Actividad 4: Explorando Bases

Oskar Arturo Gamboa Reyes

2024-08-13

1. Leer archivo con datos

```
M=read.csv("mc-donalds-menu.csv")
```

2. Analizar variables

```
cal = M$Calories
sug = M$Sugars
```

a) Datos atípicos Calorias

```
q1=quantile(cal, 0.25)
q2=quantile(cal, 0.50)
q3=quantile(cal, 0.75)
q4=quantile(cal, 1)
ri = IQR(cal)    #Rango intercuartílico de X

print("Cuartíles")
```

```
## [1] "Cuartíles"
```

```
print(q1)
```

```
## 25%
## 210
```

```
print(q2)
```

```
## 50%
## 340
```

```
print(q3)
```

```
## 75%
## 500
```

```
print(q4)
```

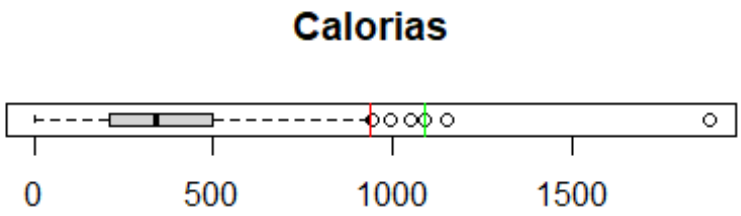
```
## 100%
## 1880
```

```
print(paste("Rango Intercuartílico: ", ri))
```

```
## [1] "Rango Intercuartílico:  290"
```

```
par(mfrow=c(2,1))  #Matriz de gráficos de 2x1
boxplot(cal,horizontal=TRUE, main="Calorias")   #y1=min en la escala del eje
```

```
Y, y2=máx en la escala del eje Y
abline(v=q3+1.5*ri, col="red")  #linea vertical en el límite de los datos
atípicos o extremos
abline(v=mean(cal)+3*sd(cal), col="green")
```

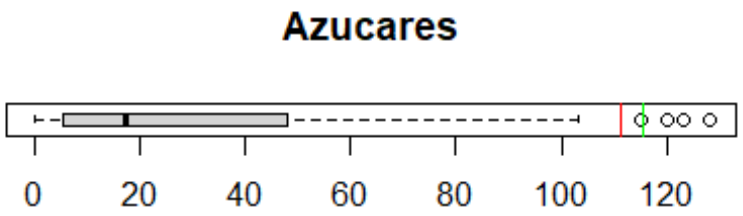


A partir de esta gráfica podemos determinar que la variable de calorías tiene algunos datos atípicos arriba de mil calorías y unos datos extremos con más de 1500 calorías. Esto puede ser ya que existen varios datos con 0 y crea un sesgo a la derecha. Desde este punto podemos ver que no va a tener una curva normal.

b) Datos atípicos Azucares

```
q1=quantile(sug, 0.25)
q2=quantile(sug, 0.50)
q3=quantile(sug, 0.75)
q4=quantile(sug, 1)
ri = IQR(sug)   #Rango intercuartílico de X

print("Cuartíles")

## [1] "Cuartíles"

print(q1)

##  25%
## 5.75

print(q2)
```

```
##  50%
## 17.5

print(q3)

## 75%
##  48

print(q4)

## 100%
##  128

print(paste("Rango Intercuartílico: ", ri))

## [1] "Rango Intercuartílico:  42.25"

par(mfrow=c(2,1))  #Matriz de gráficos de 2x1
boxplot(sug,horizontal=TRUE, main="Azucares")
abline(v=q3+1.5*ri, col="red")  #linea vertical en el límite de los datos
atípicos o extremos
abline(v=mean(sug)+3*sd(sug), col="green")
```



La gráfica de la azúcar sigue un patrón similar a las calorías, esto tiene sentido ya que una mayor cantidad de azúcar significa más calorías, sin embargo mis primeras impresiones es que tiene un poco más de variedad en los datos por lo que no tiene datos tan extremos.

