Práctica 2 - Limpieza y validación de los datos

Irene Calvo Cuesta - icalvocu

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset elegido recoge un análisis sobre los gastos médicos individuales facturados por el seguro de salud de Estados Unidos, y una serie de características personales de los asegurados. Para realizar el análisis se han recogido datos de 1338 asegurados. El objetivo de dicho análisis es investigar si se puede predecir la prima del seguro a partir de la edad del asegurado, su género, su índice de masa corporal, el número de hijos que tiene y por tanto que están cubiertos por el seguro de salud, si es fumador o no, y la región de Estados Unidos en la que vive; y que grado de influencia tienen dichas características sobre los costos del seguro.

2. Integración y selección de los datos de interés a analizar.

En primer lugar se debe realizar la carga de los datos, para ello se inspecciona el tipo de formato csv. Se puede comprobar que se usa la coma (,) como separador de valores y el punto (.) como separador decimal, por tanto se usará la funcion **read.csv()** para la lectura del fichero.

```
#Lectura de datos
asegurados <- read.csv('insurance.csv')
#Se agiliza la manipulacion de los datos, para que no sea necesario especificar el nombre del dataframe
attach(asegurados)

#Se comprueba la carga correcta del archivo
str(asegurados)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

```
#Se observan las primeras filas del conjunto
head(asegurados)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

Como he comentado, el fichero que tenemos recoge información sobre una previsión del coste de la prima de los seguros médicos en Estados Unidos y una serie de datos personales de los asegurados.

Podemos observar que el fichero contiene 7 variables que corresponden a la edad de los asegurados (**age**), al género (**sex**), al índice de masa corporal (**bmi**), el número de hijos (**children**), si fuman o no (**smoker**), la región a la que pertenecen (**region**), y los cargos de la prima del seguro (**charges**). Este fichero cuenta con 1338 observaciones.

Visualizando las 5 primeras filas del archivo se comprueba que se ha cargado correctamente.

Por otro lado voy a observar las principales características que tiene las variables del conjunto de datos:

```
#Se observan las principales características de las variables del conjunto
summary(asegurados)
```

```
##       age            sex           bmi           children     smoker
##  Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064
##  1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274
##  Median :39.00                Median :30.40   Median :1.000
##  Mean   :39.21                Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                Max.   :53.13   Max.   :5.000
##        region       charges
##  northeast:324   Min.   : 1122
##  northwest:325   1st Qu.: 4740
##  southeast:364   Median : 9382
##  southwest:325   Mean   :13270
##                  3rd Qu.:16640
##                  Max.   :63770
```

Observamos que la muestra presenta una media de edad de 39 años, un índice de masa corporal medio de 30.66 kg/m2, una media de hijos entre 1 y 2, y unas primas de seguro medias de 13270$.

Comprobamos también que los niveles de las variables categóricas son los correctos, que no es necesario realizar alguna estandarización en los nombre. La variable **sex** tiene dos niveles (female y male), la variables **smoker** tiene dos niveles (no y yes) en función de si el asegurado fuma o no, y la variable región tiene cuatro niveles correspondientes a la división por regiones de EEUU en noreste, noroeste, sureste, suroeste.

Si que voy a estandarizar los valores de la variable **bmi** en dos cifras decimales, y los de la variable **charges** en tres cifras decimales:

```
#Estandarizacion variable bmi
asegurados$bmi<- round(asegurados$bmi, 2)
#Estandarizacion variable charges
asegurados$charges<- round(asegurados$charges, 3)

#Se comprueba que se ha realizado el cambio
head(asegurados, 3)
```

```
##   age    sex   bmi children smoker    region   charges
## 1  19 female 27.90        0    yes southwest 16884.924
## 2  18   male 33.77        1     no southeast  1725.552
## 3  28   male 33.00        3     no southeast  4449.462
```

A continuación se va a comprobar si se cumple el tipo de variable estadística que debe tener asociada cada variable:

