Análisis Univariado de valores atípicos

En este análisis se utilizará el conjunto de datos llamado `Glass`, que se encuentra en la librería `mlbench` de R, y que proviene del UC Irvine Machine Learning Repository. Este conjunto de datos de 214 observaciones contiene muestras del análisis químico realizado a siete tipos diferentes de vidrio. Este análisis químico incluye el índice refractivo (RI) y los porcentajes de ocho elementos (Na, Mg, Al, Si, K, Ca, Ba, Fe), teniendo en total nueve atributos.

Utilizando un análisis univariado de valores atípicos, se determinará si existe evidencia de valores atípicos ("outliers") en alguno de los nueve atributos.

1° VALORES PERDIDOS

Analizamos la existencia de valores perdidos, encontrando que en Mg, K, Ba y Fe hay NAs o vacíos.

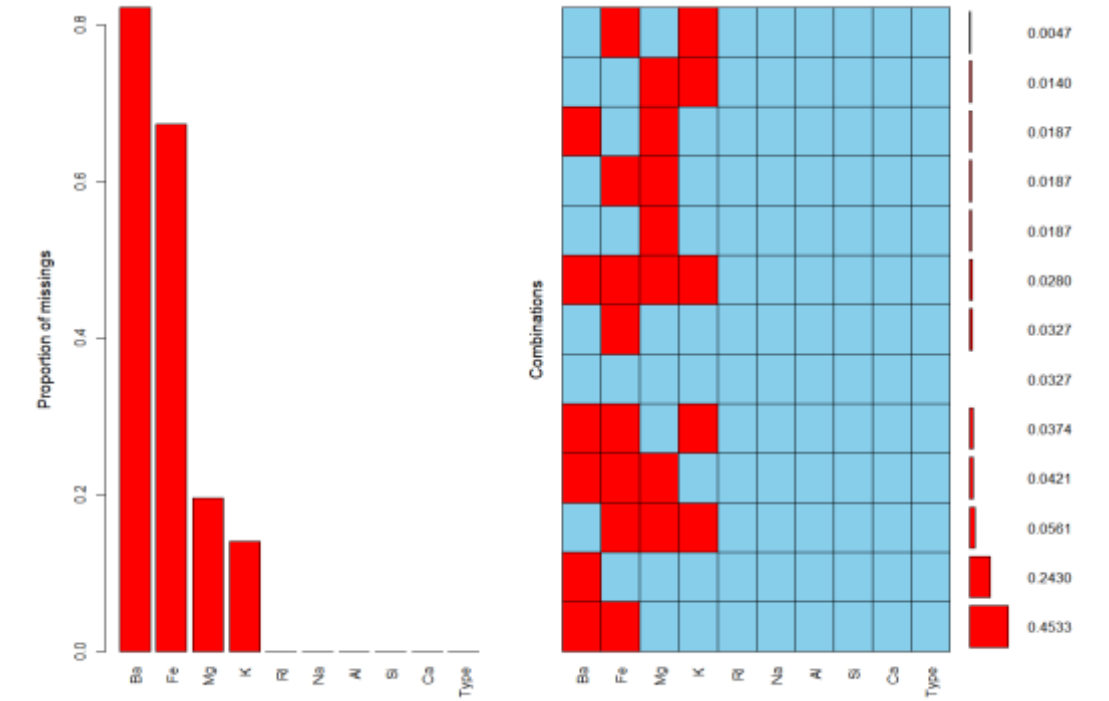
```
> sum(is.na(Glass))
[1] 392
> colmiss
Mg  K Ba Fe
 3  6  8  9
> colSums(is.na(Glass))# Muestra qué columnas tienen valores perdidos
  RI   Na   Mg   Al   Si    K   Ca   Ba   Fe Type
   0    0   42    0    0   30    0  176  144    0
```

