Práctica 2 Tipología y ciclo de vida de los datos

Carlos Herrera Carballo

27 de mayo, 2021

Contents

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

La crisis sanitaria mundial producida por el COVID-19 ha provocado un incremento sin precedentes en el número de contrataciones de seguros médicos privados por parte de la población. Si particularizamos este hecho en España, se tiene que en 2020 hay 469.750 asegurados más que en 2019 [1]. De este modo, en este contexto social surge la necesidad de desarrollar herramientas que, por medio del uso de la Ciencia de Datos, faciliten a la población la toma de decisiones a la hora de escoger un seguro de salud privado y que dicha decisión sea informada con resultados precisos. Por lo tanto, el dataset escogido para esta actividad tiene relevada importancia, pues constituye una colección de datos sociodemográficos de ciudadanos estadounidenses a partir de los cuáles y, con el uso de distintas pruebas estadísticas, nos permitirán conocer la relación de estas variables con el costo final de este tipo de seguros privados. Además, también podremos explorar las diferencias en el precio final del seguro en función del sexo de la persona asegurada, así como de la zona de residencia del mismo.

El conjunto de datos ha sido extraído de la página web www.kaggle.com [2] y está conformado por un total de 1338 observaciones y las 7 variables que definimos a continuación:

- **Age**: Edad del beneficiario principal del seguro médico contratado.
- **Sex**: Género del beneficiario principal del seguro médico contratado. (Female, Male)
- **bmi**: Índice de masa coporal del beneficiario principal del seguro médico contratado.
- **Children**: Número de niños del beneficiario principal del seguro médico contratado cubiertos por dicho seguro.
- **Smoker**: Condición de fumador o no fumador del beneficiario principal del seguro médico contratado. (Yes, No)
- **Region**: Área residencial del beneficiario principal del seguro médico contratado en EEUU. (northeast, southeast, southwest, northwest)
- **Charges**: Costes médicos anuales e individuales facturados por el seguro médico contratado.

Referencias del apartado 1:

[**1**] https://www.infolibre.es/noticias/politica/2021/04/29/los_seguros_privados_rompen_sus_records_2020_superan_los_millones_clientes_los_000_millones_facturacion_118757_1012.html

[**2**] https://www.kaggle.com/mirichoi0218/insurance

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

En primer lugar, comenzamos cargando el fichero de datos cuya extensión es de tipo csv, creando así un dataset con las variables de interés. En este caso, seleccionaremos las siete variables que componen este conjunto de datos pues todas serán necesarias para las pruebas estadísticas que realizaremos posteriormente.

```
# Carga del conjunto de datos escogido.
dataset <- read.csv("insurance.csv",
                    stringsAsFactors = FALSE,
                    fileEncoding = "UTF-8",
                     sep = ",")
```

En segundo lugar, procedemos a examinar el tipo de datos con los que R ha interpretado cada variable. En este sentido, observamos que tenemos tres variables de tipo caracter (sex, region y smoker) y cuatro variables numéricas (dos discretas: age y children y dos continuas: bmi y charges).

```
# Visualización de la estructura del dataset.
str(dataset)
```

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

Para finalizar este apartado, extraemos una estadística descriptiva de nuestro conjunto de datos para obtener una primera descripción de los datos cargados:

```
# Estadística descriptiva del dataset.
summary(dataset)
```

```
##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   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
##     smoker             region             charges
##  Length:1338        Length:1338        Min.   : 1122
##  Class :character   Class :character   1st Qu.: 4740
##  Mode  :character   Mode  :character   Median : 9382
##                                        Mean   :13270
##                                        3rd Qu.:16640
##                                        Max.   :63770
```

```
# Variables no numéricas
table(dataset$sex)
```

```
##
## female   male
##    662    676
```

```
table(dataset$smoker)
```

```
##
##   no  yes
## 1064  274
```

```
table(dataset$region)
```

```
##
## northeast northwest southeast southwest
##       324       325       364       325
```

De este modo, podemos ver que estamos tratando con datos de personas adultas, puesto que el rango de edades comprendidas en la muestra van desde los 18 hasta los 64 años. Asimismo, el número de hombres y mujeres en la muestra es similar, pues contamos con un total de 662 mujeres y 676 hombres. Además, también son similares la cantidad de individuos para cada región. No obstante, encontramos mayor diferencia entre el número de personas no fumadoras (1064) y el número de personas si fumadoras (274).

Por otra parte, la variable children tiene como valor máximo 5 (5 hijos) y mínimo 0 (asuencia de hijos por parte del asegurado), mientras que el índice de masa corporal tiene unos valores comprendidos entre 15,96 y 53,13. Finalmente, comentar que el seguro más barato de la muestra tiene un costo de 1122 dólares, mientras que el más caro es de 63770 dólares.

3. Limpieza de los datos.

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

Primero, verifiquemos si existen elementos vacíos en alguna de las variables de nuestro dataset:

```
# Estadísticas de valores vacíos
colSums(is.na(dataset))
```

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

Como podemos ver, no existen elementos vacíos en el conjunto de datos. No obstante, vamos a describir cómo actuaríamos en el caso de que existiesen.

En el caso de tener valores vacíos en las variables de texto (smoker, region y sex) una posible imputación de estos valores sería sustituir el valor faltante por la moda de la variable en cuestión. Asimismo, en el caso de variables discretas, tales como la edad y el número de hijos, se puede proceder sustituyendo los valores faltantes por la mediana de estas variables. Finalmente, en el caso de las variables continuas (bmi y charges) se puede proceder sustituyendo los valores vacíos por las medias aritméticas de dichas variables.

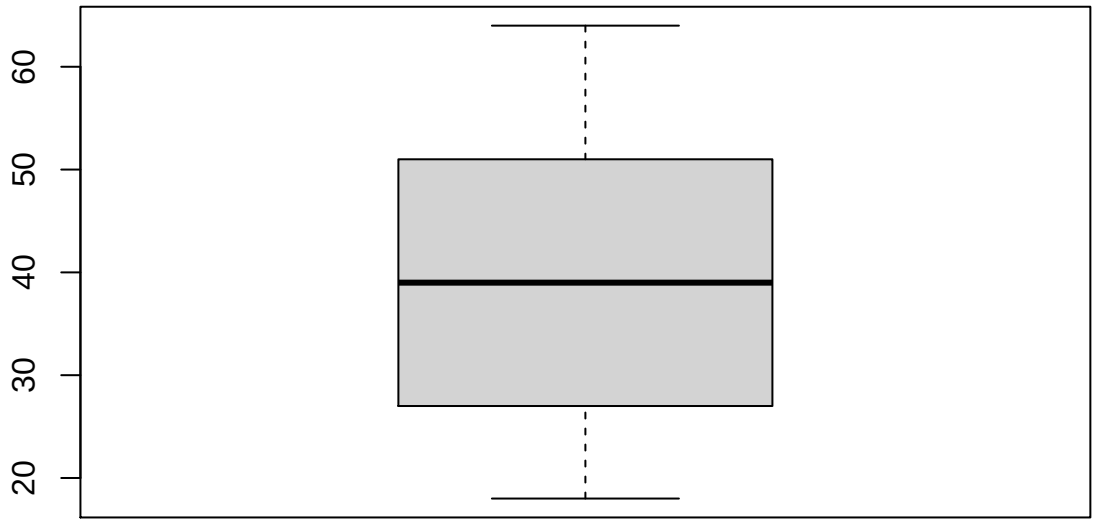
Por otra parte, dada la naturaleza de este conjunto de datos en el que hay presencia de variables mixtas (numéricas y de texto), aparte de la sustitución por medidas centrales que acabamos de mencionar, también podemos imputar los valores perdidos con técnicas más sofisticadas como podría ser el algoritmo missForest, ya que resulta un método apto y robusto para la tipología del conjunto de datos usado.

3.2. Identificación y tratamiento de valores extremos.

A continuación, llevaremos a cabo la identificación de valores extremos en cada una de las variables suceptibles a la presencia de valores atípicos. Para ello, realizaremos gráficos de cajas para localizar estos valores atípicos para luego debatir qué tipo de tratamiento es el más adecuado (en caso de que dichos valores necesiten tratramiento).

En primer lugar, observamos que la variable edad no tiene ningún valor outlier.

```
# Boxplot variable age
age.bp <- boxplot(dataset$age)
```

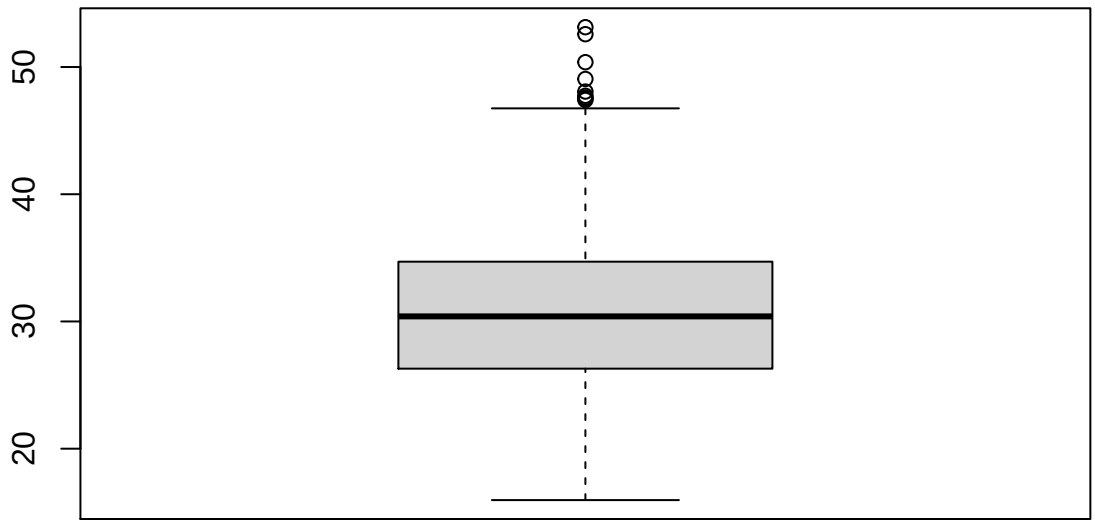


En segundo lugar, observamos que la variable bmi (índice de masa corporal) presenta un total de 9 valores extremos que superan el valor 40, a partir del cual, se considera que un paciente sufre obesidad mórbida. En este sentido, estos valores pueden ser perfectamente legítimos en el caso de que estos individuos padezcan

esta enfermedad, además, debemos tener en cuenta que se trata de una muestra tomada en EEUU donde los índices de obesidad en la población son altos y, también, las características genéticas de los habitantes pueden ser perfectamente compatibles con estos valores en el índice de masa corporal.