Las variables **sex**, **smoker**, y **region** deben ser de tipo factor (cualitativa nominal), ya que no tienen un criterio de orden; las variables **age** y **children** deben ser de tipo integer, ya que contienen valores discretos; y las variables **bmi** y **charges** deben ser de tipo numeric, ya que la naturaleza de estas variables es continua.

```r
#Se muestran los tipos de variables
sapply(asegurados, class)
```

```
##       age       sex       bmi  children    smoker    region   charges
## "integer"  "factor" "numeric" "integer"  "factor"  "factor" "numeric"
```

Se observa que todas las variables tienen asignado el tipo apropiado, y por lo tanto no necesitan una conversión para conseguir que el tipo final sea el adecuado.

3. Limpieza de los datos.

A continuación se va a llevar a cabo la limpieza de los datos. Este paso es muy importante en cualquier análisis de datos.

3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

En primer lugar se va a buscar si los datos co

```r
#Se busca que variable tiene elementos vacios
sapply(asegurados, function(x) sum(is.na(x)))
```

```
##      age      sex      bmi children   smoker   region  charges
##        0        0        0        0        0        0        0
```

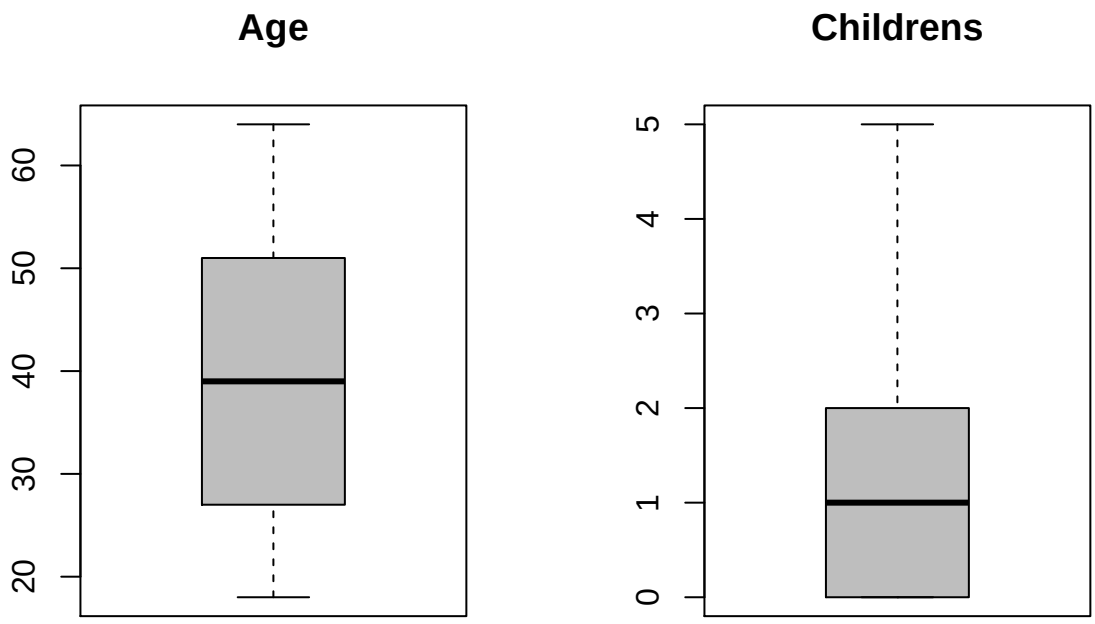
Se observa que ninguna de las variables contiene elementos vacíos, por tanto no habría que seguir ningún procedimiento adicional.

En el caso de que si que hubiese registros desconocidos una decisión sería eliminar dichos registros, siempre y cuando fuese una cantidad que no afectase a nuestra investigación; otro de los procedimientos, que sería el que yo utilizaría para gestionarlo, sería imputar los valores faltantes a partir de los k-vecinos más cercanos, utilizando por ejemplo la distancia de Gower.

