[](http://www.lamolina.edu.pe/)

**UNIVERSIDAD NACIONAL AGRARIA LA MOLINA**

*Escuela de Post Grado*

*Facultad de Economía y Planificación*

**NOTA**

*Departamento de Estadística e Informática*

**TÉCNICAS DE MINERÍA DE DATOS**

**Práctica Calificada 01**

**Apellidos y Nombres**:

GOMEZ MARIN, JAIME

GARCIA RODRIGUEZ, RENZO

RIVERA REYNA, GIANCARLOS

Cuestionario:

**Pregunta 1**

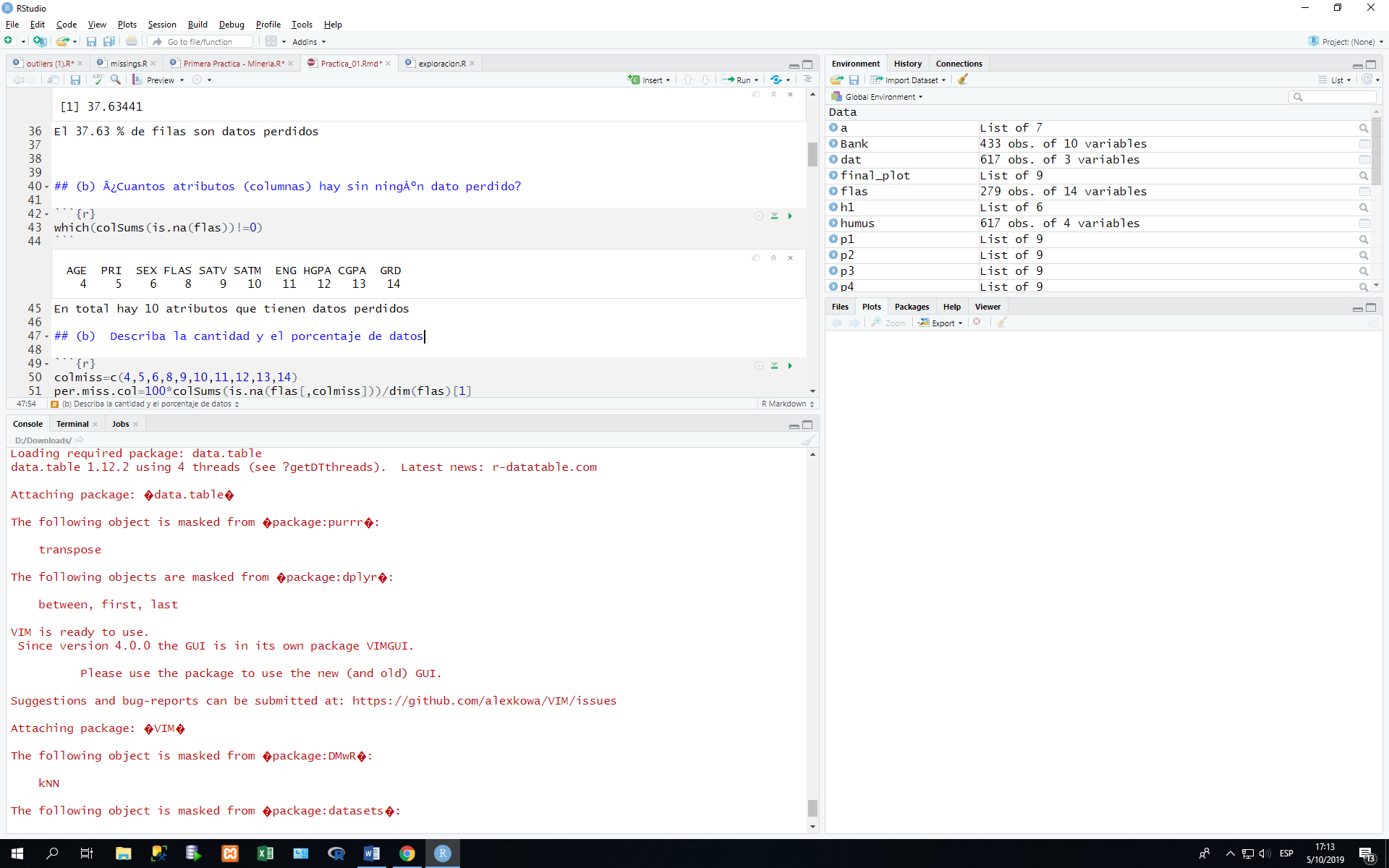
1. ¿Cuantas filas con valores perdidos hay en el conjunto de datos? ¿Qué porcentaje de todos los valores es?