Por otra parte, también debemos considerar que esta afectación trae consigo otras complicaciones de salud (enfermedades cardíacas, accidentes cerebrovasculares, diabetes, etc.) para el paciente que tienen influencia en el precio final de un posible seguro médico que quiera contratar. Por estos motivos, tomaremos la decisión de mantener estos valores extremos en el conjunto de datos a estudiar, pues pueden resultar valiosos para conocer el impacto de esta condición en el precio final de un seguro privado de atención sanitaria.

```
# Boxplot variable bmi
bmi.bp <- boxplot(dataset$bmi)
```

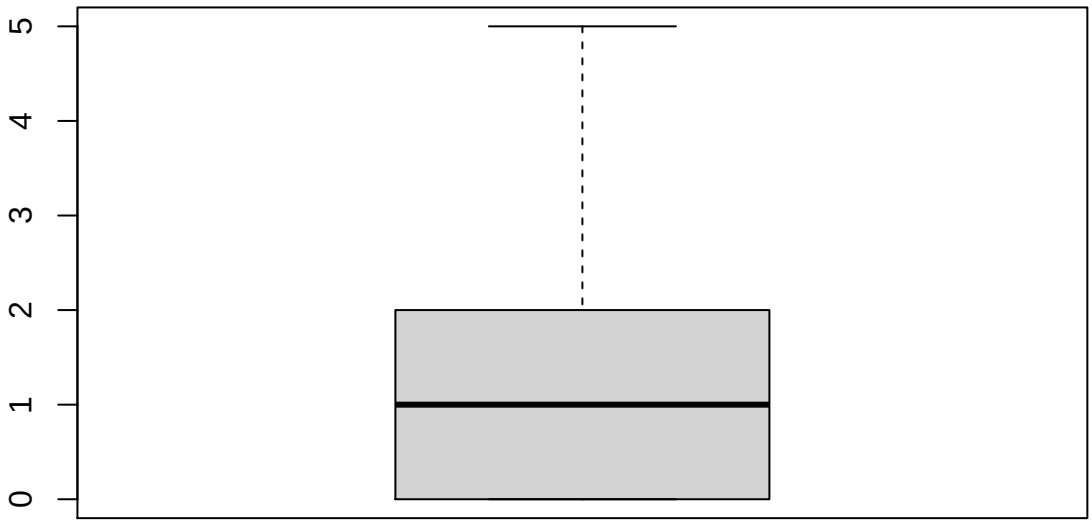



```
bmi.bp$out
```

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

En cuanto al número de hijos, tampoco encontramos ninguna observación que resulte atípica o extrema.

```
# Boxplot variable children
children.bp <- boxplot(dataset$children)
```

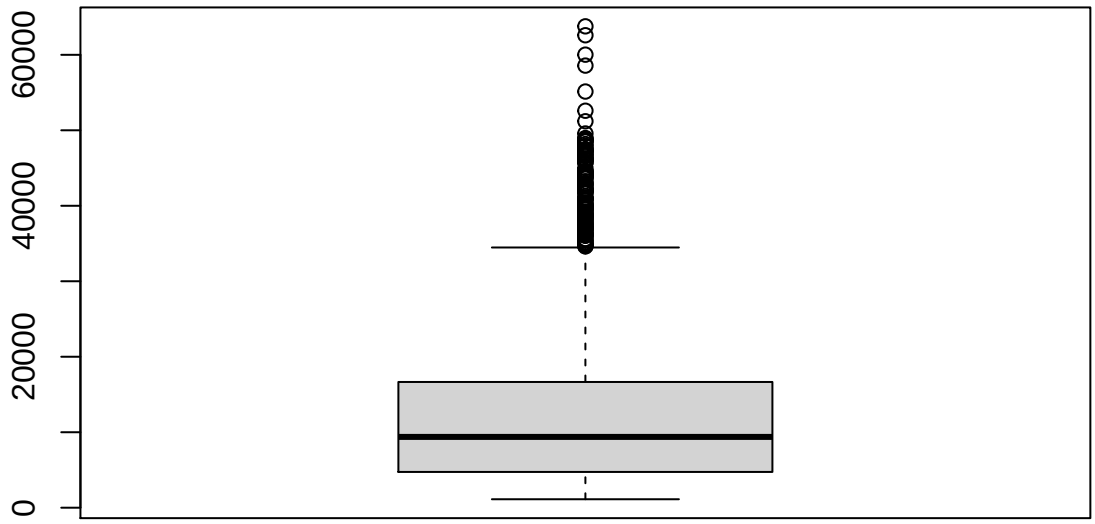


Por último, en la variable charges encontramos un total de 139 valores outliers, esto supone más de un 10% del total de observaciones (1338), luego no contemplamos la eliminación de estos registros como un procedimiento adecuado para este caso. En su lugar, hemos llevado a cabo una investigación para conocer el coste promedio (individual y familiar) de un seguro médico en Estados Unidos. En este sentido, se tiene que, en el transcurso de un año, el gasto promedio en seguro médico para una familia de cuatro miembros en los EEUU Fue de 25011 dólares en 2020 [3], mientras que este mismo gasto para una persona es de 11582 dólares [4]. Sin embargo, estas cuantías varían en función de la edad y otras condiciones del contratante del seguro.

En consecuencia de lo anterior y observando que la variable charges es asimétrica hacia la izquierda (ver histograma a continuación), en nuestra investigación tomamos la decisión de considerar como valor atípico aquellos cargos que superen la cuantía de 30000 dólares, pues entendemos que casos extremos como un cargo total de 63770.43 dólares (máximo registrado para esta variable) son situaciones inusuales que no corresponden con una tarifa promedio de este tipo de seguros y que puede deberse a numerosas particularidades del contratante en sí, de modo que, si incluímos este tipo de valores en nuestras pruebas estadísticas, los resultados se verán altamente influenciados y sesgados. Por lo tanto, procedemos a imputar aquellos valores superiores a 30000 dólares por la mediana de la variable.

Destacar que la elección de esta medida de tendencia central para la imputación se debe a la afectación que sufre la media de charges (13270.42) frente a la presencia de estos valores extremos. Por este motivo, creemos más conveniente imputar los valores outliers a través de la mediana (9382.033).

```
# Boxplot variable Charges
charges.bp <- boxplot(dataset$charges)
```



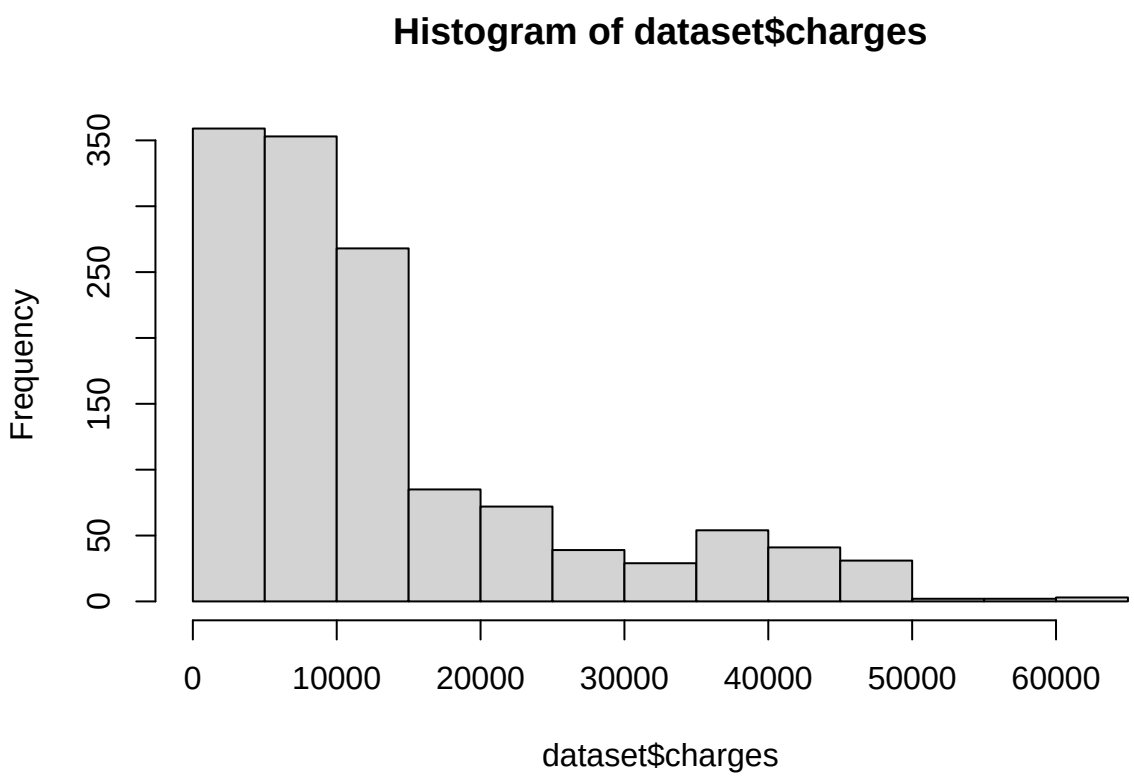

```
charges.bp$out
```

```
##   [1] 39611.76 36837.47 37701.88 38711.00 35585.58 51194.56 39774.28 48173.36
##   [9] 38709.18 37742.58 47496.49 37165.16 39836.52 43578.94 47291.06 47055.53
##  [17] 39556.49 40720.55 36950.26 36149.48 48824.45 43753.34 37133.90 34779.61
##  [25] 38511.63 35160.13 47305.31 44260.75 41097.16 43921.18 36219.41 46151.12
##  [33] 42856.84 48549.18 47896.79 42112.24 38746.36 42124.52 34838.87 35491.64
##  [41] 42760.50 47928.03 48517.56 41919.10 36085.22 38126.25 42303.69 46889.26
##  [49] 46599.11 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 42969.85
##  [65] 40182.25 34617.84 42983.46 42560.43 40003.33 45710.21 46200.99 46130.53
##  [73] 40103.89 34806.47 40273.65 44400.41 40932.43 40419.02 36189.10 44585.46
##  [81] 43254.42 36307.80 38792.69 55135.40 43813.87 39597.41 36021.01 45008.96
##  [89] 37270.15 42111.66 40974.16 46113.51 46255.11 44202.65 48673.56 35069.37
##  [97] 39047.29 47462.89 38998.55 41999.52 41034.22 36580.28 35595.59 42211.14
## [105] 44423.80 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 36898.73
## [121] 52590.83 40941.29 39722.75 37465.34 36910.61 38415.47 41661.60 60021.40
## [129] 47269.85 49577.66 37607.53 47403.88 38344.57 34828.65 62592.87 46718.16
## [137] 37829.72 36397.58 43896.38
```

```
# Número de outliers
length(c(charges.bp$out))
```

