Tarea2. Explorando Bases
Marcelo Márquez Murillo
A01720588

```
# Libreria
library(moments)
library(nortest)

M=read.csv("mc-donalds-menu-1.csv") #leer la base de datos
# M$variable   #para llamar una variable, aunque también la puedes leer con
corchetes cuadrados M[renglón, columna]

# Variables escogidas
calories = M$Calories
carbohydrates = M$Carbohydrates
```

1. Realiza pruebas de normalidad univariada de las variables (selecciona entre los métodos vistos en clase)

```
# Utilizamos prueba de Anderson Darling dado al tamaño de los datos
ad.test(calories)

##
##  Anderson-Darling normality test
##
## data:  calories
## A = 2.5088, p-value = 2.369e-06

ad.test(carbohydrates)

##
##  Anderson-Darling normality test
##
## data:  carbohydrates
## A = 4.1402, p-value = 2.547e-10
```

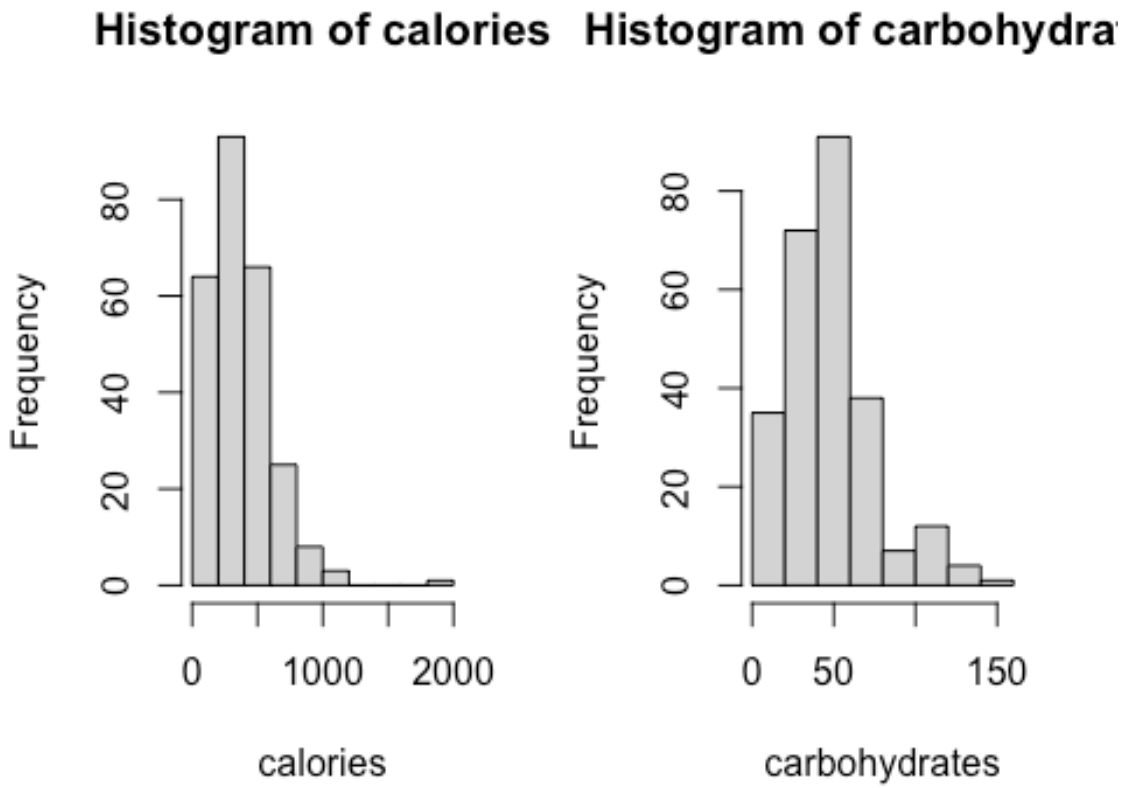
h0: Los datos se distribuyen con normalidad h1: Los datos no tienen una distribución normal

Como alpha > valor p se rechaza la hipótesis nula por lo que se concluye que los datos no siguen una distribución normal.