* Tenemos 105 filas con valores perdidos y esto representa el 37.63% de casos.

|  |
| --- |
| Para ver la cantidad datos perdidos  fila.perdidas <- which(rowSums(is.na(flas))!=0)  length(fila.perdidas) |

|  |
| --- |
| (length(fila.perdidas)/dim(flas)[1]) \* 100 |

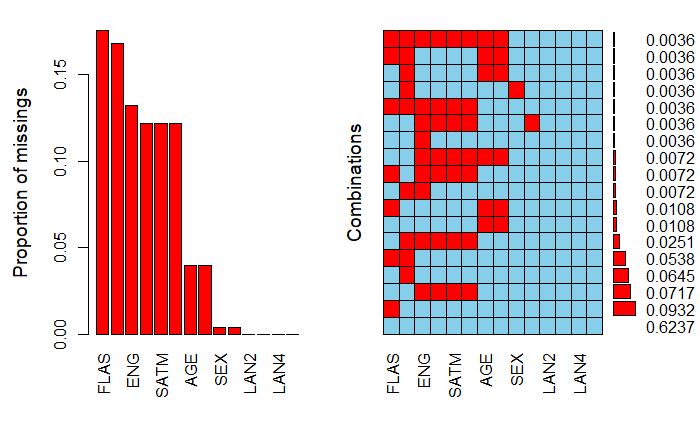
1. ¿Cuantos atributos (columnas) hay sin ningún dato perdido? Describa la cantidad y el porcentaje de datos perdidos que presenta cada atributo.



|  |
| --- |
| which(colSums(is.na(flas))!=0) |

* En total tenemos 10 atributos que tienen datos perdidos.
* Se ve que la columna AGE tiene 3.94%
* Se ve que la columna PRI tiene 3.94%
* Se ve que la columna SEX tiene 0.36%
* Se ve que la columna FLAS tiene 17.56%
* Se ve que la columna SATV tiene 12.19%
* Se ve que la columna SATM tiene 12.194%
* Se ve que la columna ENG tiene 13.264%
* Se ve que la columna HGPA tiene 0.36%
* Se ve que la columna CGPA tiene 12.194%
* Se ve que la columna GRD tiene 16.84%

(c) ¿Cuantos patrones de datos perdidos distintos presenta el conjunto de datos? ¿Cuál es el que ocurre con mayor frecuencia? Justifique su respuesta usando una gráfica apropiada.



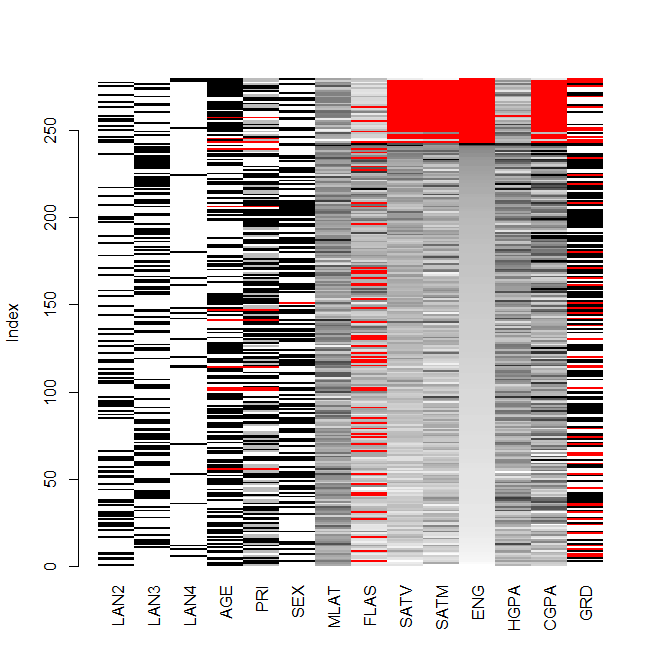
* Tengo un 18% que tiene datos faltantes para la variable FLAS
* Tengo un 17% que tiene datos faltantes para la variable GRD
* Tengo un 13% que tiene datos faltantes para la variable ENG
* Tengo un 12.5% que tiene datos faltantes para la variable SATV
* Tengo un 12.5% que tiene datos faltantes para la variable SATM
* Tengo un 12.5% que tiene datos faltantes para la variable CGPA
* Tengo un 4% que tiene datos faltantes para la variable AGRE
* Tengo un 4% que tiene datos faltantes para la variable PRI
* Tengo un 0.5% que tiene datos faltantes para la variable SEX
* Tengo un 0.5% que tiene datos faltantes para la variable HGPA