Para este análisis no voy a quitar ningún dato atípico ya que estos datos altos son representativos del menú de McDonald's. En el caso de querer normalizar las curvas sería importante acotar los datos iguales a cero ya que estos no son muy representativos del menú de McDonald's ya que principalmente son aguas y refrescos sin calorías. 3. Análisis de normalidad

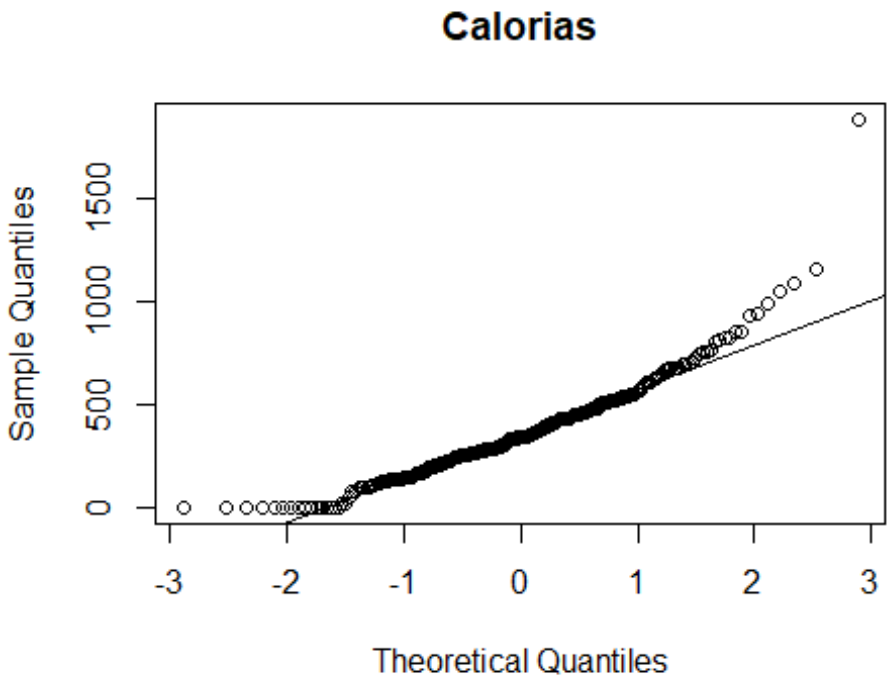
a) Pruebas de normalidad:

```
library(nortest)
ad.test(cal)

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

ad.test(sug)

##
##  Anderson-Darling normality test
##
## data:  sug
## A = 9.9899, p-value < 2.2e-16
```

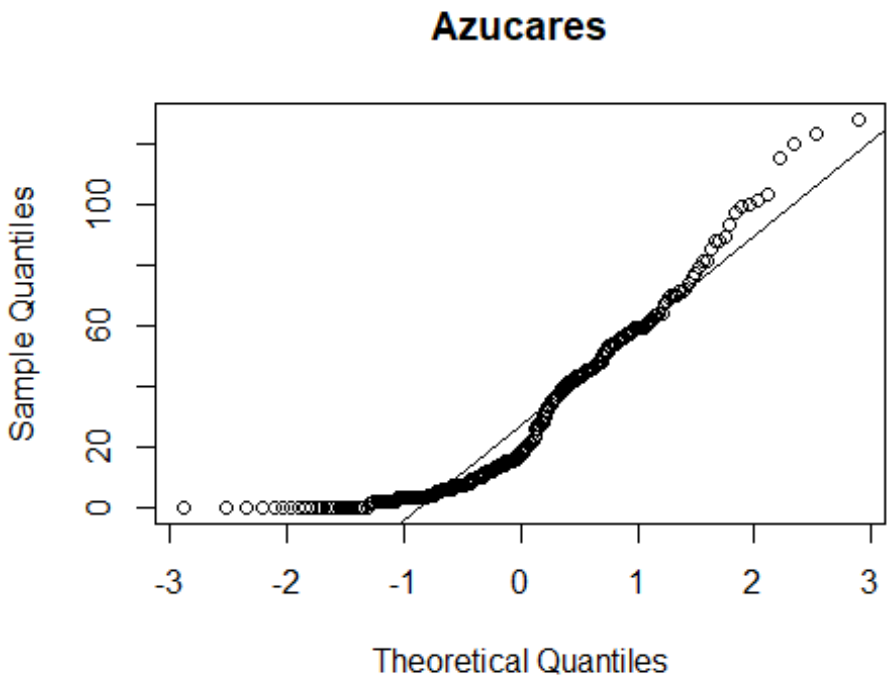
Los resultados de las pruebas de normalidad indican que los datos no siguen una curva normal, el p-value es muy pequeño en ambas pruebas.

