Actividad4_Explorando_bases

Rodolfo Jesús Cruz Rebollar

2024-08-13

```r
# Leer los datos del archivo csv

mcdonalds = read.csv("mc-donalds-menu.csv")

head(mcdonalds)
```

```
##    Category                            Item   Serving.Size Calories
## 1 Breakfast                     Egg McMuffin 4.8 oz (136 g)      300
## 2 Breakfast                Egg White Delight 4.8 oz (135 g)      250
## 3 Breakfast                 Sausage McMuffin 3.9 oz (111 g)      370
## 4 Breakfast        Sausage McMuffin with Egg 5.7 oz (161 g)      450
## 5 Breakfast Sausage McMuffin with Egg Whites 5.7 oz (161 g)      400
## 6 Breakfast             Steak & Egg McMuffin 6.5 oz (185 g)      430
##   Calories.from.Fat Total.Fat Total.Fat....Daily.Value. Saturated.Fat
## 1               120        13                        20             5
## 2                70         8                        12             3
## 3               200        23                        35             8
## 4               250        28                        43            10
## 5               210        23                        35             8
## 6               210        23                        36             9
##   Saturated.Fat....Daily.Value. Trans.Fat Cholesterol
## 1                            25         0         260
## 2                            15         0          25
## 3                            42         0          45
## 4                            52         0         285
## 5                            42         0          50
## 6                            46         1         300
##   Cholesterol....Daily.Value. Sodium Sodium....Daily.Value.
Carbohydrates
## 1                          87    750                    31
31
## 2                           8    770                    32
30
## 3                          15    780                    33
29
## 4                          95    860                    36
30
## 5                          16    880                    37
30
## 6                         100    960                    40
31
##   Carbohydrates....Daily.Value. Dietary.Fiber
```

```
Dietary.Fiber....Daily.Value.
## 1                             10             4
17
## 2                             10             4
17
## 3                             10             4
17
## 4                             10             4
17
## 5                             10             4
17
## 6                             10             4
18
##   Sugars Protein Vitamin.A....Daily.Value. Vitamin.C....Daily.Value.
## 1      3      17                        10                          0
## 2      3      18                         6                          0
## 3      2      14                         8                          0
## 4      2      21                        15                          0
## 5      2      21                         6                          0
## 6      3      26                        15                          2
##   Calcium....Daily.Value. Iron....Daily.Value.
## 1                      25                   15
## 2                      25                    8
## 3                      25                   10
## 4                      30                   15
## 5                      25                   10
## 6                      30                   20
```

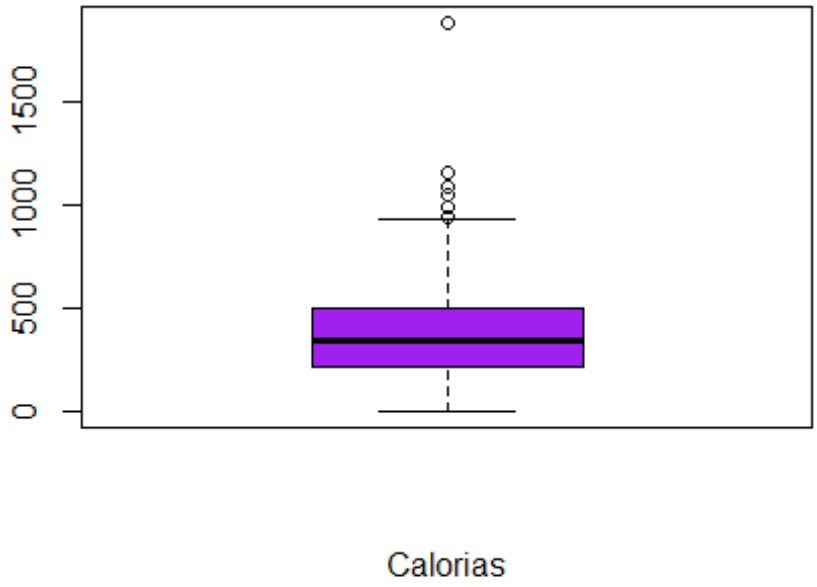
Análisis de la variable calorias

Análisis de datos atípicos

```
# Colocar los datos de calorias en una variable aparte

calorias = mcdonalds$Calories
```

Diagrama de caja y bigote de calorias

```
# Graficar el diagrama de caja y bigote para la variable calorias

boxplot(calorias, col = "purple", xlab = "Calorias")
```



En el diagrama de caja y bigote se observa principalmente que existen datos atípicos que quedan fuera del bigote superior del diagrama que se ubica un poco por debajo de 1000 calorias, además de que la caja del gráfico no tiene mucha longitud, esto indica que en general, los datos no presentan una variación mayormente significativa entre ellos, lo cual además sugiere que los datos de calorias pueden tener pocos datos atípicos, dado que la gran mayoría de ellos se encuentran concentrados dentro la zona abarcada por la caja del diagrama, mientras que al mismo tiempo, solamente otros pocos quedarán fuera tanto de la caja como de los bigotes de la misma.