Al evaluar el porcentaje de los nulos en las columnas, vemos que el porcentaje es mayor del 14% y en algunas columnas más del 67%, teniendo un impacto importante con las demás variables por lo que no se pueden eliminar estos datos y se les debe imputar.

```
> # Muestra el porcentaje de valores perdidos en las columnas
> per.miss.col=100*colSums(is.na(Glass[,colmiss]))/dim(Glass)[1]
> per.miss.col
      Mg        K       Ba       Fe
19.62617 14.01869 82.24299 67.28972
```

Visualizando gráficamente, vemos la relación de los nulos. Se ve que la relación de los nulos Ba y Fe es de 45%, mientras que hay 24% solo en función de Ba, luego hay relaciones menores al 5%.

```
Missings in combinations of variables:
       Combinations Count    Percent
0:0:0:0:0:0:0:0:0:0     7  3.2710280
0:0:0:0:0:0:0:0:1:0     7  3.2710280
0:0:0:0:0:0:0:1:0:0    52 24.2990654
0:0:0:0:0:0:0:1:1:0    97 45.3271028
0:0:0:0:0:1:0:0:1:0     1  0.4672897
0:0:0:0:0:1:0:1:1:0     8  3.7383178
0:0:1:0:0:0:0:0:0:0     4  1.8691589
0:0:1:0:0:0:0:0:1:0     4  1.8691589
0:0:1:0:0:0:0:1:0:0     4  1.8691589
0:0:1:0:0:0:0:1:1:0     9  4.2056075
0:0:1:0:0:1:0:0:0:0     3  1.4018692
0:0:1:0:0:1:0:0:1:0    12  5.6074766
0:0:1:0:0:1:0:1:1:0     6  2.8037383
```



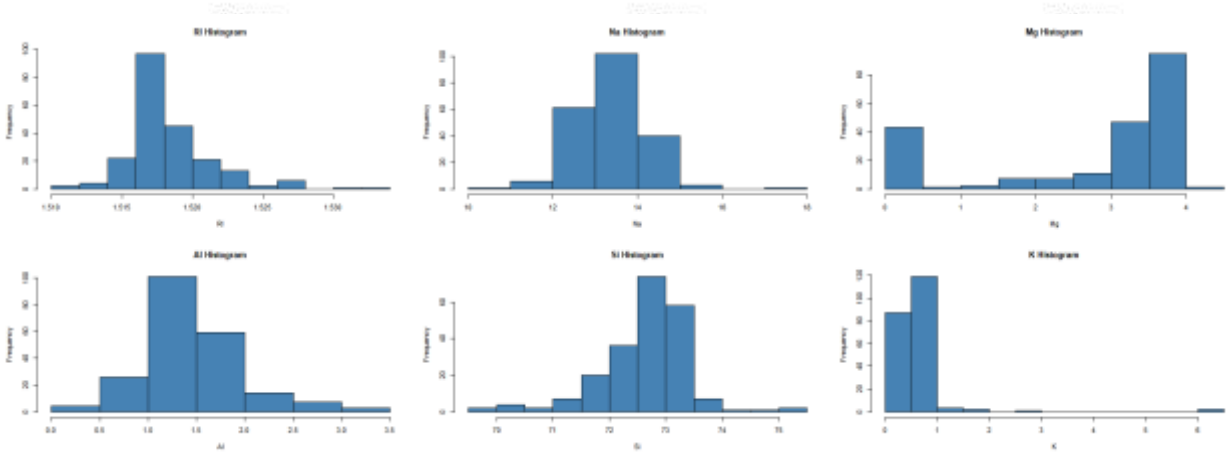
Imputamos los valores por la mediana

```
# Usando una medida de Tendencia Central
#----------------------------------------
library(DMwR)
Glass.c <- initialise(Glass,method="median")
summary(Glass.c)
colSums(is.na(Glass.c))# Muestra qué columnas tienen valores perdidos
```

