Anomalias_C1

Miguel Merelo Hernández

February 11, 2018

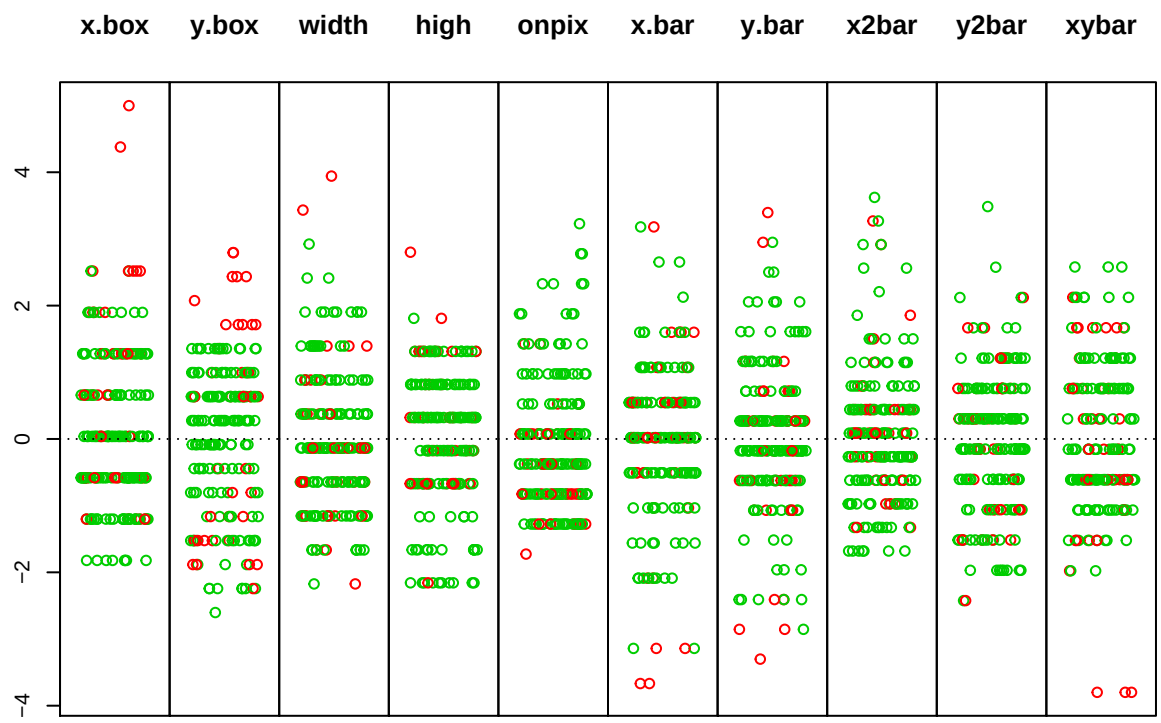
Crear datos

Utilizamos LetterRecognition del paquete mlbench. Solo utilizamos las primeras 200 líneas ya que la visualización de 20000 entradas hace imposible analizar el problema.

```
library(mlbench)
data("LetterRecognition")
mydata.numeric = LetterRecognition[1:200,-c(1)]
mydata.numeric.scaled = scale(mydata.numeric)
```

1. Obtención de los outliers multivariantes

```
alpha.value = 0.05
alpha.value.penalizado = 1 - ( 1 - alpha.value) ^ (1 / nrow(mydata.numeric))
set.seed(12)
#solo con las 10 primeras variables, uni.plot no permite mas
mvoutlier.plot<-uni.plot(mydata.numeric[1:10],symb=FALSE,alpha=alpha.value.penalizado)
```



En el gráfico tenemos representados valores escalados de las variables con los que son outliers multivariantes.

