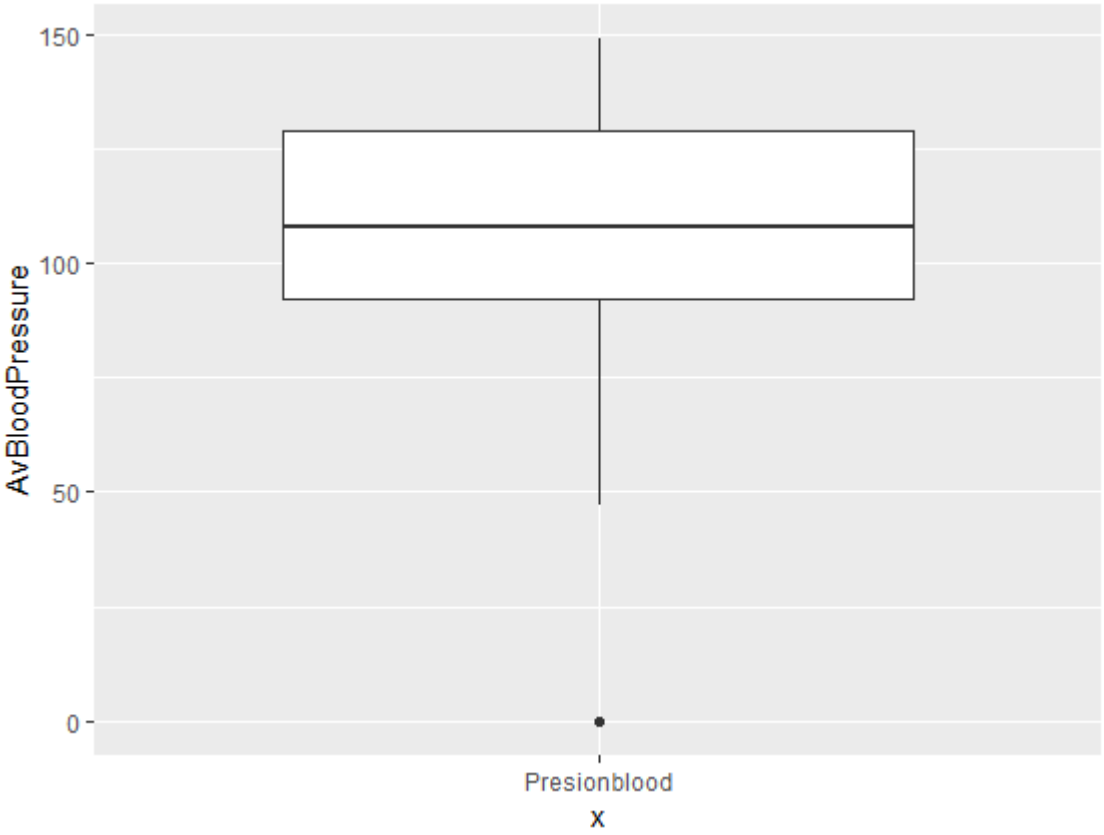
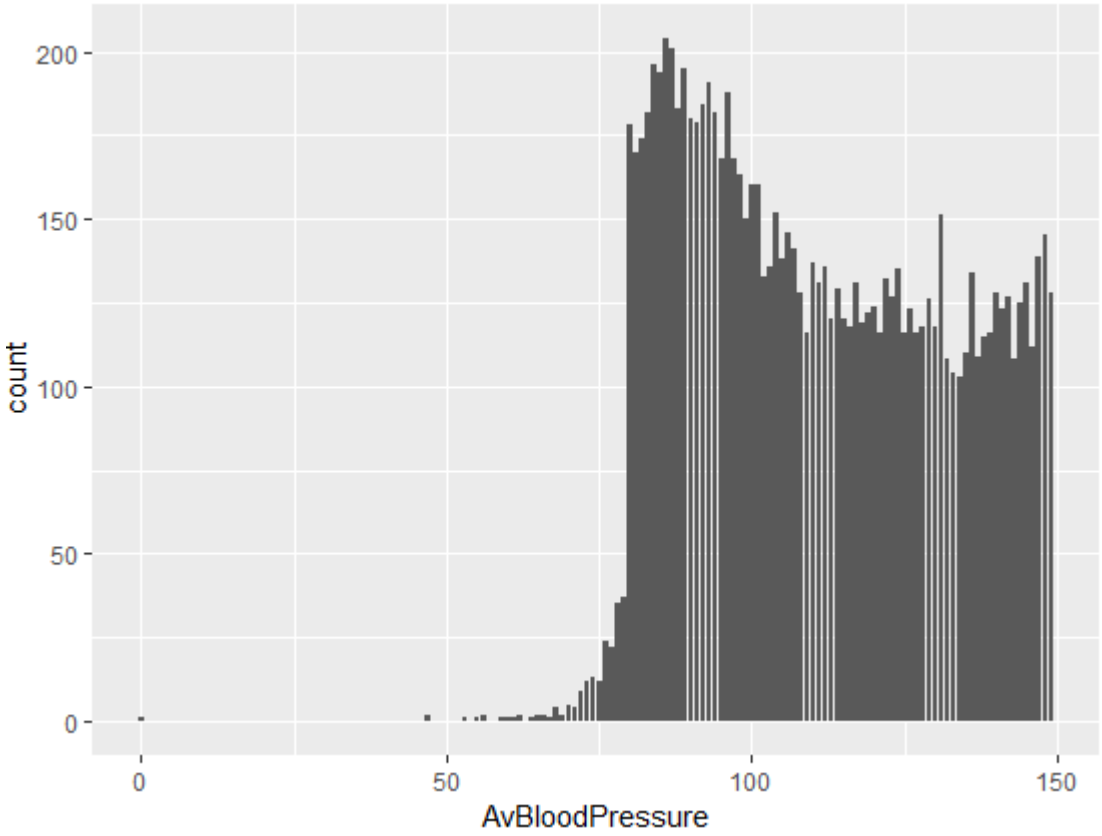
Cristiam Camilo Lopez Ruiz