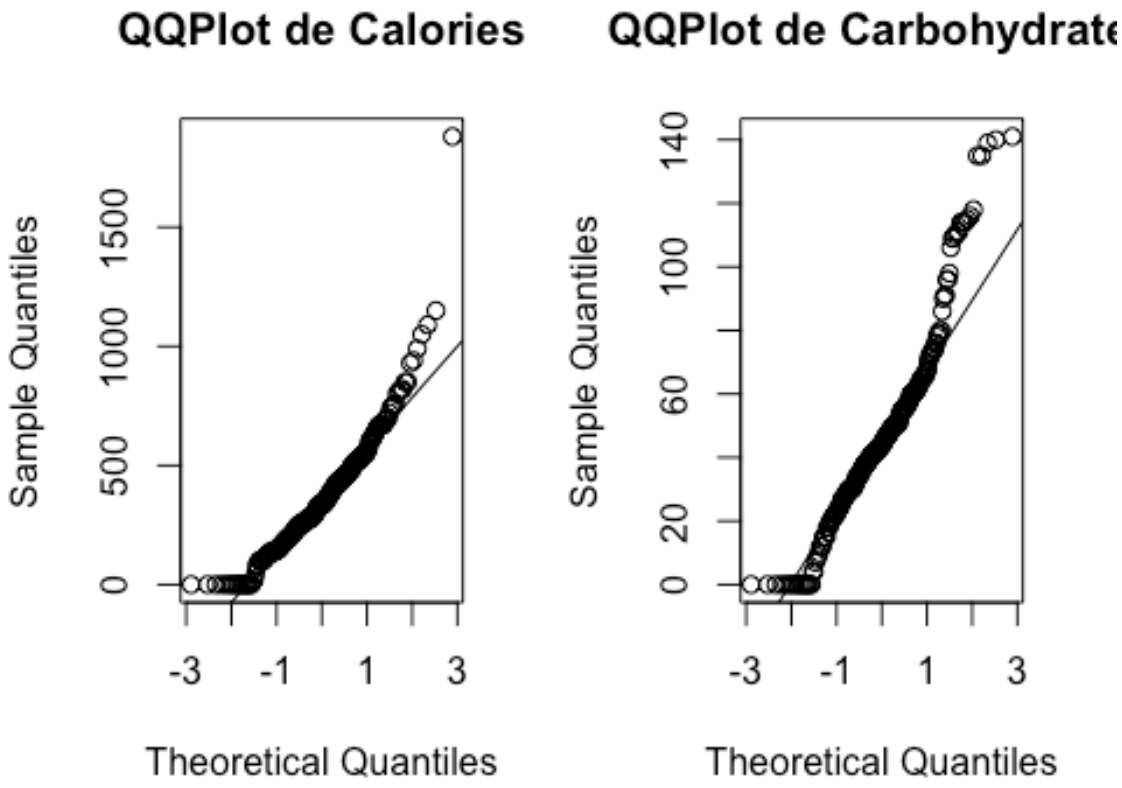
2. Grafica los datos y su respectivo QQPlot: qqnorm(datos) y qqline(datos) para cada variable

```
par(mfrow = c(1, 2))

# Grafica de Histograma
```

```r
hist(calories)
hist(carbohydrates)
```



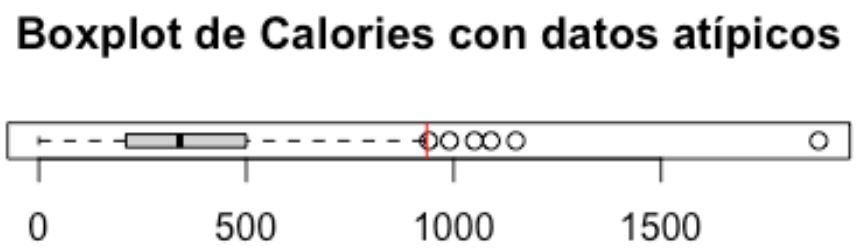

```r
# QQ-Plot de calories
qqnorm(calories, main="QQPlot de Calories")
qqline(calories)

# QQ-Plot de carbohydrates
qqnorm(carbohydrates, main="QQPlot de Carbohydrates")
qqline(carbohydrates)
```