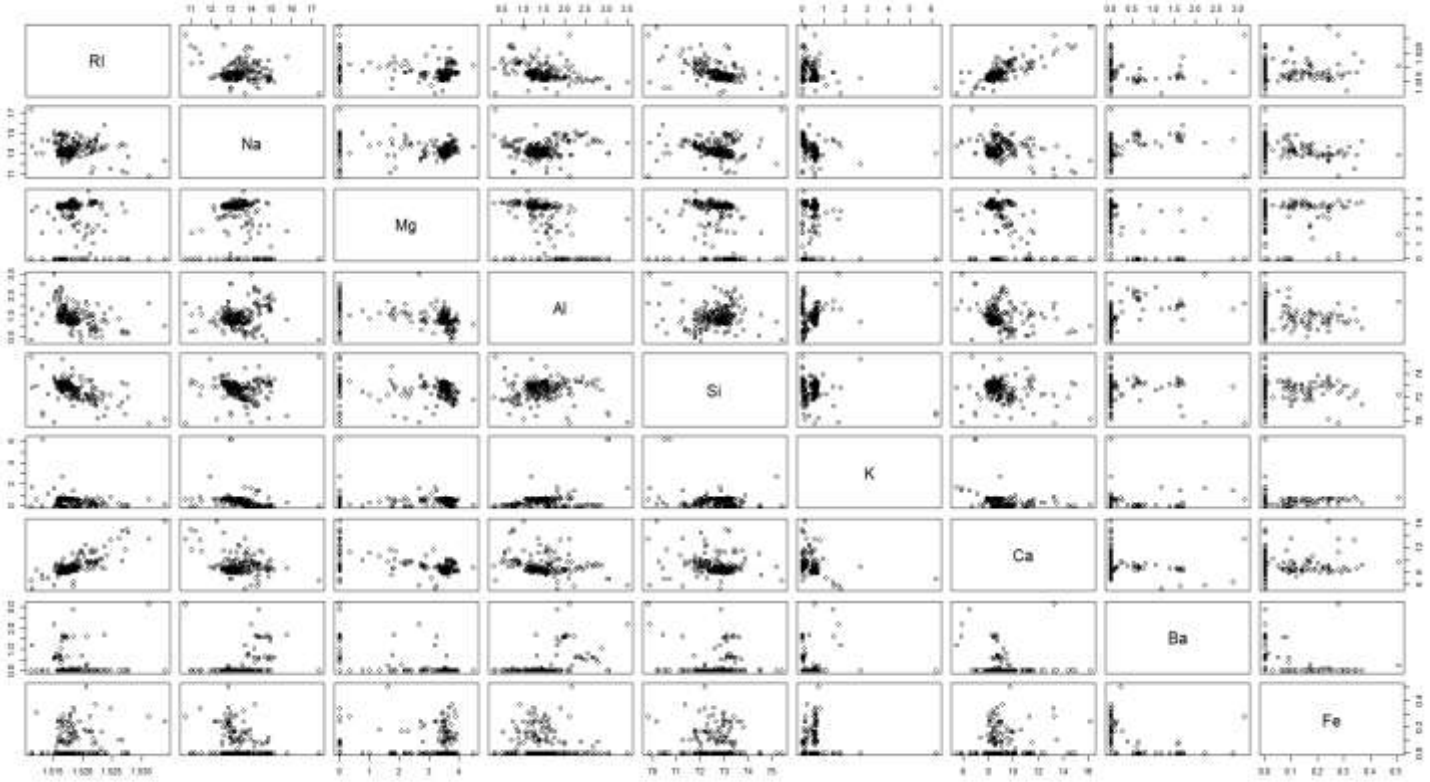
```
> sum(is.na(Glass.c))
[1] 0
> colSums(is.na(Glass.c))# Muestra qué columnas tienen valores perdidos
  RI   Na   Mg   Al   Si    K   Ca   Ba   Fe Type
   0    0    0    0    0    0    0    0    0    0
```