Enero 2019

1. Presión sanguínea se observa un outlier que tiene presión sanguínea en 0, se realiza validación de este valor para identificar cuantos registros cuentan con este error.
   En este caso solo 1 paciente presenta este inconveniente

```
Patient AvBloodPressure
1      81               0
Se procederá a colocar la media de la presión sanguínea, si bien
un solo registro no afecta a nuestro set de datos se colocará para
homogenizar nuestros datos.
```





Se observa una distribución sesgada a la derecha

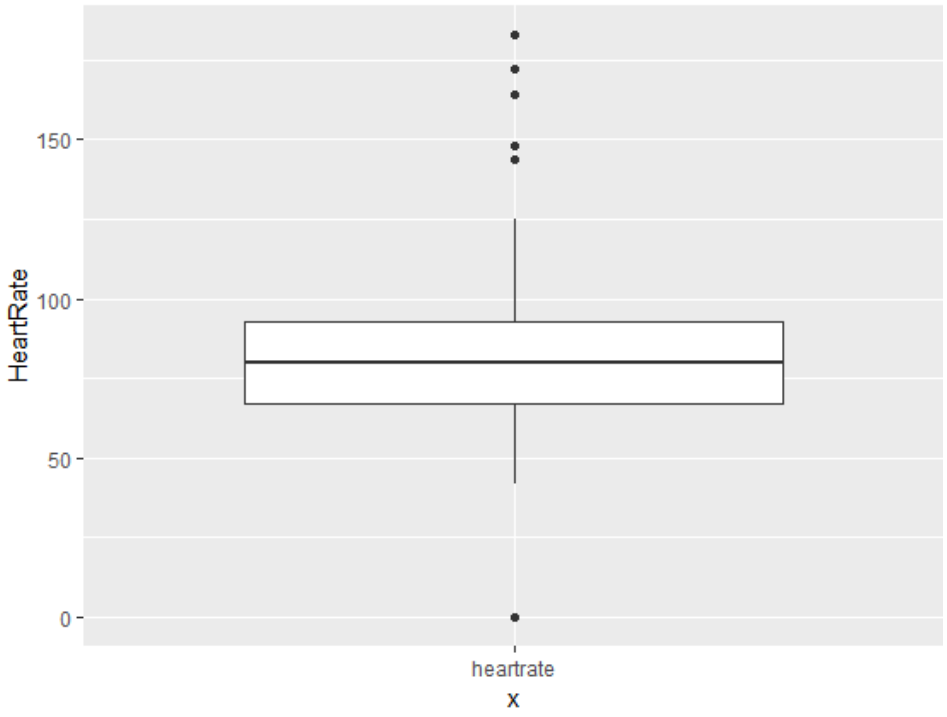
Latidos del corazón.

Se observan outliers donde se puede observar que tenemos pulsaciones que no se encuentran dentro de los índices normales, se observan a continuación la cantidad de registros de que se encuentran fuera de los estándares.

En este caso son 7 registros que contienen latidos fuera de lo normal 3 de ellos hacen parte de la misma persona, en este caso no será necesario reemplazar la data de estas 7 personas ya que puede afectar al modelo, es probable que tengamos algunas personas que en el momento de la medición

Patient	HeartRate
1	41	164
2	64	183
3	81	144
4	81	148
5	81	172
6	8	172
7	17	125

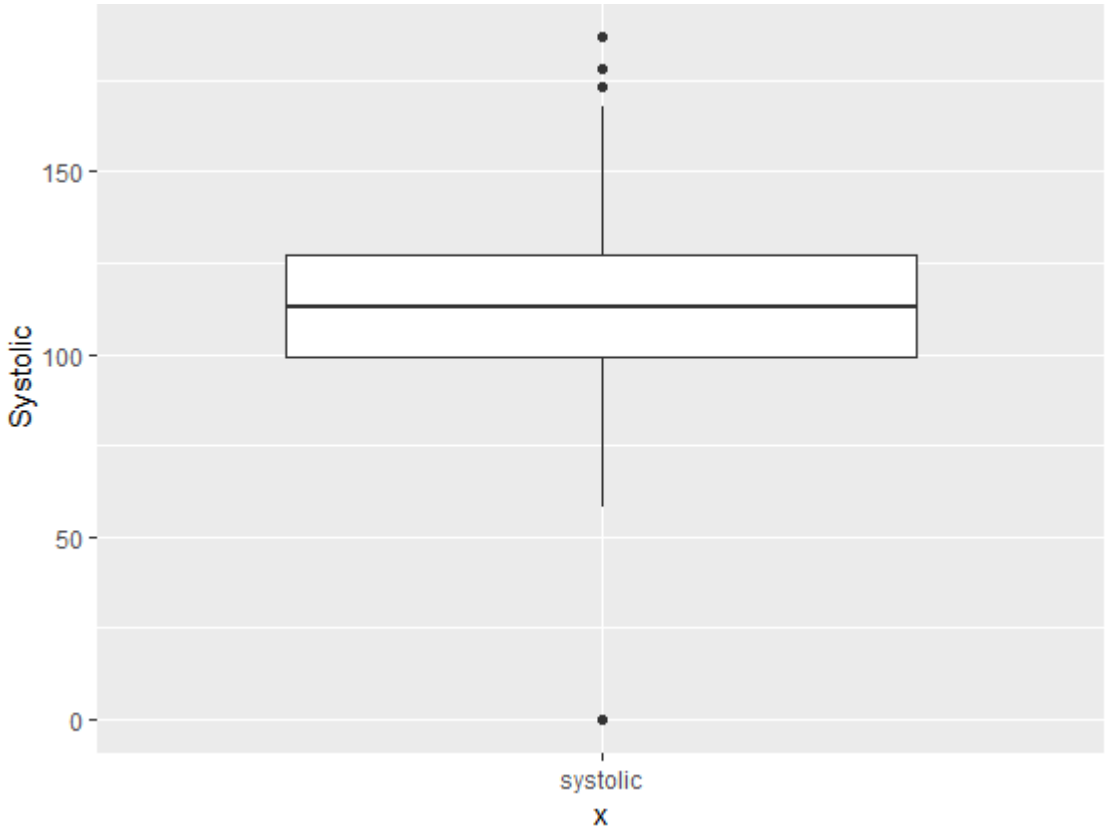
Sin embargo, el valor mínimo 0 debe ajustarse con el promedio de los datos ya que dicho valor puede generarnos ruido.