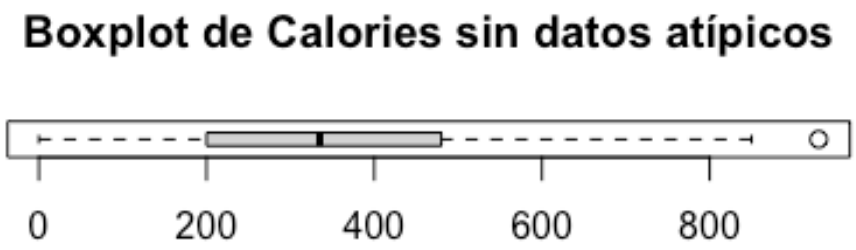


Se puede observar de los qq plots que para ambas variables la mayoría de los datos se ajustan a la línea recta por lo que se puede inferir que se tiene normalidad en los mismos. No obstante, para los datos en los extremos podemos observar que los datos se separan bastante de la línea por lo que no se tiene completamente una distribución normal.

```
par(mfrow = c(2, 1))  #Matriz de gráficos de 2x1

# Borrar datos atípicos de Calories
q1 = quantile(calories, 0.25)  #Cuantil 1 de la variable calories
q3 = quantile(calories, 0.75)  #Cuantil 3 de la variable calories
ri = IQR(calories)    #Rango intercuartílico de calories
boxplot(calories, horizontal = TRUE, main = "Boxplot de Calories con datos
atípicos")
abline(v = q3 + 1.5 * ri, col = "red")  #linea vertical en el límite de los
datos atípicos o extremos
print("Calories con datos atípicos")

## [1] "Calories con datos atípicos"

summary(calories)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   210.0   340.0   368.3   500.0  1880.0
```

```r
calories1 = M[M$Calories < q3 + 1.5 * ri, c("Calories")]  #En la matriz M, quitar datos más allá de 3 rangos intercuartílicos arriba de q3 de la variable calories
boxplot(calories1, horizontal = TRUE, main = "Boxplot de Calories sin datos atípicos")
```






```r
print("Calories sin datos atípicos")
```

```
## [1] "Calories sin datos atípicos"
```

```r
summary(calories1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   202.5   335.0   349.0   480.0   930.0
```

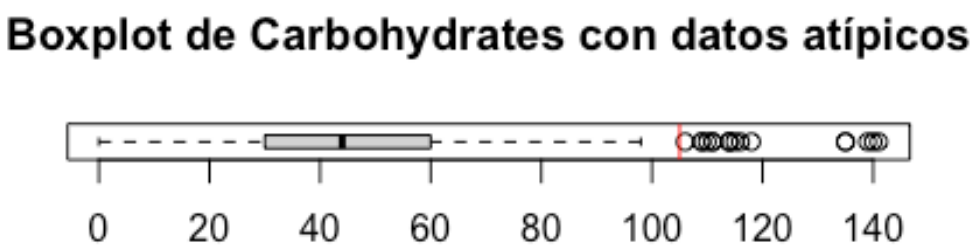
```r
# Borrar datos atípicos de Carbohydrates
q1 = quantile(carbohydrates, 0.25)  #Cuantil 1 de la variable carbohydrates
q3 = quantile(carbohydrates, 0.75)  #Cuantil 3 de la variable carbohydrates
ri = IQR(carbohydrates)    #Rango intercuartílico de carbohydrates
boxplot(carbohydrates, horizontal = TRUE, main = "Boxplot de Carbohydrates con datos atípicos")
abline(v = q3 + 1.5 * ri, col = "red")  #linea vertical en el límite de los datos atípicos o extremos
print("Carbohydrates con datos atípicos")
```

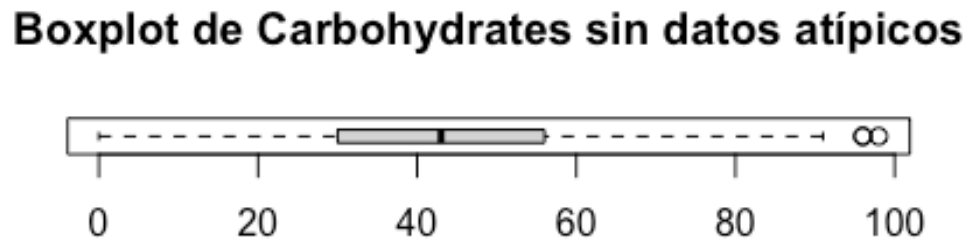
```
## [1] "Carbohydrates con datos atípicos"
```

```
summary(carbohydrates)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   30.00   44.00   47.35   60.00  141.00
```

```
carbohydrates1 = M[M$Carbohydrates < q3 + 1.5 * ri, c("Carbohydrates")]  #En la matriz M, quitar datos más allá de 3 rangos intercuartílicos arriba de q3 de la variable carbohydrates
boxplot(carbohydrates1, horizontal = TRUE, main = "Boxplot de Carbohydrates sin datos atípicos")
```






```
print("Carbohydrates sin datos atípicos")
```

```
## [1] "Carbohydrates sin datos atípicos"
```

```
summary(carbohydrates1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   30.00   43.00   42.28   56.00   98.00
```

