle)
datos<-read.csv("datos.csv")
```

Para hacernos una mejor idea de cómo son las variables que tenemos en "datos", ejecutaremos un comando para que nos muestre un **resumen** que incluye el valor mínimo, el valor máximo, la media de todos los valores, la mediana y el primer y tercer cuartil.

```
summary(datos)
```

Ilustración 1. Resultado inicial de summary(datos).

```
ï..criminalidad    terrenoresidencial   industrias          rio
Min.   : 0.00632   Min.   :  0.00     Min.   : 0.46   Min.   :0.00000
1st Qu.: 0.08232   1st Qu.:  0.00     1st Qu.: 5.13   1st Qu.:0.00000
Median : 0.26938   Median :  0.00     Median : 8.56   Median :0.00000
Mean   : 3.69038   Mean   : 11.62     Mean   :11.15   Mean   :0.07071
3rd Qu.: 3.73597   3rd Qu.: 15.00     3rd Qu.:18.10   3rd Qu.:0.00000
Max.   :88.97620   Max.   :100.00     Max.   :27.74   Max.   :1.00000
oxidonitrogeno     habitaciones    viviendasantiguas ditanciacentro
Min.   :0.3850   Min.   :3.561   Min.   :  2.90    Min.   : 1.130
1st Qu.:0.4480   1st Qu.:5.888   1st Qu.: 44.05    1st Qu.: 2.083
Median :0.5320   Median :6.211   Median : 77.70    Median : 3.280
Mean   :0.5541   Mean   :6.289   Mean   : 68.49    Mean   : 3.824
3rd Qu.:0.6275   3rd Qu.:6.627   3rd Qu.: 94.20    3rd Qu.: 5.223
Max.   :0.8710   Max.   :8.780   Max.   :100.00    Max.   :12.127
accesoautopista    impuestos      ratioalumnos   poblacionnegra
Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
1st Qu.: 4.000   1st Qu.:280.5   1st Qu.:17.00   1st Qu.:374.71
Median : 5.000   Median :330.0   Median :18.90   Median :391.23
Mean   : 9.679   Mean   :409.8   Mean   :18.42   Mean   :355.81
3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.06
Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
poblacionpobre   preciovivienda
Min.   : 1.730   Min.   : 5.00
1st Qu.: 6.925   1st Qu.:16.90
Median :11.340   Median :21.20
Mean   :12.664   Mean   :22.59
3rd Qu.:17.025   3rd Qu.:25.05
Max.   :37.970   Max.   :50.00
```

Para evitar posibles equivocaciones a la hora de realizar el análisis, **corregimos algunos nombres de variables** (en este caso, el de criminalidad y el de distanciacentro):

```
names(datos)[names(datos) == 'ï..criminalidad'] <- 'criminalidad'
names(datos)[names(datos) == 'ditanciacentro'] <- 'distanciacentro'
```

Podemos ayudarnos mediante el uso de **gráficos** para saber qué valores son más frecuentes. También podremos ver con mejor claridad qué datos encajarían mejor con rangos o con categorías. Para empezar, visualizaremos con un **histograma** y con un **diagrama de cajas** la frecuencia de valores para cada tasa de criminalidad:

hist(datos$criminalidad, main="Criminalidad", xlab="Tasa criminalidad", ylab="Número de zonas",col="lightblue",labels=TRUE)

boxplot(datos$criminalidad, main="Criminalidad", col="pink")

Ilustración 3. Histograma de criminalidad.

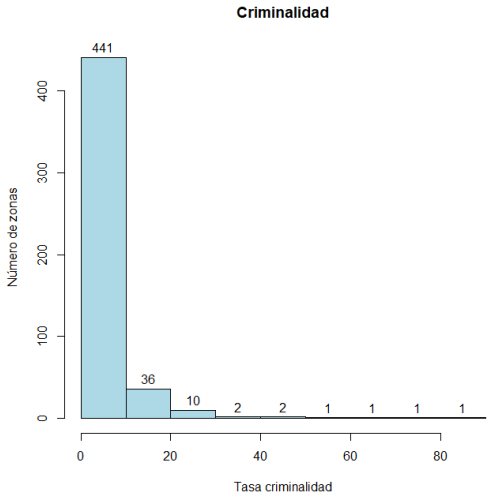
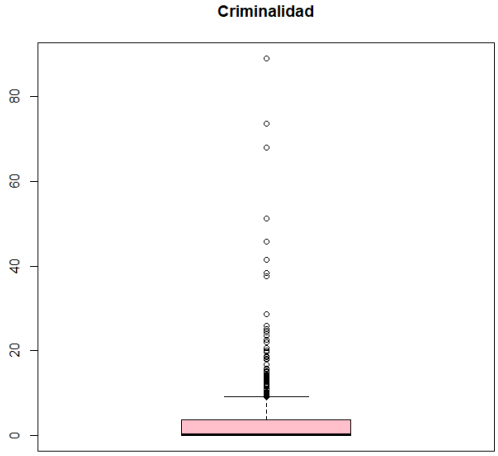


Ilustración 2. Diagrama de cajas de criminalidad.



De las 495 zonas a estudiar, vemos que **441 de ellas tienen una tasa de criminalidad entre 0 y 10.** Esto corresponde a que aproximadamente un 89% de las zonas a analizar entran dentro de este rango. Se puede observar que, a partir de una tasa aproximadamente de 10, aparecen múltiples valores atípicos.

Hacemos lo mismo para **terrenoresidencial**:

hist(datos$terrenoresidencial, main="Terreno residencial", xlab="Porcentaje terreno residencial", ylab="Número de zonas",col="lightblue",labels=TRUE)

hist(datos$terrenoresidencial, breaks = 1000, xlim=c(0,1), main="Terreno residencial", xlab="Porcentaje terreno residencial", ylab="Número de zonas",col="lightblue",labels=TRUE)

Ilustración 5. Histograma de terreno residencial.

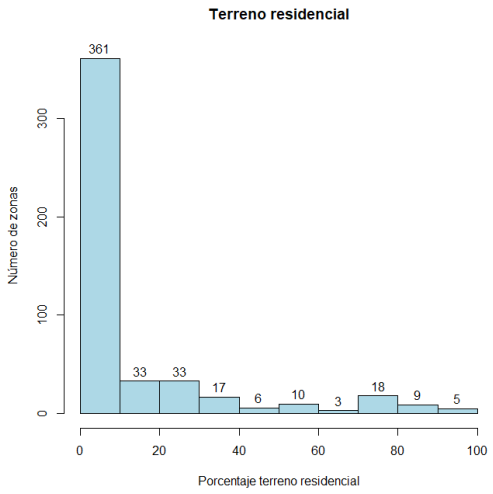
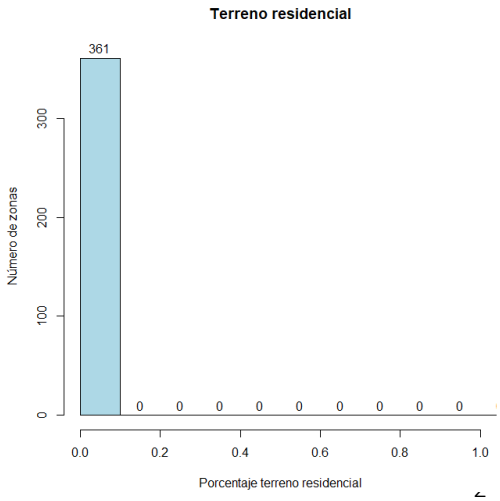


Ilustración 4. Histograma limitado de terreno residencial.



La mayoría de las zonas están entre 0 y 10% de terrenoresidencial. Si observamos este diagrama junto a los resultados obtenidos usando "summary", vemos que el valor mínimo encontrado es 0 y, por tanto, usamos el histograma entre 0 y 1 para observar qué ocurre con mayor detalle.

Como podemos ver, todas las zonas que estaban entre un 0 y 10% (361 zonas) en realidad se tratan de zonas que no tienen terreno residencial (podemos observar que las **361 zonas anteriores se encuentran en el 0**).

Realizamos las gráficas para el resto de las variables, pero se mostrarán en el Anexo B. Datos y gráficas adicionales.

Después del análisis exploratorio, procedemos a buscar que variables encajan con la definición de **variable categórica**. Tenemos la variable "**rio**" que solo va a tomar dos posibles valores (no limita con un río o sí limita con un río):

```
datos$rio<-factor(datos$rio, labels=c('No', 'Si'))
```

De la variable "terrenoresidencial", como tiene muchos valores 0, podemos sacar otra variable categórica "**habitada**" que indique si la zona está habitada (valor > 0) o no (valor = 0):

```
datos$habitada<-factor(ifelse(datos$terrenoresidencial==0, 'No', 'Si'))
```

Para comprobar cómo se distribuyen las zonas que limitan con un río según la tasa de criminalidad, podemos combinarlos en el mismo histograma haciendo uso de la librería "plotrix":