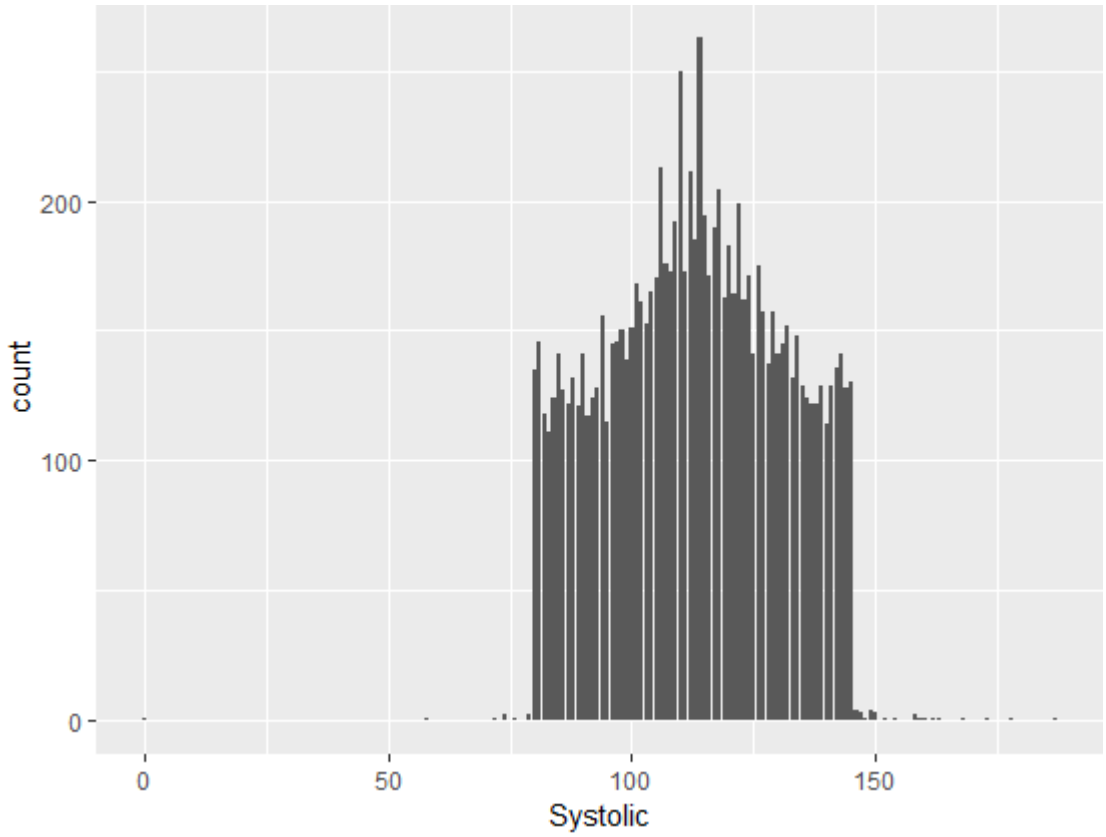


Sistole

Para esta variable se observa que existe varios outliers, para este caso tenemos 19 registros identificados como outliers ya que no se encuentran dentro del promedio y máximo de los datos.

	Patient	Systolic
1	125	149
2	35	168
3	188	158
4	81	159
5	231	150
6	81	187
7	81	162
8	81	178
9	8	150
10	259	150
11	105	154
12	267	149
13	105	149
14	105	160
15	288	152
16	296	149
17	34	158
18	8	173
19	81	163