Rango intercuartílico y cuartíles

```
# Calcular cuartíles

cuartiles = quantile(calorias, c(0.25, 0.5, 0.75))

cuartiles
```

```
## 25% 50% 75%
## 210 340 500
```

```
# Calcular rango intercuartílico

IQR = cuartiles[3] - cuartiles[1]

cat("\n", "Rango intercuartílico: ", IQR)
```

```
##
##  Rango intercuartílico:  290
```

Cota de 1.5 rangos intercuartílicos

```
# Calcular la cota para determinar su existen datos atípicos considerando
el criterio de
# 1.5 veces el rango intercuartílico

cota_1.5_inferior = cuartiles[1] - 1.5 * IQR # cota inferior

cota_1.5_superior = cuartiles[3] + 1.5 * IQR # cota superior

cat("La cota de 1.5 rangos intercuartílicos para datos atípicos es: ",
"\n",
    "Cota inferior: ", cota_1.5_inferior, "\n",
    "Cota superior: ", cota_1.5_superior)

## La cota de 1.5 rangos intercuartílicos para datos atípicos es:
##  Cota inferior:  -225
##  Cota superior:  935

# Buscar si existen datos atípicos con base al criterio de 1.5 rangos
intercuartílicos

calorias[(calorias < cota_1.5_inferior) | (calorias > cota_1.5_superior)]

## [1] 1090 1150  990 1050  940 1880
```

De acuerdo al criterio de 1.5 rangos intercuartílicos, se observa que en total existen 6 datos atípicos en la base de datos, los cuales sobrepasan la cota máxima calculada de 935 calorias, mientras que no hay datos atípicos inferiores a la cota inferior calculada de -225, esto debido a que las calorias no pueden ser valores negativos (solamente pueden ser 0 o mayores que 0), por lo que en el caso particular de las calorias, la única cota que utilizaremos para comprobar si existen datos atípicos es la superior, misma que tiene un valor de 935.

Cota de 3 desviaciones estándar

Para calcular la cota para el criterio de 3 desviaciones estándar, primero tenemos que establecer la relación de los valores de x en función de los de z, para ello, se despeja la variable x de la función $z = \frac{x-\mu}{\sigma}$:

$$z = \frac{x - \mu}{\sigma}$$

$$z\sigma = x - \mu$$

$$x = z\sigma + \mu$$

```r
# Calcular la cota para determinar valores atípicos en base al criterio
de 3 desviaciones
# estándar

# Cota inferior  (con z = -3)

cota_3sd_inferior = -3 * sd(calorias) + mean(calorias)

# Cota superior (con z = 3)

cota_3sd_superior = 3 * sd(calorias) + mean(calorias)

# Mostrar cota superior e inferior

cat("La cota de 3 desviaciones estándar para datos atípicos es: ", "\n",
    "Cota inferior: ", cota_3sd_inferior, "\n", "Cota superior: ",
cota_3sd_superior)
```

```
## La cota de 3 desviaciones estándar para datos atípicos es:
##  Cota inferior:  -352.5404
##  Cota superior:  1089.079
```

```r
# Buscar si existen datos atípicos en base al criterio de 3 desviaciones
estándar

calorias[(calorias < cota_3sd_inferior) | (calorias > cota_3sd_superior)]
```

```
## [1] 1090 1150 1880
```

Nota: Dado que las calorias no pueden ser valores negativos, solamente utilizamos la cota superior (positiva) para determinar datos atípicos y la cota inferior no se considera.

Acorde al criterio de 3 desviaciones estándar alrededor de la media, se observa que se tienen 3 datos atípicos en comparación con el criterio de 1.5 rangos intercuartílicos que arroja 6, por lo que el criterio de 1.5 rangos intercuartílicos detecta una mayor cantidad de datos atípicos que el criterio de 3 desviaciones estándar, esto principalmente debido a que en el criterio de 3 desviaciones estándar, la cota para detectar datos atípicos es de 1089.079, mientras que en el criterio de 1.5 rangos intercuartílicos, dicha cuota tiene un valor de 935, lo cual permite que éste último criterio, detecte una mayor cantidad de datos atípicos, puesto que al tener menor cota, los datos no necesitarán tener valores muy elevados como en el caso del criterio de las 3 desviaciones estándar para ser detectados.

¿Quitar, o no quitar los datos atípicos?

Por último, después de todo el análisis realizado, resulta conveniente no quitar los datos atípicos de la base de datos, debido a que nos pueden indicar cuáles productos de McDonalds en concreto tienen un alto nivel de calorías, esto con el principal

objetivo de que la empresa contribuya al cuidado de la salud de sus clientes, en especial de aquellos que padezcan alguna enfermedad como la diabetes, en el sentido de que sabiendo los alimentos que tienen más calorias, la empresa pueda desarrollar versiones con una cantidad reducida de calorias de los alimentos que venden habitualmente, sin necesidad de modificar radicalmente todo su menú de alimentos para adaptarlo a las necesidades de salud de sus clientes.

