Práctica 2
Limpieza y análisis de datos

Alejandra Cristina Marrero Suarez
Luis Miguel Moreno López

ÍNDICE

1 Descripción del DataSet

El *dataset* *insurance.csv* contiene 1338 registros constituidos por 7 columnas o variables.

En concreto, las columnas que conforman el dataset son:

```
> doc_csv <- read.csv("INSURANCE.csv",header = TRUE)
> head(doc_csv)
  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
> str(doc_csv)
'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 ...
```

Ilustración 1- Columnas que conforman el dataset

- **Años:** edad del beneficiario principal.
- **Sexo:** género del beneficiario principal (mujer u hombre).
- **BMI:** índice de masa corporal, que proporciona una comprensión del cuerpo, pesos que son relativamente altos o bajos en relación con la altura, índice objetivo de peso corporal (kg / m ^ 2) utilizando la relación entre la altura y el peso, idealmente 18,5 a 24,9
- **Hijos:** número de hijos cubiertos por el seguro médico / Número de dependientes
- **Fumador:** si es fumador o no.
- **Región:** el área residencial del beneficiario en los EE. UU., noreste, sureste, suroeste, noroeste.
- **Cargos:** costos médicos individuales facturados por el seguro médico.

A partir de los datos ofrecidos por este dataset se pretende resolver cuestiones cómo: ¿Qué género de la muestra proporcionada fuma más?, ¿Qué género fuma más?, ¿Cómo afecta fumar a los costos médicos?

2 Integración y selección de los datos

A partir de los datos proporcionados por el dataset, es interesante estudiar como una serie de variables; edad, sexo, número de hijos de un individuo ... influyen en los costos médicos.

Se podrían crear modelos que ayuden a predecir los costes médicos de un individuo a partir de las características de edad, sexo … que este disponga.

La importancia de este estudio recae en la futura capacidad de las empresas de seguros de poder establecer primas de riesgo y cuotas más acertadas a sus clientes en función de las características de estos.

La única variable que se podría descartar del dataset para este estudio es la región, ya que es la única que se podría considerar como no intrínseca a la propia persona, además de que los valores que puede tomar son demasiado reducidos en espectro.

3 Limpieza de los datos

Una vez realizada una valoración de aquellos atributos y datos del dataset más relevantes para nuestros estudios, continuamos con la limpieza de los datos.

3.1 Valores nulos

Las variables seleccionadas no contienen ceros ni elementos vacíos. Para asegurarnos hemos lanzado la función colSums creando una tabla que nos muestre la suma de valores nulos de cada columna. Al ser 0 para todas nos hemos asegurado de que no tenemos valores nulos.

```
> colSums(is.na(doc_csv))
     age      sex      bmi children   smoker   region  charges
       0        0        0        0        0        0        0
> colSums(doc_csv =="")
     age      sex      bmi children   smoker   region  charges
       0        0        0        0        0        0        0
> |
```

Aun así, vamos a dar un paso más viendo los posibles valores tomados por algunas de estas variables.
La variable sexo es una variable cualitativa conformada por dos opciones, female y male.

```
> table(doc_csv$sex)

female   male
   662    676
```

Ilustración 2- Valores y concurrencia de estos de la variable sex en el dataset

No hay que eliminar espacios ni estandarizar la variable.
Igual ocurre para la variable smoker. No contiene datos nulos, no tiene que ser estandarizada.

```
> table(doc_csv$smoker)

  no  yes
1064  274
```

Ilustración 3- Valores y concurrencia de estos de la variable smoker en el dataset

3.2 Identificación y tratamiento de valores extremos

Se presenta un boxplot de cada variable cuantitativa. Además, se realiza una tabla con las estimaciones robustas y no robustas de tendencia central y dispersión para cada variable cuantitativa.

Primero se contempla obtener qué variables son cuantitativas.

```
> res <- sapply(doc_csv,class)
> kable(data.frame(variables=names(res),clase=as.vector(res)))


|variables |clase   |
|:---------|:-------|
|age       |integer |
|sex       |factor  |
|bmi       |numeric |
|children  |integer |
|smoker    |factor  |
|region    |factor  |
|charges   |numeric |
>
```

Ilustración 4- Tipología de las variables

```
res <- sapply(doc_csv,class)
kable(data.frame(variables=names(res),clase=as.vector(res)))
res <- which(res=="integer" | res == "numeric")
```

Ilustración 5- Seleccionamos solo las variables cuantitativas

```
for(i in 1:4){
  boxplot(doc_csv[,res[i]],main=names(doc_csv)[res[i]],col="gray")
}
```

Ilustración 6- Código generación boxplot

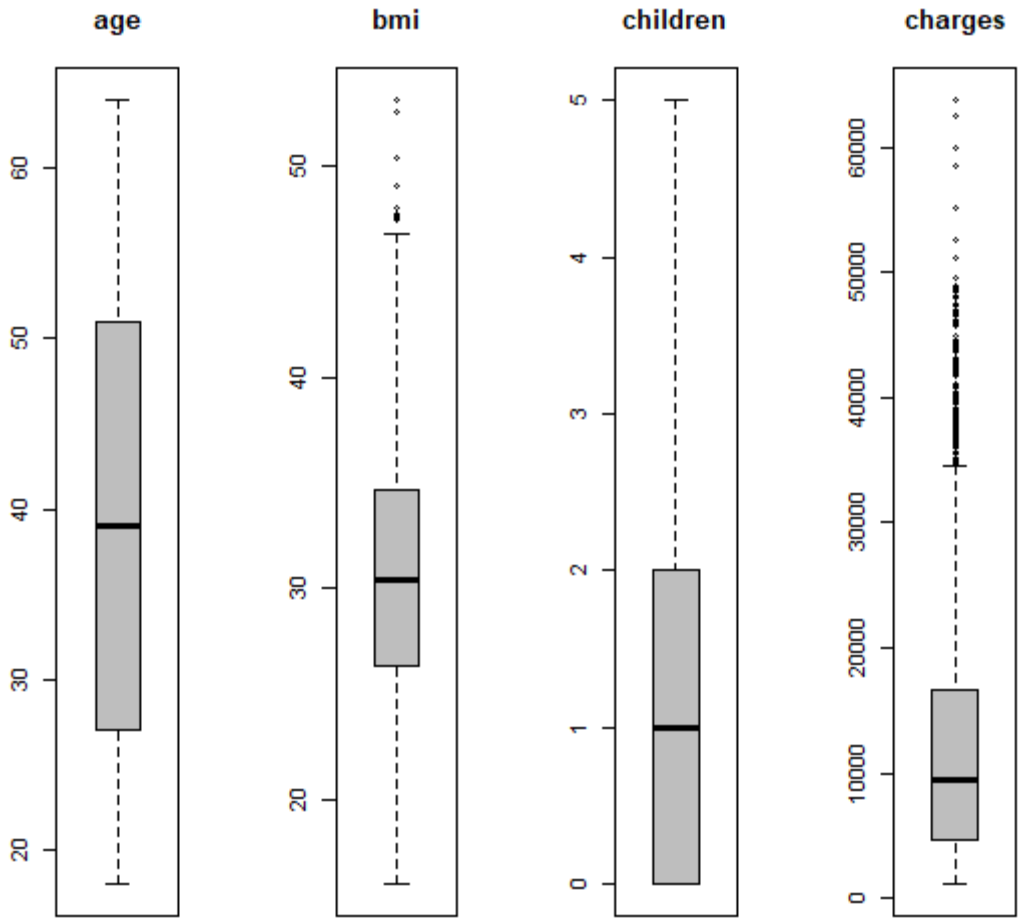


Ilustración 7- Boxplot variables cuantitativas dataset

Una vez visualizadas las cuatro variables cuantitativas podemos ver que solo dos de ellas cuentan con valores de carácter atípico: bmi y charges. Además, en estos gráficos podemos ver los cuartiles y la mediana.

```
> clus <- doc_csv[,c("age","sex","bmi","children","smoker","region","charges")]
> clus
     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
7     46 female 33.440        1     no southeast  8240.590
8     37 female 27.740        3     no northwest  7281.506
9     37   male 29.830        2     no northeast  6406.411
10    60 female 25.840        0     no northwest 28923.137
```