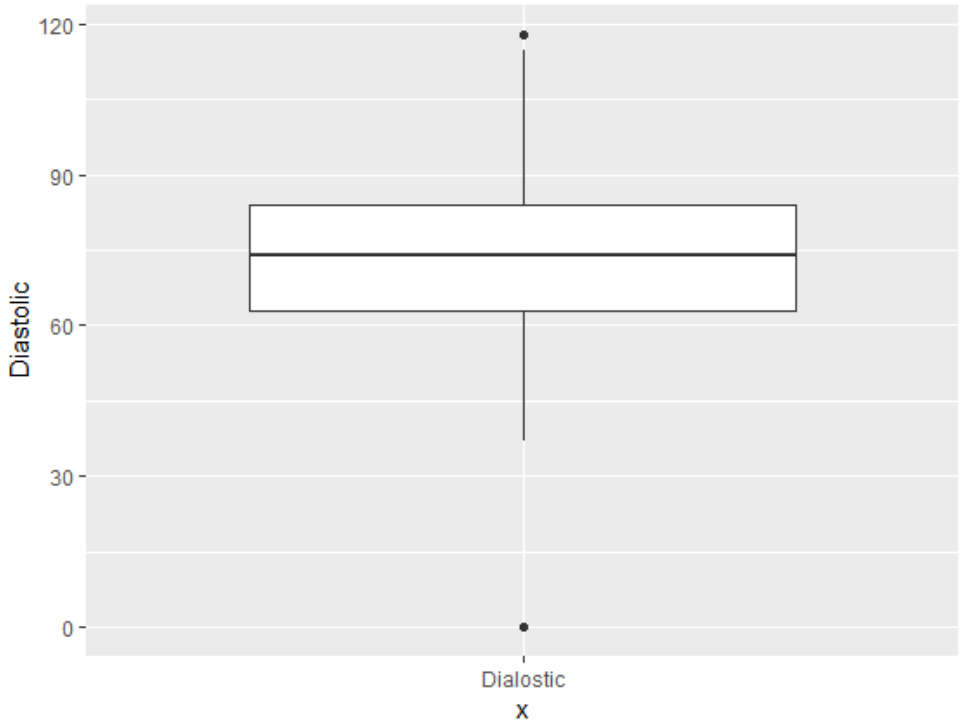


Se observa que la distribución es normal y se observan los distintos outliers entre ellos el valor cero el cual parecer ser una constante del primer registro de este dataset.

Dialostic

Para esta variable se observa un comportamiento normal con dos outliers fuera de medidas normales ubicados en 118 y un outlier con valor 0.

```
+   filter(Diastolic >= 115)
  Patient Diastolic
1     125       118
2      81       118
3     105       115
```



Análisis conjunto glucosa.

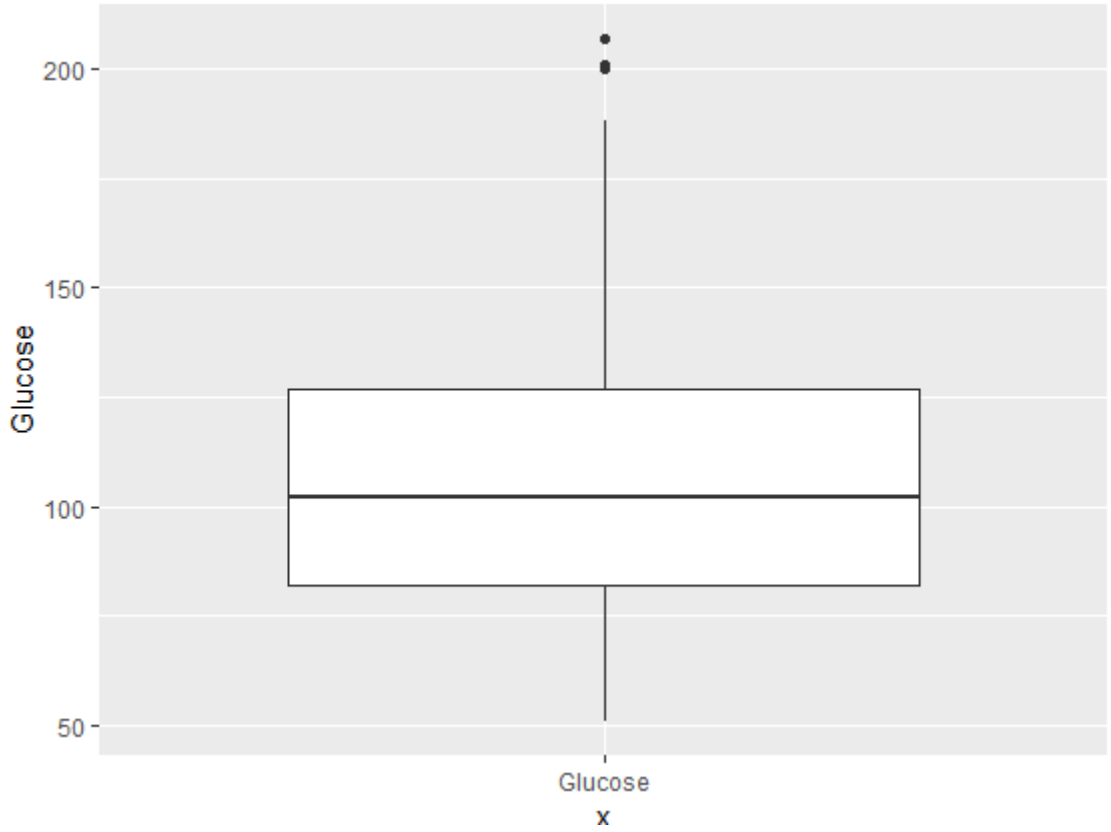
Se observan 10.046 observaciones y solamente 3 variables.

```
'data.frame':  10074 obs. of  3 variables:
 $ Patient: int  1369 1410 1156 663 1198 740 574 787 623 1116 ...
 $ Glucose: int  61 82 89 68 135 105 130 81 138 137 ...
 $ Date   : Factor w/ 9163 levels "1/1/2013 10:41",..: 28 29 30 31 32 33 34 35 36 37 ..
```

```
> summary(glucose)
    Patient          Glucose                  Date
 Min.   :   1.0   Min.   : 51.0   2/1/2015 20:55:    5
 1st Qu.: 426.0   1st Qu.: 82.0   2/1/2015 9:55 :    5
 Median : 736.0   Median :102.0   2/5/2015 14:58:    5
 Mean   : 726.2   Mean   :105.6   2/9/2015 10:49:    5
 3rd Qu.:1076.0   3rd Qu.:127.0   2/1/2015 12:45:    4
 Max.   :1453.0   Max.   :207.0   2/1/2015 19:25:    4
                                  (Other)       :10046
```