Para ambas variables se observa que al eliminar los datos atípicos el 50% de los datos de cada variable se acomodan en el centro del rango.

3. Calcula el coeficiente de sesgo y el coeficiente de curtosis de cada variable.

```r
# Coeficiente de sesgo y curtosis de Calories
cat("Coeficiente de sesgo de calories con datos atípicos:",
skewness(calories), "\n")
```

```
## Coeficiente de sesgo de calories con datos atípicos: 1.444105
```

```r
cat("Coeficiente de curtosis de calories con datos atípicos:",
kurtosis(calories), "\n")
```

```
## Coeficiente de curtosis de calories con datos atípicos: 8.645274
```

```r
cat("Coeficiente de sesgo de calories sin datos atípicos:",
skewness(calories1), "\n")
```

```
## Coeficiente de sesgo de calories sin datos atípicos: 0.3490549
```

```r
cat("Coeficiente de curtosis de calories sin datos atípicos:",
kurtosis(calories1), "\n")
```

```
## Coeficiente de curtosis de calories sin datos atípicos: 2.716828
```

```r
# Coeficiente de sesgo y curtosis de Carbohydrates
cat("Coeficiente de sesgo de carbohydrates con datos atípicos:",
skewness(carbohydrates), "\n")
```

```
## Coeficiente de sesgo de carbohydrates con datos atípicos: 0.9074253
```

```r
cat("Coeficiente de sesgo de carbohydrates con datos atípicos:",
kurtosis(carbohydrates), "\n")
```

```
## Coeficiente de sesgo de carbohydrates con datos atípicos: 4.357538
```

```r
cat("Coeficiente de sesgo de carbohydrates sin datos atípicos:",
skewness(carbohydrates1), "\n")
```

```
## Coeficiente de sesgo de carbohydrates sin datos atípicos: -0.02861759
```

```r
cat("Coeficiente de sesgo de carbohydrates sin datos atípicos:",
kurtosis(carbohydrates1), "\n")
```

```
## Coeficiente de sesgo de carbohydrates sin datos atípicos: 2.931357
```

4. Compara las medidas de media, mediana y rango medio de cada variable.

```r
# Media, mediana y rango medio de Calories
cat("Media de Calories:", mean(calories), "\n")
```

```
## Media de Calories: 368.2692
```

```r
cat("Mediana de Calories:", median(calories), "\n")
```