```
library(plotrix)
histStack(datos$criminalidad,datos$rio, xlim=c(0,1),legend.pos="topright")
```

Ilustración 7. Histograma para criminalidad respecto a rio.

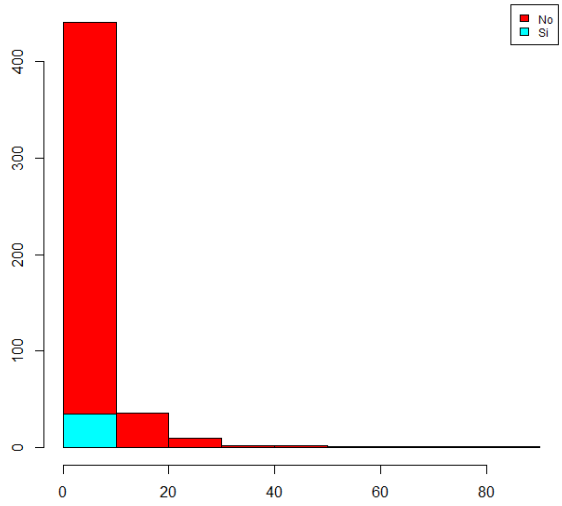
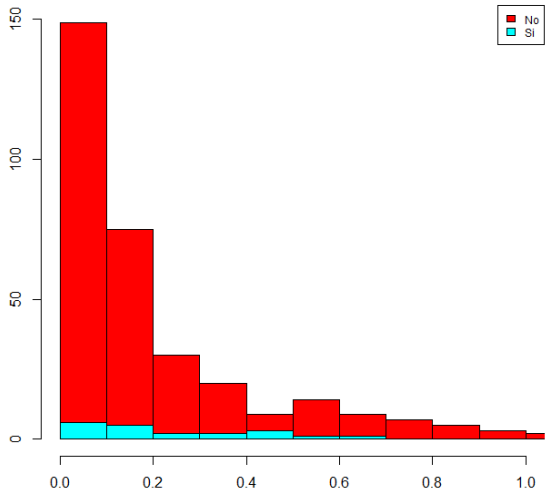


Ilustración 6. Histograma limitado para criminalidad respecto a rio.



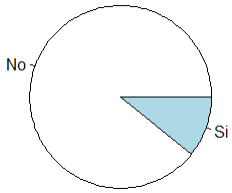
Se puede concluir que puede haber una ligera tendencia a cometer más crímenes en aquellas zonas que no limitan con ríos, pero es una diferencia muy poco significativa. De igual forma observamos que esta tendencia se parece sospechosamente a una función de la forma $\frac{1}{\log_a x}$.

También creamos otra variable categórica para separar aquellas tasas de criminalidad que son especialmente altas (corresponden a los **valores atípicos** vistos en el diagrama de cajas de criminalidad):

datos$altamentepeligrosa<-factor(ifelse(datos$criminalidad<10, 'No', 'Si'))

pie(table(datos$ altamentepeligrosa), main="Zona altamente peligrosa")

Ilustración 8. Diagrama de sectores de altamente peligrosa.



Con **tablas de contingencia** podemos ver más claras las relaciones entre las distintas variables categóricas:

table(datos$rio,datos$habitada)

table(datos$altamentepeligrosa,datos$rio)

Ilustración 9. Tablas de contingencia.

	No	Si
No	333	127
Si	28	7

La mayoría de las zonas estudiadas son no habitadas y sin limitar con un río.

	No	Si
No	406	35
Si	54	0

Ninguna de las zonas que limitan con un río pertenecen a la categoría de altamente peligrosa.

	No	Si
No	307	54
Si	134	0

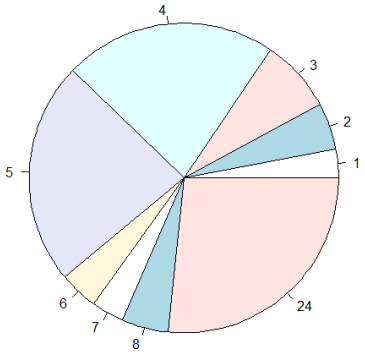
Ninguna de las zonas habitadas es considerada como altamente peligrosa.

Ilustración 10. Diagrama de sectores de acceso autopista.

Por último, también consideramos la variable **accesoautopistas** como categórica, ya que siempre toma los mismos valores:

datos$accesoautopista <- factor(datos$accesoautopista, labels=c('Uno', 'Dos', 'Tres', 'Cuatro', 'Cinco', 'Seis', 'Siete', 'Ocho', 'Veinticuatro'))

pie(table(datos$accesoautopista), main="Acceso autopista")



Finalmente, creamos unas variables que corresponderán a los diferentes **rangos** de valores que más ocurren para las variables de criminalidad y terrenoresidencial para poder visualizar mejor los datos:

datos$rangoterrenoresidencial <- cut(datos$terrenoresidencial,c(-1,0,20,40,60,80,100))

datos$rangocriminalidad <- cut(datos$criminalidad,c(0,1.5,10,Inf))

Gracias a tenerlas como rangos, podemos verlas de manera más clara en **diagramas de sectores**:

pie(table(datos$rangocriminalidad), main="Criminalidad")

pie(table(datos$rangoterrenoresidencial), main="Terreno residencial")

Ilustración 12. Diagrama de sectores de rangocriminalidad.

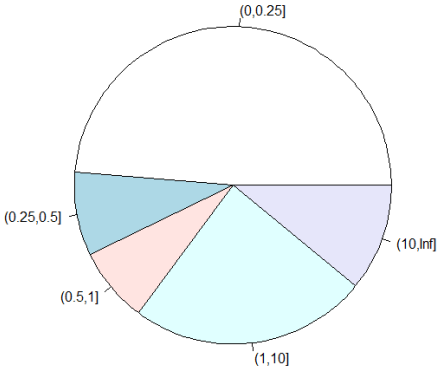
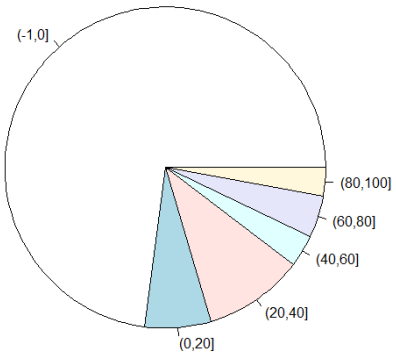


Ilustración 11. Diagrama de sectores de rangoterrenoresidencial.



Si hacemos ahora un summary(datos), observaremos todas las variables nuevas.

Ilustración 13. Resultado final de summary(datos).

```
 criminalidad      terrenoresidencial   industrias    rio      oxidonitrogeno
Min.   : 0.00632   Min.   :  0.00     Min.   : 0.46   No:460   Min.   :0.3850
1st Qu.: 0.08232   1st Qu.:  0.00     1st Qu.: 5.13   Si: 35   1st Qu.:0.4480
Median : 0.26938   Median :  0.00     Median : 8.56            Median :0.5320
Mean   : 3.69038   Mean   : 11.62     Mean   :11.15            Mean   :0.5541
3rd Qu.: 3.73597   3rd Qu.: 15.00     3rd Qu.:18.10            3rd Qu.:0.6275
Max.   :88.97620   Max.   :100.00     Max.   :27.74            Max.   :0.8710
 habitaciones   viviendasantiguas distanciacentro  accesoautopista
Min.   :3.561   Min.   :  2.90    Min.   : 1.130   Min.   : 1.000
1st Qu.:5.888   1st Qu.: 44.05    1st Qu.: 2.083   1st Qu.: 4.000
Median :6.211   Median : 77.70    Median : 3.280   Median : 5.000
Mean   :6.289   Mean   : 68.49    Mean   : 3.824   Mean   : 9.679
3rd Qu.:6.627   3rd Qu.: 94.20    3rd Qu.: 5.223   3rd Qu.:24.000
Max.   :8.780   Max.   :100.00    Max.   :12.127   Max.   :24.000
  impuestos      ratioalumnos   poblacionnegra   poblacionpobre
Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.730
1st Qu.:280.5   1st Qu.:17.00   1st Qu.:374.71   1st Qu.: 6.925
Median :330.0   Median :18.90   Median :391.23   Median :11.340
Mean   :409.8   Mean   :18.42   Mean   :355.81   Mean   :12.664
3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.06   3rd Qu.:17.025
Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.970
preciovivienda  habitada altamentepeligrosa rangoterrenoresidencial
Min.   : 5.00   No:361   No:441            (-1,0]  :361
1st Qu.:16.90   Si:134   Si: 54            (0,20]  : 33
Median :21.20                              (20,40] : 50
Mean   :22.59                              (40,60] : 16
3rd Qu.:25.05                              (60,80] : 21
Max.   :50.00                              (80,100]: 14
 rangocriminalidad
(0,0.25]  :241
(0.25,0.5]: 42
(0.5,1]   : 38
(1,10]    :120
(10,Inf]  : 54
```

Análisis descriptivo.

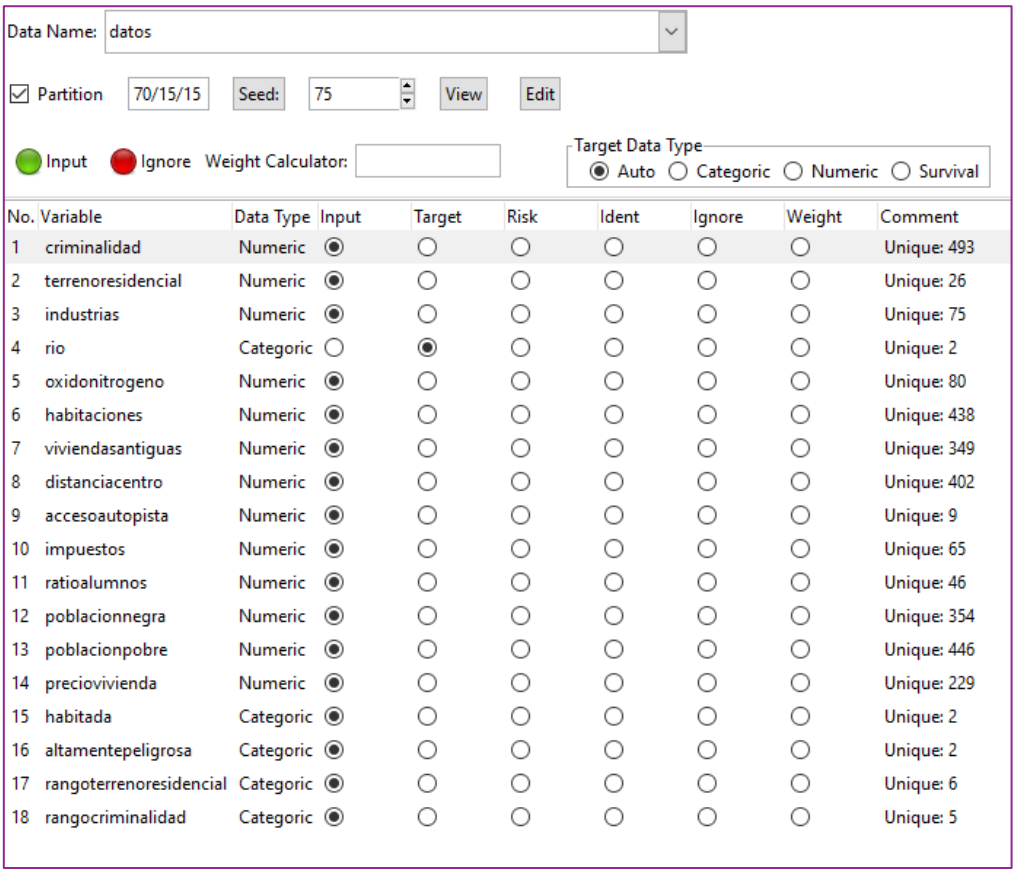
Para realizar este análisis, agruparemos las diferentes zonas de la ciudad en diferentes **clústers** (grupos de elementos que tienen similitudes entre ellos), usando **kmeans**. De esta forma, podremos describir los datos que previamente hemos explorado organizándolos en grupos, ya que se "reduce" el conjunto de datos mejorando su comprensión.

Para empezar, hacemos uso de la función rattle:

rattle()

Una vez iniciado, cargaremos los datos como un "R Dataset", seleccionando el conjunto "datos" en el desplegable de "Data name". Una vez hecho esto, le damos al botón de ejecutar para cargar los datos.

Ilustración 14. Ventana con los datos cargados en rattle.



No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	criminalidad	Numeric	◉	○	○	○	○	○	Unique: 493
2	terrenoresidencial	Numeric	◉	○	○	○	○	○	Unique: 26
3	industrias	Numeric	◉	○	○	○	○	○	Unique: 75
4	rio	Categoric	○	◉	○	○	○	○	Unique: 2
5	oxidonitrogeno	Numeric	◉	○	○	○	○	○	Unique: 80
6	habitaciones	Numeric	◉	○	○	○	○	○	Unique: 438
7	viviendasantiguas	Numeric	◉	○	○	○	○	○	Unique: 349
8	distanciacentro	Numeric	◉	○	○	○	○	○	Unique: 402
9	accesoautopista	Numeric	◉	○	○	○	○	○	Unique: 9
10	impuestos	Numeric	◉	○	○	○	○	○	Unique: 65
11	ratioalumnos	Numeric	◉	○	○	○	○	○	Unique: 46
12	poblacionnegra	Numeric	◉	○	○	○	○	○	Unique: 354
13	poblacionpobre	Numeric	◉	○	○	○	○	○	Unique: 446
14	preciovivienda	Numeric	◉	○	○	○	○	○	Unique: 229
15	habitada	Categoric	◉	○	○	○	○	○	Unique: 2
16	altamentepeligrosa	Categoric	◉	○	○	○	○	○	Unique: 2
17	rangoterrenoresidencial	Categoric	◉	○	○	○	○	○	Unique: 6
18	rangocriminalidad	Categoric	◉	○	○	○	○	○	Unique: 5

En la ventana donde se nos muestran los datos cargados, seleccionaremos como "Ignore" aquellas variables fuera del grupo que nos interesa analizar, dejando como "Input" aquellas que sí queremos agrupar. Como inicialmente no tenemos clara la influencia que tienen las distintas variables en la tasa de criminalidad, las agrupamos todas excepto las categóricas. Volvemos a darle a "Ejecutar" y procedemos a ir a la pestaña de "Cluster". En ésta, basándonos en el **criterio del codo**, iremos probando diferentes números de clústeres hasta encontrar aquel donde haya una "subida". Justamente el número donde empieza a crecer la función es aquel que usaremos para dividir los clústeres. Mientras estemos probado diferentes números de clústeres, dejamos marcada la casilla de "Iterate Clusters".

Ilustración 15. Pestaña de Cluster en rattle.

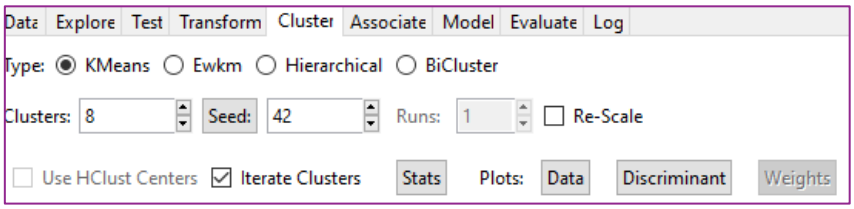
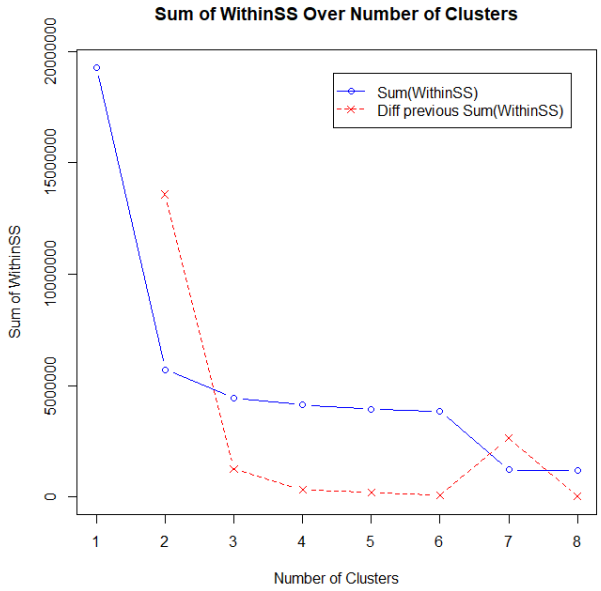


Ilustración 16. Función número de clústeres.



Como la subida comienza en 6, desmarcamos "Iterate Clusters" y le indicamos que queremos 6 clústeres.

Ilustración 17. Resultado inicial de análisis descriptivo.