Glucosa

Al realizar el análisis de glucosa se observa que existen valores por encima de su medida máxima en este caso son 160 registros que cuentan con medidas mayores a la máxima de los datos, esto se obtiene con la formula Q3+1.5*IQR que en este caso tiene como resultado valor de 194.5 aquellos valores por encima de este valor se consideraran outlier, para este caso estos datos corresponden al 0.015% de los datos es decir un 1.5, dicha medida no afectaría nuestra medición así que se procede a retirar estos registros.



Análisis fuente Oximetría

Para este dataset se evidencia una estructura como se muestra a continuación:

#'data.frame': 9872 obs. of 4 variables:

$ Patient : int 1369 1410 1156 663 1198 740 574 787 623 1116 ...

#$ SpO2 : int 96 73 85 66 88 86 67 63 78 70 ...

#$ HeartRate: int 73 61 57 87 75 62 73 103 106 71 ...

#$ Date : Factor w/ 8984 levels "1/1/2013 10:41",..: 28 29 30 31 32 33 34 35 36 37 ...

En su resumen los datos muestran las siguientes medidas:

#Patient SpO2 HeartRate Date

#Min. : 1.0 Min. : 44.00 Min. : 44.00 2/1/2015 20:55: 5

#1st Qu.: 438.0 1st Qu.: 71.00 1st Qu.: 66.00 2/1/2015 9:55 : 5

#Median : 749.0 Median : 83.00 Median : 80.00 2/5/2015 14:58: 5

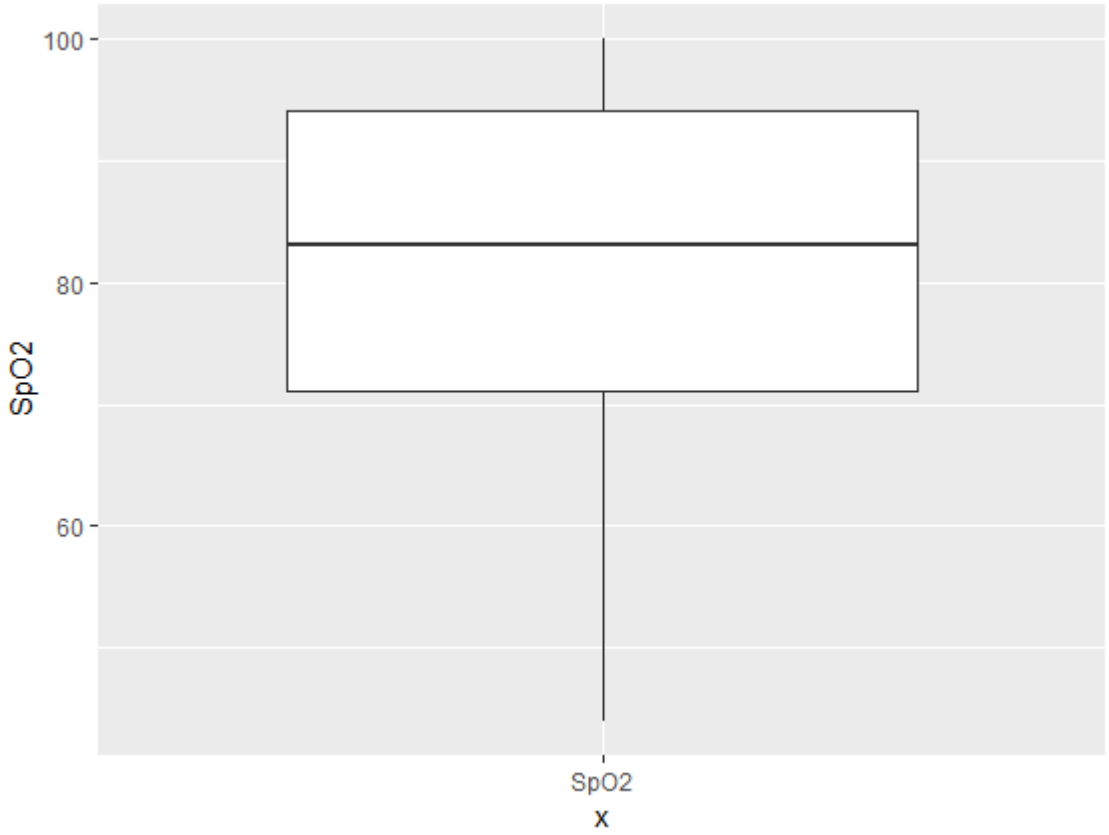
#Mean : 739.8 Mean : 81.94 Mean : 80.36 2/9/2015 10:49: 5