Ilustración 8- Selección de columnas del dataset

Para la normalización de las variables en la estructura de datos clus, copiamos su contenido en clus2 y reemplazamos las columnas del clus2 por las normalizadas.

```
> clus2[,c("age")] <- (clus$age-mean(clus$age))/sd(clus$age)
```

Ilustración 9- Normalización variable age

```
> clus2[,c("children")] <- (clus$children-mean(clus$children))/sd(clus$children)
> str(clus2)
'data.frame':   1338 obs. of  7 variables:
 $ age     : num  -1.438 -1.509 -0.798 -0.442 -0.513 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi     : num  -0.453 0.509 0.383 -1.305 -0.292 ...
 $ children: num  -0.9083 -0.0787 1.5803 -0.9083 -0.9083 ...
 $ 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 ...
```

Ilustración 10- Normalización variable children y visualización clus2

Realizamos algunas representaciones gráficas para describir las variables normalizadas y comprobamos que la nube de puntos representada es igual a la original, lo único que cambia es la escala de los ejes.

```
> plot(clus2[c("smoker","charges")], xlab="Fidelidad",ylab="Experiencia")
> xk <- clus2[,c("age","bmi","children","charges")]
```

```
> plot(xk[c("age","charges")], xlab="Fidelidad",ylab="Experiencia")
> title(main="nube de puntos normalizados",col.main="blue",font.main=1)
```

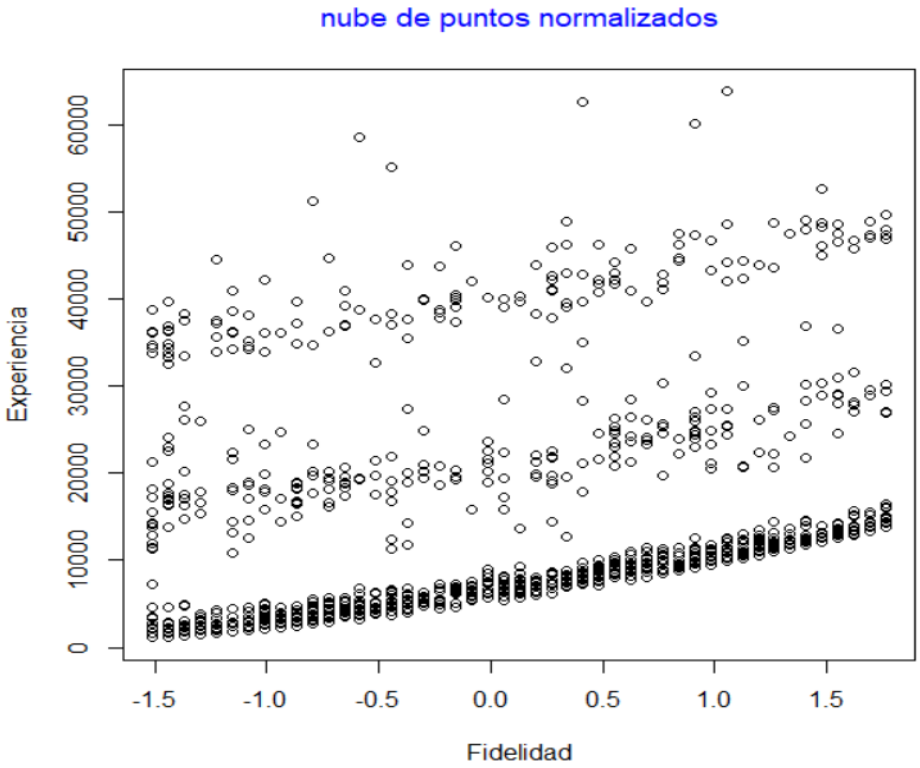


Ilustración 11- Representación gráfica normalización variables age y children

Los algoritmos de segmentación no supervisados, como es el kmeans(), requieren que el analista determine cuál es el número de clústers (grupos) a formar, de hecho, la función kmeans() incorpora como parámetro el número de clústers (centers=).

Utilizamos la función kmeans() para formar 3 grupos de individuos y visualizamos algunos resultados como son: los centros de grupos, la suma de cuadrados totales, las sumas de cuadrados dentro de cada grupo y para todos de forma conjunta y la suma de cuadrados entre grupos.

```
> clus2_k3 <-kmeans(xk,center=3)
> clus2_k3$centers
         age         bmi      children   charges
1 -0.2617116 -0.02902627  0.007566516  5933.822
2  0.5386878 -0.24008426 -0.033961644 17331.353
3  0.1431968  0.64037943  0.034157723 40279.623
> clus2_k3$totss
[1] 196074225579
> clus2_k3$withinss
[1] 7704468668 8275419963 7028870420
> clus2_k3$tot.withinss
[1] 23008759051
```

Ilustración 12- Kmeans 3 resultados

Para la selección del número de clústers existen criterios objetivos los cuales están basados en la optimización de un criterio de ajuste. Los criterios de ajustes en el kmeans() se basan en los conceptos de sumas de cuadrados entre grupos (betweens) y dentro de grupos (withins).

La suma de cuadrados entre grupos (betweenss) más las sumas de cuadrados dentro de grupos (tot.withinss) nos proporciona la suma de cuadrados totales (tots).

```
> #suma de cuadrados entre grupos
> kmeans(xk,2)$betweenss
[1] 148059301016
> kmeans(xk,3)$betweenss
[1] 1.731e+11
> #suma de cuadrados dentro de grupos
> kmeans(xk,3)$tot.withinss
[1] 23008759051
> #suma de cuadrados total
> kmeans(xk,3)$totss
[1] 196074225579
> kmeans(xk,2)$totss
[1] 196074225579
```

Ilustración 13- Kmeans 2 resultados

Creamos la gráfica que representa la suma de cuadrados entre grupos mirando el número de grupos para saber la varianza de crecimiento.

```
> bss <- kmeans(xk,centers=1)$betweens
> for(i in 2:10) bss[i] <- kmeans(xk,center=i)$betweens

> plot(1:10,bss,type="l",xlab="Número de grupos", ylab="Sumas de cuadrados entre grupos")
```