b) Datos y QQPlot:

```
qqnorm(cal,main="Calorias")
qqline(cal)
```




```
qqnorm(sug,main = "Azucares")
qqline(sug)
```



Podemos ver que los datos no se apegan a la normalidad en las extremidades y por la forma podemos notar que hay un sesgo a la derecha en ambas variables.

c) Sesgo y Curtosis

```
library(e1071)
print("Skewness Calorias")

## [1] "Skewness Calorias"

skewness(cal)

## [1] 1.435782

print("Kurtosis Calorias")

## [1] "Kurtosis Calorias"

kurtosis(cal)

## [1] 5.5789

print("Skewness Azucares")

## [1] "Skewness Azucares"

skewness(sug)

## [1] 1.020064

print("Kurtosis Azucares")
```

```
## [1] "Kurtosis Azucares"
```

```
kurtosis(sug)
```

```
## [1] 0.460967
```

El resultado del sesgo de las dos variables indica que hay un sesgo a la derecha. Mientras que una curtosis alta indica que hay una mayor concentración de datos lo que crea una campana más pronunciada.

d)Media, mediana y rango medio

```
print("Media de Calorias")
```

```
## [1] "Media de Calorias"
```

```
mean(cal)
```

```
## [1] 368.2692
```

```
print("Mediana de Calorias")
```

```
## [1] "Mediana de Calorias"
```

```
median(cal)
```

```
## [1] 340
```

```
print("Rango Medio de Calorias")
```

```
## [1] "Rango Medio de Calorias"
```

```
IQR(cal)
```

```
## [1] 290
```

```
print("Media de Azucares")
```

```
## [1] "Media de Azucares"
```

```
mean(sug)
```

```
## [1] 29.42308
```

```
print("Mediana de Azucares")
```

```
## [1] "Mediana de Azucares"
```

```
median(sug)
```

```
## [1] 17.5
```

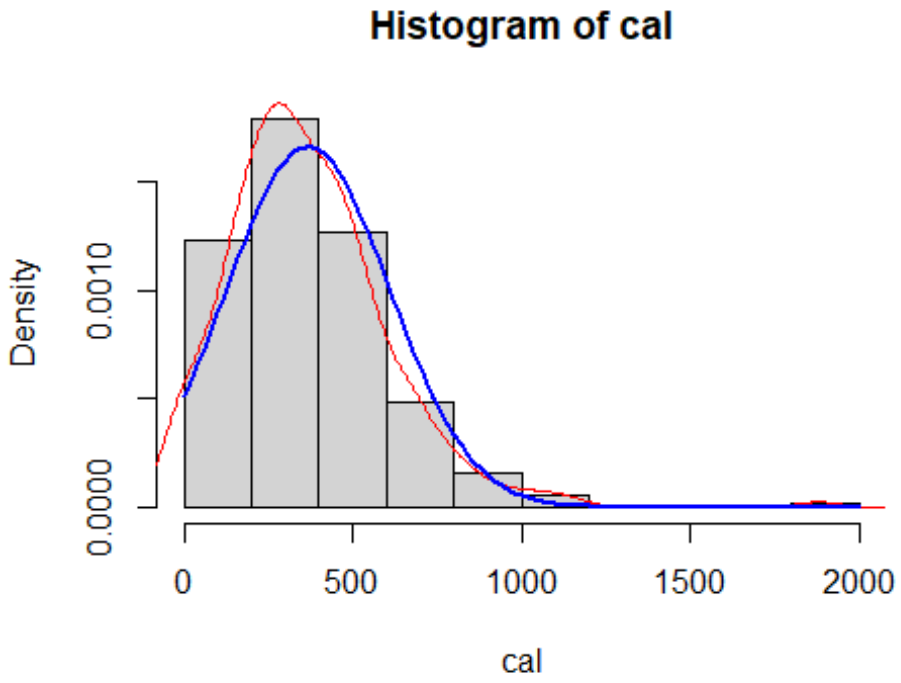
```
print("Rango Medio de Azucares")
```

```
## [1] "Rango Medio de Azucares"
```