#3rd Qu.:1082.0 3rd Qu.: 94.00 3rd Qu.: 94.00 2/9/2015 13:25: 5

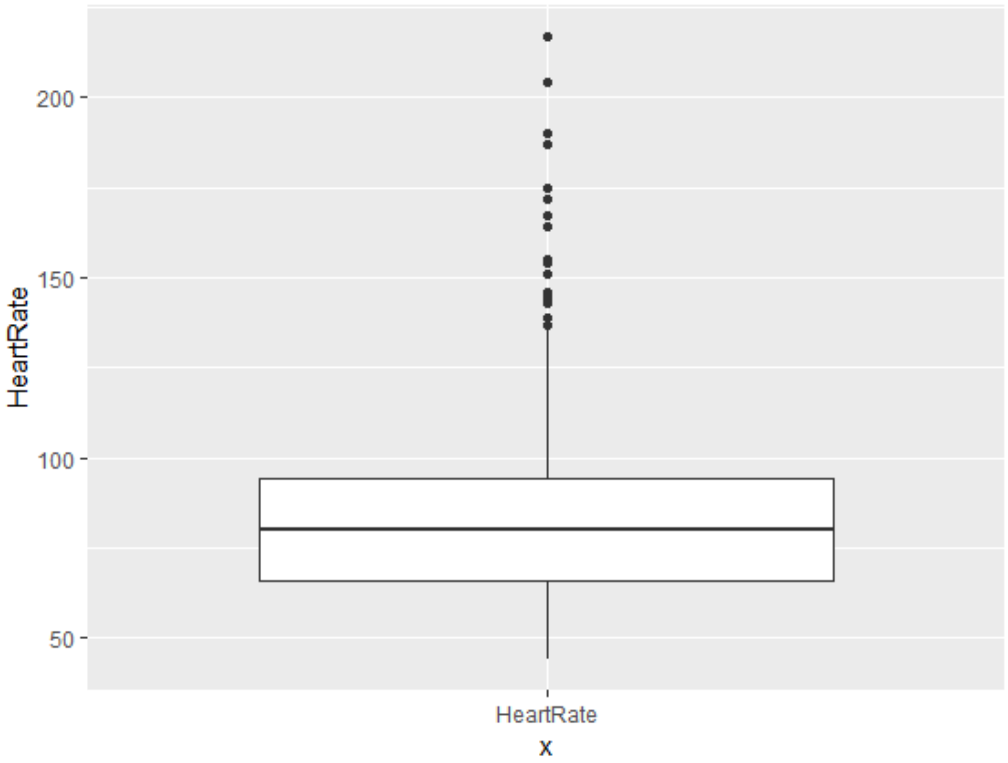
#Max. :1453.0 Max. :100.00 Max. :217.00 2/1/2015 12:45: 4

#(Other) :9843

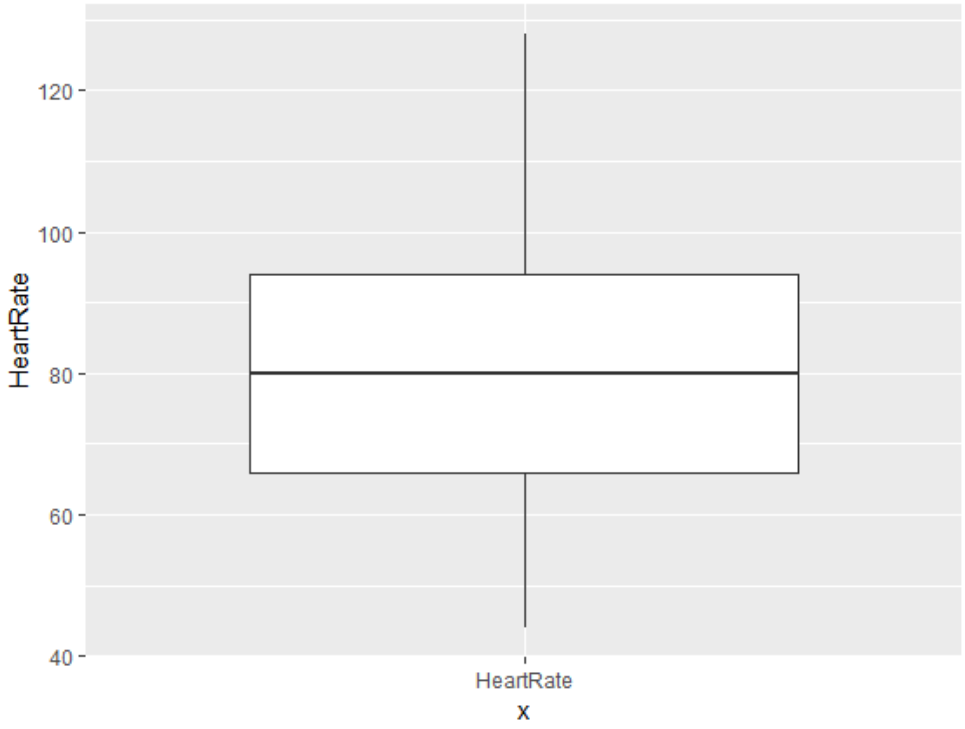
Se logra identificar que la variable SpO2 es una variable sesgada a la izquierda con datos normales.



Sin embargo, la variable heart rate presenta 26 outliers los cuales corresponde a valores mayores a su limite superior.