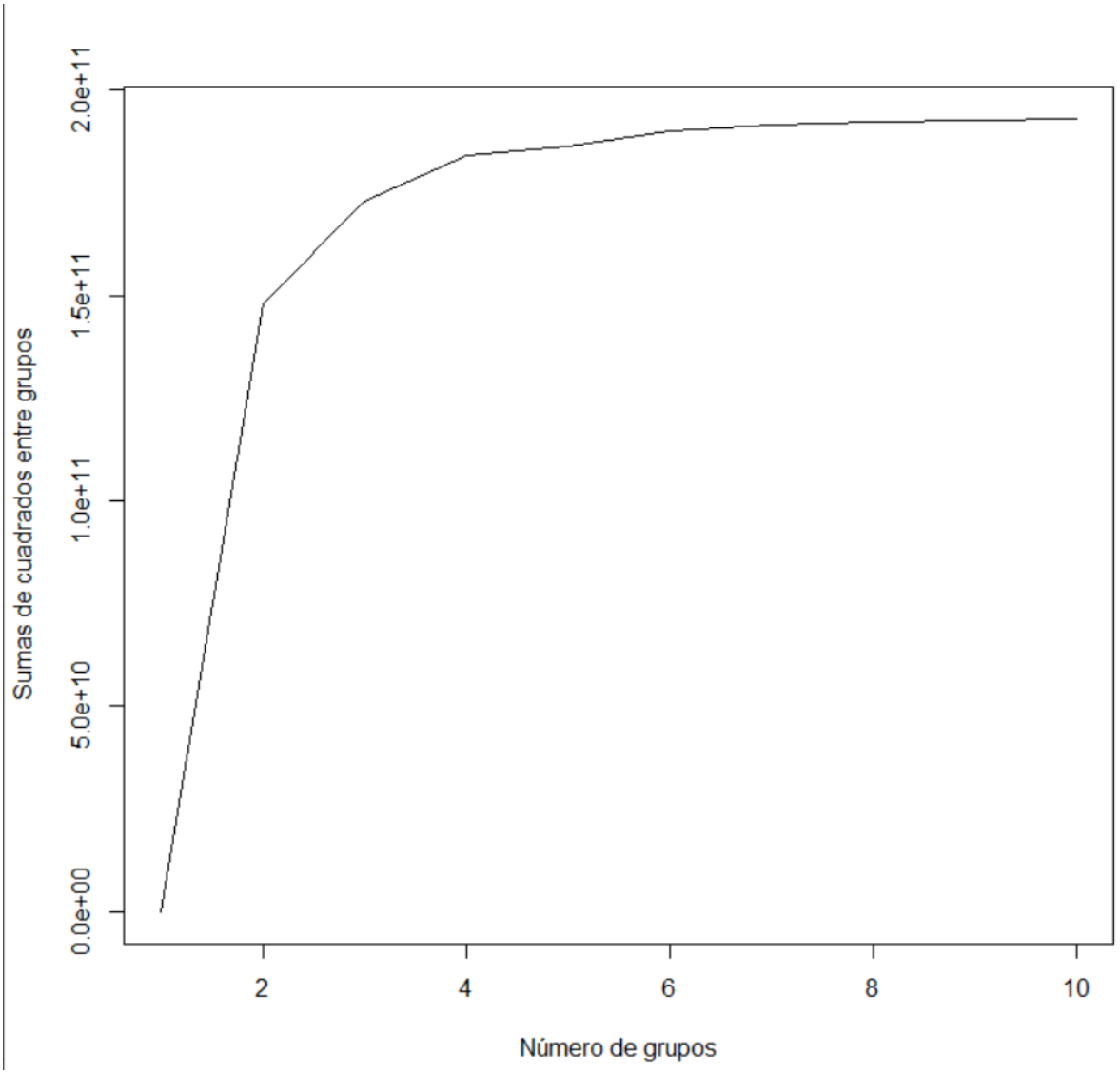


Ilustración 14- Representación media distancia entre centros de los clusters ($betweens)

Ahora debemos elegir el K intentando que la suma de cuadrados sea lo mayor posible entre grupos.

```
> tot_w <- kmeans(xk,centers=1)$tot.withinss
> for(i in 2:10) tot_w[i] <- kmeans(xk,center=i)$tot.withinss
> plot(1:10,tot_w,type="l",xlab="Número de grupos", ylab="Sumas de cuadrados entre grupos")
```

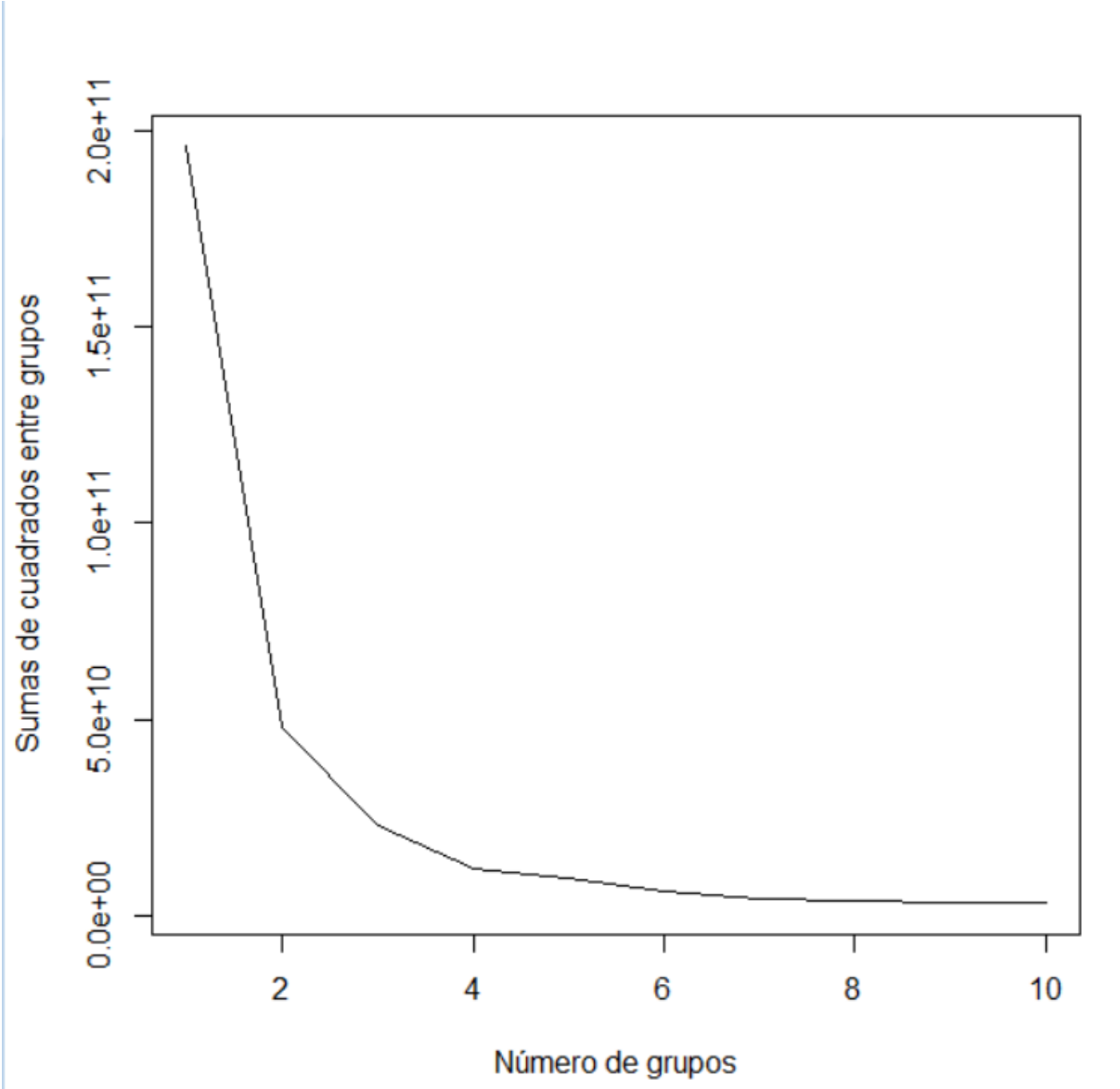


Ilustración 15 Representación ($tot.withinss)

Como se puede ver en estas dos gráficas a partir de k=2 la pendiente deja de ser tan pronunciada, por lo que a partir de ahí no es tan provechoso realizar otros grupos.