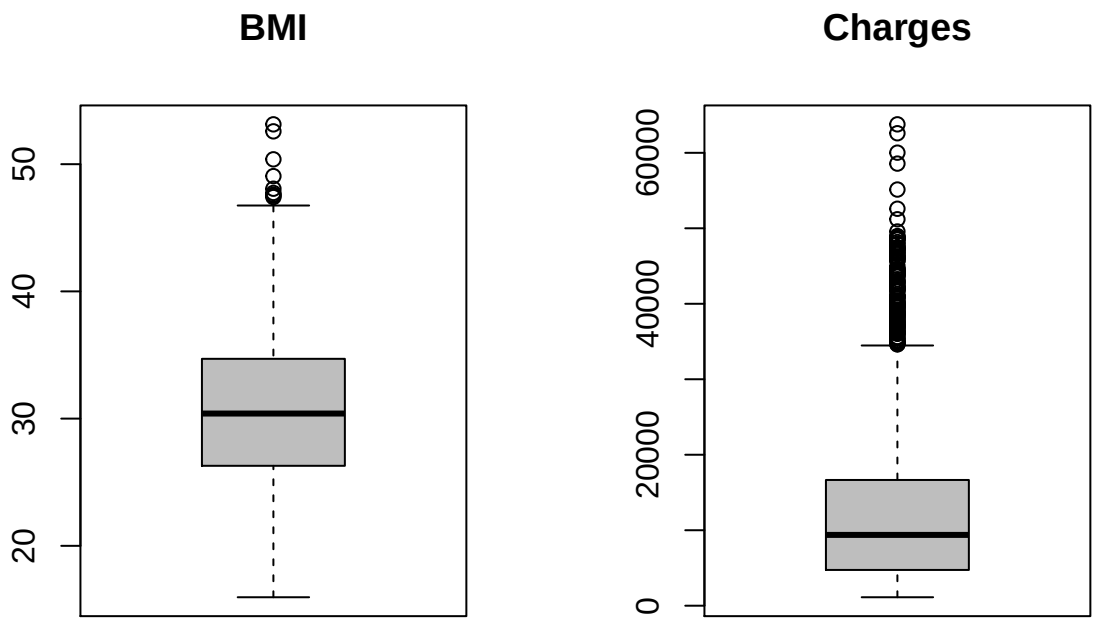
3.2 Identificación y tratamiento de valores extremos

Para identificar los valores extremos voy a presentar un boxplot para cada variable cuantitativa. En este dataset las variables cuantitativas que tenemos son: **age**, **children**, **bmi**, y **charges**.

```r
par(mfrow = c(1,2))
boxplot(asegurados$age, main='Age', col = 'grey')
boxplot(asegurados$children, main='Childrens', col = 'grey')
```




```
boxplot(asegurados$bmi, main='BMI', col = 'grey')
boxplot(asegurados$charges, main='Charges', col = 'grey')
```