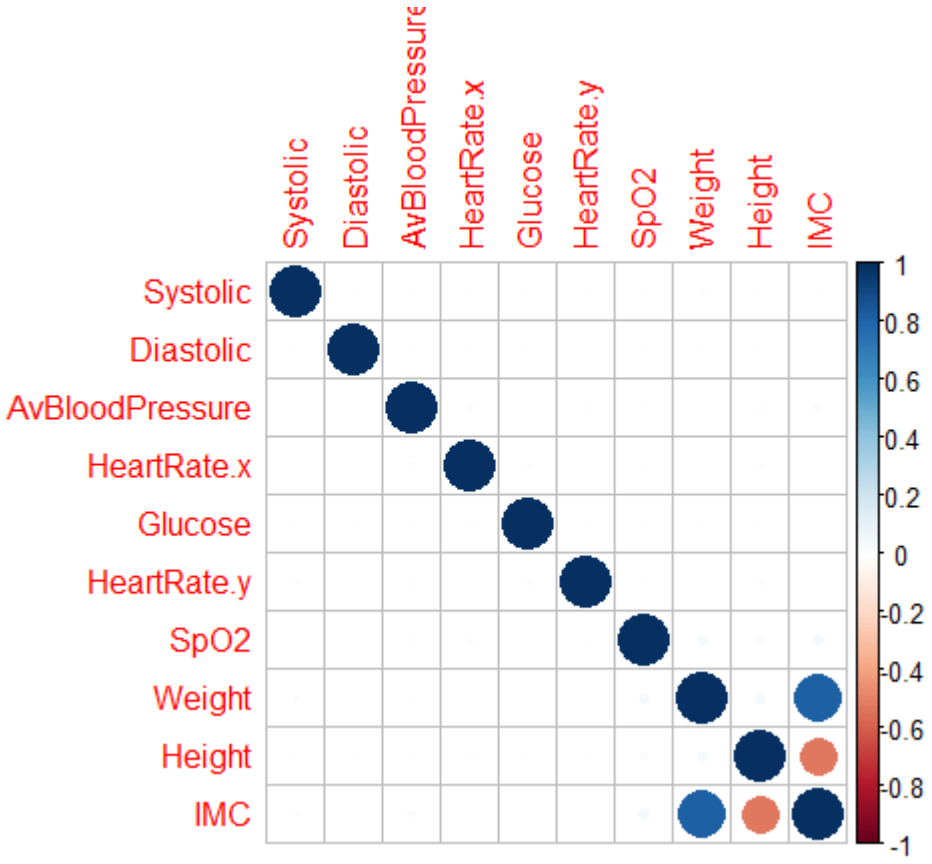


Luego de haber limpiado eliminado los outliers, se observan los datos normalizados.