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

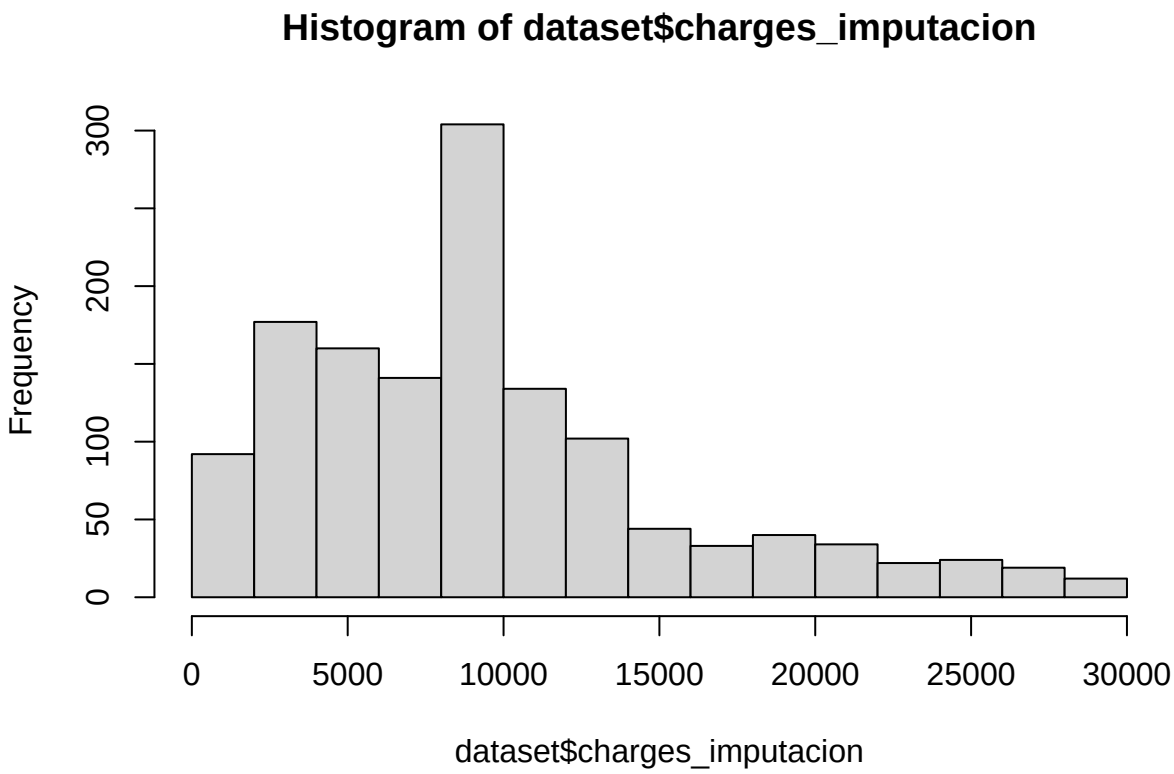
```
# Histograma para estudiar la distribución de valores de charges.
hist(dataset$charges)
```



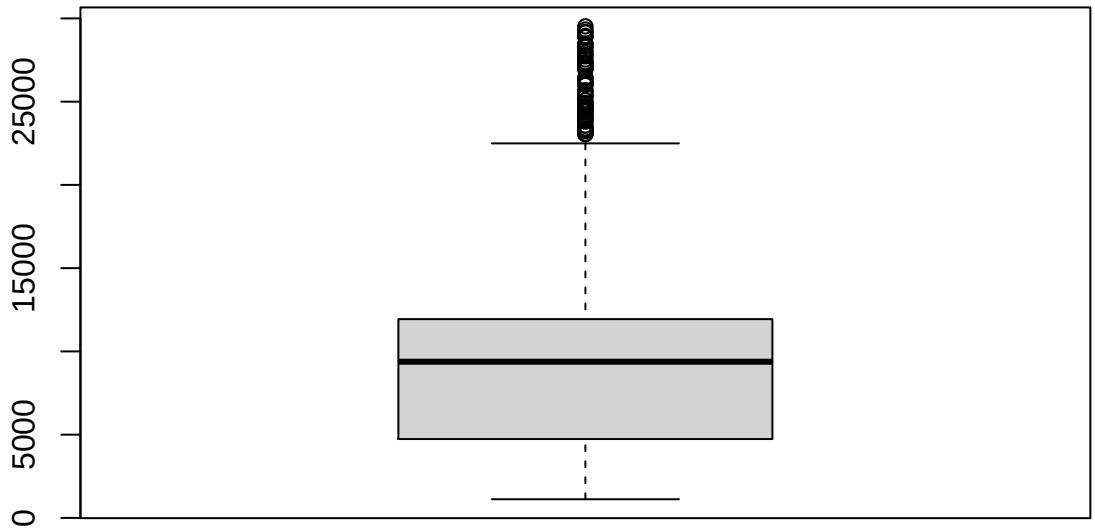
Así, procedemos a la imputación de estos valores extremos creando una nueva variable charges con el tratamiento de dichos outliers:

```
# Imputación de outliers por la mediana.
dataset <- dataset %>%
mutate(charges_imputacion = case_when(charges > 30000 ~ median(dataset$charges),
                                      charges <= 30000 ~ charges))

# Comprobación de la imputación.
hist(dataset$charges_imputacion)
```




```
length(c(boxplot(dataset$charges_imputacion)$out))
```




```
## [1] 68
```

Referencias del apartado 2:

[**3**] eHealthInsurance.com. "ACA Index Report on Unsubsidized Consumers in the 2020 Open Enrollment Period." Consultada en Mayo. 17, 2021.

[**4**] CMS.gov. "National Health Expenditure Data, Historical." Consultada en Mayo. 17, 2021.

4. Análisis de los datos.

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

Llevaremos a cabo un total de tres análisis estadísticos distintos. En primer lugar, realizaremos un contraste de hipótesis con el objetivo de conocer si la media de los costes del seguro médico (variable charges) para las mujeres es significativamente mayor que la de los hombres. En segundo lugar, también estudiaremos si existe diferencias de medias para la variable charges en función de la zona de residencia del contratante del seguro (variable region) a través de un ANOVA. Por último, trataremos de ajustar un modelo de regresión lineal que nos ayude a predecir el costo de un seguro médico (charges) en función de las características demográficas de la persona en cuestión (variables age, sex, smoker, bmi y children).

De este modo, para la regresión y el ANOVA no debemos hacer una selección previa, pues a la hora de aplicar estos métodos estadísticos nombraremos directamente las variables necesarias para ello. No obstante, en el caso del contraste de hipótesis, debemos separar nuestra muestra original en dos submuestras: una de mujeres y la otra de hombres. Por lo tanto, procedemos a realizar la selección de estas submuestras:

```
# Creación de submuestras para contraste de hipótesis.
mujeres <- filter(dataset, sex == "female")
hombres <-  filter(dataset, sex == "male")
```

Otra acción que debemos realizar para llevar a cabo nuestro análisis de regresión será recodificar las variables cualitativas region, sex y smoker. Las recodificaciones a realizar serán las siguientes:

sex: 0 = "male", 1 = "female". smoker: 0 = "no", 1 = "yes". region: 1 = "northeast". 2 = "southeast", 3 = "southwest", 4 = "northwest".

Por tanto, procedemos a ello y a su posterior etiquetado:

```
library(car)

# Recodificación sex
dataset$sex <- factor(case_when(
                 dataset$sex == "male" ~ 0,
                 dataset$sex == "female" ~ 1,
                  ))

levels(dataset$sex) <- c("male", "female")

# Recodificación smoker
dataset$smoker <- factor(case_when(
                 dataset$smoker == "no" ~ 0,
                 dataset$smoker == "yes" ~ 1,
                  ))
levels(dataset$smoker) <- c("no", "yes")

# Recodificación region
dataset$region <- factor(case_when(
                        dataset$region == "northeast" ~ 1,
                        dataset$region == "southeast" ~ 2,
                        dataset$region == "southwest" ~ 3,
                        dataset$region == "northwest" ~ 4))
levels(dataset$region) <- c("northeast", "southeast", "southwest", "northwest")
```

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

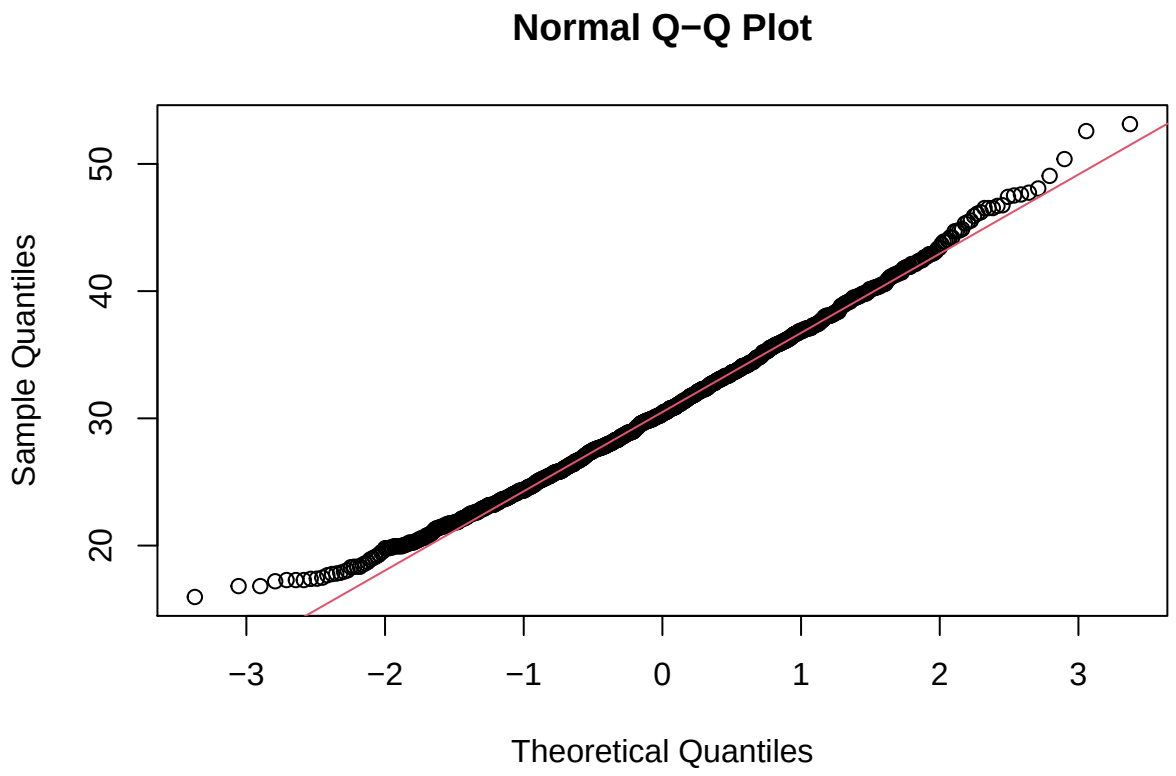
Antes de aplicar las pruebas estadísticas mencionadas en el apartado anterior, realizaremos un estudio de la normalidad y homogeneidad de la varianza de las variables cuantitativas. Esto, nos ayudará a también a saber si las pruebas estadísticas planteadas pueden llevarse a cabo.