Por lo que fijamos el k=2.

En este último paso de la agrupación usando k-means con las variables age, bmi, children y charges dibujamos la representación de los grupos generados.

```
> x <- xk
> fit <- kmeans(x,2)
> clusplot(x,fit$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```

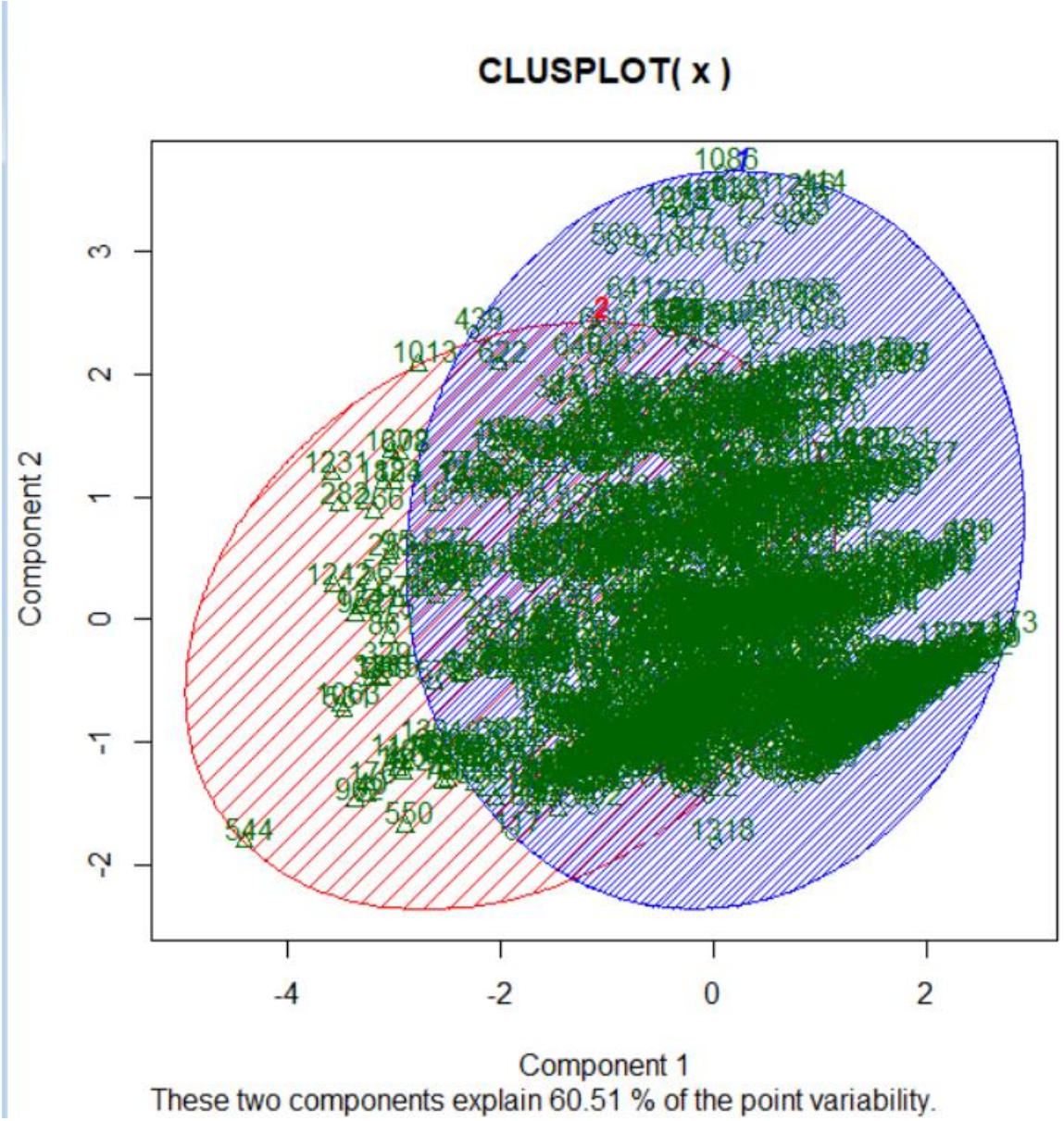


Ilustración 16- Representación grupos obtenidos con K-means

4 Análisis de los datos

Supuestos de normalidad y Contraste

Charges

Supuesto de normalidad

Con el gráfico Q-Q se hace una primera aproximación visual de si hay o no normalidad. Hay que tener en cuenta que este gráfico es meramente descriptivo.

Interpretación: La nube de puntos se sitúa sobre la recta. En un principio, visualmente se aprecia que nuestros datos no cumplen el supuesto de normalidad.

```
> qqnorm( doc_csv$charges )
> qqline( doc_csv$charges )
>
```

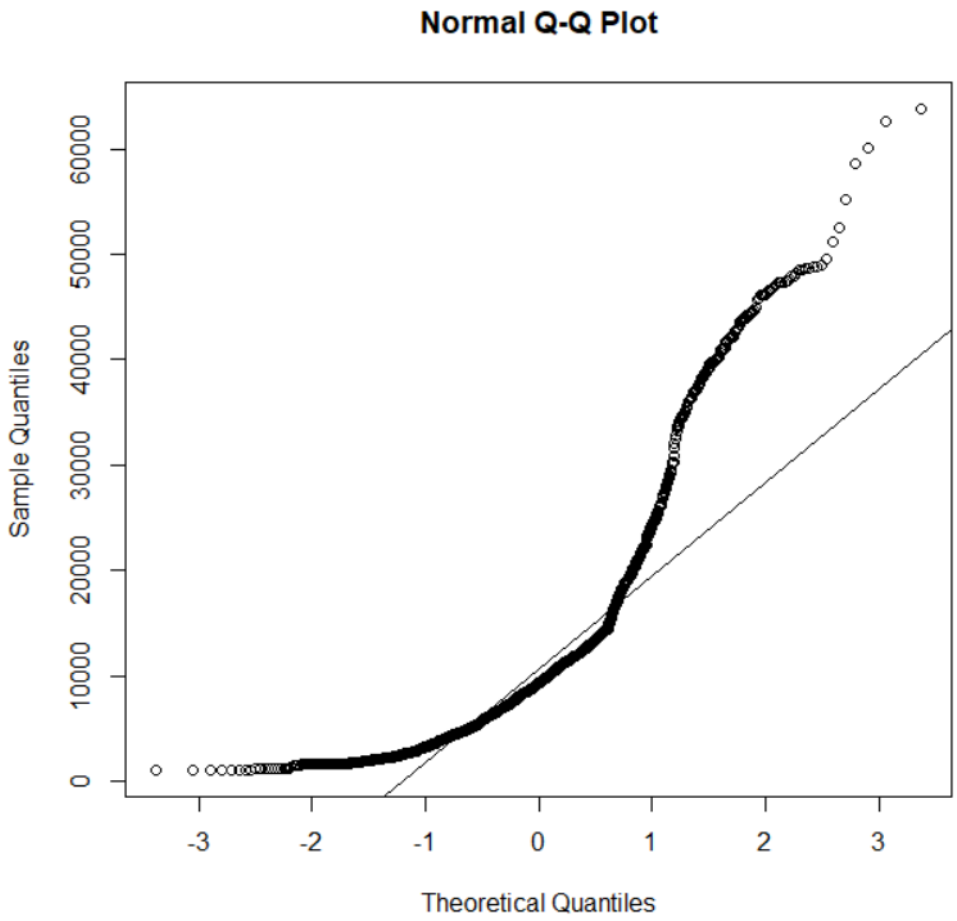


Ilustración 17- Representación normal Q-Q

Contraste normalidad

Realizar el **contraste para normalidad**. En este contraste *la hipótesis nula es la hipótesis de normalidad*, esto es, no hay diferencias entre nuestra distribución y una distribución normal con esa media y esa desviación típica. Para contrastar la normalidad usamos el test de Shapiro-Wilk, con la función shapiro.test( ).