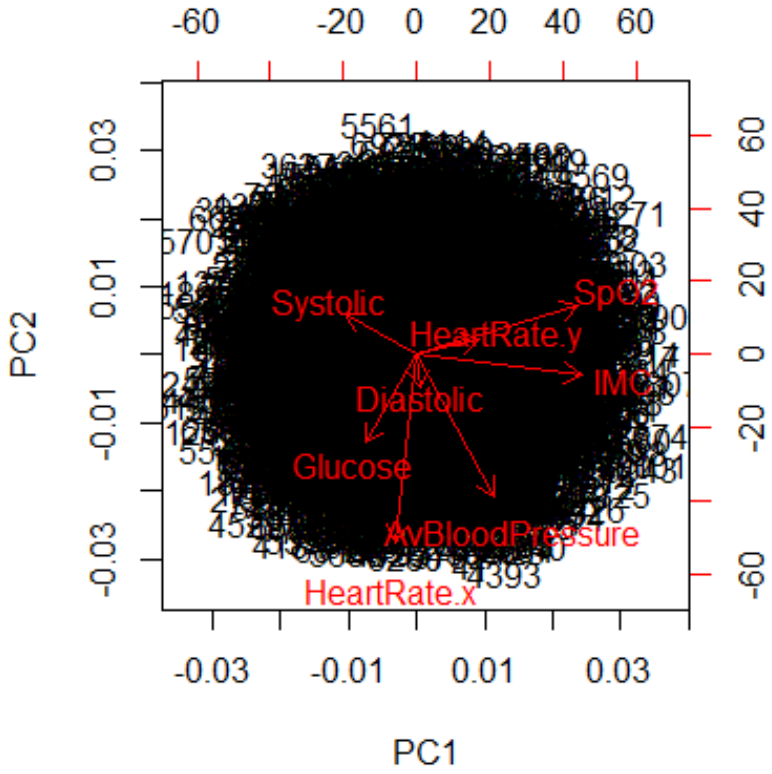


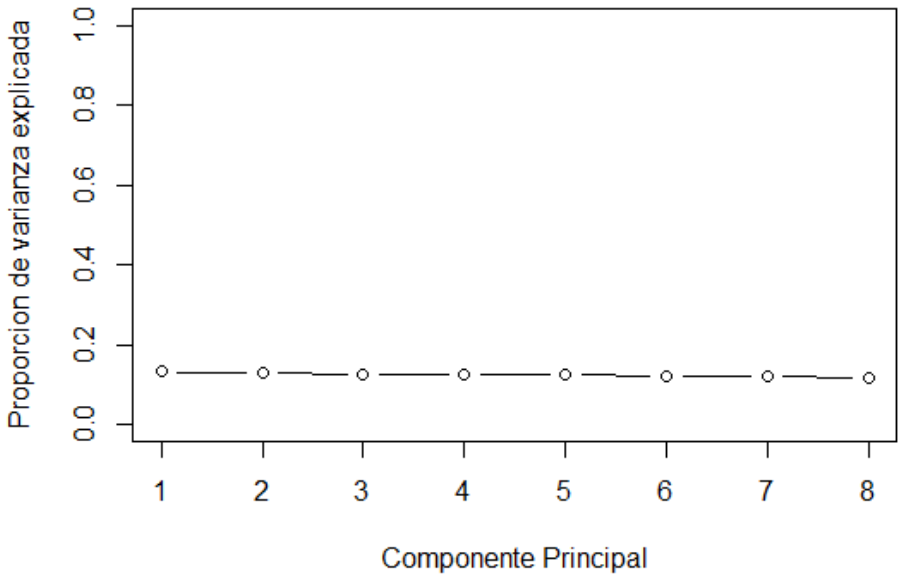
Filtrado de datos

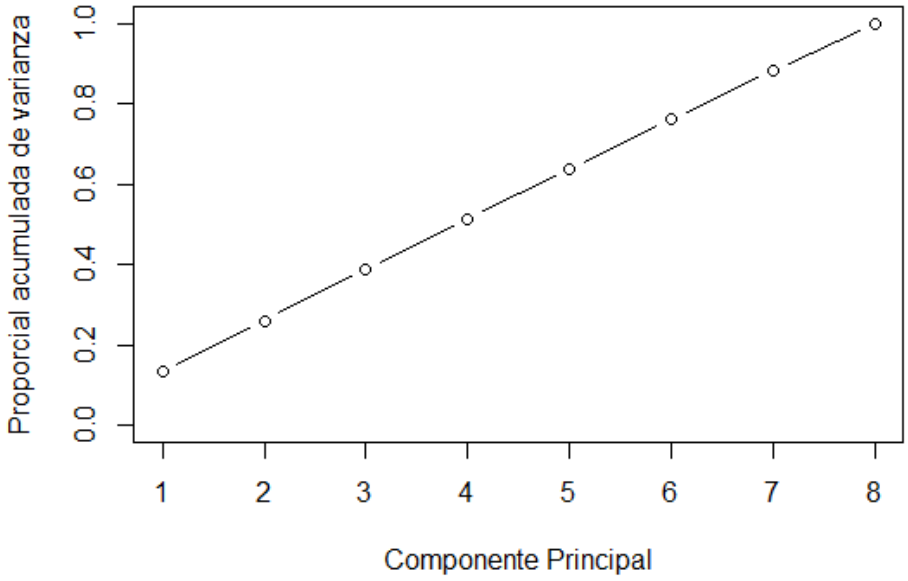
Se realiza análisis de correlación de variables, el cual muestra que tenemos alta correlación entre IMC Heigth y Weigth, para nuestro análisis debemos dejar solo una de las tres y el índice de masa corporal IMC es la variable que reúne condiciones de peso y altura y retiraremos el peso y altura.



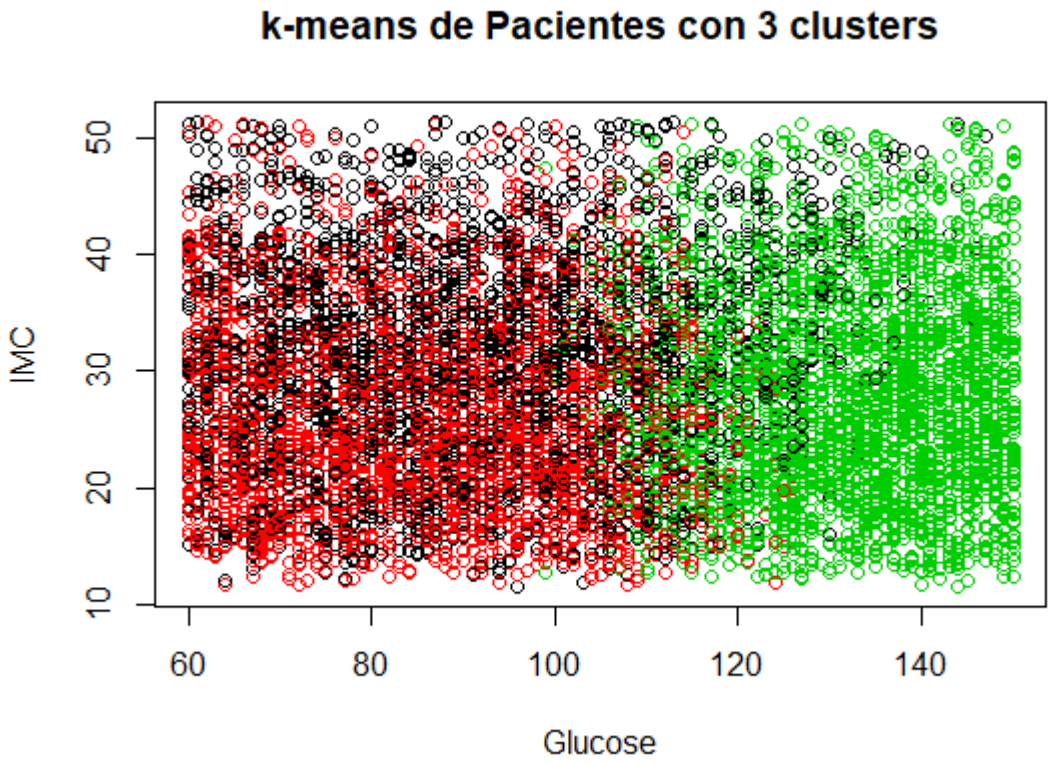
Análisis de componentes principales




Proporcion de varianza explicada
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
Componente Principal


Proporcial acumulada de varianza
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
Componente Principal

K-means clusters.



Comparación contra Componentes principales