En primer lugar, estudiaremos la normalidad de las variables cuantitativas a través del test de Kolmogorov-Smirnov (con la corrección Lilliefors), ya que contamos con un número de observaciones mayor a 50, y la visualización de gráficos QQ:

```
# Test Kolmogorov-Smirnov (con la corrección Lilliefors) bmi
lillie.test(dataset$bmi)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataset$bmi
## D = 0.0261, p-value = 0.03272
```

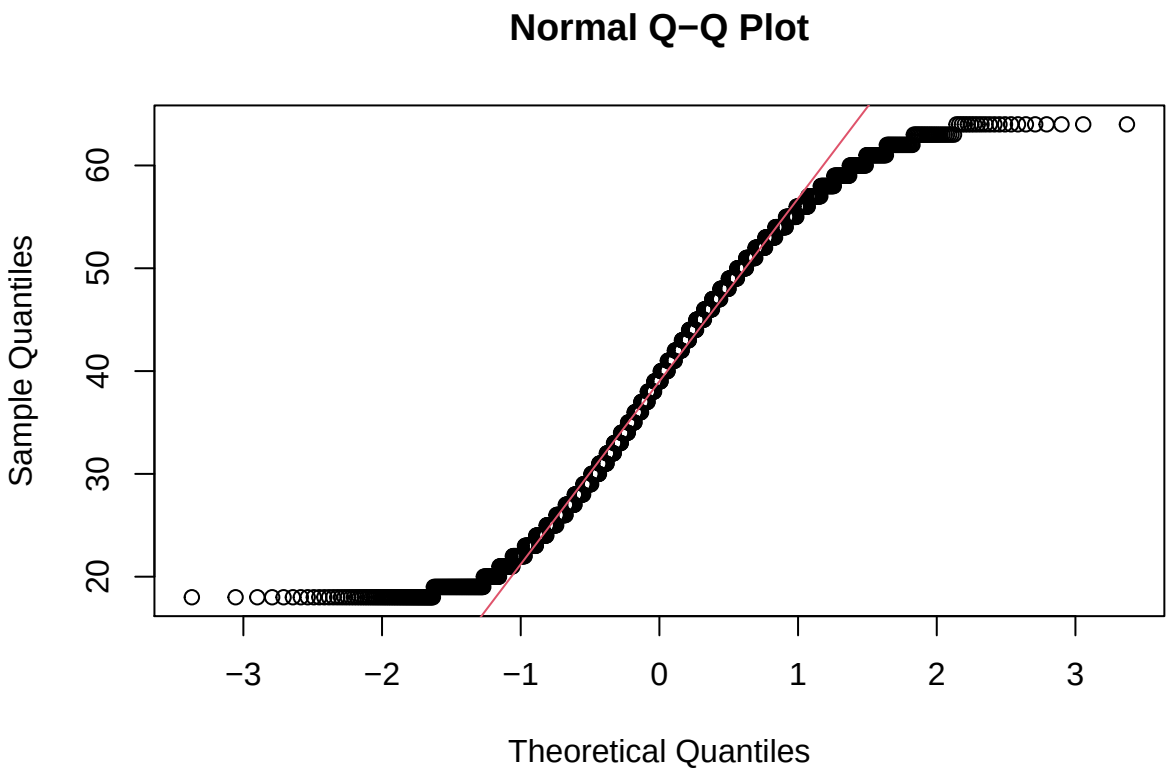
```r
# QQ plot bmi
qqnorm(dataset$bmi)
qqline(dataset$bmi, col = 2)
```




```r
# Test Kolmogorov-Smirnov (con la corrección Lilliefors) age
lillie.test(dataset$age)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataset$age
## D = 0.078945, p-value < 2.2e-16
```

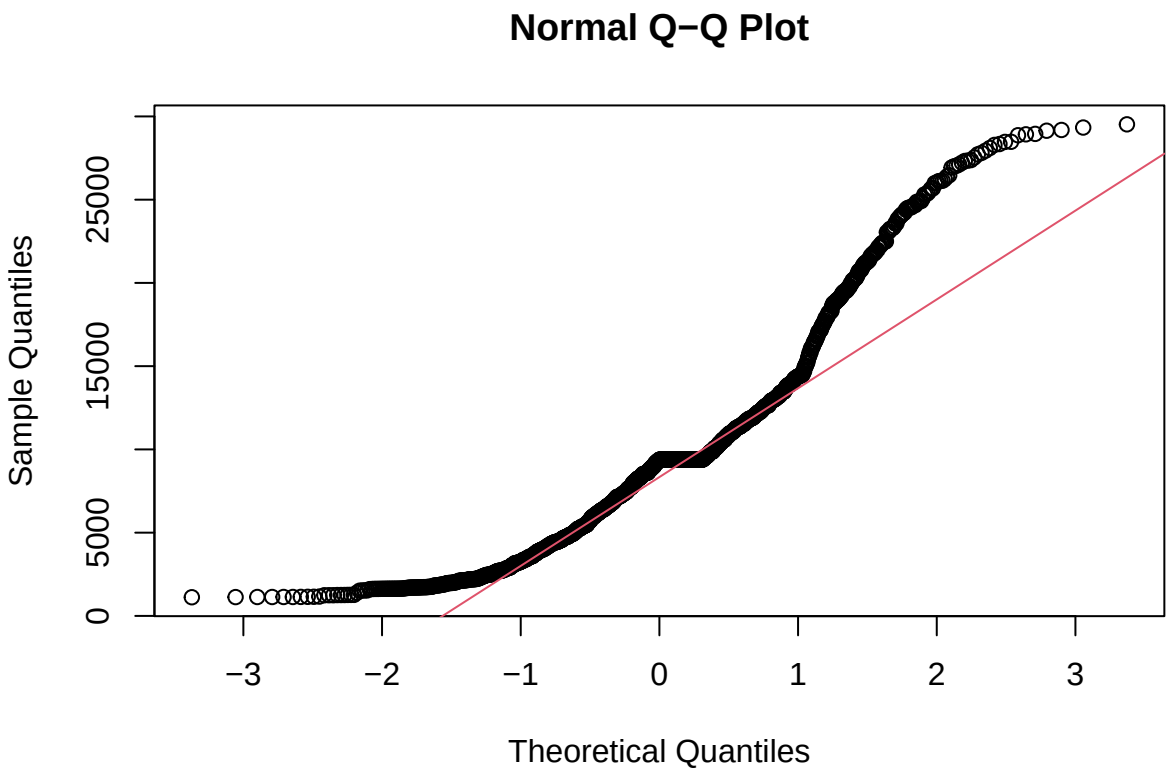
```r
# QQ plot age
qqnorm(dataset$age)
qqline(dataset$age, col = 2)
```




```
# Test Kolmogorov-Smirnov (con la corrección Lilliefors) charges
lillie.test(dataset$charges_imputacion)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dataset$charges_imputacion
## D = 0.12785, p-value < 2.2e-16
```

```
# QQ plot charges
qqnorm(dataset$charges_imputacion)
qqline(dataset$charges_imputacion, col = 2)
```



A partir de los resultados gráficos y de los test calculados, es claro que las variables age y charges_imputation carecen de normalidad, pues los test de Kolmogorov-Smirnov han resultado significativos con p-valores menores al nivel de significación impuesto (0.05). Además, en los QQ plots para estas variables se detecta un claro desvío con respecto a un comportamiento normal, siendo este desvío más acusado en los extremos de dichos gráficos. Por otra parte, la variable bmi carece de normalidad según el test de Kolmogorov-Smirnov, ya que el p-valor obtenido (0.03272) nos permite rechazar la hipótesis nula. Sin embargo, en la evaluación gráfica del QQ-plot correspondiente a esta variable, observamos que la mayoría de los cuantiles muestrales de esta variable son muy similares a los teóricos de la distribución normal, aunque en los extremos algunas observaciones se desvían del carácter normal. Pese a los resultados del test, consideraremos que la variable bmi es de carácter normal y entenderemos que el resultado del test puede haberse visto influenciado por los valores outliers no tratados en el apartado anterior. Así, concluímos que las variables age y charges_imputation carecen de normalidad, mientras que la variable bmi se distribuye normalmente.

Llegados a este punto, debemos realizar una correción en el planteamiento de los análisis hechos anteriormente dada la condición de no normalidad de la variable charges_imputation. En este sentido, como esta variable continua charges_imputation carece de normalidad, esto no nos permite cumplir uno de las hipótesis a tener en cuenta en un estudio ANOVA, por lo tanto, esta prueba estadística no puede llevarse a cabo. En su lugar, procederemos a llevar a cabo una prueba estadística no paramétrica, en particular, emplearemos el test de Kruskal-Wallis (equivalente no paramétrico del ANOVA).