```
Cluster sizes:

[1] "59 137 57 77 46 119"

Data means:

      criminalidad terrenoresidencial         industrias     oxidonitrogeno       habitaciones
         3.6903817         11.6161616         11.1463030          0.5541428          6.2889758
 viviendasantiguas    distanciacentro    accesoautopista          impuestos       ratioalumnos
        68.4872727          3.8235697          9.6787879        409.8121212         18.4208081
    poblacionnegra     poblacionpobre     preciovivienda
       355.8065859         12.6636566         22.5943434

Cluster centers:

  criminalidad terrenoresidencial industrias oxidonitrogeno habitaciones viviendasantiguas distanciacentro
1   0.09569373           9.347458   7.022203      0.4799271     6.445780          56.62203        4.622090
2  12.29916168           0.000000  18.451825      0.6701022     6.006212          89.96788        2.054470
3   0.09398439          38.342105   5.935263      0.4379474     6.626018          28.37719        5.710188
4   0.99156714           0.974026  15.573247      0.6243117     6.103753          84.58961        2.491813
5   0.08265196          50.619565   3.786957      0.4223261     6.570174          28.11304        7.390041
6   0.42519420           5.121849   7.256807      0.5186471     6.386479          74.04034        4.043774
  accesoautopista impuestos ratioalumnos poblacionnegra poblacionpobre preciovivienda
1        3.559322  217.6441     18.19492       391.4593       9.800339       26.96780
2       23.270073  667.6423     20.19635       291.0391      18.674526       16.27226
3        4.157895  273.7193     17.46842       389.2488       6.665263       28.85263
4        4.831169  411.8182     17.77532       355.2904      14.071818       21.07143
5        4.304348  353.2826     17.11304       388.2748       6.568261       26.65435
6        4.924370  294.0000     17.86807       384.4591      11.481429       24.12269
```

Observamos que todas las zonas tienen una **tasa de criminalidad** menor a 1, pero se localiza una **excepción con un valor de 12.30**. Para este valor tan alto, observamos el resto de las variables del clúster para tratar de encontrar un patrón que lo justifique. En primer lugar, vemos que es el único clúster que **carece de terreno residencial**, esto nos indica que estamos en presencia de una **zona industrial**, como podría ser un polígono. A continuación, vemos que obtiene el mayor valor de óxido de nitrógeno el cual nos indica una **alta contaminación**, lógico teniendo en cuenta que se trata de una zona industrial. A su vez sorprende que siendo un área de este tipo se trate de la **zona más próxima al centro**. Este clúster también presenta el **mayor número de accesos a autopistas** con 23, una gran diferencia con respecto al resto que están por debajo de 5. Destaca de igual forma por tener los **mayores impuestos, ratio de alumnos por profesor y mayor porcentaje de población pobre**, **así como de viviendas antiguas**, sin embargo, tiene el **menor índice de población negra y precio de vivienda.**

En cuanto a las tasas de criminalidad inferiores a 1, la más alta es 0.991. Para esta se observa que apenas tiene terreno residencial, pasa un caso parecido al anterior con la cantidad de industrias, a su vez también tiene la segunda mayor contaminación, impuestos, población pobre y viviendas antiguas mientras que tiene la segunda menor cercanía al centro, precio de vivienda y cantidad de población negra. En cuanto al número de accesos autopista es similar al resto de clústeres exceptuando el previamente comentado, al igual que sucede con la ratio de alumnos similar al resto.

La menor tasa de todos los clústeres es de 0.082 y es el clúster que **mayor porcentaje de terreno residencial, distancia al centro y población negra** tiene. También es el que **menos industria, menos contaminación y menos viviendas antiguas**. En cuanto a los impuestos está en la media en comparación al resto de clústeres.

Para la media de habitaciones vemos que todas las zonas tienen aproximadamente 6, la cual es la media sin agrupar.

En criminalidad, podemos deducir que la media sube a 3 por culpa del valor "extremo" de 12, puesto que la mayoría están entre 0 y 1. En cuanto al terreno residencial, concluimos que cuanto menor es hay una mayor criminalidad. Por tanto, **en aquellos conjuntos donde abunden residencias, disminuye considerablemente la criminalidad.** Obtenemos que los **valores de 0 en terreno residencial van de la mano con valores más altos de industria y óxido nitrógeno**, lo cual tiene sentido si estamos hablando de zonas como polígonos, que tendrán más zona industrial y, por tanto, una mayor contaminación del aire. **En viviendas antiguas, vemos una tendencia a mayor criminalidad** cuando aumenta esta dimensión. Podría tener relación con las mayores tasas de criminalidad son producidas también con zonas más céntricas y no una influencia directa en la criminalidad. El índice de autopistas solo influye para esas zonas de criminalidad excesiva, al igual que la ratio de alumnos, aunque éste último no aumenta demasiado. **Unos impuestos muy altos influyen significativamente en la tasa de criminalidad**, sin embargo, para aquellas zonas donde los impuestos son más bajos hay una mayor tasa de criminalidad respecto a aquellas zonas donde se acercan más a la media de impuestos. Se comprueba también que, **a menor población negra, mayor tendencia al crimen** y que **cuanta más población pobre y cuanto menor sea el precio de las viviendas, probablemente debido a que son zonas con peor caché, mayor es la tasa de criminalidad.**

Ilustración 18. Resultados clúster terreno, industrias, oxido, centro, autopista.

```
Cluster sizes:

[1] "58 132 233 72"

Data means:

terrenoresidencial         industrias     oxidonitrogeno    distanciacentro    accesoautopista
        11.6161616         11.1463030          0.5541428          3.8235697          9.6787879

Cluster centers:

  terrenoresidencial industrias oxidonitrogeno distanciacentro accesoautopista
1          69.181034   3.069655      0.4171759        7.011690        3.517241
2           0.000000  18.100000      0.6724167        2.061254       24.000000
3           7.457082   6.498026      0.4955386        4.452689        4.742489
4           0.000000  19.946389      0.6372917        2.450374        4.361111
```

Al realizar un **clúster con las variables terrenoresidencial, industrias, oxidonitrogeno, distanciacentro y accesoautopista**, se ven con más claridad las relaciones entre dichas variables que teníamos en duda. A menor terreno residencial, mayor industria, contaminación y una menor distancia al centro.

Si al anterior clúster le añadimos la criminalidad, nos muestra que, efectivamente**, la criminalidad aumenta en aquellas zonas industriales**.

Ilustración 19. Clúster zonas industriales y criminalidad.