Se encuentran varios valores atípicos en las variables **bmi** y **charges**. Por tanto voy a inspeccionar dichos valores

```
#lista de valores atípicos
boxplot.stats(asegurados$bmi)$out
```

```
## [1] 49.06 48.07 47.52 47.41 50.38 47.60 52.58 47.74 53.13
```

```
boxplot.stats(asegurados$charges)$out
```

```
##   [1] 39611.76 36837.47 37701.88 38711.00 35585.58 51194.56 39774.28
##   [8] 48173.36 38709.18 37742.58 47496.49 37165.16 39836.52 43578.94
##  [15] 47291.06 47055.53 39556.50 40720.55 36950.26 36149.48 48824.45
##  [22] 43753.34 37133.90 34779.61 38511.63 35160.14 47305.31 44260.75
##  [29] 41097.16 43921.18 36219.40 46151.12 42856.84 48549.18 47896.79
##  [36] 42112.24 38746.36 42124.51 34838.87 35491.64 42760.50 47928.03
##  [43] 48517.56 41919.10 36085.22 38126.25 42303.69 46889.26 46599.11
##  [50] 39125.33 37079.37 35147.53 48885.14 36197.70 38245.59 48675.52
##  [57] 63770.43 45863.21 39983.43 45702.02 58571.07 43943.88 39241.44
##  [64] 42969.85 40182.25 34617.84 42983.46 42560.43 40003.33 45710.21
##  [71] 46200.99 46130.53 40103.89 34806.47 40273.64 44400.41 40932.43
##  [78] 40419.02 36189.10 44585.46 43254.42 36307.80 38792.69 55135.40
##  [85] 43813.87 39597.41 36021.01 45008.96 37270.15 42111.67 40974.17
##  [92] 46113.51 46255.11 44202.65 48673.56 35069.38 39047.29 47462.89
##  [99] 38998.55 41999.52 41034.22 36580.28 35595.59 42211.14 44423.80
```

```
## [106] 37484.45 39725.52 44501.40 39727.61 48970.25 39871.70 34672.15
## [113] 41676.08 44641.20 41949.24 36124.57 38282.75 46661.44 40904.20
## [120] 36898.73 52590.83 40941.29 39722.75 37465.34 36910.61 38415.47
## [127] 41661.60 60021.40 47269.85 49577.66 37607.53 47403.88 38344.57
## [134] 34828.65 62592.87 46718.16 37829.72 36397.58 43896.38
```

```
length(boxplot.stats(asegurados$bmi)$out)
```

```
## [1] 9
```

```
length(boxplot.stats(asegurados$charges)$out)
```

```
## [1] 139
```

Se observa que hay 9 valores extremos en la variable **bmi**, y 139 valores extremos de la variable **charges**, de un total de 1338 observaciones, por lo que se va a considerar que no son valores erroneos sino que al ser escogidos aleatoriamente se han podido dar en distinto asegurados.

Una vez realizada la limpieza de los datos se va a proceder a su análisis.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Voy a realizar una agrupacion por condicion de fumador, que me será util a la hora de analizar los datos.

```
asegurados.fumadores <- asegurados[asegurados$smoker=='yes',]
asegurados.nofumadores <- asegurados[asegurados$smoker=='no',]
```

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Para comprobar la normalidad de la varianza voy a applicar el **test de Shapiro Wilk** en las variables cuantitativas (**age**, **children**, **bmi**, y **charges**)

```
asegurados.cuantit <- asegurados[, c(1,3,4,7)]

alpha = 0.05
for(i in 1:length(asegurados.cuantit)){
  if(shapiro.test(asegurados.cuantit[,i])$p.value<alpha){
    cat('Se rechaza la hipótesis nula. La variable', names(asegurados.cuantit)[i], 'no sigue una distri
  }
  else{
    cat('No se rechaza la hipótesis nula. La variable', names(asegurados.cuantit)[i], 'sigue una distri
  }
}
```

```
## Se rechaza la hipótesis nula. La variable age no sigue una distribución normal.
## Se rechaza la hipótesis nula. La variable bmi no sigue una distribución normal.
## Se rechaza la hipótesis nula. La variable children no sigue una distribución normal.
## Se rechaza la hipótesis nula. La variable charges no sigue una distribución normal.
```

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.

Voy a contrastar el hecho de que el valor medio de las primas del seguro de los no fumadores sea inferior al valor medio de las primas del seguro de los fumadores.

```
t.test(asegurados.nofumadores$charges, asegurados.fumadores$charges, mu=0, conf.level = 0.95, alternati
```

```
##
##  Welch Two Sample t-test
##
## data:  asegurados.nofumadores$charges and asegurados.fumadores$charges
## t = -32.752, df = 311.85, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -22426.4
## sample estimates:
## mean of x mean of y
##  8434.268 32050.232
```

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

Antes de finalizar se va a crear el archivo de datos corregido, con todos los cambios hasta el momento:

```
write.csv(asegurados, file = 'insurance_clean.csv', row.names = FALSE)
```

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Después de la limpieza, análisis, y representación de los datos se puede concluir

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

El código se encuentra en el archivo 'insurance.Rmd'