En segundo lugar, y con motivo de la no normalidad de las variables anteriores, procedemos a evaluar la homogeneidad de la varianza de las variables a través del test de Levene. Este test es más robusto y menos sensible cuando los datos se alejan de una distribución normal. Por ello, procedemos a usarlo para estudiar la homogeneidad de la varianza de las variables tomando como factor de agrupamiento la variable region, es decir, vamos a comprobar la homogeneidad de las varianzas entre las distintas regiones de procedencia de los individuos muestrales:

```r
# Test de Levene para charges
leveneTest(dataset$charges_imputacion, group = dataset$region)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    3  1.8522 0.1359
##       1334
```

```r
# Test de Levene para age
leveneTest(dataset$age, group = dataset$region)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    3  0.0153 0.9974
##       1334
```

```r
# Test de Levene para bmi
leveneTest(dataset$bmi, group = dataset$region)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    3  6.2004 0.0003502 ***
##       1334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En base a los resultados obtenidos anteriormente, hemos detectado homocedasticidad en las variables charges_imputation y age pues los test de Levene para ambas variables han sido no significativos, esto es que el p-valor obtenido en ambos test es mayor al nivel de significación impuesto (0.05) y, por tanto, no podemos rechazar la hipótesis nula. Por el contrario, para la variable bmi hemos detectado heterocedasticidad pues, en este caso, el test de Levene es significativo y rechazamos la hipótesis nula confirmando que se encuentran diferencias significativas en las varianzas de la variable para los distintos grupos establecidos.

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. Aplicar al menos tres métodos de análisis diferentes

1. Contraste de hipótesis.

Comenzaremos este apartado del informe llevando a cabo un contraste de hipótesis que nos permita responder a la siguiente pregunta: ¿la media de los costes del seguro médico para las mujeres es significativamente mayor que la de los hombres?

Procedemos a realizar un contraste de hipótesis asumiendo las siguientes hipótesis nula e hipótesis alternativa:

$$H_0 : \mu_1 = \mu_2$$
$$H_0 : \mu_1 > \mu_2$$

siendo μ_1 la media de la variable charges_imputation en la población de mujeres y μ_2 la media de esta misma variable pero en la población de hombres.

El tipo de contraste que vamos a realizar es un contraste para dos muestras independientes. Además, aunque la variable charges_imputation no es una variable normal, debemos tener en cuenta que contamos con dos muestras cuyo número de individuos es mayor a 30 (662 para mujeres y 676 para hombres), por lo tanto, el Teorema Central del Límite nos permite asumir normalidad en la distribución de la media muestral de la variable charges_imputation. A su vez, se asumirán desconocidas las desviaciones típicas poblacionales y se

asume homocedasticidad, puesto que el test de Levene planteado a continuación resulta ser no significativo y, por ende, no podemos rechazar la hipótesis nula, es decir, no existen diferencias entre las varianzas para la muestra de mujeres y hombres.

```
# Test de Levene para bmi
leveneTest(dataset$charges_imputacion, group = dataset$sex)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    1  1.1259 0.2888
##       1336
```

Por lo tanto, procedemos a calcular el contraste de hipótesis unilateral planteado:

```
# Contraste de hipótesis
t.test( mujeres$charges_imputacion, hombres$charges_imputacion, paired=FALSE, alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  mujeres$charges_imputacion and hombres$charges_imputacion
## t = 1.4802, df = 1332.1, p-value = 0.06952
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -55.78862       Inf
## sample estimates:
## mean of x mean of y
##  9722.824  9224.664
```

```
test_hip <- t.test( mujeres$charges_imputacion, hombres$charges_imputacion, paired=FALSE, alternative="
```

Dado que se obtiene un p-valor de 0.06952, es decir, un valor mayor que 0.05, siendo 0.05 el nivel de significación impuesto, se tiene que no podemos rechazar la hipótesis nula. En consecuencia, podemos afirmar, con un nivel de confianza del 95%, que no es cierto que las mujeres tienen un coste medio mayor del seguro médico en comparación con los hombres.

2 Test de Kruskal-Wallis.

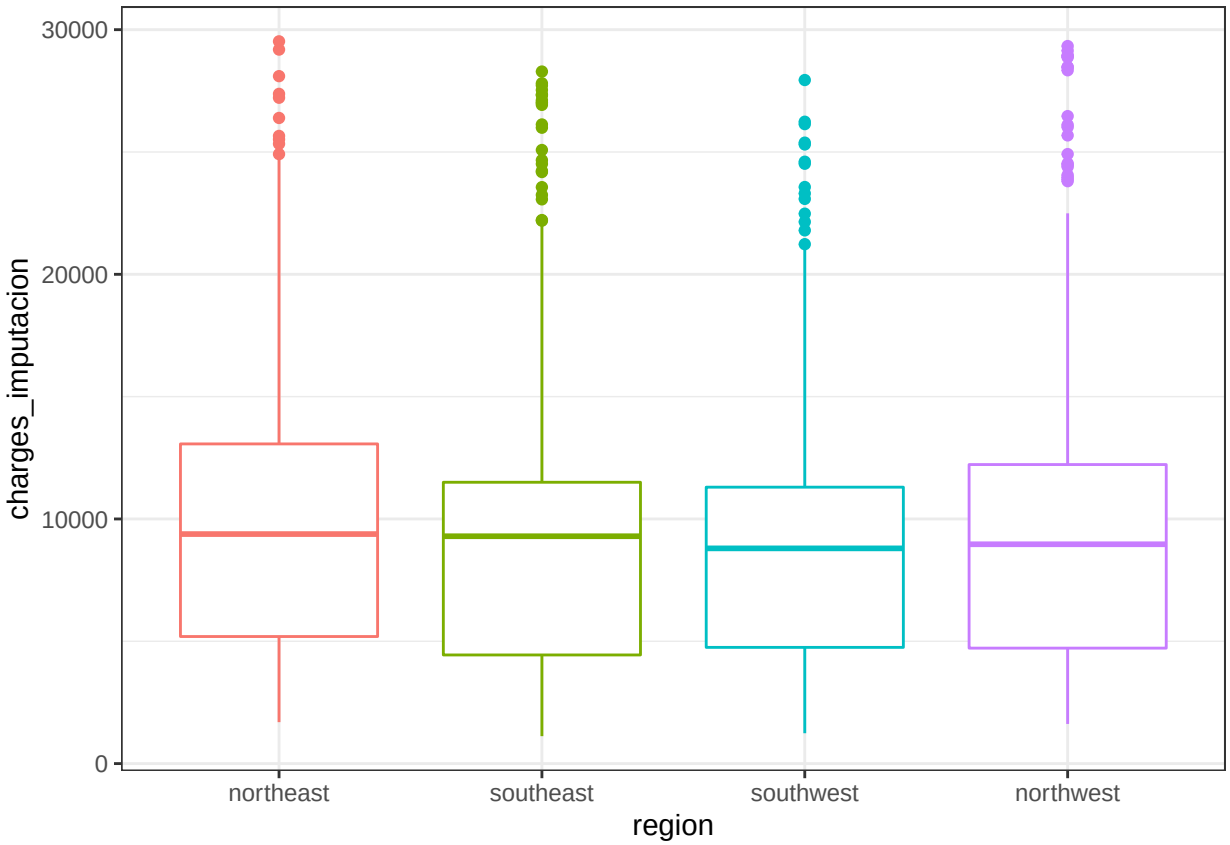
Como comentamos en apartados anteriores, la segunda prueba que realizaremos será un Test de Kruskal-Wallis en sustitución del ANOVA planteado inicialmente. Esto se debe a la falta de normalidad de la variable charges_imputacion, lo cual, nos obliga a realizar una prueba no paramétrica como esta. El objetivo que nos planteamos con esta prueba es detectar si existen diferencias en la mediana de la variable charges_imputacion entre las distintas regiones existentes en nuestra población. Con ello, lo que tratamos de estudiar es si el costo de un seguro médico privado difiere en función del lugar de residencia del contratante de dicho servicio.

Dentro de las condiciones necesarias que deben cumplirse para llevar a cabo el Test de Kruskal-Wallis, ya hemos verificado la existencia de homocedasticidad de la variable charges_imputacion en los distintos grupos que conforma la variable region (northeast, northwest, southeast y southwest). Este hecho podemos intuirlo también reflejado en el siguiente diagrama de cajas a partir del tamaño de la caja y bigotes:

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.4
```

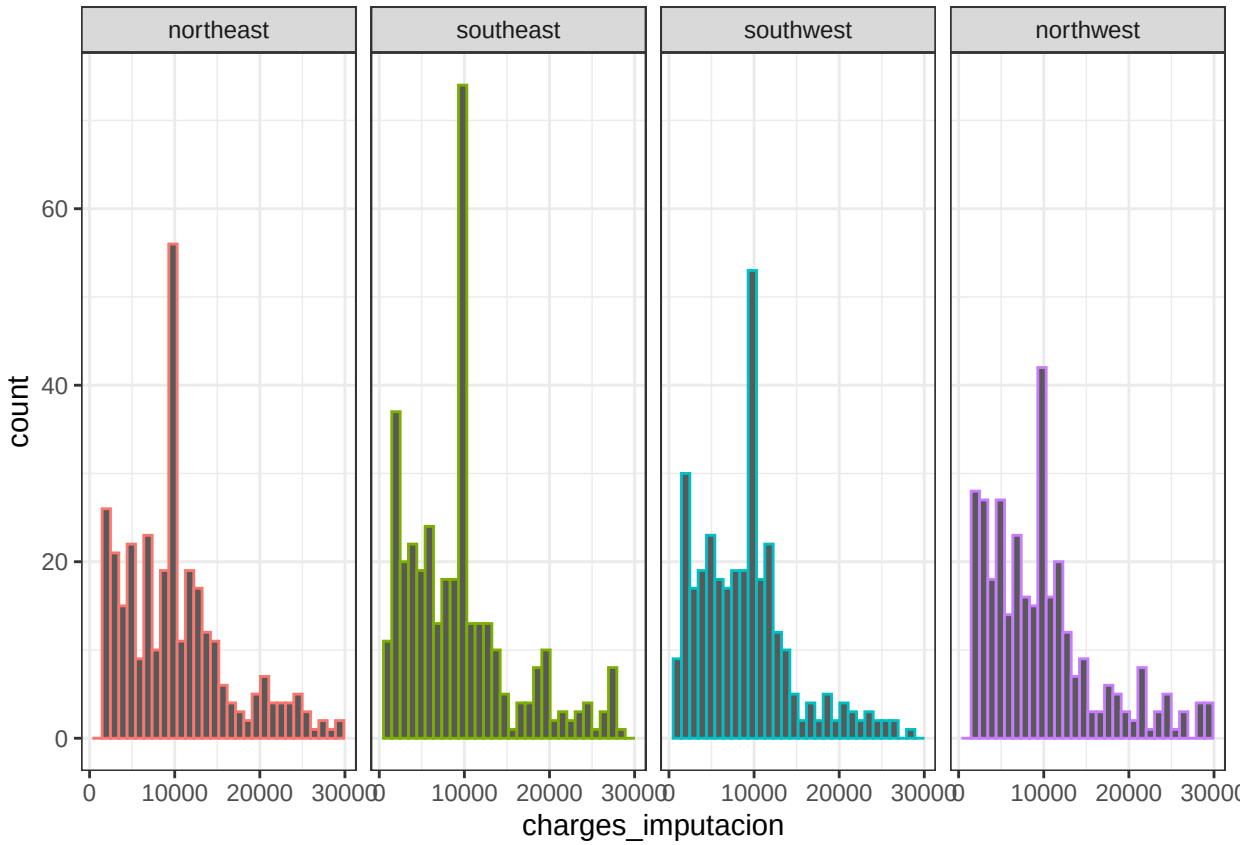
```
# Boxplot de la variable charges_imputacion por cada región.
ggplot(data = dataset, mapping = aes(x = region, y = charges_imputacion, colour = region)) +
    geom_boxplot() +
    theme_bw() +
    theme(legend.position = "none")
```