```
Cluster sizes:

[1] "363 132"

Data means:

      criminalidad terrenoresidencial         industrias     oxidonitrogeno    distanciacentro
         3.6903817         11.6161616         11.1463030          0.5541428          3.8235697
   accesoautopista
         9.6787879

Cluster centers:

  criminalidad terrenoresidencial industrias oxidonitrogeno distanciacentro accesoautopista
1    0.3925966           15.84022   8.617686      0.5111342        4.464412        4.471074
2   12.7592909            0.00000  18.100000      0.6724167        2.061254       24.000000
```

En esta misma ventana, tenemos las opciones para “Plots” de “Data” y “Discriminant” que nos permiten visualizar de una forma más clara la separación entre los clústeres.

Ilustración 20. Plot data de clústeres criminalidad en zonas industriales.

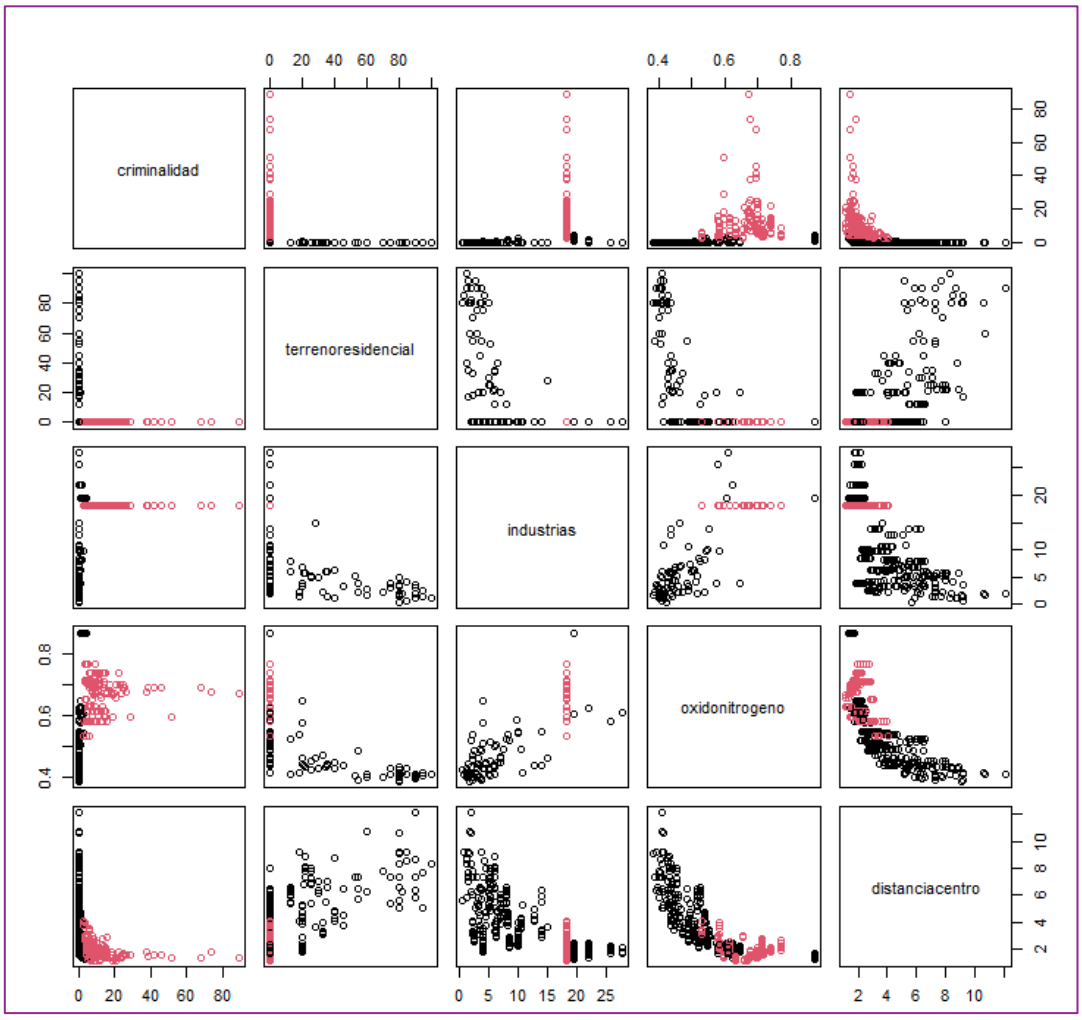
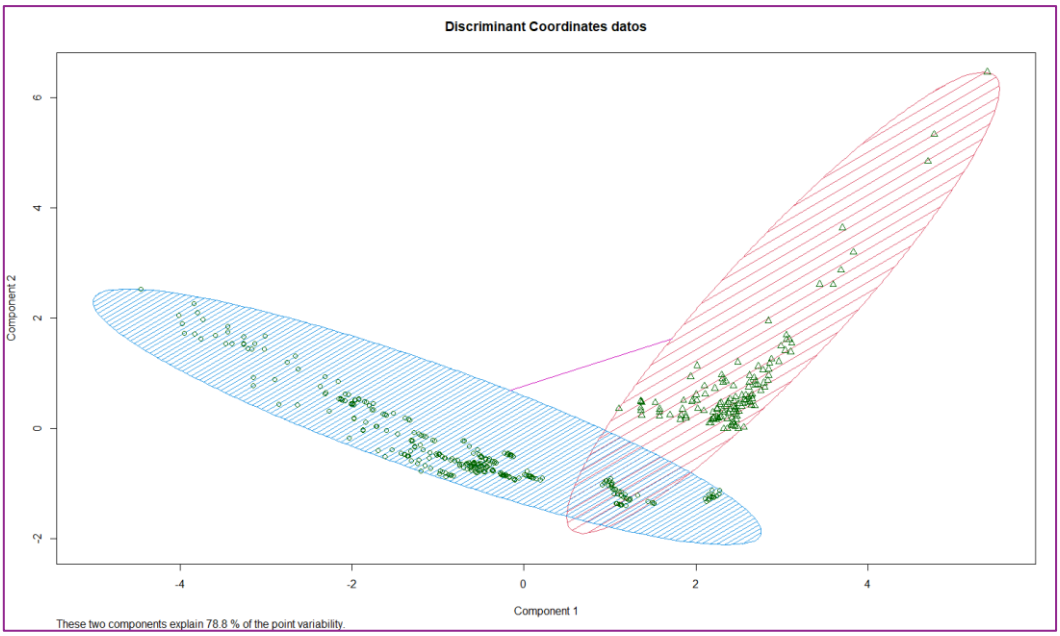


Ilustración 21. Plot Disriminant de clústeres criminalidad en zonas industriales.

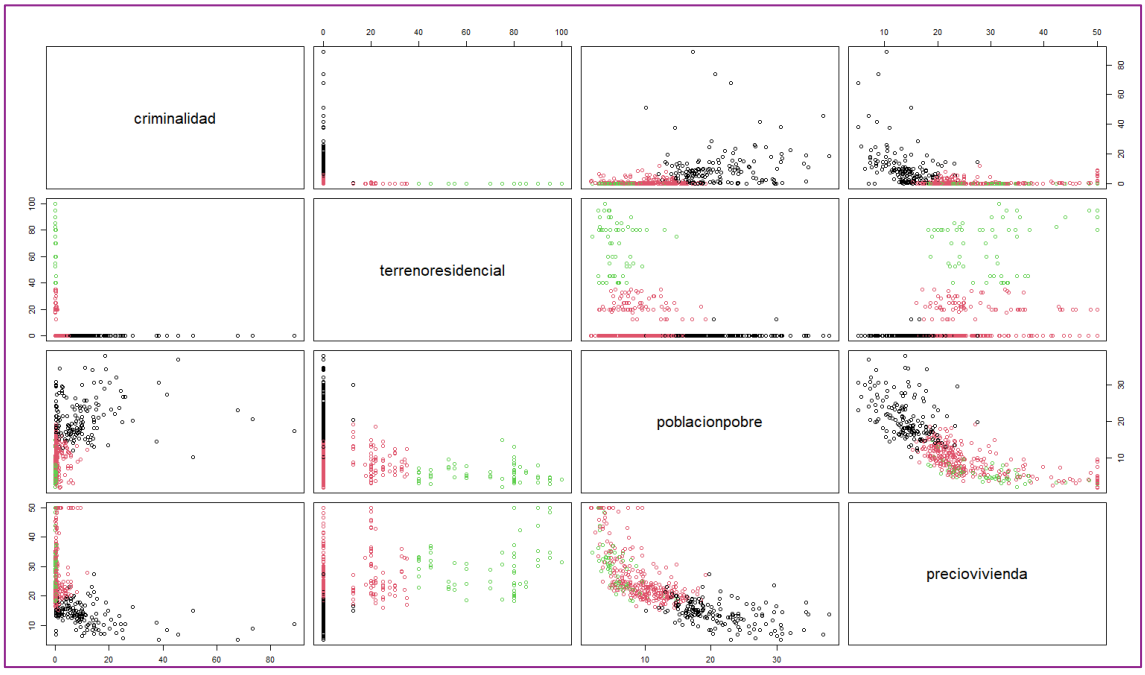


Si hacemos un **clúster para criminalidad, terrenoresidencial, poblacionpobre y preciovivienda**, se ven observa que claramente hay una relación entre la criminalidad y una mayor pobreza. Además, se puede ver que esta situación se da en zonas con menor precio de vivienda y menor terreno residencial.

Ilustración 22. Clúster criminalidad y pobreza.

```
Cluster sizes:

[1] "155 282 58"

Data means:

      criminalidad terrenoresidencial     poblacionpobre      preciovivienda
          3.690382          11.616162          12.663657          22.594343

Cluster centers:

  criminalidad terrenoresidencial poblacionpobre preciovivienda
1  10.29796619          0.1612903      21.032065       14.09290
2   0.80846996          6.0726950       9.491986       25.75851
3   0.04423569         69.1810345       5.720690       29.92931
```

Ilustración 23. Plot Data de clústeres criminalidad población pobre.



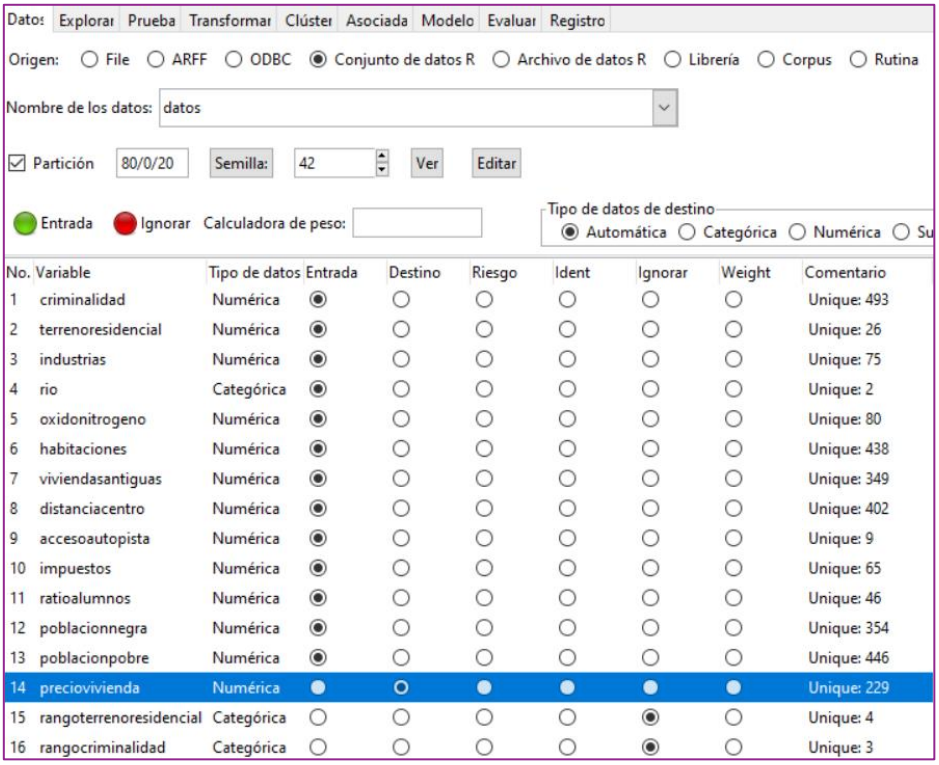
Predicción mediante regresión lineal.

Se solicita realizar una **predicción sobre el precio de la vivienda** de la zona en función del resto de variables. Para ello, empleado "rattle", aplicaremos **regresión lineal múltiple**.

El primer paso, al igual que en el análisis descriptivo, es cargar los datos en "rattle". Esto se realiza seleccionando la opción conjunto de datos R y el nombre de los datos en el desplegable en nuestro caso estos se llaman homónimamente. Una vez hecho esto los cargamos pulsando el botón ejecutar, una vez cargados procedemos con la elección de variables del primer modelo, para ello seleccionamos todas las variables a excepción de las variables correspondientes a los rangos, que se ignoran, en el caso de la variable a predecir, precio vivienda, la marcamos como variable destino.

Ejecutamos nuevamente con la casilla **partición** seleccionada con los valores **80/0/20** que se reparten como porcentajes de los conjuntos de entrenamiento, validación y test respectivamente. No asignamos valor al conjunto de validación puesto que usaremos la **técnica k-fold**.

Ilustración 24. Ventana con las variables iniciales y partición seleccionada.



A continuación, nos dirigimos a la pestaña Modelo, en la que tendremos que elegir la opción lineal junto a numérica, indicando de esta forma que se quiere realizar una regresión lineal numérica.

Ilustración 25. Pestaña "Model" en rattle.

Una vez seleccionamos estas opciones, ejecutamos para generar el modelo. Como nos indica, el número de "*" que contiene a la derecha de la variable determina la importancia de esta en el cálculo del modelo, esto lo usaremos para dirimir que variables deben o no continuar en el modelo.

Ilustración 26. Modelo inicial de regresión lineal.