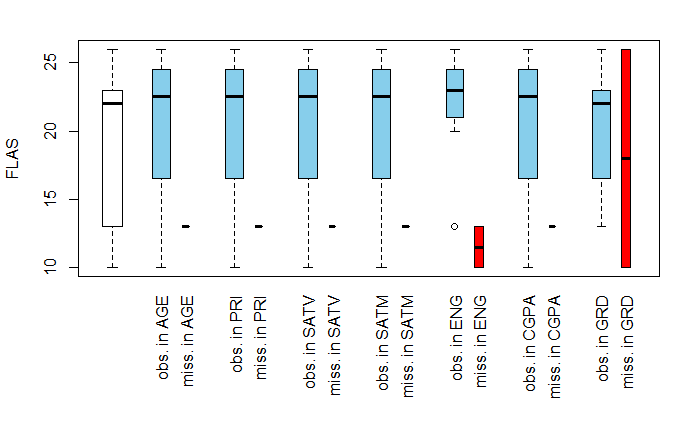
|  |
| --- |
| library(ranger)  library(laeken)  library(robustbase)  library(VIM)  # Aggregation plot  a=aggr(flas,numbers=T) |

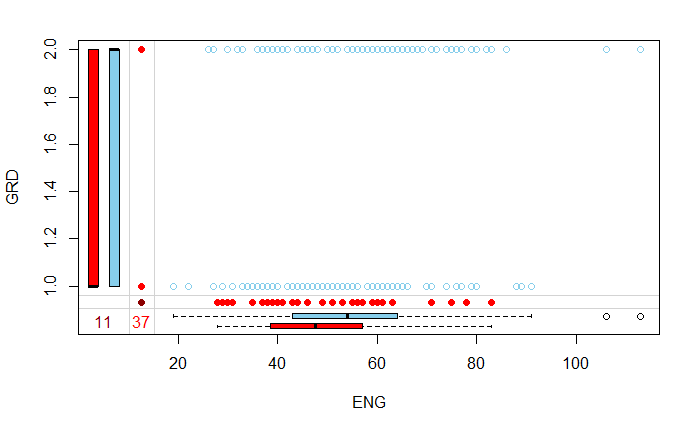
(d) ¿Es posible identificar algún patrón de datos perdidos? (por ejemplo, la perdida de información en alguna variable parece estar asociada a otra.) Justifique su respuesta

* Se observa que cuando las notas son bajas, se tiene datos perdidos en la matricula del curso de inglés, podría ser un mecanismo de MAR
* En cambio, hay datos perdidos en la matricula del idioma alemán sin importar la nota obtenida, esto podría ser un mecanismo MCAR

Cuando se pierde información de la variable Inglés también se pierden en las variable SATV y SATM y CGPA







|  |
| --- |
| marginplot(flas[,c("ENG", "GRD")]) |

(e) Se desea realizar la imputación de la variable FLAS. Identifique a qué tipo de mecanismo se puede atribuir la presencia de datos perdidos. Justifique su respuesta usando técnicas de visualización y la prueba estadística t. Luego, sobre la base del mecanismo identificado, realice la imputación de los datos perdidos.

Supuestos para hacer la prueba de medias

- Los valores deben ser obtenidos datos aleatorios

- Deben venir de una distribución normal

|  |
| --- |
| data: FLAS by is.na(ENG)  t = 5.3669, df = 53.534, p-value = 1.76e-06  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  3.090576 6.777752  sample estimates:  mean in group FALSE mean in group TRUE  24.96447 20.03030 |

|  |
| --- |
| data: FLAS by is.na(GRD)  t = 3.302, df = 36.651, p-value = 0.002148  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  1.545674 6.459473  sample estimates:  mean in group FALSE mean in group TRUE  24.76119 20.75862 |

|  |
| --- |
| # Prueba t de medias  t.test(FLAS ~ is.na(ENG), data=flas)  # Prueba t de medias  t.test(FLAS ~ is.na(GRD), data=flas) |

* Como el pvalor es casi cero en ambos casos, se tiene evidencia para descartar la hipótesis nula, por lo tanto, podemos asumir que la medias son distintas y se acepta el mecanismo MAR.
* Sobre los mecanismos identificados se procedió a realizar la imputación de los datos con el método de regresión lineal ya que los datos perdidos representan el 17%.