2. Análisis de los outliers

```
is.MCD.outlier<-mvoutlier.plot$outliers
numero.de.outliers.MCD<-sum(is.MCD.outlier)
numero.de.outliers.MCD
```

```
## [1] 53
```

En nuestro conjunto de datos tenemos 53 outliers multivariantes.

```
indices.de.outliers.en.alguna.columna<-
  vector_claves_outliers_IQR_en_alguna_columna(mydata.numeric)
indices.de.outliers.en.alguna.columna<-
  indices.de.outliers.en.alguna.columna[!duplicated(indices.de.outliers.en.alguna.columna)]
indices.de.outliers.multivariantes.MCD<-which(is.MCD.outlier)
indices.de.outliers.multivariantes.MCD.pero.no.1variantes<-
  setdiff(indices.de.outliers.multivariantes.MCD,indices.de.outliers.en.alguna.columna)
nombres.de.outliers.multivariantes.MCD.pero.no.1variantes<-
  names(is.MCD.outlier[indices.de.outliers.multivariantes.MCD.pero.no.1variantes])
indices.de.outliers.multivariantes.MCD.pero.no.1variantes
```

```
##  [1]   4   9  12  16  18  45  49  56  59  63  70  71  82 108 123 127 132
## [18] 152 156 179 184 189 190 200
```

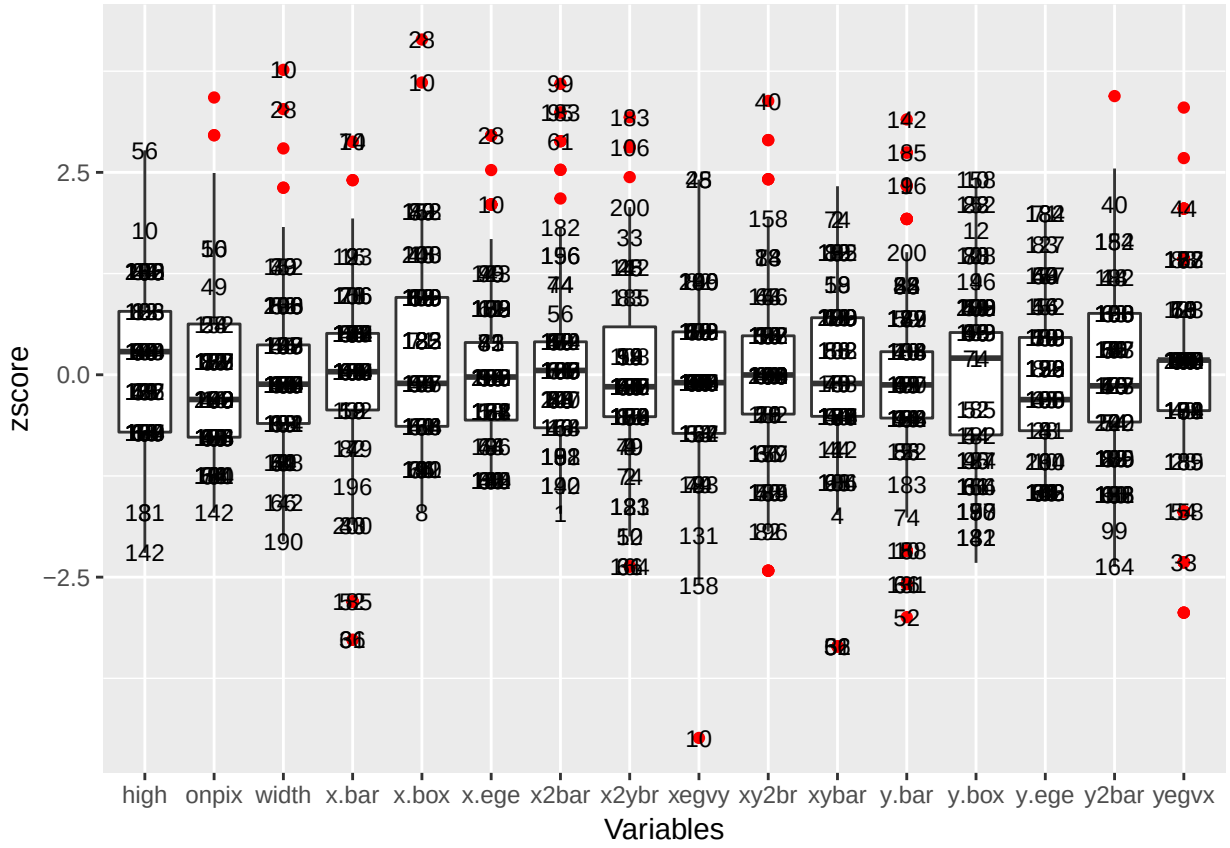
Indices de los outliers que son únicamente multivariantes.

```
data.frame.solo.outliers<-mydata.numeric.scaled[is.MCD.outlier,]
head(data.frame.solo.outliers)
```

```
##        x.box      y.box      width       high      onpix      x.bar
## 1 -1.1680108  0.2054706 -1.0873763 -0.2113252 -1.2380388  0.5110552
## 2  0.4247312  1.4699052 -1.0873763  0.7831463 -0.7717341  1.4574537
## 4  1.4865592  1.1537966  0.3689313  0.2859105 -0.3054295 -0.9085426
## 7 -0.1061828 -1.6911813 -0.1165046 -0.7085609  0.1608751  0.5110552
## 8 -1.6989248 -2.0072899 -1.0873763 -1.7030323 -1.2380388  0.5110552
## 9 -1.1680108 -1.6911813 -0.6019405 -0.7085609 -0.7717341  1.4574537
##        y.bar       x2bar      y2bar      xybar      x2ybr        xy2br
## 1  2.3343823 -1.71716256  0.3109080 -0.9197064  1.3318240 -0.002417871
## 2 -0.9448202  0.05310812 -0.5837913  1.9226533 -1.2578338  0.481156231
## 4  0.6947810 -0.30094601  0.3109080 -1.7318092 -0.8878827  0.964730333
## 7 -0.1250196  0.40716226  0.3109080 -0.5136550 -0.1479804 -0.969566075
## 8 -2.1745212 -1.00905429 -1.4784905 -0.1076036 -1.6277849 -0.002417871
## 9 -0.5349199 -1.00905429  0.3109080  1.5166019 -0.8878827 -0.002417871
##        x.ege       xegvy       y.ege      yegvx
## 1 -1.3059334 -0.09718502 -1.45895955  0.1809413
## 2 -0.4537746 -0.09718502  0.07678734  1.4288120
## 4  1.2505431  1.15681519 -0.69108610  0.1809413
## 7 -0.4537746 -0.09718502  1.22859751  1.4288120
## 8 -0.8798540 -1.35118522 -0.69108610 -0.4429941
## 9 -0.8798540 -1.35118522 -1.07502282 -0.4429941
```