Análisis de normalidad

Prueba de hipótesis para verificar la normalidad:

H_0: la variable calorias sigue una distribución normal.

H_1: la variable calorias no sigue una distribución normal.

Dado que la base de datos tiene una cantidad de registros mayor a 50, se utilizará la prueba de Kolmogorov Smirnov y la prueba de Jarque Bera para verificar la normalidad de los datos.

Prueba de Kolmogorov-Smirnov

```
# Librería para pruebas de normalidad univariada

library(nortest)

# Realizar prueba de normmalidad de Kolmogorov Smirnov considerando la
modificación de
# Lilliefors

lillie.test(calorias)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  calorias
## D = 0.073753, p-value = 0.001611
```

Prueba de Jarque Bera

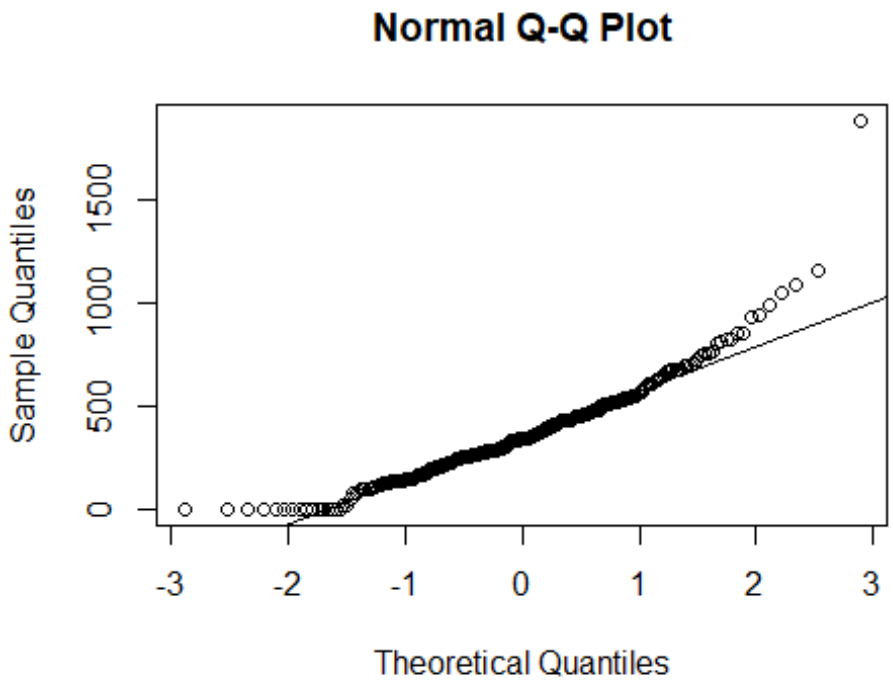
```
# Librería moments para realizar el test de Jarque Bera basado en sesgo y
curtosis

library(moments)

# Realizar el test de Jarque Bera

jarque.test(calorias)
```

```
##
##  Jarque-Bera Normality Test
##
```

```
## data:  calorias
## JB = 435.62, p-value < 2.2e-16
## alternative hypothesis: greater
```

QQplot de calorias

```
# Graficar los puntos del qqplot con el comando qqnorm

qqnorm(calorias)

# Graficar la línea recta del qqplot que simboliza el caso ideal (los
datos siguen distribución normal)

qqline(calorias)
```