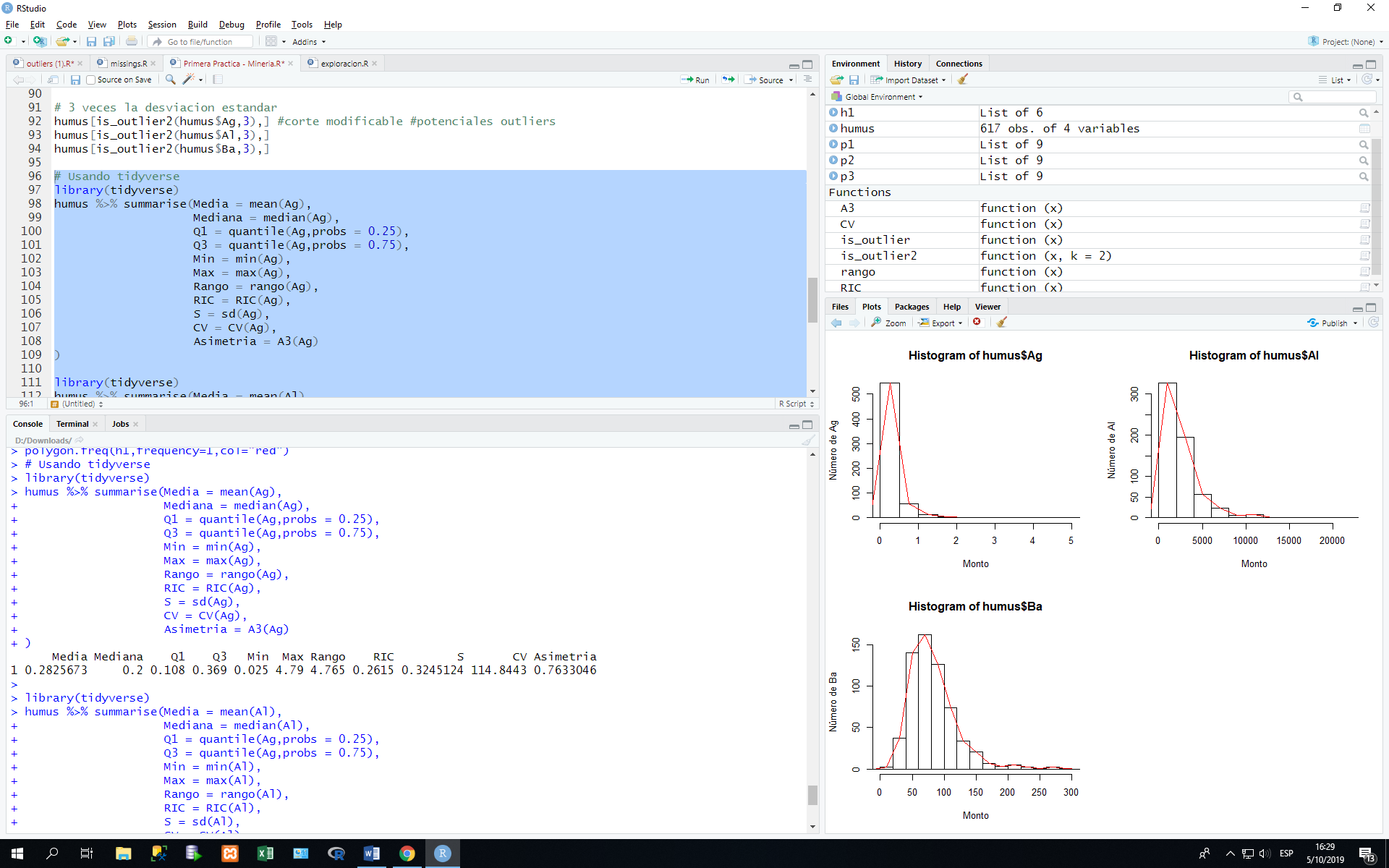
|  |
| --- |
| library(simputation)  flas.i <- impute\_lm(flas, ENG + GRD ~ 1)  flas.j <- impute\_lm(flas.i, FLAS ~ ENG + GRD, add\_residual = "normal")  flas[c(7,8,12),]$FLAS  flas.j[c(7,8,12),]$FLAS  [1] 21.18164 19.42445 29.17382 |