Por otro lado, para llevar a cabo este test también debe cumplirse que la distribución de la variable, en nuestro caso charges_imputation, sea la misma para cada grupo, es decir, que los datos provienen de la misma distribución. Por lo tanto, procedemos a verificar esta condición gráficamente mediante la realiación de histogramas de la variable charges_imputation para cada categoría de la variable region:

```
# Histogramas de la variable charges_imputacion por cada región.
ggplot(data = dataset, mapping = aes(x = charges_imputacion, colour = region)) +
    geom_histogram() +
    theme_bw() +
    facet_grid(. ~ region) +
    theme(legend.position = "none")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Así, observamos que efectivamente la variable charges_imputacion tiene una distribución asimétrica hacia la izquierda en cada una de las categorías de region (northeast, northwest, southeast y southwest).

Antes de llevar a cabo la prueba, debemos definir las hipótesis nulas e hipótesis alternativa de esta prueba estadística. En este sentido se tiene lo siguiente:

H_0: Igualdad de medianas (todas las muestras provienen de la misma población)

H_1: Existe diferencias entre alguna de las medianas (Al menos una muestra proviene de una población con una distribución distinta)

Finalmente, procedemos a realizar el cálculo del test:

```
# Test de de Kruskal-Wallis.
kruskal.test(charges_imputacion ~ region, data = dataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  charges_imputacion by region
## Kruskal-Wallis chi-squared = 8.2188, df = 3, p-value = 0.0417
```

Como se obtiene un p-valor $= 0.0417$ (< 0.05) podemos rechazar la hipótesis nula y afirmar que las diferencias entre algunas de las medianas son estadísticamente significativas, es decir, existe significancia en la diferencia entre, almenos, dos grupos de la variable region, no obstante, debemos tener en cuenta que el p-valor obtenido es muy cercano al nivel de significación.

Dado que el test de Kruskal-Wallis ha detectado diferencia significativa en almenos dos grupos, vamos a realizar una comparación a posteriori dos a dos para conocer qué grupos difieren. Por tanto, llevamos a cabo

un Test de Dunn entre cada par de grupos con corrección de significancia, en este caso, esta correción será la de Bonferroni:

```
library(FSA)
```

```
## Warning: package 'FSA' was built under R version 4.0.5
```

```
## ## FSA v0.8.32. See citation('FSA') if used in publication.
## ## Run fishR() for related website and fishR('IFAR') for related book.
```

```
##
## Attaching package: 'FSA'
```

```
## The following object is masked from 'package:car':
##
##     bootCase
```

```
# Test de Dunn por pares de grupos (Correción de Holm).
dunnTest(charges_imputacion ~ region,
         data=dataset,
         method="bonferroni")
```

```
## Dunn (1964) Kruskal-Wallis multiple comparison
```

```
##   p-values adjusted with the Bonferroni method.
```

```
##              Comparison         Z    P.unadj      P.adj
## 1 northeast - northwest 1.4742778 0.14040682 0.84244091
## 2 northeast - southeast 2.4330752 0.01497119 0.08982715
## 3 northwest - southeast 0.9184608 0.35837766 1.00000000
## 4 northeast - southwest 2.5367210 0.01118961 0.06713766
## 5 northwest - southwest 1.0632627 0.28766284 1.00000000
## 6 southeast - southwest 0.1744800 0.86148824 1.00000000
```

Contrariamente a lo obtenido en el Test de Kruskal-Wallis, en estas comparaciones dos a dos no se evidencian diferencias significativas, pues los p-valores ajustados por la correción de Bonferroni son superiores a 0.05. Esto puede deberse a que este ajuste de Bonferroni es muy conservador. En consecuencia, vamos a realizar este mismo Test pero sin ningún tipo de ajuste:

```
# Test de Dunn por pares de grupos (Sin ajuste).
dunnTest(charges_imputacion ~ region,
         data=dataset,
         method="none")
```

```
## Dunn (1964) Kruskal-Wallis multiple comparison
```

```
##   with no adjustment for p-values.
```

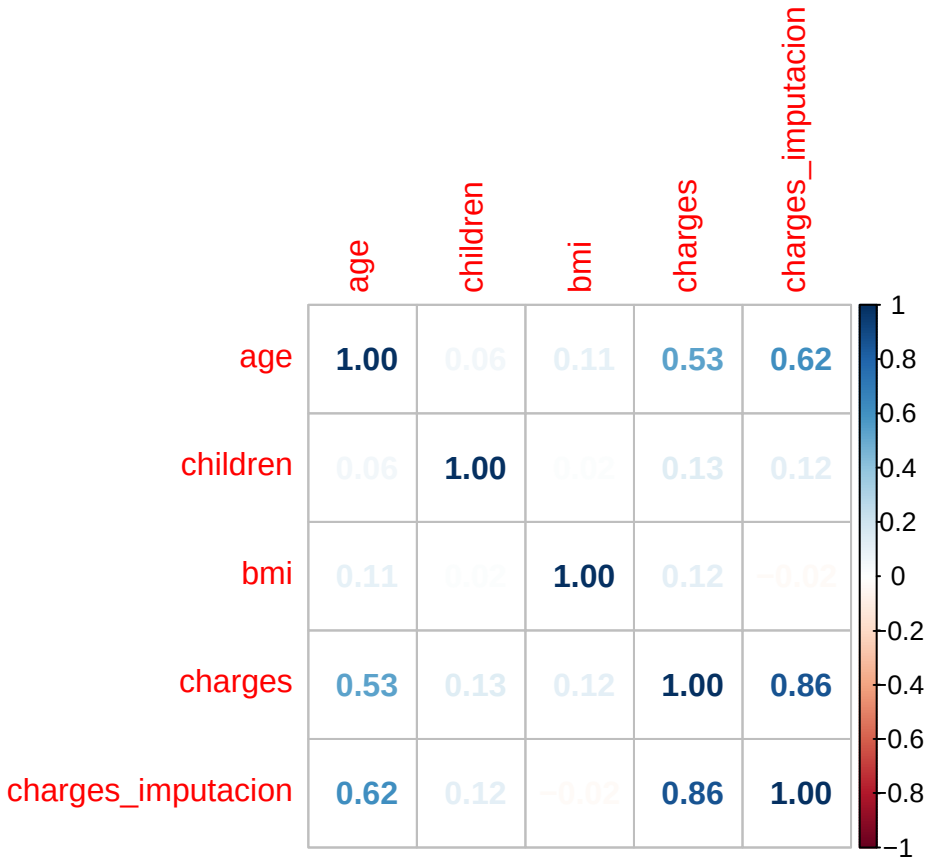
```
##              Comparison         Z    P.unadj      P.adj
## 1 northeast - northwest 1.4742778 0.14040682 0.14040682
## 2 northeast - southeast 2.4330752 0.01497119 0.01497119
## 3 northwest - southeast 0.9184608 0.35837766 0.35837766
## 4 northeast - southwest 2.5367210 0.01118961 0.01118961
## 5 northwest - southwest 1.0632627 0.28766284 0.28766284
## 6 southeast - southwest 0.1744800 0.86148824 0.86148824
```

En este caso, sí se detectan diferencias significativas, concretamente, entre las regiones northeast y southeast (p-valor $= 0.01497119 < 0.05$) y entre las regiones northeast y southwest (p-valor $= 0.01118961 < 0.05$).

3 Modelo de regresión lineal.

En primer lugar, realicemos un análisis de correlaciones para conocer qué relación existe entre la variable dependiente charges y el resto de variables. En este caso, dado que charges no tiene carácter normal, decidimos calcular el coeficiente de correlación de Spearman.

```
matriz_correlacion <- as.matrix(cor(select(dataset, age, children, bmi, charges,
                                           charges_imputacion),
                                    method = "spearman"))
corrplot(matriz_correlacion , method = "number")
```



Obviando la alta, y lógica, correlación entre charges y charges_imputacion, parece ser que la mayor correlación se establece entre la variable age y charges_imputacion siendo ésta positiva y con un valor de 0,62. Para el resto de variables, parece no existir un alto grado de correlación.

De este modo, primero calcularemos el modelo de regresión lineal con todos los predictores demográficos que tenemos en el dataset. Posteriormente, analizaremos la significancia de estas variables predictoras para decidir cuáles mantenemos en el modelo.

```
# Predictores cuantitativos.
edad <- dataset$age
IMC <- dataset$bmi
hijos <- dataset$children

# Predictores cualitativos.
sexo <- dataset$sex
fumador <- dataset$smoker
region <- dataset$region
```

```r
# Variable dependiente.
coste.seguro <- dataset$charges_imputacion

# Modelos de regresión lineal.
modelo1 <- lm(coste.seguro ~ edad + IMC + hijos + sexo + region +
    fumador , data = dataset)
summary(modelo1)
```

```
##
## Call:
## lm(formula = coste.seguro ~ edad + IMC + hijos + sexo + region +
##     fumador, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11027.2  -2161.9  -1115.3    406.4  24778.4
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3095.455    768.295   4.029 5.92e-05 ***
## edad               218.815      9.145  23.927  < 2e-16 ***
## IMC               -124.083     21.981  -5.645 2.02e-08 ***
## hijos              371.307    105.911   3.506  0.00047 ***
## sexofemale         705.813    255.890   2.758  0.00589 **
## regionsoutheast   -515.795    367.906  -1.402  0.16116
## regionsouthwest  -1015.426    367.323  -2.764  0.00578 **
## regionnorthwest   -299.858    366.049  -0.819  0.41283
## fumadoryes        6374.711    317.535  20.076  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4659 on 1329 degrees of freedom
## Multiple R-squared:  0.4305, Adjusted R-squared:  0.4271
## F-statistic: 125.6 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Así, se observa que en términos generales, la variable region no es significativa, pues dos de las variables dummies generadas en el primer modelo referidas a las categorías southeast y northwest de esta variable no son significativas. En consecuencia, procedemos a eliminarla del modelo y generar un segundo modelo sin ella.

```r
# Modelo de regresión lineal (sin variables no significativas)
modelo2 <- lm(coste.seguro ~ edad + IMC + hijos + sexo + fumador , data = dataset)
summary(modelo2)
```

```
##
## Call:
## lm(formula = coste.seguro ~ edad + IMC + hijos + sexo + fumador,
##     data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11202.9  -2142.3  -1182.3    409.7  24760.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2842.887    741.306   3.835 0.000131 ***
```

```
## edad         219.059      9.155  23.928  < 2e-16 ***
## IMC         -130.996     21.087  -6.212 6.97e-10 ***
## hijos        365.411    106.020   3.447 0.000585 ***
## sexofemale   703.675    256.377   2.745 0.006138 **
## fumadoryes  6387.925    317.259  20.135  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4668 on 1332 degrees of freedom
## Multiple R-squared:  0.427,  Adjusted R-squared:  0.4249
## F-statistic: 198.6 on 5 and 1332 DF,  p-value: < 2.2e-16
```

Tras la eliminación de la variable region (no significativa) obtenemos un coeficiente de determinación ajustado $R = 0.4249$, lo cual significa que nuestro modelo de regresión lineal explica el 42.49% de la varianza de las observaciones. Este hecho nos demuestra que no es un modelo muy potente a la hora de explicar la relación entre el costo del seguro y las variables predictoras.