2° VALORES OUTLIERS

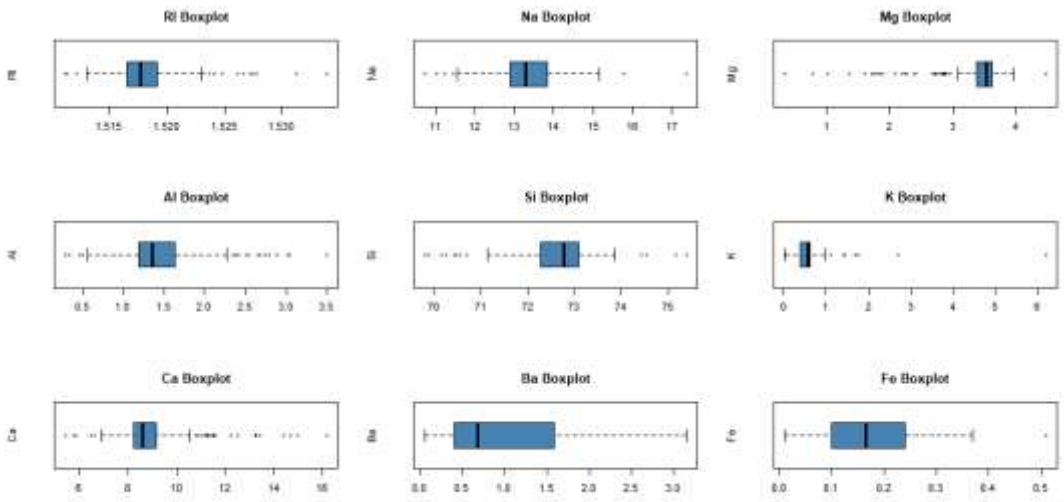
Mediante la visualización de histogramas vemos que RI, Na, Al y Si tienen un comportamiento relativamente normal(simétrica).



En la gráfica de dispersión por pares, no parecen demostrar correlaciones fuertes, pero se observa una relación positiva entre RI y Ca, y una relación negativa entre RI y Si.



Mediante el diagrama de cajas se observa que todas las variables aparentemente muestran valores atípicos a excepción de Ba. Se observa también sesgo en las variables.