Coeficientes de sesgo y curtosis

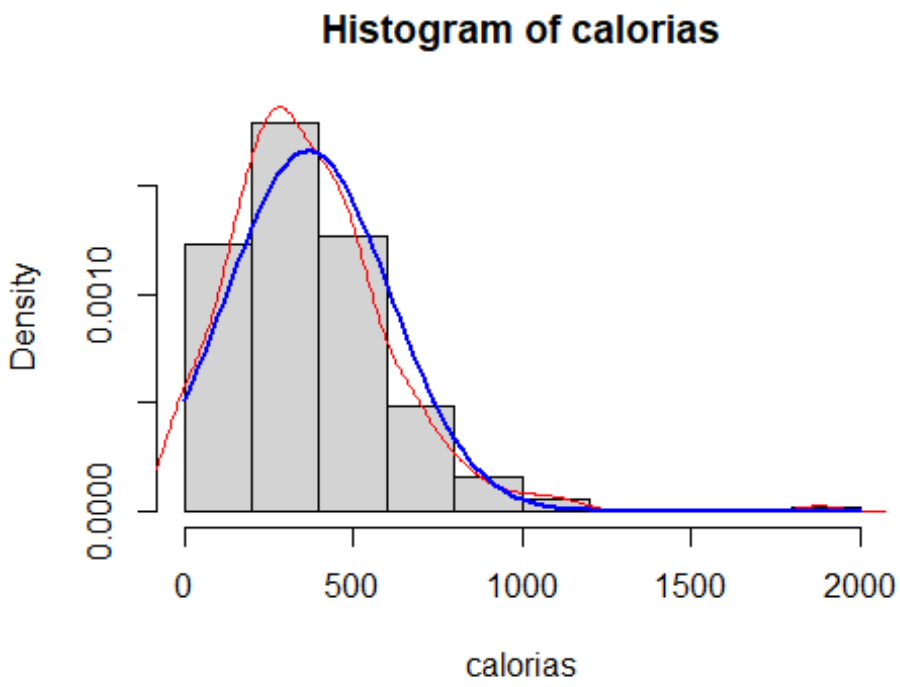
```
# Calcular el coeficiente de sesgo de calorias

cat("Sesgo de calorias: ", skewness(calorias), "\n")

## Sesgo de calorias:  1.444105

# Calcular coeficiente de curtosis de calorias

cat("Curtosis de calorias: ", kurtosis(calorias))

## Curtosis de calorias:  8.645274
```

Media, mediana, rango medio

```
# Media de calorias

cat("Media de calorias: ", mean(calorias), "\n")

## Media de calorias:  368.2692

# Mediana

cat("Mediana de calorias: ", median(calorias), "\n")

## Mediana de calorias:  340

# Rango medio

cat("Rango medio de calorias: ", mean(c(max(calorias), min(calorias))))

## Rango medio de calorias:  940
```

Histograma de calorias y su distribución teórica

```
# Histograma de calorias

hist(calorias, freq = FALSE)

# Graficar distribución empírica (a partir de los datos) y teórica (cómo
debería ser idealmente)

lines(density(calorias), col = "red") # distribución empírica

curve(dnorm(x,mean=mean(calorias), sd=sd(calorias)), from = 0, to = 2000,
add = TRUE,
      col = "blue", lwd = 2) # distribución teórica
```



Influencia de datos atípicos en la normalidad de los datos

En términos generales, después de realizar las pruebas de hipótesis para normalidad junto con algunos gráficos, entre ellos, el QQplot además de la gráfica de la distribución empírica contra la teórica de los datos, es posible afirmar que en cuanto a la influencia de los datos atípicos en la normalidad de los datos en general, dichos datos atípicos provocan que la distribución de todo el conjunto de datos experimente una traslación hacia la región izquierda del gráfico, por lo que la distribución general de los datos queda con un sesgo a la derecha, dejando una cantidad mínima de datos en dicha región de la gráfica, esto sucede principalmente debido a que los datos atípicos en cuestión son en realidad datos extremos, ubicados a más de 3 desviaciones estándar de la media de todos los datos, por lo que la distribución considerando el total de datos busca ajustarse a dichos datos extremos o influyentes, para lo cual, si dichos datos influyentes se encuentran muy a la izquierda o derecha del gráfico, la distribución de los datos se desplazará para ese lado del histograma.

Conclusión final sobre normalidad de calorias

Adicionalmente, en cuanto a los gráficos elaborados, en primer lugar, en el QQplot se puede observar que aunque hay una concentración de datos significativa sobre la línea recta que representa a la distribución normal, sin embargo, también se puede apreciar que hay ciertos datos en los extremos de la línea recta que se salen de ella, lo cual es un indicativo de que los datos en general no siguen una distribución normal. Además de lo anterior, también es importante mencionar que de acuerdo al gráfico de la comparación entre la distribución empírica de los datos y la teórica, se observa que

la línea roja (correspondiente a la distribución empírica) presenta ciertos alejamientos con respecto a la distribución teórica (caso ideal de normalidad), lo cual indica también que la distribución de los datos en cuestión se alejan de la normal. Por otra parte, otro recurso empleado para determinar la normalidad de los datos radica en el cálculo de la media, mediana y el rango medio de dichos datos, esto debido a que en el caso ideal de que los datos siguieran una distribución normal, las tres medidas tendrían valores iguales, por lo que son un criterio útil para determinar si la distribución de los datos es normal, o no, por lo tanto, para el caso particular de los datos de calorias, se observa que el valor de la media es 368.2692, mientras que el de la mediana y el rango medio son 340 y 940 respectivamente, por lo que al contrario, los valores de la media, mediana y rango medio de la distribución de los datos analizados son diferentes entre sí, motivo por el cual, también indica que los datos no siguen distribución normal, apoyando así la conclusión obtenida a partir de los gráficos que radica en el hecho de que los datos de calorias no siguen distribución normal. Finalmente, otras herramientas empleadas para analizar si los datos de calorias se distribuyen de forma normal son los tests de normalidad de Kolmogorov Smirnov y de Jarque Bera, los cuales se usan para analizar muestras de datos grandes, por lo que son aplicables en el caso de los datos analizados al tener más de 50 registros (260), debido a ello, al momento de realizar el test de Kolmogorov Smirnov, la prueba arroja un valor p de 0.001611, lo cual al ser menor que 0.05, se rechaza H_0, por lo tanto, se concluye que en base al test de Kolmogorov Smirnov, la variable calorias no sigue una distribución normal. De manera similar, al realizar el test de Jarque Bera, se obtiene un p valor menor que 2.2e-16, lo cual al ser mucho menor que 0.05, se rechaza igualmente H_0, por lo que de acuerdo al test de Jarque Bera, también se concluye que la variable calorias no sigue una distribución normal, además de que también el sesgo de calorias es de 1.444105 y su curtosis es de 8.645274, por lo que al tener valores de sesgo y curtosis elevados, se respalda el hecho de que calorias no sigue distribución normal, por lo que al considerar todas las evidencias presentadas, se concluye finalmente que la variable calorias no sigue una distribución normal.