Nótese que estamos realizando este análisis de regresión con la variable dependiente charges_imputacion, que corresponde a la variable original charges con los valores outliers asumidos anteriormente imputados por la mediana. Esta decisión, aunque ha sido válida para las pruebas estadísticas anteriores, ya que nos ha permitido llevar a cabo un estudio comparativo para valores de costes de seguros cercanos al promedio nacional, puede estar afectando a este análisis de regresión. Por tanto, para finalizar este apartado vamos a ajustar el mismo modelo de regresión, pero tomando como variable objetivo en este caso la variable charges original (sin imputación).

```
# Modelo de regresión lineal sin imputación de outliers.

# Variable dependiente.
coste.seguro <- dataset$charges
# Variable dependiente.
coste.seguro <- dataset$charges

# Modelo de regresión lineal.
modelo3 <- lm(coste.seguro ~ edad + IMC + hijos + sexo +
    fumador , data = dataset)

summary(modelo3)
```

```
##
## Call:
## lm(formula = coste.seguro ~ edad + IMC + hijos + sexo + fumador,
##     data = dataset)
##
## 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) -12181.10     963.90 -12.637  < 2e-16 ***
## edad           257.73      11.90  21.651  < 2e-16 ***
## IMC            322.36      27.42  11.757  < 2e-16 ***
## hijos          474.41     137.86   3.441 0.000597 ***
## sexofemale     128.64     333.36   0.386 0.699641
## fumadoryes   23823.39     412.52  57.750  < 2e-16 ***
## ---
```

```
## 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
```

En este caso, el modelo generado ha mejorado considerablemente, pues el coeficiente de determinación ha aumentado su valor hasta 0.7488, es decir, ahora nuestro modelo tiene la capacidad de explicar el 74,88% de la variabilidad de las observaciones. En este sentido, podemos decir que el tercer modelo es un buen modelo para predecir el costo de un seguro en base a las variables demográficas empleadas.

Ahora, usaremos este modelo para predecir el costo del seguro para un individuo de 30 años, con índice de masa corporal igual a 20, 2 hijos y no fumador:

```
datos <- data.frame(edad = 30, IMC = 20, hijos = 2, sexo = "male", fumador = "no")
predict(modelo3, newdata = datos)
```

```
##        1
## 2947.054
```

Se ha obtenido un coste del seguro igual a 2947.054 dólares.

Ahora realizaremos la misma predicción pero para un individuo de las mismas características pero, en este caso, que sí sea fumador:

```
datos <- data.frame(edad = 30, IMC = 20, hijos = 2, sexo = "male", fumador = "yes")
predict(modelo3, newdata = datos)
```

```
##        1
## 26770.45
```

Como podemos ver, el coste del seguro de salud privado calculado (26770.45) es ahora mucho mayor para la persona fumadora.

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

Primeramente, presentamos la siguiente tabla en la que se exponen los resultados obtenidos en el primer contraste de hipótesis realizado, en el cual, estudiamos si exístia diferencia significativa entre las medias de la variable coste del seguro para hombres y mujeres:

```
# Contenido de la tabla 1.
contenido.t1 <- data.frame("Estadístico" = c(test_hip$statistic),
                           "P-valor" = c(test_hip$p.value),
                           "IC"= c("(-55.79, Inf )"),
                           "Nivel de Confianza" = "95%",
                           "Significación" = "0.05")

# Generamos la tabla con los resultados.
knitr::kable(contenido.t1, digits = 4, align = "c",
caption = "Resultados test de hipótesis: Diferencias de medias en el coste del seguro por género." )
```

Table 1: Resultados test de hipótesis: Diferencias de medias en el coste del seguro por género.

	Estadístico	P.valor	IC	Nivel.de.Confianza	Significación
t	1.4802	0.0695	(-55.79, Inf)	95%	0.05

En segundo lugar, llevamos a cabo un test de Kruskal Wallis para comprobar si existían diferencias en la variable charges_imputacion en función de la región del individuo. A continuación, realizamos un gráfico de barras en el que mostramos las diferencias existentes y, también, evidenciamos aquellas regiones en las que encontramos diferencias con respecto al valor del seguro médico: northeast - southeast (color rojo) y northeast - southwest (color rosa).

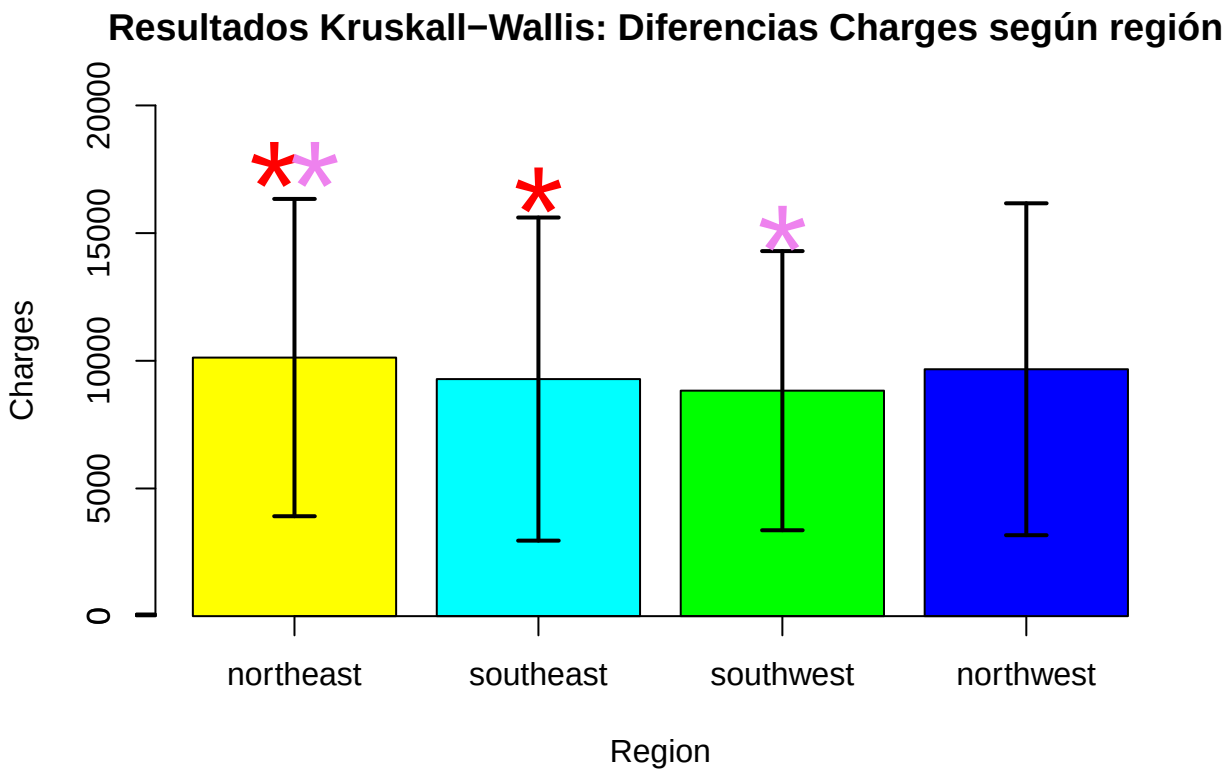
```
# Gráfico de las diferencias del coste del seguro por región.

# Cálculo de la media y desviación típica de charges_imputacion por región.
med.charges<-ddply(dataset,.(region), summarize, mean=mean(charges_imputacion))
sd.charges<-ddply(dataset,.(region), summarize, sd=sd(charges_imputacion))

# Barras del gráfico basado en las medias de charges_imputacion en función de la región.
BARRAS<-barplot(med.charges$mean, axes=TRUE,axisname=FALSE, ylim=c(0,20000),
        col=c("yellow", "cyan", "green", "blue"),
        main=" Resultados Kruskall-Wallis: Diferencias Charges según región",
        xlab="Region", ylab="Charges")
axis(1,labels=c("northeast", "southeast", "southwest", "northwest"), at=BARRAS)
axis(2,at=seq(0,100,by=10))

#Segemtnos de errores
segments(BARRAS-0.1,med.charges$mean-sd.charges$sd,BARRAS+0.1,med.charges$mean-sd.charges$sd,lwd=2)
segments(BARRAS-0.1,med.charges$mean+sd.charges$sd,BARRAS+0.1,med.charges$mean+sd.charges$sd,lwd=2)
segments(BARRAS,med.charges$mean-sd.charges$sd,BARRAS,med.charges$mean+sd.charges$sd,lwd=2)

# Etiquetas de las regiones que difieren.
text(1.9,17000,labels="*",cex=4, col = "red")
text(0.6,18000,labels="*",cex=4, col = "red")
text(0.8,18000,labels="*",cex=4, col = "violet")
text(3.1,15500,labels="*",cex=4, col = "violet")
```



En tercer lugar, procedimos a ajustar tres modelos de regresión lineal estableciendo como variable dependiente el costo del seguro médico y, como predictoras, el resto de variables de carácter sociodemográfico que teníamos en el dataset. Así, tras ajustar el primer modelo, observamos la no significancia de la variable region y ajustamos un segundo modelo sin dicho predictor. Finalmente, sospechamos de la influencia que tenía la imputación que habíamos realizado en un principio sobre la variable dependiente y, en consecuencia, ajustamos un modelo sin estas imputaciones.