Mediante la regla de Turkey hallamos los valores Outliers

```
Browse[1]> Glass.c$RI[is_outlier(Glass.c$RI)]
 [1] 1.52667 1.52320 1.51215 1.52725 1.52410 1.52475 1.53125 1.53393 1.52664 1.52739 1.52777
[12] 1.52614 1.52369 1.51115 1.51131 1.52315 1.52365
Browse[1]> Glass.c$Na[is_outlier(Glass.c$Na)]
[1] 11.45 10.73 11.23 11.02 11.03 17.38 15.79
Browse[1]> Glass.c$Mg[is_outlier(Glass.c$Mg)]
 [1] 4.49 3.85 3.84 2.87 2.84 2.81 2.71 3.86 3.87 3.09 2.88 2.96 2.85 2.72 2.76 3.15 2.90
[18] 3.97 3.89 3.90 2.28 2.09 1.35 1.01 3.98 3.93 3.85 3.90 3.18 3.90 2.68 1.85 1.88 1.71
[35] 1.61 0.33 2.39 2.41 2.24 2.19 1.74 0.78 3.20 2.20 1.83 1.78
Browse[1]> Glass.c$Al[is_outlier(Glass.c$Al)]
 [1] 0.29 0.47 0.47 0.51 3.50 3.04 3.02 0.34 2.38 2.79 2.68 2.54 2.34 2.66 2.51 2.42 2.74
[18] 2.88
Browse[1]> Glass.c$Si[is_outlier(Glass.c$Si)]
 [1] 70.57 69.81 70.16 74.45 69.89 70.48 70.70 74.55 75.41 70.26 70.43 75.18
Browse[1]> Glass.c$K[is_outlier(Glass.c$K)]
 [1] 0.06 0.15 0.06 0.03 0.11 0.11 0.17 0.02 0.13 0.09 0.14 0.12 0.23 0.18 0.09 0.19 0.19
[18] 0.23 0.12 0.16 1.10 0.07 0.08 0.08 0.12 0.10 0.06 0.19 0.11 0.06 0.11 0.16 0.23 1.68
[35] 0.97 6.21 6.21 0.13 1.76 1.46 0.04 0.08 0.14 0.04 0.05 2.70 1.41 0.08
Browse[1]> Glass.c$Ca[is_outlier(Glass.c$Ca)]
 [1] 11.64 10.79 13.24 13.30 16.19 11.52 10.99 14.68 14.96 14.40 11.14 13.44  5.87 11.41
[15] 11.62 11.53 11.32 12.24 12.50 11.27 10.88 11.22  6.65  5.43  5.79  6.47
Browse[1]> Glass.c$Ba[is_outlier(Glass.c$Ba)]
 [1] 0.09 0.11 0.69 0.14 0.11 3.15 0.27 0.09 0.06 0.15 2.20 0.24 1.19 1.63 1.68 0.76 0.64
[18] 0.40 1.59 1.57 0.61 0.81 0.66 0.64 0.53 0.63 0.56 1.71 0.67 1.55 1.38 2.88 0.54 1.06
[35] 1.59 1.64 1.57 1.67
Browse[1]> Glass.c$Fe[is_outlier(Glass.c$Fe)]
 [1] 0.26 0.11 0.24 0.24 0.17 0.07 0.19 0.14 0.22 0.06 0.30 0.16 0.10 0.16 0.11 0.09 0.24
[18] 0.31 0.11 0.11 0.07 0.17 0.17 0.16 0.03 0.12 0.32 0.14 0.09 0.10 0.09 0.22 0.19 0.15
[35] 0.24 0.22 0.20 0.34 0.28 0.24 0.08 0.14 0.10 0.29 0.21 0.12 0.17 0.17 0.18 0.10 0.15
[52] 0.28 0.12 0.17 0.25 0.24 0.35 0.10 0.17 0.09 0.24 0.37 0.51 0.28 0.09 0.09 0.08 0.07
[69] 0.05 0.01
```