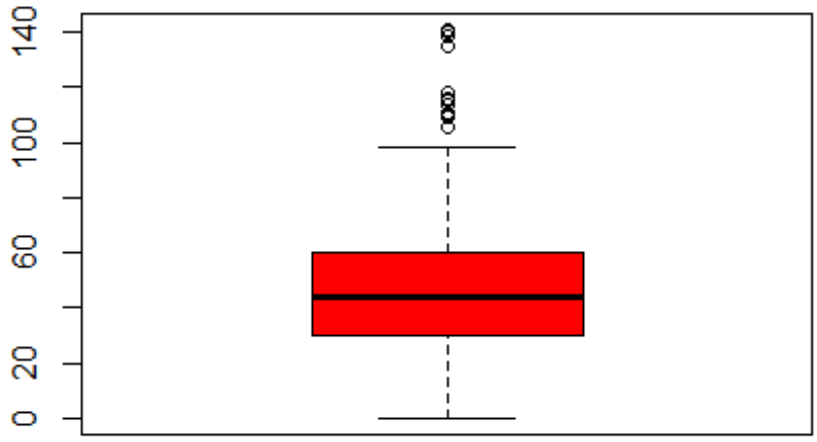
Análisis de la variable Carbohidratos

```
# Guardar los datos de carbohidratos en una variable a parte

carbohidratos = mcdonalds$Carbohydrates
```

Análisis de datos atípicos

Diagrama de caja y bigote de carbohidratos

```
# Graficar el boxplot de la variable carbohidratos

boxplot(carbohidratos, col = "red", xlab = "Carbohidratos")
```



En el diagrama de caja y bigote para los datos de carbohidratos, se observa principalmente que hay algunos puntos (datos) que se encuentran fuera de la caja y de su bigote superior, mismo que se ubica alrededor del valor de 100, lo que sugiere que los posibles datos atípicos serán mayores que 100, además, también se aprecia que la caja del diagrama no es muy larga, indicando que no existe una variación altamente significativa en los datos correspondientes a los carbohidratos de los alimentos, adicionalmente, dado que la mayoría de los datos se encuentran dentro de la zona abarcada por la caja del diagrama, en realidad puede existir solamente una cantidad reducida de datos que estén fuera de dicha zona, por lo cual, es posible que mediante el criterio de 1.5 rangos intercuartílicos, se logren encontrar cuáles son esos datos atípicos que aparecen arriba del bigote superior del boxplot.

Rango intercuartílico y cuartíles

```
# Calcular cuartiles (Q1, Q2, Q3)

cuartiles_carbo = quantile(carbohidratos, c(0.25, 0.5, 0.75))

cuartiles_carbo
```

```
## 25% 50% 75%
##  30  44  60
```

```
# Calcular rango intercuartílico Q3 - Q1

IQR_carbo = cuartiles_carbo[3] - cuartiles_carbo[1]
```

```r
cat("\n", "Rango intercuartílico: ", IQR_carbo)
```

```
##
##  Rango intercuartílico:  30
```

Cota de 1.5 rangos intercuartílicos

```r
# Calcular cota inferior

cotainf_itq = cuartiles_carbo[1] - 1.5 * IQR_carbo

# Calcular cota superior

cotasup_itq = cuartiles_carbo[3] + 1.5 * IQR_carbo

cat("La cota para 1.5 rangos intercuartílicos es: ", "\n", "Cota
inferior: ",
    cotainf_itq, "\n", "Cota superior: ", cotasup_itq)
```

```
## La cota para 1.5 rangos intercuartílicos es:
##  Cota inferior:  -15
##  Cota superior:  105
```

```r
# Buscar datos que sean menores a -15 o mayores a 105 (datos atípicos)

carbohidratos[(carbohidratos < -15) | (carbohidratos > 105)]
```

```
##  [1] 111 116 110 115 118 111 109 135 114 140 114 141 109 135 139 106
114
```

En base al criterio de 1.5 rangos intercuartílicos, se observa que dicho método arroja en total 17 datos atípicos, cuyos valores son todos mayores que 105 (límite superior de la cota), mientras que el límite inferior de la cota (-15) no se toma en cuenta ya que es un valor negativo y los datos de carbohidratos no pueden ser negativos, solamente pueden ser 0 o positivos, por lo que en realidad para determinar los datos atípicos mediante éste criterio, solamente se utiliza el límite superior de la cota (105).