En la siguiente tabla, se exponen los coeficientes de determinación obtenidos en cada uno de los modelos comentados anteriormente. Así, se observa que tras eliminar la variable region la variabilidad explicada sufre un leve descenso y, también, al llevar a cabo el tercer modelo planteado, el coeficiente de determinación se sitúa en un valor de 0.7497. Por este motivo, elegimos este tercer modelo como el mejor entre los tres modelos.

```
# Contenido de la tabla 2.
contenido.t2 <- data.frame("Modelo"=c("Modelo 1", "Modelo 2", "Modelo 3"),
                           "Coef. Determinación" = c(summary(modelo1)$r.squared,
                                                     summary(modelo2)$r.squared,
                                                     summary(modelo3)$r.squared)

                           )

# Generamos la tabla con los resultados.
knitr::kable(contenido.t2, digits = 4, align = "c",
caption = "Coeficientes de determinación obtenidos en los distintos modelos" )
```

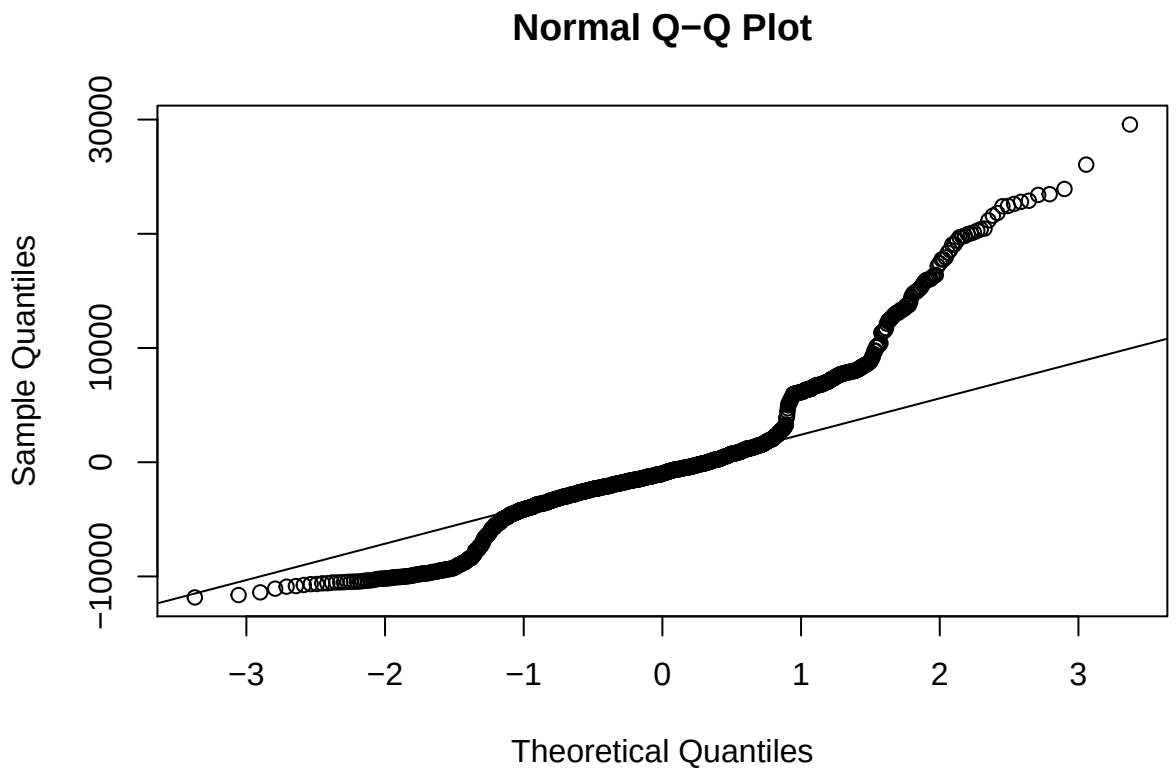
Table 2: Coeficientes de determinación obtenidos en los distintos modelos

Modelo	Coef..Determinación
Modelo 1	0.4305
Modelo 2	0.4270
Modelo 3	0.7497

A continuación, se exponen los gráficos que nos permitirán analizar la viabilidad de nuestro tercer modelo, el cual fue elegido porque obtuvo el mayor coeficiente de determinación como vimos en la tabla anterior. Por un lado, realizamos un QQ plot para determinar la normalidad en la distribución de los residuos y, por otro, un gráfico de los residuos vs los valores predichos por el modelo para analizar la existencia de homocedasticidad.

De esta forma, en el QQ plot observamos que los residuos no se comportan de manera normal, mientras que en el gráfico de los residuos frente a los valores predichos podemos intuir cierta independencia en la distribución de puntos que nos hacen confirmar la presencia de homocedasticidad.

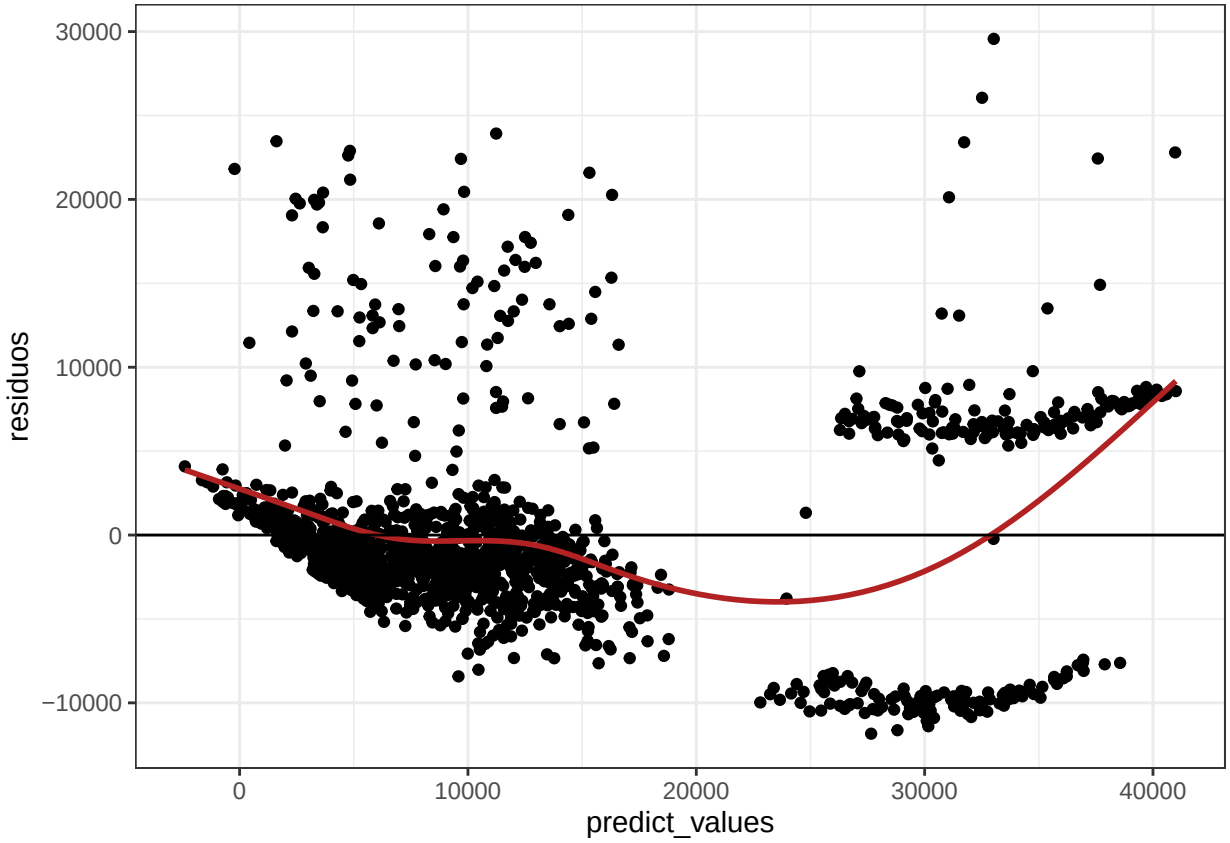
```
# Q-Q Plot
qqnorm(modelo3$residuals)
qqline(modelo3$residuals)
```




```
# Gráfico Residuos vs Predichos
ggplot(data = data.frame(predict_values = predict(modelo3),
                         residuos = residuals(modelo3)),
       aes(x = predict_values, y = residuos)) +
    geom_point() +
```

```
    geom_smooth(color = "firebrick", se = FALSE) +
    geom_hline(yintercept = 0) +
    theme_bw()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



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

Tras realizar estas tres pruebas estadísticas procedemos a determinar las conclusiones extraídas de los resultados anteriormente expuestos. Así, el primer test de hipótesis realizado nos ha permitido afirmar, con un 95% de confianza, que no es cierto que las mujeres tienen un coste medio mayor del seguro médico en comparación con los hombres.

Por otro lado, a partir de los resultados obtenidos en el test de Kruskal Wallis hemos sido capaces de estudiar la existencia, o no, de diferencias significativas en el coste medio de un seguro de salud privado en función del lugar de residencia del contratante, obteniendo que efectivamente se encuentran diferencias. Posteriormente, hemos detectado que dichas diferencias se encuentran entre dos pares de regiones: northeast - southeast y northeast - southwest.

Finalmente, en cuanto al tercer modelo de regresión lineal múltiple ajustado, hemos obtenido un alto coeficiente de determinación, sin embargo, la condición de normalidad en la distribución de los residuos en dicho modelo no se verifica. Este hecho nos obliga a no considerar como válido el modelo, y, por tanto, no ha podido responder a nuestra pregunta inicial en la que tratábamos de conocer una relación (válida y fiable) entre

las variables demográficas del dataset elegido y el coste de un seguro privado de salud. Debido a esto, debe realizarse otro tipo de modelo que se ajuste mejor a los datos y asegure el cumplimiento de las premisas necesarias, para que verificar su validez. Otra opción, sería continuar con el enfoque de la regresión lineal pero realizando una transformación logarítmica para tratar de normalizar las variables no normales.

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 con el que se ha realizado la limpieza, análisis y representación de los datos se muestra en el prensete informe, aunque puede ser consultado en el siguiente link:

https://github.com/cherreracar/Practica2_Tipologia-.git/

Exportación de los datos analizados.

```
write.csv(dataset, file="insurance_clean.csv")
```

Tabla de contribuciones al trabajo

Contribuciones	Firma
Investigación previa	CHC
Redacción de las respuestas	CHC
Desarrollo código	CHC

Figure 1: Tabla de contribuciones