```
> shapiro.test (doc_csv$charges)

        Shapiro-Wilk normality test

data:  doc_csv$charges
W = 0.81469, p-value < 2.2e-16
```

Ilustración 18- Shapiro test

Interpretación: Con un p-value =2.2e-16 mayor de 0.05 no podemos rechazar la hipótesis nula (hipótesis de normalidad). Por lo tanto, podemos concluir que nuestros datos cumplen el supuesto de normalidad.

Contraste de hipótesis

Se supone normalidad en nuestros datos, podemos realizar el contraste.
Definimos nuestras hipótesis. Queremos probar si la media de los valores en el mes inicial es 10. Por lo tanto, tenemos que:

- Hipótesis nula es H0: μ=10
- Hipótesis alternativa es H1: μ≠10

Realizamos el contraste. La prueba t para una muestra se utiliza cuando tenemos una variable de medida y un valor esperado para la media, y se supone normalidad de los datos (o muestra grande). Para este contraste sobre una media utilizamos el t.test:

```
> t.test( doc_csv$charges,
+         mu = 10,
+         alternative = "two.sided" )

        One Sample t-test

data:  doc_csv$charges
t = 40.054, df = 1337, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 10
95 percent confidence interval:
 12620.95 13919.89
sample estimates:
mean of x
 13270.42
```

Ilustración 19- Contraste de hipótesis

Interpretamos los resultados. Con un *p-value = 2.2e-16* mayor de 0.05 no podemos rechazar la *hipótesis nula* H0. El intervalo de confianza excluye el 10 (12620.95 13919.89).

Charges - Sex

¿Debemos aceptar o rechazar la diferencia de la media de los cargos "charges" según el sexo Sex, para α=0.05?

Estamos ante un contraste para **dos muestras independientes** (hombres y mujeres). Para dos muestras independientes se debe comprobar el supuesto de normalidad y el supuesto de homocedasticidad. Después se realiza el contraste sobre lo que queremos probar

- PREPARAMOS NUESTROS DATOS. Creamos HombresIni solo con los datos del (charges) de hombres (Sex == "male"), y creamos MujeresIni solo con los datos del mes inicial (charges) de mujeres (Sex == "female"). Serán nuestras dos muestras independientes.

Supuesto de normalidad

```
> HombresIni <- doc_csv$charges[doc_csv$sex == "male"]
> MujeresIni <- doc_csv$charges[doc_csv$sec == "female"]
```

Ilustración 20- Selección de valores de charges para cada sexo

Con el **gráfico Q-Q** se hace una primera aproximación visual, y con el test de Shapiro-Wilk se realiza el **contraste para normalidad**. La normalidad se comprueba para cada una de las muestras (Hombres y Mujeres).

Supuesto de normalidad para los Hombres:

```
> qqnorm( HombresIni )
> qqline( HombresIni )
```

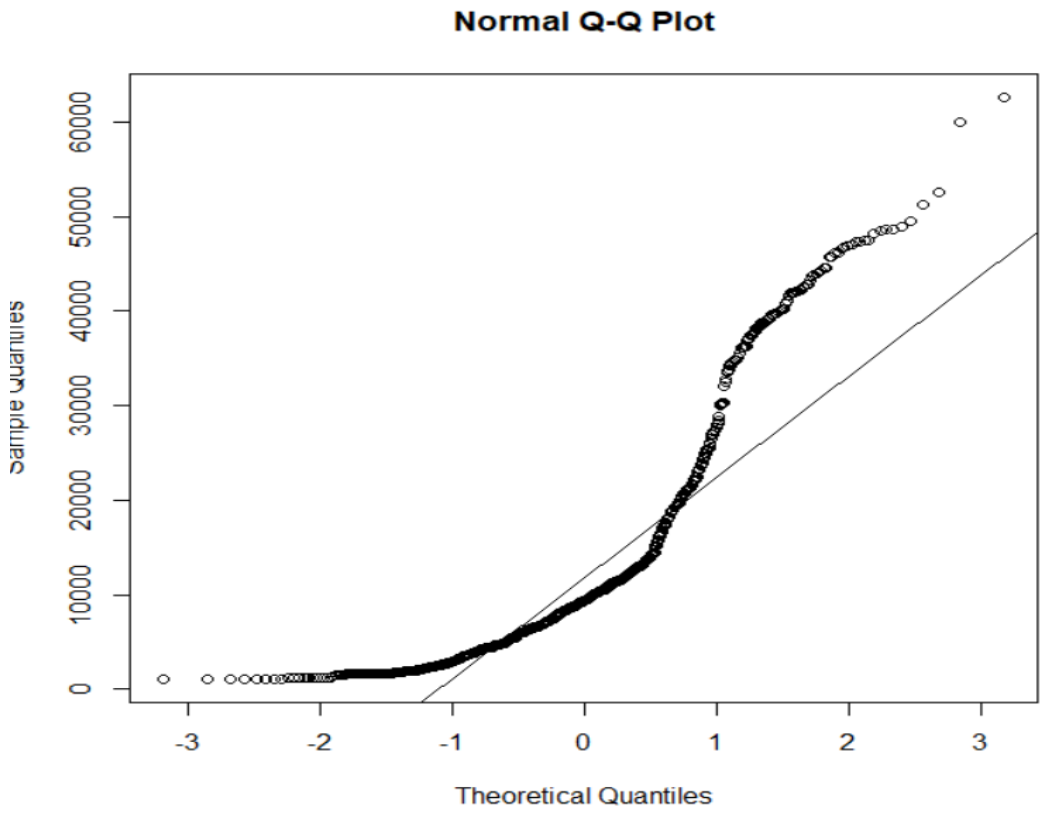


Ilustración 21- Representación normal Q-Q

```
> shapiro.test ( HombresIni ) # contraste de normalidad

        Shapiro-Wilk normality test

data:  HombresIni
W = 0.82281, p-value < 2.2e-16
```

Ilustración 22- Shapiro test

Interpretación Hombres:

La nube de puntos se ordena cerca de la recta. En un principio, visualmente se aprecia que nuestros datos cumplen el supuesto de normalidad.

Con un p-value =2.2e-16, mayor de 0.05, no podemos rechazar la hipótesis nula. Por lo tanto, podemos concluir que **nuestros datos cumplen el supuesto de normalidad**.

Supuesto de normalidad para las Mujeres:

```
> qqnorm( MujeresIni )
> qqline( MujeresIni )
```

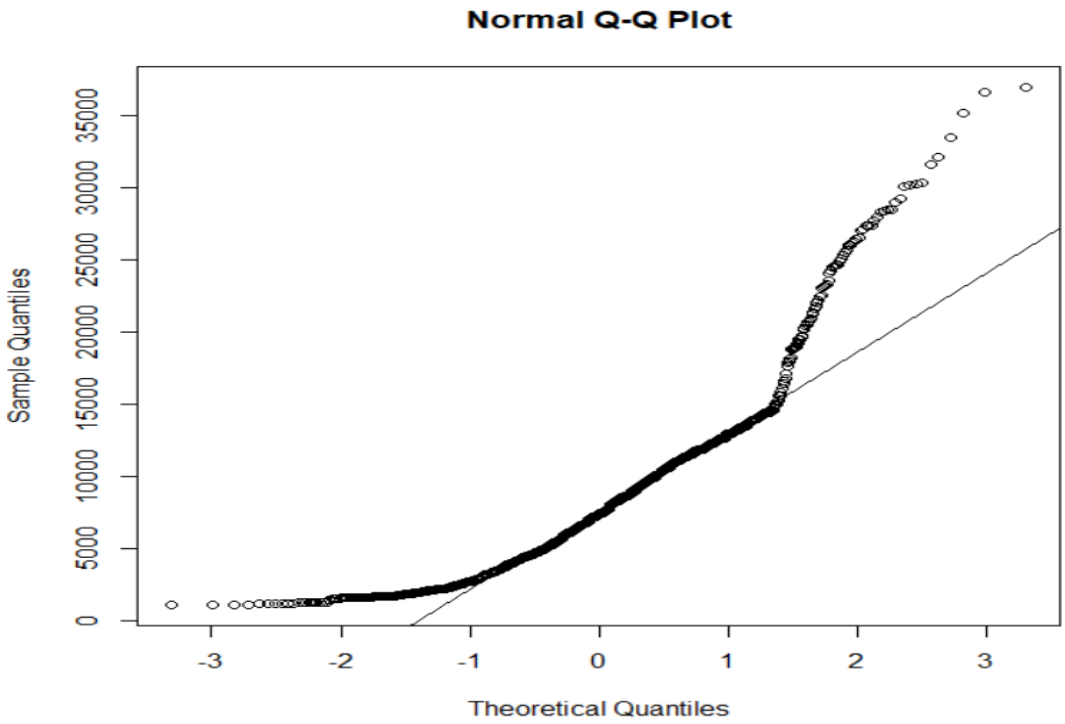


Ilustración 23- Representación normal Q-Q

```
> shapiro.test ( MujeresIni ) # contraste de normalidad

        Shapiro-Wilk normality test

data:  MujeresIni
W = 0.87286, p-value < 2.2e-16
```

Ilustración 24- Shapiro test

Interpretación Mujeres:

La nube de puntos se ordena cerca de la recta. En un principio, visualmente se aprecia que nuestros datos cumplen el supuesto de normalidad.

Con un p-value =2.2e-16, mayor de 0.05, no podemos rechazar la hipótesis nula. Por lo tanto, podemos concluir que **nuestros datos cumplen el supuesto de normalidad**.

Supuesto de homoscedasticidad (homogeneidad de varianzas).

En el contraste de homogeneidad de varianzas la *hipótesis nula* es que la *varianza es constante (no varía) en los diferentes grupos*. Para contrastar podemos utilizamos el test F de Snedecor con var.test( ), que se aplica cuando solo hay dos grupos.

```
> var.test( HombresIni, MujeresIni )

        F test to compare two variances

data:  HombresIni and MujeresIni
F = 3.7079, num df = 273, denom df = 1063, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 3.087603 4.500341
sample estimates:
ratio of variances
          3.707885
```

Ilustración 25- Contraste de homogeneidad de varianzas

Interpretación:
Con un p-value = 2.2e-16, mayor de 0.05, no podemos rechazar la hipótesis nula. Por lo tanto, **suponemos homogeneidad de varianzas**.

Charges - Smoker

¿Debemos aceptar o rechazar la diferencia de la media de los cargos "charges" según si son fumadores o no (Smoker), para α=0.05?

Estamos ante un contraste para **dos muestras independientes** (fumador y no fumador). Para dos muestras independientes se debe comprobar el supuesto de normalidad y el supuesto de homocedasticidad. Después se realiza el contraste sobre lo que queremos probar

- PREPARAMOS NUESTROS DATOS. Creamos SmokerYes solo con los datos del (charges) de fumadores (smoker == "yes"), y creamos SmokerNo solo con los datos del mes inicial (charges) de no fumadores (smoker == "no"). Serán nuestras dos muestras independientes.

```
> smokerYes <- doc_csv$charges[doc_csv$smoker == "yes"]
> smokerNo <- doc_csv$charges[doc_csv$smoker == "no"]
> qqnorm( smokerYes )
> qqline( smokerYes )
>
```

Supuesto de normalidad para fumadores:

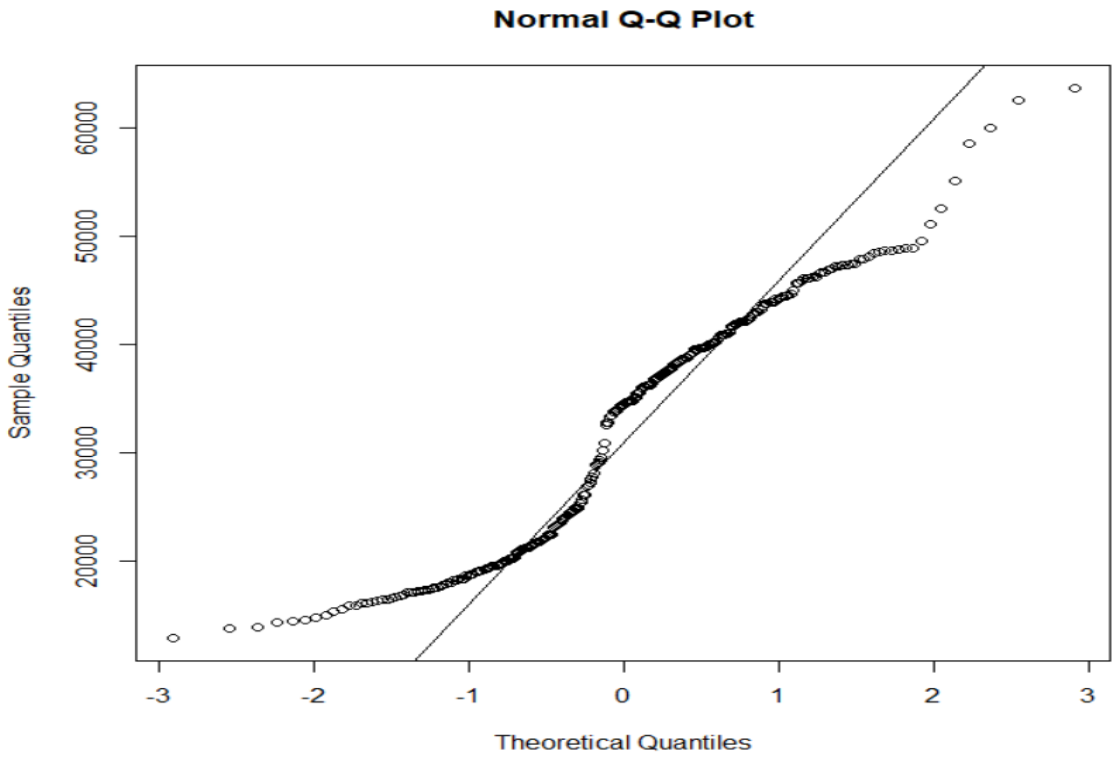


Ilustración 26- Representación normal Q-Q

Interpretación Fumadores:

La nube de puntos se ordena cerca de la recta. En un principio, visualmente se aprecia que nuestros datos cumplen el supuesto de normalidad.