Cota de 3 desviaciones estándar

Para calcular la cota para determinar los datos atípicos a partir del criterio de 3 desviaciones estándar alrededor de la media, se considerará la siguiente relación entre los valores de x y de z, obtenida en el análisis de datos atípicos de la variable calorias:

$$x = z\sigma + \mu$$

```r
# Dado que los datos atípicos son aquellos ubicados a más de 3
desviaciones estándar de la
# media, se procederá a calcular los valores de x correspondientes a z =
-3 y z = 3
```

```r
# Para z = -3

cotasd_z1_carbo = -3 * sd(carbohidratos) + mean(carbohidratos)

# Para z = 3

cotasd_z2_carbo = 3 * sd(carbohidratos) + mean(carbohidratos)

cat("La cota de 3 desviaciones estándar para datos atípicos es: ", "\n",
    "Cota inferior: ", cotasd_z1_carbo, "\n", "Cota superior: ",
cotasd_z2_carbo)
```

```
## La cota de 3 desviaciones estándar para datos atípicos es:
##  Cota inferior:  -37.41054
##  Cota superior:  132.1028
```

```r
# Buscar datos que estén por debajo de -37.41 o por encima de 132.1

carbohidratos[(carbohidratos < cotasd_z1_carbo) | (carbohidratos >
cotasd_z2_carbo)]
```

```
## [1] 135 140 141 135 139
```

Al buscar datos atípicos mediante el criterio de las 3 desviaciones estándar, se aprecia que éste criterio solamente consigue identificar 5 datos atípicos, lo cual es una cantidad considerablemente menor que la identificada por medio del criterio de 1.5 rangos intercuartílicos, esto sucede nuevamente porque en el caso de 1.5 rangos intercuartílicos, la cota utilizada para determinar los datos atípicos es de 105, mientras que en el criterio de 3 desviaciones estándar dicha cota es de 132.1, lo cual permite que el criterio de los rangos intercuartílicos detecte más datos atípicos que el de las desviaciones estándar.

¿Quitar o no quitar datos atípicos?

Por último, es conveniente no quitar los datos atípicos de carbohidratos porque es posible que jueguen un papel importante al momento de querer estudiar con más profundidad el efecto de la cantidad de carbohidratos en otras variables medibles de los alimentos, o bien, si en algún momento la empresa desea desarrollar algún modelo predictivo que prediga los valores de alguna otra propiedad de los alimentos a partir de la cantidad de carbohidratos de los mismos, será importante conservar los datos atípicos de carbohidratos, ya que pueden aportar información relevante para que las predicciones derivadas del modelo sean lo más precisas y confiables posible y que de no incluirlos en el modelo, tanto el rendimiento como la confiabilidad del modelo se vean notablemente disminuidos.

Análisis de normalidad

Prueba de hipótesis:

H_0: La variable carbohidratos sigue una distribución normal.

H_1: La variable carbohidratos no sigue una distribución normal.

Test de normalidad de Kolmogorov Smirnov

```
# Realizar prueba de normmalidad de Kolmogorov Smirnov considerando la
modificación de
# Lilliefors para la variable carbohidratos

lillie.test(carbohidratos)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  carbohidratos
## D = 0.098548, p-value = 2.081e-06
```