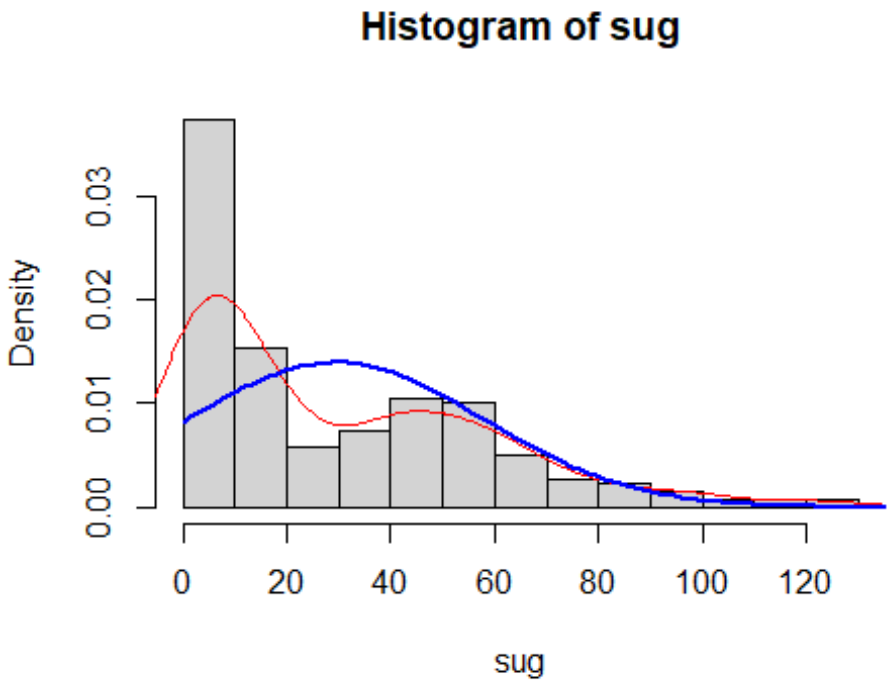
```r
IQR(sug)

## [1] 42.25
```

e) Histograma y distribución teórica de probabilidad

```r
hist(cal,freq=FALSE)
lines(density(cal),col="red")
curve(dnorm(x,mean=mean(cal),sd=sd(cal)), from=0, to=2000, add=TRUE,
col="blue",lwd=2)
```




```r
hist(sug,freq=FALSE)
lines(density(sug),col="red")
curve(dnorm(x,mean=mean(sug),sd=sd(sug)), from=0,
to=150,add=TRUE,col="blue",lwd=2)
```



Ahora que ya tenemos un histograma podemos ver el sesgo a la derecha en las dos gráficas y que tiene una gran cantidad de datos iguales a 0 lo que produce una alta curtosis. Finalmente, podemos determinar que las variables NO siguen una curva normal.