```
Summary of the Linear Regression model (built using lm):

Call:
lm(formula = preciovivienda ~ ., data = crs$dataset[crs$train,
    c(crs$input, crs$target)])

Residuals:
     Min       1Q   Median       3Q      Max
-15.2087  -2.7684  -0.5504   1.8325  25.2891

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       37.147803   5.663434   6.559 1.76e-10 ***
criminalidad      -0.099656   0.041113  -2.424  0.01582 *
terrenoresidencial  0.039153   0.015708   2.493  0.01310 *
industrias         0.041376   0.068162   0.607  0.54420
rioSi              1.972263   0.920559   2.142  0.03279 *
oxidonitrogeno    -19.005512   4.230436  -4.493 9.34e-06 ***
habitaciones       3.745355   0.451063   8.303 1.77e-15 ***
viviendasantiguas  0.010290   0.014931   0.689  0.49115
distanciacentro   -1.437331   0.225001  -6.388 4.89e-10 ***
accesoautopista    0.299748   0.073389   4.084 5.39e-05 ***
impuestos         -0.012093   0.004079  -2.964  0.00322 **
ratioalumnos      -0.947359   0.152840  -6.198 1.48e-09 ***
poblacionnegra     0.009056   0.002933   3.087  0.00217 **
poblacionpobre    -0.563708   0.055047 -10.241  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.738 on 382 degrees of freedom
Multiple R-squared:  0.7426,    Adjusted R-squared:  0.7338
F-statistic: 84.78 on 13 and 382 DF,  p-value: < 2.2e-16
```

Una vez generado el modelo, éste debe ser evaluado. Para ello, nos movemos a la pestaña evaluar y elegimos el tipo calificación y el conjunto de datos de prueba. Seguidamente, ejecutamos el programa y generará un **archivo con las predicciones**. Una vez guardado el archivo, abandonamos rattle para movernos a la consola de R y, aquí, calcularemos el error del modelo. Este se calculará mediante el uso del **error cuadrático medio**, esto es, la diferencia cuadrática promedio entre el valor real y el estimado, que será mejor cuanto más se aproxime a 0. Para obtenerlo, primero cargamos los resultados desde el archivo con el primer comando para, después, calcular el error con el otro comando. En este caso asignamos el error a una variable:

```
prueba<-read.csv("prueba.csv")
errorprueba<-sqrt(mean((prueba$preciovivienda - prueba$glm)^2,na.rm=TRUE))
```

Por tanto, partiendo de un primer modelo con todas las variables a excepción de los rangos, obtenemos los siguientes modelos, los cuales son evaluados y comparados en la tabla posterior:

- Modelo 1 ➡ (criminalidad, terrenoresidencial, industrias, rio, oxidonitrogeno, habitaciones, viviendasantiguas, distanciacentro, accesoautopista, impuestos, ratioalumnos, poblacionnegra, poblacionpobre)
- Modelo 2 ➡ (criminalidad, terrenoresidencial, rio, oxidonitrogeno, habitaciones, distanciacentro, accesoautopista, impuestos, ratioalumnos, poblacionnegra, poblacionpobre)
- Modelo 3 ➡ (oxidonitrogeno, habitaciones, distanciacentro, accesoautopista, ratioalumnos, poblacionpobre)

- Modelo 4 ➡ (oxidonitrogeno, habitaciones, distanciacentro, ratioalumnos, poblacionpobre)

Tabla 1. Resultados para diferentes valores de semilla

Modelo	42	43	44	45	46	47	48	Media
1	4.821029	4.558702	4.385251	5.609852	5.972888	4.524956	3.600816	4,781928
2	4.805823	4.522392	4.387044	5.610653	5.961524	4.513360	3.574578	4,767910
3	5.240272	4.411376	4.472344	5.878720	6.141704	4.558558	3.935647	4.948374
4	5.228729	4.388212	4.455443	5.891013	6.150639	4.554335	3.887842	4.936602

Adicionalmente a estos modelos, probamos otros utilizando los rangos descritos, pero no se registran debido a que tras breves pruebas variando la semilla se observa un peor rendimiento de forma clara. Conforme con esto y las medidas arriba expuestas elegimos como **el mejor el modelo 2 por ser el de menor error cuadrático medio**. Por tanto, podemos determinar que las variables utilizadas en este modelo son las de mayor influencia en la determinación del resultado, mientras que las que no están son las de menor influencia.

Predicción mediante regresión logística.

Nos interesa **predecir si la tasa de criminalidad de una zona es alta o baja**. Según lo que vimos en el análisis exploratorio, la **mediana** de esta variable corresponde a un 0.26. Basándonos en esto, podemos considerar que una criminalidad es alta a partir de dicho valor. Creamos, pues, la siguiente variable categórica tanto para el conjunto de datos como para el de test:

```
datos$criminalidadalta<-factor(ifelse(datos$criminalidad<0.26, 'No', 'Si'))
test$criminalidadalta<-factor(ifelse(test$criminalidad<0.26, 'No', 'Si'))
```

A diferencia de la predicción lineal numérica, hacemos una **partición de 70/15/15** con el conjunto "datos" (se divide también para un conjunto de validación). El procedimiento sigue siendo el mismo que en el anterior apartado; comenzamos con todas las variables inicialmente, excepto criminalidad y las categóricas) y vamos probando hasta obtener el menor error posible. En este caso, la variable "Target" debe ser la variable categórica "**criminalidadalta**" que creamos previamente. Después de ejecutar con la partición, vamos a la pestaña "Model" y seleccionamos las opciones "Linear" y "Logistic" para que nos haga la predicción logística y nos saldrán los datos de una forma similar a como pasaba en la predicción lineal (más asteriscos al lado de aquellas variables más significativas). Para comprobar el acierto y error de la predicción, nos dirigimos a la pestaña de "Evaluate" y seleccionamos la opción de "Error Matrix". Con estas variables iniciales, se obtiene un error medio de 6.7%. Si quitamos la variable preciovivienda seguimos obteniendo la misma precisión. Al quitar poblacionpobre, disminuye el error hasta un 5.35%. Si quitamos poblacionnegra, viviendasantiguas,

habitaciones y terrenoresidencial sigue siendo **5.35%. Es, por tanto, el porcentaje más bajo de error que podemos obtener.**

Ilustración 27. Variables usadas en regresión logística.

Data Name: datos

Partition 70/15/15 Seed: 42 View Edit

Input Ignore Weight Calculator:

Target Data Type: Auto Categoric Numeric Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	criminalidad	Numeric	○	○	○	○	●	○	Unique: 493
2	terrenoresidencial	Numeric	○	○	○	○	●	○	Unique: 26
3	industrias	Numeric	●	○	○	○	○	○	Unique: 75
4	rio	Categoric	○	○	○	○	●	○	Unique: 2
5	oxidonitrogeno	Numeric	●	○	○	○	○	○	Unique: 80
6	habitaciones	Numeric	○	○	○	○	●	○	Unique: 438
7	viviendasantiguas	Numeric	○	○	○	○	●	○	Unique: 349
8	distanciacentro	Numeric	●	○	○	○	○	○	Unique: 402
9	accesoautopista	Numeric	●	○	○	○	○	○	Unique: 9
10	impuestos	Numeric	●	○	○	○	○	○	Unique: 65
11	ratioalumnos	Numeric	●	○	○	○	○	○	Unique: 46
12	poblacionnegra	Numeric	○	○	○	○	●	○	Unique: 354
13	poblacionpobre	Numeric	○	○	○	○	●	○	Unique: 446
14	preciovivienda	Numeric	○	○	○	○	●	○	Unique: 229
15	habitada	Categoric	○	○	○	○	●	○	Unique: 2
16	altamentepeligrosa	Categoric	○	○	○	○	●	○	Unique: 2
17	rangoterrenoresidencial	Categoric	○	○	○	○	●	○	Unique: 6
18	rangocriminalidad	Categoric	○	○	○	○	●	○	Unique: 3
19	criminalidadalta	Categoric	○	●	○	○	○	○	Unique: 2

Ilustración 28. Resultados predicción logística.

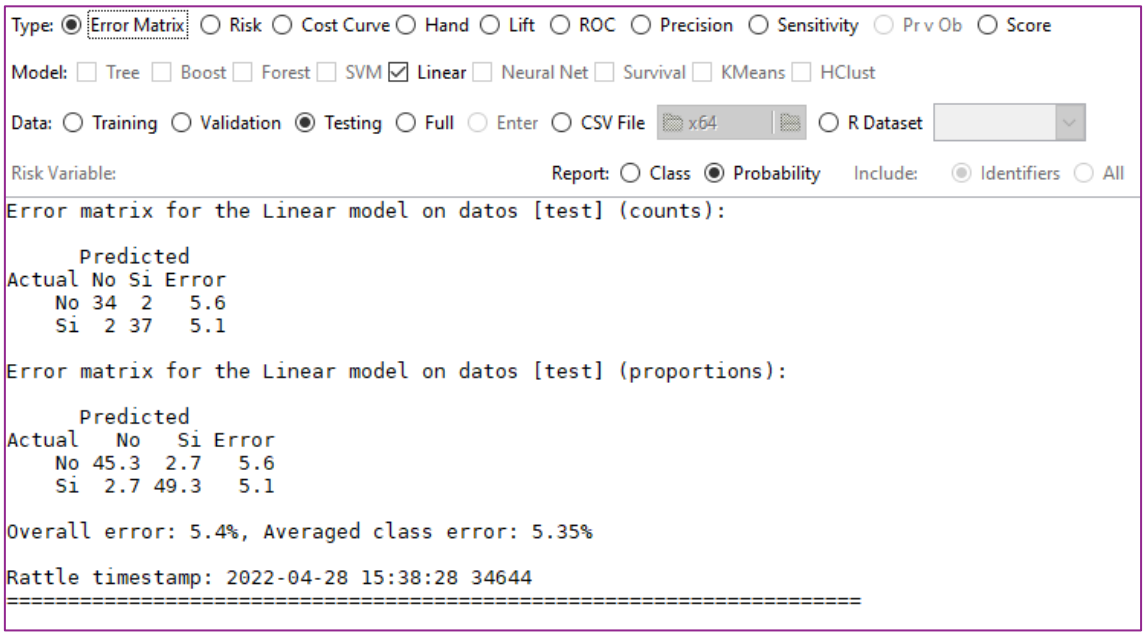
Type: ○ Tree ○ Forest ○ Boost ○ SVM ◉ Linear ○ Neural Net ○ Survival ○ All

○ Numeric ○ Generalized ○ Poisson ◉ Logistic ○ Probit ○ Multinomial

Plot

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.39949  -0.17842   0.00003   0.00558   2.74640

Coefficients:
                  Estimate Std. Error z value     Pr(>|z|)
(Intercept)     -33.581871   6.006565  -5.591 0.0000000226 ***
industrias       -0.143466   0.055913  -2.566      0.01029 *
oxidonitrogeno   57.631308  10.157876   5.674 0.0000000140 ***
distanciacentro   0.208860   0.184713   1.131      0.25817
accesoautopista   0.632825   0.156827   4.035 0.0000545601 ***
impuestos        -0.009480   0.003234  -2.932      0.00337 **
ratioalumnos      0.181828   0.106580   1.706      0.08800 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 479.55  on 345  degrees of freedom
Residual deviance: 152.12  on 339  degrees of freedom
AIC: 166.12

Number of Fisher Scoring iterations: 9

Log likelihood: -76.062 (7 df)
Null/Residual deviance difference: 327.429 (6 df)
Chi-square p-value: 0.00000000
Pseudo R-Square (optimistic): 0.85239902

==== ANOVA ====

Analysis of Deviance Table

Model: binomial, link: logit

Response: criminalidadalta

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           345     479.55
industrias      1  124.816       344     354.74 < 2.2e-16 ***
oxidonitrogeno  1  150.766       343     203.97 < 2.2e-16 ***
distanciacentro 1    0.969       342     203.00   0.32502
accesoautopista 1   38.138       341     164.86 6.591e-10 ***
impuestos       1    9.784       340     155.08   0.00176 **
ratioalumnos    1    2.956       339     152.12   0.08558 .
```