Test de normalidad de Jarque Bera

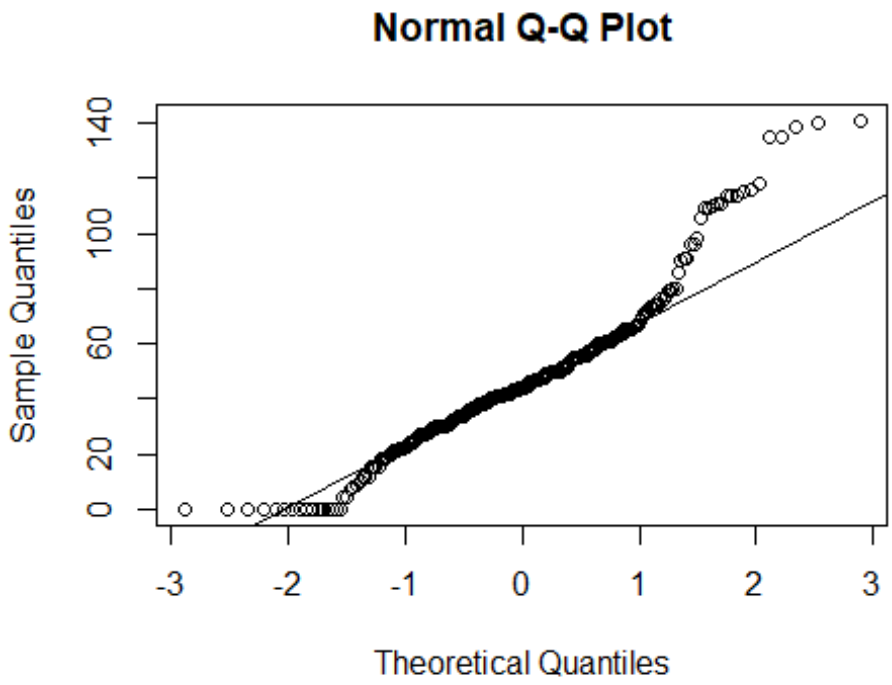
```
# Realizar el test de Jarque Bera de la variable carbohidratos

jarque.test(carbohidratos)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  carbohidratos
## JB = 55.646, p-value = 8.251e-13
## alternative hypothesis: greater
```

En base a los tests de normalidad anteriores, se puede apreciar que en el test de Kolmogorov Smirnov, el p valor arrojado por el test es de 2.081e-06, lo cual es mucho menor que 0.05, por lo tanto, se rechaza H_0 y se concluye que la variable carbohidratos no sigue una distribución normal. Por otro lado, en el test de normalidad de Jarque Bera, el p valor del test fue de 8.251e-13, lo cual al ser mucho menor que 0.05, se rechaza nuevamente H_0, motivo por el cual, se concluye que la variable carbohidratos no sigue una distribución normal.

QQplot de carbohidratos

```
# Graficar los puntos del QQplot (percentiles)

qqnorm(carbohidratos)

# Graficar la línea recta que corresponde a la distribución ideal
(normal)

qqline(carbohidratos)
```



En el gráfico QQplot de los datos de carbohidratos, se observa que a pesar de que hay muchos datos ubicados sobre la línea recta (distribución normal), hay otros que en cambio, se desvían de dicha recta, en concreto aquellos ubicados en ambos extremos de la misma, por lo que al alejarse de la línea recta, prácticamente significa que esos datos se alejan de la distribución normal, por lo que en resumen, el QQplot sugiere que los datos no siguen una distribución normal, apoyando así la conclusión hecha en base a los resultados de los tests de normalidad previos (los carbohidratos no siguen una distribución normal).

Coeficientes de sesgo y curtosis

```
# Calcular el coeficiente de sesgo para los carbohidratos

cat("Coeficiente de sesgo de carbohidratos: ", skewness(carbohidratos),
"\n")

## Coeficiente de sesgo de carbohidratos:  0.9074253

# Calcular el coeficiente de curtosis para los carbohidratos

cat("Coeficiente de curtosis de carbohidratos: ",
kurtosis(carbohidratos))

## Coeficiente de curtosis de carbohidratos:  4.357538
```

En base al coeficiente de sesgo y al de curtosis, se puede apreciar que la distribución de los datos posee un nivel de sesgo positivo (a la derecha) medianamente

significativo (0.9074253), no obstante, su nivel de curtosis es considerablemente elevado (4.357538), mientras que una distribución que sea normal, posee un coeficiente de sesgo y curtosis muy bajo y cercano a 0, por lo que el sesgo y la curtosis de la distribución en cuestión poseen valores que se encuentran alejados de lo ideal, además, dados los valores de sesgo y curtosis, la distribución de los carbohidratos resulta ser asimétrica con un sesgo a la derecha, y mayormente delgada pero con una altura considerable debido al valor elevado de curtosis, por lo que al ser una distribución con cierta asimetría, la distribución de los carbohidratos no es normal, respaldando los resultados obtenidos en el gráfico de QQplot y en los tests de normalidad.