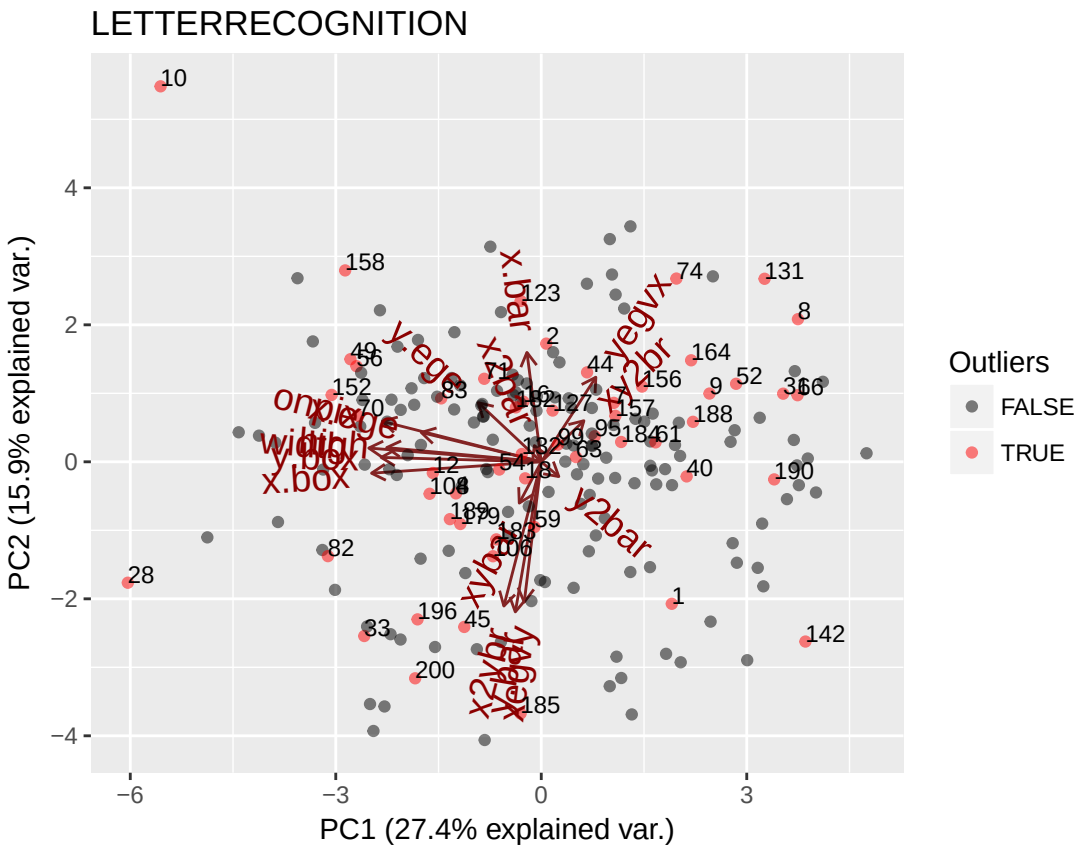
Valores normalizados de las instancias con outliers.

```
set.seed(12)
MiBoxPlot_juntos(mydata.numeric,is.MCD.outlier)
```