```
> shapiro.test ( smokerYes ) # contraste de normalidad

        Shapiro-Wilk normality test

data:  smokerYes
W = 0.93955, p-value = 3.625e-09
```

Ilustración 27- Shapiro test

Con un p-value =2.2e-16, mayor de 0.05, no podemos rechazar la hipótesis nula. Por lo tanto, podemos concluir que **nuestros datos cumplen el supuesto de normalidad**.

Supuesto de normalidad para No fumadores:

```
> qqnorm( smokerNo )
> qqline( smokerNo )
```

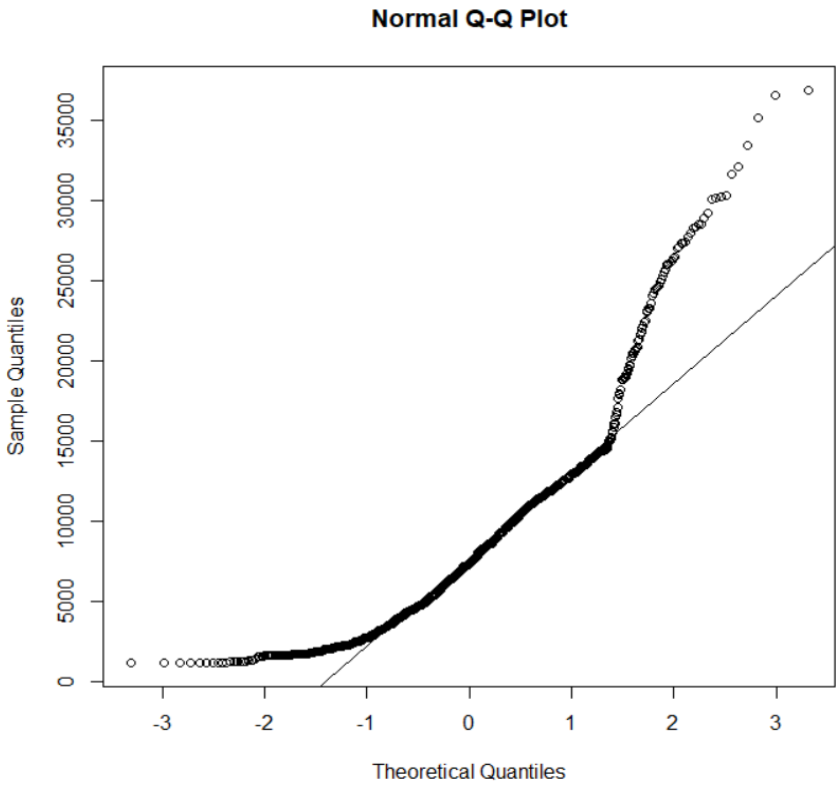


Ilustración 28-Representación normal Q-Q

```
> shapiro.test ( smokerNo ) # contraste de normalidad

        Shapiro-Wilk normality test

data:  smokerNo
W = 0.87286, p-value < 2.2e-16
```

Ilustración 29- Shapiro test

Interpretación No fumadores:

La nube de puntos se ordena cerca de la recta. En un principio, visualmente se aprecia que nuestros datos cumplen el supuesto de normalidad.

Con un p-value =2.2e-16, mayor de 0.05, no podemos rechazar la hipótesis nula. Por lo tanto, podemos concluir que **nuestros datos cumplen el supuesto de normalidad**.

Supuesto de homoscedasticidad (homogeneidad de varianzas).

En el contraste de homogeneidad de varianzas la hipótesis nula es que la varianza es constante (no varía) en los diferentes grupos. Para contrastar podemos utilizamos el test F de Snedecor con var.test( ), que se aplica cuando solo hay dos grupos.

```
> var.test( smokerYes, smokerNo )

        F test to compare two variances

data:  smokerYes and smokerNo
F = 3.7079, num df = 273, denom df = 1063, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 3.087603 4.500341
sample estimates:
ratio of variances
          3.707885
```

Ilustración 30- Contraste de homogeneidad de varianzas

Interpretación:
Con un p-value = 2.2e-16, mayor de 0.05, no podemos rechazar la hipótesis nula. Por lo tanto, **suponemos homogeneidad de varianzas**.

Reglas de Asociación

Vamos a aplicar reglas de asociación a priori para el análisis del dataset. Lanzamos las reglas apriori gracias a una librería, para ello hemos fijado las reglas de confianza y soporte.

Recordamos: "El soporte indica cuantas veces se han encontrado las reglas {lsh => rhs} en el dataset, cuanto más alto mejor. La confianza habla de la probabilidad de que {rhs} se de en función de {lhs}. El lift es un parámetro que nos indica cuánto de aleatoriedad hay en las reglas. Un lift de 1 o menos es que las reglas son completamente fruto del azar."

```
#Correlaciones
library(arules)
trans <- as(doc_csv, "transactions")
rules <- apriori(trans,parameter = list(support = 0.01, confidence = 0.5))
inspect(head(sort(rules, by = "confidence"), 3))
```

Tras aplicar las reglas, las filtramos por confianza y mostramos las tres primeras.

```
> inspect(head(sort(rules, by = "confidence"), 3))
    lhs                             rhs                                 support   confidence coverage  lift     count
[1] {smoker=yes}                 => {charges=[1.28e+04,6.38e+04]} 0.2047833 1          0.2047833 3.000000 274
[2] {charges=[6.27e+03,1.28e+04)} => {smoker=no}                   0.3333333 1          0.3333333 1.257519 446
[3] {charges=[1.12e+03,6.27e+03)} => {smoker=no}                   0.3333333 1          0.3333333 1.257519 446
>
```

Ilustración 31- Reglas de Asociación

Nos centramos en la primera norma, nos llama la atención el valor de lift tan alto que posee, lo que nos hace pensar que no se trata de una norma casual. Analizamos la norma; viene a trasladar que si eres fumador tus gastos médicos se sitúan entre 12800 y 63800 dolares.

A continuación, vienen las otras dos normas con más confianza que asocian ser no fumador con costes médicos más bajos.

Por tanto, quizá se podría afirmar que ser fumador hace que los costos médicos del paciente puedan ser más elevados.

Modelo de regresión lineal múltiple (regresores cuantitativos)

Estimaremos por mínimos cuadrados ordinarios un modelo lineal que explique los cargos (charges) de un individuo en función de tres factores cuantitativos: los años (age), el índice de masa corporal (bmi), y el número de hijos asegurados (children).

```
> Model.1.1<- lm(charges~bmi+children+age, data=clus2 )
> summary(Model.1.1)

Call:
lm(formula = charges ~ bmi + children + age, data = clus2)

Residuals:
   Min     1Q Median     3Q    Max
-13884  -6994  -5092   7125  48627

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13270.4      310.9  42.684  < 2e-16 ***
bmi           2025.1      312.9   6.472 1.35e-10 ***
children       654.4      311.3   2.102   0.0357 *
age           3371.9      313.2  10.767  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11370 on 1334 degrees of freedom
Multiple R-squared:  0.1201,    Adjusted R-squared:  0.1181
F-statistic: 60.69 on 3 and 1334 DF,  p-value: < 2.2e-16
```

El coeficiente de la bondad de ajuste es 0.1201 y el coeficiente ajustado es: 0.1181. Además, se observa que el test global de la regresión es significativo.

Dado que la multicolinealidad entre las variables explicativas es un factor de inestabilidad en la estimación de los coeficientes de regresión. Es interesante explorar la matriz de correlación entre regresores.