Media, mediana y rango medio de carbohidratos

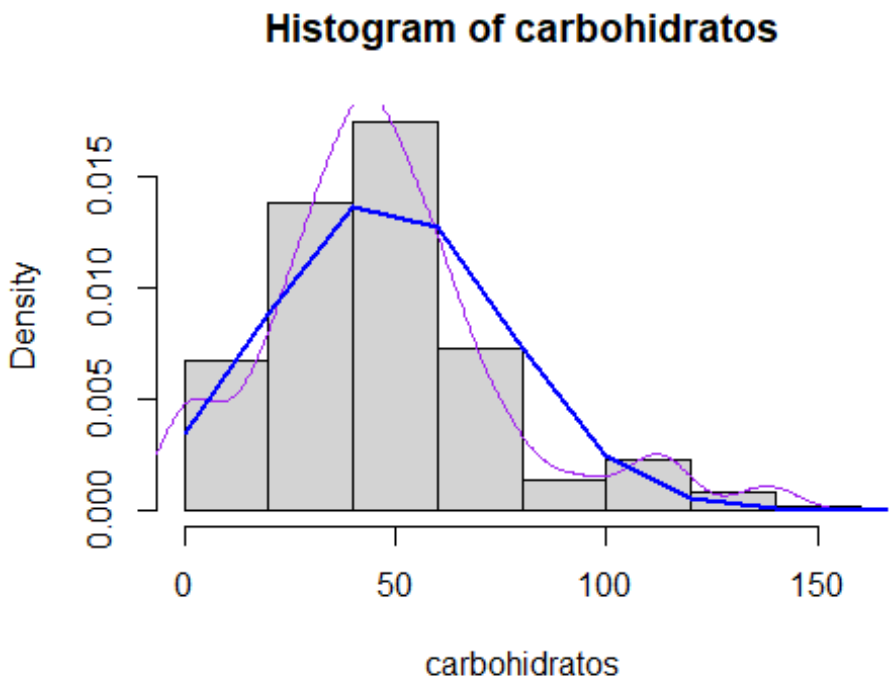
```
# Calcular media/promedio de carbohidratos

cat("Media de carbohidratos: ", mean(carbohidratos), "\n")

## Media de carbohidratos:  47.34615

# Calcular mediana de carbohidratos

cat("Mediana de carbohidratos: ", median(carbohidratos), "\n")

## Mediana de carbohidratos:  44

# Calcular rango medio de carbohidratos

cat("Rango medio de carbohidratos: ", mean(c(min(carbohidratos),
max(carbohidratos))))

## Rango medio de carbohidratos:  70.5
```

Después de calcular la media, mediana y rango medio para los datos de carbohidratos, se observa que aunque la media y la mediana tienen valores muy cercanos entre sí, el rango medio de los datos se encuentra considerablemente alejado de la media y la mediana de los mismos, por lo que en general, las 3 medidas calculadas poseen valores distintos entre sí, por lo tanto, a partir de ello, se puede concluir que la distribución de la variable carbohidratos no es normal.

Histograma y distribución teórica de carbohidratos

```
# Histograma de carbohidratos

hist(carbohidratos, freq = FALSE)

# Graficar distribución empírica (a partir de datos) y teórica (cómo
debería ser idealmente)

lines(density(carbohidratos), col = "purple") # distribución empírica
```

```r
curve(dnorm(x,mean=mean(carbohidratos), sd=sd(carbohidratos)), from = 0,
to = 2000,
      add = TRUE, col = "blue", lwd = 2) # distribución teórica
```



En histograma anterior se aprecia que la distribución empírica de los carbohidratos (gráfica color morado), presenta alejamientos mayormente significativos con respecto a la distribución teórica (gráfica color azul), además de que también en el histograma en sí mismo se aprecia que los datos de carbohidratos siguen una distribución bastante asimétrica que está muy lejos de ser normal, lo cual a su vez respalda la conclusión obtenida en gráficos y pruebas anteriores de que la variable carbohidratos no sigue distribución normal.

Influencia de los datos atípicos en la normalidad de los datos

En términos generales, los datos atípicos tienen un alto grado de influencia sobre el conjunto total de datos, esto principalmente debido a que modifican significativamente el valor tanto de la media como de la desviación estándar de los datos, introduciendo de esa manera sesgo en la distribución de los mismos y como resultado, dicha distribución se vuelve en su mayoría asimétrica con cierto sesgo, ya sea positivo, o negativo, por lo cual, al momento de llevar a cabo los tests de normalidad, al detectar variaciones en ciertos parámetros de la distribución de los datos, tales como el sesgo, los tests ya no asocian las características de la distribución en cuestión con las de una distribución normal, por lo que arrojan un valor p mayormente pequeño (menor que 0.05), lo que conduce a rechazar la hipótesis nula de la prueba de hipótesis para normalidad (la variable sigue una distribución normal)

y por consiguiente se concluye que la variable en cuestión, en este caso carbohidratos, no sigue una distribución normal.

Conclusión final sobre normalidad de carbohidratos

En conclusión, después de analizar tanto los gráficos (QQplot y distribución empírica vs teórica) como los tests de normalidad (Kolmogorov Smirnov y Jarque Bera), se concluye de manera general que la variable carbohidratos no sigue una distribución normal, puesto que en los gráficos elaborados se observa un alejamiento de los datos respecto a la recta normal en el caso del QQplot y también se aprecian alejamientos de la distribución empírica de los datos respecto a la teórica. Por otra parte, también al realizar los tests de normalidad mencionados, el valor p obtenido en ambas pruebas de normalidad fue significativamente inferior a 0.05 (nivel de significancia de las pruebas), por lo que en ambos tests se rechazó la hipótesis nula que establece que los carbohidratos siguen una distribución normal, por lo que derivado de todo lo anterior, se concluye que la variable carbohidratos no sigue una distribución normal.