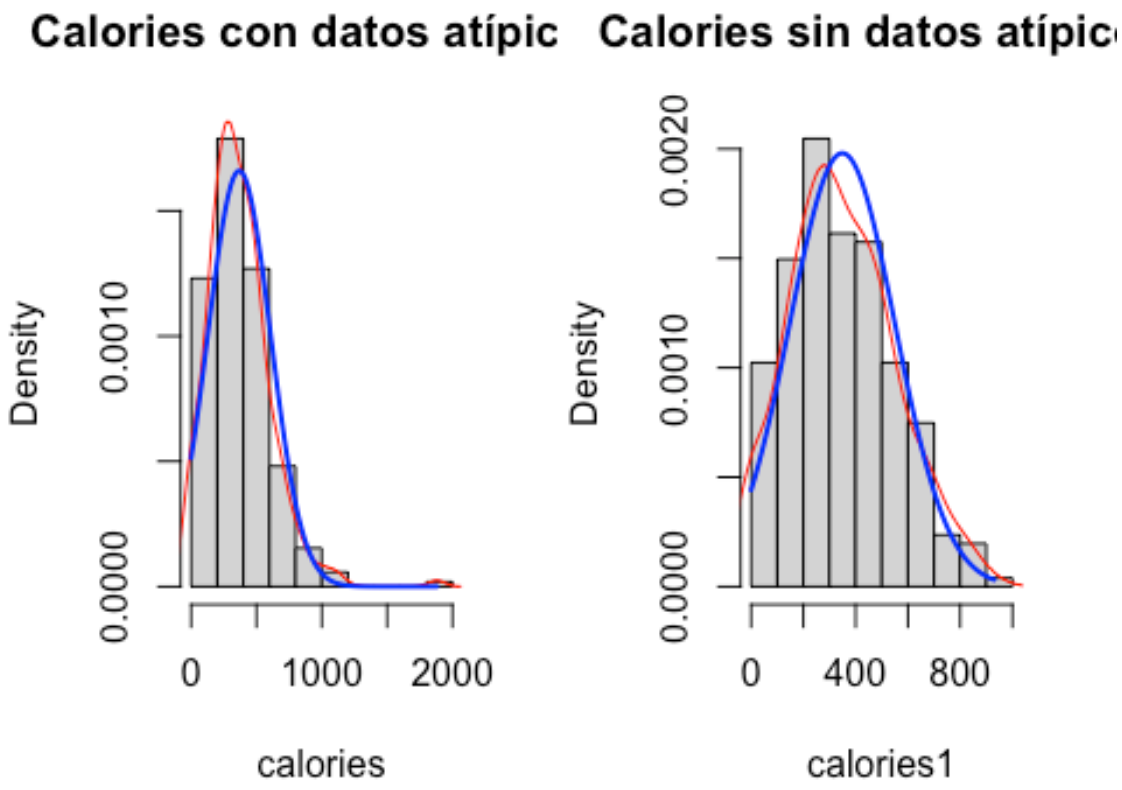
```
## Mediana de Calories: 340
```

```r
cat("Rango Medio de Calories:", IQR(calories), "\n")
```

```
## Rango Medio de Calories: 290

# Media, mediana y rango medio de Carbohydrates
cat("Media de Calories:", mean(carbohydrates), "\n")

## Media de Calories: 47.34615

cat("Mediana de Calories:", median(carbohydrates), "\n")

## Mediana de Calories: 44

cat("Rango Medio de Carbohydrates:", IQR(carbohydrates), "\n")

## Rango Medio de Carbohydrates: 30
```