Con la puntuación Z, se genera una distribución normal de media 0 y desviación estándar de 2, de este modo vemos que hace un ajuste para solo extraer los datos bien extremos o que no se relacionen con los demás valores de la variable considerando como valores atípicos los valores por encima de 2 desviaciones estándar.

```
Browse[1]> Glass.c[is_outlier_z(Glass.c$RI,3),]
          RI    Na    Mg   Al    Si    K    Ca   Ba    Fe Type
107 1.53125 10.73 3.535 2.10 69.81 0.58 13.30 3.15 0.280    2
108 1.53393 12.30 3.535 1.00 70.16 0.12 16.19 0.68 0.240    2
113 1.52777 12.64 3.535 0.67 72.02 0.06 14.40 0.68 0.165    2
Browse[1]> Glass.c[is_outlier_z(Glass.c$Na,3),]
          RI    Na    Mg   Al    Si    K    Ca   Ba    Fe Type
107 1.53125 10.73 3.535 2.10 69.81 0.58 13.30 3.15 0.280    2
185 1.51115 17.38 3.535 0.34 75.41 0.57  6.65 0.68 0.165    6
Browse[1]> Glass.c[is_outlier_z(Glass.c$Mg,3),]
          RI    Na   Mg   Al    Si    K    Ca   Ba    Fe Type
130 1.52020 13.98 1.35 1.63 71.76 0.39 10.56 0.68 0.180    2
131 1.52177 13.75 1.01 1.36 72.19 0.33 11.14 0.68 0.165    2
175 1.52058 12.85 1.61 2.17 72.18 0.76  9.70 0.24 0.510    5
176 1.52119 12.97 0.33 1.51 73.39 0.13 11.27 0.68 0.280    5
182 1.51888 14.99 0.78 1.74 72.50 0.57  9.95 0.68 0.165    6
Browse[1]> Glass.c[is_outlier_z(Glass.c$Al,3),]
          RI    Na    Mg   Al    Si    K   Ca   Ba    Fe Type
164 1.51514 14.01 2.680 3.50 69.89 1.68 5.87 2.20 0.165    5
172 1.51316 13.02 3.535 3.04 70.48 6.21 6.96 0.68 0.165    5
173 1.51321 13.00 3.535 3.02 70.70 6.21 6.93 0.68 0.165    5
Browse[1]> Glass.c[is_outlier_z(Glass.c$Si,3),]
          RI    Na    Mg   Al    Si    K    Ca   Ba    Fe Type
107 1.53125 10.73 3.535 2.10 69.81 0.58 13.30 3.15 0.280    2
108 1.53393 12.30 3.535 1.00 70.16 0.12 16.19 0.68 0.240    2
164 1.51514 14.01 2.680 3.50 69.89 1.68  5.87 2.20 0.165    5
185 1.51115 17.38 3.535 0.34 75.41 0.57  6.65 0.68 0.165    6
189 1.52247 14.86 2.200 2.06 70.26 0.76  9.76 0.68 0.165    7
202 1.51653 11.95 3.535 1.19 75.18 2.70  8.93 0.68 0.165    7
```

```
Browse[1]> Glass.c[is_outlier_z(Glass.c$K,3),]
          RI    Na    Mg   Al    Si    K   Ca   Ba    Fe Type
172 1.51316 13.02 3.535 3.04 70.48 6.21 6.96 0.68 0.165    5
173 1.51321 13.00 3.535 3.02 70.70 6.21 6.93 0.68 0.165    5
202 1.51653 11.95 3.535 1.19 75.18 2.70 8.93 0.68 0.165    7
Browse[1]> Glass.c[is_outlier_z(Glass.c$Ca,3),]
          RI    Na    Mg   Al    Si    K    Ca   Ba    Fe Type
106 1.52475 11.45 3.535 1.88 72.19 0.81 13.24 0.68 0.340    2
107 1.53125 10.73 3.535 2.10 69.81 0.58 13.30 3.15 0.280    2
108 1.53393 12.30 3.535 1.00 70.16 0.12 16.19 0.68 0.240    2
111 1.52664 11.23 3.535 0.77 73.21 0.57 14.68 0.68 0.165    2
112 1.52739 11.02 3.535 0.75 73.08 0.57 14.96 0.68 0.165    2
113 1.52777 12.64 3.535 0.67 72.02 0.06 14.40 0.68 0.165    2
132 1.52614 13.70 3.535 1.36 71.24 0.19 13.44 0.68 0.100    2
Browse[1]> Glass.c[is_outlier_z(Glass.c$Ba,3),]
          RI    Na    Mg   Al    Si    K    Ca   Ba    Fe Type
107 1.53125 10.73 3.535 2.10 69.81 0.58 13.30 3.15 0.280    2
164 1.51514 14.01 2.680 3.50 69.89 1.68  5.87 2.20 0.165    5
208 1.51831 14.39 3.535 1.82 72.86 1.41  6.47 2.88 0.165    7
Browse[1]> Glass.c[is_outlier_z(Glass.c$Fe,3),]
          RI    Na    Mg   Al    Si    K    Ca   Ba   Fe Type
106 1.52475 11.45 3.535 1.88 72.19 0.81 13.24 0.68 0.34    2
146 1.51839 12.85 3.670 1.24 72.57 0.62  8.68 0.68 0.35    2
163 1.52211 14.19 3.780 0.91 71.36 0.23  9.14 0.68 0.37    3
175 1.52058 12.85 1.610 2.17 72.18 0.76  9.70 0.24 0.51    5
```