**Pregunta 2**

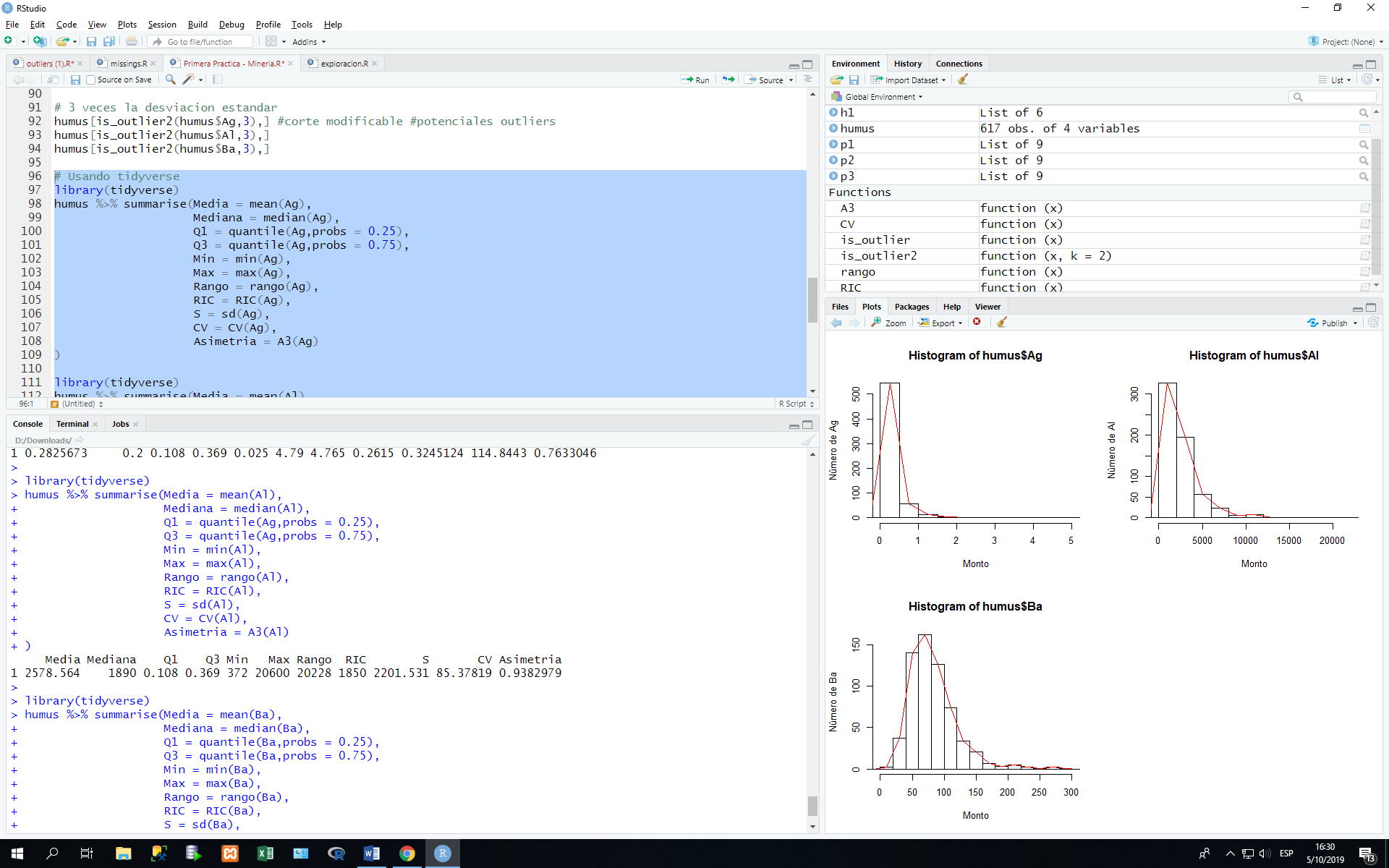
1. Sobre la base de un análisis univariado, ¿Existen evidencias de outliers en alguna de las variables? Justifique su análisis usando técnicas de visualización y la puntuación Z.

Se realizó un análisis exploratorio para identificar el tipo de distribución. Adicionalmente se calculó el estadístico de asimetría para identificar cuál de las 3 variables es más asimétrico.