```
> is_number <- sapply(clus2,is.numeric)
> a <-cor(clus2[,is_number])
> a
                age       bmi    children     charges
age       1.0000000 0.1092719 0.04246900 0.29900819
bmi       0.1092719 1.0000000 0.01275890 0.19834097
children  0.0424690 0.0127589 1.00000000 0.06799823
charges   0.2990082 0.1983410 0.06799823 1.00000000
```

Modelo de regresión lineal múltiple (regresores cuantitativos y cualitativos)

Ahora se añade las variables Sex y Smoker. Usaremos como categoría de referencia de la variable Sex la categoria “female” y de la variable Smoker la categoría “yes”.

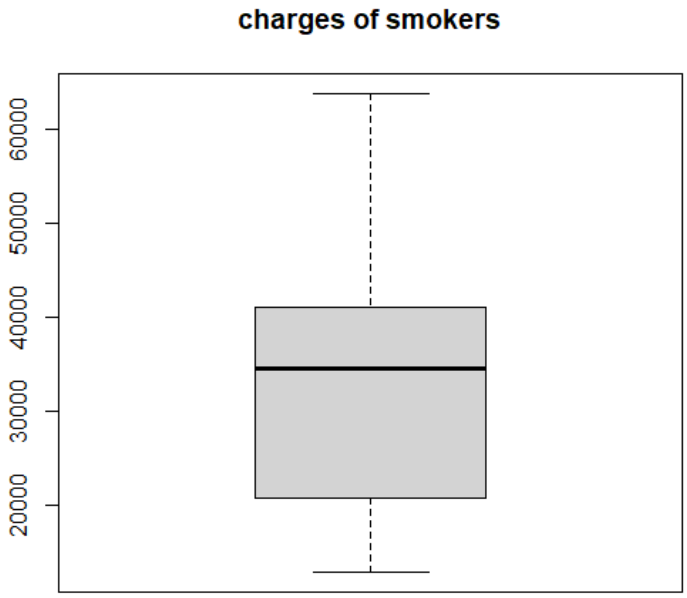
```
> clus2$SexR=relevel(clus2$sex, ref = 'female')
> clus2$SmokerR=relevel(clus2$smoker, ref = 'yes')
> str(clus2)
'data.frame':   1338 obs. of  9 variables:
 $ age     : num  -1.438 -1.509 -0.798 -0.442 -0.513 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi     : num  -0.453 0.509 0.383 -1.305 -0.292 ...
 $ children: num  -0.9083 -0.0787 1.5803 -0.9083 -0.9083 ...
 $ 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 ...
 $ SexR    : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ SmokerR : Factor w/ 2 levels "yes","no": 1 2 2 2 2 2 2 2 2 2 ...
> Model.1.2<- lm(charges~bmi+children+age+SmokerR+SexR, data=clus2 )
> summary(Model.1.2)

Call:
lm(formula = charges ~ bmi + children + age + SmokerR + SexR,
    data = clus2)

Residuals:
     Min       1Q   Median       3Q      Max
-11837.2  -2916.7   -994.2   1375.3  29565.5

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  32280.2      414.5  77.871  < 2e-16 ***
bmi           1965.8      167.2  11.757  < 2e-16 ***
children       571.9      166.2   3.441 0.000597 ***
age           3621.2      167.2  21.651  < 2e-16 ***
SmokerRno   -23823.4      412.5 -57.750  < 2e-16 ***
SexRmale      -128.6      333.4  -0.386 0.699641
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6070 on 1332 degrees of freedom
Multiple R-squared:  0.7497,    Adjusted R-squared:  0.7488
F-statistic:   798 on 5 and 1332 DF,  p-value: < 2.2e-16
```

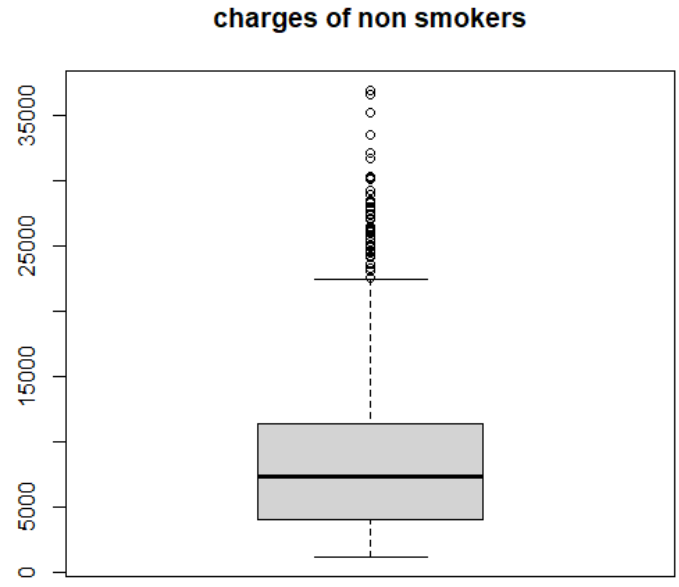
El coeficiente de la bondad de ajuste del primer modelo es 0.1201 y del segundo es 0.7497. Por tanto, el mejor modelo es el que tiene un coeficiente ajustado superior. Dado que el segundo modelo es mejor podemos concluir que las variables SexR y SmokerR introducen ciertas diferencias en el modelo predictivo.

Por otra parte, han sido significativos los test parciales sobre los coeficientes de los regresores bmi, children, age, SmokerRno y SexRmale. Siendo las estimaciones de sus coeficientes 1965.8, 571.9, 3621.2, -23823.4, -128.6. El signo negativo en los coeficientes indica que dichos coeficientes tienen un efecto de disminución de los cargos (charges), en cambio los coeficientes con signo positivo indican un efecto incrementador de la variable Charges.

5. Representación de los resultados

Hemos optado por representar todos los valores de los costos médicos con un diagrama de bigotes, separando en fumadores y no fumadores. **Así representamos uno de los principales resultados de este estudio**; los costos médicos de un paciente serán mayores si este es fumador.





6. Conclusiones.

En este estudio hemos centrado nuestros esfuerzos en observar y analizar datos referentes a pacientes médicos; en particular hemos analizado como condiciones de su propia persona; en especial su sexo o si fuma o no, influyen en los costos médicos que este paciente produce; intentando generalizar estos datos no solo con un paciente sino a traves de una muestra de ellos. Obteniendo y buscando obtener por tanto conclusiones de amplio espectro.

Hemos podido saber cómo influyen ciertas variables en los costos médicos. Afirmando que fumar eleva claramente estos gastos. También se han estudiado la influencia de otras variables como el sexo.

De forma contundente se puede afirmar que los resultados de las regresiones y de las normas de asociación dan resultados que pueden ser aplicados directamente por los aseguradores para modelar mejor la prima de sus afiliados.

7 Bibliografía

- **https://www.kaggle.com/mirichoi0218/insurance?select=insurance.csv**

- Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.

Contribuciones	Firma
Investigación previa	Luis Miguel Moreno López Alejandra Cristina Marrero Suarez
Redacción respuestas	Luis Miguel Moreno López Alejandra Cristina Marrero Suarez
Desarrollo código	Luis Miguel Moreno López Alejandra Cristina Marrero Suarez