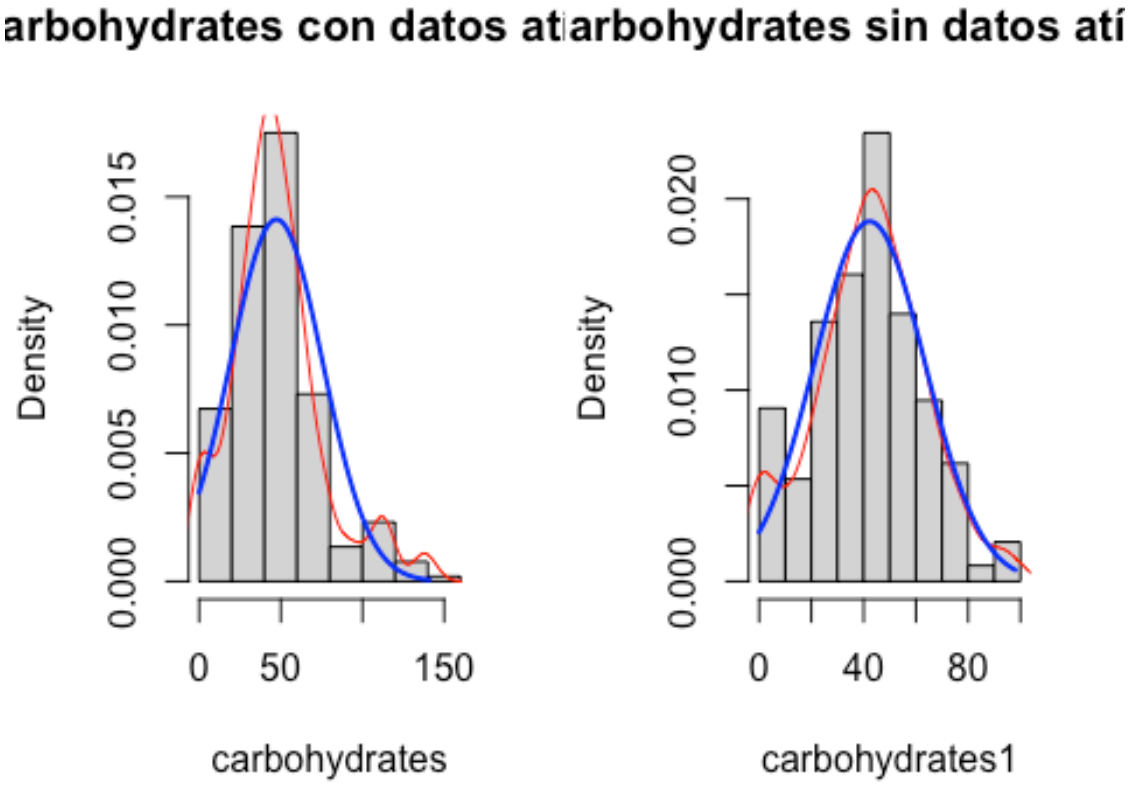
5. Realiza el histograma y su distribución teórica de probabilidad (sugerencia, adapta el código:
   - hist(datos,freq=FALSE)
     lines(density(datos),col="red")
     curve(dnorm(x,mean=mean(datos,sd=sd(datos)), from=-6, to=6, add=TRUE, col="blue",lwd=2)

```
par(mfrow = c(1, 2))

# Calories
x = seq(min(calories), max(calories), 0.1)
hist(calories, freq = FALSE, main = "Calories con datos atípicos")
lines(density(calories), col = "red")
curve(
  dnorm(x, mean = mean(calories), sd = sd(calories)),
  from = min(calories),
  to = max(calories),
  add = TRUE,
  col = "blue",
  lwd = 2
)

x = seq(min(calories1), max(calories1), 0.1)
hist(calories1, freq = FALSE, main = "Calories sin datos atípicos")
lines(density(calories1), col = "red")
curve(
  dnorm(x, mean = mean(calories1), sd = sd(calories1)),
  from = min(calories1),
  to = max(calories1),
  add = TRUE,
  col = "blue",
  lwd = 2
)
```




```r
# Carbohydrates
x = seq(min(carbohydrates), max(carbohydrates), 0.1)
hist(carbohydrates, freq = FALSE,  main = "Carbohydrates con datos atípicos")
lines(density(carbohydrates), col = "red")
curve(
  dnorm(
    x,
    mean = mean(carbohydrates),
    sd = sd(carbohydrates)
  ),
  from = min(carbohydrates),
  to = max(carbohydrates),
  add = TRUE,
  col = "blue",
  lwd = 2
)

x = seq(min(carbohydrates1), max(carbohydrates1), 0.1)
hist(carbohydrates1, freq = FALSE,  main = "Carbohydrates sin datos
atípicos")
lines(density(carbohydrates1), col = "red")
curve(
  dnorm(
```

```
    x,
    mean = mean(carbohydrates1),
    sd = sd(carbohydrates1)
  ),
  from = min(carbohydrates1),
  to = max(carbohydrates1),
  add = TRUE,
  col = "blue",
  lwd = 2
)
```



Al eliminar los datos atípicos, se puede observar que para las variables calories y carbohydrates, los datos tienen una distribución normal.

6. Comenta los gráficos y los resultados obtenidos con vías a interpretar normalidad de los datos.

En esta actividad pude observar la distribución de los datos para verificar si se cumple el supuesto de normalidad por medio de una prueba de normalidad, en este caso se realizó la prueba de Anderson Darling. Al haber rechazado la hipóteis nula y concluido que no hay normalidad en los datos, elaboré unas gráficas “QQ Plots” con las que, de manera visual podemos explorar si existe normalidad en los datos. En las gráficas se puede observar

claramente que los datos atípicos están influyendo en la normalidad de los datos para ambase variables para ambos extremos.

Posteriormente, se realizaron gráficos de caja y bigotes para observar los quanitles y eliminar los datos atípicos. De esta manera, se puede obsevar cómo al eliminar los datos atípicos los datos se para esta base se acomodaron en el centro del rango.

Al exameniar el sesgo y la curtosis se pudo observar una mejora al eliminar los datos atípicos para calories y carbohydrates. Finalmente, se utilizó el recurso del histograma para observar el comportamiento de la distribución de los datos con la linea de densidad ilustrando dicha distribución. Se observa de nuevo cómo la eliminación de datos atípicos ayudó en esta circunstancia a conseguir una distribución normal en las variables de calorías y carbohidratos.