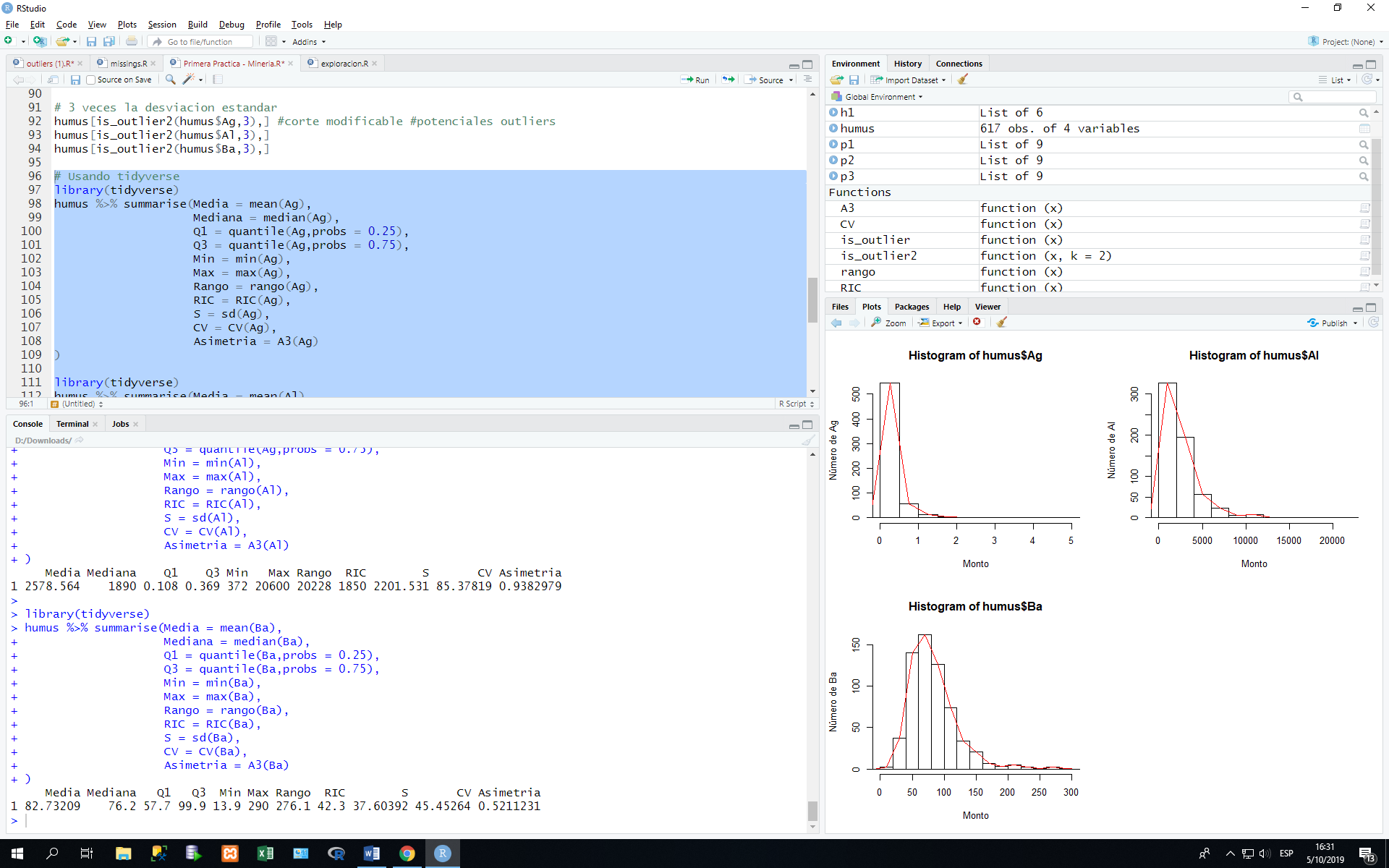
**Resumen: Variable Ag**



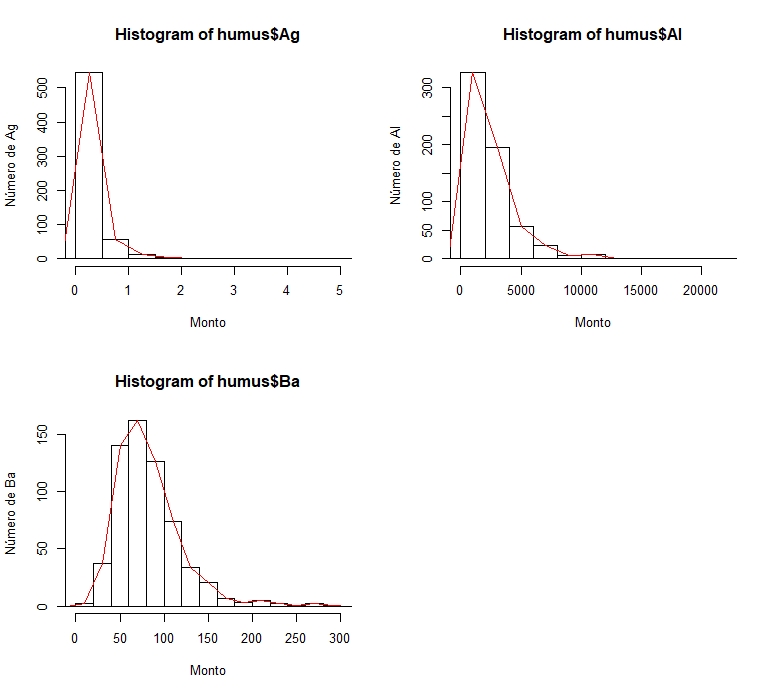
**Resumen: Variable Al**

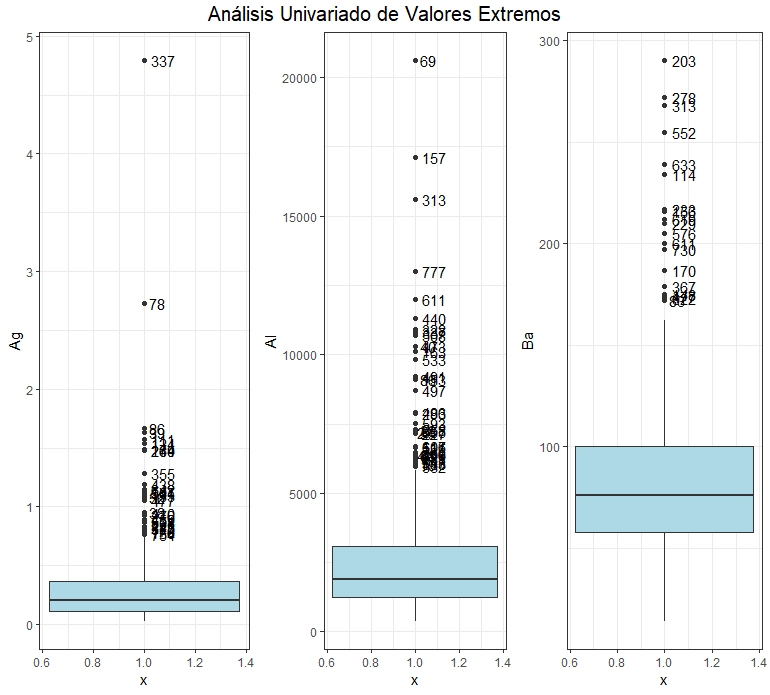


**Resumen: Variable Ba**



La variable Al es la que presenta mayor asimetría y la variable Ba es la que presenta menor asimetría. Se realizaron los gráficos de histograma correspondientes visualizando la distribución de cada variable.





Identificando Outliers con la Puntuación Z

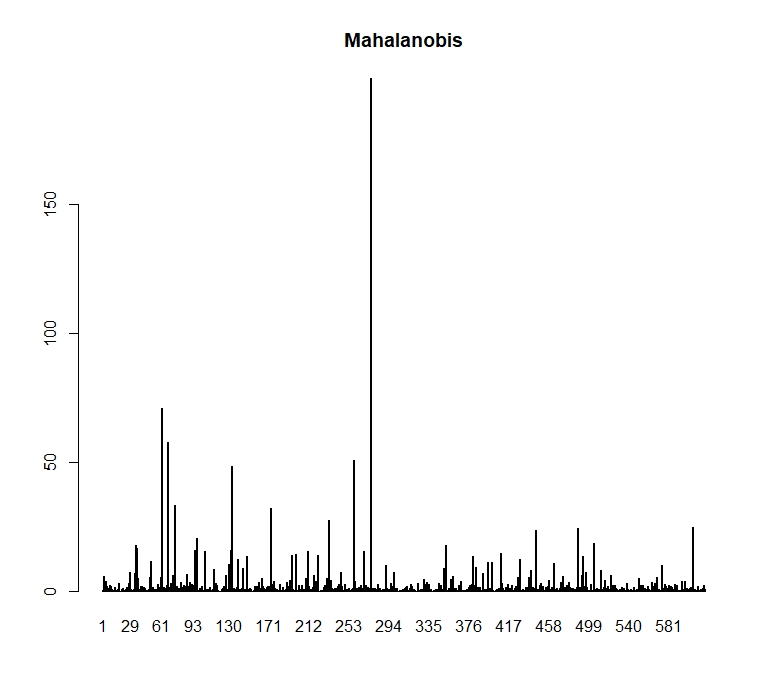
|  |
| --- |
| > humus[is\_outlier2(humus$Ag,4),]  ID Ag Al Ba  35 39 1.63 3640 140.0  67 78 2.73 4170 143.0  75 86 1.67 9090 20.8  275 337 4.79 10800 128.0 |

|  |
| --- |
| > humus[is\_outlier2(humus$Al,4),]  ID Ag Al Ba  61 69 0.261 20600 137  133 157 0.092 17100 109  258 313 0.249 15600 268  487 611 0.250 12000 200  605 777 0.129 13000 119 |

|  |
| --- |
| > humus[is\_outlier2(humus$Ba,2),]  ID Ag Al Ba  50 56 1.080 644 159  72 83 0.209 3010 172  97 114 0.037 1840 234  127 148 0.583 3320 175  132 156 0.106 1560 216  144 170 0.800 3460 187  164 192 0.242 1940 160  173 203 0.382 3700 290  194 229 0.137 1920 210  198 233 0.432 1910 217  232 278 0.456 2530 272  258 313 0.249 15600 268  299 367 0.475 2260 179  383 477 1.050 5050 174  426 531 0.140 2070 160  437 544 0.194 2090 161  444 552 0.214 2530 255  462 576 0.438 5710 205  487 611 0.250 12000 200  492 619 0.118 3110 212  495 622 0.127 1840 173  503 633 0.236 3890 239  568 724 0.160 2450 162  573 730 0.179 3390 197 |