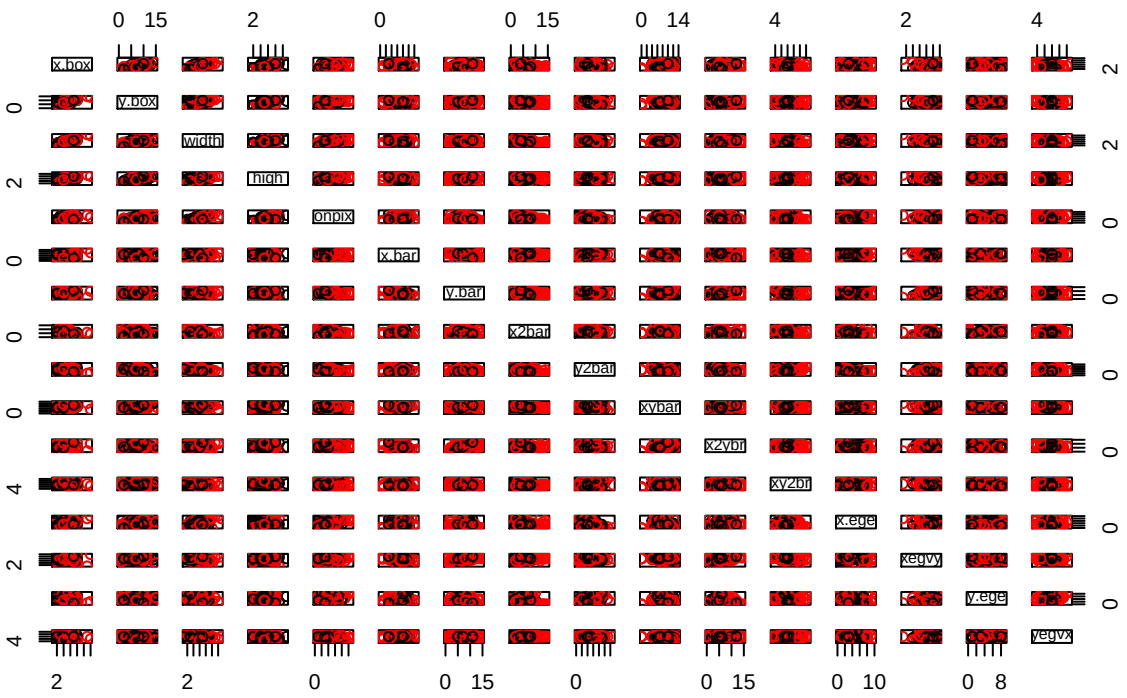
Podemos ver como se muestran en rojo los outliers simples y los multivariables. Por ejemplo, en x.ege podemos ver algunos entre el primer y segundo cuartil.

```
set.seed(12)
MiBiPlot_Multivariate_Outliers(mydata.numeric,is.MCD.outlier,"LETTERRECOGNITION")
```



En este gráfico se colorean en rojo aquellos valores que son outliers multivariantes.

```
set.seed(12)
MiPlot_Univariate_Outliers(mydata.numeric,
                           indices.de.outliers.en.alguna.columna,
                           "LETTERRECOGNITION")
```



Mostramos las variables dos a dos y marcamos aquellos puntos pertenecientes a instancias que tengan algún outlier en alguna variable. Por la cantidad de instancias con outliers que tenemos, casi el 50%, es imposible sacar una conclusión de este gráfico.