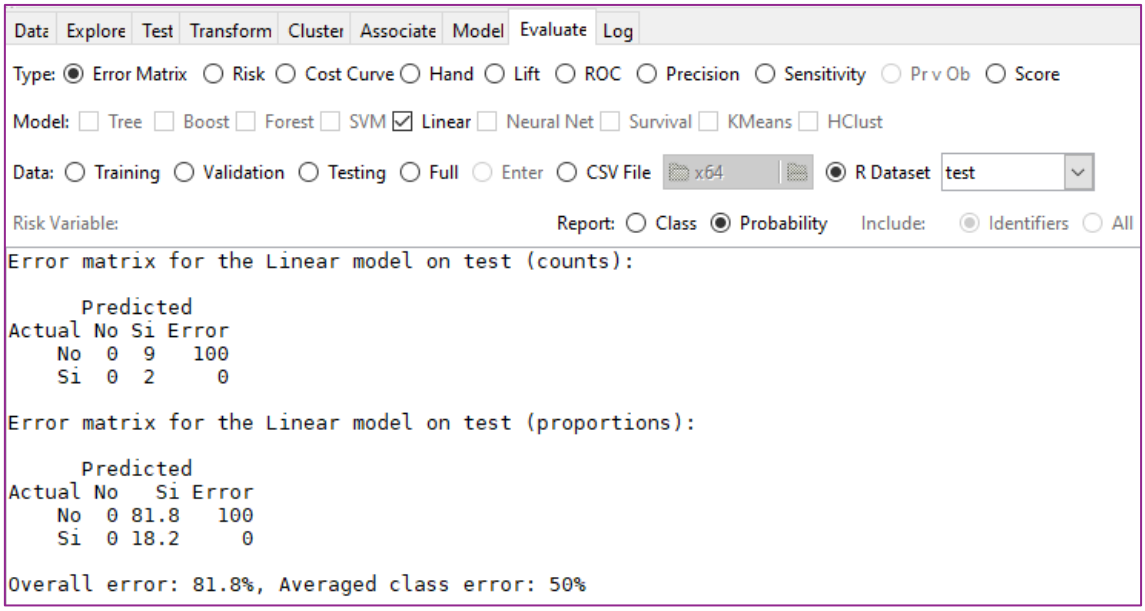
Ilustración 29. Resultados matriz de error.



El modelo ideal quedaría entonces como **(industrias, oxidonitrogeno, distanciacentro, accesoautopista, impuestos, ratioalumnos).**

Probamos el mismo procedimiento, pero **para el conjunto de test** que se nos ha proporcionado. En este caso, el error es muy alto. Esto posiblemente se deba a que los datos no son de buena calidad y/o son escasos.

Ilustración 30. Resultados matriz de error con conjunto test.



Anexo A. Comandos.

Para el conjunto datos:

```
library(rattle)
datos<-read.csv("datos.csv")
names(datos)[names(datos) == 'i..criminalidad'] <- 'criminalidad'
names(datos)[names(datos) == 'ditanciacentro'] <- 'distanciacentro'
datos$rio<-factor(datos$rio, labels=c('No', 'Si'))
datos$habitada<-factor(ifelse(datos$terrenoresidencial==0, 'No', 'Si'))
datos$altamentepeligrosa<-factor(ifelse(datos$criminalidad<10, 'No', 'Si'))
datos$rangoterrenoresidencial <- cut(datos$terrenoresidencial,c(-1,0,20,40,60,80,100))
datos$rangocriminalidad <- cut(datos$criminalidad,c(0,1.5,10,Inf))
datos$criminalidadalta<-factor(ifelse(datos$criminalidad<0.26, 'No', 'Si'))
summary(datos)
rattle()
```

Para el conjunto test:

```
test<-read.csv("test.csv")
names(test)[names(test) == 'i..criminalidad'] <- 'criminalidad'
names(test)[names(test) == 'ditanciacentro'] <- 'distanciacentro'
test$criminalidadalta<-factor(ifelse(test$criminalidad<0.26, 'No', 'Si'))
test$terrenoresidencial<-NULL
test$rio<-NULL
summary(test)
rattle()
```

Gráficos:

```
hist(datos$criminalidad, main="Criminalidad", xlab="Tasa criminalidad", ylab="Número de zonas",col="lightblue",labels=TRUE)
boxplot(datos$criminalidad, main="Criminalidad", col="pink")
hist(datos$terrenoresidencial, main="Terreno residencial", xlab="Porcentaje terreno residencial", ylab="Número de zonas",col="lightblue",labels=TRUE)
hist(datos$terrenoresidencial, breaks = 1000, xlim=c(0,1), main="Terreno residencial", xlab="Porcentaje terreno residencial", ylab="Número de zonas",col="lightblue",labels=TRUE)
datos$rio<-factor(datos$rio, labels=c('No', 'Si'))
datos$habitada<-factor(ifelse(datos$terrenoresidencial==0, 'No', 'Si'))
library(plotrix)
histStack(datos$criminalidad,datos$rio, xlim=c(0,1),legend.pos="topright")
```

```
datos$altamentepeligrosa<-factor(ifelse(datos$criminalidad<10, 'No', 'Si'))

pie(table(datos$ altamentepeligrosa), main="Zona altamente peligrosa")

datos$accesoautopista <- factor(datos$accesoautopista, labels=c('Uno', 'Dos', 'Tres', 'Cuatro', 'Cinco', 'Seis', 'Siete', 'Ocho', 'Veinticuatro'))

pie(table(datos$accesoautopista), main="Acceso autopista")

pie(table(datos$rangocriminalidad), main="Criminalidad")

pie(table(datos$rangoterrenoresidencial), main="Terreno residencial")

hist(datos$industrias, breaks = 10, main="Industrias", col="lightblue", xlab="Proporción de industrias de la zona", ylab="Número de zonas", labels=TRUE)

boxplot(datos$industrias, main="Industrias", col="pink")

pie(table(datos$rio))

hist(datos$oxidonitrogeno, breaks=10, main="Óxido de nitrógeno", col="lightblue", xlab="Cantidad de óxido de nitrógeno", ylab="Número de zonas", labels=TRUE)

boxplot(datos$oxidonitrogeno, main="Óxido de nitrógeno", col="pink")

hist(datos$habitaciones, breaks=10, main="Habitaciones", col="lightblue", xlab="Número medio de habitaciones por vivienda", ylab="Número de zonas", labels=TRUE)

boxplot(datos$habitaciones, main="Habitaciones", col="pink")

hist(datos$viviendasantiguas, breaks=10, main="Viviendas antiguas", col="lightblue", xlab="Proporción de viviendas antiguas", ylab="Número de zonas", labels=TRUE)

boxplot(datos$viviendasantiguas, main="Viviendas antiguas", col="pink")

hist(datos$distanciacentro, breaks=10, main="Distancia al centro", col="lightblue", xlab="Distancia al centro desde la zona", ylab="Número de zonas", labels=TRUE)

boxplot(datos$distanciacentro, main="Distancia centro", col="pink")

hist(datos$accesoautopista, xlim=c(0,25), breaks=10, main="Acceso autopista", col="lightblue", xlab=" Índice de acceso autopista ", ylab="Número de zonas", labels=TRUE)

boxplot(datos$accesoautopista, main="Acceso autopista", col="pink")

hist(datos$impuestos, main="Impuestos", xlim=c(100, 800),col="lightblue", xlab=" Impuestos sobre bienes inmuebles", ylab="Número de zonas", labels=TRUE)

boxplot(datos$impuestos, main="Impuestos", col="pink")

hist(datos$ratioalumnos, main="Ratio alumnos", col="lightblue", xlab=" Número de alumnos por profesor", ylab="Número de zonas", labels=TRUE)

boxplot(datos$ratioalumnos, main="Ratio alumnos", col="pink")

hist(datos$poblacionnegra, main="Población negra", col="lightblue", xlab=" Índice de población negra", ylab="Número de zonas", labels=TRUE)

boxplot(datos$poblacionnegra, main="Población negra", col="pink")

hist(datos$poblacionpobre, main="Población pobre", col="lightblue", xlab=" Porcentaje de población pobre", ylab="Número de zonas", labels=TRUE)

boxplot(datos$poblacionpobre, main="Población pobre", col="pink")

hist(datos$preciovivienda, main="Precio vivienda", col="lightblue", xlim=c(1,50), xlab=" Precio medio de viviendas (en miles de euros)", ylab="Número de zonas", labels=TRUE)

boxplot(datos$preciovivienda, main="Precio vivienda", col="pink")
```

Anexo B. Datos y gráficas adicionales.

Tabla 2. Datos iniciales del problema.