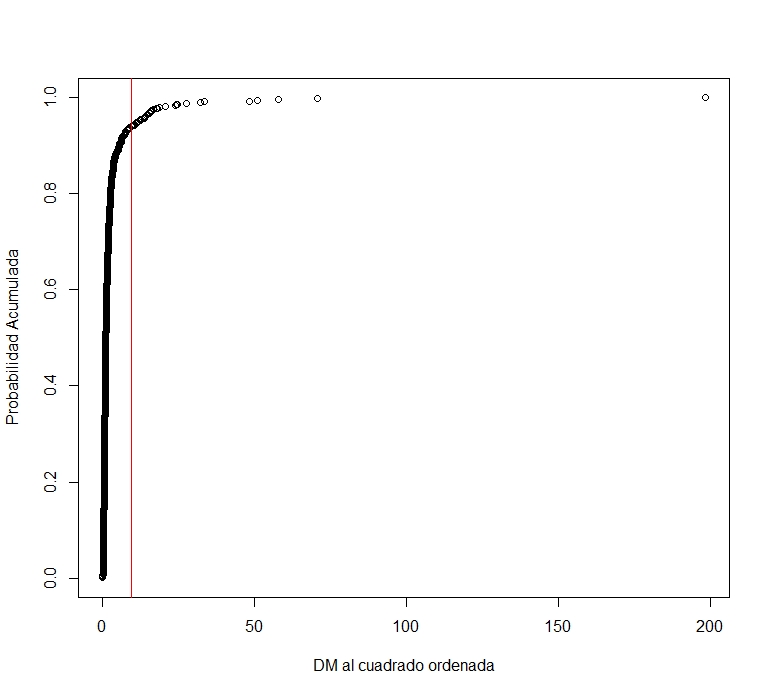
(b) Evalué la presencia de outliers multivariados usando la distancia de Mahalanobis al cuadrado. Justifique su análisis usando técnicas de visualización (grafica de las distancias, grafica Q-Q y de frecuencias acumuladas usando la distribución chi-cuadrado)

Distancia de Mahalanobis - Se visualizar el elemento 275 como posible outliers.



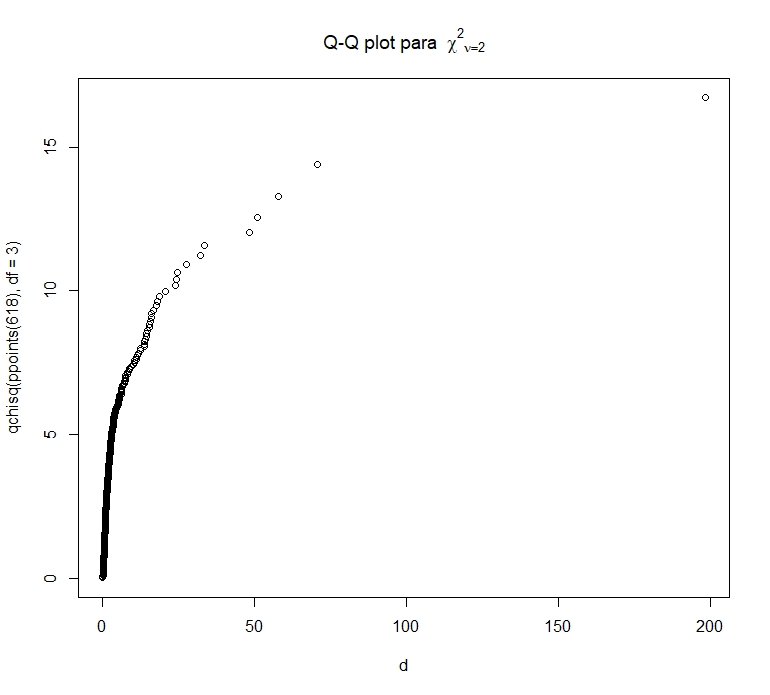
Frecuencias Acumulados

|  |
| --- |
| #ojiva  plot(sort(d), ppoints(617), xlab ="DM al cuadrado ordenada", ylab = "Probabilidad Acumulada")  Q <- qchisq(p = 0.975, df = 3)  abline(v=Q, col = "red") |



Q-Q Plot

|  |
| --- |
| # QQ-plot:  qqplot(d, qchisq(ppoints(617), df = 3),  main = expression("Q-Q plot para" ~~ {chi^2}[nu == 2])) |



(c) Evalué la presencia de outliers multivariados usando el método basado en densidad local (LOF)

considerando 10 vecinos. Compare los resultados obtenidos con el método anterior.

**Evaluando presencia de outliers multivariados**

**Resultados de Outliers con Método de Densidad Local**

|  |
| --- |
| ID Ag Al Ba  61 69 0.261 20600 137  133 157 0.092 17100 109  132 156 0.106 1560 216  50 56 1.080 644 159  232 278 0.456 2530 272  258 313 0.249 15600 268  444 552 0.214 2530 255  97 114 0.037 1840 234  495 622 0.127 1840 173  198 233 0.432 1910 217 |

|  |
| --- |
| ### Calculo los Outliers con el Método de Densidad Local  library(DMwR)  lof=lofactor(dat,10)  lof  humus[order(lof,decreasing=T)[1:10],] |