PRACTICA 2: ¿Cómo realizar la limpieza y análisis de datos?

Autores

03/01/2023

Índice

1.Descripción del dataset

2.Integración y selección

3. Limpieza de los datos

En este apartado se corregirán los registros del csv que sean erróneos.

3.1. ¿Los datos contienen ceros o elementos vacíos?

Podemos comprobar que no se encuentran elementos nulos en el csv con el siguiente código:

```
sapply(df, function(x) sum(is.na(x)))
```

```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh
##        0        0        0        0        0        0        0        0
##     exng  oldpeak      slp      caa    thall   output
##        0        0        0        0        0        0
```

—TODO Discretizar edad —-

3.2. Identifica y gestiona los valores extremos

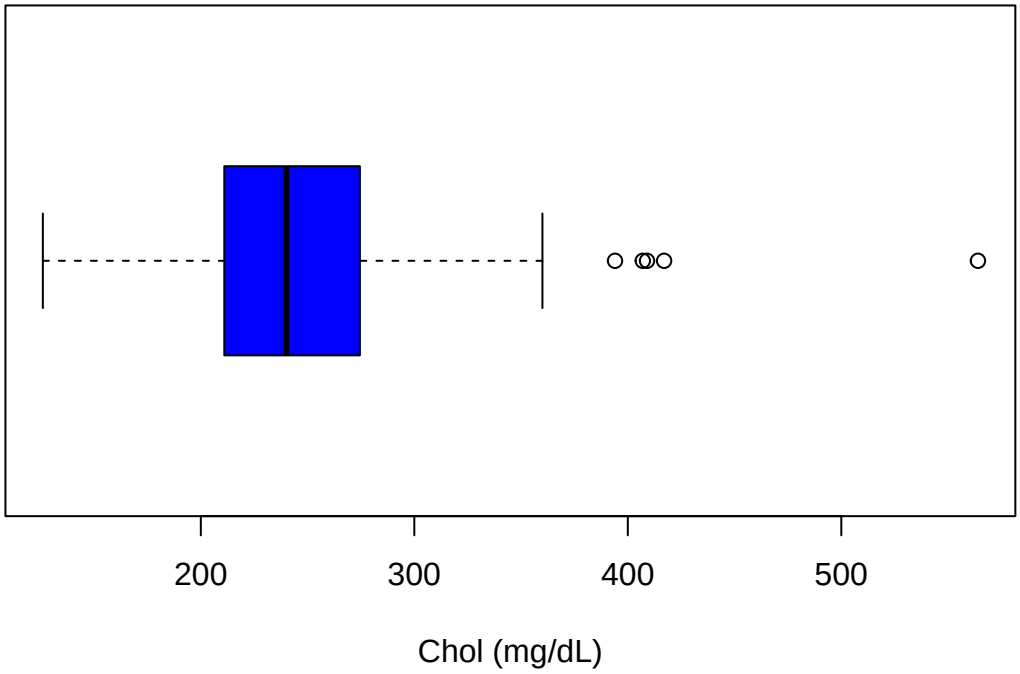
Los valores extremos (*outliers*) son aquellos que se encuentran fuera del rango [Lo,Ho], donde:

· Lo (Lower Outlier) = Q1 – (1.5 * IQR)

· Ho (Higher Outlier) = Q3 + (1.5 * IQR)

y IQR = Q3 - Q1. También hay una fórmula que trata outliers los que sobrepasan la media +/- la desviación estándar, pero optaremos por fórmula descrita.

Encontramos varios outliers en los datos. Visualizamos los outliers de la columna **chol**. Que según definido en el dataset es el colesterol en mg/dl capturado con un sensor BMI.



TODO Quiza plotear todas las columnas pero que se puedan leer. Problema es que no estan en la misma escala (0-1 y 80-200)

4. Análisis de los datos

4.1. Selección de los grupos de datos que se quieren analizar/compara

4.2. Comprobación de la normalidad y homogeneidad de la varianza

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos

5. Representación de los resultados a partir de tablas y gráficas

6. Resolución del problema

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```