VARIABLE	DESCRIPCIÓN
criminalidad	tasa de criminalidad en la zona
terrenoresidencial	porcentaje de terreno residencial en la zona
industrias	proporción de industrias grandes en la zona
rio	variable con 1 si limita con un río, 0 si no limita
oxidonitrogeno	óxido de nitrógeno (partes en 10 millones)
habitaciones	número medio de habitaciones por vivienda
ratioalumnos	número de alumnos por profesor en los colegios de la zona
poblacionnegra	índice (sobre 1000) relacionado con la población negra (cuanto más alto es el valor de este índice, indica que hay un mayor porcentaje)
poblacionpobre	porcentaje de población de estado económico bajo
preciovivienda	precio medio de las viviendas ocupadas (en miles de euros

Tabla 3. Muestra de los datos (Excel).

criminalidad	terrenoresidencial	industrias	rio	oxidonitrogeno	habitaciones	viviendasantiguas	ditanciacentro	accesoautopista	impuestos	ratioalumnos	poblacionnegra	poblacionpobre	preciovivienda
632	18	231	0	538	6575	652	409	1	296	153	3969	498	24
2731	0	707	0	469	6421	789	49671	2	242	178	3969	914	216
2729	0	707	0	469	7185	611	49671	2	242	178	39283	403	347
3237	0	218	0	458	6998	458	60622	3	222	187	39463	294	334
6905	0	218	0	458	7147	542	60622	3	222	187	3969	533	362
2985	0	218	0	458	643	587	60622	3	222	187	39412	521	287
8829	125	787	0	524	6012	666	55605	5	311	152	3956	1243	229
14455	125	787	0	524	6172	961	59505	5	311	152	3969	1915	271
21124	125	787	0	524	5631	100	60821	5	311	152	38663	2993	165
17004	125	787	0	524	6004	859	65921	5	311	152	38671	171	189
22489	125	787	0	524	6377	943	63467	5	311	152	39252	2045	15
11747	125	787	0	524	6009	829	62267	5	311	152	3969	1327	189
9378	125	787	0	524	5889	39	54509	5	311	152	3905	1571	217
62976	0	814	0	538	5949	618	47075	4	307	21	3969	826	204
63796	0	814	0	538	6096	845	44619	4	307	21	38002	1026	182
62739	0	814	0	538	5834	565	44986	4	307	21	39562	847	199
105393	0	814	0	538	5935	293	44986	4	307	21	38685	658	231
7842	0	814	0	538	599	817	42579	4	307	21	38675	1467	175
80271	0	814	0	538	5456	366	37965	4	307	21	28899	1169	202
7258	0	814	0	538	5727	695	37965	4	307	21	39095	1128	182
125179	0	814	0	538	557	981	37979	4	307	21	37657	2102	136
85204	0	814	0	538	5965	892	40123	4	307	21	39253	1383	196
123247	0	814	0	538	6142	917	39769	4	307	21	3969	1872	152
98843	0	814	0	538	5813	100	40952	4	307	21	39454	1988	145
75026	0	814	0	538	5924	941	43996	4	307	21	39433	163	156
84054	0	814	0	538	5599	857	44546	4	307	21	30342	1651	139
67191	0	814	0	538	5813	903	4682	4	307	21	37688	1481	166
95577	0	814	0	538	6047	888	44534	4	307	21	30638	1728	148
77299	0	814	0	538	6495	944	44547	4	307	21	38794	128	184
100245	0	814	0	538	6674	873	4239	4	307	21	38023	1198	21
113081	0	814	0	538	5713	941	4233	4	307	21	36017	226	127

Ilustración 33. Histograma de industrias.

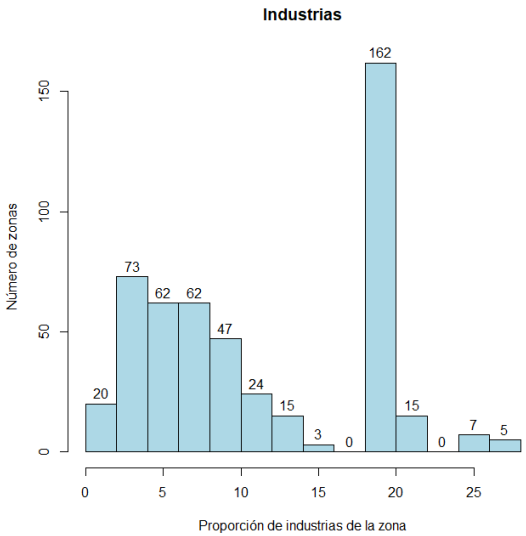


Ilustración 32. Diagrama de cajas de industrias.

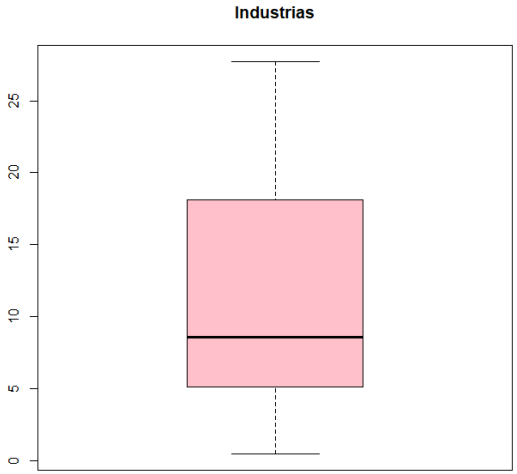


Ilustración 31. Diagrama de sectores de río.

Ilustración 35.Histograma de Óxido de nitrógeno.

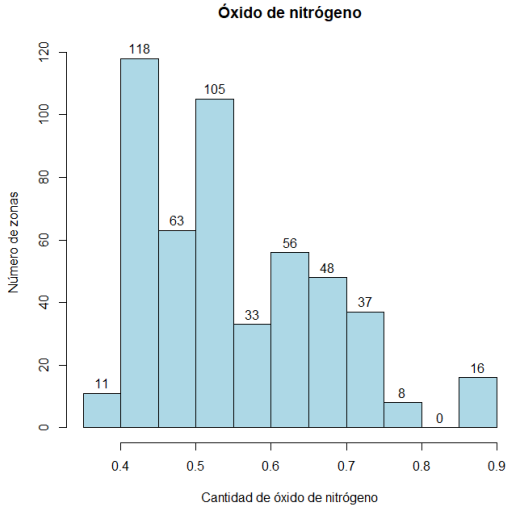


Ilustración 34. Diagrama de cajas de óxido de nitrógeno.

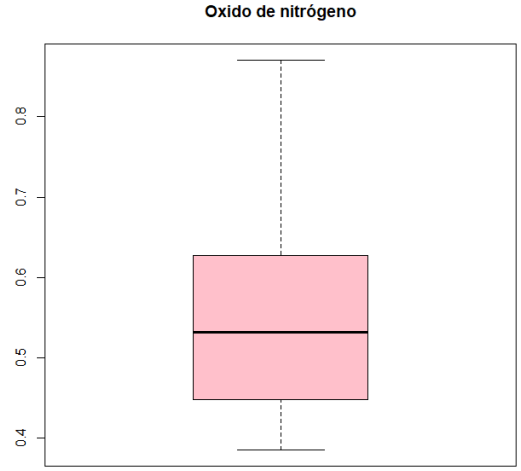


Ilustración 36. Histograma de habitaciones.

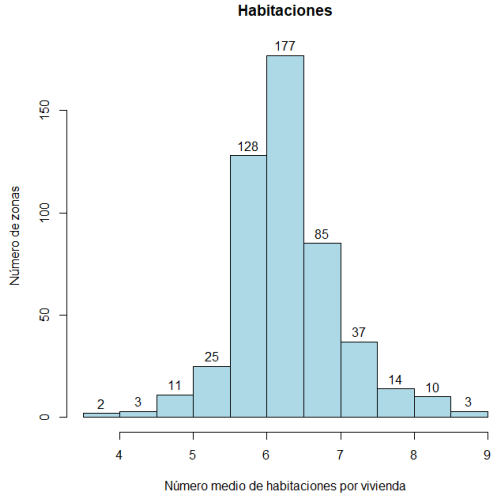


Ilustración 39. Diagrama de cajas de habitaciones.

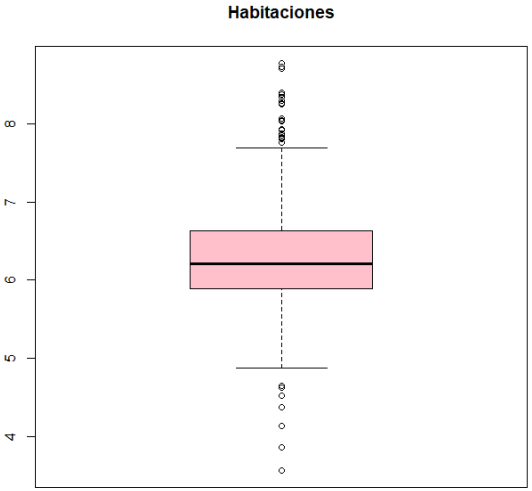


Ilustración 38. Histograma de viviendas antiguas.

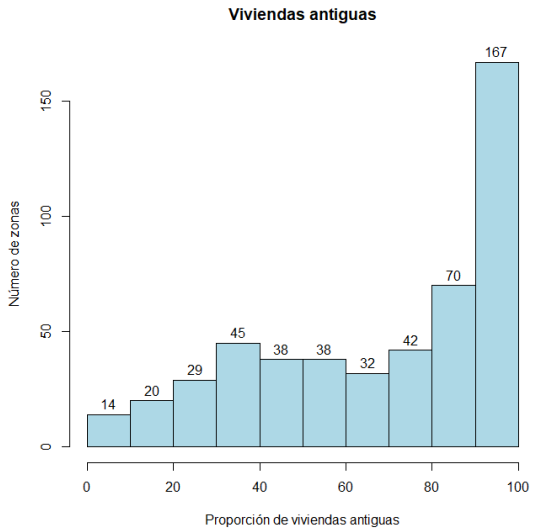


Ilustración 37. Diagrama de cajas de viviendas antiguas.

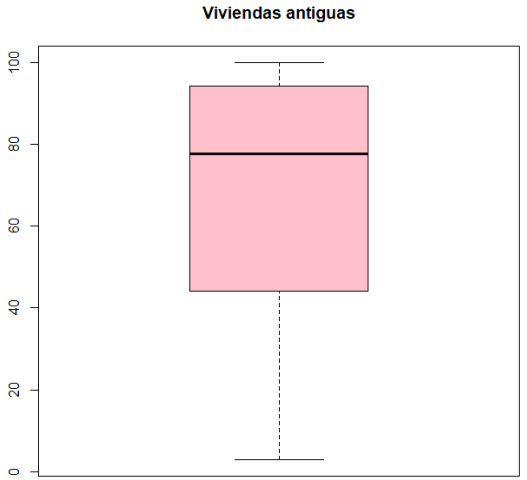


Ilustración 41. Histograma de distancia al centro.

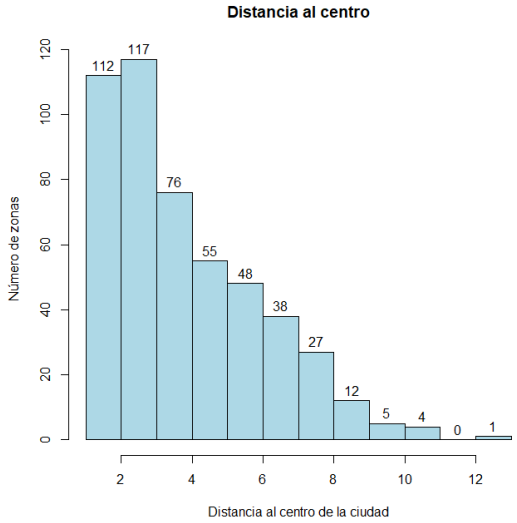


Ilustración 42. Diagrama de cajas de distancia centro.

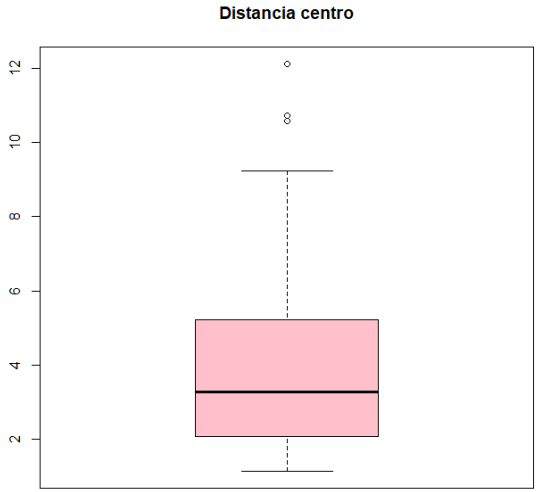


Ilustración 40. Histograma de acceso a autopista.

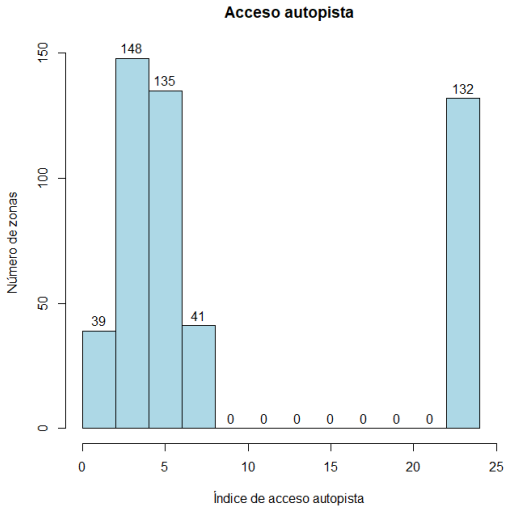


Ilustración 45. Diagrama de cajas de acceso autopista.

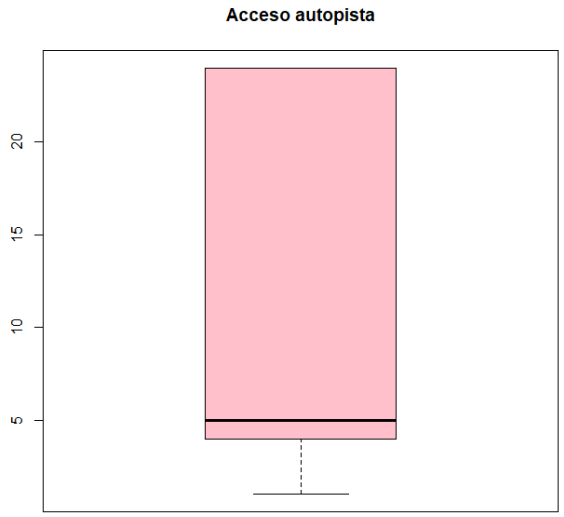


Ilustración 44. Histograma de impuestos.

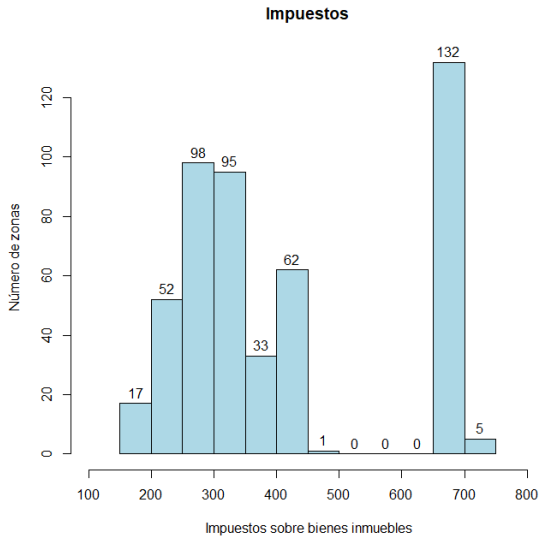


Ilustración 43. Diagrama de cajas de impuestos.

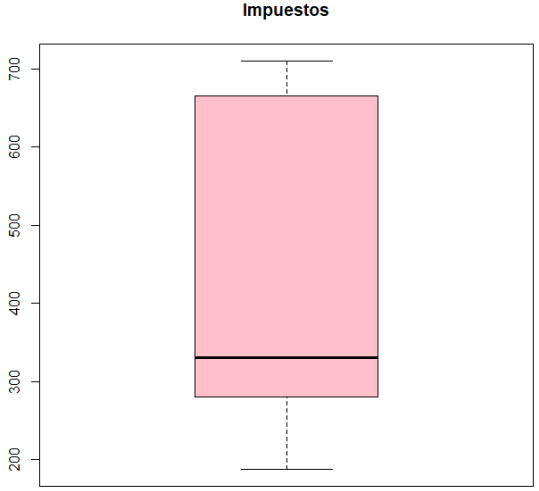


Ilustración 48. Diagrama de cajas de población negra.

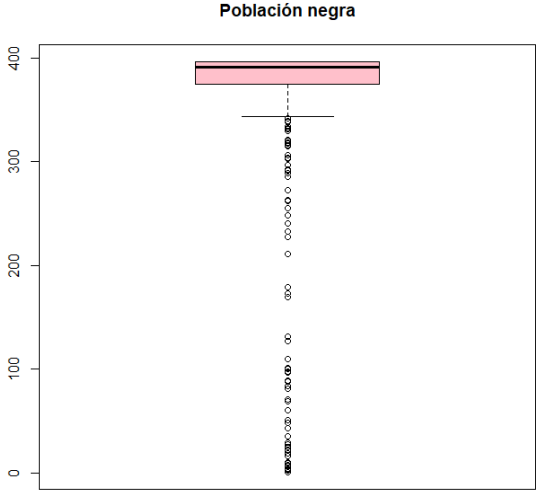


Ilustración 47. Diagrama de cajas de ratio alumnos.

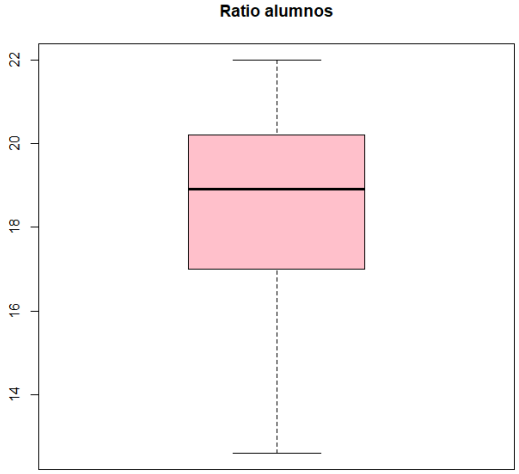


Ilustración 46. Histograma de población negra.

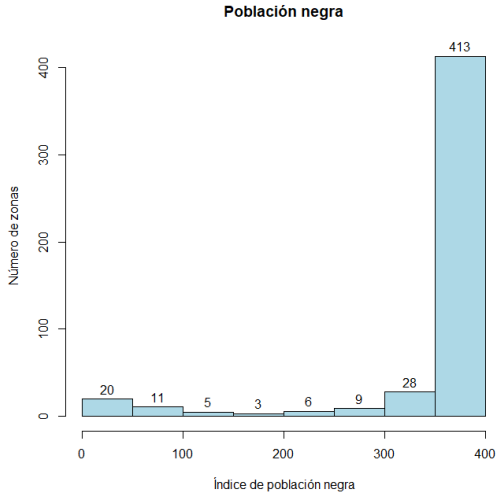


Ilustración 50. Diagrama de cajas de precio vivienda.

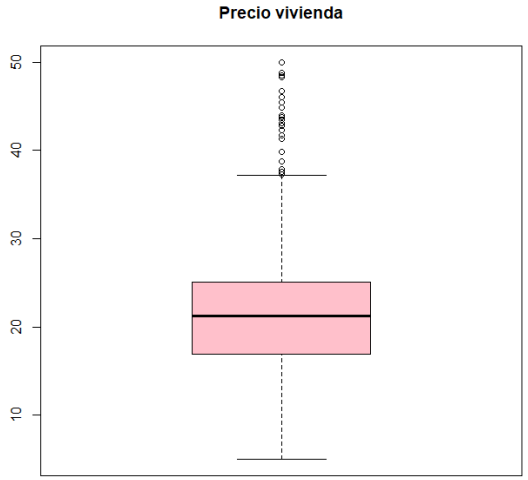


Ilustración 49. Diagrama de cajas de población pobre.

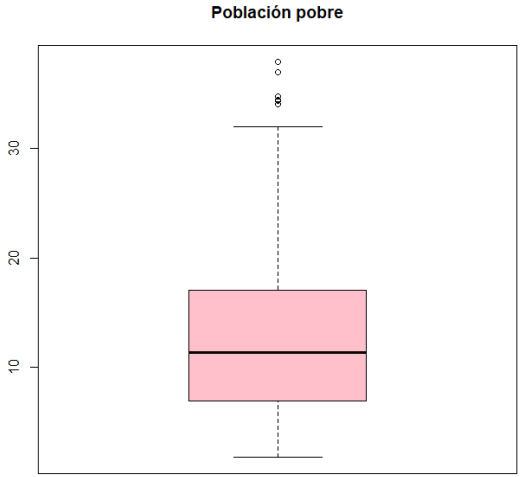


Ilustración 52. Histograma de población pobre.

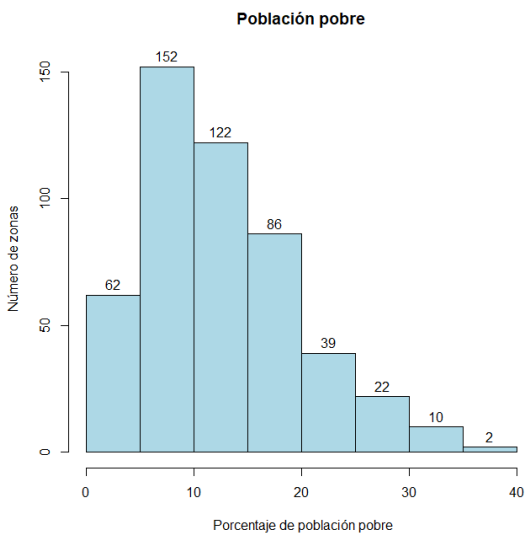


Ilustración 51. Histograma de precio de vivienda.